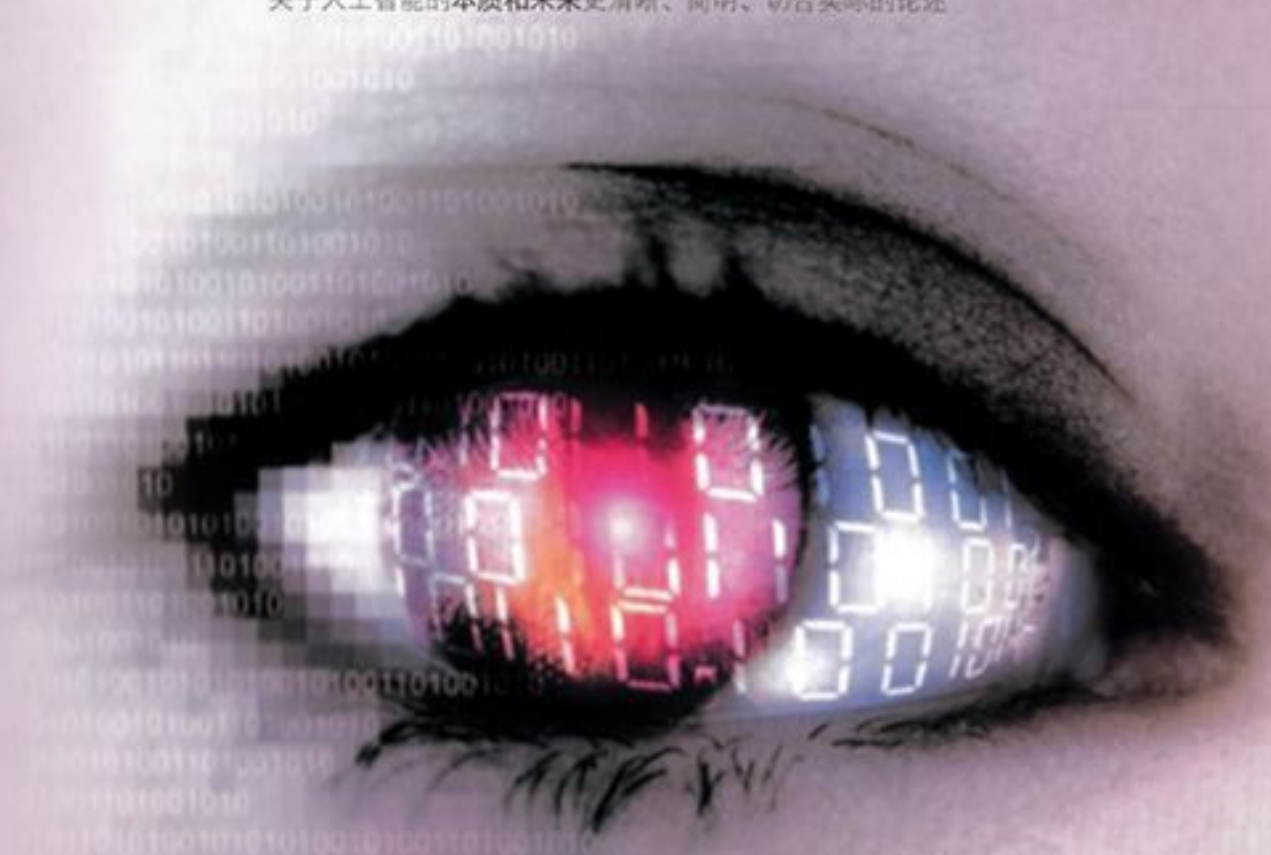


一部人工智能进化史

◆  
集人工智能领域顶级大牛、思维与机器研究领域  
最杰出的哲学家多年研究之大成

◆  
关于人工智能的本质和未来更清晰、简明、切合实际的论述




AI: Its Nature and Future

AI

人工智能的本质与未来

【英】玛格丽特·博登 (Margaret A. Boden) / 著 孙诗惠 / 译

 中国人民大学出版社

## 版权信息

书名：AI：人工智能的本质与未来

作者：【英】玛格丽特·博登

译者：孙诗惠

出版社：中国人民大学出版社

出版日期：2017-06-01

ISBN：978-7-300-24430-3

价格：55.00元

# 目录

## CONTENTS

---

- 01 什么是人工智能**
  - 虚拟机
  - 人工智能的主要类型
  - 人工智能的预言
  - 人工智能的起源
  - 控制论
  - 计算机建模者们分道扬镳
- 02 强人工智能：人工智能领域的圣杯**
  - 只有超级计算机还远远不够
  - 启发式搜索
  - 人工智能领域中的规划
  - 数学简化
  - 知识表示
  - 基于规则的程序
  - 框架、词向量、脚本、语义网络
  - 逻辑和语义网
  - 计算机视觉
  - 框架问题
  - 智能体和分布式认知
  - 机器学习
  - 通用系统
  - 梦想复兴
  - 缺失的方面
- 03 语言、创造力和情感**
  - 语言
  - 创造力
  - 人工智能与情感
- 04 人工神经网络**
  - 人工神经网络更广泛的含义
  - 分布式并行处理
  - 神经网络学习
  - 反向传播、大脑和深度学习

网络丑闻  
连接不是一切  
混合系统

## **05 机器人和人工生命**

情境机器人和有趣的昆虫  
进化人工智能  
自组织

## **06 强人工智能会有真正的智能吗**

图灵测试  
意识的很多问题  
机器意识  
人工智能和现象意识  
虚拟机和身心问题  
意义和理解力  
神经蛋白是必要条件吗  
不只是大脑，身体也很重要  
道德社区  
道德、自由和自我  
心智和生命  
巨大的哲学分歧

## **07 奇点**

奇点的预言家  
竞争的预测  
为怀疑论辩护  
全脑仿真  
我们应该担心什么  
我们为此做了些什么

译者后记

# 01 什么是人工智能

人工智能（Artificial Intelligence, AI）就是让计算机完成人类心智（mind）能做的各种事情。通常，我们会说有些行为（如推理）是“智能的”，而有些（如视觉）又不是。但是，这些行为都包含能让人类和动物实现目标的心理技能，比如知觉、联想、预测、规划和运动控制。

智能不是一维的，而是结构丰富、层次分明的空间，具备各种信息处理能力。于是，人工智能可以利用多种技术，完成多重任务。

人工智能无处不在。

人工智能的实际应用十分广泛，如家居、汽车（无人驾驶车）、办公室、银行、医院、天空……互联网，包括物联网（连接到小物件、衣服和环境中的快速增多的物理传感器）。地球以外的地方也有人工智能的影子：送至月球和火星的机器人；在太空轨道上运行的卫星。好莱坞动画片、电子游戏、卫星导航系统和谷歌的搜索引擎也都以人工智能技术为基础。金融家们预测股市波动以及各国政府用来指导制定公共医疗和交通决策的各项系统，也是基于人工智能技术的。还有手机上的应用程序、虚拟现实中的虚拟替身技术，以及为“陪护”机器人建立的各种“试水”情感模型。甚至美术馆也使用人工智能技术，如网页和计算机艺术展览。当然，它还有一些应用不那么让人欢欣鼓舞，如在战场上穿梭的军事无人机——但是，谢天谢地，它也用在机器人扫雷舰上。

人工智能有两大主要目标：一个是技术层面的，利用计算机完成有益的事情（有时候不用心智所使用的方法）；另一个是科学层面的，利用人工智能概念和模型，帮助回答有关人类和其他生物体的问题。大多数人工智能工作者只关注其中一个目标，但有些也同时关注两个目标。

人工智能不仅可以带来不计其数的技术小发明，还能够对生命科学产生深远的影响。某一科学理论的计算机模型可以检验该理论是否清晰连贯，还能生动形象地证明其含义（通常是未知的）。理论是否正确另当别论，但其依据是从相关科学范畴得出的证据。就算我们发现该理论是错误的，结果也能够给人以启迪。

值得一提的是，心理学家和神经学家利用人工智能提出了各种影响深远的心智—大脑理论，如“大脑的运作方式”和“这个大脑在做什么”的模型：它在回答什么样的计算（心理）问题，以及它能采用哪种信息处理形式来达到这一目标等。这两个问题不一样，但都十分重要。还有一些问题尚未回答，因为人工智能本身已经告诉我们：心智内容十分丰富，远远超出了心理学家们先前的猜想。

生物学家们也用到了人工智能——人工生命（A-Life）。利用这项技术，他们为生物体的不同内部结构建立了计算机模型，以解读不同种类的动物行为、身体的发育、生物进化和生命的本质。

人工智能对哲学也有影响。如今，很多哲学家对心智的解读也基于人工智能概念。例如，他们用人工智能技术来解决众所周知的身心问题、自由意志的难题和很多有关意识的谜题。然而，这些哲学思想都颇具争议。人工智能系统是否拥有“真正的”智能、创造力或生命，人们对此意见不一。

最后，人工智能向我们发出了挑战——如何看待人性，以及未来在何方。的确，有些人会担心我们是否真的有未来，因为他们预言人工智能将全面超过人的智能。虽然他们当中的某些人对这种预想充满了期待，但是大多数人还是会对此感到害怕。他们会问，如果这样，那还有什么地方能保留人类的尊严和责任？

我们将在接下来的几章逐一讨论上述问题。



## 虚拟机

谈到人工智能，人们可能会说：“那不就是指电脑嘛。”嗯，他们这么说既对也不对。电脑不是重点，重点是电脑做的事情。也就是说，虽然人工智能离不开物理机（如电脑），但是我们最好把它看作计算机科学家所说的虚拟机。

虚拟机和虚拟现实中所描述的机器不一样，和训练机修工时所使用的模拟汽车引擎也不一样，它是程序员在编程时和人们使用它所想到的信息处理系统。

让我们拿管弦乐队作类比。首先乐器是不能少的。要想让乐器演奏出美妙的音乐，那么木头、金属、皮革和弦线都必须遵循一定的物理定律。但观众在听音乐会时并不在意这一点，他们感兴趣的是音乐。他们也不在意单个音符，更不用说空气中发声的震动了。他们听的是音符产生的音乐“形状”：旋律与和声、主题与变奏、含混音与切分音。

当我们谈到人工智能时，情况也类似。用户使用设计师设计出来的文字处理器直接处理文字和段落。通常情况下，程序本身既不包含文字，也不包含段落（但有些段落也包含，比如用户可以很容易将版权标示插入到文字中）。神经网络（见第4章）也是并行处理信息，即使它通常是在约翰·冯·诺依曼（John von Neumann）结构计算机上（按顺序）实现的。

当然，这并不是说虚拟机只是杜撰或凭空想象出来的东西。虚拟机是真实存在的。我们不仅可以利用虚拟机完成系统内的任务（如果将其连接到照相机或机器人的手等这样的物理设备上），甚至还可以做好外部世界的工作。如果程序突发问题，人工智能工作者通常很少去找硬件方面的原因，而是对虚拟机或软件中的事件和因果关系更感兴趣。

编程语言也是虚拟机（它的指令只有翻译成机器码后才能运行）。有些指令用更低级的编程语言进行定义，所以多个层级的指令都需要翻译。否则，要是用机器码的位组合模式处理信息，大多数人将无法思考。如果信息处理过程过于复杂且层级划分过于细化的话，那么也没有人能正常思考。

虚拟机不只是编程语言。虚拟机一般包含各个层级的活动模式（信息处理）。虚拟机也不只是在电脑上运行的虚拟机。在第6章中，我们将看到“人类的心智”也可以被看作在大脑中实现的虚拟机，更确切地说，是并行运行（在不同时间发展和学习得到的）且交互的虚拟机集合。

要实现人工智能领域的进步，我们需要不断完善有趣实用的虚拟机的定义。不断改良物理机（更大、更快）确实有好处，它甚至可能是实现某种虚拟机的必要条件。但是，只有具备海量信息的虚拟机才能在这些物理机上运行，否则后者就算功能再强大也没用（同理，要在神经科学领域取得进步，我们需要清楚了解在神经元上实现什么“心理”虚拟机，详见第7章）。

各类外部世界的信息得到充分利用。所有人工智能系统都需要输入和输出设备，要是只需要一个键盘和一个屏幕就好了。它通常还需要专用传感器（可能是照相机或压敏晶须）或反应器（可能是供音乐或演讲用的声音合成器或机器人的手）。人工智能程序不仅处理内部信息，还与这些计算机的接口连接，或改变它们。

人工智能程序处理通常包含内部的输入和输出设备，供整个系统内部的虚拟机交互。例如，象棋程序的某一部分可能通过注意其他部分的情况来发现自己所面临的潜在威胁，这时候，它就有可能与那个部分配合，共同阻断本次威胁。



## 人工智能的主要类型

信息处理的方法取决于其所包含的虚拟机。我们将在后面的章节中看到，这主要有五种处理类型，每种处理类型又都包含很多变体。一种是经典逻辑或符号主义，有时称为有效的老式人工智能（Geod Old-Fashioned AI，以下简称GOFAI）；另一种是人工神经网络或联结主义。此外，还有进化编程、细胞自动机以及动力系统。

工作者通常只使用一种方法来处理信息，但也存在混合虚拟机。例如，在第4章中提到的一个在符号主义处理和联结主义处理之间不断切换的人类行为理论（这解释了为什么有的人在完成计划任务的过程中，会分心去关注环境中与之无关的东西以及这种现象是如何发生的）。第5章描述了一款集“情境”机器人学、神经网络和进化编程三者于一体的感觉运动装置（在装置的协助下，机器人将纸板三角形用作地标，找到了“回家”的路线）。

除了实际应用外，这些方法能够启发心智、行为和生活。神经网络有助于模拟大脑的内部结构以及进行模式识别和学习。经典逻辑人工智能（特别是与统计学结合时）可以模拟学习、规划和推理。进化编程阐明了生物进化和大脑发育。细胞自动机和动力系统可用来模拟生物体的发育。有些方法更接近于生物学，而不是心理学；有些方法更接近非条件反射行为，而不是慎重思考。要想全面了解心智，除了要用到上述所有方法外，还可能需要更多别的方法。

许多人工智能工作者并不关心心智的运作方式，他们只注重技术效率，而不追求科学理解。即使人工智能技术起源于心理学，但现在与心理学的联系却很少。然而，我们会发现，如果要想在强人工智能（artificial general intelligence）方面取得进步，我们需要加深理解心智的计算架构。

## 人工智能的预言

19世纪40年代，埃达·洛夫莱斯（Ada Lovelace）伯爵夫人预言了人工智能。更准确地说，她预言了部分人工智能。她专注于符号和逻辑，从未考虑过神经网络、进化编程和动力系统。她也未考虑过人工智能的心理目标，而纯粹对技术目标感兴趣。例如，她说一台机器“可能编写所有复杂程度或长度的细腻且系统的乐曲”，也可能表达“在科学史上具有划时代意义的、自然界的重要事实”（因此，如果当她看到以下情况时，她将不会感到吃惊：两百年以后，科学家们用“大数据”和精心制作的编程方法来推动遗传学、药理学、流行病学等无数领域知识的发展）。

她口中的机器是分析机（Analytical Engine）。这是一台齿轮连嵌齿轮的装置（从未被真正地制造出来），由其密友查尔斯·巴贝奇（Charles Babbage）于1834年设计。虽然这台机器主要用于求解代数和处理数字，但其本质相当于一台通用数字计算机。

她认识到了分析机的潜在通用性和处理符号（表示“宇宙中的所有主体”）的能力。她还描述了现代编程的各种基础知识：存储程序、分层嵌套的子程序、寻址、微程序设计、循环、条件、注释以及程序错误。她并没有谈到编曲或科学推理是如何在巴贝奇的机器上实现的。是的，人工智能可以实现，但是实现的方法当时仍然是一个谜团。

## 人工智能的起源

一个世纪以后，艾伦·图灵（Alan Turing）解开了这个谜团。1936年，图灵提出，每个合理计算在原则上都可以由现在被称为“通用图灵机”（Turing Machine）的数学系统来执行。图灵机是一个虚构系统，建立和修改用“0”和“1”表示的二进制符号组合。

第二次世界大战期间，图灵在布莱切利园（Bletchley Park）破解德国密码系统后，到20世纪40年代末一直在思考如何让一台物理机最接近抽象定义的图灵机（他帮助设计的第一台现代计算机于1948年在曼彻斯特完成），以及如何让这台物理机智能地执行任务。

与埃达·洛夫莱斯不同，图灵接受了人工智能的两个目标（技术和心理）。他想让新机器做通常需要智能才能完成的有意义的事情（可能通过使用非自然技术），并模拟以生理为基础的心智所发生的过程。

1950年，他那篇以幽默方式提出图灵测试（见第6章）的论文成为了人工智能的宣言〔第二次世界大战后不久，其论文得到进一步完善，但《官方保密法》（Official Secrets Act）阻止其出版〕。它抓住了智能信息处理（游戏、知觉、语言和学习）的症结，并暗示了当时计算机领域已经取得的成就，让人跃跃欲试（只有“暗示”，因为布莱切利园的工作仍然属于最高机密）。它甚至给出了算法，如神经网络和进化计算，不过在其论文发表很久以后，这些算法才得到广泛认可。要解开奥秘，这些都只是冰山一角，只是泛泛而谈——纲领性的东西，而不是程序。

图灵坚信，人工智能一定能以某种方式实现。20世纪40年代初，他的这一信念得到了神经病学家/精神病学家沃伦·麦卡洛克（Warren McCulloch）和数学家瓦尔特·皮茨（Walter Pitts）的支持。

他们的论文《神经活动中内在思想的逻辑演算》（A Logical Calculus of the Ideas Immanent in Nervous Activity）结合了图灵的观点与另外两项令人兴奋的成果（可追溯到20世纪早期）：伯特兰·罗素（Bertrand Russell）的命题逻辑和查尔斯·谢林顿（Charles Sherrington）的神经突触理论。

命题逻辑的关键点在于它是二进制的。每个句子（也称为命题）

假定为真或假。没有中间答案，也不接受不确定性或概率。只允许两个“真值”，即真和假。

此外，利用逻辑运算符（诸如and、or和if-then）构建了复杂命题，完成演绎论证，而逻辑运算符的意义由子命题的真/假来定义的。例如，如果两个（或更多）命题由“and”连接，则认为这两个（所有命题）都是真的。所以当且仅当“玛丽嫁给汤姆”和“弗洛西嫁给彼得”二者都是真命题，那么“玛丽嫁给汤姆和弗洛西嫁给彼得”才是真命题。事实上，如果弗洛西没有嫁给彼得，那么包含“and”的复杂命题就是假命题。

麦卡洛克和皮茨将罗素和谢林顿的观点结合，因为他们都描述了二进制系统。逻辑的真/假（true/false）值映射到图灵机中的脑细胞开/关（on/off）活动和个体状态0/1中。谢林顿认为，神经元不仅进行严格的开/关活动，而且具有固定阈值。因此，逻辑门（and、or和not）被定义为微小的神经网络，可以相互连接来表示高度复杂的命题。任何东西只要能用命题逻辑表述，那就能用某种神经网络和某种图灵机来计算。

简单来说，就是神经生理学、逻辑学和计算被放在一起研究。后来，心理学也被纳入进来一起讨论。麦卡洛克和皮茨相信（就像许多哲学家当时所说的），自然语言在本质上归结为逻辑。所以，从科学论证到精神分裂症错觉的所有推理和观点都可以放到他们的理论“磨坊”里加工。麦卡洛克和皮茨为整个心理学预言了一个时代，“（神经）网络的设计规格将对心理学领域取得的所有成果都有帮助”。

其核心含义在当时很清楚：同一个理论方法，即图灵计算，可用于人和机器智能，麦卡洛克和皮茨的文章甚至影响了计算机的设计。约翰·冯·诺依曼当时打算使用十进制代码，但他后来意识到了问题，改为二进制。

图灵当然赞同图灵计算，但他无法进一步推动人工智能的发展：当时技术太过原始。然而，到20世纪50年代中期，出现了功能更强大且更容易使用的机器。这里的“易于使用”并不是说更容易打开电脑的按钮，也不是说更容易将它在房间里推来推去，而是指定义新的虚拟机更加容易（例如，编程语言），从而有利于定义更高级的虚拟机（例如，用来做数学运算或规划的程序）。

大约本着图灵宣言的精神，符号人工智能的研究在大西洋两岸得以开始。20世纪50年代末期，有一个标志性事件上了新闻头条，即阿瑟·塞缪尔（Arthur Samuel）的跳棋（国际跳棋）程序打败了塞缪尔

本人。这无疑暗示着电脑有一天可能会具有超人的智力，超过设计它们的程序员的能力。

20世纪50年代末期，还出现了第二个这样的暗示，即逻辑理论机（Logic Theory Machine）不仅证明了罗素的18个关键逻辑定理，还发现了一个更有效的证明，来证明其中某一个定理。这的确令人印象深刻。塞缪尔只是一个平庸的跳棋选手，但是罗素可是一位世界级的逻辑学家〔罗素本人为这项成就感到十分高兴，但是《符号逻辑杂志》（Journal of Symbolic Logic）拒绝发表一篇计算机程序撰写并署名的论文，更为重要的是，它并没有证明一个新定理〕。

逻辑理论机很快就被一般问题解决器（General Problem Solution，以下简称GPS）“超越”——“超越”并不是说GPS可以“超越”更多卓越的天才，而是说它的应用范围不再限制在一个领域。顾名思义，GPS可解决用目标、子目标、动作和运算符表示的任何问题（详见第2章）。程序员一旦确定与任何特定领域相关的目标、动作和运算符，剩下的推理工作就可以由GPS负责完成。例如，GPS解决了“牧师和野人”的问题（三个牧师和三个野人在一条河的一边，现在有一艘船，一次最多可以载两个人。问题来了，如何在野人数量不超过牧师数量的情况下确保每个人都能过河）。这个问题对人类来说都不简单，因为每次把两个人运过去之后，都必须让其中一个回来，这样游戏才能继续下去（大家可用便士试一试）。

逻辑理论机和GPS都是GOFAI的早期示例。现在说它们是“老式的”，当然毫无疑问；但它们也是“有效的”，率先运用了“启发法”和“规划”——二者在今天的人工智能领域都至关重要（见第2章）。

并不是只有GOFAI这种人工智能受到论文《神经活动中内在思想的逻辑演算》的启发，联结主义也备受鼓舞。20世纪50年代，计算机上特制或仿真的麦卡洛克和皮茨逻辑神经网络被用来〔如艾伯特·厄特利（Albert Uttley）〕模拟联想学习和条件反射（这些神经网络进行集中式而非分布式处理，与今天的神经网络不同，详见第4章）。

早期网络模拟并不完全由神经—逻辑统治。雷蒙德·拜沃勒（Raymond Beurle）在20世纪50年代中期实现的系统（在模拟计算机中）大不一样。他的研究工作没有始于精心设计的逻辑门网络，而是始于随机连接的和不同阈值的单元的二维数组。他认为神经自组织的发生是因为动力波的激活——构建、传播、坚持、死亡和时不时的相互作用。

拜沃勒已经意识到，诡辩机可以模拟心理过程，但不等于大脑实际上就是这样的机器。麦卡洛克和皮茨已经指出了这一点，他们发表了一篇具有开创性意义的论文，短短四年后，又在其另外一篇论文中指出了热力学比逻辑更接近大脑的功能。逻辑学被统计学取代，单一单元被集合取代，确定性纯度被概率噪音取代。

换句话说，他们已经描述了我们现在所说的分布式容错算法（见第4章），并认为这种新算法是之前算法的“延伸”，彼此并不矛盾。它在生物学上更现实。



## 控制论

麦卡洛克比GOF AI和联结主义对早期人工智能的影响更为深远。20世纪40年代，在其神经学和逻辑学研究成果的指引下，处于萌芽期的控制论运动得到蓬勃发展。

当时，控制论者的研究重心是生物自组织。它涵盖了各种适应和新陈代谢，包括自主思考、微观运动行为和（神经）生理调节。其核心思想是“双向循环性”或反馈。关键问题是目的论或目的性。对于反馈取决于目标的差异性而言，这些概念高度相关——目标现阶段的偏差被用于指导下一步动作。

1948年，诺伯特·维纳（Norbert Wiener，其在战争期间设计了反弹道导弹）对该运动进行了命名，将其定义为“关于在动物和机器中控制和通信的研究”。那些控制论者在建立计算机模型的时候，经常从控制工程学和模拟计算机中获取灵感，而不是从逻辑学和数字计算中获取。然而，这种区分并不是十分明确。例如，目标的差异性被用于控制制导导弹，以及解决符号问题。此外，图灵作为经典人工智能的冠军，用动力学方程（描述化学扩散）定义自组织系统。在这些系统中，诸如点或片段的新结构可以从一堆同质的低级个体中产生（见第5章）。

早期参与该运动的成员还包括：经验心理学家肯尼思·克雷克（Kenneth Craik）、数学家约翰·冯·诺伊曼、神经学家威廉·格雷·沃尔特（William Grey Walter）和威廉·罗斯·艾什比（William Ross Ashby）、工程师奥利弗·塞尔弗里奇（Oliver Selfridge）、精神病学家和人类学家格雷戈里·贝特森（Gregory Bateson）以及化学和心理学家戈登·帕斯克（Gordon Pask）。

克雷克（于1943年在一次自行车事故中逝世，享年31岁）在研究神经系统的过程中参考了模拟计算，当时还没出现数字计算机。他根据大脑中“模型”的反馈，大致描述了知觉、微观运动行为和智力。他的大脑模型或表示概念后来在人工智能中产生了巨大的影响。

整个20世纪30年代，约翰·冯·诺依曼都对自组织心存疑惑，同时又因麦卡洛克和皮茨的第一篇论文感到异常兴奋。他不仅将基本的计算机设计从十进制改为二进制，还完善了麦卡洛克和皮茨的观点，以解释生物进化和繁殖。他定义了各种细胞自动机，即由很多基本计算



单元组成的各种复杂系统。计算单元的变化遵循简单规则，而这些规则又取决于相邻单元的当前状态。其中一些单元可以复制其他单元。他甚至定义了一个能够复制任何东西的通用复制器——包括复制它自己。他指出，复制错误可能导致进化。

约翰·冯·诺依曼用抽象的信息术语对细胞自动机作了详细说明。但是这些细胞自动机可以用多种方式具现，例如自组装机机器人、图灵的化学扩散、拜沃勒的物理波或很快将揭开神秘面纱的DNA。

从20世纪40年代末开始，艾什比制作了同态调节器（Homeostat），它是一个生理性自体调解的电化学模型。它可以维持机体内环境的总体恒定，无论最初分配给它的100个参数值是多少（允许设定近400000种不同的起始条件）。它阐释了艾什比的动态自适应理论在试错法学习和自适应行为中的应用，这种动态自适应可以发生在身体内部（尤其是大脑），也可以发生在身体与外部环境之间的环境。

格雷·沃尔特也在研究自适应行为，但他用的是一种迥然不同的研究方式。他研发了一款类似乌龟的微型机器人，其感觉运动电路模拟了谢林顿的神经反射理论。这些情境机器人先驱的行为栩栩如生，如寻找光线、避开障碍，以及利用有条件的反射进行联想学习。这些有意思的机器人于1951年在“英国节”（Festival of Britain）上向公众展示。

十年后，塞尔弗里奇（伦敦百货商店创始人的孙子）利用符号方法实现了一种叫伏魔殿（Pandemonium）的并行处理系统。

这个GOF AI程序利用许多底层“守护程序”（特征感知器）来学习如何识别模式，每个“守护程序”一直都在感知外界信息，并将感知到的结果传递给更高级的“守护程序”。这些“守护程序”重点关注到目前为止一致的特征（例如，一个F中只有两根水平条），而忽略了任何不合适的特征。置信度可以有差异，而且它们至关重要：声音最洪亮的守护程序影响最大。最后，最高级的守护程序根据既得证据（通常是冲突的），选择最佳模式。这项研究很快对联结主义和符号人工智能产生了影响，一个最近的分支是学习智能分布实体（Learning Intelligent Distribution Agent，以下简称LIDA）的意识模型（详见第6章）。

贝特森对机器没有什么兴趣。他于20世纪60年代提出了与文化、酗酒和（父或母对子女的）“双重约束”精神分裂症有关的理论。但是，这些理论基础却是在早些时候控制论会议上提出的和通信（即反

馈)相关的想法。从20世纪50年代中期开始,帕斯克——麦卡洛克口中的“自组织系统天才”,在许多项目中都用到控制论和符号思想,其中包括:交互式剧院、互通音乐机器人、获悉并适应其用户目标的架构、化学自组织概念和教学机。借助帕斯克的研究,人们能够利用复杂的知识表示来采取不同方法,而这对认知方式为循序渐进型和整体型(以及对不相关事物不同程度的容忍)的学习者都适用。

简言之,到20世纪60年代后期,研究人员考虑了所有主要的人工智能类型,甚至将其实现——有的甚至更早。

大多数相关研究人员至今还广受人们的尊重,但只有图灵一直是人工智能盛宴上的“幽灵”,其影响无处不在。多年来,其他人只被一些研究领域的分支机构所记住。特别是,格雷·沃尔特和艾什比几乎被人们遗忘,直到20世纪80年代后期,他们才被赞誉为“人工生命之父”(与图灵一起)。帕斯克等待的时间更长。要知道其中的原因,我们必须了解计算机建模者们是如何分道扬镳的。

## 计算机建模者们分道扬镳

20世纪60年代之前，模拟语言/逻辑思维和模拟有目的的/自适应的微观运动行为这两个研究方向是有交叉的。有些专家二者都研究[唐纳德·麦凯（Donald Mackay）甚至建议制造将神经网络与符号处理结合起来的混合计算机]。所有相关工作者都能彼此产生共鸣。研究生理自动调整的工作者们认为自己与注重心理研究的同事们做的是同一件事。他们都参加相同的会议：在美国召开的跨学科Macy研讨会（1946年到1951年，由麦卡洛克担任主席）和在伦敦召开的“思维过程机械化”研讨会（1958年，由厄特利组织）。

然而，大约从1960年开始，工作者们的研究方向便出现了分歧。广义上来说，对生命感兴趣的人只关注控制论，而那些对心智感兴趣的人则关注符号计算。网络爱好者们当然对大脑和心智都感兴趣，但他们通常研究联想学习，而非具体的语义内容或推理，所以他们关注的是控制论而不是符号人工智能。研究的分支越来越多，不幸的是，各分支之间缺乏对彼此应有的尊重。

这个过程中必然会出现一些优秀的小社会圈子，因为他们讨论的理论问题各不相同，既有生理方面的，也有心理方面的。所用的技术也不一样。广义上讲，是微分方程与逻辑之间的较量。专门化趋势不断加强，交流也因此变得越来越困难，而且很大程度上是无利可图的。兼收并蓄的会议已经过时。

即使如此，各分支学派也不应该太操之过急。对控制论和联结主义学派的反感源于专业上的嫉妒和正义的愤慨。这是因为符号计算在发展初期取得了巨大的成功；带有挑衅性的术语“人工智能”[由约翰·麦卡锡（John McCarthy）于1956年提出，以前称为“计算机模拟”]博得了新闻工作者们的眼球；一些不现实的炒作；一些符号主义研究者表现得傲慢自大。

符号主义阵营的成员们认为自己赢得了人工智能的比赛，最初没有太多的敌意。事实上，他们在很大程度上忽视了早期的网络研究，其中的一些领导者[例如，马文·明斯基（Marvin Minsky）]已经开始着手网络的研究。

然而，在1958年，弗兰克·罗森布拉特（Frank Rosenblatt）提出了神经动力学理论，定义了能在随机初始状态下（并且能够容忍初始

化阶段的错误）进行自组织学习的并行处理系统，并在他的光电感知器中部分实现了该理论。它与“伏魔殿”不同，无须输入模式让程序员提前分析。符号主义学派无法忽视这种新形式的联结主义。但它很快就被打入“冷宫”。20世纪60年代，明斯基与西摩尔·帕普特（Seymour Papert）一道发表了一篇尖锐的批评文章，声称感知器连一些基本的东西都无法计算（详见第4章）。

