Advances in Psychology 心理学进展, 2021, 11(4), 940-950
Published Online April 2021 in Hans. http://www.hanspub.org/journal/ap
https://doi.org/10.12677/ap.2021.114107

Hans汉斯

经济博弈中的公平理论及公平规范执行的群体偏见

肖　沿[1]，朱海东[1,2]，孙桂芹[1,2]，孟欢蕾[1]，张　俊[1]

[1]石河子大学师范学院，新疆　石河子
[2]石河子大学心理应用研究中心，新疆　石河子
Email: 86775709@qq.com, xiaoyan@stu.shzu.edu.cn

收稿日期：2021年3月21日；录用日期：2021年4月12日；发布日期：2021年4月22日

摘　要

公平是群际互动中最重要的规范之一，它是群体成员在一定的社会情境下如何行为的规范。历年来，经济学家和社会心理学家运用行为博弈的经典范式——最后通牒博弈范式及其相关变式探究个体的公平偏好、利他、互惠等行为。对于经济博弈中个体违背“理性人假设”的现象，学者们分别提出了不平等厌恶、互惠偏好、平等－互惠－竞争三种理论来解释其背后的心理机制。群体认同是指个体对其所属群体身份的知觉及其所付诸于该群体身份上的价值与情绪，直接影响着群际互动中人们的公平规范执行行为。研究者发现在群体偏见影响公平规范执行时会产生内群体偏爱(In-group Bias)和黑羊效应(the Black Sheep Effect, BSE)两种现象，表现为相比于外群体，个体更愿意接受内群体成员的不公平提议或者更多拒绝内群体成员的不公平提议。当前研究者分别通过纯粹偏好理论(Mere Preference Theory, MPT)和规范聚焦假设(Norms-Focused Hypothesis, NFH)来解释上述相悖的现象。未来研究应侧重促进公平理论与群体偏见与公平规范执行理论的交叉共融，拓宽群体偏见与公平规范研究的群体范围，增加得失框架下群体偏见与公平规范的研究。

关键词

经济博弈范式，公平理论，群体偏见，公平规范执行，内群体偏爱，黑羊效应

Fairness Theory in Economic Games and Group Bias of Fairness Norm Enforcement

Yan Xiao[1], Haidong Zhu[1,2], Guiqin Sun[1,2], Huanlei Meng[1], Jun Zhang[1]

[1]Normal College, Shihezi University, Shihezi Xinjiang
[2]Center for Psychological Application, Shihezi University, Shihezi Xinjiang
Email: 86775709@qq.com, xiaoyan@stu.shzu.edu.cn

Received: Mar. 21st, 2021; accepted: Apr. 12th, 2021; published: Apr. 22nd, 2021

文章引用: 肖沿, 朱海东, 孙桂芹, 孟欢蕾, 张俊(2021). 经济博弈中的公平理论及公平规范执行的群体偏见. *心理学进展, 11(4),* 940-950. DOI: 10.12677/ap.2021.114107

Abstract

Fairness is one of the most important norms in group interaction. It is a norm for how group members behave in a certain social situation. Over the years, economists and social psychologists have used the classic paradigm of behavioral games—Ultimatum game paradigm and its related variants to explore individual behaviors such as fairness preferences, altruism, and reciprocity. Regarding the phenomenon that individuals violate the "rational man hypothesis" in economic games, scholars have put forward three theories to explain the psychological mechanism behind it, such as inequality aversion, reciprocity preference and equality-reciprocity-competition. Group identity refers to some knowledge of one's group membership together with the value and emotional significance attached to that membership, which directly affects people's fairness norm enforcement during inter-group context. Researchers have found that when group biases affect the implementation of fair norms, there will be two phenomena: In-group Bias and the Black Sheep Effect, which are manifested in that individuals were more likely to accept unfair offer from in-groups or more likely to reject unfair proposals offered by in-group members compared with out-group members. Currently, norms-focused hypothesis and mere preferences theory have usually been used to explain the above contradictory phenomena. Based on this review, future research should focus on promoting the integration of fairness theory and group bias of fairness norm enforcement theory, broaden the group scope of group bias and fairness norm research, and enhance the research on group bias and fairness norm under the framework of gains and losses.

Keywords

Economic Game Paradigm, Fairness Theory, Group Bias, Fairness Norm Enforcement, In-Group Bias, Black Sheep Effect

1. 经济博弈的范式

公平(Fairness)是人类社会生活的基本规范之一，其实质是个体自我利益与他人利益之间进行的权衡。它出现在人类社会生活的各个方面，权利的公平、社会资源分配的公平、人际交往中社会成员相处关系的不偏不倚，都是一种公平的体现(李炜，2019)。我国自古以来就有“不患寡而患不均”的治国思想，在当今社会，公平也是社会主义核心价值观的基本内容。自 Adams (1965)提出“公平理论”后，国内外的学者便开始了在公平道路上的实践探索，“公平理论”的主旨在于比较自身和他人在过去或者现在的投入－产出比，从而判断一个阶段的公平状态。

研究者们借用资产分配任务来研究人类对公平的追求，其中最经典的研究范式是由 Güth，Schmittberger 和 Schwarze (1982)提出的最后通牒博弈范式(Ultimatum game，简称 UG)，该范式由两名被称之为提议者(proposer)和回应者(responser)的个体共同参与，两名参与者得到一定数额的金钱，由提议者对这笔金钱进行分配，若回应者接受提议者给出的分配方案，则两人各自得到相应数额的金钱；若回应者拒绝此分配方案，则二者均得不到钱。传统经济学理论的“理性人假设”(Neumann & Morgenstern, 1953)预期提议者会将自身利益最大化从而考虑将所得的金钱最小化的分配给回应者，而回应者则不会拒绝提议者任何给自己大于 0 的分配方案，因为拒绝意味着一分钱也没有(Camerer, 2003)。而 UG 范式驳斥了该

