·专题一:双清论坛“行为科学与经济政策设计”·

社会困境下提升合作水平的机制设计:理论与证据*

彭咏淳[1] 郑 捷[2**]

1. 清华大学 经济管理学院,北京 100084
2. 山东大学 经济研究院,济南 250100

[摘 要] 坚持合作、不搞对抗是构建人类命运共同体的时代要求。在现实社会的复杂环境中,探究什么是有利于合作的机制,以及如何执行该机制的研究问题变得愈发重要。社会困境中的合作一直是行为与实验经济学领域的重大问题之一。本文综合国内外的研究动态,总结了八类常见的合作机制:声誉、示范、渐进、承诺、奖惩、评价、互惠与反馈,并从行为经济学决策理论的角度分析了不同机制的优缺点及其背后的行为动机。最后本文列出了四点未来关于合作机制的研究展望。为提升合作机制的效果,结合现实挑战,本文提出在现有机制基础上发展创新的方向,为广大研究学者推进合作机制的研究提供有益的借鉴。

[关键词] 社会困境;合作机制;行为与实验经济学;博弈论;机制设计

1 社会困境中的合作问题

习近平总书记在建党100周年的“七一”重要讲话中指出,“坚持合作、不搞对抗,坚持开放、不搞封闭,坚持互利共赢、不搞零和博弈”[1]。合作取代对抗是历史潮流,改革开放的成果已经证明了开放与合作的重要性。在经济全球化的时代,全球产业链分工体系逐渐成熟,个人、企业甚至国家之间以合作为主题的活动,正以前所未有的势头推进。这体现在,国际间高校合作的科学基金项目数量不断增加,“一带一路”中的技术交流成果数量日益提升[2,3]。关于未来的“十四五”规划文件也明确强调要“实行高水平对外开放,开拓合作共赢新局面”[4]。

合作行为一直是经济学与心理学等领域的重要课题,对于人类合作行为的演进研究从社会困境诞生以来未曾停止[5]。“现代经济学之父”Adam Smith认为人们在良好的市场条件下通过分工合作,尽管每个人都只考虑最大化自己的利益,最后也能最大化社会整体的福利。但是现实中仍普遍存在个体利益与社会利益互相冲突的情况,构成了对合作的挑战。这种现象被称为社会困境(Social Dilemma)。一个经典例子就是囚徒困境,其最早是在1950年由美国RAND公司的Merrill Flood与Melvin Dresher拟定,后来被Albert Tucker使用囚徒的方式完整化表述。囚徒困境表述的是两名囚徒

郑捷 山东大学经济研究院教授、博士生导师,山东大学特聘教授,山东大学理论与实验经济学研究中心主任。国际学术期刊 *Journal of Economic Behavior and Organization* 副主编,*Research in Economics* 副主编,担任多个SSCI期刊客座主编。研究领域包括信息经济学、实验经济学、行为经济学、产业经济学,研究主题涵盖机制设计、市场设计、信息设计等经典问题和合作协调、信任互惠、参照依赖等行为问题。

彭咏淳 清华大学经济管理学院博士生。研究领域为行为经济学、实验经济学,研究主题包括参照依赖、风险偏好、社会困境和合作机制,研究工作获中国行为与实验经济学论坛优秀论文奖。

收稿日期:2023-08-16;修回日期:2023-10-01
* 本文根据第338期“双清论坛”讨论的内容整理。
** 通信作者,Email: zhengjie@sdu.edu.cn
本文受到国家自然科学基金项目(72073080, 71873074)的资助。

在无沟通的情况下，可以都选择沉默以达到较轻的判决结果，但是因为每名囚徒都可以举报对方的罪行来使自己单独获得最轻的判决，因此双方的最优策略都是举报对方，从而达成了对于双方整体来说不是最优、而是互相举报的纳什均衡。社会困境还有猎鹿人博弈，公共品困境等多种类型，其在政治、经济甚至自然界都有着相当广泛的应用。

在现实的社会困境中，合作时刻面临着沟通困难、利益分配、文化差异和利益冲突等挑战。为了在不同情境中设计合适的机制以促进合作行为，避免分裂对抗，我们急需建立关于各类合作机制的理论与实验体系，实现因地制宜、对症下药。本文总结了声誉、示范、渐进、承诺、奖惩、评价、互惠和反馈八类常见合作机制，基于目前的国内外研究进展，回顾并分析了各类机制的深层原理与实际表现，并评价其优缺点。进而提出了四点关于未来合作机制研究方向的展望。

2　促进合作的八种机制

合作与对抗博弈中的参与对象可以简单分类为发起方、接收方与第三方。需要注意的是，发起与接收是相对己方动作而言的，在同时博弈中，发起方同时也可以是对方的接收方。第三方虽然不直接参与博弈，但可以为双方提供客观的协助。八种合作机制与主体之间的关系如图 1 所示。其中，声誉、示范、渐进和承诺是由发起方主动实现的，而奖惩、评价、互惠和反馈则是接收方的回应。而第三方可以参与声誉、示范、奖惩和评价四种机制，为双方提供必要的协助。下文将分别阐述八种机制。

