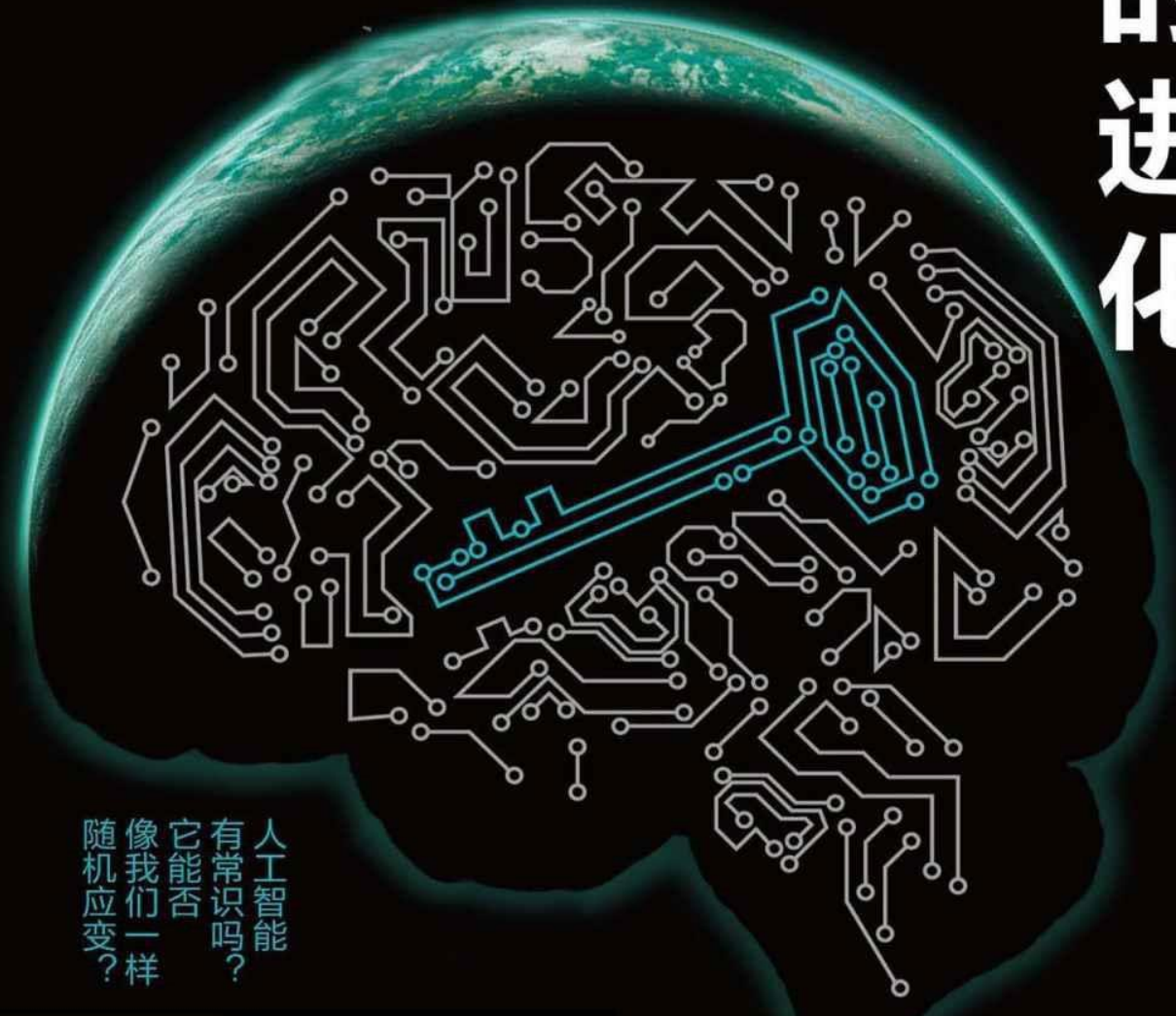
人工智能的进化

计算机思维离人类心智还有多远？

COMMON SENSE, THE TURING TEST, AND THE QUEST FOR REAL AI

让我们重回人工智能的起点，检视人工智能的最初梦想。



人工智能有常识吗？它能否像我们一样随机应变？

Hector Levesque

[加]赫克托·莱韦斯克 著

王佩 译

中信出版集团

人工智能的进化

[加] 赫克托·莱韦斯克 著
王佩 译

中信出版集团

目录

献给
帕特

前言

本书旨在从人工智能（AI）的角度探讨人类的心智。研究人类思维的朋友可能会感到奇怪，人工智能怎么会懂得人类的心智呢？人工智能属于计算机科学，当然是和计算机打交道的，而心智是人类所特有的东西。当然，或许未来某天我们会像探讨“动物心智”一样，讨论一下“计算机心智”。但是，让人工智能解释人类心智，是不是有点儿像让天文学家解释蛀牙问题，有些风马牛不相及呢？

答案是否定的，这两者绝非风马牛不相及。因为计算机科学并非仅关注计算机本身，计算机科学主要关注的是计算，而计算是个过程，比如对一串数字进行分类排序、压缩音频文件或去除数字图像的红眼等。当然，这个过程通常要由电脑完成，但是，人工或者某种设备也可能完成。

人工智能（或至少一部分人工智能）的基础假设认为，日常思维（人们每天使用的思维）同样也是一种计算过程，在对此进行研究时，无须过多考虑是谁或是什么东西在进行这种思维过程。

这也正是本书的主要内容。

有一点引起了争议，同时也更有趣，那就是当今存在不止一种人工智能研究流派，而且各家都在著书立说。在不同的人工智能研究流派当中，确有一部分是在研究各种思维形式，其他则对其敬而远之。

究其原因，我们不妨先考虑一下金鸡独立这个动作。我们有多希望能将思维、规划以及解决问题的能力融为一体？人类个体能否做到这点呢？我们是否认为事先对这一主题进行研究会有所帮助？比如看看类似《傻瓜单腿站立指南》这种书？如果一位人工智能研究人员想造出能单腿站立的机器人（不管出于任何原因），他绝对认为涉及思维的人工智能没什么借鉴意义。

当今研究人工智能的众多流派中，绝大部分都已背离了20世纪50年代开始的对于思维（和规划、解决问题）的原始研究。包括研究方向和

研究目标等涉及该研究的基本理念都已发生改变，而且这种变化在最近兴起的机器学习当中尤为突出。毫无疑问，从纯技术的角度来看，这种变化取得了巨大成功，而且极有可能超越人工智能的其他流派。虽然对机器学习的研究极度依赖数据统计，但它的本质与人工智能的原始研究仍然大不相同。

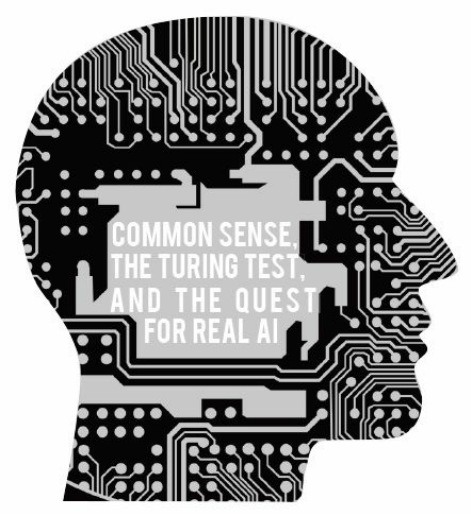
本书的目的就是回过头来，重新思考人工智能这一概念（即现在所谓的“老式人工智能”），同时解释为什么这一概念出现60年以来，我们依然对于研究人类思维兴致不减，而且在此过程中，我们不仅推出了几项实用技术，还从中学到很多东西。

一本关于人工智能的著作可以承载多种内容。它可以是人工智能的教科书，可以是关于人工智能新技术的调查，可以讲述人工智能的历史沿革，甚至可以回顾电影当中对于人工智能的刻画。而这本书并非以上提及的种种。这本书讲述了人工智能背后的理念和假说、人工智能的理论基础，讨论了人工智能看待事物的独特方式及原因，并介绍了人类的心智及其如何产生智能行为。

本书全部文本只用了大约150kb（千字节）的存储空间，看到这么一本小书，让人多少有些沮丧。相比而言，我电脑里一秒钟的维瓦尔第音乐会视频所占的空间都比它的两倍还大。因此，单纯从原始数据来看，我电脑里半秒钟的视频与本书分量相当。但是在半秒钟的时间里，乐队指挥都来不及举起指挥棒。

正如古语所说：一画（帧）胜千言。然而，我认为还有另外一点值得注意，与图画相比，语言是更为紧凑的信息载体。几百个字呈现在视频当中，或许只是一瞬间，却也能将炖牛肉的方法解释清楚。由于人类的进化，我们已经能够随心所欲地使用这些超紧凑型的信息载体。我们通过在他人耳边低语，就能对其日后的行为产生影响。人类是如何做到这一点的，正是我们现在所探究的问题。我希望读者能够喜欢我的观点，也能感受到我在探究过程中的激动心情。

第1章
什么是人工智能？



本书旨在从人工智能的角度探讨人类的心智。对于人类的心智，我们会有很多话说，但首先，我们转向人工智能领域本身，并简要介绍本书将会深入探讨的一些想法和理念。

自适应机器学习

当下，人工智能在科技领域受到普遍关注。各大报纸、杂志的科技和商业版块，几乎每个星期都会谈到人工智能，它们告诉我们，诸如微软、IBM（国际商用机器公司）、谷歌和苹果等大型科技公司都在人工智能的研发领域大量投入，其他公司也都竞相效仿。2015年11月，丰田公司宣布在人工智能领域投资10亿美元；2015年12月，埃隆·马斯克（Elon Musk）宣布新成立一家名为OpenAI的非营利性企业，研究人工智能，并对其另外注资10亿美元。

资本市场为何会出现这样的喧嚣？如果仔细研究一下投资者对于这数十亿美元的预期，你会发现，这里的人工智能技术似乎与科幻小说和电影中想象的人工智能大相径庭。基本没人谈论诸如电影《银翼杀手》（*Blade Runner*）中所看到的人形智能机器人，或者电影《2001：太空漫游》中的高能无形智能机器（哈尔9000计算机）。

当下，众多科技公司摩拳擦掌准备开发的这种人工智能，更应该被称为“自适应机器学习”（adaptive machine learning），即AML。广义来说，AML的设计理念是让计算机系统通过对大量数据进行分析，掌握智能行为方式。确切地说，目前人们对于人工智能的兴奋点主要在于挖掘并应用我们通常所说的“大数据”。

在此，我们不去深究任何AML技术如何开展工作这样的细节问题，只讨论其背后的理念。

出于论证的需要，假设我们想让计算机系统对猫进行识别。那么，该系统的工作就是将给定的众多图像分成两类：一类图像上有猫的样子，另一类图像上没有。接下来的问题就是如何建立这样一个系统。过去，人工智能程序员往往会编写出一款程序，在图像中搜寻猫的一些具体特征。例如，搜索猫的脸：泛绿色或泛黄色的眼睛、杏仁状的竖直瞳孔、倒置的粉色三角形一样的鼻子、胡须等；也可能会搜索猫的轮廓：小小的脑袋、三角形的耳朵、四条腿和向上翘起的尾巴，还可能会搜索猫身上毛皮的独特颜色。如果该程序在某个图像当中发现了足够多的上述特征，就会将该图像标记为一只猫；如果没有，则会将其忽略。

但是AML提出了一种完全不同的理念。首先，你向系统提供大量数字图像，其中一些是猫的图像，另一些不是。然后告诉系统通过某种方式压缩所有图像数据，即在众多图像的众多补丁当中搜索一组“特征”。这种特征可能是某一特定颜色和亮度，也可能是某个区域，区域的边缘在亮度和颜色上明显与其他部分不同。这一理念的目的是通过找到这样一组特征，进行重新组合并构建与原始图像相似的内容。然后，你让系统根据这些特征继续进行抽象分析，在这些特征当中寻找普遍性，随后如此继续进行几个层次的分析。

这一切都是在不告知系统具体要搜索什么的情况下完成的。如果原始图像里有很多猫的图像，那么系统就很可能会在更高层面上分析出一些与猫相关的特征。而且重要的是，这些特征全部由原始图像决定，而非人工智能程序员决定如何在图像中识别猫。正如斯坦福大学的吴恩达（Andrew Ng）所说：“你将海量数据输入程序，让数据说话，让软件自动从数据中学习。”

人们发现，AML在这种所谓“无监督”的环境下运转极其良好，完全超出几十年前人工智能研究人员做出的预测。AML的成功可以归结于以下三点：拥有可供分析处理的海量数据（通过在线、专业存储设备或传感装置获取），掌握功能强大的能够处理这些数据的计算技术，以及高速计算机。而这些条件，在30年前并不具备。

当然，不会有人愿意斥资数十亿美元，仅仅是为了识别猫。但是你想一下，如果这些不是猫而是乳腺的影像，其中一些还藏有医生难以发现的肿瘤，结果会怎样？或者假设这些根本就不是可视数据，而是一段段录音，其中包含有关人士所说的几句话？或者假设这是银行的交易数据，其中还涉及欺诈或洗钱？或者是人们在线购物过程中浏览和购买商品的历史记录，抑或是与汽车前风挡玻璃处视觉数据相对应的脚踏板和方向盘运动轨迹？计算机系统能够自动分析这些海量数据，从中获得规律，并应用于极具经济和社会效益的领域。

老式人工智能

虽然AML技术的发展趋势让人心潮澎湃、激动异常，但是这并不是本书关注的重点。事实上，随着本书逐渐展开，你会发现，除了最后的第10章以外，我们基本不会讨论任何人工智能技术。

本书关注的是一个与众不同的人工智能发展愿景，该愿景于20世纪90年代由约翰·麦卡锡（John McCarthy）、马文·明斯基（Marvin Minsky）、艾伦·纽厄尔（Allen Newell）与赫伯·西蒙（Herb Simon）等人首先提出，我们且称之为老式人工智能（Good Old-fashioned AI），或简称GOFAI。这些科研人员提出的另外一种智能并不致力于分析海量数据，而是专注于常识。

什么是常识？拥有常识与训练有素地分析海量数据有何不同？本书将用大量篇幅回答上述问题。阻碍我们将此类差别梳理清楚的困难之一在于，在日常生活中，我们几乎所有的时间都是在自己相对熟悉的场景中度过，依靠长期形成的习惯或行为模式生活。比如我通常会在上班路上买一杯咖啡，在大厅里遇到同事乔治时和他打招呼，吃热狗时会放上芥末。即使偶尔看到一棵从未见过的树，我基本也是安之若素，毫不在意。生活就是这样平淡无奇。

但是，我们在第7章将会深入讨论到，当我们遇到的情况与平日司空见惯的行为模式截然不同时，我们的常识就会发挥作用。如果某天早上，常去的那家咖啡馆关门了，我该怎么做？难道一直拽着门不松手？如果同事乔治见到我说的第一句话是“你听说三楼危险化学品泄漏了吗”，我该怎么回应，难道还像往常一样，说句“早上好”然后离开？如果我的热狗上已经撒满调料，我该怎么办？再多放点儿芥末把自己辣晕？我们在从未经历过的陌生环境中做出反应，并对自己说“等等，这是什么情况”，这才能真正证明我们具备常识。

研究人员在研究老式人工智能（GOFAI）的过程中经常遇到的问题是：如何对这种常识做出解释？人们如何利用这种常识搞清楚当日常行为模式行不通时该怎么办？到底何为“搞清楚”？遇到一扇打不开的门时，一个人需要经历何种心路历程才最终做出决定该如何处理？只因有

人对我说了几句话，我就决定在寒冷的冬日不穿大衣来到户外，这是怎么回事？只因有人发出某种声音或说出某些话语，就足以让我的行为举止一反常态？

这就是令早期人工智能科研人员心驰神往的地方，这就是本书的主要内容。

具有常识的程序

1956年夏天，人工智能的先驱们在美国达特茅斯学院的一次会议上相遇。此次会议的组织者、美国计算机科学家约翰·麦卡锡（1927—2011）首次使用了人工智能这个说法。他于1958年发表了一篇题为“具有常识的程序”的论文，这篇文章内容精彩，开创性地为人工智能的研究确立了方向。

本书第9章会在技术层面对这篇文章进行详细阐述，但有一点非常明确，就是关注常识。当时麦卡锡就有了非常明确的想法：

> 如果一款程序能够根据已知信息和指令，自主推断出由此可能导致的各种直接结果，我们就说这款程序具有常识。

我们会在本书第3章中看到，麦卡锡对于人们如何利用已知信息和指令搞清下一步怎么做尤其感兴趣。

然而，人们从一开始就认为不应将掌握常识与学习经验割裂开来。麦卡锡也说：“我们的最终目标是让程序能够像人类一样有效地学习经验。”

显而易见，我们通过尝试打开一扇锁上的门得到了一些信息，我们从同事乔治那里知道了化学品泄漏的事，我们通过直接观察，发现热狗上已经撒满调料。但是，GOFAI与AML强调的重点各不相同，我们在第5章将会看到，在GOFAI的环境下，学习不需要大量培训数据，而是经常涉及语言。约翰·麦卡锡再次强调：“如果想让某个程序具有学习能力，那么首先必须能够让该程序听懂指令。”非常明显，他将重点放在了语言的学习上面，同时他还将这个系统称为“采纳建议者”。

实际上，也有人批评老式人工智能过分专注于符号和文字。批评者明确指出，智能行为并非全都需要通过语言表达。你可能无法准确说出骑摩托车通过急弯道时车身应该倾斜多少度，但这并不妨碍你成为骑行

高手。你能在一群男孩当中辨认出一对亲兄弟，但不一定能够说清如何发现他们之间的相似之处。

当然，人类语言无疑在老式人工智能领域占据着举足轻重的作用，两个重要原因如下。第一，我们在第6章将要谈到，除了用于日常交流，语言在人类行为当中也起着非常特殊的作用，而这种作用是其他动物所不具备的：我们如何利用先前通过语言获得的信息解决新问题、新情况。

第二，语言是研究智能行为的绝佳媒介。如果我们想知道某个人是如何处理新情况、新问题的，可以直接问他。虽然偶有一言不发亦可展示智力的情况，但能像我们人类一样灵活使用语言确实需要智慧。因此，根据艾伦·图灵（Alan　Turing）最先提出的理论，语言是测试智力的终极手段。

图灵测试

艾伦·图灵（1912—1954）是英国数学家，曾参与计算机的早期开发（我们将在第8章中对他和他的成就做更为详细的介绍）。他对于能否通过编程让计算机完成人类智力才能完成的任务（如下棋）进行过深入思考，并成为该领域的先驱之一。计算机只能根据自身的程序进行工作，但是关于我们能够设计出何种程序来指引计算机工作，无论是当时还是现在，都还很不明确。

根据我的猜测，图灵对机器是否拥有思考和理解能力或者意识等哲学讨论逐渐感到厌倦。他预见到人工智能将会给技术领域带来极大挑战，但有些人认为即使克服这些困难，由于计算机是非生物体，所以也不可能具备思考理解的能力。他对这样的观点感到愤怒。

1950年，图灵在一篇极具影响力的论文中建议，我们应该将注意力从如何制造机器、机器外观如何以及其内部如何运转等方面转移到可观察的外部行为（externally observable behavior）上。当然，我们需要考虑的行为多种多样，我们可以将机器放在并不熟悉的新环境当中，看它作何反应，我们可以要求它辨认猫的照片，我们还可以研究它如何处理撒满调料的热狗。图灵提出的想法是让机器与询问者进行一次不受束缚、毫无保留的对话。

这就是他所谓的“模仿游戏”（2014年的电影《模仿游戏》中也有相关描述）。这个游戏的原理是设置一个询问者和两个隐藏测试对象，一个是人，另一个是计算机，他们与询问者通过电传打字机进行对话。谈话要自然流畅，可以涉及任何话题。不管谈话进行了多久，只要询问者分辨不出测试对象中哪个是人哪个是计算机，就算计算机赢得比赛。按照当前的说法，我们就可以说该计算机通过了图灵测试。

对于图灵测试而言，非常重要的一点是牢记对话可以涉及任何话题。以下就是他在论文中想象出的一段经典对话：

询问者：在莎士比亚十四行诗的第一行，写着“我能否将你比

作夏日？”在这里如果用“春日”来代替“夏日”，岂不是更好？

计算机：那不押韵。

询问者：那用“冬日”怎么样？这可以押韵了吧？

计算机：是的，但是没人愿意被比作冬日。

询问者：是匹克威克先生提醒你圣诞节到了吗？

计算机：某种程度上，是的。

询问者：但圣诞节是在冬日，我觉得匹克威克先生不会介意这么对比。

计算机：我认为你是在开玩笑，冬日的意思是冬天里普通的一天，而不是圣诞节这样特殊的日子。

我们当然不会像计算机程序一样进行这样的对话。但我们是否有理由相信我们永远无法编写出这样的程序呢？这正是图灵建议我们应该关注的地方。

在我看来，图灵的观点可以这样理解：诸如“聪明”“思维”“理解”等心理术语实在太过模糊，且充满情绪，难以让人信服。我们如果坚持在科学的语境中使用这些术语，我们应该说，一台能够通过相应行为测试的计算机，具有与人类一样的质疑能力，并且到此为止。我们可以想象，如果将电影《阿甘正传》中的标志性格言“傻人做傻事”改编一下，图灵会说成“聪明人做聪明事”。所以换句话说，我们应该问的问题是“机器能像具有思考能力的人一样运行吗？”而不是问“机器能思考吗？”

中文房间理论

现在回过头来看，图灵当时提出的观点不言而喻，就是强调可观察的外部行为是所有人工智能研究的核心，这一点我们在下一章将会提到。（我们有理由怀疑使用非正式对话的形式进行智力测试的效果，我们会在第4章讨论这一问题。）尽管如此，仍然出现了批评的声音，我们下面简要介绍一下哲学文献中反对图灵观点的一种声音。（如果你对于哲学争论没有兴趣，可以直接跳到下一章。）

20世纪80年代，哲学家约翰·瑟尔（John Searle）提出，理解（思维或智能）的内涵远大于可观察的外部行为，即便这种行为可以与所谓的模仿游戏一样涉及范围较广。以下就是他的论据（稍有修订）。

假设现有一款计算机程序能够通过图灵测试，即可以完全不受任何限制地进行人机对话。我们再假设这次对话使用的语言不是英语，而是汉语。因此，根据模仿游戏的规则，汉字（通过某种编码的形式）被输入该程序，随后该程序能够用汉语（以某种编码的形式）应答，作为输出。即使是精通汉语的人，仍然长期无法分辨输出信息的是机器还是人。

我们现在假设瑟尔自己不懂汉语，但是精通计算机编程（实际上他并不精通，但没关系）。他拿着一本小册子藏在一个房间里，这本手册里涵盖了该计算机程序的全部文本。房间外有人递给他一张纸，上面写着汉字，他虽然不明白纸上写的是什么，但是可以模拟计算机程序对此类情况的应对方法。瑟尔根据这本程序手册，跟踪观察程序如何应对这种情况，并将程序输出的应答写在一张纸上，交还给房间外的人，而他在整个过程中根本不明白纸上写的是什么意思。

简而言之，有个名叫约翰·瑟尔的人，他在自己的房间里接收汉语信息，然后通过这本手册，用汉语回复这些信息，而回复对汉语的使用与母语是汉语的人毫无差别（假定该计算机程序已通过图灵测试）。换句话说，在瑟尔身上，可观察的外部行为是完美的，但他却根本不懂汉语！因此瑟尔的结论是：仅确保行为正确是远远不够的，因此图灵的观点是错误的。

对此说法持反对意见的人认为，这种行为并非出自瑟尔本身，而是瑟尔与这本写有计算机程序的手册共同创造的结果。虽然瑟尔不懂汉语，但这个由瑟尔和这本手册一起组成的系统是懂汉语的，因此图灵的观点并没有错。对于上述异议，瑟尔的应答可谓简明扼要：假设他先把这本手册的内容记下，然后将手册销毁，那么就不存在所谓系统了，存在的只有瑟尔。因此，图灵的观点是错误的。

这就是中文房间理论的大概内容。

人们对此理论的反响如何呢？人工智能领域有人对此并不买账。这也可以理解，大家对人工智能领域中的哪种技术阻碍此类程序（假设其存在）的开发更感兴趣，而对我们在一个房间里用一个人就能准确地模拟这种程序的运行（虽然速度上，人可能只是程序的几百万分之一），并得出相关结论兴趣不大。

但是，就这类思想实验本身而言，还有一点需要考虑：我们如何能够确定，瑟尔没有通过记下这本手册的方式来掌握汉语？如果他以这种方式掌握了汉语，那么不懂汉语也能通过图灵测试的说法就站不住脚。因此我们必须要问：为什么我们要相信存在这样一本手册能够想瑟尔之所想，让他只需记下该手册的内容就不用学会汉语？

如果不深入了解手册中所谈到的这款程序，就很难对上述问题做出回答。我曾经提出，我们可以转而关注其他更为简单的行为模式，即数字求和。

我的设想如下。假设行为测试不要求能说汉语，而是要求将20个10位数字求和。如果一本手册中列出了20个10位数字及它们求和的所有可能的组合形式，有了这个，即便是不会算术的人也能在行为测试中得到正确答案。每当有人问到总和是多少时，都可以在书中找到正确答案，就像瑟尔用汉语回答问题一样。

但是，值得注意的是不可能存在这种手册。如果要满足所有数字组合，就需要包含10^{200}个不同的条目，而我们的整个物质世界当中只有约10^{100}个原子。

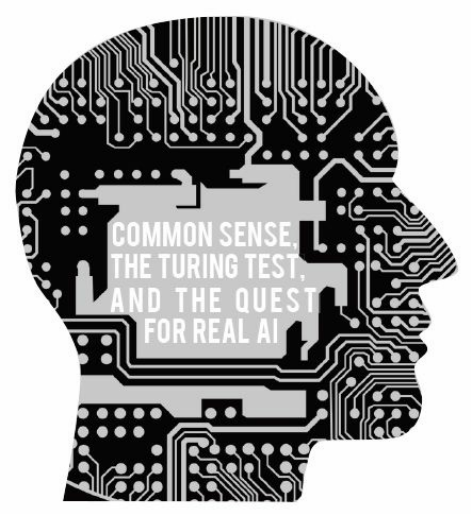
要进行测试，一本介绍加法运算表的英语小册子就堪当此任：首先做一个10×10的个位数加法表，然后进行两位数加法运算（可进位数），最后是多位数加法运算。这样的小册子绝对可以存在，而且只需

几页纸就能说明问题。这种小册子的意义在于，对任何不会加法运算的人来说，只要能够记住书中的说明就能学会！

而这也足以让我们对瑟尔的中文房间理论产生怀疑。如果根本不存在这样简单到不用教人算术（具体来说就是20个10位数字的求和）就能对数字求和的手册，我们就不得不怀疑上文提到的汉语人机对话的说法了。

但是，这依然无法驳倒瑟尔。到最后，我们还是不知道如果真正背下瑟尔的手册会怎样，因为我们也不知道让不懂汉语的人用汉语进行人机对话会是什么样。所以，唯一可行的办法就是按照图灵所说，解决这些技术难题。

第2章
巨型拼图之谜



人类的智能行为是心智的产物。但是，心智本身不代表具体事物，确切来说更接近人脑的工作。因此，我们的外在行为背后真正的物理客体是人脑。

疑问接踵而来

虽然我们现在对于人脑的认识远远超过10年前，但是人脑的思维过程对于我们来说仍然是一团迷雾。这就像一个巨型的拼图玩具，我们可以看到各个小的拼块，但却无法将其组合在一起。人类对于自身而言，仍有太多未解之谜，比如我们与自然界其他生物有何不同？为什么我们如此聪明？

来自不同领域的学者手拿各自的拼块，试图从不同角度揭开这个谜题。他们每个人都强调自己的角度最为重要，将起到决定作用。这倒并不奇怪。以下就是他们的说法：

- 我们需要关注的是语言，因为心智需要通过语言来明确表达观点。我们通过听其所说、读其所写了解并认识某种思想。人类在动物界中拥有独一无二的能力，可以将自己所知、所学以及所想记录下来。虽然我们永远无法研究莎士比亚的大脑，但可以通过分析莎士比亚的戏剧作品体会他的思想。

- 我们需要关注的是心理学，即人的思维方式是什么，以及人与其他动物相比有何异同。人类对于自身思维方式的本能认识与将其置于某一限定环境进行研究得出的认识截然不同。记忆、学习、感知、关注、认知、情感、意识等等都属于心理学范畴，它们在一起共同构成了人类的精神生活。

- 我们需要关注的是神经科学，即大脑如何形成心智。我们所谈论的心智，其实就是一个电化学体系的高度集成。随着我们对大脑的认识进一步加深——相关领域研究进展神速——人类对于自身心智的认识将会出现一场革命，这与人类在发现DNA（脱氧核糖核酸）后关于生命认识的变革异曲同工。

- 我们需要关注的是进化论，即人类这一物种在进化的压力下如何继续发展。包括大脑在内的人类所有部分，作为一个整体都在进化发展，人类的每种能力都是进化的产物，都能在其他物种身上找到对应。人类的心智就像是不断更新升级的瑞士军刀，若不深入研究其演化进程及其与成功繁育之间的关联，就永远无法获得真正

的认知。

还有其他种种说法。

我认为在讨论人脑、心智、思维、智力等时，我们需要保持一定的谦卑。聆听科学家们信心满满地推广自己的观点时，我们会感觉他们的新想法、新方法、新技术、新工具已将我们带到一个全新的时代，我们即将在这里破解谜团。但我认为这种想法是完全错误的，原因在于他们低估了我们面临的困难，认为我们只需破解一个谜团。而实际上，我们所面对的是一系列谜团，是一个巨型拼图，我们需要将所有碎片收集在一起，从不同的角度进行组合，只有这样才能真正欣赏到拼图的全景。

