

第三方惩罚的演化与认知机制*

谢东杰 苏彦捷**

(北京大学心理与认知科学学院, 行为与心理健康北京市重点实验室, 北京, 100871)

摘要 第三方惩罚一般是指当违反社会规范的行为与自身利益无关时, 个体牺牲自我利益来惩罚违规者的行为。最近的研究发现, 对于个体而言, 第三方惩罚是一种具有适应性的信号, 包括建立良好的声誉以及威慑违规者。不同情境下第三方惩罚的功能有所差异, 它是一种特殊的利他行为。第三方惩罚依赖于多个系统, 涉及情绪反应、共情等社会认知能力以及认知控制等中央执行能力。探讨该行为的演化和认知机制也能贡献于个体、群体和人际层面的社会心理服务工作。

关键词 第三方惩罚 互惠 威慑 共情 认知控制

1 引言

社会规范 (social norm) 在促进和维护人类合作中扮演着重要角色 (Fehr & Fischbacher, 2004a)。涉及到社会规范的利他行为一般可以分为两类, 遵守规范 (norm compliance) 和维护规范 (norm enforcement) (Bendor & Swistak, 2001; Bernhard, Fischbacher, & Fehr, 2006)。公平分配这种常见的遵守规范行为在演化上的意义 (Brosnan & de Waal, 2014) 及其社会认知机制 (Wu & Su, 2014) 等方面已经得到了大量的探讨。然而, 人们为何以及如何做出维护规范行为等一系列问题还有待深入研究。

个体维护社会规范最常见的方式是惩罚。当一方违反社会规范时, 利益受损方 (自我利益卷入方) 会对违规者采取惩罚措施, 这种行为被称为第二方惩罚 (second-party punishment) (Fehr & Gächter, 2000; Jensen, Call, & Tomasello, 2007), 也就是“以牙还牙”。人类社会中还存在着第三方惩罚 (third-party punishment, TPP), 它是指当违规行为与自身利益无关时, 个体牺牲自我利益来惩罚违规者的现象 (Fehr & Fischbacher, 2004b)。这一现象普遍存

在于各个文化当中 (Henrich et al., 2006)。有研究者还会将内疚 (guilt) 视为第一方惩罚 (first-party punishment) (Nelissen & Zeelenberg, 2009b); 它是一种“自我惩罚”, 指的是个体真实或假想的行为违背自身标准后产生的负性情绪体验 (Eisenberg, 2000)。

第二方惩罚和第三方惩罚都是他人对违规者做出的惩罚行为, 均需个体牺牲自我利益来维护社会规范, 并且对群体中的大多数其他人有利, 因此研究者将它们统称为利他性惩罚 (altruistic punishment) (Fehr & Fischbacher, 2003; Raihani & Bshary, 2015)。在这两类利他性惩罚中, 第三方惩罚尤其受到研究者的关注。一方面是因为在维护社会规范方面, 第三方惩罚比第二方惩罚更加有效 (廖玉玲, 洪开荣, 张亮, 2015)。很多时候违规者只侵犯了少数人的利益, 甚至有时并没有直接的利益受损方, 若社会中只存在第二方惩罚, 那么惩罚违规者的比例将会很低, 因此第二方惩罚能够维护的社会规范种类十分有限, 而第三方惩罚则能扩大维护规范的范围 (Fehr & Fischbacher, 2004b)。另一方面, 相比于利益直接受损方做出的第二方惩罚, 第三方

* 本研究得到国家自然科学基金面上项目 (31872782, 31571134) 的资助。

** 通讯作者: 苏彦捷。E-mail: yjsu@pku.edu.cn

DOI:10.16719/j.cnki.1671-6981.20190132

惩罚是与施害者、受害者都无关的中立个体做出的行为,从演化的角度来看,该行为对于个体而言具有何种适应性更加难以解释。

2 第三方惩罚是一种具有适应性的信号

自然选择理论(natural selection theory)很难解释这一行为是如何演化的(Pennisi, 2005)。有研究者基于间接互惠理论(Nowak & Sigmund, 1998),提出第三方惩罚对于个体而言是一种具有适应性的信号(陈欣,赵国祥,叶浩生,2014; Jordan, Hoffman, Bloom, & Rand, 2016; Krasnow, Delton, Cosmides, & Tooby, 2016)。信号的内容包括两个方面:观察者知觉到的信号为第三方惩罚者富有公德心、值得他人信赖;而违规者知觉到的信号则是,第三方惩罚者不是“软柿子”、不可“搭其便车”。