2.1　声誉机制

声誉是构成社会身份的一个重要特质，代表了他人对于自己是否值得信赖的衡量。在互联网时代，不止是个人，企业机构乃至国家都愈发注意塑造并维护自己的品牌或组织形象。在完全的陌生关系中，声誉作为一个公共信号可以发挥类似于社会背书的作用。因为声誉来源于过去的行为，在缺乏当前信息的情况下，过往的历史有助于帮助他人识别合作或背叛的行为，在历史中保持高声誉的个体也往往能得到多方的信赖[6]。声誉系统的核心在于建立信任关系。最为有效的信任关系是由双方直接建立的，但是这要求双方是重复互动的，现实中更多的可能是高流动性的交流。此时间接的声誉机制也能在一定程度上发挥替代作用[7]。在这类间接互惠的声誉系统中，即使双方之间没有任何直接的联系，但他们能通过之前与其他人建立的声誉间接地建立信任关系。无论是对方信任他人或者是对方被他人信任的历史都有助于建立双方之间的信任关系[8]。Stahl[9]就通过建立一个简单的颜色标注信誉机制，成功提高了美国大学生在重复囚徒困境中的合作率。该颜色声誉系统使用绿色和紫色的标签来代表声誉的好坏，随着重复博弈经验的增加，原本逐渐下降的合作率反而开始上升。该理论在现实中存在诸多应用，如国务院办公厅指示要加强个人诚信体系建设。中国人民银行建立了征信中心，为个人和企业提供信用记录；法院会公布失信被执行人名单；商会等行业组织也建立了会员信用档案。这些声誉机制都为商业活动的正常进行提供了可靠的保障[10]。

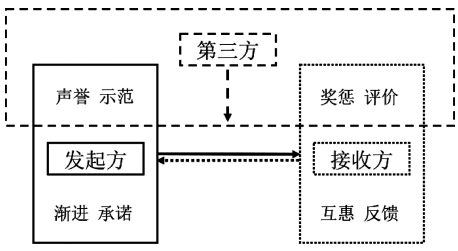

图 1　合作机制与主体关系图

虽然建立声誉系统是一种广受认可且行之有效的实现合作的方法，但仍然存在一些缺点。首先声誉系统的设置不是一成不变的，需要因地制宜。Abraham 等[11]从信息交流成本和主客观评价标准等方面对于声誉系统的设置作了详细的测试。他们发现信息传输成本的增加会相应减少声誉系统的使用。客观信息在被大量使用时能有效增加个体之间的信任与合作。而主观评价在私人交流中则无明显效果，只有将其放到公共层面才能有所作用。过去的声誉信息本身并不一定能产生足够的影响，因为对于部分始终坚持利己主义的个体，声誉系统的存在与否并不一定会左右他们的选择，即难以对完全不合作的群体产生激励[12]。

2.2　示范机制

在社会活动中，人们总是在观察别人的行为，并与自己进行比较。因为比较对象的不同，合作决策也可能产生不同的变化。例如在三角结构中，人们更愿意和朋友而不是与敌人合作。尤其是朋友和敌人同时存在时，因为存在参照对比，两者的合作水平显著不同[13]。常见的参照对象包括领导者、朋友和多数人的行为。特别是当人们缺乏主见，不清楚该做出何种行为时，经常会模仿显眼的个体或群体中多数人的行为，这分别被称为领头羊效应和同伴效

应。大量证据表明，人们会愿意服从道德劝说，去做所谓对的事情[14]。

当缺乏普遍共识时，领导者可以直接将理想的目标标注出来，作为一个焦点(Focal Point)，告知所有人，这种目标设定的方法又被称为助推(Nudge)。Barro 等[15]认为焦点可以为不同的个体提供清晰直接的统一参照点，从而更容易完成协作，提升整体的合作率。现实组织中常见的示范来自组内最显著的个体，即领导者。无论是在实验室还是现实社会，领导者对于跟随者的影响都得到了验证。Gächter 等[16]在公共品博弈中发现，作为领导者的个体在初期对于其他成员的信念和贡献有着强烈的影响，后期人们才会将注意力更多地转至其他成员身上，而此时群体中产生的路径依赖效应已经难以扭转，从而导致了从领导者到群体的过渡，削弱了领导者个人的示范效应。周业安等[17]认为无论是搭便车的追随者，还是不作为的领导者，都有可能是领导者榜样作用有限的原因。领导者和追随者实际策略的不确定性给示范机制的设计带来了很大的挑战。行为实验中，参与者的选择范围常常是有限的，但现实中的领袖们具备通过选举、宣传和公开发言等多种手段来增强自身影响力。Jack 等[18]在南美洲的玻利维亚进行的实地实验发现，政府当局对于居民参与公共品的贡献具有很强的正向示范效应，这种影响来源于其正式的领导地位、政策的特点和当局的实际贡献等方面。

另一种示范机制，即同伴效应，取决于身边多数人的行为。俗语“近朱者赤，近墨者黑”的一种解释就来源于人们的服从一致性[19]。当人们决策前知晓了其同伴的行为，其后续的行为也会与同伴的平均行为更加接近。与合作意愿强的同伴一起，自己选择合作的可能性越高，而在搭便车的环境中，受到自私行为影响的程度则会更严重[20]。因此，正如孟母三迁的典故所揭示，选择或打造合适的环境对于助推大部分人的行为尤为重要。

2.3 渐进机制

如果一步实现所有目标比较困难，那分为多个阶段一步一步实现就相对容易。渐进机制的原理在拍卖机制中已经有了较为成熟的运用[21]。Ye 等[22]比较了一步到位式和循序渐进式的协作过程，发现后者具有显著更高的合作水平。因为在渐进式环境下，团队先从较容易实现的目标开始合作，此时个体对于对方贡献的估计可能性较高。随着协作不断推进，目标的难度逐渐升高，但是过往成功合作的路径依赖和顺序溢出效应，使得循序渐进机制相较于一步达成最高目标更具有可行性。

