心理科学进展 2014, Vol. 22, No. 12, 1935–1943
Advances in Psychological Science DOI: 10.3724/SP.J.1042.2014.01935

自我惩罚：影响因素、模型与展望*

朱睿达 张 燊 申学易 刘 超

(北京师范大学认知神经科学与学习国家重点实验室, 北京 100875)

摘 要 自我惩罚是个体在违反社会规范后, 自愿使自己承受伤害或蒙受损失的行为。自我惩罚会受到负性情绪、补救机会、代偿机会和性别的影响。情绪模型和互惠模型分别从情绪和互惠的角度阐述了自我惩罚的认知机制。根据近期的研究结果, 可以推测自我惩罚与前扣带回、脑岛、右外侧眶额叶皮层、背内侧前额叶皮层、伏隔核、腹内侧前额叶皮层有关。未来值得研究的方向有: 进一步确认自我惩罚涉及的情绪成分、为互惠模型收集更多实验证据、探究自我惩罚对集体合作的影响以及开展跨文化研究。

关键词 自我惩罚; 情绪; 互惠; 合作

分类号 B849:C91

1 自我惩罚的概念

自我惩罚(self-punishment)最早是在弗洛伊德的精神分析理论中被提及的, 指的是寻求疼痛、承受痛苦, 以缓解无意识内疚的倾向(Freud, 1953)。弗洛伊德认为自我惩罚作为一种道德上的受虐现象, 是超我和自我之间斗争冲突的产物。自我惩罚的思辨解释具有一定的价值, 但随着实验研究的发展, 研究者们对自我惩罚有了更多的认识, 如：Wallington (1973)发现, 抑郁与自我惩罚有着紧密的联系, 越具有抑郁行为表现的被试越倾向于自我惩罚。后续的研究者们将抑郁这一概念细化, 发现其包含的内疚是影响自我惩罚的核心情绪, 个体的内疚程度越高就越倾向于自我惩罚, 并且观察到个体的自我惩罚可以缓解内疚(Bastian, Jetten, & Fasoli, 2011; Inbar, Pizarro, Gilovich, & Ariely, 2013)。同时, 这方面的研究并没有仅仅局限于研究情绪对自我惩罚的影响, 而是不断拓展研究的范围。Nelissen (2011)的研究表明, 受害者是否在场也会影响个体的自我惩罚行为。这表明自我惩罚可能是一种用于修复违规者与受害者的关系的方法。另外, 不同的实验范式和新的研究思路也更新着人们对自我惩罚概念的理解。如：自我惩罚的途径不再局限于忍受疼痛, 还可以是损失金钱(Nelissen & Zeelenberg, 2009)、耗费时间(De Hooge, Nelissen, Breugelmans, & Zeelenberg, 2011); 自我惩罚的作用也不再仅仅是减少内疚感, 还可能包括修复社会关系、促进集体合作(Gintis, 2000)。

总的来看, 尽管研究者们对自我惩罚有一些探索, 但这方面的研究仍处于初期阶段, 相关的研究并不丰富。为使将来的研究者们更加明确自我惩罚的概念, 有利于更多研究的开展, 结合近期的一些研究(Bastian et al., 2011; Inbar et al., 2013; Nelissen & Zeelenberg, 2009; Nelissen, 2011), 我们把自我惩罚概括为：个体在违反社会规范后, 自愿使自己承受伤害或蒙受损失的行为。

要准确理解自我惩罚的概念, 还必须弄清它与利他惩罚(altruistic punishment)的联系和区别。利他惩罚, 也称高代价惩罚(costly punishment), 指的是个体以牺牲自己利益为代价, 惩罚享受群体成果却不为群体合作出力的成员的行为(Egas & Riedl, 2008; Fehr & Gächter, 2002; Gächter, Renner, & Sefton, 2008)。自我惩罚与利他惩罚的相似之处在于个体都主动牺牲了自身的利益, 尝试去维护某种社会规范。但两者也存在一定的差异, 自我惩罚是个体对自己的违规行为施加惩罚,

收稿日期：2013-11-25

* 973 计划(2011CB711000, 2013CB837300), 国家自然科学基金(31170971, 61210010), 国家社科重大项目(12&ZD228)。

通讯作者：刘超, E-mail: liuchao@bnu.edu.cn

而利他惩罚是个体牺牲自己的利益对别人的违规行为施加惩罚。目前来看，有关利他惩罚的研究已较为深入系统，而关于自我惩罚的研究并不多且相对零散。对自我惩罚的相关研究结果进行归纳梳理，指明未来的研究方向，具有一定的理论与现实意义。

2 自我惩罚的影响因素

自我惩罚是一种复杂的社会心理行为，会受到各种心理、生理因素的影响。现有的研究表明，负性情绪、补偿机会、代偿机会和性别是影响自我惩罚的主要因素。

2.1 负性情绪

在违反社会规范的情况下，个体会产生负性情绪，这些负性情绪的强度会影响个体自我惩罚的程度。Wallington (1973)发现，相对于没有做出欺骗行为的被试，做出欺骗行为的被试在行为上有更多的抑郁表现，更倾向于在随后的任务中高强度电击自己，即进行更严厉的物理性自我惩罚。Watanabe 和 Ohtsubo (2012)首先让被试在分钱游戏中不得不做出不公平的分配，分给自己的钱多于分给别人的，然后让被试填写一份内疚量表。接着询问被试，由于前面的分配不公平，他现在是否愿意从自己分得的钱中扣除一部分，并强调只是单纯的扣除，而不是将钱送给其他游戏伙伴。结果显示，内疚程度越高的被试扣除自己钱的数额越大，即进行更强的经济性自我惩罚。

