世纪前沿

The Philosophy of Artificial Intelligence

[英] 玛格丽特·A·博登 编著

Margaret A. Boden

刘西瑞 王汉琦 译

人工智能哲学

上海世纪出版集团

人工智能哲学

[英] 玛格丽特·A·博登 编著　刘西瑞　王汉琦 译

世纪出版集团　上海译文出版社

图书在版编目(CIP)数据

人工智能哲学/(英)博登(Boden, M. A.)编著;
刘西瑞、王汉琦译. —上海:上海译文出版社
2006.7
(世纪人文系列丛书)
书名原文: The Philosophy of Artificial Intelligence
ISBN 7—5327—4054—4

Ⅰ. 人 … Ⅱ. ①博…②刘…③王… Ⅲ. 人工智能—技术哲学—文集 Ⅳ. TP18—02

中国版本图书馆 CIP 数据核字(2006)第 061800 号

责任编辑 赵凤珍

装帧设计 陆智昌

人工智能哲学
[英]玛格丽特·A·博登 编著
刘西瑞 王汉琦 译

出 版 世纪出版股份有限公司 上海译文出版社
(200001 上海福建中路 193 号 www.ewen.cc www.yiwen.com.cn)
发 行 上海世纪出版股份有限公司发行中心
印 刷 上海商务联西印刷有限公司
开 本 635×965mm 1/16
印 张 31.75
插 页 4
字 数 427 000
版 次 2006 年 7 月第 1 版
印 次 2006 年 7 月第 1 次印刷
ISBN 7—5327—4054—4/B·249

定 价 43.00 元

出版说明

自中西文明发生碰撞以来，百余年的中国现代文化建设即无可避免地担负起双重使命。梳理和探究西方文明的根源及脉络，已成为我们理解并提升自身要义的借镜，整理和传承中国文明的传统，更是我们实现并弘扬自身价值的根本。此二者的交汇，乃是塑造现代中国之精神品格的必由进路。世纪出版集团倾力编辑世纪人文系列丛书之宗旨亦在于此。

世纪人文系列丛书包涵“世纪文库”、“世纪前沿”、“袖珍经典”、“大学经典”及“开放人文”五个界面，各成系列，相得益彰。

“厘清西方思想脉络，更新中国学术传统”，为“世纪文库”之编辑指针。文库分为中西两大书系。中学书系由清末民初开始，全面整理中国近现代以来的学术著作，以期为今人反思现代中国的社会和精神处境铺建思考的进阶；西学书系旨在从西方文明的整体进程出发，系统译介自古希腊罗马以降的经典文献，借此展现西方思想传统的生发流变过程，从而为我们返回现代中国之核心问题奠定坚实的文本基础。与之呼应，“世纪前沿”着重关注二战以来全球范围内学术思想的重要论题与最新进展，展示各学科领域的新近成果和当代文化思潮演化的各种向度。“袖珍经典”则以相对简约的形式，收录名家大师们在体裁和风格上独具特色的经典作品，阐幽发微，意趣兼得。

遵循现代人文教育和公民教育的理念，秉承“通达民情，化育人心”的中国传统教育精神，“大学经典”依据中西文明传统的知识谱系及其价值内涵，将人类历史上具有人文内涵的经典作品编辑成为大学教育的基础读本，应时代所需，顺时势所趋，为塑造现代中国人的人文素养、公民意识和国家精神倾力尽心。“开放人文”旨在提供全景式的人文阅读平台，从文学、历史、艺术、科学等多个面向调动读者的阅读愉悦，寓学于乐，寓乐于心，为广大读者陶冶心性，培植情操。

“大学之道，在明明德，在新民，在止于至善”（《大学》）。温古知今，止于至善，是人类得以理解生命价值的人文情怀，亦是文明得以传承和发展的精神契机。欲实现中华民族的伟大复兴，必先培育中华民族的文化精神；由此，我们深知现代中国出版人的职责所在，以我之不懈努力，做一代又一代中国人的文化脊梁。

