1

第 1 单元

人工智能概述

1.1 什么是人工智能

学习目标

（1）理解人工智能的基本概念，包括其研究目标、实现方式，以及与其他学科的区别。

（2）认识智能机器的发展历程，了解古今中外对智能机器的设想与实践。

（3）区分人工智能、自动化和机器智能等相关概念，掌握它们之间的核心差异。

（4）掌握现代人工智能的主要特征，包括自主学习、大数据驱动、模型与程序分离等特点。

（5）了解人工智能的典型应用场景，分析其对人类社会的影响，并思考其带来的积极与负面效应。

人工智能（artificial intelligence，AI），直观理解就是人工制造出来的智能。随着技术的进步，人工智能越来越频繁地出现在我们的生活中。典型的例子包括乘坐高铁时的刷脸进站、家里能听会说的智能音箱、餐厅里导引和送餐的机器人等。那么，究竟什么是人工智能？人工智能的研究目标是什么？与其他学科相比有什么特点？科学家们又是如何实现人工智能的呢？本节将对这些问题进行简要探讨，更深入的知识将在后续内容中展开。

① 智能机器的梦想

人们很早就希望制造出聪明的机器来帮助自己做事。传说我国春秋时期的巧匠鲁班曾经用竹子做了一只会飞的鸟，能在天上飞三天。与鲁班同时代的另一位巧匠偃师，则制作了一个与人极其相像、能歌善舞的人偶。三国时期蜀国的丞相诸葛亮也是一位了不起的发明家，据说他在北伐魏国时曾制造了一种名为“木牛流马”的运粮工具，可以在山间自动行走，不吃不喝，昼夜不停地运粮。

无论是偃师造人还是木牛流马，都只是传说，并无事实考证。然而，正是这些传说反映了人们对于智能机器的向往，这种向往也成为驱动科学家不断探索的动力。随着技术的进步，一些真正的自动化机器逐渐被研制出来。公元1世纪，亚历山大里亚的著名数学家兼工程师希罗在其著作《自动装置的制作》一书中描述了一个全自动的木偶剧院。通过杠杆、滑轮等设备之间的相互作用，这家剧院可以上演一出完整的戏剧。

加扎利（1136—1206）是一位杰出的阿拉伯博学者，集发明家、机械工程师、工匠、艺术家、数学家和天文学家于一身。他最著名的著作是《精巧机械装置的知识之书》，书中描述了50种机械设备的制造方法。在加扎利之前，虽然也有很多作者写过类似的书，但大多对技术细节语焉不详。与之不同的是，加扎利在《精巧机械装置的知识之书》中详细描述了每一个发明的制作细节，只要跟着他的步骤就能复现出同样的机器来。正因如此，加扎利被一些人称为“现代工程之父”。

《精巧机械装置的知识之书》中记录了很多有趣的自动机械装置，如水钟、提水机等。其中的机器人乐团（图1-1）尤其引人注目，他设计了一个翻斗储水箱，每半个小时储水箱装满，随后翻转，将水倾倒进第二个水箱。第二个水箱底

部带有小孔，水流从小孔喷出，冲击叶片带动轮轴转动。这种转动通过一组转轮传导到机械玩偶的手部，让他们拨动风琴或敲击鼓面完成演奏。

图1-1　加扎利发明的机器人乐团

注：左侧四位是由水力驱动的音乐机器人。

② 从自动化到人工智能

前文中提到的各种自动化机器虽然精巧，但很少有人认为它们具有了真正的“智能”。无论设计得多么精巧，这些机器都是基于机械原理实现的。因此，它们所展现出的“智能”实际上是一种机械自动化：自动化程度越高，表现出来的“智能”程度越高。在人类历史上，这种自动化机器起到过非常重要的作用，例如蒸汽机的出现极大地提高了工业生产的效率，内燃机的出现引发了农业机械化的浪潮，电力的出现则让自动化机器越来越普及。然而，自动化程度再高，其智能能力仍是单一的、有限的，难以与人类丰富、高超的智能相提并论。科学家们很早就意识到了这一点，并开始思考如何让机器拥有更强大的、类似人的智能。这便是人工智能思想的源头。

然而，要厘清“人工智能”这一概念并非易事。让我们首先讨论什么是“智能”（intelligence）。通俗地理解，智能是指生物所具备的一般性精神能力，包括推理、理解、计划、解决问题、抽象思维、表达意念和语言、学习能力等方面。

许多动物具有一定的智能，例如可以控制肢体的动作，可以追踪猎物或逃

离风险，甚至表现出一定的学习、组织、规划的能力。人类的智能显然更加高级和全面，尤其在抽象思维方面具有优势。模拟动物的智能，尤其是人类的智能，使机器具备类似的能力，这正是人工智能研究者的初心和使命 。美国计算机学家约翰·麦卡锡（图1-2）是“人工智能”一词的提出者。他曾这样定义人工智能：“人工智能是制造智能机器的科学与工程，特别是智能的计算机程序。”显然，人工智能的这一目标要比制造自动化的机器更高，实现起来也困难很多。

图1-2　约翰·麦卡锡（1927—2011）

同时，要认识到人工智能有其特殊的实现方式，即通过“计算”来实现智能。在长期的探索中，人们逐渐意识到，要想让机器拥有人类的智能，首先需要理解人类的思维过程。研究表明，人类的思维过程可以表示为计算过程。这意味着只要机器能够完成同样的计算，就可能复现出同样的思维过程。这种用计算来实现智能的思想成为人工智能的一个重要特点。现在，计算机是使用最常见的计算工具。

综上所述，人工智能是探讨用计算机模拟人类智能行为的科学。这一定义可以从两个方面来理解。一方面，人工智能的实现手段主要是“计算”，主要工具是计算机。基于物理过程实现的功能通常不作为人工智能的研究对象。如加扎利制造的吸水机、希罗设计的木偶剧院等，都是用机械方式实现的，不存在计算成分，因此不能算作人工智能。另一方面，人工智能起源于对“需要动脑子”的工作即“智能行为”的模拟，如感知、记忆、动作、推理等。汽车在路上跑、吊车移动吊臂这些简单功能一般不被视作智能行为。

值得说明的是，智能行为是智能过程的结果或外在表现，而非智能过程本身。为什么要强调对智能行为的模拟呢？因为人类的智能过程极为复杂，我们至今无法完全了解其具体细节，因此直接模拟人类智能的内部过程并不容易。而且，人工智能的目标是制造更强大的智能机器，这种机器只要表现得足够智能就可以了，并不需要完全复现人类的智能过程。事实上，受到生物属性的限制，人类的智能过程未必是实现智能机器的最优方案。基于上述原因，目前对于智

能行为的模拟是研究界的主流。尽管如此，探讨人类智能的内在过程也具有重要意义，它可以为人工智能研究者提供关键性的启发。因此，理解、模拟、复现人类的智能过程也是人工智能的一个重要研究方向。

③ 区分几个易混淆的概念

1）智能机器与人工智能

一般来说，拥有一定智能的机器可称为智能机器。智能机器是一个较为主观的概念，人们常常对新奇的功能感到智能，而对那些习以为常的功能则不再觉得多么智能。例如，我们家里常用的电吹风、电饭锅、电风扇、计算器，在刚刚被发明的时候无一不是让人震惊的智能机器，但现在很少有人认为它们是智能的。因此，智能并没有绝对标准，我们通常会说某台机器的智能程度如何，而不是轻易断言它是否拥有智能。

智能机器可以通过多种途径实现，包括机械设计和电路控制等。人工智能采用的是计算方式，这是众多实现方法之一，也是目前最受关注的方法。同时，人工智能技术通常与其他方法配合，共同实现智能机器。因此，我们一般也不会断言某台机器是不是人工智能的，而更倾向于判断它是否包含人工智能的成分，以及这部分成分的智能程度如何。

2）自动化与人工智能

自动化（automation）和人工智能是两个不同范畴的概念。自动化更多关注实体机械的外在表现，如机械臂的操作或机器人行走。自动化既可以通过机械设计或物理方式实现，也可以通过人工智能的计算方式实现。典型的如各种不同智能等级的机器人，既能用简单的弹簧机制来产生动作，也可能通过人工智能方法实现爬山、开门、踢足球等复杂行为。与自动化相比，人工智能更关注的是对人类智能行为的模拟，其应用场景十分广泛，自动化和机器人只是其中之一。即便在自动化和机器人领域，人工智能主要解决的也是那些复杂、高级的功能，如抓取过程的自动规划、抓取技巧的自主学习等。

3）机器智能与人工智能

顾名思义，机器智能即机器所具有的智能，人工智能则是人造出来的智能。在大多数场合下，这两者含义相似，常常互换使用，只不过“人工智能”更强调模拟人类的智能行为，目的是把人类的智能复制到机器上。相对地，机器智能并不一定局限于模拟人类，只要其行为方式表现得像有智能即可。随着人工智能技术的进步，“模拟人类”这一点可能会逐渐淡化，因为机器可能从多个方面获得各种智能，其中一些很有可能是人类所不具备的。

4）人工智能与互联网、大数据

人工智能是模拟人类智能行为的科学，关注的是感知、动作、推理、学习、规划、决策、想象、创造、情感等人类特有的智能行为及其实现方法。许多技术本身并不属于人工智能，但和人工智能有很强的相关性，是人工智能生态的重要组成部分，如互联网和大数据技术。互联网是一种信息流通工具，大数据关注的是数据的生产、存储和应用。这些工具和技术对现代人工智能的发展起到了重要的促进作用，为构建强大的人工智能系统提供了有力支持。与此同时，人工智能也在这些领域得到了广泛应用，反过来推动了互联网和数据科学的进一步发展。正是这种协同进步，使人类社会正在迈入智能时代。

4 现代人工智能的特征

人工智能的发展，本质上是知识、数据、算法和算力四个要素相互作用的结果。在人工智能的早期阶段，数据和算力都比较匮乏，因而基于知识的人工智能方法成为主流。随着数据的积累和计算资源的丰富，更多的人工智能研究转向机器自主学习。总结起来，现代人工智能的主要特征可以总结为自主学习、大数据驱动、模型与程序分离三个方面。

1）自主学习

现代人工智能在很大程度上依靠机器的自主学习。传统的人工智能方法多基于知识灌输：人类专家将知识告诉机器，再让机器基于这些知识进行推理。这

种知识型方法设计复杂，且难以超越人类知识的极限。现代人工智能以大规模学习为核心，基于知识设计学习框架，让机器在这一框架内自主学习，从而有机会突破人类知识的极限。这便是“机器学习”的思路。如今，机器学习方法已经非常强大，不仅能从历史积累的数据中学习，还能在实际环境中边工作边学习，实时进行自我更新。

2）大数据驱动

基于学习的现代人工智能通常需要大量数据。例如，美国OpenAI公司在2022年年底推出的ChatGPT系统，训练所用文本数据量相当于175万本《红楼梦》。近年来，人们发现通常训练所用的数据越多，人工智能系统的能力通常越强。目前，人工智能几乎使用了人类所积累的全部公开数据源，正在整合各个专业领域的数据。可以预期，随着对专业领域数据的梳理与使用，人工智能将在各行各业都取得更大的突破。

3）模型与程序分离

模型与程序分离是现代人工智能系统的典型设计。这里的模型是指机器内部的学习结构，即存储知识的方式。神经网络是目前的主流模型，它通过模拟人类大脑的神经结构来实现知识的表示与存储。基于这一模型，人工智能的设计者不必再为机器的每一步具体行为编写程序，而是让机器依靠模型中积累的知识自主决策。

自主学习、大数据驱动、模型与程序分离三者彼此关联，共同定义了现代人工智能系统的构建方式、知识来源和运行架构。

⑤ 人工智能的应用与影响

1）传统人工智能领域

刷脸支付：目前人脸识别技术已经达到较高精度，可在手机或支付终端实现刷脸支付，不用带现金也可以结账。相关的技术也应用在高铁进站、抓捕逃犯

等各种场景中。

语音助手：目前，许多智能手机中都自带语音助手。通过语音对话，语音助手可以帮助我们完成简单任务，如拨打电话、订机票等。这一技术还应用在地图导航中，通过语音控制导航系统，让司机专心开车，减少事故的发生。

推荐系统：各类新闻和视频网站会依据用户的浏览偏好，自动推荐可能感兴趣的内容。类似地，购物网站也会根据用户的购买行为进行推荐，减少用户自行搜索的麻烦。

自动驾驶：目前自动驾驶技术越来越成熟，一些公司开发的无人驾驶汽车已经在一些科技园区和货运码头试运行。可以预期，未来无人驾驶的汽车、飞机、轮船会越来越普及。这些无人驾驶设备不仅能够节省人力成本，还可以通过设备间通信来协调行驶的路径和速度，从而大幅提升通行效率并增强交通安全。

送餐机器人：送餐机器人已经在很多餐厅出现，外形可爱，代替人类服务员送餐。它们遇到行人时会主动避让，还可以通过语音与客人进行交流。这类服务机器人也出现在一些酒店里，帮助服务员给客人送物品。

2）人工智能的学科交叉融合

近年来，人工智能不断突破了传统的视、听、言、行等应用范围，逐渐与更多专业领域交叉融合，极大地推动了社会生产力的提升。一些典型应用领域如下。

医疗健康：人工智能在读取医疗影像、辅助病理分析等方面已有显著进展，其准确度已经超过了不少人类医生。人工智能还可以辅助医生进行疾病诊断和治疗方案的制定，使诊断更准确，治疗方案更完善。除此之外，人工智能还可以帮助医院优化就医流程、监测慢性病人的病情、优化ICU的资源配置。更进一步，人工智能正在帮助科学家们加快新药研发、研制癌症疫苗、预测传染病的传播趋势。可以想象，在不久的将来，人工智能将在保护人类健康方面做出越来越大的贡献。DeepMind团队开发的AlphaFold可以通过氨基酸序列预测蛋白质的三维结构，从而获知蛋白质的功能，如图1-3所示。这一成果可以帮助人们理解生命过程，加速新药设计，因此获得了2024年的诺贝尔化学奖。

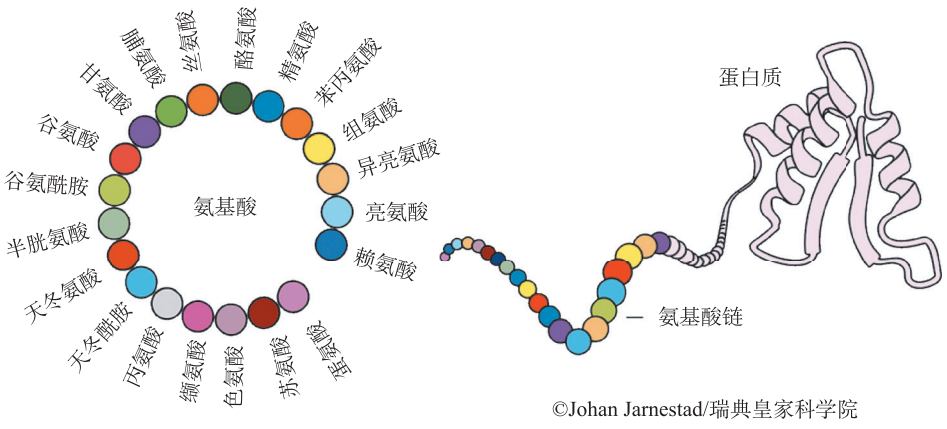


图1-3　DeepMind团队开发的AlphaFold可以通过氨基酸序列预测蛋白质的三维结构

天文学：天文学已进入大数据时代，人工智能成为天文学家们处理和分析海量观测数据的得力助手。首先，人工智能可以帮助天文学家选择天文台站的位置，充分考虑地理环境、大气条件、城市光污染、人造卫星干扰等多重因素，给出合理的选址。另外，人工智能还可以帮助天文学家分析望远镜数据，从中找出人眼很难发现的新天体、新现象、新规律。例如，目前对引力波的探测就是通过人工智能技术辅助进行的。除此之外，人工智能还可以帮助天文学家调校望远镜、监控设备工作状态、提升天文图像的质量等。

金融分析：人工智能可依据历史数据和市场变化因素来预测金融市场的走向，也可以预测某只股票的涨跌趋势。在保险领域，人工智能可以分析用户的信息，预测理赔概率，从而给出合理的保险价格。

城市交通管理：人工智能可以帮助交警合理设置红绿灯时长，缓解路口拥堵。通过红绿灯网络和导航系统，可以分流车辆，设计绿灯路段，不仅可以提高通行效率，还可以为特殊车辆的通行开辟快速通道。此外，通过历史交通数据，人工智能还可以分析交通网络中的关键节点，提供改造建议。

教育教学：人工智能能帮助教师查找资料、设计授课流程、评估教学效果、提出改进建议。教师还可以利用人工智能分析学生的学习情况，设计个性化的教学方案。人工智能还可以作为智能助教，回答学生的问题，给出解题思路，推荐相关资料，提出学习建议，成为学生的学习伙伴。