2.2 补救机会

一般而言，个体在伤害了他人后的反应是尝试补救，而不是伤害自己。Trivers (1971)认为个体违规后的内疚感会促进弥补行为，并由此修复受损的人际关系。不少研究都发现，个体违规后的内疚感与个体的亲社会表现(如：补偿他人、帮助他人)显著相关(Brown & Cehajic, 2008; Brown, González, & Zagefka, 2008; Ghorbani, Liao, Çayköylü, & Chand, 2013; Xu, Bègue, & Bushman, 2012)。那么当存在补救机会的时候，个体是否还会进行自我惩罚。Nelissen 和 Zeelenberg (2009)设计了一个 2×2 的被试间实验，要求被试想象：自己曾经向父母承诺好好学习，但最后由于自身没有努力(内疚条件)或老师教学不当(中性情绪条件)没能通过考试；被试还有一次补考机会(有补救机会条件)或被试不得不向父母要钱再交一次学费(无补救机会条件)。随后，被试收到一次去滑雪的邀请，询问被试自己在多大程度上会去参加这次快乐的滑雪旅行。研究者将被试拒绝体验快乐的机会视为自我惩罚。结果显示，无补救内疚组的自我惩罚程度最高，其他三组自我惩罚程度较低且两两间没有显著差异。

Nelissen 和 Zeelenberg 从情绪的角度对该结果进行了解释，认为无补救内疚组没有为自己错误开脱的借口，也没有其他方式来减轻自己所造成的危害，所以只能通过自我惩罚来减少内疚感。然而也可以从修复人际关系的角度对结果进行解释。如果是因为老师教学不当导致被试考试失败，父母不会怪罪于被试，父母与被试的关系不会受损。如果是因为被试不努力而考试失败，被试就违背了向父母做出的承诺，那么双方关系受损，需要修复。在该情况下补救是一种优于自我惩罚的互惠关系修复方式。如果被试抓住补救机会补考通过，不仅维护了父母的利益，修复了双方关系，还避免了自我惩罚给自己带来的损失。所以当存在补救机会时，被试出于利益最大化的原则，会寄希望于补救而不是选择自我惩罚。但当没有补救机会时，被试只能通过自我惩罚来表示自己的懊悔，以期获得原谅，来保证自己的长期利益。由此看来，自我惩罚服务于个体利益最大化原则，是用来修复人际互惠关系的方法之一。

2.3 代偿机会

是否存在代偿机会也对自我惩罚产生影响。De Hooge 等人(2011)在第一个实验中，先让内疚组被试想象，由于自己在分工作业中的消极行为使得 A 必须重修课程；让控制组被试想象尽管自己在分工作业中的行为消极，但 A 还是通过了课程考核。接着，让被试根据自己的意愿，将 50 欧元分给 A 和自己。结果发现相比于控制组，内疚组分给自己的钱较少，分给 A 的钱较多。该结果与经典的内疚组被试的自我惩罚程度高的实验结果一致。在第二个实验中，被试的任务与第一个实验中的任务基本相同，只是在分钱环节中增加了一个分配对象 B。结果发现，内疚组和控制组分给自己的钱没有显著差异，内疚组分给 A 的钱多于控制组，内疚组分给 B 的钱少于控制组。也就是说在该情况下，内疚组没有牺牲自己的利益，不会做出自我惩罚行为，而是将原本会分给 B 的

钱转而分给 A 了，牺牲 B 的利益去弥补自己对 A 犯下的过错，即让 B 代替自己对 A 进行补偿。在个人利益的视角下，个体让第三方进行代偿的方法不仅不需要牺牲自己的利益，还可以修复与原先的受害者的关系，符合利益最大化的原则。因此，在存在代偿机会的情况下，个体放弃自我惩罚而选择他人代偿具有一定的合理性。但潜在的问题在于，这种行为或许会损伤个体与代偿者的互惠关系。总的来说，自我惩罚、补救和代偿都是用于修复与他人互惠关系的方法，个体会在利益最大化原则的指导下根据需求对不同方法进行取舍。

2.4 性别

性别是另一个自我惩罚行为的影响因素。Inbar 等人(2013)和 Wallington (1973)均发现，当自我惩罚的形式是对自己施加电击时，女性被试给自己施加的电击强度要显著大于男性被试给自己施加的电击强度。对于该现象, Aronfreed (1964)的观点是，接受心理管教方式(如：讲道理、引发内疚、撤销关爱等)的孩子，比接受物理管教方式的孩子的更容易形成内化的良知，并在违规后进行自我惩罚。而父母多以心理管教对待女孩，以物理管教对待男孩，所以女性会给自己更多的自我惩罚。但与之矛盾的是，当研究中的自我惩罚的形式是非物理性刺激时，研究者们并没有报告自我惩罚存在性别差异(Nelissen & Zeelenberg, 2009; Watanabe & Ohtsubo, 2012)。该问题还有待进一步的研究。

