中图分类号: B84　　文献标识码: A　　文章编号: 1004-8634(2014)01-0146-(07)

惩罚对信任与合作的影响: 争论与解释

刘国芳[1],辛自强[2]

(1. 北京师范大学 发展心理研究所,北京 100875; 2. 中央财经大学 社会发展学院心理学系,北京 100081)

摘　要:　对背叛者与失信者进行惩罚一直是促进社会信任与合作的重要手段,有研究显示惩罚对信任与合作的确具有促进作用。然而,越来越多的研究显示,惩罚释放了不信任或合作水平较低的信号,同时会将个体信任与合作的动机由内部的道德动机转变为外部的工具性动机,因而破坏信任与合作,这种破坏作用对信任水平较高的亲社会型个体更甚。因而,为了促进信任与合作,要谨慎使用惩罚手段,并且,应用惩罚时首先要建立惩罚的合法性,同时要考虑到信任类型、个体原初的信任水平等的影响。

关键词:　惩罚; 信任; 合作

信任与合作对人类生存与发展有着重要意义,探求促进信任与合作的方式也一直是学者关注的焦点。为了提高社会的信任与合作水平,不断有人呼吁对失信者、背叛者、见死不救者等进行惩罚。那么,惩罚是否真的可以促进信任与合作呢? 有关研究显示,惩罚对信任与合作并非是简单的促进或阻碍作用,惩罚的积极影响与消极影响均存在。那么,惩罚与信任、合作间究竟是何关系? 我们又该如何合理地使用惩罚机制以促进信任与合作呢?

一、惩罚促进信任与合作的证据及其解释

惩罚是人们促进信任与合作的常用方式之一。在日常生活中,人们会对那些失信者或背叛者进行道德上的谴责,同时也不愿意与之进行社会与经济交往。在德国、法国等国家,更是明确将见死不救等行为定为犯罪的一种。可见,为了增强社会中的信任与合作,人们经常求助于惩罚手段,研究者也的确发现惩罚是促进信任与合作的一条重要途径。[1~5] 例如,考德威尔(Caldwell)发现,在囚徒困境博弈中,对那些不进行合作的个体施加惩罚可以有效促进合作。[6] Eek 等[7] 的研究也显示,如果背叛行为会得到惩罚,则背叛行为会显著减少,与惩罚其他背叛者相比,当自己被惩罚时群体内的背叛行为下降的最为明显。不仅惩罚能够促进信任与合作,更重要的是,在漫长的进化过程中,生物体已经进化出了利用惩罚促进合作的文化。研究者还发现,人们不仅会惩罚背叛自己的人,还会作出利他主义惩罚,即惩罚背叛他人的人,即使这种惩罚对自己有一定的损失而并无任何益处,利他主义惩罚是促进合作的一个重要因素。[8,9] 除了对背叛者进行涉及经济利益的惩罚可以促进合作外,研究者也发现,在不能直接对

收稿日期: 2013-05-12
基金项目: 教育部新世纪优秀人才支持计划(NCET-10-0869); 国家社会科学基金项目(11CSH046)
作者简介: 1. 刘国芳,男,河南新乡人,北京师范大学发展心理研究所博士研究生,主要从事社会心理研究。
2. 辛自强,男,山东费县人,中央财经大学社会发展学院心理学系教授,主要从事社会心理研究。

背叛者进行经济惩罚的情景中,群体成员会对背叛者进行声誉上的惩罚,即背叛者的声誉会由于其背叛行为而受损,这种惩罚威胁同样会促进信任与合作行为。[10] 可见,惩罚的确能够起到促进信任与合作的积极作用。

