社会合作的行为经济学解释评述*

史丹　汪崇金

内容提要:人类与生俱来就具有与他人合作并维护伦理规范的倾向。尊重并激发人类的这一特质,已是当前中国推进社会治理的一个有效突破口。本文紧扣经济学实验、演化仿真和脑成像行为经济学三大研究方法,从静态视角系统梳理了人类这一特质的证据,并从动态视角勾画其可能的演进路径。本文有助于人们正确理解个体在社会互动中的行为动机与方式,强化人们对他人合作态度的乐观判断,以及对他人维护伦理规范倾向的积极预期,从而在借助他律的同时,践行慎独,自觉地维护社会合作和良好规范,更好地促进社会合作。

关键词:强互惠行为　实验经济学　脑科学　演化仿真

一、引言

现实生活中随处可见人与人之间的合作,但与此不相称的是,主流经济学长期以来以"经济人假设"为起点,以"竞争"为主线,专注于研究稀缺资源的有效配置,忽视了对人类合作行为的研究(黄少安、韦倩,2011)。实际上,人类所以能够取得今天的成就,并不是由于人类与其他动物一样具有竞争的本性,而是与之相反,在于人类与其他动物不同的特点——高度的合作能力(孟昭勤、王一多,2004)。认识到这一点,对于方方面面的制度设计与安排意义重大。大的方面关乎如何推进"一带一路""环境治理"等国际合作,小的方面关系如何加强"社区治理""组织管理"等人际互动。近年来,学术界甚至出现了一种呼声,即经济学有从传统的资源配置理论走向合作理论的必要(黄少安,2000;张维迎,2015)。

研究人类合作行为,"如何认识'人'"是绕不过的槛。因为要理解经济如何运行,懂得如何管理经济并促进经济繁荣,就必须关注人们的某些思维模式(阿克洛夫、席勒,2009)。不过,正如卢梭在《论人类不平等的起源和基础》的序言中所说,"人类的各种知识中最不完备的,就是关于'人'的知识"。其中,关于人性的讨论由来已久。古今中外,概不例外。在中国传统文化中,管仲有"夫凡人之情,见利莫能勿就,见害莫能勿避"的感叹,而孟子则有"人无有不善"的乐观。在西方文化中,对人性探索可追溯到马基雅弗利和孟德维尔,但影响最为深远的当属经济学之父亚当·斯密。他在《国富论》中的一段论述被尊为"经济人假设"的始源。不过,斯密在强调人的"自爱"的同时,还强调了"克己"和"谨慎",自爱的经济人本身包含了以"同情"为内容的伦理范畴(朱富强,2009)。毫无疑问,如果只有"自利"或"自爱",人类怎能破解"囚徒困境"、走出"霍布斯丛林"?令人欣喜的是,近几十年来,行为经济学借助实验、仿真和脑成像等技术,迅速积累了大量的证据,系统地证实了人们并不是具有同质的自利偏好,而是深刻地受到生活环境、社会规范和文化传统的复杂影响,具有异质的社会偏好(World Bank,2015)。

通俗地讲,社会偏好是指一些感觉。它包括,人们愿意与志趣相投的人合作,可以从中获得快乐,或者感到对这种行为抱有义务;人们也喜欢惩罚那些盗用他人合作成果的人,或者感到有义务这么做(鲍尔斯、金迪斯,2015)。人们这种喜欢合作、讨厌不合作者的社会偏好,在行为上则体现为条件性合作(conditional cooperation)[①]和利他性惩罚(altruistic punishment)。在桑塔费学派(Santa Fe Institute)

* 史丹,中国社会科学院工业经济研究所,邮政编码:100836,电子邮箱:shidan01@163.com;汪崇金,中国社会科学院财经战略研究院,邮政编码:100028,山东财经大学财政税务学院,邮政编码:250014,电子邮箱:wangchjin@126.com。本文为国家社会科学基金重大项目(13&ZD168)、国家社会科学基金一般项目(14BGL148)、中国博士后科学基金资助项目(2015M581266)的阶段性成果。感谢匿名审稿人的宝贵建议,当然文责自负。

的语境中，这些行为又被定义为积极的强互惠（positive strong reciprocity）和消极的强互惠（negative strong reciprocity）。[②] 基于异质性社会偏好的强互惠理论为我们描述了这样一幅图景：在一个群体中，强互惠者会积极尝试着与他人合作，但仅此并不足以维系合作，因为难免存在一些搭便车者，如果不对他们加以约束，搭便车行为会进一步蔓延。好消息是，倘若允许个体间相互监督与惩罚，即使没有预期利益作为补偿，强互惠者也会不惜花费个人成本惩罚那些搭便车者，合作则得以维系。在社会学、人类学等领域的学者看来，强互惠者的合作倾向与对违规、卸责、搭便车等机会主义行为的利他惩罚，是维系伦理、道德、习俗、禁忌、礼仪、规矩等非正式制度的根本力量，也是强化法律、法规、合同等正式制度的重要支撑，现已成为理论界破解“社会合作何以可能”这一难题的重要突破口。

强互惠理论强调了人类行为动机的多样性和社会性，对“人”的抽象更符合实际。不过，强互惠理论毕竟是在新近才发展起来的，尚有诸多质疑，对社会实践的指导潜力也尚待挖掘。鉴于如此，本文着力从静态视角，系统梳理强互惠特质的经济学实验证据与脑科学研究发现，并尝试从动态视角勾画这一特质可能的演进路径。借此评述，以期进一步宣传强互惠理论、彰显强互惠力量，强化人们对他人的合作与利他惩罚的预期，引导人们在借助他律的同时，践行慎独，自觉地参与到规范维系、社区治理、环境保护，以及食品安全等方面的公共利益维护和公共事务管理中来，从而更好地促进不同领域的社会合作。

二、强互惠行为的实验证据

经济学实验凭借其较好的可控制性和可复制性，能有效地测度变量之间的因果关系，为强互惠理论提供了一系列极具说服力的行为方面的证据。其中，公共品实验被认为最为适合模拟人们在现实状态下的互动（Chaudhuri，2011），为此本文着重讨论公共品实验的证据。具体而言，这些实验证据可分为以下两大类。[③]

