

多主体系统中的知道推理

陈慕泽

(中国人民大学哲学系, 北京 100872)

摘要: 多主体认知逻辑, 特别是其中关于多主体之间互知推理的研究, 是近十年来国际上新发展起来的研究领域, 对于经济学、军事学、博弈论、人工智能和计算机科学的发展, 具有重要的价值。本文围绕实例分析, 介绍了一种通过构造“知道”模型, 对多主体互知进行形式刻划和语义分析的方法, 意在引起国内读者特别是学者对于这一研究领域的兴趣和关注。

关键词: 多主体系统; 互知推理; 共同知识; 协同知识; “知道”模型

中图分类号: B81

文献标识码: A

单主体系统和多主体系统

某人知道下午有雨, 中午出门时他就会提醒自己带伞。

这是一个单主体系统。一般地说, 单主体系统中主体 **A** 的某个恰当决策, 取决于他是否知道事实 (或原理) **P**。

曹操败走华容道岔口, 大路并无动静, 小路几处烟起。曹操不走大路, 偏走小路华容道, 因为他知道诸葛亮深知“用兵虚则实之, 实则虚之”, 结果反而中了诸葛亮的埋伏。这里, 问题出在曹操光知道诸葛亮知道“用兵虚则实之, 实则虚之”, 但不知道诸葛亮知道曹操知道诸葛亮知道“用兵虚则实之, 实则虚之”。诸葛亮胜在他的“知道”比曹操的高出了一阶。

这是一个二主体系统。一般地说, 二主体系统中主体 **A** 的恰当决策, 不光取决于 **A** 是否知道事实 **P**, 而且取决于 **A** 是否知道另一个主体 **B** 是否知道 **P**, 如果答案都是肯定的, 那么还要取决于 **A** 是否知道 **B** 知道 **A** 知道 **B** 知道 **P**, ..., 等等。奕棋, 商务谈判, 都是典型的二主体系统。

如果主体超过两个, 是三个, 四个, 甚至更多, 这样的多主体系统中关于知道的推理及基于这种推理之上的恰当决策, 将变得极为复杂。

例如, 假设: 汽车驾驶员只有在确信别的驾驶员都遵守交通规则时自己才遵守规则, 那么, 一个汽车驾驶员光知道“红灯停, 绿灯行”的交通规则是不会按该规则开车的, 因为他并不知道所有驾驶员都知道这条规则。现在的问题是, 如果一个驾驶员不光自己知道这一规则, 而且知道所有的驾驶员都知道这

条规则，他是否一定遵守规则呢？回答是仍然不一定。因为他虽然知道所有的驾驶员都知道这条规则，但是并不一定知道所有的驾驶员都知道所有的驾驶员都知道这条规则。如果事实上他并不知道所有的驾驶员都知道所有的驾驶员都知道这条规则，那么，他有理由认为，除了他以外的所有驾驶员虽然都知道这条规则，但可能都不知道别的驾驶员知道这条规则，因而都不遵守规则，因而他自己也不遵守规则。这样的问题可以类似地问下去，回答都是“不一定”，直到问题中的“知道”重复的遍数等于驾驶员的人数。也就是说，假设这样的驾驶员有 n 个，这 n 个驾驶员就组成了一个 n （元）主体系统，在这个系统中，每个驾驶员都遵守“红灯停。绿灯行”的交通规则的必要条件是：所有的驾驶员都知道所有的驾驶员都知道所有的驾驶员都知道……（重复 n 遍）这一规则。（当然，这里假设每个驾驶员都有完备的推理能力）

多主体互知推理的挑战性

多主体之间的互知推理的复杂性在于，这种推理的对象中，不仅包括对象世界的知识，而且包含多个具有推理能力的主体；推理者对其他主体的思考及其结果进行推理，这些主体同样对推理者的思考及其结果进行推理。这使得推理的素材是弹性的，动态的，随着推理的过程不断变化的。这种推理，是对人的日常思维能力的挑战，也是对逻辑学的挑战。

