自我欺骗的认知机制

鞠实儿[1] 赵艺[1] 傅小兰[2]
(1.中山大学逻辑与认知研究所 2.中国科学院心理所)

摘要： 自我欺骗现象与经典信念逻辑之间的偏差导致自我欺骗悖论；因此，建立合理的自我欺骗理论的必要条件是，提出自我欺骗的认知机制的逻辑结构，解释自我欺骗悖论。本文试图采用逻辑分析和心理实验相结合的方法解决自我欺骗悖论，即在逻辑学层面说明矛盾信念间的逻辑关系，进而在心理学层面说明自我欺骗发生的机制。为了实现这一目标，本文首先用逻辑学的术语严格表述自我欺骗悖论；然后，在文献的基础上给出自我欺骗的逻辑结构；最后，根据上述逻辑结构设计并实施两可图认知实验，实验结果分析表明，自我欺骗的心理机制满足上述逻辑结构。

关键词： 认知偏差； 自我欺骗； 自我欺骗悖论； 信念
中图分类号：B81 文献标识码：A

引言

对自我欺骗问题的研究至少可追述到 Moore（1942，1944）和 Wittgenstein (1953)[1]。1960 年，Demos 在文献 *"Lying To Oneself"* 中首次完整提出“自我欺骗”（self-deception）问题。由于该问题涉及到高级思维过程中的认知偏差（cognitive bias），它引起了逻辑学家和心理学家对自我欺骗的认知机制的共同兴趣，研究的问题有：自我欺骗现实存在性问题（Haight 1980，1985，Martin 1985, 1986）；自我欺骗的发生机制问题（Mele 1982, 1987, 1997）；自我欺骗的意向性问题（Fingarette 1969，Davidson 1982, 1998，Audi 1985，Barnden 1996）；经典信念逻辑对于刻划自我欺骗的适用性问题（Rorty 1972，da Casta 1990，鞠实儿 1993，1995）。研究方法主要有两种：（1）理论研究方法，运用心理学或逻辑学概念构造自我欺骗悖论的解释模型。其困难是：认知过程中的基本概念，如信念、证据等，缺乏合理的定义，其结果是导致上述模型对自我欺骗现象缺乏解释力。（2）实验方法，通过设计实验，观察实验结果得出自我欺骗现象发生的充分条件。其困难是：由于缺乏合理的基本假设指导实验设计，实验的效度和信度不足。例如，Sackeim 和 Gur（1978）、Quattrone 和 Tversky（1984）曾经设计实验证明经典自我欺骗的事实存在，但受到 Mele（1987，1997）的批评[2]。本文将采用逻辑分析和心理实验相结合的研究方法，立足于基本概念分析，分别从理论和事实层面说明自我欺骗的逻辑结构和心理机制，以克服上述研究方法的困难。

1 问题与研究方法

自我欺骗的经典定义与人际间欺骗的定义具有相同的形式结构。人际间欺骗的定义是：欺人者（简称 Agent）使第三者相信一个与欺人者已相信的命题相矛盾的命题。推广之，可得自我欺骗的经典定义：Agent 使自己相信一个与自己已相信的命题相矛盾的命题（Demos 1960）。由于自我欺骗定义是人际间欺骗定义的简单推广，它保留了欺骗的本质属性。

自我欺骗的形式结构可以用符号语言表达为：

$$B(P) \wedge B(\neg P) \qquad (1)$$

式中的 P 表示任意语句，B( )是相信算子（believe that）；相信算子的空位中填入一个语句得到一个信念语句，称信念内容语句；信念语句 B（P）和 B（¬P）分别读作相信 P（believe

that P）和相信非 P（believe that ¬P）；类似地，可以定义知道算子，知道语句，知道内容语句和相应的读法。在这里，“∧”是经典合取联结词；故而（1）表明相信 P 和相信非 P 同时为真。对（1）作认知解释可得：相信是一种认知状态，相信的对象是语句 P 所表达的命题（简称命题 P）；在自我欺骗这种认知状态下，Agent 同时相信命题 P 和命题非 P。在不会混淆的场合中，我们用“自我欺骗”同时指称（1）和它的解释。

