进化视角下的间接互惠行为*

——评《道德体系生物学》及其开启的间接互惠行为研究

杜　月

亚历山大在《道德体系生物学》(*The Biology of Moral Systems*)(Alexander, 1987)中首次提出了间接互惠的概念,并用它来解释人类大规模合作行为:助人者为受助者提供帮助,对这种恩惠的报答不一定来自受助者,而可能来自被其他助人者帮助的其他受助者,此即间接互惠的机制(Alexander, 1987)。亚历山大还创造性地提出了他人在场所形成的名誉机制,蕴含了一种进化的视角。今年恰逢《道德体系生物学》出版20周年,本文拟用一种进化的视角来详细评论它所开启的间接互惠行为研究领域的当代发展。本文的评论专注于近年来间接互惠行为的研究进展,详细梳理当前有关间接互惠的理论论争:名誉问题、人类的复杂心智、公共物品的悲剧,以及人类宽容的美德,等等。间接互惠作为一种在自然选择中胜出并不断延续的行为策略,人类在其中体现了与任何物种相同的对于适应性的追求;与此同时,间接互惠作为人类特有的一种行为,又涵括了人类卓越的智慧与复杂的心智。

一、间接互惠行为在当今的研究价值

在经济领域,长期与固定熟人交易的情况越来越少,全球经济逐渐趋向于频繁快捷的、一次性的交易方式,与陌生人交手的情况越来越多。个体无法完全依据自己的经验做出判断,他必须求助于其他个体

* 感谢方文老师的信任和指导;应星老师也给予了中肯的评论和修改意见。文献的收集得到清华大学社会学系刘月雯同学的帮助,北京大学信息科学技术学院王一然同学提供了制图方面的意见。谨此致谢。

来了解陌生的交易对方。在这种情况下,一个个体在群体中的声誉以及他给对方带来的信任感成了成功的关键因素。近年来,部分经济学家逐渐提倡实践进化生物学的研究视角,这种视角带来了许多重要的启示(Hammerstein &Hagen, 2005:604—609)。

进化生物学感兴趣的问题是:人类独特的道德情感究竟源于何处?与其他完全社会性(eusocial)动物物种(如蜜蜂和蚂蚁等)不同,人类展现了大规模的非亲属间的互助(Wilson, 1975; Trivers, 1985)和一种强烈的道德情感,如对背叛者的唾弃和对助人者的颂扬。人习惯于评论他人的行为,即使这种行为和自己并没有直接的关系,并且能够有效利用舆论做出决策。因而这种间接互惠行为有可能指示了人类道德系统的起源。亚历山大提出"人类的道德系统就是间接互惠的系统",并且给出了令人信服的阐释(Alexander, 1987)。

在政治科学中,法律中心论曾经占据重要的位置,国家被认为是执行规则的主要渊源;由于个体根据纳什平衡而追求个人的利益最大化,社会合作成为不可能,个体无助地陷于社会困境的陷阱之中。完全理性的个人既无法组成社会,外部力量的干预则成为必要的。然而大量社会事实表明,人们并不完全依赖于法律制约,体现在互动过程中的非中心化的社会力量对于社会构成也具有强有力的影响(Ellickson, 1991)。基于名誉机制的间接互惠行为就是一个很好的例子,它启示了自我管理和民主的可能性(Ostrom, 1998:1—22)。

数学与计算机科学对于间接互惠模型的建立有重大贡献,这种人类美德也向建模的技术发起了挑战。作为人类特有的一种行为,间接互惠行为涉及到名誉的评估、流言、信息不完整等等变量,如何在模型中有效体现和操纵这些变量是一个很具挑战性的问题。

二、间接互惠思想的萌芽:亚历山大的洞察力

根据威尔逊(Wilson, 1975)和特里弗斯(Trivers, 1985),利他行为指付出一定的代价给对方带来一定的利益。在进化生物学中,这些代价和利益都是以适应性,即繁殖能力来衡量的。直接互惠是指两个个体间交换利他行为,从而达到两者共同受益的结果(Trivers, 1971:35—57)。在最简单的模型中,个体付出 c 使对方获得收益 b,b> c 以保证

互助的情况下达到双赢。在重复博弈的情况下,这就是著名的"囚徒困境"(Axelrod, 1984):当双方都选择帮助时,都可以得益 b－c,但是双方都拒绝帮助时只能得到 0,但当一方提供帮助而另一方拒绝时,一方的收益为 b 而另一方收益为－c。因而在对方付出时拒绝帮助能够获得最大的收益,但当双方都拒绝时谁也得不到任何收益。在这种语境下,"以牙还牙"(tit for tat)策略,即在第一轮互动中选择合作,在此后的每一轮中严格复制对方的行为,是达成合作的一种有效手段。触发策略(trigger strategy)是另一种可能的策略,它是指个体保持合作,一旦对方拒绝合作一次,个体从此拒绝合作到底。因此个体必须仔细衡量一次掠夺对方的利益和从此失去对方帮助的损失孰轻孰重,以此迫使个体选择合作。

