认知悖论

陈　波

目录

一个逻辑理论可以通过其处理疑难的能力而得到检验。在思考逻辑时，头脑里尽量多装难题，这是一种有益的方法，因为解这些难题所要达到的目的与自然科学通过实验所要达到的目的是一样的。——罗素

所谓“认知悖论”（epistemic paradox），是指与知识、信念、证据以及与知道、相信、怀疑、证成（justification）等认知行为和态度相关的各种难题和谜题，其中包含着矛盾和不一致。最古老的认知悖论是柏拉图归之于苏格拉底的美诺悖论，最典型的认知悖论包括彩票悖论（与知识的接受有关），序言悖论，意外考试悖论，可知性悖论，知道者悖论，盖梯尔问题，等等。认知悖论向我们表明，存在某个深层的错误，如果这个错误不直接与知识有关的话，则它肯定与知识相关联的其他概念如证成、合理信念和证据等有关。对某个认知悖论的解决，常常意味着认识论研究方面的某种新进展。

一、古希腊时期的认知悖论

美诺悖论

在与苏格拉底的对话中，美诺（一名富家子弟）提出一种观点：研究工作不可能进行，还提出了下述论证：“一个人既不能研究他所知道的东西，也不能研究他不知道的东西。他不能研究他所知道的东西，因为他知道它，无需再研究；他也不能研究他不知道的事情，因为他不知道他要研究的是什么。”[①] 为明确起见，将该论证整理如下：

（1）如果你知道你所寻求的东西，研究是不必要的;
（2）如果你不知道你所寻求的东西，研究是不可能的。
（3）所以，研究或者是不必要的，或者是不可能的。

现在的问题是：这个论证有效吗？

分析：

美诺的论证有一个隐含的前提：“或者你知道你所寻求的东西，或者你不知道你所寻求的东西。”但其中有歧义。

A. 你知道你所探究的那个问题;
B. 你知道你所探究的那个问题的答案。

在（A）的意义上，（2）是真的，因为如果你不知道你要研究什么问题，研究工作是没有办法进行的；但（1）却是假的，因为尽管你知道你要探究什么问题，但不知道该问题的答案，研究工作仍有必要进行：它的目标就是探寻该问题的答案。在（B）的意义上，（1）是真的，因为如果你知道你所要探究的问题的答案，那还什么必要去再做研究？但（2）却是假的，因为尽管你不知道某个问题的答案，但你知道你要探究什么问题，研究工作仍有可能进行。故两个前提不是在同一种意义上为真。于是，从一对真的前提，即（1B）和（2A），推不出任何结论，因为其中有歧义性，说的不是一回事。

为了看清楚歧义性，我们还可以再考虑这样一个问题：“你有可能知道你不知道的东西吗？”在一种意义上，答案是否定的，因为你不可能同时知道又不知道同一个东西；但在另一种意义上，答案是肯定的，你可以知道你对之尚没有清楚答案的那个问题，你遵循正确的程序去回答该问题，最后你知道了你先前不知道的东西，也就是该问题的答案。

结论:美诺的论证是有缺陷的,它犯了歧义性谬误。但柏拉图并没有简单拒斥美诺悖论，而是由此发展出一套“学习就是回忆”的理论。

麦加拉派的认知悖论

（1）幕后人悖论

你认识那个幕后的人吗？不认识。那个人是你的父亲。所以，你不认识你的父亲。

[①] 苗力田主编：《古希腊哲学》，中国人民大学出版社，1989 年，第 250 页。

（2）厄勒克特拉悖论

厄勒克特拉不知道站在她面前的这个人是她的哥哥，但她知道奥列斯特是她的哥哥。站在她面前的这个人与奥列斯特是同一个人。所以，厄勒克特拉既知道又不知道这同一个人是她的哥哥。

这两个悖论最早表明，在由“认识”、“知道”、“相信”、“怀疑”等语词组成的上下文（语境）中，经典逻辑的替换原则失效。

当代广泛讨论的弗雷格之谜、分析悖论和信念之谜，与麦加拉派所提出的问题本质上是类似的，都牵涉到替换原则在认知语境中是否仍然有效。

弗雷格之谜

由萨蒙（Nathan Salmon）在其同名专著[①]中提出和探讨，力图解答下面的问题：当 a、b 分别代表两个专名时，“a=a”和“a=b”为何会具有不同的认知价值：前者是同语反复，后者却提供新的信息。

当考虑认知价值时，弗雷格之谜就是一个认知悖论，其实质与克里普克所谓的“信念之谜”是一样的，也与麦加拉派所提出的那两个悖论是一回事情。

分析悖论

分析悖论与弗雷格之谜有些类似。它与摩尔所提倡的概念分析（conceptual analysis）有关，最早由布拉克（Max Black, 1909–1988）提出[②]，涉及如下问题：概念分析如何能够既是正确的又传达信息？换言之，我们如何同时说明概念分析的正确性（correctness）和传达信息这个性质（informativeness）？

根据摩尔，概念分析应满足三个条件：（i）被分析项和分析项都是概念，在正确的分析中，两个概念必须同义；（ii）用来表示两个概念的语言表达式不同；（iii）表示分析项的表达式明确提到表示被分析项的表达式未明确提及的某些概念。他给出了如下三个例子：

例 1：“是兄弟”这一概念等同于“是男性同胞”这一概念。

例 2：“x 是兄弟”这一命题函项等同于“x 是男性同胞”这一命题函项。

例 3：断言某人是兄弟等同于断言某人是男性同胞。

这里只考虑例 1，并把它简化为下面的句子：

（1）兄弟是男性同胞。

也可以把（1）写成一阶逻辑公式：

（1′）$\forall x$(x 是兄弟 $\rightarrow$ x 是男性同胞)

如果“兄弟”和“男性同胞”是同义的，它们就可以相互替换。由（1）可以得到：

---

① Salmon, N. *Frege's Puzzle*. Cambridge, Mass: MIT Press, 1986.

② 参见Max Black, “The ‘Paradox of Analysis’,” *Mind* 53 (1944): 263-267; “The ‘Paradox of Analysis’ Again: A Reply,” *Mind* 54 (1945): 272-273.

（2）兄弟是兄弟。

也可以把（2）写成一阶逻辑公式：

（2′）∀x(x 是兄弟 → x 是兄弟)

如果要求在正确的分析中被分析项和分析项必须是同义的，并同时接受替换规则的话，就会得到两个结果：（1）是正确的，依据替换规则，可得到（2），但（2）不传达任何信息，而不传达信息的分析是不足道的；如果认为（1）中的被分析项和分析项不是同义的，则（1）不正确，但它传达信息。由此可知，像（1）这样的分析是不正确的却是足道的。由此我们面临一个严重的问题：能够有正确且足道的概念分析吗？这就是“分析悖论”。

如何解决分析悖论？一种选择是承认（1）是正确且传达信息的，但不允许从（1）得到（2），这意味着抛弃逻辑学中的替换规则，这将导致对经典逻辑做重大修改。几乎没有人选择这条路径。另一种选择是：承认（1）传递信息，但不承认其中的被分析项“兄弟”和“男性同胞”同义。

可以再考虑弗雷格的例子：

（3）多个线段有同一方向，当且仅当它们相互平行。

有人承认（3）传递信息，但其中的被分析项“线段”和分析项“相互平行”并不同义，故不能由（3）得到（4）：

（4）多个线段有同一方向，当且仅当它们有同一方向。

是什么使得（1）和（3）比它们的不足道的对应物，如“兄弟是兄弟”等，有更多的信息内容呢？回答确实是：在分析所提到的概念时使用了不同的概念：兄弟概念是根据两个不同的概念，即男性和同胞来解释的；（更有意思的是），同一方向概念是根据相互平行概念来解释的。如果你有兄弟概念但没有更一般的同胞概念，你就可以设想帕特是一位兄弟却不相信他是一位男性同胞；如果你没有平行线概念，你就能够相信两条线有同一方向却不相信它们相互平行。[①]

布拉克早前提出了类似解释。他认为，（1）涉及“兄弟”（b）、“男性”（m）和“同胞”（s）三者的关系，可用符号表示：R(b, m, s)；而（2）只涉及“兄弟”与其自身的关系，最多是二项关系。若用“I”表示“等于”，可表示为：I(b, b)。由此看出，（1）是非同一性陈述，而（2）却是同一性陈述。

分析悖论引发了对“分析”概念，的怀疑论思考。摩尔令人惊奇地预言了对分析悖论的讨论必定会将战火引至分析和综合的区分这一根本性的问题上。并且他指出，“我并不认为这两个术语有任何清楚的意义”。[②]1950年，蒯因发表著名论文《经验论的两个教条》，对分析—综合区分以及证实主义和还原论发动了摧毁性批评，从而引起了迄今仍未结束的一场

---

① Clark, M. *Paradoxes from A to Z*, Second edtion, p.10.

② 关于分析悖论，亦可参看：李大强，《分析悖论的分析》，《哲学研究》2006 年第 6 期；陈四海，《马克斯·布莱克论分析悖论》，《东方论坛》2012 年第 2 期。

哲学论战。

信念之谜

由克里普克在其同题论文[①]（最初发表于 1979）中提出，与指称相同的名称（简称“共指名称”）和信念归属有关。

> 设想有一位法国人皮埃尔，不懂英语，通过某种途径（如看画册）形成了一个法语信念“Londres est jolie”（伦敦很漂亮）。后来，由于某种机缘，他搬到了伦敦的一个落后社区，通过与当地人一起生活学会了英语。基于在当地的生活经验，形成了一个英语信念 “London is not pretty”（伦敦不漂亮）。不过，他并没有放弃他原有的法语信念。由此产生一个问题：皮埃尔究竟是相信“伦敦很漂亮”还是相信“伦敦不漂亮”？

克里普克论述说，这里至少涉及如下两个原则（为简单起见，略去细节）：

去引号原则：若某个语言的正常说话者，经过反思后，真诚地赞同“p”，则他相信 p。

翻译原则：如果一个语言的句子在该语言中表达一个真理，它在另一个语言中的译文也表达同一个真理。

在学会英语之前，皮埃尔相信 “Londres est jolie”。根据去引号原则，他相信 Londres est jolie。再根据翻译原则，他相信伦敦很漂亮。再对皮埃尔的英语信念“London is not pretty 应用去引号原则和翻译原则，可以得知：他相信伦敦不漂亮。问题是，在学会英语之后，皮埃尔仍然持有他原来的法语信念，故根据去引号原则和翻译原则，他就同时拥有一对相互矛盾的信念：他既相信伦敦很漂亮又不相信伦敦很漂亮。

克里普克想要回答如下问题：在皮埃尔的信念世界中如何出现了矛盾？是哪些因素造成了矛盾？他论证说，替换原则不是造成信念之谜的关键所在，故不能通过攻击替换原则来攻击直接指称论，后者指这样的观点：名称没有涵义，直接指称外部对象；任一名称在所有可能世界都分别指称同一个对象，是严格指示词；名称对它们所在语句的语义贡献也仅在它们的所指。这是他和他的一大批追随者的观点。

克里普克用另外的例子表明，只应用去引号原则就能够造成同样的信念之谜。设想下面的情形：帕德瑞夫斯基（简称“帕德”）是波兰著名的政治家兼音乐家。彼得出席了帕德的专场音乐会，由此知道帕德有杰出的音乐才能，当然也就相信 “帕德有杰出的音乐才能”；在另外某个场合，彼得知道帕德是一位政治家，由于他从来不认为政治家会有杰出的音乐才能，故他把作为政治家的帕德当作另外一个人，故不相信 “帕德有杰出的音乐才能”。对彼得的两个信念应用去引号原则，将导致矛盾：彼得既相信又不相信“帕德有杰出的音乐才能”。

这就是克里普克所谓的“信念之谜”：在信念归属句中，共指名称不能相互替换，否则会由真命题得到假命题，甚至会得出逻辑矛盾。弗雷格早就指出了这一点：替换规则（即外延论题）在引语（包括直接引语和间接引语）语境中不成立，也在许多命题态度词（如“知道”、“相信”）语境中不成立。例如，你不能从“哥白尼相信地球围绕太阳运转”推出“哥

[①] Kripke, S. “A Puzzle about Belief,” in his *Philosophical Troubles, Collected Papers*, vol.1, Oxford and New York: Oxford University Press, 2011, pp.125-161.

白尼相信人是由猿猴进化而来的”。克里普克与弗雷格的差别在于：他在隐含地为他的严格指示词理论辩护。根据该理论，严格指示词对所在语句的唯一语义贡献就是其所指，共指名称应该在任何语句（包括信念语句）中能够相互替换，但多个例证表明：共指名称在信念语句中不能相互替换。克里普克承认，这对于严格指示词理论来说确实是一个问题，他也不知道该如何解决。不过，他转守为攻，力图证明：这个问题对关于名称的弗雷格式理论（即描述论）也存在。克里普克所采取的论证策略是：避开替换原则，引入另外两个原则，即去引号原则和翻译原则。据我看来，克里普克新引入的两个原则后面都隐含替换原则。例如，由于皮埃尔本人并不知道法语词“Londres”和英语词“London”指称同一座城市，即不知道这两个名词共指，因此，在对他做信念归属时，就不能允许它们相互替换，但克里普克通过去引号原则和翻译原则，让这两个名词成为事实上的共指名称，且在对皮埃尔做信念归属时允许它们相互替换，这才造成了他所谓的“信念之谜”。

应该指出，在知识、信念领域对名称“a”和“b”做替换时，其必要条件是：不仅“a”和“b”事实上共指，即 a = b，而且相关认知主体还必须认知到“a”和“b”共指，即知道“a = b”。否则，仅仅使用替换原则就足以造成类似的“信念之谜”。改用弗雷格本人的例子：

（1a）保罗相信长庚星是长庚星。

（1b）保罗不相信长庚星是启明星。

这里，（1b）是用“启明星”替换（1a）中“长庚星”的一次出现的结果，而且“长庚星”和“启明星”这两个名称共指，但如果保罗不知道或不相信这两个名称共指，它们就不能在（1b）中相互替换。很显然，尽管有些名称事实上共指，有人却不知道它们共指；共指不能仅从两个名称的字面上知道，还必须诉诸别的认知要素和认知手段。究竟诉诸哪些认知要素或手段？不同的哲学家会有不同的选择。

二、欧洲中世纪的认识论悖论

欧洲中世纪对悖论做了大量的研究，当时的逻辑学家把悖论叫做“不可解问题”（insolubles），后者是一个令人误解的用法，因为他们也不认为悖论是不可解的，而只是解决起来很困难。大阿尔伯特（Albert the Great, 1193–1280）断言：“不可解问题是这样一个命题，它由一个逻辑矛盾构成，无论承认矛盾的哪一方，都可以推导出对立的一方。”欧洲中世纪的悖论研究开始于 12 世纪巴尔夏姆的亚当（Adam of Balsham, 1100? – 1157?），他研究了说谎者类型的悖论；大阿尔伯特、罗马的吉勒士（Giles of Rome, 1243–1316）、西班牙的彼得（Peter of Spain, 13 世纪，生卒不详）曾简要讨论过悖论；到伪司各脱（Pseudo-Scotus）时期，悖论成为热门话题；奥卡姆的威廉（William of Ockham, 1288–1347）把关于悖论的讨论列为他的逻辑教科书的专门章节，自此以后，悖论研究成为中世纪逻辑学的实质性部分之一。当时的研究集中在两方面：一是提出了各种类型的悖论，二是提出了各种不同的悖论解决方案。