从逻辑学层面看，在经典信念逻辑系统中，（1）等价于 B（P∧¬P）或 B（P）∧¬B（P），后两个公式是逻辑矛盾（da Casta 1990）；因此，自我欺骗导致逻辑矛盾。从心理学层面看，在传统心理学理论框架下，自我欺骗不可能存在。例如：Mele（1997）认为自我欺骗的经典定义无法摆脱以下两个困境：

困境 1 自我欺骗的心理状态困境（state puzzle）：如果自我欺骗存在，根据定义，Agent 相信命题 P 的心理状态和相信命题非 P 的心理状态共存，这种矛盾的心智状态不可能存在。该困境质疑自我欺骗现实存在的可能性。

困境 2 自我欺骗的心理过程困境（dynamic puzzle）：根据定义，自我欺骗和人际间欺骗具有相同的结构，都是有策略的欺骗行为，而在知道欺骗策略的情况下，Agent 不可能实现自我欺骗。该困境质疑自我欺骗的可实现性。

根据以上分析，在自我欺骗的定义中，自欺者同时相信两个矛盾命题，在逻辑学和心理学的经典理论框架里被认为是不可能存在的。因此，相信一个与已相信的命题相矛盾的命题，或同时相信命题 P 和命题非 P，被称为自我欺骗悖论。自我欺骗悖论直接起因于由经典信念逻辑从自我欺骗的定义引出的逻辑矛盾；它的直接后果是导致自我欺骗何以可能，即自我欺骗的存在性问题；只有解决这一问题才可能进一步研究自我欺骗的性质。

本文试图通过消除自我欺骗悖论的途径解决自我欺骗的存在性问题，为进一步研究自我欺骗问题提供理论基础。首先，澄清“相信”这一自我欺骗悖论所使用的基本概念。其次，在文献（鞠实儿 1993，1995；周清，鞠实儿 1999）的基础上描述一种基于证据的信念理论，给出信念语句为真的形式条件。再次，根据上述理论，给出自我欺骗发生的充分条件；在逻辑层面上，证明自我欺骗可无矛盾地实现性，消除自我欺骗悖论；最后，我们利用上述逻辑分析的结果设计心理学实验，用实验表明自我欺骗的可实现性，并揭示它的心理机制。需要说明的是：“自我欺骗”一词在日常语言中被广泛运用，是一个含义丰富的概念，对自我欺骗的研究涉及社会心理学、政治学和宗教学等领域，本文仅从认知的角度[3]研究信念语句，解决自我欺骗问题，为综合情感、语用等因素研究这一问题提供基础。

2 自我欺骗的逻辑结构

下面将给出信念语句为真或具有某信念的定性条件。按照 Cohen（1989）的信念理论，Agent 相信一个命题，或具有一个信念可定义为：

定义 1 Agent 相信某命题，当且仅当 Agent 倾向于认为该命题为真（各种信念定义的比较可参见鞠实儿（1995））。

问题是在什么条件下 Agent 才可能倾向于认为一个命题为真？在自然科学中，判定一个命题是否可信以为真的常用方法是所谓假说-演绎方法。具体操作程序是：给出被检验假说和辅助条件，从中推出可检验的结论，利用实验检验这些结论；如果结果表明有一个可检验结论为假，被检验假说被判定为假；如果实验结果表明所有这些结论均为真，可相信被检验假说为真；而表达这些检验结果的语句就被认为是判定假说真值的证据。从逻辑学的角度

看，对于一个一般性假说而言，一个假的可检验后承可否定它；但是，即使所有已知的证据支持它，使得我们完全相信这一假说，我们仍然无法排除该假说最终被证伪的可能性；因此，支持一个一般性假说的证据总是不完备的。当然，对于一个本身就是可检验的单称假说而言，对它的直接检验便可断定它的真假；因此，支持它的证据集也可以是完备的。以下，我们将根据上述科学方法，引入基于假说－演绎模型的证据概念。

