·西方哲学研究·

人工智能科学在十七、十八世纪欧洲哲学中的观念起源

徐英瑾
(复旦大学　哲学学院,上海　200433)

[摘　要]　尽管人工智能科学是在二战后才在西方科技界涌现的,但其思想根苗至少可以上溯到十七、十八世纪的欧洲哲学。具体而言,人工智能的哲学"基础问题"可被一分为二:第一,建立一个能够展现真正人类智能的纯机械模型,在观念上是否可能?第二,若前述问题的答案是肯定的,怎样的人类心智模型才能够为这种模型的建立提供最佳的参照?本文将论证,笛卡尔和莱布尼茨对上述第一个问题都给出了否定的回答,而霍布斯则给出了肯定的回答。至于第二个问题,休谟关于心智构架的重构工作,就可以被视为当代AI科学中的联结主义进路的先驱,而康德在调和直观和思维时所付出的努力,则为当代AI专家整合"自下而上"进路和"从上至下"进路的种种方案所应和。一言以蔽之,十七、十八世纪的欧洲哲学实乃AI科学的一个潜在的"智库",尽管AI界的主流对此并无清楚之意识。

[关键词]　人工智能　符号AI　类比　联结主义　"自下而上"进路　"从上至下"进路

一、导　　论

在很多人看来,"人工智能"(Artificial Intelligence,简称AI)是一个工程技术色彩浓郁的学术领域,哲学研究则高度思辨化和抽象化,二者之间应当是风马牛不相及的。但这实际上是一种误解。芝加哥大学哲学教授郝格兰的著作《人工智能概念探微》(特别是第一章)[①]以及加州大学伯克利分校的哲学教授德瑞福斯的著作《计算机依然不能做什么》(特别是第67—69页),[②]都留出了一定的篇幅,用以挖掘AI的设想在西方哲学史中的根苗。而英国女哲学家兼心理学家博登的浩瀚巨著《作为机器的心灵——认知科学史》,[③]则以更大的篇幅讨论了AI科学和整个西方科技史、思想史之间的互动关系(尤其是第二章)。不过,令人遗憾的是,在汉语哲学界,将西方哲学史的视角和AI哲学的视角相结合的研究成果,相对还比较罕见,因此拙文将在这个方向上作出一番小小的开拓性努力。另外,笔者也希望能够通过这种"架桥"工作,帮助读者看到那些看似新锐的科技问题和相对古老的哲学争议之间的密切关联,并为缓解目前在汉语学界业已过于紧张的"科学—人文"关系,献上绵薄之力。

为了能够集中讨论,本文将只选取西方哲学史中的一个片段——十七、十八世纪欧洲哲学——为参考系,来讨论哲学和AI之间的关系。

[收稿日期]　2010-05-13

[作者简介]　徐英瑾,哲学博士,复旦大学哲学学院副教授。

*　本课题的研究得到了2006年国家社科基金项目"维特根斯坦哲学视野中的人工智能问题"(项目批准号:06CZX011)的资助。

① John Haugeland, Artificial Intelligence: The Very Idea (MIT Press, 1985).

② Herbert Dreyfus, What Computers Still Cannot Do (MIT Press, 1992).

③ Margaret Boden, Mind as Machine: A History of Cognitive Science (Oxford University Press, 2006).

由于篇幅限制，在下面我们只能选择五位哲学家予以概要式的讨论：笛卡尔（Rene Descartes，1596—1650）、莱布尼茨（Gottfriend Wilhelm von Leibniz，1646—1716）、霍布斯（Thomas Hobbes，1588—1679）、休谟（David Hume，1711—1776）和康德（Immanuel Kant，1724—1804）。他们可被编为三组：

第一组：笛卡尔和莱布尼茨。其特点是：他们通过卓越的哲学想象力，明白地预报了后世 AI 科学家通过被编程的机械来实现智能的设想。但他们又同样明确地提出了反对机器智能的论证。从这种意义上说，他们虽不可能为今日符号 AI 的技术路线投赞成票，却明确地表述出了“人工智能哲学”的基本问题意识：制造人类水平的智能机器，是不是先天可能的？

第二组：霍布斯。他处在笛卡尔和莱布尼茨的对立面。具体而言，他虽没有明确地提到机器智能的可实现性问题，但是他对于人类思维本性的断言，却在逻辑上等价于一个弱化的“物理符号假设”。因此，他可被视为二十世纪的符号 AI 路线在近代哲学中的先祖。

第三组：休谟和康德。从现有文献来看，他们并未明确讨论过“机器智能的可实现问题”。然而，他们各自的心智理论却在一个更具体的层次上引导了后世 AI 专家的技术思路，因此也算作是 AI 科学的先驱。

在所有的这些哲学家中，我会留给康德最多的篇幅，因为他的思想最为深刻，可供 AI 挖掘的材料也最多（尽管认识到这一点的只有侯世达等少数 AI 专家）。

二、笛卡尔和莱布尼茨：机器智能的反对者

从表面上看来，与下文所要提到的霍布斯相比，笛卡尔和莱布尼茨似乎更有资格充当符号 AI（也就是经典 AI）在近代哲学中的先驱。摆得上桌面的理由有：

其一，此二君都属于广义上的“唯理派”阵营，都主张人的心智活动的实质在于符号表征层面上的推理活动（为了宣扬这个观点，莱布尼茨还专门写了一本《人类理智新论》，和经验论者洛克打起了笔仗）；

其二，符号 AI 路数一般都重视数理研究和一般意义上的科学研究，而笛、莱两人的学术造诣也都体现了这样的特征。具体而言，笛卡尔是直角坐标系的发明人，在物理学（特别是光学）领域小有斩获，也喜欢搞生理学。莱布尼茨则是微积分的发明人之一，是柏林科学院的创始人；

其三，与人工智能直接相关的一些计算机技术，和莱布尼茨有直接关联。他在 1764 年于巴黎建造的乘法运算机（改进于帕斯卡的运算机），以及他对于二进制的推崇，都是为计算机史家所津津乐道的实例。

然而，笔者却并不认为这些理由能够充分地担保他们会赞同机器智能的可能性。

首先，成为计算机技术的先驱并不等于成为人工智能的先驱。一个计算机科学家完全可能拒绝实现人类水平上的机器智能的可能性，而仅仅把计算机视为人类的工具。因此，莱布尼茨对于帕斯卡运算机的改进，并不担保他会成为 AI 的同道；

