开篇 什么是人工智能

2016 年，正值人工智能诞生 60 周年，一场举世瞩目的围棋人机大战正在 DeepMind 公司研发的围棋软件 AlphaGo 和来自韩国的世界著名围棋手李世石之间开展。由于围棋一直被认为是难于被人工智能突破的堡垒，此次人机大战吸引了全世界的关注。最终 AlphaGo 以 4:1 的成绩战胜了李世石，轰动了全世界。

小明从头至尾观看了这场激动人心的人机大战，对于人工智能为什么能取得如此辉煌的成就，既感到震惊又觉得不可思议。究竟什么是人工智能呢？小明找到一直从事人工智能研究的艾博士，向艾博士请教究竟什么是人工智能，人工智能是如何实现的。

0.1 人工智能的诞生

一见到艾博士，小明就开门见山地说：艾博士好！您肯定观看了这场人机大战吧？这 AlphaGo 也太厉害了，竟然战胜了李世石，真是令人震惊，我想请艾博士讲讲究竟什么是人工智能。

艾博士：我全程观看了这场比赛，AlphaGo 确实令人震级。什么是人工智能呢？这确实是一个不好回答的问题，我们从人工智能的历史慢慢讲起吧。很早人类就有制造智能机器的幻想，比如传说中的木牛流马，就是诸葛亮发明的一种运输工具，解决几十万大军的粮草运输问题。在指南针出现之前，作为行军打仗指引方向的装置，发明了指南车。车上有个木人，无论车子如何行走，木人的手指永远指向南方，“车虽回运而手常指南”。这些可以说就是最早的机器人。

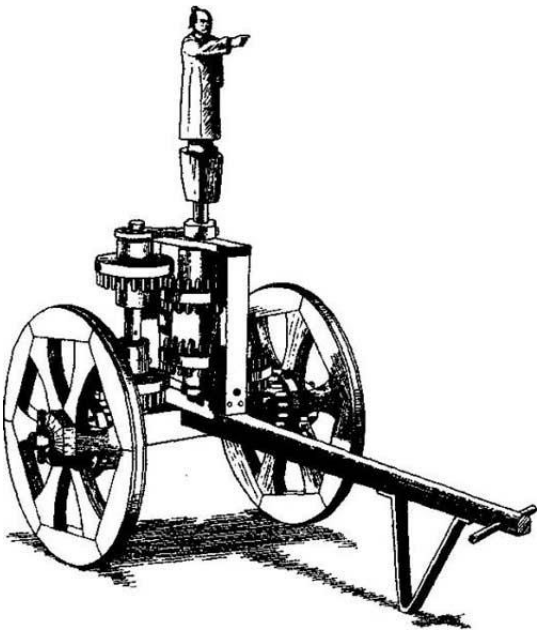

图 0.1 指南车

说到机器人，小明知道机器人英文怎么说吗？

小明不加思考地回答说：机器人的英文是 robot。

艾博士又接着问：机器人的英文为什么是 robot 呢？

小明手摸着头不好意思地说：这我就不知道了，英文课上老师说的。

艾博士哈哈大笑：robot 一词最早来源于 1921 年的一部捷克舞台剧《罗素姆万能机器人（Rossum's Universal Robots)》，用来代表剧中的“人造劳役”，从而诞生了“robot”一词，用来表示“机器人”。

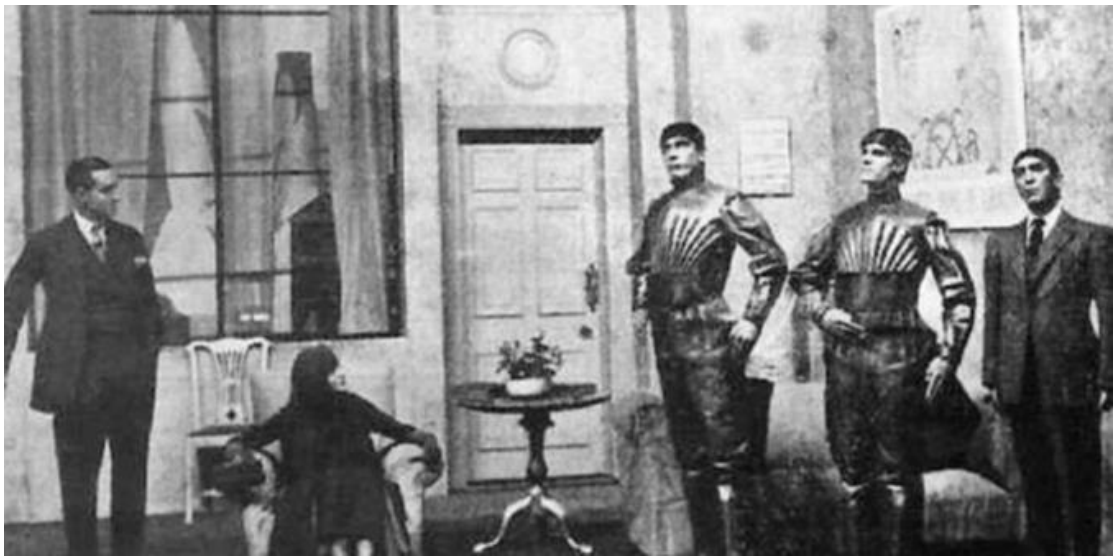

图 0.2 《罗素姆万能机器人》剧照

小明：原来 robot 一词是这么来的。

艾博士：计算机科学之父图灵很早就对智能机器进行过研究，于 1950 年发表了一篇非常重要的论文《计算机与智能》(Computing machinery and intelligence)，文中提出了一个“模仿游戏”，详细地论述了如何测试一台机器是否具有了智能，这就是被后人称作“图灵测试”的测试，并预测 50 年之后可以建造出可以通过图灵测试的智能机器。当然现在来看，图灵的这个预测失败了，目前还没有一般意义下能通过图灵测试的人工智能系统。

小明：听说过图灵测试，原来这么早就提出了图灵测试。

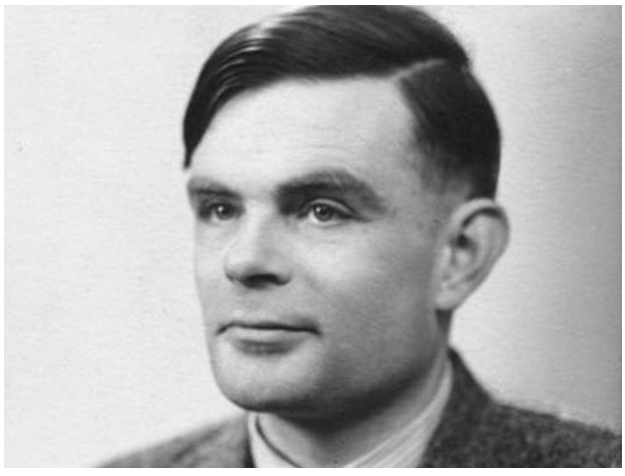

图 0.3 计算机科学之父图灵

艾博士：虽然很早就有建造智能机器的幻想，但苦于没有合适的工具，直到电子计算机的诞生，人们突然意识到，借助于计算机也许可以实现建造智能机器的梦想。正是在这样的背景下，诞生了人工智能一词，开创了一个至今仍然火热的研究领域。

那是在 1956 年的夏天，一群意气风发的年轻人聚集在达特茅斯大学，利用暑假的机会召开了一个夏季讨论会，推论会前后长达 2 个月时间，正是在这次讨论会上，第一次公开提出了人工智能，标志着人工智能这一研究方向的诞生。当年参加达特茅斯会议的大多数学者是年龄 20 几岁的年轻人，他们年轻气盛，敢想敢干，很多人后来成为了人工智能研究的著名学者，多人获得计算机领域最高奖图灵奖。这次讨论会的发起者是来自达特茅斯大学的助理教授约翰·麦肯锡，也是他最早提出了人工智能这一名称。1971 年约翰·麦肯锡教授因在人工智能方面的突出贡献获得图灵奖。

图 0.4 达特茅斯会议会址

图 0.5 一群具有青春活力的年轻人聚集在达特茅斯

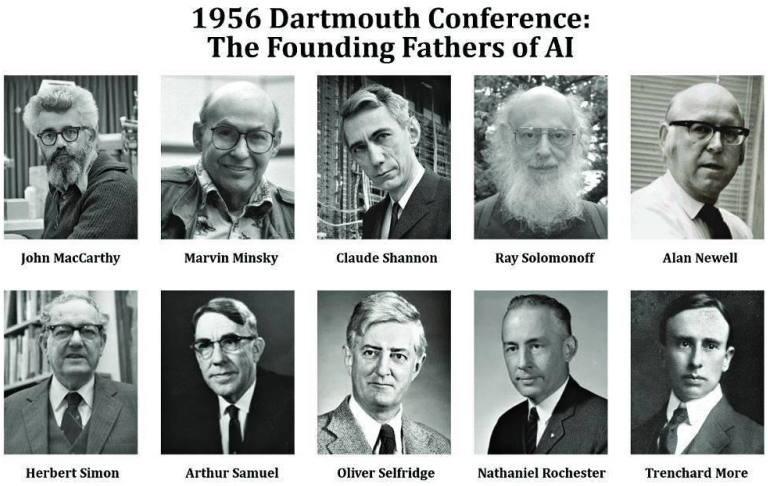

图 0.6 曾经参加 1956 年达特茅斯会议的部分学者

图 0.7 达特茅斯会议组织者、图灵奖获得者约翰·麦卡锡

小明：为什么刚好在这个时期提出了人工智能的概念，与当时的背景有什么渊源吗？

艾博士：就像刚才提到的，虽然人们一直有建造智能机器的想法，但是苦于没有合适的工具。到了 1956 年，现代计算机已经出现了有几年时间，相对于以前的各种计算工具，计算机的计算能力得到了很大的发展，借助于这样一个强大的计算工具，应该开展哪些新的研究工作呢？正是在这样的背景之下，召开了达特茅斯夏季讨论会。当时一些学者已经开展了一些与人工智能相关的研究工作，在讨论会上就有人报告了有关定理证明、模式识别、计算机下棋的一些成果。研究方向是明确的，但是应该对这一方向起一个什么样的名称充满了争议，开始时研究者对人工智能一词并没有取得共识，比如有学者建议用复杂信息处理，英国等一直用机器智能表示，若干年之后大家才逐渐接受了人工智能这一说法。