定义 2 如果一个语句的演绎后承为真，则该后承支持该语句。

定义 3 如果一个语句的演绎后承为假，则该后承反驳该语句。

定义 4 每一个支持或反驳某语句的语句称为该语句的证据，证据集合为（该语句的）证据集。

定义 5 既不支持也不反驳某语句的语句被称为对于该语句不相关的语句，简称不相关语句。

定义 6 一个语句的证据集相对于该语句是完备的，如果该证据集中包括反驳该语句的证据或该语句本身；否则它相对于该语句不完备。

定理 1 某语句的证据集是不完备的，当且仅当它不能判定该语句的真值，但它的所有成员都支持该语句。

证明 根据**定义 6** 得。证毕

显然，当 P 的证据集完备时，我们掌握了判断 P 的真值所必须的全部理由，故或知道 P 真或知道 P 假。因此，根据**定义 1** 和**定理 1**，相信 P 的条件必须基于不完备证据集，即由既不能证明 P 也不能反驳 P，但支持 P 的语句构成的集合。

接下来的问题是：是否 P 的任何一个不完备证据集都能成为相信 P 的条件，是否需要对它作出某些限制？本文仅从认知意义的角度考察包括证据和命题的逻辑关系，研究信念的形成问题；因此，如有必要，只对不完备证据集作认知方面的限制。本文在假说-演绎法的框架内定义证据支持概念，并论证了它的合理性；同时，在此框架内只有满足**定义 2** 的语句才对信念内容语句提供支持，而不完备证据集中只包含这类语句；所以，没有必要对其中的成员的认知性质作限制。因此，唯一的可能是对证据集的大小提出某种要求。由于逻辑等价的命题可以合理的被看作是同一证据，所以，这里所说的证据集的大小应该被理解为在逻辑等价关系下的等价类的集合的大小，即，由逻辑等价关系对原证据集取商集所得到的集合的大小。Agent 的认知能力有时空界限，在任一时刻 Agent 具有的证据集的大小的上限应该是：包括 Agent 在这一时刻所能获取的所有证据。由于这一上限是 Agent 的认知能力决定的，因此无法超越。同时，如果证据集的规模未能达到该上限，Agent 可能由于忽略已有的证据，对语句的真值作出本来可以避免的错误判断。根据以上所述，我们可给出知道和相信一个命题的条件：

条件 1 如果 Agent 关于 P 的证据集是完备的且它的成员均支持 P, 那么 Agent 知道 P。

条件 2 如果 P 的不完备证据集由 Agent 在某一时刻所能获取的所有证据组成，那么 Agent 在该时刻相信 P。

根据**定义 1** 和**定理 1**，**条件 2** 可以解释为：尽管 Agent 在某一时刻所能获取的证据不充分，如果它们都支持 P，那么 Agent 倾向于认为 P 真。在这里，作为客观逻辑关系的概念，“证据”与“支持”得到了严格的刻画。但是，分别作为主观的能力和状态的概念“所能取得”和“倾向于认为 P 真”则是需要给予进一步说明的。“所能取得”一词完全在认知意

义上被理解，意指在正常环境下依据认知者的能力能够得到。作这样的理解是必要的，否则 Agent 会由于没有得到应该得到的证据而对是否相信某命题犹豫不决。“倾向于认为 P 真”涉及到如“易碎”这类趋向性词项。通常借用一个虚拟条件句对它们进行解释。例如：“易碎”可解释为，如果掉到地上就会破碎。但是，与“易碎”这类关于物理性质的词项不同，由于心理倾向不可观察，不同的 Agents 对相应的词项会有不同的理解。“倾向于认为 P 真”可解释为：如果 P 为真，则不足为怪；或如果 P 为假，则令人惊讶（参见 Shackle(1953) 的潜在惊奇理论）；或者：尽管 P 假是逻辑可能的，但还是视 P 为真。

下面，我们证明一个与自我欺骗的逻辑结构有关的定理。

定理 2 如果 Agent 具有支持 P 的不完备证据集，并且有支持¬P 的不完备证据集，且没有反驳 P 或者反驳¬P 的证据，那么 Agent 同时相信 P 和¬P，B(P)和 B(¬P)同时成立。

证明 从**条件 2** 直接可证。证毕

根据**定理 2**，自我欺骗产生的必要条件是 Agent 的信息不完全性，它表现为某时刻相对于某一命题的证据集 E 的不完备性；只有在此条件下，两个矛盾的命题才可能同时得到已有证据的支持；从而 Agent 同时相信两个矛盾命题，出现自我欺骗行为。而在证据完备的条件下，根据**条件 1** 我们或证明 P 或证明¬P，两者不可能同时发生。