2.1 可信度假说

高成本信号理论(costly signaling theory)认为,行动者会用高成本行为来向他人暗示自己的个人品质,而观察者则将高成本行为视为一种可靠的暗示信号(Camerer, 2003; Spence, 1974)。第三方惩罚是一种高成本的利他行为(Baumard, André, & Sperber, 2013; Gintis, Bowles, Boyd, & Fehr, 2003),行动者可以用它来暗示自己拥有公平公正的积极品质,而观察者则能通过这一行为来推测行动者是否值得信赖、能否作为未来的合作伙伴。因此,第三方惩罚者比不作为者在后续的社会交往中更有可能获得优势。

Jordan 等人(2016)使用演化计算建模和匿名互动游戏相结合的方法对可信度假说进行了验证。他们让一部分被试(行动者)先扮演惩罚者参与第三方惩罚游戏(third-party punishment game, TPPG),然后扮演信托人(trustee)参与信任游戏(trust game, TG);而另一部分被试(观察者)在TPPG中只需要观察行动者的选择(是否惩罚),然后扮演投资人(investor)在TG中决定给不同的信托人分别投资多少代币。结果发现,观察者会给第三方惩罚者投资更多代币,表明他们认为第三方惩罚者的可信度高于不作为者;而且,第三方惩罚者也确实更加值得他人信赖,因为他们在TG中会返还给观察者更多代币(Jordan et al., 2016)。

然而,第三方惩罚并非暗示个体可信度的最佳信号。它的利他性(altruism)在只有“惩罚违规者”和“不作为”的有限选项时才能体现出来(Raihani & Bshary, 2015)。当个体还有“帮助受害者”的选

项时,第三方惩罚的利他性将被弱化(Jordan et al., 2016),个体可能将之视为一种“损人不利己”的行为,而帮助或慷慨才是暗示个体可信度更加有效的信号(Przepiorka & Liebe, 2016)。

2.2 威慑假说

“威慑假说(the deterrence hypothesis)”则认为,第三方个体会基于违规者当下的行为推测出他(或她)在未来可能也会伤害自己或重要他人,出于对违规者的威慑,第三方个体会做出惩罚行为(Krasnow et al., 2016)。虽然标准的第三方惩罚游戏只进行一次,而且与惩罚者交互的对象都是匿名的,但是从演化的视角来看,人类生活在规模较小的世界里(Dunbar, 1993),那么个体很有可能与陌生人再次发生面对面的社会交互活动(Krasnow, Delton, Tooby, & Cosmides, 2013)。因此,第三方个体惩罚他人的违规行为虽然在当下看似有损自身利益,但是却可以阻止违规者在未来不伤害自己以及重要他人,尤其是在高强度连接性和低水平流动性的情境下(Roos, Gelfand, Nau, & Carr, 2014)。

Krasnow 等人(2016)发现,第三方个体在不知道分配者未来是否会伤害自己时,他们会基于分配者此时对他人的自私分配,推测出分配者在未来也很有可能对自己进行自私分配,并且第三方个体的这一推测能够预测他们此时对分配者的惩罚行为;但是,如果第三方个体此时有机会知道分配者未来将对自己进行公平分配,那么他们此时惩罚分配者的倾向则会有所减弱。这说明,第三方惩罚者很有可能是为了向未来的合作搭档释放警告和威慑的信号——惩罚者不接受自私等违反社会规范的行为。

2.3 小结

虽然这两个假说对于第三方惩罚这一利他行为的信号内容进行了不同的解释和验证,但是它们可以整合到一个理论框架中。人类合作的演化模型可以分为两个部分,搭档选择(partner choice)和搭档控制(partner control)(Baumard et al., 2013)。在搭档选择的子模型中,人际交互范围较大,旁观者会偏爱第三方惩罚者而非不作为者,将其考虑为未来的合作搭档(Jordan et al., 2016),此时第三方惩罚这一信号的受众主要是行为发生时周围众多的观察者。而在搭档控制的子模型中,人际交互范围较小,第三方个体很有可能与违规者发生直接的社会交互活动,因此,个体希望以第三方惩罚来警示违规者不要损害自己和重要他人的利益(Delton &