现实社会中人群的流动性很大,两个个体长期重复博弈的几率很小(Boyd & Richerson, 1988: 337－356),因而个体间的重复博弈并不能很好地解释大规模的合作行为。亚历山大在《道德体系生物学》一书中提出了人类大规模合作的一种更合理的解释:助人者为受助者提供帮助,对这种恩惠的报答不一定来自受助者,而可能来自被其他助人者帮助的其他受助者,即间接互惠的机制(Alexander, 1987)。从此一批学者致力于研究从直接互惠到间接互惠模型的过渡。博伊德和里谢尔森用一种触发策略的变体(Boyd & Richerson, 1988)和环状的互助结构模型(Boyd & Richerson, 1989: 213－236)都没能够达到整个群体的合作,他们于是认为只有在互动密切的小群体中间接互惠才可能是一种比较重要的现象。然而这些结构上的变化并不能完成过渡,它在等待一种新的机制的出现——名誉机制。还是亚历山大富于启发性地提出:"间接互惠是在有兴趣的观众在场情况下直接互惠的结果"(Alexander, 1987),这些"有兴趣的观众"起着名誉评定的作用。在此启发下,波洛克和迪加特因(Pollock & Dugatkin, 1992: 25－37)在以牙还牙策略中引入名誉的机制,即个体在对对方毫无了解的情况下依据以牙还牙策略进行互动,但是一旦发现对方在上次互动时拒绝提供帮助就拒绝帮助对方。这种以牙还牙策略的变体被证明是一种进化稳定策略(evolutionarily stable strategy)。在此基础上诺瓦克和西格蒙德提出了更加精确的间接互惠模型。

三、口碑评定策略及其批评

诺瓦克和西格蒙德(Nowak &Sigmund, 1998a: 573—577)突破了以囚徒困境为代表的典型的直接互惠模型, 他们将亚历山大(Alexander, 1987)提出的间接互惠中非常重要的因素——声誉——引入模型, 用计算机模拟出几种不同的策略相互博弈的过程。在相当于一个古代人类社会群体人数的个体(100 人)中分布运用三种策略的行动者: 合作者(cooperator)、背叛者(defector)和辨别者(discriminator), 每个个体在一生中随机地被指定为助人者和被助者, 两个互动过的个体重逢的可能性被排除以保证相同两个个体重复博弈的直接互惠不能发生。每个个体都有被量化的口碑分数(image scores), 在生命起始时 s=0。当两个个体互动时, 若助人者提供帮助则其 s 增加一个单位, 其收益减少 c, 受助者的收益增加 b(b> c); 若助人者拒绝提供帮助, 则其 s 减少一个单位, 助人者与受助者的收益都不变。辨别者按口碑评定策略(image scoring)决定是否提供帮助: 每个个体都有不同的策略(计为 k), 当受助者的口碑分数 s 大于等于 k 时才提供帮助, 否则拒绝提供; 合作者运用合作策略, 总是无条件地对所有人提供帮助(k 值很小); 背叛者运用背叛策略, 拒绝帮助所有人(k 值很大)。在生命结束时, 个体留下个数与其适应性(用收益表示)成正比的子代。子代继承亲代的策略, 继续进行博弈。经过 166 代之后, 其他两种策略消失, 使用口碑评定策略的辨别者胜出并稳定进化。然而引入突变这一进化因素之后, 在数千代的博弈中, 辨别者策略被合作者取代, 而后者又被背叛者策略取代, 辨别者策略接着取代背叛者策略, 这种浮动反复且规律地出现, 口碑评定策略并不能稳定胜出。在真实的互动中, 除非在很小因此互动异常频繁的群体中, 个体的口碑分数是不太可能被所有群体成员感知的。这篇论文也分析了口碑信息的不完整对于互动的影响: 使用口碑评定策略的辨别者胜出的可能性与群体人数呈反比。

这篇论文虽然依旧利用简单的二人互动模式来研究互惠行为, 但已经打开了新的研究视野。基于追求适应性这一永恒的进化主题, 声誉与信任对于互动的影响以及语言传递信息的独特作用向着人类的独特性更迈进了一大步, 其开启的若干研究命题, 如信息的不完整性、突变与自然选择的相互作用、自身口碑对于互动策略的影响, 以及信任的

进化意义等都成为此后研究的突破口。

计算机模拟适于分析各种因素的加入对于结果的影响，而更精确的量化分析则依靠两极化模型(two score)来完成。在另一篇论文中，诺瓦克和西格蒙德(Nowak &Sigmund，1998b：561－574)利用两极化模型来进一步分析口碑评定策略的稳定性。个体的口碑只有"好"与"坏"两种，取决于其上一次作为助人者是否提供了帮助。

口碑评定策略最容易被质疑的一点在于：如果个体是否能够得到帮助完全取决于个体自身的口碑分数，理性的个体就应该在任何情况下都选择合作而不顾及对方的口碑是好是坏。莱马尔和哈默斯坦认为口碑评定策略之所以成为稳定策略是诺瓦克和西格蒙德(Nowak & Sigmund，1998a，1998b)所构造的孤立小群体中基因漂移所致，而在进化过程中基因漂移的作用又是很微小的。他们对口碑评定策略的另一点疑义在于只有当付出 c 相对于收入 b 足够小的情况下，该策略才成为稳定的策略(Leimar &Hammerstein，2001：745－753)。