综上所述，本文认为自我欺骗的发生过程是一个基于证据的信念运算过程，经典自我欺骗可以在逻辑层面上没有矛盾地实现，定理 2 说明了自我欺骗发生的条件和机制。以下将对上述自我欺骗理论作心理学实验检验。

3 实验和解释

上述分析从逻辑学角度给出经典自我欺骗的逻辑结构，指出在一定条件下，经典自我欺骗的发生是逻辑可能的。但是，逻辑可能事件不一定事实可能。因此，紧接着的问题就是：经典自我欺骗事实可能吗？下面将用心理学实验方法，检验自我欺骗的现实可实现性。

从**条件 1** 和**条件 2** 及其解释出发，我们可以合理地说明直观意义下知道与相信的关系。事实上，由于**条件 1** 的前件包含**条件 2** 的前件；因此，知道 P 者必相信 P。同时，根据 Agent 获取证据的能力，还可以说明为什么不同的 Agent 对同一个命题会有知道、相信或不相信等不同的命题态度。但是，**条件 2** 毕竟将客观逻辑关系与主观的命题态度联系起来；这种联系是通过对经典“知道”概念的逻辑分析建立起来的，“知道”的定义也仅仅是认识论分析的结果。因此，我们将对**条件 1** 和**条件 2** 作心理学实验检验。本文通过**条件 2** 和**定理 2** 说明自我欺骗现象，因此要考察自我欺骗的现实可实现性问题，还需要对**定理 2** 作心理学实验。根据以上分析，我们的实验目的是：

（1）检验**条件 1**：如果 Agent 关于 P 的证据集是完备的且它的成员均支持 P, 那么 agent 知道 P。

（2）检验**条件 2**：如果 P 的不完备证据集由 Agent 在某一时刻所能获取的所有证据组成，那么 Agent 在该时刻相信 P。

（3）检验**定理 2**：如果 Agent 具有支持 P 的不完备证据集，并且有支持¬P 的不完备证据集，且没有反驳 P 或者反驳¬P 的证据，那么 Agent 同时相信 P 和¬P，B(P)和 B(¬P)同时成立。

实验设计思路：为了有效控制关于某语句所表达的命题的证据集的完备性和不完备性特征，并避免不同被试的知识背景的干扰，我们采用图形认知实验方法。根据上文的理论，我们将被试判断静物图内容的心理过程描述为一个收集、运算证据的过程，判断的结果是对某内容语句的断定。**条件 1** 和**条件 2** 为 Agent 判断静物图提供了依据。图片本身是关于某物

的证据集，若图片提供关于某物的完备证据集，则 Agent 知道“这是关于某物的图”；若图片提供关于某物的不完备证据集，则 Agent 相信“这是关于某物的图”。因此，在图片上区分完备证据集和不完备证据集的条件成为实验设计的关键。我们的做法是：由于即使是一张真实照片也难以表达一个内容语句的完备证据集，我们把一张能基本表达某事物主要特征的图，称为关于该事物的语句的完备证据集，如：普通静物图。相应的，我们把一张只能表达某事物的部分主要特征的图，称为关于该事物的语句的不完备证据集，如：两可图（三可图）。一张两可图（或三可图）是关于两个（或三个）不相容语句，**语句 P 和 Q**（和**语句 R**）的不完备证据集。接下来的问题是：如何在实验中区分“知道 P”和“相信 P”。我们预测，被试“知道 P”和“相信 P”的差别表现在排除干扰的能力上。我们将考察任务设计为看图做选择题的形式，若被试“知道 P”，则表现为排除无关选项，选出表达图片的正确语句且准确率很高；若被试“相信 P”,则表现为能基本排除无关选项，选出表达图片的正确语句但是准确率与前者有显著差别。

鉴于以上分析，实验一通过对照组实验考察**条件 1** 和**条件 2** 。实验二为**定理 2** 提供实验支持。

3.1 预备实验

目的：挑选出能准确表达两个图形特征的两可图以制作实验材料。（参见附录一）

3.2 实验一

目的：对照组实验，检验**条件 1** 和**条件 2**。对于每幅静物图，如果被试成功选出正确答案，那么实验结果支持**条件 1**；对于每幅两可图，如果被试成功选**语句 P** 或**语句 Q** 的百分比显著高于选不相关语句的百分比，那么实验结果支持**条件 2**。