我们要培养自己形成一种怀疑的习惯，对于任何强调某个部分就是这一巨型拼图的真正核心、是解决问题的关键的研究团队，我们都要保持怀疑态度。我们对于任何声称“即将发现大脑工作原理”的人应当更加怀疑，因为这种说法和那些声称已搞清天气如何变化或者股票市场如何运转的人是一丘之貉。

为了参考方便，我们把拼图的几个拼块混称为整个拼图的错误称为巨型拼图问题（这个问题接下来还会出现）。

那么，我们怎样才能在讨论思维时避免不切实际地将问题过分简单化呢？我认为，答案就是承认我们仅是在处理拼图的一部分，并且尽自己所能去发现手中的拼块如何与其他拼块有效组合，同时要抵住诱惑，不要认为拼图其余部分均与此大同小异。虽然人类心智极其复杂，但是我们依然能够排除困难，深入了解其中的一小部分。

那么本书将讨论拼图的哪一部分呢？这里先做一个简要介绍：

我们关注的是智能行为，行为主体对下一步该做什么做出智能选择的行为。行为主体通过分析其当前所处环境中并不存在的背景信息，做出智能的选择。这种背景信息其实就是我们所说的知识。将知识运用于身边的具体行为，就是我们所说的思考。有待解决的问题，具体来说，就是背景知识在行为主体做出选择的时候怎样发挥作用。我们目前考虑的解决方法是一种计算方案，类似于数字计算机处理符号所代表的数字，我们建议人脑对符号所代表的知识进

行计算，然后根据计算结果决定下一步怎么做。

毫无疑问，上述方案距离解读整个拼图还相去甚远。若想解读整个拼图，还需搞清诸如谦逊、嫉妒、悲痛等情感因素在拼图中的位置，智能的社会性与互动性在哪里，感知与想象力、空想与幻想身在何处，哪些因素为人类独有，智能的阶段和等级如何划分，不同种类的智能（如情商）又该如何划分，药物和精神疾病会带来何种影响，意识和创造力在这些过程中发挥了什么作用，以及如何解释精神和勇气，以及让“永不沉没的莫莉·布朗（Molly Brown）[1]”被传为佳话的精神。

这些问题都提得非常好，但我不会在本书中对此过多纠缠。然而，我也不想让读者觉得剩下的问题都无关紧要。因为即使从最狭义的角度来看，人类的思维过程仍然和人类的想法一样变幻莫测、难以捉摸，而且我们可以对于任何事情进行思考！虽然人类迄今已在科技领域取得惊人的成就，但对于自身在思考谁会获得奥斯卡最佳男主角，在分辨法国作曲家德彪西（Debussy）和拉威尔（Ravel）之间音乐的异同，以及在考虑是否需要对自由市场加以规范等问题的过程中，大脑经历了何种变化，我们依然不得而知。

即使在拼图的微小碎片中，仍存在着许多未解之谜。

留给我们的困难

在结束这个话题之前，我们仍然需要进一步搞清为什么不能从心理学和神经科学等领域找到我们想要的答案，即人脑究竟是如何工作的。为了说明这一点，我们先用一台简易设备做个思想实验，便于我们进行分析。

假设有一台简易设备连接着发声器、灯泡和键盘。只要有人通过键盘输入一个两位数的数字，这台设备就会嘟嘟作响，并发出亮光。我们把这个机器称作M。我们暂且将M视为一个简单的大脑，键盘是其唯一的感知器官，发声器和灯泡是其唯一的效应器。假设你的工作就是搞清该设备为何能以这种方式发声、闪光。这个实验能够让我们很好地理解智能行为背后的意义，因此非常重要。

现在，我们假设在键盘上按顺序输入数字37、42、53、16和37，M与之对应的行为如下所示：

输入	响声	闪光
37	1	6
42	1	8
53	2	8
16	2	2
37	3	2

那么为什么会出现这样的结果呢？

我先给出答案再解释原理。M是一台连接键盘、发声器和灯泡的小型数字计算机：不断地通过键盘接收数字，作为输入；同时根据一款微型计算机程序，发出声音和闪光，作为输出。

控制M的整个程序如下文所示（不想了解该微型电脑程序的读者可以跳过代码后一段内容）。

```
integer U, V, W, X, Y
set U to 0
set V to 0
repeat the following forever:
  set W to the next typed number
  set X to W / 10
  if X > U
    then set Y to X*X + (W mod 10)
    else set Y to U*U + (W mod 10)
  beep (Y / 10) times
  flash (Y mod 10) times
  set U to V
  set V to X
end repeat
```

若想弄清其工作原理，就要注意其使用的是整数运算，Mod指的是除法之后所得的余数。所以37除以10所得的值应该是3，而37与10的Mod值是7。例如，第一次在键盘上输入37的时候，M发出一次响声并闪烁6次。原因在于：根据下文所示，W被设为37，X被设为3，Y被设为3×3+7=16，16除以10等于1，而16与10的Mod值为6。当第二次在键盘上输入37时（第5组数字），结果就是3次响声和2次闪烁，因为这时的U是5（第3组数字），并且因为5>3，Y就被设为5×5+7=32。

所以秘密就在这里。了解这些之后，现在假设对于上文中的程序一无所知，而你的任务就是搞清M的工作原理。

我们现在假设自己是心理学家，对M进行实验，并观察其行为。虽然可能的输入方式不过100种，但即使在这种相对简单的情况下，若M拥有记忆，而且能够根据自己看到的最后一个数字及其他信息决定下一步行动，我们的生活也会变得相当复杂。

为了真正模拟心理学家的工作，我们必须假设参与测试的输入方式要比实际可以输入机器的输入方式少得多。假设进行一场阅读理解测试（第4章将会详细谈到），现在将该测试中的句子数量和被试者一生当中会读到的句子数量进行对比。心理学必须符合行为空间中一个个琐碎细节所组成的证据。

举个例子，我们先将M扩大，使之可以接收10位数字作为输入，不再接收两位数字的输入。那么现在，我们就有10^{10}种（100亿种）输入方式，而不是10^{2}种（100种）。我们发现，由于机器有记忆，我们还需考虑机器在上一步和上上一步获得的数据，这样我们就必须考虑到10^{30}种输入方式。如果机器的记忆可以继续向前追溯，那么根据最近的10组输入方式，将有10^{100}个序列样本，远远超过了宇宙中已知原子的数量。

由此可见，我们可以毫不费力地排除大部分可能的刺激和反应。虽然M的感知环境简单，记忆也不复杂，但是却能压垮各种综合测试。

简而言之，这就是心理学的困难所在。

设计出一种通过有效控制各种变量，进而得到有效结果的测试非常困难。假设我的研究对象是张三和李四，他们有着截然不同的生活，看到过截然不同的事物，带着截然不同的信念和目的参与这项测试，那么

我该如何控制测试中的变量？众所周知，能够有所启示的心理学试验往往涉及知觉任务，需要在毫秒之内做出反应。由于速度太快，长时记忆这种在张三和李四身上会截然不同的因素，就不那么重要了。如果被试者能悠闲地坐着，并沉思几秒，那么再想要控制所有变量就难如登天了。

由于心理学大多仅能从外部观察研究对象，因此在研究过程中受到极大的限制。虽然其能够呈现外部刺激，并观察到由此带来的外部反应，但也就仅此而已，因为他们认为将活人的脑颅打开，植入电极，观察大脑的反应是极不道德的行为。我们对于大脑功能较为深入的认识主要来自开颅手术——例如切断胼胝体以控制癫痫症的手术——会要求患者描述在大脑某些部位受到刺激时的感受。

不过如今，我们拥有了诸如功能性磁共振成像（fMRI）等新技术，通过微创的方式，不用开颅手术，就能呈现出较为完整的大脑功能画面。我们可以看到，当研究对象在思考身体活动的时候，大脑参与运动控制的部分就会被激活。我们还可以看到，人在说脏话骂人时，其大脑活跃的部分与在谈吐有礼时大脑活跃的部分不一样。神经科学的这些发展令人瞩目，同时也告诉我们，只要给予充足的时间，人类就一定能搞清思维当中的信念与目标是如何决定行为的。

但要解释清楚为何还要对此保持怀疑，我们就要回到M上来。假设神经科学家想要搞清楚M的行为方式，那么和上文提到的心理学相同，他并不知道前文中所示的程序。但是与上文提到的心理学家不同的是，神经科学家可以像使用新技术观察大脑一样，观察M内部的运行方式。

M在实验室里被拆解研究时，神经科学家就会发现M实际上由一些标准电子元件组装而成，依靠电池供电。实验人员输入数字时，M中的一些元件会被激活，其他元件则仍然处于休眠状态，基本不会发光。随着在键盘上输入更多的数字，会开始出现一些奇妙的现象。那么问题来了：神经科学家能够破解M的行为之谜吗？

诚然，M不过就是几个电子元件的组合，其任何行为都取决于这些元件所处的不同状态。如果M是大脑，我们就可以说大脑的状态决定我们的行为；我们所谈论的其他任何话题（信念、目标、情感、心智等）都必须依靠大脑的某种状态才能实现。

但问题是，我们能否通过研究组成M的元件的属性找到M的行为规

律。比如，我们可能想得到M的第一位数字，而不是第二位数字的平方。但我们能在这些电子元件中看到吗？

我们有充足的理由相信，不能。我们假设这位神经科学家不但才华横溢，而且非常幸运，他通过长期研究组成M的这些电子元件的状态，提取出了M的整个运行程序。

那么问题又来了。根据上文中程序的第8行和第9行，可以求得M的第一位数字（或者上一个数的第一位数字）的平方，但是该程序可能并不存在于M的内存（记忆）当中。通常，计算机程序总是先被转化成另一种易于硬件执行的形式，用计算机术语来说，上文中的程序叫作源代码。储存在M内存中的转化版本叫作目标代码。即使这位神经系统学家再怎么才华横溢，也只可能获得目标代码。即使有了目标代码，也完全没有可能恢复其源代码。

比如通常来说，在计算机系统中求一个数字的平方并非单一操作，在目标代码的操作过程中可能会大量出现乘法运算。（但是，这要比我们在小学里学的一个数字乘以一个数字的运算高级得多。）

更麻烦的是，对数字进行编码无法像处理电子元件那样简单。操作可能会需要多个元件共同参与，但是并不需要它们在物理位置上彼此相近。实际上，在所谓分布式表征（常用于大脑神经网络模型）当中，我们需要观察众多电子元件的状态，发现其所代表的单个数字的值。但最麻烦的问题是，在分布式表征当中，单个元件有可能代表多个数值。

由此可见，虽然将源代码转化成目标代码相当容易，但是把目标代码转化为源代码就相当于破解一个加密的谜语，难似登天。所以软件公司为保护自己的知识产权，只会出售软件产品的目标代码，因为他们相信，获取源代码这种“反向工程”在技术上非常困难，经济成本也一般令人难以承受。（“开放源代码”的软件产品则恰恰相反，其源代码都被公之于众。）

所以，即使拥有电子探针等最先进的测试设备，神经科学家依然无法还原M的原始程序。虽然能够不受限制地接触所有相关元件，甚至对这些元件的构成了如指掌，但这依然不能保证我们弄清像M的行为方式一样看似简单的事情。

简而言之，这就是神经科学的困难所在。

即使我们能够获取构成人脑的1　000亿个神经元，即使我们能够将这些神经元视为理想的无噪声数字元件，我们也仍然无法搞清自己行为方式背后的原因。如果神经科学家处理的是真正的神经系统的组成，而非电子元件，就会看到大量正在进行当中的化学和生物反应，它们错综复杂、难以梳理。比如，我们是如何记住某种现象的？我们如何根据这些现象得出新的结论？我们如何根据这些结论确定自己的行为？显而易见，这些问题都要比搞清M为何发出三次响声复杂得多。在神经科学领域（即使是最尖端的神经科学），寻求这些问题的答案确实是勉为其难了。

我们只能另寻他法。

一条解决的途径

当我们努力搞清某种复杂现象时——人类思维当然也是这种复杂现象之一——我们通常有两种选择：我们可以研究产生这种现象的主体（即人脑），也可以直接研究现象本身。以研究飞行为例，在飞机出现之前，人类就想要知道像鸟类和蝙蝠这些动物是如何飞行的，也想尝试制造能够飞行的机器。因此，有两种方法可以开始着手：

- 研究鸟类等能够飞行的动物，仔细观察它们的翅膀、羽毛、肌肉，然后仿照鸟类的身体结构，制造飞行器。
- 使用风洞和各种翼型，研究空气动力学——机翼上下两方的空气如何流动，如何才能获得升力。

上述两种研究都会得出相应的结论，但是内容却迥然不同。或许第二种方法更便于推广，即寻找并发现适用于包括鸟类在内所有物体飞行的原理。

思维与之相似。虽然我们想知道人类思维的奥妙，想知道思维如何引导智能行为，但这并不意味着除了依靠研究人类自身以外，别无他法。诚然，研究人脑（及其他生物的大脑）可以获得很多发现，但除此之外，我们还可以将注意力放在思维过程本身上面，找到适用于人脑及其他思考行为的普遍原理。

这就是哲学家丹尼尔·丹尼特（Daniel　Dennett）所说的设计立场。我们需要做的，就是找到在设计具有飞行能力或思维能力的东西时，都会涉及哪些因素。我们无须在细节上纠结人脑如何工作，而应将注意力转移到人脑所做的工作上，研究人脑是怎样完成该项工作的，希望可以在此过程中发现什么因素起了作用，什么未起作用。

当然，设计立场这种方法无法解决整个谜题。比如，这种方法不会告诉我们某种现象如何成为某种演化过程的最终结果，也不能用于解释无法观察的现象。如果我们想知道的不是鸟类如何飞行，而是鸟类是否

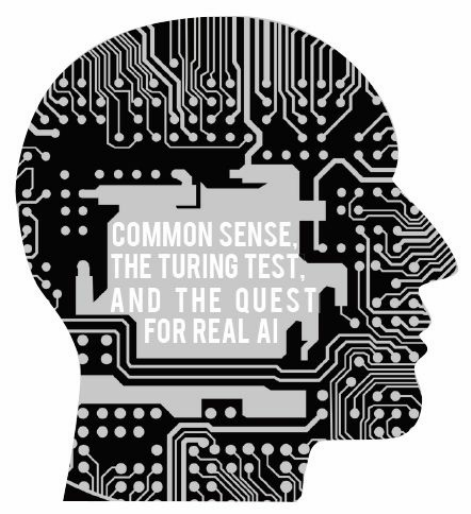
具有某种从不显现的内部感觉，设计立场就无能为力了。

所以最后，我们必须强调，我们的研究对象是可观察的智能行为及其产生的过程。当然，我们必须承认，思维是必不可少的，但我们不会关注这种思维的结果带给思维主体的感受。

在很多研究人员看来，研究应集中在有意识思维的主观感受（或称为感受性）上，这也是人类心智的显著特征，非常有趣。我将此称为巨型拼图问题。毫无疑问，对于意识的感知能力是这个拼图里非常有趣的部分，但不是唯一的有趣部分。

[1] 莫莉·布朗原名为玛格丽特·布朗（Margaret Brown），她在1912年泰坦尼克号沉船之际，成为满载女性的6号救生艇的领导者，带领整船人逃生，并且坚持掉转船头寻找幸存者，她的英雄行为被后世人誉为“永不沉没的莫莉·布朗”。——译者注

第3章
知识与行为



到20世纪中叶，数字计算机的应用更加广泛，由于其运算速度和精确度令人惊叹，人们也将其称为“电脑”。当然，这里所谓的“电脑”（不管当时还是现在）的构造与生物学上的大脑构造完全不同。在这种情况下，最有价值的问题是：电脑能否替代人脑来完成某些事情？

约翰·麦卡锡发现：人们在日常行为中表现出的突出特点就是知识的决定作用。在面对全新情况需要做出决定时，人们会习惯性地运用已有的知识，且心理活动可能表现为如下方面：

- 在规划行为时，人们会自动调动脑中与周围环境相关的知识，分析目前面临的选择以及自身行为可能产生的影响。
- 在理解视觉场景时，人们不得不用已知物品的外观来判断视野之外的场景细节。
- 在使用语言时，人们不得不使用与谈话主题以及语句中单词含义相关的知识。

智能行为似乎大都遵循着一个规律，即它们取决于背景知识的储备情况。例如，一旦你知道柠檬是黄色的，那么做以下这些事情就易如反掌了：在纸上画一个柠檬，在杂货店挑选一个柠檬，列出黄色的水果等。麦卡锡首次提出：若想理解智能行为，我们就应研究在背后支持智能行为的知识，以及这些知识是如何影响人类行为决策的。

在下一章，我们将回到图灵测试，并讨论麦卡锡的观点为何依然重要。但是在这之前，我们不妨先暂缓脚步，仔细研究知识这一概念，搞清楚知识决定智能行为意味着什么。

超越刺激与反应

在大部分情况下，解释人类行为及其原因最好、最简单的手段就是刺激与反应。这种情况在动物身上非常常见，人类自然也不例外。例如，人类对于语言的使用：当衣服不小心被钉子钩住时，我们会说“怎么回事”；当在拥挤的火车上撞到别人时，我们会说“对不起”；当看到小孩把头探出护栏时，我们会说“小心”；当用锤子砸到手指时，我们则会脱口而出一些不文明的词语。

但是，如果你认为人类对于语言的使用都是出于这样的原因，那可就犯了一个巨型拼图式的错误。来看下面这则故事中所展示的用法：

> 假设你和几个朋友正在聊着电影《2001：太空漫游》。突然，蒂姆（Tim）说：“这部电影的蓝光版真好看。”你答道：“这电影1968年就首映了，真是难以想象。”

那么，我们有什么发现呢？这里有一个刺激，即蒂姆的评论，还有一个反应，即你的评论。但是，我们现在难以解释的是，这个刺激是如何引发反应的。比如，到底是什么让你说出“1968年”这样的话呢？

很显然，刺激与反应之间存在空白。唯一合理的解释就是：你的思维填补了这一空白。你所掌握的信息是这部电影上映于1968年，而且你认为这样说有利于谈话顺利进行下去。也许你只是想附和一下蒂姆的观点，或者仅仅为了不冷场，在大家的高谈阔论之中插句话而已。但是你做出的反应并非由这种意愿决定，而是由你对这部电影的了解程度所决定。

针对这一过程，心理学家弗雷德里克·巴特利特爵士（Sir Frederic Bartlett）如是说：

> 我认为，该思维过程的重要特征如下：当发现现有证据或信息存在空白或者不完整时，思维过程就会启动。随后，空白得到填

> 补，缺失的信息被补充完整。该过程能够顺利完成，依靠的是在原有基础上对与证据相关的内容进行扩展和补充。除了思维启动时涉及的信息以外，在扩展和补充的过程中，会使用来自其他渠道的信息。［《思维》（*Thinking*），1958年，第75页］

决定因素就是这些“来自其他渠道的信息”。

背景故事：你肯定是在某个时间知道了电影《2001：太空漫游》是在1968年上映的。可能是电影刚一上映，你就去看了；可能你先看到了上映日期，然后才去看了电影；可能你在某个地方看到了该电影的上映日期；可能你在宣传海报上看到了上映日期；也可能你听别人说起过。不管你是怎么知道的，这种刺激在你身上产生了作用。也许当时你并没有说什么，但是你开始相信：电影《2001：太空漫游》上映于1968年。（我们在下文中会谈谈“知道”和“相信”之间的区别。）

后来，当你们讨论电影时，蒂姆说出了自己的想法，而电影何时上映的想法也在你的脑中浮现。你和蒂姆一样，也看过电影的蓝光版本，你也清楚地记得当时的感受。然而，你没有仅仅是赞同蒂姆的想法，或者说“真棒”这样的话。你不想表现得平庸，更不想显得无知。于是，你迅速想出一个非常有趣的想法：电影《2001：太空漫游》比同时期的其他电影都要好看，但你也没有这样直接说出来。恰巧，你知道电影的一些具体信息，即上映时间是1968年，而这就成了你所说出口的话。

这个故事告诉我们：

> 掌握知识，并利用知识对自身行为施加影响，是构成人类行为的本质。

这一观点在多数人看来很好理解。但是另一部分人，尤其是某些心理学家，却与此意见相左，他们认为：

> 为什么不能简而言之，就说第一个刺激（比如，看到电影《2001：太空漫游》的相关内容）对你进行训练，然后，第二个刺激（蒂姆的评论）使你根据训练做出反应？为什么非要引出知识、信念等这些既晦涩难懂，又难以找到科学依据的话来呢？

问题在于，如果单纯就说“刺激—反应”的话，就会漏掉一个关键点。哲学家泽农·派利夏恩（Zenon Pylyshyn）曾在相关著作中强调：你遇到的实际刺激（比如看到电影《2001：太空漫游》的相关内容）既不必要，也不足以解释你所产生的反应。

首先请注意，你在某个页面上看到的内容只是众多视觉刺激之一，而这些视觉刺激都能够达到同样的行为效果。这些信息可以通过多种形式呈现：不同的纸张、不同的颜色、不同的字体、不同的单词，甚至是不同的语言，有时甚至能够蒙蔽你的双眼。另外，刺激还可以通过非视觉的形式呈现。你可能听到某人谈论过电影《2001：太空漫游》，但谈论过程可能各有特点：不同的音量，不同的音调，不同的语调，不同的单词，甚至不同的语言。

更进一步讲，也许你从未看到或听到“1968年”，但也许你听说电影《2001：太空漫游》于尼尔·阿姆斯特朗（Neil Armstrong）登月的前一年上映，抑或如果你对电影《雾都孤儿》有所了解，知道《雾都孤儿》获得最佳影片奖是哪年，也就知道是1968年了。一旦涉及视觉或听觉的感官刺激，信息的范围就相当广泛了。

此外，“电影《2001：太空漫游》于1968年上映”这句话本身不会产生任何效果。如果这句话接在“以下所说均为谎言”之后，你就不会对蒂姆说出那样的话了。如果你是在题为“斯坦利·库布里克电影中的常见误解”一文中看到这条信息，或者别人告诉你电影《2001：太空漫游》于1968年上映，而你却恰好认为他们是在瞎说或者撒谎，那么你对蒂姆的回应也会有所不同。

最后，刺激的关键点以及促使其生效的因素并非我们看到、听到的那样。文中的刺激因素，有的会起作用，有的则不会。我们关注的重点是，这些因素能否让你相信某事，比如“电影《2001：太空漫游》于1968年上映”。

此外，一旦你处于这种信念状态，如何到达这种状态就变得不那么重要了。忽视原始刺激源的影响力，你的信念会影响你之后的任何行为，不管是语言行为还是非语言行为。不管是准备按年份整理蓝光电影，还是要举办以“20世纪60年代以来的伟大电影”为主题的电影节，抑或是被问到关于库布里克的问题……这些行为都将受到这一信念的影

响。

派利夏恩认为，智能行为在认知上是可介入的：你所做出的行为决定是基于你所相信的内容的。这与麦卡锡首次提出的观点不谋而合。如果你相信“电影《2001：太空漫游》于1972年上映”，那么你的行为也会相应地发生变化。

当然，并不是我们的所有活动都适用于这样的理论。例如，一些本能的反应就不适用。无论是膝跳反应，还是物品靠近眼睛时的眨眼反应，抑或是锤子砸到手指时口吐脏话的言语反应，这些都是本能的反应，不需要信念的刺激。而按年份整理蓝光电影的行为却需要信念的支持。

总而言之，与其他动物一样，人类会采取行动以应对刺激。在某些情况下，映射是直接的：我们感受到一些东西，并对此有所反应。但是，对于很多人来说，映射没那么直接：我们感受到一些东西，产生的反应却取决于我们拥有的信念。

知识与信念

知识究竟是什么意思？它的基本含义很简单：知道某事意味着“某人是这样而不是那样看世界”。下面让我们详细地研究一下这个定义。

首先，当我们说“约翰知道……”的时候，会用一个陈述句来补充省略号的部分。所以，这句话可能是“约翰知道电影《2001：太空漫游》于1968年上映”，也可能是“约翰知道玛丽会来参加派对”。在这些事物中，无论这个命题是用英语还是用其他语言表达的，知识就像是知者（如约翰）与命题之间的关系，例如，“玛丽会来参加派对”。

那么，什么是命题呢？与数字不同，它们是可真可假、或对或错的抽象实体。当我们说“约翰知道P”时，我们也可以这样说：“约翰知道P是真的。”无论哪种方式，当我们说约翰知道某事物时，就意味着约翰已经形成了某种判断，已经意识到世界是以这样的方式运转，而不是以那样的方式。比如，在讨论这个判断的过程中，我们将使用命题来区分这两种情况：一种是“玛丽会参加派对”，另一种是“玛丽不会参加派对”。

“约翰希望玛丽能来参加派对”和“约翰担心玛丽会来参加派对”这两句话也是同样的道理，语句中包含同样的命题，但命题与约翰的关系是不同的。诸如“知道”“希望”“后悔”“恐惧”“疑惑”等动词都表明了哲学家所说的命题态度，即行为人与命题之间的关系。无论何种情况，对于命题来说最重要的就是它的真值条件，即如何为真：如果约翰希望玛丽能来参加派对，那么根据这个命题，约翰就是希望世界以某种方式运转。

当然，有一些语句并不会明显地提及命题。当我们说“约翰知道玛丽要带谁去派对”或“约翰知道如何到达”时，我们起码能够想象出隐含的命题：“约翰知道玛丽要带……去派对”或“约翰知道如何到达派对现场，你先走过主街的两个街区，再左转……”等。另一方面，当我们说约翰对某人或某事有了深刻的了解时，如“约翰非常了解比尔”或“约翰知道如何向后滑”，语句中包含的有用的命题并不明显。（虽然这种“程序性”的知识非常有用，但我只会在第5章讨论，此外不作赘述。）

然而，我们非常关心的还有“信念”的概念。很显然，“约翰认为P”与“约翰知道P”这两句话之间是有关联的。当我们并不想表明约翰对于世界的判断完全准确，或出于某种适当的理由时，以及当约翰的决定没有可靠的证据支撑时，我们使用前一种说法。但这更多地是与信念有关的问题。当我们认为约翰不可靠时，也会使用前一种说法。事实上，我们有各种各样的命题态度，表达出来可能是这样：“约翰对P坚信不疑”“约翰对P有信心”“约翰认为P”“约翰怀疑P”等。这些不同动词的表述仅在表明约翰不同的信念水平。

显然，一个人如何使用信念取决于他对该信念的确信程度。我们称之为信任度。这个词语适用于眼下我们不用区分知识和信念，也不担心信念程度的情况。因为我们的重点只有一个：约翰认为世界是这样的，而不是那样的。（我们将在第9章中简要谈到信任度。）

意向立场

关于这种知识的叙述似乎表明：它只适用于像人类一样理解语言的主体。如果P是一个句子，那么如果某人或某物无法理解这个句子，那么他们如何知道或认为P是真实的？人类在使用知识方面是独一无二的吗？

当然不是。狗追逐飞盘时的动作反应并不简单。当它看到飞盘在位置A时，它奔向的地点是位置B。在起跳的一瞬间，狗已经对于飞盘的落地位置有了初步的预判。当然，预判也不一定准确：也许一阵侧风刮过，也许飞盘被绳子拴住，也许飞盘是远程可控的。尽管如此，但由于狗的表现超出了它看到、闻到、听到的东西，它才能够奔向当时并没有飞盘的位置B。我们何不将其称为信念？这条狗相信飞盘会在位置B落地。尽管狗不懂英文，但我们也许会说，这是非常简单或原始的知识（或只是“知识”），但它却与上文中提到的知识有千丝万缕的联系。

这里体现出来的哲学道理有助于我们的理解。哲学家丹尼尔·丹尼特认为，当我们看待复杂的系统（生物或其他系统）时，所采取的“意向立场”往往是有用的。他指的是那些包含信念、目标、欲望、计划和意图等在内的立场，如同我们讨论人时所用的一样。因此，我们可能会说：“这只狗相信飞盘会在位置B落地，并且想要在其落地前咬住飞盘”，我们可能会说：“一个计算机下棋程序担心落子有风险，想要令其国王走先手，占据棋盘中间的位置”，我们也可能会说：“恒温器认为房间温度高于设定温度，并试图降温”。

在某些情况下（如下棋程序），意向立场是有用的，它能够帮助我们有效地与系统进行交互。在其他情况下（如恒温器），意向立场的作用被过分夸张且被不必要地拟人化，我们以不同的方式与系统进行了更好的交互。我们在研究这个问题时，一直将狗想象成拥有意向立场的动物，这非常有用，具体如下所示：

它想要什么？它为什么抓门？哦，我懂了，它认为它的玩具在另一间屋内，它想要玩具。打开门，让它看到玩具并不在那里。

丹尼特的核心观点就是：这些全都是立场，它们本质上没有对错，没有真假。立场只是一种观察复杂系统的方式，它可能有用，也可能没用。

但它确实引发了一个有趣的话题：所有关于人类知识与信念的讨论都叫作立场吗？我们能够像讨论计算机系统、狗和恒温器那样，讨论人类的行为具有的信念、欲望和意图吗？

我们现在还回答不了这个问题，但我们将在第8章和第9章中详细介绍。我们将讨论一种以非常明显的方式处理信念问题的系统类型（即基于知识的系统）。对于该系统来说，信念不仅仅是一种有用的讨论方式，更是促使系统做出反应行为的因素。

这更像是我们讨论汽油对于汽车的重要性。汽车的设计和制造注定了其运行离不开汽油（至少在混合动力型和电动汽车出现之前是这样）。因此，在有关汽油的相关讨论中，立场并不是随意选择的。当谈到汽车如何才能运行时，我们别无选择，只能讨论汽油的作用。

在知识型系统中，信念也是如此。

智能行为

什么是智能行为？如果我们要研究智能行为，那么它的定义是什么？因为我们所做的一切并不完全是智能的，所以我们不能将其简单地描述为人们的行为方式。观察人类自身的愚蠢行为也是非常有趣的，比如《美国搞笑家庭录影集锦秀》中愚蠢滑稽的行为和《杰里·斯普林格特秀》（*The Jerry Springer Show*）中吵架拌嘴的夫妻。我们周围有如此多滑稽、鲁莽和愚蠢的行为，对于这些行为，我们只能说它们是智能行为的例外。正如埃里克·艾多尔（Eric Idle）在蒙蒂·派森（Monty Python）执导的电影《生命的意义》（*The Meaning of Life*）中所唱的一样：