理论假设并将开始研究的目光着眼于公平规范的执行，研究者既研究个体作为提议者的行为，也研究个体作为接受者的行为，通常，提议者的分享行为可作为一种研究策略的契机，因为分配者可能会因为害怕接受者的拒绝导致收入减少而更多的分配给接受者一定数额不多于自己的金钱，对接受者的研究则主要探查的是个体对公平的态度。

在此基础上，Kahneman，Knetsch 和 Thaler (1986)发展出独裁者博弈范式(Dictator Game，简称 DG)，在该范式中回应者只能接受提议者提出的方案，没有表达观点的机会，这使得提议者成为了独裁者，从而更有可能对回应者做出不公平分配，这也更可能引起回应者的不公平感，研究者通常采用此种范式研究分享行为和利他水平(Bašić, Falk, & Kosse, 2020; Brañas-Garza et al., 2009; Engel, 2011)。

另一个研究公平决策的范式是第三方惩罚范式(Third-Party Punishment Game，简称 TPPG)，该范式中，旁观者(bystander)作为第三方，对提议者给出的分配方案进行回应和干预，旁观者本身不获得任何金钱数额，不受其直接影响，而是目睹他人遭受不公平，回应者获得的金钱数额完全由旁观者决定。该研究范式通常用于研究个体维护公平规范的倾向(Bašić et al., 2020; El Zein, Seikus, De-Wit, & Bahrami, 2019; Liu et al., 2018)。

在 UG 范式的基础上，还发展出另一种不公平范式(Inequity Game，简称 IG)，这时旁观者作为第三方，对另外两名参与者进行金钱分配，其中一名作为回应者，另一名作为被动接受者(Mcauliffe, Blake, Steinbeis, & Warneken, 2017)。相对于拥有惩罚权力的 UG 任务，在被剥夺惩罚权力的 IG 任务下面对不公平分配，回应者甚至会产生更强的不公平感，这一范式通常用于探究人们对公平准则的感知(Cheng et al., 2015)。

在众多关注公平的研究中，不公平状态又可以分成两类，一类是劣势不公平(disadvantageous inequity, DI)，即自己的分配收入少于他人；另一类是优势不公平(advantageous inequity, AI)，即自己的分配收入多于他人。上述四种公平范式中，研究者多采用不同的角度研究公平规范的执行。UG 和 DG 范式通常从提议者和回应者两方角度探查个体的合作、利他、公平感知等行为状态。有所差异的是，UG 范式通过从回应者的拒绝行为中探查个体的公平感知，DG 范式从提议者角度关注个体的公平和利他行为。因此，UG 回应者对不公平分配的拒绝率可以作为规避 DI 的指标，而 UG 和 DG 提议者分给对手的分配是否公平可以作为规避 AI 的指标。

TPPG 范式提供了一个新的视角，从旁观者角度，规避理性“经济人”的社会角色，提供了一个当事人不获益的视角，将研究的重点集中于公平规范的执行，这可能使得关注公平规范的旁观者倾向于拒绝任何不公平的方案，即拒绝优势不公平(AI)和劣势不公平(DI)的方案。同样，IG 范式也从第三方角度，尝试关注社会的公平分配，同时规避 AI 和 DI 而探讨单一的公平偏好。上述几种经济博弈范式及其研究的旨意如表 1 所示：

Table 1. The purpose of the study of different economic game paradigms
表 1. 不同经济博弈范式的研究旨意

经济博弈范式	研究对象		
	提议者	旁观者	回应者
最后通牒博弈(UG)	利他偏好 互惠偏好 公平偏好	/	互惠偏好 公平偏好
独裁者博弈范式(DG)	利他偏好 公平偏好	/	公平偏好
第三方惩罚范式(TPPG)	公平偏好		/
不公平范式(IG)	公平偏好		公平偏好

有趣的是，当研究者从个体视角来关注公平行为时，即从提议者角度关注实验研究时发现，在 UG 范式中，提议者会更多的将一定数额的金钱分配给回应者，而回应者也会拒绝一定数量的分配方案。具体来说，提议者分配金额的中数或者众数常在 40%~50%之间，平均数在 30%~40%之间；而低于分配总额 20%的分配方案被回应者拒绝的概率为 40%~60%，40%~50%份额的方案很少会被拒绝，且份额越大被反应者拒绝的概率越低(Glimcher, Camerer, & Fehr, 2008; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003)。也就是说，提议者倾向于给出较为公平的分配方案，而回应者倾向于拒绝不太公平的分配方案。针对这一违背“理性人”现象的出现，研究者提出了以下三种模型进行解释。

2. 公平理论

2.1. 不平等厌恶模型(Inequity Aversion)

由 Fehr 和 Schmidt (1999)提出，其内涵在于在资源分配中，人们既是追求利益最大化的“经济人”，又是关心他人收益，甚至为了惩罚那些破坏公平规范的人而甘愿付出一定代价的“社会人”。该模型由两人姓氏首字母命名为 FS 模型(FS model)，其假设人们会因为收益差距产生不公平感，进而损伤自身效用，并且在自身收益高于别人时，损伤较小，而在自身收益低于别人时不公平感更为强烈，损害较大，FS 模型个体 i 对收入 x 的效用函数式如下：

$$U_i(x)=x_i-\alpha_i\frac{1}{n-1}\sum_{j\neq i}\max\left\{x_j-x_i,0\right\}-\beta_i\frac{1}{n-1}\sum_{j\neq i}\max\left\{x_j-x_i,0\right\}$$

注：$U_i(x)$ 代表个体i实际的效用值，x_i 代表绝对收入的效用，$\alpha_i\frac{1}{n-1}\sum_{j\neq i}\max\left\{x_j-x_i,0\right\}$ 代表 DI 引发的效用损失，$\beta_i\frac{1}{n-1}\sum_{j\neq i}\max\left\{x_j-x_i,0\right\}$ 代表 AI 引发的效用损失。