上海世纪出版集团
世纪人文系列丛书编辑委员会
2005年1月

目录

导　言

人工智能(AI)有时被定义为：研究怎样制造计算机，并(或)为其编程，使其能做心灵所能做的那些事情。这些事情中有一些被公认为是需要智能的：开药方和(或)作医嘱，提供法律或科学咨询，证明逻辑或数学定理。另外一些事情则不同，它们与教育背景无关，是所有正常的成年人都能做到的(有时甚至人类以外的动物也能做到)，其特点是不受意识支配，如看到阳光下的物体和影子，找到穿过复杂地形的小路，把木桩塞进洞里，用母语讲话，以及运用自己的常识。

由于上述定义涵盖了与这两类心理能力有关的 AI 研究，所以它胜过把 AI 说成是让计算机去做“人类需要运用智能才能做的事情”的定义。然而它有一个预设假定：计算机所能做的就是人脑所能做的，计算机真的可以开处方，提建议，做推理，善理解。如果将 AI 定义代之以 AI 是“计算机的发展，而这些计算机的外在性能具有我们认为是属于人类心理过程的那些特征”，我们就有可能回避这一尚存争议的假定(同时也避开了计算机在做这些事情时是否采用了与我们**相同的方式**这一问题)。这一适度的特征描述可以为一些 AI 工作者所接受，尤其是那些把眼光投向为商业目的而生产技术工具的人们。

但是也有不少人偏爱一个更有争议的定义，即把 AI 看作是**一般性的智能科学**，或更确切地说，看作是认知科学的智力内核。这样，它的目标就是提供一个系统的理论，该理论既可以解释(也许还能使我们复制)意向性的一般范畴，也可以解释以此为基础的各种不同的心理能

力。它不仅要包括地球上各种生物的心理，而且还要包括全部可能存在的心灵。它必须告诉我们，智能是仅仅体现于那些具有大脑般的基本构造(包括由关联细胞组成的网络中的并行处理过程)的系统之中，还是也可以用某种别的方式来实现。这样，由于“计算机”已退出了这一定义，它们与这样一门科学的特殊关系必须加以确证。这一雄心勃勃的事业是否能够成功(如果能够，又怎样成功)，抑或它是根本错误的想法，这个问题引发出许多与 AI 相关联的哲学问题。

因此，AI 哲学(这里把 AI 看作是一般性的智能科学)同心灵哲学、语言哲学以及认识论紧密相联，同时又是认知科学哲学，特别是计算心理哲学的核心。计算心理学家们共同约定了四个哲学假定。对待心灵和智能，他们采用机能主义的方式，认为心理过程是能够被精确说明的过程，而心理状态则取决于它们与感觉输入、动作行为以及其他心理状态的因果关系。他们把心理学看成是对心理表象所藉以构成、解释和变换的那些计算过程的研究。他们把大脑视为一种计算系统，所关心的是它体现出何种函数关系，而不是哪些脑细胞体现出这些关系，或大脑生理机能怎样使这种体现方式成为可能。他们虽然并不认为(AI 工作者也是如此)哪一些 AI 概念和计算机模型的方法论在认识智能方面可能是最有帮助的，但是他们都认为，某种 AI 概念一定会成为心理学理论基本内容的组成部分。

以与 AI 概念极为类似的概念对智能加以解释，是哲学家们长久以来的梦想，可以认为从柏拉图开始就是如此(见第 13 章)。在过去的许多个世纪中，这个梦想孕育了形而上学理论，产生出对心理机能的形式说明，甚至是解释模型——我们的脑海中会浮现出霍布斯、莱布尼兹和巴比奇这些人物。到了 20 世纪，这一思想的智力资源因三个方面的进展而更加丰富：形式计算理论，为实现形式上规定的计算而设计的功能计算机，以及神经元的发现。