小明：在达特茅斯会议上主要讨论了哪些问题呢？

艾博士：在讨论会的建议书中，罗列了以下几方面的内容：

- 自动计算机（这里的自动指可编程）
- 编程语言
- 神经网络

- 计算规模理论（指计算复杂性）
- 自我改进（指机器学习）
- 抽象
- 随机性与创造性

从这些内容可以看出，达特茅斯会议上讨论的内容是十分广泛的，涉及到了人工智能的方方面面，很多问题到现在也还处于研究之中。

0.2 人工智能的 4 个发展时代

艾博士：人工智能诞生 60 多年了，经历过几次高潮和低谷，既有成功又有失败。60 多年来，人工智能的研究一直在曲折地前进，大体上我们可以将人工智能划分为以下 4 个时代：

- 初期时代
- 知识时代
- 特征时代
- 数据时代

这 4 个时代主要是以处理对象的不同划分的，每个时代代表了当时人工智能主要的研究方法。下面我们就分别简述一下每个时代的主要代表性研究工作。

0.2.1 初期时代

初期时代，也就是人工智能诞生的 1956 年前后，当时人们对人工智能研究给予了极大热情，研究内容涉及人工智能的很多方面，从多个方面积极探索人工智能实现的可能性。

是赫伯特·西蒙和艾伦·纽厄尔开发了一个定理证明程序“逻辑理论家”，在达特茅斯会议上二人曾经演示了这个程序，可以对著名数学家罗素和怀特海的名著《数学原理》第二章 52 个定理中的 38 个定理给出证明。后来经过改进之后，可以实现第二章全部 52 个定理的证明。据说其中有一个定理，还给出了一种比之前人类的证明方法更加简练的证明方法。

听到这里小明不禁赞叹到：在当时的条件下就可以取得这样的成绩真是了不起。

艾博士：在“逻辑理论家”的基础上，赫伯特·西蒙和艾伦·纽厄尔又进一步开发了一个称作通用问题求解器（GPS：General Problem Solver）的计算机程序，试图从逻辑的角度，构造一个可以解决多种问题的问题求解器，其逻辑基础就是赫伯特·西蒙和艾伦·纽厄尔提出的逻辑机。从原理上来说，这种求解器可以解决任何形式化的符号问题，比如定理证明、几何问题、下棋等，经形式化后，都可以统一在通用问题求解器这个框架下得以解决。

1975 年赫伯特·西蒙和艾伦·纽厄尔两人同获图灵奖，赫伯特·西蒙后来还获得了诺贝尔奖的经济学奖，成为一代传奇人物。


图 0.8 图灵奖获得者赫伯特·西蒙和艾伦·纽厄尔

小明：赫伯特·西蒙一人获得图灵奖和诺贝尔奖，可太厉害了。

艾博士：下棋可以认为是人类的一种高级智力活动，从一开始就被当作人工智能研究的对象，在 1956 年的达特茅斯夏季讨论会上，就曾经演示过计算机下棋。图灵很早就对计算机下棋做过研究，信息论的提出者香农早期也发表过论文《计算机下棋程序》，提出了极小极大算法，成为计算机下棋最基础的算法。图灵和香农还一起就计算机下棋问题进行过探讨。约翰·麦卡锡在 50 年代提出了 α－β 剪枝算法的雏形，Edwards、Timothy 于 1961 年、Brudno 于 1963 年分别独立提出了 α－β 剪枝算法。在相当长时间内， α－β 剪枝算法成为了计算机下棋的主要算法框架。1963 年一个采用该算法的跳棋程序，战胜了美国康涅狄格州的跳棋大师罗伯特·尼尔利，这在当时可以说是非常辉煌的成绩。1997 年战胜国际象棋大师卡斯帕罗夫的深蓝也是采用的 α－β 剪枝算法。

小明：没想到在人工智能的初起时代就取得了这么多的成果。

艾博士：机器翻译也是当时的一个研究热点。当时把这个问题看得有些简单化，认为只要建造一个强大的电子词典，借助于计算机的强大计算能力，就可以解决世界范围内的语言翻译问题了。

小明：结果怎么样呢？

艾博士：翻译问题当然不是只靠词典就可以解决的，结果自然是以失败告终。

在初期时代，人工智能开展了很多研究，虽然取得了一些很好的成果，但是由于对人工智能研究的困难认识不足，很快就陷入了困境之中。如何走出困境成为人们思考的问题。

小明：怎么做才能走出人工智能研究的困境呢？

艾博士：科学研究总是在与困难的搏斗中前进的。遇到问题并不可怕，关键是要找到为什么会出现这样的问题，以便想办法去战胜困难，解决问题。那么什么是走出困境的关键所在呢？科学家们开始认真反思以往的研究工作存在的问题。

前面介绍过在初期时代机器翻译是一个研究热点问题，但遇到了困难。比如对于这样一个英文句子：

The spirit is willing but the flesh is weak.

请小明说一下这句英文是什么意思？

小明回答说：这句话翻译成中文就是：

心有余而力不足。

艾博士称赞到：小明英文很好，翻译得很准确。

为了检验机器翻译的效果，有人将这句英文输入到一个英-俄翻译系统中，翻译得到一句俄语，然后又将翻译得到的俄语输入到一个俄-英翻译系统中，再次得到一句英语。如果

翻译系统靠谱的话，前后两句英文的意思应该差不多，然后最后得到的确是如下的一句英文：

The vodka is strong but meat is rotten.

请小明再说一下这句英文是什么意思？

小明看后哈哈大笑说：这句英文翻译成中文就是：

伏特加酒虽然很浓，但肉是腐烂的。

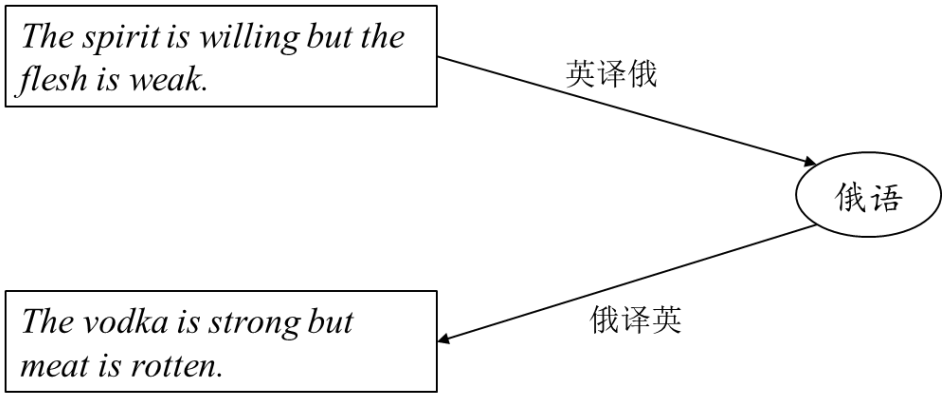

图 0.9 一个机器翻译结果示意

小明非常不解地问到：为什么是这样的结果呢？前后两句英文句子完全不是一个意思。

艾博士：因为当时的机器翻译缺乏理解能力，只是机械地按照词典进行翻译。而一些词具有多个含义，不同的搭配下具有不同的意思，如果不加以区分就会出现翻译错误。

比如这里的 spirit 一词，字典上有两个意思，一个是“精神的”，一个是“烈性酒”，在这句英文中，正确的含义应该是指“精神的”，显然机器翻译系统把它当成“烈性酒”了。如果按照“烈性酒”理解，翻译成“伏特加酒”还是比较确切的，因为“伏特加酒”是俄罗斯的一种烈性酒，但是这里的意思是“精神的”，所以造成了翻译错误。

小明：原来是这样的啊。

艾博士说：也有人说这并不是一个真实的例子，而是根据当时的机器翻译水平人为构造的一个例子。无论是真实例子还是人造的例子，其实都反应了当时机器翻译系统的一个痛点问题，即单凭构造一个庞大的字典是不能解决机器翻译问题的。翻译需要理解，而理解需要知识。就好比我要翻译一本有关人工智能的书，译者有两个候选，一个是工智能专业的学生，一个是英文专业的学生。一般来说，英文专业的学生其英语水平应该远胜于人工智能专业的学生，但是我会首选人工智能专业的学生做翻译，因为他懂得人工智能方面的知识，这些知识可以辅助他正确理解书的内容，而英语专业的学生虽然英语水平很高，但是由于不懂专业，很可能犯一些类似上面例子这样的可笑错误。