（一）条件性合作的实验室证据

实验经济学家对强互惠行为的兴趣首先源于对“最后通牒博弈”的分析。在最后通牒博弈中有两个参与者，分别称为提议者和回应者，他们进行一定数量的现金分配。提议者首先提出一个分配方案，回应者然后决定接受或拒绝该方案，如果回应者接受，则双方按照分配方案分得现金；如果回应者拒绝，双方收益均为零。按照自利偏好假设，在一次性匿名博弈中，提议人给对方任意一个非常小的正的单位收益，响应者将接受提议并达成均衡。但是，大量的最后通牒博弈实验显示，大多数提议者会分给回应者 40%～50% 的现金（Güth et al，1982；Bolton & Zwick，1995；List & Cherry，2000）。这是自利偏好假设无法解释的现象，Güth et al（1982）称之为“最后通牒博弈悖论”。

类似地，按照主流经济学的博弈分析，在一次性公共品博弈中，参与者不会向公共账户中贡献自己的禀赋。但是，在实验经济学不算长的历史中，已开展的 200 多个公共品实验均显示：被试的公共品投资显著不为零（Isaac & Walker，1988；Andreoni，1988）。[④] 除此之外，在独裁者博弈实验、信任博弈实验、礼物交换博弈实验中，被试也都呈现出传统的自利偏好假设无法解释的合作倾向。

这些现实观察与理论预测的“不一致”吸引了人们对超越自利偏好假设的研究，顺势而生的社会偏好（social preference）理论引起学界的重视（陈叶烽，2010）。尽管社会偏好概念的雏形可以追溯到 Veblen（1934）、Duesenberry（1949）等，但要严格地给社会偏好下一个定义并非易事。文献中一般有四种具体形式的社会性偏好：纯粹利他（pure altruism）、光热效应（warm glow）、互惠（reciprocity）和不平等厌恶（inequality averse）。Ashley et al（2010）、周业安和宋紫峰（2008）、陈叶烽（2009）等还曾运用计量分析方法尝试给予进一步界定。从本文掌握的资料来看，社会偏好的进一步具体化并未引起学界的更多关注，而是广泛用于表示人类的亲社会情感。相似的表达还有亲社会偏好、涉他偏好等。

条件性合作是指人们在预计他人合作时也会还以合作的行为特征，是社会偏好这一心理动机的行为表现。Fischbacher et al（2001）在开篇即提出，“一些人可能是出于某种形式的社会性偏好而表现为条件性合作”，从而免于对社会偏好具体形式的纠缠。[⑤] 在此之后，条件性合作成为一个更为中性的概念，用于描述人们愿与志同道合者合作的特质。而且，Fischbacher et al（2001）的两阶段公共品实验设计现已发展成为定量分析异质性社会偏好的基本范式。他们基于 Selten（1967）的策略性方法，通过激励相容约束，要求被试回答其在他人的公共品贡献量分别为 0、1、2……20 等情况下的公共品贡献量。然后依据这两个序列之间的相关性，将被试划分为

条件性合作者、搭便车者、倒U型合作者等类型。在此之后，Herrmann & Thöni(2009)、Rustagi et al(2010)、Fischbacher & Gächter(2010)、Volk et al(2011)、汪崇金等(2012)、周业安等(2013)、周晔馨等(2014)先后基于这一范式，以不同经济发展水平和文化背景的个体为实验对象，得到了较为一致的实验结论，为条件性合作提供了更有力的实验证据。[6]

以条件性合作为主要内容的异质性社会偏好假设能够有效地解释重复多期的标准公共品实验中的“非零贡献”与“合作退化”(周业安、宋紫峰，2008)。实验中，一部分人是强互惠者，表现出条件性合作倾向，他们在实验伊始就尝试着与他人合作，因此我们观测到实验中的公共品贡献量不为零。不过，还有一部分人是搭便车者，他们不顾强互惠者的努力，而一直选择搭便车。强互惠者的回应便是减少或拒绝合作，再次表现出条件性合作特征，因此我们观测到实验中的公共品贡献量随着实验的重复进行而下降。这进一步佐证了条件性合作假说的合理性。

(二)利他惩罚的实验室证据

强互惠行为的另一方面为利他惩罚，利他惩罚的实验证据也应从“最后通牒博弈”谈起。前述分析已经提到，在大量的实验中，多数提议者表现得相当慷慨，这是自利偏好假设无法解释的。这里需要补充的是，这些实验还显示，对于提议者的吝啬(比如说低于禀赋的30%)，响应者常常会拒绝接受，导致双方都一无所获。响应者的拒绝实质上是对提议者的惩罚，当然响应者自身也为这样的惩罚支付成本，因为他本可以获得一个正的收益，只不过在他看来有点少而已。这里有点“宁为玉碎，不为瓦全”的情绪宣泄，体现了利他惩罚的特征。

利他惩罚首份公共品实验证据来自Fehr & Gächter(2000)。他们是以现实中的一个悲剧事件开始的。在1979年“油荒”期间，卡特政府出台了一系列汽油配给与价格控制的措施，导致购油司机加油时需要排长队等候。排队的人群中常常因插队而产生殴斗、叫骂，一位乘汽车旅行的人甚至因为插队而被一位素不相识的卡车司机枪杀。这显然是一个极端案例，但现实中类似的“路见不平、拔刀相助”的行为时有发生，例如最近发生的“瓜子哥”“项链姐”等事件。这些都反映了这样一种现象：人们厌恶破坏合作规范、搭便车等不合作行为，有时甚至不惜花费个人成本施以惩罚。为了验证这种利他惩罚，Fehr & Gächter(2000，2002)在标准公共品实验中新增了一个环节，允许被试之间相互监督，对搭便车的队友实施有成本的惩罚(下文称这一实验设计为F&G设计)。[7]他们的实验结论显示，对于搭便车行为的利他惩罚普遍存在；搭便车程度越严重，遭受队友的利他惩罚就越大。这也进一步解释了标准公共实验中“合作退化”现象。强互惠者遭遇搭便车后之所以减少或拒绝合作，是因为这是他们惩罚搭便车者的唯一手段(Fehr & Gächter，2000)。更为重要的是，Fehr & Gächter(2000)还发现，利他惩罚能够维系较高水平的合作。

在随后的十多年里，Bochet et al(2006)、Carpenter(2007)、Sefton et al(2007)、宋紫峰和周业安(2011)等基于F&G设计，以不同文化背景的个体为实验对象，证实了广泛存在的利他惩罚，一遍遍复述着与Fehr & Gächter(2000)相同的乐观故事。