针对惩罚的上述积极作用,研究者提出了不同的解释,最主要的有如下两种。一种解释是惩罚改变了博弈的收益结构,使得背叛的成本要高于其可能带来的收益,因而当惩罚存在时,人们不再具有背叛的动机。[11] 从博弈论的视角看,人际交往就是各种博弈,人们会合作还是背叛取决于两种行为的成本收益比,当背叛的收益高于成本时,人们就会选择背叛。[12] 从该视角来看,惩罚起作用的原因就在于惩罚改变了博弈的收益结构,使得背叛的成本大于收益。例如,在囚徒困境博弈中,当对方选择合作时,自己背叛时的收益(如 8 元)要大于合作时的收益(如 5 元),所以理性的个体应该进行背叛。然而,如果允许对背叛行为进行惩罚(如处以 4 元的罚金),此时,背叛时的个体收益就降为 4 元,要低于合作时的收益。由于惩罚使得背叛不再是优势反应,因而理性的个体会选择合作。在通过破坏背叛者的声誉进行惩罚的情景中,该解释同样适用。在这种情景中,尽管不能够对背叛者进行直接的经济惩罚,但是背叛者声誉上的损失会降低其未来得到帮助或合作的机会,会降低其未来的收益。[10] 因而,理性的个体应该认识到这一点,并选择进行合作。

惩罚促进信任与合作的另一种解释是从进化论的角度提出的。在进化心理学家与进化生物学家看来,合作是经过自然选择的进化稳定策略,在进化过程中能够战胜其他的策略而被个体普遍采用,具有巨大的生存意义。[13~15] 而要维持信任与合作,惩罚是一个重要的、甚至是关键性的因素。[1,16,17] 例如,阿克塞尔罗德(Axelrod)曾模拟了不同交往策略的进化适应性,发现当个体间可以进行重复交往时,具有惩罚威胁的“以牙还牙”策略是进化稳定策略。[18] “以牙还牙”指的是在交往中,个体复制对方上一次交往中的策略,如果对方在上一次交往中选择了背叛,在未来的交往中他将得不到别人的信任与合作,因而“以牙还牙”策略对个体的背叛行为是一种威胁。而当个体间没有机会进行重复交往时,利他主义惩罚则是促进合作的重要因素,[8] 同时,背叛行为还会损害个体的声誉,而声誉对未来交往中是否能得到他人的信任与合作是至关重要的,所以,声誉机制的存在也是一种惩罚威胁。[19] 可见,惩罚机制的存在是具有进化适应性的,正是由于惩罚能够促进信任与合作,所以才能得到进化与完善,被个体普遍采用。

尽管从博弈论和进化论的角度都可以为惩罚促进合作提供有效的解释框架,但是这两种解释都存在各自的问题。首先,博弈论建基于个体是完全理性的假设之上,假设个体能够对所需信息进行理性的加工,以作出最合理的反应。事实上,个体是有限理性的,一方面,个体的认知能力是有限的,无法把握所有的信息; 另一方面,个体往往不愿意对信息进行完全的理性计算,利他主义偏好等同样发挥着重要作用。[20] 所以,博弈论的解释并不完备。不同于博弈论,进化心理学家和进化生物学家并没有对人的理性程度等作预先假设,往往直接对各种交往策略进行进化模拟,以发现最具适应性的策略。[18] 这种研究和解释方式完全是数据驱动的,可以更加客观地证明惩罚的积极作用并作出解释。然而,这种解释具有还原论的危险,容易陷入循环论证的框架: 为什么惩罚能够促进信任与合作? 因为进化模拟的结果显示,惩罚策略是具有进化适应性的。为什么惩罚是具有进化适应性的? 因为惩罚能够促进信任与合作。

二、惩罚阻碍信任与合作的证据及其解释

尽管上述研究都揭示了惩罚对信任与合作的促进作用,但是也有研究显示惩罚会阻碍信任与合作,[21~23] 且惩罚并不能提高群体收益,博弈中表现最好的个体也不使用惩罚策略。[24] 例如,费尔(Fehr)和罗肯巴赫(Rockenbach)[25] 为了研究惩罚对信任与合作的影响,对投资博弈进行了改进。在研究中,Fehr 等设置了两种博弈情景,一种是标准的投资博弈,[26] 一种是动机条件下的投资博弈。在标准的投资博弈中,信任者向被信任者投资一定数额的金钱,被信任者会得到 3 倍的收益,并可以向信任者返还收益中任何比例的金钱; 在动机条件下,信任者向被信任者提出期望返还额,并决定是否惩罚那些不满足自己期望的被信任者。结果发现,在动机条件下,决定对被信任