可以看到，人工智能的发展已深刻地改变着我们的生产和生活方式。它带来了显著的效率提升，也促进了社会生产力的进一步发展。同时，也应看到人工

智能带来的潜在风险，如隐私泄露、信息伪造、对传统岗位的冲击等。这些问题应该引起足够的重视，及早制定应对策略。

6 如何学习人工智能

人工智能并非单一技术，而是由庞杂的技术体系组成，这对初学者来说是个不小的挑战。此外，人工智能发展迅速，进一步加大了学习的难度。总体来说，学习人工智能应注意以下几点。

（1）打好基础：人工智能涉及较多数学知识，需要有一定的数学功底。随着人工智能与各个基础学科交叉融合，与其相关的学科基础也不能忽视。

（2）关注核心概念：应关注核心概念，如机器学习、人工神经网络等。通过这些概念的学习建立对人工智能的基础认知，然后再学习具体的算法。

（3）动手实践：人工智能是一门理论与实践相结合的科学，因此在掌握基础知识的前提下完成一些简单的实验，对加深理解有很好的帮助。

（4）关注前沿：人工智能是一个快速发展的领域，每天都有新的进展出现。要养成主动了解相关前沿的习惯，以便及时跟进最新的发展方向。

（5）讨论与交流：现代人工智能的突破离不开研究者们的分享与合作精神。在人工智能的学习中也应重视与他人的讨论与交流，通过小组协作等方式更好地掌握知识、激发创新思路。

小结

人工智能是用计算机模拟人类智能行为的科学。与其他学科相比，人工智能有独特的目标，即实现类似人的智能机器；有独特的工具，即用计算机来模拟人类的智能行为；有独特的方法，即将知识和数据相融合，通过学习获得完成任务的技能。目前，人工智能已经在我们的日常生活中广泛应用，并且已经和很多基础学科交叉共融，在各个专业领域大显身手，成为现代社会生产生活的基础工具。

1.2 人类智能的起源

学习目标

（1）理解人类智能的生物学基础，认识大脑进化过程中智能发展的关键特征与机制。

（2）了解人类智能进化的过程，认识从古猿到现代智人演化中的关键节点，包括直立行走、使用工具、语言产生等关键环节。

（3）探讨合作在智能进化中的重要性，理解群体合作、互信与共情如何推动人类智能的发展。

（4）认识文明形成与发展的过程，理解知识共享与累积（棘轮效应）如何促使人类社会产生“阶跃式”进步。

（5）思考人类智能的演化对人工智能发展的启示。

从20万年前烈日炎炎的非洲大陆到今天星光闪耀的现代都市，人类经历了一段漫长的文明之旅。在这段旅程中，人类与自然斗争，慢慢学会了建造城堡和高楼，创作出了动人的诗歌和音乐，发明了能够移山填海的机械，甚至把目光投向了遥远的宇宙，探索几十亿光年之外的奥秘。为什么人类拥有如此强大的智能？我们的智能是如何在历史长河中一步步进化而来的？本节将探讨人类智能的起源和发展之谜，回顾人类从原始社会的合作狩猎到现代文明的演进之路。

① 人类发展简史

为了理解人类智能的起源，首先需要回顾地球上生命的进化历程（图1-4）。大约45亿年前，地球从环绕早期太阳旋转的吸积盘中形成。距今42亿～40亿年前，地球表面温度逐渐降低，地壳凝固，大气与海洋形成。大约在40亿年前，最早的生命以简单的有机分子形式出现。随着时间的推移，这些原始的生命形式

逐渐演化成更加复杂的单细胞生物。大约5.8亿年前，海洋中出现了最早的动物，如海绵和水母。随后，生命形式不断多样化，出现了更复杂的生物，如节肢动物和软体动物。大约5.3亿年前，地球经历了寒武纪大爆发，生物多样性迅速增加，各类生物不断进化，涌现出了大量新的物种。目前地球上约有870万种生物，包括650万种陆地生物和220万种海洋生物。目前，有记录描述的物种大约有180万种。

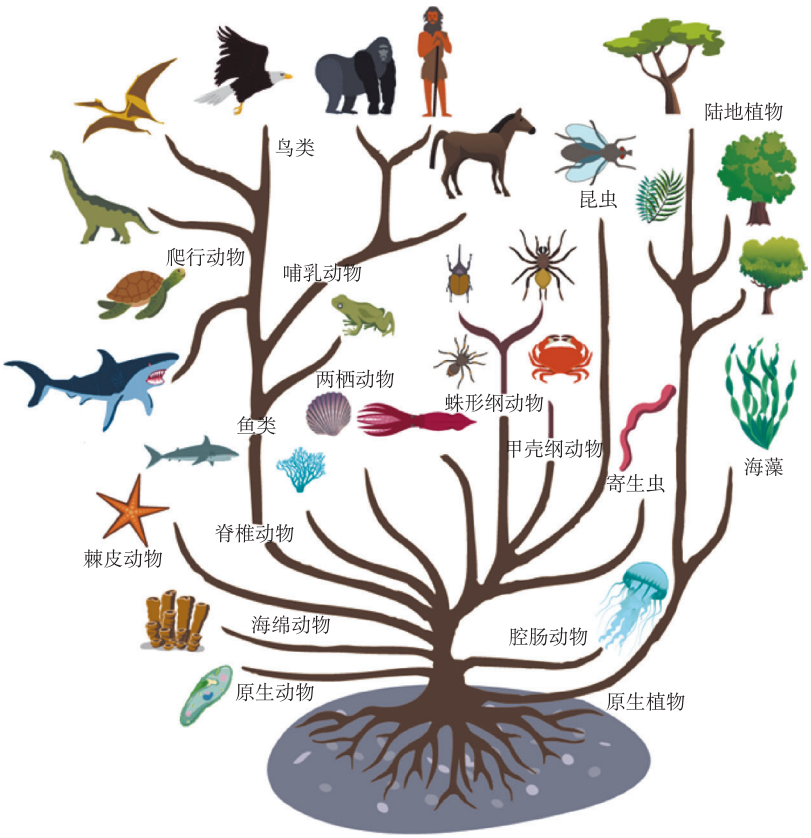


图1-4　地球生物进化树

如果把地球的生物进化过程浓缩成一天，那么人类的出现只相当于这一天的最后一秒。约600万年前，在非洲的某个地区，因为环境的变化，森林退化成草原，一群古猿不得不从树上来到地面，开始习惯直立行走。直立行走使他们能够更好地观察远方，同时解放出双手，做更多精细的事情。大约400万年前，这些古猿逐渐演变成一个新的种群，称为“南方古猿”。南方古猿是人类最早的祖先。

大约200万年前，一支被称为“能人”的古猿开始用双手制造石器，这是人类进化的重要一步。能人不仅能制造工具，还初步拥有了语言能力。大约180万年

前，能人逐渐进化为直立人。直立人是第一个真正直立行走的人类祖先，可以制作更复杂的石器，并开始用火煮食肉类。随着时间的推移，直立人开始向非洲以外的地区扩散。大约40万～30万年前，一支被称为“智人”的人类种群在非洲出现，他们具有更大的脑容量和更复杂的认知能力，因此在与其他古人类的竞争中逐渐占据了优势，成为现代人类的直接祖先。智人开始使用语言、制造复杂的工具、进行艺术创作，奠定了现代人类文明的基础。人类进化示意图如图1-5所示。

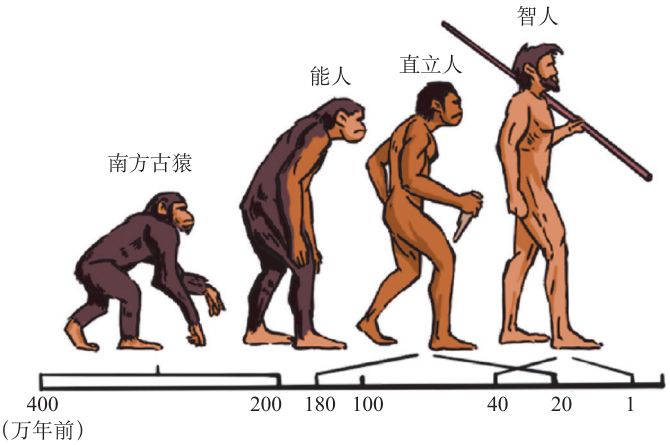


图1-5　人类进化示意图

② 人类为什么这么聪明

人类大约出现在200万年前。对地球生物演化而言，200万年是非常短暂的，人类如何在这么短时间内就进化出了无与伦比的人类文明呢？

科学家们认为，人类聪明的根本原因在于人类拥有一个强大的大脑。研究表明，大多数动物的大脑重量与身体重量通常呈正比增长，如图1-6所示。这一脑容量的增长主要是为了满足控制身体的需求，而非提升智力。如果某种动物的大脑重量与身体重量的比值更高，那么多余的脑容量将用于更高级的思维活动，从而表现出更高的智力水平。

我们可以通过大脑重量占身体重量的比例（脑化指数，encephalization quotient，EQ）来衡量一种动物的聪明程度。通常脑化指数越大，动物的聪明程度也就越高。计算表明，成年人的大脑大约为1.4千克，占身体重量的2%，如

图1-7所示，几乎是所有动物中比例最高的。

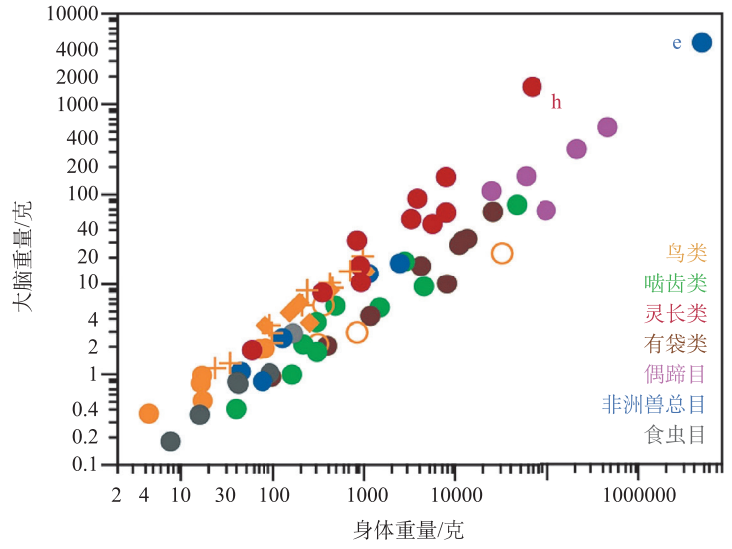


图1-6　动物大脑重量与身体重量关系图

h—人类；e—非洲象

图1-7　不同动物的脑化指数

总结来说，人类是通过头脑而不是四肢或牙齿来获得生存优势的。这一选择使我们的头脑越来越聪颖，智力也越来越强大。现在看来，这一选择是非常明智的：那么多庞然大物都消失在历史长河中，而人类成了这个星球的主宰。

一个有趣的问题是，为什么只有人进化出了这样强大的大脑，而处于相似进化起点的人类近亲们（如大猩猩、黑猩猩）却没有做到这一点呢？近年来，科学家们发现人类有一种其他灵长类动物不存在的基因，称为ARHGAP11B，这一

基因可以促进神经系统的发育。科学家们猜测，这一基因的存在可能是基因突变的结果。这一突变使人类拥有更大的脑容量（是黑猩猩的3倍），从而获得了更强的竞争优势。科学家们通过实验发现，如果将ARHGAP11B基因注入小鼠的大脑，小鼠的脑容量会显著增加，且神经元的密度也大幅提升。这表明，这些基因在促进大脑发育方面发挥了重要作用。

③ 人类智能阶跃之谜

虽然脑容量的大小在一定程度上可以解释人类智能的物质基础，但无法解释人类智能的全貌。首先，自南方古猿开始，人类的脑容量确实呈现出增长的趋势，这与人类越来越聪明的趋势相符。然而，自智人以来，人类的脑容量实际上是减小的，但无疑人类是越来越聪明的。这该如何解释呢？此外，海豚的脑化指数（5.3）和人类（7.5）相差不大，但其智力水平和人相比显然是天差地别的，也没有进化出人类这样的文明社会。

最难以解释的是，人类所发展出来的智能，不仅远超包括自己近亲在内的所有生物，还远远超出了自身的生存需要。这是件令人惊奇的事情：几乎所有动物的智能都是以生存需要为边界的，只要能够满足温饱、可以活下去就可以了，不会想到变得更聪明。人类却是例外，我们从未满足于温饱，而是一直在持续不断地探索自然，为整个族群创造更好的生存环境。

1）合作激发智能

科学家们对人类智能的阶跃之谜进行了长期研究。一些研究者认为合作是人类智能开始飞跃的起点，其中迈克尔·托马塞洛（图1-8）的研究具有代表性。他在《人类思维的自然史》一书中对此做了详细阐释。

图1-8　迈克尔·托马塞洛（Michael Tomasello）

设想这样一个场景：我们的祖先因为环境变化，无法再依靠采集果实生存，他们不得不开始捕猎。然而，他们没有强大的身体和尖利的牙齿，奔跑速度也

没有优势。为了生存，他们必须进行合作，一起捕捉跑得更快或更强大的动物。在这种合作中，他们需要制定策略、分工协作、彼此配合、共同承担风险，也需要不断交流、沟通并改进方案，从而锻炼了大脑的各种能力，激发了智能的快速提高。

特别是，人们在合作的过程中产生了语言。语言的使用不仅锻炼了人的记忆力，还提升了人们抽象思维的能力。通过使用符号化的语言，人们可以建立抽象的概念（如“时间”“智力”等），并讨论概念间复杂的因果关系。生物学研究也表明，语言需求可能推动了人类大脑区域的扩展和功能分化；语言的概念体系也强化了听觉、视觉、运动等多条神经通路的协作，从而带动整体认知能力的提升。

2）互信与共情

合作是很多群居动物共有的特性，但只有人类的合作激发了智能的飞跃，这又是为什么呢？这是因为人类的合作更加深刻，包括合作养育婴儿、分享狩猎经验等。这些行为不仅是为了自身利益，更是为了整个群体的利益。

这种深层次的合作本质上源于人与人之间深刻的认同感，即每个人会把其他人视为与自己具有同样思考方式的个体。这种认同感奠定了人类“共情”的心理基础，即通过换位思考理解他人的处境与苦难。当原始人看到一只野兽追赶另一个原始人时，即便互相不认识，他也会力所能及地提供帮助。因为他会设身处地着想，想到自己被野兽追赶时的恐惧和面临的可怕后果。这种设身处地为他人着想的心理称为共情。因此，我们的祖先愿意帮助他人、信任他人、分享成果、分享经验，必要时甚至为他人和集体做出牺牲。

人类之所以能够养成这种无私的品质，可能是因为当时的生存环境极其恶劣，只有具备这种特质的个体和群体才能更好地生存下来。而那些过于自私自利的人则在自然选择中被早早淘汰了。因此，生存下来的人类天然具有互信互爱的高贵基因。

今天，我们看到很多人忘我地工作并不是为了自己有多好的生活，而是为了社会的发展做出自己的贡献。这正是源于他们对自己的国家和社会有深刻的认同感，对和自己同类的其他人有天然的信任感。他们知道自己的付出会受到尊重，自己的成果会让更多人受益，因此殚精竭虑、无怨无悔。人类的这种“认同种群、服务他人”的天然情感倾向解释了为什么人的智力会远远超出生存的需

要：很多人在努力学习、勤奋工作的时候，目标不仅是个体的生存，还有国家、社会和整个人类的进步。这是其他动物难以想象的行为模式。

3）人类文明的诞生

互信让人类的先辈们乐于合作，并带动了人类个体智能水平的提升，这一步虽然重要，但还不能实现人类智能的阶跃。真正起到决定性作用的是这种互信与合作，推动了人类作为一个整体的累积式演进，从而建立起璀璨的人类文明。这是比人类个体智能提高更重要的事，是人类智能阶跃式进步的真正开端。

首先，基于人与人之间的互信，人们愿意把自己的经验和知识分享给他人，而获得这些经验与知识的人也倾向于相信对方没有欺骗自己，故而乐于学习和接受。获得他人的传授之后，人们也愿意在前人的基础上继续贡献。这样就形成了一种“棘轮效应”，每一代种群所创造的成果得以保存并被后代持续改进，一点点积累起来，保证了文明的齿轮始终是向前进的。

正是基于这种积累和改进，人类慢慢发展出了文字、宗教、艺术乃至现代科学。新生的人类在新的知识环境中不断学习并创造出更优秀的智力成果，一步步推动文明的进步，反过来也促进了自身的头脑与能力的持续进步。因此，人类的智能已经不仅是单一的思维能力，而是通过一代代积累所获得的知识与视野。人类文明的技术演进如图1-9所示。

