信念与心理博弈：理论、实证与应用 *

姜树广　韦倩

内容提要：对物质财富的追求并非人类的全部动机，人们的日常决策行为还受到道德、情感情绪、社会规范等左右而偏离物质收益最大化的目标。以信念依赖动机为基础的心理博弈论开辟了情感情绪决策研究的热潮。在心理博弈中，参与者的效用不仅依赖于最终的物质支付，还取决于参与者所持信念引发的心理状态。本文对以信念依赖动机为核心的研究进行了系统的述评。在简要梳理信念及心理博弈论基本理论框架的基础上，对基于意图的互惠、失望与内疚厌恶、自我形象与社会形象，以及焦虑和源自预想的效用等信念依赖的基本人类行为动机进行了总结评价，并梳理了相关实验与神经科学的证据。预期信念与心理博弈研究的深入将对情感情绪决策、信任问题、制度研究、非市场决策行为，以及多学科融合产生重要影响。

关键词：信念依赖动机　心理博弈论　情感情绪　信念引出实验

一、引言

涉及不同主体相互作用的决策模式普遍存在于经济社会中，但建立在理性自利人假设基础上的传统经济学理论主要关注市场行为，而将他人行为和意图的直接影响进行了最简化处理。虽然博弈论是用来分析相互策略的，但传统博弈论仍在理性自利人的前提下，不仅假设理性还需要共同知识，确保参与者能够预期对方的策略，这样才得以分析均衡。传统博弈理论的预测遭到了大量实验证据的反驳，实验中受试者表现出显著的对自利最大化行为的背离 。①

涉他偏好开始得到重视并涌现丰富的理论，如温情效应（warm glow）（Andreoni，1990）、分配公平偏好（Fehr & Schmidt，1999；Bolton & Ockenfels，2000）、基于意图的互惠（Rabin，1993；Dufwenberg & Kirchsteiger，2004）、内疚厌恶（guilt aversion）（Battigalli & Dufwenberg，2007）、身份认同（identity）（Akerlof & Kranton，2000），以及社会形象（social image）或社会尊重（social esteem）（Bernheim，1994；Benabou & Tirole，2006；Ellingsen & Johannesson，2008）等。这些理论一个共同特征在于捕捉到人类不仅在乎物质（货币）收益，而且受到“心理效用”的驱动，只是不同的理论中构成心理效用的成分不同。对物质财富的追求并非人类的全部动机，人类行为同时受到内在动机和外在动机（Bénabou & Tirole，2003）的驱动，重视道德、情感情绪、社会规范等。众多情感情绪都具有重要的经济意义，这可能是传统经济学在解释世界和进行预测中忽略的最主要部分，也是导致经济学脱离现实和指导实践失败的重要原因（Elster，1998）。②

情感情绪等的重要特征在于它们都是由信念（belief）引起的。在实际决策中（特别在互动决策中），参与者的信念可能直接构成选择行为的依据。为了弥补传统博弈理论对表达各种信念依赖动机的不足，Geanakoplos et al.（1989）（以下用三位作者姓氏首字母简称 GPS）提出“心理博弈”

* 姜树广，山东大学经济研究院博士研究生，邮政编码：250100，电子信箱：jsg123000@sina.com；韦倩，山东大学经济研究院，清华大学社科学院在站博士后，邮政编码：250100，电子信箱：weiqian1979@163.com。感谢国家社科基金一般项目（12BJL014）、教育部人文社科青年基金项目（10YJC790270）、山东大学自主创新基金青年团队项目（IFYT12096）的资助。作者感谢黄少安教授的悉心指导，感谢匿名审稿人的宝贵意见，但文责自负。

① 见韦倩（2010）、陈叶烽等（2012）对相关实验证据的梳理。

② Elster（1998）列举了大量受信念驱动的情感情绪，包括愤怒、怨恨、内疚、羞耻、自豪、后悔、欣喜、失望、得意、恐惧、希望、欢乐、悲伤、嫉妒、惊喜、担心、挫折等等。

的概念，将信念直接纳入效用函数分析。Battigalli & Dufwenberg（2009）（以下简称 BD）推广并扩展了 GPS 的模型，提供了一个可以用来分析信念依赖动机博弈的一般框架。心理博弈论为基于信念的决策分析提供了统一化的工具。但心理博弈并非适宜信念决策的唯一分析框架，其他一些模型从不同角度入手尝试了将信念纳入决策分析，经济学对信念的重视才刚刚起步。[①]将信念纳入决策分析是经济学最重要的课题之一，一方面开辟了对情感情绪决策分析的广阔领域，另一方面对经典的理性人假设可能带来最重要的修正，使经济学得以研究活生生人的选择行为。

余文第二部分是信念与心理博弈的基本理论；第三部分分析信念依赖的基本人类行为动机，主要关注基于意图的互惠、失望与内疚厌恶、自我形象与社会形象问题，以及焦虑与源自预想的效用等；第四部分梳理相关的实证研究；最后对信念与心理博弈进行总结性的评价与应用展望。

二、信念与心理博弈的基本理论

（一）信念与信念依赖动机

信念是一个范围很广的类名。正如英国哲学家罗素所言，这一词语“带有一种本身固有的和不可避免的意义上的模糊不清”。罗素在《人类的知识》（2009，第 179-185 页）中认为，信念是身体上或心理上或者两方面兼有的某一种状态。他列举出五种不同种类的信念：（1）那种以动物性推理补足感觉的信念；（2）记忆；（3）预料；（4）只凭证据不经思考就得出的信念；（5）那种得自有意识推理的信念。在当代哲学中较普遍的将信念描绘为“命题态度”（propositional attitude），即个体认为某种情况属实的心理状态（Schwitzgebel，2006）。

信念在经济学中的用意同样广泛：（1）在微观行为分析中，信念常常是对不确定或不可知事件（物质结果）的估计，或对未来事件、博弈中他人行为与信念的预期或猜测，这两种意义上的信念实质等同于主观概率测度。（2）文化信念是特定群体成员共有的观念或思想，用于支配成员之间或与其他群体的互动交往（Greif，1994），经济学关注文化信念对经济绩效或社会组织的影响。（3）价值或道德信念，人们认为正当合理的，坚信正确的观念认识，这在社会规范的研究中尤为重要。（4）直觉信念，对可能发生结果直观的感觉。直觉（快速而情绪化的）作为区别于慎思（慢速而有克制的）的基本人类决策机制，在人际互动决策中扮演重要角色（Kuo et al.，2009）。

对上述复杂的信念概念进行抽象，根本上可以理解信念为个体进行判断的一种心理状态。当决策动机不仅仅是物质收益最大化，还直接依赖于信念（即持有信念的心理状态本身构成最终效用的一部分）时，我们说个体具有信念依赖的动机（偏好）。[②] 信念依赖动机为人类所普遍持有。信念以目的、动机的形式贯穿于人类活动中，并与情感、意志相结合，形成一种稳固的支配人类行动的心理倾向。持有信念的对错决定人生成败与哀乐，直接影响人们的福利状况。一个人是表现出利他还是充满恶意，取决于对他人的信念（Levine，1998），亲社会行为也是依赖于信念的。另外，个体非常重视自我形象（self-image）与社会形象，在乎他人如何看待自己，注重身份和社会（及自我）认同，关注社会地位等等，这些都明显受到所持信念的影响。在经济决策中，由于个体也会从对未来的期望中获得当前效用，向前看的决策者会扭曲信念，如个体会选择信念以最小化对坏结果的恐惧带来的不快乐（Akerlof & Dickens，1982）。[③] 期望偏差会导致错误的决策和

① Colman（2003）提出了一个不同于 GPS 的心理博弈概念用来分析社会互动合作问题，Camerer （1997）则提出行为博弈论（behavioral game theory）的概念以分析传统博弈不能解决的问题。另外，本文涉及的许多模型也没有直接使用心理博弈论。