者施以惩罚威胁的信任者投资额要少于不进行惩罚威胁的信任者,即他们的信任水平更低;同时,与没有受到惩罚威胁的被信任者相比,受到惩罚威胁的被信任者返还了更少的金钱,表现出了更低的合作水平。可见,对被信任者的惩罚既破坏了对方的合作意愿,又对个体自身的信任水平有消极影响。而且研究者还发现,随着信任者借助于惩罚对被信任者要求的提高,惩罚的消极影响也变得更大。[20,25] 不仅即时交往中存在的惩罚会破坏信任与合作,个体的信任与合作还会受到交往历史中惩罚经验的影响。穆德(Mulder)等[11]使用两阶段的移除惩罚范式研究了惩罚经验对以后的信任与合作的影响。在实验中,被试需要完成两阶段的公共品博弈,在第一阶段,若干被试组成一个群体,需要向公共品中捐赠一定数量的金钱,在惩罚条件下,捐赠最少的两个成员将被罚 5 欧元,无惩罚条件下则没有该威胁;在第二阶段,两种条件下都不存在惩罚威胁。结果发现,在第一阶段,惩罚条件下被试的信任与合作水平要高于无惩罚条件。然而,在第二阶段,当移除惩罚威胁时,惩罚条件下的被试的信任水平出现了明显下降,合作水平同样出现了一定程度的下降。而最初没有经历过惩罚威胁的被试的信任与合作行为没有这种变化。该现象在中国大学生身上同样存在。[27]

为什么惩罚会破坏信任与合作呢?一种解释是惩罚释放了不信任的信号或营造了敌对的氛围。[28,29] 在群体层面上,惩罚可能暗示了群体内的信任与合作水平较低,导致群体成员作出与群体规范一致的行为。在交往中,人们可以根据一些微弱的线索来估计群体水平上的行为表现和群体规范,并调整自己的行为以适应群体规范。[30] 在信任与合作情景中,被试会努力对群体层面的信任与合作水平作出估计,当惩罚存在时,被试会将其知觉为一种群体的信任或合作水平较低的信号,因而也不愿作出信任与合作行为。在个体层面上,当惩罚威胁存在时,个体会认为对方不信任自己,由于自我实现预期的存在,使得个体不愿意进行合作,并形成恶性循环。福尔克(Falk)和考思费尔德(Kosfeld)[31] 发现,当个体对交往对象的行为施加控制时,对方会表现出更低的合作水平,同时报告更多的不信任、不自主的负面情绪。实际上,惩罚也是一种对他人行为的控制策略,当个体面对惩罚威胁时,其可能的反应会受到限制,不能自主决定自身行为,在这种情况下,个体往往会进行反抗,表现出更低的信任与合作水平。

对惩罚的消极影响的另一种解释是动机转变,也就是说,惩罚将个体信任与合作的动机由内部动机转变为了外部动机,将个体的行为由伦理性、道德性考量转变为工具性、计算性考量。[28,32] 一方面,泰布罗塞尔(Tenbrunsel)和梅西克(Messick)[23] 发现,制裁或惩罚系统使得个体认为决策是与交易相关的,因而会精于计算而忽略了道德考量,这种变化会破坏合作。这种解释也得到了认知神经科学研究的支持,Li 和同事[22] 发现,惩罚威胁会降低大脑中涉及社会奖赏评价区域的激活水平,增强顶叶皮层的激活,顶叶皮层是与理性决策相关的。可见,惩罚的确会使得个体在决策时更多地考虑成本收益,进行理性判断,而信任与合作往往是非理性的行为。[12] 另一方面,根据认知失调理论,[33] 个体会力图使自己的行为与态度保持一致。当不存在惩罚威胁时,个体会认为自己的信任或合作行为是出于自己的道德、亲社会性或利他主义偏好;而当惩罚存在时,个体将自身的亲社会行为归因为外部动机,即为了避免惩罚。因而,当移除惩罚之后,个体的信任与合作水平会出现较大程度的下降。[11,27]

