

载《经济研究》2011年08期：65-77页。

## 我们为何偏好公平：一个演化视角的解释\*

董志强

**内容提要：**当代行为经济学已经确认人类存在公平心理偏好，但并未回答它从何而来。本文提出一个演化解释：人类公平偏好的根源可能在于人类早期的进化过程。通过一个演化博弈模型和随机演化仿真模型表明：①在一个完全封闭的族群中，公平偏好的单态社会是随机稳定的演化均衡，不公平偏好的双态社会也是演化稳定的但并非随机稳定的；②若同时考虑族群之间的竞争，则公平偏好的单态社会将是唯一的演化稳定均衡。其原因在于，合作机会多少与合作利益大小是此消彼长的，公平行为最能够平衡合作机会与合作利益对生存竞争的影响，从而成为个体-族群两个层面的生存竞争中最具适存性优势的行为模式。人类的公平偏好，可能源于本能性的公平行为在人类早期进化中的适存性优势。这一认识有助于为行为经济学的公平偏好假设提供逻辑支持，也有助于从新的视角思考人们的非理性行为，以及探索经济理性的边界。当然，本文的逻辑仍有待古人类学、进化心理学和生物学等学科的研究来提供更充分的证据事实。

**关键词：**公平心理 演化 进化心理机制 多主体仿真 理性

### 一、引言

当代方兴未艾的行为经济学研究已经确认，个人的动机和行为常常与公平心理偏好有紧密的联系(Fehr and Schmidt, 1999; Fehr and Falk, 2002; 凯莫勒, 2006)。尽管“公平”一词有丰富而多样的含义,但在本文它主要指的是合作利益分配上的平等,推而广之也有行为的“对等性”(reciprocity)之意。公平偏好的存在,使人们不仅会努力追求公平,也会试图通过惩罚不公平行为来维护公平,尽管从事这些惩罚可能会让自己承担代价。比如 Gintis(2000)、Fehr and Rockenbach(2003)、Bowles and Gintis(2004)、Sánchez and Cuesta(2005)以及 Fehr and De Quervain et al(2004)均发现,为了维护合作的对等性,人们会积极惩罚不合作的个体。特别是 Fehr and De Quervain et al.(2004)发表在《科学》杂志上的研究,从决策的神经基础方面支持了利他惩罚。此外,公平偏好对宏观经济的影响也受到了重视,一些行为宏观经济学家认为公平偏好是影响经济波动的重要力量(Akerlof and Shiller, 2009);世界银行(2006)也在其发展报告中不但回顾了人类具有公平偏好的证据,更强调了公平对于发展至关重要。正因为公平偏好对经济行为有重要影响,最近十多年不少经济学家试图将公平偏好整合进经济理论,其中流传甚广的比如 Rabin(1993)的动机公平模型、Fehr and Schmidt(1999)的不平等规避模型、Bolton and Ockenfels(2000)的相对支付模型、Dufwenberg and

---

\* 董志强, 华南师范大学经济与管理学院, 华南市场经济研究中心, 邮政编码: 510006, 电子信箱: d\_zq@163.com。感谢国家社科基金项目(编号: 07CJL019)和广东省高校学科建设创新团队重大项目(编号: 051500020100925)研究资助。本文曾在重庆大学、华南师范大学、广东金融学院宣讲, 感谢与会者的有益评论。非常感谢匿名审稿人的良好意见和建议, 但文中的错失完全由作者负责。

Kirchsteiger(2004)扩展的公平均衡模型等；国内研究者亦有此方面的尝试，比如蒲勇健(2007)尝试将公平偏好引入了委托-代理模型。

上述研究工作中，实验研究证实了公平偏好的存在，理论研究则以公平偏好的存在为前提假设。它们很少询问和关注这样一个问题：人类的公平心理偏好是怎样产生的，或者说它从何而来？而这个问题的回答在理论上可能是很重要的。因为，一方面人们对世界的探索不仅希望“知其然”，也希望“知其所以然”，我们不会仅仅满足于了解公平偏好的存在，更希望了解它从何而来、为什么能够长期稳定存在。确实，对于当代行为经济学发现的事实和命题，已有经济学家试图探索其进化根源，比如 Dasgupta and Maskin(2005)和马斯金(2007)就从生存进化的视角解释了“双曲线贴现”存在的原因，叶航、汪丁丁和罗卫东(2005)也从演化视角对利他行为如何可以成为内生偏好进行了探索。另一方面，实验表明即使在不需要公平行为的时候(比如独裁博弈实验中)人们也会表现出一定程度的公平行为，而且这种看似不理性的行为在许多领域都相当普遍，以至于足以对宏观经济产生某种程度的影响(Akerlof and Shiller, 2009)。对这种“非理性行为”追根溯源，将有助于更好地理解人类的经济行为——实际上，许多看似非理性的行为在人类漫长的生存斗争史上可能曾经是“理性”的。

本文的主要目的在于，试图从演化的视角，对公平心理偏好的根源提供合乎逻辑的解释。我们将偏好视为演化力量的结果。具体地，我们假设人的行为并非来自理性计算，而是出于本能。本能乃是由基因携带和遗传下来的，携带着那些更具适存性(fitness)的本能的基因在进化过程中将更容易生存下来，从而某些行为模式和心理机制实际上是长期演化的结果。在本文，我们通过演化博弈论和演化仿真技术从逻辑上证明：公平心理偏好(以及该偏好驱动的公平行为)，在合作机会至为关键的人类漫长的生存进化斗争史上，的确是比非公平偏好或行为更具有适存性的一种本能。残酷的生存竞争和复杂的行为互动之背后，是简单的“物竞天择”之法则：具有公平偏好及行为之本能的个体，可以拥有尽可能最优的合作机会和合作利益之组合，因而可以有最大的适存性；具有公平行为模式的群体，比遵循不公平行为模式的群体有更大的合作期望赢利，因而可以在群体竞争中胜出。

应当承认，人类行为相当复杂，一方面受到本能的驱使，另一方面又的确会理性算计；既遗传有生物基因，又会从社会中学习(Flinn, 1997)，即遗传“文化基因”。很难说，人们的行为有多大比例来自本能，又有多大程度来自理性。但我们相信现代人都能够接受这一看法：人类在蒙昧时期的行为，多来自本能，而非理性。站在现代社会回望最初的人类社会，就如同回望我们的婴儿时代。初生婴儿的行为，完全来自本能而非理性。这在一定程度上可以作为本文采取本能行为假设的辩护意见。另一方面，如果公平偏好及行为乃人类长期进化中形成的本能，那么我们应该可以在一些缺乏人类算计理性的动物身上也观察到公平行为模式。的确，现代生物学研究发现了动物具有公平行为模式的事实(Trivers, 1983)，对猴子的研究发现猴子也有公平偏好(Markey, 2003; Brosnan and De Waal, 2003)。这为本文采取本能行为假设提供了一定程度的支持。对于人类，Burnham(2007)一项很有意思的研究也发现了公平偏好的生物学证据：在最后通牒博弈中，睾酮素较高的男性比睾酮素较低的男性更倾向拒绝不平等的提议。就睾酮素与身体状况和攻击性相联系这一点来说，这可能提供了一个基因联系，表明公平偏好确实有人类本能的成份。不过，本文

采取本能行为假设并不仅仅因为考虑到人类早期行为演化模式，也是因为考虑到生物学上的生存优势先于文化过程这一事实；即先有“自然为人类立法”，而后才有“人类为社会立法”（汪丁丁等，2006；董志强，2008）。文化过程本身亦应是生存竞争演化的产物。