在公共品博弈中，一种常见的分类是离散与连续(实时)的贡献机制，区别在于个体能否实时地调整自己每次贡献的水平。在实时贡献调整机制中，个体能自主地从较低的贡献水平开始，视他人贡献情况逐步提升自身贡献水平。这样的策略能够避免因他人背叛而导致自身损失过多的情况。研究发现，个体在观察到他人的投入水平后，会不断调整自己的贡献到比组内最低稍微高一些的水平，从而实现组内平均贡献水平如同棘轮一般的连续增长[23, 24]。但是渐进机制也有失败的设计：Gallier 等[25]在公共品博弈中外生规定本轮的贡献不能低于上一轮的贡献，个体因为会预计到未来不能下调贡献水平，所以将起始贡献严格限制在很低的水平，且增长缓慢，后续的贡献增长也并没有弥补前期的损失。对于渐进机制，个体的自主探索和调整最终可能实现整体的合作率增长，而严格的限定反而会强行造成上升的压力，从而降低个体的合作积极性。

2.4 承诺机制

承诺代表个体对于自己将来一定要做出某种行为的约定表述。遵守承诺在各类文化里都被认为是传统美德。成语一诺千金也谕示人们对于承诺的看重。相反对于经常违背承诺的人，信任与合作都是更加困难的事情。如果要求人们对于自己的行为发誓，那么事后违背誓言的概率会有所降低。即使个体事前决定要违背誓言，事后也会花费更多成本来进行权衡取舍[26]。Belot 等[27]使用荷兰一档名为“是否会分享”电视节目的数据发现，有过承诺的一方比起未承诺的一方的合作可能性要高 50%。人们不是自愿撒谎，但在被迫的情况下并不对撒谎感到愧疚。类似的，在公共品博弈中，发誓机制也能将公共品的平均贡献比例提升 33%[28]。

为了便于被试承诺，实验者通常在实验前加入交流环节。在没有复杂策略性考虑的情形下，即使是单项的交流也能促进双方的合作[29]。但是另一方面，关于合作的承诺也会使得利己主义者更容易找到背叛的目标，从而使合作承诺变成一种危险的信号，进而使整个承诺机制崩溃[30]。因为承诺本身并没有统一的格式，所以可能受到性别差异和时间压力等多种因素的影响，从而缺乏可比较的标准[31, 32]。一种解决方法是通过加入撒谎成本或承诺保险，将承诺的激励内生化。个体如果违背了之前的承诺，那么就需要付出相应的物质代价[33]。例

如，个体在参与博弈前先为自己的行为提交保证金，如果自己没有做出与承诺不同的行为，保证金就能原额退回，而如果违反了承诺，保证金就会被没收。这类带有物质激励的承诺保证会改变社会困境的利益结构，从而创造新的合作均衡[34, 35]。

2.5　奖惩机制

团体组织中，对于为谋求私利而损害集体利益的行为常常有惩罚规则，而对于为集体服务的贡献行为也有对应的奖励规则。国家的法律鼓励公民惩恶扬善，军队中赏罚分明的纪律能有效地规范军人的行为，球队会让表现不好的球员“坐板凳”，也就是暂停上场，这在公共品博弈中也有类似的应用[36]。在社会困境中，奖励与惩罚是实现合作的重要机制之一[37]。即使该奖励和惩罚需要团队成员自己去付出成本，也有成员愿意自发去执行该规则[38]。从个体的角度看，奖惩的动机并不一定是为了促进共同的合作，事实上相当一部分非合作个体为了提升自身收益也会有动机惩罚搭便车行为。因此，合作与奖惩是本质不同的两类行为[39]。Leibbrandt 等[40]认为，大部分人支持实行奖惩规则的动机是基于互惠心理和对于不公平的厌恶，即使奖惩会损害自己的部分利益，但被视为执行正义的必须成本。不过，无论是奖励还是惩罚措施都并不一定需要额外的实施成本，Yang 等[41]发现让成员自主分配内部的税收收入就能在预算平衡的情况下促进群体成员的合作行为。

因为组织中的奖惩规则通常是由成员制定的，所以如何在团体中设置奖励和惩罚与成员的构成情况有着天然的密切关系。对于都由类似的合作者所构成的团体，成员构成信息本身就能起到提示作用：该团体并不需要太多的惩罚机制。相反，对于由类似的搭便车者所构成的团体，成员构成信息也能帮助他们更早地认识到需要严格的惩罚机制以保证合作[42]。人们会因为规则不同，而对不同团体有不同偏好，这种现象被称为机构偏好（Institutional Preference）。Yang 等[43]在中国西北地区农村的实地实验发现，人们更倾向于选择奖励而不是惩罚机制。Sutter 等[44]认为给予成员自主选择奖惩规则的权利能有效提升团体的合作水平。类似地，Marci 等[45]也发现如果是由内部推举的个体所执行的惩罚机制会更加温和，即使惩罚力度比外生指定的更低，也能达到同样的合作水平。但另一方面，宋紫峰等[46]发现基于群体的惩罚机制的威慑效果并不如任何个体都能独立做出惩罚决策的情况，因为后者可以被视为一个更难达成的门槛公共品。