Krasnow, 2017; Krasnow et al., 2016), 此时第三方惩罚这一信号的受众主要是违规者。

因此, 我们认为, 行动者通过第三方惩罚既可以向观察者暗示自己的可信度, 也可以威慑违规者不要侵犯自己和重要他人的利益, 具体的作用取决于行为发生的具体情境, 未来的研究还需要进一步操纵情境(如人际交互范围)来检验这两个假说的适用条件。

3 影响第三方惩罚的因素

第三方惩罚的演化意义说明了它是一种特殊而复杂的利他行为。不同情境下其功能有所差异, 并且仅在特定情境下它才会被知觉为利他行为。它的复杂性也体现在行为的发生依赖于多个系统, 包括内疚和愤怒等情绪系统、共情等社会认知系统以及认知控制等中央执行系统。当然, 情境也会影响第三方惩罚发生的比例和过程, 这提示第三方惩罚与帮助等其他利他行为之间也存在相似之处, 具有一定的狭隘性(parochialism)(Bernhard et al., 2006)。

3.1 共情

共情(empathy)作为利他行为的主要驱动力之一(de Waal, 2008), 在特定情境下对于促进第三方惩罚的发生可能有重要作用(李佳, 蔡强, 黄禄华, 王念而, 张玉玲, 2012)。共情一般是指分享和理解他人的感受并对他人的处境做出适当反应的能力(Decety, Bartal, Uzefovsky, & Knafo-Noam, 2016)。违规行为的受害者会引发第三方个体的共情(Ciaramidaro et al., 2018), 因此他们会干预(帮助受害者或惩罚违规者)与自己利益无关的违规情况(Liu, Li, Zheng, & Guo, 2017)。当第三方个体同时拥有“帮助受害者”和“惩罚违规者”的选项时, 共情关注(empathic concern)较强的个体会更多地选择帮助受害者, 更少地惩罚违规者(Hu, Strang, & Weber, 2015)。这意味着第三方惩罚与共情关注之间的关系可能不像帮助等其他利他行为与共情关注之间的关系那样简单: 高共情关注的个体更有可能预期自己若未能帮助受害者或维护正义将产生内疚情绪(Tangney, 1991), 因而更倾向于帮助受害者而非惩罚违规者; 当只有“惩罚违规者”和“不作为”的有限选项时, 高共情关注的个体才更有可能做出第三方惩罚。

有研究者使用了催产素(oxytocin)来进一步

探讨第三方惩罚(Krueger et al., 2013)。他们给予被试外源性催产素, 发现催产素选择性地提高了第三方个体对受害者被害程度的评价, 但是并未影响第三方个体惩罚违规者的意愿。这可能是因为催产素与共情等社会认知能力紧密相关(尚思源, 苏彦捷, 2016), 它提高了第三方个体的共情关注水平, 使之更加关注受害者的苦难和需求(Bartz et al., 2010)而非关注违规者的反社会性。这一结果也进一步说明, 第三方惩罚与慷慨等其他利他行为有所不同(Pornpattananangkul, Zhang, Chen, Kok, & Yu, 2017), 提升个体的催产素水平并不一定会提高其第三方惩罚倾向(Stallen et al., 2018)。

3.2 认知控制

然而, 共情并不足以解释所有个体的选择, 共情水平相似的个体也会做出不同的选择(Liu et al., 2017)。第三方惩罚还需要认知控制的参与(丁毅, 纪婷婷, 陈旭, 2012; Friehe & Schildberg-Hörisch, 2018; Glass, Moody, Grafman, & Krueger, 2016)。认知控制是指个体控制自己的思想和行为以实现目标的能力(Diamond, 2013)。它可能从两个方面来影响第三方个体的选择。一方面, 避免自我利益受损和维护社会规范之间存在认知冲突, 个体需要牺牲自身利益才能惩罚违规者或帮助受害者(Chavez & Bicchieri, 2013; Fehr & Fischbacher, 2003; Jordan et al., 2016), 而认知控制可以帮助个体抑制自利倾向。第三方个体即使拥有惩罚违规者或帮助受害者的利他动机, 倘若无法抑制自利倾向, 仍然可能选择不作为从而避免自我利益受损。