上述两种解释分别侧重了不同的方面。动机转变的解释更多地关注惩罚对个体信念、态度的影响,而惩罚释放不信任信号的解释则更多地关注惩罚对人际互动与群体规范的影响。人是社会性动物,总是力图作出符合群体规范的行为,这也正是文化对行为影响的来源。如果惩罚的确释放了群体不信任的信号,并且影响个体对群体规范的判断,那么,这种消极影响就必须得到足够的重视,因为群体规范作为一种文化对个体的影响是弥散性的,所有群体成员都会受到影响,随着交往的展开,这种不信任的规范会迅速扩展开,产生累积性的破坏作用。而且,群体规范一旦形成就非常难以改变,个体还会主动维护这种规范。[34] 当然,强调惩罚对群体规范的影响并不是说动机转变的解释就不重要或不合理。动机转变的理论基础是费斯廷格的认知失调理论,该理论由于其对人类行为与动机的解释效力而在心理学中占据了极其重要的地位,这足以说明该解释的重要性。而且,一旦个体将自身的亲社会行为建基于成本

收益分析,并解释为外部动机的激发,那么,要维持这种亲社会行为就需要外部激励的持续存在和加强。同时,当惩罚变为一种对信任与合作的外部激励后,惩罚就不能被移除了。研究者也的确发现了人们对惩罚机制的依赖性。[35]

三、惩罚影响信任与合作的调节因素

尽管惩罚经常被用来促进信任与合作,但是研究者却发现,惩罚对信任与合作的影响并不确定,积极的促进作用和消极的妨碍作用都存在。那么,为什么会出现这些差异呢? 针对这些差异,我们又该如何有针对性地、正确地使用惩罚机制呢? 我们认为,正确认识这些差异和使用惩罚机制的前提之一是识别影响惩罚效应的具体因素,即惩罚影响信任与合作的调节因素,这里列出了一些主要的调节因素。考虑到这些因素的存在,能够为合理利用惩罚机制提供支持。

1. 信任类型

信任指的是个体基于对交往对象意图和行为的积极预期而具有的愿意向对方暴露自身弱点,并且不担心被利用的心理状态和行为。[36~38] 根据信任的建立基础,研究者往往将信任分为人际信任和条件信任(制度信任) ,人际信任多基于对他人善心、诚意等的积极预期,而条件信任则强调对信任行为的成本收益分析,强调外部制度与规范对信任的保障。[39,40] 研究者发现,惩罚能够激发条件信任,[41] 但是会破坏人际信任。[42,43] 例如穆德等[35] 发现: 惩罚会削弱人际信任水平,减弱合作程度; 然而,惩罚会激发条件信任从而加强合作,条件信任会随着惩罚的取消而消失。

惩罚之所以对不同类型的信任有不同的影响,关键在于人际信任和条件信任有着不同的心理基础。人际信任的心理基础是个体对他人善心、诚意等的积极预期,此时个体愿意承担信任带来的风险,建立在此基础上的信任是非理性的或有限理性的。当惩罚存在时,会导致个体将决策与交易相联系,将被试的行为由伦理性、道德性转变为工具性、计算性,这对人际信任是不利的。[23,28,32] 相对于人际信任,条件信任的心理基础就是对信任进行成本收益分析。[44,45] 惩罚或者可以保障个体的信任不被利用,或者可以改变决策的收益结构使得信任与合作成为优势选项。因而惩罚会促进条件信任。信任类型的影响与惩罚对信任的不同作用及其解释是密切相关的。从前文可以看到,研究者对惩罚促进信任的解释更多地强调人是理性的,会对行为的收益成本进行计算; 而对惩罚消极影响的解释则强调惩罚对个体信任行为的内部动机、信任偏好的破坏作用。