2.6　评价机制

人类是社会性动物，会在意他人对自己的评价，即使该评价本身并不会造成物质财富上的影响。社交软件基本都具备分享功能，例如微信中的朋友圈功能与 QQ 中的空间功能，可以被视作现实社交在互联网中的延伸。人们乐于在自己的交友圈中分享工作、生活、时事等内容。其他好友既可以通过“点赞”的二元评价方式来表达认同，也可以通过在评论区留言的评价方式来参与互动。在微博、知乎、小红书等网络平台，甚至陌生人也可以参与点赞分享。在分享与点赞的过程中，尽管所有人都没有收获直接的物质利益，但是人们依然希望自己的分享能够得到更多的点赞与正面留言，以寻求社会认同感，从而获得精神上的满足。

社会评价可以被视为来源于他人的非物质奖惩机制，在社会困境中能有效抑制自利行为。在更强调集体主义的社会中，无私奉献的行为会得到称赞，自私自利的行为则容易受到指责。即便是使用简单的高兴符号或愤怒符号来分别代表正面和负面评价，也能提升公共品博弈中的合作率水平[47]。尽管从长期来看物质方面的奖惩有较强的效果，但是社会评价系统具有执行成本更低的优点，因此两者各有优劣[48]。例如 Handgraaf 等[49]在节碳的实地实验中发现，社会层面的精神表彰比个人层面的物质表彰还要更加有效。一般而言，人们都希望自己的行为能得到更多人的认同，而不是批评，因此人们在社会活动中会自觉地约束自己的行为，避免受到他人的指责。即使该评价是事后的且不会有物质上的惩罚，人们依然不希望得到关于自己的负面评价[50]。

社会评价系统发挥作用的核心点在于评价与行为身份的一一绑定。在公共品博弈中，将人们的身份与贡献展示出来，使之受到社会的评价，会造成人群中合作行为的增多[51]。但如果是团队整体而不是个人的决策，社会评价的作用就会变得十分有限[52]。可见，社会评价的压力如果分摊到多人甚至集体层面，其效果远不如指向个体的单独评价。正因如此，当负面行为具有群体性特征时，“法不责众”的侥幸心理还是在人们的认知中广为流传[53]。

2.7　互惠机制

互惠心理一直被视为人类行为中的一类重要动机，能解释很多现实中非完全自利的现象[54]。在日常生活中，大部分个体更倾向于投桃报李，即报答曾

经给予我们恩惠的一方，并且希望能收获对方对自己善意行为的回报[55]。在社会困境中，如果自己善待了对方，也会期望对方能够同等程度地回报。此时的互惠行为就促进了双方的合作，达成了共赢。但如果目标是独赢，那么个体比起回报对方，会优先提升自己的收益，且会采取背叛或者搭便车的策略。互惠行为并不局限于两方互动的情景，多方之间也可以存在间接互惠的机制[7,8]。此时声誉机制与评价机制能提供重要的参考信息，以保证间接互惠机制能正确发挥作用。近年来相关机制的研究也代表了间接互惠行为的一个重要研究方向。另一类多方互动如多方公共品博弈的自愿贡献机制中，个体也倾向于把他人的贡献程度作为自己的标准，此时互惠行为可以被视为一种顺从他人的行为，或者称为条件性合作[56]。

一种观点认为互惠行为的动机来源于重复博弈中的策略性考虑。即使个体本身并没有利他倾向，也可能会为了长期更高的合作收益而选择互惠策略[57]。这类观点将互惠行为看作策略性考虑的结果，即便是利己的个体，在不同的环境中也会有意愿克服自利冲动，因为长期收益而选择互惠。另一类观点认为互惠心理与个人特质的相关度很高，例如Sabater-Grande等使用大五人格量表(The Revised NEO Personality Inventory)发现其中的宜人性(Agreeableness)指标能在很大程度上解释互惠行为[58]。即使不存在长期的利益考虑，人们也会因为希望双方都能共同更好而采取互惠行为。前一类观点强调个体对于长期利益的策略性考虑，后一类则注重个体内在的共情心理，两者都可能是互惠行为的产生来源。

2.8 反馈机制

反馈指的是由客体对于主体行为的信息呈现。不同于相对主观的评价机制，反馈机制侧重于客观的信息披露。互联网时代中，即时反馈在网络平台上越来越容易实现，该机制能显著加强买家与卖家之间的信任关系与合作[59]。例如消费者在网购后可以填写对于商品的具体问题反馈，例如“商品少件”“包装破损”“售价与实付金额不一”“发票未开”等问题。而商家一方也高度重视收到的反馈，售后部门根据收到的具体反馈为消费者提供替代或补偿方案。

在社会困境中，人们乐于为对方的行为提供反馈。Andreoni和Rao将其描述为“询问的力量”。在专断博弈中，接收方只能被动接受专断者的分配方案。但如果接收方有反馈的机会，即使该反馈只是来自接收方单方面的询问，该询问也能在一定程度上激发专断者的同情心，从而提高分配的公平性[60]。同一个市场中，商家愿意提供高成本反馈机制的行为本身是一个正面的信号，而相反如果商家缺乏相应的反馈机制，其一直保持沉默的行为就会带来负面的影响[61]。所以互联网平台中，头部商家的流量聚集愈发明显，而无人问津、缺乏反馈的平台则很难得到消费者的信任，从而加剧了平台中流量的马太效应。