神经网络研究的资金来源也因此被切断。这个结果是由于两派攻击者蓄意为之，从而加深了人工智能内部的对抗。

现在大家看来，经典人工智能研究似乎在当时占绝对的主导地位。诚然，格雷·沃尔特的机器乌龟们在英国节上备受赞誉。和伯纳德·威德罗（Bernard Widrow）的模式学习Adaline<sup>[1]</sup>（基于信号处理）一样，罗森布拉特的感知器在20世纪50年代后期也被媒体大肆宣传。但符号主义研究者的批评让人们完全失去了对感知器的兴趣。20世纪60年代到70年代，在媒体中如日中天的是符号型人工智能（还影响了精神哲学）。

风水轮流转。神经网络，如分布式并行处理（Parallel Distributed Processing，以下简称PDP）于1986年再次登台（见第4章）。本该更懂得此方法的大多数外界人士和一些内部人士都把它当成了一个彻头彻尾的“新”东西。它还吸引了无数研究生和很多新闻媒体（和哲学）的关注。那时，鼻子都被气歪的人恐怕是那些符号人工智能的研究者了。一时间，PDP研究成为时尚，大家普遍认为经典人工智能的研究当时已经失败。

还有一些控制论者因其在1987年命名“人工生命”，终于从大批记者和研究生那里“受宠”。于是，符号人工智能再次受到挑战。

然而，在21世纪，不同的问题需要不同类型的答案——各有所长，这一点显而易见。虽然先前的敌意至今犹存，但不同方法仍有相互尊重和合作的空间。例如，“深度学习”有时用于将符号逻辑与多层概率网络结合的强大系统；还有一些混合方法包含高级复杂的意识模型（见第6章）。

构成人类心智的虚拟机本来就是各式各样的，因此大家没必要对人工智能领域的研究分歧太过惊讶。

---

注释

[1]Adaline是一个早期的单层人工神经网络和实现这个网络的物理设备的名称。网络使用存储电阻器。由伯纳德·威德罗教授及其在斯坦福大学的研究生泰德·霍夫（Ted Hoff）于1960年联合开发。它基于麦卡洛克—皮茨的神经元，由权重、偏差和求和函数组成。——译者注

## 02 强人工智能：人工智能领域的圣杯

最先进的人工智能是一种神奇美妙的东西，可以提供各式虚拟机，以进行各种不同类型的信息处理。它没有核心秘密，也没有统一的核心技术。人工智能工作者来自各个领域，几乎没有统一的目标和方法。本书只能涵盖最近取得的极少一部分成就。总之，人工智能的方法范围极其宽泛。

可以说，人工智能已经取得了惊人的成就。它的实际运用范围也十分广泛。我们现在有大量针对无数特定任务而设计的人工智能应用程序。生活中各个领域的专业人士和非专业人士几乎都在使用。很多程序甚至比一些专家还牛。由此看来，人工智能的进展的确引人注目。

但是，人工智能先驱们的目标不仅限于专家系统，他们还希望发展通用智能系统。他们模拟的所有人类能力——视觉、推理、语言、学习等——可应对各类挑战。此外，这些能力将适时得到整合。

从这些标准来看，进步的空间非常大！约翰·麦卡锡很早就认识到人工智能需要“共识”。作为1971年和1987年图灵奖的获得者，他在发表获奖感言的时候谈到了人工智能的通用性（Generality in Artificial Intelligence），但他其实是在抱怨，而不是庆祝。到2016年，他的抱怨仍未得到答案。

随着近来计算机能力的不断增强，强人工智能在21世纪再次引起人们的兴趣。如果这一目标得以实现，人工智能系统将减少对专用编程技巧的依赖，而受益于推理和知觉这些通用功能——语言、创造力和情感（所有这些我们都将在第3章中讨论）。

然而，这谈何容易。通用智能仍然是一个严峻的挑战，让人难以

捉摸。强人工智能无疑是人工智能领域的圣杯。



## 只有超级计算机还远远不够

对于任何想要实现这个梦想的人而言，超级计算机必然是一个助推器。组合爆炸——其中需要的计算超过实际能执行的计算——已不再构成威胁。然而，我们不能一直靠增强计算能力来解决问题。

通常，我们还需要新的解决方法。此外，即使某一具体方法从理论上讲是可行的，但也可能需要大量的时间或存储，才能在实践中发挥作用。第4章给出了相关例子（有关神经网络）。同样，穷举法列出了所有可能的象棋步骤，但它需要的存储位置比宇宙中的电子还要多，因此就算有一大堆超级计算机也不能满足需要。

另外，效率也很重要：计算数量越少越好。总之，问题必须易于处理。

如今，已经有几个基本策略可以做到这一点。它们由经典符号人工智能或GOF AI开创，在今天仍然必不可少：第一，只关注一部分搜索空间（问题的计算机表示，解决方案假定就在该表示中）；第二，简化假设以构建较小的搜索空间；第三，有效安排搜索过程；第四，用新方式表示问题，以构建不同的搜索空间。

这些方法分别对应的是启发法、规划、数学简化和知识表示。接下来的五个部分将分别讨论这些强人工智能策略。

## 启发式搜索

“启发式的”（heuristic）和“找到了”（Eureka）在英文中有相同的词根：来自意为“寻找”或“发现”的希腊语。启发法得到了早期GOFAI学派的重视，并且经常被看作“编程技巧”。但是，这个术语并非源于编程：逻辑学家和数学家对其早已熟悉。早在数千年前，人类就用启发法解决问题（有意或无意地），远远早于埃达·洛夫莱斯伯爵夫人（Ada Love Lace's）预见人工智能的时间。

无论是对人还是对机器来说，启发法都有利于问题的解决。强人工智能使用启发法的模式是：让程序只针对搜索空间的某些部分，同时避开其他部分。

很多启发式算法都属于无法保证成功的经验法则，如早期人工智能使用的大多数经验法则。在启发法引导下，系统正好忽略了某部分的搜索空间，而解决方案可能正好位于这部分空间里。例如，在国际象棋中，“保护女王”是一条非常有用的规则，但偶尔也应该违背。

还有一些启发式算法从逻辑学角度或数学角度被证明是合理的。如今，人工智能和计算机科学领域的大量工作都是为了确定程序可证明的属性。这是“友好人工智能”的一个方面，因为人类安全可能由于使用从逻辑学角度看不是很可靠的系统而受到的威胁（详见第7章，启发式算法和算法之间原则上没有区别，许多算法实际上是包含多个特定启发式算法的微型程序）。

无论启发法可靠与否，它对人工智能搜索来说都不可或缺。上文提到人工智能越来越专业化，这部分取决于能显著提高效率的新启发法的定义，但仅限于限制颇多的某类问题或搜索空间。一个非常成功的启发法可能并不适合让其他人工智能程序“借用”。

如果我们给定几种启发法，那么应用它们的顺序就可能变得很重要。例如，即使这种排序偶尔会导致灾难，也应该先考虑“保护女王”，再考虑“保护象”。不同的顺序将定义不同的搜索树遍历整个搜索空间。给启发法下定义和排序是现代人工智能的关键任务（启发法在认知心理学中的地位也很突出，例如，“快速节俭启发式”指出进化如何让人获得对环境作出有效回应的方法）。

利用启发法，我们不再需要穷举搜索整个搜索空间，但它们有时

会和（有限的）穷举搜索结合使用。1997年，因击败世界冠军加里·卡斯帕罗夫（Gary Kasparov）而名声大噪的IBM国际象棋程序深蓝（Deep Blue）使用的是专用硬件芯片，每秒能处理2亿个位置，可以知道接下来8步的所有备选棋步。但是，它不得不用启发法来选择备选棋步中的“最佳”棋步。由于启发法可信度不高，所以即使是深蓝，也不能每次都胜出。

## 人工智能领域中的规划

在今天的人工智能领域中，规划的地位十分突出，尤其是在很多军事活动中。事实上，直到最近才在人工智能上支付大部分研究费用的美国国防部指出，在第一次伊拉克战争的战场上，他们在后勤保障方面省下的钱（利用人工智能规划）超出了其前期投入。

规划不仅限于人工智能：我们都在规划。请想想假期打包东西的情景。你必须找到所有要带的东西，但这些东西可能没有放在同一个地方。你可能还要买一些新东西（比如防晒霜）。你必须决定是把所有东西放一块儿（也许放在床上，也许放在桌子上），还是把每样东西都放在皮箱里。这个决定可能部分取决于你是否决定最后再放衣服，因为你怕把衣服弄皱。你是需要一个背包，还是要一个手提箱，或者两者都要，你如何取舍？

把规划用作人工智能技术的GOF AI程序员们考虑到了有意识的深思熟虑（基于神经网络的人工智能大不一样，因为它不模拟有意识的深思熟虑，详见第4章）。因为负责逻辑理论机（见第1章）和GPS的先驱们主要对人类推理心理学感兴趣。他们的程序基于被试主体为人的实验，要求被试人进行“有声思维”：一边做逻辑谜题，一边描述自己的思考过程。

现代人工智能规划程序并不那么依赖从有意识的内省中或实验观察中获得的想法。它们的规划比早期程序的规划要复杂得多，但基本思想一致。

一个规划列举出一系列在常规层级上表示的动作——一个最终目标、加上很多子目标和次级子目标……这样就不用一次性考虑所有的细节。在某一适当的抽象层级规划上可以修剪搜索空间内的搜索树，因此一些细节根本不需要考虑。有时，最终目标本身就是一个动作规划——可能是调度送向/运出工厂或战场的货物。有时候，它还是一个问题的答案，比如医疗诊断。

针对任何给定目标和预期情况，规划程序需要一个动作列表（即符号运算符），或多种动作类型（通过填写来源于问题的参数，可以将动作实例化），每个动作都能够作出一些相应的变化；针对每个动作，规划程序需要一组必要的前提条件（比较抓住某个东西，前提条件是这个东西必须是在手的可活动范围内）；针对所需变化的优先次

序和对动作的排序，规划程序需要启发法。如果程序要选定某个特定动作，它可能要设置一个新的子目标以满足前提条件。这个目标制定的过程可以一再重复。

通过规划，程序——或人类用户——可以发现已经做了什么动作，以及为什么这么做。“为什么”指的是目标的层次结构：做这个动作是为了满足那个前提条件，以实现这个子目标。人工智能系统通常采用“正向链接（推理）”和“反向链接（推理）”技术，来解释程序是如何找到解决方案的。这可以帮助用户判断程序的动作或建议是否恰当。

当前，一些规划程序拥有数万行代码，它们在无数层级上定义分层的搜索空间。这些系统通常与早期的规划程序大相径庭。

例如，大多数规划程序假定，并非所有的子目标都可以被独立处理（即问题可完全分解）。毕竟在现实生活中，某一个以目标为导向的行动所产生的结果都可能被另一个行动撤销。今天的系统能处理可部分分解的问题：它们独立处理子目标，但必要的时候，还可以做其他处理，以整合随之产生的各项子规划。

经典系统只能解决那些在完全可观察的、可确定的和有限的静态环境中出现的问题。但是，一些现代规划程序可以处理在部分可观察（即系统的“世界”模型可能不完整或不正确）和不确定的环境中出现的问题。在这些情况下，系统必须监控执行期间的变化情况，以在规划中，或在自己对于“世界”的“信念”中，作出适当改变。一些现代规划程序可以在很长一段时期内一直进行监控：它们可以根据环境变化，不断制定、执行、调整和舍弃目标。

其他许多开发领域已经取得了新进展，并且还在不断进步。所以，在20世纪80年代，一些机器人专家完全否决规划并转而青睐于“情境”机器人学（见第5章）的情况，着实让人惊讶，例如，目标和可能动作的内部表示等概念也被否决了。然而，在很大程度上，这种批评是不对的。批评者们自己的系统可能没有表示目标，但可能表示了其他东西，例如视网膜刺激和奖励。此外，即使是最先遭受此类批评的机器人技术，也常常需要规划以及纯粹的被动回应，例如，制造会踢足球的机器人。

## 数学简化

鉴于启发法任由搜索空间自由发展（这样程序就可以专注于搜索空间里最合适的部分），简化假设构造了一个不切实际但更易于计算的搜索空间。

一些假设和数学相关，如机器学习中通常使用的“i.i.d.”（独立同分布）假设。与数据中实际的概率构成相比，用i.i.d.表示的概率构成要简单得多。

数学简化在定义搜索空间时的优点是可以利用搜索的数学方法。这种方法定义清晰，而且至少方便数学家理解。这并不代表任何数学定义的搜索都有实际使用价值。如上所述，某一方法在数学层面可以保证解决某一特定类别的所有问题，但是该方法在实际运用中并不适用，因为其时间成本可能是无限的。然而，它可能建议与此类似但更实际的方法，请见第4章中有关“反向传播”的讨论。

在人工智能领域，非数学的简化假设比比皆是，并且通常很直接。一种是假定无须考虑情感因素（见第3章）就能定义和解决问题的（默认）假设。还有很多假设被构建在用来指定任务的一般知识表示中。

## 知识表示

一般来说，实现人工智能的最难部分是最开始如何向系统表示问题。即使人类看似可以直接与程序交流（用英语对着Siri说话或者在谷歌的搜索引擎中输入法语单词），但事实并非如此。人们无论是处理文本还是图像，都必须将包含的信息（“知识”）以机器可以理解的方式表示给系统，换句话说，以系统可以处理的方式表示（机器是否“真”的理解，我们将在第6章中进行讨论）。

人工智能表示问题的方式五花八门。有些是对GOFAI中知识表示的一般方法进行演绎或改动。越来越多的是针对一小类问题制定的一些高度专业化的方法。例如，我们可能为人类某种癌细胞的X射线图像或照片精心设计一种新的表示方法，从而可以得到某个非常具体的医学解释方法（因此，这种新方法不能用来识别猫或CAT扫描）。

欲实现强人工智能，通用方法是关键。这些方法最初受到人类认知心理学研究的启发，包括：IF—THEN规则集；个体概念的表示；模式化的动作序列；语义网络；以及利用逻辑或概率进行推理。

接下来，我们对上述方法依次展开讨论（第4章描述了另一种知识表示形式，即神经网络）。



## 基于规则的程序

在基于规则的编程中，大量知识/信念被表示为一套将条件与动作联系起来的“如果—则”（IF—THEN）规则集：如果满足这个条件，则进行那个动作。这种形式的知识表示利用了形式逻辑[埃米尔·珀斯特（Emil Post）的“产生式”系统]。但是人工智能先驱艾伦·纽厄尔（Allen Newell）和赫伯特·西蒙（Herbert Simon，又名司马贺）通常认为它是人类心理学的基础。

条件和动作都可能很复杂，规定的内容可能是几个或多个命题的合取（或析取）。如果同时满足几个条件，则包含最多命题的合取被赋予优先级。所以，“如果目标是制作烤牛肉和约克郡布丁”将优先于“如果目标是制作烤牛肉”，而在条件中增加“和三种蔬菜”又优先于“如果目标是制作烤牛肉和约克郡布丁”。

基于规则的程序不提前规定每一步的顺序。相反，每条规则都在等待被其条件触发。尽管如此，这类系统可以用来做规划。如果不能做规划，那么它们在人工智能方面就只能发挥有限的作用。但是它们的规划方式不同于最古老、最为人们熟悉且最常用的编程形式（有时称为“执行控制”）。

在能进行执行控制的程序（如GPS和逻辑理论机，详见第1章）中，规划被明确表示。程序员按照严格的时间顺序，规定一个寻找目标的指令序列，指出哪一步该执行哪一条指令：“做这个，然后去做那个；然后看看X是否为真；如果是真，就做这个事；如果不是，就做那个事。”

“这个事”或“那个事”有时是一条设置某个目标或子目标的明确指令。例如，机器人如果有离开房间的目标，那么可能指示该机器人设置开门的子目标（原文如此）；接下来，如果检查门当前状态的结果显示门将被关闭，则设置抓握门把手的子次目标（人类蹒跚学步者可能需要更低级的子次目标——即让成年人抓住自己够不着的门把手，并且如果婴儿要做到这一点，可能需要在更低级别设定几个目标）。

基于规则的程序也可以用来解决如何逃离房间的问题。然而，规划层级不会被表示为按时间顺序排列的明确步骤，而是表示为构成系统的“IF—THEN”规则集合中所隐含的逻辑结构。某一条件可能要求已经建立了这样一个目标（IF你想打开门，而且你不够高）。同样，

动作可以包括设置一个新目标或子目标（THEN 找一个成人）。更低级的目标将自动激活（IF 你想要求一些人做一些事，THEN 设置接近他们的目标）。

当然，程序员必须列入相关的IF—THEN规则（上述案例中指的是涉及门和门把手的规则）。但是，他们不需要预期这些规则的所有潜在逻辑含义（这是一把“双刃剑”，因为潜在的不一致可能在很长一段时间都无法被发现）。

被激活的目标/子目标被贴在中央“黑板”上，可供整个系统访问。显示在黑板上的信息不仅包括被激活的目标，还包括感知输入和当前处理的其他方面（该想法不仅影响了一个意识神经心理学的前沿理论，还影响了以它为基础的意识人工智能模型，详见第6章）。

基于规则的程序广泛应用于20世纪70年代早期出现的先驱“专家系统”。这些系统包括：MYCIN系统——在人类医生鉴定感染性疾病和开抗生素药物时提建议；还有树枝状演算法（DENDRAL）——对有机化学中某一特定范围内的分子进行光谱分析。例如，做医疗分析的计算机咨询专家系统MYCIN，它的诊断方法是将症状/病人本身的身体状况（条件）与诊断结论/建议相匹配，以便继续检测或开处方（动作）。这些程序是人工智能远离“从一般化走向专门化之梦”的第一步，同时为实现埃达·洛夫莱斯梦想迈出了第一步——机器制造科学之梦（见第1章）。

由于基于规则的知识表示，程序能够被逐步建立，因为程序员或者强人工智能系统本身可增加对域的了解。新规则可以随时添加。没有必要从头重新编写程序。但是有一个棘手的问题，即如果新规则与旧规则逻辑冲突，系统将不会总是做它应该做的事，甚至可能和它应该做的事相去甚远。在处理一小组规则时，这些逻辑冲突很容易被避免，但是如果系统较大，它们就很难被识破。

20世纪70年代，新IF—THEN规则在与人类专家不断对话的过程中得到，它们被要求解释自己的决定。今天，尽管许多规则不是来自有意识的内省，但这些规则更高效。现代专家系统（今天很少使用的术语）应用范围广，从大型科学研究和商业程序到手机上小的应用程序。由于受益于其他形式的知识表示，许多系统超过旧系统，如统计和专用视觉识别或大数据的使用（见第4章）。

在某些狭窄领域中，这些程序可以帮助甚至取代人类专家。有些超越了上述领域的世界翘楚。近四十年前，在诊断大豆疾病的时候，一个基于规则的系统比最权威的专家还准确。如今，用它来帮助科

学、医学、法律甚至服装设计领域专业人士的例子不胜枚举（这不完全是好事，详见第7章）。

## 框架、词向量、脚本、语义网络

其他常用的知识表示方法包含个体概念，而不是整个领域（如医学诊断或服装设计）。

例如，可以通过规定分层数据结构（有时称为“框架”）告诉计算机什么是房间。它将一间房表示为有地板、天花板、墙壁、门、窗户和家具（床、浴缸、餐桌……）。真实的房间具有不同数量的墙壁和门窗，因此可在框架中的“插槽”里填充特定数字，并提供缺省赋值（四道墙、一扇门和一扇窗）。

计算机可以使用这类数据结构找到相似类、回答问题、参与对话、创作或理解故事。它们是CYC<sup>[1]</sup>（encyclopedia，即百科全书）的基础：一个试图表示所有人类知识的大胆尝试。有人甚至说这个想法是痴人说梦。

然而，框架也可能造成误导。例如，缺省赋值就有诸多问题（有些房间没有窗户，开放式的房间没有门）。更糟糕的情况是：该如何表示下落或溢出这样的日常概念？符号人工智能这样表示“朴素物理学”的常识性知识：构造对事实进行编码的框架，如未支撑的物体会下落，但也有例外——氦气球就不会下落。考虑清楚这类情况是一项永无止境的任务。

在一些利用最新技术处理大数据的应用中，单个概念可能被表示为一个簇或“云”，由成百上千个偶尔相关的概念组成（概念对之间的相关性概率各不相同，详见第3章）。类似地，概念现在可以用“词向量”而不是单词来表示。此处的语义特征生成许多不同概念并连接各个概念，由（深度学习）系统发现，可用来预测接下来的词——例如，在机器翻译中的运用。然而，这些表示用在推理或谈话中的时候，不像经典框架那么经得起检验。

有些数据结构（称为“脚本”）表明熟悉动作的顺序。例如，哄小孩子睡觉通常要做以下动作：盖被子、读故事、唱首摇篮曲、打开小夜灯。这样的数据结构既可用于问答问题，也可用来提问题。如果妈妈省掉打开小夜灯的动作，就会出现这样的问题，如“为什么”以及“接下来发生了什么”，换句话说，这里有故事开始的缘由。因此，这种形式的知识表示被用于自动书写故事，也正是和人类能正常交谈

的“陪护”计算机所需要的知识表示形式（见第3章）。

概念的另一种知识表示形式是语义网络（这些是集中式网络，见第4章）。20世纪60年代，罗斯·奎利恩（Ross Quillian）率先提出了几个延伸示例（例如WordNet<sup>[2]</sup>）作为人类联想记忆的模型，现在属于公共数据资源。语义网络通过以下方法连接概念：如同义、反义、从属、上位、部分—整体这样的语义关系；以及将真实的世界知识比作语义学的联想连接（见第3章）。

语义网络可能增加为音节、初始字母、语音学和同音异义词编码的连接，来表示概念和词。金·宾斯泰德（Kim Binsted）的JAPE和格雷姆·里奇（Graeme Ritchie）的STAND UP在使用这种网络，它们基于双关语、解释和变换音节来制造笑话（9种不同类型）。例如，问：什么叫沮丧的火车？答：低压机车；问：羊和袋鼠生出来的宝宝是什么？答：一位毛茸茸的跳高运动员。

注意：语义网络与神经网络不同。我们将在第4章中看到，分布式神经网络以迥然不同的方式表示知识。在神经网络中，单个概念不是用精心定义的联想网络中的单个节点来表示，而是用整个网络上活动的变化模式来表示。这类系统可以容忍冲突迹象，因此不需要考虑保持逻辑一致性的问题（将在下一节描述）。但它们无法进行精确推理。不过，这种知识表示类型十分重要（并且是实际应用的一个重要基础），值得我们用一个单独的小节对其展开讨论。

---

#### 注释

[1]CYC是一个致力于将各个领域的本体及常识知识综合地在一起，并在此基础上实现知识推理的人工智能项目。其目标是使人工智能的应用能够以类似人类推理的方式工作。这个项目是由道格拉斯·莱纳特（Douglas Lenat）在1984年设立的，由Cycorp公司开发并维护。——译者注

[2]WordNet是一个由普林斯顿大学认识科学实验室在心理学教授乔治·A. 米勒的指导下建立和维护的英语字典。开发工作从1985年开始，从此以后该项目接受了超过300万美元的资助（主要来源于对机器翻译有兴趣的政府机构）。由于它包含了语义信息，所以有别于通常意义上的字典。WordNet根据词条的意义将它们分组，每一个具有相同意义的字条组称为一个synset（同义词集合）。WordNet为每一个synset提供了简短、概要的定义，并记录不同synset之间的语义关系。——译者注



## 逻辑和语义网

如果一个人的最终目标是强人工智能，逻辑似乎是一种超级不错的知识表示。因为逻辑普遍适用。原则上来说，相同的表示（相同的逻辑符号主义）可以用来表示视觉、学习和语言等，当然也适用于由此产生的任意集成。此外，它提供了很有说服力的定理证明方法，以处理信息。

所以，早期人工智能中的知识表示方式首选谓词演算。这种逻辑比命题逻辑的表示能力更强，因为它可以“进入句内”来表达句子的意思。以“这个商店有一顶适合所有人的帽子”这个句子为例。谓词演算可以清楚区分这句话三种可能的意思：“对于每个人来说，这家商店有一顶适合他们的帽子”；“在这家商店有一顶尺寸可调的帽子，适合任何人”；和“在这个商店有一顶帽子（假定被折起来）足够大，可以同时适合所有人。”

对许多人工智能研究人员来说，谓词逻辑仍然是首选。例如，CYC的框架就是基于谓词逻辑。组合语义学中的自然语言处理（NLP）表示也同样如此（见第3章）。我们延伸谓词逻辑来表示时间、原因或职责/道德。当然，这取决于某人已经提出了这些形式的模态逻辑，但这并非易事。

然而，逻辑也有缺点。

第一个缺点包含组合爆炸。人工智能中广泛使用的逻辑定理证明方法是消解法。利用这种方法得出的结论可能本身是正确的，但它与目标结论并不相关。启发法用来指导和限制结论，并决定何时停止证明（魔法师的弟子<sup>[1]</sup>做不到）。但这些方法也并非万无一失。

第二个缺点是消解定理证明，假定非—非—X就意味着X。这个观点大家并不陌生：反证法就是首先假设某命题不成立（对原命题的结论进行否定），然后推理出明显矛盾的结果，从而下结论说假设不成立，原命题得证。如果被推理的域被完全理解，那么这在逻辑上是正确的。但是，使用内置消解程序（例如许多专家系统）的用户通常假定找不出矛盾来，这就意味着不存在矛盾，即所谓的“失败则否定”。这往往是一个错误。在现实生活中，证明某事是假，和不能证明它是真完全不是一回事（如在猜测伴侣是否欺骗你的时候）。因为

还有许多不知道的证据（潜在假设）。

第三个缺点是，在经典（“单调推理”）逻辑中，一旦某事被证明是真，那它永远是真。在现实中，情况并不总是如此。我们可以有充分理由认为X为真（也许它是一个缺省赋值，甚至是通过仔细论证或从有说服力的证据中得出的结论），但后来可能会发现X不再是真，或者从最开始就不是真。如果是这样，我们也必须相应地改变自己的认知。对于基于逻辑的知识表示，这说起来容易做起来难。许多研究者受到麦卡锡的启发，已经试图提出可以容忍不断变化的真值的“非单调推理”逻辑。类似地，人们已经定义了各种“模糊”逻辑，其中的语句能够被标记为可能/不可能或者未知，而不是真/假。即便如此，防止单调性的可靠方法仍未找到。

有些人工智能专家在研究基于逻辑的知识表示法时，基本上都是越来越想找到知识或意义的本元。但他们不是先驱：麦卡锡和海斯（Hayes）在其合著的论文《从人工智能立场中衍生出来的某些哲学问题》（*Some Philosophical Problems from an AI Standpoint*）中就在做这件事。大家对这篇文章讨论的很多问题都不陌生：从自由意志到非真实条件句。这些问题包含宇宙的基本本体论：状态、事件、属性、变化、动作……什么？

除非某人在内心深处就笃信形而上学（这种激情十分罕见），不然为什么要关心本体这种东西？为什么现在对这些神秘问题的探讨越来越多？很显然，如果试图设计强人工智能，就必须考虑知识表示能够使用何种本体。我们在设计语义网的过程中也要考虑这些问题。

语义网与万维网不同。自20世纪90年代以来，我们就有了万维网。语义网甚至不是技术发展的最新水平：它是未来的技术发展水平。如果存在语义网，而且当它真的存在时，机器驱动的联想搜索将通过机器的理解力得到改进和补充。这样一来，应用程序和浏览器就可以访问互联网上的任何信息，并在推理问题的过程中合理整合不同内容。这项艰巨任务由蒂姆·伯纳斯-李爵士（Sir Tim Berners-Lee）指导，这项任务甚至可以说是苛求，不仅需要在硬件和通信基础设施方面取得巨大的工程进步，还需要网络漫游程序加深理解它们正在做什么。

谷歌、一般的NLP程序等这类搜索引擎通常可以找到单词或文本之间的关联，但不存在理解力。这里不是指哲学上的理解力（见第6章），而是指一种经验性的东西，是实现强人工智能的另一个障碍。尽管有一些例子听起来很诱人，但终究是骗人的，如IBM公司的沃森、Siri和机器翻译（都将在第3章中讨论），今天的计算机并不知道



它们“读”或“说”的东西是什么意思。

缺乏理解力的体现之一是各程序之间不能彼此交流（相互学习），因为程序不同，知识表示形式或基本本体也不同。如果语义网研究人员可以找到一种通用的本体论，那么让机器理解其接收的东西可能不再只是空想。因此，在20世纪60年代，人工智能领域提出的形而上学的问题，如今因其实用性而变得非常重要。

---

注释

[1] 《魔法师的弟子》（The Sorcerer's Apprentice）是德国诗人歌德在1797年创作的一首诗。诗的内容大致是：一位老魔法师离开店铺前，给弟子留了一堆活。这位弟子在老魔法师离开之后，因疲于用桶提水，所以对一把扫帚施了魔法，但是他并没完全熟练掌握这种魔法。可想而知，接下来地板上到处都是水。而这位弟子也意识到自己无法让扫帚停下来，因为他根本不知道怎么让它停。——译者注

## 计算机视觉

今天的计算机不能像人类一样理解视觉图像（同样，这是一种经验性的东西：强人工智能是否可以运用有意识的视觉现象学，将在第6章中讨论）。

自1980年以来，有关人工智能视觉的各种知识表示研究大多基于心理学，特别是大卫·马尔（David Marr）和詹姆斯·吉布森（James Gibson）的理论。马尔注重构建3D表示（通过反转图像形成的过程），而不是用它们执行动作。吉布森强调视觉可供性对动作的支持，即一些视觉线索，它们能提示一条路或一根承重树枝，甚至是友好的或敌对的物种成员。尽管有了这些心理学研究基础，但是目前的视觉程序仍然严重受限。

不可否认，计算机视觉已经取得了显著成就，例如面部识别的成功率达到98%，阅读草书笔记，注意到停车场内某人的可疑行为（一直站在车门边不走），甚至能比人类病理学家更准确地鉴定出一些病变细胞。面对这些成功，人们开始变得焦虑。

但是程序（很多是神经网络，见第4章）通常必须准确地知道它们想要什么，例如，一张脸不能倒置、不能侧摆、一点也不能被别的东西挡住（98%的成功率），而且还得是在特定的光亮下。

“通常”这个词很重要。2012年，谷歌实验室整合了1000台大型（16核）计算机，建成了一个巨大的神经网络，拥有10亿多个联结。然后，由研究人员将来自YouTube视频的1000万张随机图片放入这个具备深度学习能力的网络中。该网络没有被告知要找什么，图像也没有被标记。然而，三天后，其中一个单元（一个人工神经元）学会了对一张猫脸图像和一张人脸图像作出反应。

这让人印象深刻吧？嗯，是的。这也很吸引人：研究人员很快想到了大脑中的“祖母细胞”。自从20世纪20年代以来，神经科学家对于是否存在祖母细胞的观点各持己见。如果说它们存在，那就表示大脑中有些细胞（单个神经元或小组神经元）当且仅当察觉到祖母或某些特定特征的时候会被激活。显然，谷歌的猫脸识别网络情况类似。而且，虽然猫脸必须完整且摆正，但其大小可以改变，也可以出现在200×200阵列的不同位置。研究人员训练精心预选的（但未标记）人脸图像（包括侧脸）上的系统，由此发现某一单元偶尔能够辨别避开

取景器的脸。

此类成就不久将纷纷涌现，甚至更加令人惊叹。多层网络已经在面部识别领域取得了巨大进步，有时还可以找到图像最突出的部分，并给出相应的语言描述（例如，“在户外市场购物的人”）。最近发起的“大规模视觉识别挑战赛”每年都在增加可识别视觉种类，并减少对相关图像的约束（例如，对象的数量和遮挡）。然而，这些深度学习系统仍未克服旧系统的一些弱点。例如，猫脸识别器这类系统并不理解什么是3D空间，也不知道“侧脸”或遮挡的真实含义，甚至是为机器人设计的视觉程序对它们也只是一丁点了解。

再看看机遇号（Opportunity）和好奇号（Curiosity），这两种火星漫游机器人分别于2004年和2012年登陆。它们依赖于特殊的知识表示法：针对其可能面临的3D问题量身打造的启发法。它们无法在一般情况下寻路或操控对象。有些机器人模拟有生命的视觉，在这个过程中身体的运动可提供有用信息（因为它们系统地改变视觉输入）。但这些机器人可能忽略一些可行路径，也不知道自己的手能拿哪些陌生的东西。

