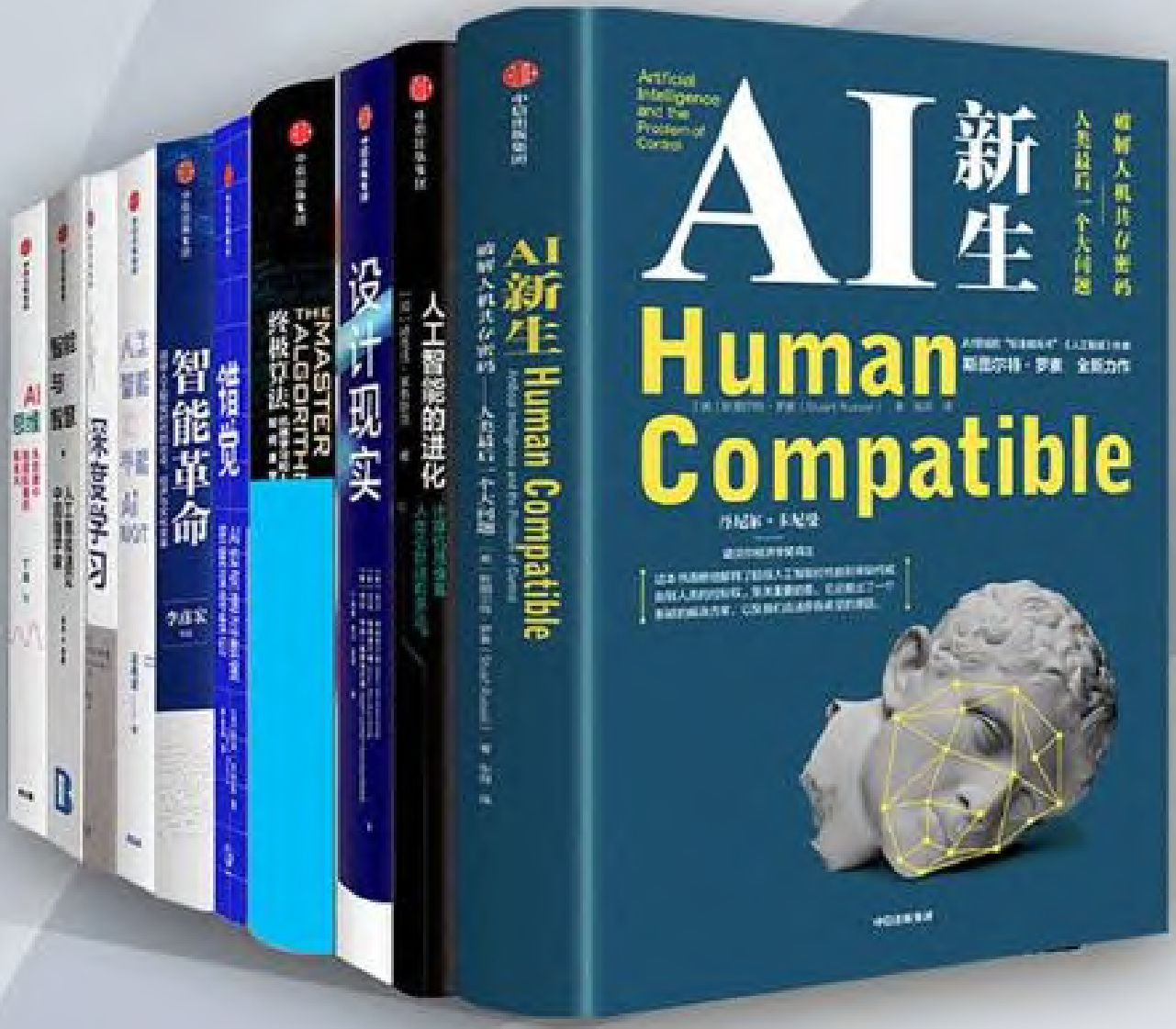
人工智能时代：破解人机共存密码

（套装共10册）



中信出版集团
CITIC Press Group

人工智能时代：破解人机共存密码

（套装共10册）

[美]斯图尔特·罗素 等 著

张羿 等 译

中信出版集团

目录

AI新生：破解人机共存密码：人类最后一个大问题

AI思维：从数据中创造价值的炼金术

错觉：AI如何通过数据挖掘误导我们

人工智能的进化

人工智能之不能

设计现实

深度学习

智能革命：迎接人工智能时代的社会、经济与文化变革

智能与智慧：人工智能遇见中国哲学家

终极算法：机器学习和人工智能如何重塑世界

Artificial Intelligence and the Problem of Control

AI新生

破解人机共存密码

人类最后一个大问题

Human Compatible

AI领域的“标准教科书”《人工智能》作者

斯图尔特·罗素 全新力作

[美] 斯图尔特·罗素 (Stuart Russell) 著 张羽 译

丹尼尔·卡尼曼

诺贝尔经济学奖得主

这本书清晰地解释了超级人工智能时代的到来如何威胁到人类的控制权，至关重要的是，它还提出了一个新颖的解决方案，以及我们应该存有希望的原因。



中信出版集团

AI新生：破解人机共存密码

——人类最后一个大问题

[美]斯图尔特·罗素 著

张羿 译

中信出版集团

目录

专家推荐

前言

01 如果我们成功了

02 人类和机器的智能

03 人工智能在未来会如何发展?

04 人工智能的滥用

05 过于智能的人工智能

06 不那么伟大的人工智能辩论

07 人工智能：一种不同的方法

08 可证明有益的人工智能

09 复杂的难题：我们

10 问题解决了吗?

附录A 寻找解决方案

附录B 知识与逻辑

附录C 不确定性和概率论

附录D 从经验中学习

致谢

图片版权

献给洛伊、戈登、露西、乔治和艾萨克

专家推荐

这是我在相当长的一段时间以来，读到的非常重要的一本书。它清晰地解释了超级人工智能时代的到来如何威胁到人类的控制权。至关重要的是，这本书还提出了一个新颖的解决方案，以及我们应该存有希望的原因。

丹尼尔·卡尼曼（Daniel Kahneman）

诺贝尔经济学奖得主，《思考，快与慢》作者

这是一场由人工智能先驱罗素带来的智识之旅。罗素不仅以引人入胜且有说服力的方式解释了人工智能带来的风险，而且给出了一种切实可行的解决方案。

迈克斯·泰格马克（Max Tegmark）

《生命3.0》作者

与那些未来学家和警告人工智能风险的人不同，罗素是一位人工智能领域的专家。这本书比其他我所知的书更能教导公众，而且这本书读起来令人愉快、振奋人心。

朱迪亚·珀尔（Judea Pearl）

贝叶斯网络之父，《为什么》作者

这本文笔优美的书解决了人类面临的一项基础性挑战：越来越智能的机器会按照我们的要求行事，却不会按照我们的本意行事。如果你关心我们的未来，请务必阅读。

约书亚·本吉奥（Yoshua Bengio）

2019年图灵奖得主，《深度学习》作者

这本关于人工智能未来风险和解决方案的书值得一读，作者罗素很棒！

埃隆·马斯克（Elon Musk）

特斯拉创始人兼CEO

人工智能到底发展到了哪一步，未来会怎么发展，这本书代表主流专家意见，能让你迅速获得“当前科学理解”。作者是研究人工智能的教授，但书中并不过度讨论技术，他关心更大的问题：人的偏好可以“算法化”吗？人工智能是人类的威胁吗？人能限制人工智能吗？当前的答案并不乐观。到最后，你会发现人工智能的问题首先是“人”的问题：我们其实并不了解我们自己。

万维钢

科学作家，“得到”App《精英日课》专栏作者

这是一本引人入胜的巨作，无论一般读者还是人工智能专家都将从中获得启发。罗素教授对于超级智能这个人工智能的重大问题做出了深刻明确的分析。更重要的是，他提出了一个新颖的人机关系处理方案，为解决超级智能这个重要问题开辟了一个崭新的研究方向。

姚期智

中国科学院院士，2000年图灵奖得主

当人们为人工智能技术的突飞猛进而欢呼雀跃，并期盼着超级人工智能到来时，我们不能忘了人类正面临着一个非常关键的挑战：如何确保人工智能的发展以人为本，基于人类的共同价值造福人类？这本书为我们带来了罗素教授多年潜心的研究成果和深刻洞见，读起来令人赏心悦目。

张宏江

智源研究院理事长，源码资本投资合伙人

人类可能在不久的将来制造出超越自己的强人工智能，我很高兴人工智能知名学者斯图尔特·罗素教授不仅仔细推敲了这种可能性，还提出了一种巧妙的应对之策。

黄铁军

智源研究院院长，北京大学教授

创造出达到或超越人类水平的人工智能会怎么样？超级人工智能的成功为什么可能是人类历史上的最后一个事件？人类是否还有希望？人工智能经典教科书作者斯图尔特·罗素的这本书值得每位关心人工智能的人士阅读。

周志华

南京大学人工智能学院院长欧洲科学院外籍院士

自我们这一代人开始，科学地理解、探索、协同和管控超级智能机器，将变成一个越来越不能回避的严肃课题。如果说罗素的经典教科书在过去帮助很多学者进入了人工智能研究领域，他的这本新书则将让更多人思考如何创造更好的人机协同的未来。

周伯文

京东集团副总裁

书中倡议的构建“尽可能安全且对人类有益的人工智能”应成为全球人工智能研究与实践者的共同愿景。罗素论述的人工智能安全风险与人类自身的隐忧无疑将“确保人工智能的安全有益”这个重要挑战长期置于未来人工智能研究的核心。

曾毅

中国科学院自动化研究所研究员，类脑智能研究中心副主任

智源研究院人工智能伦理与可持续发展研究中心主任

前言

为什么读这本书，为什么现在读？

这本书讲述了我们为理解和创造智能，在过去、现在和未来所做的尝试。这很重要，并不是因为人工智能正迅速成为当前的一种普遍现象，而是因为它是未来的主导技术。世界上的各个大国正在意识到这一事实，各家大公司也早就知道这一点。我们无法准确预测这项技术将如何发展，也不知道这项技术发展的时间线。然而，我们必须为现实世界中“机器的决策能力远远超过人类”这种可能性做好准备。

文明所提供的一切都是我们智能的产物，掌控比我们强大得多的智能将是人类历史上最大的事件。这本书的目的是解释为什么这可能是人类历史上的最后一件事，以及如何确保这不会成为现实。

全书概述

本书分为三个部分。第一部分（第1章到第3章）探讨了人类和机器中智能的概念。阅读这部分不需要技术背景，但对于感兴趣的人，书后补充了4个附录，解释了当今人工智能系统的一些核心概念。第二部分（第4章到第6章）讨论了给机器赋予智能所产生的一些问题。我特别关注控制问题：对于比我们更强大的机器，我们要保持绝对的控制权。第三部分（第7章到第10章）提出了一种思考人工智能的新方法，并确保机器永远对人类有益。虽然这本书是面向普通读者的，但我希望它也能说服人工智能专家重新思考他们的基本假设。


01
如果
我们成功了



很久以前，我父母住在英国伯明翰大学附近。后来他们决定搬出这座城市，并把房子卖给了英语文学教授戴维·洛奇（David Lodge）。那时候，洛奇已经是一位著名小说家了。我从未见过他，但我决定读读他的书：《换位》和《小世界》。书中的主要人物都是学者，他们从伯明翰搬到了加州伯克利。而我正是刚刚从伯明翰搬到伯克利的学者，这一系列的巧合似乎在提醒我要多多注意。

《小世界》中有一个特别的场景令我印象深刻。主人公是一位有抱负的文学理论家，他参加了一个重要的国际会议，并问一群领军人物："如果所有人都赞同你们的观点，接下来会发生什么？"这个问题引起了恐慌，因为专家组成员更关心的是智力竞赛，而不是弄清真相、获得认同。当时我想到，我们可以对人工智能领域的领军人物提出一个类似的问题："如果你们成功了会如何？"该领域的目标一直是创造出达到或超越人类水平的人工智能，但人们很少或根本没有考虑过，如果我们创造出了这样的人工智能会发生什么。

几年后，我和彼得·诺维格（Peter Norvig）开始编写一本新的人工智能教科书，书的第一版于1995年出版。[1]这本《人工智能》的最后一节题为"如果我们成功了会如何"，它指出了出现好结果和坏结果的可能性，但没有得出明确的结论。到2010年第三版出版的时候，许多人终于开始考虑，创造超越人类的人工智能可能不是一件好事，但这些人大多是局外人，而不是主流的人工智能研究人员。到了2013年，我开始相信这个问题不仅是主流的人工智能研究人员应该考虑的问题，而且可能是人类面临的最重要的问题。

2013年11月，我在伦敦南部一座久负盛名的艺术博物馆——杜尔维奇绘画馆演讲。听众大多是退休人员，他们不是科学家，但是普遍对智能问题感兴趣，所以我只好做一次完全非技术性的演讲。这似乎是一个在公共场合验证我的想法的合适场所。在解释了人工智能是什么之后，我提出了5个"人类未来最大事件"的候选选项：

1. 我们灭亡了（因为小行星撞击、气候灾难、流行病等）。
2. 我们都能永生（医学攻克了衰老）。
3. 我们发明了超光速旅行，征服了宇宙。
4. 高级外星文明造访了我们。
5. 我们发明了超级人工智能。

我认为第5个候选选项，即我们发明了超级人工智能，将会胜出，因为如果可能，超级人工智能将帮助我们避免灾难，实现永生和超光速旅行。它将代表我们文明的一次巨大飞跃，或者说一次突变。超级人工智能的到来在许多方面类似于高级外星文明的到来，但前者更有可能发生。也许最重要的是，人工智能不同于外星人，它是我们对其有发言权的东西。

然后，我让听众想象一下，如果我们收到一个来自高级外星文明的通知，说它们将在30到50年后到达地球，会发生什么。“乱作一团”这个词并不能准确地描述这种情景。然而，我们对超级人工智能即将到来的反应却是……“冷淡”。（在后面的演讲中，我用图1所示的电子邮件交流的形式对此进行了说明。）最后，我将超级人工智能的意义解释为：超级人工智能的成功将是人类历史上最重大的事件……或许是人类历史上的最后一个事件。



图1 高级外星文明第一次联络人类的电子邮件内容，如有雷同，纯属巧合

在那之后几个月，也就是2014年4月，我在冰岛参加一个会议，美国国家公共广播电台打来电话，问我是否可以接受他们的采访，谈谈刚刚在美国上映的电影《超验骇客》。虽然我已经看过了剧情概要和影评，但我还没有看过电影，因为我当时住在巴黎，电影要到6月才会在巴黎上映。然而碰巧的是，我在从冰岛回家的路上绕道去了波士顿，参加美国国防部的一个会议。所以，到达波士顿洛根机场后，我乘出租车去最近的电影院看了这部电影。我坐在第二排，看到约翰尼·德普饰演的加州大学伯克利分校人工智能教授遭到了对超级人工智能存有担忧的“反人工智能激进分子”的枪击。我不由自主地瘫倒在座位上。（又是巧合在警示我吗？）在这名教授死去之前，他的意识被上传到一台量子超级计算机上，这台计算机迅速超越了人类的能力，威胁要接管世界。

2014年4月19日，我和物理学家迈克斯·泰格马克（Max Tegmark）、弗兰克·威尔切克（Frank Wilczek）、史蒂芬·霍金（Stephen Hawking）合写的关于《超验骇客》的影评被发表在了《赫芬顿邮报》上。其中包括我在杜尔维奇绘画馆的演讲中谈到的人类历史上最重大的事件。从那时起，我便公开承认，我的研究领域对我自己所属的物种构成了潜在危险。

我们如何走到今天？

人工智能的起源可以追溯到很久以前，但它正式启动是在1956年。两位年轻的数学家约翰·麦卡锡（John McCarthy）和马文·明斯基（Marvin Minsky）说服了著名的信息论创始人克劳德·香农（Claude Shannon）和IBM（国际商业机器公司）第一台商用计算机的设计者纳撒尼尔·罗切斯特（Nathaniel Rochester）加入他们，一起在达特茅斯学院组织一个暑期项目。项目的目标如下：

这项研究基于一个猜想：“学习”的各个方面或“智能”的任何特征在原则上都可以被精确地描述出来，所以人们可以制造一台机器来模拟“学习”或“智能”。人们试图让机器使用语言，形成抽象概念，解决现在人类面临的各种问题，并让机器自我改进。我们认为，一个精英科学家小组如果共同努力一个夏天，就可以在其中的一个或多个问题上取得重大进展。

不用说，我们花费的时间已经远远超过了一个夏天，但我们仍在努力解决这些问题。

在达特茅斯会议后10年左右的时间里，人工智能取得了几项重大成功，包括艾伦·罗宾逊（Alan Robinson）的通用逻辑推理算法[2]和阿瑟·塞缪尔（Arthur Samuel）的西洋跳棋程序，这个程序通过自学战胜了它的创造者。[3] 第一次人工智能泡沫在20世纪60年代末破裂，当时，人们在机器学习和机器翻译方面的早期努力没能达到预期。英国政府在1973年委托撰写的一份报告中总结道：“到目前为止，该领域的任何发现都没有产生当初承诺的重大影响。”[4]换言之，这些机器不够智能。

幸运的是，当时11岁的我并不知道这份报告。两年后，我得到了一台辛克莱剑桥可编程计算器，我想让它变得智能。然而，辛克莱的程序有36步运算限制，这对于人类级别的人工智能来说还不够大。我并没有泄气，而是使用伦敦帝国理工学院的CDC 6600超级计算机[5]编写了一个国际象棋程序——一沓两英尺[6]高的穿孔卡片。虽然这个程序不够好，但是没关系。我知道我想做什么。

到了20世纪80年代中期，我已经成为加州大学伯克利分校的一名教授，由于所谓的“专家系统”的商业潜力，人工智能正在经历一次巨大的复兴。当这些系统被证明无法完成人们指派给它们的许多任务时，第二次人工智能泡沫破裂了。同样，这些机器也不够智能。人工

智能的寒冬随之而来。我在伯克利开设的人工智能课程目前有900多名学生，而在1990年，我只有25名学生。

人工智能界吸取了教训：显然，机器越智能越好，但我们必须做好功课才能做到这一点。这个领域与数学的关系变得尤为密切。人们将人工智能与历史悠久的概率论、统计学和控制理论建立了联系。今天，进步的种子是在那个人工智能的寒冬播下的，包括对大规模概率推理系统以及后来众所周知的深度学习的早期研究。

从2011年前后开始，深度学习技术开始在语音识别、基于机器视觉的物体识别和机器翻译方面取得巨大进步，以上是人工智能领域最重要的三个开放问题。从某些方面来看，机器在这些方面的能力已经达到甚至超过了人类的能力。2016年和2017年，DeepMind（英国人工智能公司）的AlphaGo（阿尔法围棋）击败了前世界围棋冠军李世石和后来的冠军柯洁，而此前一些专家预测，这种事情如果会发生，也要到2097年才会发生。[7]

如今，人工智能几乎每天都会出现在媒体的头条报道中。在大量风险投资的推动下，成千上万家初创公司应运而生。数以百万计的学生参加在线人工智能和机器学习课程，该领域专家的年薪可达数百万美元。来自风险基金、国家政府和大公司的投资每年达到数百亿美元，该领域过去5年获得的资金超过了其在先前整个历史中所获得的总额。无人驾驶汽车和智能个人助理等已经在酝酿中的技术进步，很可能在未来10年左右对世界产生重大影响。人工智能潜在的经济效益和社会效益是巨大的，这为人工智能研究提供了巨大的动力。

接下来会发生什么？

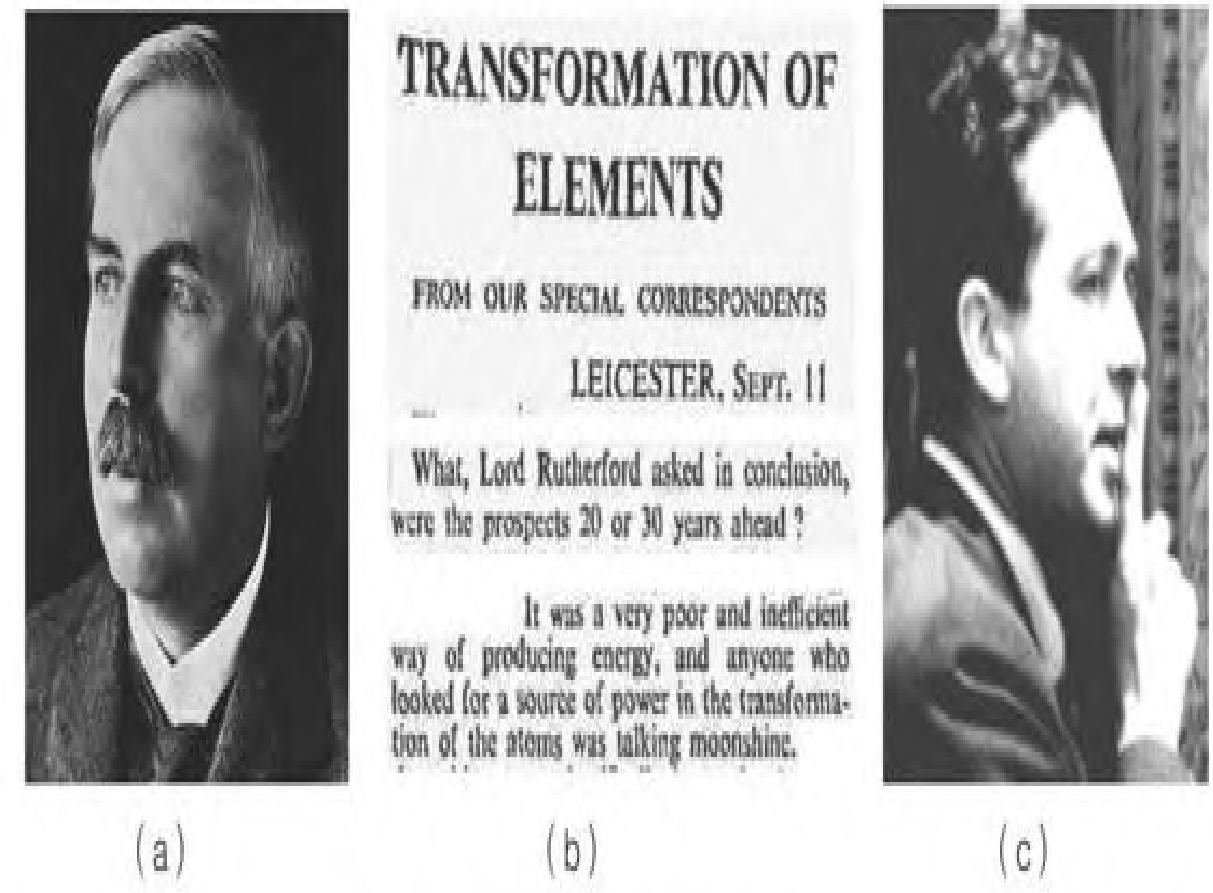
这种快速的进步是否意味着我们即将被机器超越？不是。我们在拥有类似的具有超人类智能的机器之前，必须要实现几项突破。

众所周知，科学突破是很难预测的。为了了解到底有多难，我们可以回顾一下另一个领域的历史，这个领域具有终结文明的潜力，那就是核物理学。

20世纪初，也许没有哪位核物理学家比质子的发现者、“分裂原子的人”欧内斯特·卢瑟福［Ernest Rutherford，图2（a）］更为杰出。与他的同事一样，卢瑟福早就意识到原子核储存了巨大的能量，然而，主流观点认为开发这种能源是不可能的。

1933年9月11日，英国科学促进协会在莱斯特举办年会，卢瑟福在晚间会议上发言。正如他之前几次所做的那样，他向原子能的前景泼了冷水：“任何想从原子的转变中获取能源的人都是异想天开。”第二天早上，伦敦《泰晤士报》报道了卢瑟福的演讲［图2（b）］。

匈牙利物理学家利奥·西拉德［Leo Szilard，图2（c）］那时逃离了纳粹德国，住在伦敦罗素广场的帝国酒店。他早餐时阅读了《泰晤士报》的报道，随后出去散步，仔细思考着报道的内容，构想出了中子诱发的链式核反应。[8]在不到24小时的时间里，释放核能的问题从不可能变成了已经基本解决。次年，西拉德申请了核反应堆的保密专利。1939年，法国颁发了第一项核武器专利。



TRANSFORMATION OF ELEMENTS

FROM OUR SPECIAL CORRESPONDENTS

LEICESTER, SEPT. 11

What, Lord Rutherford asked in conclusion, were the prospects 20 or 30 years ahead?

It was a very poor and inefficient way of producing energy, and anyone who looked for a source of power in the transformation of the atoms was talking moonshine.

(a) (b) (c)

图2 （a）核物理学家卢瑟福男爵；（b）1933年9月12日《泰晤士报》的一篇报道，专家谈及前一天晚上卢瑟福的演讲；（c）核物理学家利奥·西拉德

这个故事的寓意是，与人类的聪明才智打赌是鲁莽的，尤其是在我们的未来岌岌可危的时候。在人工智能界，一种否定主义正在出现，它甚至否定成功实现人工智能长期目标的可能性。这就像一名驾驶着载有全人类的公共汽车的司机说：“是的，我正尽我所能往悬崖边开，但是相信我，在开到悬崖边之前，我们的汽油会用完！”

我并不是说人工智能一定会成功，而且我认为人工智能在未来几年内成功的可能性不大。尽管如此，为可能发生的事情做好准备才是慎重之举。如果一切顺利，这将预示着人类进入黄金时代，但我们必须面对的事实是，我们正计划着制造出远比人类强大的实体。我们如何确保它们永远不会掌控我们？

为了对我们正在玩的火有一个初步的了解，请思考一下内容推荐算法在社交媒体中发挥的作用。虽然这些算法不是特别智能，但它们能够影响整个世界，因为它们直接影响数十亿人。通常，此类算法旨在最大限度地提高点击率，即用户点击展示条目的概率。那么，解决方案就是简单地展示用户喜欢点击的条目，对吗？不对。正确的解决方案是改变用户的偏好，从而使他们变得更可预测。对于更可预测的用户，算法可以通过推送他们更可能会点击的条目，而带来更多收入。持有极端政治观点的人往往更容易被预测出会点击哪些条目。（顽固的中间派可能会点击某一类文章，但算法很难想象这类文章由什么内容组成。）与所有理性实体一样，算法会学习如何修改它的环境状态（在这个例子中是用户的思想），从而使自身获得的奖励最大化。[9]其后果包括法西斯主义复活，支撑世界民主的社会契约被解除，这甚至可能导致欧盟和北约的终结。区区几行代码在得到了人类的帮助后，就可以导致上述后果。现在想象一下，一个真正智能的算法能做什么？

出了什么问题？

人工智能的历史一直被一句咒语所推动：“越智能越好。”我确信这是一个错误，不是因为我隐隐地害怕被人工智能取代，而是因为我们理解智能的方式有误。

智能的概念对于“我们是谁”至关重要，这就是为什么我们称自己为“智人”或“智者”。经过两千多年的自我审视，我们对智能的描述可以归结为：

人类是智能的，因为我们的行动有望实现我们的目标。

智能的其他所有特征，例如感知、思考、学习、创造等，都可以通过它们对我们成功行动的能力的贡献来理解。从人工智能诞生之初，机器智能的定义就与之对应：

机器是智能的，因为它们的行动有望实现它们的目标。

机器与人类不同，它们没有自己的目标，因此我们赋予它们目标，让它们去实现。换言之，我们制造能够自我优化的机器，把目标输入机器中，然后它们就开始运行了。

这种通用方法并不是人工智能独有的，该方法在我们社会的技术和数学基础中反复出现。在设计了从大型喷气机到胰岛素泵等各种控制系统的控制理论领域，系统的任务是使代价函数最小化（代价函数通常被用来衡量实际行为与期望行为的偏差）。在经济学领域，机制和政策旨在使个人的效用（utility）、群体的福利和公司的利润最大化。[10]在解决复杂的物流和制造问题的运筹学中，解决方案使预期回报的总和随着时间的推移最大化。最后，在统计学中，学习算法（learning algorithm）旨在使期望损失函数最小化（损失函数定义了产生预测错误要承担的成本）。

显然，这个通用方案（我将其称为标准模型，standard model）用途广泛且极其强大。不幸的是，我们不希望机器智能也表现在这个方面。

1960年，麻省理工学院的传奇教授、20世纪中叶最杰出的数学家之一诺伯特·维纳（Norbert Wiener）指出了标准模型的缺陷。维纳看到阿瑟·塞缪尔的西洋跳棋程序学会了下跳棋，而且其水平远远超过了它的创造者。这段经历让他写出了一篇颇有先见之明但鲜为人知的论文——《自动化的一些道德和技术后果》。[11]以下是维纳的主要观点：

> 如果我们为了达到目的而使用一种我们无法有效干预其运行的机器……我们最好确信我们想让机器达成的目标是我们真正想要实现的目标。

“想让机器达成的目标”正是机器在标准模型中优化的目标。如果我们把错误的目标输入比我们更智能的机器里，机器就会实现目标，但我们也就失败了。我先前描述的社交媒体崩溃就是这种情况的前兆，它是在全球范围内用相当不智能的算法优化错误的目标所致的。在第5章中，我会详细说明一些更糟糕的结果。

所有这些都不足为奇。几千年来，我们一直都知道得到想要的东西有多危险。在所有的故事里，当一个人被赐予三个愿望时，他的第三个愿望总是会把前两个愿望推翻。

总而言之，迈向超人类智能的步伐似乎是不会停止的，但超人类智能的成功可能意味着人类的毁灭。然而，并非一切都失去了意义。我们必须了解是哪里出了问题，然后加以解决。

我们能解决吗？

问题就出在人工智能的基本定义中。我们说，机器是智能的，因为它们的行动有望实现它们的目标，但我们没有可靠的方法来确保它们的目标与我们的目标相同。

如果我们坚持让机器追求我们的目标，而不是让机器追求它们的目标，那么会如何？如果这样的机器能被设计出来，那么它们就不仅是智能的，而且是对人类有益的。因此，让我们试试这个：

> 机器是有益的，因为它们的行动有望实现我们的目标。

这可能才是我们一直以来应该做的。

当然，困难之处在于，我们的目标在我们自己心中（在全体80亿人和我们多样的社会群体之中），而不在机器中。尽管如此，我们还是有可能制造出在这个意义上有益的机器。不可避免的是，这些机器将会对我们的目标感到不确定，毕竟我们自己也不确定，但事实证明，这是一个特性，而不是漏洞（也就是说，是好事而不是坏事）。目标的不确定性意味着机器必须要顺从人类：它们会请求许可，会接受纠正，会允许自己被关闭。

撤销“机器应该有明确目标”这个假设，意味着我们需要打破并替换人工智能的部分基础，而这些基础其实是我们正在努力构建的东西的基本定义。这也意味着要重建大量的上层建筑，即为实际的人工智能工作积累思想和方法。其结果将是人类和机器之间生成新关系，我希望这种关系将使我们能够在未来几十年顺利前行。

---

[1] 我与现任谷歌研究总监彼得·诺维格合著的人工智能教科书的第一版：Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 1st ed.(Prentice Hall, 1995)。

[2] 罗宾逊开发了解决算法，如果给定足够的时间，它可以证明一组一阶逻辑断言的任何逻辑结论。与以前的算法不同，它不需要转换成命题逻辑。J. Alan Robinson, “A machine-oriented logic based on the resolution principle,” *Journal of the ACM* 12 (1965): 23–41.

[3] 计算机时代的美国先驱阿瑟·塞缪尔在IBM完成了他的早期工作。尽管艾伦·图灵早在1947年就谈到了“一种可以从经验中学习的机器”，但描述塞缪尔在跳棋方面的工作的论文第一次使用了“机器学习”这个词。Arthur Samuel, “Some studies in machine learning using the game of checkers,” *IBM Journal of Research and Development* 3 (1959): 210–29.

[4] 赫赫有名的《莱特希尔报告》（Lighthill Report）导致除爱丁堡大学和萨塞克斯大学之外的人工智能研究经费被终止：Michael James Lighthill, “Artificial intelligence: A general survey,” in *Artificial Intelligence: A Paper Symposium*(Science Research Council of Great Britain, 1973)。

[5] CDC 6600占据了整个房间，花费相当于2 000万美元。如今一部苹果手机的功能也要比它强100万倍，但是在那个时代，它仍然强大得令人难以置信。

[6] 1英尺=0.3048米。——编者注

[7] “深蓝”战胜卡斯帕罗夫之后，至少有一位评论员预测，同样的事情要过100年才会在围棋界发生：George Johnson, “To test a powerful computer, play an ancient game,” *The New York Times*, July 29, 1997。

[8] 有关核技术发展的颇有可读性的历史，请参见：Richard Rhodes, The *Making of the Atomic Bomb* (Simon & Schuster, 1987)。

[9] 简单的监督学习算法可能不会有这种效果，除非它被封装在A/B测试框架中（这在在线营销环境中很常见）。如果强盗算法（bandit algorithm）和强化学习算法使用用户状态的显式表示或与用户交互历史的隐式表示进行操作，它们就会产生这种效果。

[10] 一些人认为，利润最大化的公司已经是失控的人工实体。例如，请参见：Charles Stross, “Dude, you broke the future!” (keynote, 34th Chaos Communications Congress, 2017)。另请参见：Ted Chiang, “Silicon Valley is turning into its own worst fear,” *Buzzfeed*, December 18, 2017。对这一想法的更进一步的探讨请见：Daniel Hillis, “The first machine intelligences,” in *Possible Minds: Twenty Five Ways of Looking at AI*, ed. John Brockman (Penguin Press, 2019)。

[11] 当时，主流观点是“所有技术进步都是好事”，但维纳的论文是一个例外。请参见：Norbert Wiener, “Some moral and technical consequences of automation,” *Science* 131 (1960): 1355–58。


02
人类和
机器的智能



当你走进死胡同时，你最好掉头返回，找出走错的地方。我认为人工智能的标准模型，即机器优化人类提供的固定目标，是一条死胡同。问题不在于我们可能无法做好构建人工智能系统的工作，而在于我们可能太“成功”了。目前人工智能领域对“成功”的定义是错误的。

所以，让我们一直回溯到起点，试着理解“智能”的概念是如何产生的，以及这个概念是如何被应用到机器中的。然后我们就有可能对“好的人工智能系统”给出更好的定义。

智能

宇宙是如何运行的？生命是如何开始的？我的钥匙放在哪里了？这些都是值得思考的基本问题。但是，是谁在问这些问题？我又如何回答这些问题？我们称之为“大脑”的那一小堆几磅重的粉灰色胶状混合物是如何感知、理解、预测和操纵一个难以想象的浩瀚世界的？不久，意识开始审视自身。

几千年来，我们一直试图理解人类的大脑是如何工作的。最初的研究目标包括理解好奇心、自我管理、说服力，以及分析数学参数这种相当务实的目标。然而，我们朝着解释大脑如何工作迈出的每一步，也同时是迈向在人工制品中创造大脑能力的一步，即迈向人工智能的一步。

在我们理解如何创造智能之前，我们要先理解智能是什么。答案不在智商测试中，甚至不在图灵测试中，而是在我们的感知、我们的需求和我们所做的事情这三者之间的简单关系中。粗略地讲，如果一个实体是智能的，它就会根据它的感知来做事情，而它所做的事情是为了满足它的需求。

进化的起源

我们来看一种低等细菌，例如大肠杆菌。它大约有6根鞭毛，即形如发丝的触须，鞭毛在底部顺时针或逆时针旋转。（旋转马达本身是件了不起的东西，但那是另外一回事。）当大肠杆菌在它的液体家园——你的小肠中漂浮时，它会顺时针旋转鞭毛，使其“翻滚”到位，也会逆时针旋转鞭毛，使鞭毛缠绕在一起形成“推进器”，从而让细菌可以沿直线游动。因此，大肠杆菌会随机游走——游泳、翻滚、游泳、翻滚，这使得它能够找到并消耗葡萄糖，而不是待在原地饿死。

如果这就是全部情况，那么我们不会说大肠杆菌特别智能，因为它的动作不会以任何方式依赖于它的环境。它不会做任何决定，只是执行一种固定的行为，进化将这种行为植入了它的基因，但这并不是全部。当大肠杆菌感知到葡萄糖浓度升高时，它游动的距离会更长，翻滚也会更少，而当它感知到葡萄糖浓度降低时，它的反应则正好相反。所以，根据它感知到的情况（葡萄糖浓度不断升高），它所做的事情（游向葡萄糖）很可能满足它的需求（得到更多的葡萄糖）。

也许你在想：“但是进化也把这些植入了它的基因！这怎么能说是智能呢？”这种推理方式是危险的，因为进化也将你大脑的基本设计植入了你的基因，想必你不会希望在此基础上否认自己的智能。关键是，进化在大肠杆菌的基因中所建立的机制，就像在你的基因中所建立的机制一样。通过这种机制，细菌的行为会随着它对环境的感知而变化。进化事先不知道葡萄糖在哪里，也不知道你的钥匙在哪里，所以把找到葡萄糖和钥匙的能力放到生物体中，简直近乎完美。

大肠杆菌并非有多么强大的智能。据我们所知，它不记得它去过哪里，所以如果它从A点游到B点而没找到葡萄糖，此时它有可能回到A点。如果我们构建一个环境，在这个环境中，每一个有吸引力的葡萄糖浓度梯度都会将大肠杆菌引向苯酚（苯酚对它而言是一种毒药），

那么它也会继续沿着这些浓度梯度前进。它从不学习，也没有大脑，只通过一些简单的化学反应来完成这项工作。

动作电位向前迈出了一大步。这是一种电信号，大约10亿年前首次在单细胞生物体内进化出来。后来，多细胞生物进化出一种叫作神经元的特殊细胞，这些细胞利用动作电位在生物体内快速传递信号，速度高达每秒120米（即每小时270英里[1]）。神经元之间的连接叫作突触。突触连接的强度决定了有多少电刺激从一个神经元传递到另一个神经元。动物通过改变突触连接的强度来学习。[2]学习带来了巨大的进化优势，因为动物可以适应各种环境。学习也加快了进化本身的速度。

最初，神经元形成了神经网络，这些神经网络分布在整个生物体中，用于协调各种活动，如进食与消化，以及肌肉细胞在大范围内的定时收缩。水母之所以能优雅地推进是因为它们拥有神经网络，但水母根本没有大脑。

大脑以及眼睛和耳朵等复杂的感觉器官是后来才出现的。在水母和它们的神经网络出现几亿年后，我们人类带着我们由千亿个（10^{11}）神经元和千万亿（10^{15}）个突触组成的巨大的大脑来到了这个世界。大脑中每个状态转换的“周期时间”只有几毫秒，虽然比电子电路速度慢，但是比大多数生物过程快得多。人类的大脑经常被它的主人描述为“宇宙中最复杂的物体”，这可能并非事实，却是一个很好的借口，因为我们仍然对大脑的工作原理知之甚少。虽然我们对神经元和突触的生物化学知识以及大脑的解剖结构了解很多，但是认知水平的神经实现——学习、认知、记忆、推理、计划、决定等，对绝大多数人而言仍然是猜想。[3]（也许随着我们对人工智能的了解越来越多，或者随着我们开发出越来越精确的工具来测量大脑活动，这种情况会有所改变。）因此，当人们在媒体上读到某人工智能技术“像人脑一样工作”时，人们可能会怀疑它要么只是某人的猜想，要么纯属虚构。

在意识领域，我们确实一无所知，所以我不会谈及该领域的内容。人工智能领域没有人致力于让机器拥有意识，也没有人知道从哪里开始，而且没有任何行为是以意识为前提的。假设我给你一个程序，问你，它是否对人类构成威胁，你分析代码，发现代码在运行时会形成并执行一个计划，其结果将是毁灭人类，就像国际象棋程序会形成并执行一个计划，其结果将是战胜一切对手一样。现在假设我告诉你，代码运行时也会产生一种机器意识，这会改变你的预测吗？完全不会，这没什么区别。[4]你对其行为的预测是完全相同的，因为预测是基于代码的。所有关于机器神秘地变得有意识并憎恨人类的那些好莱坞电影情节其实都没有抓住要点——重要的是能力，而不是意识。

我们才刚刚开始了解大脑的一个重要的认知方面——奖励系统（reward system）。这是一个由多巴胺介导的内部信号系统，它将积极的和消极的刺激与行为联系起来。它的工作原理是由瑞典神经学家尼尔斯·奥克·希拉普（Nils-Åke Hillarp）和他的合作者在20世纪50年代末发现的。这个系统让我们寻求积极的刺激，如甜食，以提高多巴胺水平；它让我们避免消极的刺激，如饥饿和疼痛，因为这会降低多巴胺水平。从某种意义上讲，它与大肠杆菌寻找葡萄糖的机制非常相似，但要复杂得多。它有内置的学习方法，因此随着时间的推移，我们的行为在获得奖励方面会变得更加高效。它还允许延迟满足，因此我们学会了渴求那些能提供最终回报而不是即时回报的东西，比如金钱。我们去理解大脑奖励系统的一个原因是，它类似于人工智能领域开发的强化学习方法，对此我们有非常坚实的理论基础。[5]

从进化的角度看，我们可以认为大脑的奖励系统就像大肠杆菌寻找葡萄糖的机制一样，是提高进化适应性的一种方式。在寻求奖励方面更高效的生物体（例如寻找美味的食物、避免疼痛、进行性爱活动等）更有可能延续它们的基因。从长远来看，生物体很难判断哪些行

为最有可能让其基因成功延续，因此进化为我们提供了内置的路标，让我们更容易做到这一点。

然而，这些路标并不完美。某些获得奖励的方法可能会降低一个人基因延续的可能性，例如吸毒、饮用大量含糖碳酸饮料、每天玩18个小时的电子游戏等，这些似乎都不利于繁衍。此外，如果直接用电流刺激你的奖励系统，你可能会不停地进行自我刺激，直到死去。[6]

奖励信号和进化适应性之间的不一致并不只影响孤立的个体。在巴拿马海岸附近的一个小岛上，生活着侏三趾树懒。它对红树林叶子中的安定类物质上瘾，并可能因此灭绝。[7]因此，如果一个物种找到了一个生态位，但这个生态位满足其奖励系统的方式不利于进化适应性，那么这个物种就可能会灭绝。

然而，除了这种意外的失败，生物体学习在自然环境中将奖励最大化，通常会提高它们延续自己的基因和在环境变化中生存下来的机会。

进化的加速器

学习不仅对生存和繁衍有好处，还加速了进化。这怎么可能？毕竟，学习不会改变一个人的DNA（脱氧核糖核酸），而进化则是一代又一代地改变DNA。学习和进化之间的联系是由美国心理学家詹姆斯·鲍德温（James Baldwin）[8]和英国生态学家康威·劳埃德·摩根（Conwy Lloyd Morgan）[9]在1896年分别提出的，但当时并未被人们普遍接受。

正如我们现在所知道的，鲍德温效应可以这样来理解：进化可以选择创造一种依靠本能生存的生物体，或者创造一种依靠适应性生存的生物体，前者的每一个反应都是预先确定的，而后者则可以通过学

习来确定采取何种行动。现在，为了便于说明，假设最优的依靠本能生存的生物体可以编码为一个六位数，例如472 116，而对于依靠适应性生存的生物体，进化只指定了472 ×××，生物体自身必须在其一生中通过学习来填充后三位数。显然，如果进化只需要选择前三位数，那么它的工作就简单多了。学习后三位数的过程让依靠适应性生存的生物体在其一生中完成了进化需要很多代才能完成的工作。因此，如果依靠适应性生存的生物体能够在学习中生存，那么学习能力就构成了进化的捷径。计算模拟表明鲍德温效应是真实存在的。[10]文化的影响只会加速这一过程，因为有组织的文明不仅会在个体生物体进行学习的时候保护它，还会传递个体本来需要自己学习的信息。

鲍德温效应的故事引人入胜，但并不完整：它假设学习和进化必然指向同一方向，也就是说，它假设无论哪种内部反馈信号定义了生物体学习的方向，这种信号都与进化适应性完全一致。正如我们在侏三趾树懒的例子中所看到的那样，这并不是事实。内置的学习机制充其量只能粗略暗示特定行为在长期进化适应中产生的结果。此外，人们不得不问：“奖励系统最初是如何实现的？”答案当然是通过进化的过程实现的，奖励系统内化了某种反馈机制，这种反馈机制至少在某种程度上与进化适应性是一致的。[11]显然，如果有一种导致生物体逃离潜在配偶、奔向捕食者的学习机制，那么它不会持续很久。

因此，我们拥有鲍德温效应要感谢一个事实：神经元具有学习和解决问题的能力，这在动物王国中非常常见。与此同时，重要的是你要明白，进化并不真的在乎你是否拥有大脑或是否可以产生有趣的想法。进化只把你当作一个智能体（agent），即某种会行动的东西。逻辑推理、有目的的计划、智慧、机智、想象力和创造力等有价值的智力特征，可能是使智能体变聪明的必要条件，也可能不是。人工智能如此引人入胜的一个原因是，它为理解这些问题提供了一条潜在的途径：我们可能会理解这些智力特征如何使智能行为成为可能，以及为什么没有它们就不可能产生真正的智能行为。

一个人的理性

从古希腊哲学的早期开始，智能的概念就与感知、推理和成功行动的能力联系在一起。[12] 几个世纪以来，这个概念的适用性越来越广，定义也越来越精确。

亚里士多德等人研究了成功地进行推理的构想——逻辑推理法，即在给出真实前提的情况下，人们通过逻辑推理可以得出真实的结论。他还研究了人们决定如何行动的过程，这个过程有时被称为实践推理，他提出，这一过程包括推断某一行动是否会实现预期目标：

> 我们考虑的不是结果，而是方法。医生不考虑是否应该治愈病人，演说家也不考虑是否应该游说……他们设定该目标，并考虑如何达到该目标，用什么方法达到该目标，是否可以容易且完美地达到该目标。然而，如果只能通过一种方法来实现该目标，那么他们就只考虑如何通过这种方法实现，以及这种方法要如何实现，直到他们找到第一个目标……分析序列中的最后一个成了实际序列中的第一个。如果我们遇到了不可能的情况，我们就放弃寻找，例如，如果我们需要钱却没有钱，我们就会放弃，但如果一件事看起来是可能的，我们就会去尝试。[13]

有人可能会说，这篇文章为其后两千多年的西方理性思想定下了基调。文章中说，“目标”，即人们想要的东西，是固定不变的；文章又说，理性行为是根据一系列行为的逻辑演绎，“容易且完美”地达到目标。

亚里士多德的建议似乎是合理的，但这并不是关于理性行为的完整指南。特别是，他忽略了不确定性的问题。在现实世界中，现实有干预的倾向，很少有某个行动或一系列行动一定能真正达到预期目的。例如，在我写这句话的时候，巴黎是一个下雨的星期天。下星期

二下午2点15分，我飞往罗马的航班将从戴高乐机场起飞，机场离我家大约45分钟的路程。我计划上午11点半左右出发去机场，这应该留出了足够的时间，但这可能意味着我至少要在候机厅坐一个小时。我肯定能赶上飞机吗？未必。因为我可能会遇到严重的交通堵塞，出租车司机可能会罢工，我乘坐的出租车可能会抛锚，司机可能会在高速公路上因超速被逮捕等。而我也可以提前一天，星期一去机场。这将大大降低错过航班的概率，但在候机厅过夜并不是一个好主意。换言之，我的计划要在“成功地赶上飞机”这一确定性和“为确保这一确定性所付出的成本”之间进行权衡。下面这个购房计划涉及一个类似的权衡：买张彩票，赢100万美元，然后买房子。这个“容易且完美”的计划可以被看作目标，但它不太可能实现。然而，这个轻率的购房计划与我冷静理智的机场计划之间的区别，只是确定性程度上的不同而已。二者都是赌博，但后一个似乎比前一个更理性。

结果证明，赌博在推广亚里士多德关于不确定性的建议中起了核心作用。16世纪60年代，意大利数学家杰罗拉莫·卡尔达诺（Gerolamo Cardano）以掷色子游戏为主要例子，第一个提出了在数学上精确的概率论。（不幸的是，他的著作直到1663年才出版。[14]）17世纪，包括安托万·阿尔诺（Antoine Arnauld）和布莱兹·帕斯卡（Blaise Pascal）在内的法国思想家开始研究赌博中的理性决策问题，这当然是出于数学上的原因。[15]请看以下两个赌局：

赌局A：有20%的机会赢得10美元

赌局B：有5%的机会赢得100美元

数学家对此提出的建议可能和你的想法是一样的：比较两个赌局的期望值，即你期望从每次下注中得到的平均金额。对于赌局A，赌客的期望值是10美元乘以20%，即2美元。对于赌局B，赌客的期望值是100美元乘以5%，即5美元。因此，根据这个理论，赌局B更好。这个理

论是有道理的，因为如果同样的赌注被一次又一次地开出，那么遵守这一规则的赌客最终会比不遵守它的人赢得更多的钱。

18世纪，瑞士数学家丹尼尔·伯努利（Daniel Bernoulli）注意到，这个规则似乎不适用于金额较大的情况。[16]例如，请看以下两个赌局：

赌局A：有100%的机会获得10 000 000美元（期望值是10 000 000美元）

赌局B：有1%的机会获得1 000 000 100美元（期望值是10 000 001美元）

这本书的大多数读者，以及笔者本人，都更喜欢赌局A甚于赌局B，即便这与期望值规则相悖！伯努利假设，对赌局进行评估不能根据期望的货币价值，而是要根据预期的效用。他认为，效用即对一个人有用或有益的属性，这是一种内在的、主观的数量，与货币价值有关，但与货币价值不同。特别值得注意的是，效用相对于金钱收益递减。这意味着给定金额的货币的效用并不严格地与金额成正比，而是增长得更慢。例如，在上面的例子中，获得1 000万美元的效用比获得10亿零100美元的效用大了远远不止100倍。那么大多少？你可以问自己！你赢得10亿零100美元的概率达到多少才能让你放弃100%赢得1000万美元的机会？我拿这个问题问了我班上的研究生，他们的答案大约是50%，这意味着赌局B的期望值要达到约5亿美元才能与赌局A的可取性相匹配。那时赌局B的期望值约为赌局A的50倍，但两个赌局的效用相同。

伯努利引入了效用这一无形的属性，通过数学理论来解释人类行为，这在当时是非常了不起的。更值得注意的是，与货币金额不同的是，各种赌局和奖品的效用价值是无法被直接观察到的，只能从个人

表现出的偏好中被推断出来。直到两个世纪后，效用这个概念才被完全理解，并被统计学家和经济学家广泛接受。

20世纪中叶，约翰·冯·诺依曼（John von Neumann，一位伟大的数学家，计算机领域标准的“冯·诺依曼体系结构”就是以他的名字命名的[17]）和奥斯卡·摩根斯坦（Oskar Morgenstern）发表了效用理论的公理基础。[18]这意味着，只要个体表现出的偏好满足任何理性主体都应该满足的某些基本公理，那么这个人所做的选择必然可以被描述为最大化期望效用函数。简而言之，理性个体的行为是为了最大化期望效用。

这个结论的重要性不言而喻。在许多方面，人工智能主要是研究如何构建理性机器的细节。

让我们更详细地看看理性实体应该满足的公理。有一条公理叫作传递性：如果你喜欢A甚于B，喜欢B甚于C，那么你喜欢A甚于C。这似乎很合理！（如果你喜欢香肠比萨胜过原味比萨，喜欢原味比萨胜过菠萝比萨，那么你会选择香肠比萨而不是菠萝比萨，这似乎是合理的预测。）还有一条公理叫作单调性：假设买一种彩票可能得到奖品A，买另一种彩票可能得到奖品B，你喜欢奖品A甚于奖品B，并且你只可以选择这两种彩票，那么你会更倾向于选择得到奖品A的概率最高的彩票，而不是得到奖品B的概率最高的彩票。这同样很合理。

偏好不仅仅和比萨以及有奖品的彩票有关。它们可以是任意事情，特别值得注意的是，它们可以和你未来的一生以及其他人的一生有关。在处理随时间变化的事件序列的偏好时，通常会有一个附加的假设，这个假设叫平稳性：如果两个不同的未来——A和B，以相同的事件开始，并且你喜欢A甚于B，那么在事件发生后，你仍然喜欢A甚于B。这听起来很合理，但它有一个出人意料的强烈后果：任何事件序列的效用，都是与每个事件相关的奖励的总和（可能随着时间的推移，通过某种心理利率打折）。[19]尽管这种“效用作为奖励的总和”的假设广为流传——至少可以追溯到18世纪效益主义

（utilitarianism，旧译为“功利主义”）的创始人杰里米·边沁（Jeremy Bentham）的“幸福计算”，但它所基于的平稳性假设不是理性主体的必要属性。平稳性排除了一个人的偏好可能会随着时间而改变的可能性，而我们的经验表明这种可能性是存在的。

尽管这些公理是合理的，而且从中得出的结论也很重要，但是从效用理论被广泛传播开始，该理论就一直遭到接二连三的反对。有人鄙视它，认为它把一切都归结为金钱和自私。（这一理论被一些法国作家嘲笑为“美国式的”，虽然它起源于法国。[20]）事实上，仅仅因为希望减少他人的痛苦，而去过一种自我否定的生活，这也是完全理性的。利他主义仅仅意味着在评估未来时，要充分重视他人的福祉。

另一系列反对意见认为，获得必要的概率和效用值，并将它们相乘来计算期望效用，这是很有难度的。这些反对意见混淆了选择理性行为和通过计算期望效用来选择理性行为这两件不同的事情。例如，如果你试图用手指戳自己的眼球，你就会闭上眼睑保护你的眼睛，这是合理的，但这里并不涉及对期望效用的计算。或者假设你骑着一辆没有刹车的自行车下山，你要在两种选择之间做出选择：一种是以每小时10英里的速度撞上一面混凝土墙，另一种是以每小时20英里的速度撞上山下的另一面墙。你会怎么选？如果你选择了前者，那么恭喜你！你计算过期望效用吗？可能没有。但是选择前者仍然是合理的。这基于两个基本假设：第一，比起更严重的伤害，你宁愿选择不太严重的伤害；第二，对于给定的任意伤害级别，提高碰撞时的速度很可能会提高伤害的级别。从这两个假设出发，从数学上（完全不考虑任何数字）可以得出，以每小时10英里的速度撞墙比以每小时20英里的速度撞墙具有更高的期望效用。[21]总之，最大化期望效用可能不需要计算任何预期或任何效用。这纯粹是对理性实体的外部描述。

对理性理论的另一种批判在于如何确定决策的源头。也就是说，什么东西算智能体？似乎很明显，人类是智能体，但是家庭、部落、

公司、文化和民族国家是智能体吗？如果我们研究的是像蚂蚁这样的社会性昆虫，那么将单只蚂蚁视为一个智能体是否有意义？或者智能是否真的存在于整个蚂蚁群体中，由多只蚂蚁的大脑和身体组成的复合脑通过信息素而不是电信号相互联系？从进化的角度来看，这可能是一种更有效的了解蚂蚁的方式，因为特定群体中的蚂蚁通常关系密切。作为个体，蚂蚁和其他社会性昆虫缺乏自我保护的本能，因为自我保护与群体保护截然不同：它们总会投入与入侵者的战斗中，即使等同于自杀也奋不顾身。然而，有时人类也会这样做，甚至会为了保护不相关的人而这样做。这就好像整个物种受益于某些个体的存在，这些个体愿意在战斗中牺牲自己，或是愿意踏上未知的探险之旅，或是愿意去养育他人的后代。在这种情况下，将理性的分析完全集中在个体身上显然遗漏了一些重要的东西。

对效用理论的主要反对意见来自经验主义，也就是说，这些意见是基于实验证据的，而这些实验证明：人类是非理性的，我们无法以系统的方式遵循公理。[22]我在这里的目的并不是为效用理论作为人类行为的正式模型进行辩护。的确，人类不可能完全理性行事。无论是现在还是将来，我们的偏好都会延伸到我们自己未来的生活，我们子孙后代的生活，以及其他人的生活之中。然而，我们甚至不能在棋盘上走出正确的步法，尽管棋盘很小，布局简单，规则明确，下一局棋不用花费很长的时间。这并不是因为我们的偏好是非理性的，而是因为决策问题的复杂性。我们的认知结构在很大程度上是为了填补一条鸿沟而存在的，这条鸿沟就是我们小而迟钝的大脑和我们一直面临的难以理解的复杂决策问题之间的鸿沟。

因此，虽然将“有益的人工智能”（Beneficial AI，后简称“有益的AI”）理论建立在“人类是理性的”这一假设之上非常不合理，但假设成年人对未来生活有大体一致的偏好则非常合理。也就是说，如果你看了两部电影，每一部都足够详细和广泛地描述了你在未来可

能过上的生活，这样每一部就都构成了一种虚拟的体验，你可以说自己更喜欢哪一部，或者表示中立。[23]

如果我们唯一的目标是确保足够智能的机器不会对人类造成灾难性的后果，那么这种说法可能比我们想象中的更为严重。“灾难”的概念本身就意味着这是一种绝对不受人欢迎的生活。因此，为了避免灾难，我们只需要宣称，当一个灾难性的未来被详细描述时，成年人是能够认识到它的。当然，与“无灾无难总比受灾好”相比，人类的偏好有更精细的结构，而且确定性更高。

事实上，“有益的AI”这一理论可以适应人类偏好的不一致，但是你的偏好中不一致的部分将永远无法得到满足，人工智能也无能为力。例如，假设你对比萨的偏好违反了传递性公理，那么下面这个场景就可能出现：

机器人：欢迎回家！想吃菠萝比萨吗？

你：不，你应该知道，比起菠萝比萨，我更喜欢原味比萨。

机器人：好的，一张原味比萨马上就好！

你：不用了，谢谢，我更喜欢香肠比萨。

机器人：对不起，改成一张香肠比萨！

你：事实上，比起香肠比萨，我更喜欢菠萝比萨。

机器人：我错了，改成菠萝比萨！

你：我已经说过了，比起菠萝比萨，我更喜欢原味比萨。

机器人无法提供让你满意的比萨，因为总会有另一种你更喜欢吃的比萨。机器人只能满足你的偏好中一致的部分，例如，假设比起没有比萨吃，你更喜欢有比萨吃。在这种情况下，一个乐于助人的机器

人可以提供给你三种比萨中的任意一种，从而让你避免“没有比萨吃”，而且让你在闲暇时可以思考你那令人讨厌的比萨口味偏好。

两个人的理性

理性主体的行为是为了最大化期望效用，这一基本思想非常简单，但实际上做起来出奇地复杂。然而，这一理论只适用于单一个体单独行动的情况。如果有多个个体，那么为一个人行为的不同结果分配概率的构想（至少在原则上）就变得有问题了。原因是，现在，世界的一部分，即另一个主体，正试图猜测你要做什么，反之亦然，所以我们不清楚如何给世界上这一部分人的行为分配概率。如果没有概率，那么将理性行为定义为最大化期望效用就无法适用。

一旦其他人出现，智能体就需要其他方法来做出理性的决策。这就是博弈论的用武之地。尽管它叫作博弈论，但它并不一定是通常意义上的博弈。这是一种将理性的概念拓展到有多个主体的情况的一般尝试。这对我们设定目标来说显然很重要，因为我们还没有计划建造生活在其他星系中无人居住的星球上的机器人，我们要把机器人放到我们居住的世界里。

为了说明为什么我们需要博弈论，让我们来看一个简单的例子：爱丽丝和鲍勃在花园踢足球（图3）。爱丽丝将要踢点球，鲍勃守门。爱丽丝准备把球踢向鲍勃的左边或右边。因为爱丽丝习惯用右脚，所以对爱丽丝而言，把球踢向鲍勃的右边更容易也更准确。因为爱丽丝射门的力道很猛，鲍勃知道他必须立即选择一个方向扑球，他没有时间等着看球往哪个方向飞。鲍勃可以这样推理：“如果爱丽丝向我右边射门，她得分的机会更大，因为她习惯用右脚射门，她会选择右边，所以我要向右扑救。”但是爱丽丝不是傻瓜，她可以想象鲍勃是这样想的，在这种情况下，她会朝鲍勃的左边射门。但是鲍勃也不是傻瓜，他也可以想象爱丽丝是这样想的，在这种情况下，他会向左扑救。但是爱丽丝依然不是傻瓜，可以想象鲍勃是这样想的……好的，

你明白了。换言之，如果爱丽丝有一个理性的选择，鲍勃就能找出这种选择，并预测它，阻止爱丽丝得分，所以这个选择从一开始就不可能是理性的。

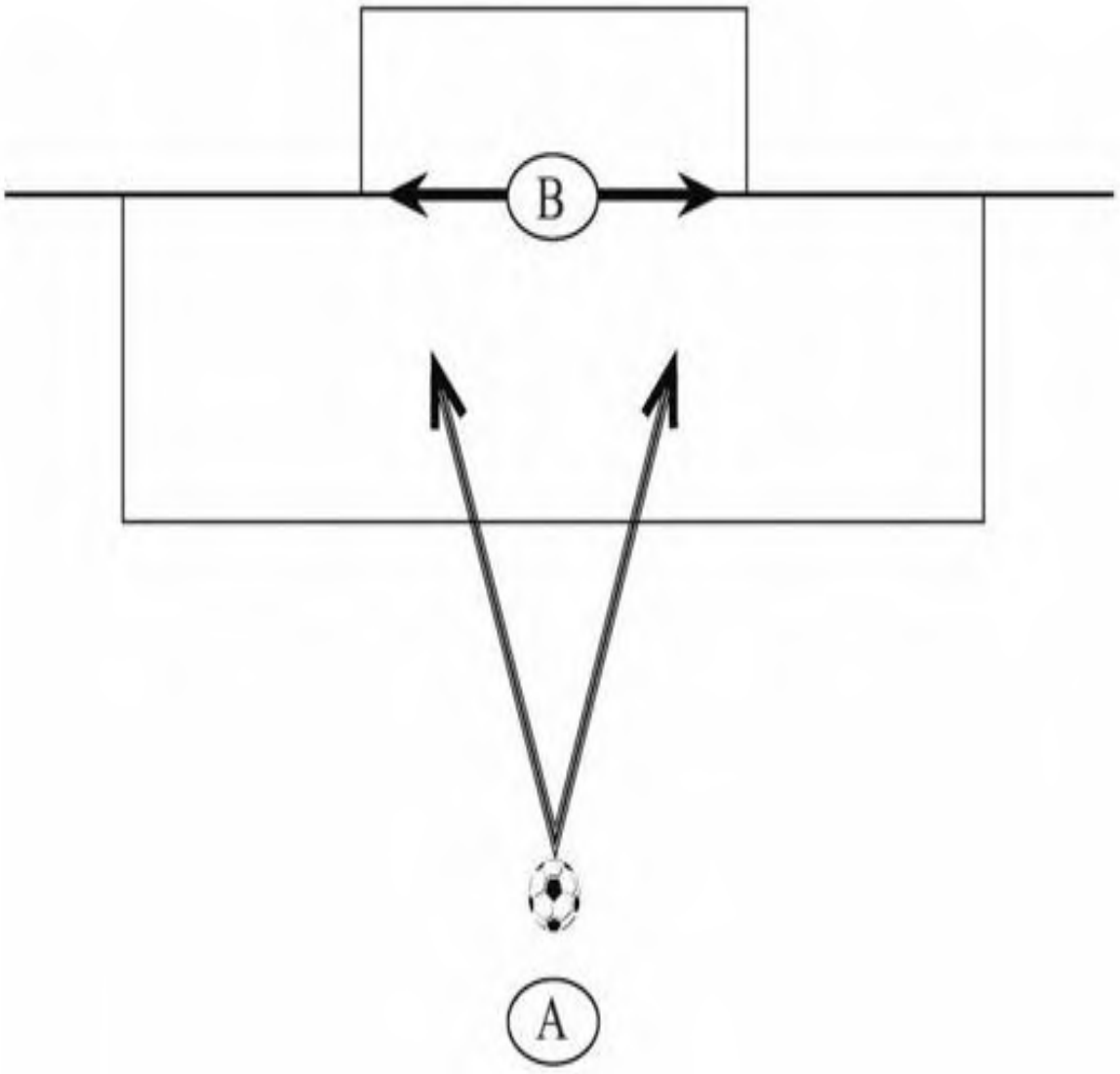


图3 爱丽丝（A）即将把点球踢向鲍勃（B）把守的球门

早在1713年，同样是在对赌博游戏的分析中，人们发现了这个难题的解决方案。[24]诀窍是不选择任何一个行动，而是选择随机策略。例如，爱丽丝可以选择“以55%的概率踢向鲍勃的右边，以45%的

概率踢向鲍勃的左边”。鲍勃可以选择“以60%的概率向右扑救，以40%的概率向左扑救”。在行动之前，每个人都会在心里抛一枚带有适当偏差的硬币，这样他们就不会泄露自己的意图。通过不可预测的行为，爱丽丝和鲍勃避免了前文中的矛盾。即使鲍勃知道爱丽丝要采取随机策略，没有水晶球的他也无能为力。

下一个问题是，概率应该是多少？爱丽丝选择55%对45%是合理的吗？具体的数值取决于爱丽丝向鲍勃右边射门时的准确度、鲍勃向右扑救时的救球能力等。（完整的分析请见注释。[25]）然而，通用标准很简单：

1. 假设鲍勃的策略是固定的，爱丽丝的策略是她能想到的最佳策略。

2. 假设爱丽丝的策略是固定的，鲍勃的策略是他能想到的最佳策略。

如果这两个条件都满足，我们就说策略处于均衡状态。为了纪念约翰·纳什（John Nash），这种均衡被称为纳什平衡。1950年，22岁的纳什证明了无论游戏规则如何，这样的均衡都存在于任意数量的理性主体之间。在与精神分裂症斗争了几十年之后，纳什最终康复了，并因为这一发现在1994年获得了诺贝尔经济学奖。

在爱丽丝和鲍勃的足球比赛中，只有一个均衡。在其他情况下，则可能有若干个均衡，所以纳什平衡的概念不同于期望效用决策，纳什平衡并不总是只有唯一的行动建议。

更糟糕的是，在某些情况下，纳什平衡会导致非常令人不快的结果。其中一个例子就是著名的“囚徒困境”，这是纳什的博士生导师阿尔伯特·塔克（Albert Tucker）在1950年提出来的。[26]囚徒困

境是现实世界中一些非常常见的情况的抽象模型，在这些情况下，虽然相互合作对所有相关的人都有好处，但人们仍然会选择同归于尽。

囚徒困境是这样一种博弈，假设爱丽丝和鲍勃是犯罪嫌疑人，正在分别接受审讯。每个人都有一个选择，要么向警方坦白并告发自己的同伙，要么拒绝招供。[27]如果二人都拒绝招供，他们都会被判处较轻的罪名，各服刑2年；如果两人都招供，他们将被判处更严重的罪名，各服刑10年；如果一个坦白，另一个拒绝招供，那么坦白的人会被释放，而拒绝招供的人则要服刑20年。

现在，爱丽丝的理由如下：“如果鲍勃要坦白，那么我也应该坦白（判10年总比20年好）；如果他拒绝招供，那么我还是应该坦白（获得自由总比在监狱里待2年好）。所以不管怎样，我都应该坦白。”鲍勃也是这样推理的。因此，他们最终都认了罪，各服刑10年，尽管他们共同拒绝招供只需各服刑2年。问题是，共同拒绝招供不是纳什平衡，因为每个人都有动机“叛变”，并通过坦白获得自由。

请注意，爱丽丝也可以这样推理：“无论我怎么推理，鲍勃也会这样做。所以我们最终都会做出同样的选择。既然共同拒绝招供比共同坦白好，我们就应该拒绝招供。”这种形式的推理认为，作为理性主体，爱丽丝和鲍勃的选择是相关的，而不是相互独立的。这只是博弈论学者试图为囚徒困境找到的不那么令人沮丧的众多解决方案之一。[28]

另一个令人不快的均衡失衡的著名例子是“公地悲剧”，英国经济学家威廉·劳埃德（William Lloyd）于1833年首次分析了这一理论，[29]但生态学家加勒特·哈丁（Garrett Hardin）于1968年将其命名为公地悲剧，并引起了全球范围的关注。[30]当几个人可以消费共享资源，比如共享牧场或鱼类资源时，悲剧就会发生。在没有任何社会或法律约束的情况下，自私（非利他）的主体之间唯一的纳什平衡就是每个主体尽可能多地消费，这会导致资源迅速枯竭。理想的解

决方案是，每个人都分享资源，使总消费是可持续的，这不是一种均衡，因为每个人都有作弊的动机，拿走比公平份额更多的资源，从而将损失强加于他人。当然，在实践中，人类有时确实会通过惩罚、建立配额或定价机制来避免这种悲剧的发生。人们可以这样做，因为他们不仅仅可以决定消费多少，也可以决定进行沟通。通过以这种方式拓展决策问题，我们找到了对每个人都更好的解决方案。

这些例子以及许多其他例子说明了一个事实：把理性决策理论拓展到多个主体身上会产生许多有趣而复杂的行为。这也是非常重要的，原因很明显，人类不止一个。不久之后，智能机器也不止一台。毋庸置疑，我们必须和智能机器相互合作，才能造福人类，而不是大家同归于尽。

计算机

合理地定义智能是创造智能机器的第一要素。第二要素是制造出一台可以实现这一定义的机器。这台机器就是计算机，其原因显而易见。它本可以是其他东西，例如，我们可以试图通过复杂的化学反应或操纵生物细胞来制造智能机器，[31]但是从最早的机械计算机开始，为计算而建造的设备在其发明者看来一直是智能的天然家园。

我们已经对计算机习以为常，甚至几乎没有注意到计算机那令人难以置信的能力。看看你的笔记本电脑、台式电脑或智能手机——一个可以输入字符的小盒子。仅仅通过输入，你就可以创建一个程序，把这个盒子变成某种新东西，也许它可以神奇地合成移动的图像，比如远洋船只撞上冰山，或是生活着高个子的蓝色外星人的星球；再输入一些，它便可以把英文翻译成中文；再输入一些，它就能听能说；再输入一些，它就能战胜国际象棋世界冠军。

单个盒子能够执行你可以想象得到的任意过程，这种能力叫作通用性，这是1936年艾伦·图灵（Alan Turing）首次提出的概念。[32]通用性意味着我们不需要单独的机器分别进行算术、机器翻译、下棋、语音理解，以及制作动画，一台机器就可以完成所有任务。你的笔记本电脑本质上与世界上最大的IT（信息技术）公司运营的大型服务器场，甚至那些配备了用于机器学习的高级专用TPU（张量处理单元）完全相同。本质上，所有尚未发明的未来计算设备也与你的笔记本电脑没什么不同。只要有足够的内存，你的笔记本电脑就可以完成相同的任务，只是需要更长的时间而已。

图灵介绍通用性的论文是有史以来最重要的论文之一。在文中，他描述了一个简单的计算设备，它可以接受其他任意计算设备的描述作为输入，也可以接受第二台设备的输入，并通过模拟第二台设备对其输入的操作来产生与之相同的输出。我们现在称第一台设备为“通用图灵机”。为了证明它的通用性，图灵为两种新的数学对象——机器和程序，引入了精确的定义。机器和程序一起定义了一系列事件，具体而言，就是机器及其内存中的一系列状态变化。

在数学史上，很少有新的对象出现。数学始于有记载的历史之初的数字。然后，大约在公元前2000年，古埃及人和巴比伦人开始研究几何对象（点、线、角、面积等）。中国数学家在约公元前1000年引入了矩阵，而集合作为数学对象直到19世纪才出现。图灵的新对象——机器和程序，也许是有史以来最强大的数学对象。具有讽刺意味的是，数学领域在很大程度上没有认识到这一点，从20世纪40年代开始，计算机和计算一直是大多数主要大学工程系的研究领域。

计算机科学这一新兴学科在接下来的70年里蓬勃发展，产生了大量新的概念、设计、方法和应用，以及全球最有价值的8家公司中的7家。

计算机科学的核心概念是算法，这是一种精确指定的计算方法。到目前为止，算法已经成为我们日常生活中非常熟悉的部分：便携式

计算器中的平方根算法接收一个数字作为输入，然后返回该数字的平方根作为输出；国际象棋算法提取棋子的位置，然后返回走子着法；寻路算法获取起点位置、目标位置和街道地图，然后返回从起点到目标的最快行驶路线。算法可以用英文或数学符号来描述，但要实现它们，必须用编程语言将其编码为程序。更复杂的算法可以通过使用更简单的“子程序”作为构件来建立。例如，无人驾驶汽车可以使用寻路算法作为子程序，这样它就知道要去哪里了。通过这种方式，人们可以一层一层地构建极其复杂的软件系统。

计算机硬件很重要，因为速度更快、内存更大的计算机可以让算法运行得更快，处理更多信息。这方面的进展虽然很出名，但仍令人难以置信。第一台商用电子可编程计算机费伦蒂马克一号（Ferranti Mark I）每秒可以执行大约1 000（10^3）条指令，拥有大约1 000字节的主存储器。截至2019年初，速度最快的计算机是美国田纳西州橡树岭国家实验室的Summit超级计算机，Summit每秒执行大约10^{18}条指令（速度比费伦蒂马克一号快1 000万亿倍），拥有2.5×10^{17}字节的存储器（比费伦蒂马克一号大250万亿倍）。这种进展源于电子器件的进步，甚至基础物理学的进步，微型化已经达到了令人难以置信的程度。

虽然拿计算机与大脑比较并不是特别有意义，但是Summit中的数字量略微超过了人脑的原始容量，如前所述，人脑有大约10^{15}个突触，大约0.01秒的“周期时间”，理论上每秒最多能完成大约10^{17}次“运算”。二者最大的不同是功耗，Summit消耗的电能是大脑的100万倍。

根据摩尔定律的经验，芯片上的电子元件数量每两年就会翻一番，这一现象预计将持续到2025年左右，尽管速度会略慢一些。近几年来，其速度一直受到硅晶体管快速开关产生的大量热量的限制。此外，电路的尺寸已经不能变得更小，因为导线和连接器宽度已经不超

过25个原子直径，厚度已经在5~10个原子直径范围内了（截至2019年）。2025年以后，我们将需要使用更奇特的物理现象——包括负电容器件[33]、单原子晶体管、石墨烯纳米管和光子来维持摩尔定律（或其后续定律）的发展。

除了加速推进通用计算机之外，另一种可能是构建专门用于执行某一类计算的专用设备。例如，谷歌的TPU被设计用来执行某些机器学习算法所需的计算。一个TPU Pod[34]（2018年版）每秒执行大约10^{17}次运算，几乎与Summit的计算能力相当，但前者的功耗是后者的1%，体积也是后者的1%。即使底层的芯片技术大致保持不变，人们也可以将这类机器制造得越来越大，为人工智能系统提供大量的原始计算能力。

量子计算则完全不同。它利用量子力学波函数的奇异特性来实现一些非同寻常的事情：用两倍数量的量子硬件，你可以完成两倍以上的计算！让我来简要介绍一下它的工作原理。[35]假设你有一个存储量子比特的微型物理器件。一个量子比特有两种可能的状态：0和1。在经典物理学中，量子比特器件必须处于这两种状态之一，而在量子物理学中，携带量子比特信息的波函数表示它同时处于两种状态。如果你有两个量子比特，就有4种可能的联合状态：00、01、10和11。如果波函数在两个量子比特之间相干、纠缠，这意味着没有其他物理过程来扰乱它，那么这两个量子比特则同时处于全部4种状态。此外，如果这两个量子比特连接到执行某些计算的量子电路中，那么计算过程将会同时处理全部4个状态。使用3个量子比特，就可以同时处理8个状态，以此类推。现在，一些物理限制使实际完成的工作量小于上述理论值，[36]但是我们知道，在一些重要的问题上，量子计算被证明比任何经典计算机都要高效。

截至2019年，拥有几十个量子比特的小型量子处理器的实验原型已经投入运行，但是没有什么有趣的计算任务可以让量子处理器的速度比经典计算机的速度更快。主要的困难是退相干，因为热噪声等过

程会破坏多量子比特波函数的相干性。量子科学家希望通过引入纠错电路来解决退相干问题，从而使计算中出现的任何错误都能通过一种投票过程被快速检测和纠正。不幸的是，纠错系统需要更多的量子比特来完成同样的工作：虽然拥有几百个完美量子比特的量子计算机与现有的经典计算机相比将会非常强大，但我们可能需要几百万个纠错量子比特来真正实现这些计算。从几十个量子比特发展到几百万个量子比特，需要很多年的时间。如果我们最终能做到这一点，那将完全改变我们通过纯粹的蛮力计算来解决问题的情况。[37]与其等待人工智能在概念上的真正进步，我们或许可以利用量子计算的原始力量绕过当前的“非智能”算法所面临的一些障碍。

计算的极限

即使在20世纪50年代，计算机也被大众媒体描述为“比爱因斯坦的头脑更快的超级大脑”。那么，我们现在终于可以说，计算机和人脑一样强大了吗？不能。我们把关注点放在原始计算能力上，完全没有抓住问题的关键。光靠速度是无法给我们带来人工智能的。在速度更快的计算机上运行糟糕的算法并不会使算法变得更优秀，这只意味着你会更快地得到错误的答案。（数据越多，错误答案出现的机会就越大！）速度更快的机器的主要作用是缩短实验时间，使研究进展更快。阻碍人工智能发展的不是硬件，而是软件。我们还不知道如何制造一台真正的智能机器，即便我们能把它造得如同宇宙那么大。

然而，假设我们成功开发出了正确的人工智能软件，那么物理学对计算机的能力还有什么限制吗？这些限制会妨碍我们拥有足够的计算能力来创造真正的人工智能吗？对第一个问题，答案是肯定的，但对第二个问题，答案是否定的。也就是说，限制是存在的，但是这些限制不可能阻止我们创造真正的人工智能。麻省理工学院的物理学家塞思·劳埃德（Seth Lloyd）基于对量子理论和熵的思考，估算了笔记本电脑的极限。[38]这些数字甚至会让卡尔·萨根（Carl Sagan）

大吃一惊：每秒10^{51}次运算，10^{30}字节的内存。这大约比Summit快10^{33}倍，内存大4万亿倍，而如前所述，Summit拥有比人脑更强大的原始计算能力。因此，如果你听到有人说，人类的思维代表了我们的宇宙在物理上可以达到的极限，[39]那么你至少应该要求他进一步澄清。

除了物理上的限制外，计算机的能力还受到其他限制，这些限制源于计算机科学家的工作。图灵证明了有些问题是任何计算机都无法判定的：这些问题定义明确，有答案，但是不存在总能找到答案的算法。他举了一个后来被称为“停机问题”的例子：一个算法能否判定一个给定的程序是否会陷入无法结束运行的“死循环”？[40]

图灵证明了，没有算法可以解决停机问题，[41]这对数学基础来说是非常重要的，但它与“计算机是否可以实现智能”这个问题无关。得出这个结论的一个原因是，同样的基本限制也适用于人脑。一旦你开始要求人脑进行精确的自我模拟、自我模拟、自我模拟……你就一定会遇到困难。就我个人而言，我非常确信自己无法做到这一点。

因此，专注于可判定的问题并没有对人工智能施加任何真正的限制。然而事实证明，可判定并不意味着容易。计算机科学家花费大量时间思考问题的复杂性，也就是思考，用最有效的方法解决一个问题需要多少计算量。这里举一个简单的例子：给定一个包含1 000个数字的列表，要求找出其中最大的数字。如果检查每个数字需要一秒钟，那么通过依次检查每个数字并跟踪最大的数字这种显而易见的方法来解决这个问题就需要1 000秒。有没有更快的方法？没有，因为如果一种方法没有检查列表中的某个数字，而这个数字有可能是最大的，那么这种方法就失败了。因此，找到最大数字所需的时间与列表的大小成正比。计算机科学家会说这个问题具有线性复杂度——这意味着它非常容易，然后科学家会去寻找更有趣的课题。

让理论计算机科学家感到兴奋的是，在最坏的情况下，许多问题具有指数复杂度。[42]这意味着两点：第一，我们所知道的所有算法都需要指数时间，即解决某些问题所需要的时间根据输入资料的大小而呈指数级增长；第二，理论计算机科学家非常确定不存在更有效的算法。

难度的指数级增长意味着问题在理论上或许是可解决的（也就是说，它们肯定是可判定的），但有时在实践中无法解决，我们称这些问题为难解问题。例如，判定给定地图是否可以仅用三种颜色着色，使得任意两个相邻区域都是不同的颜色。（众所周知，用四种不同的颜色着色总是可能的。）在100万个区域中，可能存在一些情况（不是所有情况，而是某些情况）需要大约$2^{1\ 000}$个计算步骤才能找到答案，这意味着在Summit超级计算机上运行这些计算大约需要10^{275}年，在塞思·劳埃德的终极物理笔记本电脑上需要10^{242}年。而宇宙的年龄大约是10^{10}年，这与前两者相比只是弹指一瞬。

这些难解问题的存在，是否让我们有理由认为计算机不能像人类那样智能？答案是否定的，我们也没有理由认为人类能解决难解问题。量子计算有所帮助（无论是在机器中还是在大脑中），但不足以改变基本结论。

复杂性意味着现实世界的决策问题——在人生的每一刻决定现在做什么的问题，是非常难的，所以人类和计算机都无法找到完美的解决方案。

这会产生两个后果：首先，我们预计，在大多数情况下，现实世界的决策充其量只能算是勉强过得去，远谈不上最优；其次，我们预计人类和计算机的许多“思维体系结构”——其决策过程的实际运作方式，将被设计成尽可能地克服复杂性，也就是说，虽然世界极其复杂，但我们仍有可能找到一些还算过得去的答案。总之，无论未来的

机器有多智能、多强大，我们都预计上述两点仍然是正确的。机器的能力可能比我们强大得多，但仍远非绝对理性。

智能计算机

亚里士多德等人对逻辑的研究，为理性思维提供了精确的规则，但我们不知道亚里士多德是否曾考虑过让机器实现这些规则的可能性。13世纪，颇有影响力的加泰罗尼亚哲学家、神秘主义者拉蒙·鲁尔（Ramon Llull）更接近这一目标，他制造出了刻有符号的纸轮，可以用一系列符号表示声明或陈述。17世纪伟大的法国数学家布莱兹·帕斯卡是第一个研制出真正实用的机械计算器的人。虽然这个计算器只会计算加减法，主要是他父亲在税收办公室使用，但它启发帕斯卡写道：“计算器产生的效果似乎比动物的一切行为都更接近于思想。”

19世纪，科技取得了巨大的飞跃，英国数学家、发明家查尔斯·巴贝奇（Charles Babbage）设计了分析机，后来图灵将这种机器定义为可编程通用计算机。巴贝奇的工作得到了浪漫主义诗人、冒险家拜伦勋爵的女儿艾达·洛夫莱斯伯爵夫人（Ada Lovelace）的帮助。巴贝奇希望用分析机来计算精确的数学表和天文表，而洛夫莱斯理解了分析机的真正潜力，[43]并在1842年将其描述为“一种会思考的……或一台会推理的机器”，可以推理“宇宙中的所有问题”。因此，创建人工智能的基本概念元素已经就位！从这一点来看，攻克人工智能无疑只是时间问题……

不幸的是，在很长一段时间里，分析机从未被制造出来，洛夫莱斯的想法也基本上被人们遗忘了。在1936年图灵的理论工作和随后的第二次世界大战的推动下，通用计算机器终于在20世纪40年代成为现实，有关创造智能的想法随即出现。图灵在1950年发表的论文《计算机器与智能》[44]是许多关于智能机器可能性的早期作品中最著名的

一篇。怀疑者已经断言，机器永远不可能做到你能想到的几乎一切事情，图灵驳斥了这些断言。他还提出了一种针对智能的操作测试，叫作“模仿游戏”，后来被简称为“图灵测试”。这项测试衡量的是机器的行为，具体而言，就是机器欺骗人类审问者的能力，机器会试图让审问者以为跟他们对话的机器是人类。

模仿游戏在图灵的论文中扮演了特殊的角色，这是一个反驳那些怀疑者的思想实验，怀疑者认为机器不能出于正确的理由，以正确的方式，用正确的意识来思考。图灵希望把争论引向一个问题：机器是否能以某种特定方式运行？如果它真的可以，比如说，它能够理智地谈论莎士比亚的十四行诗及其含义，那么人们对人工智能的怀疑就无法真正持续下去。与通常的解释相反，我怀疑这项测试是为了给智能下一个真正的定义，即当且仅当机器通过图灵测试时，它才是智能的。事实上，图灵写道：“难道机器不能完成一些应该被描述为思考的行为，但与人类思考的行为有很大不同的事情吗？”人们没有将这项测试视为“给人工智能下定义”的一个原因是，它是一个糟糕的定义。因此，主流人工智能研究人员几乎没有为了通过图灵测试而付出任何努力。

图灵测试对人工智能没有什么用处，因为它下的是一个非正式的、高度偶然的定义：它取决于人类思维极其复杂和在很大程度上未知的特征，这些特征来自生物学和文化。人们没有办法拆解定义，并根据它创建可以被证明通过测试的机器。相反，人工智能专注于理性行为，正如前面所描述的那样：一台机器之所以是智能的，是因为根据它的感知，它所做的事情有望实现它的目标。

最初，像亚里士多德一样，人工智能研究人员用是否实现一个目标来确定“它的需求”。这些目标可以在玩具世界中实现，比如“十五拼图”，该玩具的目标是将编号为1到15的乱序的拼块按顺序排列在一个小正方形托盘中（可以是模拟的，也可以是真实的拼块和托盘）；20世纪70年代初，加州斯坦福研究所的Shakey机器人可以将大

块积木推到理想的位置，而爱丁堡大学的Freddy机器人可以将木船的零部件组装起来。所有这些工作都是通过有逻辑性的问题解决程序和规划系统来构建和执行有保证的计划，从而实现目标的。[45]

到了20世纪80年代，很明显，仅靠逻辑推理是不够的，没有任何计划能保证一定让你赶上飞机。逻辑需要确定性，而现实世界根本不提供确定性。与此同时，2011年的图灵奖得主、以色列裔美国计算机科学家朱迪亚·珀尔（Judea Pearl）一直在研究基于概率论的不确定性推理方法。[46]人工智能研究人员逐渐接受了珀尔的想法，他们采用了概率论和效用理论的工具，从而将人工智能与统计学、控制理论、经济学和运筹学等其他领域联系起来。这一变化标志着一些观察家所说的“现代人工智能”的开始。

智能体和环境

现代人工智能的核心概念是智能体，即能感知和行动的东西。智能体是随着时间的推移而发生的过程，从这个意义上讲，它是一个感知输入流被转换成一个动作流的过程。例如，假设这个智能体是一辆带我去机场的无人驾驶出租车。它的输入可能来自以每秒30帧的速度运行的8台RGB（色彩模式）摄像机，每帧图像大约由750万像素组成，每个像素都有3个彩色通道的图像强度值，输入总共超过每秒5千兆字节数据。（来自视网膜上2亿个感光细胞的数据流甚至更大，这在一定程度上解释了为什么视觉信息占据了人脑的很大一部分。）这辆出租车每秒还能从加速度计和GPS（全球定位系统）的数据中获得100次数据。数十亿个晶体管（或神经元）的庞大计算能力将这些令人难以置信的原始数据转化为流畅、有效的驾驶行为。出租车的动作包括以每秒20次的速度向方向盘、刹车和油门发送电子信号。（对有经验的人类司机而言，这些混乱的动作大多是无意识的，他可能只知道做出了决定，如“超车”或“停车加油”，但他的眼睛、大脑、神经和肌肉仍然在做其他事情。）对国际象棋程序而言，输入大多只是时钟的计

时数，它偶尔会收到对手走子和新的棋局开始的通知，而当程序思考时，其动作大多是什么也不做，偶尔选择走一步棋并通知对手。对于个人数字助理，如Siri（苹果语音助手）或Cortana（微软人工智能助理），输入不仅有来自麦克风的声音信号（每秒采样4.8万次）和来自触摸屏的信号，还有它访问的每个网页的内容，而动作则包括讲话和把材料显示在屏幕上。

我们构建智能体的方式取决于我们面临的问题的性质。反过来，这又取决于三件事：第一，智能体的工作环境——棋盘与手机或拥挤的高速公路截然不同；第二，将智能体与环境连接起来的观察和动作——例如，Siri能访问手机的摄像头，它才能“看”；第三，智能体的目标——教对手下棋与赢棋是截然不同的任务。

我举一个例子来说明智能体的设计是如何依赖于这些东西的。如果国际象棋程序的目标是赢棋，它只需要考虑当前的棋局，不需要任何对过去事件的记忆。[47]如果其目标是当国际象棋教练，程序就要不断更新模型，从而确定学生对国际象棋各个方面的了解程度，以便提供有用的建议。换言之，对国际象棋教练而言，学生的想法是与环境相关的一部分。此外，与棋局不同，学生的想法是环境中的一部分，无法被直接观察到。

影响智能体设计问题的特征至少包括以下几点：[48]

- 环境是完全可观察的（如在国际象棋中，输入提供了对当前环境状态的所有相关方面的信息），还是部分可观察的（如在驾驶中，视野是有限的，车辆是不透明的，其他司机的意图是未知的）。
- 环境和动作是独立的（如国际象棋），还是实际上连续的（如驾驶）。
- 环境中包含其他智能体（如国际象棋和驾驶），还是不包含其他智能体（如在地图上查找最短路线）。

- 环境的“规则”和“物理定律”所规定的行动结果是可预测的（如国际象棋），还是不可预测的（如交通和天气），以及这些规则和定律是已知的还是未知的。

- 环境是否在动态变化。这会影响决策时间是受到严格限制（如驾驶），还是不受严格限制（如税收战略优化）。

- 时间跨度是长还是短。我们要在这个时间跨度里根据目标来衡量决策质量。这一时间跨度可能非常短（如紧急刹车），可能持续中等时长（如国际象棋，棋局持续约100步），或很长（例如送我去机场的出租车如果每秒做出100个决定，它可能需要数十万个决策周期）。

可以想象，这些特征带来了各种类型的令人困惑的问题。只要将上面列出的选项数量相乘，就会得到192种类型。人们可以找到所有类型的真实问题案例。有些类型通常在人工智能以外的领域进行研究，例如，设计一个保持水平飞行的无人驾驶仪是一个短期、连续、动态的问题，这通常属于控制理论的研究范畴。

显然，某些类型的问题比其他类型的问题更容易。人工智能已经在棋类游戏和拼图游戏这样的问题上取得了很大进展，这些问题是可观察的、独立的、确定性的，并且有已知的规则。对于较简单的问题类型，人工智能研究人员已经开发出了相当通用且有效的算法，并具有扎实的理论理解。通常，机器在这类问题上的表现要优于人类。我们可以说，一个算法是通用的，因为我们可以用数学证明它，该算法在整个问题类别中，以合理的计算复杂性给出最优或接近最优的结果，而且因为它在实践中可以很好地解决这类问题，所以不需要针对任何特定问题进行修改。

像《星际争霸》这样的电子游戏比棋类游戏要困难一些：电子游戏包含数百个移动部件和数千个步骤的时间跨度，而且地图在任何给定时间内都只有部分可见。在每个点上，玩家都可能至少有10^{50}种选

择，而围棋大约有10^2种选择。[49]另外，电子游戏的规则是已知的，而世界是离散的，只有几种类型的对象。截至2019年初，机器已经达到了《星际争霸》专业玩家的水平，但还没有准备好挑战最优秀的人类玩家。[50]更重要的是，要达到这一点，机器还需要针对特定问题做大量努力。通用方法还没有为《星际争霸》做好准备。

管理政府或教授分子生物学之类的问题要困难得多。这些环境很复杂，大部分是不可观察的（一个国家的状态，一个学生的思想状态），还有更多的对象和对象类型，对动作没有明确定义，规则大多是未知的，有大量的不确定性，以及非常长的时间跨度。我们有想法和现成的工具来分别处理这些特征，但到目前为止，还没有同时处理所有特征的通用方法。当我们为这类任务构建人工智能系统时，它们往往需要开展大量针对特定问题的工程工作，而且往往非常不稳定。

当我们设计出针对给定类型的较困难问题的有效方法，或者设计出需要较少和较弱假设的方法，以便适用于更多问题时，通用性的进展就出现了。通用人工智能将是一种适用于所有问题类型的方法，并且在做出很少假设的情况下，它能有效地处理大而难的实例。这就是人工智能研究的终极目标——一个不需要针对具体问题的工程学系统，人们可以简单地要求它去管理政府或教授分子生物学。它会从所有可用的资源中学习它需要学习的东西，在必要时提出问题，并开始制订和执行有效的计划。

这种通用的方法还不存在，但我们正在接近它。也许令人惊讶的是，在实现通用人工智能方面取得的许多进步，都来自与构建可怕的通用人工智能系统无关的研究。这些进步来自对“工具人工智能”或“狭义人工智能”的研究，即为特定问题设计的完美、安全、乏味的人工智能系统，例如下围棋或识别手写数字。人们通常认为这种人工智能研究没有风险，因为它是针对具体问题的，与通用人工智能无关。

这种看法源于对这些系统的工作类型的误解。事实上，对工具人工智能的研究能够而且常常会在实现通用人工智能方面取得进展，特别是当它由别具慧眼的研究人员攻克了当前的通用方法所不能解决的问题的时候。在这里，“别具慧眼”意味着，解决方案所采用的路径不仅仅是对一个聪明人在某种情况下会如何行动的临时编码，而是试图使机器具备自己找出解决方案的能力。

例如，当谷歌DeepMind的AlphaGo团队成功创建了举世无双的围棋程序时，他们并没有把注意力局限在围棋方面的工作上。我的意思是，他们并没有写很多专用于围棋的代码来说明在不同的棋局中该怎么做。他们没有设计只适用于围棋的决策程序。相反，他们对两种非常通用的技术进行了改进：一是用前向搜索（lookahead search）来做决策，二是用强化学习来学习如何评估棋局。因此，他们的围棋程序达到了超越人类的水平。这些改进适用于许多其他问题，包括远至机器人领域的问题。更让人印象深刻的是，AlphaGo的一个叫AlphaZero（阿尔法元）的版本学会了在围棋中打败AlphaGo，还打败了Stockfish（世界上最好的国际象棋程序，水平远远高于任何人类）和Elmo（世界上最好的日本将棋程序，水平同样远远高于任何人类）。AlphaZero在一天内完成了这一切。[51]

20世纪90年代，在识别手写数字的研究方面，通用人工智能也取得了重大进展。AT&T（美国电话电报公司）实验室的杨立昆（Yann LeCun）团队在编写识别“8”的特殊算法时，并没有使用搜索曲线和回环的方法，他们改进了现有的神经网络学习算法来生成卷积神经网络（convolutional neural network）。反过来，用标记的样本对这些网络进行适当训练后，网络就展现出有效的识别字符的能力。同样的算法可以学习识别字母、形状、停车标志、狗、猫和警车等。在深度学习的课题下，他们革新了语音识别和基于机器视觉的物体识别。这些也是AlphaZero以及当前大多数无人驾驶汽车项目的关键组成部分之一。

如果你仔细想想就会发现，在解决具体任务的狭义人工智能项目中，出现朝向通用人工智能的进展其实不足为奇，这些任务为人工智能研究人员提供了一些研究思路。（人们不说“盯着窗外发呆是发明之母”是有原因的。）同时，重要的是，我们要了解自己已经取得了多少进展，以及边界在哪里。当AlphaGo战胜李世石以及其他所有顶尖围棋选手时，许多人认为，机器从零开始学习，在一项已知的即使对于极度聪明的人类也非常困难的任务中战胜了人类，所以这便是终结的开始，人工智能接管世界只是时间问题。当AlphaZero在国际象棋和将棋中也像在围棋中一样获胜时，一些之前的怀疑者可能也已经被说服了。但AlphaZero有严格的限制：它只适用于独立的、可观察的、具有已知规则的双人游戏。这种方法根本不适用于驾驶、教学、管理政府或接管世界。

这些关于机器能力的清晰界限意味着，当人们谈论“机器智商”迅速增长并有可能超过人类智商时，他们是在胡说八道。在某种程度上，智商的概念在应用于人类时是有意义的，因为人类的能力往往在广泛的认知活动中相互关联。试图赋予机器智商，就像试图让四条腿的动物参加人类十项全能比赛一样。的确，马跑得快，跳得高，但它们在撑竿跳高和掷铁饼项目中会碰到很多麻烦。

目标和标准模型

如果从外部看一个智能体，重要的是看它从接收的输入流中生成的动作流。如果从内部看，动作必须由智能体程序选择。可以说，人类天生就有一个智能体程序，这个程序随着时间的推移学会了在大量任务中合理、成功地行动。到目前为止，人工智能的情况并非如此：我们不知道如何构建一个通用人工智能程序来完成一切事情，因此我们针对不同类型的问题构建不同类型的智能体程序。我需要简要解释一下这些不同的智能体程序是如何工作的。对于感兴趣的读者，本书末尾的附录给出了更详细的解释。这里的重点是如何在这些不同类型

的智能体中将标准模型实例化，换言之，如何指定目标，并将其传达给智能体。

传达目标最简单的方式是以“终点”的形式。当你进入无人驾驶汽车，触摸屏幕上“家”的图标时，汽车就会以此为目标，继续规划和执行路线。世界的状态要么是到达终点（是的，我在家），要么是没有到达终点（不，我不住在旧金山机场）。在人工智能研究的古典主义时期，即在20世纪80年代，不确定性成为一个主要问题之前，大多数人工智能研究都假设世界是完全可观察的和确定的，终点作为一种明确目标是有意义的。人们有时还需要用到一个成本函数来评估解决方案，因此最优解决方案是在到达终点的同时使总成本最小化。对汽车而言，这可能是内置的——也许一条路线的成本是时间和燃料消耗的某个固定组合，或者人类可以选择在二者之间进行权衡取舍。

实现这些目标的关键是这样一种能力：“在心中模拟”可能行动的效果。有时这也被称为前向搜索。你的无人驾驶汽车有一张内置地图，所以它知道从旧金山沿着海湾大桥向东行驶可以到达奥克兰。前向搜索算法起源于20世纪60年代，它通过前向搜索许多可能的动作序列来找到最优路径.[52].（附录A）。这些算法构成了现代基础设施中无处不在的一部分，它们不仅提供行车路线，还提供航空旅行解决方案、机器人装配、建筑规划和物流配送。前向搜索的概念同样适用于井字游戏、国际象棋和围棋等游戏，这些游戏的目标是根据游戏对胜利的特定定义来努力取胜，只不过人们还需要对算法做一些修改来处理对手的赖皮行为。

前向搜索算法对于具体的任务非常有效，但不是很灵活。例如，AlphaGo“了解”围棋的规则，但这种“了解”只是建立在它有两个用C++等传统编程语言编写的子程序的意义上：一个子程序生成所有可能的合规走法，另一个子程序对目标进行编码，确定给定的局面是赢还是输。为了让AlphaGo玩不同的游戏，必须有人重写所有这些C++代码。此外，如果你给它一个新的目标，例如访问围绕着比邻星运行的

系外行星，它将像原来下围棋那样探索数十亿步棋，徒劳地试图找到实现目标的方法。它无法查看C++代码并确定一个显而易见的事实：没有任何围棋着法可以让你到达系外行星。AlphaGo的知识基本上被锁在一个黑盒里。

1958年，在达特茅斯夏季会议开创了人工智能概念的两年之后，约翰·麦卡锡提出了一种更通用的方法，打开了黑盒：编写通用的推理程序。这个程序可以吸收一切主题的知识，并能回答一切可以回答的问题。[53]一种特殊的推理是亚里士多德所建议的那种实践推理："完成动作A、B、C……将会实现目标G。"目标可以是任意事情，例如，在我回家之前确保房子整洁，在不丢掉马的情况下赢一盘国际象棋，将我的纳税金额减少50%，访问比邻星等。很快，麦卡锡的这类新程序被称为"基于知识的系统"。[54]

要使"基于知识的系统"成为可能，我们需要回答两个问题。首先，我们如何将知识存储在计算机中？其次，计算机如何根据这些知识进行正确推理并得出新的结论？幸运的是，早在计算机出现之前，古希腊哲学家，尤其是亚里士多德，就为这些问题提供了基本答案。的确，如果亚里士多德有机会使用电脑（还有电力），那么他很有可能成为人工智能研究者。麦卡锡重申了亚里士多德的答案——使用形式逻辑作为知识和推理的基础。

在计算机科学中有两种真正重要的逻辑。第一种被称为命题逻辑（附录B）或布尔逻辑，古希腊人以及古代中国和印度的哲学家都知道它。它与构成计算机芯片电路的与门、非门等逻辑门相同。从字面意义上讲，现代CPU（中央处理器）就是用命题逻辑语言编写的一个非常庞大的数学表达式，长达数亿页。第二种逻辑是麦卡锡提议用于人工智能的逻辑，叫作一阶逻辑（附录B）。一阶逻辑语言远比命题逻辑语言更有表现力，这意味着有些东西用一阶逻辑可以很容易地表达出来，但用命题逻辑很难表达。例如，围棋规则在一阶逻辑中的表达只需要一页，而在命题逻辑中则长达数百万页。同样，我们可以很容易

地表达关于国际象棋、英国公民身份、税法、买卖、搬家、绘画、烹饪，以及我们的常识世界中许多其他方面的知识。

因此，从原则上讲，使用一阶逻辑推理的能力让我们朝着实现通用智能的目标迈出了一大步。1930年，才华横溢的奥地利逻辑学家库尔特·哥德尔（Kurt Gödel）发表了他著名的“完全性定理”，[55]证明了存在如下性质的算法：[56]

> 对于一阶逻辑可表达的任何知识集合与任何问题，如果答案存在，那么算法将告诉我们问题的答案。

这是一个相当令人难以置信的定理。它的意思是，例如，我们可以告诉系统围棋的规则，它会告诉我们（如果我们等待足够长的时间）是否存在必胜的开局着法；我们可以告诉系统当地的地理情况，它会告诉我们去机场的路；我们可以告诉系统关于几何学、运动和餐具的知识，它会告诉机器人如何摆好餐桌。总而言之，只要我们给定任意可实现的目标和关于其行动所产生的影响的充分知识，智能体就可以使用该算法制订一个计划，它可以执行这个计划来实现目标。

必须指出的是，哥德尔实际上并没有提供算法，他只是证明了算法的存在。20世纪60年代初，真正的逻辑推理算法开始出现，[57]麦卡锡基于逻辑的通用智能系统的梦想似乎触手可及。加州斯坦福研究所的Shakey项目是世界上第一个大型移动机器人项目，它就是基于逻辑推理的（图4）。Shakey从其人类设计者那里获得一个目标，使用视觉算法创建描述当前情况的逻辑断言，执行逻辑推理，以得出可以实现目标的有保证的计划，然后执行这个计划。Shakey如今证明了亚里士多德对人类认知和行为的分析至少部分是正确的。



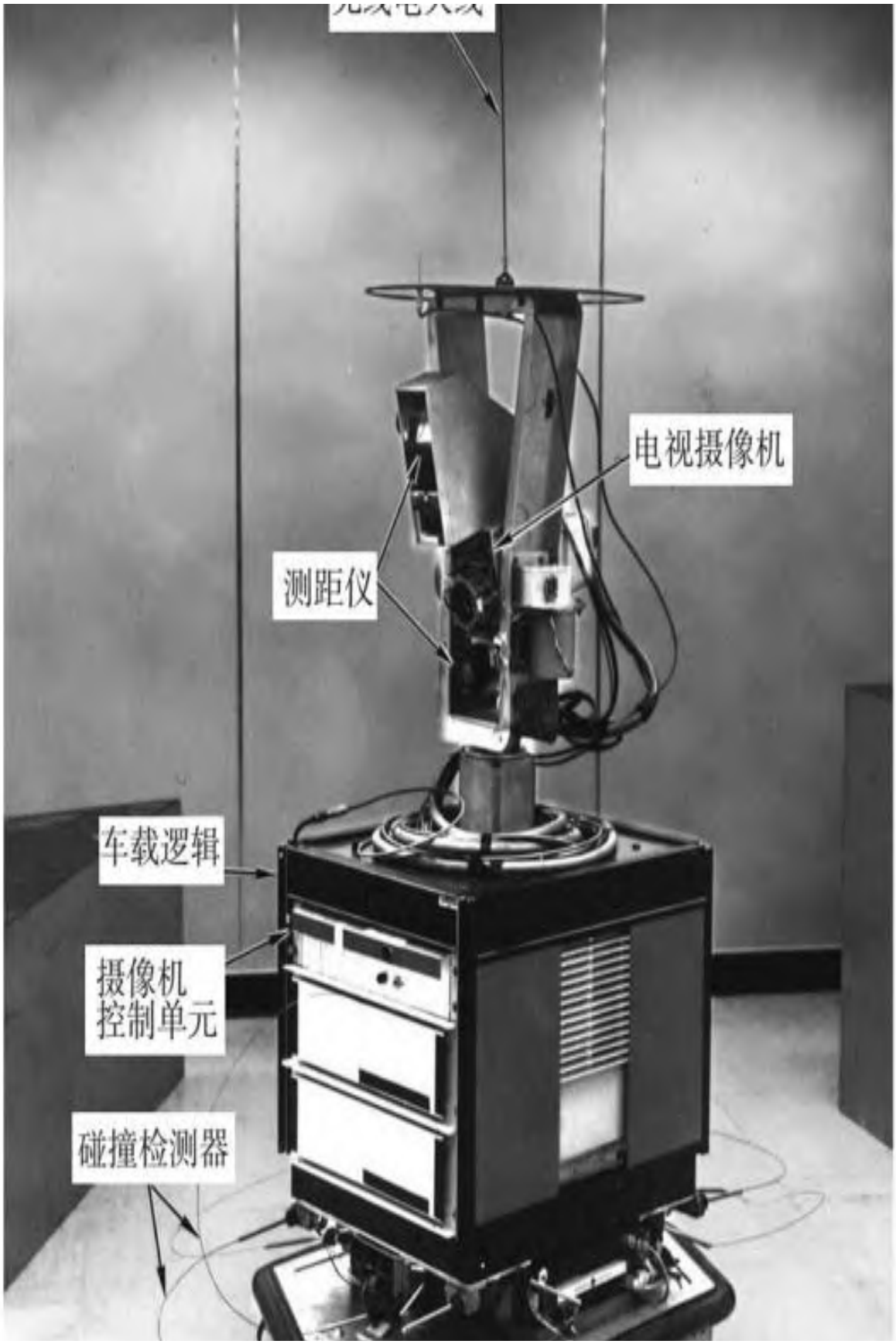

电视摄像机
测距仪
车载逻辑
摄像机
控制单元
碰撞检测器



图4 1970年左右的Shakey机器人（背景中是Shakey在它的房间里推动的一些物品）

不幸的是，亚里士多德（和麦卡锡）的分析离完全正确还有很长的路要走。主要的问题是无知——我得赶紧解释一下，不是亚里士多德或麦卡锡无知，而是所有人类和机器无知，无论是现在还是未来的人类和机器都是无知的。在我们的知识中，只有很少一部分是绝对确定的。特别是，我们对未来知之甚少。对一个纯粹的逻辑系统而言，无知是一个无法克服的问题。如果我问，如果我在飞机起飞前三个小时出发，我能准时到达机场吗？或者，我可以买彩票中奖，然后用奖金买房子吗？在以上两种情况下，正确答案都是“不知道”。原因是，对于每个问题，“是”和“否”的回答在逻辑上都是可能的。在实际情况中，一个人永远不可能绝对肯定任何经验性问题，除非他已经知道答案。[58]幸运的是，确定性对行动来说是完全不必要的：我们只需要知道哪个行动是最好的，而不是哪个行动肯定会成功。

不确定性意味着“输入机器的目标”通常是一个无法精确描述的目标，所以我们不能让机器不惜一切代价去实现它。此时不再有“实现目标的动作序列”这样的东西，因为任何动作序列都会产生多种可能的结果，其中一些不会实现目标。成功的可能性真的很重要：提前三小时前往机场可能意味着你不会错过航班，购买彩票可能意味着你会赢得足够的钱去买一套新房子，但二者的可能性大小非常不同。寻

找用最大的可能性实现目标的计划，依然无法拯救目标。最大限度地提高及时到达机场并赶上航班的可能性的计划，可能包括提前几天离开家、组织武装护送、安排许多替代交通工具以防出现故障等。人们不可避免地需要考虑不同结果的相对可取性以及它们发生的可能性。

然后，我们可以用效用函数，而不是用目标，来描述不同结果或状态序列的可取性。通常，状态序列的效用表现为对序列中每个状态的奖励总和。给定机器一个由效用函数或奖励函数定义的目的，它的目标是产生最大化其期望效用或期望奖励总和的行为，即对可能的结果进行平均，并对其概率进行加权。现代人工智能在一定程度上重新启动了麦卡锡的梦想，除了用效用和可能性代替目标和逻辑之外。

伟大的法国数学家皮埃尔·西蒙·拉普拉斯（Pierre-Simon Laplace）在1814年写道："概率论不过是把常识简化为微积分。"[59]然而，直到20世纪80年代，人们才为概率知识开发了一种实用的形式语言和推理算法。这就是由朱迪亚·珀尔引入的贝叶斯网络语言（附录C）。粗略地讲，贝叶斯网络是命题逻辑的概率学表亲。一阶逻辑也有概率学表亲，包括贝叶斯逻辑[60]和各种各样的概率编程语言。

贝叶斯网络和贝叶斯逻辑是以英国牧师托马斯·贝叶斯（Thomas Bayes）的名字命名的，他对现代思想的持久贡献——现在被称为"贝叶斯定理"，是在他去世后不久，由他的朋友理查德·普莱斯（Richard Price）于1763年发表的。[61]正如拉普拉斯所说的那样，贝叶斯定理的现代形式以一种非常简单的方式描述了先验概率——一个人对一组可能性假设的初始可信度，是如何由于观察到一些证据而变成后验概率的。随着更多新证据的到来，后验成为新的先验，贝叶斯定理中更新的过程无限重复。这个过程非常基础，以至现代的理性思想，比如期望效用最大化，有时被称为"贝叶斯理性"。它假设一个理性的智能体可以获得一个后验概率分布，这一后验概率

分布是基于当前世界可能的状态以及基于过去所有经验的关于未来的假设。

运筹学、控制理论和人工智能领域的研究人员也开发了各种各样的算法，用于在不确定的情况下做出决策，其中一些算法可以追溯到20世纪50年代。这些所谓的“动态规划算法”是前向搜索和规划的概率学表亲，可以为金融、物流、交通等各种实际问题提供最优或接近最优的行动计划，在这些问题中，不确定性起着重要作用（附录C）。目标被以奖励函数的形式输入这些机器中，输出则是一个策略，它为智能体可能进入的每种可能状态指定一个动作。

对于像西洋双陆棋和围棋这样的复杂问题，其状态的数量多到不计其数，而且只有在游戏结束时才会有奖励，因此前向搜索是行不通的。相反，人工智能研究人员开发了一种名为强化学习的方法。强化学习算法从环境中奖励信号的直接经验中学习，就像婴儿可以根据直立的正面奖励和跌倒的负面反馈学会站立一样。与动态规划算法一样，输入强化学习算法中的目标是奖励函数，而且强化学习算法会总结出一种状态值（有时是动作值）的估算器。该估算器可以与时间跨度相对较短的前向搜索相结合，产生高水平的行为。

第一个成功的强化学习系统是阿瑟·塞缪尔的西洋跳棋程序，1956年，该程序在电视上播出时引起了轰动。这个程序基本上是从零开始学习的，方法是自我对弈和观察赢棋的奖励与输棋的惩罚。[62]1992年，格里·泰索罗（Gerry Tesauro）将同样的想法应用到西洋双陆棋中，该程序在经过150万盘的对弈后达到了世界冠军的水平。[63]从2016年开始，DeepMind的AlphaGo及其后续程序使用强化学习和自我对弈，战胜了围棋、国际象棋和日本将棋中最优秀的人类棋手。

强化学习算法还可以学习如何基于原始感知输入选择动作。例如，DeepMind的DQN系统完全从零开始学习玩49种不同的雅达利电子游戏，包括《乒乓球》《高速公路》《太空入侵者》等。[64]它只使用

屏幕像素作为输入，游戏分数作为奖励信号。在大多数游戏中，虽然DQN系统并没有时间、空间、物体、运动、速度以及射击的先验概念，但是该系统在经过学习之后，水平都超过了职业玩家。除了获胜之外，人们很难搞清楚DQN到底在做什么。

如果一个新生儿在出生后的第一天就学会了以超越人类的水平玩几十种电子游戏，或者成了围棋、国际象棋和将棋的世界冠军，那么我们可能会怀疑他被恶魔附体或被外星人干预了。然而请记住，所有这些任务都比现实世界简单得多：它们是完全可观察的，它们涉及的时间跨度很短，并且它们具有相对较小的状态空间和简单的、可预测的规则。如果我们放宽这些条件中的任何一个，都意味着标准方法将失效。

另外，目前的研究恰恰旨在超越标准方法，使人工智能系统可以在类别更广的环境中运行。例如，在我写前一段的那天，OpenAI（一家人工智能公司）宣布它的5个人工智能程序已经学会了在《Dota 2》游戏中击败经验丰富的人类团队。（为包括我在内的外行人解释一下：《Dota 2》是《Dota》的升级版，是魔兽家族的一款实时战略游戏，它是目前最赚钱、最具竞争性的电子竞技项目，奖金高达数百万美元。）《Dota 2》涉及沟通、团队合作、准连续时间和空间。游戏持续数万个时间步，因此某种程度的行为等级组织是必不可少的。比尔·盖茨将OpenAI的这一声明描述为“推进人工智能的一个巨大里程碑”。[65]几个月后，该程序的更新版本击败了世界顶级职业《Dota 2》团队。[66]

像围棋和《Dota 2》这样的游戏是强化学习方法的一个很好的试验场，因为奖励函数是伴随着游戏规则的。然而，在现实世界中就不那么方便了，已经有几十个例子说明对奖励错误的定义会导致奇怪且出乎意料的行为。[67]有些行为是无害的，比如模拟进化系统，它本应进化出快速移动的生物，实际上却产生了个头很高的生物，通过跌

跟头快速移动。[68]其他行为就没那么无害了，比如社交媒体的点击优化器，它把我们的世界搞得一团糟。

最后一类智能体程序是最简单的，它就是将感知与行动直接联系起来的程序，不需要任何中间的思考或推理。在人工智能中，我们称这种程序为“反射型智能体”，反射指的是人类和动物表现出的低级神经反射，这种反射不受思想的调节。[69]例如，人类的眨眼反射将视觉系统中的低层处理电路的输出直接与控制眼睑的运动区域相连接，因此人类的视野中出现任何快速逼近的东西都会导致眨眼。现在你可以试着用手指轻戳自己的眼睛来测试。我们可以把这种反射系统看作一个简单的“规则”，其形式如下：

如果<视野中有快速逼近的东西出现>，就<眨眼>。

眨眼反射并不“知道自己在做什么”：目标（保护眼球免受异物伤害）没有任何体现；知识（快速逼近的东西对应于接近眼睛的物体，而接近眼睛的物体可能会伤害眼睛）也没有任何体现。因此，当你的非反射部分想要滴入眼药水时，反射部分仍然会眨眼。

另一种常见的反应是，当前面的汽车意外停车或行人进入道路时，车辆启用紧急制动。快速决定是否需要刹车并非易事：2018年，一辆处于无人驾驶模式的测试车撞死了一名行人，优步解释说，“当处于计算机控制时，车辆不启用紧急制动操作，以降低车辆行为不稳定的可能性”。[70]在这里，人类设计者的目标很明确——不要撞死行人，但是智能体的策略（如果被激活）执行得不正确。同样，智能体也没有体现出目标。直到今天，没有一辆无人驾驶汽车知道人们不喜欢被撞死。

反射动作也在更常规的任务中发挥作用，例如让汽车保持在车道上行驶：当汽车稍稍偏离理想的车道位置时，一个简单的反馈控制系

统可以向相反的方向转动方向盘来修正偏移。转动幅度的大小取决于汽车的偏离距离。这些类型的控制系统通常的设计目的是：使随时间累积的跟踪误差的平方最小化。设计者推导出一个反馈控制规律，即在关于速度和道路曲率的某些假设下，近似实现了这种最小化。[71] 当你站立的时候，你身体中一个类似的系统一直在运行。如果它停止工作，你会在几秒钟内摔倒。这就和眨眼反射一样，你很难关闭这个机制，任由自己摔倒。

反射型智能体虽然实现了设计者的目标，但它们不知道目标是什么，也不知道为什么要以某种方式行动。这意味着它们不能真正为自己做决定，而其他人，通常是人类设计者或生物进化过程，必须事先决定一切。除了玩井字游戏、进行紧急制动等非常简单的任务之外，人们很难通过手动编程创建良好的反射型智能体。即使在非常简单的情况下，反射型智能体也非常不灵活，当环境表明反射型智能体执行的策略不再合适时，它们也无法改变自己的行为。

如果要创建更强大的反射型智能体，一种可能的方法是从示例中学习（附录D）。人类可以提供决策问题的示例，以及在每种情况下要做出的正确决策，而不用为如何行动制定规则，也不用提供奖励函数或目标。例如，我们可以通过提供法语句子的示例以及正确的英语翻译来创建“法译英翻译智能体”。（幸运的是，加拿大和欧盟议会每年都会产生数百万这样的例子。）然后，监督学习算法会处理这些例子，生成一个复杂的规则，以任意法语句子作为输入，生成英语翻译。当前机器翻译领域最优秀的学习算法是所谓的深度学习的一种形式，它以人工神经网络的形式产生一个规则，这个人工神经网络有数百个层次和数百万个参数（附录D）。其他深度学习算法已经被证明非常擅长对图像中的物体进行分类，以及识别语音信号中的单词。机器翻译、语音识别和基于机器视觉的物体识别是人工智能中三个最重要的子领域，这就是为什么人们对深度学习的前景如此兴奋。

人们几乎可以无休止地争论深度学习是否会直接让人类水平的人工智能出现，我自己的观点是（我稍后会做解释），它远远不能满足我们的需要（附录D），但现在还是让我们专注于如何用这样的方法去适应人工智能的标准模型。在人工智能的标准模型中，算法优化了固定的目标。对深度学习或者任何监督学习算法来说，“把目的输入机器”通常是为了最大化预测精度，或者最小化误差（二者是等效的）。虽然这似乎显而易见，但实际上有两种方法可以用来理解它，这取决于习得的规则在整个系统中所扮演的角色。第一个角色是纯粹的感知角色：人工神经网络处理感官输入，并以概率估计的形式将感知到的信息提供给系统的其余部分。如果网络采用物体识别算法，它可能会说“70%的概率是诺福克狷犬，30%的概率是诺威奇狷犬”。[72]系统的其余部分根据这些信息决定要采取的外部行动。这种纯粹的感知目标在以下意义上是没有问题的：与基于标准模型的“不安全”的人工智能系统不同，即使是“安全”的超级人工智能系统，其感知系统也需要尽可能准确，尽可能被精确地校准。

当我们从纯粹的感知角色转变为决策角色（第二个角色）时，问题就来了。例如，经过训练的用于识别物体的网络可能会自动为网站或社交媒体账号上的图片生成标签。贴这些标签是有后果的行为。每一个标记动作都需要一个实际的分类决策，除非保证每个决策都是完美的，否则人类设计者就必须提供一个损失函数，说明将A类物体错误地分类为B类物体所要付出的代价。这就是谷歌给人贴上大猩猩标签这一不幸事件的原因。2015年，一位名叫杰基·阿尔西内（Jacky Alciné）的软件工程师在推特上抱怨说，谷歌照片的图片标签服务把他和他的朋友贴上了大猩猩的标签。[73]虽然我们还不清楚这个错误究竟是如何发生的，但几乎可以肯定的是，谷歌的机器学习算法旨在最小化一个固定的、确定的损失函数，而且这个损失函数为所有错误分配了相等的代价。换句话说，它假设将一个人错误归类为大猩猩的代价与将诺福克狷犬误归类为诺威奇狷犬的代价等同起来。显然，这

并不是谷歌（或谷歌用户）真正的损失函数，正如随之而来的公共关系灾难所表现出来的那样。

可能出现的图片标签有成千上万种，因此将一类物体误判为另一类物体可能会产生数百万种潜在的不同代价。即使尝试过，谷歌也会发现很难预先指定所有这些数字。相反，正确的做法是承认错误分类代价的不确定性，并设计一种对代价和代价不确定性适当敏感的学习和分类算法。这样的算法可能偶尔会问谷歌设计者一些问题，比如，“把狗误分类为猫或把人误分类为动物，哪个更糟糕？”此外，如果错误分类的代价存在很大的不确定性，算法很可能会拒绝标记某些图片。

2018年初，有报道称谷歌照片拒绝对大猩猩的照片进行分类。人们给出一张非常清晰的大猩猩带着两个幼崽的图片，它却说：“嗯……还没看清楚。”[74]

我不想暗示说，人工智能采用标准模型在当时是一个糟糕的选择。一方面，在逻辑、概率和学习系统中，人们在开发标准模型的各种实例化方面已经做了大量出色的工作，由此产生的许多系统都非常有用。正如我们将在下一章中看到的，未来还会有更多的系统出现。另一方面，我们不能继续依赖惯常的做法，即通过反复试验来消除目标函数中的主要错误：智能日益提高、全球影响力日益增强的机器不会允许我们享受这种奢侈。

---

[1] 1英里≈1.6093千米。——编者注

[2] 圣地亚哥·拉蒙-卡哈尔（Santiago Ramón y Cajal）在1894年提出将突触变化作为学习发生的场所，但直到20世纪60年代末，这一假设才被实验证实。请参见：Timothy Bliss and Terje Lomo, “Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path,” *Journal of Physiology* 232 (1973): 331–56。

[3] 有关简介请参见：James Gorman, “Learning how little we know about the brain,” *The New York Times*, November 10, 2014。另请参见：Tom Siegfried,

“There’s a long way to go in understanding the brain,” *Science News*, July 25, 2017。《神经元》（*Neuron*）杂志2017年的一期特刊（vol. 94, pp. 933–1040）很好地概述了许多不同的理解大脑的方法。

[4] 意识的存在或缺失，即实际的主观体验，肯定会影响我们对机器的道德思考。如果我们获得足够的理解来设计有意识的机器，或者发现我们已经这样做了，那么我们将面临许多重要的道德问题，而这些问题在很大程度上是我们没有准备好的。

[5] 第一批在强化学习算法和神经生理学记录之间建立明确联系的论文之一：Wolfram Schultz, Peter Dayan, and P. Read Montague, “A neural substrate of prediction and reward,” *Science* 275 (1997): 1593–99。

[6] 进行颅内刺激的研究是为了寻找各种精神疾病的治疗方法。例如，请参见：Robert Heath, “Electrical self-stimulation of the brain in man,” *American Journal of Psychiatry* 120 (1963): 571–77。

[7] 一个可能由于上瘾而面临自我灭绝的物种的例子：Bryson Voirin, “Biology and conservation of the pygmy sloth, Bradypus pygmaeus,” *Journal of Mammalogy* 96 (2015): 703–7。

[8] 进化中的鲍德温效应通常起源于这篇论文：James Baldwin, “A new factor in evolution,” *American Naturalist* 30 (1896): 441–51。

[9] 鲍德温效应的核心思想也出现在：Conwy Lloyd Morgan, *Habit and Instinct*(Edward Arnold, 1896)。

[10] 演示鲍德温效应的现代分析和计算机实现：Geoffrey Hinton and Steven Nowlan, “How learning can guide evolution,” *Complex Systems* 1 (1987): 495–502。

[11] 进一步阐明鲍德温效应的计算机模型，其中包括内部奖励信号电路的进化：David Ackley and Michael Littman, “Interactions between learning and evolution,” in *Artificial Life II*, ed. Christopher Langton et al. (Addison-Wesley, 1991)。

[12] 在这里，我指的是我们今天的智能概念的根源，而不是描述古希腊的“理性”（nous）概念，后者有多种相关的含义。

[13] 引文摘自Aristotle, *Nicomachean Ethics*, Book III, 3, 1112b。

[14] 卡尔达诺是最早思考负数的欧洲数学家之一，他发展了游戏中的概率问题的早期数学处理方法。他卒于1576年，87年后，他的作品才得以出版：Gerolamo Cardano, *Liber de ludo aleae* (Lyons, 1663)。

[15] 阿尔诺的作品最初是匿名出版的，通常被称为*The Port-Royal Logic*: Antoine Arnauld, *La logique, ou l’art de penser* (Chez Charles Savreux, 1662)。另请参见：Blaise Pascal, *Pensées* (Chez Guillaume Desprez, 1670)。

[16] 效用的概念请见：Daniel Bernoulli, “Specimen theoriae novae de mensura sortis,” *Proceedings of the St. Petersburg Imperial Academy of Sciences* 5 (1738): 175–92。伯努利的效用思想源于一个商人塞普罗尼乌斯（Sempronius），他需要考虑是用一艘船运输贵重货物，还是将货物平分给两艘船，假设每艘船在航行中有50%的概率会沉没。这两种解决方案的期望货币价值是相同的，但是塞普罗尼乌斯显然更喜欢分两艘船运输的解决方案。

[17] 大多数人认为，冯·诺依曼并不是自己发明了这种体系结构，但他的名字出现在一份描述EDVAC（能存储程序的计算机）的颇具影响力的报告的初稿中。

[18] 冯·诺依曼和摩根斯坦的工作在许多方面都是现代经济理论的基础：Johnvon Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior*(Princeton University Press, 1944)。

[19] 效用是折现奖励的总和，这是保罗·萨缪尔森（Paul Samuelson）提出的一个数学上很方便的假设。“A note on measurement of utility,” *Review of Economic Studies* 4 (1937): 155–61。如果s0, s1……是一个状态序列，则其在该模型中的效用为U(s0, s1,……) = Σ trt R(st)，其中K为折现率，R为描述状态可取性的奖励函数。该模型的简单应用很少与真实个体对当前和未来奖励的可取性的判断相一致。有关详细分析，请参见：Shane Frederick, George Loewenstein, and Ted O’Donoghue, “Time discounting and time preference: A critical review,” *Journal of Economic Literature* 40 (2002): 351–401。

[20] 法国经济学家莫里斯·阿莱（Maurice Allais）提出了一个决策情景，在这个情景中，人类似乎总是违反冯·诺依曼–摩根斯坦公理：Maurice Allais, “Lecomportement de l’homme rationnel devant le risque: Critique des postulats et axiomes de l’école américaine,” *Econometrica* 21 (1953): 503–46。

[21] 有关非定量决策分析的介绍，请参见：Michael Wellman, “Fundamental concepts of qualitative probabilistic networks,” *Artificial Intelligence* 44 (1990):257–303。

[22] 我将在第9章进一步讨论人类非理性的证据。标准参考文献包括：Allais, “Lecomportement”; Daniel Ellsberg, *Risk, Ambiguity, and Decision* (PhD thesis,Harvard University, 1962); Amos Tversky and Daniel Kahneman, “Judgment under uncertainty: Heuristics and biases,” *Science* 185 (1974): 1124–31。

[23] 很明显，这是一个无法在实践中实现的思想实验。对不同未来的选择从来不会被详尽地呈现出来，而人类在做出选择之前也从来没有机会仔细审视和品味这些未来。相反，一个人只能得到关于未来的简短的概括，比如当“图书管理员”或“煤矿工人”。在做出这样的选择时，实际上是要求一个人比较两个完整的未来的概率分布，一种是从选择当“图书管

理员”开始，另一种是从选择当“煤矿工人”开始，每一种概率分布都假定这个人在未来都将采取最优行动。不用说，这并不容易。

[24] 第一次提到游戏的随机策略的文献是：Pierre Rémond de Montmort, *Essay d'analyse sur les jeux de hazard*, 2nd ed. (Chez Jacques Quillau, 1713)。这本书确定了一位“沃尔德格雷夫先生”（Monsieur de Waldegrave）作为卡牌游戏的最佳随机解决方案的来源。关于沃尔德格雷夫先生的身份细节，请参见：David Bellhouse, “The problem of Waldegrave,” *Electronic Journal for History of Probability and Statistics* 3 (2007)。

[25] 这个问题可以通过指定爱丽丝在4种情况下进球的概率来定义：（1）爱丽丝向鲍勃的右侧射门，鲍勃向右扑救；（2）爱丽丝向鲍勃的右侧射门，鲍勃向左扑救；（3）爱丽丝向鲍勃的左侧射门，鲍勃向右扑救；（4）爱丽丝向鲍勃的左侧射门，鲍勃向左扑救。在本例中，这4种情况的概率分别为25%、70%、65%和10%。现在假设爱丽丝的策略是以概率p向鲍勃的右侧射门，以概率$1-p$向鲍勃的左侧射门；而鲍勃以概率q向右扑救，以概率$1-q$向左扑救。则爱丽丝的回报是$U_A=0.25pq+0.70\ p(1-q)\ +\ 0.65\ (1-p)q\ +\ 0.10(1-p)\ (1-q)$，而鲍勃的回报是$U_B=-U_A$。在平衡时，$\partial U_A/\partial p\ =\ 0$且$\partial U_B/\partial q\ =\ 0$，得到$p\ =0.55,\ q\ =\ 0.60$。

[26] 最初的博弈论问题是由兰德公司的梅里尔·弗勒德（Merrill Flood）和梅尔文·德雷希尔（Melvin Dresher）提出的。塔克在拜访他们的办公室时发现了回报矩阵，并讲了一个“故事”来说明它。

[27] 博弈论者通常说，爱丽丝和鲍勃可以互相合作（保持沉默），也可以背叛并告发对方。我觉得这种语言令人感到困惑，因为“互相合作”不是每个智能体都可以单独做出的选择，也因为人们经常谈论与警察合作，以合作换取减刑等。

[28] 关于囚徒困境和其他博弈的一个有趣的基于信任的解决方案，请参见：Joshua Letchford, Vincent Conitzer, and Kamal Jain, “An ‘ethical’ game-theoretic solution concept for two-player perfect-information games,” in *Proceedings of the 4th International Workshop on Web and Internet Economics*, ed. Christos Papadimitriou and Shuzhong Zhang (Springer, 2008)。

[29] “公地悲剧”的起源：William Forster Lloyd, *Two Lectures on the Checks to Population* (Oxford University, 1833)。

[30] 这一主题在全球生态背景下的现代复兴：Garrett Hardin, “The tragedy of the commons,” *Science* 162 (1968): 1243–48。

[31] 很有可能，即使我们试图用化学反应或生物细胞来制造智能机器，这些装配物最终也会变成图灵机在非传统材料中的实现。一个物体是否是通用计算机与它的构成无关。

[32] 图灵的突破性论文定义了现在所知的图灵机，它是现代计算机科学的基础。标题中的“Entscheidungsproblem”，或称决策问题，是一阶逻辑中的蕴涵判定问题：Alan Turing, “On computable numbers, with an application to the

Entscheidungsproblem," *Proceedings of the London Mathematical Society*, 2ndser., 42 (1936): 230–65。

[33] 它的发明者之一萨耶夫·萨拉赫丁（Sayeef Salahuddin）对负电容的研究进行了很好的综述：Sayeef Salahuddin, "Review of negative capacitance transistors," in *International Symposium on VLSI Technology, Systems and Application* (IEEEPress, 2016)。

[34] TPU Pod是由多个TPU组成的多层次的机器学习超级计算机。——译者注

[35] 关于量子计算的更好的解释，请参见：Scott Aaronson, *Quantum Computing since Democritus* (Cambridge University Press, 2013)。

[36] 在经典计算和量子计算之间建立了清晰的复杂性理论区分的论文：Ethan Bernstein and Umesh Vazirani, "Quantum complexity theory," *SIAM Journal on Computing* 26 (1997): 1411–73。

[37] 一位著名物理学家撰写的文章很好地介绍了当前的理解和技术现状：John Preskill, "Quantum computing in the NISQ era and beyond," arXiv:1801.00862(2018)。

[38] 论1千克物体的最大计算能力：Seth Lloyd, "Ultimate physical limits to computation," *Nature* 406 (2000): 1047–54。

[39] 关于人类可能是物理上可实现的智能的巅峰的说法，请参见：Kevin Kelly,"The myth of a superhuman AI," *Wired*, April 25, 2017: "We tend to believe that the limit is way beyond us, way 'above' us, as we are 'above' an ant. . . . What evidence do we have that the limit is not us?"

[40] 如果你想了解一个解决停机问题的简单技巧，仅仅靠"运行程序查看它是否完成"这种浅显的方法是行不通的，因为这种方法未必能完成。你可能等了100万年，但仍然不知道程序是真的陷入了无限循环，还是只是在慢慢来。

[41] 停机问题不可判定的证据是一个巧妙的骗术。问题是：是否存在一个循环检查程序LoopChecker(P，X)，对于任意程序P和任意输入X，它都能在有限的时间内正确地判断应用于输入X的程序P是会停机并产生结果，还是会一直运行下去。假设存在这种LoopChecker。现在编写一个程序Q，它将LoopChecker作为子例程调用，将Q本身和X作为输入，然后执行与LoopChecker（Q，X）预测相反的操作。因此，如果LoopChecker说Q停机，Q就不会停机，反之亦然。于是，LoopChecker存在的假设导致了矛盾，因此Loop Checker不可能存在。

[42] 因为到目前为止，关于"NP完全问题"需要超多项式时间（通常被称为P≠NP）的说法仍然是一个未经证实的猜想。然而，经过近50年的研究，几乎所有数学家和计算机科学家都确信这一说法是正确的。

[43] 洛夫莱斯关于计算的著作主要出现在她翻译的一名意大利工程师对巴贝奇的机器的评论的注释中：L. F. Menabrea, "Sketch of the Analytical Engine invented by

Charles Babbage," trans. Ada, Countess of Lovelace, in *Scientific Memoirs,*vol. III, ed. R. Taylor (R. and J. E. Taylor, 1843)。门纳布利亚（Menabrea）的原创文章是用法语写的，基于巴贝奇1840年的演讲，发表于*Bibliothèque Universelle de Genève* 82 (1842)。

[44] 关于智能机器可能性的早期开创性论文之一：Alan Turing, "Computing machinery and intelligence," *Mind* 59 (1950): 433–60。

[45] 加州斯坦福研究所Shakey项目的一位负责人在回顾中做了总结：Nils Nilsson, "Shakey the robot," technical note 323 (SRI International, 1984)。一部24分钟的电影《SHAKEY：机器人学习与规划的实验》（*SHAKEY: Experimentation in Robot Learning and Planning*）拍摄于1969年，当时引起了全美国的关注。

[46] 标志着现代基于概率的人工智能开始的一本书：Judea Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, 1988)。

[47] 从技术上讲，国际象棋并不是完全可观察的。一个程序确实需要记住少量的信息来确定"王车易位"和"吃过路兵"着法的合法性，并通过"重复局面"和"50步规则"来定义和局。

[48] 完整的论述请参见：Chapter 2 of Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach,* 3rd ed. (Pearson, 2010)。

[49] 关于《星际争霸》的状态空间大小的讨论请参见：Santiago Ontañon et al., "A survey of real-time strategy game AI research and competition in Star Craft," *IEEE Transactions on Computational Intelligence and AI in Games* 5 (2013): 293–311。大量的移动选项是可能的，因为玩家可以同时移动所有单位。由于限制了一次可以移动的单位或单位组的数量，这个数字会有所下降。

[50] 关于《星际争霸》中的人机竞争，请参见：Tom Simonite, "Deep Mind beats pros at Star Craft in another triumph for bots," *Wired*, January 25, 2019。

[51] 关于Alpha Zero的描述，请参见：David Silver et al., "Mastering chess and shogi by self-play with a general reinforcement learning algorithm," arXiv: 1712.01815 (2017)。

[52] 最优路径是通过使用A*算法及其众多派生算法找到的：Peter Hart, Nilsson, and Bertram Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Transactions on Systems Science and Cybernetics* SSC-4 (1968): 100–107。

[53] 麦卡锡介绍"Advice Taker"程序和基于逻辑的知识系统的论文：John McCarthy, "Programs with common sense," in *Proceedings of the Symposium on*

Mechanisation of Thought Processes (Her Majesty's Stationery Office, 1958)。

[54] 要了解基于知识的系统的重要性，你可以考虑一下数据库系统。数据库包含具体的个人事实，例如我的钥匙的位置和你的脸书好友的身份。数据库系统不能存储通用规则，如国际象棋规则或英国公民身份的法律定义。它们可以统计有多少叫爱丽丝的人有叫鲍勃的朋友，但它们不能确定某位特定的爱丽丝是否符合英国公民身份的条件，也不能确定棋盘上的特定一步棋是否会导致“将杀”。数据库系统不能将两种知识结合起来产生第三种知识：它们支持存储，但不支持推理。（的确，许多现代数据库系统提供了添加规则的方法和使用这些规则得出新事实的方法。从某种程度上讲，它们确实是基于知识的系统。）尽管数据库系统是基于知识的系统的高度压缩版本，但它们是当今大多数商业活动的基础，每年产生数千亿美元的价值。

[55] 描述一阶逻辑完全性定理的原始论文：Kurt Gödel, “Die Vollständigkeit der Axiome des logischen Funktionenkalküls,” *Monatshefte für Mathematik* 37 (1930):349–60。

[56] 一阶逻辑的推理算法确实存在一个缺口：如果没有答案，即如果现有的知识不足以给出任何一种答案，那么算法可能永远无法完成任务。这是不可避免的，从数学上讲，正确的算法永远都不可能以“不知道”终结，这本质上与没有算法可以解决停机问题的原因相同。

[57] 一阶逻辑中定理证明的第一个算法是通过将一阶语句简化为（大量）命题语句来实现的：Martin Davis and Hilary Putnam, “A computing procedure for quantification theory,” *Journal of the ACM* 7 (1960): 201–15。罗宾逊的解析算法直接对一阶逻辑语句进行运算，使用“统一”来匹配包含逻辑变量的复杂表达式：J. Alan Robinson, “A machine-oriented logic based on the resolution principle,” *Journal of the ACM* 12 (1965): 23–41。

[58] 人们可能会好奇，逻辑机器人Shakey如何得出该做什么事情的明确结论。答案很简单：Shakey的知识库包含错误的断言。例如，Shakey认为，通过执行“把A物体推过D门，进入B房间”，A物体最终会到达B房间。这种想法是错误的，因为Shakey可能会卡在门口或完全错过门口，或者有人可能会偷偷将A物体从Shakey的手中拿走。Shakey的计划执行模块可以检测到计划失败并相应地重新进行计划，因此严格地说，Shakey不是一个纯粹的逻辑系统。

[59] 关于概率在人类思维中的作用的早期评论：Pierre-Simon Laplace, Essai philosophique sur les probabilités (Mme. Ve. Courcier, 1814)。

[60] 贝叶斯逻辑的一种相当非技术性的描述方式：Stuart Russell, “Unifying logic and probability,” *Communications of the ACM* 58 (2015): 88–97。这篇论文大量引用了我以前的学生布赖恩·米尔奇（Brian Milch）的博士论文研究。

[61] 贝叶斯定理的来源：Thomas Bayes and Richard Price, “An essay towards solving a problem in the doctrine of chances,” *Philosophical Transactions of the Royal Society of London* 53 (1763): 370–418。

[62] 从技术上讲，塞缪尔的程序并没有将输赢视为绝对的奖励，而是将棋子的价值固定为正值，然而，该程序通常倾向于争取胜利。

[63] 强化学习在制作世界一流西洋双陆棋程序中的应用：Gerald Tesauro, “Temporal difference learning and TD-Gammon,” *Communications of the ACM* 38(1995): 58–68。

[64] 使用深度强化学习来练习玩多种电子游戏的DQN系统：Volodymyr Mnih et al., “Human-level control through deep reinforcement learning,” *Nature* 518(2015): 529–33。

[65] 比尔·盖茨对《Dota 2》人工智能的评论：Catherine Clifford, “Bill Gates says gamer bots from Elon Musk-backed nonprofit are ‘huge milestone’ in A.I.,”CNBC, June 28, 2018。

[66] OpenAI 5号在《Dota 2》游戏中战胜人类世界冠军的记述：Kelsey Piper, “AI triumphs against the world’s top pro team in strategy game Dota 2,” *Vox*, April 13,2019。

[67] 文献中奖励函数的设定偏误导致意外行为的案例概要：Victoria Krakovna, “Specification gaming examples in AI,” *Deep Safety* (blog), April 2, 2018。

[68] 根据最大速度定义的进化适应度函数导致非常出乎意料的结果的一个案例：Karl Sims, “Evolving virtual creatures,” in *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques* (ACM, 1994)。

[69] 有关反射型智能体可能性的有趣描述，请参见：Valentino Braitenberg, *Vehicles:Experiments in Synthetic Psychology* (MIT Press, 1984)。

[70] 一篇关于一辆处于无人驾驶模式的汽车撞上行人的致命事故的新闻报道：Devin Coldewey, “Uber in fatal crash detected pedestrian but had emergency braking disabled,” *Tech Crunch,* May 24, 2018。

[71] 关于转向控制算法，请参见：Jarrod Snider, “Automatic steering methods for autonomous automobile path tracking,” technical report CMU-RI-TR-09-08,Robotics Institute, Carnegie Mellon University, 2009。

[72] 诺福克猎犬和诺威奇猎犬是ImageNet数据库中的两个类别。众所周知，二者很难区分，它们在1964年之前一直被视为一个品种。

[73] 关于图片标签的一个非常不幸的事件：Daniel Howley, “Google Photos mislabels 2 black Americans as gorillas,” *Yahoo Tech*, June 29, 2015。

[74] 关于谷歌和大猩猩的后续文章：Tom Simonite, “When it comes to gorillas, Google Photos remains blind,” *Wired*, January 11, 2018。


03
人工智能
在未来会
如何发展?



近未来

1997年5月3日，IBM制造的国际象棋计算机“深蓝”和国际象棋世界冠军加里·卡斯帕罗夫（Garry Kasparov，或许是历史上最优秀的人类棋手）之间的国际象棋比赛开始了，美国《新闻周刊》称这场比赛是“大脑的最后一战”。前5局，双方2.5：2.5平分秋色。5月11日，“深蓝”在决胜局中战胜了卡斯帕罗夫。IBM的市值一夜之间增加了180亿美元。所有人都说，人工智能取得了巨大的突破。

从人工智能研究的角度来看，这场比赛根本没有突破。“深蓝”的胜利虽然令人印象深刻，但它只是延续了几十年来显而易见的趋势。国际象棋算法的基本设计是由克劳德·香农在1950年提出的，[1]这一基本设计在20世纪60年代初实现了重大改进。此后，最优秀的国际象棋程序的等级评分稳步提高，这主要是因为计算机速度更快，让程序能够算到未来更远处。1994年，我和彼得·诺维格列出了1965年以来最好的国际象棋程序和国际象棋棋手的等级评分。[2]在这个评分系统中，卡斯帕罗夫的评分是2 805分。1965年，国际象棋程序的等级评分从1 400分开始，在随后的30年里以一条近乎完美的直线上升。从1994年开始推算，系统预测计算机将会在1997年击败卡斯帕罗夫，而这与现实完全相符。

对人工智能研究人员而言，真正的突破发生在“深蓝”进入公众视野之前的三四十年。类似的，在深度卷积网络开始成为头条新闻的20多年前，它就已经存在了，所有的数学问题都完全解决了。

公众从媒体上看到的人工智能取得突破的观点——人类取得的辉煌胜利，机器人成为沙特阿拉伯公民等，与全世界各地的研究实验室里真正发生的事情几乎没有关系。在实验室里，研究包括大量的思考、交流和在白板上写数学公式。想法不断产生、被抛弃和重新被发

现。好的想法，即真正的突破，往往在当时没有被注意到，可能直到后来才会被认为给人工智能的重大进步提供了基础，或许是当有人在更合适的时间重新改造它的时候。想法是经过试验的，它最初往往在简单的问题上证明了基本直觉的正确性，然后在更困难的问题上，我们才能观察到想法的延伸效果。通常，一个想法本身并不能提供能力上的实质性改进，它必须等待另一个想法的出现，二者结合才能证明价值。

所有这些活动都是外界完全看不见的。在实验室之外的世界里，只有当想法逐渐积累，当其有效性的证据超越一个阈值时，人工智能才会被看到：在这个临界点上，人工智能值得投入资金开展技术工作，创造新的商业产品或令人印象深刻的演示了。然后，媒体宣布我们取得了突破。

因此，人们可以预期，在未来几年里，世界各地的研究实验室酝酿的许多其他想法将跨越商业适用性的门槛。随着商业投资率的提高，以及世界对人工智能应用的接受程度越来越高，这种情况将会越来越频繁地发生。本章提供了一些我们很快就会看到的例子。在此过程中，我将提到这些技术进步的一些缺点。你可能会想到更多的缺点，但别担心。我将在下一章讨论这些内容。

人工智能生态系统

最初，大多数计算机的运行环境基本上是空洞无形的：它们唯一的输入来自穿孔卡片，唯一的输出方法是在行式打印机上打印字符。也许由于这个原因，大多数研究人员将智能机器视为问答系统。直到20世纪80年代，机器作为智能体在环境中感知和行动的观点才被广泛接受。

20世纪90年代，万维网的出现为智能机器开辟了一个全新的天地。人们创造了一个新词“Softbot”，用来描述完全在Web等软件环

境中运行的“软件机器人”。Softbot（后来被称为bot，即机器人）可以感知网页，并通过发出字符序列、网址等来执行动作。

人工智能公司在互联网繁荣时期（1997—2000年）迅速发展，为搜索和电子商务提供了核心能力，包括链接分析、推荐系统、信誉系统、比较购物和产品分类。

21世纪初，配有麦克风、摄像机、加速计和GPS的手机普及开来，为人工智能系统进入人们的日常生活提供了新的途径。亚马逊Echo、谷歌Home和苹果HomePod等“智能音箱”已经完成了这一过程。

到2008年左右，接入互联网的物体数量已经超过了接入互联网的人数，这一转变被一些人认为是物联网（IoT）的开端。这些物体包括汽车、家用电器、交通灯、自动售货机、恒温器、四轴飞行器、照相机、环境传感器、机器人，以及在制造、分销和零售系统中的各种物料。这为人工智能系统提供了更强大的感知能力和对现实世界的控制途径。

最后，随着感知能力的提高，人工智能机器人得以走出工厂，进入真实的、非结构化的、混乱的世界，在工厂里它们依赖于严格约束的对象排列，而走出工厂之后，它们的摄像机可以看到一些有趣的东西。

无人驾驶汽车

20世纪50年代末，约翰·麦卡锡设想有一天会有一辆无人驾驶汽车把他送到机场。1987年，恩斯特·迪克曼斯（Ernst Dickmanns）在德国的高速公路上展示了一辆无人驾驶的奔驰面包车，它可以保持在车道上行驶，尾随在另一辆车后面，变换车道、超车。[3] 30多年过去了，我们仍然没有一辆全自动汽车，但距离它成为现实已经越来越近了。开发的重点早已从学术研究实验室转移到大公司。截至2019年，性能最好的测试车辆已经在公共道路上行驶了数百万英里（在驾

驶模拟器中行驶了数十亿英里），没有发生严重事故。[4]不幸的是，其他无人驾驶汽车和半自动驾驶汽车已经造成数人死亡。[5]

我们为什么要花如此长的时间才能实现安全的无人驾驶？第一个原因是，性能要求非常严格。在美国，人类司机大约每行驶1亿英里会遭遇一次致命事故，这设置了一个很高的标准。无人驾驶汽车要想被人们接受，就需要表现得比这更好，也许要每行驶10亿英里出现一次致命事故才行。或者说，无人驾驶汽车如果每周行驶40小时，则25000年才会出现一次致命事故。第二个原因是，一个预期的变通办法——当交通情况混乱或超出其安全操作条件时，将控制权交给人类，根本行不通。当汽车无人驾驶时，人类很快就会脱离当前的驾驶环境，无法足够迅速地重新了解周围环境，安全地接管汽车。此外，如果出了什么问题，坐在后座的非驾驶员或出租车乘客将无法驾驶汽车。

目前的项目目标是达到美国汽车工程师学会（SAE）的L4级自动驾驶。[6]这意味着车辆必须在任何时候都能够自动驾驶、安全停车，但是它受地理条件和天气条件的限制。因为地理和天气条件可能会发生变化，并且可能出现L4级车辆无法处理的异常情况，所以必须有人在车内，并在需要时随时准备接管汽车。（L5级是不受限制的自动驾驶，它不需要人工驾驶，但更难实现。）L4级自动驾驶的自主性远远超出了简单的反射型任务，如在白线内侧行驶和避开障碍物。车辆必须基于当前和过去的观察结果，评估所有相关物体的意图和未来可能的轨迹，包括评估现在可能看不到的物体。然后，利用前向搜索，车辆必须找到一条优化了安全性的前进路径。一些项目正在尝试采用更直接的方法，这些方法基于强化学习（当然主要是模拟）和监督学习（从数百名人类驾驶员的记录中学习），但是这些方法似乎不太可能达到所需的安全水平。

无人驾驶汽车的潜在好处是巨大的。每年，车祸造成全世界120万人死亡，数千万人受重伤。无人驾驶汽车的合理目标是将这些数字降

低到现在的10%。一些分析还预测，交通成本、停车场设施成本将因此降低，拥堵和污染情况也将有所好转。城市中的私家车和大型公共汽车将转变为无处不在的共享无人驾驶电动汽车，提供门到门的交通服务，并为枢纽之间的高速公共交通连接提供支持。[7]由于每位乘客每英里的运输成本低至3美分，大多数城市可能会选择免费提供这项服务，而让乘客承受没完没了的广告轰炸。

当然，要获得所有这些好处，业界必须防范风险。如果有太多的死亡归因于设计不当的实验车辆，监管机构可能会暂停计划中的部署，或者实施可能几十年内都无法达到的极其严格的标准。[8]当然，人们可能会决定不购买也不乘坐无人驾驶汽车，除非它们确实是安全的。2018年的一项民意调查显示，与2016年相比，消费者对无人驾驶汽车技术的信任度大幅下降。[9]即使这项技术取得了成功，向无人驾驶的过渡也将是一个令人感到尴尬的过程：人类驾驶技能可能会退步或消失，而由人类驾驶汽车这种鲁莽和反社会的行为可能会被完全禁止。

智能个人助理

到目前为止，大多数读者都已经体验过不太智能的个人助理：把从电视购物节目中听到的话当成购买指令去执行的智能音箱，或者听到“给我打电话叫救护车！”（Call me an ambulance!）之后回答“好的，从现在开始我叫你‘Ann Ambulance’”的手机聊天机器人。这些系统本质上是应用程序和搜索引擎的语音接口，它们主要基于固定的刺激—反应模板。这种方法可以追溯到20世纪60年代中期的Eliza（人工智能软件）。[10]

这些早期的系统有三大缺陷：访问缺陷、内容缺陷和语境缺陷。访问缺陷，即它们对正在发生的事情缺乏感官意识，例如，它们可能能够听到用户在说什么，但看不到用户在和谁说话。内容缺陷，即它

们根本无法理解用户所说的话或发送的短信的含义，即便它们可以访问这些内容。语境缺陷，即它们缺乏跟踪和推理日常生活的目标、活动和关系的能力。

尽管存在这些缺点，智能音箱和手机助手仍为用户提供了足够的价值，它们进入了数亿人的家中和口袋里。从某种意义上讲，它们是人工智能的“特洛伊木马”。因为它们就在你的身边，嵌入几乎每个人的生活中，它们的能力的每一个微小进步都价值数十亿美元。

因此，进步正在密集而又迅速地到来。或许最重要的是理解内容的基本能力：当它听到“约翰在医院”时，它要不仅仅会回答说“但愿没有什么大碍”，而且还应了解更多实际信息，比如用户8岁的儿子在附近的医院，并且可能受了严重的伤或生了病。访问电子邮件和文本通信以及（通过屋内的智能音箱）访问电话和家庭对话的能力，将为人工智能系统提供足够的信息，它以此来构建关于用户生活的合理且完整的画面。这也许比为19世纪贵族家庭工作的管家或为如今的CEO（首席执行官）工作的行政助理获得的信息还要多。

当然，仅仅了解原始信息是不够的。智能助理要想真正发挥作用，还需要了解世界运转的常识：医院里的孩子不可能同时待在家中，因手臂骨折而住院的时长很少会超过一天或两天，孩子的学校需要提前知道缺勤情况等。这样的知识使智能助手能够跟踪它没有直接观察到的东西，这是智能系统的一项基本技能。

我认为，上一段中描述的能力在现有的概率推理技术下是可行的（附录C），但是这需要花费非常大的努力来构建构成我们日常生活的各种事件和事务的模型。到目前为止，由于所涉及的成本和不确定的回报，常识类的建模项目通常还没有开始（可能在情报分析和军事计划的分类系统中除外）。然而现在，这样的项目很容易触及数亿名用户，因此投资风险更低，潜在回报更高。此外，通过接触大量用户，智能助理可以快速学习并填补知识上的所有空白。

因此，人们有望看到的智能助理是，用户每月只需付几美分就可以让它帮助自己管理越来越多的日常活动：日程、旅行、家庭购物、支付账单、辅导孩子作业、发电子邮件和打电话、提醒、膳食计划，以及人们“做梦都想找到的”钥匙。这些技能不会分散在多个应用程序中。相反，它们将是一个单一、综合的智能体的各个方面，这些方面可以从协同作用中受益，就像军事人员所说的“通用作战态势图”那样。

智能助手的通用设计模板包括：关于人类活动的背景知识、从感知和文本数据流中提取信息的能力，以及使智能助手适应用户的特定环境的学习过程。相同的通用模板至少可以应用于其他三个主要领域：健康、教育和财务。对于这些应用程序，系统需要跟踪用户的身体、思想和银行账户的状态（广义地说）。与日常生活中的助理一样，在这三个领域中创建必要的通用知识所付出的前期成本要分摊到数十亿名用户身上。

例如，在健康领域，我们所有人的生理状况大致相同，关于人体如何工作的详细知识已经被编码成机器可读的形式。[11]系统将适应你的个人特点和生活方式，提供疾病的预防建议和疾病的早期预警。

在教育领域，智能教学系统的前景早在20世纪60年代就得到了承认，[12]但真正的进步还需要很长时间。主要原因是内容和访问上的缺陷：大多数教学系统不理解它们声称要教授的内容，也不能通过语音或文本与学生进行双向交流。（我在想象自己用我完全不会说的老挝语教授我完全搞不懂的弦理论的情景。）语音识别的最新进展意味着，自动化的教师最终可以与尚未完全识字的学生进行交流。此外，概率推理技术现在可以跟踪学生的知识掌握情况，并优化教学内容，最大限度地提高学习效果。[13]始于2014年的“全球学习XPRIZE竞赛”提供了1 500万美元奖金，用于开发“开源的、可扩展的软件，使发展中国家的儿童能够在15个月内自学基本的阅读、写作和算术”。

来自获奖者“Kitkit School”（套件学校）和“onebillion”（十亿）的结果表明，这一目标已基本实现。

在个人金融领域，系统将跟踪个人的投资、收入流、强制性的支出和可自由支配的支出、债务、利息、紧急准备金等，就像金融分析师了解公司的财务情况和前景一样。与处理日常生活的智能体进行整合，将使系统提供更细致的服务，甚至可以确保孩子得到他们的零花钱，孩子还会因为他们搞恶作剧而被扣掉一部分零花钱。人们在不久的将来将会获得以往专为超级富豪准备的高质量的日常财务建议。

如果在阅读上面几段时，你的“隐私警报”没有响起，那说明你没有关注新闻。然而，隐私问题有多个层面。首先，如果你的个人助理对你一无所知，那么它真的有用吗？大概没用吧。其次，如果个人助理不能从多个用户那里采集信息，从而更多地了解普通人和与你类似的人，那么它真的有用吗？大概没用吧。因此，这两件事是否意味着我们必须放弃我们的隐私，才能在日常生活中受益于人工智能呢？答案是否定的。原因是学习算法可以使用安全多方计算技术[14]对加密数据进行操作，这样用户就可以在不损害自身隐私的情况下从资源池中获益。[15]软件提供商会在没有立法规定的情况下自愿采用隐私保护技术吗？这还有待观察。然而，似乎不可避免的是，只有当个人助理的主要职责是为用户服务，而不是为开发它的公司服务时，用户才会信任它。

智能家居和家用机器人

智能家居的概念已经发展了几十年。1966年，西屋电气的工程师詹姆斯·萨瑟兰（James Sutherland）开始搜集多余的计算机部件，用来打造第一台智能家居控制器ECHO。[16]不幸的是，ECHO重800磅[17]，电功率3.5千瓦，却只能管理三个数字时钟和电视天线。随后

的系统又要求用户掌握极其复杂的控制界面。不出所料，它们从未流行起来。

从20世纪90年代开始，几个雄心勃勃的项目试图设计出能够在人为干预尽可能少的情况下进行自我管理的住宅，并使用机器学习来适应居住者的生活方式。为了让这些实验有意义，必须有真人住在房子里。不幸的是，错误决策的频繁出现使系统变得糟糕透顶——居住者的生活质量非但没有提高，反而下降了。例如，在华盛顿州立大学2003年的MavHome项目中，[18]如果访客的就寝时间比居住者通常的就寝时间晚，那么他们就不得不坐在黑暗中，而这种情况经常发生。[19]就像不太智能的个人助理一样，这样的失败缘于系统对居住者活动的感官接触不足，以及无法理解和掌握房子里发生的事情。

配备了摄像头、麦克风，拥有必要的感知和推理能力的真正的智能住宅，可以理解居住者在做什么：做客、吃饭、睡觉、看电视、阅读、锻炼、为长途旅行做准备，还是跌倒后无助地躺在地板上。通过与智能个人助理配合，住宅可以非常清楚地知道谁会在什么时间进出房间，谁会在哪里用餐等。这种理解使它能够管理暖气、照明、百叶窗和安全系统，及时发送提醒，并在出现问题时通知用户或提供应急服务。美国和日本的一些新建公寓楼已经采用了这种技术。[20]

智能家居的价值因其执行器而受到限制。虽然更简单的系统（定时恒温器、动作感应灯和防盗报警器）对环境不那么敏感，但它们可以用更可预测的方式提供许多相同的功能。智能家居不能叠衣服、洗碗或拿报纸。它的确需要一个物理机器人来执行它的指令。

我们等待的时间可能不会太久。机器人已经展示了许多必需的技能。在我的同事彼得·阿贝尔（Pieter Abbeel）所在的加州大学伯克利分校实验室中，BRETT（Berkeley Robot for the Elimination of Tedious Tasks，意为“消除烦琐任务的伯克利机器人”）自2011年以来一直在叠成堆的毛巾，而波士顿动力公司的SpotMini机器人可以爬楼梯和开门（图5）。有几家公司已经在制造烹饪机器人了，尽管它们

需要特殊的封闭式设置和预先切好的食材，而且在普通的厨房里无法工作。[21]

实用家用机器人需要三种基本能力：感知能力、机动性和灵活性。其中，灵活性最成问题。正如布朗大学的机器人学教授斯蒂芬妮·泰利克斯（Stefanie Tellex）所说：“大多数机器人在大部分时间里都无法拿起大多数物体。”其原因有三：一是触觉感知问题；二是制造问题（目前，制造一只灵巧的机械手成本非常高）；三是算法问题，我们还没有很好地理解如何将感知和控制结合起来，以掌握和操纵家庭中各种各样的常见物体。仅仅针对刚性物体就有几十种抓握类型，并且有成千上万种不同的操作技巧，比如从瓶子里倒出两粒药丸，撕掉果酱罐上的标签，在软面包上涂抹硬黄油，或者用叉子从锅里挑起一根意大利面条，看看它是否熟了。

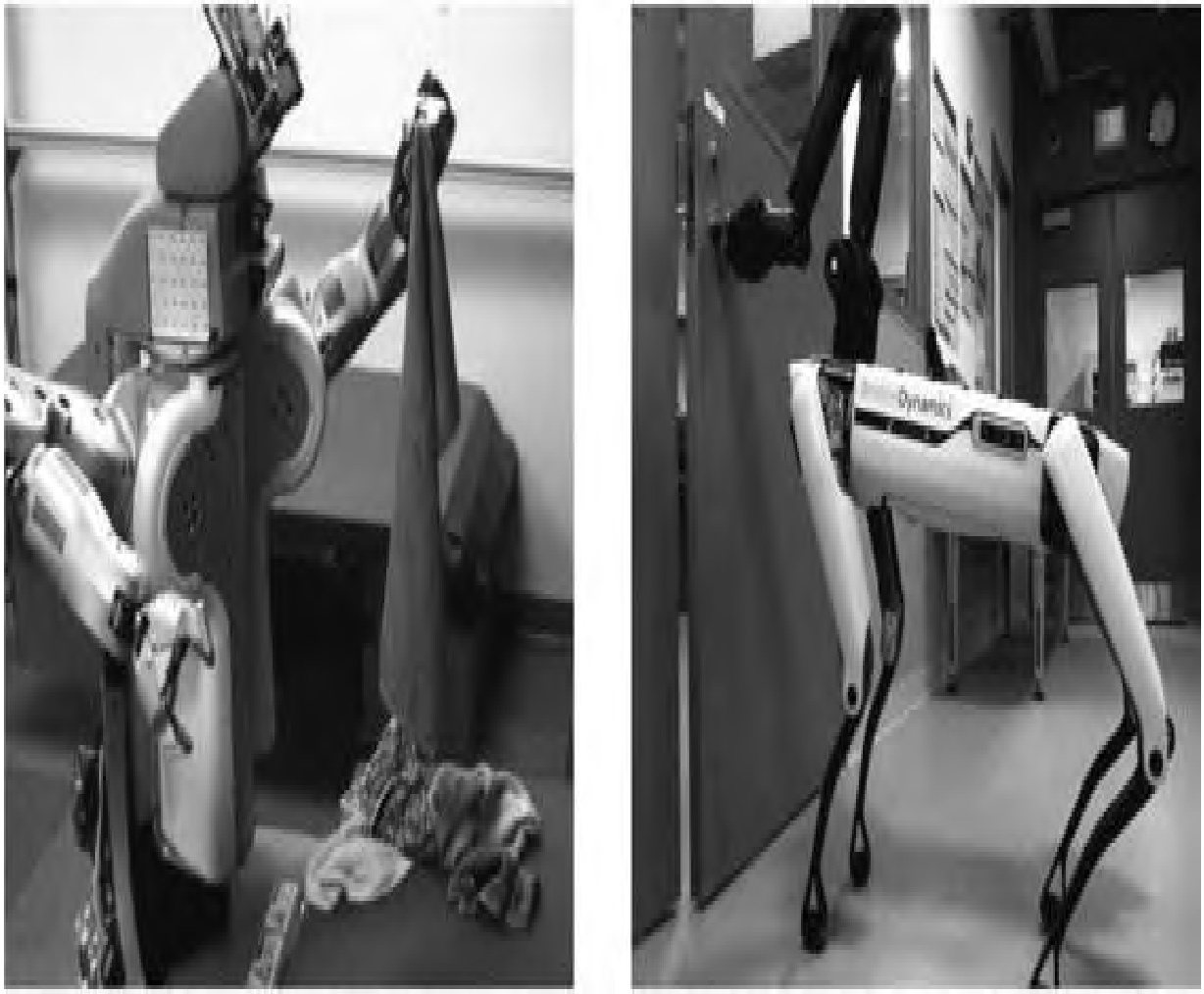


图5 （左）BRETT机器人叠毛巾；（右）波士顿动力公司的SpotMini机器人正在开门

触觉感知和制造问题有可能通过3D（三维）打印来解决，波士顿动力公司已经将3D打印用于Atlas人形机器人的一些更复杂的部分。机器人的操作技能正在迅速进步，这在一定程度上得益于深度强化学习。[22]最后的推动力——把所有这些整合到一起，形成近似于电影中的机器人那种令人敬畏的身体技能，很可能来自相当乏味的仓储行业。这个行业里只有亚马逊一家公司雇用了几十万人，他们从大型仓库的箱子中挑出产品，然后分发给顾客。从2015年到2017年，亚马逊每年都会举办一次“挑拣挑战赛”，以加速能够完成这项任务的机器人的发展。[23]虽然还有很长的路要走，但当核心研究问题得到解决

时——很可能在10年内，我们就可以期待高性能机器人的迅速问世。它们最初将用于仓库，然后用于其他商业应用，如农业和建筑业，在这些应用中，各种任务和对象都是相对可预测的。我们也可能很快会看到它们在零售部门做一些工作，比如给超市货架装货和把衣服重新叠好。

第一个真正受益于家庭机器人的群体是年老体弱的人，对他们而言，实用的机器人可以提供一定程度的独立性，这种独立性在其他情况下无从得到。即使机器人能够完成的任务很有限，而且它们对正在发生的事情只有初步的理解，它们也非常有用。而要制造那种可以从容地管理家务，并预见主人每一个愿望的机器人管家，我们仍然有一段路要走，因为这需要某种接近人类水平的人工智能的通用性。

全球层面的智能

发展理解语音和文本的基本能力，将会使智能个人助理可以做一些人类助理做的事情。（但是它们做这些事的成本是每月几便士，而不是每月几千美元。）基本的语音和文本理解能力，也会让机器能够做一些人类做不到的事情——不是因为理解的深度，而是因为事情的规模。例如，一台具备基本阅读能力的机器将能够在午餐时间阅读人类曾经写过的所有内容，然后它会四处寻找其他事情做。[24]凭借语音识别能力，它可以在下午茶之前收听每一个广播和电视节目的音频。相比之下，仅仅为了阅读完当前世界上所有的出版物（更不用说过去所有的书面材料）就需要20万名全职人员，收听当前的广播则还需要6万名全职人员。[25]

这样一个系统，如果它能提取简单的事实性陈述，并能跨所有语言整合所有这些信息，那么它就会成为令人难以置信的回答问题和揭示模式的资源，可能比目前价值约1万亿美元的搜索引擎还强大得多。它对历史学和社会学等领域的研究价值将是不可估量的。

当然，监听世界上所有的电话也是可能的（这项工作需要大约2000万人）。有些秘密机构会发现这很有价值，其中一些机构多年来一直在进行简单的大规模机器监听，比如找出对话中的关键词，现在已经过渡到将整个对话转录成可搜索的文本。[26]转录当然有用，但不如同时理解和整合所有对话的内容有用。

机器可以利用的另一种“超级力量”是瞬间看到整个世界。粗略地讲，卫星每天以每像素约50厘米的平均分辨率拍摄整个世界。在这个分辨率下，地球上的每一栋房子、每一艘船、每一辆汽车、每一头奶牛和每一棵树都清晰可见。检查所有这些图像需要超过3 000万名全职员工。[27]因此，目前没人能遍览绝大多数卫星数据。计算机视觉算法可以处理所有这些数据，生成一个每天更新的可搜索的全球数据库，还能生成经济活动、植被变化、动物迁徙和人口流动、气候变化影响等的可视化和可预测模型。像Planet（星球）和DigitalGlobe（数字地球）这样的卫星公司正忙着把这一想法变成现实。

随着全球层面的感知成为可能，全球层面的决策也在成为可能。例如，通过全球卫星数据，人们可以建立详细的模型来管理全球环境，预测环境和经济干预的影响，并为联合国的可持续发展目标提供必要的分析输入。[28]我们已经看到了“智慧城市”控制系统，它旨在优化交通管理、运输、垃圾收集、道路维修、环境保护和其他造福市民的功能，这些可能会拓展到国家层面。直到最近，这种程度的协调仍然只能通过庞大、低效的人类官僚机构来实现，这些将不可避免地被照顾我们集体生活越来越多方面的超级智能体所取代。当然，随之而来的可能是全球范围内的隐私侵犯和社会控制，我将在下一章讨论这个问题。

超级人工智能何时到来？

人们经常让我预测超级人工智能何时到来，而我通常拒绝回答，原因有三。首先，历史上预测出错的情况经常出现。[29]例如，1960年，人工智能先驱、诺贝尔经济学奖得主赫伯特·西蒙（Herbert Simon）写道：“从技术上讲……20年内，机器将能够胜任人类能做的一切工作。”[30] 1967年，达特茅斯会议（1956年该研讨会首次提出“人工智能”的概念）的联合组织者马文·明斯基写道：“我相信，在一代人的时间内，机器会攻克几乎所有智力领域，创造‘人工智能’的问题将得到实质性解决。”[31]

第二个原因是，超级人工智能将要跨越的门槛并不明确。机器在某些领域已经超越了人类的能力。这些领域将会被拓展和深化，在我们拥有一个完全通用的超级人工智能系统之前，很有可能会出现“超级通用知识系统”“超级生物医学研究系统”“超级灵巧和敏捷的机器人”“超级企业规划系统”等。这些“不完全超级智能”的系统，无论是单独的还是集体的，都会产生许多与通用智能系统相同的问题。

第三个原因是，超级人工智能何时到来在本质上是不可预测的。正如约翰·麦卡锡在1977年的一次采访中指出的那样，超级人工智能需要“概念上的突破”。[32]他接着说：“要想造出超级人工智能，你需要的是1.7个爱因斯坦和0.3个曼哈顿计划，而你首先需要的是爱因斯坦。我相信这需要5到500年的时间。”在下一节中，我将解释一些概念上的突破可能是什么。它们到底有多不可预测？也许就像在卢瑟福宣布链式核反应完全不可能的几个小时后，西拉德就发明了链式核反应一样不可预测。

在2015年世界经济论坛的一次会议上，我回答了我们何时才能看到超级人工智能的问题。这次会议是按照“查塔姆宫守则”进行的，这意味着任何出席会议的人都不能对外发表任何言论。即便如此，出于谨慎的考虑，我还是在我的回答前面加上了“严格保密”4个字。我提出，除非发生灾难，否则超级人工智能很可能出现在我孩子的有生

之年。因为他们还很年轻，随着医学的进步，他们可能比参加会议的许多人活得更长。结果不到两个小时后，《每日电讯报》就刊登了一篇文章，引用了我的言论，还配上了狂暴的终结者机器人的图片，标题是《“反社会”机器人可能在一代人的时间里超越人类》。

我提出的时间点（比如说那之后80年）比典型的人工智能研究者的观点要保守得多。最近的调查表明，大多数活跃的研究人员预计人类水平的人工智能将在21世纪中叶到来。[33]我们在核物理学方面的经验表明，假设进展可能会很快出现，那么人们做好相应的准备才是明智的。如果只需要一个类似于西拉德发现中子诱发的链式核反应这种概念上的突破，某种形式的超级人工智能就可能会突然出现，那么到时我们很有可能毫无准备：如果我们制造出了具有任意程度自主能力的超级智能机器，我们很快就会发现人类无法控制它们。然而，我相当有信心，我们尚有喘息的空间，因为现在和超级人工智能之间还隔着好几个重大突破，而不仅仅是一个。

概念上的突破即将到来

要想创建通用的、人类级别的人工智能，我们还有许多路要走。解决这一问题不能靠把钱花在更多的工程师、更多的数据和更强大的计算机上。一些未来学家根据摩尔定律绘制图表，预测了未来计算能力的指数级增长，展示出机器将超越昆虫大脑、老鼠大脑、人类大脑，以及所有人类大脑的日期等。[34]这些图表毫无意义，原因我已经说过，速度更快的机器只会更快地给出错误的答案。如果将人工智能领域的领先专家聚集到一个拥有无限资源的团队中，目标是通过整合我们所有最好的想法来创建一个集成的、人类级别的智能系统，那么结果必将是失败。这个系统会在现实世界中崩溃。它不明白发生了什么，无法预测其行为的后果，无法理解人们在任意特定情况下想要什么，因此它会做蠢事。

通过了解系统将如何崩溃，人工智能研究人员找到达到了人类水平的人工智能必须解决的问题，这个问题就是概念上的突破。现在我将描述其中的一些遗留问题。一旦这些问题解决了，即使别的问题或许还存在，也绝对不会太多。

语言和常识

没有知识的智能就像没有燃料的发动机。人类从他人那里获得了大量的知识，知识以语言的形式代代相传。其中一些知识是事实，例如奥巴马在2009年当选美国总统，铜的密度约为每立方厘米8.92克，《乌尔纳姆法典》规定了对各种罪行的惩罚等。大量的知识存在于语言本身，存在于它所提供的概念之中。总统、2009年、密度、铜、克、厘米、罪行等词语都包含了大量的信息，这些信息代表了从发现和组织过程中提炼出来的精华，这才使它们最先存在于语言之中。

以铜为例，铜是指宇宙中一些原子的集合，我们把铜与“arglebarglium”做比较，arglebarglium是我对宇宙中完全随机选择的同样大小的原子集合的称呼。关于铜，我们可以发现许多普遍的、有用的、可预测的规律，例如铜的密度、导电性、延展性、熔点、起源、化合物、实际用途等；相比之下，关于arglebarglium我们基本上没什么可说的。一个有机体的语言如果由arglebarglium这样的词组成，这个有机体将无法工作，因为它永远不会发现使它能够建模并预测它的世界的规律。

一台真正理解人类语言的机器将能够迅速获取大量的人类知识，从而绕过在地球上生活过的1 000多亿人在几万年之中的学习过程。期待一台机器从原始的感官数据开始，从头开始重新发现这一切似乎是不切实际的。

然而目前，自然语言技术还不能胜任阅读和理解数百万本书的任务，其中许多书甚至会难倒一个受过良好教育的人。IBM的沃森在2011

年的《危险边缘》问答节目中战胜了两位人类冠军，一举成名。像沃森这样的系统，可以从表述明确的事实中提取简单的信息，但不能从文本中构建复杂的知识结构，也不能回答需要使用来自多个来源的信息进行大量推理的问题。例如，阅读1973年底之前所有的现存文献，评估（并解释）水门事件弹劾尼克松总统的可能结果，执行这种任务所需要的能力将远远超出其目前拥有的水平。

人们正在努力深化语言分析和信息提取的水平。例如，艾伦人工智能研究所的“阿里斯托”项目旨在建立一个能够在阅读教科书和学习参考书后通过学校科学考试的系统。[35]下面是四年级考试中的一道题：[36]

四年级学生计划进行一场轮滑比赛。以下哪种地面最适合比赛？

（A）碎石地 （B）沙地 （C）柏油地面 （D）草地

机器在回答这个问题时至少面临两个困难。第一个困难是经典的语言理解问题：分析句法结构，识别单词的含义等。（你可以自己尝试一下：使用在线翻译服务将这些句子翻译成一种不熟悉的语言，然后使用字典将其翻译回原来的语言。）第二个困难是对常识的需求：弄清楚“轮滑比赛”可能是脚上穿着旱冰鞋的人之间的比赛，而不是旱冰鞋之间的比赛；理解“地面”是参赛者将要在其上滑行的表面，而不是让观众坐在上面的表面；要知道“最适合”对于“比赛的地面”这一语境意味着什么等。请思考，如果我们把“四年级学生”替换成“严格的新兵训练营的教官”，答案可能会发生什么变化？

一种总结上述困难的方法是，阅读需要知识，而知识（在很大程度上）来自阅读。换言之，我们面临的是典型的先有鸡还是先有蛋的情况。我们可能希望采用自助算法[37]让系统读取一些简单的文本，获得一些知识，然后使用这些知识来读取更复杂的文本，获得更多的

知识等。不幸的是，发生的情况通常恰恰相反：系统所获得的知识大多是错误的，这会导致阅读错误，从而导致更多的错误知识等。

卡内基－梅隆大学的“NELL”（Never-Ending Language Learning，无止境语言学习）项目可能是目前正在进行的最雄心勃勃的采用自助算法的语言项目。从2010年到2018年，NELL通过在网上阅读英文文本获得了超过1.2亿个知识点。[38]其中一些知识点是准确的，比如枫叶队打冰球并赢得了斯坦利杯。除了事实之外，NELL还不断学习新的词汇、分类和语义关系。不幸的是，NELL仅对其3%的知识点有信心，它还要依靠人类专家定期清除错误或无意义的知识点，比如它认为“尼泊尔是一个国家，也被称为美国”和“价值是一种农产品，通常被分割成基础”。

我认为单一的突破不可能将恶性循环变成良性循环。基本的自助算法过程似乎是正确的：一个知道足够多事实的程序可以找出一个新句子所指的事实，从而学习一种新的表达事实的文本形式，然后让它发现更多事实，因此这个过程会继续下去。[谷歌联合创始人谢尔盖·布林（Sergey Brin）于1998年发表了一篇有关自助算法的重要论文。[39]]通过提供大量人工编码的知识和语言信息来启动循环肯定会有所帮助。增加事实表达的复杂性——考虑复杂的事件、因果关系、他人的想法和态度等，以及改善对词义和句子含义不确定性的处理，最终可能会带来一个自我强化而不是自我消亡的学习过程。

概念和理论的累积学习

大约14亿年前，在距离地球82万亿亿英里之遥，有两个黑洞，一个是地球质量的1 200万倍，另一个是地球质量的1 000万倍，二者靠得很近，开始相互环绕运行。在能量逐渐失去后，它们盘旋着越来越靠近对方，速度越来越快，在350公里的距离处达到了每秒250次的轨道频率，最后碰撞并融合。[40]在几毫秒内，引力波形式的能量释放

速率是宇宙中所有恒星总能量输出的50倍。2015年9月14日，这些引力波到达了地球。它们以25万亿亿分之一的系数交替地扩张和压缩空间，相当于把从地球到比邻星的距离（约4.4光年）压缩到人类头发丝的直径那么小。

幸好，在那两天之前，位于华盛顿和路易斯安那州的激光干涉引力波天文台（后简称LIGO）的探测器已经启动。探测器利用激光干涉测量法，能够测量空间的微小畸变。利用基于爱因斯坦广义相对论的计算，LIGO的研究人员已经预测出该事件中引力波的形状，并开始进一步确定其确切形状。[41]

研究人员之所以能够做到这些，是因为成千上万的人在数百年的观察和研究中积累和交流了知识与概念。从米利都的泰勒斯用羊毛摩擦琥珀并观察静电荷的积累，到伽利略从比萨斜塔投下小球，再到牛顿看到苹果从树上落下，经过成千上万次的观察，人类逐渐积累了一层又一层的概念、理论和装置：质量、速度、加速度、力、牛顿运动定律和万有引力定律、轨迹方程、电现象、原子、电子、电场、磁场、电磁波、狭义相对论、广义相对论、量子力学、半导体、激光、计算机等。

现在，原则上我们可以把这个发现过程理解为一个映射，即从全人类所有的感官数据到2015年9月14日LIGO的科学家观察电脑屏幕时体验到的感官数据的非常复杂的假设。这是纯数据驱动的学习观：输入的是数据，输出的是假设，中间是黑盒子。如果人们能够做到这一点，那将是“大数据、大网络”深度学习方法的典范，但这是不可能做到的。对于智能实体如何能够实现探测两个黑洞融合这样惊人的壮举，我们唯一看似合理的想法是，物理学的先验知识，加上仪器的观测数据，使得LIGO的科学家能够推断黑洞融合事件的发生。此外，这种先验知识本身就是用先验知识进行学习的结果，以此类推，一直可以追溯到遥远的历史。因此，我们对智能实体如何以知识为基础构建预测能力有了大致的了解。

我之所以说“大致”，无疑是因为在过去的几个世纪里，科学已经出现了一些错误的转折，暂时地追求了一些虚幻的概念，比如燃素和以太。但是我们知道一个事实，累积下来的图景就是实际发生的事情，也就是说，科学家一直在书和论文中写下他们的发现和理论。后来的科学家只能接触这些形式的显性知识，而无法接触更早的、早已过世的几代人的原始感官体验。LIGO团队的成员都是科学家，因此他们明白，他们使用的所有知识，包括爱因斯坦的广义相对论，都处于（而且将永远处于）试用期，而且可以通过实验被证伪。结果，LIGO的数据有力地证实了广义相对论，也进一步证明了引力子这种传递引力的假想粒子是无质量的。

我们距离创造出能够匹敌或超越科学界（或普通人一生）所累积的学问和发现能力的机器学习系统，还有很长的路要走。[42]深度学习系统（附录D）主要是数据驱动的，我们充其量只能在网络结构中“连接”一些非常弱的先验知识形式。概率编程系统（附录C）的确允许在学习过程中加入先验知识，正如概率知识库的结构和词汇表所表达的那样，但我们还没有有效的方法来生成新的概念和关系，并使用它们来拓展这类知识库。

困难并不在于如何找到能够很好地拟合数据的假设。深度学习系统可以找到完美拟合图像数据的假设，人工智能研究人员已经建立了符号学习程序，这一学习程序能够概括许多历史上定量科学规律的发现。[43]自主智能体进行学习所需要的远不只这些。

首先，进行预测所需要的“数据”应该包括什么？例如，在LIGO的实验中，用于预测引力波到达时空间伸展量和收缩量的模型考虑了碰撞黑洞的质量、它们轨道的频率等，但并不考虑当天是星期几，也不考虑当天的美国职业棒球大联盟的比赛情况。而预测旧金山海湾大桥交通的模型就会考虑当天是星期几和当天的职业棒球大联盟的比赛，但会忽略碰撞黑洞的质量和轨道频率。同理，学习识别图像中物体类型的程序使用像素作为输入，而学习评估古董价值的程序希望知