莱马尔和哈默斯坦认为口碑评定策略失败的主要原因在于出于正义拒绝提供帮助的个体承受了不公正的惩罚。他们依据萨格登(Sugden，2004)提出的"好名声"概念提出了另一种名声策略模型(standing)：起初所有个体都拥有好名声，拒绝帮助名声好的个体会使助人者丢掉好名声，即拒绝帮助有坏名声的个体并不给助人者带来任何名声上的损失。失去好名声的个体可以通过助人(不论受助者名声好坏)来重新获得好名声。[①] 所以一个助人者应该在以下两个条件具备其一或是同时具备时帮助受助者：1. 助人者本身失去了好名声；2. 受助者有好名声。

我们在生活中并非作为完全的理性人在活动，各种行为偏差是不可避免的。在模型中加入偏差是使模型更加接近真实的一种方式，一种策略在各种偏差(表 1)的作用下是否依然可以成为稳定的策略是研究的一项重要命题。口碑评定策略与名声策略的较量中偏差分析是很重要的一环。

① 在另一种同时提出的严格评定策略(judging)中，失去好名声的个体只能通过帮助名声好的个体赢回自己的好名声。

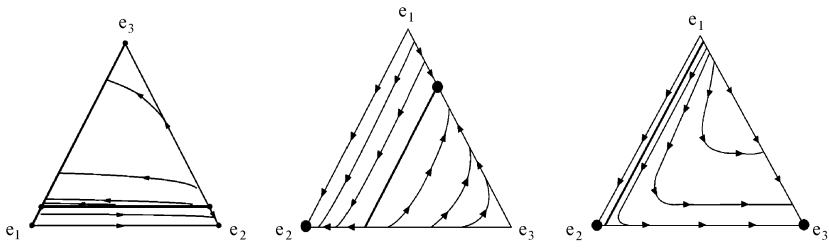
表 1 执行偏差与认知偏差

执行偏差: 个体根据自身运用的策略意欲提供帮助但由于各种原因无法提供帮助。在这里一般不考虑无意提供帮助客观上却提供了帮助的情况。
认知偏差: 助人者对于受助者的名声认知存在偏误, 进一步对自身的名声估计错误。如助人者拒绝帮助一个他认为名声坏的(实际上是好的)受助者进一步导致自身丢掉好名声而却没有意识到这一点。认知偏差会使分析过程变得十分复杂, 所以大部分研究都会忽略它。
应该注意的是, 这些行为偏差更应该看作疾病或意外等不稳定的状况, 阻止了个体按其真实意愿行事, 而非可遗传的特征。

资料来源: Panchanathan & Boyd, 2004: 499—502; Leimar & Hammerstein, 2001.

当模型中没有引入各种行为偏差时, 只要采取口碑评定策略的辨别者在群体中所占比例高于某值则背叛者策略将消失, 无条件合作策略和口碑评定策略将占据人群; 相反, 若辨别者的比例低于此值则背叛者将会战胜其他两种策略(图 1a), 在无条件合作策略和口碑评定策略间自然选择是中性的, 二者的适应性相当。当引入执行偏差时, 两种合作策略的适应性发生变化, 自然选择在二者之间确定了一个稳定点, 而这个比例恰好也是背叛者入侵的条件。因而任何轻微的干扰如基因漂移等等都会使辨别者的比例发生浮动从而导致背叛者入侵并统治(图 1b), 口碑评定策略不再是进化稳定策略。当加入执行偏差时, 名声策略却依然是一个进化稳定策略(图 1c), 合作者的比例同样也要高于一个特定值以剔除背叛者。

使用名声策略的个体需要有效地辨别拒绝提供帮助者的意图, 即是出于自私(为节省自己的资源而拒绝提供帮助)还是出于公正(拒绝帮助名声不好的受助者作为对其惩罚)。因而与口碑评定策略相比, 名声策略需要更高的认知能力, 需要同时辨别行动者的行为及其深层的意图。这种对于认知能力的更高要求呼应了间接互惠行为只见于人类的局限性, 甚至也很可能有助于揭示间接互惠行为的进化起源(Panchanathan & Boyd, 2003: 115—126)。但相关的质疑也接踵而至。首先, 一个个体要想分析其互动对象拒绝他人的意图, 还要了解其互动对象的受助者的名声如何。在如此紧密的环环相扣的信息传递过程中, 一旦发生认知偏差, 整个认知系统就会一塌糊涂。其次, 个体在运用名声策略进行互动时所需要的信息量明显增加, 在实际互动中人类是否能够掌握如此大的信息量是一个很大的问题。最后, 人类真的会



注：e1—合作者策略　　e1—合作者策略　　e1—合作者策略
e2—背叛者策略　　e2—背叛者策略　　e2—背叛者策略
e3—口碑评定策略　　e3—口碑评定策略　　e3—名声策略