小明：您说的很有道理，我也会选择人工智能专业的学生翻译这本书。

艾博士：经过总结经验教训，研究者认识到知识在人工智能中的重要性，开始研究如何将知识融入到人工智能系统中，这就进入了人工智能的知识时代。

0.2.2 知识时代

艾博士：知识时代最典型的代表性工作就是专家系统。

小明：什么是专家系统呢？

艾博士：一个专家之所以能成为某个领域的专家，因为他充分掌握了该领域的知识，并具有运用这些知识解决本领域问题的能力。如果将专家的知识总结出来，以某种计算机可以使用的形式存贮到计算机中，那么计算机也可以使用这些知识解决该领域的问题。存储了某

领域知识，并能运用这些知识像专家那样求解该领域问题的计算机系统称作专家系统。

小明：原来专家系统是这样的含义。

艾博士：斯坦福大学的爱德华·费根鲍姆开发了世界上第一个专家系统 DENDRAL，该系统可以帮助化学家判断某待定物质的分子结构。接着又开发了帮助医生对血液感染者进行诊断和药物治疗的专家系统 MYCIN，可以说 MYCIN 奠定了专家系统的基本结构。在此基础之上，爱德华·费根鲍姆又进一步提出了知识工程，并使得知识工程成为人工智能领域的重要分支。在这个时期，专家系统几乎成为了人工智能的代名词，也是最早应用于实际、并取得经济效益的人工智能系统。

爱德华·费根鲍姆因在专家系统、知识工程等方面的贡献，于 1994 年获得图灵奖。

图 0.10 图灵奖获得者爱德华·费根鲍姆

这个时代主要的研究内容包括知识表示方法和非确定性推理方法等。首先为了让计算机能够使用知识，必须将专家的知识以某种计算机可以使用的形式存储起来，以便于计算机能够使用这些知识求解问题。为此提出了很多种知识表示方法，比如常用的知识表示方法有规则、逻辑、语义网络和框架等。其次，现实生活中的问题大多数具有非确定性，而计算机擅长求解确定性问题，如何用善于求解确定性问题的计算机完成具有非确定性问题的求解，也是专家系统研究中遇到的问题，为此很多学者从不同角度，提出了很多非确定推理方法等，像 MYCIN 系统采用的置信度方法就是非确定性推理的典型方法。

专家系统的出现让人工智能走向了使用，XCON 是第一个实用商用并带来经济效益的专家系统，该系统拥有 1000 多条人工整理的规则，帮助 DEC 公司为计算机系统配置订单。美军在伊拉克战争中也使用了专家系统为后勤保障做规划。战胜国际象棋大师卡斯帕罗夫的深蓝，在“浪潮杯”首届中国象棋人机大战中战胜柳大华为首的 5 位中国象棋大师的浪潮天梭等，均属于专家系统的范畴。

图 0.11 与深蓝对战的国际象棋大师卡斯帕罗夫

图 0.12 与浪潮天梭对战的中国象棋大师柳大华

小明：建构专家系统，专家知识的获取非常关键，如何有效地获取专家知识呢？

艾博士：是的，小明提出了一个非常重要的问题。一个专家系统能否成功，很多程度上取决于是否足够地整理了专家知识，这是一个非常困难的任务，也是构建专家系统时最花费精力的地方。一方面，领域专家一般并不懂人工智能，专家系统的构建者也不懂领域知识，双方沟通起来非常困难。另一方面，专家可以解决某个问题，但是很多情况下，专家又难于说清楚在具体解决这个问题的过程中，运用了哪些知识。因此知识获取成为了建造专家系统的瓶颈问题。如果不能有效地获取到专家的知识，那么建造的专家系统也就没有任何意义。

小明有些疑惑地问到：为什么专家可以解决问题，却说不出来运用了哪些知识呢？

艾博士解释说：我们举一个例子说明这个问题。小明你会骑自行车吧？

小明不太明白艾博士为什么问这样的问题，回答到：我会骑自行车，每天都是骑车去上学。

艾博士：假设说我不会骑自行车，一上车摇摆几下就摔倒了，你能告诉我为什么你可以平稳地骑车，总结出一些知识来告诉我，以便让我会骑自行车吗？

小明想了想说：我也不知道为什么我骑车就不到，我也不是一开始就会骑车的，慢慢练习就会了，说不出个所以然来。

艾博士：很多专家也是类似，他们长期地从事某个领域的工作，积累了大量的经验，但是却很难将知识整理出来，存在“只可意会不可言传”的问题，从而如何有效地获取知识成为了专家系统建构过程中的瓶颈问题。这极大地影响了专家系统的研发和应用。

专家的一个特点就是善于学习。我们人类一生都在学习，从中小学到大学，再到工作中，一直都在学习，我们所有的知识都是通过学习得到的。那么计算机是否也可以想人类那样学习呢？通过学习获得知识？在这样的背景下，为了克服专家系统获取知识的瓶颈问题，提出了机器学习，也就是研究如何让计算机自己学习，以便获取解决某些问题的知识。

小明：这是一个很好的想法，如果计算机自己会学习，就可以实现知识自动获取了。

艾博士：实现机器学习是一个很好的想法，但是如何实现却是一个很难的问题。早期提出过很多的机器学习方法，比如归纳学习、基于解释的学习等，虽然取得了一些研究成果，但距离实用还差的比较远，直到统计机器学习方法的提出，才使得机器学习走向了实用。这就进入了特征时代。

0.2.3 特征时代

艾博士：机器学习是通过执行某个过程从而改进系统性能的方法，统计机器学习是运用数据和统计方法提高系统性能的机器学习方法。统计机器学习方法的提出让人工智能走向了更广泛的应用，并随着互联网的发展，网上内容越来越多，为人工智能应用提供了用武之地。毫不夸张的说，统计机器学习方法的提出和互联网的发展拯救了人工智能，将滑向低谷的人工智能从崩溃的边缘又拉了回来，并逐步走向发展高潮。像 IBM 公司的“沃森”在美国电视智力竞赛节目《危险边缘》中战胜两位人类冠军选手、清华大学的中文古籍识别系统实现《四库全书》的数字化等，都是采用了统计机器学习方法实现的。

图 0.13 “沃森”在《危险边缘》竞赛中

图 0.14 大型中文古籍《四库全书》识别

小明：统计机器学习方法具有这么大的功劳啊，都有哪些统计机器学习方法呢？

艾博士：研究者提出了很多不同的统计机器学习方法，常用的方法有朴素贝叶斯方法、决策树、随机森林、支持向量机等，这些都是在实际工作中经常使用的方法。莱斯利·瓦利安特和朱迪亚·佩尔两位学者做了很多基础理论方面的研究工作，为机器学习研究建立了理论基础，二人分别于 2010 年和 2011 年获得图灵奖。

图 0.15 图灵奖获得者莱斯利·瓦利安特和朱迪亚·佩尔

小明：为什么把这个时期称作特征时代，而不是统计机器学习时代呢？

艾博士：前面曾经提到过，我这里对时代的划分是从处理对象的角度考虑的。统计机器学习方法具有很多种不同的方法，但是它们的共同特点是将特征作为处理对象，也就是输入是抽取的特征，对特征数据进行统计分析和处理。所以把这一时代称作是特征时代。比如用统计机器学习方法做汉字识别的话，首先需要我们编写程序抽取出汉字的特征，然后再运用统计机器学习方法对汉字的特征数据进行处理，从而实现汉字识别。而这里的特征是人为定义的。

特别需要强调的是，计算机用的特征与我们人类用的特征并不一定一致，需要定义计算机可以用的特征。还是以汉字识别为例，我们人认识汉字靠的是偏旁部首、横竖撇捺等特征，但是这些特征并不能用于计算机识别汉字，因为无论是偏旁部首还是横竖撇捺特征都很难抽取出来，这些特征的抽取难度并不亚于汉字识别的难度。因此定义的特征必须是容易抽取并且具有一定区分度的特征。

在利用统计机器学方法做应用时，最主要的问题就是如何抽取特征问题。然而寻找一个计算机可以使用、容易抽取并具有一定区分度的特征，并不是一件容易的事情。比如语音识别，我们很容易听懂别人在说什么，但是其特征是什么？什么特征可以区分出每一个音节？我们很难说出来。很多研究者对语音识别特征抽取做了大量的研究，然而并没有找出一个有效的特征，很长时间内语音识别的错误率居高不下。这就遇到了特征抽取的瓶颈问题。

小明：科学研究真是不容易啊，客服了一个困难又遇到了新的困难。

艾博士：我一直强调，遇到困难并不可怕，关键是要找出克服困难的方法，努力去攻克这些困难。我们人类很容易区分猫和狗，也很容易区分自家的猫和别人家的猫，哪怕是同一个品种的猫。在这个过程中并没有人告诉我们如何区分，用的特征是什么。一个小孩刚开始可能并不能准确地做出这些区分，但是慢慢地看得多了，自然就会区分了，所谓的见多识广。计算机是否也可以从原始数据中自动抽取特征呢？这就进入了数据时代。