② 在心理博弈论的文献中，处理的信念概念一般与传统博弈论并无本质差异，基本可以作为一个概率分布处理，不同之处主要在于信念是否进入效用函数。这里并不试图为信念提供精确的定义，而是将其作为一个开放的概念，重点关注于信念依赖动机。不论是主观概率，还是文化信念，都可能为个体决策的终极动机所依赖。

③ Akerlof & Dickens（1982）是最早的一篇将信念纳入效用函数分析的文献，对后续研究具有重要影响。

非合意结果的出现，如在投资中决策者会过高估计投资的回报而做出非理性的决策（Brunnermeier & Parker，2005）。不论在微观层面还是宏观层面，主观信念对经济决策结果都具有重要影响。

在具有信念依赖动机的博弈中，信念有两个渠道影响参与者行为：一是传统博弈论中所具有的关于对手的信念（或修正的信念）对自身策略偏好的影响；二是信念对终点结偏好直接心理的影响。心理博弈论正是专门发展来分析信念依赖动机的工具。

（二）心理博弈主要理论模型

传统博弈论完全信息下参与者的效用仅依赖于终点结的支付。在 GPS 心理博弈的分析框架下，参与者的效用不仅取决于终点结的支付，还取决于他们持有信念（包括对选择的信念、对信念的信念或对信息的信念）的策略互动的情景。GPS 标准形式的心理博弈 $G=(A_1,\cdots A_n;u_1,\cdots u_n)$ 对参与者集合 $N=\{1,\cdots,n\}$ 的每个参与者 i 包括一个行动集合 A_i 和一个效用函数 u_i：$\overline{B}_i\times\Sigma\to\Re$。$\overline{B}_i$ 代表参与者 i 的全局一致（collectively coherent）信念，是 i 原初信念的无限阶层（infinite hierarchies）空间，$\Sigma=\times_{i\in N}\Sigma_i$，$\Sigma_i=\Delta(A_i)$ 是参与者 i 混合策略的集合。参与者 i 的一阶信念是对其他参与者混合策略的概率估计，其一阶信念集合为 $B_i^1=\Delta(\Sigma_{-i})$。二阶信念是对 $\Sigma_{-i}\times B_{-i}^1$ 的概率测度，涉及到对对手策略和对对手一阶信念的关系。合理的二阶信念应该与一阶的信念相一致，以此类推高阶的信念应与低阶的全部保持一致，理性的决策者也可以预期他人的信念也都是一致的，信念一致性被作为参与者的共同知识。

由于 GPS 的信念只涉及行动前的“原初信念”，而许多重要的信念依赖动机需要随博弈展开进行信念更新，故 GPS 的框架只能模型化策略环境中一些特定信念依赖的动机，而将许多其他合理的信念依赖动机排除在分析之外。在汲取多阶条件信念表达方法的基础上，BD 对 GPS 的模型进行了扩展，意在提供一个更基本的分析框架，而将传统博弈论和 GPS 的心理博弈作为特殊情况包括在内。BD 主要在三个方面对 GPS 的信念概念进行了扩展：（1）允许更新的高阶信念、他人的信念、行动的计划和不完全信息都可以影响动机；（2）解决了随着博弈的展开，参与者对他人信念的信念是如何修正的问题；（3）定义了心理序贯均衡的概念，这一般化了传统博弈论中序贯均衡的概念，并证明了在温和假设下的存在性，并特别允许对博弈进行非均衡分析。

BD 扩展形式的心理博弈（可测度且有界）的心理支付函数形式可表示为：$u_i=Z\times M\times S_{-i}\to\Re$。支付函数中 Z 为终点结的物质支付，S_{-i} 是除 i 外其他参与者的策略集合，M 是所有参与者的全局一致阶层条件信念的集合，支付函数还可以表述为形式 u_i：$Z\times M_i\times\prod_{j\neq i}(M_j\times S_j)\to\Re$，其中 M_j 是 j 关于其他对手策略和条件信念的可能条件信念的集合，S_j 是 j 的策略集合。

Battigalli & Dufwenberg（2011）提供了一个简洁形式信念依赖动机动态博弈的分析思路。假设两个参与者张三（A）和李四（B），周期 *t = 1; 2; …; T*，令 $m_A(z)$ 和 $m_B(z)$ 为博弈的物质支付，$\mu_A^0\in\Delta(M)$ 为张三对结果的原初信念，$\mu_A^t\in\Delta(M)$ 为张三在 t 期末的信念，假设张三对于结果 m

的效用依赖于张三和李四信念的时间序列：$u_A((\mu_A^0,\mu_B^0),...,\mu_A^T,\mu_B^T),m)$。张三的信念$\mu_A^t$会引发其自身对结果预期的情感（可能是正面的或负面的），如焦虑或兴奋，这会影响其t期的效用；另一些t期的情感（如失望）则可能依赖于前期的信念$\mu_A^k(k<t)$；张三也可能在乎李四的情感，如对李四伤害的内疚、羞耻，对李四焦虑的关注等，对这类情感的预期也会影响行为。虽然基于多层阶条件信念的动态博弈理论比较复杂，但通过假设心理效用取决于参与者自己和他人的一阶信念可以对许多类型的信念依赖偏好建立简洁的模型进行分析，尤其对定性的分析是足够的，这也是目前对情感情绪分析中常用的方法。

（三）心理博弈的进一步解释与心理博弈均衡

1. 心理博弈与传统不完全信息博弈的区别

心理博弈论与传统博弈论最根本的区别在于参与者的最终支付不仅取决于每个人怎么做，还取决于每个人怎么想。容易混淆的是传统不完全信息动态博弈的支付也与阶层的信念有关，但两种情况下信念的根本意义是不同的。正如开篇所述，传统博弈论的基本假定前提是理性自利人，不完全信息博弈的参与人并没有心理效用。所以传统博弈论在完全信息的情况下并不涉及信念问题，不完全信息情况下，参与者的效用函数是类型（如参与者的能力或品味）依存的，即参与者所拥有的私人信息$\theta\in\Theta$进入效用函数，参与者有关于θ的信念，这个信念是一个概率分布，为了分析均衡，通过海萨尼转换引入自然。由于参与者类型是自然选择的外生参数，对参与者类型的阶层信念也就是外生的。动态博弈中信念体系通过贝叶斯法则从策略组合中导出，参与者随博弈展开修正后验概率，均衡时要求信念与均衡策略相容。不完全信息动态博弈中概率的修正或说信念的更新，甚至博弈中改变对手的固有信念，但最终必然以某种方式依附于物质支付，信念没有直接作为效用的一部分。这样传统博弈论就难以分析大量信念依赖的心理动机如惊喜、自信、失望等。心理博弈中，终点结的支付依赖于对策略的信念以及对此信念的信念，等等类推的多阶层信念，进入效用函数的参与者信念是内生的变量，信念问题在完全信息的情况下同样存在。

2. 心理博弈的均衡解

心理博弈的均衡概念与传统博弈论类似。GPS 和 BD 实际上一般化了传统博弈的均衡概念以考虑信念进入参与者效用函数的情形。这样在处理均衡行为时，就产生两个额外需考虑的问题（Attanasi & Nagel，2007）：（1）“信念正确”的条件要求“表述的效用正确”。给定分析中信念的不同顺序，明确强制它们在均衡时正确。在均衡中，参与者根据对行动持有的正确信念以及对对手信念的测度，最大化他们的总效用。同时，不同顺序的信念应符合参与者对总效用形式正确计算的最佳反应。（2）随着博弈展开，关于他人信念的信念会修正，参与者的信念更新也会导致效用的更新。在一个心理博弈中，为了决定最优的行动步骤，参与者可能需要形成对其他参与者信念的无限阶层的信念。