其次，他们在数学和自然科学方面的贡献，也并不担保他们会赞成机器智能的可实现性（理由由上一点类推）；

第三，是否赞同符号 AI，和是否处在“唯理派”阵营中，并无直接关联。这是因为，唯理派的立场仅仅是“心智活动的实质在于符号表征层面上的推理活动”，但对于符号 AI 来说，更为有用的一个论题则是“任何被恰当编程的、符号表征层面上的推理活动都是心智活动”。很显然，从逻辑上看，即使已经有了前面这个立场，也并不担保后一个论题就能够被推出。

进而言之，笛卡尔和莱布尼茨还各自提出了一个论证，明确反对机器智能的可能性。

先从笛卡尔说起。我们知道，在“身心关系”问题上笛卡尔是二元论者，即认为人是占据广延的物质实体和不占据广延的灵魂实体的复合体。而在关于动物的本性的问题上，他倒是一个比较

彻底的机械唯物论者，即认定动物只是“自动机”，毫无灵魂。从这个立场出发，他显然是不可能认为我们有可能制造出具有人类智能水平的机械装置的，因为从他的二元论立场来看，“智能”——或者说“灵魂”——的形式，和物理世界的配置形式无关，通过改变后者，我们是不可能得到前者的。不过，这样的一种反机器智能的论证本身就已经预设了二元论立场，因此非二元论者未必会买他的账。好在笛卡尔还有一个形而上学负荷更少的反机器智能论证。此论证见于其名著《方法论》：

> 假若真有这样的一些机器，其具有猿猴（或其他缺乏理性的动物）的所有器官和外形，那么，恐怕我们就毫无理由断言，这些机器并不完全具有那些被模仿动物的本性。但请再试想这样一种情况：假若有一些机器，其在技术允许的范围内竭力仿造我们的身体，并试图模拟我们的行为，那么，它们是否为真人？答案是否定的，而且我们总能通过两个途径来获得这个否定性的答案。第一个途径是：我们会发现，它们总不会使用语词和记号，或像我们那样把语词和记号组合在一起，以便向他人传达出我们的思想。为何这么说呢？我们可以设想一台从表面上看来可以表达语词的机器，甚至可设想，其表达的语词是匹配于一些将最终导致相关器官变化的身体行动（比如，当你触及其某一部分的时候，它就会问你，是不是想和它说些啥；而当你触及其另一部分的时候，它就会大哭，抱怨你弄疼了它）。但即使如此，它却无法给予语词以不同的排列方式，以便应对人们在面对它时所能够说出的种种不同的话——尽管最笨的人也能够胜任这个任务。第二个途径是：尽管这些机器能够执行很多任务，并在执行某些任务的时候表现得比人类更为出色，但它们必定会在执行另外一些任务时出洋相。这样一来，我们就会发现，这些机器并不是根据知识来运作的，而是根据其器官部件自身的倾向来运作的。这又是为何呢？因为人类理性乃是在诸种问题语境中皆有用武之地的万能器具，而这些器官部件呢，则只不过是分别为特定的问题语境而定制的专门器具。这样一来，如果我们要让这些机器能够对付所有的问题语境的话，那么我们就得让它配备有大量的器官部件，其中的每一个都对应着一个特定的语境——否则，它就无法像我们人类运用理性所做的那样，应付生命中层出不穷的种种偶然事态。很显然，从实践角度看，这样的机器设计思路是行不通的。[①]

笛卡尔的这个论证其实可以分为两个部分。第一部分的要点是：从“机器能够表达语词”出发，我们推不出“机器能够根据环境的变化而调整语义输出策略”，而后者则被笛卡尔视为“真正智能存在”的充分必要条件。我认为这个论证比较弱，因为是否能够根据环境的变化调整语义输出策略，乃是一个程度性的概念，而不是一个非黑即白的概念。在今天的 AI界，能够根据环境的变化而有限调整语义输出策略的程序，并不是做不出来，在这个问题上笛卡尔的确太低估后世 AI工程师的能力了。[②] 若按照笛卡尔的标准，这些程序的问世显然就意味着机器智能的实现——但直觉却告诉我们，这些程序的表现依然和真人智能行为大有差距。由此看来，在第一个论证中，笛卡尔关于“真正智能”的标准设置过低，这就使得他关于机器智能之不可能性的论断很容易被反例所驳倒。

笛卡尔的第二个论证的要点是：如果我们真的要做出一台“智能”机器，我们就需要把所有的问题解决策略预存在其内置方法库中，但在实践上这是不可能的。和前一个论证相比，我认为这个论证质量高得多，因为笛卡尔在此已经预见到了符号 AI的核心思路——在机器中预置一个巨大的

---

① Rene Descartes, Philosophical Essays and Correspondence, Roger Ariew (ed) (Hacket Publishing Company, 2000) 72

② 比如各种专业的拟人聊天程序（chatterbot），就能够在一定时间内欺骗人类认为他们是在和另一个人对话。其中比较有名的程序乃是 1966年问世的自然语言处理程序 ELIZA，发明人是犹太裔美国科学家魏岑鲍恩（Joseph Weizenbaum, 1923－2008），以及 1972年问世的偏执型精神分裂症患者模拟程序 PARRY，发明人是美国精神病医生科尔比（Kenneth Colby, 1920－2001）。

方法库，并设计一套在不同情境下运用不同方法的调用程序[①]——尽管符号 AI 的正式出现（1956 年）乃是笛卡尔的《方法论》出版（1637 年）三百多年之后的事情了。另外，笛卡尔在此也天才地预见到了，真正的智能将体现为一种“通用问题求解能力”，而不是特定的问题求解能力的一个事后综合。这种通用能力的根本特征就在于：它具有面对不同问题语境而不断改变自身的可塑性、具有极强的学习能力和更新能力，等等。这种“智能”观，也比较符合我们一般人的直觉。但笛卡尔的问题却在于，他认为这种“通用问题求解能力”是人类所独有的。但相关论证呢？很显然，从“所有可被我们设想的机械不具有通用问题求解能力”这个前提出发，我们是得不出笛卡尔所欲求的如下结论的：所有机械都不具有通用问题求解能力。前提和结论之间的跳跃性在于，哲学家关于机械制造可能性的设想很可能是有局限的，甚或会充溢着培根所说的“四假相”。在这里，笛卡尔显然对自己的想象力过于自信了。不过，自信归自信，他对人类理性和机器智能之间差异的提示，的确也算是一条攻击机器智能可能性的思路。在二十世纪，该路数最重要的后继者乃是美国哲学家德瑞福斯，尽管他本人并不是一个笛卡尔式的唯理派哲学家，而是一位现象学家（请参看他的著作《计算机依然不能够做什么？》）。