三、对强互惠理论的质疑

大量的公共品实验显示，一些人具有强互惠特质，已成为破解“人类合作何以可能”这一难题的突破口。当然，对于这一乐观判断，也有很多学者并不信服，提出了诸多质疑。[8]在后续的拓展性研究中，有些质疑得以化解，而有些却不断强化并更具颠覆性。

(一)“非零贡献”是“合作”还是“迷糊”？

尽管异质性社会偏好假说能够解释公共品实验中的“非零贡献”与“合作退化”两大经典现象，并得到广泛的认可，但持怀疑态度的人也不在少数。例如，Andreoni(1995)、Houser & Kurzban(2002)等研究指出，一些被试没有选择搭便车，是“混乱”(confusion)或“失误”(error)所致。这些迷糊的(confused)被试贡献了总量的50%左右，这一比例远高于强互惠理论支持者测算的6%～10%(Fischbacher et al，2001；Fischbacher & Gachter，2010)。不过，随着实验经济学的发展，实验程序更为规范、实验技术更为成熟，“迷糊”一说曾一度消失，但近年来再次风起。持怀疑态度的代表性人物有M. Burton-Chellew、S. West等。他们的实验研究再次提出，“非零贡献”不是因为合作，而是出于利益最大化的动机、不断试错的结果。

比如，Burton-Chellew(2016)得到了一些有别于以往的实验结论：(1)当被试在人机博弈时，同样表现出异质性的社会偏好，社会偏好类型分布与Fischbacher et al(2001)、Fischbacher & Gächter(2010)中以人作为博弈玩家的实验结论基本相同。

(2)无论是与计算机博弈还是与人博弈,被试在策略性实验中表现出的社会偏好都可以解释他们在一次性博弈中的公共品供给,这与 Fischbacher & Gächter(2010)以人为被试对象的实验结论也是一致的。[9](3)个体利益最大化策略应该是贡献零单位公共品,而与他人贡献量的多少无关,但在实验中,只有搭便车者是这样认为的,而条件性合作者一般会相信自己利益最大化策略与他人的贡献量有关。作者强调,尽管设置了标准的控制性问题,但还不能确保被试能够正确理解博弈,实验方法的可靠性仍值得怀疑,行为经济学实验中的基本假设——"选择"显示"动机",并不必然成立。

再比如,Burton-Chellew et al(2015)假设了三种学习规则:基于利益的学习(payoff-based learning)、亲社会的学习(pro-social learning)、条件性合作(conditional cooperation)。基于利益的学习规则是指被试仅关心自己的收益;亲社会的学习规则是指被试不仅关心自己的收益还关心他人的收益;而条件性合作规则假设被试不仅关心自己的收益还关心他人的贡献量。他们开展了三种设计的实验:黑盒子(black box)设计、标准设计、强化设计。三者的区别仅在于信息的多寡不一。黑盒子设计实验中的信息量最少,仅告诉被试将按照某一个数学公式计算个人所得,再无其他提示信息;在标准设计实验中,告诉被试其收益以及其他三位队友的公共品贡献量,这与 F&G 设计一致;而强化设计实验中提供的信息比在标准设计中多了两条,即其他队友的公共品收益和总收益。[10]理论上讲,强化设计实验提供了更多的信息,不确定性会有所下降,人们因未知而学习模仿他人的可能性也会有所下降。但实际上,在标准设计实验中,被试更明显地表现为条件性合作,而在强化设计实验中,更多的信息没有改善合作,反而具有反社会效果(anti-social consequence),[11]公共品供给水平下降趋势更为明显。这说明被试表现出的条件性合作不是出于社会性偏好,而是社会学习的结果;更为重要的是,仅有基于利益的学习规则能够解释全部的三种设计实验的数据。他们据此判断,条件性合作主要是因为困惑或失误所致,而不是亲社会偏好的体现,公共品实验不能证实人类所拥有的利他性。

尽管 Burton-Chellew 等人开展了卓有成效的研究,但他们引述的文献中,除了自己团队的研究成果,如 Kummerli et al(2010)、Burton-Chellew & West(2013)之外,剩下的只有上文提及的 Andreoni (1995)、Houser & Kurzban(2002)。可以说,Burton-Chellew 等人的质疑尚未在更大范围内引起共鸣。当然,"学习"是人类活动的基本特征,实验中被试可能存在学习活动,这也难怪"学习"假说(learning hypothesis)由来已久却难以排除。演化心理学、演化博弈论、生物学和有限理性论一致认为,人类能够快速地习得和有效地运用互惠规范和社会规则,正是这个强大的学习能力,使得个体能够在大量的社会困境中通过其积极行动获得收益(奥斯特罗姆,2010)。也许正如 Muller et al(2008)所言,被试在实验中的自愿供给行为的变化反映了他们尝试探索对他们最为有效的策略,但这种变化并非一直朝着个体利益最大化的方向变动。换言之,被试并非简单地学习如何最大化个人利益。这说明 Burton-Chellew 等人的质疑尚不能否定社会偏好假说,反而为深入研究社会偏好提出了新的视角。

(二)利他惩罚实验果真客观描述了现实生活?

强互惠理论的支持者们声称,利他惩罚实验解释了狩猎聚居部落、游牧民族等小型社会的自发合作(Bowles & Gintis,2002;Richerson & Boyed,2005)。但在 Guala(2012)看来,这样的声称过于随意,并提出利他惩罚实验缺乏现实证据的质疑。Guala(2012)的质疑引发了学术界就利他惩罚的大讨论,诸多著名学者,如 E. Ostrom、N. Nikiforakis 等都参与进来,桑塔菲学派"四君子"S. Bowles、R. Boyd、H. Gintis、E. Fehr 也加入了论战。2012 年 2 月发表的《行为与脑科学》(Behavior and Brain Science)还以专题的形式集中收录了这些讨论。总体来看,争议主要集中在以下几个方面。