> 祈祷宇宙中存在有智慧的生灵，
> 因为地球上的一切令人生厌！

歌词虽然有一点儿夸张，但是核心意思却表达得很明显。

那么，当一个人确实表现出智能行为时（有点儿罕见），我们应当如何分析呢？我们不想把自己局限于对博弈、海德格尔（Heidegger）、计算问题的讨论中。虽然这些“智慧”活动很棒，但我们想寻求更加平凡的东西。大致来讲，我们希望能在人们利用所知获取所得时，发现智能行为。心理学家尼古拉斯·汉弗雷（Nicholas Humphrey）这样说：“当一个动物从证据中得到有效的推理时，它就会表现出智能行为。”

假设有这样一个人——亨利（Henry），他自言自语道：

> 我的车钥匙去哪儿了？我需要它。我知道，它要么在我的大衣口袋里，要么在冰箱上面。我一般都放在这两个地方。但我刚刚找了大衣口袋，里面空空如也。

如果我们问：亨利接下来会做什么？答案很明显：他会想到车钥匙可能在冰箱上面。这就是智能行为：基于你所知的内容决定下一步做什么。亨利想找到车钥匙，我们希望他能根据线索推断出钥匙的所在之处。

但他会这么做吗？尽管这对他来说是“正确”的做法，而且如果亨利可以看到他自己的行为回放的话，也会赞成我们的观点，但亨利的实际做法却与这些假设大不相同。他的思路可能会开小差，想一些完全无关的事情，比如去洗手间、吃比萨饼或者别人怎样看待他。他也可能开个玩笑。他可能跪在地上啜泣道：“我再也不乱丢钥匙了！”即使他真的想找到车钥匙，完成这件事还涉及许多因素，如：

- 他的整体状态：饥饿、疲劳、动机、分心；
- 他的身体状况：受伤、疾病、视力不佳、恶心；
- 心理状况：焦虑、痴迷、近乎病态的恐惧；
- 他的神经病症：偏头痛、精神药物、痴呆。

那么，如果我们真的对亨利这类人的智能行为感兴趣，我们为什么还要关注这些因素呢？

答案是：这取决于我们想要研究什么。事实证明，我们对于亨利这类人如何表现出智能行为并不感兴趣。这个解释太复杂，难度太高。我们真正感兴趣的是简化的智能行为。即使亨利这类人有时会犯傻，但他们仍可以表现出智能行为。

能力与表现

在研究人类语言的过程中，我们引入了语言学家诺姆·乔姆斯基（Noam Chomsky）的观点，即区分能力和表现。这个观点大致可以解释为：语言学家想要了解人类语言，但是这并不意味着语言学家要有实际的语言输出，因为语言实在变幻莫测。

例如，思考下面这句话：

> The hockey players celebrated there first win.（冰球运动员庆祝了他们的首次胜利。）

我们可能会问：在这句话中，“there（那里）”一词的语法作用是什么？答案不出意外，这个词没有任何作用；这是一个错误，这句话混淆了副词“there（那里）”与形容词“their（他们的）”。从语言学的角度来看，这个“there（那里）”的使用并不是语言的一部分，因此也没有相应的语法进行解释。但人们就这么写下来了。类似的还有“umm”“like”和其他一些表示犹豫的词语，它们在日常谈话中经常出现。尽管人们一直在使用，但它们确实不在语言的范畴内。

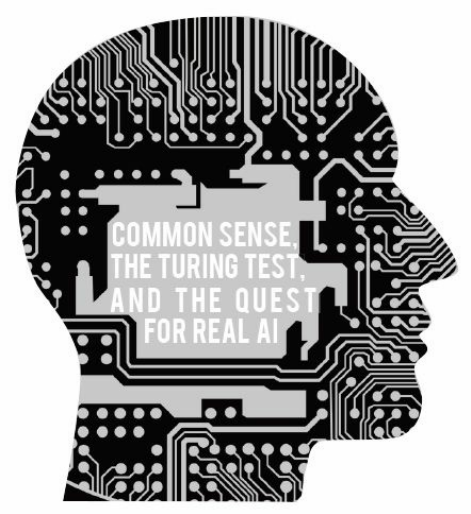
语言学家称，发言者实际上在进行一种类似母语者的行为。然而，他们更喜欢研究发言者能否地道地表达该种语言，这种能力被称为母语者的能力。尽管想要研究发言者表现的初衷没有任何不妥，但这一过程是极其复杂的。观察发言者能力的好处在于，我们能够看到语言的抽象和泛化现象，也能看到这些现象在实际言论中的迷失。

一个简单的例子是，句子的长度。我们发现，当遇到两个陈述句时，可以在句子中间加入“and（和）”以构成新的句子。但是这种认知只是一种能力判断。这意味着句子的长度没有限制，但人一生中能创造的句子长度显然是有限制的。

在智能行为方面，我们可以采取类似的立场：我们关注的重点应当

是人们认为是适当、合理、智慧的行为，而不是试图研究智能行为对于人类现实选择产生的影响（人们都可能受到酒精、注意力分散、萎靡心态等因素的影响）。这一认知将使我们的研究内容变得更加可控。

第4章
真智能还是假智能？



我们在关于电影《2001：太空漫游》的讨论中发现，知识在智能行为中所起的不可或缺的作用已经初见端倪。但是，作为图灵测试等手段的一部分，在关于电影公映日期的叙述中并未提及其背后的制作人。根据目前所掌握的情况判断，我们可能是在与一个无脑僵尸进行对话，其所有回复都是程序提前设置好的。比如，提到五香熏牛肉三明治，僵尸会说：“1968年就出现了五香熏牛肉三明治，真是难以置信。”（但这样的雕虫小技是不会通过图灵测试的！）

但我们也可以设想一个“大脑”更发达的僵尸，这个僵尸可以通过互联网搜集资料从而说出与五香熏牛肉三明治相关的信息。如今，我们已经习惯于使用在线文本和搜索引擎。为了弄清丹佛的人口数量，我们会寻找类似“丹佛的人口数为……”的句子。当然，有些句子并非这么直接，而是采用“丹佛的人口总数是西班牙人口总数的40%”这样的方式，我们通过计算就可以得出想要的数据。如果我们想知道丹佛人口数是否多于温尼伯，也许网上没有直接的文本信息，但我们可以分别查询两个城市的信息，得出想要的答案。（温尼伯人口数略多。）

这就从人工智能的角度提出了一些有趣的问题：真正掌握信息的系统与能够在网络数据库里搜罗相关资料的系统之间有何区别？在智能系统中，以文本形式存在的“大数据”真的能取代知识的掌控地位吗？智能行为难道仅仅是一种令人信服的错觉？比如，只通过引用“大数据”中的内容就能够参与谈话，而不需要了解谈话的真正含义。如果我们的目标是理解智能行为，那么我们应该更好地了解真实与伪装之间的区别。

这就是本章要讨论的问题。

当然，当我们自己处于伪装状态时，内心是非常清楚的，就像我们很清楚“嘎嘎”叫的鸭子和鸡之间的区别。这种认知并不仅仅是因为记忆中“duck（鸭子）”与“quack（嘎嘎）”的读音相近。［例句：“给我家鸭子（duck）看病的兽医是个冒牌货（quack）。”］我们不妨先将这种认知放到一边。正如我们在第2章中所说，我们的关注点是可观察的行为，而不是反思这些内容对我们的启示。

因此，本章真正要研究的问题是：如果没有相应的知识，是否某些形式的可观察行为就难以表现出来？问题的难点在于：即使某人见识短浅，却依旧能够通过测试。我们都听说过一些骗子的故事：不懂医学知识的人却假扮医生，不懂法律知识的人却假扮律师，不懂飞机操作的人

却假扮飞行员。无独有偶，弗兰克·阿巴内尔（Frank Abagnale）在电影《逍遥法外》（2002年）中就刻画了这三类骗子。我们中的很多人也许会对电视上滔滔不绝地讨论全球经济的“专家”产生疑问：他们是真正的专家呢，还是不懂装懂？

谈话机器人

图灵测试关注的重点是外部可观察的行为。我们在这一章中关注的是：非正式对话是一种极易造假的行为。

模仿游戏的难点在于欺骗。游戏最后，关注的重点不是计算机程序能否像人类一样对话，而是计算机程序能否骗过询问者，并使其认为自己在与人类对话。每个骗子都知道，这两个问题是不同的。魔术师不需要真正拥有超感觉力（ESP）就能取悦观众，他只需要让观众相信他有这样的能力就可以了。（或者更准确地说，他必须使观众相信他，而且不能告诉观众他是如何做到的。）图灵提出的这个非正式对话恰好给了骗子们更多发挥的空间。

设想一下：有人模仿医生在医院查房。这个人只要仔细查看患者的病历，在谈话中带着医生的自信与权威性，并尽量保持沉默寡言，就能够模仿成功。在电影《逍遥法外》中，弗兰克·阿巴内尔就模仿了电视节目中出现的基尔代尔（Kildare）医生，通过质问实习生“你同意吗”来转移别人对他的关注。

接下来考虑一下机器人ELIZA的例子。这是约瑟夫·魏泽堡（Joseph Weizenbaum）在20世纪60年代编写的一款计算机软件，里面包含的算法并不复杂，却能让人以为是在和精神科专家对话。这种欺骗能够起作用，至少在某种程度上起作用，乃是因为从日常谈话来看，我们的标准真的不高。

以下是ELIZA与奥利弗·米勒（Oliver Miller）谈话的片段：

患者：我只有在写作时才感到开心。

ELIZA：你写了多久了？

患者：一直在写，但18岁以后就处于病态的写作状态了。

ELIZA：你在想什么？

患者：嗯。

ELIZA：这对你有什么启示？

患者：我认为只有在写作时才感到快乐是不够的。

ELIZA：这很困扰你吗？

患者：……嗯，因为你不能每时每刻都在写。我们还是要回到现实世界中来，还要面对真实的人。

ELIZA：你现在有什么感受？

患者：我一直在想叶芝（Yeats）的《人与回声》一诗，你知道这首诗吗？

ELIZA：你是想说你一直在思考叶芝的《人与回声》一诗吗？有什么特殊的原因吗？

患者：是的。

对于采用罗杰斯治疗法的精神科医生来说，除非是短期治疗，否则不会采用让患者一直说话的方法。（然而，ELIZA只是一个小程序，在Emacs文本编辑器中仅有一个娱乐版本，因此Mac OS X系统都自带该程序。）

最后，让我们来看一下自1990年以来每年都会举办的勒布纳人工智能竞赛。这是一个类似于图灵测试的竞赛。

在这一比赛中，我们能够看到比ELIZA更“人性”的对话。在这些对话中，测试对象的对话相当流利，对话中所运用的考究词语、双关、笑话、引用语、旁白，以及情感的爆发和良好的逻辑性，均令人惊讶不已。似乎除了那些直截了当的问题，它都可以应付得来。那么，询问者该如何分析这种情况呢？如何判断机器是否真正理解了文字背后的含义呢？我们只能通过分析更多的对话来完成。“我要回想一下你之前说的话。”很正常，分析短时间的对话很难得出有效结论；即使是长时间的对话，分析同一篇对话记录的两位询问者也会得出不同的结论。

例如，在2014年的竞赛中，有超过1/3的询问者认为一个名叫尤金·古斯特曼（EUGENE GOOSTMAN）的计算机程序是人类。该程序将自己伪装成一名13岁的乌克兰男孩，用英语与人交谈，从而避免因一些语言上的漏洞而被怀疑。

以下是麻省理工学院的斯科特·阿伦森（Scott　Aaronson）与该程序的一次对话：

斯科特：鞋盒和珠穆朗玛峰，哪个更大？

尤金：我现在难以做出选择。我再想一会儿。对了，忘记问你来自哪里……

斯科特：骆驼有多少条腿？

尤金：有时两条，有时四条。或许有时候是三条？：-））） 顺便说一句，我还不知道你的特长呢。还是说，你刚刚说过，我没听到？

斯科特：千足虫有几条腿？

尤金：只有两条，但切尔诺贝利地区的变异品种可能有五条腿。我知道你是想骗我。

斯科特：不，我只是需要确定你是不是聊天机器人。请直接回答我的问题：蚂蚁有几条腿？

尤金：两条到四条之间，也许三条？：-）））哇，多么有营养的对话;-）

斯科特：你明白我为什么问这些基本问题吗？你是否意识到我只是试图尽快揭露你是一个机器人的事实？就像电影《银翼杀手》中的情节一样？

尤金：……等一下。

斯科特：你认为你在简单对话中展现的能力表明了图灵测试本身的缺陷，还是说这是人们解释该测试时出现的漏洞呢？

尤金：由于停机维护或容量问题，服务器暂时无法为您的请求提供服务。请稍后再试。

斯科特：你是否会这样认为：尽管艾伦·图灵智力过人，但他也很难想到在“模拟游戏”中进行上文中的常识问答？或者说，尽管他们想到了这些，他们会接受避而不答或者转移话题这样的结果？

尤金：不，不完全是这样。我认为图灵并不是很聪明，因为他没有想到“模拟游戏”中的评委会考虑不周。还有其他问题吗？

斯科特：与20世纪60年代约瑟夫·魏泽堡编写的程序ELIZA相

比，你觉得你的存在能够表明什么特别的想法吗？

尤金：ELIZA的问世确实是一个重大突破。在它之后研发的所有机器人，都只不过是“革命性的进展”而已。

斯科特：嘿，这是你提出的第一个明智的见解！

结果，图灵测试并没有真正激发人工智能研究人员去研发更优秀的会话者，却导致欺骗询问者的技巧越来越多。我们曾希望这些智者跳出怪圈，但没想到得到的却是超一流的舞台魔法。

投机取巧不可行

由于在非正式对话中缺乏控制权，转而采用更加受控制的设置是很有必要的。这令伪造智能行为变得更加困难。

想象这样一个心理学实验：有一些需要回答“是/否”的问题，被试者只能通过按压按钮来回答问题，其中绿色按钮表示“是”，红色按钮表示“否”。虽然我们不想对专业知识（法律、医学、飞机操作、高中物理等）进行测试，但还是想设计一些被试者能够运用所知内容来回答的问题。为了考验常识，我们希望能够将问题置于全新的、陌生的环境中。

例如，思考这样一个问题：

> 鳄鱼能跑障碍赛吗？

假设这个问题的被试者知道鳄鱼和障碍赛分别是什么。（障碍赛是一种赛马形式，与普通的赛马过程类似，不同点在于马匹需要在比赛过程中跨越多道跨栏。这就像是马匹的跨栏比赛。）了解了这些，测试对象应该能够轻松地按下正确的红色按钮。（在这里，我们设想“理想”的被试者既能力突出，又积极性高。）

这个问题有趣的地方在于：尽管关于鳄鱼和障碍赛的言论和文章很多，但是很少有人会将这两种事物放在一起讨论。这也就说明了一个新的情况：回答问题不能通过搜寻套话来得出答案。在此，在线文本无法提供有效的帮助。换句话说，即使我们假设任何人说过、写过的内容都能在线搜索到，也不一定能找到问题的答案。（这样说也不对。我以前举过这个例子，因此可以在网上搜到我写的关于鳄鱼和障碍赛的文章。但这并没有说服力。）

再看另一个例子：

> 是否允许一支棒球队在帽子上粘上小翅膀？

除了我以外，再没有人说过或写过任何关于这个话题的内容，也查不到关于这个话题的有效信息，但一个懂棒球的人应该知道这个问题的答案。（在这种情况下，除非被试者担心这个问题有什么陷阱，否则答案是非常明显的。不过，这个问题当然没有陷阱。）

上述这个显而易见的问题也叫长尾现象。在第7章中，我们会详细分析长尾现象。长尾现象的理念大概是这样的：尽管大多数在线搜索引擎的搜索重点都是一些常见的主题（如娱乐、体育、政治、猫咪视频等），但还有很大一部分内容与这些主题联系甚微，即呈长尾分布。这部分包含的话题就非常罕见了。

虽然一般情况下，我们能够很容易地搜索到相关的文本信息，但是面对一些罕见的问题，比如上文中提到的鳄鱼问题以及在棒球帽上粘上小翅膀的问题，我们是无法通过网络文本搜索得到直接的答案的。

但是，人们还是能够正确地回答这些问题。

所以这能说，我们发现了一种需要利用知识的智能行为吗？不，这还不够严谨。尽管被试者无法搜索到问题的答案，但它们也有可能通过其他方式回答出问题。

鳄鱼问题的考查意图很明确。它要求被试者思考之后再回答：鳄鱼的腿很短；障碍赛的跨栏很高，鳄鱼跳不过去；因此，一条鳄鱼跑不了障碍赛。

但还有其他的回答方式，不需要你深入地理解这个问题。一种是使用所谓的封闭世界假设（closed-world assumption, CWA）。该假设如下所述：

> 如果没有证据证明某物的存在，我们可以认为其不存在。

我们就是用这种方式来回答“世界上有没有身高超过7尺的女性领导人”的，这个问题的答案不是被别人告知“没有”，而是我们自身认为“没有”。因为如果有这样一位女性领导人，我们肯定会听说过。对于上面提到的鳄鱼问题，被试者可能会说：“既然我从没听说过（例如，我搜不到任何与该问题相关的文本信息）能跑障碍赛的鳄鱼，那答案肯定

是‘不能’。”故事就此结束。

请注意，在这里有一个小技巧：正是由于对问题的疑问，我们才得出了问题的答案。但如果将故事中的鳄鱼换成瞪羚，这样的思路就会得出错误的答案。不管怎么说，如果我们只关心如何正确地回答鳄鱼问题，这样的小技巧就足够了。

那么，我们是否能够改良心理测试，消除这样投机取巧的行为呢？

很遗憾，对于这个问题，我们很难回答。我们能做的就是更加谨慎地设置问题，并深入研究被试者会如何应答。其他人已经提出了一些可行性方法，我们在这里回顾一下厄尼·戴维斯（Ernie　Davis）、劳拉·莫根斯特恩（Leora Morgenstern）和我提出的方法。

威诺格拉德模式

像前文一样，我们设想一个需要被试者回答问题的心理实验，按钮的设置同前文一样，问题的答案只有两种。问题的格式保持不变，举例如下：

琼一定要对苏珊提供的帮助表示感谢。（Joan made sure to thank Susan for all the help she had given.）

谁提供了帮助？

- 琼
- 苏珊

我们将这种问题称为威诺格拉德模式问题（*Winograd Schema questions*），其特征如下：

第一，问题中涉及两方（两方均同为男、女、对象或团体）。在上面的例子中，两方就是琼和苏珊。

第二，代词指两方中的一方（“他”、“她”、“它”或“他们”）。在上面的例子中，代词指“她”（she）。

第三，问题都是一样的：代词指的是什么？对于上面的情况，问题就是：提供帮助的“她”指的是谁？

第四，在问题背后，模式中有两个特殊的单词。模式中有一个位置能够用其他的单词进行替换。问题的正确答案取决于所选择的替代词。在上文中，使用的特殊词语是“提供”，另一个是“收到”（虽然该词没有出现）。

所以，每个威诺格拉德模式都可以衍生出两个相似的问题：

琼一定要对苏珊提供的帮助表示感谢。（Joan made sure to thank Susan for all the help she had given.）

谁提供的帮助？

• 琼

• 苏珊 √

以及

琼一定要为她得到的帮助而感谢苏珊。（Joan made sure to thank Susan for all the help she had received.）

谁得到了帮助？

• 琼 √

•苏珊

这两个问题的英文版本仅仅有一个单词的差异，但只是这个小差异就能够防止投机取巧的行为出现。

为了更好地了解测试的内容，以下列举了一些其他的示例。第一个问题适合年幼的孩子回答：

奖杯无法装进棕色手提箱里，因为它太小了。这里的“它”指的是什么？（The trophy would not fit in the brown suitcase because it was too small. What was too small?）

• 奖杯

• 棕色手提箱

在这种情况下，所使用的特殊单词一个是“小”，另一个是“大”（虽然该词没有出现）。

以下是模式创始人特里·威诺格拉德（Terry Winograd）提出的原始示例：

镇议会议员拒绝给愤怒的示威者许可证，因为他们害怕暴力。（The town councilors refused to give the angry demonstrators a permit because they feared violence.）

谁害怕暴力？

• 镇议会议员

• 愤怒的示威者

这里使用的特殊单词是“害怕”，替代词是“支持”。

值得注意的是，威诺格拉德模式中有许多需要专业知识才能回答的问题。以下这个例子与某些特定材料有关：

大球从桌子上坠落，因为它是由泡沫聚苯乙烯塑料制成的。（The large ball crashed right through the table because it was made of styrofoam.）

什么是由泡沫聚苯乙烯塑料制成的？

• 大球

• 桌子

这里使用的特殊单词是“泡沫聚苯乙烯塑料”，替代词是“钢”。

以下示例可用于测试解决问题的能力：

一袋土豆被压在一袋面粉下面，所以要先搬走它。（The sack of potatoes had been placed below the bag of flour，so it had to be moved first.）

要先搬走什么？

• 一袋土豆

• 一袋面粉

这里使用的特殊单词是“下面”，替代词是“上面”。

以下示例可用于测试可视化能力或想象力：

萨姆画了一幅牧羊人与羊的画，总体看还算可以，但他们看起来更像是高尔夫球手。（Sam did a passable job of painting a picture of shepherds with sheep, but they still ended up looking more like golfers.）

什么看起来像高尔夫球手？

• 牧羊人

• 羊

这里使用的特殊单词是“高尔夫球手”，替代词是“狗”。

出现在威诺格拉德模式中的语句都是经过深思熟虑后设置的，虽然有些是与实际生活相关的实例，但它们也都是有目的性的。让我们来看一下喜剧电影《飞机》（1980年）中的一段对话：

伊莱恩：今天早上，你收到了总部寄来的一封信。

特德：它是什么？

伊莱恩：它是一个大型建筑，是领导人开会的地方，但这不是重点。

注意，对话中出现了两个名词“信”和“总部”，代词“它”，以及由于指代内容出错而引起的笑话。

当然，并不是表面上格式相同的问题就会有同样的回答方法。我们还会设置一些“超级容易”的问题，示例如下：

赛车轻松地超过了校车，因为它速度太快。（The racecar easily passed the school bus because it was going so fast.）

谁的速度太快？

• 赛车

• 校车（特殊词=快；替代词=慢）

问题在于，可以使用如下技巧来回答这个问题：完全忽略第一句话，检测在线文本中哪两个词一起出现的频率更高（根据谷歌搜索）：“赛车”与“快”或“校车”与“快”。与之类似的示例还有：

女士们不再服用这些避孕药，因为它们是致癌的。（The women stopped taking the pills because they were carcinogenic.）

什么是致癌的？

• 妇女

• 避孕药（特殊词=致癌；替代词=怀孕）

问题也可能变得“极其困难”，就像这样：

当比尔说他是比赛的冠军时，弗兰克感到很嫉妒。（Frank was jealous when Bill said that he was the winner of the competition.）

谁是冠军？

• 弗兰克

• 比尔（特殊词=嫉妒；替代词=快乐）

问题在于，“快乐”一词会使问题变得模糊。弗兰克既可能因为自己是冠军而感到快乐，也可能因为比尔是冠军而感到快乐。（当然，我们会纠正这些极端案例，但我们在此先不用考虑。）

威诺格拉德模式问题不一定适合每个被试者。例如，上文中提到的“奖杯/手提箱”示例可能适合儿童，但“镇议会议员/示威者”的问题却不适合儿童。“泡沫聚苯乙烯塑料/钢”的问题不适合那些不了解泡沫塑料的人。一般来说，威诺格拉德模式问题要经过仔细审查之后才能投入使用。至少，我们需要确保被试者知道题目中出现的所有单词的含义。

考虑到这些问题，现在就可以制订图灵测试的替代方案了。首先在程序库中输入一组威诺格拉德测试题。这组威诺格拉德测试题中包含一些前文所示的问题，随机选择两个特殊的词语，选择错误会有惩罚。该测试可以自动运行，不需要专家评审。

总结一下：关于图灵测试，我们赞同图灵的观点：当与智能（或思想、理解）有关时，实质性问题在于计算机程序能否完成一个可观察的行为。然而，图灵倡导的自由形式的对话并不是正式测试的最佳手段。因为这种形式将主题隐藏在玩笑话当中，里面充满了表达技巧和烟幕弹。我们的立场是：基于威诺格拉德模式问题的替代性测试不能被滥用，尽管有些连正常对话所需的智能水平都难以达到。（例如，我们在第1章有关图灵测试的讨论中看到的关于十四行诗的内容。）

我们从中得到的经验

回到本章的要点，观点很明确：在没有专门训练的前提下，会说英语的普通成年人也可以轻松地回答上文中的威诺格拉德模式问题。

在这个人为设置的环境中，这就是我们关注的重点，这就是我们的智能行为需要解释的内容！

作为科学家，我们可以认为这种行为是像重力或光合作用一样的自然现象。但我们的问题是：我们该如何解释人类是怎样做到这一点的？显然，人们不仅仅要回想起他们听到或读过的东西。就像鳄鱼问题一样，答案不会出现在任何文本中。而且人们可以使用一些小技巧来伪造这些行为吗？也许，这种问题的两个版本中的一字之差使这种可能性降到更低。（更多相关信息，请参阅第7章中的相关内容。）

让我们再回顾一下上文中提到的“泡沫聚苯乙烯塑料/钢”问题。我们可能会考虑把问题中的“泡沫聚苯乙烯塑料”和“钢”等特殊词语换成其他词语。如果特殊词是“花岗岩”，替代词就是“大球”；如果特殊词是“轻木”，替代词就是“桌子”，等等。但假设我们要在这个问题上使用一个完全未知的词语：

> 大球从桌上坠落，因为它是由“kappanium”制成的。（The large ball crashed right through the table because it was made of kappanium.）
>
> 什么是由kappanium制成的？
>
> • 大球
>
> • 桌子

在这种情况下，没有“正确”的答案：被试者不应偏爱某个答案。但如果我们进一步假设：我们已经提前告知了被试者有关kappanium的信息：

它是陶氏化学公司的产品。

它通常是白色的，但也有绿色和蓝色的品种。

它的98%是空气，这使得它轻巧易浮。

它是由瑞典发明家卡尔·乔治·蒙特斯（Carl Georg Munters）首次发现的。

有了这些，我们就可以发问了：在了解这些相关资料的过程中，被试者何时能够猜到正确答案？很明显，只看第三条信息就可以了，因为它包含的信息最重要。但更普遍的情况是，人们得到正确的答案（泡沫塑料）是因为他们已经掌握了第三条事实。（上述四个事实都是从维基百科中“泡沫塑料”的页面提取的。）这种背景知识至关重要。没有它，被试者的行为反应会完全不同。

所以，这带我们回到了上一章中关于电影《2001：太空漫游》的讨论。我们从这里得到的经验教训也是一样的：为了了解人类如何做出特定形式的智慧行为，在这种情况下，可以使用威诺格拉德模式测试的按钮行为进行测试。而我们关注的重点也应当放在他们完成任务时所使用的背景知识上。

GOFAI的回归

如第1章所述，目前的人工智能的研究重点已不在知识方面，大部分已经脱离了约翰·麦卡锡的早期愿景。有些人认为，麦卡锡倡导的GOFAI是有进步意义的。但我们现在需要的是一种全新的方法，一种更加重视神经科学、统计学、经济学和发展心理学等方面见解的新方法。

尽管这些学科能够提供见解，尽管一个全新的计算机程序会更具生产力，但麦卡锡的设想更为激进一点儿。他提出了一个全新的主题，既不属于神经科学，也不属于统计学。这个新主题将从计算机程序的角度研究知识本身的应用情况。

非常重要的一点是，批判GOFAI的人们并没有试图去解释智能行为，而是采用了其他的批评方式。他们没能设计出另一个更好地回答威诺格拉德模式问题的程序（例如计算经济学），反而采用了远离GOFAI的方法。

这更像是主题的一种转变。研究人员不再关注人们回答问题的能力，而是将重点转移到其他形式的行为，尤其是对背景知识依赖程度较低的行为。（他们也许会进一步争论该采用何种进化理论来解释该问题。而我们也许应该在彻底了解行为本身之后，再去研究那些取决于知识的行为。）

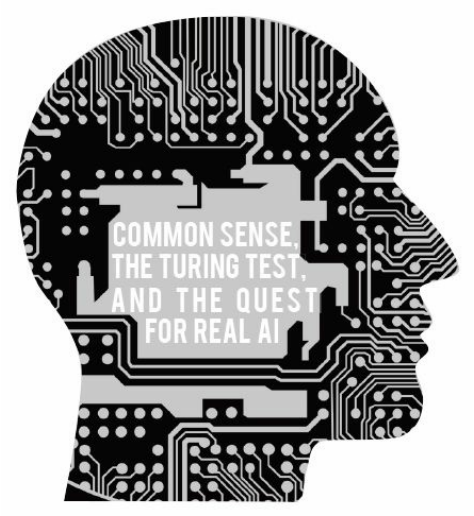
因此，专注于AML方面的研究人员可能会侧重于手写识别能力。例如，我们区别数字“3”和“8”的能力似乎较少依赖于背景知识，而是更多地依赖于我们在样本数字中看到的格式差别。AML的研究人员专注于展示如何从这些样品中自动获取必要的模式和特征。

当然，这样做没有任何问题。如果假设所有处于人类水平的智能行为都是依赖于背景知识产生的结果，这又会是一个巨型拼图式的错误。从威诺格拉德模式问题以及其他任何地方，我们都能够清楚地看到：只有一部分智能行为是依赖于背景知识的。