再来看莱布尼茨。从莱布尼茨的整个形而上学背景来看，他对于机器智能的抵触其实应当比笛卡尔还大。笛卡尔毕竟还是半吊子的机械唯物主义者，可莱布尼茨的“单子论”却是彻彻底底反唯物主义的。在他看来，构成世界的最终实体，乃是一些缺乏广延、形状和可分性的精神性单子，而物质世界所赖以存在的空间关系乃是通过诸单子的彼此知觉而产生的。站在这个立场上看，“通过机械的空间配置来产生智能”这种说法，自然就完全无法和莱布尼茨的整个哲学立场相容了。不过，和笛卡尔一样，莱布尼茨也提出了一个不那么依赖其形而上学预设的反机器智能论证（简称为“磨坊论证”），见于《单子论》第十七节（因为《单子论》篇幅很短，所以我们这里不再给出引文的页码）：

> 此外也不能不承认，知觉以及依赖知觉的东西，是不能用机械的理由来解释的，也就是说，不能用形状和运动来解释。假定有一部机器，构造得能够思想、感觉、具有知觉，我们可以设想它按原有比例放大了，大到能够走进去，就如同走进一个磨房似的。这样，我们察看它的内部，就会只发现一些零件在彼此推动，却找不出什么东西来说明一个知觉。因此，应当在单纯的实体中，而不应当在复合物或机器中去寻找知觉。因此，在单纯实体中所能找到的只有这个也就是说，只有知觉和它的变化。也只有在这里面，才能有单纯实体的一切内在活动。

我们前面刚提到，在笛卡尔看来，外部行为和人类一样具有灵活性和变通性的推理机器是造不出来的。和他的论证策略不同，莱布尼茨则玩弄了一把“欲擒故纵”的把戏，即预先假定我们已经造出了这样的一台机器。而他的论证要点则是：即使该假定本身是真的，从中我们也推不出真正的智能的存在。因为在莱布尼茨看来，真正的智能需要知觉的介入，而在机械运作的任何一个层面，我们都看不到这样的知觉的存在。所以，即使一台机器所表达出来的“输入—输出关系”和人的“输入—输出关系”完全吻合，前者依然不能算作真有智能的。

但笔者认为这个论证有很大的问题。我们姑且可以同意莱布尼茨的前提，即“知觉的存在对于智能的存在来说是不可或缺的”。但是，仅仅通过对于智能机械的内部观察，我们又如何确定知觉是否存在于这台机器中？知觉本身——而不是伴随着知觉的外部物理运作——毕竟不是掉在地

---

① 符号 AI 乃是 AI 最正统的技术思路。其基本设想是：我们在设计智能程序的时候，需要先把一个问题领域内的所有备选解题思路全部在自然语义层面上罗列出来，然后再搞清楚“先调用哪些解决策略，后调用哪些解决策略”的逻辑次序。尔后，我们再把所有这一切转化为程序，即使得一个“物理符号系统”可执行之。这是一条“从上至下”的技术思路，和本文后面提到的“联结主义”不同。

上的怀表和挂在墙上的背包，是可以在第三人称立场上被经验地观察到的。换言之，从“我们观察不到知觉的存在”，我们实在推不出那个对莱布尼茨有用的结论：知觉本身不存在。按照他的标准，我们甚至不能说人类也是有智能的，比如，我们不妨设想把莱布尼茨本人的大脑放大到上海世博园区那么大，并同时保持其中各个部件之间的比例关系不变。我们若进入这个超级大脑，看到的恐怕也只会是一些纯粹的生物化学反应，而观察不到知觉。然而，由此我们就能够推出莱布尼茨的大脑没有知觉，没有灵魂吗？这显然是荒谬的。

尽管这个论证很荒谬，但是它却直接引导了后世的塞尔提出了反对机器智能的“汉字屋论证”，[①]因此也是具有一定的思想史地位的。

三、霍布斯：符号 AI之真正哲学前驱

霍布斯是近代唯物主义哲学家的代表人物之一，但这并不是他在这里被我们提到的首要原因。这是因为，尽管 AI的理想（即制造出某种智能机器）必然会预设某种版本的唯物主义，但反过来说，从唯物主义的哲学立场中我们却未必能够推出 AI的理想。说得更清楚一点，一种关于 AI的唯物主义必须得满足这样的条件：它除了泛泛地断定心理层面上的人类智能行为在实质上都是一些生物学层面上的物理运作之外，还必须以某种更大的理论勇气，去建立某种兼适于人和机器的智能理论，以便能指导我们把特定的智能行为翻译为某些非生物性的机械运作。在这方面，拉·美特里（他可能是近代西方哲学史中最著名的唯物主义者）对于 AI的价值恐怕就要小于霍布斯，因为前者关于“人（是）机器”（L'hommeMachine）的主张，实质上并没有直接承诺智能机器实现的可能性。毋宁说，拉·美特里只是给出了一个关于人的生物属性和心理属性之间关系的局域性论题，其抽象程度要低于符号 AI的基本哲学假设：被恰当编程的符号运算，就是真正智能活动的充分必要条件（我们简称此假设为“物理符号假设”，其提出者是 AI专家司马贺和纽厄尔[②]）。

霍布斯就不同了。与迷恋医学和解剖学的拉·美特里不同，他更迷恋的乃是抽象的几何学，并致力于给出一种关于人类思维的抽象描述。他在其名著《利维坦》中写道：

> 当人进行推理的时候，他所做的，不外乎就是将各个部分累加在一起获得一个总和，或者是从一个总和里面扣除一部分，以获得一个余数。……尽管在其他方面，就像在数字领域内一样，人们还在加减之外用到了另外一些运算，如乘和除，但它们在实质上还是同一回事情。……这些运算并不限于数字领域，而是适用于任何可以出现加减的领域。这是因为，就像算术家在数字领域谈加减一样，几何学家在线、形（立体的和平面的）、角、比例、倍数、速度、力和力量等方面也谈加减；而逻辑学家在做如下事情的时候也做加减：整理词序，把两个名词加在一起以构成断言，把两个断言加在一起以构成三段论，或把很多三段论加在一起以构成一个证明，或在一个证明的总体中（或在面对证明的结论时）减去其中的一个命题以获得另外一个。政治学的论著者把契约加在一起，以便找到其中的义务；法律学家把法律和事实加在一起，以找到个体行为中的是与非。总而言之，当有加减施加拳脚的地方，理性便有了容身之处，而在加减无所适从的地方，理性也就失去了容身之所。[③]

尽管霍布斯并不可能了解后世 AI专家所说的“物理符号系统”的技术细节，但从这段引文看，他已经很清楚地意识到了，看似复杂的人类的理性思维，实际上是可以被还原为“加”和“减”这两

---

① 这是当代反对机器智能的最有名的论证，但我们这里没有篇幅讨论之。请参看 John Searle，“Minds，Brains and Programs，”TheBehavioralandBrainSciences3，pp.417－424

② 请参看 AllenNewellandHebertSimon，“ComputerScienceasEmpiricalInquiry：SymbolsandSearch，”Haugeland(ed)，MindDesign(Bradford Books，1981)41.“司马贺”是 HebertSimon先生的汉译名，此名在其生前曾得到其确认。

③ 见该书第五章开头。ThomasHobbes，Leviathan(TheUniversityofOregen Press，1999).