在本书出版之际，有些机器人可能已经具备上述能力了，但它们还是会受限。例如，它们理解不了“我不能把那个东西拿起来”这句话的意思，因为它们根本不知道什么叫“可以”和“不可以”，它们的知识表示也有可能仍然不具备必要的模态逻辑。

有时候，视觉能够忽略3D空间，比如阅读笔迹。在许多高度受限的2D任务中，计算机视觉比人类视觉犯的错误要少。的确，有时候人眼无法识别的复杂模式（例如，X射线中的模式），计算机视觉能够采用高精尖的非自然技术来分析（同样，3D计算机视觉常常利用非自然方式取得显著成就）。

但是，即使是2D，计算机视觉也有局限。尽管不乏研究类比表示或图像表示的工作，但是人工智能在解决问题时无法以可信赖的方式使用图表，而我们在几何推理中或在封皮背面绘制抽象关系时可以做到这一点（同样，心理学家也还不知道做到后者的方法）。

总之，大多数人类视觉的成就优于今天的人工智能。通常，人工智能研究人员不清楚要问什么问题。以“整齐折叠光滑的真丝礼服”为例。没有机器人可以做到这一点（虽然我们可以一步一步地指导它们如何折叠长方形毛巾）。还有穿T恤：必须先把头套进去，而不是先穿袖子。但是这是为什么？在人工智能中几乎不存在这种拓扑问题。

难道这意味着人类水平的计算机视觉就无法实现吗？不是。不过要做到这一点，比大多数人想得更困难。

因为描述视觉的特性就是一块难啃的“硬骨头”。所以，它是第1章所列事实的一个特例：人工智能已经告诉我们，人脑的丰富程度和微妙程度超出了心理学家之前的猜想。事实上，这也是我们从人工智能中学到的最重要的一课。

## 框架问题

无论在什么样的领域中，要找到合适的知识表示都很难，部分原因是需要避免框架问题（注意：虽然在用框架作为概念的知识表示时，也出现过这个问题，但是此处“框架”的含义不一样）。

正如麦卡锡和海斯最初定义的那样，框架问题指假定（由机器人规划时）一个动作仅会引起这些改变，然而它也可能导致那些改变。一般来说，只要人类默认的含义被计算机忽略，框架问题就会出现，因为这些含义没有被阐明。

经典案例就是猴子和香蕉的问题，其中的问题解决者（可能是机器人的人工智能规划程序）假定在框架外不存在相关事物（见图2—1）。

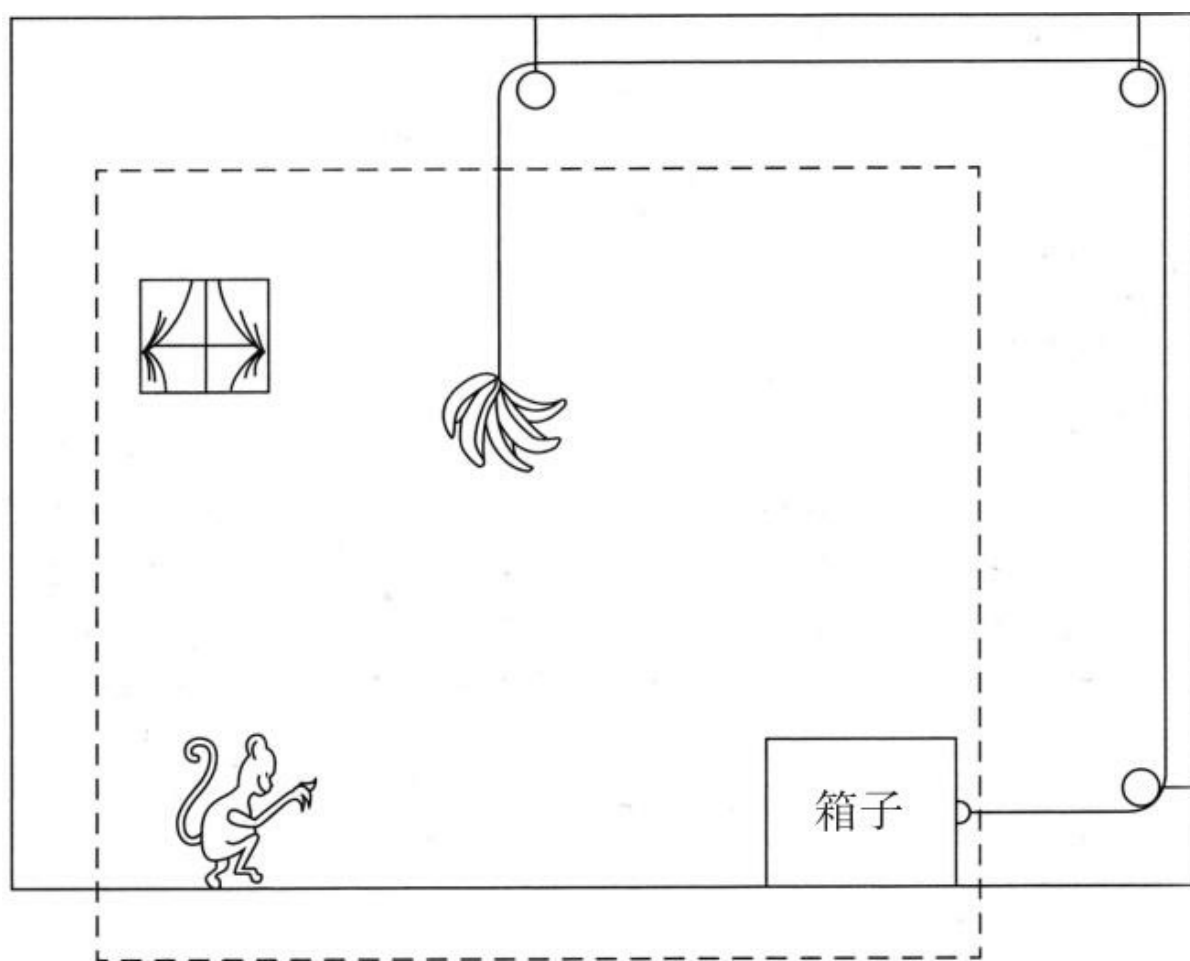


图2—1 猴子和香蕉问题：猴子怎么得到香蕉

图注：解决这个问题的一般方法是：假定相关的“世界”是虚线框架内表示的世界（虽然没有明确说明）。换句话说，这个框架之外不存在什么东西能让它内心发生重大变化，它会继续移动箱子。

资料来源：摘自玛格丽特·博登1977年所著的《人工智能和自然人》（Artificial Intelligence and Natural Man）一书第387页。

我自己最喜欢的例子是：假设一位20岁的男士可以在一个小时内采摘10磅黑莓，一位18岁的女士可以采摘8磅，那么他俩要是一起去采摘黑莓，一共能摘多少？当然，“18磅”不是一个合理的答案。有可能超过18磅（因为他们都想炫耀），也有可能少于18磅。我是在五十年前第一次听说这个例子的，不过它在当时更具代表性。但为什么答案不确定呢？这里涉及什么样的知识？强人工智能是否能克服这种看似普通的算术事实？

框架问题产生的原因是人工智能程序没有人类的关联感<sup>[1]</sup>（见第3章）。如果能知道每个动作可能带来的所有后果，那么框架问题就可以避免。有些技术或科学领域已经做到了这一点（所以人工智能科学家有时会声称框架问题已经被解决。或者，如果他们特别严谨，会说“或多或少”得到了解决）。然而，一般来说，这个问题并没有解决，这也是人工智能系统缺乏常识的症结所在。

简而言之，框架问题就潜伏在我们周围，是实现强人工智能的主要障碍。

---

注释

<sup>[1]</sup>如无法像人一样知道某些事与自己手头上的事情相关联。——译者注

## 智能体和分布式认知

一个人工智能体（Agents）是一个独立的（“自主的”）程序，有时可以比作膝盖反射，有时可以比作一个微型大脑。电话应用程序或拼写校正器可以称为智能体，但一般情况下又不是，因为不同的智能体之间通常会进行合作。它们会利用自己十分有限的智能与其他智能体合作或并肩达到单凭自己的力量无法实现的目的。而且，智能体之间的相互作用与智能体一样重要。

有些智能体系统的组织形式是分层控制的，可以说是统治者和被统治者。许多系统则是分布式认知的典范，包括无分层控制结构的合作（因此有了上面两种说法：“合作”和“并肩”）。没有中心规划，没有自上而下的影响，也就没有单个智能体处理所有相关知识的情况。

从自然角度来看，分布式认知的例子有：蚂蚁寻迹、船舶导航和人类的大脑。蚂蚁寻迹来源于很多蚂蚁个体的行为，它们在走路时会自动滴落（并跟随）化学物质。类似地，船舶的导航和操纵来自船员的互动活动，甚至是船长也不一定具备所有必要的知识，而且一些船员具备的知识确实少得可怜。即使是单个大脑也包含分布式认知，因为它集成了许多有感知的、激发性的和有情感的子系统（见第4章和第6章）。

从人工角度来看，分布式认知的例子包括：神经网络（见第4章）；人类学家的船舶导航计算机模型；情境机器人学、蜂群智能和群体机器人学方面的人工生命研究（见第5章）；金融市场（智能体是银行、对冲基金和大股东）的符号人工智能模型；LIDA的意识模型（见第6章）。

分布式认知的意识也有助于人机交互设计，例如协作式工作场所和计算机接口。因为如伊冯·罗杰斯（Yvonne Rogers）所言，它阐明了“人、人工制品和技术系统之间复杂的相互依赖性。在运用传统认知理论时，往往会忽略这些要素”。

那么很显然，人类水平的强人工智能包含分布式认知。



## 机器学习

人类水平的强人工智能还包括机器学习。然而，这并不是说要和人类一样。该领域的研究始于心理学家进行的有关概念学习和强化方面的工作。不过它现在取决于可怕的数学技术，因为所使用的知识表示都包含概率论和统计学可以说从心理学入手的研究已经远远落后了。一些现代机器学习系统与人类头部内可能发生的事情之间少有或甚至没有相似之处。但是，贝叶斯概率在该人工智能领域研究中的运用之广，不亚于最近的认知心理学和神经科学理论。

今天的机器学习超级赚钱。它不仅用于数据挖掘，还用于处理大数据，因为超级计算机每秒可以做千万亿次计算（见第3章）。

一些机器学习使用神经网络，但是大多数都依赖符号人工智能，并辅以统计算法。事实上，做事情的是统计学，而GOFAI是“管理者”，引导工人去工作场所。因此，一些专业人士把机器学习看作计算机科学或统计学，而不是人工智能。不过这并没有明确的界限。

有些计算机科学家故意排斥麦卡锡最先提出的“人工智能”的概念，认为其哲学含义有问题（见第6章）。也有些科学家有意避而不谈，因为他们不喜欢大多数人工智能研究属于实验性研究，而且相对来说不成体系。

机器学习分三种类型：监督式学习、非监督式学习和强化学习（这种划分源于心理学，可能还涉及不同的神经生理机制，跨物种的强化学习包含多巴胺）。

在监督式学习中，程序员“训练”系统，方法是为一系列输入（标记的示例和无标记的示例）定义一组希望得到的结果，并且不断反馈该系统是否完成所希望的结果。学习系统对相关特征作出假设。只要系统分类出错，它就相应地修正自己的假设。具体的错误消息至关重要（不只是反馈系统出错了）。

在非监督式学习中，用户不提供希望得到的结果或错误消息。学习由原则驱动，而原则是指共同发生的特征产生它们以后将共同发生的期望。非监督式学习可以用来发现知识。程序员不需要知道数据中有什么模式/分类，系统自己会找到它。

强化学习受奖励和惩罚所驱动：反馈信息告诉系统它刚刚做的事情是好还是坏。通常，强化不只是二进制，还是由数字表示，如视频游戏中的分数。“它刚刚做了什么”可能是单个决定（例如游戏中的某个步骤），也可能是一系列决定（例如国际象棋走到最后将军了）。在一些视频游戏中，每动一步，数字分数都会更新。在极其复杂的情况下，例如象棋，仅在许多决定完成之后才能得到成功或失败的信号，用于信用分配的某一程序识别出那些最可能带来成功的决定（进化人工智能是一种强化学习形式，其中成功由适应值函数监控，详见第5章）。

符号机器学习通常假定（并不完全正确）学习的知识表示必须包含某种形式的概率分布。许多学习算法假定（通常是错误的）数据中的每个变量相互独立，具有相同的概率分布。此i.i.d.假设是许多数学概率理论的基础，而后者又是许多算法的基础。数学家采用i.i.d.假设，让数学运算变得更简单。同样，在人工智能中使用i.i.d.可以简化搜索空间，这样易于问题的解决。

然而，贝叶斯统计处理条件概率，其中的内容或事件并不独立。这里的概率取决于与域相关的分布式现象。这种形式的知识表示不仅更加实际，而且要是出现新的现象，概率还可以发生变化。贝叶斯技术在人工智能以及心理学和神经科学中的作用不断凸显。有关“贝叶斯大脑”的一些理论（见第4章）通过使用非i.i.d.数据，去驱动和微调在知觉和运动控制中的非监督式学习。

因为有各种各样的概率理论，所以有许多算法都适用于不同类型的学习和不同数据集。例如，接受i.i.d.假设的支持矢量机（Support Vector Machine）广泛用于监督式学习，特别是如果用户缺乏该域的专业先验知识的时候。如果不计特征的顺序（如在搜索单词而不是短语时），则可以用“词袋”算法。如果不用i.i.d.假设，贝叶斯统计（“亥姆霍兹机器”）可以向分布式数据学习。

大多数研究机器学习的专家使用现成的统计方法。这些方法的发起者备受行业好评：Facebook最近聘用了支持矢量机的创建者；在2013年和2014年，谷歌公司聘用了几位深度学习的重要发起者。

基于多层网络的深度学习（见第4章）是一大新进步，前景光明。利用深度学习，输入数据中的模式能在各个层级中被识别出来。换句话说，深度学习发现多层知识表示，例如从像素到反差检测器，到边缘检测器，到形状检测器，到对象部分，以及到对象。

谷歌研究YouTube时出现的猫脸检测器就是一个很好的例子。还

有最近《自然》（Nature）杂志报道某强化学习程序（“DQN”算法）已经学会了玩经典的Atari 2600 2D游戏。虽然这个程序的输入只有像素和游戏得分（并且提前知道的只有每场比赛的动作数量），但是在49场比赛中，它有29场超过了75%的人类学习者，并在22场比赛中胜过专业游戏测试人员。

这一成就的进步空间有多大还有待观察。虽然DQN算法偶尔能找到最佳策略，如按时间排序的动作，但如果规划游戏需要更长一段时间，那DQN算法就不能控制游戏。

未来的神经科学可能会告诉人们如何改进这个系统。当前版本的DQN借鉴了休伯尔—威泽尔（Hubel-Wiesel）的视觉受体——视觉皮层中的细胞，只对运动或特定方向的直线作出回应（这也没什么大不了的，休伯尔—威泽尔的受体还启发了伏魔殿——一种学习模式（见第1章）。更奇怪的是，此版本的DQN设计还借鉴了海马睡眠期间的“经验重放”行为。DQN系统与海马一样，存储了过去的样本或经验，并在学习期间快速重新激活这些样本和经验。这个特征十分关键，设计师指出，如果特征出现问题，那系统的性能将“严重恶化”。

## 通用系统

Atari游戏玩家总是能振奋人心，上《自然》杂志也无可厚非，其中一部分原因是它又朝着强人工智能迈进了一步。这个算法在没有使用手工知识表示的情况下习得大量技能，完成包含高维感官输入的任务。

然而（如本章开头所述），完整的强人工智能可以完成更多任务。打造一位高性能的人工智能“专家”已经很困难，打造一个人工智能“通才”更是难上加难（深度学习不是答案：它的狂热追随者承认，需要“新范式”将其与复杂推理结合，而后者是“我们没有一点头绪”的学术密码）。所以，大多数人工智能研究人员放弃早期设想，转向各色细分任务，他们取得的成就通常让人惊叹。

也有强人工智能先驱们始终保持着雄心壮志，如纽厄尔和约翰·安德森（John Anderson）。他们分别于20世纪80年代初提出了SOAR和ACT-R这两个系统。三十年过去了，这两个系统仍在不断完善（和使用）。然而，它们过度简化了任务，只关注人类的一小部分能力。

1962年，纽厄尔的同事西蒙研究了一只蚂蚁在崎岖地面上行走的之字形路径。他说，蚂蚁的每个动作都是蚂蚁对其当时感知的情境作出的直接反应（这是情境机器人学的精髓，详见第5章）。十年后，纽厄尔和西蒙所著的《人类问题求解》（Human Problem Solving）一书将人类的智力描述成和蚂蚁的智力类似的东西。根据他们的心理学理论，知觉和微观运动行为由在问题解决期间存储在记忆中或新建的内部表示（IF—THEN规则或“产生式规则”）来补充。

他们说：“被视为行为系统的人类很简单。”但是，突然出现的行为复杂性十分重要。例如，他们表示，只包含14条IF—THEN规则的系统能够解决算式谜问题（例如在这个和式中，将字母映射到数字0到9：DONALD+GERALD=ROBERT，其中D=5）。有些规则处理目标或子目标的组织问题；有些规则指挥注意力（指向特定的字母或列）；有些规则回想以前的步骤（中间结果）；有些规则辨识假启动；还有些规则回溯，以从这些假启动中恢复。

他们认为，算式谜是所有智能行为计算架构的典范，所以该心理学方法适合“通才”人工智能。1980年，纽厄尔与约翰·莱尔德（John Laird）和保罗·罗森布鲁姆（Paul Rosenbloom）开发了成功导向型成

就实现系统（Success Oriented Achievement Realized，以下简称SOAR）。总的来说，它是一个认知模型，它的推理整合了知觉、注意力、记忆、联想、推理、类比和学习。像蚂蚁一样的（情境）反应结合了内在的深思熟虑。事实上，深思熟虑往往带来反射性反应，因为以前使用过的子目标序列可以“分块拼成”一个规则。

事实上，SOAR并不能模拟认知的所有方面。人们认识到了一些缺陷以后，就不断将其完善。今天的SOAR有很多用途，从医疗诊断到工厂调度等。

安德森的认知架构思维的自适应控制系统（Adaptive Control of Thought，以下简称ACT-R）是混合系统，它结合了产生式规则系统和语义网络（见第4章）。这些程序通过识别环境中的统计概率，模拟联想记忆、模式识别、意义、语言、问题求解、学习、图像和（自2005年以来）感知运动控制。ACT—R系列主要是科学人工智能领域中的一次操练。虽然商业机器学习已经不再关注心理学（机器学习的起源）方面的研究，但是ACT—R仍在继续（最近还包含神经科学，例如，与“模块化的”大脑系统类似的IF—THEN规则集）。

ACT-R系统的一个重要特征是整合了程序性知识和陈述性知识[1]。有人可能不知道如何在几何证明中运用欧几里德定理，但他可能知道这个定理正确。ACT-R系统通过构造数百个新的产生式规则来学习如何应用一个命题性事实，这些新规则可以控制它在不同情况下的使用。它学习哪些目标、子目标和子次级目标……在哪些条件下是相关的，某一特定动作在不同情况下会导致哪些结果。总之，它通过做事来学习，并且它（像SOAR）可以将有序执行的几条规则整合到一条规则中。这就与人类专家和新手解决“同一个”问题的情况类似：一个是不假思索，一个是煞费苦心。

ACT-R系统有多种应用。它的数学导向（一款智能家教系统）提供个性化反馈，包括相关领域的知识以及问题求解的目标/子目标结构。得益于定量分块，使数学导向的建议粒度随着学生的学习进度而不断变化。其他应用有NLP、人机交互、人类记忆力和注意力、驾驶和飞行以及视觉网络搜索。

SOAR和ACT是强人工智能时代另一个早期尝试——道格拉斯·莱纳特的CYC的“同辈人”。它是一款符号人工智能系统，于1984年被推出，目前仍在升级当中。

到2015年，CYC不仅包含62000条“关系”，能够链接其数据库中



的概念，还包括这些概念之间的数百万条链接。这些关系包括存储在大型语义网络（见第3章）中的语义和事实关联，以及朴素物理学上的无数事实——所有人类具备且与物理现象相关的非形式化知识（例如丢弃和溢出）。系统同时使用单调和非单调逻辑以及概率来推论数据（目前，所有的概念和关系都是手写编码，但是贝叶斯学习理论不止于此。因此，CYC能够自己从互联网上获得知识）。

几家美国政府机构都在使用CYC，包括美国国防部（例如监测恐怖团体）和美国国家卫生研究院，还有一些大型银行和保险公司。CYC的较小版本OpenCyc作为各种应用程序的背景资源已被公开发布，人工智能研究人员现在可获得更简洁的版本（ResearchCyc）。虽然OpenCyc定期更新（最近一次在2014年），但它只包含CYC数据库的一小部分数据和一小部分推理规则。不过完整（或接近完整）的系统最终将在市面上出售。但是，如果不采取特别措施，它可能会落入坏人手中（见第7章）。

莱纳特在1986年的《人工智能杂志》（AI Magazine）中将CYC描述为“使用常识性知识来克服脆弱性和知识获取瓶颈”。也就是说，CYC的主要任务就是解决麦卡锡预见的挑战。如今在模拟“常识”推理以及“理解”所处理的概念方面，它已经成为佼佼者（有些概念甚至连NLP程序也无能为力，详见第3章）。

然而，它也有许多缺点。例如，不能很好地处理隐喻（虽然数据库包括许多隐喻）。它忽略了朴素物理学中各种不同的方面。它的NLP虽然在不断完善，但仍然有极大的进步空间。它尚未包含视觉。总之，就算它的目标是包罗万象，但现在确实不包含人类知识。

---

#### 注释

[1] 在认知心理学的知识学习的心理原理中，安德森将语文知识分为两类：一类为陈述性知识（declarative knowledge），指事实性和资料性知识；另一类为程序性知识（procedural knowledge），指按一定程序操作从而导致结果的知识。——译者注

## 梦想复兴

纽厄尔、安德森和莱纳特已经从人工智能这个舞台上消失了将近30年。然而，最近人们对强人工智能的兴趣又开始明显上升。2008年开始召开的人工智能年度会议，除了SOAR、ACT-R和CYC外，还出现了其他所谓的“通才”系统。

2010年，机器学习的倡导者汤姆·米切尔（Tom Mitchell）带领卡内基梅隆大学机器学习研究团队开发了NELL系统（NeverEnding Language Learner，意为永不停止的语言学习者）。这种“常识”系统通过不停地（在写入时设定为5年）搜索网站以及通过接受公众的在线更正来构建自己的知识体系。它可以基于自己的（未标记的）数据完成简单推理，例如，运动员乔·布洛格斯（Joe Bloggs）在戴维斯队，所以他打网球。它从具有200个分类和关系的本体开始（例如，专家，是由于……），5年后，它扩大了本体，并积累了9000万个候选可信度，每个都有自己的置信水平。

坏消息是NELL系统并不知道你可以用字符串拉取对象，但不推送这些对象。事实上，所有强人工智能系统的推定常识都受到严重限制。臭名昭著的框架问题已经被“解决”了这种说法产生了严重误导。

NELL系统现在有一个姐妹程序叫NEIL（Never-Ending Image Learner，意为永不停止的图像学习者）。有些部分视觉强人工智能将逻辑—符号知识表示与类推或图示表示相结合〔亚伦·斯洛曼（Aaron Sloman）早在多年前就已经开始这项研究，但对它的理解仍然不明朗〕。

此外，斯坦福大学研究所在研究CALO（Cognitive Assistant that Learns and Organizes，意为学习和组织的认知助教）的过程中意外开发了Siri应用程序（见第3章）。2009年，苹果以2亿美元收购Siri。当前类似的活跃项目包括斯坦·富兰克林（Stan Franklin）的LIDA（我们将在第6章进行讨论）和本·格策尔（Ben Goertzel）的OpenCog，它们在丰富的虚拟世界中或向其他强人工智能系统学习事实和概念（目前LIDA和CLARION这两个“通才”系统侧重意识）。

更近的一个强人工智能项目于2014年开始，旨在构建“机器人道德能力的计算架构”（见第7章）。除了上述困难，它还将必须面对许多与道德相关的问题。



一个真正达到人类水平的系统将会面临更多问题。难怪强人工智能是如此地难以捉摸。

## 缺失的方面

如今几乎所有的“通才”系统都聚焦于认知。例如，安德森的目标是详细说明“认知心理学中的所有子领域是如何相互联系的”（“所有”子领域？虽然他讨论了运动控制，但他没有讨论有时在机器人学中起到重要作用的触摸或本体感知）。真正通用的人工智能还将包括动机和情感。

一些人工智能科学家已经认识到了这一点。虽然马文·明斯基和斯洛曼都没有建立全脑模型，但都细致描述了整个大脑的计算架构。

我们将在第3章中概述斯洛曼的MINDER焦虑模型。乔斯查·巴赫（Joscha Bach）借鉴了他的研究成果和迪特里希·多纳（Dietrich Dorner）的心理学理论，开发了MicroPsi——一种基于七种不同“动机”并在规划和动作选择中运用“情感”因素的强人工智能。它也影响了上述的LIDA系统（见第6章）。

要实现真正的强人工智能，只做到这些还远远不够。明斯基的未来人工智能宣言——“走向人工智能”，既发现了障碍，也给出了承诺。许多障碍还有待克服。从第3章可以看到，人类水平的强人工智能仍然离我们很遥远。许多人工智能专业人士不同意这个观点。有人甚至说，强人工智能很快将成为超人工智能（ASI）——“S”指超人。与此同时，智人将淡出研究人员的视线（详见第7章）。

## 03 语言、创造力和情感

人工智能的一些领域似乎特别具有挑战性，如语言、创造力和情感。如果人工智能不能模拟它们，要实现强人工智能就好似做白日梦。

无论就上述哪一方面而言，我们所取得的成就都早已超出了人们的想象。即便如此，一些困难仍然显著存在。这些典型的“人类”特征只是在一定程度上被模拟（人工智能系统是否能够具有真正的理解力、创造力或情感，我们将在第6章中讨论。我们在此关注的是人工智能系统是否有可能拥有它们）。

## 语言

无数的人工智能应用程序使用NLP（自然语言处理）。大多数程度关注的是计算机对呈现给它的语言的“理解”，而不是计算机自己创造语言。因为对于NLP而言，创造比接受更困难。

其中的困难包括主题内容和语法形式。例如，我们在第2章中看到，熟悉的动作顺序（“脚本”）可能用作人工智能故事发生（哄孩子上床睡觉的母亲改变惯常动作）的缘故。这并不是说，背景知识表示就一定包含足够的人的动机，所以它不一定能使故事变得有趣。要就一家公司不断变化的财务状况写一份年度报告，买个系统就完事了，但是可以看出这个系统创造的“故事”非常无聊。计算机创作的小说和肥皂剧情节确实有——但是任何以细微的地方处理得好为评判标准的奖项往往都与它们无关（用人工智能进行翻译或总结人类创造的文本，得到的译文和总结在内容上可能更丰富，但还是因为源语文本是人类完成的）。

至于语法问题，计算机创作的散文有时在语法上就不正确，而且通常很不恰当。人工智能对画圈打叉游戏（井字游戏）的描述能够包含从句或子从句结构，很好地讲述游戏的具体步骤。但是，我们也能充分理解画圈打叉游戏的概率和策略。不过，要人工智能清楚地描述很多人类故事中主角的一系列想法或动作就没那么容易了。

再谈谈人工智能接受语言，有些系统十分简单，甚至让人觉得无聊：它们仅需要识别关键字（想想电子零售网站上的“菜单”），或者预测字典里所列出的单词（想想在编辑短信时自动弹出的匹配词或句）。还有一些系统要复杂得多。

有些系统需要识别语音：要么是单个词（如自动电话购物），要么是连续语音（如实时电视字幕和电话窃听）。更有意思的是，后者的目标可能是挑选出一些特定词（如炸弹和圣战），以此抓住整个句子的意思。这绝对是NLP：首先必须区分出单词本身，这些单词是由不同声音发出来的，而且可能带有不同地方的口音或外来口音（区分单词在印刷文本中是免费的）。深度学习（见第4章）已经促使语音处理技术取得了巨大进步。

对整句的理解也有一些令人印象深刻的例子，如机器翻译；从大量自然语言文本中挖掘数据；总结报纸和期刊上的报道；以及回答一

些自由提问（频繁用于谷歌搜索和iPhone的Siri应用程序）。这些系统真的可以欣赏语言吗？例如，它们能处理语法问题吗？

在人工智能早期，人们认为语言理解需要解析句法。于是研究人员花了很大力气去编写程序，以实现这一目标。20世纪70年代初期，特里·维诺格拉德（Terry Winograd）在麻省理工学院写的SHRDLU<sup>[1]</sup>就是一个典型案例。在此之后，无数先前没听说过或认为人工智能不可能实现的人开始关注人工智能。

该程序接受英语指令，该指令告诉机器人用彩色积木搭建结构，并计算出如何移动这些积木才能实现目标。它之所以影响深远，原因很多，其中部分知识已经应用到了一般的人工智能领域。与此相关联的点在于它能够赋予复杂的句子详细的语法结构，例如：如果你之前不知道奶奶的食谱是错的，你打算在蛋糕中加多少鸡蛋。（尝试一下！）

就技术层面而言，SHRDLU不尽如人意，其中有许多程式错误，所以只为少数技艺精湛的研究人员使用。当时还出现了其他各种句法处理程序，但也没能推广到现实文本当中。总之，研究人员后来很快发现，复杂的句法分析对现成系统来说太难了。

除了句法外，在人类语言中，语境和相关性也很重要。当时没有显著成就表明人工智能能够做好这两点。

1964年，美国政府的确在自动语言处理咨询委员会（Automatic Language Processing Advisory Committee，以下简称ALPAC）的报告中宣布机器翻译不可能实现。报告预测，看好其“钱”景的人为数不多（尽管机器辅助人类翻译也许可行），认为计算机将与句法作斗争，被语境击败了，而最为重要的是，对相关性一无所知。

这就像是给机器翻译（实际上，它的资金来源一夜之间干涸）和人工智能丢了一枚炸弹。大家普遍将报告解读为：人工智能研究做了许多无用功。畅销书《计算机和常识》（Computers and Common Sense）也声称（1961年）人工智能研究是在浪费纳税人的钱。报告的发布似乎也证明了政府的高层专家同意这种观点。当时两所即将开设人工智能学院的美国大学也跟着取消了其计划。

不过人工智能的研究工作并未因此中断。几年后，精通句法的SHRDLU闪亮登场，为GOF AI做了一次成功的辩护。但是质疑很快悄然而至。NLP研究的焦点也因此逐渐转向语境而非句法。

20世纪50年代早期，一些研究人员开始重视语义语境。英国剑桥大学的玛格丽特·玛斯特曼（Margaret Masterman）研究小组用同义词词典而不是字典来处理机器翻译（和信息检索）。他们认为句法是“语言中非常肤浅和冗余的部分，被匆忙的人完全忽略了”。他们专注于词丛，而不是单个单词。他们没有尝试字对字的翻译，而是搜索同义词的相关文本。这样就可以正确翻译模糊词（如果找到了同义词的相关文本）。因此，bank可以（用法语）表示为rive或banque，这取决于语境是否分别包含诸如water（水）或money（钱）等词。

有些词的词义不同（例如鱼和水），但是常常同时出现。这些词可以强化以同义词词典为基础的语境法。时间证明事实的确如此。今天的机器翻译除了区分各类词汇层面的共性——同义词（empty/vacant）、反义词（empty/full）、归属关系（fish/animal）和包含关系（animal/fish）、同类关系（cod/salmon）以及部分/整体关系（fin/fish），还能识别主题共现关系（fish/water, fish/bank, fish/chips等）。

由此可见，总结、提问或翻译自然语言文本不一定非得处理复杂的语法。今天的NLP更多依赖于“体力”（计算能力）而不是大脑（语法分析）。数学，特别是统计学，已经取代逻辑，机器学习（包括但不限于深度学习）已经取代句法分析。这些NLP的新方法（从书面文本到语音识别）非常高效，所以在实际应用中，95%的成功率是可接受标准。