资料来源：图 1a 引自 Nowak & Sigmund, 1998b；图 1b 与图 1c 引自 Panchanathan & Boyd, 2003。

图 **1**　两种策略的比较

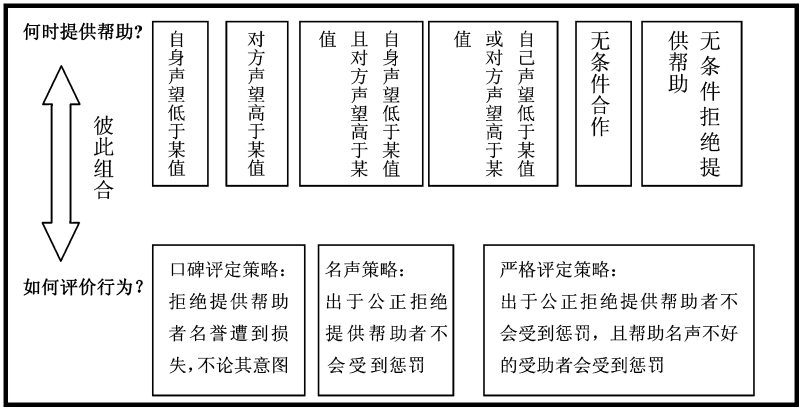
自觉加工更深一层的信息吗？米林斯基等(Milinski et al., 2001: 2495—2501)在实验室实施的真人试验发现个体并不会自觉地利用更深一层的信息去辨别拒绝提供帮助者的意图，他们趋向于同等对待公正和自私的拒绝提供帮助者。不过也许这种辨别失败是由于实验设计者的失误：只提供了行动者的行为历史而忽略了语言这一具有进化意义的交流工具(Panchanathan & Boyd, 2003)。

随着模型的进一步完善，出现了一些整合性的工作(Brandt & Sigmund, 2004: 475—486; Ohtsuki & Iwasa, 2004: 107—120)。把个体决定是否提供帮助的前提与系统评价行为的方式组合起来得到许多种不同策略。大月和岩狭(Ohtsuki & Iwasa, 2004)通过分析 4096 种策略的稳定性得出了八种稳定策略，它们共同的特点是：1. 帮助名声好的受助者会获得好的名声，相反名声会受损；2. 拒绝帮助名声坏的受助者应该被看作是好的行为，获得好的名声(见表 2)。

名声策略在每方面都显得比口碑评定策略要强，那是因为当惩罚行为没有任何代价时，基于惩罚的合作必然会得以加强。但是勃兰特和西格蒙德(Brandt & Sigmund, 2004)要强调的是，有代价的惩罚才是真实的，而口碑评定策略即使在有行为偏差出现的情况下还足以支撑间接互惠的发展。同时他们还指出潘查南森和博伊德(Panchanathan & Boyd, 2003)借以质疑口碑评定策略的两极化模型对行为偏差极其敏感

和脆弱,因而是一种阻碍合作行为发展的人为限制。

表 2 整合工作——帮助前提与评价方式的组合



资料来源: Brandt & Sigmund, 2004.

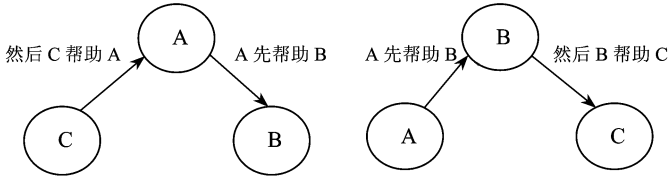
口碑评定策略的提出开启了间接互惠行为的深层研究,在对它进行讨论和批评的过程中,大量的研究面向涌现出来,如行为偏差,信息的不完整性,以及实验室中用于检测理论的真人试验。这些真人试验在后期突破了检验理论的思路,用于独立研究或提出新的命题与理论,有了很大的发展。

四、模拟人群与互动的结构

为防止小群体中基因漂移现象的过高频率发生,莱马尔和哈默斯坦(Leimar & Hammerstein, 2001)提出了一种模拟人群的构造方式,人群由 n 个小群体组成,每个小群体含有 g 个个体。在产生子代个体时,计算出某种基因型在本群及在全部人群中的适应性,并成比例地留下相应的子代个数。一个子代个体的基因有 p 的可能性取自本群,而有 1 —p 的可能性取自全体人群。由此繁殖能力强的群体将会对全部人群的基因有更大影响,但每个小群体对于个体产生作用的 p 是相同的。

根据亚历山大(Alexander, 1987),间接互惠行为的互动结构对于理解人类大规模的合作行为是有重要意义的。间接互惠模型中常见的随

机抽取一对助人者和受助者的方式确实有待改善。依循亚历山大提出的两种间接互惠模型(见图2),间接互惠行为可以在闭合的环状结构中模拟——间接互惠就是"长短不一的相互连通的环"(Boyd & Richerson, 1989)。一个典型的例子是在一个闭合的环中,个体接受来自左手边个体的帮助并决定是否为右手边的个体提供帮助(Greiner & Levati, 2005: 711—731)。很有意思的是穆赫塔奇米和梅(Mohtashemi & Mui, 2003: 523—531)提出的一种名为"集体记忆"(collective memory)的网状互动结构。社会互动结构就像一张信任的大网,结点代表了每个个体,网的边则代表了朋友和熟人之间信任的多少。个体获得信息的一个关键来源就是咨询朋友和熟人。这张信任之网是包含顺序的,因此信息的取回和发散都是选择性而非随机性的。网状结构的另一个特点是具有增长性。当一对互动者互动时,助人者成为受助者及受助者熟人的单向熟人(助人者并没有从互动中得知受助者的信息),根据集体记忆的定义,助人者的熟人也会成为受助者及其熟人的单向熟人。在受助者下一回合作为助人者时可以咨询新旧熟人的意见进行决定。但是这种网状结构也存在问题:助人者在拒绝帮助时也会成为受助者的熟人,受助者在做出决定时会同等地参考背叛者和合作者的意见吗?熟人与熟人之间的信赖程度也是不同的。另外在一对熟人之间也可能存在信任程度不对等的情况。因此个体对于不同熟人不同程度的信赖以及信任关系的不对称性将是未来研究的方向。



资料来源: Nowak & Sigmund, 2005: 1291—1298.