另一方面, 认知控制可能会调节共情与第三方惩罚之间的关系。认知控制较强的个体更能够关注他人的想法和感受, 而认知控制较弱的个体则可能沉浸于自己的消极情绪之中, 反而无法做出第三方惩罚、帮助等具有他人指向性的利他行为(黄翥青, 苏彦捷, 2012; Eisenberg, Duckworth, Spinrad, & Valiente, 2014)。一些使用人际反应指针(interpersonal relation index, IRI)测量共情的研究发现, 个人悲伤(personal distress)子维度与利他行为之间存在负相关(Eisenberg & Fabes, 1990)。这些证据在一定程度上提示我们, 缺少认知控制的共情可能无法顺利转化为第三方惩罚等利他行为; 而且在有机会帮助受害者的情况下, 有认知控制参与的共情可能促进第三方个体帮助受害者而非惩罚违规者。

3.3 情境

除了个体能力, 情境因素也会影响第三方惩罚发生的比例。通过操纵群体身份 (group membership), Delton 和 Krasnow (2017) 发现, 第三方在外群体成员 (分配者) 对内群体成员 (接受者) 做出自私分配时惩罚意愿最强; 其次是外群体成员对外群体成员做出自私分配以及内群体成员对内群体成员做出自私分配的情况; 当内群体成员对外群体成员做出自私分配时, 第三方个体的惩罚意愿最弱。还有研究者考察了匿名 (anonymity) 程度对第三方惩罚发生比例的影响。相比于完全匿名 (观察者不知道被试的决策也不知道其姓名) 的条件, 在半匿名 (观察者只知道被试的决策但不知道其姓名) 或不匿名 (观察者既知道被试的决策也知道其姓名) 的条件下, 第三方个体更倾向于惩罚违规者 (Piazza, 2008)。

情境还会影响第三方惩罚发生的过程。相比于责任集中情境 (只有被试作为第三方), 责任分散情境 (有其他第三方个体在场) 不仅会减弱被试做出第三方惩罚的行为倾向, 还会降低他们面对违规情况时前脑岛 (anterior insula, AI) 的激活程度, 提高与推测他人心理状态有关的腹内侧前额叶 (ventromedial prefrontal cortex, vmPFC)、楔前叶 (precuneus) 和背内侧前额叶 (dorsomedial prefrontal cortex, dmPFC) 的激活程度 (Feng et al., 2016)。这说明, 责任分散的情境下, 第三方个体可能不仅会评价违规行为的严重程度, 还会推测他人是否会做出第三方惩罚, 从而决定自己是否惩罚违规者。关于情境影响第三方惩罚的证据都一致地表明, 第三方惩罚作为利他行为, 也会表现出一定的狭隘性 (Bernhard et al., 2006), 与其他利他行为之间存在共性 (Levine, Prosser, Evans, & Reicher, 2005)。

3.4 小结

面对违规行为, 个体往往需要结合违规者的行为意图和行为结果来综合考虑违规者在多大程度上应当受到指责 (Buckholtz & Marois, 2012)。即使第三方个体认为违规者应当受到指责, 也不一定会做出第三方惩罚, 因为共情、认知控制等个体能力以及违规者或受害者的群体身份、匿名性等情境因素都会影响最终第三方个体的行为。第三方惩罚不像帮助等其他利他行为具有普遍认可的利他性, 它的利他性仅在第三方个体只有惩罚和不作为的选项时才能体现出来。因此, 个体选择第三方惩罚的自动