个机械操作的。这个讲法，在精神上和经典 AI的思想是很接近的（而我们今天已经知道了，所谓的“加法”和“减法”，其实都可以通过一台万能图灵机来加以模拟）。不难想见，如果霍布斯是对的话，那么“加”和“减”这样的机械操作就成了理性存在的充分必要条件——也就是说，一方面，从加减的存在中我们就可以推出理性的存在，而在另一方面，从前者的不存在中我们也就可以推出后者的不存在（正如引文所言，“当有加减施加拳脚的地方，理性便有了容身之处，而在加减无所适从的地方，理性也就失去了容身之所”）。很明显，如果我们承认这种普遍意义上的加减的实现机制不仅包含人脑，也包含一些人造机械，那么他对于“理性存在”的充分必要条件的上述表达，也就等于承诺了机器智能的可能性。换言之，霍布斯的言论虽然没有直接涉及人工智能，但是把他的观点纳入到人工智能的叙事系统之内，在逻辑上并无任何突兀之处。另外，就“哪些知识领域存在有加减运作”这个问题，霍布斯也抱有一种异常开放的态度。根据上述引文，这个范围不仅包括算术和几何学，甚至也包括政治学和法律学。这也就是说，从自然科学到社会科学的广阔领域，相关的理性推理活动竟然都依据着同一个机械模型！这几乎就等于在预报后世 AI专家设计“通用问题求解器”[①]的思路了。也正鉴于此，哲学家郝格兰才把霍布斯称为“人工智能之先祖”。[②] 而考虑到他的具体建树和符号 AI更为相关，笔者更情愿将其称为“符号 AI之先祖”。

但需要指出的是，符号 AI的基本哲学预设——“物理符号假设”——只是在霍布斯那里得到了一种弱化的表达，因为该假设原本涉及的是一般意义上的智能行为和底层的机械操作之间的关系，而霍布斯则只是提到了理性推理和这种机械操作之间的关系。换言之，他并没有承诺理性以外的心智活动——如感知、想象、情绪、意志等——也是以加减等机械运作为其存在的充分必要条件的。而从文本证据上来看，在正式讨论理性推理之前，《利维坦》对于“感觉”、“想象”、“想象的序列”等话题的讨论，也并未直接牵涉到对于加减运作的讨论。

那么，如何把一种机械化的心灵观从理性领域扩张到感性领域，并由此构建一种更为全面的、并对 AI更有用的心智理论呢？这关键的一步是由休谟走出的。有意思的是，走出这一步，却使得他和 AI阵营中相对新潮的一派——联结主义——攀上了亲。

四、休谟：联结主义的哲学前驱

在此笔者默认读者已经具有了休谟哲学的背景知识，并将不再过多依赖他自己的哲学术语来重构他的思想。笔者下面的重构将主要依赖当代认知心理学[③]的语言框架。

从认知心理学的视角来看，休谟的心智理论的基本思想是：一种更为全面的心智理论应当弥补前符号表征层面和符号表征层面之间的鸿沟，否则就会失去应有的统一性（而缺乏这种统一性，恰恰就是霍布斯的心智理论的毛病）。而他采取的具体“填沟”策略则是还原论式的，即设法把符号表征系统地还原为前符号的感觉原子。在《人性论》中，这些感觉原子被他称为“印象”，而符号表征则被称为“观念”。

更具体地说，他实际上是把整个心智的信息加工过程看成是一个“自下而上”的进路：

第一，人类的感官接受物理刺激，产生感觉印象。它们不具有表征功能，其强度和活跃度是物

---

① “通用问题求解器”（General Problem Solver，简称 GPS）是司马贺和纽厄尔长期投入研究的一个科研项目。顾名思义，研究者赋予 GPS的终极理想便是：它应当能像人脑一样，对各种各样的问题给出解答（而在实际的 GPS研究中，本来意义上的“通用问题求解器”其实只是一个理论基础而已，以便为更为专业的专用求解器的衍生提供一般的技术支持）。请参看：Ernst G. & Newell A., GPS: A Case Study in Generality and Problem Solving (New York: Academic Press, 1969); Newell A. & Simon H., Human Problem Solving (Englewood Cliffs, NJ: Prentice-Hall, 1972).

② John Haugeland, Artificial Intelligence: The Very Idea (MIT Press, 1985) 23

③ 现代的认知心理学大约和 AI同时诞生，其早期核心教条也是“物理符号假设”。学界一般把此派心理学主张称为“心灵的计算理论”（computational theory of mind），其核心思想便是：心灵实际上就是一个信息处理系统，而思维过程，无非就是某种形式的计算过程。

理刺激自身强度的一个函数(不过休谟不想详细讨论这个过程,因为他觉得这更是一个生理学的题目,而不是他所关心的心理哲学的题目)。

第二,感觉印象的每一个个例(token)被一一输入心智机器,而心智机器的第一个核心机制也就随之开始运作了,这就是抽象和记忆。记忆使得印象的原始输入得以在心智机器的后续运作中被妥善保存,而要做到这一点,记忆机制就首先需要对印象的个例加以抽象,以减少系统的信息储存空间,并以此提高系统的工作效率。这种抽象的产物乃是"感觉观念"。它们具有表征功能,其表征对象就是相应的印象个例。在这个抽象形式中,每一个原始个例的特征都被平均化了,而其原有的活跃程度则被削弱。

第三,每一个感觉观念本身则通过第二个心智核心机制——想象力——的作用,得到更深入的加工。想象力的基本操作是对感觉观念加以组合和分解(类似于霍布斯所说的加减运算),而这些组合或分解活动所遵循的基本规律则是统计学性质的,也就是说,观念 A和观念 B(而不是 A和 C)之所以更有机会被联想在一起,乃是因为根据系统所记录的统计数据,A的个例和 B的个例之间的联接实例要多于 A和 C之间的联接实例。由此一来,一个观念表征的所谓"含义",在根底上就可被视为对原始输入的物理性质的一种统计学抽象,而观念表征之间的联系,则可被视为对输入之间实际联系的一种统计学抽象。当然,休谟本人并没有使用笔者现在用的这些术语,他只是提到,A和 B的联接之所以被建立,乃是"习惯"使然——但这只是同一件事情的另一个说法。从技术角度看,一个模式之所以会成为习惯,就是因为该模式的个例在系统的操作历史已经获得了足够的出现次数——或者说,关于 x的"习惯"的强度,乃是关于 x的个例的出现次数的函数。

但以上所说的这些,和 AI又有何关系?