图 **2**　两种间接互惠模型

五、对于名誉的深入探讨

经过试验验证,有助他行为历史的受助者确实得到更多帮助

(Wedekind &Milinski, 2000: 850—852)。这表明有两个动力可以促使个体进行利他行为: 1. 实施利他行为给个体带来的名誉; 2. 由好的名声进一步得到的实际利益。那么这两个动机孰轻孰重? 为了研究这个问题, 研究者将助人者分为两组, 一组可以通过给与帮助获得好名声而另一组不可以, 结果, 前一组的帮助率达 74%而后一组达 37%(Engelmann &Fischbacher, 2002)。可见人类对于两者的诉求处于几乎同样重要的地位。

孔雀开屏, 鸟类殷勤地为对方梳理羽毛, 这些过于铺张浪费的求偶行为无疑降低了个体的适应性, 然而这些行为作为一种信号却可以诚实地反映个体的生命质量, 即其适应性(Zahavi, 1975: 205—214)。扎哈维(Zahavi, 1995: 1—3; Zahavi &Zahavi, 1997)坚持人类的利他行为符合与上述动物行为相同的阻碍原则, 即通过牺牲短期的适应性(如帮助他人或向集体捐赠)来获得长期的更大的适应性补偿(如高质量的配偶与同盟), 且利他行为应该看作是竞争性而非合作性的。

间接互惠理论与这种理论的最大差别在于名誉在互动中的作用。在间接互惠模型中, 个体通过帮助他人获得好名声, 以进一步得到他人的帮助, 而在扎哈维的理论中, 个体如果拥有了好名声, 不仅可以得到相应的作为回报的帮助, 还可以得到一些其他的利益, 如特权、配偶与好的联盟。间接互惠理论中名誉只是获得回报的帮助的一种手段, 而在扎哈维的理论中, 当额外获得的利益远大于回报的帮助时, 短期内名誉就是个体追求的目标。个体通过各种利他行为竞争获得稀缺的名誉资源, 这时受助者的名声好坏已不再重要, 利他行为以这种竞争而非互惠的形式进行着。帮助行为的回报与支出的比例也不再重要, 个体只需确定最终所获的利益足以偿还利他行为降低的适应性即可(Roberts, 1998: 427—431)。

最近的一个试验发现在有其他顾客鱼观察的时候一种清洁夫鱼干活会分外卖力和小心, 而顾客鱼在观察后趋向于选择为其他顾客鱼热情服务的清洁夫鱼[①] (Bshary &Grutter, 2006)。这很有可能是一种种内竞争压力促成的利他行为。两个物种之间(如顾客鱼与清洁夫鱼之间)

① 清洁夫鱼帮助顾客鱼啃食掉其体表附着的寄生物, 但同时要面临将其美味的黏液层一起咬下来的诱惑。实验发现顾客鱼在观察后会选择干活更加卖力的清洁夫鱼, 而清洁夫鱼一旦感受到有顾客鱼在场则会挡住诱惑只啃食寄生物(啃食黏液层引起的骚动会吓走其他顾客鱼)。

的直接互惠共生关系有可能沿着这条竞争的轨迹过渡到间接互惠这一更高的层次。这种围绕名誉而展开的利他行为的竞争性对于间接互惠理论是很有启发性的。首先,好的配偶、特权、或联盟等不同于一般回报性帮助的行为,有助于大幅度提高适应性的事物应该受到重视;其次,竞争性利他行为理论通过利他者的集聚与对背叛者的排斥来模拟行为,这种联盟的形成更贴近人类真实活动过程,也很值得借鉴。

六、人类的复杂心智

间接互惠是人类独特的利他行为,与之相应的是人类复杂的心智,对于自己处境的分析与对于外部环境和气氛的感知都会通过人类独特的心理因素发挥作用并影响其行为。

除去常见的基于对方名声的间接互惠模型外,亚历山大曾经提出过间接互惠的另一种可能的形式:个体只知道自己受到帮助的状况而并不知道其要帮助的个体或群内任何其他个体的名声即合作的态度。个体依据自己的经历决定是否对受助者给与帮助及给予多少帮助。这种情况下个体做出决策的意图源于对于公平而非对方名声的考虑。博尔顿等(Bolton et al., 2005:1457－1468)在真人试验中发现上一轮得到帮助的助人者提供帮助的比例达71.0%,而上轮没有得到帮助的助人者提供帮助的比例为32.9%,可见个体自身的经历是其考虑是否给与帮助的必要因素。在一项环状网络结构的间接互惠模型的真人试验中,个体只知道其左手边的个体给与其帮助的多少,并以此决定是否提供帮助给在其右边的个体(Greiner &Levati, 2005)(见图2)。试验结果表明,无论如何改变决策的环境,如互动群体的大小,决策的顺序和个体重新组合的方式,平均给与的实验货币数(稍后可按比例兑换成实际货币)都是正数,由此证明了排除一切策略性考虑,仅仅依据公平原则形成和保持间接互惠行为的可能性。由于个体自身的经历可以潜在地反映合作在整个人群中的比例,引发人类关于公正的心理机制,这基于公正的第二种间接互惠模型也是极具启发性的。