2. 惩罚的合法性

影响惩罚效果的另一关键因素在于惩罚的合法性。肖(Xiao) 和豪泽(Houser) [46] 认为,惩罚的重要作用之一在于表达规则的存在和有效性,如果规则的有效性受到挑战,那么惩罚的作用就不存在了。据此,他们认为,只有当惩罚机制和效果是公开的,才可能传递规则的信号。也就是说,公开的惩罚具有合法性,能够促进信任与合作。为了检验惩罚的公开性对惩罚效果的影响,肖和豪泽设置了三种博弈情景: 标准的公共品博弈; 秘密惩罚博弈以及公开惩罚博弈。在标准的公共品博弈中,群体成员需要向公共品中捐赠一定量的金钱,公共品中的金钱会被乘以数倍,然后在群体成员间均分。在该博弈中,捐赠最少的个体会获得最大的收益。在两种惩罚情景中,群体中捐赠最少的个体会被处以一定比例的罚金,群体中的其他成员获得博弈任务所规定的收益。秘密惩罚和公开惩罚博弈的差别在于,在秘密惩罚博弈中,惩罚信息仅受惩罚者知晓,而在公开惩罚博弈中,所有个体都知道有人受到了惩罚。结果发现,相对于标准博弈,公开惩罚可以促进群体合作,而秘密惩罚会降低群体合作水平。可见公开惩罚具有合法性,能够传递群体规则,因而促进了合作; 而受到秘密惩罚的个体并不能明确识别自己受到惩罚是由于自己破坏了群体规则,不具合法性的惩罚只能引发反抗,而不能促进合作。惩罚释放了群体不信任或合作水平较低的信号正是由于惩罚没有明确传递一种规范,因而个体将惩罚知觉为对方的不信任或群体的信任水平较低。

影响惩罚合法性的另一因素是有关惩罚者的因素,那些具有良好行为的个体施加的惩罚是合法的,能够促进惩罚; [47] 当惩罚不是为了追求自身利益时,惩罚是有效的。[48] 研究者发现社会中存在反社会惩罚现象,即不良行为的个体会惩罚那些行为良好的个体,而这是有害于群体合作的。[49] 费勒(Faillo) 等[47] 认为,在群体内,只有那些行为良好的个体对行为不良的个体施加的惩罚

才是合法的,才能促进合作。为了研究惩罚合法性对惩罚效果的影响,费勒等比较了三种条件下人们对公共品的捐赠水平: 人人都可以惩罚他人、完全信息条件下的合法惩罚、不完全信息条件下的合法惩罚。实验分两个阶段进行,在第一阶段,群体成员向公共品进行捐赠,在第二阶段,群体成员获得群体捐赠信息并可以对他人施加惩罚。在人人都可以惩罚他人的条件下,群体内成员在第二阶段知道所有成员的捐赠信息,并可以任意惩罚其他人; 在完全信息条件下的合法惩罚组中,惩罚者同样知道所有成员的捐赠信息; 在不完全信息条件下,惩罚者只知道群体平均捐赠水平与比自己捐赠少的个体的信息。两种惩罚条件下的惩罚者可以惩罚比自己捐赠少的人。结果发现,在完全信息条件下,人们的捐赠水平最高,而另两组没有差异。这说明,惩罚合法性对于促进合作是必要的,但是信息公开同样是必须的,这与肖和豪泽[46]的研究结果相一致。这些结果说明惩罚要起到促进信任与合作的作用,惩罚必须具有合法性,即惩罚是公开的、公正的,惩罚者要具有良好的行为以及动机。

3. 个体信任或亲社会水平

惩罚对信任与合作的影响还受到个体最初信任或亲社会水平的调节,尽管人们使用惩罚是为了提高社会信任与合作水平,但一些研究却发现,惩罚对信任有着负面影响,尤其对那些本来信任水平较高或具有亲社会倾向的个体而言。在穆德等[11]的实验 2 中,他们首先使用问卷调查了被试的信任水平,然后使用移除惩罚范式考察了惩罚对信任与合作的影响,他们发现,惩罚只对那些本来信任水平就较高的个体有作用。在移除惩罚后的第二阶段,与本来信任水平较低的被试相比,本来信任水平就较高的被试的信任和合作水平有了更大的下降; 惩罚强度越高,本来信任水平较高的被试的信任与合作水平下降得越明显。这一结果在王沛与陈莉[27]的研究中得到了进一步的证实,他们发现,惩罚对亲社会型被试的影响更大,经历过惩罚的亲社会型被试在惩罚取消后的人际信任水平与合作程度更低。