反馈系统对于信息客观性的要求非常严格。无成本的反馈机制在长期内会因为缺乏监管和道德约束而逐渐失效[62]。失败的反馈机制还可能会造成更加恶劣的效果。反馈机制能发挥作用的关键在于信息是否公开客观。Lumeau等[63]比较了公共可见与私人可见的两类反馈机制，结果发现公共可见的机制更明显地提升了个体之间的信任，而私人可见机制在长期并不具备足够的影响力。因此，只有信息公共且客观的反馈机制才能正常地发挥反馈作用。

3 合作机制的研究展望

关于合作机制的研究数量庞大且种类丰富，但多数研究的主题都可以总结为两个问题。第一个问题是什么类型的合作机制是更有效的。在解决第一个问题后，第二个问题是应该如何去执行该有效的合作机制。过往研究多聚焦于单独的某类机制所能达到的效果，而在已有研究的数量和种类已经较为丰富的情况下，研究者可以总结出不同机制对应的优缺点以及根据应用环境的变化而具有的限制条件。因而未来关于合作机制的研究应该着力于开发不同合作机制的组合优势，更加全面地关注合作机制从设计到执行的动态过程。基于以上两个研究问题，本文提出了如图2所展示的四点研究展望。

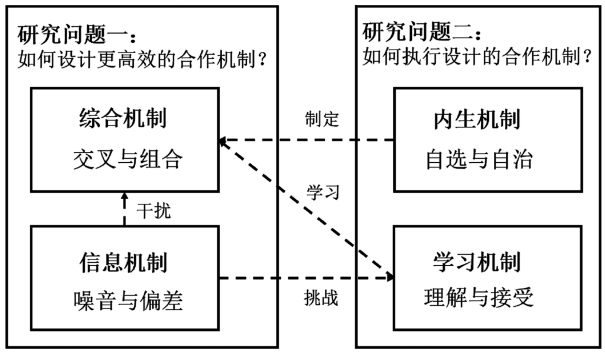

图2 研究展望关系图

3.1　综合机制

在第一、二部分中，我们总结了八类常见合作机制，其中部分机制的组合存在一些优势。例如，Wu 等[64]发现将示范机制与奖惩机制结合能从多方面更好地解释公共品困境中人们顺从行为的内在动机，从而促进整体的合作。近年来一些学者也开始考虑不同合作机制的对比效果。比如 Ye 等[22]比较了连续或离散投入情况。Masclet 等[48]比较了非物质的反馈与物质的奖惩的效果。连洪泉等[65]指出单纯的惩罚机制可能会引来被惩罚者的报复而弱化合作，而引入其他非惩罚机制可以降低惩罚机制的破坏效应。目前关于综合机制的研究还相对较少。如图 1 所示，部分机制在实施主体与方式上是比较类似的，而且多数机制是可以替代或组合使用的。例如，何韵文等[66]在投资博弈中实行的分期拨付机制综合了渐进、承诺和反馈机制的优点，从而更好地保护了投资委托人的利益。不同类型的合作机制交叉是否会产生冲突或协同效用？如何发挥组合优势，避免组合冲突？这些问题都还有待解决。因为机制设计上合理的组合有望突破单个机制的局限性，所以综合机制是最有望得到新发现的领域之一。

3.2　信息机制

已有的合作机制激励措施在很大程度上依赖于完全信息，这一点在声誉、评价与反馈机制中尤为明显。声誉机制侧重于建立自我的历史信息系统，评价机制针对主体本身整体性信息提供主观评定，反馈机制则对主体行为提供客观信息披露。可见三种机制尽管在具体的实行方式上有所不同，但本质都是为了使多方之间的信息更加透明，消除或降低信息不对称。减少信息不对称有助于促进合作行为，例如，何韵文等[67]在具有时效性的公共品供给困境的研究中发现，完美信息动态博弈中的活动发起时间和个体加入时间均要早于不完美信息动态博弈；可实时观测的信息机制能带来更高水平的社会福利。连洪泉等[68]也发现关于个体或群体异质性的公开信息可有效协调异质公平偏好个体的行动策略，从而更容易达到均衡。相反，如果观察到的信息存在噪声，就会降低信息的可靠性，从而减弱激励措施的有效性。例如电商平台中如果商家可以通过贿赂消费者来获取虚假的正面评价，或者消费者能通过策略性评价来打击商家，那么其他消费者就难以判断反馈的客观性，从而削弱了双方的信任关系[69-71]。因此，不完全信息环境中的噪音与偏差构成了对以信息为核心的合作机制的主要挑战。

3.3　内生机制

不同于静态的实验设定，现实中的规则是由个人单独或群体共同制定的。一个需要关注的现实问题在于什么规则是具有可行性的，即能使群体成员自觉遵守。即使是同一个规则，在由群体成员内生决定或者由外部强行给定的两种情况下，可能会有不同的实施效果。例如 Sutter 等[44]和 Marcin 等[45]就在研究机构偏好时发现了这一现象。内生决定规则的优势在于能将群体成员配属到所偏好的规则中，从而有效缓解群体内成员之间的偏见[68]。成员能发挥主观能动性，也就更愿意遵守自己选择的规则。周晔馨等[72]同时使用学生传统实验和工人实地实验发现，惩罚制度在不同群体中可能存在方向和机理都不一致的内生溢价，即相较于外生机制更能促进合作。内生机制也并不一定总是有效。例如，He 等[73]发现内生机制产生的纯自愿领导者的不合作行为，和领导者之间的内部利益冲突都会导致整体效果更差，反而不如随机指定领导者的机制。这是因为，在内生机制下人们会倾向于选择更有利自我的情境，从而在选择阶段就提前陷入社会困境。