如上文所述,基于间接互惠模型的实验室真人试验近年来有很大发展。由最初的验证间接互惠理论在人群中的真实存在(Wedekind & Milinski, 2000;Milinski et al., 2001)到独立研究间接互惠理论架构下的

问题，如互动的结构（Greiner &Levati，2005）、信息可知性对互动的影响（Bolton et al.，2005）、间接互惠与公共物品（Milinski et al.，2002：424－426）等等，再到对传统试验方法进行反思和质疑（Haley &Fessler，2005：245－256）：人类真实的心智使真人试验具有无可替代的价值，也正是基于这种真实的心智，实验室试验也要不断进行自我反思。毋庸置疑的是，在实验室试验与真实生活中都存在一些无法用基于名誉的考虑来解释的个体行为：在试验中，个体能够根据实验条件提供的形成自身好名声的机会适当改变自己的行为，然而即使在一些消除名誉形成机制的试样中，一定程度的利他行为依然存在。一些学者趋向于用早期人类社会发生的某种群体选择过程来解释这一现象（Gintis，2000：169－179；Gintis et al.，2003：153－172），然而在一个近期的实验室试验中，黑利和费斯勒（Haley &Fessler，2005：245－256）验证了是人类进化过程中他者在场的机制导致了利他行为。亚历山大也曾经提到“间接互惠是在有兴趣的观众在场情况下直接互惠的结果”。在场的也许并不一定是有兴趣和意向的观察者，人类基于长期进化而发展的视觉或听觉上的刺激也许就足以促成他人的在场。黑利和费斯勒就是利用耳塞与电脑屏幕上的眼状图标来进行验证的：戴上耳塞可以略微降低合作的程度，而眼状图标却可以激发多达两倍数量的帮助。由此试验设计中除去外显的试验环境，最微小的操作细节也可以影响到个体的决策。他者的存在，这种在人类漫长进化过程中不断显现的影响因素应该在实验设计和理论研究中受到重视（方文，2006）。

不断推测外界环境对自身未来行动的反应也是人类的复杂心智之一。在实际的互动中，我们免不了要问自己一个问题：如果这回不提供帮助对我会有什么不利的影响？用模型表达出来就是：每个个体都在度量当自己的口碑分数为 s 时得到他人帮助的可能性 QS，和不提供帮助这一举动在避免降低适应性的同时所损失的得到帮助的可能性，即 QS＋1－QS－1，每个个体都有一个承受限度 ΔQ，当损失超过这个限度时就提供帮助（Leimar &Hammerstein，2001）。

七、如何解除“公共物品的困境”

分享食物、共同狩猎和战争等活动都涵盖了大规模人群，根据定

义,公共物品可以为群体中的任何成员所享用,无论其是否对公共物品做出了贡献。因此每个个体都有搭便车的动力,这就是所谓的"公共物品的困境"(public good dilemma)。个体只有在认定其他个体会投资的情况下才会对公共物品做出贡献。在一次性的公共物品试验中,尽管预料到自私的个体会不做任何付出,但一般个体通常会贡献出自己在试验开始时所得到的财产的40%—60%(Dawes, 1980:169—193,转引自Fehr &Fischbacher, 2003:785—791),然而当试验连续重复10次时,个体投资普遍处于很低的水平(Isaac &Walker, 1988:179—199)。由于个体会基于他人的投资而考虑自身投资的多少,一小部分自私的群体就足以使一个起初合作者占大部分的群体达到零合作率的均衡(Fehr &Schmidt, 1999:817—868)。

如何才能解除"公共物品的困境"?对于名誉的追求和对于惩罚的畏惧促成了最核心的解决方法。如果可以将惩罚直接指向在公共投资中表现自私的个体,即在直接互惠行为中对其进行惩罚,群体的合作水平将有很大提高(Fehr &Gachter, 2000:980—994)。另一种惩罚不是直接的,而是依靠在互动中拒绝为在公共投资中表现自私的个体提供帮助,即对背叛者的排斥来实现惩罚,这种间接惩罚借助了间接互惠中的名誉机制。真人试验表明,在公共物品试验中加入间接互惠博弈的环节有助于提高群体的合作水平(Milinski et al., 2002:424—426)。潘查南森和博伊德(Panchanathan &Boyd, 2004:499—502)运用分析的方法系统地揭示了间接互惠与公共物品投资之间的关系:首先,不光是公共物品的投资,只要个体实现一种行为的损失小于其在稍后的间接互惠环节中的所得,这种行为与间接互惠行为连接在一起时都可以达到稳定。其次,将间接互惠与公共物品投资行为连在一起的策略并不能取代单纯进行间接互惠行为的策略,因此尽管间接互惠行为有助于提高公共物品的投资水平,但两者并不能作为一个混合体同时进化。因此我们还需另外寻找公共物品投资行为的来源。