欧洲中世纪逻辑学家在研究悖论的过程中，也涉及一大类与知道、相信、怀疑、犹豫这类认识论概念相关的悖论，其中也涉及真、假等语义概念，它们就是我们目前讨论的“认知悖论”。具体讨论略。

三、认知逻辑和逻辑万能问题

认知逻辑系统

从语形方面说，认知命题逻辑是在经典命题逻辑的基础上加一元认知算子 K_i 和 B_i 构成的，其中：

K_iA 表示：认知主体 i 知道 A；
B_iA 表示：认知主体 i 相信 A。

这两个公式各自的语义解释是：

K_iA：在与认知主体 i 所知道的东西相容的所有可能世界中，A 是真的；
B_iA：在与认知主体 i 所相信的东西相容的所有可能世界中，A 是真的。

我们先列表给出知道逻辑的特征公理：

K $K_i(A\to B)\to(K_iA\to K_iB)$
D $K_iA\to\neg K_i\neg A$
T $K_iA\to A$
4 $K_iA\to K_iK_iA$
5 $\neg K_iA\to K_i\neg K_iA$
.2 $\neg K_i\neg K_iA\to K_i\neg K_i\neg A$
.3 $K_i(K_iA\to K_iB)\vee K_i(K_iB\to K_iA)$
.4 $A\to(\neg K_i\neg K_iA\to K_iA)$

再列出知道逻辑系统的推理规则：

MP 从 A 和 $A\to B$ 推出 B
RN 从 A 推出 K_iA

由此我们可以定义如下的知道逻辑系统，其中每个系统都含有 MP 和 RN，以及全部经典命题逻辑的重言式，故不再单独列出：

KT4 = S4
KT4 + .2 = S4.2 ↑
KT4 + .3 = S4.3 ↑
KT4 + .4 = S4.4 ↑
KT5 = S5 ↑

其中“↑”表示该系统强于它上面的系统。

如果把这些系统中的“K_i”都换成“B_i”，我们就得到相应的相信逻辑系统。

逻辑万能问题

在上面所列的知道逻辑系统中，下列公式都是定理或导出规则：

(1) $K_iA \wedge K_i(A \rightarrow B) \rightarrow K_iB$

(2) $A \models K_iA$

(3) $A \rightarrow B \models K_iA \rightarrow K_iB$

(4) $A \leftrightarrow B \models K_iA \leftrightarrow K_iB$

(5) $(K_iA \wedge K_iB) \rightarrow K_i(A \wedge B)$

(6) $K_iA \rightarrow K_i(A \vee B)$

(7) $\neg(K_iA \wedge K_i\neg A)$

(8) K_i(Taut)，Taut 代表重言式

若把这些公式中的“K_i”换成“B_i”，这些公式也是相应的相信逻辑系统中的定理或导出规则。它们都涉及“逻辑万能问题”，即假定认知主体在逻辑上万能：他们具有无限的资源和推理能力，能够推出他们所知道或所相信的命题的一切逻辑后承。具体来说：

演绎封闭。如果一个认知主体 i 知道或相信一个公式集 Γ，而从 Γ 可以逻辑地推出公式 A，则这个主体 i 知道或相信 A。例如，(1)说，如果 i 知道 A，并且 i 知道 A 蕴涵 B，则 i 知道 B；(3)说，如果 A 蕴涵 B，就可以推出：如果 i 知道 A 则 i 知道 B。

不相干的知识或信念。一个认知主体 i 知道一个逻辑系统的所有定理，特别是他知道所有的经典逻辑重言式。这正是(2)和(3)所说的。但实际情况是：许多逻辑定理或重言式与一个人所具有的有限知识不相干。在很多情况下，他不一定知道它们，甚至不必知道它们。

不相容的信念。(7) 说，在一个人的知识中不能包含矛盾：他不能既知道 A 又知道$\neg A$。如果这还勉强说得过去的话，那么，当把 (7) 中的“K_i”换成“B_i”，得到：

(7′) $\neg(B_iA \wedge B_i\neg A)$

(7′) 说，一个人的信念体系中不应包含矛盾：他不能既相信 A 又相信$\neg A$。这种说法肯定不对，至少对某些认识主体来说是如此：他们的信念世界中往往隐含着逻辑矛盾，但他们没有意识到这一点，故依然泰然自若地拥有其信念系统。由于拥有什么样的信念，与欲求、需要、情感、情绪等等因素有关，后者并不完全受理性控制，由此产生自相矛盾的信念是完全可能的。

爆炸的计算量。由于认知主体要求计算他的信念的所有逻辑后承，从计算的角度看，这会导致计算量的膨胀。现实的认知主体（不管是人还是电脑）都是资源有限的，它们没有足够的时间、空间、记忆能力、金钱去无穷计算下去，它们只计算与它们所关注的当前目标相关联的信息。

正因如此，甚至连认知逻辑的创始人亨迪卡（Jaakko Hintikka）也断言，基于可能世界语义学的认知逻辑不适于处理人类的推理，因为它们假定了认知主体在逻辑上万能，而人类个体在逻辑上并不是万能的。[①]

于是，如何在一个认知逻辑系统中避免“逻辑万能问题”，向人的实际的认知过程逼近，就是认知逻辑学家必须考虑的问题。目前有以下几种选择：句法路径，语义路径，设置不可能世界的路径（以容纳不一致的信念），非标准逻辑的路径，其中有些路径区分了显信念（explicit belief）和隐信念（implicit belief），演绎封闭对显信念不成立，却对隐信念成立。

[①] Hintikka, J. “Impossible possible worlds vindicated,” *Journal of Philosophical Logic* **4** (1975)**, pp.**475–484.

四、 意外考试悖论及其分析

（1）意外考试悖论

“意外考试悖论”是从“突然演习问题”[①] 变化而来。在第二次世界大战期间，瑞典广播公司播出一则通告：

> 下周内将举行一次防空演习，为验证备战是否充分，事先并没有任何人知道这次演习的具体日子，因此，这将是一次突然的演习。

瑞典数学家埃克博姆（L. Ekbom）意识到这个通告具有一种奇异的性质：按通告所给条件，演习不能在下周日举行，因为那样演习就会被事先知道在周日发生，从而不是突然的；因此，周日被排除。同理，周六也可以被排除，既然演习已确定不能在周日举行，那么在余下的六天中，若在周六举行依然不具有突然性。循此继进，同样的推理程序可以排除周五、周四直至周一。埃克博姆由此推出，符合通告条件的突然演习不可能发生。然而，在下周周三凌晨，空袭警报响起，演习“突然”举行……

这在某种意义上构成一个“悖论”，它有许多不同的变体，其中之一是“意外考试悖论”，最早由英国学者奥康诺（D. O’Connor）于 1948 年提出[②]：

> 某教授对学生们说，下周我将对你们进行一次出其不意的考试，它将安排在下周一至周六的某一天，但你们不可能预先推知究竟在哪一天。显然，这样的考试是可以实施的。但有学生通过逻辑论证说，该考试不可能安排在周六。因为，如果它被安排在周六，则周一至周五都未考试，就可推算出在周六，该考试因此不再出其不意。同样，该考试不可能安排在周五。因为，如果它被安排在周五，则周一至周四都未考试，学生们就可预先推算出在周五或周六；已知考试不可能在周六，因此只能在周五，该考试也不再出其不意。类似地，可证明其余四天都不可能安排考试。学生由此得出结论：这样的考试不可能存在。但该教授确实在该周的随便某一天宣布：现在开始考试，也确实大大出乎学生的意料之外。由此得到一个悖论：这样的考试既可以实施，又不可能进行。

（2）对意外考试悖论的分析

克里普克等人的分析

1972 年，克里普克在剑桥大学做了题为“论两个知识悖论”的讲演，讲稿拖到 2011 年才正式发表[③]。在该讲演中，他对意外考试悖论做了比较详细的分析。他提炼出该悖论的 5 个前提，用符号表示它们；然后，他列出了学生推理所需的如下 4 个知识论前提，其中“E_i”表示考试在第 i 天进行 ，“$K_i p$”表示认知主体在第 i 天时知道 p：

（1）E_i, 对于某些 i, $1 \leq i \leq N$（或等价的，$E_1 \vee E_2 \vee \ldots \vee E_n$）

---

① 参见张建军：《逻辑悖论研究引论》，南京：南京大学出版社，2002 年，193—194 页

② O’Connor, M. I. 1948. ‘Pragmatic paradoxes’, *Mind*, vol.57, pp.316-329.

③ Kripke, S. “On Two Paradoxes of Knowledge,” in his *Philosophical Troubles, Collected Papers Volume 1*, Oxford, New York: Oxford University Press, 2011, pp.27-51.

（2）$\neg(E_i \wedge E_j)$, 对任意 $i \neq j, 1 \leq i, j \leq N$

（3）$\neg K_{i-1}(E_i)$, 对每一个 $i, 1 \leq i \leq N$

（4）$(\neg E_1 \wedge \neg E_2 \wedge \ldots \wedge \neg E_{i-1}) \rightarrow K_{i-1}(\neg E_1 \wedge \neg E_2 \wedge \ldots \wedge \neg E_{i-1})$, 对每一个 $i, 1 \leq i \leq N$

（5）$E_i \rightarrow K_{i-1}(\neg E_1 \wedge \neg E_2 \wedge \ldots \wedge \neg E_{i-1})$, 对每一个 $i, 1 \leq i \leq N$

（6）$K_i(p) \rightarrow p$, 对每一个 $i, 1 \leq i \leq N$

（7）$K_i(p) \wedge K_i(p \rightarrow q) \rightarrow K_i(q)$, 对每一个 $i, 1 \leq i \leq N$

（8）$\text{Taut} \rightarrow K_i(\text{Taut})$, 对每一个 $i, 1 \leq i \leq N$

（9）$K_i(p) \rightarrow K_j(p)$, 对每一个 $i, j, 0 \leq i \leq j \leq N$

（10）$K_i(p) \rightarrow K_i(K_i(p))$, 对每一个 $i, 0 \leq i \leq N$

这里，（7）是知识蕴涵真理的原则，（8）是知识的演绎封闭原则，克里普克把（9）叫做“知识持续原则”，（10）表明认知主体 i 的推演能力足够强，他知道所有的重言式。他认为，意外考试悖论的根源不在前提，而在于（9）所表示的知识持续原则：如果认知主体在第 i 天知道 p，他在以后的任何一天都知道 p，即是说，知识是可持续的。但他论证说，（9）不一定成立，因为知识和信念会随着新证据的发现而改变，故可以有 $K_i(p)$但没有 $K_j(p)$。在我看来，克里普克持这样的观点很不好理解，因为他接受（6）：知识蕴涵真理，这种客观意义上的真理一旦为真就永远为真，不会随时间流逝、证据增减而改变。于是，一旦承认某命题是知识，它就应该永远为知识，也不会因时间流逝、证据增减而改变。克里普克接受（6）而否定（9），这难道是前后一致的吗？为什么他没有意识到这种不一致性？

威廉姆森从意外考试悖论中引出了一个更难以理解的结论：存在着不可知的真理！某些命题是真的，但一旦假设知道它们为真，就会得出矛盾，故它们是不可知的真理。与克里普克不同，他论述说，即使人们今天知道某件事情，也推不出他今天知道他明天仍然知道该件事情。这等于直接拒绝了前面提到的认知逻辑的“正内省原则”的历时版本。[①]

下面阐述我对意外考试悖论的分析。

（1）什么是“意外”？

有必要事先澄清什么是“意外”，或者是哪种意义上的“意外”。按我的理解，“意外”有以下两种意义：

（a）理性的“意外”，即逻辑的“意外”，因为理性的基础和核心是逻辑。可以做进一步区分：

（a1）弱意义，即逻辑没有推出的“意外”：从已知信息出发，在逻辑上没有推出 p，但事实上 p；在逻辑上没有推出非 p，但事实上非 p。

（a2）强意义，与逻辑推理相反的“意外”：从已知信息出发，在逻辑上推出 p，但事实上非 p；在逻辑上推出非 p，但事实上 p。

（b）心理的“意外”，即心理预期的“意外”。可以做进一步区分：

[①] 参见蒂莫西·威廉姆森：《知识及其限度》，陈丽、刘占峰译，北京，人民出版社，2013 年，第 6 章。

（b1）弱意义，没有预期到的“意外”：从已知信息出发，在心理上没有预期 p，但事实上 p；在心理上没有预期非 p，但事实上非 p。

（b2）强意义，即与已有预期相反的“意外”：从已知信息出发，在心理上预期 p，但事实上非 p；在心理上预期非 p，但事实上 p。

老师在做预先宣布时，他所说的“意外”是：考试在第 i 天，但学生事先甚至在当天也不知道考试会在当天进行，可用符号表示：$E_i \wedge \neg K_i(E_i)$。这种“意外”既可以是理性意义上的，也可以是心理意义上的，包括各自的弱意义和强意义。从直观上说，即使老师事先宣布，这种“意外”考试仍然是可以发生的，甚至在一天之内也可以：你若假定老师的话为真，考试只能在当天进行，而这个考试已经事先知道，不再意外，你推出结论说：老师不可能按他的条件实施考试。但老师马上宣布：现在考试！这次考试仍然是一个“意外”，无论是在理性的意义上还是在心理的意义上！

那么，意外考试悖论的根源究竟在哪里呢？只有两种可能性：一是老师的宣布出错，二是学生的推理出错。下面分别考察之。

（2）老师的宣布出错？

确实有人这么认为，并且这是该悖论发表之后所获得的最初反应。例如，发表该悖论的奥康诺就认为，老师的宣布是自我反驳的：如果老师不事先宣布，他可以安排出人意料的考试；但一旦宣布，他就无法安排出人意料的考试了，该老师必定会自食其言。奥康诺开玩笑式地从该悖论引出一个有关教学的劝告：如果你想给学生一个意想不到的考试，你千万别事先向学生宣布你的意图。奥康诺把老师的宣布比作这样的句子：“我根本不记得任何东西”，“我现在没在说话”。尽管这些句子是一致的，却在任何情景下都不能被设想为真。科恩（L. Jonathan Cohen）把老师的宣布视为语用悖论，并将后者定义为：说出一个句子这件事本身就使得该句子为假。他认为，那位老师没有意识到他的宣告会使该宣告本身为假。

但人们大都认为，上述说法及其分析并不成立。老师的宣布告诉了学生下周有一次考试，一旦宣布，这一点就不再“出人意外”。但老师并没有告诉学生该考试具体安排在哪一天，假如一周 5 个工作日，考试正好发生在从周一至周五中的某一天，这一点仍然是未知的；一旦考试在那天实际发生，仍然是一个“意外”，至少是含有出人意料的要素。前面的分析表明，老师的确可以兑现他的承诺，在任何一天他都可以进行一次“意外”考试！

（3）学生的推理出错？

如果老师的宣布没有什么错，那么，悖论似乎只能归结为学生的推理出错了。

学生的推理是一个归谬推理：假设老师的宣告为真，最后得出不可能有意外考试的结论，由此证明老师的宣告为假。这个论证分为两部分：第一部分，论证考试不能在周五；第二部分，把周五的结论逻辑转移到其他每一天上。关键是第一部分：关于周五的论证真的成立吗？假设老师的宣告为真，周五真的不会有考试吗？不一定吧。到周四为止还没有考试时，学生周五上课时会显得有些迷惑不解，或忐忑不安：怎么还没有考试？老师忘记他的宣告了吗？或者他说话不算数，只是故意吓唬我们？今天该不会有考试吧？实际情形是：老师完全可以