3 自我惩罚的认知机制

在不同文化背景的社会中，都可以发现违规后的自我惩罚现象，如：欧洲的苦行僧为弥补自己罪行，游历各个乡镇，并在其间不断鞭打自己；美国的职业篮球教练因无法带领球队取得好成绩，而主动请辞；日本的武士在战败后，以切腹的方式以死谢罪。很显然，违规者的自我惩罚是损害个体的自身利益的，那么，人们为何要进行自我惩罚？为回答这一问题，研究者们分别从情绪和互惠的角度提出了两大理论模型。

3.1 情绪模型

情绪模型强调情绪在自我惩罚中起到的作用，认为情绪是自我惩罚行为的核心因素。情绪模型主要有三个论点：(1)负性情绪引发自我惩罚，自我惩罚调控负性情绪；(2)自我惩罚中涉及的主要负性情绪是内疚；(3)个体在通过自我惩罚调节负性情绪时，会产生一些积极副作用。

负性情绪引发自我惩罚，自我惩罚调控负性情绪。当个体做出的行为与内心行为规范不符，自我与超我诉求产生矛盾时，负性情绪就会产生。这种负性情绪会引起生理和心理上的不适，使个体会产生洗刷自身的罪恶(Zhong & Liljenquist, 2006)、摆脱负性情绪负担的需要。自我惩罚正是满足该需要的一种方式。个体本能地认为，接受惩罚可以抵消犯下的罪过(Glucklich, 2001)。因此，在没有外界惩罚的情况下，个体会主动对自己实施惩罚，寻求内心的平衡，降低负性情绪。Bastian 等人(2011)的研究支持了“负性情绪引发自我惩罚”的观点。他们首先通过让内疚组回忆自己排斥他人的经历引发被试的内疚情绪。然后告诉被试，他们将再参加一个身体敏感性的测试，要求被试将手尽可能长时间的放在冰水里。结果发现，内疚组的被试比中性情绪组的被试自愿将自己的手放入冰水的时间更长，即进行更强的物理性自我惩罚。另外, Bastian 等人(2011)还发现，自我惩罚有助于被试内疚感的减轻。在引发被试内疚情绪后，相比于让被试自愿将手放入温水中一段时间(非自我惩罚条件)，让被试自愿将手放入冰水中一段时间(自我惩罚条件)，能更有效的降低被试的内疚感。

自我惩罚中涉及的负性情绪主要是内疚。支持情绪模型的研究主要探索情绪与自我惩罚的关系，其对被试情绪的引发主要是通过让被试回忆或做出违规行为的范式来实现的。值得注意的是，这类范式除了可以引发研究者关心的目标情绪外，还会引发许多其他的负性情绪。因此，虽然大量研究都报告了自我惩罚与内疚情绪相关的现象，但还是有必要将内疚与其他负性情绪对自我惩罚的影响进行区分。Inbar 等人(2013)发现，在自愿接受惩罚性刺激时，内疚组被试比悲伤组、中性情绪组被试自愿给予自己更多的电击惩罚。悲伤组和中性情绪组给予自己的电击强度无显著差异。由此可以说明悲伤并不能显著影响自我惩罚。此外，羞耻与内疚同属道德情绪，都出现在违规的行为之后，具有高度的相似性，将两者进行区分也十分必要。内疚与羞耻的关键差异之一在于，内疚主要与个体违反规范从而伤害他人的事件相

关，羞耻则主要和个体在众人面前表现出自己无能的事件相联系(Smith, Webster, Parrott, & Eyre, 2002)。以此为依据, Nelissen 和 Zeelenberg (2009)为避免引发被试的羞耻感，在引发内疚的过程中特意让被试展现了自己的能力。实验结果依然显示，相比于控制组被试，内疚组被试的自我惩罚更严厉。另外, Nelissen (2011)在关于自我惩罚的实验中使用量表直接测量了被试的内疚和羞耻情绪，该结果也表明与自我惩罚相关的情绪是内疚而非羞耻。

个体在通过自我惩罚调节内疚感时，会产生一些积极副作用。根据自我肯定理论(self-affirmation theory) (Steele, 1988)，可以将自我惩罚解释为一种情绪管理策略，成功的情绪调控有助于个体实现自我肯定。在违反社会规范后，负性情绪的产生会导致个体内心的失衡，自我否定机制启动，自我认同感和自我形象受损。而当自我惩罚成功减少违规者的负性情绪后，个体的自我否定机制停止。这将有助于自我肯定的重新建立。也就是说，在自我惩罚在调节自身情绪的同时，会产生有利于个体重新自我肯定的积极副作用。自我惩罚的积极副作用不仅局限于个人，还会作用于集体。不少研究都发现，个体会主动牺牲自己的利益去制止不道德、不公平的行为(Carpenter, Bowles, Gintis, & Hwang, 2009; Fehr, Fischbacher, & Gächter, 2002; Gintis, 2000)。而这种行为会促进个体所在群组的合作行为，使该群组在进化中的组间选择中胜出并存留下来。自我惩罚就是这种牺牲自身利益，起到维护社会规范作用的行为。但一般认为，这种行为的产生并不是因为个体的“高尚”，而是因为个体面对不公平状况时自然产生的负性情绪(如：Fehr et al., 2002)。根据情绪模型，个体的自我惩罚只是单纯由内疚驱动的，而最后该行为给群体带来好处是个体开始没有认识到的。