GPS 提出了一个心理纳什均衡和子博弈精炼纳什均衡的概念，但只允许先于行动的原初信念进入参与者的效用函数。BD 提出了心理序贯均衡的概念来处理信念更新问题。在传统不完全信息动态博弈中均衡的一个适当定义必须涉及到状态（assessments），即行动策略和条件（一阶）信念的组合，一个状态当是一致性的且满足序贯理性的时候是一个序贯均衡。BD 在序贯均衡概念的基础上加入第三个要求以包含高阶的信念，需要参与者在每个阶段持有共同的、正确的对彼此信念的信念。

在心理博弈中，由于正确判断其他参与者的心理倾向非常困难，所以完全信息的条件比传统博弈更难满足。在不完全信息情况，用$\theta=(\theta_A,\theta_B)$概括表示并非共同知识的与博弈所有支付相

关的参数向量。$\theta_i(i=A,B)$ 是仅归参与者 i 知道的部分（如其自身对某种心理动机的敏感度）。θ 属于一个参数空间 $\Theta=\Theta_A\times\Theta_B$ 是共同知识。Θ 的元素称为自然状态。在动态心理博弈中，简单起见假设参与者并没有随博弈展开得到更多关于自然状态的信息，而只观察到博弈先前阶段选择的行动，这就容易一般化信念空间的结构以包括对自然状态的信念。如 B 的互惠或内疚的敏感性参数 A 都不知道，A 可以推测 θ_B 来自一个混合分布，对其有一个先在的判断。这样就可以类似传统博弈论处理不完全信息的方法处理心理博弈的不完全信息问题。另外，BD 特别强调心理博弈非均衡分析的重要性。

三、信念依赖的基本人类行为动机

理解人类根本的行为动力机制，不仅有助于解释普遍的亲社会行为来源，[①]还有益于认识人类真实决策的内在机制，从而为具体经济问题提供合理解释和正确的策略指导。心理博弈论为将心理的、社会的、文化的多种因素纳入正式的决策分析框架提供了有用的工具。另一些模型虽然没有明确使用心理博弈论，但也以信念构成心理效用来分析，与心理博弈论异曲同工。

（一）基于意图的互惠动机

Rabin（1993）基于意图的互惠是心理博弈论最早最著名的应用。此模型中博弈参与者会对友善（不友善）的行为友善（不友善），这里的关键概念“善意”（kindness）依赖于信念。Rabin（1993）双人博弈互惠模型中参与者 i 的效用函数形式为：

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

右式第一项 $\pi_i(a_i,b_j)$ 是 i 最终获得的物质效用，第二项为互惠情感带来的心理效用，$f_i(a_i,b_j)$ 是参与者 i 对 j 的善意程度；$\tilde{f}_j(b_j,c_i)$ 则刻画了 i 对 j 对自身善意的感知。a_i 为 i 采取的策略，b_j 是参与者 i 对 j 采取策略的信念，c_i 则是 i 关于 j 对 i 策略信念的信念（二阶信念）。

Dufwenberg & Kirchsteiger（2004）则发展了扩展形式博弈的互惠理论。在动态情况下，信念依赖于参与者在不同历史得到的新信息，而不仅是原初的信息。这样参与者的策略选择在博弈的所有阶段都是最优反应，并在 Rabin 两人博弈基础上考察了多人博弈的情形，并提出了一个“序贯互惠均衡”的解概念。Falk & Fischbacher（2006）构造的互惠模型同样使用了心理博弈论的方法来刻画善意函数。另外，Segal & Sobel（2007）等也分别建立了不同的基于信念依赖的互惠模型。对互惠行为的分析是心理博弈论最活跃的研究领域，以下借用 Attanasi & Nagel（2007）以信任游戏来具体分析互惠决策的心理博弈情境。

考虑如图 1a 的信任游戏：参与者 A（委托人）和 B（代理人）作为搭档获得 2 单位的总收入。参与者 A 先决定是否信任 B 而继续合作，如果 A 不信任 B 而结束游戏，按合约两人平分收益。如果 A 继续总收益会翻倍到 4 单位，之后由 B 决定与 A 分享或是独吞收益。按传统博弈论我们知道唯一的子博弈精炼纳什均衡是 A 选择不信任，如果 A 继续 B 选择独吞。

① 解释亲社会行为的社会偏好理论引发了经济学界的广泛关注，但着重在“利他”或“公平”等稳定偏好，笔者认为“公平偏好”等对亲社会行为的解释只是冰山一角，下文的内疚厌恶等模型同样可用于解释亲社会行为的实验证据，公平或利他等偏好本身也是依赖于信念的，信念是更为根本的动机来源。

现在将互惠的心理动机纳入，仍然假设 A 是自利者，则 A 的总效用函数简化为物质支付，无心理效用部分。假设 B 是互惠者，B 的总期望效用函数为：

$$u_B((\alpha_B, s_B); \beta_B) = \pi_B(\alpha_B, s_B) + \theta_B^r \cdot E_B[K_A; \beta_B] \cdot \pi_A(\alpha_B, s_B).$$

右式第一部分为 B 总效用的物质部分，第二部分为 B 总效用的心理部分。$\theta_B^r \geq 0$ 是互惠敏感程度的测度，$E_B[K_A; \beta_B]$ 为 B 对 A 的善意感知，π_A 为 A 的物质支付函数。α_B 是 B 认为 A 会选择继续的初始一阶信念，条件二阶信念 $\beta_B = E_B[\alpha_A \mid \text{继续}]$ 度量 B 知道 A 已经继续时，B 关于 A 对他信任的预期，其中 α_A 为 A 选择继续后相信 B 会分享的初始一阶信念。K_A 为 A 对 B 的善意。善意和感知的善意取决于信念。A 的善意取决于 A 的一阶信念，而 B 的善意感知取决于 B 的二阶信念（B 关于 A 信念的信念）。当 A 选继续，他越少考虑 B 会在继续后分享，他就越善意，在 A 选结束的情况下可以直观的认为他对 B 不友善。因此，如果 α_A 值较低继续可能被认为“善意”。 当 β_B 值较低继续被 B 感知为善意。如果 β_B 低并 B 是有强烈互惠动机考虑的，他会回报 A 而选择分享。

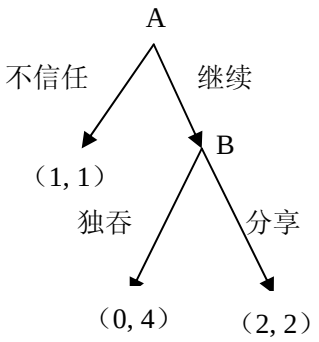

图 1a 完全自利的信任游戏

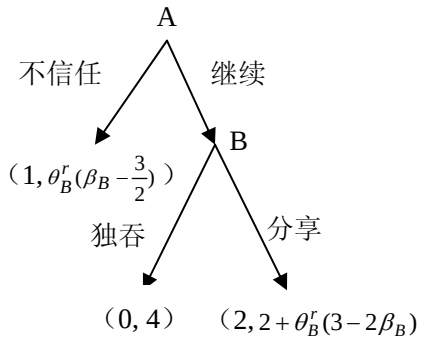

图 1b 带互惠的信任游戏

A 对 B 的“善意”为 A 给予 B 的期望支付和公平支付 $\pi_B^{e_A}$ 的差值：

$$K_A(s_A, \alpha_A) = E_A[\pi_B; s_A, \alpha_A] - \pi_B^{e_A}(\alpha_A).$$

其中公平支付在这里可取给定 A 策略下 B 的最大与最小期望支付的均值。具体地，当选择继续时 A 的善意为：

$$K_A(\text{继续}, \alpha_A) = 2\alpha_A + 4(1-\alpha_A) - \frac{1}{2}(2\alpha_A + 4(1-\alpha_A) + 1) = \frac{3}{2} - \alpha_A.$$