在该模型中，个体的收入效用在绝对收入 x 的基础上，排除了喜欢受到不公平待遇或者自豪偏好倾向的人，除去了两种不公平状态引发的效用损失，从而厘清个体厌恶不公平的行为特点(Klaus et al., 2012；徐富明，李欧，邓颖，刘程浩，&史燕伟，2016)。

2.2. 互惠偏好模型

由 Rabin (1993)提出，其理论基础在于人们不是仅仅追求自己的物质利益，而是有额外的“社会”目标：人们通常对帮助他们的人表现出善意行为，对伤害他们的人表现出敌意行为，这是从社会互动的角度去探讨公平行为的产生。Rabin 开创性地在效用函数中加入了“友善函数”(Kindness Function)，其内涵在于心理博弈中，博弈人的反应函数不再只取决于博弈双方的战略选择，还与参与者对对手的信念(Belief)紧密相关(Geanakoplos, Pearce, & Stacchetti, 1989)。互惠偏好模型个体i的效用函数形式为：

$$U_i\left(a_i,b_j,c_i\right)=\pi_i\left(a_i,b_j\right)+\tilde{f}_j\left(b_j,c_i\right)\left[1+f_i\left(a_i,b_j\right)\right]$$

注：$U_i\left(a_i,b_j,c_i\right)$ 代表个体i的效用值，$\pi_i\left(a_i,b_j\right)$ 代表参与者i获得的物质效用部分，$\tilde{f}_j\left(b_j,c_i\right)\left[1+f_i\left(a_i,b_j\right)\right]$ 代表互惠偏好作用下给参与者带来的心理效用。

在该模型中，个体的收入效用在所获得的物质效用 $\pi_i\left(a_i,b_j\right)$ 基础上，增加了心理效用，其中友善函数 $I\left(f_i\left(a_i,b_j\right)\right)$ 和 $II\left(\tilde{f}_j\left(b_j,c_i\right)\right)$ 用于衡量博弈对方的善意程度，具体通过比较参与者的实际收益与期望收益的实际差距得到。

上述两种解释“理性人”偏差的模型分别从结果和动机上解释了经济博弈中的社会偏好，不平等厌

恶模型(IA)从自利和追求公平的动机出发，验证了个体的公平偏好，并在 UG、TPPG、IG 等范式中得到验证(Biella & Sacchi, 2018; Sun, Tan, Cheng, Chen, & Qu, 2015; Yamagishi et al., 2012)。互惠偏好模型从对方的动机和自己的期望角度证实了个体互惠偏好的存在，并在 UG 范式得到验证(Fehr, Fischbacher, & Gächter, 2002)。通过上述两种模型的解释，我们不难发现，上述效用函数的构造与经典的以自利公理为基础的“理性人”假设收益不同的一个最重要的特征是，在偏好中不仅关注自己的行为和自身的利益，还关注自身的行为对他人的影响，以及他人的行为过程和理由。

除此之外，我们需要知道的是经济博弈本就是掺杂社会互动的复杂行为过程，上述两种模型较为清晰地分离了博弈进程中动机和结果的交互作用，但仍然难以覆盖所有的经济博弈范式，例如 DG 范式。由于回应者只能接受提议者提出的方案，没有表达观点的机会，理性的“经济人”假设，提议者应该能最大化自己利益收入，而回应者收入只能是最小化，然而，根据 Forsythe 等人(1994)研究中呈现的结果显示，独裁者游戏中，提议者分配给回应者的金钱数额并不总是 0，也不总是趋于公平的分配方案，其余大量类似的研究也发现了一致的结果。这既无法用互惠偏好的理论来解释，也无法适用于公平偏好模型。因此，Bolton 和 Ockenfels (2000)提出了以下模型来解释此种现象。

2.3. 平等－互惠－竞争模型(Equity, Reciprocity, and Competition，简称 ERC)

该模型在关注个体自身的收益同时，相比于 FS 模型，同时关注个体对群体平均值的偏离，即与他人的相对份额比例(唐俊，2011，徐富明等，2016)。同时，在动机的角度，基于互惠偏好，更加关注个体纯粹的利他动机意图。从而驳斥了人们看起来是以自我为中心的公认的利己主义理论，是利他偏好在 DG 范式中的完整体现(Weiland, Hewig, Hecht, Mussel, & Miltner, 2012)。

3. 群体偏见与公平规范

经济博弈属于社会活动的一种，而个体的社会属性是众多影响经济博弈中个体公平感知及行为决策的因素中不容忽视的因素之一(韩小丽，田孟奇，田嘉乐，&苏文亮，2020；李雪莹&贾宁，2020)。公平作为群体最重要的社会规范，受到群体身份的影响。在经济博弈中，个体作为回应者对来自内群体的成员和来自外群体成员的金钱分配方案多产生截然不同的回应(Biella & Sacchi, 2018; Mcauliffe & Dunham, 2016; Mendoza, Lane, & Amodio, 2014; Schiller, Baumgartner, & Knoch, 2014; Wang et al., 2017; Wu, Leliveld, & Zhou, 2011；王益文等，2014；王珍珍&蒋文明，2016；张瀚月&赵玉芳，2018)。在上述研究中，通过比较被试对内、外群体成员在不同公平水平方案的接受情况，结果发现，就不公平的分配条件下，被试对内群体成员与外群体成员的接受率有显著差异，被试更多地选择接受内群体成员提出的不公平方案或者给予更弱的惩罚(Biella & Sacchi, 2018; Schiller et al., 2014; Wang et al., 2017；王益文等，2014；张瀚月&赵玉芳，2018)，而少部分研究者发现内群体成员的公平违背会导致个体作出更多的拒绝反应，进而增强了公平规范的执行(Abrams et al., 2014; Mendoza et al., 2014; Wu et al., 2011)。例如，Wang et al. (2017)通过 UG 范式发现，相比于中等不公平和公平的分配方案，被试对于来自内群体的不公平分配方案接受率更高。而 Wu 等人(2011)使用 DG 范式发现被试对于来自朋友的不公平分配方案态度更加消极。通过 TPPG 范式，Schiller 等人(2014)验证了在公平规范的执行中，个体表现出的群体偏见同时由对内群体的偏爱和对外群体的贬损造成。在公平规范的执行中，群体在不同经济博弈范式中表现的两种截然不同的现象，可分别称之为内群体偏爱(Ingroup Bias)和黑羊效应(Black Sheep Effect)。