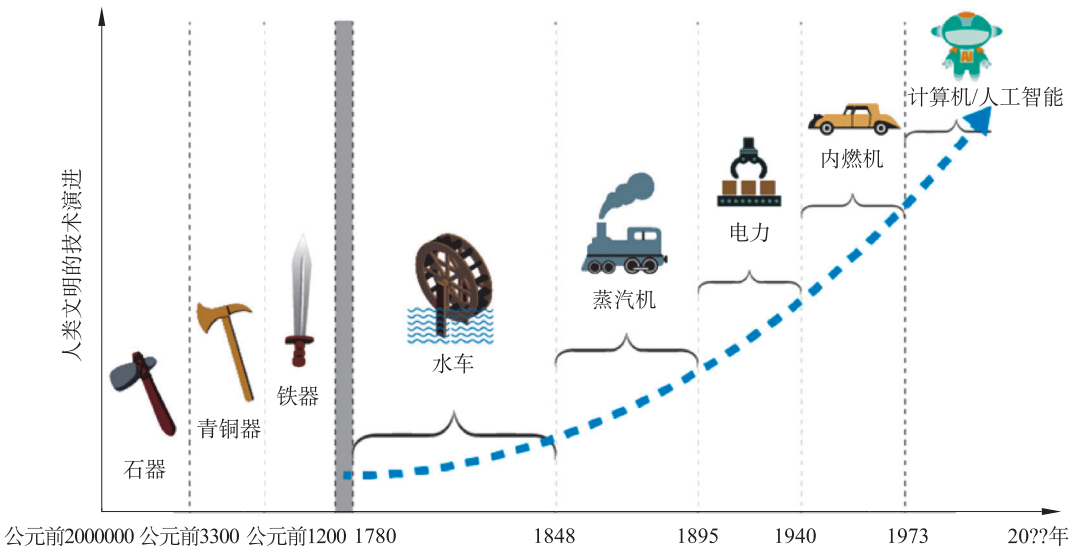


图1-9　人类文明的技术演进

小结

从生物进化到文明演进，人类智能的发展是一个充满奥秘而精彩的过程。在这一过程中，起到关键作用的是人类对同伴的天然认同感，由此产生了人与人之间的互信与共情，这成为人类互相合作乃至形成社会和国家的基础。人与人之间的合作是深刻的，不仅满足了个人的生存需要，而且超出了个体私利，以接纳、创新的心态为整个种群贡献自己的力量，从而铸就了今天伟大的人类文明。

关于人类智能起源的探讨让我们对人工智能的发展有了更深刻的认识。首先，智能需要有物质基础，人类的大脑就是人类智能的物质基础。今天，人工智能的一个基础思路就是借鉴人类大脑的工作方式，通过模拟大脑中的神经元网络实现了强大的智能。其次，作为人类整体，个体的智能只是起点，更重要的是个体之间的认同、合作与群体积累。这也启发人工智能的学者们开始思考如何开启机器智能的群体演化之路。未来，机器也会互相合作，共同探索，开启人工智能的新篇章。

1.3 人工智能的起源：数理逻辑

学习目标

（1）理解形式逻辑的基本概念，掌握亚里士多德的三段论及其在推理中的作用和局限性。

（2）认识思维数学化的历史意义，了解霍布斯、莱布尼茨、布尔等人的贡献，特别是布尔代数的基本原理。

（3）认识数理逻辑的建立过程，了解弗雷格、罗素、希尔伯特、哥德尔等人对逻辑体系完善的贡献。

（4）思考人工智能的核心目标，理解“模拟人的思维能力”这一研究初心及其面临的挑战。

人工智能是一门既古老又崭新的科学。作为一门独立学科，它的历史只有六七十年，但它的源头可以追溯到两千多年前亚里士多德所建立的逻辑学。逻辑学总结了人类的思维过程，形成了理性的思维框架，从而为机器模拟人的思维过程提供了理论基础。20世纪40年代以后，电子计算机出现，人工智能有了强大的计算工具，才最终实现了模拟人类思维的梦想。今天的人工智能已经超越了对逻辑思维的模拟，开始全面模拟人类的感知、认知、联想、创造等各种复杂的智能行为。人工智能始于逻辑推理，成于逻辑演算，并拓展到了全面的人类智能。本节将以“思维”和“计算”作为两条主线，来回顾人工智能的历史起源以及这一过程中那些站在地平线上的伟人。

① 开端：形式逻辑

要理解人工智能的起源，我们首先要回到2300多年前的古希腊，了解伟大的哲学家亚里士多德和他所构建的逻辑学。逻辑学研究的是人类的思维规律，这是让机器复制人类智能的第一步。

亚里士多德（Aristotle，公元前384—公元前322）是古希腊著名的哲学家和博学家，柏拉图的学生、亚历山大大帝的老师。他在众多领域做出了开创性工作，包括逻辑学、伦理学、政治学、经济学、天文学、物理学、心理学、生物学、地质学等。亚里士多德是人类历史上极具天赋的伟人，是古希腊科学发展的代表性人物。至今哈佛大学的校训依然是“与柏拉图为友，与亚里士多德为友，与真理为友”，他“吾爱吾师，吾更爱真理”的名言至今激励着一代代学子勇往直前，打破权威，探索未知世界。

亚里士多德在他的著作《工具论》中提出了今天我们称为“三段论”的思维规律。如图1-10所示，三段论通过一个大前提和一个小前提进行推理。例如，大前提是“所有人都会死亡”，小前提是“苏格拉底是人”，因此可以推理出“苏格拉

底会死”。亚里士多德认为，所有理性的人都会承认这一推理过程是正确的、毋庸置疑的。因此，如果大前提和小前提是正确的，那么结论必然是正确的。

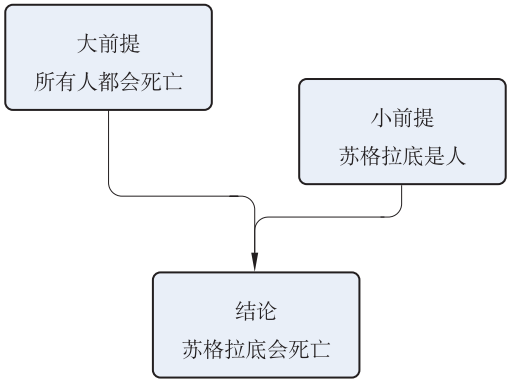


图1-10　三段论举例

三段论看似简单，但它将思维形式（过程）和思维对象（内容）区分开来，是人类对自身思维规律的第一次理性总结。亚里士多德的工作奠定了形式逻辑的基础，使我们能够用理性的方式理解人类的思维，也为人工智能模拟人的思维过程提供了可能性。

值得强调的是，三段论是一种推理工具，它只保证推理过程是正确的，但并不保证推理结果是正确的。例如下面的推理：

大前提：所有的鸟都会飞
小前提：企鹅是鸟
结论：所以企鹅会飞

显然，企鹅是不会飞的。是三段论错了吗？不是，是因为大前提错了，并不是所有的鸟都会飞。这体现了思维形式（过程）和思维对象（内容）之间的独立性。思维形式保证“如果大前提和小前提都成立，则结论成立”，但并不保证大前提和小前提本身的正确性；大前提和小前提是思维的对象，其正确性需要独立验证。

总而言之，三段论把思维形式和思维对象进行了区分，这是人类认识自我的历史性飞跃。从此以后，人类才开始科学地认识自己的思维，梳理思维规律，填补思维漏洞。这奠定了逻辑学的基础，也为哲学和科学的发展奠定了理性思

维的基石。同时，逻辑学的诞生，也为机器模拟人类思维提供了可能性，成为人工智能最初的起点和源头。

② 进阶：思维的数学化

亚里士多德的形式逻辑多以自然语言表述，容易产生歧义。例如，“如果你获得馈赠，那么你应该感谢”，这里“应该”到底是强制要求，还是道德上的劝说？在自然语言中并不完全清晰。

思维的数学化就是用符号表示事实和命题，用符号演算表示思维过程。如果这些符号定义精确，演算规则清晰可靠，那么任何人面对同样的符号和演算规则无论过程多么复杂，都会得出一致的结论。这就是思维数学化的意义。

英国哲学家托马斯·霍布斯（图1-11）在其著作《利维坦》一书中提出，人类的思维可以表示为一个数学计算过程，简单地说，“推理即计算”。随后，德国哲学家戈特弗里德·莱布尼茨（图1-12）在《发现的艺术》（1685）一书中同样主张用数学来表达思维。他写道：“如果人们发生了争执，那么很简单。来，让我们来算算，看看谁是对的。”

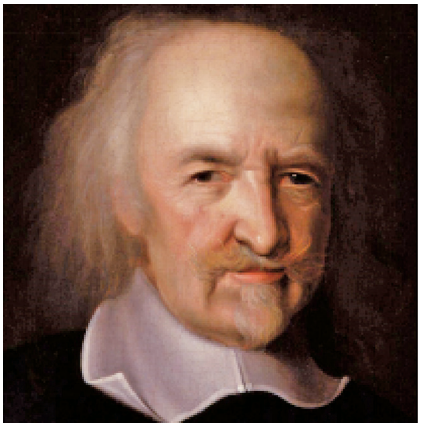


图1-11　托马斯·霍布斯（1588—1679）

图1-12　戈特弗里德·莱布尼茨（1646—1716）

思维数学化的目的是对思维过程进行精确、无歧义地描述。然而，直到19世纪，数学家乔治·布尔才发明了描述思维的数学工具——布尔代数，用数学符号来表示事实，用逻辑运算来表示思维推理，奠定了现代计算机科学和人工智能的基础。

③ 完善：数理逻辑的确立

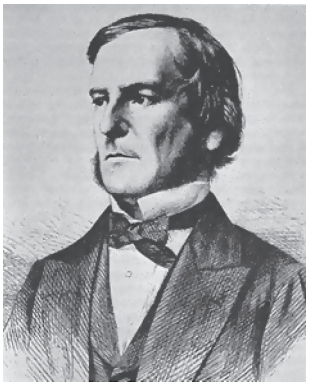


图1-13　乔治·布尔（1815—1864）

1854年，英格兰数学家乔治·布尔（图1-13）出版了《思维规律》一书，完成了逻辑符号化的开创性工作。他用符号代表事实，用符号演算表示从既有事实到未知事实的推理过程。这一体系被后人称为布尔代数。布尔的工作证明了基于明确定义的符号和运算规则，可以表达形式逻辑的推理过程，从而模拟人的思维。因此，他在《思维规律》一书的序言中写道："本书论述的是探索心智推理的基本规律。"

在布尔的演算系统中，事实用字母表示，如 q、p、r。这些字母只有"是"和"否"两种取值，分别表示事实成立或不成立。布尔将"是"和"否"分别表示为1和0。进一步，布尔定义了×和+两种演算，分别表示逻辑中的"并且"和"或者"两种关系；定义了符号−，表示逻辑中的"非"；定义了符号=，表示逻辑上的"推论"。基于这些定义，就可以确立一系列演算规则，如图1-14所示。这些演算规则与算术中的加法和乘法规则类似，只不过布尔在这里定义的是逻辑演算，而不是数量演算。

例如，用p表示"明天下雨"，q表示"明天刮风"，r表示"明天下雪"，则命题"明天下雨或刮风，且下雪"就可以表示成$(p+q)\times r$。根据图1-14中第五个演算规则，可以对$(p+q)\times r$进行推论，得到$(p\times r)+(q\times r)$，即"明天下雨且下雪，或者刮风且下雪"。

$$p\times q=q\times p$$
$$p+q=q+p$$
$$p\times(q+(-q))=p$$
$$p+(q\times(-q))=p$$
$$p\times(q+r)=(p\times q)+(p\times r)$$
$$p+(q\times r)=(p+q)\times(p+r)$$

图1-14　布尔演算规则

布尔开创了用符号和符号演算表示逻辑过程的先河。随后，弗里德里希·弗雷格（图1-15）在《概念文字》一书中进一步完善了布尔的演算系统。他定义了"任何""存在"这样的量词，极大地扩展了布尔代数的表达能力，不仅可以表示我们基本的逻辑推理过程，还可以通过

逻辑演算完成“三加五等于八”等更通用的计算。后来，经过阿佛列·怀特黑德（图1-16）、伯特兰·罗素（图1-17）、大卫·希尔伯特（图1-18）、库尔特·哥德尔（图1-19）等数学家的努力，数理逻辑正式确立。

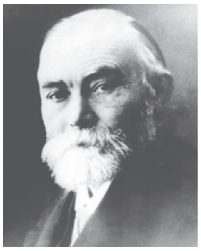


图1-15　弗里德里希·弗雷格（Friedrich Frege，1848—1925）

图1-16　阿佛列·怀特黑德（Alfred Whitehead，1861—1947）

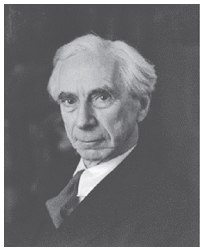


图1-17　伯特兰·罗素（Bertrand Russell，1872—1970）

图1-18　大卫·希尔伯特（David Hilbert，1862—1943）

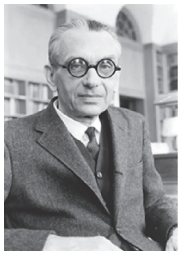


图1-19　库尔特·哥德尔（Kurt Gödel，1906—1978）

数理逻辑的建立，为形式化、精确地描述人类的思维提供了坚实的理论支撑，也成为人工智能学科的第一块基石。

小结

人工智能是用计算机模拟人类智能行为的科学。人工智能起源于人类对自身思维规律的探索，这一探索最早可以追溯到古希腊时代，当时一大批智慧的先贤们开始了对世界的理性思考。比如，毕达哥拉斯对“数”的看重，认为稳定、完美的世界应该表示为分数，比如行星的轨道、音乐的音阶。苏格拉底对“普遍定义”进行了深入探讨，认为普遍定义是一种必然的、确定性的知识，只能通过理性的思维活动来触达。柏拉图对于“理念世界”

有更深刻的认识。它认为存在一个完美的理想国度，我们的世界只是这个理想国度的投影，只有通过智能才能达到这个理想的世界。亚里士多德继承了这些前辈们对于理性思考的重视，并开始研究这些理性背后的思维规律。这是逻辑学的起点，也是人类对自身的思维能力的第一次系统性总结。亚里士多德以后，科学家们沿着他开创的道路继续探索，完成了逻辑形式化、数学化的伟业，最终建立起了数理逻辑，成为人工智能的第一块基石。

通过学习人工智能的起源，我们可以清晰地看到这门学科建立的初衷，理解它区别于其他学科的独特之处，从而建立清晰的学科边界。事实上，人工智能从一开始建立，其目的就是为了模拟人的思维能力，即人们“动脑子”的能力。这是一个非常宏大的目标，因为人类之所以能远远超出其他物种成为万物之灵，根本上就是因为我们有聪明的头脑。如果机器真的能模仿人类的动脑能力，意味着人能做到的事机器都可以做到，包括发展出新的科学技术。从这个角度上看，人工智能在目标上是超越现有学科的，是“科学之上”的科学。当然，这一目标的实现也是极为艰难的，直到今天也没有完全实现。这也是人工智能在历史上饱受质疑甚至嘲讽的原因，甚至在人工智能已经取得了辉煌成就的今天，质疑和观望也依然存在。无论如何，模拟人的思维能力就是人工智能研究者的最初动因和终极理想。

1.4 人工智能的起源：计算机的诞生

学习目标

（1）理解图灵机的基本概念与工作原理，认识其在计算理论和人工智能发展中的重要性。

（2）认识数字电路的基本原理与逻辑门的概念，理解香农如何将布尔逻辑应用于计算机电路的设计。

（3）了解计算机诞生的历史过程，认识ENIAC和冯·诺依曼存储程序结构的关键意义。

（4）明确计算机与人工智能之间的关系，理解计算机如何支持人工智能的发展。

数理逻辑的建立奠定了人工智能的第一块基石，使得人类的思维可以用计算的方式来模拟。然而，实现这种模拟还需要一种强大的计算机器。历史上，人们设计了很多计算机器，比如中国古代的算盘。但是，这些机器并不是通用的计算机器，只能用于特定的计算任务。这一问题直到英国数学家、逻辑学家艾伦·图灵提出图灵机模型后才解决。图灵机模型成为计算机的理论模型，后来经过艾伦·图灵、香农、冯·诺依曼等科学家的努力，最终研制成功通用电子计算机，为人工智能的发展奠定了第二块基石。

1 图灵机模型

1936年，年仅24岁的英国科学家艾伦·图灵（图1-20）提出了一种称为图灵机的计算模型。这一模型展示了通过简单的读/写操作可以处理极为复杂的逻辑演算。

图1-20　艾伦·图灵

艾伦·图灵的基本思想是用机器来模拟人们用纸笔进行数学运算的过程。他把这样的过程看作下列两种简单的动作：①在纸上读出或写上某个符号；②把注意力从纸的一处移动到另一处。在这一计算过程，下一步要采取什么样的动作，依赖于纸上当前所关注位置的符号和当前思维的状态。