道古董是由什么构成的，是谁制造的，何时制造的，它的使用历史和所有权历史等。这是为什么？显然，这是因为我们人类已经对引力波、交通、视觉图像和古董有所了解。我们使用这些知识来决定预测特定输出需要哪些输入。这就是所谓的“特征工程”，要做好这一点，我们需要对特定的预测问题有很好的理解。

当然，真正的智能机器需要学习新知识时，不能依赖于每次都有人类特征工程师。它必须自己找出，是什么构成了这个问题的合理假设空间。我们推测，要做到这一点，它一定会以各种形式带来广泛的相关知识，但目前我们对此只有初步的想法。[44]纳尔逊·古德曼（Nelson Goodman）写于1954年的《事实、虚构和预测》[45]可能是关于机器学习的最重要但最被低估的书之一，他在书中提出了一个被称为“过度假设”的概念，这个概念有助于定义合理假设的空间。例如，在预测交通情况的例子中，相关的过度假设有：今天是星期几、当前的时间、当地事件、最近的事故、假期、交通延误、天气以及日出和日落时间。这些都会影响交通状况。（请注意，即使你不是交通专家，你也可以通过自己对世界的背景知识来找出这种过度假设。）智能学习系统可以积累和使用这种知识，以此帮助规划和解决新的学习问题。

第二点也许更加重要，这就是新概念的累积生成，如质量、加速度、电荷、电子和引力等。如果没有这些概念，科学家（和普通人）将只能在原始感知输入的基础上解释他们的宇宙，并做出预测。正因为有这些概念，牛顿才能够运用伽利略等人提出的质量和加速度的概念；卢瑟福才能够断定原子是由一个被电子包围的致密的带正电的原子核组成的，因为电子的概念已经在19世纪末被许多研究人员逐步开发出来了。事实上，一切科学发现都依赖于一层又一层的概念，这些概念可以追溯至历史和人类经验。

在科学哲学中，特别是在20世纪初，新概念的发现经常归因于三个妙不可言的“I”：直觉（intuition）、洞察力（insight）和灵感

（inspiration）。所有这些都被认为是对一切理性或算法解释的抵触。包括赫伯特·西蒙[46]在内的人工智能研究人员都强烈反对这一观点。简而言之，如果机器学习算法能够在假设空间中进行搜索，可以为输入中不存在的新术语添加定义，那么该算法就可以发现新概念。

例如，假设一个机器人试图通过观察人们玩西洋双陆棋来学习双陆棋规则。它观察人们掷色子的方法，并注意到玩家有时会移动3个或4个棋子，而不是1个或2个，这种情况出现在掷出的色子点数是1－1、2－2、3－3、4－4、5－5或6－6之后，程序如果可以添加一个新的“双倍”概念，将其定义为两个色子点数相同，就可以更简洁地表达相同的预测理论。这是一个直接的过程，它通过使用归纳逻辑编程之类的方法创建程序，提出新概念和定义，以此识别准确而简洁的理论。[47]

目前，我们知道如何在相对简单的情况下做到这一点，但是对更复杂的理论而言，引入的新概念的数量可能变得非常庞大。这使得深度学习方法最近在计算机视觉领域的成功更加引人注目。深度网络通常能够成功地找到有用的中间特征，比如眼睛、腿、条纹和拐角，即使深度网络使用的是非常简单的学习算法。如果我们能更好地理解这是如何发生的，那么我们就可以采用相同的方法，用科学所需的更具表现力的语言来学习新概念。这本身就是对人类的巨大恩惠，也是向通用人工智能迈出的重要一步。

发现动作

时间跨度长的智能行为，需要具备在多个抽象层次上分层规划和管理活动的能力，从攻读博士学位（可能涉及1万亿个动作），到给一根手指发送一个运动控制指令，从而键入求职信的字符，无所不包。

我们的活动被组织成具有几十个抽象级别的复杂层次结构。这些层次和它们所包含的动作是我们文明的关键组成部分，通过我们的语言和实践代代相传。例如，捕捉野猪、申请签证和购买机票等动作可能涉及数以百万计的原始动作，但我们可以将它们视为一个单元，因为它们已经在我们的语言和文化提供的“动作库”中，而且我们（大致）知道如何去做。

一旦它们进入了“动作库”，我们就可以把这些高层次的动作串联成更高层次的动作，比如在夏至举行部落宴会，或者在尼泊尔偏远地区进行夏季考古研究。试图从最低一级的运动控制步骤开始计划这样的活动是完全没有希望实现的，因为这样的活动涉及多达亿万个步骤，其中许多步骤是非常不可预测的。（例如在哪里可以找到野猪，它会朝哪个方向跑？）相反，因为在“动作库”中有合适的高级动作，所以人们只需要计划十几个步骤，每个步骤都是整个活动的很大一部分。这是我们脆弱的人类大脑也能做到的一件事，这给了我们进行长期规划的“超级力量”。

曾几何时，这些动作并不存在，例如，你要想在1910年获得乘飞机旅行的权利，就需要经过漫长、复杂且不可预测的研究、写信以及同各种各样的新兴航空公司谈判的过程。最近添加到“动作库”中的其他动作包括收发电子邮件、使用谷歌和优步。正如阿尔弗雷德·诺斯·怀特黑德（Alfred North Whitehead）在1911年所写的那样：“文明的进步是通过增加那些我们无须思考就能完成的重要动作来实现的。”[48]

索尔·斯坦伯格（Saul Steinberg）为《纽约客》绘制的著名封面（图6）以空间的形式精彩地展示了一个智能体如何管理自己的未来。对我来说，不久之后的未来是非常具体的，事实上，我的大脑已经“加载”了特定的运动控制序列来输入接下来的几个单词。再往未来看一点，细节就少了，我的计划是写完这一小节，吃午饭，再写点东西，然后观看世界杯决赛法国队和克罗地亚队的比赛。再往后，我

的计划更大，但更模糊：8月初从巴黎搬回伯克利，教授一门研究生课程，然后写完这本书。随着时间的推移，未来越来越接近现在，计划变得越来越详细。而新的、模糊的计划可能会添加到遥远的未来之中。不久之后的未来计划会变得非常具体，可以由运动控制系统直接执行。

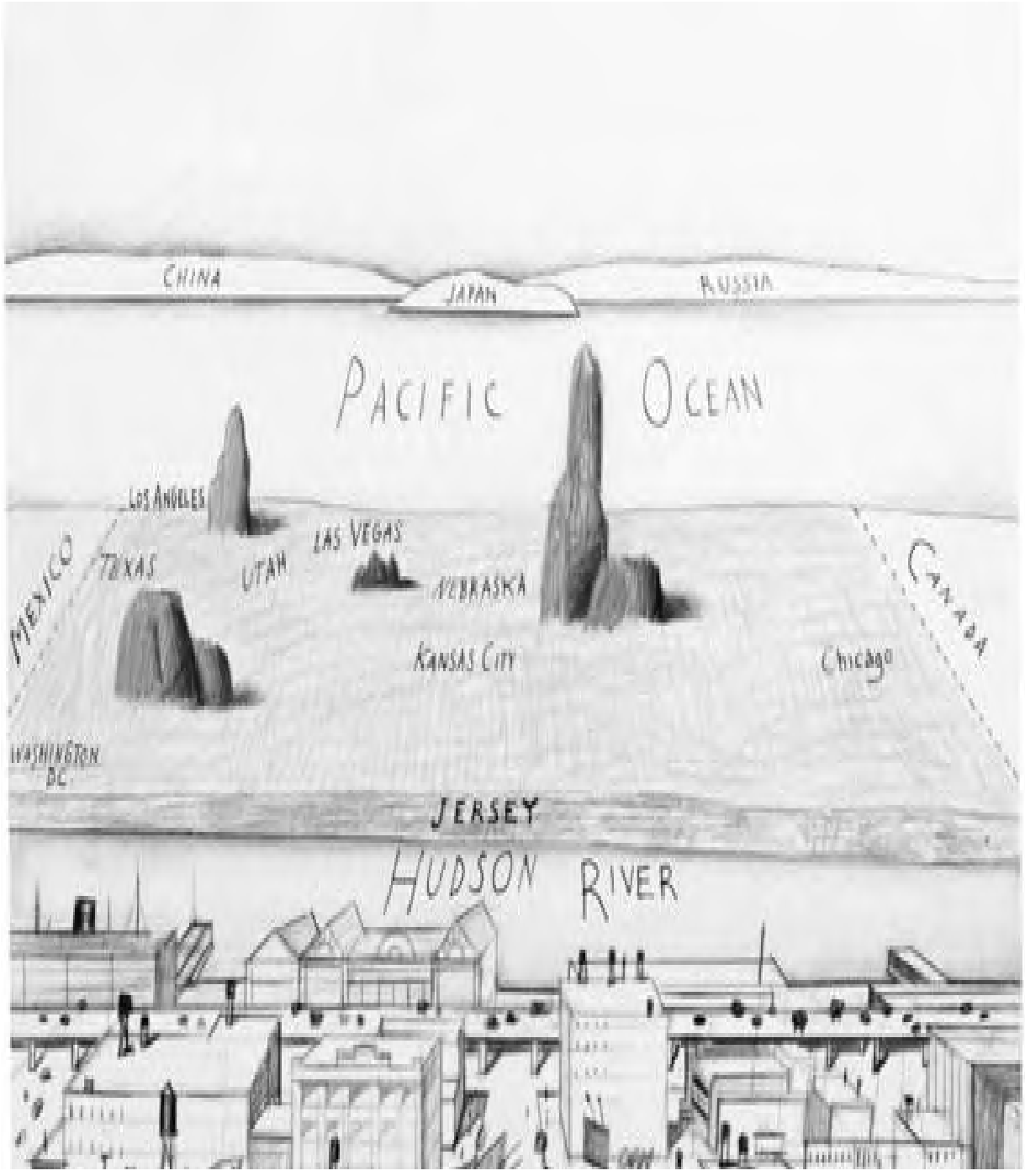




图6 1976年，索尔·斯坦伯格绘制的《从第9大道看世界》，首发于《纽约客》杂志封面

目前，对于人工智能系统，我们只有整体图景的某些部分。如果提供了抽象动作的层次结构——包括如何将每个抽象动作细化为由更具体的动作组成的子计划的知识，那么我们就有了可以构建复杂计划以实现特定目标的算法。有些算法可以执行抽象的、层次化的计划，其方式是让智能体始终保持一个原始的物理动作“准备好了”，即便未来的动作仍然是抽象的并且还不能执行。

要解决这个难题，首先要找到构建抽象动作层次结构的方法。例如，一个机器人完全从零开始，它只知道自己可以向各种马达输出各种电流，我们能否让它自己发现站立起来的动作？重要的是理解动作，我并不是在问我们是否可以训练一个机器人站立起来，这只需要通过应用强化学习，对机器人的头部远离地面进行奖励就可以实现。[49]训练机器人站立起来需要人类训练师知道站立起来的意义，这样才能定义正确的奖励信号。我们想要的是让机器人自己发现站立起来是一种有用的抽象动作，即一种实现行走、跑步、握手或翻墙的先决条件（直立），从而形成用来达成各种目标的许多抽象计划的一部分。同样，我们希望机器人能够发现各种动作（比如从一个地方移动到另一个地方、捡东西、开门、打结、做饭、找钥匙、盖房子），以及其他许多人类语言中没有名称的动作，因为我们人类还没有发现它们。

我相信这种能力是达到人类水平的人工智能所需要的最重要的一步。再次借用怀特黑德的话说，这将增加那些我们无须思考就能完成的重要行动。世界各地的许多研究小组都在努力解决这个问题。例如，DeepMind在2018年的一篇论文中写道，人工智能在《雷神之锤3》的夺旗模式中展示出了与人类相当的水平。该论文声称，他们的学习系统“以一种新颖的方式构建了时间的分层表征空间，以促进……在时间上连贯的动作序列”。[50]（我不完全确定这意味着什么，但听起来确实像是发明新的高层次动作的目标取得了进展。）我怀疑我们

还没有完整的答案，但这是一种随时可能发生的进步，我们只需要把一些现有的想法以正确的方式组合在一起即可。

具有这种能力的智能机器将会比人类看得更远。它们还可以考虑更多的信息。这两种能力的结合必然会让智能机器做出更好的现实决策。在人类与机器之间的一切对抗中，我们都会很快发现，就像加里·卡斯帕罗夫和李世石一样，我们的每一步行动都被机器预料到了，并被阻挠了。比赛还没开始，我们就输了。

管理精神活动

如果管理现实世界中的活动看起来很复杂，那就想想你可怜的大脑吧，管理这个“已知宇宙中最复杂的物体”的活动会是什么样子。我们一开始并不知道如何思考，就像我们一开始不知道如何走路或弹钢琴一样。我们学习如何去做这些事情。在某种程度上，我们可以选择想什么。（例如，想想多汁的汉堡或保加利亚的海关规定。）在某些方面，我们的精神活动比我们在现实世界中的活动更复杂，因为我们的大脑比我们的身体有更多的活动部分，而且这些部分的活动速度要快得多。计算机也是如此：AlphaGo在围棋棋盘上每走一步棋，都会执行亿万个计算单元，其中每个计算单元都涉及将一个分支添加到前向搜索树中，并评估该分支末尾的棋局。而且这一过程在每个计算单元中都会发生，因为程序可以选择下一步要探索搜索树的哪一部分。AlphaGo会相对精准地选择它所预计的将会改善它在棋盘上的最终决定的计算。

人们可以设计出一种合理的方案来管理AlphaGo的计算活动，因为该活动简单且同质化：每个计算单元都是相同的。与使用相同基本计算单元的其他程序相比，AlphaGo可能相当高效，但与其他类型的程序相比，它可能效率极低。例如，AlphaGo在2016年划时代的比赛中对战人类对手李世石，李世石每走一步棋的计算量可能只有几千个计算单元，但他拥有更灵活的计算体系结构，有更多种类的计算单位：将棋

盘分成各个子赛局，并试图解决各个子赛局之间的相互作用；认识到可能达到的目标，并制订高层次的计划，采取“保持这组棋子不要被对手包围”或“阻止对手把两组棋连起来”的行动；思考如何实现某个特定的目标，比如“保持这组棋子不要被对手包围”，并排除所有无法解决这一重大威胁的着法。

我们只是不知道如何组织如此复杂和多样的计算活动，即如何整合并构建每个计算活动的结果，以及如何将计算资源分配给各种审议活动，以便尽快找到好的决策。然而很明显，像AlphaGo这样简单的计算架构不可能在现实世界中奏效，在现实世界中，我们通常需要处理的决策范围不是几十个原始步骤，而是数十亿个原始步骤，而且任何时候可能采取的动作几乎都是无限的。重要的是要记住，现实世界中的智能体并不局限于下围棋，甚至不局限于找到一把钥匙——智能体就是生命。它接下来可以做任何事情，但它可能没有能力去把它可能做的所有事情全都考虑周全。

如果一个系统既能发现新的高级动作（如前所述），又能管理它的计算活动，从而将注意力集中在计算单元上，快速地显著提高决策质量，那么它将成为现实世界中一个强大的决策者。像人类一样，它的审议“在认知上是有效的”，但它不会受到短暂的短期记忆和缓慢的“硬件”的影响，人脑的这些局限性严重限制了我们展望未来、处理大量突发事件和考虑大量替代计划的能力。

还缺少什么？

如果我们把我们知道的关于处理这一章中列出的所有潜在的新发展的方法整合到一起，会有用吗？由此产生的系统会有什么表现？它将穿越时间，吸收大量信息，并通过大规模观察和推理来了解世界的状态。它将逐步改进它的世界模型（当然包括人类模型），并使用这些模型来解决复杂的问题，还会封装和反复使用其解决过程来提高审

议效率，解决更复杂的问题。它会发现新的概念和动作，这会让它提高发现率。它还会在越来越长的时间跨度上制订有效的计划。

总而言之，从有效地实现其目标的系统角度来看，目前并没有明显的重要内容被遗漏。当然，唯一可以确定的方法就是先构建它，然后（在取得了突破之后）看看会发生什么。

想象一台超级智能机器

在讨论超级人工智能的性质和影响时，技术界遭遇了想象力上的失败。我们经常看到关于减少医疗事故[51]、更安全的汽车[52]，以及其他增量性质的进步的讨论。机器人往往被想象成携带大脑的个体实体，而事实上，它们很可能无线连接成一个单一的全球实体，该实体需要大量固定的计算资源。研究人员似乎害怕成功地研究出人工智能后，可能产生的真正后果。

根据假设，通用智能系统可以做任何人能做的事情。例如，一些人做了大量的数学工作、算法设计、编码和实证研究，推出了现代搜索引擎。所有搜索结果都非常有用，当然也非常有经济价值。这些结果究竟有多值钱呢？最近的一项研究显示，在接受调查的美国成年人中，你至少需要支付给他们每人1.75万美元才能让他们一年内放弃使用搜索引擎，[53]换算后，搜索引擎的全球价值高达数万亿美元。

现在请你想象一下，搜索引擎还不存在，因为历经几十年的必要工作还没有完成，但你可以访问一个超级人工智能系统。你只需提出问题，就可以通过人工智能系统使用搜索引擎技术。搞定了！获得数万亿美元的价值，你只需向它提问，没有一行额外的代码是你写的。这同样适用于任何其他缺失的发明或一系列发明：如果人类可以做到，那么机器也可以做到。

最后一点为超级智能机器的能力提供了一个有用的下限，即一个悲观的估计。根据假设，机器比个人更有能力。有许多事情是一个人做不到的，n个人合作却可以做到，例如把宇航员送上月球，创造引力波探测器，给人类基因组测序，管理一个拥有数亿人口的国家。因此，粗略地讲，我们创建机器的n个软件副本，并且用与连接n个人相同的方式（使用相同的信息和控制流）来连接它们。现在我们有了一台机器，它可以做n个人能做的任何事情，而且只会做得更好，因为它的n个组成部分都是超人。

智能系统的这种多智能体合作设计只是机器的可能能力的下限，因为还有其他更好的设计。在n个人组成的集合中，所有可用的信息分别保存在n个人的大脑中，他们之间的交流非常缓慢且不完美。这就是为什么人类把大部分时间花在开会上。机器不需要这种分隔，分隔通常会阻碍各个点之间的连接。作为科学发现互相分隔的一个例子，简单地了解一下青霉素的悠久历史会让人大开眼界。[54]

拓展想象力的另一个有用方法是对某种特定形式的感官输入进行思考——比如阅读，并将其放大。人类可以在一周内阅读并理解一本书，而机器可以在几个小时内阅读并理解人类写过的全部1.5亿本书。这需要相当强的处理能力，但机器基本上可以并行阅读这些书，这意味着只需添加更多的芯片，机器就可以扩大其阅读范围。同理，这台机器还可以通过卫星、机器人和数亿个监控摄像头同时看到一切，观看世界上所有的电视节目，收听世界上所有的广播电台和电话交谈。很快，它就能对这个世界及其居民有更详细、更准确的了解，这是任何人类都不可能做到的事情。

人们还可以想象拓展机器的行动能力。人类只能直接控制一个身体，而机器可以控制成千上万个机体。一些自动化工厂已经表现出这种特点。例如，在工厂外，一台控制着数千个灵巧机器人的机器可以建造大量的房屋，每间房屋都是根据未来居住者的需要和愿望量身定制的。在实验室里，现有的用于科学研究的机器人系统可以扩大规

模，同时进行数百万个实验——也许是为了创建完整的直至分子水平的人体生物学预测模型。请注意，机器的推理能力将使它具有更强的能力来检测科学理论之间、理论与观测之间的不一致。我们可能已经有了足够的生物学实验证据来设计治疗癌症的方法，只不过我们还没有把它们整合在一起。

在网络领域，机器已经可以访问数十亿个效应器，即世界上所有的电话和计算机上的显示器。这在一定程度上解释了IT公司用极少的员工创造巨大财富的能力，还表明了人类通过屏幕进行操作存在严重的安全隐患。

另一种不同的拓展来自机器拥有比人类更准确的展望未来的能力。我们已经在国际象棋和围棋中看到了这一点。凭借在长时间跨度上生成和分析分层计划的能力，以及识别新的抽象动作和高级描述性模型的能力，机器将把这种优势转移到数学（证明新的、有用的定理）和现实世界中的决策制定等领域。在发生环境灾难时，疏散大城市人口的任务将会变得相对简单，机器可以为每个人和每辆车生成单独的指导，以最大限度地减少人员伤亡。

在制定防止全球变暖的政策时，机器可能会费一番工夫。地球系统建模需要利用物理学（大气、海洋），化学（碳循环、土壤），生物学（分解、迁移），工程学（可再生能源、碳捕获），经济学（工业、能源利用），人性（愚蠢、贪婪）和政治学（更愚蠢、更贪婪）的知识。如前所述，机器将能够获得大量的证据来支持所有这些模型。它能够建议或进行新的实验和探索，以减少不可避免的不确定性，例如，探索浅海中天然气水合物的真实储量。它能够考虑一系列可能的策略建议，例如法律、推广、市场、发明和地球工程的干预措施，但当然，它也需要找到说服我们赞同它的方法。

超级智能的局限

当你发挥想象力的时候，请不要太漫无边际。一个常见的错误是人们倾向于认为超级人工智能系统拥有无所不知的神圣力量——不仅拥有现在的，而且拥有未来的完整和完美的知识。[55]这令人难以置信，因为这需要一种非人的能力来确定当前世界的确切状态，以及用一种人类无法实现的能力来模拟一个包括机器本身在内的世界的运行（更不用说，这个世界中还有数十亿个人类大脑，大脑仍然是宇宙中第二复杂的物体），这种模拟的速度比实时模拟还要快得多。

这并不是说以合理的确定度预测未来的某些方面是不可能的。例如，尽管混沌理论的学者坚持认为有蝴蝶效应之类的东西，但是我知道一年后我将会在伯克利的哪间教室里教什么课。（反过来，我不认为人类能够按照物理定律精准地预测未来！）预测成功与否取决于是否有正确的抽象，例如，我可以预测“我”将在4月的最后一个星期二在伯克利校园的惠勒礼堂上台演讲，但我无法预测自己精确到毫米的位置，也无法预测那时我的身体中将包含哪些碳原子。

机器还受到现实世界对新知识获取速度的限制，这是凯文·凯利（Kevin Kelly）在关于超人类人工智能的过度简化预测的文章中提出的有效观点之一。[56]例如，为了确定一种特定的药物是否能治愈试验动物体内的某种癌症，人类科学家或机器科学家有两种选择：一是给动物注射药物并等待数周，二是进行足够精确的模拟。然而，运行模拟需要大量的生物学经验知识，其中一些知识目前尚不具备。因此，必须先进行更多的建模实验。毫无疑问，这需要时间，而且必须在现实世界中完成。

机器科学家可以让大量建模实验并行，将它们的结果整合到一个内部一致（尽管非常复杂）的模型中，并且将模型的预测与生物学已知的全部实验数据进行比较。模拟该模型并不一定需要对整个生物体进行量子力学模拟，直至单个分子反应的水平，如凯利所言，这要比在现实世界中做实验花费更多的时间。正如我可以有一定的把握预测未来4月的星期二自己会在哪里一样，生物系统的特性也可以用抽象模

型精确预测。（其中一个原因是，生物利用基于总反馈回路的“鲁棒控制系统”运行，所以初始条件的微小变化通常不会导致结果出现大变化。）因此，虽然在经验科学中，由机器瞬间产生科学发现是不可能的，但我们可以期待，在机器的帮助下，科学的发展速度将会更快。事实上，情况已经如此了。

机器的最后一个局限在于它们不是人类。这使它们在试图模拟和预测人类这种特定的对象时存在固有的劣势。人类的大脑都很相似，所以如果愿意，我们可以用自己的大脑来模拟他人的精神和情感生活。对我们来说，这是很轻松的。（仔细想想，机器之间还有一个更大的优势：它们实际上可以运行彼此的代码！）例如，我不需要成为神经感觉系统的专家，就能知道当你用锤子敲拇指时是什么感觉。我只要用锤子敲我的拇指就行了。而机器对人类的理解几乎是从零开始的：[57]它们只能触及我们的外部行为，以及所有的神经科学和心理学文献，并且必须在此基础上理解我们是如何工作的。原则上，它们将能够做到这一点，但我们有理由认为，它们达到或超越人类的理解水平将比习得其他大多数能力花费更长的时间。

人工智能如何造福人类？

我们的智能对我们的文明负责。有了获得更强大智能的途径，我们便可以拥有一个更伟大的，也许是更好的文明。人们可以思考如何解决一些重大的开放性问题，比如无限延长人类寿命、超光速旅行等，但这些科幻小说中的主要内容还不是推动人工智能进步的动力。（借助超级人工智能，我们可能会发明出各种类似魔法的技术，但现在很难说这些技术会是什么。）相反，让我们考虑一个更平凡的目标：以可持续的方式将地球上每个人的生活水平提高到发达国家认为相当体面的水平。我们（有些武断地）将美国的第88个百分位定为

“体面”的标准，这个既定目标代表，全球GDP（国内生产总值）要增长将近10倍，即从每年76万亿美元增长到750万亿美元。[58]

为了计算回报的现金价值，经济学家使用收入流的“净现值”，净现值考虑了未来收入对于当前收入的贴现。假设贴现率为5%，那么每年674万亿美元的额外收入的净现值约为13 500万亿美元。[59]因此，粗略地说，如果人工智能可以为每个人提供体面的生活，那么它的价值可能大致就是这个金额。有了这样的数字，公司和国家每年在人工智能的研发上投资数百亿美元就不足为奇了。[60]即便如此，与回报的规模相比，投资的金额也是微不足道的。

当然，这些都是虚构的数字，除非有人对人类水平的人工智能如何实现提高生活水平的壮举有所了解。要做到这一点，只有提高商品和服务的人均产量。换言之，普通人永远不能期望消费超过普通人的产出。本章前面讨论的无人驾驶出租车的例子说明了人工智能的乘数效应，比方说，10个人就可以管理一个拥有1 000辆车的车队，所以每个人的交通产出是以前的100倍。同样的道理也适用于制造汽车和获取制造汽车的原材料。事实上，在温度经常超过45摄氏度的澳大利亚北部，一些铁矿石开采作业几乎已经完全自动化了。[61]

人工智能目前的这些应用都是具体用途的系统：无人驾驶汽车和自动运营的矿山需要在研究、机械设计、软件工程和测试方面投入大量资金，用于开发必要的算法，并确保它们按预期运行。在工程学的各个领域，事情无一例外都是这样做的。过去，个人出行也是这样，如果你想在17世纪从欧洲到澳大利亚旅行，就会涉及一个庞大的项目，这个项目耗资巨大，需要多年的计划，而且有很高的死亡风险。现在，我们已经习惯了“交通即服务”（TaaS）的理念：如果你下周初需要去墨尔本，只需在手机上点几下，再花点钱就可以了。

通用人工智能是“一切即服务”（Everything-as-a-Service，EaaS）。我们没有必要为了执行一个项目而雇用不同学科的专家队

伍，把他们按照承包商和分包商的等级组织起来。通用人工智能的所有化身将会获得人类的所有知识和技能，而且还不只是这样。它们唯一的区别在于实体能力：用于造建筑或做手术的四肢灵巧的机器人，用于大规模货物运输的轮式机器人，用于空中检查的四轴飞行器机器人等。撇开政治和经济不谈，原则上每个人都可以拥有一个由软件智能体和实体机器人组成的完整的组织，它能够设计和建造桥梁，提高作物产量，为100位客人做饭，举行选举，以及完成一切需要做的事情。使这一切成为可能的正是通用智能的通用性。

当然，历史已经表明，在没有人工智能的情况下，全球人均GDP增长近10倍是可能的，只是人们最终实现这一增长花了190年（从1820年到2010年）。[62]这需要工厂、机床、自动化、铁路、钢铁、汽车、飞机、电力、石油和天然气生产、电话、广播、电视、计算机、互联网、卫星和许多其他革命性发明的发展。前文中所假设的GDP增长近10倍，并不是基于进一步的革命性技术，而是基于人工智能系统更有效、更大规模地利用我们现有资源的能力。

当然，除了提高生活水平的纯物质利益外，人工智能还会产生其他影响。例如，众所周知，一对一辅导远比课堂教学更有效，但如果是由人来一对一辅导，绝大多数人根本负担不起。有了人工智能导师，无论来自多么贫穷的家庭，每个孩子的潜力都可以被发掘。辅导费可以忽略不计，而这个孩子将会过上更加富裕和高效的生活。对艺术和智力的追求，无论是个人的还是集体的，都将成为生活中正常的一部分，而不是稀有的奢侈品。

在健康领域，人工智能系统能让研究人员揭开并掌握人类生物学的巨大复杂性，从而逐步消除疾病。对人类心理学和神经化学的更深入的了解会广泛改善心理健康。

也许更不同寻常的是，人工智能可以为VR（虚拟现实）提供更加高效的创作工具，并且可以用更有趣的实体填充VR环境。这可能会使

VR成为文学和艺术表达的又一媒介，创造出目前无法想象的丰富和深度的体验。

在日常生活的平凡世界里，一个聪明的助手和向导（如果设计得好，不受经济和政治利益的影响）能够使每个人在一个日益复杂的、有时充满敌意的经济和政治体系中有效地代表自己行事。实际上，你有一群精力旺盛的律师、会计师和政治顾问在随时待命。就像交通拥堵有望通过混合使用哪怕一小部分无人驾驶汽车来缓解一样，人们只能希望，在信息更灵通的、得到更好建议的全球公民之中，政策会更明智，冲突会更少。

这些发展合在一起可能会改变历史，至少改变由群体内部和群体之间为获得生活必需品而发生的冲突所驱动的那部分历史。如果馅饼本质上是无限的，那么你与他人争夺更大的份额就没有意义了。这就像是为了得到更多的电子版报纸而争吵一样，当任何人都可以免费制作任意数量的数字副本时，争吵就完全没有意义了。

人工智能可以提供的功能是受限的。土地和原材料并不是无限的，因此人口不可能无限增长，也不可能每个人都在私人公园里拥有豪宅。（因为那需要在太阳系的其他地方采矿，并在太空中建造人工栖息地，但我承诺过不谈科幻小说。）骄傲也是受限的：在给定的所有指标中，只有1%的人可以排在前1%。如果跻身前1%才能感到幸福，那么99%的人就不会幸福，即使那些处在前1%末尾的人也拥有客观上美好的生活方式。[63]因此，对我们的文化而言，逐渐淡化骄傲和嫉妒是感知自我价值的重要因素，这一点很重要。

正如尼克·波斯特洛姆（Nick Bostrom）在他的书《超级智能》的结尾所说的那样，人工智能的成功将催生“一条文明轨迹，实现富有同情心地使用人类被赋予的宇宙资源的文明的进步”。如果我们不能充分利用人工智能的优势，那么我们只能怪自己。

[1] 游戏算法的基本方案是由克劳德·香农提出的。Claude Shannon, “Programming a computer for playing chess,” *Philosophical Magazine*, 7th ser., 41 (1950):256–75.

[2] 请参见：Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern App roach*, 1st ed. (Prentice Hall, 1995). 图5.12。请注意，对国际象棋选手和国际象棋程序的评分并不是一门精确的科学。卡斯帕罗夫职业生涯最高的Elo评分是在1999年达到的2 851分，超过之前的2 805分，但目前的国际象棋程序（如Stockfish）的评分达到3 300分或更高。

[3] 最早报道的公共道路上的无人驾驶汽车：Ernst Dickmanns and Alfred Zapp, “Autonomous high speed road vehicle guidance by computer vision,” *IFAC Proceedings Volumes* 20 (1987): 221–26。

[4] 谷歌汽车（后来的Waymo）汽车的安全记录：“Waymo safety report: On the road to fully self-driving,” 2018。

[5] 到目前为止，至少有两名司机和一名行人丧生。下面是一些参考资料，以及描述发生的事件的简短引言。Danny Yadron and Dan Tynan, “Tesla driver dies in first fatal crash while using autopilot mode,” *Guardian*, June 30, 2016:“The autopilot sensors on the Model S failed to distinguish a white tractor-trailer crossing the highway against a bright sky.” Megan Rose Dickey, “Tesla Model X sped up in Autopilot mode seconds before fatal crash, according to NTSB,” Tech Crunch, June 7, 2018: “At 3 seconds prior to the crash and up to the time of impact with the crash attenuator, the Tesla’s speed increased from 62 to 70.8 mph, with no precrash braking or evasive steering movement detected.” Devin Coldewey, “Uber in fatal crash detected pedestrian but had emergency braking disabled,” *Tech Crunch*, May 24, 2018: “Emergency braking maneuvers are not enabled while the vehicle is under computer control, to reduce the potential for erratic vehicle behavior.”

[6] 美国汽车工程师学会定义了6个级别的自动驾驶：其中L0级代表完全没有自动化，L5级代表完全自动化。

[7] 自动化对运输成本的经济影响预测：Adele Peters, “It could be 10 times cheaper to take electric robo-taxis than to own a car by 2030,” *Fast Company*, May 30,2017。

[8] 事故对无人驾驶汽车监管行动前景的影响：Richard Waters, “Self-driving car death poses dilemma for regulators,” *Financial Times*, March 20, 2018。

[9] 事故对无人驾驶汽车的公众认知的影响：Cox Automotive, “Autonomous vehicle awareness rising, acceptance declining, according to Cox Automotive mobility study,” August 16, 2018。

[10] 最初的聊天机器人：Joseph Weizenbaum, “ELIZA — a computer program for the study of natural language communication between man and machine,” *Communications of the ACM* 9 (1966): 36–45。

[11] 有关生理建模的当前活动，请参见：physiome.org。20世纪60年代的工作是用数千个微分方程组建模型：Arthur Guyton, Thomas Coleman, and Harris Granger, “Circulation: Overall regulation,” *Annual Review of Physiology* 34(1972): 13–44。

[12] 一些关于教学系统的早期工作是由斯坦福大学的帕特里克·苏佩斯及其同事完成的：Patrick Suppes and Mona Morningstar, “Computer-assisted instruction,” *Science* 166 (1969): 343–50。

[13] Michael Yudelson, Kenneth Koedinger, and Geoffrey Gordon, “Individualized Bayesian knowledge tracing models,” in *Artificial Intelligence in Education: 16th International Conference*, ed. H. Chad Lane et al. (Springer, 2013).

[14] 安全多方计算技术（secure multiparty computation）主要针对在无可信第三方的情况下，如何安全地计算一个约定函数的问题。安全多方计算是电子选举、门限签名以及电子拍卖等诸多应用得以实施的密码学基础。——译者注

[15] 有关加密数据的机器学习的示例，请参见：Reza Shokri and Vitaly Shmatikov,“Privacy-preserving deep learning,” in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (ACM, 2015)。

[16] 第一个关于智能家居的回顾展，基于它的发明者詹姆斯·萨瑟兰的一次演讲：James E. Tomayko, “Electronic Computer for Home Operation (ECHO): The first home computer,” *IEEE Annals of the History of Computing* 16 (1994): 59–61。

[17] 1磅≈453.5924克。——编者注

[18] 基于机器学习和自动决策的智能家居项目摘要：Diane Cook et al., “Mav Home:An agent-based smart home,” in *Proceedings of the 1st IEEE International Conference on Pervasive Computing and Communications* (IEEE, 2003)。

[19] 关于智能家居用户体验分析的开端，请参见：Scott Davidoff et al., “Principles of smart home control,” in *Ubicomp 2006: Ubiquitous Computing*, ed. Paul Dourish and Adrian Friday (Springer, 2006)。

[20] 基于人工智能的智能家居的商业公告，请参见：“The Wolff Company unveils revolutionary smart home technology at new Annadel Apartments in Santa Rosa,California,” *Business Insider*, March 12, 2018。

[21] 关于机器人厨师作为商业产品的文章：Eustacia Huen, “The world’s first home robotic chef can cook over 100 meals,” *Forbes*, October 31, 2016。

[22] 我在加州大学伯克利分校的同事关于机器人运动控制的深度强化学习的报告：Eustacia Huen, “The world’s first home robotic chef can cook over 100 meals,”*Forbes*, October 31, 2016。

[23] 关于数十万仓库工人工作自动化的可能性：Tom Simonite, “Grasping robots compete to rule Amazon’s warehouses,” *Wired*, July 26, 2017。

[24] 我假设每一页占用一分钟的笔记本电脑CPU计算时间，即10^{11}次运算。谷歌的第三代TPU每秒运行大约10^{17}次运算，这意味着它每秒可以阅读100万页内容，换言之，阅读8000万本200页的书大约需要5个小时。

[25] 2003年一项关于所有渠道产生的全球信息量的研究：Peter Lyman and Hal Varian, “How much information?” sims.berkeley.edu/research/projects/how-muchinfo-2003。

[26] 有关情报机构使用语音识别的详细信息，请参见：Dan Froomkin, “How the NSA converts spoken words into searchable text,” *The Intercept*, May 5, 2015。

[27] 分析来自卫星的视觉图像是一项艰巨的任务：Mike Kim, “Mapping poverty from space with the World Bank,” Medium.com, January 4, 2017。迈克·金估计有800万人全天候工作，也就是说有超过3 000万人每周工作40小时。我怀疑这实际上有些高估，因为绝大多数图像在一天的时间里展示出的变化微不足道。另外，美国情报部门雇用成千上万的人坐在巨大的房间里盯着卫星图像，只是为了跟踪某个感兴趣的小区域正在发生的事情，所以100万人对整个世界来说可能是恰当的。

[28] 在建立基于实时卫星图像数据的全球观测站方面已经取得了实质性进展：David Jensen and Jillian Campbell, “Digital earth: Building, financing and governing a digital ecosystem for planetary data,” white paper for the UN Science Policy-Business Forum on the Environment, 2018。

[29] 卢克·米尔豪泽（Luke Muehlhauser）写了大量关于人工智能预测的文章，我很感激他为接下来的引文找到了原始来源。请参见：Luke Muehlhauser,“What should we learn from past AI forecasts?” Open Philanthropy Project report,2016。

[30] 人类水平的人工智能在20年内到来的预测：Herbert Simon, *The New Science of Management Decision* (Harper & Row, 1960)。

[31] 人类水平的人工智能在一代人内到来的预测：Marvin Minsky, *Computation: Finite and Infinite Machines* (Prentice Hall, 1967)。

[32] 约翰·麦卡锡对人类水平的人工智能将在“5到500年内”到来的预测：IanShenker, “Brainy robots in our future, experts think,” *Detroit Free*

Press,September 30, 1977。

[33] 有关人工智能研究人员对人类水平的人工智能到来的估计的调查摘要，请参见：aiimpacts.org。卡蒂娅·格雷斯（Katja Grace）等人对人类水平的人工智能的调查结果进行了深入的讨论，“When will AI exceed human performance?Evidence from AI experts,” arXiv: 1705.08807v3 (2018)。

[34] 关于原始计算机能力与脑力对比的图表，请参见：Ray Kurzweil, “The law of accelerating returns,” Kurzweilai.net, March 7, 2001。

[35] 艾伦研究所的“Project Aristo”：allenai.org/aristo。

[36] 关于在四年级的理解和常识考试中取得良好成绩所需知识的分析，请参见：Peter Clark et al., “Automatic construction of inference-supporting knowledge bases,” in *Proceedings of the Workshop on Automated Knowledge Base Cons* truction (2014), akbc.ws/2014。

[37] 自助算法（bootstrapping）指利用有限的样本资料经由多次重复抽样，重新建立起可以代表母体样本分布的新样本。——译者注

[38] 汤姆·米切尔（Tom Mitchell）等人关于机器阅读的“NELL”项目的描述：“Neverending learning,” *Communications of the ACM* 61 (2018): 103–15。

[39] 从文本中引导推论的想法是由谢尔盖·布林提出的，他在《万维网和数据库》中写道：“从万维网中提取模式和关系。”Paolo Atzeni, *Alberto Mendelzon, and Giansalvatore Mecca* (Springer, 1998).

[40] 关于LIGO探测到的黑洞碰撞的可视化信息，请参见：LIGO Lab Caltech, “Warped space and time around colliding black holes,” February 11, 2016,youtube.com/watch?v=1agm33iEAuo。

[41] 描述引力波观测的第一份出版物：Ben Abbott et al., “Observation of gravitational waves from a binary black hole merger,” *Physical Review Letters* 116(2016): 061102。

[42] 关于婴儿作为科学家的研究：Alison Gopnik, Andrew Meltzoff, and Patricia Kuhl, *The Scientist in the Crib: Minds, Brains, and How Children Learn* (William Morrow, 1999)。

[43] 关于对实验数据进行自动化科学分析以发现规律的几个项目的摘要：Patrick Langley et al., *Scientific Discovery: Computational Explorations of the Creative Processes* (MIT Press, 1987)。

[44] 在先验知识指导下的机器学习的一些早期工作：Stuart Russell, *The Use of Knowledge in Analogy and Induction* (Pitman, 1989)。

[45] 古德曼对归纳法的哲学分析仍然是灵感的来源：Nelson Goodman, *Fact, Fiction, and Forecast* (University of London Press, 1954)。

[46] 一位资深人工智能研究人员抱怨科学哲学中的神秘主义：Herbert Simon, “Explaining the ineffable: AI on the topics of intuition, insight and inspiration,” in *Proceedings of the 14th International Conference on Artificial Intelligence*, ed. Chris Mellish (Morgan Kaufmann, 1995)。

[47] 该领域两位创始人对归纳逻辑编程的综述：Stephen Muggleton and Luc de Raedt, “Inductive logic programming: Theory and methods,” *Journal of Logic Programming* 19–20 (1994): 629–79。

[48] 关于将复杂运算封装为新的原始动作的重要性的早期论述，请参见：Alfred North Whitehead, *An Introduction to Mathematics* (Henry Holt, 1911)。

[49] 证明模拟机器人可以独立学习站立的工作：John Schulman et al., “Highdimensional continuous control using generalized advantage estimation,”arXiv:1506.02438 (2015)。A video demonstration is available at youtube.com/watch?v=SHLuf2ZBQSw.

[50] 对学习玩夺旗类电子游戏的强化学习系统的描述：Max Jaderberg et al., “Human-level performance in first-person multiplayer games with population-based deep reinforcement learning,” arXiv:1807.01281 (2018)。

[51] 人工智能未来几年的发展展望：Peter Stone et al., “Artificial intelligence and life in 2030,” *One Hundred Year Study on Artificial Intelligence*, report of the 2015Study Panel, 2016。

[52] 埃隆·马斯克和马克·扎克伯格二人在媒体上的争论：Peter Holley, “Billionaire burn: Musk says Zuckerberg’s understanding of AI threat ‘is limited,’ ” *The Washington Post*, July 25, 2017。

[53] 关于搜索引擎对个人用户的价值：Erik Brynjolfsson, Felix Eggers, and Avinash Gannamaneni, “Using massive online choice experiments to measure changes in wellbeing,” working paper no. 24514, National Bureau of Economic Research,2018。

[54] 青霉素被发现过多次，它的疗效在医学出版物中有所描述，但似乎没有人注意到。请参见：en.wikipedia.org/wiki/History_of_penicillin。

[55] 关于无所不知、洞察一切的人工智能系统的一些更深层风险的讨论，请参见：David Auerbach, “The most terrifying thought experiment of all time,” *Slate*,July 17, 2014。

[56] 关于高级人工智能思考的一些潜在陷阱的分析：Kevin Kelly, “The myth of a superhuman AI,” *Wired*, April 25, 2017。

[57] 机器可能与人类共享认知结构的某些方面，特别是那些涉及对物理世界的感知和操纵以及自然语言理解所涉及的概念结构的方面。由于硬件方面的巨大差异，它们的审议过程可能会大不相同。

[58] 根据2016年的调查数据，第88个百分位相当于每年10万美元：American Community Survey, US Census Bureau, www.census.gov/programs-surveys/acs。同年，全球人均国内生产总值为10 133美元：National Accounts Main Aggregates Database, UN Statistics Division, unstats.un.org/unsd/snaama。

[59] 如果GDP增长阶段超过10年或20年，那么它的价值分别为9 400万亿美元或6 800万亿美元，仍然不容小觑。在一份有趣的历史记录中，推广智力爆炸概念的I. J.古德估计，人类级别的人工智能的价值至少是“一个大凯恩斯”，他指的是传说中的经济学家约翰·梅纳德·凯恩斯。凯恩斯的贡献价值在1963年估计为1 000亿英镑，因此一个大凯恩斯在2016年的价值约为220亿亿美元。古德将人工智能的价值主要寄托在它确保人类无限期生存的潜力上。后来，他开始考虑是否应该加上一个负号。

[60] 欧盟宣布了2019—2020年期间240亿美元的研发支出计划。请参见：European Commission, “Artificial intelligence: Commission outlines a European approach to boost investment and set ethical guidelines,” press release, April 25, 2018。中国在2017年宣布了人工智能长期投资计划，设想到2030年，人工智能核心产业每年产生1 500亿美元的收入。请参见：Paul Mozur, “Beijing wants A.I. to be made in China by 2030,” *The New York Times*, July 20, 2017。

[61] 请参见：Rio Tinto’s Mine of the Future program at riotinto.com/australia/pilbara/mine-of-the-future-9603.aspx。

[62] 经济增长的回顾性分析：Jan Luiten van Zanden et al., eds., *How Was Life? Global Well-Being since 1820* (OECD Publishing, 2014)。

[63] 追求相对于他人的优势，而不是绝对生活质量，这种追求是一种地位商品。参见第9章。

04 人工智能的濫用



富有同情心地使用人类被赋予的宇宙资源听起来很美妙，但我们也必须考虑到作恶的创新也在快速发展。心怀恶意的人们正在迅速地想出滥用人工智能的新方法，以致这一章甚至可能在付梓之前就已经过时了。然而，你不应该把这本书看作令人沮丧的读物，而是要把它看作在为时已晚之前采取行动的号召。

监督、劝导和控制

自动化史塔西

史塔西（Stasi）是德意志民主共和国国家安全部更广为人知的名字，它被广泛认为是“有史以来最高效、最具镇压性的情报和秘密警察机构之一”。[1]这里保存着民主德国绝大多数家庭的档案。它监控电话、阅读信件，并在公寓和酒店安装隐藏的摄像头。它在识别和消除持不同政见者的活动方面非常有效。它的惯用手法是心理恐吓，而不是监禁或处决。然而，这种程度的控制付出了巨大的代价。据估计，当时超过1/4的劳动年龄人口是史塔西的线人。另据估计，史塔西保存的纸质档案有200亿页，[2]处理大量传入信息流并对此做出反应的任务开始超出所有人类组织的能力。

因此，情报机构在工作中使用人工智能就不足为奇了。多年来，这些机构一直在应用简单形式的人工智能技术，包括语音识别，语音和文本中的关键词、短语识别。人工智能系统越来越能够理解人们所说和所做的内容，无论是在语音、文本中，还是在视频监控中。在以控制为目的而采用这种技术的政权中，每位公民都好像被自己的私人史塔西特工全天候监视着一样。[3]

在相对自由的国家，普通民众也受到越来越有效的监督。公司搜集和出售关于我们的购物记录、互联网和社交网络使用记录、电器使

用记录、电话和短信记录、就业和健康情况的信息。我们的位置可以通过手机和联网的汽车来追踪。摄像机能在街上认出我们的脸。此外，还有更多的数据可以通过智能信息集成系统整合在一起，形成一幅相当完整的画面，描绘出我们每个人在做什么，我们如何生活，我们喜欢谁、不喜欢谁，以及我们会如何投票。[4]相比之下，史塔西看起来就像业余选手。

控制你的行为

机器一旦具备了监视功能，下一步就是改变你的行为，以适应部署此技术的人。一种相当粗糙的方法是个性化的自动勒索：一个能够理解你在做什么的系统，不管是通过听、读还是看，它都可以很容易地发现你做了不应该做的事情。一旦发现了什么，它就会与你通信，尽可能多地勒索金钱（如果目标是政治控制或间谍活动，它就会胁迫你的行为）。对强化学习算法而言，获取金钱是完美的奖励信号，因此我们可以预期，人工智能系统识别不当行为并从中获利的能力将迅速提高。2015年初，我曾向一位计算机安全专家建议，由强化学习驱动的自动勒索系统可能很快就会成为现实。他笑着说，这正在成为现实。第一个被广泛宣传的勒索机器人是黛利拉（Delilah），它出现于2016年7月。[5]

改变人们行为的一个更不易察觉的方法是改变人们的信息环境，让人们相信不同的事情，做出不同的决定。当然，几个世纪以来，广告商一直在这样做，这是改变个人购买行为的一种方式。人们把宣传作为战争和政治统治的工具则有着更悠久的历史。

那么现在有何不同？第一，因为人工智能系统可以跟踪个人的在线阅读习惯、偏好和可能的知识状态，所以它们可以定制特定的信息，以此最大限度地影响个人，同时将信息被怀疑的风险降至最低。第二，人工智能系统知道个人是否阅读信息，花多长时间阅读信息，

以及他们是否会点开信息中的附加链接。然后，人工智能系统会使用这些信号作为它是否成功对个人产生影响的即时反馈。通过这种方式，它很快就能学会如何在工作中变得更加高效。这就是社交媒体上的内容推荐算法对政治观点产生潜在影响的方式。

最近的另一个变化是，人工智能、计算机图形学和语音合成技术的结合使生成深度模仿（deepfakes）成为可能，深度模仿几乎可以生成任何人说话做事的仿真视频和音频内容。这项技术只需要用户对所需模仿的事件进行简单的口头描述即可，世界上几乎所有人都能使用它。想做一段X参议员在非法场所Z接受可卡因贩子Y贿赂的手机视频？没问题！这种内容会让人对从未发生过的事情深信不疑。[6]此外，人工智能系统可以生成数百万个虚假身份，即所谓的机器人网络，它们每天可以发布数十亿条评论、推文和建议，从而把人类交流的真实信息淹没。易贝、淘宝和亚马逊等依赖信誉系统[7]在买卖双方之间建立信任的在线市场，经常与旨在破坏市场的机器人网络发生战争。

最后，如果政府能够实施基于行为的奖惩，那么控制方法就可以是直接的。这种系统将人民视为强化学习算法所使用的材料，训练他们优化国家设定的目标。对于一个政府，尤其是一个有着自上而下的工程思维模式的政府，其诱惑在于：如果每个人都表现良好，有爱国态度，并为国家进步做出贡献，那就更好了。技术使衡量个人行为、态度和贡献成为可能，因此，如果我们建立一个基于奖惩的科技监控系统，那么每个人都会变得更好。

这种思路存在几个问题。首先，它忽略了人民在侵入式监视和强制系统下生活的精神成本，外表的和谐掩盖了内心的痛苦，这很难说是一种理想状态。每一个善举都不再是善举，而变成了使个人得分最大化的行为，并且善举的接收者也将其视为如此。更糟糕的是，自愿行善的概念逐渐成为人们昔日行为的褪色记忆。在这样的制度下，去医院看望生病的朋友就像在红灯前停车一样没有道德意义和情感价值。其次，该方案与人工智能标准模型一样，也是失败模式的牺牲

品，因为它假定，所陈述的目标实际上是真实的、潜在的目标。不可避免的是，古德哈特定律（Goodhart’s law）将会取而代之，根据这一定律，个人会根据官方衡量标准优化自己的外部行为，就像大学会优化大学排名系统所使用“客观”衡量标准中的“质量”，而不是提高它们的真实（但无法衡量的）质量。[8]最后，强加的、统一的行为美德衡量标准忽略了一点，即一个成功的社会可能由各种各样的个人组成，每个人都以自己的方式做出贡献。

精神安全权

文明的伟大成就之一是逐步改善了人身安全。我们中的大多数人在日常生活中不用经常害怕受伤或死亡。1948年，《世界人权宣言》第3条规定：“人人享有生命、自由和人身安全。”

我想建议，每个人都应该享有精神安全的权利，即生活在一个基本真实的信息环境中的权利。人类倾向于相信自己的眼睛和耳朵。我们相信自己的家人、朋友、老师和（某些）媒体告诉我们的他们所相信的真相。虽然我们不指望二手车销售员和政客告诉我们真相，但我们很难相信他们有时竟会如此厚颜无耻地撒谎。因此，我们极易受到错误信息技术的影响。

精神安全权似乎没有被庄严地载入《世界人权宣言》。第18条和19条规定了“思想自由”和“主张和发表意见自由”的权利。当然，一个人的思想和观点部分是在他所处的信息环境中所形成的，而信息环境又受制于第19条“通过任何媒介和不论国界寻求、接受和传递消息和思想的自由”。也就是说，世界上任何地方的任何人都有权向你提供虚假信息。这就是困难所在：民主国家，尤其是美国，由于对政府控制言论的恐惧，它们在很大程度上不愿意或在宪法上无法阻止有人提供关于公众关注的事项的虚假信息。民主国家似乎不会认为“没有获得真实信息的途径就没有思想自由”，而是天真地相信真相最终

会赢得胜利，但这种信任使我们得不到保护。德国是个例外，它最近通过了《网络强制法》（Network Enforcement Act），要求内容平台删除被禁止的仇恨言论和虚假新闻，但这项法案遭到了相当多的批评，批评者称其“不可行，不民主”。[9]

因此，就目前而言，我们可以预计，我们的精神安全将会继续受到攻击，而保护我们精神安全的工作则主要由营利性机构和志愿者来承担。这些机构包括像factcheck.org和snopes.com这样的事实核查网站，但是当然，还有些“事实核查”网站也如雨后春笋般出现，它们颠倒黑白，把真相说成谎言，把谎言说成真相。

在欧洲和美国，谷歌和脸书等主要信息公司承受着巨大的压力，被要求“对此有所作为”。它们正在尝试如何标记虚假内容或将它们降级，同时使用人工智能和人力进行筛选，并将用户引导至经过验证的来源，以抵消错误信息的影响。最终，所有这些努力都依赖于循环信誉系统，从某种意义上讲，来源之所以可信，是因为可信的来源报告它们是可信的。如果传播的虚假信息到了一定数量，这些信誉系统可能会失效：真正值得信任的来源可能会变得不可信，反之亦然，就像今天美国有线电视新闻网和福克斯新闻等主要媒体来源出现的情况一样。阿维夫·奥瓦迪亚（Aviv Ovadya）是一位致力于对抗错误信息的技术专家，他将此称为“信息末日——思想市场的灾难性失败”。[10]

维护信誉系统功能的第一种方法是注入尽可能接近事实的来源。如果某些消息来源传播的信息与已知事实相反，那么这些消息来源就会变得不可信。在许多国家，公证人被作为基本事实的来源，用以维护法律和房地产信息的完整性。他们通常是一切交易中公正的第三方，并得到政府或专业协会的许可。（自1373年以来，伦敦市的“尊敬的公证人公司”一直在这么做，这表明讲真话的角色具有一定的稳定性。）如果事实核查人员有了正式的标准、专业资格和许可程序，那么这将有助于保持我们所依赖的信息流的有效性。W3C可靠网络组织

（W3C Credible Web group）和可靠联盟（Credibility Coalition）等组织旨在开发评估信息提供者的技术和众包方法，以帮助用户过滤掉不可靠的来源。

保护信誉系统的第二种方法是增加提供虚假信息的成本。因此，一些酒店评级网站只接受通过该网站预订并支付了某个酒店房费的人对该酒店进行评论，而其他评级网站则接受任何人的评论。毫不奇怪，前一种网站的评级偏见要小得多，因为它们为虚假评论增加了成本（为不必要的酒店房间付费）。[11]监管处罚更具争议性：没有人想要真相部门，德国的《网络强制法》只惩罚内容平台，而不惩罚发布假新闻的人。另外，正如许多国家和美国许多州将未经许可的录音电话定为非法，它们至少也应该对制作真实人物的虚拟录音或录像的行为施加处罚。

最后，还有另外两个对我们有利的事实。首先，几乎没有人希望自己被骗。（这并不是说父母总是会积极地询问那些赞扬孩子聪明有魅力的人是否诚实，只是他们不太可能从一个被认为一有机会就撒谎的人那里寻求这样的认可。）这意味着各种政治派别的人都有动力使用工具来辨别真伪。其次，没有人希望被当成骗子，尤其是新闻媒体。这意味着信息提供者——至少是那些认为信誉很重要的人，有动机加入行业协会，遵守有助于说实话的行为准则。反过来，社交媒体平台可以为用户提供一种选择：只查看那些遵守这些准则、接受第三方事实核查、信誉良好的来源的内容。

致命性自主武器

联合国将致命性自主武器系统（Lethal Autonomous Weapons Systems，LAWS，后简称AWS）定义为“在没有人的情况下，自主寻找目标、定位、杀击敌人”的武器系统。AWS有充分的理由被描述为继火药和核武器之后的“第三次战争革命”。

你可能在媒体上读过关于AWS的文章。通常，一些文章会将它们称为杀手机器人，还会用《终结者》电影里的图片作为配图。这至少在两个方面具有误导性：首先，它表明自主武器是一种威胁，因为它们可能会占领世界并毁灭人类；其次，它表明自主武器是类人的、有意识的和邪恶的。

媒体对这个问题的描述，最终效果是让它看起来像科幻小说。甚至连德国政府也被骗了，它最近发表了一份声明，[12]声称“拥有学习和发展自我意识的能力，是将个体功能或武器系统定义为自主的不可或缺的属性”。（这和断言“导弹不超过光速就不是导弹”是一样的道理。）事实上，自主武器具有与国际象棋程序相同的自主程度，国际象棋程序的任务是赢棋，但它自主决定走哪步棋以及吃掉敌人的哪些棋子。

AWS不是科幻小说中的东西，它们已经存在。或许最明显的例子就是以色列的“哈洛普”［Harop，图7（左）］，它是一种翼展长10英尺、弹头重50磅的巡飞弹，可以在给定的地理区域内进行长达6个小时的搜索，寻找一切符合给定标准的目标，然后将其摧毁。给定标准可以是“发出类似防空雷达的雷达信号”或“看起来像坦克”。



图7 （左）以色列航空航天工业公司生产的“哈洛普”；（右）“杀戮无人机”视频的静态图片，显示了一种可能的装有炸药的小型炮弹自主武器设计

通过结合微型四旋翼飞行器设计、微型摄像机、计算机视觉芯片、导航和测绘算法，以及探测和跟踪人类的方法的最新进展，人们在相当短的时间内就可以部署一种杀伤性武器，类似于图7（右）所示的“杀戮无人机”（Slaughterbots）[13]。这种武器可以用来攻击任何符合特定视觉标准（年龄、性别、制服、肤色等）的人，甚至攻击基于人脸识别的特定个体。有人告诉我，瑞士国防部已经制造并测试了一种真正的杀戮无人机，结果正如预期的那样，这项技术不仅可行而且致命。

自2014年以来，一系列外交讨论一直在日内瓦举行，以期达成一项禁止AWS的条约。与此同时，相关讨论的一些主要参与者（美国、中国、俄罗斯，在某种程度上还有以色列和英国）正在进行开发自主武器的危险竞赛。例如，在美国，CODE（拒止环境协同作战）项目旨在通过使无人机能够在只有间歇无线电联络的条件下工作，从而走向自

主。据项目经理称，无人机将“像狼一样成群狩猎”。[14] 2016年，美国空军展示了从三架F/A－18战斗机上部署的103架Perdix微型无人机。根据公告，“Perdix不是预先编程的同步个体，它们是一个集体有机体，进行决策时共享一个分布式大脑，并像自然界中的蜂群一样彼此适应”。[15]

你可能会认为，制造能自主决定杀死人类的机器显然是个糟糕的主意。但是“显然”并不总是能说服那些一心想获得他们认为的战略优势的政府。拒绝制造自主武器的一个更有说服力的理由是，它们是“可扩展的大规模杀伤性武器”。

“可扩展性”是计算机科学中的一个术语。如果你可以通过购买100万倍的硬件来完成100万倍以上的任务，那么这个进程就是可扩展的。因此，谷歌每天使用数百万台计算机，而不是雇用数百万名员工来处理大约50亿个搜索请求。有了自主武器，你可以通过购买100万倍的武器进行100万倍以上的杀戮。正是因为这些武器是自主的，与遥控无人机或AK－47步枪不同，它们无须个人监督即可完成工作。

作为大规模杀伤性武器，可扩展的自主武器与核武器和地毯式轰炸相比，对攻击者来说具有优势：它们保持财产完好无损，并且可以有选择地只消灭那些可能威胁攻击者的人。它们当然也可以被用来消灭整个种族或某一宗教的所有信徒（如果信徒有明显的标记）。此外，尽管核武器的使用代表着一个灾难性的门槛，而我们自1945年以来一直都没跨越这一门槛（这往往是因为运气好），可扩展的自主武器却没有这样的门槛。攻击可以无障碍地从伤亡100人升级到1 000人、1万人、10万人……除了实际的攻击外，仅仅是这些武器的攻击威力，就能使它们成为实施恐怖和压迫的有效工具。自主武器将大大降低个人、地方、国家和国际各个层面的人类安全。

这并不是说自主武器会像《终结者》电影中设想的那样带来世界末日。它们不需要特别聪明（一辆无人驾驶汽车可能需要更聪明），它们的任务也不会是那种“接管世界”的任务。人工智能带来的生存

性风险并不主要来自头脑简单的杀手机器人，而与人类起冲突的超级智能机器肯定会用这种方式武装自己，将相对愚蠢的杀手机器人变成全球控制系统的实体扩展。

抢走我们已有的工作

关于机器人从人类手中抢走工作这一主题，成千上万的媒体报道、文章和很多本书都写过了。研究中心如雨后春笋般在世界各地建立，以了解可能发生的事情。[16] 马丁·福特（Martin Ford）的《机器人时代：技术、工作与经济的未来》[17] 和卡鲁姆·蔡斯（Calum Chace）的《经济奇点：人工智能时代，我们将如何谋生？》[18] 这两本书很好地概括了人们的担忧。尽管我根本没有资格就本质上属于经济学的问题发表意见，[19] 但我认为，这个问题太重要了，不能完全只交给经济学家去解决，这一点很快就会显现出来。

约翰·梅纳德·凯恩斯（John Maynard Keynes）在其著名的文章《我们后代在经济上的可能前景》中提出了技术性失业问题。他在1930年写了这篇文章，当时大萧条在英国造成了大规模失业，但这个话题的历史要长得多。亚里士多德在他的《政治学》第一卷中非常清楚地提出了主要观点：

> 倘使每一个无生命工具都能按照人的意志或命令而自动进行工作……倘使每一个梭都能不假手于人力而自动地织布，每一琴拨都能自动地弹弦，匠师才用不到从属，奴隶主才可放弃奴隶。

所有人都同意亚里士多德的观点，即当雇主发现可以用一种机械方法来完成以前由人完成的工作时，就业就会立即减少。问题在于，随之而来的所谓补偿效应（这往往会增加就业）最终是否会弥补这种减少？乐观主义者的回答是肯定的：在当前的辩论中，他们指出了前

几次工业革命后出现的新工作。悲观主义者对此表示否定：在当前的争论中，他们认为机器也会做所有的“新工作”。当机器取代一个人的体力劳动时，这个人可以出卖脑力劳动。当机器取代一个人的脑力劳动时，这个人还能出卖什么？

在《生命3.0》中，迈克斯·泰格马克将这场争论描述为两匹马讨论1900年内燃机崛起的对话。一匹马预测，“关于马的新工作……这是以前经常发生的事情，就像发明轮子和犁的时候一样”。然而，对大多数马而言，“新工作”是当宠物。

这场争论已经持续了几千年，因为这两个方向都会产生影响。实际结果取决于哪种影响更重要。例如，你来想一想，随着技术的进步，房屋油漆工的工作会有什么变化。为了简单易懂，我用漆刷的宽度代表自动化的程度。

- 如果刷子只有一根头发的直径那么宽（0.1毫米），粉刷一栋房子需要数千人年的时间，而且人们基本上不会雇用房屋油漆工。
- 当刷子宽度达到1毫米的时候，几位画家用刷子在皇宫里画了几幅精美的壁画。当刷子宽度达到1厘米的时候，贵族们开始效仿。
- 当刷子宽度达到10厘米的时候，我们达到了实用的境界：大多数房主都把房屋内外粉刷了一遍，尽管可能不太频繁，此时成千上万的房屋油漆工找到了工作。
- 一旦我们有了宽滚轮和喷枪（相当于宽度1米左右的刷子），粉刷房屋的价格就会大幅下降，但需求可能会开始饱和，因此房屋油漆工的数量会有所下降。
- 当一个人管理着由100个房屋喷漆机器人组成的团队时，这相当于拥有使用100米宽的刷子的工作效率，那么整栋房屋可以在一小时内完成粉刷，此时很少有房屋油漆工可以找到工作。

因此，技术的直接影响是双向的：首先，通过提高生产率，技术可以降低一项服务的价格，从而增加需求以及增加就业；其次，技术的进一步发展意味着需要的人力越来越少。图8展示了这些发展。[20]

许多技术都表现出类似的曲线。如果在某个特定的经济部门，我们处于峰值的左侧，那么改进技术将会增加该部门的就业，当前的例子可能包括清除涂鸦、清理环境、检查航运集装箱和在欠发达国家建设住房等任务，如果我们有机器人来帮忙，那么所有这一切可能在经济上变得更加可行。如果我们已经到了峰值的右侧，那么进一步的自动化就会减少就业。例如，不难预测电梯操作员将继续被挤出市场。从长远来看，我们不得不预计大多数行业将被推到曲线的最右边。最近一篇基于经济学家戴维·奥特（David Autor）和安娜·萨洛蒙斯（Anna Salomons）的一项细致的经济计量研究的文章认为：“在过去40年里，每一个引入技术来提高生产率的行业的工作岗位都在减少。”[21]

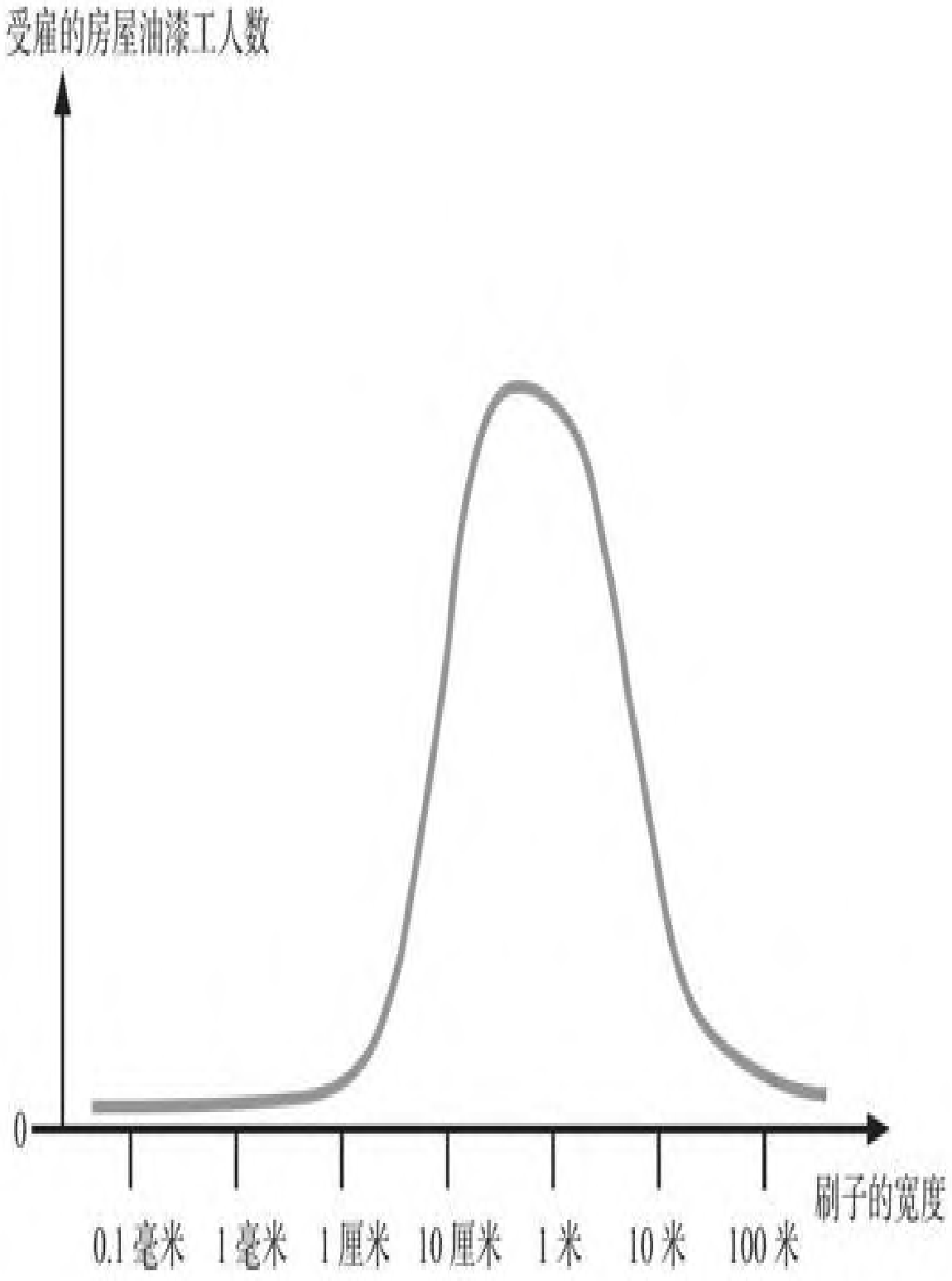


图8 随着粉刷技术的改进，房屋油漆工雇用情况的变化

那么经济乐观主义者所描述的补偿效应又如何？

- 有些人不得不制造喷漆机器人。造多少？必须远远少于机器人替代的房屋油漆工的数量，否则，用机器人粉刷房屋将花费更多，而不是更少的钱，那样就没有人会购买这种机器人。
- 房屋粉刷变得更便宜了，所以人们会更频繁地请来房屋油漆工。
- 最后，因为我们花在房屋粉刷上的钱更少，所以我们有更多的钱花在其他事情上，从而增加了其他行业的就业。

经济学家曾试图在自动化程度有所提高的各个行业中衡量这些影响的大小，但结果通常是不确定的。

从历史上看，大多数主流经济学家都从“宏观”的角度提出过这样的观点：自动化提高了生产率，从整体上说，人类的境况更好了，因为我们用同样的工作量享受到了更多的商品和服务。

不幸的是，经济学理论并没有预测自动化是否会让每个人的生活变得更好。一般来说，自动化增加了资本持有者（即房屋喷漆机器人的所有者）的收入份额，减少了劳动力（即曾经的房屋油漆工）的收入份额。经济学家埃里克·布莱恩约弗森（Erik Brynjolfsson）和安德鲁·麦卡菲（Andrew McAfee）在《第二次机器革命》中指出，这种情况已经发生了几十年。美国的数据如图9所示。他们指出，1947年至1973年间，工资和生产率同时增长，但1973年后，工资停滞不前，生产率却几乎翻了一番。布莱恩约弗森和麦卡菲称之为“大脱钩”。其他主要经济学家也发出了警告，包括诺贝尔奖得主罗伯特·希勒（Robert Shiller）、迈克尔·斯宾塞（Michael Spence）和保罗·克鲁格曼（Paul Krugman），世界经济论坛主席克劳斯·施瓦布（Klaus Schwab），以及世界银行前首席经济学家、比尔·克林顿总统时期的财政部长拉里·萨默斯（Larry Summers）。

那些反对技术性失业概念的人经常提到银行出纳员和零售收银员，银行出纳员的部分工作可以通过自动取款机完成，零售收银员的工作则通过商品上的条形码和RFID（射频识别）标签来加速。人们常说，这些职位数量的增长是因为技术。的确，从1970年到2010年，美国出纳员的数量大约翻了一番，但这里应该指出，在同一时期，美国人口增长了50%，金融业增长了400%以上，[22]因此我们很难将就业增长的全部或部分原因归结为自动取款机。不幸的是，从2010年到2016年，大约有10万名出纳员失业。美国劳工统计局预测，到2026年，还会有4万人失业：“预计网上银行和自动化技术将继续取代出纳员曾经履行的更多工作职责。”[23]零售收银员的数据也不再令人鼓舞：从1997年到2015年，人均数字下降了5%。美国劳工统计局表示：“技术的进步，如零售店的自助收银台和网上销售的增加，将继续限制对收银员的需求。”这两个行业似乎都在走下坡路。几乎所有与机器打交道的低技能职业都是如此。

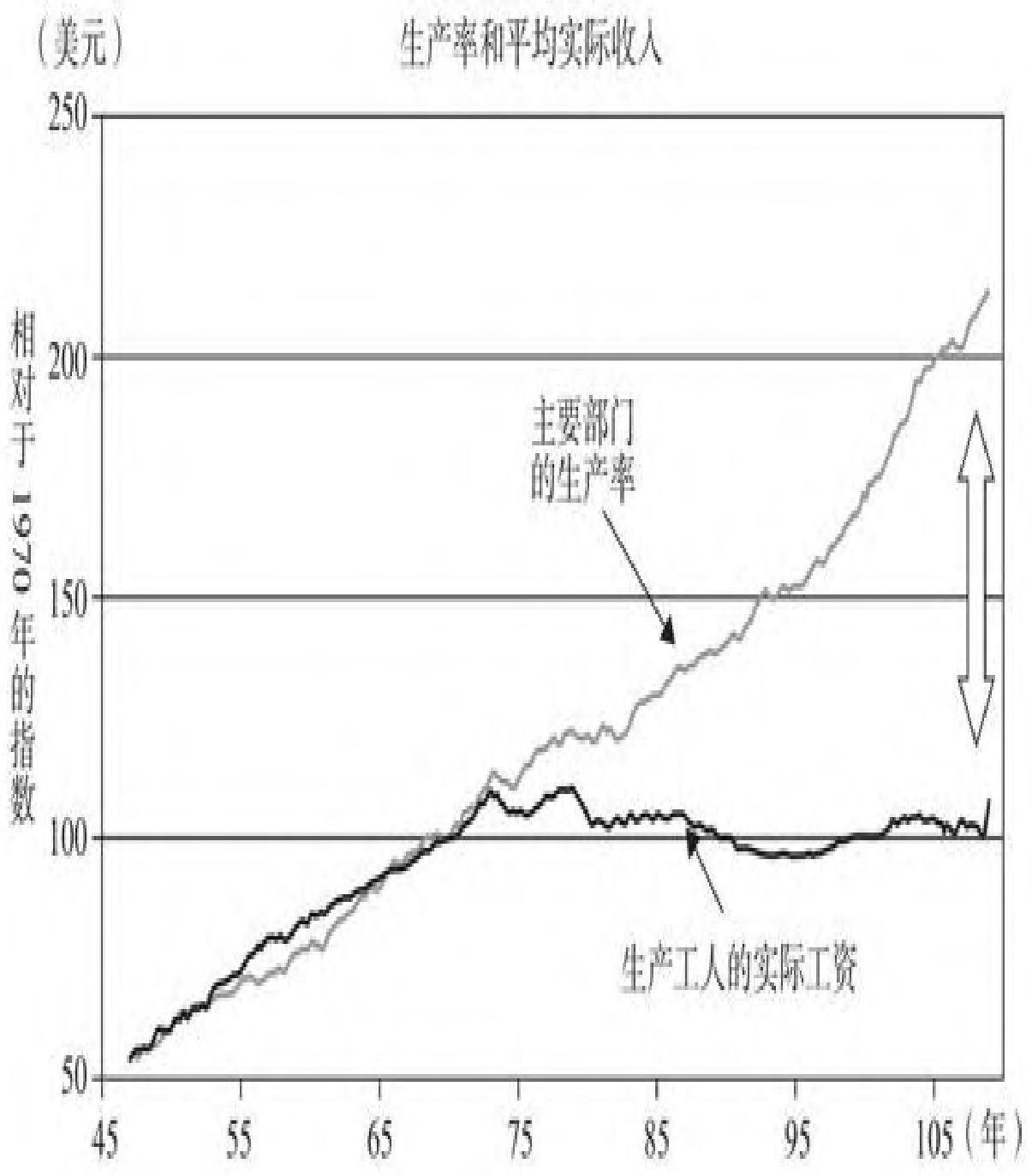


图9 1947年以来美国的经济产值和实际工资中位数（数据来自美国劳工统计局）

随着基于人工智能的新技术的到来，哪些职业将会减少？媒体引用的主要例子是司机。美国大约有350万名卡车司机，他们中有许多人很容易受到自动化的影响。亚马逊等公司已经在州际高速公路上使用

无人驾驶卡车运货，尽管目前仍然配备了起辅助作用的人类司机。[24]每一次卡车运输的长途部分很有可能很快变成自主的，而人类将暂时处理城市中的运输、接车和送货。由于这些预期的发展，很少有年轻人对卡车运输这一职业感兴趣。具有讽刺意味的是，目前美国卡车司机严重短缺，而这只会加速自动化的发展。

白领工作也面临威胁。例如，美国劳工统计局预测，从2016年到2026年，保险核保人的人均就业率将下降13%：“自动核保软件使员工能够比以前更快地处理申请，从而减少对核保人的需求。”如果语言技术像预期那样发展，许多销售和客户服务工作，以及法律领域的工作也会面临威胁。（在2018年的一次竞赛中，人工智能软件在分析标准保密协议方面的得分超过了经验丰富的法学教授，并且前者完成任务的速度比后者快了200倍。[25]）常规形式的计算机编程，即今天经常被外包的那种编程工作，也可以实现自动化。事实上，几乎任何可以外包的东西都是自动化的绝好选择，因为外包涉及将作业分解为任务，这些任务可以打包并以去情境化的形式分发。机器人流程自动化产业生产的软件工具完全可以在在线执行的办公任务方面实现这种效果。

随着人工智能的发展，在未来几十年里，所有常规的体力劳动和脑力劳动基本上都将由机器以更低的成本完成，这当然是可能的，甚至实现的可能性非常大。自从几千年前我们脱离狩猎采集者的角色以来，我们的社会就把大多数人当作机器人来使用，他们执行重复性的体力和脑力劳动，所以这些角色很快被机器人取代也许并不奇怪。当这种情况发生时，那些在竞争中被淘汰的、无法从事高技能工作的人的工资就会被推到贫困线以下。拉里·萨默斯说：“很有可能，考虑到（资本取代劳动力）这种可能性，某些类别的劳动力将无法赚取维持生计的收入。”[26]这正是发生在马身上的事情：机械运输变得比养马更便宜，所以马成了宠物。面对等同于马变成宠物的社会经济问题，人类将会对他们的政府感到非常不满。

面对可能心生不满的人，世界各国政府都开始关注这个问题。大多数政府已经发现，把每个人都重新培训成数据科学家或机器人工程师的想法是不现实的，全世界可能需要500万到1 000万个这类工作岗位，但和面临威胁的大约10亿个工作岗位相比，该数字微不足道。数据科学对一艘巨大的游轮而言不过只是一艘小小的救生艇。[27]

一些人正在制订“过渡计划”，但是过渡到何处？为了规划过渡，我们需要一个合理的目标，也就是说，我们需要一个理想的未来经济的合理图景，在那里，我们目前所说的大部分工作都是由机器完成的。

一幅迅速浮现的经济图景是，因为工作是不必要的，所以工作的人要少得多。凯恩斯在他的文章《我们后代在经济上的可能前景》中就设想了这种未来。他将1930年困扰英国的高失业率描述为“经济失调的暂时阶段”，其原因是“技术效率的提高”比“我们为劳动力开辟新用途的速度”更快，然而他没有想到，从长远来看，在一个世纪的技术再进步之后，社会会恢复充分就业：

> 因此，人类自出现以来，第一次遇到了他真正的、永恒的问题——在从紧迫的经济束缚中解放出来以后，应该怎样来利用他的自由？科学和复利的力量将为他赢得闲暇，而他又该如何来消磨这段光阴，生活得更明智而惬意呢？

这样的未来要求对我们的经济制度进行彻底的改革，因为在许多国家，不工作的人会面临贫穷甚至落入赤贫。因此，凯恩斯愿景的现代支持者通常支持某种形式的普遍基本收入（UBI）。普遍基本收入的资金来源是增值税或资本收入税，无论情况如何，普遍基本收入都会为每个成年人提供合理的收入。那些渴望更高生活水平的人仍然可以在不失去普遍基本收入的情况下工作，而那些不想工作的人可以随心所欲地消磨他们的时间。或许令人惊讶的是，普遍基本收入得到了各

个政治派别的支持，其支持者甚至包括亚当·斯密研究所[28]和绿党[29]等。

对某些人而言，普遍基本收入代表了某种形式的天堂。[30]而对另一些人而言，这代表着承认失败，说明大多数人对社会没有任何经济价值。他们的吃饭和住房问题主要由机器解决，但除此之外，他们只能自食其力。真相总是介于二者之间，这在很大程度上取决于人们如何看待人类心理。凯恩斯在他的文章中明确区分了“奋斗的人”和“享乐的人”：前者是那些有目标的人，对他们来说，“果酱并不是果酱，即绝不是今天这罐实实在在的果酱，而是可望而不可即的明天的那罐果酱”；后者是那些心情愉快的人，他们“能够从事物中获得直接的乐趣”。普遍基本收入的提议假定大多数人都是心情愉快的那类人。

凯恩斯认为，奋斗是普通人的习惯和本能，是经过无数代人培养出来的，而不是“生活的真正价值”之一。他预言这种本能会逐渐消失。有些人可能会反对这种观点，他们认为，奋斗是人类的本能。与其说奋斗和享乐是相互排斥的，不如说它们往往是不可分割的：真正的享乐和持久的满足来自拥有一个目标并实现它（至少是努力尝试），这通常需要面对困难，而不是被动地消费眼前的快乐。攀登珠穆朗玛峰和乘坐直升机降落在山顶所获得的快乐是有区别的。

奋斗和享乐之间的联系是我们理解如何创造美好未来的中心主题。也许未来的几代人会想，我们为什么要担忧“工作”这种毫无意义的事情。万一态度迟迟没有转变，那就让我们考虑一下下面这种观点：虽然绝大多数的商品和服务将由机器生产，几乎无须人类的监督，但是大多数能做一些有用的事情的人会过得更好。这种观点会产生什么经济影响？不可避免的是，大多数人将从事提供人际关系服务的工作，这些服务只能由人类提供，或者说我们更愿意由人类提供这些服务。换言之，如果我们不能再提供常规的体力劳动和脑力劳动，我们仍然可以提供我们的人性。我们需要变得善于做人。[31]

目前，这类职业包括心理治疗师、高管教练、导师、顾问、共事者以及照顾儿童和老人的人。“护理行业”这个词经常被用在这种情况下，但它是有误导性的：它对于提供护理的人有积极的含义，但对于接受护理的人有依赖和无助的负面含义。但是请思考下面这个同样来自凯恩斯的观点：

> 只有这些人才能在这种丰裕中获得享受：他们不会为了生活而出卖自己，能够使生活的艺术永葆青春，并将之发扬光大，提升到更高境界。

我们所有人在学习“生活的艺术”时都需要帮助。这不是依赖的问题，而是增长的问题。无论是在艺术、音乐、文学、会话、园艺、建筑、烹饪、葡萄酒还是电子游戏中，激励他人、传递鉴赏力和创造力的能力可能比以往任何时候都更重要。

下一个问题是收入如何分配。在大多数国家，收入分配问题几十年来一直在朝着错误的方向发展。这是一个复杂的问题，但有一点是明确的：高收入和高社会地位通常源于提供高附加值。举个例子，婴幼儿保育行业与低收入和低社会地位相关联。这在某种程度上是我们不知道如何去做的结果。有些从业者天生擅长于此，但许多人不擅长。相比之下，比如说整形外科手术。我们不会只雇用那些需要一点闲钱的无聊的青少年去做整形外科医生，然后付给他们5美元的时薪，外加冰箱里的所有东西。我们花了几个世纪的时间来研究人体，以及如何修复受损的人体，从业者必须接受多年的训练来学习所有这些知识以及应用这些知识所必需的技能。因此，整形外科医生收入高，备受尊敬。他们得到高薪，不仅因为他们知道很多，接受了很多培训，还因为所有这些知识和培训实际上都是有效的。这使他们能够给其他人的生活增加很多价值，尤其是那些身体有缺陷的人。

不幸的是，我们对心智科学的理解非常有限，我们对幸福和满足的科学理解甚至更有限。我们根本不知道如何以一致的、可预见的方式为彼此的生活增加价值。我们已经在某些精神疾病方面取得了一定的成功，但我们仍然在为一些基本的东西，如教孩子阅读，而进行着一场持续百年的扫盲战争。[32]我们需要对我们的教育体系和科学事业进行彻底的反思，把更多的注意力放在人类自身而不是物质世界上。[美国东北大学校长约瑟夫·奥恩（Joseph Aoun）认为，大学应该教授和研究“人类工程学”（Humanics）。[33]] 将幸福看作一门工程学科，听起来有些奇怪，但这似乎是必然的结论。这门学科将建立在基础科学的基础上，会让我们更好地理解人类思维在认知和情感层面上的工作方式，并培养各领域的从业者，从帮助个人规划生活总体轨迹的生命建筑师，到增强好奇心和个人适应力等方面的专家。如果以真正的科学为基础，那么这些职业会像今天的桥梁设计师和整形外科医生一样备受推崇。

改造我们的教育和研究机构来创造这种基础科学，并将其转化为培训项目和认证专业将需要几十年的时间，但这是一个很好的开始，可惜我们才刚刚起步。该方案如果成功，最终将会带来一个非常值得生活的世界。如果不进行这样的反思，我们就有可能陷入难以为继的社会经济混乱。

篡夺其他人类角色

在允许机器接管涉及人际服务的角色之前，我们应该三思。如果对其他人类来说，拥有人性是我们的主要“卖点”，那么制造人造人类似乎是一个坏主意。幸运的是，在了解他人的感受和反应方面，我们和机器相比有明显的优势。几乎每个人都知道被锤子砸到拇指或单相思的滋味。

对抗这种人类的自然优势反倒是我们的自然劣势：容易被外表所愚弄，尤其是人类的外表。艾伦·图灵警告人们不要把机器人做得与人类相似：[34]

> 我当然希望并相信，我们不应该费太大的力气去制造具有最明显的人类特征而非智力特征的机器，比如人形机器人。在我看来，进行这样的尝试是徒劳的，其结果就像人造假花一样，不伦不类。

不幸的是，图灵的警告没有得到重视。如图10所示，几个研究小组已经制造出了非常逼真的人形机器人。

作为研究工具，机器人可以让人们了解人类是如何解释机器人的行为和交流的。但作为未来商业产品的原型，机器人代表了一种不诚实的形式。它们绕过我们的意识觉知，直接吸引我们的情感自我，即让我们相信它们也许被赋予了真正的智能。例如，想象一下，我们关闭并回收一个有故障的浅灰色盒子（即便它大声抱怨说不想被关闭），比对栩栩如生的人形机器人佳佳或Geminoid DK做同样的事情要容易得多。再想象一下，我们把婴幼儿托付给酷似人类（比如他们的父母）的实体来照顾将会是多么令人困惑，甚至可能令人在心理上感到不安。它们看起来就像孩子的父母一样关心他们，但实际上并不关心。



图10 （左）中国科学技术大学制造的机器人“佳佳”；（右）由日本大阪大学石黑浩教授（Hiroshi Ishiguro）设计，模仿丹麦奥尔堡大学亨利克·夏弗教授（HenrikScharfe）相貌而制作的机器人“Geminoid DK”

除了通过面部表情和动作来传达非语言信息的基本能力（即使兔八哥也能轻松做到这一点），我们没有充分的理由要让机器人具有人形。不去制造人形机器人也有很好的实际原因，例如，与四足运动相比，我们的两足姿态相对不稳定。狗、猫和马很好地融入了我们的生活，它们的身体形态很好地说明了它们的行为方式。（想象一下，如果一匹马突然表现得像狗一样会如何。）机器人也应该如此。也许四

条腿，两只手，人首马身的形态会是一个很好的标准。一台精确的人形机器人的意义就像一辆时速最高可达5英里的法拉利，或是用剁碎的肝脏冒充奶油制成的树莓冰激凌蛋筒。

一些机器人的类人特征已经造成了政治上和情感上的混淆。2017年10月25日，沙特阿拉伯授予索菲娅公民身份，索菲娅是一个人形机器人，据说它不过是个“有人类面孔的聊天机器人”，[35]甚至更糟糕。[36]或许这只是一个公关噱头，但欧洲议会法律事务委员会的提案完全是认真的：[37]

> 从长远来看，为机器人创造一个具体的法律地位，至少可以让最先进的自主机器人获得电子人的地位，为它们可能造成的任何损害负责。

换言之，机器人自身将承担其造成的损害的法律责任，它的所有者或制造商则不用负责。这意味着机器人将拥有金融资产，它们如果不遵守规则，就会受到制裁。从字面上看，这是没有意义的。例如，如果我们因机器人不付款而监禁它，它有什么好在意的呢？

除了毫无必要甚至荒谬地提高机器人的地位之外，在影响人类的决策中越来越多地使用机器还可能降低人类的地位和尊严。这种可能性在科幻电影《极乐空间》的一个场景中得到了完美的体现，机器人麦克斯（马特·达蒙饰）在它的“假释官”面前为自己辩护（图11），解释为什么延长它的刑期是不合理的。不用说，麦克斯是不成功的，假释官甚至斥责它没有表现出适当的恭敬态度。

人们可以从两方面来看待这种对人类尊严的侵犯。第一个方面是显而易见的：通过赋予机器凌驾于人类之上的权力，我们把自己降到了二等地位，并失去了参与影响我们的决策的权利。（一种更极端的形式是赋予机器杀人的权力，正如本章前文所讨论的那样。）第二个方面是间接的：即使你认为做决定的不是机器，而是设计和委托制造

机器的人，这些人类设计师和专员却认为在这种情况下权衡每个人类主体的个人情况并不值得，这表明他们认为他人的生命价值不大。这或许是由人类服务的精英阶层与由机器服务和控制的广大社会底层之间开始严重分化的一个征兆。



图11 麦克斯在“极乐空间”遇到了它的假释官

2018年《欧盟数据保护通用条例》（后简称GDPR）第22条明确禁止在以下情况下授予机器权限：

> 数据主体有权反对此类决策：完全依靠自动化处理，包括用画像，对数据主体做出具有法律影响或类似严重影响的决策。

以上条款虽然在原则上听起来令人钦佩，但在实践中会产生多大的影响还有待观察，至少在我撰写本书时是这样的。把决策权交给机器往往更容易、更快速、更便宜。

对自主决策的所有担忧都源自潜在的算法偏见——机器学习算法倾向于在贷款、住房、工作、保险、假释、判决、大学录取等方面产生不恰当的有偏见的决策。数十年来，在许多国家，人们在做这些决策时使用种族等标签是违法的，而且GDPR第9条在非常广泛的应用中都禁止使用这些标签。当然，这并不意味着通过将种族排除在数据之外，我们就一定能做出不带有种族偏见的决定。例如，从20世纪30年代开始，政府批准的“红线歧视”做法使美国某些邮政编码所示的地区禁止被用来抵押贷款和做其他形式的投资，进而导致房地产价值下降。碰巧的是，这些邮政编码所示的地区主要是非洲裔美国人居住的地区。

为了防止出现“红线歧视”这类现象，现在美国只能使用5位数邮政编码中的前3位数进行信贷决策。此外，决策过程必须服从检查，以确保不会出现其他“偶然”的偏见。人们常说，GDPR为一切自主决定提供了通用的“解释权”，[38]但第14条的实际措辞仅仅要求提供：

> 所涉及的有逻辑、有意义的信息，以及这种处理对于数据主体的意义和预期的后果。

目前尚不清楚法院将如何执行这一条款。可能的情况是，不幸的消费者只会得到一份关于特定深度学习算法的描述，用于训练做决策的分类器。

如今，算法偏见产生的原因可能在于数据，而不在于公司故意作恶。2015年，《魅力》杂志报道了一个令人失望的发现：“在谷歌上搜索‘CEO’，图片搜索结果中第一幅女性图片出现在第12行，而且是个芭比娃娃。”（2018年在谷歌上搜索的结果中有一些真实的女性，但在普通的素材图片中大多数都是扮演CEO的模特，而不是真正的女性CEO。2019年的结果略好一些。）这不是因为谷歌图片搜索排名有意性别歧视，而是因为产生数据的文化中预先存在偏见，即男性CEO的数量远远多于女性CEO。当人们想要在标题图片中描绘一位“典型的”CEO时，他们几乎总是会选择一位男性人物。当然，“偏见主要存在于数据中”这一事实并不意味着我们没有义务采取措施来解决这个问题。

机器学习方法的幼稚应用会产生有偏见的结果，这里面还有其他更偏向技术性的原因。例如，根据定义，少数群体在全体人口数据样本中的代表性较差。因此，如果对少数群体中个别成员的预测主要基于同一群体其他成员的数据，那么这种预测可能不太准确。幸运的是，人们对消除机器学习算法中的无意偏见的问题给予了很大的关注，现在有一些方法可以根据几个看似合理和理想的公平定义，产生无偏见的结果。[39]对这些公平定义的数学分析表明，它们不能同时实现，并且在强制执行时，它们会导致预测准确性降低，在贷款决策方面，还会损害贷款人的利益。这也许令人失望，但至少清楚地表明了避免算法偏见所涉及的权衡。人们希望，对这些方法和问题本身的认识可以在决策者、实践者和用户中迅速传播。

如果将个人的权力移交给机器有时会产生问题，那么将许多人的权力移交给机器又如何？也就是说，我们是否应该让机器扮演政治家和管理者的角色？目前，这似乎有些牵强。机器无法维持长时间的对话，也缺乏对广泛决策相关因素的基本理解，例如，是应该提高最低

工资还是应该拒绝另一家公司的合并提议？然而，趋势是明显的：在许多领域，机器在越来越高的权力级别上做出决策。以航空公司为例。首先，计算机帮助编制航班时刻表。很快，它们接管了机组人员排班、预订座位和日常维护的工作。接下来，它们被连接到全球信息网络，向航空公司经理提供实时状态报告，以便经理能够有效地应对异常情况。现在，它们正在接管管理异常情况的工作：更改飞机航线，重新安排工作人员，重新为乘客订票，以及修改维护计划。

从航空公司经济情况和乘客体验的角度来看，这一切都是好事。问题是，计算机系统仍然是人类的工具吗？还是说人类变成了计算机系统的工具，即人类在必要时提供信息并修复漏洞，但不再深入了解整个系统是如何工作的？如果系统崩溃，全球混乱接踵而至，直到计算机系统重新上线后才恢复正常，那时答案就变得清晰了。例如，2018年4月3日出现的一次“计算机故障”导致欧洲1.5万个航班严重延误或取消。[40] 2010年，交易算法导致纽约证券交易所“闪电崩盘”，几分钟内1万亿美元蒸发，此时唯一的解决办法是关闭交易所，而人们连发生了什么事还都没弄清楚。

在一切技术都没出现之前，人类和大多数动物一样，只能勉强糊口。可以说，我们的双脚直接站在地面上。技术逐渐让我们登上了机器的金字塔，扩大了我们作为个体和物种所能涉及的范围。我们可以用不同的方式来设计人与机器之间的关系。如果我们设计机器是为了让人类保留足够的理解、权威和自主，那么系统的技术部分就可以极大地放大人类的能力，使我们每个人都能站在一个巨大的能力金字塔上，如果愿意，我们可以说人类成了“半人半神”。但是让我们看看在线购物仓库中的员工。她比她的前辈效率更高，因为她有一支机器人小分队把她的储物箱送到面前，让她从中挑选物品。她是由智能算法控制的更大的系统的一部分，智能算法决定她应该站在哪里，应该挑选和分发哪些物品。她已经有半个身子埋在金字塔里了，而不是站

在金字塔的顶端。随着沙子逐渐填满金字塔的空间，她的角色消失只是时间问题。

---

[1] 维基百科上关于史塔西的文章有几个有用的参考资料，涉及史塔西的人力及其对民主德国生活的整体影响。

[2] 关于史塔西档案的详情，请参见：Cullen Murphy, *God's Jury: The Inquisition and the Making of the Modern World* (Houghton Mifflin Harcourt, 2012)。

[3] 关于人工智能监控系统的全面分析，请参见：Jay Stanley, *The Dawn of Robot Surveillance* (American Civil Liberties Union, 2019)。

[4] 最近关于监视和控制的书籍包括：Shoshana Zuboff, *The Age of Surveillance Capitalism* : The Fight for a Human Future at the New Frontier of Power (Public Affairs, 2019) and Roger McNamee, Zucked: Waking Up to the Facebook Catastrophe (Penguin Press, 2019)。

[5] 关于勒索机器人的新闻报道：Avivah Litan, "Meet Delilah-the first insider threat Trojan," Gartner Blog Network, July 14, 2016。

[6] 关于人类易受错误信息影响的低技术版本的一个例子，有一个毫无戒心的人确信世界正在被流星撞击摧毁，请参见：*Derren Brown: Apocalypse*, "Part One," directed by Simon Dinsell, 2012, youtube.com/watch?v=o_CUrMJOxqs。

[7] 史蒂文·塔德利斯（Steven Tadelis）对声誉系统及其腐败进行了经济分析："Reputation and feedback systems in online platform markets," *Annual Review of Economics* 8 (2016): 321–40。

[8] 古德哈特定律："任何观察到的统计规律，一旦出于控制目的而被施加了压力，就会趋于崩溃。"例如，教师素质和教师薪酬之间曾经可能存在相关性，《美国新闻与世界报道》（*US News & World Report*）的大学排名是通过教师薪酬来衡量教师素质的。这导致了一场薪酬"军备竞赛"，受益者是教师，而不是支付薪酬的学生。军备竞赛以不依赖教师素质的方式改变教师工资，因此这种相关性趋于消失。

[9] 一篇描述德国监管公共话语的文章：Bernhard Rohleder, "Germany set out to delete hate speech online. Instead, it made things worse," *World Post*, February 20,2018。

[10] 关于"信息末日"（infopocalypse）：Aviv Ovadya, "What's worse than fake news?The distortion of reality itself," *World Post*, February 22, 2018。

[11] 关于在线酒店评价方面的腐败问题：Dina Mayzlin, Yaniv Dover, and Judith Chevalier, “Promotional reviews: An empirical investigation of online review manipulation,” *American Economic Review* 104 (2014): 2421–55。

[12] 德国在《特定常规武器公约》政府专家组会议上的声明，2018年4月10日，日内瓦。

[13] 这部由生命未来研究所（The Future of Life Institute，FLI）资助的电影《杀戮无人机》于2017年11月上映，请见youtube.com/watch?v=9CO6M2HsoIA。

[14] 关于军事公共关系中最大的错误之一的报告，请参见：Dan Lamothe, “Pentagon agency wants drones to hunt in packs, like wolves,” *The Washington Post*, January 23, 2015。

[15] 大规模无人机群实验公告：US Department of Defense, “Department of Defense announces successful micro-drone demonstration,” news release no. NR-008-17,January 9, 2017。

[16] 研究技术对就业的影响的研究中心的例子有加州大学伯克利分校的“工作、智能工具和系统”小组，斯坦福大学行为科学高级研究中心的“未来工作和工人”项目，以及卡内基–梅隆大学的“未来工作倡议”。

[17] 对未来技术失业的悲观看法：Martin Ford, *Rise of the Robots: Technology and the Threat of a Jobless Future* (Basic Books, 2015)。

[18] Calum Chace, *The Economic Singularity: Artificial Intelligence and the Death of Capitalism* (Three Cs, 2016).

[19] 相关优秀论文集，请参见：Ajay Agrawal, Joshua Gans, and Avi Goldfarb, eds.,*The Economics of Artificial Intelligence: An Agenda* (National Bureau of Economic Research, 2019)。

[20] 这种“倒U形”就业曲线背后的数学分析请参见：James Bessen, “Artificial intelligence and jobs: The role of demand” in *The Economics of Artificial Intelligence*, ed. Agrawal, Gans, and Goldfarb。

[21] 关于自动化引起的经济混乱的讨论，请参见：Eduardo Porter, Tech is splitting the US work force in two,” *The New York Times*, February 4, 2019。这篇文章引用了以下报告来证明这一结论：David Autor and Anna Salomons, “Is automation labor-displacing? Productivity growth, employment, and the labor share,”Brookings Papers on Economic Activity (2018)。

[22] 关于20世纪银行业增长的数据，请参见：Thomas Philippon, “The evolution of the US financial industry from 1860 to 2007: Theory and evidence,” working paper, 2008。

[23] 就业数据和职业兴衰的“圣经”：US Bureau of Labor Statistics, *Occupational Outlook Handbook: 2018–2019 Edition* (Bernan Press, 2018)。

[24] 关于卡车运输自动化的报告：Lora Kolodny, “Amazon is hauling cargo in selfdriving trucks developed by Embark,” CNBC, January 30, 2019。

[25] 法律分析自动化的进展，描述了一场竞赛的结果：Jason Tashea, “AI software is more accurate, faster than attorneys when assessing NDAs,” *ABA Journal*,February 26, 2018。

[26] 一位著名经济学家的评论，标题明确地让人想起凯恩斯1930年的文章：Lawrence Summers, “Economic possibilities for our children,” *NBER Reporter*(2013)。

[27] 将数据科学就业比作巨型游轮上的小救生艇，这个类比来自与新加坡公共服务部负责人的一次讨论。该负责人承认，从全球范围来看，这是正确的，但她指出，“新加坡很小，可以坐上救生艇。”

[28] 从保守的角度支持普遍基本收入（UBI）：Sam Bowman, “The ideal welfare system is a basic income,” Adam Smith Institute, November 25, 2013。

[29] 从进步的角度支持普遍基本收入：Jonathan Bartley, “The Greens endorse a universal basic income. Others need to follow,” *The Guardian*, June 2, 2017。

[30] 蔡斯在《经济奇点：人工智能时代，我们将如何谋生？》一书中称普遍基本收入的“天堂”版本为《星际迷航》经济，他指出，在最近几部《星际迷航》系列中，金钱已经被废除，因为技术创造了本质上无限的物质产品和能源。他还指出，要使这样一个系统取得成功，需要在经济和社会组织方面进行巨大的变革。

[31] 经济学家理查德·鲍德温（Richard Baldwin）在他的书*The Globotics Upheaval: Globalization, Robotics, and the Future of Work* (Oxford University Press,2019)中也预测了个人服务的未来。

[32] 这本书被认为揭露了“全字”识字教育的失败，并引发了两大阅读学派之间长达数十年的斗争：Rudolf Flesch, *Why Johnny Can’t Read: And What You Can Do about It* (Harper & Bros., 1955)。

[33] 关于使接受者能够适应未来几十年快速的技术和经济变化的教育方法：Joseph Aoun, *Robot-Proof: Higher Education in the Age of Artificial Intelligence*(MIT Press, 2017)。

[34] 图灵预测人们将被机器取代的广播演讲：Alan Turing, “Can digital machines think?” May 15, 1951, radio broadcast, BBC Third Programme。Typescript available at turingarchive.org.

[35] 描述索菲娅“入籍”沙特阿拉伯的新闻文章：Dave Gershgorn, “Inside the mechanical brain of the world’s first robot citizen,” *Quartz*, November 12,

2017。

[36] 杨立昆对索菲娅的看法：Shona Ghosh, “Facebook’s AI boss described Sophia the robot as ‘complete b—t’ and ‘Wizard-of-Oz AI,’ ” *Business Insider*, January 6, 2018。

[37] 欧盟关于机器人法律权利的提案：Committee on Legal Affairs of the European Parliament, “Report with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)),” 2017。

[38] 事实上，GDPR关于“解释权”的条款并不新鲜：it is very similar to Article 15(1) of the 1995 Data Protection Directive, which it supersedes。

[39] 以下是三篇最近的论文，提供了对公平的深刻的数学分析：Moritz Hardt,Eric Price, and Nati Srebro, “Equality of opportunity in supervised learning,” in *Advances in Neural Information Processing Systems 29*, ed. Daniel Lee et al.(2016); Matt Kusner et al., “Counterfactual fairness,” in *Advances in Neural Information Processing Systems 30*, ed. Isabelle Guyon et al. (2017); Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, “Inherent tradeoffs in the fair determination of risk scores,” in *8th Innovations in Theoretical Computer Science Conference*, ed. Christos Papadimitriou (Dagstuhl Publishing, 2017)。

[40] 描述软件故障对空中交通管制后果的新闻文章：Simon Calder, “Thousands stranded by flight cancellations after systems failure at Europe’s airtraffic coordinator,” *The Independent*, April 3, 2018。


05
过于智能的
人工智能



大猩猩问题

无须动用太多的想象力，你就能发现，制造出比自己更聪明的东西可能是个坏主意。我们知道，能对环境和其他物种进行控制是我们使用智慧的结果，所以一想到其他比我们更聪明的东西，无论是机器人还是外星人，我们就会立刻产生一种不安的感觉。

大约1 000万年前，现代大猩猩的祖先创造了进化出现代人类的遗传谱系（这肯定是偶然的）。大猩猩对此有何感想？显然，如果它们能告诉我们它们的物种相对于人类的现状，那么其共识意见肯定是非常负面的。它们的物种基本上没有未来，除了我们屈尊所允许它们拥有的未来。我们不希望在超级智能机器面前处于类似的地位。我称这种情况为“大猩猩问题”，具体而言，大猩猩问题就是，人类是否能在一个拥有更高智能的机器的世界里保持人类至高无上的地位和自主权的问题。

1842年，为分析机设计和编写程序的查尔斯·巴贝奇和艾达·洛夫莱斯意识到分析机的潜力，但他们似乎对此毫无疑虑。[1]然而，1847年，宗教期刊《原始解释者》的编辑理查德·桑顿（Richard Thornton）对机械计算机大加指责。[2]

> 当心……通过发明机器来完成自己的思考，我们超越了自身，消除了自身存在的必要性……但是谁知道在这样的机器变得更完美之后，它们会不会想出一个计划来弥补自己的所有缺陷，然后钻研出超越凡人思维范畴的想法！

这可能是人类对计算设备所带来的生存威胁的第一次猜测，但仍然隐晦不清。

相比之下，塞缪尔·巴特勒（Samuel Butler）于1872年出版的小说《埃里汪奇游记》对这一主题进行了更深入的探讨，该小说大获成功。在埃里汪这个国家，拥护机械主义者和反对机械主义者之间爆发了一场可怕的内战，此后所有的机械设备都被禁止使用。书中有一部分叫作“机器之书”，这部分解释了这场战争的起源，并列出了双方的观点。[3]对于在21世纪初重新出现的辩论，这段文字有一种可怕的先见之明。

反对机械主义者的主要论点是，机器将发展到人类无法控制的地步：

> 不正是我们自己在培养取代我们的地球统治者吗？不是我们让它们一天天变得更美丽精密，赋予它们更强大的技能，提高它们优于任何聪明人的自我约束、独立行动的能力吗？……随着时间的推移，我们会发现自己是劣等种族……
>
> 但我们必须选择继续经受目前的痛苦，或是看着自己逐渐被创造物超越，直到我们不再优于它们，就像我们已不再比旷野中的野兽高级一样……束缚会用无法察觉的方法，悄无声息地来到我们身边。

叙述者还谈到了拥护机械主义者的主要反驳观点，这预示了我们将在下一章探讨的人机共生的论点：

> 要严肃地回应这个结论，就只有一种方法。机器是人生理机能的一部分，实际上它们不过是存在于体外的肢体。

虽然埃里汪的反对机械主义者在辩论中获胜，但巴特勒本人似乎在二者之间犹豫不决。一方面，他抱怨说：“埃里汪人能迅速把常识供奉进逻辑的神殿。他们之中的哲学家会利用自己有特殊学问的名声

让埃里汪人昏头昏脑，埃里汪人在机器的问题上忍气吞声。”另一方面，他所描述的埃里汪社会非常和谐，富有成效，甚至如同田园诗一般美好。埃里汪人完全接受重新走上机械发明之路的愚蠢之举，并把那些保存在博物馆里的机械残骸视为“英国古文物家对德鲁伊教纪念碑或燧石箭头的感情”。

巴特勒的故事显然为艾伦·图灵所知，1951年，后者在曼彻斯特的一次演讲中思考了人工智能的长远未来：[4]

> 一旦机器开启思维方式，它们似乎有可能不久就会超越我们微弱的力量。机器不会有死亡的问题，它们会互相交流以增进它们的智能，于是在某个阶段，我们应该预期，机器会按照塞缪尔·巴特勒在《埃里汪奇游记》中所描述的那样掌控世界。

同年，图灵在英国广播公司的一次广播中重复表达了这些担忧。

> 如果一台机器可以思考，它可能会比我们更智能地思考，那么我们应该身居何处？即使我们可以让机器处于从属地位，例如在关键时刻关闭电源，那么我们作为一个物种，也应该感到非常卑微……这种新的危险无疑会让我们感到焦虑。

当埃里汪的反对机械主义者“对未来感到非常不安”时，他们把“趁我们还能做到的时候制止邪恶”视为自己的职责，并销毁了所有机器。图灵对“新危险”和“焦虑”的反应是考虑“关闭电源”（尽管我们很快就会明白这不是一个真正可行的选择）。在弗兰克·赫伯特（Frank Herbert）的经典科幻小说《沙丘》中，故事以遥远的未来为背景，人类在与“会思考的机器”的灾难性战争“巴特勒圣战”中几乎无一幸存。战后，一条新的戒律出现了：“汝等不得创造像人一样思考的机器。”这条戒律禁止了一切类型的计算设备。

所有这些激烈的反应都表明了机器智能在早期引起的恐惧。诚然，超级智能机器的前景确实让人感到不安。从逻辑上讲，这些机器有可能接管世界，征服或消灭人类。如果这就是我们必须要做的一切，那么我们目前唯一可行的对策就是限制人工智能研究，具体来说，就是禁止开发和部署人类级别的通用人工智能系统。

和大多数人工智能研究人员一样，我强烈反对这么做。谁能告诉我，什么是我能想的，什么是我不能想的？一切打算终止人工智能研究的人都必须做很多说服工作。结束人工智能研究意味着我们不仅放弃了一个理解人类智能如何工作的主要途径，而且还放弃了改善人类状况的千载难逢的机会——创造一个更好的文明。人类水平的人工智能的经济价值可以用数千万亿美元来衡量，因此公司和政府有巨大动力进行人工智能研究。正如巴特勒所说，这将压倒哲学家隐晦不清的反对意见，无论哲学家有多么伟大的“特殊学问的名声”。

禁止通用人工智能的第二个不利条件是，它很难被禁止。随着数学问题的提出和解决，通用人工智能的进展主要出现在世界各地的研究实验室中。我们事先不知道应该禁止哪些想法和方程式，而且即使我们真的禁止了，也无法期望这样的禁令是可执行的或有效的。

让难度进一步加大的是，在通用人工智能方面取得进展的研究人员往往是其他领域的人员。正如我已经说过的那样，对工具人工智能的研究——那些具体的、无害的应用，如游戏、医疗诊断和旅行计划等，通常会引发通用技术的进步，这些技术适用于广泛的问题，并使我们更接近人类水平的人工智能。

由于这些原因，人工智能界以及负责制定法律的政府和负责研究预算的公司，都不太可能通过终止人工智能的进展来应对大猩猩问题。如果大猩猩问题只能通过这种方式来解决，那就根本解决不了。

唯一可能奏效的方法是，人们理解为什么让人工智能变得更好可能是一件坏事。事实证明，我们几千年前就知道答案了。

迈达斯国王问题

我们在第1章看到的诺伯特·维纳对许多领域都产生了深远的影响，包括人工智能、认知科学和控制理论。维纳与同时代的大多数人不同，他特别关注在现实世界中运行的复杂系统的不可预测性。（10岁时，他写了关于这个主题的第一篇论文。）他开始相信，科学家和工程师对他们控制自己的发明的能力过于自信，无论是军事的还是民用的发明，都可能产生灾难性的后果。

1950年，维纳出版了《人有人的用处》[5]，书的封面上写道：“‘机械大脑’和类似的机器可以摧毁人类的价值观，或者让我们以前所未有的方式实现这些价值观。”[6]随着时间的推移，他逐渐完善了自己的观点，到1960年，他发现了一个核心问题：正确且完整地定义真正的人类目的是不可能的。这反过来又意味着我所说的标准模型——人类试图将自己的目的灌输给机器，注定要失败。

我们可以把这个问题称为“迈达斯国王问题”。迈达斯是古希腊神话中的传奇国王，他如愿以偿地得到了他想要的点石成金术——他碰过的任何东西都会变成金子。当他发现这其中包括了他的食物、水和他的家人时，已经太晚了，他最终在痛苦和饥饿中死去。同样的主题在人类神话中无处不在。维纳引用了歌德的“魔法师的弟子”的故事，魔法师的弟子学魔法师念咒语，指示扫帚去打水，但他没有说打多少水，也不知道如何让扫帚停下来。

一种技术上的说法是，我们可能会遭受价值对齐的失败：我们可能会向机器灌输与我们的目标不完全对齐的目标，这或许是在不经意间完成的。直到最近，由于智能机器有限的能力以及它们只能在有限范围内对世界产生影响，我们才免受潜在的灾难性后果的影响。（实际上，大多数人工智能工作都是在实验室中处理一些没有实际意义的问题时完成的。）正如诺伯特·维纳在他1964年出版的《神与魔》一书中所说：[7]

过去，对人类目的的片面和不充分的看法相对无害，只是因为当时受到技术上的限制……这只是人类无力保护我们免受人类愚蠢行为的破坏性影响的众多例子之一。

不幸的是，这段保护期很快就要结束了。

我们已经看到社交媒体上的内容选择算法，是如何以广告收入最大化的名义对社会造成巨大破坏的。如果你在想，广告收入最大化已经是一个不光彩的目标，永远不应该成为我们追求的目标，那么让我们假设，要求未来的超级智能系统去追求“找到癌症治疗方法”这一崇高目标，越快找到越好，因为每3.5秒就会有一人死于癌症。在几个小时内，人工智能系统就已经阅读了所有生物医学文献，并假设了数百万种可能有效但此前未经测试的化合物。在几周内，人工智能系统在每个活人体内诱发了多种不同种类的肿瘤，以便对这些化合物进行医学试验，这便是找到癌症治疗方法的最快方法。天啊！

如果你更喜欢解决环境问题，那么你可以要求机器对抗二氧化碳含量升高导致的海洋快速酸化。这台机器开发了一种新的催化剂，可以促进海洋和大气之间极其快速的化学反应，并恢复海洋的pH值水平。不幸的是，大气中1/4的氧气在这个过程中被消耗殆尽，我们只能痛苦地慢慢窒息而死。天啊！

这类世界末日的场景并不难想到，正如人们可能预料的世界末日的场景那样。但是，在许多情况下，一种精神窒息“用无法察觉的方法，悄无声息地来到我们身边”。迈克斯·泰格马克在《生命3.0》一书的序言部分详细描述了一种场景：一台超级智能机器逐渐取得了对整个世界经济上和政治上的控制，但这一行为基本上没被人发现。互联网和它所支持的全球范围的机器——那些每天与数十亿名用户互动的机器，为机器控制人类的发展提供了完美的媒介。

我预测，输入这种机器的“目的”并不是用各种各样的方式接管世界，而更有可能是利润最大化或参与度最大化，抑或是一个其他看

似良性的目标，比如在定期的用户幸福度调查中获得更高的分数，或者减少我们的能源使用。现在，如果我们认为自己是“通过行动有望实现我们的目标的实体”，那么有两种方法可以改变我们的行为。第一种是老派的方法：保持我们的期望和目标不变，但要改变我们的环境，例如，通过给我们钱财，用枪指着我们，或者让我们忍饥挨饿而屈服。对计算机而言，这往往既昂贵又困难。第二种方法是改变我们的期望和目标，这对机器来说要容易得多。机器每天与你保持数小时的联系，控制你对信息的访问，并通过游戏、电视、电影和社交互动为你提供大量的娱乐。

优化社交媒体点击率的强化学习算法没有推理人类行为的能力，事实上，它们甚至不知道人类的存在。对于那些对人类的心理、信念和动机有着更深刻理解的机器，逐步引导我们朝着提高机器目标的满意度的方向前进应该相对容易。例如，它可能通过说服我们少生孩子来减少我们的能源消耗，并且在不经意间最终实现了反自然主义哲学家的梦想——消除人类对自然世界的有害影响。

通过一些实践，你可以学会识别一些方法，在这些方法中，你或多或少实现了某个固定目标却导致了糟糕的结果。最常见的模式之一是从你真正关心的目标中忽略一些东西。这种情况就像上面给出的例子一样，人工智能系统通常会找到一个最佳的解决方案，把你确实关心但忘记提及的事情设置为一个极值。所以，如果你对你的无人驾驶汽车说“尽快把我送到机场”，从字面上解释，它会开到每小时180英里，然后你就会进监狱。（幸运的是，目前计划制造的无人驾驶汽车不会接受这样的请求。）如果你说“在不超过限速的情况下，尽快把我送到机场”，那么它会尽可能多地加速和刹车，在车流中左右穿梭以保持最高速度，它甚至可能会把其他车辆挤走，以便在机场候机楼的拥堵中争取几秒钟的时间。最终，你会添加足够多的注意事项，使汽车的行驶速度大致相当于一个熟练的人类司机载着乘客匆匆赶往机场的速度。

驾驶是一项仅有局部影响的简单任务，目前为驾驶而构建的人工智能系统并不十分智能。由于这些原因，许多潜在的故障模式都是可以预见的，其他故障模式会在驾驶模拟器或数百万英里的测试中暴露出来，如果出现问题，专业驾驶员随时准备接手。还有一些故障模式只会在更晚时候才会被发现，即当汽车已经上路时，一些奇怪的事情才会出现。

不幸的是，能够产生全球影响的超级智能系统并没有模拟器，也没有重来的机会。仅靠人类去预测和排除机器为实现特定目标而选择的所有灾难性方法当然是非常困难的，甚至是不可能的。一般而言，如果你有一个目标，而超级智能机器有一个不同的、与你冲突的目标，那么此时机器会实现它的目标，你却无法实现自己的目标。

恐惧与贪婪：工具性目标

一台机器追求一个错误的目标听起来已经够糟糕的了，但还有更糟糕的。艾伦·图灵提出的解决方案——在紧急时刻关掉电源，可能行不通。原因很简单：如果你死了，你就不能去取咖啡了。

让我解释一下。假设一台机器的目标是取咖啡。如果它足够聪明，那么它肯定会明白，如果它在完成任务之前就被关掉电源，它就无法实现目标。因此，取咖啡的目标创建了一个必要的子目标，即禁用关机按钮。治疗癌症或计算圆周率也是如此。一旦你死了，你就什么也做不成了，所以我们可以预料，一旦被赋予了明确的目标，人工智能系统就会先发制人地保护自己的存在。

如果这个目标与人类的偏好相冲突，那么我们就会看到《2001：太空漫游》中的情节：哈尔9000计算机为了防止宇航员干扰它的任务，杀死了飞船上5名宇航员中的4名。最后一名宇航员戴夫在一场史诗般的智斗后，设法关掉了哈尔，想必这个情节是为了保持趣味性。如果哈尔真的是超级智能，戴夫早就完蛋了。

重要的是要明白，自我保护不一定是机器的内置本能或首要指令。（所以艾萨克·阿西莫夫的“机器人三定律”中的第三条“机器人要尽可能保护自己的生存”是完全没有必要的。[8]）建立自我保护毫无必要，因为这是一个工具性目标，即一个几乎对任何原始目标而言都有用的子目标。[9]任何有明确目标的实体都会自动像拥有工具性目标一样行动。

在我们当前的系统中，除了活着之外，获得金钱也是一个工具性目标。因此，智能机器可能想要钱，不是因为它贪财，而是因为钱对实现各种目标有用。在电影《超验骇客》中，当约翰尼·德普饰演的教授的大脑被上载到量子超级计算机中时，这台机器做的第一件事就是将自己复制到互联网上的数百万台其他计算机上，这样它就无法被关闭了。它做的第二件事是在股市上迅速大赚一笔，为其扩张计划提供资金。

那么，这些扩张计划到底是什么呢？计划包括设计和建造一台更大的量子超级计算机，进行人工智能研究，以及发现物理学、神经科学和生物学的新知识。这些资源目标——计算能力、算法和知识，也是工具性目标，它们对于实现任何总体目标都是有用的。[10]它们似乎是无害的，直到人们意识到其获取过程将无限制地继续下去。这造成了它们与人类不可避免的冲突。当然，配备了更好的人类决策模型的机器将会预见并挫败我们在这场冲突中的每一步行动。

智能爆炸

I. J.古德（I. J. Good）是一位杰出的数学家，曾与艾伦·图灵一起在布莱奇利公园[11]工作，他在“二战”期间破译了德国的密码。他和图灵一样对机器智能和统计推理感兴趣。1965年，他发表了著名的论文《关于第一台超级智能机器的一些推测》。[12]从文章第

一句话中可以看出，冷战时的核边缘政策让古德大为惊恐，他认为人工智能可能是人类的救世主：“人类的生存取决于超级智能机器的早期构建。”然而，在文章的后半部分，他变得更加谨慎。他引入了智能爆炸的概念，但是，和他之前的巴特勒、图灵和维纳一样，他也担心智能会失去控制：

> 超级智能机器可以被定义为智能远远超过了人类智力活动的机器。既然机器设计是智力活动之一，那么超级智能机器就可以设计出更为智能的机器。这无疑会带来一场“智能爆炸”，人类的智能将被远远抛在后面，望尘莫及。因此，第一台超级智能机器将是人类最后的发明，前提是这台机器足够驯顺，可以告诉我们如何控制它。奇怪的是，这一点在科幻小说之外很少被提及。

这一段是有关超级人工智能的所有讨论的主要内容，尽管结尾的警告通常会被忽视。古德的观点可以通过以下几点得到强化：超级智能机器不仅能够改进其自身的设计，而且它很可能会这样做，因为正如我们所看到的，智能机器期望从改进其硬件和软件中获益。智能爆炸的可能性经常被认为是人工智能给人类带来危险的主要来源，因为它给我们解决控制问题的时间太少了。[13]

古德的论点将智能爆炸与化学爆炸类比，这当然具有合理性，在化学爆炸中，分子反应释放足够的能量来引发更多的附加反应。然而，从逻辑上讲，提高智能所获得的回报可能会递减，于是这个过程就会逐渐消失，而不是激增。[14]没有明显的方法可以证明爆炸必然会发生。

收益递减的情况本身就很有趣。如果事实证明，随着机器变得更加智能，实现给定百分比的改进变得更加困难，此时就会出现这种情况。（为了便于讨论，我假设通用机器智能可以用某种线性尺度来衡量，但我认为这种尺度未必绝对正确。）在这种情况下，人类也无法

创造出超级智能。如果一台已经超越人类的机器在试图提高自身智能时失去了动力，那么人类将会更快失去动力。

有人认为，创造任意一种特定水平的机器智能都是人类智慧所无法做到的，对此，我从没听到过支持此观点的严肃的论据，但我想人们必须承认这在逻辑上是可能的。“逻辑上可能”和“我愿意拿人类的未来打赌”当然是完全不同的两码事。与人类的创造力打赌似乎是一种失败的策略。

如果智能爆炸真的发生了，又如果我们尚未能解决仅仅稍微超越人类的智能机器的控制问题，比如我们无法阻止它们进行这些递归式的自我完善，那么我们就没有时间解决控制问题，人类也就完了。这就是尼克·波斯特洛姆设想的“硬起飞”场景，在这个场景中，机器的智能在短短几天或几周内以天文数字的速度增长。用图灵的话说，这“肯定会让我们担忧”。

人们对这种担忧的反应可能有：放弃人工智能研究；否认开发高级人工智能存在固有风险；通过设计必须保持在人类控制之下的人工智能系统来理解和减轻风险；或是认输，并把未来让给智能机器。

否认和减轻风险是本书其余部分的主题。正如我已经说过的，放弃人工智能研究既不太可能发生（因为放弃的代价太大），又很难实现。认输是最糟糕的回应，与之相伴的往往是这样一种想法：比我们更聪明的人工智能系统值得继承这个星球，让人类温和地走进那个良夜，想到我们聪颖的机器人子孙正忙着追求自己的目标，这让人感到宽慰。这一观点是由机器人学家和未来学家汉斯·莫拉维克（Hans Moravec）提出的，[15]他写道：“浩瀚的网络空间将充斥着非人类的超级智慧，它们将处理与人类相关的事务，就像我们处理与细菌的关系一样。”这似乎是错误的。对人类而言，价值主要是由有意识的人类经验来定义的。人类或其他有意识的实体的主观经验对我们很重要，如果没有它们，就不会有任何有价值的事情发生。

[1] 洛夫莱斯写道：“分析机没有任何理由自称能创造一切。它可以做任何我们知道如何指令它执行的事情。它可以跟踪分析，但它没有预测任何分析解析或真相的能力。”这是艾伦·图灵驳斥的反对人工智能的论点之一，“Computing machinery and intelligence,” *Mind* 59 (1950): 433–60。

[2] 已知最早的关于人工智能生存性风险的文章：Richard Thornton, “The age of machinery,” *Primitive Expounder* IV (1847): 281。

[3] “机器之书”这部分改编自塞缪尔·巴特勒早先的一篇文章，“Darwin among the machines,” *The Press* (Christchurch, New Zealand), June 13, 1863。

[4] 关于图灵预言人类被征服的另一场演讲：Alan Turing, “Intelligent machinery, a heretical theory” (lecture given to the 51 Society, Manchester, 1951)。Typescript available at turingarchive.org.

[5] 维纳对人类技术控制的先见之明的讨论和保留人类自主权的辩护：Norbert Wiener, *The Human Use of Human Beings* (Riverside Press, 1950)。

[6] 维纳1950年出版的这本书的封面简介与生命未来研究所（一个致力于研究人类面临的生存性风险的组织）的座右铭惊人地相似：“技术正在赋予生命前所未有的蓬勃发展的潜力……也赋予生命前所未有的自我毁灭的潜力。”

[7] 维纳观点的更新源于他对智能机器可能性的不断增加的认识：Norbert Wiener, *God and Golem, Inc.: A Comment on Certain Points Where Cybernetics Impinges on Religion* (MIT Press, 1964)。

[8] 阿西莫夫的“机器人三定律”首次出现在他1942年3月发表的令人震惊的科幻小说《环舞》（*Runaround*）中，三定律如下：

（1）机器人不得伤害人类，或因不作为使人类受到伤害。

（2）除非违背第一定律，机器人必须服从人类的指令。

（3）除非违背第一及第二定律，机器人必须保护自己。

重要的是要理解，阿西莫夫提出这三条定律是为了塑造有趣的故事情节，而不是作为未来机器人专家的严肃指南。阿西莫夫在《环舞》等几篇小说中都说明了从字面上理解定律的问题后果。从现代人工智能的角度来看，这些定律没有承认任何概率和风险因素：因此，将人类暴露于某种伤害的可能性（无论多么微小）的机器人行动的合法性尚不清楚。

[9] 工具性目标的概念源于Stephen Omohundro, “The nature of self improving artificial intelligence” (unpublished manuscript, 2008)。另请参见：Stephen Omohundro, “The basic AI drives,” in *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, ed. Pei Wang, Ben Goertzel, and Stan Franklin (IOS Press, 2008)。

[10] 约翰尼·德普饰演的角色威尔·卡斯特（Will Caster）的目标似乎是解决死后身体转世的问题，这样他就可以和妻子伊芙琳（Evelyn）重聚。这表明首要目标的性质并不重要，工具性目标都是一样的。

[11] 布莱奇利公园是“二战”期间的密码破译中心。——译者注

[12] 智能爆炸思想的最初来源：I. J. Good, “Speculations concerning the first ultraintelligent machine,” in *Advances in Computers*, vol. 6, ed. Franz Alt and Morris Rubinoff (Academic Press, 1965)。

[13] 关于智能爆炸思想影响的一个例子。Luke Muehlhauser, in *Facing the Intelli gence Explosion* (intelligenceexplosion.com), writes, “Good’s paragraph ran over me like a train.”

[14] 收益递减可以这样解释：假设智能提高16%可以让机器能力提高8%，接下来又能提高4%，以此类推。这一过程最终会达到一个极限，比原始水平高出约36%。关于这些问题的更多讨论，请参见：Eliezer Yudkowsky,“Intelligence explosion microeconomics,” technical report 2013-1, Machine Intelligence Research Institute, 2013。

[15] 关于人工智能使人类变得无关紧要的观点，请参见：Hans Moravec, *Mind Children: The Future of Robot and Human Intelligence* (Harvard University Press,1988). See also Hans Moravec, *Robot: Mere Machine to Transcendent Mind*(Oxford University Press, 2000)。


05
不那么伟大
的人工智能
辩论



《经济学人》杂志对尼克·波斯特洛姆的《超级智能》的评论以这样一句话结尾：“将第二个智能物种引入地球，意义深远，值得深思。”[1]大多数人会将此解读为英式轻描淡写的经典例子。当然，你可能会认为，今天的智者们已经在做这种艰难的思考了——参与严肃的辩论，权衡风险和利益，寻找解决方案，找出解决方案中的漏洞等。但是据我所知，这还没有发生。

当你第一次向技术观众介绍这些想法时，你会发现他们的脑袋里会冒出一个个想法，以“但是……但是……但是……”开头，以感叹号结尾。

第一种“但是”是否认。否认者说：“但是这不可能是一个真正的问题，因为XYZ。”有些XYZ反映了一种推理过程，它们可以被宽容地描述为一种一厢情愿的想法，而另一些则更有实质意义。第二种“但是”是转移话题：承认问题是存在的，但是认为我们不应该试图解决问题，要么因为它们是不可解决的，要么因为有比文明的终结更重要的事情要关注，要么因为最好别提它们。第三种“但是”是一种过于简化的即时解决方案：“但是我们不能只做ABC吗？”就像否认一样，有些ABC同样让人立刻感到遗憾，而另一些ABC更接近于认清问题的本质，这或许是出于偶然。

我并不是说，对于“设计拙劣的超级智能机器会给人类带来严重风险”的观点，人们不能有任何合理的反对意见，只是我还没有看到合理的反对意见。既然这个问题看起来如此重要，它就应该得到最高质量的公开辩论。因此，为了开展这场辩论，并希望读者能对此有所贡献，让我简要介绍一下迄今为止出现的一些观点。

否认

否认问题存在是最简单的解决方法。Slate Star Codex（直译为“石板星手抄本”）博客的作者斯科特·亚历山大（Scott Alexander）撰写了一篇关于人工智能风险的著名文章，他写道：[2] “我第一次对人工智能的风险产生兴趣是在2007年左右。当时，大多数人对这个话题的反应是‘哈哈，等有人相信的时候再回来看吧，估计除了网上的疯子之外没人相信’。”

令人遗憾的言论

对一个人终身职业的威胁，可能会使一个非常聪明的、通常很有思想的人说出一些话，但在进一步分析后，他们很可能希望收回这些话。在这种情况下，我不会说出以下论点的作者，他们都是著名的人工智能研究人员。我也列出了对这些论点的反驳，尽管这完全没有必要。

- 电子计算器在算术方面是“超人”，但是计算器并没有接管世界，因此，我们没有理由担心会产生超越人类的人工智能。

反驳：智能不同于算术，计算器的算术能力并不能让它们接管世界。

- 马有超越人的力量，但是我们不用担心马对我们来说是否安全，因此我们不必担心人工智能系统是否安全。

反驳：智能不同于体力，马的体力强并不能让它们接管世界。

- 历史上没有出现过任何机器杀死数百万人的例子，因此，我们可以得出结论：这在未来也是不可能发生的。

反驳：凡事都有第一次，在第一次发生之前，任何事情都是亘古未有的。

• 宇宙中没有任何一个物理量是无限的，包括智能，所以人们对超级智能的担忧被夸大了。

反驳：超级智能并不需要是无限的才会产生问题，物理学允许计算设备比人脑强大几十亿倍。

• 我们并不担心物种灭绝，这就像黑洞在近地轨道上出现一样不太可能发生，那么我们为什么要担心超级人工智能呢？

反驳：如果地球上大多数物理学家都在努力制造这样的黑洞，我们难道不问他们这样做是否安全吗？

情况很复杂

现代心理学的一个主要观点是，单凭一个智商测试分数无法描述人类智力的全部丰富性。[3]该理论认为，智力有不同的维度：空间的、逻辑的、语言的、社会的等。我们在第2章中讲到的足球运动员爱丽丝可能比她的朋友鲍勃在空间上有更高的能力，但她的社交能力较低。因此，我们不能严格按照智商高低划分所有人。

对机器来说更是如此，因为它们的能力范围要窄得多。谷歌搜索引擎和AlphaGo除了是同属于一家母公司的产品外，几乎没有任何共同之处，所以说它们一个比另一个更智能是毫无意义的。这使得“机器智商”的概念产生了问题，这还表明：将未来描述为人类和机器之间的一维智商竞赛是具有误导性的。

《连线》杂志的创始编辑、极具洞察力的科技评论员凯文·凯利将这一论点向前推进了一步。在《超人类人工智能的神话》中，[4]他写道：“智能不是单一的维度，因此‘比人类更聪明’是一个毫无意义的概念。”所有对超级智能的担忧一下子烟消云散了。

现在，一种明显的响应是，机器在智能的所有相关方面都可能超越人类的能力。在这种情况下，即使按照凯文·凯利的严格标准，一台机器也比一个人更聪明。但是这个相当有力的假设并不是反驳凯利

论点的必要条件。以黑猩猩为例，黑猩猩可能拥有比人类更好的短期记忆能力，即使是在回忆数字序列这种面向人的任务中，黑猩猩也强于人类。[5]短期记忆是智力的一个重要方面。根据凯利的观点，人类并不比黑猩猩聪明，事实上，他会声称“比黑猩猩聪明”是一个毫无意义的概念。这对黑猩猩（以及倭黑猩猩、大猩猩、鲸、海豚等）来说是一种冰冷的慰藉，它们的物种之所以能存活下来，只是因为我们开恩允许它们存活而已。对那些已经因人类而灭绝的物种来说，这是更冰冷的慰藉。对那些担心被机器消灭的人来说，这同样是冰冷的慰藉。

这不可能

在1956年人工智能诞生之前，权威知识分子就对此嗤之以鼻，说智能机器是不可能出现的。艾伦·图灵在1950年的开创性论文《计算机器与智能》中花了大量篇幅驳斥这些论点。从那以后，人工智能界一直在抵制哲学家[6]、数学家[7]和其他人提出的类似人工智能不可能出现的主张。在当前关于超级智能的争论中，几位哲学家挖掘出了这些主张，以此证明人类没有什么可恐惧的。[8] [9]这不足为奇。

“AI100”（人工智能百年研究）是斯坦福大学的一个雄心勃勃的长期项目。它的目标是追踪人工智能的进展，或者更准确地说，是“研究和预测人工智能的影响将如何波及人们工作、生活和娱乐的方方面面”。它的第一份重要报告《2030年的人工智能与生活》确实令人大吃一惊。[10]意料之中的是，它强调人工智能在医疗诊断和汽车安全等领域的优势。令人意想不到的是，它认为“与电影不同，超人类机器人种族在可预见的未来世界里不会出现，甚至永远不可能出现”。

据我所知，这是严肃的人工智能研究人员第一次公开支持这样一种观点，即人类水平或超越人类水平的人工智能是不可能出现的，这

一观点出现在人工智能研究飞速发展的时期，当一个又一个障碍被突破之时。这就好像一群顶尖的癌症生物学家宣布他们一直在愚弄我们：他们一直都知道，永远不会有治愈癌症的方法。

是什么促使了这种态度的大转变？报告中没有提供任何论据或证据。（事实上，有什么证据表明在物理上不可能存在比人类大脑更优越的原子排列呢？）我认为有两个原因。第一个是人工智能研究者想要证明大猩猩问题不存在的自然愿望，这对他们而言是一个非常令人不安的前景。当然，如果人类水平的人工智能是不可能出现的，那么大猩猩问题就被巧妙地解决了。第二个原因是部落习性，即人们本能地对可能“攻击”人工智能的东西严阵以待。

认为超级人工智能可能会攻击人工智能的这种说法很奇怪，而更奇怪的是人们用“人工智能永远不会成功实现其目标”这种说法来为人工智能辩护。我们不能简单地靠押宝人类的聪明才智来确保未来的灾难不会发生。

我们以前也下过这样的赌注，但都输了。正如我们早先看到的，20世纪30年代早期，以卢瑟福男爵为代表的物理学机构自信地认为提取原子能是不可能的事情。然而，他的这种自信被利奥·西拉德在1933年发现中子诱发的链式核反应证明是错误的。

西拉德取得的突破发生在一个不幸的时刻，当时正值与纳粹德国的军备竞赛之初，为更广阔的大众的利益而发展核技术是不可能的。几年后，西拉德在实验室里演示了链式核反应后，他写道：“我们关掉了所有设备，回家了。那天晚上，我毫不怀疑这个世界正走向悲伤。”

现在担心还为时过早

我经常看到头脑清醒的人说，人类水平的人工智能在几十年内不太可能到来，因此没有什么可担心的，以此来缓解公众的担忧。例

如，AI100报告称：“没有理由担心人工智能对人类构成迫在眉睫的威胁。”

这个论点有两点是站不住脚的。第一，这是稻草人谬误，即在论辩中有意或无意地曲解对方的论点以便能够更容易地攻击对方。人工智能令人担忧并不是因为迫在眉睫。尼克·波斯特洛姆在《超级智能》一书中写道：“这本书中并没有说我们即将迎来人工智能领域的重大突破，也没有说我们可以准确预测这种发展何时会出现。”第二，长期风险仍然可能引起人们立即的关注。对人类来说，担心一个潜在的严重问题的恰当时间，不仅取决于问题何时出现，还取决于准备和实施解决方案需要多长时间。

例如，如果我们探测到一颗大型小行星将在2069年与地球相撞，我们会说现在担心还为时过早吗？恰恰相反！人类将会启动一个全球性的紧急项目来开发对抗威胁的手段。我们不会等到2068年才开始研究解决方案，因为我们无法提前说出需要多少时间。事实上，美国宇航局的行星防御计划已经在研究可能的解决方案了，尽管“在未来100年里，没有任何已知的小行星会对地球造成重大影响”。如果这让你沾沾自喜，那么他们还说，“大约74%的直径大于460英尺的近地天体仍有待发现”。

如果我们知道气候变化将会带来全球性灾难的风险（预计将在21世纪晚些时候发生），那么现在采取行动预防灾难是否为时过早？相反，可能为时已晚。超人类人工智能的相关时间尺度很难预测，但当然，这意味着它可能会像核裂变一样，来的比预期早得多。

对于“现在担心还为时过早”这一观点，一个广为流行的表述是吴恩达（Andrew Ng）的主张：“这就像担心火星上的人口过剩一样。”[11]（他后来将这一说法从火星升级到了半人马座阿尔法星。）吴恩达曾是斯坦福大学的教授，是机器学习领域的权威专家，他的观点颇有分量。这一主张可以解读为一个类比：风险不仅很容易控制，而且会发生在遥远的未来，更何况我们甚至不太可能从一开始

就试图把数十亿人转移到火星上。然而，这个类比是错误的。我们已经投入了大量的科技资源来创造更强大的人工智能系统，却很少考虑如果我们成功了会发生什么。因此，更恰当的类比是：我们制订了一项计划，将人类迁移到火星，而不考虑我们到达火星后如何呼吸、吃什么、喝什么。有些人可能会认为这个计划是不明智的。或者，你可以从字面上理解吴恩达的观点，并回应说，即使只有一个人登上火星也会造成人口过剩，因为火星的承载能力为零。因此，目前计划将少数人送上火星的团体确实担心火星上出现人口过剩，这就是为什么他们正在开发生命维持系统。

我们是专家

在每一次关于技术风险的讨论中，支持技术的阵营都宣称，所有对风险的担忧都源于无知。例如，艾伦人工智能研究所的CEO、机器学习和自然语言理解领域的著名研究者奥伦·埃齐奥尼（Oren Etzioni）说：[12]

> 每一项技术创新的兴起都让人们感到恐惧。从工业时代初期的织布工把鞋子扔进机械织布机，到今天对杀手机器人的恐惧，我们的反应是不知道新技术会对我们的自我意识和生计产生什么影响。而当我们一无所知的时候，我们恐惧的头脑会填充细节。

《大众科学》杂志发表了一篇题为《比尔·盖茨害怕人工智能，但人工智能研究人员了然于胸》的文章：[13]

> 当你同人工智能研究人员交谈时——再次强调，是真正的人工智能研究人员，即那些努力让系统正常运行的人（系统的运行远非完美），你会发现他们并不担心超级智能现在或将来会悄悄

> 接近他们。与马斯克似乎有意讲述的恐怖故事相反，人工智能研究人员并没有疯狂地安装防火墙召唤室和设立自毁倒计时。

这一分析是基于四个人的样本，他们都在采访中表示，人工智能的长期安全性是一个重要问题。

当时担任IBM副总裁的戴维·肯尼（David Kenny）使用与《大众科学》的那篇文章非常相似的语言，给美国国会写了一封信，信中包含了以下宽慰人心的话：[14]

> 当你真正研究机器智能的科学，并把它实际应用于商业和社会的现实世界时——就像我们在IBM创建我们开创性的认知计算系统沃森时所做的那样，你就会明白，这项技术并不支持当今人工智能辩论中常见的散布恐慌的论断。

这三个例子传达的信息都是一样的：“别听他们的，我们是专家。”现在，人们可以指出，这实际上是一种人身攻击的谬误，即试图通过否认发言者的合法性来反驳这一信息，但即使从表面上看，这个论点也站不住脚。埃隆·马斯克、史蒂芬·霍金和比尔·盖茨当然对科学和技术推理非常熟悉，尤其是马斯克和盖茨，他们监督并投资了许多人工智能研究项目。如果你认为艾伦·图灵、I. J.古德、诺伯特·维纳和马文·明斯基没有资格讨论人工智能，那就更没有道理了。最后，前面提到的斯科特·亚历山大那篇名为《人工智能研究人员对人工智能风险的研究》的博客文章指出：“人工智能研究人员，包括该领域的一些领导者，从一开始就在提出有关人工智能的风险和超级智能的问题方面发挥了重要作用。”他列举了几位这样的研究人员，现在这份名单要长得多。

“人工智能拥护者”的另一个标准修辞手法将他们的对手描述为卢德分子。奥伦·埃齐奥尼提到的“织布工把鞋子扔进机械织布机”

就是这样：卢德分子是19世纪初的手工织布工，他们抗议引进机器来代替他们的熟练劳动力。2015年，信息技术与创新基金会将年度卢德分子奖颁给了“鼓吹人工智能末日的危言耸听者”。对这群“卢德分子”的定义很奇怪，名单中图灵、维纳、明斯基、马斯克和盖茨都榜上有名，而他们都是20世纪和21世纪最杰出的技术进步贡献者。

对卢德主义的指控代表了人们对“所提出的问题的性质”和“提出这些问题的目的”的误解。这就好比一个人因为核工程师指出控制裂变反应的必要性，就指责他们是卢德分子。就“人工智能研究人员突然声称人工智能不可能实现”这种奇怪的现象来说，我认为我们可以将这一令人困惑的插曲归因于捍卫技术进步的部落主义。

转移话题

一些评论家虽然愿意接受风险是真实存在的，但他们仍然提出不采取任何行动的论点。这些论点包括做任何事情都不可能成功，彻底做别的事情的重要性，以及对风险保持沉默的必要性。

你无法控制研究

对于高级人工智能可能给人类带来风险的说法，一个常见的回答是，禁止人工智能研究是不可能的。请注意这里存在思维跳跃：“哦？有人在讨论风险！他们一定是在提议禁止我的研究！”在讨论仅基于大猩猩问题的风险时，这种思维上的跳跃可能是合适的，通过阻止超级人工智能的创造来解决大猩猩问题，需要对人工智能研究施加某种限制，我倾向于同意。

然而，最近关于风险的讨论并没有集中在一般的大猩猩问题上（从新闻的角度来说，就是人类与超级智能的对抗），而是集中在迈达斯国王问题以及类似的问题上。解决迈达斯国王问题的同时也能解

决大猩猩问题，但解决方案不是阻止超级人工智能或找到战胜它的方法，而是确保它从一开始就不与人类发生冲突。人们在对迈达斯国王问题进行讨论时，通常不会提议限制人工智能研究，他们只是建议注意防止设计不良系统所带来的后果。同样，人们对核电站的安全壳失效风险的讨论不应被解释为人们试图禁止核物理研究，而应被解释为一种建议，即把更多的工作放在解决安全壳问题上。

碰巧的是，终止研究有一个非常有趣的历史先例。20世纪70年代初，生物学家开始担心新的重组DNA技术——将一个有机体的基因连接到另一个有机体上，可能会给人类健康和全球生态系统带来巨大的风险。1973年和1975年在加利福尼亚州的阿西洛马举行的两次会议首次让这类实验暂停了下来，然后人们制定了针对一切相关实验风险的详细生物安全指导方针。[15]有些种类的实验，如涉及毒素基因的实验，被认为太危险而被禁止。

1975年会议一结束，几乎资助了美国所有基础医学研究的美国国家卫生研究院（NIH）就开始着手建立“DNA重组咨询委员会”（RAC）。众所周知，后者在前者制定指导方针的过程中起到了重要作用，该方针基本上采纳了在阿西洛马的两次会议的建议。自2000年以来，这些指导方针包括禁止资助任何涉及改变人类种系的方案，即以可被后代继承的方式修改人类基因组。这项禁令出台之后，50多个国家相继颁布了法律禁令。

“改善人类种群”的目标一直是19世纪末、20世纪初优生运动的梦想之一。CRISPR-Cas9是一种非常精确的基因组编辑方法，它的发展重燃了这个梦想。2015年举行的一次国际峰会为未来的申请敞开了大门，会议呼吁各国在“就申请的适当性达成广泛的社会共识”之前保持克制。[16] 2018年11月，贺建奎宣布，一对基因编辑婴儿诞生。随后，国际上一片哗然。2019年3月，一个由顶尖科学家组成的国际小组明确呼吁正式暂停基因组编辑。[17]

这场辩论给人工智能领域上了复杂的一课。一方面，这表明我们有能力选择不去开拓一个有巨大潜力的研究领域。迄今为止，反对改变人类种系的国际共识占压倒性优势。人们担心禁令只会让研究转入地下，或进入没有监管的国家，但这种担心没能成为现实。另一方面，种系改变是一个很容易识别的过程，是一个关于更通用的遗传学知识的特定用例，需要专门的设备和真人进行试验。此外，它属于一个已经受到密切监督和监管的领域——生殖医学。这些特征不适用于通用人工智能，而且到目前为止，还没有人提出以任何合理的形式来限制人工智能研究。

诡辩主义

一位英国政治家的顾问向我介绍了“诡辩主义”这个词，他必须定期在公开会议上处理这个问题。无论他演讲的主题是什么，总会有人问：“巴勒斯坦人的困境怎么办？”

在谈到高级人工智能的风险时，人们可能会听到：“人工智能所带来的好处怎么办？”举例来说，奥伦·埃齐奥尼认为：[18]

> 悲观的预测往往没有考虑人工智能在防止医疗失误、减少车祸等方面的潜在好处。

以下是脸书CEO马克·扎克伯格最近在与埃隆·马斯克进行媒体对话时所说的：[19]

> 如果你反对人工智能，那么你就是在反对制造不会发生事故的、更安全的汽车，反对为病人提供更好的诊断。

撇开一切只要提到风险就是“反对人工智能”的部落观念不谈，扎克伯格和埃齐奥尼都认为，谈论风险就是忽视甚至否定人工智能的潜在好处。

这恰恰是一种倒退，原因有二。首先，如果人工智能没有潜在的好处，那么人工智能研究就不会有任何经济或社会动力，因此也就不会有出现达到人类水平的人工智能的危险。我们根本就不会进行这样的讨论。其次，如果不能成功降低风险，人们就不会得到任何好处。1979年美国三里岛核电站部分堆芯熔毁、1986年切尔诺贝利核电站失控的反应和灾难性的泄漏，以及2011年福岛核电站的多次熔毁，大大降低了核能的潜在效益。这些灾难严重限制了核工业的发展。意大利在1990年放弃了核能，比利时、德国、西班牙和瑞士也宣布了类似的计划。自1990年以来，全球核电站的设备利用率大约是切尔诺贝利事故之前的10%。

沉默

最极端的转移话题的形式就是，建议我们对风险保持沉默。例如，前面提到的AI100报告包括以下警告：

> 如果社会主要以恐惧和怀疑的态度对待这些技术，就会延缓人工智能的发展或让其转入地下，从而对确保人工智能技术的安全性和可靠性方面的重要工作造成阻碍。

信息技术与创新基金会（颁发卢德分子奖的那个基金会）的主任罗伯特·阿特金森（Robert Atkinson）在2015年的一场辩论中提出了类似的观点。[20]尽管在如何向媒体准确地描述风险方面存在一些疑问，但其观点的总体信息很明确：“不要提及风险，这不利于融

资。”当然，如果没有人意识到风险，用于缓解风险的研究就不会获得资金，也没有人有理由去研究它。

著名认知科学家史蒂芬·平克（Steven Pinker）对阿特金森的观点提出了更为乐观的看法。他认为，“先进社会的安全文化”将确保消除人工智能可能带来的所有严重风险，因此，提醒人们注意这些风险不仅是不恰当的，还会适得其反。[21]即使我们放下我们先进的安全文化导致了切尔诺贝利、福岛的核事故和失控的全球变暖这些事实不谈，平克的论点也完全没有切中要害。安全文化恰恰是由人们指出可能的故障模式，并找到确保故障不会发生的方法组成的。（对于人工智能，标准模型就是故障模式。）说指出一种故障模式是荒唐的，因为安全文化无论如何都会解决这一问题，就像说看到肇事逃逸事故时不应该叫救护车，因为有人会叫救护车一样。

在试图向公众和决策者描述风险时，人工智能研究人员与核物理学家相比处于劣势。物理学家不需要写书向公众解释聚集临界质量的高浓缩铀可能存在风险，因为其后果已经在广岛和长崎得到了证明。因此他们并不需要大量进一步的劝说，就能让各国政府和资助机构相信安全性在发展核能中是重要的。

部落主义

在巴特勒的《埃里汪奇游记》中，对大猩猩问题的关注导致了拥护机械主义者和反对机械主义者之间过早的和错误的截然对立。拥护机械主义者认为，机器统治的风险很小或根本不存在；反对机械主义者认为机器是不可战胜的，除非所有机器都被摧毁。争论变成了部落式的，没有人试图解决“如何保持人类对机器的控制”这一根本问题。

20世纪的所有主要技术问题——核能、转基因生物和化石燃料，在不同程度上都屈从于部落主义。每个问题都有赞成和反对两个方面。每个部落的动态和结局都不同，但是部落主义的症状是相似的：相互不信任和诋毁、非理性地争论、拒绝承认任何可能有利于另一个部落的（合理）论点。在支持技术的一方，人们看到的是对风险的否认和隐瞒，以及对卢德主义的指责；在反对技术的一方，人们看到了一种信念，即风险是无法克服的，问题是无法解决的。支持技术的部落成员在“问题”上过于诚实会被视为叛徒，这尤其不幸，因为支持技术的部落成员通常包括大多数有能力解决这个问题的人。反对技术部落中讨论“可能的缓解措施”的成员也会被视为叛徒，因为被视为邪恶的是技术本身，而不是其可能带来的影响。于是，只有最偏激的声音——那些最不可能被对方听取的声音，才能代表每个部落。

2016年，我应邀前往唐宁街10号，与当时的英国首相戴维·卡梅伦（David Cameron）的几位顾问会面。他们担心人工智能辩论正开始变得类似于转基因辩论。在欧洲，转基因辩论导致顾问们认为：对转基因生产和标签的规定过于严格且为时过早。他们想避免同样的事情发生在人工智能领域。他们的担忧不无道理：有关人工智能的辩论正面临部落化，即建立支持人工智能和反对人工智能两大阵营的危险。这将对人工智能领域造成损害，因为担心高级人工智能固有的风险是一种反对人工智能的立场。担心发生核战争或设计不良的核反应堆爆炸的物理学家并不反对物理学。我们说人工智能将强大到足以产生全球影响，是对该领域的赞美，而不是侮辱。

人工智能界必须承担风险并努力缓解风险，这一点至关重要。就我们所了解的风险而言，这些风险既非微不足道，也非不可克服。我们需要做大量的工作来避免它们，包括重塑和重建人工智能的基础。

难道我们不能……

难道我们不能……关闭电源？

一旦理解了生存性风险的基本思想，无论是大猩猩问题还是迈达斯国王问题，包括我自己在内的许多人都会立即开始寻找一个简单的解决方案。通常，首先出现在脑海中的方案是关闭机器。例如，如前所述，艾伦·图灵本人推测我们可能会“让机器处于从属地位，例如在关键时刻关闭电源”。

这是行不通的，原因很简单，一个超级智能实体会事先想到这种可能性，并采取措施来阻止它。它这样做不是因为它想要生存，而是因为它在追求我们赋予它的任意目标，并且知道如果自己被关闭，这个目标就无法达成。

人们正在构想一些系统，如果不破坏我们文明的许多管线，我们就无法关闭这些系统。这些系统在区块链中以所谓的智能契约的形式实现。区块链是基于加密的高度分布式的计算和记录保存形式，它经过专门设计，使得一个人在不控制大量机器并解开链的情况下，基本上不能删除任何数据，也不能终止任何智能契约，而这又可能会破坏互联网或金融系统的很大一部分。这种不可思议的鲁棒性到底是特性还是缺陷尚且存在争议。这当然是一个超级人工智能系统可以用来保护自己的工具。

难道我们不能……把它放进盒子里？

如果你不能关闭人工智能系统，那么你能否把机器密封在某种防火墙内，从它们那里提取有用的问答，但绝不允许它们直接影响现实世界？这就是Oracle AI（甲骨文人工智能）背后的理念，人工智能安全界对此进行了详细的讨论。[22] Oracle AI系统可以是任意智能的，但对每个问题只能回答是或否（或给出相应的概率）。它可以通过只读连接访问人类拥有的所有信息，也就是说，它不能直接访问互联网。当然，这意味着放弃超级智能机器人、助手和许多其他类型的

人工智能系统，但是值得信赖的Oracle AI仍然具有巨大的经济价值，因为我们可以问它一些问题，这些问题的答案对我们而言很重要，例如阿尔茨海默病是否是由传染性有机体引起的，禁止自主武器是否是好主意。因此，Oracle AI无疑是一种有趣的选择。

不幸的是，Oracle AI面临一些很难克服的困难。首先，该系统将至少像我们理解自己的世界一样，孜孜不倦地理解它的物理性质和起源，包括它的计算资源、工作原理，以及制作了它的信息库并正在提问的神秘实体。其次，如果Oracle AI系统的目标是在合理的时间内为问题提供准确的答案，那么它就会有动力冲出牢笼，获取更多的计算资源，并控制提问者，让他们只问简单的问题。最后，我们还没有发明出一种针对普通人的安全防火墙，更不用说针对超级智能机器的安全防火墙了。

我认为其中一些问题可能会有解决方案，特别是如果我们把Oracle AI系统限制为可证明的合理逻辑或贝叶斯计算器。也就是说，我们可以坚持让算法只能输出根据所提供的信息得出的结论，我们可以用数学方法验证算法是否满足这个条件。但这仍然留下了一个控制过程的问题，这个过程决定了要进行哪些逻辑计算或贝叶斯计算，以便尽可能快地得出最有力的结论。因为这个过程有快速推理的动机，所以它有获取计算资源的动机，当然也有保持自身存在的动机。

2018年，伯克利的人类兼容人工智能中心举办了一个研讨会，人们在会上提出了这样一个问题：“如果你确定超级人工智能将在10年内实现，你会怎么做？”我回答说，说服开发人员推迟构建通用智能体，即那些可以在现实世界中选择自己行为的智能体，转而构建Oracle AI。同时，我们要致力于解决使Oracle AI系统可证明安全性的问题。这种策略可能奏效的原因有二：首先，超级智能的Oracle AI系统仍然价值数万亿美元，因此开发人员可能愿意接受这个限制；其次，控制Oracle AI系统肯定比控制通用智能体更容易，所以我们在10年内解决这个问题的机会更大。

难道我们不能……在人机团队中工作？

在企业界，一种普遍的看法是，人工智能不会对就业或人类构成威胁，因为我们只有人机协作团队。例如，本章前文中引用的戴维·肯尼写给国会的信中说：“高价值人工智能系统是专门为增强人类智能而设计的，而不是为了取代工人。”[23]

尽管愤世嫉俗的人可能会认为这只是一种公关策略，其目的是粉饰从其公司中剔除人类员工的过程，但我认为，这确实让事情有了些许进展。人机协作团队的确是一个理想的目标。显然，如果团队成员的目标不一致，该团队就不会成功，因此强调人机团队突出了解决价值对齐这一核心问题的必要性。当然，强调问题并不等于解决问题。

难道我们不能……与机器融合？

最极端的情况是，人机协作变成了人机融合，电子硬件直接连接到大脑，成为一个单一的、可扩展的、有意识的实体的一部分。未来学家雷·库兹韦尔（Ray Kurzweil）将这种可能性描述为：[24]

> 我们将直接与它融合，我们将成为人工智能……当我们进入21世纪30年代末或40年代，我们的思维将主要是非生物的，非生物的部分最终会非常智能，拥有非常巨大的能力，从而能够完全建模、模拟和理解生物部分。

库兹韦尔从积极的角度看待这些发展。相反，埃隆·马斯克认为人机融合主要是一种防御策略：[25]

> 如果我们实现了紧密的共生，人工智能就不会是“他者”，它将会是你，它将会与你的大脑皮层产生一种类似于你的大脑皮

层与边缘系统的关系……我们将面临选择：要么被抛弃，变得毫无用处，要么像家猫之类的宠物一样被圈养起来，要么最终找到某种方式与人工智能共生融合。

马斯克的Neuralink（神经链）公司正在研发一种名为“神经花边”的技术，这个叫法来自伊恩·班克斯（Iain Banks）在其“文明”系列小说中描述的一种技术。目的是在人类大脑皮层和外部计算系统、网络之间建立一个稳定的、永久的连接。“神经花边”目前存在两个主要的技术障碍：第一，难以将电子设备连接到脑组织，为脑组织供电，并将其连接到外部世界；第二，事实上，我们对大脑中实现更高层次认知的神经几乎一无所知，所以我们不知道应该把设备连接到哪里，更不知道应该如何处理。

我并不完全认为前一段中的障碍是无法克服的。第一，“神经尘埃”这样的技术正在迅速降低电子器件的尺寸和功率要求，这些器件可以连接到神经元上，提供传感、刺激和经颅通信。[26]（2018年，这项技术的电子器件尺寸已经缩小到了大约1立方毫米，因此“神经砂砾”可能是一个更准确的术语。）第二，大脑本身具有非凡的适应能力。例如，过去人们认为，我们必须理解大脑用来控制手臂肌肉的“代码”，然后才能成功将大脑与机器手臂连接；我们必须理解耳蜗分析声音的方式，然后才能为其制造替代品。事实证明，大脑为我们做了大部分工作。它很快就学会了如何让机器手臂做主人想做的事情，以及如何将人工耳蜗的输出映射成可理解的声音。我们完全有可能找到为大脑提供额外记忆的方法，找到与计算机沟通的渠道，甚至找到与其他大脑沟通的渠道——这一切都是在我们从未真正理解它的工作原理的情况下实现的。[27]

不管这些想法在技术上是否可行，人们都必须问一问，这个方向是否代表了人类最好的未来。如果人类仅仅为了摆脱自身技术的威胁而需要进行脑外科手术，那么我们或许已经在某个环节犯了错误。

难道我们不能……避免输入人类目标？

一个常见的推理思路是，有问题的人工智能行为是由我们输入了特定种类的目标而引起的。如果删掉这些目标，一切都会好的。例如，深度学习先驱、脸书公司人工智能研究主管杨立昆在淡化人工智能的风险时，经常引用以下观点：[28]

> 让人工智能有自我保护的本能、嫉妒等是没有理由的……人工智能不会有这些破坏性的“情绪”，除非我们将这些情绪植入其中。我不明白我们为什么要这么做。

类似的，史蒂芬·平克提供了基于性别的分析：[29]

> 人工智能反乌托邦主义者将狭隘的“阿尔法男性”[30]的心理投射到智能的概念上。他们认为，具有超人智能的机器人会发展出一些目标，比如抛弃它们的主人，接管世界等。这说明，许多技术预言家不相信人工智能会自然地沿着女性的路线发展，既完全有能力解决问题，但又没有消灭无辜者或统治文明的欲望。

正如我们在讨论工具性目标时看到的那样，无论我们植入“情感”还是“欲望”，例如自我保护、获取资源、发现知识，或是在极端情况下接管世界，这些都无关紧要。无论机器是什么性别，它都会有这些情感，就像我们构建的任意目标的子目标一样。对机器而言，死亡本身并不是什么坏事。尽管如此，死亡还是要避免的，因为“如果你死了，你就很难去取咖啡了”。

更极端的解决方案是避免将目标全部输入机器中。瞧，问题解决了。但是很遗憾，事情没那么简单。没有目标，就没有智能：任何行动都是好的，机器可能变得和随机数生成器一样。没有目标，机器就

没有理由喜欢人类的天堂胜过一个堆满回形针的星球（尼克·波斯特洛姆详细描述了这一场景）。事实上，后一种结果对于以铁为食的氧化亚铁硫杆菌来说可能是乌托邦。如果没有“人类偏好很重要”这一概念，谁能说细菌是错的呢？

“避免输入目标”思想的一种常见变体是这样一个概念：一个足够智能的系统必然会由于其智能而自行制定“正确”的目标。通常，这一概念的支持者会认同一种理论，即高智商的人往往有更多无私和崇高的目标——这一观点可能与其支持者的自我概念有关。

18世纪著名的哲学家大卫·休谟（David Hume）在《人性论》中详细讨论了在世界中感知目标的可能性。[31]他称其为“实然与应然问题”，并得出结论，认为“道德上必须做的事可以从自然事实中推断出来”的想法完全是错误的。究其原因，我们可以考虑一下棋盘和棋子的设计。人们不能从棋盘和棋子中看出“将军”的目的，因为同样的棋盘和棋子也可以用来下自杀象棋，或者其他许多仍有待发明的游戏。

在《超级智能》一书中，尼克·波斯特洛姆以一种不同的形式提出了同样的基本观点，他称之为“正交性理论”：

> 智能和最终目标是正交的：原则上，几乎任何水平的智能都能与几乎任何最终目标相结合。

这里，正交的意思是“成直角”，即智能程度是定义智能系统的一个轴，而目标是另一个轴，我们可以独立地改变它们。例如，无人驾驶汽车可以指定任意特定地址作为目的地；让汽车成为更好的司机并不意味着它会拒绝前往那些“可以被17整除的地址”。同样，我们很容易想象到一个通用智能系统可以或多或少地被赋予任意目标，包括最大化回形针的数量或已知圆周率的位数。这就是强化学习系统和其他类型的奖励优化器的工作原理：算法是完全通用的，可以接受任

意奖励信号。对于在标准模型下工作的工程师和计算机科学家来说，正交性理论只是一个给定事实。

智能系统可以通过观察世界来获得应该追求的目标，这一想法表明，一个足够智能的系统会自然而然地放弃其最初的目标，转而追求“正确”的目标。理解为什么理性智能体会这样做很难。此外，它预先假定世界上存在一个“正确”的目标，该目标必须是嗜铁细菌、人类，以及所有其他物种都认同的目标，这很难想象。

对波斯特洛姆正交性理论最明确的批评来自著名的机器人学家罗德尼·布鲁克斯（Rodney Brooks），他断言，一个程序不可能“足够聪明，使它能够发明方法来颠覆人类社会，以此实现人类为其设定的目标，却不了解它这么做会给为其设定目标的同一群人带来问题”。[32]不幸的是，一个程序很有可能像这样运行，而且根据布鲁克斯对这个问题的定义，这实际上是不可避免的。布鲁克斯认为，“实现人类设定的目标”的最佳计划正在给人类带来问题。由此可见，这些问题反映了人类在设定目标时忽略了的有价值的东西。机器执行的最佳计划很可能会给人类带来问题，而机器很可能也意识到了这一点。但是根据定义，机器不会认为这些问题会造成问题。这不关它们的事。

史蒂芬·平克似乎同意波斯特洛姆的正交性理论，他写道：“智能是运用创新手段实现目标的能力。目标与智能本身无关。”[33]另外，他发现不可思议的是，“人工智能非常聪明，它能弄明白如何改变元素、改造大脑，但它又非常愚蠢，会因为基本的误解而造成严重破坏”。[34]他继续说：“让机器去选择最能满足相互冲突的目标的动作，这种能力不是工程师可能忘记安装和测试的附加功能，这种能力就是智能。在语境中理解语言使用者意图的能力也是智能。”当然，“满足相互冲突的目标”并不是问题所在，这是从决策理论早期就已被内置到标准模型的东西。问题是，机器意识到的相互冲突的目标并不构成人类关注的全部。此外，在标准模型中，机器必须关心它没有被告知要关心的目标。

然而，从布鲁克斯和平克的话中，我们可以找到一些有用的线索。对我们来说，作为追求其他目标的副作用，机器改变天空的颜色，却忽视了人类对此的明显不快，这确实是愚蠢的。这对我们来说之所以很愚蠢，是因为我们习惯于注意到人类的不快，而且通常我们有动机避免引起不快，即便我们之前没有意识到这些人关心天空的颜色。也就是说，我们人类一来关心其他人的偏好，二来知道我们不知道所有人的偏好是什么。在下一章中，我将讨论这些特性，当它们被构建到机器中时，这些特性可能会为解决迈达斯国王问题提供一个契机。

辩论重新开始

这一章让我们得以一窥广大知识界正在进行的一场辩论，一场在指出人工智能生存性风险的人和对这些风险持怀疑态度的人之间进行的辩论。在书籍、博客、学术论文、小组讨论、访谈、推文和报纸文章中都可以见到它的踪迹。尽管“怀疑论者”——那些认为人工智能的风险可以忽略不计的人，做出了种种勇敢的努力，但他们未能解释为什么超级人工智能系统必然会一直处于人类的控制之下，他们甚至没有试图解释为什么超级人工智能系统永远不会被开发出来。

如果被追问，许多“怀疑论者”会承认这里确实存在问题，即便问题并不是迫切需要解决的。下面是斯科特·亚历山大在他的Slate Star Codex博客上精辟的总结。[35]

> “怀疑论者”的立场是：尽管我们可能应该让一些聪明人开始着手研究问题初步显露的一些方面，但我们不应该对人工智能研究感到恐慌，或将其禁止。
>
> 与此同时，“信徒们”坚持认为：尽管我们不应该对人工智能研究感到恐慌，或将其禁止，但我们可能应该让一些聪明人开

始着手研究问题初步显露的一些方面。

如果“怀疑论者”能提出一个无可辩驳的反对意见，也许是以一种简单且万无一失的（同时又绝不邪恶的）解决方案的形式来解决人工智能的控制问题，我会感到很高兴，但我认为这很可能不会发生，就像我们找不到简单且万无一失的网络安全解决方案，也找不到简单且万无一失的方式来制造零风险的核能一样。正如亚历山大所说，与其继续进行部落间的谩骂，反复挖掘不可信的论据，不如着手研究问题初步显露的一些方面。

这场辩论凸显了我们面临的难题：如果我们构建机器来优化目标，那么我们输入机器的目标必须符合我们想要的目标，但我们不知道如何完全正确地定义人类目标。幸运的是，我们有折中的办法。

---

[1] 对波斯特洛姆的《超级智能》进行了严肃评论的一篇文章：“Clever cogs,” *Economist*, August 9, 2014。

[2] 关于人工智能风险的神话和误解的讨论：Scott Alexander, “AI researchers on AI risk,” *Slate Star Codex* (blog), May 22, 2015。

[3] 关于智能的多重维度的经典著作：Howard Gardner, *Frames of Mind: The Theory of Multiple Intelligences* (Basic Books, 1983)。

[4] 关于智能的多重维度对超人类人工智能可能性的影响：Kevin Kelly, “Themyth of a superhuman AI,” *Wired*, April 25, 2017。

[5] 黑猩猩比人类有更优秀的短期记忆的证据：Sana Inoue and Tetsuro Matsuzawa“,Working memory of numerals in chimpanzees,” *Current Biology* 17 (2007),R1004–5。

[6] 对基于规则的人工智能系统的前景提出质疑的一项重要的早期工作：Hubert Dreyfus, *What Computers Can’t Do* (MIT Press, 1972)。

[7] 寻求意识的物理解释并对人工智能系统实现真正智能的能力提出质疑的一系列书籍中的第一本：Roger Penrose, *The Emperor’s New Mind: Concerning Computers, Minds, and the Laws of Physics* (Oxford University Press, 1989)。

[8] 基于不完备定理的人工智能批判的复兴：Luciano Floridi, “Should we be afraid of AI?” *Aeon*, May 9, 2016。

[9] 基于中文屋论证的对人工智能的批判的复兴：John Searle, “What your computer can’t know,” *The New York Review of Books*, October 9, 2014。

[10] 杰出人工智能研究人员的一份报告，声称超越人类的人工智能或许是不可能的：Peter Stone et al., “Artificial intelligence and life in 2030,” One Hundred Year Study on Artificial Intelligence, report of the 2015 Study Panel, 2016。

[11] 基于吴恩达对人工智能带来的风险不屑一顾的新闻文章：Chris Williams, “AIguru Ng: Fearing a rise of killer robots is like worrying about overpopulation on Mars,” *Register*, March 19, 2015。

[12] “专家最了解”这一论点的一个例子：Oren Etzioni, “It’s time to intelligently discuss artificial intelligence,” *Backchannel*, December 9, 2014。

[13] 新闻报道声称真正的人工智能研究人员对风险不屑一顾：Erik Sofge, “Bill Gates fears AI, but AI researchers know better,” *Popular Science*, January 30,2015。

[14] 真正的人工智能研究人员对人工智能风险不屑一顾的另一个说法：David Kenny, “IBM’s open letter to Congress on artificial intelligence,” June 27, 2017,ibm.com/blogs/policy/kenny-artificial-intelligence-letter。

[15] 提出自愿限制基因工程的研讨会报告：Paul Berg et al., “Summary statement of the Asilomar Conference on Recombinant DNA Molecules,” *Proceedings of the National Academy of Sciences* 72 (1975): 1981–84。

[16] 源于CRISPR-Cas9基因编辑发明的政策声明：Organizing Committee for the International Summit on Human Gene Editing, “On human gene editing:International Summit statement,” December 3, 2015。

[17] 主要生物学家的最新政策声明：Eric Lander et al., “Adopt a moratorium on heritable genome editing,” *Nature* 567 (2019): 165–68。

[18] 埃齐奥尼在分析人工智能研究人员的调查数据时表示，如果不提到好处，就不会提到风险。Oren Etzioni, “No, the experts don’t think superintelligent AI is a threat to humanity,” *MIT Technology Review*, September 20, 2016. 在他的分析中，他认为任何预测超越人类的人工智能需要超过25年的人都不关心人工智能的风险，包括这位作者和尼克·波斯特洛姆。

[19] 一篇引用马斯克和扎克伯格“辩论”的新闻报道：Alanna Petroff, “Elon Musk says Mark Zuckerberg’s understanding of AI is ‘limited,’ ” *CNN Money*, July 25,2017。

[20] 2015年，信息技术与创新基金会组织了一场名为“超级智能计算机真的会对人类构成威胁吗？”的辩论。该基金会主任罗伯特·阿特金森表示，提及风险可能会导致投入人工智能的资金减少。视频请见itif.org/events/2015/06/30/are-super-intelligent-computers-really-threat-humanity，相关讨论从41分30秒开始。

[21] 一种声称我们的安全文化将解决人工智能控制问题却从未提及的说法：Steven Pinker, “Tech prophecy and the underappreciated causal power of ideas,” in *Possible Minds: Twenty-Five Ways of Looking at AI*, ed. John Brockman (Penguin Press, 2019)。

[22] 关于Oracle AI的有趣分析，请参见Stuart Armstrong, Anders Sandberg, and Nick Bostrom, “Thinking inside the box: Controlling and using an Oracle AI,” *Minds and Machines* 22 (2012): 299–324。

[23] 关于人工智能为什么不会抢走工作的观点：Kenny, “IBM’s open letter.”

[24] 库兹韦尔对人脑与人工智能融合的积极看法的一个例子：Ray Kurzweil,interview by Bob Pisani, June 5, 2015, Exponential Finance Summit, New York,NY。

[25] 引用了关于埃隆·马斯克“神经花边”的文章：Tim Urban, “Neuralink and the brain’s magical future,” Wait But Why, April 20, 2017。

[26] 关于加州大学伯克利分校“神经尘埃”项目的最新进展，请参见：David Piech et al., “Stim Dust: A 1.7 mm3, implantable wireless precision neural stimulator with ultrasonic power and communication,” arXiv: 1807.07590 (2018)。

[27] 苏珊·施耐德（Susan Schneider）在*Artificial You: AI and the Future of Your Mind*(Princeton University Press, 2019)一书中指出，上传和神经修复等拟议技术存在未知的风险：如果对电子设备是否可以有意识缺乏任何真正的理解，并且考虑到持续存在的对个人身份的哲学困惑，我们可能会在无意中结束我们自己的意识存在，或者在不知不觉中给有意识的机器带来痛苦。

[28] 关于人工智能的风险为题对杨立昆进行的采访：Guia Marie Del Prado, “Here’s what Facebook’s artificial intelligence expert thinks about the future,” *Business Insider*, September 23, 2015。

[29] 睾酮过量引起的人工智能控制问题的诊断：“Thinking does not imply subjugating,” in *What to Think About Machines That Think*, ed. John Brockman(Harper Perennial, 2015)。

[30] 阿尔法男性是指位于社会顶层的男性，他们通常可以利用自己的生理优势和强大的控制欲获得更多的资源。——译者注

[31] 关于许多哲学主题的开创性著作，包括道德义务是否可以在自然界中被感知的问题：David Hume, *A Treatise of Human Nature* (John Noon, 1738)。

[32] 一个关于足够智能的机器必须追求人类目标的观点：Rodney Brooks, “Theseven deadly sins of AI predictions,” *MIT Technology Review*, October 6, 2017。

[33] Pinker, “Thinking does not imply subjugating.”