与本文有一定关系的文献也许还包括 Seabright(2006)、Gintis(2006)和 Young(1998,1996)。Seabright 和 Gintis 基于 Binmore(2005)的自然正义讨论过公平规范的演化<sup>①</sup>；Young 基于随机演化博弈讨论了地主和农民之间五五分成的公平规则何以会成为最可能的分配惯例。但本文有所不同，要考察的是人类在其生存进化的斗争中，如何“无意识”地进化出了公平偏好，这中间没有道德的评判，也没有面向他人的学习，公平偏好将是最大化生存能力的进化结果，而不是道德的或者策略性的选择。或者说，本文要强调的是，自然选择的压力如何迫使人进化出了“公平”这种有利于合作的偏好。

本文接下来安排如下：第二部分将建立一个基本的演化(博弈)模型，基于简单的假设情形，讨论在封闭的原始族群中公平偏好的单态社会如何成为随机稳定的演化均衡，以及考虑族群竞争下公平偏好的单态社会何以成为唯一的演化稳定均衡；第三部分则建立随机演化的多主体仿真模型(multi-agent based simulation model)，在更为复杂的情形下验证第二部分的结论。最后是全文的总结，同时简要讨论作为演化产物的公平偏好对于经济学的意义。

## 二、基本模型及分析

公平偏好演化起源的建模，首先遭遇的一个难题就是模型环境的设置。迄今我们并没有任何证据表明人类的公平偏好产生于何时。但可以推测的是，如果公平偏好是人类早期演化而来的本能，那一定是在人类早期的合作过程中逐渐演化而成的。倘若没有合作，就很难谈得到上利益分配上的公平。但人类合作的起源和演化本身尚有许多未解之谜(汪丁丁等, 2006; Bowles and Gintis, 2003)，本文也只好回避困难，假设人类已经具有一定的相互合作的能力，并且事实上已经通过合作来争取更大的生存机会了。

从人类进化的历史来看，即便以 180 万年前直立人从非洲向亚洲扩散(第一次大迁徙)算起，到 8000-10000 年前农业革命（第一次社会大分工），人类处于茹毛饮血的狩猎时代的时间超过了人类迄今为止 99.9%的时间。人类进化而来的心理机制，与狩猎社会密不可分。演化心理学家认为，大型狩猎活动通常需要狩猎者之间的合作与交流，从而成为人类进化的主要推动力，而且还衍生出诸多其他结果：制造工具、使用工具、语言、脑容量增加(巴斯, 2007)。

因此，我们的演化博弈模型和随后的仿真模型将设定在人类的狩猎社会时代。这个时代人们以狩猎为生，族群已经形成但规模并不太大，合作对于生存至关重要，没有集中的治理结构(国家、司法体系或者权威等)，人们的地位无甚差异，也尚未发展出公平分配观念——族群成员根本不知道“公平”二字，更不懂得道德和分配正义之类的哲学。

---

<sup>①</sup> Binmore(2005)宣称演化压力促成了正义、公平等道德规范。Seabright(2006)认为在 Binmore 的自虑(self-regarding)人假设下难以解释道德直觉如何得以演化，因此道德分析须考虑人的他虑(other-regarding)偏好。Gintis(2005,2009)则认为，由无名氏定理描述的多人重复博弈均衡缺乏动态稳定性，故 Binmore 以无名氏定理为分析基础得到的正义理论是有失误的；道德法则并不能解决纳什议价问题，这些问题的解决需要求助于人类的亲社会性和利他观念等演化而来的人性。本文的立场更接近于 Gintis，因为本文试图证明公平偏好乃演化而来的人性之一。

我们考虑两种层面的生存竞争：1)单个族群内部的生存竞争；2)族群之间的生存竞争。在历史上，这两个不同层面的生存竞争都是时刻存在的，而且是同时演化的。出于分析技术上的可行性，我们将其分成两个不同的层面分别分析。

### 1. 单个族群内部的生存竞争和公平偏好演化

考虑一个原始族群，成员以狩猎为生。在每个时期  $t$ ，任何成员单独行动将不能获得猎物(一个人杀不了猛犸象!)，任意两个成员合作则一定可以捕得 1 单位猎物。对猎物消费的多寡决定了成员繁殖后代的能力(以此作为个体适存性的度量)。每个时期  $t$ ，族群的成员在领地内随机游走，当任何两个成员相遇，他们可以决定是否合作。当然，达成合作的条件是：这两个成员要求分割猎物的比例之和不能超过 1。由于族群社会尚未发展出公平分配观念，任何成员都只奉行一种简单的谈判方式：每人报告一个希望分割猎物的份额(以下称“要价”)，若二人要价之和超过 1，则不能达成合作，每人的赢利(payoffs)为 0；若二人要价之和不超过 1，则每人得到其要价(若二人要价之和不足 1，则假设未分配的剩余被浪费掉<sup>①</sup>)。这里，假设要价是受成员的基因控制(即出自一种本能)，因而与观念无关。每个成员的要价都是基因或本能层面的无意识选择。设全部成员共有  $I \equiv \{1, 2, \dots, I\}$  种基因，对应于要价集合  $\alpha \equiv \{\alpha_1, \dots, \alpha_I : 0 < \alpha_i < 1, \alpha_i \neq \alpha_j, i \in I, j \in I\}$ ；则  $\alpha_i \in \alpha$  刻画了  $i$  型基因成员的贪婪程度；每种贪婪性为  $i \in I$  的成员在时期  $t$  占总人口比例记为  $s_i$ 。假设每个成员只存活一期，每个成员留下的后代数量是由各自得到猎物的份额所决定的，且后代数量随猎物份额单调递增。那么，考虑每期时间无穷小，则可以得到要价为  $\alpha_i$  的贪婪本性为  $i$  的个体在时间维度  $t$  上的变化模式：

$$\dot{s}_i = s_i \left[ \alpha_i \sum_{j \in A_i} s_j - \sum_{j \in I} s_j \left( \alpha_j \sum_{k \in A_j} s_k \right) \right] \quad i \in I, j \in I, k \in I \quad (1)$$

式(1)被称为复制动态方程，它根据生物学中将赢利(payoffs)作为适存性的一种度量，且考虑个体的(单性繁殖)后代数量由个体的赢利决定(满足单调递增性)，而得到这个方程；复制动态方程的具体推导可参阅(VeGa-Redondo, 2003)。方括号中， $\alpha_i \sum_{j \in A_i} s_j$  是  $i$  型个体的预期赢利，其中  $A_i$  是  $i$  型个体可遭遇的潜在合作对象的集合， $A_i \equiv \{j : \alpha_j \leq 1 - \alpha_i, j \in I\}$ ，即所有要价不超过  $1 - \alpha_i$  的  $j$  型个体都可以与  $i$  型中的某一个体合作，注意某些时候允许  $j=i$ ，譬如： $\alpha_i \leq 0.5$  的时候， $i$  型个体之间也是可以达成合作的。 $\sum_{j \in I} s_j (\alpha_j \sum_{k \in A_j} s_k)$  是族群全体成员(包括了各种贪婪程度的个体)的平均赢利，其中  $A_j \equiv \{k : \alpha_k \leq 1 - \alpha_j, k \in I\}$ 。

由于  $i \in I$  是任意的，所以式(1)实际上是  $I$  个联立的微分方程组；加上其函数形式比较复杂，求解这样的方程组几乎是不可能完成的任务。但是，我们可假设相对简单而有一定代表性的情形