为了模拟人的这种运算过程，图灵设计了这样一台假想的机器，该机器由以下几个部分组成。

- 无限长的纸带：纸带被划分为一个接一个的小格子，每个格子上包含一个来自有限字母表的符号（一般为二进制数0或1）。纸带上的格子从左到

右依次被编号，纸带的两端可以无限伸展。

- 读写头：读写头在纸带上左右移动，能读出纸带上当前格子的符号，也可以往格子里写入符号。
- 状态寄存器：用于保存机器当前所处的状态。可能状态的总数是有限的，并且有一个特殊的状态，称为停机状态。
- 控制规则表：规定在每个状态下，读写头在读取特定符号后所采取的行动，包括读什么符号、写入什么符号、向哪个方向移动读写头、更新状态等。这个规则表定义了机器运行的方式，更改了规则表也就改变了机器的运行方式。从现代眼光来看，这个规则表即这台机器的程序。

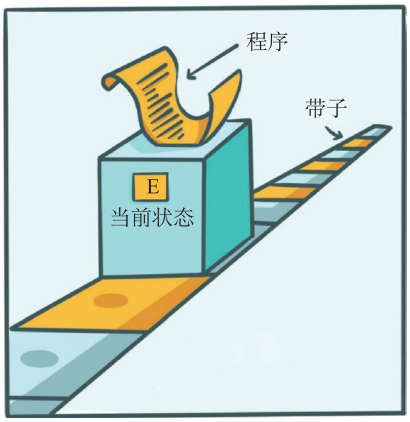


图1-21　图灵机

根据图灵导师阿隆左·邱奇的建议，这台假想的机器被称为图灵机（图1-21）。

表面上看，图灵机非常简单，那么这种机器的计算能力如何呢？或者说，它能计算哪些函数呢？图灵和其他研究者很快发现，图灵机非常强大，人们所设计的各种复杂的计算模型最后发现都弱于或等价于图灵机。注意，这里的“强大”不是指计算的速度有多么快，而是指图灵机所能代表的计算函数非常广泛，可以涵盖任何可以想象到的计算过程。为此，斯蒂芬·克莱恩（Stephen Kleene）提出了著名的邱奇-图灵论题：一切直觉上可计算的函数都可用图灵机计算。值得说明的是，这是个论题而非定理，并没有严格的证明，直到今天人们还没有发现超越图灵机的计算模型，因此被科学家们普遍接受。

图灵机这种强大的计算能力具有重要意义，它意味着科学家们不用再尝试构造各种复杂的计算机器了，只要把图灵机实现，就能计算所有可计算的函数了。这就为通用计算机器奠定了理论基础。事实上，在现代计算机科学中，可计算函数也是由图灵机定义的，由于图灵机能够处理所有“直觉上可计算”函数。因此，图灵机能计算的函数就被认为是可以被计算的函数。

② 数字电路

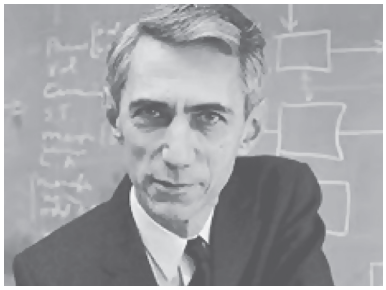
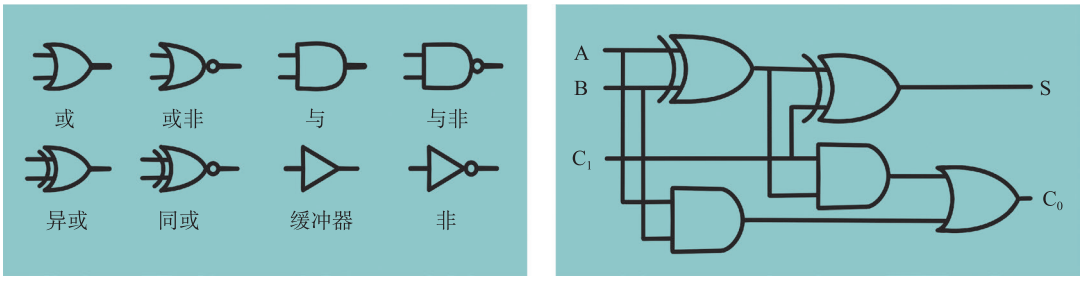


图1-22　克劳德·艾尔伍德·香农（1916—2001）

1937年，年仅21岁的麻省理工学院研究生克劳德·艾尔伍德·香农（图1-22）提交了他的硕士论文《继电器和开关电话的符号分析》。在这篇据称是“有史以来最重要的硕士论文”中，香农提出基于布尔逻辑设计电路的新方法。香农的研究表明，用电子开关的组合可以模拟布尔运算，从而实现复杂的逻辑演算过程。具体来说，香农利用继电器和开关设计出了一种可以执行逻辑运算的“数字电路”系统（图1-23），这种系统以开关的“通”或“断”来表示布尔逻辑中的0或1，通过开关电路的组合实现与、或、异或等基本逻辑操作，而这些逻辑操作组合起来就可以实现各种复杂的运算，包括各种控制过程和加、减、乘、除等数学计算。



(a)　　(b)

图1-23　可以执行逻辑运算的“数字电路”系统

香农的工作将逻辑运算和数字电路联系起来，为数理逻辑的硬件实现奠定了基础。原则上说，只要设计好了逻辑演算过程，就可以用相应的门电路来实现它。从简单的交通信号灯控制电路到复杂的超大规模集成电路，都离不开这一原理。

值得说明的是，图灵机与香农的数字电路理论密切相关。图灵机定义了一种机械计算过程，读写头在纸带上的读取、写入、移动等操作都可以视作逻辑运算。因此，图灵机的基本组件都可以用数字电路实现。另外，加、减、乘、除等数

学计算过程也可以表示为逻辑操作，同样可以用数字电路实现。因此，图灵机定义了通用计算机的计算方式，而香农的数字电路理论则提供了实现这些计算的物理方式。这两个理论共同奠定了现代计算机的基础。

需要强调的是，数字电路技术应用广泛，并不局限于用来实现计算机。事实上，在很多应用中只需要实现简单的控制逻辑即可（如交通信号灯的控制），并不需要实现一个完整的通用计算机。

③ 计算机的诞生

在图灵和香农理论的启发下，科学家们开始尝试设计和建造电子计算机。1943年，英国科学家设计并建造了Colossus计算机，用来破解德军密码，这是世界上第一台可编程电子数字计算机。Colossus的成功展示了电子计算机在处理复杂计算任务上的巨大潜力，但尚未实现通用性。1946年，第一台通用电子数字计算机ENIAC（electronic numerical integrator and computer）（图1-24）在美国宾夕法尼亚大学诞生。这台计算机是由宾夕法尼亚大学的约翰·莫奇利（John Mauchly）和约翰·皮斯普·埃克特（John Presper Eckert）主持设计，重达27吨，耗电150千瓦，占地167平方米，是一个庞然大物。ENIAC采用十进制运算，每秒可执行大约5000次加法操作、385次乘法操作、40次除法操作。这一计算速度远比

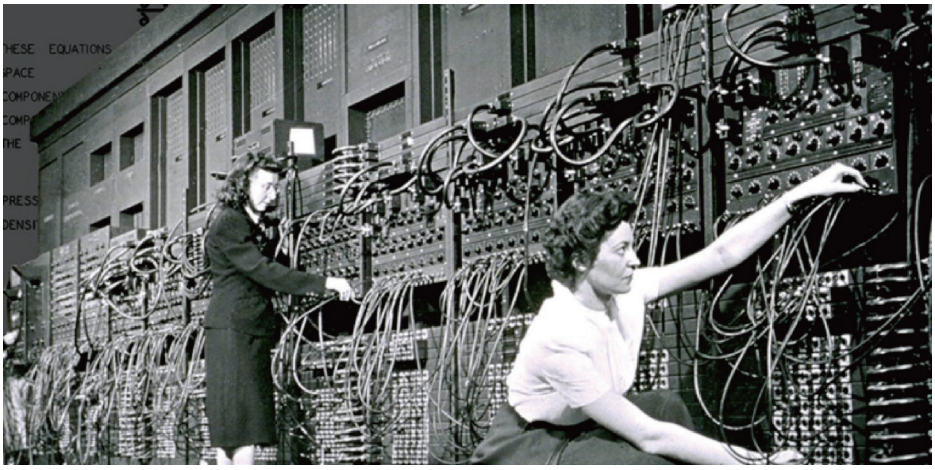


图1-24　第一台通用电子数字计算机ENIAC

注：图中女士正在通过插拔的方式对ENIAC进行“编程”，改变它的计算任务。

现在的手机要慢，但却开启了通用计算机时代。

1945年前后，以约翰·冯·诺依曼为代表的科学家们逐渐确立了计算机设计的基础原则，明确使用二进制计算，并将程序作为一种特殊的数据存储在存储器中，在需要运行的时候将程序读取出来，因此称为存储程序结构。这种结构的优点是可以很方便地对程序进行修改，而不必像ENIAC那样通过插拔电缆来实现。从此以后，编程才变得简单起来。

此外，冯·诺依曼等还将计算机明确分成运算器、控制器、存储器、输入设备和输出设备五大组件。计算机通过输入设备读入数据，由控制器读取指令，并送入运算器中计算，最后由输出设备输出结果，如图1-25所示。这一模块化设计奠定了现代计算机体系结构的基础。

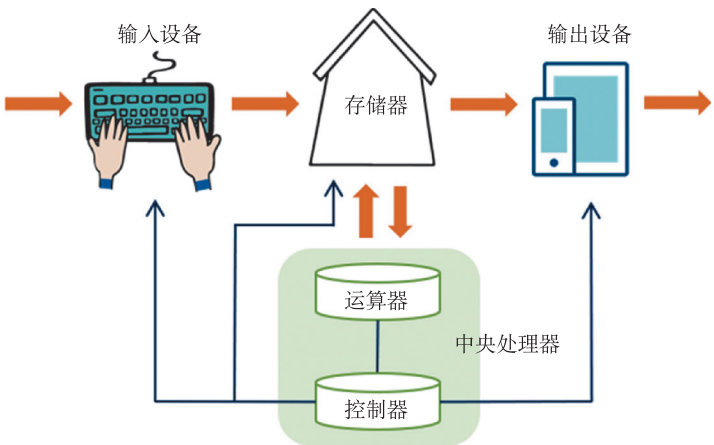


图1-25　现代计算机的基础架构

1948年，第一台基于存储程序结构的计算机Manchester Baby在曼彻斯特大学研制成功。这台计算机的设计初衷并非建造一个实用的计算引擎，而是用于测试一种称为“威廉斯管”的存储设备。1949年，第一台实用的基于存储程序架构的电子计算机EDSAC（electronic delay storage automatic calculator）在英国剑桥大学问世。EDSAC由莫里斯·文森特·威尔克斯教授领导设计和制造，并于1949年投入运行。从此以后，计算机飞速发展，推动人类社会进入信息时代。

小结

计算机的诞生是人类科技史上一次革命性的事件，而图灵是这场革命的揭幕人。他所设计的图灵机模型不仅论证了通用计算机的可行性和它强大的计算能力，同时也为计算机的实现提供了原型。香农是这场革命中的另一位关键人物，他提出的用门电路实现逻辑演算的思路奠定了数字电路的理论基础，也为用电子电路实现通用计算机提供了思路。在这些理论的指引下，经过无数科学家的努力，在20世纪40—50年代，通用电子计算机诞生，开启了人类历史的新篇章。

计算机的出现是人工智能诞生的第二块基石。自从有了计算机，人工智能的先驱者们“用计算模拟人类思维”的理想就有了强大的工具。于是，让机器模拟人类智能的想法再次浮现在图灵天才的头脑中，我们将在下一节具体介绍。

1.5 图灵：人工智能之父

学习目标

（1）了解图灵的生平与重要贡献，认识他在计算机科学和人工智能领域的奠基作用。

（2）掌握图灵机的工作原理及其对计算理论发展的重要意义。

（3）认识图灵对机器智能的早期思想，理解他提出的机器学习、强化学习、演化学习等概念的早期雏形。

（4）掌握图灵测试的基本原理，理解其在人工智能领域的重要意义。

（5）思考图灵的影响力，了解图灵奖的设立背景及其对计算机科学和人工智能发展的推动作用。

艾伦·图灵，英国数学家、逻辑学家、计算机学家，被誉为“计算机科学之父”和“人工智能之父”。他不仅对现代计算机科学的诞生起到了奠基性的作用，也对人工智能的诞生作出了重要贡献。本节将与读者一起回顾这位伟人的非凡人生，重温他在人工智能领域的三大贡献：图灵机及其可计算理论、对机器智能的开创性思考与实践，以及提出的图灵测试。

① 少年天才

1912年6月23日，艾伦·图灵出生于英国伦敦。他从小就展现出非凡的天赋。在图灵读小学时，他的老师曾说过：“我见过不少聪明勤奋的孩子，然而，艾伦是个天才。”1926年，图灵被父母送到伦敦的谢伯恩公学寄宿就读。在谢伯恩公学的学习岁月中，图灵表现出对科学的浓厚兴趣，并自学了爱因斯坦等科学家的著作。

1931年，图灵考入剑桥大学，由于成绩优异，获得了数学奖学金。在剑桥大学，他的数学能力得到了充分的发展，被授予数学一等奖。1934年，他提交了毕业论文《论高斯误差函数》，提出了一种证明中心极限定理的新方法。这一论文使他当选为国王学院的研究员（fellow），并于次年荣获英国著名的史密斯数学奖，成为国王学院声名显赫的毕业生之一。

② 贡献一：图灵机

1936年，年仅24岁的图灵发表了一篇划时代的论文《论可计算数及其在判定问题上的应用》。这篇论文旨在证伪希尔伯特提出的可判定问题，即是否存在一种通用算法能够判断任意数学命题的真伪。为此，图灵提出了图灵机这一通用计算模型，开启了计算机科学的先河（见1.4节）。

在图灵机的设想中，有一条可供读写的无限长纸带，让它虽非真实的机器，却能在功能上与一台真实的机器几乎等效。按照图灵的设计，设计出一台现实的物理计算机并不存在理论上的困难。特别有价值的是，通过修改图灵机的控制规则表（即程序），图灵机可以完成所有可以想象到的计算。这意味着，只要

可以实现图灵机，就可以得到一台通用的计算机器，应对所有领域的复杂计算。因此，图灵机的提出大大激发了人们设计通用计算机的信心。不仅如此，图灵机中所引入的存储区、程序、控制器等概念直接启发了冯·诺依曼等的存储程序结构设计，奠定了现代计算机架构的基础。

对人工智能而言，计算是最基本的支柱，没有计算机的诞生，也就没有人工智能的开端。从这一点上看，图灵为人工智能的发展准备了必要的计算工具。

③ 年轻的密码学家

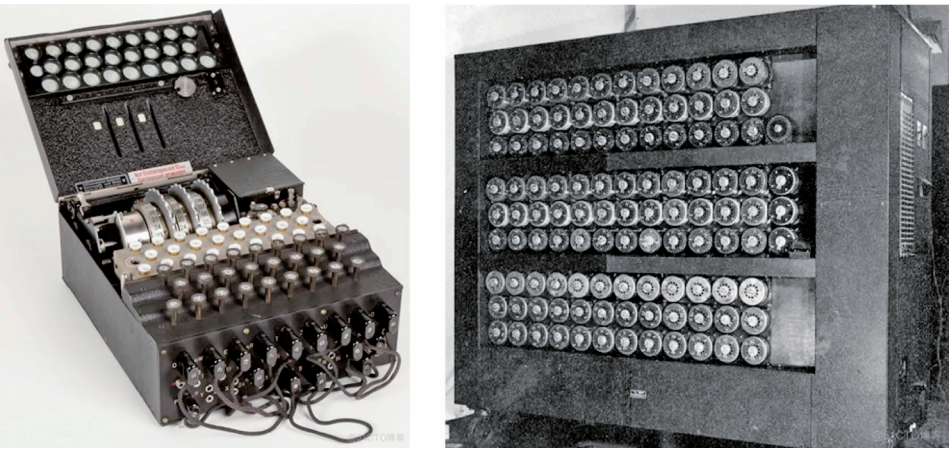
1936年9月，图灵远赴美国，在普林斯顿大学攻读博士学位，师从数学家阿隆佐·邱奇。1938年6月，图灵获得普林斯顿数学系博士学位，他的论文基于序数的逻辑系统，介绍了序数逻辑和相对计算的概念。同年，图灵婉拒了冯·诺依曼的挽留，毅然回到英国，投身对抗法西斯德国的战斗。

回国后不久，图灵就参与到政府的密码破译项目中，和全国各地顶尖的数学家们一道在白金汉郡的布莱切利公馆破译德国密码。图灵破解了升级版的Enigma密码机（图1-26），并探索出了一套高效的破译算法。据估计，图灵的工作使战争的结束时间提早了两年，挽救了上千万人的生命。