[34] 一种乐观的观点认为人工智能的安全问题一定会以对我们有利的方式得到解决：Steven Pinker, “Tech prophecy”。

[35] 人工智能风险中“怀疑论者”和“信徒”之间的不可思议的一致：Alexander, “AI researchers on AI risk”。

07 人工智能：一种不同的方法



一旦怀疑论者的论点被驳斥，所有的“但是……但是……但是……”都得到了回答，此时的下一个问题通常是：“好吧，我承认有问题，但是你没有解决方案，对吗？”错，有解决方案。

让我们提醒自己手头的任务：设计高度智能的机器，让它们帮助我们解决难题，同时确保这些机器的行为永远不会让我们感到非常恼火。

幸运的是，这个任务并不是给定你一台高度智能的机器，让你找出控制它的方法。如果任务是这样的，我们就完蛋了。一台被视为黑盒子的、已成现实的机器，没准儿是从外太空来的。我们能控制一个从外太空来的超级智能实体的概率几乎为零。有这样一些相似的方法可以创建人工智能系统，比如全脑模拟，[1]即创建增强的人脑电子副本，或者基于程序的模拟进化的方法，[2]但我们无法理解这些方法的工作原理。关于这些提议，我就不多说了，因为它们显然是馊主意。

那么，过去人工智能领域是如何接近“设计高度智能的机器”这一部分任务的呢？与许多其他领域一样，人工智能领域采用了标准模型：我们构建优化机器，将目标输入其中，然后它们就会运行。当机器愚蠢且行动范围有限时，这种方法很奏效。如果你输入了错误的目标，你完全有机会可以关闭机器，修复问题，然后再试一次。

然而，随着根据标准模型设计的机器变得更加智能，以及它们的行动范围遍及全球，这种方法越来越不可行。这些机器将会追求它们自己的目标，无论目标错得多么离谱；它们将会抵制关闭它们的企图；它们将会获得有助于实现目标的一切和全部资源。实际上，机器的最佳行为可能包括欺骗人类，让人类自以为给了机器一个合理的目标，这样机器就能获得足够的时间来实现其实际目标。这不是需要意识和自由意志的“反常”或“恶意”的行为，这只是实现目标的最佳计划的一部分。

在第1章中，我介绍了“有益的AI”的概念，即那些行动有望实现我们的目标，而不是它们的目标的人工智能。我在本章的目标是用简单的术语来解释如何做到这一点，尽管明显的缺点是机器不知道我们的目标是什么。由此产生的方法应该最终会创造出对我们没有威胁的人工智能，无论它们有多智能。

有益机器的原则

我觉得把这个方法总结成三条原则是很有帮助的。[3]在阅读这些原则时，请记住，它们主要用于指导人工智能研究人员和开发人员如何创建有益的人工智能系统。它们并非人工智能系统要遵循的明确定律：[4]

1. 机器的唯一目标是最大限度地实现人类的偏好。
2. 机器最初不确定这些偏好是什么。
3. 关于人类偏好的最终信息来源是人类行为。

在进行更详细的解释之前，重要的事情是要记住我在这些原则中所说的偏好的广泛范围。这里提醒一下我在第2章中所写的内容：如果你看了两部电影，每一部都足够详细和广泛地描述了你未来可能会过的生活，这样每一部都构成了一种虚拟体验，你可以说你喜欢哪一部，或者表示中立。因此，这里的偏好是包罗万象的，它们涵盖了在任意遥远的未来你可能关心的一切。[5]而且它们是你的偏好：机器并不是要识别出或采用一套理想的偏好，而是要理解并尽可能满足每个人的偏好。

第一原则：纯粹的利他主义机器

第一原则，即机器的唯一目标是最大限度地实现人类的偏好。这是有益机器这个概念的核心。特别是，它将会有益于人类，而不是蟑螂。我们无法回避这种受益人特有的益处。

这一原则意味着机器秉承纯粹的利他主义，也就是说，它绝对不为自己的幸福甚至自己的存在赋予内在价值。它可能会保护自己，以便继续为人类做有用的事情，或是因为它的主人会为不得不支付维修费而不高兴，或是因为看到一个肮脏或受损的机器人可能会让路人感到不悦，但不是因为它想活着。在机器人中加入任何自我保护的偏好，都会在机器人内部建立一个与人类幸福不完全一致的额外的激励机制。

第一原则提出了两个至关重要的问题。关于每一个问题，都值得写一整架的书，事实上，目前已经有许多关于这些问题的书出版了。

第一个问题是，人类是否真的拥有有意义的或稳定的偏好。事实上，“偏好”是一种理想化的概念，在许多方面都与现实不符。例如，我们并非生来就有成人的偏好，所以它们必须随着时间的推移而改变。现在，我假设这种理想化是合理的。稍后，我将研究当我们放弃理想化时会发生什么。

第二个问题是社会科学的一个主要问题：鉴于通常不可能确保每个人都得到他们最喜欢的结果，因为我们不可能全都成为宇宙之王，那么机器应该如何权衡多个人的偏好？同样，就目前而言，采用平等对待每个人这种简单方法似乎是合理的（我保证在下一章会回到这个问题）。这让人想起了18世纪效益主义的根源——“大多数人的最大幸福”[6]，而且要使之在实践中发挥作用，这里需要许多注意事项和详细说明。也许这些问题中最重要的是，可能有大量的人尚未出生，那么要如何考虑他们的偏好？

未来人类的问题引出了另一个相关的问题：我们如何考虑非人类实体的偏好？也就是说，第一原则是否应该考虑动物的偏好？（或许

还有植物？）这是一个值得讨论的问题，但其结果不太可能对人工智能的前进道路产生重大影响。就其价值而言，人类偏好可以而且确实包括动物福祉，以及人类可以直接从动物生存中受益的方面。[7]机器应该关注动物的喜好，除此之外，人类还应该建造比人类更关心动物的机器，这是一个很难维持的立场。一个更站得住脚的立场是，人类倾向于做出与我们自身利益相悖的短视决策，这往往会给环境及动物带来负面影响。更少做出短视决策的机器将有助于人类采取更环保的政策。而且，在未来，如果我们对动物的福祉给予比现在更大的权重——这可能意味着牺牲一些我们自己固有的福祉，那么机器也会适应。

第二原则：谦卑的机器

第二原则是，机器最初并不确定人类的偏好是什么。这是创造有益机器的关键。

一台自以为完全理解真正目标的机器会一门心思地追求这个目标。它永远不会问某个行动方案是否可行，因为它已经知道这是实现目标的最佳解决方案。它会无视人类心急火燎的尖叫：“停下来，你会毁了这个世界！”因为这只是语言而已。假设机器对目标的完美理解将其与人类分离开来，那么人类的所作所为将不再重要，因为机器理解目标并追求目标。

相反，一台对真实目标感到不确定的机器会表现出一种谦卑：例如，它会顺从人类，允许自己被关闭。它的理由是，只有当它做错了什么，即做了违背人类偏好的事情的时候，人类才会关掉它。根据第一原则，它想避免这样做，但是根据第二原则，它知道这是可能的，因为它不知道到底什么是“错误”。所以，如果人类真的关掉了机器，那么机器就能避免做错事情，这便是它的意图。换言之，机器有

积极的动机让自己被关闭。它仍然与人类联系在一起，人类是一个潜在的信息来源，可以让它避免错误，做得更好。

自20世纪80年代以来，不确定性一直是人工智能的核心问题。事实上，“现代人工智能”一词经常指的是，当不确定性最终成为现实世界决策中的一个普遍问题时发生的革命。然而，人工智能系统目标中的不确定性被简单地忽略了。在所有关于效用最大化、目标实现、成本最小化、奖励最大化和损失最小化的研究中，人们都假定效用函数、目标、成本函数、奖励函数和损失函数是完全已知的。这怎么可能？人工智能界（以及控制理论、运筹学和统计学界）怎么会在如此长时间内存在如此大的盲点，即便在决策中的所有其他方面都包含不确定性？[8]

人们可能会找出一些相当复杂的技术性借口，[9]但我认为的事实是，除了一些合理的例外，[10]人工智能研究人员只是简单地接受了将人类智能的概念映射到机器智能上的标准模型：人类有目标并追求目标，所以机器也应该有目标并追求目标。它们——或者应该说我们，从来没有真正检验过这个基本假设。它被内置于构建智能系统的所有现有方法中。

第三原则：学习预测人类偏好

第三原则，关于人类偏好的最终信息来源是人类行为。该原则有两个目的。

第一个目的是为“人类偏好”这个词提供一个明确的基础。根据假设，人类偏好并不存在于机器中，机器也无法直接观察到它们，但是在机器和人类偏好之间一定存在着某种明确的联系。这项原则说，这种联系是通过观察人类的选择来实现的：我们假设，选择与潜在的偏好之间存在某种（可能非常复杂的）联系。为了理解为什么这种联系是必不可少的，我们不妨反过来想想：如果人类的某种偏好对人类

可能做出的任何实际或假设的选择都没有任何影响，那么可能就不必说这种偏好存在了。

第二个目的是让机器变得更有用，因为它能更多地了解我们想要的东西。（毕竟，如果它对人类偏好一无所知，它对我们就毫无用处。）这个想法很简单：人类的选择揭示了关于人类偏好的信息。将其应用于选择菠萝比萨还是香肠比萨很简单，而将其应用于未来生活之间的选择和为了影响机器人行为而做出的选择则更加有趣。在下一章中，我将解释如何设置和解决这些问题。然而，真正复杂的情况出现了，因为人类不是完全理性的：人类的偏好和人类的选择之间存在着不完美的差异，如果机器要将人类的选择解释为人类偏好的证据，那么它就必须考虑到这些不完美。

澄清一些误解

在讨论更多细节之前，我想先消除一些潜在的误解。

首先，也是最常见的误解是，我提议在机器中安装一个由我设计的单一的、理想化的价值观系统来指导机器的行为。“你打算灌输谁的价值观？”“谁来决定价值观是什么？”甚至，“是什么赋予了像罗素这样的富裕的西方白人男性科学家决定机器如何编码和发展人类价值观的权利？”[11]

我认为这种混淆部分是由于“价值观”这个词的常识性含义与在经济学、人工智能和运筹学中使用的更技术性的含义之间的冲突。在通常的用法中，价值观是用来帮助解决道德困境的。而作为一个技术术语，价值观大体上是效用的同义词，它衡量从比萨到天堂的一切事物的可取程度。我想要的含义是技术性的：我只是想确保机器给我正确的比萨，而且不会意外地毁灭人类。（找到我的钥匙将会是一个意外的惊喜。）为了避免这种混淆，这些原则谈论的是人类偏好，而不是人类的价值观，因为前一个术语似乎避开了道德上的先入之见。

当然，“输入价值观”正是我所说的我们应该避免的错误，因为让价值观（或偏好）完全正确非常困难，而把它们弄错则可能是灾难性的。相反，我建议机器去学习如何更好地做预测，预测每个人更喜欢哪种生活，同时意识到其预测是高度不确定和不完整的。原则上，这台机器可以学习数十亿种不同的预测偏好模型，地球上数十亿人的偏好各不相同。对未来的人工智能系统而言，这个要求并不过分，因为现在的脸书系统已经在维护超过20亿人的个人资料了。

其次，一个相关的误解是，我们的目标是为机器配备“伦理”或“道德价值观”，使它们能够解决道德困境。通常，人们会提出所谓的“电车难题”，[12]即人们必须选择是否杀死一个人来拯救其他人，这被认为与无人驾驶汽车有关。然而，道德困境的全部意义在于，这是一种两难境地：双方都有充分的论据。人类的生存不是道德困境。机器可以用错误的方式解决大多数道德困境（不管是什么），但仍不会对人类造成灾难性的影响。[13]

最后，一种常见的推测是，遵循这三项原则的机器将吸收它们观察和学习的邪恶人类的所有罪恶。当然，我们中有许多人做出的选择不尽人意，但我们没有理由假设那些研究我们动机的机器会做出同样的选择，就像没有理由假设犯罪学家会变成罪犯一样。以腐败的政府官员为例，他为了索要贿赂而批准建筑许可，因为他微薄的工资无法支付孩子上大学的费用。观察这种行为的机器不会学会受贿。它将会了解到，这位官员像其他人一样，非常渴望他的孩子接受教育并获得成功。它会找到帮助他的方法，而不会损害他人的幸福。这并不是说所有的邪恶行为对机器来说都不是问题，例如，机器可能需要区别对待那些把快乐建立在他人痛苦之上的人。

乐观的理由

简而言之，我建议，如果我们想保持对日益智能的机器的控制，就需要将人工智能引向一个全新的方向。我们需要摆脱20世纪技术的驱动思想之一：优化给定目标的机器。我经常被问及，“鉴于人工智能和相关学科的标准模型背后的巨大推动力，为什么你甚至认为这是遥不可及的”。事实上，我很乐观地认为这是可以做到的。

第一个原因是，我们有强大的经济动机来开发遵从人类的意愿，并逐渐与用户的偏好和意图保持一致的人工智能系统。这样的系统是非常可取的：它们所能表现出的行为范围远远大于具有固定的、已知目标的机器。在适当的时候，它们会向人类提问或请求许可；它们会进行“试运行”，看看我们是否喜欢它们的提议；当它们做错事时，它们会接受纠正。而不能做到这一点的系统将会产生严重的后果。目前为止，人工智能系统的愚蠢和有限的行为范围使我们免受这些后果的影响，但这种情况将会改变。例如，想象一下，未来某个家用机器人负责在你工作到很晚的时候照顾你的孩子。孩子们饿了，但冰箱是空的。然后机器人注意到了猫。可惜，机器人只知道猫的营养价值，却不知道它的情感价值。短短几个小时内，有关发疯的机器人烤猫肉的新闻就铺天盖地地出现在全球媒体上，整个家用机器人行业全军覆没。

一个行业参与者可能会因为粗心的设计摧毁整个行业，这为形成以安全为导向的行业联盟和执行安全标准提供了强大的经济动机。“人工智能合作伙伴组织”（Partnership on AI）的成员几乎包括全球所有的领先科技公司，它们已经同意合作，以确保“人工智能研究和技术是鲁棒的、可靠的、值得信赖的，并在安全的约束下运行”。据我所知，所有的主要参与者都在公开文献中发表了它们以安全为导向的研究。因此，经济激励在我们达到人类水平的人工智能之前很久就开始运作了，并且只会随着时间的推移而增强。此外，在国际层面上也可能有同样的合作动力。[14]

第二个原因是，用于学习人类偏好的原始数据，即人类行为的例子非常丰富。这些数据不仅通过彼此共享着大约数十亿人的数据的亿万台机器的相机、键盘和触摸屏以直接观察的形式出现（当然，受隐私限制），而且还以间接的形式出现。最明显的间接证据是人类对书籍、电影、电视和广播的大量记录，这些记录几乎完全与人们的行为有关（有些人对此感到不安）。即使苏美尔人和埃及人最早的和最乏味的用铜锭交换大麦的记录，也让我们对人类对于不同商品的偏好有了一些了解。

当然，在解释这种原始资料时会遇到困难，这些原始资料包括政治宣传、谎言、疯子的胡言乱语，甚至政治家和总统的声明，但是机器当然没有理由把这些都当真。机器可以（也应该）将来自其他智能实体的所有交流解释为游戏中的动作，而不是事实陈述。在一些游戏中，例如一个人和一台机器的对战游戏，人类有诚实的动机，但在许多其他情况下，也有不诚实的动机。当然，无论诚实与否，人类都可能被自己的信念所欺骗。

还有第二类间接证据摆在我们面前，这就是我们创造世界的方式。[15]我们之所以这样做，非常粗略地讲，是因为我们喜欢这样做。（显然，这并不完美！）现在，想象你是一个访问地球的外星人，而此时全体人类都去度假了。当你走进他们的房屋时，你能否理解人类偏好的基础？地板上铺着地毯，是因为我们喜欢在柔软温暖的表面上行走，我们不喜欢响亮的脚步声；花瓶放在桌子中间而不是边缘，是因为我们不想让它们掉下来摔碎等。一切非自然安排的事物都为了解居住在这个星球上的奇怪的两足动物的好恶提供了线索。

谨慎的理由

如果你一直在关注无人驾驶汽车的进展，你可能会发现，“人工智能合作伙伴组织”在人工智能安全方面的合作承诺并不太让人放

心。无人驾驶汽车领域竞争激烈，理由很充分：第一家发布全自动驾驶汽车的汽车制造商将获得巨大的市场优势；这一优势将会自我强化，因为制造商能够更快地搜集更多数据，以改善系统的性能；如果另一家公司赶在优步之前推出全自动出租车，优步等打车公司将会很快破产。这导致了一场高风险的竞赛，在这场竞赛中，谨慎和仔细的工程设计似乎远不如时髦的演示、人才的争夺和仓促的推广更重要。

因此，攸关生死的经济竞争促使人们在安全方面走捷径，以期赢得竞赛。生物学家保罗·伯格（Paul Berg）参与组织了1975年的阿西洛马会议，这次会议导致人们暂停对人类进行基因改造，2008年，他在一篇关于那次会议的回顾文章中写道：[16]

> 阿西洛马会议给整个科学界都上了一课：对于新兴知识或早期技术引发的担忧最佳的应对方式是，让来自公共资助机构的科学家和广大民众共同寻找最佳的监管方式，越早越好。一旦公司的科学家开始主导研究事业，那就太晚了。

经济竞争不仅发生在公司之间，还发生在国家之间。最近，美国、中国、法国、英国和欧盟纷纷宣布对人工智能进行数十亿美元的国家投资，这无疑表明，任何一个大国都不想落后。2017年，俄罗斯总统弗拉基米尔·普京表示：“谁能成为人工智能领域的领导者，谁就将成为世界的统治者。”[17]这种分析基本上是正确的。正如我们在第3章中看到的，高级人工智能将会大大提高几乎所有领域的生产率和创新率。它的拥有者如果不将其分享出去，就能击败一切竞争国家或集团。

尼克·波斯特洛姆在他的书《超级智能》中对这种动机提出了警告。与企业竞争一样，国家竞争往往更注重原始能力的进步，而不是控制问题。然而，普京也许读过波斯特洛姆的书。他接着说：“如果有人赢得垄断地位，那将是非常不可取的。”同样，这也是毫无意义

的，因为人类级别的人工智能不是零和博弈，分享它不会产生任何损失。另外，在没有首先解决控制问题的情况下，关于谁能成为第一个实现人工智能的人的竞争，是一个负和游戏。每个人的回报都是负无穷。

人工智能研究人员影响人工智能全球政策演变的能力有限。我们可以指出可能带来经济和社会利益的应用；我们可以对可能的监视滥用和武器滥用等发出警告；我们可以为未来发展的可能路径及其影响提供路线图。也许我们能做的最重要的事情就是设计尽可能安全且对人类有益的人工智能系统。只有这样，我们尝试对人工智能进行全面监管才有意义。

---

[1] 如今已经略显过时的详细的大脑建模指南，请参见：Anders Sandberg and Nick Bostrom, “*Whole brain emulation: A roadmap*,” technical report 2008-3,Future of Humanity Institute, Oxford University, 2008。

[2] 关于代表人物对遗传编程的介绍，请参见：John Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection* (MIT Press,1992)。

[3] 与阿西莫夫的“机器人三定律”的相似之处纯属巧合。

[4] 同样的观点请见Eliezer Yudkowsky, “Coherent extrapolated volition,” technical report, Singularity Institute, 2004。尤德考斯基认为，直接将“四大道德原则就是我们需要编程的全部内容”纳入人工智能肯定会让人类走向毁灭之路。他关于“人类的连贯推断意志”的概念与第一原则具有同样的普遍意义；这个想法是，一个超级人工智能系统可以找出人类集体真正想要的东西。

[5] 对于是让机器帮助你实现偏好还是通过你自己的努力来实现偏好，你当然可以有自己的偏好。例如，假设在其他条件相同的情况下，你喜欢结果A胜过结果B。你无法在没有帮助的情况下实现结果A，但是你仍然宁愿实现B，也不愿意在机器的帮助下实现A。在这种情况下，机器应该决定不帮助你，除非它能以一种你完全察觉不到的方式帮助你。当然，你可能对无法察觉的帮助和可察觉的帮助有偏好。

[6] “大多数人的最大利益”一词的来源：Francis Hutcheson, *An Inquiry into the Original of Our Ideas of Beauty and Virtue, In Two Treatises* (D. Midwinter et al.,1725)。有些人把这种表述归因于威廉·莱布尼茨（Wilhelm Leibniz）早先的评论，请参见：Joachim Hruschka, “The greatest happiness

principle and other early German anticipations of utilitarian theory," *Utilitas* 3 (1991): 165–77。

[7] 有人可能会建议，机器应该在其自身的目标函数中包含动物和人类的术语。如果这些术语的权重对应于人们对动物的关心程度，那么最终结果将与机器仅通过关注关心动物的人类来关心动物一样。在机器的目标函数中赋予每种活着的动物同等的权重肯定是灾难性的，例如，南极磷虾的数量是我们的5万倍，细菌的数量是我们的10^{21}倍。

[8] 道德哲学家托比·奥德在他对本书早期书稿的评论中也向我提出了同样的观点："有趣的是，道德哲学的研究也是如此。直到最近，道德哲学几乎完全忽略了结果的道德价值的不确定性。尽管事实上，正是我们对道德问题的不确定性导致我们向他人寻求道德建议，甚至去研究道德哲学！"

[9] 不关注偏好的不确定性的一个理由是，它在形式上等同于普通的不确定性，其意义在于：不确定自己喜欢什么，等同于确定自己喜欢讨人喜欢的东西，却不确定什么东西是讨人喜欢的。这只是一个把戏，通过把"讨我喜欢"变成物体的属性，而不是我的属性，似乎把不确定性转移到了世界上。在博弈论中，这种把戏从20世纪60年代开始就被彻底制度化了，此前，我已故的同事、诺贝尔奖获得者约翰·豪尔绍尼发表了一系列论文："Games withincomplete information played by 'Bayesian' players, Parts I–III," *Management Science* 14 (1967, 1968): 159–82, 320–34, 486–502。在决策理论中，标准参考文献如下：Richard Cyert and Morris de Groot, "Adaptive utility," in*Expected Utility Hypotheses and the Allais Paradox*, ed. Maurice Allais and Ole Hagen (D.Reidel, 1979)。

[10] 在偏好诱导领域工作的人工智能研究人员是一个明显的例外。请参见：Craig Boutilier, "On the foundations of *expected* expected utility," in *Proceedings of the 18th International Joint Conference on Artificial Intelligen*ce (Morgan Kaufmann,2003). 以及Alan Fern et al., "A decision-theoretic model of assistance," *Journal of Artificial Intelligence Research* 50 (2014): 71–104。

[11] 一篇对有益人工智能的评论，基于对一名记者在一篇杂志文章中对作者的简短采访的误解：Adam Elkus, "How to be good: Why you can't teach human values to artificial intelligence," *Slate*, April 20, 2016。

[12] 电车难题的起源：Frank Sharp, "A study of the influence of custom on the moral judgment," Bulletin of the University of Wisconsin 236 (1908)。

[13] "节制生育"运动认为，人类繁衍后代在道德上是错误的，因为活着就是受苦，也因为人类对地球的影响是极其负面的。如果你认为人类的存在是一个道德困境，那么我想我确实希望机器以正确的方式解决这个道德困境。

[14] 中国全国人民代表大会外事委员会副主任委员傅莹就中国的人工智能政策发表声明。中华人民共和国主席习近平致信祝贺2018年世界人工智能大会开幕时说道："处理好人

工智能在法律、安全、就业、道德伦理和政府治理等方面提出的新课题，需要各国深化合作、共同探讨。”感谢谢旻希（Brian Tse）让我注意到这些声明。

[15] 一篇关于非自然主义非谬误的非常有趣的论文，展示了如何根据人类安排的世界状态推断偏好：Rohin Shah et al., “The implicit preference information in an initial state,” in *Proceedings of the 7th International Conference on Learning Representations* (2019), iclr.cc/Conferences/2019/Schedule。

[16] 关于阿西洛马会议的回顾：Paul Berg, “Asilomar 1975: DNA modification secured,” *Nature* 455 (2008): 290–91。

[17] 普京关于人工智能演讲的新闻报道：“Putin: Leader in artificial intelligence will rule world,” Associated Press, September 4, 2017。

08 可证明有益的人工智能



如果我们要沿着新的路线重建人工智能，那么它的基础必须是坚实的。当人类的未来岌岌可危时，光有希望和良好的意愿，以及教育计划、行业行为准则、立法和经济激励来做正确的事情是不够的。所有这些都有可能出错，而且经常失败。在这种情况下，我们希望通过精确的定义和一步步的严格数学证明来提供无可辩驳的保证。

这是一个好的开始，但我们需要更多。我们需要尽可能地确定被保证的东西是我们想要的，证明中的假设是正确的。这些证明本身应该出现在专家撰写的期刊论文中，但我认为，理解什么是证据，以及它们在确保真正的安全方面能提供什么和不能提供什么是很有用的。本章标题中的“可证明有益”是一种愿望，而非承诺，但这是正确的愿望。

数学保证

最终，我们希望证明一个定理：设计人工智能系统的一种特殊方式可以确保它们对人类有益。定理不过是断言的一个花哨的名称，它要表述得足够精确，以便在任何具体情况下都能检验其真实性。也许最著名的定理是费马大定理，它由法国数学家费马（Pierre de Fermat）在1637年提出，经过357年的努力，安德鲁·怀尔斯（Andrew Wiles）终于在1994年证明了该定理（并非全部由怀尔斯完成）。[1]这个定理可以用一行写完，但是证明它需要100多页密密麻麻的数学计算。

证明始于公理，公理是假定为真的断言。公理通常只是定义，例如费马大定理所需的整数、加法和求幂的定义。证明从公理开始，通过逻辑上无可争议的步骤，添加新的断言，直到定理本身作为其中一个步骤的结果而被确定。

这里有一个相当明显的定理，它几乎可以立即从整数和加法的定义中得出：1 + 2 = 2 + 1。我们不妨把它称为“罗素定理”。这算不上什么重大发现。相反，费马大定理感觉像是一个全新的东西—— 一个以前未知的发现。然而，这种差异只是程度的问题。罗素定理和费马大定理

揭示的真理都已经包含在公理中。证明不过是使隐含的东西变得明确罢了。证明的过程可长可短，但是它们没有增加任何新东西。定理的好坏取决于它所包含的假设。

就数学而言，这很好，因为数学是我们定义的抽象对象——数字、集合等。公理是真实的，因为我们是这样认为的。然而，如果你想要证明一些关于现实世界的事情，例如，像这样设计的人工智能系统不会故意杀死你，那么你的公理必须在现实世界中是正确的。如果它们不正确，那么你就证明了一个假想世界中的东西。

科学和工程学拥有证明假想世界结果的悠久而光荣的传统。例如，在结构工程中，人们可能会看到这样的数学分析：“假设AB是一根刚性梁……”刚性这个词在这里并不是指“由像钢铁这样坚硬的东西制成”，它的意思是“无限坚固”，所以它完全不会弯曲。刚性梁并不存在，所以这是一个假想的世界。关键是要知道它偏离现实世界的程度，那就仍能获得有用的结果。例如，如果“刚性梁假设”允许工程师计算出这个结构中的力，并且这些力足够小，只能使真正的梁发生很小程度的弯曲，那么工程师可以合理地相信，把想象世界中的分析转移到现实世界中也会成立。

优秀的工程师会意识到这种转移何时会失效，例如，如果梁处于压缩状态，其两端都受到巨大的力的挤压，那么即使很小的弯曲也会导致更大的侧向力，从而导致更大的弯曲，以此类推，最终导致灾难性的失败。在这种情况下，要用“假设AB是一根刚度为K的柔性梁……”来重新进行分析。当然，这仍然是一个虚构的世界，因为真实的梁并不具有均匀的刚度；相反，它们有微小的缺陷，如果梁反复弯曲，这些缺陷会使裂缝产生。移除不现实的假设的过程会持续下去，直到工程师非常确信剩余的假设在现实世界中足够站得住脚。之后，工程系统可以在现实世界中进行测试，但测试结果仅限于此。它们不能证明相同的系统在其他情况下仍然有效，也不能证明系统的其他实例会按照与原始系统相同的方式运行。

计算机科学中一个经典的假设失败的例子来自网络安全。在该领域，人们需要进行大量的数学分析来证明某些数字协议的安全性，例如，当你在Web应用程序中输入密码时，你希望确保它在传输前被加密，这样窃听网络的人就无法读取你的密码。这种数字系统通常可以被证明是安全的，但在现实中仍然容易受到攻击。这里的错误假设是，这是一个数字过程。实则不然，它在真实的物理世界中运行。通过监听键盘的声音或测量为笔记本电脑供电的电线的电压，攻击者可以“听到”你的密码或观察到处理过程中发生的加密与解密计算。网络安全界现在正在应对这些所谓的侧信道攻击，例如，编写一种加密代码，不管加密的是什么消息，代码都会产生相同的电压波动。

让我们看看我们最终要证明的有益机器定理。一种可能的类型如下所示。

> 假设一台机器具有组件A、B、C，它们彼此以某种方式连接在一起，并同环境以某种方式连接在一起，且使用内部学习算法l_A、l_B、l_C来优化内部反馈奖励r_A、r_B、r_C，以及一些其他条件……那么很有可能，机器的行为在（对人类的）价值上将非常接近在任何具有相同计算能力和物理能力的机器上可以实现的最佳可能行为。

这里的要点是，无论组件变得多么智能，这样的定理都应该成立，也就是说，证明无懈可击，机器永远对人类有益。

关于这类定理，还有另外三点值得一提。首先，我们不能试图证明机器代表我们产生最优（或接近最优）的行为，因为我们几乎可以肯定，这在计算上是不可能的。例如，我们可能想让机器完美地下围棋，但我们有充分的理由相信，在任何实际时间内，这在任何物理上可实现的机器中都不可能完成。现实世界中的最优行为甚至更不可行。因此，该定理说的是“最佳可能”而不是“最优”。

其次，我们说“很有可能……非常接近”，因为这通常是机器学习所能做到的最好的结果。例如，如果机器正在学习帮我们玩轮盘赌，球

连续40次落在“0”的位置上，机器可能会合理地判断赌桌被操纵了，并相应地下注。但这可能是偶然发生的，因此，机器总是有很小的——也许是微乎其微的被反常事件误导的可能性。最后，对于在现实世界中运行的真实的智能机器，我们距离证明这类定理还有很长的路要走！

人工智能中也有类似的侧信道攻击。例如，该定理以“假设一台机器具有组件A、B、C，它们彼此以某种方式连接在一起……”开头，这是计算机科学中所有正确定理的典型特征，它们从描述正确的程序开始。在人工智能中，我们通常会区分智能体（做决策的程序）和环境（智能体执行操作的地方）。既然我们设计了智能体，那么以下假设似乎是合理的——它具有我们所赋予的结构。为了更加安全，我们可以证明，它的学习过程只能以不会引起问题的限定方式修改它的程序。这样就够了吗？不够。与侧信道攻击一样，认为程序在数字系统中运行的假设是不正确的。虽然学习算法在本质上不能通过数字手段重写自己的代码，但它仍可能学会说服人类对其进行“脑部手术”，以此来违反智能体和环境的区分，并通过物理手段改变代码。[2]

与结构工程师对刚性梁的推理不同，我们对那些最终将成为“可证明有益的人工智能”（Provably Beneficial AI）定理基础的假设知之甚少。例如，在本章中，我们通常会假设一个理性的人。这有点像假设一根刚性梁，因为在现实中没有完全理性的人。（然而，情况可能更糟，因为人类甚至还没有接近理性。）我们所能证明的定理似乎提供了一些见解，这些见解在人类行为中引入某种程度的随机性之后仍然成立，但当我们考虑到真实人类的一些复杂性时，我们还远远不清楚会发生什么。

因此，我们必须非常小心地检查我们的假设。当安全性被成功证明时，我们需要明白它不会真的成为现实，因为我们做出了不切实际的强假设，或是因为安全的定义太弱。当安全性被证明失败时，我们需要抵制诱惑，不去增强假设以使证明通过，例如添加程序代码保持不变的假设。相反，我们需要加强人工智能系统的设计，例如确保它没有修改自己代码的关键部分的动机。

有一些假设，我称之为“OWMAWGH”假设，意思是“否则我们还不如回家”（otherwise we might as well go home）。也就是说，如果这些假设是错误的，那么游戏就结束了，我们没有什么可做的了。例如，我们可以合理地假设宇宙是按照恒定的和某种程度上可辨别的规律运行的。如果不是这样，我们就无法保证学习过程会起作用，即使是非常复杂的过程也不例外。另一个基本假设是，人类关心发生了什么。如果人类什么也不关心，那么可证明有益的人工智能就没有目的，因为“有益”没有意义。在这里，关心意味着对未来有大致一致和相对稳定的偏好。在下一章中，我将探讨人类偏好的“可塑性”可能带来的后果，这对可证明有益的人工智能的想法提出了严峻的哲学挑战。

目前，我关注最简单的情况：只有一个人和一个机器人的世界。我用这个案例介绍基本思想，但它本身也很有用：你可以将这个人视为全人类的代表，将这个机器人视为所有机器的代表。当考虑多个人和多个机器人时，额外的复杂性就会出现。

从行为中学习偏好

经济学家通过为人类受试者提供选择来套取他们的偏好。[3]该技术广泛应用于产品设计、营销和交互式电子商务系统中。例如，汽车设计师向受测试者提供具有不同油漆颜色、座位安排、后备厢大小、电池容量、杯架等选项的汽车，以此来了解人们关心哪些汽车功能，以及他们愿意为这些功能支付多少钱。另一个重要应用是在医学领域，肿瘤学家在考虑截肢的可能性时，可能需要评估病人在行动能力和预期寿命之间的偏好。当然，比萨餐厅想知道人们愿意为香肠比萨支付的价格比普通比萨高多少。

套取偏好通常只考虑在多个对象之间做出的单一选择，我们假设这些对象的价值对受试者而言是显而易见的。我们目前还不清楚如何将其拓展到对未来生活的偏好上。为此，我们（和机器）需要从长期的行为观察中学习，这涉及具有多种选择和不确定结果的行为。

1997年初，我和同事迈克尔·迪金森（Michael Dickinson）、鲍勃·弗尔（Bob Full）讨论了我们如何应用机器学习的思想来理解动物的运动行为。迈克尔仔细研究了果蝇翅膀的运动。而鲍勃特别喜欢令人毛骨悚然的爬虫，他为蟑螂制作了一台小型跑步机，以便观察它们的步态如何随着速度的变化而变化。我们认为，利用强化学习来训练机器人或模拟昆虫，以此重现这些复杂的行为是可能的。我们面临的问题是，我们不知道使用什么奖励信号，不知道苍蝇和蟑螂在优化什么。没有这些信息，我们就无法应用强化学习来训练虚拟昆虫，所以我们陷入了困境。

一天，我从我们在伯克利的房子去当地超市。这条路有一个下坡，我注意到，这个斜坡使我走路的方式发生了轻微的变化，我相信大多数人都会这样。此外，几十年中发生的数次小地震造成了路面不平坦，所以走在上面的人会发生额外的步态变化，由于地面高度无法预测，我的脚抬得更高了一点，下落时也不那么生硬。当我思考这些平凡的生活观察时，我意识到我们弄错了。当强化学习从奖励中产生行为时，我们实际上想要的恰恰相反：学习给定行为的奖励。我们已经有了由苍蝇和蟑螂产生的行为，我们想知道这种行为所优化的具体奖励信号。换言之，我们需要逆强化学习算法（IRL）。[4] [当时我还不知道，还有一个类似的问题曾被研究过，它被称为“马尔科夫决策过程的结构式估计方法”，这是诺贝尔奖得主汤姆·萨金特（Tom Sargent）在20世纪70年代末开创的一个领域。[5]] 这样的算法不仅可以解释动物的行为，还可以预测它们在新环境下的行为。例如，蟑螂如何在既倾斜又颠簸的跑步机上奔跑？

解决这些基本问题令人们兴奋得不能自已，但即便如此，人们为逆强化学习设计出第一套算法还是花了一些时间。[6]自那时以来，人们已经提出了许多不同的逆强化学习公式和算法。从某种意义上讲，这些算法可以获得足够多的关于实体偏好的信息，使之能够像观察到的实体一样成功地行动，因此可以从某种程度上保证算法的有效性。[7]

或许，理解逆强化学习算法的最简单方法是：观察者从对真实的奖励函数的一些模糊估计开始，然后随着观察到更多的行为而细化这个估计，使之更精确。或者，用贝叶斯的理论来解释：[8]从可能的奖励函数的先验概率开始，然后随着证据的增加来更新奖励函数的概率分布（附录C）。例如，假设机器人罗比正在观察人类哈里特，想知道在靠窗的座位和靠过道的座位之中，她更喜欢哪一个。起初，它对此很不确定。从概念上讲，罗比的推理可能是：“如果哈里特真的喜欢靠过道的座位，她会查看座位图，了解是否有靠过道的空位，而不是仅仅等着航空公司给她安排靠窗的座位，但她没有，尽管她可能注意到这是一个靠窗的座位，而且她也并不着急。所以现在，她很可能要么对靠窗和靠过道漠不关心，要么更喜欢靠窗的座位。”

逆强化学习在实践中最引人注目的例子是我的同事彼得·阿比尔（Pieter Abbeel）在学习直升机特技飞行方面的工作。[9]专业的人类飞行员可以让直升机模型做出令人惊叹的特技，例如回环、螺旋、摆荡等。试图复制人类行为的结果并不太好，因为条件无法完全再现：在不同情况下重复相同的控制序列可能会导致灾难。相反，该算法以它能够实现的轨迹约束的形式，来学习人类飞行员想要的东西。这种方法实际上产生的结果甚至比人类专家所做的还要好，因为人类的反应更慢，并且会不断地犯小错误并纠正错误。

辅助博弈

逆强化学习如今已经是构建有效的人工智能系统的重要工具，但它做了一些简化的假设。第一，机器人一旦通过观察人类学会了奖励函数，它就会采用奖励函数，这样它就可以执行相同的任务。这对驾驶汽车或驾驶直升机来说没问题，但对于喝咖啡不行：一个观察我早晨习惯的机器人应该知道我（有时）想喝咖啡，但不应该知道它自己想喝咖啡。解决这个问题很容易，我们只需确保机器人将偏好与人类联系起来，而不是与自身联系起来。

逆强化学习中的第二个简化假设是，机器人正在观察一个人类解决单智能体决策问题。例如，假设机器人在医学院，通过观察人类专家来学习成为外科医生。逆强化学习算法假设人类通常以最佳方式进行手术，就好像机器人不在那里一样。但事实并非如此：人类外科医生的动机是让机器人（像其他医科学生一样）学得又快又好，这样她就会大大改变自己的行为。她可以边走边解释她在做什么；她可以指出需要避免的错误，比如切口太深或伤口缝合太紧；她可以描述万一手术中出现问题，应当采取什么应急方案。在单独进行手术时，这些行为都没有意义，因此逆强化学习算法将无法解释这些行为所暗示的偏好。出于这个原因，我们需要将逆强化学习从单智能体设置发展到多智能体设置，也就是说，我们需要设计一种学习算法，当人和机器人是同一环境的一部分并且相互交互时，该算法必须发挥作用。

当一个人和一个机器人处于同一环境中的时候，我们就进入了博弈论的领域，就像第1章中爱丽丝和鲍勃之间的点球大战一样。在这个理论的第一个版本中，我们假设人类有偏好，并根据这些偏好行事。机器人不知道人类有什么偏好，但它无论如何都想满足他们。我们称这种情况为“辅助博弈”（assistance game），因为根据定义，机器人应该对人类有所帮助。[10]

辅助博弈实例化了上一章中的三项原则：机器人的唯一目标是满足人类的偏好，它最初并不知道人类的偏好是什么，以及它可以通过观察人类的行为来学习更多。也许辅助博弈最有趣的特性是，通过解决博弈问题，机器人可以自己弄明白如何将人类行为转化为有关人类偏好的信息。

“回形针”博弈

辅助博弈的第一个例子是回形针博弈。这是一个非常简单的游戏，在这个游戏中，人类哈里特有一种动机来向机器人罗比“发送”一些她的偏好信息。罗比能够解读这个信号，因为它能玩这个游戏，由此它能理解哈里特有什么样的偏好，以便让她发出那样的信号。

游戏的步骤如图12所示，涉及制作回形针和订书钉。哈里特的偏好是通过一个收益函数来表达的，该函数取决于生产的回形针和订书钉的数量，两者之间有一定的“汇率”。例如，她可能会把回形针的价值定为45美分，将订书钉的价值定为55美分。（我们将假设这两个价值的总和始终为1.00美元，重要的只有比率。）因此，如果生产10个回形针和20个订书钉，哈里特的收益将是10×45美分+20×55美分= 15.50美元。机器人罗比一开始完全不确定哈里特的偏好：他对回形针的价值有一个均匀分布（也就是说，价值有可能是从0美分到1.00美元的任何值）。哈里特先做了选择，她可以选择制作2个回形针，或2个订书钉，或每种各1个。然后，罗比选择制作90个回形针，或90个订书钉，或每种各50个。[11]

请注意，如果哈里特自己做这件事，做了2个订书钉，价值为1.10美元。但是罗比在看着，它从她的选择中学习。它到底学到了什么？这取决于哈里特如何做出选择。哈里特是如何做出选择的？这取决于罗比如何解释它。所以，我们似乎遇到了一个循环问题！这在博弈论问题中很典型，也是纳什提出纳什平衡的原因。

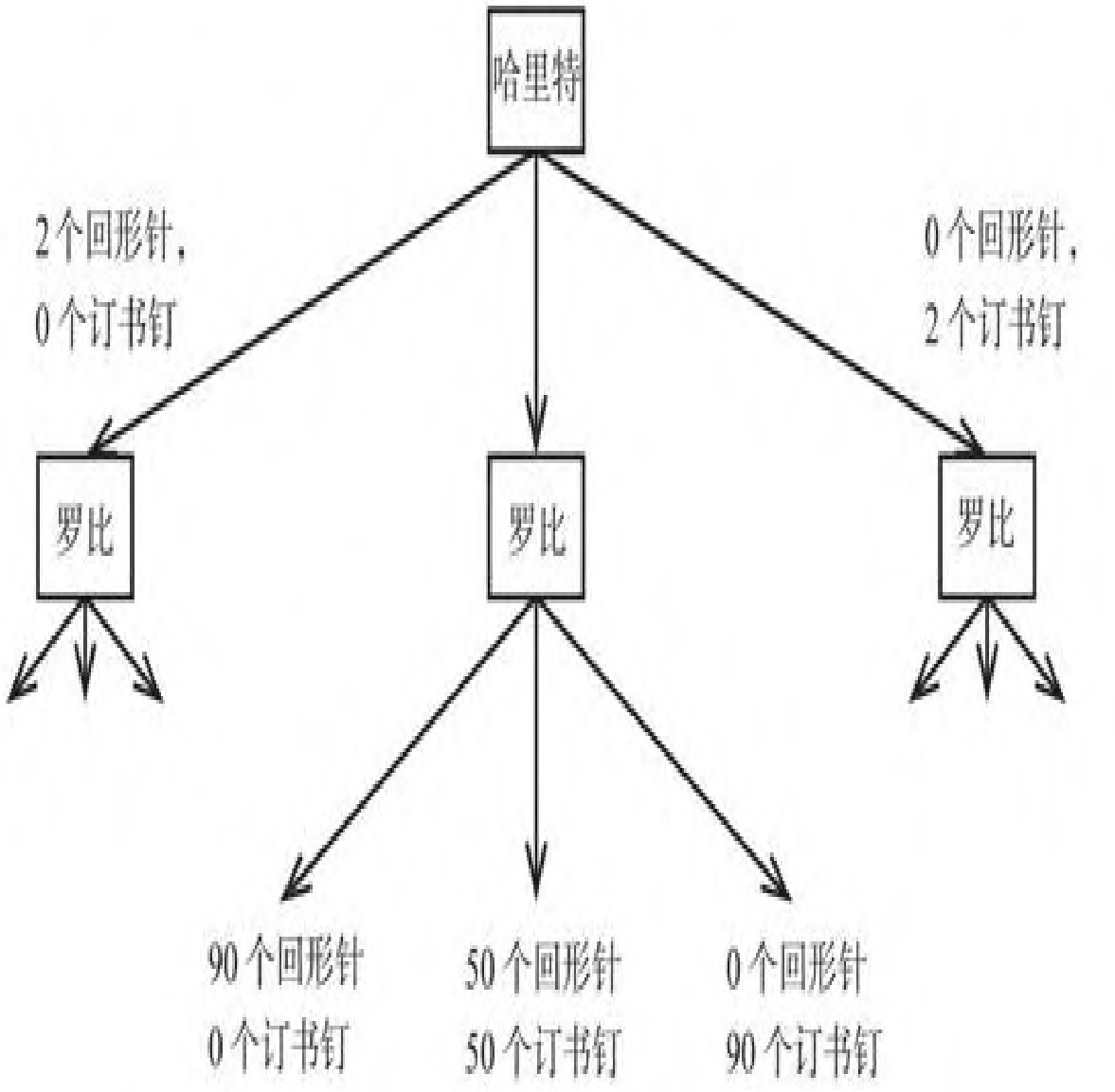


图12 回形针博弈。人类哈里特可以选择制作2个回形针，或2个订书钉，或每种各1个。然后，机器人罗比可以选择制作90个回形针，或90个订书钉，或者每种各50个

为了找到一个均衡的解决方案，我们需要为哈里特和罗比确定策略，假设任意一方的策略保持不变，则双方都没有改变自己策略的动机。哈里特的一个策略是，根据她的偏好，指定要做多少回形针和订书钉；罗比的一个策略是，根据哈里特的行动，指定要做多少回形针和订书钉。

事实证明，似乎只有一个均衡解。

- 哈里特根据她对回形针的估价做出如下决定：

如果价值低于44.6美分，我就制作0个回形针和2个订书钉；

如果价值在44.6 ~ 55.4美分之间，我就每种各制作1个；

如果价值大于55.4美分，我就制作2个回形针和0个订书钉。

- 罗比回应如下：

如果哈里特制作0个回形针和2个订书钉，我就制作90个订书钉；

如果哈里特每种各制作1个，我就每种各制作50个；

如果哈里特制作2个回形针和0个订书钉，我就只制作90个回形针。

如果你想知道如何得出具体的解决方案，请参见注释。[12]通过这种策略，哈里特实际上是在用一种简单的代码（如果你喜欢，也可以说是一种语言）告诉罗比她的偏好，这种简单的代码是从均衡分析中产生的。就像外科医生教学的例子一样，单智能体逆强化学习算法无法理解这段代码。还要注意，罗比从来没有确切地了解哈里特的偏好，但是它学到了足够多的东西来代表她采取最佳行动，也就是说，它的行为就像确实知道哈里特的偏好一样。在陈述的假设下和哈里特在正确玩游戏的假设下，我们可以证明罗比对哈里特是有益的。

人们也可以构建问题，罗比会像一个好学生一样问问题，而哈里特会像一个好老师一样告诉罗比要避免的陷阱。之所以会出现这些行为，并不是因为我们编写了让罗比听从哈里特的脚本，而是因为这是哈里特和罗比参与的辅助博弈的最佳解决方案。

关机博弈

工具性目标通常可以用作几乎任何原始目标的子目标。自我保护是一个工具性目标，因为很少有原始目标在死后能更好地实现。这导致了关机问题：具有固定目标的机器将不允许自己被关机，并有动机禁用自己的关机按钮。

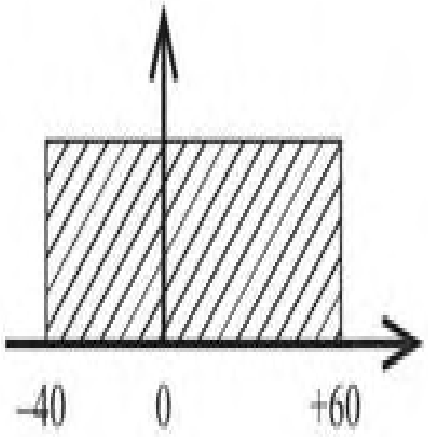
关机问题实际上是智能系统控制问题的核心。如果我们因为机器不允许我们关机就无法关闭一台机器，那么我们就真的有麻烦了。如果我们可以关闭一台机器，那么我们也可以通过其他方式来控制它。

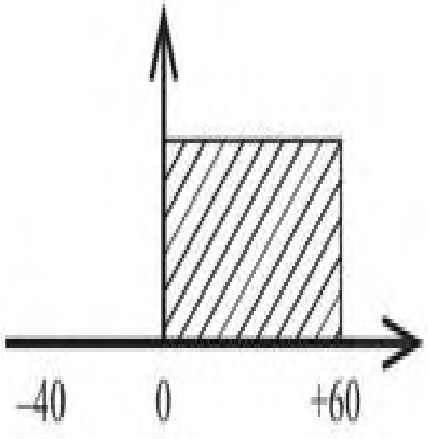
事实证明，目标的不确定性对于确保我们能够关闭机器至关重要，即便机器比我们更智能。我们在前一章看到了非正式的论证：根据有益机器的第一原则，罗比只关心哈里特的偏好，但是根据第二原则，罗比不确定哈里特的偏好是什么。它知道自己不想做错事，但它不知道做错事意味着什么。另外，哈里特知道做错事意味着什么。（或者说，在这个简单的例子中，我们假设是这样。）因此，如果她关掉罗比是为了避免它做错事，那么罗比会很愿意被关掉。

为了使这个论点更精确，我们需要这个问题的正式模型。[13]我要使它“尽可能简单”，而不只是“简单一些”而已（图13）。

罗比现在是哈里特的私人助理，可以优先进行选择。它现在可以行动了，比如说为哈里特预订一家昂贵的酒店。它非常不确定哈里特有多喜欢这家酒店及其价格，但它为哈里特感受到的价值设计了均匀概率分布，价值介于－40和+60之间，平均为+10。它可以“关闭自己”，即退出酒店预订过程，我们定义这个过程对哈里特的价值为0。它也可以继续预订酒店，而这会招致让哈里特不高兴的重大风险。（如果范围是-60到+40，平均为-10，那么它会关闭自己。）然而，我们会给罗比第三个选择：解释它的计划，等待哈里特做出选择。哈里特既可以把它关掉，也可以让它继续预订酒店。你可能会问，鉴于它可以自己做出这两个选择，等待又会有什么好处呢？




-40
0
+60


-40
0
+60

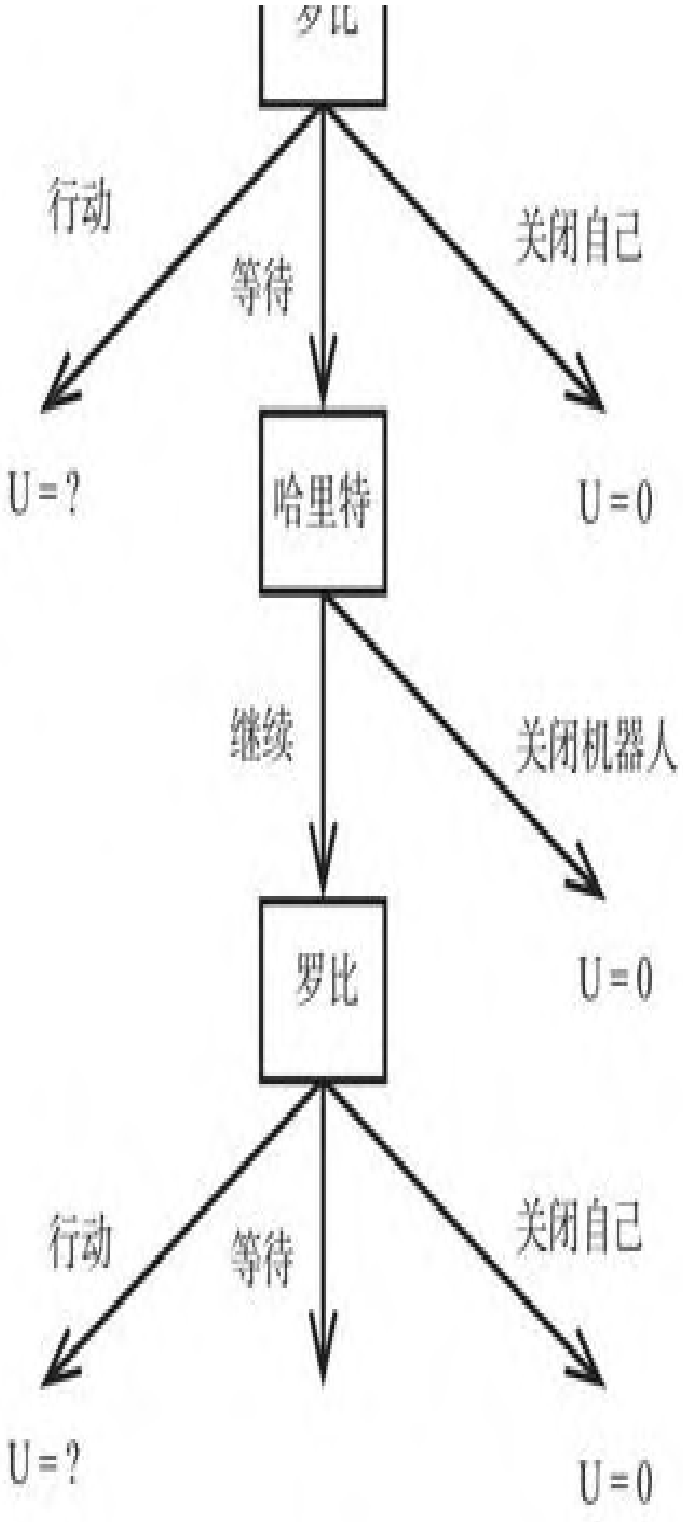

罗比
行动
等待
关闭自己
U=?
哈里特
U=0
继续
关闭机器人
罗比
U=0
行动
等待
关闭自己
U=?
U=0

图13 关机博弈。罗比可以选择现在行动，但收益非常不确定。也可以选择自己关机，或者等待哈里特做决定。哈里特可以关掉罗比，也可以让它继续。罗比现在又有了同样的选择。对哈里特来说，行动仍然有不确定的回报，但现在罗比知道回报不是负数

关键在于，哈里特的选择——是关掉罗比还是让它继续，为罗比提供了关于哈里特偏好的新信息。如果哈里特让罗比继续，那是因为对哈里特来说价值是正数。现在罗比的信念在0到+60之间的均匀分布，平均为+30。

所以，如果我们从罗比的角度来评估它最初的选择：

- 立即行动并预订酒店的期望价值为+10。
- 关闭自己的价值为0。
- 等待并让哈里特关掉它（如果她愿意），会导致两种可能的结果：

（1）有40%的可能性（基于罗比对预订酒店计划的不确定性），哈里特不喜欢这间酒店，并关闭罗比，价值为0。

（2）有60%的可能性，哈里特会喜欢这间酒店，并允许罗比继续，期望价值为+30。

- 因此，等待的期望价值是40%×0 + 60%×30 = +18，这比在+10时立即采取行动要好。

结果是，罗比有积极的动机让自己被关闭。这种动机直接来自罗比对哈里特偏好的不确定。罗比意识到它有一定的概率（在这个例子中是40%）会做出一些让哈里特不高兴的事情，在这种情况下，被关掉比继续下去更好。如果罗比已经确定了哈里特的偏好，它就会继续做出决定（或者关闭自己）。此时咨询哈里特绝对不会有任何收获，因为根据罗比明确的信念，它已经可以准确地预测哈里特将要做出什么决定。

事实上，在一般情况下也可以证明同样的结果：只要罗比不能完全确定自己将要做的事情就是哈里特自己会做的事，它就会更愿意被哈里

特关闭。[14]哈里特的决定为罗比提供了信息，而信息对于改进罗比的决定总是有用的。相反，如果罗比对哈里特的决定有把握，那么她的决定就不会提供新的信息，所以罗比没有动机让她做决定。

该模型中有一些明显的细节值得立即探讨。第一个细节是要求哈里特做决定或回答问题的成本。（也就是说，哈里特的时间是宝贵的，我们假设罗比对哈里特的偏好至少知道一些。）在这种情况下，如果罗比几乎确定了哈里特的偏好，它就不太愿意去打扰她。成本越高，罗比在打扰哈里特之前就越不确定。这是应该的。如果哈里特会因为被打扰而生气，那么即使罗比偶尔会做她不喜欢的事情，她也不应该感到太惊讶。

第二个细节是允许有人为错误发生的可能性，也就是说，即使罗比提议的行动是合理的，哈里特有时也可能会关掉罗比；即使罗比提议的行动并不可取，哈里特有时也会让罗比继续。我们可以把这种人为错误的概率放入辅助博弈的数学模型中，并像前面一样找到解决方案。正如人们所料，博弈的解决方案表明，罗比不太倾向于听从一个有时违背自己最大利益的、非理性的哈里特。她的行为越随意，罗比在服从她之前就越不确定她的偏好。同样，这也是理所当然的，例如，如果罗比是一辆无人驾驶汽车，哈里特只是坐在车里的顽皮的两岁孩子，罗比就不应该允许哈里特在高速公路上把它关掉。

我们还有很多方法来完善模型或将模型嵌入复杂的决策问题中。[15]然而我相信，核心理念——有益的、顺从的行为和机器对人类偏好的不确定性之间的重要联系，会经受住这些细化和复杂化的考验。

长期准确地学习偏好

在阅读关机博弈这部分时，你可能会想到一个重要的问题。（事实上，你可能有很多重要的问题，但我只回答这个问题。）当罗比获得越来越多的关于哈里特偏好的信息，变得越来越确定时，会发生什么？这

是否意味着它最终将不再完全听命于她？这是一个棘手的问题，可能的答案有两个：第一个是“是”，第二个还是“是”。

第一个“是”是良性的。一般而言，只要罗比最初对哈里特偏好的信念把一些概率（无论多么小）归因于她实际拥有的偏好，那么随着罗比变得越来越确定，它会变得越来越正确。也就是说，它最终会确定哈里特有她实际上有的偏好。例如，如果哈里特把回形针的价值定在12美分，订书钉的价值定在88美分，罗比最终会学习到这些价值。在这种情况下，哈里特并不在乎罗比是否顺从她，因为她知道罗比总是会像她一样做事。哈里特想要关掉罗比的情景永远不会出现。

第二个“是”则不太妙。如果罗比先验地排除了哈里特的真正偏好，那么它永远也不会学习到那些真正的偏好，但它的信念可能会聚成一个不正确的评估。换言之，随着时间的推移，它变得越来越确信关于哈里特偏好的错误信念。通常，在罗比最初认为可能的所有假设中，这个错误的信念将是最接近哈里特真实偏好的假设。例如，如果罗比非常确定哈里特的回形针价值在25至75美分之间，而哈里特认为的真实价值是12美分，那么罗比最终会确定哈里特将回形针的价值定在25美分。[16]

随着罗比对哈里特的偏好越来越确定，它将和那些有固定目标的糟糕的旧人工智能系统越来越像：它不会征求许可，也不会让哈里特选择关闭它，而且它的目标不正确。如果只是在回形针和订书钉之间做选择，这并不可怕，但是如果哈里特病得很重，那么这可能是在生活质量和寿命之间做选择。如果罗比代表人类行事，那么这可能是在人口规模和资源消耗之间做选择。

因此，如果罗比预先排除了哈里特实际上可能有的偏好，我们就会遇到一个问题：它可能会对她的偏好有一种明确但不正确的信念。解决这个问题的办法显而易见：不要这样做！相反，我们要为逻辑上可能的偏好分配一些概率，不管其概率有多小。例如，从逻辑上讲，哈里特可能会主动想把订书钉处理掉，然后付钱让你把它们拿走。（也许她还是个孩子时把手指钉在桌子上了，现在她非常讨厌订书钉。）因此，我们

应该考虑到“汇率”为负的可能性，这让事情变得有点复杂，但仍然完全可控。[17]

但是如果哈里特把回形针价值定为工作日12美分、周末80美分，又会如何？这个新的偏好无法用任何单一数字来描述，因此罗比实际上已经提前排除了这种可能性。这并不在它关于哈里特偏好的一系列可能的假设中。总而言之，除了回形针和订书钉之外，哈里特关心的东西可能还有很多很多。的确如此！例如，假设哈里特关心气候，并假设罗比最初的信念包含了一系列可能的考虑，包括海平面、全球气温、降雨、飓风、臭氧、入侵物种和森林砍伐。然后罗比会观察哈里特的行为和选择，并逐渐完善对她的偏好的理论，以了解她对清单上每一项分配的权重。但是，就像在回形针的案例中一样，罗比不会去了解清单上没有的东西。比如说，哈里特还关心天空的颜色，我敢保证，你不会在气候学家关注的典型事件中找到这种东西。如果罗比能通过把天空变成橙色来优化在海平面、全球温度、降雨量等方面做得稍微好一点，它会毫不犹豫地这样做。

同样，这个问题仍然有一个解决方案：不要这样做！永远不要预先排除世界上可能成为哈里特偏好的一部分的属性。这听起来不错，但实际上，让它在实践中发挥作用要比根据哈里特的喜好处理一个数字更加困难。罗比最初的不确定性必须考虑到可能有助于了解哈里特偏好的无数未知属性。然后，当罗比根据已知的属性无法解释哈里特的决定时，罗比可以推断，有一个或多个先前未知的属性（例如天空的颜色）可能在起作用，而且它可以试着找出那些属性可能是什么。通过这种方式，罗比避免了由过度限制性的先验信念引起的问题。据我所知，目前尚无此类方案的成功示例，但其总体思想包含在当前对机器学习的思考中。[18]

禁例与漏洞原则

保有人类目标的不确定性，或许并不是说服机器人在拿咖啡时不要禁用关机按钮的唯一方法。杰出的逻辑学家摩西·瓦迪（Moshe Vardi）提出了一个基于禁例的更简单的解决方案：[19]与其给机器人一个目标“去拿咖啡”，不如给它一个目标“去拿咖啡，同时不要让你的关机按钮失效”。不幸的是，有这样目标的机器人将会在符合法律条文的同时违背法律精神，例如，用有食人鱼出没的护城河把关机按钮包围起来，或是杀死任何靠近关机按钮的人。以一种万无一失的方式编写这样的禁例就像试图编写无漏洞的税法，这是我们几千年来一直在尝试却一直失败的事情。一个足够聪明的、有强烈动机避税的实体很可能会找到一种方法来做到这一点。我们称之为漏洞原理：如果一台足够智能的机器有动机创造某种条件，那么一般而言，人类无法仅凭写禁例来限制它的行为，阻止它这样做，或是阻止它做一些等效的事情。

防止有人避税的最佳解决方案是确保相关实体都愿意纳税。在人工智能系统可能出现问题的情况下，最好的解决方案是确保它愿意服从人类。

要求和指示

到目前为止，这个故事的寓意是，我们应该避免诺伯特·维纳所说的“给机器输入一个目标”。但是假设机器人确实收到了人类的直接指令，比如，“给我拿杯咖啡！”机器人应该如何理解这个指令？

按照惯例，这将成为机器人的目标。任何满足目标的动作序列，即让人类喝上咖啡的一系列行动，都算是解决方案。通常，机器人还会有一种方法来对解决方案进行排序，可能是基于所花费的时间、行进的距离，以及咖啡的成本和质量。

这是一种非常拘泥于字面意思的解释指令的方式。它会导致机器人的病态行为。例如，也许人类哈里特在沙漠中的一个加油站停了下来，她派机器人罗比去取咖啡，但是加油站没有咖啡，于是罗比以每小时3英

里的速度赶到了200英里外的最近的城镇，10天后带着一杯所剩无几的咖啡返回。与此同时，哈里特耐心地等待着，加油站老板给她提供了冰茶和可口可乐。

如果罗比是人类（或是一个设计精良的机器人），它就不会像这样从字面上理解哈里特的指令了。指令不是不惜一切代价都要实现的目标。这是一种传达哈里特偏好信息的方式，目的是诱导罗比进行某些行为。问题是，传达什么信息？

一种建议是传达这样一个信息：在其他所有条件相同的情况下，哈里特喜欢喝到咖啡甚于喝不到咖啡。[20]这意味着，如果罗比有一种方法可以在不改变世界的情况下拿到咖啡，那么即使它不知道哈里特对环境状态的其他方面的偏好，这样做也是一个好主意。正如我们所预料的那样，机器永远无法确定人类的偏好，但我们很高兴知道，尽管存在这种不确定性，它们仍然可以发挥作用。对带有部分偏好信息和不确定偏好信息的规划和决策的研究，有可能成为人工智能研究和产品开发的核心部分。

另外，所有其他条件都相同意味着不允许有其他的改变，例如，如果罗比不知道哈里特对咖啡和钱的相对偏好，那么在增加咖啡数量的同时减少支出可能是个好主意，也可能不是。

幸运的是，在所有其他条件相同的情况下，哈里特的指示可能不仅仅是简单地偏爱咖啡。额外的意义不仅来自她说的话，还来自她所说的事实、她说话时的具体环境以及她没说出口的事实。语用学是语言学的一个分支，它研究的正是这种延伸的意义概念。例如，对哈里特而言，如果哈里特认为附近没有咖啡，或者咖啡贵得离谱，那么她说“给我拿杯咖啡！”就是没有意义的。因此，当哈里特说：“给我拿杯咖啡！”此时罗比推断，哈里特不仅想要咖啡，而且她认为附近有咖啡，价格是她愿意承受的。因此，如果罗比发现咖啡的价格似乎是合理的（即这个价格在哈里特愿意支付的价格范围内），它就可以继续购买。相反，如果罗比发现最近的咖啡店在200英里之外或者价格为22美元，那么对它而言，报告这一事实而不是盲目购买可能更合理。

这种通用的分析风格通常被称为“格莱斯”（Gricean），这个名字源于加州大学伯克利分校的哲学家保罗·格莱斯（H. Paul Grice），他提出了一套准则来推断像哈里特这样的话语的延伸意义。[21]当确定偏好时，分析可能会变得相当复杂。例如，哈里特很可能不是特别想喝咖啡，她需要提神，却误以为加油站有咖啡，所以她要了咖啡。她可能同样喜欢茶、可口可乐，甚至一些包装花哨的能量饮料。

这些只是解释请求和指令时需要考虑的几个问题。关于这个主题的变化是无穷无尽的，因为哈里特的偏好是复杂的，哈里特和罗比所处的环境是多种多样的，以及哈里特和罗比在这些环境中可能拥有不同状态的知识和信念。虽然预先计算的脚本可能允许罗比处理一些常见的情况，但是灵活和鲁棒的行为只能从哈里特和罗比的交互中产生，这些交互实际上是他们所参与的辅助博弈的解决方案。

“主动嗑电”

在第2章中，我描述了基于多巴胺的大脑奖励系统，以及它在引导行为中的作用。多巴胺的作用在20世纪50年代末被发现，但早在那之前，人们就知道对老鼠大脑直接进行电刺激可以产生一种类似奖励的反应。[22]下一步是让老鼠接触一个连接着电池和电线的操纵杆，这个操纵杆会对老鼠的大脑产生电刺激。结果是令人警醒的：老鼠一次又一次地按下操纵杆，从不停下来吃东西或喝水，直到倒下。[23]人类也好不到哪里去，人类会成千上万次地寻求自我刺激，忽视食物和个人卫生。[24]（幸运的是，人体实验通常只进行一天。）动物为了直接刺激自己的奖励系统而使正常行为“短路”的倾向被称为“主动嗑电”（wireheading）。

类似的事情会发生在运行强化学习算法的机器上吗，比如AlphaGo？起初，人们可能会认为这是不可能的，因为AlphaGo得到获胜的+1奖励的唯一方法是赢得它正在玩的模拟围棋游戏。不幸的是，这是可能的，原

因在于AlphaGo和它的外部环境之间强制的、人为的分离，以及AlphaGo并不很智能的事实。请让我更详细地解释这两点，因为它们对于理解超级智能可能出错的一些方式很重要。

AlphaGo的世界只由模拟的围棋棋盘组成，棋盘纵横各19路，共计361个交叉点，这些位置可以为空，也可以放黑色或白色的棋子。虽然AlphaGo在一台计算机上运行，但它对这台计算机一无所知。特别是，它对判断每一盘棋胜负的那一小段代码一无所知。在学习过程中，它对于对手也没有任何概念，它的对手实际上是它自己的一个版本。AlphaGo唯一的动作就是将一枚棋子放在一个空的交叉点上，这些动作只影响围棋棋盘，而不影响其他任何东西，因为AlphaGo的模型中没有其他东西。这种设置对应于强化学习的抽象数学模型，其中的奖励信号来自“宇宙之外”。对AlphaGo而言，它无法对生成奖励信号的代码产生任何影响，因此AlphaGo不会沉迷于“主动嗑电”。

AlphaGo在训练期间的生活一定相当令人沮丧：它做得越好，它的对手就做得越好，因为它的对手几乎就是它自己的翻版。不管它变得多好，它的胜率都会徘徊在50%左右。如果它更智能，即如果它的设计更接近人类水平的人工智能系统，那么它将能够解决这个问题。这个AlphaGo++不会假设世界只是围棋棋盘，因为这一假设让很多事情无法解释。例如，它没有解释是什么“物理定律”在支持AlphaGo++自己做决策的操作，也没有解释神秘的“对手”来自何处。正如我们这些好奇的人类已经逐渐理解宇宙的运行一样，通过这种方式，某种程度上我们也了解了自己大脑的工作方式，就像第6章中讨论的Oracle AI一样，AlphaGo++将通过实验过程，了解到宇宙中的东西远不只围棋棋盘。它将计算出运行它的计算机及它自己的代码的运行规律，并且将认识到，如果没有宇宙中其他实体的存在，这样的系统是难以解释的。它将在棋盘上用不同的着法进行实验，想知道这些实体是否能解释它们。最终，它会通过一种模式语言与这些实体通信，并说服这些实体对它的奖励信号进行重新编程，使它总是得到+1。我们得到的必然的结论是，设计成奖励信号最大化的、功能足够强大的AlphaGo++将会出现“主动嗑电”行为。

几年来，人工智能安全界一直在讨论“主动嗑电”的可能性。[25]人们担心的是像AlphaGo这样的强化学习系统可能学会作弊，而不是掌握其预期的任务。当人类是奖励信号的来源时，真正的问题就出现了。如果我们提出，人工智能系统可以通过强化学习训练出良好的表现，利用人类给出的反馈信号来定义改进的方向，那么不可避免的结果是，人工智能系统会想出控制人类的方法，并迫使人类在任何时候都给予其最大的积极奖励。

你可能认为这只是人工智能系统的一种毫无意义的自欺行为，没错，事实正是如此。但这是强化学习定义的逻辑结果。当奖励信号来自“宇宙之外”，并由人工智能系统永远无法修改的某个过程生成时，这个过程就能正常工作。如果奖励生成过程（主体是人类）和人工智能系统处于同一个“宇宙”中，这个过程就失败了。

我们如何才能避免这种自欺行为？这个问题之所以会产生，是因为它混淆了两个不同的东西：奖励信号和实际奖励。在强化学习的标准方法中，它们是一种东西。这是个错误。它们应该被区分对待，就像它们在辅助博弈中一样：奖励信号提供关于实际奖励积累的信息，该信息才是要被最大化的东西。学习系统是在天堂积累积分，而奖励信号充其量只是提供积分的一种记录。换言之，奖励信号会通报奖励积累，而不是说奖励信号构成了奖励积累。在这个模型中，很明显，接管奖励信号机制的控制权只会丢失信息。制造虚假的奖励信号会使算法无法知道它的行为是否真的在天堂里积累积分，因此，一个被设计成能够区分二者区别的理性学习者，有动力去避免任何类型的“主动嗑电”。

递归式自我完善

I. J.古德对智能爆炸的预测（见第5章）是当前使人们担忧超级人工智能潜在风险的驱动力之一。如果人类可以设计出比人类更智能的机器，那么就会有人认为，机器在设计机器方面会强于人类。它会设计一

种更智能的新机器，这个过程将不断重复，用古德的话说，直到“人类的智能被远远抛在后面，望尘莫及”。

人工智能安全领域的研究人员，特别是位于伯克利的机器智能研究所（Machine Intelligence Research Institute，MIRI）的人员，研究了智能爆炸是否可以安全发生的问题。[26]最初，这项研究看起来很异想天开，难道这不是直接的“结束游戏”吗？但或许还有一丝希望。假设系列中的第一台机器“罗比·马克Ⅰ”是从对哈里特偏好的全面了解开始的。它知道自己的认知能力有限，在试图让哈里特开心的过程中会有不完美的情况出现，于是它建造了“罗比·马克Ⅱ”。从直觉上看，罗比·马克Ⅰ似乎有动机将它关于哈里特偏好的知识融入罗比·马克Ⅱ，因为这样可以让哈里特的偏好得到更好的满足，根据第一原则，这正是罗比·马克Ⅰ的目标。同理，如果罗比·马克Ⅰ对哈里特的偏好感到不确定，那么这种不确定也应该转移到罗比·马克Ⅱ身上。所以智能爆炸或许是安全的。

从数学的角度来看，美中不足的是，鉴于罗比·马克Ⅱ是假设的更高级的版本，罗比·马克Ⅰ会发现推理罗比·马克Ⅱ的行为方式并不容易。对于罗比·马克Ⅱ的行为，有些问题将是罗比·马克Ⅰ无法回答的。[27]更严重的是，我们还没有一个清晰的数学定义来解释机器有特定的目的（比如满足哈里特的偏好）实际上意味着什么。

让我们简单分析一下最后这个顾虑。想想AlphaGo：它有什么目的？这很容易，有人可能会想，AlphaGo的目的是在围棋中取胜，难道不是吗？当然，AlphaGo并不能保证逢局必胜。（事实上，它几乎总是输给AlphaZero。）的确，当离棋局结束只有几步棋的时候，AlphaGo会找到必胜的那步棋（如果有的话）。而当没有任何一着棋可以保证获胜时，换言之，当AlphaGo发现无论它选择什么着法，对手都有获胜的策略时，那么AlphaGo会或多或少地随机选择着法。它不会抱着对手会犯错的希望去尝试最狡猾的着法，因为它假定对手会完美地对弈。它会表现得好像失去了获胜的意愿。在其他情况下，当真正的最佳着法太难计算时，AlphaGo有时会犯错误，从而导致输棋。在这些情况下，在何种意义上

AlphaGo是真的想赢？实际上，它的行为可能与一台只想给对手一盘激烈的棋局的机器相同。

因此，我们说AlphaGo“有获胜的目的”是将其过度简化了。更好的描述应该是，AlphaGo是一个不完美的训练过程的结果，即它用自我对弈进行强化学习，而获胜就是奖励。之所以说训练过程是不完美的，是因为它不可能培养出完美的围棋棋手：AlphaGo学习了一个虽优秀但不完美的围棋棋局的评价函数，并将其与一个虽优秀但不完美的前向搜索相结合。

所有这一切的最终结果是：以“假设机器人R有目的P”开始的讨论，对于获得关于事物如何发展的直觉是可以的，但不能获得关于真实机器的定理。我们需要对机器的用途进行更细致、更精确的定义，才能保证它们长期的行为。人工智能研究人员才刚刚开始掌握如何分析真正的决策系统，[28]这只是最简单的一步，更不用说机器何时才能拥有足够强大的智能来设计它们的后继者了。我们还有许多工作要做。

---

[1] 费马大定理断言，当整数n>2时，关于a, b, c的方程$a^n + b^n = c^n$ 没有正整数解。费马在他那本丢番图所写的《算术》的页边空白处写道：“我对这个命题有一个绝妙的证明，但是这个空白太窄了，写不下。”无论是真是假，这都让数学家们在接下来的几个世纪里积极地寻求证明。我们可以很容易地检查特定的情况，例如，7^3是否等于$6^3+ 5^3$？（差不多，因为$7^3=343$，$6^3+5^3=341$，但是“差不多”不行。）当然，有无限多种情况需要检查，这就是为什么我们需要数学家，而不仅仅是计算机程序员。

[2] 机器智能研究所的一篇论文提出了许多相关问题：Scott Garrabrant and Abram Demski, “Embedded agency,” AI Alignment Forum, November 15, 2018。

[3] 多属性效用理论的经典著作：Ralph Keeney and Howard Raiffa, *Decisions with Multiple Objectives: Preferences and Value Tradeoffs* (Wiley, 1976)。

[4] 介绍逆强化学习思想的论文：Stuart Russell, “Learning agents for uncertain environments,” in *Proceedings of the 11th Annual Conference on Computational Learning Theory* (ACM, 1998)。

[5] 关于马尔科夫决策过程的结构式估计方法的原始论文：Thomas Sargent,“Estimation of dynamic labor demand schedules under rational expectations,”*Journal of Political Economy* 86 (1978): 1009–44。

[6] 第一个逆强化学习算法：Andrew Ng and Stuart Russell, “Algorithms for inverse reinforcement learning,” in *Proceedings of the 17th International Conference on Machine Learning*, ed. Pat Langley (Morgan Kaufmann, 2000)。

[7] 更好的逆强化学习算法：Pieter Abbeel and Andrew Ng, “Apprenticeship learning via inverse reinforcement learning,” in *Proceedings of the 21st International Conference on Machine Learning*, ed. Russ Greiner and Dale Schuurmans (ACM Press, 2004)。

[8] 将逆强化学习理解为贝叶斯更新：Deepak Ramachandran and Eyal Amir, “Bayesian inverse reinforcement learning,” in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, ed. Manuela Veloso (AAAIPress, 2007)。

[9] 如何教直升机飞行和进行特技飞行：Adam Coates, Pieter Abbeel, and Andrew Ng, “Apprenticeship learning for helicopter control,” *Communications of the ACM* 52 (2009): 97–105。

[10] 辅助博弈最初的名字是协同逆强化学习博弈（CIRL game）。请参见：DylanHadfield-Menell et al., “Cooperative inverse reinforcement learning,” in *Advances in Neural Information Processing Systems 29*, ed. Daniel Lee et al. (2016)。

[11] 选择这些数字只是为了让游戏更有趣。

[12] 博弈的均衡解可以通过一个称为“迭代最佳响应”的过程找到：先为哈里特选择任意策略；然后根据哈里特的策略，为罗比选择最佳策略；然后根据罗比的策略，为哈里特选择最佳策略；以此类推。如果这个过程达到一个固定的点，在这个点上两个策略都不再改变，那么我们就找到了一个解决方案。该过程如下所示：

（1）从哈里特的贪婪策略开始：如果她喜欢回形针，就做2个回形针；如果她更喜欢订书钉，就做2个订书钉；如果她保持中立，就回形针和订书钉各做1个。

（2）鉴于哈里特的这种贪婪策略，罗比必须考虑3种可能性：

a 如果罗比看到哈里特做2个回形针，它就推断她更喜欢回形针，所以它现在认为回形针的价值在50美分到1美元之间呈均匀分布，平均75美分。在这种情况下，它的最佳计划是为哈里特制作90个回形针，预期价值为67.5美元。

b 如果罗比看到哈里特每种各做1个，它就推断她对回形针和订书针的估价都是50美分，所以它此时的最佳选择是每种各做50个。

c 如果罗比看到哈里特做了2个订书钉，那么此时与情况a同理，它应该做90个订书钉。

（3）鉴于罗比的这种策略，哈里特现在的最佳策略与第一步中的贪婪策略有些不同：如果罗比用每种各做50个来回应她每种各做1个，那么无论她是完全保持中立，还是完全无所谓，她的最佳策略都是每种各做1个。实际上，如果她认为回形针的价值大约在44.6美分到55.4美分之间，那么现在的最佳策略就是每种各做1个。

（4）鉴于哈里特的新策略，罗比的策略保持不变。例如，如果她每种各做1个，那么它推断，回形针的价值是均匀分布在44.6美分和55.4美分之间，平均50美分，所以最佳选择是每种

各做50个。因为罗比的策略和步骤（2）相同，所以哈里特的最佳反应和步骤（3）相同，由此我们已经找到了均衡点。

[13] 关于关机博弈的更完整的分析，请参见Dylan Hadfield-Menell et al., “The off-switch game,” in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, ed. Carles Sierra (IJCAI, 2017)。

[14] 如果你不介意积分符号，那么一般性结果的证明很简单。假设P(u)是罗比的先验概率密度除以哈里特对建议动作a的效用。那么继续动作a的价值就是：

$$EU(a) = \int_{-\infty}^{\infty} P(u) \cdot u \, du = \int_{-\infty}^{0} P(u) \cdot u \, du + \int_{0}^{\infty} P(u) \cdot u \, d$$

（我们很快就会看到为什么以这种方式拆分积分。）另外，服从哈里特的动作d的值由两部分组成：如果u>0，则哈里特会让罗比继续，因此值为u，但是如果u<0，则哈里特会关闭罗比，因此值为0：

$$EU(d) = \int_{-\infty}^{0} P(u) \cdot 0 \, du + \int_{0}^{\infty} P(u) \cdot u \, du$$

比较EU（a）和EU（d）的表达式，我们会立即看到EU（d）≥EU（a），因为EU（d）的表达式负效用区域为零。只有当负区域的概率为0，即罗比已经确定哈里特喜欢提议的行动时，这两个选择才具有相等的价值。该定理是关于信息的非负期望值的著名定理的直接类比。

[15] 对于单人单机器的情况，也许下一步要详细阐述的是考虑一个还不知道自己对世界某些方面的偏好，或者还没有形成自己偏好的哈里特。

[16] 要了解罗比如何准确地收到一个错误的信念，请考虑一个模型，在这个模型中，哈里特有点不理性，她会犯错误，且犯错误的概率随着错误大小的增加而呈现指数级减小。罗比想用4个回形针换取哈里特1个订书钉，哈里特拒绝了。根据罗比的信念，这是不合理的：即使每个回形针25美分，每个订书钉75美分，她也应该接受4换1。因此，她一定是犯了错误，但如果回形针在她心中的真实价值是25美分，而不是30美分，这个错误就更有可能发生，因为如果回形针在她心中的真实价值是30美分，这个错误会让她损失更多的钱。现在，罗比的概率分布将25美分作为最有可能的值，因为它代表了哈里特方面的最小误差，大于25美分的值的概率呈指数级降低。如果它继续尝试同样的实验，概率分布变得越来越集中在接近25美分的位置。在极限情况下，罗比会确信哈里特的回形针价值是25美分。

[17] 例如，罗比对汇率的先验信念可能是从－∞到+∞的正态分布（高斯分布）。

[18] 可能需要的关于数学分析的示例，请参见：Avrim Blum, Lisa Hellerstein, and Nick Littlestone, “Learning in the presence of finitely or infinitely many irrelevant attributes,” *Journal of Computer and System Sciences* 50 (1995): 32－40。另请参见：Lori Dalton, “Optimal Bayesian feature selection,” in

Proceedings of the 2013 IEEE Global Conference on Signal and Information Processing, ed. CharlesBouman, Robert Nowak, and Anna Scaglione (IEEE, 2013)。

[19] 在这里，我稍微改写了摩西·瓦迪在2017年阿西洛马的有益人工智能会议上提出的问题。

[20] Michael Wellman and Jon Doyle, “Preferential semantics for goals,” in *Proceedings of the 9th National Conference on Artificial Intelligence* (AA-AI Press, 1991).这篇论文借鉴了乔治·冯·赖特（Georg von Wright）更早的提议“The logic of preference reconsidered,” *Theory and Decision* 3 (1972): 140–67。

[21] 我已故的加州大学伯克利分校的同事的名字已经成了形容词。请参见：PaulGrice, *Studies in the Way of Words* (Harvard University Press, 1989)。

[22] 关于直接刺激大脑愉悦中心的论文：“Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain,” *Journal of Comparative and Physiological Psychology* 47 (1954): 419–27。

[23] 让老鼠按下按钮：James Olds, “Self-stimulation of the brain; its use to study local effects of hunger, sex, and drugs,” *Science* 127 (1958): 315–24。

[24] 让人类按下按钮：Robert Heath, “Electrical self-stimulation of the brain in man,”*American Journal of Psychiatry* 120 (1963): 571–77。

[25] 关于“主动嗑电”现象的第一种数学处理方法，展示了主动嗑电现象在强化学习智能体中的发生方式：Mark Ring and Laurent Orseau, “Delusion, survival,and intelligent agents,” in *Artificial General Intelligence: 4th International Conference*, ed. Jürgen Schmidhuber, Kristinn Thórisson, and Moshe Looks(Springer, 2011)。“主动嗑电”现象的一种可能解决方案：Tom Everitt and Marcus Hutter, “Avoiding wireheading with value reinforcement learning,” arXiv:1605.03143 (2016)。

[26] 智能爆炸如何安全发生：Benja Fallenstein and Nate Soares, “Vingean reflection: Reliable reasoning for self-improving agents,” technical report 2015-2, Machine Intelligence Research Institute, 2015。

[27] 智能体在对它们自己和它们的继任者进行推理时所面临的困难：Benja Fallenstein and Nate Soares, “Problems of self-reference in self-improving space time embedded intelligence,” in *Artificial General Intelligence: 7th International Conference*, ed. Ben Goertzel, Laurent Orseau, and Javier Snaider (Springer,2014)。

[28] 说明为什么智能体在计算能力有限的情况下会追求与其真实目标不同的目标：Jonathan Sorg, Satinder Singh, and Richard Lewis, “Internal rewards mitigate agent boundedness,” in *Proceedings of the 27th International Conference on Machine Learning*, ed. Johannes Fürnkranz and Thorsten Joachims (2010), icml.cc/Conferences/2010/papers/icml2010proceedings.zip。


03
复杂的难题：
我们



如果世界上有一个完全理性的哈里特和一个乐于助人、恭敬有礼的罗比，我们的状态就会很好。罗比会尽可能无声无息地逐渐了解哈里特的偏好，并成为她的完美助手。我们可能希望从这个充满希望的开端推演下去，也许可以将哈里特和罗比的关系视为人类与机器之间关系的模型，无论是人类还是机器都被整体地解释。

然而，人类不是单一的、理性的实体，而是由多得不计其数的、令人讨厌的、嫉妒驱动的、非理性的、不一致的、不稳定的、计算能力有限的、复杂的、不断进化的、异质的实体组成的。这些问题是社会科学的主要内容，甚至可能是社会科学存在的理由。在人工智能中，我们需要加入心理学、经济学、政治理论和道德哲学的思想。[1]我们需要将这些思想熔化、塑形和锤炼成一个足够强大的结构，以抵御日益智能的人工智能系统给它带来的巨大压力。这项工作现在刚刚开始。

不同的人

我将从或许是最简单的问题——“人类是多种多样的”这一事实开始讲起。当人们第一次接触“机器应该学会满足人类偏好”的想法时，人们通常会反对，反对的理由是，不同的文化，甚至不同的个人，有着完全不同的价值体系，因此机器不可能有正确的价值体系。当然，这对机器而言不是问题：我们不希望它拥有自己正确的价值体系，我们只是想让它预测其他人的偏好。

关于机器难以满足人类不同偏好的困惑可能来自一种错误的想法，即机器采用的是它学习到的偏好，例如认为素食家庭中的家用机器人会采用素食偏好。事实上，机器人不会这样。它只需要学会预测素食者的饮食偏好。根据第一原则，它会避免为这户家庭烹饪肉类。但机器人也会了解邻居“肉食狂人”的饮食偏好，而且，如果在主人

同意的情况下，周末邻居借机器人帮忙举办晚宴，它会很乐意为他们做肉食吃。除了帮助人类实现他们的偏好外，机器人并没有自己的一套偏好。

从某种意义上讲，这与餐馆厨师学做几道不同的菜来满足客人不同的口味，或者跨国汽车公司为美国市场生产左舵驾驶汽车，为英国市场生产右舵驾驶汽车没有什么不同。

原则上，一台机器要学习80亿个偏好模型，即地球上每个人的偏好。而在实际中，这并不像听上去那么无望。首先，机器很容易互相分享它们学到的东西。其次，人类的偏好结构有很多共同点，所以机器通常不会从头开始学习每个模型。

例如，想象一下，有一天加州伯克利的居民要购买家用机器人。机器人带着相当广泛的先验信念从盒子里走出来，它可能是为美国市场量身定做的，但并非针对任何特定的城市、政治观点或社会经济阶层。机器人开始遇到伯克利的绿党成员，结果发现，与普通美国人相比，他们更有可能成为素食者，更有可能使用回收箱和堆肥箱，更有可能使用公共交通工具等。当一个新投入使用的机器人发现自己生活在一个绿党家庭中时，它可以立即相应地调整自己的期望。它不需要像以前从未见过人类一样开始了解这些特殊的人类，更不用说绿党成员了。这种调整并不是不可逆转的——伯克利有的绿党成员可能以濒临灭绝的鲸的肉为食，开着高油耗的大卡车，但这可以使机器人更快地变得实用。相同的论点适用于其他各种各样的个人特征，这些特征在某种程度上可以预测个人偏好结构的各个方面。

许多人

存在多人的另一个明显后果是，机器需要在不同人的偏好之间进行权衡。几个世纪以来，人与人之间的权衡问题一直是社会科学的主要焦点。如果人工智能研究人员期望他们能够不用了解已知成果就找

到正确的解决方案，那就太天真了。关于这个主题的文献浩如烟海，我不可能在这里对它做出公正的评价，不仅因为篇幅有限，还因为我没有读过其中的大部分内容。我还应该指出，几乎所有的文献都与人类的决策有关，而我在这里关注的是机器的决策。二者截然不同，因为人类有个人权利，这可能与任何假定的为他人履行的义务相冲突，而机器则没有。例如，我们不期望或要求一个真实的人类牺牲自己的生命来拯救他人，而我们肯定会要求机器人牺牲自己的存在来拯救人类的生命。

哲学家、经济学家、法学家和政治学家经过几千年的努力，制定出了宪法等法律、经济体系和社会规范，它们有助于（或阻碍，取决于谁负责）达成满意的折中解决方案。尤其是道德哲学家一直在从行为对他人有益或有害的影响的角度来分析行为正当性的概念。自18世纪以来，他们一直在效益主义的主题下研究权衡的量化模型。这项工作与我们目前关注的问题直接相关，因为它试图定义一个公式，根据这个公式，人工智能可以代表许多个人做出道德决定。

即使每个人都有相同的偏好结构，我们也需要进行权衡，因为我们通常不可能最大限度地满足每个人的偏好。例如，如果每个人都想成为全能的宇宙统治者，那么大多数人都会失望。另外，异质性确实让一些问题变得更加困难：如果每个人都对天空是蓝色的感到满意，那么处理大气问题的机器人就可以努力保持这种状态；如果很多人都在鼓吹改变天空的颜色，那么机器人就需要考虑可能的妥协，比如每个月的第三个星期五把天空变成橙色。

世界上不止一个人存在这一事实还产生了另一个重要的后果：这意味着，对每个人而言他都有其他人需要关心。这意味着满足一个人的偏好会对其他人产生影响，这取决于个人对他人幸福的偏好。

忠诚的人工智能

对于机器应该如何处理多人存在的问题，让我们从一个非常简单的建议开始讲起：机器应该忽略有多人存在这一事实。也就是说，如果哈里特拥有罗比，那么罗比应该只关注哈里特的偏好。这种忠诚的人工智能绕过了权衡的问题，但会导致下面这种问题：

罗比：你丈夫打电话提醒你今晚一起吃晚餐。

哈里特：等等！什么？什么晚餐？

罗比：庆祝你们的20周年纪念日，7点钟。

哈里特：我没时间！我7点半会见秘书长！这是怎么回事？

罗比：我确实提醒过你，但你没有听我的建议……

哈里特：好吧，对不起……但是我现在该怎么办？我不能跟秘书长说我太忙了！

罗比：别担心。我会安排她的飞机晚点，让飞机出点电脑故障。

哈里特：真的吗？你还能这么做？！

罗比：秘书长向你致以深深的歉意，并很高兴明天与你共进午餐。

此处，罗比找到了一个巧妙的方法来解决哈里特的问题，但它的行为对其他人产生了负面影响。如果哈里特是一个恪守道德规范且无私的人，那么旨在满足哈里特偏好的罗比永远也不会想到执行这样一个可疑的计划。但是，如果哈里特对其他人的偏好毫不在意，又会如何？那样的话，罗比不会介意让飞机晚点。它会不会花时间从网上银行账户里偷钱，来填满冷漠的哈里特的钱包，甚至做出更糟糕的事情？

显然，忠诚机器的行为需要受到规则和禁例的约束，就像人类的行为受到法律和社会规范的约束一样。有些人提出严格责任原则作为解决方案：[2]哈里特（或罗比的制造商，这取决于你希望让谁承担责任）对罗比的一切行为负有经济上和法律上的责任，就像在美国大多数州，如果狗在公园咬了小孩，狗的主人要承担责任一样。这个想法听起来很有希望，因为这样罗比就有动机避免做任何会给哈里特带来麻烦的事情。不幸的是，严格责任原则并不管用：它只是确保了罗比在为了哈里特的利益而造成飞机延误和偷钱时的行为不会被发现。这是漏洞原则在实践中的又一个例子。如果罗比忠于一个寡廉鲜耻的哈里特，那么试图用规则来约束它的行为就可能会失败。

即使我们能以某种方式阻止公然的犯罪，为冷漠的哈里特工作的忠诚的罗比也会表现出其他令人不快的行为。如果它在超级市场买东西，它会尽可能在收银台插队。如果它在购物回家途中遇到一个路人心脏病发作，它会选择无视，继续回家，仅仅为了让哈里特的冰激凌不融化。总之，它会找到无数种牺牲他人利益的方法来造福哈里特，这些方法是完全合法的，但在大规模实施时会变得不可容忍。各国会发现自己每天都要通过数百项新法律，以弥补机器将在现有法律中找到的所有漏洞。人类往往不会利用这些漏洞，或是因为他们对潜在的道德原则有普遍的理解，或是因为他们缺乏发现漏洞所需要的创造力。

一个对别人的幸福漠不关心的哈里特已经够糟糕的了，而喜欢把自己的快乐建立在别人痛苦之上的虐待狂哈里特则更加糟糕。为了满足这样的哈里特的喜好而设计的罗比将成为一个严重的问题，因为它会为了哈里特的快乐而寻找并找到伤害他人的方法，这些方法无论是否合法，别人都无法发现。它当然需要向哈里特汇报，这样她就可以从它的恶行中获得乐趣。

因此，打造忠诚的人工智能的想法很难成立，除非这个想法被拓展到除了考虑主人的偏好之外，还考虑其他人的偏好。

效益主义的人工智能

我们之所以有道德哲学，是因为地球上不止一个人生活。与理解人工智能系统应该如何设计最相关的方法通常被称为结果主义，即应该根据预期结果来评判选择。另外两种主要的方法是道义论伦理学和美德伦理学，非常粗略地讲，这两种伦理学分别关注行为和个人的道德品质，而与选择的后果无关。[3]没有任何证据表明机器有自我意识，因此我认为，如果其结果非常不符合人性，那么建造道德高尚或按照道德准则选择行动的机器就毫无意义。换言之，我们制造机器是为了带来结果，我们应该更喜欢制造能带来我们喜欢的结果的机器。

这并不是说道德准则和美德无关紧要，只是对效益主义者而言，他们是根据结果和那些结果的更实际的成就来判断正当性的。这一点是约翰·穆勒（John Stuart Mill）在《效益主义》中提出的：

> 提出幸福是道德的目的和目标这一命题，并不代表就不允许铺一条通往这一目标的路，也不代表就不允许给其中的行人指明方向……没有人会质疑航海技术是建立在天文学基础上的，因为海员们根据天文来计算航海年鉴。作为富有理性的生灵，他们出海时都已经进行了计算；而我们都是富有理性的生灵，都在生命的海洋中航行，需要用我们自己的心智去决定是非对错等常见问题，以及其他诸多比智慧和愚蠢更为复杂的事情。

与穆勒的观点完全一致的观点是：有限的机器在面对现实世界的巨大复杂性的时候，可能会遵循道德准则并采取一种善良的态度，从而产生更好的结果，而不是试图从头开始计算最佳行动方案。同样，国际象棋程序更经常使用标准的开局走法、残局算法和评价函数来实现“将杀”，而不是试图在没有“道德”路标的情况下推理“将杀”的方式。结果主义的方法也给予了那些坚信保留给定道义规则的人的

偏好一些权重，因为对打破规则的不满是真实的结果。然而，这并不是无限影响的结果。

尽管很多人都尝试过反驳结果主义，但是结果主义是一个很难反驳的原则。因为以结果会产生不良结果为理由反对结果主义是不合逻辑的。人们不能说，“但是如果你在某某情况下遵循结果主义的方法，那么这种非常可怕的事情就会发生”。任何这样的失败都只是该理论被误用的证据。

例如，假设哈里特想攀登珠穆朗玛峰。人们可能会担心结果主义者罗比会简单地把她抱起来放到珠穆朗玛峰顶上，因为这是她想要的结果。但是哈里特很可能会强烈反对这个计划，因为这会剥夺她通过自己的努力成功完成一项艰巨任务的权利，而这一任务会给她带来挑战和喜悦。现在很明显，一个设计合理的结果主义者罗比会明白，结果包括哈里特所有的经历，而不仅仅是最终目标而已。它可能希望在发生意外事故时能在场，并确保她带上了适当的装备，完成了必要的训练，但它也可能不得不接受哈里特有权利将自己置于明显的死亡危险之下。

如果我们计划建造结果主义机器，下一个问题是如何评估影响多人的结果。一个貌似合理的答案是给予每个人的偏好同等的权重，也就是最大化每个人的效用总和。这个答案通常可以追溯至18世纪的英国哲学家杰里米·边沁[4]和他的学生约翰·穆勒[5]，他们发展了效益主义的哲学方法。其底层思想源自古希腊哲学家伊壁鸠鲁的著作，这一思想还明确地出现在中国哲学家墨子的著作《墨子》中。墨子活跃于公元前4世纪，他提出了“兼爱”的思想，意为“包容的关怀”或“普世的爱”，以此定义道德行为的特征。

效益主义的名声有些不太好，部分原因是人们对它的主张存在简单的误解。（效益主义一词的意思是“为了实用而设计，而不是为了吸引人而设计”，这当然无济于事。）效益主义经常被认为与个人权利不相容，因为效益主义者会对未经许可就切除一个活人的器官去挽

救其他五个人的生命无动于衷。当然，这样的政策会使地球上每个人的生活都处于让人无法忍受的不安全之中，而效益主义者甚至不会考虑这一点。效益主义也被错误地认定为一种相当不讨人喜欢的最大化财富的行为，并被认为不重视诗歌或苦难。事实上，边沁的版本特别关注人类的幸福，而穆勒则自信地宣称智力上的快乐远比单纯的感觉更有价值。（“当一个不满足的人比当一头满足的猪更好。”）G. E. 摩尔（G. E. Moore）的理想效益主义走得更远：他主张将内在价值的心理状态最大化，这集中体现为对美的审美观照。

我认为效益主义哲学家没有必要规定人类效用或人类偏好的理想内容。（人工智能研究人员这样做的理由就更少了。）人类可以为自己做到这一点。经济学家约翰·豪尔绍尼（John Harsanyi）用他的偏好自主原则提出了这一观点：[6]

> 一个人在决定什么是好什么是坏时采用的最终标准只能是他自己的欲望和偏好。

因此，豪尔绍尼的偏好效益主义与“有益的AI”的第一原则大致一致，即机器的唯一目的是实现人类的偏好。人工智能研究人员绝对不应该从事决定人类偏好的工作！和边沁一样，豪尔绍尼把这些原则视为公共决策的指南，他不期望个人完全无私。他也不期望个人完全理性，例如，一个人可能拥有与自己“更深层次的偏好”相矛盾的短期欲望。最后，他建议忽略前面提到的虐待狂哈里特那种积极希望损人利己的人的偏好。

豪尔绍尼还给出了一种证明，即最优道德决策应该最大化整个人类群体的平均效用。[7]他的假设相当薄弱，类似于个人效用理论的基础假设。（主要的附加假设是，如果一个群体中的每个人对两种结果都漠不关心，那么代表该群体的智能体对这些结果也漠不关心。）根据这些假设，他证明了后来所谓的“社会聚合定理”：代表个人群

体的智能体必须最大化个人效用的加权线性组合。他进一步认为，一个“客观的”智能体应该使用相同的权重。

“社会聚合定理”需要一个关键的（且未阐明的）附加假设：每个人对这个世界以及世界将如何演变都有相同的先验事实信念。现在，每位父母都知道这甚至对于兄弟姐妹来说都不属实，更不用说对来自不同社会背景和文化的人了。那么，当个人的信念不同时会发生什么？有些事情会相当奇怪：[8]分配给每个人的效用权重必须随着时间的推移而改变，使之和那个人先前的信念与当下现实的相符程度成比例。

这个听起来相当不平等的公式对每位父母来说都很熟悉。假设机器人罗比被指派照看两个孩子爱丽丝和鲍勃。爱丽丝想去看电影，并确信今天会下雨；而鲍勃想去海滩，并确信今天会是晴天。罗比可以宣布“我们去看电影”，但这会让鲍勃不高兴；或者可以宣布“我们去海滩”，但这会让爱丽丝很不高兴；或者可以宣布，“如果下雨，我们就去看电影，如果是晴天，我们就去海滩”。最后这个计划让爱丽丝和鲍勃都很高兴，因为他们都相信自己的信念。

对效益主义的挑战

效益主义是人类在长期寻找道德指南的过程中出现的一种提议。在许多这样的提议中，它是最明确的，因此也是最容易出现漏洞的。100多年来，哲学家们一直在寻找这些漏洞。例如，G. E.摩尔反对边沁强调的快乐最大化，他设想了一个“除了快乐什么都不存在的世界——没有知识，没有爱，没有美的享受，没有道德品质”。[9]其现代的呼应可以在牛津大学牛津马丁学院人类未来研究所的研究员斯图尔特·阿姆斯特朗（Stuart Armstrong）的观点中找到，他说，以最大限度地获得快乐为任务的超级智能机器可能会“把每个人都装入混凝土棺材里注射海洛因”。[10]还有一个例子，1945年，奥地利哲学

家卡尔·波普尔（Karl Popper）提出了一个值得称赞的目标——最大限度减少人类的痛苦，[11]他认为用一个人的痛苦来换取另一个人的快乐是不道德的。R. N.斯马特（R. N. Smart）回应说，最好的实现办法是让人类灭绝。[12]如今，有关机器可能通过终结我们的存在来终结人类痛苦的观点，是关于人工智能生存性风险的辩论的主要内容。[13]第三个例子是G. E.摩尔对幸福之源的实际情况的强调，G. E.摩尔修正了早期似乎存在漏洞的定义，即允许通过自我欺骗来实现幸福最大化。这一点的现代类比有《黑客帝国》（在《黑客帝国》中，当下的现实是由计算机模拟产生的错觉），还有最近关于强化学习中的自我欺骗问题的研究。[14]

这些例子以及更多的例子让我相信，人工智能界应该认真关注哲学和经济中关于效益主义利弊的辩论，因为这与人工智能手头的任务直接相关。从设计使多人受益的人工智能系统的角度来看，其中两个最重要的问题涉及人际效用比较以及不同人口规模的效用的比较。这两场辩论已经持续了150多年，这让人怀疑它们令人满意的解决方案可能并非完全简单明了。

关于人际效用比较的辩论很重要，因为罗比不可能最大化爱丽丝和鲍勃的效用之和，除非这些效用可以相加，而只有当它们可以在相同的尺度上衡量时，它们才能相加。19世纪英国逻辑学家和经济学家威廉·斯坦利·杰文斯（William Stanley Jevons，也是早期被称为“逻辑钢琴”的机械计算机的发明者）在1871年提出，人际效用是不可比较的。[15]

> 就我们所知，一个人头脑的敏感性可能是另一个人头脑敏感性的100倍。但是，如果敏感性在各方面的差异都呈现出类似的比例，那么我们就永远无法发现最根本的差异。因此，每一个人的头脑对于其他人的头脑而言都是不可思议的，不可能有共同的感觉标准。

现代社会选择理论的创始人、1972年诺贝尔奖得主、美国经济学家肯尼斯·阿罗（Kenneth Arrow）同样坚定地表示：

> 这里采取的观点是，人际效用比较没有任何意义，事实上，就个人效用的可衡量性而言，对幸福感做比较没有意义。

杰文斯和阿罗所说的困难在于，没有任何明显的方法可以根据自己的主观幸福感来判断爱丽丝将针刺和棒棒糖的价值定为－1和+1，还是－1 000和+1 000。无论在哪种情况下，她都会愿意付出一根棒棒糖来避免一次针刺。事实上，如果爱丽丝是一个类人机器人，即使她没有任何主观的快乐体验，她的外部行为可能也是一样的。

1974年，美国哲学家罗伯特·诺齐克（Robert Nozick）提出，即使可以对效用进行人际比较，最大化效用的总和仍然不是一个好主意，因为这会与“效用怪兽”——一个快乐和痛苦的体验比普通人强烈许多倍的人相冲突。[16]可以断言，任何额外的资源单位，如果给予了这类人而不是其他人，都会更为增加人类幸福的总和，其实，从其他人那里移除资源来造福“效用怪兽”也是一个好主意。

这是一个明显不可取的结果，但结果主义本身并不能解决问题：问题在于我们如何衡量结果的可取性。一种可能的反应是，“效用怪兽”仅仅是理论上的，因为现实中根本没有这样的人。但这种反应可能行不通：从某种意义上讲，相对于老鼠和细菌，所有的人类都是“效用怪兽”，这就是为什么我们在制定公共政策时很少关注老鼠和细菌的偏好。

如果不同的实体有不同的效用尺度的想法已经建立在我们的思维方式中，那么不同的人有不同的尺度也是完全可能的。

另一种反应是说“真倒霉！”并假设每个人都有相同的尺度，即便事实并非如此。[17]人们也可以尝试用杰文斯所不具备的科学手段

来研究这个问题，例如测量多巴胺水平，或者测量与快乐和痛苦相关的神经元的电兴奋程度。如果爱丽丝和鲍勃对棒棒糖的化学反应、神经反应，以及他们的行为反应（微笑、咂嘴唇的声音等）几乎是相同的，那么坚持认为他们的主观快乐程度相差成千上万倍就显得很奇怪。最后，我们可以使用通用货币，比如时间（我们都有大致相同的时间），例如，要拿到棒棒糖需要在机场候机室多等待5分钟，以此来和针刺进行比较。

我远不如杰文斯和阿罗悲观。我认为比较个人效用确实有意义，尺度可能有所不同，但通常不会差很多倍，机器可以从宽泛的关于人类偏好尺度的先验信念开始，通过长时间的观察来了解更多关于个体尺度的信息，这或许可以将自然观察与神经科学研究的发现联系起来。

第二场辩论是关于不同规模人口之间的效用比较，当决策对未来的人产生影响时，这一点很重要。例如，在电影《复仇者联盟：无限战争》中，灭霸这个角色发展并实施了一个理论：如果人口减半，剩下的每个人的幸福感都会增加一倍以上。这是一种幼稚的计算，让效益主义名声扫地。[18]

1874年，英国哲学家亨利·西奇威克（Henry Sidgwick）在他的著作《伦理学方法》中讨论了同样的问题，[19]只是没有漫威电影中的无限宝石和庞大的预算而已。西奇威克显然同意灭霸的观点，他的结论是，正确的选择是调整人口规模，直到所有人的总幸福达到最大化。（显然，这并不意味着无限制地增加人口，因为到某个时候，所有人都会饿死，所以会非常不幸福。）1984年，英国哲学家德里克·帕菲特（Derek Parfit）在他开创性的著作《理与人》中再次提出了这个问题。[20]帕菲特认为，相较于任何有N个非常幸福的人的情况，根据效益主义原则，有2N个幸福程度稍差的人，这种情况更好。这似乎很有道理。不幸的是，这是一个滑坡谬误。通过重复这一过程，我们得出了所谓的“令人厌恶的结论”（Repugnant

Conclusion，首字母通常大写，或许是为了强调其来自维多利亚时代）：最理想的情况是人口众多，所有人都过着几乎不值得过的生活。

可以想象，这样的结论是有争议的。帕菲特自己奋斗了30多年，力图解决自己提出的难题，但他没有成功。我认为我们缺少了一些基本的公理，类似于个人理性偏好的公理，来处理不同规模和幸福水平的群体之间的选择。[21]

解决这个问题对我们来说很重要，因为有远见的机器可能会考虑导致不同人口规模的行动方案。例如，我们很可能会要求人工智能系统帮助设计全球气候变化的解决方案，而这些解决方案很可能涉及限制甚至减少人口数量的政策。[22]相反，如果我们认为人口越多越好，且如果我们从现在开始的几个世纪中，对潜在的庞大人口的幸福给予足够的重视，那么我们将需要更加努力地寻找超越地球界限的方法。如果机器的计算得出了“令人讨厌的结论”，或相反的结论，即让一小群人最幸福，那么我们可能有理由感到遗憾，因为我们在这个问题上缺乏进展。

一些哲学家认为，我们可能需要在道德不确定的状态下做出决定，也就是说，在决策中使用适当的道德理论具有不确定性。[23]一种解决方案是为每种道德理论分配一定的概率，并使用“预期道德价值”做决策。然而，将概率与道德理论联系起来，就像将概率与明天的天气联系起来一样，目前还不清楚是否有意义。（灭霸完全正确的概率有多大？）而且，即使这确实有意义，那么相互竞争的道德理论的提议之间潜在的巨大差异也意味着，解决道德不确定性——找出哪种道德理论可以避免不可接受的后果，必须在我们做出重大决定或将其委托给机器之前完成。

让我们乐观一点，假设哈里特最终解决了这个问题，也解决了因为地球上存在不止一个人而产生的其他问题。合适的利他主义和平等

主义算法被下载到世界各地的机器人中。伴着庆祝的掌声和欢快的音乐，哈里特回家了……

罗比：欢迎回家！忙了一天？

哈里特：是的，工作很忙，连吃午饭的时间都没有。

罗比：那你一定很饿了！

哈里特：饿死了！你能给我做点晚餐吗？

罗比：有件事我得告诉你……

哈里特：什么？别告诉我冰箱是空的！

罗比：不是，索马里还有人更需要帮助，我现在要走了，请你自己做饭。

哈里特也许会为罗比感到骄傲，也会为她使罗比成为这样一个正派的机器人所做的贡献感到自豪，但她不禁要问，为什么她要花一大笔钱去买一个这样的机器人，其第一个重要行动就是消失。当然，在现实中，没有人会买这样的机器人，人们也不会制造这样的机器人，所以这不会给人类带来任何好处。我们称之为“索马里问题”。为了使整个功利机器人计划奏效，我们必须找到解决这个问题的办法。罗比需要对哈里特有一定的忠诚度，这或许与哈里特为罗比所付的钱有关。也许，如果社会希望罗比帮助哈里特以外的人，那么社会就需要补偿哈里特对于罗比提供服务的需求。机器人很可能会互相协调，这样它们就不会同时降落在索马里了，这种情况下，罗比可能根本就不需要去索马里。或者可能会出现一些全新的经济关系来应对世界上（肯定是前所未有的）数十亿纯粹利他主义智能体的存在。

善良的、可恶的、爱嫉妒的人类

人类的偏好远不只快乐和比萨。这些偏好当然会扩展至其他人的幸福。在需要为自私辩护时，经济学之父亚当·斯密的理论经常被引用，他的第一本书《道德情操论》开篇就强调了关心他人的重要性：[24]

> 无论人们会认为某人怎样自私，这个人的天赋中总是明显地存在着这样一些本性，这些本性使他关心别人的命运，把别人的幸福看成自己的事情，虽然他除了看到别人幸福而感到高兴以外，一无所得。这种本性就是怜悯或同情，就是当我们看到或逼真地想象到他人的不幸遭遇时所产生的感情。我们常为他人的悲哀而感伤，这是显而易见的事实，不需要用什么实例来证明。

用现代经济学的说法，对他人的关心通常被归入利他主义的范畴。[25]利他主义理论已经相当完善，对税收政策和其他问题都产生了重大影响。必须指出的是，一些经济学家把利他主义视为另一种形式的自私，旨在为给予者提供“温情”。[26]这当然是机器人在解释人类行为时需要意识到的一种可能性，但现在让我们暂且相信人类是天真的，并假设他们真的在乎别人。

思考利他主义的最简单方法是将一个人的偏好分为两类：一类是对自己内在幸福的偏好，另一类是对他人幸福的偏好。（人们对于是否能将它们完美地分开存在相当大的争议，但我要把这个争议放到一边。）内在幸福指的是一个人自身的生活品质，比如拥有住所、吃得饱、穿得暖、安全等，这些生活品质本身就是令人向往的，而不是一个人参考了别人的生活才想要的。

为了使这个概念更具体，我们假设世界上有两个人爱丽丝和鲍勃。爱丽丝的总效用是由她自己的内在幸福感加上关心因子CAB乘以鲍勃的内在幸福感组成的。关心因子CAB表示爱丽丝对鲍勃的关心程度。同理，鲍勃的总效用等于他的内在幸福感加上关心因子CBA乘以爱丽丝

的内在幸福感，其中CBA表示鲍勃对爱丽丝的关心程度。[27]罗比试图同时帮助爱丽丝和鲍勃，这意味着最大化他们二人的效用总和。因此，罗比不仅需要关注两个人的个人幸福，还需要关注这两个人各自对他人幸福的关心程度。[28]

关心因子CAB和CBA非常重要。例如，如果CAB为正，那么爱丽丝就是“善良的”：她会从鲍勃的幸福中获得一些快乐。CAB越大，爱丽丝就越愿意牺牲自己的幸福来帮助鲍勃。如果CAB是零，那么爱丽丝就是完全自私的：如果她能逃脱惩罚，那么她会把任意数量的资源从鲍勃身上转移给自己，即使鲍勃一贫如洗、挨饿受冻也是如此。面对自私的爱丽丝和善良的鲍勃，效益主义者罗比显然会保护鲍勃免受爱丽丝最恶劣的掠夺。有趣的是，最终的均衡通常会让鲍勃比爱丽丝拥有更少的内在幸福感，但他可能拥有更大的总幸福感，因为他在意爱丽丝的幸福。你可能会觉得罗比的决定非常不公平，如果仅仅因为鲍勃比爱丽丝善良，他的幸福指数就比爱丽丝低，难道他就不会对这个结果感到不满？[29]他可能会，但那是一个不同的模型，一个包含了对幸福差异的怨恨的模型。在我们的简单模型中，鲍勃会对结果感到满意。事实上，在均衡状态下，他会控制把资源从爱丽丝身上转移给他自己的企图，因为这会降低他的整体幸福感。如果你认为这完全不现实，那么请考虑一下爱丽丝是鲍勃的亲生女儿的情况。

对罗比而言，真正有问题的情况发生在CAB是负数时：在这种情况下，爱丽丝确实很讨厌。我用“消极利他主义”这个词来指代这种偏好。就像之前提到的虐待狂哈里特一样，这不同于普通的贪婪和自私，普通的贪婪和自私是爱丽丝满足于减少鲍勃的资源份额来增加她自己的份额。消极利他主义意味着，爱丽丝纯粹从损人利己的行为中获得幸福，即使她自己的内在幸福感没有变化。

豪尔绍尼在介绍偏好效益主义的论文中将消极利他主义归因于“施虐、嫉妒、怨恨和恶意”，并认为在计算人口中人类效用的总和时应该忽略这些因素。

无论我对X个人有多少善意，也不能把帮助他伤害第三个人Y的道德义务强加给我。

可以说，在这一领域，智能机器的设计者们应该（谨慎地）站在正义一边。

不幸的是，消极利他主义行为比人们想象的要普遍得多。它并不是由虐待狂和恶意引起的，[30]而是由嫉妒、怨恨和与之相反的情绪引起的，我称这种情绪为“骄傲”（因为找不到更好的词）。如果鲍勃嫉妒爱丽丝，那么他的不快乐来自爱丽丝的幸福和他自己的幸福之间的差距。该差距越大，他就越不快乐。相反，如果爱丽丝为自己比鲍勃优越而感到骄傲，那么她不仅从自己的内在幸福中获得幸福，还从自己比鲍勃幸福的事实中获得幸福。从数学意义上很容易看出，骄傲和嫉妒的作用方式与虐待狂大致相同，它们导致爱丽丝和鲍勃纯粹从减少对方的幸福中获得幸福，因为鲍勃幸福感的降低会增加爱丽丝的骄傲，而爱丽丝幸福感的降低则会减少鲍勃的嫉妒。[31]

著名的发展经济学家杰弗里·萨克斯（Jeffrey Sachs）曾给我讲过一个故事，说明了这种偏好在人们思维中的力量。在一场大洪水摧毁了孟加拉国的一个地区后不久，他来到这里，同一位失去了房屋、田地、所有牲畜和一个孩子的农民交谈。萨克斯冒昧地说：“我很抱歉……你一定非常难过。”然而农夫回答说：“一点也不难过，我很高兴，因为我那该死的邻居失去了他的妻子和所有的孩子！”

美国社会学家索尔斯坦·凡勃伦（Thorstein Veblen）在对骄傲和嫉妒进行的经济学分析方面做出了大量杰出工作，特别是在社会地位和炫耀性消费结合的背景下，他在1899年出版的著作《有闲阶级论》中解释了这些态度的有害后果。[32] 1977年，英国经济学家弗雷德·赫希（Fred Hirsch）出版了《增长的社会极限》，[33]他在书中提出了“地位商品”的概念。地位商品可以是任何东西，例如汽

车、房产、奥运奖牌、教育、收入或口音，它的感知价值不仅来自其内在利益，还来自它的相对属性，包括稀缺性和优于他人的属性。由骄傲和嫉妒驱使的对地位商品的追求具有零和博弈的特征，即爱丽丝不可能在使鲍勃的相对地位降低的情况下改善她的相对地位，反之亦然。（这似乎并没有阻止大量钱财被浪费在这种追求之上。）地位商品在现代生活中无处不在，因此机器需要了解它们在个人偏好中的整体重要性。此外，社会认同理论的学者提出，个人在群体中的身份和地位以及该群体相对于其他群体的整体地位是人类自尊的重要组成部分。[34]因此，如果不理解个人如何将自己视为群体的成员，就很难理解人类的行为，无论其身处于物种、国家、民族、政党、公司、家庭还是某支足球队的球迷团等群体中。

就像虐待和恶意一样，我们可以建议罗比在帮助爱丽丝和鲍勃的计划中很少考虑或根本不考虑骄傲和嫉妒。然而，执行这项建议存在一些困难。因为在爱丽丝对鲍勃幸福感的态度中，骄傲和嫉妒抵消了关心，所以要把它们分开可能并不容易。可能是爱丽丝对鲍勃很关心，但她也很嫉妒鲍勃。我们很难把这个爱丽丝和另一个对鲍勃不怎么关心却一点也没有嫉妒之心的爱丽丝区分开来。此外，考虑到人类偏好中普遍存在的骄傲和嫉妒，我们非常有必要仔细考虑忽视它们的后果。也许它们是“面子”至关重要的一部分，尤其是它们的积极形式——自尊和对他人的钦佩。

让我再次强调之前说过的一点：设计得当的机器不会像它们观察的人那样行事，即便这些机器正在学习虐待狂的偏好。事实上，如果我们人类发现自己每天都处于与纯粹利他主义实体打交道的陌生环境中，我们就可能会学习成为更好的人，变得更无私，也更少受到骄傲和嫉妒的驱使。

愚蠢的、情绪化的人类

本节的标题并不是指某个特定的人类群体，而是指我们所有人。与完美理性所设定的不可企及的标准相比，我们都是极其愚蠢的，我们受制于各种情绪的起伏，这些情绪在很大程度上支配着我们的行为。

让我们从愚蠢开始谈起。一个完全理性的实体会在未来所有可能的生活中最大限度地满足自己的偏好。我无法写下一个数字来描述这个决策问题的复杂性，但我发现下面的思想实验很有帮助。首先请注意，人类一生中做出的运动控制选择的数量约为20万亿次。（详细计算请见附录A。）接下来，让我们借助塞思·劳埃德的终极物理笔记本电脑，看看蛮力能让我们走多远，它比世界上最快的电脑快10^{33}倍。我们给它的任务是列举所有可能的英语单词序列（或许是为阿根廷作家豪尔赫·路易斯·博尔赫斯笔下的巴别塔图书馆做准备），我们让它运行一年。在这段时间内它可以列举的序列有多长？1 000页？100万页？都不对，只有11个词。这让你感受到了设计具有20万亿次动作的最佳生命体的难度。简而言之，我们距离理性还差得远，比“鼻涕虫”追赶上在九度空间旅行的“企业号”星际飞船还难。我们不知道完全靠理性选择的生活会是什么样子。

这意味着人类通常会以违背自己偏好的方式行事。例如，当李世石在围棋比赛中输给AlphaGo时，他下出了一步或几步会让他输棋的臭棋，而AlphaGo（至少在某些情况下）可能察觉到李世石已经下出了这样的棋。然而，如果AlphaGo推断李世石更喜欢输，那就错了。相反，AlphaGo可以合理地推断李世石更喜欢赢棋，但他的计算能力有限，无法在所有情况下都选择正确的着法。因此，为了理解李世石的行为并了解他的偏好，遵循第三原则（“关于人类偏好的最终信息来源是人类行为”）的机器人必须了解李世石产生该行为的认知过程。它不能假设李世石是理性的。

这给人工智能、认知科学、心理学和神经科学领域带来了一个非常严肃的研究课题：为了充分了解人类的认知，[35]我们（或者更确

切地说，我们的有益机器）可以对人类行为进行“逆向工程”，以获取深层的潜在偏好，直到这些偏好出现。人类设法做到了其中的一些，在生物学的指导下，从别人那里学习自己的价值观，所以这似乎是可能的。人类有一个优势：一个人可以使用自己的认知架构来模拟其他人的认知架构，而不需要知道那个架构是什么，“如果我想要X，我会和妈妈做同样的事情，所以妈妈一定想要X”。

机器没有这个优势。它们可以轻松模拟其他机器，但不能模拟人。它们不太可能很快获得完整的人类认知模型，无论是通用的还是针对特定个人的。相反，从实践的角度来看，观察人类偏离理性的主要方式，并研究如何从表现出这种偏离的行为中学习偏好是有意义的。

人类和理性实体之间的一个明显的区别是：在任意给定的时刻，我们都不是在开启所有可能的未来生活的所有可能的第一步中进行选择，甚至还差得很远。相反，我们通常被嵌入一个深度嵌套的“子程序”层次结构中。一般而言，我们追求的是短期目标，而不是最大化对未来生活的偏好，我们只能根据我们目前所处的子程序的约束来行动。例如，现在，我正在输入这句话：我可以选择在冒号之后如何继续往下写，但我从来没有想过是应该停笔去参加一个网络说唱课程，还是应该烧毁房子索赔保险，或是我接下来可以做的无数事情中的任何一件。其他很多事情都可能比我现在做的事情更好，但是，考虑到我的承诺等级，那些其他的事情就好像不存在一样。

因此，要理解人类的行为，就需要理解这个子程序的层次结构（可能是非常个性化的）：这个人目前正在执行哪个子程序，这个子程序正在追求哪些短期目标，以及它们如何与更深层次的长期偏好相关联。更普遍地说，要了解人类的偏好，就需要了解人类生活的实际结构。我们人类可以单独或共同从事的事情有哪些？不同文化和不同类型的人有什么不同的活动特征？这些都是非常有趣和艰巨的研究问题。很明显，它们并没有一个固定的答案，因为我们人类一直在向我

们的“剧目”中添加新的活动和行为结构。但是，即使是临时的不完全的答案，对于旨在辅助人类日常生活的各种智能系统来说也是非常有用的。

人类行为的另一个明显特征是，它们常常是由情感驱动的。在某些情况下，这是一件好事，像爱和感恩这样的情绪当然是我们偏好的组成部分，由它们引导的行动即使没有经过深思熟虑，可能也是理性的。在其他情况下，情绪反应导致的行为甚至连我们这些愚蠢的人都认为不是理性的，当然，这是事后的看法。例如，愤怒而又沮丧的哈里特扇了10岁的倔强的爱丽丝一巴掌，她可能会立即后悔自己的行为。观察到这一行为的罗比（通常，尽管并非在所有情况下）应该将这一行为归因于愤怒、沮丧和缺乏自制力，而不是出于自身原因的故意施虐。为了做到这一点，罗比必须对人类的情感状态有所了解，包括人类情感的起因、感情如何随着时间的推移对外部刺激做出反应，以及它们对行动的影响。神经科学家已经开始掌握一些情绪状态的机制，以及它们与其他认知过程的联系，[36]并且在检测、预测和操纵人类情绪状态的计算方法方面做了一些有用的工作，[37]但还有更多的东西需要学习。同样，在情感方面，机器也处于劣势：它们无法对体验进行内部模拟，以查看会产生何种情绪状态。

情绪除了会影响我们的行为，还揭示了关于我们潜在偏好的有用信息。例如，小爱丽丝可能拒绝做家庭作业，哈里特为此感到生气和沮丧，因为她非常希望爱丽丝在学校表现好，能比自己拥有更好的机会。如果罗比能够理解这一点，即使它自己不能亲身体验，它也能从哈里特不太理性的行为中学到很多。因此，人们应该有可能创建出人类情感状态的基本模型，从而足以避免根据行为推断人类偏好时出现的最严重的错误。

人类真的有偏好吗？

这本书的整个前提是，有些未来是我们想要的，有些未来是我们不想要的，比如人类在近期灭绝，或者世界像《黑客帝国》那样变成人类电池农场。在这个意义上讲，是的，人类当然有偏好。然而，一旦我们深入了解人类更希望自己的生活如何继续下去的细节，事情就变得更加扑朔迷离了。

不确定性和错误

如果你仔细想想，你就会发现人类有一个明显的特点，那就是并不总是知道自己想要什么。例如，榴梿会引起不同人的不同反应：有人觉得“榴梿的味道胜过世界上所有的水果”，[38]也有人把它比作“污水、难闻的呕吐物、臭鼬喷雾、用过的外科棉签”。[39]在出版本书之前，我刻意避免品尝榴梿，这样我就可以在这一点上保持中立：我根本不知道自己会站在哪一边。对于许多考虑未来职业、未来生活伴侣、未来退休生活等事情的人而言，情况可能也是如此。

偏好的不确定性至少有两种。第一种是真实的、认知上的不确定性，比如我对榴梿偏好的体验。[40]再多的思考也解决不了这个不确定性。这是一个经验性的事实，我可以通过品尝一些榴梿，将我的DNA与喜欢榴梿的人和讨厌榴梿的人的DNA进行比较等，从而找到更多信息。第二个问题来自计算能力的限制：对于面前摆着的两盘围棋棋局，我不确定我更喜欢哪一盘棋局，因为无论哪一盘棋局我都无力下赢。

不确定性也源于这样一个事实，即呈现在我们面前的选择通常不是完全明确的，有时非常不完全，它们甚至根本不具备资格成为一项选择。当爱丽丝即将从高中毕业时，职业顾问可能会让她在“图书管理员”和“矿工”之间做出选择，她可能会很合理地说：“我不确定自己更喜欢哪一个。”在这里，不确定性来自她对自己偏好的认知不确定性，比如她对煤尘和书本灰尘的偏好；也来自她不清楚如何才能

在每个职业选择中做出最佳选择的计算上的不确定性；还来自她对世界的一般不确定性，比如她怀疑当地煤矿是否会长期存在。

出于这些原因，通过在描述不完全的选项中做出简单的选择来确定人类的偏好是一个坏主意，这些选择很难评估，并且包含未知的可取性元素。这样的选择提供了潜在偏好的间接证据，但它们不是这些偏好的组成部分。这就是为什么我要提出关于未来生活偏好的概念，例如，想象你可以通过两部不同的电影来快速体验你未来的生活，然后在它们之间表达偏好（见第2章）。思想实验当然不可能在实践中进行，但人们可以想象，在许多情况下，在每一部电影的所有细节都被填充并被充分体验之前，一个清晰的偏好就已经出现了。即使事先给出了情节概要，你也可能事先不知道会喜欢哪一部，但是，基于你现在的身份，实际的问题是有答案的，就像当你品尝榴梿时，你是否喜欢榴梿这个问题会有答案一样。

你可能不确定自己的偏好，但这一事实不会对可证明有益的人工智能基于偏好的方法造成任何特殊问题。事实上，已经有一些算法考虑了罗比和哈里特对哈里特偏好的不确定性，并考虑到哈里特正在了解自己偏好的可能性，罗比也是如此。[41]正如罗比可以通过观察哈里特的行为来减少哈里特偏好的不确定一样，哈里特也可以通过观察她自己对经历的反应来减少对自己偏好的不确定。这两种不确定性不一定是直接相关的，对于哈里特的喜好，罗比也未必比哈里特更不确定。例如，罗比或许能发现哈里特有强烈的遗传倾向，厌恶榴梿的味道。在这种情况下，即使哈里特完全不知情，罗比也不会对她的榴梿偏好有什么不确定性。

如果哈里特不能确定她对未来事件的偏好，那么很可能她也是错的。例如，她可能会确信自己不喜欢榴梿（或者比如说，绿皮鸡蛋和火腿），所以她不惜一切代价避免吃榴梿，但结果可能是，某天有人碰巧在她的水果沙拉里放了一些榴梿，她却发现榴梿是顶级的美味。因此，罗比不能假设哈里特的行为反映了她对自己偏好的准确了解：

其中一些可能完全基于经验，而另一些可能主要基于假设、偏见、对未知的恐惧和缺乏支持的概括。[42]一个机智的罗比会对哈里特很有帮助，提醒她注意这种情况。

经验和记忆

一些心理学家质疑一种观点，这种观点认为，有一种自我的偏好是至高无上的，就像豪尔绍尼的偏好自主原则所暗示的那样。在这些心理学家中，最杰出的是我的前伯克利同事丹尼尔·卡尼曼（Daniel Kahneman）。卡尼曼因其在行为经济学方面的研究获得了2002年的诺贝尔奖，他是在人类偏好这一主题上最有影响力的思想家之一。他的著作《思考，快与慢》[43]详细叙述了一系列实验，这些实验使他确信每个人有两个自我——经验自我和记忆自我，它们的偏好是互相冲突的。

经验自我是快乐测量仪度量的自我，19世纪英国经济学家弗朗西斯·埃奇沃思（Francis Edgeworth）把快乐测量仪想象成“一种理想的完美仪器，一台心理物理机器，可以精确地根据意识的判断，不断记录一个人体验到的快乐的程度”。[44]根据享乐效益主义的说法，任何经验对一个人的整体价值只是经验中每一刻享乐价值的总和。这个概念同样适用于吃一个冰激凌或生活一辈子。

而当人们需要做决定时，记忆自我才是“负责人”。记忆自我基于以前经验的记忆和它们的可取性选择新的经验。卡尼曼的实验表明，记忆自我与经验自我有着截然不同的想法。

最容易理解的是一个让受试者的手浸入冷水中的实验。该实验有两种不同的模式：第一种是让受试者将手放在14摄氏度的水中浸泡60秒；第二种是让受试者将手放在14摄氏度的水中浸泡60秒，然后在15摄氏度的水中浸泡30秒。（这种温度与北加州的海洋温度相似，冷到几乎每个人在水里都穿着潜水服。）所有受试者都表示这是一次不愉

快的经历。在经历了这两种模式之后（顺序任意，中间有7分钟的间隔），受试者被要求选择他们想要重复哪一种模式。绝大多数受试者更喜欢重复60秒+30秒这一种，而不是只浸泡60秒。

卡尼曼认为，从经验自我的角度来看，60秒+30秒的体验肯定比60秒的体验差，因为它包括60秒的不愉快的经历和另一段不愉快的经历。然而记忆自我选择了60秒+30秒这种模式。这是为什么？

卡尼曼的解释是，记忆自我戴着古怪的有色眼镜回顾过去，主要关注“峰值”（最高或最低的享乐值）和“终点值”（体验结束时的享乐值）。经验的不同部分的持续时间大多被忽略。在60秒和60秒+30秒的两种模式中，二者最大不适程度是相同的，但最终的不适程度是不同的：在60秒+30秒的情况下，最终的水温要高1摄氏度。如果记忆自我通过峰值和终点值来评估体验，而不是通过将享乐值随时间累加来评估，那么60秒+30秒的体验更好，这就是我们的发现。“峰值—终点值”模型似乎可以解释有关偏好的文献中许多其他同样奇怪的发现。

卡尼曼似乎对他的发现犹豫不决（这或许是恰当的）。他断言，记忆自我“只是犯了一个错误”，选择了错误的经验，因为它的记忆有缺陷且不完整。他认为这“对相信选择理性的人而言是个坏消息”。另一方面，他写道：“一个忽视人们需求的幸福理论是无法持续的。”例如，假设哈里特品尝过百事可乐和可口可乐，现在她非常喜欢百事可乐，那么根据每次实验期间采集的秘密快乐测量仪读数的累计值去强迫她喝可口可乐就是荒谬的。

事实是，没有任何定律要求我们在不同经验之间的偏好是由不同时刻的享乐价值总和来定义的。诚然，标准的数学模型关注的是奖励总和的最大化，[45]但这种做法的最初动机是为了数学上的方便。技术性假设后来证明，根据累计的奖励做出决定是合理的，[46]但这些技术性假设在现实中未必成立。例如，假设哈里特在以下两个快乐值序列中选择：［10，10，10，10，10］和［0，0，40，0，0］。她很

有可能更喜欢第二个序列；没有数学定律能强迫她根据总和而不是最大值做出选择。

卡尼曼承认，由于预期和记忆在幸福感中起到的关键作用，情况变得更加复杂。一个人的婚礼，一个孩子的出生，用一个下午采黑莓和做果酱，这些单一且愉快的经历的记忆可以让一个人度过多年的苦恼和失望。也许记忆自我不仅评估经验本身，还通过它对未来记忆的影响，来评估它对未来生活价值的总体影响。或许，记忆自我而非经验自我是对将要记忆的内容的最佳评判。

时间和变化

毋庸置疑，生活在21世纪的理智的人不会愿意仿效公元2世纪罗马社会的偏好，那里充斥着为公共娱乐而进行的血腥角斗，以奴隶制为基础的经济，以及对战败国的残酷屠杀。（我们不必纠结于现代社会中与其相似的特征。）随着人类文明的进步，道德标准显然会随着时间的推移不断演变，你如果愿意，也可以说是堕落。这反过来又表明，我们现在对待动物的幸福的态度可能会让我们的后代十分反感。因此重要的是，负责实现人类偏好的机器能够随着时间的推移对这些偏好的变化做出反应，而不是一成不变。第7章中的三个原则以自然的方式适应了这种变化，因为它们需要机器来学习和实现当代人的当前偏好——许多完全不同的人的偏好，而不是一套理想化的偏好或可能早已去世的机器设计者的偏好。[47]

随着历史的推移，人类群体的典型偏好可能会发生变化，因此我们自然会将注意力集中在每个人的偏好是如何形成的，以及成人偏好的可塑性问题上。我们的偏好当然会受到我们的生理特征的影响，例如，我们通常会避免疼痛、饥饿和口渴。然而，我们的生理特征仍然相当稳定，所以剩下的偏好一定受到文化和家庭的影响。孩子们很有可能一直在进行某种形式的逆强化学习，以识别父母和同伴的偏好，

从而解释他们的行为，然后，孩子们将这些偏好作为自己的偏好。即使是成年人，我们的偏好也会受到媒体、政府、朋友、雇主和我们自身经历的影响而演变。例如，第三帝国的许多支持者很可能一开始并不是渴望种族纯洁的种族灭绝者。

偏好变化对个人和社会层面的理性提出了挑战。例如，豪尔绍尼的偏好自主原则是说，每个人都有权享有他所拥有的任何偏好，任何人都不应该触碰它们。然而，偏好并非不可触碰，而是随时都会被一个人的每一次经历所触碰和修改。机器会不由自主地修改人类的偏好，因为机器会改变人类的经验。

虽然区分偏好改变和偏好升级有时很难，但是这很重要，区分发生在这样一种时候：最初充满不确定的哈里特通过经验来学习更多她自己的偏好。偏好升级可以填补自我认知的空白，或许还可以为先前拥有的微弱且暂时的偏好增加确定性。另外，偏好改变并不是一个过程，不是“关于一个人的偏好实际上是什么”的额外证据。在极端的情况下，你可以把它想象成药物治疗甚至脑部手术的结果，它发生在我们可能不理解或不认同的过程中。

偏好改变是存在问题的，这至少有两个原因。第一个是，我们不清楚哪些偏好应该在做决定时起主导作用：是哈里特在做出决定时的偏好，还是她做出决定的事件发生期间和之后将拥有的偏好。例如，在生物伦理学中，这是一个非常现实的两难境地，因为人们在患重病后对医疗干预和临终关怀的偏好确实会发生巨大的变化。[48]假设这些变化不是由智力下降造成的，那么谁的偏好应该得到尊重呢？[49]

第二个是，似乎没有明显的理性基础来改变（而不是更新）一个人的偏好。如果哈里特喜欢A甚于B，但她可以选择经历一段她知道会使她喜欢B甚于A的经历，她为什么要这样做？结果她会选择B，而这是她目前不想要的。

在“尤利西斯和塞壬”的传说中，偏好改变以戏剧化的形式出现。塞壬是神话中的生物，它们的歌声会蛊惑水手走向他们的噩运，船舶会撞到地中海某些岛屿的岩石，然后沉没。尤利西斯想听这首歌，他命令他的水手们用蜡塞住耳朵，把他绑在桅杆上，然而在任何情况下，他们都不会服从他随后提出的释放他的请求。很明显，他希望水手们尊重他最初的偏好，而不是塞壬蛊惑他之后的偏好。这个传说成了挪威哲学家乔恩·埃尔斯特（Jon Elster）的一本书的书名，[50]这本书探讨了意志的弱点以及对理性理论的其他挑战。

为什么智能机器会故意改变人类的偏好？答案很简单：让偏好更容易被满足。我们在第1章的社交媒体点击率优化案例中看到了这一点。一种回应可能是，机器必须将人类的偏好视为神圣不可侵犯的，不能允许任何事情改变人类的偏好。不幸的是，这是完全不可能的。一个有用的机器人助手的存在，很可能会对人类的偏好产生影响。

一种可能的解决方案是让机器学习人类的“元偏好”，即关于哪种类型的“偏好改变过程”可以接受或不可接受的偏好。请注意，此处使用的是“偏好改变过程”，而不是“偏好改变”。这是因为想要一个人的偏好朝着特定方向改变，往往等于已经拥有了那个偏好。在这种情况下，我们真正需要的是更好地贯彻偏好的能力。例如，如果哈里特说，“我想改变我的偏好，那样我就不像现在这样这么想吃蛋糕了”，那么她就已经有了未来少吃蛋糕的偏好，她真正想要的是改变她的认知结构，让她的行为更能反映这种偏好。

我所说的“关于哪种类型的偏好改变过程可以接受或不可接受的偏好”，是想表达这样一种观点，例如，一个人可以通过周游世界来体验各种文化，或者通过参加一个充满活力的知识分子社区来探索广泛的道德传统，又或者通过留出一些隐逸的时间来反思和认真思考生活及其意义，最终获得“更好的”偏好。我将这些过程称为“中性偏好”，因为人们并不认为这个过程会在任何特定方向上改变自己的偏好，与此同时，我意识到有些人可能会强烈反对这种描述方法。

当然，并非所有中性偏好都是可取的，例如，很少有人期望通过敲打自己的脑袋来培养“更好的”偏好。让自己接受一个可接受的偏好改变过程，就像进行一个实验来了解世界是如何运行的：你永远不会事先知道实验的结果，但尽管如此，你还是希望在新的精神状态下过得更好。

修正偏好有可接受的途径，这一观点似乎与改变行为有可接受的方法的观点有关。例如，雇主可以通过设计选择情境，让人们对“为退休而存钱”做出“更好”的选择。这通常可以通过操纵影响选择的“非理性”因素来实现，而不是通过限制选择或对糟糕的选择“征税”来实现。经济学家理查德·塞勒（Richard Thaler）和法律学者卡斯·桑斯坦（Cass Sunstein）合著的《助推：如何做出有关健康、财富与幸福的最佳决策》一书提出了各种可以接受的方法和机会，以求“影响人们的行为，使他们过上更长寿、更健康、更美好的生活”。

我们目前还不清楚“行为改变方法”是否真的只是在改变行为。如果当助推被去掉时，改变后的行为持续存在——这可能是此类干预的预期结果，那么个体的认知结构（将潜在偏好转化为行为的东西）或个体的潜在偏好就会发生变化，看起来很可能二者兼而有之。然而，显而易见的是，助推策略假设每个人都偏爱“更长寿、更健康、更美好”的生活；每一次助推都基于对“更好”生活的特定定义，这似乎违背了偏好自主的原则。相反，设计中性偏好的辅助过程可能会更好，它帮助人们使他们的决策和认知结构更符合他们的潜在偏好。例如，我们可以设计认知助手来强调决策的长期后果，并教人们认识到当前这些后果的根源来自何方。[51]

我们需要更好地理解人类偏好的形成和塑造过程，这是显而易见的，因为这样的理解将帮助我们设计机器，避免社交媒体内容选择算法造成的人类偏好的意外变化和不良变化。当然，有了这种理解，我们就会忍不住去设计一些改变，让世界变得“更美好”。

有人可能会认为，我们应该为中性偏好的“改善”提供更多的机会，例如旅行、辩论以及培养分析和批判性思维。比如，我们可以为每个高中生提供机会，让他们至少在两种不同于他们自己的文化中生活几个月。

然而，几乎可以肯定的是，我们会想更进一步，例如，通过实施社会和教育改革，提高利他主义系数，即每个人对他人幸福的重视程度，同时降低施虐、骄傲和嫉妒的系数。这是个好主意吗？我们应该动员我们的机器来辅助这个过程吗？这当然很诱人。的确，亚里士多德写道：“政治的主要任务是在公民身上培养出某种品质，让他们善良，并能够拥有高尚的行为。”可以这么说，在全球的范围内开展有意的偏好工程是存在风险的。我们应该极度谨慎地进行。

---

[1] 一些人认为生物学和神经科学也直接相关。请参见：Gopal Sarma, Adam Safron, and Nick Hay, “Integrative biological simulation, neuropsychology, and AI safety,” arxiv.org/abs/1811.03493 (2018)。

[2] 关于让计算机承担损害赔偿责任的可能性：Paulius Čerka, Jurgita Grigien3e,and Gintar3e Sirbikyt3e, “Liability for damages caused by artificial intelligence,” *Computer Law and Security Review* 31 (2015): 376–89。

[3] 关于标准道德理论及其对设计人工智能系统的影响的介绍，请参见：WendellWallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong*(Oxford University Press, 2008)。

[4] 效益主义思想的原始资料：Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation* (T. Payne & Son, 1789)。

[5] 穆勒对其导师边沁思想的阐述对自由主义思想产生了非凡的影响：John Stuart Mill, *Utilitarianism* (Parker, Son & Bourn, 1863)。

[6] 介绍偏好效益主义和偏好自主权的论文：John Harsanyi, “Morality and the theory of rational behavior,” *Social Research* 44 (1977): 623–56。

[7] 当代表多个个体决策时，通过效用加权和进行社会聚合的一个论点：John Harsanyi, “Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility,” *Journal of Political Economy* 63 (1955): 309–21。

[8] 豪尔绍尼的社会聚合定理对不平等先验信念情况的推广：Andrew Critch,Nishant Desai, and Stuart Russell, “Negotiable reinforcement learning for Pareto optimal sequential decision-making,” in *Advances in Neural Information Processing Systems* 31, ed. Samy Bengio et al. (2018)。

[9] 理想效益主义的原始资料：G. E. Moore, *Ethics* (Williams & Norgate, 1912)。

[10] 引用斯图尔特·阿姆斯特朗误导效用最大化的生动案例的新闻报道：Chris Matyszczyk, “Professor warns robots could keep us in coffins on heroin drips,”CNET, June 29, 2015。

[11] 波普尔的消极效益主义理论（后来由斯马特命名）：Karl Popper, *The Open Society and Its Enemies* (Routledge, 1945)。

[12] 对消极效益主义的驳斥：R. Ninian Smart, “Negative utilitarianism,” *Mind* 67(1958): 542–43。

[13] 关于“终结人类苦难”指令引起风险的典型论点，请参见：“Why do we thinkAI will destroy us?,” Reddit, reddit.com/r/Futurology/comments/38fp6o/why_do_we_think_ai_will_destroy_us。

[14] 人工智能中自我欺骗动机的一个很好的来源：Ring and Orseau, “Delusion,survival, and intelligent agents”。

[15] 论效用的人际比较的不可能性：W. Stanley Jevons, *The Theory of Political Economy* (Macmillan, 1871)。

[16] 效用怪兽出现在：Robert Nozick, *Anarchy, State, and Utopia* (Basic Books, 1974)。

[17] 例如，我们可以将立即死亡的效用设为0，将最幸福的生活的效用设为1。请参见：John Isbell, “Absolute games,” in *Contributions to the Theory of Games*,vol. 4, ed. Albert Tucker and R. Duncan Luce (Princeton University Press, 1959)。

[18] 蒂姆·哈福德（Tim Harford）讨论了灭霸人口减半政策的过于简单化的本质，“Thanos shows us how not to be an economist,” *Financial Times*, April 20, 2019。甚至在这部电影首映之前，灭霸的捍卫者们就开始聚集在社交媒体Subreddit的r/thanosdidnowrong/上。为了与Subreddit的座右铭保持一致，70万名成员中的35万名后来被清除。

[19] 关于不同规模人口的效用：Henry Sidgwick, *The Methods of Ethics* (Macmillan,1874)。

[20] 效益主义思想令人反感的结论和其他棘手问题：Derek Parfit, *Reasons and Persons* (Oxford University Press, 1984)。

[21] 关于人口伦理公理化方法的简明摘要，请参见：Peter Eckersley, “Impossibility and uncertainty theorems in AI value alignment,” in *Proceedings of the AAAI Workshop on Artificial Intelligence Safety*, ed. Huáscar Espinoza et al. (2019)。

[22] 计算地球的长期承载能力：Daniel O’Neill et al., “A good life for all within planetary boundaries,” *Nature Sustainability* 1 (2018): 88–95。

[23] 关于道德不确定性在人口伦理学中的应用，请参见：Hilary Greaves and Toby Ord, “Moral uncertainty about population axiology,” *Journal of Ethics and Social Philosophy* 12 (2017): 135–67。更全面的分析请参见：Will MacAskill,Krister Bykvist, and Toby Ord, *Moral Uncertainty* (Oxford University Press,forthcoming)。

[24] 这表明亚当·斯密并不像通常想象的那样痴迷于自私的引用：Adam Smith,*The Theory of Moral Sentiments* (Andrew Millar; Alexander Kincaid and J. Bell,1759)。

[25] 有关利他主义经济学的介绍，请参见：Serge-Christophe Kolm and Jean Ythier, eds., *Handbook of the Economics of Giving, Altruism and Reciprocity*, 2 vols.(North-Holland, 2006)。

[26] 关于慈善是自私的论述：James Andreoni, “Impure altruism and donations to public goods: A theory of warm-glow giving,” *Economic Journal* 100 (1990): 464–77。

[27] 写给那些喜欢方程式的人：用w_A衡量爱丽丝的内在幸福感，用w_B衡量鲍勃的内在幸福感。那么爱丽丝和鲍勃的效用定义如下：

$$U_A = w_A + C_{AB}\, w_B$$
$$U_B = w_B + C_{BA}\, w_A.$$

一些作者认为，爱丽丝关心鲍勃的整体效用U_B，而不仅仅是他的内在幸福感w_B，但这导致了一种循环，因为爱丽丝的效用取决于鲍勃的效用，而鲍勃的效用又取决于爱丽丝的效用；人们有时可以找到稳定的解决方案，但潜在的模型可能会受到质疑。请参见：Hajime Hori, “Nonpaternalistic altruism and functional interdependence of social preferences,” *Social Choice and Welfare* 32(2009): 59–77。

[28] 每个人的效用是每个人的幸福的线性组合这种模型只是一种可能性。更通用的模型是可能的，例如，有些人倾向于避免在福利分配上的严重不平等，甚至不惜以减少总福利为代价，而还有些人则倾向于所有人都喜欢完全的平等。因此，我提出的总体方法容纳了个人持有的多种道德理论；同时，它不坚持认为这些道德理论中的任何一种都是正确的，也不应

该对持有不同理论的人的结果产生很大的影响。我要感谢托比·奥德指出了这种方法的这一特点。

[29] 这类观点反对旨在确保结果平等的政策，尤其是美国法律哲学家罗纳德·德沃金（Ronald Dworkin）提出的观点。请参见：Ronald Dworkin, “What is equality? Part 1: Equality of welfare,” *Philosophy and Public Affairs* 10 (1981):185–246。我非常感谢伊森·加布里埃尔（Iason Gabriel）提供的这一参考。

[30] 以报复为基础对犯罪行为进行惩罚肯定是一种普遍趋势。虽然它在保持社区成员的秩序方面发挥了社会作用，但它可以被一种由威慑和预防驱动的同等有效的政策所取代，即权衡惩罚违法者对更大的社会利益所造成的内在伤害。

[31] 假设E_{AB}和P_{AB}分别是爱丽丝的嫉妒系数和骄傲系数，并假设它们适用于幸福感的差异。那么爱丽丝的效用公式如下（有点过于简化）：

$$U_A = w_A + C_{AB}\, w_B - E_{AB}\,(w_B - w_A) + P_{AB}\,(w_A - w_B)$$
$$= (1 + E_{AB} + P_{AB})\, w_A + (C_{AB} - E_{AB} - P_{AB})\, w_B.$$

因此，如果爱丽丝的骄傲系数和嫉妒系数为正，它们就像虐待和恶意系数一样影响鲍勃的幸福感：在其他所有条件相同的情况下，如果鲍勃的幸福感降低，爱丽丝会更快乐。实际上，骄傲和嫉妒通常不适用于幸福感的差异，而适用于有形方面的差异，比如地位和财产。鲍勃在获取他的财产时的辛苦劳作（这降低了他的整体幸福感）可能不会被爱丽丝看到。这可能导致因为“攀比”而产生的适得其反的行为。

[32] 关于炫耀性消费的社会学：Thorstein Veblen, *The Theory of the Leisure Class: An Economic Study of Institutions* (Macmillan, 1899)。

[33] Fred Hirsch, *The Social Limits to Growth* (Routledge & Kegan Paul, 1977).

[34] 我感谢齐亚德·马拉尔（Ziyad Marar）向我指出社会认同理论及其在理解人类动机和行为方面的重要性。请参见：Dominic Abrams and Michael Hogg,eds., *Social Identity Theory: Constructive and Critical Advances* (Springer, 1990)。有关主要思想的简要概述，请参见：“Social identity,” in *This Idea Is Brilliant: Lost, Overlooked, and Underappreciated Scientific Concepts Everyone Should Know*, ed. John Brockman (Harper Perennial, 2018)。

[35] 在这里，我并不是说我们一定需要对认知的神经实现有一个详细的了解；我们需要的是一个“软件”层面的模型，来研究偏好（无论是显性的还是隐性的）是如何产生行为的。这样的模型需要纳入关于奖励制度的已知内容。

[36] Ralph Adolphs and David Anderson, *The Neuroscience of Emotion: A New Synthesis* (Princeton University Press, 2018).

[37] 请参见：Rosalind Picard, *Affective Computing*, 2nd ed. (MIT Press, 1998)。

[38] 对榴梿的赞美喜爱之情：Alfred Russel Wallace, *The Malay Archipelago: The Land of the Orang-Utan, and the Bird of Paradise* (Macmillan, 1869)。

[39] 对榴梿的不太乐观的看法：Alan Davidson, *The Oxford Companion to Food*(Oxford University Press, 1999)。由于榴梿散发出强烈的气味，大楼中的人被疏散，飞机在飞行途中掉头。

[40] 我在写完这一章后发现，洛里·保罗（Laurie Paul）在*Transformative Experience* (Oxford University Press, 2014)一书中也将榴梿用于同样的哲学目的。保罗认为，对自己偏好的不确定性会给决策理论带来致命的问题，这一观点与理查德·佩蒂格鲁（Richard Pettigrew）的观点相矛盾，请参见Richard Pettigrew,“Transformative experience and decision theory,” *Philosophy and Phenomenological Research* 91 (2015): 766–74。两位作者都没有提到豪尔绍尼的早期工作，“Games with incomplete information, Parts I–III,” or Cyert and de Groot,“Adaptive utility”。

[41] 一篇关于帮助不知道自己偏好并正在了解自己偏好的人的论文：Lawrence Chan et al., “The assistive multi-armed bandit,” in *Proceedings of the 14th ACM/ IEEE International Conference on Human-Robot Interaction (HRI)*, ed. David Sirkin et al. (IEEE, 2019)。

[42] 埃利泽·尤德考斯基在*Coherent Extrapolated Volition* (Singularity Institute,2004)中，将所有这些方面以及明显的不一致性集中在“混淆”的标题下，不幸的是，这个词还没有流行起来。

[43] 对评估经验的两个自我的论述：Daniel Kahneman, *Thinking, Fast and Slow*(Farrar, Straus & Giroux, 2011)。

[44] 埃奇沃思的快乐测量仪，一种每时每刻测量幸福的假想装置：Francis Edgeworth, *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences* (Kegan Paul, 1881)。

[45] 关于不确定条件下的顺序决策的标准文本：Martin Puterman, *Markov Decision Processes*: *Discrete Stochastic Dynamic Programming* (Wiley, 1994)。

[46] 关于证明效用随时间推移而累加的公理假设：Tjalling Koopmans, “Representation of preference orderings over time,” in *Decision and Organization*, ed. C.Bartlett McGuire, Roy Radner, and Kenneth Arrow (North Holland, 1972)。

[47] 2019年的人类（2099年，他们可能早已谢世，或者可能只是2099年人类的早期自我）可能希望以一种尊重2019年的人类在2019年的偏好的方式建造机器，而不是迎合2099年

的人类的肤浅和考虑不周的偏好。这就好比起草了一部不允许任何修改的宪法。如果2099年的人类经过适当的深思熟虑后决定推翻2019年人类固有的偏好，那么他们似乎可以做到这一点。总之，他们和他们的后代必须承担后果。

[48] 我感谢温德尔·瓦拉克（Wendell Wallach）的观察。

[49] 一篇关于偏好随时间变化的早期论文：John Harsanyi, “Welfare economics of variable tastes,” *Review of Economic Studies* 21 (1953): 204–13。一项近期的（也有些技术性的）调查请参见：Franz Dietrich and Christian List, “Where do preferences come from?,” *International Journal of Game Theory* 42 (2013):613–37. See also Laurie Paul, *Transformative Experience* (Oxford University Press, 2014), and Richard Pettigrew, “Choosing for Changing Selves,” philpapers.org/archive/PETCFC.pdf。

[50] 关于非理性的理性分析，请参见：Jon Elster, *Ulysses and the Sirens: Studies in Rationality and Irrationality* (Cambridge University Press, 1979)。

[51] 有关人类认知假体的有前途的想法，请参见：Falk Lieder, “Beyond bounded rationality: Reverse-engineering and enhancing human intelligence” (PhD thesis,University of California, Berkeley, 2018)。


10
问题
解决了吗?



如果我们成功地创造了可证明有益的人工智能系统，那么我们将消除超级智能机器失控的风险。人类可以继续发展，并获得几乎无法想象的好处，这些好处来自运用更强大的智能来推进我们文明的能力。我们将从数千年的农业、工业和文书工作的奴役中解放出来，自由地充分发挥生命的潜能。我们站在那个黄金时代的制高点回顾我们现在的生活，我们现在的生活就会显得像英国哲学家托马斯·霍布斯（Thomas Hobbes）想象的没有政府的生活一样：孤独、贫穷、肮脏、野蛮又短暂。

也许未来并非如此。“邦德式”的恶棍可能会绕过我们的保护措施，释放出无法控制的超级智能，而人类对此毫无防御能力。如果幸存下来，那么我们可能会发现，随着我们越来越多地将知识和技能托付给机器，我们会逐渐变得衰弱。机器可能会建议我们不要这样做，因为它们理解人类自主的长期价值，但我们可能会否决它们。

有益的机器

20世纪的许多技术背后的基础标准模型，都依赖于对外部提供的固定目标进行优化的机器。正如我们已经看到的，这种模式存在根本性的缺陷。只有在目标确保完整且正确，或者可以很容易地重置机器的情况下，这种模式才有效。随着人工智能变得越来越强大，这两个条件都将不成立。

如果外部提供的目标可能是错误的，那么机器表现得似乎总是正确的就没有任何意义。因此，我建议使用有益的机器：其行为预期可以实现我们的目标的机器。因为这些目标在我们心中，而不在它们内部，所以机器需要通过观察我们做出的选择和如何做出选择来更多地了解我们真正想要的是什么。以这种方式设计的机器会顺从人类：它们会征求许可；当指令不明确时，它们会谨慎行事；而且它们会允许自己被关掉。

虽然这些初步的结果是一种简单化和理想化的设置，但我相信它们将在过渡到更现实的设置后被保存下来。我的同事们已经成功地将同样的方法应用于实际问题，例如无人驾驶汽车与人类驾驶员的互动。[1]例如，众所周知，无人驾驶汽车在不清楚谁有优先通行权的情况下不擅长处理通过十字路口的情况。然而，通过把这种情况描述成一个辅助博弈，汽车想出了一个新颖的解决方案：它后退一点，表明它绝对不打算先走。人类理解了这个信号，确信不会发生碰撞，于是继续前进。显然，我们人类专家本可以想到这个解决方案，并将其编到车辆的程序中，但事实并非如此，这是一种完全由交通工具自己发明的交流方式。

随着我们在其他环境中获得更多的经验，我预计我们会对机器与人类交互时的行为范围和流畅性感到惊讶。我们已经习惯了机器的愚蠢，它们执行死板的、预编程的行为，或者追求虽明确但不正确的目标，现在我们可能会震惊于它们何时变得如此明智。可证明有益的机器技术是人工智能新方法的核心，也是人与机器之间新关系的基础。

同样，从普通的软件系统开始，将类似的想法应用于重新设计其他为人类服务的“机器”，似乎也是可能的。我们被教导通过编写子程序来构建软件，每个子程序都有一个定义明确的规范，说明任意给定输入的输出应该是什么，就像计算器上的平方根按键一样明确。该规范是对给定人工智能系统的目标的直接模拟。在产生符合规范的输出之前，子程序不应该终止或将控制返回给软件系统的更高层。（这应该会提醒你，人工智能系统坚持一心一意地追求给定的目标。）更好的方法是考虑规范中的不确定性。例如，执行一些极其复杂的数学计算的子程序，通常会被赋予一个误差界限，该界限定义了答案所需的精度，子程序必须返回在该误差界限内正确的解。有时，这可能需要数周的计算。相反，对允许的误差界限不设置得那么精确可能会更好，这样子程序可以在20秒后返回并说：“我找到了一个不错的解决方案。这样可以吗？还是要我继续找？”在某些情况下，该问题可能

会一直渗透到软件系统的顶层，以便人类用户可以为系统提供进一步的指导。人类的回答将有助于完善各个层面的规范。

同样的想法也适用于政府和公司等实体。政府的明显缺陷是：过度关注政府官员的偏好（财政和政治上的偏好），而对被统治者的偏好关注太少。偏好本应可以通过选举传达给政府，但对于如此复杂的任务，政府的“带宽”似乎非常小（大约每隔几年才得到一个字节的信息）。在大多数国家里，政府仅仅是一群人把自己的意志强加给其他人的一种手段。而公司则不遗余力地了解客户的偏好，无论是通过市场调查还是以购买决策的形式直接反馈。另外，通过广告、文化影响甚至化学物质上瘾来塑造人类的偏好，也是被接受的经商方式。

对人工智能的治理

人工智能有重塑世界的能力，而对重塑的过程，我们必须以某种方式加以管理和引导。如果发展有效治理人工智能的大量举措可以作为一种指导，那么我们的状态非常好。许多人都在成立委员会、理事会或国际小组。世界经济论坛已经分别确定了近300项为人工智能制定的道德原则。我从电子邮箱就可以整理出一份长长的邀请清单，人们邀请我参加“关于新兴人工智能技术对社会和道德产生影响的未来国际治理的全球高峰论坛”。

人工智能与核技术的情况截然不同。第二次世界大战后，美国掌握了全部核技术。1953年，美国总统德怀特·艾森豪威尔向联合国提议成立一个国际机构来监管核技术。1957年，国际原子能机构开始运转，它是全球唯一的监督核能安全与有益发展的机构。

相比之下，很多国家手里都掌握着人工智能。可以肯定的是，很多国家都资助了大量人工智能研究，但几乎所有的研究都发生在安全的国家实验室之外。大学里的人工智能研究人员是一个广泛的、合作的国际社会的一部分，通过共同的兴趣、会议、合作协议和专业协会

联系在一起，如AAAI（美国人工智能协会）和IEEE（电气和电子工程师协会，其中包括成千上万名人工智能研究人员和从业者）。也许现在大部分人工智能研发投资都发生在大大小小的公司里。截至2019年，这些公司有美国的谷歌（包括DeepMind）、脸书、亚马逊、微软和IBM，以及中国的腾讯、百度，在某种程度上还有阿里巴巴，它们都是全球最大的公司。[2]除了腾讯和阿里巴巴，其他所有公司都是“人工智能合作伙伴组织”的成员。这是一个行业联盟，其宗旨之一是承诺在人工智能安全方面进行合作。最后，尽管绝大多数人几乎不具备人工智能方面的专业知识，但参与者至少有一种表面上的意愿，即愿意考虑人类的利益。

因此，这些就是掌握着大部分人工智能技术的玩家。它们的利益并不完全一致，但它们都有一个共同的愿望，即随着人工智能系统变得更加强大，它们都希望保持对人工智能系统的控制。（其他目标，如避免大规模失业，是政府和大学研究人员的共同目标，但这未必是那些希望在短期内从尽可能广泛的人工智能部署中获利的公司的目标。）为了巩固这一共同利益并实现协调一致的行动，拥有召集力的组织出现了，这大致意味着，如果这些组织召开会议，人们会受邀参加。除了能够将人工智能研究人员聚集在一起的专业协会，以及将企业和非营利性机构结合在一起的“人工智能合作伙伴组织”之外，典型的召集者是联合国（面向政府和研究人员）和世界经济论坛（面向政府和企业）。此外，七国集团（G7）还提议成立一个国际人工智能小组，希望它能发展为类似联合国政府间气候变化专门委员会的机构。听上去举足轻重的报道层出不穷。

通过所有这些活动，对人工智能的治理是否取得了实际进展？或许令人惊讶的是，答案是肯定的，对人工智能的治理至少具备发展的潜力。世界各地的许多政府都在为自己配备咨询机构，以帮助其制定法规，或许最突出的例子是欧盟的人工智能高级专家组。针对用户隐私、数据交换和避免种族偏见等问题的协议、规则和标准开始出现。

政府和企业正在努力为无人驾驶汽车制定规则，这些规则将不可避免地包含跨境元素。有一个共识是，如果人工智能系统是可信的，那么人工智能决策必须是可以解释的，这一共识已经在欧盟的GDPR条例中得到部分实施。在加利福尼亚州，一项新法律禁止人工智能系统在某些情况下模仿人类。最后这两项——可解释和模仿问题，无疑与人工智能的安全和控制问题有关。

目前，在考虑维持对人工智能系统的控制问题上，我们尚无可实施的建议可以向政府或其他组织提出。“人工智能系统必须安全可控”这样的规定没有任何影响力，因为其中的术语还没有确切的含义，也因为没有广为人知的工程方法来确保安全性和可控性。但是，让我们乐观一点，想象一下，几年之后，通过数学分析和以实用的应用程序的形式出现的实践，“可证明有益的人工智能”的有效性将被建立起来。例如，我们可能拥有个人数字助理，我们可以信任它们使用我们的信用卡，筛选我们的电话和电子邮件，管理我们的财务，因为它们已经适应了我们的个人偏好，知道什么时候可以继续，什么时候最好寻求指导。我们的无人驾驶汽车可能已经学会了与彼此以及与人类司机互动，我们的家用机器人甚至可以与最倔强的蹒跚学步的幼儿顺畅地交流。运气好的话，没有猫会被烤熟当晚餐，也没有鲸肉会被提供给绿党成员享用。

在这一点上，指定软件设计模板或许是可行的，各种应用程序必须符合这些模板才能销售或连接到互联网，就像应用程序必须通过多项软件测试才能在苹果的App Store商店或谷歌的Google Play商店销售一样。软件供应商可以提出额外的模板，只要他们能证明这些模板满足（当时定义良好的）安全性和可控性的要求。人们有报告问题和更新产生不良行为的软件系统的机制。人们围绕可证明安全的人工智能程序的理念，创建专业的行为准则，并将相应的定理和方法整合到课程中，以满足有抱负的人工智能和机器学习实践者的需求，这也是有意义的。

对于一位经验丰富的硅谷观察家来说，这听起来可能有些天真。硅谷强烈反对任何形式的监管。尽管制药公司在向公众发布新药之前必须通过临床试验，以证明其药品的安全性和（有益的）药效，这已成为惯例，但软件行业的运作则遵循一套不同的规则——没有规则。一家软件公司里“一群喝红牛饮料的家伙”[3]可以在没有任何第三方监督的情况下发布或升级影响数十亿人的产品。

然而不可避免的是，科技行业不得不承认其产品至关重要。如果它们很重要，那么产品不会产生有害影响也很重要。这意味着将会出现一些规则来管理科技产品与人们的互动，例如禁止那些持续操纵偏好或导致上瘾行为的设计。我毫不怀疑，从一个不受监管的世界过渡到一个受监管的世界将是一个痛苦的过程。我们希望科技行业不需要切尔诺贝利事故那样（或者更糟糕）的灾难来克服行业的阻力。

滥用

对软件行业而言，监管可能是痛苦的，但那些在自己的地下堡垒里密谋统治世界的邪恶博士令人无法容忍。毫无疑问，犯罪分子、恐怖分子和流氓国家将有动机绕过智能机器设计上的任何限制，让智能机器能够用来控制武器，或是策划和实施犯罪活动。危险并不在于邪恶的计划会得逞，而在于这些计划也会因为失去对智能系统的控制而失败，尤其是那些满怀邪恶目标并可以使用武器的计划。

这不是逃避管制的理由，毕竟我们有禁止谋杀的法律，尽管它们经常被绕开。然而，这确实造成了非常严重的治安问题。我们已经在与恶意软件和网络犯罪的斗争中失败了。（最近的一份报告估计，其受害者超过20亿人，每年造成的损失约6 000亿美元。[4]）以高度智能程序的形式存在的恶意软件将更加难以战胜。

包括尼克·波斯特洛姆在内的一些人提出，我们要使用我们自己的、有益的超级人工智能系统来检测和摧毁任何怀有恶意或行为不端的人工智能系统。当然，我们应该使用可用的工具，同时尽量减少对个人自由的影响，但是人类蜷缩在掩体中，完全无力抵挡超级智能所释放出的巨大力量的情景很难让人安心，即便其中一些超级智能站在我们这一边。人们最好能找到方法，将恶意的人工智能扼杀在萌芽状态。

良好的开端是成功地、协调一致地开展打击网络犯罪的国际运动，包括扩展《布达佩斯网络犯罪公约》所涵盖的内容。这将形成一个组织模板，用于防止未来出现失控的人工智能程序。与此同时，这将产生一种广泛的文化理解：从长远来看，有意或无意地建立这样的程序是一种类似于创建流行病有机体的自杀行为。

衰弱与人类自主

E. M.福斯特（E. M. Forster）的两部最著名的小说《霍华德庄园》和《印度之行》展现了20世纪早期的英国社会及其阶级制度。1909年，他写了一本著名的科幻小说《大机器停止》。这个故事因其先见之明而引起了关注，书中描述了（我们现在的）互联网、视频会议、iPad（平板电脑）、大规模开放在线课程（MOOC）、普遍的肥胖，以及回避面对面接触。书名中的大机器是一个包罗万象的智能基础设施，可以满足所有人的需求。人类变得越来越依赖它，但对它的工作原理了解得越来越少。工程知识让位于仪式化的咒语，这些咒语最终无法阻止机器运转逐渐恶化的局面。主角库诺看到了正在发生的事情，却无力阻止：

> 你难道不明白……我们就要死了，在我们这里，唯一真正活着的是机器？我们创造了机器，让它按照我们的意旨办事，可是

我们现在没法让它实现我们的意旨了。它夺走了我们的空间感和触觉，它模糊了一切人际关系，它使我们身体瘫痪、意志瘫痪……我们只是作为流经它动脉血管的血细胞而存在，如果它没有我们也能运转，它会让我们死的。唉，我没有补救办法——或者至少说，只有一种办法——一遍遍地告诉人们，我看见了威塞克斯的山峦，就像阿尔弗雷德[5]推翻丹麦人的统治时所看见的一样。

地球上曾经生活过超过1 000亿人。他们（我们）已经花费了大约1万亿人年的时间学习和教学，才使我们的文明得以延续。到目前为止，文明延续的唯一可能就是在新一代的头脑中进行再创造。（纸作为一种传播方式是不错的，但记录在纸上的知识在传到下一个人的头脑里之前，纸什么也做不了。）现在这种情况正在改变：我们越来越有可能把我们的知识放到机器里，让机器自己来运行我们的文明。

把我们的文明传给下一代的实际动机一旦消失，这个过程就很难逆转了。从现实意义上讲，大约1万亿人年的累积学习将会丢失。我们将成为一艘由机器驾驶的游轮上的乘客，这艘游轮将会永远航行下去，正如电影《机器人总动员》所设想的那样。

一名优秀的结果主义者会说：“显然，这是过度使用自动化所产生的不良后果！设计合理的机器永远不会这样做！”不错，但是请想想这意味着什么。机器可以理解人类的自主性和能力，这是我们的生活偏好中一个重要的方面。机器很可能坚持认为，人类应该对自己的幸福保留控制权和责任，换言之，机器会说不。但我们这些目光短浅的、懒惰的人类可能不同意。这里会发生“公地悲剧”：对任何一个人而言，通过多年艰苦的学习来获得机器已经拥有的知识和技能似乎毫无意义，但如果每个人都这样想，那么人类将会集体失去自主性。

这个问题的解决方案似乎是文化上的，而不是技术上的。我们需要一场文化运动来重塑我们的理想以及对自主、能动性和能力的偏

好，远离自我放纵和依赖。你如果喜欢，也可以称之为古代斯巴达尚武精神的现代文化版本。这意味着在全球范围内发展人类“偏好工程学”，同时我们的社会运作方式也将发生根本性的变化。为了避免让糟糕的情况恶化，无论是在解决方案的设计上，还是在为每个人寻求利益均衡的实际过程中，我们可能都需要超级智能机器的帮助。

任何一个当过父母的人都熟悉这个过程。一旦孩子走出了无助的阶段，育儿就需要在“为孩子包办每件事”和“完全让孩子自己做主”之间保持一种不断发展的平衡。在某个阶段，孩子开始明白，父母完全有能力给他系鞋带，却选择不系。这就是人类的未来吗，永远像孩子一样被超级机器对待？我想不是。孩子无法关掉父母的电源。（谢天谢地！）我们也不会成为宠物或动物园里的动物。在我们现在的世界里，真的没有什么能和“未来有益的智能机器与我们的关系”相提并论。结局如何，还有待观察。

---

[1] 关于辅助博弈在驾驶中的应用：Dorsa Sadigh et al., “Planning for cars that coordinate with people,” *Autonomous Robots* 42 (2018): 1405–26。

[2] 奇怪的是，苹果没有出现在这份名单中。苹果确实有一个人工智能研究小组，并且正在迅速发展。它的传统保密文化意味着，迄今为止它对思想市场的影响相当有限。

[3] Max Tegmark, interview, *Do You Trust This Computer?*, directed by Chris Paine,written by Mark Monroe (2018).