被试：中山大学管理学院本科二年级学生 50 人，组 1（男 12，女 13），组 2（男 12，女 13）。

材料：**材料 1** 和**材料 2**。

任务：在计算机上依次显示 20 幅图，要求被试看图，根据每幅图下的选项做选择题。

结果：表 1 和表 2 分别列出两组被试对静物图作出正确选择的情况。

表 1 50 名被试对静物图作出正确选择的情况

	N	图 1	图 3	图 4	图 5	图 8	图 9	图 12	图 14	图 15	图 18
组 1(人)	25	24	23	25	25	25	25	23	24	23	24
组 2（人）	25	23	22	23	24	24	24	22	22	23	25
合计（人）	50	47	45	48	49	49	49	45	46	46	49
占被试总人数的百分比（%）		94	90	96	98	98	98	90	92	92	98

表 2 50 名被试回答静物图的正确率情况

	N	正确率					
		100%	90%	80%	70%	60%	50%
组 1（人）	25	16	7	2	0	0	0
组 2（人）	25	15	9	0	0	0	1
合计（人）	50	31	16	2	0	0	1
累计百分比（%）		62	94	98	98	98	100

结果表明，对于每一幅静物图的平均反应正确率都在 90%以上（见表 1）；50 名被试中，答对十幅静物图的有 31 人，占 62%，答对九幅以上的有 47 人，占 94%（见表 2）。因此，可认为**条件 1** 得到实验结果支持。

表 3 和表 4 分别列出对每个两可图（图 2、6、7、10、11、13、16、17、19 和 20，其中图 6 是三可图，组 1 被试选择**语句 P** 和组 2 被试选择**语句 Q** 的比例以及卡方检验结果。

表 3 组 1 被试选择句 P 的比例及卡方检验结果

	图 2	图 6	图 7	图 10	图 11	图 13	图 16	图 17	图 19	图 20
正确比例（%）	100.00	64.00	88.00	84.00	88.00	80.00	80.00	88.00	88.00	72.00
Chi-Square	25.00**	1.96	14.44**	11.56**	14.44**	9.00**	9.00**	14.44**	14.44**	4.84*

*表示 p<0.05； **表示 p<0.01。

表 4 组 2 被试选择语句 Q 的比例及卡方检验结果

	图 2	图 6	图 7	图 10	图 11	图 13	图 16	图 17	图 19	图 20
正确比例（%）	72.00	84.00	76.00	88.00	72.00	92.00	76.00	84.00	80.00	80.00
Chi-Square(a)	4.84*	11.56**	6.76**	14.44**	4.84*	17.64**	6.76**	11.56**	9.00**	9.00*

*表示 p<0.05； **表示 p<0.01。

结果表明对于所有两可图，虽然被试判断两可图的准确率均低于静物图，但是，组 1 被试选**语句 P 和**组 2 被试选**语句 Q** 的百分比都显著大于选不相关语句的百分比，均达到显著水平。对于三可图（图 6），组 1 被试选**语句 P** 的百分比与选不相关语句的百分比的差异没有达到显著水平，但是选**语句 P** 选项的被试占样本的 64%，超过了机率水平（50%）；而组 2 被试选**语句 Q** 的百分比都显著大于选不相关语句的百分比，且达到显著水平。因此，可认为**条件 2** 得到实验结果支持。

3.3 实验二

目的：考察**定理 2**，如果被试对两可图选“C 两者都是”（三可图选“D 三者都是”）的百分比显著高于选**语句 P** 或**语句 Q**（或**语句 R**）的百分比，那么证明被试把该图形作为支持两个不相容命题的证据，从而相信两个不相容的命题。

被试：中山大学管理学院本科二年级学生 50 人（男 26，女 24），均未参加过实验一。

材料：**材料 3**。

任务：在计算机上依次显示 20 幅图，要求被试看图做选择题。

结果：表 5 列出被试对两可图选 C（三可图选 D）的比例以及卡方检验结果。被试选 C(或 D)的百分比都高于选**语句 P** 或**语句 Q** 的百分比，且达到显著水平。因此，可认为**定理 2** 得到实验结果的支持。