惩罚对具有不同信任水平的个体产生的不同影响可能源于惩罚释放的信号与个体预期间是否具有一致性。皮洛特(Pillutla) 和陈(Chen) [50]的研究显示,当个体所了解到的他人的行为与自己的预期不一致时,个体的行为就会受到影响。也就是说,在交往中,如果存在惩罚机制,个体会将其知觉为群体不信任或对方不信任自己的信号,这种信号与高信任或亲社会型被试的期望是不一致的,这种不一致会导致被试的行为发生改变。然而,对信任水平本来就较低的个体而言,他们本来就预期他人会表现较低的信任与合作水平,惩罚所释放的信号与他们的信念是一致的,因而惩罚不会影响到他们的行为。此外,从动机转变的角度而言,亲社会型被试的信任与合作行为更多地出于内部动机以及道德上的考量,而信任水平较低的个体的信任与合作行为更多地出于外部动机或工具性考量,因而,惩罚会使得信任水平较高的亲社会型被试发生动机转变,破坏其信任与合作。

四、总结

人们一直力求建立一个高信任与合作水平的社会,然而,有证据显示,世界的信任与合作水平都出现了不同程度的下降。[51~53]为了提高社会信任与合作水平,人们致力于寻求促进信任与合作的方式,惩罚就是其中之一。尽管社会中惩罚机制被广泛采用,但是研究却显示惩罚对信任与合作的影响并不明确,尤其是惩罚可能破坏信任与合作,这种负面影响必须得到足够的重视。尤其是考虑到惩罚对信任水平较高的亲社会型个体有更大的影响,以及人们对惩罚的依赖性时,[35]对惩罚机制的使用需要更加谨慎。

当然,尽管惩罚对信任与合作有负面影响,其具体的影响机制与调节因素也还没有得到完全揭示,但是基于已有的研究成果,惩罚作为促进信任与合作的手段还是有用的,关键在于如何使用。第一,要有针对性地应用惩罚机制。在社会流动不断加速的现代社会,对条件信任的需求不断增加,这为应用惩罚提供了现实基础,但是,完全没有人际信任也无法建立高水平的条件信任,因而,应用惩罚要谨慎。在使用时,还应考虑到惩罚对象的特征、惩罚强度等的影响。第二,要充分发挥惩罚的积极作用,降低可能的负面影响,必须要建立惩罚的合法性,通过公开的、公正的、透明的惩罚来促进信任与合作,并且要建立惩罚者或惩罚机构的权威性,因为只有那些高信任个体或机构

作出的惩罚才具有合法性。[47]

参考文献：

[1] E. Fehr, U. Fischbacher. *Social Norms and Human Cooperation*. Trends in Cognitive Sciences, 2004, 8(4): 185 – 190.

[2] E. Fehr, S. Gächter. *Cooperation and Punishment in Public Goods Experiments*. The American Economic Review, 2000, 90(4): 980 – 994.

[3] C. McCusker, P. J. Carnevale. *Framing in Resource Dilemmas: Loss Aversion and the Moderating Effects of Sanctions*. Organizational Behavior and Human Decision Processes, 1995, 61(2): 190 – 201.

[4] A. Wit, H. Wilke. *The Presentation of Rewards and Punishments in a Simulated Social Dilemma*. Social Behaviour, 1990, 5(4): 231 – 245.

[5] E. Xiao, D. Houser. *Emotion Expression in Human Punishment Behavior*. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(20): 7398 – 7401.

[6] M. D. Caldwell. *Communication and Sex Effects in a Five – person Prisoner's Dilemma Game*. Journal of Personality and Social Psychology, 1976, 33(3): 273 – 280.

[7] D. Eek, P. Loukopoulos, S. Fujii, et al. *Spill – over Effects of Intermittent Costs for Defection in Social Dilemmas*. European Journal of Social Psychology, 2002, 32(6): 801 – 813.

[8] E. Fehr, S. Gächter. *Altruistic Punishment in Humans*. Nature, 2002, 415(6868): 137 – 140.

[9] R. M. A. Nelissen, M. Zeelenberg. *Moral Emotions as Determinants of Third – party Punishment: Anger, Guilt, and the Functions of Altruistic Sanctions*. Judgment and Decision Making, 2009, 4(7): 543 – 553.

[10] 刘国芳，辛自强. 间接互惠中的声誉机制：印象、名声、标签及其传递 [J]. 心理科学进展，2011，19(2)：233 – 242.

[11] L. B. Mulder, E. Van Dijk, D. De Cremer, et al. *Undermining Trust and Cooperation: The Paradox of Sanctioning Systems in Social Dilemmas*. Journal of Experimental Social Psychology, 2006, 42(2): 147 – 162.