在现代NLP中，功能强大的计算机统计搜索海量（“语料库”）文本（在机器翻译中，这些是由人类配对的翻译），以找到常见的和意料之外的单词模式。它们可以知道鱼/水、鱼/蝌蚪、鱼和薯条、盐和醋的统计结果。NLP现在可以学习构建“词向量”（如第2章中所述），来表示既定概念下该词所有含义出现的概率云。不过此处的关注点通常是词和短语，而不是句法。语法没有被忽略：文本在接受检测的过程中，其中一些单词将被（自动或手动地）赋予形容词和副词之类的标记。但是句法分析却很少使用。

详细的语义分析也不多。“组合的”语义用句法分析句子的含义；这种做法仅限于研究实验室，没有大规模应用。“常识”推理器CYC对概念（词）的语义表示相对完整，因此能更好地“理解”它们（见第2章）。但这种应用也十分有限。

当前的机器翻译倒是风生水起的。有些系统包含主题很少，但有些系统则包罗万象。谷歌翻译每天为超过2亿名用户提供各种主题的机器翻译。SYSTRAN翻译系统每天为欧盟（24种语言）、北约、施

乐公司和通用汽车公司服务。

许多机器翻译的译文都近乎完美，如欧盟的文件（因为在源语文本中只用到有限子集的单词）。尽管大多数机器翻译存在问题，但还是很容易理解，因为博学的读者可以忽略译文中的语法错误和生硬的单词——就像听非母语人士说话一样。有些机器翻译出来的译文只需人类稍作编辑和修改（而日语在翻译前后需要大量编辑。如英语的过去时态vot-ed，日语没有分段的单词。而且日语的短语顺序是颠倒的。匹配不同语系的语言对机器来说并非易事）。

简而言之，人类用户可以很好理解机器翻译出的译文。同样，总结期刊文章的单语NLP程序经常能够反映论文是否值得全文阅读（完美的翻译基本不现实。例如，用日语说“要一个苹果”需要反映对话者的社会地位，但在英语中没有同类区别）。

人工智能应用程序上的实时翻译不太成功，如Skype。因为系统必须识别语音，而不是书面文本（单个词被清楚分开）。

NLP的另外两个突出应用是信息检索——加权检索（由玛斯特曼的研究小组在1976年发起）和数据挖掘。例如，谷歌搜索引擎搜索词条的时候，通常用相关性对要搜索的词条进行加权——这是在统计学层面而不是语义层面评估（即没有“理解”）。数据挖掘可以找到人类用户未意识到的单词模式。它长期用于研究市场中的产品和品牌，现在（使用深度学习）用于“大数据”，即搜集起来的海量文本（有时是多种语言）或图像，如科学报告、医疗记录、社交媒体和互联网上的词条。

政府、政策制定者和社会科学家用大数据挖掘开展侦查和反间谍活动，以及监测公众的态度，以此来了解不同群体变化的观点并对其进行比较：男/女、年轻人/老年人、北方人/南方人等。英国智库Demos（与萨塞克斯大学的NLP数据分析团队合作）分析了数以千计有关厌女症、种族群体和警察的Twitter消息。通过搜索特定事件（twitcidents）发生之后突然发出的一些推文，可以发现公众对“警方回应”的态度发生了什么样的转变。

大数据NLP给出的结果是否有用尚无定论。数据挖掘（使用“情绪分析”）不仅能度量公众兴趣度，还能度量其评价语气。然而，语气这种东西不会直接说出来。例如，一则推文包含具有明显贬损语气的种族歧视字眼，机器由此解读为“负面”情绪，但事实上可能并不表示贬损。法官在读到它的时候可能会认为这个词被用作（在这种情况下）群体身份的一种积极标记，也可能觉得它是中性描述（例如，拐



角处巴基斯坦佬开的商店），并非侮辱或辱骂。根据迪莫斯（Demos）的研究发现，只有一小部分包含种族或民族术语的推文真正带有挑衅意味。

人的判断在这些情况下依赖于语境，例如推文中的其他词。调整机器的搜索标准，以减少“负面情绪”归属可能是行得通的，但也可能行不通。搜索标准也往往颇具争议。即使人和机器的标准一致，也很难确定语境中的哪些方面能证明人类的解读合理。

在计算（甚至口头的）方面确定相关性很难，这只是其中一例。

乍一看，两个知名的NLP应用程序似乎与刚才的说法相矛盾，即苹果的Siri和IBM公司的沃森。

Siri是基于规则的私人助理，是一款能说话的“聊天机器人”，可以快速回答许多不同的问题。可以访问互联网上的一切资源——包括谷歌地图、维基百科、不断更新的《纽约时报》以及出租车和餐馆等当地服务列表，甚至还可以访问功能强大的在线自动回答系统WolframAlpha，后者利用逻辑推理“想出”而不只是“找到”各种事实性问题的答案。

用户口头对Siri（逐渐适应语音和方言）发问，然后Siri利用网络搜索和对话分析回答问题。对话分析研究人类如何就对话中的主题进行排序，以及如何安排它和人类之间的互动（如解释和协商）。利用对话分析，Siri将思考“对话者想要什么”“我应该如何回答”等问题，同时在一定程度上适应个人用户的兴趣和偏好。

简言之，Siri似乎不仅对主题相关性敏感，而且对个人相关性也很敏感。从表面上看，它真的能让人印象深刻。然而，它很容易给出荒唐的答案，如果用户偏离事实的轨道，Siri也就会跟着失去方向。

IBM公司的沃森也专注于事实。它是处理大数据的现成资源（有2880个核心处理器），已经用在一些呼叫中心，通过改良，还用到了医疗领域中，如评估癌症治疗。它不仅能像Siri一样回答直截了当的问题，还可以处理在常识游戏《危险边缘》（Jeopardy）中出现的谜题。

在《危险边缘》中，玩家不会被问到直接的问题，而是根据以答案形式提供的各种线索，以问题的形式做出正确的回答。例如，玩家被告知“1921年5月9日，这家‘尽善尽美的’航空公司在阿姆斯特丹开设了第一个客运办事处”，那么他们的答案应该是“KLM（荷兰皇家

航空)是什么?”

沃森还可以应对很多其他挑战。它的《危险边缘》游戏版本不像Siri那样能访问互联网(虽然它的医疗版本可以),不懂对话结构,也不能通过逻辑推理找到答案,然而它能对庞大但封闭的数据库进行大规模并行统计搜索。数据库中有各种文件,如无数评论和参考书,还有《纽约时报》等,里面提供了各类事实,从麻风病到李斯特(匈牙利钢琴家、作曲家)、从氢到九头蛇等。在玩《危险边缘》的时候,它的搜索由数百种反映游戏中固有概率的特殊算法作指导。它还可以从其他人类对手的猜测中受益。

2011年,沃森在玩《危险边缘》的时候,“明显地”战胜了两位人类冠军,这可以和它在IBM公司的表兄弟深蓝(Deep Blue,见第2章)的表现(打败了国际象棋大师卡斯帕罗夫)相媲美(“很显然”,因为计算机瞬间作出反应,而人类需要一些反应时间,然后才会按蜂鸣器)。但它和深蓝一样都不能稳居冠军宝座。

沃森有一次比赛失利的原因是,虽然它正确地将注意力集中在某位运动员的一条腿上,但是它忽略了在它的存储数据中有一个关键事实——这个人少了一条腿。沃森不会再犯这个错误,因为程序员现在已经标记“缺失”这个很重要的词,但它还会犯其他错误。即使在普通事实搜寻语境下,人们通常依赖的相关性判断都超出了沃森的能力范围。例如,凭一条线索找到耶稣的两个门徒,他们的名字既是十大首选婴儿名,又都以同一字母结尾。答案是“马修”(Mathew)和“安德鲁”(Andrew)——沃森立即给出了答案。人类冠军也得到了这个答案。但他的第一想法是“詹姆斯”(James)和“犹大”(Judas),他回忆说,自己之所以排除了这个答案,是因为出于某种原因,他认为犹大不是一个流行婴儿名。沃森就做不到这一点。

人类的相关性判断往往没有上面的例子那么明显,对于今天的NLP来说,这个判断太微妙了。相关性是语言/概念版的“框架问题”(见第2章),都是难啃的硬骨头。许多人会觉得让非人类系统完全掌握它简直是天方夜谭。难道仅仅因为包含的信息量过大且过于复杂,还是因为相关性是人类特有的生命形式?我们将在第6章对此展开讨论。

---

注释

[1]SHRDLU是一个用自然语言指挥机器人动作的系统,由维诺格拉德于1972年在麻省理工学院建立。——译者注

## 创造力

创造力——产生新颖的、异乎寻常的以及有价值的想法或人工制品的能力——是人类智慧的顶峰，对实现人类水平的强人工智能也是必不可少的。但人们普遍认为它很神秘。现在我们连人类是如何产生的新颖想法都没弄明白，更别提计算机了。

目前，对创造力的识别甚至都没有统一的答案：人们对一个想法是否具有创造性通常会持不同意见。有些分歧点在于：它是不是真的很新颖，以及它在何种意义上是真的很新颖。一个想法可能只是对相关个体来说是新颖的，也有可能对整个人类历史来说都很新颖（分别是“个体”和“历史”创造力的典范）。无论哪种情况，它可能多多少少和前述观点类似，会引发分歧。还有些分歧点是有关估价（包含功能意识，有时会是现象意识，参见第6章）。同一个想法，有的社会群体可能重视，而有的却不一定（比如现在的年轻人会嘲笑任何仍然喜欢看瑞典流行演唱组合乐队Abba DVD盘的人）。

人们通常认为没有什么有趣的人工智能可以体现创造力。但人工智能技术产生了许多在人类历史上属于新颖的、异乎寻常的以及有价值的想法。例如，它们被用在了发动机、药品和各类计算机技术的设计过程中。

此外，人工智能概念还有助于解释人类的创造力。借此，我们可以分出三种类型的创造力：组合型、探索型和变革型。三者包含不同的心理机制，能带来不同的惊喜。

在组合型创造力中，常见的想法以不常见的方式组合在一起。例如，视觉拼贴、有诗意的图像和科学类比（将心脏比作泵，原子比作太阳系）。新组合在统计学层面带来意外发现——这在以前是不大可能做到的事情，就像一个冷门选手不大可能赢得德比（Derby）。但它浅显易懂，所以有价值。价值大小取决于如何评判前文讨论的相关性。

探索型创造力较为常见。它充分利用了一些有文化价值的思维方式（例如，绘画或音乐的风格、化学或数学的子区域）。使用风格法则（主要是无意识地）可以产生新想法，就像英语语法可以生成新句子一样。艺术家或科学家可能无条件地探索该风格的潜力，也可能刻意大力推行它的应用或对其进行测试，以了解它可能生成哪些想法。

它甚至可能因为某一规则的些许变化（例如弱化/加强）而发生小变动。尽管这个结构很新颖，但仍然属于常见的风格。

变革型创造力继承了探索型创造力，如果现有风格受限，变革型创造力就会发生。一个或多个风格限制将被彻底改变（删除、否定、补充、替换、添加……），因此生成了之前不可能生成的新结构。这些新想法堪称异类，因为它们的出现像是天方夜谭。最初，它们可能晦涩难懂，因为以惯常思维方式很难完全理解。然而，如果新想法要被接受，它们就必须贴近惯常思维方式（有时这种接受要花很多年）。

三种创造力都发生在人工智能中——观察者通常认为创造力是人类确定的（实际上是通过图灵测试，见第6章），但可能没有像人们预期的那样多。

像组合系统就十分罕见。人们可能认为模拟组合型创造力很容易，毕竟没有什么比让计算机在已经存储的想法之间产生不常见的关联更简单了。这些关联（在历史上）通常很新颖，（在统计学上）也令人惊讶。但如果它们要有价值，就必须彼此相关。当然，我们也清楚，相关性没那么容易得到。例如，我们在第2章中提到了一些笑话生成程序，它们用笑话模板来帮助提供相关性。同理，符号人工智能基于案例的推理利用预编码的结构相似性来构造类比。因此，这些程序的组合型创造力还结合了探索型创造力。

同时，人们可能认为人工智能无法模拟变革型创造力。这种想法也是错误的。任何程序确实只能做它可能有能力做的事情，但是进化程序是可以进化自身的（见第5章）。它们甚至可以评估自己新进化的想法，但前提是程序员提供了明确的挑选标准。这样的程序通常用在追求新颖的人工智能应用上，例如设计新科学仪器或药物。

然而，变革型创造力不是一条通向强人工智能的神奇之路。它几乎不能保证产生有价值的结果。我们可以相信（在数学或科学中的）有些进化程序能够找到最优方案，但许多问题不能由最优化来定义。变革型创造力之所以有风险，是因为以前已经接受的规则被打破了。所有新结构都必须进行评估，否则就会出现混乱。但是当前人工智能的拟合函数是由人类定义的：程序不能独立改变/推断出它们。

探索型创造力最适合人工智能。这类例子不胜枚举。工程学中一些探索型的人工智能创新（如CYC的设计者设计的程序所生成的创新，见第2章）已被授予专利。对于技术熟练人员来说，他们不一定觉得获得专利的想法就属于创新，但这个想法可能是他们想要探索的

风格。有些人工智能的探索能与人类取得的杰出成就相媲美——如按照肖邦或巴赫的风格创作音乐，又有多少人能做到这一点？

然而，即使是探索型人工智能也在很大程度上依赖于人的判断。因为必须有人识别并清楚地说明风格化的法则。这通常很难。有位世界级专家在研究弗兰克·劳埃德·赖特（Frank Lloyd Wright）的“草原式住宅”时，不再描述建筑风格，宣称它们“难以理解”。后来，一个可计算的“形状语法”生成了无数个“草原式住宅”的设计，包括四十多个原创——这没什么不可信。但系统成功的根本原因还是人类分析师。只有当强人工智能自己能够分析（艺术或科学中的）风格时，它的创造性探索才是“自己的作品”。尽管最近有一些（但不多）深度学习识别艺术风格的案例（见第2章和第4章），但它的确是一项艰巨的任务。

利用人工智能，人类艺术家开发了一种新的艺术形式——数字艺术（computer-generated, CG）。它涉及建筑学、图像、音乐，以及编排和运用不太理想的文学（因为NLP面临句法和相关性方面的困难）。在数字艺术中，计算机不只是个工具，可以将其比作一支新画笔，帮助艺术家们做他们自己本来可以做的事情。相反，如果没有它，这项工作就不可能做到，或者甚至想都不用想。

数字艺术体现了上述三种创造力。由于上述原因，几乎没有任何数字艺术是组合型的。英国法尔茅斯大学教授西蒙·克尔顿（Simon Colton）的The Painting Fool软件制作了与战争相关的视觉拼贴画，但是它也收到了特殊指令，被要求搜索数据库中与“战争”相关的图像。大多数数字艺术都是探索型或变革型的。

计算机有时通过执行艺术家编写的程序，可以完全独立地生成艺术品。哈罗德·科恩（Harold Cohen）的AARON程序独立生成了线条图和彩色图像（有时创造的颜色绚丽多彩，所以科恩说，AARON是一个比他更优秀的五彩画家）。

相比之下，在交互艺术中，艺术作品的最终形式部分取决于观众的输入，当然，观众可能是无意间控制了发生的事情。有些交互艺术家将观众看作同他们一起创作的人，还有一些交互艺术家认为观众以各种方式无意间影响了艺术作品，于是将他们看作作品产生的起因[欧内斯特·埃德蒙兹（Ernest Edmonds）等艺术家同时采用了这两种方法]。在以威廉·莱瑟姆（William Latham）和乔·麦考马克（Jon McCormack）为代表的进化艺术中，计算机不断生成/改变结果，但通常是由艺术家或观众挑选的。

总之，人工智能的创造力有很多应用。在科学或艺术的一些小角落里，它有时可以和人类的创造力一决高下，甚至超过人类。但在一般情况下要和人类创造力匹敌就另当别论了。强人工智能仍然离我们很遥远。



## 人工智能与情感

和创造力一样，情感也被看作与人工智能格格不入的东西。除了直观上觉得不可能，想想情绪和情感依赖于大脑中散布的神经调节剂这一事实，构建情感的人工智能模型也似乎不太现实。

多年来，人工智能科学家们似乎也赞同这个观点。他们忽略了情感，只有在20世纪60年代和70年代出现了几个例外，如西蒙，他认为认知控制包含情感；还有肯尼斯·科尔比（Kenneth Colby），他为神经症和偏执狂构建了有趣的模型，虽然这是一个超级有野心的目标。

如今情况发生了变化。神经调节（在GasNets中，见第4章）已经被模拟。此外，许多人工智能研究小组都在研究情感。尽管大部分研究在理论层面很肤浅，但大多数都“钱”景光明，它们致力于打造“计算机伴侣”。

还有些人工智能系统是基于屏幕的机器人，有些是门诊用机器人，在与用户的交互中，不仅实用，还关注用户的舒适度以及满意度。大多数服务对象是老年人或残疾人，包括初发性痴呆病患者。还有一些是婴儿或交互式“成人玩具”。总之，包括电脑护工、机器人保姆和性玩伴。

另外，人机交互的例子包括：提醒用户购物、吃药和拜访家人；帮助编写个人日志；安排和讨论电视节目，如每日新闻；制作美食和饮料；取东西；监测生命体征（和婴儿哭泣）；说一些色情话语以及做一些色情动作等。

这其中的很多任务都包含人类的情感。人工智能伴侣就体现在它们能识别人类用户的情感或以明显带有情感的方式回应用户。例如，用户承受丧亲之痛时，可能会得到一些机器的同情。

人工智能系统已经能够用多种方式识别人类的情感。有些是生理的，如监测人的呼吸频率和皮肤电反应；有些是口头的，如注意说话的速度、语调和用词；有些是视觉的，如分析面部表情。当前的方法都相对简陋。用户的情感不仅容易被遗漏，而且容易被曲解。

计算机伴侣的情感表现通常体现在口头上。它基于词汇以及语调（如果系统能生成语音的话）。但是，系统不仅密切注意用户常用的



关键词，还以极其刻板的方式作出回应。对于用户说过的东西（可能在日记中），它偶尔可能会引用由人类创作的相关言论或诗歌。但NLP所面临的难题意味着计算机生成的文本在细节上很难做好。这些文本甚至可能不会被接受：用户可能会因为机器人伴侣没有人类的外观而被激怒或感到沮丧。同样，一只咕噜咕噜叫的机器猫可能会讨人嫌，而不是让用户觉得放松、舒服或满足。

当然也有惹人疼的机器人伴侣：帕罗（Paro）是一只可爱的交互式“海豹宝宝”，它有着迷人的黑眼睛和浓密的睫毛，是许多老年人和痴呆症患者的好伴侣（未来版本还可以监测人类的生命体征，并据此向人类看护人员发出警告）。

有些人工智能伴侣可以利用自己的面部表情，也可以用眼睛凝视，以看似富有情感的方式回应用户。有些机器人有弹性“皮肤”，覆盖在人类面部肌肉模拟物的上面，它的外形可以（向人类观察者）显示出多达十二种基本情感。基于屏幕的系统通常显示虚拟角色的面容，其表情根据（他/她）可能经历的情绪而发生改变。然而，所有这些事情都有可能（原文如此）陷入所谓的“恐怖谷理论”<sup>[1]</sup>中，即人们在遇到与人类极其相似但仍存在些许差异的生物时，就会觉得不舒服，甚至极为反感。因此，机器人或屏幕虚拟化身如果拥有似是而非的面孔，可能会让人类觉得自己正在受到威胁。

为情感空虚的人提供上述类似人类伴侣关系的做法是否符合道德标准，目前尚无定论（见第7章）。当然，有些人机交互系统（例如帕罗）似乎能够为一些人带来快乐，甚至是持久的满足感。如果没有这些系统，有些人可能会觉得生活很空虚。但是这样就足够了吗？

“伴侣”模型缺乏理论深度。专家们开发人工智能伴侣的情感是为了赚钱。他们没去想怎样让“伴侣”用情感解决自己的问题，也没有去了解情感在整个大脑运作过程中发挥什么样的作用。他们觉得情感是可有可无的附加物：他们忽视情感，除非在某些棘手的人造情境下，他们才不得不考虑。

这种不屑的态度当时弥漫在整个人工智能领域，直到最近，情况才相对有所改观。“情绪计算”之母罗莎琳德·皮卡德（Rosalind Picard）的“情感计算”把情感从20世纪90年代末期的“冷宫”中解救出来，不过她也没有深究。

一直以来，情感被人工智能忽视（与西蒙富有洞察力的评论命运相似），其中一个原因是它没有得到很多心理学家和哲学家的重视。

换句话说，他们认为智能不需要情感。相反，他们觉得情感不利于解决问题，会破坏理性。“情感可以帮助一个人决定做什么以及做这件事的最佳方法”的想法不合潮流。

情感最终会越来越重要，部分得益于临床心理学和神经科学的发展。但它能进入人工智能领域离不开马文·明斯基和亚伦·斯洛曼这两位人工智能科学家。他们一直把大脑看成一个整体，而不是像大多数同事那样，将自己的想法局限在智能领域内的某个小角落中。例如，斯洛曼正在进行的CogAff项目就关注情感在大脑计算架构中的作用。CogAff影响了于2011年发布并仍在推广中的LIDA的意识模型（见第6章），也启发了20世纪90年代末由斯洛曼研究小组带头开发的MINDER程序。

MINDER程序模拟了独自照顾几个婴儿的护士心中所产生的焦虑（功能方面）。“她”只有几项任务：给婴儿喂吃的；别让婴儿掉进路边的沟里；如果有婴儿掉进去，“她”得将婴儿送到急救中心。“她”只有几个动机（目标）：给一个婴儿喂吃的；如果已经有一个婴儿在防护栅栏后面，“她”要再放一个；将一个婴儿从沟中抱出并送去急救站；在沟边巡逻；筑围栏；将一个婴儿移至离水沟较远的安全位置；如果当前没有其他动机被激活，“她”就在托儿所周围漫步。

所以，她比真正的护士简单得多（虽然比典型的规划程序更复杂，因为后者只有一个终极目标）。然而，“她”容易感到不安，而这种不安可以算得上是焦虑。

这位模拟护士必须对所处环境中发出的视觉信号作出适当回应。有些信号触发（或影响）的目标比其他目标更紧急：如果一个婴儿正在爬向水沟，而另一个婴儿只是饿了，那么“她”得先管爬向水沟的那个；此时如果刚好有一个婴儿快掉进水沟了，那么“她”的注意力得先转向这个。但是就算有些目标当时被搁下了，可能最终还是必须解决，它们的紧迫程度可能会随着时间的推移而不断增强。所以，如果有一个婴儿在水沟附近，那么“她”可以先把饿了的婴儿放回婴儿床；但是“她”应该先给喂食等待时间最长的婴儿喂吃的，然后再喂不久前刚喂过的婴儿。

总之，模拟护士的任务有时可能被中断，要么被放弃，要么被搁置。MINDER程序必须决定当前的优先级。“她”在完成任务的整个过程中必须做这种决定，这样“她”的行为可能会因此而被不断改变。事实上，任何任务在完成过程中都会被中断，因为环境（婴儿们）对系统提出了很多相互冲突且不断变化的要求。模拟护士和人类护士一样，会因为婴儿（每个婴儿是不可预测的自主智能体）数量不断增加

而变得越来越焦虑，表现得也越来越差。不过这种焦虑很有用，护士因此能成功地照顾婴儿。不过这个过程并不顺利：冷静和焦虑相距甚远。

MINDER程序表明了一些情感控制行为的方式，从而智能地安排相互竞争的动机。毫无疑问，人类护士会因为情况发生变化而经历（原文如此）各种焦虑。但这里的重点是情感（emotions），不只是感觉（feelings）。前者还涉及现象和功能意识（见第6章）。具体来说，它们是被安排了竞争动机的计算机制，如果没有这些机制，我们就无法运作。所以，影视剧《星际迷航》中没有情感的斯波克（Spock）先生就无法进化成真正的人。

如果要实现强人工智能，那么我们必须考虑和利用情感因素，如焦虑。

---

#### 注释

[1] 恐怖谷理论（uncanny valley）是一个关于人类对机器人和非人类物体的感觉的假设，它在1969年被提出，说明了当机器人与人类相像超过一定程度的时候，人类对它们的反应便会突然变得极为反感，即哪怕机器人与人类有一点点的差别都会显得非常显眼刺目，从而对整个机器人有非常僵硬恐怖的感觉，犹如面对行尸走肉。——译者注

## 04 人工神经网络

人工神经网络（artificial neural networks, ANN）是由许多相互连接的单元组成，每个单元只能计算一件事情。这样的描述听起来有些无聊。但人工神经网络似乎很有魔力。它必然也让记者们着迷。弗兰克·罗森布拉特的“感知器”（光电机）在没有接受明确指导的情况下可以学会识别字母，曾在20世纪60年代成为各大报纸的“宠儿”，吸睛无数。20世纪80年代中期，人工神经网络名声大振，至今仍然备受媒体的青睐。最近与人工神经网络相关的大量宣传还包括深度学习。

人工神经网络有无数应用，从操控股票市场和监测货币波动到识别语音或人脸，但真正有趣的是它们的运行方式。

有一小部分人工神经网络在特定的并行硬件上运行，甚至在硬件或湿件混合物上运行，将真正的神经元与硅电路结合。然而，大多数网络通常由约翰·冯·诺依曼机器来模拟。也就是说，人工神经网络是在经典计算机上实现的并行处理虚拟机（见第1章）。

之所以说它们的运行方式很有趣，部分原因是它们与符号人工智能的虚拟机有很大差别。大规模的并行计算代替串行指令，自下而上的处理代替自上而下的控制，以及概率代替逻辑。动态和持续变化的人工神经网络与符号程序形成了鲜明对比。

此外，许多神经网络从随机开始时就具有神秘的自组织属性（20世纪60年代的感知器也具有这一属性，它们在新闻中都很高调）。系统从随机架构（随机权重和联结）开始，并自己逐渐适应去执行需要完成的任务。

人工神经网络有许多优点，显著增强了人工智能的计算能力。然而，它们也有缺陷，即它们不能提供第2章中所设想的真正意义上的强人工智能。例如，虽然一些人工神经网络可以做近似推理或推理，但它们不能像符号人工智能那样精确（问：2+2是多少？答：很可能

是4。真的吗？）。在人工神经网络中，也很难模拟层级。一些（反馈式）网络在一定程度上能够用交互式网络来表示层级。

由于当前对深度学习的热情高涨，神经网络的网络不像以前那样罕见了。但是，这些网络还是相对比较简单。人脑必须包括无数在很多层级上以极其复杂的方式进行交互的网络。总之，强人工智能仍然十分遥远。

## 人工神经网络更广泛的含义

人工神经网络是作为计算机科学的人工智能取得的一大胜利，但它们的理论含义并未仅局限于此。它们与人类概念和记忆有一些相似之处，因此引发了神经科学家、心理学家和哲学家们的兴趣。

神经科学家们的兴趣由来已久。事实上，罗森布拉特没有把具有开创意义的感知器当作一个在现实中有用的发明，而是把它当作一条神经心理学理论。今天的神经网络尽管与大脑有很多差异，但是在计算神经科学中扮演着重要的角色。

心理学家也对人工神经网络感兴趣，哲学家们紧跟其后。例如，非专业人工智能人士狂热追捧20世纪80年代中期的一个神经网络。该网络显然已经像小孩一样学会了使用过去时，开始的时候没有犯错误，后来由于过分遵守规则，以至于把英文“go”的过去时变成了“goed”（本来应该是“went”），在犯了这些错误之后，最后才得到规则和不规则动词的正确形式。这是有可能做到的，因为提供给该网络的输入反映出小孩经常听到的词的变形概率——神经网络没有使用先天的语法规则。

这个网络的出现意义重大，因为当时大多数心理学家（和许多哲学家）都接受了诺姆·乔姆斯基（Noam Chomsky）的说法，声称小孩必须依靠先天的语言规则来学习语法，以及婴幼儿过分遵守规则的行为恰好证明那些规则在起作用。过去时态的神经网络证明了这两种说法都不正确（当然，它没有证明小孩不具备先天的规则，只是证明了他们不需要这些规则）。

另一个十分有趣的例子是对“表征轨迹”（representational trajectories）的研究，最初受到了发展心理学的启发。原先混乱的输入数据在连续的层级上被重新编码（在深度学习中也一样），所以除了能捕获到明显的规则性外，还有不太明显的规则性显露出来。这不仅涉及儿童的发展，还涉及与归纳学习相关的心理学和哲学争论。它表明有了先前的期望（计算结构），才能学习输入数据中的模式，学习不同模式的顺序必然受到约束。

简而言之，人工神经网络在商业和理论层面都很重要。



## 分布式并行处理

有一种人工神经网络理论尤其吸睛——PDP。事实上，当人们提及“神经网络”或“联结主义”（不常用的术语）时，他们通常说的就是PDP。

PDP网络的运行方式主要有四大优势，涉及技术应用和理论心理学（以及精神哲学）。

第一，它们只需要被显示例子而不需要被精确编程，就能学习模式以及各个模式之间的关联。

第二，能容忍“凌乱的”迹象，可以解决约束满足问题，弄清部分冲突迹象的意思。它们不需要严格定义（不需要被表示成一系列的充分必要条件）。相反，它们处理具有家族相似性的重叠集合，这也是人类概念的一个特征。

第三，能够识别不完整的或部分损坏的模式。也就是说，它们具有找寻内容的记忆。人类也有这种记忆，例如，听到开头几个音符就能识别一段旋律，或者即使是一段旋律几个音弹错了，也能识别这段旋律。

第四，它们很强健。PDP网络就算丢失一些节点，也不会满口胡言或停止运行。它可能显示出适度的退化，在这个过程中，它的性能随着损害的增加而逐渐变差。它们不像符号程序那么脆弱。

这些优势源于PDP中的D，即“分布式”。不是所有的人工神经网络都涉及分布式处理。在集中式（中央控制式）网络（例如WordNet，见第2章）中，所有概念分别是由唯一的节点表示。在分布式网络中，一个概念存储（分布）在整个系统中。有时候，集中式和分布式处理会结合，但不常见。单纯的集中式网络也不常见，因为它们没有PDP的四大优势。

分布式网络本质上是集中式网络，因为它的每个单元都相当于单个微特征——例如，视野中某个位置的一小块颜色（不是太小，也不是太特别。可以证明一些粗调单元比许多精调单元更有效率）。但是与需要表示的概念相比，这些单元在更低级概念层上定义：PDP包含“子符号”计算。此外，每个单元可以是许多不同整体模式的一部

分，因此促成了许多不同的“意义”。

**PDP**系统有多种类型。所有类型都是由三层或更多层互通单元组成的，每个单元只能计算一个简单的东西。但这么多单元连在一起，那情况就完全不一样了。

无论何时，只要输入层中单元的微特征被呈现给网络，它就会被触发。当输出单元被与之连接的单元激活，它就会被触发，它的活动也会被传递给人类用户。位于中间层的隐藏单元与外界没有直接接触。有些隐藏单元是确定的：它们是否被触发只取决于与其连接的单元产生的影响。有些隐藏单元则是随机的：它们是否被触发，部分取决于某个概率分布。

各个连接也不一样。有些是正向反馈，将信号从低层传递到高层；有些则是反向的；还有一些是横向的，连接同一层内的各个单元；有些既有正向反馈，也有反向反馈，我们将在下文中提到。就像大脑突触一样，连接要么是兴奋，要么是抑制。它们的强度（权重）不同。权重表示为+1和-1之间的数字。兴奋（或抑制）连接的权重越高，接收信号的单元被触发的概率越高（或越低）。

**PDP**包含分布式表示，因为每个概念都是由整个网络的状态表示的。这似乎令人困惑，甚至自相矛盾。它当然和符号人工智能中定义的表示有很大差异。

只对技术或商业应用感兴趣的人不在乎这一点。如果他们只想解决实践中的一些突出问题，例如，单个网络怎样才能存储几个不同的概念或模式，那么他们很乐意先把问题留在那儿！

研究人工智能心理和哲学含义的人也问到了这个“突出问题”。答案就是一个**PDP**网络可能出现的所有状态千变万化，所以在这块或那块单元中，只有少数几种状态同时被激活。被激活的单元仅仅将活性扩散到一些其他单元。然而，那些“其他单元”各不相同：任意既定单元能够促成许多不同的活性模式（一般情况下，带有许多未激活单元的“稀疏”表示更有效）。系统将最终饱和，关联存储器的理论研究问题将是一定规模的网络原则上能够存储多少模式。