情绪模型主要关注情绪与自我惩罚的关系，其优点在于能够简单而清晰的阐述自我惩罚的产生原因和效用。但它的问题在于，情绪模型将自我惩罚产生的非情绪性积极作用均解释为自我惩罚情绪调控的副产品，过度简化了人类的高级认知过程，一定程度上低估了个体对自身行为的认识和运用能力。并且与之相反的是，已有研究发现个体能够有意识的、策略性的运用自我惩罚来为自己的利益服务(De Hooge et al., 2011; Nelissen & Zeelenberg, 2009; Nelissen, 2011)。综上所述，仅依据情绪说是无法对自我惩罚行为进行全面合理解释的，需要其他的理论模型对其进行补充。

3.2 互惠模型

与将自我惩罚视为一种情绪驱动行为的情绪模型不同，互惠模型认为个体的自我惩罚是互惠利益关系驱动的。人类作为群体性动物，需要通过与他人的合作和积极互动来获得更多的利益。一些研究发现个体会通过牺牲短期利益做出慷慨助人、维护社会规范等积极行为来提升自身声誉，以期获得或维持优质的合作互惠关系(Bereczkei, Birkas, & Kerekes, 2010; Gintis, Smith, & Bowles, 2001)。在互惠模型中，自我惩罚被认为是违规者为了改变自己在别人眼中原有的消极违规形象，修复受损的互惠关系，以获取与别人再次合作的可能，从而保护自身利益的行为。互惠模型认为，自我惩罚的本质是个体放弃短期利益，以获得更多的长期利益，实现利益的最大化的策略性行为。根据互惠关系的不同，可以将自我惩罚产生的原因和效用分为：直接互惠关系修复和间接互惠关系修复。

直接互惠简单来说就是，“我帮你，然后你帮我”。直接互惠理论(direct reciprocity theory)认为在多次交往中，直接互动的双方互助互惠，获得较大的长期利益(Miyaji, Tanimoto, Wang, Hagishima, & Ikegaya, 2013; Rand, Ohtsuki, & Nowak, 2009; Trivers, 1971)。违规行为出现后，违规者和受害人的关系急剧恶化，直接互惠关系终止，违规者的利益受到损害。此时违规者会做出一些补救措施，如：用自我惩罚来表达歉意。Nelissen (2011)先让被试误以为自己做出了违规行为，然后给被试电击自己的机会，在被试电击自己时有三种情况：无人在身边、受害者在身边、无关人员在身边。结果发现，当受害者在被试身边时，被试给予自己的电击强度显著高于无人在身边和无关人员在身边的情况。如果自我惩罚只是因为违规所引起的内疚导致的，在三种情况下高强度的自我惩罚都应该出现，然而实验结果显示，高强度的自我惩罚有且仅在有受害者在被试身边时出现。Nelissen (2011)认为自我惩罚是一种违规者针对受害者发出的懊悔信号，违规者以此寻求受害者的原谅。而违规者针对受害者发出懊悔信号和寻求受害者原谅的实质是在尝试修复与受害者的关

系, 以期保存再次与受害者建立互惠关系的可能, 实现自己未来利益的最大化。研究表明这种有代价的懊悔展示确实是有效的。Petersen, Sell, Tooby 和 Cosmides (2012)发现当罪犯表现出懊悔时, 普通人会认为其作为未来伙伴的价值增加且更有可能在未来与之合作互惠, 相应的, 对罪犯采取的制裁方式会较为温和。Ohtsubo 和 Watanabe (2009)发现相比于无代价的懊悔展示, 被试会将有代价的懊悔展示理解为更真诚的道歉。

间接互惠理论(indirect reciprocity theory)认为, 个体在帮助别人的过程中会获得好的社会声誉(reputation), 而人们更愿意与有积极声誉的个体合作互惠(Milinski, Semmann, & Krambeck, 2002; Nowak & Sigmund, 1998; Stanca, Bruni, & Mantovani, 2011)。个体的违规行为不仅会影响其与受害者的关系, 还会损害自身的社会声誉。社会声誉受损后, 违规者将不得不承担被社会群体排斥以及遭到其他社会成员惩罚的风险(Cashdan, 1980)。同时, 社会群体还可能因此不再将违规者纳入互惠活动中去。出于总体利益的考虑, 当自我惩罚的损失小于被他人惩罚的损失或自我惩罚的损失小于未来互惠的获利时, 违规者会以自我惩罚的方式向作为非受害者的其他社会成员传递一种“我愿意服从社会规范”的信号, 从而避免他人的惩罚和获得群体的原谅(Bastian et al., 2011)。虽然还没直接的实验证据支持个体的自我惩罚是为了提升名誉, 维持间接互惠关系, 但人们以提升自身声誉为目的而牺牲自己利益的行为是存在的。如：在私下秘密的情况下, 即难以获得声誉提升的情况下, 仅有少数人愿意为陌生人提供帮助; 而在其他社会成员面前, 即可以获得声誉提升的情况下, 更多的人愿意向陌生人伸出援手(Bereczkei et al., 2010)。而在关于自我惩罚的研究中(如:Bastian et al., 2011; Nelissen & Zeelenberg, 2009), 被试很有可能将主试视为非受害者的第三方社会成员, 尝试以自我惩罚的方式向主试传达懊悔信息。并且这种被试对主试评价的顾虑, 是很难消除的(Watanabe & Ohtsubo, 2012)。关于自我惩罚的间接互惠关系修复功能的问题, 还有待实验的验证。