<sup>①</sup> 当然也可这样假设，如果甲要价  $\alpha \in [0, 1]$ ，乙要价  $\beta \in [0, 1]$ ，在  $(\alpha + \beta) \in (0, 1)$  时按照比例  $\alpha/(\alpha + \beta)$  和  $\beta/(\alpha + \beta)$  进行分配。这样的分配方式，相当于假设甲乙两人若未能索要全部猎物( $(\alpha + \beta) < 1$ )时，则他们就剩下的部分又按照自己要价的比例进行分配……由于  $(\alpha + \beta) < 1$  所以剩余始终存在，但是经过无穷次就剩余进行分配后，他们各自得到的猎物份额极限正好是  $\alpha/(\alpha + \beta)$  和  $\beta/(\alpha + \beta)$ 。由于考虑了演化，任何未分配剩余  $1 - \alpha - \beta$  在长期演化中一定不会存在，即演化稳定均衡的时候，将有  $\alpha + \beta = 1$ ，从这个角度看假设将未分配剩余浪费掉也并无不妥。浪费假设并不会影响模型的结果，但对分析却更简单，故本文做此假设。

来加以分析。具体地，考虑成员的贪婪程度类型只有三种： $\alpha_1=x<0.5$ ， $\alpha_2=0.5$ ， $\alpha_3=1-x$ （这里请考虑前文脚注提及二人要价和不为1所产生的剩余在长期演化中不会存在，才能理解到假设 $x$ 和 $1-x$ 的合理性）。这三种类型分别代表了“吃亏”的基因(从来只索要一小半猎物)、绝对“公平”的基因(总是索要刚好一半猎物)，以及“贪婪”的基因(永远索要大部分猎物)。则个体的要价情形可表示如下：

$$A = \begin{bmatrix} x & x & x \\ 0.5 & 0.5 & 0 \\ 1-x & 0 & 0 \end{bmatrix}, \quad x < 0.5 \quad (2)$$

考虑到这里的群体状态属于二维单形(即 $s_1+s_2+s_3=1$ )，因此式(1)的复制动态方程(组)可由两个频率刻画，我们选择 $s_1$ 和 $s_2$ 来刻画：

$$\begin{cases} \dot{s}_1 = s_1 \left\{ x - [s_1x + 0.5s_2(s_1 + s_2) + (1-s_1-s_2)(1-x)s_1] \right\} \\ \dot{s}_2 = s_2 \left\{ 0.5(s_1 + s_2) - [s_1x + 0.5s_2(s_1 + s_2) + (1-s_1-s_2)(1-x)s_1] \right\} \end{cases} \quad (3)$$

令 $\dot{s}_1 = 0 = \dot{s}_2$ 可得到5个均衡点<sup>①</sup>： $(0,0)$ ， $(0,1)$ ， $(1,0)$ ， $(\frac{x}{1-x}, 0)$ ， $(\frac{x}{1-x}, \frac{(1-2x)x}{1-x})$ 。

对于这些均衡点的演化稳定性，我们可以利用ESS(演化稳定策略)概念来检验。随机匹配下的单种群演化博弈中，一个ESS应具有这样的性质：对于原来的单态群体来说，一小部分采取不同策略的另一类型个体(“突变”)是否会造成永久性的扰乱(“入侵”)(Smith and Price, 1973; Smith, 1982)。而一个演化稳定策略组合，在复制动态中将是渐进稳定的(Hofbauer and Schuster et al., 1979)。或者简单地说，ESS能够对微小的扰动自我校正(Bowles, 2004)。

先考虑点 $(\frac{x}{1-x}, 0)$ ，这个点上要价类型向量 $(x, 0.5, 1-x)$ 对应的频率向量为 $(\frac{x}{1-x}, 0, \frac{1-2x}{1-x})$ ，群体的平均赢利为：

$$\bar{\alpha} = x \left( \frac{x}{1-x} \right) + \left( \frac{1-2x}{1-x} \right) (1-x) \left( \frac{x}{1-x} \right) = x \quad (4)$$

对于0频率的公平型( $\alpha_2=0.5$ )基因个体增加一个非常小的频率 $\epsilon>0$ ，则该类型的预期赢利为： $\bar{\alpha}_2=0.5(\frac{x}{1-x}+\epsilon)$ ，只要 $\epsilon < \frac{(1-2x)x}{1-x}$  (由于 $x<0.5$ 故这样的 $\epsilon>0$ 是一定存在的)，则有 $\bar{\alpha}_2 < \bar{\alpha}$ ，即公平型个体的平均适存性低于种群平均适存性，小规模公平基因“入侵”并不能伤害种群的状态分布。同样， $\alpha_1=x$ 吃亏型基因个体增加 $\epsilon>0$ 的频率其预期赢利仍是 $x$ ，并不能提高其适存性； $\alpha_3=1-x$ 贪婪型基因个体增加 $\epsilon>0$ 的频率(则吃亏型基因频率将下降)，其个体预期赢利为 $\bar{\alpha}_3=(1-x)(\frac{x}{1-x}-\epsilon)<x$ ，即其平均适存性反而低于种群平均适存性。总结起来，我们可宣称均衡点 $(\frac{x}{1-x}, 0)$ 是演化稳定的。

再考虑点 $(\frac{x}{1-x}, \frac{(1-2x)x}{1-x})$ ，可以计算此时种群平均赢利仍为 $x$ 。此时，若 $\alpha_2=0.5$ 公平型个体频率增加 $\epsilon>0$ ，其预期赢利变化为：

$$\bar{\alpha}_2 = 0.5 \left( \frac{(1-2x)x}{1-x} + \frac{x}{1-x} + \epsilon \right) = x + 0.5\epsilon > \bar{\alpha} = x \quad (5)$$

即公平型个体频率略微增加也会更大地提高其适存性。故均衡点 $(\frac{x}{1-x}, \frac{(1-2x)x}{1-x})$ 不是演化稳定的。根据同样的方法，我们还可检验出 $(0,0)$ 和 $(1,0)$ 两个均衡点也不是演化稳定的，而 $(0,1)$ 点则是

<sup>①</sup> 本来是6个解，其中一对解是重根解，即 $\{s_1=0, s_2=0\}$ 和 $\{s_1=0, s_2=0\}$ 。

演化稳定的。总结上述全部分析，我们有：

**命题 1**：式(2)要价情形的单一原始族群博弈中，存在两个演化稳定的均衡点：

I. 均衡点(0,1)，即全部成员均为公平型成员，要价类型均为  $\alpha_2=0.5$ ；

II. 均衡点( $\frac{x}{1-x}, 0$ )，即社会没有完全公平型的成员，不会有要价类型为  $\alpha_2=0.5$  的成员；有  $\frac{x}{1-x}$  比例的个体为吃亏型成员，他们的要价类型  $\alpha_1=x<0.5$ ；有  $\frac{1-2x}{1-x}$  比例的个体为贪婪型成员，他们的要价类型为  $\alpha_3=1-x>0.5$ 。