3.4　学习机制

“忠言逆耳，良药苦口”。即使我们在理论上发现了更有效的合作机制，人们也可能因为理解或操作困难，不愿意遵守规则，从而无法有效发挥机制作用[35]。参与者的智力水平也是影响合作行为的关键因素，不同人可能对信息的接受度不同，从而增大了均衡实现的难度[74]。特别是在多类合作机制同时存在时，过于复杂的机制会增加人们的理解成本。此外，机制的高复杂度也可能催生复杂的策略性考虑，从而使机制难以达到原本的设计效果。对此，一个有效的解决方法就是事前模拟。实验证明，无论是面对真人还是与计算机互动，人们都能取得合格的学习效果。例如，Wang 等[75]发现参与者在与计算机匹配的重复囚徒困境博弈中，会通过学习逐渐适应计算机程序中的敲诈或慷慨策略来提高自己的收益。因此，如何设计既有效又让个体易接受的机制，并且提供有效的学习机制，引导参与者理解并遵守规则，避免机制设计成为空中楼阁，是机制设计与执行过程中不可缺失的一个环节。

4　总　结

全球化呼吁着个人、机构和国家之间开展全方面的合作。合作可以共享风险和资源，形成互惠关系。如同博弈论中的社会困境，只有各方通力合作，

克服私利，才能跨越困难，达成社会最优。为实现合作，机制的作用至关重要。本文总结了八类合作机制并提出了四点研究展望，作为合作机制研究的一个阶段性总结，希望为后续的研究提供有益的借鉴。目前关于合作机制的研究大部分还处于各自为营的状态，不同机制的交叉与融合仍是较少被关注的研究方向，而且无论是对于合作与竞争行为的内在动机与外在因素都缺乏足够的实验证据。如果我们能基于行为经济学理论与实验经济学方法，构建合作动机与行为的理论范式，就能为在各类社会困境中如何设计利于合作的机制的问题提供解决方案。建立一套关于合作机制的完整研究体系，能在各类复杂情境中，帮助设计者识别判断环境条件，合理运用不同机制的优势，为群体成员提供适宜合作的环境，避免掉入社会困境的陷阱。通过合理设计的合作机制，我们能发挥团结一致的优势，共同创造更加开放、和谐与高效的社会经济环境。

参 考 文 献

[1] 习近平. 在庆祝中国共产党成立 100 周年大会上的讲话. 党建，2021，7：4—9.

[2] 高扬，李冬，李崧维，等. 从国家自然科学基金看“世界一流大学建设高校”国际(地区)合作研究发展现状. 中国科学基金，2021，35(4)：650—656.

[3] 纪军，史翊翔，张永涛，等. 推进“一带一路”国家生物质资源化利用研究：基于中国和泰国组织间国际合作研究项目的分析. 中国科学基金，2019，33(2)：191—196.

[4] 中华人民共和国中央人民政府. 中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要. (2021-03-13)/[2023-07-31]. https://www.gov.cn/xinwen/2021-03/13/content_5592681.htm.

[5] 黄少安，张苏. 人类的合作及其演进研究. 中国社会科学，2013，7：77—89.

[6] Lunawat R. An experimental investigation of reputation effects of disclosure in an investment/trust game. Journal of Economic Behavior & Organization，2013，94：130—144.

[7] Bohnet I，Huck S. Repetition and reputation：implications for trust and trustworthiness when institutions change. American Economic Review，2004，94(2)：362—366.

[8] Charness G，Du NH，Yang CL. Trust and trustworthiness reputations in an investment game. Games and Economic Behavior，2011，72(2)：361—375.

[9] Stahl DO. An experimental test of the efficacy of a simple reputation mechanism to solve social dilemmas. Journal of Economic Behavior & Organization，2013，94：116—124.

[10] 中华人民共和国中央人民政府. 国务院办公厅关于加强个人诚信体系建设的指导意见. (2016-12-23)/[2023-07-31]. https://www.gov.cn/gongbao/content/2017/content_5160241.htm.

[11] Abraham M，Grimm V，Neeß C，et al. Reputation formation in economic transactions. Journal of Economic Behavior & Organization，2016，121：1—14.

[12] Kamei KJ，Kobayashi H，Tse TTK. Observability of partners' past play and cooperation：experimental evidence. Economics Letters，2022，210：110186.

[13] 连增，彭咏淳，郑捷. 三角结构中社会关系如何影响合作意愿：基于信念估计与社会偏好的视角. 第五届中国行为与实验经济学论坛，2023.

[14] Dal Bó E，Dal Bó P. “Do the right thing：” The effects of moral suasion on cooperation. Journal of Public Economics，2014，117：28—38.

[15] Barron K，Nurminen T. Nudging cooperation in public goods provision. Journal of Behavioral and Experimental Economics，2020，88：101542.

[16] Gächter S，Renner E. Leaders as role models and ‘belief managers’ in social dilemmas. Journal of Economic Behavior & Organization，2018，154：321—334.

[17] 周业安，黄国宾，何浩然，等. 领导者真能起到榜样作用吗？——一项基于公共品博弈实验的研究. 管理世界，2014，10：75—90.

[18] Jack BK，Recalde MP. Leadership and the voluntary provision of public goods：field evidence from *Bolivia*. Journal of Public Economics，2015，122：80—93.