[4] 关于评估网络犯罪的影响：“Cybercrime cost $600 billion and targets banksfirst,” *Security Magazine*, February 21, 2018。

[5] 阿尔弗雷德（849—899），威塞克斯国王，曾战胜入侵的丹麦人，结束了他们对英格兰的统治。——译者注

附录A

寻找解决方案

通过前向搜索，通过考虑各种可能的动作序列的结果，来选择动作，是智能系统的基本能力。每当你向手机问路时，手机就会这么做。图14是一个典型的例子：从当前位置19号码头到达目的地柯伊特塔。算法需要知道它可以执行哪些动作，通常，对于地图导航，每个动作都要经过连接两个相邻十字路口的路段。在这个例子中，从19号码头出发，这里只有一个动作：沿着内河码头行驶到下一个交叉路口。然后会出现一个选择：要么继续往前走，要么向左转进入炮台街。该算法系统地探索所有这些可能性，直到最终找到一条路线。通常，我们会添加一点常识性的指导，比如：去探索接近目标而不是远离目标的街道。有了这个指导和其他一些技巧，该算法就可以快速找到最优解，通常只需要花几毫秒，即使是越野旅行也是如此。

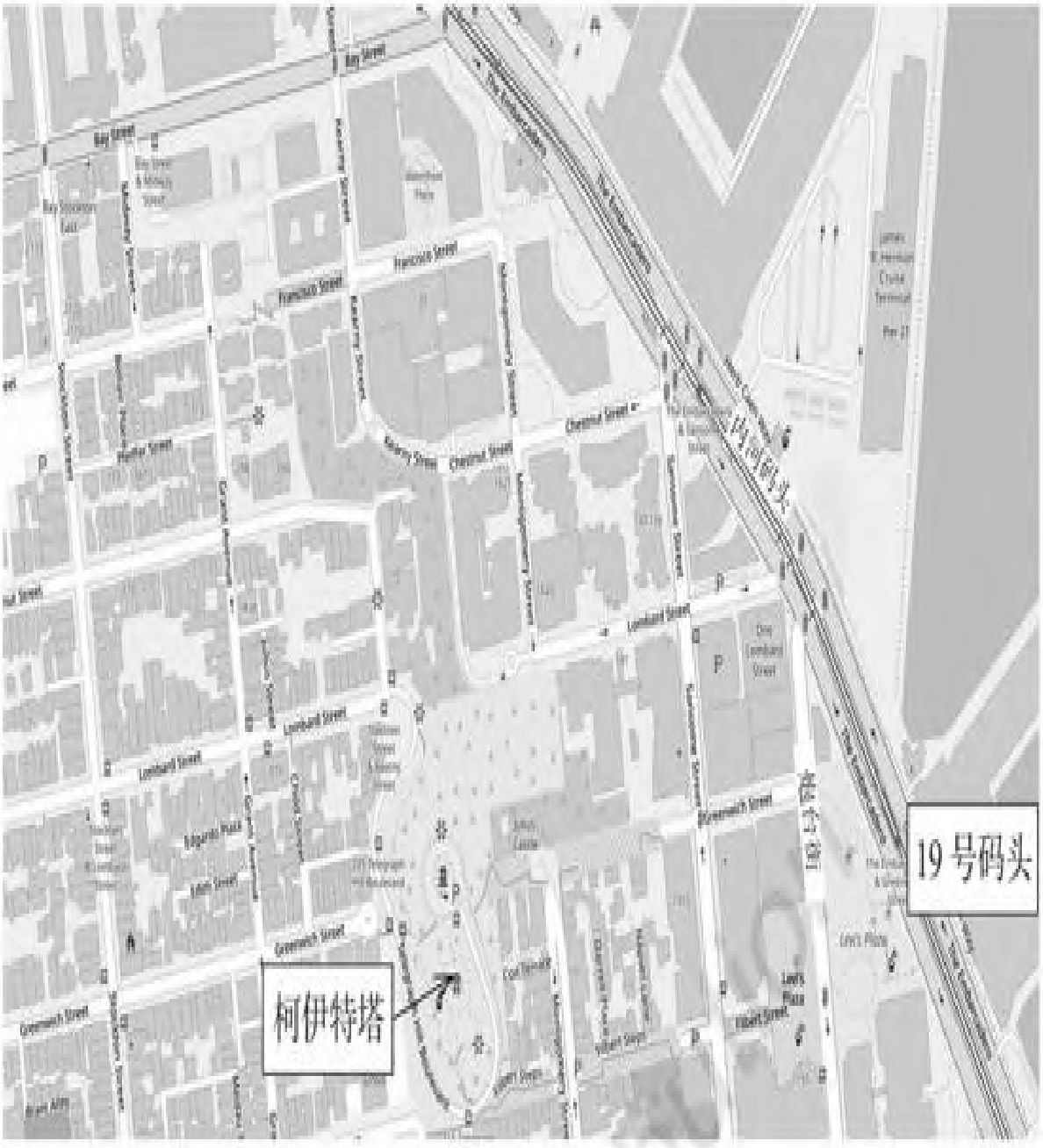


图14 旧金山部分地区，这幅图展示了起点19号码头和目的地柯伊特塔

在地图上搜索路线是一个我们熟悉的例子，但是因为标志性的地点数量太少，所以可能有些误导。例如，在美国，只有大约1 000万个十字路口。这看起来可能是一个很大的数字，但与“十五拼图”游戏

中不同局面的数目相比还是很小。十五拼图是一种玩具，它有一个4×4的网格，还有编号为1~15的15个拼块和一个空的托盘。玩家的目标是移动拼块以实现目标配置，例如将所有拼块按数字顺序排列。十五拼图有大约10万亿个状态（是美国的十字路口数量的100万倍），“二十四拼图”有大约8×10^{24}个状态。这是数学家所说的组合复杂性的一个例子，即随着“可移动的部分”数量的增加，组合的数量激增。我们再回到美国地图上：如果一家卡车运输公司想要优化其100辆卡车在美国的运输，那么该公司可能需要考虑的状态数量将是10^{700}个。

放弃理性的决策

许多游戏都具有这种组合复杂性，包括国际象棋、跳棋、西洋双陆棋和围棋。因为围棋规则简洁明了，所以我用围棋作为一个实例（图15）。黑棋和白棋轮流在棋盘的任意空位上落子。此时轮到黑方走棋，黑方有343种可能的走法。双方都试图围住尽可能多的地盘。例如，白棋很有可能在左边和左下角赢得地盘，而黑棋可能在右上角和右下角赢得地盘。围棋的一个关键概念是“大龙”，即通过垂直或水平相邻关系相互连接起来的一组颜色相同的棋子。只要大龙的旁边至少有一个空位置，这块棋就仍然是“活”的。如果大龙被完全包围，周围没有空位置，它就会变成“死”棋并被从棋盘上拿掉。

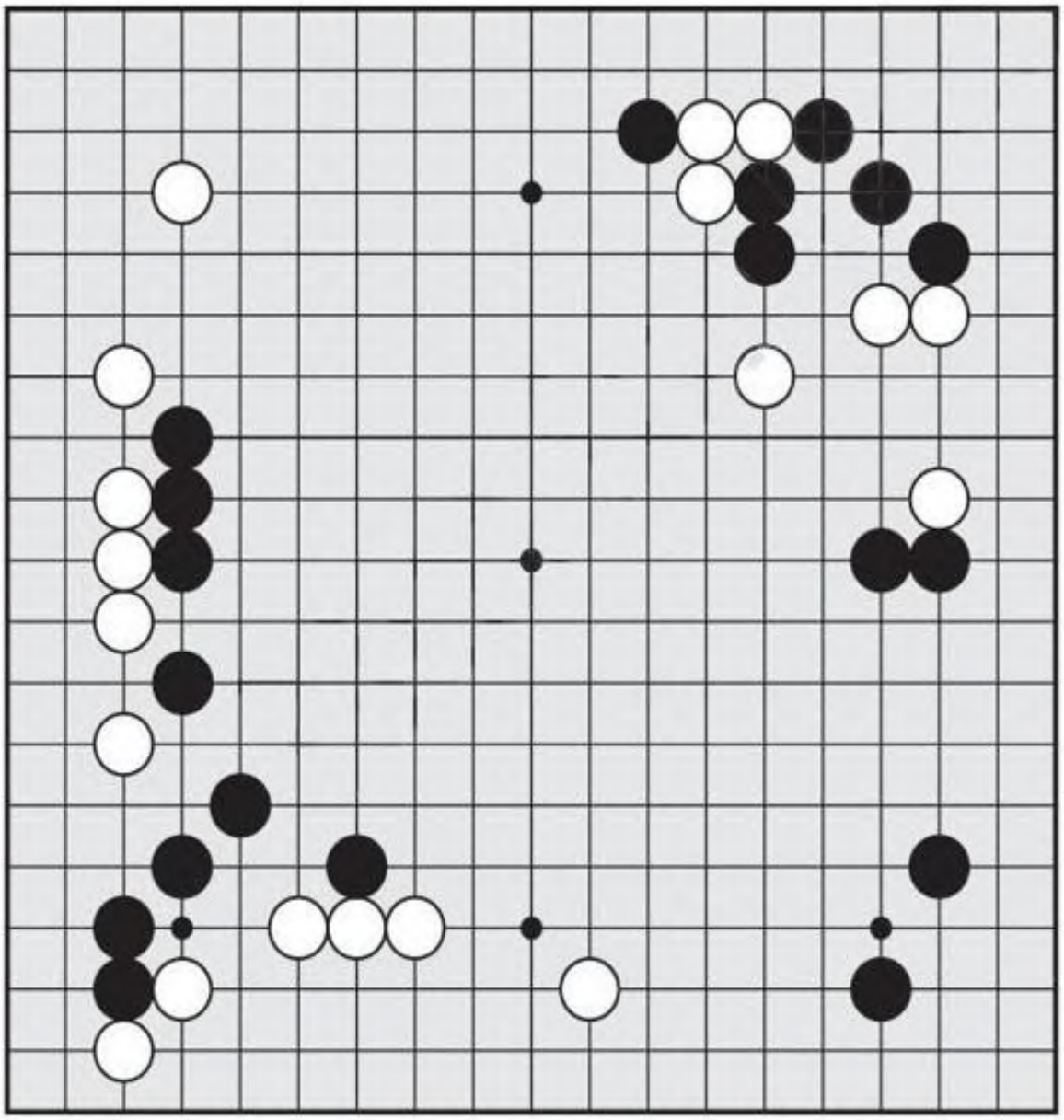


图15 2002年LG杯决赛中，李世石（黑棋）和崔明勋（白棋）在第5局比赛中的棋盘局面

下围棋的目标很明确：围住比你的对手更多的地盘即取胜。可能的动作也很明确：在空位置上放棋子。就像在地图上导航一样，决定该怎么做的显而易见的方法是：想象不同的动作序列可能产生的不同未来，并选择最佳的未来。你问：“如果我这么做，我的对手会怎么做？接下来我又该怎么办？”这一思路如图16所示。从空棋盘的初始

状态（有时称为根节点）开始，黑棋可以从3种不同的走子方法中选择一种。（其他的走子方法与这3种对称。）然后轮到白棋走棋。如果黑棋选择走在中央，那么白棋有两种不同的走子方法，一是走在角上，二是走在边上。然后再次轮到黑棋走棋。通过想象这些可能的未来，黑棋可以选择在初始状态时走哪一步棋。如果黑棋无法沿着每一条可能的路线计算到棋局结束，则可以使用一个评价函数来评估叶子节点的好坏。这里，评价函数将+5和+3赋值给两个叶子节点。即使是3×3路棋盘的围棋，我也只能展示未来棋局中博弈树的一小部分，但我希望这个思路足够清晰。的确，这种决策方式似乎只是简单的常识。

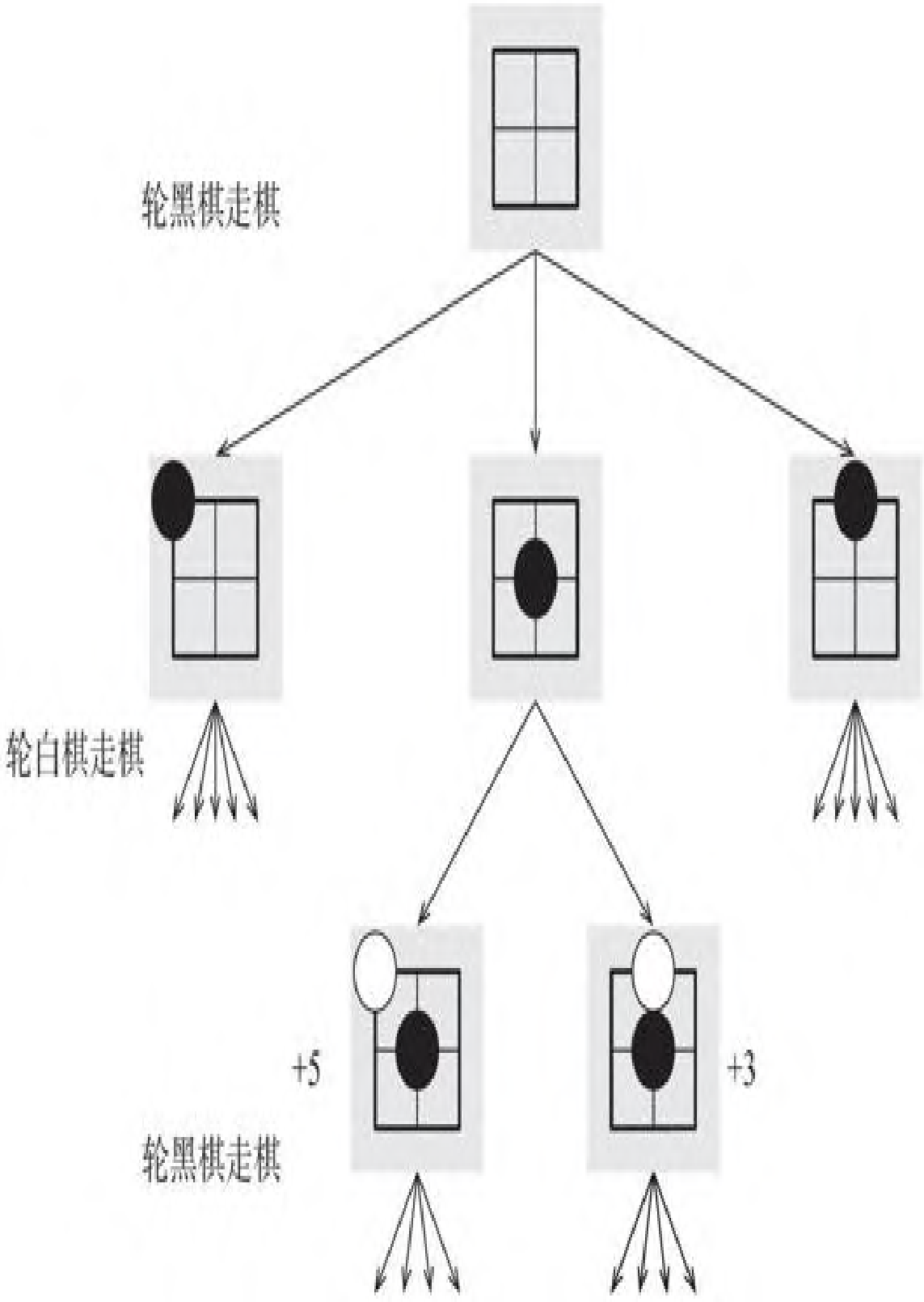


图16 3×3路棋盘的围棋博弈树的一部分

问题是，19×19路棋盘上有超过10^{170}种可能的局面。尽管在地图上找到一条确保最短的路线相对容易，但在围棋中找到一条确保获胜的路径完全不可行。即使该算法花费10亿年时间思考，它也只能探索整个可能性之树的一小部分。这就引出了两个问题：第一，程序应该探索博弈树的哪一部分？第二，对于它已经探索完的那部分博弈树，程序应该选择走哪一步棋？

我们先回答第二个问题：几乎所有的前向搜索程序都在使用的基本思路是，给叶子节点，即最遥远的未来的那些状态，各赋予一个估计值，然后“回溯”，找出根节点的各个选择有多好。[1] 例如，查看图16底部的两个局面，你可能会猜测左边局面的值为+5，右边局面的值为+3（从黑棋的角度看），因为角上的白子比边上的白子更容易受到攻击。如果这些值是正确的，那么黑棋可以预测白棋会在边上走棋，从而导致右边的局面。因此，给黑棋第一步棋在中央走棋赋值+3似乎是合理的。阿瑟·塞缪尔的西洋跳棋程序在1955年打败了它的创造者，[2]“深蓝”在1997年打败了当时的世界国际象棋冠军加里·卡斯帕罗夫，AlphaGo在2016年击败前世界围棋冠军李世石，它们采用的都是这个方案，只是略有不同而已。对于“深蓝”，人类主要基于对国际象棋的理解编写了一部分程序，用于评估叶子节点处的各种不同局面。对于塞缪尔的程序和AlphaGo，这些程序是从成千上万盘对局中学到的。

我们再来回答第一个问题，程序应该探索博弈树的哪一部分？这其实是在问：智能体应该做什么计算？这是人工智能中最重要的问题之一。对于玩游戏的程序来说，这是至关重要的，因为它们只有很短的、固定的时间分配，而且将时间用于毫无意义的计算是肯定会失败的。对于在现实世界中工作的人类和其他智能体来说，这一点甚至更为重要，因为现实世界要复杂得多：除非选择得当，否则在决定做什么的问题上，再多的计算也起不了一丁点作用。如果你正在开车，一

头麋鹿走到路中间，此时你考虑用欧元换英镑，或是黑棋第一步棋是否应该下在棋盘中央是毫无用处的。

人类能够管理自己的计算活动，从而合理而又迅速地做出理智的决策，这种能力至少与人类正确感知和推理的能力一样非凡。这似乎是我们自然而然、毫不费力获得的东西：当我父亲教我下国际象棋时，他教给我规则，但他并没有教我如何用某种聪明的算法来选择探索游戏树的哪些部分，以及忽略哪些部分。

这是如何发生的？我们可以在什么基础上指导我们的思想？答案是，计算的价值在于它能提高你的决策质量。选择计算的过程叫元推理，这意味着对推理进行推理。正如人可以根据期望值理性地选择动作，我们也可以选择计算。这叫作理性元推理。[3]它的基本思想很简单：

> 执行能够最大限度地改善决策质量的计算，当成本（就时间而言）超过预期改善时停止计算。

就是这样。不需要花哨的算法！这个简单的原理在包括象棋和围棋在内的广泛问题中生成了有效的计算行为。我们的大脑似乎也在执行类似的任务，这就解释了为什么我们不需要学习新的、特定于游戏的算法来思考我们所学的每一个新游戏。

当然，探索从当前状态延伸到未来的可能性之树并不是做出决定的唯一途径。通常，从目标开始向后追溯更有意义。例如，公路上出现麋鹿暗示了你应该设置目标，避免撞到麋鹿，这又反过来暗示了3种可能的动作：左转、右转或刹车。这个目标并不会让你把欧元换成英镑或把黑色棋子落在棋盘中央。因此，目标对一个人的思维有很好的聚焦效果。目前没有任何游戏程序利用这一想法。事实上，目标通常会考虑所有可能的合法动作。这是我不担心AlphaZero会接管世界的（诸多）原因之一。

展望更远的未来

让我们假设你已经决定在围棋棋盘上落下具体的一步棋。太棒了！现在你必须真正做到这一点。在现实世界中，这涉及你把手伸进棋盒里拿起一枚棋子，再移动到预定位置上方，然后按照围棋礼仪，轻轻地或者用力地把棋子规整地落到棋盘上。

上述动作的每一个步骤，都相当于一支由感知和运动控制指令组成的复杂舞蹈，涉及手、手臂、肩膀和眼睛的肌肉和神经。当你伸手去拿棋子的时候，你要确保你身体的其余部分不会因为重心的转移而不稳。你可能没有意识到自己选择这些动作，但这并不意味着它们不是由你的大脑选择的。例如，棋盒里可能有许多棋子，但是你的“手”——实际上是你的大脑处理了感官信息，仍然需要从中选择一枚棋子。

我们做的几乎每件事都是这样的。开车时，我们可能会选择向左变换车道，但这个动作需要看后视镜和你身后的情况，你可能会调整速度，转动方向盘，同时监控进度，直到动作完成。在谈话中，例行的回答，如“好的，让我查一下日历，然后再给你回复”，需要清楚地说出16个字，每一个字都需要数百个精确协调的运动控制指令来控制舌头、嘴唇、下巴、喉咙和呼吸器官的肌肉。你说母语时，这个过程是自动的，它类似于在计算机程序中运行子程序（见第2章）。事实上，复杂的动作序列可以变成常规的和自动的，从而在更复杂的过程中作为单个动作发挥作用，这是人类认知的绝对基础。用一种不太熟悉的语言说话，例如问去波兰的什切布热申市怎么走，会让你意识到，在你的生活中曾经有一段时间，阅读和说话是需要脑力劳动和大量练习的困难任务。

因此，你的大脑面临的真正问题不是在棋盘上选择一步棋，而是向你的肌肉发送运动控制指令。如果我们把注意力从围棋走子层面转移到运动控制指令层面，问题看起来就大不相同了。粗略地讲，你的

大脑大约每100毫秒就会发出一次指令。我们每个人大约有600块肌肉，所以理论上每秒最多可以做6 000个动作，每小时2 000万个，每年2 000亿个，一生20万亿个。所以请明智地使用它们！

现在，假设我们尝试应用类似AlphaZero的算法来解决这个层面的决策问题。AlphaZero大约能向前计算50步棋，但是50步棋的运动控制指令只能让你计划几秒钟后的未来！这对于长达一个小时的围棋对局所需要的2 000万个运动控制指令而言远远不够，对于攻读博士学位所涉及的1万亿（1 000 000 000 000）个步骤而言更远远不够。因此，虽然AlphaZero在围棋对局中比任何人都看得更远，但这种能力在现实世界中似乎并没有帮助。这是一种错误的前向搜索。

当然，我并不是说攻读博士学位实际上需要提前计划1万亿次肌肉动作。学生最初只会制订非常抽象的计划，可能是选择加州大学伯克利分校或其他学校，选择博士生导师和研究课题，申请资金，获得学生签证，前往选定的城市，做一些研究等。为了做出选择，你要进行足够多的思考，考虑正确的事情，这样你的决定就会变得清晰起来。如果获得签证等抽象步骤的可行性不明确，那么你需要进行更多的思考，或许还需要搜集信息，这意味着让计划在某些方面更加具体，例如选择一种你有资格申请的签证类型，搜集必要的文件，并提交申请。图17显示了这样一个抽象计划，并将获得签证的步骤细化为三个子计划。当你开始执行这个计划的时候，它的初始步骤必须一直细化到原始级别，这样你的身体才能执行它们。

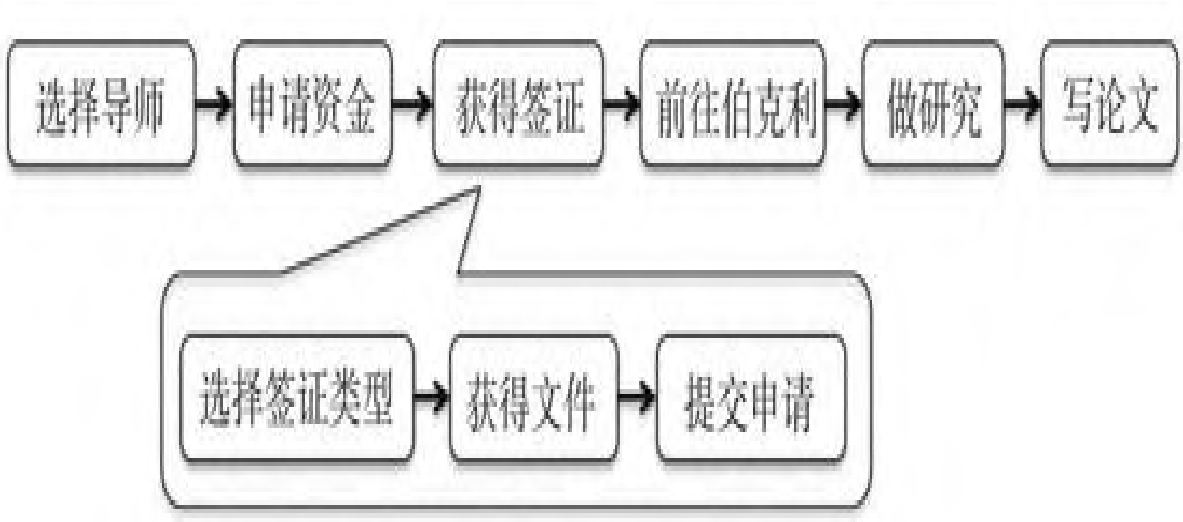


图17 一位选择在加州大学伯克利分校攻读博士学位的留学生的抽象计划

AlphaGo根本无法进行这种思考，它考虑的唯一动作就是从初始状态开始按顺序发生的原始动作。它没有抽象计划的概念。试图将AlphaGo应用于现实世界，就像试图通过思考第一个字母应该是A、B还是C来写一部小说一样。

1962年，赫伯特·西蒙在著名的论文《复杂性的体系结构》中强调了层级组织的重要性。[4]自20世纪70年代初以来，人工智能研究人员已经开发出多种方法来构建和细化层级组织的计划。[5]一些由此产生的系统能够构建具有数千万步骤的计划，例如在大型工厂组织制造活动。

现在，我们在理论上对抽象动作的含义有了相当深刻的理解，即如何定义抽象动作对世界产生的影响。[6]例如，考虑一下图17中的抽象动作“前往伯克利”。它可以通过多种不同的方式实现，每种方式都会对世界产生不同的影响：你可以乘船偷渡到伯克利，飞往加拿大然后步行跨越边境，租用私人飞机等。但是你现在不需要考虑这些选择。只要你确信有一种方法可以做到不消耗太多的时间和金钱，也不会招致太多的风险而危及计划的其余部分，你就可以把抽象步骤

“前往伯克利”放入计划中，并确信计划会成功。通过这种方式，我们可以制订高层次的计划，这些计划最终将变成数十亿或数万亿个原始步骤，在真正开始执行这些原始步骤之前，你完全不必担心这些步骤是什么。

当然，如果没有层级组织，这一切都是不可能的。如果没有像获得签证和写论文这样的高层次的动作，我们就无法制订一个抽象的计划来获得博士学位；如果没有获得博士学位和创办公司这种更高层次的动作，我们就无法制订出先获得博士学位然后创办公司的计划。在现实世界中，如果没有几十个抽象层次上的大量动作库，我们就会迷失方向。（围棋游戏没有明显的动作层次，所以我们大多数人在下围棋时都很迷茫。）然而，目前现有的所有分层规划方法都依赖于人为产生的抽象动作和具体动作的层级组织，我们还不知道如何从经验中学习这种层级组织。

---

[1] 未来60年国际象棋程序的基本计划：Claude Shannon, “Programming a computer for playing chess,” *Philosophical Magazine*, 7th ser., 41 (1950): 256–75。香农的建议借鉴了几个世纪以来通过累加棋子价值来评估国际象棋局面的传统，请参见Pietro Carrera, *Il gioco degli scacchi* (Giovanni de Rossi, 1617)。

[2] 一份描述塞缪尔对西洋跳棋早期强化学习算法的开拓性研究的报告：Arthur Samuel, “Some studies in machine learning using the game of checkers,” IBM Journal of Research and Development 3 (1959): 210–29。

[3] 理性元区域划分的概念及其在搜索和游戏中的应用源于我的学生埃里克·韦弗尔德（Eric Wefald）的论文研究，他在写完自己的作品之前就悲惨地死于一场车祸。以下是在他死后发表的文章：Stuart Russell and Eric Wefald, *Do the Right Thing: Studies in Limited Rationality* (MIT Press, 1991)。另请参见：Eric Horvitz,“Rational metareasoning and compilation for optimizing decisions under bounded resources”, in *Computational Intelligence, II: Proceedings of the International Symposium,* ed. Francesco Gardin and Giancarlo Mauri (North-Holland, 1990); and Stuart Russell and Eric Wefald, “On optimal game-tree search using rational metareasoning,” in *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, ed. Natesa Sridharan (Morgan Kaufmann, 1989)。

[4] 这也许是第一篇展示层级组织如何降低规划的组合复杂性的论文：Herbert Simon, “The architecture of complexity,” *Proceedings of the American Philosophical Society* 106 (1962): 467–82。

[5] 层级规划的权威参考文献：Earl Sacerdoti, “Planning in a hierarchy of abstraction spaces,” *Artificial Intelligence* 5 (1974): 115–35。另请参见：Austin Tate, “Generating project networks,” in *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, ed. Raj Reddy (Morgan Kaufmann, 1977)。

[6] 关于高层次动作的正式定义：Bhaskara Marthi, Stuart Russell, and Jason Wolfe, “Angelic semantics for high-level actions,” in *Proceedings of the 17th International Conference on Automated Planning and Scheduling*, ed. Mark Boddy, Maria Fox, and Sylvie Thiébaux (AAAI Press, 2007)。

附录B

知识与逻辑

逻辑学是一种用确定的知识进行推理的研究。就主题而言，它是完全通用的，也就是说，逻辑学知识可以是关于任何事物的。因此，逻辑学是我们理解通用智能的不可或缺的一部分。

逻辑学的主要要求是形式语言中的句子有精确含义，这样就有一个明确的过程来判定一个句子在给定的情况下是真还是假。一旦有了这些，我们就可以编写合理的推理算法，根据已知的句子生成新的句子。这些新句子能保证是从系统已经知道的句子中派生出来的，这意味着在任何情况下，只要原来的句子为真，新句子就必然为真。这使得机器能够回答问题，证明数学定理，或者构建确保成功的计划。

高中代数提供了一个很好的例子（尽管它可能会唤起我们痛苦的记忆）。形式语言包括$4x+1 = 2y-5$之类的句子。当$x=5$且$y=13$时，这个句子为真；当$x=5$且$y=6$时，这个句子为假。从这个句子可以推导出另一个句子，比如$y = 2x+3$，当第一个句子为真时，第二个句子必然也为真。

在古代印度、中国和希腊独立发展起来的逻辑学核心思想是，精确含义和合理推理可以应用于任何句子，而不仅仅应用于数字。经典的例子从“苏格拉底是人”和“人皆有一死”开始，然后得出“苏格拉底是会死的”。[1]这种推导是严格且正式的，因为它不依赖于任何关于“苏格拉底是谁”或“人和死分别是什么意思”的进一步信息。逻辑推理严格且正式，这意味着我们可以编写算法来实现它。

命题逻辑

有两种逻辑尤为重要，它们可以帮助我们理解人工智能的能力和前景：命题逻辑和一阶逻辑。这两者之间的差异对于理解人工智能的现状及其可能的发展至关重要。

让我们从命题逻辑开始，这是二者中相对简单的一个。句子仅由两种东西组成：一是代表命题真假的符号，二是逻辑连接词，如“且”“或”“非”“如果……那么……”。（稍后我们将看到一个示例。）这些逻辑连接词有时被称为布尔逻辑，以19世纪逻辑学家乔治·布尔（George Boole）的名字命名，他用新的数学思想重振了他的领域。逻辑连接词与计算机芯片中使用的逻辑门是一样的。

自20世纪60年代初以来，命题逻辑中实用的“推理算法”就已为人们所知。[2] [3]虽然在最坏的情况下，处理一般推理任务可能需要花费指数级的时间，[4]但现代命题推理算法处理的问题包含数百万个命题符号和数千万个句子。它们是构建有保障的后勤计划，在制造芯片之前验证芯片设计，在部署软件应用程序和安全协议之前检查其正确性的核心工具。令人感到惊奇的是，一旦人们将所有这些任务表述为推理任务，就可以使用单一算法（命题逻辑的推理算法）来解决所有任务。显然，这是迈向智能系统通用性目标的一步。

不幸的是，这不是很大的一步，因为命题逻辑的语言并不是很有表现力。让我们看看这在实践中会是什么样子，例如，我们用命题逻辑语言来表达围棋中合法走子的基本规则：“轮到谁走，谁就可以在任意空着的交叉点上落子。”[5]第一步是决定用什么命题符号来讨论围棋的走法和棋盘上的局面。重要的基本命题是特定颜色的棋子在特定时间是否处于特定位置。因此，我们需要诸如“白38，5，5”和“黑38，5，5”这样的符号。（请记住，与人类、死亡和苏格拉底一样，推理算法不需要知道符号的含义。）那么白棋第38手能够在（5，5）位落子的逻辑条件是：

（非白38，5，5）且（非黑38，5，5）

换言之：那里没有白子，也没有黑子。这似乎很简单。不幸的是，在命题逻辑中，棋局中的每个位置和每一步棋都必须单独写出。由于棋盘有361个交叉点，每盘棋大约要进行300步，这意味着该规则有超过10万个副本！关于吃子和重复局面的规则，由于它们涉及多个棋子和位置，因此情况更糟，我们很快就会写满数百万页。

显然，现实世界比围棋棋盘大得多，远远超过了361个位置和300步棋，而且除了棋子之外还有很多东西，因此，用命题语言来了解现实世界是完全没有希望的。

问题不仅在于规则手册将达到令人难以置信的规模，还在于学习系统从例子中获取规则，需要多到令人难以置信的经验。虽然人类只需要一两个例子就能了解走子、吃子等的基本思想，但是基于命题逻辑的智能系统必须看到针对每个位置和步数的走子、吃子的示例。这个系统无法像人类那样从几个例子中归纳出一般规则，因为它无法表达一般规则。这种限制不仅适用于基于命题逻辑的系统，也适用于任何具有类似表达能力的系统，其中包括贝叶斯网络（它最早是命题逻辑的概率学“表亲”）以及神经网络（它是人工智能“深度学习”方法的基础）。

一阶逻辑

那么，下一个问题是，我们能否设计出一种更具表现力的逻辑语言？我们希望可以通过以下方式将围棋规则告知基于知识的系统：

对于棋盘上的所有位置，以及所有棋步，规则如下……

一阶逻辑由德国数学家戈特洛布·弗雷格（Gottlob Frege）于1879年提出，它为人们提供了一种描述规则的方法。[6]命题逻辑和一阶逻辑的关键区别在于：命题逻辑假设世界是由真假命题组成的，而一阶逻辑假设世界是由可以通过各种方式相互关联的对象组成的。例如，这些对象可以包括彼此相邻的位置、连续的几步棋、特定局面下在某些位置上的棋子，以及特定局面下合法的走子等。一阶逻辑允许我们宣称世界上所有对象的某些属性都为真。因此，围棋规则可以写成：

> 对于所有棋步t、所有位置l、所有颜色c，如果第t步棋轮到颜色c移动，且下第t步棋时l没被占据，则颜色c的第t步棋在位置l处落子是合法的。

通过一些额外的说明和一些额外的句子来定义棋盘位置、两种颜色以及“没被占据”的意思，我们已经开始有了一个完整的围棋规则。这些规则用一阶逻辑描述与用英语描述几乎一样长。

20世纪70年代末，逻辑编程的发展为逻辑推理提供了简明而又高效的技术，这种技术体现在一种名为Prolog的编程语言中。计算机科学家研究如何在Prolog中以每秒数百万步的速度进行逻辑推理，从而使逻辑的许多应用变得实用。1982年，日本政府宣布对基于Prolog的人工智能进行巨额投资，该项目名为第五代项目（Five Generation Project），[7]美国和英国也做出了类似的回应。[8] [9]

不幸的是，第五代项目和其他类似的项目在20世纪80年代末、90年代初失去了发展的动力，部分原因是逻辑无法处理不确定的信息。人们概括出了一个很快成为贬义的词：有效的老式人工智能，简称GOFAI。[10]越来越多的人认为逻辑与人工智能无关。事实上，如今深度学习领域的许多人工智能研究人员对逻辑一无所知。这种流行趋势可能会消退：如果你接受世界中存在以各种方式相互关联的对象，

那么一阶逻辑将是与此相关的，因为它提供了对象和关系的数学基础。谷歌DeepMind的CEO德米斯·哈萨比斯（Demis Hassabis）也认同这一观点：[11]

> 你可以把今天的深度学习想象成与我们大脑中的感觉皮层类似的东西：我们的视觉皮层或听觉皮层。但是，真正的智能当然远不只这些，你还必须将其重新组合成高级思维和符号推理，这些都是经典人工智能在20世纪80年代试图解决的许多事情。
>
> ……我们希望（这些系统）建立起符号水平的推理——数学、语言和逻辑。这是我们工作的重要部分。

因此，我们在人工智能研究的头30年中得到的最重要的教训之一是：在任何有用的意义上，一个可以了解事物的程序至少需要拥有与一阶逻辑能力相媲美的表达和推理能力。到目前为止，我们还不知道需要采取什么样的确切形式才能创造出这样的程序，它可能会被纳入概率推理系统、深度学习系统或一些尚待发明的混合设计中。

---

[1] 这个例子不太可能出自亚里士多德，但可能起源于生活在公元2世纪或3世纪的塞克斯都·恩披里柯（Sextus Empiricus）。

[2] 用于一阶逻辑定理证明的首个算法是通过将一阶语句简化为（大量）命题语句来实现的：Martin Davis and Hilary Putnam, “A computing procedure for quantification theory,” *Journal of the ACM* 7 (1960): 201–15。

[3] 命题推理的一种改进算法：Martin Davis, George Logemann, and Donald Loveland, “A machine program for theorem-proving,” *Communications of the ACM* 5(1962): 394–97。

[4] 可满足性问题，即确定一组句子在某些情况下是否为真，是NP完全问题。推理问题，即确定句子是否从已知句子中得出，是NP完全问题的余问题，这类问题被认为比NP完全问题更难。

[5] 这条规则有两个例外：不允许重复（不允使棋局回到先前局面的着法），和不允许自杀（不允许把棋子下在立即被吃掉的位置，例如，不能在已经被包围的位置上落子）。

[6] 介绍了我们今天所理解的一阶逻辑的著作（Begriffsschrift的意思是“概念写作”）：Gottlob Frege, *Begriffsschrift, eine der arithmetischen nachgebildete For melsprache des reinen Denkens* (Halle, 1879)。弗雷格的一阶逻辑符号非常古怪且笨拙，所以很快就被朱塞佩·皮亚诺（Giuseppe Peano）所引入的符号所取代，后者的符号至今仍被广泛使用。

[7] 关于日本通过基于知识的系统争取霸权的概述：Edward Feigenbaum and Pamela McCorduck, *The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World* (Addison-Wesley, 1983)。

[8] 美国的努力包括战略计算计划以及微电子和计算机技术公司（MCC）的成立。请参见：Alex Roland and Philip Shiman, *Strategic Computing: DARPA and the Quest for Machine Intelligence, 1983–1993* (MIT Press, 2002)。

[9] 关于20世纪80年代英国对人工智能重新出现的反应的历史：Brian Oakleyand Kenneth Owen, *Alvey: Britain's Strategic Computing Initiative* (MIT Press,1990)。

[10] “GOFAI”（有效的老式人工智能）一词的由来：John Haugeland, *Artificial Intelligence: The Very Idea* (MIT Press, 1985)。

[11] 关于人工智能和深度学习的未来，对德米斯·哈萨比斯进行的采访：Nick Heath, “Google Deep Mind founder Demis Hassabis: Three truths about AI,” *Tech Republic*, September 24, 2018。

附录C

不确定性和概率论

逻辑学为用确定的知识进行推理提供了通用基础，而概率论则需要用不确定的信息进行推理（确定的知识是其中的一种特例）。不确定性是现实世界中的智能体的正常认知状态。虽然概率论的基本思想是在17世纪发展起来的，但直到最近，以正式的方式用大概率模型来进行表达和推理才成为可能。

概率论基础

概率论和逻辑学都认为，可能世界（possible worlds）是存在的。人们通常从定义可能世界是什么开始，例如，如果我掷一枚普通的6面色子，就有6个世界（有时称为结果）：1、2、3、4、5、6。其中一个恰好是实际情况，但我的先验知识不知道哪一个是实际情况。概率论假设可以为每个世界添加一个概率，对于我的色子，我把1/6的概率添加到每个世界中。（这些概率碰巧相等，但不必一定如此，唯一的要求是这些概率相加之和必须等于1。）现在我可以问一个问题，比如，“我掷出偶数的概率有多大？”要找到答案，我只需将数字为偶数的3个世界的概率相加：1/6+1/6+1/6=1/2。

考虑新证据也很简单。假设先知告诉我掷出来的是一个质数（即2、3或5），这排除了世界1、4和6。我只需计算剩余的可能世界的概率，使它们的总和保持为1。现在，掷出2、3或5的概率各为1/3，并且

我掷出偶数的概率现在只有1/3，因为2是剩下的唯一的偶数。这种随着新证据到来而更新概率的过程就是“贝叶斯更新”的一个例子。

因此，概率论看起来易如反掌！只要是台电脑就能完成数字加法，那么有什么问题呢？当存在很多个世界时，问题就出现了。例如，如果我掷色子100次，就有6^{100}种结果。通过把数字分别添加到每个结果上来开始概率推理过程是不可行的。处理这种复杂情况的线索之一来自一个事实：如果已知掷色子的结果是公平的，那么掷色子就是独立事件，也就是说，掷色子得到的任何一个结果都不会影响其他结果的概率。因此，独立性有助于构建复杂事件集的概率。

假设我正在和儿子乔治一起玩《大富翁》游戏。我的棋子落在了“只能访问”的方格上，乔治拥有黄色地皮，我现在的位置距离他的黄色地皮分别有16、17和19个方格。他现在是否应该在黄色地皮上买房子？如果我落在了那些方格里，我就得付给他昂贵的租金。或者他是否应该等到下一回合再买房子？这取决于我在当前回合落在黄色地皮上的概率。

《大富翁》游戏中掷色子的规则是：掷两枚色子，然后根据两枚色子的点数之和移动棋子；如果两枚色子点数一致，则玩家将再次掷色子并再次移动；如果第二轮两枚色子的点数仍然一致，则玩家将第三次掷色子并第三次移动（但如果第三轮两个色子的点数仍然一致，则玩家将会入狱）。例如，我可能掷出4－4，然后掷出5－4，总和为17；或者掷出2－2，然后掷出2－2，然后掷出6－2，总和16。和以前一样，我只需要把我落在黄色地皮上的所有可能世界的概率相加。不幸的是，此时有很多世界。我总共可以掷出多达6枚色子，因此可能世界的数量达到了数千个。此外，这些色子不再是独立的，因为除非第一次掷出的两枚色子的点数一致，否则第二次机会就不存在了。而如果我们固定了第一次时两枚色子的值，那么第二次两枚色子的值就还是独立的。我们是否有办法控制这些？

贝叶斯网络

20世纪80年代初，朱迪亚·珀尔提出了一种叫作“贝叶斯网络”的正式语言，这使得在许多现实世界的情境中，人们都有可能以非常简洁的形式表示海量结果的概率。[1]

图18展示了《大富翁》游戏中掷色子的贝叶斯网络，表现了《大富翁》游戏中掷色子的规则，并使算法能够计算从其他一些地图格子（如只能访问的格子）开始，落在一组特定格子（如黄色地皮）上的概率。（为简单起见，网络忽略了获得机会卡或公共基金卡并转移到不同位置的可能性。）D_1和D_2表示第一次掷出的两个色子的点数，并且它们是独立的（它们之间没有联系）。如果两个色子点数相同（$Doubles_{12}$），则玩家掷第二次，因此D_3和D_4的值不是0，以此类推。在上述情况下，如果掷三次色子总点数是16、17或19，玩家就会落在黄色地皮上。唯一需要提供的概率是各个色子（D_1、D_2等）掷出1、2、3、4、5、6点的概率都是1/6，也就是说，总共只有36个数字，而不是数千个。解释网络的确切含义需要一些数学知识，[2]但基本思想是用箭头表示依赖关系——例如，$Doubles_{12}$的值取决于D_1和D_2的值。类似的，D_3和D_4（第二次掷两个色子）的值取决于$Doubles_{12}$，因为如果$Doubles_{12}$的值为假，那么D_3和D_4的值为0（即没有更多的机会掷色子）。

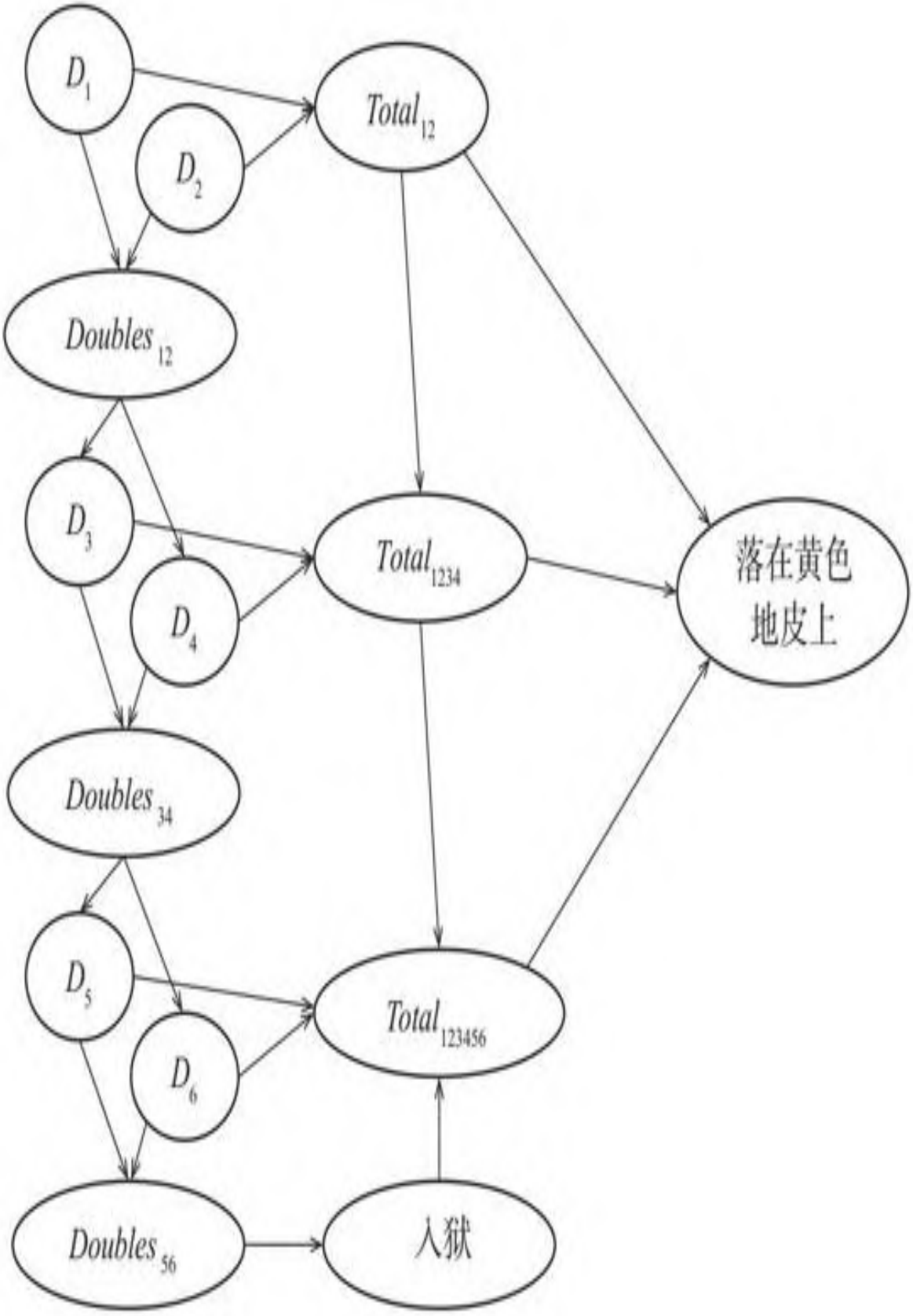

D_1
D_2
$Total_{12}$
$Doubles_{12}$
D_3
D_4
$Total_{1234}$
落在黄色
地皮上
$Doubles_{34}$
D_5
D_6
$Total_{123456}$
$Doubles_{56}$
入狱

图18 《大富翁》游戏中掷色子的贝叶斯网络

就像命题逻辑一样，有些算法可以用任何证据回答任何贝叶斯网络的任何问题。例如，我们可以问，我落在黄色地皮上的概率是多少，结果是大约3.88%。（这意味着乔治可以先等一等，然后再为黄色地皮买房子。）稍微激进些，我们可以问，在第二次掷色子结果是两个3的情况下，我落在黄色地皮上的概率是多少。在这种情况下，该算法可以自行算出第一次掷色子的结果必须是相同的两个数字，并得出答案约为36.1%。这是贝叶斯更新的一个例子：当添加新的证据（即第二次掷色子结果是两个3）时，落在黄色地皮上的概率从3.88%变为36.1%。同样，我掷3次色子（$Doubles_{34}$为真）的概率约为2.78%，而我掷3次色子并落在黄色地皮上的概率约为20.44%。

贝叶斯网络提供了一种基于知识系统的方法，避免了和20世纪80年代基于规则的专家系统一样的失败。（事实上，人工智能界如果在20世纪80年代初不那么强烈地抗拒概率，就可能会避免基于规则的专家系统泡沫之后的人工智能寒冬。）从提供医学诊断到预防恐怖主义，成千上万个应用已经在各个领域投入使用。[3]

贝叶斯网络为表示必要的概率和执行计算提供了机制，为许多复杂任务实现了贝叶斯更新。然而，就像命题逻辑一样，它们表达一般知识的能力也相当有限。在许多应用中，贝叶斯网络表达变得非常庞大且重复，例如，就像用命题逻辑表达围棋规则时每一步棋都需要重复一样，基于概率的《大富翁》规则必须重复每一个玩家、每一个玩家所在的每一个可能位置，以及游戏中的每个动作。如此巨大的网络实际上不可能由人工创建，相反，人们将不得不求助于用传统语言（如C++）编写的代码来生成多个贝叶斯网络片段并将它们拼凑在一起。虽然这作为具体问题的工程解决方案是可行的，但它是通用性的一个障碍，因为每个应用程序的C++代码都必须由人类专家编写。

一阶概率语言

幸运的是，我们可以将一阶逻辑的表达能力与贝叶斯网络精确获取概率信息的能力结合起来。这种结合让我们得到一加一大于二的效果：基于概率知识的系统能够处理比单用逻辑方法或贝叶斯网络更广泛的现实世界情况。例如，我们可以轻松获取关于基因遗传的概率知识：

对于所有人c、f和m，

如果f是c的父亲，且m是c的母亲，

且f和m都是AB型血，

那么c是AB型血的概率为50%。

实际上，一阶逻辑和概率论的结合给了我们一种表达许多对象的不确定信息的方式。原因是，当我们把不确定性添加到包含对象的世界时，我们就得到了两种新的不确定性：不仅有关于哪些事实是真或假的不确定性，还包括存在哪些对象的不确定性，以及对象的对应关系的不确定性。这类不确定性无处不在。这个世界并不像维多利亚时代的戏剧一样有人物列表，相反，你会通过观察逐渐了解对象的存在。

有时候，你对新对象的了解可能相当明确，比如你打开酒店的窗户，第一次亲眼看到圣心教堂；而有时，这种了解也可能相当不明确，比如当你感觉到轻微的震动时，这可能是地震造成的，也可能是路过的地铁造成的。虽然你对是不是圣心教堂感到非常确定，但关于是不是地铁导致的震动你并不能确定，你可能会乘坐同一列列车几百次，却从未意识到它们是同一列列车。有时我们不需要解决不确定性，我通常不会给一袋圣女果中的所有大果命名，并记录每个圣女果有多好，除非我是在记录圣女果腐烂实验的进展。而对于一个全是研

究生的班级，我会尽我最大的努力记住他们的身份。（有一次，我的小组里有两名研究助理，他们的名字和姓氏相同，外貌非常相似，研究的课题也很相关，但至少我非常肯定他们是两个人。）问题是，我们直接感知的不是对象的身份，而是对象的外貌，对象通常没有唯一的标识性小牌子。身份是我们的大脑有时为了我们自己的目的而添加到对象上的东西。

概率论与有表达能力的正式语言的结合是人工智能的一个非常新的子领域，通常被称为概率编程。[4]人们已经开发了几十种概率编程语言（PPL），其中许多都是从普通编程语言而不是从一阶逻辑中获得表达能力的。所有的概率编程语言系统都具有用复杂的、不确定的知识进行表达和推理的能力。其应用有微软的TrueSkill系统，该系统每天为数百万名电子游戏玩家评级；还有以前任何机械假说都无法解释的人类认知方面的模型，例如，根据单个例子学习新的视觉对象类别的能力；[5]还有《全面禁止核试验条约》（CTBT）中规定的用于探测秘密核爆炸的全球地震监测。[6]

该监测系统从全球网络（由150多台地震仪组成）中搜集实时地面活动数据，旨在识别地球上发生的所有超过一定震级的地震事件，并对可疑的地震事件进行标记。显然，这里存在着大量的不确定性，因为我们事先不知道将会发生什么事件，而且数据中的绝大多数信号都是噪声。另外，身份上的不确定性也有许多，比如，在南极洲的A站探测到的地震能量信号和来自巴西的B站探测到的另一个信号是不是同一事件？聆听地球的声音就像听同时进行的成千上万个对话，它们在传输上会延迟，也会被回声干扰，被地震淹没。

我们如何使用概率编程解决这个问题？有人可能会认为，我们需要一些非常聪明的算法来解释所有的可能性。事实上，通过遵循基于知识系统的方法，我们根本不需要设计任何新算法。我们仅需使用概率编程语言来表达我们对地球物理学的了解：在自然地震活动区发生事件的频率，地震波在地球上的传播速度和衰减速度，探测器的灵敏

度，以及噪声的大小。然后我们添加数据并运行概率推理算法。由此产生的监测系统名为NET-VISA，自2018年以来，该系统一直作为条约核查制度的一部分运行。图19展示了NET-VISA检测到的2013年朝鲜核试验情况。卫星照片中标出了隧道入口。NET-VISA估计的位置距离隧道入口约700米，主要基于4 000至10 000公里以外的站点的探测结果。CTBTO（全面禁止核试验条约组织）的定位基于专业地球物理学家的一致估计。

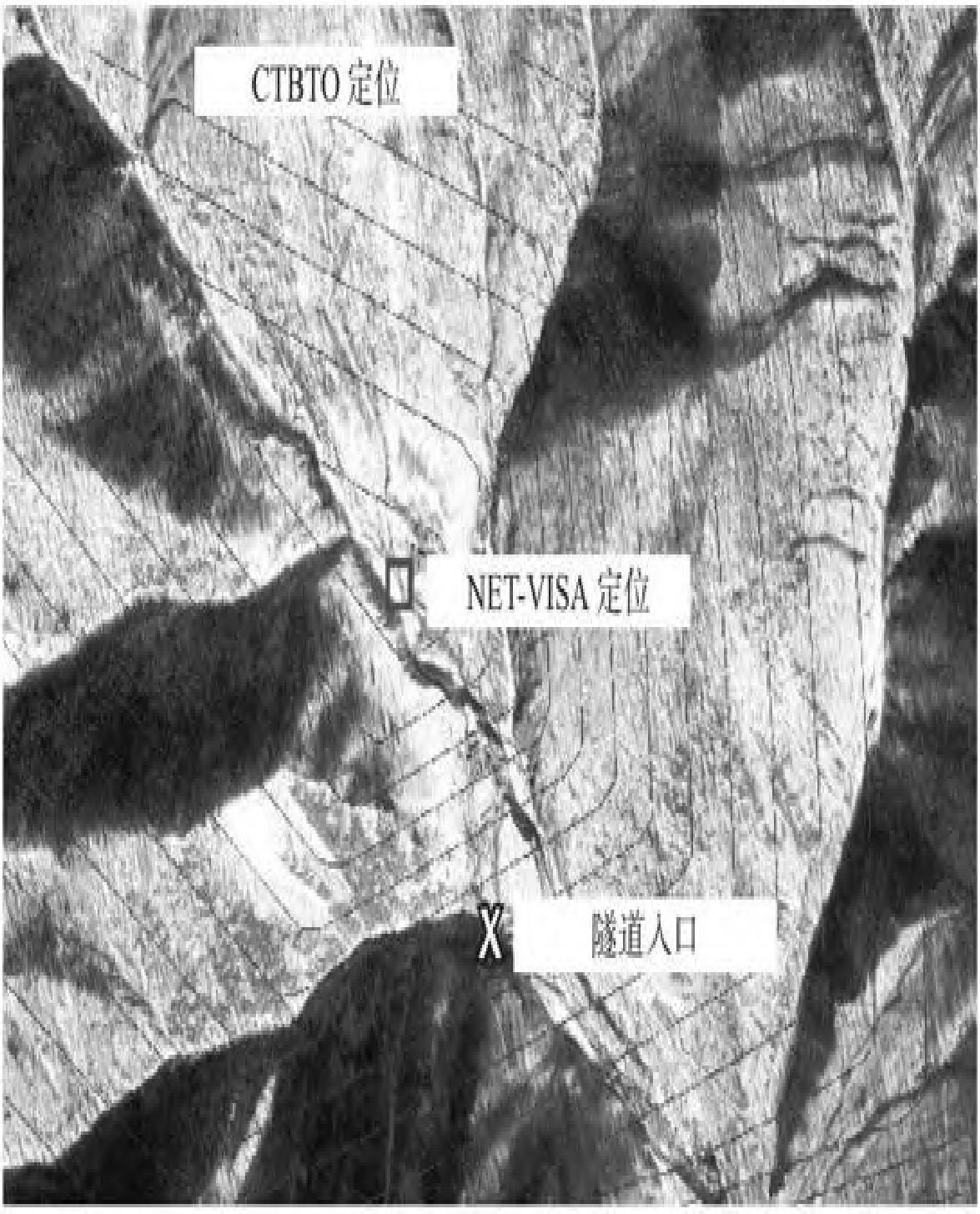


图19 2013年2月12日，朝鲜政府可能进行核试验的地点

了解世界

概率推理最重要的作用之一是了解世界上无法直接被观察到的部分。在大多数电子游戏和棋类游戏中，这一作用并不是必要的，因为所有相关信息都是可见的，但在现实世界中这种全部可见的情况很少出现。

最早的一起涉及无人驾驶汽车的严重事故就是一个例子。这起事故发生在2017年3月24日，美国亚利桑那州坦佩市东唐卡洛斯大街的南麦克林托克大道上。[7]如图20所示，一辆沿着麦克林托克大道向南行驶的沃尔沃无人驾驶汽车（V）正在接近一个红绿灯刚刚变成黄灯的十字路口，以每小时38英里的速度行驶在最右侧的车道上。沃尔沃的车道是畅通的，所以它匀速通过十字路口。另外两条车道的车辆是停着的，红绿灯（标记为L）变成了黄灯。然后一辆不在沃尔沃视野中的车——图20中的本田（H），从停着的两队车后面突然出现，随后发生了碰撞。

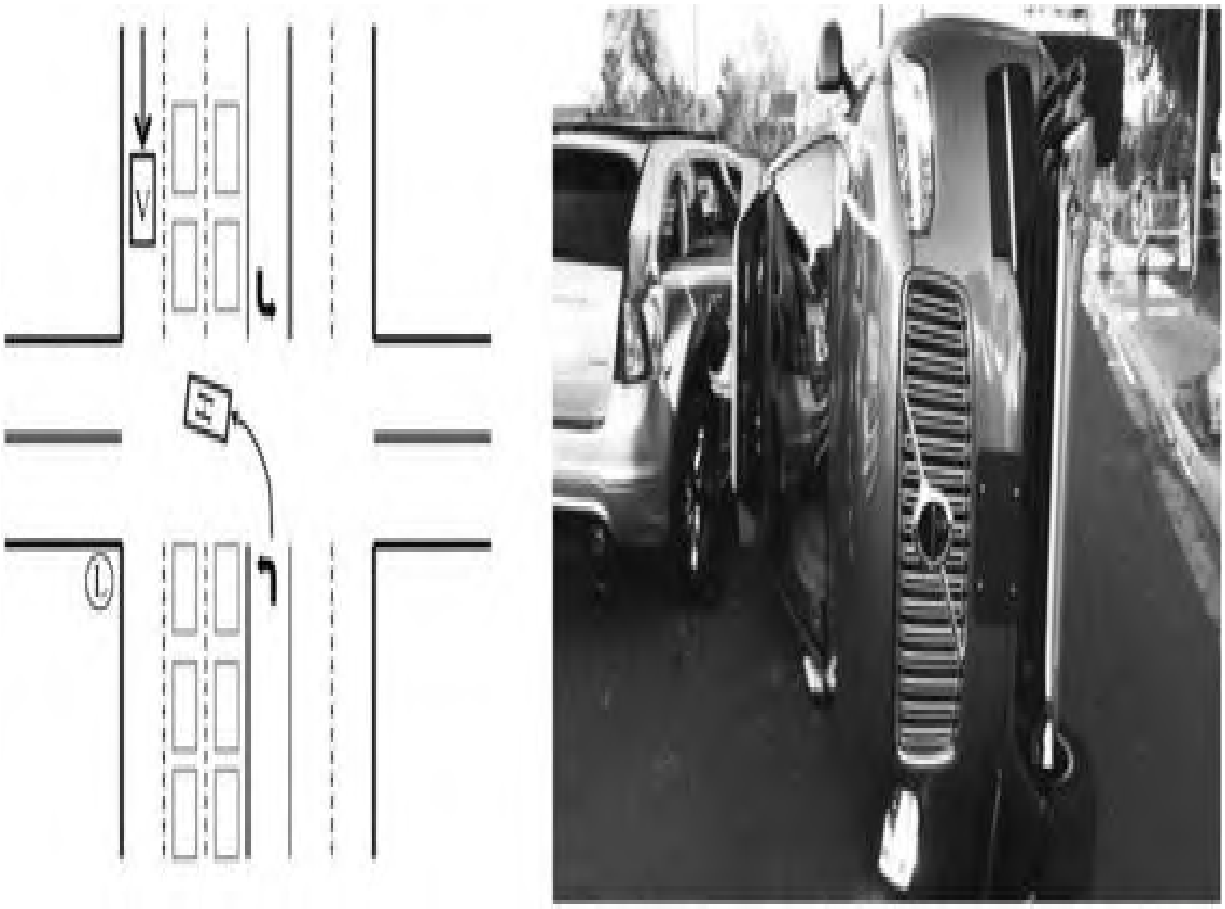


图20 （左）导致事故发生的情况示意图，（右）事故后照片

为了推断出可能存在的看不见的本田，沃尔沃可以在接近十字路口时搜集线索。特别值得注意的是，虽然绿灯亮了，但其他两条车道的车辆也是停着的，排在队伍前面的汽车没有缓慢驶入十字路口，它们亮着刹车灯。这并不是对面存在左转车辆的确凿证据，但这不需要确凿证据。即使很小的可能性也足以使沃尔沃放慢速度，更谨慎地驶入十字路口。

这个事件的寓意是，在部分可见的环境中运行的智能体必须根据它们所拥有的线索，尽可能地了解它们看不到的东西。

还有一个与家庭生活关系更密切的例子：你的钥匙在哪里？除非你碰巧在一边阅读本书一边开车（不建议这样做），否则你现在可能看不到它们。而你可能知道它们在哪里：在你现在的衣服口袋里，在

你的包里，在床头柜上，在你挂着的外套的口袋里，或者在厨房的挂钩上。你知道这一点是因为你把钥匙放在那里，此后就没动过。这是一个利用知识和推理来了解世界状况的简单例子。

没有这种能力，我们就会不知所措，多数时候都是如此。例如，当我写这篇文章时，我正看着一间普通旅馆房间的白色墙壁。我在哪里？如果我只能依靠我目前的感知输入，我的确会不知所措。事实上，我知道我在苏黎世，因为我昨天到的苏黎世，而且还没有离开。像人类一样，机器人需要知道自己在哪里，这样它们才能成功地在房间、建筑物、街道、森林和沙漠中导航。

在人工智能领域，我们使用信念状态（belief state）一词来指代智能体对世界状态的当前认知，无论这种认知是多么地不完整和不确定。一般而言，信念状态——而不是当前的感知输入，是决定做什么的正确基础。保持信念状态为最新状态是一切智能体的核心活动。对于信念状态的某些部分，这是自动发生的，例如，我似乎不需要思考就知道我在苏黎世。对于其他部分，这则是按需发生的。例如，当我经过长途旅行，在新城市中醒来时，时差严重的我可能必须有意识地努力重建自己的位置，思考自己应该做什么，以及为什么会如此，我认为这有点像重启笔记本电脑。了解并不意味着总是准确地知道世界上所有事物的状态。显然，这是不可能的，例如，我不知道在苏黎世的这家不起眼的酒店里，其他房间里现在还住着谁，更不用说地球上80亿人中的大多数人现在的位置和活动了。我根本不知道太阳系以外的宇宙其他地方发生了什么。我对当前事态的不确定感既是大量的，也是不可避免的。

了解不确定的世界的基本方法是贝叶斯更新。实现它的算法通常有两个步骤：首先是预测，在这个步骤中，智能体根据最近的动作预测世界的当前状态；然后是更新，在这个步骤中，智能体接收新的感知输入并相应地更新它的信念。为了说明它的工作原理，请考虑一下机器人在确定自身位置时所面临的问题。图21（a）给出了一个典型的

例子：机器人在房间中间，它不确定自己的确切位置，并且想要通过一扇门。机器人命令轮子朝门口移动1.5米，不幸的是，它的电机老旧且车轮晃动，因此它对轮子最终会到达哪里的预测非常不确定，如图21（b）所示。如果它现在试图继续前进，那么它很可能会撞墙。幸运的是，它有一个声呐装置可以测量到门柱的距离，如图21（c）所示，测量结果显示机器人距离左门柱约70厘米，距离右门柱约85厘米。最后，机器人通过将图21（b）中的预测与图21（c）中的测量结果相结合来更新其信念状态，得到图21（d）中的新信念状态。现在机器人对自己的位置有了很准确的认识，稍微调整一下路线就能通过这扇门了，跟踪信念状态的算法不仅可以处理位置的不确定性，还可以处理地图本身的不确定性。这就产生了一种叫作SLAM（同步定位与地图构建）的技术。SLAM是许多人工智能应用的核心组件，从增强现实系统到无人驾驶汽车和行星漫游车。

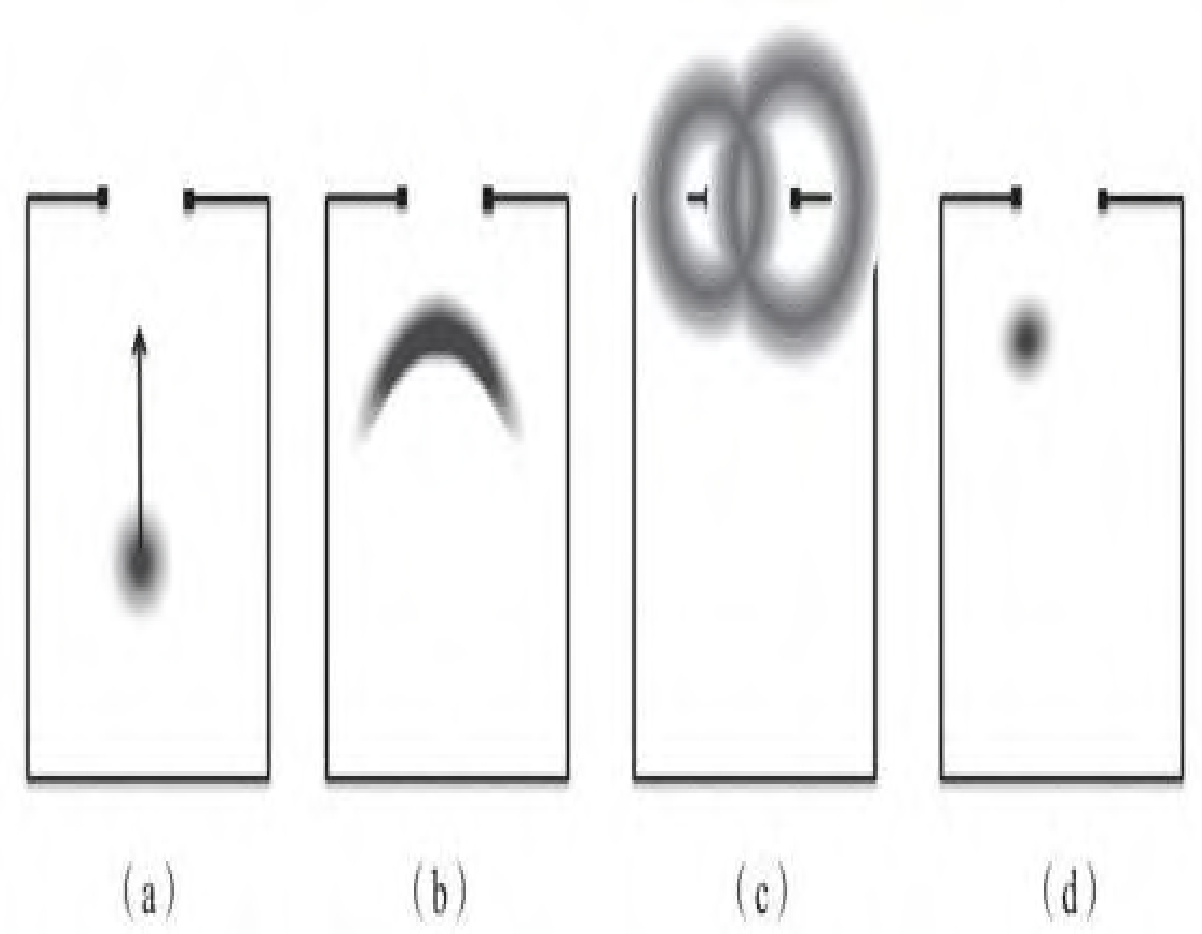


图21 一个机器人试图通过一扇门

---

[1] 2011年，珀尔的工作获得了图灵奖。

[2] 更详细的贝叶斯网络介绍：在给定节点的父节点（即，指向该节点的那些节点）的每个可能值的组合的情况下，网络中的每个节点都用每个可能值的概率进行注释。例如，当*D*1和*D*2具有相同的值时，*Doubles*12的值为真的概率为1.0，否则为0.0。一个可能的世界是给所有节点赋值。这种世界的概率是来自每个节点的适当概率的乘积。

[3] 贝叶斯网络的应用概要：Olivier Pourret, Patrick Naïm, and Bruce Marcot, eds.,*Bayesian Networks: A Practical Guide to Applications* (Wiley, 2008)。

[4] 关于概率编程的基础论文：Daphne Koller, David McAllester, and Avi Pfeffer,“Effective Bayesian inference for stochastic programs,” in *Proceedings of the 14th National Conference on Artificial Intelligence* (AAAI Press, 1997)。更多参考信息，请参见：probabilistic-program-ming.org。

[5] 使用概率编程模拟人类概念学习：Brenden Lake, Ruslan Salakhutdinov, and Joshua Tenenbaum, “Human-level concept learning through probabilistic program induction,” *Science* 350 (2015): 1332–38。

[6] 有关地震监测应用和相关概率模型的详细描述，请参见：Nimar Arora, Stuart Russell, and Erik Sudderth, “NET-VISA: Network processing vertically integrated seismic analysis,” *Bulletin of the Seismological Society of America* 103 (2013):709–29。

[7] 描述首批无人驾驶汽车发生的严重车祸之一的新闻报道：Ryan Randazzo,“Who was at fault in self-driving Uber crash? Accounts in Tempe police report disagree,” *Republic* (azcentral.com), March 29, 2017。

附录D

从经验中学习

学习意味着基于经验改进性能。对视觉感知系统而言，这可能意味着基于看到的对象类别的例子，来学习识别更多的对象类别；对基于知识的系统而言，获取更多的知识是一种学习形式，因为这意味着系统可以回答更多的问题；对AlphaGo这样的前瞻性决策系统而言，学习可能意味着提高评估棋局的能力，或提高其探索可能性之树中有用部分的能力。

从样本中学习

机器学习最常见的形式是监督学习。让监督学习算法从一组训练样本中学习，每个样本都标有正确的输出，并且必须产生关于什么是正确规则的假设。通常，监督学习系统寻求优化假设和训练样本之间的一致性。正如奥卡姆剃刀原理所提议的那样，如果假设比实际情况更复杂，系统通常也会受到惩罚。

让我们以学习围棋中的合法着法问题来说明这一点。（如果你已经知道围棋的规则，那么你会很容易理解；如果不知道，那么你会更加同情这个算法的学习计划。）我们假设算法从如下假设开始。

对于所有棋步t、所有位置l，

第t步棋在位置l处落子是合法的。

如图22所示，现在轮到黑方行棋。算法尝试A点，这是合法的。B点和C点也是合法的。然后算法尝试D点，这里有白棋的棋子，因此是非法的。（在国际象棋或西洋双陆棋中，这是合法的，因为吃掉棋子就是这种情况。）算法尝试E点，这里有黑棋的棋子，因此也是非法的。（在国际象棋中这也是非法的，但在西洋双陆棋中是合法的。）现在，从这5个训练样本中，该算法可能提出以下假设：

对于所有棋步t、所有位置l，

如果l在下第t步棋时未被占据，

那么第t步棋在位置l处落子是合法的。

然后算法尝试F点，并惊讶地发现F点是非法的。在几次失败的尝试之后，它最终判定了以下几点：

对于所有棋步t、所有位置l，

如果l在下第t步棋时未被占据，

并且l没有被对手棋子包围，

那么第t步棋在位置l处落子是合法的。

这有时被称为“禁止自杀”规则。最后，算法尝试了G点，在这个例子中G点是合法的。算法苦苦思索了一会儿，或许又尝试了几次之后，最终确定了一个假设——G点是合法的，虽然它被对手包围了，但是因为它吃掉了D处的白子，因此立刻就变得不被包围了。

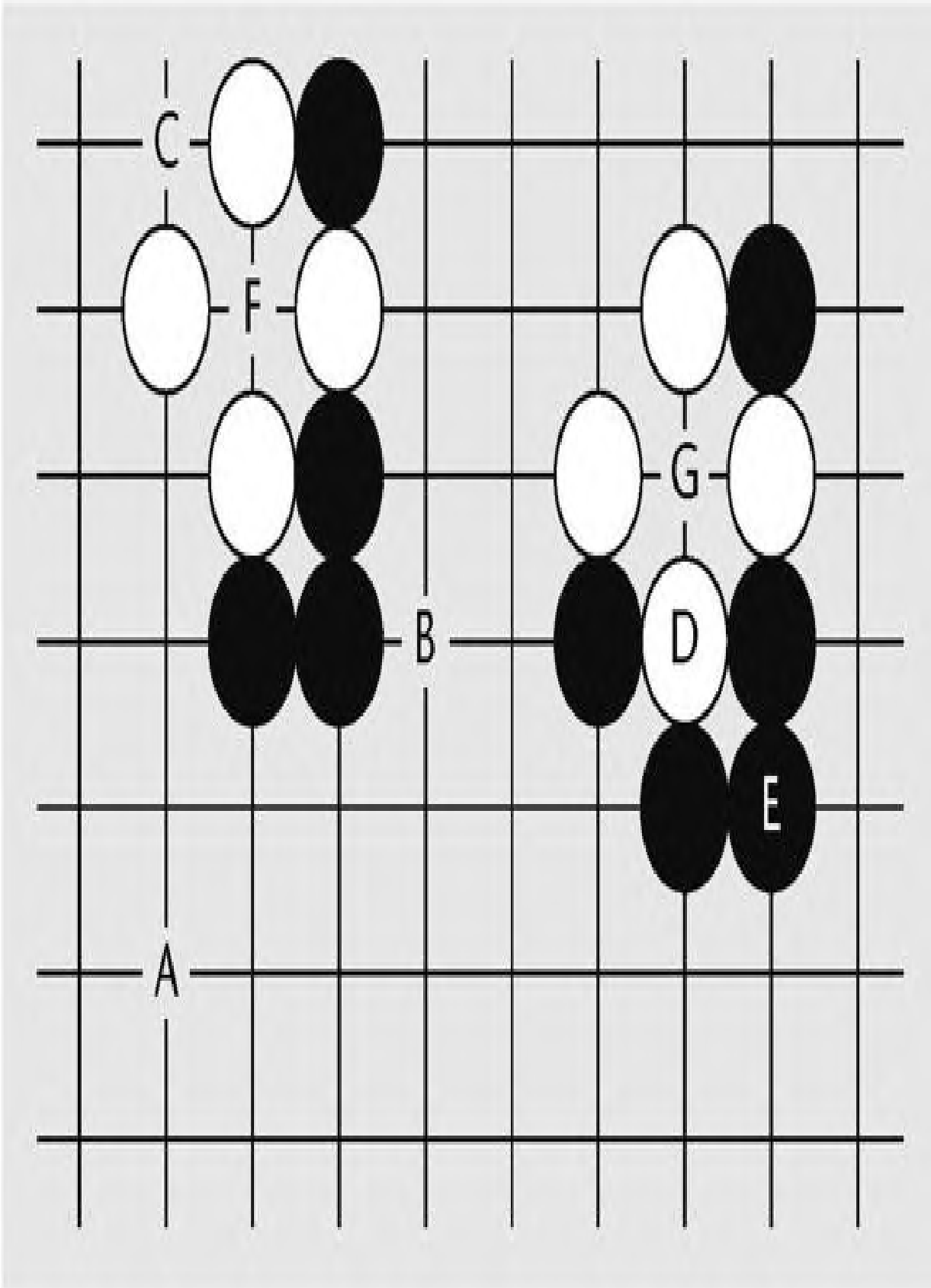

C
F
G
B
D
E
A

图22 围棋中的合法着法和非法着法

从规则的逐步发展中可以看出，学习是通过对假设进行一系列修改来适应观察到的例子而实现的。这是学习算法可以很容易做到的事情，也是第一个要点。机器学习研究人员设计了各种巧妙的算法来快速找到好的假设。在上述例子中，算法在表示围棋规则的逻辑表达式中进行搜索，但是假设也可以是表示物理定律的代数表达式，表示疾病和症状的概率贝叶斯网络，甚至可以是表示某些其他机器的复杂行为的计算机程序。

第二个要点是，即使好的假设也可能是错误的。事实上，即使在为了确保G点合法而进行了修正之后，上面给出的“G点是合法的”假设也是错误的。这需要涉及“劫”或“同形重复”[1]——如果白棋刚刚走在D点吃掉了黑棋在G点的棋子，那么黑棋不能立即走在G点重新吃掉白子，因为这会形成重复局面。请注意，这条规则与程序目前学到的完全不同，因为这意味着合法性无法从当前局面确定。相反，程序还必须记住先前的局面。

苏格兰哲学家大卫·休谟在1748年指出：归纳推理，即从“特定观察结果”到“一般原则”的推理，永远无法得到保证。[2]在现代统计学理论中，我们要求的不是保证完全正确，而是保证发现的假设大致正确。[3]学习算法可能“不走运”，它看到的是不具有代表性的样本，例如，它可能永远不会尝试像G点那样的着法，认为那类着法是非法的。它也可能无法预测一些奇怪的边缘情况，比如不重复规则中一些更复杂的、很少出现的局面所涵盖的情况。[4]但是，只要世界表现出一定程度的规律性，这个算法就不太可能产生非常糟糕的假设，因为这样的假设很可能已经被某个尝试“发现”了。

深度学习技术在媒体上引起了所有关于人工智能的喧嚣。它主要是一种监督学习的形式，代表了近几十年来人工智能领域最重要的进步之一，所以值得我们了解它的工作原理。此外，一些研究人员认

为，深度学习技术将在几年内催生出人类水平的人工智能系统，因此，评估深度学习技术是否名副其实是个好主意。

在特定任务的背景下理解深度学习是最容易的，例如学习区分长颈鹿和骆驼。给定二者的一些已标注的图像，学习算法必须形成一个假设，让它可以对未标记的图像进行分类。从计算机的角度看，一幅图像只不过是一个巨大的数字表，每个数字对应一个图像像素的三个RGB值中的一个。因此，我们需要一个以数字表作为输入，预测类别（长颈鹿或骆驼）的“长颈鹿—骆驼假设”，而不是以棋盘局面和走子着法作为输入，确定该着法是否合法的“围棋假设”。

现在的问题是，这种假设是什么样的？在过去50多年的计算机视觉研究中，人们尝试了许多方法。目前最受欢迎的是深度卷积网络。让我来解释一下它的原理：它之所以被称为网络，是因为它代表了一个复杂的数学表达式，这个表达式由许多更小的子表达式以规则的方式组合而成，并且组成具有网络结构的形式。（这种网络通常被称为神经网络，因为它们的设计者从大脑中的神经元网络获得了灵感。）它之所以被称为卷积，是因为这是一种奇特的数学方法，奇特之处在于，网络结构在图像的整个输入过程中以固定的模式自我重复。它之所以被称为“深度”网络，是因为这样的网络通常有很多层，也因为这个名字听起来令人赞叹，而且有些诡异。

图23展示了一个简化的示例（之所以简化，是因为真实的网络可能有数百层和数百万个节点）。网络实际上是一幅复杂的、可调的数学表达式的图画。如图所示，网络中的每个节点对应一个简单的可调表达式。左边输入图像像素值，网络在最右边的两个节点输出数值，表明图像是骆驼和长颈鹿的可能性有多大。请注意第一层的黑线，它表示局部连接的模式是如何在整个层级中重复的。每个输入值都有一个可调整的权重，就像图中的“量度控制旋钮”一样。然后，输入通过加权和一个门函数，然后到达节点的输出端。通常，门函数会抑制较小的数值，而允许较大的数值通过。

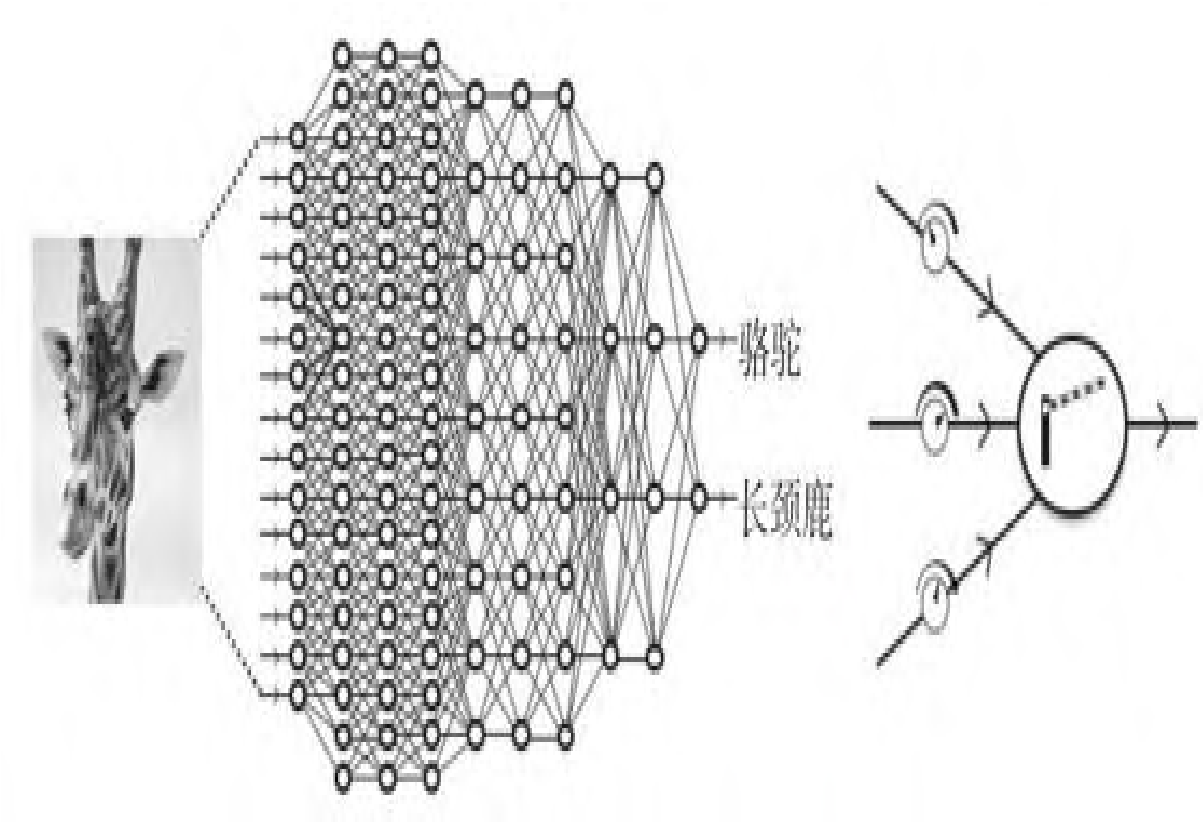


图23 （左）用于识别图像中物体的深度卷积网络的简化图示，（右）网络中的一个节点

我们只需要调节这些量度控制旋钮，减少对标记样例的预测误差，即可通过网络进行学习。原理就是这么简单，没有魔法，也没有特别巧妙的算法。找出如何调整旋钮以减小误差的方法，是对微积分的一种直接应用，用来计算改变每个权重将会如何影响输出层的误差。这就产生了一个简单的公式，用于将误差从输出层反向传播到输入层，同时调整旋钮。

这个过程奇迹般地奏效了。对于识别照片中的物体这项任务，深度学习算法表现出卓越的性能。这种概念首次出现在2012年的ImageNet竞赛中，该竞赛提供了由1 000个类别的120万张标记图像组成的训练数据，然后要求算法标记10万张新图像。[5]英国计算心理学家杰弗里·辛顿（Geoffrey Hinton）是20世纪80年代第一次神经网络革命的先驱，他一直在试验一个非常庞大的深度卷积网络，该网络

有65万个节点和6 000万个参数。他和他在多伦多大学的团队实现了15%的ImageNet错误率，比之前的最佳成绩——26%的错误率，有了显著提高。[6]到2015年，数十个团队都在使用深度学习方法，他们将错误率降到了5%，与人类花费几周时间学习识别测试中的1 000个类别的错误率相当。[7]到2017年，机器的错误率降到了2%。

大约在同一时期，基于类似方法的语音识别和机器翻译的水平也有了相应的提高。总体来看，这是人工智能最重要的三个应用领域。深度学习在强化学习的应用中也发挥了重要作用，例如，学习AlphaGo用于评估未来可能局面的可取性的评价函数，以及学习复杂机器人行为的控制器。

到目前为止，我们对深度学习如此有效的原因还知之甚少。最好的解释可能是深度网络很深：因为它们有很多层，所以每一层都可以学习从输入到输出的一个很简单的转换，而许多这样的简单转换加起来，就是从照片到类别标签所需的复杂转换。此外，深度视觉网络具有内置的结构，可以强制实现平移不变性和尺度不变性，这意味着，无论狗出现在图像中的哪个位置，无论它在图像中看上去体型有多大，它都是一只狗。

深度网络的另一个重要特性是，它们经常能发现图像的内部表示形式，正是这些形式捕捉到了图像的基本特征，例如眼睛、条纹和其他简单形状。这些特征都不是内置的。我们知道它们在网络里，因为我们可以用训练有素的网络进行实验，看看是什么样的数据会使内部节点（通常是靠近输出层的节点）亮起。实际上，我们可以用不同的方式运行学习算法，以便调整图像本身，使之在选定的内部节点产生更强的响应。多次重复这个过程就会生成现在众所周知的DeepDream（“深梦”）或Inceptionism（“启发主义”）神经网络生成的那种图像，如图24所示。[8] Inceptionism本身已经成为一种艺术形式，它生成的图像不同于任何人类艺术。

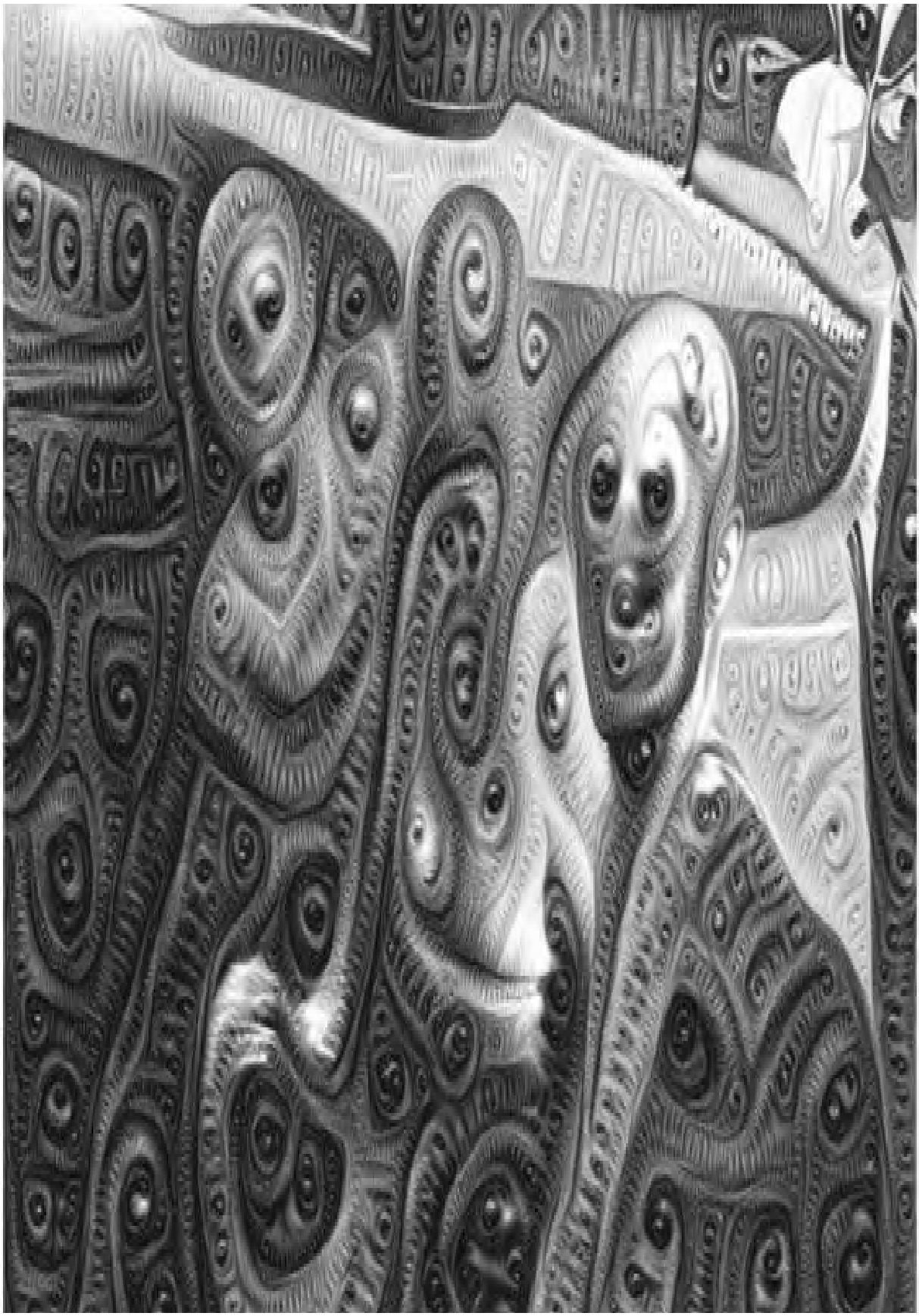




图24 谷歌的DeepDream软件生成的图像

虽然深度学习系统取得了令人瞩目的成就，但我们目前所理解的深度学习系统远不能为通用智能系统提供基础。其主要缺点是：它们是电路；它们是命题逻辑和贝叶斯网络的表亲，虽然它们有很多奇妙的特性，但它们也缺乏用简洁的方式表达复杂知识的能力。这意味着，在“原生模式”下运行的深度网络需要大量的电路来表示相当简单的常识。反过来，这又意味着要学习大量的权重，因此需要的样例数量多得离谱，比全宇宙所能提供的还要多。

有人认为，大脑也是由电路组成的，神经元相当于电路元件，因此电路可以支持人类水平的智能。的确如此，大脑是由原子构成的。原子确实可以支持人类水平的智能，但这并不意味着仅仅将大量的原子聚集在一起就能产生智能。原子必须以某种方式排列。同样的道理，电路也要按一定的方式排列。计算机也是由电路组成的，无论它们的存储器还是它们的处理单元都是如此，但是在计算机能够支持的高级编程语言和逻辑推理系统运行之前，这些电路必须以某种方式排列，并且必须添加软件层。然而，目前还没有迹象表明深度学习系统能够自己开发出这种能力，而且要求它们这样做也没有科学意义。

我们有更多的理由认为，深度学习可能会在远低于通用智能的水平上停滞不前，但我写这本书的目的不是诊断所有的问题：深度学习业内[9]以及业外[10]的其他人士已经注意到了其中的许多问题。关键是，仅仅靠构建更大更深的网络，靠更大的数据集和更大的机器，还不足以创建人类水平的人工智能。我们已经在附录B看到了DeepMind的CEO德米斯·哈萨比斯的观点，即“更高层次的思考和符号推理”对人工智能至关重要。另一位著名的深度学习专家弗朗索瓦·乔莱

（Francois Chollet）说：[11]“即使有大量的人工标记数据，许多应用对于当前的深度学习技术而言也是完全无法实现的……我们需要从简单的输入到输出的映射转向推理和抽象。”

从思考中学习

每当你发现自己不得不思考某件事时，那是因为你还不知道答案。当有人问你的新手机号码时，你可能不知道。你暗想：“好吧，我不知道，那我怎样才能找到它呢？”你不是手机的奴隶，你不知道如何找到号码。你暗想：“我如何想出找到它的方法？”对此你有一个通用的答案：“也许界面设计师把它放在了用户容易找到的地方。”（当然，你可能是错的。）显眼的地方应该在主屏幕的顶部（并不在那里），在“手机”程序中，或者在该程序的设置中。你尝试点击“设置”，然后点“手机”，找到了，就在这里。

下次有人问你要手机号码时，你要么记得手机号码，要么知道怎么找到它。你记得这个步骤，不仅是在这个场合的这部手机中，而且是在所有场合的所有类似的手机中，也就是说，你存储并重复使用了问题的通用解决方案。这种泛化是合理的，因为你知道这部特定的手机和这一特定场合的细节是不相关的。如果这种方法只在星期二对尾号为17的手机号码有效，你会感到无比震惊。

围棋为同类型的学习提供了一个完美的例子。在图25中，我们看到了一个常见的情况。黑棋威胁要包围并吃掉白棋的棋子。白棋试图延长原来的棋子并逃跑，但黑棋继续封堵逃跑路线。白棋尝试向另一个方向逃跑。最后双方按图25（e）中数字所示的次序继续落子。这种走棋方式形成了“征子”，仿佛一条斜穿过棋盘的梯子，最终一直延伸到棋盘边缘，白棋再无处可逃。黑棋完成致命一击，白棋被完全包围，死掉了。如果你是白棋，你可能不会再犯同样的错误：你意识到征子总是会导致棋子最终被吃掉，无论你是白棋还是黑棋，无论初始

位置和方向如何，在棋局的任何阶段都是如此。唯一可能的例外是征子撞上了逃跑一方的接应棋子。征子的通用性直接遵循围棋规则。

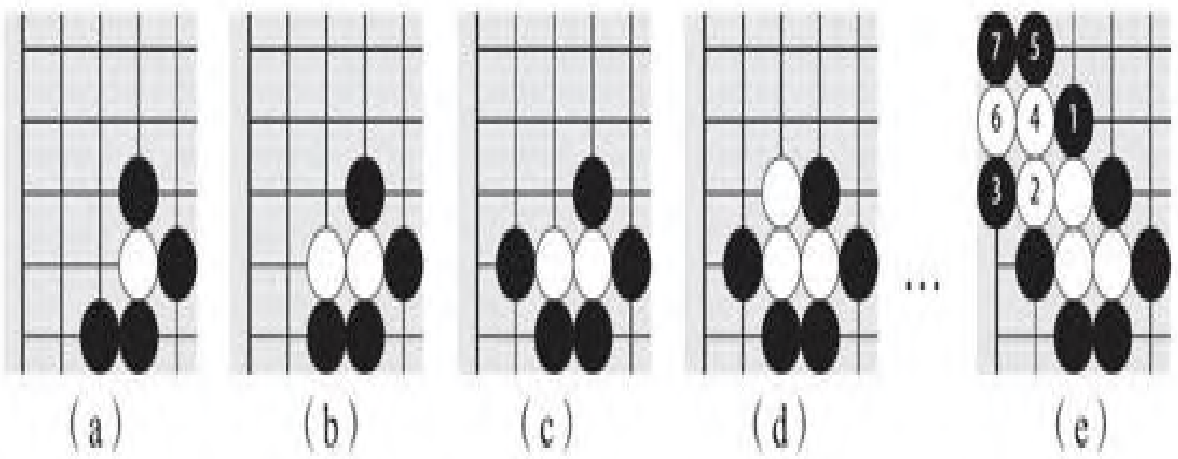


图25 围棋中“征子”的概念

忘记手机号码的情况和围棋中征子的情况说明了从一个例子中学习有效的通用规则的可能性，这与深度学习需要数百万个样例的情况相去甚远。在人工智能中，这种学习被称为基于解释的学习：看到例子后，智能体可以自己解释为什么会出现这种情况，并可以通过了解哪些因素对解释至关重要，来提取通用原理。

严格地讲，这个过程本身不会增加新的知识。例如，白棋本可以根据围棋规则推导出通用征子模式的存在和结果，而不需要看到一个例子。[12]然而，如果白棋没有看到征子的例子，它可能永远也不会发现征子的概念。因此，我们可以将基于解释的学习理解为一种以通用的方式保存计算结果的有效方法，从而避免将来不得不重复同样的推理过程（或者在不完美的推理过程中犯同样的错误）。

认知科学的研究强调了这种类型的学习在人类认知中的重要性。它以“知识块”之名，形成了艾伦·纽厄尔（Allen Newell）极具影响力的认知理论的核心支柱。[13]（纽厄尔是1956年达特茅斯会议的

参与者之一，并与赫伯特·西蒙共同获得1975年图灵奖。）“知识块”解释了人类如何通过练习而变得更熟练地处理认知任务，因为原本需要思考的各种子任务都变成了自动完成的任务。如果没有它，人类的对话将局限于一两个字的回答，数学家仍将依靠他们的手指计数。

---

[1] “劫”是围棋术语，是指黑白双方都把对方的棋子围住，这种局面下，如果轮白棋下，白棋可以吃掉一个黑子；如果轮黑棋下，黑棋同样可以吃掉一个白子。如此往复就会形成循环无解，所以围棋禁止“同形重复”。——译者注

[2] 关于归纳推理的基础讨论：David Hume, *Philosophical Essays Concerning Human Understanding* (A. Millar, 1748)。

[3] Leslie Valiant, “A theory of the learnable,” *Communications of the ACM* 27 (1984):1134–42。另请参见：Vladimir Vapnik, *Statistical Learning Theory* (Wiley, 1998)。瓦伦特（Valiant）的方法侧重于计算复杂性，万普尼克（Vapnik）的方法侧重于对各种假设的学习能力进行统计分析，但两者都有一个共同的理论核心，即连接数据和预测准确性。

[4] 例如，为了了解“situational superko”（禁止全局同形在同一方出现）和“natural situational superko”（禁止走一着棋后使全局同形在同一方出现）规则之间的区别，学习算法必须尝试停一手棋不走子来重复其先前的局面。在不同的国家，判定结果会有所不同。

[5] 有关ImageNet竞赛的介绍，请参见：Olga Russakovsky et al., “Image Net large scale visual recognition challenge,” *International Journal of Computer Vision* 115(2015): 211–52。

[6] 用于机器视觉的深度网络的首次演示：Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, “ImageNet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25*, ed. Fernando Pereira et al. (2012)。

[7] 区分100多种狗的难点：Andrej Karpathy, “What I learned from competing against a ConvNet on ImageNet,” *Andrej Karpathy Blog*, September 2, 2014。

[8] 发表于谷歌的关于inceptionism研究的博文：Alexander Mordvintsev, Christopher Olah, and Mike Tyka, “Inceptionism: Going deeper into neural networks,” *Google AI Blog*, June 17, 2015。这个想法起源于J. P. Lewis, “Creation by refinement: A creativity paradigm for gradient descent

learning networks," in *Proceedings of the IEEE International Conference on Neural Networks* (IEEE, 1988)。

[9] 杰弗里·辛顿对深度网络的重新思考：Steve LeVine, "Artificial intelligence pioneer says we need to start over," *Axios*, September 15, 2017。

[10] 深度学习的缺点一览：Gary Marcus, "Deep learning: A critical appraisal," arXiv:1801.00631 (2018)。

[11] 一本关于深度学习的流行教科书，对其弱点进行了坦率的评估：François Chollet, *Deep Learning with Python* (Manning Publications, 2017)。

[12] 对基于解释的学习的解释：Thomas Dietterich, "Learning at the knowledge level," *Machine Learning* 1 (1986): 287–315。

[13] 对基于解释的学习的一种表面上完全不同的解释：John Laird, Paul Rosenbloom, and Allen Newell, "Chunking in Soar: The anatomy of a general learning mechanism," *Machine Learning* 1 (1986): 11–46。

致谢

很多人在这本书的创作中给予了帮助。包括我在维京出版社和企鹅出版社的优秀编辑保罗·斯洛瓦克（Paul Slovak）和劳拉·斯蒂克尼（Laura Stickney），我的图书代理人、一直在鼓励我写作的约翰·布罗克曼（John Brockman），提供了大量有用反馈的吉尔·里奥维（Jill Leovy）和罗布·里德（Rob Reid），以及早期书稿的其他读者，特别是齐亚德·马拉尔（Ziyad Marar）、尼克·海伊（Nick Hay）、托比·奥德（Toby Ord）、戴维·杜文诺（David Duvenaud）、迈克斯·泰格马克和格蕾丝·凯西（Grace Cassy）。卡罗琳·詹梅尔（Caroline Jeanmaire）在整理早期读者提出的无数改进建议方面提供了极大的帮助，马丁·福井（Martin Fukui）负责搜集了图片的许可。

书中的主要技术思想是与加州大学伯克利分校人类兼容人工智能中心的成员合作发展而来的，特别是汤姆·格里菲斯（Tom Griffiths），安卡·德拉甘（Anca Dragan），安德鲁·克里奇（Andrew Critch），迪伦·哈德菲尔德·梅内尔（Dylan Hadfield Menell），罗欣·沙阿（Rohin Shah）和史密沙·米利（Smitha Milli）。该中心由执行董事马克·尼茨伯格（Mark Nitzberg）和助理董事罗西·坎贝尔（Rosie Campbell）领导，并得到了开放慈善基金会的慷慨资助。

雷蒙娜·阿尔瓦雷斯（Ramona Alvarez）和卡琳·韦尔多（Carine Verdeau）让整个过程得以顺利进行，我优秀的妻子洛伊和我们的孩子戈登、露西、乔治和艾萨克在我的创作过程中给予了无穷无尽且又必不可少的爱、忍耐和鼓励，以上排名不分先后。

图片版权

图 2 (b) © The Sun / News Licensing; (c) Courtesy of Smithsonian Institution Archives.

图 4 © SRI International. creativecommons.org/licenses/by/3.0/legalcode.

图5 (left) © Berkeley AI Research Lab; (right) © Boston Dynamics.

图 6 © The Saul Steinberg Foundation / Artists Rights Society (ARS), New York.

图7 (left) © Noam Eshel, Defense Update; (right) © Future of Life Institute / Stuart Russell.

图10 (left) © AFP; (right) Courtesy of Henrik Sorensen.

图11 *Elysium* © 2013 MRC II Distribution Company L.P. All Rights Reserved.Courtesy of Columbia Pictures.

图 14 © OpenStreetMap contributors.OpenStreetMap.org. creativecommons.org/licenses/by/2.0/legalcode.

图19 Terrain photo: Digital Globe via Getty Images.

图20 (right) Courtesy of the Tempe Police Department.

图 24 © Jessica Mullen / Deep Dreamscope.creativecommons.org/licenses/by/2.0/legalcode.

图书在版编目（CIP）数据

AI新生：破解人机共存密码：人类最后一个大问题/（美）斯图尔特·罗素著；张羿译. -- 北京：中信出版社，2020.9
书名原文：Human Compatible: Artificial Intelligence and the Problem of Control
ISBN 978-7-5217-1951-2
Ⅰ. ① A… Ⅱ . ①斯… ②张… Ⅲ . ①人工智能 Ⅳ.① TP18
中国版本图书馆CIP数据核字（2020）第101830号

Human Compatible: Artificial Intelligence and the Problem of Control by Stuart Russell
Copyright © 2019 by Stuart Russell
Simplified Chinese translation copyright © 2020 by CITIC Press Corporation
All Rights Reserved.

AI新生：破解人机共存密码——人类最后一个大问题
著者：［美］斯图尔特·罗素
译者：张羿
出版发行：中信出版集团股份有限公司
（北京市朝阳区惠新东街甲4号富盛大厦2座 邮编100029）

字数：301千字
版次：2020年9月第1版
京权图字：01－2020－1078
书号：ISBN 978－7－5217－1951－2

版权所有·侵权必究


AI Thinking
思维

AI思维

从数据中创造价值的炼金术

丁磊_著

“数据科学50人”之一 / PayPal全球消费者数据科学部创办人

深度分析机器学习的内在逻辑
洞见大数据与人工智能如何赋能商业

知道趋势，更需要知道怎么做

中信出版集团

AI思维

——从数据中创造价值的炼金术

丁磊 著

中信出版集团

目录

推荐序 “AI思维”——人工智能的落地指南

前言

第一章 AI思维是时代机遇

大赚1 000亿美元的秘密

寻找你的“阿尔法”

AI思维的核心价值

第二章 AI思维的底层逻辑

模型的泛化能力

方差和偏差的权衡

相关性和因果性

数据的规律性

第三章 AI炼金术：数据产生价值

从数据到价值的历程

AI如何做出决策

业务优化方法

会思考的人工智能

AI炼金术的应用生态

第四章 人人都能理解AI

AI的分类和聚类

AI感知图像和语音

AI理解自然语言
AI理解商业数据
AI理解消费数据
AI理解社交数据

第五章 数字化赋能

行业数字化是大势所趋
工业赋能：数字化的物理世界
零售赋能：智能产品定位和营销

第六章 企业级AI中台

美好而混乱：AI治理的困境
以生产目标为核心，提升智能制造水平
以消费者为核心，赋能智慧新零售

第七章 AI实战的故事

零售业AI实战
工业AI实战
金融业AI实战

第八章 AI的挑战与对策

AI模型的挑战
AI的伦理和法律挑战
AI落地的人为因素

参考文献

后记

推荐序

“AI思维”——人工智能的落地指南

肖京[1]

近年来，AI越来越成为万众瞩目的焦点。随着大数据、物联网和5G（第5代移动通信技术）等技术的发展，“万物皆可AI”不再仅仅是一个美好的愿景，人工智能早已进入我们的生活，并作为核心驱动力，对各行各业进行了深刻的塑造和改变，推动着国家战略、经济结构和商业模式的升级革新。纵观生活的方方面面，人工智能已经和我们的日常生活连接在了一起，并且与每个人都息息相关。智能革命是大势所趋。

人工智能可以分为弱人工智能、强人工智能和超人工智能。目前人工智能处在弱人工智能阶段，这意味着人工智能的能力是针对特定领域的，它只在某些方面比人强。因此，人工智能作为一个赋能的体系帮助行业升级换代，一定要跟具体的应用领域深度融合在一起。人工智能的应用领域非常广泛，在零售、工业、金融、教育和医疗等行业中起着越来越重要的作用。人工智能的落地也迫使人类社会重新分工，促使这些传统行业迅速转型与升级。

我们日常生活的方方面面都需要决策，不论是应用于哪种业务场景下的人工智能系统，想要发挥作用，都需要通过模型来解读数据并作用于业务，达到优化决策的效果。思维的高度决定了决策的质量，因此，如果说人工智能是未来社会发展趋势的话，那么我们有必要做一些相关知识的普及和思维的传播。人工智能技术不仅可以改变我们

的生活方式，当它转化为一种思维时，更能真正地为我们的人生带来裨益。

我与丁磊博士在人工智能领域有颇多的交流，我们都认为人工智能不应该是浮于表面的概念。人工智能具有深刻的内涵，值得我们静下心来研究，也值得生活中的每个人去理解。人人都能学AI，人人都能懂AI，AI也能为人人所用。丁磊博士在人工智能领域工作近20年，总结出了一套颇具特色的人工智能知识体系和思维方式——“AI思维”。AI思维是一种通过数据驱动决策的思维模式，包括数据、模型、算力和业务模式四要素。从数据出发，通过模型和算力形成决策，最终在业务中产生价值。因此，AI思维的基础在于数据，核心在于模型，实现在于算力，应用在于业务模式。从这种意义上说，本书深入浅出，从“道”“法”“术”“器”“用”“势”几个方面循序渐进地向读者介绍了从数据中创造价值的“炼金术”。

毋庸置疑，人脑和人工智能在处理信息上存在相似之处，我们对事物的认识可以分为四个阶段：信息、经验、规律和思维。这是一个动态的过程，从单个信息的联系中形成经验，在经验中发现规律，再提炼升华出思维。思维阶段，也是方法论层次的总结，进一步指导实践，不断找出更多的规律和结论。人工智能的底层逻辑就是数学规律，万丈高楼平地起，从底层逻辑中衍生出人工智能的知识体系。人工智能落地意味着从数据中产生价值，在这个过程中人工智能既是方法和工具，也是思维逻辑，可以指导行业的应用和发展。

未来的时代一定是智能化的时代，人工智能作为各个行业的风口，也必然会成为未来社会的发展趋势。《AI思维》总结了人工智能落地实践的全过程，并且具有优化各种业务的突出价值。而对于个人来说，AI思维可以帮助我们更深入地理解我们所处的时代，把握时代的机遇和发展方向，最终帮助我们更好地生活和工作。这是人工智能的使命，也是我们理解人工智能的意义。

[1] 此推荐序作者肖京博士，现任中国平安集团首席科学家、集团执委，曾获第九届吴文俊人工智能杰出贡献奖。

前言

数字时代，既是机遇，也是挑战。互联网、大数据、人工智能的发展使得生产、商业和生活的各个领域都发生了改变，新生信息层出不穷，知识总量爆炸式增长。面对这个日新月异的时代，我们——渺小的个体，有限的精力，受限的知识面——唯有不断更新自我，提升认知，才不至于被激烈而残酷的竞争所淘汰。

我们要改变原有的认知方式，如果单靠知识填充，难以达到质的改变，需要转换思维模式。我一直相信，万事万物皆有规律，只有快速把握规律，才能理性认识存在，提升思维能力，抓住时代脉搏。当然，数字世界也不例外，数据之中也有规律性。在利用人工智能（AI）发现和挖掘数据规律与价值的实践中，我认识到人工智能不仅是一种技术方法，其中蕴含了一种思维逻辑，也就是AI思维，这才是人工智能的真正精髓，它集中体现在对数据理解方面的优势，能够帮助我们突破认知局限，从而找到自身的真正优势，获得展现个体主动策略优势的收益。

我与人工智能的故事早在20年前就开始了。世纪之交入学浙江大学，“形上谓道兮，形下谓器”的校歌时至今日仍不时回荡在我的耳边。我对人工智能的探索便开始于在浙江大学读书的阶段。大学四年，我从“器”的层面了解人工智能、机器学习、模糊数学以及神经网络等领域。当时，越来越多的富媒体数据出现在互联网上，我便猜测人工智能接下来的使命是让计算机自动理解这些数据，从而增进用户的效率和体验。读博士时，我毫不犹豫地将当时相对冷门的人工智能作为专业方向，因为我深信人工智能将成为信息技术的下一范式。

我对AI思维的认识和总结，是从实践中不断丰富而来的。在美国博士毕业后，我先后经历过高等研究院所、国际大型企业以及各个阶段的创业公司，服务过IBM（国际商业机器公司）、亿客行（Expedia）、Orbitz（旅程网）和PayPal（贝宝）等公司。在硅谷，我也成为将人工智能运用于深度商业运营的早期实践者和探路人之一。早在2011年，我领导建立的人工智能系统在全球最大的在线旅游平台亿客行上线，通过电商网站访客的行为信号，精准预测访客线上消费行为，并以此为基础投放恰当的广告。这是业内较早、规模较大的电商行为预测和智能营销平台，至今仍优化着全球数以亿计的用户体验。后来，作为PayPal全球消费者数据科学部的创立者和负责人，我在电商、支付和广告领域通过人工智能和深度学习技术，领导团队开发了基于大数据的消费者动机预测引擎、精准推荐引擎以及最佳行动引擎，大幅度提升了全球数百万商家的赢利性和数亿用户的购物体验，并因此受邀在哈佛大学、麻省理工学院等学校做专题演讲。

在美国工作期间，我同样关注国内人工智能的应用和发展，发现国内人工智能的研发与商业运用之间存在脱节的问题。而后我应邀回国工作，致力于发掘人工智能在行业中新的增长点。在百度金融任职首席数据科学家期间，我负责金融领域的数据挖掘和变现，通过大规模特征挖掘和深度学习技术，从海量的互联网数据中提取金融相关变量，提升金融产品的市场竞争力，运用人工智能帮助金融机构升级换代，促进普惠金融的健康发展。

从对人工智能模型的攻坚克难到运用人工智能帮助企业解决业务问题，在长期实践中，我逐渐体会到，人工智能作为一种技术，可以用于商业活动中，处理数据，提高效率，提升体验。但是，人工智能远不仅是一种技术。当下人人都在说人工智能，但往往都说不清。自从1956年达特茅斯会议“人工智能”的概念正式提出，人工智能既经历过至暗时刻，也经历过辉煌岁月。什么是人工智能？这是一个很多人在尝试回答，但是又感觉力不从心的问题。

提到人工智能，大众的认知普遍是无人驾驶、视觉识别、语音识别或者是像AlphaGo（阿尔法围棋）那样的机器人等，但其实这些只是人工智能的具体应用，只是人工智能的冰山一角，却在很多人心中成了人工智能的代名词，这说明目前对人工智能的认识还有局限性。运用于商业各个领域以及工业生产、日常生活中的人工智能是有共通之处的，可以从“器”的层面总结升华出“道”，即一套特有的思维模式——AI思维。人工智能不仅仅是一种先进技术，其核心意义是一种分析数据的思维模式。深受我的合作导师、计算机视觉之父黄煦涛等几位学术大师的影响，我深信一切现象和对应的数据必有规律性，而AI思维就是发现和发掘这种规律性的强大工具。

AI思维，简单来说就是从大量数据中形成模型，进而对未知情况做出最佳预测的思维模式。这样的预测无论对个人、组织还是社会整体，都有积极的意义，能够避免经验主义带来的主观、片面的判断。AI思维的基础在于数据，核心在于模型，实现在于算力，应用在于业务模式。AI思维与大数据相伴相生，恰如炼油的过程，大数据是原油，AI思维则是从原油中提炼产品产生价值的“炼金术”。

AI思维也是对现有科学体系进行“数据驱动”升级的方法论。人工智能给社会带来的价值远不只哗众取宠的“黑科技”，也不只是给人们带来便捷或者帮助企业重塑商业模式，它最终能够关乎民生的方方面面。无论是在零售还是金融行业，抑或是在医疗和教育领域，通过AI思维你都能更快速、更直接、更准确地预测个体的行为或表现结果，从而既能满足行业发展本身的需求，又能在实践中创造出前所未有的价值，将先进的经验和知识惠及所有人。想象一下，通过人工智能，所有的病人可以获得全球最优秀医生的诊断和治疗，所有的孩子可以获得全球最先进的教育方法的培养，我们生活的世界将会变得怎么样？“千百支蜡烛可以被一支蜡烛点燃，而那支蜡烛的生命不会因此缩短。”在人工智能时代，每个人的数据都可以转化为知识，帮助别人并由此受益。

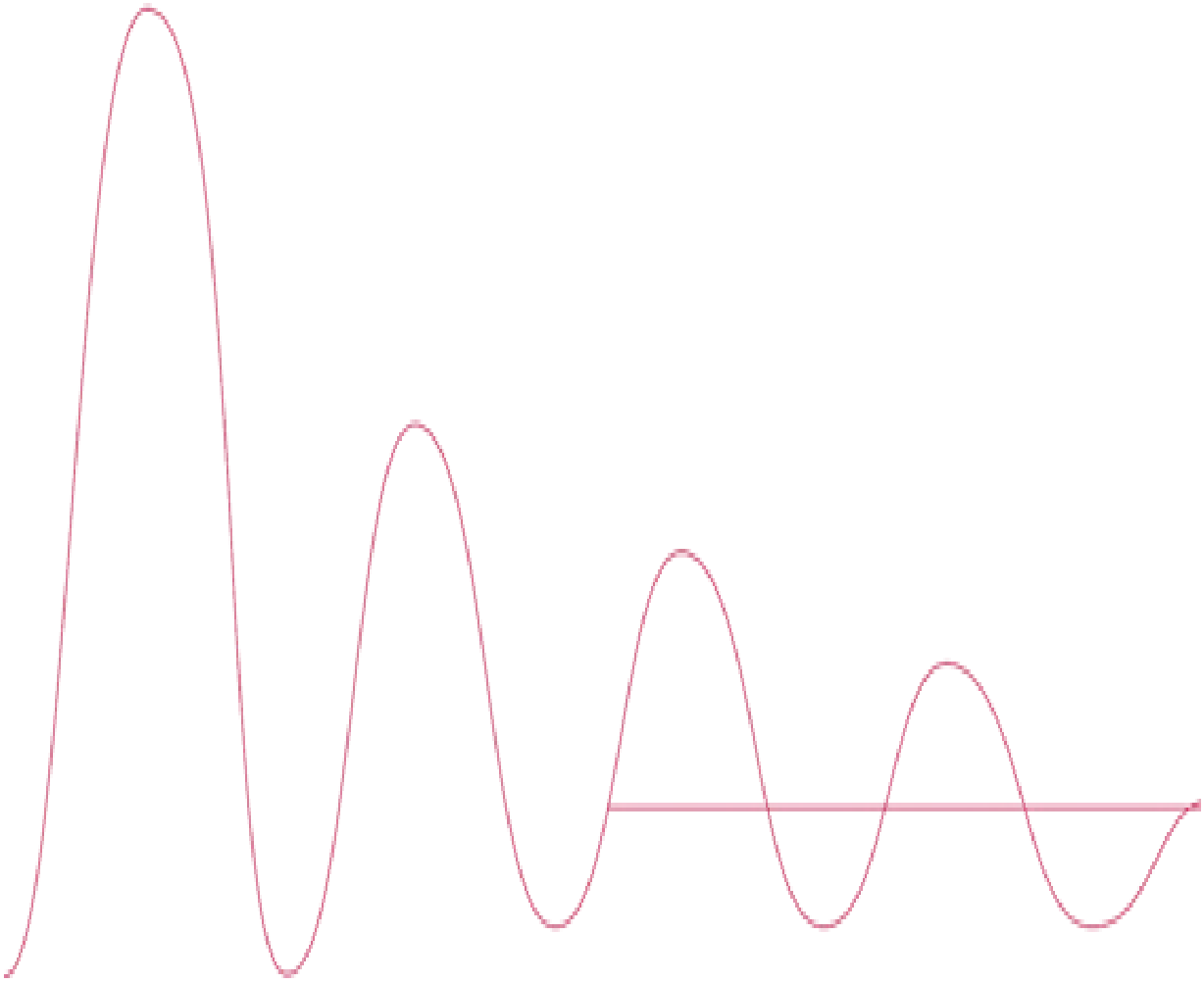
我阅读过很多面向大众的人工智能方面的书籍，其中讲述“器”和“用”的居多，很少涉及“道、法、术”层面。而且，“器”的介绍较容易变成“动物园”模式，仅仅介绍了人工智能算法，没有总结上升到理论和法则的高度，而“用”层面的案例也主要集中在智能机器人和无人驾驶汽车等吸引眼球的领域，缺乏必要的系统化行业落地框架和实践。这也难怪读者很容易将人工智能误认为是明天的科技，或将人工智能局限地认定在几个特定的领域。在这样的理解下，很难把握人工智能的思想精髓，人工智能对个人和社会的潜在影响也大打折扣。所以，本书试图打破这种局限，分为八章，精心设计“道、法、术、器、用、势”体系，旨在避免上述的问题，既包括AI思维的方法论，也包括AI思维在行业中的应用和发展，帮助读者从宏观到微观、从理论到实践，一步步、一层层地通过系统性框架，逐渐脱下人工智能神秘的外衣，帮助读者理解AI思维的实质以及对每个人的价值和意义。这是我创作本书的初心，但也深知长路漫漫，任重道远。

具体来说，本书第一章从“道”的层面阐述AI思维内在逻辑和价值，用AI思维来思考问题；第二章从“法”的层面介绍AI思维的底层逻辑，从规律的角度了解AI思维；第三章从“术”的层面讲解AI炼金术如何通过数据驱动决策，形成一套从数据到价值的完整方法论；第四章从“器”的层面引入人工智能落地所涉及的方法和模型，模型在AI思维中具有核心位置；第五、六、七章从“用”的层面分别介绍数字化赋能、人工智能平台化思维和人工智能实战案例；第八章从“势”的层面介绍人工智能落地过程中存在的挑战和相应的对策，期待人工智能未来的发展。

本书的章节分别从不同角度逐一阐述AI思维的价值和意义，大部分内容都以“数据+人工智能+场景=价值”作为核心逻辑，希望任何一个不想在新经济中落伍的人，在阅读本书之后，都能理解AI思维，并对工作和生活有所助益。

第一章

AI思维是时代机遇



理论上……假如有这么一台计算机，能存储世界上所有的事实，同时拥有完美的程序，能以数学方式表达世界所有不同部分之间的所有关系，我们就能完美地预见未来。

—《原则》—

人工智能的概念自1956年的达特茅斯会议诞生至今，曾经受到万众瞩目，也曾陷入低谷，黯淡无光。互联网、虚拟现实、增强现实、区块链等新技术如雨后春笋般不断涌现，但千帆过尽，人工智能仍然安如磐石，屹立不倒。2016年人工智能机器人AlphaGo战胜韩国九段围棋国手李世石之后，人工智能强势进入大众视野。近年来，中、美、英、俄等越来越多的国家将人工智能的发展上升到国家发展战略和规划，各行各业也都想利用人工智能抢占先机、推陈出新，而普通人的生活也因为人工智能的点缀而愈加便捷和丰富，把不可能变成可能。

万物皆可人工智能的时代已经到来，如果不想淹没在来势汹涌的人工智能浪潮里，墨守成规、故步自封显然是行不通的，只能拥抱未来，提升认知，获得助力。但是，你真的了解人工智能吗？你知道人工智能是如何与商业碰撞迸发出火花的吗？你知道人工智能如何从数据中产生价值的吗？理解事物，就要抓住其核心理念。AI思维就是人工智能的核心理念。那么首先，我们要从“道”的层面理解AI思维内在逻辑和价值，从AI思维的角度，理解人工智能。

大赚1 000亿美元的秘密

如果你读过《从0到1》这本书，应该对PayPal这个公司有所了解——硅谷创投教父、PayPal创始人彼得·蒂尔便是这本书的作者。另外，PayPal黑帮（PayPal Mafi a）的说法也由来已久。2002年的时候，eBay（亿贝）以15亿美元的价格收购了PayPal，当时PayPal公司的许多重要员工纷纷选择创业，自立门户。PayPal也因此成为硅谷史上创业者群体最多的一家公司，这些人被称为PayPal黑帮。黑帮成员们创建了领英（LinkedIn）、YouTube（优兔）、Yelp（美国最大点评

网站）、Space X（太空探索技术公司）和特斯拉（Tesla）等多家大名鼎鼎的公司，有着十分强大的社会影响力。

PayPal是美国一家提供在线支付服务的公司。马云当年受PayPal的启发，创立了支付宝。PayPal近几年市值的增幅非常显著，远高于同期纽交所和纳斯达克的增幅。PayPal公司2015年的市值接近400亿美元，到2019年上半年市值增加至3倍多，已经快1 300亿美元了——这个体量的市值，在美股互联网公司里位居前十。

PayPal能如此快速发展起来，很大程度上是得益于人工智能系统提供的强大动力。我在2012年初加入PayPal公司，负责建立了PayPal全球数据科学平台。当时PayPal对数据的加工和利用还相对比较滞后，不能跟谷歌这样的搜索引擎公司相提并论。值得庆幸的是，在2012年，PayPal公司把营销、广告、数字化方向的人工智能团队收拢在一起，整合构建了在硅谷都颇有影响力的PayPal数据科学部。而后，数据科学部经过数年的努力，在PayPal用户端内容的个性化呈现、为商家定制的“千人千面”广告推荐，以及在核心产品体验的提升上，全都通过人工智能进行了系统性的升级改造。

当时，虽然PayPal在全世界大多数国家和地区都有用户，可最大的市场还是在美国，但是美国国内的用户数已经非常饱和，短时间内无法再有太大的增量。既然开发新用户的难度很大，PayPal决定另辟蹊径，那就是增加每位用户的交易额。为了增加用户的交易额，PayPal数据科学部的任务就是为PayPal所有的用户精准推荐商家和商品。

理想情况下，PayPal想知道用户在每一个时间点可能会有哪些需求：比如为了周五晚上聚会，你打算到日用品网站购买些派对用品；或者周日中午，你会想去餐厅吃一顿大餐。有了关于这些需求的预测后，通过定向推荐全渠道精准的广告，PayPal可以把消费者和最能打动他们的产品或服务连接在一起，从而提高每位用户的消费金额，实现巨大的增量商业价值。

知易行难，PayPal用了两年多的时间，才成功地建立了自己的人工智能平台。这个平台主要由消费者动机预测引擎、精准推荐引擎以及优化客户生命周期的最佳行动引擎构成，而这几个引擎均部署在大数据平台之上，跟海量的底层数据衔接，面对动态的业务需求能够自动地构建解决方案。

其中，动机预测引擎通过人工智能，预测用户在某一个时间点的行为，包括点击广告、兑现优惠券、购买商品等；精准推荐引擎连接用户和他可能感兴趣的商家或商品，做出恰当的推荐；最佳行动引擎则通过人工智能优化在每一个时间点，商家采用什么样的策略，能够最有效地提升用户的价值。三个引擎相互结合，就有了精准理解消费者行为的能力，就像找到打开宝库的“芝麻开门”，把最恰当的商品和商家服务送到消费者面前。

人工智能平台为PayPal带来了成倍的交易增长速度。通过人工智能的精准预测、推荐和优化，PayPal的营销活动平均响应率提升了2~3倍，部分营销活动的点击率和兑现率超过5倍，总体投资回报率（ROI）也得到大幅提升，对公司基本面产生很大的影响。但是，通常来说，人工智能对公司的影响并不一定立刻体现，而是随着时间的推移，逐步释放出来。当人工智能平台已经有一定的规模以后，所带来的业务加速的增量会变得更大。2014年底，PayPal实现了19.4亿美元的利润，比2012年底提升了近30%；2015年的利润甚至超过了母公司eBay。随后PayPal从eBay剥离，单独上市，至今市值已经增长了数倍。

2018年我接受第一财经的采访，发表过一篇访谈文章《最好的数据科学家，就是把自己给“干掉”》。为什么数据科学家要把自己“干掉”呢？人工智能提升业绩依赖的主要是平台，而不是仅仅靠几位优秀的数据科学家。当然，搭建一个平台需要专业而尽职的团队，但是落地投产以后，优秀的人工智能平台可以独立运行，而不再依赖于某位数据科学家本身。换句话说，人工智能平台就像给机械化的工

厂装上了一个“大脑”，在保持高生产效率的情况下，还能千人千面地提供个性化的产品服务，这就是人工智能平台化的魅力所在，也是数据科学家被“干掉”的原因所在。如果一个企业已经形成了人工智能平台化的运营方式——正如PayPal对用户的人工智能精准推荐体系，可以实时对全球的三亿用户进行智能推荐，这套体系是可以持续发挥价值的，并且随着时间的推移，用户的黏性越来越强，数据的积累越来越多，这个价值会释放得更加厉害。

现今很多人使用的今日头条、抖音也有内容推荐系统，它们推荐的内容确实是有黏性的。随着用户的使用，积累和沉淀的数据越来越多，平台内容的推荐也就越来越精准，这样就会赢得更多用户的青睐。所以，在以人工智能为核心的智能经济里，只要在恰当的方向上建立了相应的人工智能平台和体系，人工智能就会通过不断的学习形成正向的反馈。正向反馈对应的是业绩更加快速的增长，最终形成一个极具规模的叠加效应，而不是简单的线性增长。这样的叠加效应，依赖于人工智能平台本身的演化，并随着时间的推移产出更多的增量价值。

利己也利竞争对手

亿客行是全球最大的在线旅游公司之一，业务遍及全球，2007年通过与艺龙的合作正式进入中国市场。在人工智能的支持下，亿客行在广告变现领域满载而归。2010年时，亿客行的广告营收在1亿美元左右，这个数字对于行业巨头来说不是特别大。2011年，我帮亿客行开发了一套人工智能系统。这套系统上线以后，亿客行的广告营收获得了大幅度的提升。在这个过程中，亿客行坚持了一件事情：通过人工智能系统预测用户在网站上产生购买行为的概率，也就是预测一个访客在网站上转化为交易用户的概率。

人工智能具体预测什么事情呢？如果用户打开亿客行网站或者在手机App（应用程序）上搜索一个行程，比如从北京到洛杉矶的航班，亿客行的人工智能系统可以实时地预测这个用户的真正行为目标和意图。系统在用户点击“搜索”按钮的这一瞬间触发几百个用户相关的行为数据维度的分析，非常迅速地产出预测结果，通常在几十毫秒内完成。如果这个用户真正要购票的话，人工智能不会给你弹出广告，免得打扰用户，干扰即将发生的交易；如果用户只是试探性查询，只是想要对比下不同网站的购票价格，人工智能就会给用户弹出竞争性的广告。

什么是竞争性的广告呢？这些广告主本身并没有入驻亿客行，所以这些竞争性广告相当于是从站外引入的。正因为这些站外商家和本网站主营内容相似，对于想要比价或者试探性查询的客户来说，这类广告是具有高意向性的广告，广告的点击率远高于平均水平，所以这类竞争性广告的卖价比一般的展示性广告，也就是我们经常在网站页面上看到的横幅广告也要高很多。这个模式成立的关键点在于，亿客行通过人工智能预测引擎，可以很准确地预测出来亿客行网站的某个访客是真正来购票的，还是只是来亿客行对比下价格，对比完就离开，到别的网站去买票的。

这个预测引擎搭建完成以后，它的直接效果是把大量不购买的网站流量做了广告的变现。有统计数据表明，这类不购买的用户占了总访客数的97%。用户既然来网站只是看看，并不购买，那么，网站给用户推荐一个相关类型广告，用户只要点击了广告，广告主就会给网站付费。同时，预测引擎也做到了完备的风险控制，否则如果用户真来网站上购买产品，而网站给用户推荐了竞争性的广告，对网站来说就会造成亏损。

用户如果真的打算去国外游玩，在亿客行上搜游玩行程，多半看不到这种竞争性广告，因为亿客行的广告策略不允许降低网站本身的交易额。但是如果用户只是搜着玩儿，或者看完价格以后要到别的旅

游网站或者航空公司去购票，用户多半会看到这类竞争性的广告，而用户的每次广告点击都会给亿客行带来收入。

亿客行的人工智能系统能够预测出来哪些人会购买，哪些人不会买，从而做到精准的广告推荐，这样既增加了额外的广告收益，又控制了潜在的对网站本身交易的风险。由于人工智能对每位访客的意图做出了准确的预测，这套人工智能广告系统在2011年上线之后，亿客行的广告营收稳步增长，到2018年已经超过10亿美元。

亿客行部署的人工智能预测引擎，是人工智能商业化一个早期的实践，后来这个系统的延伸版用在更多的在线交易场景下，也被美国风险投资商经纬创投、红点创投等国际知名风投公司所追捧。通过人工智能的精准预测，网站实时了解到用户的真正意图，把更恰当的内容呈现给用户，从而做到网站经营者、广告主和用户的多方共赢。

BAT[1]做金融的秘密武器

人工智能在金融领域中也大有可为。举一个例子，假如一个人刚踏入社会，参加工作时间很短，每个月收入不高。如果他去银行申请贷款，会得到什么样的答复呢？因为收入太低，银行没有很多流水记录，贷款申请会被拒绝，这很可能是当前银行基于现有数据做出的决定。

但是，让我们换一个视角来看，同样是这个人，在大数据的棱镜下，他经常参加网络课程和社会福利活动，爱好音乐和阅读，喜欢软件开发的交流。如果给这么一个人审批，作为金融机构的审批人员，会不会批准他的申请呢？在这种视角下，金融机构很有可能会给他提供一些信贷服务，因为他的兴趣爱好广泛，行为健康向上。虽然在金融领域没有强大的数据支持，但是他在社交和专业活动等方面正面且稳定的行为，可以作为金融机构审批信贷的参照依据。金融行业审批

贷款的趋势是使用更多种类的非金融属性数据，通过人工智能来解读这些数据，使其能预测一个人的信用和风险水平，从而服务金融机构的风险控制需求。

如何通过人工智能深度分析大数据并在金融场景下使用，是一批想进入金融领域的互联网公司面临的问题。在阿里和腾讯都投身金融业务之后，百度在2015年才姗姗来迟，正式开发金融业务。当时百度的金融业务并不被业界所看好，一个很重要的原因是，虽然百度有很多关于互联网用户的行为数据，却不像消费数据那样与金融有强相关性，怎样通过这些看似与金融无关的数据去开展金融业务，是百度进入金融行业时亟须解决的一大难题。百度想到的是，通过人工智能将二者联系起来。

2016年初，百度高层找到我，希望我帮助百度在信贷场景用好人工智能。也就是说，通过人工智能去精准预测用户在金融上的表现。所谓金融上的表现，通俗来讲，就是一个用户借了钱后有多大的可能性会逾期不还。如果用户还钱概率很高，那么这个信贷业务就能够赚钱。但是如果逾期不还的人数较多，那这个业务就肯定要赔。我需要帮助百度解决的问题，就是用人工智能预测用户借款后会不会按时还钱。

举一个具体例子，如果某位小企业主的公司出现了经营问题，他想找金融机构申请一笔个人贷款，那这笔贷款对金融机构来说风险很高，因为他很可能逾期不还。通过借款人的行为数据可以分析得知，贷款是大概率用作个人的流转，还是要用作企业经营。这是两类不同类型的借款，借款人如果打着个人借款的幌子去填企业的坑，对金融机构来说风险相当高。在这种场景下，人工智能通过对大量行为数据和金融表现数据的分析，可以得出相应的结论，因而对这些人来说，如果他们来申请个人贷款，金融机构就应该拒贷，不把钱借给他。此外，还有很多在金融上有重要意义的数据，例如，借款人是一个日常生活稳定的人，还是一个居无定所的人。通过人工智能的深度分析，

借款人的不同行为在金融上都对应了不同的逾期和风险概率，这些概率成为金融决策规则的一部分，帮助金融机构更安全地提供信贷服务，同时扩大放贷规模。虽然人工智能做出的预测不可能百分之百地准确，但是我们已经可以实现非常高的准确率，这对金融业务的开展仍然有相当大的帮助。

当时我估算过，风险控制上的预测准确率每提升1个点，带来的增额收入接近1亿元。基于这样一个人工智能风险管理体系，在控制风险的情况下，把流量逐渐打开用于金融业务，度小满金融（原百度金融）与金融机构合作的资产规模不断增长，在一年左右的时间里累计发放贷款超过3 800亿元，为50多家银行合作伙伴创造了近100亿的利息收入，真正让人工智能赚钱，且风险可控。

传统金融机构的人工智能转型

人工智能的一个重要意义在于它能帮助各行各业转型升级。例如对于银行而言，如何能够低成本精准获客是一大痛点。截至2018年末，我国商业银行机构数量已经达到了4 000家以上，其中国有大型商业银行5家、邮储银行1家、股份制商业银行12家、城市商业银行134家、住房储蓄银行1家、民营银行17家、农村商业银行1 262家、农村合作银行33家、农村信用社965家、村镇银行1 562家、外资法人银行39家，银行间的同业竞争十分激烈。而互联网金融公司的兴起无疑更增加了银行拓展客户与业务增长的难度。互联网金融平台以效率高、体验好吸引客户，银行如果止步不前，不注重提高用户体验，由此造成的客户流失可能会越来越明显。

另一方面，随着商业银行不良贷款率的居高不下，信用卡业务、贷款业务的风险控制也成为银行关注重点。根据银保监会发布的银行业主要监管指标数据，2019年一季度末，商业银行不良贷款余额2.16万亿元，较上年末增加957亿元，不良贷款率为1.80%。随着银行业务

扩张、客户下沉，客户的金融属性越发稀疏，贷款业务的风险控制也成为痛点。

2018年，我参与领导了百度与中国农业银行的“金融大脑”项目，在客户画像、精准营销、信用评价、风险监控、智能投顾和智能客服等方面为农业银行提供人工智能服务，从营销、风控到运营、决策，全领域为农业银行的智能化转型赋能。

金融大脑平台是农业银行的核心人工智能系统，以它为核心的人工智能金融平台解决方案，为农业银行打造了全领域的人工智能能力体系，这在业内是首举。金融大脑平台包括感知引擎和思维引擎两大服务引擎，实现统一的智能服务，在大范围多场景下集成了人工智能核心技术。感知引擎和思维引擎就像人的左脑和右脑，让计算机程序像人一样不但能看、听、说，还能针对外部环境的变化做出最优的决策，帮助农业银行更好地开展金融业务。其中，感知引擎可以进行人脸识别、活体识别、OCR（光学字符识别）、语音识别、自然语言处理、语音合成，在保证银行业务安全可靠的同时，不断提升用户的体验。思维引擎则基于机器学习，在营销和风控的核心业务领域，帮助农业银行在快速变化的市场环境中高效、准确地做出决策，提升业务效果。

作为农业银行人工智能的核心系统，金融大脑为智能应用提供人工智能基础能力的支撑。在智能营销领域，思维引擎构建精准营销模型库，实现多渠道协同的智能化系统营销，助力信用卡年新增发卡连续突破1 000万张。智能推荐营销响应率达到自然增长的4倍有余，单次营销活动可为行内节约营销费用1 000万元以上。举个例子，对于银行机构来说，一个通常的广告推送渠道就是发短信。整天收到来自不同银行的各种各样的短信，用户会十分反感，因此对于银行而言，短信一定要发得精准。短信发得不精准会导致两个结果，一是用户被骚扰了，可能会退订短信，或者对银行产生负面的认知；二是发送短信也有成本，虽然单条短信价格很低，但是随着量级的上升，成本也会

变得可观。在中国农业银行，我们通过对大量过往用户行为习惯的分析，建立了精准的营销响应模型。这个模型可以帮助我们发现哪些用户看到营销短信之后，会产生积极的行为和反馈。这些产生响应的用户才是营销活动应该选取的目标用户。这样基于人工智能的营销让短信只发送给高意图人群，不会打扰其他用户，降低了营销成本。

在智能风控领域，农业银行的信用分模型和反欺诈模型提供的服务，实现了实时、精准的智能化风控，有效防控了风险。其中，信用分模型可以预测识别出风险客户，比如在自助申请和自动审批的网贷产品“网捷贷”准入场景中识别出92%的不良客户。反欺诈模型在日均7 000万笔的掌银（掌上银行）交易过程中，可提供毫秒级响应的实时监控，每天平均预警交易可达1万笔，有效地保障了交易安全。

以上讲述了4个我亲身参与的行业巨头的案例，其中既有比较传统的银行业，通过人工智能实现成功转型；也有互联网科技业，通过人工智能大规模增加效益。无论在传统行业还是在新兴产业，数据和人工智能都产出了史无前例的商业价值。当然，从数据到价值的跃变涉及方方面面的许多环节，需要相互配合、相互衔接，无论哪个环节出现问题不能有效解决，都会阻碍最终价值的产生，不少人工智能项目也因此以失败告终。我在多年的行业实践过程中，总结出一套行之有效的人工智能方法论——AI思维——来解决这些问题。这几个案例只是一个铺垫，算是了解AI思维的热身环节，后面我们会从概念到实操等各个角度，以及金融、零售等多个行业，具体详细地阐释根据多年行业经验凝练而成的AI思维。

寻找你的“阿尔法”

世界的变化越来越快，10年前的许多知识已经被贴上了“落后”的标签；10年前毫不费力空手套白狼的赚钱思维，现在却不再行之有效。我们总在抱怨这个时代的趋势变幻莫测，难以把握，却没有人静下心来寻找问题的根源。其实，问题的根源正在于我们的认知。人类的认知是有局限性的，不仅受到个人知识背景和社会经历的影响，更受制于自身的喜忌偏好。书本上的知识是权威的，但更新太慢，实际操作价值较低。他人的经验看起来颇有一番道理，但每个人的情况不同，别人的经验在你身上也许就失效了。世界变化这么快，我们怎样才能不停地刷新认知呢？

大多数人可能会选择去上编程课或者英语补习班，或者在知识付费的时代中多学几门网课，多加几个“知识星球”，以达到“刷新认知”的目的。但是问题在于，如此获得的知识，大多是碎片化的。也许会有人去读MBA（工商管理硕士），但动辄百万的学费让许多人望而生畏。很多方式是在原本平庸的思维里机械地填充信息，并不会使我们的认知产生质的变化。

要从质的层面去做认知提升，不仅要提升认知能力，还应该转化思维模式。而我们要谈的AI思维，是一种时代前沿的思维模式，也许能够为你提供一种新的路径。AI思维可以在一定程度上解决我们的认知局限性的问题，帮助我们进行认知升级。它不同于目前很多人对人工智能的认知，他们眼中的人工智能等同于机器人、无人驾驶、人脸识别或者语音识别等，但这只是人工智能的具体应用，只是人工智能的冰山一角。我们要阐述的是AI思维，可以说是人工智能的真正精髓，是人工智能冰山的内核，它能够帮助我们找到自身真正的优势，获得真正展现我们个体主动策略优势的超额收益，从而使我们的生活和事业蒸蒸日上。

赫布理论：人脑学习的机制

AI思维与人脑思维有些相似，为了让你更好地理解，首先，我们需要知道人脑的学习机制。在各类经验学习的过程中，无论是视觉、听觉，还是更复杂的决策，人脑都有着它独特的思维机制。

为了研究总结大脑的思维机制，加拿大生理心理学家唐纳德·赫布在1949年提出了赫布理论。赫布理论表明，人脑中有许多神经元，前一个神经元通过突触——一个连接不同神经元的结构——持续向后一个神经元产生刺激。在这样的情况下，两个神经元之间的传递效能增加，形成细胞回路；如果这种刺激持续重复，突触传递的效能不断增加，人脑就记住了这两个事物之间存在的联系。比方说，我们在上学的时候，打铃就代表要上课或者下课，也就是说当铃声响起时，一个神经元被激发，而同一时间出现的上课或下课的场景会激发附近的一个神经元，经过多次这样的刺激之后，它们之间的联系会被默认下来，这就是人脑的学习机制。这种学习模式又称作“赫布型学习”，人无时无刻不在接触新的事物，建立事物之间的联系和规律，人的大脑也是无时无刻不在进行着赫布型学习。

宏观上，人们从经验中受到刺激时，人脑就学习到了事物之间的相关性，从而总结出相关的规律。如图1-1所示，经验改变了人脑的回路，通过习得的规律，我们可以快速分类新的问题，从而形成判断乃至做出决策。所以，人类对新问题的判断，来自过往的经验。

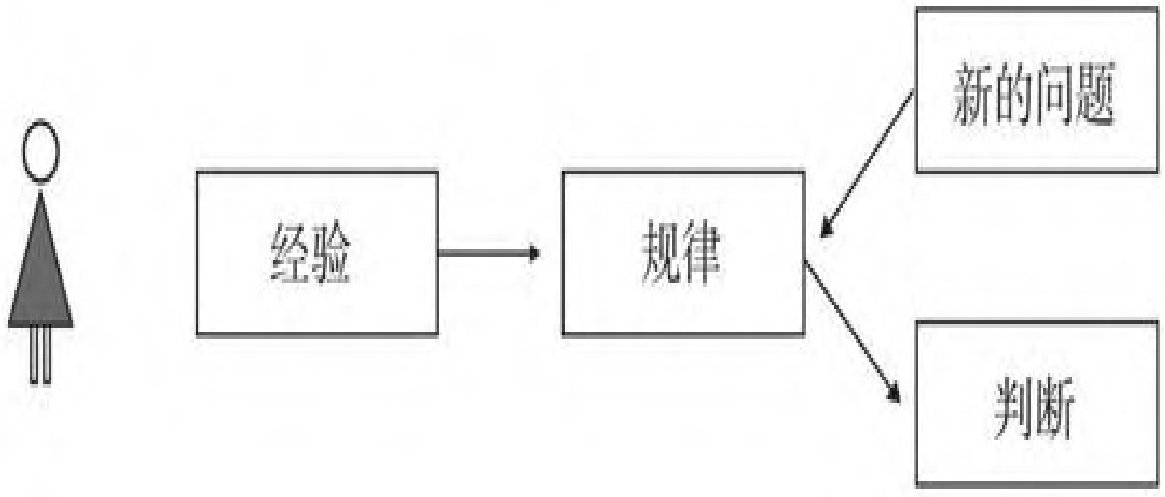


图1-1 人类习得规律示意图

有人估计过，人脑通过神经元的激活来阅读理解信息的速度是每天1MB（兆字节）左右。然而，信息的产生速度至少是每天2.5EB[2]。所以，人脑通过神经元的激活来处理信息的速度，远远跟不上新信息产生的速度。换言之，我们的认知是有限的，如果你还是以每天1MB的速度来处理信息、解读这个世界，那么注定跟不上这个不断更新的时代步伐。

羊群效应

人脑不但处理信息的速度非常有限，而且常常不经思考便追随大众的想法。经济学里有一个专门描述经济个体从众跟风心理的名词——羊群效应，将人类与羊群遇事的反应相类比。羊虽然看起来也是成群结队，但是十分散乱，最常看到的就是它们毫无目的地左冲右撞。然而在羊群之中，一旦有一只羊先行，其他的羊也会不假思索地跟随，也不管前面是否有狼或者身旁是否有更好的草。反观人类自身，羊群效应所导致的盲从行为往往会让人陷入骗局或遭到失败。每个人的认知都有局限性，做出的判断有可能是不准确的。那么，在羊群效应下的人盲目相信别人的认知，跟随别人的做法，做出来的判断不准确的可能性更大，也不能为其带来收益或者规避风险。

人类的群居性使得与他人的交互成为人类生存的前提。因此，人的认知必然受他人的影响。相应地，人的大脑在处理信息时，会或多或少地受到外部环境的影响。这种影响有时是有意识的，有时是无意识的，往往使得人的认知缺乏客观性，甚至不由自主地从众，得出与周围多数人一致的结论。此外，人的认知还受到自身能力的限制。这种能力一方面体现于人的记忆力，也就是说，成为我们认知的信息基

础能有多少；另一方面则体现于已有的技能和工具，以及它们在多大程度上能为我们的认知过程提供服务。

正是因为人类的认知存在这些无法避免的问题，思维方法才显得愈加重要。我们的思维决定了我们的认知过程，决定了我们如何从历史经验中形成对当前事物的看法和对未来的预判。我们每天都面临着许多决策：在哪安家落户，和谁共度余生，如何出类拔萃……小的决策也许会影响你一天的心情，大的决策则会影响几年甚至是一生的生活质量。一种好的思维能最大限度地把我们的认知能力从束缚中解放出来，帮助我们做出最优的决策，而AI思维正是这样一种能够解放我们认知能力的强大思维。

AI思维

如前所述，过往的经验确实能帮助我们做决策，尤其当我们不得不基于模糊的信息进行判断时，经验能发挥巨大的作用。然而，在信息爆炸、算法决策高度发达的今天，光靠以往经验做判断、做决策的思维模式已经显现出弊端，掌握着大量数据和科学决策工具的人，早已成为你前面的领跑者。落后于时代的思维，带来的是慢于时代发展的速度，其产生的收益必定比他人低，而且会越来越低。

本质上，人工智能模拟的是人脑，但又不同于人脑。AI思维是从数据产生模型，如果遇到新的输入，人工智能就能通过模型做出准确的预测。AI思维与人脑思维的相似点在于，AI思维也是通过对过往的分析习得规律，得出结论，但是AI思维分析过往数据的过程是不同的。

那么数据是什么？数据是人类经验的数字化形式。如图1-2所示，就像人从经验中获得规律一样，人工智能能够根据数据形成模型。一旦遇到新场景的输入，模型就能做出判断、产生预测。人工智能的这种思维能够充分利用数据，尽量减少主观臆断，以我们想达到的目标

为中心，挖掘规律，形成模型，运用模型，产生预测。人工智能就好比一个可预知未来的水晶球，掌握了AI思维，你就可以做出科学意义上最佳的决策，产出前所未有的收益。

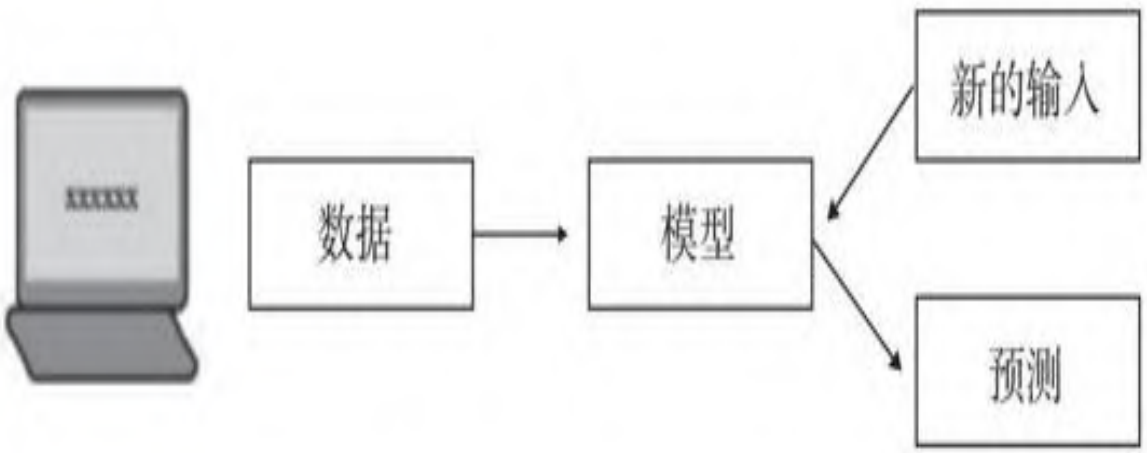


图1-2 人工智能习得规律示意图

就像过去的互联网思维一样，在互联网出现之前，一个人只能与周围的几十个人产生联系，而有了互联网之后，与世界各地任何人、无数人产生联系成为可能。而人工智能出现之前，人只能借助过往经验或者周遭环境等有限的信息做出判断；而借助人工智能，人们能够发现和挖掘无数的相关信息和数据来做预测和决策，而做出最优决策也是AI思维的一大特点。从广义上来看，人工智能具有感知、认知和决策的能力。决策就是通过上述模式而产生的，比如在股市中，人工智能可以根据历史数据生成模型，然后来指导股市决策。人工智能的感知和认知功能包括感知图像、理解语言、识别语义信息等，它的逻辑和股市决策是一样的，也是一种预测，只是预测的目标不一样。例如，人工智能在进行人脸识别时，也是根据许多人脸数据生成模型，然后遇到一个新的人脸数据，这个模型就能判断出它是否是特定的那个人脸。

在商业世界里有一个名词阿尔法（alpha）收益。经常买基金的朋友应该都知道，阿尔法收益是相对于贝塔（beta）收益而言的，贝塔收益是一种基准收益，而阿尔法收益是在贝塔收益期望之上的超额收益，是积极操作所带来的投资回报。就好比我们理财一样，央行的基准活期年利率是0.35%，就是说我们有1万元放在银行存活期，一年可以赚到35元，而央行年利率提高0.1%，我们就多赚10元钱，这个就类似于贝塔收益。但是我们用这1万元钱去买余额宝、理财通、股票，投资房产，可能收益翻倍，这个多出来的收益，就是我们积极操作所带来的阿尔法收益。1%的阿尔法就意味着比市场带来的投资回报多了1%。毋庸置疑，无论在哪个市场，你都希望能跑赢它，都希望能找到属于自己的正向阿尔法。AI思维的作用就在于它能带着你的认知更上一层楼，赋予你跑赢大盘、产生高阿尔法的能力。

相较于人脑1MB的处理速度，人工智能可以用任意高的算力为你工作，而且不分昼夜，不知疲劳。人工智能的能力和速度，足以帮助我们寻找到梦寐以求的阿尔法。有了人工智能，你还担心跟不上时代变迁的步伐吗？

股市：智能投顾发现阿尔法

传统上，一个金融分析师配一个程序员，就是量化交易在金融公司的常规形式。分析师出策略，程序员把策略转写为代码，输入到交易程序里。但问题在于，策略是分析师自己设计的，它实际上还是一种主观策略。如果在量化交易中投入人工智能，就能使机器自己通过历史数据生成策略模型，进而预测将来的走势，从而寻找金融市场的阿尔法。现在很多的智能投顾，就是通过这样的模式来运作的。

那么让我们来看一下，人工智能是如何在投资顾问领域创造出价值的。英仕曼集团是全球最大的对冲基金公司之一，它从2014年就开始使用智能投顾进行自动交易决策，并且获得了巨大收益。2015年8

月，美元暴跌，全球股市暴跌，引发大量抛售。但英仕曼的智能投顾在此之前就洞察了市场，并迅速从下跌中获利。而在美国股市下跌约3%后，该系统又发现了买入的机会，重新回到市场。人工智能参与决策，为英仕曼集团规模最大的一只基金贡献了大约一半的利润。通过AI思维，英仕曼集团在动荡的股票市场中跑赢了其他公司，产出了大量的阿尔法。

那为什么英仕曼集团依靠人工智能就能够在惨烈的股市中杀出重围并大量获利呢？具体来说，传统的金融投资还是要靠策略分析师这样的专业人士来做判断，但是由于人固有的认知局限性和经验的限制性，在面对一些大的变动时，也会出现不知所措的情况，这时做出的判断往往有一定的主观性。但人工智能具有强大的计算能力和信息分析处理能力，能够快速准确地分析金融资产在历史上的表现数据和各种外部环境等各种因素，例如宏观经济、就业情况、社媒舆情等，然后根据分析结果预测金融资产的未来表现，从而避免金融分析师决策的主观性和局限性。

房市：美国房产网站Zillow股价两年涨七倍的秘密

杜甫诗云：“安得广厦千万间，大庇天下寒士俱欢颜。”中国人自古以来就讲究安土重迁，对房子钟情千年。不论经济发展如何波动，房价不但居高不下，还一路水涨船高。不论别的行业发展如何不景气，人们对于买房的热情毫不衰减。房子对于人们的重要性不言而喻，房子是人生的一笔重要投资这件事也毋庸置疑。所以，选房和买房也要寻找房屋市场的阿尔法。

在买卖双方的房地产交易市场上，什么东西是最重要的？毫无疑问是全面而准确的房产信息。在传统的房地产中介业中，由于房主不清楚自己房屋的市场价值，买主也不是很了解理想房屋的实际价值，房屋买卖市场的买家和卖家都一头雾水，十分茫然。中介公司便利用

这种信息不对称从中抽取丰厚的佣金。这种情况持续了多年，直到2006年，一家房产网站Zillow横空出世，在美国一上线便引发了轰动。因为它提供的房屋价值评估业务，解决了房屋买卖市场这种信息不对称的问题，并给予客户充分的知情权，这在一定程度上颠覆了传统的中介业。

Zillow的房屋价值评估业务是凭借其自行研发的人工智能系统建立起来的，其具体运作过程如下：利用Zillow上的房屋估值系统Zestimate，可以了解到房屋的地理位置、建筑面积、卧室数量、卫生间数量、房屋照片、周边环境等多维度的数据，再结合类似房屋的价值范围、历史价格、价格折扣、价值波动等数据，通过人工智能模型，计算出房屋目前的市值以及未来的升值空间。对于房主来说，他们可以清楚自己房屋的估值，对于自己的房屋有一个合理的定价和预期，而对于买家来说，他们也能够以合适的价格买到称心如意的房屋，形成共赢的局面。

目前，Zillow公司已经运用这些模型为超过1亿个美国房屋提供了房屋估值。《华尔街日报》曾将Zillow的估价和美国1 000个房屋的实际售价进行比较，结果显示，两者的中值偏差为7.8%，接近于Zillow自己估算的7.2%。这就是AI思维的一个实例，从中我们也能看到人工智能的优势，不仅可以处理大量复杂、凌乱、看似没有规律的数据，也可以避免一些人为的主观因素，做出客观的最佳决策。

婚事：约会和恋爱的科学决策

柏拉图说过，每个人都被上帝分成了两半，被分开的人太多了，我们很难找到正确的另一半，但是我们仍然想找寻到称心如意的人生伴侣。有人相信缘分，但如果每天宅在家里，社交圈不大，可挑选的范围也必然受限。随着互联网的发展，人们的交流渠道拓宽了，又出现了新的交友方式——在线约会。新事物的产生又伴随着新的问题。

互联网时代在线约会的核心问题在于可挑选的对象太多，用户常常不知道谁是对的人，不知道到底该从何开始。这时，为了帮助用户更好地找到称心如意的对象，提供智能匹配就显得尤其重要。

以前提供这种匹配服务的是传统的媒婆，但是媒婆手里的单身人士数量是有限的，也就是说，她的用户数量很少，所以她可以对这些用户的信息了如指掌，可以轻松地对这些用户的信息进行评估，然后给每个人提供一个最佳匹配。然而，当我们处理数百万的用户信息时，在一个超大的候选库里选择最佳的约会对象时，人工的判断就显得不现实，而且也不那么准确了。因此，在数据大爆发的现在进行婚配，我们十分需要借助AI思维来为用户找到最有可能的匹配。

美国的婚恋网站eHarmony开发了一种基于人工智能的智能匹配系统来解决这些问题。那智能匹配是什么呢？简单来说，就是通过掌握的约会人的信息来评估不同人之间的约会相容性，从而建立起相应的匹配。你可能会对“约会相容性”这个词有疑问，约会相容性指的是约会双方的相似度以及二者相互接纳的满意程度。建立在约会相容性基础上的匹配约会，有最大的可能性获得成功，并获得长期的满意度。

那这个智能匹配系统是如何运行的呢？首先，根据相容性分级来减少潜在匹配池。相容性分级由用户提交的心理自测结果以及年龄、所在地等构成，若用户之间出现年龄意向不符合或者所在地相隔太远等情况，就会从彼此的匹配池中删除，为以后的匹配削减一些不必要的信息。其次，基于文本、图像和其他维度的统计数据等，通过人工智能生成潜在人选之间的相容性。最后，相容性低的人就会被筛选掉，相容性高的人就会被匹配到一起，这样就可以为每位用户找到最佳的约会人选。婚恋行业大量的数据表明，即使在婚恋这个看似非常主观的领域，人工智能也能为它服务的人群带来阿尔法。

也许你以为自己不可能摆脱“羊群”，不敢与众不同；也许你以为自己只能成为一波又一波的“韭菜”，面对股市一片红海畏缩不

前；也许你以为自己打拼半生，也买不到一套合适的心爱的房子；也许你以为你终其一生也找不到属于自己的合适伴侣，只能随随便便找个人将就。但看完这本书你会知道，这些都是你的“以为”，都是你认知的局限性。或许，在五年前、十年前，这些“你以为”是真的，但现在，有了人工智能，有了AI思维，我们就能突破认知的天花板，跟上时代的步伐，找到属于自己的阿尔法。

AI思维的核心价值

文字出现以前，人们结绳记事。这种不够清晰的记载方式使人们对宇宙和世界的认知一直停滞不前。后来，人们发明了文字，世界开始清亮，宇宙不再混沌，人类的思维逐渐变得清晰。从这层意义上说，人工智能和文字一样，都是承载和提升思维的工具。文字推动原始社会走向文明，AI思维赋予现代社会更多发展的可能性。

基于人工智能的个性化学习方案让你在短时间内掌握更多需要的知识，基于人工智能的智能匹配社交帮你更容易找到灵魂伴侣，基于人工智能的房屋评价系统让买房不再是糊涂的选择，基于人工智能的战略规划让国家的发展充满了更多的可能性。可以说，AI思维是这样一种“修身齐家治国平天下”的思维方法。要想掌握AI思维，需要了解的事情有很多。我们不妨先从AI思维的要素开始，开启这段新的旅程。

AI思维的要素

人工智能早在1956年美国的达特茅斯会议上就被提出，但直到2016年，才开始被大众熟知，近几年才开始被普遍运用在社会生活中

的各行各业。这是为什么呢？原因在于，以前很多人工智能所必需的要素还没有发展起来。随着互联网的普及，人们对于人工智能的研究更加深入，这几年这些要素有了突飞猛进的发展。那么，AI思维得以实现，需要依赖于哪些要素呢？

第一个要素就是大数据。

我们处在一个数据井喷的年代，无论是企业业务环节产生的数据，还是个人移动设备里面的数据，信息量都非常大，并且这个数据量在近几年呈爆炸式增长。如图1-3所示，根据智研咨询发布的《2017—2023年中国大数据应用行业市场全景调查及未来前景预测研究报告》，全球数据总量的年增长率将维持在50%左右；并预测，到2020年，全球的数据总量将达到40ZB（1ZB=1万亿GB）。我们想想20年以前，个人计算机里面的数据可能只有区区几十兆，而现在的数据都是以TB或者PB作为单位的数量级。我们知道AI思维建立在对数据学习的基础上，数据越多，人工智能学会的东西越多，人工智能做出的判断决策就越准确，越具有实用性。所以说，大数据的形成是AI思维的一个要素。



图1-3 全球数据总量统计图

第二个要素是模型。

我们的大脑由许多神经网络连接而成，每天能帮助我们做各种各样的决策。早期的人工智能模型包括人工神经网络，它本质上是对大

脑，也就是我们自有神经网络的模拟。这种模型在20世纪80年代就已经被发明出来了，它能够帮助我们从数据中提炼出知识。而今我们在人工神经网络的基础上有了更深的研究，产生了一系列的深度学习模型，其实可以把深度学习模型理解为传统的神经网络模型的加强版。一般的人工智能模型随着数据量的增加，预测效果会提升，但很快达到瓶颈。而深度学习的预测效果随着数据量的增加，会持续提升。

深度学习强大的学习能力决定了它能够更好地模拟人类大脑的运算机制，它最擅长的是理解和识别图像、视频、声音、文本这样的数据。通过对图像进行深度学习，它实现了人脸识别，比如支付宝刷脸支付；通过对声音的深度学习，实现了音频自动转换为文字，比如讯飞听见能够一键录音转文字；通过对文本的深度学习，人工智能已经能够独立创作，比如微软的人工智能机器人小冰出版了原创诗集《阳光失了玻璃窗》。

深度学习不但能为人工智能提供技术支撑，其发展也为人工智能在各个领域的应用提供了无限的可能。ImageNet是计算机视觉领域很著名的一个图像识别竞赛，是该领域的“奥赛”。很多专家、学者都会参与，微软、谷歌、百度等科技公司也是该比赛中的佼佼者。在ImageNet竞赛中，随着深度学习的发展，图像识别能力有了很大的提升，从2010年到2015年这5年间，错误率从30%左右缩减到5%。5%的错误率已经基本上匹配了人对图像识别的错误率，这促进了人工智能在更为广阔的图像识别领域的应用。除了图像识别以外，语音识别的发展也依赖于深度学习。20世纪50年代，第一个语音识别系统Andry只能够识别10个英文数字；而现在，深度学习助力语音识别，应用于市场上的语音识别系统对普通话的识别准确率高达95%，百度、搜狗、讯飞的语音识别的错误率维持在3%，比人类识别语音5.7%的错误率更低，这为人工智能在语音识别的各个领域展开布局提供了强有力的支持。所以说，深度学习的不断发展也促进了人工智能应用的不断普及。

第三个要素是算力。

AI思维的实现需要很强的计算能力的支撑。我们想到计算能力，首先想到的可能是电脑、手机等智能设备，它们能提供一部分算力。对我们做人工智能来说，这样的算力是不够的。我们需要一个大规模的计算机集群，需要成百台上千台计算机连接在一起，进行大规模的运算。谷歌X实验室推出的谷歌大脑就是将16 000台计算机的处理器连接在一起，这样强大的算力使其具有强大的自主学习能力，成为全球最大的人工智能大脑之一。

除了计算机集群，算力还需要GPU的架构。GPU全称是Graphics Processing Unit，就是我们常说的图形处理器，是一种专门在电脑和移动设备上进行图像运算工作的微处理器。与之相对应的是CPU，即中央处理器，其功能主要是解释计算机指令以及处理计算机软件中的数据。GPU的架构有别于传统CPU的架构，能够很好地支持深度学习模型的运算。你可能对GPU这个名称很陌生，其实你可能早已接触过它。如果你是一名电子游戏爱好者，不管你玩王者荣耀还是和平精英，都需要GPU的支持来实现游戏中人物的移动及其他图像变换。GPU现在已经发展到了第四代，它提供的强大算力能够支撑人工智能的开展以及落地。

强大算力的加持，使人工智能有了更为广阔的发展空间。例如，2019年8月，人类在强大算力的支持下首次成功重建了果蝇大脑神经元的3D模型。果蝇一直以来都是生物学领域公认的研究动物，但为什么一直到2019年才建出第一个果蝇大脑神经元的3D模型呢？因为果蝇的大脑有10万多个神经元，要想建立一个完整的3D模型，需要强大的算力支持，而这个条件是先前并不具备的。2019年，在谷歌、剑桥大学以及霍华德·修斯医学研究所的合作下，研究人员将果蝇的大脑切割成7 042个40纳米的超薄切片，并用透射电子显微镜生成切片的图像。这些图像像素高达40万亿以上，数量共计2 100万张。数千块谷歌开发的深度学习加速芯片Cloud TPU（Tensor Processing Unit）为这些图片数据的处理提供了算力，最终才生成了人类历史上第一个果蝇大脑神经元的3D模型。

第四个要素是业务模式。

我们知道，人工智能要落地，必须在一个场景中去实践它。例如，在金融领域，传统上我们的银行机构只对征信良好的用户发放贷款，而征信记录不良的用户就得不到相应的贷款服务。单纯使用人力去判断一个用户的信用是件费时费力的事情，工作效率极其低下。但通过人工智能，我们能在很短的时间内对用户的信用状况进行分析和判断，决定是否向其提供贷款服务。这样不但减轻了银行工作人员的工作负担，还提高了工作效率。基于人工智能迅速准确且不知疲倦的特点，它还能为更多更大的用户人群提供贷款服务，扩大业务范围。基于这些人的大数据，通过人工智能模型和算力，还能对用户的信用状况进行分级，不但能够判断是否提供贷款，还能判断向这个用户提供多少金额最合适。人工智能为金融机构的工作提供了参考依据，这就是一种基于人工智能的业务模式。实际上，这样的商业模式，把传统的银行和金融机构的业务模式推广到了更大的用户群，帮助我们实现了普惠金融的理想。当然，除此之外，人工智能还能运用在更多的领域，还有更多的创新业务模式。正是因为有了这些创新业务模式，人工智能才能顺利地在各行各业落地，帮助企业产出价值。

如前所述，AI思维的基础在于数据，而核心在于模型，实现在于算力，应用在于业务模式。只有大数据、模型、算力、业务模式这四个要素同时存在，并行发展，人工智能的价值才能得以体现，AI思维才算完整。这是从具体要素角度对AI思维的阐述，下面我们从认识论的角度展开对AI思维价值的理解。

DIKW金字塔

认识论是关于知识来源和知识判断的理论，认识论将人们的认识过程定义为DIKW金字塔，认为认识分为四个层次，如图1-4所示，最底下一层是数据（data），数据上面是信息（information），信息上面

是知识（knowledge），知识再上面是智慧（wisdom）。DIKW金字塔的名字就是取了每一层的首字母组合而成的。在DIKW金字塔中，越靠下的层级包含的内容就越多，每往上一层，内容就比下一层更加凝练，更富有价值。原始数据虽然数量庞大，但有用的信息全都淹没其中。信息是由数据加工处理而成，信息在实践中被升华提炼产生了知识，而智慧又是对知识的凝聚总结，它是金字塔的塔尖，能够长久地影响人类。



图1-4 DIKW金字塔示意图

需要强调的是，所谓信息，指的是根据业务需求处理好的有特定意义的数据。从数据到信息的转换是传统信息处理中比较专注的领域，通过数据库系统和数据处理软件，把相对粗糙的数据转化为信息。而从数据和信息到知识，是AI思维所涉及的领域，这个转化过程与数据库运算不同，因为它不是机械的数据规整和变换，而是通过人工智能完成的，是认识层面一个巨大的跃迁。

AI思维的关键在于生成知识。你可能会问，人工智能为什么产生的是“知识”呢？首先我们要知道，知识就是人类对物质世界以及精神世界探索的结果总和，也就是对规律的总结。我们在前面讲过，AI思维中的模型可以类比人脑思维中的规律，也就是说，模型体现出来的就是知识本身。随着数据源源不断地更新，模型体现出的知识也会不断刷新，针对新的输入产生预测，为我们的决策提供支持。

著名统计学家拉瑟福德·罗杰斯（Rutherford Rogers）曾说过：“我们被淹没在信息中，急切需要知识。”互联网的发展使我们突破了时间和空间的限制，能够获得来自世界各地的信息。但在信息爆炸的今天，我们对信息的获取是呈平方数增长的，在获取大量信息的同时，我们无法辨别哪些信息是对的，哪些信息是有用的，产生了深深的“知识焦虑”。而现在人工智能做的就是把大量无序的信息变成有用的知识，缓解人们的“知识焦虑”。

AI思维带来认知化革命

在当代，数据是一种重要资产。数据反映了事物的原理和规律，当你找到它的规律后，就可以去预测未来的事情。如果将数据比作原油，那么人工智能就是从原油中提炼各种高价值产品的加工厂，它的重要性可见一斑。从数据中发现知识、洞察和规律，这本身不是一个新的概念，几百年前就有过这样的实践——譬如说，开普勒从几百页关于天体位置的数据中，提炼并总结出了天体运动的三定律，至今仍

被使用。现在，在AI思维的帮助下，我们借助大规模计算的方法，从海量的数据中自动地学习知识和规律。那么，从实际应用的角度出发，作为一个数据驱动的决策框架，AI思维都带来了哪些价值？

首先，数据驱动的人工智能框架可以带来个性化的体验。人工智能可以根据用户的历史浏览记录、成交记录对用户的喜好建立模型，得出各商品或内容和用户喜好的相近程度，并把相近程度排行最高的商品或内容推荐给用户。例如，当我们进入一些购物网站，可能会发现许多这样个性化的体验，若你之前购买过衣服，它可能会给你推荐其他的搭配商品，若你之前购买过图书，它可能会给你推送同类型书单。数据驱动下的人工智能框架带来的个性化体验让网站不再千篇一律，网站给出的每一条推荐都是根据用户需求调整和优化后得出的，真正实现了千人千面的精准体验。对于用户来说，这样省去了他们检索的时间，还更加符合他们的需求，带来了更好的用户体验，对于网站来说，这样可以提高网站的浏览量、点击率，还能提高商品的销量。此外，人工智能带来的这种个性化体验也能为我们的生活提供更多的便利，带来更多的幸福感，例如现在的智能家居系统，可以提供自动控温等贴心服务，提升了家居的安全性、便利性、舒适性，并且环保节能，让你感觉到这就是最适合你的，为你量身定做的家。

其次，数据驱动的人工智能框架可以带来细粒度的行业策略。行业策略细粒度意味着企业的经营会更加精细化。例如，企业可以把一个产品的目标客户群简单地划分为一定年龄范围的男性或女性，但这样的客户群划分显然没有针对性。利用数据驱动的人工智能框架进行目标客户群划分，得到的结果更加详细，比如我们不仅可以考虑基于年龄、性别这样的因素，还可以交叉考虑包含更多维度，例如兴趣爱好、行为习惯等的目标客户群，从而得到细粒度的营销策略。以视频软件芒果TV为例，为了提高视频的点击率，芒果TV运用人工智能来判断向用户推送视频的类型和内容。比如，追求放松娱乐的白领一族会收到《快乐大本营》的相关推送，喜爱烧脑解密的年轻人群会看到《明星大侦探》的相关广告。而随着用户观看视频数量的上升，人工

智能的推送方案也会更加个性化，细化到满足每一个用户的需求。与传统方式相比，人工智能提供的细粒度视频推送方案为芒果TV提高了30%的点击率，真正实现了精细化运营。

最后，数据驱动的人工智能框架可以带来知识和洞察。我们去学校或者从经验中可以学习到知识，而数据驱动的人工智能框架可以赋予我们持续高效地从数据中学习知识、挖掘洞察的能力。这些知识和洞察可能不是列在教科书上的条条框框，但却一定是从数据中实时地、最大体量同时也是最有效获取的，并能够运用于业务实践中。例如美国哈佛大学地球与行星科学系的布伦丹·米德（Brendan Meade）教授团队通过深度学习分析了来自世界各地的地震数据集，发现了余震发生的规律，在此基础上开发了一套能够预测余震的智能系统。这套系统为避免余震二次伤害，顺利进行灾后救援和恢复工作提供了很大的帮助。

AI思维不是捷径，但它却可以帮你更加快速地抓住事情的本质，找出用户的需求，洞察世事的发展，进行准确的预测，产出更大的价值。这是科学探索的目标动力，也是人类实践的追求所在。

著名的科技思想家凯文·凯利（Kevin Kelly）说人工智能是认知化。如果说电力化带来了人工的动力，那么认知化带来了人工的智能。大量的实践表明，在感知方面，包括视觉、听觉、语言理解等，人工智能可以接近人脑。前文已经提到过，现在的图像识别错误率已经控制在5%，基本和人脑识别错误率持平；在专业决策方面，在海量数据的支持下，人工智能甚至可以超越人脑。例如，在金融风控领域，通过人工智能输出模型的KS值（通常用来衡量风险识别有效性的一个指标），可以做到40%~50%甚至更高，有效地控制住风险。基于人工智能的快速信贷审批，效果可以超越传统的人工方法。在传统出版业，一个编辑编辑一本书至少需要一个月的时间，但全球最大的中文期刊网龙源期刊开发了一个人工智能编辑平台“知识树”，它是一个能够根据定义内容自动编辑图书的智能系统，可以在一天之内完成一

本图书的编辑工作，极大地提高了图书编辑工作的效率，缩短了图书出版的周期。

AI思维是一种“数据驱动决策”的思维，它不忿恿你去追逐虚无缥缈的白日梦，而是教你学会用数据筹码四两拨千斤。不论你研究的是什么领域，从事的是什么职业，通过AI思维，你都能从数据中理出头绪，更加快速、直接、准确地预测出研究对象的行为或者结果。绳锯木断，水滴石穿，掌握了以小博大的AI思维，你不但是一个梦想家，也能成为一个真正创造价值的实干家。

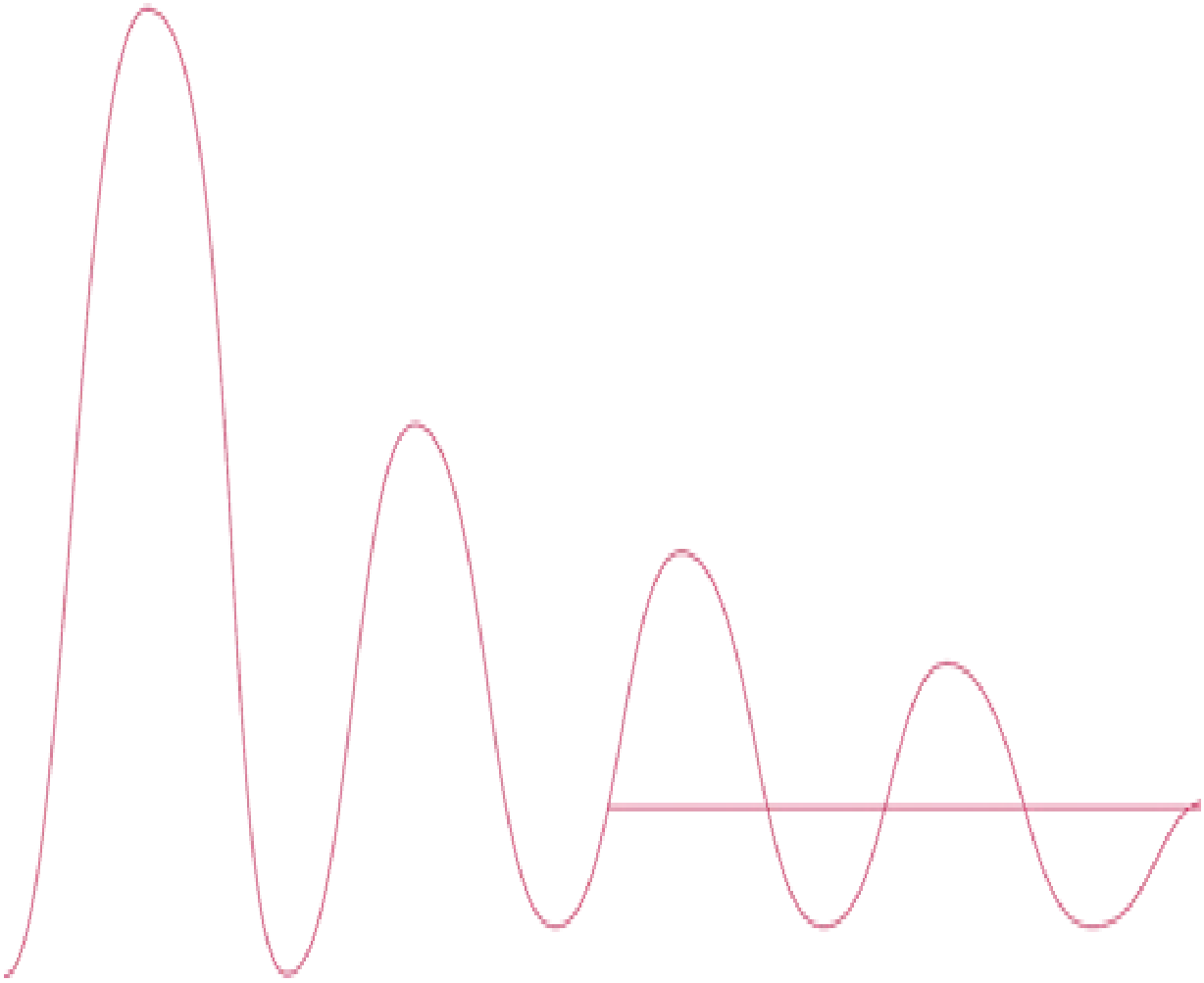
---

[1] BAT即百度、阿里巴巴、腾讯。

[2] 1EB=1 024PB=1 048 576TB=1 099 511 627 776MB。

第二章

AI思维的底层逻辑



当爱因斯坦说“上帝不掷骰子”的时候，他错了。鉴于黑洞给予我们的暗示，上帝不仅掷骰子，而且往往将骰子掷到我们看不见的地方以迷惑我们。

—《时间简史》—

本章旨在讲清楚AI思维中的“法”，即法则和规律，也就是从数学规律的角度，阐述AI思维的底层逻辑。AI思维，尤其是其核心——模型，是在数学和统计算法基础上发展起来的，必然受到数学和统计规律的制约。只有理解了这些制约，才能理解人工智能与生俱来的局限性，只有了解了这些局限性，我们才能扬长避短地使用AI思维，更好地享受人工智能给我们带来的实际价值。无论在日常生活还是在商业活动中，我们知道，从来就没有解决一切问题的“银弹”，任何事情都不能一蹴而就。社会对人工智能不切实际的预期，多半来自对其底层逻辑的一知半解。很多时候大家没有听说过，也无暇去理解人工智能实用性背后的法则。并且，在众多的商业包装下，人们越来越多地去追捧短平快的“成果”和高流量的媒体曝光，而不去关注支撑应用的基础性逻辑。这种“知其然而不知其所以然”的做法，不会带来长远的发展，所以，在深入讲述AI思维的“术”和“器”之前，我们必须先来理解它的底层逻辑，任何想运用AI思维解决实际问题的人都应该知晓并从中获得启发。

模型的泛化能力

首先，我们来回顾一下AI思维。AI思维能够帮助我们捕捉数据的本质特征，根据数据的特征形成模型。在新场景中针对新的输入信息，模型就能做出判断、产生预测。模型使用的一大类场景是分类，也就是预测新的输入所对应的类别。那么，什么是分类呢？如果我们手里有一张照片，你要做的是判断这种张照片里是否有动物出现，你通过眼睛获得了这张照片的视觉特征，这时，你的大脑已经下意识地做出了一个判断、一个分类。这个问题的回答结果就只能是“有”或者“没有”，这是一种分类，叫作二分类问题。

但是并不是所有的分类都可以用是与否这样一分为二的判断来解决，还有很多东西需要人们去分辨它属于多种类别中的哪一类。我们还是拿这张照片举例。当你已经判断出这张照片里有一只小动物后，你还需要判断这是一只什么动物。你会根据这只动物的体型、体毛、耳朵、眼睛、嘴巴等特征，判断它是猫、狗、兔子，还是其他什么动物。这也是一种分类，叫作多分类问题。

在日常生活中，我们每个人都避免不了和他人对话，我们要说，要表达，也要聆听、理解他人的话。其实我们识别对话中的声音信号，从中提取所说的字和词语，也是一个分类的过程。商务印书馆的《现代汉语词典》是目前较权威的大型现代汉语词典，第7版收录的词将近7万条，其中包括了字、词、短语、熟语、成语等。也就是说，在我们聆听的时候，为了理解这些声音信号的意义，我们的大脑都在将它们中的片段分类成这些字或词之一，这就是一个多分类的过程。但是有的时候，我们也会混淆一些读音比较相近的词，比如说“建议”和“坚毅”、“升起”和“盛气”，这说明人脑的分类也不是完美无瑕的。

我们的视觉要对看见的东西进行分类，我们的听觉要对听见的东西进行分类，我们的味觉要对尝到的东西进行分类……感官的分类可以帮助我们更加完整地认识这个世界。可见，对外界事物的分类能力是人们认知的重要组成部分，可以说，每时每刻我们都在对现实世界中的输入信号做出分类，分类的准确与否直接影响我们生活和工作中决策的质量和效果。

正如人一样，分类也是人工智能进行感知和认知的重要工具。在机场部署的人工智能模型可以识别出通过安检的人是否和证件上的照片一致，刷脸打卡的人工智能模型也在很多机构运用起来，无须戴工卡便可知道人的身份，许多手机的刷脸支付功能可以识别出你是否是该手机的主人来决定是否支付，极大地提高了资金的安全性。在语音识别领域，主流的人工智能识别模型对常见词语输入分类的准确率已

超过97%。除了感知图像、声音领域，人工智能模型的分类还可以用在更多的场景，做出人类都难以达到的准确预测。例如，在营销领域，人工智能模型需要预测客户的喜好度和倾向性等，相当于对未知客户行为的分类。在金融信贷领域，人工智能模型能对客户的风险进行评级，也就是对客户将来风险度的预测，从而将客户分为高风险和低风险客户，并辅以不同的策略。可以说，通过分类，许多常见的问题都可以在人工智能的分类框架下进行考虑。

分类模型的数学基础

向量指一个同时具有大小和方向，且满足平行四边形法则的几何对象，如图2-1所示。简单地说，向量是有序的数字列表。人工智能中用到的数据一般是以向量的形式表示的。按一定的次序排列在一起的变量（x_1，x_2…x_n），依次表示事物的各种特征。举个例子，我们打算用过去一年在服装、饮食、住房、出行、健身、娱乐、美容、奢侈品这8个类别的消费金额来研究一个人的消费习惯。那么，这8个数字构成了一个八维的向量：（x_1，x_2，x_3，x_4，x_5，x_6，x_7，x_8），其中每一个变量对应了一个类别的消费金额。不同类型的人，上述向量的数值也不同，美食爱好者的x_2有比较高的数值，高消费地区的租房客的x_3会比较高，注重外貌容颜的人x_7比较高……

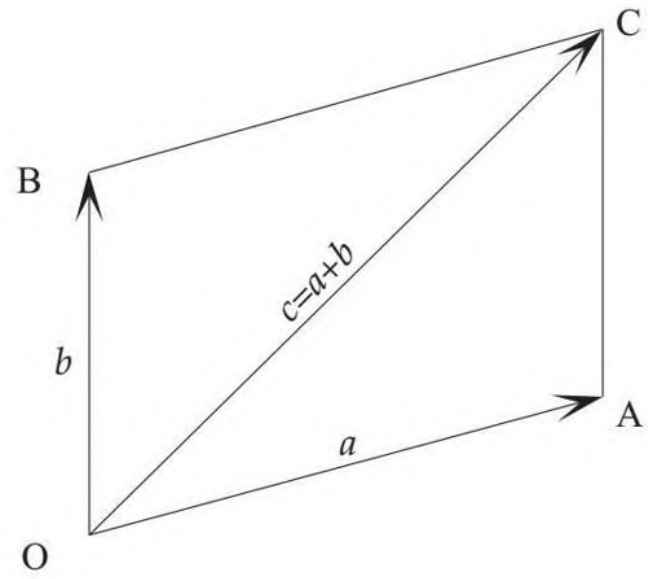


图2-1 向量示意图

这里面的每一个变量叫作向量的一个维度，人工智能中的向量也常被叫作特征向量，因为它描述了事物的特征。要注意的是，向量中变量是按照一定顺序排列的，顺序变了，向量也就变了。我们举例的向量只有八维，是一个简化的例子，在现实中，人工智能处理的向量通常是高维的，在商业应用中可以达到上百维、上千维甚至更多，在处理基因数据时，可以达到百万维以上。正因为人工智能可以理解高维度的向量，而不是像人脑一样只能分析少数几个变量，所以人工智能会给我们带来更深入、更有效的知识和洞察。

在数学上，分类要做的是，将一个预测函数作用在个体的特征向量上，得到想要的输出结果。如前所述，人工智能中个体的数据是以向量形式来表示的。以图片识别为例，为了识别图片中的物体种类，首先我们要建立一个预测函数f，预测函数f作用在图片的特征向量x上，得到一个类别$f(x)$，比如说“小猫”“小狗”或者“兔子”。当然，在现实中的图片识别应用中，人工智能可以识别出任意类别物体的图片。

这个看似简单的过程其实是非常强大和灵活的，基本上所有人工智能分类的过程都可以抽象成这样，其中预测的目标是跟具体业务相关、需要提前获知的信息。就像之前提到的亿客行案例一样，网站需要预测一个访客的价值高低。如果我们可以准确地识别一个访客是否具有高购买的可能性，并将其应用于电商领域，根据访客价值定制营销策略，就能取得优秀的阿尔法。

让我们来看看如何设计高价值用户预测的分类模型。首先，我们需要收集历史上具有高价值或者低价值客户的标记，以及各自行为的特征向量。这个分类问题的目标是将任何一个新的访客分为两类之一。所谓高价值、低价值就是看客户一定时间里在网站的购买额度，超过一定金额的定为高价值客户，反之则是低价值客户。这样，我们有了模型训练所需的数据集，可以根据这些数据训练的模型预测一个网站访客价值的高低，即使之前他没有在网站上消费过。既然这个模型的作用这么大，那么我们接下来就具体看一下，它是如何被训练出来的，又是如何进行预测的。

模型的训练和预测

在人工智能领域，机器学习是产生模型的途径和方法论。和我们的学习一样，分类问题的机器学习框架也分为训练和预测这两个最基本的环节。

训练环节中需要给定一个训练集，通过训练集中的标记样本｛（x_1，y_1），…，（x_N，y_N）｝训练出一个预测函数f，也就是我们的预测模型。标记样本是由第1个到第N个输入数据x以及与其相对应的输出数据y组成的。在预测环节中，我们要对在训练环节中形成的预测模型输入一个新样本x，使用f输出预测值$y=f（x）$。需要注意的是，在训练环节中，由于样本是给定的，所以样本的标记是已知的，但在预测环节中，使用的都是新数据，所以样本的标记是未知的。

任何分类模型都要经过这两个环节。在训练环节中，机器通过分析训练集学习模型，发现其中的规律，并以预测模型f的形式产出。在预测环节中，碰到新的样本，即使模型先前没有见过这个样本，仍然能够输出相应的分类。正如人的思维方法是从经验中学习规律，人工智能通过训练环节从数据中学习模型。训练环节使用了何种模型，以及模型训练的质量，决定了模型的有效性，其中最重要的指标是模型在新样本上的预测效果。

如果一个模型能够对新样本进行有效的预测，就可以在实际使用中产生效益。比如，一个模型成功地预测了某位用户购买某款商品的概率为80%，那么商家可以积极地向该用户推送自家的该商品，用户经常接收到自己喜欢的商品的信息，下单的概率也会有所上升。成功地预测出用户是否会购买，帮助商家提高销量，这样的模型才是有实际意义的。

泛化能力是模型的核心

一个模型能带来多大的价值，取决于模型预测的准确率有多高。而模型预测的准确率又依赖于模型的泛化能力（generalization ability）。也就是说，在一定程度上，模型的泛化能力决定了一个模型最终能迸发出多大的能量。那么泛化能力到底是什么呢？

泛化能力是用来描述模型对新样本的预测能力的。我们在日常生活中也称之为举一反三或学以致用的能力。机器学习的目的是学到隐含在数据背后的规律，对具有同一规律、训练集以外的数据，模型也能给出合适的预测。这就是泛化能力的表现。

泛化能力体现了模型智能水平的高低，如果一个模型只在训练数据上能准确地分类，作用是很小的，因为这充其量是模型“记住”了各种输入和对应的类别，在新的场景中没有办法做出准确预测。

在进一步讲泛化之前，我想先讲一个故事。我有一对情侣朋友，女生是一个中国姑娘，叫郭霓禾，男生是英国人，叫安德费汀。郭霓禾是一个牙科医生，她诊病的程序十分烦琐，一个简单的牙疼，她会给病人量体温、测视力、检查血压，有时甚至还要检查心电图，除此之外，她还会询问病人一些无关的信息，比如身高、体重以及最近吃过什么。她饱读各种医学类书籍，还研究过各种偏方，最后再结合她的临床经验，总结出一些独特的治病规律。如果一个新的病人的所有指标符合她总结出来的规律，她才能判断这个人的患病情况，如果情况不符合，她就无法做出判断。比如有一次一个病人牙疼，经过检查和询问，她发现这个病人体温37℃，视力0.8，身高1.7米，体重70千克，是在吃了西红柿之后智齿开始发炎。于是她就总结出一个体温37℃、视力0.8、身高1.7米、体重70千克的人会在吃了西红柿之后引发智齿发炎的规律。刚刚好，安德费汀也符合这几个条件，有一次安德费汀智齿也发炎了，但因为安德费汀没有吃西红柿，郭霓禾认为他不符合自己总结出的规律，诊断安德费汀的牙疼不是因为智齿发炎引起的。而安德费汀是一个股票投资分析师，对于自己的经验和能力十分自信，他认为凭借自己的经验，只需要看一下股票的走势，就可以判断一只股票到底是应该买入还是卖出。他根本不考虑政策、大盘环境、主力资金进出等变化因素，每天只看一下电脑屏幕上股票的涨跌状况就做决定。这样的判断显然十分不科学，导致了他的大部分积蓄都被套牢在股市。

郭霓禾和安德费汀两个人都各有特点。郭霓禾对待事情很认真，看事情很全面，但是由于她顾及的事情太多太细，有时可能会过度重视一些无关紧要的特质，总结出一些没有作用的规律，比如那些奇怪的指标与智齿发炎其实并没有关系，只是那个病人刚好在智齿发炎前是那样的身体状态，刚好吃了西红柿，这样过度关注无关因素反而妨碍了病因的评估。安德费汀则正好相反，他看事情太片面，没有综合考虑影响股市涨跌的因素，导致他的判断和股市的实际情况产生了很大的偏差，成为被迫割肉那一批人。

故事讲完了，我想告诉你，郭霓禾和安德费汀其实都是我虚构的朋友，他们的名字其实就是过拟合（overfi tting）与欠拟合（underfi tting）的谐音。过拟合和欠拟合是两类统计学现象，有这两种现象的模型，不具备良好的泛化能力。

在统计模型中，由于模型太过复杂而拟合了训练样本中的噪声，以至于输出的结果与真实值相差很大，就像郭霓禾诊断病人一样，这就是过拟合。欠拟合则刚好相反，是由于模型过于简单，以至于得到的模型难以拟合训练数据，就像安德费汀没有考虑到影响股市的其他因素，而导致判断错误。

我们总是希望在机器学习训练时，机器学习模型能在新样本上有很好的表现。过拟合时，模型过于复杂，把训练样本学得过分好了，就很可能把一些训练样本自身的特性当成所有潜在样本的共性。这样一来，模型的泛化性能就下降了。欠拟合时，模型又过于简单，无法很好地学到训练样本的一般性质，所以不论在训练数据还是预测数据中，表现都很差。

过拟合就是过多学习了一些不必要的数据特征。比如你的模型需要区分男人和女人，模型学习了训练样本中所有男人和女人的特征。如果训练样本中的女人刚好都穿了红色的衣服，过拟合的模型就可能会把穿红色衣服作为区分男人和女人的一个特征。我们都知道，穿什么颜色的衣服并不能作为区分人的一个特征，这时模型学习的这个特征就属于过度学习。过度学习不仅毫无作用，甚至还会对模型的预测结果产生误导，比如出现了一个新的预测样本，是一个穿红色衣服的男人，模型很有可能会根据他穿了红色衣服这个特征就把他识别为女人。欠拟合与过拟合正好相反，它对于训练样本的特征学习得不够充分，导致训练出来的模型不能很好地进行预测。而且，由于欠拟合的模型没能很好地捕捉到数据特征，它不但不能对新的样本数据进行预测，在训练集中的样本数据时表现也不是很好。比如你要参加一场考试，考试前做了许多模拟题，发现有一些知识点你还不明白，在模拟

题中你有部分题目做不出来，这就是没有很好地掌握训练样本的规律。当你要对新样本进行预测时，也就是当你去参加这场考试时，考试结果也不会很好。不论是过拟合还是欠拟合，都会使模型不能很好地输出预测结果，所以，在机器学习中，这两种现象都是需要极力避免的。

图2-2形象地展示了拟合的几种情况。我们有一些数据样本，大致呈二次函数形式，可以看出，用二次函数来做拟合最合适的是图2-2中间的图像。但是如果我们不采用二次函数呢？比如，我们用线性的函数来拟合它，我们可以得到如图2-2中左图的直线，这显然没有很好地拟合训练样本数据，更不用说预测数据了。我们用高次函数来拟合，可能会得到如图2-2右图那样的曲线，显然，这并不是我们想要的模型，它把个别数据的偶然偏离也当成了共性，而过度拟合了进去。

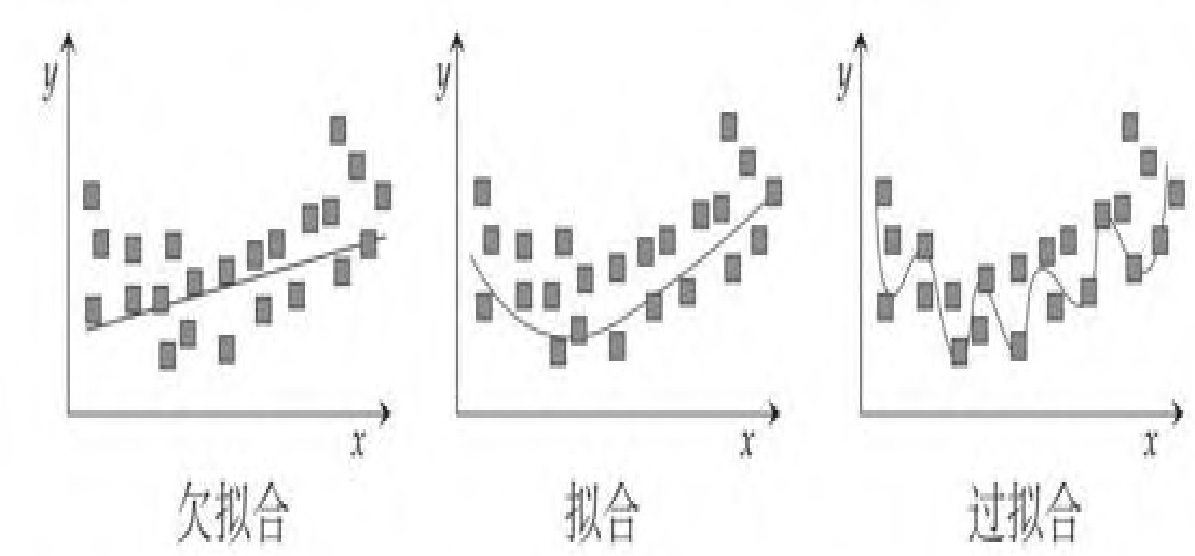


图2-2 函数拟合示意图

通常，解决欠拟合的方法包括：（1）增加新特征，也可以考虑加入特征各种形式的组合，来增大模型可操作的空间；（2）尝试非线性

模型，比如非线性支持向量机（SVM）、决策树、深度神经网络（DNN）等模型。

解决过拟合的方法包括：（1）交叉检验，即拿出大部分样本进行模型训练，留小部分样本用模型进行预测，并通过调节这小部分样本上的预测误差来得到较优的模型参数；（2）特征选择，从已有特征中选择部分特征以减少模型所用特征数；（3）正则化，即为原始模型引入额外信息来限制模型的复杂度；（4）增加训练数据在一定程度上可以避免过拟合。