对于知道推理的形式分析，最早可追溯到 20 世纪 50 年代初。芬兰哲学家冯·赖特（G.H.vonWright）在 1951 年出版的《模态逻辑》一书，首次提出和论述了认知逻辑的思想；芬兰哲学家辛提卡（J.Hintikka）继续了这方面的研究，他于 1962 年发表的《知识与信念》一书，为认知逻辑奠定了理论基础。但一般地说，认知逻辑所研究的有关知道（以及信念）的性质及其推理，还只限于单主体系统。运用现代逻辑的方法，对多主体系统中的知道特别是互知推理的研究，是近十年的事。推动上述研究的主要动力，直接来自于经济学、军事学、计算机科学和人工智能等学科对逻辑实用性的要求。经济学中的市场和战争学中的战场，都是典型的多主体系统，和中国做飞机生意的美国波音公司，就是一个站在华容道前的曹操，他不光要知道波音飞机的实际价值，而且要知道中国是否知道这种实际价值，而且要知道中国是否知道美国知道中国知道这种实际价值…。关于知道的推理，特别是多主体系统中关于互知推理的研究成果，首先受到了上述这些学科站在实用立场上的重视和欢迎。反映这方面研究成果的一本代表性论著，是由美国学者罗纳德·费金（Ronald. Fagin）等四人所著的《关于知道的推理》，（《Reasoning About Knowledge》），1995 年出版，1996 年再版。该书出版后，立即受到了学术界的重视和高度评价，有的经济学家甚至认为，该书对于“互知（interactive knowledge）”的研究，是经济学近十年的最重要的成果之一。

本文将围绕典型实例分析，通过构造关于知道的模型，看看逻辑语义学的方法，是如何成功地运用于分析多主体之间的互知的。

额上沾泥巴的孩子

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有举起了手。假设所有的孩子都有最佳的逻辑分析能力，那么，他们是如何思考并得出结论的？

先运用日常思维的方式来求解。

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。

细心的读者会发现一个很有意思的问题。在老师进教室之前，事实上每个孩子都已经知道他们之中至少有一人额上有泥巴，因此，老师所说的“你们之中至少有一人额上有泥巴”，似乎是句新信息量为零因而完全可以不说的废话；但另一方面，没有老师的这句话，老师的提问即使重复一万遍，也不会有孩子举手。这是怎么回事呢？

事实上，如果至少有两个孩子额上有泥巴，那么，所有孩子都知道至少有一个孩子额上有泥巴，但是，如果恰有两个孩子有泥巴，那么，虽然所有的孩子都知道至少有一个孩子额上有泥巴，但并非所有的孩子都知道所有的孩子都知道至少有一个孩子有泥巴（额上有泥巴的孩子 A 会认为，如果自己额上无泥巴，那么额上有泥巴的孩子 B 就会猜测，没有孩子额上有泥巴，这就是说，A 并不知道所有的孩子都知道至少有一个孩子额上有泥巴）；一般地，如果 n 个孩子额上有泥巴，那么“所有的孩子都知道…（重复 $n-1$ 遍）至少有一个孩子额上有泥巴”成立，但“所有的孩子都知道…（重复 n 遍）至少有一个孩子额上有泥巴”不成立。而在上例中，“所有的孩子都知道…（重复 n 遍）至少有一个孩子额上有泥巴”，是 n 个额上有泥巴的孩子作出正确判断的必要条件。老师当众说的那句话，使得“所有的孩子都知道…（重复 $\geq n$ 遍）至少有一个孩子额上有泥巴”成立。因此，老师的话当众而说，对于该题的解决是必不可少的。

如果说前一段的分析尚在日常思维的能力之内的话，那么后一段则有点超出这种能力了。这就需要提供形式的工具。建立“知道”的语义模型，就是这样一种尝试。

形式语言 K

在形式语言 K 中：

1, 2, ..., n 分别表示系统中 n 个不同的主体。

p, q, r, s ... 分别表示如“旧金山正在下雨”、“冰冰额上有泥巴”这样一些原子命题，它们的集合 Φ 构成作为主体认知和推理对象的外部世界的知识。

$\neg$ 读作“并非”， $\wedge$ 读作“合取”，表示“并且”。

模态算子 K_i 表示“主体 i 知道...”。因此， $K_i p$ 读作“主体 i 知道 p”，如果 p 表示“旧金山在下雨”，那么 $K_i \neg p$ 就表示“主体 i 知道旧金山并不在下雨”；而 $\neg K_i p$ 则表示“主体 i 不知道旧金山在下雨”；