八、宽容的美德

每个个体都有不同的生活状况(体现了其适应性的高低),这种质量的差异很可能通过合作行为显现和交流(Leimar, 1997:1209—

1215)。莱马尔和哈默斯坦(Leimar &Hammerstein, 2001)提出了一种依据状态的间接互惠模型(state-dependent reciprocity),即个体处于较好的生活状态时其给与他人帮助时自身的损失相对较低,这时个体会选择提供帮助,而状态较差时则相反:即使按照名声策略个体应该提供帮助,也因为代价过大而放弃提供帮助。这种行为差异有助于体现个体的质量,当个体的状态相对稳定时,依据状态的间接互惠行为将导致同类质量的个体汇聚形成小群体。与互惠行为显示个人质量的观点相对,费什曼(Fishman, 2003: 285—292)强调同一个体在生命的不同时段会有不同的状态,一个很有力的问题是:无辨别能力的合作者如果能在任何时候为他人提供帮助,这样的人自己又怎么会需要他人帮助呢?因此间接互惠是这样一个行为系统:当提供帮助给自己带来的损失很低的时候个体提供帮助,并在获得帮助收益很高的时候寻求他人帮助。当个体处于较差的境遇,即自己处于需要他人帮助的状态时,即使他主观上愿意提供帮助也无法实践,这就是所谓的无意背叛(involuntary defection)或表型背叛(phenotypic defection)(Lotem et al., 1999: 226—227; Fishman et al., 2001:: 87—95; Sherratt &Roberts, 2001: 313—317)。这种背叛与上文提到的执行错误相似,会导致在有辨别能力和无辨别能力的合作者之间进行选择时偏向前者。[①] 但有意思的是,当引入这种无意的背叛时,辨别合作策略只有与非辨别合作策略共存时合作才能达到稳定。这两种策略的混合有两种可能的情况:1. 信息不完整,群体由有辨别能力的合作者组成,但当他们不知道对方的名声时会采取信任策略[②],即非辨别合作策略,从而达到两种策略的混合;2. 群体中存在与两种策略的比例相应的行动者(即有辨别能力和无辨别能力的合作者)。不论是信息的不完整还是存在无条件合作者,都说明一定程度上对于口碑评定策略的偏离有助于合作行为达到稳定,即间接互惠是基于不完美个体的互动行为,一定程度的宽容是必要的。

道金斯(Dawkins, 1976)提出的绿胡子效应(green beard effect)启示了另一种合作进化的可能性,即个体不但继承了长绿胡子的基因,同时

① 试想一个人群只由这两种合作者构成,由于他们都有好的名声因而总是为对方提供帮助,这时选择是中性的,而一旦引入这种无意的背叛,无辨别能力的合作者将会因为无意的拒绝合作而受到有辨别能力的合作者的惩罚,而非相反。

② 在信息不完整的情况下,个体不了解对方的名声时,采取信任的态度更有利于个体的适应性发展(Noark &Sigmund, 1998b)。

继承了为其他绿胡子个体提供帮助的基因,这种基于亲缘选择理论的相似个体间的互助可以为合作行为提供一种解释。廖洛和阿克塞尔罗德(Riolo &Axelrod, 2001:441－443)进一步提出一种可能的解释:每个个体都被赋予一个固有的特质,即容忍限度,零容忍限度表明只帮助与自己极其相似的个体,最大化的容忍限度表明帮助任何人,这种特质被子代所继承。电脑模拟的结果显示,一定程度的合作可以建立起来。一定程度上的非随机安排,比如相似个体的聚集和互动,可以促进合作行为的发展(Axelrod &Hamilton, 1981:1390－1396)。因此当容忍限度适当时,通过同类个体的聚集可以使群体的合作行为达到稳定;但是当容忍限度过低时,有辨别能力的合作者很有可能不再用全部力量去惩罚背叛者而是毫无意义地将矛头指向合作者内部(Sigmund &Nowak, 2001:403－405)。因而,在间接互惠的环境中,一定限度的宽容有益于合作行为的稳定。

总之,亚历山大的《道德体系生物学》开启了间接互惠行为研究的进化视角。我们已经发现这是一个在对适应性普遍追求的框架内不断填充人类独特性质的过程。处于进化树最高端的人类将从根求所,以共通性与特异性的双重眼光去了解和审视真实的人类自身。

参考文献:

方文,2006,《宗教行动者——一种宗教资格论》,未刊稿。

Alexander, R. D. 1987, *The Biology of Moral Systems*. New York: Aldine de Gruyter.

Axelrod, R. 1984, *The Evolution of Cooperation*. New York: Basic Books.

Axelrod, R. &W. D. Hamilton 1981, "The Evolution of Cooperation." *Science* 211.

Bolton, G. E., E. Katok &A. Ockenfels 2005, "Cooperation among Strangers with Limited Information about Reputation." *Journal of Public Economics* 89.

Boyd, R. &P. J. Richerson 1988, "The Evolution of Reciprocity in Sizable Groups." *Journal of Theoretical Biology* 132.

—— 1989, "The Evolution of Indirect Reciprocity." *Social Networks* 11.

Brandt, H. &K. Sigmund 2004, "The Logic of Reprobation: Assessment and Action Rules for Indirect Reciprocity." *Journal of Theoretical Biology* 231.

Bshary, R. &A. S. Grutter 2006, "Image Scoring and Cooperation in A Cleaner Fish Mutualism." *Nature* 441.

Dawes, R. M. 1980, "Social Dilemmas." *Annual Review of Psychology* 31.

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

Ellickson, R. C. 1991, *Order Without Law: How Neighbors Settle Disputes*. Cambridge, Massachusetts: Harvard University Press.

Engelmann, D. & U. Fischbacher 2002, "Indirect Reciprocity and Strategic Reputation Building in An Experimental Helping Game." Working Paper 132, Institute for Empirical Research in Economics, University of Zurich.