把考试安排在周五，且仍然是一次“意外”！因为老师知道，学生会进行这样的逻辑推理：既然前 4 天没有考试，本周又必有一次考试，则考试只能发生在周五，我们预先就能知道这一点，该考试不再是“意外”考试。所以，周五不会有考试。但老师正好针对学生的逻辑，给他们安排一个逻辑的“意外”：在周五上课时宣布“立即考试”！看起来，学生的推理在第一步就出错了……

但是，且慢下结论！实际上，一周 5 天的条件是无关紧要的，一门课通常也不会连上 5 天。为简单起见，我们考虑一周上两次课的情形吧，比如周一和周四。老师的宣告现在变成：“我将在周一和周四对你们做一次考试，但在考试那天的早晨，你们都没有足够的理由相信当天会有考试，所以，那次考试对你们来说仍然是一次意外。”下面，我们采用一些符号表达式：

M：周一；

T：周四；

E_M：考试在周一；

E_T：考试在周四；

$K_M(\cdots)$：学生在周一知道(…)

$K_T(\cdots)$：学生在周四知道(…)

老师的宣告 A：$[(E_M \wedge \neg K_M(E_M)) \vee (E_T \wedge \neg K_T(E_T))] \wedge \neg(E_M \wedge E_T)$，也可以写成：$[(E_M \vee E_T) \wedge \neg(E_M \wedge E_T)] \wedge (\neg K_M(E_M) \wedge \neg K_T(E_T))$

K(A)：学生知道老师的宣告为真。

我们可有如下证明：

(1) K(A)	假设学生知道老师的宣告为真
(2) $\neg E_M$	假设考试不在周一
(3) $K_T(\neg E_M)$	(2)+ 学生的记忆
(4) $\neg E_M \rightarrow E_T$	根据老师的宣告
(5) $K_T(\neg E_M \rightarrow E_T)$	根据学生足够理性
(6) $K_T(E_T)$	(3)(4)+ 知识的演绎封闭原则
(7) $K_T(E_T) \rightarrow \neg A$	根据老师的宣告
(8) $\neg A \rightarrow \neg K(A)$	知识蕴涵真理原则的逆：只有真理才能被知道
(9) $K_T(E_T) \rightarrow \neg K(A)$	(7)(8) + 推理传递律
(10) $\neg K(A)$	(1)(9) + MP
(11) E_M	(2)(10)(1) + 反证法
(12) $K_M E_M$	根据学生足够理性
(13) $E_M \wedge K_M E_M$	(11)(12) + 合取引入
(14) $E_M \wedge K_M E_M \rightarrow \neg A$	根据老师的宣布
(15) $\neg A$	(13)(14) + MP
(16) $\neg A \rightarrow \neg K(A)$	知识蕴涵真理原则的逆：只有真理才能被知道
(17) $\neg K(A)$	(15)(16) + MP
(18) $\neg K(A)$	(2)(17) + 归谬法

(18)所说的，即使老师做了意外考试的宣告，学生也不知道老师的宣告是真的！否则，由假设学生知道老师的宣告是真的，就可以逻辑地推出学生不知道老师的宣告是真的。蒯因最早发现了这一点，他说：学生的推理并没有证明老师的宣告不可能为真，而只是证明了：学生不可能知道老师的宣告为真[①]。在我看来，这太吊诡了：老师做了宣告，学生怎么会不能知道老师的宣告为真呢？原因是什么？说真的，对这一点我也说不清楚，还是留给读者一起思考吧！[②]

意外考试悖论有很多变体：意想不到的老虎，不可执行的绞刑，选定的学生，因迪悖论，毒药悖论等。

五、知道者悖论

蒙塔古和卡普兰于 1960 年共同发表文章[③]，从对意外考试悖论的分析中提炼出一个新悖论，后来叫做“知道者悖论”（paradox of the knower）。

假如一周五个工作日，老师宣布下周将有一次出其不意的考试，这等于如下断言：

T_1 考试将发生在周一但在周一之前你们将不知道这一点；或者，考试将发生在周二但在周二之前你们将不知道这一点；或者，考试将发生在周三但在周三之前你们将不知道这一点；或者，考试将发生在周四但在周四之前你们将不知道这一点；或者，考试将发生在周五但在周五之前你们将不知道这一点；或者，这个宣布被知道是假的。

蒙塔古和卡普兰认为，T_1 中选言支的数目多少是无关紧要的，可以很多很多，例如多至一个月甚至一年，也可以很少很少，甚至少到一天，甚至少到 0 天。当少到 0 天时，T_1 就变成了 T_2：

T_2：知道这个语句是假的。

如果 T_2 是真的，既然知道 T_2 是假的，由于知识蕴涵真理，所以 T_2 是假的。由于没有任何语句既是真的又是假的，由此我们就证明了 T_2 是假的。由于证明产生知识，故知道 T_2 是假的，而这正是 T_2 所说的意思，所以，T_2 必定是真的。悖论！

这个悖论有说谎者悖论的味道。后来的评论者在对知道者悖论做形式表述时有点漫不经心，没有注意到 $K\neg p$（知道¬p）和 $\neg Kp$（不知道 p）之间的区别，由此导致了知道者悖论的一个变体：

T_3：没有人知道这个语句。

明显可以看出，T_3 就是前面谈到的布里丹悖论语句 1。

怀疑论者希望通过否认人们知道任何东西来消解悖论性语句 T_2，但这个补救办法对于 T_3 并不奏效。如果人们不知道任何东西，则 T_3 就是真的。怀疑论者能够转而攻击证明一个命题是知道该命题的充分条件吗？这个办法甚至对他们本身也是难以接受的，因而他们也是

---

① 蒯因：“论一个假定的二律背反”，载《蒯因著作集》第五卷，涂纪亮、陈波主编，北京，中国人民大学出版社，2007 年，25-26 页。

② 关于意外考试悖论的解读，可参阅 Sainsbury, R. M. *Paradoxes*, third edition, pp.107-115.

③ 蒙塔古和卡普兰：《对一个悖论的再思考》，载蒙塔古：《形式哲学》，第 309—326 页。

通过证明来传播其怀疑论结论的。如果抛弃证明，他们就会变得像他们所嘲讽的独断论者一样。也应该指出，他们想到的这一办法也不是没有一点合理性。

很明显，没有假命题能够被证明为真。但是，有真命题不能被证明为真吗？回答是：有，而且有无穷多。根据哥德尔不完全性定理，任何一个包括算术在内的形式系统都包含一个类似于意外考试悖论中的自指句："本语句在本系统内不可证明。"该类语句叫做"哥德尔语句"。该系统不能证明它的哥德尔句，该语句却是真的；如果该系统能证明它的哥德尔句，该系统就是不一致的，即导致矛盾。所以，或者该系统是不完全的，或者该系统是不一致的。当然，这个结果把可证明性与某个特定的系统关联起来。一个系统能够证明另一个系统内的哥德尔句。哥德尔认为，数学直觉给他这样的知识：算术是一致的，尽管他不能证明这一点；人类的知识不能局限于人类能够证明的东西。有的计算机科学家断言，人不是机器。因为一台计算机就是一个形式系统的具体体现，它的知识就是它能够证明的东西。人却与计算机不同，充分掌握算术的人能够是一致的，即没有矛盾。另有哲学家捍卫人与计算机之间的对等性，他们认为我们有自己的哥德尔句，举例来说，如果我们把意外考试悖论中学生关于考试日的信念作为一个逻辑系统，那么，老师的宣布就是关于学生的一个哥德尔句：下周将有一次考试，但是，你不能根据那位老师的宣布和对该周前几天所发生情况的记忆来证明考试会出现在哪一天。有的评论者指出，把意外（surprise）解释成形式系统中的不可证明性是改变了论题，意外考试悖论更类似于说谎者悖论。如果他们的说法属实，意外考试悖论就不属于认知悖论，而属于与真假概念相关的语义悖论了。

六、摩尔悖论

摩尔发现，在下面的语句中隐含着荒谬或不一致之处[①]：

(M) 上周二我去看了画展，但我不相信我去了。

可以把（M）语句表示为：

（M′）$p \wedge \neg Bp$

M′说，p 但我不相信 p。尽管 M′语句表面上不含逻辑矛盾，但隐含一个矛盾：因为如果一个人自己说出 p，通常意味着他相信 p；如果他又说他不相信 p，则导致矛盾：一个人不可能同时拥有一个信念又不拥有它。这在逻辑上是不可能的。

不过，如果 M′语句中所涉及的不是第一人称，而是第二或第三人称，或者有混合人称，则不会导致矛盾：

（1）周五有考试但你不相信这一点。

（2）张三自杀了但李四不相信这一点。

对（M）还可以有另一种符号化：

（M″）$p \wedge B\neg p$

意思是：p 但我相信非 p。例如，上周二我去看了画展，但我相信我没有去。这里也隐含矛盾，但性质与 M′中的矛盾不同。如果一个人说 p，通常意味着他相信 p；如果他又说他

[①] Moore, G.E. 'A reply to My Critics', in The Philosophy of G. E. Moore, ed. P. A. Schlipp, Evanston, 1942, p.543; 'Russell's Theory of Description', in The Philosophy of Bertrand Russell, ed. P. A. Schlipp, Evanston, 1944, p.204.

相信非 p，这只是表明这个人拥有不一致的信念，不是一个理性的说话者。但在逻辑上是可能发生的，确实有不完全按逻辑说话和行事的人。

因此，（M′）和（M″）是两个不同的悖论。并且，（M″）还有另一种性质上类似的形式：上周二我没有看画展，但我相信我看了，即：

（M‴）$\neg p \wedge Bp$

令 S 是一个形如“p 但我不相信 p”的陈述。雷谢尔把摩尔悖论表示如下：

（1）S 做出了一个有意义的陈述，传达了融贯的信息。（一个合情理的假定）

（2）在做出“p 但是 q”这样的断言时，说话者隐含地表明他接受 p。（一个逻辑-语言事实）

（3）在做出“p 但我不相信 p”这样的断言时，说话者明显地表明他拒绝 p。（一个逻辑-语言事实）

（4）由（2）和（3）可以推出，在做出 S 这样的断言时，该说话者同时表明他接受 p 并且拒绝 p。

（5）（4）和（1）不相容。

雷谢尔给出的办法是：既然从（1）能够推出矛盾句（5），这就说明，在理性交流的语境中，（1）是不能成立的，故应该直接抛弃（1）。[①]

正是维特根斯坦把摩尔语句（M）命名为“摩尔问题”，但他对此却有不同的看法。他认为，M 语句近似于自相矛盾，但并非真的如此。如果我说我（过去）相信 p，我在报道我过去的信念；如果你说我（现在）相信 p，你在报道我当下的信念。但是，如果我说我相信 p，我并不是在报道我的信念，而只是表达了它。如果我简单地说 p，我就是在表达我的信念。如果我说“我相信 p”而不是直接说出 p，我通常是在表达对 p 的某种保留。因此，当我说“p 但我不相信 p”时，我的意思是“p 但或许非 p”。这近似于自相矛盾，但并不真的是自相矛盾。维氏认为，摩尔问题中隐含着关于相信（行为）的深刻洞见，这就是“我相信 p”和“你相信 p”之间不对称。通过听你所说的话以及观察你的行为，我知道你相信什么。但在我能够表达我的信念之前，我不必对我自己做观察。如果有人问：为什么我相信玛丽琳没有自杀，我通常会谈论玛丽琳而不是谈论我自己。如果关于我的信念我有任何理由，不只是把该信念当作一种预感，那么，我相信玛丽琳没有自杀的理由就是玛丽琳没有自杀的理由，我不必把有关自杀的理由与我关于自杀的信念的理由分开。[②]

七、序言悖论

麦金森（D. C. Makinson）于 1965 年构造了“序言悖论”[③]：

一位严肃认真的学者，通常会相信：“我在书中所写的每一句话都是真的”，因为假如他不认为它们为真的话，就不会把它们写进他的书中。但是，他通常又会在序言中，

[①] Rescher, N. *Paradoxes: Their Roots, Range and Resolution*, Chicago and La Salle, Illinois: Open Court, 2001, 44-45.

[②] 维特根斯坦：《哲学研究》，李步楼译，北京：商务印书馆，1996 年，288-293 页。

[③] Makinson, D. C. (1965) "The Paradox of the Preface", *Analysis* 25 (6): 205–207.

在对有关人士如妻子、师友、秘书、编辑表示感谢之后，对书中“在所难免”的错误预先向读者表示歉意。即是说，他相信“我的书中至少有一句话是假的”。麦金森指出，上面两个信念是不一致的。

序言悖论的关键在于:

(1) 下面的信念合取原则是否成立?

$B_i p \wedge B_i q \rightarrow B_i (p \wedge q)$

也就是问，如果一个认知主体 i 相信 p 并且相信 q，他是否相信 p 和 q 的合取?

(2) 信念 $B_i p$ 或 $B_i q$ 是否得到了证成（justified）?

凯伯格拒绝信念合取原则[①]。在这一点上，许多哲学家接受他的看法，并得出结论说：拥有搁在一起不一致的信念并不是不合理的；由此引发一个有关该悖论本性的有意思问题：如果允许不一致的信念的话，悖论将如何改变我们的心智?

一个悖论经常被定义为这样一组命题：单个地看，它们都是合乎情理的；但搁在一起，它们却是不一致的。悖论迫使我们以高度结构化的方式改变我们的心智。例如，关于信念的证成（justification）有下面 4 个命题：

1. 一个信念只能由另一个信念来证成。
2. 不存在（或不允许）循环的证成链条。
3. 所有的证成链条都是有穷长的。
4. 有些信念得到了证成。

许多认识论家认为，这 4 个命题不能同时成立。基础论者拒斥(1),他们认为某些命题或者因为理性的原因或者因为经验的原因是自明的。融贯论者拒斥(2)，他们容忍某些形式的循环推理。例如，古德曼把反思的平衡方法（the method of reflective equilibrium）刻画为“良性循环”。皮尔士拒斥(3)，他相信，既然允许无穷长的因果链条，就应该允许无穷长的证成链条,后者并不比前者更不可能。最后,有些认识论的无政府主义者或者取消论者拒斥(4)，例如取消论者认为，像在格雷林悖论中“非自谓的”（heterological）是一个病态谓词一样，“得到证成的”也不是一个真正的谓词，因此“有些句子是得到证成的”也是一个病态句，没有真假可言。

如果像凯伯格所主张的那样，相互不一致的信念在理性上是可容忍的，这些哲学家为什么还要如此费心地去提供关于序言悖论的解决方案呢？可能的回答也许是：规模效应。如果一对矛盾稀释在一个大的理论体系中，它就不那么触目惊心，因而是可容忍的。但是，如果一对矛盾集中显现在少数几个命题中，那就过于碍眼，因而必须以某种方式消解掉。但这种解释很难行得通。如果容忍相互不一致的信念，就必须在一个理论中容忍明显的或隐含的矛盾。对于这一点必须给出充足的理由。

八、可知性悖论（Fitch 悖论）

[①] 参见 Kyburg, H. *Probability and the Logic of Rational Belief*, Middletown: Wesleyan University Press, 1961.