互惠模型以利益为着眼点对自我惩罚行为的起因和效应进行了阐述, 对某些无法用情绪模型来理解的自我惩罚现象进行了解释, 为理解自我惩罚提供了另一个视角。但其问题在于缺乏实验证据, 尤其是尚未有直接的实验证明个体会使用自我惩罚维持间接互惠关系。总的来说, 情绪模型和互惠模型分别从情绪和利益的角度提出了相应的理论观点, 两者相互补充, 有助于研究者们对自我惩罚形成更加深刻的认识。

4　自我惩罚的神经机制

目前对于自我惩罚的研究集中于行为层面, 而关于其神经机制的研究还基本处于空白状态。此处根据已有的行为研究结果和理论, 对自我惩罚的相关脑区进行合理推测和简介, 以期引起研究者们对自我惩罚脑机制的研究兴趣。

4.1　情绪相关脑区

根据情绪模型, 自我惩罚行为与负性情绪, 尤其是内疚, 紧密相关。与内疚体验相关的脑区有：前扣带回(anterior cingulate cortex, ACC)、脑岛(insula)、右外侧眶额叶皮层(rightlateral orbitofrontal cortex, RLOFC)、背内侧前额叶皮层(dorsomedial prefrontal cortex, DMPFC)。

ACC　已有不少研究发现, 预期或想象会引发内疚情绪的情景会引起 ACC 的活动增强(Basile et al., 2011; Chang, Smith, Dufwenberg, & Sanfey, 2011; Shin et al., 2000)。Yu, Hu, Hu 和 Zhou (2014)则研究了被试在更为真实的人际互动中的内疚情绪神经机制。他们让被试与另一名玩家(实际上为假被试)玩一个点数估计的游戏, 只要两人中的任意一人估计点数错误, 双方就需要共同接受总量一定的电击。在看到双方估计点数的正误情况后, 被试可以选择自己将承担总电击量中的多少, 而剩余的电击量将由对方接受。相比于双方都估计错误, 仅被试单独估计错误时, 被试的内疚感和主动承担的电击量都更高。将被试单独错误条件与双方错误条件下的脑激活情况进行对比, 发现在被试单独错误条件中, aMCC (anterior middle cingulate cortex)的激活强度更高。另外, 他们还发现 aMCC 会通过中脑核团(midbrain nucleus)的完全中介作用对被试自愿接受多少电击产生影响。这些研究表明 ACC 是自我惩罚行为中重要的情绪脑区之一。

脑岛　脑岛不仅负责身体感觉的加工, 如温度、疼痛等(Craig, 2002), 在负性情绪的加工中也起重要作用。当个体面临惩罚的威胁(Spitzer,

Fischbacher, Herrnberger, Grön, & Fehr, 2007)、经历不公平事件(Tabibnia, Satpute, & Lieberman, 2008)或具有负性情绪预期(Herwig, Kaffenberger, Baumgartner, & Jäncke, 2007; Herwig, Baumgartner, et al., 2007)时, 研究者通常可以观察到脑岛的激活。具体到内疚而言, Chang 等人(2011)发现, 个体在反事实任务中报告的内疚程度越高, 内疚敏感性越高, 其脑岛在涉及内疚的任务中的激活水平越高。相似的, Baumgartner, Fischbacher, Feierabend, Lutz 和 Fehr (2009)发现, 当个体计划违背承诺(将体验内疚情绪)时, 脑岛活动增强。

RLOFC 和 DMPFC Wagner, N'Diaye, Ethofer 和 Vuilleumier (2011)利用文字让被试回忆生活中的特定典型事件, 在不同的典型事件中被试会分别体验内疚、羞耻、悲伤三种不同情绪。将被试在回忆内疚事件时的脑区激活状态与被试回忆羞耻、悲伤事件时的相比, 结果发现仅当被试处于内疚状态时, 被试的 RLOFC 和 DMPFC 独特性激活。因此, 可以将 RLOFC 和 DMPFC 的激活作为区分内疚、羞耻、悲伤的一个指标, 这对验证自我惩罚行为的核心情绪为内疚, 有一定的应用价值。

4.2 奖赏计算和预期相关脑区

根据互惠模型, 个体违规后的自我惩罚是为了自己将来利益的最大化。因此个体可能会对自我惩罚的成本、修复关系所得利益进行计算和预期。伏隔核(nucleus accumbens, NAcc)和腹内侧前额叶皮层(ventromedial prefrontal cortex, VMPFC)在奖励预算和预期中起重要作用。