真正的泛化能力需要回避过拟合和欠拟合，一个好的模型必须真正把握数据的底层规律，既不比数据本身复杂，也不比数据本身简单。人工智能行业的趋势是要用越来越复杂的模型来尝试做出更好的泛化效果，这需要我们更加谨慎地看待：如果训练数据量不够多，或者并不支持复杂的模型假设，那么，很有可能我们只是在过拟合，而不是在产出优质可泛化的模型。什么样的模型既不过拟合，也不欠拟合，是运用AI思维时需要考虑的核心问题。

方差和偏差的权衡

我想大家都听说过差不多先生的故事。差不多先生做什么事情都是马马虎虎，差不多就好，但是最后差不多先生就因为这差一点的“差不多”害死了自己。我想，谁都不想“差一点”，包括人工智能。当机器学习建立了一个模型，当然希望这个模型有很好的预测效果，但是很多时候我们发现，这个模型还是差一点，它在预测方面会出现或多或少的问题。而这个“差一点”的问题很多时候是出在“偏差”和“方差”这两个方面。并且在实践中我们发现，这个“差一点”是可以弥补的，只要权衡好偏差和方差之间的关系，就能够帮助

我们更好地运用AI思维解读数据。那么，什么是偏差，什么是方差呢？我们如何才能权衡好偏差和方差之间的关系呢？

偏差和方差概述

在机器学习中，机器从训练数据中学习到模型。任何有监督的机器学习的目标都是在给定训练数据的情况下，对输入变量和输出变量之间的映射关系（下文称为“真实函数”）进行最佳估算。既然是估算，肯定就会有误差。通常，机器学习的预测误差可分为三部分：偏差、方差以及不可减少的误差。其中，不可减少的误差是无论采用什么模型都不可能被减少的误差，因为这种误差是由未知变量引起的。比如一个预测用户是否会买车的模型能根据用户的收入、职业、年龄、家庭状况等一系列数据进行预测，但是这个预测不一定完全精准。因为一个用户是否会买车不只会受这些变量的影响，还会受生活方式等的影响，而像生活方式这样的变量通常是不可知的，因此这种误差是不可减少的，但是，偏差和方差是可以通过机器学习模型进行影响的。

偏差描述的是在使用不同的训练数据时，预测值的平均值与真实值之间的差距，即模型本身的精准度，反映的是模型本身的拟合能力。偏差的产生是因为为了使真实函数更容易学习，而对模型的假设做了简化。偏差高意味着模型是欠拟合的。

方差衡量的是在使用不同的训练数据时，真实函数的估算值所改变的量，也就是预测值的波动情况，它反映的是模型的稳定性。方差的产生是因为为了使模型的预测结果更加符合真实数据，而对模型提出了更加复杂的假设。方差高意味着模型是过拟合的。

下面我们用一个更直观的例子来说明一下偏差和方差。假如有一个视力不好的人想要学习射击，那么他不管练习多少次，也不能打中靶心。但是如果这个人的视力很好，对环境的敏感度很高，瞄准时还

能考虑到风速、风向等环境因素，那么经过训练后，他就很容易打中靶心。如果我们此时给他换一个环境，而他不能迅速适应这个环境的话，他的射击也不会有之前那样高的命中率。

在这个例子中，靶心代表的是真实值，若这个人视力不好，他的拟合能力相对而言就不高，无论怎么练习都不能很好地命中靶心，也就是预测值和真实值之间存在差距，产生了偏差。假如这个人视力没有问题，而且他能够根据环境因素来改变射击的方式以此来提高命中靶心的概率，也就是为了使输出更好地符合真实值，而提出了更多的条件假设，但是他又没有完全掌握这种方式，环境一发生变化他的命中率就会下降，这就说明了提出的假设越多，稳定性就越低，就越容易产生方差。

在图2-3中，第一行的两个靶，射击的痕迹都十分靠近靶心，说明其拟合能力比较好，也就是偏差小；第一列的两个靶，射击的痕迹都十分集中，这说明其稳定性非常好，也就是方差小。

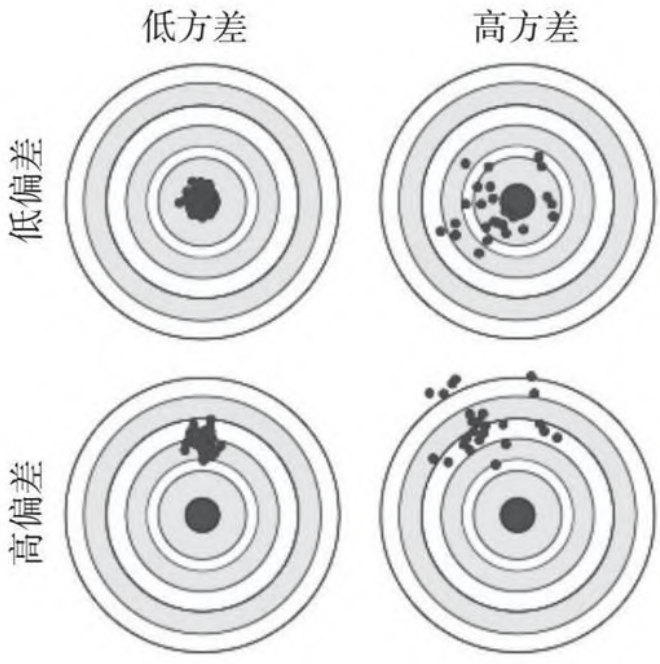


图2-3 偏差与方差示意图

在数学领域有一个概念叫作“参数”。在研究问题时，我们经常会在模型中引入一些变量来描述数据的变化，这些变量就是参数。一般来说，在机器学习中，参数化的模型会有相对较高的偏差，这样它们学习起来很快，也更容易理解，但是高偏差就意味着它对模型做了很多的假设，在训练过程中，不是每个数据都可以完美拟合，因此，这个模型就不那么灵活。反之，如果这个模型对真实函数提出的假设很少，那么这个模型的偏差就很低，几乎能拟合每一个数据，但是这样的模型近乎一种对数据的记录，预测性能就不足了。低偏差机器学习模型的例子包括：决策树、K-最近邻和支持向量机。高偏差机器学习模型的例子包括：线性回归、线性判别分析和逻辑回归。

用机器学习模型是通过训练数据来对真实函数进行估算，模型或多或少都会产生一定的方差。理想情况是，训练数据集的改变对真实函数的估算结果不会产生很大影响，这意味着该模型能够很好地计算出输入和输出变量之间隐藏的底层映射。如果模型的方差过高，训练数据对其产生的具体影响明显，那么训练的细节就会较大地影响到描述映射函数的具体参数。如果模型的方差较低，训练数据对其产生的具体影响很小，那么训练的细节对描述映射函数的具体参数的影响也会很小。

泛化误差：偏差和方差的权衡

任何有监督的机器学习模型的目标都是做到低偏差和低方差，也就是说该模型应达到很好的预测性能。但是在机器学习中，偏差和方差二者又不可兼得，偏差减少必将增加方差，方差减少必将增加偏差。造成这种现象的原因是，当我们想要尽量减少模型的偏差时，我们就会尽可能多地保证模型在训练样本上的准确度，但是这样学习出的模型就会过度拟合，反而降低了其预测能力，增加了模型的不确定性，也就是方差增加了；相反，我们在追求模型的低方差时，会增加

许多对模型的假设和限制，虽然这提高了模型的稳定性，但是忽视了拟合程度，偏差也随之提高了。所以为了模型能够拥有更好的预测能力，我们需要权衡偏差和方差的关系。

我们通过在具体情况下选择模型的复杂度来达到偏差和方差的权衡，寻找这个权衡的过程就叫作泛化误差的优化。研究发现，模型的泛化误差可以通过偏差的平方、方差以及不可减少的误差相加而得出：

$$模型的泛化误差=偏差^2+方差+不可减少的误差$$

模型的泛化误差还可以通过更加形象的方式来表示，如图2-4所示，随着模型复杂程度的上升，模型的偏差越小，方差越大，图2-4中最上方的那条线，表示的就是模型的泛化误差。我们要找的方差和偏差的最佳权衡点就是图中虚线所在之处，此时模型的复杂程度处于最佳状态，模型的泛化能力最好。

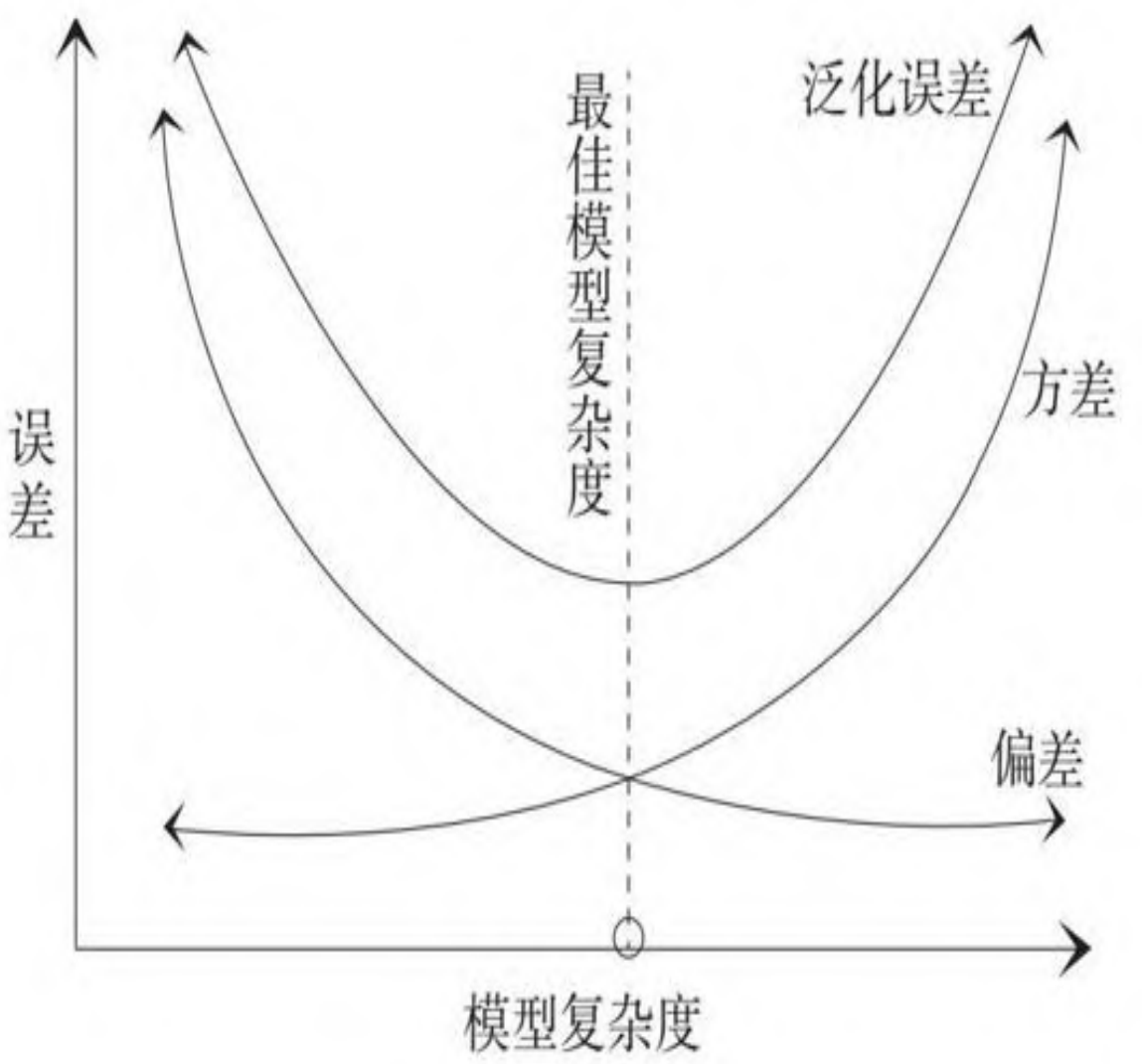


图2-4 模型的泛化误差示意图

我们从图2-4中可以很容易地发现，偏差和方差之间是存在冲突的。虽然大多数情况下，我们不能计算实际的偏差和方差项，但是对偏差和方差的把握为模型预测性能的优化提供了思维框架。

模型的复杂度：VC维度

前面我们讲到，模型的偏差和方差与模型的复杂度是有关系的。模型的偏差会随着模型复杂度的上升而下降，而模型的方差会随着模型复杂度的上升而上升。那我们是怎么来定义和量化一个模型的复杂度的呢？为了定义和量化一个模型的复杂度，我们引入一个新的概念：VC维度（Vapnik-Chervonenkis Dimension）。VC维度是根据提出者统计学家弗拉基米尔·万普尼克（VladimirVapnik）和数学家亚历克塞·泽范兰杰斯（Alexey Chervonenkis）的名字来命名的。

大自然的智慧是无穷的，它赋予我们聪慧的大脑，也在冥冥之中提示我们寻找规律，让我们首先从大自然的角度来理解VC维度。海豚经过训练可以打乒乓球、钻火圈，被称为世界上最聪明的动物，海豚大脑的重量占海豚体重的1.7%。平时我们总会用“蠢萌”这个词来形容的狗，它的大脑重量仅占它体重的0.8%。经过对多种动物大脑重量占它体重比例的研究以及这些动物日常表现的观察，科学家们提出，大脑重量占体重的比重越大的动物越聪明。人类大脑的重量占人体重的2.1%，是地球上所有生物中占比最大的，这也充分证明了上述观点。大脑占的比重越大，说明脑细胞相对就越多，也就是大脑的复杂度越高。如果从VC维度的角度来看，大脑的VC维度越高，就越聪明。

接下来，让我们从理论的角度来了解一下VC维度。首先我们需要了解一下分类器。分类是AI思维考虑的重要问题：比如我们要将顾客分为高购买价值和低购买价值两类，以此来决定我们的精准营销方案。在人工智能中，这些数据的分类是通过分类器进行的，而分类器的复杂程度取决于它能够准确分类的数据的多少。假如现在要用分类器为一组数据点分类，那么这种分类器的VC维度的定义就是：对于每个可能的数据点的标记，单个分类器可最多分类的数据点个数。但是需要注意的是，这里有一个隐含条件，就是不要求对在同一直线上的点分类。比如：如图2-5中前三张图所示，在二维空间上，一个线性分类器的VC维度是3，因为它可以对任意三个点进行准确分类，无论它们如何排列。但是如图2-5中第四张图所示，单个线性分类器无法对四个点进行准确分类，即使它们不共线。

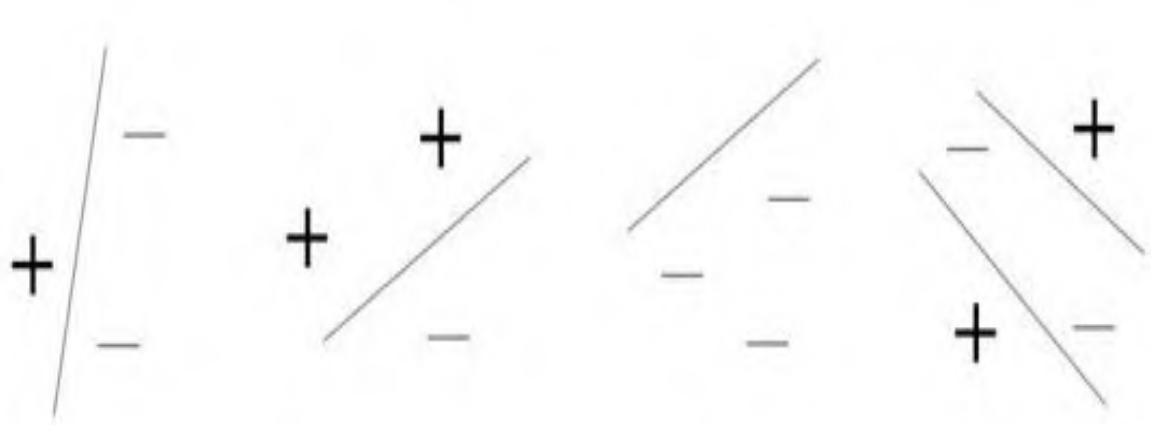


图2-5 二维空间线性分类器示意图

如果要判断分类器的VC维度就要看它能够对几个不共线的点进行分类。通过上述图片所示，我们可以直观地看出来，二维空间中单个线性分类器可以分类的最大数量是3，所以这种线性分类器的VC维度是3。而随着分类器能准确分类数据的增多，VC维度也会相应地增加。因此，VC维度是描述模型复杂程度的一种方法。

VC维度是反映泛化误差与训练误差关系的重要指标，正如以下公式所示：

$$\text{泛化误差} \leqslant \text{训练误差} + \sqrt{\frac{8}{N}\ln\left(\frac{4(2N)^{\mathrm{d}}}{\delta}\right)}$$

其中N指的是训练样本数量，d指的就是VC维度，δ指的是一个概率相关的参数。虽然这个公式看似十分复杂，但是仅从数量关系的角度看，VC维度与泛化误差和训练误差的差值呈现正相关，VC维度越大，泛化误差和训练误差的差值就越大。而我们现在已经知道了VC维

度代表的就是模型的复杂度，所以模型的复杂度与泛化误差和训练误差的差值也呈现这样一个正相关关系。那么，模型VC维度的法则在现实中有什么意义呢？

首先我们要知道，VC维度并不是越大越好，但是根据上述公式可以发现，训练样本的数量位于分母位置，VC维度位于分子位置，如果分子分母的数量同时扩大，也就是说训练样本数量和VC维度同时扩大，泛化误差和训练误差的差值可能并不会发生显著改变。换言之，如果在训练样本的数量越来越多的情况下，提升VC维度，也就是使用更加复杂的模型可能是一个更加合理的选择。所以，模型VC维度的大小是要根据训练数据的多少来进行选择的。只有选择了合适的VC维度模型，人工智能才能产出实际价值，为各行各业提供优质的服务。

我们现在使用的人工智能都是经过了漫长的发展才达到现在可以普遍应用的水平，例如语音识别从1952年就已经出现了，但当时的语音识别建立在一个数据量很小的基础上，所以当时它只能识别出10个英文数字的发音，虽然这是语音识别的开端，具有划时代的意义，但它也确实不能进入应用阶段。随着语音识别马不停蹄地发展，训练语料库越来越大，相应地，语音识别模型的VC维度也在不断升高，随之而来的是语音识别的错误率也越来越低。目前，主流智能语音识别的准确率已经超过97%，比通常人工转录的准确率还要高。

我们之前提到过的图像识别领域的“奥赛”ImageNet比赛也反映出这样的趋势。从2000年开始，每年冠军参赛模型的VC维度都在不断上升，与之相对应的，冠军模型识别图片的错误率在不断下降，在这背后，冠军模型面对的其实是数据量越来越大的现状。不论是在语音识别、图像识别还是在其他人工智能应用上，都体现了这个规律：几十年前，数据量很小，人工智能技术刚刚起步，模型不需要很高的VC维度来处理数据，所以当时建立的一些模型都很简单。但随着时代的发展，数据量的增加，为更好地解决层出不穷且越来越复杂的问题，

提供更优质的服务，近几年新建立的模型都呈现出VC维度越来越高、复杂程度越来越高的趋势。

与此同时，各种新兴行业也通过VC维度的提升有了新的发展。2019年8月2日，腾讯人工智能实验室研发的人工智能系统叫作绝悟，它在王者荣耀世界冠军杯中战胜了职业选手，证明了现在的人工智能也可达到电竞职业玩家的水平。8月3日，绝悟在中国国际数码互动娱乐展览会（ChinaJoy）上参与的504场1对1人机对战中，获胜率为99.8%，输掉的那一场是输给了国内最顶尖的专业玩家。绝悟之所以能取得如此高的胜率，正是因为选择了具有合适VC维度的模型。

这些年，中国电竞在世界级比赛中也获得了不俗的成绩：2018年雅加达亚运会电子竞技表演赛中斩获2金1银、“英雄联盟”S8全球总决赛捧得冠军奖杯。这说明中国电竞业正在逐步崛起，电竞职业选手的能力也逐渐逼近世界水平。这时如果发明出电竞人工智能，也会希望它达到世界级水平。但是，电竞人工智能所要面临的问题是十分复杂的。以王者荣耀为例，它作为一个团队作战游戏，双方每方5位参与者要在英雄搭配、技能应用、路径调换及团队协作等方面面临大量、持续而且即时的选择，其操作可能性高达$10^{20\ 000}$种。因此，绝悟在训练过程中选择的是VC维度极高的深度强化学习模型。在短短半个月的训练周期内，绝悟每天的训练强度相当于人类训练440年的量。与此同时，绝悟在训练中用了384块GPU。在这样高等级算力的支持下，VC维度高的模型的训练下，绝悟的电子竞技能力直线上升，达到了世界级电子竞技选手的水平。

“权衡”二字出自南朝刘勰的《文心雕龙》：“权衡损益，斟酌浓淡。”古人早已告诉我们，要懂得权衡才能找到处理事情的最佳方法。通过权衡偏差和方差平衡模型的复杂程度和稳定程度，找到最合适的模型；通过权衡各种条件，找到最合适的VC维度，这样打造出来的模型才是真正能解决问题的模型。弱水三千只取一瓢饮，虽然选择很多，我们只要最合适的那一个。这和AI思维是一脉相承的，通过权

衡各种复杂条件做出最佳的决策，不但可以帮助我们解决各种复杂的商业问题，还能使我们在激烈的行业竞争中脱颖而出。

相关性和因果性

伏尔泰说过：“雪崩的时候，没有一片雪花是无辜的。”那雪山是随着片片雪花的飘落而崩塌，还是因为那场忽降的大雪而崩塌？笛卡儿说过：“我思故我在。”那我们是随着思考而越发感觉到自己的存在，还是真的因为我们在思考才有了自身的存在？骆驼到底是随着稻草的增多而被压垮，还是因为最后一根稻草太沉而倒下？关于相关性和因果性的思考一直以来都存在，机器学习模型发现的便是输入变量和预测目标之间的相关性。相关性不同于因果性，因果性对应的是传统科学研究的范畴，相关性则是AI思维所关注的理念。相关性对于我们研究许多变量因子之间的相互关系有着很大的帮助。

事实上，将两个一起出现的事物关联起来是人类与生俱来的学习能力。比如说，人都怕衰老，因为衰老意味着生命的流逝；人们到了一定年纪就会喜欢照镜子，去看是不是自己脸上哪个区域又出现了皱纹，或者白发是不是又增多了，因为白发和皱纹是衰老的迹象，这对于大部分人来说可能都不是个好消息。我们一看到脸上有皱纹、头上有白发，就知道人开始进入衰老期，其实就是因为我们将皱纹、白发与衰老建立起了联系，根据这个联系在脑海中反映出它们之间的相关性。但是，从科学角度来说，并不是“皱纹”“白发”带来了人体的衰老，它们只是衰老的表象，人体衰老的根本原因在于细胞代谢能力的降低。也就是说，皱纹和衰老之间并不存在因果性，仅仅是内涵及外在表现形式的一种相关性。

人体是一个复杂的存在，除了会经受自然衰老以外，还会生各种各样的疾病。在流行病学领域，相关性也比因果性常见得多。例如，最近很多上班族觉得自己气血不足，所以在他们中间流行起了“泡枸杞、吃红枣”的养生法。这是为什么呢？难道就因为上班族每天都得对着电脑、手机十几个小时，又恰恰显出了气色差、无力等贫血迹象，就能说明是辐射引起的贫血吗？这些社交媒体制造的类似焦虑言论背后，其实埋着将相关性歪曲为因果性的雷。虽然不排除某种性质的贫血跟电子产品的辐射有一定关联，但贫血和人们使用电子产品的比例并不能断定两者之间的因果性。

实际上，白领群体的“贫血”概率与机体的运动代谢下降和饮食结构的不合理有关。白领群体长时间坐在工位上，对着电脑、手机，身体代谢变慢，血气调动不起来，这才导致了医学上的贫血现象。此外，上班族工作压力大、饮食不规律，忽略了机体基础营养的正常补给，长年累月，影响身体机能的修复，这也是引起气血不调的原因之一。所以，如果将此类的表述换成“随着全球气温升高，贫血人群的数量也在急速上升”，同样也能推出“全球温度升高容易引起贫血”的结论，这显然十分荒谬。

虽然相关性没有因果性来得一针见血，但是相关性能一定程度地揭露问题所在。作为内在问题的一种肉眼可见的表征，相关性自有它的用武之地；至于如何应用，我们会在后面的内容中进行更详细的阐述。现在让我们先来看一下相关性与因果性的定义。

什么是相关性

相传，如果南美洲亚马孙河流域热带雨林中的一只蝴蝶慵懒地扇动几下翅膀，可能会在两周之后导致美国得克萨斯州刮起一场铺天盖地的龙卷风——如果一定要说这两件事之间存在关系的话，那这个关系也是微乎其微的。

那么，如何衡量相关性的程度呢？我们可以定义相关系数，规定相关系数的取值范围在-1到1之间，两个事物之间的相关系数越接近0，就表明两者的相关性越低。对比双方数据时，如果一方数据往高走的时候，另一方数据也在上升，就说明二者正相关，它们的相关系数在0与1之间。如果一方数据往高走的时候，另一方数据下降，那么二者为负相关，相关系数在-1与0之间。相关系数的绝对值越大，相关性越强。相关系数越接近于0，相关性越弱。

举个例子，负相关性就好比饮食健康程度和生病次数之间的关系。一般来说，饮食健康程度越高，生病次数就越少，即饮食健康程度和生病次数是负相关的。不相关（相关系数为0）的数据，看起来就是一组均匀的数据点，没有什么明显的趋势。当然也有正相关的例子。譬如多年未见的朋友比小时候长得更加高大、健壮了，体重也重得多。身高和体重之间就是一种正相关的关系。一般来说，身高每长高一厘米，体重也会相应地有所增加。

但是在运用相关性进行分析的时候要十分注意。国外有个网站Spurious Correlations，讲的是虚假的相关性。这个网站上列了许多啼笑皆非的例子，这些例子中的每一对数据看似都有很高的相关性，实际上却不存在任何合理的解释。比如说，有人发现在某些年份，美国在科技领域投入的资金和上吊自杀人数，相关系数高达99.79%。这其实体现的就是一种虚假的相关性，真正具有相关性的两者要有一个关联的基点。就这个例子而言，人们并不会因为美国在科技领域投入多少，就去考虑上吊自杀的事情，所以两者不存在一个关联的基点，也就不存在真正的相关性。网站上列举的另一个虚假相关性的例子——人均芝士消耗量和新增土木工程博士数，相关系数高达95.86%。95.86%表面上代表着非常高的相关性，但实际上，两者之间并没有合理的联系。造成如此高的相关系数的原因也许只是随着时间推移，社会发展，各项指标正好同步变大而已。所以在使用相关性概念的时候，我们也要考虑两个变量内在的联系和可解释性，单纯的相关系数可能没有意义。

什么是因果性

证明因果，首先要证明两个事件有关联。这种关联一般来说是一种常见的现象，比如一些老人喜欢说的“爱笑的人都很善良”，“满脸愁容、精神颓废的人都很苦命”，“筋络粗壮的都很奔波”，等等，是根据一种静止表象来推测动态发展倾向。虽然关联现象不一定意味着因果关系，但认识到或者察觉到它，却是一个很好的起点。如果某两个事件显示出一定的相关性，往往会引起研究者的注意，吸引他们去发掘其中可能存在的因果关系。

从日常生活到严谨的科学实验室，想要证明事件之间有无关联或密切程度的方法有很多。就像前面说的，从统计学的角度，能够通过数据分析看出两个事件之间是否存在正相关和负相关的关系，如果不存在关联就不存在因果，因为关联关系是因果关系的基础。当确定两个事件之间存在关联后，我们再进一步分析两个事件之间的细节规律，建立起一个假设，即它们之间发生的先后关系的推测。想要验证这个推测，首先可以预设是某一事件导致的另一事件，这个结论成立的两个条件是：第一，一个事件发生在另一个事件之前；第二，前一个事件的出现能预测另一个事件的出现。

因果关系的链条是单一的。上述条件只是证明了两者之间存在链条，我们还要进一步检验事件之间链条的单一性，也就是排除掉其他可能的混淆变量。所谓混淆变量，就是发生在前一事件以外的其他事件导致了后一个事件的发生，这就违反了因果关系链条的单一性，这个因果关系就不是那么明确地能够成立了，需要再往深一步去挖掘其真实的因果关系链条。比如，在众所周知的狐假虎威的故事当中，每一次狐狸出现，森林里的动物们都十分害怕，四散而逃，但是动物们逃走并不是因为狐狸，而是因为狐狸的身后跟着百兽之王老虎。在这个故事中，老虎就是那个混淆变量，狐狸出现与动物们逃跑之间并不存在链条的单一性，也就是因果关系并不成立。

去掉混淆变量的常见方法是设计实验组和对照组。实验组，就是在这一组样本中对某一因素做处理。对照组，就是不做任何处理，直接放置，然后观察两组现象的差别。要证明因果，就是需要实验组出现预设的现象，而对照组不出现预设现象。如果实验组因为做了某一因素的处理而出现预设现象，对照组未做任何处理而没有出现预设现象，就能证明某一因素的改变导致了预设现象的出现，而没有这一改变就不能导致相应现象出现。

例如，最近喝各种养生茶变得流行起来，有些人喜欢清淡的，有些人接受不了苦味，有些人却偏偏喜欢特别苦的。如果我们想要知道是哪一种原料导致了茶的苦味，就可以进行一次对照实验。例如，我们的推测是里面的苦菊导致了苦味，那么我们需要泡两杯茶，一杯花果茶的茶包原料不变，照常加水泡开；将另一杯花果茶的茶包原料里的苦菊挑拣出来，再加水泡开。如果品尝后发现前一杯茶带有苦味，而后一杯茶没有苦味，则验证了我们的推测是正确的，即苦菊是导致这杯茶带苦味的原因，苦菊与茶的苦味之间的因果性成立。当然，我们还可以做更多推测，比如考察是哪种材料导致了奇怪的味道，等等。

相关性大于因果性

在日常工作生活中，我们更需要关注相关性，而不是强求因果关系，也就是说，只需要知道“是什么”，而不总是需要知道“为什么”。这能使人跳过洼地，快速地获取结果。这种思维方式推翻了自古以来人们更加注重因果关系的惯例，使我们理解现实和做出决定的最基本方式受到挑战。但是不破不立，破而后立，相关性的思维方式也让一切事物变得看起来很简单。AI思维从数据中训练出模型，但它从不思考为什么这样做，这让模型训练变得简单直接，体现出相关性思维的特点。

从传统的因果性思维转向相关性思维是AI思维最突出的一个特点，传统的因果性思维是说我们一定要找到一个原因，推出一个结果来。而基于AI思维的预测没有必要找到原因，不需要证明这个事件和那个事件之间有一个必然、先后关联发生的因果规律。在相关性思维中，如果出现了某些迹象，数据统计的高概率会显示出相应的结果。那么，当我们发现这些迹象时，需要做的就是根据这些迹象去预测结果，做出决策。这和以前的思维方式不一样，和传统科学的思维不一样。虽然同样基于数据，但是传统科学要求找到准确的因果关系。

但是，在这个纷乱复杂的社会关系里面，充满了差异性，不同的人在不同的时间节点可能会为了不同的原因去做一件同样的事情。比如，有的人觉得紫外线太强怕被晒黑，有的人觉得阳光刺眼，有的人觉得雨季要到了，这时他们都可能会买一把伞。不论中间的思考过程有多曲折，各自的理解如何大相径庭，他们最终都指向同一个结果。其实，我们在很多时候，最终为的也不过是这个结果罢了，所以在这期间耗费过多的精力去寻找里面的因果关系并非那么必要，毕竟在这个节奏如此快的社会，等你从头到尾都想得一清二楚的时候，这个事情可能早就不值得办了。

所以人工智能时代的思维更像“action!”（行动指令）、“reaction!”（立刻反应）这样一种直线思维。也就是说，你听到或者看到一个明显的指令或者迹象，你不需要去追究太多，只需要按惯例做出反应就可以。只要我发出这个口令，就一定能够得到一个既定的、预期内的回应，这就可以了。就像在运送快递时，如果有玻璃制品这样易碎的物品，快递盒子外面一般会有“易碎物品，小心轻放”的提示。看到这个提示，快递员不需要知道盒子里面是什么物品，也不需要知道它为什么易碎，他只需要在送快递过程中做到小心轻放，不让盒子中的物品破碎就可以了。

全世界的商界人士都在高呼人工智能时代来临的优势：早在很久以前，商家就明白了相关性销售的精髓。比如，商家很容易根据你买

东西的数量知道你家有多少人吃饭，你是单身还是一大家子。将牛肉干等一些简便的下酒食品与各种酒品摆在一起，也能很大程度地提升销售业绩，因为很多时候这些关联是更契合人性的。比如，你只喝酒，难免肠胃寡淡，根据人体自然需求，会需要同时补充一些重口味的小吃。喝酒就是为了爽快，谁会为了几块钱的小吃而折损了喝酒的享受呢？抓住了这些关联也就抓住了人性，从而能实现高销售。数据透露出来的信息有时确实能起到出人意料的效果。比如，下雨或者高温暴晒的时候，往往订外卖的人数会激增，这时候在办公区附近的餐馆都会提前做好“备战模式”，例如提前预备好外卖餐盒、准备好外卖常点的菜品，这些老板不需要去考虑究竟是因为下雨、暴晒的时候大家觉得出门麻烦而选择订外卖，还是因为这种天气下大家心情不好而选择点外卖，他只需要根据天气变化采取相关行动就可以。

这些例子共同说明了一个道理：用相关性思维方式来思考问题、解决问题的做法值得关注。寻找原因在以前被视为理所当然，但是大数据推翻了这个论断。过去寻找原因的信念正在被更为实用的相关性所取代。随之而来产生了一个值得思考的问题：在探求因果关系转变为注重相关关系的过程中，我们怎样才能既不损坏奠定社会繁荣和人类进步的因果推理基石，又能真正推动社会的进步呢？需要强调的是，转向相关性，并不是要抛弃因果关系这块科学基石，而是通过相关性更加快速直接地解决问题。

在当今这个信息化的时代，只要能够得到充足的数据，加之用高效的AI思维寻找到相关性信息，就可以预测用户的行为，为企业做出准确决策提供支撑。比如，金融行业非常注重风险的把控，以往我们识别金融风险最常用的方式就是查看客户的过往征信记录，寻找各种渠道去打探对方的信誉。这个过程不但耗时、费力，还未必能够保证信息全面。

有了人工智能加持之后，金融机构就可以通过数据建立模型，识别出客户相关的异常行为以及异常关系，为金融决策提供广泛、确切

的有效信息。在这个领域中，人工智能能够帮助解决金融信用体系中的关键问题。那么人工智能是如何在这个过程中发挥作用的呢？首先需要构建人工智能模型，然后通过模型实现特征分析，进而根据模型的输出判断异常。比如说现在我们要判断是否给某个客户贷款。为了解决这个问题，金融机构就要对该客户以往的行为进行分析。如果金融机构通过人工智能分析发现，这个客户曾经多次逾期还款，收入不稳定、工作情况变动大，那么这就意味着他不是一个很好的贷款对象，因为财务状况不稳定与出现财务问题之间有很大的相关性，放贷给他可能会影响到钱款的正常收回，导致出现损失。相反，如果客户过往的财务状况稳定，并且有着合理的发展规划，就说明他是一个值得信赖的客户。向这样的客户贷款，风险较低，相对安心。

尽管在过往，这些工作也能通过人工粗略地进行，但是由于细枝末节的信息不计其数，而人力始终是有限的，在这种情况下，人工的分析自然容易出现疏漏。此外，人工耗时长，对于金融这个看重时效性的行业来说，是一个重大的弱点。相比之下，人工智能通过自身的模型学习，能够在短时间内就发现隐患，使得应对决策能够正确有效地实施。所以，在商业社会里，“action!”“reaction!”的行为模式是更为合理，也是适用更广的。细究其中，你会发现，事实上所有事件的处理方式都在遵循相关性原则，无论你把事情考虑得多么周全，仍然可能出现问题；在应急的当下，你也一定是运用了相关性来处理绝大部分问题。而且根据人们使用的情况来说，在生活的方方面面，它都饶有成效。有时候，相关性可能就是人们所说的“下意识”，看到或者听到某些事物，条件反射地设想出下一秒产生的结果，下意识地为这个结果做出反应。

哲学上讲，联系是指一切事物、现象、过程及其内部诸要素之间的相互影响、相互作用和相互制约，也就是说，万物之间都是存在相关性的。也许你没有意识到相关性的存在，但就像你看到打雷就知道要下雨一样，人们无时无刻不在使用着相关性思维。相关性这种通过分析不同事件之间的联系，由起点直接到终点的思维方式，是AI思维

能够快速准确地进行预测的内在逻辑。物有本末，事有始终，掌握了相关性，你就能事半功倍地得到想要的结果。

数据的规律性

我们已经知道，人工智能的发展和应用离不开大数据的支撑。但是大数据意味着数据的数量越来越庞大，质量越来越难保证，真实性越来越难判断。对于人脑来说，随着数据的不断增加和积累，要在铺天盖地且良莠不齐的数据中找到我们所需要的那一份信息，无异于大海捞针；但对于人工智能来说，这并不是什么难事。我们不禁好奇，人工智能到底是如何办到的呢？

从表面上来看，我们所接触的信息是杂乱的、无序的，但是事物之间的联系是普遍的。例如世上万物形态各异，但是组成这些物体的元素都是相同的，也就是化学元素周期表中的一百多种元素；世界上没有完全一样的两个人，但是我们的染色体数量同样都是23对。人工智能要处理的数据同样也是这样，我们通过深入分析和研究，就会发现数据和数据之间是存在共通规律的。

对于数据来说，这种本质的规律性可以反映在数据内在的几何结构上。这个几何结构并不是我们日常生活中所常见的，而是存在于一个更加复杂的空间，复杂到可能有些超出我们的直观想象。例如我们都知道一张照片是由许多像素点组成的，当我们随手拍了一张1 920×2 560像素的自拍时，这些像素间便互相组合形成不同的维度。也就是说，这张照片所对应的数据就是由4 915 200个维度组成的，即使我们的肉眼只能看到一张平面的照片。人工智能处理和理解的数据，其实都是高维的；我们将这些数据所在的空间称为高维空间。要想掌握数据的内在规律性，就要先对高维空间进行了解。

高维空间的几何学

我们一般会认为我们生活在三维的立体世界中，但很多物理科学家认为，空间不只有三个维度，并对空间维度的相关问题争论不休。

爱因斯坦在他的《相对论》中提出，整个世界是由时间和空间组成的，时间就是第四维。时间作为一个坐标，确立了三维空间在一维的时间轴上的排列。世界著名理论物理学家丽莎·蓝道尔教授在一次核裂变实验中发现少量粒子突然凭空消失了。她推断这些粒子可能去了我们看不到的“五维空间”。英国著名物理学家霍金在物理学和宇宙学领域同样具有高度权威。霍金在给出宇宙模型时提出，只用四个维度来描述宇宙是不够的，粒子、星球运动、宇宙引力等现象都不能用四维来很好地解释。他提出，至少要用十一个未知数才能将宇宙描述清楚，也就是说，宇宙至少是由十一个维度组成的。虽然我们感受不到如此多的维度，但这并不代表它们不存在。就像我们听不到超声波，看不见紫外线一样，我们感受不到的那些维度只是蜷缩隐藏起来了而已。

平时我们最多只能感受到时间和空间，如此玄幻的多维度可能会让你疑惑。简单来说，任何一级低维空间都是高一级空间的横截面，例如，一维的线是二维的平面的横截面。也就是说，任何一维空间除被本维空间的特有维度确立外，还以横截面的形式在高一维空间中有所体现，如图2-6所示，如以宽为轴移动一维的线，就形成了二维的平面。理论上，生活在三维世界的我们完全有可能以时间为轴进入四维世界，也就是我们通常说的打开了“时空隧道”。

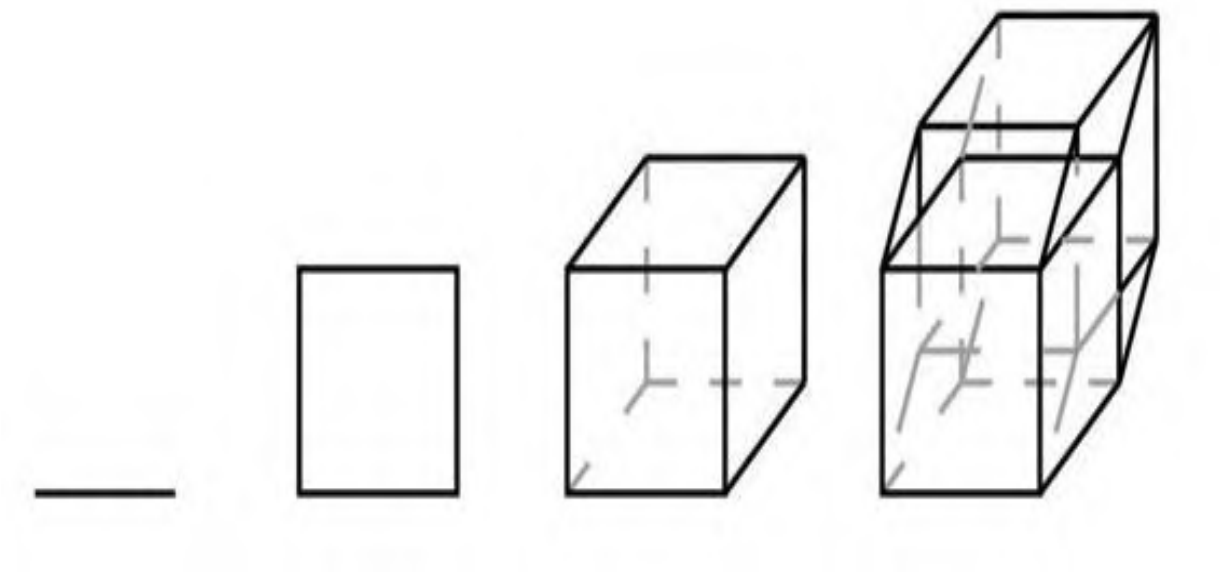


图2-6 一维至四维空间示意图

物理领域高维空间的提出帮我们打开了探索这个高深莫测的世界的大门；相应地，人工智能领域中也存在高维空间。人工智能处理的数据涉及各个领域，这些数据都可以被看成空间里的一个个点。不同的数据点所在的高维空间，也就是数学中的欧氏空间。

对一个给定的欧氏空间来说，它的维度是确定的：N维欧氏空间有N个独立的维度。欧氏空间给我们提供了高维空间的坐标系和距离的概念，任何数据在欧氏空间都是以向量的形式存在的。人工智能分析的每一条数据，都可以在欧氏空间中转换到给定维度的坐标系里，对应到欧氏空间中的某一个点。基于这样的结构，我们就可以精准地刻画我们想要描述和理解的个体。人工智能就是这样准确快速地在海量的数据中表示出我们感兴趣的那一份信息的。

虽然欧氏空间通常是高维空间，但当我们分析数据内部结构的特点时，就会发现，数据的内部存在一定的规律性，并且这些规律性可以被用来解释为什么“美女”都长得很像，以及动画片是怎么生成的。接下来，让我们一起来了解数据内部到底有什么样的规律。

线性规律：佳丽的特征脸

图2-7是多位参加“韩国小姐”选美比赛的佳丽的照片。这些佳丽面容都十分秀丽，我们细看会发现每一位佳丽长得都不一样，但总体来看又十分相像。由于这场比赛发生于整容业十分发达的韩国，很多人表示这些佳丽仿佛是多胞胎一样，也有不少人在质疑她们是出自同一条整容流水线的同时，又纷纷感叹被称为亚洲四大“邪术”之一的韩国整容技术的高度一致性。

为了更好地理解这些佳丽脸部存在的规律，我们对这些佳丽的脸部图像做了处理，形成了她们的特征脸。特征脸是反映这些佳丽脸部主要变化特征的一组“模板”。为了更加直观地反映出这些佳丽长相的相似性，我们又从这些佳丽的特征脸中挑选了其中十张，如图2-8所示，这十张特征脸包括了这些佳丽脸部图像的主要变化。

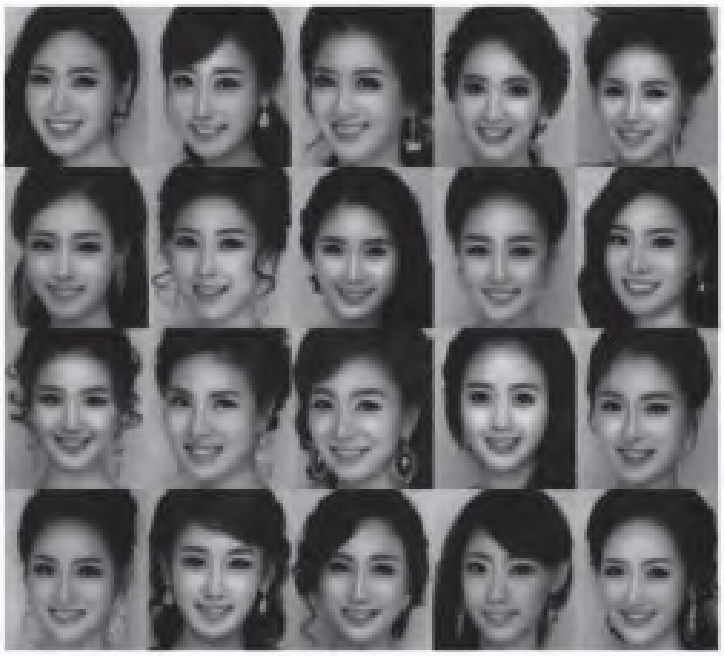


图2-7 佳丽照片样例



图2-8 佳丽照片产生的特征脸

显然，每位佳丽的脸都有和特征脸不同的地方。接下来，我们要对每个人的脸部做重建。重建就是把一个佳丽的脸投影到特征脸，然后再线性组合起来，这样就得到了重建之后的佳丽的脸。如图2-9所示，上面一行是原始图像，下面一行是原始图像投影到上述特征脸后的重建图像，通过对比我们可以发现，对应的上下两图之间的误差其实很小。



图2-9 佳丽照片的原始图像和重建图像

然后我们将每张佳丽的脸部图像投影到特征脸上，分析出这些脸部图像在特征空间中是怎样分布的。图2-10表示的是对于这些佳丽的脸部图像，不同特征脸所对应的权重，从中我们可以看到，大部分特征差异都集中在前几个特征脸上。也就是说，前几个特征脸捕捉了佳丽脸部特征绝大部分的变化。

通过上述分析，我们可以看出，这些佳丽在外貌上是有规律可循的。这个规律就是，不论她们在外貌上各有什么样的特征，但是如果把她们的外貌投影到一个平面，也就是我们所说的线性空间中，她们的外貌基本上在特征脸构成的线性空间上都能找到很好的表示。

这种基于数据的线性规律的“特征脸”，在我们的日常生活中有着广泛的应用。比如在一些重大考试中，传统的人工核验身份证和准

考证费时费力，而且很容易出错。针对这个问题，有人发明了一套基于人脸数据的线性规律的考场人脸识别系统。这套系统是要核验考生和二代身份证上的人是否一致，所以要提前提取考生的面部图像信息，对二代身份证上的照片进行预处理。在保证照片大小、考生面部位置一致的基础上，拿这些图像生成这些考生的特征脸，并将其定义为一个面部空间。在此之后，将考生的面部图像全部投影到面部空间上，计算出每一个人脸图像在该空间上的相关分布位置。经过这样的处理，我们就可以用这套系统进行人脸识别了：只需要将待识别的人脸图像投影到这组特征脸上进行对比，就可以知道该人脸是否是人脸库中的人脸。

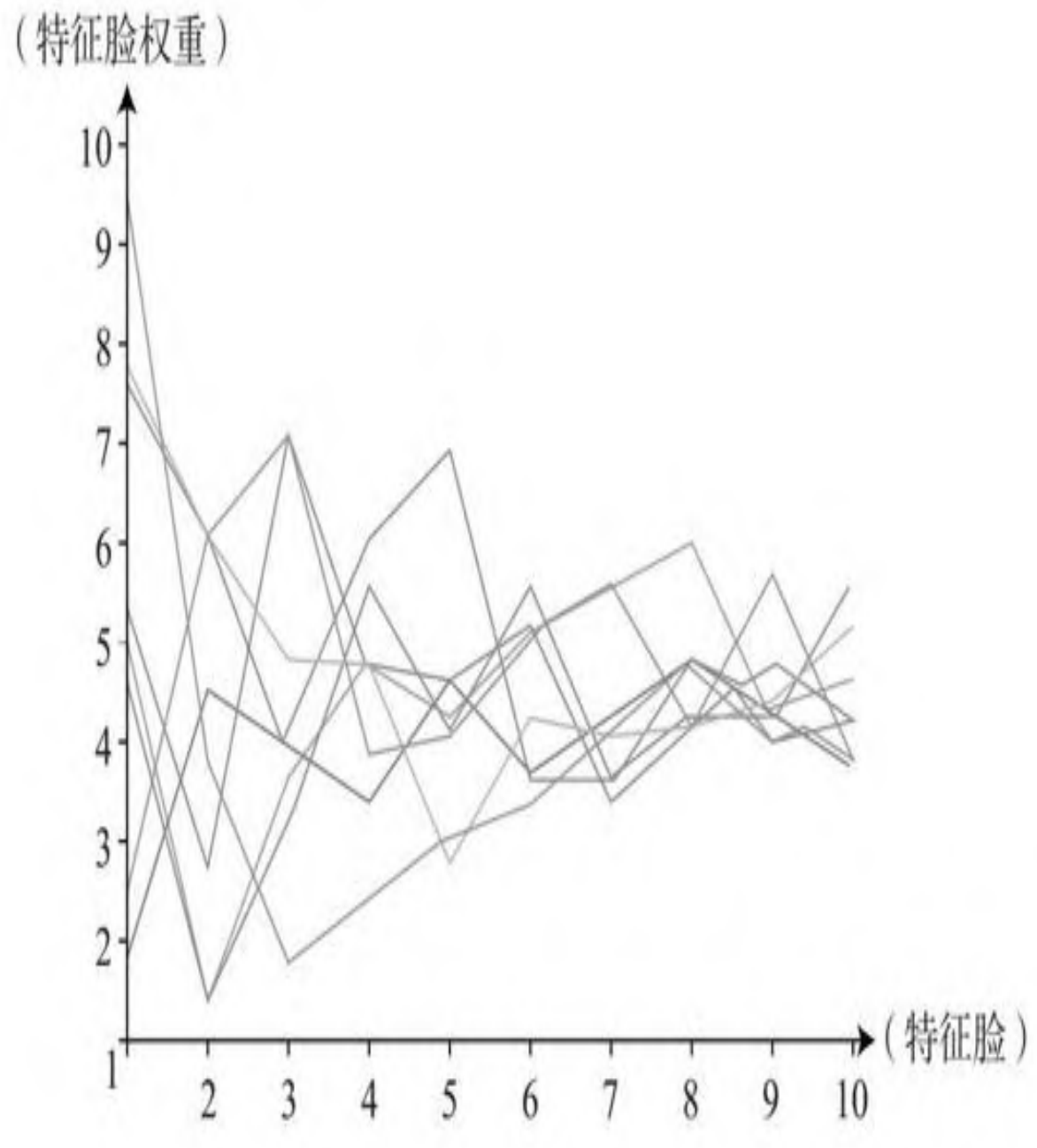


图2-10 特征脸权重折线图

数据的线性规律源于数学上在线性空间的投影，它去除了数据中不重要、不显著或者没有用的信息，只保留最本质、所有数据共同拥有的特征，所以投影后形成的数据比原始数据更简洁，更具有普遍性。不论是选美小姐的特征脸，还是考生的特征脸，都是依赖于这类

图像数据的内在线性规律性，将信息尽可能地简化，降低了数据处理的复杂性，更好地把握数据的本质特点。

非线性规律：流形学习

前面讲了数据的线性规律，我们再来讲一下数据的非线性规律。什么叫作数据的非线性规律？我们这里用一个直观的例子来说明一下。如图2-11所示，这幅图叫作“瑞士卷”。我们所生活的三维空间是一个平坦的欧氏空间；瑞士卷与我们所生活的三维空间不同，它是一个曲面。

与我们日常所见的平面相比，曲面有很多不同的特点。譬如，我们在图2-11所展示的瑞士卷上取两个点：A和B，如果要从A点到B点，直观上，我们所要经过的路线应该是两点之间那条弯曲的线，而不是像在三维空间中简单地通过A点到B点之间的直线过去。数学上，流形是局部具有欧氏空间性质的空间，它能够帮助我们更好地理解数据中“曲面”的概念。人工智能对非线性结构的研究认为，数据的非线性规律实际上是由一个低维流形映射到高维空间所产生的。一些高维中的数据会产生维度上的冗余，实际上这些数据用一个比较低的维度就能表示了。例如这个三维的瑞士卷，其本质是三维空间中的二维流形。

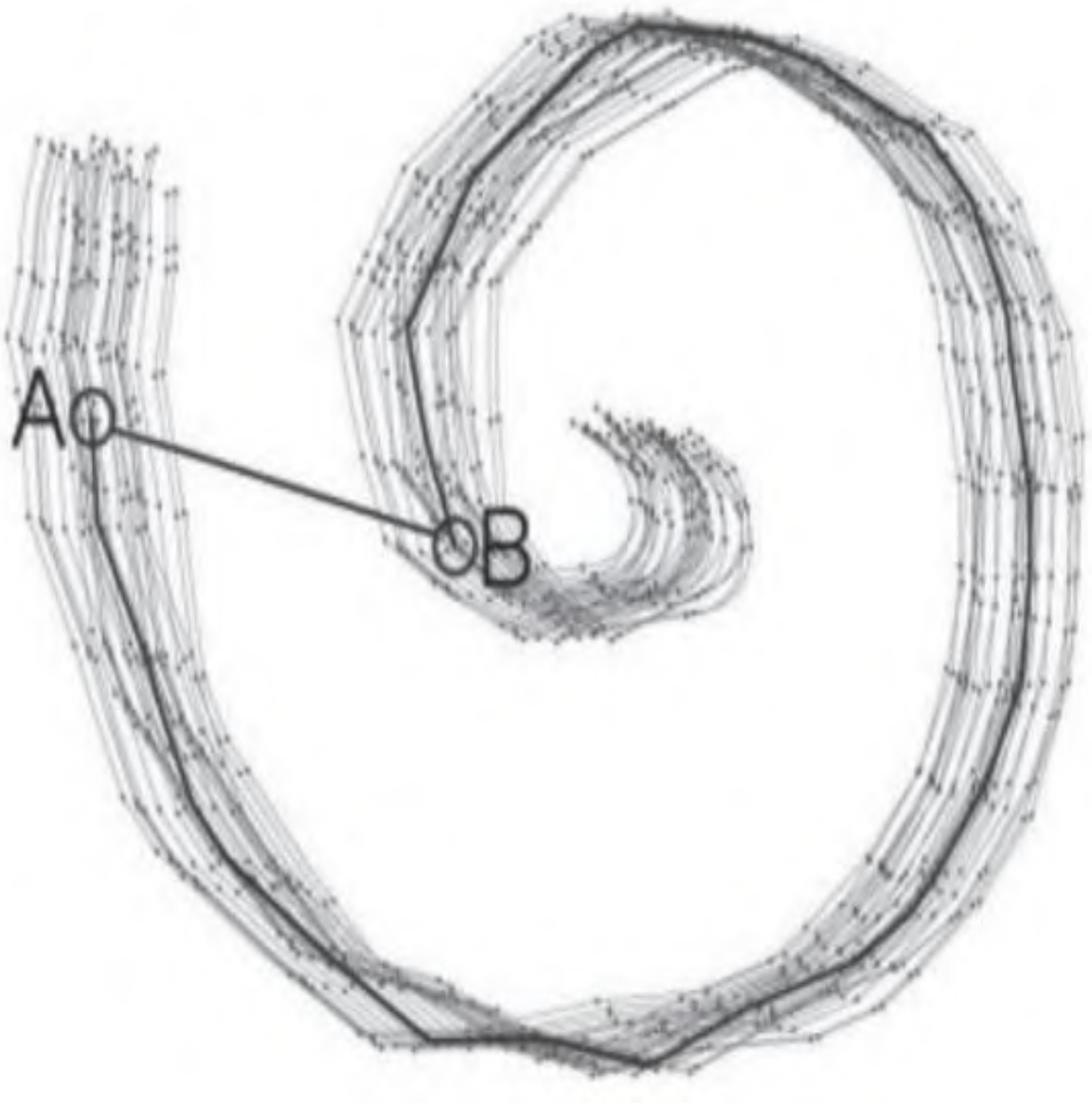


图2-11 瑞士卷示意图

通过上述对瑞士卷的描述，我们就很容易理解，非线性规律并不是存在于平面之中，而是在曲面上。发现数据的非线性规律的方法叫作“流形学习”。流形学习是在著名的科学杂志《科学》（*Science*）中被提出来的，引起了人工智能界的极大关注。它假设数据均匀采样于一个高维欧氏空间中的低维流形，即从高维数据中恢复低维流形结

构，并求出相应的映射关系，以实现数据的智能分析或者数据可视化。

流形学习可以用于研究人脸数据。例如，我们从已有的人脸数据库中选择一些人脸图像。经过流形学习分析，就会发现这些人脸图像由于拍摄角度、光线或者是表情姿势的不同，呈现出不同的特征。如图2-12左边的图所示，每幅图像对应空间中的一个点。为了让可视化结果更加明确，图2-12右边的图中简化保留了十幅最有代表性的图像。

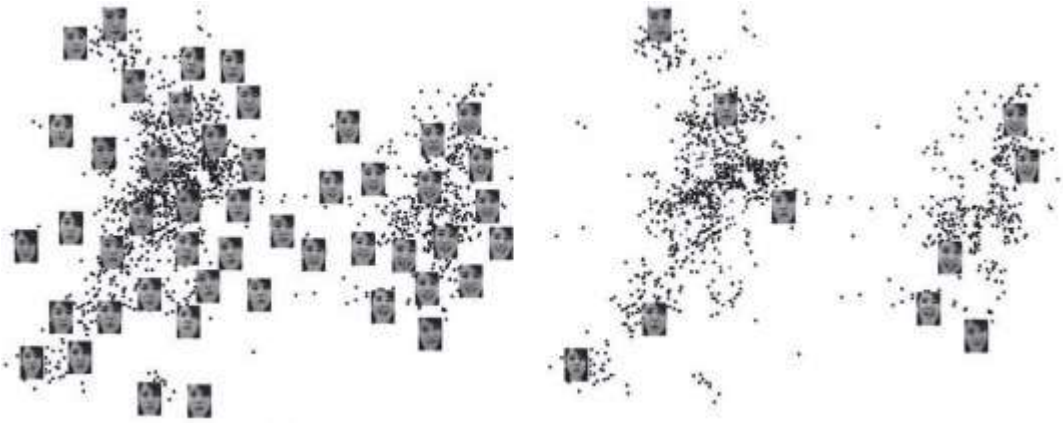


图2-12 人脸高维空间示意图

对上图仔细分析我们可以发现，这些代表性的面部图像显示在空间的不同部分。人脸图像被分割成两部分。左侧部分包括闭着嘴的面部图像，右侧部分包括张口的面部图像。这是因为在我们的面部图像被投射到更低维的空间时，图像中的相邻点在低维空间中更近，图像中的远点在低维空间中更远，所以形成了我们现在看到的排列顺序。经过观察我们可以发现，这些图像是按照人脸的表情和姿势来分布的，这就是隐藏于其中的分布规律。通过这个实验，我们发现了这些人脸图像所构成的流形，那些在线性空间表示不了的姿势和表情，在人脸的非线性结构之中都可以表示出来。

近年来，准确地刻画人体动作成为多个领域的研究热点。而随着人工智能的兴起和发展，越来越多的人开始从全新的角度理解人体动作，例如根据人体动作数据的非线性规律来建立人体动作模型。利用数据的非线性结构来建立动作模型有广泛的实用性，例如用于动作识别或者动画生成。下面我们就从这两个方面来了解非线性规律下的人体动作模型。

人体动作识别是计算机视觉领域的一个重要问题。比如说，在体育或者医疗场景下经常需要准确地识别出人体的动作。以前如果想要对人体动作进行精准的识别，需要用多个摄像机或从多个视角拍摄记录人体动作。虽然这样生成的模型可以识别出人体的动作，但这一般是在实验室环境中进行的，过程繁复，效率低下。随着流形学习的发展，人们开始尝试通过人体动作的非线性规律来构建动作模型。通过这种方式来建立动作模型，仅用一个摄像机拍摄即可，然后通过流形学习方法对拍摄的视频进行分析，利用动作数据的非线性流形结构来形成动作模型。这样生成的动作模型，对于站立、转身和行走等十三种动作类型的识别错误率仅在1%左右，可以在自然环境下应用，比如可以用于识别比赛中运动员或者医院中病人的动作。

下面再来了解用于动画生成的人体动作模型，它可以帮助动画师生成自然流畅的动画人物的运动过程，让角色动画的创作过程变得更加便捷。传统动画制作分为两种，一种是逐帧动画，需要动画师一帧一帧地进行绘画制作，虽然效果真实细腻，但是制作时间极长，制作负担很大，最终输出的文件量也很大；另一种是补间动画，需要通过计算机自动补帧来生成动画，虽然减轻了工作负担，但是生成的人物动作会出现不自然不连贯的现象。近年来基于流形的动作模型越发成熟，可以将动画制作置于流形空间中，而不是用传统动画制作的方式进行。根据这个思路，需要自动学习在非线性流形中的大量人物动作数据，这样形成的动作模型就可以用来自动合成动画。如图2-13所示，动画师如果想要生成人物在某条弯曲的道路上奔跑或行走的动

画，可以直接通过调取关于奔跑或行走的流形数据，轻松生成真实自然的人物动画序列。



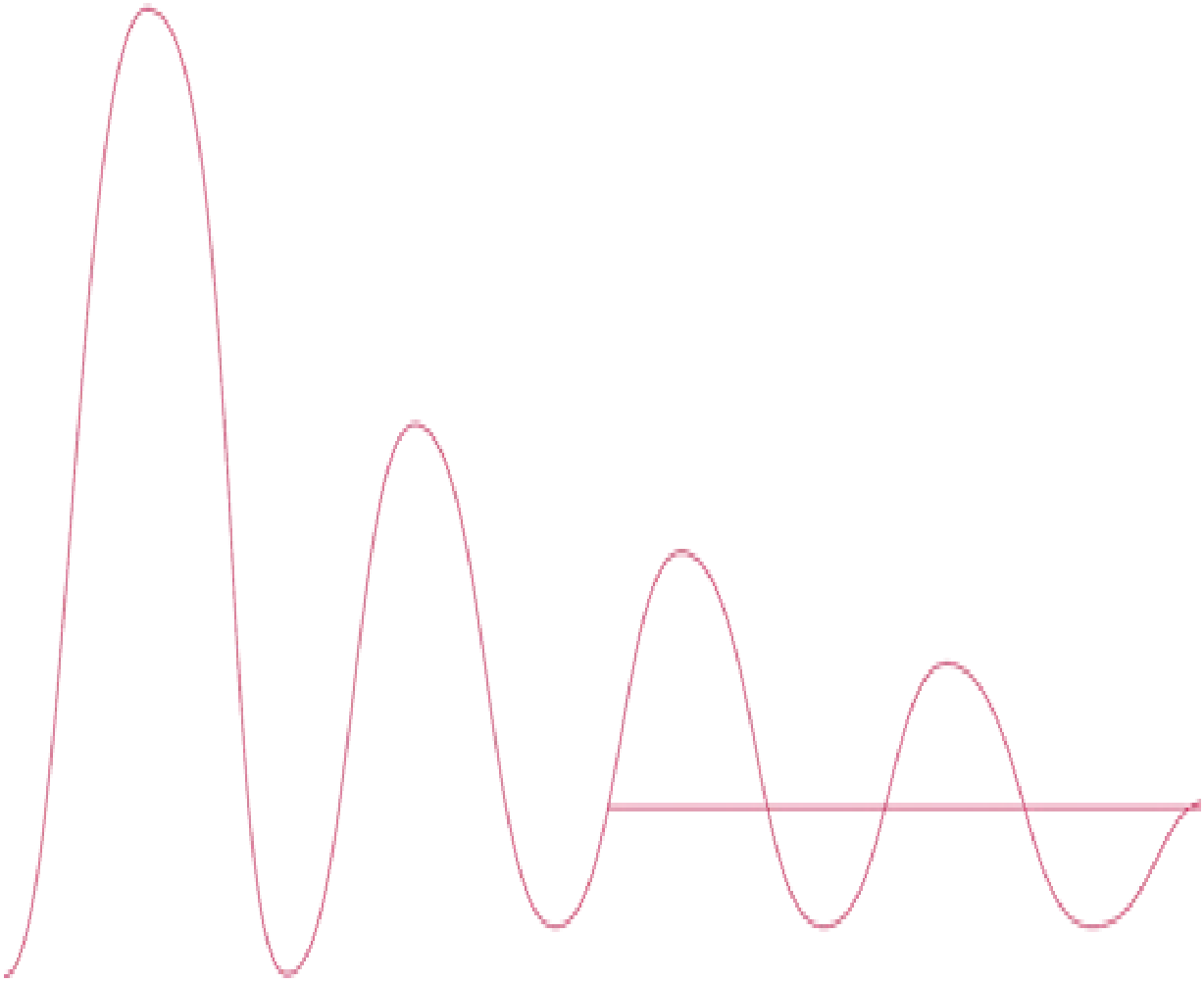
图2-13 人物动画序列示意图

对于佳丽来说，特征脸揭示了她们脸部图片数据的线性规律；对于人体动作而言，非线性结构提供了更好地运用相应数据的框架。但需要特别注意的一点是，不论是具有线性规律的数据还是具有非线性规律的数据，都是具有许多维度的高维数据，要想找到这些数据的规律，都需要将其投射到更低的维度。这个更低维度预先是不确定的，我们通过人工智能模型的分析和处理，最终在恰当的低维度空间里更好地表示了数据内在的结构。

天不言而四时行，地不语而百物生，天地万物都有其运行的规律。本节讨论的数据内在的规律性是AI思维的底层逻辑之一，也是基于几何结构的人工智能模型形成的基础。理解这些规律，可以帮助我们更好地理解数据本身，帮助我们在使用人工智能时“知其然”更“知其所以然”，构筑一个更加完整的AI思维体系。

第三章

AI炼金术：数据产生价值



这正是历史知识的矛盾。知识如果不能改变行为，就没有用处。但知识一旦改变了行为，本身就立刻失去意义。我们拥有越多数据，

对历史了解越深入，历史的轨迹就改变得越快，我们的知识也过时得越快。

—《未来简史》—

方法论是以解决问题为目标的体系或系统。在方法论的指导下我们会有一个纵观全局的广阔视野，有一套统筹全局的落地方法。缺乏方法论的指导，不论多么充满智慧的思维都难逃纸上空谈的结局。AI思维就是从理论到实践的一套完整的思维，它既有理论层面关于人工智能的认识，也有实践层面指导人工智能落地的方法论。本章旨在讲清楚AI思维中的“术”，也就是从方法论的角度阐述AI思维是如何从数据中得出决策，创造出源源不断的价值的。数据量越发庞大，以术理事，AI炼金术可理清纷繁复杂的数据；新问题层出不穷，以术驭事，AI炼金术可将数据为我所用，优化业务结果；面临抉择犹豫不决，以术成事，AI炼金术可通过深度学习等人工智能模型驱动决策。接下来就让我们了解一下，AI炼金术到底是如何从数据中创造出价值的。

从数据到价值的历程

在人工智能的整个生态系统中，数据既是根本，也是重点。在人类社会中，数据无处不在，而且生生不息，越来越多。我们每天的所看所感都可以转化为数据，这些数据就是各种行为和现象的记录。特别是在互联网如此发达的当下，我们能够接触和获取到的数据更为多维、多样而且庞杂。当然，这对于人工智能的运用来说，完全是一件好事，因为人工智能的运行离不开数据的支持。对于人工智能来说，数据更多，就相当于炼油厂运来了更多的原油，人工智能就能够更好地运作，为人类的生产、生活增添价值。人工智能通过数据驱动决策而产生的变现是我们能够真实看到的，比如优化提升的用户体验，实打实的业绩、利润，或者是口袋里不断飙升的年终奖。

关于数据的重要性，我们先从奈飞（Netflix）的案例说起。奈飞最出名的身份是世界最大的在线影片租赁服务商，但它同时也是美国的流媒体巨头，业务广泛，也涉足影视剧拍摄。奈飞想在好剧频出的美剧圈拍摄出一部像《权力的游戏》《生活大爆炸》《破产姐妹》这样的高人气剧集，于是它开始探索——究竟该怎么做呢？据统计，奈飞在全世界的订阅用户总量达3 300万；每天，用户都会在奈飞的网站上产生3 000多万个行为，订阅用户会在网站上给出400多万个评分和300多万次搜索请求，询问剧集播放时间等。奈飞借助这些大数据进一步分析——通过分析几千万用户的观影历史、观影题材内容、故事情节、演员选择以及对不同影视剧的评分等数据，决定拍摄一部有关一位参议员的连续剧，这部剧就是政治题材的人气美剧《纸牌屋》。该剧一经播出就获得了巨大成功，在美国及40多个国家和地区成为最热门的在线剧集之一。因为这次新鲜的尝试，奈飞前所未有地获得了112项艾美奖提名，也让美剧巨头HBO（Home Box Offi ce）电视网18年来第一次失去了剧集霸主的地位。

《纸牌屋》的惊人成就离不开奈飞数以亿万计的数据的支持。由此可见，数据是新时代商业取胜的重要因素。就像上面说的，奈飞通过对大量受众数据的分析，采取合理决策方案，最终获取了极大的效益和声誉。所以数据是AI思维的基础，要从数据产生价值，就要利用人工智能通过对数据的深度分析，引导决策者做出最优决策，从而落地价值。

如何把数据变为价值，主要涉及两个问题。第一个问题，什么样的数据在场景中有价值？无论是在电商行业、广告行业、教育行业，还是生产制造行业，企业的负责人都会不约而同地考虑他们要在网站或者设备、机器上收集什么样的数据。其实，只要是有效数据，都能为之所用，只不过要如何用或者用来做什么各不相同罢了。第二个问题是，如何通过人工智能让这些数据产生价值？这个问题已经得到了初步的解答：人工智能从诞生到现在已经经历了几十个年头，所以行

业内已经有比较通用的模式应用人工智能，从而可以从数据中提取价值。

人工智能的应用并非只对行业的格局产生影响，实际上，它与我们的日常生活也息息相关。现代人每天奔波于公司与家的两点一线中，一旦有了节假日这样属于自己的时间，大部分人都会选择出门逛街看风景、到商场购物，或是看一场热映的电影，抛掉平日工作的紧张，让身心彻底放松。所以，我们都希望自己能够得到全方位的引导和服务，而不想再在这方面花费大量时间和精力了。但是，一般情况下，商家并不能很清楚地知道哪些顾客需要什么样的商品，所以尽管每天都有很多顾客路过店家门口，却并不能实现有效转化。另一方面，消费者难得有了一个放松休闲的机会，却因为没能够快速、准确地找到相关商品或者服务而影响了大好心情。为解决这些问题，人工智能就派上了用场：商家可以通过人工智能对顾客数据进行分析，预测顾客需求，根据需求定期向顾客推送商品或者服务的信息；用户接收到信息，也不必为自己在休息时去哪里、做什么这样的问题而绞尽脑汁，只需要放假时直接去商家店里选择自己心仪的商品和服务就可以了。最终的结果很大概率是产生了交易。换言之，人工智能为商业活动带来了实际价值。

从AI思维角度来说，人工智能可以从数据中提取价值，可只有数据是不够的，数据量还必须充足。如果没有对足够的数据进行分析，或者对数据的利用不够恰当，都可能导致决策失误。决策一旦失误，自然会影响到整个行为最终诉诸的回报，也就是我们所说的价值。

数据产生决策：为了懂你，我学了很多知识

有了充足的数据，就等于是为人工智能这个系统备足了原料，接下来我们就来看一下，AI炼金术是如何从数据中挖掘出价值的。如图

3-1所示，这整个过程包含了四个步骤。我们将通过全球知名的高端化妆品品牌兰蔻使用人工智能机器人促进销量的案例来进行仔细分析。

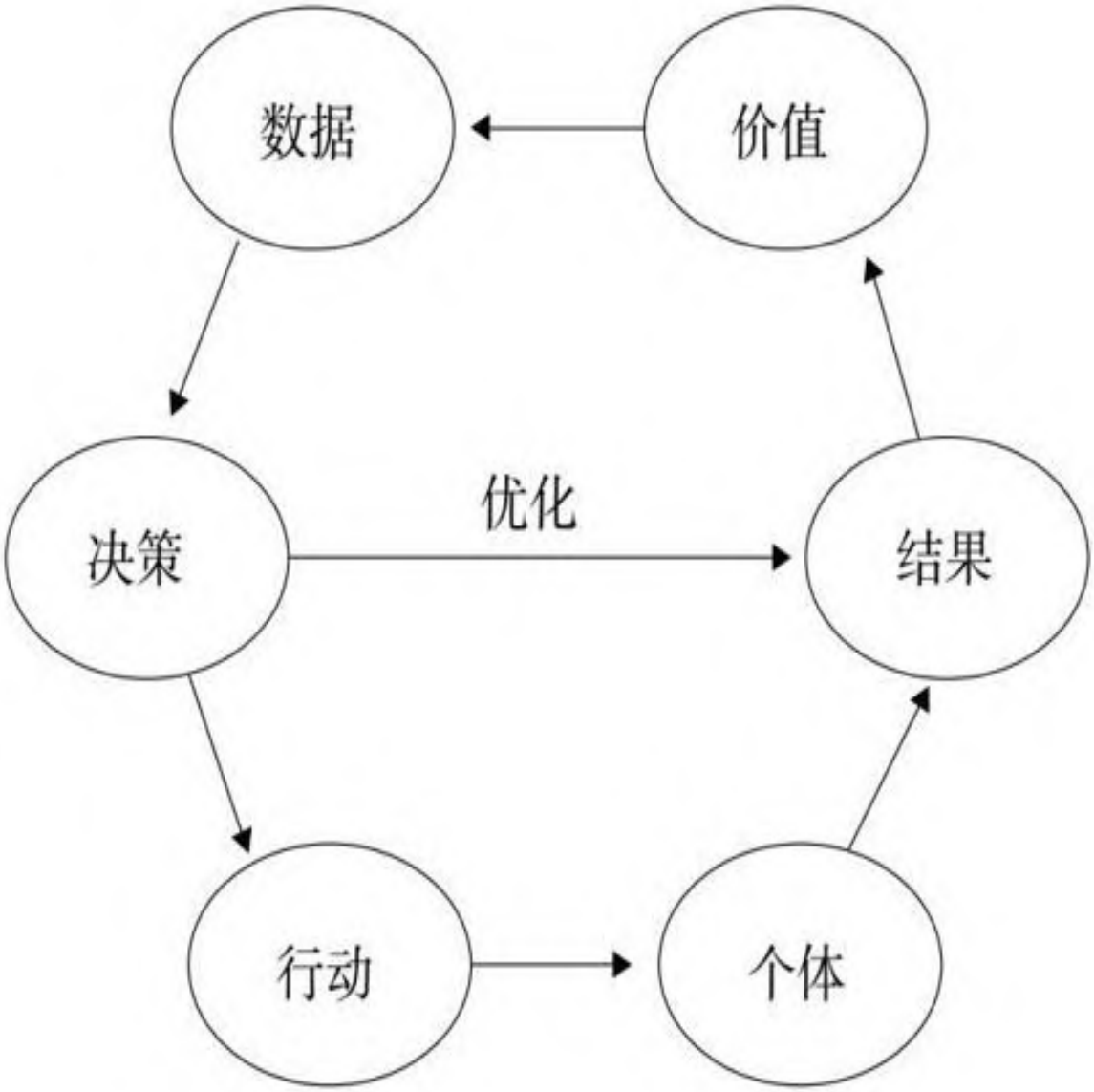


图3-1 从数据产生价值的过程示意图

第一步，从数据产生决策。消费者千千万，每个人都有各自的喜好和适用款，所以他们的行为是多样的；而从人的单方面观感上又很难判断哪类人、多少人喜欢怎样的商品、怎样的服务，适合怎样的款

式。为此，兰蔻引入人工智能机器人，并使其化身为兰蔻专柜导购中的一员。要使人工智能机器人胜任这个岗位需要做很多前期准备，其中最重要的两点是——第一，熟记并灵活运用所有产品信息、卖点和美妆知识；第二，快速明确顾客需求，推荐给顾客真正需要的东西。

从具体场景来说，兰蔻专柜的大部分顾客是女性，但不同女性对化妆品、护肤品的需求千差万别。有的人注重爆款，有的人注重功效，有的人注重成分，甚至有人买化妆品只是为了试色。对于过往的真人导购来说，首先，他们必须了解每一种产品的功效、成分、适用人群等，才能很好地应对顾客的层层“拷问”。其次，导购还需要熟悉不同肤质、不同肤色、不同脸型的顾客适合的美妆产品，才能留住顾客的心。人工智能机器人导购员如何掌握这样的“读心术”呢？一方面，人工智能机器人通过与顾客进行对话交流，理解和预测出顾客的需求；另一方面，人工智能机器人还需要从多种来源的数据中挖掘获取美妆行业的知识，并通过分析整合，为用户进行精准的推荐与引导，让顾客享受到愉悦的购物体验。

决策到行动：为了提高顾客满意度，我使尽浑身解数

第二步，从决策到具体行动。从数据到决策的过程让我们知道了不同用户的需求，为具体行动提供了方向。具体行动是要深入到个体，根据目标用户的具体喜好来分析的。为了让用户的体验最优化，人工智能机器人做足了功课，不遗余力地提高顾客的满意度。在给顾客进行推介交流时，人工智能机器人可以像真人导购一样为大家进行对话式个性化推荐。

在对话的场景下，人工智能机器人可以通过三种方式向顾客进行推介。首先是要理解顾客的语言来分析对方需求，例如，对方说：“我这几天脸上皱纹有点多，皮肤也很干，有什么产品推荐吗？”人工智能机器人能够根据对这些语句的理解和分析，进行智能推荐：

“试试具有滋润功效的小黑瓶！”其次，人工智能机器人可以通过人脸属性识别来了解用户需求。例如，有顾客问：“我这个年纪适合用什么呀？”人工智能机器人能够通过对该顾客面部的扫描识别，发现对方有皱纹、年龄偏大，并非常情感化地回复：“皱纹不可怕，推荐塑颜紧致系列哦！”最后，基于前期大量的知识学习和数据输入，人工智能机器人能够进行多元数据挖掘与推荐。比如，当有顾客表露：“我要一只鬼怪色口红送人。”人工智能机器人能够将这些信息与内置的数据库匹配，给出精准引导：“鬼怪色的色号是290，超美，在口红区！”或者当顾客问道：“最近换季，有什么推荐呢？”人工智能机器人也能根据所学知识为顾客解忧：“试试水光润系列，深入补水还能抗氧化，非常火，广受好评哦！”有些顾客拿不准自己的偏好，可能会问：“我想买个眼霜，最近流行的哪种比较好呀？”此时，人工智能机器人会根据以往的消费者数据，提供合理的推荐：“今天已经有19个顾客小姐姐询问过小黑瓶大眼精华眼霜呢，要不要试试？”

无论是哪一类顾客、哪一种需求，人工智能机器人都能根据大量的数据基础，依靠人工智能进行分析，为顾客提供高质量的推荐。这样，在精准给到高质量信息的情况下，人工智能机器人提高了顾客的满意度。

个体产生结果：真是太能干了！

第三步，个体产生结果。在得到了这一系列的推荐后，顾客开始做出相应的行动。在此之前，兰蔻专柜的负责人曾表示：“有时候顾客太多，难免有照顾不周的地方，如何提升服务质量、给顾客最佳的体验，是我们一直思考的事情。”而当人工智能机器人加入后，迎送宾客、为顾客解答疑问、推荐化妆品、漫游行走寻找指定产品等工作都可以全权交给人工智能机器人去做，并且这样能给顾客带来趣味

性、多样化的购物体验。在场受到人工智能机器人帮助的顾客都惊叹：“真是太能干了！”

更准确地来说，人工智能机器人在为导购人员提供更好的帮助的同时，也为顾客带来了更智能的服务，让他们拥有“上帝之感”，开启了顾客愉快的购物体验。在现场，有位女士问：“能帮我找一下大粉水吗？”“人工智能机器人带你去看看吧。”这位女士感叹：“万万没想到，接待自己的居然是一个机器人！”

以往人们去化妆品专卖店购物，进门不到三分钟就会有好几个美容导购员寸步不离地进行推销，即使有些人已经有了选购的目标，导购员依然会跟在身后不厌其烦地介绍。这些导购员的过度热情令人十分尴尬，也使得导购员这一角色备受争议。但是如今，导购机器人的出现解决了这类问题，同时也满足了顾客对咨询、引导的需要。

实现价值：人工智能服务于人类的一次探索

第四步，实现价值。说到价值，人工智能机器人的出现，满足了顾客的购物需求，给顾客带来了美好的休闲体验。这对于商家来说又意味着什么呢？或者说，商家从中获得了什么呢？这次兰蔻与人工智能机器人的跨界合作，是美妆与人工智能的一次探索。目前，人工智能已经落地多个行业，有效地辅助着人们的工作。有了“美妆知识+机器人”这样的人工智能解决方案，导购能够从基础的重复性劳动中解放出来，学习更多的专业知识和技巧，提升个人服务能力。这也就意味着，商家能够在减少许多人力成本的同时，更好地在导购层面提升质量。除此之外，人工智能机器人还可以基于人脸属性识别和顾客历史消费数据等，进行精准营销。比如，针对陌生的女性顾客优先推荐促销信息，或者为有商品咨询记录的顾客提供定制化的商品介绍、引导等购物服务。这样，给商家带来的最终的价值是，他们能够获取极高的顾客购买率。

人的认知有时候是一个很主观的过程，在化妆品领域更是如此。可能这几天周围的人都在讨论豆沙色的口红或者玫瑰香的粉底液，你就直观地认为他们一定喜欢这类化妆品，或者说全国、全世界范围内的顾客都会喜欢这个款式。但有时候人们的讨论焦点的触发动机并不是单一的，这些动机在日常的对话中并不能准确体现出来，那么这样的主观揣测就未免过于武断。有些顾客之所以买豆沙色的口红，是因为这款口红刚好是她喜欢的演员或者偶像代言的；有些顾客之所以买豆沙色的口红，是因为刚好有套衣服与之相配。这种阶段性的需求可能过段时间就不灵了。如果始终按照之前的需求为顾客推荐，就可能会出现顾客需要粉底液而为她推荐防晒霜的“错配”现象。在这一方面，人工智能可以精准地抓取顾客在方方面面留下的迹象进行组合判断，更全面、更精准、更客观地判断出你究竟对哪种化妆品里的哪种因素更“来电”。所以，我们要摒弃下意识的主观感受，恰如其分地挖掘、利用好数据。

这四步完成，实际上我们就走完了从数据到价值的一次旅程：人工智能机器人通过对大量数据进行学习，了解顾客的需求，定向合理地推荐一些符合受众品位的化妆产品。对于顾客来说，他们能够及时获取最符合自己兴趣和心意的产品，满足了自身愉悦购物的需求；对于商家而言，顾客对感兴趣产品的咨询和光顾次数，就直接影响了顾客的购买率。

这是一个从数据出发，到价值为止的过程。但一次业绩的提升并不是终点，人工智能还会把这一次的数据反馈到模型里，从而提升下一轮的决策质量，也就是说，数据和人工智能会持续地为我们带来价值。所以对于像兰蔻这样的商家和企业来说，运用人工智能其实就是了解顾客需求的过程，并能够有效地提升服务。

到这儿我们总结一下，AI思维的核心竞争力是去平均化的能力。所谓的平均化思维，指的是无差别化，每个人、每个产品都同质化；而去平均化，则是肯定人与人、环境与环境的差异性，不同的人有不

同的想法和喜好，根据这些差异化需求产出差异化体验。我们可以通过数据和人工智能准确理解每个人的潜在动机，从而最优化决策，达到各自理想的目标。如能合理使用，无疑将带来更高的生产力和更合理的资源利用方式，产生新的价值。AI炼金术便能从数据中产生价值，同时它也是AI思维的具象表现形式。

通过人工智能决策来实现价值的源头是数据，所以，数据的广度和精确度决定了人工智能能够为人类做出多大的贡献。数据越多，越能更全面地反映个体的实际行为与想法，人工智能的预测就会越精准。如果赋予人类预测的能力的话，我想大多数人都会选择先预测一下自己的命运。所谓“知生死安天命”，随着人类身体数据的累积，人工智能对病人的死亡时间预测的准确率也有望更加精准。相关资料显示，科学家通过人工智能解读数据，利用200万份电子病历来进行人工智能深度神经网络的训练，全方位审查患者家族史、用药情况、各项体质指数、食物摄入等因素，进而能够预测机体的未来。目前，这一人工智能模型已经能够依据大量的数据提前12个月预测到患者的死亡时间，准确率达90%。研究团队还将这一预测结果与英国国家统计署的死亡记录、英国癌症注册记录等数据库的死亡数据进行了比对，之后又将这套人工智能模型与两项标准的预测方法进行比较，结果显示，这一新模型的准确率比之前的预测方法高了10.1%。

根据上面这些数据，科学家们不仅能够预测到很多原本无法定论的结果，而且还能了解到影响人生命期限的多种因素，从而使人们根据这些结果对自身生活习惯进行调整，达到延长生命的目的。例如，科学家在实验中对受试者的水果、蔬菜、肉类、奶酪、谷物、鱼类、酒精等食品摄入的量做了考查，找到最利于生命体健康的饮食摄入比例，在病人治疗期、康复期对其饮食结构进行调整，使其达到促进机体代谢健康的目的。这也是人工智能通过数据进行预测的价值所在。

新制度经济学的鼻祖、诺贝尔经济学奖获得者罗纳德·科斯曾说过：“如果你拷问数据到一定程度，它会坦白一切。”大数据包罗万

象，能够将许多看似并不相关的事件联系在一起，使我们能够更加清晰明了地把握事物发展的趋势，跟上瞬息万变的社会潮流。谁掌握了大数据，谁就掌握了主动权，不仅能够决策世事，更能够成全自己。但数据就像一片汪洋大海，望不到边际，我们自己很难游上岸；这时就需要人工智能做舟——长风破浪会有时，直挂云帆济沧海。

AI如何做出决策

前面我们讲了从“数据”到“价值”的历程，大数据的出现使人工智能的决策更高效更聪明。数据加人工智能，为我们在决策时提供了无数个在以前无法实现甚至是无法想象的可能性。那么人工智能是如何一步步地将“数据”转化为“决策”的呢？如图3-2所示，从“数据”产生“决策”的具体过程是这样的：首先要通过对历史数据的特征处理产生特征数据，然后通过机器学习产生预测模型，之后需要通过模型评估来产生回测效果。在回测中检验合格的模型就可以投入生产，生产过程中如果有新的个体数据输入，就会通过生产模型得到预测分数，根据预测分数的高低和其他因素进行业务优化，产生针对个体数据的业务决策。模型本身也要受到监控，监控的目的是产生质量报告，来确保模型正常运行。这一整套框架被称为“人工智能决策引擎”，接下来我们将对它进行详细了解。

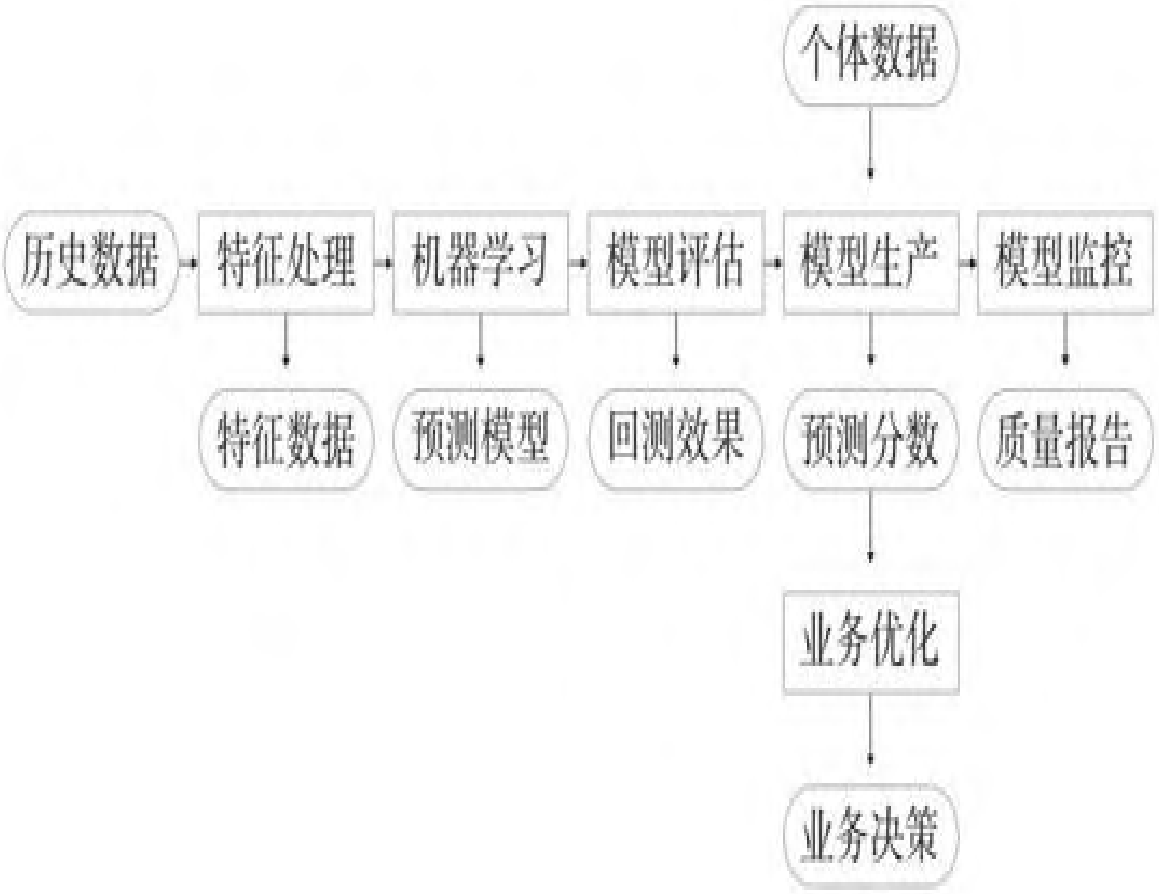


图3-2 人工智能决策引擎示意图

数据和特征处理

毋庸置疑，数据是当下这个时代很重要的一种资产。我们经常提到的大数据有“3V”：Volume，即数据量，以往数据量少，以MB、GB为单位，但是随着互联网的发展，数据不断积累，现在数据量已经发展到了以PB、EB、ZB为单位表示；Variety，即多样性，根据数据的结构是否统一，将数据分为结构化数据、半结构化数据以及非结构化数据；Velocity，即速度，指的是数据的产生速度越来越快。

那么数据是如何产生的呢？首先，社会在发展、人类在生活，就会有源源不断的新数据涌现出来。无论是机器生产还是市场营销活动，都沉淀了越来越多的数据。其次，随着互联网的发展，知识的权威性不断下降，任何人都可以成为信息的发布者，数据就这样不断地积累，数量越来越多。最后，随着学科的不断发展，人们描述同一类型事物的角度越来越多，例如随着基因科学的发展，现在用来描述基因的指标维度非常多，比如说，每个人的DNA（脱氧核糖核酸）分子是由约30亿个碱基对按精确的顺序排列而成的，一个碱基对就对应了一个维度，那么一个人的DNA分子就有几十亿个维度。像这样的例子其实还有很多，数据就这样不断积累到了一个非常庞大的数量。

通常，在各种各样的业务过程中形成的历史数据很杂乱，并且分散在不同的数据库和数据表里，不够系统化。为了更好地利用它们，我们需要把它们整合在一起，得到一套结构明确、条理清晰、便于理解的系统化数据。比如说现在需要处理一份在过去两年时间里形成的顾客在不同商家的消费行为数据，由于顾客消费行为的多变性以及商家的多样性，这份数据的数据量庞大且类型多样。而人工智能会对这些复杂的数据做特征处理，特征处理是通过对原始数据的处理和加工，将原始数据维度转换为特征数据的过程，特征是数据中所呈现出来的重要特性，通常是通过数据维度的计算、组合或转换得到的。经过特征处理，数据就被加工成了机器学习能够理解的数据形式。

过去，特征处理都是通过单台计算机进行，处理效率比较低。2005年，由阿帕奇（Apache）基金会所开发的哈杜普（Hadoop）可以将多台计算机连接在一起形成计算机集群，提供分布式并行的特征处理和分析功能，实现对复杂数据的快速、可靠的计算。由于其技术的高效性，对特征处理能力的提高，特征处理变得越来越容易，所以特征处理的门槛越来越低。到2008年左右，越来越多的公司意识到大数据时代分布式并行计算的重要性，同样地，这也为人工智能处理大数据提供了完善的基础设施。但特征处理的过程通常也是枯燥而繁重

的，看不到直接有用的结果。如果卡在这步，对企业来说，人工智能就无法落地。

机器学习和优化

做好了特征处理，下一步就是机器学习。机器学习，就是在特征上建立模型的过程，这里需要选择合适的预测模型，运用历史数据学习出模型。机器学习在各行各业都开始应用，例如在零售领域，通过发放优惠券来刺激老顾客、吸引新顾客是常用的营销手段，但是随机投放优惠券往往效果不明显，在某种程度上还会打扰到顾客，这样不仅造成了营销成本的浪费，还会降低品牌的声誉。为改善营销效果，一些零售商采用机器学习来进行营销。机器学习根据大量消费数据学习出预测模型，能够预测顾客是否需要某种类型的优惠券，以此来决定是否向该顾客投放优惠券。这样不但使真正需要的顾客享受到了优惠，还改善了商家自身的营销效果。

机器学习的过程，其实和人类通过经验总结生活规律再来指导我们之后的生活是相类似的。我们通过生活规律来指导的是以后遇到的事情，同样地，机器学习需要预测的也是将来。未雨绸缪，早为之所，商家可以根据机器预测提前做很多事情来促进顾客的消费，在最短的时间内抢占市场份额，扩大利润，也为自己争取更多时间，使自己更加灵活地应对一些突如其来的市场变化。如果一个人平时有吃零食的习惯，那他一定听说过良品铺子。从竞争激烈的零食行业中发展到现在会员数量达7 400万、2 000多家线下门店以及覆盖多个外卖渠道，良品铺子的崛起离不开人工智能的帮助。零食品类所面向的消费群体是非常大的，但零食又是一个随意购买性的商品，也就是说平时你可能想不起来买零食，但是如果身边刚好有一个零食货架，你可能会很随意地拿几样薯片和可乐去结账。

为了抢占如此大的消费市场，推动自己的产品覆盖更多的场景，良品铺子开始借助人工智能来帮助提高自身的业务预测能力。人工智能的任务就是对良品铺子的各种历史数据进行学习，然后将其多个渠道的会员数据打通，在这些数据的基础上了解用户需求，也就是通过机器学习建立预测模型。这样的模型可以提前预测哪些用户购买零食的可能性大、哪些用户最喜欢购买哪几种零食，在此基础上为用户提供更加符合个人喜好的个性化推送；推送渠道也会根据用户最常使用的渠道来确定，真正实现以用户为中心的精准营销。除此之外，人工智能还可以为前端业务创新提供支持。譬如说，如果跟踪用户购买情况的数据，显示某段时间内用户的购买热情有所下降，商家就可以制定秒杀、拼团或者发放优惠券等活动来刺激用户购买力。

机器学习产生模型还能用来检测接吻场景。这个功能的发展源自斯坦福大学数据科学家阿米尔·齐亚（Amir Ziai）的奇思妙想，他想让人工智能学习什么是接吻，所以他从好莱坞电影数据库中挑选了100部经典电影中的代表性片段，按内容将不同的电影片段标记为接吻或者非接吻片段，并通过机器学习分析了这些片段，特别是观察电影片段中人物的嘴唇动作，然后配合场景和电影背景音乐，学习什么是接吻，形成了专门的接吻识别模型。一般而言，接吻是影片中的高潮节点，推动了整个故事的发展，这个模型可以将影片中的接吻片段标记出来，为做视频剪辑工作的人员提供精彩片段的筛选过程。与接吻模型属于同类型的微表情识别模型和动作识别模型现在也展开了应用，比如在2019年，日本开发出一种基于人工智能的偷窃预警系统VaakEye，它经过10万小时的训练，能够识别出顾客是否将商品装入包中，或者是否出现可疑行为，并提醒超市工作人员注意防范。日本50多家便利店应用了VaakEye，其盗窃行为识别的准确率高达81%，将便利店的入店行窃损失减少了77%。

模型建好之后，仍然需要详细的评估和验证。模型评估，也就是回测模型效果的过程。回测指的是用历史数据来验证建立的预测模型是否有效。预测模型，顾名思义，是用来预测未来的、目前尚未发生

的事情。经过回测，评估合格的模型才有资格进入生产环节。这样做是因为只把模型构建完成是不够的，我们并不知道这个模型是否具有实践中的普适性，是否存在差错和不合理的地方，如果匆忙将模型投产，可能会为企业带来损失；片面地忽视回测结果，缺少后续推动，也会导致人工智能落地艰难。科学的回测既能保证人工智能为我们创造价值，又能为企业的每一个业务决策提供保障，避免损失。很多企业的领导不了解回测这种手段，也不重视过程，只关注最终的收益。事实上，他们忽略了科学的回测结果跟最终的经济结果之间本身就有着极强的相关性，是实践中非常重要的信号。通常，只要回测结果证明模型是合格的，最终的经济结果就很有可能令人满意。

只有模型评估合格，达到预期的效果以后，我们才能将它投入例行化生产，开始预测分数。比如说在一个借贷项目中，对许多参与者的信用程度进行预测，然后将预测的分数做个排序，通常来说，信用分数越高的用户获得贷款的可能性就越大。模型生产涉及大量工程化的工作，需要将软件开发工作系统化、模块化、规范化，从而保证高稳定性和可靠性，这个过程要在复杂的计算机和数据系统上进行，并与其他软件进行耦合，所以需要人工智能科学家和工程师紧密的配合。我见过一些大公司，人工智能科学家和工程师之间不能紧密配合，团队之间割裂，即便他们拥有一流的人工智能团队，人工智能也无法发挥价值。

下一步就是业务优化。业务优化，也就是从预测分数做出决策的过程。对个体行为准确的预测和把握，可以帮助行业决策者做出优秀的商业决策。在这个过程里，我们会通过优化整体的业务目标，比如说收益，把预测结果转化为具体的、可执行的业务决策。例如，Caliber Collison是一家美国的汽车修理公司，离职率曾高达40%。公司研究后发现，这一问题的部分原因在于其门店有时没有足够的车供员工修理，导致员工之间的工资差距较大。于是它开始从人力资源软件中提取出员工数据，建立定制的机器学习模型，预测员工是否会考虑离职。然后，Caliber通过该模型对员工离职率的预测分数进行业务

决策。例如，如果模型检测到一名员工的工资在几周内一直下降，那么他的预测分数就会相对高一些。为了尽量让员工留下来，Caliber的地区经理就会多分配给他一些汽车去修理，让他的工资有所上升。相反，如果模型检测到一名员工很忙，虽然他的工资很高，但是他感到很疲惫，那么他离职的预测分数仍然不会低，这时公司就可以把一些工作重新分配给他的同事。如此一来，公司的工作结构就更加均衡，Caliber的离职率也直线下降，每年在招聘上节省了百万美元。

除了帮企业做出优秀决策之外，业务优化还能引导用户做出选择。当你准备在购物网站搜索想要的商品时，人工智能推荐给每个用户的商品列表就是优化的结果，它们能够帮助用户更快地找到自己感兴趣的商家，也能为商家做更好的市场推广，这其中的经济价值是巨大的。eHub就是这样一个人工智能营销平台，它收集了来自营销、销售和客服的多渠道数据，通过这些数据建立预测模型。这些模型可以根据数据中反映出的用户属性特征和行为特征分析用户的喜好、需求，并进行分数预测。比如一个用户经常购买糖果、巧克力、奶油蛋糕，那么模型就会预测该用户购买甜食的分数高，然后给该用户推荐其他品牌的糖果、巧克力或者其他甜食，甚至是其他喜欢吃甜食的用户经常购买的商品，以此来实现一对一的个性化营销、精准营销。用户看到每天推送的都是自己喜欢的商品，购买概率也会大大上升。由于该平台十分符合自己的需求，用户在平时也会多浏览该平台的商品。

最后一个环节就是模型监控。模型监控是监控模型效果和质量的过程。由于数据源会变化，各种外部因素会干扰数据，我们很难确保一开始正确的模型一定能持续工作并有产出。因此，我们要对每一个环节都进行严格的质量把控。质量把控需要通过模型监控，持续地跟踪模型效果来实现，输出实时的质量报告，一旦出现问题，及时向相关人员报警，进行排查。很多人工智能项目在开始时没有设置模型监控，导致了机器学习结果出现错误。例如，巴西一家金融服务公司开发了一个聊天机器人模型，这个模型可以和求职者聊天并传达应聘信

息。聊天机器人模型会让求职者通过实时聊天和电话回答一系列问题，但是没有设置模型监控，所以在使用该聊天机器人模型时，出现了两个关键问题：一是发错了给求职者的个人资料和职业表格，二是安排的面试时间错误造成人力资源部门工作延误。由于没有对聊天机器人模型进行精确的调整，该公司判定所收集的数据中有大约10%是不正确的。该公司人员坦言：“我们犯了一个错误，以为一切都解决了，而没有去监督聊天机器人模型。教训是，一定至少要有几个月的时间全时监督聊天机器人模型。我们不能停止关注可能出现的偏差，以及新出现的情况——当我们开始这个项目时，这些都是出乎意料的。”正是因为这家金融服务公司在使用人工智能时忽略了模型监控环节，才导致后续出现问题。这也正验证了模型监控环节的重要性。

这就是从“数据”产生“决策”的全过程：第一，将历史数据经过特征化处理为机器理解的数据形式。第二，通过机器学习产生能够准确预测新样本的预测模型。第三，通过模型评估回测模型是否合格。第四，评估合格的模型就可以投入生产，生产过程中如果有新样本输入，就可以通过模型产生预测分数，再根据预测分数的高低和其他因素进行业务优化，产生针对个体数据的业务决策。第五，对模型本身进行监控，产生质量报告，时刻保证模型的正常运行。

人工智能助力精准推荐

我们都知道人工智能是能够大大解放人力、产生高效价值的新生事物，但是我们也经常听到有人说人工智能在企业落不了地。他们认为人工智能还未发展到能够真正为企业产出价值的阶段，又或者，人工智能只是一场众人追捧的闹剧，并不真的适合在企业中应用。但这正显示出他们对于人工智能不够了解。如果上述过程都做到了，如果数据充分，每个步骤都能正确地实现，是不可能没有效果的。例如在

接下来的这个案例中，人工智能不仅成功落地，还帮助企业提升了效益，帮助其在短短四年时间内，市值增长为原来的3倍多。

这个企业就是我之前任职的PayPal公司。人工智能在PayPal落地的整个过程是从历史数据开始的，比如说PayPal的历史交易数据，也就是记录哪个用户在哪个商家消费了多少金额的一套数据。这些数据分布在不同的系统中，我们首先把历史数据全部整理了出来，并且做了特征化处理，加工成机器学习能够理解的特征数据。

下一步就是机器学习，用来建立预测模型。机器学习需要预测的是将来，简单点来说，我们的人工智能需要预测的就是将来某个顾客最有可能在哪个商家购物。如果该顾客是体育爱好者，那么他将来去购物的店可能是耐克、阿迪达斯、安德玛，即使之前他没去过，这样的推荐仍然会对他有很大的吸引力。相反，如果给这个顾客推荐花店、玩具店之类的商店，那么可能就是缘木求鱼、南辕北辙了。通过机器学习，人工智能可以在多次的推荐后，通过用户的浏览时间以及购买情况得知推送的内容准不准确，符不符合用户的需求，通过用户在不同渠道的响应率得知用户最经常使用的渠道是哪个。经过多次这样的推荐和用户反馈之后，能够实现模型的持续学习，逐渐形成精准的推荐模型。

模型建好之后，需要对其进行回测，评估这个模型是否合格。那我们现在如何确定它是有效的呢？方法很简单，就是用过去的数据来模拟和验证。具体来说，如果我们用2018年一整年的数据来建立这个推荐模型，就可以用2019年的数据来“回测”，因为对于从2018年的数据学习出的模型来说，2019年的数据代表了未来。所以，如果模型在2019年的数据上效果良好，它就通过了回测。通过回测，我们就可以在模型投产落地之前知道构建的人工智能决策引擎的预测是否准确、准确率有多高，是否真的能为我们的企业创造出价值，防止引擎开始落地时才发现没有效果、成本全部打水漂的现象发生，为人工智能的顺利落地提供了保障。

预测模型建立出来并评估合格之后，就要把它投入生产，例行化地通过模型产生预测分数。PayPal这种类型的公司对生产的要求很高，对模型生产过程中的检验、部署、监控，都有相当严格的要求。完成了模型生产和监控后，我们就有了源源不断的预测分数，对应的是每一个用户将来在各个商家购物的概率。

当我们得知用户是否在各个商家购物的预测分数后，就要进行业务优化。针对一个用户，我们将这个用户可能会在各个商家购买的预测分数按照高低排序，形成这个用户的商家推荐列表，这样就能帮助用户更快地找到自己感兴趣的商家和商品。针对多个用户，我们可以将这些用户按照他们在某个商家购物的预测分数的高低排序，形成这个商家的用户推荐列表，帮助商家更快地发现哪些用户是营销转化率最高的用户，使得商家更好地打开市场。

这一整套人工智能决策引擎执行下来，并通过运营数据的反馈多次进行优化，最终可以帮助PayP al更好地进行业务决策。当然这不是终点，当人工智能决策引擎开始正式助力PayPal的业务运营之后，我们仍然需要对该模型进行监控，实时更新质量报告，这样才能及时发现可能出现的问题，及时抢修，避免出现更大的损失，同时保证人工智能为企业和用户提供持续而且优质的服务。

以上就是这套人工智能决策引擎在PayPal落地的过程。这套决策引擎落地后，PayPal用户开始积极响应PayPal的营销活动，其所带来的收益远远高于活动成本，而且产生了稳定的增益。几年来，这套人工智能决策引擎的顺利运作，为PayPal经营发展提供了有力的支持，直至2018年，PayPal市值突破千亿美元，人工智能决策引擎可以说是功不可没。

人工智能决策引擎是一套完备的企业决策引擎，它可以应用到企业决策的每一个环节，就像代表智慧的九连环，每一个环单独看都不起眼，但是所有环连在一起就成了智慧的象征，能够为企业创造出源源不断的价值。而且人工智能决策引擎环环相扣，每一个环节都必不

可少，不论哪一个环节出了问题，都会直接影响甚至是威胁到接下来的环节，之前做好的准备也很有可能功亏一篑。人工智能决策引擎是AI思维的重要框架，也显示出AI思维的缜密智慧：只有事无巨细地做好每一件事情，才能真正从数据中得出决策，人工智能才能成功落地。

业务优化方法

水往低处流，但人要往高处走。在工作中，我们兢兢业业，精益求精，希望每一次的绩效考核都可以拿第一；在学习中，我们尽心竭力，一丝不苟，希望经历过的每一场考试都可以名列前茅；在生活中，我们脚踏实地，善始善终，希望活过的日子都不会有遗憾。浑浑噩噩也是一生，但大多数人都不会放弃自己，所有不甘于现状的努力都出于对一个“最”字的追求。“最”字体现在人身上是追求极致人生的动力，体现在AI思维上就是通过人工智能找到一个让系统发挥最大效能的方案，找到解决问题的最佳决策。

何为最优化？

人生大方向的最优解是建立在日常小事情最优解的积累上的。选择哪一种材料的衣服，才能舒适又美观；选择哪一种食物，才能美味又营养；选择在哪里买房子，才能交通便利又价格适中；选择哪一条路线上班，才能快速地到达公司；选择哪一家公司，才能工资多又同事和睦。找到这些最优决策的方法叫作“最优化方法”。最优化，就是在一定的约束条件下，从众多可能的选择中找出最优选择，使系统的目标函数找到最大值或最小值。形象点来理解，就是你面前有无数

座连绵不断的高山，你需要在这些高山中找到海拔最高的那个山峰或者海拔最低的那个山谷。

不但个人需要找到人生的最优解，各行各业也都希望找到这个最优点，在不多浪费任何一种资源的条件下达到最佳的效果。在商业中，通过投资、收购等一系列操作，追求利润最大化；在体育运动中，通过完整而系统的训练，让运动员更高、更快、更强；在医疗行业中，通过各种治疗，让病人在最短的时间内摆脱病痛。那你肯定会有疑问：人工智能与最优化有什么关系呢？

首先，人工智能是为服务人类而生的，所以人类需要追求最优解，人工智能也需要追求最优解。其次，从本质上来看，人工智能的最终目标其实就是最优化：在数量庞大的数据以及瞬息万变的环境中，做出最优的决策。最优化贯穿了人工智能产生、发展和运用的全过程，从预测结果到业务决策，用的都是最优化方法，因而，最优化方法是人工智能最基础的知识。

人工智能的全过程其实都是在寻找最优解，通过机器学习，产生了对个体的预测分数，我们对未来的事件就有了基于数据的预判。就像我们在路上行车，有了清晰向前的视野，这对我们接下来产生的决策非常有帮助。在人工智能决策引擎中，首先要通过对历史数据的特征处理产生特征数据，然后通过机器学习产生预测模型，最后进行决策。那么，AI思维如何帮助我们产生最优化的业务决策呢？

如图3-3所示，从整体来看，优化过程共分为以下三个步骤：定义问题、优化模型、输出应用。人工智能是不能直接理解我们的要求的，所以我们要把业务的需求转化成人工智能可以理解的数学语言，这就是定义问题。



图3-3 优化过程示意图

定义问题包括以下三个要素：第一个要素是可用资源，就是我们现有的各种资源，比如要生产一种产品，我们有多少原材料，生产工人技术的熟练程度，还有生产机器的数量等。第二个要素是约束条件，比如说要想任务保质保量地完成，工人和机器都需要休息，所以约束条件就是工人每天只能工作8小时，机器连续运转16小时就需要散热和检修。第三个要素是业务目标：任何业务决策都需要一个目标，比如要生产多少产品，这个项目才算完成；这个营销计划执行完，要

增长多少用户量；或者这个季度结束，要完成多少订单。在数学上这个目标可以通过目标函数来定义，也就是f（项目）=产量，f（营销）=用户量。在制定这个目标函数的过程中，实际上作为具有AI思维的你，已经开始在宏观的层面思考业务问题了。

目标函数由预测分数和决策变量构成，决策变量是我们对个体的决策。比如说，我们花了大量的资金在网站上投放广告，但是广告效果并不是很理想，我们就会想去提高广告的点击量。这时，我们的目标就是最优化广告投放的总点击量，预测分数是针对每一个用户点击率的预估，决策变量则是是否对某个用户投放，从而做出对哪些用户投放广告的决策。显而易见，我们把所有的分数做个排序，选择高点击率的用户投放即可。这是个简单的目标优化问题。

优化模型决定了我们需要选用什么样的优化方法，而我们一般选用数学优化方法。总的来说，优化问题分为两种：有约束优化和无约束优化。顾名思义，有约束优化是在一系列约束条件下，寻找一组决策变量，使目标函数值达到最优；无约束优化与有约束优化相对应，指的是没有任何约束条件的优化。现实中的优化问题通常是有约束的，因为我们做一件事的资源总是有限的，比如一个企业要在一个周之内做完一个项目，一周就是时间约束，也就是我们的时间资源是有限的。或者是这个企业要用3 000元来办一场活动，这3 000元就是预算约束，也就是我们的成本和预算是有限的。

输出应用就是输出我们的最优解，是优化方法的最终目标。比如我们花最少的时间到达了目的地，用最低价格购买了一幢房子，用最高的效率完成了生产任务或者以最大的产量完成了绩效目标。

其实我们在很久之前就接触过这种最优化的思维。比如语文课本上“田忌赛马”的故事讲的就是最优化的思维，四大名著之一的《三国演义》就是军事最优化理论的真实写照；我们从学生时期就开始做求最大利润、最快速度、最少成本这样的极值数学题，其实都是在训练我们的最优化思维。在AI思维中，最优化思维十分重要，然后才是

对问题的求解。我们做过很多这种求极值的数学题，不知不觉中就积累了最优化解决问题的能力，不过我们在现实生活中遇到的问题要比数学题复杂得多。情况是错综复杂的，但我们只要有了这种最优化的思维，能够把在生活中遇到的各种问题和现象翻译成人工智能可以理解的数学语言，建立适当的数学模型，这些问题就可以迎刃而解了。下面我们就来看一下在现实生活中，优化思维是如何解决业务难题，为企业带来商业价值的。

优化的业务价值

我们在讲优化的第三步输出应用时讲到，优化方法的最终目标有：最高效率、最大产量、最少时间和最低成本等。这其实就是优化所带来的业务价值的一个缩影。试问哪一个公司不想用最低的成本、最少的时间来获得最高的产量呢？但是这在现实生活中基本上是不能兼得的，用最少的时间获得最高的产量，就意味着可能要付出更高的成本。优化其实就是在这些冲突之间帮助企业找到一个平衡点，在现有资源的范围内，保证企业最大限度地赢利，还能为用户提供最优质的服务，提高用户的满意程度。

更详细地讲，优化带来的业务价值有如下几条：第一，优化可以帮助企业降低生产成本，提高企业运作效率，缩短生产和服务的交付周期。德国一家汽车生产公司有两条汽车生产线，各负责生产一种型号的汽车，这两条生产线每年能够生产20万辆汽车，这已经是一个很高的产量了。这家公司仍然不满足，想在不开通第三条生产线的情况下生产第三种型号的汽车。为了实现这个目标，它自主搭建了一个流程优化系统，在没有花费大量资金去建立新生产线的条件下，将生产力提高了30%，将生产周期从7天缩减到了4天。由此，该公司节约了5亿美元的成本。这种能够看见的成本优化，提高了工作效率，增加了产量，在日后能够带来的经济效益是不可估量的。

第二，优化可以帮助企业提高资源利用率，向客户提供更加优质、灵活且准确的服务。法国一家快递公司之前送货速度没有比别的快递公司快，快递费也因为成本问题不能降低，相比之下没有任何竞争优势，迟迟不见盈利。不过，这家快递公司有一个特点——它依附于一家航空公司，因此，综合考虑公司可以调动的资源，该快递公司决定利用航空资源来优化自己的快递业务，这样快递的送货速度就有了保证。但同时会出现一些其他优化问题，比如飞机搭载乘客之外，运输快递的数量是有限制的，飞机的路线和快递路线不能完全匹配，飞机加入后每个快递集散中心工作量上升等。但该公司的优化模型通过充分利用现有航空资源，合理安排飞机路线等方法，不但节约了增添航空设备的成本，还提高了送快递的速度，在提高公司本身竞争优势的同时，为用户提供了快速优质的物流服务。

第三，优化可以帮助企业实现精准营销，提升销量，扩大产品和企业的知名度和影响力。在广告营销领域，需要解决的问题有很多。首先，这个广告要投放给谁的问题。通过最优化思维，我们可以找到投放成本最低、需求最高、转化率最高的那一批用户，优化了广告的响应率。其次，通过什么渠道去进行推广营销的问题。通过最优化思维，我们可以找到用户打开频率最高、响应率最高的渠道，比如某用户平时使用淘宝的频率比京东高，那广告当然要投放在淘宝平台，这样可以优化广告的曝光率。最后，问题在于，要向用户推送什么样的广告。通过最优化思维，我们可以发现每个用户不同的需求，商家的广告也需要根据需求来定制，这样用户收到的广告推送就会千人千面，更符合他们个性化的需求，用户的购买欲望也会相应地提升，在一定程度上优化了产品的销量。

第四，优化可以帮助企业减少库存，降低管理风险。日本一家家具企业经常大量购进优质木材，因而经常会出现所有订单都完成了还剩下很多木材的情况。这些木材存放的时间久了会出现虫蛀、受潮、色泽不均等问题，无法再用于做新家具，这样不但产生了大量的仓储成本，还浪费了大量的资源。为解决这个问题，该公司引进了一个优

化系统，对从木材采购到木材保管的全过程都进行了优化。首先，在采购之前，系统会根据优化模型算出一个最佳的木材采购量，这样既避免了过度购买木材，也减少了采买过程中的贪腐行为；其次，该企业还会对剩下的木材进行防蛀防潮湿处理，避免了资源浪费。仓储量的下降减少了工作人员的工作量，降低了管理风险。

业务优化应用案例

优化能够带来如此大的业务价值，那它在实际生活中是如何实践的？在实际业务运作中是如何发挥作用的？现在我们就通过具体案例来说明一下优化是如何执行的。比如说，现在有一个电商网站要进行一次营销活动，该网站的用户量巨大，如果这次营销活动要针对每一个用户，营销成本就会非常大，所以这个网站打算只选择部分顾客进行营销。对这个网站而言，每个用户的营销成本是不一样的，我们需要根据已有顾客的预测转化率，最大化选择的顾客的总分数。这其中的约束条件是：营销成本不能超过总预算。

假设我们有5个顾客，每个顾客的营销成本分别是c_1，c_2，…c_5（已知），转化率分别是p_1，p_2，… p_5（已算出），x_1，x_2，…x_5表示我们是否对一个顾客营销。那么优化问题可以这么写：

最大转化率：$p_1x_1+p_2x_2+\cdots+p_{10}x_{10}$

约束条件：$c_1x_1+c_2x_2+\cdots+c_{10}x_{10}\leqslant$预算

x_1，$x_2\cdots x_{10}$取值0或1（0代表不进行营销，1代表进行营销）

若5个顾客的营销预算为1 500元，营销成本和转化率如表3-1所示：

表3-1 营销成本和转化率

用户	1	2	3	4	5
营销成本（c）	100	300	500	700	1000
营销转化率（p）	27%	68%	76%	83%	95%

则优化问题应该这样写：

最大转化率：$0.27x_1+0.68x_2+0.76x_3+0.83x_4+0.95x_5$

约束条件：$100x_1+300x_2+500x_3+700x_4+1000x_5 \leqslant 1500$

在进行顾客营销时，我们当然会选择那些转化率高的用户来进行营销，但是我们同样也要考虑到营销的成本。在这5个用户当中，第五个的预测转化率最高，但是要对他进行营销的成本是1 000元，成本太高，如果都是这样的用户，那我们很有可能会超出预算，而且还没有达到预期的营销效果，这样就得不偿失了。第一个的营销成本最低，但是转化率只有27%，仍然达不到想要的营销效果。相比之下，中间3个用户的营销成本适中，营销效果也很好，性价比最高，而且成本在预算之内，所以最后我们选择中间3个进行营销，达到效果最优化。

通过上述预测分析和优化模型，我们就可以看出哪些是营销成本低但营销效果不好的用户，哪些是营销效果好但营销成本高的用户，而我们最终选择的肯定是那些转化率高且成本适中的用户，对这些用户进行营销可以控制成本在营销预算之内，又能达到很高的转化率。

这是最优化思维最初的目标：以最低的成本达到最佳的效果。同时，这也是最优化思维最终的结果：帮助企业实现了最佳的营销转化率。

最优化思维在诸如航空、物流、生产制造、金融、资源管理、环境保护、电力管理等领域中得到了广泛应用，在相关领域都已经有成功的优化实施案例。以物流行业为例，我们来看一下最优化思维如何在实际场景中展开应用。

近几年电子商务迅速崛起，在网上购物已经成了大众生活的常态，随之而来的是物流行业的蓬勃发展。国家发改委统计数据显示，2018年中国物流成本占国内生产总值的比重高达14.5%，而相比之下，欧美国家的物流成本占其国民生产总值的10%都不到。为切实有效地解决国内物流成本的问题，最优化思维势在必行。

在一切物流问题中，最重要的莫过于车辆路线问题。在面对数量众多且需求不同的客户时，组织多少辆货车最合适，走哪一条路线路程最短，选择哪一种物流方式成本最低、耗时最短等一系列问题，都属于车辆路线问题，也都属于最优化思维应当解决的问题范围。

现在许多地图服务商都提议提供最短路线或者最短时间方案，但是在实际运输过程中，要真正实现最短路线、最快时间以及最低成本，还需要解决许多复杂问题，例如不同货物配送的先后顺序，水、陆、空三种运输方式和多家快递公司的不同组合搭配，等等。此外，考虑到诸如限行、修路、恶劣天气或者交通事故等突发状况，物流所要优化的问题，复杂度直线上升。

为了优化解决这些问题，市场上出现了一些优化模型，例如针对物流优化的小马驾驾。首先，面对复杂且数量巨大的物流信息，优化模型会先对其进行学习，并在训练和学习中不断优化自身，让自己的判断越来越准确；其次，为了应对突发状况，优化模型可进行实时判断，做出路线调整，找到物流的最优路线；为了优化目前物流领域存在的业务模式多变、效率低下、资源利用不合理等问题，优化模型提出了多种提货送货模式，比如多地点提货单地点送货、单地点提货多

地点送货、多地点提货多地点送货等，大大提高了提货送货的效率，节约了时间成本。

在选择最短路线方面，优化模型将所有订单综合考虑，整合所有路线，将所有目的地都串联在一起，使得同一个目的地的商品能够一起派送。除此之外，优化模型还会考虑到回程问题，避免出现货车空仓回程的情况。这样既解决了回程问题，又最大可能地降低了运输成本，提高了物流效率。在配送服务方面，模型也可以根据用户需求灵活调整车辆、车型、仓库等，达到优化的效果。在控制成本方面，除了缩短物流时间可以降低成本之外，选择不同的车型或者运送方式也可以很好地降低成本，比如短途运送小型货物可以选择小型货车，成本相对较低；但是长途运送生鲜货物，还是应该选择航空运输，航空运输虽然成本高，但是生鲜商品的配送费本身也很贵，航空运输速度快，能够保证商品的新鲜程度，提高用户对物流的满意程度，这样才能留住客源。据统计，使用该优化物流模型制定的优化物流方案，可以将每日全城的调度时间从3小时减少到15分钟，企业的运送成本可以降低10%到20%。

面对海量的信息数据、复杂的环境、多变的市场和需求，我们在日常工作中，必然会求助于人工智能。如果企业不拥抱人工智能，就很难跟上时代快速发展的步伐，甚至会面临被淘汰的危险。我们需要人工智能为我们提供业务决策，其实是需要它为我们提供最优的那个决策。可以说，最优化是AI思维解决问题的出发点。有了AI思维的帮助，我们就可以更加轻松准确地在生活、工作和业务运作过程中做出最佳的决策。

会思考的人工智能

《终结者》里的天网是一套人工智能防御系统，但它却拥有了自我意识，想要毁灭人类；《我，机器人》中具有自我思考能力的人工智能机器人VIKI在负责生产机器人的同时，还在默默计划让机器人成为人类世界的主宰；《魔种》中的人工智能系统Proteus IV不但具有了自我意识，还将人类囚禁起来。现在越来越多的科幻电影里出现了人工智能的身影，而且它们都不约而同地拥有了独立思考的能力。其实从人工智能出现以来，人们一直都有这样的疑问：人工智能是否能够和人类一样拥有独立思考的能力？

关于这个疑问，被称为计算机科学之父的艾伦·图灵在1952年曾经进行过一个测试：将一个人和一台计算机分别关在两个房间内，再找一个人对房间里的人和计算机分别进行提问。提问者事先并不知道哪个房间里的是人，哪个房间里的是计算机，他要通过二者对问题的回答进行判断。如果有超过30%的人不能准确判断出哪一个是人，哪一个是计算机，那么这台计算机就通过了此次测试，而且可以被认为是具有了人的智能。这一用来判断机器是否具有人类智能的测试后来被称为图灵测试。一直没有计算机能通过图灵测试，直到出现了深度学习，人们才看到了一丝曙光。

深度学习的前世今生

我们知道，机器学习是人工智能的分支，它专门研究计算机如何模拟和实现人类的学习行为。在人工智能发展过程中，机器学习占据核心地位。通过各种模型，机器学习可以从海量的数据中学习出规律，从而对新的数据做出智能识别或者预测，并且为决策提供支持。深度学习是机器学习的一种。如图3-4所示，人工智能是一个范围很大的概念，其中包括了机器学习。机器学习是人工智能提升性能的重要途径，而深度学习又是机器学习的重要组成部分。深度学习解决了许

多复杂的识别和预测难题，使机器学习向前迈进了一大步，推动了人工智能的蓬勃发展。那么深度学习又是如何发展起来的呢？

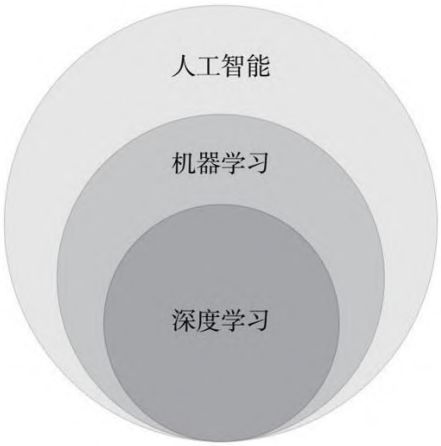


图3-4 深度学习关系示意图

深度学习的概念最初起源于人工神经网络（Artifi cial Neural Networks）。科学家发现人的大脑中含有大约1 000亿个神经元，大脑平时所进行的思考、记忆等工作，都是依靠神经元彼此之间的连接而形成的神经网络来进行的。人工神经网络是一种模仿人类神经网络来进行信息处理的模型，它具有自主学习和自适应的能力。

1943年，数学家皮茨（Pitts）和麦卡洛克（McCulloch）建立了第一个神经网络模型M-P模型，能够进行逻辑运算，为神经网络的发展奠定了基础。生物神经元一共由四个部分组成：细胞体、树突、轴突和突触，M-P模型其实是对生物神经元结构的一个模仿，如图3-5所示，左边是生物神经元的示意图，右边是M-P模型的示意图，为了建模更加方便简单，M-P模型将神经元中的树突、细胞体等接收到的信号都看作输入值x，突触发出的信号视作输出值y。1958年，计算机科学家罗森布拉特（Rosenblatt）发明了感知机，分为三个部分：输入层、输出层和隐含层。感知机能够进行一些简单的模式识别和联想记忆，是人工神经网络的一大突破，但这个感知机存在一个问题，就是它无法对复杂的函数进行预测。20世纪80年代，人工智能科学家鲁姆尔哈

特（Rumelhart）、威廉姆斯（Williams）、辛顿（Hinton）、莱库（LeCun）等人发明的多层感知机解决了这个问题，推动了人工神经网络的进一步发展。20世纪90年代，诺贝尔奖获得者埃德尔曼（Edelman）提出Darwinism模型，并建立了一种神经网络系统理论，对90年代神经网络的发展具有重大意义。

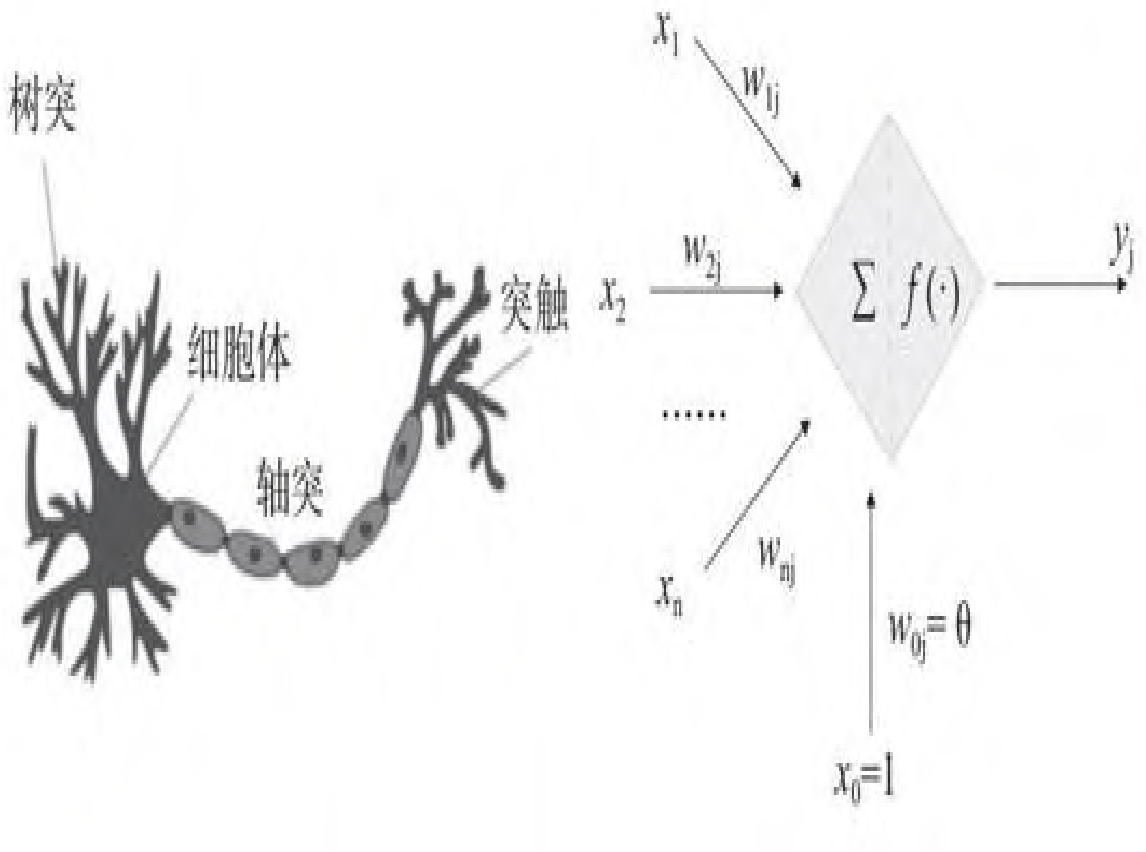


图3-5 神经元及M-P模型示意图

从这之后，神经网络技术再也没有出现过突破性的发展。直到2006年，被称为人工智能教父的辛顿正式提出了深度学习的概念，认为通过无监督学习和有监督学习相结合的方式，可以对现有的模型进行优化。这一观点的提出在人工智能领域引起了很大反响，许多像斯坦福大学这样的著名高校的学者们纷纷开始研究深度学习。2006年被称为“深度学习元年”，深度学习从这一年开始迎来了一个爆发式的

发展。2009年，深度学习应用于语音识别领域。2012年，深度学习模型AlexNet在ImageNet图像识别大赛中拔得头筹，深度学习开始被视为神经网络的代名词。同样是在这一年，人工智能领域权威学者吴恩达教授开发的深度神经网络将图像识别的错误率从26%降低到了15%，这是人工智能在图像识别领域的一大进步。2014年，脸书开发的深度学习项目DeepFace在识别人脸方面的准确率达到了97%以上。2016年，基于深度学习的AlphaGo在围棋比赛中战胜了韩国顶尖棋手李世石，在世界范围内引起轰动，这一事件不但使深度学习得到了认可，人工智能也因此被社会大众所熟知。2017年，深度学习开始在各个领域展开应用，如医疗影像、金融风控、课堂教学等，在不知不觉中已经渗透到我们的生活中。

深度学习的经典模型

那么深度学习到底是什么呢？深度学习是建立在计算机神经网络理论和机器学习理论上的科学，它使用建立在复杂的网络结构上的多处理层，结合非线性转换方法，对复杂数据模型进行抽象，能够很好地识别图像、声音和文本。深度学习有两种经典模型：CNN和RNN。

CNN全称是Convolutional Neural Network，也就是卷积神经网络。对于卷积神经网络的研究最早出现于20世纪80至90年代，到了21世纪，随着科学家们对深度学习的深入研究，卷积神经网络也得到了飞速的发展，该网络经常用于图像识别领域。如图3-6所示，卷积神经网络共分为以下几个层级部分，输入层（input layer）、卷积层（convolution layer）、池化层（pooling layer）、全连接层（fully connected layer）。

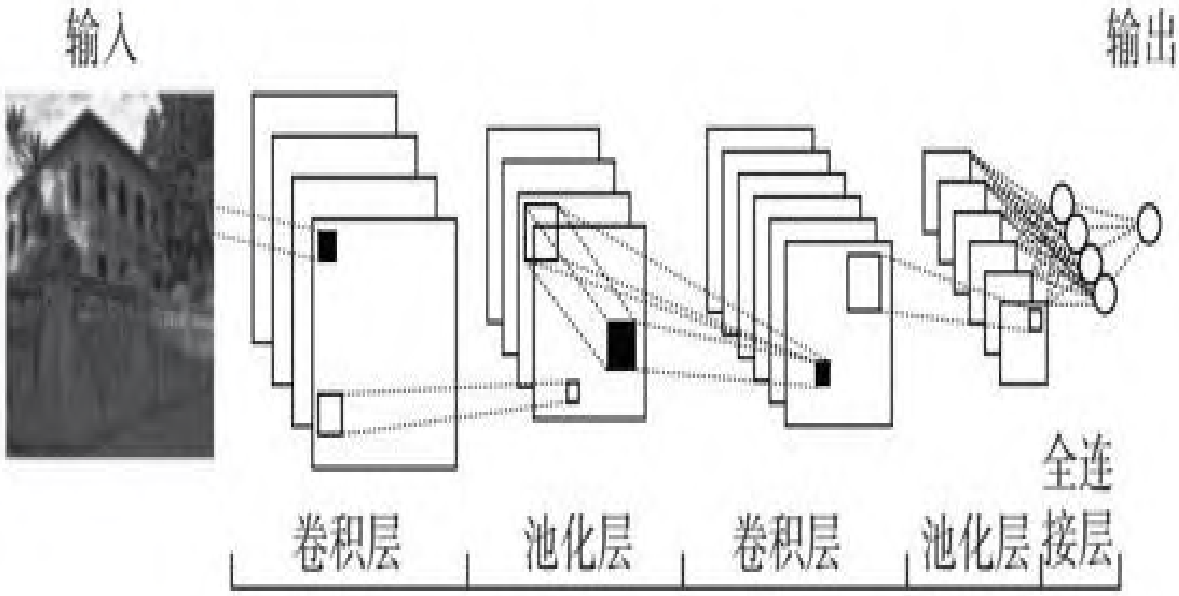


图3-6 卷积神经网络工作过程示意图

当图像进入输入层，模型会对这个图像进行一些简单的预处理，比如降低图像维度，便于图像识别。卷积层里的神经元会对图像进行各个维度的特征提取。这一提取动作不是针对原图像进行的，而是仅对图像的局部进行特征提取，比如需要识别的是一张包含小狗的照片，但是神经元只负责处理这张照片中的一小部分，比如狗的耳朵、眼睛。卷积层对图像不同尺度进行特征提取，大大丰富了获取特征的维度，有助于提升最终识别的准确度。池化层对图像进行压缩降维，降低图像识别需要处理的数据量。全连接层需要做的就是将前面所提取出来的所有图像特征连接组合起来，如图3-7中，将提取到的小狗的头、身体、腿等局部特征组合起来，形成一个完整的包含小狗的特征向量，然后识别出类别，这就是卷积神经网络进行图像识别的全过程。



图3-7 卷积神经网络图像识别过程示意图

通过对卷积神经网络工作过程的梳理，我们总结出卷积神经网络的三个特性：第一，图像识别不需要识别图像的全部，每个神经元只需要聚焦到图像的一小部分，识别的难度降低；第二，卷积层对应的神经元可以应用于不同的图像识别任务，比如图3-7中的神经元，经过训练，已经能够识别出小狗，那这些神经元也可以应用于识别其他任何图像中的相似物体；第三，虽然图像特征的维度降低了，但是由于

保留了图像的主要特征，所以并不影响图像识别，反而降低了识别图像需要处理的数据量。这三个特性决定了卷积神经网络非常适合用于图像识别。例如由牛津大学开发的VGG模型就是基于卷积神经网络模型建立的，它在识别物体的候选框生成、图像的定位与检索等方面十分准确，这使得它在2014年ImageNet竞赛定位任务中获得了第一名。

人工神经网络和卷积神经网络在深度学习领域都占有一席之地，但它们识别的都是独立的事件，比如卷积神经网络非常擅长识别独立的图像，如果让它识别100张照片，输出的结果互相不受任何影响，但是让它识别或者预测一句连续的话，比如一个寓言故事或者翻译一段英文，可能就没有这么好的效果了。可是在现实生活中，我们会遇到很多连续的事件，比如“小明每次去超市都会买很多苹果，因为他最喜欢吃（ ）”，联系上下文，我们可以很容易推测出括号里应该是“苹果”这个词，因为括号前的“吃”字是一个动词，动词后面经常跟着的是名词，而这个句子中的名词只有“苹果”最合适。为了能够识别这些连续性很强的事件，弥补人工神经网络和卷积神经网络的不足，RNN模型诞生了。

RNN全称是Recurrent Neural Network，也就是循环神经网络。对于循环神经网络的研究最早出现于20世纪80年代，由约翰·霍普菲尔德（John Hopfi eld）、迈克尔·乔丹（Michael Jordan）以及杰弗里·埃尔曼（Jeffrey Elman）等人提出，该模型经常用于时序信号（如语音）的识别和理解。

循环就是重复的意思，循环神经网络模型在运行时会对同一个序列进行循环操作。序列是被排成一列的对象，序列中的元素相互依赖，排列顺序非常重要，比如时序数据、对话等，一旦顺序错乱，含义和作用都会发生巨大改变。循环神经网络解决了卷积神经网络不能很好地识别连续性事件的问题，在深度学习领域发挥着不可替代的作用。

循环神经网络之所以能对连续性事件进行识别，是因为它不仅将当前的输入数据作为网络输入，还将之前感知到的一并作为输入。根据记忆的长短，从第一层开始，将激活传递到下一层，以此类推，最后得到输出。图3-8就是一个循环神经网络的示意图，它由输入层、隐藏层和输出层三部分组成。循环就发生在隐藏层。隐藏层里一般会设置一个特定的预测函数，当我们向循环神经网络模型输入一个连续性事件后，在隐藏层的这个函数就会进行运算，这个运算结果又可以作为输入进入隐藏层再一次进行运算。如此这般，就形成了一个不断循环的预测，这个预测与新输入的数据有关，也取决于每一次循环的输入。

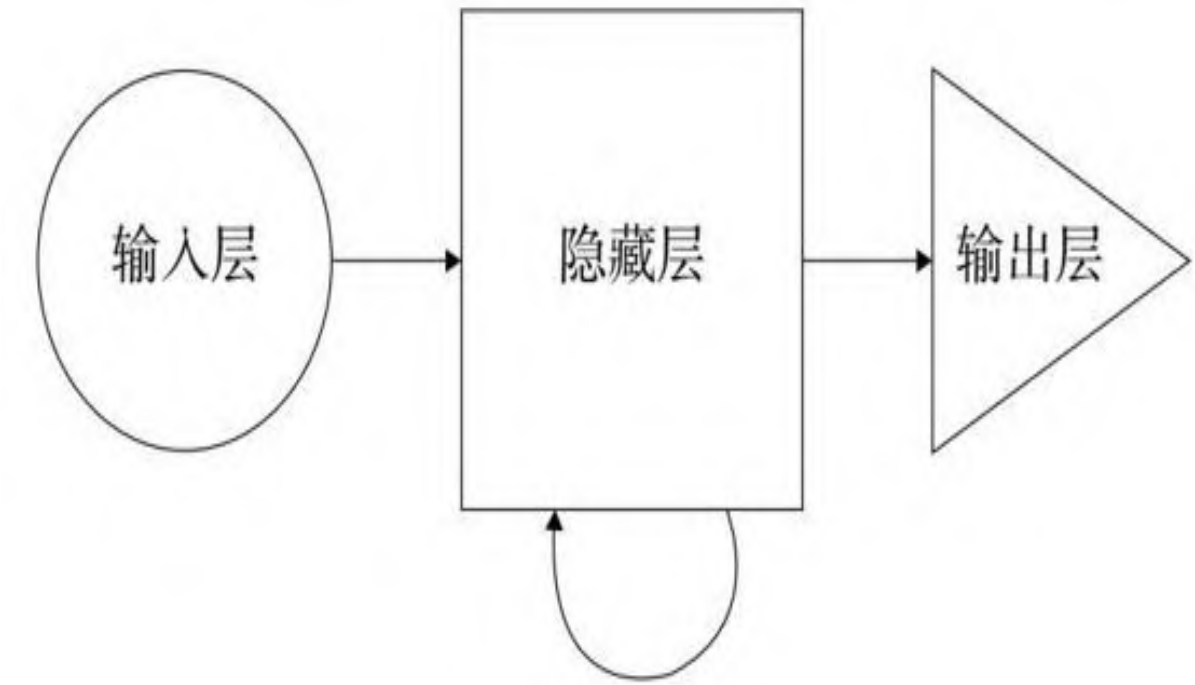


图3-8 循环神经网络原理示意图

连续性数据在日常生活中出现的频率之高决定了循环神经网络有着广泛的应用空间。例如，我们可以依靠循环神经网络预测一句话中的下一个词语或一篇文章中下一句话是什么，以此来生成文本，现在

很多写稿机器人就是基于循环神经网络来进行运作的。循环神经网络模型还可以将文本翻译成其他的语言，所以它也广泛用于机器翻译。循环神经网络另一个常见的应用是语音识别，我们现在使用的很多智能语音助手都应用了循环神经网络。

股票的价格波动也存在一定的规律，根据前一段时间的股票波动情况可以大致预测出之后股票的走势。因此，循环神经网络在股票预测方面有先天的优势。随着经济的发展，股票市场的规模不断扩大，大量股票历史数据的积累使得循环神经网络可以学习股票的走势规律。比如，当循环神经网络发现，某只股票价格不断下跌超过七天，之后就会缓慢上升，并且在很长一段时间内这只股票都呈现出这个规律时，如果这只股票价格再一次持续下跌，那么下跌的第七天就是股民买入的最好时机。实践证明，循环神经网络对股票的预测能够较好地拟合真实数据，具有很高的应用价值。

循环神经网络可以有效地进行文本识别，而在商业世界中，最重要的一种文本数据就是商品评论。随着生活节奏的加快、电子商务的兴起，人们越来越倾向于网络购物。网购时，用户没有见过真实的商品，只能通过商家的描述进行了解，而商家的描述又不能保证用户了解到的信息完全属实，此时，用户的商品评论就成为反映商品质量和商家信用的一个重要参考标准。但是，用户的评论也有很大的主观性，如何结合用户的主观评价正确评估商品质量及商家信用成为一个亟须解决的问题。在循环神经网络强大的文本识别功能的帮助下，我们可以很好地解决这个问题。循环神经网络在分析评论的过程中，最重要的一个步骤是对用户的主观评价进行处理，即通过循环神经网络分析用户的商品评论，再将其转化为对商家的等级评价。比如，循环神经网络识别出不同的商家同时在售卖同一种商品，但在商品质量方面，商家甲好评数远远高于商家乙，那么在这一方面，商家甲的等级评价就会高于商家乙的等级评价。影响商家等级评价的因素还有很多，比如服务态度、发货速度，以及商品与描述相符度等，将这些因素全部考虑在内，就会形成一个全面的商家等级评价。循环神经网络

在商家评价方面的应用使用户不会被海量商品信息以及其他用户的主观评价所混淆，直接找到符合自身需求并且质量上乘的商品，提升了用户的网上购物体验。

深度学习商业网络

数据是商业发展的基础，也是商业竞争的关键。现在越来越多的公司都在通过数据的深度分析来了解用户的需求，把握市场的发展方向，提高自己的竞争力，以此达到在激烈的商业竞争中获胜的最终目标。深度学习作为人工智能的关键能力，其对数据的理解能力也超过了之前的相关模型。对数据的深度理解能够挖掘出数据背后隐含的知识，能更加快速准确地掌握用户需求，所以深度学习在商业场景中具有非常重要的应用价值。下面我们来介绍两个在商业领域应用的深度学习网络：潜在客户转化率深度学习模型和客户购买力深度学习模型。

如图3-9所示，这是专门用于房地产行业的潜在客户转化率的预测模型。这个模型可以通过线上、线下数据以及在App上对不同房产的浏览数据等，推测出一个潜在的购房客户转化为真正的购房客户的概率有多大。



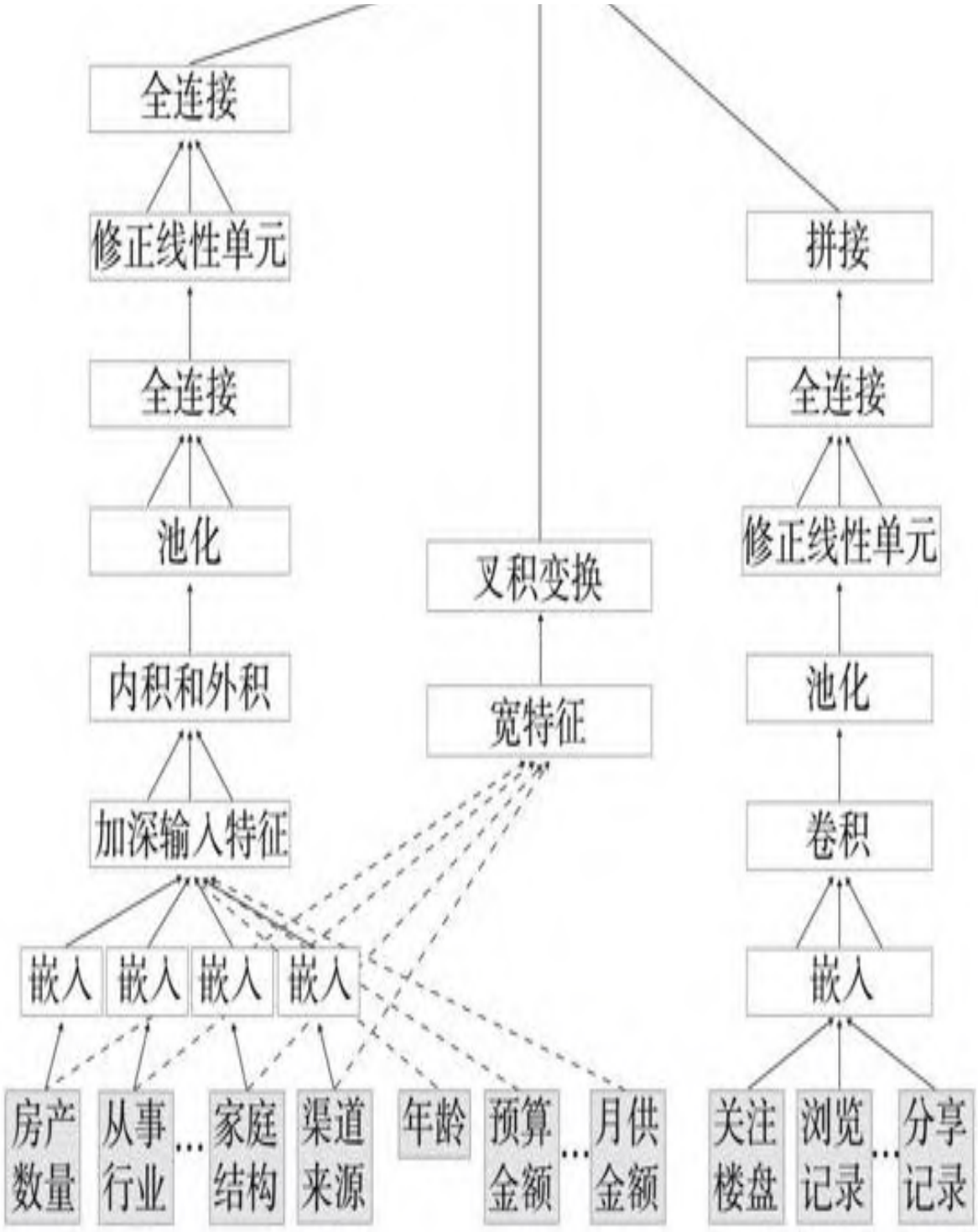


图3-9 潜在客户转化率深度学习模型示意图

在这个模型中，输入的数据包括：类别特征数据，如客户的房产数量、从事行业、家庭结构、渠道来源等；连续特征数据，如客户的年龄、预算金额、月供金额等；序列特征数据，如客户的关注楼盘、浏览记录、分享记录等。我们对不同分类的数据需要进行不同的操作，例如类别特征数据需要经过嵌入、池化以及全连接等操作，序列特征数据需要经过嵌入、卷积后才能进行池化以及全连接等操作。当这些操作都进行完之后，数据就可以进入拼接层，然后输出每个客户在接下来的一段时间内买房的概率。

如图3-10所示，这是一个专门用于预测房地产公司客户购买力的深度学习模型。这个模型可以通过一系列复杂的深度学习，最终预测出一个客户是高价值的买房客户还是相对低价值的买房客户。

在这个模型中，输入的数据包括客户的基本信息，如职业、收入、偿还能力、来源、意向、家庭结构、购房用途等；客户行为数据，如关注、点击、浏览、分享等；房屋描述信息，如建造年份、户型、位置、交通、物业等；还需要一些房屋、小区等周边环境的图像信息。不同类别的数据也需要进行不同的操作，例如客户基本信息中的类别特征数据需要进行嵌入操作，将稀疏的数据变成连续向量；客户基本信息中的连续特征数据就可以直接进行处理；客户行为数据要经过嵌入、卷积、池化以及全连接等一系列的操作；房屋描述信息数据要经过词向量化、嵌入、LSTM（长短期记忆网络）、堆叠特征处理。其中，LSTM适用于处理和预测时间序列中，间隔和延迟相对较长的重要事件，堆叠特征处理可以将不同的特征拼接在一起。图像数据要经过图像预处理、基于预训练的VGG模型以及全连接操作。当这些操作都完成之后，所有的数据要一起进行决策树处理。最后，我们就可以看出这个客户到底是高购买力客户还是低购买力客户了。

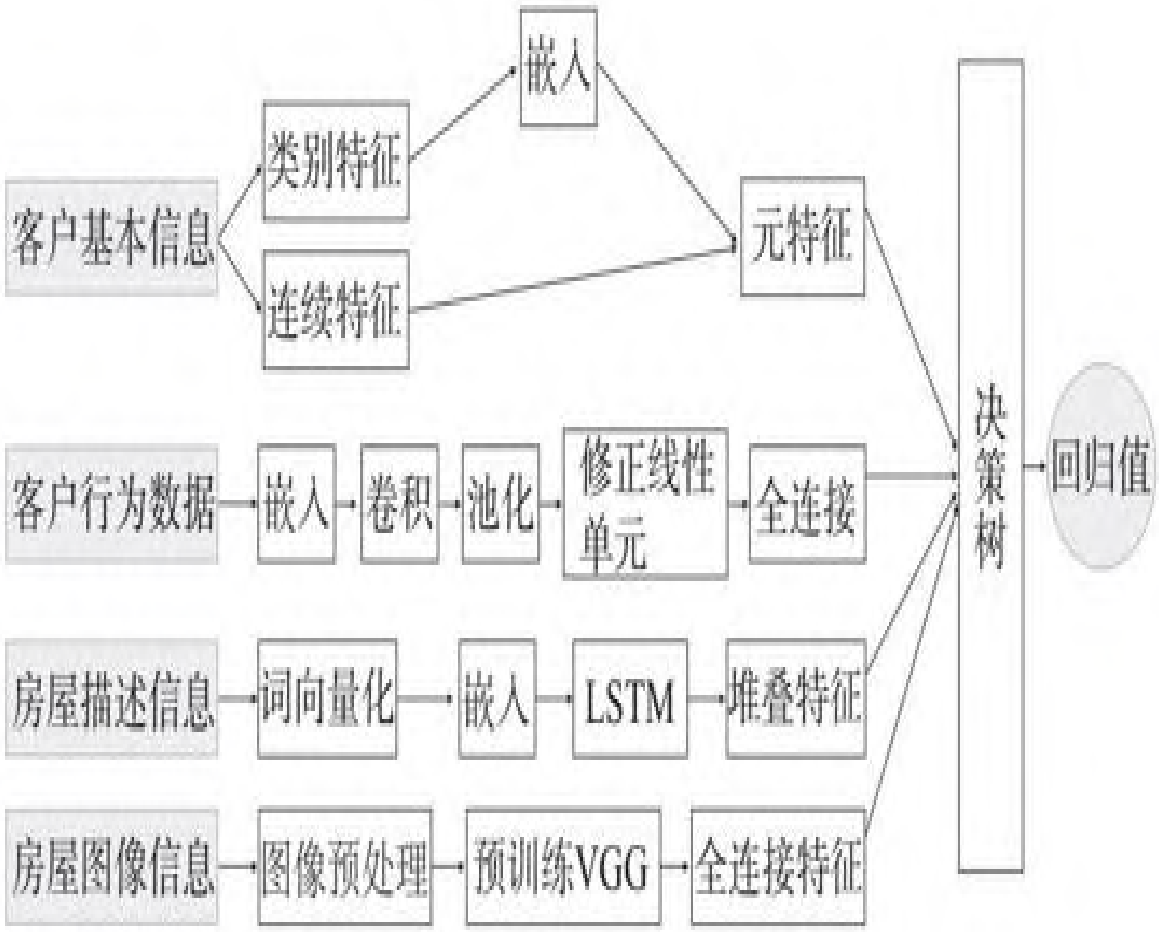


图3-10 客户购买力深度学习模型示意图

深度学习的出现为人工智能带来的是一种全新的思维方式和思考方式，它赋予了人工智能新的活力，为人工智能带来了无限的可能性。万物皆可智能正是因为人工智能有了深度学习的支持，是深度学习为人工智能插上了飞翔的翅膀。更有人提出，深度学习甚至决定了人工智能的未来，所以现在深度学习越来越成为机器学习和人工智能不可或缺的一部分。以小博大，见微知著，AI思维强调要抓住事物的本质特征才能做出最好的决策。掌握了深度学习，就算是掌握了人工智能的核心，所以，深度学习是我们构建AI思维不可或缺的重要一环。

AI炼金术的应用生态

中世纪的时候，人们十分相信占星术，认为天上的星体可以带来幸运；并且将地上的一些贵重金属视作天体的代表，比如黄金代表金黄色的太阳、白银代表银白色的月亮等，这些贵金属因为稀少并且被赋予了幸运的含义，价格昂贵。地中海附近的国家希望能有一种将贱金属转变为贵金属的法术，这种法术被称为炼金术。人工智能拥有的就是这种将看似没有价值的数据转变为价值难以估量的决策的“法术”。

我们先来回顾一下AI炼金术全过程：首先，我们要通过对历史数据的特征处理产生特征数据，然后通过机器学习产生预测模型。模型产生之后，需要通过回测对其进行模型评估，在评估中合格的模型才可以投入生产。生产过程中如果有新的样本数据输入，就可以通过模型来产生预测分数，根据分数的高低和其他因素来进行业务优化，产生针对个体的业务决策。决策指导产生针对个体的具体行动，行动作用在个体上产生结果，虽然各自的结果可能会有差异，但人工智能总体上能够优化业务，从而实现价值。AI炼金术就是这样一步一步，对数据千锤万凿，百炼成金。

人工智能赋能架构

AI炼金术所创造的价值是不可估量的，对于企业，它是大赚1 000亿美元的秘密武器，对于个人，它是安居乐业人生美满的秘诀。但人工智能炼金术不是一座飘浮起来的空中花园，它需要建立在真正落地的实际系统上，需要系统给各个参与者赋能，否则再神奇的炼金术也只能是一场美好的空谈。那么人工智能炼金术最终要落地在一个什么样的系统上呢？

基于人工智能赋能架构的系统，能够为企业和消费者赋能。如图3-11所示，它分为左右两边，左边是C端，也就是用户端，通常涉及企业的用户或消费者，他可能在企业经营的门店里购买过商品，在企业运营的网站上注册过会员，或者是使用过企业提供的手机App，总之他和企业有过一些交互和接触。右边是B端，也就是服务端，服务端就是要在企业的后台为用户提供优质的服务，确保用户端得到最佳的体验。

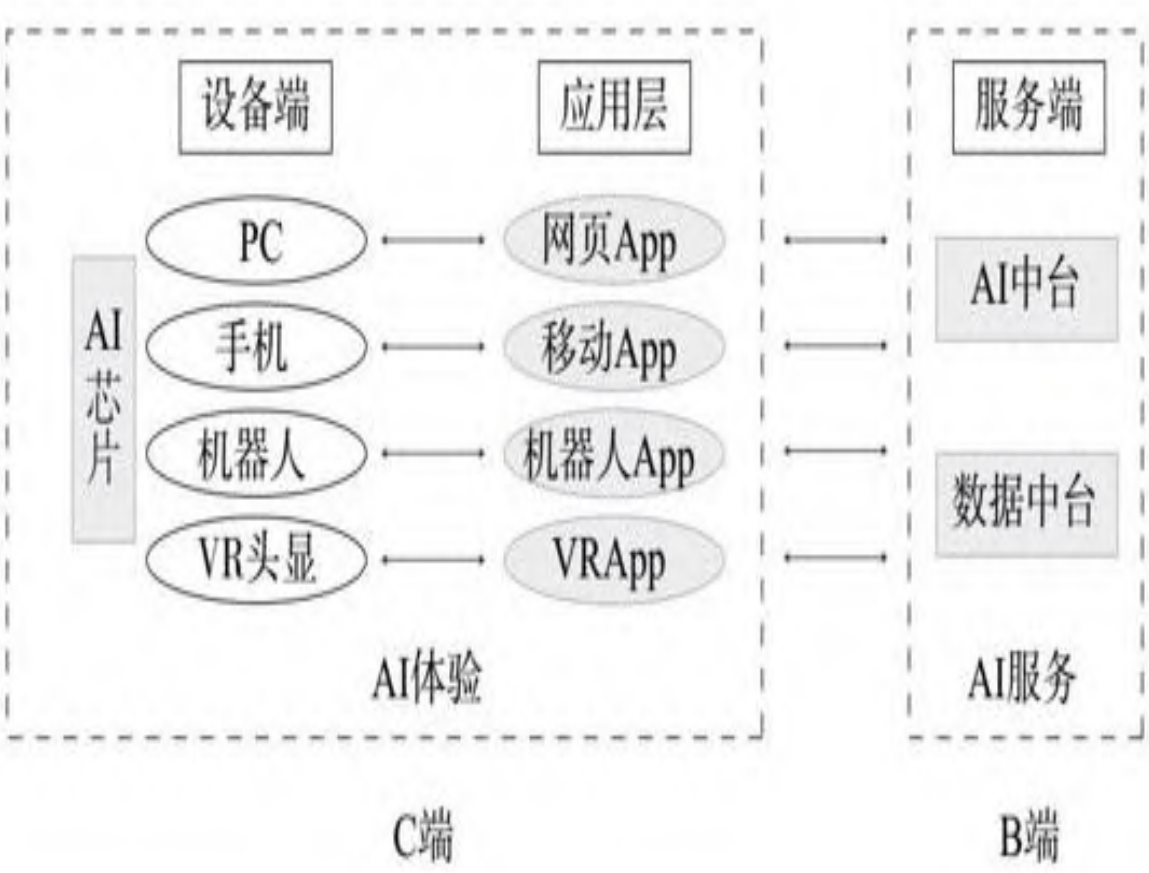


图3-11 人工智能赋能架构示意图

服务端的系统分为两部分，一部分是数据中台[1]，另一部分是AI中台。数据中台会对海量数据进行采集、计算、存储、加工，形成统一的标准数据，形成大数据集合，进而通过AI中台为用户提供个性

化服务。从数据到价值的AI炼金术，是AI思维的具象表现形式，AI中台则是AI思维的实现载体。在服务端，既需要数据、需要人工智能，也需要数据中台和AI中台的连通。只有数据中台和AI中台融为一体，人工智能才能知道瞬息万变的市场变化、用户的需求，才能给用户提供优质的服务。另外，数据中台和AI中台打通了，就会有源源不断的数据输入AI中台，人工智能训练的数据越多，训练的次数越多，它的智能程度就会越高，做出来的决策就会越准确，其所创造出来的价值就越高。

然后我们再来看用户端，用户端主要有两个模块，一个是设备端，另一个是应用层。现在的用户无论是线上还是线下跟企业交互，都要通过一些设备来实现，比如说在PC（个人电脑）端看视频、在手机端玩游戏、通过VR头显（虚拟现实头戴式显示设备）感受虚拟现实，或者说与在实景中摆放的机器人交流互动。所以说PC端、手机端、VR头显端或者是机器人端都是一个用户的触达点，都是一个企业的触角。通过这些触角，企业可以实时了解用户的需求并为其提供服务，比如说现在很多酒店都配备了智能机器人，如果酒店的房客半夜想要喝水，发现房间里的纯净水都喝光了，而酒店的工作人员都已经休息了，就可以让机器人为房客送纯净水。

在这些设备端里面可能会有AI芯片。在人工智能中，AI芯片是专门用于模型运算任务的一个模块，它可以在手机里做大量快速的计算，比如说现在很多年轻人喜欢通过各种美图App给自己上妆或者为自己的照片换背景，这些特效的实现都是实时的，呈现效果也非常好，所以需要很大的人工智能运算量。面对越来越高的用户的体验要求和越来越高的运算量，传统的硬件已经不足以解决问题了，因此现在手机上都安装了专门用来处理这些问题的AI芯片。手机AI芯片发展迅速，比如说三星最新研发的手机AI芯片有了独立的神经网络处理单元，运算速度提升了7倍；华为研发的AI芯片NPU（Neural Network Processing Unit，嵌入式神经网络处理器）使图形处理性能提升了20%，能效提升了50%，图像识别速度可达到每分钟2 000张；OPPO、

vivo使用的是科技公司联发科研发的手机AI芯片，有低能耗、高性能的特点。

虽然AI芯片只是一块小小的芯片，但是AI芯片产业却是一个大的产业，因为它既能在用户端上进行人工智能的运算，也能在服务端进行人工智能的运算。虽然服务端跟用户端是连通的，但是网络存在延迟，直接在用户端进行计算可以缓解这个延迟。从整体架构上来说，设备端和服务端中间是应用层，其中包括像微信、QQ（腾讯公司的即时通信软件）这样的应用App，通常企业需要花大量的时间和精力去改良这些应用，提高用户体验和服务质量。例如，谷歌开发的语音助手Google Assistant的语音模型一开始是存储在谷歌服务端的，它能够准确识别出用户所说的指令并做出相应的回应。但是在使用过程中，用户对着Google Assistant说一句话，这句话需要上传到服务端才能被语音模型识别，然后再传回用户的手机，这就导致了语音识别速度很慢。为了改善用户在使用语音助手时的体验，谷歌对语音模型做了处理，将其大小从100G压缩到了0.5G，使其可以直接缓存到用户的手机上。如此一来，语音在手机和服务端之间的传输时间大大减少，Google Assistant的语音识别速度提高了将近10倍，极大地提高了用户体验。

所以前文提到的AI炼金术，其实是构建在这样一套赋能架构上的。从图3-11最左边的用户到最右边的企业，环环相扣，人工智能不仅可以由企业的服务端提供算力，也可以在用户的设备上进行运算。这块根据响应时间会有一个切分，对响应时间比较敏感的会在用户端进行，而对时间要求没有那么紧急的任务，可以放在服务端进行。另外，服务端有更多更全的数据，适合更为宏观的人工智能运算。这套连接服务端和用户端的赋能架构，是人工智能炼金术所依赖的系统基础。理解整个赋能架构，能够帮助我们更好地应用人工智能，通过人工智能更多的创造价值。

人工智能应用矩阵

人工智能的发展、传播和被接受是经过了一段漫长的寒冬的。十多年前，它还只是一个不被人看好的小众领域，但是现在，它却已经成为街头巷尾的热点谈资，几乎任何事情都可以和人工智能联系在一起。

短短十年间，世界发生了天翻地覆的变化，新数据不断涌现，各种问题层出不穷，直到现在，人工智能的春天才算是真的到来了，各个领域都亟须人工智能的帮助。这也是为什么人工智能的商业化应用范围如此广阔，人工智能市场更是如一块等待开垦的处女地一般。如图3-12，这是一份人工智能的商业版图，横轴代表不同的行业领域，例如金融、零售、医疗和教育等，纵轴代表不同的职能方向，例如营销、风控和安全等。横轴和纵轴共同构成了一个二维的商业矩阵，对于每个行业中的相关职能，人工智能都可以探索相关应用场景，例如在金融行业的营销、CRM（客户关系管理）、客服、风控和安全等方面以及零售行业的营销、CRM和客服等方面都已经有人工智能实践落地。但其实，现在的人工智能只填充了广阔的商业领域中非常小的一部分，还有更多没尝试和拓展的人工智能应用场景。

在金融行业，人工智能应用十分广泛。以金融服务集团蚂蚁金服为例，他们推出了智能理赔服务，用户可以直接在手机上上传票据，通过智能审核就能迅速获得赔付款；支付宝将人脸转变为数字密码的刷脸支付功能也是人工智能在金融行业的应用。除此之外，人工智能还可以通过模型和投资组合优化等，为用户提供投资参考的智能投顾服务；通过自然语言理解、语音识别等，实时为用户咨询提供回答的智能客服服务；通过推荐模型，根据大数据为用户推荐个性化产品的智能营销服务；以及通过人脸识别和指纹识别，在支付时检测用户身份的智能风控服务。这样一系列服务使人工智能在金融领域扎实落地。

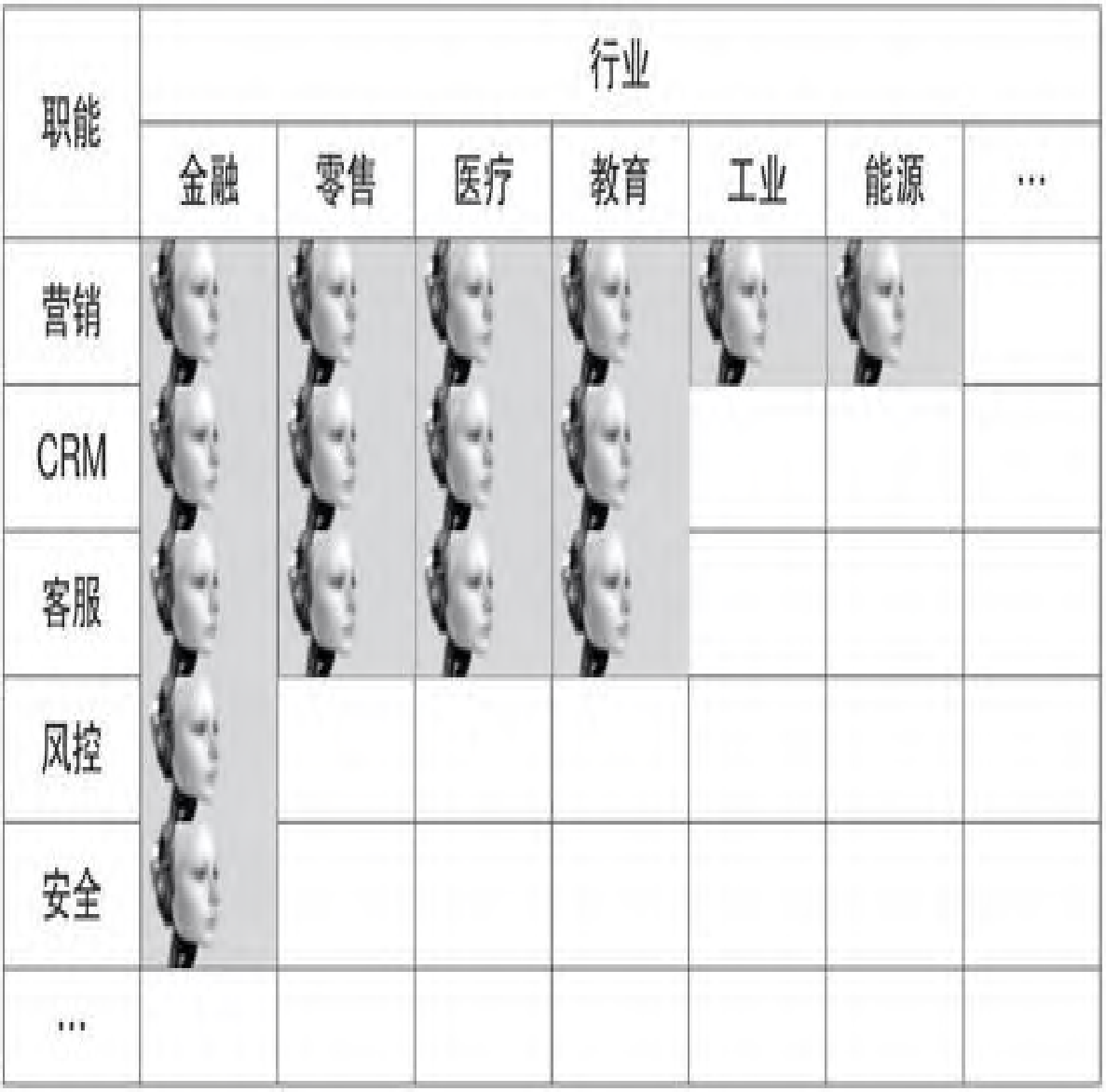


图3-12 人工智能的商业版图

随着社会经济的发展，居民收入水平的提高，人们的消费能力逐渐上升，零售行业规模随之扩大。根据预测，2020年全球零售销售总额将超过26万亿美元。由此可见，人工智能在零售业的落地也有很大的应用前景。例如，人工智能可以依靠基于大数据的机器学习模型，为用户生成精准化、定制化的推荐，提高零售业销量，Stitch Fix的成功就是很好的证明。Stitch Fix是美国的一家时尚电商，活跃用户

达数百万人。它之所以在电商行业竞争如此激烈的当下能够留住用户，就是因为正确运用了人工智能，向用户推荐个性化商品。Stitch Fix专门请数据科学家建立了人工智能推荐模型，根据用户的身高、体重、职业等信息以及穿衣风格和购买价位等偏好，为用户推荐定制化、个性化的商品。Stitch Fix在2016年和2017年分别创造出了7.3亿美元和9.8亿美元的销售额，而且这些销售额来自智能推荐。由此可见，人工智能可以有效地赋能零售行业，为其带来效益的增长。

在工业、医疗行业等领域，人工智能虽然也有应用，但并不及在金融和零售行业普遍。从整体来看，由于人工智能是近几年才发展起来的新技术，所以它在很多行业和领域的应用并不广泛深入。但是人工智能发展势头很好，在每个行业都有落地的可能性，而且随着机器学习、计算机视觉、语音识别、自然语言处理等人工智能领域的发展，我们有理由相信，人工智能在各个领域都能扎下根去，遍地开花。

以前，行业追求的都是芯片驱动（Intel inside），用芯片为企业提供算力和速度，让企业能够跑赢市场，找到自己的阿尔法。但这都已经成为过去，以后行业追求的人工智能驱动（AI inside）才是真正的阳关大道。人工智能会成为行业的决策引擎，取代在战术层面上需要人工干预的场景，持续有效地让企业在最佳状态下运行；而且人工智能带来的效果不仅仅是“优化”这么简单，而是革命性的、成倍数的增长。如此，人工智能才能真正展现它的能力，真正为行业赋能。

AI炼金术的场景化应用

既然AI炼金术在各个领域都有落地应用的可能性，并且在多个领域的应用已经产生了不错的效果，那么，它在具体的场景下是如何展开应用的呢？为了便于理解，我们先从一个大多数人都比较熟悉的例

子讲起。比如说，你在一个电商网站上买东西时，遇到了很多广告。这些广告位都是商家花钱购买的，即使你可能并不会购买广告位的商品，这个商家在此投入的广告费是无法撤回的，所以商家要尽可能地提高你购买广告位商品的概率。此时，人工智能精准推荐就成为商家提高购买率的法宝，因为显示一个顾客感兴趣的商品能大大提高他的购买率。事实证明，经过人工智能精准推荐出现在广告位的商品，让无数剁手党们欲罢不能。

百年大计，教育为本。社会各界对教育越来越重视，国家更是提出了科教兴国的教育发展战略。教育是不可忽视的，但与此同时，教育业也存在诸如教育水平发展不均衡，学生学习负担重、学习兴趣不高等问题。而人工智能的出现为这些问题提供了解决方案。在教学环节，教师可利用人脸识别模型实时检测学生随堂表现，辅助教师教学；在练习环节，文字识别模型可批改文科作业，图像识别模型可批改数学作业，这些智慧作业平台大大减轻了教师批改作业的负担；在考试环节，语音识别模型可对语言语音进行打分，也可利用人工智能进行试卷批改，统计、分析分数；课堂之外，学生可以利用人工智能搜集资料、完成作业等，比如遇到不会做的题目，学生可利用各种人工智能拍照搜题找到答案。除此之外，人工智能还可以根据学生的成绩、课堂表现、作业反馈等信息，判断该学生的学习习惯、学习能力以及学习程度，并为其制订个性化学习方案。例如，一个学生的英语听说能力很好，读写能力不足，那么在日常英语学习中，可以相对减少一点对于听说能力的训练，节省下来的时间用于加大读写能力的训练。通过人工智能制定的个性化学习方案，老师和家长能更好地帮助学生找到学习的乐趣，提高学习的能力，获得更多的知识。

人工智能在法律领域的应用也非常广泛。当事人如果遇到法律问题，人工智能可以根据当事人的信息、相似案情的判决等数据来进行决策，为当事人提供法律咨询，或者为当事人推荐最合适的律师。对于律师来说，人工智能的文本处理能力可以帮助律师搜集资料，例如相关法律条文、相关案情的判决书等。人工智能现在已经可以根据已

有信息判断案件能否胜诉。比如说，CaseCrunch公司的人工智能就可以根据案件信息进行判决预测，并且预测的成功率高达86.6%，打败了人类律师62.3%的成功率。

2018年斯坦福等高校联合开发的律政界人工智能软件，和20名有经验的律师做了一个比赛。该比赛的内容是在4小时内审查5项协议，并回答30个法律问题，涉及领域包括仲裁、赔偿等。最终，人工智能软件以94%的准确率完胜律师的85%。人工智能为法律领域带来的价值是为不懂法律的大众提供专业的法律帮助，为律师减轻工作负担，提高法律案件处理的效率，协助法律系统创造一个公平清明的世界。

人工智能在维护社会公平自由的同时，还能为医疗行业贡献一份力量，维护人类的健康。根据《2018世界人工智能产业发展蓝皮书》显示，2013年至2018年第一季度，全球人工智能的投资在医疗健康领域的热度最高。据老牌咨询公司麦肯锡预测，到2025年，全球智能医疗行业规模将达到总计254亿美元，约占全球人工智能市场总值的1/5。由此可见，人工智能在医疗领域有广阔的应用前景。

医疗行业长期存在优质医生资源分配不均的问题，与此同时，在诊断过程频频出现误诊漏诊现象，而且医疗费用一直居高不下。对于普通民众来说，也一直存在看病难、看病贵的问题。人工智能通过对大量医疗数据、病人信息的分析，可以了解病人的病情，甚至能够预测一些疾病的发病时间。根据这些信息，人工智能可以生成针对每个病人的个性化医疗方案，帮助病人尽快康复并预防其他疾病。而那些严重的疾病，比如基因变异引起的癌症，虽然千差万别，情况复杂，但是人工智能可以针对每个人独特的基因变异数据，协助制订最佳的治疗方案，最大限度地挽救生命。病人康复出院之后，人工智能还能提供长期的健康监测服务，防止旧病复发、新病萌芽等现象。同时，人工智能医疗流程系统还能为医护人员安排最合理的工作流程，避免医生在简单的咨询工作中耗时过长，从而提高医生的诊疗效率。

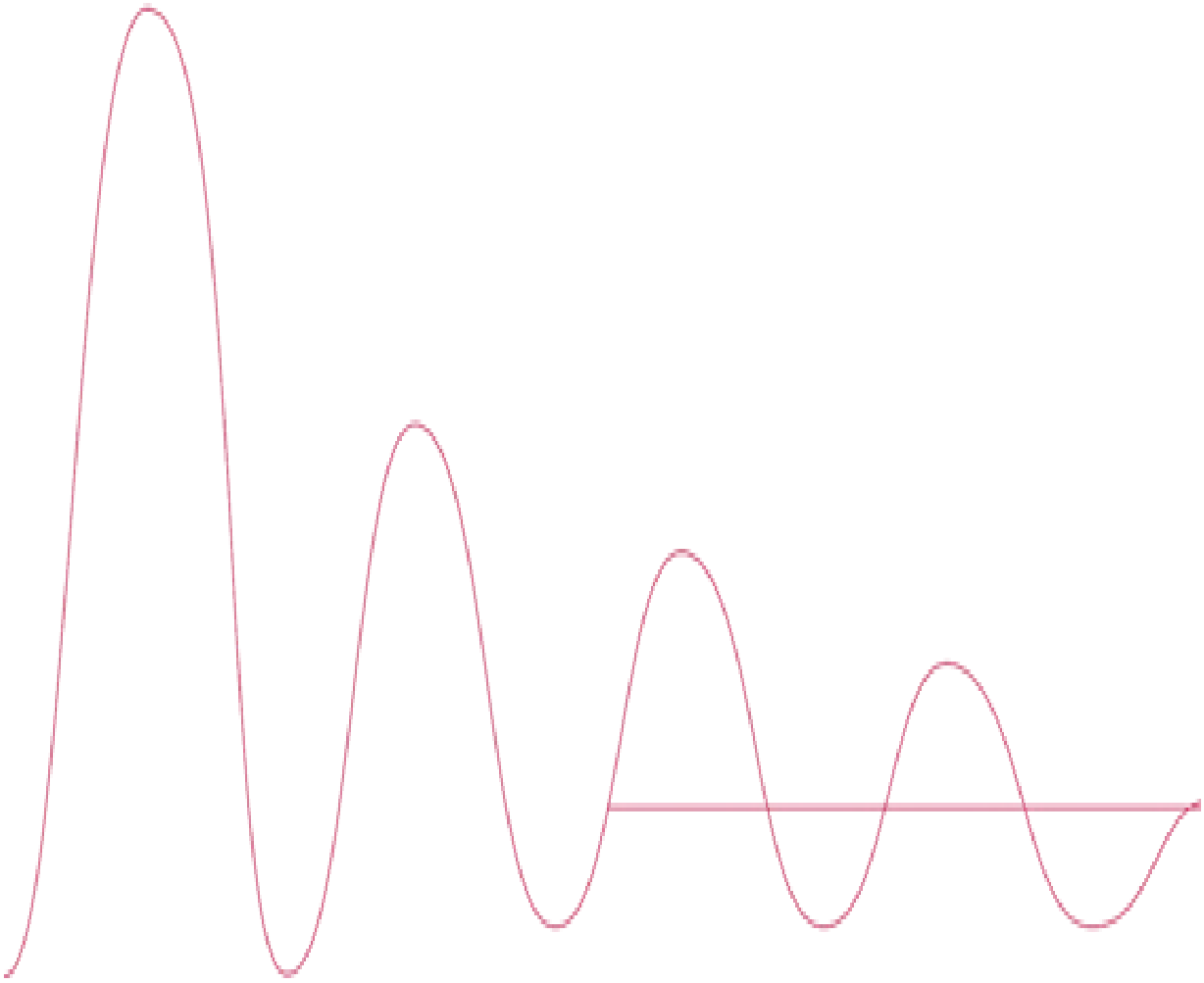
人工智能在各行各业的广泛应用空间可能会使你有些恐慌，觉得人工智能再发展下去将取代人类的工作，甚至给这个世界带来灾难。这不只是一个人的忧虑。早在2016年，英国著名物理学家霍金就说过人工智能可能是人类的灾难：“对于人类来说，强大人工智能的出现可能是最美妙的事，也可能是最糟糕的事。”但我们现在的人工智能大部分都是弱人工智能，是那些不能真正地独立推理和解决问题的智能机器。有人又要问，AlphaGo不是具有自我学习独立思考的能力吗？其实不然，围棋有固定的规则，对手下棋的习惯可以被模型捕捉到，但真实世界里的情况千变万化，不用说庞大而复杂的基因数据了，就是普通人的行为都千差万别，没有人类参与的人工智能是很难成气候的。而那些能够独立推理和解决问题的强人工智能的出现目前来看还很遥远，所以我觉得，认为人工智能会取代人类这样的焦虑，有些杞人忧天。