休谟并没有直接讨论人工智能系统的可能性,也许他从来都没有想过这个问题。不过,他对于人类心智模型的建构,却非常契合于后世 AI界关于联结主义进路的讨论。那什么叫"联结主义"呢?这是 AI学界内部一个相对新颖的技术流派,从上世纪八十年代开始风靡。其核心思想是:若要建立一个专门用于"模式识别"[①]的人工智能系统,不必像经典的符号 AI所建议的那样,从上至下地构建出一个内置的方法库和方法调用程序,而可以采纳一个新的技术进路:用数学办法建立起一个人工神经元网络模型,让该模型本身具有自主学习功能。这些人工神经元的底层计算活动本身并不具有符号表征功能,而只有在对整个网络的整体输出做出一定的统计学抽象之后,我们才能够将这个总结果映射到一个语义上。

今日的联结主义进路和休谟的心智模型之间的共通处体现在二者都严厉拒绝了传统的符号 AI的一层重要意蕴:我们可以先把智能体的问题求解策略尽量完美地再现出来,然后再设法把这些理性反思的产物程序化,换言之,先有符号表征描述,尔后才能够有前表征的底层运算。很显然,该想法本身就预设了:的确存在着一个为所有智能体的同类问题求解过程所共享的一般符号描述,而不同智能体实现这个抽象描述的不同运算过程,实际上只是同一轮月亮倒影在不同山川中的不同月影而已。但在休谟主义者和联结主义者看来,那一轮月亮的实在性不是被给定的东西,而至多是被构造出来的东西。用休谟的话语框架来说,那些高高在上的符号(观念),只不过就是前符号的感觉材料(印象)在心理学规则(特别是联想机制)的作用下,所产生的心理输出物而已。考虑到智能系统本身的输入历史将决定性地影响其最后形成的符号体系的结构,两个彼此不同的输入历史就必然会导致两个不同的观念表征系统——这样一来,不同智能系统在不同环境中所执行的不同的底层运作,就很难被映射到一个统一的符号层面上,并由此使得符号层获得起码的自主性和实在性。与休谟相呼应,在后世的联结主义模型建构者看来,人工神经元网络的拓扑学构架在很大程

① "模式识别"(pattern recognition)的实质,就是对一些原始信息输入加以处理,以便对这些输入进行一种合适的抽象描述。比如,通过望远镜中敌舰的侧影来判断其舰型,通过龙飞凤舞的字迹辨认出普通人可读的字符,都是"模式识别"的实例。

度上也是在前符号表征层面上运作的，而被输出表征的性质，则在根本上取决于整个网络“收敛”之前训练者所施加给它的原始输入的性质。换言之，两个识别任务相同但训练历史不同的人工神经元网络的输出结果，并不必然会（且往往不会）指向同一个语义对象。后者就像休谟眼中的“观念”一样，在整个人工神经元网络构架中处于边缘位置。

另外，休谟关于观念之间联系产于“习惯”的观点，也部分地吻合于联结主义进路对于人工神经元节点间的联系权重的赋值方式，其细节笔者就不再加以赘述了。但由于科学视野的局限，休谟并没有在神经科学的层面上重新理解心智对于前符号信息的加工过程；而他所给出的描述成果只是采用了模糊的哲学语言，没有使用定量的数学模型。这些地方也都正是今日的联结主义超越于休谟主义之处。

五、康德：“从上至下”进路和“自下而上”进路的整合者

稍有西方哲学史常识的读者都知道，康德在《纯粹理性批判》中提出了一套整合经验论和唯理论的心智理论。关于他的这套整合策略，哲学史研究方面的文献早已是汗牛充栋了。但如何跳出哲学史叙事的惯常视角，从 AI的角度来重新解读康德的这种整合策略呢？在这方面，美国 AI科学家侯世达[①]、澳大利亚哲学家查尔莫斯等人联合撰写的论文《高阶知觉、表征和类比——对于人工智能方法论的批评》[②]就颇有参考价值。文章起头部分有一段评论直接和康德相关：

> 很早人们就知道知觉活动是在不同层面上进行的。伊曼纽尔·康德将心智的知觉活动划分为两个板块：其一是感性能力，其任务是拣选出那些感官信息的原始输入，其二是知性能力，其任务是致力于把这些输入材料整理成一个融贯的、富有意义的世界经验。康德并不对感性能力很有兴趣，并将主要精力投向了知性能力。他孜孜以求，最终给出了一个关于高阶认知的精细模型，并通过该模型将知性能力区分为十二个范畴。尽管在今天看来，康德的这个模型多少显得有点叠床架屋，但他的基本洞见依然有效。依据其洞见，我们可以将知觉过程视为一道光谱，并出于方便计，将其区分为两个构成要素。大约和康德所说的感性能力相对应，我们划分出了低阶知觉。这主要指的是这样一个过程：对从不同感官通道搜集来的信息进行早期处理。另外，我们还划分出了高阶知觉——通过这种知觉，主体获得了对于上述信息的一种更为全局性的视角，并通过和概念的沟通而抽象出了原始材料的意义，最终在一个概念的层次上使得问题求解的情景具有意义。这些问题求解情景包含：对象识别、抽象关系把握，以及把某个具体环境辨识为一个整体。

从这段引文看，康德对于 AI科学家的启发就在于：知觉的“从上至下”进路（“知性”或“高阶知觉”）和“自下而上”进路（“感性”或“低阶知觉”）都是不可或缺的，因此一个更全面的人工认知模型将囊括这两者。但这里的问题是：凭什么说两者都不可或缺呢？或者说，仅仅遵从休谟式的“自下而上”思路，或者仅仅遵从霍布斯式的“从上至下”思路，为何就行不通？

康德本人对于这个问题的解答是：如果我们仅仅遵从“自下而上”的思路，我们就很难解释，为何人类的心智仅仅凭借经验联想，就能够构成普适性的“先天综合判断”（回答不了这个问题，我们将陷入对于普遍性知识的怀疑论）；如果我们仅仅遵从“从上至下”的思路，我们很难解释，为何我

---

① 道格拉斯·霍夫斯塔德（Douglas R. Hofstadter, 1945—），其汉化名字为“侯世达”，美国著名学者、计算机科学家，印第安纳大学伯明顿分校计算机科学和认知学教授，观念与认知研究中心主持人。