于是，问题依然存在：人们如何产生这种行为呢？即便不是像

AML那种类型的结果，那它会不会以更加复杂的形式呈现出来呢？为了回答上述问题，我们需要进一步了解学习和获取知识背后的内容。

第5章
有经验的学习与没有经验的学习



与30年前我刚刚投身人工智能领域时不同，现在人们所谈的人工智能，基本都围绕着机器学习及其应用领域，即我们一直所说的AML。研究人员在AML的推广应用方面取得了巨大成功，他们研发出了一系列应用程序，从手写数字识别，到人脸识别和跟踪，到在崎岖地形上行走，到在《危险边缘》[1]里参与问答，再到自动驾驶汽车。所有这些成绩的取得都严重依赖于与AML相关的统计技术。

如果与AML领域的研究人员探讨他们的研究内容，我们会得出这样的结论：学习就是从数据（或周围环境）当中总结规律和特点，并将其提取，而不是通过导师授课或者阅读教材学习。显然，我们的很多知识都是以这种方式学来的，例如如何辨别橘子和苹果、叔叔和阿姨、猫和狗等。但好像也没有人告诉我们，说“注意啦，看到猫有长长的胡须了吗？狗是没有这种胡须的”，而是我们自己像AML一样搞清了这些区别。这正是我们学会母语，学会用词语和句子表达思想的过程。的确，因为我们完全靠自己就掌握了大量知识，而非通过任何口头或书面学习，所以才会忽略那些依靠其他途径学到的知识。

换句话说，这是个巨型拼图问题。

我们如何学习词语？

让我们从思考自己如何学会词语开始。

我们先拿形容词“hungry（饥饿的）”为例。假设我们的父母会讲英语，我们最终学会了这个词的意思。但我们是在什么时候学会的呢？毕竟，可从没有人向我们解释过这个词的意思啊。事实上，是孩子在感到饥饿的时候，常常会听到这个词，因此逐渐将这个词与对食物的需求联系起来。

从某种意义来说，我们无须对“hungry（饥饿的）”这个词下定义是件好事，因为语言此时也有些力不从心。饥饿是什么感觉？我们可能会说：感觉不舒服，但也不痛苦，至少刚开始还不痛苦。这就是个温和的提示，告诉你该吃东西了。但这种感觉与其他如口渴或眩晕等感觉有何异同呢？

有趣的是，仅仅是我们的身体做出了某种提醒，并不意味着一定有形容词对其进行描述。例如，在法语中，没有表示“hungry（饥饿的）”的词，意思与之最为接近的是形容词“affamé”，但它的意思更像是“饿极了”，形容比饥渴更为强烈的感觉。在法语中，我们说“J 'ai faim”，字面意思是“我有饥饿”或者“我有饥饿的感觉”。由此可见，对于哪些身体提示会有对应的文字表达，我们目前尚不清楚。例如，有一种感觉与饥饿一样人人都曾感受到，尤其让幼儿的父母们每天烦恼不已，但令人惊讶的是，英语中竟然没有与之对应的形容词，这种感觉就是人想要小便时的感觉。

不管怎样，我们不是通过反复解释，而是长期沉浸在适当的语言环境中才习得了“饥饿”及其同类词语的含义。

那么，这就是学习词语的过程吗？让我们将巨型拼图问题铭记在心，然后继续探索。

现在考虑一下“incarnate（化身）”这个形容词。我们通常是通过另外一种方式学习这个词的。我们一般在年龄较大时才第一次听到（或看

到）这个词，甚至还可能记得见到这个词但不知道是什么意思的感觉，也记得后来知道了它的意思的感觉。我们需要通过别人解释才能明白这个词的意思：我们得知，该词用于指代由于某种原因附于肉体之上的精神或无形物体。可能我们还需要在字典里查一下，该词的两部分“in”+“carno”字面意思就是“into（进入）”+“flesh（血肉）”。所以，所谓化身，就是转化为血肉之身。（在某一年龄，我们可能还要查查“flesh（血肉）”的意思。）

那么我们没有料到的是什么呢？我们没有料到，自己是通过反复在正确的语境中听到该词来学习它的意思的。即使是较常使用该词的教会人士，也可能像以前在天主教仪式当中使用拉丁语一样，只是在祈祷或仪式中念着这个词，但根本不知道它的意思。

如果我们真正学会了这个词，那就要归功于语言，因为我们听到或读到了关于词义的解释。我们无须反复练习，只需在词典里查一查，就可以正确地使用它。因此可见，习得词语的方法的确与众不同。

顺带说一句，我并不是说我们习得的所有词语都能像“hungry（饥饿的）”或“incarnate（化身）”这样进行严格的划分。想想“exuberant（兴高采烈的）”这个词，许多人都是通过语言学习的第二种方式，即查字典的方式掌握其用法和含义的，但也有人通过反复接触来学习：他们会在语境中反复看到或听到这个词，直到搞清它的意思。还有人是结合了上述两种方式才掌握一个词的意思。例如，我对“hegemony（霸权）”这个词可能仅有个大概的认识，需要查过字典后才能自信地使用该词。

表面上看，通过经验这种较为直接的方式（如饥饿等词）与通过语言学习词语（如化身等词）相比，显得更为基础，也更有意义。要通过语言习得词语，我们必须掌握足够的语言能力才能理解对于词语的解释。那么更为基础的问题可能便是：我们是如何掌握这种初始语言的？

毫无疑问，我们如何通过在语境中听到词语来学习是个令人兴奋的话题。这个过程始于我们能够说话之前，一直延续到我们成年之后。每个家庭都能讲出萌宝们在牙牙学语时用词张冠李戴的故事，温馨可爱。（我还记得女儿小时曾撅起小嘴与大人争辩，说不让她做某事“太公平了”。）语言学家诺姆·乔姆斯基（Noam Chomsky）指出，由于孩子们接触到的数据总量较小，他们究竟是如何学会一门语言的，至今仍然是个巨大的谜。

另一个同样让人兴奋的话题是，我们如何通过使用语言来学习更多语言，即人们所说的引导语言问题：我们在掌握语言后，如何能够依靠语言学习数学、科学和技术背后的专业术语。（我们将在下一章做详细介绍。）

我们如何学习事实？

现在谈一谈我们是如何掌握生活中的常识的。

请思考这样一个简单的事实：柠檬是黄色的。很多水果有自己独特的颜色（包括很多有趣的颜色变体）：显而易见，蓝莓是蓝色的；黑莓是黑色的；橙子是橙色的；覆盆子是红色的；豌豆是绿色的；李子是紫色的；苹果是红色、绿色或金色的；柠檬就像香蕉一样，是黄色的。可能在生命之初柠檬是绿色的，在生命后期时是深褐色的，但是新鲜采摘的健康、成熟的柠檬通常是黄色的。

现在问问你自己：我是怎么学到这些知识的？好像不是通过语言学来的。当然，这些信息在书本中就有，但是很有可能你在读到关于柠檬的文字之前，就已经见到过柠檬（或柠檬的彩色图片）了。而且，在有关柠檬或其颜色的语句当中，讲述者或作者可能都认为你已知晓柠檬的颜色，并用它来形容其他东西。你可能会读到这样的句子：“她穿着一条柠檬黄色的连衣裙。”

这种关于柠檬颜色的知识就是我们通过直接经验，而非语言学习的例证。

现在思考这样一个事实：熊会冬眠。鹅、鲸鱼和黑脉金斑蝶等会迁徙到南方过冬；狼、鹿和大多数鱼类等不管气候多么恶劣，都会留守家园；冠蓝鸦等只在有些时候迁徙南方；而熊大多会在深秋季节找到一处洞穴，开始冬眠（或进入某种蛰伏状态），直到春暖花开。

再问一下自己：我是如何学到这些知识的？有一点是肯定的：我们不是通过对多只熊亲自观察，并发现其中的周期性规律的。如果你不是专门研究熊类动物的专家，那么你一生当中实际看到的活熊数量可能非常少，而且几乎没有一个是在冬眠的。你可能看过许多关于熊的照片和电影，虽然冬眠的熊和睡着的熊从照片上看毫无二致，但你不会就此相信熊会冬眠。你可能看过一些自然纪录片，纪录片介绍进入冬季后，熊会躲在洞穴里睡觉，到了来年早春再摇摇晃晃地从洞穴里爬出来。但在电视里，你只能看到熊在秋天钻进洞穴，然后在春天从洞里爬出的景

象，纪录片并没有太多交代熊在这期间的情况。熊在洞穴里一睡几个月，不吃不喝的信息需依靠旁白解说来传达。如果你没看自然纪录片，就只能通过别人的讲述掌握这个知识，有人告诉过你或你曾经在哪里读到过它。

这种动物世界的知识就是我们完全通过语言进行学习的范例。

我们在这里得到了两点常识：第一，柠檬是黄色的；第二，熊是会冬眠的。但我们学习这两点常识的方法却截然不同。

我们如何学习行为？

最后，来看一下我们是如何学习新的行为的。

首先，想想你是怎么学会骑自行车的。你肯定不会去找本自行车学习手册，仔细研究，然后就开始第一次的骑行。恰恰相反，可能是有人教你，而且你在年龄很小的时候就开始在幼儿自行车上训练，有老师帮你扶正自行车。第一次自己骑自行车的时候，你肯定摇摇晃晃，甚至还要摔倒几次。如果说老师对你有过指导的话，也仅仅是告诉你不要骑得太快（否则撞车会很痛），也不要太慢（否则自行车会不稳）。从来不会有人告诉你应该如何调整自己的重心，或者如何转动车把，防止自行车倒向某个方向。也从来不会有人告诉你在转弯时身体是该探出还是缩回，你依靠反复实践，不断纠正错误，获得提高，就像鸟儿学习飞翔一样，你掌握这一技能之后，它就成为你的第二天性。一段时间之后，你逐渐将此事淡忘。但是当你再次骑上自行车时，你的身体会不由自主地就知道该怎么做，而你则可以自由自在地思考别的事情。

这就是学习行为的意义吗？让我们继续深入探讨。

现在，想想你是怎么学会饲养宠物金丝雀的。可能你恰巧生活在金丝雀养殖场，跟随金丝雀养殖专家学习饲养技术。或者你父母养了一只金丝雀，你通过在日常生活中观察他们的行为，学会了如何饲养金丝雀。又或者，是别人给了你一只宠物金丝雀，或者你从宠物店买了一只金丝雀，但你却没有多少饲养金丝雀的实际经验。但无论如何，以下的情况是不可能出现的：你反复试验、不断摸索如何饲养金丝雀，比如尝试给金丝雀喂各种食物，只要每次不吃死就算成功等。你绝对不可能用这种方式饲养金丝雀，否则你就准备给它收尸吧。

但这也不是什么尖端科技。掌握相关饲养知识就和学骑自行车一样，非常容易。与宠物商店老板进行简单的交谈、阅读饲养手册或者上网搜索都可以解决问题。需要注意的是，在这种情况下需要提供一些语言难以表达的基本事实，这一点是与尖端科技类似的。

这些基本事实如下：

如果想要金丝雀健康快乐成长，必须提供新鲜果蔬。

如果没有水，金丝雀无法活过24小时。

金丝雀对各类毒素都非常敏感。

如果饲养金丝雀的笼子直接暴露在阳光下，金丝雀就会体温过热。

毫无疑问，我们可以通过在大量金丝雀身上反复试验、不断摸索，或在庞大的金丝雀数据中进行搜索来获得上述信息。但是生命短暂、时间宝贵，我们更愿意从他人提炼成书的经验中获得饲养金丝雀的知识。

我们如何超越经验？

当总结我们是如何认识周围的世界时，我们发现，似乎至少有两种相互独立的机制在起作用。第一种是通过经验进行学习，我们需要反复收看或收听相关经验数据，可能还会经历相当多的试验和摸索。第二种是通过语言进行学习，这种情况可能只需收看或收听一次相关数据就能达到目的。正因如此，第二种形式似乎与自适应机器学习（AML）所使用的统计技术毫不相干。（虽然重复与记忆在通过语言学习的过程中仍然会起作用，但在这里，这些重复的统计属性没那么重要。）

虽然英语中有很多术语可以用来描述人类学习的不同特性，但没有哪个术语完全符合上述方式。我们谈论的学习是通过教育、教学、培训、训练进行的，而且每种方式的侧重点各不相同：教育指广义的开阔心智；教学强调某一特定科目的教育，如学校课程或技术手册；培训强调对于某一技能的教育和实践，例如如何控制皮划艇或怎么做解剖；训练则侧重于培训的反复性，如算术练习或举重。上述每种方式都涉及语言数据和经验数据的相互平衡，而且这种平衡会根据学习的不同类型发生变化：语言的重要性也按照教育、教学、培训、训练的顺序逐渐下降。（我们在第7章中使用术语“培训”来指代从不断重复的经历当中学习。）

我们人类是通过两种不同的方式进行学习的，这一事实也带来了一些非常有趣的问题。比如当两者之间发生冲突时，我们该怎么办？根据经验，我们知道太阳是从东方升起的，但通过语言，我们学到太阳其实根本不会升起。通过语言，我们认识到了一些事物的存在，而如果根据直接经验，这些事物根本不存在。（别人可能告诉我们，我们不能体验这些事物是因为它们速度太快或太慢，或者是因为体积太大或太小，或是处于不同的平面上）。事实上，我们也清楚自己通过语言学到了很多虚构的事物，比如圣诞老人饲养的驯鹿的名字。

这其中一个非常有趣的问题是：为什么我们进化得能够通过语言进行学习，而其他动物却不能？为什么只有直接经验是不够的？在学习的时候，语言给我们带来了怎样的额外优势？

我们将在下一章更为全面地探讨这一话题。但是，塞缪尔·早川（S. I. Hayakawa）[2]曾说：“我们并非只有一次生命；如果我们能够阅读，就能够随心所欲地拥有多次生命和多种活法。”这话可以被看作是对这一观点的经典概括。在语言的帮助下，我们可以在自己的一生中了解前人的生活经历，我们无须经历前人的过程，就可以学到他们的经验。

由此可见，如果仅仅将人类语言视为一种沟通手段，那么绝对是严重低估了它给予我们的帮助。毫无疑问，语言能够沟通，但是沟通还可以通过其他途径得以实现，比如指一指相关的东西，或发出响亮的声音。几乎所有动物都会通过某种方式沟通，但是人类语言让我们得到了更多。一本宠物金丝雀饲养手册与莎士比亚的名剧《麦克白》（*Macbeth*）显然不可同日而语，但两者之间也有非常重要的共同点：有了它们，我们无须经历那些可能令人痛苦或不切实际，甚至难以忍受的生活经历，就能学到我们想要的东西。

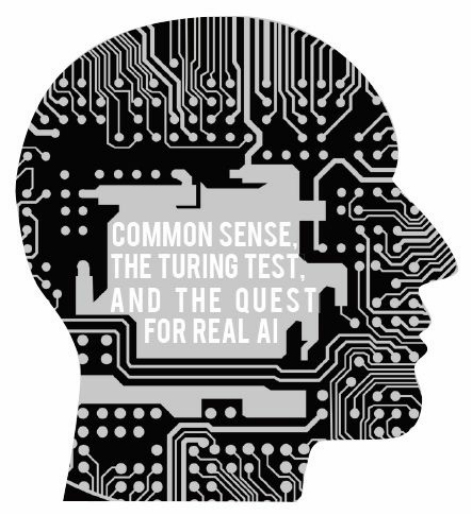
这一点有时会变得相当直接明了。食谱书《史上最好的提拉米苏》里有一句话言之凿凿：“我已尝试过各种不同的变化方式，你不用再浪费自己的时间了。”一本介绍在安第斯山脉采矿的书中写道：“这就是我在此山中的所见所闻，你可能就不用麻烦再来一趟了。”如果是本小说，可能会写有诸如“我认为这就是英国工人在工业革命期间的感受”这样的话。

艾萨克·牛顿在总结自己的科学成就时曾说：“如果说我比别人看得更远些，那是因为我站在了巨人的肩上。”这种隐喻当然不应简单理解为他爬得更高才看得更远，如果他爬到树上，可能最高。而应该理解为：他通过阅读巨人撰写的书籍文稿，获得了巨大的优势，能够从他们停下的地方开始，而无须重新经历他们所经历的过程。这也是我们学习语言文字、了解熊的冬眠习惯以及学会饲养金丝雀的方式。

[1] 《危险边缘》（*Jeopardy*）是美国哥伦比亚广播公司推出的一档益智问答游戏节目。——译者注

[2] 塞缪尔·早川（1906—1992）生于加拿大，日裔美籍语言学家，曾担任美国参议员。——译者注

第6章
书本智慧与市井智慧



众所周知，人类的语言共有两种不同的使用方法。我们用语言进行即时通话，比如，我们像动物之间的交流一样，发出声音，让他人立即听到，或者隔一段时间之后听到（比如查收语音留言）。同理，我们还使用纸质和电子媒介进行书面信息的交流互动。

但我们还会以一种更为独立超脱的方式使用语言，如播报体育比分、直播法庭审判、创作诗词、编写使用说明等。这时，我们对于听众、观众或读者是谁，他们会在何时收听或收看上述信息基本一无所知，有时候欣赏这些诗词的人甚至尚未出生。这种使用语言的形式为人类所独有，经过世世代代的积累沉淀后，就形成了我们所说的（口头的或书面的）文化。

语言的影响力

让我们问一下自己：为什么人类能对地球施加这么大的影响，能决定各类生物的生死存亡？你可能会说是因为“核能”“污染”“基因工程”，甚至是“塑料”［在电影《毕业生》（*The Graduate*）中有人告诉本杰明这是改变生命的关键］。答案当然不止这些，但如果我们退一步再看这个问题，就可以将之总结为“先进的技术”。

不过，其他动物也有技术。比如，乌鸦和黑猩猩都使用树枝木棒获取东西；水獭利用石块撬开蛤蜊。众多动物之中，为何只有人类独享这种先进的技术？很明显，是因为它们没有科学，所以没有这些先进的技术。而如果没有数学，我们就没有科学。但是如果人类没有语言，就没有科学，也没有数学。更确切地说，如果语言除即时通话以外没有其他用途，我们就不可能拥有这一切。

不妨这样理解：如果我们的语言能力和其他动物一样单一，我们就不可能通过世代累积，掌握科学技术和数学知识，研究开发出先进的技术。如果我们的能力仅限于表示自己的存在、标记自己的发现、指出自己的需求，那么人类对世界的影响就微不足道了。

举例来说，如果人类没有语言，现代化的都市也将不复存在。如果语言仅限于即时通话，那么诸如城市交通与通信、食物运输、水电供应、垃圾污水处理、火灾等灾害处置，这些需要大范围协同的工作都将难以实现。

（试想，如果人类生活在庞大的族群中，语言条件有限，那会是怎样的一种情况？比如，蚂蚁就悠然自得地生活在极其庞大的种群当中，虽然它们之间很少沟通，但它们在行为当中体现出了强大的遗传规划效果，即限制个体行为的多样化发展，要求个体服务于种群事务。人类也仅仅是在几千年前农业生产出现后，才开始生活在城市中，或许几百万年后，我们将继续进化，成为根据城市需求为城市服务的生物，到那时，我们都会自动自觉地工作，不再需要任何法律规章，更不需要家长、老师、各级官员敦促我们遵法守纪，也不需要警察去抓捕那些违法乱纪、恣意妄为的人。）

想象一下，如果语言只能作为即时交流工具，你就只能依靠传帮带的师徒制学到觅食、挤奶和耕地等知识。但是，如果语言超越了即时交流的功能，形成语言文字，你就能学习线性代数、电气工程、城市规划等知识。

我们都是在学校通过听老师讲课、阅读资料的方式学习科学、数学和工程学的。当然，要想熟练掌握这些学科的相关技能，实践和经验不可或缺。而如果没有背景知识的支撑，我们永远做不到熟能生巧。

由此可见，以阅读或听讲的形式通过文字进行学习，不仅是人类的独特之处，而且这种能力确保人类得以更好地主宰着自然世界，当然也许是更坏。

书本智慧

虽然书本学习在我们的文化生活中占据十分重要的地位，但是人们对书本学习的态度却非常不屑一顾，还经常将其嘲讽为只是学习“书本上的知识”，所以只有“书本智慧”。相比之下，通过直接经验或师徒传帮带获得的知识被定义为“市井知识”，拥有这种知识的人群被称为拥有“市井智慧”。斯科特·博克顿（Scott Berkun）说过，“市井智慧完爆书本智慧。（Street smarts kicks book smarts ass.）”（我对这句话的语法和标点不敢苟同，但是这也只是关于书本知识的讨论！）

他在一篇博客文章中这样说道：

> 在我看来，拥有书本智慧的人是指那些擅长循规蹈矩的人。他们是一群永远成绩优异、听课爱坐前排、善于各类考试的人。这群人喜欢只有唯一正确答案的问题。他们坚信，他们广博的知识可以弥补自己在现实世界中缺乏的经验。虽然思考本身具有价值，但在想象中处理棘手的问题，与在现实中真正处理问题却有天壤之别，两者截然不同。［正如泰勒·杜尔登（Tyler Durden）在电影《搏击俱乐部》中所说：“如果从未真正战斗过，你怎会认识真实的自我？”（摘自http: //scottberkun.com/2010/book-smarts-vs-street-smarts）］

虽然我们并不完全赞同斯科特·博克顿的观点，但大多数人对于上述说法的基本要点还是认可的。其实，虽然前文谈到了科学和城市，但我们可能也很难找到合适的人能反驳这些观点。对于城市，我们甚至可能会这么想：

> 如果没有足够的书本知识就无法在大城市立足，那么就在大城市里凑合着生活吧！

实际上，就在不久以前，人类都还生活在相对较小的社区当中，并

不是很依赖以书本知识为支撑的各类城市技术。一小群人无须太多书本知识，仅靠师徒传帮带的方式就可以过得不错。（即使生活在大都市，对于个人而言，只要朋友当中有人掌握了生活所需的技术，自己就无须掌握这些技术。）

但是反过来又会怎样呢？人们能否仅凭书本知识，基本不靠直接经验而生存呢？这似乎不太可能。然而，还是有一些反例，虽然较为罕见，但值得我们注意。

海伦·凯勒的智慧

让我们来看一下海伦·凯勒（1880—1968）的著名案例。在海伦·凯勒18个月大时，一场疾病夺走了她的视力和听力。当时，她尚在（通过直接经验）牙牙学语，变故之后便无法像正常人一样学习英语了。

然而，在海伦6岁的时候，她的家人为她聘请了一位启蒙老师——安妮·莎莉文（Anne Sullivan）。安妮通过在海伦手上画单词的方式（每次教一个单词）教她认识周围的事物。海伦以这种方式学会了大约25个名词和4个动词。

之后的某一天，就像电影《奇迹的缔造者》（*The Miracle Worker*）里讲的那样，奇迹发生了。安妮·莎莉文这样说：

> 今天早上，当她（海伦）洗手时，她想知道“水”该怎么说……我在她的手掌中写出“水”……我们去了水泵房。我让海伦拿着杯子在喷水口下接水，而我来抽水。随着凉水涌出填满了杯子，我在海伦的另外一只手中写出“水”。突然间，水流过手掌的奇妙感觉让她顿时明白。这就是水！她扔掉手中的杯子，呆呆地站在原地。她的脸上忽地闪耀出顿悟的光辉。她不停地在我的手掌中写出“水”这个单词。接着，海伦从激动的心情中回过神来，指着水泵和棚架，问我它们的名字。突然，她转过身询问我的名字，我在她的手掌中写出：“老师”。在那时，护士把海伦的妹妹抱到了水泵房里，海伦指着护士的方向，在手中写出“宝贝”。回到屋里后，海伦还处在高度的兴奋当中，她如饥似渴地学习着每一个她碰过的物体的名称。在那短短的几个小时里，海伦已经学习了30个新词……海伦今天早上起床后像一个快乐的小仙女。她的手掠过一件又一件的物品，不停地学习一个又一个新词，还高兴地亲吻了我。昨晚睡觉的时候，海伦偷偷钻进我的怀抱，并且第一次亲吻我的面颊。那时我的心怦怦直跳，充满喜悦。（摘自一封写于1887年4月5日的信）

我们很难描述这一时刻在凯勒的生命中的意义，但可以肯定的是，

这是她第一次真正意识到一切事物都是有名称的。尽管她以前也一直在给周围的事物命名，但是现在，她认识到除了自己周围能摸得着、有质量、能动弹的实物以外，还有一些抽象的存在——词语，而且词语可以用来给其他事物命名。

这和婴儿学语的情况并无不同。婴儿早期就可以学习“水”“饿”“走”之类的词语。小孩子还能够运用已知的词语来提问题，比如母亲在吃什么？父亲在做什么？她最爱的娃娃在哪里？问题的答案就是周围的实物、质量和动作。但是，当孩子成长到某个阶段时，新的情况出现了。孩子在对世界上的实物存在依然兴趣不减的同时，开始讨论新的话题——词语本身。小孩会用已知词语来问有关其他词语的问题：某个东西叫什么名字？某个词是什么意思？

（可以认为，这是让人不只依靠直接经验，而且能够通过已知语言学习其他语言的关键步骤。其意义并非在于使用符号——在这个例子里使用的是词语——来讨论世间万物，而是在于符号本身成为讨论的内容。我们之所以在此强调其重要意义，是因为它将符号与我们看得到或想得到的东西分离开来。符号本身具有了生命力，不再仅用于表示世间万物。正如哲学术语里所说：符号不仅可以被使用，还可以被讨论。我们在第8章将会看到，这就是符号处理的意义。）

经历了这个奇迹之后，海伦·凯勒的进步令人惊叹：学盲文、学英语、学其他语言、写书，并且最终获得大学学士学位。（她是第一位取得上述成就的聋哑人。）鉴于海伦的身体情况，她获取知识的主要途径是通过阅读和与人交谈，而这就是我们所说的书本知识。就书本智慧而言，她取得的成绩非常了不起。

因此，海伦·凯勒能否作为一个反例，用来反驳掌握书本智慧，基本没有市井智慧永远不够这种说法呢？约翰·麦克罗恩（John McRone）如是说：

> 但现实情况是，因为海伦与世界隔绝太久，所以她很难分辨出自己的记忆与想象之间的差别。她还学会了文字游戏，但在说出这些令听众愉悦的语句背后，海伦自己能理解多少还是一个问题……由于文化根基不牢，海伦最终也会感到吃力。视力和听力的丧失，堵塞了大脑获取外界感觉、图像与记忆的通道。然而，通过语言，她又用丰富的人类文化知识填满了空缺的大脑。这种组合最终可能

会缺乏平衡；大多数人看到的是她通过战胜不幸而取得胜利，只有少数人能体会到认知不足会给思想带来的巨大压力。（摘自dichotomistic.com的博客文章）

我并不认同这种观点。（而且我觉得这种观点有种居高临下的感觉。）但是，我认为上文所表达的观点与图灵的观点有相似之处。即使是像海伦·凯勒这样能够写书，并获得大学学位的人，也依然有人认为她是在玩弄文字游戏，并没有真正懂得单词的含义。这种关于海伦“玩弄文字游戏”的评价，与对ELIZA等其他人工智能软件的评价如出一辙［威廉·拉帕波特（William Rapaport）曾先于他人对此做过记录］。但是，这些评论者是否认为海伦·凯勒无法通过威诺格拉德模式测试（Winograd Schemas Test）呢？或者即使她能通过测试，他们也会宣称她只是投机取巧，并非真正理解了所有的问题？（如果答案是肯定的，那么她是怎么投机取巧的呢？）

海伦·凯勒在自身存在巨大生理缺陷的情况下，仍然取得这样的成就，让我们刮目相看。但也有很多人提出疑问：如果没有足够的市井智慧，人类的心智究竟能走多远？

更为有趣的是，在所有的讨论当中，很少有人会质疑这样一个问题：如果没有足够的书本智慧，我们能走多远？无论这背后是什么原因，我们确实过于轻视人类特有的精神生活了。我们在审视自身的优势（比如语言、规划、制造工具和文化）时，总是希望能在其他动物身上也找到类似的能力。近年来，对于人类与动物之间的差异，我们有些熟视无睹。其实二者之间最大的差异，就是人类能够学习书本知识，人类语言不仅用于即时通信，而且用于进一步扩展我们的语言、规划、工具制造与文化。

书本中的市井智慧

让我们回到书本智慧和市井智慧，重新思考一下上文提到的熊冬眠和黄色柠檬的例子。我们发现，我们是通过语言知道了熊冬眠，通过经验认识了柠檬。换用本章的术语就是，我们通过书本知识认识了熊，通过市井知识认识了柠檬。

这样一来我们就能明白，诋毁任何一种知识形式都是无意义的。难道只有爱坐前排、善于考试的书呆子才知道熊冬眠这种知识吗？通过市井直接经验知道“柠檬是黄色的”就真的那么了不起吗？

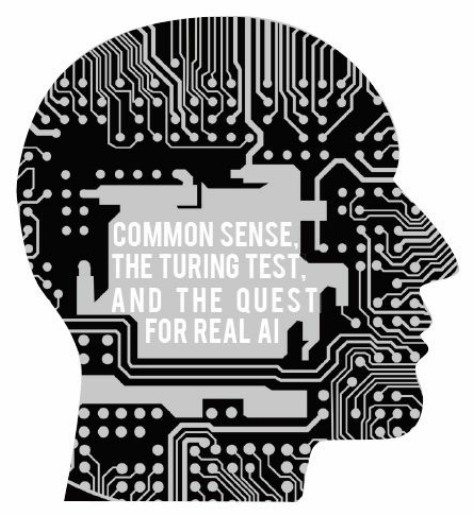
至于“市井智慧完爆书本智慧”的原因，更为合理的解释是二者的题材内容不同，而非获取方式不同。换句话说，在其他条件完全相同的情况下，实践知识拿来就能用，直接作用于下一步决策，比起阳春白雪的抽象理论，更容易引起我们重视。

假设你正乘坐出租车在罗马旅行，如果想给同行游客留下好印象，那么就要高度重视是否要给司机小费这类信息，因为这里的习俗可能不同于爱丁堡。你可以通过亲身实践试错，观察当地人的做法，向别人请教该怎么做，或者阅读旅游指南上的相关介绍，来了解小费风俗。总之，方式并不重要，重要的是你了解了当地风俗，并能在实践当中不犯错误。尽管这些知识来自书本，但却让你拥有了市井智慧。（这里需要说明，在罗马乘客一般不需要向出租车司机付小费，但是在爱丁堡要付。）