原子命题是公式；如果 A 是公式，则 $\neg A$ 是公式；如果 A 和 B 是公式，则 $A \wedge B$ 是公式。如果 A 是公式，则 $K_i A$ 是公式。

$A \vee B$ （读作“A 析取 B”，表示“A 或者 B”）定义为 $\neg (\neg A \wedge \neg B)$ ；

$A \rightarrow B$ （读作“A 蕴含 B”，表示“如果 A，那么 B”），定义为 $\neg A \vee B$ ；

$A \leftrightarrow B$ （读作“A 当且仅当 B”），定义为 $(A \rightarrow B) \wedge (B \rightarrow A)$

T 是 $p \vee \neg p$ 这样的永真公式（称为重言式）的缩写，表示“真”；F 定义为 $\neg T$ ，表示“假”。

现在，我们可以把在自然语言中非常复杂的关于知道的命题表述得十分简明。例如，以下公式

$$K_1 K_2 p \wedge \neg K_2 K_1 K_2 p$$

表示“主体 1 知道主体 2 知道 p，但是主体 2 不知道主体 1 知道主体 2 知道 p”。

我们可以用“知道”来定义主观模态“可能”：主体 i 认为 A 是可能的，当且仅当主体 i 不知道 $\neg A$ ，即 $\neg K_i \neg A$ 。而象“主体 i 不知道是否 p”这样的断定，实际上是说“主体 i 认为 p 和 $\neg p$ 都是可能的”，也就是说“主体 i 既不知道 $\neg p$ 也不知道 p”，即 $\neg K_i \neg p \wedge \neg K_i p$ 。

考虑下面这个有关水门事件的断定：迪恩不知道尼克松是否知道迪恩知道尼克松知道麦卡德偷窃了奥博林在水门的办公室。令主体 1 表示迪恩，主体 2 表示尼克松，p 表示“麦卡德偷窃了奥博林在水门的办公室”，则该断定可表达为

$$\neg K_1 \neg (K_2 K_1 K_2 p) \wedge \neg K_1 \neg (\neg K_2 K_1 K_2 p)$$

形式语言 K 的语义解释，关于“知道”的模型

以上讨论的是形式语言 K 的语法，下面讨论 K 的语义，即要构造模型，用以判定 K 语言中公式的真假。

这样的模型，记为 M ，是一个克里普克结构 $(W, V, R_1, \dots, R_n)$ ，其中， W 是一可能世界集； V 是一个解释，它给任一可能世界，指派以一个确定的真值赋值，即对任一 $w_i \in W$ ，和任一原子命题 $p \in \Phi$ ， $V(p, w_i) = T$ 或 $V(p, w_i) = F$ ，但不能二者。如果 p 表示“旧金山正在下雨”，则 $V(p, w_i) = T$ 表示在可能世界 w_i 中，旧金山正在下雨； R_i 是 W 上的二元关系。如果 w_i 和 w_j 有关系 R_i ，记为 $w_i R_i w_j$ ，表示主体 i 依据在可能世界 w_i 中的信息，认为可能世界 w_j 是可能的。（一个世界，是一个事件集，只要其中不包括矛盾事件，就是一个可能世界；但一个可能世界，对于某个主体来说，完全可能是不可能世界，如果这个主体知道这个可能世界中某个事件的矛盾事件）这里，我们进一步规定 R_i 是同时满足自返、对称和传递关系的等价关系，这样，如果主体 i 在可能世界 w_i 中觉得 w_j 是可能的，这说明在可能世界 w_i 和 w_j 中，主体 i 具有对外部世界同样的信息，从而对他来说，这两个世界是无法区分的。因此， $w_i R_i w_j$ 也表述为“主体 i 无法区分 w_i 和 w_j ”。