[12] 王则柯，李杰. 博弈论教程 [M]. 北京：中国人民大学出版社，2010.

[13] R. Axelrod, W. D. Hamilton. *The Evolution of Cooperation*. Science, 1981, 211(4489): 1390 – 1396.

[14] M. A. Nowak. *Five Rules for the Evolution of Cooperation*. Science, 2006, 314(5805): 1560 – 1563.

[15] J. Smith, J. D. Van Dyken, P. C. Zee. *A Generalization of Hamilton's Rule for the Evolution of Microbial Cooperation*. Science, 2010, 328(5986): 1700 – 1703.

[16] R. Boyd, H. Gintis, S. Bowles. *Coordinated Punishment of Defectors Sustains Cooperation and can Proliferate When Rare*. Science, 2010, 328(5978): 617 – 620.

[17] R. Boyd, P. J. Richerson. *Punishment Allows the Evolution of Cooperation (or anything else) in Sizable Groups*. Ethology and Sociobiology, 1992, 13(3): 171 – 195.

[18] 约翰·梅纳德·史密斯. 演化与博弈论 [M]. 潘春阳译. 上海：复旦大学出版社，2008.

[19] H. Brandt, C. Hauert, K. Sigmund. *Punishment and Reputation in Spatial Public Goods Games*. Proceedings of the Royal Society of London. Series B: Biological Sciences, 2003, 270(1519): 1099 – 1104.

[20] D. Houser, E. Xiao, K. McCabe, et al. *When Punishment Fails: Research on Sanctions, Intentions and Non – cooperation*. Games and Economic Behavior, 2008, 62(2): 509 – 532.

[21] S. Gächter, B. Herrmann. *The Limits of Self – governance When Cooperators Get Punished: Experimental Evidence from Urban and Rural Russia*. European Economic Review, 2011, 55(2): 193 – 210.

[22] J. Li, E. Xiao, D. Houser, et al. *Neural Responses to Sanction Threats in Two – party Economic Exchange*. Proceedings of the National Academy of Sciences, 2009, 106(39): 16835 – 16840.

[23] A. E. Tenbrunsel, D. M. Messick. *Sanctioning Systems, Decision Frames, and Cooperation*. Administrative Science Quarterly, 1999, 44(4): 684 – 707.

[24] A. Dreber, D. G. Rand, D. Fudenberg, et al. *Winners don't Punish*. Nature, 2008, 452: 348 – 351.

[25] E. Fehr, B. Rockenbach. *Detrimental Effects of Sanctions on Human Altruism*. Nature, 2003, 422(6928): 137 – 140.

[26] J. Berg, J. Dickhaut, K. McCabe. *Trust, Reciprocity, and Social History*. Games and Economic Behavior, 1995, 10(1): 122 – 142.

[27] 王沛，陈莉. 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 [J]. 心理学报，2011，43(1)：52 – 64.

[28] E. Fehr, A. Falk. *Psychological Foundations of Incentives*. European Economic Review, 2002, 46(4): 687 – 724.

[29] B. S. Frey. *Motivation as a Limit to Pricing*. Journal of Economic Psychology, 1993, 14(4): 635 – 664.

[30] I. Seinen, A. Schram. *Social Status and Group Norms: Indirect Reciprocity in a Repeated Helping Experiment*. European Economic Review, 2006, 50(3): 581 – 602.

[31] A. Falk, M. Kosfeld. *The Hidden Costs of Control*. The American Economic Review, 2006, 96(5): 1611 – 1630.

[32] U. Gneezy, A. Rustichini. *A Fine is a Price*. Journal of Legal Studies, 2000, 29(1): 8 – 19.

[33] L. Festinger. A Theory of Cognitive Dissonance (Vol. 2). Stanford, CA: Stanford university press, 1962.

[34] W. M. Baum, P. J. Richerson, C. M. Efferson, et al. *Cultural Evolution in Laboratory Microsocieties Including Traditions of Rule Giving and Rule Following*. Evolution and Human Behavior, 2004, 25(5): 305 – 326.