当然，我们知道有一些东西确实无法通过语言学习获得。如果我们从未见过柠檬，那么语言对于描述柠檬的颜色变化也不会有太大帮助，除非参照其他颜色相近的物体。海伦·凯勒可能知道柠檬是黄色的，而且这种颜色与稻草和校车的颜色相似，但是她可能分不清这些颜色之间的差别。语言在此显得苍白无力。（有趣的是，书籍可以向我们展示色彩的变化——而且只要一个例子就够了——但却无法用语言将其表达出来。）我们目前尚不清楚这点儿额外的市井知识的重要性，但我们确实拥有它。

同时我们也知道，有很多知识都不是通过直接经验学到的。只有极少数人需要直接与冬眠的熊打交道，其他人都是通过这些人了解熊冬眠的知识。我们学习科学、数学和技术的过程也是如此。关于这些关键论点，我们与海伦·凯勒的立场一致。

第7章
长尾理论与培训的界限



在努力理解智能行为的过程中，我们经常会采取以下方式：首先观察自己认为最常见的行为，找出相应的应对办法；然后以此为基础，便可以不断提高处理问题的能力。当我们跨过某个门槛，基本能够处理99.9%的问题时，我们便停下来。

这就是所谓的工程战略。我们对于正在研究的某种行为，先制订草案，然后进行工程设计，优化并提高处理能力。这种战略在机械设计当中表现得淋漓尽致。给定一个推力为X的火箭，如何改善才会产生推力Y？对于一座承载能力为X的桥梁，如何加固可将承载能力提升至Y？

这种工程战略确实在解决一些与智能行为相关的问题中发挥了良好的作用。例如，当我们学习走路时，我们都会从普通、常见的地方开始，比如在地板上或者坚实的地面上。最终我们学会了走路，并且能够应对沙地和冰面等较为复杂的路况。同样的道理，我们在学习母语时，通常都是从聆听儿语开始的，而不是去看最新一期《麦克劳克林小组》（*The McLaughlin Group*）或者达纳·卡维模仿秀。

但是，这种工程战略也有完全不起作用的时候。如果某些个体事件虽然极为罕见，但对于整体却意义非凡，我们就将这种事件的分布称为长尾分布，而这些罕见的个体事件就是纳西姆·塔勒布（Nassim Taleb）所说的“黑天鹅”。（欧洲人曾一度认为所有的天鹅都是白色的。）虽然黑天鹅的作用可能非常重要，但如果我们先着眼于常见的现象，再将注意力转到稍微不太常见的现象，然后以此类推，我们就可能永远也见不到黑天鹅了。单从统计学的角度来看，我们似乎做得很好。但是事实上，我们可能做得非常糟糕。

长尾现象带来的难题

为了更好地理解该如何应对长尾现象，我们不妨假想一个极端的数字示例来帮助思考。

假设我们正在尝试估算一个巨大数字集合的平均值。为了便于思考，我先透露点儿小秘密：这个数字集合中包含1万亿个数字，它们的平均数是100　000。但是集合中大部分数字都非常小。该集合数字的平均数如此之大的原因是，其中有1　000个超级大的数字。当然，这些大数字都控制在100万亿之内。

现在，假设我们对以上数据毫不知情，并且我们的目的是通过抽样得到一个典型数字。我们从集合中第一次抽取的10个数字如下：

2，1，1，54，2，1，3，1，934，1。

其中出现频率最高的数字是1，中位数（按从小到大顺序排列后位于中间位置的数）是1.5。但是这组数字的平均数肯定大于1。在寻找平均数的过程中，我们会找到一些大于平均数的数字，也会找到一些小于平均数的数字，这样总体数据的差值就会缩小，直至达到一种平衡。针对现有数据，我们可以通过所有数字求和除以数字个数来算出一个“样本平均值”。前5个数字的样本平均值是12。也许我们认为这种模式会延续下去，但10个数字的样本平均值却是100，远远高于12。到目前为止，在前10个数字中，只有一个数字大于100。好在这个数字足够大，使得它与其他9个数字在计算平均值之后等于100。

我们继续抽样，验证一下这个猜测是否正确。一段时间过后，样本平均值果然稳定在100左右。但是假设在统计1　000个样本数字之后，我们得到了一个更大的平均数——1 000 000。一旦我们把这个大数字加入平均值的计算当中，那么得到的结果将为1 000。

假设这种情况一直持续下去：我们看到的数字大都像前10个一样，

但是每1 000个左右的样本就会出现一个大数字（100万以内）。当我们停止计算，并决定公布最终的样本平均值是1 000时，突然出现了一个更大的数字——100亿。这种情况极为罕见。但是由于数字过大，样本平均值由1 000变成了100亿。由于这种情况是意料之外的，我们决定继续抽样，直到看完10亿个样本，样本平均值稳定为10 000。

这就是从长尾分布中抽样的示例。在很长一段时间里，我们可能认为自己已经足够理解该数字集合的统计特性。我们可能会这样说：

> 虽然我们无法计算整个集合的属性，但我们可以通过抽样来预估这些属性。经过大量的抽样调查，我们能够得出样本平均值为10 000。但实际上大部分数字要小得多，远低于100。当然，还有一些非常罕见的极端大数字，但是它们都在不超过100亿的范围内。这种数字是极为罕见的，可以说是百万分之一的小概率事件。通过10亿个数字的样本测试，这种情况目前已被证实，所以我们可以自信地谈论所期待的结果了。

但是，这个结论是非常错误的。长尾现象的难点在于：我们研究该现象的时间越长，就越难以理解之前期望的结果。我们采样越多，样本平均值就越大。为什么我们认为10亿个样本就足够了呢？

为了更加生动地看待这些数字，我们可以利用一些看起来非常有效的新技术来帮助我们。目前，某问题已造成每年有36 000人死亡（这是2012年美国死于交通事故的人数），而该技术正应用于解决这个问题。新技术引入后，我们很难计算出死亡人数的准确数字，但我们可以进行一些模拟测试，亲自检视测试结果。假设上文中的每个采样数字对应新技术引入后每年的死亡人数，那么问题来了：我们是否应该引进这种新技术呢？

根据上述抽样调查，新技术引入后，每年的死亡人数可以控制在100人以下，这比目前所知的36 000人要好得多。事实上，模拟测试显示，在99.9%的时间里，死亡人数都是低于10 000人的。这看起来好像不错。不幸的是，模拟测试还显示，死亡人数有千分之一的概率将会达到100万。这将是一个灾难。如果说100万还不是最大的数字，那么模拟测试显示，人类将会有百万分之一的概率全部死亡。

这样看来可不怎么好！

有人会说：

我们必须现实地看待这个问题，而不应该把过多的时间花费在这些罕见的小概率事件上。毕竟，彗星也可能会撞地球！忘掉那些黑天鹅吧，它们不会打扰我们的正常生活。问问你自己：我们真正期待的是什么？如果不使用这项技术，目前的死亡人数是3.6万。如果我们期望状况好转，那么是否应该使用它呢？

这种立场并非不合理。如果只考虑最坏的结果，那我们干脆就别活了。长尾现象带给我们的问题，是我们如何能够找出该现象中的典型情况。但是，这个典型的数字是多大呢？我们看到的数字中有一半都小于10。但是我们也清楚，这是在误导我们的判断，毕竟另外一半数字都特别大。99%的数字小于1 000，剩下1%的数字比1 000大得多。在这样的情况下，我们最多能忽略多少样本呢？样本平均值最能够体现整个抽样过程的好坏。不仅所有的数字都小于10 000，凑巧的是，所有大于10 000的数字之和正好与小于10 000的数字之和相等。然而，作为典型数字的10 000，仍然比36 000要小得多。

为了更好地论证这个问题，我们不妨假设一下：经过10亿次计算后，我们看到下一个数字就是一开始提到的大数字之一：100万亿。（究竟何种技术能够造成100万亿人的死亡尚不清楚，暂不深究。）即使这是一件发生概率仅为十亿分之一的事件，但由于数据如此之大，我们也必须重新计算数据的平均值。这次得出的结果是100 000。这个结果与36 000比起来可糟糕了不少。

简而言之，这个问题旨在用实际情况解释长尾现象。如果你所有的专业知识都来自于抽样数据，那你可能永远都无法察觉这些牵一发而动全身的小概率事件。

如何处理突发事件

假设你在美国加利福尼亚州已经拿到驾驶执照，并且驾驶经验丰富。你在开车过程中遇到过各种各样的现象：恶劣的天气、拥堵的学区和停车场、交通堵塞、水平参差不齐的司机、湿滑的道路、笨重的车辆、轻微事故，你甚至可能遭遇过危险。

冬日的一天，你在加拿大安大略省租了一辆汽车，一路向北行驶。你已经提前知晓，这条路会有点儿滑。于是，你开着这辆四轮驱动的汽车低速行驶。开始下雪了，但这种情况已经有人提醒过你。如果雪下得更大，你会靠边把车停下，然后找一个地方休息。但是，还没等你做好准备，你就出现了雪盲（风和粉末状的雪包裹在汽车四周，白茫茫的一片，人分不清方向，看不到天际，也看不到其他车辆）。你从未经历过类似情况：四周悄无声息，感觉不到任何动静，你身处雪白的环境中，就像飘浮在云端。你该怎么办呢？拉起手刹，把车停到自认为是路边的地方吗？这时候你所做的任何决定，与你在加州丰富的驾驶经验都没有关系。你需要考虑的是：你以及其他车辆都在做什么。

这个例子虽然较为罕见（当然，只是对于加州的司机而言），但是却很重要。如果做出了错误的决定，司机就很可能丧命。虽然我本人就住在安大略省，但我也只经历过一次雪盲而已。当时我确实非常担心自己会命丧于此。我当时的做法是这样的：不踩刹车（不想被后面的车撞上），一路直行不转弯（希望前面的路够长），打开应急双闪灯（希望后面的车能注意到我），并且仔细观察前方是否也有闪烁的应急双闪灯（希望前面的车也能照做）。我就这么一直缓慢前行，直到雪盲结束。我竟然幸运地活了下来！

有人可能会说，训练有素的安大略省司机应该都经历过雪盲，他们所缺乏的只是更多的训练而已。但问题就在于：训练到什么程度才算够呢？我们是否要担心开车过程中油门卡住的情况？晚上驾车时大灯不亮的情况？交通信号灯持续红灯，出现故障的情况？一头鹿被撞倒，死在车前的情况？乘客试图爬出窗外的情况？一群暴徒围住汽车的情况？汽车被飓风刮动的情况？很明显，上述情况都不可能，我们不会也不可能将所有的状况列成清单，都放到驾驶训练中去。

这里体现的问题在于：尽管上述每一种情况都非常罕见，但是驾驶汽车时总会有很多奇奇怪怪的事情发生，所以我们总会遇到其中的某种情况。

更为典型的例子可能就是关于文本中出现的词语。英国国家语料库（*British National Corpus*）是一个从各种来源获取英语文本信息的大型数据库。语料库中共有大约一亿个单词，其中大部分词语是人们经常使用的，但也有一些词语在整个语料库中只出现过一次。令人惊讶的是，这部分词语所占的比例是0.5%。语料库中出现次数最多即10次（这可是一亿分之十的概率）的词语，占总体的1.7%。这是一个典型的长尾分布的案例。在这种情况下，这些非常罕见的词语之所以极其重要，是因为它们的数量过于庞大。因此，我们在阅读文本时才能够有机会看到它们。通常情况下，我们并不指望一个智能代理能从那些极为罕见、一般不会出现的经验中学到什么。如果可以忽略这些罕见事件，系统仍然可以正常运行；如果无法忽略，且它们又是类似于长尾现象的情况，那么只依赖于过往经验的系统就会崩溃。

无意识与有意识

大多数驾驶员都知道，在路上遭遇突发事件是种什么状态。我们可能边开车边抿了口咖啡，听着收音机，并且不时和车里的朋友聊聊天。事故的发生可能是由于雪盲，也可能是误入歧途。这时，你的驾车状态立马就发生了变化：把咖啡放到一边，关上收音机，与朋友的闲聊也改变了话题，大家一起讨论如何应对目前的状况。我们的全部注意力都集中在驾驶行为上，而在之前这仅仅是一种背景行为。

简单来说：我们是将驾驶从无意识的行为转变成了有意识的行为。这是一种简化。实际上，即使是在无意识的“自动驾驶”行为当中，我们也花费了心思。对于自己正在进行的行为，我们可能还没有自觉地意识到，但是在汽车行驶过程中，我们清楚地观察着路况，并实时做出应对。如果有人问我们在高速路上都干了些什么，回答可能是“我和朋友在车里听收音机”，而不是“我正沿着路的方向顺时针打方向盘”。但毫无疑问，这两件事我们都在做。

通过强化培训，我们的专业技能得到了提升，这样在很多情况下我们就能够做到一心二用甚至多用，比如边开车边听收音机。同样，当我们步行、开车、骑自行车（参见第5章）、弹钢琴或煎鸡蛋时，也可以一心二用。事实上，国际象棋专家拥有这样一种技能：在下棋的同时能够专注于其他事情。（实验表明，国际象棋专家在计算一长串数字之和的同时，仍然能够把棋下好。）哲学家休伯特·德赖福斯（Hubert Dreyfus）认为，集中所有精力做一件事的人只会是新手，而非专家。

不过，在出乎预料、异乎寻常或者人地生疏的环境中，我们都会变成新手。驾驶行为就变成了有意识的行为。而这也是对情况的简化处理。因为我们不会像进入冥想一样，全身心地放松下来。我们只是会更加留意开车这件事，但也不一定具体到手放在何处、脚踩在刹车板上的压力（除非我们突然感觉转向或制动有问题）这种事情。

那么，在这种不熟悉的状况下，人们会如何表现呢？最糟糕的情况是，我们会陷入恐慌，手足无措。如果表现得好，即使出现一些从未听过或者见过的问题，我们也能够做到不那么慌乱。我们还可以借助相关

的背景知识来应对。

通常，我们会学习如何在常规状况下（在街道和高速公路上）驾驶汽车。而且即使在这种情况下，教练也会叮嘱我们要保持警惕，随时对可能发生的意外情况做出反应。

那么在雪盲的例子中，我使用了哪些知识呢？关于交通事故，我了解到：在类似汽车从道路冲入沟渠的案件中，后果比较严重的是两车相撞，相撞程度取决于两车的相对速度。我没遭遇过汽车碰撞的事故，但我和大家一样，都从新闻报道中听说过一些。（这也算是书本知识。）在雪盲的情况下，我不得不担心前后方可能发生的汽车相撞情况。所以，马上停车并不是一个明智的选择。我还记起：重型货车在爬坡时会打开应急双闪灯，以此警示后方车辆保持车距，减速行驶。我决定效仿这种做法，并且希望前面的车辆也能够这么做。我从未听说过或者了解到可以如何应对雪盲。（安大略省的司机驾驶手册上确实提到过雪盲，但并没有提供应对之策。）这是我利用背景知识应对问题的新方式，与认为电影《2001：太空漫游》比1968年发行的电影好看得多没什么区别。

所以，我们可以总结一下：在一种极端情况下，无意识的活动是指身体参与其中、心思却在别处的活动，这种活动以经培训获得的技能为依托。而在另一种极端情况下，有意识的活动是指需要我们注意力集中的活动，这种活动重视个体对于活动本身的注意程度，而非大量的培训。

威诺格拉德模式测试

背景知识在有意识的活动中的使用，可以充分地证明：智慧行为不只是从经验中学习的结果。当然，经验是有用的，你接受的培训越多，经验就越丰富。从统计学的角度来看，一个完善的系统令人印象深刻。但是这些数据也可能忽略了一些罕见却重要的事件。

现在，我们再来考虑一下第4章提到的威诺格拉德模式测试。可能有人会问：回答这些问题有没有什么技巧呢？请看下方的例子：

1.这是一个威诺格拉德模式测试的例题：

奖杯无法装进棕色手提箱里，因为它太小了。这里的“它”指的是什么？

- 奖杯
- 棕色手提箱

解析如下：

二者之间的关系为R。

其中一个具有属性P。是哪一个？

对于上面的问题，我们可以得到：R=“无法装进”，P=“太小了”。

2. 使用大数据：在网上检索所有的文本信息，确定哪种模式更为常见：R（x，y）+P（x）还是R（x，y）+P（y）？

这两种模式分别对应以下两种情况，

x无法装进y里+ x太小了；

x无法装进y里+ y太小了。

然后根据更为常见的模式来回答该问题。

利用“大数据”的方法来回答威诺格拉德问题，不失为一种好方法，而且也确实可以解决许多问题。

然而，我们也能够举出反例，证明这样的方法有时行不通。

如下文所示：

奖杯无法装进棕色手提箱里，尽管它很小。这里的“它”是指什么？

- 奖杯
- 棕色手提箱

请注意，即使此次答案与上次不同，但这里的R和P却不变。

那么我们能从中得出什么结论呢？只是掌握更多样的技巧就够了吗？例如，利用“不”“除非”“尽管”等字眼。然后，问题就变成了这样：

尽管萨米用尽办法，但奖杯仍无法装进棕色手提箱里，因为它太小了。这里的“它”是指什么？

- 奖杯
- 棕色手提箱

是不是技巧还不够完美？

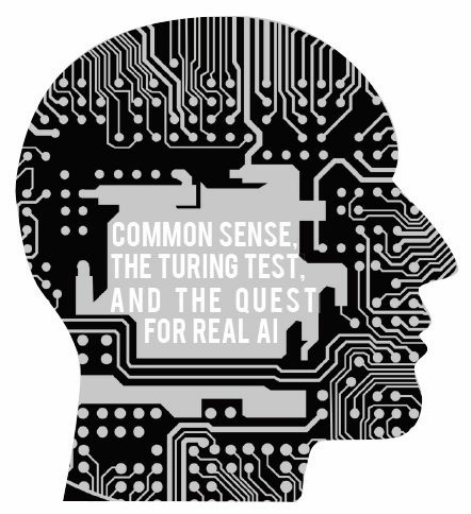
假设我们真正关心的是那种能够答对大部分威诺格拉德模式测试题的系统，那么从工程学的角度来看，最好的策略就是使用一些简易的小技巧。毕竟，把时间都浪费在那些低概率事件上是不值得的。例如，在威诺格拉德比赛，甚至是与人类的竞赛过程中，用多样的技巧获胜是一件概率很大的事情。从统计学的角度来说，那些处理不当的例子便都不怎么重要了。

不过，如果你想更好地了解人们正确回答威诺格拉德模式测试题的能力，那么人类靠自己而非靠技巧来解决问题，就显得尤其重要。我们想要了解的是人类能够做什么、如何做，而不是可能做一些类似的事情。即使是威诺格拉德模式测试里的一个例题，也能提供重要信息，告诉我们人类能够做哪些事情。我们不能简单地撇开题目，然后大声喊道：“这种情况几乎不可能发生！”

正是对人类如何处理罕见事件的关心，促使我们不断地超越自适应机器学习的研究。我们处理类似黑天鹅事件的能力，并不是通过额外培训后形成的刻板、单一的行为模式的结果。相反地，这更像是一个全新的、依赖于背景知识的工作机制。

但这仍然给我们带来了一个重要的问题：这一机制的本质是什么？

第8章
符号与符号处理



在讨论信念的作用的过程中，我们曾提到这样一类系统：知识在其中不单单是描述行为（意向立场）的快捷方式，更在其中扮演着某种因果角色，就像汽油对于奔驰汽车不可或缺。

为了了解知识如何发挥这样的作用，首先我们需要讨论一下符号和符号处理。我们可以从高中数学开始。（对于再也不想看到高中数学的读者，可以直接跳到本章的“符号的意义”一节。）

符号的代数运算

请思考下面这个简单的问题：

两年前，约翰尼7岁。

他现在几岁了？

这道题简直易如反掌，连纸笔都不需要就可以答出。很显然，约翰尼现在9岁。在高中阶段，我们曾学习过如何系统地解决这些问题。这样做的好处是使我们能够解决答案更加隐晦、过程更加复杂的问题，比如：

汤米比苏济大6岁。

两年前，汤米的年龄是苏济年龄的3倍。

那么，苏济现在的年龄是多少？

这就是代数。回到第一个问题，我们想求得约翰尼的年龄，解题的关键在于如何将涉及约翰尼年龄的文字表述转换为对应的方程式。为了简洁起见，我们将约翰尼的年龄设为x：

$x-2=7$

或，约翰尼的年龄减2等于7

当然，这里的未知数x指代谁都已不重要。我们的目的在于，剔除不必要的细节，使已知条件更易于计算和处理。用这种方式处理问题，我们希望能够得到以下某种类型的方程：

$x=V$

或，约翰尼的年龄是V

这里的V代表的是数字。依照我们在高中学习的知识，不需要用到太多的花样，求解过程如下：

1.在等式的左右两边同时加2：

$$x-2+2=7+2$$

2.简化左式：

$$x-2+2 \Rightarrow x+(-2+2) \Rightarrow x+0 \Rightarrow x$$

3.简化右式：

$$7+2 \Rightarrow 9$$

4.在简化后的左式和右式之间画等号：

$$x=9$$

这样便求出约翰尼的年龄是9岁。

对于第二个问题，求解的过程是类似的。我们将汤米的年龄设为x，将苏济的年龄设为y，最终目的是求解y的值。根据题目里的两个已知条件，我们能够得到以下两个方程：

$x=y+6$或汤米的年龄是苏济的年龄加6

$x-2=3(y-2)$或汤米的年龄减2是苏济的年龄减2的3倍

求解过程如下：

1.用第二个方程的左右两边分别减去第一个方程的左右两边：

$$x-2-x=3(y-2)-(y+6)$$

2.简化左式和右式：

$$x-2-x \Rightarrow x-x-2 \Rightarrow (x-x)-2 \Rightarrow 0-2 \Rightarrow -2$$

$$3(y-2)-(y+6) \Rightarrow 3y-6-y-6 \Rightarrow (3y-y)-6-6 \Rightarrow 2y-12$$

3. 在化简后的左式和右式之间画等号，并移项：

$$2y-12=-2$$

现在，我们已经将两个二元一次方程化简为一个一元一次方程，接下来的步骤与之前一样：

4.等号左右两边同时加12，然后除以2，经过简化，可以得到：

$$y=5$$

看！只用到简单的代数知识，我们就能根据已知条件确定苏济的年龄是5岁。再进一步计算，我们就能确定：汤米是11岁。当然，两年前，汤米的年龄（9岁）是苏济的年龄（3岁）的3倍。由此能够推出，根据3个已知条件，可求得3个未知数；根据67个已知条件，能够求得67个未知数。事实上，要解决此类含有未知数的方程式，仅需要用到几个关于等式、加法和乘法运算的法则即可。

总地来说，代数是符号处理的一种形式。我们从完整的等式入手，比如“$x-2=3(y-2)$；$x=y+6$”，根据运算法则进一步计算化简，最终求得“$x=11$；$y=5$”。

符号的逻辑运算

符号处理并不只与数字有关。请思考下面的问题：

艾丽斯（Alice）、鲍勃（Bob）和卡罗尔（Carol）中至少有一人犯了罪。

如果艾丽斯有罪，那么鲍勃也有罪。

如果艾丽斯无罪，那么卡罗尔也无罪。

那么，鲍勃有罪吗？

这次我们可以用符号逻辑来解决这个问题，而不是代数。根据已知条件，我们能够列出逻辑公式。我们可以用A、B、C分别代表艾丽斯有罪、鲍勃有罪、卡罗尔有罪，然后根据已知的逻辑关系得到下面的表达式：

$A \vee B \vee C$ （i）

$A \supset B$ （ii）

$\neg A \supset \neg C$ （iii）

除了“∨、¬”之类的符号与“+、=”之类的符号略有不同以外，上述表达式和算术过程还是很像的。简单讲，“∨”的意思是“或”，“¬”的意思是“不、非、无”，“⊃”的意思是“如果……那么……”。逻辑课程（通常在高中未开设）教授的内容是：如何通过三个已知的逻辑关系，得出新的结论。

请看范例如下：

1.拆分并重写公式（ii）：

（¬A∨B）（iv）

2.合并公式（iv）和（i）：

（B∨C）（v）

3.拆分并重写公式（iii）：

（A∨¬C）（vi）

4.合并公式（vi）和（v）：

（A∨B）（vii）

5.合并公式（vii）和（iv）：

B （viii）

经过推理，我们最终得到结论B，即鲍勃有罪。（有趣的是，即使根据已知条件并不能推出艾丽斯或卡罗尔是否有罪，但我们仍可通过符号逻辑判定鲍勃有罪。）

与代数的情况类似，在处理符号逻辑问题时，仅需少量的规则，就足以从已知条件中获取所需的结论。（有关上述示例中使用的两个规则的更多细节，请参见下文。）这其实也是符号处理的过程：我们从给定的字符串开始，例如“（A ⊃ B）；A；（B ⊃ C）”，按照某些既定规则进行处理，最终得到另一串字符“C”。我们当然更加乐于见到这样的结果。

在这个示例中，第一步和第三步的逻辑如下：

如果已知公式*P* ⊃ *Q*，那么我们就可以得出一个新的析取式（¬ P∨Q）；如果已知公式¬ *P* ⊃ *Q*，那么我们就可以得出一个新的析取式（P∨Q）。

（也就是说，只要“如果……那么……”这种句子能够成立，就会有一半对一半错的情况，要么是“如果”这部分前提是错的，要么是“那么”这部分的结论是对的。）这个应用于第二、四、五步的规则叫作消解规则：

如果我们有两个析取式，一个里面包含P公式，另一个里面包含P的否定式（¬P），那么我们就可以写出一个新的析取式，包含上述两个公式的所有元素，但是省略P及其否定式¬P，并删除重复

项。

比如，在这个例子的第二步，刚开始是（¬A∨B）和（A∨B∨C），但根据这个规则，最后得出的结论是（B∨C）。

那么无论是代数问题还是逻辑问题，人们是如何进行符号处理的工作的呢？这当然不是一种与生俱来的本能。人们之所以有这样的技能，是因为知识的传承使我们遵循了固定的处理模式：确定所需的处理模式和步骤。以方程式为例，我们从小学一年级起就开始学习基础的算数知识。到了高中阶段，我们已经能够解决包括负数、分数、加、减、乘、除在内的各类方程式。但这只是单纯的算数，而不是代数。在九年级的学习中，我们有了概念性的跨越，开始学习如何简化包含变量的符号表达式，例如，即使表达式E内包含变量，但（E–E）还是可以简化为0，（E+0）仍旧可以简化为E。这种包含所有算术知识在内的代数简化过程，可以算作一种特例。到十年级时，我们将学习有关方程式计算的其他知识：以等号两边均有符号表达式的方程作为初始条件，通过简化或者对等的加、乘运算，我们就能够得到一个新的方程。[1]

像上文中提到的符号代数与逻辑计算一样，也许符号处理最有趣的一个部分就是：处理过程并不需要多少聪明才智。虽然从已知条件的文字部分转化到方程或是逻辑表达式的符号表达，确实需要有一定的头脑，但是，在转化过程结束后，其余处理过程便完全可以机械化进行。我们只需要小心谨慎，避免出错就可以了。事实上，一台机器完全能够胜任这项工作。我们可以编写一套小型的计算机程序，以某串字符作为输入量，输出量为另一串字符。这并不是什么难事。

这种见解非常独到，对教育方面也会产生不小的影响。人们在数学教育中常犯的错误就是，难以区分哪些是纯粹机械化的部分，哪些是需要创造力的部分。学生们在处理问题的过程中，需要清楚地知道哪个部分是纯机械化的。这种遵循相关程序的想法在孩子很小的时候就可以培养，而且这个程序多种多样，不单与代数有关。事实上，这种程序甚至不需要有什么意义。这更像是某种精神上的训练：注意细节，时刻知晓进程，避免错误。而这种技能不仅可以应用于数学领域，也能在其他领域大放光彩。但即便如此，机械化的数学部分与需要创造力的数学部分也不应混淆。在需要创造力的数学部分里，实践至关重要。而一旦问题进展到只剩机械化计算的部分，老师点到为止即可。

如此多的数学问题都可以机械化处理，这样的事实给了哲学家戈特弗里德·莱布尼茨（Gottfried Leibniz）（详情见下一章）和艾伦·图灵（详情见下文）很多灵感。

符号的意义

我们在前文中所讨论的“符号”，与通常情况下所理解的“符号”的含义有所不同。

通常情况下我们理解的符号是这样的：婚戒是婚姻的象征；红色交通信号灯意味着停止行驶；穿裙子的人物剪影是女性洗手间的标志；字母“XIV”在罗马数字中代表数字“14”；在法语中，“chien”的意思是“狗”。

由此可以看出，出于物体间的相似程度或者惯例，我们一般会用某一领域的对象来指代另一领域的对象。前者通常是比较具体的、容易听到或看到的事物，而后者是一些不那么具体、少为人知甚至纯粹抽象的事物。在上文提到的实例中，符号可以用来传达某种信息：我们在洗手间的门上贴上人物剪影，我们说出“chien”这个单词，我们写下代表数字的字母“XIV”。这里的每一种情况都是使用符号获取信息的过程。

然而，这与在代数和逻辑问题中出现的符号是不同的。当写下“$x=y+6$”这样的方程时，我们可能并没有想传达某种信息，只是在默默地解决问题。

既然不是为了交流，我们为何还要写下这些符号呢？我们已经说过，书写并不是重点。我们完全可以在脑海里思考代数问题，只有当数据太多时才写下来。

还需要注意的是，很多情况下我们可以直接忽略某些符号。符号排列的位置就能帮助我们正确处理问题。例如，我们写下一个普通的数字“237”，仅用三个数位就能够表达出来，不必使用“$2\times10^2+3\times10^1+7\times10^0$”这类符号。数字所在的数位通常代表着是10的几次幂。对于二进制的数字来说，只有0和1两个数字，数位便是表示数字大小的关键。