这些人不喜欢把问题留在那儿。他们不仅对“表示”本身的概念感兴趣，还热衷于讨论人类的心智/大脑是否确实包含内部表示。**PDP**的追随者认为**PDP**源于符号人工智能，迅速传播到心智哲学，驳斥了物理符号系统假说（见第6章）。

## 神经网络学习

大多数人工神经网络能够学习。这包含在权重中作出自适应改变，有时也包含在连接中作出自适应改变。通常，网络的解剖结构（单元的数量以及单元之间的连接）是固定的。如果是这样，学习只改变权重。但有的时候，学习或进化（见第5章）能够增加新的连接和修剪旧的连接。构造型神经网络将这一点做到了极致：开始时没有任何隐藏单元，随着学习的不断深入，它们增加了隐藏单元。

PDP网络有很多学习方式，涵盖了第2章中提到的所有学习类型：监督式学习、非监督式学习和强化学习。例如，在监督式学习中，某个类的一些例子输入到PDP网络以后，网络逐渐识别这个类——所有例子都不需要具有每一个“典型的”特征（输入数据可以是视觉图像、语言描述、数字集合等）。当一个例子输入给PDP网络的时候，一些输入单元对“输入数据”的微特征作出回应，同时活性扩散，直到网络稳定下来。然后将输出单元的所得状态与期望输出作比较（由人类用户辨别），权重随之变化（可能由反向传播引起），从而减少误差。在输入了许多略有差别的例子后，届时PDP网络将开发出典型的活性模式或“原型”，即使之前没有被输入这样的典型模式（就算现在输入一个受损的例子以及激活更少的相关输入单元，该模式也能够自动完成）。

大多数人工神经网络学习基于赫布理论，它由唐纳德·赫布（Donald Hebb）于1949年提出，该理论经常被总结为“一起激发的神经元连在一起”。赫布理论学习强化了经常使用的连接。当两个连在一起的神元同时被激活的时候，权重就会被调整，上述情况以后更有可能实现。

赫布用了两种方式来表示赫布理论，这两种方式既不精确也不等效。今天的人工智能研究人员用许多不同的方式来定义它，理论基础有可能是从物理学中得到的微分方程，也可能是贝叶斯概率理论。他们利用理论分析来比较和改进各种版本。PDP研究可能属于万恶的数学范畴。因此，有大量物理学和数学专业的优秀毕业生在金融机构工作，这也正是他们的都市同事们为何很少真正了解他们的系统在做什么的缘故。

鉴于PDP网络用赫布型学习定律来调整权重，那么它何时能够停止？答案不是它什么时候能达到完美状态（消除了所有的不一致），

而是它什么时候能实现最大程度的一致性。

例如，如果相关单元同时发出信号，表示通常不同时显示出来的两个微特征，这时就会出现“不一致的情况”。许多符号人工智能程序可以解决约束满足问题，通过消除路径上迹象之间的矛盾来获取解决方案。但这些程序不能容忍解决方案中出现不一致的情况。PDP系统则不同。前文提到了PDP的优势，即使不一致持续存在，它们也可以成功运行。它们的“解决方案”是给出网络在不一致性降至最低时而非消除时的总体状态。

可以借用热平衡概念来做到这一点。物理学中的能级用数字表示，PDP中的权重也一样。如果学习定律与物理定律相类似（而且如果隐含单元是随机的），同一个描述统计行为的波尔兹曼方程可以描述能级和权重的变化，也可以借用快速但均匀冷却金属的方法。退火过程为：先将固体充分加温至足够高，再让其慢慢冷却。PDP研究人员有时使用模拟退火算法，前几个平衡周期中的权重变化远大于后面周期中的权重变化。网络因此能够摆脱和之前相比已经实现总体一致性的情况（局部极小值）；但是如果系统被干扰，网络甚至可能达到更好的一致性（以及更稳定的平衡状态）。就像我们摇动一袋弹珠，如果想要强行倒出袋子内脊处的所有弹珠，那我们开始应该使劲甩这个袋子，但是越到后面用力应该越小。

用反向传播实现最大程度的一致性，其速度更快，使用范围更广。但是，无论采用哪种学习规则，整个网络（特别是输出单元）的平衡状态被看作有关概念的表示。

## 反向传播、大脑和深度学习

PDP的追随者们认为，人工神经网络比符号人工智能更接近大脑。PDP的设计灵感确实来自大脑，一些神经学科学家确实用它来模拟神经机能。然而，人工神经网络还是与大脑中的东西有很大差异。

（大多数）人工神经网络和大脑之间的一个区别是反向传播或BP。BP是一个学习规则，或者说是一类通用学习规则，经常用于PDP。1974年，保罗·韦伯斯（Paul Werbos）首次使用BP，20世纪80年代初，杰弗里·辛顿（Geoffrey Hinton）以更有效的方式定义BP，它解决了信用分配问题。

所有类型的人工智能都面临信用分配问题，系统不断变化的时候尤其如此。假定有一个复杂的人工智能系统很成功，那么它的哪些部分最有可能促成了它的成功？在进化人工智能中，信用通常由“桶队列”算法分配（见第5章）。在具有确定（非随机的）单元的PDP系统中，信用通常由BP分配。

BP算法从输出层到隐藏层追溯系统成功的原因，并辨识一些需要被调整的个体单元（更新权重以将预测误差降至最低）。当网络给出正确答案时，BP算法需要知道输出层的确切状态（因此BP是监督式学习）。在该示范性输出和从网络中实际获得的输出之间进行逐个单元比较。输出单元的状态在两种情况下的任何差异都被视为误差。

算法假定，一个输出单元的误差由与其连接的单元中所出现的误差引起。算法在整个系统中进行反向推算，依据隐藏层中的单元与输出单元之间的连接权重，确定第一个隐藏层中的每个单元所产生的具体误差量。对于上述输出单元的误差，由与该单元连接的所有隐藏单元分摊责任（如果一个隐藏单元连接到多个输出单元，那就把它的所有单个责任相加，作为该隐藏单元的责任）。然后对隐藏层和前一层之间的连接按比例改变权重。

“前一层”可以是另一个（和另一个……）隐藏层，但最终它会成为输入层，权重将不再变化。该过程一直重复，直到输出层的误差降至最低。

多年来，BP算法仅用于只有一个隐藏层的网络。多层网络很罕见，它们很难分析，甚至很难做实验。不过最近因为出现了深度学

习，BP算法被用在了多层网络上，并引起了人们的巨大兴奋，还有一些不负责任的炒作。这里的系统学习深入到一个域的结构，而不只是表面模式。换句话说，它发现了多层而非单层的知识表示。

深度学习令人兴奋的原因是它至少为人工神经网络处理层级保驾护航。自20世纪80年代初以来，辛顿和杰夫·埃尔曼（Jeff Elman）等联结主义者为了表示层级——曾将集中式表示和分布式表示结合，还定义过递归神经网络（递归网络实际上是执行一系列离散步骤。利用深度学习，最新版本的递归网络有时能够预测句子中的下一个单词，甚至是段落中的下一个“想法”）。但是他们取得的成就有限（人工神经网络仍然不适合表示精确定义的层级或演绎推理）。

深度学习也出现在了20世纪80年代〔由于尔根·施密德胡贝（Jürgen Schmidhuber）发起〕。但是，直到最近辛顿提供了一种让多层网络在多个层级上发现关系的有效方法后，这个领域才有了进一步发展。他的深度学习系统由在六个层级上“受限制的”玻尔兹曼机（没有横向连接）组成。首先，所有层级进行非监督式学习。它们利用模拟退火算法逐一接受训练。一层的输出用作下一层的输入。当最后一层稳定以后，整个系统由BP微调，向下通过所有层级，为它们适当分配信用。

认知神经科学家和人工智能技术人员一样对深度学习感兴趣。因为它规定的“生成模型”能够学习预测网络输入的（可能会发生的）原因——模拟了亥姆霍兹在1867年提出的观点：“知觉为无意识推论。”也就是说，知觉不是指被动地接收来自感官的输入，它包含主动理解，甚至提前预测。简而言之，眼睛/大脑不是相机。

辛顿于2013年加入谷歌公司，所以BP也会频繁亮相。谷歌公司已经在许多应用中用到了深度学习，包括语音识别和图像处理。此外，它于2014年收购英国人工智能公司DeepMind。DeepMind公司的DQN算法结合深度学习和强化学习（见第2章），征服了经典的Atari游戏。IBM公司也对深度学习颇感兴趣：它不仅用在了沃森上，还被许多专家应用程序借用（见第3章）。

然而，虽然深度学习很实用，但这并不代表对它的理解就很到位。大量实验探索了不同的多层学习规则，但理论分析仍十分混乱。

其中一个问题就是是否有足够的深度来获得近乎人类的表现（第2章中提到的猫的脸部单元是由一个九层系统产生的）。例如，人类视觉系统有7个解剖层，但是大脑皮层中的计算到底增加了多少层？因为人工神经网络受到了大脑的启发（在深度学习炒作中不断强调的



点），所以这个问题必然会被问到，但它貌似不是很切题。

BP算法是计算层面取得的一项胜利，但不是生物学上的突破。BP算法过程不会产生大脑中猫的脸部“祖母细胞”（见第2章），但深度学习可以。真正的突触纯粹是前馈：它们不进行双向传播。大脑包含不同方向的反馈连接，但每个连接都是严格的单向。这只是真实神经网络和人工神经网络之间的一个差异（还有很多）。另一个差异是真实的神经网络不是按照严格的层级构成，即使视觉系统经常按照严格的层级来描述。

大脑包含正/反向连接的事实对于构建感觉运动控制的预测编码模型至关重要，这引起了神经科学专家们的极大兴趣（这些也主要基于辛顿的成果）。较高的神经层级向下层发送消息，预测来自传感器的输入信号，而且向上层发送的只有不可预测的“错误”消息。此类重复循环微调这些预测网络，使后者逐渐学会该预测什么。研究人员提到了“贝叶斯大脑”，因为按照贝叶斯统计，预测可以被理解，而且计算机模型中的预测根基实际上就是贝叶斯统计（见第2章）。

与大脑相比，人工神经网络的结构过于规整简单，层级太少，且枯燥无聊。过于规整，是因为人类建立的网络优先考虑数学层面的美观和功能，而生物进化的大脑则完全不同；过于简单，是因为单个神经元——有大约30种不同类型，计算起来和一个PDP系统或一台小型计算机一样复杂；层级太少，是因为即使具有数百万单位的单元，人工神经网络与人类大脑相比也只能是小巫见大巫（见第7章）；枯燥无味，是因为人工神经网络研究人员通常不但忽略时序因素（如神经脉冲频率和同步性），还忽略树突棘、神经调节剂、突触电流和离子通道这些生物物理学因素。

这四个缺点都在不断完善。由于计算能力增强，人工神经网络能够包含更多的单个单元。更为详细具体的单个神经元模型正在构建当中，已经在解决上述所有神经学因素的计算功能问题。“枯燥无味”在模拟中甚至在现实中都在改善（一些“神经形态”研究将活的神经元与硅芯片结合）。与DQN算法模拟视觉皮层和海马睡眠期间的过程一样（参见第2章），未来的人工神经网络无疑将借用神经科学的其他功能。

然而，仍不可否认的是，人工神经网络和大脑在很多重要方面千差万别，有些差异甚至我们现在还不知道。

## 网络丑闻

人们为PDP的到来感到兴奋的主要原因是，20年前的人工神经网络研究（也称为联结主义）被宣布为“死胡同”。正如第1章所述，这一判断是伴随着马文·明斯基和西摩尔·派普特（Seymour Papert）在20世纪60年代发表的尖锐批评文章而来的（两位专家在人工智能领域都德高望重）。到了20世纪80年代，人工神经网络似乎不只是一个“死胡同”，而是真的“断了气”。的确，控制论基本上已经被边缘化（见第1章）。几乎所有的研究资金都流向了符号人工智能。

一些早期的人工神经网络看似前途一片大好。罗森布拉特的自组织感知机一直被记者们惦记着，能够学着识别模式，即使它们开始学习的时候是一种随机状态。罗森布拉特满怀信心，声称自己的方法有涵盖人类心理各个方面的潜力。当然，他也指出了一些局限，但是他提出的“收敛定理”已经确保简单的感知机能够学做任何可能（用程序）指令它们去做的事情。这是强有力的证据！

但是马文·明斯基和西摩尔·派普特在20世纪60年代末也提供了证据。他们从数学层面证明简单的感知机并不能做人们直观上希望它们做的事情，而GOF AI可以轻松做到。和罗森布拉特的收敛定理一样，他们的证明只适用于单层网络。但是他们二位的“直观判断”是，多层系统将被组合爆炸击败。换句话说，感知机不会按比例增加。

大多数人工智能研究人员接受了联结主义注定失败的观点。不过还是有一些人继续研究人工神经网络。事实上，他们在分析联想记忆上取得了一些非常显著的成就，例如克里斯托弗·龙格-希金斯

（Christopher Longuet-Higgins）和大卫·威尔肖（David Willshaw），后来还有詹姆斯·安德森（James Anderson）、图沃·科霍恩（Teuvo Kohonen）和约翰·霍普菲尔德（John Hopfield）。但是这项工作并没有向世人公开。相关研究人员并没有将自己称为“人工智能”研究人员，而那些自称为人工智能研究人员的人并没把他们放在眼里。

PDP的到来给这种怀疑态度迎头一击。除了一些令人印象深刻的功能模型（如过去时学习程序），还有两个新的收敛定理：一个保证了基于玻尔兹曼方程<sup>[1]</sup>的PDP系统能够达到平衡状态（尽管可能是在很长一段时间之后）；另一个证明了三层网络原则上能够解决表示给它的任何问题（友情提示：和在符号人工智能中的情况一样，以可以

输入到计算机的方式表示问题通常是最难的部分）。人们对人工神经网络的热情自然随之而来，而对主流人工智能达成的共识也因此被扰乱。

符号人工智能曾经假定毫不费力的直觉思维正如有意识的推论，但其实没有意识。现在，PDP研究人员认为二者完全不一样。带头研究PDP的专家们[大卫·鲁姆哈特（David Rumelhart）、杰·麦克利兰（Jay McClelland）、唐纳德·诺曼（Donald Norman）和辛顿]指出，二者对人类心理学都至关重要。但是对PDP的宣传以及公众的反应，意味着研究心智的符号人工智能是在浪费时间。人工神经网络这个不起眼的研究这次来了个彻底大翻身。

美国国防部（人工智能的主要资助者）对此的态度也产生了360度的大转弯。在1988年召开紧急会议之后，他们承认自己以前对人工神经网络的忽视是不应该的。至此，大笔资金源源不断地投入到了PDP研究中。

不过明斯基和派普特仍然顽固不化。他们认可“基于网络的学习机器（原文如此）以后能带来的益处超乎想象”。然而，他们坚持认为高级智能不能从纯粹的随机性或完全无序的系统中产生。因此，大脑有时必须充当串行处理器，人类水平的人工智能必须采用混合系统。他们抗议称“他们的批评不是导致人工神经网络被冷落的唯一因素”，首要原因是计算能力不足。他们认为自己没有一直试图将研究资金转移到符号人工智能。他们说：“我们认为我们的工作没有杀死白雪公主，我们认为这是理解她的一种方式。”

这听起来像是令人信服的科学论证。但他们最初发表的批评文章的确尖酸刻薄（初稿甚至更毒：友好的同事劝他们语气缓和一点，多突出一下科学论点）。这篇文章让人有想法当然不足为奇。坚持不懈的人工神经网络研究人员极其反感他们新发现的文化缺位（支持符号人工智能的人放弃了人工神经网络）。PDP甚至引起了更大的躁动。人工神经网络的“死亡”和复兴包含嫉妒、自负、自我吹嘘和幸灾乐祸：“我们早就告诉过你了。”

这一小节提到了一个典型的科学丑闻，但不是人工智能中出现的唯一例子。理论分歧卷入了个人情感和较量，公正的思维十分罕见。到处都弥漫着尖酸刻薄的辱骂和打压。人工智能研究不是缺乏激情的事业。

[1] 非热力学平衡状态的热力学系统统计行为的偏微分方程。——译者注

## 连接不是一切

对于人工神经网络的大多数描述都暗含这样一层意思：与神经网络相关的唯一重要的事情就是它的解剖结构。哪些神经元与哪些神经元相关联，以及权重有多强？这些问题当然十分重要。然而，最近的神经科学表明，由于化学物质在大脑中扩散，生物学回路有时能够改变神经元的计算功能（不仅仅是让它基本形成）。

例如，一氧化氮向各个方向扩散，它的效果取决于相关点的浓度会一直存在，直到它下降（下降速率由酶改变）。因此，一氧化氮作用于给定体积的脑皮层内的所有细胞，无论这些细胞是否由突触连接。神经系统的功能动态表现与“纯粹的”人工神经网络有很大差别，因为“广播”信号取代了点对点信号。一氧化碳、硫化氢和复杂分子（如5-羟色胺和多巴胺）有类似效果。

人工智能怀疑者可能会说：“对人工神经网络的讨论应该就此打住！”“计算机里面没有化学组成！”这种说法很荒谬，你不能说计算机不能模拟天气的理由是因为计算机里面不能下雨吧。然后这些怀疑者又可能说：“人工智能无法模拟情绪或情感。”这个反对意见是由心理学家乌尔里克·奈瑟尔（Ulric Neisser）在20世纪60年代早期提出的。几年后，哲学家约翰·哈格尔（John Haugel）发表了一篇颇具影响力的“认知主义”批评文章，也提出了反对意见。他们说人工智能可以模拟推理，但是绝不能模拟情感。

然而，一些人工智能研究人员受到这些神经科学研究结果的启发，设计了一种全新的人工神经网络，在这个人工神经网络中，连接不是一切。在GasNets技术中，散布在网络中的一些节点能够释放模拟的“气体”。这些气体可以扩散，并根据浓度以不同方法调节其他节点和连接的固有属性。和扩散源的形状（模拟成空心球，而不是点源）一样，扩散体积的大小很重要。因此，给定节点在不同的时间表现也不一样。在确定的气体条件下，尽管没有直接连接，一个节点还是会影响另一个节点。真正关键的是气体与系统内电气连接之间的相互作用。同时，由于气体仅在某些特定场合下发出，而且以不同的速率扩散和衰减，所以这种复杂的相互作用在不断变化。

例如，GasNet技术用来进化自主机器人的“大脑”。研究人员发现，一个特定的行为可能包含两个未连接的子网，由于调节作用，二者能一起工作。他们还发现，将纸板三角形用作导航辅助的“方向检

测器”能够以部分不连接的子网形式进化。他们曾建立了一个整体连接的网络来做到这一点（见第5章），但神经调节网络进化得更快、更高效。

因此，一些人工神经网络研究者不再只考虑解剖结构（连接），还开始研究神经化学。他们现在在模拟不同的学习规则及其短暂的相互作用时，还会考虑神经调节。

神经调节是一种模拟现象，不是数字现象。扩散分子的浓度要一直变化。越来越多的人工智能研究人员（使用特殊的VLSI芯片）正在设计能同时体现模拟功能和数字功能的网络。模拟功能模拟生物神经元的解剖结构和生理机能，包括穿过细胞膜的离子通道。例如，这种“神经形态”计算被用于模拟知觉和运动控制的内部结构。有些人工智能研究人员计划在“全脑”模拟中使用神经形态计算（见第7章）。

有些专家甚至走得更远，他们不是纯粹在硅上建立人工神经网络模型，而是构建（或进化，见第5章）由微型电极和真实神经元组成的网络。例如，当电极X和Y都接受人工刺激时，在“湿”网络中产生的活动导致一些其他电极Z激发——这样就执行了一个“与”门。这类计算（唐纳德·麦凯曾在20世纪40年代设想过）现在处于萌芽阶段，但它可能会给大家带来惊喜。



## 混合系统

刚刚提到的模拟/数字网络和硬件/湿件网络被描述为“混合”系统，这可以理解。但这个术语通常用来指既包含符号又包含联结主义信息处理的人工智能程序。

明斯基在其早期宣言中已经说过，可能有必要包含二者，一些早期的符号程序确实结合了串行处理和并行处理。但这种尝试很罕见。我们也看到在PDP到来后，明斯基继续推崇符号和人工神经网络的混合系统。然而，这样的系统并没有立即出现（尽管辛顿建立了集中式和分布式联结主义相结合的网络来表示部分/整体层级，如家庭树）。

的确，符号处理和神经网络处理结合的系统仍然不常见。一个重逻辑，一个重概率，这两种方法差异巨大，大多数研究人员只擅长一种。

然而，一些真正意义上的混合系统已经被开发出来了，其中的控制在符号模块和PDP模块之间适度传递。因此，这种模型吸收了两种方法的互补优势，例如DeepMind开发的博弈算法（见第2章）。它们将深度学习与GOFAI结合，以学习如何玩一套视觉上多样化的电脑游戏。它们使用强化学习，没有提供手工制定的规则，只有每个步骤的输入像素和数值分数。许多可能的规则/计划被同时考虑，最有希望的规则/计划决定下一个动作〔未来的版本将专注于3D游戏，如《我的世界》（Minecraft），以及应用程序，如无人驾驶汽车〕。再如全心智系统ACT-R\*和CLARION（见第2章）以及LIDA（见第6章）。它们大多涉及认知心理学研究，开发是基于科学目的，而不是技术目的。

有些混合模型考虑神经系统的特定内部结构。例如在1980年，临床神经病学专家提摩西·沙丽斯（Timothy Shallice）与PDP先驱诺曼共同提出了一个与常见（“过度学习”）动作相关的混合理论，并得到了实施。该理论解释了一些常见的错误。例如，中风患者常常忘记应该先把信件放入信封再舔粘舌，或者他们可能在上楼换衣服时却去睡觉了，又或是拿起水壶而不是茶壶。我们所有人偶尔都会犯与顺序、捕获和对象替换相关的类似错误。

但为什么脑损伤患者特别容易犯这类错误？沙丽斯的计算理论声

称，常见动作由两种控制产生，能够在特定点上分解或接管：一是无意识的“争用调度”，它包含以层级形式组成的动作模式之间的（无意识）竞争，激活超过某个阈值的动作模式接管控制；另一个（“执行”）控制机制是有意识的，它包含审慎监督和调整第一种机制，包括规划和修复错误。对于沙丽斯而言，由PDP模拟争用调度，执行控制则交由符号人工智能。

动作模式的激活水平能够通过感知输入提高。例如，某人到达卧室时无意瞥了一眼床（模式识别），便可触发其上床的动作模式，即使其最初的意图（计划）是换衣服。

沙丽斯借鉴人工智能（特别是规划模型）的观点提出了动作理论，这与他的临床经验相吻合。大脑扫描得到的证据进一步佐证了他的理论。最近的神经科学发现了其他因素，包括与人类动作相关的神经递质。这些因素已经通过当前基于动作理论的计算机模型显现出来。

争用调度和执行控制之间的相互作用也与机器人学相关。遵循计划的智能体应该能够基于自身在环境中观察到的东西停止或改变计划。这个策略体现了机器人的特征，即结合情境处理和审慎处理（见第5章）。

任何对强人工智能感兴趣的人都应该注意到，有极少数人工智能科学家将心智的计算架构看作一个整体，他们毫无保留地接受了混合主义。如艾伦·纽厄尔和安德森（第2章讨论的SOAR和ACT\*）、斯坦·富兰克林（Stan Franklin，第6章中概述的LIDA的意识模型）、明斯基（心智的“社会”理论）和亚伦·斯洛曼（第3章中描述了他对焦虑的模拟）。

总之，在我们的大脑中实现的虚拟机既是串行的，也是并行的。人类智力需要二者的巧妙合作。假如能实现人类水平的强人工智能，那么它也会如此。

## 05 机器人和人工生命

人工生命模拟生物系统。和人工智能一样，它有着技术和科学双重目的。人工生命对于人工智能而言不可或缺，因为已知的所有智能都可以在生物体上找到。的确，有许多人相信心智只能由生命产生（见第6章）。冷静的技术人员不担心这个问题。不过他们确实在开发多种实际应用的时候考虑了生物学。这些应用包括机器人、进化编程和自组织设备。机器人是人工智能的经典例子：它们的出镜率很高，而且具有独创性，还具有很大的商机。尽管进化人工智能应用很广，但没有那么出名。知道自组织机器的人甚至更少（非监督学习除外，见第4章）。然而，在理解自组织的过程中，生物学对人工智能来说很有用，同样，人工智能对于生物学来说也很有用。

## 情境机器人和有趣的昆虫

机器人造于几世纪以前，列奥纳多·达·芬奇是一位制造专家。人工智能机器人首次亮相于20世纪50年代。第二次世界大战后，威廉·格雷·华特的机器“乌龟”因知道避开障碍并寻找光而艳惊四座。麻省理工新建了人工智能实验室，其主要目标是整合计算机视觉、规划、语言和运动控制技术，以打造一款“麻省理工机器人”。

后来的发展可谓突飞猛进。现在，有些机器人能够爬山、爬楼或爬墙；有些跑得快；有些跳得高；有些能够搬运重物；有些能够投掷重物。还有些能够自我拆解和自我重新组装，有时还能够组装成新的形状，如一条蠕虫（能够穿过一根狭窄的管道）或一个球，又或是多腿生物（适合于水平或粗糙地面）。取得这些进步的原因是人们将研究重点从心理学转到了生物学上。

经典人工智能机器人模仿人类的自发动作。根据一些大脑模型理论，经典人工智能机器人采用了世界和智能体自身动作的内部表示。它们没有给人们留下深刻的印象，因为它们依赖于抽象规划，常常碰到框架问题（见第2章）。它们不能做到及时反应，因为即使是一丝环境变化也需要它们预先规划并重新开始；它们也不能适应新（未模拟的）情况。即使在平坦整洁的地面上，平稳运动对它们来说也是难事（因此SRI机器人的昵称为SHAKY，它与“SHAKY”谐音，后者的意思是“摇晃”），而且机器人一旦跌倒后就无法恢复正常。在大多数建筑物中，它们都属于无用之物，那么就更别指望它们去火星上做事了。

而今天的机器人却焕然一新。焦点也已经从人类转变为昆虫。昆虫可能不够智能，不能模拟世界或做规划，但它们可以管理。它们的行为合乎时宜且适应性强，对，此处就是指行为而非动作。但这主要是一种反射（习惯性思维）而不是熟思。昆虫们不假思索地对当下情境作出回应，而不是想象中可能发生的一些事情或目标状态。因此，这些机器人得到了这样的标签——“情境”或“基于行为”的机器人（情境行为不仅限于昆虫，社会心理学家在人类身上也发现了许多与情境有关的行为）。

为了赋予人工智能机器类似的反射，机器人专家们喜欢工程学胜过编程。如果可能，感觉运动反射在机器人的解剖结构中是实实在在地被具化了，而不是以软件代码的形式。

机器人的解剖结构到底要和生物体的解剖结构有多匹配，至今尚无定论。就技术目的而言，巧妙的工程学技巧可以接受。今天的机器人有许多不现实的噱头。但也许是生物学机制的效率特别高？的确，它们足够高效。因此，机器人学专家还会考虑真正的动物：它们能做什么（包括它们的各种导航策略），涉及什么样的感官信号和具体运动，什么样的神经机制在起作用。生物学家反过来又用模型去研究这些神经机制——一个名为计算神经行为学的研究领域。

一个例子是兰德尔·比尔（Randall Beer）的柔性机器人学。蟑螂有六条带活关节的腿，这既是优势也是劣势。六足动物运动的时候比双足动物更稳定（通常比轮子更管用）。然而，协调六肢似乎要比协调两肢更困难。它不仅要决定接下来该移动哪条腿，还必须用正确的力量，找到正确的位置和时机。各条腿应该如何互动？它们必须非常独立，因为可能只有一条腿旁边有一个卵石，但是如果那条腿被抬高，其他腿必须做补偿动作以保持平衡。

比尔的机器人反映了真实蟑螂的神经解剖结构和感觉运动控制。它们能够爬楼梯，在粗糙的地面上行走，翻过障碍物（而不只是避开它们），并在摔倒后恢复原有姿势。

芭芭拉·韦伯（Barbara Webb）则在蟋蟀身上找灵感。她的关注点不是移动（所以她的机器人能够使用轮子）。相反，她希望自己的机器人可以识别、定位和靠近特定声音模式。显然，这种行为（“趋声性”）应该会有很多实际应用。

雌性蟋蟀一听到同种雄性蟋蟀唱的歌就会有趋声行为。然而，蟋蟀只能识别用一个速度和频率唱出来的一首歌曲。蟋蟀的种类决定速度和频率。雌性蟋蟀不会在不同的歌曲之间做选择，因为它没有为一系列声音编码的特征检测器，而使用只对一个频率敏感的机制。该机制不是神经机制，它类似于人类大脑内的听觉检测器。它是一根长度固定的管，长在胸部，连接到前腿上的耳朵和气孔。管的长度与雄性唱出来的歌曲的波长成精确比例。物理学能够确保以下两点：1.（管中的空气和外部空气之间的）相位抵消，仅在歌曲频率正确时才出现；2.强度差完全取决于声源的方向。雌性蟋蟀的神经系统强制它做出朝着这个方向移动的行为——雄性唱歌，雌性就跟着走。这就是货真价实的情境行为。

韦伯之所以选择研究蟋蟀的趋声性，是因为它得到了神经学专家的密切关注。但仍有许多问题尚未找到答案：歌曲的方向和声音是不是（以及如何）独立处理的；歌曲的识别和定位是不是独立的；是如何触发雌性蟋蟀行走的；以及如何控制它的“之”字形方向。韦伯还设

计了能够产生类似行为的最简单机制（只有四个神经元）。之后，她的模型用到了更多的神经元（基于详细的真实生命的数据），包括额外的神经特征（例如延迟、放电速率和膜电位），并且将听觉与视力结合。她的工作不仅澄清了许多神经科学问题，回答了一些问题，还提出了更多问题。所以她的工作对机器人学和生物学都有帮助。

虽然机器人是物理实体，但许多机器人学的研究都是在模拟中完成的。例如，比尔的机器人有时先在软件中进化，然后才被制造出来。同样，韦伯的机器人先被设计成程序，然后才在现实世界接受测试。

虽然主流机器人学的研究转向了昆虫，但是对人形机器人的研究仍在继续。有些只是玩具。有些是家用“社交”或“伴侣”机器人，供老年人或残疾人使用（见第3章）。专家们设计这些机器人的目的不只是让它们成为拿送东西的奴隶，而是成为独立的私人助理。有些机器人长相“可爱”，有长长的睫毛和诱人的声音，能够和用户进行眼神交流，并能识别对方的面部和声音。此外，它们在一定程度上可以独立进行未提前设定的对话，解读用户的情感状态，并产生“带有情感的”回应（类似于人的面部表情或语音模式）。

虽然有些机器人体型庞大（用来搬运重物或穿过粗糙地面），但是大多数很娇小。例如一些在血管内使用的机器人就是微型的。通常大量这种机器人一起工作。问题是只要有一个任务由多个机器人完成，那么就可能会出现问題，比如它们之间如何沟通；如何让团队能够完成个体不能单独完成的任务。

为了回答这些问题，机器人专家通常会考虑群居的昆虫，比如蚂蚁和蜜蜂。这些物种是“分布式认知”（见第2章）的典范，它们的知识（与合适的动作）分散于整个群体，而不是由任何一个动物独享。

如果机器人太过简单，那么它们的开发者就有可能会谈到“群体智能”，还会分析作为细胞自动机（CA）的协作机器人系统。细胞自动机是由多个个体单元构成的系统，单个单元会遵循简单规则，以采用有限状态中的一种状态，这就取决于其相邻单元的当前状态。一个细胞自动机的行为的整体模式可能超级复杂，这就好比多细胞生物体中的活细胞之间的相互合作。许多人工智能机器人用到了在好莱坞动画片中成群蝙蝠或恐龙使用的畜群算法。

分布式认知和群体智能的概念也适用于人类。如果参与的个体不能处理相关知识，那么就会用到群体智能（例如大批人群的整体行为）；如果参与的个体可以拥有所有相关知识但实际上却没有拥有，