NAcc 和 VMPFC 一直被认为是经典的奖赏脑区, 其活动会受到个体所面对的奖惩得失的影响(Knutson, Adams, Fong, & Hommer, 2001; Knutson, Fong, Bennett, Adams, & Hommer, 2003; Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004)。近期一些研究进一步发现, NAcc 和 VMPFC 与价值计算(Rangel, Camerer, & Montague, 2008)以及一级和二级奖励的加工和预期(Dreher & Tremblay, 2009)紧密相关。在未来的自我惩罚脑机制研究中, NAcc 和 VMPFC 是值得重点关注的区域。

5 自我惩罚的研究展望

5.1 进一步确认自我惩罚涉及的情绪成分

虽然大量的行为实验结果支持了内疚引起自我惩罚这一观点, 但我们还是应该谨慎的对待已有的研究结果。Nelissen (2011)认为使用问卷测量被试的情绪存在一定问题, 被试可能主观上难以准确区分相似的情绪(如：内疚和羞耻)。该问题很难在行为研究中克服, 但或许可以通过神经机制研究来解决。例如, (Wagner et al., 2011)通过让被试回忆典型事件来引发不同情绪, 利用磁共振成像技术记录下被试的脑神经活动情况, 然后将被试在不同情绪状态时的脑激活情况进行比较。结果发现, 相比于羞耻和悲伤, 被试在内疚状态时 RLOFC 和 DMPFC 独特激活。因此, 可以通过记录个体自我惩罚过程中的脑活动, 看是否能观察到 RLOFC 和 DMPFC 的激活, 来寻找"内疚是自我惩罚涉及的关键情绪"的支持性证据。相比于使用被试的主观评分, 使用生理指标来区分被试的情绪状态更加的敏感、明确和直接。未来的研究可以通过更多的情绪神经机制研究来进一步确认自我惩罚涉及的具体情绪成分。

5.2 为互惠模型收集更多实验证据

目前关于自我惩罚的主流研究都是围绕情绪模型进行的, 情绪模型的主要观点是内疚情绪引发了个体的自我惩罚。与情绪模型不同, 互惠模型尝试从利益关系的角度解释自我惩罚, 提出了自我惩罚是一种利益策略行为的观点。违规者在做出违规行为之后, 违规者与受害者以及看到违规者违规行为的第三方的关系恶化, 违规者与他人的互惠关系很可能终止。出于总体利益的考虑, 违规者可能以自我惩罚的方式来表示自己悔过的意愿, 并借此来修复与他人的互惠关系, 从而实现利益的最大化。互惠模型是十分必要的, 因为它可以合理解释一些情绪模型无法解释的现象, 如：当受害者在身边时, 个体的自我惩罚程度更高(Nelissen, 2011)。如果互惠模型正确, 个体的自我惩罚是为了实现利益最大化, 那么一段关系的利益价值越大, 个体就会越愿意在此关系受损时表现出自我惩罚。根据该推断, 验证互惠模型一条可行的思路是, 观察受损关系的利益价值对自我惩罚的影响。总的来说, 互惠模型对自我惩罚有一定的解释力, 但支持它的实验证据却略显单薄, 将来需要更多的实验来对其进行验证。

5.3 研究自我惩罚对集体合作的影响

在关于合作的研究领域中, 利他惩罚对群体合作的影响一直是一个研究热点。不少研究发现, 如果群体中存在利他惩罚机制能显著增加群体的

合作水平(Egas & Riedl, 2008; Fehr & Gächter, 2002; Gächter et al., 2008)。但通过利他惩罚的方式促进合作有其缺点。一方面, 利他惩罚的成本很高, 惩罚者和被惩罚者双方的收益都会被减少, 因此最终利他惩罚的代价可能高于其所促进合作的获利(Dreber, Rand, Fudenberg, & Nowak, 2008; Herrmann, Thöni, & Gächter, 2008)。另一方面, 被惩罚者可能对惩罚者心存怨恨, 而对惩罚者实施报复, 进行反社会惩罚(antisocial punishment)。最终导致双方恶意地相互惩罚(Herrmann et al., 2008)。再者, 研究发现一旦利他惩罚机制被撤除, 个体之间不能相互惩罚后, 群体的合作水平立刻降低(Fehr & Gächter, 2002)。

自我惩罚与利他惩罚行为的相似性可能使得自我惩罚也具有促进群体合作的功能。而它与利他惩罚的不同点甚至可能会使其成为一种更优的合作促进手段。首先, 两种惩罚模式自身的机制决定了, 利他惩罚会给惩罚者和被惩罚者双方造成利益损失, 而自我惩罚只会对自身造成损失。这使得自我惩罚行为的成本低于利他惩罚。其次, 自我惩罚不会对他人的利益造成损害, 惩罚者不必担忧遭到报复。避免相互的恶意惩罚对群体的整体利益都是有益的。再次, 利他惩罚的作用模式是“惩恶”。个体对不合作者进行惩罚, 使其不敢在做出违规行为。其潜在的问题在于, 一旦利他惩罚机制消失, 不合作行为会迅速反弹。与之不同的是, 自我惩罚的作用模式是“扬善”。个体在自己违反社会规范后, 以自我惩罚的方式向其群体发出一种 “我愿意为我的错误负责, 我会在下一回合选择合作”的信号, 倡导承担责任和公平合作。这更容易让其他成员形成对遵守社会规范的内在认同, 使得整个团体的合作氛围更持久。最后, 现实生活中不是在所有条件下都存在惩罚他人的条件, 有时个体无法做出利他惩罚。相比之下自我惩罚具有更高的可实施性。通过上面的分析, 将利他惩罚与自我惩罚对群体合作的影响进行比较是未来一个值得研究的问题。