这三个发展奠定了整个 AI 的基础，虽然有些研究对其中某一方面的发展利用得比较明显。当前的 AI 研究一般分为两大类别：“传统”

类［或 GOFAI[1]，即“有效的老式 AI”（Haugeland 1985）］和“联结论”。虽然它们之间的理论关系尚有争议，但是它们的历史关系是清楚的：它们是从同一个根上生长出来的分支，共同发轫于由神经心理学家兼精神病学家 W · 麦卡洛克和数学家 W · 皮茨合著的开创性之作（见第 1 章）。

麦卡洛克和皮茨合著的文章题为“神经活动内在概念的逻辑演算”，这一题目本身就表明了传统 AI 和联结论 AI 的共同继承权。他们关于实施“逻辑演算”的设想，影响到冯 · 诺伊曼对数字计算机的设计，同时鼓舞着 AI 的先驱者们尝试建立思维的形式模型。他们对于“神经活动”的讨论使赫布的细胞组合生理心理学理论获益匪浅，并促成了多种多样的神经网络模型——这正是今日联结论系统的先驱。

这篇文章之所以具有重大影响，主要是因为它虽有臆想的成份，但决不仅仅是一种推测。毋庸讳言，文章作者关于目的、学习、精神病学的神经体现方式——更不必说认识论、实在论、普遍性、价值以及数字的神经体现方式（McCulloch 1965）——的大胆构想只是提纲挈领式的论述。但是麦卡洛克和皮茨并不是简单地持有一般的唯物主义的立场，认为智能是由大脑实现的，他们**证明了**：一定类型的（可严格定义的）神经网络，原则上能够计算一定类型的逻辑函数。

他们知道，神经系统是由相互联系的细胞组成的，这些细胞的激活表现为全或无的形式，并取决于阈值和其他细胞的活动性。他们也了解图灵关于可计算数字的文章（Turing 1936），以及罗素和怀特海在命题演算方面的工作。他们在整合这些不同资料的基础上，证明了有关理想化神经网络逻辑特性的各种定理。例如，每个命题演算函数都可以由某种（类型相当简单的）网络来实现；每个网络计算出一个可由一台图灵机计算的函数；同时每个图灵可计算函数可以通过某个网络来计算。图灵机具有一条无限长的纸带，也就是说，它是一种数学上的理

[1] GOFAI 是 Good Old Fashioned AI 的缩写。——译者

想形式，而不是一台实际的机器。 由于神经网络是有限的，我们不可能通过证明一种一般化的、也许是无法实现的可能性，来恰当地解释被体现的智能。 相反，我们必须确定哪些网络能够实现一些特殊的功能。 这样，理论心理学的任务就是要设计出具有计算能力的网络，而这种计算原先是由心灵完成的。

AI 的任务就是确定和设计这种网络，并通过构造工作模型获得补充，而作为特例的人类心理也包括在这项(关注实际的和可能的心灵的)任务之中。 如果把“网络”看作是真实的神经联结的近似形式，那么我们就得到一个广泛的联结论的研究纲领。 由于被解释成神经活动高度抽象的理想形式，我们的主要着眼点是二值逻辑，而不是真正的细胞联结性和阈值，所以典型的传统 AI 是以数字式信息加工方式出现的。麦卡洛克和皮茨此文的研究成果对于这两种类型的 AI 研究都具有开创性意义。 60 年代后期，神经网络研究一度低落，部分原因是将只适用于一小类网络的批评作了过分的普遍化(Minsky and Papert 1969)。 本书后面第 11—13 章所讨论的内容即是神经网络研究再度兴起后的成果，它(与 AI 和心理学中大多数联结论模型一样)并没有试图从构造上与可识别的神经联结相对映，这一艰巨任务常常是由神经科学家来承担的(见第 14 章)。