Fehr, E. & K. M. Schmidt 1999, "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114.

Fehr, E. & S. Gachter 2000, "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90.

Fehr, E. & U. Fischbacher 2003, "The Nature of Human Altruism." *Nature* 425.

Fishman, M. A., A. Lotem & L. Stone 2001, "Heterogeneity Stabilizes Reciprocal Altruism Interactions." *Journal of Theoretical Biology* 209.

Fishman, M. A. 2003, "Indirect Reciprocity among Imperfect Individuals." *Journal of Theoretical Biology* 225.

Gintis, H. 2000, "Strong Reciprocity and Human Sociality." *Journal of Theoretical Biology* 206(2).

Gintis, H., S. Bowles, R. Boyd & E. Fehr 2003, "Explaining Altruistic Behavior in Humans." *Evolution and Human Behavior* 24(3).

Greiner, B & M. V. Levati 2005, "Indirect Reciprocity in Cyclical Networks: An Experimental Study." *Journal of Economic Psychology* 26.

Haley, K. J. & D. M. T. Fessler 2005, "Nobody Is Watching? Subtle Cues Affect Generosity in An Anonymous Economic Game." *Evolution and Human Behavior* 26.

Hammerstein, P. & E. H. Hagen 2005, "The Second Wave of Evolutionary Economics in Biology." *Trends in Ecology and Evolution* 20.

Isaac, R. M. & J. M. Walker 1988, "Group-size Effects in Public-goods Provision-the Voluntary Contributions Mechanism." *Quarterly Journal of Economics* 103.

Leimar, O. & P. Hammerstein 2001, "Evolution of Cooperation through Indirect Reciprocation." *Proceedings of the Royal Society of London Series B-Biological Sciences* 268.

Leimar, O. 1997, "Reciprocity and Communication of Partner Quality." *Proceedings of the Royal Society of London Series B-Biological Sciences* 264.

Lotem, A., M. A. Fishman & L. Stone 1999, "Poor Phenotypes Stabilize Indirect Reciprocity by Image Scoring." *Nature* 400.

Milinski, M., D. Semmann & H. J. Krambeck 2002, "Reputation Helps Solve the 'Tragedy of the Commons'." *Nature* 415.

Milinski, M., D. Semmann, T. C. M. Bakker & H. J. Krambeck 2001, "Cooperation through Indirect Reciprocity: Image Scoring or Standing Strategy?" *Proceedings of the Royal Society of London Series B-Biological Sciences* 268.

Mohtashemi, M. & L. Mui 2003, "Evolution of Indirect Reciprocity by Social Information: The Role of Trust and Reputation in Evolution of Altruism." *Journal of Theoretical Biology* 223.

Nowak, M. A. & K. Sigmund 1998a, "Evolution of Indirect Reciprocity by Image Scoring." *Nature* 393.

—— 1998b, "The Dynamics of Indirect Reciprocity." *Journal of Theoretical Biology* 194.

—— 2005, "Evolution of Indirect Reciprocity." *Nature* 437.

Ohtsuki, H. & Y. Iwasa 2004, "How Should We Define Goodness? Reputation Dynamics in Indirect Reciprocity." *Journal of Theoretical Biology* 231.

Ostrom, E. A. 1998 "A Behavior Approach to the Rational Choice Theory of Collective Action." *The American Political Science Review* 92.

Panchanathan, K. & R. Boyd 2003, "A Tale of Two Defectors: The Importance of Standing for the Evolution of Reciprocity." *Journal of Theoretical Biology* 224.

—— 2004, "Indirect Reciprocity Can Stabilize Cooperation without the Second-order Free Rider Problem." *Nature* 432.

Pollock, G. & L. A. Dugatkin 1992, "Reciprocity and the Emergence of Reputation." *Journal of Theoretical Biology* 159.

Riolo, R., M. D. Cohen & R. Axelrod 2001, "Evolution of Cooperation without Reciprocity." *Nature* 414.

Roberts, G. 1998, "Competitive Altruism: From Reciprocity to the Handicap Principle." *Proceedings: Biological Sciences* 265.

Sherratt, T. N. & G. Roberts 2001, "The Importance of Phenotypic Defectors in Stabilizing Reciprocal Altruism." *Behavioral Ecology* 12.

Sigmund, K &M. A. Nowak 2001, "Tides of Tolerance." *Nature* 414.

Sugden, R. 2004, *The Economics of Rights, Cooperation and Welfare*(2nd ed.). New York: Palgrave Macmillan.

Trivers, R. 1971, "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology* 46.

—— 1985 *Social Evolution*. Menlo Park, California: Benjamin Cummings.

Wedekind, C. &M. Milinski 2000, "Cooperation through Image Scoring in Humans." *Science* 288.

Wilson, E. O. 1975, *Sociobiology*. Cambridge, Massachusetts: Harvard University Press.

Zahavi, Amotz & Avishag Zahavi 1997, *The Handicap Principle*. New York: Oxford University Press.

Zahavi, Amotz 1975, "Mate Selection-A Selection for a Handicap." *Journal of Theoretical Biology* 53.

—— 1995, "Altruism as A Handicap-The Limitations of Kin Selection and Reciprocity." *Journal of Avian Biology* 26.

作者单位：北京大学社会学系

责任编辑：罗　琳