② Chalmers D. J. et al. (1991). Chalmers D. J., French R. M., Hofstadter D., “High-Level Perception, Representation, and Analogy”. Journal of Experimental & Theoretical Artificial Intelligence 4.3 (1992): 185–211. 这篇文章的署名以查尔莫斯居先，但实质上更多体现的是侯世达的思想（查尔莫斯现在是英语世界最当红的哲学家，但是他当年曾是侯世达的博士研究生）。这一点是笔者的朋友、计算机科学家王培先生（他也曾做过侯世达的博士研究生）告诉笔者的。所以，下面的行文将默认此文的实际第一作者为侯世达。

们心智机器的最终输出能够和外部输入发生关联（回答不了这个问题，我们将陷入“观念实在论”或“哲学独断论”）。不过，康德的这些解释带有过重的知识论气味，而且还负载了很多哲学预设（比如，他预设“哲学怀疑论”和“哲学独断论”肯定都是错的）。站在 A.I.或者认知科学的立场上看，我们需要的，其实是一种哲学预设更少的对于整合式路径的辩护方案。

侯世达等人的相关辩护方案则机智地绕开了“先天综合判断”这个麻烦话题，而以“类比”为切入点。他们的问题是：如果要在一个人工智能系统里实现“类比推理”的话，编程者的编程思路，到底要遵循“自下而上”的进路，还是“从上至下”的进路呢？或是二者的整合进路？

那么，为何要以“类比”为切入点呢？这当然是因为类比推理对于提高智能系统的工作效率很重要。不难想见，一个智能系统若能够在表征 A和表征 B之间建立起合适的类比关系的话，那么只要系统已经预存了一套关于表征 B的问题求解策略 C，那么它就能够用 C来解决关于表征 A的新问题。系统由此获得的问题求解效率，自然将大大高于其从头搜索 C的效率。类比推理的一般形式就是：

方法C —解决→ 问题A

问题A ←类比→ 问题B

方法C —解决→ 问题B

不过，要建立起这样的一个类比关系，却不是易事。请考虑对如下类比关系的建构过程：[①]

类比一：孔明之于刘玄德，可类比于管仲之于齐桓公。

假设一个智能系统已经把握了“管仲”、“齐桓公”、“孔明”和“刘玄德”这四个表征的含义（但下面我们将马上提到，即使要满足这个假设，也非易事。另外，关于什么叫表征的“含义”，我们暂且不表），但这不等于它很快就能够建立起我们所欲求的这种类比关系。不难想见，系统的知识库里还存有很多别的表征，比如“张飞”、“蒋干”、“貂蝉”、“董卓”，等等。换言之，在建立“类比一”之前，系统实际上需要做一道选择题：

孔明之于（　），可类比于管仲之于（　）。

A. 张飞、B. 蒋干、C. 董卓、D. 貂蝉、E. 齐桓公……

而面对这些杂乱无章的选择项，系统完全也可能建立起错误的类比关系，比如：

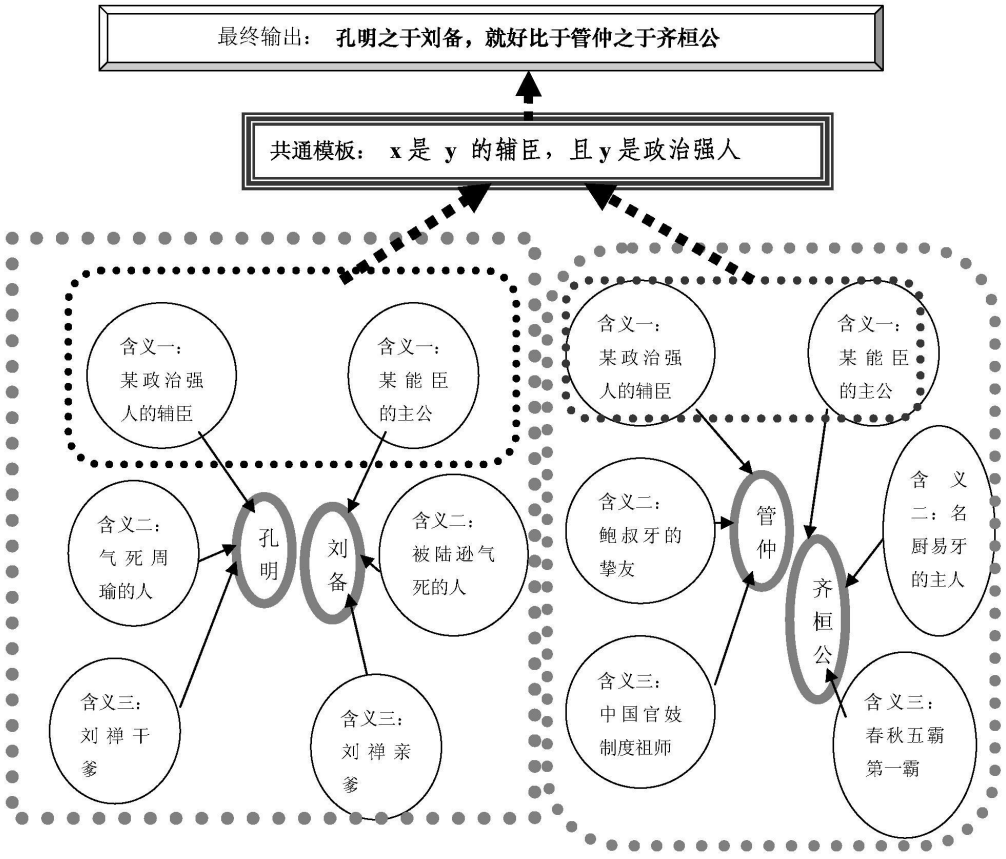
类比二：孔明之于董卓，可类比于管仲之于貂蝉。

怎么避免这一点呢？休谟主义者在面对这个问题时或许又会祭出“习惯”的法宝，也就是说，如果系统检测到“孔明—刘玄德”关系和“管仲—齐桓公”关系有比较多的共现次数的话，那么系统就会在“孔明—刘玄德”关系和“管仲—齐桓公”关系之间建立起一种更高阶的类比关系。但这种统计学的策略有两个根本缺陷：第一，很多对问题求解有用的新类比关系，往往是缺乏统计数据支持的（否则就谈不上是新类比关系）；第二，该策略对于系统输入历史的这种高度依赖性，将大大削弱系统对于输入信息的主动鉴别能力。比如，若系统恰好发现“貂蝉—董卓”关系和“管仲—齐桓公”关系有比较多的共现次数的话，那么它就会随波逐流地在这两者之间建立起一种更高阶的类比关系。但如此一来，系统又如何有机会对这种错误的建构做出主动修正呢？