[35] L. B. Mulder, E. van Dijk, D. De Cremer, et al. *When Sanctions Fail to Increase Cooperation in Social Dilemmas: Considering the Presence of an Alternative Option to Defect*. Personality

and Social Psychology Bulletin, 2006, 32(10): 1312 – 1324.

[36] M. Deutsch. *Trust and Suspicion*. The Journal of Conflict Resolution, 1958, 2(4): 265 – 279.

[37] J. B. Rotter. *A New Scale for the Measurement of Interpersonal Trust*. Journal of Personality, 1967, 35(4): 651 – 665.

[38] D. M. Rousseau, S. B. Sitkin, R. S. Burt, et al. *Not So Different After All: A Cross – discipline View of Trust*. Academy of Management Review, 1998, 23(3): 393 – 404.

[39] 卢曼. 信任:一个社会复杂性的简化机制[M]. 瞿铁鹏,李强译. 上海:上海人民出版社,2005.

[40] L. G. Zucker. Production of Trust: Institutional Sources of Economic Structure, 1840 – 1920. In B. M. Staw, L. L. Cummings (Eds). Research in organizational behavior (vol. 8). Greenwich, CT: JAI Press, 1986: 53 – 111.

[41] S. B. Sitkin, N. L. Roth. *Explaining the Limited Effectiveness of Legalistic "Remedies" for Trust/Distrust*. Organization Science, 1993, 4(3): 367 – 392.

[42] C. K. W. De Dreu, E. Giebels, E. Van de Vliert. *Social Motives and Trust in Integrative Negotiation: The Disruptive Effects of Punitive Capability*. Journal of Applied Psychology, 1998, 83(3): 408 – 422.

[43] D. Malhotra, J. K. Murnighan. *The Effects of Contracts on Interpersonal Trust*. Administrative Science Quarterly, 2002, 47(3): 534 – 559.

[44] 严进. 信任与合作[M]. 北京:航空工业出版社,2007.

[45] R. Hardin. Trust and Trustworthiness. New York: Russell Sage, 2002.

[46] E. Xiao, D. Houser. *Punish in Public*. Journal of Public Economics, 2011, 95(7): 1006 – 1017.

[47] M. Faillo, D. Grieco, L. Zarri. *Legitimate Punishment, Feedback, and the Enforcement of Cooperation*. Games and Economic Behavior, 2013, 77(1): 271 – 283.

[48] E. Xiao. *Profit – seeking Punishment Corrupts Norm Obedience*. Games and Economic Behavior, 2013, 77(1): 321 – 344.

[49] B. Herrmann, C. Thöni, G. Gächter. *Antisocial Punishment Across Societies*. Science, 2008, 319: 1362 – 1367.

[50] M. M. Pillutla, X. P. Chen. *Social Norms and Cooperation in Social Dilemmas: The Effect of Context and Feedback*. Organizational Behavior and Human Decision Processes, 1999, 78(2): 81 – 103.

[51] 马得勇. 信任,信任的起源和信任的变迁[J]. 开放时代,2008,(4): 72 – 86.

[52] 什托姆普卡. 信任:一种社会学理论[M]. 程胜利译. 北京:中华书局,2005.

[53] 辛自强,周正. 大学生人际信任变迁的横断历史研究[J]. 心理科学进展,2012,20(3): 344 – 353.

The Effects of Punishment Impacting on Social Trust and Cooperation: Controversy and Interpretation

LIU Guofang[1], XIN Ziqiang[2]

(1. Beijing Normal University, Institute of Developmental Psychology, Beijing 100875, China;
2. Central University of Finance and Economics, Department of Psychology at School of Social Development, Beijing 100081, China)

Abstract: In social interactions, punishing defectors usually is an important approach to improve trust and cooperation, and its validity has ever been proved by some researchers. Recently, however, more and more studies have found that punishment may undermine individuals' trust and cooperation, especially reduce trust level of who is pro – social or has a high initial trust level. This destructive effect is due to that 1) punishment releases signals of distrust and low cooperation; and 2) punishment changes individual's motives of trust and cooperation from ethical motives to more calculative motives. Thus, people should keep cautious when using punishment to improve trust and cooperation, and need to construct the legality of punishment first, as well as take the effect of trust types, individuals' initial trust level etc. into account.

Key words: punishment, trust, cooperation

(责任编辑:申　浩)