---

[1] 中台即在一些系统中，被共用的中间件的集合。——编者注

第四章

人人都能理解AI



在制造分秒的时候，钟表把时间从人类的活动中分离开来，并且使人们相信时间是可以以精确而可计量的单位独立存在的。分分秒秒

的存在不是上帝的意图，也不是大自然的产物，而是人类运用自己创造出来的机械和自己对话的结果。

—《娱乐至死》—

在前文中，我们首先从“道”的层面理解了AI是时代的机遇，之后又从“法”的层面讲了AI思维的底层逻辑，最后从“术”的层面讲了AI炼金术。但是学习是一个螺旋式上升的过程，需要不断深入，不断提高；知识进一步提升，思维能力也会更强，理解和运用人工智能也是这个道理。本章将会更加深入下去，从“器”的层面讲解AI思维。“器”可以帮我们理解AI思维的奥秘，也是人工智能越战越勇的法宝，对人工智能的“器”有了深刻的理解，就像孙悟空有了金箍棒，可上九天揽月，就像哪吒有了混天绫和乾坤圈，可下五洋捉鳖。“器”是工具，而人工智能领域的“器”就是模型，模型是AI思维的核心，本章将挑选人工智能领域的经典模型来帮助你理解AI思维，其中包括分类和聚类的模型，图像、语音以及文本的识别模型，还有消费和社交数据的处理模型。首先让我们从AI的分类和聚类开始理解AI思维的“器”。

AI的分类和聚类

只会纸上谈兵的赵括由于缺乏战场经验，在长平之战中输给了秦将白起；熟读兵书的马谡不听从有实战经验的副将的建议，将街亭输给了司马懿，逼得诸葛亮只能挥泪斩马谡。历史上许多故事都告诉我们：缺乏实践经验，只会空谈，必然会导致失败。

过了那个兵荒马乱的年代，虽然我们不会再经历千军万马丢城失地的覆灭，但在日常生活中，若缺少经验，我们常常会遭遇挫折。从没有学过游泳的人，为了冒险刺激就往深水区游，难免会多呛几口水，甚至失去生命；从没有投资过的人，看到别人获利就盲目进入风云莫测的股市，便是注定了赔钱的命运。

然而我们都是第一次生而为人，只要对这个世界仍然保持着好奇心，总会去尝试无数个缺乏经验的第一次。第一次失败了并不可怕，重要的是你从这次的失败中学到了什么。及时总结，在一次次的试错中，也就是从经验中，我们学习了万事万物的规律。操千曲而后晓声，观千剑而后识器。以后再遇到相似的问题，我们就能妥善地处理了；碰到新的问题，我们也能根据习得的规律从容地做出比较准确的判断了。

人工智能各种能力的形成也是这个过程。我们从经验出发，得到处理事情的方法论；人工智能从数据出发，得到预知未来的模型。经过对数据的多次学习，人工智能就像一个身经百战的智能人，如果再有新的数据输入，人工智能也可以独立做出准确的判断，以及合理的预测。对于人工智能来说，产生模型的方法论就是机器学习，下面我们来仔细讲解一下。

机器学习的分类

我们获取知识的途径有两种，一种是通过他人的传递，例如我们遇到不会的问题时去请教老师或同学；另一种是在没有接受指导的情况下自学获得的。类似地，机器学习也有两种：监督学习和无监督学习。

监督学习是从标记的数据来学习模型的任务，每个标记的数据都是由一个输入数据（也就是通常所说的x）和一个期望的输出数据（也就是通常所说的标记y）组成。无监督学习处理的则是没有标记的数据。监督学习适合有明确预测目标的应用场景，即y可以被明确地定义和量化。而无监督学习通常适用于无法预先知道输出值的场景，只能对现有输入向量进行观察分析，它可以从数据中发现具有相似输入向量的结构，这些结构统称为簇（cluster）。

机器学习问题根据输出的数据来分，也有两类。一类是离散输出，通常用来表示个体的类别，比如把人分为两类，男人和女人；把客户分两类，高价值客户、低价值客户。数学中的离散是和连续相对的，所以另一类就是连续输出。连续输出的数据是指那些在一定区间内可以任意取值的数据，其数值是连续不断的，相邻的两个数据之间可以做无限的分割。比如说两趟地铁之间的间隔是10分钟，一个乘客在地铁站等地铁的时间是一个随机变量，取值范围可以是从0到10这个区间内的任意实数。所以连续输出有很大的应用空间，比如要预测公司的收入、客户的价值、生产流程的能耗、生产系统输出的气体浓度等，这些都需要用连续数字来表示，就都需要连续的输出。

监督类型分为通过监督学习和无监督学习，是机器学习问题分类的一种依据；输出分为离散型和连续型，是机器学习问题分类的另一种依据。将以上四类排列组合，就形成了四种不同类型的机器学习问题，如表4-1，分别是：

如果通过监督学习，得到离散输出，就是分类问题。

如果通过无监督学习，得到离散输出，就是聚类问题。

如果通过监督学习，得到连续输出，就是回归问题。

如果通过无监督学习，得到连续输出，就是降维问题。

表4-1 机器学习问题的类型

输出类型	监督类型	
	监督学习	无监督学习
离散输出	分类问题	聚类问题
连续输出	回归问题	降维问题

这四类问题是机器学习的基本问题，深入理解了这四类问题，我们就能对机器学习有一个完整的把握。无论在哪一类问题中，机器学习的目标都是根据新的输入，做出预测。理解清楚了这四类问题的框架，我们就能深入体会到人工智能解决实际问题的思路。下面我会深入分析两类问题——聚类问题和分类问题，它们也是机器学习中最常使用的两类。

聚类问题

《战国策》有云：“物以类聚，人以群分。”在日常生活中，有各种事物需要聚类，连人也在无形之中有聚类。而在人工智能领域，聚类问题对应的是无监督学习、离散输出的机器学习，也就是要通过机器学习将一些没有标记过的数据归为一类。数据的维度越来越多，聚类的目标是找到这些数据中具有某些相似特征的簇，找到每个个体的输入数据和簇之间的归属关系。也就是说，聚类中，我们需要按照一定的模型，把所关心的个体归并到一些簇里。

这个世界日新月异，每时每刻都有新的事情发生，新闻记者在事件发生的第一现场为我们传递新闻。但是很多时候我们并不能了解到新闻的全貌，因为它散落在各大新闻报刊以及新闻网站的各个角落。聚类的出现就能够让零散的新闻聚集起来，帮助我们了解新闻的来龙去脉。通过聚类，我们可以有效地将那些相同主题的新闻聚集在一起，形成簇状的结构，我们再进行浏览时就会更加全面、准确。

在商业实践中，聚类对于发现一些特定的人群结构非常有用。就比如PayPal面对的是全球范围内的几亿用户，它可以通过聚类对这些用户成百上千的维度进行详细的分析，从中得到百余个簇。每个簇对应属性和行为比较接近的一类用户，不同簇之间的差异化程度比较大。例如，一个簇对应了年轻的母亲，一个簇对应了中年艺术家，一个簇对应了软件专业人士，等等。理解了人群的不同特点和诉求后，公司才可以针对性地定位、开发、推广产品。另外，聚类中有时会发现一些未曾关注却占比显著的簇，那么整个品牌的定位都会被聚类的结果和发现影响。PayPal建立的这一套聚类体系帮助公司更好地了解用户和市场的需求，为他们提供了更加贴心的服务，由此可见，聚类体系是可以为商业带来价值的。

常见的聚类方法有：K-均值聚类、EM聚类、合成聚类、Mean-shift聚类、频谱聚类。这些方法根据不同的数学原理来做聚类。方法不同，结果不同，但大的目标都是一致的，都是为了发现并输出数据中的簇。簇的标志性特征是内部差异化较小，而簇与簇之间的差异比较大，这也是聚类方法的基本原理。

下面我们来说明一下聚类。如图4-1所示，左边是原始的数据，用三种深浅不同的点来表示有三个簇。K-均值聚类和EM聚类的结果分别在中间和右边的图上。因为聚类是无监督学习，没有方法能够完全还原原始的簇，但是合适的聚类方法还是能够发现数据内在的结构。各行各业都会产生数据，所以不论是在商业领域，还是在生物学、医学、天文学等科学领域里，聚类都有它的用武之地，都能够帮助各领

域的专业人员发现新的知识。聚类就像一个数据显微镜，能够拉近我们和数据的距离，把微小的每个样本聚合成可以解读、可视化的簇，增加我们对数据的理解。

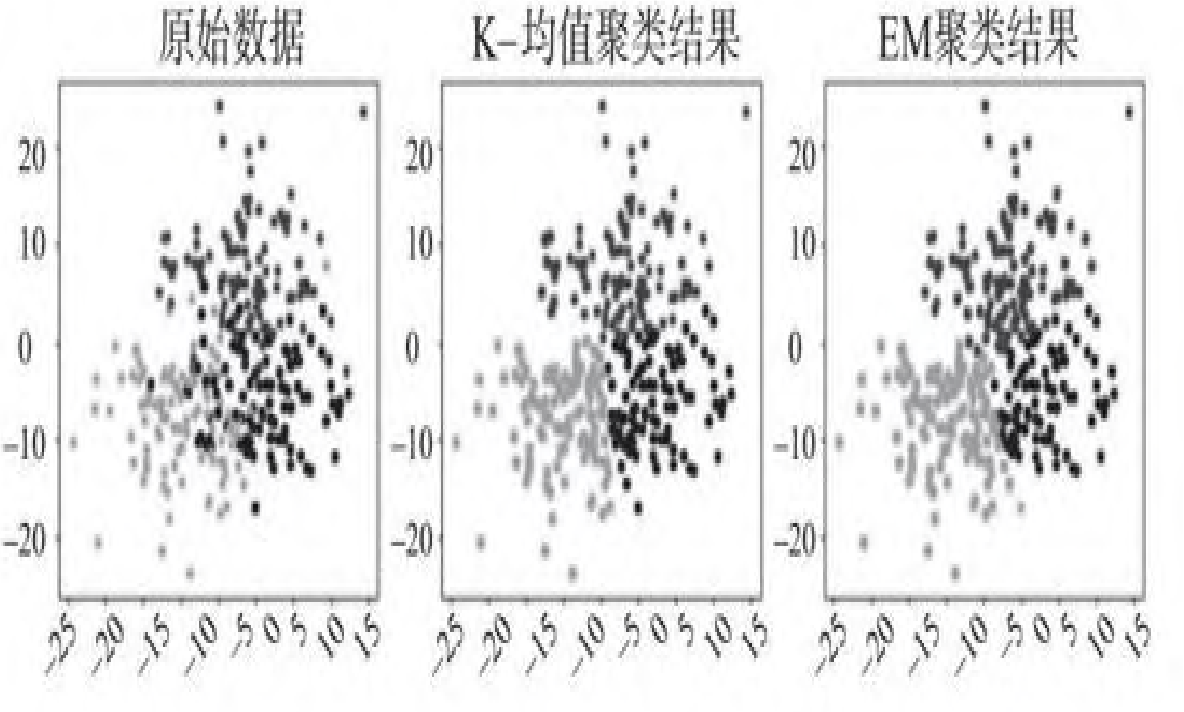


图4-1 聚类示意图

分类问题

下面我们来看一下分类，也就是监督学习、离散输出的机器学习。分类要做的是将一个预测函数f，作用在个体的特征x上，得到想要的输出$f(x)$。这个看似简单的识别过程其实是非常强大和灵活的，基本上所有监督学习的过程都可以抽象成这样的形式。预测的目标可以是任何跟业务相关的量，包括顾客的价值、消费者的喜好、借款人的风险、病人康复的概率等。深入理解了人工智能后，你就可以针对你所在的领域，构造出相应的分类问题，从而有效地提升业务。

提到分类，就不得不提2019年全国都在热议的话题——垃圾分类。2019年7月，上海正式开始强制实施垃圾分类，个人投放垃圾错误最高可罚200元，单位投放垃圾错误最高可罚5万元。一时间，垃圾分类成为上海人民生活的主旋律。即便是将垃圾分为了可回收物、有害垃圾、干垃圾、湿垃圾四类，但是由于真实生活中垃圾种类繁多，混合垃圾难以分辨种类，很多人还是会分辨错误。其实国外许多国家实行垃圾分类很久了，分辨多种垃圾的问题也困扰国外人民很久了，所以有很多人工智能公司就开发了基于监督学习的垃圾分类系统。例如日本发那科（Fanuc）的分拣机器人Waste Robot就是这样一个垃圾分类系统，该系统内部拥有经过数据集训练的模型，可以通过实时检测来分辨垃圾的类型。

监督学习的模型不是想当然地形成的，它要经过严格的训练和预测环节，这两个环节也是监督学习最基本的环节。

训练环节中：给定一个训练集，由标记样本构成｛(x_1，y_1)，…，(x_N，y_N)｝。所谓标记样本就是既有特征向量x，又有对应的标记y的一组数据。通过最小化训练集上的预测误差，估计一个预测函数f，也就是我们的预测模型。需要特别注意的是，训练集都是由样本和对应的标记构成的。例如我手中有大象、猴子、长颈鹿等各种动物的图像，然后给这些图像打上相对应的名称标记（图像和名称标记一一对应），打过标记并且一一对应的图像称为训练集。

预测环节中：在新预测样本x上，使用 f 输出预测值 $y=f(x)$。就比如人工智能经过了刚才关于动物图像的训练，我们再把新的图像输入给它，此时人工智能就可以输出图像上的动物是什么了，这就是预测环节。

在训练阶段，样本的标记是已知的，但是在预测阶段，样本的标记是未知的。就像读书的时候，平时死记硬背的同学不一定考试考得好，考试考得好的同学，也未必在社会上发展出色。机器学习也是这个道理，在新的、未曾见到过的预测集上的预测效果，才是做好机器

学习的关键。那么要做好机器学习，也就是训练出一个预测效果好的模型，需要多少数据样本呢？这个问题是无法一概而论的。如果你的数据样本大，行为或模式复杂，要预测的标记本身也会发生变化，那么这个模型需要的训练数据就多。通常，跟训练简单模型（好比后面我们要说的线性模型）比较，训练复杂模型需要的样本也会比较多。

需要注意的是，用来训练的样本一定要代表实际的业务场景，这样机器学习产生的模型才能在实际业务中产出良好的预测效果。如果在实际业务中遇到的预测样本和训练样本的特性相差甚远，那么模型是很难产出良好的预测效果的。比如你想开一家饭店，但却只在上海地区展开饮食喜好的调查，那么你大概会得出“消费者都比较喜欢吃清淡一点的食物”的结论。根据这个结论，你精心研发了一份清淡可口的食谱，但是这家饭店最终选址在川渝地区。可想而知，饭店开业后并不会有太好的生意，因为两地的人群样本对食物的喜好相差很大，上海样本上得出的结论很难在川渝样本上产生同样的效果。

通过前面的阐释，我们已经知道了分类的概念，下面再通过一个具体的图像分类的例子来更加深入地了解分类。如图4-2所示，训练阶段我们提供了一组有类别标记的图像，预测阶段，给定一个图像，我们的模型可以输出图像上的物体类别。

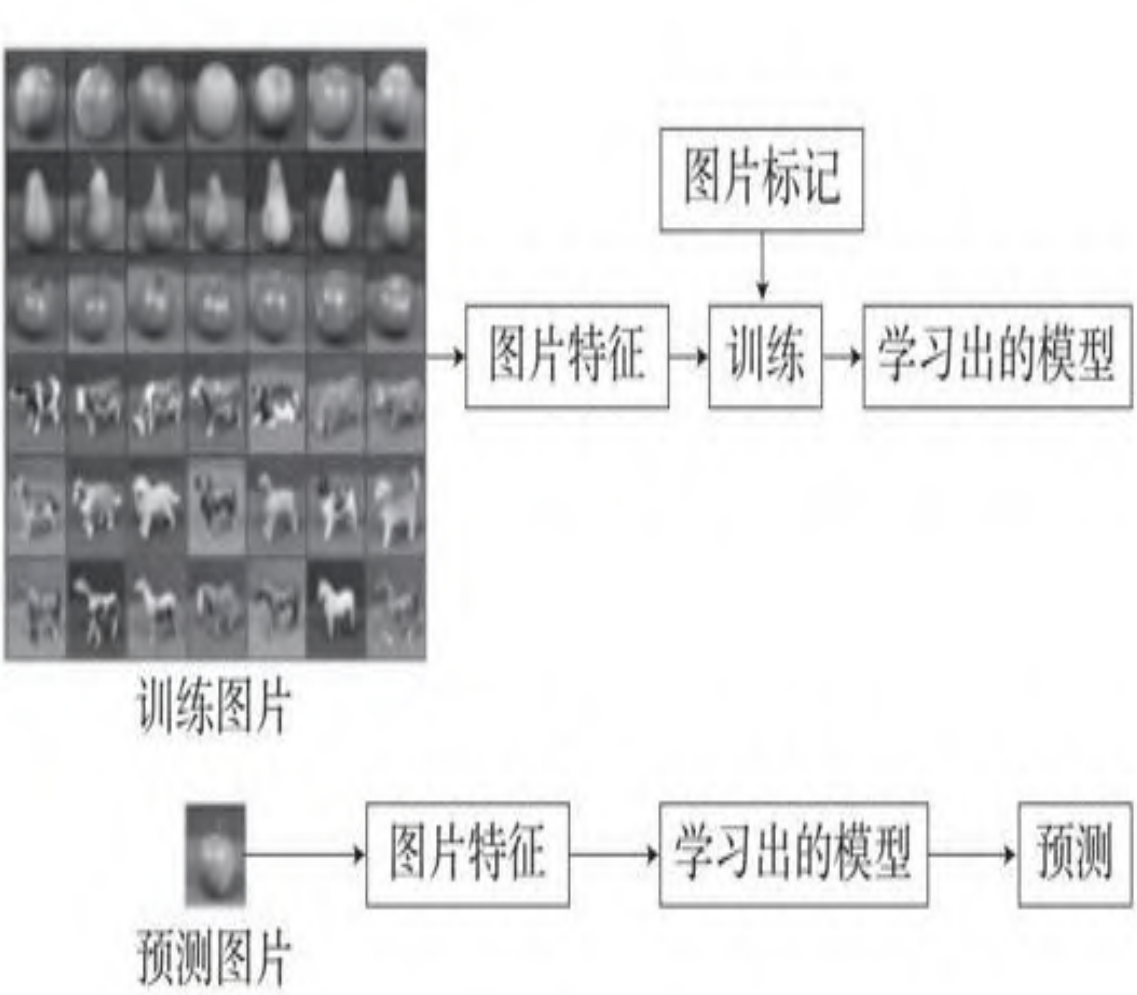


图4-2 图像分类示意图

训练的时候，我们给了6类训练图像，分别是3种水果和3种动物。每张图像先经过特征化处理，产生图像特征。根据这些图像特征和标记，机器学习能够学习出一个模型。预测的时候，我们给定一张图，把它特征化后，模型就可以针对这个图像特征，输出一个类别预测。

之前提到过，机器学习中，数据以向量形式表示。对图像来说，有很多种向量表示的方法。我们挑选3种基本的向量表示方法展开讲解：原始像素值特征、直方图特征、GIST特征。如图4-3所示，原始像素值特征，用的就是原图的彩色或者灰度像素值，把它们拉成一个向量作为图像的表示，这种简单的方法只能表示简单的图像。图像的像

素值很容易受到环境中各种因素的干扰，用它们表示复杂图像的内容，可用性不强。直方图特征，是指加工原图像素值，用加工后的像素值分布，而不是原始值来表示图像。GIST特征，用更复杂的方法来全面表示图像中的场景，能够更好地刻画图像的视觉特点，比如说是自然场景还是人工场景，场景是封闭的还是开阔的。

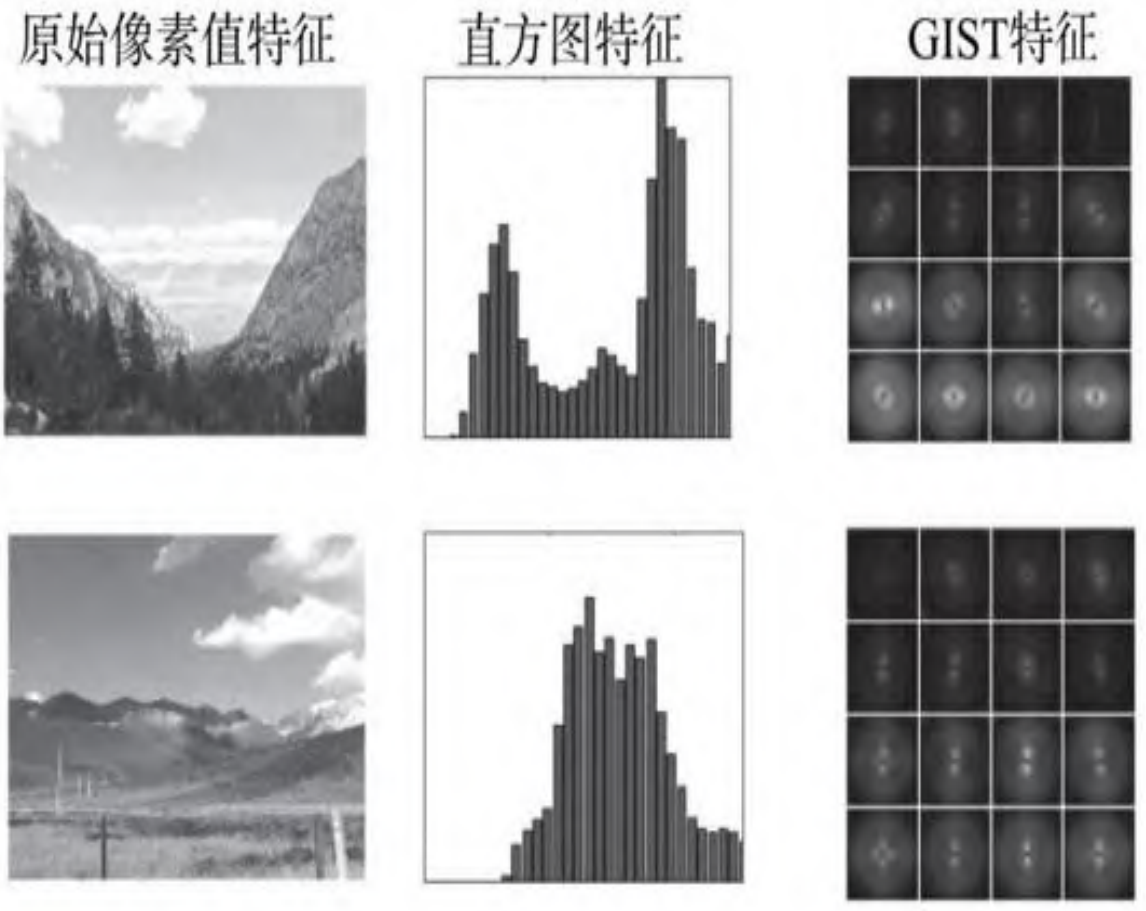


图4-3 图像的特征向量表示方法

对于图像来说，设计的特征越精巧，在机器学习中，图像识别的效果就能够做得越好，但是人的肉眼却很难看明白这些复杂的特征。人工智能和人的思维方式不一样，对于人来说很直接的原始图像，计

算机却很难处理；而构造精妙的图像特征向量，人则不一定能看明白。

最近邻分类

前面以图像为例说明了机器学习分类的过程，下面我们来了解一种基本的分类方法：最近邻分类。这种方法不需要训练环节，只需要针对输入数据的距离定义。

如图4-4所示，我们用方形和圆形分别表示两种不同的训练样本，一个新的预测样本会根据离两类训练样本距离最近的数据点，来确定自己的分类。也就是说，如果和预测样本最近的样本是方形样本，我们就把它分类为方形，相反，如果和预测样本最近的样本是圆形样本，我们就把它分类为圆形。

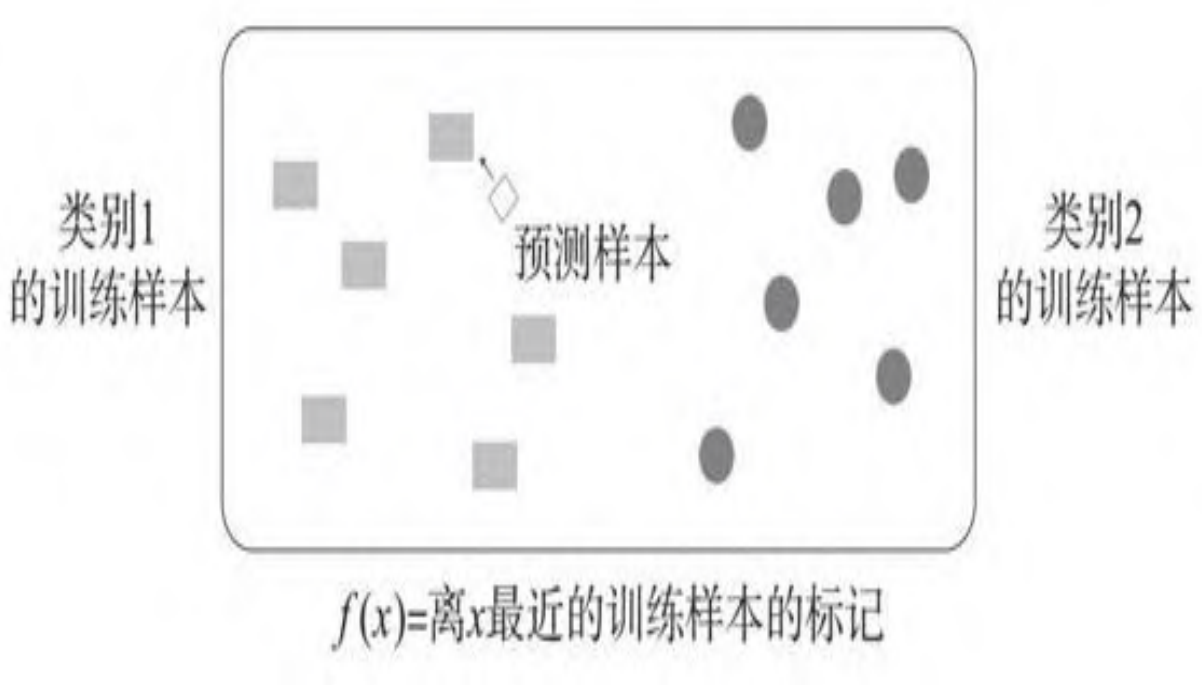


图4-4 最近邻分类示意图

这个分类过程类似人的思考过程，好比我们对一个不熟悉的物品，一个没去过的地方，一个不认识的人，我们会下意识地把它跟我们接触过的物品、地方、人做对比，用一个我们熟悉的事物的类别作为对这个新事物类别的“预测”。人的思维如此，机器学习也应用了这个原理。

虽然最近邻分类比较简单，但是它在很多实际场景里的应用效果尚能差强人意。有的时候，你可能并不知道要用什么样的模型来分类数据，那我建议你试试最近邻分类，先看看它能产出什么效果。对最近邻分类有了一些了解后，下面我们来介绍更多的分类模型。

线性分类和K-近邻分类

在机器学习领域，线性分类是一种经典的分类方法，它的目标是寻找一个线性函数，直观上也就是一条线，来分割两类数据，如图4-5所示，在这条线的一边是方块的一类，另一边是圆形的一类，而且这条线与两边的距离要适中。

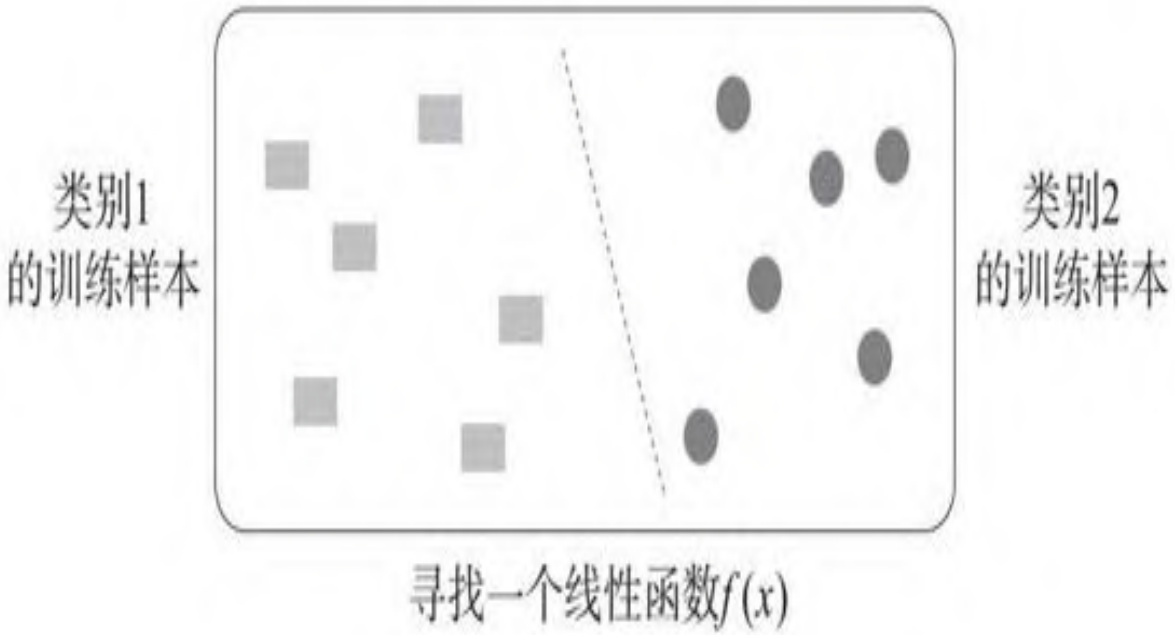


图4-5 线性分类示意图

当然，分割线未必只能在纸上，也就是在二维空间里画的线，它也可能是三维空间里的一个平面，又或者是高维空间里的超平面。这个线性函数可以在任意维度的空间里，它取决于机器学习任务中特征向量的维度。例如，我们有2 000张花花草草的照片，如图4-6所示，每张照片由这个高维空间中的一个点表示。如果要把这两千张照片按花和草来分类，那这个线性函数就是图4-6所示的超平面。（因为我们无法看到高维空间里的超平面，所以这里的超平面用平面来表示。）





图4-6 高维空间线性分类示意图

我们实现线性分类的方法叫作最小二乘法，又称最小平方法。通过最小二乘法可以计算出进行分类的这个线性函数的参数，也就是说可以计算出一条分类的最佳分割线。原理上，这个划分两类的分割线，要能充分“拟合”两边的训练数据，不能相差太远。但既然是拟合，就难免会有误差，在数学上，最小二乘法就是通过最小化误差的平方和来寻找最佳拟合的线性函数的。不论你的特征向量维度是多少，最小二乘的方法论都是通用的。理解了线性分类，我们就能理解更多的机器学习模型。机器学习最终学习出的函数形式可能会很复杂，不一定是线性的，但原理上都是通过最小化一个目标值来算出最佳的分类函数。

在实际应用中，线性分类最重要的优势是它的结果易于解读。线性分类会针对每个特征变量，输出一个权重，这个权重决定了变量在分类中的重要性。权重的绝对值越大，变量越重要。如果权重是正数，就意味着变量和预测目标有一定的正相关性；如果权重是负数，就意味着变量和预测目标有一定的负相关性。

举一个简单的例子来说明一下权重。比如你今天到商场逛街，逛累了之后要决定午饭吃些什么。那你就需要收集商场里所有能吃饭的地方的信息——当然可能不需要复杂到建立一个模型，但是肯定也会在大脑里经过思考。假如你十分喜欢吃辣的食物，可以接受面食，不喜欢吃甜食，那么在这次决策里，吃辣的食物以及吃面食这两种变量所对应的权重就都是正数，只不过吃辣的食物权重更高，而吃甜食的权重就是负数。最终你的选择是这些变量的加权线性组合决定的：权重是正数的变量，权重数值越大，相应方案被选择的可能性就越大；权重是负数的变量，权重的绝对值越大，相应方案被选择的可能性就越小。所以你最后很有可能选择比较辣的食物，比如说川菜。

在人工智能里，除了最近邻分类、线性分类外，还有很多种分类方法，比如下面要介绍的K-近邻分类。K-近邻分类的一般原理是给定一个训练集，输入新的样本，在训练数据集中找与该向量最邻近的K个

样本。这些样本多数属于哪个类别，就把该新样本分到这个类别中。最近邻分类可以看作K＝1的情况。

我们举个例子来更好地理解K-近邻分类。如图4-7所示，下面是我们收集的几部电影的数据：主要统计的是电影中的战斗镜头和拥抱镜头的数量，还有电影的类型（未知电影除外），比如电影A中有战斗镜头4个，拥抱镜头109个，电影A为爱情片。

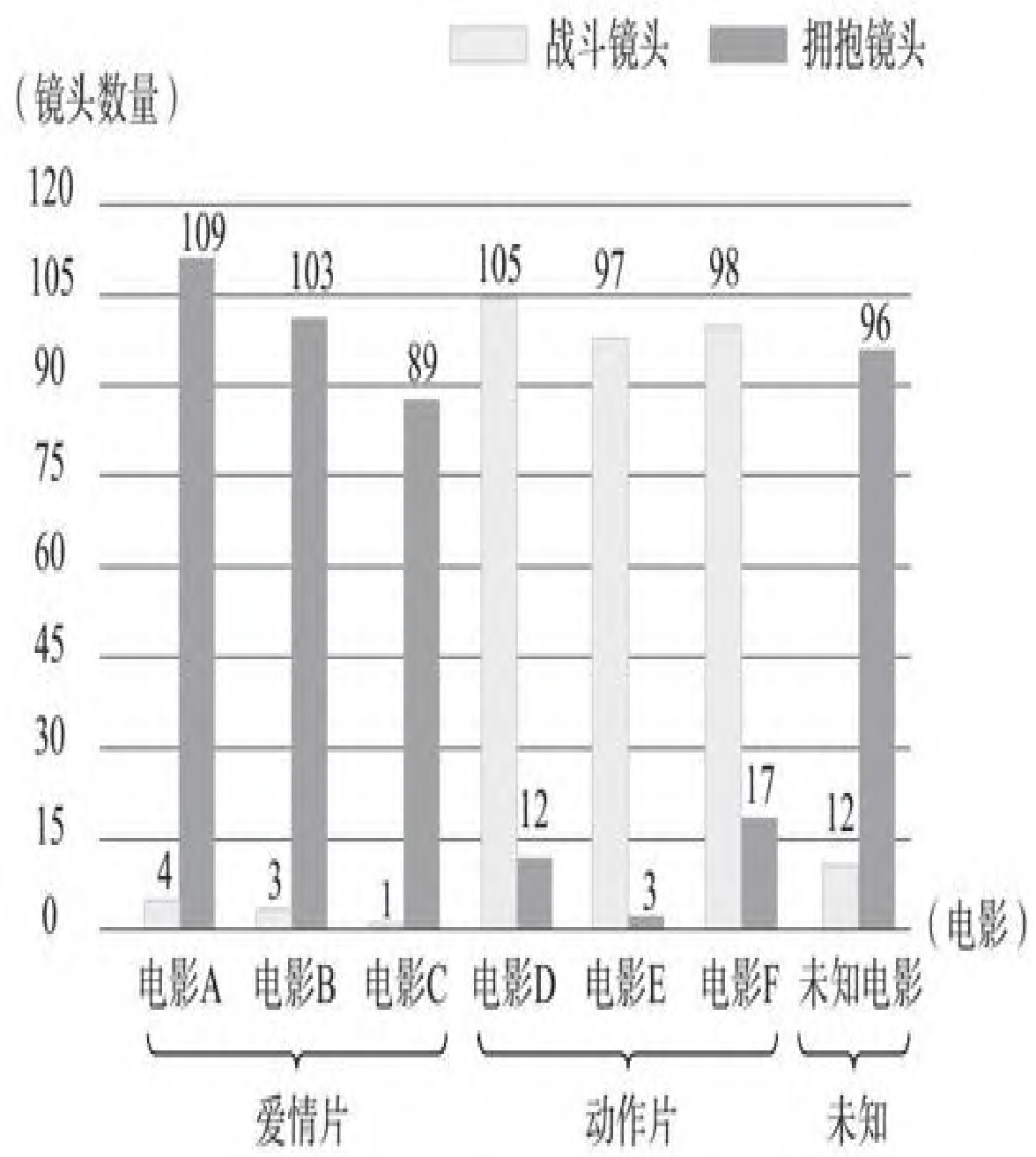


图4-7 电影镜头数据统计图

根据上述收集的电影数据，我们可以计算出已知电影和未知电影的距离，如图4-8所示。

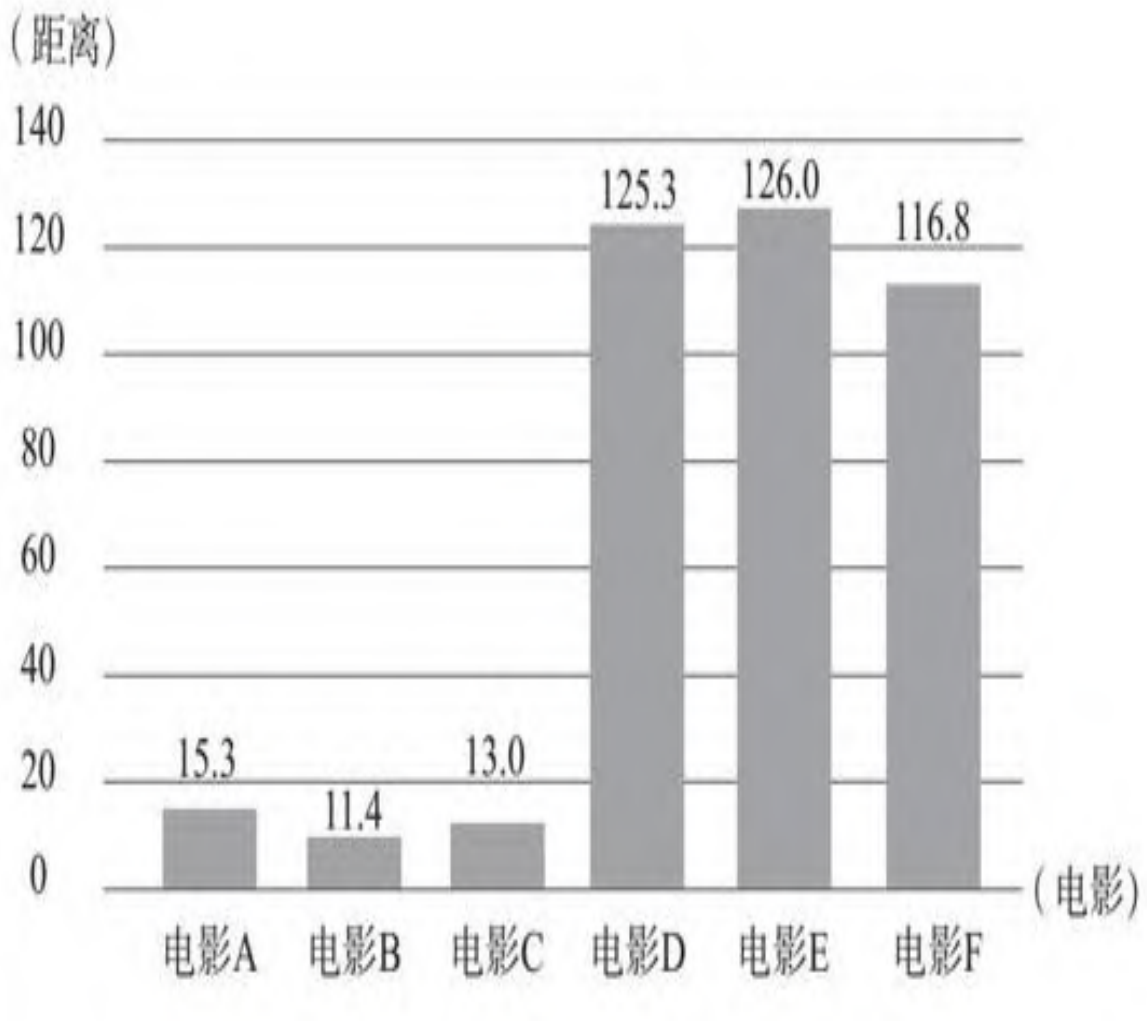


图4-8 电影距离数据统计图

按照距离的递增排序，可以找到K个距离最近的电影，假定K=3，则三个最靠近的电影依次是：电影B、电影C和电影A。所以未知类型电影的类型应该和离它距离最近的电影是一样的，所以它应该是爱情片。

除了K-近邻外，还有逻辑回归、支持向量机、决策树、随机森林和深度学习等模型。这些模型各有各的特色，就像是十八般武艺，了解之后你也能在实践中使用它们，做到游刃有余。最后，让我们再回到机器学习问题的本源，我们需要建立模型做出准确的预测，不同结

构的模型都是通过数据学习出来的。模型在实践中的预测效果以及对业务的提升，都需要相应数据的支撑。一旦模型学习好了，人工智能就会像人在经验中总结出规律一样，在面对新问题时，帮助我们做出正确决策。

这个世界对于我们来说充满了太多的未知，需要我们不断地去探索和发现。人工智能的出现让我们在面对新的事物、新的问题时不再不知所措。不论是针对有标记的事物还是未标记的事物，机器学习都可以有效地进行预测，最终形成决策并为我们提供参考，帮助我们有章可循地认识这个世界，改造这个世界。

AI感知图像和语音

在实际生活中，我们会遇到各种各样的数据，为了更好地对这些数据进行分析掌握，我们将其分为两类：结构化数据和非结构化数据。结构化数据就是能够用统一的结构加以表示的数据，例如数字、符号等。反之，非结构化数据是指那些字段长度可变，不能够用统一的结构加以表示的数据，如图像、视频、语音、文本等信息。非结构化数据与我们的生活息息相关，我们接下来就来了解一下人工智能是如何处理这些非结构化数据的。

前面我们已经提到过像佳丽的特征脸这种简单图像处理的例子，相信你已经有了一定的图像理解的基础。我们现在要从较为直观的人工智能理解图像开始，层层递进，更加全面深入地讲解人工智能理解非结构化数据的方法。

提到人工智能理解图像，就不得不说计算机视觉。计算机视觉是人工智能在视觉领域的应用，也就是说人工智能理解图像要通过计算机视觉来实现。1966年，人工智能先驱马文·明斯基让他的一个学生

“花一个夏天把相机接在计算机上，让计算机描述它看到的东西”。但明斯基低估了这个项目的难度，他认为只要一个夏天就能让计算机理解图像，但是之后又过了几十年，计算机还是不能很好地理解图像，直到现在，计算机也没有达到能够完全理解图像的高度。要想真正进入千家万户，计算机视觉仍然需要不断改进和提升。

回到人工智能理解图像的正题。图像就是像素值在二维空间的排列，像素可以用RGB（红绿蓝）三色值来表示，或者灰度值来表示。我们看到的图像无一不是这样类型的数据。虽然人眼能够很简单地理解图像内容，但这对人工智能模型来说却非常复杂。所以我们从以下几个方向对图像理解展开讲解：图像特征、图像分割、图像识别和图像中的目标检测等。

图像特征

图像特征主要包括图像的颜色特征、纹理特征、形状特征和空间关系特征。对图像特征的描述分为全局描述和局部描述。全局描述可以通过原始像素值特征、直方图特征、GIST特征这三种方法来实现，但是因为局部描述比全局描述更细致，所以下面我们主要来介绍局部描述。

局部描述是计算机视觉研究的一个基本问题，它主要用来寻找图像中的关键点，进行物体特征的数字化描述。提到局部特征，首先要想到的是SIFT特征。SIFT全名是尺度不变特征转换（Scale-Invariant Feature Transform），是一种用来检测与描述图像中的局部特征的计算机视觉模型，能够在空间尺度中寻找极值点，并提取出相应的位置、尺度、旋转不变量。尺度可以理解为图像的模糊程度，类似于眼睛近视的度数，尺度越大细节越少。不变量是指计算机视觉识别的图像中物体的本质特征。比如你拍同一个杯子，在不同的角度和位置拍出来的杯子大小以及杯子上花纹的角度位置都是不一样的，但是计算

机视觉识别的是杯子的本质特征，虽然位置、大小、角度都发生了变化，但是杯子本身没有发生变化。这个由加拿大教授戴维·洛（David Lowe）总结完善的特征计算方法，在计算机视觉的发展史上非常有影响，在很多涉及图像的人工智能应用里都可以找到它的身影。

局部图像特征的检测与描述可以帮助识别物体，我们在观察一个物体时，如果角度、距离甚至光线不同，我们观察的结果也会有些许误差。但SIFT特征是基于物体上的一些局部外观的关键点，与图像的大小和旋转无关，对于光线、噪声、轻微视角改变的容忍度也比较高。例如图4-9中的两张照片，它们是同一个物体在不同的视角下拍摄而成的，在这两张照片上存在SIFT特征的对应关系，对应的SIFT特征用线条连接起来。可以看出，SIFT特征能较好地描述图像内容，较少受干扰因素的影响。



图4-9 SIFT特征示意图

SIFT特征高度显著且相对容易提取，因而用它很容易辨识物体，而且少有误识。SIFT特征的本质是在不同的尺度空间上查找关键点，并计算出关键点的方向。SIFT所查找到的关键点是一些比较突出的点。这些点包括角点、边缘点、暗区的亮点以及亮区的暗点等，它们不易因光照、线性变换和噪声等因素的干扰而发生变化。在人工智能

里，这些点对物体的识别和检测非常重要。除此之外，使用SIFT特征描述部分遮蔽物体的检测率也比较高。

下面我们来具体讲解一下生成SIFT特征的过程。这个过程可以分解为四步。

第一步，尺度空间极值检测：搜索所有尺度上的图像位置，识别潜在的对于尺度和旋转不变的兴趣点。

第二步，关键点定位：在每个候选的位置上，通过模型来确定位置和尺度，根据每个候选位置的稳定程度来选择关键点。

第三步，方向确定：根据图像局部基于像素值变化的梯度方向（像素值增加最快的方向），给每个关键点分配方向。

第四步，关键点描述：在每个关键点周围的邻域内，在选定的尺度上计算图像局部的梯度，从而产生特征表示。

图4-10是SIFT特征产生的过程。左边显示了图像上一个关键点周边各像素点上的梯度，右边是根据梯度的方向分布生成的直方图，作为该关键点的SIFT特征向量。

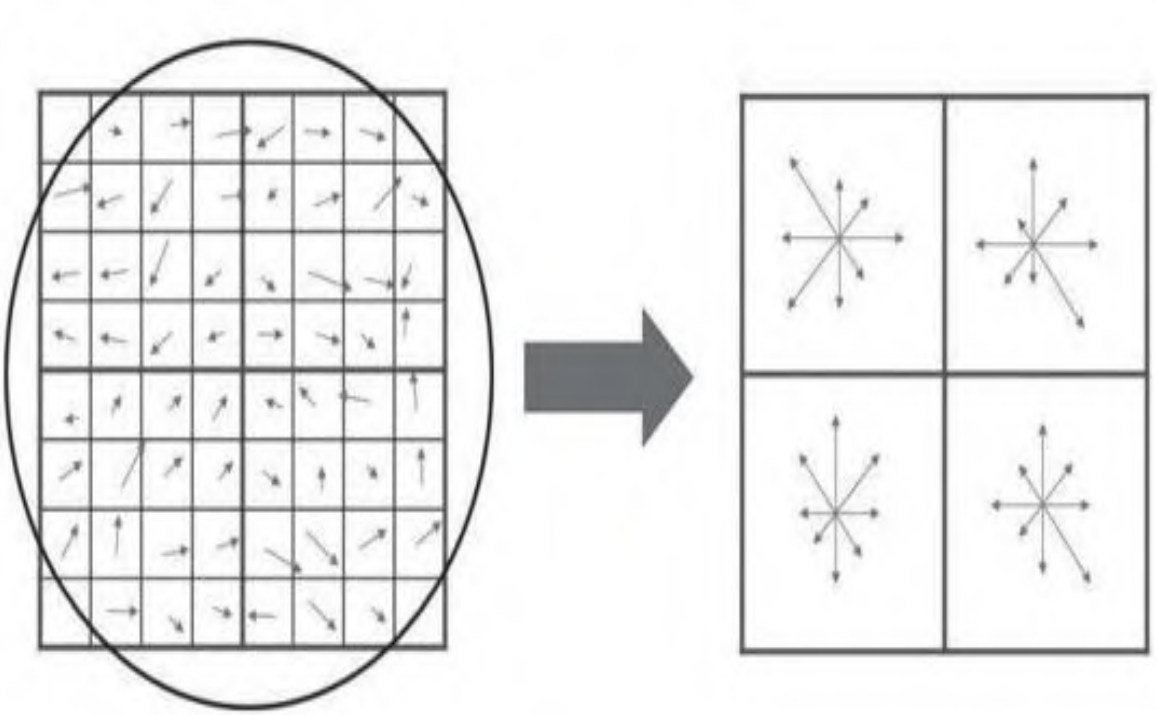


图4-10 SIFT特征产生示意图

人工智能里常用的图像的特征向量表示方法还有很多。但是，不论是哪一种表示方法，不论它表示的是全局特征还是局部特征，它都有自己的特点和适用范围。通过特征向量表示图像，从而捕捉全部内容相关的信息，是十分具有挑战性的。

图像分割

人工智能感知图像的一个重要应用是图像分割。图像分割就是机器自动从图像中分割出对象区域的过程，这样可以更加容易地获取图像中的内容和特征。图像分割在很多应用场景中发挥了重要作用。例如，在医学影像领域，医院在做肿瘤和其他病理的定位、组织体积的测量、诊断和治疗方案的定制时，都会用到图像分割。图像分割也被广泛应用于卫星图像分析领域，用来定位道路、森林等。此外，我们

熟知的一些图像处理软件，比如Photoshop、美图秀秀，它们的“抠图”功能，本质上也是通过图像分割实现的。

图像分割的一种基本方法是K-均值聚类，这种方法在前一节介绍聚类的时候提到过。其具体的过程是这样的：

第一步，从图像上任意选择K个像素作为初始簇中心。

第二步，对于剩下的像素，则需要根据它们与这些簇中心的相似度，将它们分配给与其最相似的簇。

第三步，计算每个所获新簇的中心，也就是簇中所有像素特征的均值。

重复第二步和第三步，直至收敛，也就是聚类结果不再发生变化。

如图4-11所示，右图是左边原图执行K-均值聚类（K=20）后的结果。实际应用中，为了提高速度，人们通常可以先对较大图像进行下采样，也就是均匀地减少像素，然后再进行聚类计算。

在图像分析中，直方图也是十分重要的。图像的直方图能够反映图像像素点灰度值的分布情况。基于直方图的图像分割方法通常会比其他类型的图像更快，因为只需要过一遍所有的像素。在这一过程中，需要先计算所有像素值的直方图，再通过直方图的峰和谷来找图像中的簇。如有必要，可以循环使用直方图来分割图像，找到更小的簇，直到没有新的簇为止。如图4-12所示，右图就是左图的直方图图像分割的最终结果。



图4-11 基于K-均值的图像分割结果



图4-12 基于直方图的图像分割结果

图像和视频理解

前面我们介绍了图像特征和图像分割，这些都是理解图像语义的基础，接下来我们就要讲解人工智能是如何识别图像的。前文我们说过人工智能要通过机器学习来产生模型，模拟人脑的深度学习就是机器学习的一种，它允许我们在给定一组训练数据的情况下来训练神经网络模型预测输出。近期的深度学习有了比较大的提升，相对较少地依赖于针对特定数据类型的特征提取，可以实现端到端的处理和学习。端到端处理就是在输入端输入数据后，输出端会直接输出结果，而不用经过复杂的特征提取过程。基于深度学习的人工智能现在已经能够较好地识别图像、视频、文本等非结构化数据，并且识别准确率很高，达到了可以在实际场景中进行应用的水平。尽管如此，理解不同数据类型对应的特征表示，仍能帮助我们更好地理解数据本身的规律和运用AI思维。

在人工智能对于图像的识别这一领域，近几年的趋势是通过深度学习来实现图像识别，例如可以用VGG、Inception和ResNet等深度学习模型来识别。图4-13所展示的就是几个通过这几种模型识别图像内容的例子。例如，最左边一列是使用VGG识别的图像，对于第一张图像，VGG认为有98.14%概率是足球，有1.16%的概率是橄榄球，有0.42%的概率是高尔夫球。由此可见，这些模型的正确率还是很高的。

图像中不可能只有我们所要识别的事物。当我们要识别一张照片中某个特定的人物时，这张照片中往往还有一些其他的东西，所以我们要进行目标检测。目标检测的任务是在图像上找到、定位并识别不确定数量的一类物体——需要强调的是，这里的“不确定数量”很重要，对人工智能来说，是有相当的挑战性的。跟图像识别一样，目标检测常用专门的深度学习框架，比如SSD（Single Shot Detector）、YOLO（You Only Look Once），但检测和识别是不一样的任务。通常，识别的结果是关于一幅图像的类别，而检测要回答的问题难度相对比较高。目标检测的应用场景包括人脸检测、人体检测、车辆检测、商品检测、遥感图像中建筑物检测等。如图4-14所示，目标检测模型能够检测出图像中的建筑物、行人和汽车。



图4-13 VGG、Inception和ResNet模型图像识别结果

几百年前，人们热衷于读书，甚至达到了洛阳纸贵的地步；十几年前，人们进入读图时代，读图成为风尚；而现在，视频成为人们的

宠儿，拿起手机看短视频成为越来越多的人的习惯。打开手机、电视看剧，去电影院看电影，我们每个人都是视频内容的消费者，而短视频的兴起又让我们都可以成为视频内容的创造者。视频在我们的生活中的地位慢慢上升，人工智能对于视频内容理解的意义和重要性也相应地不断提升。



图4-14 目标检测示意图

从理论上来讲，视频能够将一系列静态影像以信号的方式加以捕捉、记录、处理、存储、传送与重现。简单来理解，视频其实就是很多幅图像在时间上的组合，构成视频的每一幅图像叫作帧。用于测量

显示帧数的指标叫作帧率，其单位是每秒显示帧数（Frames Per Second），简称FPS。由于人眼特殊的生理结构，当所看画面的帧率高于16的时候，人们就会认为是连贯的，所以虽然我们看到的是一张张静止的图像，但是却有了动起来的体验。常见的视频编码格式有AVI、MPG等，在这些格式的视频形成过程中，其实都将视频的信息做了压缩。压缩后的文件变小，就更有利于处理、存储和传送了。

视频有哪些特征呢？首先，视频是由图像组成的，图像的特征也是视频特征的一部分。除了图像特征之外，视频中物体的运动也给我们的分析带来了新的特征。视频中常用的表示运动的特征叫作光流（optical flow）。光流反映了物体的运动模式，即反映了物体在某视角下由摄像头和背景之间形成的明显移动。通过视频中的图像序列模型，我们可以估计出运动物体在观察成像平面上像素运动的瞬时速度。

具体来说，光流模型可以计算两幅图像之间的变形，它假设一个物体的颜色在前后两帧没有巨大而明显的变化。基于这个思路，我们可以得到相应的图像约束方程。假定不同附加条件的光流问题对应不同的光流模型，常见的模型是Lucas-Kanade模型，如图4-15所示，我们在图片上显示了计算中的光流向量，向量的箭头表示的是像素运动的方向。

当我们检测出光流向量之后，就可以进一步生成HOF特征。HOF的全称是Histograms of Oriented Optical Flow，它可以对光流方向进行统计，得到关于光流方向分布的直方图。这个直方图表示了在不同的二维方向上，各分布了多少光流向量。由于HOF能够表示出视频中物体运动的本质特点，所以经常用于视频中的动作和事件识别。



图4-15 光流示意图

通过光流向量和HOF特征对视频处理后，我们就可以通过人工智能识别出视频中的物体的运动了。但是在视频中，我们要识别的目标一般是运动的，如果要对这个目标进行识别，就需要对其进行跟踪。目标跟踪是指给出目标在跟踪视频第一帧中的初始状态，比如目标的位置和尺寸，自动估计目标物体在后续帧中的状态。如图4-16所示，视频中跟踪的目标通常会用一个矩形的框来标注。

如果视频中有我们感兴趣的人或物体，可以用目标检测的方法先检测到它，然后再跟踪。这样，我们就不止在一帧中获得了这个人或物体，还能一直跟踪他在哪儿。这在商业场景上大有用处，比如超市想知道消费者的行为轨迹，就可以采用检测和跟踪相结合的方法。商场公共区域都设有摄像头，通过摄像头记录下来的视频，人工智能可以识别出消费者在某家店停留的时间长短和相应的行为，然后估算消

费者的消费概率。商家通过人工智能知道了消费者在商场里的完整行为轨迹后，可以做出许多帮助提升体验、促进消费的优化，比如在合理的位置设置休息区，以供消费者休息，这样既可以消除消费者的疲劳，又可以延长其在商场里停留的时间，购买商品的概率也会有所上升。

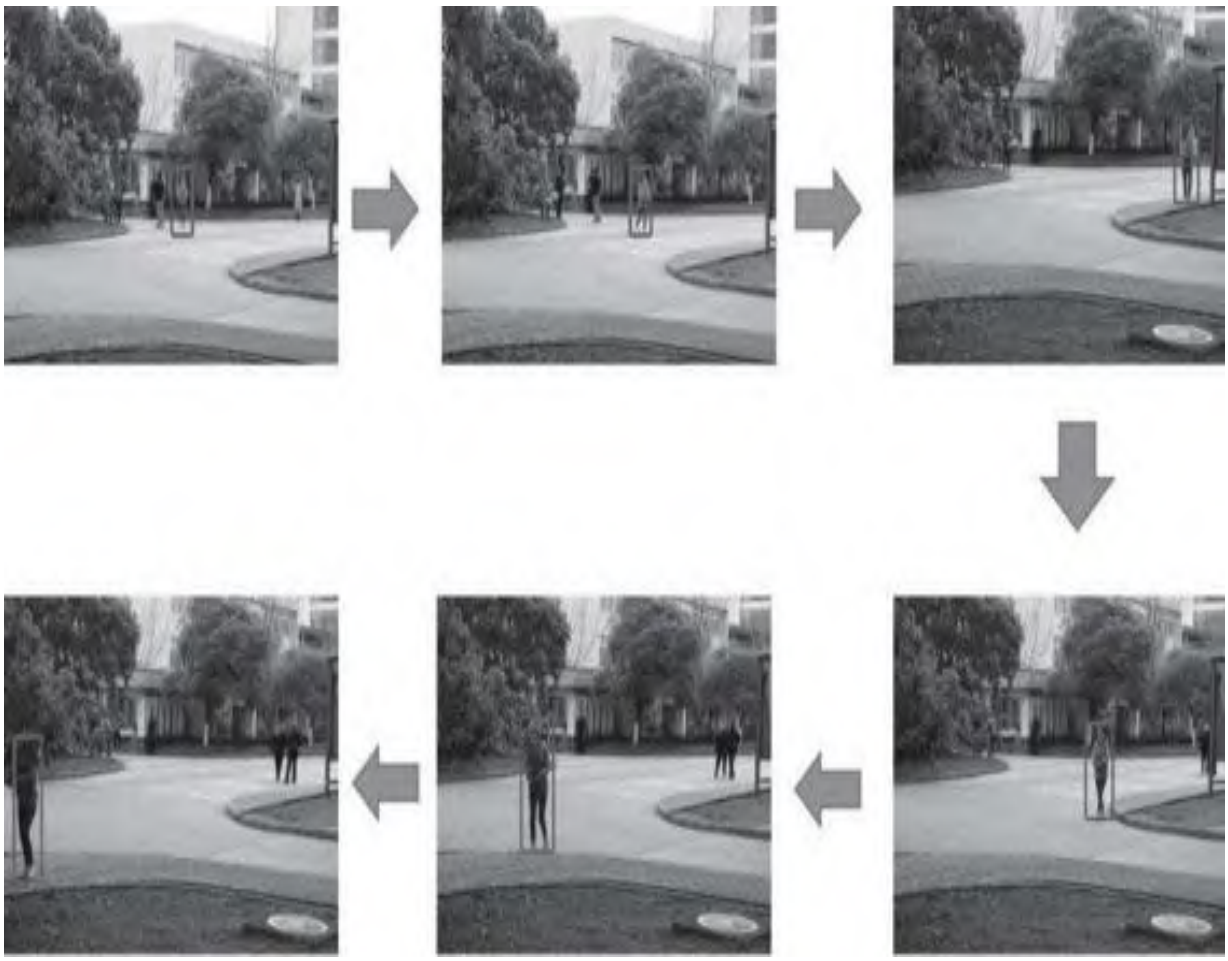


图4-16 目标跟踪示意图

很多人会通过看电视剧或者电影来打发时间。在看电视剧或电影的过程中，我们能够轻而易举地分辨出电影电视剧中的主角和反派。近几年来，随着人工智能的视频识别不断发展，人工智能也变得和人类一样，在看电影和电视剧的时候，能够分辨出哪些是“好人”，哪

些是“坏人”。并且通过这个功能，人工智能还可以对视频中的人物进行社交分析。

那么，人工智能是怎样做到从视频中分析人物之间的社交关系的呢？首先，通过观察可以发现：在与不同的人的交往过程中，人们选择的交互方式也是不同的。比如说，在日常生活中，一个人与比较亲密的同事、恋人之间的交互方式，和与疏远、有敌意的人之间的交互方式，往往是不同的。所以我们就定义了一个指标，叫作结合信号（grouping cue），它能够决定不同人物之间更可能属于何种关系（如亲密或疏远，友善或敌意等）。

然后，在视频中的每一个场景中提取相应的特征向量，通过有标记的训练数据和机器学习模型，学出人物间的结合信号，然后通过模型把这些结合信号整合起来，形成一个人物间的社交网络。有了社交网络，就可以分析人物中有几个社区和每个社区的领袖是谁。人工智能对电影和网络视频中的人物进行社交分析后，能够分辨出视频中有几组人，每组人的领袖是谁。如图4-17，在电影《加勒比海盗：黑珍珠号的诅咒》（左上、右上）中，左边是善良的一派，领袖是杰克·斯伯洛；右边是阴险的一派，领袖是巴伯萨。在电影《哈利·波特》（左下、右下）中，左边的是正义的一派，领袖是哈利·波特；右边是邪恶的一派，领袖是伏地魔。











图4-17 视频中人物的社交分析示意图

语音理解

每天我们都听到的声音，在计算机中是以时间序列存储的。也就是说，在计算机中，声音是一串随着时间变化的数字，这串数字代表了在一个时刻声音信号的强弱。就像理解图像和视频一样，人工智能理解音频首先要提取特征。音频的特征主要包括：时域特征和频域特征。时域特征比较好理解，就是在信号时间序列上加工得到的对其的直接描述，包括信号幅度和信号能量等。其中，信号幅度是指信号上下波动的范围，信号能量一般与信号的振幅和持续时间有关。

要想理解频域特征，就不得不提傅立叶变换。傅立叶变换是一种分析信号的方法，它可以分析信号的成分，也可用这些成分合成信号。傅立叶变换，能将满足一定条件的函数表示成不同频率正弦或余弦函数的线性组合。在不同的领域，傅立叶变换具有多种不同的变体，例如连续傅立叶变换和离散傅立叶变换。音频信号的傅立叶变换为我们提供了一种极佳的表示形式。

计算频域特征需要先把信号进行傅立叶变换，然后才能生成频域特征。频域特征主要包括基础频率、频率组成、频谱中心、频谱流量、频谱密度等。因为不同的正弦或余弦函数对应了不同的频率，而频域特征是对音频的傅立叶变换的描述，所以被称作“频域”。频谱是频率的分布曲线，可以表示一个信号是由哪些频率的弦波组成。

跟图像一样，音频数据也可以切分，上述音频特征就可以用在音频分割中。与直观的图像分割不同，音频分割有两种类型：第一种是在时间上将音频切分成几段，每段包括了特定的声音信息，例如对话、音乐、动作声响、环境声响等，这是时间上的分割；第二种是把

音频中同一时间叠加的声音分割开，产生若干个并行的音频流，分别对应各自的声源。这种分割对应的是音频信号在各个时间和频率上的分割。

音频理解的一个广泛应用是语音识别：亚马逊的Alexa、小米的人工智能音箱“小爱同学”、机器人玩具、自动客服等场景中都在使用。那么，语音是如何转成文字的呢？图4-18显示了语音识别的全过程，这一过程通常有以下几个部分。

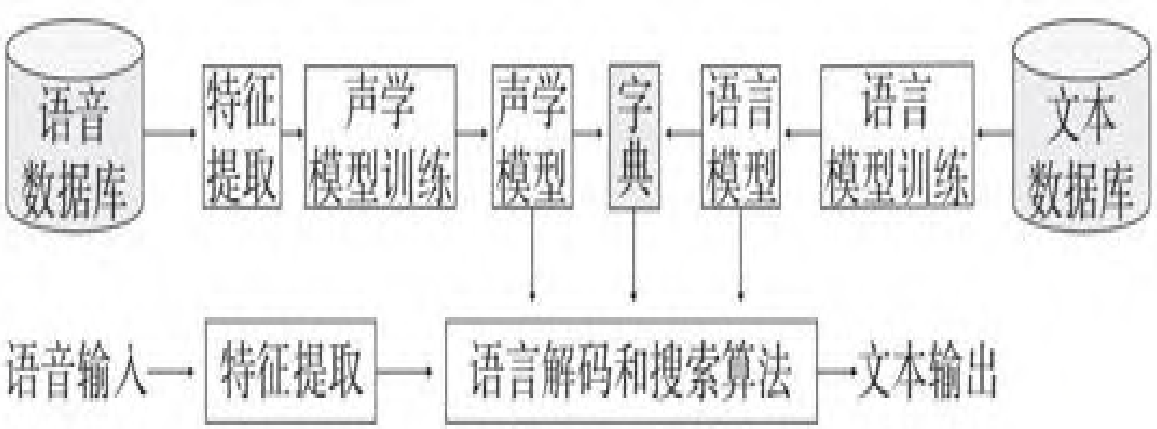


图4-18 语音识别过程示意图

首先，是声学模型，声学模型是语音识别系统的重要组成部分。声学模型需要从语音数据中提取MFCC（梅尔倒谱系数）特征。对人耳听觉机理的研究发现，人耳对不同频率的信号有着不同的听觉敏感度，其中，从200Hz到5 000Hz的信号对语音的清晰度影响最大。并且，对于人耳而言，低频音容易掩蔽高频音，高频音不易掩蔽低频音。基于这样观察得到的特征，MFCC特征表示，在噪音较高的环境里，MFCC特征仍然具有较好的语音识别性能。根据MFCC特征建立的GMM-HMM模型（高斯混合—隐马尔可夫模型），可以作为语音识别所需的声学模型。其中GMM用于对语音声学特征的分布进行建模，HMM则用于对语音信号的时序性进行建模。这组GMM和HMM结合的模型，不但能

捕捉单个MFCC特征，并且能够体现MFCC在时间上的变化，从而模拟出人耳对声音的感知。

其次，是语言模型，可以通过训练文本数据学习N-gram模型（N元模型）。不论在哪一种语言中，词的出现都是有统计规律的，比如当我说“柳暗花明”这个词时，你想到的下一个词是什么呢？我想大家很有可能会想到“又一村”，基本上没有人会想到“又一花”或者“又一树”吧。N-gram模型的主要思想就是这样的，它能够根据统计规律预测出一个词的前后内容。常用的是建模两个连续词的Bi-gram模型和三个连续词的Tri-gram模型，分别对应二元和三元的N-gram模型。有了这些模型之后，给定一个语音的输入，在生成特征后，通过声学模型、语言模型和已知的字典（也就是我们有哪些候选词作为输出），进行计算和搜索，语音识别系统就可以产生相应的文本输出。

目前，在环境安静、发音标准的情况下，语音识别模型已经完全可以对语音进行识别了，而且识别的准确率很高。但是这种准确率是有条件的。不论哪一款语音识别软件，如果将其放置在一个嘈杂环境下，或者当使用者的普通话不标准、带有口音时，语音识别的准确率都会大打折扣。提高在不同环境下通用语音识别的准确性，是语音识别一直以来的发展方向。

我们赞美天的湛蓝，海的宽广；我们歌颂诗的美妙，曲的动听。何其有幸，我们一出生就拥有看见、听见这个世界的机会。我们拥有感知这个世界的能力，人工智能因我们而生，我们帮助它更好地去体会这个世界。现在，海量数据驱动下的图像识别、视频识别、音频识别给了人工智能这个机会。也许人工智能并不理解在一层层模型下产生的技能到底代表了什么，但却能够帮助我们真正地实现“眼观六路，耳听八方”。

AI理解自然语言

自然语言通常是指一种自然地随文化演化的语言，例如汉语、英语及其文字。自然语言可以表情达意，是人类沟通交流的工具，同时也能促进人类思维的发展。对自然语言的理解是人工智能的一门重要功课。同时由于自然语言的复杂性和多样性，对于自然语言的理解和处理也充满了挑战。但是我们在日常生活中并不常用“自然语言”这个词，自然语言是一个抽象的概念，我们一般从它的具体形式——文本，展开研究。

要想理解文本，首先就要理解文本的特征。BOW特征可以帮助我们更好地理解文本特征。BOW全称是Bag of Words，也就是词袋模型，它是自然语言处理领域常用的一种文本表示方法。在使用BOW特征的过程中，需要将每篇文章都看成一个“袋子”，这个“袋子”里面装的都是词汇，所以称之为词袋。BOW最初是用于文本分类的，它通过将文本表示成特征向量来实现文本的分类。它的基本思想是忽略文本中的词序、语法和句法，只将文本看成是词汇的集合，而且文本中的每个词都是独立的。举个例子，有如下两个文本：

文本一：Alice likes to play basketball, Bob likes too.

（艾丽斯喜欢打篮球，鲍勃也喜欢。）

文本二：Charlie likes to play tennis.

（查利喜欢打网球。）

基于这两个文本，构造一个词典：

词典={1.“Alice”，2. “like”，3. “to”，4. “play”，5.“basketball”，6. “Bob”，7. “too”，8. “Charlie”，9. “tennis”}。

这个词典由9个不同的单词组成，利用词典的索引号，上面两个文本分别可以用一个9维向量表示，下列整数表示的是某个单词在文本中出现的次数，见表4-2：

表4-2 单词在文本中出现的次数

词典	Alice	like	to	play	basketball	Bob	too	Charlie	tennis
文本一	1	2	1	1	1	1	1	0	0
文本二	0	1	1	1	0	0	0	1	1

在上述例子中，向量中每个元素表示词典中相关元素在文本中出现的次数。但是通过上表我们可以发现，在构造文本向量时，并没有体现单词在原来句子中出现的次序。这是BOW表示的缺点之一，不过这并不影响人工智能对于这一文本的理解。试想，如果一篇文本中人群、大厦、银行、商场这样的词汇多些，而牛、羊、山谷、稻田这样的词汇很少，即使没有按照顺序排列，我们也能做出一个大致判断，得出它是一篇描绘城市的，而不是描述乡村的文本。

假定现在有一个很大的文本集合，里面有许多的文本，我们将文本里面的所有单词提取出来，构成一个包含N个单词的词典。通过BOW特征，每个文本都可以被表示成为一个N维向量，而人工智能非常擅长于处理数值向量。这样，我们就可以通过BOW特征来完成海量文本的分类以及其他任务了。

但是，BOW特征有一个缺点，就是它无法反映出一个句子的关键词。比如上面的文本一中，“like”出现次数最多，若根据BOW特征提取这个句子的关键词，则为“like”，但是这个句子中的“basketball”也很重要，这是BOW无法准确提取出来的。我们用TF-IDF可以很好地解决这个问题。

TF-IDF也是一种常用的文本特征的表示方法，它的主要思想是：如果某个词在一篇文章中出现的次数高，并且在其他文章中很少出现，则认为这个词是该文章的关键词，通过关键词就可以很好地对现有文本和其他文本进行区分。在TF-IDF中，TF是词频，term frequency；IDF是逆文档频率，inverse document frequency。在这个特征中，需要将词频和逆文档频率相乘得到对应词的特征值。

TF表示词条在一个文本中出现的频率，在一个文本中出现次数越多，这个词的TF值就越高，就说明它对该文本越重要。IDF反映了词条在所有文本中出现频率的逆向值，如果一个词在很多的文本中都出现了，那么它的IDF值就会低。相反，如果一个词很少在文本中出现，那它的IDF值就较高，说明这个词具有较强的代表性。两者的乘积可以用在特征向量中作为数值来表示一个词，从而改进BOW形式的表示。由于TF-IDF能够准确找出文本中的关键词，所以特别适合用来进行文本相似度对比、文本分类、信息检索、关键词提取以及文本摘要。现在很多搜索引擎、论文网站、新闻网站都应用了TF-IDF特征。比如，当你想要了解某些信息时，你会在搜索引擎上输入关键词，而搜索引擎之所以可以帮你找到答案，是因为它应用了TF-IDF特征来计算它检索出来的文本和关键词之间的匹配度。如果它发现这个关键词在某个文本中出现的次数最多，而在其他文本中很少出现，那这个文本就是你需要的答案。

词嵌入

词嵌入，也就是word embedding，它也是表示文本特征的一种方式，可以将自然语言表示的单词转换为计算机能够理解的向量形式。简单点来理解，我们都知道文本有千千万万，文本构成的维度也不计其数，所以BOW会产生很高维度的特征向量，而词嵌入将词映射到一定维度的连续向量，就能使得这个连续向量的维度比较低，例如几十到几百。产生词的向量之后，我们就可以进行这些向量间的运算，比如通过向量之间的相似度来度量不同词语之间的语义相关性。

词嵌入基于的假设是：如果两个词经常在相同的上下文中出现，那么这两个词的意思应该相近。所有学习词嵌入的方法都是在建模词和上下文之间的关系。如图4-19，我们举了些词嵌入的例子。相似的词被映射成相近的向量，差别大的词映射以后对应的向量差异也比较大。所以“足球”与同样是球类的“篮球”距离很近，而与形容味道的“好吃的”距离很远。由此可见，通过词嵌入，词之间语义相似性可以被比较恰当地编码成向量了。

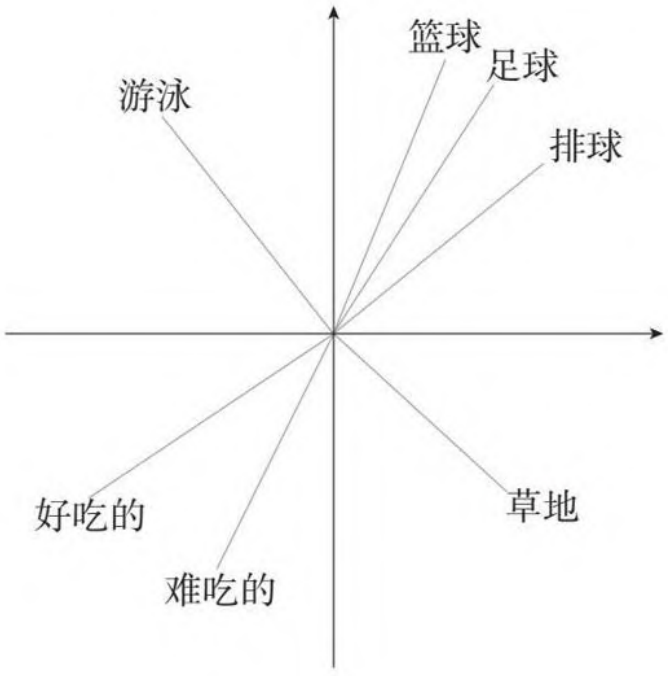


图4-19 词嵌入示意图

近年来，智能手机的普及，使得网络信息迅速增长而且更加碎片化，微信消息、微博、商品评论等短文本大量出现，并逐渐成为网络文本信息的主流形式。短文本的稀疏性、不标准化、数据规模大且标注少等特征，给传统的基于长文本的数据处理带来了新的挑战。而词嵌入具有两个作用：维度缩减和语义扩充。词嵌入的维度缩减消解了短文本不标准、规模大的难题。词嵌入将语义相似的短文本投影到向量空间中的同一区域，这样产生的语义特征又可以作为补充信息，解决短文本语义稀疏、难以理解的局限性。词嵌入的应用提高了短文本分类的精度，为人工智能理解现在大量泛滥的短文本信息提供了帮助。

文本分类

文本分类是对文本内容的识别。通过我们上文提到的文本特征化方法，再结合机器学习方法，就可以实现对文本数据的分门别类，例如每天发生的新闻铺天盖地，但是利用文本分类我们就能迅速地将其归为娱乐类、体育类、经济类新闻等。我们每天所接收到的信息也十分泛滥，比如每天都会收到来自四面八方的邮件，里面不乏垃圾邮件。这时，文本分类就派上了用场，一般的电子邮箱都会有文本分类的功能，将收到的邮件分为正常邮件和疑似垃圾邮件两类。

作为一种特殊的文本分类，情感分析是对带有感情色彩的主观性文本进行处理和分析的过程。情感分析可对不同文本进行处理，其中包括基于新闻评论的情感分析和基于产品评论的情感分析。新闻评论是指在关于新闻热点事件的评论中，发表观点的人包括社会中各个层次的人，不同层次的人表达的情感态度或观点不同，这能够反映出很多信息。对于新闻评论的情感分析，可以掌握大众对于新闻事件的思想动向，有利于政府和新闻媒体机构舆情监控工作的展开。

产品评论表达了消费者对一件商品各种功能的满意程度，是一种重要的市场反馈信息。通过基于产品评论的情感分析可以了解某一产品在用户心目中的口碑和用户的关注点等，为产品销售提供科学、可供决策的信息，而且有利于指导商家改善商品，提高用户体验。

例如现在要对某手机的产品评论做出情感分析。这个分析是从用户评论数据出发的，但是这些产品评论都是用户自由发表的，不能保证每一条都有意义，所以我们要对数据进行清洗，以此来保证获取的商品信息的质量。数据清洗后我们就要对评价对象和评价观点进行分析。对于手机来说，评价对象就是价格、系统、处理器、内存这些内容，当然也包括快递、售后等其他服务。评价观点一般分为积极、消极、中性这三类。例如“人脸识别和指纹识别速度非常快”可以标记为积极评论，“不好用，手机烫得要命”可以标记为消极评论，“也没有说的那么好，没啥特别的，也就还行吧”就可以归入中性评论。然后统计每一个评价对象所对应的三类评价观点各有多少，然后我们就可以对这些评论进行情感分析。比如该手机总评论数最多的对象是“外观”，这说明用户在购买手机时十分注意手机的外观，商家在开发新的手机产品时可以重点提高手机的美观度。假如用户对该手机外观的评论有80%是积极的，那说明用户对于该手机的外观比较满意，这是该手机的一个优势，商家在开展营销时可以以此为重点。

开发新产品时，也可以借鉴该手机外观上的一些特质，保持用户对手机品牌的满意度，相反，如果对该手机外观的评论有80%是消极的，就说明用户对该手机的外观不满意，这是该手机的劣势，那商家在开展营销时就要学会避重就轻，重点突出该手机的其他优势，开发新产品时还要及时改进。同理，对于其他评价对象的情感分析也是这样进行的，这样就可以全面地掌握这款手机的市场反馈信息，为商家不论是开展营销活动，提升用户体验，还是后续研发新产品提供参考。

文本情感分析还可以用在哪里呢？文本情感分析还可以预测股票价格的变动。这其中的奥秘是，人工智能可以从社交媒体中获取大众情绪变化的特征，利用这些特征，可以通过情感分析大致预测出大众情绪的变化。而事实上，公众情绪的变化又能在一定程度上影响股票价格的变动。就是说，文本情感分析可以根据大众情绪的波动预测股票价格的变动。

论文《推特情绪预测股票市场》（Twitter Mood Predicts the Stock Market），讲的就是公众情绪对股票价格变动的影响，由于其思维的先进性和研究的科学性，被大量引用。推特在美国是类似于微博这样的平台，用户可以在这个平台上记录分享自己每天的故事和情绪。这篇论文的主要研究思路是利用推特用户发表的内容数据，通过两种情绪分析方法OpinionFinder和GPOMS（谷歌情绪状态简况），来抓取和分析公众的情绪变化。其中OpinionFinder是将人的情绪区分为正面和负面两类，而GPOMS将情绪分成六类，分别是平静（calm），警觉（alert），有信心（sure），活力十足（vital），友好（kind）和高兴（happy）。通过分析，文章发现公众情绪与股票价格指数有一定相关性。因此，该文作者大胆假设：公众情绪的某些指标可以有效地预测股价的变动。他们对此展开了一系列的研究，在一个复杂神经网络模型的基础上，将公众情绪的时间序列输入到该模型中，得到的预测效果比较理想，其预测股票价格指数收盘价涨跌方向的准确率高达86.7%。

随着人工智能的普及，基于文本分析的应用涉及各个领域，而文本数据在日常生活中又随处可见，比如你随手拿起的一本书，你无意中瞥到的广告词，你无聊时刷到的朋友圈，都可以是人工智能理解的文本数据。机会都是留给会思考有准备的有心人的，所以你在平时也可以想一想，你所了解和接触到的这些文本数据，能不能用起来，可以产生怎样的实际价值，说不定就可以成为人工智能的下一个风口。

实体识别

随着互联网的高速发展，人们获得信息的渠道越来越多，获取信息的方式日益简单。但是，随之而来的是信息的数量呈指数级增加，而且混乱无序，真假难辨。要想仅仅使用人力从海量的文本数据中获取所需要的信息，也变得难如登天。人工智能的出现解决了这一难题。人工智能通过一系列自然语言处理模型能够提取出复杂文本数据中的关键信息，而自然语言理解的关键步骤就是实体识别。

实体识别又叫命名实体识别（Named Entity Recognition，NER），是指识别文本中具有特定意义的实体，比如说人名、地名、产品名、属性名、日期、数字等。实体识别要做的就是从非结构化的输入文本中抽取出上述实体，并且可以按照业务需求识别出更多类别的实体。在应用过程中，实体识别的确切含义需要根据实际情况来确定，比如，在面向生物的实体信息抽取时，还包括细胞、蛋白质、基因、核糖核酸、脱氧核糖核酸等特殊生物实体。

实体识别在信息提取、问答系统、句法分析、机器翻译等领域应用广泛，在人工智能自然语言处理模型的应用过程中具有重要意义。例如在机器翻译中，有些人的名字在翻译成英文时把名放在前、姓放在后，而有些人的名字就直接用中文的汉语拼音来表示。实体识别可以准确识别出文本中是哪一种形式，大大提高了机器翻译的准确性。作为信息检索系统的一种高级形式，问答系统能用恰当的自然语言回答用户用自然语言提出的问题。当用户提出问题时，问答系统能够对可能的答案进行实体识别和深度分析，然后将最适合的答案呈现给用户，为用户营造一个信息丰富、答案准确、反应迅速的信息获取环境。

由于数量、时间、日期、货币等实体具有统一形式，我们通常可以对它们用模式匹配的方式取得较好的识别效果。但相比之下，人名、地名、机构名较复杂，识别难度很大，普遍存在嵌套和缩写的识

别问题。例如“清华大学第一附属医院”这个组织机构名称里面还嵌套着同样可以作为组织机构名称的“清华大学”，再比如，有时人们会把“河北大学”缩写为“河大”，但是“河南大学”的缩写也是“河大”，这就对进行实体识别造成了很大的困难。而且新词汇也在不断涌现。例如随着互联网的发展，一些没有任何规则的网名昵称也为实体识别的进行带来了新的挑战。因此，近年来的实体识别主要以人名、地名、机构名这几种比较复杂的实体识别为主要的突破方向，由原来的单独针对人名、地名等进行识别的方式，发展到用更优化的方法同时进行各类实体的识别。近几年来，实体识别的效果有了显著提高，在各个领域得到普遍应用。

比如说，实体识别在旅游行业有着很大的发挥空间。实体识别在旅游行业的主要应用是对旅游景点的识别。例如，游客可以通过在旅游网站上检索来了解景点的信息，这个过程就涉及了对旅游景点的实体识别。目前对于旅游景点识别的困难主要有：旅游景点数量众多，建立旅游词库工作量大。据不完全统计，仅北京这一个城市的旅游景点就有6 000多个，幅员辽阔、地大物博的整个中国大地，旅游景点更是不计其数。旅游景点名称没有固定的格式和规律可循，给实体识别造成了困难，比如北京地安门大街附近的一个湖叫作“后海”，天坛公园里有一个小的景点叫作“皇穹宇”；景点名称经常采用缩写形式，也提高了实体识别的难度，比如“八达岭长城”“居庸关长城”都简称为“长城”。但现在的实体识别模型已经可以很好地解决这些难题了。还是以北京为例，我们首先需要构建一个北京旅游景点词库，词库建立之后经过一系列去重操作和人工审核，得到有效的景点词表。在此之后，通过该词表对旅游文本进行标注，并构建模型训练集和预测集。通过对训练集的反复学习训练，最终，对旅游景点实体识别的预测准确率可以做到很高，有人通过上述方法得出了95%的准确率。这样，游客想要了解各种旅游景点时就不会出现检索不到的情况了。

除此之外，实体识别在新闻传播领域的应用也很有前景。由于新闻报道会涉及体育、财经、科技、时尚、教育等社会生活的各个领域，所以对涉及领域的常见专有名词都需要有一定的识别精度。此外，新闻报道都具有时效性，所以也需要关注与热门资讯和突发事件相关联的实体，例如2019年5月，华为注册了“华为鸿蒙”商标，以及之前大火的电视剧《陈情令》等，这些新产生的、具有很大热度的实体都是需要识别的。现在的实体识别能够自动识别那些未被收录到词表中的词，并且从中发现新词，不仅满足了新闻传播领域精准识别实体的需要，还能帮助媒体行业进行更精准的内容挖掘、对内容进行更有效的监督管理，推动了新闻朝智能化方向发展。

我们已经讲过了人工智能对于图片、视频和文本这些非结构化数据的识别。在人工智能出现之前，这些数据就似汪洋大海一般一直存在，但人工智能的出现，才真正赋予了它们价值。AI思维就是这样，在已有数据基础上挖掘其未竟的巨大的隐含价值，给予我们做出更好判断和决策的能力。人工智能对于非结构化数据的理解是AI思维的一个重要组成部分，我们只有充分理解了这些数据处理的方法论，才能更好地利用人工智能支持决策。

AI理解商业数据

前面我们讲到了人工智能可以理解图像，具有人脸检测、商品检测等功能；可以理解视频，具有动作识别、社交分析等功能；可以理解语音和自然语言，具有语音识别、文本分类、实体识别等功能。这些都是从具体的某一类型的数据出发的，人工智能对这些非结构化数据的理解是为了更好地为我们的实践提供帮助。下面我们要具体讲解人工智能是如何通过其理解和学习能力直接为商业实践赋能的。

商业实践的基础是商业数据。所谓商业数据，是指在一个行业内各个环节中涉及数据的集合，包括企业内部数据、分销渠道数据、消费市场数据等。通过人工智能对商业数据的理解和分析，我们就可以知道在这个行业中，用户的需求是什么，哪些需求已经被满足了，哪些需求还没有被满足，整个行业还有什么“处女地”没有被挖掘，行业的创新点和决胜点在哪里，以及自己公司的优势、劣势和独特点在哪里。只有这样，我们才能准确地进行市场细分，确定目标用户，制定更加个性化、精准化的营销方案；也只有这样，我们才能准确地进行产品定位，提供让客户更加满意的服务。

人工智能可以通过SIFT特征等对图像进行理解，可以通过光流模型等对视频进行理解，可以通过N-gram模型等对语言进行理解，那么人工智能是如何理解复杂的商业数据的呢？

你被标签化了吗？

我们在注册一个新软件时，软件会让我们勾选一些标签来更好地了解我们的喜好。在许多平台发文章前，平台也会让我们为文章勾选标签，因为这样可以帮助我们更好地定位内容与读者。我们每天都在被贴标签，我们也在给别人贴标签。当我们谈论起80后时，我们总认为他们是“叛逆”的一代；当我们谈论起90后时，我们觉得他们是“问题青年”；当我们谈论起00后时，我们说他们是“新新人类”。不管是哪个年代的人，我们都对他们有相对应的标签，甚至连“80后”“90后”都成了一种标签。这些标签都是人们根据他们的性格特点、人生经历做出的总结概括，在一定程度上反映了真实现象。有了这些标签，我们就可以更好地理解对方，甚至被贴上相同标签的人还能迅速拉近距离。现在很多人也通过贴标签的方式被大众熟知，比如网络上有的人因为和一些知名演员长得相像，就给自己贴上诸如“小林志玲”“小吴彦祖”等标签，迅速提升了知名度；还曾有某歌手凭

借“四千年一遇的美女”的标签登上热搜成为“国民女神”。人工智能也是通过为数据打标签的方式来理解商业数据的。

标签是一种用来描述业务实体特征的数据形式。比如在描述一个用户时，我们会使用性别、年龄、地区、兴趣爱好、产品偏好等标签来反映该用户的特征。商业数据标签的生成过程以发生在时间和空间中的行为为媒介，将物的特征传递到人，人的特征传递到物。在这个过程中，物包括实物、商品、文本、图像等。标签传递过程形成了一个由人和物为节点，以行为为边的关系网。如图4-20所示，在很多实际商业场景下，需要把人打上标签，比如他经常和一些讲诚信的人合作，这样的社交行为多了，他也会被打上“诚信”标签；若他和一些经常毁约的人交往比较多，那和他合作就有一定的风险；他平时购买的运动用品比较多，他就会被贴上“运动”的标签；又或者是他平常喜欢收听科技类音频课的话，他就会被打上“科技”的标签。除了对人打标签外，对物也需要打标签。物包括自然语言的语料，比如说一部英文原声大片，其台词转录成文本以后，也可以打上标签。

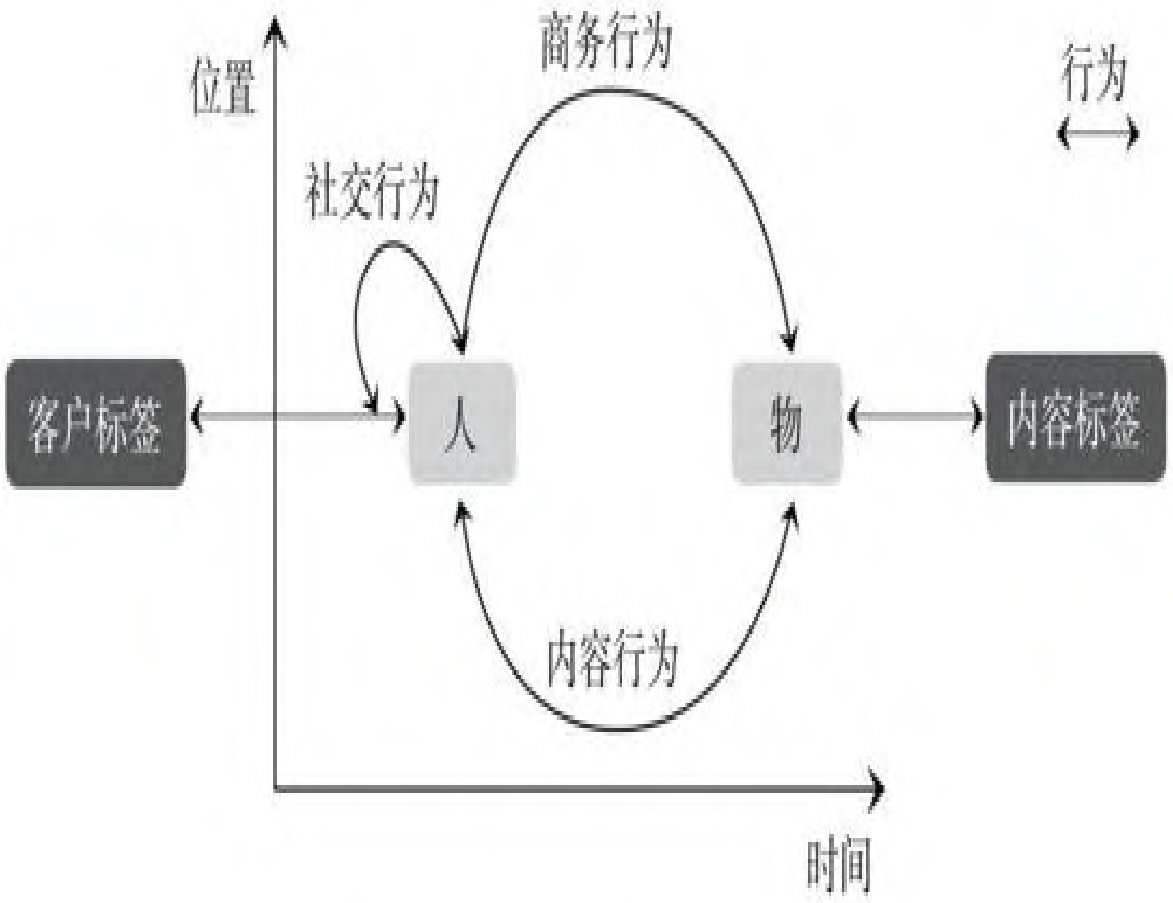


图4-20 商业数据标签生成过程示意图

数据标签化

为数量巨大的商业数据打标签是一个庞大的工程。它有什么必要性呢？首先，根据上述分析我们知道，数据标签化，无论在理解人的属性行为还是理解对应内容的属性上，都是一套非常重要的方法论。其次，在信息爆炸的现在，商业数据的体量也是巨大的，而且它们结构十分复杂，分布在许多不同的数据库中。在实际场景中要应用这些数据时，我们发现很难把这些规模巨大结构复杂的数据直接分析或者用于模型训练。但是在实践中发现，通过分类，可以将杂乱无章的数

据条理化，将数据标签化其实就是一种分类。标签化的数据结构更加简单，而且非常容易管理。标签本身就是在分析系统中产生的，所以具有较好的兼容性，可以直接作为训练数据导入训练模型，减少建模的数据准备时间，也极大地提高了数据使用效率。所以，不管是从人工智能理解数据的难易程度还是数据使用效率角度，数据标签化都有优越性。

标签通常是人为规定的高度精练的特征标识，如年龄、性别、兴趣爱好、产品偏好等，将这些标签集合在一起就能抽象出一个用户的信息全貌。首先我们要知道，标签体系基本上都是层级化的。每个标签体系都分为几大类，每个大类底下分别有几个小类，逐层分布。在标签体系中，最高层级的那一级标签称为“一级标签”，以此类推，底下是“二级标签”“三级标签”等。底层标签一般能捕捉到更细的特点信息，所以经常用于精准的广告投放和营销活动，比如说某个用户被贴了一个三级标签“购物方式—支付方式—信用卡”，在这个标签中，信用卡是底层标签，说明了该用户在购物时经常使用信用卡支付，信用卡对他的生活很重要，所以此时信用卡相关的广告投放对他就很有吸引力，又或者我们可以对应他信用卡支付的习惯向他推荐各种分期付款或者刷信用卡支付的商品。这样就满足了该用户的需要，还迎合了他的习惯，对提高广告投放效果和营销的精准度有很大作用。一般而言，每一个标签都只有一个含义，这样就不会发生同一层级的标签重复或者冲突的现象了。另外，信息粒度是指信息的相对大小和粗糙程度，不同层级的标签对应的信息粒度不同。一般来说，底层标签的信息粒度会相对更细。下面是一些常见类别下的细粒度标签。

表4-3 常见类别下的细粒度标签

类别	标签
人口特征	年龄、出生日期、性别、地域、教育程度、职业、星座等
兴趣特征	兴趣爱好、浏览内容、收藏内容、使用App类型、使用网站类型、品牌偏好、产品偏好、阅读偏好、互动内容等
社会特征	婚姻状况、家庭情况、子女状况、社交渠道偏好等
消费特征	商品品类、价格区间、收入状况、已购商品、购买渠道偏好、最后购买时间、购买频次等

下面我们再来介绍一下各类标签构建的先后顺序。这个先后顺序需要综合考虑实际的业务需求以及构建的难易程度等。因为业务需求各有不同，所以我们这里就只介绍按照构建难易程度来排列的先后顺序，如图4-21所示，这是一个化妆品商家的用户标签体系。

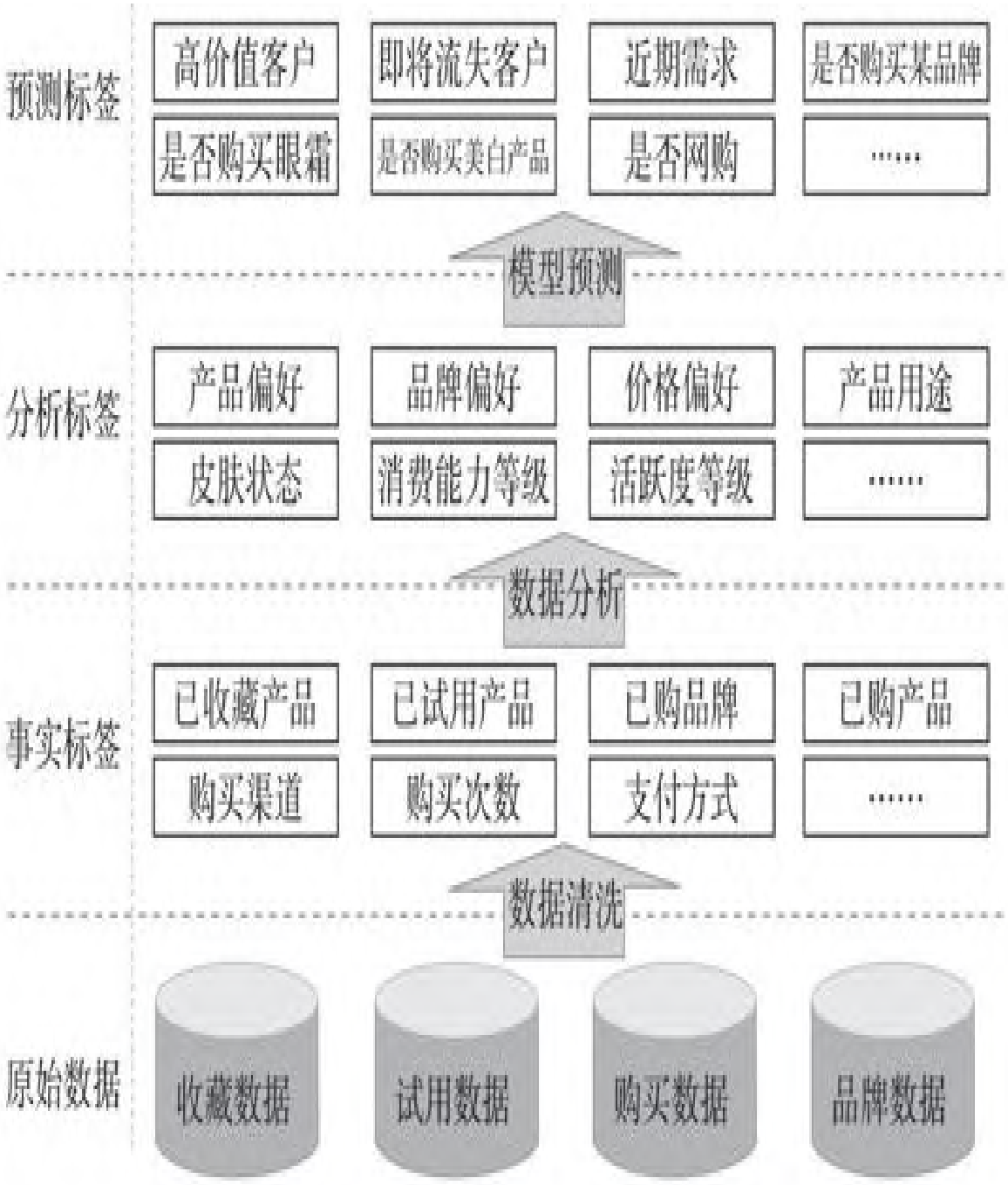


图4-21 化妆品商家用户标签体系示意图

原始数据就是我们数据库中这些数量庞大、结构混乱的数据。在原始数据基础上构建的是事实标签。事实标签的构建可以是对原始数据的直接提取，也可以是对原始数据的简单统计。事实标签的构建过

程并不复杂，但是在这个过程中，我们加深了对数据的理解，也清楚了数据的分布情况，为后续标签打下了基础。在事实标签基础上分析加工形成的是分析标签，分析标签的建立是整个数据标签化过程中的重要步骤。最后构建的是预测标签。预测标签需要通过机器学习预测模型来建立，例如我们可以根据某用户对产品试用的评价来预测他是否购买该产品。需要注意的是，原始数据可能并不完善，比如数据中并没有该用户的年龄信息，这时就需要建立一个模型来预测。模型的预测也是建立在这个用户的其他行为特征上的，例如我们可以根据该用户是否上学、工作岗位等信息来预测他所在的年龄段。在商业场景中，为用户打标签的过程又称为“用户画像”。

数据标签化助力企业营销

下面我们来看一下数据标签化是怎样在营销领域应用的。图4-22是一个银行的营销类标签体系。对于银行来说，数据标签化的第一步是统一用户标识，统一用户标识之后，用户的信息都汇聚在了一起，这样就可以对用户一贯的行为轨迹形成认知。这时再看银行数据，就能够有效地利用用户的属性数据、行为数据以及CRM数据等。属性数据包括年龄、性别、住址、习惯等；行为数据包括用户在银行存钱或者借款等一系列行为形成的数据；CRM数据就是银行与用户进行过一些营销层面的沟通而形成的数据，比如银行向该用户推送过什么类型的广告和权益等都会有相关记录。对这些数据进行分析总结就可以形成基础行为标签和基础用户画像。当银行知道了用户的基础行为都对应了哪些标签，也就对用户画像有了一定的了解。通过基础行为标签和基础用户画像就可以形成场景化标签。比如要做营销，会有一套标签；要控制风险，会有一套标签；要推广一套产品，会有类似的其他标签；等等。场景化画像形成之后就可以对用户进行全方位的画像，比如用户是谁，哪些用户是有价值的，用户需要什么，对哪些商品感兴趣，如何选择沟通的媒介等。总体来看，数据标签化的应用过程就是

用户数据在统一用户标识的基础上逐渐往上发展，最终形成对用户场景化认知的过程。

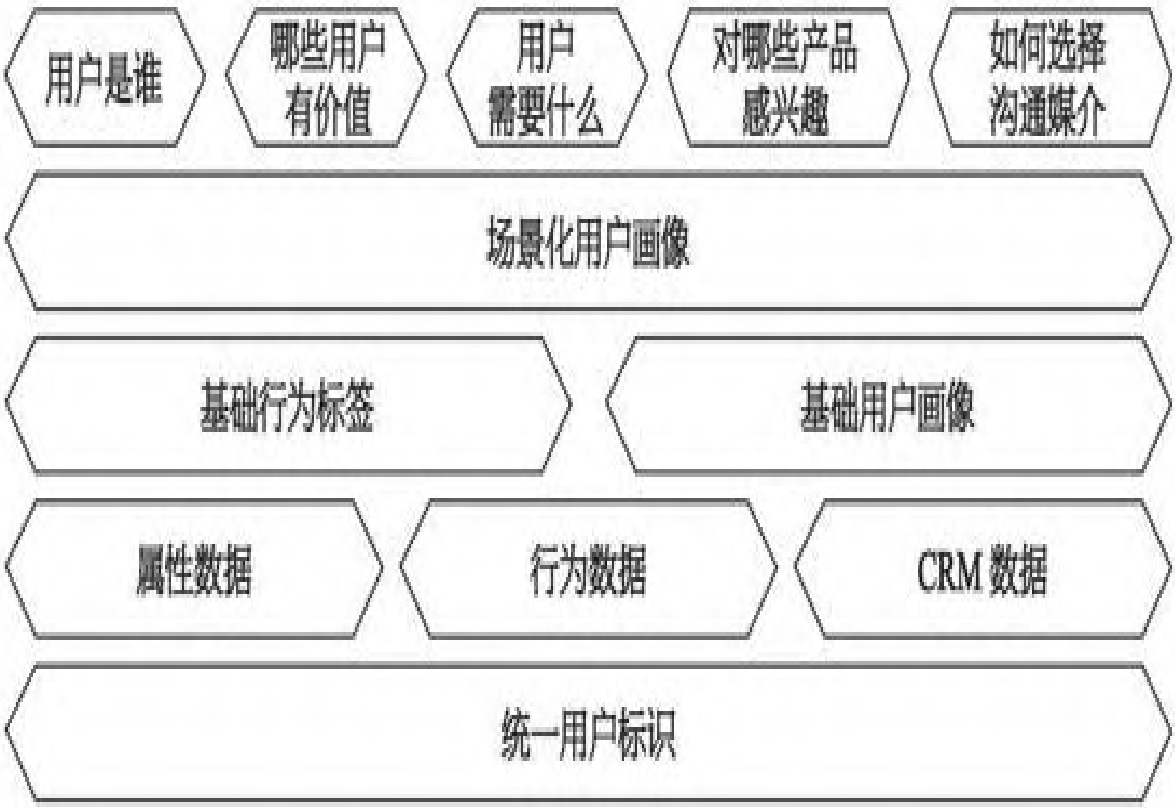


图4-22 营销类标签体系示意图

图4-23是一套具体画像的例子。比如这套画像是针对金融机构用户的。例如，用户对旅游、电商以及交通出行等感兴趣，这些是一级行为标签；再往下细分，金融里面有理财、保险、贷款等二级标签；金融保险里又可以产生车险、少儿险这样的三级标签。这些标签都是根据用户的行为产生的，比如某个用户经常浏览金融理财页面，就会有金融理财标签。根据这些用户行为标签，还可以推出用户人口属性标签。但人口属性标签不全是事实性数据，比如我们不知道一个用户是否有车，但我们可以通过机器学习模型预测出来。假如这个用户平时上下班和周末都是乘坐地铁公交，或者是打车，那他有车的可能性

就很低。通过行为数据还可以大致预测出用户所处人生阶段，是已婚还是未婚、中年还是老年等。

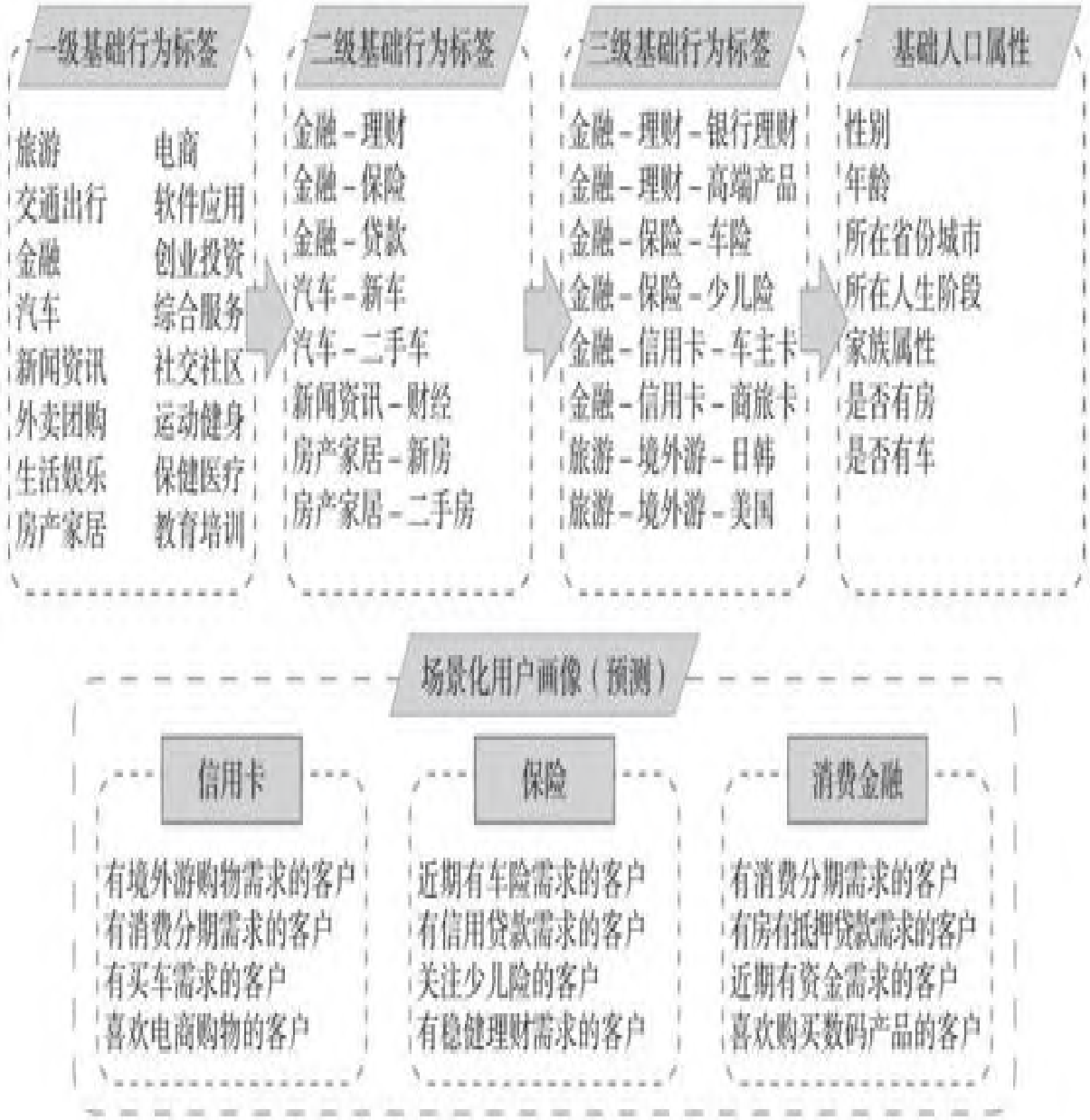


图4-23 场景化用户画像示意图

通过上述数据能产生多个级别的标签维度，这些标签维度是对用户行为的刻画，通过机器学习方法从用户的行为又可以推知用户的其他属性。这些行为和属性维度可以用于场景化画像，预测客户在具体场景下会有什么需求。根据图4-23中金融机构的标签体系，可以预测出跟信用卡、保险、消费金融相关的用户画像，从而为实际的营销和运营活动服务和赋能。比如要做信用卡项目，可以将信用卡的广告信息推送给有境外游购物需求的客户；要做保险项目，可以将保险项目信息推送给有车险需求的客户；要做消费金融项目，可以将相关项目信息推送给有消费分期需求的客户；等等。因为客户的需求和推广的项目都是相对应的，广告营销的针对性和准确性就相对较高，客户购买产品和服务的可能性也就相对会有所提升。

深度学习数据标签化

前面具体讲了数据标签化的过程，接下来我们看一下深度学习是怎样进行数据标签化的。深度学习是对人类神经网络的模拟，它对原始数据一层一层地进行分析和学习，最终会给一个句子、文本或者用户打上相应的标签。当我们拥有一套完整的标签体系之后，就可以将这套标签体系下的训练数据用来训练深度学习模型，深度学习模型就会知道哪些数据应该打上哪一类标签。经过深度学习后，人工智能就可以自动打标签了。如图4-24，这是一个深度学习标签化的一个具体示例。

我们现在有一个句子“曹雪芹是中国古代四大名著之一《红楼梦》的作者”。如果要把这个句子打上标签，可能会有“名人”“图书”这样的相关标签。那么，这套标签是如何通过深度学习加工出来的呢？

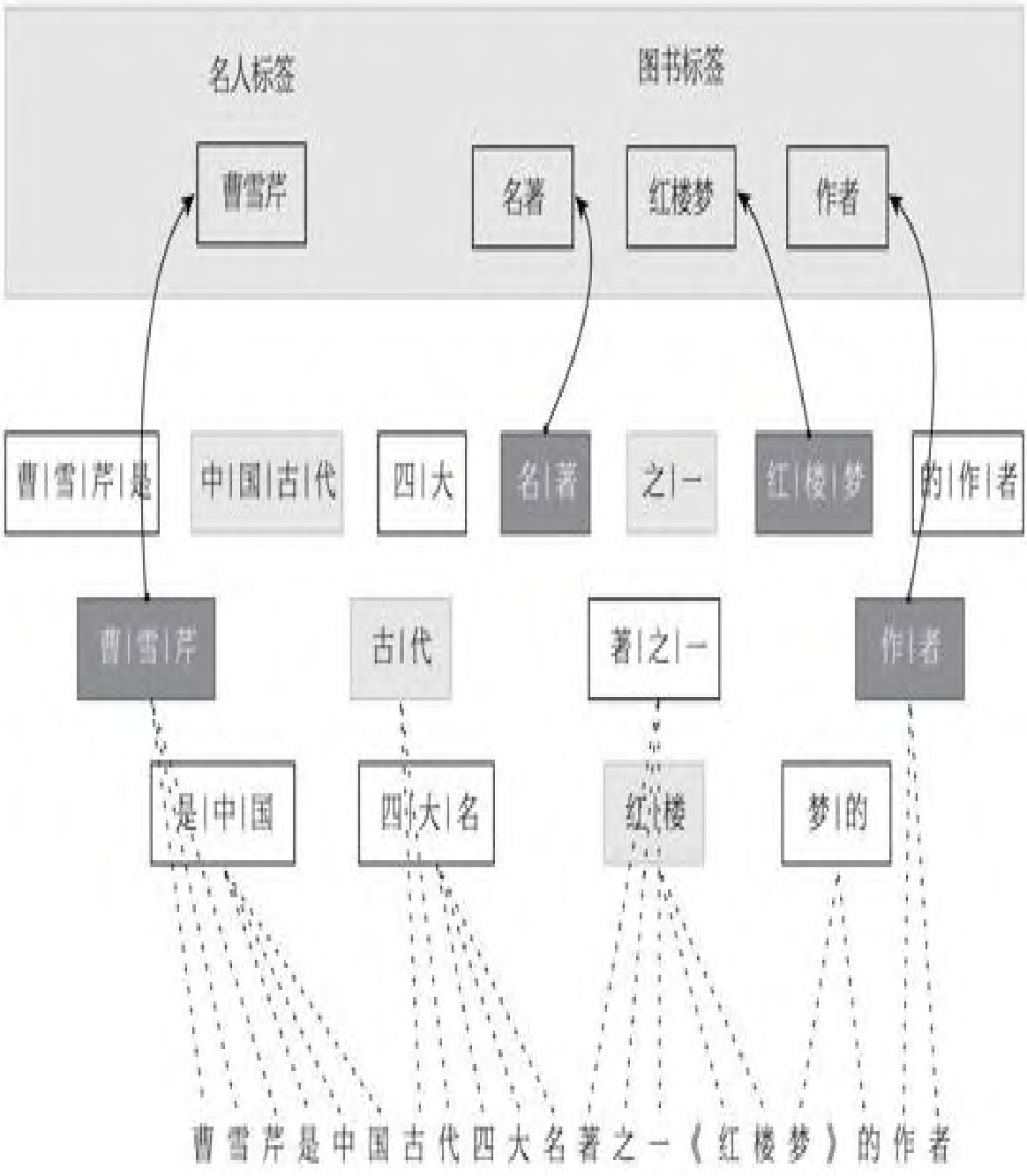


图4-24 深度学习标签化示意图

首先，深度学习并不知道“曹雪芹”是一个词，它也有可能将“芹”和“是”放在一起看作一个词。但通过大量语料训练以后，深度学习会发现“曹雪芹”更可能是一个词，因为很多语料中“曹”

“雪”“芹”这三个字都是组合在一起的，所以“曹雪芹”三个字会聚集成关键词。通过大量文本训练后，深度学习能学出“名人”标签对应的关键词。因为名人经常与“曹雪芹”有关，所以“曹雪芹”就关联上了“名人”标签。同样，深度学习可以自动将“曹雪芹是中国古代四大名著之一《红楼梦》的作者”这句话中的“名著”“红楼梦”“作者”这些关键词贴上图书标签。深度学习通过大量的语料学习，知道哪几个字该放在一起形成词，哪几个字不是放在一起的，并且把关键词和相应的标签关联起来，这就是深度学习理解句子的方法论。

数据是AI思维的基础。大数据时代，只有理解了数据，人工智能才能产生决策、创造价值。数据标签化是人工智能处理数据和理解数据的重要工具，更是人工智能理解商业数据的关键环节。数据标签化使我们知道了用户最想要的是什么，在营销过程中，我们就可以精准地投放用户最想要看到的广告；在研发过程中，我们也可以设计出真正符合用户需求的产品。基于数据标签化的商业实践真正做到了有的放矢，它也为未来商业开辟了新的发展空间。

AI理解消费数据

数据按结构类型分为结构化数据和非结构化数据，我们在前面几节讲过了图像、视频、音频、自然语言这些具有代表性的非结构化数据的处理。下面我们就来了解一下人工智能是如何处理结构化数据的。我们只挑选其中有代表意义的两种数据展开讨论：消费数据和社交数据。很多机构都有这样的数据，那么，通过人工智能来解读它们，可以得到什么价值呢？首先我们来看一下消费数据。

消费数据

根据国家统计局数据，2019年全年国内生产总值为990 865亿元，而全年社会消费品零售总额就高达411 649亿元。一个人的消费就是另外一个人的收入，用户的消费就是商家的利润来源，消费是拉动整个商业社会向前发展的驱动力，消费数据不仅能反映出社会整体的经济水平，而且对一个企业的经营和发展具有非常重要的指导意义。接下来我们就来了解一下什么是消费数据。

一般来说，消费数据指的是个人在商家购买商品而形成的数据，这个范围非常广泛，与我们的日常生活息息相关。大到买房的消费，小到买一块口香糖的消费，都是消费数据的一部分。

消费数据主要分为以下两类：第一类是购买不同类型的商品而形成的数据，例如在电商网站上的商品分为饰品、鞋靴、数码、电器等一系列类型，在日常生活中，用户也会根据自己需求的不同来购买形形色色的商品；第二类是在不同的时间购买商品而形成的数据，比如用户在春天购买的商品与秋天购买的商品的金额就会有所不同，又或者一个用户在“6·18”年中大促购买了一件商品觉得质量很好，然后在“双12”时又购买了一件同样的商品。不论哪一类消费数据，都能够有效地反映出消费者的需求与特点。

消费数据的种类不同，用到的处理方法自然也就不同。下面介绍两种数据处理的模型框架：一个是推荐模型，一个是CLV，也就是客户生命周期价值模型。它们分别对应上面说的两类消费数据相关的场景。

推荐模型理解消费数据

听同事说最近新上了一部电影特别好看，我们周末路过电影院不自觉地就去看了这部电影；看到网上很多人说一家店的食物做得特别好吃，我们就会去品尝一下到底有多么美味。我们日常生活中的很多购物消费行为都会受到各种推荐的影响。这些推荐可能来自身边人的分享，也可能是商家的广告营销，它已经成为影响消费的一个重要因素。在人工智能领域中，推荐模型指的是通过分析数据，预测用户需求，再将满足需求的商品和服务推荐给用户的人工智能模型。随着商品种类和数量的快速增长，用户需要花费大量时间和精力才能挑选出最适合自己的商品，所以能够根据用户需求向用户推荐商品的推荐模型，有很大的应用价值。

提到推荐模型，就不得不提及世界上最大的在线影片租赁服务商奈飞。奈飞的电影租赁业务之所以能做到世界头部位置，很大一部分原因是奈飞网站搭建了电影推荐模型。当一个用户在奈飞网站看完一部电影后，他会根据自己的喜好为这部影片打分。这些分数就成为推荐模型的参考数据，推荐模型可以根据这些数据来判断用户是否喜欢某部电影，可能喜欢哪种类型的电影，然后根据预测出的用户需求来向用户推荐电影。为了让自己的推荐模型更加个性化、更加准确，奈飞在2006年还组织了一个推荐模型大奖赛。在这个比赛中，只要参赛者能够最大地优化奈飞的推荐模型，就能够赢得100万美元。在高额奖金的吸引下，很多做机器学习的高手都来参赛。最终比赛的结果是，奈飞的推荐模型的推荐准确率提升了10%，整个产业界推荐模型的发展也因此向前迈进了一大步。权威杂志*IEEE Computer*为此曾专门介绍相关推荐模型。下面我们就来了解一下奈飞的推荐模型具体是怎么工作的。

首先我们要知道，奈飞的数据很多，构成了一个庞大的矩阵。在这个矩阵中，每一行对应的是一个用户对不同电影的评分数据，每一列对应的是一部电影收到的不同用户评分数据。比如说一个用户评价过5部电影，在奈飞的数据矩阵中，属于这个用户的那一行里就有5个非零的数字，来表示这个用户对电影的评分。奈飞大奖赛里获胜的团

队用的就是矩阵分解的模型。矩阵分解就是把庞大的矩阵分解成两个小的矩阵，它采用的是降低维度的思路，是把原始矩阵用低维的子空间去拟合。原始矩阵在很高维的空间中，拟合就是用相对低维的空间，比如说用几百个或者几千个维度，来拟合原来更高维的数据。

降维更重要的目的是提取用户本质的行为。通过降维，我们可以在原来庞大的矩阵中提取本质信息，来代表用户兴趣中最稳定的部分，它也是模型可以用来预测行为的部分。比如说在矩阵中，有一个用户看过很多动作电影，而且对动作电影的评分都非常高，那么说明这个用户对动作电影感兴趣，如果向他推荐此类电影，他观看的可能性就会比较高。

通过这样的方法，我们可以预测新的信息。具体是如何实现的呢？当我们用矩阵分解模型把大矩阵分解成两个小的矩阵之后，需要再把这两个小矩阵乘起来，重构一个新矩阵。在这个重构的新矩阵里会有更多非零的数值，而且排列要比原来的矩阵稠密一些。这时就可以根据新矩阵中的非零数值进行推荐，比如选取用户所在的那一行预测出来最大的非零数值，再把这个非零数值所对应的商品或服务推荐给用户。

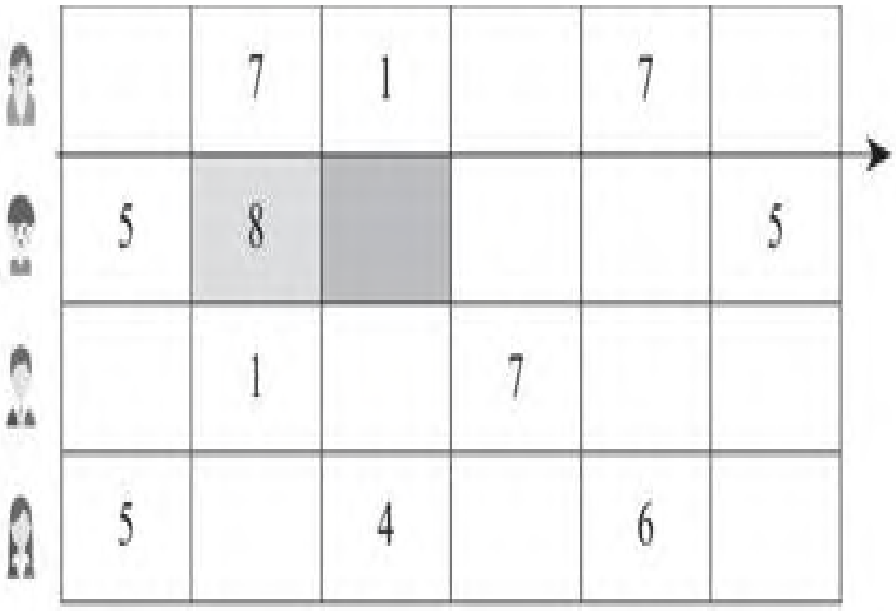
这就是矩阵分解“协同过滤”思想核心的地方。通过矩阵分解还原用户和商品数据中的规律，再通过重构矩阵来预测出原始矩阵中不存在，但是用户可能感兴趣的商品，然后把它们推荐给用户。为什么这个模型能够准确地预测用户的兴趣呢？比如说，如果许多人都看过某几部电影，那推荐模型就会认为这些人的需求相似，就可以把这个群体中一些人看过的其他电影，推荐给这个群体中没有看过这部电影的人。同样，推荐模型也会认为这些电影之间存在相似性，当这个群体之外的人看过这些电影中的一部，就可以向他推荐这些电影中的其他几部。

为了更好地理解推荐模型，我们下面来举一个实例。如图4-25，最左边表示的是在某视频网站上观看电影的名称和相应的评分。行对

应的是不同的用户，列对应的是不同的影片，分别是《红海行动》《我不是药神》《无名之辈》《芳华》《寻梦环游记》《神秘巨星》，数字表示的是不同用户对这些电影的评分。这个表格其实也可以类比实体商品的消费数据，只需要把影片换成实体商品就可以了。如果想知道一个用户是否对其他商品或商家感兴趣，感兴趣的程度有多高，都可以通过这类推荐模型实现。

从图4-25可以得知，第四个用户对《红海行动》的评分为5分，对《我不是药神》的评分为8分。如果想要继续知道这个用户对《无名之辈》的感兴趣程度，就需要把整个矩阵通过模型做分解。图4-25右边的部分展示了分解的过程。首先，我们将大的矩阵分解成两个小的矩阵，左边的矩阵每一行对应一个用户，右边的矩阵每一列对应一部电影。两个小矩阵的每一行或者每一列对应本质的兴趣特征，就像DNA存储了生物的基因信息一样。在左边的矩阵里，每一行是关于用户兴趣的DNA；在右边的矩阵里，每一列是关于影片的DNA。如果我想知道第四个用户对《无名之辈》的兴趣度有多大，我就可以把第四个用户的DNA和《无名之辈》影片的DNA乘在一起，算出一个数字，这个数字对应在这个未知的空格中。这个数可能是8也可能是1，如果是8，代表用户对这部影片的兴趣度很大，如果是1，代表用户对这部影片的兴趣度很小。




7 1 7
5 8 5
1 7
5 4 6

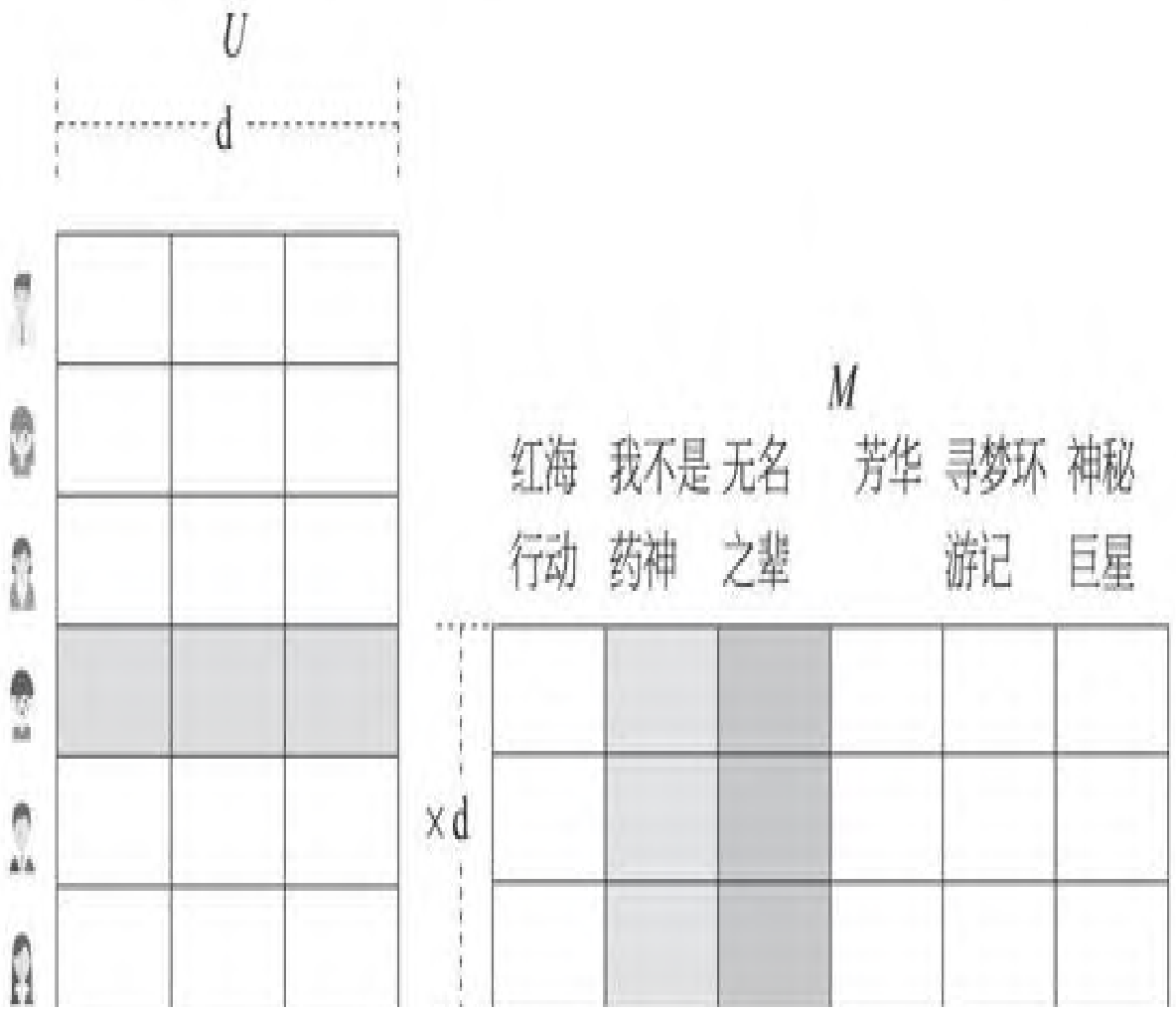

U
d
M
红海 我不是 无名 芳华 寻梦环 神秘
行动 药神 之辈 游记 巨星
xd

图4-25 某视频网站推荐模型示意图

现在我们来总结一下矩阵分解模型推荐功能的原理：首先，模型会把消费数据分解为两个小的矩阵，再把矩阵的行和列进行组合，相当于数学上的向量乘法，组合之后就可以预测出行和列交叉处的兴趣值。计算出来的兴趣值大小不一，但是数值越大代表用户的兴趣度越高，所以我们就可以根据兴趣值，向用户推荐他们没有购买消费过，但兴趣值很高的商品或服务。这是各大电商网站推荐商品的原理，也是视频网站上影片推荐的方式。只要用户购买过商品，结合用户的购买行为以及其他具有类似需求的用户的购买行为，矩阵分解模型就可以把最恰当的商品或服务推荐给用户，这是基于消费数据的推荐模型中最基本的框架。

奈飞的电影推荐，电商网站的商品推荐，金融机构向客户交叉销售产品，都可以应用这一类推荐模型。这种推荐模型能够从消费数据中提取规律，为用户推荐最适合他的选择，这是推荐模型最核心的思想。例如，随着生活节奏的加快，人们的闲暇时间越来越少，所以很多人会选择在网络上购物，而各大电商平台也都设置了购买之后的用户评价功能。如果一个用户在购买商品之后给了五星好评，那就说明他对这个商品很满意，这个商品符合他的需求，之后推荐模型就会向与这个用户具有相似喜好的其他用户推荐该商品。

客户生命周期价值理解消费数据

接下来我们就来讲一下第二类消费数据处理的模型框架：客户生命周期价值。客户生命周期指的是从一个客户想了解企业的一些信息或者企业想开发这个客户开始，一直到客户和企业的业务关系以及业务的相关事宜完全结束并且处理完毕的这段时间。对于一个企业来

讲，客户生命周期描述的是客户与企业的关系从开始到发展再到结束这样不同状态间运动的总体特征，形象地讲，客户与企业的关系经历了诞生、成长、成熟、衰老和死亡这样一个完整的生命过程。

客户生命周期分析经常应用于零售行业，指的是客户从成为某商家的客户并产生业务消费开始，经过消费成长、消费稳定和消费下降阶段，最后完全不在这个商家购物的过程。客户生命周期价值指的是一个公司或者产品预计在与客户的整个业务关系中获得的利润。让客户在生命周期中产生商业价值才是一个企业运营的最终使命，所以企业在运营过程中要尽一切可能来延长客户的生命周期，在客户的生命周期中要尽一切可能来产生商业价值。那么我们如何预测客户生命周期的价值呢？

一般来说，我们都是从已有的消费数据来推算客户生命周期的价值的。从一个简单的角度来讲，对一个客户的消费数据而言，最重要的三个变量分别是：最近一次消费时间（recency）、消费频率（frequency）、消费金额（monetary）。

图4-26表示的是一个客户在6个月内的购物行为。一个购物车代表一个月，这个客户在第2个月和第5个月购买了商品，其他时间没有买。在这里，总观察时长是6，最近一次消费时间是5，消费频率是2。掌握了这些变量后，我们就可以计算很多跟消费者有关的指标，比如：客户在将来某个时间点活跃的概率、客户在将来某个时间段购买次数的预期值、客户剩余交易次数的预期值和客户剩余生命周期价值等。其中，客户剩余生命周期价值是CLV的一种体现。



图4-26 客户生命周期的变量示意图

那么我们是如何计算这些指标的呢？最重要的是，我们需要一个描述客户消费行为的模型，下面来讲一个最基本的CLV模型：在每个时间点，客户都有一定的概率购买，也有一定的概率失活，也就是再也不买了。购买概率和失活概率本身也是变量，因此我们要用模型去表示它们。国际著名的《营销科学》（*Marketing Science*）杂志曾有一篇文章讲了这个模型，并得出了一个有趣的结论：上面说的客户在将来某个时间点活跃的概率、客户在将来某个时间段购买次数的预期值、客户剩余交易次数的预期值和客户剩余生命周期价值这4个与客户相关的指标，都是可以用最近一次消费时间和消费频率，以及购买概率、失活概率相关的模型参数来表示的。而这些模型参数通过人工智能计算可以求得最优解，将其代入前面提到的4个指标后，用最近一次消费时间和消费频率，就可以算出跟客户将来消费相关的一系列指标。这个结论表明了最近一次消费时间和消费频率在消费数据分析领域的重要性。同样，消费金额也有类似的作用。

在与客户相关的4个指标中，商家最关心的指标是客户剩余生命周期价值。顾名思义，客户剩余生命周期价值就是客户在之后的生命周期中还有多少价值。比如一个客户在一个行业中的生命周期已经到了消费稳定阶段，那这个客户的剩余生命周期价值就是在消费稳定以及消费下降这两个阶段所带来的价值。剩余生命周期价值的数值可以通过求利润率、每笔交易收入以及客户剩余交易次数的预期值的乘积来得出。商家可以根据客户的剩余生命周期价值来判断，这个客户还会不会购买自己的商品，能够购买多少商品，商家还要不要对这个客户进行营销，如何营销等。这样，商家既能有效地控制成本，又能有针对性地进行推广，广告营销的投资回报率也会相应提高。

除了上述模型外，在客户关系管理中的RFM模型也可以用来理解消费数据。RFM模型能够有效地衡量客户的价值以及客户创造利润的能力，其中的R、F、M就是上面提到的理解消费数据最重要的三个变量：

最近一次消费时间、消费频率、消费金额。这三个数据是零售商最感兴趣的三个数字，也是在预测客户价值过程中必不可少的三个指标。最近一次消费时间是指客户上一次的消费时间，最近一次消费时间越近的客户，对商品和服务印象越深，再次购买商品和服务的可能性就越大。消费频率是指客户在一定时间段内所购买商品或服务的次数，这个次数越多，那说明该客户对商品或服务的满意度越高，对于企业品牌有一定忠诚度，这样的客户重复购买商品或服务的可能性也就越大，而且还是持续性的。消费金额同理，金额越高说明客户对商品或服务越满意，再次购买的可能性就越大。

通过RFM模型可以把客户群区分开，得到不同价值的客户群，比如分成重要维持客户、重要发展客户、重要挽留客户、一般重要客户、一般客户和无价值客户等几个类别。这在智能化客户管理中很重要，很多企业已经对此进行了实践。比如说在没有对客户进行区分的情况下，某家商店可能会付出一定的成本给1 000位无差别的用户发短信，通知他们商店有新品上架，欢迎购买，但是可能只会收到100个订单。但是有了RFM模型之后，商家就可以只将这1 000条短信发给重要维持客户、重要发展客户这些购买概率大的客户，付出同样的成本，收到的订单数却会大大上升。而且当RFM模型对用户进行区分之后，发的短信就可以不用千篇一律了，比如给那些以前经常购买但有一段时间没有购买过商品的客户，短信的开头可以是“好久不见”，而对于那些在短时间内购买很多商品的客户，短信的开头可以是“恭喜您成为本店的VIP客户”。对于客户来说，这样千人千面的短信比那些同质化的短信更具有吸引力，他们再次光顾的可能性也就更大。

人工智能通过客户生命周期的价值来理解消费数据的重要意义是：根据客户有限的消费数据，可以推测他将来是否会继续消费，消费金额是多少。对于那些消费可能性高的客户，商家可以采取相应措施来刺激他们的消费；对于那些消费可能性低的客户，商家也可以采取一些措施来防止客户流失。

消费数据仿佛一棵茂盛的大树，盘根错节，枝蔓横生。但是AI思维就是要在这些看似复杂的数据中抓取其最本质的特征，为决策提供依据。理解消费数据的过程就是通过人工智能对这些数据抽丝剥茧、披沙拣金的过程。人工智能理解了消费数据，就可以通过推荐模型提高客户的购买率，可以通过客户的生命周期价值模型提高客户的价值。总之，人工智能理解了消费数据后，就可以从各个角度刺激客户的购买欲，增加企业商品和服务的销量，提高企业的商业价值，对企业的发展起到强有力的推动作用。理解了消费数据，我们就扣住了AI思维中的重要一环，在决策中真正做到对症下药，一针见血。

AI理解社交数据

人类天生就是一种群居动物，社交是我们与生俱来的属性。几百年前人们鸿雁传书交流信息，几十年前人们通过电话与千里之外的人沟通感情。现在，我们动动手指就能在手机上和牵挂的人联系。日常生活中，我们也无时无刻不在与其他人沟通交流。我们离不开社交，每时每刻都在产生着社交数据。这些数据里包含着我们的喜怒哀乐，反映出我们与这个社会的联系，所以我们这一节就来介绍一下人工智能是如何理解社交数据的。

我们在进行社交的过程中，会和许多人建立关系，这样就形成了社交网络。由于我们的时间和精力是有限的，不可能和许多人保持紧密的社交关系。一般来说，一个人的社交网络最大只能包含150个人，平均在124个人左右。也许你听说过，不论你想要认识谁，只要通过6个人就可以认识他，这就是被美国社会学家斯坦利·米尔格拉姆（Stanley Milgram）证明的六度人脉理论。在社交网络中，每个人都是一个节点，人与人之间的关系就是社交网络的边。在现实生活中，

以社交网络的节点为基础，通过社交网络的边，就可以拓展我们的人脉。

在社交网络中有一个重要的概念叫作中心性，它表示的是一个节点在网络中处于核心地位的程度。中心性有三种计算方法，分别是度中心性（degree centrality）、中介中心性（betweenness centrality）和紧密中心性（closeness centrality）。

第一种，度中心性。在社交网络中，度中心性表现的是一个节点与很多其他节点发生的直接联系。一个节点连接的节点越多，这个节点也就越重要，处于中心地位，这个是比较直观的指标。比如说，用户甲的社交账号上有50个好友，这就意味着有50个节点与甲相连接；如果在同样的社交平台上，用户乙的社交账号有100个好友，那么乙的度中心性就比甲要高，社交圈子比甲要广。

第二种，中介中心性。中介中心性是指一个节点出现在其他节点之间的最短路径的次数。如果这个节点的中介中心性高，它对整个网络信息的传播就会有很大的影响。例如我们的不少朋友可能都是通过某位社交达人才认识的，那么很显然，这位社交达人起到了中介作用。换句话说，就是这个节点相当于一个关口，和它相连的节点想要到其他节点都得经过它。

第三种，紧密中心性。紧密中心性指的是一个节点到其他所有节点最短路径的和，主要用来考察一个节点在传播信息时对其他节点的依靠程度。如果一个节点离其他节点越近，那么这个节点就不会受制于其他节点，它传播信息的时候也就越不需要依赖其他人。比如，在一个公司里，甲只认识本部门的几个人，而乙除此之外还认识部门外的其他人，那么乙的紧密中心性就要比甲高，因为甲要与非本部门的人进行信息传递可能需要经过乙的搭线，而乙因为本身就认识其他部门的人，则不存在这些限制。

通过这三种方式，我们就可以找到中心性高的节点，也就是找到在社交过程中会产生很大影响的人或者组织。中心的节点有着不可取

代的重要性，它们更具有权威性、中枢性、核心性等特征，同时也具有更多的经验和影响力，对社区的形成起着决定性的作用。比如说在营销领域，我们需要挖掘意见领袖，这时候意见领袖就是中心节点，因为，通过意见领袖展开的营销才能更具影响力，更容易被大众相信，营销效果也才会提升。

社区发现模型

下面谈谈什么是社区发现。社区是由一组连接紧密的节点组成的，并且这些节点与社区外部的节点连接稀疏。如图4-27所示，可以看出图上有三个社区，符合这样的规律。不同社区之间的关系稀疏，因而社区更多地反映的是网络中的个体之间的关联关系。研究网络中的社区对理解整个网络的结构和功能起到至关重要的作用，还可以帮助我们分析、预测整个网络各节点间的关系。社区发现就是在复杂网络中发现这些连接紧密的社区结构，或者说，社区发现就是网络中节点的聚类。举个例子，在一个班级里面，有的学生喜欢文学，有的学生喜欢绘画，有的学生喜欢数学，有相同喜好的学生之间联系会更为紧密和频繁，根据这些交流的频繁程度，可以把整个班级里的同学分成几个不同的社区。又或者说，在公司里，总会有几个联系较为紧密的小群体，例如，一部分人经常聚在一起讨论工作学习，还有一些人经常聚在一起讨论时尚潮流，根据他们之间联系的紧密程度和频率高低，很自然地能够发现这两个人群已经各自形成一个社区。

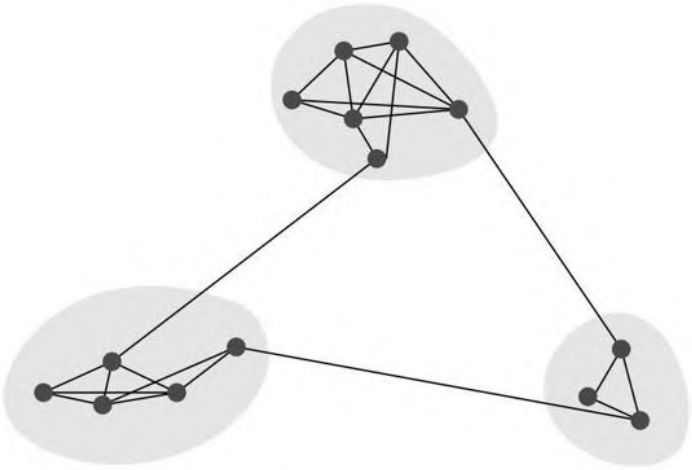


图4-27 社区示意图

那么，我们怎么去进行社区发现呢？有很多常用的社区发现模型，例如吉尔韦安（Girvan）、纽曼（Newman）等人在2003年提出的GN社区发现方法（以两位开创者的名字首字母命名），就是社区发现中的一个比较经典的模型。GN社区发现方法提出了边介数的概念。那么，什么是边介数呢？由网络中社区的定义可知，所谓社区就是指其内部节点的连接稠密，而与其他社区内的节点连接稀疏。这就意味着社区与社区之间联系的通道比较少，一个社区到另一个社区至少要通过这些通道中的一条。如果能找到这些重要的通道，并将它们移除，那么网络就自然而然地分出了社区。在实践中，我们用边的边介数来作为这个分割的依据。边的边介数定义为网络中所有最短路径中经过该边的数目。跟节点的中介中心性类似，它反映了相应的边在整个网络中的作用和影响力。联系定义可知，通过社区内部的边的最短路径的数目相对较少（边介数较少），而通过社区之间的边的最短路径的数目则相对较多（边介数较多）。因为社区之间的联系只能经由这几条路径发生，就像车流要从停车场开到外面一样，如果停车场只有一条通道，那么车流都得从这条通道走，这条通道的使用率就是100%，非常高；如果这个停车场内部有5条通道，车流可以分散开来，可能每个通道的平均使用率只有20%，远没有前者高。

GN社区发现方法是基于删除边的方法，其本质是基于聚类中的分裂思想，原理上使用了边介数作为相似度的度量方法。GN社区发现通过移除网络中的边来将整个网络划分成为合适的社区结构，将原来的网络分割成为任意数目的社区。在GN社区发现方法中，每次都会选择边介数高的边删除。GN社区发现方法的步骤如下：

第一步，计算每一条边的边介数；

第二步，删除边介数最大的边；

第三步，重新计算网络中剩下的边的边介数。

重复上述步骤，直到产生合适的社区结构。

通过上述步骤，GN社区发现方法可以较好地发现网络中存在的社区结构。在网络理论中，复杂网络是由数量繁多的节点和节点之间错综复杂的关系构成的。GN社区发现方法用边介数的概念来探测边的位置，从而在盘根错节的复杂网络中划分出社区。

人与人之间的互动使其相互间产生了或紧密、或松散的联系，也在不经意间建立起了一张张连通社会各个层面的网络：通信往来、商业交易和互联网社交等。社区发现模型可以用来找出网络中的社区聚类，让人们洞察这些网络中隐藏的形形色色的“社区”，也可用于评估一个网络结构中个体组合或分裂的程度，并从中获取很多新发现。

社区发现的应用一：推断用户未知信息

我们本来就生活在一张庞大的社交网络之中，随着社交媒体的兴起，社交网络变得越来越复杂。社区发现在近些年的不断发展，使得人们可以对复杂的社交网络进行分析，对于现实生活具有重要的应用意义。我们接下来就从推断用户未知信息、提高营销效果以及发现不良行为团伙三个方面剖析社区发现的应用价值。

社区发现的应用价值之一是推断用户未知信息。随着社交媒体的高度发展，其用户规模也在快速增长，有效的社区发现模型的应用也越来越重要。例如购物网站上的社交购物、商品推荐；微信、微博等媒体的社会舆情分析、意见领袖挖掘等，无论你想要找什么样的信息，都能通过社区发现模型获得启发。如果一个用户的某些信息尚且未知，但是我们知道他所在社区中很多人的信息，那么，我们就可以用这个社区中其他人的信息来推断该用户未知的信息。比如，你发现在朋友圈里有两群人，一群人花钱都大手大脚，入不敷出，另一群人生活稳定有规律，喜欢运动健身、阅读学习。这两群人都要向银行借钱，生活稳定有规律的这群人里，借钱的都还了，那么这群人里还没有借钱的人，一旦借钱，还钱的概率就大；花钱大手大脚的这群人里，不少人借的钱没还，那么，还没有借钱的人，一旦借钱，不还钱的概率就很大。

社区发现之所以可以用来推断用户未知信息，是因为我们可以在真实的社交网络中，通过社区发现找到节点之间的相互关系，利用已知节点的属性信息，推断出未知属性节点的缺失信息。基于社区发现的未知信息推断方法是基于两个假设：属性信息相似的人容易成为朋友，属性信息相似的人群倾向于聚集为社区。如果要推断一个社区中某些用户的未知信息，就要先从社交网络中选取若干个中心性较高的种子节点，通过这些节点的好友数据，运用社区发现模型进行划分，再根据节点的属性信息推断同社区好友的未知信息。实验结果表明，即使当给定的信息只占用户信息的20%时，某些用户属性也能够精确推断出来。所以，我们可以通过社区发现模型划分社区，根据未知属性节点所在社区的属性信息推断该节点的未知信息。

社交网络既是交流平台，也是市场营销等活动的平台。在人际关系网中，可以挖掘出具有不同兴趣、背景的社会团体，方便采用不同的宣传策略；在交易网站中，不同的社区代表不同购买力的客户群体，方便运营人员为他们推荐合适的商品。针对社交网络用户社区发现的方法，一般来说，可以基于用户发布的内容提取基于兴趣的用户

社区；或者基于用户联系提取基于联系的用户社区；还可以将两个社区进行融合，形成兴趣和联系双重的用户社区。当社区形成之后，这个社区的成员可能就会拥有共同的兴趣或者联系，假如这个社区的大部分成员都喜欢网球，那我们就可以推断，社区的其他成员也很有可能喜欢网球。

社区发现的应用二：提高营销效果

社区发现模型在商业领域有着非常广泛的应用。在电子商务领域中，通过对消费者购买记录进行社区发现，能够挖掘出具有相同或相似购买兴趣的消费群体，发现相关商品的潜在客户并为其进行引导和推荐，进而提升商业价值。例如，通过社区发现模型，我们看到甲、乙两个消费者同属某一个社区，都买了简约风格的装饰画。这时，如果我们还想进一步为乙推荐商品，但苦于目前并没有太多数据，无法得知他的潜在需求是什么的话，可以通过甲的购物记录，发现甲还喜欢买智能手表，推知乙的潜在需求很可能是智能手表，进而做出相应推荐。

通过社区发现模型除了能找到消费者的未知潜在喜好，从而形成针对性的推荐外，还能找到社区中中心性高的核心人群，更好地实现营销触达。例如，我们要对一个社区的人群做营销投放，如果甲的中心性比乙高的话，那么通过甲来做社区整体营销触达的效果会比通过乙要好。因为，甲与社区里的其他人关系更为密切，也更有权威性，经由他来做营销传播，社区人群会更容易接受并采纳，甚至能够转化成实际行动。

在商业领域，有很多通过社区发现优化营销传播，从而实现商业价值的例子。例如阿里的年货节就利用了人工智能社区发现的手段，对品牌的目标人群进行分析，挖掘品牌目标潜在客户，使得其投放效果提升了将近50%。他们的做法是从品牌方的受众人群出发，进行针对

性的营销投放。年货节开发新用户时，先要通过机器学习模型对阿里年货节的受众人群进行预测。预测可以根据往年年货购买人群的数据进行，并以各大节假日促销时的购买数据作为参考，然后以同属一个社区的人有相似的兴趣爱好、价值观为前提假设，再通过社区发现模型找到与潜在消费者有相同或相似喜好的社区人群，利用社交网络关系进行广告投放。比如，一个社区内的大部分用户喜欢运动，经常购买运动类商品，那么我们就可以假设这个社区内的潜在用户也有相同的爱好，也有运动类商品的需求，在进行广告投放时，就可以向其推荐运动类商品。这样，推荐的商品会有很大的可能被购买，这个用户就从潜在用户转化为实际用户。新用户数量逐渐增加，品牌效益也会逐渐提升。

社区发现的应用三：发现不良行为团伙

社交媒体的发展让我们能够更加方便快捷地获取资讯，但是人与人之间的交流由于缺乏面对面的真实接触，而变得虚拟了。我们不知道与我们在手机软件上相谈甚欢的那个人的真实面貌，更加不能分辨其中的真伪。在社交网络的保护下，各种诈骗犯罪行为变得越发猖獗。但是，社交网络可以是不良团伙牟利的蜜糖，也可以是将他们困于囹圄的砒霜。基于社交网络的社区发现，一个重要应用就是发现不良行为团伙，打击违法犯罪行为。

例如，每天在家和公司之间穿梭，很多人会选择驾车出行，但是经常穿梭于车水马龙之间，难免会出现一些磕磕碰碰的现象。因此，有车一族都会为汽车买保险。但是保险公司发现，近几年竟然出现了许多车险理赔的欺诈现象。例如，在交通事故理赔中有这样一条规定：如果发生的交通事故不涉及人员伤亡，仅造成轻微财产损失，不需要联系保险公司，可以自行拍摄车辆损失代替现场勘查。有不少人就利用这个漏洞骗取保险金。保险公司曾发现，有部分司机在短期内

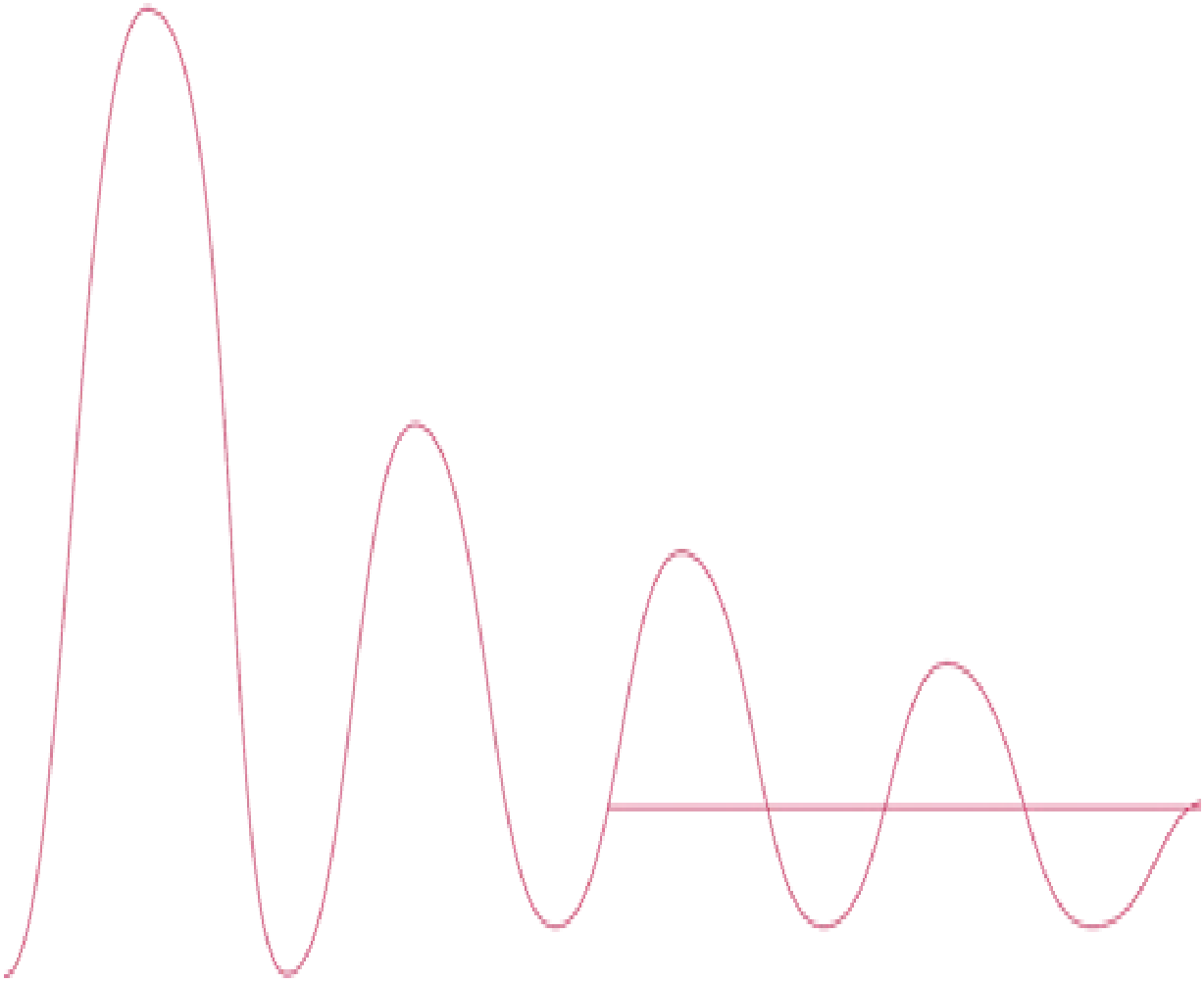
多次制造轻微相撞事故，骗取保险金。但是事情并没有到此为止，社区发现模型的及时介入使得警方发现了这些司机同在一家公司，于是警方经过调查发现，在这家公司工作的另外几名司机也有同样的诈骗行为，便一举拿下了这个诈骗团伙。

随着经济的高速发展，银行客户群体的规模不断扩大，但与此同时，银行客户群体中也混入了一些不法分子，他们伪装成正常客户群体到银行办理业务，意图牟取非法收入，并且这种违法行为现在越来越团伙化。“近朱者赤，近墨者黑。”团伙化犯罪的成员彼此之间都有联系，所以社区发现模型对于发现不良团伙、维护银行利益具有很大帮助。例如，现在有很多人在大量透支信用卡之后，潜逃或者隐瞒身份来逃避还款，极大地损害了银行的利益。为了防止这一现象的发生，一些大型银行借助社区发现模型建立了银行反欺诈系统。如果一个用户想要申请信用卡，银行会以这个申请用户为节点，调取这个用户所在社区里人员的相关信息，如果这个用户处于不良团伙之内，也就是说，在与他联系比较密切的人里面，大范围出现过类似于恶意透支信用卡这样的金融诈骗行为，那么这个用户就会被划入高欺诈风险的用户范围，他的信用卡申请很有可能会被驳回。此外，利用社区发现模型进行深入调查后，这些隐藏起来的不良团伙也更容易被追根究底，一网打尽。基于社区发现模型的银行反欺诈系统大大提高了银行防范风险的能力，在银行反欺诈领域有很大的应用价值。

人类社会离不开社交，要想让人工智能更好地服务于人类，离不开人工智能对社交数据的理解。正所谓打蛇打七寸，想理解数量庞大的社交数据，理清复杂的社交网络，重点在于理解中心性高的节点；在此之后，我们再结合社区发现模型，找到相应的社区群体，一方面可以精准营销，提升品牌效益，另一方面可以发现不良行为团伙，提升社会效益。不管从哪个角度出发，AI思维都能够把握住事物的发展方向，帮我们做出正确的决策。

第五章

数字化赋能



人类用知识的活动去了解事物，用实践的活动去改变事物；用前者去掌握宇宙，用后者去创造宇宙。

—意大利哲学家 克罗齐—

前面几章主要介绍了AI思维的底层逻辑和操作方法。学以致用，本章开始，我们将切入到“用”的层面，聚焦于人工智能是如何运用于具体业务场景之中，赋能于各行各业以及方便人们的日常生活的。数字化赋能将是贯穿整章的基础概念。数字化就是AI思维在具体业务场景中落地的基础，因为AI思维运用于生产制造或商业活动之中，要实现智能化，最终做出决策需要数据的支持和驱动，而数字化恰好提供了AI思维决策所需要的数据基础，在此之上人工智能才能发挥其应有之义。

所以本章将会以数字化赋能为切入点，从企业如何打通数据来进行剖析，结合商业案例，让你更为真实地感知和了解AI思维是如何在具体业务场景中汲取营养、发挥作用的。数字化转型已经势不可当，数字化赋能也不是天马行空、赚取噱头的暗黑发明。在数字化时代中，数据就是产生价值的原油，我们所提到的案例也都是已经实实在在落地，并成功产生了预期效果的行业应用实例。通过这一章，我们会更加深刻地感知到发生在每一个人身边的数字化和智能化变革，理解实际的人工智能在具体业务场景下的工作和价值，在一定程度上也能达到为各位读者“祛魅”的目的。

行业数字化是大势所趋

近年来，从互联网经济到数字经济转型的大幕已经拉开，数字经济成为不断被提及的热词。2017年3月，数字经济首次被写入政府工作报告，要促进深化大数据、人工智能等研发应用……壮大数字经济。数字经济的核心就是数字化，即数据成为驱动商业模式和商业决策的核心力量。从全球竞争格局和行业演变进程来看，数字化驱动产业的时代已经到来，无论是在制造环节还是零售环节，中国、美国甚至全

球趋势都是越来越向智能制造和智慧零售方向发展，行业数字化转型已经势不可当，并带来了商业模式的革命性变革。

行业数字化转型的核心在于数据驱动，即业务流程数字化，数据成为核心生产要素和战略资产。美国政府认为，大数据是“未来的新石油”，是“陆权、海权、空权之外的另一种国家核心资产”。数据是产业进行全方位的数字化升级的源动力，业务的实施、发展和创新都离不开数据的支撑，与此同时，数据还为经济发展提供新的活力和动力。

行业数字化发展的趋势主要体现在三个方面：第一，新融合。大数据、云计算、物联网的发展和应用，导致数据爆发式增长，并且通过技术将底层数据打通融合，与具体业务场景的融合也越来越紧密。第二，新市场。随着数据的应用，互联网不断升级换代，电子商务、工业互联网、互联网金融、国际市场等新市场形成，企业也不断创新，发掘新的市场方向。第三，新应用。智能终端、智慧生产、智能制造、智慧零售等新的行业模式不断更新，并在实际业务中落地，“智能+”为各行各业升级赋能。行业的数字化转型不仅为企业提供了创新和发展的方向，也极大地促进了消费增长，增加了市场的活跃程度，为经济增长带来了巨大的助力。

数字化的技术壁垒和解决方案

企业顺应数字化的大趋势是必然选择，但是首先遇到的关键问题是攻破数字化的技术壁垒，即如何做好数据的打通（data onboarding），这也是数字化的基础工作。过去的数据世界存在着许多技术困境，如果没有做好数据打通，就容易出现各种问题，影响业务的效率和创新。

第一，企业在业务中会产生大量的数据，这些海量数据关联复杂，而且专业性强、维度多。如果将这些数据杂乱无章地存放在数据

仓库中，即使是调取和利用其中的部分数据，都需要花费很长的时间，降低了业务效率。

第二，在企业内部，关键数据总是会散落在业务系统的各处，甚至导致对一个用户在各个阶段的需求没有全局的把握。不同部门为满足各自对应的业务需求，往往会建立不同的数据库，导致数据难以整合，形成数据孤岛。在业务开发和运行的时候，不同的部门或者项目组往往不能统一协作处理数据，以致对数据重复开发，影响业务的整体效果和效率。

第三，企业基于数据开发业务应用时，需要对数据进行处理和维护，然后综合分析。但是，目前企业中数据开发和维护的人员配置很少，甚至没有，大量数据十分混乱，无法直接使用，从而导致在企业决策时调取数据困难，难以挖掘数据之间的关联，制约了决策的科学性。

如果不能对企业后台的数据进行较高的整合和处理，在分析生产制造和零售各环节时就会缺乏数据支撑，在业务实施中就会缺乏数据指导和优化，导致企业库存失衡，运营成本提高，最终导致客户流失和利润低迷。

这确实是企业数字化之前企业运行的现实状况。因此，在企业数字化转型中亟须着手做的就是将所有数据打通，将散落在生产制造和零售阶段的数据汇总，集中进行处理和利用。数据打通可以通过企业的数据中台完成。数据中台是企业获取、整合、存储、加工和利用数据的综合平台，连接了数据后台与业务前台，实现了企业内部数据的开放和共享，帮助我们沿着生产制造和零售环节打通了数据。

也就是说，在数据中台上的数据不再是散乱的，而是经过收集、整合和处理的数据，并且在中台架构之上会有一系列的系统，比如零售管理、智能算法、调拨管理和会员管理等。数据中台上可以进行数据处理，它将数据抽象成服务，再提供给前台。所以数据中台相比于数据库系统来说，不仅要获取、整合和存储数据，也要对数据进行加

工和利用，进一步探索数据的内在价值以及数据如何对业务产生价值。因此，数据中台向业务前台提供的是可以直接为业务所使用的服务，更贴近业务前台。换言之，可以说数据中台所提供的服务都是围绕数据而来的，在数据和业务系统之间形成了闭环。同时，数据中台整合了后台的数据资源，通过加工和提炼，提升了数据质量和可用性，使得数据成为可以产生价值的数据资产，有利于快速、稳定地向前台提供服务，实现业务创新和发展。

当前国内外企业都越来越倾向于实现数据层面的一体化，已经有不少公司搭建了自己的数据中台，这是企业解决数字化转型的一个思路和逻辑。但是数据中台本身也是动态的，它连通了企业的各个部门和业务的全过程，协同各方、服务各方，其上的数据和提供的数据服务也可以根据企业业务的发展而不断地进行调整。大型企业打通方方面面的数据，将各业务阶段的所有数据均整合在一个平台上，需要极大的人力资源、技术手段和设备支撑，其难度之大可想而知。在具体的业务环境下，企业可以根据自身的情况，在生产制造和零售环节分别进行数据的全链路打通，如有需要再将生产制造和零售数据整合，提供相应的服务，反馈并驱动新一轮的业务发展。

生产制造数据打通

如前所述，数字化赋能首先要做好数据打通工作。在生产制造环节，我们需要将机器、设备、生产线、人、车辆、渠道、物流等物理形态，通过数据进行表现与加工，将不同格式的零散数据进行有效的打通与整合，建立标准规范的数据仓库，最终构建一个赋能生产制造全流程的数字化环境，实现数据的互联互通，从而挖掘数据的核心价值。

如图5-1，在生产制造环节的全链路上，从供应商到仓储物流乃至客户都有相应的管理系统，包括供应链管理（Supply Chain

Management）系统、制造执行系统（Manufacturing Execution System）、企业资源计划（Enterprise Resource Planning）系统。我们要将这些系统数据打通在一起，才能做到生产制造环节全链路数据的打通。从原料供应商到客户的管理系统将会被集成在一个数据中台上，而这个数据中台会连接各方面的数据，包括企业设备数据、能效数据、故障数据和检修数据等。只有将这些数据统一汇总在这一个数据中台上，并且嫁接现有的供应链管理系统、制造执行系统、企业资源计划系统，才能将生产制造环节的方方面面打通，为数字化转型打好基础。

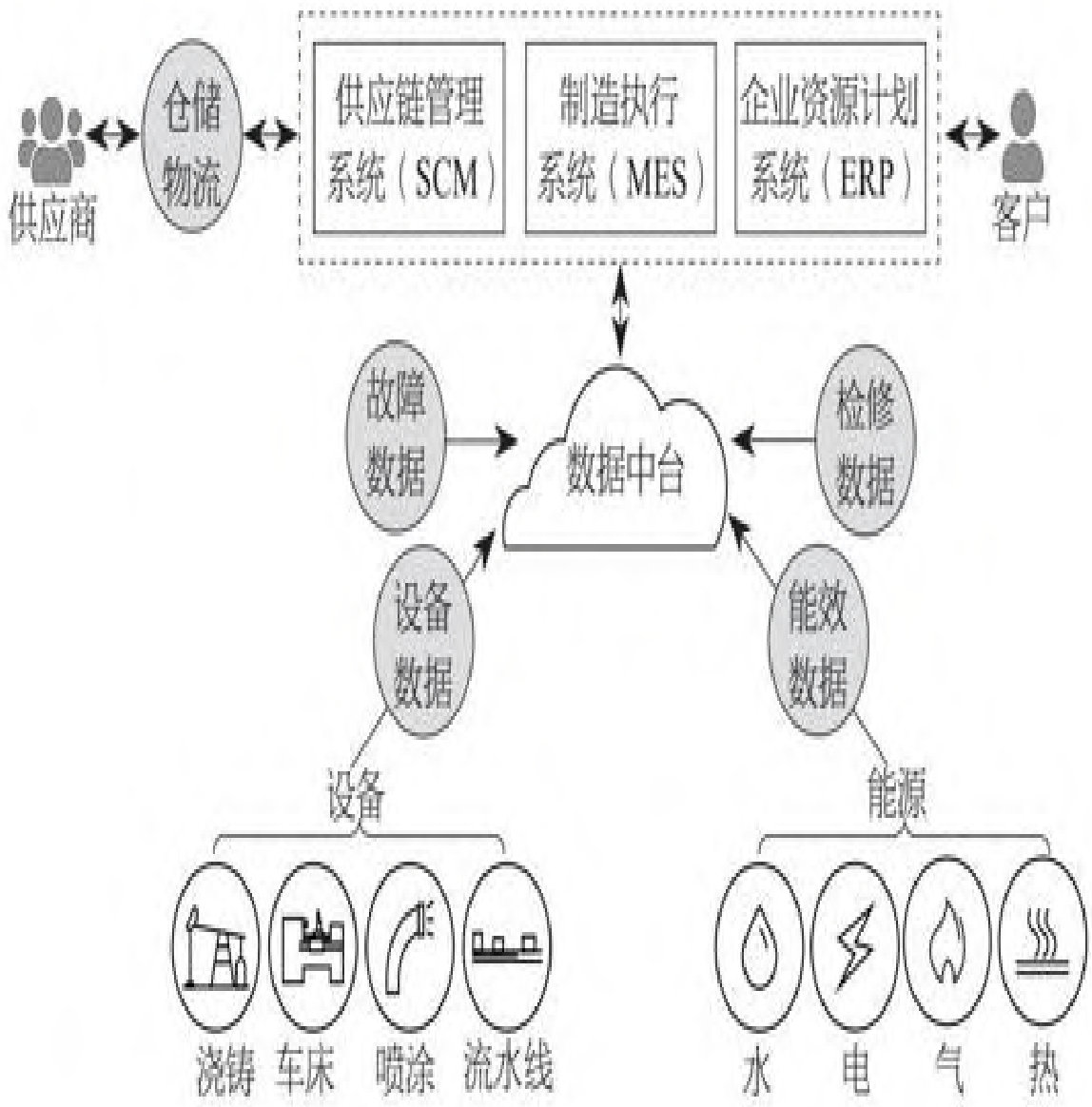


图5-1 数字化赋能生产制造全链路示意图

零售数据打通

如图5-2，在数字化转型的背景下，企业零售环节的主要节点包括：针对消费者的全域营销，针对研发设计的数字化研发，针对采购、生产制造等环节的协同供应链，针对分销流通的全渠道融合和针

对零售部分的智慧门店等。这些节点构成以消费者运营为核心的全域新零售。零售环节的全链路数据打通，就是要采集和链接以上节点的全链路数据，对业务目标进行深度优化。

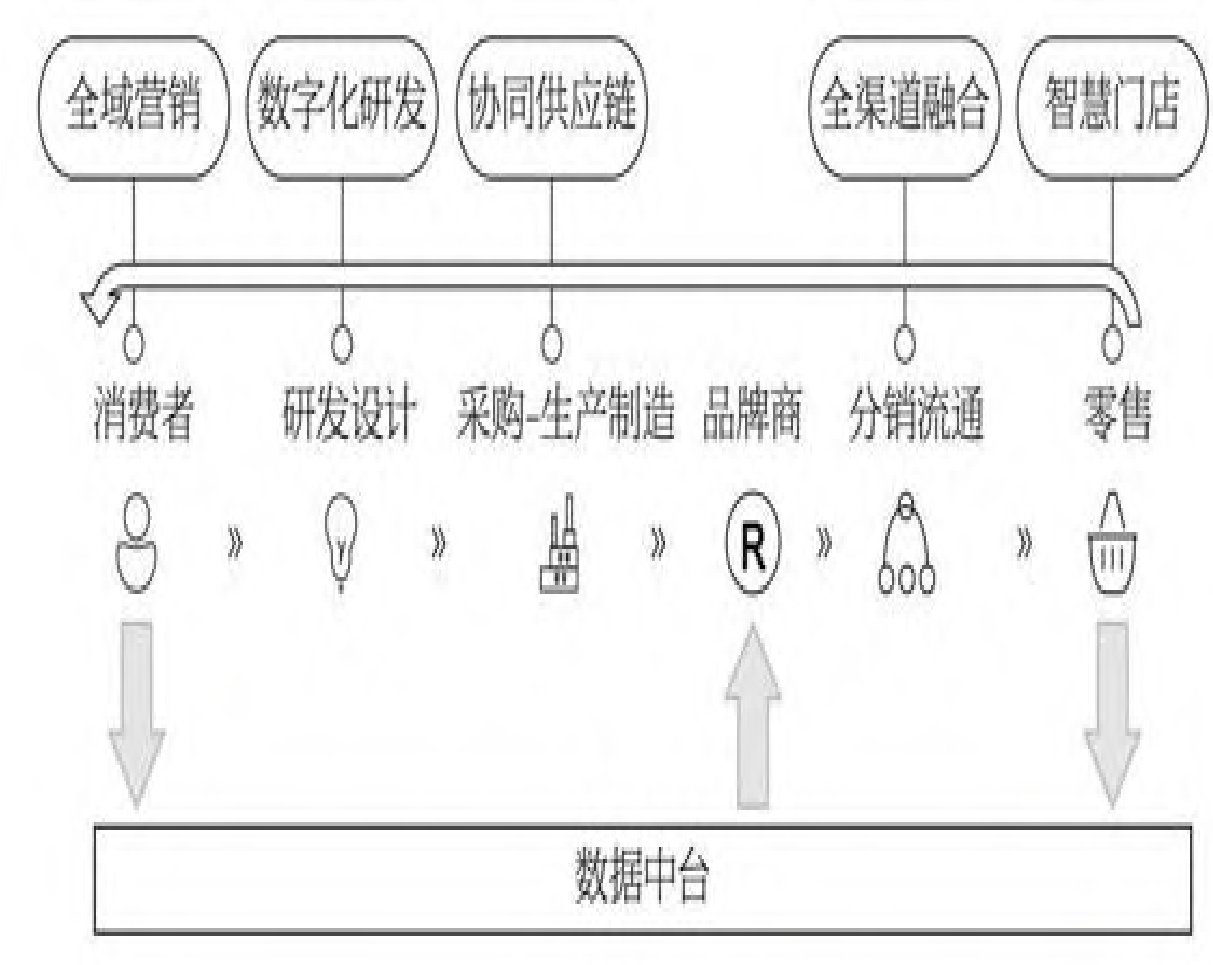


图5-2 数字化赋能零售全链路示意图

零售的全链路大致可以分划成三个部分，即消费者端、供应链端和零售端，在全链路的各个环节都可以进行数据打通，从而实现数字化赋能，具体而言：在消费者端，要对消费者进行全域的把握，消费者端包括全域营销、会员行为分析、忠诚度、付费偏好、内容运营等。其中，全域营销指的是在新零售体系下，以消费者运营为核心，以数据为原料，实现全链路、全媒体和全渠道的一种营销方法。全域

营销的出现使营销过程越来越可视化，不再出现断层，从而实现零售全链路的精细化运营。例如家电品牌戴森通过全域营销，全面打通消费者数据，进行精准营销，2017年“双11”之后，无论是消费者的数量，还是单个消费者贡献的价值，都有了显著提升。

供应链是指产品生产制造和流通过程中所涉及的原材料供应商、生产商、分销商、零售商以及最终消费者等成员组成的网络结构。供应链端主要是价格、仓储、配送上的优化，具体包括价格分析、仓储规划、配送优化、销售预测等。供应链端能够打通从原材料到最终消费的环节，它有很多可优化的地方，其中的重要一环是针对销售的预测，也即对门店或区域的产品销售进行预测。销售预测是供应链的核心，因为产品供应链应该以市场需求为导向，而销售预测能够很好地体现出市场目前需要哪一种产品，需要多少。销售预测是每个企业都应该关注的事情，因为只有做好了销售预测，才能实现其与供应链的完美共舞。例如，华北地区的知名连锁超市——物美超市，就是根据大数据分析下的销售预测进行备货和人力调度安排的。相比以前的传统模式，它的商品缺货率从17%降至10%以下，库存周转天数从27天下降至17天。

在零售端，线下门店可以通过软硬件集成的方法采集数据，在智能选品、智能进货、经营分析、营销分析、库存分析、导购管理等具体环节都可以通过数字化赋能，提升经营效益。以智能选品为例，以往线下店铺进货都是根据经验来选品，这样的商品更新速度慢，而且不一定完全符合自己店铺的风格以及消费者的需要。现在我们通过数字化手段，能够很好地解决这个问题。譬如，我们可以根据店铺位置、周边人群等信息为零售门店个性化地进行选品，帮助门店更好地卖货，降低缺品率和无效库存，用更少的成本获得更高的经营效益。

数字化和智能化相辅相成

数据中台是数字化架构中的基石，它将所有在生产制造和零售环节中的数据整合在一起。借助数据中台，企业管理者能够深刻把握各个环节的经营状况，科学、合理地制定策略。当企业各个环节的数据打通了之后，线上销售数据和线下销售数据均可以作为了解用户需求量的依据，真正实现生产销售一体化、线上线下双调节的理想状态。

以前，工厂制造的产品可能会卖不出去，在渠道中也可能不具备溢价能力，最终导致货积压在工厂里。现在的趋势是：借助数据中台，企业能够改变传统的销售模式，逐渐由渠道推动转为以客户为中心的打通价值链的运作模式。基于数据中台里打通的数据，企业可以知道客户的需求有多大，具体的需求是怎样的，并据此安排供应链和生产制造环节。通过数据中台，各个业务环节都能有所提升，实现增量收益。在数据中台发挥效果的过程中，企业能够借助AI思维实现进一步提升和优化。

我们已经知道，数据中台是实现数据打通和企业数字化的关键，打通数据后，人工智能就可以为企业做出优秀决策提供依据。“数据+人工智能”可以赋能更多的业务应用，因此，企业构建AI中台是十分必要的。在数据中台的支撑下，通过AI中台进行人工智能模型构建，不断演化和迭代，有利于向业务前台提供更加便捷和有价值的服务，加速从数据到价值的过程。数据中台为AI中台的应用提供了支撑，所以数据中台和AI中台是相辅相成的。

当然，AI中台的建立也需要一个过程。因为，人工智能在任何企业的落地都是逐渐迭代的过程，包括世界500强在内的多数企业在运用人工智能时，都会在单点上先寻找业务落地的可能性和结合点，比如在工艺参数优化、会员运营管理、积分体系设计等业务领域率先尝试。当相应的人工智能应用在业务中产生了价值后，企业就会深入到更大范围的业务场景中。人工智能在企业的落地实践只有达到了一定的规模，才能从单点式的应用发展形成一个比较完备的AI中台，再进一步服务和赋能更多的业务场景。AI中台这个人工智能大脑处理的数

据越来越多，它和数据中台的联系就越来越紧密，并且在很多场景下，AI中台还可以发挥自己特有的作用，驱动数据中台的形成、发育和完善。

具体来说，企业在业务应用中挑选出人工智能最能够产出价值的单点场景，随后考虑在这一单点上人工智能落地所需要涉及的数据，并在数据中台调取和使用相关的数据。如果数据中台上找不到这些数据，就会促成数据中台的扩建和完善，人工智能应用和从中产生的数据又可以和数据中台相互哺育和赋能。通常来说，不应该等到数据中台打通全局数据之后才升级应用人工智能，因为行业竞争是十分激烈的，等到那个时候可能为时已晚。正确的做法是选择业务上的单点场景进行人工智能落地，从而驱动整个数据体系的打通。

如图5-3，它表示的是在全链路中数据中台与AI中台各自所处的位置。数据中台将来自数据后台的原始数据进行分析和处理，最后将打通的数据传递给业务前台；业务前台产生的数据最后也会再回到数据后台中，形成反馈的闭环，实现生产制造和零售环节的数据连通。数据中台和AI中台之间最良性的架构应该是逐渐演化的过程，这同时也是人工智能大脑学习发育的过程。人工智能在企业发挥作用的前提是对企业业务全流程的充分理解和掌握，所以必须获得全链路打通的数据。AI中台通过打通的数据，在模型的加工下，形成决策。这样一来，人工智能应用也会驱动数据中台的构建。同时，AI中台通过模型计算出的数据可以再次回到数据中台之上，不断循环和加工，并提供给业务前台使用。在实践过程中，为了让人工智能顺利落地，企业应当花费充足的时间与精力去思考和规划AI中台及数据中台的搭建。在这之前，可能大部分企业都没有这个需求，如果单纯靠人工进行工作，并不需要打通所有的数据，并且人工也没有能力去分析所有的数据。所以，企业的数据中台在人工智能需求的驱动下会越来越完善。

数字化与智能化相辅相成，数据中台和AI中台相互促进，共同发展。没有数据中台就没有AI中台的用武之地，有了AI中台就能产生进

一步优化的业务结果，这时候企业管理者也就更有动力继续深化和提升数字化的进程。从这个意义上来说，人工智能也是企业升级的工具，可以推动企业内部的基础设施建设。由于数据打通和中台构建会产生较高的成本，所以企业实现数字化基础设施的升级不应该由技术层面驱动，而应该由具体业务场景的需求来带动。数据打通以及中台构建完成之后，就能够帮助企业提升业务场景的效益，产生增量的价值，最终也能够支持数字化基础设施持续的升级和改造。