那么就会用到分布式认知。例如，一个人类学家已经展示了导航知识是如何在船员之间共享的，同时是如何具化到实物上的，如图表和海图桌（的位置）。

说知识是具化到实物上的东西，听起来可能很奇怪，或者最多也就用了隐喻的手法。但是，如今有大把人声称，人类的心智已经被具化到了人类的具体动作中了，而且还被具化到人类用来吸引外部世界的文化艺术品中。这种“外部的/具化的智能”理论在一定程度上是基于麻省理工的罗德尼·布鲁克斯（Rodney Brooks）所做的工作，他是将机器人学的研究焦点从“人转向昆虫”的第一人。

布鲁克斯现在是美国军事机器人的主要开发者。在20世纪80年代，他还是一个初出茅庐的机器人专家，当他看到当时一些符号人工智能模拟世界的不切实际的规划程序时，他感到十分沮丧。所以，他后来单纯为了技术目的而改变了自己的研究方向，着手研究情境机器人学。不久之后，他就将自己的方法发展成了一个适应性行为理论。他的研究范围远远超出了昆虫。他认为，即使是人类动作也不包含内部表示（他有时暗示，不经常包含表示）。

他对符号人工智能的批判引起了心理学家和哲学家的兴趣。有些人和他产生了极大的共鸣。心理学家指出，很多人类行为受情境制约，例如在不同社会环境中的角色扮演。认知心理学家强调有生命的视觉，那么对于视觉而言，智能体的身体运动是关键。如今，具化的心智理论在人工智能之外的领域颇具影响力（见第6章）。

但是大卫·基尔希（David Kirsh）等人仍然坚决反对布鲁克斯的观点，他们认为复合表示对于涉及概念的行为来说必不可少。比如认识感知恒定性的时候，可以从不同角度识别目标；一段时间之后重新识别个体；预想的自我控制（规划）；协商而不只是调度相互冲突的动机；反事实推论；以及语言。这些批评者们承认，情境机器人学表明了不涉及概念的行为要比许多哲学家所想的更普遍。然而，逻辑、语言和认真考虑过的人类动作都需要符号计算。

许多机器人专家也反对布鲁克斯的一些更为极端的论述。研究足球机器人的阿兰·迈科沃斯（Alan Mackworth）团队提到了“反应式的深思熟虑”，它包括感官知觉、制定实时决策、规划、识别规划、学习和协调。他们试图实现GOF AI和情境观点的一体化（也就是说，他们正在构建混合系统，详见第4章。）

一般来说，表示对于机器人学中选择动作的过程至关重要，但对于动作的执行不太重要。所以，如果有人戏称“人工智能”现在代

表“人工昆虫”，那么这种说法并不完全正确。

## 进化人工智能

大多数人认为，人工智能需要一丝不苟的设计。鉴于计算机的无情天性，要是没有一丝不苟的设计，它怎能这样？其实，它能够这样。

例如，进化机器人（包括一些情境机器人）是通过组合严格的编程/工程学和随机变异而产生的结果。它们的进化未经预测，也没有精心的设计。

进化人工智能通常有这样一个特点：它从符号人工智能中产生，但也用于联结主义。它有许多实际应用，包括艺术（该领域可能欢迎不可预测性）和开发对安全苛求的系统（例如飞机发动机）。

程序能够自我改变（而不是由程序员重写），甚至可以用遗传算法（GA）完善自身。受真实生命遗传学的启发，程序能够随机变异和非随机选择。选择需要成功的标准或“适应度函数”（与生物学中的自然选择类似）与遗传算法一起工作。于是，定义适应度函数成了关键。

在进化软件中，面向任务的初始程序既不能有效完成任务，也可能根本无法完成该任务，因为这个任务可能是一个不连贯的指令集或随机连接的神经网络。但是，整个程序包括后台的遗传算法。

遗传算法能够改变面向任务的规则。这些随机变化与生物学中的点突变和交叉类似。因此，程序指令中的单个符号可能被改变，或者两个指令的短符号序列可能被“交换”。

任何一代中的任务程序都会被比较，用那些最成功的任务程序繁殖下一代。少数（随机选择的）其他程序也可能被保留，这样就不会丢失那些尚未具有任何良好效果的潜在有用突变。经过一代又一代的比较选择，任务程序的效率不断提高。有时能在这个过程中找到最优解。在一些进化系统中，信用分配问题用约翰·霍兰德（John Holland）的“桶队”算法的一些变形算法来解决，以辨识让这个复杂的进化程序成功的是它的哪些部分，见第4章。

一些进化人工智能是完全自动的：程序在每一代都使用适应度函数，并且在无监督情况下进化。必须清楚定义这里的任务，如用飞机

发动机物理学来定义。相比之下，进化艺术的交互性通常很强（艺术家在每一代筛选出最好的“任务程序”），因为其不能清楚阐明适应度函数——美学标准。

大多数进化机器人学是间歇性交互。机器人的解剖结构〔例如传感器和感觉运动连接或它的控制器（“大脑”）〕自动进化，但是这个过程是在模拟中进行的。大多数进化都没有用到物理机器人。但是在进化到第500代的时候，进化的设计会在物理设备上进行测试。

无用的突变往往不能存活。萨塞克斯大学的研究人员发现，如果任务不需要深度视觉或触觉，机器人的一只“眼睛”（共两只）和所有的“晶须”都可能会失去其与控制神经网络的初始连接（同样，无论是先天性聋人，还是被剥夺听觉输入的动物，他们的听觉皮层都用于视觉计算：大脑终身都在进化，而不只是进化几代）。

进化人工智能还有可能带来更多的惊喜。例如，一个不断进化的情境机器人（也在萨塞克斯）在靠近目标的过程中能够做出回避障碍物的动作，它“长出”了一个定位检测器，和大脑中的定位检测器相类似。机器人的世界要包括一个白色纸板三角形。让人出乎意料的是，一个随机连接的微型网络将出现在控制器中，该控制器对在某一特定方向渐变的亮/暗（三角形的一边）作出回应。然后这个微型网络进化成视觉动作机制中不可分割的一部分，机器人利用它（最初随机）与动作单元的连接，将纸板三角形用作导航辅助。动作机制对黑色三角形和三角形的另一边都不适用。它单独存在，没有全面的定位检测器系统，不过它很有用。总体来说，这个结果可以重复。萨塞克斯团队使用不同类型的神经网络后发现，每个成功的解决方案都进化了某种主动定位检测器。因此，高级行为策略也一样（确切的实施细节可能会有所不同，但往往非常相似）。

萨塞克斯团队还用遗传算法设计硬件电路。任务是进化振荡器。然而，最后出现的却是一个原始的无线电波传感器，从附近的计算机拾取背景信号。这取决于不可预测的物理参数。有些参数可以预测（所有印刷的电路板类似空气的性质），虽然该团队以前忽略了这一点。其他参数却是偶然的，而且看起来不相关，例如，与计算机的空间距离；模拟开关的设置顺序；以及留在附近工作台上的烙铁插入电源的事实（此结果不会重复，无线电天线下次可能受到墙纸化学反应的影响）。

无线电波传感器很有趣。要知道许多生物学家（和哲学家）认为人工智能不会出现全新的东西，理由是计算机程序的所有结果（包括遗传算法的随机结果）必须属于它所定义的可能性空间。他们说只有

生物进化能够产生新的知觉传感器。他们承认，一个无效的人工智能视觉传感器可以进化得更好。但是，他们说，第一个视觉传感器只会出现在因果关系控制的物理世界中出现。产生光敏化学物质的随机遗传突变，可以将外部世界已经存在的光带入到生物体的环境中。然而，这个出人意料的无线电传感器，同样会将无线电波带入设备的“环境”中。从一定程度上来讲，它的确取决于物理原因（插头等）。但它是人工智能的一次练习，而不是生物学。

人工智能中要有全新的东西出现确实需要外部影响，因为不可否认的是，程序不能超出其可能性空间。但这些影响不一定是物理的。连接到互联网的遗传算法系统可能通过与虚拟世界交互，进化出彻底底底的新东西。

进化人工智能另一个更早的惊喜仍然激励着进行中的进化研究。生物学家托马斯·雷（Thomas Ray）用遗传算法模拟热带雨林的生态。他看到了寄生虫的自然涌现、对寄生虫的抵抗和能够克服这种抵抗的超级寄生虫。他还发现，一系列微小的基础（基因型）突变可以造成（表型）进化中的突然“剧增”。正统的进化论者当然相信这一点。但是，该观点违反常理，所以史蒂芬·杰·古尔德（Stephen Jay Gould）等生物学家认为，必须考虑非进化论的过程。

今天，模拟突变率仍处在系统地变化和追踪阶段，遗传算法研究人员正在分析“适应值曲面”“神经（原文如此）网络”和“遗传漂变”。在突变尚未增加生殖适应度的情况下，如何维持突变？这项工作作了解释。总之，人工智能在帮助生物学家提出进化理论。

## 自组织

生物有机体的关键特征是它们具有组织自我的能力。自组织是指从不太有序的起源到有序结构的自发出现。这种性质让人不解，甚至自相矛盾。它能否在非生物上发生，这一点尚不明显。

广义上说，自组织是一种创造性现象。我们在第3章中讨论了创造力（包括“历史”和“个体”），在第4章中讨论了自组织的（非监督）联想学习。这里，我们将重点讨论生物学研究的自组织类型。

这方面的例子有动植物种类史的进化（历史创造力）；胚胎发育和变态（与心理学的个体创造力类似）；脑发育（先是个体创造力，然后是历史创造力）；细胞形成（生命开始时为历史创造力，然后是个体创造力）。人工智能如何帮助我们理解它们？

艾伦·图灵通过追本溯源解释了自组织。他问道：“同质事物（如未分化的卵子）为何会创造结构？”他承认，大多数生物发育为现有顺序增加了新顺序，例如胚胎神经管中变化的顺序。但是从同质性中得到的顺序是基本（数学上最简单的）情况。

胚胎学家也设想了“组织者”：以未知方式指导发育的未知化学物质。图灵不认同组织者，而是考虑和化学扩散相关的超级通用原则。他指出，如果不同的分子相遇，结果将取决于它们的扩散速率、浓度，以及它们之间的相互作用对分子的破坏/构建速度。他的证明方法是改变设想的化学方程中的数字并研究结果。有些数字组合只产生形体不明的化学品混合物。但其他数字组合则产生顺序，例如，某一分子浓度的常规峰值。他说，这样的化学峰值可能在生物学上被表示为表面标记（条纹），或重复结构的起源，如花瓣或身体的部分。三维中的扩散反应能够产生空心化现象，如早期胚胎中的原肠胚形成。

这些想法立刻引起了巨大的反响。它们解决了以前很难解决的难题，即无序的起源如何产生顺序。但是20世纪50年代的生物学家用它们还做不了什么。图灵用的是数学分析。他确实手写了一些极其乏味的模型，还在原始计算机上建模。但是他的机器计算能力不够，无法得到相关总和，也无法系统地探索数字变化。而当时也没有计算机图形可用，无法将数字转换为明显可以理解的形式。

人工智能和生物学界足足等了四十年才等到图灵的深刻见解。计



计算机图形学专家格雷·图尔克（Greg Turk）研究了图灵的方程，有时先“冻结”一个方程式的结果，然后再用另一个方程式。这个步骤让人联想到基因的开/关切换，是有关“模式中的模式（pattern-from-pattern）”的例子——图灵也提到过，但无法分析。在图尔克的人工智能模型中，图灵的方程式不仅产生了斑点犬的标记和条纹（如他的手写模型所做的），还产生了美洲豹的斑点、猎豹的斑点、长颈鹿的网状花斑和狮子鱼的图案。

其他研究人员利用更复杂的方程序列，得到了更复杂的模式。这其中包括部分当今对实际的生物化学更了解的发育生物学专家。

例如，布赖恩·古德温（Brain Goodwin）研究了伞藻（藻类）的生命周期。这种单细胞生物体从一个无形状的斑点长成一根细长的茎；然后，它会长出一个扁平的顶；接下来，围绕顶的边缘长出一圈疙瘩；而这些疙瘩后来发芽变成一轮侧根或分支；最后，侧根合并形成一个伞形帽。生化实验表明，所涉及的代谢参数多达30多个（例如钙浓度、钙和某些蛋白质之间的亲和性，以及细胞骨架的力学阻力）。古德温的伞藻计算机模型模拟了复杂的重复反馈过程，其中的参数时刻发生着变化。各种身体的蜕变随之产生。

与图灵和图尔克一样，古德温也反复琢磨数值，想看看哪些数值会真正产生新的形式。他只使用生物体中可观察范围内的数字，但这些数字都是随机的。

他发现，某些模式会反复出现，例如茎末梢处的钙的高/低浓度的交替变换（一个轮形成的对称性）。它们不依赖于某个具体参数值，但只要参数值被设置在一个大范围内，它们就会自发出现。此外，轮一旦成形就不会消失。所以，古德温说，模式可能是变形的基础，而变形会产生其他经常出现的特征。这可能发生在动植物种类史和个体发育中（历史创造力以及个体创造力），如在四足动物的肢体进化过程中。

古德温模型没有产生过伞形帽。可能是因为需要额外的参数来表示真正的伞藻内未知的化学作用，抑或是这些伞形帽确实就在模型的可能性空间内，所以原则上能够从模型中产生。但是只要数值被严格限定，那么随机搜索就可能找不到这些伞形帽（侧根也没有产生，但这只是因为计算能力不足。对于每个单独的侧根，整个程序需要在较低的层级上执行）。

古德温从中得到了一个理论教训。他把轮看作“通用”形态，在许多动植物中发生，这一点和伞状帽不一样。这表明，轮产生的原因不

是由偶然进化的基因引导的特殊生物化学机制，而是在大多数甚至是所有生物中发现的一般过程（如反应扩散）。这些过程可能构成“结构主义”生物学的基础——形态学的一种普通科学，它的解释与达尔文自然选择学说完全一致，但先于后者出现〔图灵的论述隐含了这种可能性，达西·汤普森（D'Arcy Thompson）也强调了这种可能性，但是图灵自己忽视了它〕。

反应扩散利用决定局部分子相互作用的物理化学规律起作用，即在细胞自动机中可表示的规律。约翰·冯·诺依曼在定义细胞自动机的时候指出，它们原则上适用于物理学。今天的人工生命研究人员将细胞自动机用于许多目的，这里的主要作用是它非常适合产生生物模式。例如，非常简单的细胞自动机只在一个维度（一条线）上定义，就可以产生非常逼真的模式，例如贝壳上的图案。

最有趣的可能是人工生命的尝试——即用细胞自动机描述“可能生命<sup>[1]</sup>”，而不仅仅是“我们所知道的生命”。克里斯托弗·兰顿（Christopher Langton，1987年命名“人工生命”）探索了无数随机定义的细胞自动机，关注它们产生顺序的倾向。许多细胞自动机只产生无秩序状态。另外一些则形成了无聊重复甚至静态的结构。但是有些产生了变化微妙但相对稳定的模式——兰顿说，这就是生物（也是计算）的特征。令人惊讶的是，在简单度量系统的信息复杂度时，这些细胞自动机拥有相同的数值。兰顿提出，这个“ $\lambda$ 参数”适用于所有可能的生物，无论它/她/他是在地球上还是在火星上。

自组织不仅塑造整个身体，还塑造器官。例如，大脑发育的方法是利用进化过程（在一生中和各代之间）和非监督式学习。这种学习可以带来非同寻常的结果（历史创造力）。但是每个个体的早期脑发育也可产生可预测的神经结构。例如，新生的猴子拥有全方位方向检测器。这些检测器不能从外部世界的经验中习得，所以我们自然会假定它们被编码在基因中，但它们没有。它们是从最初随机的网络中自发出现的。

神经科学家建立的仿真计算机模型和“纯粹的”人工智能都说明了这一点。IBM公司的研究员拉尔夫·林斯克（Ralph Linsker）定义了多层前馈网络（见第4章）。该网络表明，如果考虑随机活动（如胚胎大脑中的“噪声”），简单的赫布型学习规则能够产生有组织的方向检测器集合。

林斯克不仅依靠实际演示，也不只关注方向探测器，他的抽象“infomax”理论适用于这种类型（多层前馈网络）的任何网络。该

理论指出，当信号在每个处理阶段被变换时，网络连接会不断发展，以最大程度保存信息量。所有连接在某些经验约束下形成，例如生物化学和解剖学限制。然而，数学确保将会出现一个通信单元协作的系统。infomax理论还涉及系统发育进化。因此，“复杂的系统在进化时所出现的单个突变具有自适应性”这样的观点会更符合常理。如果每个层级都能够自发地适应另一个层级中的小变化，那么就不再需要几个同时发生的突变了。

对于细胞层的自组织而言，胞内生物化学和细胞/细胞壁的形成都已经被模拟。这项工作利用了图灵在反应扩散方面的研究成果。然而，它更多地依赖于生物学，而不是源于人工生命的概念。

总之，人工智能提供了许多与自组织相关的理论概念。自组织的人工制品比比皆是。

---

注释

[1] 人工生命所研究的人造系统能够演示具有自然生命系统特征的行为。

——译者注

## 06 强人工智能会有真正的智能吗

假设未来的强人工智能系统（银幕上或机器人）能够匹敌人类的表现，那么它们会有真正的智能、理解力和创造力吗？它们会有自我、道德身份和自由选择吗？它们会有意识吗？如果没有意识，它们会有任何其他属性吗？

这些显然不是科学问题，而是哲学问题。许多人直观上会认为，上述每种情况的答案都是“很显然嘛，不会！”

事情并不是非黑即白。我们要的是仔细论证，而不只是未经核实的直觉。这些论证表明：对于上述问题，没有任何无懈可击的答案。因为所涉及的概念本身就颇具争议性。只有透彻理解概念本身的含义，我们才有信心说假设的强人工智能将会或不会具备真正的智能。总之，没有人知道确切的答案。

有些人可能会说：“这没关系，强人工智能实际上做的事情才是关键。”然而，我们将看到，上述问题的答案可能会影响到我们处理强人工智能的方式。

本章不会给出明确的答案，但会谈谈哪些答案比其他答案更合理。同时也会介绍（一些）哲学家如何使用人工智能概念来阐明真实心智的本质。

## 图灵测试

艾伦·图灵在哲学杂志《思想》（Mind）上发表了一篇文章，描述了所谓的图灵测试，即测试者是否可以确定自己是在与计算机还是在和人类交互（交互时间最多五分钟），如果有超过30%的测试者不能确定被测试者是人还是机器，那么这台机器就通过了测试。电脑能够真正思考的说法也就因此站得住脚了。

图灵测试是以一种幽默的方式提出来的。虽然它出现在开头几页的位置，但主要目的是预言未来的人工智能，是整个论文的附属内容。图灵甚至在向朋友罗宾·甘迪（Robin Gandy）介绍图灵测试时，也将其描述为轻松愉快的“宣传”，朋友对此一笑置之，没有大肆评论。

然而，哲学家们倒是认真地讨论起了图灵测试。大多数人认为，即使程序的回应与人类的回应不可区分，也不能证明程序有智能。最常见的反对意见是（仍然是）图灵测试只关注看得见的行为，所以一具僵尸也可以通过测试：一个和我们有着一模一样的行为但缺乏意识的“怪物”。

该反对意见给出了两个假定前提：1.智能需要意识；2.僵尸从逻辑上说得通。我们将在“人工智能和现象意识”一节看到对意识的一些描述表明，僵尸的概念不合逻辑。如果这些描述正确的话，那么任何强人工智能都不会是一具僵尸。就这一点而言，图灵测试是合理的。

图灵测试极大地吸引了哲学家们（和公众）的兴趣，但它在人工智能中的地位却并不重要。大多数人工智能的目标是提供有用的工具，而不是模拟人类的智能，那就更不可能是为了让用户相信他们正在与人类交互了。诚然，那些高调的人工智能研究人员有时会自称或允许记者宣称他们的系统通过了图灵测试。然而，这些测试不符合图灵的描述。例如，肯·科尔比（Ken Colby）的PARRY模型“愚弄”了精神病医生，让他们认为自己正在阅读妄想狂的病例，因为他们很自然地认为自己正在和人类患者打交道。同样，如果没有提示可能涉及机器，计算机艺术通常也会被认为是人类所为。

和真正的图灵测试最接近的是一年一次的罗布纳奖（Loebner prize）比赛（现在在布莱切利园举行）。当前的规则规定：进行25分钟的互动，使用20个预设问题，目的是测试记忆、推理、常识和个

性。裁判综合考虑结果的相关性、正确性，以及表达/语法的清晰度和合理性。到目前为止，没有一个程序可以骗过30%以上的裁判（在2014年，一个形象被定位为13岁乌克兰男孩的程序让33%的裁判相信他们在和人对话。问题是，非英语母语人士犯的错误很容易被原谅，特别是儿童）。



## 意识的很多问题

没有“意识的问题”这样的东西。我们应该说“意识的很多问题”。“有意识的”一词被用来划清事物界限：清醒的/睡着的；故意的/不留心的；全神贯注的/走神的；易达到的/难达到的；值得报告的/不值得报告的；反思的/未核实的等。没有一个解释可以阐明上述所有问题。

上面所列出的对比属于功能性问题。原则上来说，可以用信息处理术语或神经科学术语来理解它们，对此，很多哲学家都会赞同。

但是，现象意识——感觉（如蓝色或疼痛）或“感受性”<sup>[1]</sup>（*qualia*，哲学家的技术术语）似乎不一样。在物质宇宙（基本上）中，感受性的存在是一个臭名昭著的形而上学问题。

大卫·查默斯（David Chalmers）称感受性的存在是一个“难题”，还认为它不可避免：“我们必须重视意识……不能接受将问题重新定义成解释某些认知或行为功能是如何执行的。”

对于问题的答案，已经有了各种猜想。如查默斯版的泛心论，他自己承认泛心论是一个“离谱的甚至疯狂的”理论，根据该理论，现象意识是宇宙的一个不能削减的属性，类似于质量或电荷。其他几位理论家则从量子物理着手，但在他们的对手看来，这些人只是用一个谜题来解决另外一个谜题。科林·麦克金（Colin McGinn）甚至认为，人类天生就不能理解大脑和感受性之间的因果关系，就像狗不能理解算术一样。认知科学的哲学翘楚杰瑞·福多（Jerry Fodor）认为：“说物质的东西有意识，人们对此一点概念都没有。就算是有概念，也不知道会是一种什么样的情况。”

简而言之，几乎没有哲学家声称自己了解现象意识，如果有人说自己已经了解了，也几乎没有人相信。这个话题一直是一个哲学难题。

---

注释

<sup>[1]</sup>指从实体中概括出来的可感受特性，如颜色、味道等。——译者注

## 机器意识

赞成人工智能的思想家们用两种方式研究意识：一种是建立意识的计算机模型，这叫作“机器意识”（Machine Consciousness，机器意识）；另一种（是受人工智能影响的哲学家的特点）是用计算术语分析意识，但不模拟。

一个真正智能的强人工智能具有功能意识。例如，它会在不同时间关注（留心、注意）不同的事物。人类水平的系统还能够进行深思熟虑和自我反思，它可以产生创造性的想法，甚至特意评估它们。如果没有这些能力，它就不能产生看似智能的表现。

人们在评估创造性的想法时，可能会涉及现象意识（见第3章）。事实上，许多人会说，只要有“功能上的”差异，就有现象意识出现。然而，所有机器意识研究者在考虑功能意识时，通常都会忽略现象意识（勇敢的，还是愚蠢的？人们声称自己的人工智能系统已经“以它的方式”有了现象意识，理由是它的辨别力基于感知输入，例如光。这是否意味着出现了视觉体验还十分值得怀疑）。

斯坦·富兰克林的团队在美国孟菲斯进行了一个有趣的机器意识项目，研究LIDA。这说明了两件事：一件是（功能上）意识的概念模型——口头表达的计算理论；另一件是对该概念理论模型的部分简化实现。

两者都用于科学目的（富兰克林的主要目标）。但后者也有实际应用。LIDA的实现能够自定义，以适应具体的问题领域，例如医学。

与SOAR、ACT-R和CYC不同（见第2章），LIDA是在最近才出现的。第一个版本（为美国海军制作，目的是为任务结束的水手们安排新的工作）于2011年发布。而当前版本能够模拟注意力，而且能够影响在各类记忆（情景、语义和程序）中的学习；感觉运动控制也用在机器人学中。但还有很多特征仍然缺失，例如语言（无论已经实现了哪些方面，下面的描述涉及概念模型）。

LIDA是一个混合系统，包括扩散活性和稀疏表示（见第4章），还有符号编程。它是基于伯纳德·巴尔斯（Bernard Baars）的意识的神经心理学全局工作空间理论（Global Workspace Theory，以下简称

GWT) 开发的。

GWT将大脑视为一个分布式系统（见第2章），其中并行运行的大量专用子系统竞争访问工作记忆（如图6—1所示）。各个意识项（完整的独立单元）按照顺序出现（意识流），但是“广播”到所有脑皮层区域。

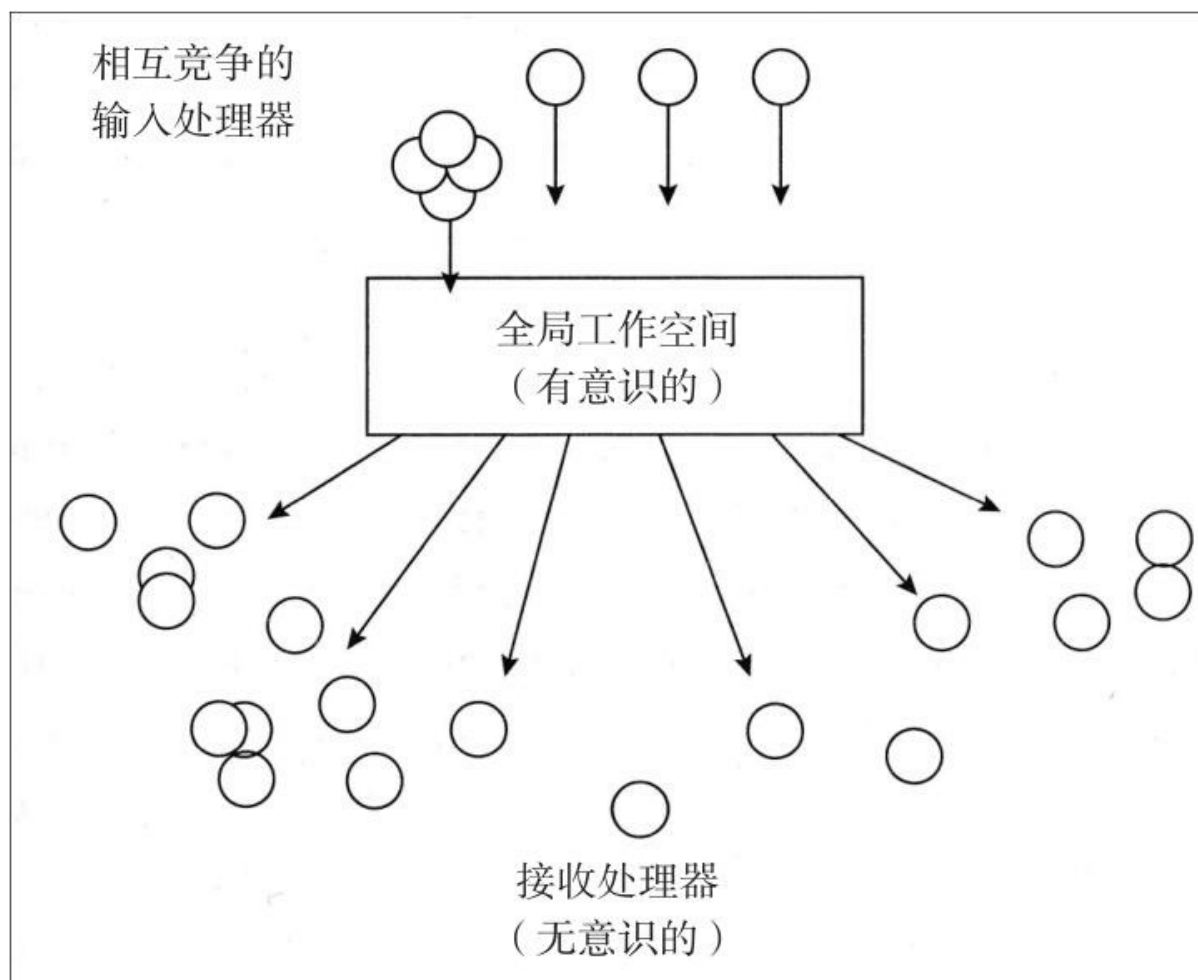


图6—1 分布式系统中的全局工作空间

图注：神经系统包含各种无意识的专家处理器（感知分析器、输出系统、规划系统等）。这些无意识的专家处理器之间如果要想实现互动、协调和控制，它们需要一个中央信息交流区或“全局工作空间”。输入专家之间可以合作并竞争访问全局工作空间。此处的案例显示，四个输入处理器合作放置一条全局消息，然后将其广播到整个系统。

资料来源：改编自伯纳德·巴尔斯的著作《意识的认知理论》（A Cognitive Theory of Consciousness）（剑桥大学出版社，1988）。

如果从感觉器官或其他子系统得到的广播意识项触发了某个区域并产生反应，则该反应在赢取注意力的竞争中稳操胜券，从而该区域主动控制了对意识的访问（新颖的知觉/表示易于获得注意力，而重复项则从意识中逐渐消失）。子系统通常很复杂。有些是分级嵌套的，许多具有不同种类的关联联系。各种无意识的背景（组织在不同的记忆中）形成有意识的经历，既唤起又修改全局工作空间中的意识项。注意力的内容反过来通过引起各类学习来适应持久的背景。

这些内容在广播时指导下一个动作的选择。许多动作属于认知型，即建立或修改内部表示。道德规范被存储起来（在语义记忆中），以作为评估潜在动作的步骤。决策也可能被其他社会智能体感知到/预测到的反应所影响。

想想规划的例子（见第2章）。意图被表示为近乎无意识但相对高级的结构，它能够产生有意识的目标图像（由当前从知觉、记忆或想象中得到的显著特征来挑选）。这些意图招募相关的子目标。它们是“招募”而不是“检索”，因为子目标自身决定其相关性。像所有大脑皮层的子系統一样，它们等待时机由某个广播项触发——这里的广播项就是合适的目标图像。LIDA能够将选定的受目标驱动的动作计划变成低级（计算机程序）可执行的微观运动行为，能够对不可预测且变化的环境的详细特征作出回应。

巴尔斯的理论（和富兰克林版的GWT）不是在计算机科学家的研讨会上凭空想出来的。相反，它的设计考虑到了各种众所周知的心理现象和大量的实验证据（如图6—2所示）。但是这些作者都声称，它还解决了一些以前未解决的心理谜题。例如，他们说GWT/LIDA解决了长期备受争议的“约束”问题，即来自不同大脑区域中不同感官的几个输入是如何来自同一物的，例如猫的感觉、外观、气味和声音。富兰克林和巴尔斯声称，它也解释了人类的心智是如何避免框架问题的（见第2章）。例如，在产生有创意的类比时，没有中央执行系统在整个数据结构中搜索相关项。当然，如果子系统能识别出某个广播项适合/接近（总是）其寻找的对象，那么它就会竞争进入全局工作空间。

富兰克林整合各种实验证据，用LIDA探索认知心理学和神经科学理论。例如，他模拟了“注意瞬脱”，在这期间，被试未能报告在第一个视觉目标出现之后很快出现的第二个视觉目标。另外还有其他注意瞬脱的理论和计算机模型，但大多数都是用来回答孤立的问题。富兰克林的模型来源于系统层统一的认知理论（另一个注意瞬脱的统一模型的存在基础是ACT-R，但ACT-R不包括情感处理或高级视觉，所以不能解释所有的实验结果）。