化程度可能与其他利他行为有所不同, 它不完全是自动化的、受到内疚等情绪驱动的启发式反应。

Krueger 和 Hoffman (2016) 结合已有的认知神经科学证据, 提出了第三方惩罚的神经心理框架: 首先, 前脑岛、前扣带回 (anterior cingulate cortex, ACC) 和杏仁核 (amygdala) 会参与个体面对违规行为时的情绪反应, 后扣带回 (posterior cingulate cortex, PCC) 和颞顶联合区 (temporoparietal junction, TPJ) 会参与个体对违规行为的意图判断; 之后内侧前额叶 (medial prefrontal cortex, mPFC) 会整合两部分的信息, 判断违规者是否应该受到指责, 形成“指责信号”; 然而, “指责信号”要转变为实际的惩罚行为还需要依赖于背外侧前额叶 (dorsolateral prefrontal cortex, dlPFC) 和后顶叶 (posterior parietal cortex, PPC)。

4 研究展望

第三方惩罚是一种特殊而复杂的利他行为, 其利他性的体现依赖于情境中第三方惩罚是否为个体唯一的利他选择。未来可以进一步考察第三方惩罚的自动化程度与其他利他行为的差别。例如, 通过比较做出第三方惩罚的反应时和做出帮助等其他利他行为的反应时 (Rand, Greene, & Nowak, 2012) 之间的差异, 来揭示第三方惩罚这一利他行为的特殊性以及它在不同情境下的适应性。

研究者还可以借助多种技术手段深入理解第三方惩罚的行为动机和发生机制, 例如, 通过眼动和脑成像技术来区别由内疚 (Nelissen & Zeelenberg, 2009a) 或愤怒 (Fehr & Gächter, 2002) 情绪诱发的第三方惩罚之间的异同, 通过事件相关电位 (event related potential, ERP) 和时频分析 (time-frequency analysis) 来考察个体在进行第三惩罚决策时的动态过程 (Wang, Jing, Zhang, Lin, & Valadez, 2017)。

5 政策建议

在“加强社会心理服务体系建设” (习近平, 2017) 的过程中, 心理学研究人员可以从个体、群体和人际三个层面进行社会心理服务工作 (俞国良, 谢天, 2018), 将心理学研究应用到社会治理 (social governance) (辛自强, 2018; 杨玉芳, 郭永玉, 2017) 等实际问题的解决当中。在个体层面, 第三方惩罚的认知机制提示我们, 个体需要拥有较高的能力素质, 才能更好地做出第三方惩罚等利他

行为。因此,在进行思想道德教育时,除了传授道德规范、核心价值观等知识之外,还可以增加提升个体基本认知能力(如抑制控制)和社会认知能力(如共情、观点采择)的训练内容(Schonert-Reichl et al., 2015),从而促进利他行为的实现,减少“知行不合一”(knowledge-behavior gap)的现象(Blake, McAuliffe, & Warneken, 2014)。

在群体层面,考虑到特定的情境可以促进第三方惩罚等利他行为的发生,那么政策制定者可以考虑设置引导机制,例如借助媒体宣传中华民族的群体身份来减弱利他行为的狭隘性,宣传核心价值观,引导积极的社会心态(俞国良,谢天,2018),培养公民的社会责任感(陈思静,马剑虹,2011),从而提高公民做出第三方惩罚等利他行为的倾向性。政策制定者也可以设置过滤机制来营造公正平等的社会环境。例如,考虑到第三方惩罚等利他行为具有一定的狭隘性,那么在社会公共服务中应当减少不必要信息(如群体身份、民族等)的混入,从而提升社会心理服务的公正性。

在人际层面,第三方惩罚的演化研究提示我们,未来可以借助云技术(cloud technology),将第三方惩罚、帮助等利他行为作为衡量行动者可信度的动态指标之一;同时,第三方也可以参与标记他人的违规行为。政府等机构还可以建立公开、公正、合法的信用管理系统,在更加频繁的社会交互活动中优化决策,扬善惩恶,防范失信行为,促进社会信任与合作(刘国芳,辛自强,2014)。