图灵关于可计算数字的文章(Turing 1936)，对两种 AI 研究的途径来说，都堪称理论上的奠基之作，文中将计算定义为： 应用形式规则，对(未加解释的)符号进行形式操作。 “有效过程”———种可严格定义的计算过程——的一般观念，是通过数学演算的例子来说明的，但是(正如麦卡洛克和皮茨所意识到的那样)，这就意味着，如果智能可以普遍地用在大脑中实现的有效过程来解释，那么一台普适的图灵机，或是某种与之近似的实际机器，就可以对其进行模拟。 1950 年时，图灵，还有其他人，已经制造出通用数字计算机，他们被用来模拟智能的某些方面。 在“计算机器与智能”(第 2 章)一文中，他特地提出了这种机器能否思维的问题。

他指出，对这个问题的回答，不应当依据预设的(很可能是武断的)“思维”定义，而是应当问一问某种可构想出的计算机是否能够表演“模仿游戏”，才能作出判断。无论是做加法还是阅读十四行诗，一台计算机能以无法与人类回答相区别的方式来回答提问者的问题吗？这个问题(常常表达为一台计算机是否能通过“图灵检验”)包括三个方面：某个未来的计算机真的有能力以所设想的方式回答问题吗？无论在人类还是在计算机中，有效过程原则上能够生成这种性能吗？这种性能足以使计算机具备智能属性吗？图灵本人对每一问题的回答都是：“是的。”

图灵的立场受到来自三个迥然不同方面的攻击(这三个方面既不相互排斥，也没有必然的联系)，每一方面都存在许多重要的变化形式。

第一类攻击采用了一套常见的反行为主义的论据，拒绝作为智能充分判据的模仿游戏——此外，并未拿出任何专门与 AI 有关的东西。然而，即使如反行为主义者所坚持的那样，意识体验对于智力来说是一个必要条件，如果不能在未加论证的直觉以外提供新的论据，以说明意识显然不可能产生于计算机的话，就不能证明智能计算机是不可能的，一般说来，反行为主义的论据所能证明的，充其量不过是：一台高性能的计算机并**不需要**是智能的。

支持 AI 的人会表示赞同，因为他们采取了机能主义的立场，认为智能必须包含某种系统式的因果过程(计算)。然而，行为，无论在表面上给人以何等深刻的印象，仍是来自某个庞大的事先存储的查寻表，而不同于结构式的(有可能反映出习俗心理的精神范畴的)过程和表象，因此行为不能看作是智能的(Sloman 1986)。从图灵的判据来看，他只规定了，作为智能行为基础的原因是**某种**有效过程。而且，由于他没有明确指出，思维**必须**包括思想者内部的基本原因，他的判据就不排除通过魔术或偶然地引起的行为：果实从被风吹动的树上落下，掉到电传机键盘上，可能会“愚弄”一个正在玩模仿游戏的提问者。

对图灵立场的第二类攻击，集结着这样或那样主张计算机**不可能**有

智能的另一些看法。有一种看法认为，图灵依据于言语行为甚至比通常的行为主义更加不可取，因为不仅缺少运动行为，甚至没有活的身体外形，确切地说，心理属性不能归属于计算机(Dreyfus 1979)。另一种反对意见认为，即使计算机能够像图灵设想的那样去行事(诸如阅读十四行诗)，它也并非真的具有智能，因为不能设想计算机真的会思考和理解：没有意向性，就没有智能。这一指责对于作为技术，甚至作为执行模仿的 AI 并不构成威胁，因为它承认完全具有人类特点的计算机性能是有可能存在的。此外，它并不否认，计算机模型在心理学中(像在其他科学中一样)可起到厘清理论的作用。但是它仍然坚持，AI 的概念内容不能帮助哲学家或心理学家去描述或解释心理过程本身，因为心灵具有意向性，而计算机没有，也**不可能**有。