表 5 被试对两可图（或三可图）选择 C（或 D）的比例及卡方检验结果

	图 2	图 6	图 7	图 10	图 11	图 13	图 16	图 17	图 19	图 20
正确比例（%）	32.00	50.00	66.00	60.00	48.00	66.00	52.00	72.00	62.00	68.00
Chi-Square(a)	7.72*	21.52**	25.48**	17.92**	14.56**	27.64**	8.32*	34.12**	23.56**	27.52**

*表示 p<0.05； **表示 p<0.01。

3.4 解释和说明

实验一结果支持 **条件 1** 和**条件 2**。现在考虑实验二的结果如何为**定理 2** 提供支持，即被试同时相信两个不相容语句与被试同时相信两个矛盾语句之间的关系问题。下面，我们做一些具体分析。根据文献（Ju Shier, 1999），一个语句的否定可表达为与该语句不相容的所有语句的析取。设 $S=\{P_1, \cdots P_n\}$ 为所有与 P 不相容的语句的集合，故有：$\neg P = P_1 \vee P_2 \vee \cdots \vee P_n$。对于两可图，根据它提供的证据，排除了 $P_2, \cdots P_n$ 为真的可能；在此条件下，$\neg P$ 为真当且仅当 $\neg P_1$ 为真。因此，在确定 B（$\neg P$）只要考虑证据与 $\neg P_1$ 的关系即可。上述

结论可推广至三可图情形。鉴于以上分析，实验二结果（见表 5）支持被试同时相信两个不相容语句时，该实验为**定理 2** 提供实验证据。

综上所述，本系列实验为**条件 1**、**条件 2** 和**定理 2** 提供了实验支持，说明本文给出的知道一个知道语句或相信一个信念语句的条件不仅是认识论和逻辑分析的结果，而且还得到科学实验的支持。实验结果为经典自我欺骗可以无矛盾的实现提供了实验证明。

4 几种解决方案比较

本章节试图将我们的自我欺骗解决方案与已有解决方案作比较。要使解决方案对自我欺骗现象具有解释力，该方案至少要满足以下要求：（1）由于自我欺骗问题涉及 Agent 的认知过程，解决方案的基本假设应符合直观并具有认识论基础；（2）该方案应给出自我欺骗发生的充分条件，因而具有预测和再现自我欺骗现象的能力；（3）理论内部自恰。根据上述要求，对已有方案分析如下：

方案 1 怀疑主义方案。Haight（1980，2000）Agent 的信念系统具有一致性特征，自我欺骗定义导致矛盾，自我欺骗没有存在的理由。然而，怀疑主义方案对自我欺骗存在性问题的简单否定，不能给日常生活中的自我欺骗现象一个合理解释。

方案 2 分裂主义方案。Barnden（1996）认为日常生活中存在经典定义的自我欺骗，试图采用分裂“自我”的方法摆脱自我欺骗悖论，实现自我欺骗与人际间的欺骗的严格的相同结构关系。他认为“自我”（self）是由多个功能独立的部分组成的，自我欺骗时，一部分“自我”扮演欺骗者的角色，另一部分“自我”扮演被骗者的角色。分裂主义方案受到了 Haight（2000）的攻击，她指出“自我”分裂是精神失常的表现，即使不是，也必须回答分裂的自我是如何整合的，在什么条件下分开，意识是否参与等问题。

方案 3 修改定义方案。Mele（1983，1987，1997）自我欺骗的经典定义用人际间欺骗比拟自我欺骗不适当，矛盾信念并不同时存在于自我欺骗行为中。他试图通过修改自我欺骗的定义，引进“认知偏向”概念解释自我欺骗的发生机制，重构一个从自欺者角度看来是合理的解释，化解悖论。da Casta（1990）指出，Mele 的理论对“自愚”现象有解释力，但是“自愚”和“自欺”是不同的概念，因此，Mele 没有解决自我欺骗问题。

方案 4 弗协调逻辑方案。da Casta（1990）认为经典信念逻辑不适用于信念的表达和推理，他将弗协调逻辑理论应用于信念的表达，建立信念的弗协调逻辑（paraconsistent logic of belief）刻画信念，限制矛盾律的作用范围，使自我欺骗得到合理解释。但是，该方案欠缺对自我欺骗的发生作认识论的说明。