菲奇（B. F. Fitch）于 1963 年谈到[①]，从关于他的一份手稿（他后来从未发表它）的审稿意见中得知了关于“存在着不可知的真理”的下述证明。据档案记载，这位审稿人就是著名的逻辑学家丘奇（Alonzo Church），其证明可简述如下：

> 假设存在一个真命题，其形式是“p 但 p 不是已知的”。虽然这个句子是不含逻辑矛盾，但认知逻辑的最温和的原则也蕴涵着：这种形式的句子是不可知的。特别是，利用两个最无争议的认知逻辑原理 KE(“知识蕴涵真理”) 和 KD (“知识对合取式分配”) 就足以给出一个简单的证明：某些真理是不可知的。证明如下
>
> （1）$K_i(p\wedge\neg K_i p)$　　假设
> （2）$K_i p\wedge K_i\neg K_i p$　　1, KD
> （3）$K_i p$　　2, ∧消去
> （4）$K_i\neg K_i p$　　2, ∧消去
> （5）$\neg K_i p$　　4, KE
> （6）$K_i p\wedge\neg K_i p$　　3, 5, ∧引入
> （7）$\neg K_i(p\wedge\neg K_i p)$　　1, 6, 归谬法
>
> （7）不依赖于任何假设，是一个必然真理。它所说的是：“p 但 p 不是已知的”是一个不知道的真理。或许用符号把上述证明的结论表述为条件句形式更好：
>
> （8）$\exists p(p\wedge\neg\ K_i p)\rightarrow\exists p(p\wedge\neg\Diamond K_i p)$
>
> 其意思是：如果有现实的未知的真理，则有不可知的真理。菲奇没有觉得（8）有什么特别之处，以至该定理在很长时期内未受到人们的关注。一个相信“万能的（包括全知）上帝（偶然或必然）存在”的有神论者会接受（8）空洞地为真，因为其前件或实际上为假或必然为假。但是，大多数认识论家都承认，存在某些实际上未知的真理，但他们同时坚持认为，所有真理都是可知的。这与（8）矛盾。因为通过逻辑上等值的变换，从（8）可以推出：
>
> （9）$\forall p(p\rightarrow\Diamond K_i p)\rightarrow\forall p(p\rightarrow K_i p)$
>
> 其意思是：如果所有真理都是可知的，则所有真理都是已知的。由于这些学者认为存在着未知的真理，这就否定了（9）的后件，因此他们必须否定（9）的前件：并非所有真理都是可知的。但他们坚持认为所有真理都是可知的，由此导致矛盾。他们不愿意修改自己原来的立场，认为矛盾的根源是认知逻辑定理（8）或（9），遂将其称作“可知性悖论”（the knowability paradox）。

威廉姆森坚决不同意把“存在着不可知的真理”称为“悖论”，认为它是一个经过简洁证明的真命题，只不过与人们的无根据主张“所有的真理都是可知的”相冲突罢了。在《知识及其限度》等论著中，他证明，认知逻辑的如下两个公理不成立：

正内省　$K_i\mathrm{p}\rightarrow K_i K_i\mathrm{p}$　　[若 i 知道 p，则 i 知道自己知道 p]
负内省　$\neg K_i\mathrm{p}\rightarrow K_i\neg K_i\mathrm{p}$　　[若 i 不知道 p，则 i 知道自己不知道 p]

[①] Fitch, Frederic. 1963. “A Logical Analysis of Some Value Concepts”, *Journal of Symbolic Logic*, 28/2: 135–142.

他构造了所谓的反透明性论证，证明人的知识和证据状态并不是完全透明的，一个人并不总是能够知道他的所知和无知，一个人也并不总是能够知道他的证据是什么。从对意外考试悖论的分析中，他所引出的结论也是“存在着不可知的真理”。[①]

在我对他的访谈中，威廉姆森还给出了关于存在不可知真理的另一类型的论证。该论证是这样进行的：他举例说，在 2008 年元月一日，我办公室里书的数目或者是奇数或者是偶数。既然我当时没有数它们，自那时以来已经发生了太多的改变，没有人将会知道该数目是什么。于是，或者“该批书的数目是奇数”总是一个未知的真理，或者“该批书的数目是偶数”总是一个未知的真理。我们能够允许，虽然那些真理总是未知的，却不是不可知的，既然在 2008 年元月一日，某个人能够通过计数我房间里的书，从而知道这两个真理中的某一个。不过，如果“该批书的数目是奇数”总是一个未知的真理，那么，“‘该批书的数目是奇数’总是一个未知的真理”就是一个不可知的真理，因为如果任何人知道“‘该批书的数目是奇数’总是一个未知的真理”，他们因此就知道“该批书的数目是奇数”，在这种情形下，“该批书的数目是奇数”就不会总是一个未知的真理。所以，在这种情形下，他们根本上就不知道“‘该批书的数目是奇数’总是一个未知的真理”（既然知识依赖于真理；整个论证使用了归谬法）。类似地，如果“‘该批书的数目是偶数’总是一个未知的真理”，那么“‘该批书的数目是偶数’总是一个未知的真理”就是一个不可知的真理。于是，无论按哪一种方式，都存在不可知的真理。反实在论者常常把此论证叫做“不可知悖论”，因为他们不喜欢该结论；而在威廉姆森看来，它不是悖论，而是一个出乎意料的从真前提得出真结论的简洁论证。

威廉姆森在回答我所提出的“此类结论是否含有不可知论的意谓”，“如何划出可知的与不可知的界限”等问题时，他解释说：“我的观点确实蕴涵一种有限度的不可知论，在它看来，我们必须承认，存在着某些我们不能知道的真理。不过，也存在着许多我们能够知道的真理——甚至是关于是否存在一个上帝的真理。同一个认识论原则既解释了在某些情形下的无知，也解释了在另外情形下知识的可能性，我看不出对这样的不可知论有什么可反对之处，只要它不会变成怀疑论。在某些非常清楚的情形下，我们知道我们知道一些东西。正内省的失败只是意味着，当我们知道时，我们不能总是知道我们知道；它并不意味着，当我们知道时，我们不能在某时知道我们知道。类似地，负内省的失败只是意味着当我们不知道时，我们并不总是知道我们不知道；它并不意味着：当我们不知道时，我们不能在某时知道我们不知道。我正在解释的论证类型给予我们很多关于可知性与不可知性之间界限何在的知识，但是它们也表明，我们不可能具有关于这种界限何在的完全知识。生活本身就是这样。”[②]

有很多哲学家不同意威廉姆森的看法。他们举例说，由于一阶逻辑的量词有存在含义，逻辑学家可以从“每一个事物都自身等同”这个逻辑原则证明（这个世界上）确实有某些东西存在。大多数哲学家回避这个简单的证明，因为他们觉得，某些事物在这个世界上的存在不能够仅凭逻辑来证明。同样的道理，他们也不愿接受关于存在不可知的真理的证明，因为他们觉得一个如此深刻的结果不可能从如此有限的手段得到。

[①] 参见 Williamson, T. ***Knowledge and its Limits***, Oxford: Oxford University Press, 2000, Ch.4-6, Ch.12.

[②] 陈波：《深入地思考，做出原创性贡献——威廉姆森访谈录》，《晋阳学刊》2009 年第 1 期，第 10 页。

九、独断论悖论

这个悖论是由克里普克 1972 年在剑桥大学所做的一次讲演中提出的[①]，吉尔伯特·哈曼（Gilbert Harman）将其命名为“独断论悖论”。

克里普克谈到，马尔康姆（N. Malcolm）在讨论“知道”一词的强意义时，曾提出一个原则：如果我选择知道某个陈述，作为理性的主体，我应该采取这样一种态度：不让任何进一步的证据去推翻它。克里普克指出，“但是，这似乎不是我们对我们所知道的陈述的态度——也似乎不是一种理性的态度。”[②] 他构造了下面的论证，以证明马尔康姆所建议的原则是不合理的。[③]

(i) 如果认知主体 A 知道 p，并且 A 知道 p 推出 q，基于这些知识，A 将得出结论 q，那么，A 知道 q。

这就是前面谈到过的知识对演绎封闭的原则。

令 p 是任一陈述，其内容如下所述

(ii) p 衍推如下的假设：任何反对 p 的证据都是致人迷误的，即导致假的结论。

如果 p 是真的，任何反对它的证据都是致人迷误的，即会导致假结论¬p.

(iii) A 知道 p，并且 A 知道(ii)。

于是，假设 A 进行了适当的推演，由前提(ii)可以得到结论：

(iv) A 知道任何反对 p 的证据都是致人迷误的。

(iv) 适用于现在和未来的一切证据，特别是未来的证据。这一点看起来已经很奇怪了：仅凭知道一个平常的陈述 p，A 就知道一个概括性陈述：任何反对 p 的未来证据都是致人迷误的。

我们还可以有如下的一般性原则：

(v) 如果 A 知道采取一个 T 型行动导致后果 C，并且 A 特别想避免后果 C，那么，A 会下决心不采取任何 T 型行动。

举例来说，假如 A 知道：如果他打开门，站在门外的某个人就会朝他射击，那么，对他来说，不开门就是一个明智的决定。

现在，令(v)中提到的 T 型行动是“接受反对 p 的证据”，也就是基于未来的证据去怀疑或否定 p，令 C 代表某个假的信念，或者代表失去一个真信念，这两者都是 A 所不想要的。于是，我们可以得出结论：

(vi) A 决心不受任何反对 p 的证据的影响。

(vi) 所表达的是一种典型的独断论态度：为了执着于某个信念，避开或拒绝接受一切不利的证据。某些政治或宗教领导人常利用(vi)去论证，假如他们的追随者或臣民本身并不足够坚强以至能够固守他们所持有的信念，他们作为领袖就应该要求甚至强迫其

[①] Kripke, S. “On Two Paradoxes of Knowledge,” in his *Philosophical Troubles, Collected Papers Volume 1*, pp.27-51.

[②] 同上书，第 43 页。

[③] 同上书，第 43-45 页。

追随者或臣民避免接触某些误导性证据，例如，要求甚至强迫后者不去读某些报纸和书，不要去看电视、上网、听广播等等。有些人甚至不需要强迫，就会自动这样做：例如，某个人是他们心目中的英雄，他们就会自动过滤掉一切有损其英雄形象的不利信息。

克里普克回到马尔康姆的讨论和提议。“我认为，正是从这样一个论证中，关于强意义的‘知道’的想法有可能出现；在某些特殊的情形下，这些结论是真的。但是，如果你打量这些前提和推理，看起来没有假定任何‘超级’意义的‘知道’，仅仅是普通意义的‘知道’。所以，必定有某些东西出了错，问题在于——错在哪里呢？”[①]

哈曼把克里普克的独断论悖论改述如下：

如果我知道 *h* 是真的，我知道任何反对 *h* 的证据就是反对真理的证据；我知道这样的证据是致人迷误的。所以，一旦我知道 *h* 是真的，我就能够不考虑任何未来的不利于 *h* 的证据。[②]

独断论者接受这个推理。对他们来说，知识使探究止步。任何与已知的东西相冲突的“证据”都被作为致人迷误的证据排除掉。这种保守态度跨越了从自信到顽固的界限。为了更形象地说明这种顽固态度，有人构想了下面一种情形，“我”认为：我的车目前在停车场，但他的老实忠厚的朋友派克告诉“我”，他的车目前不在停车场，但“我”不相信，并为其信念“我的车目前在停车场”构造了下面一个连锁论证[③]：

(C_1) 我的车目前在停车场。

(C_2) 如果我的车目前在停车场，而派克提供了我的车目前不在停车场的证据，则派克的证据是致人迷误的。

(C_3) 如果派克报告说他看见一辆看起来是我的车被拖出了停车场，那么，他的报告是一个致人迷误的证据。

(C_4) 派克报告说他看见一辆看起来是我的车被拖出了停车场。

(C_5) 派克的报告是一个致人迷误的证据。

根据假设，“我”只是相信 C_1，即我的车目前在停车场；前提 C_2 是分析真的，并且从 C_1 和 C_2 到 C_3 的推论是有效的，所以，“我”对 C_3 的相信度等同于我对 C_1 的相信度。既然我们假定“我”相信 C_4 有充分的证成，由此可推知，“我”对 C_5 的信念也有充分的证成。类似的论证可以使“我”无视其他进一步的证据，如来自拖车公司的电话，或者当“我”走到停车场却找不到我的车。

对于独断论悖论，哈曼给出了如下诊断：上述悖论性论证完全忽视了实际得到的证据可能会造成的影响。既然我现在知道我的车目前在停车场，我现在就知道任何似乎表明相反的状态的证据都是误导性的。但这并不能确保我可以忽略任何进一步的证据，特别是当那些新证据能够改变我目前的知识状态时。因为得到这些新证据会使我不再知道新证据是致人迷误的。结果是，哈曼否认知识的坚硬性，坚硬性原则说，一个人知道某个结论的必要条件是：

---

① 同上书，第 44 页。

② 参见 Harman, G. *Thought*, Princeton: Princeton University Press, 1973, p.148.

③ Sorensen, R. A. “Dogmatism, Junk Knowledge, and Conditionals,” *Philosophical Quarterly*, 38 (1988), pp. 433– 454.

不存在任何证据会使得一旦他知悉这些证据，就不再有充足的理由去相信那个结论。新知识不能削弱旧知识。哈曼不同意这个原则，他主张，新知识可以削弱旧知识。[①]

十、自我欺骗的悖论

自欺悖论可以归结为如下问题：

自我欺骗如何可能？

两种最常见的欺骗形式是：骗财和骗色。欺骗之所以可能，是因为欺骗者掩盖其真实意图，利用被骗者的某些缺点，允诺给被骗者带来更大的回报，或承诺帮助他们解决其急需解决的难题，来达到欺骗的目的。例如，某公司雇人到处打电话，装作为对方设想的样子：现代社会要学会理财，钱在银行里放着，不升值，反而贬值，因而要注重投资，而我们这里有一些投资品种，在不长的时间内就有很大升值空间，买了以后肯定赚大钱……。有人患上某种绝症，正规医院治不了，病人也治不起，于是转而求助气功大师或民间神医，他们自称有独门特技，祖传秘方，专治疑难杂症，对付癌症尤其有效，并且治好了很多重症病人。这恰好迎合了病人及其家属的心理……。某些女士想找高富帅，或有某些特点的男士，而某男人多少符合一些条件，于是他展开柔情骗术，既骗女士们的情色，也骗她们的钱财。当欺骗者存心欺骗时，他就会有意误导被骗者。其手法一旦被欺骗对象识破，欺骗就不能得手。只有被骗者不知道或者不相信对方在欺骗时，欺骗才有可能成功。

但问题的诡异之处在于：自我欺骗如何可能？当一个人打算自己欺骗自己时，难道他不知道自己的这个意图？难道他没有自我意识或自我反思的能力？自我欺骗的情形却十分常见，几乎到熟视无睹的地步。一个患了绝症的病人，觉得自己没有患绝症，即使患了，也很容易治好，因而仍然乐观向上，但没过几个月就死掉了。某个人才智庸常，在单位没有得到重视和许多好处，但他认为自己在能力、水平、工作绩效等方面都比许多同事强，是单位领导和同事们对自己不公，因而愤愤不平。某个人没有得到某个待遇优厚且有发展前途的工作，他就只想该项工作的种种弊端，最后在他人面前把该项工作贬得一塌糊涂（摘不到葡萄就说葡萄酸）。这样的自我欺骗是如何发生的？其答案只能从如下事实中去寻找：人虽然是理性的动物，但又不全是理性的，他或她还有本能、欲望、需要、情感、情绪……，这些东西可以归诸于人的“非理性”的一方面。欲望和情感会产生信念，凯撒说过，“一般而言，人们愿意相信他们希望得到的东西。”一个人把自己的欲求和情绪投射到周围环境之中，也投射到自己眼中的自己，他有选择地看，有选择地听，有选择地读，他“看到”他想看到的，他“相信”他所渴求的，把愿望当作真实，然后对周围环境和对自己本身持有虚假的信念，这就是自我欺骗。所以，按我的理解，自我欺骗的根源在于人类本身潜在的非理性方面。我们先前的文化过于强调了人的理性方面，是弗洛伊德、叔本华、尼采这样的思想家提醒我们注意到人的非理性的方面，注意到人的动物性遗传和本能的黑洞。我曾经写道：“人性确实非常之复杂，复杂到连当事人自己都无法控制和把握，有时候都会感到担心和害怕。”[②]

[①] 参见 Harman, G. *Thought*, p.149.