因此当 A 选择继续时 B 感知的善意为：$E_B[\text{继续}, \beta_B] = \frac{3}{2} - \beta_B$，由此可计算出各终点结上 B 的效用水平如图 1b。该图表达了自利的委托人和具有互惠情感的代理人信任游戏的心理博弈，出现在终点结的不是物质支付，而是总效用水平。对比图 1a 和 1b，具有互惠情感代理人的出现改变了整个博弈的结构，使合作解更容易实现。

（二）失望与内疚厌恶

失望厌恶模型建立的思想是个体对于他们期望获得结果的背离是敏感的，当所得低于期望水

平时会有一个心理的损失，而高于期望时会得到快乐。失望厌恶者i的一个简单效用函数可表示为：$u_i = m_i - \theta_i Max\{E_{\mu_i^0}[\tilde{m}_i] - m_i, 0\}$，$m_i$为$i$最终的物质支付，对$m_i$的期望依赖于初始的信念$\mu_i^0$，$\theta_i$为对失望的敏感程度。期望结果$E_{\mu_i^0}[\tilde{m}_i]$与实现结果$m_i$的差值构成个体心理效用的变量。这个简单的模型中没有考虑损失厌恶，实际个体在期望支付处是损失厌恶的，即相对期望的损失造成的痛苦远高于同等量的收益获得的快乐。个体在行为决策时会对未来的损失与收益作出预期，对失望的厌恶和对惊喜的期待会影响决策，期望的参照点水平也是取决于信念的（Koszegi & Rabin，2006）。[①] Gill & Prowse（2012）在一个真实努力竞争性的实验环境研究个体的失望厌恶行为，沿着Koszegi & Rabin（2006）期望依赖参照点模型的思路，使用动态心理博弈工具为失望厌恶者建立模型。实验中发现失望导致显著的气馁效应（Discouragement effect）而影响真实努力，显示这种情绪在经济管理中具有重要意义。

与失望密切相连的一种人类情绪是内疚（愧疚），当人们对他人造成伤害时自身会感到内疚。伤害他人的一个普遍形式是使他人失望，如果一个人认为自己使他人失望时会遭受心理效用损失，我们称其为内疚厌恶者。Battigalli & Dufwenberg（2007）使用心理博弈论提出了一个内疚厌恶的一般理论并给出其序贯均衡解。

仍使用图 1a 的游戏框架，假设委托人 A 仅受自利动机驱动，B 为内疚厌恶者，B 会考虑 A 的失望情绪（当 A 在继续后获得的物质支付与其期望水平不符时）。A 选择继续后相信 B 会分享的初始一阶信念为α_A，α_A度量了游戏开始时 A 对 B 的信任。定义 B 当 A 继续时会分享的条件二阶信念为$\beta_B = E_B[\alpha_A \mid 继续]$。因此$\beta_B$度量了 B 知道 A 已经选择了继续下，B 对 A 对自身信任的预期。这样 A 在继续后的期望物质收益为$2\cdot\alpha_A + 0\cdot(1-\alpha_A) = 2\alpha_A$，当（继续，独吞）发生时 A 失望的确切值是$-2\alpha_A$，即两种情况的支付差：（继续，分享）-（继续，独吞）。

B 的内疚是由他对 A 失望的预期给定的，当 A 已经选择了继续就是对$-2\alpha_A$的预期。B 在 A 继续而自己独吞时确切的心理效用是$-2\beta_B$乘以 B 的内疚敏感程度$\theta_B^g \geq 0$。B 的总效用就是$4 - \theta_B^g 2\beta_B$，即总物质效用和心理效用的和。具有自利委托人和内疚厌恶代理人信任游戏的心理博弈可表示如图 2。

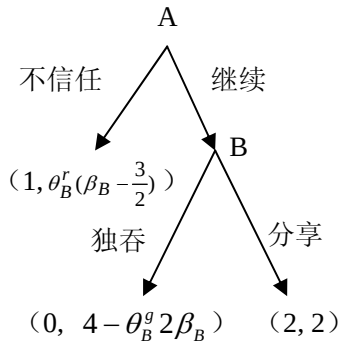

图 2 具有内疚厌恶的信任游戏

① Koszegi & Rabin（2006）提出的参照点依赖偏好（reference-dependent preferences）中代理人的总效用函数具有形式$u(c \mid r) = m(c) + n(c \mid r)$，$m(c)$是消费效用，$n(c \mid r)$是依赖于参照点 r 的得失（gain-loss）效用，参照点由概率信念形成，虽然没有明确提到“失望”，但考察的情感是类似的。

与互惠情感类似，内疚厌恶情绪也会导致博弈结构的改变。为了不让他人失望，个体更倾向在社会合作中表现出信任和合作的行为，有助解释了实验中对自利行为的背离。

（三）自我形象与社会形象问题

假设一个人具有的能力或性格是一个并非所有人知道的参数 θ_i，i 的感觉受到自身对 θ_i 信念的影响（自我形象或自信问题），并且可能会在乎别人对自身的看法，即别人对 θ_i 的信念（社会形象问题）。个人可以直接从对自我的正面评价中获得效用，Koszegi（2006a）称为自我效用（ego utility），认为自我效用直接依赖于信念并为之建立了模型。自我效用与个人的自信问题密切相关，由于自信带给个体效用，个体会有自我形象保护（self-image protection）的动机和自我形象提升（self-image enhancement）的动机（Koszegi，2006a），并有过度自信的倾向而可能引起决策偏差。

受 Koszegi（2006a）模型的启发，Rabin（2010）提出了一个简洁的自我效用模型。假设有两种类型人，擅长制小工具者和不擅长者，张三的效用函数是 $u = w - e - \varphi\sqrt{p}$，参数 $\varphi \geq 0$，w 是他的收入，e 是他工作努力的成本，p 是认为自己擅长制工具的概率。当 $\varphi = 0$ 时就是普通的效用表达，当 $\varphi \geq 0$ 即认为张三有自我效用。假设张三有一个关于擅长制工具原初的信念 $\bar{p} \in [0,1]$，当他工作时 $e = k > 0$ 就会确切发现是否擅长，如果擅长则 $w = 1$，不擅长则 $w = \alpha < 1$。如果不工作则 $e = 0$，$w = 0$，也不能得到类型的信息。这样工作的效用就是 $u = \bar{p}[1 + \varphi\sqrt{1}] + (1 - \bar{p})[\alpha + \varphi\sqrt{0}] - k$，不工作的效用是 $u = \varphi\sqrt{\bar{p}}$。张三是否工作取决于两种效用的权衡，$\varphi$ 和 p 值的大小对决策具有决定影响，而不再是简单取决于成本与收益的权衡。

人们对他人看待自己的在乎也会对行为有重要影响。利他或公平等亲社会行为可能并非出于本能，而是出于别人如何看待自己的考虑。Bernheim（1994）分析了人们对社会规范的遵从以维护自己身份地位的内在动机，作者虽然没有明确使用心理博弈论，但实质探查的是一种信念依赖的动机。Benabou & Tirole（2006）考察了人们对社会声誉或自尊的关注，认为在委托代理关系中物质的激励可能会破坏来自尊重的激励。行为代理理论（behavioral agency theory）认为企业人力资源管理中需同时兼顾物质激励和对雇员的尊重，员工对于其他同事和老板对自身的看法非常在乎（Ellingsen & Johannesson，2007）。

Ellingsen & Johannesson（2008）建立的社会尊重模型的核心假设是参与者会对其他人怎么看自己有骄傲感，他们会喜欢被别人认为是一个利他的人而不是自私的人，模型中不论是尊重还是被尊重的感觉都取决于对参与者类型的信念。Andreoni & Bernheim（2009）则假设人们喜欢被他人认为是公平的，并将这种内在需求加入效用函数中以解释普遍存在的五五分派的社会规范。Tadelis（2011）认为人们在乎他人看法而具有羞耻厌恶的动机，这对于亲社会行为意义重大。