面对同样的问题，霍布斯主义者的表现或许会更为狼狈。霍布斯—经典 A.I.思想路线的要点就在于，整个认知系统必须在符号表征的层面上运作，换言之，他们都默认了正确表征的存在已然不成为问题。但在真实的“类比关系匹配”任务中，成为问题的，往往就是如何找到正确的表征形式。再以“孔明之于刘玄德，可类比于管仲之于齐桓公”为例。现在我们暂且遵循弗雷格以来的语言哲学传统，把一个词项的含义看成是把该词项映射为一个外部对象的函数。比如，“孔明”的含义，就是把该词项映射为历史上真实存在过的那个人的函数。这样的映射方式肯定很多，比如你可以将

① 下面的例子自然不是侯世达等人原本所举的。笔者根据汉语读者的阅读习惯和文化背景，做了一些改写。

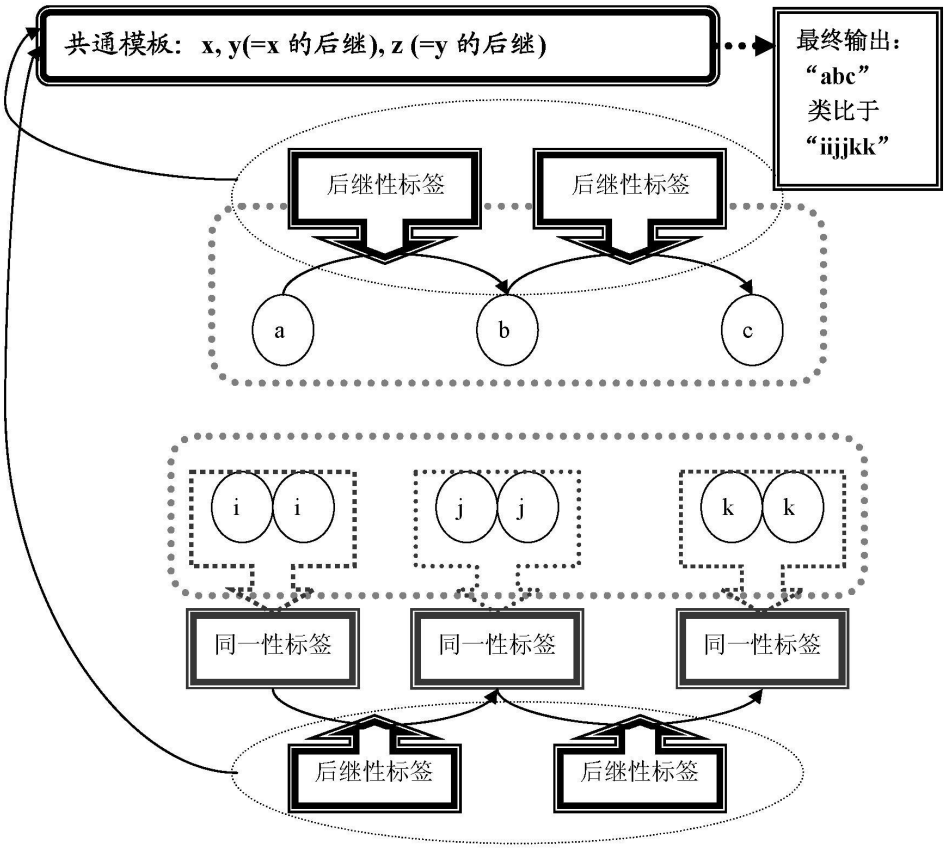
“孔明”视为“刘禅的亚父”、“三国时蜀国的丞相”、“《隆中对》的作者”、“刘备最有名的文臣”，等等（其中的每一个都能够把“孔明”映射到同一个对象上去）。而现在的问题就是，若要建立“孔明之于刘玄德，可类比于管仲之于齐桓公”这个类比关系，我们需要的又是其中怎样的一种表征形式呢？依据一般中国人的历史常识来判断，答案显然就是“刘备最有名的文臣”，因为这样我们就可以将其匹配于管仲的表征形式“齐桓公最有名的文臣”，并在这种匹配的基础上建立起我们所需要的类比关系。该匹配流程可示意如下：



但麻烦的是，我们又到底如何能在“刘备—孔明”关系属性集以及“管仲—齐桓公”关系属性集中，找到一个为两集所共享的成员呢？很显然，这个关键性的表征形式并不会自动跳出来让系统注意到自己。而要让系统用野蛮搜索的方式来自上而下地逐一寻找它，则又显得过于耗时。因此，系统就需要用某种自动搜索程序来发现它。欲建立这种搜索程序，我们就得为系统设计出一个低层次人工知觉能力以模拟康德的“感性”能力，并由此迅速检索与任务求解更为相关的表征形式；同时，让高层次的人工知觉能力（类似于康德的“知性”能力）实时地参与其中，构成高—低互动。换言之，无论是霍布斯—经典 AI的道路，还是休谟—联结主义的道路，都无法引导我们设计出能够正确地建立起所需类比关系的系统。只有康德式的整合式策略，才是我们努力的方向。

在康德哲学的启发下，侯世达等建立了一个专门的类比关系搜索程序，名字叫“照猫画虎”（Copycat）。“照猫画虎”的工作环境是一大串字母串，每一串字母串构成了系统的一个原始输入，比如“abc”、“iijkk”、“eejkk”等。系统的任务是找出每个输入的内部结构规律，并在此基础上将一个输入看成是另一个输入的类比物。比如，“abc”和“iijkk”之间就有这种类比关系，因为前者由三个单元“a”、“b”、“c”构成，每个单元的右边都是自己在字母表中的后继者（同样的关系也存在于“ii”、“jj”、“kk”之间）。很显然，同样的类比关系就不存在于“abc”和“eejkk”之间，因为“c”的后

继不是“ ŷ”, 而是“ f”。请看如下示意图 (笔者根据原文精神自绘):



而要让系统也能够辨识出这种类比关系, 我们就得一一建立系统中的如下构成要素:

1. 人工“感性”能力。也就是说, 系统的输入系统必须有能力辨识出每一个字母串的记号构成, 并辨识出一个输入和另一个输入之间的界限。这一步比较简单, 没有什么可以说的。

2. 人工“想象力”。在康德的心智理论中, “想象力”是介于“感性”和“知性”之间的一种能力, 其任务是对感官输入进行初步处理, 以便为知性的高级操作做准备。从分类上看, 它可以从属于一种更为宽泛的“感性” (实际上, 上文所谈的“感性”就已包含了“想象力”)。在“照猫画虎”程序中, 这就对应于这样一个设计: 系统配置有一些自动运作的“短码算子” (codelet), 其任务是对“人工感性”输送来的信息进行初步结构分析。这就为人工范畴表的运作提供了基础。