在这种攻击中，一个颇有影响的例子是 J · 塞尔的文章“心灵、大脑与程序”(第 3 章)，该文运用图灵自己的计算概念来反驳图灵的观点：一台配有适当程序的计算机是有智能的。塞尔并没有把批评直接对准图灵的文章，而是指向该文章提出的两个纯理论条款：“强”AI(尝试通过编程构造真正的心理能力)和计算心理学(在这方面 AI 对心理学理论的内容有所贡献)。

塞尔的第一个论点，包括他构想出来的“中文屋”，认为 AI 程序和计算机模型当然地是纯形式句法的(和图灵机一样)。基于这一点，他认为，一个系统不可能纯粹借助完成计算而达到理解。所以计算心理学决不可能解释我们的心理能力，任何一个程序**更加**不可能将智能赋予计算机。塞尔的第二个论点是，智能或意向性不仅需要心灵式的行为，同时还需要作为这一行为的基础的“正确的因果能力”。如前所述，对这样表达的论断，支持 AI 的人们是可以接受的。然而塞尔在定义这种因果能力时，根据的不是特性或功能，而是材料**质别**。此外，他还认为，从直觉上看，显然神经蛋白可以生成意向性，而金属和硅则不能。

我在反驳中指出(见第 4 章“逃出中文屋”)：即使最简单的程序也

并不是纯形式主义的，而是具有某种相当本原的语义特性，所以从根本上说，计算理论并非不能解释意义。 此外，只要大脑生成意向性的能力是清楚的，而不是完全反直觉的(头盖骨里面那些黏糊糊的物质如何**可能**进行理解?)，这种认识所采用的信息加工方式同样可以用于计算机。 这样，AI的概念就完全有理由被用作心理学理论的基本组成部分，同时某些想象之中的计算机也可以具有与意向性和智能十分近似的能力。 任何一台计算机，即使它的内部计算组织与我们的完全等同，是否可以**丝毫不含引喻**地将其称为有智能的，则又是另一回事情，排除引喻不仅需要对事实的识别，还需要人类自身的道德评判(Boden 1987：423－5)。

攻击图灵立场的第三条阵线(第5—14章与之有种种联系)持有的观点是：与图灵的假定相反，要使计算机的表现在深度、广度以及灵活性上与人类心智相媲美，在原理上和(或)实践上，都是不可能的。 一台阅读十四行诗的计算机，不管是真有智能，或者仅仅只能模仿智能，都决无存在的可能。 这种攻击所依据的常常就是图灵文章所驳斥的那些观点的变种：行为、创造性以及哥德尔定理的“非形式特性”(不能约简为规则的特性)的观点(Dreyfus 1979；Lucas 1961)。 此外，技术性AI并非不能存在，实际上已经制造出实用的AI系统，但是AI和计算心理学的最高目标——人类心理过程的详尽的计算机模型——是不可能的，也是(或是)不可行的。

有些哲学家或许会反对说：这与不可行性无关，这里的问题是逻辑的可能性，而不是经验的可能性。 这种回答忽略了“经验上可能的”在较为抽象的和较为实际的意义之间的差别，对于单独存在的基本科学原理和**受到非常普遍的现实世界的制约**的科学原理，应当区别对待。

图灵机在两种意义上都表现为经验上的不可能性，因为它有一个无限长的纸带。 其他一些计算机器，就像传说中的地狱雪球一样，虽然没有被基本的科学原理所否认，但在现实世界中，由于时间和(或)空间

的限制，也是无法实现的。例如，将野蛮搜索的算法用于下棋，每走一步都可能需要天文数字般的时间长度(虽然是有限的)。同样，即使加工单元的反应能够比神经元快得多，许多视觉任务也只有采用大规模并行处理方式，才可能在实际时间内完成。原则上讲，这种处理方式可以用一台串行计算机来模拟(所以某些理论问题能够在不规定串行或并行实现方式的情况下被提出来)。但是在实践中，只有相对比较小的并行系统才能如此完成。既然我们的大脑不是用天文数字般的时间工作的，我们就不应当将我们的心灵归结为实施过程需要上千年时间的计算形式。现实世界进一步的限制是我们进化的起因，正如克拉克所说(Clark 1987a)，如果某些类型的计算与进化是一致的，而另一些不一致，这就不仅牵涉到心理学，也牵涉到心灵哲学和 AI 哲学。