0.2.3 数据时代

艾博士：数据时代的典型代表就是深度学习，实际上是采用深层神经网络实现的一种学习方法，其特点是直接输入原始数据，深度学习方法可以自动地抽取特征。不仅是自动抽取特征，还可以抽取不同层次、不同粒度的特征，实现深层次的特征映射，获得更好的系统性能。

深度学习的概念首先由多伦多大学的杰弗里·辛顿教授提出，实际上就是一种多层的神经网络。神经网络的研究起始于上个世纪 40 年代，80 年代中期随着反向传播算法（BP 算

法）的提出又一次掀起研究热潮。由于受当时计算条件的限制，以及统计机器学习方法的崛起，有关神经网路的研究很快落入低谷，不被人看好。但是以辛顿教授为代表的少数研究者一直坚持自己的理念，在不被看好、得不到研究经费、发表不了论文的情况下，依然“固执”地从事相关研究，直到 2006 年辛顿教授在《科学》期刊上发表论文提出深度学习的概念，才再一次受到业界的重视。

在这个过程中，两件事情引起了研究者对深度学习的广泛关注，推动了深度学习的发展。第一件事是辛顿教授与微软合作，将深度学习应用于语音识别中，在公开的测试集上取得了非常惊人的成绩，使得错误率下降了 30%，如同一石激起千层浪，让沉默了多年的语音识别看到了新的希望，在此之前，多年来语音识别没有什么大的进展，识别错误率每年以不到 1% 的水平下降。第二件事是辛顿教授组织学生用深度学习方法参加 ImageNet 比赛。ImageNet 比赛是一个图象识别任务，需要对多达 1000 个类别的图象做出分类。在比赛中辛顿教授及他的学生率先使用线性整流单元激活函数（ReLU）和舍弃正则化方法（Dropout）提升了深度卷积神经网络的性能，首次参赛就以远高于第二名的成绩取得了第一名，分类错误率几乎降低了一半。自此以后 ImageNet 比赛就成为了深度学习的天下，历届前几名均为深度学习方法，并最终达到了在这个数据集上分类错误率小于人工分类的结果。

深度学习的提出，极大地推动这一次人工智能的发展浪潮，先后战胜了世界顶级围棋手李世石、柯洁的 AlphaGo，就是在蒙特卡洛树搜索的基础上，引入了深度学习的结果。围棋曾经被认为是计算机下棋领域的最后一个堡垒，战胜世界顶级围棋手，这在以前是不可想象的。清华大学与搜狗公司合办的“天工”智能计算研究院研发的“汪仔”，在浙江卫视智力竞赛“一站到底”中，多次战胜人类，并最终战胜五年巅峰战的人类冠军，采用的也是深度学习技术。

图 0.16 与 AlphaGo 对弈的李世石、柯洁

图 0.17 参加《一站到底》节目的汪仔（右）

在神经网络和深度学习的发展过程中，与 4 位研究者的贡献是分不开的，除了前面提到的辛顿教授外，另 3 位研究者分别是纽约大学的杨立昆教授、蒙特利尔大学的约书亚·本吉奥和瑞士人工智能实验室（IDSIA）的于尔根·施密布尔博士（Jürgen Schmidhuber）。这三位学者均出生于 60 年代初期，而辛顿教授则年长他们近 20 岁左右。

纽约大学的杨立昆（Yann LeCun）教授曾经跟随辛顿教授做博士后研究，杨立昆是他自己确认的中文名。杨立昆教授在卷积神经网络方面做出了特殊贡献，早在上个世界 90 年代就开展了有关卷积神经网络的研究工作，实现的数字识别系统取得了很好的成绩，用于支票识别之中。现在卷积神经网络已经成为了深度学习中几乎不可或缺的组成部分。

蒙特利尔大学的约书亚·本吉奥教授曾经于 90 年代提出了序列概率模型，将神经网络与概率模型（隐马尔科夫模型等）相结合，用于手写识别数字，现代深度学习技术中的语音识别可以认为是该模型的扩展。2000 年本吉奥教授发表了一篇具有里程碑意义的论文“神经概率语言模型”，通过引入高维词嵌入技术实现了词义的向量表示，将一个单词表达为一个向量，通过词向量可以计算词的语义之间的相似性。该方法对包括机器翻译、知识问答、语言理解等在内的自然语言处理任务产生了巨大的影响，使得应用深度学习方法处理自然语言问题成为可能，相关任务的性能得到大幅度提升。本吉奥教授的团队还提出了一种注意力机制，直接导致机器翻译取得突破性进展，并构成了深度学习序列建模的关键组成部分。本吉奥教授与其合作者提出的生成对抗网络（GAN），引发了一场计算机视觉和图形学的技术革命，使得计算机生成与原始图像相媲美的图像成为了可能。

鉴于辛顿教授、杨立昆教授和本吉奥教授三人对深度学习的贡献，2018 年三人同时获得图灵奖，这也是图灵奖历史上仅有的三次三人同时获奖。

图 0.18 图灵奖获得者杨立顿、辛顿和本吉奥（从左至右）

非常难能可贵的是，在神经网络、深度学习遭遇学界质疑，甚至不被看好的情况下，三位教授仍然坚持研究，经过三十几年的不断努力，终于克服了种种困难，取得了突破性进展。如今计算机视觉、语音识别、自然语言处理和机器人技术以及其他应用取得的突破，均与他们的研究探索有关，并引发了新的人工智能热潮。

小明：听了您的介绍，这些科学家可真了不起，在那么困难的情况下，仍然坚持研究，并最终取得了这么了不起的成就，真是令人敬佩。

艾博士：在神经网络、深度学习的发展过程中，另一位值得一提的是瑞士人工智能实验室（IDSIA）的于尔根·施密布尔博士。1997 年施密布尔博士和塞普·霍克利特（Sepp Hochreiter)博士共同发表论文，提出了长短时记忆循环神经网络(Long Short-Term Memory，LSTM)，为神经网络提供了一种记忆机制，可以有效解决长序列训练过程中的梯度消失问题。由于其思想过于超前，在当时并没有得到学界的理解和广泛关注。后来的实践证明这项技术对于自然语言理解和视觉处理等序列问题的处理，起到了非常关键的作用，广泛应用于机器翻译、自然语言处理、语音识别、对话机器人等任务。2016 年、2021 年 IEEE 神经网络先驱奖分别授予了施密布尔博士和霍克利特博士。

对于 2018 年图灵奖颁发给辛顿、杨立昆和本吉奥三位教授，施密布尔博士多次表达过不满，认为现在很多神经网络和深度学习的工作是在自己以前工作的基础上发展起来的，忽略了自己在神经网络方面做出的贡献，曾发表长文列举自己在 90 年代的 20 项有关神经网络方面的研究工作，以及这些工作与现在的深度学习方法的关系。

图 0.19 施密布尔博士

小明：这一时代称作数据时代，是不是因为深度学习的处理对象是原始数据的原因？

艾博士：小明理解的非常正确。深度学习方法不需要人为提取特征，直接输入原始数据，实现自动特征抽取，不但解决了特种抽取的瓶颈问题，其效果还远好于人为抽取的特征，因为深度学习方法可以抽取多层次、多粒度的特征。

艾博士最后总结说：前面我们根据人工智能不同时期的发展特点，从处理对象的角度，将人工智能划分为 4 个时代，每个时代具有每个时代的特点。人工智能具有很多研究方向，在 60 多年的发展史上，提出了很多种不同的方法，这里只是简单地列举了一些每个时代的主要方法，试图让大家对人工智能的发展有个大概的了解。

小明：我觉得您的介绍挺好的，让我对人工智能有了一个总体了解，大概知道了人工智能是如何一步步发展起来的，也了解了其中的艰辛和不容易。人工智能之所以有今天的结果，是很多科学家长期不懈地努力的结果。艾博士，我想问一下，从知识、特征到数据时代，人工智能有各种不同的方法，那么这些方法之间是否有所联系和共同点呢？

艾博士：我们先通过一个男女同学分类的例子，看看不同时代是如何解决这个问题的。

图 0.20 男女同学分类问题

在知识时代，如果用专家系统解决这个问题的话，需要总结大量相关知识，并以规则的形式表达出来。比如可以总结如下规则：

如果 长发 并且 带发卡 则 是女同学

如果 短发 并且 穿短裤 则 是男同学

如果 穿高跟鞋 则 是女同学

通过这些知识实现男女同学的分类。需要总结很多知识才有可能建立一个具有一定分类能力的专家系统。

在特征时代，不需要总结知识，只需给出不同的特征即可。每种特征只需要具有一定的分类能力就可以，不需要完全 100%的区分能力。比如头发长度、鞋跟高度、衣服颜色等都可以作为特征使用。也不需要给出特征的组合，这些都交给统计机器学习方法求解即可。比起总结知识来，抽取特征相对容易的多。

在数据时代，只需收集数据就可以了，找来足够多的男女同学照片，并分别标注哪些照片是男同学，哪些照片是女同学。收集好数据之后，提交给深度学习进行训练就可以了。比起总结知识、抽取特征来，收集数据是件容易的多的事情。

从这个例子可以看出，不同方法解决问题的角度是不同的，但它们也存在共同之处。从实现的角度，人工智能一直在研究如何定义问题、描述问题，然后再结合具体的表示方法加以求解。这样我们可以将人工智能表示如下：