首先,利他惩罚实验设计究竟在多大程度上刻画了真实世界?比如,Güney & Newell(2012)指出,实验中无须真正地付出努力,实验所得类似于意外之财,这在现实中相当少见,因此实验不能用于模拟现实生活、解释现实问题。更具颠覆性的是,上述乐观结论都是基于 F&G 设计,而这种设计事先排除了包括报复在内的反社会惩罚(Anti-social punishment)。无论是在经济学实验中(Denant-Boemont et al,2007;Nikiforakis,2008),还是在演化博弈模型中(Hauert et al,2007;Janssen & Bushman,2008;Rand et al,2010),一旦加入反社会惩罚,上述乐观故事都将被改写。由此可见,无视反社会惩罚显然有损利他惩罚实验的效度。

其次,现实生活中存在利他惩罚吗?这也是回应实验效度质疑的关键问题。Guala(2012)重新审

视了强互惠理论支持者所声称的人种学证据，并指出现实中一些惩罚是无须惩罚者支付成本的，而另一些所谓的高成本惩罚(costly punishment)往往是由集体完成的，惩罚成本为所有成员公摊。他强调，这些惩罚都与实验中的利他惩罚不一致，不可认定为利他惩罚的现实证据。Guala(2012)对利他惩罚缺乏现实证据的质疑也得到 Binmore(2005)、Ross(2006)、Johnson(2012)等的广泛支持和认同。

毫无疑问，Guala 等人的质疑极具挑战，但还没有严重到让我们否认社会偏好是重要行为动机的判断。虽然允许反社会惩罚的公共品实验和演化博弈分析再次得出悲观结论，似乎又应验了霍布斯、洛克等先哲的预言，但在新近开展的允许交流(Ostrom，2012)、增加信息供给(Kamei & Putterman，2013)的公共品实验中，还是得到了支持利他惩罚能够维系社会合作的结论。就利他惩罚缺乏现实证据这一质疑而言，一些学者指出，由于现实生活多处于均衡状态(Johnson，2012；Gächter，2012)，再加上存在不确定性(Bereby-Meyer，2012；Gehrig et al，2007)，因此很难观测到利他惩罚，但这不能否定利他惩罚在一次性交往中的作用。Fudenberg & Pathak(2010)的发现就是一个佐证，他们以美国大学生为被试对象的实验显示，仅仅是利他惩罚威胁就足以维系较高水平的社会合作。换言之，日常生活中往往无须真正地发生利他惩罚。另外，Balafoutas & Nikiforakis(2012)新近在希腊雅典的一个地铁站组织了有关利他惩罚的自然现场实验(natural field experiment)，实验者故意违反车站的公共秩序，他们发现许多人会出面指责制止，这为利他惩罚实验提供了新的有利证据。

四、强互惠理论的脑科学证据

从上述研究来看，无论是实验经济学研究还是演化经济学研究，均尚存分歧，未能为强互惠理论提供令人信服的证据，近年来，脑科学有了长足发展，从另一个角度为强互惠理论提供了新的证据。现有研究表明，人类的大脑由一系列专门的模块组成，这些模块是按照早期人类所处环境的特殊需求而逐渐被塑造出来的(福山，2015)。脑科学家基于这样的认识，运用脑功能成像(functional neuroimaging)、功能性磁共振成像(fMRI)等工具，迅速积累了大量的脑科学数据，就人类的信任、互惠交换等社会行为背后的神经系统展开了深入的研究，其中不乏对强互惠行为的探索。

(一)条件性合作的脑科学解释

由著名的行为神经科学教授 J. Rilling 领衔的研究团队对合作行为背后的神经系统做了大量研究(Rilling et al，2002，2007，2008)。其中，Rilling et al(2002)发现，被试与他人合作而非背叛时，包括伏隔核(nucleus accumbens)、尾核(caudate nucleus)等在内的纹状体(striata)被激活。[12]纹状体大约形成于7000 万年前，是与决策行为有关的重要脑区，尤其是与奖赏系统有关，包括金钱回报和愉悦情绪(Schultz & Romo，1988；Kawagoe et al，1998；Doherty et al，2004)。纹状体被激活说明被试从合作行为中获得了额外收益。当然，J. Rilling 等人的系列研究基本上都是基于固定匹配的重复囚徒困境实验(fixed matching repeated PD)。这种实验设计可能存在这样一个问题，由于博弈对象是固定不变的，被试在看到自己当期实验收益时可能也在谋划下期是合作还是背叛。因此将难以区分所观测到的脑区变化究竟是对实验收益的反应还是对行为决策过程的映射(Suzuki et al，2011)。

为此，Suzuki et al(2011)开展了随机匹配的重复囚徒困境实验(random matching repeated PD)。实验中，被试是随机匹配的，实验者会告知被试与其随机相遇的队友究竟是“合作的”、“非合作的”，还是“未能确定类型的”。实验者的这一判断是根据被试往期的贡献情况总结而成的。然后，实验者分别扫描被试在决策时和观测到实验收益时的功能性磁共振成像。他们发现，相对于非合作的队友而言，被试更愿意与合作的队友或未能确定类型的队友合作，表现出条件性合作特征。而且，当遇到不合作的队友时，被试右部的前额叶侧背部(dorsolateral prefrontal cortex，简称 DLPFC)、[13]双侧的后颞上沟(posterior superior temporal sulcus，简称 pSTS)和颞顶交界区(temporo-parietal junction，简称 TPJ)更为活跃。他们进一步指出，合作是被试的优势反应(pre-potent response)，但遇到非合作队友时，会抑制优势反应而选择背叛，DLPFC、pSTS/TPJ 等脑区被激活反映的正是这一认知抑制过程。已有研究显示，其中的 DLPFC 关乎对犯罪行为是否实施惩罚的研判(Knoch et al，2006；Buckholtz et al，2008)。不难看出，Suzuki et al(2011)的脑科学证据与 Fehr & Gächter(2002)的调查结论是一致的，即拒绝或减少合作是强互惠者对不合作者的一种惩罚。

(二)对利他惩罚的脑科学解释

在得不到物质补偿的情况下，人们为什么肯不

惜花费个人成本去惩罚那些违反合作规范的人呢？这是强互惠理论的核心问题。E. Fehr、T. Singer 等人的两份有关囚徒困境博弈与信任博弈的脑功能神经成像研究对此做了解释。De Quervain et al (2004)的研究显示，如果被试在遭遇不公平对待时还以利他惩罚，那么他们大脑中纹状体背侧(dorsal striatum)的尾核会被激活；而且，尾核的活跃程度与其用于惩罚他人的成本呈正相关性。前文已指出，纹状体是哺乳动物权衡损益的主要脑结构。换言之，人们可以从利他惩罚这种行为本身获得满足(叶航等，2005)。De Quervain et al(2004)这一文献在国内学界流传已久，叶航等(2005)、韦倩(2010)、韦倩和姜树广(2013)、汪崇金(2013)等均有所译介，在此不再赘述。