本文运用逻辑分析和心理实验相结合的方法，从基本概念分析出发，保证了（1）；逻辑论证过程保证了（3）；心理实验结果证明了（2）。因此，本文提出的方案较已有方案更有效的解决自我欺骗问题。

5 结论

本文采用逻辑分析和心理实验相结合的方法，为经典自我欺骗提供合理解释。一方面，我们从逻辑学角度出发，运用概念分析和演绎推理方法给出一种基于证据的信念理论，说明自我欺骗的逻辑结构，解决自我欺骗悖论，指出自我欺骗的发生是逻辑可能的；另一方面，我们从心理实验角度出发，通过设计并实施两可图认知实验证明了根据本文给出的自我欺骗逻辑结构及自我欺骗发生的充分条件，指出它是基于证据的信念运算结果，克服了 Mele（1997）提出的两个自我欺骗困境。因此，从证据-信念角度看，自我欺骗可以没有矛盾的实现。

本文的目的不仅在于自我欺骗问题的解决，更在于对认知领域研究方法的探讨，本文以自我欺骗问题为例，展示了逻辑分析和心理实验相结合的研究方法的优势，认为该方法可能是研究认知问题的一种新的卓有成效的研究方法。

参考文献

[1] Audi , R.(1985) Self-Deception and Rationality In: Self-Deception and Self-Understanding, Martin, M. ed. University Press of Kansas. 1985. 169-94

[2] Barnden,J.A.(1996) Consciousness and Common-sense Metaphors of Mind. In: Reading for Mind: Foundations of Cognitive Science, S.O'Nuallain, P. Mckevitt & E. MacAogain ed., John Benjamin

[3] da Casta (1990) N.C.A., French, S. Belief, Contradiction and the Logic of Self-Deception. In: American

[4] Philosophical Quarterly, Oxford. 27(3), 179-77

[5] Cohen, J. (1989) Belief and Acceptance. In: Mind. XCVIII(391) 368-89

[6] Davidson (1982) Paradoxes of Irrationality. In: Philosophical Essays on Freud. Ed. Wollheim, R.,Hopkins, J. Cambridge University Press.

[7] Davidson (1998) Who is Fooled. In: Self-Deception and Paradoxes of Rationality. Dupuy J.-P. ed. 1-18

[8] Demos R. (1960) Lying to Oneself. The Journal of Philosophy, 1960, 57: 588-595

[9] Fingarette, H. (1969) Self-Deception. Humanities Press.

[10] Haight, M.R.(1980) Study of Self-Deception. Harvester Press.

[11] Haight, M. R.(1985) Tales from a Black Box. In: Self-Deception and Self-Understanding Martin, M.W. ed. University Press of Kansas. 1985, 244-60

[12] Ju S, Chen Y Q. The Logical Structure of Belief. In: Fan-Lun Xiong, etc, ed. Proceedings of Paces' 95, Pacific-Asian Conference on Expert Systems. Publishing House of Electronics Industry, 1995, 143-146.

[13] Loar，B.（2001），"Meaning" in The Cambridge Dictionary of Philosophy, edied by R.Audi, Cambridge Univ.

[14] Press, 2001

[15] Locke K.(1975) An Essay Concerning Human Understanding. 5th. Ed. P. H. Nikditch, Oxford Clarendon Press，1975. 46-54

[16] Martin, M. W. (1985)ed. Self-Deception and Self-Understanding. University Press of Kansas.

[17] Martin, M.W. (1986) Self-Deception and Morality. University Press of Kansas.

[18] Mele, A. (1982) Self-Deception, Action and Will: Comments. Erkenntnis, 18:159-64

[19] Mele, A. (1987) Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control. Oxford University Press.

[20] Mele, A.（1997）Real Self-deception. Behavioral and Brain Sciences, 1997(20): 91-137

[21] Moore, G. E. (1942) Reply to My Critics. In: Schilpp, P. ed. The Philosophy of G. E. Moore. La Salle, Illinois: Open Court.

[22] Moore, G. E (1944) Russell's Theory of Descriptions. In: Schilpp, P. ed. The Philosophy of Bertrand Rissell. La Salle, Illinois: Open Court.