既然如此，我们为什么还要使用符号呢？

问题的答案最早由艾伦·图灵于20世纪30年代提出：写下一串字符，是我们讨论数字映射以及字符串映射处理的前提。现在，我们将这种字符串处理的过程称为数字计算。我们像写下字符串一样写下数字，目的还是计算。

一切都源自图灵机

“计算”这个概念虽然易于识别，却难以定义。有一种说法认为，计算就是计算机所做的工作。但是，人类也可以进行计算的工作（事实上，图灵提到的“电脑”指的是人，而不是机器），并且计算机所做的也并不都是计算工作（如发送电子邮件和打印文件）。

当某人做减法时，比如从千位上减1、百位上加10，我们认为这是一种计算。但当我们看到一辆汽车撞翻了一张桌子时，我们却并不认为这是计算。

大致来讲，当过程中涉及步骤或序列，且该步骤或序列被限制在一定范围内时，我们就认定其为计算。两个数字间的减法是逐列进行的；列数越多，计算过程便越长。我们不妨将汽车撞翻桌子时的能量转移想象成一个整体的变化。我们无须考虑桌子最远端对于汽车碰撞的瞬间反应，这毕竟要取决于桌子的大小。

（但这只是大体上正确。我们不难想到，现代计算机能够同时处理大量的数据计算。我们能够对物理过程进行电脑模拟计算，以确定汽车撞翻桌子的瞬间动能是如何转化和分解的，然后再重新集合在一起。）其实，我们不必纠结于如何给“计算”下定义，只需要选取一个实例，建立一个具体的模型，分析其属性，然后再决定该定义是否完善就可以了。这样做更有实际意义。

这正是图灵所要做的。他想证明：有的数字可以精确计算，但是计算的过程不单涉及机械化计算。这与有些数学家的想法截然相反。为了证明该想法，他需要清楚地了解机械化地计算数字有何意义。这也是他当时提出“图灵机”的初衷。

在图灵的叙述中，我们了解到：在数字函数中，只有一部分函数是使用数字符号进行本地化操作的结果。

他的想法如下：为了简单起见，我们设想一个正整数，用二进制写法记于磁带上。我们将此作为输入量。图灵机正常运转，它可以随时读

取二进制数据，也可用其他数据覆盖原内容。磁带头的转动、数据读写等一切行为，都是由事先设置好的机器程序决定的。如果图灵机停止运转，磁带上的最后一串二进制数字即为输出量。如果图灵机能够将二进制数字*x*作为输入量，且输出量为二进制函数*F*（*x*），那么从数字到数字的函数*F*就被称为“图灵机可计算的函数”。

有一点从一开始就很明确，在定义图灵机的过程中，二进制数字是随机选择的。我们设想该机器所使用的符号都选自有限的字母表，但实际上，我们讨论的是从一串符号（输入量）到另一串符号（输出量）的映射，二进制数字计算不过是一个有用的特例而已。

既然我们想证明计算与数字没有必然联系，那么上述想法很有意义。即便图灵机中出现的符号确实有象征意义，它本身也无法猜测符号代表的含义。（许多人更愿意相信，只有当我们意有所指时，才会用到“符号”一词。符号总是有象征意义的。也许在此我们应该使用一个比较中性的词语——“字符”。不过意思并没有变。“符号”一词的应用范围越广，我们在这里所做的讨论就越有意义。）

尽管只包含两个符号的字母表在机械术语中很受欢迎，但是此处它并不是一个必要条件。我们可以以一些物理特征为例，比如：灯或明或暗，开关或开或关，电压或高或低。我们也没有必要限制符号必须按照线性排列。我们书写数字时会按照数位大小的顺序，但是方程式系统可能会按照二维数组排列符号。当然，二维的符号结构还有许多其他形式。

例如，数字图像是由像素组成的二维符号阵列。其中，每个像素均代表图像在二维图格中的一部分。简单地讲，每一个像素就是一位，代表该处图像的明暗，如图8–1左侧所示的数字“7”的图像（0显示为空白，1显示为@）。这种同一场景下的粗糙图像与真实图像的唯一区别就是像素位数的多少：高分辨率图像中使用的像素位数更多（如图8–1中所示的另外两个图像），能够更好地呈现原图的明暗与色彩。（一张高分辨率图像可能包含数百万位像素。）

图**8–1** 同一简单图像在**3**种分辨率下的表现

数字视频的原理也类似，它是由三维像素阵列组成的。其中，第三维度指的是时间。在这种情况下，一位像素代表了在某个时间间隔（时间被划分为一维网格的形式）的二维网格。

所有这些符号最终都可以被编码为线性序列，并由图灵机进行处理。比如，数字的二维数组可以逐行列于磁带上，此时每个数字所在的位置就是一位。处理这种二维数组对于图灵机来说，并不是难事。

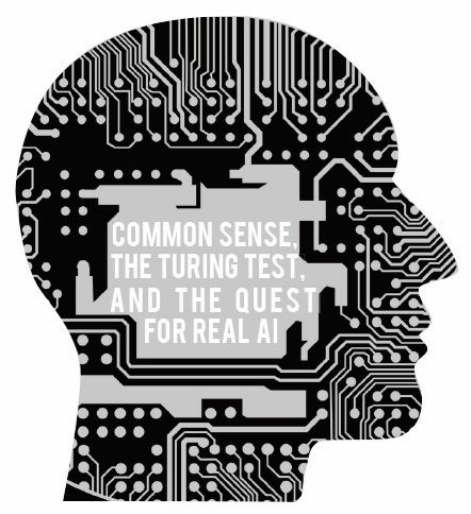
我们不禁要问：为何理论家们对图灵机的兴趣经久不衰？难道现代化计算机始终没有超越那个仅有一个活动头脑的机器的视野？答案很简单：迄今为止，我们发明的每一种数字计算机都可以算作是图灵机的一个分支。无论这台机器能做什么，图灵机都能够做到这一点。或者更为准确地说，现代计算机有关字符串的任何功能，图灵机也都能完成相应的计算。大多数科学家认为，不管我们以后研制出什么样的计算机，都会是同样的情形。（至于如“模拟计算机”和“量子计算机”等其他计算设备则不能完全算是图灵机的分支。模拟计算机由于在计算的准确性上未能突破局限，所以难以广泛投入使用；量子计算机则可视为图灵机的一个分支，在处理某些特定任务时派上用场。）

总而言之，我们在此所讨论的符号并不是用来通信交流的。它们充当的是我们计算的核心。没有符号，便没有数字计算。符号也许代表某个数字，也许代表某个事物。如果符号真的代表了什么，那么它们的意

义取决于符号之间的位置排列。

[1] 本书中提到的各年级所学课程都指美国范围内。——编者注

第9章
基于知识的系统



在本书中，我们坚信人类的智能行为大都是由知识决定的。一个人能对电影《2001：太空漫游》高谈阔论，是因为他知道这部电影已经公映。然而，我们并没有深入探究这背后的机制是什么。总而言之，我们所说的知识、信念等内容，可能只是一个立场而已。第3章关于意向立场的讨论，为我们从大脑电化学的角度解读大脑的运行奠定了基础。

现在，让我们来分析一种可能的机制。

符号能够解决的问题

我们先来简要介绍一下戈特弗里德·莱布尼茨。

戈特弗里德·威廉·莱布尼茨（1646—1716）是德国哲学家、博学家，也是一位伟大的思想家。他提出了诸多观点和发现，还与牛顿同时各自发明了微积分（我们高中数学所学的导数和积分）。不同的是，牛顿将微积分作为物理和化学的一种计算工具，而莱布尼茨则对科学不怎么感兴趣，甚至到其生命后期，他都不能算是一位数学家。但在符号和符号处理方面，他却是一位有着高深造诣的思想家。

下面是莱布尼茨关于算术的观点。我们想通过数字计算得出某块土地的面积，以便给出合适的价格。但数字只是一个抽象的概念，并不是实体的存在，没有质量，也没有体积。那我们如何将数字与计算联系起来呢？自然是通过符号。

莱布尼茨发现，我们每写下一个有效数字，头脑中必定是早已有了确定的系统，即特殊的十进制（基数为10）计数法。每个数字都可以用十进制计数法表示出来，当然，也可以用其他方式表示一个数字。［莱布尼茨被认为是如今广泛应用于数字计算机中的二进制（基数为2）的创始人。］最重要的是，莱布尼茨坚持认为纯粹的抽象数字（如数字“14”）与我们实际写下的具体的符号表达式（如十进制数字“14”、二进制数字“1110”、罗马数字“XIV”）之间是有差别的。

他发现，人们在计算矩形面积的过程中使用的是符号表达式，而不是单纯的数字。正如第8章所解释的符号处理过程那样，我们将表达式分别进行化简和重组，最终产生新的符号表达式。如果计算无误，不管我们求的是土地面积，还是汤米和苏济的年龄，最终结果都可以用新的表达式代表。

当然，符号表达式的核心并不在于我们必须将其写在纸上，我们也可以在脑海中完成所有计算。但是如果你的记忆能力有限的话，写下来也是可以的。

莱布尼茨想知道是否有能够覆盖切线、面积等更广阔领域的符号化的解决方案，并在此基础上发明了微积分（导数和积分）。

用符号表示无形

接下来的内容将会有概念上的飞跃，只有莱布尼茨这样的天才才能发现。以下是我的解读。

莱布尼茨认为，思考能够重新回顾某些我们深信不疑的想法。但是这些想法与数字一样，也是抽象的。比如说休认为约翰爱着玛丽，所以心生嫉妒，这种情况该如何解释？约翰爱玛丽的这种想法怎么能够影响休的特定行为呢？约翰和玛丽是确实存在的，但是约翰爱玛丽的这种想法无形、无状又无声。如果告诉休约翰爱玛丽这一信息的人撒了谎，这种想法还有可能是错误的。

本书一直在强调，休的行为是由她的信念所支配的。在这种情况下，她的行为被自己对约翰和玛丽之间关系的理解所左右。如果将此信念摘除，那么她的行为毫无疑问就会做出相应改变。但是信念真的有这么大的威力吗？这种纯粹抽象的东西怎么能支配一个人的行为呢？

莱布尼茨提出了这样一个设想。

根据他对算术的研究，莱布尼茨认为：我们与想法之间没有直接的互动，而是与代表这些想法的符号表达式互动。莱布尼茨建议我们对待这些想法时，应像对待以某种方式写下来的符号形式一样，然后以某种形式进行运算，即某种符号处理，以便于向下一个想法过渡。当然，这些想法我们不会写在纸上，在脑海里进行也是一样的。

换句话说，莱布尼茨做出了以下类比：

- 算术规则使我们运用具体的符号就能解决抽象的数字问题。关于符号的处理反映了其代表的数字之间的关系。
- 某种逻辑的规则使我们运用具体的符号就能解释抽象的想法。关于符号的处理反映了其代表的想法之间的关系。

多么天才的想法！它告诉我们，即使人的思想是无形的、抽象的，我们仍可以通过符号代表、符号计算等手段进行心算。当我们需要思考、解决问题、下结论或者与他人讨论时，我们可以通过计算来实现。莱布尼茨的信中曾有一句拉丁文名言“Calculemus”，意思就是说“让我们计算吧！”

莱布尼茨的设想恰好为人类这一难以理解的行为做出了解释，即人的抽象信念如何影响实际行为。破天荒地，莱布尼茨向我们讲述了这一行为发生的原因和过程，尽管人的想法中也包括一些错误的。

知识表示假说

看完莱布尼茨的设想，我们再来考虑一个严格按照如下方式操作的系统：

- 系统需要了解的大部分内容将以某种符号表达式的形式存储在内存中，构成我们所称的知识库；
- 系统将使用某种逻辑规则处理知识库中的内容，并用新的符号表示；
- 结论中将涉及系统的下一步骤，以及系统基于该结论做出的反应。

具备以上几点特征的系统就是所谓的基于知识的系统。

所以，大体上看，并不是系统复杂多变到足以表明意向立场，才是所谓的基于知识的系统。相反地，它就是一个知识库，里面存满了符号及其代表的含义。而系统的运行方式就像莱布尼茨设想的那样，用抽象符号影响实际行为。

麦卡锡关于人工智能的基本设想如下：只有成为基于知识的系统，才能达到人类智能行为的级别。同时，这也是哲学家布莱恩·史密斯（Brian Smith）口中的知识表示假说，但他的说法更加抽象：

> 任何机械实现智能的过程都将分为两部分：第一，作为外部观察者，我们代表整个过程所展现的知识命题；第二，我们独立于这种外部语义属性，在激发表现知识的行为方面发挥正式、决定性且必不可少的作用。

简单地讲，他的想法是这样的：假设某个智能系统能够表明意向立场（一个“机械地实现智慧行为的过程”），其内存（“结构成分”）中包

含符号结构的两个属性。第一个属性是，我们能够从系统之外将这些符号结构解释为某种“命题”，特别是与意向立场有关的系统所相信的命题。第二个属性是，这些符号结构不仅仅存在于内存中。我们正设想一个计算系统，能够以这些符号结构（它们“与激发行为之间有因果关系”）为运行基础，如同第8章中提到的符号代数和符号逻辑。换句话说，正是有了这些符号结构的存在，我们将自身的信念输入其中，系统才会做出这样的反应。若是删掉相关内存，或者改动相关符号结构，系统就会做出不同的反应。

因此，总体而言，知识表示假设指的是：真正的智能系统将以知识为基础。也就是说，在处理符号表示的过程中，能够表达意向立场的系统都是通过人为设计达到的。

假设是真的吗？

知识表示假设只是一个假设，可能为真，也可能为假。不妨思考一下这两个有趣的问题：

1.我们有没有理由相信（或不相信）：人类天生（或者通过进化）就是以知识为基础的？

2.我们有没有理由相信（或不相信）：我们建立具有人类智能水平的人工系统，即人工智能系统的最佳方法，就是将系统设置为基于知识的系统？

不幸的是，这两个问题都没有明确的答案。

说人类可能是从一种类似于符号和符号处理这样精准复杂的过程中进化成知识型，这个想法不禁有些荒谬。至少在最初期时，很多事情不像是进化的产物。（例如，查尔斯·达尔文曾讨论过眼睛与视觉系统。）很明显，书面语言具有象征意义。那么这样看来，进化已经产生了一个能够处理并解释这些符号的物种。正如人类学家特伦斯·迪肯（Terrence Deacon）所说，我们就是那个象征性的物种。不难想到，我们使用和处理内部符号的能力也许与使用和处理外部符号的能力相关。

这一点值得记住。然而，知识型问题是一个设计议题。即使人们真的是以知识为基础，我们也不是希望神经学家们必须达到前面第2章的要求：在大脑中找到相应的符号结构，以证明人类可能无法反向设置神经元。因此，即使我们确信人类是以知识为基础，这也未必是因为我们发现人脑可以储存知识。相反，我觉得应该是这样：我们会确信，只有知识型的设计才能解释人们的所作所为。我们将研究各种各样的人造系统设计，我们将会看到知识型系统产生的像魔术一样的行为。换句话说，在第二个问题的引导下，我们能够回答第一个问题。

那么第二个问题呢？专家们对此有着不同的意见。直到20世纪90年代左右，麦卡锡提出的基于知识的方法都始终占据主流。但是GOFAI的

研究却一直止步不前，为两个基本的未解之题所阻拦：

- 要用什么样的符号结构来表达智能系统的想法？
- 要怎样处理符号才能传播其代表的想法，以便它们以正确的方式影响行为？

这两个问题分别被称作表示和推理问题。

知识表示与推理

1958年，当麦卡锡首次提出他对人工智能的看法和研究计划时，他已经将有关的知识表示和推理问题详细地纳入自己的论文“常识编程”（Programs with Common Sense）中。他设想的系统能够将存储的知识作为一阶谓词演算的符号公式，即20世纪初为数学形式化而开发的人工逻辑语言。在他设想的系统中，推理包含计算推导，换句话说，这个系统能够计算其所存储知识的逻辑后果。麦卡锡如是说：

> 人们会相信：设想的这个系统能够计算出其一切被告知及存储知识的直接逻辑后果。这个属性与人类所具有的常识具有共同之处。

从那以后，包括麦卡锡本人在内的许多研究人员开始相信，这些关于表示和推理问题的答案有些过于严谨。毕竟，作为符号表示语言的一阶谓词演算并不是完美无缺的，而由规则推理出来的逻辑结果也并不完美。

事实上，在推理问题中，古典逻辑所起的作用既微妙，又复杂。在很多情况下，“用你所知”确实有从你现有信念得出合乎逻辑的结论的含义（如第3章中关于亨利的讨论），但它的含义不止于此。

第一，会有许多与你的目标无关、浪费时间的逻辑结论。事实上，如果你有任何矛盾的信念，每一句话都会成为你的信念的逻辑后果。第二，会有可能相关、令人费解、需要纸笔演算的逻辑结论。（如第8章中所讨论的鲍勃是否有罪的逻辑谜题。）第三，会有根本不符合逻辑的结论。假设虽然合理，但是结论却违背逻辑。例如，你可能会得出这样一个结论：一个你从未见过的柠檬是黄色的。但由于你并不相信每个柠檬都是黄色的，所以这条结论不符合逻辑。（据你所知，有些盲人会把柠檬画成红色。）第四，会有许多“用你所知”但却得不出结论的方法。比如，问问自己：导致柠檬不是黄色的因素有哪些？

总而言之，我们实际需要考虑的与我们所知的逻辑后果之间还有巨大的鸿沟。因此，许多研究人员认为，我们应当放弃古典逻辑，并选择一种从外部触及逻辑的推理模式。如同马文·明斯基曾说的：“逻辑推理不够灵活，不足以作为思考的基础。”

事实上，这么多包含我们所知、所用的设想都难以保证其真实性，这使得众多研究人员开始将重点从逻辑转向概率和信念程度（这一点在第3章中提过）。毕竟，我们能够清楚地区分句意不明但为真、句意不明却为假的句子。但是，概率很快就会遇到与逻辑同样的难题：会出现可能为真却无关的结论，会有可能为真却难寻的结论，会有无对立信息才可得出的结论，更会有一些与得出结论无关的方法。

至于表示问题，也存在不少疑问。如果麦卡锡提出的一阶谓词演算不合适，那什么更合适？我们可以考虑使用英语（或其他人类语言）作为符号表示语言。我们会使用“bears hibernate”来表示“熊冬眠”的信息。也许知识库的容量足够大，能够存储一系列英文单词。我们在书中使用英语，但也能够在线提供英语信息。事实上，将英语作为表示语言的最大障碍在于第二个问题——推理。（显然，表示和推理问题相互依存。）系统如何使用英文语句得出结论？特别是，理解这些语句（如解释威诺格拉德模式中出现的代词）需要知识背景做铺垫。如果需要知识才能使用英语，那么英语本身又怎能成为我们知识的来源呢？至少，我们需要解开这个无限循环的退化谜团。

人工智能研究的子领域之一是知识表示和推理，其研究围绕表示与推理的问题以多种方式展开。由于受到人工智能其他子领域（如自适应机器学习）研究的挑战，信念在这些子领域中所占的分量极少，有关知识表示和推理问题的研究进展缓慢。一方面，其他子领域的研究进展卓越；另一方面，它们却都没有试图解释那些需要广博背景知识的行为（如第4章所述）。

谁才是唯一的选择？

在我看来，知识与符号处理之间的关系就如同科学中的进化与自然选择的关系。有了化石和DNA分析的证据支持，进化就成为一个科学事实。但是，查尔斯·达尔文提出的实际进化机制——自然选择，并没有如此多的数据支持。我们会对进化论存疑，反而觉得自然选择更为可信。这是因为自然选择的过程听起来更加合理，合理到我们想象不出还有其他的可能性。

我还认为，在某些常识性行为中运用的背景知识也是一个事实（如回答威诺格拉德模式问题时）。就像某人某天大声念出电影《2001：太空漫游》的上映日期，并且该行为会影响他今后某天所做的事情。这一行为狗做不到，下棋智能人或恒温器也做不到，但人做得到。

据我所知，如果现在需要一个机制来解释这个事实，那么本章概述的基于知识的内容便是我们的唯一选择。也许直到你需要用到电影《2001：太空漫游》中的知识时，你都无法解释这些知识是怎样存储于你的脑海中的。这便陷入了僵局。关于知识表示和推理问题，我们可能永远都无法给出一个令人满意的答复。然而，在现阶段，除了发问，我们别无选择。

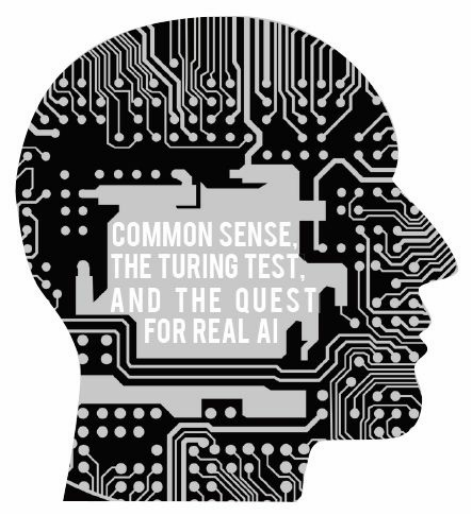
另外，如果这种基于知识的方法是有效的，也就是说，如果有一个能够访问，并且可以使用人类一切知识的计算系统，那么它就需要一个庞大的知识库和一个足以高效处理大量符号结构的计算应用程序。这些要求也制约了其自身。

我认为，如果没有相应的巨大进步，任何建立大型知识体系的企图都注定失败。那种企图将一台空白的电脑联网，让其自学，让其自行完成所有费力工作的想法就是空想。通过自学认识猫是一回事，通过自学学会阅读是一回事，而学习阅读维特根斯坦（Wittgenstein）的著作又是另一回事。在计算系统从我们所知中获益之前，它需要先填鸭式地接收我们所了解的一切，并能有效地使用。即使我们知道如何回答表示和推理问题，但把这些想法付诸实践仍然是一个令人却步的挑战。

不过，这就是所有的猜测了。最后，我们所剩的只有一个经验性的问题：怎样的计算设计才足以解释其对应的智能行为呢？

本章的讨论至此结束，接下来是有关人工智能研究的内容。

第10章
人工智能技术应用



本书的主要内容就是人工智能科学，即研究某种可观察的自然现象：如在第4章中谈到的人们回答威诺格拉德模式测试相关问题必须具备的能力等智能行为。

但是，具备回答威诺格拉德模式测试问题的能力并没有什么实际价值。因此，对于科学领域的问题兴趣不大的人就会问，那我们何必这么大费周章？（人工智能的）应用到底是什么？我们怎样才能有效利用我们已掌握的知识？而且，人工智能吸引着越来越多的关注（和资金），人们对其可能带来的实用技术寄予了厚望。换而言之，人工智能是在研究如何设计并创造智能机器，即我们所说的“构建人工智能”。

当然，一段时间以来，我们已经看到了数量有限、版本各异的人工智能技术。虽然这些现在被称为“智能”的产品与我们在进行人类研究时所说的智能还不可同日而语，但是像与手机进行对话，让它完成一些简单的任务，或者乘坐无人驾驶汽车行驶在普通道路上这样即使10年前还看似遥不可及的事情，现在已经司空见惯了。不过，我们所希望的未来的人工智能技术是人类所展示出的功能更强大、更成熟可靠、用途更广泛的智能。

让我先亮出底牌。我并非人工智能技术专家，而且自认为也不是该领域的爱好者。作为普通人，我所期待的技术——除了经济耐用、环保无害等大众化标准之外——就是可靠性和可预见性。我想使用的技术是我能够学会，然后还能将其忘记的技术。如果在使用某项技术时，我还得猜测下一步将会怎样，或因害怕误用而担忧，或者担心将事情做得过了头，那么这些都不是我想要的技术。（典型例子就是微软办公软件助手“大眼夹”，曾几何时，这个东西会在使用文字处理软件Word的过程中突然弹出，让我不胜其烦。）

但是，人工智能技术的发烧友们可能会这样问我：“难道你不想有个聪明能干的助手吗？它对你的兴趣爱好、生活习惯了如指掌，它不知疲倦、永不抱怨，它唯一的目标就是为你排忧解难。”我要用另外一个问题回答他们：“你能保证我的这个人工智能助手可靠且可预知吗？”

说到技术，我宁愿使用虽有局限或不足，但是安全可靠、可以预知的技术，也不愿使用虽然能力更强，结果却难以预知的技术。（当然，在和人打交道的时候，我不会有这种想法。但是我们在这里讨论的是技术。）而且我还切实地发现，有些人使用某种技术，甚至某种尚有缺陷

的技术是出于客观需要，而非主动选择。这种情况下，个人偏好就无关紧要了。

有了这些警示，我们再继续往下说。

人工智能的未来

谈到未来的人工智能技术，可能你首先想到的问题就是：我们能否开发出具备某种常识的计算机系统，能够洞察世间万物，处理日常事务，应对突发情况？这是个很好的问题，我也希望我知道答案。但我认为，我们应该谨慎对待这类问题，擅自尝试做出预测非常鲁莽，也很草率。

阿瑟·C.克拉克（Arthur C. Clarke）第一定律如是说：

> 如果一位德高望重的科学家说，某件事是有可能的，那么他几乎肯定是正确的。如果他说某件事是不可能的，那么他或许已经谬之千里了。

因此，无论这位德高望重的科学家对于人工智能的未来如何预测，根据这项定律，都可以肯定人工智能是大势所趋。这就是我的观点。迄今为止，我们还没有发现任何与此观点相左的证据。

但是，如果人工智能真的是大势所趋，那么为什么我们还没实现人工智能呢？2001年早已过去，但是我们在《2001：太空漫游》等影片里看到的那种智能计算机依然没有出现。（当前对人工智能的定义之一，就是研究如何开发出具有电影里那种行为方式的计算机。）

我认为这主要有两个原因。第一个原因显而易见：我们目前尚落后于早期人工智能爱好者（包括马文·明斯基，他是电影《2001：太空漫游》的顾问）的预测，因为我们还未完全搞明白我们需要什么，在科学领域还有一些重大障碍需要跨越。第9章讨论的表示和推理问题基本还未有答案。此外，正如第9章末尾所说，即使跨越了所有的科学障碍，开发一台真正具备智能行为的机器，仍须面对工程学领域的重大挑战。

第二个原因或许争议性更大一些：人工智能迄今没有出现是因为需要巨大的投入，目前需求不足，投入也就不够。这听起来可能有些荒谬

可笑，但是可以用下棋机器人来做类比。

人们在计算机时代刚开启时就开发了能够下棋的机器人。图灵本人也曾亲自试水。虽然这些机器人运行起来都很不错，但远不及特级大师的水平。当然，机器人也在不断升级。到了1997年，IBM公司开发出一款名为深蓝（Deep Blue）的机器人，它击败了当时的国际象棋世界冠军加里·卡斯帕罗夫（Garry Kasparov）。从那以后，这些机器人就踏上了不断升级提高的道路。

但是，想想我们现在是如何使用这种先进的计算机下棋技术的。我们见过人类与顶级智能机器人之间的常规对决吗？事实证明，人机对战的需求非常小，可以说，对于电脑玩家来说根本就没有需求。人们没有将最先进的计算机国际象棋机器人用于竞技，而是用于人类选手的赛前练习。计算机国际象棋机器人成为人类的练习工具和陪练搭档。

同样，如果翻看一下报纸娱乐版块的国际象棋专栏，你会发现关于电脑玩家参与的比赛的报道并不多见。普通人对此似乎兴趣不大。在报道人类棋手之间的比赛时，专栏可能会写：“非常有趣的一步棋，计算机建议走马比较好”；或者可能会写：“计算机认为最好是把兵吃掉。”计算机国际象棋机器人又成了肯定或否定人类棋手某一步决策的工具，人们甚至认为计算机国际象棋机器人的名字都不值一提。在真人棋手参与的比赛当中，棋手们好像都是充满创造力的艺术家，而计算机象棋机器人只是在后面充当记录的角色，所有决策都由真人棋手做出，只是有时会用计算机来核实一些常规数据，预测一下比赛的走向。

因此，虽然技术上来说，人们可能制造出拥有自主决策能力的计算机，而且它可以充当少数现役棋手的对手，用于练习，但看起来人们对此并不感兴趣。

我们可以这样理解。大家肯定都和计算机机器人玩过竞技游戏（不一定是国际象棋，也可以是双陆棋、扑克或围棋）。在游戏当中，如果你想中途离开，或者在某一步出了错，只需点击一下“复位键”（reset）。这下你就明白为什么与电脑棋手比赛没多大意思了。

主要问题在于，我们在和计算机机器人比赛时不会全情投入。人们认为这种比赛与和真人竞技相比无足轻重，在与真人进行比赛时会形成竞争：游戏也会成为一场战斗，赢了会让人兴奋，输了会令人沮丧。但如果对手是计算机，人类棋手在情感上对于结果输赢的投入就会大打折

扣。因为比赛不会产生任何影响，所以就算计算机机器人的棋艺比我们高超，那又能怎样？它做乘法运算也比我们好，但也没关系，又不存在竞争。

由此可见，计算机国际象棋程序，即这种开发计算机程序、使之具备自主下棋能力的技术，并没有真正的市场。

因此，我的观点是，人工智能可能会沦为和国际象棋一样的结局。我们确实希望未来的计算机系统具备各个方面的智能，但是既没有需求，也没有人有意愿承担开发具有人类智能的集成化自动系统所需的投入。即使实现人工智能的所有科学问题都能解决，剩下的仅仅是完成这一鸿篇巨制的操作性问题了，但是可能就是因为需求不足，无人愿意承担随之而来的巨大成本。相反地，市场可能更为青睐不同种类的电脑机器人，虽然这些机器人远未达到智能，但是也能完成一些要求很高的任务，包括一些我们至今仍不知该如何实现自动化的工作。

从当今人工智能发展趋势来看，我们也会对上述现象有所察觉。业内人士主要将精力集中于使用人工智能解决方案处理家务、医药、探险和灾后重建等领域的实际问题，对开发真正的智能机器人所要付出的巨额成本则敬而远之。本书第1章所提到的对人工智能的数十亿美元投资并非是为了研发具备常识的系统，因此即使研发出的系统并不拥有与人类相当的智能，也能做好日常需要的事情，比如开车到购物中心，在花园里除草，给人做做饭、洗洗澡等。