一个公式 A 在一个结构(模型) M 的一个给定的可能世界 w_i 中真，记作 $(M, w_i) \models A$ ，读作 (M, w_i) 满足 A 。

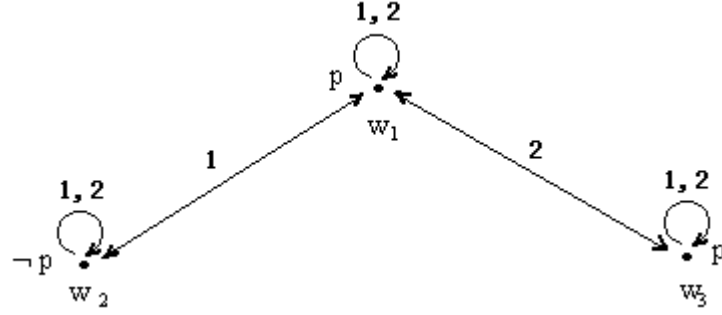
$$(M, w_i) \models p \text{ 当且仅当 } V(p, w_i) = T.$$

$$(M, w_i) \models A \wedge B \text{ 当且仅当 } (M, w_i) \models A \text{ 并且 } (M, w_i) \models B$$

$$(M, w_i) \models \neg A \text{ 当且仅当 } (M, w_i) \not\models A$$

$$(M, w_i) \models K_i A \text{ 当且仅当对任一 } w_j \in W, \text{ 如果 } w_i R_i w_j, \text{ 则 } (M, w_j) \models A.$$

上述模型所表达的核心意思是：主体 i 知道 p ，当且仅当 p 在主体 i 认为可能的所有可能世界中都真。我们用一个实例的图示来描述这一点，克里普克结构的优点之一是可图示的。



(图 1)

上图所示的模型 $M = (W, V, R_1, R_2)$ ，其中， $W = \{w_1, w_2, w_3\}$ ， p 在 w_1 和 w_3 中真，而在 w_2 中假。主体 1 不能区分 w_1 和 w_2 （即主体 1 在 w_1 认为 w_2 是可能的，由 R 的对称性，自然在 w_2 同样认为 w_1 是可能的，即 $w_1 R_1 w_2$ 和 $w_2 R_1 w_1$ 成立），主体 2 不能区分 w_1 和 w_3 。标有 1, 2 的线段在 w_i ($i = 1, 2, 3$) 从自身指向自身，表示 R 关系的自返性，即表示 $w_i R_j w_i$ ($i = 1, 2, 3; j = 1, 2$)，例如表示 $w_3 R_1 w_3$ ；标有 1 的线段的两端指向 w_1 和 w_2 ，表示主体 1 不能区分 w_1 和 w_2 ，并表示 R 关系的对称性。同样，标有 2 的线段表示主体 2 不能区分 w_1 和 w_3 。

令 p 表示“北京天晴”，则依据上图，可得出以下结论：

结论 1，在可能世界 w_1 ，北京天晴，但主体 1 并不知道这一点，因为他在 w_1 中认为 w_1 和 w_2 都是可能的（或者说依据他在 w_1 的知识，他无法确定 w_1 和 w_2 究竟哪个是真实世界，即无法区分 w_1 和 w_2 ），而 p 在 w_1 中真，但在 w_2 中假。

结论 2，主体 2 在可能世界 w_1 知道北京天晴，因为在可能世界 w_1 ，主体 2 认为可能的世界是 w_1 和 w_2 ，在这两个可能世界中， p 都是真的。

结论 3，主体 2 在可能世界 w_2 知道并非北京天晴，因为主体 2 在 w_2 中认为可能的世界只有 w_2 自身，而在 w_2 中， $\neg p$ 真。同理，主体 1 在可能世界 w_3 中知道北京天晴。

结论 4，在可能世界 w_1 ，主体 1 知道主体 2 知道北京是否天晴，因为在可能世界 w_1 ，主体 1 认为可能的两个世界是 w_1 和 w_2 ，在这两个世界中，主体 2 都知道北京的天气（见结论 2 和结论 3）。也就是说，虽然在可能世界 w_1 ，主体 1 并不知道北京是否天晴，但是他知道主体 2 知道这一点。

结论 5，和结论 4 成为对比的是，在可能世界 w_1 ，虽然主体 2 知道北京天晴（结论 2），但是他不知道主体 1 不知道这一点。因为在可能世界 w_1 ，主体 2 认为可能的两个世界是 w_1 和 w_3 ，在 w_1 中，主体 1 不知道北京天晴（结论 1），但在 w_3 中，主体 1 知道北京天晴（结论 3）。

以上结论，可以用一个逻辑表达式概括：

$$(M, w_1) \models p \wedge \neg K_1 p \wedge K_2 p \wedge K_1 (K_2 p \vee K_2 \neg p) \wedge \neg K_2 \neg K_1 p$$