[19] Thöni C，Gächter S. Peer effects and social preferences in voluntary cooperation：a theoretical and experimental analysis. Journal of Economic Psychology，2015，48：72—88.

[20] Isler O，Gächter S. Conforming with peers in honesty and cooperation. Journal of Economic Behavior & Organization，2022，195：75—86.

[21] 何韵文，郑捷. 拍卖机制与竞价行为：基于付费竞价式拍卖的理论与实验. 经济研究，2021，56(11)：192—208.

[22] Ye ML，Zheng J，Nikolov P，et al. One step at a time：does gradualism build coordination? Management Science，2020，66(1)：113—129.

[23] Dorsey RE. The voluntary contributions mechanism with real time revisions. Public Choice，1992，73(3)：261—282.

[24] Kurzban R，McCabe K，Smith VL，et al. Incremental commitment and reciprocity in a real-time public goods game. Personality and Social Psychology Bulletin，2001，27(12)：1662—1673.

[25] Gallier C，Sturm B. The ratchet effect in social dilemmas. Journal of Economic Behavior & Organization，2021，186：251—268.

[26] Jacquemet N，Luchini S，Rosaz J，et al. Truth telling under oath. Management Science，2019，65(1)：426—438.

[27] Belot M，Bhaskar V，van de Ven J. Promises and cooperation：evidence from a TV game show. Journal of Economic Behavior & Organization，2010，73(3)：396—405.

[28] Hergueux J，Jacquemet N，Luchini S，et al. Leveraging the honor code：public goods contributions under oath. Environmental and Resource Economics，2022，81(3)：591—616.

[29] Koukoumelis A, Levati MV, Weisser J. Leading by words: a voluntary contribution experiment with one-way communication. Journal of Economic Behavior & Organization, 2012, 81(2): 379—390.
[30] Camera G, Casari M, Bigoni M. Binding promises and cooperation among strangers. Economics Letters, 2013, 118(3): 459—461.
[31] Kleinknecht J. A man of his word? An experiment on gender differences in promise keeping. Journal of Economic Behavior & Organization, 2019, 168: 251—268.
[32] Zhang C, Rao YL, Houser D, et al. Trusting promises under pressure. Economics Letters, 2023, 225: 111046.
[33] Bahel E, Ball S, Sarangi S. Communication and cooperation in prisoner's dilemma games. Games and Economic Behavior, 2022, 133: 126—137.
[34] Huang S, Lien JW, Zheng J. Self-commitment for cooperation. ECNU Industrial Organization and Behavioral Economics Workshop, 2023.
[35] Hong F, Lien JW, Zheng J. Committed deduction for cooperation: theory and experiment. Wuhan University Behavioral and Experimental Economics Workshop, 2023.
[36] Lien JW, Zheng J. Getting benched when you give the least: an effective mechanism for public good provision. Beijing Normal University Conference on Experimental Economics, 2019.
[37] Chen JN, Lian Z, Zheng J. Self-serving reward and punishment: evidence from the laboratory. Scientific Reports, 2023, 13: 13997.
[38] Fehr E, Gächter S. Cooperation and punishment in public goods experiments. American Economic Review, 2000, 90(4): 980—994.
[39] Albrecht F, Kube S, Traxler C. Cooperation and norm enforcement-The individual-level perspective. Journal of Public Economics, 2018, 165: 1—16.
[40] Leibbrandt A, López-Pérez R, Spiegelman E. Reciprocal, but inequality averse as well? Mixed motives for punishment and reward. Journal of Economic Behavior & Organization, 2023, 210: 91—116.
[41] Yang CL, Zhang BY, Charness G, et al. Endogenous rewards promote cooperation. Proceedings of the National Academy of Sciences of the United States of America, 2018, 115(40): 9968—9973.
[42] Bühren C, Dannenberg A. The demand for punishment to promote cooperation among like-minded people. European Economic Review, 2021, 138: 103862.
[43] Yang XJ, Nie ZH, Qiu JY, et al. Institutional preferences, social preferences and cooperation: evidence from a lab-in-the-field experiment in rural China. Journal of Behavioral and Experimental Economics, 2020, 87: 101554.
[44] Sutter M, Haigner S, Kocher MG. Choosing the carrot or the stick? endogenous institutional choice in social dilemma situations. Review of Economic Studies, 2010, 77(4): 1540—1566.
[45] Marcin I, Robalo P, Tausch F. Institutional endogeneity and third-party punishment in social dilemmas. Journal of Economic Behavior & Organization, 2019, 161: 243—264.
[46] 宋紫峰，周业安. 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，2011，34(10)：35—54.
[47] Peeters R, Vorsatz M. Immaterial rewards and sanctions in a voluntary contribution experiment. Economic Inquiry, 2013, 51(2): 1442—1456.
[48] Masclet D, Noussair C, Tucker S, et al. Monetary and nonmonetary punishment in the voluntary contributions mechanism. American Economic Review, 2003, 93(1): 366—380.
[49] Handgraaf MJJ, Van Lidth de Jeude MA, Appelt KC. Public praise *vs.* private pay: effects of rewards on energy conservation in the workplace. Ecological Economics, 2013, 86: 86—92.
[50] López-Pérez R, Vorsatz M. On approval and disapproval: theory and experiments. Journal of Economic Psychology, 2010, 31(4): 527—541.
[51] Rege MR, Telle K. The impact of social approval and framing on cooperation in public good situations. Journal of Public Economics, 2004, 88(7/8): 1625—1644.
[52] Christens S, Dannenberg A, Sachs F. Identification of individuals and groups in a public goods experiment. Journal of Behavioral and Experimental Economics, 2019, 82: 101445.
[53] Guo F, He Y, Zheng J. Disapproval rating towards individuals versus towards groups. Economic Science Association World Meeting, 2019.
[54] Fehr E, Gächter S. Reciprocity and economics: the economic implications of *Homo Reciprocans*. European Economic Review, 1998, 42(3/4/5): 845—859.
[55] Croson R, Fatas E, Neugebauer T. Reciprocity, matching and conditional cooperation in two public goods games. Economics Letters, 2005, 87(1): 95—101.
[56] He Y, Lien JW, Yang Y, et al. A goose feather from a thousand miles away: a theory and experiment on reciprocity. Xiamen University Experimental Economics Workshop, 2023.
[57] Carrasco JA, Harrison R, Villena MG. Strategic reciprocity and preference formation. Journal of Economic Behavior & Organization, 2022, 203: 368—381.
[58] Sabater-Grande G, García-Gallego A, Georgantzís N, et al. The effects of personality, risk and other-regarding attitudes on trust and reciprocity. Journal of Behavioral and Experimental Economics, 2022, 96: 101797.
[59] Dellarocas C. The digitization of word of mouth: promise and challenges of online feedback mechanisms. Management Science, 2003, 49(10): 1407—1424.
[60] Andreoni J, Rao JM. The power of asking: how communication affects selfishness, empathy, and altruism. Journal of Public Economics, 2011, 95(7/8): 513—520.
[61] Gazzale RS, Khopkar T. Remain silent and ye shall suffer: seller exploitation of reticent buyers in an experimental reputation system. Experimental Economics, 2011, 14(2): 273—285.