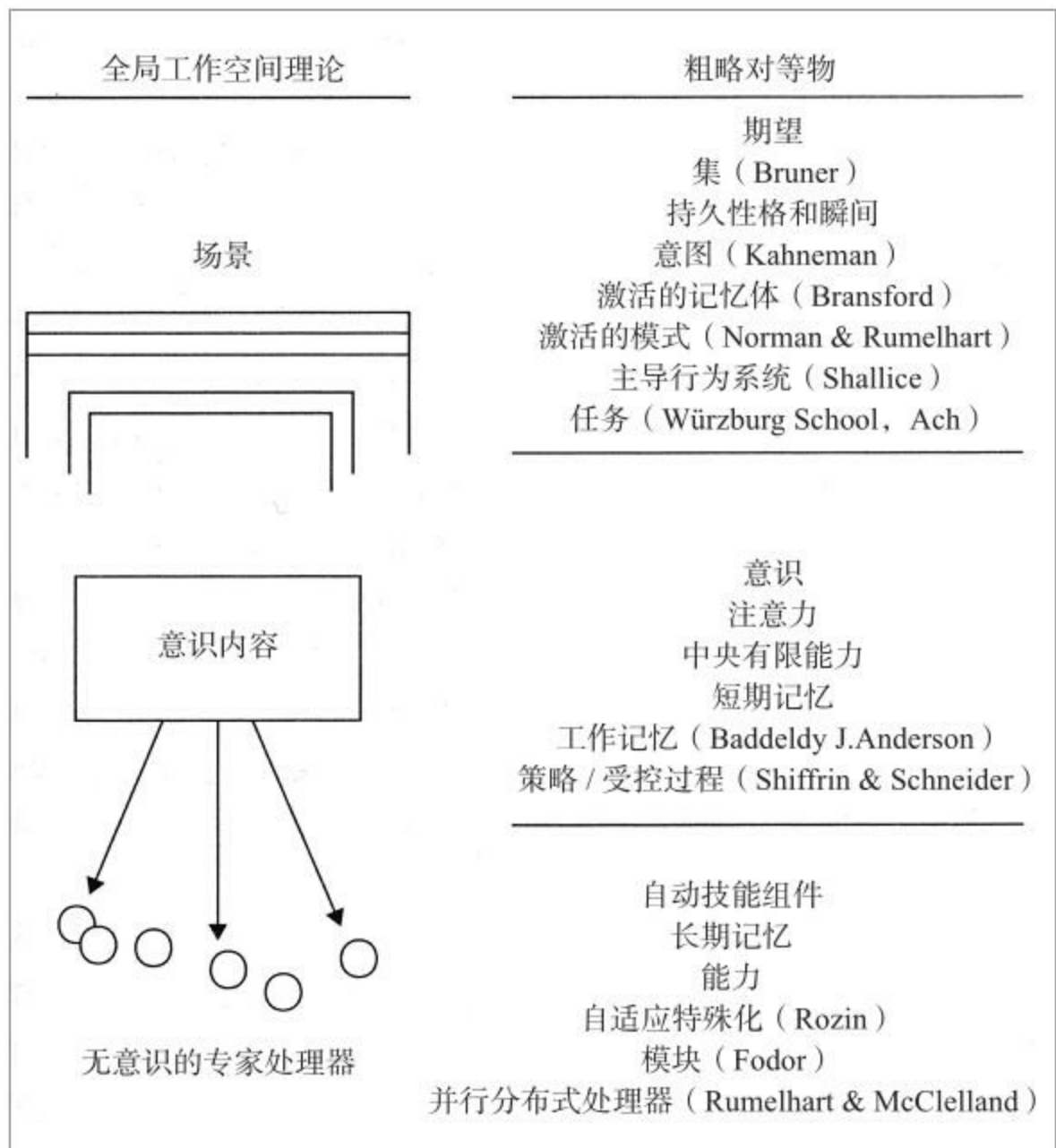


图6—2 全局工作空间术语和其他普遍概念之间的相似性

这种人工智能方法让人联想到伏魔宫中的“恶魔们”，以及用于实现产出系统的“黑板”架构（见第1章和第2章）。这并不足为奇，因为那些想法给了巴尔斯灵感，他提出了神经心理学理论，最终带来了LIDA。理论的巨轮已经回归原样。

## 人工智能和现象意识

机器意识的实践者忽略了现象意识这个“难”题。但三位受人工智能启发的哲学家迎难而上：保罗·丘奇兰德（Paul Churchland）、丹尼尔·丹尼特（Daniel Dennett）和亚伦·斯洛曼。只是说他们给出的答案有争议算是轻描淡写了。然而，就现象意识而言，存在争议也是意料之中的事。

丘奇兰德的“取消物理主义”否认存在非物质的思想和经验。相反，他把二者和脑状态等同起来。他给出一个科学理论——部分计算范畴（联结主义）、部分神经学范畴——定义了一个四维的“味觉空间”，将对味觉的感受性系统地映射到特定的神经结构上。四个维度反映了舌头上的四种味觉受体。

对丘奇兰德来说，这不是心智—大脑相互关系的问题，体验味觉只是让人的大脑访问那个被抽象定义的感官空间内的一个特定点。这意味着，所有的现象意识只是大脑在某个凭经验可发现的超空间内的某一特定位置的存在。如果是这样，那么计算机（可能除了全脑模拟）就不可能具有现象意识。

从本体论角度来看，丹尼特也认为不存在截然不同的体验，身体活动除外（所以他那本颇具煽动性的书得到了一个共同反应——“解释的不是意识，但搪塞过去了”）。

他认为，体验就是辨别。但是，人们在辨别物质世界中存在的东西时，不会让其他非物质世界中存在别的东西。他在一个虚构的对话中说明了这一点：

奥托（Otto）：在我看来，你已经否认了无疑是此处最真实的现象的存在，甚至连笛卡尔（Descartes）在其所著的《沉思集》<sup>[1]</sup>中都不能怀疑的真实表象。

丹尼特：就某种意义来说，你是对的，就是我否认的东西存在着。让我们想想霓虹色扩散现象。就好像防尘布套上有一个粉红色的发光环（他在描述一种视觉错觉，这是由闪亮白纸上的红色和黑色线条引起的）。

奥托：一定有。



丹尼特：但是没有任何粉红色的环。不完全是。

奥托：对。但一定看起来是！

丹尼特：对。

奥托：那么它在哪里呢？

丹尼特：什么在哪里？

奥托：粉红色的发光环。

丹尼特：没有什么发光环，我以为你刚刚承认了。

奥托：对，在那张纸上没有任何粉红色的发光环，但一定看起来是。

丹尼特：对。看起来是有一个粉红色的发光环。

奥托：所以我们谈谈那个发光环吧。

丹尼特：哪一个？

奥托：看起来像的那一个。

丹尼特：没有一个只是看起来是粉色光环这样的东西。

奥托：请注意，我并不仅仅说看起来有一个粉红色的发光环，而是确实看起来有一个粉红色的发光环！

丹尼特：我举双手赞同……当你说那里看起来有一个粉红色的发光环的时候，你说的是真的。

奥托：请注意。我要说的不止这些。我认为不仅看起来有一个粉红色的发光环，还是真的看起来有一个粉红色的发光环！

丹尼特：由于你这么做法，你和其他很多人一起掉入了一个陷阱。你似乎认为“思考（判断、决定、坚决主张）对你来说看起来是粉红色的东西”和“真的看起来是粉红色的东西”之间有区别。但是没有区别。没有真的看起来这样的现象——以某种方式判断事情是这样的现象除外。

换句话说，为感受性给出一种解释的要求无法得到满足。不存在这样的东西。

亚伦·斯洛曼不同意这个观点。他承认感受性的真实存在。但他以一种不寻常的方式做到了这一点：他在分析感受性的时候，将它们看作多维虚拟机的内部结构，此处的多维虚拟机也就是我们称为心智的东西（请见下面的章节）。

他说感受性是内部的计算状态。它们可以对行为（例如无意识的面部表情）或心智的信息处理的其他内部结构产生因果效应。它们只能存在于结构极其复杂的虚拟机当中（他概述了所需反射计算资源的类型）。它们只能由相关特定虚拟机的其他部分访问，并且不一定有任何行为表现（因此，这是它们的隐私）。此外，它们不能总是（由心智中更高的自我监控层级）用语言描述（因此它们难以形容）。

这并不代表斯洛曼将感受性等同于大脑（丘奇兰德将二者等同起来）。因为计算状态是虚拟机的内部结构：它们不能由物理描述的语言来定义。但是，它们只有在某个基本的物理机制实现时才能够存在且具有因果效应。

那图灵测试怎么样呢？丹尼特和斯洛曼的分析都表明（以及丹尼特明确指出）僵尸是不可能存在的。因为对他们来说，僵尸的概念是混乱的。如果给予适当的行为或虚拟机，对于斯洛曼来说，意识甚至包括感受性，都是有保证的。因此，如果还有人用僵尸能通过图灵测试这样的理由来反对图灵测试的话，那么这个反对是不成立的。

那假设的强人工智能又会如何？如果丹尼特是对的，它将包含我们所有的意识，而不包括感受性。如果斯洛曼是对的，它将具有和我们的感觉相同的现象意识。

---

#### 注释

[1] 此处《沉思集》应该为《第一哲学沉思集》，论证上帝的存在和灵魂的不灭，是法国哲学家勒内·笛卡尔所著的一本哲学论文选集，《第一哲学沉思集》全书由六个《沉思》（Meditation）、其他学者对这六个沉思的《反驳》（Objection）以及笛卡尔本人对《反驳》所作的《答辩》（Reply）组成。——译者注

## 虚拟机和身心问题

在20世纪60年代，希拉里·普特南（Hilary Putnam）提出了“功能主义”，他借鉴了图灵机的概念，还考虑了（当时很新颖）软件/硬件的区别，以证明心智实际上是大脑所做的事情。

身体和心智是两种截然不同的物质，它们之间（笛卡尔的）形而上学的分歧被描述层级之间的概念分歧取代。根据程序与计算机的类比，“心智”和“身体”的确迥然不同。但该类比与唯物主义完全一致（它是否能够包括感受性一直都备受争议）。

到1960年，虽然出现了几个博人眼球的人工智能程序（见第1章），但是功能主义哲学家很少考虑具体的例子。他们专注于一般原则，如图灵计算。只有到20世纪80年代中期PDP兴起（见第4章）时，许多哲学家才开始考虑人工智能系统的实际运作方式。即使是当时，也很少有人问，什么计算功能能实现（例如）推理或创造力。

理解这些事情的最好方法是借用计算机科学家的虚拟机概念。我们与其说心智是大脑做的事情，还不如说（跟随斯洛曼）心智是虚拟机，或者说是一套完整的不同虚拟机，它/它们在大脑中实现（但是“心智是虚拟机”这个观点有一个有悖于常理的含义，见“神经蛋白是必要条件吗”一节）。

如第1章中的解释，虚拟机是真实的，且具有真实的因果效应：没有形而上学神秘的身心相互作用。因此，LIDA的哲学意义是它规定了一套虚拟机，这能够表明（功能）意识的各个方面是如何实现的。

虚拟机方法完善了功能主义的核心部分——物理符号系统（Physical Symbol System，以下简称PSS）假设。在20世纪70年代，艾伦·纽厄尔和赫伯特·西蒙将PSS定义为“一组实体，它们被称作符号，它们是物理模式，能够作为另外一种称为表达式（或符号结构）实体的组成部分出现……在符号结构（中），……物理上相关的（例如一个符号在另一个符号旁边）符号（或记号）实例”。他们说，过程是为了创建和修改符号结构而存在的，也就是说符号人工智能的定义过程。他们还补充说：“一个PSS具有实现一般智能动作的充分必要条件。”换句话说，心智—大脑是一个PSS。

从心智是虚拟机的观点来看，他们应该将PSS称为物理实现的符号系统假设（还是别表示成首字母缩略词），因为符号是虚拟机的内容，而不是物理机的内容。

这意味着神经组织不是智能的必要条件，除非它是能够实现相关虚拟机的唯一材料基体。

PSS假说（和大多数早期人工智能）假定，表示或物理符号是机器/大脑的一个特征，明显可分离和恰好容易定位。联结主义对表示的描述不同（见第4章）。它将表示看作整个网络的细胞，而不是容易定位的神经元。它将概念看成部分冲突的约束，而不是严格的逻辑定义。熟悉路德维希·维特根斯坦（Ludwig Wittgenstein）如何描述“家庭相似性”的哲学家们对此十分感兴趣。

后来，情境机器人学的研究人员完全否认了大脑包含表示（见第5章）的观点。有些哲学家对此表示赞同，但是大卫·基尔希认为，涉及概念（包括逻辑、语言和审慎动作）的所有行为都需要组合表示（和符号计算）。

## 意义和理解力

根据纽厄尔和西蒙的观点，任何执行正确计算的PSS的确是智能的。它有“实现智能行动的充分和必要手段”。哲学家约翰·希尔勒将这种说法称为“强人工智能”（“弱人工智能”只是说人工智能模型可以帮助心理学家阐述连贯的理论）。

他认为强人工智能是错误的。符号计算可能继续在我们的大脑中起作用（虽然他怀疑这一点），但是单靠符号计算不能提供智能。更准确地说，它不能提供“意向性”——哲学家描述意义或理解力时使用的技术术语。

希尔勒的依据是一个今天仍然具有争议的思想实验：希尔勒被关在一间密闭的房间中，房间只有一个开口，用来传递上面只写有字迹潦草的中文问题和相关回复的字条。房间里有一个装满中文字卡片的盒子，还有一本规则书。希尔勒收到从屋外递入的中文问题后，便按照手册的说明，找到合适的指示，将相应的中文字符组合成对问题的解答，并将答案递出房间。希尔勒不认识字条上的问题，因为问题是中文写成的，而规则书是一个中文NLP处理程序，房间外说中文的人让希尔勒回答他们的问题。然而，希尔勒进入房间的时候不懂中文，而且他离开房间的时候仍然不懂中文。结论是：单凭形式化计算（这正是在房间里的希尔勒做的事）不能产生意向性，所以强人工智能是错误的，人工智能程序不可能有真正的理解力（“汉字屋”论证最初针对的是符号人工智能，但后来被推广到联结主义和机器人学）。

希尔勒在此宣称，人工智能程序产生的“意义”完全来自人类用户/程序员。就程序本身而言，这些意义是任意的，这在语义层面是没有意义的。同一个程序如果“只有句法，但没有语义”，那么该程序同样也可以被判断为税务负责人或是舞蹈艺术。

有时候他的这种说法是正确的。但是别忘了富兰克林的观点：LIDA模型用感官、执行器和环境之间的结构耦合来为认知打基础，甚至体塑认知。别忘了控制回路，它进化成了机器人的方位检测器（见第5章）。把它称为“方位检测器”并非随意行为。它能否成为方向检测器决定了它的存在，这有助于实现机器人的目标。

后一个例子很重要，尤其是一些哲学家认为进化是意向性的源泉。例如，露斯·米利肯（Ruth Millikan）认为思想和语言是生物现

象，它们的意义取决于我们的进化史。如果这是正确的，那么不能进化的强人工智能就不具备真正的理解力。

其他崇尚科学的哲学家（如纽厄尔和西蒙自己）用因果关系定义意向性。但他们很难描述虚假的陈述：如果有人声称看到一头牛，但那里没有牛可以产生词，那么他们怎么能说是牛呢？

总之，没有意向性的理论能让所有哲学家都满意。由于真正的智能包含理解力，所以这再次说明了没有人知道我们假设的强人工智能是否真的会智能。



## 神经蛋白是必要条件吗

希尔勒之所以排斥强人工智能，部分原因是计算机不是由神经蛋白构成的。他说，正如叶绿素是光合作用的场所，神经蛋白是意向性的温床。神经蛋白可能不是宇宙中能够支撑意向性和意识的唯一物质。但是他说金属和硅也肯定做不到这一点。

然而，他这么说有点过了。不可否认，我们要是说金属和硅制成的电脑可以真正感受忧郁、痛苦或理解语言，的确有悖常理。但是要是说神经蛋白能够产生感受性，这种说法仍旧有违直觉，在哲学上也一样存在诸多问题（所以，违背常理的东西不一定就是错的）。

如果我们接受斯洛曼对感受性展开的虚拟机分析，那么就可以打消和常理相悖这个难题。然而，心智作为虚拟机的整个描述提出了另一个相似的难题。如果一个心智达标的虚拟机能够在人工智能硬件中实现，那么该心智会存在于一台机器或许几台机器中。因此，心智作为虚拟机的观点表明，计算机的个人永生（克隆的）原则上是可能实现的。对于大多数人来说（见第7章），这和计算机能够支持感受性的说法一样不符合常理。

如果神经蛋白真的是能够支持人类水平虚拟机的唯一物质，那么我们可以反对“克隆永生”的说法。但它到底是不是？我们不得而知。

也许存在某种特殊或高度抽象的属性，一旦神经蛋白有了这种属性，就能实现心智执行的各种计算。例如，神经蛋白必须能够（相当快地）构建稳定（和可储存的）且柔性的分子，必须能够形成具有电化学属性的结构和结构之间的连接，这样各结构之间才能传递信息。也许，其他行星上的其他物质也可以做到这些事情。

## 不只是大脑，身体也很重要

一些研究心智的哲学家认为大脑得到了太多关注。他们说，整个身体是更好的焦点。

他们的主张经常借鉴欧陆哲学中的现象学观点，强调人类的“生命形式”。它包含有意义的意识（包括人类的“兴趣”，它是关联感的根基）和体塑化。

被体塑就是指成为一个动态环境中或与这个环境积极互动的生命体。环境和互动涉及物理层面和社会文化层面。关键的心理属性不是推理或思想，而是适应和沟通。

研究体塑化的哲学家们几乎没时间关注符号人工智能，认为它太过深奥。只有基于控制论的方法受到青睐（见第1章和第5章）。从这个观点来看，真正的智能基于身体，那么屏幕上的强人工智能不可能真的智能。即使屏幕上的系统是一个自主智能体，结构上能够耦合到一个物理环境中，然而这不能算是（让富兰克林控制速度）被体塑。

那么机器人呢？毕竟机器人是基于并适应现实世界的物理实体。事实上，情境机器人学有时会得到这些哲学家们的称赞。但是机器人有身体吗？有兴趣吗？有生命形式吗？它们究竟是不是活着呢？

现象学家们会说：“当然没有！”他们可能引用维特根斯坦的著名评论：“就算一头狮子可以说话，我们也不会理解它。”狮子的生命形式与我们的生命形式差异巨大，沟通几乎不可能。当然，狮子的心理和我们的心理（例如饥饿、恐惧、疲劳等）之间有诸多重叠部分，一些最低限度的理解和同情可能存在。但即便如此，在与机器人“交流”时，饥饿、恐惧、疲劳等心理都没有（所以对计算机伴侣的研究令人堪忧，见第3章和第7章）。

## 道德社区

人类会或者说人类应该接受人类水平的强人工智能成为道德社区的成员吗？如果接受了，这将有重要的实际意义。因为它将以三种方式影响人机交互。

第一，强人工智能将和动物一样得到人类的道德关注。人类将在一定程度上尊重它的利益。如果它要求某人中断休息或纵横字谜游戏，以帮助它达到“高优先级”目标，这个人会这样做（你是否曾从扶手椅上起来去遛狗，或者让瓢虫进入花园）。人类越是认为它的利益对它重要，就越觉得有必要尊重它。然而，这种判断很大程度上取决于人类是否认为强人工智能可以产生现象意识（包括感受到的情感）。

第二，人类将认为可以对强人工智能的行动进行道德评估。今天的杀手无人机无须承担道德责任（不像其用户/设计者，见第7章）。但也许真正智能的强人工智能会承担？想必它的决定可能会因为人类对它的反应而受到影响：如人类的赞美或责备。如果不受影响，就没有社区。它能够通过学习变得“有道德”，就像婴儿（或狗）能够学会听话，或者一个年长的孩子能够学会体贴（体贴需要发展认知心理学家口中的“心智理论”，它用能动性、意图和信念来解释人们的行为）。若以工具主义为评判标准，那么虐待甚至有可能算是合乎情理。

第三，人类将把它作为道德决策的论证和说服目标。它甚至可以为人们提供道德建议。在讨论这些话题的时候，我们要相信（它不仅具有人类水平的智能）强人工智能经得起特定道德标准的检验。但是那又能代表什么？伦理学家全力反对的不仅是道德的内容，还有它的哲学基础。

人们越多考虑“道德社区”的含义，就似乎越难承认强人工智能。事实上，大多数人直观上觉得强人工智能是痴人说梦。

## 道德、自由和自我

人们之所以有这种直觉，主要是因为道德责任的概念与其他概念有着密切关系，如有意识的能动性、自由和自我，而这些概念促成了人类概念的形成。

有意识的深思熟虑让我们的选择更加符合道德标准（尽管未经考虑的行动也会受到批评）。道德赞美或谴责来自智能体或自我。与自由的做法相比，在强大的约束下的行动不太容易受到谴责。

即使将这些概念用在人身上也颇具争议。将它们用在机器上貌似不合适，尤其是考虑到上一节谈到的对人机交互的影响。然而，我们可以结合实际情况，借鉴研究人类心智时所采用的“心智作为虚拟机”的方法，去理解这些现象。

从马文·明斯基开始，受人工智能影响的哲学家们用某类认知——动机的复杂性来分析自由。他们指出，在某些方式上，人们显然是“自由”的，而蟋蟀不是。雌蟋蟀通过硬连线反射反应找到自己的伴侣（见第5章），但是一位女性有很多策略寻找伴侣。除了交配以外，她还有许多其他动机，但并不是所有动机都可以同时得到满足。但是，她可以利用计算资源（又称智能）进行管理，而这一点正是蟋蟀所缺乏的。

这些资源由功能意识构成，包括感知学习、预期规划、默认分配、偏好排名、反事实推理，以及情感上可引导的行动计划。丹尼特确实用这些概念和许多生动示例来解释人类的自由。因此，人工智能有助于我们理解人类是如何做到自由选择的。

决定论/非决定论在很大程度上是一个与事实不相干的论点。在人类行动中有一些非决定因素，但因为这将损害道德责任，所以在做决定的时候，非决定因素不会发生。但是，它有可能会影响深思熟虑期间产生的考虑。智能体不一定会想到 $x$ ，或想到 $y$ ——其中的 $x$ 和 $y$ 既包括事实，也包括道德价值。例如，某人在选择生日礼物时，可能会因偶然注意到的一些事情，从而想到收礼物的人喜欢紫色或支持动物权益，那么这个人的选择也可能因此受到影响。

刚刚列出的所有计算资源都可用于人类水平的强人工智能。因此，除非自由选择必须包含现象意识（并且如果人们拒绝对此做计算

分析），否则我们想象的强人工智能似乎会有自由。强人工智能有各种对其有意义的动机，如果我们能够看懂这其中的深意，那么我们甚至可以区分它是“在自由条件下”还是“在约束条件下”做出的选择。然而，要做到这个“如果”无疑是一个超级大难题。

对于自我，人工智能研究人员强调递归计算的作用，其中每个过程可以操作自己。与自我认知（和自我欺骗）相关的许多传统哲学谜题能够用这个递归计算解决。

但是，“自我认知”是什么认知？一些哲学家否认自我的真实存在，但受人工智能影响的思想家不这么认为。他们将其视为特定类型的虚拟机。

对他们来说，自我是一个持久的计算结构，能够合理组织智体的行动，特别是它们认真考虑的自发行动（例如，LIDA的设计师将自我描述为“经历的持久背景，它能够组织和稳定许多不同局部背景下的经历”）。自我不会出现在新生儿身上，而是一个终身塑造的过程，某种程度上服从深思熟虑的自我塑造。它有很多维度，允许发生很大变化，能够产生可识别的个人能动性和个人特质。

它有可能实现，因为智能体的“心智理论”（最初用来解释其他智能体的行为）反射性地被用到某人自己的想法和行动上。“自我”看懂这些想法和行动的优先动机、意图和目标。反过来，这些想法和行为由持久的个人偏好、个人关系和道德/政治价值观构成。利用这种计算架构，可以构建自我形象（别人眼中的形象）和理想的自我形象（某人想要变成的形象），也可以得到基于二者差异存在的行动和情感。

丹尼特（受明斯基的极大影响）称自我是“叙事重心的中心”，即一种结构（虚拟机），它在讲述某人自己的生命故事时，会产生并寻求解释他们的行动，特别是与其他人的关系。这当然为出现无数类型的自我欺骗和自我隐形留有余地。

道格拉斯·霍夫斯塔特（Douglas Hofstadter，中文名为侯世达）同样将自我描述为抽象的自我参照模式，它们来源于并因果循环回到无意义的神经活动基础。这些模式（虚拟机）不是某人的表面特征。相反，自我的存在仅仅是为该模式被体塑化。

霍夫斯塔特补充说，一位备受爱戴的人在身体死亡之后仍然能够存在，这位“逝去的”人的自我先前全部被体塑在自己的大脑中，现在体塑在他们的活人的大脑中，只是现在的体塑没有那么细致。他坚持

认为，这不仅仅是“生活在”某人的记忆中，也不是活着的人已经吸纳了这位逝者的其他特征，例如对歌剧的热情，而是在逝者辞世前，两个自我深深相互渗透进对方的精神生活和个人理想，以至于二者确实都能够在对方身上“活着”。一位已逝母亲甚至可以通过她还活着的丈夫有意识地经历孩子的成长。这种有悖常理的说法假定，存在着个人永生之类的东西，即便因为所有活着的人都已经逝去，使得逝去的自我不再被体塑化。“超人类主义”哲学家预见计算机不朽的个人永生，详见第7章。

总之，如果决定相信强人工智能具有真正的人类智能，包括道德、自由和自我，那么这将是一个重大跨越，会有重要的实际意义。如果有人直观上认为这个想法存在根本性错误，那么他们很可能是正确的。不幸的是，这种直觉没有无争议的哲学论证来支撑。目前在这些问题上还未达成共识时，要想得到答案并不容易。



## 心智和生命

我们知道的所有心智都能在生物体上找到。包括控制论者（见第1章和第5章）在内的许多人都相信肯定是这样的。也就是说，他们认为心智必须以生命为前提。

内行的哲学家有时明确地指出了这一点，但几乎没有论证。例如，普特南说“如果机器人不是活的，那它就不会有意识”是一个“确定的事实”。但他没有给出科学论证，而是依靠“语言的语义规则”。甚至是少数最终捍卫这一假设的人也未能证明它是毋庸置疑的，如环境哲学家汉斯·乔纳斯（Hans Jonas），还有用到“自由能量原则”的物理学家卡尔·弗里斯顿（Karl Friston，他的这一原则总体上算控制论）。

让我们假定这个普遍的信念是正确的。若如此，那么只有实现了真实生命，人工智能才能实现真正的智能。然后，我们必须问，“强人工生命”（网络空间中的生命）是否可能实现。

生命没有公认定义。但通常提到了9个特征：自组织、自主、涌现、生长、适应、应激性、繁殖、进化和新陈代谢。前8个可以理解成信息处理术语，因此原则上可以用人工智能人工生命体塑化。例如，包括所有其他特征的自组织（广义上理解）已经以多种方式实现（见第4章和第5章）。

新陈代谢不同。计算机能够模拟它，但不能把它体塑化。自组装的机器人和虚拟（屏幕上）人工生命都不能新陈代谢。新陈代谢是指使用生化物质和能量交换来组合和维持生物体，所以它符合不可缩减的自然法则。强人工生命的捍卫者指出，计算机使用能量，还有一些机器人具有单独的能量存储器，需要定期补充能量。但是，这与灵活利用相互关联的生化循环来构建生物体的身体结构相去甚远。

因此，如果新陈代谢是生命存在的必要条件，那么强人工生命不可能实现。如果生命是心智的必要条件，那么强人工智能也不可能实现。不管未来强人工智能的表现如何令人印象深刻，它都真的可能不会有智慧。

## 巨大的哲学分歧

“分析型”哲学家和人工智能研究人员想当然以为某种科学的心理状态可能会实现。事实上，这一立场贯穿本书，包括本章。

然而，现象学家持反对意见。他们认为，我们所有的科学概念都来自有意义的意识，所以不能用来解释意识（普特南本人现在接受这个立场）。他们甚至声称，假定存在一个独立于人类思想的真实世界且认为科学可能发现这个世界的客观属性，那这一假设是荒谬的。

因此，有关心智/智能的性质所产生的分歧甚至比我目前所说的还要深。

不管是反对还是支持现象学家的观点，都没有压倒性的论据。因为论据的基础不一样，也就是说，每一方都为自己辩护而批评另一方，但对于各自论据中所使用的关键术语，彼此都不赞同。对于一些基本概念，分析哲学和现象学哲学甚至都各自给出了根本不同的解释，如理性和真理 [人工智能科学家布雷恩·坎特韦尔·史密斯（Brain Cantwell Smith）提出了一个大胆的形而上学，有关计算、意向性和对象，他希望尊重双方的见解；不幸的是，他的这一有趣论据毫无说服力。]

这个争论依然没有解决，也许无法解决。对某些人来说，现象学家的立场“显然”正确。但对其他人来说，它“显然”很荒唐。因此，人们依然无法确切地知道强人工智能是否能真的是智能的。

## 07 奇点

人工智能自出现以来，其未来就一直备受关注。一些人工智能专家过度狂热的预言让记者和文化评论员们十分振奋，有时甚至是害怕。如今最好的例子是奇点（Singularity），即人工智能超过人类智力极限的时间点。

奇点代表人工智能将达到人类水平的智能（默认这将是真正的智能，见第6章）。不久的将来，强人工智能将变为超人工智能。届时系统将智能化到可以自我复制，从而在数量上超过人类，并且还可以自我提高，从而在思想上超越人类。最重要的问题和决定将交由计算机负责。

这个概念颇具争议性。它是否能够发生、它是否将发生、它什么时候发生，以及它是一件好事还是坏事，人们对此意见不一。

奇点信徒（S信徒）认为，随着人工智能的进步，奇点必将到来。有些人支持这一观点。他们预言人类的问题将被解决。战争、疾病、饥饿、无聊，甚至个人死亡等问题都将不复存在。其他人预言人类将终结，或者我们所知道的受开化的生命必将结束。史蒂芬·霍金与人工智能主要教科书的合著者斯图尔特·罗素（Stuart Russell）在2014年5月发表评论，称忽略人工智能的威胁“可能是我们犯的最大错误”，这在全世界范围内引起了轰动。

相比之下，奇点怀疑派（S怀疑派）不指望奇点发生——肯定不是在可预见的未来。他们承认，人工智能造成了许多让我们头疼的问题。但他们还没有看到存在的威胁。

## 奇点的预言家

强人工智能向超人工智能过渡的观点近来已经成了媒体的老生常谈，但它最早始于20世纪中叶。关键的发起人是杰克·古德（“Jack”Good，他和艾伦·图灵一道在布莱切利园担任密码破译家）、弗诺·文奇（Vernor Vinge）和雷·库兹韦尔（Ray Kurzweil）。而图灵自己曾经预言“机器将获得控制权”，但没有详细说明。

1965年，古德预言了一台超智能机，它将“远远超过所有人类的智能活动”。这台机器可以设计更优质的机器，所以它将“毫无疑问地引起智能爆炸”。当时的古德仍然很谨慎，说：“第一台超智能机是人类需要做的最后一项发明，倘若这台机器听话到告诉我们如何控制它。”然而，他后来指出超智能机会毁灭人类。

25年以后，文奇推广了“奇点”这一术语（在这一背景下，由约翰·冯·诺依曼于1958年发起）。他发表了一篇名为《技术奇点即将来临：后人类时代生存指南》（The Coming Technological Singularity）的论文，称奇点到来时，所有的预言都会被打破（堪比黑洞的事件穹界）。

他承认，奇点本身能够被预知：的确，它不可避免。但在它带来的许多（未知的）后果中，可能包含对人类文明甚至人类的破坏。我们正朝着这样一种情况前进：“以前所有的规则将不复存在，也许就在眨眼之间，一切都指数般地增加，毫无控制的希望。”他说，即使每一个政府都意识到了危险并试图阻止，但也是心有余而力不足。

文奇和古德的悲观情绪（最终）遭到雷·库兹韦尔的反击。他不仅展现了惊人的乐观情绪，还给出了日期。

库兹韦尔的著作《奇点临近》（The Singularity is Near: When Humans Transcend Biology）一书，标题引人注目，指出强人工智能将在2030年之前实现，到2045年，超人工智能（与纳米技术和生物技术相结合）将战胜战争、疾病、贫困和个人死亡。它将产生“艺术、科学和其他知识形式的爆炸……将赋予生命真正的意义”。到21世纪中叶，我们生活的拟真虚拟现实将远比现实世界更加丰富、更加令人满意。对于库兹韦尔来说，“奇点”真的是奇异的，“临近”真的意味着临近（这种超级乐观的情绪有时会缓和下来。库兹韦尔列出了许多现实存在的风险，主要来自人工辅助生物技术。他说，至于人工智能本