3. 人工“范畴表”。康德心目中的知性范畴表, 大致对应于“照猫画虎”程序中的“滑溜网” (slipnet)。所谓“滑溜网”, 就是由不同的范畴所结成的一个网络, 其中的每一个范畴都对应着一个更低层面上的短码算子 (比如, 若在更低的层面上有“同一性短码算子”, 那么在“滑网”中就必定有一个“同一性”范畴与之呼应)。该网和诸短码算子之间的相互协作方式乃是这样的: 一方面, 一个短码算子的工作输出的性质构成了与之对应的那个网络范畴节点的激发条件 (这是一个由下而上的进路); 另一方面, 一个网络范畴节点的激发状态又反过来决定了系统的资源应当倾向于那些短码算子 (这是一个从上至下的进路)。

综上所述, 诸“短码算子”的自主运作为范畴节点的启动提供了条件, 而后者的启动又会反过来指导前者的资源分配方向。两个层面相辅相成, 合力完成了建立类比关系的任务。就这样, 康德的名言“概念无直观则空, 直观无概念则盲”, 在 AI时代便获得了这样一种全新的诠释形式: “滑溜网无短码算子则空, 短码算子无滑溜网则盲”。这种“无心插柳柳成荫”的效果, 恐怕是康德本人也

始料不及的。

笔者认为，康德式的“从上至下”和“自下而上”相互整合的进路，其启发意义不仅局限于类比模型的构建，而且还可以被运用于其他的 AI 研究领域，比如机器视觉。但若要真正地做出这样一种推广，仅仅按照“照猫画虎”程序的模式去从事研究，恐怕还远远不够。比如，在“照猫画虎”程序中，系统所处的人工环境本身就已经是一个被高度数理化的世界（这个环境所提供的有效输入，都已经是字符串了）。这固然便利了程序设计员接下来的程序设计流程，却大大歪曲了康德的如下原初设想：人类的原始认知境遇，乃是一片没有数理描述形式的“混沌”——换言之，数理描述形式本身只可能是心智运作的产物，而不可能是被自然给予的。但如何能够设计出一个更基本的程序，以便让系统能把一个实际的工作环境自动转化为一个数理化的环境模拟形式呢？恰恰在这个问题上，“照猫画虎”程序的设计思路采取了回避策略。由此看来，侯世达等人的这项工作虽然很精彩，但这也只是在一个方向上体现了康德哲学的某种理论意图，而绝未穷尽康德思想库中的宝藏。

六、总　　结

笔者希望本文的讨论，能够带给读者以下三点启示：

第一，看似新锐的“AI 哲学”，其实并不是崭新的东西，而的确和西方哲学史有着密切的联系。从抽象的角度看，哲学思辨切入人工智能的方向主要有两个：其一，机器智能的实现是否先天可能？其二，怎样的心智理论才能够为机器智能的实现提供更好的参照系？而从本文的哲学史梳理结果来看，笛卡尔、莱布尼茨等哲学家实际上已经超越了自己时代的科学发展的限制，明确提出了第一个问题，并给予了其以否定性的应答（不过本文的讨论也已经表明了，他们的反机器智能的论证都是有问题的）。而霍布斯则间接地肯定了机器智能的可能性。休谟和康德虽未正面谈论该问题，但是他们各自提供的心智理论，却分别构成了 AI 中的联结主义进路和“上下整合”进路的哲学前驱，并由此为上述第二个问题提供了答案。从某种意义上说，今日在英美方兴未艾的 AI 哲学，依然没有从根底上跳出这两个问题所规定的理路。由此看来，十七、十八世纪欧洲哲学家对于相关问题的前瞻能力，乃是令人惊异的。

第二，虽然经典的 AI 进路包含着对于数理模型的高度推崇，但同样推崇数理描述方式的“唯理派”哲学家，却往往对“机器智能”持有敌意。这是因为，对于“机器智能”的赞成不仅仅依赖于对于数理模型的推崇，而且还依赖于一种对于身心关系的唯物主义观点。但由于种种文化、宗教因素，唯理派哲学家往往在身心关系问题上持反唯物主义立场。从这个角度看，近代唯理派和经典 AI 之间的亲缘关系，并没有一些论者（如德瑞福斯在其《计算机依然不能做什么？》中）所说的那么强。

第三，作为十七、十八世纪欧洲哲学的集大成者，康德虽没有直接讨论过机器智能的可实现问题，但是他的心智理论对于 AI 的启发意义却依然不容小觑。此理论的要点就是把“从上至下”和“自下而上”的两个认知进路加以打通，将其整合在一个更大的心智模型里。笔者认为，这种整合式的进路要比单纯的“自下而上”进路或“从上至下”进路更具有解释力，因此应当是未来 AI 建模的一个主要参照模式。但如何把这种哲学启发转化为更具体的编程工作，却会面临着一个巨大的理论—技术障碍，即如何把系统所在的非数理化的实际工作环境加以实时的数字化模拟（这种模拟必须由系统自己完成，而不能由程序员事先输入）。在这个问题上，侯世达等人的“照猫画虎”程序并没有为我们提供一个完美的工作模板。总之，更艰巨的任务还在等待 AI 专家们去完成。

The Prototypes of the Very Idea of Artificial Intelligence in the 17th and 18th Century European Philosophy

XU Ying-jin

(School of Philosophy, Fudan University, Shanghai 200433, China)

Abstract: Although Artificial Intelligence (AI) is a relatively new discipline emerging in the 20th century, its root can be traced back, minimally speaking, to the 17th and 18th century European Philosophy. Philosophers in these two centuries, knowing nothing about modern computer science notwithstanding, did approach to the philosophical foundation of AI by raising at least two questions: First, is it conceptually possible to build up a mechanical model which can perfectly behave as a genuine intelligent agent? Second, if the answer of the foregoing question is positive, what kind of the theory on human mind can be the best reference for the required model? I will argue that both Descartes and Leibniz offered a negative reply to the first question, while Hobbes offered a positive one. So far as the second question is concerned, Hume's reconstruction of cognitive architecture, as a theory of mind in its face-value, can be viewed as the blueprint of the contemporary connectionist approach in AI, whereas Kant's attempts to negotiate between intuition and thinking is echoed by AI scientists' attempts to integrate both the "bottom-up" and "top-down" approaches as whole. In a nutshell, the 17th and 18th century European Philosophy can be viewed as a potential "wisdom bank" for AI, although the mainstream of the AI circle is still ignorant of this fact.

Key words: Artificial Intelligence (AI); Symbolic AI; analogue; connectionism; top-down approach; bottom-up approach

[责任编辑　晓　诚]