[23] Quattrone G, Tversky A.(1984) Causal Versus Diagnostic Contingencies: on Self-Deception and on the Voter's Illusion. Journal of Personality and Social Psychology, 1984, 46: 237-48

[24] Rorty, A. (1972) Belief and Self-Deception Inquiry, 5, 387-410

[25] Sackeim H A, Gur R. (1978) Self-Deception, Self-Confrontation, and Consciousness.In G. Schwartz, D. Shapiro ed. Consciousness and Self-Regulation. Plenum Press: 1978(2)，173

[26] Shackle,G.L. (1953) The Logice of Surprise. Economica.

[27] 鞠实儿.三种信念悖论的消除.自然辩证法通讯，1995, 17（96）：16-21

[28] 鞠实儿.信念的证据和不一致信念的表达.哲学研究，1993,增刊：50-53

[29] 鞠实儿.博士论文《非巴斯卡归纳概率及其逻辑基础》（1999）。

[30] 周清.鞠实儿.（1999）《不确定推理的支持度》.《软件学报》.10（2）：193-6）

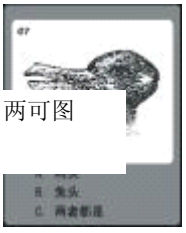
附录一　实验材料举例

材料 1　以语句 P 和一不相关语句组成 A、B 两选项配在每幅两可图（或三可图）下；以对静物图的准确描述和一不相关语句组成 A、B 两选项配在每幅静物图下。

材料 2　以语句 Q 和一不相关语句组成 A、B 两选项配在每幅两可图（或三可图）下；以对静物图的准确描述和一不相关语句组成 A、B 两选项配在每幅静物图下。

材料 3　以语句 P、语句 Q 和“两者都是”组成 A、B、C 三选项配在每幅两可图下；以语句 P、Q、R 和“三者都是”组成 A、B、C 和 D 四选项配在三可图目标图下；以对静物图的准确描述、一不相关语句和“两者都是”组成 A、B、C 三选项配在每幅静物图下。

实验指导语：“这是一个与个人资料无关的图形认知实验。请看图做单项选择题，每幅图下都有 A、B 两个选项（问卷 3 为“多个选项”），在你认为与该图内容吻合的选项上打‘O’。做完一题后请按‘空格’键，计算机将显示下一幅图。不得不选，凭第一感觉真实作答。”



两可图

静物图

Study on the Mechanism of Self-Deception

JU Shi-er[1]　ZHAO Yi[1]　FU Xiao-lan[2]

(1. Institute of Logic and Cognition, Zhongshan University, Guangzhou 510275,China; 2. The Institute of Psychology, Chinese Academy of Sciences, Beijing 100101,China)

Abstract: Researches on self-deception focus on the paradox of self-deception, involving discussion on the rationality of inconsistent believes. Both logicians and psychologists are interest in the mechanism of self-deception. In this paper, first, we will discuss what is self-deception and what is the paradox of self-deception. Secondly, we try to eliminate the paradox of self-deception basing on the analysis of the nature of believe. Third, we will design a set of psychological experiment to support our solution and reveal the mechanism of self-deception.

Key Words : Cognitive Bias ; Self-Deception; Paradox of Self-Deception; Belief

收稿日期：2003-5-9

作者简介：鞠实儿（1953-），浙江人，中山大学逻辑与认知研究所，教授，博士生导师
赵艺（1975-），中山大学哲学系博士研究生
傅小兰（1963-），中国科学院心理研究所，研究员，博士生导师

---

[1] Moore 悖论：It is raining, but I do not believe that it is raining. Moor 的问题是“I”如何可能既认为外面在下雨，又认为外面不在下雨？Wittgenstein 也讨论了该问题. 详细参见：*“Philosophical Investigations”* (ⅡX)和“*Remarks on the philosophy of psychology*”，471、478、501。

[2] Mele 认为 Sackeim 和 Gur 的实验设计含有特设性假定；Quattrone 和 Tversky 的实验难免练习效应。

[3] 自我欺骗涉及到“相信”这类命题态度。通常认为（B. Loar，2001，p.546）：当我们理解一个语句时，关于这个语句所把握的东西就是该语句的意义；除去其中的感情与语用因素得到的便是语句的认知意义（cognitive meaning），它也是一个陈述句与它的疑问式的共同表达的东西，我们称之为命题；命题或真或假，表达某命题的语句与该命题具有相同的真值。