前面已经指出，一个可能世界是一个事件集，相应的命题在其中真或假。在以上的讨论中，构成 w_1 和 w_3 的事件都是“北京天晴”，因此，似乎是两个相同的世界因而可以略去一个。但事实上却不能这样。因为一个可能世界的规定，不光基于构成它的事件，而且基于主体认为它是否可能。例如，在可能世界 w_1 ，主体 1 认为可能世界 w_2 是可能的，但在可能世界 w_3 ，他却不这么认为，这样，他在 w_1 不知道北京天晴，而在 w_3 则知道这一点。

多主体系统中的共同知识

在 n 主体系统中，如果所有的主体都知道所有的主体都知道…（重复 $\geq n$ 遍） A ，则称这 n 个主体掌握了关于 A 的共同知识（Common Knowledge），或称 A 是这 n 个主体的共同知识。在上文的实例分析中，可以看到共同知识在多主体系统的知道推理中的重要作用。这一多主体认知系统中的重要概念，最早是由路易斯在讨论“协约”时提出的，他认为，某种东西要成为多方的“协约”，必须成为缔约各方的共同知识，也就是说，缔约各方不但都要知道协约的内容，而且要知道各方都知道协约的内容，等等。

为了对共同知识进行形式刻画，需要在语言 K 中增加新的算子 E_G 和 C_G ，满足：如果 A 是公式，则 $E_G A$ 、 $C_G A$ 都是公式。

G 表示主体集 $\{1, 2, \dots, n\}$ 。 $E_G A$ 表示“ G 中每个主体都知道 A ”； $C_G A$ 表示“ A 是 G 中所有主体的共同知识”。在不引起歧义的情况下，作为下标的 G 可以省略，即 $E_G A$ 和 $C_G A$ 分别记为 EA 和 CA 。

如果 $\{1, 2, \dots, i\}$ 是 G 的一个真子集（即 $i < n$ ），则 $E_{\{1,2,\dots,i\}} A$ 表示在 $\{1, 2, \dots, i\}$ 中每个主体都知道 A 。这种写法同样用于 C 。这样， $K_3 \neg C_{\{1,2\}} p$ 就表示主体 3 知道 p 不是主体 1 和主体 2 的共同知识。

在模型 M 中作如下定义：

$(M, w_i) \models EA$ ，当且仅当对任一 $i \in G$ ， $(M, w_i) \models K_i A$ ，即在可能世界 w_i 中， EA 真，当且仅当每个主体都知道 A 。

令 $E^1 A$ 表示 EA ， $E^{k+1} A$ 表示 $EE^k A$ ，则

$(M, w_i) \models CA$ ，当且仅当 $(M, w_i) \models E^k A$ （ $k \geq n$ ），即在可能世界 w_i 中， CA 真（ A 是所有主体的共同知识），当且仅当所有的主体都知道所有的主体都知道…（重复 $\geq n$ 遍） A 。（注意，事实上，对于任意 $k > 1$ 和任意 $w_i \in W$ ，如果 $(M, w_i) \models E^k A$ ，则 $(M, w_i) \models E^{k-1} A$ ）

对共同知识可以作出一种有意思的直观图示，为此，先来定义何为从一个

可能世界到另一个可能世界可通达。(1) 对任意可能世界 w_{j1} 和 w_{j2} ，如果存在主体 i ， $w_{j1} R_i w_{j2}$ ，则称从 w_{j1} 至 w_{j2} 可通达，并称这种通达为一步可通达；(2) 对任意可能世界 w_{j1} 、 w_{j2} 和 w_{j3} ，如果从 w_{j1} 至 w_{j2} 可通达，并且从 w_{j2} 至 w_{j3} 可通达，则从 w_{j1} 至 w_{j3} 可通达。并且，如果从 w_{j1} 至 w_{j2} 是 k 步可通达，从 w_{j2} 至 w_{j3} 是 1 步可通达，则从 w_{j1} 至 w_{j3} 是 $k + 1$ 步可通达。