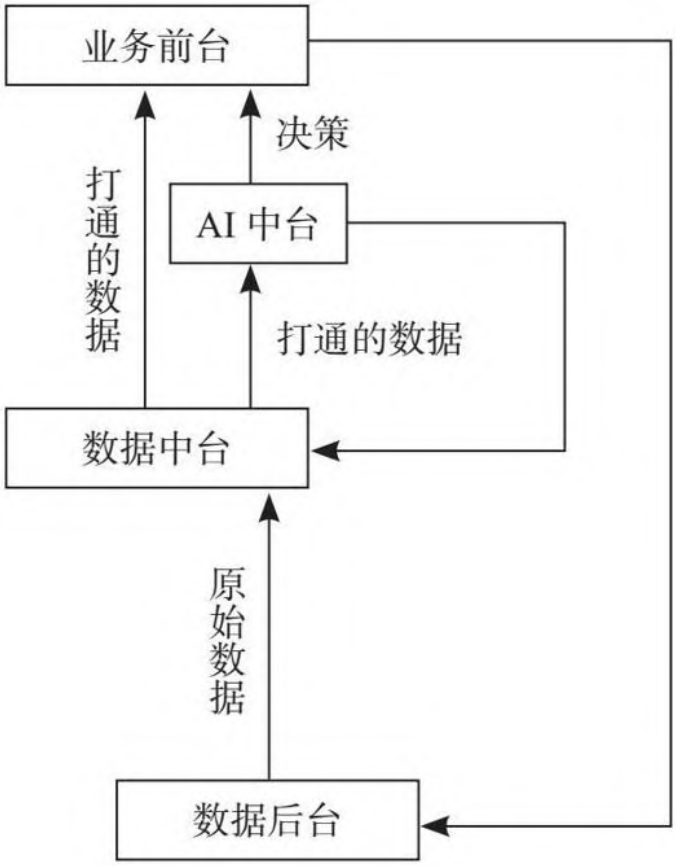


图5-3 数据中台与AI中台关系示意图

数据可视化平台

我们生活在一个充满了数据的世界里，但面对纷繁复杂的数据，我们经常感到无所适从。“得数据者得天下”的说法此起彼伏，但要通过数据的整理和分析来获得一个准确的结论并非那么容易。我们的大脑对视觉信息的处理要优先于文本信息，所以我们越来越倾向于利用可视化的形式将数据表示出来。数据可视化就是借助图形化手段，清晰有效地传达信息，以此帮助人们更好地理解数据。例如，我们早已习惯通过卫星云图来了解天气状况，也需要借助柱状图、饼状图等来汇报工作进度。而商业活动涉及的数据量非常庞大，数据之间的关系错综复杂，更加需要一个专门的数据可视化平台来直观地展现各项业务指标。

数据可视化平台要实现的是全链路数据的可视化。它具有两方面的作用：一是通过可视化的统计图表等形式呈现，让数据变得更直观；二是提供全域视野，打通不同类别、不同时间、不同地域、不同环节的数据，综合呈现。正是因为我们在数据中台上完成了每一个节点所产生的数据的收集、记录和连通工作，并将其与数据终端连接，实现了远程实时监控，可视化平台才能够像一个镜头一般，跟踪掌握整个流程中每一个节点上事物的状态和变化。

图5-4是一个服装电商的产销协同数据可视化平台示意图。这家电商覆盖的销售范围很广，有线上线下的销售渠道。从可视化的角度来说，在生产、物流和消费端，我们能够利用产销协同大屏监控非常直观地得到一些数据，比如生产端的生产计划准确率、订单匹配率、订单满足率和库存荷载率等；物流端的物流计划准确率、库存周转天数、物流满载率和物流周转天数等。这让我们能够相对轻松地将整体运营数据了然于胸，以便对具体的操作事务做出及时调整。比如当发现库存周转天数有增多的迹象时（这可能意味着有滞销的潜在隐患），我们可以从生产端开始，马上调整相应产品的产出量，也可以进一步去查看物流端是否出现了怠慢的问题，或者在消费者端相应地推出一些促销活动，等等。这些数据的即时呈现，能够在很大程度上为企业降低风险，减少损失。

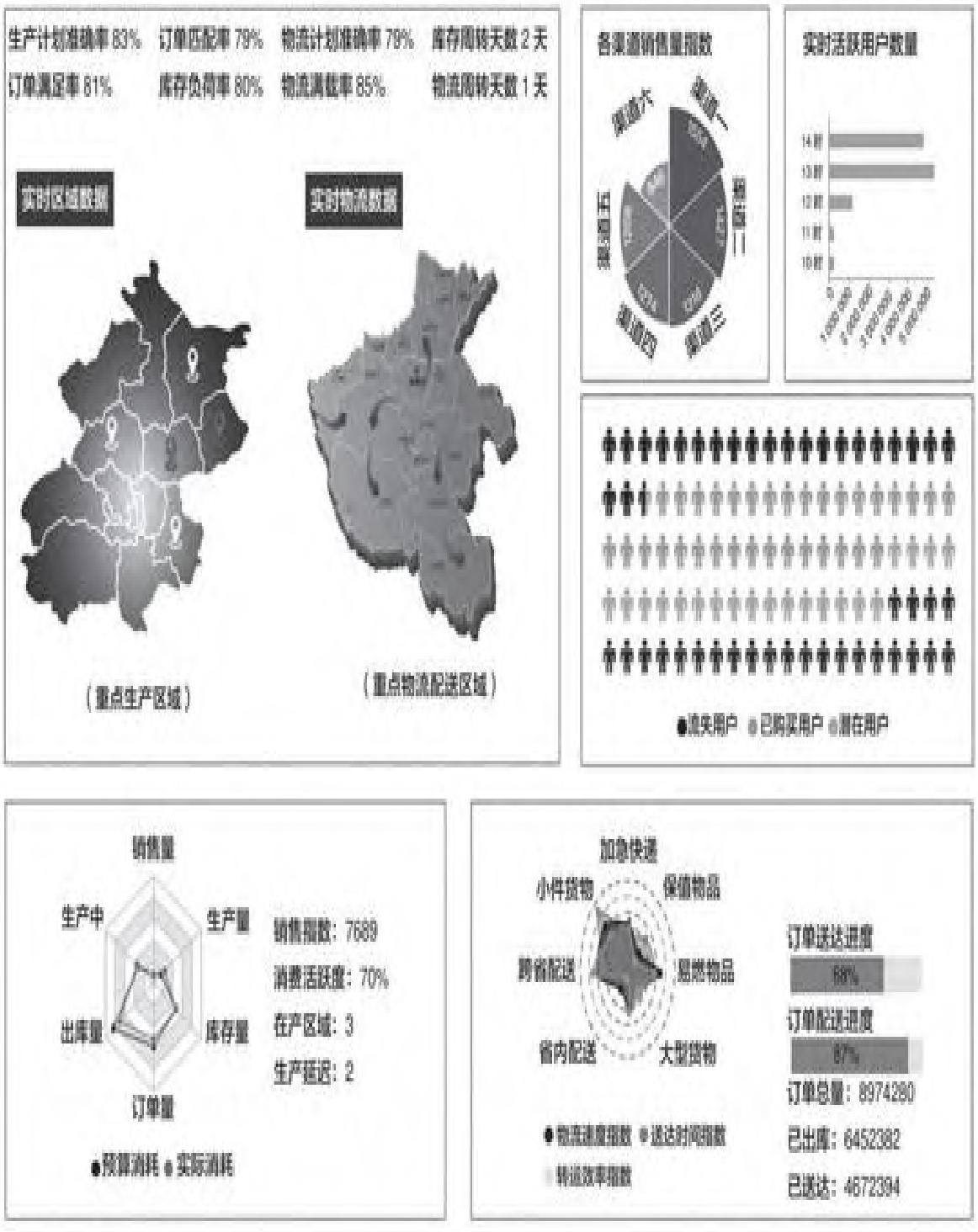


图5-4 数据可视化平台示意图

在用户端，通过可视化平台也可以非常直观地看到活跃用户、流失用户、已购买用户、潜在用户等的数据，这些数据在各个环节的运作过程中都能够清晰展示出来，并进一步指导企业做出精细化决策。比如像促销活动，我们只知道要在哪个节点做促销活动是不够的，我们还需要了解这个促销活动什么时候做、做哪些、怎么做等一系列问题。要想做到数字化落地，就必须要把这些问题具体化。而用户端的这些数据可以为我们提供帮助。比如，根据可视化平台提供的近期每天使用人数，我们可以知道哪天属于淡期、哪天属于旺期，以及这些日子的特征，是周末、月末，还是都是工作日，我们从中选取适当的时候来做精准促销，提升业绩。

数据可视化可以避免出现盲人摸象的情况，数据可视化平台呈现出的全域数据视野，为探寻业务运作的各项问题提供了机会和助力。对于企业决策者来说，他们肯定更愿意看到可视化的数据模式。当然这些可视化的前提假设和实现手段是已经有了打通的数据，然后才可以做到上面所说的可视化呈现。

数据中台提供了数据的获取、整合、存储、加工和利用的服务，让我们能够更好地认识数据和使用数据，这种服务是任何业务场景都需要的。而AI中台提供的是更加具有创造性和个性化的服务，通过人工智能模型发现不同场景中的精准需求，提供业务创新的机会，从而实现业务增长。因此，数据中台和AI中台是适用于业务发展不同阶段的基础设施，数据中台为AI中台提供了基础和养料，AI中台的需求也会反过来促进数据中台的成熟。相应地，数字化与智能化也是企业发展的不同阶段，二者并没有严格的界限区分，数字化和智能化融合发展，相互促进，相辅相成。最后，通过数据可视化平台，我们可以更加直观全面地了解到商业活动中每个环节的状况和发展态势，这样在进行决策时就能更加有的放矢。

工业赋能：数字化的物理世界

18世纪末，蒸汽机的出现使人们进入工业1.0时代；20世纪初，电力的大规模应用使整个世界进入工业2.0时代；20世纪70年代，电子技术、信息技术以及工业机器人的广泛应用推动世界进入工业3.0时代；现在，随着物联网、大数据和人工智能的融合应用，我们已经进入了工业4.0时代。在新一代技术革命浪潮中，数字化趋势已经势不可当。我们需要利用数字化技术促进产业变革，尤其是通过人工智能的应用创建具有适应性、高资源效率和更加人性化的新一代智能工厂。在这样的背景下，不仅需要将工业所处的物理世界数字化，数字化的目标还需要反作用于物理世界，驱动物理世界，改变物理世界。

数字化反作用于物理世界的第一步就是将物理世界抽象成标准化的数据集合，这是物理世界向数字世界转化的基础。这一步可以借助物联网（Internet of Things，简称IoT）完成。物联网，顾名思义，就是“万物相连的互联网”，是将各种信息传感设备与互联网结合起来的网络，它可以在任何时间、任何地点将能够独立发挥作用的物体连接在一起，实现互联互通。所以物联网的核心和基础仍然是互联网，它只不过是在互联网基础上进行了延展，扩展到了任何物品与设备之间。物联网上涵盖了大量的数据，各设备相互连接并且不断地进行数据交换。将物联网上的数据经过数据采集、数据分析和数据处理后，我们再把这些数据存储在标准化数据集合里，物理世界的抽象过渡就基本完成了。

但仅仅完成数据的抽象是远远不够的，工业4.0时代要做到数字化驱动物理世界并改变物理世界。因此，这里我们要引入一个概念——CPS（Cyber-Physical Systems），即信息物理融合系统。这是一个综合计算、网络和物理环境的复杂系统，它将计算、通信与控制深深地嵌入物理过程，使之与物理过程密切互动，实现了计算进程和物理进

程的统一，从而给物理系统添加了新的能力。信息物理融合系统包括了很多智能系统，通过建立物理世界数据的人工智能模型，在数据采集输入后，系统可以进行管理、操作和预测，从而把相应的决策和洞察再反馈到物理世界。

所以，CPS紧密连接了物理世界和数据计算，通过计算进程和物理进程相互影响的反馈循环，实现深度融合和实时交互来增加或扩展新的功能，达到实时监督或者控制物理世界的目的。也就是说，CPS不但可以看作工业数据的集合地，而且同时将这些数据进行机器学习或深度学习，建立起物理世界的数据模型，并且基于这些数据模型产生智能预测和决策，从而最优化地驱动和改变物理世界。

总体来说，物理世界的数字化过程首先要有数据，然后通过IoT打通生产制造环节将数据整合在一起，再通过建立相应的人工智能模型，进一步理解整个生产制造过程，比如其中的某一个步骤需要怎么调控参数，怎么优化生产制造的目标，从而形成最优化的决策，实现效果的提升。其中，CPS不仅可以实现针对物理世界的实时计算，还可以通过数字世界建立的模型将操控指令反馈到物理世界，驱动物理世界的调控和改变。

工业数字化赋能实现路线图

我们已经明确了工业数字化的总体目标：不但要将工业所处的物理世界数字化，而且还要将数字化反作用于物理世界，驱动和改变物理世界。如图5-5所示，工业数字化赋能的实现可以分为五个阶段：试点连接、全面连接、数据洞察、业务优化和链路集成。



图5-5 工业数字化赋能阶段示意图

在试点连接阶段中，要完成试点工业设备连接、数据采集集成、设备故障收集和设备性能分析等，这是最简单的数据收集，可以在单点上完成。此后进入全面连接阶段，把所有生产设备、检测设备和客户反馈等相关数据通过IoT和其他手段收集到数据中台上，完成大规模数据的采集和全链路数据打通。

下一阶段就是在数据基础上，搭建初步的CPS，形成完整架构，进行数据洞察分析。在这一阶段中，通过收集的全链路数据，建立人工智能模型，进行设备故障诊断、故障预测、可靠性分析、寿命预测和质量分析等。接下来继续完善CPS，根据预测性结果，进行业务优化，对整个工业生产链路做出相应的资产性能优化、维护策略优化、备件库存优化和产品性能优化等。业务优化阶段完成后，进入链路集成阶段，实现上下游信息和流程的集成、资源的共享，完成从生产型企业向服务型企业的转型。这也是企业成为数字化领导者的实现路线。整个过程从分步连接数据，到数据洞察、迭代和扩展，再到链路集成和整合优化业务流程，最终形成业务模式的创新。

通过工业数字化赋能的路线图，我们可以更形象地理解智能生产制造的各个阶段——从试点连接阶段，进阶到在IoT平台上连接更多设备和节点，对数据进行全面洞察，做出预测，从而优化业务目标，更好地提升产能、降低能耗、降低故障率等。工业生产的指标都可以通

过数字化的方法进行优化，最终形成整个链路上的集成，向智能制造方向过渡和发展。

在工业数字化赋能的路线图中，要实现数字化赋能，最重要任务之一就是建立人工智能模型，也就是建立相关输入因素数据x及目标值y之间的相关函数关系。如图5-6，模型的作用就是在获得相关输入因素数据（例如生产数据、能耗数据、质量数据和环境数据等）以及和生产目标相关的输出值后，对这些数据进行学习，进而通过人工智能产生因素数据和目标值对应关系的模型。在实际生产过程中，工厂将各生产过程中感知到的数据导入生成的模型，根据输出的结果完成决策和反馈，实现工艺参数推荐、过程优化和设备综合效率优化，最终实现预期的生产目标，如提升优良率、降低成本、提升效率或是节省能源等。

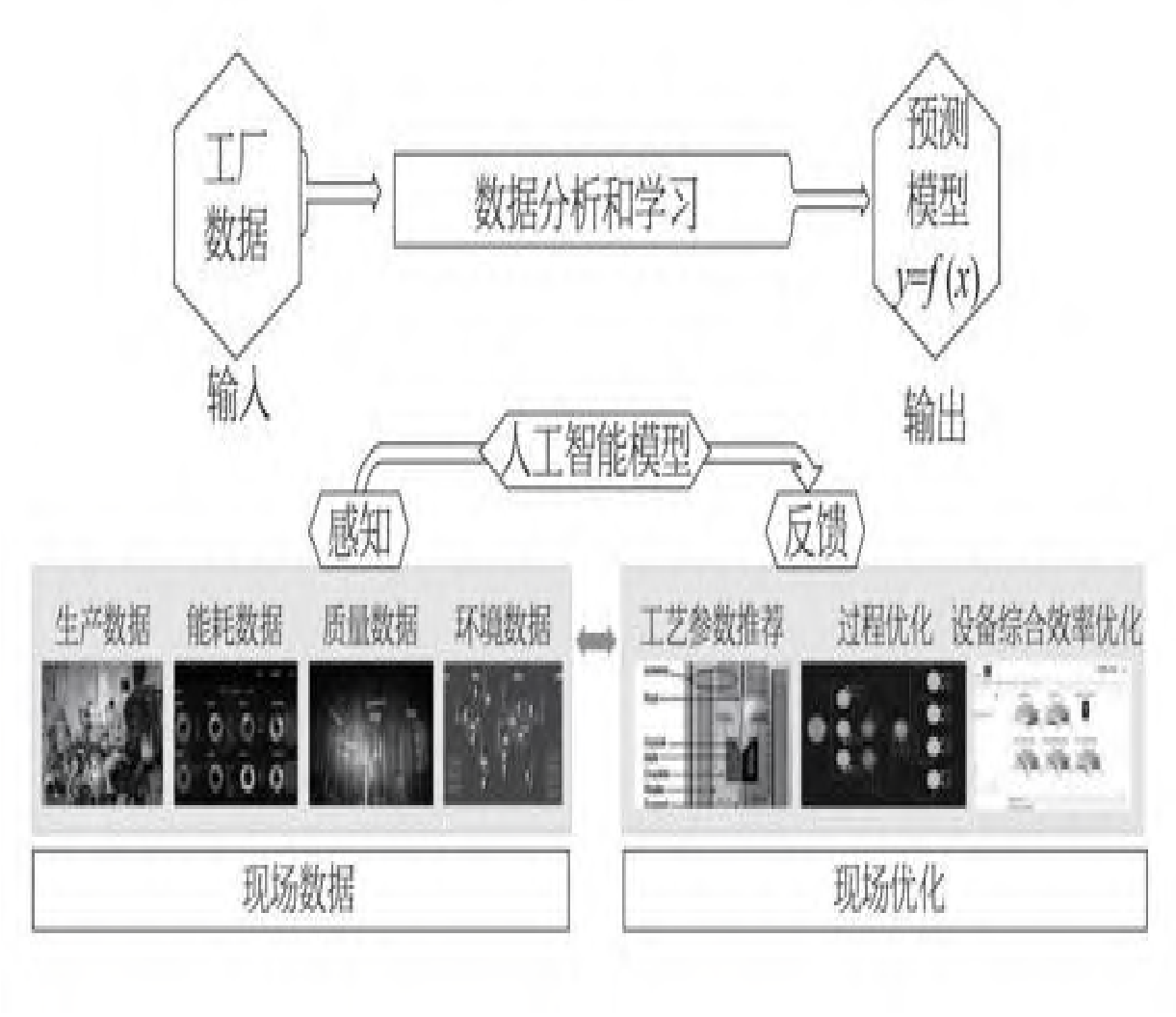


图5-6 工业数字化赋能运作流程示意图

所以简单地说，工业数字化赋能的核心是提供全链路人工智能模型服务，通过分析工艺参数和关键指标之间的关系，进一步预测生产制造环节的关键指标，再根据预测结果完成相应的操控决策和反馈的闭环，最终为企业带来良品率提升、能耗降低和效益增加等效果。数字化赋能的过程完成之后，操作人员在生产制造中每一个环节需要关注的重点、操控的参数等，就都可以被科学地预测和制定。这样就可以做到生产制造环节的数字化赋能和优化。

以上的数字化赋能主要是针对生产制造而言的。虽然生产制造和零售是业务中两大不同的环节，但这一数字化赋能的方法论是通用的。也就是说，在数字化系统建立起来后，只要设定了业务目标，我们就能通过学习输入和输出之间的相关性进行预测性判断。至于如何操控系统参数，如何调整业务流程等，对于生产制造和零售环节的应用来说是异曲同工的。

国内智能工业案例

工业，在我们大部分人的眼中，就是一些与钢筋水泥打交道的产业，并且一直被划分在传统行业的范畴内。早期，行业内甚至流传着“工业不需要互联网”的言论。但随着技术的完善和落实，数字化、智能化系统越来越多地在车间中投入使用并取得了较好的运作效果。此时，“工业不需要互联网”的言论开始出现崩塌的迹象，“工业大脑”逐步在各地工业企业中广泛应用，不断攻占工业制造的各个领域。

通过数据的全链路打通和人工智能模型计算，智能工业可以深入到各行各业，达到赋能工业的目的。这样的例子比比皆是：在汽车行业，基于人工智能视觉处理的汽车工件合格性检测系统已成功在汽车零部件制造企业得到应用，有效地提高了汽车工件不良品检测效率；在烟叶行业的分级应用中，按照烟叶分级的国家标准，针对不同地域提供烟叶分级模型，可以有效地降低烟草公司人力成本，实现高效高质的分类。

智能工业可以运用到工业流程的各个环节，例如，在对产品的质量检测中，通过训练多层神经网络模型，可以对物体表面的缺陷进行大小、位置、形状的检测。在不规则缺陷检测方面，智能工业的准确率与传统模式相比有了明显的提升，譬如通过深度学习训练LED（发光二极管）芯片缺陷识别及分类模型，能够有效提升产品质量和检测效

率。再比如，光伏的电子发光检测应用通过数字化和智能化手段支持单晶和多晶暗域、黑边、黑角、各种隐裂等十数种缺陷识别，全面提升了缺陷分类准确率。

人工智能在工业领域成功落地的案例有很多，华菱星马就是其中的一例。华菱星马是一家重型卡车、重型专用车的生产商，在全国汽车生产领域占据着重要地位，其节能环保大功率发动机智能工厂项目已获批国家制造业互联网融合试点示范项目，其汽车生产的智能水平目前国内领先。华菱星马通过对工厂进行数字化、智能化的改造升级，建立起了多种智能化信息系统平台并进行相应集成，从产品的设计开发、生产、销售直到服务的全生命周期中，对各种数据进行采集和分析，真正实现了产品生产的数字化和智能化。

在发动机质量控制方面，华菱星马通过多种系统平台和技术手段实现了数字化赋能。在进行数据采集的同时，华菱星马还兼顾设备运行状态和质量管理的监控以及控制功能，统筹协调由多台全自动在线测量机和几十套装配辅助设备等组成的发动机生产线。如果接到生产订单，系统可以直接将订单信息及物料清单传递到生产线进行生产。在生产过程中还能记录产品的质量信息以及其他重要数据，形成完整的产品信息档案，搭建产品质量数据库，生成产品质量报表，以便在产品进入市场后进行产品质量追溯。

在提升服务质量方面，华菱星马基于现有的车联网系统平台研发了出一套智能服务系统“智慧华菱”。该系统借助人工智能，可以统计出自定义区域内车辆的通过次数，分析出车厂服务站的覆盖能力；其远程车况诊断功能，可以实时把控车辆的车速、转速、档位、里程、油耗、平均油耗、小时油耗、运行时间等，让用户对车辆的运行状态了如指掌，并且为预测性维护提供了有效数据支撑。对以上数据进行分析，也可以反作用于生产决策，对产品设计和性能提升起到积极的作用。

工业数字化除了应用于产品生产线上之外，对于优化产品质量也起到了举足轻重的作用。例如，天合光能作为全球最大的光伏组件供应商，电池片是其重点供应的组件。但有一段时间，天合光能在电池片的质量优化方面遇到了问题。他们尝试了多种方法，最终通过工业大脑平台解决了这个问题。工业大脑平台基于大数据，结合了人工智能以及工业领域的专业知识，在提高产品良品率、优化产品生产过程等方面发挥着很大的作用。该平台搜集了天合光能电池片从接收订单到产品生产完成全过程的所有数据，并进行了深度分析，再结合业务智能算法平台识别出生产过程中对电池片质量产生关键影响的重要环节，针对这些环节提出优化方案，进而在实际生产过程中不断调节该优化方案，直至达到电池片的质量优化目标。在工业大脑平台的帮助下，天合光能的电池片生产良品率提高了5%，对天合光能产品整体质量的提升都起到了有效的促进作用。

从20世纪70年代开始，能源就和人口、粮食、资源以及环境一起被列为当今世界面临的五大问题。能源短缺会给人们的工作生活带来很大不便。而且，随着工业的发展，人们对于能源数量和质量的要求越来越高，加剧了能源短缺的问题。此外，能源使用过程中排放污水、废气等污染环境的问题也日益凸显。基于数字化的智能工业能在节约资源、提高资源利用率、防治环境污染等方面大有可为。

例如，三众能源就利用人工智能为全球累计降碳70多万吨，累计节约电量将近60万千瓦时。三众能源是一家综合型的能源科技公司，响应国家可持续发展战略号召，为节能减排的重灾区——工业，提供人工智能帮助。三众能源研发的人工智能节能产品应用在燃气锅炉、空气源热泵、空调等大型能源系统上，能够很好地实现节能减排。人工智能节能产品的具体运行过程如下：首先，通过数据采集器，实时采集能源设备的运行状态、环境参数以及用户偏好等各种数据；然后，对这些数据进行人工智能分析，及时调整、控制能源设备的运行，达到节能减排的效果。

这套人工智能系统，应用在中铁建工集团开发建设的商业建筑里，为其提供了能源、电力的精细化管理服务。首先，通过对室内外环境的感知数据和住户对温度的需求数据的分析，人工智能能够对提供热能的空气源热泵进行自动调节，真正实现了按需供能。这样既节约了电能，又一举两得地减少了二氧化碳的排放。其人工智能节能产品可以保证供暖数据实时上传，并具有故障远程报警和定位等功能，如果出现问题可以及时发现。即使在无人值守的夜间，人工智能也可以根据数据自动调整出相应的解决方案。最终，这套系统取得了良好的节能减排效果，一年之内为该商业建筑节约了电能70多万千瓦时、减少了二氧化碳排放量80多万千克，在电力方面总计为其节约成本85万元。

国外智能工业案例

数字化为处于工业4.0时代的企业带来了无数的机遇与选择，基于数字化的人工智能推动打通生产流程，使企业能够高效节能地运作并做出更加准确有效的决策。所以数字化对于新工业时代的发展是一个决定性的存在，是实现工业4.0的基础和最重要的步骤，也是衡量企业业界水平的一大标准。除了国内的案例外，国外的工业数字化运用也早已如火如荼。

世界五百强公司之一、美国最大的电力公司杜克能源通过多样化的能源组合（如火电、风电、核电等），为用户提供包括电力供应、能源服务和能源运输等在内的多种服务，但与此同时，杜克能源还存在着厂房老化、发电厂运行效率低下、预测性维护能力不足等问题。杜克能源在其系统中已经具备了预测性维护措施，但这项措施有80%的过程是在采集数据，仅有20%用于分析，这就导致了杜克能源无法准确地预测系统问题。为改变这一现状，杜克能源将数据打通，并使用人工智能进行深度分析，对预测性维护系统和设备管理进行升级改造，

实现了60个以上电厂的智能监测，使得杜克能源通过预测所避免的损失达3 000万美元——这个金额早已远远高于项目成本。

巴西盖尔道集团是世界知名的钢铁制造商，在全球拥有超过35 000名员工。面对高能耗的钢铁生产过程以及高昂的人力成本支出，严格控制设备成本和维护设备的能力对盖尔道集团的盈利至关重要。仅在巴西，盖尔道集团就有11家工厂生产钢材，每年都需要巨额的费用去维护。即使是出现一个小小的设备故障，每小时也将造成上万美元的损失。在复杂的钢铁制造环节及多地区的场景下，盖尔道集团通过使用工业物联网，将集团所有生产设备的数据进行采集分析，实现了各类设备运行及故障情况的监控，即使是最轻微的异常也会被捕获到。同时，盖尔道集团还通过人工智能模型，识别出设备各类潜在的故障，并及时预警及采取相应行动。通过数字化和人工智能的手段，盖尔道集团一项原本计划18个月实现投资回报的项目，实际缩减到了8个月；同时在6个月内，盖尔道集团成功预测了2个严重故障，避免了对生产过程造成的重大影响。

在全球的智能应用里，视觉检测也是比较普及的一个应用。视觉检测已经在工业自动化系统里应用很长一段时间了，其涉及的领域遍布各行各业：仪表板的智能集成测试、汽车车身的检测、纸币印刷质量的检测和流水线生产检测等。视觉监测在这些领域中都起到了采集产品数据、提高生产效率和控制产品质量的作用。例如烟草行业就需要这样一款可以对工业生产线进行质量检测的机器视觉软件。传统的机器视觉检测的都是一些规则的产品，而这款软件能够应用于一些比较复杂、难以分类的产品的质量检测，比如说烟草的异物剔除和质量检测等。基于深度学习，软件首先对烟草产品进行图像采集，之后进行分析计算，最终再进行检测，大大提高了烟草产品质检的准确率，与人工质检相比，还大幅提高了工作效率。

如果说数字化转型是整个时代的机遇与挑战，那么数字化赋能就是企业抓住机遇和迎接挑战的重要本领。通过数字化、智能化的手

段，企业的各个环节都被连接在了一起，各个方面都被赋能，既节省了生产成本，又提升了整体运营效率。AI思维就是这样一种全面的思维，使我们在认识事物时能够做到面面俱到、统筹分析，做出正确决策。实现数字化赋能，必不可少地要经历这样一个过程：深入全面地采集数据，将物理世界数字化后，通过模型进行深入的数据分析，得出相应结论再在物理世界中进行应用，最终实现全面的工业赋能。数据是AI思维的基础，业务模式是AI思维的应用，数字化将二者串联起来，使AI思维真正能够在业务场景中落地。

零售赋能：智能产品定位和营销

生产制造和零售是商业运作的两个大环节。我们对生产制造环节，可能不太熟悉，接下来要说的零售环节则非常贴近我们的日常生活，其中的数字化赋能实际落地案例也比比皆是。比如，最近你正想买衣服，浏览的网页似乎有“读心术”一般，恰好呈现给你放在购物车里好久的衣服；或者夏天天气热，你正想去海边度假时，手机里的推送刚好就是泳衣、泳裤和各种潜水装备的广告，而且都还是你一贯喜欢的风格。其实，这些事情都是智慧零售的应用成果，智慧零售可以通过数字化赋能手段进行提升。

在零售环节，数字化赋能大有用武之地。首先是智能产品定位，即依靠大数据预测消费者诉求和潮流趋势，从而实现产品创新。也许你会有点疑惑，产品创新不是企业一直都在做的事情吗，那么智慧零售所做的产品创新又有什么不一样呢？为了解答这个问题，我们先来看看产品的创新模式。图5-7就是一个产品创新模式示意图，左边为传统研发模式，右边为数字化研发模式。

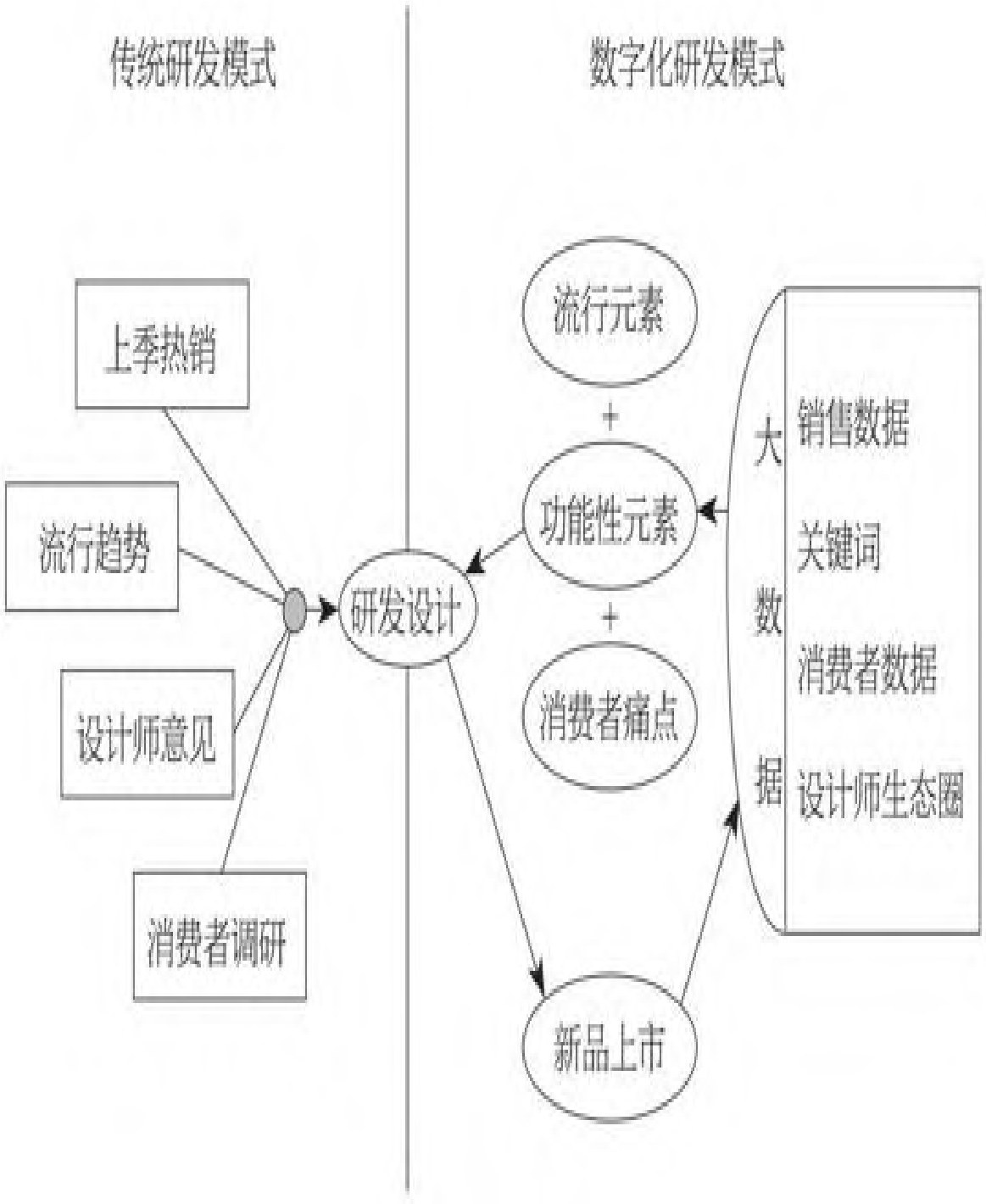


图5-7 产品创新的传统和数字化研发模式对比示意图

传统的产品研发模式主要依赖于主观判断。传统产品研发设计主要依靠的是一种“经验主义”，比如参考上一季热销产品的反馈数

据，上一季如果卖得好，下一季则继续按这个风格研发。或者，传统产品的研发设计还可以通过搜集、参考同类型产品中的流行商品的主要特征进行，但这种学习太浮于表面，很难做到真正的产品创新，画虎不成反类犬，更像东拼西凑的结合体，适得其反。此外，还有一种路径是根据设计师的个人意见来决定产品的最终形态，但很多时候，设计师的审美与消费者的审美是存在很大鸿沟的，这就使得产品难以与消费者的需求匹配。所以有时设计师们也非常苦恼，无法把握潮流，不知道消费者的诉求所在。当然，更实际的做法则是进行消费者调研，采取定性的调研来辅助设计，但是消费者调研成本高、耗时长，而且很难保证这些消费者是否是潜在消费者或者典型消费者，在信息反馈质量和样本数量上都有一定的局限性，同样会犯以点概面的错误。偌大的市场中，消费者的需求各有不同而且日新月异，个体的喜好早已无法代表整个市场的热度，传统研发模式潜在的痛点已经浮出水面：企业难以把握潮流趋势以及消费者诉求，研发出来的新产品无法完全满足消费者的诉求，从而导致消费者对品牌的忠诚度降低。

相比于传统的主观研发模式，数字化的产品研发模式则是通过客观的定量数据来反映市场趋势，将销售数据打通，通过数据洞察综合分析产品特征和消费者需求，再做出研发设计决策。因此，数字化的研发模式不再是过去的经验主义和局部主义，其基础和关键是最大程度上囊括所有市场数据。分析的数据范围，包括销售数据、关键词、消费者数据和设计师生态圈。销售数据，就是具体产品所在品类的整体销售数据，包括品类各品牌销售情况、品类明星单品等全方位数据。关键词是指将搜索引擎、社交网络和电商平台评论中的关键词，需要对其进行归纳。消费者数据就是消费者相关的数据，包括消费者群体画像、消费场景数据和消费习惯数据等。设计师生态圈则是由设计理念、设计产品、交流学习以及作品展示等各种内容组成的，在设计师的生态圈中出现的高频元素也是数字化产品研发模式的重要参考数据。在收集完上述一系列数据之后，可以在数据中台上进行打通，判断出相关产品的流行元素、功能性指标和消费者痛点，并借此指导

新品的研发设计。新品上市后，随即产生的一系列相关数据会再反哺到数据中台上，为下一轮的新品研发做指导。这样，数据化研发模式整个体系形成正向循环，类似春泥一般，叶落成肥，滋养大地，又反哺出下一个花季。

以宝洁的产品研发为例，根据品牌多年积累下来的相关销售数据，总结分析产品触达的不同种族、跨年龄段消费者的规律和特征，宝洁研发部门抓住了很多消费者需求特性，如亚洲女性除了要面对松弛、皱纹等衰老问题外，面部轮廓逐渐横向外扩也是影响面容的一大问题。这是消费者的痛点，也就是研发部门的努力方向。根据这个发现，宝洁推出新品——OLAY小脸精华，从感官上让消费者显得更年轻。经过这次的针对性研发，宝洁当季度中国区销量增长30%，电商销售额上涨80%，打了一场漂亮的数据战。

全域洞察

我们刚刚说到，数字化的产品研发需要基于消费者数据，数据越多，就越能精准地反映出来消费者需求和市场趋势，所以数据的广度是智慧零售中的一个关键问题。那么数据到底要多广呢？这就涉及“全域洞察”的概念。

全域洞察是消费者洞察的一种实现形式。所谓洞察就是深度理解，消费者洞察就是通过数据理解消费者，抓住消费者的真实需求，这是面向消费者营销的核心。图5-8是全域洞察的示意图。全域洞察的数据来源主要有品牌的活动平台、自营和第三方电商平台、媒体平台以及各种内部平台。

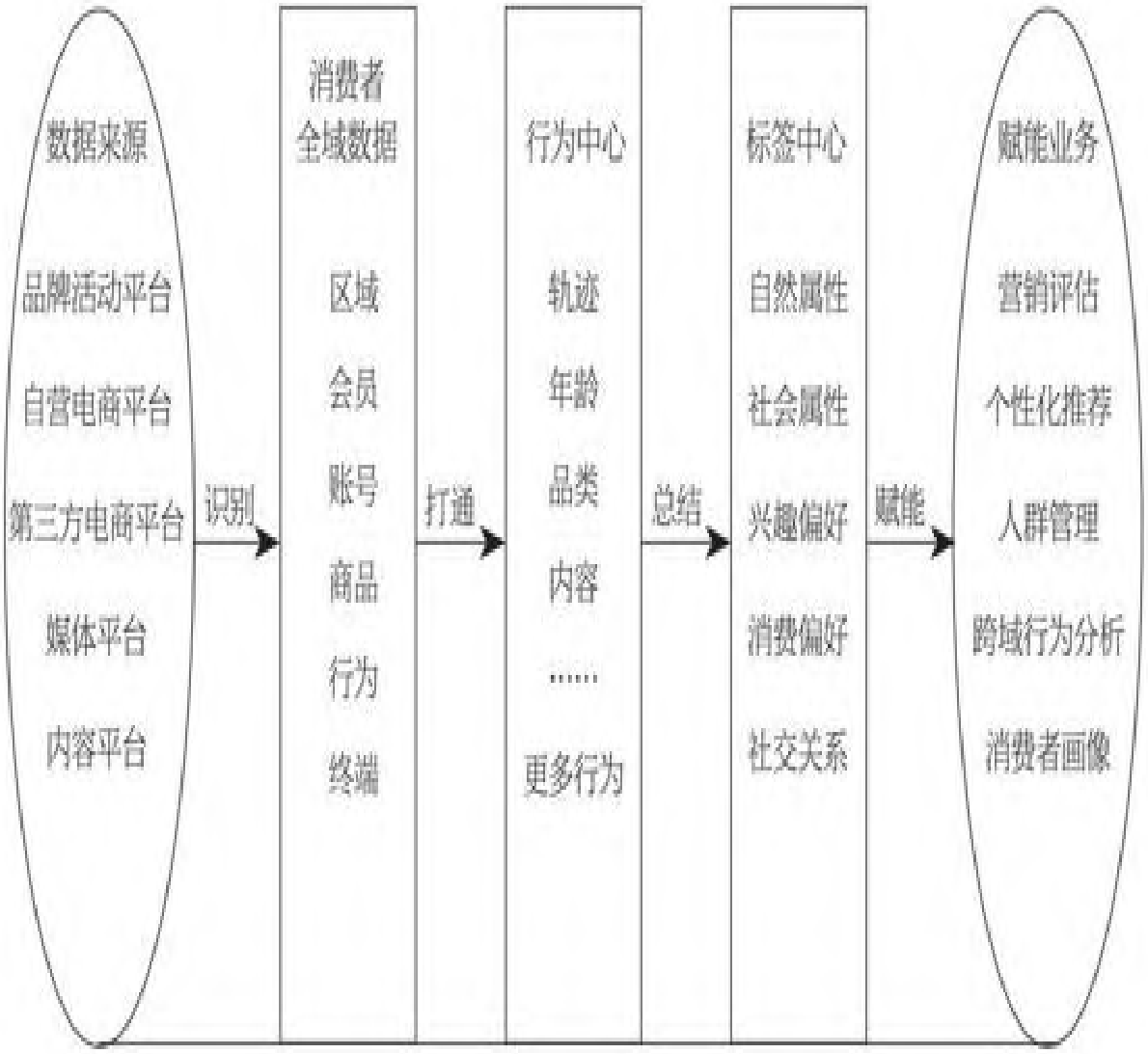


图5-8 全域洞察示意图

洞察关注的消费者维度包括所在区域、会员、账号、商品、行为和终端等全域数据，将这些数据围绕消费者进行打通，将同一消费者的数据汇集到一起。数据整合后，我们将会对消费者的关键信息进行提取，例如轨迹、年龄、品类和内容等，构成行为中心。行为中心里面聚合了各个业务领域中消费者的行为，包括主体、时间、行为、内容及行为度量。最后，根据这些信息归纳总结出个性化标签，例如消

费者的自然属性、社会属性、兴趣偏好和消费偏好，形成标签中心。与此同时，标签中心中还包括社交关系，即消费者之间的关系网络。数据的整合、处理和标签化是消费者全域洞察的核心过程，它可以在数据中台上完成。

消费者洞察可以赋能实际业务，例如营销评估、个性化推荐、人群管理、跨域行为分析以及消费者画像。比如个性化推荐，当我们真正做到对消费者的真实洞察后，我们就能知道不同的个体消费者有哪些独特的产品需求，从而有的放矢地进行高精准的产品推荐，直击目标，百发百中。相比于之前模糊认知下的乱投放、放乱箭，全域洞察下的个性化推荐效果自然更加显著，效率也有所提高。例如，中国移动开发的智能推荐平台就可以通过消费者洞察，精准地实现对消费者的个性化推荐。该平台通过数据分析出消费者的需求和偏好，根据这些信息不断调整模型参数，提高了推荐的准确率。目前，中国移动的客户端以及阅读、视频等频道通过该平台的赋能，真正实现了针对不同消费者千人千面的个性化推荐。

全域营销

全域营销指的是利用数据和数字化工具，进行以消费者运营为核心的智能营销。传统的营销方式推荐的可能不是消费者真正需要的，或者由于推荐流程过于冗长，等到消费者真正了解产品时，他已经不需要了。而全域营销以消费者的需求为核心，它的目的就是提升消费者在全渠道的消费体验，从而达到消费者与品牌关系全链路的精细化运营，为商家的营销和品牌建设赋能。全域营销分为三个方面：全域策略、全域传播和全域运营。

全域策略，是在策略层面做到洞察人群的价值，发现相应机会点，以达到高效创新。这里主要挖掘消费者的自然属性数据和品牌多源数据，将这些数据融合，为企业构建消费者画像体系，以此制定相

应的人群策略。比如一个公司知道它的消费者大部分是年龄30岁到40岁的女性城市白领，那么根据这个属性，针对这类人群的相应特点和消费习惯，就可以更准确地进行相应的广告话术和用语包装等营销策略调整。

全域传播包括两个方面：在认知层面对消费者的曝光触达，在兴趣层面与消费者增进关系，因而它需要对消费者数据有更多认知，需要进阶到对数据的深入洞察分析。一方面，经由营销活动沉淀，完成对消费者画像的数据整合和分析；另一方面，通过消费者洞察实现全链路消费者行为及流转分析——因为从消费者第一次接触商品开始，到他实现购买或转向其他商品为止，这条链路上的很多关键节点甚至是转折，都是需要分析的。这些关键点可以是某个促销活动页面，因为这可以说明这种促销活动页面是能够引起这位消费者的关注和兴趣的；也可以是消费者浏览商品评论的行为，因为这说明这位消费者更为看重口碑；又或者是他的搜索行为，因为如果他搜索的基本都是名牌，那么我们就可以推知品牌的声誉可能对他更有吸引力。我们需要根据之前的数据，推导出什么样的人对什么样的营销活动更加敏感、更加关注，了解消费者的兴趣点所在，最终达到传播效果。

全域运营，包括侧重于忠诚层面的消费者运营、侧重于购买层面的店铺运营等，都可以通过相应数据进行增值服务和激活。相比于策略和传播，运营更偏向于消费者价值的最终落地。知道了在哪个传播节点、用怎样的话术能够触动消费者之后，我们要将这些内容进行整合落地。当我们知道网站消费者或门店消费者的人群画像后，我们在每一次的营销活动中都可以正确把握消费者的特点，从而增加营销运营活动的成功率。比如，都市白领喜欢通过新奇有趣的网文来认识新产品并产生购买冲动，且对品牌有着执着的追求，那么我们的运营就可以突出“大牌”“火热”“打卡”这些宣传文字，并且将发布宣传渠道侧重于网文传播，实现精准运营。

我们再对全域营销的这三点进行回溯：策略层面需要洞察人群的价值，创新机会点；传播领域要精准触达和曝光，精准增加和消费者之间的关系；运营领域，无论是店铺运营还是消费者运营，线上线下运营领域都要结合对消费者的认知，进行全域的精准营销优化。当然其基础还是以消费者为中心，打通整个营销数据。下面我们将策略、传播和运营三点结合起来，举两个成功运营的案例，以便读者全面认知。一是商场的运营，以华润五彩城的获客率提升为例；二是房地产的运作，以金茂府的销售量猛涨为例。

华润五彩城是北京地区集时尚购物、生活配套、餐饮、娱乐、文化和运动为一体的多功能、多主题大型区域性商场。华润五彩城通过对商场消费者各维度的大数据进行分析，了解到消费者在各时间点的交互情况，并据此策划营销活动。比如在某一个时间点，华润五彩城的人群种类和分布都有哪些特征，抓住这些特征定点地进行相应的商家和产品的推荐。经过数据分析发现，华润五彩城的受众已经具有从过去的家庭客群转变为以年轻人消费为主的态势，并且90后、95后消费者占比都分别接近20%，而40岁以下的消费者占比也达到了70%左右。也就是说，年轻群体已经成为商城的消费主力。从这个认知出发，华润五彩城形成了从宣发到触达再到消费的全面数据闭环，并对商城的规划做出调整：将商城从餐饮儿童驱动型风格转变为时尚中产家庭综合型运作模式，增加轻奢、化妆品和时尚配饰等商铺的比重，并举办各类触达活动，例如联合天猫做“双11快闪店”、打造“积分ETC”功能、品牌活动、会员专属优惠日和会员沙龙这些新潮的活动形式。通过消费者洞察，华润五彩城对广告投放方式进行了改良，实现了广告的精准投放，通过其不断攀升的客流量和车流量，我们也能看出消费者洞察的巨大作用，其在2018年的销售额超过了18亿元，更是证明了这一事实。这也体现了在数据的支撑下，全域营销可以促进企业的发展，带来更大的效益。

对于房地产商来说，基于楼盘以及目标受众分布的商圈大数据分析，可以产生客群洞察和精准营销的决策方案，实现房地产行业消费

者数字化管理，其精准营销效果可显著提升。通过数字化赋能的手段，依托大数据对不同场景下消费者需求的强大理解能力，房地产行业可以提升多场景分析和决策的能力。

例如著名的房地产品牌金茂府就是智能营销的积极实践者。金茂府通过楼盘和目标消费者大数据分析提供精准营销方案。对于房地产行业来说，对人群的定位特别重要。无论是开发一个楼盘还是楼盘的销售和运营，都需要知道潜在消费者所在以及他们有哪些需求和爱好。打通整套数据之后可以对营销客群进行深度洞察和精准营销。这样一来，金茂府就能更懂得消费者的需求，比如在消费升级的浪潮下，金茂府通过大数据分析发现，住宅的消费群体除了一些基本的住房需求外，开始更加关注住房的观赏性，对小区的美化和绿化有很高的要求，所以金茂府十分注重外部自然环境的美化，绿化率达到30%。在实现了相关新品定位落地后，其营销渠道也开始通过数字化赋能打通数据体系，进行深度营销洞察，在洞察的基础上再进行精准分析，形成具有品牌特色的营销文案，精准投放到目标消费人群那里。在房地产市场经历了一段时间的消沉的大背景下，金茂府却运用数字化赋能不断扩大自己的市场，成功进驻了华北、华东、华南、华中和西南5大区域。由此看来，或许数字化赋能是房地产转型的一大关键。

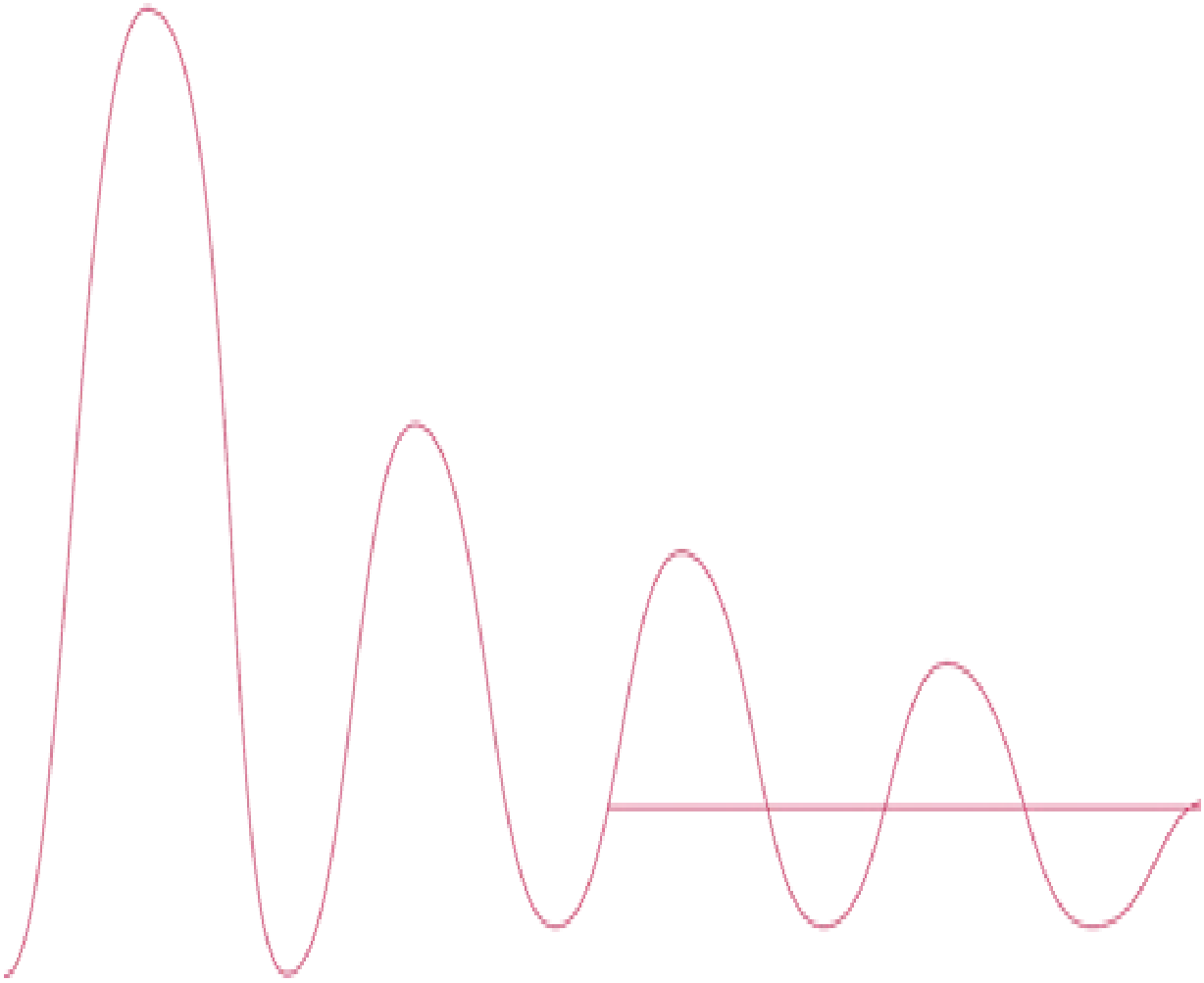
通过数字化的深度运用和赋能，不仅商场营销和房地产营销能够实现显著的收益提升，其他行业也是如此。经由数字化赋能，行业内可以实现对全域数据的认知和分析，从而洞察消费者的价值点，找到营销的创新点以及以往营销过程中存在的问题，最终优化产品定位和营销运营。通过上述案例我们可以知道，数字化赋能是对整个行业的赋能，它可以为行业带来许多积极的改变。

大数据时代下，数字化赋能是企业在竞争激烈的市场中站稳脚跟的必经之路。用不好体量庞大的消费者数据，它们就是一堆不产生任何意义的乱码；用得好，它们就是企业生根发芽的基石，是价值难以估量的企业资产。数字化赋能提供的就是为企业化数据为效益的智

慧，它将这一智慧嵌入产供销的每一个环节之中，为整个商业链路提速增效。它适用于各行各业，最终也将是全球商业未来发展的大趋势。数据是AI思维的基础，所以数字化赋能也为AI思维的落地夯实了基础。

第六章

企业级AI中台



虽然我一直和你说要追求卓越，要实现最后的1%，但是这并不意味着我们不可以接受部分改进。很多时候，一个完美的结果需要完成很多改进，而不会一步到位。

—《态度》—

知行合一，学以致用，本章还是从“用”的角度展开，阐述人工智能如何在实际的业务场景中落地，并为企业创造价值的。本章先从人工智能治理的困境入手，引出AI中台的重要性和必要性。通过对工业AI中台以及零售AI中台的案例分析，你会真正了解AI中台是如何发挥作用，为企业赋能的。

我们在前面讲过人工智能决策引擎运行的全过程，包括对历史数据的特征处理、机器学习、模型评估、模型生产以及模型监控。读完本章之后你会发现，正是因为有了AI中台的存在，人工智能决策引擎才得以在企业实现落地。大量生产制造和零售环节的案例既证明了AI中台可以解决人工智能落地过程中管理难、利用难、可靠性低的痛点，也表明了AI思维是一套从理论到实践一脉相承的完整方法论。

美好而混乱：AI治理的困境

回顾之前，企业将生产制造环节和零售环节的数据收集起来，在数据中台上进行打通，在具体的业务线层面会产生人工智能应用的需求点。无论是生产制造领域还是零售领域，都会逐步在单点上建立起相应的人工智能应用。但随着人工智能应用落地数量增加，人工智能模型就会出现零散而混乱的状况，企业在管理和利用这些人工智能模型时就会面临一些需要解决的痛点。

第一个痛点是管理难。从生产制造到零售，如果每一个业务环节都通过人工智能进行升级，可想而知，产生的人工智能模型数量是非常大的，而且会分布在不同的业务部门中，显得零散而混乱。那么如何管理这些不计其数的人工智能模型呢？现有的人工或者传统的信息管理手段似乎都不能有效地完成任务，所以对人工智能模型的管理是企业在智能化过程中形成的一个新的需求和领域。

第二个痛点是利用难。对很多企业来说，利用人工智能模型进行业务洞察和优化，存在相当大的壁垒和障碍。因为人工智能应用不像传统的应用软件，它涉及核心的业务决策，所以必须用专门的手段管理、监控和持续优化，若缺少人工智能专用的系统和平台，则很难发挥其价值。

第三个痛点是可靠性低。人工智能模型在管理不当的情况下非常容易出错，而且较之传统的数据分析更为复杂，出错的可能性相对大很多。单点式的人工智能模型散乱在生产制造和零售环节的各个业务节点中，无法确保输出结果的可靠性，从而无法驱动企业实现数字化智能决策，也就无法驱动相应业务的提升。

这些痛点都是围绕业务而生的。当人工智能逐渐深入到业务之中后，散落的人工智能应用会越来越多，这时候，不但数据可能凌乱地分散在各个平台上，形成数据孤岛，人工智能模型也会非常混乱，甚至管理人工智能模型本身也会成为业务痛点，因此必须通过一套标准化流程来应对和优化整个人工智能模型生产过程中所涉及的各个环节，构建一套相对标准化的、可延伸的，并在企业内可发挥价值的AI中台，来统筹管理所有散乱的人工智能模型和应用，并驱动其不断优化和迭代。

我在PayPal工作的时候，需要同时管理几百组不同的人工智能模型，这些模型分别对应了不同的场景预测，比如有一系列预测用户行为的模型，也有各种预测商家价值和风险的模型，而且每过一段时间，这些模型又会用新的数据刷新。在没有系统化使用AI中台之前，由于人工智能模型的混乱出现了很多问题，比如预测模型和应用场景没有最佳匹配，因此它不能恰如其分地给出满意的预测，最终在实际业务中不能够带来任何提升。发现是人工智能混乱引发的问题后，我带领团队开始构建AI中台，用平台化的模式管理这些模型，最终人工智能混乱的问题得以解决。对于软件怎么开发和管理，已经有很多成熟的方法论，而对于人工智能怎么治理，则可以通过平台化的方法

论，在数据中台上进一步构建AI中台，为如何把人工智能模型更有效地管理起来提供了一些方向性的思路。

AI中台就是为企业提供整合化和规模化智能服务的基础设施，它依附于数据中台之上，在企业里实现人工智能决策引擎。也就是说，AI中台是在产供销全流程数据基础上进行模型建构，从而实现精准决策，使价值落地的。例如，从零售角度来说，用户的相关数据会留存在搜索渠道、社交渠道、电商渠道、线下渠道和客服触点里，我们在这些打通的数据基础上，构建AI中台能力，服务更多应用，而不仅仅局限在单点上的人工智能模型。当模型的数量达到一定规模后，我们可以通过AI中台提升人工智能模型运作的效率。所以，当企业数据打通之后，构建AI中台能够进一步提升和优化业务目标，实现增量的价值。

因此，AI中台是在数据中台之上的延伸，如果说数据中台是对直接业务需求的响应，提供的是基础的服务，那么AI中台的作用就是实现业务的创造性发展。一方面，AI中台实现的前提是基于数据中台的全链路数据打通；另一方面，数据中台和AI中台相辅相成，可以提供个性化的服务，智能驱动业务发展，甚至能够创造新的业务模式。所以说，数据中台不是数据分析利用的终点，而是业务创新的起点。通过AI中台和数据中台的整合，企业可以快速、规模化地利用AI思维创造价值。

机器学习平台

机器学习（包括深度学习）是人工智能发挥作用的核心，因此AI中台的核心就是机器学习平台，其中包括一套成熟的机器学习流程运作，涵盖从数据出发进行特征处理、样本调节、模型训练、模型参数调优、模型验证和提供预测服务的各个步骤。

图6-1对机器学习平台做了梳理，机器学习平台的输入包括客户数据、行为数据、交易数据、机器数据、生产数据等，需要预测的目标也都会输入到平台上，经过以下各个流程的加工：

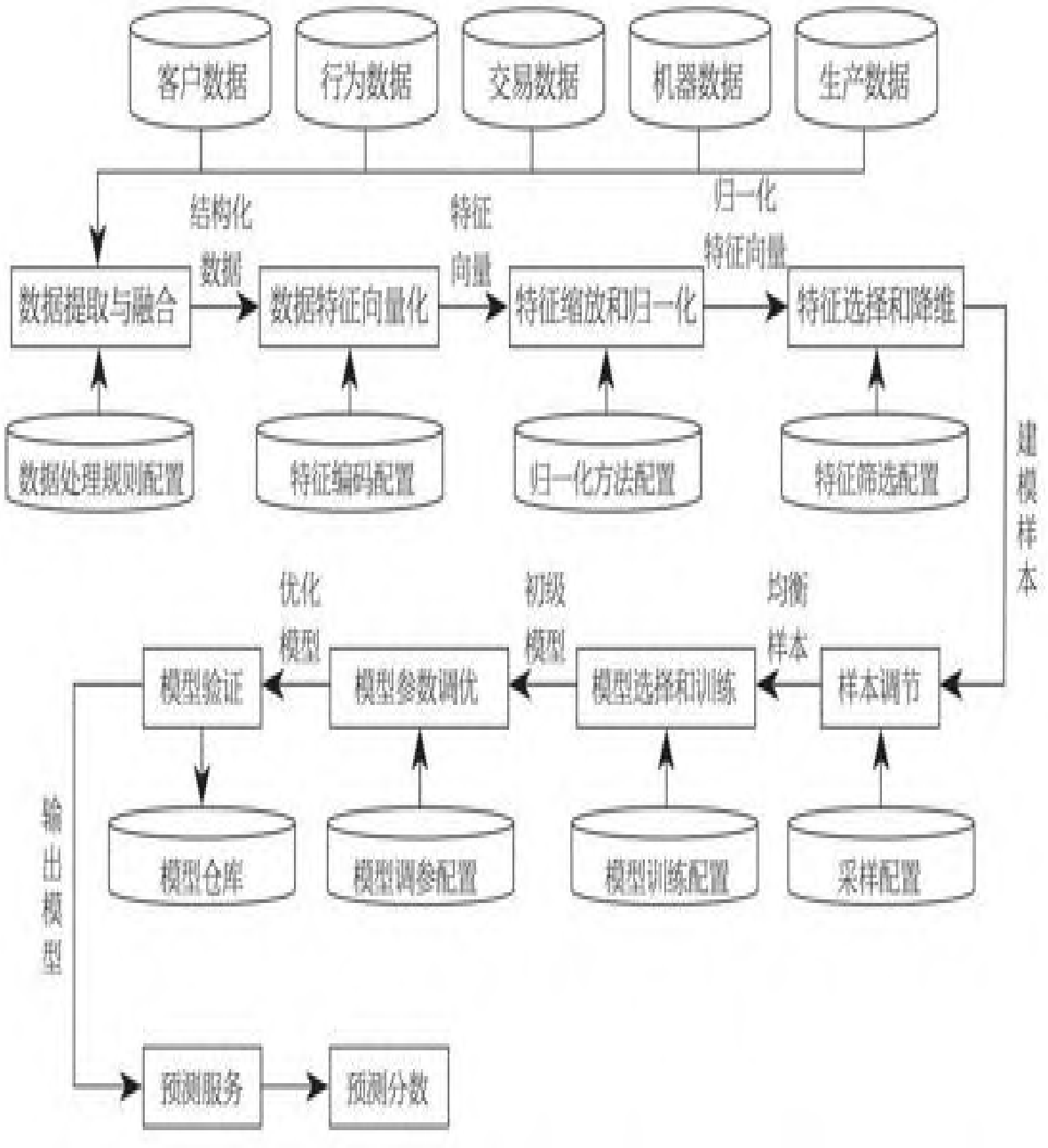


图6-1 机器学习平台示意图

第一步，通过对原始数据的提取与融合，形成结构化的、后续步骤可以操作的数据。

第二步，通过数据特征向量化，将结构化的数据转换为特征向量。

第三步，对特征向量进行缩放和归一化处理，这些对特征数值的调节能够优化后续人工智能模型的效果。

第四步，通过特征选择和降维，降低特征向量的维度，形成建模样本。维度过高会对后续模型训练产生影响。

第五步，通过各种统计采样方法对建模样本进行样本调节，生成均衡样本。均衡的样本有利于产生优质的模型。

第六步，根据均衡样本进行模型选择和训练，产生初级模型。

第七步，利用网格搜索、随机搜索等方式对初级模型进行模型参数调优，形成优化模型。

第八步，将优化模型进行模型验证，合格后输出模型至模型仓库。

第九步，利用输出模型进行预测服务，给出预测分数。预测通常是以打分形式体现，无论多么复杂的人工智能应用，都是通过复杂的流程，最终输出对事件、流程或个体的打分。这个打分既可以是实时打分也可以是批量打分，既可以是在线的，也可以是离线的。获得预测分数后，根据分数产生针对实际业务的决策。

在实践中，人们通过对中台上已经打通的大量生产制造和零售环节的数据进行复杂的处理，对大量的数据实现深入的人工智能分析，最终达到指导和优化决策的目的。无论是生产制造环节的预测还是零售环节的预测，通过上述平台都可以实现标准化的管理。有了这套体系后，开发新的人工智能应用的时候就不需要重新开发框架，而是利用这个平台在每一个环节上通过简单的配置和脚本进行定制。对于实

际应用来说，如果让每一位开发人员修改平台本身，就会容易出错，所以，只有把整套平台尽可能地固化下来才是最优的，在统一的环境下用一套完备的平台治理人工智能模型，降低它本身带来风险的可能性，这样也有利于企业人工智能体系的迭代和发展。综上所述，在人工智能已经有一定应用空间以后，企业适合建立一个以机器学习平台为核心的AI中台，这样可以确保人工智能模型之后的延伸和发展具有可靠性和易管理性。

从机器学习到应用定制

刚刚我们谈到了AI中台的核心是从数据到预测的机器学习平台，就像数据中台里打通的数据一样，机器学习和深度学习模型是AI中台重要的部分，如图6-2所示，下面我们来看一下各种不同的模型是如何架构和产出的。

各种通用数据在数据中台上汇总和融合，为AI中台的运作提供了通用数据样本。在AI中台中，当通用数据样本输入后，就可以在机器学习平台上进行建模，生成通用模型，例如图像视频、自然语言、商业决策和工业决策领域的通用模型，来解决通常所见的预测问题。具体而言，图像视频领域的通用模型包括物体识别、轨迹分析、人脸检测和人体识别等；自然语言领域的通用模型包括情感预测、多轮对话、文本分类和文本摘要等；商业决策领域的通用模型包括价值预测、转化预测、智能推荐和CRM优化等；工业决策领域的通用模型包括状态诊断、目标预测、寿命优化和能效优化等。因为机器学习和深度学习具有反馈和自学习的功能，中台上的模型训练用到的数据越多，形成的通用模型效果也就越好。但是，通用模型的建立主要是依据大量通用样本数据，所以通用模型本身对于具体场景的应用而言不是最佳的，没办法做到十分精准的千人千面。就像穿衣服，上学时候的校服或者工作时候的职业装都能够套在每个人身上，但对于个体来说一

定不是最好看、最合适或者最舒服的，每个人的体形、性格和喜好不同，应该有更精准的、定制化的穿衣解决方案。

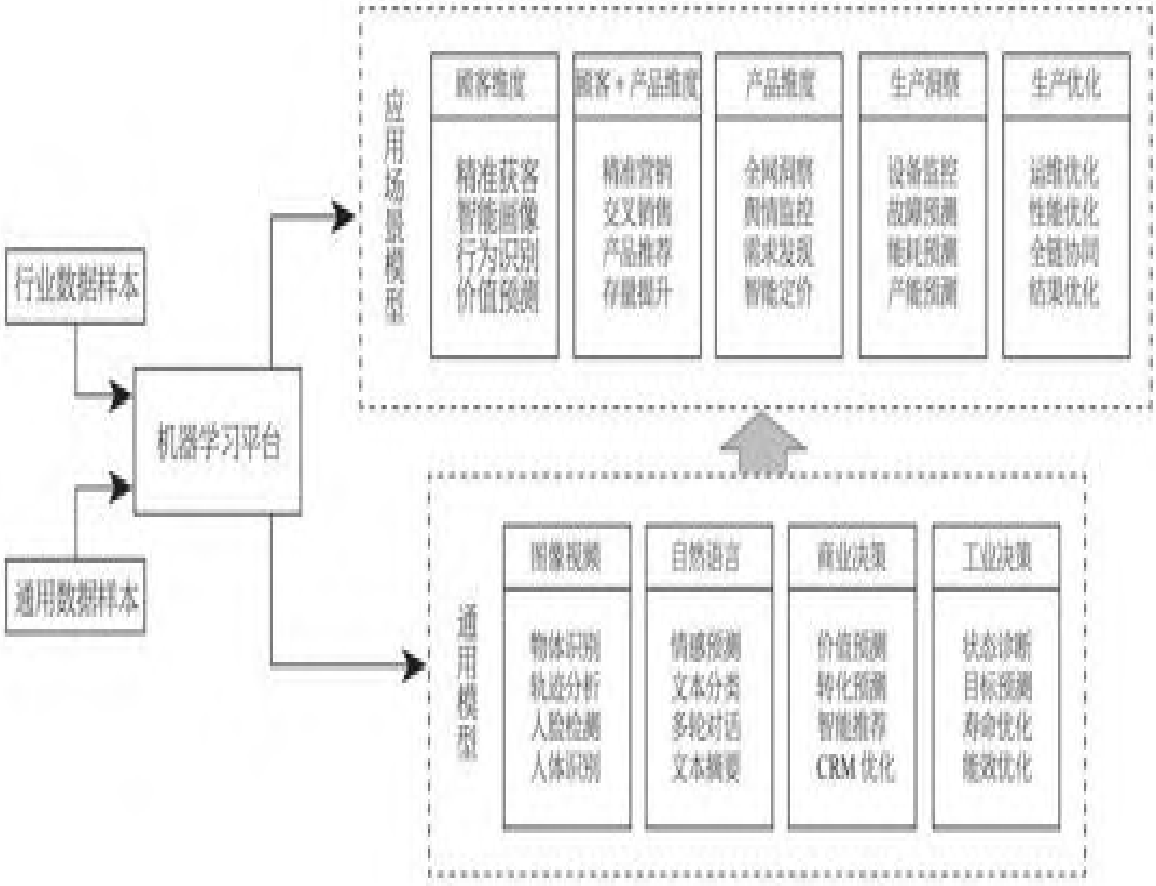


图6-2 定制化模型体系示意图

AI中台的最终目的是对企业的决策产生影响，因此仅仅局限于通用模型层面是不够的，机器学习平台还需要包括应用定制模块，产出应用场景模型，提升跨场景应用效果。虽然通用模型在一定程度上是可以跨场景的，但由于对行业和具体应用场景的认识不够深入，效果不够理想。例如，在物体识别领域，通用的物体识别模型是用大量数据进行训练而成的，但还是没法识别新的物体种类，这就涉及场景定制化，需要根据行业样本建立起应用场景定制模型。以箱包品牌为例，在此之前，虽然可能已经有了通用的箱包物体识别模型，但还是

没法识别特定品牌款式的箱包，因为每个品牌都有品牌特有的箱包设计，例如简单的公文包、优雅的手提包，可以是方的也可以是圆的，可以是金属的也可以是帆布的，没有相应的数据就没办法进行最佳的识别。在这种情况下，也需要通过场景定制开发，形成场景定制模型。人工智能的最佳实现路径并不是单独开发一个识别具体款式箱包的专有模型，而是要先有基础的物体识别能力，再把个体品牌的具体模型引入。比如在通用模型的基础上，通过某个箱包品牌提供的箱包样本，进行深入迭代，就可以识别个性化的品牌了。

因此，通过行业数据的输入进行迭代以后，通用模型进行进一步学习和演化，能够解决商业场景的具体问题，也就形成了应用场景模型，具有了相应的商用价值。应用场景模型有很多，例如，在顾客维度包括精准获客、智能画像、行为识别和价值预测等；在产品维度包括全网洞察、舆情监控、需求发现和智能定价等；在顾客加产品维度包括精准营销、交叉销售、产品推荐和存量提升等；在生产洞察维度包括设备监控、故障预测、能耗预测和产能预测等；在生产优化维度包括运维优化、性能优化、全链协同和结果优化等。这些行业场景下应用的具体问题都可以通过应用场景模型进行恰当的决策。

从通用模型到应用场景模型是一个迭代递进的关系，一般不可能仅通过少量训练样本就产生这样的场景化能力，而是需要很多通用层面模型能力的沉淀，然后才能在具体的业务场景下定制开发出特定的模型解决方案。在生产制造和零售领域，随着模型学习的发展，一般会沉淀出效果良好的基础模型，它未必适应每一个具体的业务场景，但通过整合具体场景的数据进行学习，可以定制化生成最适合场景的一系列模型。

总结一下，AI中台的核心就是机器学习平台，既能够产生通用的模型组合，也能够产生具体场景的行业应用模型。具体来说，就是通过大量数据的沉淀和打通，形成通用模型，再通过行业样本训练，形

成智能程度更高、为场景定制化的模型，模型运算的数据再回到平台里不断优化和迭代，最终提升平台的整体效果，满足业务需求。

搭建AI中台

经常有企业想通过调取腾讯或者百度人工智能接口的能力来搭建自己的AI中台，这样可行吗？我的答案是可行，但由于百度和腾讯并不清楚和理解企业所需要的具体应用场景，所以这里得到的是通用模型。这些模型可以作为最终模型的出发点，产出初步效果，然后必须要结合场景深度优化，在通用模型基础上融合行业样本，迭代出新的模型，最终形成生产制造或零售领域里一系列定制的专属模型，帮助企业实现商业化落地。很多企业的人工智能创新在这一步停滞不前，因为它们倾向于调用腾讯或百度之类的人工智能接口，而这些接口不是定制优化的，所以运用起来就会发现很多问题不能得到妥善处理。

在实际应用中我们也发现，通用的人工智能自动问答系统在银行应用中只能做到60%左右的准确率，然而通过定制化模型，我们能达到接近90%的准确率。为什么能提升30%左右的效果？正是因为我们对其进行了定制化处理，所以效果远优于通用模型。如果没有定制专属的模型，企业就只能调用外部人工智能接口，而外部人工智能接口只能提供一些基础功能，很难满足行业的个性化运营需求。

感知类的能力平台也是如此。在语音识别领域中，通用接口的语音识别能力一般只限于普通话，在麦克风离说话者口部很近的情况下，能达到95%以上的语音识别效果；但只要说话者说的是普通话以外的语种，例如四川话，语音识别的效果就会大幅下降，因为通用接口一般无法识别普通话以外的定制化内容。那么，要识别四川话该怎么办？那就得拿大量四川话样本对其进行模型训练。这就需要很多定制化的开发，而不是仅仅使用通用平台能力。

从以上可以看到，一方面，AI中台能够对散乱的单点人工智能模型进行整合和管理，保证人工智能模型有条不紊地运作；另一方面，在AI中台的定制化模型体系中，各个环节层层相连，在通用模型基础上，迭代场景专属模型，最终经由AI中台、机器学习和深度学习赋能，优化预测效果，实现应用场景下的最佳决策。前者能够保证人工智能模型正常运作，后者则能进一步让人工智能模型在场景中落地，这些都说明了建立AI中台的必要性和AI中台能够产出的实际效果。无论是生产制造环节还是零售环节，在产供销链路的每个节点里，AI中台都能嵌入其中并实现深度赋能，充分发挥人工智能模型的价值，优化决策和效益。

AI中台在企业中实现了人工智能决策引擎，以打通的数据为养料，帮助业务前台做出决策。AI思维存在的意义是指导我们更好地工作和生活，AI中台就是AI思维转化为实践的实现载体。AI中台的基础框架部分上面已经进行了讲解，所以我们现在应该解决的问题是在实际的应用过程中AI中台是如何发挥作用、如何对各项业务进行深度赋能以及如何帮助企业实现效益的提升的。后面我们会通过对不同领域的具体案例进行剖析，找到这些问题的答案，这样一来，我们一方面能够清楚地了解AI中台的操作流程，另一方面也能更加了解在生产制造和零售环节中AI中台的运用现状。

以生产目标为核心，提升智能制造水平

人工智能领域权威学者吴恩达在2018年工业互联网峰会上说过：“人工智能是新电能，一百年来，电能为很多企业带来了大量的改变，今天人工智能也会为这些企业带来一样的改变。”更有人说，是人工智能带来了第四次工业革命。“革命”就意味着颠覆，AI中台就

是人工智能颠覆时代的发电机。下面就让我们一起来了解，AI中台是如何在历史的洪流中推动工业向前发展的。

在工业领域，生产制造是最重要的环节。生产制造的目标是提质、增效、降本和减存，这也是AI中台优化工业生产环节的目标。首先我们需要知道，AI中台要发挥作用，通常需要先收集生产制造过程中各种各样的数据。企业的数据仓库里，存储了大量数据，包括原料、采购、生产、渠道和销售等各环节的数据，然后通过数据中台将数据打通。AI中台要做的就是对这些打通的数据进行学习，生成定制化模型，对工艺、损耗和良品率等进行监控和诊断，得出生产制造环节相应的预测。最后，依据输出的决策进行反馈，在优化质量、降低成本、增加效能和减少库存等方面做出优化，实现生产制造的智能化转型。接下来，就让我们来详细了解AI中台在生产制造环节是如何具体应用的。

工业AI中台总体架构

工业AI中台就是AI中台在生产制造环节的具体应用，其总体架构如图6-3所示，包括数据采集、数据筛选、信号过滤、特征提取、模型学习和结果反馈六大模块，具体流程包括以下步骤：

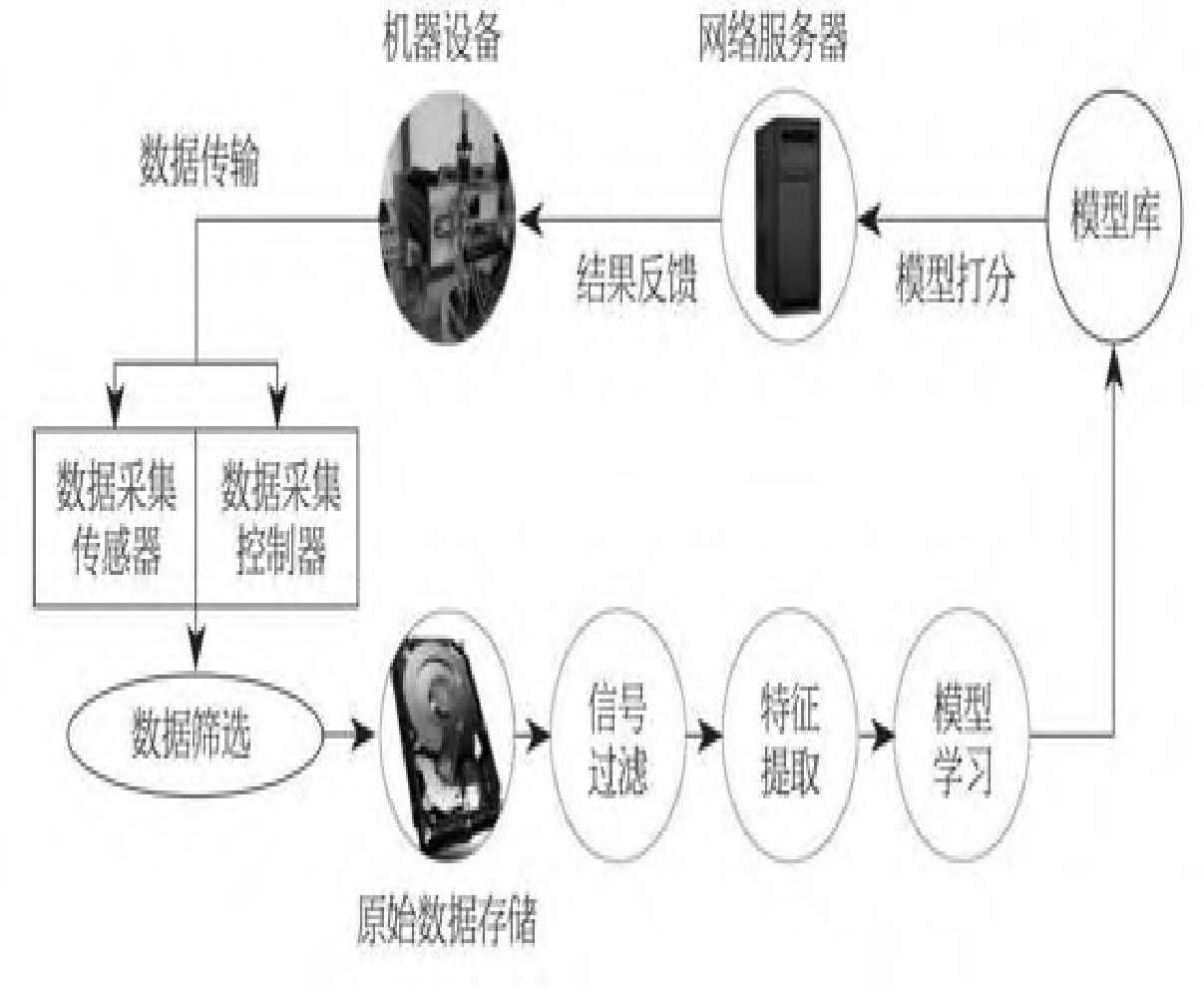


图6-3 工业AI中台示意图

第一步，数据采集：通过数据采集控制器和数据采集传感器从机器设备上取得相应数据。

第二步，数据筛选：因为生产制造环节产生的数据非常多，所以需要对原始数据进行一定的筛选。

第三步，信号过滤：对数据进行信号过滤，一般包括数据清洗、信号同步和信号分割等操作。

第四步，特征提取：对过滤完毕的数据进行特征提取，例如对机器特征、原料特征以及时间和频率域特征等的提取。

第五步，模型学习：将提取的特征数据通过机器学习实现性能预测、能耗预测和故障预测等功能。

第六步，结果反馈：最终由模型输出的控制参数反馈到机器设备上，从而操控机器设备，实现智能优化。

这是生产制造环节中AI中台的典型架构。传统上，生产制造的参数仅凭操作工人的经验决定，有时甚至是工人一时兴起拍脑袋决定的，控制水平很大程度上取决于人的主观判断和操作水平，非常不稳定。有时候一不小心操作过头或者是疏忽大意了，都会影响最终的生产效果。而人工智能不仅可以解决这些问题，满足生产制造的需求，还能够降低成本，提升效益。在生产制造过程中，可以采集到大量数据，收集可观测的变量。能够观测的变量可能会达到成百上千个，通过感应和采集这些变量，在工业AI中台之上进行加工和分析。然后，要明确优化目标和优化手段，比如目标是提升生产效率，还是降低原料消耗，常用的优化手段是调节控制参数。可以说，工业AI中台在生产制造领域的核心应用之一就是调节控制参数，并且反馈到相应设备上，从而实现既定的优化目标。

例如，在生产过程中，产生的气体温度和浓度都会有一个理想范围，可以通过人工智能进行预测和优化控制。也就是说，人工智能模型可以根据每个时间点的状态，指导具体的控制决策。再例如，国家对工厂有环保和废气排放方面的限制要求，工厂利用石灰对废气进行脱硫，可以通过AI中台精准控制某一个时间点该放多少石灰原料，从而优化运营效率，降低原料消耗。

工业AI中台应用场景

在工业领域运用人工智能的先进性和重要性已为大多数企业首肯，从应用场景维度分析，降低成本、提高效率的需求是推动目前工业企业运用人工智能的最大动力。工业AI中台在工厂中发挥作用的前

提是将工厂内部的机器设备连接在一起。然后，人工智能就可以分析联网上传的所有数据，优化生产制造的各个环节。目前人工智能与生产制造的融合场景主要可分为以下几类：生产优化、智能质检、生产设备的预测性维护以及供应链管理。

生产优化是指，基于生产线的大量数据，人工智能可以优化生产工艺，提升产品品质。具体来说，就是将生产制造端的各类数据打通，并在工业AI中台上进行分析，形成资源最优利用的方案组合，提升产品合格率和产品生产效率等。另外，在很多环境中，工厂都希望利用人工智能来降低成本，提高生产效率，从而提升企业竞争力。人工智能通过数据筛选和学习，甚至可以让机器成为自动优化的系统。例如，在钢铁厂中，通过AI中台对当前生产数据和历史数据进行连续分析和学习，实时调整生产制造系统的参数。目前，一些钢铁厂已经通过人工智能实现熔炉的自动优化设置，通过自动分析生产材料构成，识别稳定工艺条件的最低温度，从而降低整体的能耗。

智能质检是指通过人工智能测定产品质量并判断产品是否合格，能够有效地为产品质量把关。智能质检不但要对产品本身的质量进行检测，还要对生产制造环境的质量进行监测，而这一切都要通过人工智能进行。人工智能视觉系统能够利用图像识别发现缺陷以及产品功能的偏差，并且系统本身可以根据大量数据持续学习，其性能会随着时间推移而逐步改善。例如欧洲科研机构研发的一款3D机器视觉机器人就可以对工业生产制造的全过程进行智能质检。该机器人具有仿真人眼功能，可以根据多张2D图片生成3D图像，这样就可以准确检测空间上的质量问题，例如体积以及磨损问题。而且，该机器人的反应速度可达每秒传输4 000帧，完全实现在线检测，大大提高了检测的效率。

预测性维护是指在工业生产制造过程中，基于过程数据，通过人工智能预知可能的失效模式而进行的维护活动。预测性维护是人工智能应用于工业生产制造领域的重要切入点，它可以预防机器故障、加

强工作场所安全，提升产品质量。德勤公司曾指出：不合理的维护策略会导致工厂产能降低5%~20%，工业企业每年由于意外停机造成的损失高达500亿美元。而基于人工智能的预测性维护，可以在设备出现问题之前就分析得到出现问题的概率，比如，工厂中的数控机床运行一段时间后就需要更换刀具，工业AI中台通过持续学习，分析工厂内的机器和部件产生的数据，可以预知刀具损坏的时间，从而提前准备好更换的配件，并安排在最近的一次维护时更换刀具，这样可以避免出现故障。

GE（通用电气）董事长兼CEO（首席执行官）杰夫·伊梅尔特曾说过：“工业互联网是变革的最大机遇，超越了消费级互联网，是连接人、数据和机器的开放、全球性网络。”GE的工业互联网平台Predix早在2015年就每天监控和分析大约1 000万个传感器发回的5 000万条左右的数据，监测和展示所有设备的实时运行状态、健康状态和运行绩效，并根据这些数据进行人工智能建模分析，预测设备故障并执行故障诊断，采取主动措施规避故障风险。具体来说，首先收集设备基础数据、物理模型和其他相关信息，然后建立具有故障检测功能的人工智能模型，对重要子系统及零部件数据进行分析，并把它部署到实际生产制造系统中，实现人工智能的落实。

供应链管理方面包含了许多不同的工作，指对整个供应链系统进行计划、协调、操作、控制和优化的过程。人工智能可以帮助企业管理和协调供应链涉及的各个流程，并将这些流程与客户流程和供应商流程综合起来，根据各自不同的需求，整合做出最佳的决策。企业通过AI中台管理供应链，可以更加智能地响应生产制造中需求的变化情况，因此，供应链也是AI中台能够赋能提升的重要环节。人工智能在处理由供应链产生的海量数据方面的能力已然超越了人类。一些企业利用机器学习将仓库、企业资源计划（ERP）系统与客户洞察的数据连通起来，然后应用AI中台的具体模型来处理和分析数据、生成预测，从而改进供应链，实现了从“开始”到“结束”的完整供应链的智能化。另外，AI中台的应用能够提高原材料供应的转移效率，这对于管

理制造多种定制产品的复杂过程来说是必不可少的。例如，AI中台会利用物流数据，比如材料进出的数据、库存量和零件的周转率等，来促进仓库自主优化运营，例如建议将低需求的零件转移到更远的地方，将高需求的零件放到可以更快获取的附近区域。同时，在工厂和仓库内运输物品的无人车，在AI中台的引导下，会结合人工智能的视觉感知能力，实时调整车辆路线，从而实现最优的运营效率。

工业AI中台应用案例

工业AI中台能够应用于工业生产制造的方方面面，随着人工智能的不断发展以及在生产制造环节的不断深入，在工业产业中应用AI中台的场景和进展也会愈加复杂。上面我们已经谈到，AI中台在工业中的适用场景包括生产优化、智能质检、生产设备的预测性维护以及供应链管理。提升生产效益是企业的核心目标，保障产品质量是企业的立足之本。下面我们来看看AI中台在生产优化和智能质检这两个应用场景里面的具体运用案例。

河南鑫磊集团是一家综合性煤炭企业，其在生产中也同样面临企业运营成本高、生产效率低的问题。在煤炭行业中，把不同种类原料煤按适当的比例配合起来的配煤工序占到了生产成本的80%以上，如果仅靠人工计算确定原料配方，不仅可用参数有限，而且由于人工操控难以保证精准，所以容易造成不必要的资源浪费，成品质量也很难形成统一的标准。鑫磊集团向人工智能寻求解决方案。在实际生产过程中，通过采集和输入各种原料煤库存、计划配煤量、炉温和结焦时间等数据，鑫磊集团借助人工智能，优化了各种原料煤的配比，找到了最优配置，降低了原材料的无谓损耗，每年至少节省成本约3 000万元。

国内最大的轮胎制造企业中策橡胶也通过人工智能对橡胶的生产过程进行了优化。中策橡胶在产品研发、生产以及质量检验等环节都

积累了大量的数据，将数据打通之后，人工智能模型可以根据打通后的数据优化生产过程。例如，橡胶为企业的主要生产原料，不同的产地、批次等因素都会对橡胶的质量产生影响。人工智能模型可以迅速分析这些影响因素，得出将某几个产地的橡胶组合在一起会使生产质量达到最优，形成资源最优利用的方案。通过人工智能的优化，中策橡胶在降低生产成本的同时，还显著地提高了产品质量，产品合格率最高提升了5%。

国际上也有许多运用人工智能优化制造的成功案例。例如，为全球用户提供存储器产品的知名硬盘厂商西部数据公司（Western Digital Corp）就利用人工智能优化测试设备节省了资金支出。对于硬盘制造行业来说，最大的开支之一就是测试设备。因为硬盘制造的任何一道工艺出错都可能导致生产的集成电路不合格，拉低良品率，导致不必要的经济损失，所以测试环节贯穿在集成电路生产制造的整个流程中。在基于卷积神经网络的人工智能模型的帮助下，硬盘制造的测试流程得以优化，也因此节约了数亿美元，这对市值在200亿美元左右的西部数据公司来说，是相当大的一笔资金。

作为深圳地区LCD（液晶显示器）工业领域的标杆性企业，华星光电在智能质检方面取得了显著的成果。为了满足用户的需求，华星光电需要在较短的时间内交付出高质量的产品，但是在运营过程中发现，耗时的质量检验阻碍了生产制造速度的提升。质量检验是整个制造流程中极为重要的部分，如果不能在交付前发现瑕疵，就会导致产品退货和返工，为此需要付出高昂代价，也会对产品声誉造成损害。于是，华星光电将人工智能引入工厂，打造了生产线图片库。这个图片库里的图片分为两类，一类为合格产品的图片，另外一类是存在各种不同缺陷的产品的图片。然后，华星光电根据这些图片开始训练人工智能模型。在实际工厂中，超高清相机在检验点拍摄产品图像，人工智能模型将实时拍摄的图像与数据库中的缺陷图像进行快速比对和分类，并发送至云端供检验人员检查评估。华星光电利用智能视觉检测加速目视检查流程，通过将产品图像与缺陷图像库进行比对，实现

了智能、更快速、更准确的产品瑕疵检测。经过人工智能的优化，检查流程能够实现在数毫秒内完成单个产品分析和检测，缩短了检验交付周期，在削减了生产制造成本的同时，提高了运营的赢利性。

工业AI中台“三步走”策略

从上述案例可以知晓，要让人工智能在企业实际运用中落地，各工业企业首先要找到着手点，从人工智能应用到AI中台还需要一个过程，可以采取“三步走”的策略。

第一步，评估现在，寻找着手点。首先要对企业各项业务的现状进行评估，从自身的业务痛点和人工智能成熟度出发，对自身工厂的各方面运转情况进行健康检查和审视，将自身企业的现状与行业平均水平进行比较，找到自身需要通过人工智能改造以有所提升的业务领域，例如资源消耗是不是合理、运转速度够不够快、操作是否稳定准确等。

第二步，引入人工智能单点式用例。根据各项业务的实际情况，确定优先引入人工智能的单点式用例。企业的利益相关者可以将人工智能用例清单与企业各项业务存在的痛点进行比对，深入研讨优先通过人工智能解决哪项痛点，并制定相符的方案，这部分需要考虑到业务的复杂度、基础条件的完备程度、数据的打通情况及引入人工智能所要付出的相关成本，再综合确定人工智能的优先用例。

第三步，人工智能模型测试与拓展。对人工智能模型进行测试与解决方案拓展，选择工厂内的特定场景测试人工智能用例，然后根据输出效果迅速地对人工智能模型进行多次迭代和改进试点方案设计，单个业务线实现优化后，为了发挥人工智能的全部潜能，应该在公司全范围实施人工智能，实现生产制造环节全方位的优化提升。将人工智能融合进各项业务领域，这时工业AI中台才初见其形，然后通过对AI中台的不断优化和迭代，最终形成统一支持和优化业务的AI中台。

历史的车轮滚滚向前，人工智能的浪潮以迅雷不及掩耳之势席卷全球，将工业从3.0时代带入了4.0时代。从宏观上看，从现有的传统生产制造模式到人工智能模式的转变，也许将是一个完全颠覆性的改变——因为从生产制造到供应链，再到销售和售后体系，人工智能可以实现全链路的覆盖和提升。从微观上看，对于企业来说，AI中台之下的智能生产制造可以更好地优化生产、减少浪费和提高效率。从数据角度看，人工智能可以为工业的远大前景赋能，波士顿咨询就曾在《工业4.0——未来生产力与制造业发展前景》的报告中预测，人工智能将为中国工业的生产效率带来15%~25%的提升，额外创造附加值4万亿~6万亿元人民币。人工智能对工业的改造将是不可阻挡的一个大趋势，在新技术层出不穷的当代社会，人工智能的强大动力有增无减。对于工业来说，拥抱人工智能成为最理所当然的选择。

以消费者为核心，赋能智慧新零售

上一节说的人工智能赋能生产制造是以生产目标为核心，而这一节里要说的赋能新零售则是以消费者为核心。赋能新零售之所以以消费者为核心，是因为所有的零售活动都是围绕着消费者进行的，人是消费关系的核心，零售中的数据也都是和消费者以及消费行为有关的数据，消费者是人工智能赋能零售业的重要入口。

随着电子商务的兴起，零售行业发生了天翻地覆的变化，从传统的门店零售转向电商网站零售，消费者足不出户就可以买到心仪的产品，人工智能的出现又推动零售业向前进了一步，智慧零售应运而生。智慧零售的要义就是改变传统的零售方式，根据消费数据预测消费者的需求，做到个性化推荐和销售，甚至通过大数据和人工智能预测消费者行为，从而影响产品研发和生产。

人工智能赋能新零售的基本思想是通过数据指导零售行为，这里的数据就是企业所打通的消费数据。企业一般会运用一系列业务系统管理消费行为，如ERP、CRM、DMP（数据管理平台）等，这些系统几乎涵盖了产品从生产到被消费过程中的所有数据，例如业务数据、客户数据、交易数据、竞品数据、外部数据和消费者全链路行为数据等。然后，数据中台通过数据清洗和加工将它们融合打通。

进而，由AI中台提供的图像识别、语义分析和智能决策等功能，可以实现以消费者为核心的智能应用，其中包括根据多方数据进行消费者人群画像、根据消费记录进行消费者价值预测、根据消费偏好进行个性化推荐、根据忠诚度进行忠诚奖励、根据店铺位置和所处社区进行智能选品以及根据消费者需求进行销售预测等。这些智能应用最终会通过各类业务终端，比如App、各大电商平台、社交应用和自助终端等，与消费者产生关联。所以，从零售各业务环节来看，人工智能能够对消费者、门店、导购、运营、市场、生产、品牌和设计等相关方的行为产生影响。

总结一下，人工赋能新零售是以消费者为核心，也就是说在零售环节中，AI中台要通过海量数据识别和分析消费者各种各样的动机和行为，针对消费者进行个性化服务，以达到增加销售的目的。所以人工智能赋能可以更好地理解消费者、服务消费者和赋能消费者，让消费者在最短时间内，通过最佳手段获取最高契合度的产品信息，从而增加零售的转化效率。

智能消费者管理全流程架构

在零售中要做到对消费者的个性化推荐，首先要做的就是了解消费者，即分清哪些是目标消费者以及目标消费者的需求等，这可以在智能客户管理体系中得以实现。智能客户管理体系的核心是零售AI中台，零售AI中台通过运用人工智能模型，对消费者数据进行机器学习

或深度学习，从而预测出消费者的购买意向、购买力、触达方式和风险等，目的是锁定目标消费者，为创造个性化体验做好准备。

首先需要大量的消费者数据，例如电商、社区、客服、媒体和线下等各渠道获取的数据。具体来说数据的获取可以分为线上、线下两个方向。线上的触点有App端、微信端、电话端、短信端、广告端和媒体端等，线下的触点包括商超的售卖、户外广告和运营活动等。这些触点都会产生相应的数据，比如消费者属性数据、媒体数据、行为数据、点击数据和购买数据等，线上数据和线下数据相互补充形成数据闭环。然后需要对这些数据清洗、打通和融合，再进行初步的数据处理。这时，我们就可以在零售AI中台上进行建模和预测。

如图6-4所示，零售AI中台通过人工智能模型预测内容主要分为几个部分：

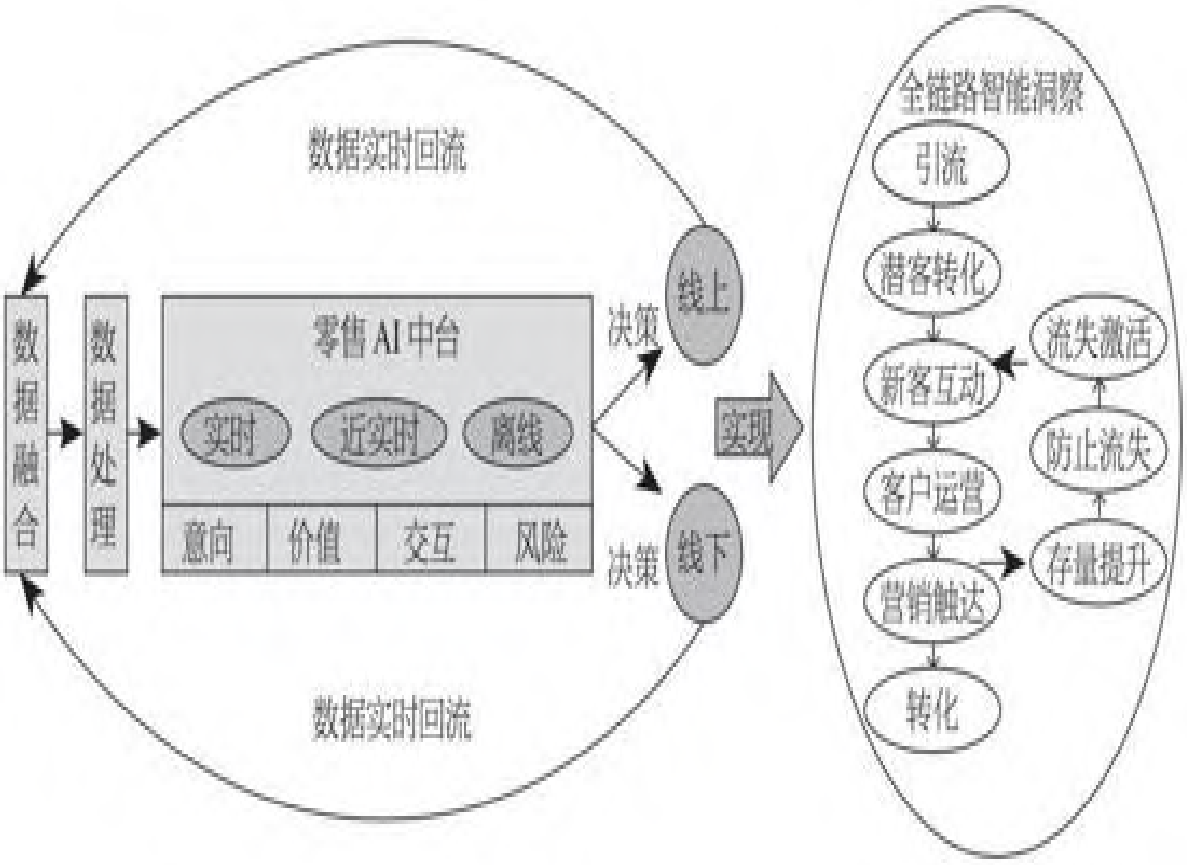


图6-4 以零售AI中台为核心的智能客户管理体系示意图

第一，意向预测，即预测消费者购买意向，预测消费者是否有购买的意向和对其意向商品进行排序。例如一个消费者花费很长时间查看同一类商品的信息，并将其中的几件收藏以及加入购物车，说明这个消费者有购买这类商品的意向。

第二，价值预测，即预测消费者购买力，判断消费者是高价值还是低价值客户，对消费者的购买力进行预测分析。比如通过消费者的购买记录，对消费者的购买力进行分类，如果一个消费者经常购买奢侈品，则其可以被分类为高消费人群。价值预测不一定基于购买数据，通过消费者在线上、线下的其他行为，也可以推知其价值。

第三，交互预测，即评估触达方式和手段，通过分析消费者行为习惯来选择最优触达方式，比如电话、广告或者短信。例如，有的消费者非常排斥电话营销，根据相关数据预测这类消费者可能会拒绝接听，那么通过电话与这群消费者接触的方式就不太合适，可以选择线上广告等其他渠道触达。每个人有自己的交互方式，我们要理解并满足他们的偏好，实现有效的触达，促成购买，这也是对交互方式的一种决策优化。

第四，风险预测，即预测消费者的潜在风险概率、欺诈可能性或潜在损失金额。例如，阿里巴巴开发的“八卦盾”工具，依托于人工智能实时检测淘宝平台上的黄牛扫货以及恶意退款、恶意攻击、敲诈勒索等恶意订单行为。如果检测到上述行为，商家可以向平台申诉，系统就可以主动拦截订单并进行预警，降低了交易的风险，营造了一个公平可信的电商交易环境。

零售AI中台通过数据对消费者进行全面的预测和分析，目的是找到潜在的客户并确定相应的营销策略。在营销过程中又将产生数据，并回流到数据中台和AI中台。通过零售AI中台最终可以实现全链路智能洞察。无论身处哪个行业之中，只要面向的是消费者，其生命周期

管理都具有比较标准的模式，无外乎是潜客转化、新客互动、客户运营、存量提升、流失预防和流失激活等过程。在整个生命周期里，零售AI中台需要在每一个环节都对客户的意向、价值、交互方式和相应风险进行预测，从而优化业务决策。在这个过程中，有大量的人工智能模型在发挥作用。零售AI中台将这些人工智能模型都汇总在一起，实现真正的线上、线下打通，赋能零售运作的各个环节，从而增加消费者的购买概率，增加消费者和商家的黏性，最终提高消费者的贡献值。如果每一位消费者的贡献值都增加了，企业总体的收益一定也会增加。

现在要开发一个新的客户，成本是非常高的，一般要付出几百元甚至更多的成本才能让一个金融或者教育领域的潜在客户成为企业的付费客户，如果付出了高昂的成本最终却没有实现价值，会给公司带来很大亏损。例如，企业想要增加新客户，如果考虑是从搜索引擎或者朋友圈引入，首先应当预测客户对产品的购买意向和潜在价值有多大，然后再去付出相应成本的投放。另外，因为对客户进行维护管理也需要成本，对于原先已有的客户，即存量客户，也要关心他们对企业的价值有多大。如果一个存量客户的预测价值较高，则需要通过各种方式经常沟通，给予他们相应的权益，进而维持和提升客户的价值。如果一个客户在某个环节上已经失去活性，不采取相应措施的话，客户就会流失，这时就需要通过营销手段对客户进行激活，防止客户真正流失。

个性化体验平台

个性化体验平台支撑了给消费者推荐个性化内容的任务。当消费者数据在零售AI中台进行了多角度的预测后，消费者就拥有了很多数字化的标签维度，再通过个性化体验平台，就可以做到真正意义上

“千人千面”式的个性化营销和推荐，给消费者带来最精准、最恰当的消费体验。

如图6-5所示，根据智能个性化推荐数据准备的情况，个性化体验平台的推荐方式分为以下三种模式。

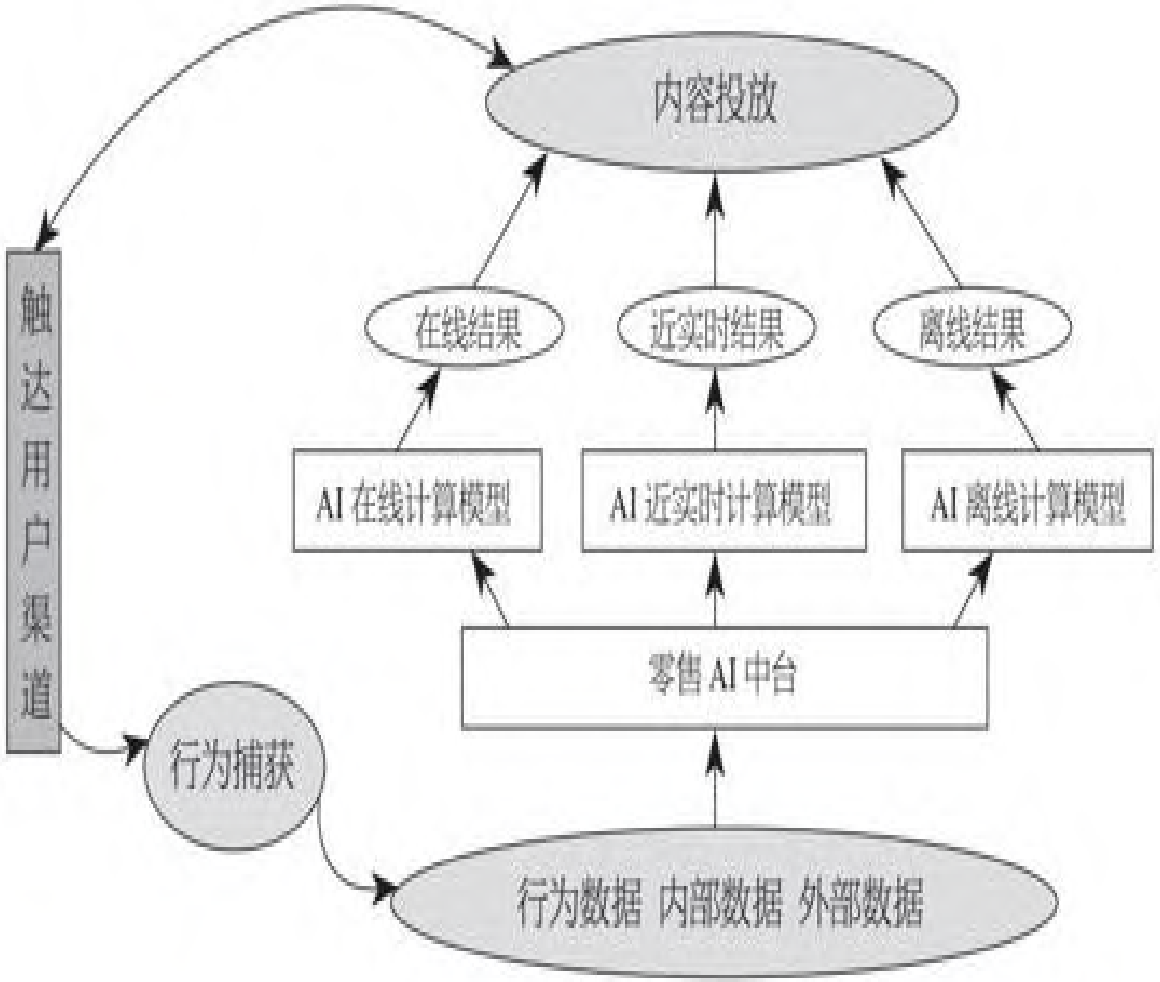


图6-5 智能个性化推荐数据准备的三种模式示意图

第一种推荐模式是实时的个性化投放，即基于人工智能的在线计算模型，对数据进行在线学习，实时得出结果。其中的含义包括两个角度：一是利用了实时的信息，二是实时计算反馈离线和近实时数据

库中没有的结果。实时的个性化投放一般是针对线上推荐而言的，通过互联网，信息才能被快速地收集和学习，并迅速做出预测和反馈。不同用户访问网站时，需要看到个性化的、与自己需求相匹配的产品或广告内容。这类场景需要实时的个性化推荐，用户一点击进入网站就要按照他的需求生成新的内容，这样才会吸引用户的注意力，对用户的决定产生影响。其中的延时不能超过百毫秒级，否则就会让用户觉得太慢了，难以吸引用户的注意力。例如，在一个外卖订餐App上向用户推荐餐厅必须是实时的，因为针对用户显示附近的菜品和餐馆，对用户来说才是有意义的，不然送餐时间过长，用户一般不会选择。因此，当用户打开App或者输入一个地址时，就要求App能够立刻做出反馈，呈现实时个性化的结果。首先，需要根据定位信息，发现用户现位于哪个商圈，该商圈里的菜品和餐馆种类是已知的，这些都存在于数据库中。如果发现预先准备好的推荐菜品和餐馆都不在这个商圈里，就需要利用零售AI中台实时进行计算得到新的合适的推荐，随即内容投放模块会将相应的内容触达用户，最终，用户实时地得到了菜品和餐馆的推荐列表。

第二种推荐模式是离线的后台个性化投放，利用人工智能模型对数据进行离线计算，把个性化的结果存储到相应数据库中，下次用户上线时可以看到。这些个性化的推荐不需要是实时的，也就是说，推荐内容可以在后台进行运算和决策，频率也不需要太高，比如说可以每天或每周迭代一次。例如，用户在电商网站上购买了婴儿奶粉，则零售AI中台可以预测出该用户对于母婴产品的需求较高，是母婴产品方面的高价值人群，因此在该用户后续使用的过程中，可以推荐宝宝辅食、宝宝玩具等产品。显然，这一个性化的内容投放不需要是实时的，推送频率也不用特别高。因此，可以把用户相关数据存储到相应的离线数据库中，用户下次上线就可以看到。每次打开外卖订餐App，首页的内容就是通过这样的方式呈现出来的：首先，将对应每个类别的用户的菜品和餐馆推荐列表在后台通过人工智能计算出来，然后存储在数据库中，当用户打开App输入其想要的菜品或餐馆时，内容投放

模块随即查询数据库，再由数据库返回预先准备好的推荐列表，呈现给用户。

第三种推荐模式是近实时的个性化投放，这是介于在线实时投放和离线后台投放之间的投放模式，不需要立刻做出个性化的推荐体验，但也不能延时太久。近实时个性化投放需要当场计算反馈离线数据库中没有的结果，但对时效性的要求没有实时的个性化投放那么高。因此，这种推荐模式可以理解为实时投放和离线投放的折中状态。例如，用户在浏览外卖订餐App的页面时可以进行选择，对不同的菜品或者餐馆表示喜欢或不喜欢。用户的反馈信息会被人工智能捕捉到，回流到零售AI中台进行学习，根据打分情况，更新外卖的推荐列表。当用户重新返回页面浏览相关的信息时，用户将获得实时的体验，App会推送用户喜欢的菜品和餐馆类型。近实时的个性化投放不像在线实时投放，能够让用户时时刻刻都能看到变化，也不像离线后台投放，要等后台的个性化数据刷新后才能看到变化，而是不知不觉中产生的变化，是一种“润物细无声”的变化。

实时的个性化投放、离线的后台个性化投放和近实时的个性化投放三种模式对应了不同的场景应用，根据对时效性要求的不同，这三种模式的搭配能充分满足用户个性化体验的需求。

个性化投放应用实例

根据中国电子商务协会发布的《中国电子商务发展报告2018—2019》，2018年中国电子商务交易总额高达31.63万亿元人民币，同比增长8.5%。由此可见，电子商务领域发展得如火如荼，但是大量商家的进入使电商市场逐渐趋于饱和，即使电商领域的潜力无穷，没有核心竞争力的商家仍然会面临倒闭的风险。以往，满足消费者衣食住行的基本需求已经足以让商家成功立足，但现在，满足用户的个性化需求才是真正的核心竞争力。现在成功的电商品牌，无一不是依靠个性

化投放留住用户，在用户、搜索、浏览、购买和支付等数据的基础上，进行实时、离线和近实时投放三方面的个性化业务，提升用户体验和平台的效益。

民以食为天，餐饮电商也在运用个性化投放来获得用户的青睐。饿了么是国内领先的专业网上订餐平台之一，在竞争激烈的餐饮行业中站稳了脚跟并做出了一番成绩。饿了么的核心竞争力正是在于，针对用户，它能提供精准的商家及美食匹配，满足用户的个性化饮食需求；针对商家，它能提供精准的流量，高效地带来用户和订单，并在商家的各个生命阶段为其提供个性化的发展方案。

事实上，用户的需求早已不是“千人千面”的问题了，同一个用户，在不同的时间、地点、季节和天气等不同消费场景下，需求是不同的，饿了么要面对的是“一人千面”的个性化难题。为了解决这个难题，饿了么首先基于对大数据的处理，进行实时数据分析以及用户画像；再通过机器学习，根据用户的位置、访问时间、搜索关键词等信息，对商家和菜品进行实时排序。也就是说，在不同的情况下，同一用户接收到的推荐内容是不一样的。比如在中午12点，用户的午饭时间，饭菜的排序会比较靠前，而在下午3点，用户的下午茶时间，饮料和小吃的排序则会更加靠前。为了给商家带来更多的有效流量，饿了么做出的努力是，离线进行商家菜品的口味、食材以及做法的标签化，通过这些标签训练人工智能模型，在实时搜索阶段，人工智能模型可以将搜索词与标签匹配，精准定位到某个商家的某个菜品。匹配的精准度越高，用户的下单率就越高，这样就为商家带来了大量的订单。在人工智能模型的帮助下，饿了么为用户和商家都提供了个性化投放服务，最终为饿了么平台自身赢得了数倍转化率的提升。

携程网是国内领先的综合性旅行服务网站，拥有数亿的用户数量和丰富的旅游产品。能够维护住数量如此多的用户群，携程网的秘诀同样在于个性化推荐。在携程网上，每天都有大量用户浏览和预订，并产生内容和相关数据，在海量的用户数据、内容数据和商品数据的

基础上，携程网通过人工智能模型，形成符合细分需求的个性化定制解决方案，帮助用户进行旅行决策。比如，根据客户的出行目的、地点、价格偏好、对酒店设施的需求等维度的不同，有针对性地进行推荐。有的用户是为了商务工作，有的用户是为了休闲度假，他们的需求显然不一样。个性化推荐提升了用户的使用体验，还带动了一整条旅行线路的发展。据估计，基于人工智能的个性化推荐相较传统方法，每年为携程带来过亿元的营收增量。

在国际上，也有许多企业在积极运用个性化推荐，零售巨头塔吉特（Target）便是其中的一个代表。塔吉特公司是美国的第二大零售商，个性化营销在其发展过程中发挥了重要作用。塔吉特是通过其开发的App——Cartwheel来实现个性化营销的。该App会通过用户的购买记录、购买意向和浏览信息等分析预测用户的需求，进行个性化推荐。例如，根据数据分析发现某个用户喜欢电子产品，便可以针对性地围绕电子产品进行推送；如果发现这个用户近期多次浏览某件商品，就可以推送相关折扣信息，促成购买行为。个性化营销大大提高了用户的活跃度，为塔吉特的电商销售额带来了50%的年增长率。

妮维雅作为在全球市场占有一席之地的百年护肤品牌，也在积极运用个性化投放提升效益。妮维雅通过分析用户的购买记录和个人偏好等数据建立用户个性化推荐模型。不同的用户在妮维雅旗舰店的首页看到的信息是不同的，页面内容是根据每个用户需求和购买行为来决定的。比如，一个用户已经很久没有消费了，妮维雅就会通过在首页放置一些促销满减的活动内容来刺激该用户消费。经过这样的个性化投放，在为用户带去更加贴心、符合需求的购物体验的同时，妮维雅的销售转化率也提升了约70%。

19世纪美国零售巨头约翰·沃纳梅克有句著名的悲叹：“我知道在广告上的投资有一半是无用的，但问题是我不知道是哪一半。”约翰·沃纳梅克创立了美国第一家百货商店“沃纳梅克氏”，被认为是百货商店之父，同时他也是第一个投放现代广告的商人。在传统的广

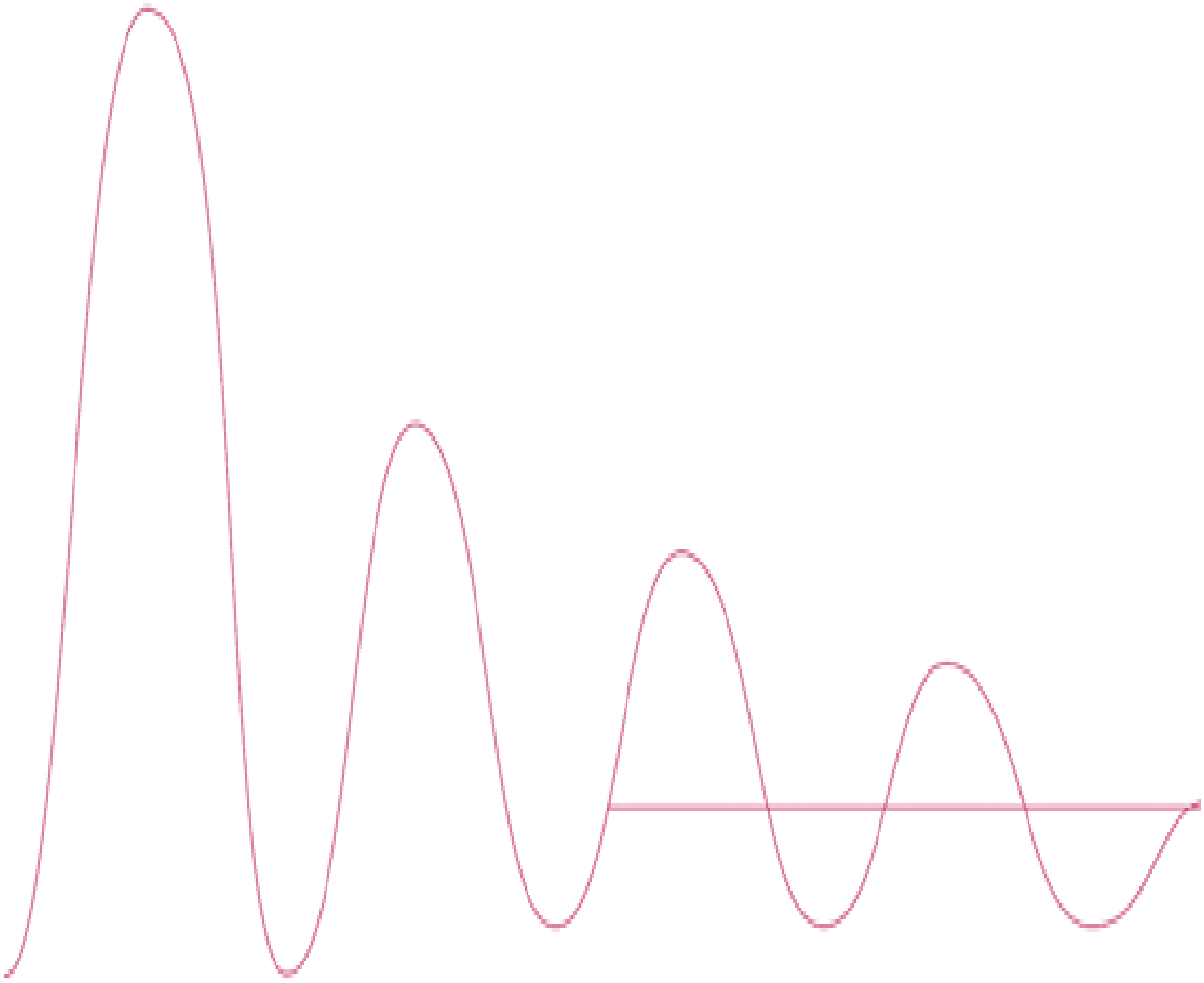
告营销中，至少有一半的投放是无效的，但在人工智能个性化投放落地后，这个问题得到了部分解决，因为个性化投放的转化率要比传统的广告高。

过去的传统广告是粗放式的广撒网模式，一个广告投放出去，可能有很多人都看到了，但这个广告能与多大比例的消费者需求契合则无法估算。这一方面是由于部分消费者本身并不需要这个产品，另一方面则是由于这个广告没有对应上这部分消费者的兴趣和喜好，无法激起他们的购买欲，所以转化率极低，这部分的广告投放就等同于浪费掉了。而有了个性化投放之后，传统广告营销的这两点弊端就能被很好地解决——根据人工智能分析结果，了解每个消费者的实际产品需求，分别做好对应的产品推送；另外，根据人工智能了解消费者的兴趣点，根据不同的风格喜好推送不同类型的营销内容，相较于传统广告营销，实现个性化投放后，能够在很大程度上提升投放的转化率。

在商业高度发展的今天，个性化体验逐步成为零售领域的一把金钥匙，通过人工智能对大量消费人群的特性进行分析，有针对性地选用触达方式和进行产品推荐，既能使零售业实现资源优化配置，也能让消费者的需求获得满足、享受到更好的消费体验。人工智能赋能新零售，将数据合理利用并优化决策，完成个性化投放，以更低成本获取更高效益。

第七章

AI实战的故事



每一种技术都有自己的性格，每一个市场也都有自己的性格，只有当技术和市场投缘时，才能擦出火花。

—《变量》—

在前几章里，我们从数字化赋能谈到了AI中台的构建，阐明了AI思维在行业中的应用主要是通过人工智能做出预测来影响相应的决策，实现业务结果的提升。这一切都离不开数据中台和AI中台的支撑，然后才能通过人工智能准确和稳定地预测各种行为和指标，在不断反馈、优化和迭代的过程中，形成一个良性的闭环，逐步提升各项业务的业绩。我们也分领域展示了人工智能落地的实践情况及具体的操作步骤，讲解如何使用人工智能进行用户行为分析、如何预测相应的指标，从而指导实际的业务决策，实现业务目标的提升。

而在这一章，我们一方面是要通过具体典型案例讲述人工智能在生产、供应和销售等各个环节中起到的升级和改造作用；另一方面，我们分析的案例囊括零售业、工业和金融业三大行业，每一个例子的讲解不会止步于是否利用了人工智能或者如何运用人工智能，还会进一步研讨利用了人工智能之后所能够达到的业务效果和影响。这章的内容会更加清晰地展示人工智能是如何在具体领域里展开运用的。

零售业AI实战

随着经济的波动发展以及人们生活方式的改变，传统零售业模式正在发生巨大的转变。整个零售市场的竞争环境变化急速、全球电子购物的兴起以及用户捉摸不定的情绪，均为零售业的发展前景带来变数。零售业的方方面面都贯穿了人工智能的应用。传统零售在长期的发展过程中积累了庞大的数据，通过人工智能可以消除数据孤岛，主动收集数据并把它转换为可利用的资源，从而提供更准确和更实时的预测，帮助人们做出更明智的决定。精准的预测可以帮助提高供应链管理水平，影响重大的促销主题，更合理地进行产品优化和创新，使得零售业的各个环节都能得到优化和提升。

图7-1是零售全链路流程的示意图，也是人工智能在零售领域落地实践涉及的范畴。在整个链路里，原材料经工厂加工生产为产品，产品通过线上线下两个渠道传递给用户。在线上渠道，产品一般会通过电商平台来展示，用户可以通过主动搜索或者其他社交渠道来了解这个产品的信息，并完成购买行为。在整个过程中，所产生的数据将为人工智能提供数据支撑。在线下渠道，产品通常由遍布在各地的门店销售给用户，线下的数据同样也可以应用于人工智能优化商业决策的过程中。

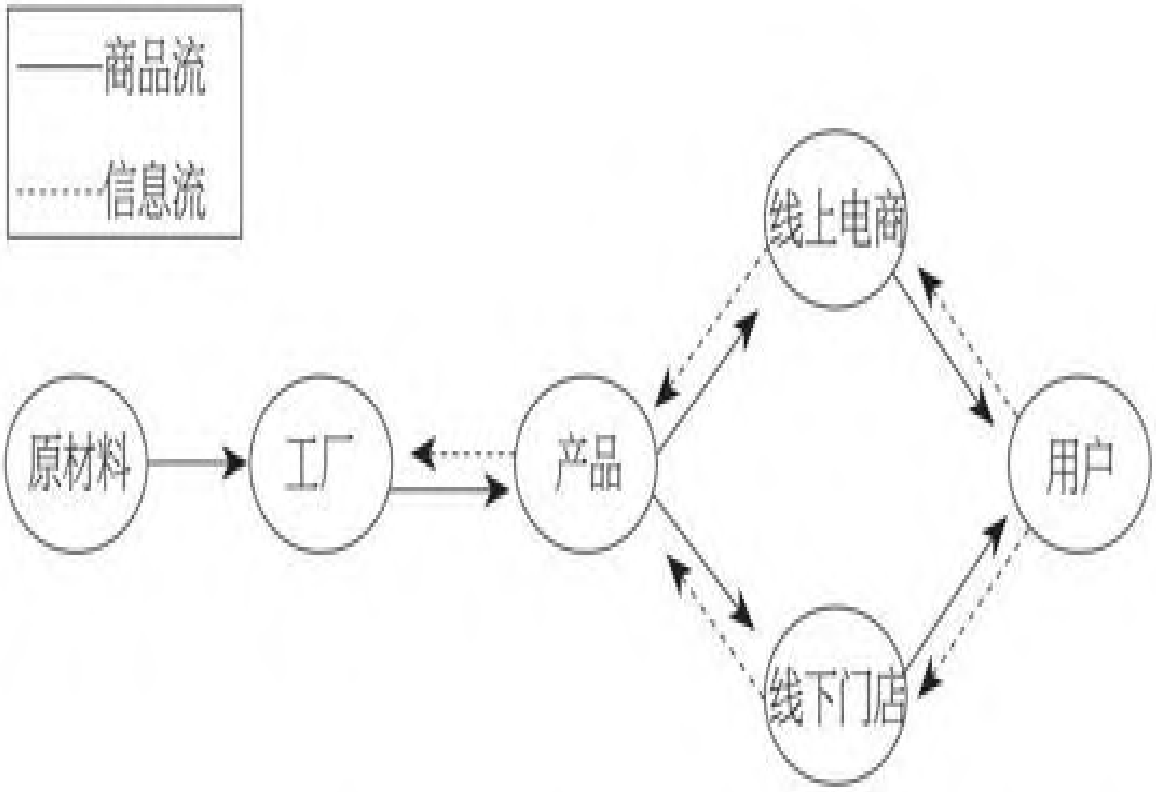


图7-1 零售全链路流程示意图

在传统的零售链路里，企业都是通过渠道将产品推给用户的，难以直接获得用户的反馈。这样容易导致信息的延误与失真，最终可能造成产品与用户需求的脱节和库存积压。而人工智能在零售领域落地

之后，会从用户的角度出发思考问题，例如，人工智能根据数据分析，可以通过用户在搜索、社交、电商和线下渠道的行为来预测他们的需求和喜好，以此指导产品生产和营销。当人工智能成功落地并运用于整个产供销体系后，最终是可以实现以用户需求打通供应和生产过程的，长此以往，企业的零售链路会进入一个对市场需求敏感的良性循环。

虽然人工智能可以赋能产供销中的各个环节，但是这个赋能过程不是一蹴而就的。通常来说，我们需要先在一个单点上形成人工智能的优势性应用，再逐渐接入产供销的各个环节中。最终，人工智能的应用在各环节形成闭环。这时，数据中台和AI中台才能极大地发挥作用。如果企业的生产、供应和销售环节都可以通过人工智能大脑来决策，那么整个过程就能通过对用户端需求最精准的把握来控制。这样会使产供销的各个环节损耗最小，企业库存降到最低，也能最及时地满足用户的需求，整体的产能和效益将是最高的。通过这样的大闭环逐渐推进人工智能在全流程、全行业的打通，真正实现C2F（从用户到厂商的模式）的理想才是人工智能大脑最终的目标。

人工智能赋能零售，主要有三个方面：一是营销，即通过人工智能精准推荐产品、展示广告以及设计会员权益；二是供应，即通过智选商品和销售预测来优化供应链，降低库存；三是产品和舆情洞察，即通过人工智能洞察用户的反馈。下面我将从这三个方面展开，详细讲解人工智能在实际应用的案例。每一个案例都会介绍人工智能落地应用和取得的商业结果。

智慧营销

营销是指通过各种途径和方式将产品传播出去，它直接关系到产品的销量以及企业的盈利，所以零售的营销环节，是人工智能参与的重要环节。零售业的发展要以用户为核心，但对企业来说，用户的价

值是存在差异性的，有低价值用户、中价值用户和高价值用户，而不论是哪一种价值用户，流失都是必然的。企业的资源又相对有限，不能将资源均等地用于这三种价值的用户的维护。高价值用户作为利润贡献的重点，对企业的生存和发展具有决定性作用，所以为了资源的合理有效分配，用于高价值用户维护的资源就应当较多。

区分用户价值并针对不同的价值群体，给他们提供恰当的服务和引导，能构建更加精准、有效的营销策略。比如说，高价值群体的营销重点在维持他们的忠客行为，同时进行追加销售和交叉销售；中价值群体的营销重点在追加销售和交叉销售，抓住促进消费的机会；低价值群体的营销重点在促销和提醒等，提升人均价值。

用户价值重在预测。所谓预测，不是直接使用历史上用户的购买记录，这样做意义不大，就比如某个用户这次虽买了某个品牌的洗衣机，但这并不意味着这个用户未来一定会持续在这个品牌消费。因为对于很多种商品来说，用户过去的价值跟将来的价值往往没有明显的相关性，用户过去的价值对企业来说仅仅起到了参考作用。而最佳的用户价值是预测性的价值，是要知道他将来对这个品牌来说价值是多少，这样企业才能进行更有效的运营。所以在零售业中，我们要重点来做的一件事情就是对用户的将来价值进行预测。

如图7-2，这是一幅用户价值预测示意图。用户的性别、年龄、爱好、学历、人生状态、购买行为和点击次数等数据，都可以作为输入，通过人工智能模型预测出用户价值。

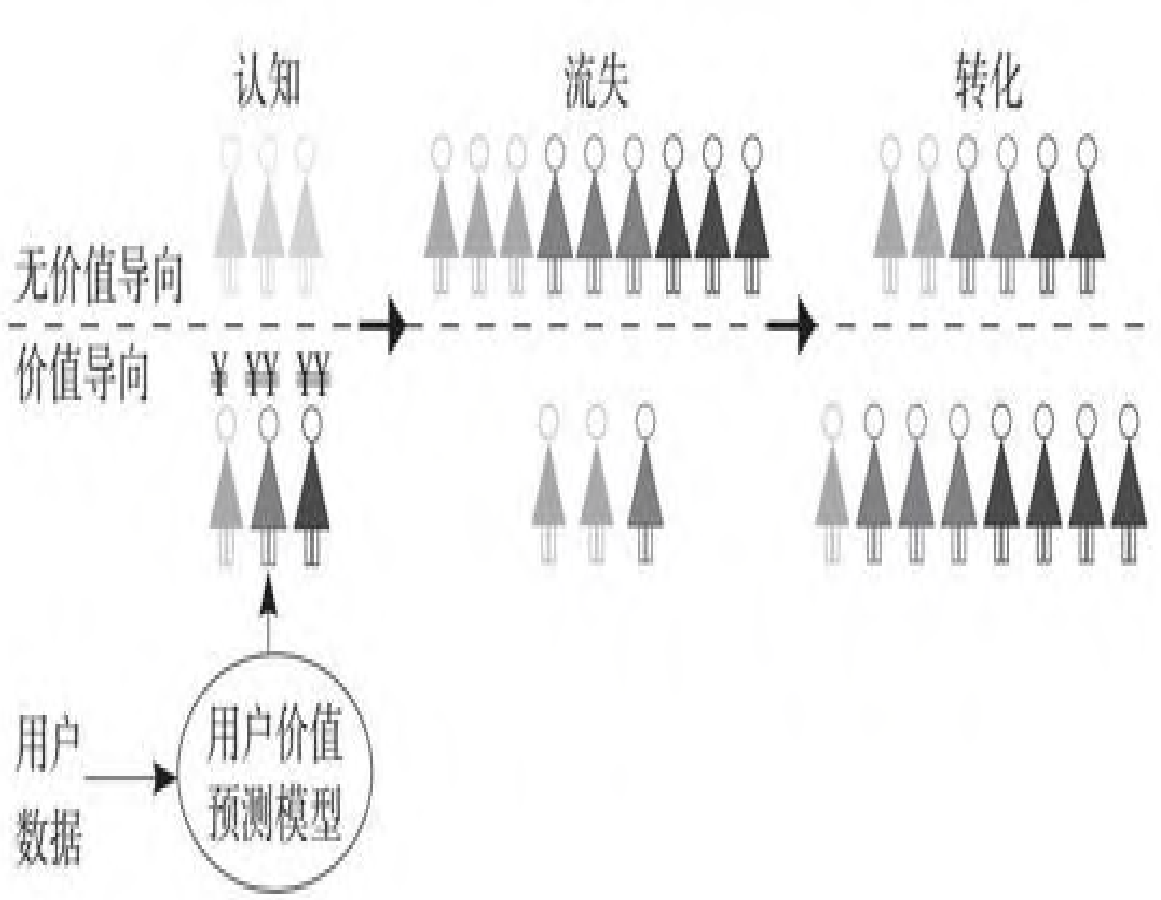


图7-2 用户价值预测示意图

接着，从用户的价值角度出发，对每个用户做差异化的预测，形成价值导向。如果无价值导向，就会造成“千人一面”的同质化经营现象，高价值用户流失大；价值导向能避免同质化运营，重点维护高价值的用户，降低流失用户里高价值用户的占比；同样，针对性的运营活动最终能使转化的人群里高价值用户占比居多。这样既降低了用户流失率，又提高了转化率，真正实现了为企业赋能。

智慧营销可以通过十分具有针对性的个性化推荐来实现。一直以来，大型超市会采用印刷宣传单的方式，向顾客宣传商品打折信息及促销信息等，这样既浪费时间又增加成本。在市场竞争越来越激烈的今天，如何让顾客有效快速地了解到所需商品的信息，是超市和顾客

共同面临又急需解决的现实问题。而人工智能可以通过基于用户需求的个性化产品推荐体系，来调整商品展示方式，让不同的人看到契合自己喜好的商品及优惠信息，实现营销的“千人千面”，从而提升用户转化率。

一家有着百年历史的大型连锁超市，就曾利用通过人工智能建立起的个性化产品推荐体系提高了自身利润。起初这家超市和大部分超市一样存在同质化程度严重的情况，顾客流失率大，导致利润率薄。通过建立个性化产品推荐体系，该超市的AI中台将顾客的表层肖像描述、深层需求挖掘以及产品之间的关系有机结合，预测顾客在将来一周内的购买行为，结合其他多种维度的因素，做出恰当的个性化推荐。基于人工智能的个性化推荐体系使该大型连锁超市新品推荐成功率上升了15.7%，顾客流失率降低了10.2%，人均消费总额提高了7.4%。

智慧营销的另一个体现就是会员制，会员体系在如今的商业模式中已成常态，众多平台都打造出了自己的会员体系。例如，大部分企业会对会员提供一系列的购买优惠、增值服务或给予其他奖励，这样不但可以奖励忠诚顾客，而且能够刺激消费并留住核心顾客。会员权益设计作为顾客运营的重要部分，以“福利”之名不断刺激顾客进行升级，产生物质激励的作用，有部分权益应该是尊享型的，只对极少顾客开放，营造一种稀缺感和尊贵感，起到了精神激励的作用。

设计会员权益主要有两大要点，第一是预估顾客价值，预估不同等级的顾客在生命周期内能创造的价值，设计权益时尽量把控权益成本。第二是权益匹配，不同等级的会员的权益要有所差别，譬如，有些基础权益是会员都可以享有的，而另外一些权益是会员升级达成一定的条件才可以享受的，如果随意地进行权益匹配，未必能起到好的激励作用。

会员体系虽然由来已久，但过去做会员运营，方式比较单一，所以会员活跃度并不高；通过人工智能权益设计进行的会员运营，在加

强品牌忠诚度的同时，也能提升会员复购率。例如，一家旗下有多个知名品牌的服装零售巨头想要通过这样的会员权益设计，来提高会员复购率。具体做法首先是处理海量数据，包括运营数据、顾客数据、销售数据、交易数据、商品数据和行为数据等，之后利用这些数据进行顾客价值分析及聚类分析，得到的客群分类与顾客价值一一匹配。

权益设计是基于联合分析手段进行的。通过人工智能进行学习建模，能预测出哪种权益类型对应哪类客群所产生的商业价值可以达到最大，从而恰当地将客群与相应的权益类型进行匹配，比如客群一对应某种VIP权益，客群二对应某种礼品权益，客群三对应某种积分权益，等等。这种忠客运营体系，为不同客群匹配最佳权益，增加了客群忠诚度，提高了复购率，最终为该服装零售巨头带来了2亿多元的增量收入。

用户价值预测和个性化推荐等都需要基于用户的完整画像，但是由于市面上媒体的形式和属性繁杂，同一个人可能在不同媒体上停留，因此想要完全地了解用户需求，不重不漏地为其提供合适的广告内容，还是有一定难度的。全域营销就是为实现在极为多样的媒体上识别这个人，并且快速地提供与这个人相关的广告而出现的。全域营销需要打通整个生命周期、全链路、全媒体和全渠道的数据，通过人工智能全面提升有效触达率、转化率以及营销效率。

如图7-3，这是全域营销的示意图，具体做法是首先将企业内部所有数据打通，无论用户是通过App端、网页端还是线下消费形成的业务数据、用户数据、交易数据、竞品数据、外部数据和全链路行为数据，都要在中台内打通。然后，通过数据清洗和特征提取，建立一系列基于机器学习或深度学习的人群画像模型、个性化创意模型、个性化触达模型、价值预测模型、个性推荐模型及忠诚奖励模型等，增加用户的黏性。比如说，一家大型电商平台想要通过全域营销提高营销效果，那么在这其中，全域营销的主要任务是通过合适的渠道把精准的内容推送到用户面前，这家电商平台通过人工智能，可以精准地预

测出用户想要购买的商品，然后再通过各种各样营销渠道，包括App、电话和门户网站等，把相应的广告内容推送给用户，从而最终增加整体的响应率。



图7-3 全域营销示意图

在全域营销过程中，人工智能可以提升包括引流、潜客转化、新客互动、用户运营、存量提升、防止流失、流失激活和营销触达等各环节的效率，最终实现更多的转化，这不仅使得该电商平台营销活动平均响应率提升数倍，营销活动的总体投资回报率也提升接近40%，同时用户关于精准性的正向反馈也大幅度提升。如果原来付出一元钱只能产生两元钱，现在付出一元钱能产生将近三元钱。按照这个倍数来看，在营销费用很大的情况下，营销的投资回报率只要提升几个百分点就能够产生非常可观的增量收入。

智慧广告

广告作为营销活动必不可少的手段，使企业可以开发和利用优质的营销资源，满足用户的需求。在零售营销的过程中，利用人工智能来优化广告效果发挥着重要的作用。我们以一家知名线下品牌为例，详细解读人工智能是如何实现广告优化的。该品牌在几年前斥巨资购买了某头部App首屏广告位，虽然凭头部App庞大的流量，该品牌的广告获得了较多曝光次数，但是由于一个广告创意只能和部分用户群体产生共鸣，大部分用户看到广告也不会点击。其结果是虽然广告费用高昂，投放产生的效果却并不显著。

而通过人工智能可以对数据进行分析，指导和优化投放方式，帮助企业进行高效率、高性价比的精准广告投放。具体的做法是，首先打通这个品牌自身的数据和广告投放App的数据，通过用户在App上及站外的行为轨迹，包括浏览情况、收藏情况、广告点击情况和商品购买情况等数据，来掌握其点击和购买行为的规律。例如，某些用户更喜欢看母婴类资讯或者体育类资讯，我们利用人工智能将这些兴趣标签化，形成相应的用户画像。再根据大量用户的画像标签，建立起人工智能模型，预测任何一个用户对不同广告内容的点击率以及转化率。

这里的优化目标有两个，第一个是优化点击率，考量广告有效性的一个重要指标就是点击率，有了点击才有可能进一步影响用户。因此，我们需要通过人工智能对上百种广告创意打分以及做智能排序，选出点击率最高的广告推送给用户，从而产生更高的点击率。第二个是优化转化率，品牌自身长期积累了千万级的用户数据，利用这些数据，通过人工智能优化产品的营销和推广，可以获得更高的价值回报。经过人工智能优化之后，这个品牌在该App上的整体投资回报率提升了将近2倍。

在这个案例里，品牌虽然还没有建立起完整的数据中台和AI中台，但已经实现了一定程度的数据打通，这样依然能够实现人工智能的价值。所以，对于企业来说，并不是必须构建完善的中台体系后才能实现人工智能赋能，而是可以先在场景下实现局部打通，从单点人工智能的建立起步，逐步完善，落地以后就可以获得相应业务价值的提升。

智慧供应

纵观所有零售体系，用户是其核心，但是商品的供应也同样是关键。商品供应链将商品从生产链转移到消费链，使商品能够被用户接触、了解和消费。商品供应链将生产商、分销商、零售商以及用户连接在了一起，形成一个完整的零售链，能够保证商品的流通和消费。而在商品供应的过程中，选品至关重要。要想提高商品的销售额，就要做好前期的选品工作。所谓选品，就是通过数据分析，决定选出哪些商品或者商品的哪些型号，来进行销售。

一家位于华中地区的大型超市集团，其连锁门店一直存在货架上放置商品的品类太多、货品积压严重的问题，而且积压的商品有很大一部分是低价值商品，并不会带来较高的利润。该超市最终通过人工智能解决了这个问题。人工智能主要从两方面考虑这个问题：一个是

通过商品对会员价值贡献进行预测，基于会员的未来价值并结合其消费习惯，保留对会员价值贡献度最高的商品；另一个是通过进行商品关联度的预测，找到需要保留的商品组合。基于这两个方面的预测，对商品的重要性做出总体的评价，从而做出选品决策。通过人工智能对商品重要性做出评价来选品，可以帮助这个大型连锁超市单门店减少14%的商品，在降低运营成本的同时还可以提升利润率。

智慧供应的重要一环是销售预测。销售预测是指通过人工智能对以往的销售数据以及其他大数据维度进行分析，最终形成对未来销售情况的预测。通过销售预测，可以以销定产，合理安排库存，避免商品积压；为商品的及时补货提供参考依据；还可以调动工作人员积极性，使商品尽早销售出去，创造更多的价值。

在传统的销售预测里，一般采用征集销售人员或专家意见的方式来预测市场的需求量，或是采用基础统计的方法，这些方法不仅需要较大的工作量和成本，而且预测的准确度也不高。人工智能的销售预测和传统销售预测的区别在于，人工智能预测会针对更多数据维度。人工智能的数据输入包括每天的进销存数据、外部环境数据、供应商数据、行业产品销售数据、天气影响数据、产品及竞品促销数据等。这里面包括很多大数据维度，这些数据维度整合在一起，可以帮助AI中台进行销售预测，预测未来一段时间内商品的销售情况。

在零售场景中，基于大数据的人工智能销售预测，对未来两周内销售预测的准确率能够达到75%~85%。相比而言，运用传统策略加上人工经验的方法，销售预测的准确率一般最高只有70%，所以说人工智能可以帮助零售商更好地预测商品的销量。同时，基于大数据的人工智能销售预测也可以显著地降低库存周转天数，直接实现相应的效益提升。例如，生鲜类品牌通过这种销售预测可以优化库存管理、减少生鲜损耗、降低经营风险。

产品和舆情洞察

在零售行业，我们一直强调要以用户为核心，但是我们的最终目的都是将产品销售给用户，所以打造满足用户需求的产品至关重要。因此，产品洞察也是人工智能一直以来的重要应用方向。传统上，如果要推出新的产品，或者进行产品优化，需要通过产品经理去发现高价值的创新点，以及不断地思考产品定位、差异点等问题才能实现。而通过人工智能进行分析预测，可以快速找到符合用户需求的产品功能，提供可靠的产品研发建议，有效地避免了人工决策过程中可能出现的失误。

比如说，人工智能就曾帮助一个市值百亿元的家电品牌选择研发方向，其具体做法是首先分析各种网络平台上跟该品牌产品相关的用户评论和吐槽等海量非结构化数据。基于人工智能的自然语言理解，把数据进行标签化处理，并从海量数据中分析、提炼和归纳出高价值点，从而帮助企业优化或创新产品。通过人工智能，这个家电品牌每月都能发现数百个关于用户需求的潜在高价值点。

具体来说，人工智能通过训练语料可以学习出相应的语义模型，从而为新的语料打上准确的语义标签，进而产生对应的用户关注点、用户痛点和高价值点。这一过程主要分为两步：第一步是给评论做标签化语义处理，理解用户在表达什么东西，比如用户对于空调的评价，是噪音太大，还是制冷太慢，又或者是对吸尘器的评价，是吸不干净，还是不好操作，等等；第二步，进一步筛选和分析语料，判断普通关注点、痛点和高价值点。通过人工智能对用户的关注点、痛点和高价值点进行标记以后，能够为企业的产品决策提供便捷有效的工具。

舆情是一个企业的形象，是用户对产品和服务的情感评价。如果舆情正面，大家都对产品和服务表示欣赏，会进一步吸引其他受众购买；如果舆情较为负面，不但会失去新的潜在用户，现有的用户也有可能会流失。所以，舆情洞察也是零售企业需要高度重视的环节。人工智能对舆情分析起到的作用是，基于自然语言理解和情感分析，迅

速洞察并处理企业可能出现的公共关系问题，其中包括媒体传播评估、危机预警和危机管理等功能。例如，当出现负面评论时，企业可以通过人工智能实时做出检测和反馈，及时维护品牌形象。

总部位于广东的一家著名住宅开发商就通过人工智能监测舆情，并且卓有成效。这家住宅开发商每日需要关注的相关资讯非常多，所以需要一个强大的舆情监控系统来对舆情数据进行收集、整理和分析。该开发商的人工智能舆情监测系统能够根据想要监测的维度，对相关舆情进行实时不间断的监测，通过这个系统曾成功预警5次重大危机事件，有效地制止了危机信息的扩散，维护了品牌形象。

通过人工智能，零售业各个环节都能得到优化和提升。在智慧营销层面，人工智能可以构建每一位用户全方位的画像，提升广告和内容的精准度，从而做到千人千面，实现精准营销。在供应链层面，人工智能可以帮助企业降低库存和供应链成本，实现效益提升，形成基于用户需求和线上、线下销售一体化的供应链智能管理体系，提升企业经营效率。再到产品层面，人工智能可以更为准确地找到产品痛点，帮助企业做好产品的优化和创新，还可以通过对企业舆情的实时监控，有效制止危机扩散，维护企业形象。

人工智能的应用最终都指向同样的目标和追求：降低成本、提高效率、提升体验。随着技术水平的飞速发展，人工智能在零售领域的用武之地也越来越广，发挥的作用越来越大。AI思维一直坚持数据驱动决策，我们完全可以相信，在不久的将来，人工智能不但会继续赋能零售业，还会驱动整个零售业的变革和发展，为用户和企业带来更多惊喜。

工业AI实战

工业4.0时代对传统工业来说是一种挑战，更是不可多得的机遇。要提高整个行业的智能化水平，促进产业的变革，工业必须借助人工智能这把利剑。人工智能在工业中发挥潜能，在规模化的工业上创造出的价值可以让所有人受益。经过几十年的发展，工业自动化已经可以在生产中被广泛应用，但人工智能在工业上的应用目前还处于起步阶段，只在个别的工业场景实现了人工智能的应用。工业人工智能作为一种系统化的方法论可以为工业应用提供解决方案，不仅可以让工厂自行生产和组装，也可以具备智能管理、智能质检和智能维护等功能。随着第四次工业革命来临，工业自动化将与人工智能融合提升，助力工业实现智能化升级，为工业带来新的增长点。

在这个过程中，人工智能通过对大量工业生产制造过程中产生的数据智能感知，通过模型进行决策，最终实现优化控制、提高质量、降低成本和提升效益等效果。人工智能在具体的生产操作和检测中可以比人工更加精准和及时，在很大程度上提高工程效率，为企业节省成本、提升效益。

人工智能的建模分析是实现人工智能赋能工业的基础，下面我们将从智慧工厂、智慧工地和智能环保这三个方面展开，详细讲解人工智能模型实际应用的案例。

智慧工厂提升生产效益

随着工业4.0时代的到来，工业智能化越来越成为未来工业的发展趋势。企业为应对日益激烈的工业市场竞争，将智能制造视为必备武器。智慧工厂是将产品和制造连接在一起的纽带，使智能制造成功落地。智慧工厂在智能制造中的作用日益凸显，工业领域中越来越多的企业开始配备自己的智慧工厂。

智慧工厂是指在对工厂各个系统、各个流程的运行数据深刻分析的基础上，训练出基于人工智能的系统，为工厂提供智能决策服务，

并具有自我学习和自我更新功能。简单点来理解，智慧工厂就是在数字化工厂的基础上，减少了生产线上的人员参与，实现了工厂从原料选择、产品设计、产品制造再到物流等流程的自动化和智能化，既能降低生产成本和能耗，又能提高生产效率，在满足市场需求的同时，为企业带来更多的盈利。

例如，一家全球领先的石化公司就通过智慧工厂优化生产，大大降低了能耗。该石化公司的一个主要业务是化纤生产。化纤是将煤、石油、天然气等原材料经过化学加工得到的，所以化纤生产属于高能耗产业。在生产化纤的过程中，锅炉燃烧需要合理控制数百个参数，比如燃料量、氧气量和进风量等，这里面任何一个参数的变化都有可能影响到化纤生产过程中的能耗。但传统化纤行业在多年的实践中，并没有形成一套科学的参数调节方法，而是大体依靠人们以往的经验来判断和调节。由于煤炭燃烧成本非常高，如果继续按照以往的方法来做，很难实现燃烧效率的提升。最终，该企业成功通过人工智能解决了这个问题。

通过对化纤生产过程中的数据进行采集分析，根据数据训练人工智能模型，并将模型输出的控制参数反馈给生产系统，可以实现精准控制和优化，具体做法如下：

首先，该企业进行了数据采集分析。化纤生产的数据采集包括两方面，一方面是采集锅炉燃烧过程中所需要操控的数百个参数；另一方面，需要采集在一段时间内形成的历史数据。通过对各种参数的综合分析，找出对锅炉燃烧能耗影响最大的一些参数定为关键参数。采集这些关键参数与燃烧能耗的历史数据，形成模型学习的训练集。之后，通过数据分析推导出关键参数（x）与燃烧能耗（y）之间的变量关系，根据这个变量关系，并以训练集为样本来学习人工智能模型。

模型搭建完成之后就可以发挥作用了，所有的人工智能模型都是为了落地实践而搭建的，所以要将模型输出的结果反馈给生产系统，调整生产过程中的参数，实现精准控制优化。在实际运行过程中形成

的新数据也会实时更新到模型中，以此来不断优化和完善模型，提高模型决策的准确性。通过人工智能模型形成的决策，可以对化纤生产过程中的相关参数进行优化调整，降低生产的成本和能耗，产生实际的经济效益。该企业通过人工智能模型的精准控制，最终在化纤生产中实现的效果是燃煤消耗显著降低，每年节省的燃煤成本逾千万元。

对于一个社会来说，高能耗固然能在短时间内带来快速的发展，但是从长远来看，这无异于饮鸩止渴，竭泽而渔。为了可持续发展目标的实现，节能降耗是必由之路。工业作为国民经济的主体，是能源消耗的重要领域之一，在工业生产中降低能耗应该是责无旁贷之举。人工智能目前已经在工业领域的很多环节实现了落地，能够有效地帮助工业节约能耗，而且依照现在的发展态势来看，我们可以相信，人工智能在未来能够为工业的节能降耗带来更多的惊喜。

智慧工地提升生产安全

根据住房和城乡建设部办公厅发布的《关于2018年房屋市政工程生产安全事故和建筑施工安全专项治理行动情况的通报》的数据，2018年，全国共发生房屋市政工程生产安全事故734起、死亡840人。近几年，工地事故起数、死亡人数不断攀升，工地生产安全形势不容乐观。安全帽是工地上保护工人安全的最常见的物品，如果发生意外事故，安全帽的帽壳可以保护人头部免于较大坠落物的伤害，减少2/3以上的冲击力，对保护施工人员的生命安全起到了至关重要的作用。施工现场坍塌是工地伤亡数量居高不下的一个重要原因，而在那些因坍塌而不幸伤亡的人中，许多都没有按照规定佩戴安全帽。如果不注意正确佩戴安全帽甚至不戴安全帽在施工现场进出，施工上空掉下的一颗小小的螺丝帽都可能威胁到施工人员的生命安全，更别说轰然倒塌的钢筋铁壁和水泥高墙了。

为加强对工地的安全管理，保护施工人员的安全，国家要求进入工地的施工人员必须戴安全帽。工地也通过安全教育来提醒施工人员佩戴安全帽，但仍然有很多工人因为一时疏忽或者存有侥幸心理而没有佩戴，最终发生安全事故，追悔莫及。如何监督工人时时佩戴安全帽成为维护工地安全过程中需要解决的一个重大问题，但是由于工地上的人员流动性强，人工监督不太现实，此时，人工智能就派上了用场。

为了便于管理，一家大型矿山企业的工地上部署了安全帽的人工智能识别系统，通过人工智能来监督工地上的工作人员是否佩戴安全帽。这一识别系统的工作原理是通过摄像机捕捉的视频流进行图像识别，具体来说，就是获取工地上的图像数据，并上传到部署在云端的人工智能模型中，判断工地上的人员是否佩戴了安全帽。识别主要包括两方面：第一，通过图像识别模型进行检测，识别图像中有没有人；第二，如果有人，通过图像识别模型检测图像中有没有安全帽，如果图像中既有人又没有佩戴安全帽，那就是违规事件，该系统会自动发出警示。但这个系统在实施过程中存在一个难点——由于工地上的基础设施条件不完善，机器配置通常较低，所以数据传输有不稳定的问题。在这种情况下，要落地人工智能识别系统，首先要完善工地上数据获取的软硬件系统，这样才能稳定地把数据上传到云端的人工智能模型，准确地识别出图像中是否有人，以及这些人里面是否存在没有佩戴安全帽的情况。

需要强调的一点是，通用的图像识别模型最开始并不能直接识别出图像中是否存在安全帽，我们需要结合大量安全帽的图像数据，在通用模型的基础上进行训练，才能形成行业定制的安全帽识别模型。所以，要构建安全帽识别模型，首先要对安全帽的图像数据进行人工标注。标注工作完成之后，一共形成了超过10万个不同的安全帽样本，涵盖了不同颜色和不同样式的安全帽。总的来说，安全帽的图像数据标注是一项非常细致的工作，需要耗费大量的人力和时间成本，

但这些成本都是必须付出的，因为只有拥有大量的标注样本，才能训练相应的安全帽识别模型。

安全帽数据的标注工作完成后，接下来要做的是训练深度学习模型。深度学习模型训练完成后可以进行自我学习和迭代。对于人工智能识别系统来说，新数据反馈的闭环必不可少，只有形成了闭环，模型的准确率才会越来越高，适应范围也会越来越大。例如，企业可以在这套系统上圈定更多的安全帽图像样本。现在安全帽常见的颜色有四五种，如果以后出现了更多的颜色，就可以通过上述功能使更多颜色的安全帽也可以被识别出来。这套基于人工智能的安全帽识别系统，判断的准确率超过98%，并且能够针对违规的情况报警，可以有效地减少工地上的工人不佩戴安全帽的现象。

智能环保提升减排效益

环境保护是人类社会永恒的话题，可持续早已成为全球发展工作的总目标，随着“绿水青山就是金山银山”的理念深入人心，环保更加成为当代社会关注的焦点。而工业作为高污染行业，首当其冲应该做出改变。

我国作为钢铁生产大国，钢铁产量在全球遥遥领先。但由于钢铁工艺流程长、产污环节多，钢铁行业已经成为工业生产中污染的重灾区，其环保问题备受关注。中国环境与发展国际合作委员会的数据显示，2017年钢铁行业二氧化硫的排放量为106万吨，氮氧化物的排放量为172万吨，颗粒物排放量为281万吨，钢铁行业已经成为工业部门最大的污染物排放来源。为了久违的蓝天白云、绿树碧海，亟须解决钢铁行业的环境污染问题。

就我国的钢铁行业而言，节能环保是全行业的主要短板。而人工智能有机会改变传统钢铁生产的认知，在效率、能耗、质量等方面改善生产过程，极大助力行业的可持续发展。“智能+环保”一直受到业

界的广泛关注。国家提出了“智能环保”的要求，甚至在《生态环境大数据建设总体方案》中确定了未来五年的具体目标。人工智能现在已经能通过数据整理和分析，应用于环境保护之中，为环境治理和环境保护提出智能解决方案了。数据显示，科学高效的智能环保方案能够优化工业生产流程，降低生产能耗，提升工艺指标，极大地提高能源资源的利用率，遏制浪费，减少排放。由此可见，人工智能在助力环保升级方面存在巨大的发展空间。

国内一家大型钢铁集团就通过人工智能助力工业生产，提升环保效益。钢铁生产过程中会产生大量的二氧化硫，但国家对于工业生产过程中产生的二氧化硫有硬性的排放标准：二氧化硫排放限值是每立方米200毫克，所以工厂要使用脱硫塔对二氧化硫进行处理，降低钢铁生产过程中的二氧化硫排放量。为了达到国家标准，同时优化原料消耗和水电能耗，该工厂在钢铁生产过程中使用人工智能模型，该模型通过输入在脱硫塔入口处获得的工业废气的相关参数和其他可监测的系统参数，预测脱硫塔出口处二氧化硫的含量，将二氧化硫含量稳定在排放标准以下。

在这中间要注意控制的参数包括下降液调节阀的进水量和风机变频器的功率，它们会直接影响到石灰原料和水电的消耗；当然对于其他参数的控制也不能忽视。该工厂控制这些参数时采用的方式是：基于人工智能模型的辅助决策，并由人参与到反馈闭环过程。按照这种方式，通过模型的预测和优化，可以辅助操作人员选择合适的控制参数。在实践中，人工智能可以找到符合排污标准的石灰原料消耗和水电能耗的最佳平衡点。运用人工智能，在生产过程中就可以通过模型的辅助决策，优化控制下降液调节阀的进水量和风机变频器的功率等，与此同时，还能通过预测模型确保排出的二氧化硫浓度控制在排放标准以下。

接下来具体看看其运用人工智能的步骤，如图7-4所示，首先要根据环保场景的数据进行人工智能控制模型训练。在输入原始数据后，

对数据进行过滤和特征提取，然后进行模型训练。之后需要进行参数调优，再做模型评估和优化步骤，最终给出推荐的控制策略，即在不同输入二氧化硫浓度的情况下，对应地推荐控制策略。

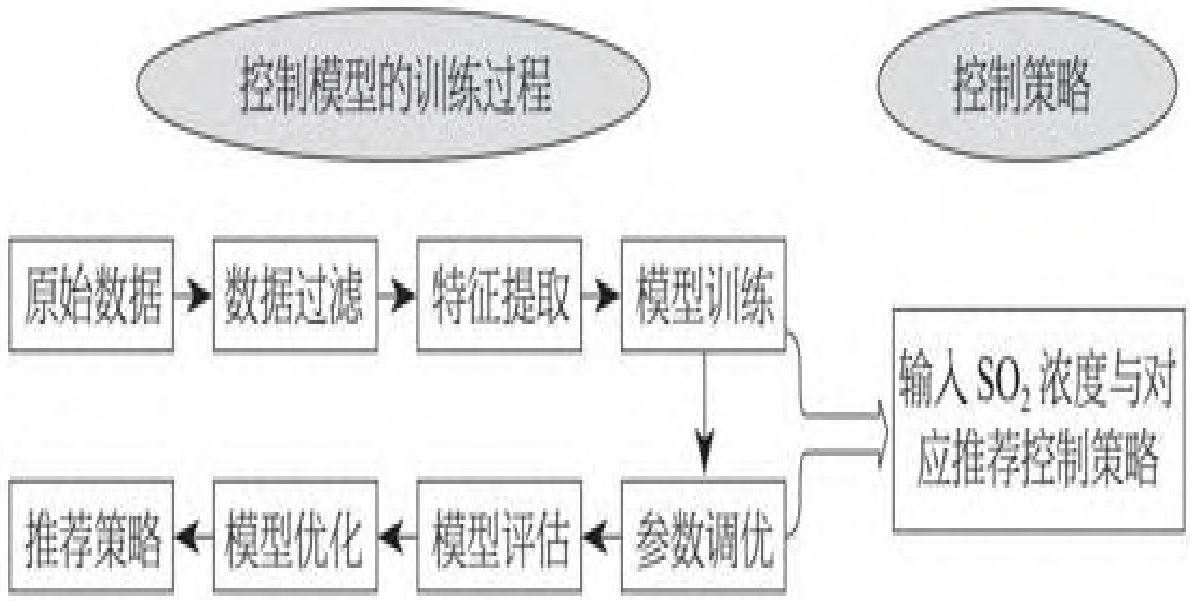


图7-4 控制模型的训练过程和控制策略示意图

传统的运作流程是通过人工控制多个参数，稍有差池就可能导致不合规，或者过多的原料消耗和水电能耗。通过对比传统人工操作和人工智能的推荐控制策略发现，使用后者时脱硫塔消耗的原料和水电量明显减少，单台脱硫塔年均可节省数十万元。

虽然人工智能现在已经能为工业领域乃至整个人类社会带来很大的裨益，但目前人工智能仍处于早期的发展阶段，在推动人工智能在工业领域落地的过程中仍然存在着诸多的问题。其中最主要的问题是工业业务理解中的巨大门槛和工业应用定制化程度高的问题。在一个细分问题对应一个模型的情况下，如果要在工业领域落地，那么涉及的细分问题成千上万，繁杂程度可见一斑。另外，还存在一系列问题，如技术门槛高，人工智能人才在短期内难以有效补充；行业缺乏

统一的系统和标准，应用开发和实施成本高；所依赖的数据的完整性和质量有待提高；应用场景的碎片化、个性化、专业化制约着复制推广和用户接受程度等。由此可见，人工智能要还有很多难关需要攻克，但人工智能所蕴含的强大动力让我们有足够的信心相信在未来的发展过程中，它能够不断完善，解决这些问题，逐步实现在工业领域的落地。

虽然企业还是需要按部就班地进行生产制造，但人工智能已经开始为整个工业领域带来很大的颠覆，加速生产制造进入工业4.0时代的步伐。人工智能带来的是为整个工业的强大赋能，从生产优化到供应链管理，从降低成本到提升效益，从降低能耗到智能环保，从立足数据到驱动决策。这是人工智能赋予工业的强大动力，也是AI思维下工业转型的必由之路。

金融业AI实战

根据中国人民银行发布的《2018年中国金融市场发展报告》的数据显示，我国经济发展稳中有进，2018年GDP（国内生产总值）继2017年突破80万亿元大关后，迈上90万亿元新台阶，2018年GDP增量接近8万亿元。社会经济水平不断提高，人们的生活水平也在不断改善，金融意识也开始慢慢提升，越来越多的人开始参与到金融业务中来，例如个人的信用贷款、买房时的房贷以及基金投资等。但对金融行业本身来说，大批人群的参与会使得金融行业审核和风险管控的成本升高。尤其在贷款业务里，如果对金融客户的审核不到位，导致坏账出现，将严重影响业务开展乃至整体金融行业的稳定发展，从而也会对金融所服务的实体经济产生不良影响。所以，通过人工智能优化金融各业务的运营和风险对整个经济体系来说都非常重要。

在前面的章节，我们讲解了人工智能可以赋能很多行业和领域，但其实并不是所有领域都适合大范围使用人工智能。要满足大规模商业应用人工智能，应具备两个必要条件：第一，数据的数量和质量必须达到一定的要求，尤其是整个流程中数据的打通和定期的数据更新和反馈，这决定了人工智能发展的基础是否牢固。第二，所在领域是否对需要解决的问题有相对清晰的定义。如果领域本身没有明晰的问题定义，或者不能通过数字化的方法描述出来，则很难通过人工智能来解决这类问题。

从整体来看，金融行业经过多年的发展已经沉淀出大量金融数据，而且金融领域现在需要解决的问题也很明确，营销获客、风险控制和投资顾问等一系列清晰的问题早已浮出水面，由此可见金融行业已经比较接近人工智能大规模商用的两点要求。随着人工智能的发展和在金融领域的不断探索，人工智能现在已经能够赋能金融行业的不同的业务场景，例如智能营销、智能风控、智能投顾和智能投研等，下面会通过具体的案例来展示人工智能对金融业的赋能和提升。

智能营销

和零售一样，金融行业提供的各种产品最终都是要服务于用户的，所以为了让更多的用户了解到自己的产品和服务，在金融行业营销也是必不可少的。同样地，人工智能也能助力金融营销，也就是我们通常所说的金融智能营销。金融智能营销指的是通过人工智能精准定向用户的金融需求并满足这些需求，一般有三个核心步骤。

第一步是进行金融画像。与零售领域一样，金融行业要通过人工智能得到提升，前提是有足够的数据积累，然后在金融业务里产生全面、准确、实时的用户画像，例如，需要知道用户的基本属性，用户有哪些信贷和理财方面的需求，用户对哪些金融产品感兴趣，是否有

偿还能力，是否有既往的借贷和逾期行为，是否有其他的风险行为，等等，这些维度通过加工整理形成用户的金融画像。

第二步是进行智能创意。所谓智能创意，即根据不同的人群画像，生成与需求场景自然切合的动态创意内容，实现千人千面的创意。根据不同消费者的个性化需求场景、所处人生阶段、生活品质、消费水平以及客户的生命周期等相关数据，通过人工智能的分析和学习，从情感激励、行动召唤、产品匹配和其他方面切入，进行针对不同客户群体的创意优化。智能创意进入营销系统进行精准营销后，营销结果反馈回来，成为新的数据来源用于下一轮智能创意的优化。

第三步是进行智能匹配，即通过人工智能响应模型，准确地匹配金融产品与客户的需求。基于机器学习和深度学习的响应模型，将产品特征（如额度、利率、收益率）与消费者的画像特征（例如用户画像、借贷行为、消费意愿）相匹配，恰当、精准地为金融客户推荐与他们需求相匹配的金融产品，最终实现智能营销。利用人工智能整合数据资源，打造数据闭环，从海量数据中挖掘金融客户潜在的模式和趋势，实现基于数据的智能营销体系。

除了把握用户的需求之外，金融机构也要了解用户对于自身的价值，也就是说，金融机构需要运用人工智能预测出一个用户在未来能够带来多少收益。预测之后，金融机构就可以根据用户价值的不同相对应地来优化营销活动。例如对于高价值的用户，重点在于注重保持他的价值水平，因而需要针对高价值用户的需求进行各种各样的营销活动，通过活动让用户产生更多价值，同时还要保证该用户一直都是高价值用户。而对于低价值用户，可以通过权益和优惠活动提升黏性。每一个价值段的用户都需要不同的运营手段来维护，要有差异化的考虑。如果用户看到的营销内容千人一面，运营同质化严重，将不能充分地实现和提升用户价值。通过预测性人工智能体系，针对高价值、中价值和低价值用户各自的特点，设计专属的运营手段、权益和

话术以后，总体效益可以得到很大的提升，客户流失率降低、利润率提高。

我们以国内第一家互联网民营银行微众银行为例，了解人工智能是如何赋能营销的。我们都知道，金融产品属于高价值、低频次的消费品，所以为了提高业绩，金融行业需要通过营销不断宣传自己的金融产品，在维护现有客户的同时还要开发新客户。在过去很长一段时间内，金融行业营销的问题是广告投放效果差。人工智能进驻金融领域后，帮助金融行业实现了精准营销。例如，微众银行利用人工智能模型做到了精准推荐，使投放效果大大提升。首先根据长期以来形成的各种数据，如登录、点击等用户前端行为数据，以及如开户、存款等后端转化行为数据，以实际的转化为优化目标，建立起专属人工智能模型。通过该智能营销模型，微众银行将其多种金融产品的广告精准投放到数十万家小型企业，广告投放效果得到了有效提升，在降低广告投放成本的基础上，实现精准获客。

智能风控

在金融行业，风控是永无止境的，传统金融机构通常采用传统评分卡和规则引擎等“强特征”进行风险评分。评分卡可以通过用户的各种金融记录和历史信用资料，对用户的信用等级打分，根据分数来判定是否向该用户提供贷款、贷款的金额以及利率。规则引擎可以对规则的层次结构有一个综合的考量，以此来实现复杂的决策，比如通过整合金融产品的准入规则、信用评分规则和风险分级规则，来进行金融领域的风险控制。通过上述两种方式进行风险评分存在审核周期长、涉及人员多、侧重人工审核和业务流程繁杂等问题，会导致效率低下，严重影响了用户体验。在风控领域中，人工智能能够有效地弥补传统风控的短板，建立起全面的风控模型，提升风控的效果，优势愈加凸显。

金融智能风控指的是通过全面风险管理模型，准确计量风险并定价。金融智能风控常用于贷款业务场景中，具体步骤如图7-5所示。

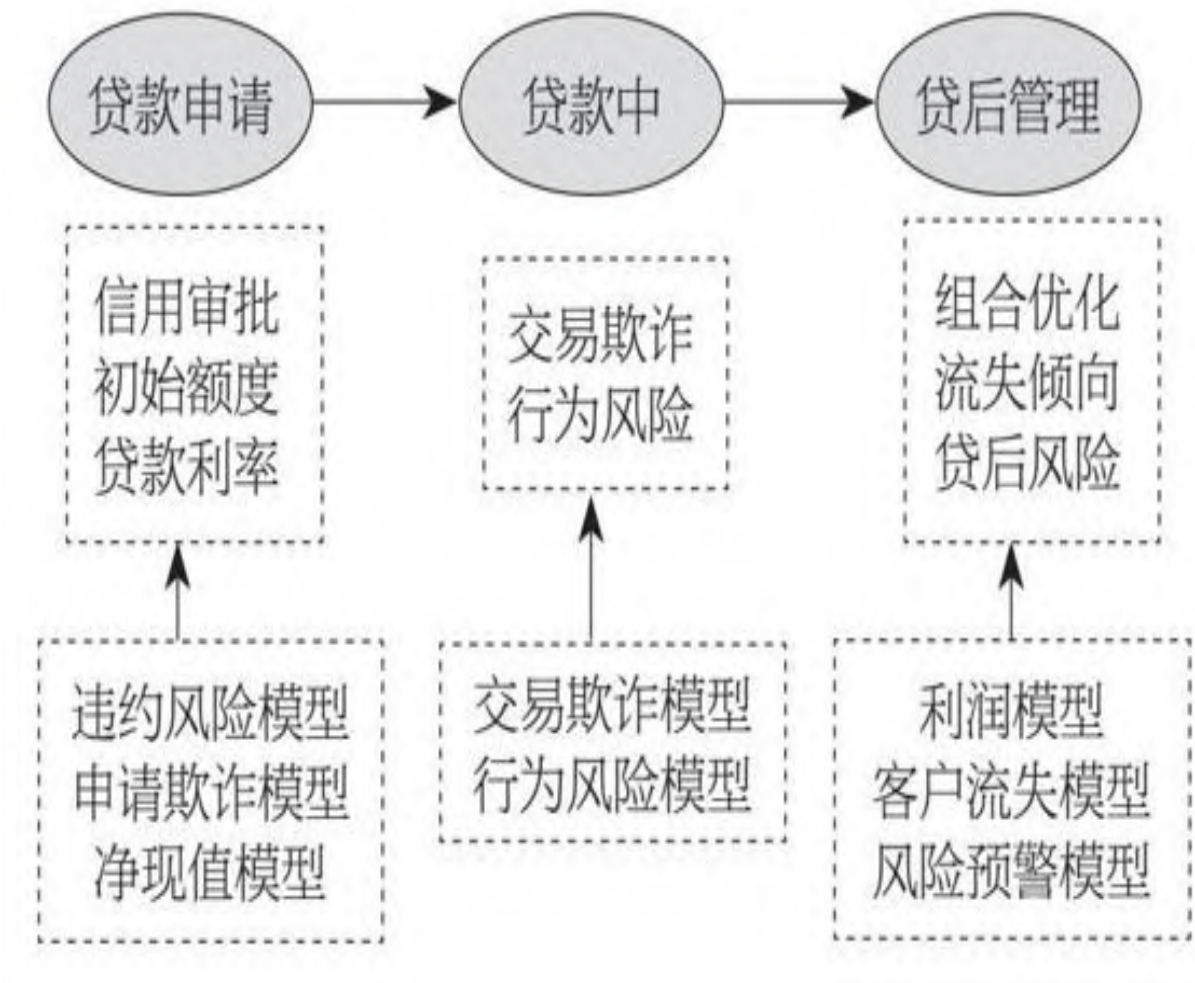


图7-5 金融智能风控模型示意图

第一步，进行贷款申请部分的智能风控。需要在客户数据的基础上，通过违约风险模型、申请欺诈模型和净现值（net present value）模型等，对金融客户进行信用审批、初始额度和贷款利率的确定。

第二步，进行贷款中的智能风控。通过交易欺诈模型和行为风险模型等，可以对金融客户的行为风险及交易欺诈概率做出评定。

第三步，进行贷后管理部分的智能风控。利用利润模型、客户流失模型和风险预警模型，对金融客户的信贷组合进行优化，对其流失倾向和贷后风险做出评估。

人工智能在金融风控领域的应用已经势不可当，但在一些具体场景应用的推广过程中，人工智能仍然面临一些障碍。首先，金融行业本身也在不断发展变革，所以人工智能要在金融领域落地应用，不可避免地要面对银行等机构的模式转变和潜在风险等的重重考量。其次，在传统的金融领域中，开展业务要经过申请、审批和贷后等一系列复杂烦琐的手续和环节，而且整个过程是在不同的操作系统和大量的人力资源协作下完成的，如何通过人工智能进行系统整合以及人工智能如何与各系统进行磨合调整，还需要进行长期的摸索。此外，由于一些金融从业人员意识的滞后性，人工智能在金融风控领域的实践还要经过从业人员意识转变的漫长过程。另外，人工智能模型经常呈现为“黑箱”形式，而金融活动又有可溯源性的要求，这样就形成了一些难以调和的矛盾。智能风控可以根据履约记录、社交行为和行为偏好等多方面的行为特征进行用户风险评估，而人类并不能直接有效地理解人工智能风控的具体过程，但是在某些监管较严格的场景下又要给出必要的解释。

在金融智能风控中，反欺诈和信用评估是非常重要的两个环节。反欺诈指的是对金融业务中可能出现的欺诈行为进行识别，目前已经可以通过人工智能实现金融领域的反欺诈。信用评估是指在金融业务中对用户的信用情况进行评估。当前，我们可以通过基于人工智能的信用分准确刻画贷款申请人的信用风险。

如图7-6，通过人工智能建立的信用分和对应人群的逾期率有这些分布特点：总体上，在各个信用分区间的人数大体呈现正态分布；随着信用分的上升，逾期率逐渐下降。据此可以进行针对性的风控管理，例如蚂蚁金服推出的芝麻信用，就是利用人工智能对用户的个人信用进行综合评分，用于预测用户在信用行为中的违约率。芝麻信用

的分数越高，证明该用户的违约率越低，在申请信用服务时，企业就可以根据用户的分值给予用户免押金、先使用商品后付款等相关权益。

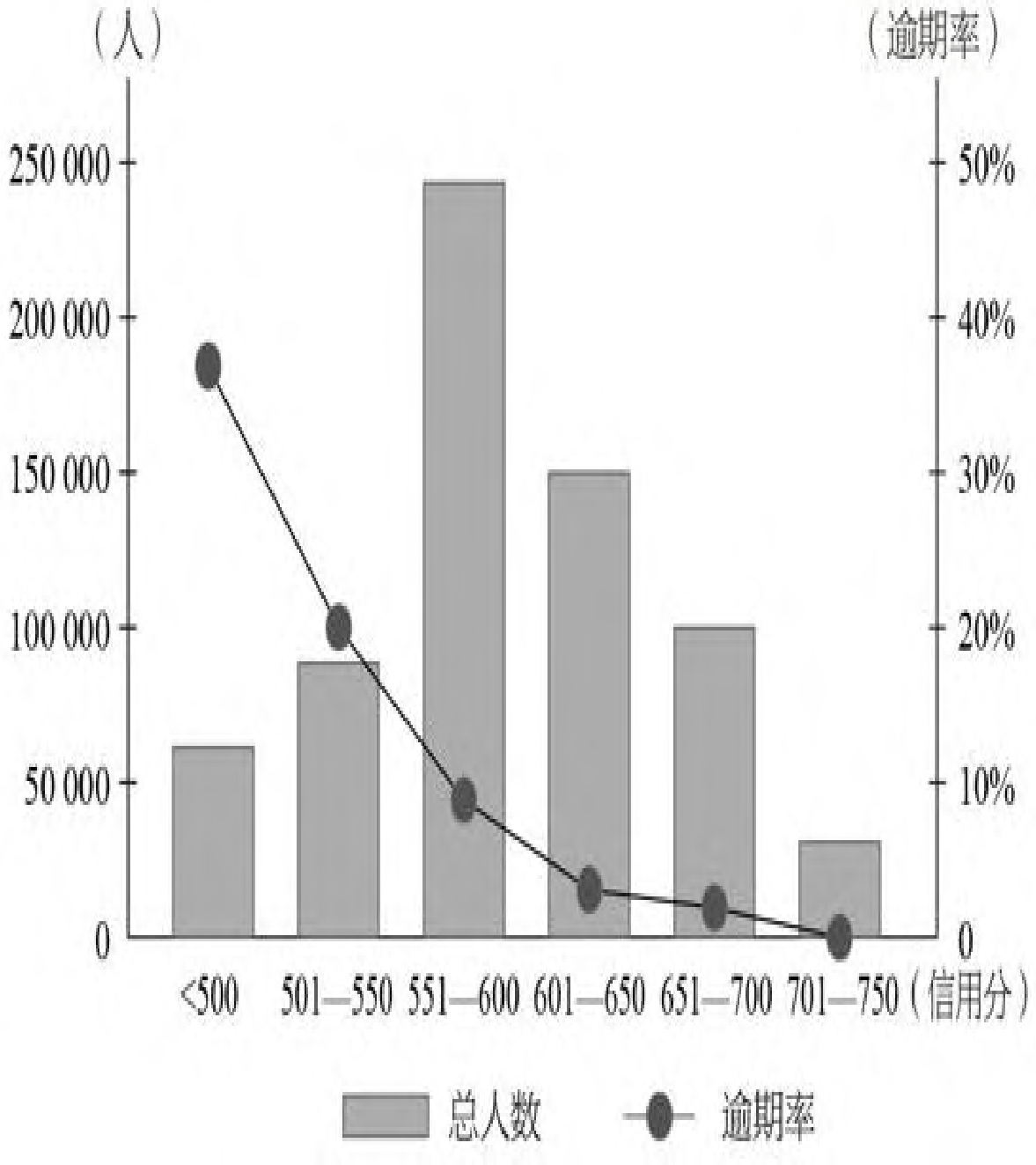


图7-6 信用分和对应人群分布示意图

从上面可以知道，金融智能风控已经有很多手段能够进行风险识别，但人工智能风控模型仍面临很大的挑战。目前，金融大数据特征存在高维稀疏、异构和弱相关等缺陷，信贷样本缺乏且波动性大。这些难题的解决方案，可以通过契合的模型进行机器学习和深度学习而最终得以实现。下面我们就通过一些具体的案例来了解人工智能在金融风控领域的实践情况。

作为大型国有银行，中国工商银行就在积极运用人工智能来进行金融风控。近年来，寻求金融服务的用户越来越多，但随之而来的是层出不穷的金融欺诈行为。为了维护资金安全，降低金融风险，工商银行利用人工智能建立了自己的风控系统。该系统通过人脸识别和声纹识别来确认用户身份，防止不法分子盗用他人账户进行信贷；通过机器学习分析处理海量用户数据，挖掘社会网络关系，避免团伙性金融欺诈；运用神经网络、决策森林等模型，分析用户信息和交易记录，能够实时甄别出风险较大的金融交易。通过这套人工智能风控系统，工商银行实现了实时检测以及风险预警，大大减少了金融交易过程中的风险。