5.4　开展跨文化的研究

个体进行自我惩罚可能是出于减少内疚的考虑, 也可能是为了修复人际互惠关系。对此, 不同文化背景中的个体是否存在差异？以欧洲、美国为代表的西方国家是一种个人主义文化, 强调独立性自我, 而以中国、日本为代表的东方国家是一种集体主义文化, 强调依赖性自我(Markus & Kitayama, 1991)。我们认为, 在西方的独立性自我构建的文化背景下, 个体会从自我的角度出发, 注重自身情感体验, 以情感指导其行为。据此, 西方人可能更注重自我惩罚减少内疚的功能。在东方的依赖性自我构建的文化背景下, 个体强调集体主义, 注重群体关系。东方人可能更注重自我惩罚修复人际关系、恢复社会声誉的功能。这些都需要通过跨文化的研究去加以证实。

参考文献

Aronfreed, J. (1964). The origin of self-criticism. *Psychological Review, 71*(3), 193–218.

Basile, B., Mancini, F., Macaluso, E., Caltagirone, C., Frackowiak, R. S. J., & Bozzali, M. (2011). Deontological and altruistic guilt: Evidence for distinct neurobiological substrates. *Human Brain Mapping, 32*(2), 229–239.

Bastian, B., Jetten, J., & Fasoli, F. (2011). Cleansing the soul by hurting the flesh: The guilt-reducing effect of pain. *Psychological Science, 22*(3), 334–335.

Baumgartner, T., Fischbacher, U., Feierabend, A., Lutz, K., & Fehr, E. (2009). The neural circuitry of a broken promise. *Neuron, 64*(5), 756–770.

Bereczkei, T., Birkas, B., & Kerekes, Z. (2010). Altruism towards strangers in need: Costly signaling in an industrial society. *Evolution and Human Behavior, 31*(2), 95–103.

Brown, R., & Cehajic, S. (2008). Dealing with the past and facing the future: Mediators of the effects of collective guilt and shame in Bosnia and Herzegovina. *European Journal of Social Psychology, 38*(4), 669–684.

Brown, R., González, R., & Zagefka, H. (2008). Nuestra culpa: Collective guilt and shame as predictors of reparation for historical wrongdoing. *Journal of Personality and Social Psychology, 94*(1), 75–90.

Carpenter, J., Bowles, S., Gintis, H., & Hwang, S. (2009). Strong reciprocity and team production: Theory and evidence. *Journal of Economic Behavior & Organization, 71*(2), 221–232.

Cashdan, E. A. (1980). Egalitarianism among Hunters and Gatherers. *American Anthropologist, 82*(1), 116–120.

Chang, L. J., Smith, A., Dufwenberg, M., & Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron, 70*(3), 560–572.

Craig, A. D. (2002). How do you feel? Interoception: The sense of the physiological condition of the body. *Nature Reviews Neuroscience, 3*(8), 655–666.

De Hooge, I. E., Nelissen, R. M. A., Breugelmans, S. M., & Zeelenberg, M. (2011). What is moral about guilt? Acting

"prosocially" at the disadvantage of others. *Journal of Personality and Social Psychology, 100*(3), 462–473.

Dreber, A., Rand, D., Fudenberg, D., & Nowak, M. (2008). Winners don't punish. *Nature, 452*(7185), 348–351.

Dreher, J. C., & Tremblay, L. (Eds.). (2009). *Handbook of reward and decision making*. Burlington, MA: Academic Press.

Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences, 275*(1637), 871–878.

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

Freud, S. (1953). The economic problem of masochism. In J. Strachey (Ed.), *The standard edition of the complete psychological works of Sigmund Freud* (Vol. 19, pp. 155–170). London, UK: Hogarth Press.

Gächter, S., Renner, E., & Sefton, M. (2008). The Long-Run Benefits of Punishement. *Science, 322*(5907), 1510.

Ghorbani, M., Liao, Y., Çayköylü, S., & Chand, M. (2013). Guilt, shame, and reparative behavior: The effect of psychological proximity. *Journal of Business Ethics, 114*(2), 311–323.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology, 206*(2), 169–179.

Gintis, H., Smith, E., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of Theoretical Biology, 213*(1), 103–119.

Glucklich, A. (2001). *Sacred Pain: Hurting the Body for the Sake of the Soul*. New York: Oxford University Press.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362–1367.

Herwig, U., Baumgartner, T., Kaffenberger, T., Brühl, A., Kottlow, M., Schreiter-Gasser, U., … Rufer, M. (2007). Modulation of anticipatory emotion and perception processing by cognitive control. *Neuroimage, 37*(2), 652–662.