在那些与图灵一样相信 AI 可行性的人之中，无论在实践上使 AI 成为现实，还是在把它用于细化的心理学问题(既有抽象的任务分析，也有细致的实验观察)上，没有人比 A · 纽厄尔和 H · 西蒙做得更多了，正如“作为经验探索的计算机科学：符号和搜索”(第 5 章)一文所阐明的那样，在计算机同心灵哲学的关系上，没有人持更不妥协的态度了：心灵**是**一个计算系统，大脑**事实上是**在执行计算的职能(计算对智能来说是**充分的**)，它与可能出现在计算机中的计算是**完全等同的**。人类智能可通过一组组控制着行为和(被逻辑上相似的行为主义心理学家所忽视的)内部信息处理的输入输出规则得到解释。由于计算机具备了正确的因果能力，它们也可以成为智能的：一台计算机——像一个大脑一样——是一个物理符号系统，而“一个物理符号系统具有对于一般智能行为来说，是必要的和充分的手段”。

纽厄尔和西蒙方法的核心，正如他们的反对者塞尔(第 3 章)所指出的，是与语义学的因果说相联系的符号论的形式句法理论。照他们看来，一个符号或计算成立的判据，是纯形式的，它的意义要借助于其因果的历史和作用来建立。一个符号就是一个物理**模式**，并以物理方式通过各种途径(如并列)同另一些模式发生联系，以构成复合“表达

式”。（由物理手段实现的）计算过程可对**模式**进行比较和修正：一个表达式作为输入，另一个作为输出。任何能够以物理方式存储并系统地变换表达模式的基底，都能行使符号的作用，但是这个基底与心理学目的无关。为达到对智能的理解，我们必须借助于指称和解释在信息处理层次上对物理符号系统加以描述。这两个语义学概念是从因果角度定义的，一个符号的意义就是这符号使系统产生的一组变化，或者达到或者响应某种（内部或外部）状态。因果相关本质上是任意的，就是说任何（非复合的）符号完全可以指称任何事物。〔这种限制不包括类比表述，因为在类比表述中，表述和被表述事物之间存在着不容忽视的相似性（Boden 1988：29－44）。〕

符号系统的这一定义可被批评为过分地物理主义，甚至那些在 AI 可行性方面与纽厄尔和西蒙有着共同信念的人，也有这种看法（Sloman 1986，待出版）。西蒙等人除了提出物质的例示对符号来说是**必不可少的**这样一个未加论证的假定（因此就排除了纯思维式的智能），他们的定义谈及的只是实际的机器，而不是虚拟的机器。所谓虚拟机，是指可被编程者**看作**正在使用的机器。它被抽象地定义为由有关系统执行的一组基本的信息加工作业。在虚拟机中，符号是抽象实体，而不是物理实体。一个计算系统之中可能存在若干个虚拟机——就像在高级编程语言由较低级语言实现的过程中，它依次被编译为汇编语言，然后再转换成机器码。心灵可能是由许多台这样的抽象符号机组成的，其中只有最基础的部分可通过脑组织以物质形式例示说明（这与某种层次较低的系统中的实际情况正相反）。然而，纽厄尔和西蒙的提法可以进行修改，以适应这一批评，批评者的观点并非认为 AI 不能实现，而是认为这比起他们两位定义中提出的那种文字形式要复杂得多。

AI 可以帮助我们认识心灵——它是什么，它是怎样工作的——这一点是所有计算心理学家都认可的，他们之中有些人同意纽厄尔和西蒙的看法：AI **就是**理论心理学（Longuet-Higgins 1987）。然而，也有些人对 AI 提出尖锐批评：D·玛尔在“人工智能之我见”（第 6 章）中指责