内群体偏爱(In-Group Bias)与黑羊效应(The Black Sheep Effect, BSE)

事实上，早在上个世纪 70 年代，社会心理学家 Brewer (1979)就发现，个体在社会互动中对群体内成

员和群体外成员的态度有巨大差别，当个体以某种身份或标签被某一群组接纳，成为该群组的成员时，会受到该群组成员的优待，得到更多的资源或者获得组内成员的正面评价，即所谓的内群体偏爱(In-group Bias)。与此同时，对外群体成员分配更少资源、产生更多的敌意和消极评价的现象，被称之为外群体贬损(Out-group Derogation/Out-group Discrimination)。与此相反，Marques，Yzerbyt 和 Leyens (1988)提出的黑羊效应(the black sheep effect, BSE)表明相比于外群体成员，个体对带来坏处的内群体成员评价更为消极。并且，这种群组间异化效应(Inter-group Differentiation)和群体内偏爱可以共存于群体内，即，在群际互动中，个体可能会同时表现出内群体偏爱和内群体贬损(In-group Derogation)。这种内群体贬损的现象则被称之为黑羊效应。

在实际的研究中，当群体偏见与公平规范相遇时，个体将面临较难的抉择，而个体如何去应对内群体的不公平行为，抑或是如何解释内群体不公平行为背后的心理机制，成为研究者关注的重点问题，目前提出了两种理论来试图分别说明内群体偏爱和黑羊效应。

4. 公平规范执行中群体偏见产生的理论基础

4.1. 规范聚焦假设(Norms-Focused Hypothesis, NFH)

在社会交往中，个人通常期望与他人合作和得到公平对待(Dick, Haslam, Tyler, & Blader, 2001; Fehr & Fischbacher, 2004)。当一个人在与内群体成员的互动关系中，相比与外群体成员的互动，对公平产生更大的期望。而当公平准则被违反时，这些违规行为在来自群体内成员时相比于来自群体外成员时造成的感受更加强烈(Abrams et al., 2014; Biernat, Vescio, & Billings, 1999; Valenzuela & Srivastava, 2012)，表现为个体对内群体成员的不公行为给与严厉的惩罚和制裁。这表明群体认同会促进公平规范的执行，具体在研究情境中，Wu 和 Gao (2017)采用最简群体范式操纵群体身份，检验群体偏见和公平规范在 3~6 岁儿童之间的冲突，发现 5~6 岁女童相比于其他年龄阶段和男童个体更倾向于惩罚内群体违规者。同样在种族差异的实验中，也证实了同种种族的成员对内群体的公平规范违背更加不满(Mendoza et al., 2014)。

规范聚焦假设认为黑羊效应，从个体层面和群体层面的解释为预期违背认知和规范维持动机(Hewig et al., 2011；张振等，2020)。个体层面上，一方面，个人希望与他人合作和得到公平对待的这些期望由群体规范来指导，另一方面，群体规范又以促进群体目标和价值观的方式协调个人的行为。群体层面上，群体规范的功能是保持群体凝聚力和促进群体利益，所以相对于群体间(Intergroup Interactions)的互动，这些规范在群体内(Intragroup Interactions)的互动中更加突出，并且更能激发个体拥护和维持其所属群体的核心规范的需求。正所谓“不以规矩，不成方圆”，群体规范是形成、运作和维持群体的必要条件，也是群体成员所认可、遵循并内化的行为准则，人们为了维持群体规范会对自私的内群体成员给予严厉的制裁从而保护规范的执行。

4.2. 纯粹偏好理论(Mere Preference Theory, MPT)

社会心理学家喜欢采用偏好来解释群体互动中的狭隘主义，认为人们对内群体成员有简单且强烈的亲社会偏好，诸如信任、合作、宽容等(Hogg, Abrams, & Brewer, 2017)。Tajfel 和 Turner (1979)提出的社会认同理论(Social Identity Theory, SIT)认为，当个体认识到他(或她)属于特定的社会群体，同时也认识到作为群体成员带给他的情感和价值意义时，会通过社会分类、社会比较、积极区分三种手段建立积极的社会认同(Social Identity)，并且以此提高自己的自尊，而这种自尊恰恰源自于内群体成员和相关的外群体成员的比较。当个体过分热衷于自己的群体，认为自己的群体比其他群体好，并在寻求积极的社会认同和自尊中体会到团体间差异时，就容易引起群体间偏见和群体间冲突，即对外群体偏见的产生是为了满足提高内群体地位和自尊的需要(张莹瑞&佐斌，2006)。

当群体划分形成并得到认同后，个体产生的“内群体依恋和积极性”(Brewer, 1999)使得人们更愿意忍受内群体成员的不公行为。这表明内群体成员的负性违规行为会被群体认同诱发的积极评价所抵消，进而有效降低被惩罚的可能性与强度。在有关公平的研究中，研究者发现当反应者分别与内、外群体成员完成最后通牒博弈实验时，人们对内群体成员的不公平行为接受率显著高于外群体成员，认为内群体积极评价增强了受测者的公平耐受性，从而抵消了由不公平分配所带来的消极感受和行为(王益文等，2014；张瀚月&赵玉芳，2018)。这也说明了群体认同的产生妨碍了公平规范的执行。