虽然一般模态逻辑都把结构中的 R 关系称为可通达关系，但这里定义的可通达关系不同于 R_i 关系。第一， R_i 关系是相对于某个主体 i 而言的，可通达关系不是相对于某个主体而言的，第二，存在可通达关系的可能世界之间，不一定有 R 关系成立。例如，图 1 中从 w_2 至 w_3 可通达，但 $w_2 R_1 w_3$ 和 $w_2 R_2 w_3$ 都不成立。

关于可通达关系，有两条重要推论。

推论 1 $(M, w_i) \models E^k A$ ，当且仅当 $(M, w_i) \models A$ 并且对所有 w_j ，如果从 w_i 至 w_j k 步可通达，则 $(M, w_j) \models A$

推论 2 $(M, w_i) \models CA$ ，当且仅当对所有 w_j ，如果从 w_i 至 w_j 可通达，则 $(M, w_j) \models A$

对 k 施归纳，易证推论 1；推论 2 是推论 1 的直接推论。

可以设想这样一个示图，其中，每个可能世界表示为一个点，任意两个一步可通达的可能世界之间用线段联接。以上两个结论的意义在于，判定 A 是否为可能世界 w_i 上的共同知识，只须看 A 是否在从 w_i 可通达的点（可能世界）上都真；判定 $E^k A$ 在 w_i 上是否为真，只须看 A 是否在从 w_i k 步可通达的点上都真。

对“额上沾泥巴的孩子”的再分析

现在，尝试构造语义模型，对“额上沾泥巴的孩子”作形式分析。假设孩子有 n 个，要证明的是，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。

自然需要假设，题目陈述的条件，例如，所有的孩子都足够聪明，对所有孩子都是共同知识。

令 $1, 2, \dots, n$ 分别表示 n 个不同的孩子。 $(x_1, \dots, x_n)$ 表示可能世界，其中任一 x_i ， $x_i = 1$ ，或者 $x_i = 0$ 。如果 $x_i = 1$ ，则表示孩子 i 额上有泥巴，否则表示没有。显然，对于 n 个孩子，这样的不同可能世界共 2^n 个。例如，如果只有 3 个孩子，则可能世界 $\{1, 0, 1\}$ 表示孩子 1 和孩子 3 有泥巴。假设这个可能世界就是真实世界。在这个世界中，在老师说话之前，孩子 1 能看到孩子 2 没有泥巴而孩子 3 有泥巴，他唯一不能确定的是自己额上是否有泥巴，因此，他认为 $(1, 0, 1)$ （即真实世界）和 $(0, 0, 1)$ 都是可能的。也就是说，孩子

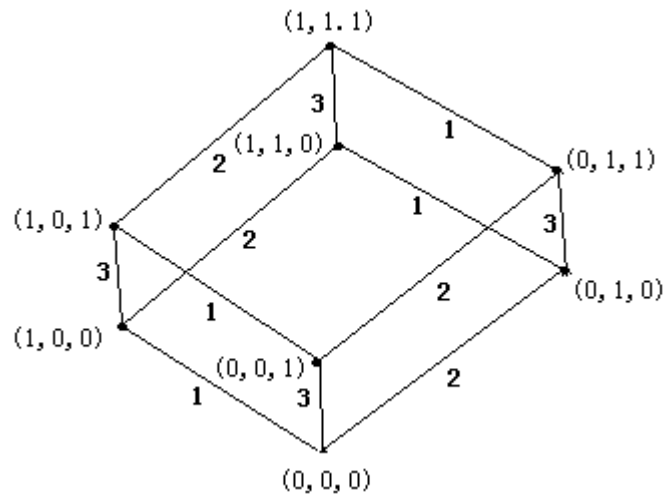
i 在可能世界 $(a_1, \dots, a_n)$ 认为可能世界 $(b_1, \dots, b_n)$ 是可能的, 即 $(a_1, \dots, a_n) R_i (b_1, \dots, b_n)$, 当且仅当除了 $a_i \neq b_i$ 以外, $(a_1, \dots, a_n)$ 和 $(b_1, \dots, b_n)$ 完全相同。

令 $\Phi = \{p_1, \dots, p_n, p\}$, 其中 p_i 表示“孩子 i 有泥巴” ($i = 1, \dots, n$), p 表示“至少有一个孩子有泥巴”。 $(M, (x_1, \dots, x_n)) \models p_i$ 当且仅当 $x_i = 1$ 。
 $(M, (x_1, \dots, x_n)) \models p$ 当且仅当存在 $x_j, x_j = 1$ 。