命题 1 的结论，在技术层面上也可通过做出相位图来发现。图 1 中(a)、(b)、(c)是设定不同  $x$  值用 Mathematica 软件绘出的方向场。

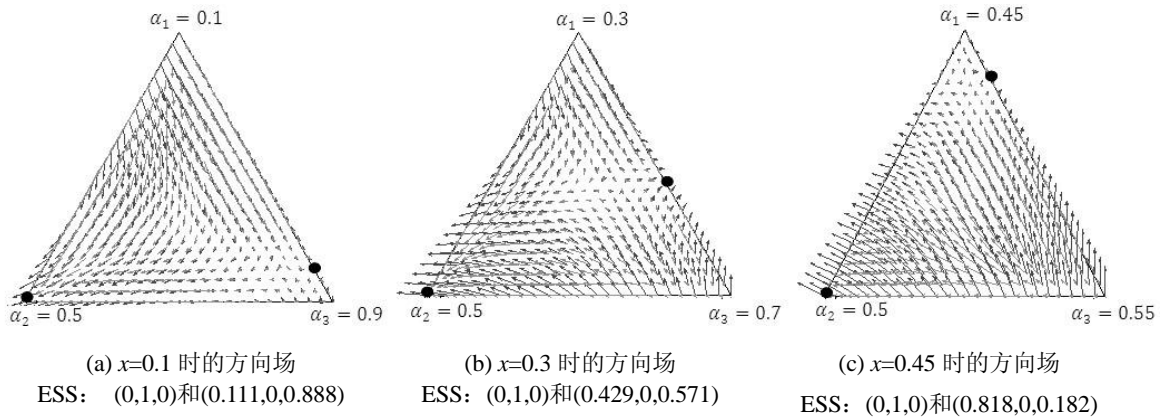


图 1 各类要价偏好演化的方向场(黑点为 ESS 均衡点)

从方向场可以清楚地看见，演化最后会收敛到两个演化稳定的均衡点。其中一个稳定均衡是单态的，即所有的成员最后都是公平型的（各图的左下角要价  $\alpha_2=0.5$  那个顶点）。另一个均衡是双态的，吃亏型与贪婪型成员并存，没有公平型成员（即要价  $\alpha_1=x$  和  $\alpha_3=1-x$  两顶点连线上的均衡点）。

方向场还提供了更多的信息：无论  $x \in (0, 0.5)$  在其定义域内取何值，在方向场的三角空间内绝大多数点出发将到达的稳定均衡都是单态均衡；可以到达双态均衡的出发点是少数；即博弈有更大的概率演化到单态均衡上。但单态均衡和双态均衡出现的概率与  $x$  有关，当  $x$  越靠近 0(图 1a) 或越靠近 0.5(图 1c)，就越可能演化到单态均衡。事实上，可近似计算出，仅有三种要价类型的博弈中，演化到双态均衡的概率约为  $\frac{(1-2x)x}{1-x}$ ，演化到单态均衡的概率约为  $\frac{1-2x+2x^2}{1-x}$ ；出现双态均衡的最大概率约为 0.17，在  $x \approx 0.29$  时取得。若主体类型更多，则双态均衡出现的最大概率会变得更小；随后的仿真实验也可证实这一点。

只要演化过程中存在突变，则演化就存在扰动过程；扰动趋于 0 时以较大频率被访问的状态即随机稳定状态(Foster and Young, 1990)。在这个意义上，我们也可以说，在一个封闭的族群中，只要存在基因突变，则公平偏好的单态社会将是随机稳定的。

## 2. 族群之间的竞争与公平偏好演化

单一的封闭族群内部的生存竞争演化过程中，公平偏好的单态社会和非公平偏好的双态社会都可以是演化稳定的；只不过公平偏好的单态社会出现的概率更大，因而是随机稳定的。然而，

一旦允许族群之间的竞争，则唯有公平偏好的单态社会才是演化稳定的。原因在于，公平偏好的单态社会比任何不公平的双态社会都具有更高的个体期望赢利，因而在生存竞争中前者也必然比后者更具有适存性。要证明这一结论很简单：假如一个族群形成 $(\alpha_1, \alpha_3)=(x, 1-x)$ 的双态社会，则族群中个体的期望赢利将是  $x < 0.5$ ，这可由式(4)得到；而  $\alpha_2=0.5$  的单态族群社会中，个体的期望赢利是 0.5。

这里有必要深入讨论考虑族群竞争的正当性。对于现代人来说，考虑族群竞争似乎理所当然，因为族群之间竞争所造成的生存压力，增进了族群内部成员之间寻求公平以促进合作的激励。不过，本文的分析乃是将公平行为视为本能层面的无意识选择来展开的，因此不允许有任何的策略性考量，在这样的情况下，引入族群竞争的正当性只能来自于族群生存竞争的事实。人类远古时代的历史，我们现在只能根据少量的证据来做大量的猜想，无论如何，族群的生存压力和族群间竞争一定是存在的。气候变化、环境恶化、瘟疫、饥荒在人类这个物种进化期间是经常的，乃至还有战争——乔伊斯·马库斯(2002)引用了 Jebel Sahaba 暴力事件遗址<sup>①</sup>后写道：“这是一个给予我们群体间竞争印象的事实”——这些都构成了自然选择的压力，它们不但筛选了个体，更筛选了群体。只要族群灭绝或自然选择的压力足够大，那么能幸存下来的只能是更具适存性优势的族群。

由于族群竞争本质上是群体选择理论的应用，因而有必要再做一点说明。由于 Williams(1966)和 Dawkins(1976,1982)的对群体选择的批判，群体选择曾一度被生物学、经济学等学科所抛弃，理由是有利于群体但不利于个体的策略不可能在个体和群体层面均演化稳定。至今关于群体选择尚存不少争议，但总的看来群体选择在分析人类社会行为方面有卷土重来之势(Bergstrom,2002)。而且，群体选择与个人主义方法论可能并不冲突(黄凯南,2008)。对于本文来说，更重要的是，无论生物学家还是经济学家几乎都已经承认，在单个群体中出现的多重均衡，群体选择将会淘汰掉那些更没有效率的均衡(Robson,2008)，这正是本文理论模型中的情况。

限于分析技术，我们这里把演化分解成了个体和族群两个层面的竞争，而对族群竞争中个体的适存性是根据个体在族群内部处于演化稳定均衡点时的预期赢利来度量的；这也是大多数群体选择理论的推理思路(可参考 Robson,2008)。但现实的演化过程中，族群内部个体之间的生存竞争和族群之间的生存竞争是同时发生的，这与本文分析所采取的度量标准有一定差异；不过随后的仿真模型中我们允许个体和群体同时演化，结果表明上述关于族群间竞争的结论仍然成立。于是我们有：

**命题 2：**考虑族群之间的生存竞争，则具有公平偏好的族群比不公平偏好的族群将更具适存性。

命题 2 实际上意味着，一旦考虑族群竞争，则命题 1 中作为演化稳定均衡出现的非公平偏好的双态社会将不再是演化稳定的，只有公平偏好的单态社会才会仍然是演化稳定的。

### 三、仿真模型及结果

---

<sup>①</sup>Jebel Sahaba 遗址是目前所知的最古老的群体暴力证据。该遗址碳测年代测定在 1.5 万年前（若加以校正则更古老），由 58 座墓葬组成，死者有男人、妇女和儿童，有一半死于系列的伏击暴力。有些人被扎了 15 到 30 枚箭镞。Jebel Sahaba 位于一处有许多狩猎采集族群居住的地方，这些觅食者显然是为了争夺沿尼罗河港湾中食物资源而死，并有各自的墓地。

为避免分析上的困难，上述基本模型中，只考虑了三种要价类型(基因)；族群内个体的竞争和族群之间的竞争分解为两个单独的层面进行。但更现实的情况应该是有多种多样的要价类型(基因)，并且族群内个体的竞争和族群之间的竞争是同时发生的。为了研究这更符合现实的情况，我们设计了公平偏好演化的基于行为主体的仿真模型。

我们采用的仿真平台是美国西北大学开发的 NetLogo。共设计了两个模型<sup>①</sup>，一个模型是只考虑单一族群内部的公平偏好演化；另一个模型则在第一个模型基础上进一步允许有多个族群(可由实验者设置)展开竞争，即个体在族群内部进行生存竞争，但同时族群之间也在进行生存竞争。