T. Singer 曾是 E. Fehr 的学生，因发现“同情心”的脑神经网络而声名鹊起(汪丁丁，2011)。[14] 她和同事于在《自然》杂志上发表的文章(Singer et al，2006)，再次佐证了 E. Fehr 等人的上述结论。她们的研究显示，当看到行事公正的队友遭遇痛苦时，被试大脑中与痛苦相关的脑区额岛皮层(fronto-insular cortex)和扣带前沟(anterior cingulate cortices)会被激活，这种反应就是亚当·斯密所说的“同情”。当看到行事不公正的队友遭遇痛苦时，至少是在男性被试中，这种同情反应(empathy-related response)会明显下降，与此相应地，他们与奖赏系统有关的脑区，如纹状体腹侧(ventral striatum)、眶额叶皮层(orbito-frontal cortex)更为活跃，活跃程度与被试自我报告的对该队友的憎恨程度密切相关。她们推测，人们对他人的同情是以其对他人社会行为的评价为基础的；特别是对于男性被试，当看到行事不公的人遭遇痛苦时，他们不会给予相应的同情，这种免于同情是对他人不公正行为的惩罚。严重的情况就是人们通常所说的“幸灾乐祸”。男性被试不同脑区活跃程度的“一降一升”说明，他们对他人的不幸本来会产生同情，但又因为他人行事不公对其实施惩罚而未同情，由惩罚产生的满足感正好弥补了未给予同情所造成的损失。她们的发现与 E. Fehr 等人的结论遥相呼应。

除此之外，针对第三方的利他惩罚也得到了脑科学研究的支持。一般而言，相对于第二方利他惩罚而言，第三方利他惩罚刺激的脑区可能更为平静(dispassionate)，但 Buckholtz et al(2008)功能性磁共振成像研究显示事实并非如此。实验中，被试阅读一份描述某一场景的书面材料后，决定是否对其中的主人公实施惩罚及其程度。与以往研究一致，被试脑内与决定是否实施惩罚以及惩罚力度的脑区，前额叶侧背部和杏仁核(amygdala)均有相应的反应。这些发现说明，第三方惩罚同样是受针对失范者的负面情感使然(Rilling et al，2011)。

上述研究显示，人类大脑对相互合作和惩罚背叛者的加工过程与其他享乐行为的过程几乎相同，人们在合作和惩罚搭便车者的过程中获得了满足感(鲍尔斯、金迪斯，2015)。脑科学研究的发现有助于人们消弭分歧，从而更为深入地理解强互惠行为。当然，人脑的各个部分既有分工又有合作，人们对于脑内的合作秩序仍然知之甚少。[15] 我们注意到，尽管上述研究多以控制回报系统的纹状体为考察对象，但他们关注的具体部位又有所不同。人脑内部结构相当精细复杂，这种微小不同或许暗示着神经系统的巨大差异。因此，现有的研究结论不仅难以在同一个层面上比对，其可信度也大打折扣，甚至给人一种盲人摸象的怀疑。这注定着脑科学方面的研究仍然是任重而道远。

五、讨论与启示

经济学实验与脑科学研究尽管仍存分歧，但给我们呈现了这么一个事实：人具有与他人合作并维护伦理规范的倾向。诚然，仅仅是这些还不够，更为重要的命题是要解释清楚，人类的这些行为倾向是如何形成的？

从上述综述来看，强互惠行为的脑科学证据并不充分，这一微观层面上的研究尚缺，不过，在社会偏好这个宏观层面上的研究颇丰。相关的研究具体分为两大类。第一类是脑科学的研究。其中，有这样一个共识：人脑有三层，分别来自不同演化阶段，具有不同的功能。当中的第二层是“外缘系统”(limbic system)，也称“情感脑”，是情感活动的策源地，被称为欲望、愿望、冲动等的心理活动都生发于此(汪丁丁，2011；福山，2015)。与此相应的是，上文提及的脑区都集中于此，由此可见，控制人类社会偏好情感的脑区是在长期演化中逐渐形成的。第二类研究是演化仿真研究，其中的“基因—文化共演化”(gene-culture coevolution)模型已广为接受。该模型假设，一个新的生物体为了更好地适应所处环境，可籍两种通道获得信息，一种是基因的信息通道，即通过父辈的基因编码获得在所处环境中持久不变或者在时间和空间中变动很慢的信息；另一种是非基因的信息通道，具体而言包括个体学习和社会学习，

即凭借自身的学习能力从所处环境中习得。对于大多数动物来说，基因传递和个体学习就是事情的全部，而对于人类而言，社会学习或称文化传播，是获取信息的重要渠道。“基因—文化共演化”模型认为，人类的社会偏好是基因影响文化演化、文化影响基因演化的动力过程的结果。这一假设得到了模拟仿真的佐证（鲍尔斯、金迪斯，2015）。

总而言之，强互惠理论以大量的实验经济学、脑科学等方面的证据，并通过演化经济学的仿真分析，逻辑自洽地提醒我们：人们在长期生活中逐渐形成了社会偏好，自愿遵守并希望他人遵循合作规范，自己做不到时会内疚，别人做不到时则会气愤，甚至会不惜花费个人成本给予惩罚（福山，2015）。通过对相关文献的梳理，我们可以从中获得下列一些启示。