[62] He Y, Wang Z, Xu B, et al. Trust under request versus trust with threat. (2022-05-06)/[2023-08-15]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4099357.
[63] Lumeau M, Masclet D, Penard T. Reputation and social (dis) approval in feedback mechanisms: an experimental study. Journal of Economic Behavior & Organization, 2015, 112: 127—140.
[64] Wu JJ, Li C, Zhang BY, et al. The role of institutional incentives and the exemplar in promoting cooperation. Scientific Reports, 2014, 4: 6421.
[65] 连洪泉，周业安，左聪颖，等. 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，2013，4：69—81.
[66] 何韵文，郑捷. 分期拨付机制与渐进式信任. 第六届中国制度经济学论坛，2023.
[67] 何韵文，郑捷. 社会信息环境、奉献行为与公共品供给困境. 第三届中国行为与实验经济学论坛，2021.
[68] 连洪泉，周业安，陈叶烽，等. 信息公开、群体选择和公共品自愿供给. 世界经济，2015，38(12)：159—188.
[69] Bolton GE, Greiner B, Ockenfels A. Engineering trust-reciprocity in the production of reputation information. Management Science, 2013, 59(2): 265—285.
[70] Bolton GE, Kusterer DJ, Mans J. Inflated reputations: uncertainty, leniency, and moral wiggle room in trader feedback systems. Management Science, 2019, 65(11): 5371—5391.
[71] Krügel JP, Paetzel F. The impact of fake reviews on reputation systems and efficiency. (2021-02-28)/[2023-08-15]. https://www.econstor.eu/bitstream/10419/242415/1/vfs-2021-pid-50192.pdf.
[72] 周晔馨，涂勤，胡必亮. 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，2014，49(10)：125—138.
[73] He Y, Zheng J. Promoting cooperation by leading: leader-selection mechanisms in public goods games. Shanghai International Studies University Economics Seminar, 2021.
[74] Proto E, Rustichini A, Sofianos A. Intelligence, personality, and gains from cooperation in repeated interactions. Journal of Political Economy, 2019, 127(3): 1351—1390.
[75] Wang ZJ, Zhou YR, Lien JW, et al. Extortion can outperform generosity in the iterated prisoner's dilemma. Nature Communications, 2016, 7: 11125.

Mechanism Design to Promote Cooperation in Social Dilemmas: Theory and Evidence

Yongchun Peng[1] Jie Zheng[2*]

1. *School of Economics and Management, Tsinghua University, Beijing* 100084
2. *Center for Economic Research, Shandong University, Jinan* 250100

Abstract Cooperation, rather than confrontation, is an indispensable requirement for a community with a shared future for mankind. Mechanism design and implementation aimed at promoting cooperation in diverse environments has become increasingly important. How to achieve cooperation in social dilemmas has been one of the most popular topics in the field of behavioral and experimental economics. This paper summarizes eight commonly studied mechanisms for cooperation based on the existing literature, which includes reputation, introducing a role model, gradualism, commitment, rewards and punishments, evaluation, reciprocity, and feedback. For each mechanism, their advantages and disadvantages, as well as the underlying behavioral motives, are discussed. Furthermore, this paper outlines four research aspects for future work on cooperation mechanism. To improve the effectiveness of cooperation mechanism and take into account the challenges in the real world, this paper proposes directions of innovation based on the existing mechanisms, providing important helpful implications for researchers in this area.

Keywords social dilemma; cooperation mechanism; behavioral and experimental economics; game theory; mechanism design

（责任编辑 崔国增 姜钧译）

* Corresponding Author, Email: zhengjie@sdu.edu.cn