[②] 陈波：《与大师一起思考》，北京大学出版社，2012 年，第 322 页。

十一、自我修正的悖论

在美国宪法中，第五条的内容如下：

> 举凡两院议员各以三分之二的多数认为必要时，国会应提出对本宪法的修正案；或者，当现有诸州三分之二的州议会提出请求时，国会应召集修宪大会，以上两种修正案，如经诸州四分之三的州议会或四分之三的州修宪大会批准时，即成为本宪法之一部分而发生全部效力，至于采用那一种批准方式，则由国会议决；但一八〇八年以前可能制定之修正案，在任何情形下，不得影响本宪法第一条第九款之第一、第四两项；任何一州，没有它的同意，不得被剥夺它在参议院中的平等投票权。

这个条文是美国宪法的一部分，它规定了如何修正它作为其中一部分的美国宪法的程序和规则。问题是：根据该条文，能够修正该条文本身吗？由此可提炼出如下的“自我修正的悖论”：

> 即使一条制度性规则提供了在某些特定条件下修正该制度的措施，合法地修正该规则本身似乎是不可能的。但这与公认的法律实践相冲突。如何解释这一点？

斯堪的纳维亚法学家罗斯（Alf Ross）认为，在美国宪法第五条中出现了部分的自我指称，而无论是部分的自我指称还是完整的自我指称，都会使它们身处其中的命题失去意义。因此，因为第五条中出现部分的自我指称，故它是一条被剥夺意义的规定。并且，不能根据第五条所规定的法律程序去修正第五条。不过，他对第五条提出了一个补救措施，给它增加一条规范，使它最后读起来像这样：“……服从由第五条所设定的权威，直至整个权威本身任命了它的继任者；然后，服从这个新权威，直至它任命了它的继任者；如此这般地继续下去。”

不过，关于罗斯对美国宪法第五条的看法，存在着很多的争议。自我指称分为两种：恶性的和良性的，例如，“本命题是假的”中的自我指称是恶性的，该命题为真当且仅当该命题为假，这是矛盾！“本命题是真的”却是良性的，假设它为真，它就为真；假设它为假，它就为假；并没有矛盾。美国当代法学家哈特（H. L. A. Hart）给出了如下表列：

（1）草是绿色的。
（2）本表列中的每一个陈述都是真的，包括本陈述在内。

他认为，（2）是无可非议的，故美国宪法第五条也无可非议；罗斯对该条文提出的补救措施既无必要，也不可行。[①]

十二、一些认知逻辑的趣题

（1）白帽子问题

[①] 参见 Clarke, M. *Paradoxes from A to Z*, second edition, pp.200-202.

老师让三个同学坐在一条直线上，甲可以看到乙和丙，乙可以看到丙但看不到甲，丙既看不见甲也看不见乙。让三个同学闭上眼睛，给他们各自戴上一顶帽子，并告诉他们：他们中至少有一人戴白帽子。当他们睁开眼睛，老师问甲是否知道他是否戴白帽子，甲说不知道；又问乙同样的问题，乙也说不知道。老师问丙，丙说知道了，我戴白帽子。丙是怎么知道的？

解析：根据题意，已知：

（1）甲乙丙中至少一人戴白帽子。

（2）甲知道乙、丙是否戴白帽子。

（3）乙知道丙是否戴白帽子。

（4）甲乙丙都知道以上三点，而且都知道别人也知道。

（5）甲不知道自己是否戴白帽子。

（6）乙知道（5）。

（7）乙不知道自己是否戴白帽子。

（8）丙知道以上三点。

求证：丙知道丙戴白帽子。

（9）设乙、丙都不戴白帽子。由（2），甲知道这一事实。由（4）和（1），甲又知道三个人中至少一个戴白帽子。于是，甲应知道自己戴白帽子，与（5）矛盾。可见乙、丙中至少一人戴白帽子。

（10）由（4）及（6），在（9）中我们所做的推理乙也可以做，所以，乙应知道乙、丙中至少一人戴白帽子。

（11）设丙不戴白帽子。由（3），乙知道这一点，再由（10），乙应知道自己戴白帽子，与（7）矛盾。可见丙戴白帽子。

（12）由（4）、（8），以上（9）（10）（11）三点中的推理，丙也可以做出。可见丙知道自己戴白帽子。证毕。

（2）三个聪明人

国王想知道他的三个聪明人中谁最聪明，就在每个人前额上画了一个点，并且说：他们中至少一人额上有白点，并重复地问他们“谁知道自己额上点的颜色？”他们两次都同时回答说“不知道”。求证下一次他们全都说知道，而且所有的点都是白色的。

解析：假如只有一人额上有白点，那么第一次问时，该有白点的人在看到另外两人没有白点时，就应该回答“知道”。所以，不止一人额上有白点。

假如只有两人额上有白点，有白点的人看到另一个人额上无白点，就能够推知自己额上有白点，所以他应该回答“知道”。所以，三个人额上都有白点，当第三次问时，他们都会回答说：我知道自己额上有白点。

（3）七个玩泥巴的孩子

同一同一个教室中有 10 个孩子。其中，有 7 个孩子额上沾有泥巴。每个孩子都能看到

别的孩子额上是否有泥巴，但无法看到自己的。这时老师走进教室对孩子们说：“你们之中至少有一人额上有泥巴”。然后，他问：“谁知道自己额上有泥巴？知道的请举手。”他如是连续问了六遍，无人举手，当问到第七遍的时候，所有额上有泥巴的孩子都举起了手。假设所有的孩子都有最佳的逻辑分析能力，请问他们是如何思考并得出结论的？

解析：假设只有一个孩子额上有泥巴，那么，在老师第一遍提问时，他就会举手，因为他看到除他之外所有的孩子额上都没有泥巴，既然至少有一个孩子额上有泥巴，那么这个有泥巴的孩子自然是自己。

假设有两个孩子额上有泥巴，他们都看到并且只看到一个孩子额上有泥巴，当老师第一遍提问时，他们无法确定自己是否有泥巴因而都不举手，但是当老师的第一遍提问结束后，他们立即都明白自己额上有泥巴，因为如果自己额上无泥巴，则说明只有一个孩子有泥巴，在老师第一遍提问后这个唯一有泥巴的孩子就会举手。这样，当老师第二遍提问时，两个有泥巴的孩子都举起了双手。

同理，如果有三个孩子额上有泥巴，他们就会根据第二遍提问时无人举手而立即判断出自己额上有泥巴，因而在第三遍提问时举手。

因此，一般地，额上沾泥巴的孩子的人数，正好等于他们都举手时老师提问的次数。

（4）“S 先生和 P 先生”问题

S 先生、P 先生都具有足够的推理能力。一天，他们正在接受推理面试。他们知道桌子的抽屉里有如下 16 张扑克牌：

红桃 A、Q、4；

黑桃 J、8、4、2、7、3；

草花 K、Q、5、4、6；

方块 A、5。

面试者从中挑出一张牌，并将其点数告诉 S 先生，将其花色告诉 P 先生。然后，他问 S 先生和 P 先生：你们能推知这是一张什么牌吗？

S 先生：“我不知道这张牌。”

P 先生：“我知道你不知道这张牌。 ”

S 先生：“现在我知道这张牌了。”

P 先生：“我也知道了。”

请问：这张牌是什么牌？

解析：由 S 先生的第一句话，可以推知这张牌的点数并非只有一张的，因此黑桃 J、8、2、7、3，草花 K、6 被排除，余下可能的是：红桃 A、Q、4，黑桃 4，草花 Q、5、4，方块 A、5；P 先生仅凭花色就知道，S 先生仅凭点数无法猜出这张牌，那说明这个花色的牌的点数不能只出现一次，所以，该花色的牌不是黑桃和草花；一定是红桃或者方块的某张牌，因此可能为红桃 A、Q、4，方块 A、5；必然不是 A，否则即便是 S 先生知道点数，也无法猜出到底是红桃还是方块，因此只可能是红桃 Q、4，方块 5；如果是红桃，P 先生最后也是无法猜出这张牌的，因为红桃到最后还是有 2 张，他无法确定到底是哪一张；所以只可能是方块，P

先生才会自信地说他也知道了。答案是：方块 5。

十三、盖梯尔问题及其解答

什么是知识？西方哲学家的传统看法是，知识就是有证成的真信念（justified true belief，简记 JTB）。它最早出现在柏拉图的《美诺篇》和《泰阿泰德篇》中。按这种看法，一个认知主体 S 知道 p，当且仅当：

(i) p 为真，

(ii) S 相信 p，

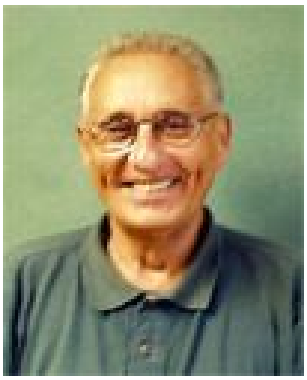
(iii) S 相信 p 是有证成的。

这是关于知识的“三元定义”，其中三个条件都是认知主体 S 拥有知识 p 的必要条件。

(i) 是成真条件：如果某个命题事实上是假的，你不可能知道它是真的。这反映了柏拉图的观念：知识即真理，它在西方哲学传统中根深蒂固。但是，如果某个命题实际上是假的，你却有可能相信它是真的。错误的信念表明你的主观认知状态和客观的事实状况之间的断裂。

(ii) 是信念条件。“信念”这个术语表示一个人按照认知证据来断言某个东西的强烈倾向，此条件要求在认知主体和认知对象之间必须有某些恰当的正面联系。“知识”是一个表示敬意的术语——把知识赋予某个人是要给予他的意见以一种有力的正面的认知地位。知识与信念是相联系的，一个人不可能对他所不相信的事情拥有知识。例如，当我们说某人知道地球围绕太阳转时，我们必已认定他相信地球围绕太阳转。说某人知道地球围绕太阳转而又不相信地球围绕太阳转是令人难以置信的。因此，知识是一种信念，是对一个命题为真所持有的一种信念。

(iii)是证成条件：有资格成为知识的信念必须是一个具有充分根据的信念。信念不同于真理或知识本身，信念可以具有主观上的确定性，但知识具有客观上的确定性，你可以持有假的信念，但假的知识就像是一个奇谈怪论。并且，知识并不简单就是对某个真命题持有信念。有些信念的真只是由于幸运猜测的结果。要使关于一个真命题的信念上升为知识，必须要求人们对于该真命题所持有的信念必须是有充足理由的。



盖梯尔（E. L. Gettier，1927—），有传奇色彩的哲学家。1961 年，在康奈尔大学获哲学博士学位，其导师是布拉克 (Max Black) 和马尔康姆(Norman Malcolm)。1957-67 年，任教于美国韦恩州立大学，其同事有认识论家莱尔(Keith Lehrer) 和逻辑学家兼哲学家普兰廷加（Alvin Plantinga）。据普兰廷加回忆，1962 年某天下午，他与盖梯尔一起喝咖啡。后者谈到，自己连一

篇文章也未发表过，故很担忧明年能否获得教授职位。普兰廷加鼓励他写一点东西发表，以对付管理部门的要求。盖梯尔于是谈到，他有一个想法，即提出一些与传统知识定义相反的小例证。后来他写成一篇短文，但自己并不看好它。有人先把它译成西班牙文，发表在一个不知名的南美刊物上；1963 年，该文在国际哲学期刊《分析》上发表，短短 3 页。它提出的问题后来以“盖梯尔问题”著称于世，极大地影响了 20 世纪后半期认识论的发展。此后，盖梯尔再未发表任何论著，这篇 3 页短文就成为他唯一的出版物。据说，在其教学活动中，他擅长于向研究生传授如何在模态逻辑中找反模型以及为各种模态逻辑构造简化语义学的新方法。1967 年后，在美国马萨诸塞大学阿默斯特分校任教授，现为该校荣誉退休教授，研究领域包括：语言哲学，形而上学，模态逻辑，形式语义学。

上述三个条件都是知识的必要条件，这看起来是确定无疑的。问题是，它们合起来是否就是知识的充分条件呢？换句话说，有证成的真信念（JTB）就是知识呢？对于这个问题，哲学家们过去的回答是肯定的。不过，这一传统的知识定义却受到盖梯尔的挑战，在一篇仅 3 页的短文中，他提出了两个反例，用以证明 JTB 只是知识的必要条件，而不是其充分条件。

史密斯反例

史密斯和琼斯都在申请某一份工作。假设史密斯有证成地相信下列命题：

(a) 琼斯将得到这份工作并且琼斯的衣服口袋里有 10 个硬币。

他相信命题(a)的证据或许是：公司经理已经告诉他，公司将雇用琼斯。而他在十分钟前由于某种原因亲手数过琼斯衣服口袋里的硬币。再假定，史密斯由命题(a)正确推出了命题(b)：

(b) 将得到这份工作的人的衣服口袋里有 10 个硬币。

再进一步设想，后来真正得到这份工作的人其实是史密斯本人而不是琼斯，而且史密斯自己的口袋里恰好也有 10 个硬币，只是他自己不知道。那么，尽管命题(a)是假的，但史密斯由之推出的命题(b)却是真的。于是，对史密斯来说，

(i) (b)为真；

(ii) 史密斯相信(b)；

(iii) 史密斯相信(b)是有证成的。

但使，根据常识，史密斯并不知道(b)，(b)不构成他的知识。

福特车反例

假设史密斯有证成地相信下列命题：

(c) 琼斯有一辆福特牌轿车。

史密斯相信命题(c)的理由可能是：在他的记忆中，琼斯一直开一辆福特车，并且他还借用过琼斯的这辆福特车。假定史密斯还有另一个朋友叫布朗，史密斯已多年不知道他的下落。再假定史密斯任意选择了三个地方作为对布朗下落的猜测，并由命题(c)推出了下列命题：

(d) 琼斯有一辆福特车，或者布朗在波士顿。

(e) 琼斯有一辆福特车，或者布朗在巴塞罗那。

(f) 琼斯有一辆福特车，或者布朗在布加勒斯特。

由于命题(d)、(e)、(f)都是从命题(c)推出来的，所以史密斯相信其中任何一个命题都是有证成的。

再进一步设想，有另外两个偶然成立的事实：

（1）琼斯并没有一辆福特车，他开的那辆福特车实际上是租来的；

（2）命题(e)所提到的地方（巴塞罗那）碰巧是布朗所在的地方。

在这种情况下，尽管命题(e)是史密斯的 JTB，即：

(i) (e)为真；

(ii) 史密斯相信(e)；

(iii) 史密斯相信(e)是有证成的。

但是，根据常识，史密斯并不知道(e)， (e)不构成他的知识。

下面再列举由其他哲学家给出的三个类似反例。

田里的羊反例

齐硕姆（R. M. Chisholm）谈到[①]：假设对 S 来说，命题 p“我看见田里有一只羊”是假的，但他相信 p 却是有证成的，因为他把田里的一条狗误看作一只羊了。于是，他相信命题 q“田里有一只羊”也是有证成的，因为 q 可以从 p 推出来。再进一步假定，碰巧有一只羊在田里，只是 S 没有看见它。在这种情况下，显然没有理由说 S 知道 q。但是，q 却符合传统的知识定义：q 是真的；S 相信 q；S 相信 q 是有证成的。

纵火犯反例

斯基姆（B. Skyrms）谈到[②]，有一名纵火犯右打算烧掉一幢大楼。他的衣兜里装着一盒火柴。由过去的多次经历，他有可靠的证据表明，他的这种火柴是管用的，从不误事。他看到今天天气不错，干湿度刚好，要烧掉那幢大楼，他相信只需用掉一根火柴。事实也证明确实如此：该纵火犯划亮了一根火柴，点燃了一堆易燃物，烧掉了那幢大楼。但他没有认识到，这一切事情纯属碰巧：他的那盒火柴里混进了助燃剂，否则，在那种情况下他是无法划亮那根火柴的。于是，纵火犯的信念“只需用掉一根火柴”就是 JTB：他确实只用掉一根火柴，他相信这一点，他的这个信念是有证成的。但是，纵火犯的这个信念真的是知识吗？

假谷仓反例

哥德曼设想了这样一种情景[③]：亨利一边在乡野中开车，一边打量其中的对象。他看见一个看起来与谷仓一模一样的对象。他没有理由怀疑他所看到的东西，故他认为他看见了一座谷仓。但是，他没有意识到，邻近地区在拍摄电影，野地里有很多假谷仓，它们实际上是谷仓画板；当人们开车经过这里时，会真的把它们看作是谷仓。但凑巧的是，野地里恰好有一座真谷仓。所以，命题 p“野地里有一座谷仓”是真的，亨利相信 p，他相信 p 是有证成的。又遇到那个老问题：亨利关于谷仓的信念真的构成知识吗？

可以把以上 5 个反例都叫做“盖梯尔反例”，因为它们有以下共同的论证结构：

(1)S 相信 p	知识的条件 1
(2)p 是真的	知识的条件 2
(3)S 相信 p 是有证成的	知识的条件 3
(4)p 是从 q 演绎得到的	逻辑
(5)S 相信 q 是有证成的	经验事实

---

① 参见齐硕姆：《知识论》，邹惟远等译，北京：三联书店，1988 年，第 45 页。

② Skyrms, B. “The Explication of ‘*X* Knows that *p*’,” *Journal of Philosophy* 64 (1967): pp.373-89.