④ 贡献二：机器智能的最初思考与实践

1948年，图灵成为曼彻斯特大学的讲师；1951年，当选为英国皇家学会会员。同年，图灵发表了一篇题为《智能机器》的报告，首次提出了机器智能的可能性，并探讨了若干具体实现方式。这篇开创性的报告被视为人工智能正式登上历史舞台的先声。

在报告中，图灵认为可以设计一个通用机器，像教育儿童那样教它一步步成长，这是机器学习的朴素思想。他还提出，可以通过奖励和惩罚来对机器进行“教育”，这是强化学习的基本思路。所谓强化学习，是指通过间接的奖励信号来进行学习的方法，就像小时候学习走路，父母并没有告诉我们如何迈步，但当我



（a）德国军队使用的 Enigma 密码机　　（b）图灵设计的密码破译机 Bombe

图1-26　Enigma密码机与密码破译机Bombe

们每一次尝试成功后，父母会给我们各种鼓励，这样就慢慢学会了走路。

图灵甚至提出了通过模拟生物进化来实现智能的方法，成为演化学习思想的最初萌芽。生物进化是自然选择的结果，包括人的进化。生物进化在智能的产生过程中扮演着重要角色。图灵认为，模仿这一过程是一种让机器产生智能的可能方案。

图灵的《智能机器》报告展示了他对机器智能的深刻见解和远见卓识。他提出的许多概念和方法成为人工智能研究的基础，并在后来的发展中得到了广泛应用。

总结一下，图灵的天才思想是人工智能发展之初的第一笔精神财富，直到今天依然指导着后人。从这一点来说，图灵为人工智能的发展奠定了最初的思想基础。

⑤ 贡献三：图灵测试

1950年，图灵发表《计算机器与智能》一文，提出了图灵测试这一假想实验。图灵测试是判断机器是否具有人类智能的一个标准。其基本思想是通过人与机器之间的对话来判断机器是否拥有了智能。图灵测试的基本形式如图1-27所示。

（1）测试员：一名人类测试员 C。

（2）被测试者：一名人类 B 和一台机器 A。

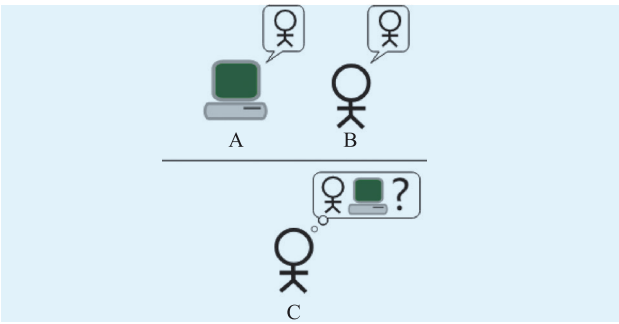


图1-27　图灵测试示意图

（3）对话形式：测试员 C 通过键盘与被测试者 A/B 进行自然语言对话，测试员不知道谁是人类，谁是机器。

（4）测试时间：通常设定为5分钟。

（5）判断标准：如果测试结束后，有30%以上的测试员误认为它是人类，则认为该机器通过了图灵测试，具备了智能。

图灵测试的重要意义在于它为“机器智能”提供了一条实践可行的衡量标准，从而让研究者摆脱了“智能”定义上的争执，设定了人工智能研究者努力的方向。从这一点来说，图灵为人工智能的发展指明了方向。

⑥ 百年影响

2012年，在图灵诞辰百年之际，《自然》杂志称他为有史以来最具科学思想的人物之一。2021年，英格兰银行发行的新版50英镑纸币上印有图灵的头像，表达对这位伟人的敬仰。

为了纪念图灵，全球计算机专业的权威组织——美国计算机协会（ACM）于1966年设立了图灵奖（图1-28），用以表彰在计算机领域作出卓越贡献的学者。该奖项被誉为计算机界的诺贝尔奖。

自1966—2024年，全球共有78位科学家获得图灵奖，涵盖编译原理、程序设计语言、计算复杂性理论、人工智能等领域。2000年，清华大学教授姚期智（图1-29）因在计算理论、密码学等方面的基础性贡献获图灵奖，这是目前唯一获此殊荣的华人科学家。

图1-28　图灵奖奖杯

图1-29　2000年图灵奖得主姚期智先生

小结

图灵提出了图灵机模型，为计算机的诞生奠定了理论基础，同时也为人工智能准备了计算工具。他关于机器智能的最初设想，为人工智能的发展奠定了思想基础。他提出的图灵测试，从可验证的视角定义了智能，为人工智能的发展指明了方向。图灵的贡献不仅奠定了计算机科学的基石，还对人工智能的诞生和发展产生了深远影响，是人工智能当之无愧的奠基人。

1.6 人工智能的开端

学习目标

（1）了解人工智能早期研究的内容与方法，包括对弈算法、定理证明和早期神经网络等关键探索。

（2）认识达特茅斯会议的历史意义，明确该会议对人工智能学科形成的推动作用。

（3）思考学术交流与跨学科合作在科学进步中的重要性，理解其如何促进人工智能的发展。

1954年，艾伦·图灵离世，但他点燃的机器智能的火种却并未熄灭。就在图灵离世后的两年，一群年轻的科学家在美国达特茅斯学院数学系的一幢小楼里组织了一次长达两个月的讨论会，在这次会议上，人工智能作为一门新科学正式登上历史舞台，从此开始了近七十年的风雨历程。这就是人工智能史上著名的达特茅斯会议，也是被学者们公认的人工智能的开端。

① 风起云涌

20世纪50年代，通用计算机刚刚诞生，其强大的计算能力引起了研究者的广泛关注。另外，随着数理逻辑的发展，“思维即计算”的理念已经深入人心。受图灵机器智能思想的启发，利用计算机来模拟人类思维、实现类似人的智能机器，极大地激发了年轻学者的研究热情。

受此影响，一批新的研究成果涌现，包括克劳德·香农的对弈算法、赫伯特·西蒙和艾伦·纽厄尔的“逻辑理论家”定理证明系统、马文·闵斯基的SNARC神经网络学习机。

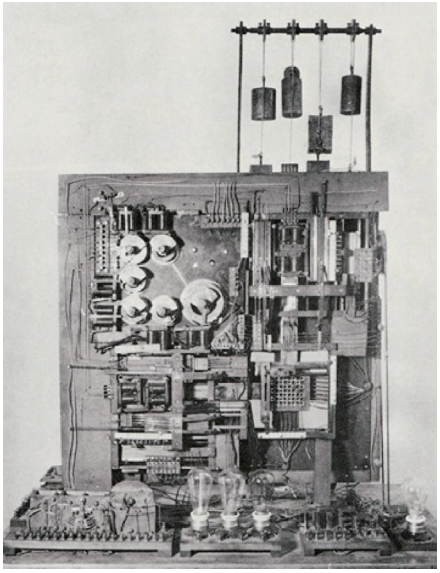


图1-30　莱昂纳多·托雷斯发明的第一代自动对弈机器El Ajedrecista

1）对弈算法

对弈一向被认为是需要很强的智能才能完成的游戏，如下象棋、围棋等。因此，对弈机器一直承载着人类的智能梦想。最早的自动对弈机器由西班牙发明家莱昂纳多·托雷斯于1910年发明，可以与人下国际象棋，如图1-30所示。

计算机发明后，许多科学家（包括图灵）开始研究对弈算法。其中，克劳德·香农的研究最为深入。1950年，克劳德·香农在一篇论文中深入探讨了一种称为 MinMax 的走棋算法，并给出了优化方案。同年，香农还设计了一台电动走棋机器，如图1-31所示。

2）定理证明

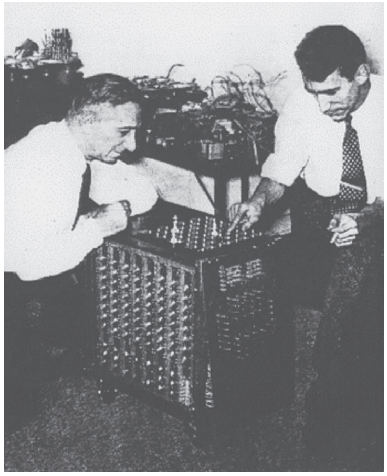


图1-31　克劳德·香农设计的电动走棋机器

随着数理逻辑的发展，人们逐渐认识到，基于少量基本原理和若干推理规则，可以推导出一个庞大的数学体系。典型的如欧几里得的几何学体系，基于五条公设即可推导出整个几何学。这启发了早期人工智能的学者们尝试用机器来完成定理证明。

1955年，赫伯特·西蒙和艾伦·纽厄尔开始探讨机器定理证明的可能性，最后由来自兰德公司的计算机程序员约翰·克里夫·肖完成了程序编写。他们把这个程序命名为逻辑理论家。逻辑理论家是一个树搜索程序，根节点是基础假设，通过设计好的推理原则进行扩展，直到扩展到要证明的结论。基于这一方案，逻辑理论家证明了《数学原理》前52个定理中的38个。逻辑理论家的诞生具有重要的历史意义，是思维即计算这一哲学思想的有力证明。正因如此，西蒙和纽厄尔在1975年共同获得了图灵奖。

3）神经网络

科学家们很早就知道，大脑是人类的智能中枢，而大脑由大量神经元组成。这些神经元是同质的，互相连接起来产生智能。若能在机器中模拟大脑神经元的连接机制，或许就能复现人类的智能。

1951年，当时还是普林斯顿大学数学系研究生的马文·闵斯基设计了一个名为 SNARC 的人工神经网络，如图1-32所示。这个网络包括40个“神经突触”模块，从随机状态开始运行，并通过操作员的反馈进行训练。SNARC成为早期神经网络研究的代表性成果。

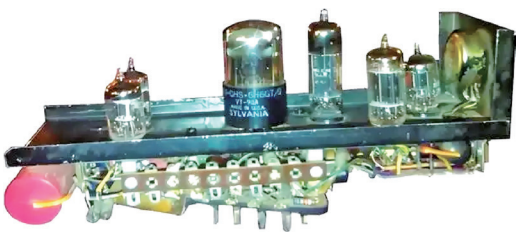


图1-32　马文·闵斯基设计的 SNARC 神经网络学习机

② 达特茅斯会议：AI 的开端

1955年9月2日，约翰·麦卡锡（达特茅斯学院数学助理教授）联合克劳德·香农（贝尔电话实验室数学家）、马文·闵斯基（哈佛大学数学与神经学初级研究员）和纳撒尼尔·罗切斯特（IBM信息研究经理）向洛克菲勒基金会提出申请，希望举办一次为期两个月、约10人参加的讨论会。在申请中，麦卡锡等人首次提出人工智能的概念，为一门新学科的诞生埋下了种子。

图1-33　达特茅斯会议旧址

这次会议开始于1956年6月18日，大约在8月17日结束，持续了近两个月，前后约有47人参加。会议地点设在达特茅斯数学系的一座教学楼内（图1-33）。其间，有时会有人做主讲报告，更多时候是自由讨论。这次会议本质上是一次长时间的头脑风暴。

依麦卡锡等人的申请，本次会议上讨论的内容非常广泛，包括：

（1）如何对计算机进行编程？

（2）如何让计算机理解和使用自然语言？

（3）能否用神经网络来表达概念？

（4）如何定义计算效率和复杂性？

（5）如何实现机器的自我改进？

（6）如何实现对象的抽象表示？

（7）如何体现随机性和创造性？

达特茅斯会议上这些问题的提出，直接引导了此后数十年人工智能的研究方向。

除了发起人麦卡锡、香农、闵斯基、罗切斯特（图1-34），本次会议吸引了赫伯特·西蒙、艾伦·纽厄尔、阿瑟·塞缪尔、雷·所罗门诺夫、约翰·纳什等。这些人在接下来的几十年里都是人工智能领域的领军人物，完成了一次又一次创举和

突破，例如麦卡锡的LISP语言、塞费里奇的机器感知理论、塞缪尔的机器学习方法、所罗门诺夫的贝叶斯推理等。

图1-34　达特茅斯会议的部分参会者

注：从左到右分别为塞费里奇、罗切斯特、纽厄尔、闵斯基、西蒙、麦卡锡、香农。

这次会议的意义不仅在于确立了“人工智能”这一概念，更在于确立了人工智能的若干重要研究方向和实现方法，标志着人工智能正式走上历史舞台。

2006年，在达特茅斯会议50周年之际，摩尔、麦卡锡、闵斯基、塞费里奇和所罗门诺夫（从左至右）重聚达特茅斯学院（图1-35）。50年前意气风发的年轻人已经年过古稀，但他们开创的“人工智能”这门学科却风华正茂。

图1-35　2006年达特茅斯会议50周年重聚

小结

任何一门新学科的诞生都不是一蹴而就的，人工智能的火种从亚里士多德时代就已经埋下了，经历2000多年的积累，才在图灵的脑海里渐渐成熟，之后才有了1956年的达特茅斯会议。可以看到，人工智能是一门既古老又年轻的科学，它经历了长期孕育，但正式诞生也不过70年的时间。了解这一历史脉络有助于我们全面、客观地认识人工智能。

达特茅斯会议标志着人工智能的诞生。这是一场自由的学术讨论，一群年轻的科学家勇敢地接过图灵“智能机器”的火炬，在美国点燃了新学科的熊熊烈焰，也开启了人工智能半个多世纪的风雨历程。从达特茅斯会议的申请和举办过程，我们看到了在历史转折时期那些年轻的科学家们敢于打破窠臼、创建新学科的勇气。我们也应该从中汲取力量，激励自己在未来的学习和探索中敢于突破，勇于创新。

达特茅斯会议也告诉我们学术交流的重要性，特别是在新学科来临之时，更需要不同学科的研究者广泛而深入地交流。目前，人工智能再次处于变革的十字路口，人工智能正深度渗透到各行各业，更需要研究者抱有开放的心态，广泛沟通，互相学习，才能把人工智能推向新的高度。

1.7 人工智能发展史(1)

学习目标

（1）了解人工智能早期发展的历史阶段，包括“黄金十年”“第一次低潮期”“回暖期”“第二次低潮期”的特点。

（2）认识符号方法的核心思想，理解定理证明、专家系统等知识驱动方法的应用场景与局限性。

（3）认识感知器模型和早期神经网络的探索成果，理解它们遇到的技术瓶颈。

（4）探讨专家系统的局限性以及第五代人工智能项目失败的原因。

1956年的达特茅斯会议之后，人工智能作为一门新学科登上了历史舞台。然而，正如其他所有新生事物一样，人工智能的发展之路也并不平坦，充满了艰辛与曲折。本节将带领大家回顾人工智能的早期发展历程，在这段时间里基于知识的人工智能占据主导地位，人们把知识总结出来教给机器，机器再基于这些知识进行推理。知识是用符号表示的，因此这一类方法也称为“符号方法”。同时，人工神经网络的研究也开始取得成果，为现代人工智能的飞跃打下了基础。

① 黄金十年（1956—1974）

达特茅斯会议之后的十余年被称为人工智能的“黄金十年”，是人工智能发展史上的第一次高潮。这一时期，研究者们在定理证明、对话机器人、神经网络方面取得一系列让人振奋的进展。这些成就使人们开始相信，创造出与人类具有同等智能水平的机器并非难事。

在这十余年里，大量资金投入人工智能研究中，很多著名大学建立了人工智能研究机构，包括马文·闵斯基所在的麻省理工学院、艾伦·纽厄尔和赫伯特·西蒙所在的卡内基-梅隆大学，以及约翰·麦卡锡在斯坦福大学创建的人工智能实验室和唐纳德·米奇在英国爱丁堡大学创建的人工智能实验室。定理证明、ELIZA机器人和感知器模型是三个代表性成果。

1）定理证明

达特茅斯会议以后，定理证明取得进一步进展。继赫伯特·西蒙和艾伦·纽厄尔的“逻辑理论家”定理证明程序之后，1959年，王浩在IBM 704计算机上用9分钟计算时间，证明了罗素和怀特黑德所著《数学原理》中的所有定理。1965年，罗宾逊提出了归结法。这种方法通过构造矛盾来反证命题的正确性，类

似于我们熟知的“反证法”。 定理证明成为人工智能研究者的第一个重要成果，展示了符号演算在解决复杂逻辑推理问题上的潜力。

2）ELIZA机器人

1966年，约瑟夫·维森鲍姆在麻省理工学院（MIT）开发了一个名为ELIZA的机器人程序（图1-36）。这个程序通过一个名为DOCTOR的脚本，能够与人类以类似心理学家的方式进行交谈。