一方面，纯粹偏好理论认为内群体偏爱的形成与社会认同有关(张振等，2020)。内群体偏爱在群体组别这一单向维度的解释为个体为提升自身地位和自尊需要产生对违反公平准则的内群体成员的积极评价和包容，抵消了公平准则被打破的消极后果，产生了内群体偏爱的现象。另一方面，研究者还认识到个体对成员的不公平行为的潜在动机、目的的解释和归因也是出现内群体偏爱的重要原因，这种现象可由心理理论(Theory of mind, ToM)的基本假设得到解释(Perner, 1983)。在群际互动中，研究者发现个体试图去合理化对方的行为和决策，这方面的研究大多在公平决策的脑机制中得到证明，即，研究者发现在面对重要他人的不公平分配时，会激活与心理理论有关的内侧前额叶皮质(medial prefrontal cortex, mPFC)的及其与冲突加工的背侧前扣带回(dorsal anterior cingulate cortex, dACC)之间的功能连接，这在一定程度上反映了人们对不公行为的解释和合理化；mPFC 与 dACC 之间功能连接的增强也反映了人们借助合理化加工来解决认知冲突的过程(Fatfouta et al., 2018)。

4.3. 群体偏见与公平决策的理论整合

正如前文所述，个体在违背“经济人”假设后，对于公平决策的选择由动机和结果两方面入手来解释，即个体有对分配能够为自己带来利益的动机(不平等厌恶模型)，并与对方产生互惠的动机(互惠偏好模型)以及对于不公平厌恶的结果导向型解释。更进一步地，平等–互惠–竞争模型在互惠偏好和不平等厌恶的假设中引入“竞争因子”——与他人的相对比例，并很好地证实了 DG 范式中利他偏好的存在(Weiland et al., 2012)，驳斥了以自我为中心的利己主义理论。社会心理学家认为内群体偏爱本质上从单维的角度来解释群际互动的意义，即个体单纯的对内群体产生了共鸣，认为相较于外群体成员，内群体成员更加友善，因此愿意在资源分配中给予内群体成员以优待。但在真实的群际互动中，个体往往会权衡利弊，并且对互动对象产生期望(预期违背认知)，使得个体在公平决策行为上显得不那么“纯粹”。因此，个体的行为理论就由单维转向为多维的理论架构，在群体水平的维度上增设了利益大小的衡量指标和对于是否遵循社会规范的要求(规范维持动机)。

将个体对公平的感知在群际互动的背景下展开时，我们不难发现群际决策事实上与公平偏好的理论解释存在一定的割裂和交叉。一方面，在整合群体偏见和公平规范的理论时，我们会发现互惠偏好模型、预期违背认知理论、心理理论都强调心理效用的影响，如图 1 所示。预期对方给出的金钱数额(预期违背认知)和友善程度(互惠偏好模型)以及个体如何去解释对方的分配方案(心理理论)都很好地解释了个体的行为决策，这三种理论都将个体的主观认知纳入到理论体系中。然而，研究者在关注主观认知的同时，却部分忽略了客观上公平规范导致的不同行为结果，在群体偏见和公平规范的解释理论中，似乎难有理论共同将个体的心理认知(个体层面)和公平规范(群体层面)纳入其中，忽视了个体诸如社会价值取向、不平等厌恶等社会偏好因子(Everett, Faber, Crockett, & De Dreu, 2015)。详尽地说，在前文所述的 7 种模型和理论中，不平等厌恶模型、平等–互惠–竞争模型、社会认同理论和规范维持动机理论分别从不平等厌恶、个体的利他动机、群体态度、社会规范等社会价值取向和社会偏好因子等方面来解释个体在群体中的决策行为，而这些社会价值取向因子与心理效用的交互作用是怎样的却并未在理论中得到有效的阐释。

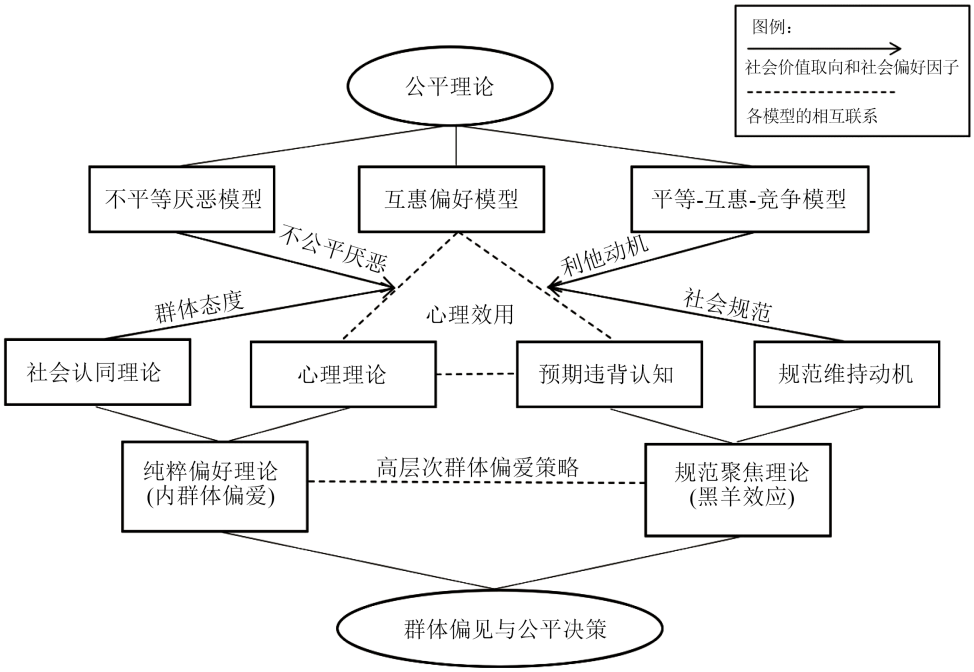

Figure 1. The theoretical integration of group bias and fair decision
图 1. 群体偏见与公平决策的理论整合