### 1. 单一族群内部公平偏好演化的仿真模型

在模型中，实验者可为这个原始族群确定一个初始人口数量（范围在 10-400，因为原始族群规模都比较小，而且过于大的主体数量会使计算机运行非常缓慢）并对系统进行初始化。系统在初始化中将为族群中每个成员随机地指派一种合作要价类型，每个成员一旦被指定类型则其类型不会发生改变，所以一种合作要价类型也可以看作是不同吃亏或贪婪程度的基因类型。系统设定的类型一共有 19 种，为要价集合{0.05,0.10,0.15,0.20,...,0.90,0.95}。系统初始化之后，可启动仿真过程，在此过程中，每个成员按照如下的方式开始相互博弈：

(1)在每一给定时期，每个成员在以自身为中心的一定半径(可设置和调整)的领域内随机游走，在可到达的地理范围内随机挑选一个没有被邀请过的成员发出合作邀请，并根据自身的类型给出对应的要价。

(2)被邀请的成员根据其自身类型回应相应的要价，若两个成员的要价之和不超过 1，则他们达成合作。若两个成员的要价之和超过 1，则不能达成合作。主动发出邀请而未能达成合作的成员还可以等待其他未被邀请过的成员邀请 1 次，并根据各自要价决定合作能否达成，如果他周围有未被其他成员邀请的成员来邀请他的话；倘若他超出了任何未被邀请过的成员之视野范围，那么他在这个时期就无法找到合作对象了。

(3)能量(即赢利)获取。每个与他人达成合作的成员得到等于其要价  $\alpha_i$  的能量(若两者要价之和小于 1，则分配后剩余的能量将浪费消失掉)，每个未能与他人合作的成员得到的能量为 0。

(4)每个成员生殖后代，生殖一个后代需要耗费一定能量  $e$ （可设置和调整），生殖后代的数量为  $\text{int}[\alpha_i/e-1]$ 。这意味着拥有能量  $\alpha_i < e$  的成员将无法繁殖后代。

(5)每个成员以  $1-\delta$  的概率将自己的类型遗传给后代，而且后代以  $\delta/18$  的概率继承其他 18 种类型中的一种。这里  $\delta$  即基因突变率，可由实验者设置。

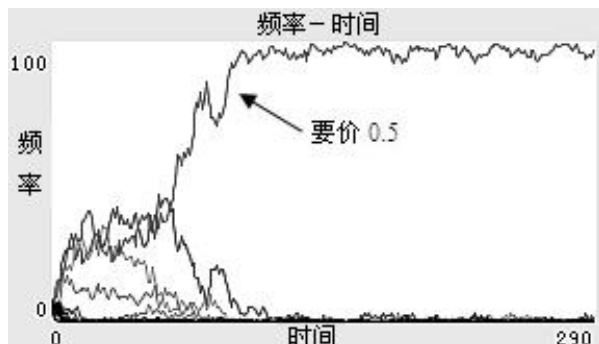
(6)每个在这一期参与博弈的父辈成员死去，博弈下一期返回(1)由后代成员作为参与人进行博弈。

为了防止成员增长过多，实验者可控制最大成员数量，每期随机地杀死一些成员使得平均程度上成员数量限制在最大数量内。当然，这也可视为原始族群的高死亡率。

单一族群内部公平演化的仿真模型实验具有如下程式化(stylized)结果，确认了前面数理模型的分析：(1)在绝大多数实验中，公平行为很快会成为族群中主导行为模式，形成公平偏好的单态社会。图 2 是一种典型的实验状况，从启动实验只需要 agent 演化几十代就形成了公平偏好社会。

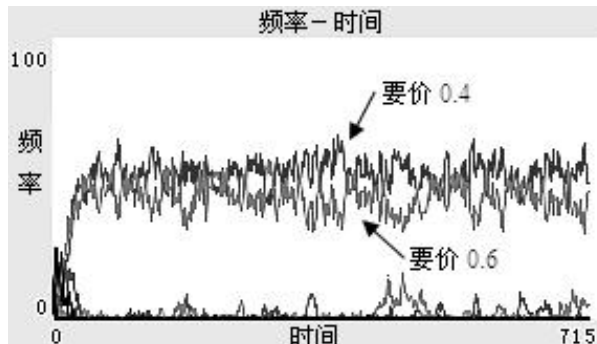
<sup>①</sup> 对仿真程序感兴趣的读者可向作者索取源程序。

(2)在少数实验中,亦出现了不公平的双态社会。在我们进行的数百次实验中,0.3/0.7和0.4/0.6的非公平社会,以及0.45/0.55这种更接近公平的公平偏好社会都出现过,但出现的次数非常少。图3是某一次实验演化了700代的0.4/0.6的非公平双态社会,图4则是某次实验演化出了0.45/0.55这种接近公平的公平双态社会。



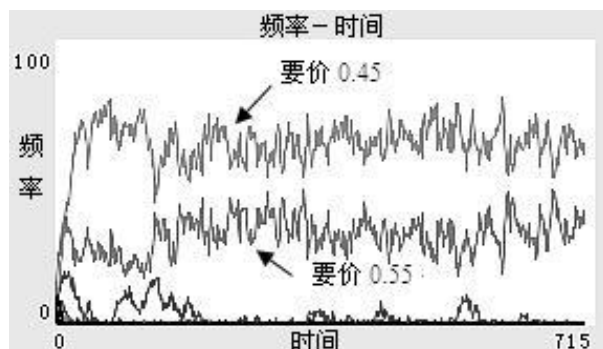
各种要价偏好在演化初期竞争,到第70代,要价0.5的公平偏好个体在族群中比例开始占绝对优势,并很快达到了公平偏好单态社会的稳态均衡。

图2 公平偏好单态社会典型实验结果



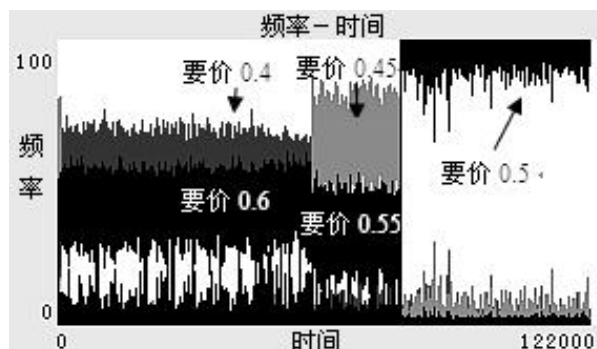
各种要价偏好在演化初期竞争,大约在第15代,要价0.4和0.6的非公平偏好个体开始胜出,大约在50代以后,0.4/0.6的非公平偏好双态社会完全形成。

图3 非公平偏好双态(0.4/0.6)社会典型实验结果



各种要价偏好在演化初期竞争,大约在第40代,要价0.45和0.55的非公平偏好个体开始胜出,大约在190代以后,0.45/0.55的非公平偏好双态社会完全形成,并成为稳态均衡。

图4 非公平偏好双态(0.45/0.55)社会典型实验结果



初期的演化形成了0.4/0.6的非公平偏好社会。但是偏好“基因”突变一直存在。大约在57300到57400代,0.45/0.55偏好入侵成功而成为均衡。大约在77445到77585代,0.5偏好入侵成功形成公平偏好单态社会。

图5 公平偏好入侵非公平偏好社会的典型实验结果

同时实验还表明,即使一个族群已经形成了不公平的双态社会,只要基因存在突变的可能(尽管可能性很小),而演化的时间足够长,公平偏好也可以成功入侵非公平偏好的稳态社会。公平偏好的单态社会的确是随机稳定的。图5反映的就是公平偏好成功入侵非公平偏好稳态社会的情况。