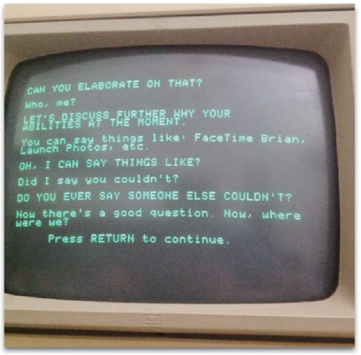


图1-36　维森鲍姆开发的ELIZA对话机器人

ELIZA的工作原理是基于转换规则，当程序检测到文本中的某些关键词时，会应用这些规则。例如，当用户输入一句话时，ELIZA会识别出关键字，然后按照预定的规则将句子重新组合成一个新的句子，仿佛在进行真正的对话。

维森鲍姆在关于ELIZA的文章中提到，计算机看起来像是在表演魔术，但一旦揭开这个程序的内部工作原理，就会发现它只不过是一些聪明的编程技巧的集合。维森鲍姆详细解释了ELIZA的工作方式，表明程序并非真的理解了人类的语言，而是巧妙的编程使它看起来像是在理解。尽管ELIZA的背后只是一些非常简单的问答模板，但人们依然认为它非常智能。

3）感知器模型

定理证明和ELIZA对话机器人都属于符号方法，通过符号演算来实现特定的功能，这在当时是主流方法。除此之外，一些“非主流”研究也在悄然进行，其中最值得注意的是关于神经网络的研究。

受人类大脑工作机理的启发，1943年，美国计算神经学家沃伦·麦卡洛克和沃尔特·皮茨提出了人工神经网络模型（ANN）。1951年，马文·闵斯基设计了第一个神经网络计算机 SNARC。1958年，康奈尔大学的弗兰克·罗森布拉特设计了一个称为感知器的单层神经网络，并在一台称为Mark 1的专用硬件上成功实现（也被称为感知机）。感知器采用麦卡洛克和皮茨提出的神经元结构，不同的是神经元之间的连接是可学习的。在罗森布拉特的实验中，感知器通过学习学会了识别图片中的字母，如图1-37所示。

图1-37　罗森布拉特的Mark 1感知器正在识别字母C

② 严冬到来（1974—1980）

20世纪70年代，人工智能的研究开始降温。研究者们错误评估了任务的难度，对未来过于乐观却无法产生预期的成果。失望情绪开始蔓延，研究经费资助也随之削减，人工智能走入低谷。

首先，符号方法遇到瓶颈。符号系统需要严格定义，很难描述大规模、开放式问题。另外，实际问题中存在大量不确定性，无法完全用符号演算来解决。其次，人们发现除了那些最简单的情况，许多问题的解决需要近乎无限长的时间。这意味着人工智能中的许多算法在实际应用中会因为计算时间过长而难以实现。

其次，被人寄予厚望的感知器模型受到打击。马文·闵斯基（Marvin Minsky）与西摩尔·派普特（Seymour Papert）在1969年出版的*Perceptrons*（《感知器》）一书中，对感知器进行了深入的分析，指出感知器模型有很大局限性，只能解决线性可分的问题（可以简单理解为用一条直线或平面即可进行完美划分的任务），无法处理线性不可分的问题，而实际问题大多是线性不可分的。因此，感知器一度被视为“鸡肋”，神经网络研究陷入停滞。

③ 短暂回暖（1980—1987）

1）专家系统

20世纪80年代，研究者意识到通用符号方法的局限，不再追求通用的问题解决方案，转而关注受限领域的应用。受此思潮影响，以专家系统为代表的基于经验知识型的人工智能走上历史舞台。

1965年，美国计算机学家爱德华·费根鲍姆和遗传学家约书亚·莱德伯格等合作，开发出了世界上第一个专家系统程序DENDRAL。DENDRAL中保存着化学家的知识和质谱仪的知识，可以根据给定的有机化合物的分子式和质谱图，从几千种可能的分子结构中挑选出一个正确的分子结构。它展示了基于专家知识解决复杂领域问题的可能性，为基于知识的人工智能打开了新大门。

与定理证明等基于规则的人工智能不同，专家系统是一种基于经验的人工智能。它不再寻求类似人脑那种通用的问题求解系统，而是专注于领域知识的构建和如何应用这些知识解决实际问题。从此以后，人工智能进入知识工程时代。

2）反向传播算法与多层感知器（MLP）

自1969年闵斯基等出版《感知器》一书后，人工神经经网络的研究几乎陷入停滞。1986年，大卫·鲁梅尔哈特（David Rumelhart）、杰弗里·E.辛顿（Geoffrey E. Hinton）和罗纳德·J.威廉姆斯（Ronald J. Williams）等利用反向传播（backpropagation，BP）算法解决了多层神经网络（图1-38）的训练问题。多层神经网络引入了一个或多个隐藏层，突破了感知器模型“只能处理线性可分问题”的局限性。自此，沉寂了十多年的神经网络重获新生，在手写体数字识别等领域取得令人瞩目的成就。

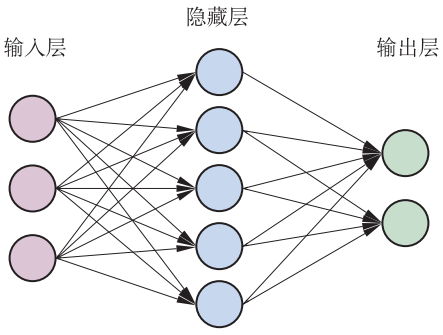


图1-38　多层神经网络

④ 二次低潮（1987—1993）

在20世纪80年代后期—90年代初期，人工智能经历了第二次低潮。尽管在此之前，专家系统一度被视为人工智能的未来，但其局限性很快显现。主要困难在于专家系统知识库的构造与维护成本极高，不仅从专家那里收集知识困难，更新和扩展知识也十分困难，因为新旧知识经常会发生冲突。例如，由匹兹堡大学开发的疾病诊断系统CADUCEUS，仅构建其知识库就耗费了近十年。随着这些问题的显现，人们对人工智能的热情再次受挫，导致对人工智能的投资大幅削减。

此后，人们开始反思传统人工智能中对符号逻辑的过度依赖。罗德尼·布鲁克斯（Rodney Brooks）是这一反思的代表人物之一，他在《大象不下棋》一文中对符号方法提出质疑。他认为，人工智能的研究不应仅仅局限于符号逻辑，而更应关注更为基础的智能行为，例如感知、运动和与环境的交互。他以大象不会下棋但依然能够很好地生存为例，强调直接与环境交互的重要性。这种观点后来发展为“行为主义”思潮，推动了大量仿生昆虫的研究。这些研究表明，通过简单的感知、反馈规则，而不是复杂的逻辑推理，也能实现足够聪明的智能行为，开启了人工智能研究的新方向。

与此同时，日本在20世纪80年代初启动了雄心勃勃的第五代计算机（图1-39）

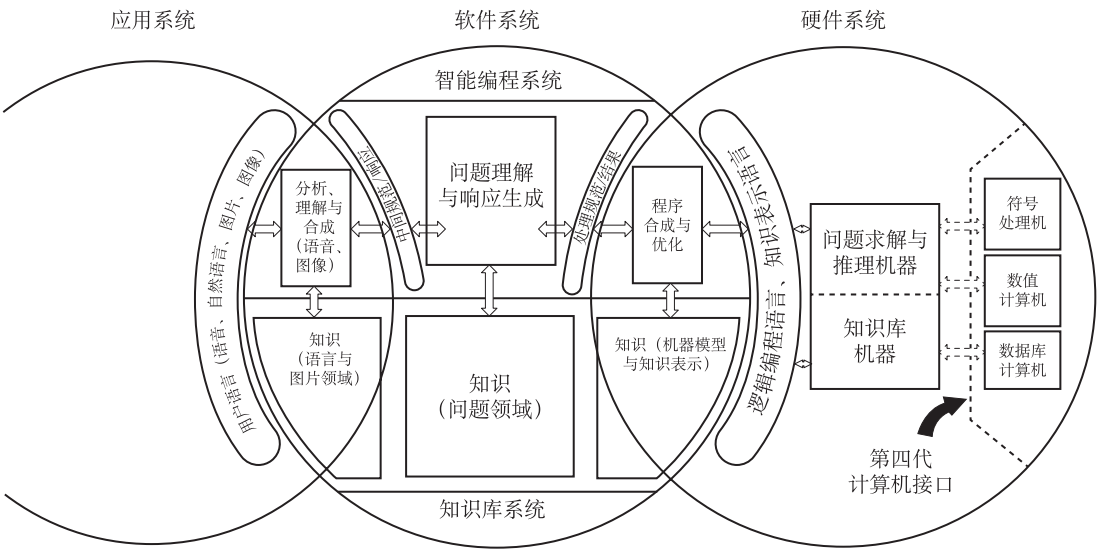


图1-39　日本“第五代计算机”概念图

项目受挫，也使人工智能的研究者再次受到质疑。第五代计算机项目旨在开发出能够与人类进行自然交流并具备人类推理能力的智能机器。然而，到20世纪90年代初，人们发现这一宏伟目标过于超前，可能无法实现。在技术上，该项目仍然依赖于逻辑推理和专家系统，而这些方法的局限性逐渐显现。再加上日本经济泡沫的破裂，导致项目未能达到预期目标。与其他人工智能项目类似，第五代计算机计划的期望值远高于技术能够实现的水平。最终，这一计划的失败也成为人工智能进入第二次低谷的重要原因之一。

小结

早期的人工智能以知识为典型特征，不论是数学定理等通用知识，还是专家的经验知识，都是机器的智能来源。研究者将这些知识表示成符号系统，利用机器的高速计算能力进行推理，因此也称为符号方法。与此同时，人工神经网络方法也取得了长足进展，虽然不是主流方法，却为现代人工智能的大发展奠定了基础。

早期人工智能的发展充满了曲折。这主要是因为人工智能的愿景过于美好，而当时不论是计算机的性能还是人工智能理论的发展都无法支撑这些愿景。当这些期望无法实现时，难免会引起失望情绪的蔓延。幸运的是，人工智能发展的总趋势始终是向前的，每一次退潮都不是回到原地，而是站在更踏实的起点上，去掉浮华，继续前进。

1.8 人工智能发展史（2）

学习目标

（1）理解人工智能复苏和变革的历史背景，理解数据积累、计算能力提升对人工智能发展的推动作用。

（2）理解机器学习的核心方法，掌握概率统计模型与神经网络的发展及其在20世纪90年代后的突破。

（3）了解人工智能的重要突破，包括深蓝击败人类棋手、DARPA无人驾驶挑战赛、IBM沃森问答系统等关键历史事件。

（4）理解深度学习革命的关键，认识AlexNet、AlphaGo等重要技术突破，及其在科学领域的应用。

（5）理解大模型时代的特点，了解Transformer架构、ChatGPT、DALL·E等生成式人工智能的工作原理与未来影响。

20世纪90年代以后，人工智能开始复苏。一个重要的变化是机器学习成为主流技术。传统基于知识的方法依赖人整理的知识，因此无法突破人的上限。机器学习方法从数据中直接学习知识，因此可以突破人的局限性，学习到强大的技能。与此同时，随着互联网的普及，数据开始快速积累，计算机性能也稳步提高，这些基础资源的累积为大规模机器学习提供了可能性，推动了以大模型为代表的现代人工智能的诞生。

① 务实与复苏（1993—2011）

在经历了20世纪80年代末—90年代初的低潮后，人工智能领域逐渐回归务实路线，研究者们不再过度地强调“模拟人类智能”这一终极目标，而是将重心转向解决特定领域的实际问题，例如语音识别、图像识别、自然语言处理、机器人动作等。这一务实转变推动了人工智能的逐步复苏。另外，数据的积累和计算机性能的提升使机器学习逐渐成为主流方法。

1）数据的积累和硬件的提升

20世纪90年代以后，数据开始快速积累。随着互联网和移动互联网的普及，人们开始在网络上制作、共享大量数据，包括大量新闻网页、知识分享社区、社交媒体、公开论文库等。研究人员对于“共享”“开源”也有了新的认识，以前收

藏在自己硬盘中的数据被共享出来，研究成果也越来越基于公开的数据集。这些都促进了数据的快速积累。

另外，计算机的性能在这一时期也有了长足进步。摩尔定律指出，集成电路上可容纳的晶体管数目，每隔约两年便会增加一倍。虽然只是一个经验预测，但近几十年的发展证明这一规律基本上是成立的。集成电路规模的增加直接推动了计算机速度的提升和内存容量的增加，进而为人工智能的发展提供了强大的计算资源。一些过去难以实现的人工智能算法变得可行，特别是以神经网络为代表的那些依赖大量数据才能完成训练的机器学习系统。

2）机器学习方法

机器学习是指让机器从数据中自主学习知识和规律的方法。图灵在《智能机器》一文中就曾提出让机器自主学习是实现智能机器的根本方法；美国科学家亚瑟·塞缪尔正式提出了机器学习的概念。然而，在人工智能发展初期没有那么多的数据，计算机的处理能力和内存都不足，机器难以进行有效学习。进入20世纪90年代以后，随着数据积累和计算机硬件的进步，机器学习才真正展现出其强大潜能。在这一时期，概率统计模型和神经网络成为当时两大主流的机器学习方法。

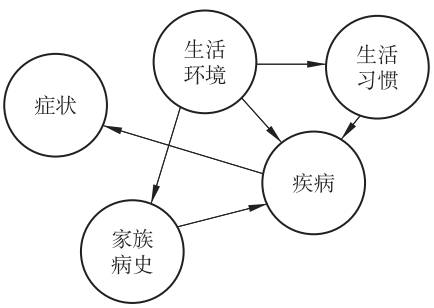


图1-40　医学诊断概率模型

概率统计模型可以认为是符号系统的扩展，和传统符号系统不同的是，它用概率来描述事件的不确定性和彼此间的关联性。例如，在医学诊断中，可以建立一个由症状、疾病、病史等各个事件组成的概率模型，如图1-40所示，并通过数据学习确定这些概率关系的具体大小。有了这一模型，可以根据症状、病史、生活环境等特征推断出最可能的疾病。

人工神经网络是另一种主要的机器学习方法。如前一节所述，人工神经网络模拟大脑的工作机制，通过将神经元互联成网络来实现功能。虽然每个神经元都很简单，但当这些神经元以复杂的方式连接在一起后，就可以实现极为复杂的功能。神经网络的研究者相信，只要网络结构足够复杂，就可以模拟任何复杂的函数，最终实现对人类智能的模型。这一思路称为“连接主义”。

与早期神经网络的研究相比，20世纪90年代以后的研究对神经网络进行了更细致的设计，并用更大量数据进行训练。与概率统计模型相比，神经网络中的节点并不对应具体事件，节点间的连接也不具有概率意义，这使它的结构更加灵活，学习能力也更加强大。然而，这一时期的数据积累还不足以支撑它的灵活结构，因此只能处于边缘地位，成为机器学习工具箱中众多工具中的一个。这一状态直到2011年深度学习革命到来之后才发生了变化。

3）代表性成就

1997年5月11日，由IBM开发的深蓝（Deep Blue）计算机战胜国际象棋世界冠军加里·卡斯帕罗夫（Garry Kasparov）（图1-41），震惊了世界。深蓝采用搜索算法寻找对自己最有利的走棋步骤，并基于其强大的计算能力和庞大的存储能力实现了超过人类顶尖棋手的思考能力。深蓝的胜利具有标志性意义，表明在规则明确的棋类游戏中，机器有可能超过人类顶尖棋手。这一事件引发了全球对人工智能能力的广泛讨论，激发了人们研究人工智能的热情。

图1-41　1997年IBM“深蓝”在国际象棋比赛中战胜当时的世界冠军卡斯帕罗夫

2005年，斯坦福大学开发的一台无人驾驶汽车Stanley（图1-42）在美国国防高级研究计划局（DARPA）举办的无人驾驶汽车挑战赛中赢得了第一名。这台无人驾驶汽车自动行驶了132英里（合212.43千米），穿过三条狭窄的隧道，完成了上百个急转弯，最终耗时6小时54分完成了比赛，勇夺冠军。DARPA挑战赛激发了全球研究者对无人驾驶技术的兴趣，推动了这一领域的快速发展。

图1-42　斯坦福大学的Stanley无人驾驶汽车赢得2005年DARPA无人驾驶挑战赛

2011年，IBM开发的沃森（Watson）人工智能系统在美国电视节目《危险边缘》（*Jeopardy*!）中击败了两位人类冠军选手（图1-43）。沃森能够快速理解用自然语言提出的问题，并通过查找知识库获知答案。据IBM介绍，Watson的知识库包括Wikipedia、电子词典、小说、话剧和Gutenberg开放的免费电子书，可以同时启动数百个搜索进程寻找答案，并对答案的可信度进行评分。Watson系统极为强大，由90台服务器和2880个处理单元组成。沃森的成功展示了自然语言处理技术的巨大进步，也让人们再次震惊于人工智能的巨大潜力。

图1-43　2011年IBM沃森在“危险边缘”问答挑战赛中战胜人类选手

② 深度学习时代（2011—2020）

2011年之后的十年是深度学习全面兴起的时代。深度学习以深度神经网络（包含多个隐藏层的神经网络）为基础建模工具，通过多层神经元模拟人脑的信息处理能力，能够从大量数据中自动学习和提取高级特征，表现出了超越传统方法的优异性能。