人工智能 = 描述 + 算法

其中“描述”指的如何定义问题、描述问题，告诉计算机做什么。“算法”则是具体的

求解方法。这就如同老师布置作业一样。老师布置作业时，要说清楚具体的作业是什么，有什么要求，这就相当于描述问题。然后同学们按照学过的方法完成作业，所学的方法就相当于“算法”。

对于人工智能来说，不同时代用不同的描述方法，比如知识时代用规则等描述问题，而特征时代用特征描述问题，数据时代就是用数据描述问题。这些必须以计算机可以处理的方式给出描述，不同的描述问题的方法，再配以相应的算法进行求解，比如数据时代用的是深度学习方法。

小明：还是感觉有些抽象，能否举一个例子说明呢？

艾博士：我们以识别猫为例说明这个问题。什么是猫呢？网上百科对猫的定义是这样的：

猫，头圆，颜面部短，前肢五指，后肢四趾，趾端具锐利而弯曲的爪，爪能伸缩，具有夜行性，行动敏捷，善跳跃，大多能攀缘上树，以伏击的方式猎捕其它动物。

这无疑是准确的描述，但是这个定义对于计算机识别猫无任何意义，因为计算机无法知道什么是头圆，什么是颜面部短等这些猫的特征，也就无法实现识别猫。在数据时代如何实现识别猫呢？这就是用数据定义猫，当然不是一两个数据，而是大量的数据，说明这些就是猫，如图 0.15 所示，告诉计算机这些就是猫，然后再利用深度学习方法，让计算机见多识广，自己去学习什么是猫，这样就可以实现如何识别猫了。

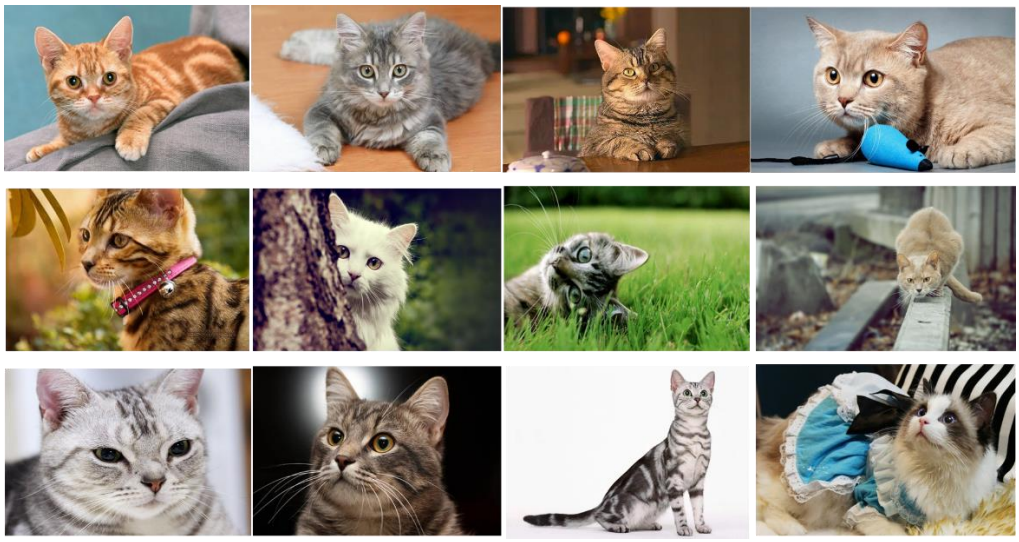

图 0.21 各种猫的数据

小明：这其实跟我们人类认识猫的过程是类似的，小孩子一开始并不认识猫，见得多了，自然就认识猫了。

从您举的例子看，数据时代的深度学习方法更具有优势，是不是就可以抛弃以前的方法了？

艾博士：虽然现在深度学习方法确实在多个方面具有优势，但是传统方法也有不可替代的作用。比如专家系统对结果比较容易控制，遇到不能求解或者求解错误的问题，容易分析出问题所在，找出问题的根源，也可以对结果给出解释。而基于特征的统计机器学习方法则具有很好的理论基础。这些都是深度学习方法不可比拟的。而深度学习方法也存在很多问题，比如不具有可解释性，理论依据不足等。所以学习人工智能的话，要多方面学习不同的方法，而不能只限于少数方法。知识面要宽，这样才有利于创新。

0.3 什么是人工智能

小明:经过您的介绍,我对人工智能有了一个初步的了解,那么究竟什么是人工智能呢?是否有一个明确的定义?

艾博士:由于智能包含了多种因素,智能的表现也是各种各样,所以如何定义人工智能也是一个难题。很多研究者从不同的角度给出了人工智能的定义,都局限于智能的某一方面,挂一漏百,因此到目前为止也没有一个能让大家都接受的统一定义。麻省理工学院人工智能实验室前任主任帕特里克·温斯顿(Patrick Winston)教授,从功能的角度将人工智能定义如下:

"人工智能就是研究如何使计算机做过去只有人才能做的智能工作。"

该定义虽然也存在一些问题,但比较通俗易懂,对初学者更容易理解。

从本质上来说,人工智能研究如何制造出人造的智能机器或系统,来模拟人类智能活动的能力,以延伸人们智能的科学。

这里有三个关键词:第一,人工智能是一个"人造"系统,第二,人工智能"模拟"人类的智能活动,第三,人工智能"延伸"人类的智能行为。这三点即是人工智能的关键因素,也反应了我们研究人工智能的目的,就是让人工智能为我们人类服务,帮助人类做更多的事情,成为人类智力的放大器。

小明:这个定义确实容易理解一些,回避了什么是智能的问题,直接从人的角度说明了什么是人工智能。

艾博士:从应用的角度来说,一个实用、受欢迎的人工智能系统应该具有如下五要素,称作"五算":

- 算据
- 算力
- 算法
- 算者
- 算景

小明:这"五算"具体是什么含义呢?

艾博士:我们先来做一个类比。小明你说一下,做一桌好的年夜饭需要哪些要素呢?

小明思考了一下回答说:我觉得首先要有好的食材,鸡鸭鱼肉样样都有,没有好的食材做不出来一桌丰盛的年夜饭,所谓的巧妇难为无米之炊。然后再有一个好的厨师,厨师的手艺很重要,否则再好的食材也做不出好的饭菜来。还有就是有一副好的灶具,灶具对厨师来说是非常重要的武器,家里普通的灶具绝对做不出饭馆的味道,主要原因是火力不够。我能想到的就是这三个要素。

艾博士:这三个要素都很重要,也确实是做一桌丰盛年夜饭的主要要素。还有一个要素就是菜谱。比如鱼可以红烧,也可以清蒸,同样是鱼,红烧和清蒸的味道完全不同。并且有的鱼适合红烧,有的鱼适合清蒸。当然这个菜谱可能是一本书,也可能完全装在厨师的脑子里。

小明:这么说的话,菜谱也很重要。

艾博士:除此以外,还有一个重要要素,就是天时地利。做任何事情都需要考虑天时地利,做年夜饭也不例外。比如说年夜饭等到了凌晨 3、4 点钟才开饭,你睡的正香甜呢,突然喊你起来吃饭,即便是满汉全席你也不会喜欢去吃。

小明:是这个理儿。

艾博士:所以一桌受欢迎的年夜饭应该具备食材、厨具、厨师、菜谱和天时地利这 5 个要素。

图 0.22 年夜饭五要素

小明不解地问到：这与人工智能有什么关系呢？

艾博士：一个实用、受欢迎的人工智能系统就如同一桌年夜饭一样，也具有与此对应的五要素，就是前面说的“五算”。

算据对应着食材，简单说就是计算的依据，包括数据、特征、知识等，是一个人工智能系统要加工的原始材料。

算力对应着灶具，就是计算的能力。现在强调大数据，对大数据的处理需要超强的计算能力，大型计算平台是必备条件。

算法对应着菜谱，就是对数据、特征、知识等进行处理的计算方法。不同的算法可以解决不同的问题，同一个问题也可以有不同的解决方法。

算者对应着厨师，是熟练掌握了算法和计算工具的人。

算景对应的是天时地利，简单说就是合适的计算场景。如同年夜饭一样，必须正确选择合适的时间、合适的场景和合适的人，才能成为一款受欢迎的人工智能系统。

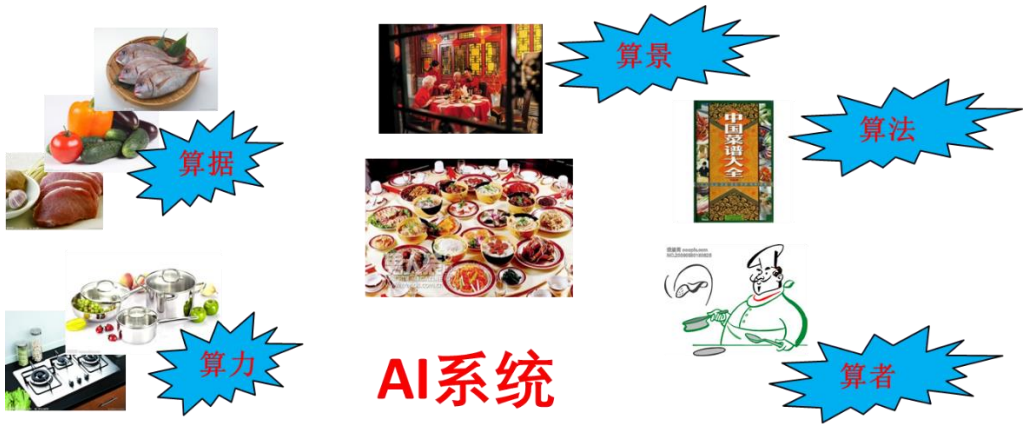

图 0.23 人工智能五要素

0.4 图灵测试与中文屋子问题

0.4.1 图灵测试

小明：艾博士，我一直有个疑问，一个人工智能系统做到什么程度就算有了智能呢？

艾博士：这是一个好问题。计算机科学之父图灵早在1950年就对这个问题进行了深入研究。在1950年发表的一篇论文中，图灵提出了著名的被后人称之为“图灵测试”的测试方法，详细讨论了这个问题。