以上实验的基本参数设置均为族群人口在200左右(人类原始族群规模都不大)、偏好“基因”突变的概率 $\delta=0.00005$ (一种偏好基因突变出其他特定类型偏好的概率则为 $\delta/18=0.0000028$ ,不到十万分之三),繁殖一个后代需要能量为0.05。我们也曾改变各个参数进行实验,结果表明前面提及的程式化结果具有相当的稳健性。

## 2. 基于个体-族群生存竞争的公平偏好演化仿真模型

考虑成员个体-族群同时演化的模型:成员在族群内部展开生存竞争;族群之间也存在生存竞争。族群之间的竞争当然可以有多种冲突形式,模型采取了一种最为直接的形式,即战争。族群的适存性取决于其成员的适存性(以获得的能量刻画),对于两个交战的族群,其胜负概率以各族

群成员的适存性总和之对比来衡量。比如：族群 1 中成员能量总和为  $E_1$ ，族群 2 中成员的能量总和为  $E_2$ ，则族群 1 获胜的概率为  $E_1/(E_1 + E_2)$ 。此外，战争在每个时期并非必然发生，而是以一定的概率发生的，即战争是偶然的。具体地，当前的模型在前述单一族群模型上做了如下扩展：

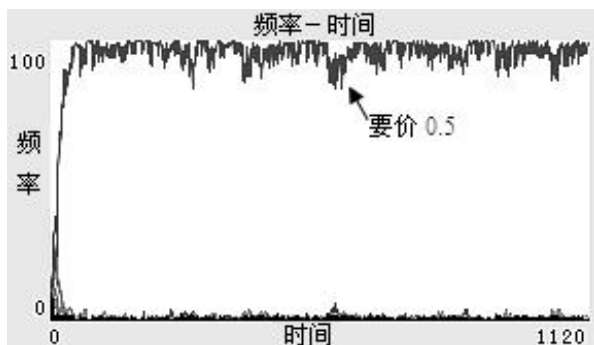
(1)初始化时，所有个体被指派到各个族群(族群数量可设置或调控)，在每一期每个个体出生时都会贴上某一个族群的标签  $g=1,2,\dots,G$ 。

(2)每一时期，每个族群内部的成员按照前述的单一族群内相互博弈的方式进行博弈，并获得相应能量和繁殖后代并死亡；

(3)每一时期，每一个族群都以一定的概率(可设置和调整)，对除自己之外的任意族群发动战争，并根据两个族群内部所有个体的当前能量总和(个体只有一期的寿命)之对比决定胜负概率。被征服的族群内个体全部死亡，胜利的族群则将自己复制一份，新产生的这个族群标记以原失败族群的标签(占据原失败群体的生存空间)，并成为完全独立的群体参与下一期的竞争。然后博弈重复进行。

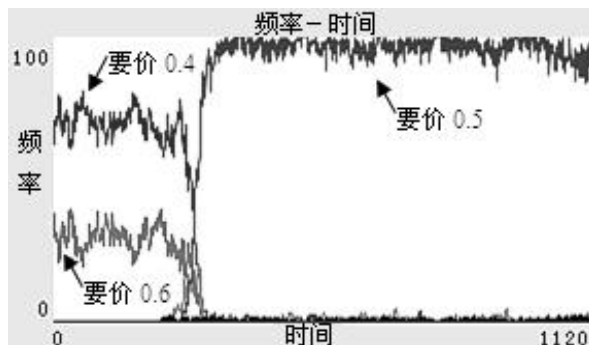
而为了保持群体间的演化竞争压力，对于因合作效率低下而人口逐渐减少直至被自然淘汰的族群，在被淘汰之后将被一个新的族群所取代，新族群初始人数设为模型初始设置中各族群人数之期望值，新族群成员的初始类型则是随机分配的。

实验结果表明，存在族群竞争压力时，公平偏好可以更快地得到演化，而且迄今为止的上百次实验中，无一例外是演化出公平偏好单态社会的族群得以生存，而未能演化出公平偏好单态社会的族群在竞争中被淘汰。图 6 是一次典型的实验结果。此外，为了检验初始状态为非公平偏好社会时，公平偏好社会能否演化出来并在竞争中胜出，我们刻意在初始状态设置所有族群的个体都是 0.4-0.6 之一种，并且不允许偏好“基因”突变以强制演化出 0.4/0.6 的非公平双态社会，然后打开偏好“基因”突变开关，允许突变。结果表明，只要允许突变，无论初始偏好状态如何，最后胜出的都是公平偏好；我们用 0.45/0.55、0.3/0.7、0.35/0.65 等进行实验结果也是一样。



个体在族群中竞争，族群之间也展开生存竞争。结果公平偏好可以很快在社会中占据绝对主导地位。

图 6 个体-群体双层次演化的一次典型实验结果



前 200 代演化中强制规定只有 0.4 和 0.6 要价偏好且不能突变，形成了 0.4/0.6 非公平稳态；然后打开突变开关，结果大约在 230-310 代要价 0.5 的公平偏好成功入侵 0.4/0.6 非公平稳态，各族群形成公平偏好社会。

图 7 个体-族群双层次竞争中的公平偏好入侵

在个体-族群双层次演化仿真模型中，实验中主要参数的设置是全部个体数量在 400 左右，划分为 20 个族群，每个族群为 20 人左右(平均意义上 20 人)；偏好“基因”突变率  $\delta=0.00005$ ，即一种偏好基因变化为其他任何一种偏好基因的概率为  $\delta/18$ ，不到十万分之三；每个时期每个族群发动战争的概率为 0.05。但是，即使调整这些参数，实验结果也表现出很高的稳健性。

### 3. “突变”的力量

在我们的模型中，如果不允许“基因”的突变，则这将是一个确定系统的演化，演化的结果就取决于初始条件。在一个非公平偏好的双态社会，若没有偏好的“突变”，就不可能演化出公平偏好所主导的社会。这反映了行为类型(行事模式)的“突变”在社会演化中的伟大力量。对于社会而言，突变常常意味着是创新；有创新的社会才更可能向有效率的社会形态改进。在我们的模型中，没有考虑有意识的学习过程。在我们看来，有意识的学习常常是对文化的一种遵循；对文化的批判和背叛，常常来自原始的冲动和不满的情感，把它归入到人的生物学因素方面也许更为合适——有些人天生就喜欢离经叛道。或许我想表达自己坚信的一个信念：正是人类的生物方面的精神、情感和冲动，使得人类免于成为纯粹社会性的、理智的傻瓜！譬如，为了破解单次囚徒困境中不合作的悲剧，常常需要引入看来不那么“理性”的利他惩罚，依靠这个社会上的大爱之人，而利他惩罚行为确实也有其神经基础(Fehr and De Quervain et al., 2004)。