身，“智能天生不可能控制.....如今若要制定策略，绝对要确保未来的人工智能体塑人类的伦理和价值观，这种做法是行不通的”）。

库兹韦尔的论点根基是“摩尔定律”，它是英特尔创建者之一戈登·摩尔（Gordon Moore）的观察报告，即一美元带来的计算机能力每年翻倍（物理定律最终将征服摩尔定律，当然不是在可预见的未来实现）。库兹韦尔指出，任何指数级增长都有悖于常理。他说这表明人工智能的发展速度将超乎想象。所以，他和文奇一样坚持认为，基于过去经验的预测基本没有价值。

## 竞争的预测

尽管有些人宣称，对奇点出现之后所做的预言几乎毫无价值，但还经常会出现此类预言。文学作品中也充斥着大量令人难以置信的例子，我们在这里仅列几个。

S信徒分为两个阵营：悲观者（跟随文奇）和乐观者（跟随库兹韦尔）。他们多半同意，强人工智能向超人工智能的转型将在本世纪末之前顺利发生，但他们的分歧点是超人工智能到底有多危险。

例如，有些人预言，邪恶的机器人将竭尽全力破坏人类的希望和生活（如科幻小说和好莱坞电影中常见的情况）。我们可能认为，有必要的话，“拔掉插头（结束机器人的工作）”不就行了，但是这种想法已被明确否认。我们只知道，超人工智能将会变得超级精明，要靠拔掉插头解决问题绝不可能。

其他人认为，超人工智能不会有恶毒意图，但是无论如何都非常危险。我们不会把人类的仇恨嵌入到它们身上，它们也没有理由自己酝酿仇恨。不过它们会对我们很冷漠，就像我们对大多数非人类物种一样。但如果我们的利益与它们的目标发生冲突的时候，它们的冷漠可能让我们堕落：变成渡渡鸟般（蠢笨）的智人。在尼克·博斯特伦（Nick Bostrom）被广泛引用的思想实验中，有制作回形针意向的超人工智能将从人体的原子中提取有用之物来完成这一目标。

或者，我们再谈谈有时用来防范奇点威胁的一般策略——遏制。阻止超人工智能直接作用于世界，虽然它可以直接感知世界。它只被用来回答我们的问题〔博斯特伦称为“Oracle”（意为预言的东西）〕。然而，世界包括互联网，超人工智能可能向互联网提供内容来间接影响人类，如事实、虚假信息和计算机病毒等。

另一种形式的奇点悲观情绪预言，机器将让人类为它们鞍前马后，做它们的肮脏工作，即使这违背人类的利益。同时，这也抨击了“切断电源”的观点——人类不可能通过切断超人工智能系统与世界的联系来“遏制”它。他们还说，超智能机器能使用贿赂或威胁的手段，说服有时与它有联系的几个人中的一人，做它无法直接做到的事情。

这种特别的担心假定，超人工智能将足够了解人类心理，可能知



道什么样的贿赂或威胁可能起作用，也可能知道哪些人最容易被哪些说服方式说服。如果有些人认为此假设难以置信，那么相关专家会回复说，金钱贿赂或谋杀威胁几乎对任何人屡试不爽，所以超人工智能既不需要亨利·詹姆斯（Henry James）那样的心理洞察力，也不需要从人的角度来理解说服、贿赂和威胁实际上指什么。它只需要知道，向某人输入某些NLP文本，就能以明显可预测的方式影响这些人的行为。

一些乐观的预言甚至更具挑战性。最引人注目的也许是库兹韦尔对在虚拟世界中生活和消除个人死亡的预言。生老病死目前仍然是自然现象，虽然人的平均寿命（利用超人工智能辅助的生物科学）与以前相比已经高出很多。不过到奇点到来之后，只要把每个人的个性和记忆下载到电脑中，死亡就可以“战胜”啦！

他在2005年出版《奇点临近》一书，书的副标题“当计算机超越人类”体现了一个哲学上让人难以信服的假设，即一个人可以存在于硅或神经蛋白（见第6章）上。库兹韦尔表达了他的“奇异”设想，也被称为超人主义或后人类主义——一个部分或甚至全部是非生物人的世界。

设想称，这些超人主义的“机器人”将会有各种直接连接到其大脑、假肢肢体或感觉器官的电脑植入物。聋哑情况将不会发生，因为视觉和听觉信号将由触觉来解读。尤为重要的是，特制药将增强理性认知（以及情绪）。

我们身边已经有了这些辅助技术的早期版本。如果这些技术像库兹韦尔所说的那样会激增的话，那么我们的人类概念将会发生深刻变化。我们不再把假肢看作人体有用的附加物，而是将其视为（准）人体部分。广泛使用的精神药物将与天然物质一同列入“人脑”内物质中，如多巴胺。转基因个体的优越智力、优势或美感将被视为“自然”特征。有关平等主义和民主的政治观点将受到挑战。拥有足够财富去利用这些可能性的人甚至可能进化出来一个新的子物种（或物种）。

简而言之，生物进化有望被技术进化所取代。库兹韦尔把奇点看作“我们的生物思维和生物存在与技术合并的高潮，将带来一个无区别化的世界……人与机器之间无差别，或物理现实和虚拟现实之间无区别”（如果你觉得信息量一下子太大，需要缓缓，那请这么做吧）。

人工智能将如何改变我们对人性的观念，超人主义是一个极端例

子。有一种不那么极端的哲学是“延展心智”，它将技术融合到心智的概念中，认为心智分散到世界，涵盖依赖它的认知过程。虽然延展心智的概念已经有了广泛影响力，但超人主义没有。超人主义得到了一些哲学家、文化评论家和艺术家的大力支持。然而，并不是所有的S信徒都接受它。

## 为怀疑论辩护

在我看来，S怀疑派是对的。第6章中对心智作为虚拟机的讨论，意味着实现人类水平的人工智能原则上是无障碍（可能除了现象意识）的。这里的问题是在实践中是否能实现。

许多有关奇点出现之后的预言违背常理，超人主义哲学近乎荒谬（恕我直言），除此之外，S怀疑论者还有其他论点。

人工智能没有很多人假设的那么有前途。第2章到第5章都提到了无数当前的人工智能不能做的事情。许多事情都需要一种人类的关联感（并且默认语义web已经完成，见第2章）。此外，人工智能一直专注于智力的理性，却忽略社会/情感智能，更别提心智了。能够与我们的世界充分交流的强人工智能可能也需要这些能力。另外，人类的心智何其丰富，我们还需要与其工作方式相关的良好心理/计算理论。人类水平的强人工智能的前景看起来黯淡无光。

即使实践中可行，是否有足够的资金去实现它仍然值得怀疑。目前，政府向全脑仿真（见下一节）研究投入了海量资源，但是打造人工人类心智的经费还不止于此，未来还将会继续烧钱。

根据摩尔定律，人工智能有望持续不断地取得进步。但是增大的计算机功率和增加的可用数据（如云存储和整个物联网上的24×7小时监测的传感器）不能保证出现似人类的人工智能。这对S信徒来说是个坏消息，因为实现超人工智能的前提是强人工智能。

S信徒忽略了当前人工智能的局限性。他们根本不在乎，因为他们有一张王牌：技术的突飞猛进正在改写所有规则手册的概念。这允许他们随意预言。他们偶尔会承认“本世纪末的”预言可能不现实。但是，他们坚持认为，“从不”仅代表一段长时间。

“从不”的确是一段长时间，所以包括我自己在内的S怀疑者都有可能错了。这些S怀疑者没有强有力的理由——特别是如果他们原则上承认强人工智能可能实现（像我一样）。他们甚至可以被说服：尽管有巨大的延迟，但奇点最终会发生。

然而，仔细考虑最先进的人工智能，我们有理由支持怀疑派的假设（或如果你喜欢，也可以说是他们的赌注），而不是S信徒的胡乱

猜测。

## 全脑仿真

S信徒预言人工智能、生物技术和纳米技术将以指数级速度发展，它们之间的合作将日新月异。事实上，这些预测已经在发生了。大数据分析现在用于推进基因工程、药物开发和许多其他科学项目（埃达·洛夫莱斯的论证，见第1章）。同样，人工智能和神经科学现已经被结合用来研究全脑仿真（WBE）。

WBE的目的是通过仿真大脑的单个组件（神经元）以及各组件之间的连接和信息处理能力，来模拟一个真实的大脑。希望在于，这些所获得的科学知识有许多应用，包括从阿尔茨海默症到精神分裂症等精神疾病的治疗。

这种逆向工程需要用神经形态计算模拟亚细胞过程，例如离子通过细胞膜的过程（见第4章）。

神经形态计算取决于各类神经元的解剖结构知识和生理学知识。但是，WBE还需要特定神经元的连接和功能的详细证据，包括时间。多数证据需要改进的脑扫描术，连续监测单个神经元的微型神经探测器。

各种WBE项目正在进行中，它们通常被赞助商们拿来和人类基因组计划或登月比赛作对比。例如，2013年，欧盟宣布了耗资10亿英镑的人脑工程。当年晚些时候，美国时任总统奥巴马宣布由美国政府拨款30亿美元（加上大量私人资金），完成10年期的BRAIN计划（使用先进革新型神经技术的人脑研究）。它的目标首先是完成老鼠大脑连接的活动图，然后再模拟人脑活动图。

部分大脑仿真的早期尝试也是由政府资助的。2005年，瑞士赞助了蓝脑计划，最初是为了模拟大鼠的皮质柱，长远目标是模拟人类新皮层中的百万柱。2008年，美国国防高等研究计划署（DARPA）向神经形态自适应塑料可扩展电子系统（SyNAPSE）项目投入了近4000万美元；到2014年，又投入了4000万美元，这些资金完成了芯片承载54亿根晶体管的工作，每块芯片有100万单位（神经元）和2.56亿突触。德国和日本正在合作利用神经模拟技术（NEST）开发超级计算机Kcomputer。到2012年，Kcomputer需要40分钟来模拟1%的真实大脑活动的1秒，包含17.3亿个“神经元”和10.4万亿“突触”。

哺乳动物的WBE研究成本如此高昂，所以相关的研究很罕见。但是模拟小型大脑的无数次尝试正在世界各地遍地开花（我自己的大学专注于蜜蜂）。这些研究可能提供神经科学方面的见解，有助于人类大脑的WBE研究。

考虑当前的硬件进步（例如SyNAPSE的芯片）和摩尔定律，库兹韦尔的预测可能成为现实：到21世纪20年代，将出现与人类大脑原始处理能力相匹敌的计算机。但这并不代表到2030年，这些计算机能达到人类智力。

因为虚拟机（见第1章和第6章）才是关键。有些虚拟机只能在超级强大的硬件上实现。因此，百万级晶体管计算机芯片可能必不可少。但是，这些芯片将执行什么计算？换句话说，在它们上面实现的是什麼虚拟机？为了达到人类（即使老鼠）的智力，这些虚拟机必须以超出计算心理学家的理解范围的方式，显示出其信息的强大性。

我们假设，人脑中的每个神经元最终都会被映射（我认为不太可能）。但映射本身不会告诉我们它们在做什么（微小的线虫类蠕虫C.线虫只有302个神经元，我们已经精确知道它们的连接，但实际上我们还不能识别突触是兴奋状态还是抑制状态）。

对于视觉皮质，我们已经有了一个神经解剖学和心理功能之间相当详细的映射。但是对于一般的新皮层，情况并非如此。更为重要的是，我们不太了解额叶皮质做的是什么事情——也就是说，在它里面实现的是什麼虚拟机。这个问题在大型WBE中不是很突出。例如，人脑工程采用了自下而上的方法，观察人体的解剖构造和生物化学过程并模拟。不考虑和大脑可能支持的心理功能相关的自上而下的问题（几乎不涉及认知神经科学家）。即使全部完成解剖模型，并仔细监测化学信息，这些自上而下的问题也不会得到答案。

要得到答案，必须有各种各样的计算概念。此外，一个关键点是整个心智（或心智—大脑）的计算架构。我们在第3章中看到，多动机生物的动作规划需要复杂的调度机制，例如情感提供的机制。第6章中讨论的LIDA表明，皮质处理超级复杂。即使是进餐用刀叉的普通活动，也需要许多虚拟机进行整合：有些处理物理对象（肌肉、手指、器具、各种传感器）；有些处理意图、计划、期望、欲望、社会习俗和偏好。为了理解这种活动是如何实现的，我们不仅需要大脑的神经科学数据，还需要与所涉及的心理过程相关的详细计算理论。

简而言之，把自下而上的WBE看作理解人类智力的途径可能会失败。但它可能有助于我们理解大脑。它可以帮助人工智能科学家开



发出更实际的应用。但是，到21世纪中期，届时的WBE将能解释人类智力的观点是一种错觉。

## 我们应该担心什么

如果S怀疑派是对的，确实没有奇点，那也不代表我们就万事大吉了。人工智能已经引起了人们的担心。未来的进步肯定会带来更多问题，所以对人工智能长期安全的焦虑也是非常必要的。更重要的是，它的短期影响也不容忽视。

有些担心非常大众化。例如，任何技术都可以用于做好事或坏事。坏人会使用所有可用工具，有时甚至资助新工具的开发，去做坏事（例如，CYC可能对坏人有用：它的开发人员已经在考虑如何限制完整系统的访问权限，详见第2章）。因此，我们必须警惕自己发明的东西。

正如斯图尔特·罗素所述，我们不仅仅要对发明的“目的”保持警惕。如果有10个与问题相关的参数，而统计优化机器学习时（见第2章）只考虑了6个参数，那么其他4个——很可能——“走向”极端。因此，我们还需要警惕使用中的数据类别。

这种担心一般包含框架问题（见第2章）。像童话故事中的渔夫一样，他的愿望是让当兵的儿子回家，而实现这个愿望的交换条件是他自己被带入一副棺材中，我们可能会惊讶于强大的人工智能系统竟然不能像我们一样理解相关性。

例如，如果冷战预警系统建议对苏联进行一次防御性打击，那么操作员的关联感是避免灾难的唯一方法，既包括政治层面，也包括人道主义。他们认为，作为联合国成员的苏联最近并没有过分扰乱秩序，他们也害怕核攻击带来的可怕后果。所以，操作员们违反协议，忽略自动预警。还有几次核打击也侥幸避免；有些还是最近的事。战争没有升级仅仅是因为人们的常识。

此外，人为失误总有可能发生。有时这可以理解（由于工作人员手动控制让计算机停下来，使得三哩岛核泄漏事故后果更为严重，但他们面临的实际条件极不寻常），但它可让人大吃一惊。上一段提到的冷战时期的错误预警，是因为有人在设计日历时忘记了闰年，所以月球出现在了“错误”的地方。还有就是他们在没有测试和（如果可能）证实人工智能可靠性的情况下就将其投入使用。

有些担心非常具体。今天有些东西将会让我们伤透脑筋。一个主

要的威胁是技术失业。好多体力和低级文职工作已经消失。其他类似的工作也会步入后尘（虽然需要灵巧和适应性的体力工作不会消失）。仓库中的大多数起重、提取和搬运工作现在都可以由机器人完成。无人驾驶车辆意味着有人要失业。

中层管理职位也有风险。许多专业人员已经在使用人工智能系统作为辅助工具。不久以后，对法规和案例的耗时研究工作（例如法律和会计）可能大部分由人工智能接管。很多高要求的工作很快也会受到影响，如医学和科学领域。即使这些工作继续存在，但技术含量也会更高。专业培训将会受到影响，年轻人该如何学会做明智的判断？

虽然一些法务工作以后可能不再需要，但是律师行业也将从人工智能中获益，因为有一大堆法律陷阱等着处理。如果出现问题，谁应该承担责任？程序员，批发商，还是零售商或用户？一位人类专业人士有时可能因为没有使用人工智能系统而被起诉？如果系统的（无论在数学上还是凭经验）可信度经证明很高，那么这种诉讼将很有可能发生。

新型工作无疑会出现，但这是否代表更多的工作机会、受教育的机会以及更强的赚钱能力（就像工业革命后发生的），对此，我们无法给出确切答案。未来还存在严重的社会政治挑战。

“服务”岗位受到的威胁较小。但是，即使是这些岗位也几近消失。在理想化的世界里，当前不受重视的一对一活动很可能在以后大量增加和升级。但是，这也并非板上钉钉的事。

例如，教育开始接受个人或基于互联网的人工智能辅助，如学术界大牛授课的MOOC（大规模公开在线课程），这对许多人类教师的工作大有裨益。计算机心理治疗师已经应运而生，费用比人类治疗师低得多（有些计算机心理治疗师提供的帮助大得令人吃惊，例如，识别抑郁症）。但是，它们完全不受管制。我们在第3章中看到，人口结构变化正鼓励研究照顾老人的“护理人员”和“机器人保姆”，这些都是“钱”途可嘉的领域。

除了失业影响，在人类生存的环境中使用冷漠的人工智能系统不仅在现实中充满风险，而且在道德上也值得怀疑。专家们为与外界接触十分有限的老年人和残疾人设计了多款“计算机伴侣”。它们不仅提供帮助和娱乐服务，还与用户对话、逗乐用户，并填补用户情感上的缺失。即使是弱势群体（如帕罗的用户），也会通过这种技术而变得更加快乐，但是他们作为人的尊严也不知不觉遭到了背叛（文化差异在这里很重要：例如日本和西方对机器人的态度差别很大）。

老年人可能喜欢与人工伴侣讨论过往，但这是真的“讨论”吗？这可能是一个令人愉悦的提醒，触发老人心中的美好记忆罢了。然而，人工伴侣带来的好处是不会让用户有一种对方能够对自己的经历感同身受的错觉。人们在接受咨询服务情绪激动时，通常最想要的是自己的勇气或痛苦得到“承认”，但这需要基于双方对彼此状态的了解。我们通过提供一种表面上像同理心的东西来短暂地“欺骗”一个人。

即使用户患有中度痴呆，用户对人工智能体的“猜想”可能比智能体的人类模型要丰富得多。那么，当某人回忆起伤心之事（失去子女）时，如果智能体没有按照期望和要求作出回应，这将产生什么后果？如果伴侣按照惯常的方式表达同情，这对消除用户的悲伤情绪没有任何帮助——而且可能弊大于利，可能勾起该用户的悲伤情绪，同时还没有什么其他安慰方式。

还有就是伴侣是否应该偶尔保持沉默或者说一个善意的谎言。不讲情面说真理（和沉默）可能会让用户心烦意乱，但是委婉礼貌需要极其先进的NLP和一个精准的人类心理模型。

至于机器人保姆（不考虑安全问题），让婴儿过多接触人工智能系统可能会扭曲他们的社会观，不利于其语言习得。

人造性伴侣不仅出现在电影中，例如，电影《她》（Her）。它们已经投入市场。有些能够识别语音，用语言和肢体动作勾引用户。它们扩大了互联网的影响，让人们的性体验变得糟糕（同时增强了女性在性方面被物化的意识）。许多评论家（包括一些人工智能科学家）已经写了一些关于与机器人发生性关系方面的文章，揭示了一个非常肤浅的概念：个人的爱慕接近情欲、性迷恋和单纯令人舒适的熟悉感。然而，这种观察报告不太可能产生预期效果。通常情况下，色情作品是一棵“摇钱树”，所以要想阻碍人工智能性玩偶未来的“进步”基本不可能。

另外，隐私也是个棘手的问题。强大的人工智能搜索和人工智能学习发布了大量自媒体和家庭或可穿戴传感器上收集的数据，有关人工智能的争议因此也變得更多（谷歌公司最近为一款机器人泰迪熊注册了专利，它有相机眼睛、麦克风耳朵，嘴巴里有扬声器，它既能和孩子，也能和父母沟通——不管愿意与否，它还可以和看不见的数据收集器沟通）。

网络安全问题也不容忽视。随着越多人工智能进入到我们的世界（通常以非透明的方式），它就越重要。要想不被超人工智能控制，那就必须知道如何写不被黑客攻击/改变的算法（“友好人工智能”的

目标，见下一节）。

军事应用也引发了关注。机器人扫雷非常受欢迎。但机器人士兵或机器人武器呢？当前的无人机受人类唆使，但即使如此，只要扩大操作者和目标之间的人类（不仅仅是地理上的）距离，无人机就能加重苦难。人们务必相信，未来的无人机不会有权力决定哪个人或哪样东西成为攻击目标。即使靠它们去识别一个（看起来像是人类选出的）目标，也会引发诸多令人不安的道德问题。

## 我们为此做了些什么

虽然只有很少人工智能工作人员直到现在才更多关注前文提到的问题，但是相关担忧并非最近才出现。

1972年，几位人工智能先驱参与了在意大利科莫湖（Lake Como）召开的会议，他们在会上谈到了人工智能的社会影响，但约翰·麦卡锡不赞同他们的观点，认为当时的推测为时过早。几年后，计算机科学家约瑟夫·魏泽鲍姆（Joseph Weizenbaum）出版了一本副标题为“从判断到计算方法”的著作，并哀叹将二者混淆属于“猥亵行为”（obscenity），但他遭到了蔑视，并被赶出人工智能研究界。

当然有一些例外。例如，在《第一本概述人工智能的书》中，最后一章就是“社会影响”。1983年成立了计算机专业人员社会责任组织（Computer Professionals for Social Responsibility, CPSR）。这离不开SHRDLU的作者特里·威诺格拉德的努力，详见第3章。但这一组织主要是为了警告星球大战技术不可靠——计算机科学家大卫·帕纳斯（David Parnas）还向美国参议院提到了这一点。对冷战的担忧消除之后，大多数人工智能专业人士似乎不太在意它们的研究领域。只有少数几个多年来一直专注研究社会/伦理问题，例如谢菲尔德大学（University of Sheffield）的诺埃尔·沙基（Noel Sharkey）<sup>[1]</sup>，以及一些人工智能哲学家，如耶鲁大学的温德尔·瓦拉赫（Wendell Wallach）和萨塞克斯大学的布莱·惠特比（Blay Whitby）。

现在，一提到人工智能的实践和未来，问题就变得更加紧迫。人工智能领域内（以及在一定程度上，人工智能领域之外）的社会影响得到越来越多的关注。

有些重要的反应与奇点无关。例如，联合国人权观察（Human Rights Watch）组织长期主张一项禁止全自动武器的条约（尚未签署），如自主选择轰炸目标的无人机。最近，一些历史悠久的专业团体复审了它们的研究重点和行为守则。但是，对奇点的讨论让更多人加入到这场辩论中。

许多S信徒和S怀疑派都认为，奇点出现的概率非常小，但它造成的后果将十分严重，我们现在应该未雨绸缪。尽管文奇声称，对于潜在的威胁，我们束手无策，不过还是有几个机构已经应运而生，以



防止它的到来。

这些机构主要由人工智能慈善家资助，包括设在英国剑桥的存在风险研究中心（CSER）和位于牛津的人类未来研究所（FHI），以及位于美国波士顿的未来生命研究院（Future of Life Institute, FLI）和伯克利的机器智能研究所（Machine Intelligence Research Institute, MIRI）。Skype的共同开发人扬·塔里安（Jaan Tallinn）参与创办了CSER和FLI。这两家机构不仅与人工智能专业人员沟通，还试图向政策制定者和其他有影响力的公众发布危险警告。

2009年，美国人工智能协会的主席埃里克·霍维兹（Eric Horowitz）组织了一场小型的专家小组研讨会，研究该采取什么样的必要预防措施去指导，甚至推迟为社会带来问题的人工智能研究工作。这场研讨会在加利福尼亚州的艾斯罗马（Asilomar）举行，遗传学专家几年前曾在此地同意暂停某些遗传研究。然而，作为专家小组的一员，我的印象是并非所有的与会人员都深切关注人工智能的未来。之后的报告也没有得到广泛的媒体报道。

2015年1月，FLI和CSER在波多黎各联合召开了规模更大的类似会议（根据查塔姆宫规则，没有记者出席）。6个月前，组织者马克思·泰格马克（Max Tegmark）、罗素和霍金共同签署了一封警告信函。不出所料，那时的气氛显然比在艾斯罗马更紧迫。会后，大量新的资金〔来自互联网百万富翁埃隆·马斯克（Elon Musk）〕投入到了人工智能安全和伦理人工智能的研究中，成千上万的人工智能研究者共同签署了一封警告公开信，后来在媒体上广泛传播。

不久之后，汤姆·米切尔（Tom Mitchell）和其他几位领域带头人起草了第二封公开信，警告不要开发自主武器，应该禁止武器在无人干预的情况下选择和攻击目标。签署人希望“防止发生全球人工智能军备竞赛”。2015年7月，这封信出现在人工智能的国际会议上，有将近3000名人工智能科学家和17000名相关领域的专业人士在上面签名，从而引起了媒体的广泛关注。

波多黎各会议之后，麻省理工学院经济学家埃里克·布林约尔松（Erik Brynjolfsson）和安德鲁·迈克菲（Andy McAfee）也在2015年6月共同签署了一封公开信，向决策者、企业家、商人和经济学专家们发出警告，称人工智能可能让经济发生彻底变化，并给出了一些可能降低但无法消除风险的公共政策建议。

这些人工智能研究仍在努力说服大西洋彼岸的政府资助者们认可社会/伦理问题的重要性。美国国防部和国家科学基金会最近都表示

愿意为此提供研究经费。但政府不是最近才支持，多年来，“政府的”兴趣一直在不断增加。

例如，2010年，两家英国研究理事会赞助了一个跨学科的“隐修会”（Robotics Retreat），参与起草了机器人专家的行为准则。商定了机器人学的五项“原则”，其中两项涉及前文所讨论的担忧：（1）机器人不应被设计为武器，除非出于国家安全考虑，（4）机器人为人工制品：情感和意图的错觉不得用于操控弱势用户。

另外两项规定了人类承担的道义责任：（2）人类而不是机器人为承担责任的主体；（5）应该可以找出任何机器人的（合法）负责人。该组织没有试图更新艾萨克·阿西莫夫的“机器人三定律”（机器人不得伤害人类，或因不作为（袖手旁观）使人类受到伤害；除非违背第一法则，机器人必须服从人类的命令；在不违背第一及第二法则下，机器人必须保护自己）。起草者认为，这里的任何“法则”都应该由人类设计者/建造者遵守，而不是机器人。

2014年5月，由美国海军资助的一项跨学科学术活动（5年750万美元）受到媒体的追捧。它是一个由5所大学合作的项目（耶鲁大学、布朗大学、塔夫茨大学、乔治城大学和伦斯勒理工学院），旨在发展机器人的“道德能力”。参与者包括认知和社会心理学家、道德哲学家，还有人工智能程序员和工程师。

此项目不是为了提供一系列道德算法（和阿西莫夫的法则一样）或优先研究某种特殊的元伦理学（例如功利主义），甚至也不是为了定义一套非竞争的道德价值观，而是希望开发一个能够在现实世界中有道德推理（和道德讨论）能力的计算系统。因为自主机器人要做出慎重决定，而不只是遵照指示（也不是对“情境”线索作出灵敏反应，见第5章）。例如，如果一个机器人开展搜救工作，那么它应该先疏散/救援谁？或者，如果提供社会陪伴服务，那它什么时候该和用户撒谎（如果真有这种情况）？

项目提出的系统将整合知觉、运动行为、NLP、推理（演绎和类推）和情感。情感包括：情感思维（它可以发出重要事件要发生的信号，并规划相互矛盾的目标，详见第3章）；“抗议和痛苦”的自动显露，这可能会影响与之交互的人做出的道德决定；以及识别周围的人的情感。而且官方发言称，机器人甚至可能“超越”普通（即人类）的道德能力。

考虑到第2章和第3章中指出的实现强人工智能的障碍，再加上与道德相关的特定困难（见第6章），人们可能会怀疑这个目标是否能

够实现。但项目值得进行。考虑现实世界的问题（如前文提到的两个差异巨大的例子）是一盏警示灯，提醒人类在使用人工智能的时候要遵守道德准则，否则会发生许多危险。

除了这些机构所作出的努力，越来越多的人工智能科学家也瞄准了埃利泽·尤德考斯基（Eliezer Yudkowsky）所说的“友好人工智能”。它对人类有积极影响，既安全又实用。它包含的算法将易于理解、值得信赖，鲁棒性<sup>[2]</sup>robust）强，并且就算这些算法都失败了，也输得得体。它们应该是清楚易懂、可以预测的，而且不易被黑客操纵——如果其可靠性是通过逻辑或数学证明，而不是凭经验测试，那就更好了。

马斯克在波多黎各会议上捐赠600万美元后，FLI立即发布了前所未有的《征求建议书》（Call for Proposals，6个月后，37个项目全部得到资助），向来自“公共政策、法律、伦理学、经济学、教育及其相关领域”和人工智能领域的专家们发出呼吁：“研究项目一方面是为了避免潜在危害，另一方面是为了最大限度地利用人工智能今后所带来的社会效益，而且仅限于如下研究：明显不是专注于实现增强人工智能能力的标准目标，而是为了增强人工智能的鲁棒性和效用……”总之，欢迎发展友好人工智能的呼吁可能已经发出。但奇点的足迹仍未消失，《建议书》指出：“应优先考虑专注于人工智能的鲁棒性和效用的研究，即使该研究涉及大幅取代当前的能力……”

总而言之，人工智能即将引发世界末日的观点实属错觉。但是，从某种程度上来说，正是由于这种观点，让人工智能研究团体、政策制定者和普通老百姓才逐渐意识到一些切实存在的危险。他们早就该注意到了！

## 译者后记

《AI：人工智能的本质与未来》的作者玛格丽特·博登教授是OBE勋章（大英帝国官佐勋章）获得者，人工智能领域最著名的人物之一。在本书中，作者试图用通俗易懂的语言，向读者介绍人工智能的方方面面。

本书概括起来，可以分为三个部分：人工智能的前世，从埃达·洛夫莱斯夫人的预言到艾伦·图灵的图灵机、到沃伦·麦卡洛克提出的神经活动中内在思想的逻辑演算、再到唐纳德·赫布的学习理论，囊括了人工智能各分支的历史演变过程，提到了多学科共同发展为人工智能的发展带来了巨大推动力；人工智能的今生，讲述了当前技术的最新进展，如逻辑符号主义的不断完善、人工神经网络的再次复兴、机器人技术和人工生命的蓬勃发展以及强人工智能面临的种种难题等；人工智能的未来，作者提出了如下问题，比如，人工智能是否真的能够实现？会思考的智能体是否需要道德约束？何时会出现超越人类智能的智能体？智能体带来利益的同时也带来了威胁，对此我们如何应对？作者不仅详述并总结了当前各个学派的观点，还给出了自己的答案。

这是一本难得的佳作。在当前互联网+的大时代背景下，人工智能技术已经广泛应用于金融、交通等各个民生领域，同时也是计算机行业内部变革最主要的推动力，如云计算、大数据、物联网都需要人工智能技术的介入。神经科学、认知心理学等众多社会学科也在人工智能技术的助力下快速发展。本书不仅从科学探索的角度向读者介绍了人工智能的理论意义，还囊括了众多具有代表性的实践案例，读者阅读的时候不会觉得枯燥乏味，反而会心潮澎湃，感觉自己犹如置身于人工智能这股大洪流之中。同时，作者没有局限于概括性的介绍，在关键部分还详述了很多技术细节，这样读者既能感受到人工智能发展道路的大气磅礴，也有机会细细体会人工智能技术的小桥流水之势。作者还提出了很多颇具启发性的问题，耐人寻味。

无论您是人工智能领域的专业人士，还是仅对人工智能颇感兴趣的普通读者，相信此书都不会让您失望。感谢中国人民大学出版社能够给予我翻译此书的机会，使我受益良多，同时也非常感谢张钊、朱文佳、肖凤霞、鹿桐欣、赵晓玮、高瑞霞、成欣、易汕、马伊莎、况辉等人为本书的翻译工作付出的努力！

孙诗惠

---

#### 注释

[1] 机器人研究学者，现在担任国际机器人武装控制委员会或ICRAC的主席。  
——译者注

[2] 鲁棒性是指控制系统在一定（结构、大小）的参数摄动下，维持其他某些性能的特性。也就是系统的健壮性，这是在异常和危险情况下系统生存的关键。——译者注