参考文献

- 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》,3,670-675.
- 陈欣,赵国祥,叶浩生.(2014).公共物品困境中惩罚的形式与作用.《心理科学进展》,22(1),160-170.
- 丁毅,纪婷婷,陈旭.(2012).利他惩罚的发生机制及其神经基础.《心理发展与教育》,6,658-664.
- 黄嵩青,苏彦捷.(2012).共情的毕生发展:一个双过程的视角.《心理发展与教育》,28(4),434-441.
- 李佳,蔡强,黄禄华,王念而,张玉玲.(2012).利他惩罚的认知机制和神经生物基础.《心理科学进展》,20(5),682-689.
- 廖玉玲,洪开荣,张亮.(2015).第三方惩罚机制与双边合作秩序的维持——来自房地产征用补偿的实验证据.《系统工程理论与实践》,35(11),2798-2808.
- 刘国芳,辛自强.(2014).惩罚对信任与合作的影响:争论与解释.《上海师范大学学报(哲学社会科学版)》,43(1),146-152.
- 尚思源,苏彦捷.(2016).催产素系统与社会行为——催产素及其受体基因的作用机制.《心理技术与应用》,4(4),224-235.
- 习近平.(2017-10-28).决胜全面建成小康社会 夺取新时代中国特色社会主义伟大胜利——在中国共产党第十九次全国代表大会上的报告(2017年10月18日).《人民日报》,1-5.
- 辛自强.(2018).社会治理中的心理学问题.《心理科学进展》,26(1),1-13.
- 杨玉芳,郭永玉.(2017).心理学在社会治理中的作用.《中国科学院院刊》,32(2),107-116.
- 俞国良,谢天.(2018).社会转型 社会心理服务与社会心态培育.《河北学刊》,38(2),175-181.
- Bartz, J. A., Zaki, J., Bolger, N., Hollander, E., Ludwig, N. N., Kolevzon, A., & Ochsner, K. N. (2010). Oxytocin selectively improves empathic accuracy. *Psychological Science*, 21(10), 1426-1428.
- Baumard, N., André, J., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, 36(1), 59-78.
- Bendor, J., & Swistak, P. (2001). The evolution of norms. *American Journal of Sociology*, 106(6), 1493-1545.
- Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature*, 442(7105), 912-915.
- Blake, P. R., McAuliffe, K., & Warneken, F. (2014). The developmental origins of fairness: The knowledge-behavior gap. *Trends in Cognitive Sciences*, 18(11), 559-561.
- Brosnan, S. F., & de Waal, F. B. (2014). Evolution of responses to (un)fairness. *Science*, 346(6207), 314-314.
- Buckholtz, J. W., & Marois, R. (2012). The roots of modern justice: Cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, 15(5), 655-661.
- Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. New York: Princeton University Press.
- Chavez, A. K., & Bicchieri, C. (2013). Third-party sanctioning and compensation behavior: Findings from the ultimatum game. *Journal of Economic Psychology*, 39, 268-277.
- Ciaramidaro, A., Toppi, J., Casper, C., Freitag, C. M., Siniatchkin, M., & Astolfi, L. (2018). Multiple-brain connectivity during third party punishment: An EEG hyperscanning study. *Scientific Reports*, 8(1), 1-13.
- Decety, J., Barta, L., Uzevovsky, F., & Knafo-Noam, A. (2016). Empathy as a driver of prosocial behavior: Highly conserved neurobehavioral mechanisms across species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371, 20150077.
- Delton, A., & Krasnow, M. (2017). The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior*, 38(6), 734-743.
- de Waal, F. (2008). Putting the altruism back into altruism: The evolution of empathy. *Annual Review of Psychology*, 59(1), 279-300.
- Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, 64, 135-168.
- Dunbar, R. I. M. (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16, 681-694.
- Eisenberg, N. (2000). Emotion, regulation, and moral development. *Annual Review of Psychology*, 51(1), 665-697.
- Eisenberg, N., Duckworth, A., Spinrad, T., & Valiente, C. (2014). Conscientiousness: Origins in childhood? *Developmental Psychology*, 50(5), 1331-1349.
- Eisenberg, N., & Fabes, R. A. (1990). Empathy: Conceptualization, measurement,