（四）焦虑与源自预想的效用（utility from anticipation）

人们会对未来可能发生的好事情充满期待，对坏事情担忧恐惧。这种预期的情感对个体福利的影响常常超过事实本身，如对于一次即将到来的长途旅行，可能会兴奋很多天，而旅行真正开始的时候反而没有那么愉快。这种包含个体对未来信念的情绪特别对涉及跨期决策的情形具有重要意义，如Loewenstein（1987）证实为避免焦虑许多个体会选择一个立刻执行的电击而不是一天后执行的电击。Caplin & Leahy（2001）建立了包括源自预想效用的模型，特别关注对不确定事实的焦虑情绪和在决策中降低焦虑的动机。在一个两期彩票选择的背景下，个体在第一期末的心理收益不仅取决于获得的物质收益，还取决于对第二期物质收益不确定的担忧。令 $y_1 = (z_1, l_2)$ 为构

成第一期心理收益的结果，z_1代表第一期物质收益结果，l_2是关于第二期收益的概率分布，ϕ是一种面对y_1时的心理状态，则两期选择的参与者在第一期末的效用函数为$V_1(y_1)=u_1(\phi(y_1))+E_{l_2}[u_2(x_2)]$。第二期末因为没有不确定性也就没有预想的心理，物质收益等同于心理收益即$z_2=x_2$（x代表心理状态）。Koszegi（2006b）以类似的思想建模，假设委托人的效用是物质结果和信念依赖情绪的函数，源自预想的效用和物质效用分别占一定比例的权重。在两期决策下，预想效用等于在第 1 期信念的条件下对第 2 期的期望物质效用，这里信念被内生的决定。①

个体会因关注他人的焦虑等情绪而影响决策。医生想提供可靠的医嘱，但是又不想吓坏病人，公司高管想做出明智的决策但是又不想破坏士气，政府要为国家面临的挑战做出准备，但又不能引起民众的恐慌和焦虑，军官、老师、父母、朋友、爱人等的行为都需要考虑士兵、学生、子女等对应者的情绪反应（Koszegi，2006b）。在 Caplin & Leahy（2001）的基础上，Caplin & Leahy（2004）建立的医生与病人博弈模型假设医生完全的富有同情心，医生的效用依赖于其关于病人福利状况的信念。假定有一个心理的生产函数，效用相关的心理状态如悬疑和焦虑受到展开的信念和实际产出的影响。信息通过影响信念而影响效用，进而影响心理状态。

其他基于信念依赖动机的情感情绪研究如同情（Sally，2001）、愤怒（Smith, 2009）、羞耻（Tadelis，2011）等正在兴起，这只是心理博弈论对人际互动交往中的情感情绪与认知可应用前景的一小部分，仍有广阔的空间需要探讨。

四、信念依赖动机的实验证据

（一）信念与行动的关系及实验中的信念引出

实验室实验是目前研究信念依赖动机的主要方法，通常在实验中设置一定的信念引出（elicitation）机制以得到参与者信念的数据。Dufwenberg & Gneezy（2000）较早使用信念引出在简单的信任实验中研究了信任回应（信任的自我实现性质）的相关性，实验中委托人可以选择拿走 0 到 20 的荷兰盾，或留下来允许代理人对 20 荷兰盾分配。在委托人和代理人同时做出选择后，有一个针对委托人（一阶信念）和代理人（二阶信念）的信念引出。通过信念数据与实际策略行为的分析发现对委托人投入有更高猜测的代理人更倾向实现信任。这为使用信念引出方法研究情感情绪决策开辟了重要方向。

如何最好地引出信念是一个棘手而重要的问题，是使用声明的（stated）信念，还是推测（inferred）的信念可能需要结合不同的实验需要（Rutström & Wilcox，2009）。按照显示偏好原理一定程度上可以通过有成本的行动来推测信念，另外对信念声明本身加入激励，使信念引出程序满足激励相容也有助于真实信念的表达，如Nyarko & Schotter（2002）等主要文献的信念引出都是通过在实验室中为参与者提供恰当的激励来得到真实的一阶信念，Gächter & Renner（2010）以直接证据表明加入激励会减少信念数据中的“噪声”量。目前最流行的信念引出程序二次得分法则（Quadratic-Scoring Rule）也是通过恰当激励的方式以引出真实信念（见Gneiting & Raftery，2007 对得分法则的综述）。②

激励有助于参与者声明真实信念，但过强的激励可能会改变整个游戏的激励结构，这涉及到

① Koszegi（2006b）模型不同于 Caplin & Leahy（2001）的特点是将物质效用和源自预想的心理效用进行了完全的分离。

② 其他的信念引出方法涉及自由法则（free rule）、简单排序法则（simple ordinal rules）、彩票法则（lottery rule）等，各种引出方法优劣的对比与问题争论可参见 Palfrey & Wang（2009）等。

信念引出程序是否会改变参与者策略行为的问题。信念引出可靠的一个前提假设是认为参与者形成信念是作为制定决策的一部分，因此让他们声明信念不应该影响行为。这部分实验证据是有分歧的。Guerra & Zizzo（2004）使用两个简单的信任游戏直接或间接地测度信任回应的稳健性。他们使用了三个实验处置：不引出信念；信念只引出不传递；信念引出并从委托人传递到代理人。信任和实现率在所有三个处置中都很类似，可以认为信任回应对任何种类的信念操纵（belief manipulation）都是显著的，因此认为信念引出是中性的。但是Rutström & Wilcox（2009）等在实验中发现引出信念会改变参与者的博弈行为。这需要研究者在实验中进一步深入检查和考虑引出程序对行为可能的影响。[①]

（二）信念依赖动机的实验证据

Rabin（1993）等基于意图的动机在理论上有助解释大量经济学实验中观测到的对自利最大化行为的背离，但是较难提供明确的实证。[②] 实验中由于参与者对意图的不同理解，导致相同结果的行动可能引起完全不同的互惠反应，一个核心的问题是个体如何评价特定行动的善意程度。一些实验是通过对比有意选择的行动与可选的替代行动来评价意图的，如在控制组先行动者可以在一组可选择的行动集中有意的选择如何行动，而在实验处置组选择是被随机决定的，在这种方法下， Falk et al.（2008）发现公平意图的归因同时对积极互惠和消极（negative）互惠行为具有显著地影响。Stanca et al.（2009）则通过在一个对称的礼物交换游戏中操纵先行动者对后行动者策略空间的信念来考察动机的本源在意图与行动的相关性验证中的作用。对给定的分配结果，基于策略动机的行动会被感受到较低的善意（A是为了B的回报而友善，相对于A的纯粹利他动机，B则感受到A较少的善意）。实验结果表明当先行动者的策略动机可被剔除时，后行动者反应出强烈的积极互惠行为，选择行为背后动机的类型对互惠行为有显著的影响。

Charness & Dufwenberg（2006）以证据表明人们努力迎合他人的预期以避免内疚，代理人认为委托人期望的水平越高，代理人实际的付出水平就越高。在信任游戏中，对内疚厌恶的敏感程度会引发代理人（B）的特定行为，实验证据显示 B 分享的二阶信念与信任的实现正相关，B 在继续后选择分享的比选择独吞的有更高的平均期望值。Vanberg（2008）通过对人们信守承诺行为的研究，认为迎合他人预期的说法不能解释承诺行为，人具有信守承诺本身的偏好，但是两种解释都可以在内疚情感中找到根据，打破承诺本身就会使人感到内疚，内疚情绪是由对于对契约或道德义务规定了的“应该做”而没有做出一致行为引发的