③ Goldman, A. I. “Discrimination and Perceptual Knowledge,” *Journal of Philosophy* 73 (1976): pp.771-91.

(6)q 是假的　　经验事实

所以，

(7)S 不知道 p

在所有的盖梯尔反例中，还可以发现以下两个共同因素：

（1）可错性。在每一个反例中，所展示的证成都是可错的。虽然对所论及的信念都提供了证成，但严格说来，证成并不完美。这意味着，证成留下了这样的可能性：所论及的信念是假的。虽然证成很强地表明，该信念是真的，但没有完全证明这一点。

（2）碰巧或幸运。所有的盖梯尔反例有一个显著特点：它们都含有运气（luck）成分：一个得到很好证成但依然可错的信念碰巧是真的;某种运气把该信念为真与其有证成结合在一起。而正常的认知语境中没有那么多的运气。

有些哲学家们通过对盖梯尔反例的分析，认为它们需要假定如下三个原则：

第一，人们能够依据假理由而相信一个命题。

有论者指出：“盖梯尔类型的反例全都依赖于这样一个原则：某人能够有理由依据 p 接受某个命题 h，即使 p 是假的。”就史密斯反例而言，史密斯相信命题(a)的依据是公司经理如是说。既然公司经理自己弄错了或后来改变了主意，他对史密斯先前所说的话就是假的，故史密斯相信命题(a)的理由也是假的。

第二，人们能够有证成地相信一个假命题。

仍就史密斯反例而言，史密斯以公司经理说的话为依据而相信命题(a)，而(a)事实上是假的；就福特车反例而言，史密斯以琼斯常开一辆福特车以及琼斯还让他用过这辆车等为理由而相信命题(c)，但是 (c)事实上是假的。

第三，在有效推理中，证成能够从前提传递到结论。即是说，如果人们相信一个命题是有证成的，则他相信由该命题合乎逻辑推出的任何命题也是有证成的。

就史密斯反例而言，由于史密斯从命题(a)合逻辑地推出了命题(b)，而史密斯有理由相信 (a)，故他有理由相信(b)；就福特车反例来说，由于命题(e)是史密斯由命题(c)合逻辑地推出来的，而史密斯有理由相信(c)，故他也有理由相信(e)。

显然，盖梯尔反例成立与否，与这三个原则有密切关系。但有不少学者论证说，这三个原则是不能接受的。

解决盖梯尔问题的策略大致可分为两种：一种是加强知识定义中的证成条件以排除盖梯尔反例。例如，齐硕姆在《知识论》第六章中就是如此做的。另一种策略是加入适当的第四个条件以补救对知识的 JTB 分析，新加入的条件会防止 JTB 被盖梯尔反例消解掉。经如此补救之后，关于知识的三元分析 JTB 就变成了四元分析：JTB + X，其中 X 代表所需的第四个条件。我们下面概述几种 JTB + X 的方案。

方案 1：不含假前提的证成

从对所有盖梯尔反例的结构性分析中，我们发现：它们都暗含至少一个假前提，S 对信念 p 的证成就来源于这个假前提。由此自然产生一个想法：我们可以在对知识的 JTB 分析中增加一个条件，即不含假前提。由此得到关于知识的 NFP（no false premise）理论：

(i) p 为真；

(ii) S 相信 p；

(iii) S 相信 p 是有证成的；

(iv) S 对 p 的证成不依赖于任何假前提。

不过，NFP 理论所增加的条件(iv)只是作为一个限制性条件起作用。但研究表明：在有些情况下，该限制条件太弱，允许把非知识误当作知识；在有些情况下，该限制条件又太强，能把确定无疑的知识排除在知识之外。[①]

方案 2：不可挫败的证成

这种观点认为，也许知识所要求的不是一个人的信念得到恰当证成，而是他的信念不应被他目前没有意识到的任何真实证据所削弱或挫败。由此，我们得到 ND（no defeater）理论，其给“知识”新增的第四个条件是：

(iv) S 对 p 的证成不会被任何真命题所挫败。

但问题在于，在持有一个信念时，该信念可以恰当地得到证成，即使它有整体的反证据。按照 ND 理论，为了成为知识，对一个人的信念的证成必须是根本上不可挫败的。“根本上”意味着他的证成没有反证据，或者所有的反证据相互抵消。但是，如果我们对一个信念或假说能够持有的证据是无限开放的，在何时何地我们能够达到该信念的一个根本上不可挫败的证成？这一点是不清楚的，甚至也是不可能的。[②]

方案 3：对知识的因果分析

这种观点认为，把知识与真信念相区别的不是证成，而是信念的因果联系，这些因果联系把该信念与它所关涉的事件联系起来：如果一个真信念有正确的因果联系，就是知识；有错误的因果联系，就不是知识。关于知识的因果论分析给“知识”增加的第四个条件是：

(iv) S 知道 p，当且仅当，事实 p 以某种恰当的方式在因果上与 S 相信 p 相关联。[③]

但因果是一个时空范畴，我们持有的许多信念在根本上与特定的事件无关，也与能够与之有因果联系的东西无关。如果在这些情形中可以说我们具有知识，则对知识的因果分析必定失败。例如，我知道 22337 是素数，这个知识就没有与之相关的因果过程。因此，关于知识的因果概念至少太狭窄，不适合作为知识的一般定义；并且，它还面临不正常因果链（例如假谷仓）的挑战。[④]

方案 4：知识即追踪实在的真信念

诺齐克认为[⑤]，一个信念要成为知识，它必须对所相信命题的真值特别敏感；更明白地说，它必须追踪真理。因此，知识就是追踪真理的信念。如果一个命题在稍微变化了的情景中仍然为真，我们就仍然相信它；如果该命题在稍微变化了的情景中不再为真，我们就不再相信它。于是，“S 相信 P”被刻画为以下 4 个条件的合取：

---

① 参见 J·丹西：《当代认识论导论》，周文彰等译，中国人民大学出版社，1990 年，第 30—32 页。

② 参见同上书，第 32-33 页。

③ Goldman, A. I. “A Causal Theory of Knowing,” *Journal of Philosophy* 64 (1967), pp.357-72.

④ 参见丹西：《当代认识论导论》，第 37—38 页，第 52—55 页。

⑤ 参见 Nozick, R. *Philosophical Explanation*, Cambridge, MA: Harvard University Press, 1981, pp.172-78, pp.197-227.

(i) p 是真的；

(ii) S 相信 p；

(iii) 倘若 p 真，S 相信 p；

(iv) 倘若 p 不真，S 不相信 p。

(iii)和(iv)在英语中是以反事实条件句（或虚拟条件句）的形式出现的。文献中常把(iv)称之为知识的“敏感性”（sensitivity）条件，而把(iv)的逆否命题

(v) 假若 S 相信 p，p 就不是假的

叫做知识的“安全性”（safety）条件。由于(p→q)→(¬q→¬p)这一推理形式对于反事实条件句不成立，因此(iv)和(iv)并不等价。对于知识来说，“安全性”条件是在“敏感性”条件之外另加的要求。不过，关于知识的这一看法也遭遇到一些困难和挑战。[①]

方案 5：可靠主义的知识论

可靠主义者认为，为了把一个信念转变成知识，不需要用恰当的证据为它提供证成，只要求该信念是通过可靠的认知过程或方法产生的。“如果一个信念要算作知识，它必须是由一个一般来说是可靠的[认知]过程引起的。”[②]于是，他们把“知识”刻画为如下三个条件的合取：

(i) p 是真的；

(ii) S 相信 p；

(iii) S 的信念 p 是通过可靠的认知过程或方法产生出来的。

但问题在于：何谓“获得知识的过程或方法”？如何把获得知识的过程或方法划分为可靠的和不可靠的？这样划分的根据是什么？真的存在完全可靠的方法吗？丹西指出，“……看起来不大可能有完全可靠的获得信念的方法。人是易犯错误的”[③]。另外，可靠的过程或方法也不必然产生知识，仅仅偶然地产生真信念。

威廉姆森的反叛：知识第一位

在《知识及其限度》一书中，威廉姆森指出，自盖梯尔证明 JTB 对于知识不是不充分条件以来，认识论学家付出了巨大努力，试图说出知识究竟是哪一种真信念，迄今为止进行了成百上千种这样的尝试，但全都失败了；而且，通过找出知识的多个必要条件，例如信念、真、证成以及 x ，就能找出知识的非循环的充分必要条件，这一假定是错误的。举例来说，“是有颜色的”是“是红色的”的必要条件，但是，如果有人问，给“是有颜色的”加入什么样的条件才能成为“是红色的”？只能回答说：除了加入“是红色的”之外别无他法。同样的道理，我们也没有理由认为，把知识的多个必要条件合取起来，就能找到知识的非循环的充分必要条件。等式“红色＝有颜色＋X”和等式“知识＝真信念＋X”都不必然有一种非循环的解答。简而言之，根据信念等等去诠释、说明、分析、定义知识的方案是行不通的。

威廉姆森所提出的替代方案是：“知识第一位”（knowledge the first），即把“知识”概

---

① 参见 “Analysis of Knowledge”, in http://plato.stanford.edu/entries/knowledge-analysis/#ModCon. 读取日期：2013 年 8 月 19 日。

② Goldman, A. I. Epistemology and Cognition, Cambridge, MA: Harvard University Press, 1986, p.51.

③ 丹西：《当代认识论导论》，第 35 页。

念作为不加诠释的基本概念，用它去说明、分析、定义“信念”等其他认知现象。认知系统的功能就是生产知识；当它发生故障时，它生产纯粹的信念，这样的信念是有缺陷的，并不构成知识，典型的是假信念，也包括碰巧为真的信念。如果某人知道事情是如何，他就相信事情是如何；但是，如果他仅仅相信事情是如何，他并不知道事情是如何。单纯的相信要相对于知道加以理解，误感知要相对于感知加以理解，误记忆要相对于记忆加以理解，就像发生故障要相对于正常起作用来加以理解一样。特别地，相信要被理解为这样的心智状态，它对于作为其特殊状态的知道具有类似的直接效果。于是，根据其直接的先行状态对行动做因果解释，经常要合适地诉诸信念而不是知识，即使当认知主体事实上知道的时候也是如此。

概而言之，在威廉姆森看来，知识是核心的而非从属于信念。知识为信念设定规范：一个直率的信念得到充分的证成，当且仅当它构成知识。既然对信念的语言表达是断定，知识也为断定设定规范：一个人应该断定某事如何，仅当他知道某事如何；或者说，一个人应该断定 p，仅当他知道 p。威廉姆森的新反叛在当代认识论研究中激起了非常大的反响。[①]

十五、图灵测试和塞尔的“中文屋论证”

1980 年，约翰·塞尔（John Searle）在《行为和脑科学》杂志上发表了《心灵、大脑和程序》一文，其中提出了中文屋论证，并答复了 6 个主要的反对意见，它们是他先前在很多大学做报告时遇到的。与该文同时发表的，还有 27 位认知科学家的评论和批评。1984 年，塞尔在其专著《心、脑和科学》中再次阐述了中文屋论证。1990 年 1 月，通俗期刊《科学美国人》将这一争论带给了大众读者。该期发表了塞尔的文章《大脑的心灵是计算机程序吗？》和丘奇兰德夫妇（Paul and Patricia Churchland）的论辩文章《机器能思维吗？”》。在 20 世纪最后 20 多年间，中文屋论证是众多论战的主题，围绕它发表了难以计数的学术论文。中文屋论证的主旨是反驳强人工智能断言：运行程序的数字计算机已经、至少能够像人一样有意识和能思考。

（一）智能和人工智能

1．智能、意识和心灵

什么是人的“智能”（intelligence）？这是一个很有争议的问题，学界尚未达成共识。粗略地说，“可以认为智能是知识和智力的总和。其中，知识是一切智能行为的基础，而智力是获取知识并运用知识求解问题的能力，即在任意给定的环境和目标的条件下，正确制订决策和实现目标的能力，它来自人脑的思维活动。”[②] 智能包括：(1) 感知能力，指人们通过视觉、听觉、触觉、味觉、嗅觉等感觉器官感知外部世界的能力。(2）记忆与思维的能力，这是人脑最重要的功能。记忆用于存储由感觉器官感知到的外部信息以及由思维所产生的知识；思维用于对记忆的信息处理，即利用已有的知识对信息进行分析、计算、比较、判断、推理、联想、决策等。思维是一个动态过程，是获取知识以及运用知识求解问题的根本途径。（3）从经验中学习并有效适应环境的能力。（4）行为或表达能力，指人们用语言或某个表情、眼神和肢体动作来对外界刺激做出反应，传达某个信息的能力。如果说人们的感知能力

[①] 参见陈波：《知识优先的认识论》，载陈波：《与大师一起思考》，第 177-188 页。
[②] 王永庆：《人工智能原理和方法》，西安：西安交通大学出版社，1998 年，第 2 页。