这样, 完成了对模型 $M = (W, V, R_1, \dots, R_n)$ 的定义。

这一模型的优点之一是基于之上可以作出清晰直观的图示解析。

令 2^n 个点表示上述 2^n 个不同的可能世界, 并在任意两个一步可通达的点之间用标有数字 i 的线段联接 (即如果孩子 i 在 w_i 认为 w_j 可能, 则用标有 i 的线段联接表示这两个可能世界的点), 这样, 长于空想想象的读者可以知道, 我们因此得到了一个 n 维立方体。下图表示的就是当 $n = 3$ (即假设只有 3 个孩子) 时这样的三维立方体。



(图 2)

图 2 中共有 8 个点, 表示所有的 8 个可能世界。每两个可能世界之间都有标有数字的线段联接, 例如, 标有 1 的线段联接 $(1, 1, 1)$ 和 $(0, 1, 1)$, 表示孩子 1 在这两个世界的任何一个中都认为另一个世界是可能的。图中也说明, 从任何一个可能世界出发, 其余的可能世界都是可通达的。

从图 2 立即可以得出 (证明) 许多结论, 例如:

结论 1 每个孩子都知道除自己外哪个孩子额上有泥巴。不妨设可能世界 $(1, 0, 1)$ 是现实世界，在这一世界中，孩子 1 认为可能的世界是 $(1, 0, 1)$ 和 $(0, 0, 1)$ ，在这两个可能世界中，孩子 3 都有泥巴，因此，孩子 1 知道孩子 3 有泥巴；同理，孩子 2 知道孩子 1 和孩子 3 有泥巴；孩子 3 知道孩子 1 有泥巴。

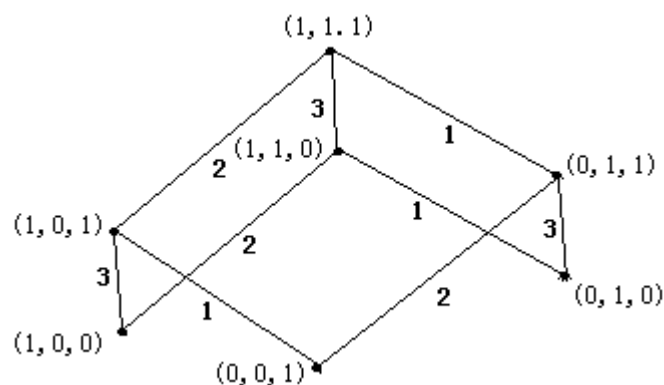
结论 2 “每个孩子都知道除自己外哪个孩子额上有泥巴”是所有孩子的共同知识。结论 1 的证明所选择的可能世界带有任意性，因此，“每个孩子都知道除自己外哪个孩子额上有泥巴”在所有可能世界中真，即在从任意一个可能世界可通达的所有可能世界中真，因此，是共同知识。

结论 3 $(M, (1, 0, 1)) \models E p$ ，即在可能世界 $(1, 0, 1)$ 中，所有的孩子都知道至少有一个孩子额上有泥巴。自 $(1, 0, 1)$ 一步可通达的可能世界有 $(1, 1, 1)$ 、 $(1, 0, 0)$ 和 $(0, 0, 1)$ ，在这四个可能世界中， p 即“至少有一个孩子有泥巴”都真。

结论 4 $(M, (1, 0, 1)) \not\models \neg E^2 p$ ，即在可能世界 $(1, 0, 1)$ 中，并非所有的孩子都知道所有的孩子都知道至少有一个孩子额上有泥巴。因为存在自 $(1, 0, 1)$ 两步可通达的可能世界 $(0, 0, 0)$ ，其中 p 假，即没有孩子有泥巴。

结论 5 在可能世界 $(1, 0, 1)$ 中，孩子 1 虽然认为 $(0, 0, 0)$ 是不可能的，但认为孩子 3 可能在 $(0, 0, 1)$ 中认为 $(0, 0, 0)$ 是可能的。