以上实验虽证实了内疚厌恶的存在性，但是很少提供内疚厌恶重要性的定量信息。Ellingsen et al.（2010）通过三个独立的实验（独裁者实验、完全信息的信任实验和隐藏行动的信任实验）分别测度内疚厌恶，发现实验中的慷慨行为与引出信念的相关性接近于零，所以认为内疚厌恶可能没想的那么重要。Bellemare et al.（2011）使用类似的外生引出二阶信念的方法对荷兰人口大样本做了一个简单的提议——回应实验。研究者通过估计内疚厌恶的结构模型来测度人群为避免令他人失望而内疚的意愿支付水平，估计结果显示显著的意愿支付来避免内疚，回应者平均愿意支付 0.4 到 0.6 欧元来避免提议者 1 欧元的失望。不同研究者相反的实验结论可能来自实验设置背景的不同或测量方法的差异，对内疚厌恶情绪的定量研究仍有待进一步的深入拓展。

Attanasi et al.（2011）构造了一个简单的机制来同时引出和传递心理博弈中参与者对内疚和互惠两种情绪的敏感性。研究者通过结构化的问卷来引出代理人（B）的信念并揭露（或不揭露）

① 信念引出对于直接验证信念对决策的影响必不可少，但是一定程度上信念依赖的效用也可以不用信念的数据来验证，如 Charness & Dufwenberg（2011）和 Cardella（2012）等。

② 如独裁者实验、最后通牒实验、信任实验、礼物交换实验和公共物品实验等类似实验的发现与传统理论预测相悖，基于分配依赖偏好的模型对此提出了部分解释，如 Fehr & Schmidt（1999）和 Bolton & Ockenfels（2000）认为个体不仅在乎自身的物质收益，还在乎最终收益与其他人的比较，这类模型只关注最终的分配结果。

给其搭档，这样内疚敏感性参数和互惠敏感性参数都可被作为参与者中的公共信息，这就可在几乎完全信息处置的情境和不完全信息处置的情境间比较参与者的行为。结果显示引出 B 的信念依赖偏好而不传递给 A 并不改变参与者的行为，这可被看做一个间接的信念依赖动机存在的证据（研究者并没有引起他们，而是仅仅把已经存在的东西引导出来）。但是，引出并传递 B 的信念依赖偏好并允许参与者进行几乎完全信息的一阶段信任心理博弈时参与者的行为明显不同（与对应的不完全信息心理博弈的设置相比）。代理人 B 的信念依赖偏好的信息导致委托人 A 感受到代理人的信念依赖动机（如内疚和互惠），这最终导致双方参与者更加合作的行为。

神经成像方法也直接用于信念依赖动机相关的研究。Bhatt & Camerer（2005）使用功能磁共振成像（fMRI）扫描了参与者在制定决策、表达一阶及二阶信念时的大脑活动。研究发现信念的表达和信念形成与特定区域的大皮层活动相关，特别的发现二阶信念的形成会激活前脑岛活动，这一过程似乎是一个做决策与形成信念的混合体，这与传统博弈论的认识明显不同。van den Bos et al.（2009）发现基于意图的互惠与大脑的颞部顶骨连接部位（TPJ）和内侧前额叶皮质（MPFC）区域的活动有关。Chang et al.（2011）使用 fMRI 研究了内疚厌恶情绪的神经元基础，发现当参与者最小化内疚，即当行为与他们对对手期望的信念一致时，脑岛、辅助运动区（SMA）、脑回前侧背面（DACC）、背外侧前额叶皮质（DLPFC）及颞部顶骨连接部位（TPJ）的活跃度增强，而当参与者返还额低于他们对对手的期望信念时，伴随腹内侧前额叶皮质（VMPFC）、双边伏隔核（NAcc）及背内侧前额叶皮层（DMPFC）的较高活跃度。对内疚敏感性系数的估计表明增强的内疚敏感性与增强的脑岛和 SMA 活跃度正相关，而与 NAcc 的活跃度负相关。这些研究为情感情绪决策提供了更为科学的证据，与理论模型和实验证据相辅相成，将共同致力于破解人类决策的黑箱。

另外，焦虑、失望、后悔等情绪的研究正在兴起，越来越多的实验设计来对各类情绪决策机制进行验证，与信念依赖动机直接相关的主要实验文献另见 Dana et al.（2007）、Tadelis（2011）和 Gill & Prowse（2012）等。

（三）影响信念与行为机制的实验考察

既然信念对行为具有决定作用，是否可以通过改变信念的机制对行为产生积极影响，从而促进信任与合作，这对与组织管理者和政策制定与实践者来说意义重大。

容易想到的是语言交流改变信念，很多研究者观察到面对面的交流可以极大的提高合作。问题是交流是否以改变信念的方式对行为产生了影响，要对此作出检验还需对其他因素严格控制。在用于检验内疚厌恶的信任实验中，Charness & Dufwenberg（2006）将行动前的交流作为一种情感引出和传递的技术，考察了无约束行动前闲聊对参与者信任和合作的影响。结果显示不同交流形式的效率水平差异明显，意图的表达对改变人们行为的感知似乎很有影响。在收集了参与者的策略决定后，研究者引出委托人的一阶信念和代理人的二阶信念。实验只允许内容简单的单方信息传递，以保证清楚的、交流功能受控的检测，并通过控制信息内容以研究是否一些特定类型的信息影响决策。结果支持交流和承诺通过导致二阶信念的改变影响了参与者的行为。Charness & Dufwenberg（2006）考察的是隐藏行动的情境下的交流作用，Charness & Dufwenberg（2011）则进一步研究了隐藏信息情况下交流是否可以促进合作，并依据“撒谎的成本”（cost-of-lying）和“责备的内疚”（guilt-from-blame）两种直接与心理博弈相关的行为模型来检验数据。承诺机制在这过程起到关键的作用，承诺起到帮助抬高期望并增加信号可靠性的作用。在一个以视频方式进行事先交流 3 人参与的独裁者实验中，Greiner et al.（2012）允许接受者对独裁者的单边视频交流，视频信息成为影响独裁者（关于接受者的期望）的信念的工具，从而影响了独裁者的分配决策。

决策框架（framing）影响选择行为的结果已被心理学家和经济学家广为接受。Dufwenberg et

al.（2011）认为框架是以通过影响信念来影响选择和行动的，如果个体是情绪化的或在乎他人的意图与需要，背景框架可以为决策者提供关于他人信念的线索，进而影响关于他人信念的信念。在同时考察互惠和内疚两种情感的公共物品实验中，研究者通过设置情境框架并引出参与者一阶和二阶信念来验证对参与者贡献的影响。在价框架（valence framing）方式中，参与者在“给”框架处置下的一阶与二阶信念及贡献都显著的比“拿”框架处置下高，这是因为框架为参与者提供了一个初始的聚点，参与者聚焦于明确指引他们去做的行为；在标签框架（label framing）方式中，“中性”的框架处置下的参与者一阶与二阶信念和贡献要高于“社区”（community）的框架处置，解释为参与者的社区环境为其提供了非合作的日常意识。社会规范、法律、合同和承诺等的暗示可能会转变个体关于他人行为和信念的信念，进而改变内在动机和行为，这对于理解社会环境如何塑造人们的行为提供了深入的理解。

五、信念与心理博弈的总结性评价与展望

围绕主观信念的探讨久存于各门社会科学与每个人的日常生活之中，经济学严格的考察信念依赖动机却是最近的事情，但已显示出旺盛的生命力，心理博弈论为分析信念依赖动机提供了基本工具。将信念纳入决策分析具有广阔的应用前景，预期至少在以下方面会产生重要影响：

一是引发情感情绪决策研究的重视。情感与理智是人类行为的两大原动力。传统的理性选择模型实际上不考虑决策者的情感情绪，与其说是描述个体如何决策的，不如说是教给个体如何决策的。将情感情绪等因素纳入决策框架有助于更准确地刻画人类行为，使经济学更趋向科学化。由于个体的情感情绪状态会影响其行为表现，对从微观的企业管理和激励制度设计到宏观的经济政策与公共管理都非常重要，情感情绪决策的现实应用研究也有待拓展。