## 四、结论及对经济学的意义

本文试图从逻辑上表明，本能的公平行为在人类早期的生存竞争演化中比不公平的行为更具有适存性，因而本能的公平行为更能穿透自然选择的筛子而保留下来。当代行为和实验经济学所证实的公平心理偏好，可能有其演化根源，乃演化而来的人性之一。虽然本能的公平行为是无意识的结果，但在今天我们已经不难理解这种行为的适存性优势——合作是人类社会必要的且更能增加个体适存性的手段，但人们在寻求合作时需要面临合作机会大小和合作利益多寡的此消彼长：太过贪婪的利益索取，虽然可在达成的单次合作中得到更大的合作利益，但却难以寻觅到合作伙伴，合作达成的机会就会变少；太过慷慨的利益让步，尽管会使达成单次合作的机会大大增加，但单次合作的利益却下降了；无论合作机会的变少，还是合作利益的下降，都会有损于个体的适存性；结果，最有利于个体适存性的合作行为一定是那种能够最佳地平衡合作机会与合作利益的行为(不管这些行为是有意的还是无意的)。这也正是本文的演化博弈和仿真模型所要揭示的道理。特别地，本文的模型还表明，在一个封闭的族群演化中，能够平衡合作机会与合作利益的行为并不必定是公平的行为(不公平的双态社会也是可以形成的，尽管其涌现的概率很小)；但一旦引入族群之间的竞争，那么更倾向于公平行为的社会将因其成就了更多的合作机会而在竞争中战胜不公平行为的社会<sup>①</sup>。人类的公平偏好，很可能就源于人类进化史上本能公平行为的适存性优势。

公平偏好乃演化而来人性，这一观念对于现代经济学有什么意义？这个问题的回答，我个人认为至少有以下几方面：

1. 为行为经济学理论模型提供更为坚实的假设基础。当代行为经济学试图在经济学中增补“理性经济人”之外的某些假设，如公平偏好、双曲线贴现、参照依赖、利他惩罚等。但如何从源头寻找这些假设的逻辑支持？一种做法就是在演化过程中去寻找。马斯金(2007)、Dasgupta and Maskin(2005)对双曲线贴现的演化解释，叶航等(2005)对作为内生偏好的利他主义的演化解释，都是这样的尝试。当然，经济学家所探索的常常只是演化的经济(适存性)逻辑，而演化的生物学和心理学证据，则还需要生物学家和心理学家的继续努力。

<sup>①</sup> 这或许可解释为什么不平等的体制在相对封闭的国家比在面临全球竞争的国家更容易稳定存在。

2.对于非理性行为的反思。今日的非理性行为，在人类演化的历史上可能是符合“理性”的；或者说，有些现在看来不符合经济理性的行为，可能曾经是符合演化理性的。这为我们看待非理性行为提供了一个新的视角。具体到公平偏好，作为一种演化而来的心理机制，人们很可能在不需公平行为的时候(比如“独裁者博弈”实验中)也表现出某种程度的公平行为。其原因在于，演化具有时间间隔(time lags)，过去演化出的心理机制，并不一定最能适应当前的环境。换句话说，我们拥有石器时代的大脑，但却生活在现代社会之中(巴斯, 2007)。现代社会的非理性行为，是石器时代理性行为的余音。

3.探索理性的边界，更好地解释人类经济行为。现实中人类行为并非只有深层的理性推理，也有本能的冲动行为。若允许将人们进化而来的本能和心理机制与人们的文化学习过程结合起来，人类的真实行为就可以得到更好的解释。因为进化而来的行为并非不可改变，当人们对心理机制了解得越多，我们就越可能改变它；我们也不能认为进化而来的心理机制是死板的，人们只是刻板地按照某项心理机制采取行动，因为人类拥有众多的进化而来的心理机制并懂得积极地对环境进行应变，拥有的机制越多，或者越懂得对环境的应变，人们行为的灵活性就越高(巴斯, 2007)。人们会经历文化教育和社会化的过程，并在一定程度上学会并展开刻意的理性推理。理解了这一点，就可以很好地理解为什么最后通牒博弈(ultimate game)、独裁者博弈中，真正五五对分的结果是极其罕见的；而且独裁者博弈中分给回应者的比例相对于最后通牒博弈又下降了许多。这说明，人们的行为确有出自本能的公平偏好考量，同时也附带了其他方面（尤其是自我经济利益和行动环境约束）的理性的考量；这很可能也说明，人们在追求公平和物质利益之间存在着权衡取舍。如果是这样，那么在经济行为研究上，就可以把公平等偏好等作为自变量纳入主体的效用函数，并运用理性行动者模型加以研究，而不是把这些偏好归结为非理性并因而抛弃理性行动者模型。这正是 Gintis(2009)试图论证的：理性的边界不是人的非理性，而是人的社会性。

4.更好地解释经济现象和制定公共政策。若公平偏好是演化而来的人性之一，我们就不能指望它会在人的短暂一生中被“经济理性”彻底地重新塑造，因而我们也就不能无视这种人性。的确，人们会在许多经济决策中表现出公平动机。公平是一个重要的动机，因而也就是理解经济现象的一个重要因素。比如 Akerlof and Shiller(2009)就坚持：对于非自愿失业、通货膨胀和总产出之间的关系这类基本的经济现象，如果把公平考虑在内，就可以容易地给出解释；反之，如果不考虑公平，那么这些现象仍将是不解之谜。并且他们在其著作中身体力行，用三章内容阐述了公平及其对经济现象的解释。另外，在公共政策制定的层面，主流经济学所坚持的基于委托-代理机制的政策设计思路通常是：政府提供一项政策(合约)，这项政策(合约)将政策接受者的利益降到其参与约束水平上。这种政策制定与“最后通牒博弈”本质上是差不多的。由此我们不难理解，为什么有些政策，特别是涉及到利益分配或利益分割的政策，在满足政策接受者的参与约束水平条件下，仍然产生了诸多的冲突行为，一个可能的原因就在于，尽管那些政策并未把人们逼到山穷水尽的地步，但人们却可能对仅仅满足自己参与约束却谈不上利益分享的政策感到不平而愤懑。正所谓“不患寡，而患不均”。在经济和社会的层面，尽管我们尚未阅读到有关公平偏好影响政策效力和引发冲突后果的实证研究成果，但确实有个案研究(Krueger and Mas, 2004)表明工人对不公平的感知可以影响到工人的态度以及产品质量和消费者安全；也有实证研究(Alesina and Di Tella

et al., 2004)表明客观的收入越不平等则个人越倾向于认为自己是不幸福的,即便控制了个人收入、一系列个体特征以及年份和国别等虚拟变量之后也是如此。

最后有必要指出,本文仅仅为公平心理的演化起源提供了一套解释逻辑,它与现有的大量心理学实验和一些生物学研究之事实可以互相支持。但对于这套逻辑的进一步的事实检验,还需要来自古人类学、进化心理学和生物学研究的更多事实,这些事实研究可能远远超出了经济学家的能力范畴,但我们却不妨对那些领域的研究拭目以待<sup>①</sup>。