另一方面，学者们认为在群体偏见和公平决策的理论中，内群体偏爱和黑羊效应的出现实际上最终的目的是一致的(Mendoza et al., 2014)。研究者将黑羊效应的出现视为一种更高层次的维持和提升内群体偏爱的重要策略，即个体通过拒绝和惩罚内群体成员的不公行为，进而巩固群体价值和维持群体凝聚力，最终维持内群体偏爱。内群体偏爱和黑羊效应看似被割裂，但实际上都是个体聚焦群体利益作出的选择，进一步提高个体的群体认同(张振等，2020)。因此在今后的研究中，处于群际互动中的公平规范执行仍然需要不断整合多种理论框架。

5. 研究展望

回顾上述文献可知，近年来，经济学家、社会学家以及心理学家从未停止探索群际互动中的公平现象，对于公平理论和群体偏见产生的对公平规范执行的理论研究也取得了较为丰硕的成果，但在该领域，仍有一些问题在未来值得我们进一步探讨。例如，在不同群体之间，这种群体偏见产生的公平决策的行为结果是否有一个动态改变的过程？或者特定的向内支持亲社会行为如何在得失背景下表现出来呢？这都是一些悬而未决的问题，有待于在未来的研究中得到检验。

5.1. 拓宽群体偏见和公平规范研究的群体范围

目前多以成人被试为对象的研究，发现群际互动和公平偏好之间的张力很强，而在解释行为结果的理论上，会涉及到诸如偏好、利他、合作等成长性的变量和因素。从个体心理发展视角出发，个体在成长过程中随着社会文化的熏陶和群际互动经验的不断累积，对群体和公平的偏向也会逐步发生偏移，但这种偏移和转化是何时发生和如何发展的还需要进一步得到研究者的探索。例如，新近研究发现 6~8 岁儿童就已经产生了群体偏见，并且随着年龄的增长，对不利群体惩罚的强度增加，但是群体偏见却逐渐减少，这表明了儿童的规范执行从一开始就有偏见，但这种偏见可以通过发展变化得到部分抵消(Jordan, Mcauliffe, & Warneken, 2014)。另外(Mcauliffe & Dunham, 2017)发现，在第二方惩罚任务中，6~10 岁儿童

作为直接受害者，尽管产生群体偏见，但不会对分配金额和反应产生影响，更多的变现为偏爱公平规范的执行，而不是内群体偏爱。Wu 和 Gao (2017)通过最简群体范式操纵分组，以中国 2~6 岁的学龄前儿童为被试，发现 3~6 岁的中国儿童拒绝利己不公平分配的次数多于公平分配，表现出强烈的公平偏好，5~6 岁的男孩相比女孩更多的惩罚外群体成员。因此，这种个体成长中社会化的差异(性别、种族等)和群体偏见与公平规范的动态化演变在未来仍值得研究者进一步探索。

5.2. 增加得失框架下群体偏见与公平规范的研究

大多数关于群体偏见和公平决策的实验研究都着眼于参与者帮助内群体(相对于外群体)成员获得积极的东西。然而认知心理学指出，在社会互动中，人们更喜欢避免损失而不是获得收益(Kahneman & Tversky, 1979)。研究表明，因为造成损失被认为比保留收益更有害、更违反公平，所以个人更有可能帮助他人避免有害的结果或帮助他人提供积极的结果(De Dreu & Kret, 2015)。此外，在群体间的背景下，外群体贬损通常表现为缺乏帮助，而不是造成伤害(Weisel & Böhm, 2015)。梳理文献发现，少有研究考察过在特定的群体间背景下得失的影响，我们已经知道，公平在群体内的互动中比与群体间的互动中更为突出，加之损失厌恶的事实，不难推测，公平问题在损失背景下比在收益背景下更突出，所以人们可能在损失背景下表现出更大的内群体偏好。然而这还需要研究者进一步加以论证。

基金项目

全国教育科学“十三五”规划课题国家青年基金课题基金项目(CBA160188)；石河子大学人文社科中青年科研人才培育基金(KC0012)。

参考文献

韩小丽, 田孟奇, 田嘉乐, 苏文亮(2020). 最后通牒博弈的研究范式述评及其对结果的影响. *中国临床心理学杂志, 28(5),* 891-896.

李炜(2019). 社会公平感: 结构与变动趋势(2006-2017 年). *华中科技大学学报(社会科学版), 33(6),* 110-121.

李雪莹, 贾宁(2020). 经济博弈中公平感知及行为决策的影响因素的述评与展望. *心理技术与应用, 8(11),* 691-700.

唐俊(2011). 行为博弈的互惠利他行为理论研究的进展. *现代经济探讨, (6),* 41-44+83.

王益文, 张振, 张蔚, 黄亮, 郭丰波, 等(2014). 群体身份调节最后通牒博弈的公平关注. *心理学报, 46(12),* 1850-1859.

王珍珍, 蒋文明(2016). 公平加工的情境依赖性: 来自行为的证据. *心理与行为研究, 14(5),* 600-604+646.

徐富明, 李欧, 邓颖, 刘程浩, 史燕伟(2016). 行为经济学中的不平等规避. *心理科学进展, 24(10),* 1613-1622.

张瀚月, 赵玉芳(2018). 社会距离对不公平行为回应的影响. *西南大学学报(自然科学版), 40(2),* 140-145.

张莹瑞, 佐斌(2006). 社会认同理论及其发展. *心理科学进展, 14(3),* 475-480.