二是深化对信任问题的研究。信任是最重要的一项社会资本，是形成和维持合作的基础。众多信念依赖动机的模型和实验都是围绕信任游戏展开的，为信任和值得信任行为提供了丰富的解释。人类的许多情感情绪（如内疚厌恶）对于维持信任行为都具有积极意义，通过一定的机制如交流（Charness & Dufwenberg，2006，2011）或对他人情感的开发（Cardella，2012）等来引致情感以提升信任水平的研究已引起大量学者的兴趣。

三是为制度研究提供新的视角。本质上，文化与社会规范是作为一定群体成员共享的信念体系（Greif，1994）。群体成员的心理博弈决定了文化与社会规范的演化与均衡，心理博弈论工具的恰当使用可能为这类问题研究开辟新的局面。Huang & Wu（1994）使用心理博弈分析了在考虑个体的情绪下法律、人情、社会规范、组织文化等对于控制腐败和维持社会秩序的不同角色。心理博弈也为腐败问题提供了深刻见解，在官僚与公众的心理博弈中，官僚的腐败与其对公众对自身期待的信念相关（Balafoutas，2011）。一个对腐败具有极高容忍度的社会可以催生高腐败，正式制度如何安排引导非正式制度达到好的自我实施均衡是重要的课题。

四是为大量非市场决策行为提供研究视角。非市场行为占据了人类活动的大部分时间，人类在非市场行为中的情感情绪体验是构成个人幸福感的主要成分。如Ruffle（1999）使用心理博弈论研究了赠送礼物行为。另外，人际关系的互动决策（如朋友、伴侣的选择问题等）对人类意义重大，其间涉及复杂的情感情绪与意图互动绝非单纯的理性最大化模型可以解释，信念依赖动机的分析有望为关系产品决策分析的兴起奠定基础。[①]

五是为多学科融合提供基本分析工具。近年来许多学者都探讨了社会科学或人类行为科学的

① 关系产品的重要意义参见 Bartolini et al.（2013）。

统一问题（Gintis，2007；韦倩，2010）。信念依赖动机几乎包含了人类行为的全部动力机制，在这一框架下，个体不仅只在乎物质利益，还在乎道德、情绪情感、身份地位、他人看法等等，从而为哲学、心理学、社会学等学科对人类动机的分析提供统一的基础。信念是比偏好更根本的动机源泉，心理博弈论则为广泛的人类行为提供了分析工具。

信念是一个复杂的概念，本文重点关注了以信念依赖动机为基础的心理博弈对情感情绪问题的分析。心理博弈的决策基础还没有完全弄明白，这需要结合认知科学，对情感、认知与行为的关系做深入的研究，不断打开人类决策的黑箱。进一步的，主观信念在多个层面与经济社会问题交织在一起。信念与习俗、社会规范、文化、制度等的关系，信念与社会资本、社会认同等的关系都有待深入探察。

参考文献：

陈叶烽、叶航、汪丁丁，2012：《超越经济人的社会偏好理论:一个基于实验经济学的综述》，《南开经济研究》第 1 期。

罗素，2009：《人类的知识——其范围与限度》，中译本，商务印书馆。

韦倩，2010：《纳入公平偏好的经济学研究: 理论与实证》，《经济研究》第 9 期。

Akerlof, G., and W. T. Dickens, 1982, “The Economic Consequences of Cognitive Dissonance”, *American Economic Review*, 72(3): 307–319.

Akerlof, G., and R. E. Kranton, 2000, “Economics and Identity”, *Quarterly Journal of Economics*, 115(3):715–753.

Andreoni, J., 1990, “Impure Altruism and Donations to Public Goods: A Theory of Warm Glow Giving”, *Economic Journal*, 100(401): 464–477.

Andreoni, J., and B. D. Bernheim, 2009, “Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects”, *Econometrica*, 77(5): 1607–1636.

Attanasi, G., P. Battigalli, and R. Nagel, 2011, “Disclosure of Belief-Dependent Preferences in a Trust Game”, Mimeo, Available at: http://didattica.unibocconi.eu/myigier/index.php?IdUte=48808.

Attanasi, G., and R. Nagel, 2007, “A Survey of Psychological Games: Theoretical Findings and Experimental Evidence”, in: A. Innocenti and P. Sbriglia (Eds.), Games, Rationality and Behaviour, Essays on Behavioural Game Theory and Experiments, Palgrave McMillan, Houndmills, 204–232.

Balafoutas, L., 2011, “Public Beliefs and Corruption in A Repeated Psychological Game”, *Journal of Economic Behavior and Organization*, 78: 51–59.

Bartolini, S., E. Bilancini, and M. Pugno, 2013, “Did the Decline in Social Connections Depress Americans’ Happiness?”, *Social Indicators Research*, 110(3): 1033–1059.

Battigalli, P., and M. Dufwenberg, 2007, “Guilt in Games”, *American Economic Review Papers and Proceedings*, 97(2): 170–176.

Battigalli, P., and M. Dufwenberg, 2009, “Dynamic Psychological Games”, *Journal of Economic Theory*, 144(1): 1–35.

Battigalli, P., and M. Dufwenberg, 2011, “Incorporating Belief-Dependent Motivations in Games”, Mimeo, Available at: http://didattica.unibocconi.eu/myigier/index.php?IdUte=48808.

Bellemare, C., A. Sebald, and M. Strobel, 2011, “Measuring The Willingness to Pay to Avoid Guilt: Estimation Using Equilibrium and Stated Belief Models”, *Journal of Applied Econometrics*, 26: 437–453.

Bénabou, R., and J. Tirole, 2006, “Incentives and Prosocial Behavior”, *American Economic Review*, 96(5): 1652–1678.

Bénabou, R., and J. Tirole, 2003, “Intrinsic and Extrinsic Motivation”, *Review of Economic Studies*, 70(3): 489–520.

Bernheim, B. D., 1994, “A Theory of Conformity”, *Journal of Political Economy*, 102(4): 841–877.

Bhatt, M., and C. F. Camerer, 2005, “Self-Referential Thinking and Equilibrium as States of Mind in Games: FMRI Evidence”, *Games and Economic Behavior*, 52: 424–459.

Bolton, G., and A. Ockenfels, 2000, “ERC: A Theory of Equity, Reciprocity, and Competition”, *American Economic Review*, 90(1): 166–193.

Brunnermeier, M. K., and J. A. Parker, 2005, “Optimal Expectations”, *American Economic Review*, 95: 1092–1118.

Camerer, C. F., 1997, “Progress in Behavioral Game Theory”, *Journal of Economic Perspectives*, 11:167–188.

Caplin, A., and J. Leahy, 2001, “Psychological Expected Utility Theory and Anticipatory Feelings”, *Quarterly Journal of Economics*, 116(1):55–79.

Caplin, A., and J. Leahy, 2004, “The Supply of Information by a Concerned Expert”, *Economic Journal*, 114: 487–505.

Cardella, E., 2012, “Exploiting the Guilt Aversion of Others-Do Agents do it and is it Effective?”, Available at SSRN: http://ssrn.com/abstract=2136514.

Chang, L., A. Smith, M. Dufwenberg, and A. Sanfey, 2011, “Triangulating The Neural, Psychological, and Economic Bases of Guilt Aversion”, *Neuron*, 70: 560–572.

Charness, G., and M. Dufwenberg, 2006, “Promises and Partnership”, *Econometrica*, 74(6): 1579–1601.

Charness, G., and M. Dufwenberg, 2011, “Participation”, *American Economic Review*, 101(4): 1211–1237.

Colman, A. M., 2003, “Cooperation, Psychological Game Theory, and Limitations of Rationality in Social Interaction”, *Behavioral and Brain Sciences*, 26(2): 139–153.

Croson, R., 2000, “Thinking Like a Game Theorist: Factors Affecting the Frequency of Equilibrium Play”, *Journal of Economic Behavior and Organization*, 41(3): 299–314.