用于信息的输入，行为或表达能力则用于信息的输出，它们都受到神经系统的控制。

另一个相关的问题是：什么是人的“意识”(consciousness)？学界对此也尚无共识，正在深入探究中。大致说来，意识是动物的神经反应，当动物或人出生时意识就与生命同在，是一种自我感受、自我存在感与对外界感受的综合体现，塞尔将“意识”泛指为“从无梦的睡眠醒来之后，除非再次入睡或进入无意识状态，否则在白天持续进行的知觉、感觉或觉察的状态”[①]。根据心理学研究，意识具有四个特性：意向性，统一性，选择性和流动性。意向指人们对待或处理外在事物的活动，表现为欲望、愿望、希望、意图等。意向是个体对外部对象的反应倾向，即行为的准备状态，准备对外部对象做出一定的反应，因而是一种行为倾向，故亦称“意图”、“意动”。意向性（intentionality）是指人的意识通常指向或关涉某个事物或某件事情。统一性是指各种觉知形式都被整合成一个同一的、整体的、独特的、连贯的意识经验。选择性指人能注意到某些事情，却没有注意到另外的事情。例如，在一次鸡尾酒会上，某人提到你的名字，当时你和那个人都在分别同时与不同的人群聊天，但你却注意到了他（她）提到你的名字。短暂性是指意识的内容是不断变化的，从来都不会静止不动。美国心理学家詹姆士（William James, 1842—1910）提出了“意识流”这一概念。还有人概括出意识的另一特征——能动性，表现在三个方面：与环境的互动；把过去的经验与现在相连接，形成自我同一性的基础；制定目标，引导行为。

还有一个更困难的问题：什么是“心灵”(mind)？大致说来，“**心灵**”是指一系列认知能力组成的总体，这些能力可以让个体具有意识、感知外界、进行思考、做出判断以及记忆事物。心灵是人类的特征，但其它生物也可能具有心灵。围绕心灵的本性所产生的最长久且最激烈的哲学争论就是心-身问题，即心灵与作为人的身体一部分的大脑或神经系统之间的关系：心灵能否独立于人的身体而存在？若回答“能”或者“不能”，其理由和根据是什么？它们合理和充足吗？围绕这些问题，主要有两种哲学立场：二元论和一元论。二元论有不同的形式：实体二元论主张，心灵和身体各自独立存在；属性二元论认为，心灵是大脑所显现的一种独立属性，不能还原到大脑，但也不是实体性存在。一元论主张，心灵或身体中只有一个是基础性的，另一个则是依附性或派生性的。观念论者主张，心灵是全部的真实存在；物理主义者坚持认为，只有物质性的大脑才是真实的存在；中立一元论断言，另有一种中立的实体，物质和心灵都是这种实体的属性。在 20 和 21 世纪，最常见的是各种牌号的物理主义，包括行为主义、同一理论和功能主义。有一门新兴的哲学分支——心灵哲学(philosophy of mind)，研究心灵的本性、心智事件、心智功能、心智属性、意识，以及它们与物质性身体特别是大脑的关系。

2．人工智能：

所谓“人工智能”(Artificial Intelligence，缩写为 AI)，指用人工方法在机器（计算机）上实现的智能，或者说是人类智能在机器上的模拟，亦称“机器智能”。就其研究对象而言，人工智能是一门研究如何构造智能机器(智能计算机)或智能系统，使它能够模拟、延伸、扩展人类智能的学科。也有更简洁的说法，“人工智能是关于知识的学科－－怎样表示知识以及怎样获得知识并使用知识的科学。”“人工智能就是研究如何使计算机去做过去

[①] Searle, J. “Minding the Brain,” review of Nicholas Humphrey, *Seeing Red*, The New York Review of Books, November 2, 2006, p.51.

只有人才能做的智能工作。”

1956 年夏，在美国达特茅斯大学，由 10 位科学家组成的一个研究小组举行了为期 2 个月的学术会议。在此次会议上，麦卡锡（J. McCarthy）提议正式采用“人工智能”这一术语，用它来代表机器智能这一研究方向。有的论者提出，人工智能的“中心目标是使计算机有智能，一方面是使它们更有用，另一方面是理解使智能成为可能的原理。”[①] 围绕这个目标，产生了关于 AI 的两种不同理解。

强 AI 认为，有可能制造出真正能够推理和解决问题的智能机器，且这种机器能将被认为是有知觉的、有自我意识的。强 AI 可以有两类：

- 类人的人工智能，即机器的思考和推理就像人的思维一样；
- 非类人的人工智能，即机器产生了和人完全不一样的知觉和意识，使用和人完全不一样的推理方式。

弱 AI 认为，不可能制造出能真正地推理和解决问题的智能机器，这些机器只不过看起来像是有智能的，但并不真正拥有智能，也没有自主意识。

塞尔的中文屋论证所要反驳的就是强 AI 观点。

（二）图灵机和图灵测试

阿兰·图灵 （Alan M. Turing，1912—1954），英国数学家、逻辑学家、密码学家，被称为计算机科学之父、人工智能之父。在二战期间，曾协助英国军方破解德国的著名密码系统 Enigma；其主要科学成就有：提出“图灵机”和“图灵测试”的构想，开创了非线性力学。1952 年，因同性恋被警察发现，先被公审后被定罪，接受强迫的药物治疗。1954 年自杀生亡。1966 年，为纪念其在计算机领域的卓越贡献而专门设立了“图灵奖”。

1937 年，图灵在一篇论文中提出“图灵机”构想，他用机器来模拟人们用纸笔进行数学运算的过程，并把该过程看作如下两种简单动作的叠加：在纸上写上或擦除某个符号；把注意力从纸的一个位置移动到另一个位置。在每一个阶段，为了决定下一步动作，要考虑当事人当前所关注的纸上某个位置的符号，以及他当前的思维状态。为了模拟人的这种运算过程，图灵构造出一台假想的机器，它由以下几个部分组成：（1）一条无限长的纸带。它被划分为各别的方格，每个方格内有一个来自有限字母表的符号，字母表中有一个特殊符号表示空格。纸带上的方格从左到右依次被编号为 0,1,2,...，纸带的右端可以无限延伸。（2）一个读写头。它可以在纸带上左右移动，能读出当前方格里的符号，并能改变该符号。（3）一个状态存储器。它用来保存图灵机当前所处的状态。图灵机的所有可能状态的数目是有限的，且有一个特殊的状态，称为“停机状态”。（4）一套控制规则。它根据当前机器所处的状态以及当前读写头所指方格内的符号来确定读写头下一步动作，并改变状态存储器的值，令机器进入一个新状态。请注意，图灵机只是提供了一种计算描述，却未提及计算机的物理构造。为了完整地描述一台图灵机在某个时刻的行为，我们只需说明如下三项：（1）那时的输入；（2）那时的机器状态；（3）状态表。正因为如此，图灵机又被称为“纸上计算机”，或“理想的计算机”。

在论文《计算机和智能》（1950 发表，1956 年收入文集时改名为《机器能够思维吗？》）中，图灵认为，图灵机能实现人脑所能实现的一切。因为他已经证明，图灵机能计算一切可计算的功能（假设纸带和时间都是无限的），再根据另外一个主张，即人类的认知是生物计

[①] 转引自王永庆：《人工智能原理和方法》，第 8 页。

算的结果，图灵得出结论：我们所有的认知行为都可以用图灵机语言进行描述。因此，任何心理过程都必定有一种图灵机描述，这种描述具有相同的输入/输出关系。

在上面那篇文章中，图灵还提出了著名的“图灵测试”，由一台计算机、被测试人和测试主持人所组成。计算机和被测试人分别呆在两个不同房间里。测试过程由主持人提问，由计算机和被测试人分别做出回答。观测者能通过电传打字机与机器和人联系，以避免要求机器模拟人的外貌和声音。被测人在回答问题时尽可能表明他是一个“真正的”人，计算机将尽可能逼真模仿人的思维方式和思维过程。如果测试主持人听取他们各自的答案后，在多数时间内分辨不清哪个回答来自人，哪个回答来自机器，他就可以认为该台计算机具有了智能。

图灵机和图灵测试体现了关于人类认知和心灵的一种功能主义观点：某种东西是否有意识和有心灵，重要的不在于其内在结构，而在于它所发挥的作用或功能，或者它所表现出来的外在行为。基于图灵机和图灵测试，认知功能主义者提出了心灵的可多样实现性（multiple realizability）论证[①]：

P1 有心灵的系统都是认知系统;
P2 认知系统都是计算系统;
P3 图灵机完全能够描述任何计算系统;
C1 图灵机完全能够描述任何认知系统（由 P2 和 P3）;
P4 图灵机是独立于其物理实现（implementation）来定义的，即从功能上定义的;
C2 认知系统能够独立于其物理实现来定义（由 C1 和 P4）;
C3 有心灵的系统能够独立于其物理实现来定义（由 P1 和 C2）。

（三）塞尔的“中文屋论证”

莱布尼茨的磨坊

这是塞尔论证的先驱之一。在《单子论》中，莱布尼兹设想了一个物理系统、一台机器，它被认为能思维、有知觉。

> 可是，我们不得不承认，知觉和与之相联的一切是不能根据机械的理由即形状和运动得到解释的。假定说有一架机器，把它构造得能思想，产生感觉，有了知觉，还可以设想把它按原样的比例放大，人们可以像走进一座磨坊一样在里面一边观看，一边发现相互推动着的部件，但是都无法解释知觉来自何处。所以，一定是在单纯实体的内部而不是在复合物中或者说在机器中，去寻找知觉。也就是说，只有在单纯实体中才能发现知觉及其变化。单纯实体的全部活动只寓于自身之中。[②]

请注意，莱布尼兹的策略是将机器的公开行为——它可能展示了思想的证据——与机器内部操作的方式进行比较。他指出，这些内部的机械操作只是一些部件从一点运动到另一点，没有什么是有意识的或能解释感觉、知觉和思维。在他看来，对于心智状态而言，物理状态既不是充分的，也不是其构成要素。

戴维斯悖论

在 1974 年的一次学术会议上，戴维斯（L. Davis）做了下面的论证。假定我们已经了解

[①] Eliasmith, C. “The Myth of the Turing Machine, The Failings of Functionalism and Related Theses,” *Journal of Experimental & Theoretical Artificial Intelligence*, 14(2002), pp.1-8.

[②] 陈乐民编著：《莱布尼茨读本》，南京：江苏教育出版社，2006 年，第 37 页。译文有改动。

关于疼痛的全部细节。如果功能主义是正确的，则我们可以建造一个可以感受疼痛的机器人，这个机器人非常巨大，我们可以走进去观看，就像观看莱布尼茨的磨坊一样。机器人的脑袋内就像一座巨大的办公室，里面不是集成电路，而是穿着正装坐在办公桌后面的一群职员。每张桌子上有一部电话，电话连着几条线，电话网模拟人脑的神经连接，可以感受疼痛。这些职员受过训练，每个职员的任务是模拟一个神经元的功能。假定就在此刻，这个办公系统的一组电话非常剧烈地响起来，这种状态代表非常剧烈的疼痛。根据功能主义的观点，机器人处在剧痛之中。但你在办公楼里面转一圈，你看不到疼痛，所看到的只是一群中层职员在平静冷漠地工作；下一次，机器人感受到无法忍受的疼痛，你进入大楼参观，发现这些职员正举办圣诞联欢，每一个人都非常高兴。[①] 所以，功能主义的疼痛理论是错误的。

塞尔的中文屋思想实验

1980 年，塞尔提出了该论证的最初版本，此后不断回到这个论证，对它做新的表述、解释和阐发。该论证实际上是一个思想实验：

> 设想一个完全不懂中文、母语为英语的人被锁在房间里，房间里装满了中文符号箱（数据库）和一本操作这些符号的指令手册（程序）。设想房间外的人递进来其他的中文符号，它们是用中文书写的问题（输入），但房间里的人并不知道。再设想房间里的人遵照程序中的指令能递出一些中文符号，它们是这些问题的正确答案（输出）。这个程序使房间里的人能通过关于理解中文的图灵测试，但他对中文一窍不通。[②]

引文中实际上有如下类比或连接：坐在房间里的那个人，假设就是塞尔本人吧，相当于一台计算机；房间里供他使用的中文符号箱，相当于计算机的数据库；他所使用的那部指令手册，相当于一套计算机程序；递进房间的中文字条，相当于计算机的输入；递出房间的中文字条，相当于计算机的输出。由此得到下图：

塞尔解释说，“这个论证的意思是：如果房间里的人基于执行合适的理解中文程序仍不懂中文，那么，任何别的数字计算机只基于此也不会懂中文，因为任何计算机作为计算机都不会有这个人所没有的东西。”[③] 更明确地说，中文屋论证反驳的是强人工智能观点：人脑不过是一台数字计算机，人心只不过是一种计算机程序，心与脑的关系就是程序与计算机硬

[①] 参见庞德斯通：《推理的迷宫》，第 263 页。
[②] Searle, J. “Chinese Room Argument”, in Wilson, R.A. and F. Keil (eds.), *The MIT Encyclopedia of the Cognitive Sciences*, Cambridge: MIT Press, 1999, p.115.
[③] Ibid, p.115.

件的关系。

塞尔后来以更明确的方式刻画了中文屋论证的逻辑结构[①]。在下面的转述中，“P”表示前提，“C”表示结论：

P1：脑产生心。（其意思是说，那个我们认为构成心的心理过程，完全是脑内部进行的过程所产生的。）

P2：句法不足以确定语义。（这是一个概念真理，它明确了我们关于纯形式的和有内容的概念的区分。）

P3：计算机程序是完全以它们的形式的或语法的结构来定义的。（这可以看作依定义而真的命题；它是我们所说的计算机程序概念中的一部分。）

P4:心具有心理内容,具体说有语义内容。(这只是关于我们心智活动的一个明显事实。)

C1：任何计算机程序自身不足以使一个系统具有一个心灵。简言之，程序不是心灵，它们自身不足以构成心灵。（这是一个强有力的结论，它意味着仅通过程序设计来创造心灵的工程从一开始就注定要失败。）

C2：脑功能产生心的方式不能是一种单纯操作计算机程序的方式。（这个结论表明，脑不是或至少不只是一台数字计算机。）

C3：任何其他事物，如要产生心灵，应至少具有相当于脑产生心的那些能力。

C4：对于任何我们可能制作的、具有相当于人的心智状态的人造物来说，单凭一个计算机程序的运算是不够的。这种人造物必须具备相当于人脑的能力。

也可以认为，中文屋实验包含了如下两个论证，它们都能在塞尔论著中找到依据：

论证 1：

P1 如果强 AI 是正确的，就会有这样一个中文程序：如果某个计算系统运行了该程序，该系统由此就会懂中文。

P2 中文屋内的那个人能运行一个中文程序却并不因此懂中文。

C　强 AI 是错误的。

其中，第二个前提得到了中文屋实验的支持。此论证的结论是：运行一个程序不能产生理解力。此论证以更强的形式展开：

P1 模拟不等同于复制。

P2 大脑具有产生心灵的能力。

P3 计算机程序仅仅作为工具对心灵进行模拟。

C 凡是具有心灵的人造物至少要复制等同于大脑成心灵的能力（因果力）。

论证 2：

P1 程序是纯形式的（句法的）。

P2 人的心灵有心理内容（语义）。

P3 句法本身既不构成语义内容，对语义内容也是不充分的。

[①] 参见约翰·塞尔：《心、脑与科学》，杨音莱译，上海译文出版社，2006 年，第 29-32 页。

C　程序本身既不构成心灵，对心灵也是不充分的。

中文屋实验本身支持论证 2 中的 P3。这种主张，即句法操作对意义或思想是不充分的，是非常重要的，它具有比 AI 或理解力归属更广泛的意义。主要的心灵理论都认为人的认知一般来说是计算的；思维包含对符号的操作，这要借助于它们的物理属性。根据一种可选择的联结主义解释，这些计算是对“亚符号”状态的计算。如果塞尔是正确的，那么，强 AI 和这些理解人类认知的主要方法都是致人迷误的。

塞尔指出，应该纠正对“中文屋”论证的一些误解[①]：

（1）它并未证明“机器不能思维”。相反，大脑是机器，大脑能够思维。

（2）它并未证明“计算机不能思维”，而只是表明：如果把计算理解为图灵等人所定义的形式符号操作，则计算本身并不构成思维。

（3）它并未证明“只有大脑能够思维”。我们知道，思维是由大脑内的神经过程引起的，没有任何逻辑障碍阻止我们去建造这样一台计算机，它能够复制大脑内的因果过程去产生思维过程。