小明：图灵测试？听说过这个说法，但是还不是太了解具体内容，请艾博士讲讲吧。

艾博士：前面我们提到过，人工智能至今没有统一的定义，不同的人从不同的角度给出了不同的定义，每种定义都是侧重了人工智能的某个方面。为什么定义人工智能这么难呢？究其根源在于什么是智能至今都无法准确说清楚。图灵早就意识到了这一点，在早期研究“机器能思维吗”问题时曾经提到：“定义很容易拘泥于词汇的常规用法，但这种思路很危险。”，“与其如此定义，倒不如用另一个相对清晰无误表达的问题来取代原问题”。正是在这样的情况下，图灵提出了后来被称为“图灵测试”的测试，以此来说明什么是机器智能，也就是后来所说的人工智能。

1950年，图灵发表了一篇题为“计算机与智能（Computing Machinery and Intelligence）”的论文，这里的Computing Machinery指的就是现在所说的计算机，由于当时Computer一词指从事计算工作的一种职业，所以图灵采用了Computing Machinery。在这篇论文中，图灵提出了判断机器是否具有智能的一种测试方法，后来被称之为“图灵测试”。

图灵测试来源于一种模仿游戏，描述图灵生平的电影《模仿游戏》片名就来源于此。游戏由一男（A）一女（B）和一名测试者（C）进行；C与A、B隔离，通过电传打字机与A、B对话。测试者C通过提问和A、B的回答，做出谁是A即男士，谁是B即女士的结论。在游戏中，A必须尽力使C判断错误，而B的任务是帮助C。也就是说，男士A要尽力模仿女士，从而让测试者C错误地将男士A判断为女士。这也是模仿游戏名称的由来。在论文中，图灵首先叙述了这个游戏，进而提出这样一个问题：如果让一台计算机代替游戏中的男士A，将会发生什么情况呢？也就是说，B换成一般的人类，机器A尽可能模仿人类，如果测试者C不能区分出A和B哪个是机器，哪个是人类，那么是不是就可以说这台机器具有了智能呢？图灵在论文中预测，在50年之后，计算机在模拟游戏中就会如鱼得水，一般的提问者在5分钟提问后，能够准确鉴别“哪个是机器哪个是人类”的概率不会高于70%，也就是说，机器成功欺骗了提问者的概率将会大于30%。后来，图灵在一次BBC的广播节目中，进一步明确说：让计算机模仿人，如果不足70%的人判断正确，也就是超过30%的测试者误以为在和自己说话的是人而非计算机，那就算机器具有了智能。这样一种测试机器是否具有智能的方法，后来被称之为图灵测试。

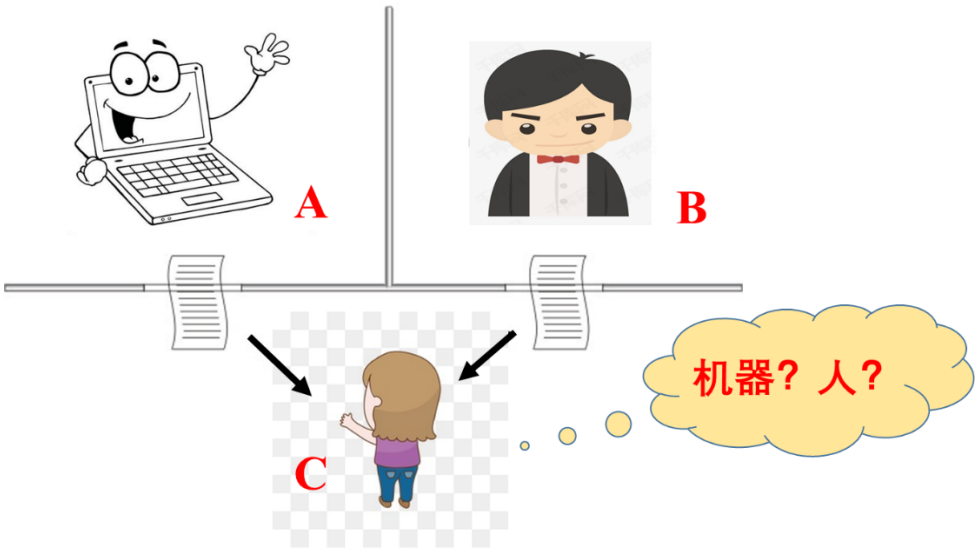

图 0.24 图灵测试

事实上，与其说图灵测试是一种测试，倒不如说是一种思想实验，是对什么是人工智能的一种定义，计算机只有达到了这样的程度，才可以说具有了智能。

小明：原来图灵测试是这么提出来的。在图灵测试中，为什么提出“5 分钟内”、“30%”这样的标准呢？又是如何确定的呢？

艾博士：据说当初男士模仿女士的游戏就是 5 分钟之后由测试者判断，而据统计，当时测试者正确区分出男女的成功率大约为 70%。也就是说约 30%的情况下，男士成功地扮演了女士，骗过了测试者。图灵也从拟人的角度，以此作为人工智能通过测试的标准。

在论文中，图灵非常详细地讨论了图灵测试的各种情况，但是在提到图灵测试时，经常会遇到一些错误的说法或用法。

小明：都有哪些错误说法或者用法呢？

艾博士：常见的错误有以下 2 种：

错误 1:将机器在某一方面的能力超过人类认作是通过了图灵测试。比如有人说 AlphaGo 在围棋比赛中通过了图灵测试。这也是不正确的说法。图灵测试要求的是模仿人类，不能让测试者很容易就分辨出它是机器，除了要求像人一样回答问题外，还要求他会伪装，不能表现出明显的超人一等的能力。因为如果一台机器具有明显的超出人类的能力，也很容易让测试者判断出他不是人类而是一台机器。就如同谷歌的 Master 在网上围棋赛上连续获胜时，很多人就已经猜测他是机器了。

图灵在论文中也已经明确的提到了这一点：“有人声称，在游戏中提问者可以向被试问几道算术题来分辨哪个是机器，哪个是人，因为机器在回答算术题时总是丝毫不差。这种说法未免太轻率了。（带模拟游戏程序的）机器并没有准备给算术题以正确的答案。它会故意算错，以蒙骗提问者。”也就是说，一个通过了图灵测试的机器，应该会蒙骗，也会像人一样出错，也不能表现出明显超出人类的能力。比如当他遇到一个复杂的计算题时，应该会适当的说不会做，或者说我需要一些时间来计算，甚至可能会算错。学会隐藏自己的实力也是智能的表现。

错误 2：将超过 30%的测试者误把机器当作人类，理解为机器的回答中超过 30%的内容与人类一致，区分不出是否为机器所答。比如在某年某市的高考试卷上就出现了这样的说法：“超过 30%的回答让测试者误认为是人类所答，那么就可以认为这台机器具有了智能”。也在新闻中看到过有公司声称自己的什么产品通过了图灵测试，给出的理由是超过 30%的内容

区分不出是否是机器所答。这显然是错误的。因为对于测试者来说，只要有一个回答有明显的问题，就可以被认作为机器所答，图灵测试的通过标准是骗过 30%以上的测试者，而不是超过 30%的回答无法确认是否是人回答的。

还有一点需要说的是，图灵测试是一个全面的测试，而不是某一单一领域的测试。在单一领域，机器水平再高，也不能说他通过了图灵测试。

小明：以上两点您总结的真好，如果不是您强调说明，我可能也会犯类似错误。

0.4.2 中文屋子问题

艾博士：关于图灵测试也一直存在一些争议，即便通过了图灵测试，就说明计算机具有智能了吗？哲学家希尔勒对此有不同看法，提出“中文屋子问题”加以反驳。

图 0.25 中文屋子问题

小明：希尔勒是如何反驳的呢？

艾博士：这要从罗杰•施安克设计的故事理解程序开始讲起，该程序可以理解用自然语言输入的一段简短的故事。

小明：如何知道这个程序理解了输入的故事呢？

艾博士：这就如同我们上课学习一样，老师怎么知道同学们是否听懂了上课内容呢？

小明：可以通过提问，看同学们是否能正确回答问题就知道同学们是否听懂了上课内容。

艾博士：对于故事理解程序也采用类似的方法，输入一段简短的故事之后，就故事内容进行提问，如果程序能正确回答问题，则说明程序理解了这段故事。提问的内容可以是故事中直接叙述的内容，也可以是故事并没有明确说明，但隐含在故事内的内容，尤其是后者更能检验程序是否理解了故事。