## 参考文献

- 埃里克·马斯金, 2007: 《最后一刻的道理》, 《比较》第33期。
- 巴斯, 2007: 《进化心理学: 心理的新科学》, 上海: 华东师范大学出版社。
- 鲍尔斯, 2006: 《微观经济学: 行为、制度与演化》, 周业安等译, 北京: 中国人民大学出版社。
- 董志强, 2008: 《制度及其演化的一般理论》, 《管理世界》第5期。
- 黄凯南, 2008: 《群体选择与个人主义方法论》, 《南方经济》第9期。
- 乔伊斯·马库斯, 2008: 《社会进化的考古学证据》(陈淳译), 《南方文物》第2期。
- 科林·凯莫勒, 2006: 《行为博弈—对策略互动的实验研究》, 贺京同等译, 北京: 中国人民大学出版社。
- 蒲勇健, 2007: 《建立在行为经济学理论基础上的委托—代理模型: 物质效用与动机公平的替代》, 《经济学(季刊)》, 第7期。
- 世界银行, 2006: 《2006年世界发展报告: 公平与发展》, 清华大学出版社。
- 汪丁丁、罗卫东、叶航, 2006: 《人类合作秩序的起源与演化(导读一)》, 上海世纪出版集团。
- 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。
- Akerlof, G. A. and R. Shiller, 2009, *Animal Spirits: How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism*, Princeton University Press.
- Alesina, A. and R. Di Tella, et al., 2004, “Inequality and happiness: are Europeans and Americans different?” *Journal of Public Economics*, 88 (9-10): 2009-2042.
- Bergstrom, T. C., 2002, “Evolution of Social Behavior: Individual and Group Selection.” *Journal of Economic Perspectives*, 16(2):67-88.
- Binmore, K. G., 2005, *Natural Justice*. Oxford, Oxford University Press.
- Bolton, G. E. and A. A. Ockenfels, 2000, “ERC: a Theory of Equity, Reciprocity and Competition.” *American Economic Review*, 90 (1): 166-193.
- Bowles, S., 2004, *Microeconomics: Behavior, Institutions and Evolution*. N.J., Princeton University Press.
- Bowles, S. and H. Gintis, 2003, “The origins of human cooperation.” in Peter Hammerstein (ed.), *the Genetic and Cultural Origins of Cooperation*, Cambridge: MIT Press: 429-443.
- Bowles, S. and H. Gintis, 2004, “Homo Economicus and Zoon Politikon: Behavioral Game Theory and Political

---

<sup>①</sup> 2009年12月22日《科技日报》消息称:日本玉川大学脑科学研究所研究员春野雅彦等人在21日的《自然·神经科学》网络版上发表文章称,他们发现当人们感到自己被不公平对待时,大脑的杏仁核会活跃起来;并且,越是强调公平的人,杏仁核越活跃。如果该项研究成立,则将是一项支持本能性公平偏好的一个证据。由于仅为新闻报道,故文中未作引证。

Behavior.” *SFI working paper*, Santa Fe Institute.

Brosnan, S. F. and F. Waal, 2003, “Monkeys reject unequal pay.” *Nature*, 425 (6955): 297-299.

Burnham, T. C. ,2007, “High-testosterone men reject low ultimatum game offers.” *Proceedings of the Royal Society B*, 274(1623): 2327.

Dasgupta, P. and E. Maskin ,2005, “Uncertainty and hyperbolic discounting.” *American Economic Review*, 95(4): 1290-1299.

Dawkins, R., 1976, *The Selfish Gene*, Oxford: Oxford University Press.

Dawkins, R. 1982, *The Extended Phenotype: the Gene as the Unit of Selection*, Oxford: Oxford University Press.

Dufwenberg, M. and A. Kirchsteiger ,2004, “A Theory of Sequential Reciprocity.” *Games and Economic Behavior*, 47(2): 268-298

Flinn, M. V. ,1997, “Culture and the evolution of social learning.” *Evolution and Human Behavior*, 18(1): 23-67.

Foster, D. and A. P. Young ,1990, “Stochastic Evolutionary Game Dynamics.” *Theoretical Population Biology*, 38 (1): 19-32.

Fehr, E. and A. A. Falk ,2002, “Psychological foundations of incentives.” *European Economic Review*, 46(4-5): 687-724.

Fehr, E. and A. B. Rockenbach ,2003, “Detrimental Effects of Sanctions on Human Altruism.” *Nature*, 422(13 Mar.): 137-140.

Fehr, E. and J. -. D. De Quervain, et al. ,2004, “The Neural Basis of Altruistic Punishment.” *Science*, 305(27): 1254-1258.

Fehr, E. and A. K. M. Schmidt ,1999, “A Theory of Fairness, Competition and Corporation.” *Quarterly Journal of Economics*, 114 (3): 817-868.

Gintis, H. ,2000, “Strong Reciprocity and Human Sociality.” *Journal of Theoretical Biology*, 206(2): 169-179.

Gintis, H. ,2006, “Behavioral Ethics Meets Natural Justice.” *Politics, Philosophy & Economics*, 5 (1): 5-32.

Gintis, H., 2009, *The Bounds of Reason: Game Theory and the Unification of Behavioral Sciences*, Princeton: Princeton University Press.

Hofbauer, J. and P. Schuster, et al. ,1979, “A Note on Evolutionary Stable Strategies and Game Dynamics.” *Journal of Theoretical Biology*, 81 (3): 609-612.

Krueger, A. B. and A. Mas ,2004, “Strikes, scabs, and tread separations: Labor strife and the production of defective Bridgestone/Firestone tires.” *Journal of Political Economy*, 112 (2): 253-289.

Markey, S. ,2003, “Monkeys Show Sense Of Fairness, Study Says.” *National Geographic News*: September 17.

Rabin, M. ,1993, “Incorporating Fairness into Game Theory and Economics.” *American Economic Review*, 83 (5): 1282-1302.

Robson, A. J., 2008, “Group Selection.” in Steven N. Durlauf and Lawrence E. Blume (ed.) *The New Palgrave Dictionary of Economics*, 2<sup>nd</sup> Edition, Eds. Palgrave Macmillan.

Sánchez, A. and A. J. A. Cuesta (2005). “Altruism may arise from individual selection.” *Journal of Theoretical Biology* , 235 (22): 233-240.

Seabright, P. ,2006, “The Evolution of Fairness Norms: An Essay on Ken Binmore's Natural Justice.” *Politics, Philosophy & Economics*, 5(1): 33-50.

Smith, J. M. ,1982, *Evolution and the Theory of Games*, Cambridge University.

Smith, J. M. and A. G. R. Price ,1973, “The Logic of Animal Conflict.” *Nature* , 246(02 Nov.): 15-18.

Trivers, R. ,1983, *The evolution of a sense of Fairness. Absolute Values and the Creation of the New World*. New York, International Cultural Foundation Press.

VeGa-Redondo, F. ,2003, *Economics and the Theory of Games*, Cambridge University Press.

Williams, G. C. 1966, *Adaption and Natural Selection: A Critique of Some Current Evolutionary Thought*, Princeton: Princeton University Press.

Young, P. ,1996, “The Economics of Convention.” *Journal of Economic Perspectives*, 10(2): 105-122.

Young, P. ,1998, *Individual Strategy and Social Structure*. N.J., Princeton University Press.

## Why Do We Prefer Fairness: An Explanation Based on Evolutionary View

Dong Zhiqiang

(School of Economics and Management, South China Normal University)

**Abstract:** Cotemporary behavioral economics make it convinced that human being has fairness preference, but do not reveal why it exists. This paper develops an evolutionary explanation to it that fairness preference of human being origins from the process of earlier human being evolution. Based on an evolutionary game model and a stochastic evolution simulation model, this paper shows that, a) in a closed group, singular state society with fairness preference is stochastic evolutionary stable equilibrium, while dual-state society with non-fairness preference is also evolutionary stable but not stochastic stable; b) if the competitions among groups are considered, the only evolutionary stable equilibrium is the singular state society with fairness preference. The reasons are as follows, there is a tradeoff between cooperation opportunities and the cooperation benefit and fair behavior can balance the effects of cooperation opportunity and cooperation benefit on survival competition, and then becomes the behavioral paradigm with optimal fitness in survival competition at both individual and group level. This notion is helpful to find a theoretical logic to support the fairness preference hypothesis in behavioral economics, as well as to rethink the human irrational behaviors from a new perspective, and explore the bounds of reason. The theoretical logic in this paper is yet to be documented by more determinative evidences from paleoanthropology, evolutionary psychology and biology.

**Key Words:** Fairness psychology; Evolution; Evolved psychological mechanisms; Multi-agent simulation; Rationality

**JEL Classification:** C91, B31, B25