- and relation to prosocial behavior. *Motivation and Emotion*, 14(2), 131–149.
- Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960), 785–791.
- Fehr, E., & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences*, 8, 185–190.
- Fehr, E., & Fischbacher, U. (2004b). Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2), 63–87.
- Fehr, E., & Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *The Journal of Economic Perspectives*, 14, 159–181.
- Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868), 137–140.
- Feng, C., Deshpande, G., Liu, C., Gu, R., Luo, Y., & Krueger, F. (2016). Diffusion of responsibility attenuates altruistic punishment: A functional magnetic resonance imaging effective connectivity study. *Human Brain Mapping*, 37(2), 663–677.
- Friehe, T., & Schildberg-Hörisch, H. (2018). Predicting norm enforcement: The individual and joint predictive power of economic preferences, personality, and self-control. *European Journal of Law and Economics*, 45(1), 127–146.
- Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior*, 24(3), 153–172.
- Glass, L., Moody, L., Grafman, J., & Krueger, F. (2016). Neural signatures of third-party punishment: Evidence from penetrating traumatic brain injury. *Social Cognitive and Affective Neuroscience*, 11(2), 253–262.
- 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.
- Hu, Y., Strang, S., & Weber, B. (2015). Helping or punishing strangers: Neural correlates of altruistic decisions as third-party and of its relation to empathic concern. *Frontiers in Behavioral Neuroscience*, 9, 24.
- Jensen, K., Call, J., & Tomasello, M. (2007). Chimpanzees are vengeful but not spiteful. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 13046–13050.
- Jordan, J., Hoffman, M., Bloom, P., & Rand, D. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, 530(7591), 473–473.
- Krasnow, M. M., Delton, A. W., Tooby, J., & Cosmides, L. (2013). Meeting now suggests we will meet again: Implications for debates on the evolution of cooperation. *Scientific Reports*, 3, 1747.
- Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological Science*, 27(3), 405–418.
- Krueger, F., & Hoffman, M. (2016). The emerging neuroscience of third-party punishment. *Trends in Neuroscience*, 39(8), 499–501.
- Krueger, F., Parasuraman, R., Moody, L., Twieg, P., de Visser, E., McCabe, K., et al. (2013). Oxytocin selectively increases perceptions of harm for victims but not the desire to punish offenders of criminal offenses. *Social Cognitive and Affective Neuroscience*, 8(5), 494–498.
- Levine, M., Prosser, A., Evans, D., & Reicher, S. (2005). Identity and emergency intervention: How social group membership and inclusiveness of group boundaries shape helping behavior. *Personality and Social Psychology Bulletin*, 31(4), 443–453.
- Liu, Y., Li, L., Zheng, L., & Guo, X. (2017). Punish the perpetrator or compensate the victim? Gain vs. loss context modulate third-party altruistic behaviors. *Frontiers in Psychology*, 8, 2066.
- Nelissen, R. M. A., & Zeelenberg, M. (2009a). Moral emotions as determinants of third-party punishment: Anger and guilt and the functions of altruistic sanctions. *Judgment and Decision Making*, 4(7), 543–553.
- Nelissen, R. M. A., & Zeelenberg, M. (2009b). When guilt evokes self-punishment: Evidence for the existence of a Dobby Effect. *Emotion*, 9(1), 118–122.
- Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, 393(6685), 573–577.
- Pennisi, E. (2005). How did cooperative behavior evolve? *Science*, 309, 93–93.
- Piazza, J. (2008). The effects of perceived anonymity on altruistic punishment. *Evolutionary Psychology*, 6(3), 487–501.
- Pornpattananangkul, N., Zhang, J., Chen, Q., Kok, B. C., & Yu, R. (2017). Generous to whom? The influence of oxytocin on social discounting. *Psychoneuroendocrinology*, 79, 93–97.
- Przepiorka, W., & Liebe, U. (2016). Generosity is a sign of trustworthiness—the punishment of selfishness is not. *Evolution and Human Behavior*, 37(4), 255–262.
- Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. *Trends in Ecology & Evolution*, 30(2), 98–103.
- Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, 489(7416), 427–430.
- Roos, P., Gelfand, M., Nau, D., & Carr, R. (2014). High strength-of-ties and low mobility enable the evolution of third-party punishment. *Proceedings of the Royal Society B: Biological Sciences*, 281, 20132661.
- Schonert-Reichl, K. A., Oberle, E., Lawlor, M. S., Abbott, D., Thomson, K., Oberlander, T. F., & Diamond, A. (2015). Enhancing cognitive and social-emotional development through a simple-to-administer mindfulness-based school program for elementary school children: A randomized controlled trial. *Developmental Psychology*, 51(1), 52–66.
- Spence, M. (1974). *Market signaling*. Cambridge, Mass: Harvard University Press.
- Stallen, M., Rossi, F., Heijne, A., Smidts, A., De Dreu, C. K. W., & Sanfey, A. G. (2018). Neurobiological mechanisms of responding to injustice. *Journal of Neuroscience*, 38(12), 2944–2954.
- Tangney, J. P. (1991). Moral affect: The good, the bad, and the ugly. *Journal of Personality and Social Psychology*, 61(4), 598–607.
- Wang, Y., Jing, Y., Zhang, Z., Lin, C., & Valadez, E. (2017). How dispositional social risk-seeking promotes trusting strangers: Evidence based on brain potentials and neural oscillations. *Journal of Experimental Psychology: General*, 146(8), 1150–1163.
- Wu, Z., & Su, Y. (2014). How do preschoolers' sharing behaviors relate to their theory of mind understanding? *Journal of Experimental Child Psychology*, 120, 73–86.