小明：听起来是一个很有意思的研究，那么罗杰•施安克的这个程序可以正确回答问题吗？

艾博士：对于比较简单的故事还是可以正确回答问题的。比如如下的两小段故事：

故事 A：

“一个人进入餐馆并订了一份汉堡包。当汉堡包端来时发现被烘脆了，此人暴怒地离开餐馆，没有付帐或留下小费。”

故事 B：

“一个人进入餐馆并订了一份汉堡包。当汉堡包端来后他非常喜欢它，而且在离开餐馆付帐之前，给了女服务员很多小费。”

这两段故事情节差不多，但是结果不同。作为对程序是否“理解”了故事的检验，可以分别向程序提问：在每个故事中，主人公是否吃了汉堡包。小明你回答一下，主人公是否吃了汉堡包？

小明：两段故事都没有明确说主人公是否吃了汉堡包，但是根据故事情节，故事 A 中主人公并没有吃汉堡包，因为该人“暴怒地离开餐馆，没有付帐或留下小费”。而在故事 B 中，主人公肯定吃了汉堡包，因为该人“非常喜欢它”、“给了女服务员很多小费”。这些都是隐含的内容，对于我们人来说理解起来并不难，但是让程序做到这一点感觉并不容易。

艾博士：小明回答的是对的。对于程序来说，这种理解确实具有难度，但是对于类似的简短故事，罗杰•施安克的程序做到了这一点。

但是，哲学家希尔勒却提出了异议。他说，能正确回答问题就是理解了吗？希尔勒背后质疑的实际是图灵测试，他认为，计算机即便通过了图灵测试，也并不代表计算机就具有了智能。

小明：不太理解希尔勒是怎样一种逻辑，难道都通过图灵测试了，还不能说计算机具备了智能吗？

艾博士：为此，希尔勒构造了一个理想实验，即“中文屋子问题”，用来阐述他的思想。

罗杰•施安克的程序本来是理解英文故事的，希尔勒认为什么语言并不重要，他假定该程序同样可以理解中文故事。

小明：为什么要换成理解中文故事呢？

艾博士笑到：可能与西方人认为中文最难有关吧？

艾博士接着讲到：既然这是一个程序，那么懂编程的人就可以看得懂这段程序，并按照程序像计算机一样进行数据处理，虽然可能很慢。希尔勒设想自己就是那个懂编程的人，把自己和程序一起关在一个称作“中文屋子”的屋子里，有人将中文故事和问题像输入给计算机一样送到屋子里，希尔勒按照程序一步步地操作，并按照程序给出答案，显然答案也是中文的，因为希尔勒一切都在按照程序操作，如果程序能给出中文回答，那么希尔勒也可以做到。如果程序可以理解这段中文故事、给出正确答案，那么希尔勒自己按照程序也同样可以给出正确答案。这似乎没有问题吧？

小明回答说：应该没有问题，如果不考虑处理所用时间的话。

艾博士：但是希尔勒最后说“我并不认识中文，也不知道这段故事讲了什么，甚至最后给出的答案是什么也不知道，但是我却通过了这个测试”。所以希尔勒提出疑问：能给出正确答案就是理解了吗？就实现了智能吗？

小明：哲学家就是不一样，通过一个简单的例子，提出了一个很有意思的问题。

艾博士：中文屋子问题提出后，引起了世界范围内有关什么是智能的大讨论，有赞同希尔勒观点的，也有反对他的观点的，公说公有理婆说婆有理，各自发表不同的见解。

小明：最终有什么结果吗？

艾博士：这类问题注定了不会有一个统一的结果，但是通过讨论，加深了人们对什么是智能、什么是人工智能的认识。

小明：艾博士您是如何看待中文屋子问题的呢？

艾博士：我个人认为，中文屋子应该当作一个整体来看待，虽然屋子里的希尔勒并没有理解这段中文故事，但是从屋子整体来说，能正确回答问题就是理解了，也就具有了智能。就如同我们人，也是从一个人的整体来讨论是否理解了问题，不能说人体里面的哪个部分理解了故事。

小明：我觉得您说的很有道理，中文屋子应当当作一个整体看待，理解故事的是屋子整体，而不是内部的某个部分。

0.5 第三代人工智能

艾博士：人工智能发展到今天虽然取得了很好的成绩，但是目前以深度学习为主导的人工智能还存在很多问题有待解决。

小明：主要存在哪些问题呢？

艾博士：我们通过一些典型的例子说明一下当前以深度学习为主导的人工智能存在的问题。

在大数据时代，人工智能需要大量的数据，但是人认识事物，并不需要太多的数据，人可以很容易做到举一反三。比如下图所示是国宝级文物东汉时期的青铜器“马踏飞燕”侧面和正面图，对于人来说，如果认识了侧面图是马踏飞燕，那么当看到正面图时，也能认出是马踏飞燕，不会由于没有见过正面图而不认识。但是对于目前的人工智能系统来说，很难做到这一点，需要学习大量不同角度的图片，才有可能正确识别出不同角度的马踏飞燕。


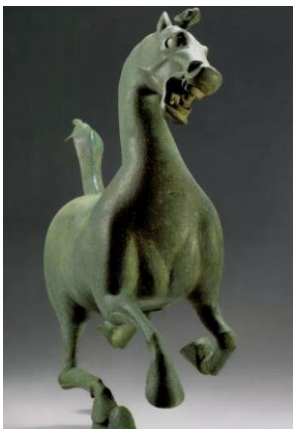

图 0.26 马踏飞燕图

小明：为什么人工智能系统不能像人一样做到举一反三呢？

艾博士：人之所以能做到举一反三，是人具有理解能力，是在理解的基础做识别，很多情况下即便不给出全图也可以正确识别。而目前的人工智能系统依靠的是“见多识广”，通过大量数据的训练形成“概念”，人工智能所谓的“认识”，其实是在猜测，由于“见过”的数据多，往往猜测的也比较准确，但也存在猜错了的风险，甚至可能错的离谱。

小明不解地问到：会有哪些风险呢？

艾博士：比如下图给出的是某自动驾驶汽车发生的车祸照片，其中图（a）是车祸后的汽车，图（b）是与汽车发生碰撞的大货车。当时该自动驾驶汽车在没有任何刹车的情况下，与大货车直接相撞，造成惨重后果。经事后分析，自动驾驶汽车将大货车识别为了立交桥，所以没有采取任何措施就撞了上去。图（b）红圈所标示的就是汽车与大货车相撞的具体位置。

小明：原来是这样啊，在自动驾驶汽车场景下，万一发生了识别错误，就可能造成严重后果。

（a）某自动驾驶汽车车祸现场

（b）发生碰撞的大货车

图 0.27 某自动驾驶汽车车祸

艾博士：还有人针对人工智能系统研究对抗样本，利用人工智能系统的脆弱性，对人工智能系统进行攻击。

小明：这里所说的攻击是什么含义呢？又怎么实现的攻击呢？

艾博士：这里说的攻击，指的是在一个原始图象上增加少量人眼无法察觉的噪声，欺骗人工智能系统发生识别错误，达到攻击的目的。下图就是一个对抗样本攻击的例子。其中左图是一个熊猫图象，中图是专为攻击构造的噪声图象，然后将噪声图象以 0.7%的强度添加到左图中，得到右图所示的添加了噪声之后的熊猫图象。小明你对比一下看，能看出左图和右图有什么差别吗？

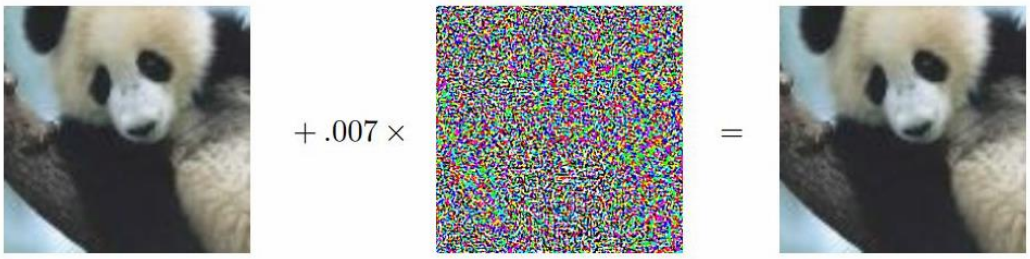

图 0.28 对抗样本举例

小明反复对比了以后说：看不出任何差别来。

艾博士：对于人来说，加上这么少的噪声不会有任何影响，即便是涂抹几下，或者部分遮挡，也不会影响到我们人类识别这是一个熊猫。但是对于人工智能系统就不同了，对于左图可以正确地识别出这是熊猫，但是却将右图识别为一只长臂猿，并且信心满满地认为是长臂猿的可信度高达 99.3%。

小明：这也太不可思议了，一点点噪声就会带来这么奇怪的事情发生。

艾博士：这就是对抗样本带来的效果，这个噪声不是普通的噪声，而是利用了目前人工智能方法的弱点，为了攻击有意构造的噪声。这件事情就更危险了，对一些人工智能的应用可能带来灾难性的后果。比如说如果自动驾驶汽车大量使用，有人对路标进行攻击，本来指引右转的路牌，攻击者通过对抗样本的方法让汽车错误地识别为向左转，而人又很难发现路牌有问题，岂不是非常危险？