1. 重视“人”的异质性是强互惠理论的重大突破。对于行为经济学来说，理解人脑三结构的功能及冲突尤为重要，因为这是解释人类行为的关键环节。人脑的三层中，除了最早演化而成的、也是在最内层的脑干和前述的“情感脑”之外，还有“理性脑”。“理性脑”是最新演化而成的、也是在最外层的大脑皮质，负责高级认识，掌管着意识、语言等功能，理性选择（对可选方案进行排序和比较，并从中选优）也发生于此（汪丁丁，2011；福山，2015）。与电脑负责精确计算不同，人的“理性脑”的理性选择过程充斥着来自“情感脑”的情感因素（福山，2015）。换言之，个体的理性决策往往会包含部分情绪（非理性）和部分非自利的成分（周业安，2015），因此既不是完全理性的，也不是完全自利的。[16] 这是人类行为的复杂性之所在，是共性方面的。除此之外，还有个性方面的，因为人们的行为方式受到其长期以来得到的教育、感受到的文化氛围、信守的道德准则等因素的影响，必然也会表现出异质性和复杂性。我们注意到，尽管社会偏好是否稳定可靠尚存争议（汪崇金、聂左玲，2015），但强互惠理论正视人类行为的复杂性，并积极沿着这个方向来理解、刻画人的复杂行为，对“人”的抽象因此更真实，是对传统自利偏好假设的重大突破。

2. 强调“人”的强互惠特质对于促进社会合作尤为重要。首先，过分强调“个人贪婪”的假设是不符合事实的，而且使得悲观的预期在个体间蔓延，这不利于实现包括公共品自主供给在内的社会合作。一个有力的例证是，相比较其他专业的学生而言，经济学专业的学生在公共品实验中表现得更为自私，一种可能的解释是他们接受的教育改变了他们的行为（Frank et al，1998）。其次，在当前的中国社会，需要通过公共教育，强化社会个体的利他惩罚预期。Wu et al（2009）、汪崇金和史丹（2016）以中国在校大学生为被试对象，分别开展了设有利他惩罚的囚徒困境实验和公共品实验。这些实验一致地证实，利他惩罚乏力、利他惩罚威胁不足，难以有效抑制违规、卸责、搭便车等机会主义行为。一个重要原因是，大多数被试，特别是搭便车者，不相信或低估他人的利他惩罚。因此，需要引导人们正确认识“人”的强互惠特质，尊重他人的善意，敬畏他人的惩罚，从而增强人们在社会互动中与他人的合作。

3. 激发“人”的强互惠特质已是当前社会治理的一个主题。由于私人契约和政府命令无论单独起作用还是联合起来，都无法为现代社会的治理提供夯实的基础，社会合作仍然是经济和社会生活的必然要求（鲍尔斯、金迪斯，2015）。我们乐见，在当前社会治理创新的背景下，个体的强互惠特质已得到重视和重用。一方面，中国积极培育和弘扬社会主义核心价值观，推进道德重建和再生，通过内化、认同和融合等心理过程，寻求道德支持的自我行为约束途径（王道勇，2014）；另一方面，“不带剑的契约不过是一纸空文，它毫无力量去保障一个人的安全”（霍布斯，1985）。中国在强化以公共权力为后盾的公共惩罚的同时，在各个领域畅通投诉举报渠道、发挥媒体舆论监督、鼓励同行监督，在私人惩罚与公共惩罚的良性互动中，充分发挥人们对违规、卸责、搭便车等机会主义行为实施利他惩罚的亲社会特质（汪崇金、聂左玲，2015）。可以说在当前社会治理实践中，在强调“放权让利”、从正向激励入手“把激励搞对”的同时，还在不断强化包括利他惩罚在内的各种形式的惩罚，[17] 着力构建多层次的惩戒体系，从负向激励入手“把激励搞对”。这一逻辑有别于以往家庭联产承包责任制改革、国有企业改革等，也是当前社会治理的一个重要突破口和显著特征。

注：

①有些文献也称之为“利他性合作”（altruistic cooperation）。

②这里的“利他性”并非强调行为主体的利他动机，而是强调行为本身会降低行为主体自己的适存度、增加群体的适存度。生物学家用后代与母代之比来测度适存度（fitness），如果一个人的生存策略有利于增加自己的适存度，那么他的后代的数目必须超过母代的数目（汪丁丁，2011）。

③尽管积极的强互惠与消极的强互惠都是受社会偏好驱使，但为了便于聚焦问题，文献梳理时分两类齐头推进。

④数据来源于 Charles Holt，“The Y2K Bibliography of Experimental Economics and Social Science”，http://www.

people. virginia. edu/～cah2k/y2k. htm。

⑤还有一种观点认为，条件性合作是个更为广泛的概念，不仅包括积极的强互惠，即 Fischbacher et al(2001)语境下的条件性合作，还包括出于直接互惠、间接互惠动机所表现出的合作行为(Suzuki et al，2011)。

⑥更为详细的总结详见连洪泉和周业安(2015)的表 1。

⑦与 Ostrom et al(1992)的不同，他们在陌生人组(Stranger-treatment)的实验中选择了随机匹配方法，也就是说，各期实验的小组成员构成不同，从而排除了被试在互动中因为直接互惠或声誉考虑而选择合作或惩罚的可能。

⑧连洪泉等(2013)提到其中的三点质疑。

⑨从不同偏好类型来看，被试在与计算机玩家或人类玩家的实验中，公共品贡献量均值也基本相同，见 Burton-Chellew(2016)的图 1。换言之，被试之间的差异可以解释为他们如何理解最大化收益方面的差异。

⑩但仔细推敲，如果被试真正理解实验，强化的设计实验额外提供的两条信息实际上是多余的，因为被试可通过简单计算获得。

⑪详见原文的表 2，其中“other's success”的系数均为负值，又见原文图 1。

⑫作者还提及 ventromedial frontal/orbitofrontal cortex，rostral anterior cingulate cortex。

⑬参照汪丁丁(2011，p193)的作法，将 dorsolateral prefrontal cortex(简称 DLPFC)译为“前额叶侧背部”。

⑭英文文献中为“empathy”。有些中文文献翻译成“共情”，而此处参照亚当·斯密的《道德情操论》(中央编译出版社出版，2011 年版)与汪丁丁(2011)的做法，将其翻译为“同情”。

⑮人类合作的扩展秩序包括三层，人脑内部合作属于其中之一(汪丁丁，2011)。

⑯我们也注意到，一些学者尝试着结合“大五”人格模型，进一步探析了个体社会偏好稳定性的心理基础(Volk et al，2011)。学术界对于人类情感的研究由来已久，但由于人格模型的脆弱性和人类情感的微妙性，这方面的研究都注定仍任重道远。