Herwig, U., Kaffenberger, T., Baumgartner, T., & Jäncke, L. (2007). Neural correlates of a "pessimistic" attitude when anticipating events of unknown emotional valence. *Neuroimage, 34*(2), 848–858.

Inbar, Y., Pizarro, D. A., Gilovich, T., & Ariely, D. (2013). Moral masochism: On the connection between guilt and self-punishment. *Emotion, 13*(1), 14–18.

Knutson, B., Adams, C., Fong, G., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *The Journal of Neuroscience, 21*(16), 1–5.

Knutson, B., Fong, G. W., Bennett, S. M., Adams, C. M., & Hommer, D. (2003). A region of mesial prefrontal cortex tracks monetarily rewarding outcomes: Characterization with rapid event-related fMRI. *NeuroImage, 18*(2), 263–272.

Markus, H., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98*(2), 224–253.

Milinski, M., Semmann, D., & Krambeck, H. J. (2002). Reputation helps solve the "tragedy of the commons". *Nature, 415*(6870), 424–426.

Miyaji, K., Tanimoto, J., Wang, Z., Hagishima, A., & Ikegaya, N. (2013). Direct reciprocity in spatial populations enhances R-Reciprocity as well as ST-Reciprocity. *PloS One, 8*(8), e71961.

Nelissen, R. M. A. (2011). Guilt-Induced Self-Punishment as a sign of remorse. *Social Psychological and Personality Science, 3*(2), 139–144.

Nelissen, R. M. A., & Zeelenberg, M. (2009). When guilt evokes self-punishment: Evidence for the existence of a Dobby Effect. *Emotion, 9*(1), 118–122.

Nowak, M. A., & Sigmund, K. (1998). The dynamics of indirect reciprocity. *Journal of Theoretical Biology, 194*(4), 561–574.

Ohtsubo, Y., & Watanabe, E. (2009). Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evolution and Human Behavior, 30*(2), 114–123.

Petersen, M., Sell, A., Tooby, J., & Cosmides, L. (2012). To punish or repair? Evolutionary psychology and lay intuitions about modern criminal justice. *Evolution and Human Behavior, 33*(6), 682–695.

Rand, D. G., Ohtsuki, H., & Nowak, M. A. (2009). Direct reciprocity with costly punishment: generous tit-for-tat prevails. *Journal of Theoretical Biology, 256*(1), 45–57.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience, 9*(7), 545–556.

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *Neuroreport, 15*(16), 2539–2243.

Shin, L. M., Dougherty, D. D., Orr, S. P., Pitman, R. K., Lasko, M., Macklin, M. L., … Rauch, S. L. (2000). Activation of anterior paralimbic structures during guilt-related script-driven imagery. *Biological Psychiatry, 48*(1), 43–50.

Smith, R. H., Webster, J. M., Parrott, W. G., & Eyre, H. L. (2002). The role of public exposure in moral and nonmoral shame and guilt. *Journal of Personality and Social Psychology, 83*(1), 138–159.

Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G., &

Fehr, E. (2007). The neural signature of social norm compliance. *Neuron, 56*(1), 185–196.

Stanca, L., Bruni, L., & Mantovani, M. (2011). The effect of motivations on social indirect reciprocity: an experimental analysis. *Applied Economics Letters*, *18*(17), 1709–1711.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 261–302). New York: Academic Press.

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.

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46*(1), 35–57.

Wagner, U., N'Diaye, K., Ethofer, T., & Vuilleumier, P. (2011). Guilt-specific processing in the prefrontal cortex. *Cerebral Cortex, 21*(11), 2461–2470.

Wallington, S. (1973). Consequences of transgression: Self-punishment and depression. *Journal of Personality and Social Psychology, 28*(1), 1–7.

Watanabe, E., & Ohtsubo, Y. (2012). Costly apology and self-punishment after an unintentional transgression. *Journal of Evolutionary Psychology, 10*(3), 87–105.

Xu, H., Bègue, L., & Bushman, B. J. (2012). Too fatigued to care: Ego depletion, guilt, and prosocial behavior. *Journal of Experimental Social Psychology, 48*(5), 1183–1186.

Yu, H., Hu, J., Hu, L., & Zhou, X. (2014). The voice of conscience: Neural bases of interpersonal guilt and compensation. *Social Cognitive and Affective Neuroscience, 9*, 1150–1158.

Zhong, C., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science, 313*(5792), 1451–1452.

Self-punishment: Contributing Factors, Theoretical Models and Research Prospects

ZHU Ruida; ZHANG Shen; SHEN Xueyi; LIU Chao

(*State Key Laboratory of Cognitive Science and Learning, Beijing Normal University, Beijing* 100875, *China*)

Abstract: Self-punishment refers to behaviors that an individual inflicts pain or imposes sanctions on himself/herself after the transgression. Self-punishment could be influenced by several different factors such as negative emotion, opportunities of remedy, opportunities of compensation at the expense of others and gender. Moreover, two core models, namely emotion model and reciprocity model, were summarized. With the development of neuroscience, it could be inferred that ACC, insula, RLOFC, DMPFC, NAcc, VMPFC are closely linked to self-punishment. Finally, authors' opinions about future research were provided.

Key words: self-punishment; emotion; reciprocity; cooperation