小明：这确实是件可怕的事情。

艾博士：最近 MIT 和伯克利的研究者发表了他们的研究成果，利用类似对抗样本的攻击方法，成功地攻击了与 AlphaGo 类似的计算机围棋系统 KataGo，通过训练得到的围棋 AI 可以 77%的胜率战胜 KataGo，而 KataGo 同 AlphaGo 一样，在围棋方面具有超越人类的能力。

小明：就是说通过对抗训练的这个围棋 AI 具有更高的下棋水平了吗？

艾博士：不是的，这个围棋 AI 水平并不高，甚至下不过普通的业余棋手，只能说是一物降一物，只对 KataGo 有效，它是通过欺骗 KataGo 犯下严重错误而获胜的，并不是真的具有什么下棋水平。

小明：艾博士，听您讲的人工智能存在的这些问题，让我想起来古希腊神话中的“阿喀琉斯之踵”。阿喀琉斯是位大英雄，在他刚出生时其母将其沉浸进冥河中做洗礼，因为相传在冥河水中洗过礼就可以做到刀枪不入、长生不老。但遗憾的是洗礼时被母亲提着的脚踝没有浸入水中，从而留下了一个死穴，最终在特洛伊战争中阿喀琉斯被帕里斯一箭射中脚踝而死。目前的人工智能可能就存在这样的“死穴”，一旦这些“死穴”被利用，就可能会带来不可预测的灾难性后果。

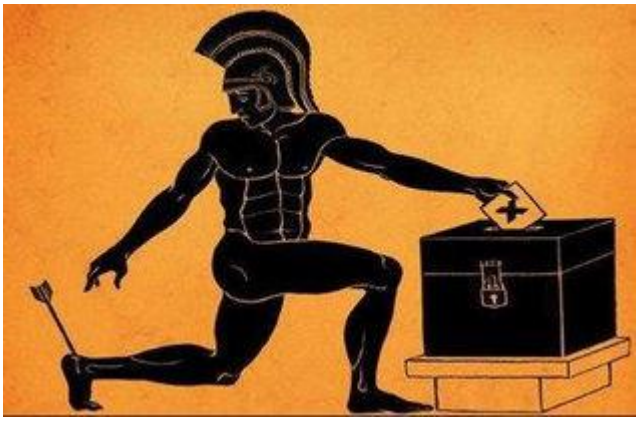

图 0.29 阿喀琉斯之踵

艾博士：这里只是通过几个例子说明了当前人工智能存在的一些典型问题，更多的问题我们就不再叙述了。这些问题在实际应用中出现就会带来不可靠、不可信、不安全等问题，究其原因是因为目前的人工智能方法靠的是猜测，缺乏理解和可解释性，无论是做对了还是做错了，都很难给出其原因所在。

为了克服这些存在的问题，清华大学的张钹院士提出了第三代人工智能的概念。张钹院士按照人工智能的发展，将目前的人工智能划分为两代。第一代是以专家系统、知识工程为代表的基于知识的人工智能，第二代是以统计机器学习、深度学习为代表的基于数据的人工智能。在这里张钹院士认为特征也是数据的一种，所以讲我们前面讲的特征时代和数据时代合并为一代。在两代人工智能发展过程中，虽然取得了很好的成绩，但是还存在诸如我们所说的各种问题。

张钹院士认为，当前的人工智能适应于求解满足如下条件的问题：

- 掌握丰富的数据或知识
- 信息完全
- 确定性信息
- 静态与结构化环境
- 有限领域与单一任务

但是在实际应用中并不能满足这样的条件，比如不足的数据、不完备的信息、动态的环境、非确定性信息等，因此一旦超出了条件所限，人工智能系统就可能出现问题。而第三代人工智能就是要解决这些问题，在数据不充分、信息不完全、信息不确定、动态环境、复合任务条件下，实现安全、可信、可靠、可扩展、可解释的人工智能。这也是人工智能今后发

展的重要方向，在其中一个或者几个方面取得进展，都将是对人工智能研究的重大突破。

小明：看起来人工智能需要研究的问题还有很多，困难与机遇并存，我要努力学习，先打好基础，学会已有的东西，在前人的基础上才有可能取得新的进展。

艾博士：小明加油！将来就靠你们了！

0.6 总结

艾博士：关于什么是人工智能就简单地讲这么多，下面请小明对这部分内容做一个总结。

小明：好的，我试着总结一下。

1956 年在达特茅斯讨论会上，第一次公开提出了人工智能这一概念，标志着人工智能的诞生。60 多年以来，人工智能研究经风历雨，几次陷入困境，在一代代研究者不畏艰难的努力之下，终于取得了今天这样的成绩。从研究对象的角度，人工智能 60 多年的研究史，可以大体上划分为 4 个时代。

第一个时代为初期时代。随着人工智能的提出，研究者们满腔热情地投入到研究中，在诸如定理证明、通用问题求解、机器博弈、机器翻译等多个方面开展了全方位的研究工作，也取得了一些成绩。但是由于对实现人工智能的困难估计不足，很快陷入了困境。通过总结经验教训，人们认识到知识的重要性，必须让计算机拥有知识，才有可能实现人工智能。

第二个时代为知识时代。一个专家之所以能够成为某个领域的专家，关键是他拥有了这个领域的知识以及运用这些知识解决领域内问题的能力。如果能将专家的知识总结出来，并以计算机可以使用的方式加以表示、存储，那么计算机也可以像专家那样求解该领域的问题。这就诞生了专家系统，专家系统是知识时代最具代表性的工作，后来又进一步发展为知识工程。

专家系统最重要的就是知识，但是如何获取专家的知识，成为了建造专家系统的瓶颈问题。

第三个时代是特征时代。人的知识是通过学习获得的，那么计算机是否可以实现自动学习呢？这就诞生了机器学习，也就是让计算机自动获取知识。曾经提出过多种机器学习方法，但都无法应用于实际之中，直到统计机器学习方法的提出才改变了这一现象，使得机器学习可以真正解决实际问题。

统计机器学习方法利用统计学方法对输入特征进行统计分析，找出特征之间的统计规律，实现对特征数据统计建模，并应用于求解实际问题。在互联网大发展、数据海量增加的情况下，为人工智能的广泛应用打下了基础，可以说是统计机器学习方法将人工智能从低谷之中拯救回来，为后来的人工智能热潮奠定了基础。

在应用统计机器学习解决实际问题过程中，除了统计机器学习方法外，最重要的就是特征抽取，各种应用研究主要围绕着针对具体问题的特征抽取方法展开，但是如何抽取特征又成为了人工智能应用中新的瓶颈问题，阻碍了人工智能的发展。

第四个时代是数据时代。能否让计算机从原始数据中自动抽取特征呢？能够从数据中自动抽取特征的深度学习方法应运而生。

简单地说，深度学习就是一种多层人工神经网络，简称神经网络。神经网络的研究起始于上个世纪四十年代，五、六十年代曾经有过很多研究，但由于缺少通用的学习方法而受到冷落。到了八十年代中期，随着 BP 算法的提出再次受到研究者的重视，并掀起新的研究热潮。但由于受诸如计算能力、数据量等客观条件的限制，有关神经网络的研究再次陷入低潮。直到 2006 年神经网络以深度学习的面貌再次出现，并在语音识别、图象识别中获得成功应用后，以深度学习为主导的人工智能取得了爆发性发展，在多个不同的领域取得快速发展和应用，重新引领了人工智能的发展热潮。

深度学习之所以能在多个方面取得好成绩，主要是因为深度学习方法具有从原始数据中自动抽取特征的能力，通过多层神经网络，可以实现不同层次、不同粒度的特征抽取，实现多层的特征映射。

如何验证一个计算机系统是否具有了智能呢？图灵对此进行了深入研究，提出了著名的图灵测试。图灵在论文中设想，有一台机器 A 和一个人 B，并有一个测试者 C。测试者 C 向机器 A 和人 B 提出问题，机器 A 和人 B 回答问题。如果经过若干轮测试之后，测试者 C 不能准确地判断出 A 是机器、B 是人，则说明机器 A 通过了测试，具有了智能。

针对通过图灵测试是否就预示着具有智能这个问题也引起过争论，“中文屋子问题”就是针对此问题而提出的。假设有一个可以理解中文的程序，一个懂的编程但并不懂中文的人，把人和程序放在一个称作“中文屋子”的房间了，提问者用中文向屋子里的人提问，屋子里的人按照程序像计算机那样“人工”执行程序。如果程序可以给出正确答案，那么屋子里的人也应该可以给出正确答案，因为他是严格按照程序操作的。虽然答案是正确的，但是屋子里的人不懂中文，他根本不知道问题是什么，也不知道回答的是什么，能说他理解了中文吗？这样的讨论推动了研究者对什么是智能、什么是人工智能的理解。

基于深度学习的人工智能虽然取得了很辉煌的成绩，但是在很多方面还存在不足，具有被攻击的风险，从而导致人工智能系统具有不安全、不可靠、不可信等。如何解决这些问题，是下一代人工智能，也就是第三代人工智能要解决的问题，也是未来人工智能的重要发展方向。