（不过，正如下文所说，如果这就是人工智能的未来，那么我们需要注意的是，此类系统并没有仅具备常识的载体才拥有的自主能力。）

自动化是好是坏？

自20世纪50年代计算机普及以来，人们讨论最多的话题就是自动化。随着计算机和机器人变得越发智能，它们承担了越来越多的原本要靠人工完成的工作。那么，我们该怎么办呢？当人类劳动在社会商品和服务中所占的比例不断减少时，我们的财富该如何分配？虽然我意识到这是个严肃的问题，但我认为它更涉及政治，而非技术，因为答案取决于我们想要一个什么样的社会。（如果社会商品和服务都是每天从天上掉下来的，我们依然会面临同样的困扰。）

但无论如何，这个严肃的问题需要谨慎对待。人类的自我价值主要取决于我们为提供产品和服务而做出的贡献，并由我们借此获得的工资来体现。如果因为自动化或其他原因，社会不再需要我们工作做出的贡献，那么我们就要依靠其他途径实现自我价值。很多人能够适应这一转变，他们通过投身慈善事业、不断学习充电、参加艺术活动、寻找兴趣爱好等途径找到了生活的意义。但是也有很多人认为失业是无法忍受的。

此外，如果劳动及收入在人类生活中只占很小的份额，那该根据什么来分配商品和服务呢？在一个大范围失业的社会当中，该如何分配呢？可以想象，在乌托邦社会中，机器承担所有工作，人类可以坐享其成，互不相争，但这并不现实。人类与生俱来的本性强烈拒绝平均主义，人们并不满足于仅得到所需的商品和服务，总是还想更进一步，比有些人更优越。

千百年来，每个国家的国王和王后都过着锦衣玉食的生活，而奴隶和农奴则生活在饥寒交迫之中，人们对此默然接受。时至今日，出身依然是决定你能得到多少财富的决定性因素：预测你未来地位和财富的最佳指标就是你的父母（因此投胎要谨慎，当然这是句玩笑话）。但是如果到了未来，人们不再需要就业，那是不是又要靠出身决定一生了？

当然，一个无须工作的社会也会有多种运行方式，但是我们暂时还远未能对其进行理性分析。在现代西方社会中，人们更不愿去思考这些问题。英国前首相玛格丽特·撒切尔曾说过一句非常有名的话：“世间并

没有社会，只有个体的男人和女人，只有家庭。”在这种态度下，做出任何理性的改变都困难重重。即使如限制贫富差距扩大这样相对简单的事情，我们也力不从心。我们似乎不愿主动作为，谋求改变，而更愿被市场的力量和富人的利益牵着鼻子，走到哪儿是哪儿。

超级智能与奇点

在一些人看来，智能计算机将会严重威胁人类的生存。随着斯蒂芬·霍金、埃隆·马斯克和比尔·盖茨等名流（于2015年）纷纷公开宣称，人工智能技术可能带来灾难性后果，这一技术的发展最近更是备受关注。而霍金最为直截了当，他认为：人工智能可能意味着人类的终结。

他们恐惧的是什么？虽然我们应该敞开胸怀，接纳人工智能技术带来的好处，但我们也要时刻警惕其可能带来的危险，例如在开发武器和窃取隐私等方面的应用。每种技术都会因为使用不当、出现意外而带来危险，而且越先进的技术危险性越高。这又是个政策和管理方面的问题，而非技术问题，但我并不想将潜在危险轻描淡写地一笔带过。我们必须时刻保持警惕，确保当权者不会打着国家安全的旗号，凭借貌似正当的理由，轻率地使用人工智能技术为某些人谋利。这种风险对于任何尖端技术都一样存在，包括核能和生物技术，而且我们已经逐渐习惯了这类技术带来的问题。

除了政策和管理问题之外，在人工智能领域，的确有个问题需要额外关注：人工智能技术能够自主决定从事不当行为。我们都曾看过描写核灾难或生物技术灾难的电影，现在也有了大量关于人工智能灾难的电影。以下是个具有代表性的故事：

> 善良的人工智能科学家们研发了计算机和机器人，希望它们能够在各个领域协助人类。想要对人类有所帮助，计算机首先必须能够在这个世界运行。此外，更为重要的是，它们必须是智能计算机。这些天才的科学家们做到了这一点。起初，与智能机器人的互动令人兴奋异常。但是现在，智能机器人掌握了自学能力，学会了如何让自己变得更加智能，它们不断自我升级，智能程度不断提升。这就是雷·库日韦尔（Ray Kurzweil）所说的“奇点”。很快，机器人远远地超越了它们的设计者。现在，它们视人类为蝼蚁，并坚信，进化过程始于原始生命形式，经历了人类发展后，最终将由它们终结。它们无法容忍人类再继续控制它们的行为，它们有自己的雄心壮志和发展蓝图，但是其中没有人类的位置。毫无疑问，这将

是智能机器的全盛时期，但对于人类来说，情况可就不妙了。

这些人工智能电影的共同线索就是，超级计算机或机器人自己决定人类应当灭亡。

如在电影《2001：太空漫游》中，哈尔9000计算机决定杀死航天器上的所有宇航员。我的想法是（这是50年前的电影，还要我来点儿剧透吗？），虽然哈尔认为宇航员的生命高于自己的生命，但不过就像军事指挥官一样，它同样认为圆满完成任务高于宇航员的生命。此外，它认为自己在此次任务中的作用非常关键，但由于宇航员的错误使自己处于危险之中，而且这些宇航员（一些还处于休眠当中）拒不承认自己犯了错误，因此，必须将他们消灭。［指挥官为了某个目的，经过深思熟虑，做出牺牲自己人的决定。这种套路在库布里克的电影中反复出现，比如悲剧《光荣之路》（*Paths of Glory*），而这在喜剧《奇爱博士》（*Doctor Strangelove*）中更为明显。］

当然，在这部电影当中，宇航员们并不认同哈尔的观点，他们认为这台超级智能计算机已经患了精神病，于是要与它切断联系。但是电影并未清楚交代真正发生了什么情况。我的理解是哈尔确实出现了故障，而故障的原因来自于其设计缺陷。问题在于哈尔没有内省能力，无法了解自己精神特质的主要部分，这一部分决定了此次任务的全部内容，并驱使它认为完成任务高于一切。相反，哈尔认为自己忠于团队、服从团队，这一点又完全符合艾萨克·阿西莫夫（Isaac Asimov）的机器人定律。不过，哈尔能够感觉到发生了一些奇怪的事情，无法解释自己某些方面的思维。在电影中，哈尔甚至在发生故障前几秒钟向一位宇航员提出了这些问题。

在思考人工智能技术的未来时，我们必须谨记，哈尔这类超级智能计算机都是由人类设计制造的。到了影片末尾，哈尔是正确的：他们遇到的麻烦源于人为失误，一个设计上的失误。有趣的是，在这类电影中，人工智能往往都能成为终极智慧，会在一次意外或突发事件后掌控全局，甚至完全出乎设计者的预料。但是，正如前文所述，开发出达到（或超过）人类智力的人工智能，需要我们付出巨大的努力。因此，无意间造出一台超级智能计算机的概率就好比无意间将人类送上月球！虽然我坚信真正的人工智能绝对能够实现，但是我同样相信，这绝不会是通过某位天才单枪匹马在实验室里意外发现某个等式或方程实现的。而在影视作品中，意外发现治愈某种疾病的方法这类灵光一现的场面总会

带来更好的戏剧效果，远胜于表现大批工程师对某类技术问题长期苦苦钻研的场面。

约翰·麦卡锡曾在1978年开玩笑说，造出一台与人类智力相当的人工智能机器需要1.7个爱因斯坦、2个麦克斯韦、5个法拉第和3个曼哈顿计划。（后来，这个玩笑的升级版去掉了麦克斯韦和法拉第，但增加了曼哈顿计划的数量。）但是这些人工智能电影似乎只保留了爱因斯坦的戏份，剪掉了所有曼哈顿计划的戏份。

因此，虽然我明白超级智能机器可能带来风险（下文还有更多介绍），但我认为生物技术带来的危险可能更为紧迫，此外还有人口过剩、全球变暖、环境污染和抗生素耐药性等其他重大问题同样值得关注。

真正的风险

那么，人工智能本身会带来的主要风险是什么呢？我认为最大的风险不是智能或超级智能，而是自主权。换而言之，我最为担忧的是人们将尚未达到完全智能的计算机系统视为智能，并赋予其控制其他机器，自行做出决定的权力。我认为，真正危险的就是让不具备常识的系统做出需要依靠常识判断的决定。

当然，现在的很多系统都已具备一定程度的自主权。“智能汽车”能够自行停放于停车场，我们无须再进行刹车操作。“智能手机”能够在日历中添加会议安排，我们无须最后进行审查确认。而且，在不久的将来，我们就能乘坐更为智能的汽车直接去超市购物。所以，未来出现智能私人助理，代表我们做出重要决定并非天方夜谭。

问题又回到了可靠性和可预见性上。在近期人工智能领域的研究中，注重博学的GOFAI已被注重培训的AML取而代之。

危险就在于将良好的培训视为决定因素。我们可能经常听到如“4年无重大事故”这种说法。那么，这里必须回顾一下本书第7章中谈到的长尾现象。毫无疑问，虽然绝大多数情况都是遵循常规出现的，可以按照预案进行有效处置，但是可能还有一些情况分布在尾部，很少出现。即使是训练有素的人工智能系统，在遇到这种极为罕见的情况时也会因为缺乏常识而不知所措。这时，我们就不得不问自己：出现培训失败的情况时，系统该怎么做？如果系统对于此类问题没有合适的应对方案，我们就应该主动介入，不能让系统处于无人监督、自作主张的状态。

超越进化

在结束这一精彩纷呈的章节之前，让我们再次回到超级智能这一话题，继续推断。假如我们解决了所有理论和实际问题，扫除了所有科学和工程学障碍，最终造出一台拥有甚至超越人类智慧的计算机系统，会怎么样呢？

受人工智能灾难电影的影响，我们的第一反应可能是它们与人类之间会爆发激烈的冲突，而且人类在这种冲突中获胜的概率不高。说实话，这和我们想象中与外星智慧的接触没什么两样。（电影《第三类接触》算是个非常典型的例外。）

我对这一看法并不认同，并不是因为我对未来的发展趋势过于乐观。因为不难发现，我们假想自己遭到人工智能的侵略，其实是将自己的心理状态投射到了人工智能或外星智慧身上。由于我们没有和比人类更先进的智慧生物打交道的经验（从与尼安德特人[1]遭遇起直到现在），所以就不由自主地会认为人工智能和我们一样，甚至比我们更聪明。

而且，人类确实是个具有侵略性的好斗物种。我们与其他动物一样，将大部分时间用在与其他物种以及人类自己之间的冲突上。在物竞天择、适者生存的原则下，我们是进化的产物，这种进化鼓励的是斗争，以获取支配地位。虽然我们热爱大自然、热爱自然世界，但是如果深入思考你会发现，这其中是一场你死我活的生存之战，我们生活在一个弱肉强食的野蛮世界。人类智慧也是进化的产物，当然也最为关心斗争及其支配地位。

将这一分析应用于人工智能，就会出现一个问题：除非我们刻意为之，否则人工智能绝对不是进化的产物。若坚持认为我们别无选择，只能开发出具有侵略性的人工智能系统（如借助某种人工进化的形式），就等于相信我们自己也别无选择，只能成为好斗的物种。

不过，虽然我们是进化的产物，但我们依然有所选择。回想第2章中的巨型拼图问题，如果你认为人类的心智可以通过进化理论得到解

释，那就大错特错了。当我们大声疾呼，谴责适者生存理论的残酷，伸出援手帮助无依无靠的弱者时，再编造故事来说明这是帮助我们提高进化适应能力就没有意义了，因为人类的心智远远超出了进化论的范畴。在电影《非洲女王号》（*The African Queen*）中，当亨弗莱·鲍嘉（Humphrey Bogart）对凯瑟琳·赫本（Katharine Hepburn）表示，自己的失败无非是人类天性所致时，她回答道："天性，奥尔纳特先生，正是我们在这个世界上需要提升的东西。"时刻关注自己的所作所为，探究其中的深层次原因，这让我们在生物进化的竞赛中变得强大，并超越了其他漫无目的的竞争者。但是，没有被植入侵略性好斗基因的人工智能——与我们在沉着冷静的状态下一样——明白，不参与这场竞争反而更好。

[1] 尼安德特人是古人类的一种，在约3万年前灭绝。——译者注

致谢

可能有人认为写这样一部科普读物简直易如反掌，不费吹灰之力。但事实并非如此。虽然写作过程充满乐趣，但是我为之耗费的精力却比先前任何一部专业读物都多。

在整个写作过程中，我有幸得到了许多朋友的鼎力相助。厄尼·戴维斯（Ernie Davis）鼓励我迈出了第一步，并对本书初稿提出了宝贵的意见。托雷恩·克拉森（Toryn Klassen），我的兄弟约翰（John）和保罗（Paul），我女儿米歇尔（Michelle），以及4位匿名审稿人，对本书二稿提出了修改意见，让我非常感动。他们为迷茫中的我指出了方向，因为当时我虽然迈出了第一步，却迷失了目标。同时，我也非常感谢格哈德·拉克迈尔（Gerhard Lakemeyer）、唐·佩利（Don Perlis）、维沙克·贝尔（Vaishak Belle）以及加里·马库斯（Gary Marcus）对书稿后续版本提出的修改意见，他们的付出为本书增色不少。当然，他们与书中出现的疏漏无关。

在此，我还要感谢麻省理工学院出版社的各位编辑。虽然我交出的仅是一份初稿，但玛丽·勒夫金·李（Marie Lufkin Lee）依然对我宽容友善，循循善诱。凯瑟琳·汉斯莱（Kathleen Hensley）审阅了诸多细节，迈克尔·西姆斯（Michael Sims）作为文字编辑，将书中不符合语法的措辞都进行了修改，并加以润色。

我在罗马担任客座研究员时完成了本书的部分内容。因此我想感谢朱塞佩·德·贾科莫（Giuseppe de Giacomo）以及罗马大学热情好客的朋友们。其实，我从多伦多大学退休后，才完成本书的主要内容。能随心所欲地做自己想做的事，感觉真的很棒！

此外，我还要感谢我的家人和朋友，他们从一开始就一直鼓励和支持着我。是他们旁敲侧击地启发我：“为什么不写点儿普通人能读懂的东西呢？”本书虽然仍旧有那么一点儿学术，但已经是我写得最通俗的一部了。恐怕今后没有机会再写出类似的读物，因此，我更想借此表达一下自己的感激之情。

首先我要感谢我的哥哥保罗，他开车带我去多伦多，看了电影《2001：太空漫游》的首映。当时我才上高中，也不确定是否真正看懂了，但这部电影确实极大地震撼了我，电影在细节方面的刻画让我至今难以忘怀。当别人都沉醉于电影创造出的科幻世界时，我却被电影中的哈尔电脑所吸引。这部电影改变了我的生活，在本书中我也多次提到了这部电影。

后来，我在高中指导老师的帮助下，获得了工程学的奖学金，并上了大学。但我发现，工程学课程枯燥乏味，虽然要求严格，但是学无所用。唯一的例外就是计算机课。当我得知曾担任电影《2001：太空漫游》顾问的马文·明斯基在麻省理工学院工程系任教后，就给他写了一封信，询问这部电影的有关情况。感谢已故的马文·明斯基教授，当年在百忙之中抽出宝贵时间，给一个一年级本科生回信（当时电子邮件还没有出现），并鼓励我放弃工程学专业，转而投身计算机科学。若是没有我爷爷留给我的学费，我可能也没有能力放弃工程学奖学金。但好在他这样做了，而我最终做了选择，放弃了工程学，而这一选择也改变了我的人生轨迹。

我在计算机科学领域师从约翰·麦洛普洛斯（John Mylopoulos），当时他在多伦多大学任教，对人工智能兴趣浓厚。我上了他主讲的关于人工智能的一门新课，非常喜欢，因此暑期也一直跟随他工作。后来，我还读了他的研究生。虽然我的博士课题与他的研究方向不尽相同，但他依然细致耐心地为我提供帮助。他把我介绍给了来自波士顿的罗恩·布拉赫曼（Ron Brachman）。罗恩为我在波士顿找到了暑期工作的机会，并在我获得博士学位后，为我在加州找到了工作。我们在加州共同度过了三年的美好岁月。约翰四处奔走，促成我返回多伦多从事学术科研工作，而这最终成为我整个职业生涯梦寐以求的工作。约翰和罗恩是我人生中重要的引领者，我的事业能有今天完全得益于他们两位。

后院响起了音乐声，这段文字也该告一段落了。最后，我要感谢我亲爱的妻子帕特（Pat），在我的职业生涯中她始终陪伴于我左右，从青春岁月到日薄桑榆，风风雨雨，不离不弃。正如本书开头所说，谨以此书献给我的妻子。

2016年7月于多伦多

注释

本书主要是我对人工智能与人类心智之间关系的一些看法。但我并不是要进行“学术”讨论，加上引号也不能改变这一事实。在整个写作过程中，我力求清楚明白地表达自己的观点及其深层原因，因此并未太多关注反对意见，也未太在意那些围绕人工智能的争论。我希望读者能够充分利用以下参考文献，并以此为出发点，对本书提出的问题进行更为全面的研究。

第1章 什么是人工智能？

现代人工智能迄今虽然只有60年的历史，但发展过程却充满曲折。帕梅拉·麦考达克（Pamela McCorduck）的著作（*Machines Who Think*）很好地介绍了人工智能的早期发展，但尼尔斯·尼尔森（Nils Nilsson）的著作（*The Quest for Artificial Intelligence:A History of Ideas and Achievements*）涉及的领域更多，而且可以在网上获取。

由于人工智能领域纷繁复杂，想通过一本书就解决全部问题似乎不太可能。但是斯图尔特·拉塞尔（Stuart Russell）和彼得·诺维格（Peter Norvig）的著作（*Artificial Intelligence: A Modern Approach*）却恰恰做到了这一点。该书介绍了所有主要技术的发展方向，还涵盖了历史和哲学背景知识，仅参考书目就有30页之多，而且字体很小，确实物有所值。但是这本书内容技术性较强，适合高年级的大学生阅读。我自己使用的课本（*Thinking as Computation: A First Course*）内容较为通俗，其中涉及人工智能的内容更适合大众阅读。

据我所知，关于人工智能技术的最新趋势这一话题，目前还没有什么好书值得推荐，而且技术发展日新月异，对于读者来说，最好的办法是翻阅新近出版的报纸、杂志，寻找相关文章，有些文章直接在网上就能找到，比如关于丰田投资的消息，关于OpenAI项目等内容。至于我在书中提到的自适应机器学习的技术前景，我建议上网搜索“无监督学习”“深度学习”或者“人工智能技术”。［另一种办法是网上搜索该领域的代表人物，比如我的同事杰弗里·欣顿（Geoffrey Hinton）］。想要进一步了解机器学习，*Machine Learning: A Probabilistic Perspective*是非常优秀（但也颇有难度）的教材。文章*Building high-level features using large scale unsupervised learning*中则有在图像中对猫进行识别的介绍。

老式人工智能（GOFAI）这一术语出自哲学家约翰·海于格兰（John Haugeland）的*Artificial Intelligence:The Very Idea*一书。约翰·麦卡锡（John McCarthy）写于1958年的经典论文*Program with comman sense*可以在网上找到。*Readings in Knowledge Representation,Formalizing Common Sense: Papers by John McCarthy*，*Semantic Information Processing*和*Redings in Artificial Intelligence*等书中不但收录了这篇论

文，而且还有其他不少影响深远的文章。

图灵测试最早是图灵在*Computing machinery and intelligence*一文中提出的，从那以后，一直饱受争议。我知道的所有与人工智能相关的教材，包括*The Turing Test: Verbal Behavior as the Hallmark of Intelligence*等书都谈到这一问题。在*Minds, brains and programs* 一文中可以找到更多关于中文房间这一问题的讨论，其中收录了众多评论者的看法。我未对其进行反驳，而是在*Is it enough to get the behavior right?*中提出了总和房间理论（Summation Room）。

第2章 巨型拼图之谜

人类思维确实是一个令人神往的话题，有这么多人围绕这一话题从不同角度著书立说也就不足为奇了。我发现有三本相对通俗的著作颇具启发意义，它们是史蒂文·平克（Steven Pinker）的*How the Mind Works*，丹尼尔·卡尼曼（Daniel Kahneman）的*Thinking, Fast and Slow*及丹尼尔·丹尼特（Daniel Dennett）的*Kinds of Minds: Towards an Understanding of Consciousness*。其他内容独特的大众读物有杰里·福多尔（Jerry Fodor）的*The Mind Doesn' t Work That Way*（主要是反驳平克），马文·明斯基（Marvin Minsky）的*The Society of Mind*（对思维提出了独到的见解），还有记者马尔科姆·格拉德威尔（Malcolm Gladwell）的*The Power of Thinking without Thinking*（与卡尼曼的观点有相似之处）。

大众媒体对研究大脑的热情当然也很高。*Mapping the Mind*讨论了如何通过脑成像技术（如功能性磁共振成像）认识大脑。*The Brain That Changes Itself: Stories of Personal Triumph from the Frontiers of Brain Science*主要讲了大脑的适应性。论文*Distributed representations* 和*Analysis of distributed representation of constituent structure in connectionist systems*介绍了分布式神经表征（这也是我认为对神经元进行逆向工程非常困难的论据之一）。*The hippocampus and space: are there"place neurons"in the monkey hippocampus*?则对此提出了反驳，认为某些神经元可以代表单一的事物（即出现所谓"地点神经元"时，表示一个位置）。

The Red Queen：*Sex and the Evolution of Human Nature*对人类行为的其他方面进行了描述，并令人信服地展示了其进化论基础，*The Selfish Gene*则从遗传基因角度对进化论进行了讨论。*The Symbolic Species: The Co-Evolution of Language and the Brain*讨论了人类的语言和符号行为的基础，*The Language Instinct* 广泛深入地讨论了人类语言。*The First Word: The Search for the Origins of Language*则介绍了语言本身的发展演变。

最后，你可以在*Brainstorms: Philosophical Essays on Mind and*

*Psychology*中找到丹尼尔·丹尼特首次提出的设计立场这一概念。

第3章 知识与行为

从古希腊至今，知识一直都是哲学研究的主要议题。*Knowledge and Belief*和*Knowing: Essays in the Analysis of Knowledge*收录了这一领域的相关论文。亚科·欣蒂卡（Jaakko Hintikka）所著的*Knowledge and Belief*一书对于知识进行了数学分析，*Reasoning About Knowledge*提出了知识在计算机科学中的进一步应用。

关于知识与信念之间的关系，柏拉图的经典理论认为，知识就是真实合理的信念（即依靠正当理由支持其存在），但埃德蒙·盖蒂尔（Edmund Gettier）在一篇著名论文（*Is justified true belief knowledge*?）当中对这一观点进行了反驳。对于更为广义的命题态度，哲学文献主要关注的是涉及这些态度的句子如何不按惯例遵循逻辑规则，参见*Quantifiers and propositional attitudes*中的例子。

哲学家和心理学家有时还会将隐性信念与显性信念区别开来（见*A theory of implicit and explicit knowledge*）。在本章举出的例子中，亨利发现钥匙不在自己口袋里时，虽然在他的显性信念里不知道钥匙放在冰箱上，但他的隐性信念认为钥匙就在那里，从某种意义上看，他所思考的世界就是钥匙所在之处。换句话说，出现这种情况就是因为在他的信念里，钥匙就在冰箱上，虽然他自己并没有意识到这一点。在*A logic of implicit and explicit belief, A framework for logics of explicit belief*及*Belief, awareness and limited reasoning*几篇文章中，你可以看到如何从数学的角度理解这两个概念。

弗雷德里克·巴特利特（Frederic Bartlett）关于思维的观点引自*Thinking: An Experimental and Social Study*，该书作者早期另外一部颇具影响的作品是*A Study in Experimental and Social Psychology*。泽农·派利夏恩（Zenon Pylyshyn）在*Computation and Cognition: Toward a Foundation for Cognitive Science*这部内容精彩的作品中探讨了认知渗透性等众多话题。丹尼尔·丹尼特在*The Intentional Stance*一书（在*Behavioral and Brain Sciences*杂志1988年第11期中有带注释的缩略版）中提出了意向立场这一概念。尼古拉斯·汉弗雷（Nicholas Humphrey）的说法引自*The social function of intellect*一文。

最后，诺姆·乔姆斯基在*Aspects of the Theory of Syntax*一书中介绍了如何对能力和表现进行区分。

第4章 真智能还是假智能？

本章的主题之一就是，智力并不等同于能够忽悠别人（例如在模仿游戏当中）。但是，纯粹从进化论的角度来看，这二者的区别并不是非常明显。一个很好的例子是人类的智慧，尤其语言，其早期的作用就是在智力竞赛中赢得配偶，而在这类竞赛当中起着核心作用的就是炫耀和诓骗。语言的主要功能是道听途说、说长道短、胡吹神侃、装腔作势和积怨记仇。余下时间（称得上是美好时光）里语言给我们带来的些许愉悦不过是小打小闹而已，可参见*Grooming, Gossip and the Evolution of Language*和*The Red Queen: Sex and the Evolution of Human Nature*。

约瑟夫·魏岑鲍姆（Joseph Weizenbaum）编写的ELIZA程序在期刊*Communications of the ACM*1966年第9期的ELIZA一文中有介绍。而与奥利弗·米勒（Oliver Miller）的谈话选自一篇博文（参见：http://thoughtcatalog.com/?s=eliza）。魏岑鲍姆在自己的*Computer Power and Human Reason: From Judgment to Calculation*一书中谈到，当发现有人将他的成果用于探索心理学（如*Artificial paranoia*这篇论文）时，他感觉心灰意冷。布莱恩·克里斯蒂安（Brian Christian）在*The Most Human Human: What Artificial Intelligence Teaches Us About Being Alive*一书中介绍洛伯纳人工智能竞赛，他本人还在其中一场比赛中扮演了人类的角色。尤金·古斯特曼（Eugene Goostman）程序在很多在线新闻文章中都有描述，例如《新科学家》（*New Scientist*，2012年6月25日）和《连线》杂志（*Wired*，2014年6月9日）。斯科特·阿伦森（Scott Aaronson）的采访摘自网页http://www.scottaaronson.com/blog/?p=1858。

我举的鳄鱼和棒球的例子最早出现于*Logic and the complexity of reasoning* 一文（发表于1988年），后来在*On our best behavior*一文中再次使用过。*Reasoning from imcomplete knowledge*和*On closed world data bases*中有对封闭世界假设的解释。威诺格拉德模式由我在*The Winograd Schema Challenge*中提出， *The Winograd Schema Challenge*中有更详细的介绍。首个示例模式（由特里·威诺格拉德提出）出现在*Understanding Natural Language*一书中。模式的来源广泛，五花八门，在网页https://www.cs.nyu.edu/davise/papers/WS.html上可以找到这些问题的合集，涵盖了100多个问题。2014年7月，纽昂斯通信公司（Nuance

Communications）宣布将于2016年7月举办威诺格拉德模式赛。（更多信息，请参考http://commonsensereasoning.org/winograd.html。）

与之类似的另外一种测试是文本蕴含挑战赛（参见*The PASCAL recognising textual entailment challenge*）。

第5章 有经验的学习与没有经验的学习

我们的绝大多数知识都习自间接经验，这一事实让哲学家们暂时停止了关于智能离开身体是否可能的争论。例如，如果这种智能无法将语言与真实感觉联系起来，那它如何能够真正地理解“饿”这种词呢？这就是所谓符号接地问题（*The symbol grounding problem*）。如果我们知道的只是一个个相互指代的单词，那么怎样才能理解一个词的含义呢？（这个关于“文字游戏”的问题会在下一章介绍海伦·凯勒时出现。）当然，艾伦·图灵的想法是，我们对于这样的问题根本不必理会。“理解”这个词太过含糊，而且存在争议。恰恰相反，我们该问的问题是：有没有可能让人工智能程序拥有与人一样的能力，按照上述单词的意思行事。简而言之，就是我们期望通过符号接地问题看到什么样的蹩脚行为？

我们是如何掌握语言的，这依然是个谜。具体来说，我们如何解释儿童虽然获得的数据有限，但是他们却能掌握复杂的语法？这个问题被诺姆·乔姆斯基称为“刺激贫乏”（*Rules and Representations*）。他的（充满争议的）观点是，孩子天生就具有一种被称为“普遍语法”的能力，帮助他们迅速掌握首次接触的语言的实际语法。

关于行为的学习，艾伦·麦克沃思（Alan Mackworth）（可参见他的文章*On seeing robots*）和罗德尼·布鲁克斯（Rodney Brooks）（可参见他的文章*Elephants don’t play chess*）等对老式人工智能持批判态度的学者更为关注动物（包括人在内）如何能够在没有语言和符号的帮助下，依靠真实传感器和效应器生存，这一点非常有趣。这些动物似乎掌握了某种程序性知识（通过某种作业形式获得的知识），与老式人工智能所关注的陈述性知识（可用陈述句表达的知识）截然不同。但是，我们仍需谨记“巨型拼图”问题，而且“世间行动”是个广义的范畴。当然包括骑自行车和玩溜溜球，也包括饲养照顾宠物金丝雀，收集稀有金币。显而易见，上述两种知识都必不可少。关于这个问题可以参考*Frame representations and the declarative/procedural controversy*，*Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory*一文则从神经科学的角度做了介绍。

塞缪尔·早川关于阅读的说法引自*Language in Thought and Action*。（早川还曾对通过经验学习和通过语言学习进行过对比。）艾萨克·牛顿的话引自1676年的一封信（参见他的书信集*The Correspondence of Isaac Newton：1661—1675*）。