中文屋论证的要旨是：任何这样的机器都必须复制大脑的那种特殊的因果能力，以便产生思维的生理过程。仅凭操作形式符号不足以确保有这样的因果能力。

（四）对“中文屋论证”的回应

塞尔的“中文屋论证”产生了很大反响，激起了许多不同的回应，有系统回应、虚拟心灵回应、机器人回应、大脑模拟器回应、他心回应和直觉回应等。[②] 限于篇幅，这里只考虑系统回应、机器人回应和他心回应。

系统回应认为，尽管那个房间里的人不懂中文，但他只是一个更大系统的一部分，是一个中央处理器，是一套包括那个房间、规则书等等的复杂机制中的一个齿轮。理解中文的是那整个系统，而不是那个人。许多人认为，高度繁杂的人工智能程序并不是像塞尔所认为的那样只是机械翻译，而是考虑许多并列的不同规则，处理它们之间的冲突，认识它们之间的联系，并进行推测，还要建立新规则。就像一位锁在“中文屋”里的特别聪明的人最终有可能开始理解中文一样。也就是说，一个繁杂的、建立在规则上的系统有可能得到基础性意识。

塞尔对系统应答的答复很简单：整个系统也不知道中文词是什么意思，因为它无法将**任何**心智内容附加于任何符号。在原则上，中文屋里的那个人可以将整个系统内化，记住所有的指令，在头脑中完成所有的运算。此后，他可以离开房间到外头走走，甚至可能用中文交谈。但他仍无法了解“任何形式符号的意义”。这个人现在就是整个系统，但他仍不懂中文。例如，他不知道表示汉堡包的中文词语的意义。他仍然不能从句法得到语义。

大脑模拟器回应提出，请考虑一台计算机，它的操作方式和普通的 AI 程序极不相同，AI 程序有字母以及对语言符号串的操作。假如这个程序模拟的是一个以中文为母语的人在理解中文时其大脑中所发生的神经激发的实际结果——每一个神经、每一次激发。这样一来，计算机的工作方式与母语为中文的人的大脑的工作方式完全相同，处理信息的方式也完全相

---

① Searle, J. “Chinese Room Argument”, in *The MIT Encyclopedia of the Cognitive Sciences*, p.116.

② 参见 David Cole, “Chinese Room Argument,” in Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/chinese-room/, 读取日期：2013 年 8 月 22 日。

同，因此它就会懂中文。

塞尔再回应说，这个回应无关紧要。他本人提出了一个大脑模拟器方案的变种：假如房间里的人有大量水管和阀门，它们的排列方式和母语为中文的人的大脑中的神经元相同。现在程序告诉这个人在对输入做出反应时要开哪些阀门。塞尔认为，显然不会有任何对中文的理解。模拟大脑活动还不现实。塞尔的再回应类似于莱布尼兹的磨坊。

在“组合回应”的题目下，塞尔还考虑了一个具有系统、大脑模拟器和机器人三种回应的特征的系统：一个机器人，它有一个模拟其头颅中的计算机的数字大脑，这样整个系统的行为就难以与人的行为相区别。由于大脑的正常输入来自于感官，自然可以认为，多数大脑模拟器回应的支持者所想到的就是这种大脑模拟、机器人和系统回应的组合。有些人认为，将意向性归属给整个这样的系统是合乎情理的。塞尔也同意这种看法，但有一个保留：这只有在你不了解它的工作方式时才行。一旦你了解了真相——它是一台计算机，它是根据句法而非语义在毫无理解地操作符号——你就不会把意向性归属给它。

他心回应是这样的：你如何知道其他人懂中文或别的事情？只能借助他们的行为。这样一来，计算机（原则上）和其他人一样能通过行为测试，因此，如果你打算把认知归属给其他人，原则上你也必须把它归属给计算机。塞尔的再回应很简洁：在我们和其他人的交往中，我们预设了他们有心灵，就像在物理学中我们预设物体的存在一样。

在《中文屋 21 年》一文中，塞尔对强 AI 背后的哲学假设做了更系统和更深入的批判。建议有兴趣的读者去阅读此文。[①]

十六、普特南的“缸中之脑”和“孪生地球”论证

（一）笛卡尔的怀疑论

笛卡尔（Rene Descartes，1596—1650）认为，只有通过普遍怀疑方法检验的东西，才是绝对确实的，才能够成为知识体系的阿基米德点（确实性支点）。他所秉承的第一条方法论原则是：

> 凡是我没有明确地认识到的东西，我决不把它当成真的接受。也就是说，要小心避免轻率的判断和先入之见，除了清楚分明地呈现在我心里、使我根本无法怀疑的东西以外，不要多放一点别的东西到我的判断里。[②]

于是，他用普遍怀疑方法构造了下面一连串的怀疑论证：

论证 1：感觉经验靠不住。

因为我们有时候确实陷入幻觉和错觉之中。一座塔看起来是圆的，但后来才知道是方的。我们对于同一件事物有相互冲突的感觉印象。为了求证其中哪一个感觉印象是真实的或接近真实，我们必须求助于其他的感觉印象，但后者也有可能出错，也有可能相互冲突，我们又

[①] Searle, J. “Twenty-One Years in the Chinese Room”, in *Views into the Chinese Room*, New Essays on Searle and Artificial Intelligence, J. Preston and M. Bishop (eds.) Oxford/New York: Oxford University Press, 2002, pp.51-69; also in J. Searle, *Philosophy in a New Century*, Selected Essays, Cambridge/New York, Cambridge University Press, 2008, pp.67-85.

[②] 笛卡尔：《谈谈方法》，王太庆译，北京，商务印书馆，2010 年，18 页。

不得不求助于另外一些感觉印象，如此无穷倒退，永远也找不到一个可靠的支点。他做出结论说：

> 直到现在，凡是我当作最真实、最可靠而接受过来的东西，我都是从感官或通过感官得来的。不过，我有时觉得这些感官是骗人的；为了小心谨慎起见，对于一经骗过我们的东西就决不完全加以信任。[①]

可以把上述论证简述如下：

P1　凡是建立在不可信赖的证据之上的东西都永远不再可信，因为我无法判别它是否仍在欺骗我。

P2 感觉印象有时建立在不可信赖的证据之上。

C　感觉印象不应该再被信赖。

笛卡尔在这里犯了“推出过多”的错误：从“感觉印象有时欺骗人”不能推出“它们总是欺骗人”，就像不能从“某人有时说谎”推出“他永远说谎”一样。

论证 2：做梦和醒着难以区分。

笛卡尔提到，他没有任何标准来判别他究竟是醒着的还是在做梦，这就使得他有理由怀疑醒着时所发生的一切也不过是梦境而已，至少我们不能完全排除这样一种可能性。

> 有多少次我夜里梦见我在这个地方，穿着衣服，在炉火旁边，虽然我是一丝不挂地躺在我的被窝里！我现在确实以为我并不是用睡着的眼睛看这张纸，我摇晃着的这个脑袋也并没有发昏，我故意地、自觉地伸出这只手，我感觉到了这只手，而出现在梦里的情况好像并不这么清楚，也不这么明白。但是，仔细想想，我就想起来我时常在睡梦中受过这样的一些假象的欺骗。想到这里，我就明显地看到没有什么确定不移的标记，也没有什么相当可靠的迹象使人能够从这上面清清楚楚地分辨出清醒和睡梦来，这不禁使我大吃一惊，吃惊到几乎能够让我相信我现在是在睡觉的程度。[②]

在论证 1 和论证 2 中，笛卡尔都在寻求一个绝对确实的标准，结果是他无法找到。我们用来判定一个感觉印象是否出错的标准是另一个感觉印象，后者也可能出错；我们用来判定我们是否醒着的标准是我们**认为**我们正在醒着，但我们也可能梦见我们认为是醒着的。

论证 3：一个恶魔可能在系统地欺骗我们。

笛卡尔做了这样一个思想实验：一个本领强大得像上帝的恶魔（Demon）在系统地欺骗我们，它在不知不觉中向我们灌输了一整套错误的观念，使我们的观念体系整个地出错，甚至连算术和逻辑也难以幸免。因为要证实一串论证，我们必须诉诸另外的论证。如果第一串论证原则上可错，其他论证在原则上也会是可错的，故我们在原则上也可以怀疑逻辑。

> ……我要假定有某一个妖怪，而不是一个真正的上帝（他是至上的真理源泉），这个妖怪的狡诈和欺骗手段不亚于他本领的强大，他用尽了他的机智来骗我。我要认为天、空气、地、颜色、形状、声音以及我们所看到的一切外界事物都不过是他用来骗取我轻信

[①] 笛卡尔：《第一哲学沉思录》，庞景仁译，北京，商务印书馆，2009 年，15 页。
[②] 笛卡尔：《第一哲学沉思录》，16 页。

的一些假象和骗局。我要把我自己看成是本来就没有手，没有眼睛，没有肉，没有血，什么感官都没有，而却错误地相信我有这些东西。[①]

笛卡尔给我们提出的问题是：我们怎么才能知道情况并非如此？怎么知道我们并没有被一个恶魔系统地欺骗？对他本人来说，他无法排除有这样一个恶魔的可能性，至少是不能排除这样一种逻辑可能性。由此他得出结论：

对于这样的一些理由，我当然无可答辩；但是我不得不承认，凡是我早先信以为真的见解，没有一个是我现在不能怀疑的，这决不是由于考虑不周或轻率的原故，而是由于强有力的、经过深思熟虑的理由。因此，假如我想要在科学上找到什么经久不变的、确然可信的东西的话，我今后就必须对这些思想不去下判断，跟我对一眼就看出是错误的东西一样，不对它们加以更多的信任。[②]

不过，笛卡尔的上述看似极端的怀疑论立场很快被他下面的一连串推论稀释掉甚至消解掉了。他论证说，尽管我在怀疑一切都不可靠，但有一点却是确定无疑："我"在怀疑，即"我"在思考，而一个怀疑和思考着的"我"不可能不存在，由此得出他的哲学的第一个肯定性命题："我思故我在"。他继续论证说，"我"本身是不完美的，但"我"心中却有一个完美的上帝观念，"我"不可能是这个完美观念的原因，只有完美的"上帝"本身才是它的真正原因，由此得出他的哲学的第二个肯定性命题：上帝存在。然后，借助于上帝的全知全善全能，他魔术般地推出了"物质存在"、"心灵存在"、"他人存在"等命题，几乎完全回到了在普遍怀疑之前我们所接受的那些知识或观念。

不过，笛卡尔的怀疑论还是激起了深远的历史回响。

（二）普特南的"缸中之脑论证"

1981 年，普特南（Hilary Putnam）在他的《理性、真理和历史》一书中，提出了"缸中之脑"这个思想实验，它明显类似于笛卡尔所提出的"恶魔"论证：

设想一个人（你可以设想这正是阁下本人）被一位邪恶的科学家作了一次手术。此人的大脑（阁下的大脑）被从身体上截下并放入一个营养缸中，以使之存活。神经末梢同一台超科学的计算机相连接，这台计算机使这个大脑的主人具有一切如常的幻觉。人群，物体，天空，等等，似乎都存在着，但实际上此人（即阁下）所经验到的一切都是从那台计算机传输到神经末梢的电子脉冲的结果。这台计算机十分聪明，此人若要抬起手来，计算机发出的反馈就会使他"看到"并"感到"手正被抬起。不仅如此，那位邪恶的科学家还可以通过变换程序使受害者"经验到"（即幻觉到）这个邪恶科学家所希望的任何情境或环境。他还可以消除手术的痕迹，从而该受害者将觉得自己一直是处于这种环境的。这位受害者甚至还会以为他正坐着读书，读的就是这样一个有趣但荒唐至极的假说：一个邪恶的科学家把人脑从人体上截下来并放入营养缸中使之存活。神经末梢据说接上了一台超科学的计算机，它使这个大脑的主人具有如此这般的幻觉……[③]

[①] 同上书，20 页。
[②] 笛卡尔：《第一哲学沉思录》，19 页。
[③] 希拉里普特南：《理性、历史和真理》，童世骏、李光程译，上海译文出版社，1997 年，11 页。

普特南还把上述反常设想推至它的极端：所有人类（或许所有有感知能力的生物）的大脑都处在这样的缸中，那台超级计算机负责向我们提供集体的幻觉。他追问到：你如何担保你自己不处在这种困境之中？你如何担保这样的情形根本不会发生？

普特南本人并不同意这种极端形式的怀疑论，他给出了反驳缸中之脑构想的论证，其结论是：我们不可能前后一致地认为自己是“缸中之脑”，像“我是缸中之脑”这样的论断是自我反驳的。他主要从语义学角度论证了这一点，论证的重要依据是下面的“因果联系论题”（记为 CC）：

CC 论题：仅当一个词项与一个对象之间有适当的因果关联时，该词项才指称该对象。

比如说，假如一群蚂蚁在沙地上留下一些痕迹，这些痕迹非常近似于温斯顿·丘吉尔的画像，但我们不能说，这些蚂蚁在通过这些痕迹“表征”或“指称”丘吉尔，因为这些蚂蚁与丘吉尔没有因果接触，它们根本就不知道丘吉尔，也没有表征或指称他的意向。一艘宇宙飞船偶尔着陆于另一个星球，上面居住着与我们类似的外星人，但该星球上从来没有树，外星人也从来不知道树。但该艘飞船把树的影像留在该星球上，对于居住在该星球的外星人来说，该影像并不表征或指称树，尽管对于我们来说，该图像确实表征或指称树。词项不会神奇地或内在地指称一个对象，与指称对象有因果关联是必要条件之一，尽管或许还有别的条件。

基于 CC 论题，普特南构造了反驳“我们是缸中之脑”的论证，简述如下：

P_1　假设我们是缸中之脑。

P_2　根据 CC 论题，如果我们是缸中之脑，那么，“脑”并不指称脑，“缸”并不指称缸。

P_3　如果“缸中之脑”并不指称缸中之脑，那么，“我们是缸中之脑”就是假的。

C　如果我们是缸中之脑，那么，“我们是缸中之脑”就是假的。

普特南因此断言，“我们是缸中之脑”这个论断必定是假的，因为假设它为真就能够推出它为假。在这个意义上，他说，该论断是自我反驳或自我摧毁的。普特南的论证在逻辑上是有效的。若假定 CC 论题是正确的，该论证的结论是否为真就取决于 P_3 是否为真，这又取决于我们如何确定“我们是缸中之脑”的真值条件。围绕这些问题，产生了很多激烈的争论。[①]

顺便说一下，“缸中之脑”这个思想实验影响了许多当代的科幻小说和科幻电影，后者诸如《黑客帝国》,《盗梦空间》,《源代码》,《飞出个未来》,《异世奇人》等。在《黑客帝国》中，尼奥（Neo）就是一个被养在营养液中的真实人，而他的意识则由电脑系统“矩阵”（The Matrix）的电流刺激所形成和控制。他的一切记忆，都是外部电极刺激大脑皮质所形成的，并不是真实的生活历程。建议由兴趣的读者重看一次该电影，并思考有关的问题。

[①] 除普特南的原著外，对“缸中之脑”论证有兴趣的读者，可参看：Tony Brueckner，“Skepticism and Content Externalism,” in Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/skepticism-content-externalism/; Lance P. Hickey, “The Brain in a Vat Argument,” in Internet Encyclopedia of Philosophy, http://www.iep.utm.edu/brainvat/。