张振, 齐春辉, 王洋, 赵辉, 王小新, 等(2020). 内群体偏爱或黑羊效应? 经济博弈中公平规范执行的群体偏见. *心理科学进展, 28(2),* 329-339.

Abrams, D., Palmer, S. B., Rutland, A., Cameron, L., & Van de Vyver, J. (2014). Evaluations of and Reasoning about Normative and Deviant Ingroup and Outgroup Members: Development of the Black Sheep Effect. *Developmental Psychology, 50,* 258-270. https://doi.org/10.1037/a0032461

Adams, J. S. (1965). Inequity in Social Exchange. *Advances in Experimental Social Psychology, 2,* 267-299. https://doi.org/10.1016/S0065-2601(08)60108-2

Bašić, Z., Falk, A., & Kosse, F. (2020). The Development of Egalitarian Norm Enforcement in Childhood and Adolescence. *Journal of Economic Behavior & Organization, 179,* 667-680. https://doi.org/10.1016/j.jebo.2019.03.014

Biella, M., & Sacchi, S. (2018). Not Fair but Acceptable… for Us! Group Membership Influences the Tradeoff between Equality and Utility in a Third Party Ultimatum Game. *Journal of Experimental Social Psychology, 77,* 117-131. https://doi.org/10.1016/j.jesp.2018.04.007

Biernat, M., Vescio, T. K., & Billings, L. S. (1999). Black Sheep and Expectancy Violation: Integrating Two Models of So-

cial Judgment. *European Journal of Social Psychology, 29*, 523-542. https://doi.org/10.1002/(SICI)1099-0992(199906)29:4<523::AID-EJSP944>3.0.CO;2-J

Bolton, G. E., & Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review, 90*, 166-193. https://doi.org/10.1257/aer.90.1.166

Brañas-Garza, P., Cobo-Reyes, R., Espinosa, M. P., Jiménez, N., Kovářík, J. et al. (2009). Altruism and Social Integration. *Games and Economic Behavior, 69*, 249-257. https://doi.org/10.1016/j.geb.2009.10.014

Brewer, M. B. (1979). In-Group Bias in the Minimal Intergroup Situation: A Cognitive-Motivational Analysis. *Psychological Bulletin, 86*, 307-324. https://doi.org/10.1037/0033-2909.86.2.307

Brewer, M. B. (1999). The Psychology of Prejudice: Ingroup Love and Outgroup Hate? *Journal of Social Issues, 55*, 429-444. https://doi.org/10.1111/0022-4537.00126

Camerer, C. F. (2003). Strategizing in the Brain. *Science, 300*, 1673-1675. https://doi.org/10.1126/science.1086215

Cheng, X., Zheng, L. et al. (2015). Power to Punish Norm Violations Affects the Neural Processes of Fairness-Related Decision Making. *Frontiers in Behavioral Neuroscience, 9*, 344. https://doi.org/10.3389/fnbeh.2015.00344

De Dreu, C. K. W., & Kret, M. E. (2015). Oxytocin Conditions Intergroup Relations Through Upregulated In-Group Empathy, Cooperation, Conformity, and Defense. *Biological Psychiatry, 79*, 165-173. https://doi.org/10.1016/j.biopsych.2015.03.020

Dick, R. V., Haslam, A., Tyler, T. R., & Blader, S. L. (2001). Cooperation in Groups. Procedural Justice, Social Identity, and Behavioral Engagement. *Zeitschrift für Arbeits- und Organisations Psychologie A&O, 45*, 212-213. https://doi.org/10.1026//0932-4089.45.4.212

El Zein, M., Seikus, C., De-Wit, L., & Bahrami, B. (2019). Punishing the Individual or the Group for Norm Violation. *Wellcome Open Research, 4*, 139-139. https://doi.org/10.12688/wellcomeopenres.15474.1

Engel, C. (2011). Dictator Games: A Meta Study. *Experimental Economics, 14*, 583-610. https://doi.org/10.1007/s10683-011-9283-7

Everett, J. A. C., Faber, N. S., Crockett, M. J., & De Dreu, C. K. W. (2015). Economic Games and Social Neuroscience Methods Can help Elucidate the Psychology of Parochial Altruism. *Frontiers in Psychology, 6*, 861. https://doi.org/10.3389/fpsyg.2015.00861

Fatfouta, R., Meshi, D., Merkl, A., & Heekeren, H. R. (2018). Accepting Unfairness by a Significant Other Is Associated with Reduced Connectivity between Medial Prefrontal and Dorsal Anterior Cingulate Cortex. *Social Neuroscience, 13*, 61-73. https://doi.org/10.1080/17470919.2016.1252795

Fehr, E., & Fischbacher, U. (2004). Social Norms and Human Cooperation. *Trends in Cognitive Sciences, 8*, 185-190. https://doi.org/10.1016/j.tics.2004.02.007

Fehr, E., & Schmidt, K. M. (1999). A Theory of Fairness, Competition and Cooperation. *Quarterly Journal of Economics, 114*, 817-868. https://doi.org/10.1162/003355399556151

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong Reciprocity, Human Cooperation, and the Enforcement of Social Norms. *Human Nature, 13*, 1-25. https://doi.org/10.1007/s12110-002-1012-7

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in Simple Bargaining Experiments. *Games & Economic Behavior, 6*, 347-369. https://doi.org/10.1006/game.1994.1021

Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological Games and Sequential Rationality. *Games & Economic Behavior, 1*, 60-79. https://doi.org/10.1016/0899-8256(89)90005-5

Glimcher, P., Camerer, C., & Fehr, E. (2008). *Decision Making and the Brain*. London: Academic Press.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior & Organization, 3*, 367-388. https://doi.org/10.1016/0167-2681(82)90011-7

Hewig, J., Kretschmer, N., Trippe, R. H., Hecht, H., Coles, M. G. H. et al. (2011). Why Humans Deviate from Rational Choice. *Psychophysiology, 48*, 507-514. https://doi.org/10.1111/j.1469-8986.2010.01081.x