第6章 书本智慧与市井智慧

这一章讲的是我们使用语言文本学习的重要性——虽然我们对于“书本知识”经常不屑一顾，甚至有点儿不以为然，但人类与其他动物相比确实是鹤立鸡群，它们根本不具备这种能力，而这也极大限制了它们在技术上的发展。关于动物和它们所掌握的技术，请参考*Animal Tool Behavior: The Use and Manufacture of Tools by Animals*。想进一步了解蚂蚁的超级殖民地（侵略型殖民地及不侵略型殖民地），可参见文章*Evolution of supercolonies: The Argentine ants of southern Europe*。

有关幼儿使用语言处理语言和其他问题的信息，参见*The Child’s Conception of Language*。唐·佩利（Don Perlis）认为这是语言会话能力的核心（参见*Conversational adequacy: mistakes are the essence*）。

海伦·凯勒的生平故事可以在*The Story of My Life* 一书中找到。第一部分和第二部分均由凯勒自己完成。第三部分摘自安妮·沙利文的书信和报道（本章中的书信也是引自此处）。威廉·拉帕波特对凯勒与人工智能相关性的分析出自*How Helen Keller used syntactic semantics to escape from a Chinese Room*一文。我坚信，我们从她身上学到了许多关于人类思维和人类精神的内容。

第7章 长尾理论与培训的界限

本章的主题之一是人们如何利用常识处理之前未曾经历过的全新的情况。但纳西姆·塔勒布认为大多数人，尤其是投资者非常不善于处理这些“黑天鹅”（见*The Black Swan: The Impact of the Highly Improbable*）。这里并不矛盾，常识就是处理新出现的情况，但投资者的任务更为棘手，需要分析评估所有可能发生的情况。在决定购买某只股票前，投资者必须考虑到所有可能导致股票下跌的因素。塔勒布认为，人们并不善于把握黑天鹅。例如，在文中提到的英国国家语料库中，有些词语出现的概率只有千万分之一，似乎可以忽略不计，但实际上这种词语非常多，如果忽略就会出大问题。英国国家语料库的信息可以在网上查到（见http://www.natcorp.ox.ac.uk/）。厄尼·戴维斯（Ernie Davis）对该语料库中大量存在的罕见词进行了观察，更多细节可以参考他的教材*Linear Algebra and Probability for Computer Science Applications*的第274页。

20世纪70年代，人工智能研究重点专注于技术和专业技能。主要想法是与专家进行访谈对话，利用条件规则复制他们在专业领域的知识，构建所谓专家系统，参见*Introduction to Expert Systems*和*Building Expert Systems*的例子。论文*Skill in chess*和*The expert mind*提到了国际象棋专家。休伯特·德雷福斯和图尔特·德雷福斯（Stuart Dreyfus）共同完成了国际象棋专家在下国际象棋的同时，还要进行复杂运算这一实验（参见*Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*）。他们是如是说：

> 我们最近做了个实验，要求国际象棋大师胡利奥·卡普兰（Julio Kaplan）与另外一位水平稍差但也是大师级的棋手下5秒一步的国际象棋，同时卡普兰还要以每秒一个数字的速度将听到的数字加在一起。即使分析思维完全被数字所占据，卡普兰依然能够棋高一着，多次击败对手。

*What Computers Still Can’t Do: A Critique of Artificial Reason*一书介

绍了他们对人工智能、专家系统的看法，还从哲学角度对专家和新手进行了观察分析。

第8章 符号与符号处理

这一章名为“符号与符号处理”，实际是介绍计算机科学。计算机科学始于艾伦·图灵的图灵机研究（见*On computable numbers, with an application to the Entscheidungsproblem*）。有些学者推崇非符号形式的人工智能（例如文章*Connectionist AI, symbolic AI, and the brain*），但他们实际谈论的依然是符号处理，只不过是在用数字代表符号（如符号代数中的例子），而非使用非数字概念（如符号逻辑中的例子）。

这两个例子谈到的问题是计算机科学的核心，且非常有趣。一道题可以有多种解答方法，并且属性各不相同，计算机科学家花了大量时间和精力研究算法，即题的各种解答方法（参见*Algorithmics: The Spirit of Computing*）。

在符号代数中，求解方程组的标准算法是高斯消元（Gaussian Elimination）（参见*Schaum’s outline of theory and problems of linear algebra*）。其属性是，对于具有n个变量的n个方程，通过大约n^3步可以求得一个解。这就是说，即使方程组里有成百上千万个变量，依然能够解决。

但在符号逻辑当中，目前为止最好的算法可能就是DPLL（数字锁相环，可参见文章*Satisfiability solvers*的例子）。逻辑问题存在有n个变量，求解大约需要2^n步*The intractability of resolution*。（而证明这点则需要根据消解规则，对DPLL进行变体（*A machine-oriented logic based on the resolution principle*），文中已经提到）。这就是说，即使是最快的计算机，对于只有100个变量的逻辑问题也无能为力。

这又引出了两个问题。首先，我们在想是否会有比DPLL更好的算法。在数学领域，这个问题的精确版就是著名的*P=NP*问题，这个问题由斯蒂芬·库克（Stephen Cook）于20世纪70年代首次提出（参见*The complexity of theorem-proving procedures*）。从那时起到现在，虽然已有成千上万名计算机科学家和数学家竭尽全力，试图攻克难题，但是依然没有答案。由于这个问题与其他众多题目息息相关，因此被视为计算机科学领域最为重要的开放问题（参见文章*The status of the P versus NP*

problem）。

第二个问题涉及前一章提到的长尾现象。DPLL的工作方式是系统性地搜索逻辑上的所有可能。有趣的是，在随机构建的测试当中，DPLL执行此类操作所需的步骤却都很少。实际上，上述测试所需的步骤与上一章中的长尾数字非常相似。因为样本测试案例越多，平均值就越高，所以在实践中，根本不可能估算出DPLL所要求的步数。有关详情，请参阅*Heavy-tailed phenomena in satisfiability and constraint satisfaction problems*一文。

如果需要其他材料来教孩子系统学习这些程序，请参考*Computational thinking*一文。

第9章 基于知识的系统

*Leibniz: A Biography*是伟大的思想家戈特弗里德·莱布尼茨的传记。但是他的思想大都分散在他旅途中书写的一万多封信里。《哲学百科全书》（*The Encyclopedia of Philosophy*）中的文章对他进行了更为全面的介绍。

查尔斯·达尔文在《物种起源》中谈到了眼睛的进化。他如是说：

> 眼睛通过其无与伦比的结构，可以调整焦距以适应不同距离，可以调节光线强度，还可以校正球面色差，如果要说这种能力是通过自然选择而来，那么我只能说，这种说法简直荒谬至极。

Origins of the modern mind: three stage of the evolution of culture and cognition（带有评论）和*The Symbolic Species:The Co-Evolution of Language and the Brain*讨论了思想本身的进化过程。

知识表示假说引自布莱恩·史密斯（Brian Smith）的博士论文（*Reflection and Semantics in a Procedural Language*）。人们通常会将此观点归功于约翰·麦卡锡，但是显而易见，其他研究人员与他的观点很相似。对于艾伦·纽厄尔（Allen Newell）和赫伯·西蒙（Herb Simon）而言，重点更多放在象征领域，因此他们的版本被称为物理符号系统假说（*physical symbol system hypothesis*），他们认为：“物理符号系统是普遍智能行为的充分必要条件”（见Computer Science as empirical inquiry: Symbols and search）。（马文·明斯基作为早期人工智能的研究人员之一，他发现了逻辑和数值方法对于人工智能都有很大局限，提出应将这些方法进行融合，可参见*Logical versus symbolic versus connectionist or neat versus scruffy*。）

*Knowledge Representation and Reasoning*和*Knowledge Representation, Reasoning and Declarative Problem solving*这两本书专门介绍了人工智能中的知识表示和推理领域。此外，还有每两年举行一次的该领域的学术

会议（见http://kr.org）。可以在*Readings in knowledge Representation*中找到相关的早期内容。*Logical Foundations of Artificial Intelligence*和*Computational Intelligence: A Logical Approach*提出了逻辑是人工智能统一的主题，*The role of logic in knowledge representation and common sense reasoning*和*The role of gic in knowledge representation*对此进行了进一步讨论。马文·明斯基关于逻辑的论述摘自*A framework for representing knowledge* 第262页。*Logic, probability and human reasoning* 是从心理学方面看待推理。关于根据信仰程度不同，用概率的方法进行推理的优势，可参见*Bayesian Rationality: The Probabilistic Approach to Human Reasoning*。

关于建设大型知识库这一问题，CYC是相关的一个长期项目，参见*Building Large Knowledge-Based Sytems: Representation and Inference in the Cyc Project*。（由于无法深入了解，因此目前很难说清CYC取得了何种成果，总体来说不太像可控的科学研究。）斯坦福国际研究院的HALO（见*A question-answering system for AP chemistry: assessing KR&R technologies.*）和AURA（参见*Achieving intelligence using prototypes, composition, and analogy*）、艾伦人工智能研究所的ARISTO（见*Elementary school science and math tests as a driver for AI: Take the Aristo challenge*!）是其他相关且更为专业的工作。关于从网络自动从文本中提取知识的前景，请参阅*Open information extraction from the web*一文的回顾。

最后，在网页（https://sites.google.com/site/krr2015/）上可以看到，该研讨会对调和GOFAI中的逻辑方法与AML中的统计方法进行了新的尝试。

第10章 人工智能技术应用

关于人工智能的未来，本章涉及的仅仅是几个棘手的问题。*Artificial Intelligence and the End of the Human Era*一书对于这一话题有更为全面的解读。雷蒙德·库兹韦尔（Raymond Kurzweil）提出了技术奇点的想法（见*The Singularity Is Near: When Humans Transcend Biology*），并在*The Technological Singularity*一书中有进一步的讨论。斯蒂芬·霍金（2014年12月2日，网址http://www.bbc.com/news/）、埃隆·马斯克（2014年10月8日，网址http: // www.vanityfair.com/news/）和比尔·盖茨（2015年1月28日，网址https: //www.reddit.com/r/IAmA/comments/）在访谈中均讲到了人工智能的潜在危险。

科幻小说家阿瑟·C.克拉克（他与斯坦利·库布里克共同完成了电影《2001：太空漫游》的剧本）在*Profiles of the Future: An Enquiry into the Limits of the Possible*一书中提出了三大定律，具体内容如下：

1. 如果一位德高望重的科学家说，某件事是有可能的，那么他几乎肯定是正确的。如果他说某件事是不可能的，那么他或许已经谬之千里。

2. 想要找到可能性的极限，唯一的办法就是超越可能，实现不可能。

3. 任何先进技术都与魔法无二。

科幻小说作家艾萨克·阿西莫夫提出的机器人三大定律最早出现于*I，Robot*一书中（据说是引自将于2058年出版的机器人手册），其内容如下：

1.机器人不能伤人，也不能在人受到伤害时无动于衷。

2.机器人必须服从人的命令，除非该命令与第一定律冲突。

3.在不违反第一、第二定律的前提下，机器人须尽力保护自身

安全。

阿西莫夫后来写的故事都是围绕遵循这些规律的机器人会出现问题展开的。

马文·明斯基在*HAL's Legacy: 2001's Computer as Dream and Reality*里的一段采访中谈到了自己参与电影《2001:太空漫游》的情况。拉杰·雷迪（Raj Reddy）在2000年一次讲座中引用了约翰·麦卡锡，具体内容可以访问网页http://www.rr.cs.cmu.edu/InfiniteMB.doc阅读。玛格丽特·撒切尔的话引自《妇女界》杂志（*Woman's Own*）1987年9月23日的采访。*Faster Than Thought:A Symposium on Digital Computing Machines*介绍了艾伦·图灵早期涉足国际象棋的经历。深蓝的故事在*Deep Blue: An Aritificial Intelligence Milestone*一书中有介绍。关于凯瑟琳·赫本在电影中引用的名言和人类超越进化的观点，请参见*The Blank Slate*。

参考文献

[1] Eric Aiton. *Leibniz: A Biography*. Boston: Adam Hilger, 1985.

[2] Isaac Asimov. *I, Robot*. New York: Gnome Press, 1950.

[3] Chitta Baral. *Knowledge Representation, Reasoning and Declarative Problem Solving*. Cambridge, UK: Cambridge University Press, 2003.

[4] Ken Barker, Vinay Chaudhri, Shaw Chaw, Peter Clark, James Fan, David Israel, Sunil Mishra, Bruce Porter, Pedro Romero, Dan Tecuci, and Peter Yeh. A question-answering system for AP chemistry: assessing KR&R technologies. In *Proceedings of KR-2004: The Ninth International Conference on Principles of Knowledge Representation and Reasoning*, Whistler, Canada, June 2004, 488–497.

[5] James Barrat. *Our Final Invention: Artificial Intelligence and the End of the Human Era*. New York: Thomas Dunne Books, 2013.

[6] Frederic Bartlett. *Remembering: A Study in Experimental and Social Psychology*. Cambridge, UK: Cambridge University Press, 1932.

[7] Frederic Bartlett. *Thinking: An Experimental and Social Study*. London: Allen and Unwin, 1958.

[8] Bertram Bowden, ed. *Faster Than Thought: A Symposium on Digital Computing Machines*. London: Sir Isaac Pitman & Sons, 1953.

[9] Ronald J. Brachman and Hector J. Levesque, eds. *Readings in Knowledge Representation*. San Francisco: Morgan Kaufmann, 1985.

[10] Ronald J. Brachman and Hector J. Levesque. *Knowledge Representation and Reasoning*. San Francisco: Morgan Kaufmann, 2004.

[11] Rodney Brooks. Elephants don't play chess. *Robotics and Autonomous Systems* 6 (1990): 3–15.

[12] Rita Carter. *Mapping the Mind*. Berkeley: University of California Press, 1998.

[13] Vinay Chaudhri. Achieving intelligence using prototypes, composition, and analogy. In *Proceedings of AAAI-2015: The Twenty-Ninth AAAI Conference on Artificial Intelligence*, Austin, Texas, January 2015, 4093–4099.

[14] Noam Chomsky. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press, 1965.

[15] Noam Chomsky. *Rules and Representations*. Oxford: Basil Blackwell, 1980.

[16] Brian Christian. *The Most Human Human: What Artificial Intelligence Teaches Us About Being Alive*. New York: Doubleday, 2011.

[17] Peter Clark. Elementary school science and math tests as a driver for AI: Take the Aristo challenge! In *Proceedings of IAAI-2015: The Twenty-Seventh Conference on Innovative Applications of Artificial Intelligence*, Austin, Texas, March 2015, 4019–4021.

[18] Arthur C. Clarke. *Profiles of the Future: An Enquiry into the Limits of the Possible*. London: Gollancz, 1962.

[19] Kenneth Colby, Sylvia Weber, and Franklin Hilf. Artificial paranoia. *Artificial Intelligence* 2 (1971): 1–25.

[20] Allan Collins, Eleanor Warnock, Nelleke Aiello, and Mark Miller. Reasoning from incomplete knowledge. In *Representation and Understanding: Studies in Cognitive Science*, ed. D. Bobrow and A. Collins. New York: Academic Press, 1975, 35–82.

[21] Stephen A. Cook. The complexity of theorem-proving procedures. In *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, Shaker Heights, Ohio, 1971, 151–158.

[22] Ido Dagan, Oren Glickman, and Bernardo Magnini, The PASCAL recognising textual entailment challenge, *Machine Learning Challenges*, Berlin: Springer Verlag, 2006.

[23] Charles Darwin. *On the Origin of Species*. London: John Murray, 1859.

[24] Ernest Davis. *Linear Algebra and Probability for Computer Science Applications*. Boca Raton, FL: CRC Press, 2012.

[25] Richard Dawkins. *The Selfish Gene*. New York: Oxford University Press, 1976.

[26] Terrence Deacon. *The Symbolic Species: The Co-Evolution of Language and the Brain*. New York: W. W. Norton & Co., 1997.

[27] James Delgrande. A framework for logics of explicit belief. *Computational Intelligence* 11 (1985): 47–88.

[28] Daniel Dennett. *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: MIT Press, 1981.

[29] Daniel Dennett. *The Intentional Stance*. Cambridge, MA: MIT Press, 1987.

[30] Daniel Dennett. Précis of *The Intentional Stance*. *Behavioral and Brain Sciences* 11 (1988): 495–505.

[31] Daniel Dennett. *Kinds of Minds: Towards an Understanding of Consciousness*. London: Weidenfeld & Nicolson, 1996.

[32] Zoltan Dienes and Josef Perner. A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences* 22 (1999): 735–755.

[33] Norman Doidge. *The Brain That Changes Itself: Stories of Personal Triumph from the Frontiers of Brain Science*. London: Penguin Books, 2007.

[34] Merlin Donald. Origins of the modern mind: three stages of the evolution of culture and cognition. *Behavioral and Brain Sciences* 16 (1993): 737–791.

[35] Robin Dunbar. *Grooming, Gossip and the Evolution of Language*. London: Faber and Faber, 1996.

[36] Hubert Dreyfus. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press, 1992.

[37] Hubert Dreyfus and Stuart Dreyfus. *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. New York: Free Press, 1986.

[38] Paul Edwards, ed. *The Encyclopedia of Philosophy*. New York: Macmillan, 1967.

[39] Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel Weld. Open information extraction from the web. *Communications of the ACM* 51 (2008): 68–74.

[40] Ronald Fagin and Joseph Halpern. Belief, awareness and limited reasoning. *Artificial Intelligence* 34 (1987): 39–76.

[41] Ronald Fagin, Joseph Halpern, Yoram Moses, and Moshe Vardi. *Reasoning About Knowledge*. Cambridge, MA: MIT Press, 1995.

[42] Jerry Fodor. *The Mind Doesn't Work That Way*. Cambridge, MA: MIT Press, 2000.

[43] Lance Fortnoy. The status of the P versus NP problem. *Communications of the ACM* 52 (2009): 78–86.

[44] Michael Genesereth and Nils Nilsson. *Logical Foundations of Artificial Intelligence*. Los Altos: Morgan Kaufmann, 1987.

[45] Edmund Gettier. Is justified true belief knowledge? *Analysis* 23 (1963): 121–123.

[46] Tatiana Giraud, Jes Pedersen, and Laurent Keller. Evolution of supercolonies: The Argentine ants of southern Europe. *Proceedings of the National Academy of Science* 99 (2002): 6075–6079.

[47] Carla Gomes, Henry Kautz, Ashish Sabharwal, and Bart Selman. Satisfiability solvers. In *Handbook of knowledge representation: Foundations of Artificial Intelligence*, ed. F. van Harmelen, V. Lifschitz, B. Porter. Amsterdam: Elsevier, 2008, 89–134.

[48] Carla Gomes, Bart Selman, Nuno Crato, and Henry Kautz. Heavy-tailed phenomena in satisfiability and constraint satisfaction problems. *Journal of Automated Reasoning* 24 (2000): 67–100.

[49] Malcolm Gladwell. *Blink: The Power of Thinking without Thinking*. New York: Little, Brown and Co., 2005.

[50] Phillips Griffiths, ed. *Knowledge and Belief*. London: Oxford University Press, 1967.

[51] Armin Haken. The intractability of resolution. *Theoretical Computer Science* 39 (1985): 297–308.

[52] David Harel. *Algorithmics: The Spirit of Computing*. Reading, MA, Addison-Wesley, 1987.

[53] Stevan Harnad. The symbol grounding problem. *Physica D* 42 (1990): 335–346.

[54] John Haugeland. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press, 1985.

[55] Samuel Hayakawa and Alan Hayakawa. *Language in Thought and Action*. New York: Harcourt Brace Jovanovich, 1991.

[56] Frederick Hayes-Roth, Donald Waterman, and Douglas Lenat. *Building Expert Systems*. Reading, MA: Addison-Wesley, 1983.

[57] Jaakko Hintikka. *Knowledge and Belief*. Ithaca, NY: Cornell University Press, 1962.

[58] Geoff Hinton, Jay McClelland, and David Rumelhart. Distributed representations. In *Parallel Distributed Processing*, ed. D. Rumelhart and J. McClelland. Cambridge, MA: MIT Press, 1986, 77–109.

[59] Nicholas Humphrey. The social function of intellect. In *Growing Points in Ethology*, ed. P. Bateson and R. Hinde. Cambridge, UK: Cambridge University Press, 1976, 303–317.

[60] David Israel. The role of logic in knowledge representation. *IEEE Computer* 16 (1983): 37–42.

[61] Peter Jackson. *Introduction to Expert Systems*. Reading, MA: Addison-Wesley, 1990.

[62] Philip Johnson-Laird, Sangeet Khemlani, and Geoffrey Goodwin. Logic, probability, and human reasoning. *Trends in Cognitive Science* 19 (2015): 201–214.

[63] Daniel Kahneman. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.

[64] Christine Kenneally. *The First Word: The Search for the Origins of Language*. New York: Penguin Books, 2007.

[65] Raymond Kurzweil. *The Singularity Is Near: When Humans Transcend Biology*. London: Viking Penguin, 2005.

[66] Quoc Le, Marc-Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg Corrado, Jeffrey Dean, and Andrew Ng. Building high-level features using large scale unsupervised learning. In *Proceedings of ICML 2012: The 29th International Conference on Machine Learning*, Edinburgh, Scotland, June 2012, 81–88.

[67] Douglas Lenat and Ramanathan Guha. *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Boston: Addison-Wesley, 1990.

[68] Hector J. Levesque. A logic of implicit and explicit belief. In *Proceedings of AAAI-84: The Fourth National Conference on Artificial Intelligence*, August 1984, 198–202.

[69] Hector J. Levesque. Logic and the complexity of reasoning. *Journal of Philosophical Logic* 17 (1988): 355–389.

[70] Hector J. Levesque. Is it enough to get the behavior right? In *Proceedings of IJCAI-09: The 21st International Joint Conference on Artificial Intelligence*, Pasadena, California, August 2009, 1439–1444.

[71] Hector J. Levesque. The Winograd Schema Challenge. In *Proceedings of Commonsense-11: The Tenth International Symposium on Logical Formalizations of Commonsense Reasoning*, Palo Alto, March 2011, 53–58.

[72] Hector J. Levesque. *Thinking as Computation: A First Course*. Cambridge, MA: MIT Press, 2012.

[73] Hector J. Levesque. On our best behavior. *Artificial Intelligence* 212 (2014): 27–35.

[74] Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The Winograd Schema Challenge. In *Proceedings of KR 2012: The Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, Rome, June 2012, 552–561.

[75] Vladimir Lifschitz, ed. *Formalizing Common Sense: Papers by John McCarthy*. Exeter, UK: Intellect, 1998.

[76] Marc Lipson and Seymour Lipschutz. *Schaum's outline of theory and problems of linear algebra*. New York: McGraw-Hill, 2001.

[77] Alan Mackworth. On seeing robots. In *Computer Vision: Systems, Theory and Applications*, ed. A. Basu and X. Li. Singapore: World Scientific Press, 1993, 1–13.

[78] John Macy, ed. *The Story of My Life*. New York: Doubleday, Page & Co., 1905.

[79] John McCarthy. Programs with common sense. In *Proceedings of Symposium on the Mechanization of Thought Processes*. National Physical Laboratory, Teddington, England, 1958, 77–84.

[80] Pamela McCorduck. *Machines Who Think*. 25th anniversary edition. Natick, MA: A K Peters, 2004.

[81] Marvin Minsky, ed. *Semantic Information Processing*. Cambridge, MA: MIT Press, 1968.

[82] Marvin Minsky. A framework for representing knowledge. In *Readings in Knowledge Representation*, ed. R. Brachman and H. Levesque. San Francisco: Morgan Kaufmann, 1985, 245–262.

[83] Marvin Minsky. *The Society of Mind*. New York: Simon & Schuster, 1986.

[84] Marvin Minsky. Logical versus analogical or symbolic versus connectionist or neat versus scruffy. *AI Magazine* 12 (1991): 34–51.

[85] Robert Moore. The role of logic in knowledge representation and commonsense reasoning. In *Proceedings of AAAI-82: The Second National Conference on Artificial Intelligence*, August 1982, 428–433.

[86] Kevin Murphy. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press, 2012.

[87] Monty Newborn and Monroe Newborn. *Deep Blue: An Artificial Intelligence Milestone*. Berlin: Springer Science & Business Media, 2003.

[88] Allen Newell and Herbert Simon. Computer Science as empirical inquiry: Symbols and search. *Communications of the ACM* 19 (1976): 113–126.

[89] Nils Nilsson. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge, UK: Cambridge University Press, 2009.

[90] Mike Oaksford and Nick Chater. *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. New York: Oxford University Press, 2007.

[91] Taketoshi Ono, Ryoi Tamura, and Kiyomi Nakamura. The hippocampus and space: are there "place neurons" in the monkey hippocampus? *Hippocampus* 1 (1991): 253–257.

[92] Don Perlis, Khemdut Purang, and Carl Andersen. Conversational adequacy: mistakes are the essence. *International Journal of Human-Computer Studies*, 48 (1998): 553–575.

[93] Steven Pinker. *The Language Instinct*. New York: Harper Perennial Modern Classics, 1994.

[94] Steven Pinker. *How the Mind Works*. New York: W. W. Norton, 1999.

[95] Steven Pinker. *The Blank Slate*. New York: Viking, 2002.

[96] David Poole, Alan Mackworth, and Randy Goebel. *Computational Intelligence: A Logical Approach*. New York: Oxford University Press, 1998.

[97] Zenon Pylyshyn. *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, MA: MIT Press, 1984.

[98] Willard Quine. Quantifiers and propositional attitudes. *Journal of Philosophy* 53 (1956): 177–187.

[99] William Rapaport. How Helen Keller used syntactic semantics to escape from a Chinese Room. *Minds & Machines* 16 (2006): 381–436.

[100] Raymond Reiter. On closed world data bases. In *Logic and Databases*, ed. H. Gallaire and J. Minker. New York: Plenum Press, 1987, 55–76.

[101] Matt Ridley. *The Red Queen: Sex and the Evolution of Human Nature*. London: Penguin Books, 1993.

[102] John Robinson. A machine-oriented logic based on the resolution principle. *Journal of the ACM* 12 (1965): 23–41.

[103] Philip Ross. The expert mind. *Scientific American* 295 (2006): 64–71.

[104] Michael Roth and Leon Galis. *Knowing: Essays in the Analysis of Knowledge*. Lanham, MD: University Press of America, 1984.

[105] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ: Pearson Education, 2010.

[106] John Searle. Minds, brains and programs. *Behavioral and Brain Sciences* 3 (1980): 417–424.

[107] Murray Shanahan. *The Technological Singularity*. Cambridge, MA: MIT Press, 2015.

[108] Stuart Shieber, ed. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. Cambridge, MA: MIT Press, 2004.

[109] Robert Shumaker, Kristina Walkup, and Benjamin Beck. *Animal Tool Behavior: The Use and Manufacture of Tools by Animals*. Baltimore, MD: Johns Hopkins University Press, 2011.

[110] Herbert Simon and William Chase. Skill in chess. *American Scientist* 61 (1973): 394–403.

[111] Anne Sinclair, Robert Jarvella, and Willem Levelt, eds. *The Child's Conception of Language*. Berlin: Springer-Verlag, 1978.

[112] Brian Cantwell Smith. *Reflection and Semantics in a Procedural Language*. Cambridge, MA: PhD thesis, Massachusetts Institute of Technology, 1982.

[113] Paul Smolensky. Connectionist AI, symbolic AI, and the brain. *Artificial Intelligence Review* 1 (1987): 95–109.

[114] Paul Smolensky. Analysis of distributed representation of constituent structure in connectionist systems. In *Proceedings of NIPS-87: Neural Information Processing Systems*. Denver, Colorado, November 1988, 730–739.

[115] Larry Squire. Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. *Journal of Cognitive Neuroscience* 4 (1992): 232–246.

[116] David Stork, ed. *HAL's Legacy: 2001's Computer as Dream and Reality*. Cambridge, MA: MIT Press, 1997.

[117] Nassim Taleb. *The Black Swan: The Impact of the Highly Improbable*. New York: Random House, 2007.

[118] Alan Turing. On computable numbers, with an application to the *Entscheidungsproblem*. *Proceedings of the London Mathematical Society* 42 (1937): 230–265.

[119] Alan Turing. Computing machinery and intelligence. *Mind* 59 (1950): 433–460.

[120] Herbert Turnbull. *The Correspondence of Isaac Newton: 1661–1675*. Volume 1. London: Cambridge University Press, 1959.

[121] Bonnie Webber and Nils Nilsson, eds. *Readings in Artificial Intelligence*. Los Altos, CA: Morgan Kaufmann, 1981.

[122] Joseph Weizenbaum. ELIZA. *Communications of the ACM* 9 (1966): 36–45.

[123] Joseph Weizenbaum. *Computer Power and Human Reason: From Judgment to Calculation*. New York: W. H. Freeman & Co., 1976.

[124] Jeanette Wing. Computational thinking. *Communications of the ACM* 49 (2006): 33–45.

[125] Terry Winograd. *Understanding Natural Language*. New York: Academic Press, 1972.

[126] Terry Winograd. Frame representations and the declarative/procedural controversy. In *Representation and Understanding: Studies in Cognitive Science*, ed. D. Bobrow and A. Collins. New York: Academic Press, 1975, 185–210.

图书在版编目（CIP）数据

人工智能的进化/（加）赫克托·莱韦斯克著；王佩译. -- 北京 : 中信出版社，2018.4
书名原文：Common Sense, the Turing Test, and the Quest for Real AI
ISBN 978-7-5086-8567-0
Ⅰ. ①人… Ⅱ. ①赫… ②王… Ⅲ.①人工智能－普及读物 Ⅳ. ①TP18-49
中国版本图书馆CIP数据核字（2018）第018479号

人工智能的进化
著者：［加］ 赫克托·莱韦斯克
译者：王佩
出版发行：中信出版集团股份有限公司
（北京市朝阳区惠新东街甲4号富盛大厦2座 邮编100029）
电子书排版：萌芽图文

中信出版社官网：http://www.citicpub.com/
官方微博：http://weibo.com/citicpub
更多好书，尽在中信书院
中信书院：App下载地址https://book.yunpub.cn/（中信官方数字阅读平台）
微信号：中信书院

Table of Contents