当老师说至少有一个孩子额上有泥巴后，这个命题立即成为所有孩子的共同知识，这样，任一孩子都不可能在任一可能世界中认为 $(0, 0, 0)$ 是可能的，这样，通向 $(0, 0, 0)$ 的可通达关系中断， $(0, 0, 0)$ 世界可因此移去，图 2 中的立方体因而坍塌成如下图所示：



(图 3)

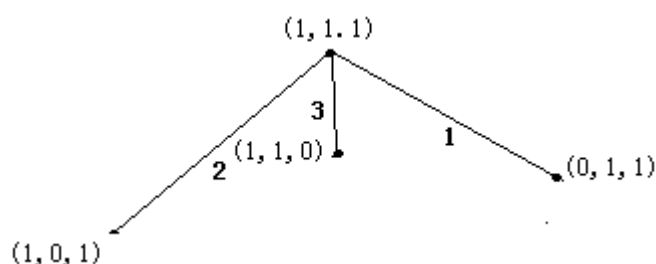
由图 3 可得：

结论 6 在可能世界 $(1, 0, 0)$ ，主体 1 知道自己额上有泥巴。因为在 $(1, 0, 0)$ ，孩子 1 认为可能的世界只有 $(1, 0, 0)$ ，（注意 R 关系的自返性。图 2、3 省略了图 1 中表示自返关系的圆弧线段）而在 $(1, 0, 0)$ 中，孩子 1 额上有泥巴。同理，在可能世界 $(0, 0, 1)$ ，主体 3 知道自己额上有泥巴，在可能世界 $(0, 1, 0)$ ，主体 2 知道自己额上有泥巴。由结论 6 直接可得：

结论 7 如果事实上只有一个孩子有泥巴（即真实世界是 $(1, 0, 0)$ 或 $(0, 1, 0)$ 或 $(0, 0, 1)$ ），那么，在老师第一遍提问后有泥巴的孩子就会举手。

结论 8 结论 7 是所有孩子的共同知识。不难验证，结论 7 在图 3 所有的可能世界中都真。例如，结论 7 在 $(1, 0, 1)$ 中真，因为在该可能世界中，有两个孩子有泥巴，因此，作为条件句的结论 7 的前件假，结论 7 自身因而真。再如，由结论 6，立即可得结论 7 在 $(1, 0, 0)$ 中真。

当老师的第一遍提问后无人举手时，由结论 8，所有的孩子都立即知道 $(1, 0, 0)$ 、 $(0, 1, 0)$ 和 $(0, 0, 1)$ 是不可能世界。这样，任一孩子都不可能在任一可能世界中认为这三个世界是可能的，这样，通向这三个世界的可通达关系中断，这三个世界可因此移去，图 3 继续坍塌而成下图所示：



(图 4)

图 4 显示，当老师的第一遍提问后无人举手时，如果只有两个孩子有泥巴，他们立即明白自己有泥巴。

一般地，在上述语义图中，如果老师的第 k 遍提问后无人举手，那么即可移去包含 k 个 1 的可能世界，并显示如果只有 $k + 1$ 个孩子有泥巴，那么，这些孩子都知道自己有泥巴。另一方面，如果老师不告诉大家至少有一个孩子有泥巴，老师的任何一次提问，都不会使上述语义图发生任何变化。

参考文献

- [1] Ronald Fagin 等: 《Reasoning about Knowledge》The MIT Press.
- [2] Halpern,J,Y: 《Reasoning about only knowing with many agents》In Proc. National Conference on AI(93)
- [3] Aumann,R,J: 《Agreeing to disagree》In Annals of Statistics.

Reasoning about Knowledge in Multi-agent System

CHEN Mu-ze

(Department of Philosophy Renmin University of China Beijing 100872,China)

Abstract: The logic of multi-agent cognition, especially the research of reasoning about multi-agent interactive knowledge is a new field developing in the past decade, which is valuable in the development of economics, strategics, game theory, artificial intelligence and computer science. With a case analysis, this paper introduces an approach formally and semantically analyzing multi-agent interactive knowledge in the base of a knowledge model in order to arouse interests and attentions to the this field.

Key words: multi-agent system; reasoning about interactive knowledge; common knowledge and distributed Knowledge; the model of knowledge

收稿日期: 2003-4-20

作者简介: 陈慕泽 (1951-) 男, 上海人, 中国人民大学哲学系教授, 中山大学逻辑与认知研究所兼职研究员