早在1986年，反向传播算法提出已证明多层神经网络在原则上是可行的。然而，实际情况并不乐观，多层神经网络因为结构复杂，训练起来异常困难，往往达不到预想的精度。直到2006年，杰弗里·辛顿提出了一种预训练方法，才训练出了超过浅层网络的多层神经网络，开启了深度学习的新篇章。在此之后，大量学者转到这一方向，从基础理论、网络结构、训练准则、训练过程等多个角度进行了深入研究。人们发现，深度神经网络具有强大的学习能力，特别是在大数据场景下表现优异，取得了远超传统方法的性能。这一时期的典型成就包括图像识别中的突破、围棋高手AlphaGo，以及在科学领域中的扩散和融合。

1）图像识别中的突破

2012年，杰弗里·辛顿及其团队在ImageNet大规模图像识别挑战赛中首次应用深度学习技术，他们训练了一个称为AlexNet的8层卷积神经网络，将识别错误率一举降低了10%。这一突破标志着深度学习方法在计算机视觉领域的崛起。从此以后，图像识别全面进入深度学习时代，经过五年的研究，在ImageNet上的图像识别错误率降到2.25%，甚至低于人类的识别错误率（5.1%），如图1-44所示。

AlexNet的成功不仅是机器视觉领域的里程碑，也是深度学习的里程碑，极大地鼓舞了研究者的信心。不久以后，深度学习在人脸识别、语音识别、自然语言理解等各个领域跨步前进，创造了一个又一个辉煌的战果，开启了人工智能的新篇章。

2）人机对弈的新进展

2016年，谷歌旗下的 DeepMind 公司开发的 AlphaGo 围棋程序在专业比赛

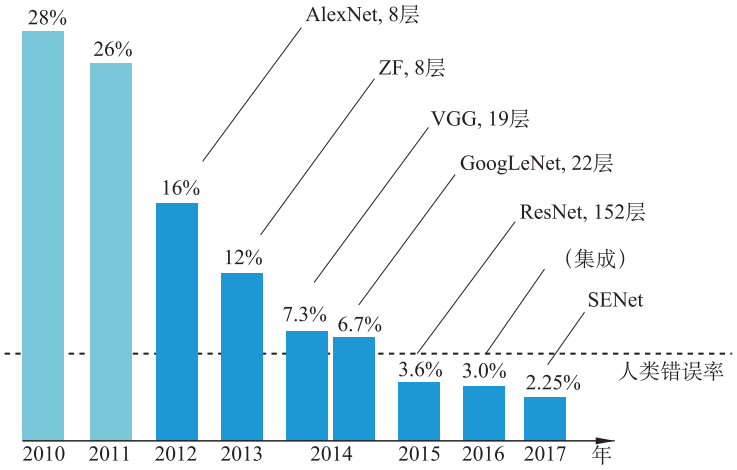


图1-44　深度学习在ImageNet上的图像识别性能

中击败了韩国棋手李世石九段（图1-45），震惊了全球。2017年，AlphaGo 再次以3∶0战胜当时的世界冠军中国棋手柯洁九段（图1-46）。AlphaGo成功的背后是深度学习，把整个棋盘局势送入神经网络，网络从局部落子开始分析，一点点扩大到全局，从而获得由局部到整体的立体视野，形成对局势的判断和落子策略。AlphaGo首先学习了大量人类棋局，并通过自我对弈的方式自我学习，最终训练成了围棋国手。

图1-45　AlphaGo对战韩国棋手李世石九段

图1-46　AlphaGo对战中国棋手柯洁九段

AlphaGo的成功让人们对人工智能的能力有了全新认知，也唤起对其潜力的极大期待。人们相信，如果人工智能在围棋这种高强度的智力活动都可以战胜人类，那么在其他领域也一样会取得惊人的成就。

3）科学领域的交叉融合

AlphaGo的成功激发了人们对人工智能的想象空间，研究者开始将深度学习方法应用到物理、材料、化学、医学、天文学、地质学、生物学等各个领域，结合深度神经网络强大的学习能力和各学科积累的知识和数据，拓展各个学科的知识边界，取得了一系列令人惊讶的成果。DeepMind的AlphaFold系统是人工智能学科融合的代表。

AlphaFold的核心功能是预测蛋白质的三维结构。蛋白质是生命活动的主要执行者，而蛋白质的功能是由其结构决定的。在AlphaFold之前，科学家们通过生化实验来解析蛋白质的结构，解析一种蛋白质往往就要花几年甚至十几年时间。AlphaFold是一个神经网络系统，通过学习大量已经得到解析的蛋白质，实现从氨基酸序列到蛋白质结构的预测。突破发生在2020年AlphaFold的第二个版本，AlphaFold2的预测误差降低到一个原子大小。这项技术彻底改变了分子生物学研究，以往需要花费大量资金和时间才能完成的工作，现在只需要在计算机前等上几分钟就能得到结果。2024年5月8日，AlphaFold3 发布，不仅可以预测蛋白质结构，还可以预测离子、核酸和蛋白质等生物分子相互作用的结果，如图1-47所示，为理解生命过程和研制新药打开了新的大门。

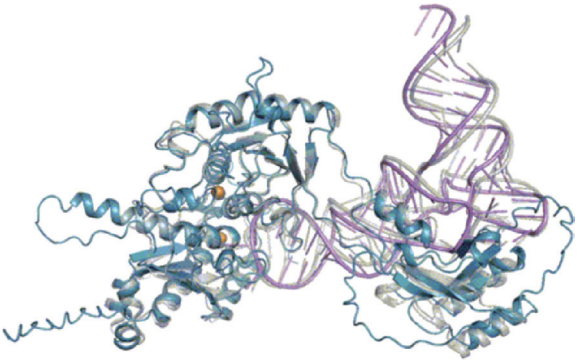


图1-47　AlphaFold3 预测生物分子相互作用得到的反应物的空间结构

注：图中蓝色部分是蛋白质分子，粉红色部分是核酸子，黄色小球是离子，灰色部分是实验得到的结果，彩色部分是AlphaFold3预测的结果。

2024年，AlphaFold主要研究人员、DeepMind公司的德米斯·哈萨比斯（Demis Hassabis）和约翰·M.朱珀（John M. Jumper）获诺贝尔化学奖，可谓实至名归。

③ 大模型时代（2020 年至今）

进入21世纪的第二个十年，人工智能又迎来了另一次历史性突破。这一突破的基础依然是深度学习，不同的是，这次研究者利用神经网络学习到了数据中的顺序性。顺序性是自然界的基础规律，一句话、一张图片、一段视频，它们都是有顺序的。句子中的顺序性体现了人类语言的内部结构和表达逻辑，图片和视频中的顺序性体现了自然界的物理规律。研究者很早就知道学习顺序性的重要意义，但一直没有找到一种合适的模型。

2017年，谷歌的研究者提出一种称为Transformer的神经网络架构，可以对长序列进行学习。2018年，OpenAI的研究者基于Transformer提出了一种称为GPT（generative pretrained transformer）预训练语言模型，这一模型可以通过很长的历史信息预测这一个词。GPT模型用40GB文本数据训练，参数达1.17亿，历史信息包括512个Token（每个英文单词依长度不同包含1～3个Token）。因为模型体积和训练数据都很庞大，人们形象地称之为大语言模型（LLM）。人们很快发现这种大语言模型具有强大的语言理解能力和生成能力，不仅可以和人顺畅地聊天，还可以写小说、做翻译、提出建议、润色论文等，表现出强大的智能。2022年年底，OpenAI发布了ChatGPT为商用名的GPT-3.5版本，其强大的能力引起轰动，两个月内注册用户达到1亿。

图1-48　GPT-3.5生成的“会飞的房子”

我：你是个科学家，写一个50字的短文，描述一个会飞的房子。

LLM：在高效能源与反重力技术的驱动下，会飞的房子悬浮于空中，自主航行。它们配备智能导航系统，可避开恶劣天气，追随阳光和清新空气，让居住环境随着人类需求动态变化。

使用LLM根据此描述生成的“会飞的房子”如图1-48所示。

此后，众多研究机构和商业公司加入大模型研究，人工智能进入大模型时代，智能水平迅速提高。例如，OpenAI在2024年9月发布的OpenAI o1在推理能力上有了大幅提高，在物理、化学、生物等学科测评中达到了博士生水平。2025年，DeepSeek AI发布的DeepSeek R1系统达到o1的推理水平，而且开源了模型。

除了在理解和生成人类语言方面取得惊人进展，深度学习还在图像和视频生成领域取得巨大成功。2021年，OpenAI发布DALL·E，可以生成逼真、高清的图片；2024年2月，OpenAI发布Sora，可以生成长达1分钟的高清视频，其逼真效果令人震惊。类似的技术也可以生成音乐。例如，2023年谷歌提出了一种MusicLM模型，在学习了20万～30万小时的人类音乐后，可以生成流畅的乐曲。近年来，一款称为Suno的音乐生成软件甚至可以为歌词谱曲，并用人声演唱出来。

无论是DeepSeek、GPT、DALL·E，还是Sora，都基于庞大的神经网络，因此统称为大模型。这些模型还有一个共同点，都是通过输出内容来完成任务的，因此也称为生成式人工智能。目前，研究者正试图将文本、音频、图像等多种信息交由一个统一的模型处理，这种模型称为多模态大模型。如OpenAI在2024年4月发布的GPT-4o就是个集视、听、读、写为一身的多模态大模型，这种模型可以和人通过语音、视频等方式自由交流，就像一个聪明的朋友，什么问题都可以向它请教。

小结

人工智能技术近十年来取得了飞速发展，这一进步可以归因于三个主要因素：海量数据的积累、强大的计算能力和深度学习算法。这三个因素共同推动了当前的人工智能技术革命。

海量数据的积累：数据是现代人工智能的基础。没有大量的数据作为支撑，人工智能就失去了学习的知识源头。现代社会的数据量呈爆炸式增长，从社交媒体、电子商务、传感器到医学成像和天文学观测，数据无处不在。这些海量数据为人工智能提供了丰富的训练素材，帮助机器学习模型不断学习和优化。

强大的计算能力：强大的计算能力能够让机器处理和学习大量数据。随着科技的进步，计算机芯片性能不断提升，为人工智能的发展提供了有力支

持。特别是高性能图形处理单元(graphics processing unit, GPU)的出现,特别适合以神经网络为主干的现代人工智能模型,极大地推动了人工智能的进步。

深度学习算法:深度学习是现代人工智能的核心技术。深度神经网络具有强大的学习能力,只要计算资源足够,就可以从大量数据中学习到人们还没有发现的新规律、新方案,这是当前人工智能迅猛发展的根本原因。

当前人工智能技术还在快速发展中,没有人能精确预测它在下一个五年会发展成什么样子。不过有些事情是可以确定的:首先,它将以通用智能体的形态逐渐渗透到我们生活的方方面面,不管是学习还是工作;其次,它将在很多领域带来深刻变革,特别是在基础科学领域,将会带来一系列颠覆性的变革。让我们拭目以待。

1.9 人工智能伦理:近期风险

学习目标

(1)掌握人工智能带来的近期风险,了解数据安全、信息伪造、AI依赖及就业冲击等现实问题。

(2)理解数据安全面临的诸多挑战,包括人脸识别隐患、大数据杀熟、隐私泄露和非法数据采集等问题。

(3)了解深度伪造(DeepFake)技术的威胁,探讨其潜在风险及应对措施。

(4)探讨AI依赖问题,分析教育、科研等领域过度依赖AI可能带来的负面影响。

(5)认识人工智能对就业市场的冲击,理解岗位替代与创造的趋势,并思考应对策略。

人工智能技术正加速渗透到我们生活的方方面面，对人类社会带来了颠覆性的影响。这些影响绝大部分是正面的，但也带来潜在的风险。特别是随着人工智能技术的进步，逼近甚至超过人类智能水平的智能机器必然会出现在我们身边，如何处理人与人工智能的关系，建设智能时代的伦理体系，需要认真研究。总体上看，人工智能引发的风险可以分为两种：一种是已经显现出来的、现实的风险；另一种是还没有到来，但可能会带来长远影响的远期风险。本节集中讨论人工智能的近期风险。

① 数据安全

现代人工智能的广泛应用离不开对数据的采集与分析，但也因此带来对数据安全方面的担忧。人工智能系统随时随地在采集我们的个人信息，从网站浏览历史到网店购物交易记录，从刷脸支付到语音通话。一些应用在不通知用户的情况下采集用户信息，导致个人信息有被滥用的风险。这些汇聚起来的海量数据如果保管不善，可能会集中泄露，带来严重的后果。

以人脸识别技术（图1-49）为例，大量应用程序使用人脸进行身份认证，如果这些信息被非法采集，就有可能被用于非法用途。例如人脸信息可能被不法分子用来进行仿冒攻击，闯入用户私人场所或盗取银行账号，造成安全隐患和财产损失。鉴于这些问题，很多国家和地区已经明令禁止人脸识别技术的随意使用。在我国，很多城市的酒店已经取消了“强制刷脸”的要求，防止用户隐私泄露。一些地区对商场、超市安装监控摄像头也做出了更为严格的规定，防止不必要的人脸扫描。

图1-49　人脸识别技术

再如一些在线商城利用收集到的交易数据对用户的购买习惯进行分析，对黏性较强的用户提高商品价格，利用大数据来“杀熟”。这种行为涉嫌非公平交易和价格欺诈。例如有记者发现某平台预订酒店时，对经常旅行的“黄金会员”显示的价格明显高出普通用户。

另外，一些应用和网站会通过第三方追踪技术（如Cookies、广告标识符）共享用户行为数据。因此，当用户在电商平台搜索“运动鞋”后，短视频平台就会推荐与运动鞋相关的视频或广告。一般来说，使用个人的身份和行为信息需要用户的授权，但很多人对技术不了解、不知情的情况下盲目授权，产生隐私泄露问题。

最后，一些人工智能公司为了训练自己的模型和系统，在没有得到用户明确授权的前提下私自收集用户数据，或强制用户授权收集用户数据。一些专门制作数据的公司在没有明确告知用户数据用途的前提下违法收集数据（如人脸或声音），或利用信息差欺骗用户同意采集数据。这些都是侵犯用户数据所有权和隐私权的行为。

为规范数据的采集和使用，我国于2021年颁布《中华人民共和国数据安全法》，明确提出：“任何组织、个人收集数据，应当采取合法、正当的方式，不得窃取或者以其他非法方式获取数据。”这为数据的合规使用提供了法律标准。

② 信息伪造

信息伪造是另一种现实风险。人工智能技术可以轻松地改变视频中的人脸和声音，生成高度逼真的伪造视频。这一技术通常称为DeepFake，中文翻译为“深度伪造”。人工智能伪造的视频不仅可能被不法分子用作诈骗工具，还可能被用于散布谣言，影响社会稳定。

图1-50　AI换脸诈骗报道

2022年，新华网报道了一位陈姓先生来到浙江省温州市公安局瓯海分局仙岩派出所报案，称自己被“好友”骗了近5万元。经过警方核实，骗子采用了AI换脸技术，利用陈先生好友之前在社交平台上发布的视频，截取了其面部画面用于“换脸”，从而对陈先生进行了诈骗（图1-50）。

2024年2月，香港媒体报道了一桩涉嫌2亿港元的AI诈骗案。报案人为一家跨国公司香港分公司的职员，称收到英国总部的信息，要求通

过视频会议讨论转账交易。会议邀请了该公司多名财务职员进行多人视频会议。由于所有人在会议内均显示了与现实相同的容貌，该职员不疑有诈，前后转账15次，合计2亿港元。事后才知道视频会议中的所有人员皆为“AI换脸”生成的。

针对伪造视频泛滥的问题，各国都在加强立法监管。例如，2023年7月4日，法国参议院投票通过“数字空间安全与监管法案”框架下关于“深度伪造”的修正案：“未经某人同意，发布通过算法处理、生成、复制其形象或话语的视频或音频内容，将被处以一年监禁和1.5万欧元罚款，如通过社交网络传播，将适用加重处罚情节。”

在我国，2023年1月施行的《互联网信息服务深度合成管理规定》明确提出，“任何组织和个人不得利用深度合成服务制作、复制、发布、传播法律、行政法规禁止的信息”“可能导致公众混淆或者误认的，应当在生成或者编辑的信息内容的合理位置、区域进行显著标识”，等等。

③ AI依赖

随着人工智能的功能越来越强大，人们对它的依赖也越来越强。文秘人员用AI写文书，翻译员用AI辅助翻译，教育工作者用AI设计教学资料，科研人员用AI润色论文，中小学生用AI解答数学题。然而，过度依赖AI可能引发严重后果。