The Evolutionary and Cognitive Mechanisms of Third-Party Punishment

Xie Dongjie, Su Yanjie

(School of Psychological and Cognitive Sciences, Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, 100871)

Abstract Third-party punishment (TPP) typically happens when uninvolved bystanders sacrifice their self-interests to punish social norm violators. Many studies have found convergent evidence that TPP serves as the fundamental mechanism to enable human with genetic heterogeneity to engage in intense cooperation, where individuals would otherwise be tempted to cheat. So far, however, how TPP has actually evolved and its underpinning cognitive mechanisms are still unresolved. Nature selection theory could hardly explain why individuals as third-parties would punish violators because it reduces their own fitness. While indirect reciprocity theory deals with interpersonal dynamics. And recent studies based on it empirically demonstrated that third-party punishers have higher levels of within-group fitness in two ways. First, TPP is a costly signal of trustworthiness in the partner-choice model. Third-party punishers are trusted more by observers than those who do not punish norm violators. Meanwhile, they are indeed more trustworthy than non-punishers. Second, TPP also functions as a deterrence to cheaters in the partner-control model. Third-parties are driven to punish norm violators in order to protect their self-interests by their inference about how norm violators will treat them in the future, and norm violators may be more deterred by third-party punishers than by non-punishers. The exact function of TPP is dependent on the specific context where TPP happens. In particular, if TPP happens in the context of large-scale interpersonal interaction, it functions as a signal of trustworthiness, and as a signal of deterrence in the context of small-scale interpersonal interaction.

According to our literature review, we found that TPP is a complex altruistic behavior involving multi-level systems, including emotional systems like guilt, socio-cognitive systems like empathy, and central executive systems like cognitive control. The relationship between empathy and TPP is not as simple as we commonly assumed, i.e., higher empathy is associated with more TPP, but is dependent on context. People with higher levels of empathy tend to help victims if they have the choice to help; instead, if to punish violators is the only choice to enforce social norms, they are inclined to do so. This illustrates how TPP is a special kind of altruistic behaviors. Moreover, cognitive control plays an important role in imposing TPP. First, with the help of cognitive control to inhibit selfishness, individuals as third-parties would be able to resolve the cognitive conflict between maximizing self-interests and sacrificing one's own interest to enforce norms. Second, cognitive control might modulate the relationship between empathy and TPP, particularly in regulating personal distress, a dimension of empathy negatively correlated with altruism. It is expected to be of great help to understand and solve the practical problems of social governance if we find more what contributes to TPP. Considering TPP is a special kind of altruistic behaviors, future studies could investigate how TPP is different from other kind of altruistic behaviors such as help. Another issue is that the underlying motivations of TPP evoked by guilt and anger may be different, and thus future studies could employ various techniques like eye-tracking, EEG, fMRI, to differentiate them in order to make contribution to the understanding of TPP.

Key words third-party punishment, reciprocity, deterrence, empathy, cognitive control