⑰如诚信体系建设中的“黑名单”制度。

参考文献：

弗朗西斯·福山，2015：《大断裂：人类本性与社会秩序的重建》中译本，广西师范大学出版社。

乔治·阿克洛夫 罗伯特·希勒，2009：《动物精神》中译本，中信出版社。

塞缪尔·鲍尔斯 赫伯特·金迪斯，2015：《合作的物种——人类的互惠性及其演化》中译本，浙江大学出版社。

埃利诺·奥斯特罗姆，2010：《集体行动如何可能？》，《华东理工大学学报》第 2 期。

陈叶烽，2009：《亲社会性行为及其社会偏好的分解》，《经济研究》第 12 期。

陈叶烽，2010：《社会偏好的检验：一个超越经济人的实验研究》，浙江大学博士论文。

程又中，2006：《国际合作研究丛书》，世界知识出版社。

胡颖廉，2014：《社会治理创新：更关注“社会”》，《学习时报》10 月 13 日。

黄少安，2000：《经济学研究重心的转移与“合作”经济学构想》，《经济研究》第 5 期。

黄少安 韦倩，2011：《合作行为与合作经济学：一个理论分析框架》，《经济理论与经济管理》第 2 期。

连洪泉等，2013：《惩罚机制真能解决搭便车难题吗？——基于动态公共品实验的证据》，《管理世界》第 4 期。

连洪泉，2014：《惩罚与社会合作——基于实验经济学的讨论》，《南方经济》第 9 期。

连洪泉 周业安，2015：《异质性和公共合作：调查和实验证据》，《经济学动态》第 9 期。

孟昭勤 王一多，2004：《论人类社会的竞争与合作》，《西南民族大学学报(人文社会科学版)》第 7 期。

宋紫峰 周业安，2011：《收入不平等、惩罚和公共品自愿供给的实验经济学研究》，《世界经济》第 10 期。

汪崇金，2013：《公共品自愿供给的实验研究——基于强互惠理论视角》，上海财经大学博士论文。

汪崇金 聂左玲，2015：《破解社会合作难题：强互惠真的够强吗？》，《外国经济与管理》第 5 期。

汪崇金 聂左玲 岳军，2012：《个体异质性，预期与公共品自愿供给——来自中国的经济学实验证据》，《财贸经济》第 8 期。

汪崇金 史丹，2016：《利他惩罚威胁足以维系社会合作吗？——一项公共品实验研究》，《财贸经济》第 3 期。

史丹 聂新伟，2014：《电力贸易的制度成本与 GMS 电力合作中的中国选择》，《财贸经济》第 9 期。

汪丁丁，2011：《行为经济学讲义：演化论的视角》，上海人民出版社。

王道勇，2014：《社会合作：现代社会治理的最大难题》，《领导科学》第 6 期。

韦倩，2010：《强互惠理论研究评述》，《经济学动态》第 5 期。

韦倩 姜树广，2013：《社会合作秩序何以可能：社会科学的基本问题》，《经济研究》第 11 期。

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》第 8 期。

张康之，2012：《合作治理是社会治理变革的归宿》，《社会科学研究》第 3 期。

张康之，2013：《论共同行动中的合作行为模式》，《社会学评论》第 12 期。

张维迎，2015：《经济学原理》，西北大学出版社。

周业安，2015：《论偏好的微观结构》，《南方经济》第 4 期。

周业安等，2013：《社会角色、个体异质性和公共品自愿供给》，《经济研究》第 1 期。

周业安 宋紫峰，2008：《公共品的自愿供给机制：一项实验研究》，《经济研究》第 7 期。

周晔馨 涂勤 胡必亮,2014:《**惩罚、社会资本与条件合作——基于传统实验和人为田野实验的对比研究**》,《**经济研究**》**第 10 期**。

朱富强,2009:《**主流经济学中的经济人:内涵演变及其缺陷审视**》,《**财经研究**》**第 4 期**。

Andreoni, J. (1995), "Cooperation in public goods experiments: Kindness or confusion", *American Economic Review* 84(4):891－904.

Andreoni, J. (1988), "Why free ride? Strategies and learning in public goods experiments", *Journal of Public Economics* 37(3):291－304.

Ashley, R. et al(2010), "Motives for giving: A reanalysis of two classic public goods experiments", *Southern Economic Journal* 77(1):15－26.

Balafoutas, L. & N. Nikiforakis(2012), "Norm enforcement in the city: A natural field experiment", *European Economic Review* 56(8):1773－1785.

Bereby-Meyer, Y. (2005), "Reciprocity and uncertainty", *Behavioral and Brain Sciences* 35(1):18－19.

Binmore, K. (2005), *Natural Justice*, Oxford University Press.

Bochet, O. et al(2006), "Communication and punishment in voluntary contribution experiments", *Journal of Economic Behavior & Organization* 60(1):11－26.

Bolton, G. E. & R. Zwick(1995), "Anonymity versus punishment in ultimatum bargaining", *Games and Economic Behavior* 10(1):95－121.

Bowles, S. & H. Gintis(2002), "Behavioural science: Homo reciprocans", *Nature* 415(6868):125－128.

Buckholtz, J. W. et al(2008), "The neural correlates of third-party punishment", *Neuron* 60(5):930－940.

Burton-Chellew, M. N. & S. A. West(2013), "Prosocial preferences do not explain human cooperation in public-goods games", *Proceedings of the National Academy of Sciences of the USA* 110(1):216－221.

Burton-Chellew, M. N. et al(2015), "Payoff-based learning explains the decline in cooperation in public goods games", *Proceedings of the Royal Society of London B: Biological Sciences* 282 (1801):2014－2678.

Burton-Chellew, M. N. et al(2016), "Conditional cooperation and confusion in public-goods experiments", *Proceedings of the National Academy of Sciences of the USA* 113(5):1291－1296.

Carpenter, J. P. (2007), "The demand for punishment", *Journal of Economic Behavior & Organization* 62(4): 522－542.

Chaudhuri, A. (2011), "Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature", *Experimental Economics* 14(1):47－83.

De Quervain, D. et al(2004), "The neural basis of altruistic punishment", *Science* 305(5688):1254－1258.

Denant-Boemont, L. et al(2007), "Punishment, counter-punishment and sanction enforcement in a social dilemma experiment", *Economic Theory* 33(1):145－167.

Doherty, J. et al(2004), "Dissociable roles of ventral and dorsal striatum in instrumental conditioning", *Science* 304 (5669):452－454.

Duesenberry, J. (1949), *Income, Saving and the Theory of Consumer Behavior*, Harvard University Press.