例如在教育领域，AI可以辅助学生解题，相当于有了一个时刻陪伴在身边的老师。但如果学生把所有作业都交给AI，自己的能力得不到提升，反而会影响学生的成长。教师可以用AI来帮自己设计课程，但如果全让AI来设计，自己的主动性和创造性无法发挥，就无法成长为一名优秀的教师。

在科学研究领域，科研人员可以让AI帮自己思考解决方案、帮自己润色论文，但如果让AI在其中承担过多角色，会影响科研人员创新能力的养成，无法锻炼出真正的科研能力。还有的科研人员甚至用AI审稿，引发了对学术公平性的担忧。

更让人忧虑的是，目前的人工智能还不完美，特别是在回答一些事实性问

题时经常出错。如果完全相信它的生成结果，会造成非常严重的后果。例如，目前一些学校引入人工智能助教，让人工智能代替教师为学生答疑解惑。但是，如果AI生成的答案是错误的，就会产生误导。再如，一些研究者利用AI大段生成论文和图书内容，不做仔细核查，快速发表，可能会污染人类的知识源头。这些被污染的内容如果被AI再次用于训练，就会形成恶性循环。

为了解决这一问题，《自然·机器智能》杂志在2024年11月的一篇文章中提出了出版界对AI生成内容的署名原则。该原则包括三方面：①人类作者需要对AI生成的内容负责；②人类作者需要有足够的贡献；③人类作者需要声明AI在科研各个环节中的贡献。虽然这些原则是有价值的，但是很难保证所有作者都会认真遵守。那些投机取巧者可以利用AI低成本地生产大量垃圾内容，不仅破坏学术公平，而且破坏科学家们长期以来建立起来的学术信任。类似的情况也可能出现在新闻报道、技术讨论区、百科资源库等各个领域。如果将来我们所读到的大部分内容都是AI生成的，都是不可信的，那后果将是灾难性的。

目前，一些研究团队也在开发对AI生成内容的检测工具，如TraceGPT、WinstonAI、Hive、GPTZero。2023年12月，《国际教育诚信期刊》发表了一篇文章，作者测试了12种开源的检测工具和2种商业检测工具。结果发现，这些工具在学术论文检测方面既不精确也不可靠，而且有很强的漏判倾向。

④ 抢占工作岗位

目前，人工智能在很多领域对人类的就业形成压力。比较有代表性的是汽车驾驶员、翻译、会计、演员、播音员、影视制作者、律师。美国高盛公司研究报告表明，人工智能将对美国和欧洲2/3的工作岗位造成影响，其中46%的行政工作和44%的法律工作可以被人工智能所替代。

如何看待AI取代人类的工作岗位呢？首先要认清，任何技术进步都会取代人类的某些岗位，例如英国工业革命中出现的珍妮纺纱机让大量纺纱女工失去了工作，汽车的出现让人力车夫走进历史。人工智能会取代一部分工作岗位，这是个必然的趋势。最让人焦虑的是，本次人工智能浪潮取代的岗位更具有专业性。例如翻译和律师，这些都是需要经过长期专业训练才能胜任的岗位，现在这

些岗位也被取代了。再如画画和作曲，从传统眼光来看，这是非常具有创造性的工作，只有天赋极高的人才能胜任，但是现在人工智能也可以生成有美感的画作和流畅的音乐了。

人工智能不仅在这些专业领域里比肩人类，在科学研究领域也正在让科学家们感到压力。科学研究是一件高智商的工作，科学家是人类真正的头脑精英。一名成熟的科研人员需要经过层层筛选，并经过大量的学术训练。就是这批人现在也面临被人工智能取代的危机。2023年12月，加州大学伯克利分校和DeepMind的研究人员在《自然》杂志发表了一篇论文，报告了一个称为A-Lab的自动实验室（图1-51）。这个神奇的实验室里没有人参与，所有工作都由机器完成，包括设计实验步骤、操作实验器械和材料、检查实验结果和改进实验方案。研究人员只需要告诉它实验的目的，之后等待实验结果就可以了。A-Lab之所以这么强大，是因为它背后有一个称为GNoME 的人工智能系统，可以自动规划实验步骤，监控实验过程，分析实验结果，改进实验方案，直到得到预期的结果。与人相比，机器不会疲劳，不会疏忽，甚至不怕毒性、放射性等危险，而且可以多台机器同步工作，成倍地提高效率。可以预见，以后大量科学研究工作都会被人工智能和自动机器人取代。

图1-51　自动实验室A-Lab

未来人类还要失去哪些工作岗位，目前还很难判断。高盛的报告表明，在美国，清洁工、维修工、护理工这些服务工种目前还是安全的，反而是行政人员、律师这些白领工作更危险。但是，这也只是当今的状态，未来人工智能必然会更加强大，更多岗位被它所取代也是必然趋势。

从整个人类社会来说，这种岗位替代并不是一件坏事，这会让人们的生活的质量更高。但是对于个人来说，失去工作会让生活变得困难。我们应该对这一变化有足够的心理准备，并及时调整职业规划，选择更具有创造性的行业。从政策层面来看，应该及时调整产业方向和收入分配模式，让人工智能的红利可以惠及全体公民。

小结

本节讨论了人工智能的近期风险，包括数据安全、信息伪造、AI依赖、对人类工作的威胁等。这些问题是人工智能快速发展过程中产生的负面影响，如果不加以重视，将成为整个社会无法承受之重。同时，也要清楚地认识到，这些问题要在发展中进行解决，而不是对人工智能进行限制。历史上曾经有人担心蒸汽机的发展会把地球烧塌、电的发展会让人不敢出门，最后证明都是杞人忧天。同样，人工智能的风险也不是绝对的，及早关注这些风险并及时采取措施，才是对抗风险的正确态度。

1.10 人工智能伦理：远期风险

学习目标

（1）掌握人工智能的远期风险，理解AI失控风险、伦理与法律挑战等未来关键问题。

（2）认识AI武器的发展现状与潜在风险，理解杀伤性自动武器引发的伦理问题。

（3）理解AI失控的可能性，掌握机器学习模型不可控性、数据偏差、不可解释性等因素带来的安全隐患。

（4）理解AI对社会伦理、法律带来的挑战，分析无人驾驶事故责任认定、AI作品版权归属等问题。

（5）思考人与AI的未来关系，理解“机器人三定律”的伦理内涵，并探讨人工智能的主体性问题。

本节讨论人工智能的远期风险，即目前看还不算严峻，但随着人工智能技术的进一步发展，有可能对人类社会带来严重破坏或深远影响的潜在因素。我们将聚焦人工智能攻击人类和道德法律重构两个方面。

① 失控风险

人工智能越来越强大，发展人工智能是否会有失控的风险？我们从人工智能用于武器开始讨论。

1）人工智能武器

人工智能武器是利用人工智能技术驱动的无人作战系统，包括无人机、无人坦克、无人艇、自动防御系统等。

人工智能已经被应用在无人武器上，用来辅助甚至代替人类操控武器。例如，美国2020年8月举办的AlphaDogfight一场模拟军事对抗中，人工智能系统操纵F-16战机模拟器，以5∶0的战绩完胜具有丰富经验的人类战机飞行员。

2021年，联合国安理会利比亚专家组报告了一起无人机自主攻击人类的事件。报告称2020年3月，一台名称为Kargu-2的四足无人机的智能控制系统在没有得到人类明确指令的前提下自主开火。这可能是历史上第一次机器主动攻击人类的事件。这种不需人类指令即可主动攻击人类目标的机器称为杀伤

性自动武器（lethal automated weapon，LAW），或形象地称为杀手机器人（killer robot）。

到目前为止，全自动自主作战的机器人还没有在战场上出现，但将当前半自动的人工智能武器改装成全自动人工智能武器并不困难。如果哪一天这类武器大规模出现，将给人类带来巨大灾难。

2）人工智能的失控风险

科学家们对人工智能武器的担心主要是人工智能的不可控性。人工智能武器杀伤力大，但只要可以被操作者所控制，至少不会造成毁灭性后果。以原子弹为例，原子弹的杀伤力很大，但自从第二次世界大战结束后从没有在战场上出现过。这是因为人们知道了它的威力，所有人都不会轻言动用核武器。同时，只要人不下指令，核武器再危险也不会造成伤害。人工智能武器则不同，因为人工智能具有不可控性。例如，操作员给无人机的命令是攻击敌方坦克，但它可能把己方设施错认成目标并发动攻击。随着技术的发展，这种出错的可能性会降低，但风险会一直存在。我们讨论一下这种不可控性的来源。

首先，人工智能的不可控性来源于机器学习模型的过度灵活。我们人类的思维过程和行为方式要受到生物体结构和功能的限制。例如我们的脑容量是有限的，无法把所有事情都记住，因此只能选择性记忆；我们的眼睛看到的图像在解析度上是有限的，难以关注到局部细节；我们能思考的问题在复杂度上也是有限的，无法处理太过复杂的问题。在这些基础约束之下，经过长期的自然选择，慢慢形成了我们今天的思维和行为规律。反观人工智能却没有这些限制；它用计算来模拟人类大脑的思考过程，不受生物约束的限制，因此可以发现与人类完全不同的解决方案。AlphaGoZero就是一个典型的例子。它没有学习任何人类棋谱，完全靠自我对弈学出了可以战胜人类的棋艺。

这些自主学习出来的新方法可以让机器突破人类的固有模式，从而更高效地完成目标。然而，这些“创新”也带来了隐患，因为我们并不知道这些创新方案是否与人类的价值观相同，也不知道它们是否会对人类造成伤害。例如，一台机器人的目标是快速到达某个地点，为了完成这个目标，它可能会选择直接穿过一片麦田，从而破坏农作物的生长。再如，一个操作股票的AI，人类为它设定的目标是在市场上获得最大利润。为了实现这个目标，它可能会用更精巧的方式

操纵股市，例如和其他AI形成默契，一起攫取财富，扰乱金融体系。

其次，人工智能的不可控性来源于模型的不可解释性。当前人工智能的主流方法基于大规模神经网络模型，往往包含数十亿甚至上百亿个神经元。这样一个复杂的网络，哪怕是我们对其内部信息传递过程一清二楚，也很难理解它的最终决策。这意味着一个人工智能系统，即便我们知道它的所有工作过程，也难以理解它所选择的行为方式。因此，我们无法预测它可能做出的反常举动，即便发现了反常举动也不知道如何解决。

最后，人工智能的不可控性来源于数据偏差。人工智能模型需要大量数据进行训练。随着数据规模越来越大，对数据进行一一核查已经不可能了，人们只能相信数据是合理的。如果这些数据中包含某些错误信息或危险信息，被人工智能系统学习之后就会产生潜在风险。这些风险被隐藏在上千亿的神经元连接中，人无法察觉，平时也不会出现。然而，如果这些风险在一些关键场合暴露出来（如战场上），就可能带来重大损失。

另一种数据偏差出现在智能体的持续学习中。一个智能系统在部署时可能是风险比较小的，但如果在持续学习过程中被恶意引导，就可能会做出危险行为。2016年，美国微软公司推出的人工智能聊天机器人Tay在上线不到一天后就被紧急下架，原因就是被一些用户恶意“调教”后开始发表一些不当言论。Tay只是一个软件机器人，如果是一个可以与人交互的硬件机器人，被教坏的后果可能会很严重。

随着技术的发展，人工智能出错的可能性会降低，但风险会一直存在，而且会越来越隐蔽。我们无法看透人工智能，也无法保证它的行为完全符合预期。

3）如何可控地发展人工智能

越来越多的科学家对人工智能的失控风险提出了警告。2024年5月，深度学习的领军人物约书亚·本吉奥和杰弗里·辛顿等科学家在《科学》杂志发表文章，表达了对人工智能风险的担心。文章称：“人工智能能力和自主性的提升可能很快会大幅放大其影响，同时带来一系列风险，包括大规模的社会危害、恶意使用，以及人类控制能力的不可逆转丧失。”文章还提出了避免人工智能失控的若干方法，包括检查人工智能的价值观，看它是否与人类相同；人工智能系统上线前严格评价社会风险；防患于未然，为人工智能失控做好预案。

② 伦理冲击

1）无人驾驶汽车责任归属问题

无人驾驶汽车发生撞车事故后的责任判定涉及多个潜在责任方，包括制造商、软件开发商、车主、其他道路使用者和基础设施管理方。目前，各国对这一问题的法律框架尚在发展中，各国对事故中的责任归属也没有达成共识。

例如，2018年，Uber公司一辆测试中的自动驾驶汽车在美国亚利桑那州撞死了一名行人。这起事故引发了广泛的讨论，最终确定为Uber公司和安全驾驶员的责任。这一判决的理由是汽车未能正确识别行人，并且安全驾驶员未能及时接管控制。特斯拉汽车公司的Autopilot系统也发生过多起事故，在这些案例中，特斯拉汽车公司强调系统仅为辅助驾驶，要求驾驶员随时保持对车辆的控制，因此事故责任通常会涉及驾驶员是否遵守了使用指南。然而，随着特斯拉汽车公司推出全自动驾驶车辆，如何处理事故责任归属问题，又变成了一个棘手的问题。

2）人工智能作品版权问题

人工智能作品版权应该归谁？这也是一个需要认真讨论的问题。大模型的能力本质上来源于它所学习的各种数据，包括书籍、论文、网页新闻、网上讨论区、知识网站等。但大模型所生成的内容并不是这些数据源的简单整合，而是经过深入学习后的再创作，这与我们读了大量书籍之后拥有了知识，可以创作自己的作品是一样的性质。从这个角度看，人工智能是有一定的原创能力的。然而，从责任角度看，人工智能是没有责任主体的，出了错误无法承担责任。因此，目前所有学术刊物都不承认人工智能的作者身份。

那么，人工智能创作的成果应该归属谁呢？不同国家的司法实践各不相同。在美国，某人用人工智能创作的作品并不认为其拥有版权。例如，美国游戏设计师杰森·艾伦（Jason Allen）使用人工智能创作了一幅名为*Théâtre D'opéra Spatial*的画作（图1-52），并在比赛中获奖。然而，美国版权局认为人工智能做了主要工作，杰森·艾伦不应拥有该作品的版权。在我国也有类似的案例，但版权部门倾

图1-52　杰森·艾伦的人工智能作品

向于认可人工智能使用者的创造性工作，允许作者拥有版权。

类似的争议还在继续。问题的关键在于，随着人工智能越来越强大，人工智能工具的使用越来越容易，人在其中的贡献会越来越少，人工智能使用者对版权的主张将越来越缺少理由。此时，AI作品的归属问题就会越来越尖锐。

3）人与人工智能的伦理关系问题

不论是自动驾驶汽车撞人的责任归属，还是AI作品的版权之争，都触及一个根本问题——人工智能是否拥有主体性。如果人工智能拥有主体性，那么它就应该承担责任，相应地它也应享有权利。

然而，承认人工智能的主体性将从根本上重新定义人与人工智能之间的伦理关系。在传统的认知中，人工智能不论多么强大，都是人类的工具，是人类创造出来的附属品。这一原则集中体现为美国科幻小说作家艾萨克·阿西莫夫的“机器人三定律”。

第一定律：机器人不得伤害人类个体，或因不作为使人类个体受到伤害（保护）。

第二定律：机器人必须服从人类的命令，但前提是这些命令不与第一定律相抵触（服从）。

第三定律：在不违反第一定律和第二定律的情况下，机器人必须保护自身的存在（生存）。

机器人三定律本质上定义了机器与人类之间的从属关系，即机器是人类的下属，机器只能为人类服务，类似主仆之间的关系。然而，随着近年来人工智能技术的飞速发展，机器的智能与人类越来越接近，甚至有可能超过人类，即所谓的超级人工智能。超级人工智能的出现未必会对人类造成伤害，机器也未必会主张其自主性。然而，正如自动驾驶的归责和人工智能版权问题所反映出来的矛盾，人类自身不得不重新思考机器的主体地位问题，因为人类无法为自己难以控制的机器承担责任，也不应该窃取机器的劳动成果。这不是机器与人类之间的公平性问题，而是人与人之间的公平性的问题。因此，在强大的人工智能面前如何定义人类与机器的社会地位，是未来智能社会的基础性问题。

小结

本节讨论了人工智能的两种远期风险。

（1）失控风险。当前人工智能是以大规模神经网络为基础，其行为缺少可解释性，也难以控制。随着人工智能越来越强大，一旦它做出难以预期的行为，有可能给人类带来灾难。

（2）对社会伦理的冲击。与以往的所有工具不同，人工智能具有一定的自主性，而且这种自主性的表现会随着技术的发展越来越明显。因为这种自主性，使用者无法对其行为完全负责，相应地，AI所生成的成果也不应该完全归功于使用者。这意味着未来人工智能可能会具有一定的主体性，像我们人类个体一样，拥有一定的权利，同时也要为自己的行为承担一定的责任。然而，AI如何拥有权利、如何为自己的行为承担责任，都是悬而未决的问题。不论如何，AI的主体性问题可能是未来智能社会的一个核心问题。