Hogg, M. A., Abrams, D., & Brewer, M. B. (2017). Social Identity: The Role of Self in Group Processes and Intergroup Relations. *Group Processes & Intergroup Relations, 20*, 570-581. https://doi.org/10.1177/1368430217690909

Jordan, J. J., Mcauliffe, K., & Warneken, F. (2014). Development of In-Group Favoritism in Children's Third-Party Punishment of Selfishness. *Proceedings of the National Academy of Sciences of the United States of America, 111*, 12710-12715. https://doi.org/10.1073/pnas.1402280111

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk Title. *Econometrica, 47*, 263-291. https://doi.org/10.2307/1914185

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the Assumptions of Economics. *Journal of Business, 59*, S285-S300. https://doi.org/10.1086/296367

Klaus, F., Phillipps, C. B., Peter, T., Marieke, S., Elger, C. E. et al. (2012). Neural Responses to Advantageous and Disadvantageous Inequity. *Frontiers in Human Neuroscience, 6,* 165. https://doi.org/10.3389/fnhum.2012.00165

Liu, Y., Bian, X., Hu, Y., Chen, Y.-T., Li, X. et al. (2018). Intergroup Bias Influences Third-Party Punishment and Compensation: In-Group Relationships Attenuate Altruistic Punishment. *Social Behavior and Personality, 46,* 1397-1408. https://doi.org/10.2224/sbp.7193

Marques, J. M., Yzerbyt, V. Y., & Leyens, J. P. (1988). The "Black Sheep Effect": Extremity of Judgments towards Ingroup Members as a Function of Group Identification. *European Journal of Social Psychology, 18,* 1-16. https://doi.org/10.1002/ejsp.2420180102

Mcauliffe, K., & Dunham, Y. (2016). Group Bias in Cooperative Norm Enforcement. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371,* Article ID: 20150073. https://doi.org/10.1098/rstb.2015.0073

Mcauliffe, K., & Dunham, Y. (2017). Fairness Overrides Group Bias in Children's Second-Party Punishment. *Journal of Experimental Psychology General, 146,* 485-494. https://doi.org/10.1037/xge0000244

Mcauliffe, K., Blake, P. R., Steinbeis, N., & Warneken, F. (2017). The Developmental Foundations of Human Fairness. *Nature Human Behaviour, 1,* Article No. 0042. https://doi.org/10.1038/s41562-016-0042

Mendoza, S. A., Lane, S. P., & Amodio, D. M. (2014). For Members Only: Ingroup Punishment of Fairness Norm Violations in the Ultimatum Game. *Social Psychological & Personality Science, 5,* 662-670. https://doi.org/10.1177/1948550614527115

Neumann, J. V., & Morgenstern, O. (1953). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Perner, W. J. (1983). Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception. *Cognition, 13,* 103-128. https://doi.org/10.1016/0010-0277(83)90004-5

Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *The American Economic Review, 83,* 1281-1302.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The Neural Basis of Economic Decision-Making in the Ultimatum Game. *Science, 300,* 1755-1758. https://doi.org/10.1126/science.1082976

Schiller, B., Baumgartner, T., & Knoch, D. (2014). Intergroup Bias in Third-Party Punishment Stems from Both Ingroup Favoritism and Outgroup Discrimination. *Evolution and Human Behavior, 35,* 169-175. https://doi.org/10.1016/j.evolhumbehav.2013.12.006

Sun, L., Tan, P., Cheng, Y., Chen, J., & Qu, C. (2015). The Effect of Altruistic Tendency on Fairness in Third-Party Punishment. *Frontiers in Psychology, 6,* 820. https://doi.org/10.3389/fpsyg.2015.00820

Tajfel, H., & Turner, J. (1979). An Integrative Theory of Intergroup Conflict. *Social Psychology of Intergroup Relations, 33,* 94-109.

Valenzuela, A., & Srivastava, J. (2012). Role of Information Asymmetry and Situational Salience in Reducing Intergroup Bias: The Case of Ultimatum Games. *Personality and Social Psychology Bulletin, 38,* 1671-1683. https://doi.org/10.1177/0146167212458327

Wang, Y., Zhang, Z., Bai, L., Lin, C., Osinsky, R. et al. (2017). Ingroup/Outgroup Membership Modulates Fairness Consideration: Neural Signatures from ERPs and EEG Oscillations. *Scientific Reports, 7,* Article No. 39827. https://doi.org/10.1038/srep39827

Weiland, S., Hewig, J., Hecht, H., Mussel, P., & Miltner, W. H. R. (2012). Neural Correlates of Fair Behavior in Interpersonal Bargaining. *Social Neuroscience, 7,* 537-551. https://doi.org/10.1080/17470919.2012.674056

Weisel, O., & Böhm, R. (2015). "Ingroup Love" and "Outgroup Hate" in Intergroup Conflict between Natural Groups. *Journal of Experimental Social Psychology, 60,* 110-120. https://doi.org/10.1016/j.jesp.2015.04.008

Wu, Y., Leliveld, M. C., & Zhou, X. (2011). Social Distance Modulates Recipient's Fairness Consideration in the Dictator Game: An ERP Study. *Biological Psychology, 88,* 253-262. https://doi.org/10.1016/j.biopsycho.2011.08.009

Wu, Z., & Gao, X. (2017). Preschoolers' Group Bias in Punishing Selfishness in the Ultimatum Game. *Journal of Experimental Child Psychology, 166,* 280-292. https://doi.org/10.1016/j.jecp.2017.08.015

Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y. et al. (2012). Rejection of Unfair Offers in the Ultimatum Game Is No Evidence of Strong Reciprocity. *Proceedings of the National Academy of Sciences of the United States of America, 109,* 20364-20368. https://doi.org/10.1073/pnas.1212126109