Fehr, E. & S. Gächter(2000), "Fairness and retaliation: the economics of reciprocity", *Journal of Economic Perspectives* 14(3):159－181.

Fehr, E. & S. Gächter(2002), "Altruistic punishment in humans", *Nature* 415(6868):137－140.

Fischbacher, U., S. Gächter & E. Fehr(2001), "Are people conditionally cooperative? Evidence from a public goods experiment", *Economics Letters* 71(3):397－404.

Fischbacher, U. & S. Gächter(2010), "Social preferences, beliefs, and the dynamics of free riding in public goods experiments", *American Economic Review* 100 (1): 541－556.

Fudenberg, D. & P. A. Pathak(2010), "Unobserved punishment supports cooperation", *Journal of Public Economics* 94(1－2):78－86.

Frank, R. (1988), *Passions within Reason: The Strategic Role of the Emotions*, Norton.

Gächter, S. (2012), "In the lab and the field: Punishment is rare in equilibrium", *Behavioral and Brain Sciences* 35 (1):26－28.

Gehrig, T. et al(2007), "Buying a pig in a poke: An experimental study of unconditional veto power", *Journal of Economic Psychology* 28(6):692－703.

Guala, F. (2012), "Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate", *Behavioral and Brain Sciences* 35(1):1－59.

Güney, S. & B. R. Newell(2012), "Is strong reciprocity really strong in the lab, let alone in the real world", *Behavioral and Brain Sciences* 35(1):29－29.

Güth, W. et al(1982), "An experimental analysis of ultimatum bargaining", *Journal of Economic Behavior and Organization* 3:367－388.

Hauert, C. et al(2007), "Via freedom to coercion: the emergence of costly punishment", *Science* 316:1905－1907.

Herrmann, B. & C. Thöni(2009), "Measuring conditional cooperation: A replication study in Russia", *Experimental Economics* 12(1):87－92.

Houser, D. & R. Kurzban(2002), "Revisiting kindness and confusion in public goods experiments", *American Economic Review* 92(4):1062－1069.

Isaac, R. M. & J. M. Walker(1988), "Group size effects in public goods provision: the voluntary contributions mechanism", *Quarterly Journal of Economics* 103(1):179－199.

Janssen, M. A. & C. Bushman(2008), "Evolution of cooperation and altruistic punishment when retaliation is possible", *Journal of Theoretical Biology* 254(3):541－545.

Johnson, T. (2012), "The strategic logic of costly punishment necessitates natural field experiments, and at least one such experiment exists", *Behavioral and Brain Sciences* 35(1): 31－32.

Kamei, K. & L. Putterman(2013), "In broad daylight: Fuller information and higher-order punishment opportunities can promote cooperation", *Journal of Economic Behavior & Organization* 120:145－159.

Kawagoe, R. et al(1998), "Expectation of reward modulates cognitive signals in the basal ganglia", *Nature Neuroscience* 1(5):411－416.

Knoch, D. et al(2006), "Diminishing reciprocal fairness by disrupting the right prefrontal cortex", *Science* 314(5800):829－832.

Kummerli, R. et al(2010), "Resistance to extreme strategies, rather than prosocial preferences, can explain human cooperation in public goods games", *Proceedings of the National Academy of Science of the USA* 107(22):10125－10130.

List, J. A. & T. L. Cherry(2000), "Learning to accept in ultimatum games: Evidence from an experimental design that generates low offers", *Experimental Economics* 3(1):11－29.

Muller, L. et al(2008), "Strategic behavior and learning in repeated voluntary contribution experiments", *Journal of Economic Behavior and Organization* 67(3):782－793.

Nikiforakis, N. (2008), "Punishment and counter-punishment in public good games: Can we really govern ourselves?", *Journal of Public Economics* 92(1－2):91－112.

Ostrom, E. (2012), "Experiments combining communication with punishment options demonstrate how individuals can overcome social dilemmas", *Behavioral and Brain Sciences* 35(1):33－34.

Rand, D. et al(2010), "Anti-social punishment can prevent the co-evolution of punishment and cooperation", *Journal of Theoretical Biology* 265(4):624－632.

Richerson, P. J. & R. Boyd(2005), *Not by Genes Alone: How Culture Transformed Human Evolution*, University of Chicago Press.

Rilling, J. K. & A. G. Sanfey(2011), "The neuroscience of social decision-making", *Annual Review of Psychology* 62(1):23－48.

Rilling, J. K. et al(2007), "Neural correlates of social cooperation and non-cooperation as a function of psychopathy", *Biological Psychiatry* 61(11):1260－1271.

Rilling, J. K. et al(2008), "The neural correlates of the affective response to unreciprocated cooperation", *Neuropsychologia* 46:1256－1266.

Rilling, J. K. et al(2002), "A neural basis for social cooperation", *Neuron* 35:395－405.

Ross, D. (2006) "Evolutionary game theory and the normative theory of institutional design: Binmore and behavioral economics", *Politics, Philosophy, and Economics* 5:51－79.

Rustagi, D. et al(2010), "Conditional cooperation and costly monitoring explain success in forest commons management", *Science* 330(6006):961－965.

Schultz, W. & R. Romo(1988), "Neuronal activity in the monkey striatum during the initiation of movements", *Experimental Brain Research* 71(2):431－436.

Sefton, M. et al(2007), "The effect of rewards and sanctions in provision of public goods", *Economic Inquiry* 45(4):671－690.

Singer, T. et al(2006), "Empathic neural responses are modulated by the perceived fairness of others", *Nature* 439(7075):466－469.

Suzuki, S. et al(2011), "Neural basis of conditional cooperation", *Social Cognitive and Affective Neuroscience* 6(3):338－347.

World Bank(2015), *World Development Report 2015: Mind, Society, and Behavior*, http://www.worldbank.org/en/publication/wdr2015.

Veblen, T. (1934), *Essays in Our Changing Order*, Viking Press.

Volk, S. et al(2011), "Temporal stability and psychological foundations of cooperation preferences", *Journal of Economic Behavior and Organization* 81(2):664－676.

Wu, J. J. et al(2009), "Costly punishment does not always increase cooperation", *Proceedings of the National Academy of Sciences* 106(41):17448－17451.

(责任编辑:李仁贵)
(校对:刘新波)