Dana, J., R. Weber, and J. X. Kuang, 2007, “Exploiting Moral Wiggle Room: Experiments Demonstrating and Illusory Preference for Fairness”, *Economic Theory*, 33(1): 67–80.

Dufwenberg, M, S. Gächter, and H. Hennig-Schmidt, 2011, “The Framing of Games and the Psychology of Play”, *Games and Economic Behavior*, 73: 459–478.

Dufwenberg, M., and U. Gneezy, 2000, “Measuring Beliefs in an Experimental Lost Wallet Game”, *Games and Economic Behavior*, 30(2): 163–182.

Dufwenberg, M., and G. Kirchsteiger, 2004, “A Theory of Sequential Reciprocity”, *Games and Economic Behavior*, 47(1): 268–298.

Ellingsen, T., and M. Johannesson, 2008, “Pride and Prejudice: The Human Side of Incentive Theory,” *American Economic Review*, 98(3): 990–1008.

Ellingsen, T., and M. Johannesson, 2007, “Paying Respect”, *Journal of Economic Perspectives*, 21(4): 135–150.

Ellingsen, T., M. Johannesson, S. TjØtta, and G. Torsvik, 2010, “Testing Guilt Aversion,” *Games and Economic Behavior*, 68(1): 95–107.

Elster, J., 1998, “Emotions and Economic Theory”, *Journal of Economic Literature*, 36(1): 47–74.

Falk, A., E. Fehr, and U. Fischbacher, 2008, “Testing Theories of Fairness: Intentions Matter”, *Games and Economic Behavior*, 62: 287–303.

Falk, A., and U. Fischbacher, 2006, “A Theory of Reciprocity”, *Games and Economic Behavior*, 54: 293–315.

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

Gächter, S., and E. Renner, 2010, “The Effects of (Incentivized) Belief Elicitation in Public Good Experiments”, *Experimental Economics*, 13: 364–377.

Geanakoplos, J., D. Pearce, and E. Stacchetti, 1989, “Psychological Games and Sequential Rationality,” *Games and Economic Behavior*, 1(1): 60–80.

Gill, D., and V. Prowse, 2012, “A Structural Analysis of Disappointment Aversion in a Real Effort Competition”, *American Economic Review*, 102(1): 469–503.

Gintis, H., 2007, “A Framework for the Unification of the Behavioral Sciences”, *Behavioral and Brain Sciences*, 30: 1–61.

Gneiting, T., and A. E. Raftery, 2007, “Stricly Proper Scoring Rules, Prediction, and Estimation”, *Journal of the American Statistical Association*, 102: 359–378.

Greiner, B., W. Güth, and R. Zultan, 2012, “Social Communication and Discrimination: A Video Experiment”, *Experimental Economics*, 15: 398–417.

Greif, A., 1994, “Cultural Beliefs and the Organization of Society: A Historical and Theoretical Reflection on Collectivist and Individualist Societies”, *Journal of Political Economy*, 102(5): 912–950.

Guerra, G., and D. J. Zizzo, 2004, “Trust Responsiveness and Beliefs”, *Journal of Economic Behavior and Organization*, 55: 25–30.

Huang, P., and H. Wu, 1994, “More Order without More Law: A Theory of Social Norms and Organizational Cultures”, *Journal of Law, Economics and Organization*, 10: 390–406.

Koszegi, B., and M. Rabin, 2006, “A Model of Reference-Dependent Preferences”, *Quarterly Journal of Economics*, 121(4): 1133–1165.

Koszegi, B., 2006a, “Ego Utility, Overconfidence, and Task Choice”, *Journal of the European Economic Association*, 4(1): 673–707.

Koszegi, B., 2006b, “Emotional Agency”, *Quarterly Journal of Economics*, 121(1): 121–156.

Kuo, WJ, T. Sjostrom, YP. Chen, YH. Wang, and CY. Huang, 2009, “Intuition and Deliberation: Two Systems for Strategizing in the Brain”, *Science*, 324: 519–522.

Levine, D. K., 1998, “Modeling Altruism and Spitefulness in Experiments”, *Review of Economic Dynamics*, 1(3): 593–622.

Loewenstein, G., 1987, “Anticipation and the Valuation of Delayed Consumption”, *Economic Journal*, 97: 666–684.

Nyarko, Y., and A. Schotter, 2002, “An Experimental Study of Belief Learning Using Elicited Beliefs”, *Econometrica*, 70: 971–1005.

Palfrey, T., and S. Wang, 2009, “On Eliciting Beliefs in Strategic Games”, *Journal of Economic Behavior and Organization*, 71(2): 98–109.

Rabin, M., 1993, “Incorporating Fairness into Game Theory and Economics”, *American Economic Review*, 83(5): 1281-1302.

Rabin, M., 2010, “Beliefs-Based Preferences”, Behavioral-Camp Lectures.

Ruffle, B. J., 1999, “Gift Giving with Emotions”, *Journal of Economic Behavior and Organization*, 39: 399–420.

Rutstrom, E., and N. Wilcox, 2009, “Stated Beliefs versus Inferred Beliefs: A Methodological Inquiry and Experimental Test”, *Games and Economic Behavior*, 67(2): 616-632.

Sally, D., 2001, “On Sympathy and Games”, *Journal of Economic Behavior and Organization*, 44(1): 1–30.

Schwitzgebel, E., 2006, “Belief”, in Zalta, Edward, The Stanford Encyclopedia of Philosophy, Stanford, CA: The Metaphysics Research Lab, http://plato.stanford.edu/entries/belief/, retrieved 2010-11-21.

Segal, U., and J. Sobel, 2007, “Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings”, *Journal of Economic Theory*, 136(1): 197–216.

Smith, A., 2009, “Belief-Dependent Anger in Games”, Working paper, Available at: http://www.hss.caltech.edu/~acs/papers/bdaig.pdf.

Stanca, L., L. Bruni, and L. Corazzini, 2009, “Testing Theories of Reciprocity: Do Motivations Matter?”, *Journal of Economic Behavior and Organization*, 71: 233–245.

Tadelis, S., 2011, “The Power of Shame and the Rationality of Trust”, Mimeo, Haas School of Business, U.C. Berkeley.

Vanberg, C., 2008, “Why Do People Keep Their Promises? An Experimental Test of Two Explanations”, *Econometrica*, 76(6): 1467-1480.

Van den Bos, W., E. van Dijk, M. Westenberg, S. A.R.B. Rombouts, and E. A. Crone, 2009, “What Motivates Repayment? Neural Correlates of Reciprocity in the Trust Game”, *Soc. Cogn. Affect. Neurosci*, 4(3): 294–304.

Belief and Psychological Games: Theory, Evidence and Applications

Jiang Shuguang and Wei Qian

（Center for Economic Research, Shandong University）

Abstract: Humans can be motivated not only by material (monetary) payoffs but also by morals, feelings, emotions or social norms in their daily decision-makings, thus exhibit deviations from the material self-interested maximizing behaviors. Psychological game theory, based on belief-dependent motivations, has led to research fervor on emotional decision-makings. In Psychological games, players’ utilities depend not only on the terminal material payoffs, but also directly depend on their psychological state triggered by their beliefs. This paper systematically reviews researches focused on belief-dependent motivations. After briefly presenting the fundamental framework of belief and psychological game theory, this paper summarizes and evaluates belief-dependent essential human motivations such as intention-based reciprocity, disappointment and guilt aversion, self-image and social image, anxiety and utility from anticipation, etc. We also discuss the main experimental and neuroscience evidences related to this topic. It is expected that as the research of belief and psychological games moves along, researches on emotional decision-makings, trust, institution, and non-market decisions will be greatly influenced, as well as the issue of multi-disciplinary unification.

Key Words: Belief-Dependent Motivations; Psychological Game Theory; Emotions; Belief-Elicitation Experiments

JEL Classification: C72, C91, D01