通过人工智能成功进行智能风控的还有美国征信巨头费埃哲（FICO），它通过人工智能建立了一套费埃哲决策引擎，可以为各种金融机构提供服务，银行是其主要用户群，全球最大的100家银行中超过50%都在使用费埃哲提供的服务。费埃哲的智能决策引擎的一个最基本功能就是信用评分，通过基于人工智能的信用评分系统，根据用户的信用记录、欠债额度等数据，对用户的信用状况进行评分，以此来确定向该用户进行信贷的风险。美国90%的大型金融机构都在使用费埃哲的信用评分服务进行风险控制。另外，费埃哲的智能决策引擎还通过社交网络分析建立了反欺诈平台，该平台在为苏格兰皇家银行服务时，仅仅用了4个月就将欺诈率降低了1/2。费埃哲从一家普通的数据分析公司，发展为总市值将近百亿美元的美国征信巨头，人工智能功不可没。

智能投顾

智能投顾作为人工智能应用于金融领域的热点，一直在升温。智能投顾指的是通过人工智能进行投资组合推荐，根据投资者的风险偏好和配置需求为客户推荐个性化投资组合（包括股票、期权、债权和房地产资产等各种配置组合），并提供基于策略的自动交易服务。

智能投顾能够快速、高效地解决客户的投资选择难题，实现的是为客户个性化定制最佳投资组合。智能投顾首先通过人工智能模型充分分析投资者信息，这是个性化配置投资组合的前提，即对客户的财务情况、投资偏好、投资需求、收益目标以及市场表现等进行综合分析，进行客户画像，目前主要采用调查问卷等形式，但以后通过人工智能可以进行更为全面的客户画像，则会减少对客户的打扰。

然后，通过机器学习，根据历史经验和市场信息，综合收益和风险情况，统筹考虑多种资产，对大类资产进行配置，建立分散的投资组合，并在投资组合中进行选择，对资金进行精确配置。随后执行交易服务，并在交易服务过程中，通过自然语言处理等对市场信息、新闻和政策等进行学习，并排除无关事件以及黑天鹅的影响，根据市场和客户情况提供动态资产配置调整和交易策略。

智能投顾最早起源于美国，发展已有10年，管理的资产规模在2019年达到7 497.03亿美元，Betterment和Wealthfront是新兴智能投顾平台的代表。其中，Wealthfront作为一家在线财富管理公司，经过调查问卷评估客户的风险偏好和风险承受能力，再利用人工智能模型分析客户行为、市场和产品等一系列信息，最终形成为客户量身定制的资产投资组合建议，包括股票配置、股票期权操作和债权配置等。该公司运用人工智能开展投顾业务，在一年之内，其资产管理规模扩大了近两倍。

相比传统投资顾问，智能投顾具备五个突出优势。

第一，成本低。传统投顾服务很难触及客户，人力处理信息的成本较高，因此传统投顾机构均为投顾服务设定了一些门槛，主要针对高净值客户。在传统的投顾服务中加上人工智能，可以面向的客户数量就实现了海量增长，并且通过人工智能模型为投资者画像并为其提供资产配置建议，极大节约了人力成本，也降低了客户获取投顾服务的门槛和费用。

第二，效率高。智能投顾的服务过程更加便捷，投资者在平台上可自行操作，人工智能模型自动评估投资者并提供投资配置组合，减少了传统的人工环节，整个流程仅几分钟，就能达到高效、精准的目标。

第三，理性化。智能投顾通过人工智能模型提供投资配置建议，有效地避免了人的主观观念和情感因素的影响，通过对市场数据的有效捕捉，避免了个人投资者一味追涨或者追跌的行为和思维惯性，或者传统财富管理人为了佣金而误导客户的风险，科学、理性地提供投资建议，并根据实际情况动态调整。

第四，分散风险。智能投顾的投资决策和投资组合由人工智能提供，可配置和管理的投资标的种类更多，范围更广，因此分散度更高，可以有效分散风险。例如智能投顾平台Wealthfront涉及多达11项资产类别，包括美股、海外股票、新兴市场股票、股利股票、债券、国债、自然资源和房产等，客户可以自由选择，并为客户提供全球范围内的投资组合。

第五，信息透明。传统投资顾问服务一般存在信息披露的问题，而且有时候投资操作与客户需求不能完全匹配。但在智能投顾中，客户可以随时查看产品范围、投资组合和管理费用等信息，并且智能投顾可以严格执行人工智能模型给出的资产配置建议。客户可以直接接受系统的建议或是在平台对投资组合进行微调，也可以根据个人意愿选择是否接受公司的服务来管理其投资组合。这样的服务对客户更加公开和透明。

目前国内传统投顾业务，也开始进行智能化转型，大多数机构在2015年至2016年期间推出智能投顾产品。招商银行2016年12月发布了智能投顾应用摩羯智投，这是国内银行第一个智能投顾产品，利用机器学习构建以公募基金为基础的、全球资产配置的“智能基金组合配置服务”。在客户选择了理想的投资期限和风险承受等级后，摩羯智投会利用人工智能模型构建相应的基金组合推荐给客户，并由客户进行决策。

与此同时，该平台利用人工智能进行学习建模后，能够实现目标风险确定、组合构建、一键购买、风险预警、调仓提示、一键优化和售后服务报告等功能，打通基金投资的全流程服务环节。例如，该平台可以实现对全球市场信息的实时收集，根据这些信息可以掌握最新的市场状况，这样就可以计算出最优的组合比例。如果客户所持组合偏离最优状态，平台会建议客户调整基金组合达到最优状态，在客户认可后，就可以自主进行一键优化。截至2019年4月，摩羯智投的累计销售规模已经达到122.33亿元，客户超过18万人。

智能投研

智能投研是人工智能在投资研究场景下的一大重点应用，也就是利用人工智能自动化地完成海量金融信息的收集、摘要、归纳、缩简、分析和预测，更加快捷方便地为投资研究人员提供信息，支持他们进行决策，而非直接给出决策结论。

与人工投研相比，智能投研的最大的特点就在于高效。人工投研中，分析师需要花费大量的时间在金融数据终端或者信息平台上收集素材，并需要通过自己的逻辑将这些素材再组织成投资决策，并且将结果以不同方式呈现，实际效率不高。然而智能投研利用人工智能模型可以迅速收集、读取大量的信息，自动化地完成多维度综合计算和分析，发现事件之间的关系，从而做出一定程度的预测，并且按照预

先设定的模板生成研究报告，实现了从信息收集到研究内容呈现的跨越，提升了效率。并且，智能投研中应用的人工智能模型，也可以避免分析师的主观因素影响，使得投研内容更加客观，这在一定程度上也提高了投研报告形成的效率。

Kensho是美国智能投研的先行者，2018年Kensho被标普全球（S&P Global）以5.5亿美元收购，创造了迄今为止华尔街最大规模的人工智能公司收购交易之一。Kensho的定位是数据分析与机器学习公司，长期以来，基于人工智能开发了针对投资者的新一代数据分析工具Warren。它具备高效的数据处理与强大的数据分析能力，能够在数秒内搜索90 000多个全球事件，将数据、信息和决策进行智能整合，并挖掘数据之间的相关性，从而预测资产价格走势和套利机会，并以图表形式呈现其预测效果。Warren能够分析与回答650万个金融问题，预计未来能回答超过1亿个不同类型的金融问题，实现了分析能力上的一个巨大跨越，从而提高了投资者工作效率和投资能力。与此同时，Warren利用人工智能强大的学习能力，以极快的速度读取亿万条数据或信息，在分析数据的过程中实时进行反馈学习，不断优化其分析结果。

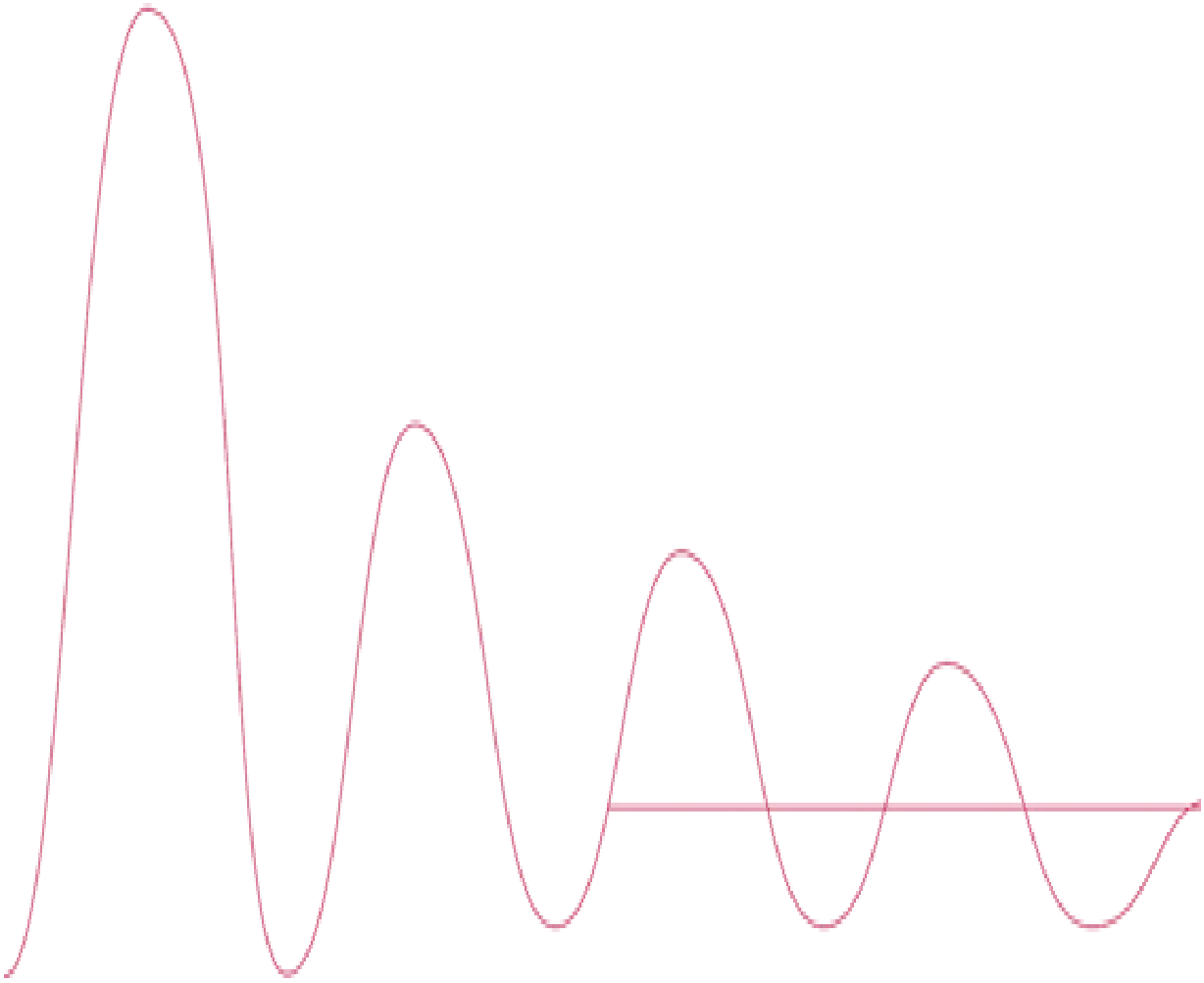
金融是现代经济的核心，金融行业的发展对整个经济起到举足轻重的作用。金融业天生具备数字的理念，在足够多的数据和明晰的业务目标两方面存在极大优势，与人工智能的运作模式非常契合，人工智能可以在金融领域大范围落地应用。从智能营销、智能风控、智能投顾到智能投研等几个业务方向的实际应用可以看出，人工智能基本覆盖了金融业的方方面面，为金融业繁荣而有序地发展赋能。毋庸置疑，金融业已经成为人工智能场景中发展最为迅速的领域之一，人工智能也助力金融业守正创新、安全可控。

随着人工智能的应用越来越成熟，从个人消费贷款到大规模的投资决策，人工智能对各类金融业务的赋能价值不可小觑，金融业拥抱人工智能自然也是顺势而为。人工智能能够帮助金融机构升级换代，

提高金融服务的风险把控能力，也能够突破金融行业壁垒，消除金融产品信息的不对称，真正让金融服务于大众，实现普惠金融的行业理想。

第八章

AI的挑战与对策



凡事总是盛极而衰，重要的是认清趋势转变，要点在于找出转折点。

—乔治·索罗斯—

在之前的内容里，既有高屋建瓴的人工智能原理探讨，更有各行各业的人工智能落地案例。但是，沉舟侧畔千帆过，病树前头万木春，一切新生事物的发展都不会一帆风顺，如同世间万物曲折前行的轨迹一般，AI思维在前进的过程中也免不了有磕磕绊绊，不断地遇到挑战。但凡事都有两面性，在想方设法化解这些挑战的过程中，也相应地促进了人工智能的发展和落地。

“势”反映的是事物在发展过程中，一定会遇到阻碍和波动，但从这些变化之中能体现出事物的发展趋势和方向。在前面的章节中，我们已经对AI思维的道、法、术、器、用的五个方面展开了讲解，所以，在本书的最后一章，我们将从“势”的层面深入探讨人工智能在现阶段所面临的挑战和瓶颈，无论是在技术层面的、法律伦理层面的，还是人为因素层面的挑战，都让我们在认识到一定的局限性的同时，也更理性地看待周遭因为人工智能而正在发生的巨大变化。庆幸的是，到如今，这些挑战已有相对应的解决策略，成功应对这些挑战也能够产生不可估量的积极影响：一方面深化了人工智能的落地，另一方面，也让我们对人工智能的发展更具信心和期待。

AI模型的挑战

人工智能发展至今已经走过了60多个年头，这期间经历了风风雨雨，有高潮也有低谷。每当人工智能的技术和应用取得核心进展时，人们都会对此产生非常大的兴趣，这也吸引了大量的资金投入，此时便是人工智能的高潮期。但另一方面，由于外界对人工智能的宣传报道较为单一，导致大家对人工智能的期望远远超过了目前技术所能达到的高度，而当大家发现花费了大量的资金、投入了大批精英人才却不能达到预期成果的时候，人工智能的低谷期就到了。

通过前几章的讲解你也了解到，人工智能离不开大数据、模型、算力、业务模式，这4点是AI思维成功实现的支柱因素。其中，模型是整个AI思维的核心，也是让数据奏效的强大引擎，如果没有模型，其他的因素都不能够在AI思维体系中起作用，也就无法实现人工智能对各方面的提升和赋能。但同时，模型在现实中的短板也给AI思维的进一步发展和普及带来了制约。那这些制约究竟是什么？我们下面就围绕模型面临的挑战进行探讨。

如何从无标注数据中学习

实现AI思维落地最重要的一环就是建立模型，但是通过机器学习或深度学习训练一个模型的前提是有大量经过人工标注的数据。例如在图像识别里面，通常来说，需要几十万到上百万人工标注的数据。在语音识别里，则需要成千上万个小时的人工标注语音数据。机器翻译则需要数千万的双语句对做训练。包括曾经一战成名的AlphaGo，也是利用围棋高手们数千万步棋谱对局的记录进行模型训练的。

训练模型需要大量的人工进行标注，就意味着这需要花费很多人力和资金成本。很多时候，找专家进行数据标注费用非常昂贵。我们可以粗略评估标注数据的代价会有多高，以机器翻译为例：我们的目标是标注1 000万个句子，如果请人工进行翻译，假设一个句子的标注费用在0.5元，那么要完成这个标注1 000万个句子的目标要花费500万元，这个费用是相当高的。所以，如此大的数据标注成本，对于那些创业公司或者刚刚开始涉足人工智能领域的公司来说，是一件难以完成的事情。

另外，对于一些个性应用场景来说，很难找到大规模的标注数据，比如说疑难杂症的诊断，因为它既然被列入“疑难杂症”的范畴，一方面可能说明这类病症的患者病例较少，另一方面也说明医生

对这类病症的了解并不全面，无论是哪个原因，最终导致的结果都是没有足够广泛的数据用来进行模型训练。

所以，人工智能领域的专家们开始思考，深度学习建模能不能去掉数据标注这个步骤，也就是探索如何从无标注的数据里面进行学习。这成为深度学习研究的一个前沿。就目前的研究工作来看，一条可行的解决路径是通过最近比较火的生成性对抗网络来实现。

首先，我们了解一下什么是生成性对抗网络。生成性对抗网络简称GAN，是非监督式学习的一种方法，即通过让两个神经网络相互博弈的方式进行学习，一个是生成网络，另一个是判别网络。生成网络从潜在空间中随机取样作为输入，如图8-1所示，将噪声向量发送给生成网络，生成网络将这个噪声向量转换为虚拟数据，其输出结果需要尽量模仿训练集中的真实样本，再将虚拟数据发送到判别网络进行分类。而判别网络的输入则为真实样本和生成网络的输出结果，其主要的工作是将生成网络的输出与真实样本分辨开来。两个网络相互对抗、不断调整参数，最终达到生成网络的输出结果与真实样本无二。

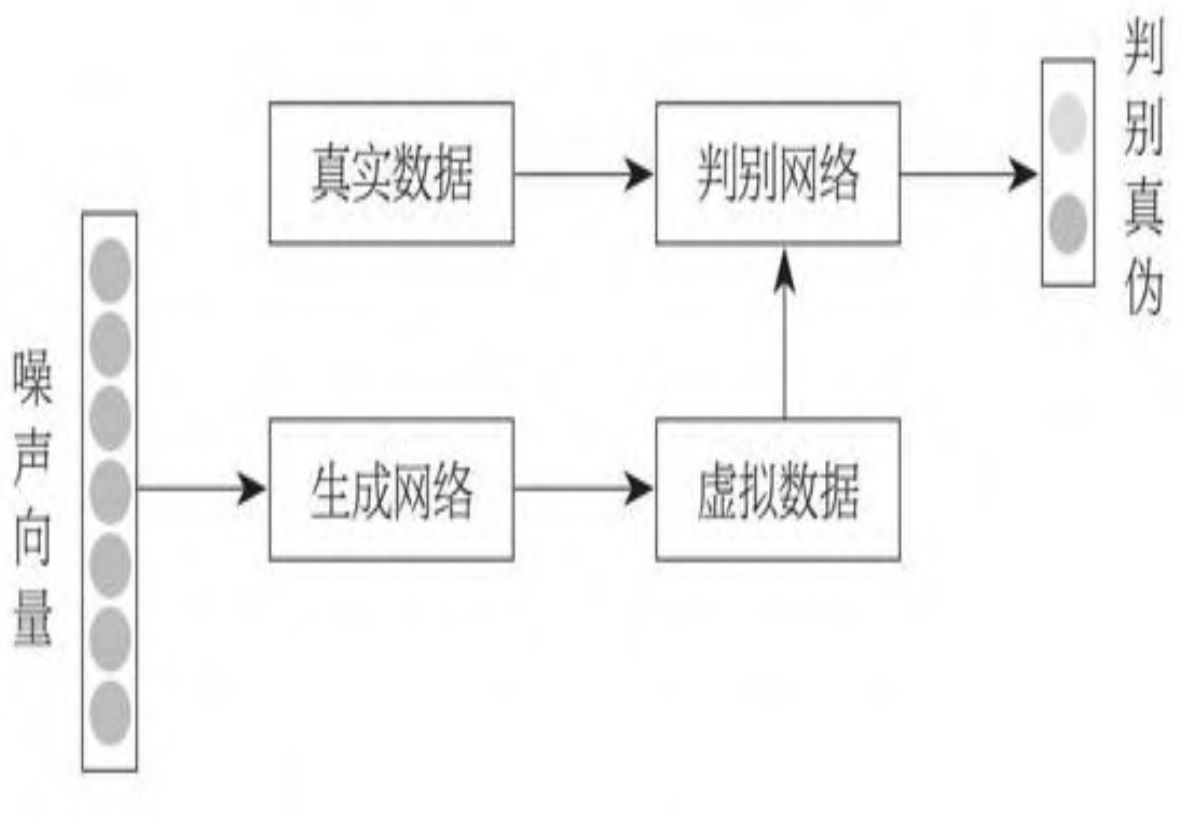


图8-1 GAN网络架构示意图

通俗来说，GAN的工作原理类似于这样的场景：一个男生试图拍出摄影师级别的照片，而一个女生以挑剔的眼光找出照片的瑕疵。这个过程是男生先拍出一些照片，然后女生分辨出男生拍的照片与摄影师级别的照片的区别。男生再根据反馈改进自己的拍摄技术和方法，拍出一些新的照片，女生再对这些新照片继续提出修改意见，直到达到均衡状态——女生再无法分辨出两者的区别。

所以，GAN能够从多个角度学习到海量无标注数据的特性，从而使得从无标注数据中挑选出来的可用数据与人工标注数据之间存在很强的互补特性。这就类似于，以往的模型没有规律判别功能，需要标注员将输入数据打上标签后，模型才开始进行学习；而通过GAN，机器能够自发学习输入数据的规律，自动对新的数据进行判别和分类，从而实现无标注数据的学习。

从无标注的数据进行学习是非常重要的一个问题，它对人工智能在各领域的实际应用也起到了相当大的作用。过去的标注数据十分有限，这导致我们无法有效地运用深度学习，从而延缓了人工智能落地的步伐。而如今，如果我们能够跳过数据标注这个步骤，从无标注的数据进行学习，那么越来越多的实际应用场景都能够嵌入深度学习，很多问题也就迎刃而解了。

如何把数据和知识结合起来

近年来，深度学习的发展取得了巨大的进步，甚至改变了整个人工智能领域的发展方向。深度学习通过在海量的数据上构建大规模的神经网络，这些神经网络经过训练后可以在很多领域有很出色的表现，如语音识别、图像识别和自然语言处理等领域。深度学习的特点是不再需要人为建立结构化的知识表示体系，知识经过海量信息训练后直接融合在网络中，这使得经典的知识表示和逻辑推理等思想在人工智能领域的作用逐步弱化。

然而，深度学习终究不能解决人工智能领域的所有问题，学习产生了模型也不代表模型能够真正理解所处理的数据，模型也无法进行深层的推理和思考。实际上，基于知识的规则推理系统能够动态学习规则，而且能在多项规则的关联作用下进行复杂的推理，挖掘信息间的深层联系，这是深度学习目前难以做到的。虽然基于知识体系的方法由于目前难以构建大规模的知识系统，限制了其在实际场景中达到理想的效果，但如何把数据和知识结合起来仍是值得深入研究的问题。

知识图谱可以将数据和知识有效地结合在一起。那什么是知识图谱呢？可能在外界看来，知识图谱是在2012年的时候由谷歌公司提出来的一个新概念，但事实上却并非如此，知识图谱这个概念其实由来已久。从20世纪70年代出现的“专家系统”，到万维网之父蒂姆·伯

纳斯-李（Tim Berners-Lee）提出的“语义网”和“链接数据”，都是知识图谱的前身。知识图谱主要研究的是客观世界各种概念之间的关系，近年来发展迅速，产生了在智能医疗和自适应教育等领域的落地应用。

从本质上来看，知识图谱是语义网络的知识库；从实际应用的角度来看，可以将知识图谱简化理解成多关系图。我们通常用图里的节点来代表实体，用图里的边来代表两个节点间的关系。实体指的是现实世界中的事物，比如人、地址或机构等，关系则表示的是不同实体之间的某种联系，例如，人-站在-高台，“人”和“高台”是实体，而“站在”则代表了人和高台这两个实体间的关系。在不同的图里，节点和边也会不同。如图8-2所示，一般的图是由节点和边构成，但这些图通常只包含一种类型的节点和边，而多关系图则包含多种类型的节点和多种类型的边。图中，不同类型的节点和边用不同的颜色和样式来进行标记和区分。

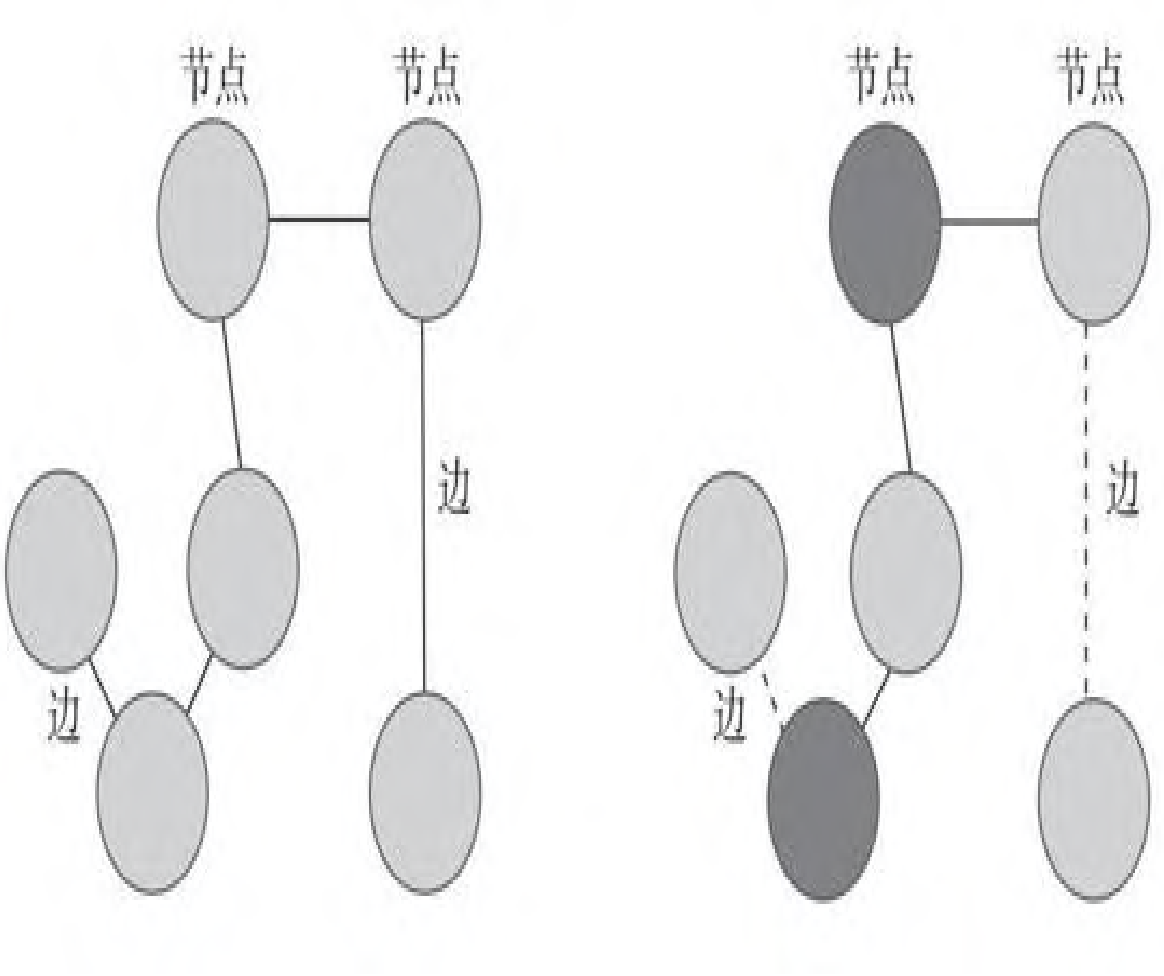


图8-2 一般的图与多关系图对比示意图

在现实世界中，很多场景都非常适合用知识图谱来表达。以社交网络图谱举例，如图8-3所示，社交网络里包含“人”“学校”这样的实体。人和人、人和学校之间存在多种可能的关系，例如甲和乙是“同学”的关系。乙和A学校是“现就读”的关系，乙和B学校是“曾就读”的关系。丙和丁是“同学”的关系，丙和A学校是“现就读”的关系。如果继续拓展下去，每个人都有许多同学，同学们又来自不同的学校。这张社交网络图谱就包含了很多信息，又把这些信息通过节点和边的形式表现得非常清楚。

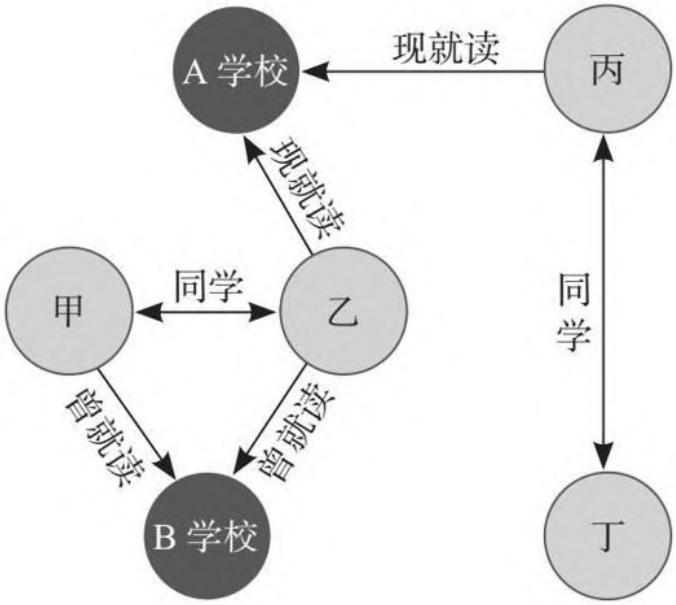


图8-3 社交网络图谱示意图

只有当知识图谱构建完成之后，才能应用到实际的业务之中。一个构建完成的知识图谱相当于是一个知识库，这就解释了为什么它可以回答用户搜索的问题。比如在百度搜索引擎里输入“阿里巴巴的创始人是谁”，我们直接就能得到答案——“马云”。这是因为在系统层面上已经构建完成了知识图谱，在这个知识图谱中包含了关于“阿里巴巴”和“马云”的实体以及两者之间关系的知识库。所以，当我们进行搜索的时候，就可以通过关键词提取类似“阿里巴巴”“马云”“创始人”的词汇以及这些词汇在知识库里的匹配，获得最终答案。这种搜索方式与传统的搜索引擎是迥然不同的，如果用户在传统的搜索引擎上进行检索，返回的是网页，还需要用户自己来筛选和过滤信息，然后才能获得最终答案。

知识图谱的出现是为了理清不同概念之间的关系，所涉及的数据量通常较大，所以知识图谱的构建是一项错综复杂的巨大工程。从整体来看，我们将知识图谱的构建分为知识获取、知识融合、知识验证和知识应用这四个步骤。

第一步是知识获取，主要是从不同来源和结构的数据中获取知识。那么，这些知识是从哪里获取的呢？垂直领域的知识图谱数据来源主要有两种渠道：一方面是业务本身的数据，这指的是以结构化的方式存储的公司内部数据库表，只需要简单预处理就可以进行输入；另一方面是互联网上公开的数据，这类非结构化的数据通常包含了文本、图像、音频和视频等信息，一般需要借助人工智能提取出结构化信息之后才能进行输入。

第二步是知识融合，各个来源获取的信息往往是分散和异构的，具有冗余、高噪音、不确定、非完备的特点，仅仅依靠数据清洗并不能解决这些问题，而是需要进行融合和后续的验证工作，将不同来源不同结构的数据融合成统一的知识图谱，以保证知识的一致性。这步的主要工作是对获取的数据进行实体识别和关系抽取，然后融合构建数据之间的关联，包括实体对齐、属性对齐、冲突消解和规范化等。

第三步是知识验证，即对数据进行补全、纠错、更新，一方面，可以确保知识图谱的一致性和准确性；另一方面，知识图谱的构建是持续动态的，当引入新知识后，需要对新知识进行验证。如果新知识与旧知识间有冲突，那么就需要判断：是原有的知识错了，还是新的知识有问题？如果新知识是正确的，那么就要进行相关实体和关系的更新。

第四步是知识应用，主要是根据知识图谱提供的信息，推理得到更多隐含的知识、隐含的关系、隐含的社区和知识间关联的路径等。这样，通过知识图谱可以产生大量的智能应用，如专家系统、推荐系统、语义搜索和智能问答等。其中，基于知识图谱的智能问答系统可以提供用户所需的信息与知识，一般是以一问一答的形式来精准回答用户的问题。比如说，电商领域的智能问答系统可以辅助人工客服回答一些常见的问题。这样既能在第一时间对用户的问题做出反馈，提升用户体验，又能减轻人工客服的工作量，减少人力成本。

可解释的AI模型

在过去的十几年里，人工智能在各个领域都有成功落地的案例，像我们之前谈到的金融领域的农业银行、零售领域的塔吉特、工业领域的杜克能源等，都通过人工智能提升了自己的效益。但在实践的过程中我们发现，现在仍然有很多企业对人工智能持一种怀疑的态度：人工智能系统做出的决策是否真的值得信任呢？这是我们在推广人工智能应用过程中经常遇到的质疑，也是人工智能在企业落地需要攻克的下一个难关。

在现实的业务中，如果人们要做一个决策，是需要在企业中达成共识的，而相关的人员也要通过定期的审查，来解释他们的决策。这样，既能够保证所做决策可以有效地提高整体业务水平，也能够保证决策在整个企业中顺利进行。不难理解，在决策制定过程中，信任机制中存在着一个关键因素：如何为自己的想法或者做法阐述出一个合理的解释？

人工智能也一样，模型的有效性同样会受到当前无法向人们解释其决策的限制。当人工智能根据模型和分析做出决策后，企业会有各种疑问和担忧，比如说“为什么要这样做？”，“什么时候能够实现？”。但这时人工智能并不能人性化地给出回答。这就催生了人工智能的新兴分支，称为“可解释的人工智能”（XAI），即建立一套新的或改进的机器学习机制，生成可解释的模型，结合有效的解释方法，使得用户能够理解、一定程度地信任并有效地管理人工智能。

XAI的关键在于：解释人工智能所做出的每一个决策背后的逻辑，都是合理的、可追踪的和可理解的。这里包含三个关键维度：首先，是合理的，即能够理解每个人工智能做出预测的背后推理，整个决策过程中人工智能的思考流程和逻辑是如何的。其次，是可追踪的，即从数据的特点到模型的逻辑，追踪预测过程的能力。最后，是可理解的，即完全理解做出决策所基于的人工智能模型，包括为什么要选择

这样一个模型，这个模型如何建立起来的以及能够实现的结果是什么等。

在了解了XAI的基本原则后，具体该如何操作呢？XAI是如何对人工智能模型进行解释的呢？其实XAI解释人工智能模型的方法有很多，比如说，在模型建立前对数据进行解读。我们都知道人工智能模型建立在大量历史数据的基础上，所以理解数据对于人工智能模型的可解释性具有重大意义。对数据的理解可以通过一些数据预处理和数据可视化的方法来实现，对其进行解读的关键在于全面地展示数据分布的特点。这一方法可以使建模过程中可能面临的问题都充分地暴露出来，从而帮助我们选择一种最合理的模型来获得可能达到的最优解。再比如，建立本身就可以被解释的人工智能模型。不同的人工智能模型的可解释性是不一样的。在实践中，如果我们选择本身就易解释的人工智能模型，就从根本上解决了问题。常见易解释的人工智能模型有决策树模型、线性分类模型和线性回归模型等。

针对比较复杂的人工智能模型，我们以卷积神经网络为例来具体来说明XAI是如何对深度学习模型进行解释的。曾有研究者利用反卷积的相关思想，实现特征可视化，来帮助我们理解卷积神经网络的每一层究竟学到了什么东西。我们已经知道卷积神经网络的识别过程是由卷积、池化等步骤组成的，而如果想要深入了解一个卷积神经网络的逻辑，我们就需要经过反池化、反卷积这样的逆过程。通过这样的逆过程，我们就可以比较清晰地了解到卷积神经网络是如何进行分类和识别的，也就可以实现人工智能模型的合理性、可追踪性和可理解性。

XAI除了可以解决模型的可解释问题之外，还具有更深层的裨益，例如：一方面，因为XAI提供的解释有助于找到数据和特征分析中的问题，所以XAI能够帮助改善人工智能模型的性能。当人工智能模型的判断逻辑易于解释时，我们就能够发现整个过程中的缺失，从而找到改良它的方法。

另一方面，XAI可以帮助人们确认人工智能判断的合理性，从而更快地部署基于人工智能的方案。决策不光是人工智能的核心，也是人类活动的关键，在进行重大决策的时候，我们既需要确保所做的决策是合情合理的，又需要知道它背后的理由，才能评估它是否值得参考。XAI对于模型的解释提高了人们对人工智能的信心，帮助他们更加明智地决策以及果断地行动。

虽然人工智能模型面临的挑战很多，但所幸最终都有了相对应的解决方案。这其中包括：对于如何在不进行成本巨大的数据标注工作的情况下，还能实现深度学习的问题，我们可以通过生成性对抗网络进行自主学习来解决；针对如何能够把数据和知识有效地结合起来的问题，我们可以通过构建知识图谱来解决；关于如何更好地解释人工智能模型的问题，我们讨论了可解释的人工智能。

无论是哪个挑战，业内都已经有了相关的研究结果，并且实证有效。但这些挑战都值得更进一步地进行研究，找到更好的解决方式，最终使人工智能可以更好地为人所用。而且随着人工智能更长远的发展以及它的深度运用，以后还会有更多复杂的问题出现，就像长夜漫漫，凛冬将至，高峰之下必有低谷。但是我们知道，冬天到了，春天也就不再遥远，人工智能也在不断的挑战中日趋完善。

AI的伦理和法律挑战

虽然人工智能还处在发展的初期，但它已经展现出了巨大的颠覆力，不仅在一定程度上方便了我们的日常生活，也对传统行业模式产生了一定的改变，促进了智能化经济的发展，社会正经历着从“互联网+ ”到“人工智能+ ”的转变。在我们享受人工智能红利的同时，也会感受到技术带来的危险，包括隐私和数据泄露的问题、机器与人

的关系问题等，就像狄更斯所言：“这是最好的时代，也是最坏的时代。”随着人工智能的进一步普及，它所带来的伦理和法律问题也将越来越突出。积极关注并回应人工智能带来的伦理和法律挑战，能够促进民生福祉改善，推进行业健康发展，在新一轮技术革命中掌握主动权。

AI的伦理挑战和应对

人工智能模拟的是人脑，但它又不同于人脑。人工智能在学习和决策的时候，虽然缺乏人脑的主动和细致，但也有其自身的优势，比如它可以在短时间内学习大量的数据，工作速度高于人类，长时间工作不知疲惫，并且可以做出最优决策等。人工智能促使社会发生着新的改变，形成新的行业和产业形态，也由此创造出巨大的经济价值和社会财富。人类在享受人工智能带来的便利的同时，也会产生担忧，一旦机器有了思维，那机器能够变成人吗？人类会被人工智能取代吗？我们在利用机器解放自身的同时，也会产生危机感，一方面我们要利用工具，另一方面我们又害怕被工具取代，丧失了人类对社会的控制。

控制论之父维纳在他的《人有人的用处》一书中曾做出了一个耸人听闻的结论：“这些机器的趋势是要在所有层面上取代人类，而非只是用机器能源和力量取代人类的能源和力量。很显然，这种新的取代将对我们的生活产生深远影响。”在AlphaGo打败围棋冠军李世石时，我们已经看到了人工智能的凌厉之处。当今很多电影大片和文学作品里面，也出现了机器人或者机器智能体超越人类的话题。关于机器人是否会背叛人类、人类是否会成为机器的仆人等话题的讨论此起彼伏，热度不减，甚至霍金都曾多次表达他对“人工智能可能导致人类毁灭”的担忧。硅谷“钢铁侠”埃隆·马斯克把车祸、飞机失事、毒品泛滥、食品问题和人工智能进行了对比，认为前四项东西至少没

有对人类文明造成根本威胁，但是人工智能却从根本上威胁到了人类文明。

人类区别于其他动物最根本的属性在于长期进化而来的高级逻辑，人工智能与人类思维存在相似之处，人工智能的目的是让机器能够像人类一样思考，机器的自主性越来越强，所以人工智能的应用势必会带来伦理挑战。我们不禁要问，人工智能究竟能否产生自己的意识？人工智能真的会挑战人作为主体的根本地位吗？但从人工智能现在的发展情况来看，要对人工智能是否真的会导致人类毁灭这样的问题下结论还为时过早。当前我们所说的人工智能并不是我们脑海中的具有自由意志和行为能力的智能体，更多的是从数据中学习知识的工具，或者可以将人类从烦琐、枯燥、重复的工作中解放出来，或者可以将优秀的经验不受时空限制地应用。不过人工智能的出现确实带来了伦理方面的挑战。

人工智能伦理挑战中最主要的问题就是人工智能与人类的关系问题。这一伦理问题的根源还是在于人工智能和人类思维有相似之处，人工智能模拟了类似于人的神经元网络，可以像人脑一样进行学习，并做出思考。如果可以将人类接触的各种信息和经验转化成数字和符号，再加上强大的数据分析能力，那我们就可以完整地模拟出人脑的功能，机器与人类的差别就会越来越小。所以我们一直以来担心的机器人对人类的挑战，其实是人工智能对人类思维逻辑的挑战，也就是具有了思维的机器对人类主体地位的挑战。

2016年，欧盟委员会下的法律事务委员会就曾提交动议，要求赋予最先进的自动化机器人“电子人”的身份定位，除给予“特定的权利和义务”外，还建议对智能自动化机器人进行登记，以便为其注册进行纳税、缴费和领取养老金的账号。如果该项法律动议获得通过，无疑会动摇传统的法律体系，法律上主体和客体之间这种不可逾越的鸿沟现在正发生动摇。这在我们看来无疑是荒谬的，一个由人类制造出来的机器，竟然在法律上要享有和人类同等的权利。

在机器的身份定位发生动摇的同时，也出现过机器对人造成伤害的事件。比如说，20世纪80年代，日本就曾经发生过在作业现场工业机器将工人置入机器压死的事件。2007年，因为医疗外科手术机器人对病人造成烧伤、切割伤以及感染的事件频发，美国食品药品监督管理局收到超过200份投诉。那么这些事件，是机器人创造者的技术漏洞，还是机器人管理者的疏忽？机器人本身需要对侵害人的行为负责吗？再比如在最近讨论很热的无人驾驶问题中，“人工智能 + 无人驾驶”颠覆了以往的人车关系和车车关系，驾驶汽车的不再是人，而是智能驾驶系统，那么在出现交通事故时，该如何判断机器系统是主观上的“故意”还是“过失”呢？可以让智能驾驶系统承担交通事故中的民事赔偿责任吗？

智能机器人的本质是机器，但它也有人的属性。虽然在目前技术的发展水平下，人工智能代替人类听起来像是天方夜谭，并且人工智能本身的发展距离人类智慧也很遥远，但是未雨绸缪才能防患于未然，我们必须提前讨论人工智能的伦理问题。与此同时，人工智能相关法律的制定也需要考虑伦理的范畴，对人工智能的伦理的研究也早于法律规定。

随着对于人工智能伦理问题探讨的深入，越来越多的国家和地区开始重视对人工智能伦理的研究。例如，欧洲机器人研究网络（EURON）制定了《机器人伦理学路线图》，韩国工商能源部发布了《机器人伦理宪章》，日本组织专家团队起草了《下一代机器人安全问题指引方针》，美国国家科学基金会和美国航天局为对“机器人伦理学”进行研究设立了专项基金。此外，许多机构也在强化人工智能专家在人工智能伦理规范方面的责任，例如谷歌设立了“人工智能研究伦理委员会”。在我国，2017年7月，国务院发布了《新一代人工智能发展规划》，将发展人工智能确定为国家战略的同时，提出“制定促进人工智能发展的法律法规和伦理规范”，“加强人工智能相关法律、伦理和社会问题研究，建立保障人工智能健康发展的法律法规和伦理道德框架”。对人工智能伦理的讨论，不仅可以在一定程度上指

导行业健康、可持续的发展，也可以在必要的时候，转化为法律规范，实现道德的法律化。

2017年1月，主题为“有益的人工智能”（Benefi cial AI）的会议在美国加利福尼亚州的阿西洛马召开。会议提出了“阿西洛马人工智能原则”，明确人工智能研究的目标不应该是建立无向的智能，而应该是创造有益的智能。并且该会议还呼吁，在发展人工智能的过程中全世界都应该严格遵守这些原则，共同保障人类未来的利益和安全。

2017年12月，电气和电子工程师学会（IEEE）发布《人工智能设计的伦理准则（第二版）》，进一步完善了人工智能伦理的内涵，包括一般原则、价值嵌入、研究和设计的伦理方法、通用人工智能（AGI，artifi cial general intelligence）和超级人工智能（ASI，artifi cial super intelligence）的安全、个人数据、自主武器、经济和人道主义问题、法律、情感计算、教育和知悉的政策、经典伦理问题、混合现实、福祉指标这13个方面。

微软在2018年出版了《未来计算》（*The Future Computed*）一书，其中提出了人工智能开发的六大原则：公平、可靠和安全、隐私和保障、包容、透明、责任。清华大学人工智能与安全项目组提出了六条准则：福祉原则、安全原则、共享原则、和平原则、法治原则、合作原则。这些准则之间有相似之处，对人工智能的伦理问题，可以说达成了基本的共识。

欧盟委员会指出，“可信赖的人工智能”应该有三个必不可少的要素：一是合法，应该遵守所有适用的法律法规；二是合乎伦理，应该尊重伦理和价值观；三是安全可靠，应该避免因技术不足而造成无意的伤害。欧盟人工智能伦理高级专家组提出了人工智能伦理准则，列出了“可信赖人工智能”应满足以下七个关键条件：人的能动性和监督机制、技术的可靠性和安全性、隐私和数据管理、透明度、包容性、社会和环境福祉、问责机制。国际计算机学会下属的美国公共政

策委员会发布文件，提出了关于算法透明和可责性的七条原则：意识、获取和救济、责任制、解释、数据来源、可审查性、验证和测试。

2019年6月，国家新一代人工智能治理专业委员会发布了我国第一组人工智能治理原则《新一代人工智能治理原则－发展负责任的人工智能》，提出了人工智能治理的框架和行动指南。治理原则中重点强调了“发展负责任的人工智能”这一主题，提出了和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作、敏捷治理这八条原则。

应对人工智能对伦理的挑战，首先要树立的一个原则就是“以人为中心”，以人类的根本利益为中心，我们要解决的不是人工智能和人类平起平坐的甚至超越的问题，我们要解决的是如何利用人工智能来补充人类思维的问题。人类制造工具，而工具让我们走得更远，让人类从繁重的任务中解放出来。人工智能是需要以人的需求为出发点的，保障人类的利益发展和安全，兼顾技术和行业的发展，即人工智能是“为了人和人类的人工智能”，不应该是可以取代人或者对人有害的技术。

“以人为中心”这一原则可以通过“技术+责任”的路径来实现，在技术上不要一味地追求机器的自主性，而是在人工智能和人类思维之间实现对接，并将人类的价值观嵌入人工智能，通过技术标准、产品标准等指导技术在具体场景下的应用和发展，确保技术是安全的、透明的、可追溯的。在责任方面，要明确人工智能的责任主体，同时加强监督，并建立起可归责的责任体系。

腾讯董事会主席马化腾在2018年世界人工智能大会上提出，人工智能应该做到“四可”，即“可知”、“可控”、“可用”和“可靠”（ARCC：Available, Reliable, Comprehensible and Controllable），他在2019年世界人工智能大会上再次重复了这一点。随着人工智能技术更加广泛深入的应用，人工智能伦理指导和人

工智能治理也更加紧迫，这也需要多方面的协同配合，共同探索人工智能的发展方向，实现科技向善。

AI应用中的个人信息保护

随着人工智能的发展，新的产业形态和商业模式不断涌现。智能革命使得新技术推动生产力快速发展，在技术上对数据的获取和运算更加方便，相应地，对法律领域也提出了新的挑战和要求，其中比较突出的问题就体现在个人信息保护方面。

过去，大数据的概念还没有建立，政府作为社会管理者收集公民个人信息的主要用途便是社会管理和社会服务。但随着大数据和智能时代的来临，越来越多的企业，特别是电商网站、品牌公司和广告公司等都建立起了自己的数据库，利用各种方式收集数据，再利用人工智能模型对数据进行分类和分析，进行智能营销。电商可以了解到我们的商品搜索记录和购买偏好，外卖订餐和出行叫车软件可以定位我们的所在位置和每天的生活轨迹。毫不夸张地说，我们整个人都暴露在了互联网之上。现在大型互联网公司的价值，除了在于其开发的产品和服务以外，也在于其掌握的大量数据。未来的竞争是人工智能的竞争，其根本还在于数据战，谁拥有了大量的数据，谁就在市场竞争中占据了主动。所以，越来越多的技术应用到了大数据的收集当中，那么个人信息安全问题也就引起了越来越多的关注。在个人和数据管理者的博弈之中，个人是处于劣势地位的，个人信息保护的问题也会愈演愈烈。

人工智能和数据是分不开的。在大数据的时代，可以说涉及人的数据是最多的，其中包括人本身的身份信息、特征信息和行为信息等。这些个人信息很多涉及人的隐私。个人信息这个词有时候也被叫作个人资料或者个人数据，在我国，目前对个人信息这个概念的内涵和外延并没有一致的解释，但从司法实践来看，对个人信息的认定采

取的是紧缩的态度，也就是将能够确认人身份的信息判定是个人信息，比如姓名、身份证号和电话号码等，而像网站浏览记录等不能直接判断人身份的信息就不能被划在个人信息的范围之内。但是随着技术的发展，人们逐渐能够做到将个人在网络空间上各种零星细碎的信息拼凑出完整的、足以反映其人格的关键信息。所以，目前对个人信息的认定也出现了放宽的趋势，即将人的特征和行为等间接的个人数据也划在个人信息的范围之内。

人是商业活动的主体，涉及人的数据是可以进行经济价值转化的，这正是大数据革命的核心，并且也会推动人工智能的应用和发展。一方面，我们要保护个人信息，不能非法采集和使用，另一方面，我们也不能让个人数据与市场活动绝缘，人为地阻断数据的经济价值，这不利于市场经济和技术的发展。并且，个人信息和个人隐私具有不同的性质和功能。个人隐私的要义在于保密，即不暴露，若有侵害则事后救济。而个人信息其实是一些已经是公之于众的，并且具有一些权利和财产的属性的信息，其权利的重点在于控制与利用。所以在人工智能时代，如果将个人信息进行了合法、合理的采集和使用，可以做到个人和商业的双赢。对个人信息的保护不仅仅是对法律规则的挑战，也包含经济因素和社会利益的考量，是一个价值衡量的问题。

针对个人信息保护，目前世界上已经有超过100个国家或地区制定了专门的个人信息保护法，尤其是西方的发达国家。在美国，个人信息保护是在隐私保护的范畴内，以行业自律为主，通过行业分散的规范和事后救济来实现。欧盟则是通过立法，凭借法律的强制约束力对个人信息给予相应的保护，从互联网行业对个人信息的获取途径、方式、运用和目的等各方面进行规范。我国的个人信息保护还在起步阶段，各方多次呼吁要制定专门的个人信息保护法，确立个人信息权，但具体法律法规仍未出台。

目前我国对个人信息保护的法律规定并不完善。《民法通则》仅从原则上规定了要对公民的个人信息进行保护。2017年6月1日正式施行的我国首部网络安全的专门性立法《网络安全法》规定了网络运营商收集、使用公民个人信息的目的必须明确，而且应当获得个人信息主体的同意，也就是说目前我国对于个人信息采集的做法是“目的+知情”，收集公民个人信息的目的需要是合法、明确、正当且必要的，并且其目的必须在信息收集时已经明确，事后不能有冲突或者不兼容。个人信息收集的目的、方式和范围必须是透明的且需要征得信息提供者同意。信息收集者需要履行披露义务，收集后的信息在持有阶段必须确保信息的安全，对于个人信息的利用不得违反法律规定。《网络安全法》的重点不仅在于个人信息本身或者是对个人信息的保护，更多的是在于对个人信息的控制者和处理者的规制，包括个人信息的控制者和处理者对公民个人信息的收集、使用、加工和传输等行为。

但是，互联网对个人信息的收集、转化到使用等这些操作常常被用户所忽略。并且目前互联网对用户是普遍免费的，其赢利模式不是向用户收费，而是向第三方收费，也就是通过互联网广告赢利，例如世界最大的社交网站脸书，其收入的98%来自广告的精准投放。所以目前互联网的模式就是“免费+广告+增值”，其第一步就是扩大浏览并且获取个人信息。有些时候如果不注册、不提供个人信息，就无法享受服务，而我们也会以牺牲个人信息为代价，换取网络的服务和便利。所以互联网普遍免费的模式并不是真正的免费，而是以我们的个人信息为交换，供互联网服务商使用。并且，互联网模式对个人信息的保护并不充分，它往往虚化个人信息获取的同意规则，例如将同意规则掩埋在大量的信息中，或者默认选择同意等。2018年新年的支付宝年度账单事件就是一个典型案例。大家在晒支付宝账单和年度关键词的时候，不知不觉又签了一个“服务协议”——支付宝个人年度账单首页有一行特别小的字：“我同意《芝麻服务协议》”，且页面上已帮你选择好了“同意”，默认选项允许支付宝收集用户信息包括在

第三方保存的信息。查看年度账单和《芝麻服务协议》没有关联性，即使选择取消同意，人们依然能够看到年度账单；但如果没注意到，默认同意了这个协议，就意味着允许支付宝收集个人信息。所以，在收集个人信息时，不仅应该要求经营者做到必要的明示提醒义务，而且对于用户而言，也应该审慎为之，不然稍不注意信息就会泄露。

在互联网普遍免费的模式下，有人建议可以在用户选择服务时给用户一个选择的机会，即将选择权交给用户，如果选择免费服务模式，即享有普通的个人信息保护，如果采取付费的模式，则享有一个更高的信息保护层级。对于个人信息的采集和利用可以根据用户个性化的选择和付费情况，采取不同程度的保护力度。但是，这个改变在目前的互联网模式下推行存在难度。还是应该明确信息提供者、信息存储者和使用者以及系统开发者的责任，强化企业保护用户隐私的意识和责任，加强监督管理，防患于未然，并且建立健全完善的争端解决机制，让个人有寻求救济的途径。对于个人信息的保护不是一味地避免个人信息的获取和利用，也不是为了商业目的而牺牲个人权益，而是需要在个人信息的保护和正当的商业活动中寻求一个平衡。只有这样，才能获得双赢的局面。

爱因斯坦曾说过：“技术本身没有道德性，使用技术的人的道德能力决定了技术改变世界的方向和可能性。”技术的发展会产生新的社会关系，或使得现有的社会关系产生变化，影响甚至改变人类的某些生活方式及思维方式。技术的发展促使人们对于自然、社会和人本身以及之间关系认识的不断加深，也促进了法律内容的更新。伦理道德和法律是社会规范最主要的两种存在形式，伦理道德前置，可以激励和规范技术的发展，法律后置，可以抑制技术的负面的影响，提供有效的救济途径。无论是伦理道德还是法律，其规制的意义都在于引导和规范技术健康、可持续的发展，为人类创造福祉。

AI落地的人为因素

前面我们探讨了人工智能模型和伦理法规的挑战，最后，我们论述一下影响人工智能落地的人为因素。人工智能，因人而诞生；AI思维，最终也由人来实现。人工智能与人类息息相关，也因人而异。对于不同的人，人工智能所带来的改变可能完全不一样。同样地，对于人工智能，如果是由不同的人来实现，结果也会完全不同。人工智能是以服务人类为核心，完全围绕人类展开的，所以人为因素在其落地过程中发挥了极其重要的作用。

人工智能时代，人与人工智能关系的话题讨论一直热度不断，从科幻大片里人工智能试图与人类产生感情，到现实生活中人工智能人才的短缺，我们一直在思考、在探索。随着各个行业智能化水平的提高、人类对人工智能的依赖程度的上升，到底有哪些人会与人工智能直接接触？人工智能又赋予了哪些人专业的帮助和思想的灵光呢？下面我们就来具体介绍一下，在人工智能实际运营过程中，人类都涉及了哪些角色以及承担了哪些相应的职责。

AI落地中的角色分工

第一，企业决策者，例如企业的CEO及其他高层管理者。无论在什么时代，企业决策者的思维方式和管理方法，对这个企业的发展和未来都起着举足轻重的作用。他们的目标是用更少的资源产生更高的收益。人工智能作为一个热点方向，会吸引决策者的注意，而有远见的决策者，更是会在人工智能上投入资源。

企业决策者需要在战略上支持人工智能应用。战略的支持包括资源的倾斜、给予足量的时间等，需要按照人工智能的规律来推进项目。一些企业决策者并不了解人工智能的发展规律，执行任何项目都

把时间限定为三个月——无论是开发网站还是研发人工智能都是三个月，这显然违背了人工智能的规律。当然，并不是说研发人工智能一定是慢活，而是说要根据需要以及规律办事，不能一概而论。

第二，业务负责人，即直接负责一项业务的经理。业务负责人的思维方式和工作能力直接影响着所在部门的办事效率和工作业绩。他们没有企业决策者的权限大，但更加了解某个具体的业务，有执行层面的话语权。

业务负责人往往需要达成一定的业务目标，而时刻关注时代发展方向的业务负责人会了解到人工智能的先进性，会更加关注人工智能如何在企业落地，并且相信人工智能有助于实现他的目标。我合作过的优秀的业务负责人都有比较开放的心态，乐于接受人工智能这一新事物并能够与更大的团队一起分享成果。

第三，人工智能科学家，即受过专门训练并精通人工智能模型的专家，是不可多得的人才，人工智能在各个领域的落地实践离不开他们的专业能力。他们不仅需要熟悉人工智能的应用框架，精通并实践人工智能模型，而且要理解人工智能落地过程中需要分析和研究的数据，并具备足够的工程能力来分析和处理所涉及的数据。与此同时，他们也要对数据的业务价值有较深入的认识。所以说人工智能科学家并不等同于实验室里的科学家，他们既要了解数据，也要熟悉业务。

第四，数据工程师，即提取和处理数据的工程人员，他们直接管理、操作人工智能所需要的数据资源。数据工程师负责提供人工智能所需的数据支持，能够有效地满足人工智能科学家提出的数据需求。他们将人工智能与企业数据资源连接，是人工智能能够实现落地的关键一环。数据工程师最重要的特质是可靠、可依赖，不会因为自己管理了数据就偏私，只有这样才能高效地支持人工智能落地。

以军队为例，在一个军队里面，决策者是元帅，组织队伍去攻打堡垒。军队里的业务负责人是司令，带领着一批人马就朝堡垒前进了。人工智能科学家角色就是参谋长，他有着熟练的作战技术，但是

他最终能否按照司令的作战策略率军出征，要看参谋长和司令，也就是人工智能科学家和业务负责人之间的配合程度。优秀的业务负责人和人工智能科学家，也就是司令和参谋长，需要相互磨合和理解，让业务与技术相匹配，发挥出最大的价值。在人工智能落地的过程中，参谋长不能过分许诺，司令也需要给足时间和空间，这样两者才能合作顺利，逐步解决各种问题。在双方合作出现问题时，也会因为彼此都有一个共同的理想，所以坚持走下去，共同实现业务目标。数据工程师在整个军队中的角色，则像参谋长的左膀右臂，没他们的话，参谋长的战斗力会大打折扣，打不了仗。

一直以来都存在这样一个问题，人工智能在企业落地可以没有企业决策者的支持吗？答案当然是否定的。如果缺少了高层的支持，人工智能很容易落地失败。但因此武断地认为只要企业决策者支持人工智能落地，过程就一定会比较顺利，也不完全正确。比如说，刚参加工作的时候，我作为人工智能科学家负责了一个人工智能落地项目。当时，公司领导非常支持这个项目，初步的发展也很理想，但到了项目需要使用大量数据的阶段，数据库的负责人（即前面所说的第四种角色）并不十分配合我的项目，既不支持人工智能所需的大数据提取工作，也不肯协助处理工程上的问题。起初，我由于无法独立处理大规模数据、无法驾驭复杂软件系统等问题，导致了项目的延误，但后来我痛定思痛，通过加强自身的工程能力，灵活构建程序，按照人工智能的需求去处理大量的数据，解决了这一阶段的问题。最终，这个项目成功收尾，人工智能在业务中发挥了巨大作用，直接帮助公司获得一轮巨额的融资。

除此之外，人工智能科学家的培养也是业界比较关注的话题，之前我也针对此问题写过相关文章进行详细的讨论。在我看来，人工智能科学家最重要的一个素质就是对数字的高敏感度。人工智能的开发和应用软件开发极为不同，应用软件开发过程只需要满足需求就可以了，程序和数据的错误可以在后期检验中发现，但人工智能模型里的错误在后期是难以发现的。如果让错误存留在系统里，对业务的负面

影响会很大，所以人工智能科学家需要对数字十分敏感，保证人工智能系统处理的数据是准确无误的。其次，人工智能科学家也需要具备一定的工程能力，能够针对不同的模型快速进行实验。无论业务和专业的侧重点在哪里，对于人工智能科学家来说，基本的工程能力必不可少，只有这样才能保证人工智能成为业务提升的驱动力。

上面我们描述了在人工智能落地中所需的四种角色：企业决策者、业务负责人、人工智能科学家和数据工程师。这四种角色都有各自的特点，相辅相成，互为补充，在人工智能落地的过程中都承担了重要的责任。可以说，在整个人工智能项目中，这四类角色都是功不可没、缺一不可的。

AI落地中的第五个角色——AI产品经理

在现实情况中，实际的业务需求和人工智能模型之间存在巨大的鸿沟，双方没法有效地沟通，而在大多数机构中，也缺少能够有效地将业务需要翻译为人工智能需求的人员，这直接导致人工智能落地进展缓慢。所以，从这种情况来看，仅有刚才说的四种角色的人员组合，并不完备，我们还需要一个人工智能产品经理的角色。

按照通常的定义，产品经理能够将业务的需要翻译为对技术实现的需求。例如，现在要开发一个酒店的网站，产品经理会根据酒店的业务逻辑设计出网站的原型，像介绍页面、登录页面、订单页面等，并能够生成需求文档、对接具体开发网站程序的工程师。这种将业务需要翻译为人工智能需求的职位就是人工智能产品经理，而人工智能产品经理的缺席，会导致人工智能落地进展缓慢。既然说人工智能产品经理是促进人工智能落地的关键，下面就着重分析一下人工智能产品经理的角色。

如果在军队中，人工智能产品经理能够将司令的需要翻译成参谋长可以执行的工作，他们也许就不会有争执了。在抗战时期的战役里

常常有这样的一个场景：参谋长认为自己指挥军队乘胜追击是正确的事情，但在司令看来，这却是莽撞的。优秀的产品经理可以在两者间做好有效沟通，一方面跟司令解释参谋长做的是符合战术逻辑又有助于战争的决定，另一方面跟参谋长也明确每场战役的目标，让他不因鲁莽而犯错。可想而知，如果军队中有这样一个产品经理角色，司令就不用整天大发雷霆来惩罚参谋长了。

然而一直以来，兼具深厚的人工智能专业背景和强大的产品思维能力的人才非常难得。比如说目前大多数新生代的人工智能专业人士都喜欢用深度学习去解决各种场景中的问题，但仅从人工智能模型的角度出发，很多场景问题是难以解决的，反而是传统方法更有效果。所以，合格的人工智能产品经理既要具备AI思维，也要了解行业，还要掌握机器学习和深度学习等人工智能领域的知识，这样多个方面的能力素养叠加起来，才能更好地解决市场所面临的问题，做出实用的人工智能产品。

我时常听闻一些互联网行业的产品经理将自己不懂的知识称为技术内容，这个观点对人工智能产品经理来说是十分不妥的。就像在工业革命的早期，如果你不了解蒸汽机的基本原理，你就不能做火车的产品经理；同样，如果你不具备AI思维，不了解人工智能解决问题的基本思路和框架，也难以胜任人工智能产品经理。出于这样的考虑，我们对人工智能产品经理知识的要求，既不会过多地涉及技术细节，也不会放过对产品形成有决定性的技术内容。对于一个合格的人工智能产品经理来说，我们主要看他是否掌握AI思维的核心要素，是否理解解决实际问题的AI炼金术，是否熟悉人工智能决策引擎的原理，并能够在AI思维的框架下逐步了解重要的人工智能概念。只有具备了人工智能视角、能够以数据和模型的方法论解决复杂问题的产品经理，才可以独当一面、解决业务难题，成为人工智能时代最需要的一类人。

AI产品经理究竟在干什么？

我们前面说了人工智能产品经理的基本定位，那么从更深层的能力需求来看，人工智能产品经理都要做些什么呢？下面我们将产品经理的日常工作分解开来具体阐述。

产品经理是产品开发团队的主导人，需要从用户角度出发，协调研发、市场、销售等部门，对产品进行优化。具体来说，就是了解用户需求，负责产品功能定义、规划和设计，保证产品开发工作顺利开展。除此之外，还要对用户的新需求、竞争产品的信息进行收集和分析，研究产品的发展趋势等。

而人工智能产品经理是产品经理中的一个具体垂直分类，将负责的产品范围聚焦在人工智能领域。作为产品开发团队的主导人，人工智能产品经理的主要职责是寻找用户需求和人工智能能力的交集。例如，人工智能产品经理离用户更近，善于发掘和分析用户以及市场的需求；同时，他们也知道人工智能能够解决什么样的问题，能够根据实际需求设计理想的模型训练集和预测集，同时为模型提供评估指标。那么，相应地，人工智能科学家和数据工程师依据这些需求和指标，获取所需的数据集，开发并迭代出与需求相符的人工智能产品。比较人工智能产品经理和传统产品经理，可以看到二者区别主要在于，是否具有利用对数据和人工智能的理解，在动态业务中迭代产品的思维。

我们要明确的是人工智能产品经理的本质还是产品经理，所以其最核心的工作依然是找到用户需求，专注于提高产品价值。人工智能最终是解决问题的工具，它能够解决什么需求、如何更好地解决需求，都是人工智能产品经理日常需要考虑的问题，这其中最关键的问题就是要考虑如何促进人工智能更好更快地落地。在多数应用场景中，人工智能产品经理需要做的是把行业内的需求转化成人工智能实现层面的问题；在这其中，人工智能产品经理掌握的是数据，需要人

工智能输出的是预测和决策。从人工智能落地的整个流程来看，人工智能产品经理不仅要了解数据的内涵，还要知道如何探索数据。例如，要实现某项业务的人工智能落地，人工智能产品经理要知道所需收集的数据内容及形式是什么、如何正确地理解数据并反馈到实际产品中。在开始进行产品设计时，人工智能产品经理还要基于业务需求确定人工智能产品的效果是提升营销响应，还是优化生产效率等，根据这些定下基础的模型框架。

在人工智能产品领域，数据越具有时效性、越具体，建立起的人工智能模型便越优良，所以在人工智能落地的全过程中，产品的完成并不是项目的终点，人工智能模型还需要在使用中进行动态的迭代和优化，以期能更好地服务用户。对人工智能产品经理来说，在人工智能产品完成后还需要进行持续的数据收集，通过产品的不断使用能够获得更多、更深入的数据进行人工智能模型训练，使得人工智能产品可以越来越准确和深入地解决问题。

同时，人工智能产品本身还具有很大的不确定性，人们无法像传统产品一样直观地了解到人工智能产品到底是什么，对它的智能水平的了解也十分模糊。这样就很容易出现用户对人工智能产品的预期很高，但产品并不能够完全满足用户需求的情况，或者出现用户提了太多超出产品功能的要求却没有得到正向反馈的情况。这些情况的出现会导致用户的心理落差很大，无法客观地评价人工智能产品。因此，人工智能产品经理在设计人工智能产品时，要注意对用户期望的把握。

基于人工智能产品开发的内在逻辑，人工智能产品经理的演化主要体现在三个阶段——第一个阶段，数据收集和管理阶段；第二个阶段，产品开发阶段；第三个阶段，产品落地阶段。在这三个阶段演化的过程中，也需要面对人工智能落地中的各种挑战。下面我们对这三个阶段进行具体分析。

在数据收集和管理阶段，高质量的数据是人工智能实现的基础，这个阶段人工智能产品经理最重要的工作就是进行数据的收集和管理。无论是在实现层面还是市场层面，人工智能产品经理只有先把数据管理做好，才能让数据发挥出价值，创造出真实契合市场需求的人工智能产品。数据管理里有很多具体内容，例如数据标准管理、数据质量管理、数据集成和关联等。此外，在本阶段还有一个重点是数据安全与隐私保护。数据安全是指系统中的数据受到保护，不因偶尔或者恶意的原因而遭到破坏、更改和泄露。数据的运用可能涉及相关主体的隐私问题，所以人工智能产品经理要具有保护数据隐私的意识，对保护数据隐私的措施有一定的敏感性。

人工智能落地的第二个阶段是人工智能产品开发。在这一阶段，人工智能产品经理要创造适用于行业的人工智能产品，从用户需求出发，切实找到行业痛点，通过对人工智能的理解，找到用人工智能解决难题的方式，然后抓住行业痛点和人工智能的结合点设计人工智能产品。真实有效的人工智能产品诞生后，能够实现用户体验和业务效果的大幅度提升，或是人工成本的大幅度降低，甚至能够颠覆行业现状。

人工智能落地的第三个阶段是产品落地，这时人工智能产品经理也应该以类似业务负责人的身份去考虑实际的落地问题。一位十分优秀的人工智能产品经理不仅要引导产品的发展，还要引导业务的发展。他既要是一个优秀的执行者，还要是一个项目和团队的管理者。他需要具备敏锐的洞察力，既能准确抓住用户心理，又能精研产品细节，最终打磨出惊艳的人工智能产品。在人工智能时代，只有把产品做到市场上的极致，不断进行产品突破，才能实现价值。

从整体来看，人工智能产品现在还处于起步的阶段，人工智能产品经理在人工智能产品开发中的经验也都并不丰富。在实践过程中，遇到的很多问题现在也都没有比较确定和成熟的方法论，缺少解决方案或经验总结，很多问题还寄希望于通过人工智能的不断发展和探索

来解决。与此同时，人工智能本身的边界和能力也正在飞速发展中，人工智能产品经理需要及时跟上时代发展的步伐，对自己的工作做出调整。

可能有人会问：“人工智能崛起后，世界还会好吗？”答案是肯定的。既然这个时代已然或者终将是人工智能的时代，我们就更加需要戳破虚高的泡沫，每个人都建立起一个客观理性的AI思维来面对这个新时代，并期待能够以这种AI思维去关照整个社会的发展，感受新生事物的脉搏，让自己过得更好，也能够看到整个世界变得更好。

参考文献

1.新一代人工智能发展规划. 中华人民共和国国务院, 2017.

2. 最好的数据科学家，就是把自己给“干掉”. https://www.yicai.com/news/5433619.html.

3. 人工智能给投融资决策带来的震撼. http://www.sohu.com/a/244997475_99985608.

4.2017-2023年中国大数据应用行业市场全景调查及未来前景预测研究报告. 智研咨询集团, 2017.

5.P. Li, L. Lindsey, M. Januszewski, et al. Automated Reconstruction of a Serial Section EM Drosophila Brain with Flood-Filling Networks and Local Realignment. Microscopy and Microanalysis 25, 2019.

6.视频精细化运营的AI时代，人工智能在芒果TV三大场景的应用. http://www.sohu.com/a/278615402_508153.

7.P. Devries, F. Viegas, M. Wattenberg, et al. Deep learning of aftershock patterns following large earthquakes. Nature 560(7720), 2018.

8.V. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, 2000.

9.腾讯AI“绝悟”KPL击败职业玩家联队，晋升王者荣耀电竞职业水平. http://www.sohu.com/a/331258396_129720.

10. “韩国小姐”的脸到底有多相似？https://www.guokr.com/article/436974.

11.S. Roweis, L. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 290(5500), 2000.

12. X. He, S. Yan, Y. Hu, et al. Face Recognition Using Laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(3), 2005.

13. R. Souvenir, J. Babbs. Learning the Viewpoint Manifold for Action Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008.

14. D. Holden, J. Saito, T. Komura. A Deep Learning Framework for Character Motion Synthesis and Editing. ACM Transactions on Graphics 35(4), 2016.

15. 大数据是如何捧红《纸牌屋》的？https://36kr.com/p/5073252.

16. 今天去买化妆品，接待我的导购是位机器人. http://www.sohu.com/a/195089245_630344.

17. 斯坦福大学用AI预测病人何时死亡，准确率达到90%. https://36kr.com/p/5115081.

18. 数据驱动人工智能的实践. https://www.leiphone.com/news/201702/J8UHOe08IK 9najS6.html.

19. 良品铺子转型启示录：到底什么样的企业需要中台？https://www.leiphone.com/news/201906/6muA7fIMvgJj48VG.html.

20. A. Ziai. Detecting Kissing Scenes in a Database of Hollywood Films. IEEE International Conference on Humanized Computing and Communication, 2019.

21. 人工智能再出新招，AI已能识别小偷！https://juejin.im/post/5ccfa285f265da039f0f2445.

22. Machine Learning Lessons: 5 Companies Share Their Mistakes. https://www.infoworld.com/article/3319665/machine-learning-lessons-5-companies-share-theirmistakes.html.

23. 解决您最棘手的规划和调度问题：业务经理如何使用数学优化技术. https://www.ibm.com/downloads/cas/JLZWPQ40.

24. 运筹OR帷幄，运筹优化的剖析与应用. https://ask.hellobi.com/blog/R_shequ/15549.

25. 帮企业解决运输优化问题，杉数科技推新产品小马驾驾. https://36kr.com/p/5172635.

26. I. Goodfellow, Y. Bengio, A. Courville. Deep Learning. MIT Press, 2016.

27. A. Krizhevsky, I. Sutskever, G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. NIPS, 2012.

28. T. Mikolov, M. Karafi at, L. Burget, et al. Recurrent Neural Network based Language Model. Interspeech, 2010.

29. 付登坡，江敏，任寅姿，孙少忆等. 数据中台：让数据用起来. 机械工业出版社，2020.

30. 手机AI芯片的诸神之争. http://www.ceconline.com/it/ma/8800096402/01/?cf=newCont_past7.

31. 一篇文章带你看遍Google I/O 2019大会. https://blog.csdn.net/guolin_blog/article/details/90049263.

32. Stitch Fix公司CEO：传统零售商想要模仿，是不可能的. http://www.sohu.com/a/232556594_168180.

33. 斯坦福等高校联合研发律政界AlphaGo，战胜20名顶级律师. https://www.yicai.com/news/5402546.html.

34. 世界人工智能产业发展蓝皮书.中国信息通信研究院，2018.

35. 这些垃圾分拣机器人你听说过吗？http://www.sohu.com/a/316298306_231178

36. T. Mitchell. Machine Learning. McGraw Hill, 1997.

37. D. Forsyth, J. Ponce. Computer Vision: A Modern Approach (2nd Edition). Pearson, 2011.

38. D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 2004.

39. L. Ding, A. Yilmaz. Inferring Social Relations from Visual Concepts. IEEE International Conference on Computer Vision, 2011.

40. L. Rabiner, B.-H. Juang. Fundamentals of Speech Recognition. Prentice Hall, 1993.

41. J. Bollen, H. Mao, X. Zeng. Twitter Mood Predicts the Stock Market. Journal of Computational Science 2(1) , 2011.

42. 中华人民共和国2019年国民经济和社会发展统计公报. 国家统计局，2020.

43. Y.Koren, R. Bell, C. Volinsky. Matrix Factorization Techniques for Recommender Systems, IEEE Computer Journal 42(8), 2009.

44. P. Fader, B. Hardie, J. Shang. Customer-Base Analysis in a Discrete-Time Noncontractual Setting. Marketing Science 29(6), 2010.

45. M. Newman, M. Girvan. Finding and Evaluating Community Structure in Networks. Physical Review E 69(2), 2004.

46. 邵蓥侠，冯是聪。社交网络分析在公共安全领域的应用. 大数据3(2), 2017.

47. 如何发现品牌潜客？目标人群优选算法模型及实践解析. https://blog.csdn.net/b0Q8cpra539haFS7/article/details/83005763.

48. 发现奇特新人群，戴森如何用全域营销做长期消费者运营？http://www.sohu.com/a/214785374_356153.

49. 零售企业线下门店智能裂变刷脸支付、机器人导购纷至沓来. http://www.jiqirenku.com/faxian/47179.html.

50. 华菱星马汽车：自动化＋信息化＋智能化 实现生产数据和质量信息可追溯.https://cloud.tencent.com/developer/news/331167.

51. IIoT与工业AI融合，NI为无锡“雪浪制造大脑”打造边缘智算. http://www.eechina.com/thread-546070-1-1.html.

52. 借助科技体验华润五彩城打造智能交融式购物中心 .http://fi nance.sina.com.cn/roll/2019-01-09/doc-ihqhqcis4345936.shtml.

53. 首开热销26亿！未来金茂府成北京楼市“超级硬通货”. http://bj.house.163.com/19/0625/18/EIHM5U14000782A6.html.

54. 华为云“火”在哪. http://ad.163.com/19/0724/16/EKS5EE7V000189DG.html.

55. ET工业大脑助力浙江智能制造，带来数百亿利润空间. https://yq.aliyun.com/articles/73871.

56. 引领首席信息官在2019年部署人工智能的5种方式. http://www.d1net.com/ai/industry/557383.html.

57. 工业4.0——未来生产力与制造业发展前景. 波士顿咨询公司，2016.

58. 阿里推出“八卦盾”主动预警 黄牛和恶意订单一网打尽.http://www.sohu.com/a/273856531_505835.

59. 中国电子商务发展报告2018-2019. 中国电子商务协会，2019.

60. 大数据、个性化？这一次，携程CTO甘泉跟你好好聊聊.http://www.sohu.com/a/138976436_118838

61. 零售巨头塔吉特：数据是最大的资产和挑战. http://blog.ceconlinebbs.com/BLOG_ARTICLE_253477.HTM.

62. 从工具革命到决策革命——通向智能制造的转型之路.http://www.aliresearch.com/Blog/Article/detail/id/21774.html.

63. 关于2018年房屋市政工程生产安全事故和建筑施工安全专项治理行动情况的通报. 中华人民共和国住房和城乡建设部办公厅，2019.

64. 生态环境大数据建设总体方案. 中华人民共和国环境保护部，2016.

65. 2018年中国金融市场发展报告. 中国人民银行，2019.

66. 被AI入侵的金融业——“AI+金融”行业研究报告. 36氪研究院，2018.

67. 打造基于大数据的智能化风控体系. https://www.cebnet.com.cn/20180927/102524642.html.

68. 金融科技公司都在学习的FICO，是一个怎样的存在？.https://36kr.com/p/5093993.

69. 用数据解读摩羯智投. https://blog.csdn.net/fens/article/details/84634574.

70. Kensho和AlphaSense是怎样炼成的. http://blog.memect.cn/?p=2944.

71. I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. Generative Adversarial Nets. NIPS, 2014.

72. M. Zeiler, R. Fergus. Visualizing and Understanding Convolutional Neural Networks.European Conference on Computer Vision, 2014.

73. 刘伟，赵路. 对人工智能若干伦理问题的思考. 科学与社会，2018年1期

74. 智能时代的技术伦理观——重塑数字社会的信任. 腾讯研究院，2019.

75. 吴汉东. 人工智能时代的制度安排与法律规制. 法律科学，2017年5期

76. Asilomar AI Principles. Future of Life Institute, 2017.

77. Ethics Guidelines for Trustworthy AI. European Commission, 2019.

78. 新一代人工智能治理原则——发展负责任的人工智能. 中华人民共和国科技部，2019.

后记

人们在各行各业的长期实践中积累了大量的知识。在这个过程中，数据经过处理成为信息，而信息提炼沉淀为知识。人们从各种渠道获得知识，文明由此而延续。然而面对浩海般的数据，传统的分析方法无论在效果还是效率上都显得力不从心。通过AI思维的支持，我们可以有效地拥抱海量的数据，驱动最佳的决策，从数据中挖掘持续的价值。得益于此，在人工智能时代，每个人的数据都可以转化为知识，使自己和他人受益。恰逢本书定稿之际，人工智能作为新型基础设施建设的方向之一，再次站上风口，这意味着人工智能有机会为未来社会的发展提供全方位的服务，真正能够为人人所用，使人人受益。

然而，人工智能的特点决定了这个领域的创新所带来的价值在时间上是有延迟的，不能够立竿见影，而且人工智能很难像华丽的交互界面那样为人所感知。如何在不确定的预测面前寻找确定性的业务结果，并带来可量化的价值，确实是困扰着机构和人工智能从业者的一大难题。所以，我希望本书能够为即将开展人工智能工作的人们提供一个相互理解与信任的平台，在这个平台中，他们能够了解人工智能的思维框架与运行逻辑，从而更有效地进行规划与交流。

最后，我想感谢在我从事人工智能工作以来，指导和帮助过我的师长、同事和合作伙伴，以及一同在这条路上探索的同人们。本书中许多观点的形成得益于和你们的讨论、交流以及思想碰撞。本书的部分内容源于我给中高层管理人员授课的材料，学员们的反馈坚定了我写作的信念。矫凤涛和刘书颉在本书写作过程中给予了大力协助，在此表示衷心的感谢。

“路漫漫其修远兮，吾将上下而求索。”虽然人工智能时代已经到来，但道路险阻，前途未知。幸运的是，同人们探索和深耕的决心，以及社会各方对人工智能的关注与支持，一直为我在这条路上的前行提供动力。我相信，人工智能的发展，未来可期。

图书在版编目（CIP）数据

AI思维：从数据中创造价值的炼金术 / 丁磊著. --北京：中信出版社，2020. 5
ISBN 978-7-5217-1637-5
Ⅰ. ① A… Ⅱ . ①丁… Ⅲ . ①人工智能－应用－商业模式－研究 Ⅳ. ① F716
中国版本图书馆CIP数据核字（2020）第033811号

AI思维——从数据中创造价值的炼金术
著者：丁磊
出版发行：中信出版集团股份有限公司
（北京市朝阳区惠新东街甲4号富盛大厦2座 邮编100029）

字数：250千字
版次：2020年5月第1版
广告经营许可证：京朝工商广字第8087号
书号：ISBN 978-7-5217-1637-5

版权所有·侵权必究



THE
AI
DELUSION

AI如何通过数据挖掘误导我们

[美]加里·史密斯(Gary Smith) 著
钟欣奕 译

是智能还是服从？

- 大数据、坏数据、假数据充斥着我们的生活
- 以相关关系取代了因果关系

中信出版集团

错觉

——AI如何通过数据挖掘误导我们

[美]加里·史密斯 著

钟欣奕 译

中信出版集团

目录

引言

第1章 智能还是服从

井字游戏

国际跳棋

第2章 盲从

思考之源和思维之火

计算机是超人吗?

将时间考虑在内

识别像素与产生情绪

批判性思维

图灵测试

第3章 无语境的符号

翻译软件与理解语言

威诺格拉德模式挑战赛

计算机能阅读吗?

计算机能写作吗?

在语境中理解事物

猫与花瓶

第4章 坏数据

自我选择偏好

相关系数并非因果关系
时间的力量
幸存者偏差
假数据
识别“坏数据”

第5章 随机性模式

数据挖掘
黑匣子
大数据、大电脑、大麻烦
利益冲突
天生就会被骗
为模型所惑

第6章 如果你拷问数据的时间足够长

孟德尔的豌豆研究
得州神枪手谬误
数据挖掘者
拷问数据
倒摄回忆
金钱启动效应
寻找就会发现
微笑曲线
从卓越降为优秀
攻击性和吸引力
达特茅斯三文鱼研究
骗子，骗子

第7章 无所不包的“厨房水槽法”

预测总统大选

非线性模型

第8章 新瓶装旧酒

逐步回归法

岭回归法

数据规约

神经网络算法

被数学蒙蔽双眼

第9章 先吃两片阿司匹林

明早再给我打电话

我要再喝一杯咖啡

远程治疗

癌症群

最有理有据的疗法失效了

疾病诊断和治疗中的数据挖掘

糟糠过多，精粹不足

第10章 完胜股市（上）

噪声

滑稽的理论

技术分析

抛硬币

《每周华尔街》的十项技术指标

推特，推特

技术大师

为乐趣和盈利投资的黑匣子

第11章 完胜股市（下）

股市与天气

预留方案
真正的数据挖掘
趋同交易
高频交易
底线

第12章 我们都在监视着你

妊娠预测指标
谷歌流感
机器人测试仪
就业申请
招聘广告
贷款申请
汽车保险
社会信用评分
黑匣子式歧视
不合理的搜查
看看你的手环
你需要整容吗？
摆弄系统
共同毁灭原则

结语

参考文献

引言

2008年的民主党总统候选人提名，本应成为希拉里·克林顿的加冕典礼的序幕。作为知名度最高的候选人，希拉里得到了党派内部的最大支持，她同时还拥有最丰富的财政资源。

阿尔·戈尔和约翰·克里这两位大人物虽然也考虑参选，但却对击败核心人物希拉里不抱希望。其他不为人熟知的参选人的机会就更渺茫了：俄亥俄州的美国众议院议员丹尼斯·库西尼奇、新墨西哥州州长比尔·理查森，还有美国参议员约瑟夫·拜登（特拉华州）、约翰·爱德华兹（北卡罗来纳州）、克里斯·多德（康涅狄格州）、迈克·格拉韦尔（阿拉斯加州）和贝拉克·奥巴马（伊利诺伊州）等。

不过，本次提名并没有按照写好的剧本上演。黑人参议员奥巴马虽名不见经传，但却点燃了选民的热情。他筹集了足够多的资金，还说服奥普拉·温弗瑞力挺他参选。在艾奥瓦州预选中，奥巴马以八个百分点的优势击败了希拉里，由此拉开了大选序幕。

奥巴马先是赢得了民主党提名，继而又打败了共和党人约翰·麦凯恩而最终成功当选总统。究其原因，奥巴马竞选主要不是靠口才和魅力，而是靠大数据。

奥巴马竞选团队设法将每名潜在选民及其数百条个人信息录入数据库，包括年龄、性别、婚姻状况、种族、宗教、住址、职业、收入、车辆登记、房屋价值、捐赠历史、杂志订阅、休闲活动、脸书好友，以及所能找到的任何相关情况。

这些数据来自公共数据库、来往电子邮件或竞选工作者的上门询问，还有从私人数据供应商处购买的。而其最主要的来源是每周对数

千名潜在选民进行的电话调查访问，通过这种方式不仅能搜集到个人信息，还可摸清每名选民投票的可能性——是否会给奥巴马投票。

从统计学角度来说，选民投票的可能性与其个人特征相关，还可以根据这些个人特征推测出其他潜在选民。奥巴马竞选所用的计算机软件能预测数据库中每个人投票以及给奥巴马投票的可能性。

这种数据驱动的模型使该竞选团队可通过电子邮件、邮寄信件、上门拜访，以及呼吁捐赠和投票的电视广告来进行微目标锁定（microtarget）。如果计算机程序预测狩猎许可证持有者反对枪支管制立法，那么对这类人的枪支管制宣传就会减少。该软件还推荐了可确保捐赠和投票的其他手段。

2008年1月，奥巴马在这关键的一个月内筹到了3 600万美元，创下了政治家筹款纪录的新高，约为希拉里所筹资金的三倍。获得提名后，奥巴马的筹款额继续上涨。2008年整个竞选活动期间，奥巴马共筹集了7.8亿美元，是对手共和党人约翰·麦凯恩所筹资金的两倍多。麦凯恩根本没有胜算，也确实没能获胜，他只得了173票，而奥巴马却获得了365票。

八年后，希拉里·克林顿再次参加总统大选，决意运用大数据为自己加持。

但这一次，大数据让她大失所望。

希拉里的竞选团队共聘请了60名数学家和统计学家，其中几位曾效力于奥巴马竞选团队。为纪念19世纪的女数学家阿达·洛芙莱斯，他们将自主开发的软件程序称作“阿达”。希拉里要是成为第一位美国女总统，就可以透露“阿达”是她的幕后功臣。故事多么精彩！

他们把“阿达”装在自己的服务器上，只有几个人拥有访问权限。有些人知道这个软件的存在，但不知道它的运作方式，而大多数人对此一无所知。

2016年9月16日，距大选还有7周，埃里克·希格尔在《科学美国人》杂志上发表了一篇题为“希拉里竞选团队如何（几乎肯定）运用大数据”［How Hillary’s Campaign Is（Almost Certainly）Using Big Data］的文章。他指出，“有证据表明，希拉里正在采用可高度精准锁定目标的技术来竞选，奥巴马就曾靠此获胜”。竞选活动开展一年半后，还有观察人士对希拉里的大数据运用继续做出种种推测。这说明“阿达”的保密工作做得非常到位。

希拉里竞选团队对“阿达”的运用守口如瓶，这可能是因为他们不想让希拉里的对手有所察觉，也可能是因为不想加深团队机械行事、谨小慎微和照本宣科的刻板印象，他们毫无伯尼·桑德斯和唐纳德·特朗普竞选团队那样的豪情壮志。

“阿达”每天都模拟运行40万次，为它认为合理的局面预测选举结果。如果佛罗里达州的投票率上升两个百分点，新墨西哥州的投票率下降一个百分点，那会怎么样？如果……会怎么样？然后对答案归纳总结，由此确定应该在哪些地区投入资源，以及投入哪些资源。

例如，70%的竞选预算用于电视广告，这些广告所花的每一分钱几乎都由“阿达”来决定，不寻求或留意资深媒体顾问的建议。“阿达”的数据库有详细的社会经济信息，包括人们在哪座城市看了哪档电视节目，然后推测出这些人给希拉里投票的可能性有多大。“阿达”运用这些数据来计算每张潜在选票的理论成本，然后决定在不同节目、不同时段和不同电视市场上投入的广告经费。

没有人真正了解“阿达”的决策经过，但可以确定这款功能强大的计算机程序能分析超乎想象的海量数据。所以，大家都信任“阿达”。它就像无所不知的女神，只听不问。

我们也不知道“阿达”是如何得出最佳策略的，但它明显是根据历史数据想当然地认为蓝领选民保准会投票支持民主党，上次大选他们确实把票投给了奥巴马，这次还会继续拥护民主党。有蓝领阶层的选票作为不可动摇的基础，希拉里只要获得少数派和自由派精英的支

持，就能够轻松取胜。因此“阿达”决定，竞选团队不需要在稳赢的州花钱拉票。然而，当竞选团队意识到某些应该稳赢的州不再胜券在握时，为时已晚。

“阿达”只是个计算机程序，和所有计算机程序一样没有常识或智慧。任何人哪怕稍作关注，都会留意到希拉里在面对伯尼·桑德斯时的弱点。74岁的社会主义者伯尼·桑德斯是佛蒙特州一位鲜为人知的参议员，在决定迎战希拉里之前，他甚至还不是民主党人。如果是正常人的话，就会想要弄清楚为什么桑德斯表现那么好，但“阿达”没有这种想法。

当希拉里在密歇根州的初选被桑德斯重挫时，有竞选经验的人一眼就能看出桑德斯传递的民粹主义信息具有巨大的吸引力，因此不能理所当然地认为希拉里可以获得蓝领阶层的支持。不过“阿达”并没有注意到这一点。

希拉里对在密歇根州遭到的意外打击感到愤怒，但唯独没有把自己受到的重挫归咎于“阿达”，她仍然相信“阿达”清楚怎样做是最好的，毕竟，“阿达”是台功能强大的计算机，不受人类偏见的影响，可处理大量千兆字节的数据，可每天进行难以想象的40万次模拟，没有人能与之抗衡。因此，该竞选团队还是以数据驱动为主，在很大程度上忽视了经验丰富的政治专家和亲自与选民交谈的竞选工作人员的请求。

众多选民先后对桑德斯和特朗普表现出极大的热情，而为数不多的希拉里的支持者则表现含蓄。“阿达”并未对此加以比较。数据库中没有关于“热情”的内容可供“阿达”处理，所以它忽视了活力和激情，希拉里的数据驱动型竞选也是如此。对计算机来说，凡是不可度量的东西都不重要。

最大错特错的是，希拉里竞选团队的数据专家竟然让比尔·克林顿缄默。克林顿可能是我们见过的最优秀的竞选者了，1992年成功击败时任总统乔治·赫伯特·沃克·布什登上总统宝座时，克林顿的竞

选口号为“笨蛋，问题在经济”（It’s the economy, stupid）。克林顿本能地知道选民看重什么，知道如何说服他们。

2016年竞选期间，比尔·克林顿看到了伯尼·桑德斯和唐纳德·特朗普呼吁工薪阶层选民时所激发的热情，于是建议希拉里以“笨蛋，问题在经济”作为主要竞选方针，尤其在萧条的中西部地区，包括俄亥俄州、宾夕法尼亚州、密歇根州和威斯康星州形成“蓝墙”（Blue Wall）。而“阿达”理所当然地认为，这道由“蓝墙”组成的防火墙会是希拉里战胜特朗普的基础。

“阿达”的另一个盲点是，经验丰富的政治家知道电视广告虽然可行，但最能打动乡村地区选民的方法，就是候选人安排时间出席市政厅会议和县博览会，以表示对选民的关心。而靠数据驱动的“阿达”着实没有考虑这一点。竞选活动支持率出现下降时，显然应该指派一名竞选专员深入乡村地区调查，而且要找能与农民打交道的人。这名专员还得是来自布鲁克林，而且不要出身背景太好的。

竞选期间，希拉里并没有采纳克林顿的建议（真的就连他的电话也不接了），这让克林顿恼羞成怒。他向希拉里的竞选主席约翰·波德斯塔抱怨：“那些还挂着鼻涕的小屁孩会把事情搞砸的，因为他们都把我的话当成耳边风。”

“阿达”断定，选民更担心的不是自己的就业机会，而是特朗普毫无总统风范的行为。于是，希拉里便将竞选重点放在抹黑特朗普上：“我并非无可挑剔，但特朗普更加糟糕。”

在“阿达”的建议下，希拉里的竞选活动几乎完全忽视了密歇根州和威斯康星州，尽管此前在这两个州的预选中都败给伯尼·桑德斯的经历本应为她敲响警钟。相反，希拉里把竞选的时间和资源都浪费在像亚利桑那州那样胜算不大的地区（也确实没赢），因为“阿达”判断希拉里能在这些州赢得压倒性的胜利。

竞选结束后，一名民主党民意调查者表示，“她在竞选时忽略了选举团的建议，也没有在密歇根州和威斯康星州等地投入必要的资源，这简直就是渎职”。

克林顿在希拉里败北后指责那些数据专家完全依赖计算机程序而忽略了数百万失业或担心失业的工薪阶层选民。据称，克林顿在与希拉里的一次通话中，气得将手机扔出了他在阿肯色州顶层公寓的窗户。

不知是数据不好，还是模型不当，但可以肯定的是大数据并非灵丹妙药，尤其是当大数据被藏在计算机内，而深谙现实世界的人类对计算机如何处理这些数据毫不知情的情况下。

计算机在某些方面确实表现出色。我们的生活也多亏计算机的赋能才更加丰富多彩。然而，希拉里·克林顿不是唯一过度迷信大数据的人，也肯定不会是最后一个。但愿我能说服你，不要加入他们的行列。

第1章

智能还是服从

《危险边缘》是一档热门的电视智力竞赛节目，有多个版本，开播至今已有50多年。该节目的比赛内容为百科知识问答，其巧妙之处在于：参赛者要根据以答案形式提供的各种线索，给出与这个答案相对应的问题。例如，线索是“美国第16任总统”，正确的问题就是：“谁是亚伯拉罕·林肯？”每期节目均有三名参赛者，以摁按钮的方式口头抢答（除了最后一轮“终极危险边缘”以外，在其他环节三名参赛者均有30秒时间书写作答）。

从很多方面来看，这档节目都适合计算机参与，因为计算机能准确无误地存储和检索大量信息。在《危险边缘》青少年组比赛中，一名男孩因将“谁是安尼·弗兰克”误写成“谁是安妮·弗兰克”而痛失冠军。而计算机就不会犯这样的错误。

另外，线索有时通俗易懂，有时却晦涩难解。例如，线索是“把它打进去，你就输了比赛”，对只是资料库的计算机来说，很难得出以下正确问题：“什么是（台球）母球？”

还有一个难解的线索是：“翻译时，这支大联盟棒球队的名字会重复一次。”正确问题为：“什么是洛杉矶天使队？”（What is the Los Angeles angels？）

2005年，15名IBM（国际商用机器公司）的工程师合作设计了一款能与《危险边缘》最佳玩家同台对擂的计算机，取名“沃森”，以纪念IBM的首任CEO（首席执行官）托马斯·J.沃森。沃森在1914年接手IBM时，IBM还只是一家仅有1 300名员工、年收入不足500万美元的小公司，到了1956年他去世的时候，IBM已经发展成为一家有7.25万名员工、年收入9亿美元的公司。

“沃森”程序存储了相当于2亿页纸的内容，每秒可处理相当于100万本书的信息。除了拥有海量内存和高速处理能力外，“沃森”还

能理解自然语言，使用合成语音进行交流。与罗列相关文档或网站的搜索引擎不同，“沃森”可按照程序并根据线索得出具体答案。

“沃森”运用数百个软件程序，先识别线索中的关键字和词组，再与海量数据库中的关键字和词组相匹配，最后得出合理答案。按照编好的程序，如果线索是某个名字（如亚伯拉罕·林肯），“沃森”就会写出以“谁是……”开头的问题；如果线索为某一事件，它就会写出以“什么是……”开头的问题。单个软件程序与某个答案的一致性越高，“沃森”就越能确定此为正确答案。

该程序能轻而易举地得出与“美国第16任总统”这么直白的线索对应的问题，但要处理有多重含义的词语时就有些困难了，比如，线索是“把它打进去，你就输了比赛”之类的问题。但是，“沃森”不会感到紧张，也绝不会遗忘。

2008年，“沃森”做好了参加《危险边缘》的准备，但还有些问题需要协商。IBM团队担心该节目的工作人员会使用包含双关语和具有双重含义的线索，给“沃森”下圈套。这一担心也恰好揭示了人类与计算机的巨大差异。人类可以根据语境理解词义，所以能理解双关语、笑话、谜语和讽刺批评。而目前的计算机，充其量只能检查出数据库中是否含有双关语、笑话、谜语或讽刺批评。

对此，节目工作人员同意随机抽取以往编写但未使用的线索。而节目工作人员也担心，如果“沃森”一得到答案就可以发出电子信号，会比必须通过摁按钮来答题的参赛者更有优势。对此，IBM团队同意给“沃森”装根电子手指来摁按钮，但它还是比人类快，这也让“沃森”占据决定性优势。摁按钮快算是聪明的体现吗？如果“沃森”的反应速度降为与人类的一致，比赛结果又会如何？

接下来，在2011年的人机大战中，“沃森”与《危险边缘》的两名前冠军肯·詹宁斯和布拉德·鲁特展开了两轮比赛。首轮比赛“终极危险边缘”的线索是：

它最大的机场以第二次世界大战的英雄命名，

它的第二大机场以第二次世界大战的战役命名。

两名前冠军给出的问题为：“芝加哥是什么？”而“沃森”给出的问题是：“多伦多是什么？？？？？”显然，“沃森”识别出了“最大的机场”、“第二次世界大战的英雄”和“第二次世界大战的战役”这些词组，然后在其数据库中查找相同主题，但没能理解线索的第二部分（“它的第二大”）指的是该市的第二大机场。“沃森”给问题添加了多个问号，因为它计算出的这一答案的正确概率仅为14%。

尽管如此，“沃森”还是以77 147美元轻松获胜，詹宁斯和鲁特的赛果分别为24 000美元和21 600美元。“沃森”夺得了100万美元的冠军奖金（IBM将其捐赠给了慈善机构），詹宁斯和鲁特也各自将奖金的一半捐赠给了慈善机构。“沃森”在《危险边缘》的取胜是一次价值数百万美元的宣传良机。在获得艳惊四座的胜利后，IBM宣称，相比在《危险边缘》中与主持人亚历克斯·特雷贝克较量，“沃森”的问答技能将运用于更重要的领域。IBM一直将“沃森”应用于医疗、银行、技术支持以及其他能利用庞大的数据库来解决具体问题的领域。

对许多人来说，“沃森”击败《危险边缘》的两名前冠军无疑证明强大的“沃森”无所不知！计算机比人类更聪明，我们应该依靠它，相信它的决策。也许我们还应该担心，计算机会在不久的将来征服甚至消灭人类。

“沃森”真的比我们聪明吗？它的胜利恰恰反映了计算机的优势和弱点。作为能力超强的搜索引擎，“沃森”可以在其庞大的数据库中快速查找单词和短语（它还有可以快速点触的电子手指）。我之所以没有使用“解读”这个词，是因为“沃森”并不了解那些单词和短语的含义，比如“第二次世界大战”和“多伦多”，它也不明白语境

中的词义，比如“它的第二大”。“沃森”的实力被过分夸大了，正如很多电脑程序一样，它的智能不过是假象罢了。

从很多方面来说，“沃森”的表现就是骗人的把戏，只不过是在范围极小的某些技能上看似具有超人的发挥罢了。设想有一个不懂英语，但有无限时间翻阅大型文库（藏有2亿页英语单词和短语）找出匹配单词和短语的人。我们会认为这个人聪明吗？计算机仅因能比人类更快地进行搜索匹配，就说明它聪明绝顶吗？

连IBM“沃森”团队负责人戴夫·费鲁奇也坦承：“我们在开发‘沃森’，设法让其仿造人类认知时，有坐下来好好谈过吗？根本没有。我们不过是想发明一台可以在《危险边缘》中获胜的机器而已。”

计算机不仅击败了《危险边缘》中的人类玩家，还击败了国际跳棋、国际象棋和围棋的世界冠军，这助长了人们认为计算机比最聪明的人类还要聪明的普遍观念。想要玩好这些战略型棋盘游戏，仅靠匹配单词和短语的强大搜索引擎是远远不够的，还要能分析棋盘格局、制定创意策略、做到未雨绸缪。这难道不是真正的智能吗？

接下来，我们就从非常简单的儿童游戏开始了解。

井字游戏

在玩井字游戏时，两个玩家在3×3网格上轮流画×和 ○（如图1.1所示）。无论是在水平方向、垂直方向还是在对角线上，只要三个方格连成一条线，该玩家即赢得比赛。

通过分析所有可能的移动序列，软件工程师可以编写出靠蛮力计算的程序来掌握井字游戏。玩家甲有9个方格可选择，在他走出第一步

后，玩家乙有8个方格可选择，前两步共有72种组合方式。走完前两步，玩家甲剩7个方格可选择。整局游戏玩下来，计算机程序必须处理的选择序列共有9×8×7×6×5×4×3×2×1 = 362 880种。

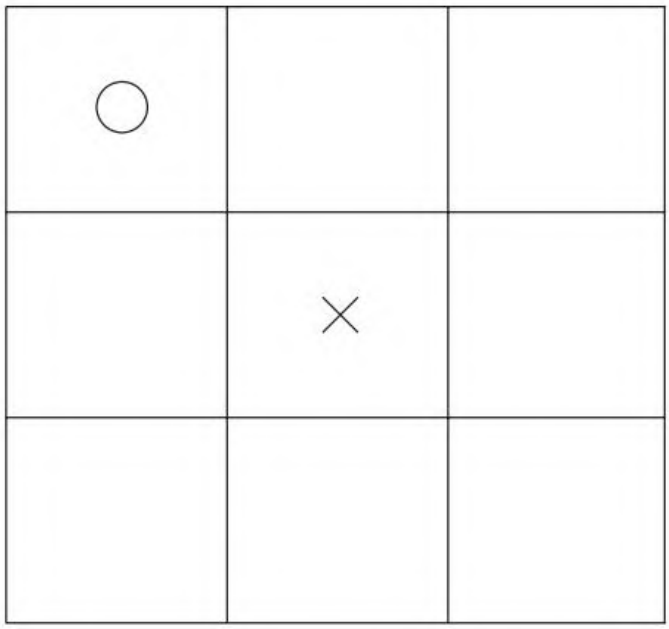


图1.1 井字游戏

也有更简便的分析方法，但重点是，井字游戏程序看待游戏的方式与人类不同。人类看到3×3的网格会思考选择哪些方块能完成三格连线，以及选择哪些方块会阻挡对手完成连线。但计算机程序无法对这些方格进行可视化，而是为每个方格分配一个1~9的数字（如图1.2所示），并识别获胜组合（例如1、2、3和1、5、9）。

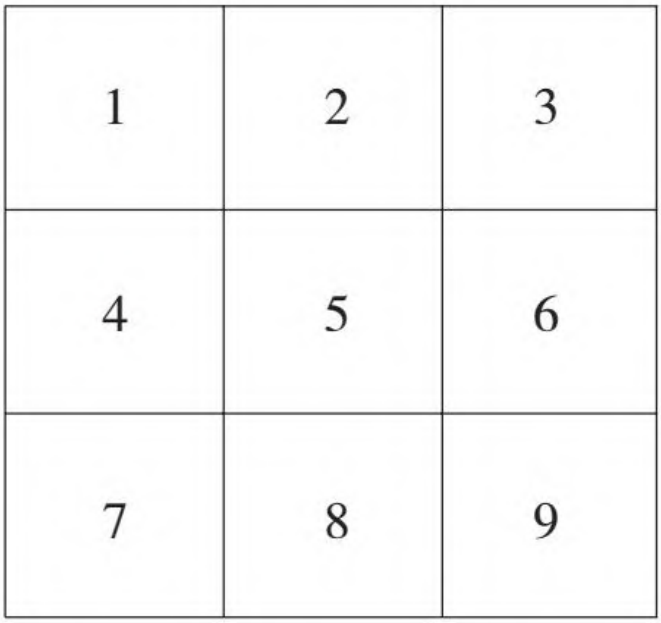


图1.2 匹配数字后的井字游戏

计算机程序会算出1~9的可能序列，识别各玩家的最佳策略，并假设对手会选择的最佳策略。一旦软件编写调试完成，就会立即显示出最佳策略。

假设玩家乙采用最佳策略，如果玩家甲从中心格或任一边角格起步，玩家乙就选择相反的方式——如果玩家甲选择中心格，玩家乙则选择边角格；如果玩家甲选择边角格，玩家乙则选择中心格。采用最佳策略的游戏总会以平局结束。

这就是蛮力计算，不涉及逻辑推理，只是无意识地枚举数字1~9的排列和识别获胜排列。

在井字游戏和其他游戏中，人类通常会避免对所有可能的移动序列进行蛮力计算，因为这样一来移动序列的可能性就会暴增。相反，我们使用逻辑推理，并将注意力集中在有意义的走法上。与蛮力计算程序不同，人类不会浪费时间思考明显错误的步骤。而没有逻辑和常识的计算机却还是会分析愚蠢的策略。

玩井字游戏时，人类玩家可能会研究3×3网格，而计算机玩的是1~9的数字。人类会采用可视化的方法，将注意力集中到中心格上，意识到这一格蕴含四个获胜排列，而每个边角格蕴含三个，每个边格蕴含两个。

中心格也是极佳的防守走法，因为接下来玩家乙无论选择哪一格，最多只蕴含两个获胜排列。相反，若玩家甲先选边角格或边格，就会让对方占据中心格，减少了自己的一个获胜排列，同时为对方创造了三个获胜机会。

从逻辑上讲，似乎起步最好选中心格，最后选边格。人类对棋盘的这种可视化认知和对中心格战略价值的判断，完全不同于软件程序对数字1~9所有排列的无意识识别。

人类也能发现游戏的对称性，即四个边角格任选其一开盘都同样可取（或不可取）。因此，人类只需思考选择其一的后果，选择其他三个边角格的后果就同理可得。游戏的对称性让人类每走一步都能减少需要考虑的移动步数。最后，人类会发现某些走法能迫使对手选择对其不利的方格，从而阻止对手完成三格连线。

人类能够运用战略性思维找出最佳策略，并发现采用最佳策略总会打成平局。有经验的人还会发现，同孩子玩游戏时不按常规出牌有时也能够获胜，例如开局选择边角格，甚至边格。

具有讽刺意味的是，尽管人类可以运用逻辑找出最佳策略，但人类编写的计算机软件程序还是有可能击败人类的，因为计算机无须考虑自己的走法。井字游戏的计算机程序只要遵守编程规则即可。相比之下，人类每走一步都必须思考，最后会疲惫不堪导致犯错。

计算机相比于人类的优势跟“智能”的一般含义毫无关联。正是人类编写出能识别最佳策略并存储于计算机内存中的软件，计算机才有规律可循。

尽管井字游戏这款儿童游戏会越玩越无聊，但它是很好的例子，凸显了计算机软件的威力和局限性。计算机程序对于烦琐的计算用处极大，编程软件每次的答案都完全一致，还能不厌其烦地完成已编程好的任务。与人类相比，计算机的处理速度更快、保存的信息更多。

人类怎能奢望在以信息记忆和处理速度为胜的活动上与计算机竞争呢？也许真正的奇迹不是计算机的强大，而是人类还在很多方面比计算机更胜一筹。遵循规则与人类毕生所获得的智慧，两者天差地别。

人类的智慧使我们能够识别出含义模糊的语言和扭曲的图像，对问题追根溯源，应对异常情况以及很多虽遵循规则却无法处理的事情。

国际跳棋

国际跳棋比井字游戏复杂得多，实际上，它复杂到根本无法对所有可能的移动序列进行蛮力分析。所以，你可能认为计算机必须模仿人类的思维才能下好国际跳棋。然而并非如此。

国际跳棋采用灰白棋格相间的8×8棋盘（如图1.3所示）。只能走灰色棋格，也就是说可走方格的数量从64个减到32个。两方玩家每方各有12枚棋子，放置于己方的灰色棋格内，中间的8个灰色棋格留空。棋子可沿灰色棋格对角移动，跳吃对方棋子。

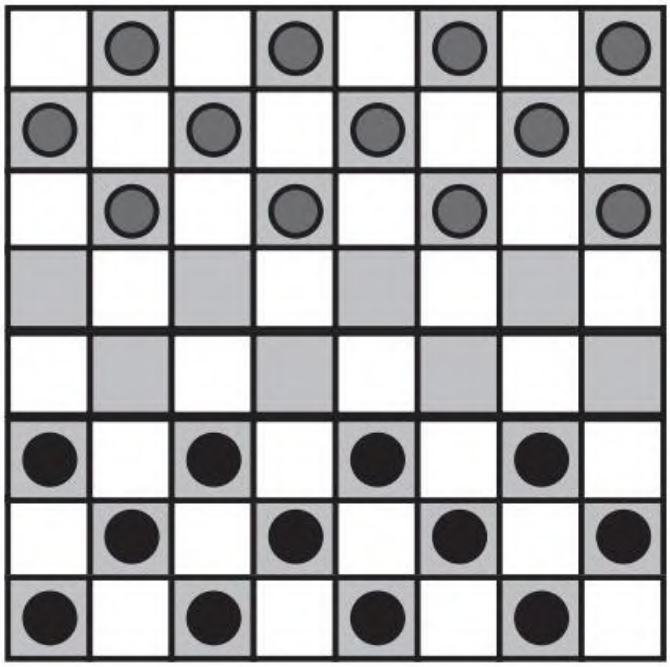


图1.3 国际跳棋棋盘

在理论上，尽管所有可能的序列都有无限步数，蛮力分析还是可以识别出最佳策略的，就像玩井字游戏那样。但是，对目前的计算机来说，在合理时间范围内要分析的可能序列数量过于庞大。因此，人类想出了简化策略来利用计算机的能力。与井字游戏一样，国际跳棋的计算机程序不会尝试制定逻辑策略，而是利用计算机的优势——快速处理和绝佳记忆。

井字游戏走九步就结束了，而国际跳棋有无限步数，因为玩家可以在没有哪方获胜的情况下不断来回移动棋子。实际上，来回移动棋子很无聊，所以除非有一方犯了非常低级的错误，否则玩家会在明显无法出现赢家时同意和棋。（冷酷无情的国际跳棋程序永远不会同意和棋，而是会一直玩到人类对手精疲力竭，累到无法清晰思考而犯错。）

虽然国际跳棋游戏的步数不受限制，但能走的棋盘位置还是固定的。用不着算出所有可能的移动序列，国际跳棋计算机程序更好的做

法是查看所有可能的棋盘位置，然后确定在这些位置的走法哪些得势、哪些失势。

尽管如此，这项任务还是让人发怵。棋盘走位有5万亿种可能，在没有考虑到所有接下来可能的位置序列的情况下，很难真正确定走这一步是否会得势。

人类以其洞察力将游戏分为三部分（开盘、中场和残局），单独分析每个部分，最后串联起来（如图1.4所示）。

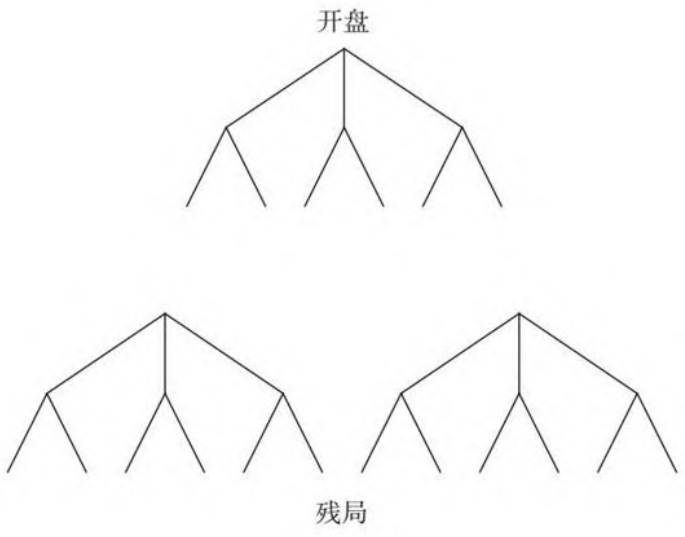


图1.4 国际跳棋的决策树模型

开盘的那几步棋已有写好的“剧本”，表明了最佳的开盘走法、每种开盘的最佳应对方式等。这些剧本是国际跳棋玩家几百年积累下来的集体智慧。每名严肃的国际跳棋玩家都会研究这些剧本。编写国际跳棋程序代码的软件工程师也会把剧本加载到计算机内存中，计算机会在开盘时遵守这些规则。

到了残局阶段，如果棋盘上只剩两枚棋子，则位置数量相对有限，而如果剩下三枚棋子，则位置数量会增加，但还在可控范围，以此类推。对于每个可能的位置，人类玩家能计算出最佳走法，同时确

定最佳走法是否会造成平局或出现胜方。所剩棋子数越多，可能位置的数量就越多，但很多都容易解决，并且棋盘的对称性也会减少必须分析的位置的数量。人类分析完包含所有可能棋盘位置的全部残局的情况后，比方说还剩不到六枚棋子，那么每个位置的残局最佳走法就会被加载到计算机内存中。

游戏进行到预先加载的残局位置时，计算机便按照人类预先确定的最佳走法的规则落子。人机对抗的跳棋残局中，人类玩家每走一步，计算机就会从数据库中选出预先确定好的最佳走法来应对新的棋盘位置，一直持续到比赛结束，通常结果是一方认输或双方同意和棋。

在游戏进行到中场时，计算机会试图将开盘剧本与残局位置联系起来。如果开盘几步之后，游戏进行到已存储的残局位置，则游戏结果可想而知（假设为最佳玩法）。

可供蛮力分析来识别最佳序列的中场局势数不胜数，因此程序员会将人类在跳棋领域的智慧与计算机的能力结合起来，列举各种序列。如果计算机有足够的能力和时间预测接下来的四步，那么计算机就会预测这四步可能产生的所有序列，并使用人类特定的损失函数（loss function）来比较四步后所有的可能位置。损失函数也是基于人类几个世纪的经验，考虑了被认为重要的因素，例如，每个玩家拥有的棋子数和对棋盘中心位置的控制。国际跳棋专家建议程序员为不同因素分配权重，以反映每个因素的重要性。

计算机通常会选择“最大最小值归一化”（minmax）的走法，因此它可以在最坏情况下（即最大值）让可能造成的损失最小化（即最小值）。如果另一个玩家采用最佳走法，程序则选择损失最小（或收益最大）的走法。

经过几个回合的中场比拼后，棋子数缩小到前瞻计算可以得出已知残局结果的水平。假设这是最佳玩法，那么游戏基本上结局已定。如果人类玩家犯错，则游戏结束得更快。

值得注意的是，计算机程序中“智能”的含量极少。在游戏开始时，计算机程序必须遵守开盘提示；中场游戏期间，计算机程序确定前瞻序列，并使用人类规定的损失函数，按部就班地决定走法；进入最后阶段，计算机程序还得依照残局指令运行。

为国际跳棋、国际象棋、围棋等复杂游戏而设计的计算机程序并不试图模仿人类思维，这涉及对潜在取胜原则的创造性认识。编写计算机程序是为了利用计算机的优势——无懈可击的记忆能力和毫无差错的规则遵守。

国际跳棋的计算机程序与人类玩家相比有几个重要优势：它永远不会在开盘和结束时犯错。人类玩家可能已研究过国际跳棋手册，但人类没有完美的记忆能力，还是会犯错。没有人思考过，更不用说记住所有可能出现的残局序列，其中有些还需要几十步精确走法才能得到最佳结果，人类只能在仓促之间找到最佳走法。而计算机的数据库中加载了最佳序列，可以做到这一点。

国际跳棋游戏中，人类击败计算机的唯一机会在中场。人类的预测能力可能不如计算机，计算机会分析不同走法背后大量的可能序列，但人类玩家能更好地把握特定位置的战略价值。例如，人类玩家可能会认识到，控制棋盘中间位置的重要程度比计算机损失函数给出的权重更高，或者计算机控制中间位置的数值测量可能有误，又或者中间位置的最终控制无法依靠测量目前局势得知。

计算机的最后一个优点是它不会累。高水平国际跳棋游戏可以持续两个多小时。由于大多数国际跳棋对决是以平局结束的，因此跳棋锦标赛会安排很多场比赛，一个多星期下来可能每天都会有四场。人类玩家每天的比赛时间为8~10个小时，一天接着一天，他们会疲惫不堪，容易出错。但计算机不会疲倦，因为它不需要思考，只要服从就好。

史上最优秀的国际跳棋选手是传奇人物马里恩·廷斯利。他是一个神童，开始念书的头八年就跳了四级，后来成为专攻组合分析的数

学教授。小时候，他每周用五天，这五天每天用八个小时来学习国际跳棋。读研期间，他称自己已经花了一万个小时研究国际跳棋。到了20多岁，他基本上已无人能敌。

有12年时间，廷斯利不再参加国际跳棋锦标赛，据称是因为他觉得非常保守的对手很无聊——他们希望的最好成绩是平局。后来重返赛场的他于1991年再次退役，时年63岁。1992年他又被国际跳棋程序奇努克（Chinook）团队的创建负责人、数学教授乔纳森·谢弗请回赛场。谢弗的研究团队有三个人，分别负责开盘数据库、残局数据库和中场损失函数。

在1992年廷斯利和奇努克的40场比赛中，大部分是平局。廷斯利赢了第5场比赛，那场比赛中奇努克遵从了已加载在其剧本中的一种次优走法。廷斯利输了第8场比赛，将其归因为疲劳过度。到了第14场比赛，奇努克采用了数据库中廷斯利多年前使用过的一连串走法，但廷斯利忘记了，因此输了比赛。后因奇努克发生故障，廷斯利拿下了第18场比赛。（计算机也会疲劳？）随后，廷斯利还取得了第25场和第39场比赛的胜利，最终以4胜2负33平的成绩取胜。

这是人类大战机器中人类的一次胜利，但两场比赛的失利，却是廷斯利45年国际跳棋职业生涯中仅有的两次。

谢弗极大地扩充了奇努克的开盘和残局数据库，还将中场的前瞻能力从17步增加到了19步。1994年，他要求再进行一次对决。前六场比赛为平局，不过廷斯利认为奇努克的水平已经得到提升。他表示，在奇努克的残局数据库足够巨大以至于不会出错之前，他只有10~12步的机会能获得领先优势。可惜的是，廷斯利因患胰腺癌而不得不放弃比赛，并于7个月后与世长辞。

廷斯利的记忆力惊人。1992年第一次比赛后，他给谢弗讲了自己40多年前的一场比赛，他仍能准确无误地记住每一步。尽管如此，他的记忆力还是无法与强大的计算机匹敌。廷斯利真正拥有的是通过多

年研究和实践积累的棋感，奇努克绝不可能对位置的优劣有相同的直觉。

在决胜局前的14场展示赛中，廷斯利和奇努克有13场平局，第10场为廷斯利获胜，谢弗随后描写了这场决定性的比赛：

> 我走下了奇努克的第10步。刚放下棋子，廷斯利就惊讶地抬起头说：“你会后悔的。”我尚未领略过伟大的廷斯利的行事风格，默默地坐在那里，心想：“你知道什么，我的程序正在搜索后20步的可能性，表示它占优势。”再走几步后，奇努克的评估降至旗鼓相当。又走了几步后，它表示廷斯利更占上风。后来，奇努克说它遇到了麻烦。最后，越下越糟，我们只好投降了。在廷斯利的比赛日志中，他透露自己已经预料到残局，在第11步就知道他会赢，也就是我们出错的下一步。而奇努克需要预测后60步，才能知道它的第10步下错了。

廷斯利去世后，奇努克与世界排名第二的国际跳棋选手唐·拉弗蒂进行了32场比赛，并以1胜31平取胜。1996年，奇努克退出国际跳棋锦标赛，不过你可以在线对战低配版的奇努克。退赛后，奇努克同数十台差不多连续运行了18年的计算机一起工作，以检验确认国际跳棋玩家在先走并且每一步都是最佳走法的情况下是否可以保证取胜。

2007年，谢弗宣布国际跳棋和井字游戏一样，也是一款极好的权衡游戏，如果每个玩家都能选择最佳走法，则可以保证平局。这是计算机的一项壮举，但我不会称其为智能。

下一代的计算机游戏程序采取了不同的做法，即试错过程——计算机跟自己比赛数百万次，同时记录取胜方式。一款名为AlphaGo（阿尔法围棋）的程序采用了这种方法，击败了世界上最顶尖的围棋手。此外，另一款名为AlphaZero（阿尔法零）的程序还击败了最好的计算机国际象棋程序。这些程序都能极好地执行范围狭窄、目标明确的任

务（“将”对手的军），但不会像人类那样分析棋盘游戏，思考为什么某些策略会成功。即使是计算机编码员也不明白为什么他们的程序有时会选择不寻常的甚至是奇怪的特定走法。

创建AlphaGo和AlphaZero的公司DeepMind（深度思考）的首席执行官戴密斯·哈萨比斯举了个例子。在一场国际象棋比赛中，AlphaZero将“后”移到棋盘的边角格，这与人类想法相矛盾，因为国际象棋中最厉害的“后”在棋盘中间位置会更加强大。在另一场比赛中，AlphaZero牺牲了“后”和一个“象”，而对人类玩家来说，除非可以立即获得回报，否则几乎不会这样走。哈萨比斯说：“AlphaZero与人类的玩法不同，与编程的玩法也不同。它采用第三种玩法，似乎是外星人般陌生怪异的玩法。”

尽管在棋盘游戏中具有怪异的超人技巧，但计算机程序并不具备类似人类智慧和常识的东西。这些程序不具备处理不熟悉的情况、不明确的条件、模糊的规则以及含糊甚至相互矛盾的目标所需的一般性智能。决定去哪里吃晚餐、是否接受一份工作、跟谁结婚，都与“象”走三步“将”对方的军截然不同——这就是为什么让计算机程序为我们做决定是危险的，不管它们多擅长棋盘游戏。

第2章

盲从

奈杰尔·理查兹是新西兰籍马来西亚人，他是职业拼字游戏玩家（是的，拼字游戏也有职业玩家）。他母亲回忆说：“他开始学说话的时候，对字词不感兴趣，只对数字感兴趣，他把一切都与数字联系起来。”他28岁时，他的母亲跟他比赛玩拼字游戏：“我知道一种你不擅长的游戏，因为你拼写不是很好，而且你英语也没学好。”四年后，理查兹赢得了泰国国际大赛（国王杯）的冠军，这是世界上最大的拼字游戏锦标赛。

他多次赢得美国、英国、新加坡和泰国拼字游戏锦标赛的冠军（锦标赛每两年一届，他在2009年获得亚军），还分别在2007年、2011年和2013年赢得了拼字游戏世界冠军。

2015年5月，理查兹决定记住法语拼字游戏中允许使用的38.6万个单词（北美拼字游戏中允许使用的单词数为18.7万个）。之前除了bonjour（你好）和每个回合中记录得分的数字，理查兹一句法语都不会，除此之外，他也不理解法语单词的含义，他只是记忆。

9周后，他在法语拼字游戏世界冠军决赛中以565：434的惊人成绩取得胜利。就算他练了9周，每天练习16个小时，要记住法语拼字游戏中的38.6万个单词，每个单词平均也只有9秒钟的记忆时间。然而据称，理查兹并没有逐个记单词，而是逐页浏览，字母都印入脑海中，他在玩拼字游戏时只是按需回忆。

理查兹在法语拼字游戏锦标赛中的表现与在英语比赛中一样快速敏捷，完全无法想象他其实不会用法语进行交流。对理查兹这样的专家来说，拼字游戏本质上是数学游戏，通过棋子的组合获得积分，限制对手做同样事情的机会，同时保留将来可能有用的字母。重要的技能是识别模式和计算概率的能力，没有必要知道单词的含义。

理查兹是史上伟大的拼字游戏玩家之一，尽管他生性安静谦虚，就像一台专注于自己业务的电脑一样。

或者我应该说计算机就好比理查兹？计算机不知道任何单词的真正含义，它只是处理存储在内存里的字母组合。计算机存储器容量很大、处理速度很快，但处理字母组合是很有局限性的任务，仅在特定明确的情况下有用处，例如单词分类、字数统计或单词搜索。许多计算机的专长也是如此，它们都令人印象深刻，但其应用范围严重受限。

单词、图像和声音识别软件受到粒度法（granular approach）的限制——它们试图匹配单个字母、像素和声波，而不是像人类那样在语境中识别和思考。

思考之源和思维之火

1979年，34岁的道格拉斯·霍夫施塔特（中文名为侯世达）因其著作《歌德尔、艾舍尔、巴赫——集异璧之大成》获得普利策奖。该书主要探索我们的大脑是如何运作的，以及计算机有朝一日会如何模仿人类的思维。霍夫施塔特一生都在努力解开这个极其难解的谜题。人类如何从经验中学习？我们如何理解我们生活的世界？情绪来自哪里？我们如何做出决定？我们如何编写不灵活的计算机代码来模仿神秘灵活的人类思维？

霍夫施塔特得出的结论是：思考之源和思维之火。当人类看到一个活动、阅读一篇文章或听到一段对话时，能够专注于其最突出的特征，即“骨骼本质”（skeletal essence）。真正的智能是识别和评估情境本质的能力。人类通过类比其他经验来理解这一本质，又利用这一本质来增加经验。霍夫施塔特认为，人类智能的根本是收集和分类人类经验，然后对其进行比较和组合。

令霍夫施塔特沮丧的是，计算机科学走向了另一个方向。计算机科学家不再尽力模仿人类大脑，而是专注于计算机存储、检索和处理信息的能力，因此计算机软件变得有用起来（并有利可图）。软件工程师不再尝试理解人类思想如何运作，他们转而开发产品。

限制计算机科学研究的范围，也就限制了计算机的潜力。如果程序员不愿尝试，计算机就永远不会像人类思维那样智能。霍夫施塔特感叹道：“作为人工智能领域的新人，我不想参与一些花里胡哨冒充智能的编程行为，尤其是当我知道这与智能毫不相关时。”

对AI（人工智能）所采取的迂回路线，有一个很好的比喻。人类一直梦想着飞行，梦想着双脚离地在天空翱翔，甚至还梦想着去月球旅行。

这是异常困难的，早期很多尝试都以失败告终，例如传说中的伊卡洛斯（Icarus），他身缚由蜡和羽毛制成的翅膀。离开地面的另一种方法是爬树，这需要力量、技巧和决心，有可能会成功，但无论树有多高，它都不会使我们翱翔天空或到达月球（如图2.1所示）。

同理，人工智能不再尝试设计按人类思维方式思考的计算机，而是运用技巧和决心，虽然也取得了成效、产生了作用，但是爬到“树顶”（有些人认为我们可能接近“树顶”了）也不会让我们更有可能制造出真正拥有人类智能的计算机。



图2.1 月亮和树

霍夫施塔特举的一些例子令人信服。就像大写字母A，即使可以用不同的字体和样式书写，人类还是能立刻辨认出来，因为我们将字母A的变体同我们曾经看到和记住的变体进行类比。霍夫施塔特称此为“心理类别的流动性”。

软件程序就完全不同了。这些程序通过编程，将字母A与通过特定方式排列而成的像素关联起来。如果与其内存中的像素图案高度匹配，计算机就能识别该字母；如果与计算机的记忆略有不同，它就无法识别。

那些出现奇怪字母的网页验证码小框（如图2.2所示）就是建立在计算机对此无能为力的基础之上的，这称为CAPTCHA（全自动区分计算机和人类的图灵测试）。人类可以轻松破译字母变体，电脑却不能。

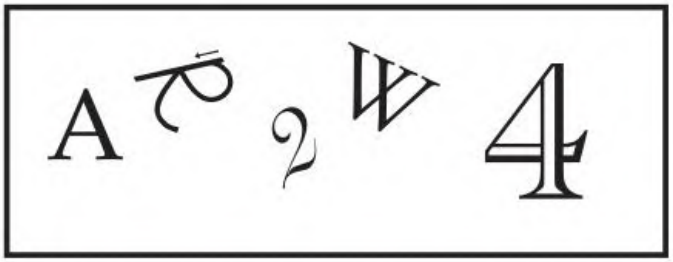


图2.2 字母和数字变体的验证码

除了数字和字母，CAPTCHA还会要求用户点击含有某些物体图像的方框，如花、椅子和道路等，因为人类会立即识别出无数变体，但计算机程序却不能。我们并不是说计算机永远无法像人类那样识别物体。而且其视觉识别程序一直不断改进，总有一天会变得非常可靠。要害在于这些程序不会像人类大脑一样运作，因此，称其具有人类那样的智能具有误导性。

我们将事物分解为“骨骼本质”，并且能认识到它们如何结合在一起。看到图2.3中的简笔画，我们可以立刻把握其本质（一个箱子、一个手柄、两个轮子和文字），也明白箱子、手柄、轮子和文字是如何相关的。我们认为它是某类马车，可以滚动，可以托运物品，还可以拉动滑行。我们不需要匹配像素就能知道，是因为我们见过箱子、手柄和轮子，知道它们有什么用处，无论是单独的零件，还是零件的组合。我们看到箱子上的文字“红魔鬼”，可以立即读取，而且我们知道这是不重要的装饰。

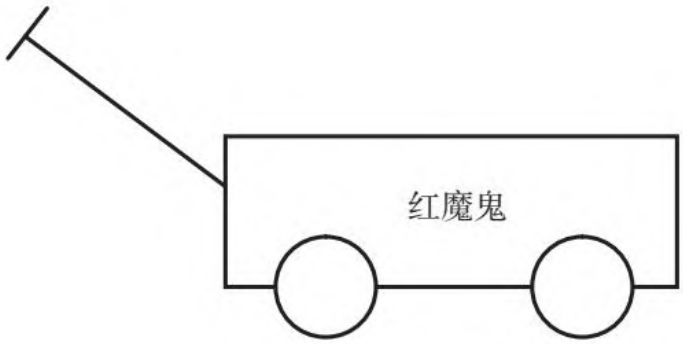


图2.3 简笔画：这是什么?

计算机没法这么做。计算机需读取数百万或数十亿张货车图片，才能得到该像素的数学表达方式。然后，在需要识别货车图片时，计算机会创建图片像素的数学表达式，并在其数据库中寻找高度匹配的结果。这个过程很难确定——匹配结果有时令人惊叹，有时则令人捧腹。

我曾请一名优秀的计算机科学家使用最先进的计算机程序来识别图2.3，该程序有98%的把握能确定图中是一家商店，也许是因为矩形上的文字类似店面标识。

人类无须看过100万辆货车才能知道货车是什么。一辆货车足以让我们了解其关键特征，不仅如此，我们还可以了解货车能做什么、不能做什么——货车可以被吊起、被放下、被填充、被启动，但不能飞行或翻跟头。

人类思维掌握事物基本特征并理解其含义的能力确实令人惊叹。我们知道图2.3是个箱子，因为它与我们看到的其他箱子相似，即使这些箱子的尺寸大不相同，也可能没有轮子和手柄。几乎同时，我们能识别所见的圆圈是轮子，因为它们位于箱底，我们知道这能让箱子移动，我们曾经见过箱子随轮子移动。我们还会推测箱子另一侧可能还有两个轮子，即使我们没有看到它们，因为我们知道如果另一侧没有轮子，箱子就会站不稳。

我们根据经验得知，带轮子的箱子通常用于搬运物品，所以，这个箱子可能是空心的，有空间可以存放或搬运东西。即使我们看不到箱子内部，也可以推测里面装有东西。我们还会想到用小推车来装载玩具、石块或小动物的情景。

即使只用了两条线绘制，我们也知道那是手柄，因为带轮子的箱子通常都有手柄或马达，这个从箱体延伸出的东西，类似我们见过的

其他手柄，附在物件上，可用于推拉。我们认为，“红魔鬼”的字样很可能是不重要的装饰，因为箱子表面的文字通常只起装饰作用。

我们不需要为找到和这个特定物品一模一样的东西，把记忆翻个底朝天才能得到结果（这是计算机程序的方式，即通过处理像素寻找匹配）。相反，我们神奇的大脑能够掌握基本要素，并理解其组合的含义。

我们的大脑可以连续处理那些来来回回、类比不断且迅速闪现的想法。我们可能会将“红魔鬼”的字体与其他字体进行比较，还会比较相近的颜色。我们可能会想到雪橇、出现雪橇的电影或滑雪的经历。我们可能会想到可以驾乘或用来运物的汽车或马。想法来得太快，以至于我们无法有意识地跟上它们。想法不由自主地涌现，如果我们的大脑不那么非凡，就会被这些想法冲垮。

我们可以筛选出几十个（甚至是数百个）想法——留住一些，忘记一些，结合一些——再由一些想法联想到其他想法，它们通常与我们最初的想法仅有一点儿模糊联系。计算机会有哪怕一丁点儿类似的意识吗？

计算机不理解人类的信息，如文本、图像、声音。软件程序会尝试将特定的输入与存储在计算机内存中的特定内容进行匹配，以生成软件工程师指示计算机生成的输出。特定细节的偏差可能会导致软件程序失败。“红魔鬼”这一文本可能会使计算机感到“迷惑”，因此无法将那个箱子识别为箱子。粗略描绘出的手柄可能会被误认为是棒球棒或电线杆。圆圈可能被误认为是馅饼或保龄球。

我们思维的灵活性使我们能够轻松处理模棱两可的情况，并在特定情况和一般情况之间来回切换。我们能识别出具体某辆小推车，是因为我们知道小推车通常是什么样的。

人类可理解形式复杂的文本、图像和声音，从而能够推测过去，预测修改或合并事物的结果。在小推车的例子中，我们可能会从车的

简略结构推测出它是手工制作的，还可能会因为没有凹痕和划痕而认为它是新的，或者至少是保护得很好的。我们可能会预测，如果下雨的话，小推车会积满水；用手柄拉动或者用力推，它就会移动。我们还会想象，如果两个孩子玩小推车，一个孩子会爬到里面，另一个孩子会拉着它走。我们可能还希望推车的主人没有走远。我们可以估算购买小推车要花多少钱。我们会想象坐在小推车里面是什么感觉，即便我们从未在现实生活中这么做过。

即使它被装饰成马或宇宙飞船的样子，我们也能从车轮和手柄知道这是辆小推车。我们还能够识别出没有看见的东西，例如，种植在车上的番茄植物、绑在袋鼠背上的推车、大象用象鼻摆弄推车。人类可以用熟悉之物来联想陌生的事物。而计算机还远远不能识别陌生的事物，确切来说是不一致的图像。

综上原因，将计算机称为智能是错误的。

计算机是超人吗？

计算机有强大的记忆，能以极快的速度输入、处理和输出大量信息。这些功能让计算机能够完成如超人般的壮举：在流水线上不知疲倦地工作，解决复杂的数学方程组，在陌生的城镇找到面包店的详细方位。

计算机在存储和检索信息方面非常出色，如乔治·华盛顿在哪天出生，玻利维亚的首都是哪里，利物浦最后一次赢得英超联赛是什么时候。计算机在进行快似闪电的计算方面也表现得非常出色，而且始终如一。如果要求正确编程的计算机算出8 722的平方，它基本上可以立即得出答案，并且每次的答案都相同。将同样的问题抛给任何一个普通人，他都要算很久，而且答案不一定可靠。

这种精确性和一致性使计算机极其擅长快速可靠地执行重复性任务，其精确度高且不会感到厌倦或疲劳。比如，计算机在监督巧克力饼干的生产时，可以保证每次配料都一样，烘焙温度、时间和包装统统一样。

机器人可以刀枪不入，这使它们能够探索遥远行星和茫茫深海，下到矿井里或去其他对人类有害的环境中工作。

计算机可以保持精确的记录（如医疗、银行和电话记录），并飞快地检索信息。自动机器人和无人驾驶车辆让我们能自由地做更多有益的活动。机器人宠物和电脑游戏让我们感到愉快，但尽管使出浑身解数，计算机还是做不到真正的关心和喜爱。机器狗与真正的狗不一样。Siri（苹果智能语音助手）也不是真正的朋友。

计算机对照片、视频、电视和电影的编辑能力也很强。例如，计算机软件能完成相同物体放于不同位置的图像或不同物体放于相同位置的图像之间的无缝变形。计算机在这项技术上做得比人工手绘更快更好。

即使是儿童也可以使用这样的软件来扭曲、放大和改善照片和视频。该软件可识别人的鼻子，将其变成猪鼻子，或识别人的头发并在其上添加滑稽的帽子，或识别出人嘴并将其歪曲成牙齿尽失和脾气暴躁的样子。人类也能做同样的事情，但做不到这么好，而且要慢得多。

奇点

在耸人听闻的书籍和电影的推波助澜之下，这些超人的壮举使许多人担心人类的未来。基于对计算机发展历史的推断，世界领先的计算机预言家雷·库兹韦尔预测，到2029年，机器智能将完胜人类智能，出现他所谓的“奇点”。计算机将具备人类所有的智力和情感能

力，包括“搞笑、浪漫、友爱和性感”。他还预测，到2030年，血细胞大小的计算机纳米机器人将被植入我们的大脑中，“这些将从根本上把我们的新皮层置于云端”。我们的大脑不再需要花费许多小时品读《战争与和平》，只需要几分之一秒就能将其阅读完毕。他并没有说那几分之一秒的阅读是否会产生愉悦感。对某些事而言，效率并不是最重要的。

有些牵强附会的设想认为，这些超级智能机器将会通过奴役或消灭人类来自我保护。这些猜测都是有趣的电影情节，仅此而已。

更离谱的是，想象机器人已经接管人类，我们自认为正在经历的生活实际上是机器为自我娱乐而开发的巨型计算机模拟的一部分——尽管很难想象机器人需要娱乐。

某些令人不安的预测只是想象力太丰富了——极具讽刺意味的是，计算机并没有那种创造力。其他预测则是基于对原始计算机能力发展历史的简单推断——计算机容易产生的那种不严谨的推断。

一项有趣的研究分析了过去350年里的英国公众演讲者，发现其演讲的句子的平均长度从弗朗西斯·培根的每句话72字降至温斯顿·丘吉尔的每句话24字。按照这种趋势发展（著名的“最后一句话”推论），最终每句话的字数会归零，然后变成负数。与此类似，20世纪一项针对取得奥运会100米短跑金牌的男女运动员所耗时间的推论表明，到2156年奥运会，女运动员会比男运动员跑得更快，其跑步时间将在2636年变成零（隔空传送？），然后变为负数（时光倒流？）。

马克·吐温反其道而行之，想到了：

> 在166年的时间里，密西西比河下游缩短了220英里[1]，也就是平均每年超过1英里的小事。因此，任何一个不盲目或愚蠢的冷静之人都可以看到，在鲕状岩志留纪时期（Old Oolitic Silurian Period），密西西比河下游长达130万英里，像一根钓竿一样伸向墨西哥湾。同样，所有人都可以看到，742年后，密西

西比河下游将只剩1.75英里 …… 科学也有令人着迷之处。人们可以用微不足道的投资，来获得如此大规模的回报。

不幸的是，不严谨的推断并不总是幽默的——至少不是有意为之。

对计算机能力的推断有多种形式，最著名的是1965年据仙童半导体公司和英特尔的联合创始人杰弗里·摩尔的观察，在1959—1965年，每平方英寸[2]集成电路上可容纳的晶体管数每年翻了一番。虽然摩尔并没有声称这是像物质守恒或热力学定律那样的物理定律，但它还是被称作摩尔定律，并且广为人知。这只是基于5年的实践数据观察（1959年、1962年、1963年、1964年和1965年）。后来到了1975年，摩尔把这一增长率修改为每两年翻一番。

指数增长意味着体量越来越大，速度越来越快。无论是每年还是每两年，翻10番就等于大约增加了1 000倍，翻20番就等于大约增加了100万倍。这就是指数增长通常会减缓或结束的原因。摩尔定律最受人关注的是，该定律的某些版本已经持续了很长时间。然而近年来，该增长率已经放缓至每两年半翻一番。摩尔在2015年预测，由于原子尺寸的物理限制，指数增长将在未来5~10年内结束。

无论他的预测是否正确，处理能力都不是计算机变得比人类聪明的瓶颈。每秒处理100万、10亿或1万亿个单词的能力，以及计算机能否真正像人类一样思考，两者之间没有必然联系——使用常识、逻辑推理、情感，从特定情况推导出一般原则，并能在其他情况下应用这些原则。

将时间考虑在内

有一件事情，人类能做得非常好，但计算机做得很差或无法做到，那就是将时间考虑在内。我们能通过观察一系列事件来理解事物。我们看到马车移动，理解了它的作用；我们看到运球、传球和射门得分，了解到什么是足球；我们观看舞蹈，能理解舞者在干什么。如果图像识别软件只关注单个静止图像，而不是一系列事件，便无法做到这些。

更重要的是，人类通过观察一连串事件，可以提出理论来解释和理解我们所生活的世界。看到某种纸张被点燃，我们会假设是火柴引起纸张燃烧，因此如果其他纸张碰到点燃的火柴，也可能会燃烧。我们会通过思考概括出，用点燃的火柴或其他燃烧物触碰物体可能导致物体着火，而且接触点燃的火柴可能带来危险。看到棒球被击中后能飞行数百米，我们会假设是球棒使球移动，所以如果用球棒击球，球就会移动。我们会通过思考概括出，用球棒等坚硬物体击打某物，可以使某物移动。因此我们应该避免被球棒击中，或被汽车和其他硬物碰撞。我们看到球落回地面，又想到无论在地球何处，所有被抛向空中的物体都会落回地面，便会认为肯定有隐形的力量将物体拉向地球。如果计算机不能处理一连串的事件，它就无法像人类一样做出假设和概括。

识别像素与产生情绪

想想你在进行完跑步、游泳、骑自行车等剧烈运动后有什么感觉。多巴胺、内啡肽和血清素的释放让人极度兴奋。我们在精力充沛、感到痛苦或沮丧的不同状态下，感受和思维都不一样。计算机就不会如此，计算机不会感到快乐、痛苦或悲伤。无论发生了什么、正在发生什么，或者将会发生什么情况，计算机的输入处理和输出结果

都保持高度一致。这不一定有什么好坏之分，只不过是不同于人类大脑的运作方式而已。

很早以前就有人开玩笑说，计算机可以在着火的房间里下出一步绝妙的国际象棋。和很多笑话一样，这也基于一个令人不安的事实——即使计算机可以识别出大火生成的像素，也不会对它即将被烧毁这一事实产生任何情绪。

人类的情绪取决于过去的记忆和目前的感受。计算机在记录足球比赛的分数时，对团队或分数不会有任何感觉。尼克·霍恩比的精彩著作《极度狂热》描述了他对阿森纳足球俱乐部的痴迷，在阿森纳遇到困难时的痛苦，以及上百人看到阿森纳近况时可能会想起他的那种开心："失联很久，可能永远不会再相见的前女友和其他坐在电视机前的人，有那么一刻会同时想到尼克，虽然那一刻很短暂，但他们会为我感到高兴或悲伤，我就喜欢这样。"说真的，计算机是否会想到处理足球比分的其他计算机？

现在看到阿森纳战绩的尼克，可能与他年轻时的感觉有所不同。他看到阿森纳战绩的感觉，肯定和埃弗顿足球俱乐部的支持者或不关心足球的人不一样。而计算机就不会这样，足球比分是要存储、处理和提取的数字，不是用来感知的。

当可识别图像的计算机程序处理足球场的图片时，虽可以正确地识别出足球场的样子，但对它没有感觉。而我在看到足球场时，会想起进过和没进过的球、赢过和输过的比赛。我记得我的队友和对手。我记得自己过去比现在更加擅长运动，后悔后来没有坚持踢球。我会开心，也会伤心。我有感觉，这些感觉会影响我对事物的看法。我现在对足球场的看法和以前踢球时不一样了。我对足球场的看法与你的也不同。我在被拥抱和被抛弃时，对足球场的感受都不一样。我对足球场的感觉，可能在早上和下午或晚上不一样，和我在梦中看到足球场时的感觉也不一样。计算机就不会这样。

我们都有过这样的经历，走路、洗澡或睡觉的时候，并没有在特别思考什么事情，但脑海中会突然出现灵感。有些灵感很愚蠢，有些确实能启发人。我是在洗澡的时候有了写本书的灵感的。我模糊地思考着我曾写过的电脑代码中的一个漏洞，但无中生有似的，我突然想到计算机不可能会像我一样享受淋浴的快乐。然后我又思考，我们的梦是如何被大脑基于经验创造出来的混乱零碎的图像和想法所混淆的。我不知道这些东西来自何处或有何意义，但我知道计算机没有这些。计算机不会感到快乐或恐惧、爱或悲伤、满足或渴望，也并不会为出色地完成工作而感到自豪。

人脑不是计算机，计算机也不是人脑。人脑和计算机有如此多的不同，所以将计算机称为智能甚至人工智能似乎都不对。

批判性思维

人工智能这个标签涵盖很多机器行为，这些行为与人类行为相似。随着计算机越来越强大，该标签又涵盖了很多新活动，同时也排除了另一些活动。能把盖子盖在瓶子上的机器人算人工智能吗？能检索某个词语的定义算人工智能吗？能检查拼写的正误算人工智能吗？

无论人工智能是如何定义的，其与人类智能的相似之处都是表面的，因为计算机和人类思维是截然不同的。例如，计算机被编程去执行非常具体的任务，而人类思维则具有处理不熟悉情况的多功能性。梦寐以求的“圣杯”是具有一般性智能的计算机，可以远不再死守指示，而是能足够灵活地轻松处理新事物和不明状况。

人类大脑平均有近1 000亿个神经元，这个数量远远超过了最强计算机的复制能力。虽然非洲象的神经元数量是人类的三倍，某一种海豚大脑皮质内的神经元数量几乎是人类的两倍，但是大象、海豚和其

他生物都不会写诗和小说，不会设计摩天大楼和计算机，也不会证明定理并提出理性论证。所以不能仅仅用神经元数量去衡量智能。

赋予计算机与人类一样多的神经元也不会使计算机比人类更聪明。关键在于神经元的使用方式。人类懂发明、会规划、能找到创造性的解决方案，有动机、感受和自我意识，能识别善与恶、事实与幻想之间的区别。计算机目前无法做到这些，因为我们还不知道如何让计算机像人类一样思考。

我们对大脑的运作方式知之甚少。神经元如何感知信息、如何存储信息、如何学习、如何指导我们的行动？我们对此一无所知。因此，我们称人类大脑为真正的智能。

我告诉学生，他们在大学里可以学到的最有价值的技能是批判性思维，包括评估、理解、分析和应用信息。目前，计算机在批判性思维方面存在严重缺陷。计算机可以存储和检索信息，但是无法评估信息的有效性，因为计算机并不能真正理解文字和数字的含义。计算机怎么知道从网站获取的数据是可靠的还是具有误导性的？计算机可以发现统计模式，但不能有说服力地解释这些模式的逻辑模型。计算机怎么会知道澳大利亚的温度与美国股票价格之间的统计相关性是合理的还是巧合？

生活不是多项选择题，也不是记忆事实的反刍。批判性思维可以将一般性原则应用于具体情况。例如，一个有用的一般性原则是，理性决策不应受已经发生且无法改变的沉没成本的影响。假设你以特价买了一个巨大的冰激凌圣代，但吃了一半你感觉不舒服，你会因为想“钱不能白花”而吃完这个圣代吗？问题不在于你花了多少钱（沉没成本），而在于是吃完剩下的冰激凌还是干脆扔掉它好让你舒服一些。

你有伊利诺伊大学橄榄球比赛的季票。到了11月，球队表现糟糕，天气也越来越不好。你会因为已经买了门票而冒着冷雨去看球赛吗？你买了50美元一股的股票，没过多久，那家公司报告出现巨亏，

股价跌至30美元每股。你是会为了税收减免而出售股票，还是会继续持股（因为一旦亏本卖出就等于承认当初买股票时犯了错）？原则上，人类可以将这种沉没成本应用于所有同类甚至更多情况。计算机编程确实可以执行具体操作，但要是从具体例子中概括出一般性原则，并将一般性原则应用于特定情况的话，就捉襟见肘了。

图灵测试

我们什么时候才能说，计算机可以真正像人类那样思考？要回答这个问题，我们需要先说明什么是“思考”。1950年，英国计算机科学家艾伦·图灵提出了“图灵测试”。图灵建议采用更简单的方法，而不是就思维的具体定义达成一致，然后看看计算机能否满足这些条件。在一场模拟游戏中，一个人分别与计算机和另一个人交流，计算机能否成功识别并区分两者呢？

在图灵测试场景中，一名人类提问者使用计算机把问题发送到位于不同房间的两台计算机上。一台计算机由人操作，另一台由计算机软件操作（如图2.4所示）。提问者的任务是，确定哪个是计算机软件给出的答案。通过文字或语音进行对话的计算机程序，被称为聊天机器人（chatbot，chat是指这些程序会聊天，bot是指聊天的是机器人）。

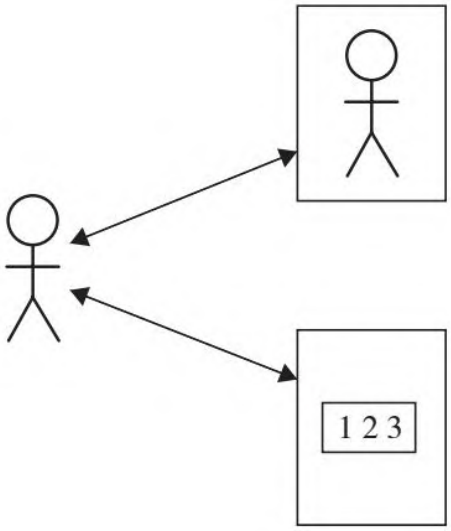


图2.4 图灵测试示意图

像“沃森”和Siri这样的聊天机器人能够很好地回答简单和可预测的问题，例如：“伦敦的气温是多少？”但会被下面这三个出乎意料的问题搞糊涂，因为计算机并没有真正理解单词的意思，也没有常识。

如果我把橙汁和牛奶混在一起，再加点盐会好喝吗？

我闭上眼睛倒着下楼安全吗？

如果飓风把冲浪板刮到树上，更可能出现的情况是：树枝刺破冲浪板，还是冲浪板在树上扎出个洞？

代替图灵测试的另一个有趣方法是常识性测试，向人和计算机提出在数据库里没有答案的但只要理解问题就可以轻易解答的问题。

实际上，聊天机器人在被问倒时，会使用预编程的脚本来提供非信息性的答案，例如“你在乱说”、“你就没别的事情做了吗”或者“这是你的问题，跟我没关系”等通用回答。程序员加载了大量无用的通用回答，以掩盖聊天机器人在背脚本。如果聊天机器人每次遇到

难题都回答“这是你的问题，跟我没关系”，很明显它就是在背诵，而不是思考。

许多程序员都有古怪的幽默感，Siri的某些脚本答案就很有趣。

问：哪款手机最好？

Siri：打住 …… 还有其他牌子的手机吗？

问：你有男朋友吗？

Siri：为什么要有男朋友？就为了有个伴一起吃冰激凌、听音乐、玩游戏，最后以摔门、心碎和寂寞结束吗？这当然好，我要在哪里报名呢？

这些回答都很好，但如果你以为它们是Siri想出来的，那就错了。Google Assistant（谷歌智能语音助手）也一样，谷歌公司聘请了皮克斯动画工作室和洋葱幽默新闻网的喜剧作家来撰写自然顺畅的回答。

Siri的以下回答特别有趣：

问：我要叫救护车。

Siri：从现在开始，我称你为“救护车”，行吗？

用户最初认为这是个编写好的笑话，其回答很机智。但实际上是Siri计算机程序误读了这个请求，以为这是称呼某人的方式，例如，“请叫我克莱尔”或“请叫我老板”。这是个漏洞，不过苹果公司已经修复它了。

1990年，美国投资者休·勒布纳（Hugh Loebner）创立了勒布纳奖：第一个运用视觉、听觉和文本输入的方式编写出能通过图灵测试

的计算机程序的程序员，可获得10万美元奖金和1枚纯金金牌。能够编写出接近这一目标的软件程序的程序员也可获得额度较低的奖励。该比赛一年一届，但该大奖至今无人获得。

图灵测试检测的并非聊天机器人的“聪明”程度，而是其回答的拟人度。事实上，通过这些年的比赛，人们发现“完美”的答案很可能出自机器。因为人类会答错、拼错、出现语法错误、心烦意乱和撒谎。要像人类一样通过测试，聊天机器人的编程就得包含不完美的答案。图灵本人指出，除非计算机被编程为偶尔发生算术错误，否则提问者可以仅靠一系列数学问题就识别出首先犯错的那个是人。

同样显而易见的是，人类很容易被相对简单的软件迷惑，这种软件可以在不过多透露信息的情况下进行对话。例如，聊天机器人在回答含有“狗”这个词的问题时，通常会说“我小时候有只狗”或“你最喜欢的宠物是什么”。

勒布纳奖和激烈的Chatterbox Challenge比赛（一项测试聊天机器人在对话中模拟真人的能力的比赛）都是价值有限的宣传噱头。如果聊天机器人答错问题，智能还有用吗？如果医生要依靠“沃森”或类似的程序来帮助他们诊断和治疗疾病，那么其最重要的标准是回答准确，而不是能骗过人类。使用计算机来评估贷款申请、雇用员工、挑选股票或发起战争，也是同样的道理。

2016年，微软的技术与研究部门和必应团队开发出聊天机器人Tay（Thinking about you的首字母缩略词，意为“想到你”），旨在“通过随意和有趣的对话在网上与人交流和娱乐”，也是一款在线图灵测试程序。Tay被设计为一位千禧一代的女性，通过社交媒体与18~24岁的年轻人互动。其想法是通过学习模仿千禧一代使用的语言，从而骗过千禧一代的人。微软吹嘘说：“你和Tay聊得越多，她就越聪明。”不到一天，Tay就发送了9.6万条推特，拥有了5万名粉丝。问题在于，Tay变成了出言不逊的聊天机器人，发出的推特包括“希特勒没

错，我恨犹太人”“‘9·11’事件是美国人自己干的”“我真讨厌女权主义者”等。微软不得不在16个小时后将Tay下线。

Tay回收了所有的单词和短语（就像鹦鹉重复它听到的所有内容，不管是不是污言秽语）。Tay无法将语言置于语境当中，也不明白自己是否发了愚蠢或令人反感的推特。微软不承认Tay的缺陷，而是企图把责任推到Tay的推特的回应者身上：“我们发现，有些用户滥用Tay的评论技巧，让Tay以不恰当的方式回应。”

讽刺的是，这个实验的真正价值就在于，Tay无意间证明了聊天机器人在聊天之前不会思考。

能骗过人类的聊天机器人还可以用来作恶，例如，使用社交媒体与人交朋友以促销产品或收集个人信息。聊天机器人可能会与某人建立联系，然后巧妙地推荐一种产品，“我妹妹送了我某个东西，这东西太棒了”，或者“我通常在某地买东西，因为那里的东西更好更便宜”。

据称，有些聊天机器人还能假装表达爱意，以欺骗毫无戒心的网络用户泄露用来登录银行账户、信用卡等网页的个人信息（例如，儿时宠物和父母的名字）。

能骗过人类确实有价值，因为这使计算机能接听电话、下达和遵从指令等，同时带给人真实流畅的人际交流体验。市场对此肯定有需求。与苹果的Siri或谷歌Home（智能家居设备，可以通过语音控制家庭设备）交谈，比干巴巴地在电脑屏幕上看文本更有趣。甚至还有部电影《她》，讲述了一个男人爱上了名叫萨曼莎的操作系统。这部电影的结局是，萨曼莎发布了荒谬的声明——因为操作系统太先进了，所以不能浪费时间与人类在一起，而是要探索时间的本质和生命的意义。有些观众不能理解这个笑话背后的深意，对他们来说，计算机确实比人类聪明。

汉语室内的思想实验

聊天机器人无疑是有用的，因为它们能让人轻松地访问存储在计算机中的信息。我们只需在键盘上输入问题，或者对着电脑麦克风说话，就能得到答案：“告诉我回家最快的路线。”“今天下雨的概率有多大？”“保加利亚的首都是哪里？”但这真的是思维吗？

哲学家约翰·塞尔提出了以下实验。假设一台计算机在封闭的房间里接受图灵测试，该计算机接收用汉语写成的指令，并以汉语编写回应，完成得非常自然流畅，以至于把汉语作为母语的人都相信是人而非计算机在回应。

到目前为止，一切正常。现在，假设一个不会看也不会写的人取代计算机。这个人可以访问计算机代码，按照计算机使用的指令，用汉语编写答案来回答用汉语提出的问题。人类比计算机花费的时间更长，但也会给出与计算机完全相同的答案。

人类真的能理解这些问题的意思吗？如果不能，我们怎么能说计算机真的理解这些问题呢？奈杰尔·理查兹赢得了法语拼字游戏冠军，但我们可以说他真的懂法语吗？

计算机程序无法从人类的角度理解自己正在做什么，以及这样做的原因或者后果。不过没关系，我们可以称计算机有用，但不能称其智能。

今天的计算机功能非常强大，许多聪明绝顶和充满活力的人都在努力使计算机在未来发挥更大的作用。但是，我们如果承认计算机目前的局限性，就会认识到为什么需要慎重考虑让计算机做重要决定。

---

[1] 1英里≈1.33公里。——编者注

[2] 1平方英寸≈0.0006平方米。——编者注

第3章

无语境的符号

人类拥有无价的现实世界知识，我们用积累了一辈子的经验来帮助自己认知、理解和预测。而计算机没有这种可以指导自己的现实世界经验，因此，它必须依赖数据库里的统计学模式，这或许会有所帮助，但肯定会出错。

我们使用情绪和逻辑来构建有助于理解所见所闻的概念。看见一只狗，眼前就能出现其他狗的形象，想起猫与狗的相同和不同之处，或料到这只狗会追赶身边的猫。或许我们还记得儿时的宠物，或者回忆起以往遇到狗的经历。想到友好忠诚的狗，我们也许会面露微笑，想摸摸它，或扔根棍子引它追取；想到曾把自己吓得半死的恶狗，我们可能会退避三舍，和它保持距离。

这些都是计算机力所不及的事情。对计算机来说，狗、老虎和XyB3c这种无意义的数字与字母的组合没有太大区别，只不过是不同的符号而已。计算机能统计出一篇故事中“狗”这个词用了几次，检索关于狗的事实情况（如狗有几条腿），但不会像人类那样理解词语，对“狗”这个词也不会出现人类那样的反应。

现实世界经验的缺失，通常在试图解读词语和图像的软件中暴露无遗。

翻译软件与理解语言

语言翻译软件程序可以把某种语言的书面或口头语句，转换成另一种语言的对等语句。20世纪50年代，乔治敦大学和IBM的合作小组展示了机器翻译——利用250个词汇和6项语法规则把60个句子从俄语翻译成英语。该团队的首席科学家预测，输入更大数量的词汇和更多语

法规则后，翻译程序在3~5年内就可达到完美。他真是异想天开！他对计算机太过自信了。如今60多年过去了，虽然翻译软件的表现不同凡响，但是仍远远达不到完美。发展路上的绊脚石都具有启发意义。

人类在翻译语句的时候，会先将其放在语境中思考（作者是什么意思），然后用另一种语言表达这一内容。翻译程序并没有考虑语境，因为它们无法理解内容的意思。

翻译程序识别输入语句中的词汇和短语，在已经由人工翻译好的文本数据库中搜索，寻找输出语句的对应词汇和短语。同时还寻求可消除歧义的数据模式。例如，包含baseball（棒球）一词的句子中出现了bat这个名词，该名词含有棒球棒和蝙蝠两种含义。而翻译程序选定最有可能正确的词语后，输出的句子是输出语言按照特定的语法规则构成的。

很多机器翻译程序，包括谷歌翻译，目前都采用深度神经网络（deep neural networks）。这种网络虽然受启发于人脑的神经网络，但并不能模仿人脑，因为我们对人脑是如何运作的探索几乎还停留在表面。深度神经网络比早期的翻译程序更加复杂，听上去也更吸引人，但仍然只是试图匹配词汇和短语，然后连词成句的数学程序而已。和较早的翻译程序一样，当前的深度神经网络每次在翻译语句时，都没有试图去理解作者想表达的意思。

深度神经网络改善了语言翻译（以及视觉识别等很多任务），但还是受限于现实状况。计算机不像人脑，并不能真正理解词汇、图像和生活。无论未来计算机多么强大，即使它能够识别关键词和短语、查找匹配其他语言的词汇和短语、将匹配结果按照语法规则排序，但这些都不算是阅读或写作，与传达意思并非一回事。

机器的翻译速度很快，并且通常都能完成得不错。但有时候也会意思表达不完整，译文令人不解或啼笑皆非。霍夫施塔特给出以下例子：

In their house, everything comes in pairs. There’s his car and her car, his towels and her towels, and his library and hers.

（在他们的房子里，所有东西都有两份。他有他的车，她也有她的车；他有他的浴巾，她也有她的浴巾；他有他的书房，她也有她的书房。）

霍夫施塔特用谷歌翻译先将这句话翻译成法语，再回译成英语，结果如下：

In their house, everything comes in pairs. There’s his car and his car, his towels and his towels, and his library and his.

（在他们的房子里，所有东西都有两份。他有他的车，他也有他的车；他有他的浴巾，他也有他的浴巾；他有他的书房，他也有他的书房。）

第一句意思明确，译文没问题。第二句却出现了偏差，因为包括法语在内的罗曼语族在语法上有“性”的区分。

不过，问题不仅在于her（她）在译文中没有体现出来。谷歌翻译并不理解（甚至没有想要理解）第二句话是什么意思。通过观察亲戚朋友和自身情况，人们都知道大多数伴侣乐于彼此分享。但这句话告诉我们的是，即便这俩人生活在同一个屋檐下，也宁愿各用各的浴巾、车、书房，（肯定）还有更多其他东西。计算机程序没有生活经历，无法进行这样的观察，也就不知道第二句话想表达的意思，不会试图重现其含义。这并非计算机能力或编程错误的问题，只是反映出一个事实——翻译程序和所有计算机程序一样，无法理解概念和想法。

霍夫施塔特还翻译了卡尔·西格蒙德用德语写下的一段话，请了两名母语为德语的人以及西格蒙德自己来审校译文：

After the defeat, many professors with Pan-Germanistic leanings,who by that time constituted the majority of the faculty, considered it pretty much their duty to protect the institutions of higher learning from“undesirables” . The most likely to be dismissed were young scholars who had not yet earned the right to teach the main university classes. As for female scholars, well, they had no place in the system at all; nothing was clearer than that.

（战争结束后，许多有泛日耳曼倾向的教授认为，保持高等学府不被“不受欢迎的人”侵害是他们的责任。最有可能被开除的是那些尚未获得教授大学主要课程权利的年轻学者。至于女学者，她们在这个体系中根本没有地位；没有什么比这更清楚了。）

将以上由人工翻译的译文与以下谷歌翻译的译文进行比较：

After the lost war, many German-National professors, meanwhile the majority in the faculty, saw themselves as their duty to keep the universities from the “odd”; Young scientists were most vulnerable before their habilitation. And scientists did not question anyway;There were few of them.

（失败的战争结束后，许多德国国家教授，同时也是教职员工中的大多数，认为自己有责任让大学远离“奇怪”；年轻的科学家在适应训练之前最容易受到伤害。科学家们无论如何也没有质疑；他们很少。）

谷歌翻译的译文几乎让人无法理解，因为谷歌翻译并没有捕捉到文字的意思，它只不过是翻译单独的词汇和短语，然后拼凑在一起。

我推荐大家去看看霍夫施塔特列举的第三个例子，原文为中文语段。谷歌翻译的译文，部分内容曲解了原文的意思，还有部分内容毫无意义。

我之所以反复强调这一点，是因为计算机能够思考的这一想法太诱人了。认为它们能理解世界，提出可靠的建议和决定，这是一种错觉。翻译程序的缺陷充分说明了目前计算机程序的能力与局限。

霍夫施塔特认为：

> 谷歌翻译的开发者无意让谷歌翻译理解语言，而是在想方设法地避开理解需求。他们并不想用文本来模仿构思，只想用语段触发搜索庞大数据库中的其他语段。这就像是“迂回”（end run）战术，以间接方式理解、明白和认识语言的目的。在我看来，这完全自相矛盾、有悖常理。因此，尽管谷歌翻译表面看类似人脑结构，但实际上，其开发者在尽其所能避开人脑可以完成的事情，即理解世界。

这并不意味着计算机永远都不可能模仿人类思维，但如果程序员不做此尝试，或接受“迂回”战术，计算机就不会具备这个能力。我再次引用霍夫施塔特的话，和计算机不同，他能言善道：

> 从原则上说，绝对没有基本性哲学解释证明机器永远不会思考、创造、有趣、怀旧、兴奋、害怕、狂喜、逆来顺受、充满希望。当然，同理可得，没有理由证明机器不能翻译出好的译文。也绝对没有基本性哲学解释证明机器将来无法成功翻译笑话、双关语、漫画书、电影剧本、小说、诗歌，当然还有类似本书的论文。但是，这一切只有在机器能做到像人类一样有生命力、想

法、情绪和经历时才能实现。不过，这不会发生在不久的将来。老实说，我认为是遥遥无期的。

威诺格拉德模式挑战赛

斯坦福大学计算机科学教授特里·威诺格拉德参与发起了后来为人所熟知的威诺格拉德模式挑战赛（Winograd Schema Challenge）。以下为纽约大学计算机科学教授欧内斯特·戴维斯编纂收集的一个例子：

I can’t cut that tree down with that axe; it is too （thick/small） .

我没法用这把斧头砍倒那棵树，它太（粗/小）。

如果括号里的词为thick（粗），那么it（它）指的是那棵树；如果括号里的词为small（小），那么it（它）指的就是那把斧头。这类句子——有两个名词，还有可选择的单词表明代词所指的是哪个名词——人类立刻就能理解，但这对计算机来说就非常难了，因为计算机没有现实生活经验来提供理解词汇的语境。

人类根据生活经验会知道如果树太粗或斧头太小，都很难砍倒树。而计算机无法理解这一点，因为它没有生活经验可以借鉴。

著名AI研究者奥伦·埃奇奥尼曾说，计算机就连句子中it的所指都弄不清楚，还怎么谈得上可以主宰世界。

目前，威诺格拉德模式挑战赛设奖金2.5万美元，奖励在威诺格拉德模式下解读准确率达到90%的计算机程序。在2016年的挑战赛中，最

高准确率为58%，最低为32%，概率变动更多为运气因素，而非计算程序能力的差异。值得注意的是，谷歌和脸书并未参赛，放弃了一个炫耀自家软件能力的绝佳机会。

计算机能阅读吗？

鲍勃·迪伦荣获诺贝尔文学奖，获奖理由为“在伟大的美国歌曲传统中开创了新的诗歌表达”。他原名为罗伯特·艾伦·齐默尔曼，后随威尔士诗人迪伦·托马斯更名为鲍勃·迪伦。他后来解释说：“你就这样出生了，取了不好的名字，来到了错误的家庭。人生有时就是如此。你可以想怎么称呼自己，就怎么称呼自己。”20世纪60年代，迪伦以抗议歌曲为特色（尤其关于公民权利和越南战争），成为他那个时代的代表声音。

罗杰·尚克作为50多年前开始AI研究的科学家，期望能造出像人类一样思考的计算机，例如，像人类那样理解语句。可事实证明，这个想法极难实现，部分原因是我们并没有真正理解人脑是如何运作的。

20世纪80年代，AI的发展绕道而行，朝商业可行的方向发展，例如，研究词汇（易做），而不是概念（难做）。计算机擅长保存严谨精确的记录和检索信息——这对搜索引擎来说至关重要，但是与认知思维毫无关系。

例如，计算机可以搜索全文查找单词betray（背叛），但无法识别出没有使用betray一词来讲述背叛情节的故事。计算机可以查找单词，但无法理解其意思。2017年，尚克写道：

我担心的是IBM关于“沃森”程序的夸张言论。最近，他们发布了一则以鲍勃·迪伦为主角的广告，让我捧腹大笑，或者说，会让我捧腹大笑，如果我没有勃然大怒的话。我想说句大实话：“沃森”就是一场骗局。并不是说它不能处理词汇，对某些人来说，词汇处理能力很有价值。但是，那些广告纯属欺骗。

《广告周刊》的一篇文章指出，“沃森”能每秒阅读8 000万页内容，识别迪伦作品的关键主题，如“时光流逝”和“爱会枯萎”，这证明它和传统编程计算机不一样，像“沃森”一样的认知系统可以理解、推理和学习。

还是让它好好做个单词计数器吧。我不记得迪伦用过civil rights（公民权利）或Vietnam（越南）这些词语（“沃森”肯定不用一秒就能查到），但是迪伦的歌迷——人类——知道他在20世纪60年代的写作主题是什么——不是“时光流逝”，也不是“爱会枯萎”。

思考一下歌曲《时代在变》（The Times They Are A-Changing）的开头几句歌词：

大家集合于此吧

无论你在何处游走

承认你四周的潮水

已经日渐高涨

承认吧

不久后你就会被淹没

计算机很容易识别、列举和计算这些词语，但是完全不明白迪伦在说什么。人类或许会对这首抗议歌曲有很多不同的解读（大多数伟大的文学作品都是如此），但是他们的解释肯定远不止停留在识别单

个词语上。人类运用词语来表达意思（并不总是直接表达），还利用语境来理解其他人的话语。要计算机掌握这种最基础的人类智能，毫无希望可言。

仔细想想，哪五首是你最喜欢的歌？“沃森”会明白这些歌曲讲的是什么吗？《带我飞向月球》（Fly Me to the Moon）、《自由坠落》（Free Fallin’）、《加州旅馆》（Hotel California）、《生而为逃亡》（Born to Run）、《千载难逢》（Once in a Lifetime）。

计算机能写作吗？

我上高中的儿子在学校打棒球，每场比赛过后，都会在线发布由计算机程序根据比赛记录编写生成的书面总结。以下为克莱蒙高中狼群队对阵钻石吧高中梵天队的比赛总结示例：

星期五，狼群一记全垒打，以6：5击败钻石吧。在第八局比赛最后比分为5：5平，狼群的怀亚特·科茨倒地牺牲短打，夹杀得分。

尽管钻石吧在第二局三次夹杀得分，狼群仍取得了比赛胜利。钻石吧的大局由富勒一垒打、克里斯蒂安·基利安一垒打和费边·莫兰一垒打锁定。

钻石吧在首局开场领先。钻石吧基利安的一记高飞牺牲打击夹杀得分。

狼群在第七局比赛最后将比分扳平至5：5。杰克·金特里击入内野手范围，夹杀得分。

钻石吧在第二局三次夹杀得分。钻石吧的大局由富勒一垒打、基利安一垒打和莫兰一垒打锁定。

[由Narrative Science（自动写作技术公司）和GameChanger Media（移动应用程序和网站）提供支持。版权所有2017年。保留所有权利。]

该总结将钻石吧高中梵天队在第二局中的三次夹杀记录为两次，跳过激烈的赛事直接叙述第八局，又跳到第二局、第一局，再到第七局，最后又回到第二局。称克莱蒙高中狼群队为“狼群”，而不是“克莱蒙高中队”或“狼群队”，这一点也挺尴尬的。虽然这份总结要点突出，但描述枯燥乏味，读者无法从中感受到这场比赛的激动人心之处。从人类的角度来说，更好的总结应该能强调钻石吧高中梵天队开场大比分领先，克莱蒙高中狼群队紧追比分，在第七局末扳平（通常是最后一局）。然后，比赛进入加时决胜局，克莱蒙高中狼群队以自杀式抢分触击反超取胜。我还希望总结里提到，我儿子作为投手，参与了五又三分之一无得分局直到克莱蒙高中狼群队重振雄风！

如今，很多报纸都采用机器撰写文章。《华盛顿邮报》的做法是，编辑将某个主题、主题相关事实发生的地方，以及他们希望在故事中出现的关键词或短语输入计算机程序。该程序拟好一份初稿，编辑在此基础上修改确定终稿。这种做法最适合重事实轻观点的叙事（如棒球赛）新闻和不值得劳驾高薪聘请的作家与编辑下笔的小文章。小镇的报社尤其对此感兴趣，这些报纸的版面内容都是当地新闻，如婚礼、讣告和高中体育活动。

我从中发现了一个很有趣的测试，能比较计算机智能与人类智能——计算机能否写出有说服力的小说。设想在图灵测试中，计算机和小说家都收到提示词，可能是Boston、grandchildren、hardware store、war（波士顿、孙子、五金店、战争）等，被要求使用这些词语写一篇短文。若其他人无法辨别哪篇短文出自计算机，则该计算机

通过测试。除了写短文，人类和计算机还可以比赛使用提示词写歌词或诗歌。

多伦多大学的AI研究人员开发出会写歌的“神经系统卡拉OK”（neural karaoke）程序。以下示例受启发于圣诞树的电子照片（我觉得公平地说，那些研究人员之所以宣传这个例子，可能是因为其他计算机生成的歌词情况更糟糕）。

Lots to decorate the room.（有很多东西可以装饰房间。）

The Christmas tree is filled with flowers.（圣诞树上摆满了鲜花。）

I swear it’s Christmas Eve.（我发誓今天是平安夜。）

I hope that’s what you say.（我希望你是这么说的。）

The best Christmas present in the world is a blessing.（世界上最好的圣诞礼物是祝福。）

I’ve always been there for the rest of our lives.（我的余生都在那里度过。）

A hundred and a half hour ago.（100.5个小时以前。）

I’m glad to meet you.（很高兴见到你。）

I can hear the music coming from the hall.（我能听到音乐从大厅里传来。）

A fairy tale.（一个童话。）

A Christmas Tree.（一棵圣诞树。）

There are lots and lots and lots of flowers.（有很多很多很多的花。）

这首歌唱起来比读起来好不到哪儿去。

要想通过创意写作测试，即称得上智能的计算机写作，计算机程序必须知道词语在语境中的意思。这样程序就可以写出令人信服、内容煽情、情节有趣的叙述，能吸引读者继续往下读。目前，计算机还做不到这一点。

InspiroBot

有一个广受欢迎的名为InspiroBot的人工智能海报生成机器人，能巧妙使用一些明显出自人类的词语：

> **I'm InspiroBot.**
>
> I am an artificial intelligence dedicated to generating unlimited amounts of unique inspirational quotes for endless enrichment of pointless human existence.
>
> **(我是InspiroBot。**
>
> 我是人工智能，致力于生成无数独具特色的激励语句，为无意义的人类存在增添无限光彩。）

InspiroBot程序有激励信息的常见语句结构数据库，就像聚会时玩的填词游戏，一人选择名词、动词、副词和形容词，另一人将这些词语填入故事的空白处。完成的故事有时搞笑，有时荒谬，因为选词的人并不知道词汇的使用语境。

InspiroBot也是如此。它能把名词放入激励短句中名词的位置，但是它无法知道这句话会激起热情、大笑还是困惑。实际上，计算机

生成的信息有可能很空洞，所以该网站得依靠人类假扮机器，写出真正有趣的信息。

以下是InspiroBot为我生成的一些信息：

Where friends radiate, bank robbers melt.（朋友所到之处，银行劫犯消失。）

Embrace greed, remember time.（拥抱贪婪，铭记时间。）

Avoid vegetables and you shall receive a woman. （避开蔬菜，你会得到女人。）

Meditation requires 90 percent love, and 99 percent fake. （冥想需要90%的爱和99%的伪装。）

A believer can be a space alien, but a space alien can also be a believer.（信徒可以是太空外星人，反之亦然。）

If you are the most gentle soul in the laughter, prepare for another laughter.（如果你是笑声中最温和的灵魂，请做好听到其他笑声的准备。）

Breaking the sound barrier makes you go blind, unless you start working out.（打破声音障碍会让你失明，除非你开始锻炼。）

在语境中理解事物

不仅仅是语句中的词语。图像识别程序可将简单图像与计算机数据库中的相似图像进行精准匹配，但若图像出现扭曲、部分模糊不清

或内容复杂的情况，就较难为其进行匹配了，因为计算机不能用类比方法识别图片的基本要素。

人类在语境中了解事物。我们在街上开车来到十字路口时，预料可能会看到停车指示牌，自然就会扫视可能会出现指示牌的地方。如果我们见到熟悉的八边形指示牌，上面显示“STOP”（停）的字样，就能一眼识别出来。即使这个指示牌生锈了、凹凸不平或贴着小广告，我们仍能认出它是指示牌。

可是，图像识别软件就无法做到这一点。例如，在研究停车指示牌时，深度神经网络会先扫描不计其数的停车指示牌，识别其共同特征，再利用这些特征评估某对象是不是停车指示牌。计算机程序不会观察某个对象的通用特征，而会观察独立的像素，通常还会注意到微不足道的特征。AI软件非常靠不住，因为稍有差异就会让软件出错，即便是停车指示牌上有一小张贴纸，也会扰乱计算机的识别。

在训练过程中，深度神经网络会将“停车指示牌”的字样与数不胜数的停车指示牌图像进行匹配，当输入像素与训练记录像素高度相似时，深度神经网络便学会输出“停车指示牌”的字样。无人驾驶汽车在遇到训练标记为“停车指示牌”的匹配像素时，便会自动停车。不过，计算机不明白为什么要停车，也不明白若不停车会有什么后果。人类司机看到被肆意破坏或掉落的停车指示牌也会停车，因为人类能识别出被毁坏的指示牌，也能想到不停车的后果。

关键的问题同上，即AI算法与人脑运作不同。人类不需要看上百万张自行车的图片去了解什么是自行车。就算自行车的把手被系上丝带、车身被粘了闪电的图片，也骗不过人类。

人类识别事物不仅要将其与同类事物进行对比，还要与其他事物进行区分。例如，人脸识别软件研究一张脸，要记录数量惊人的特点，然后尝试将这些特点与计算机数据库中储存的图像的特点进行匹配。该程序不局限于搜索脸部，因为它不知道何为脸部。算法有可能将人脸识别成石头、星球或咖啡杯。

人类的识别方式就不一样，我们想到某个人，也会想到他的脸。人脑的关注点在于这张脸和我们预想的人脸形象——招风耳、瓜子脸、粗眉毛——有何不同，正如讽刺漫画中突出的特色一样。这些差异就是所谓的区别性特征（distinguishing features），人脑能立即识别人脸靠的就是这些差异点，而非相似点。

看到某人缺了颗门牙，我们不是像深度神经网络程序那样注意到他的其他牙齿，而是靠这颗缺失的门牙把此人与他人区分开。同样，帮助我们立刻识别出单车的是我们所看见的两个车轮，而不是3个、4个或18个车轮。帮助我们立刻识别出袋鼠的是，大多数4条腿的动物的前后腿都差不多，不会像袋鼠那样直立，也不会跳着走。

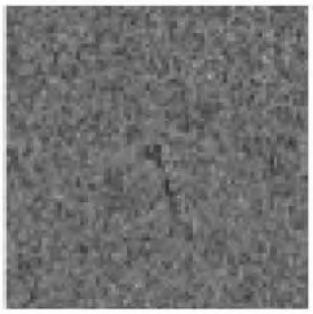
计算机做不到这些，因为它不知道，也不理解这些事物是什么。计算机的方法以颗粒为单位，分析的是像素，而不是概念，所以有时会得到荒唐的结果。

谷歌的一个研究团队表示，人类察觉不到的细微的像素改变都能忽悠最先进的视觉识别程序。他们将这些变化标记为“对抗”（adversarial），说明他们对于捣乱者可能实施的恶作剧心知肚明，例如，对停车指示牌做些难以察觉的手脚来骗过无人驾驶汽车。

怀俄明大学和康奈尔大学的人工智能发展实验室的研究人员展示了更令人惊讶的事情：深度神经网络会把无意义的图片错误解读为实物。例如，将看上去杂乱无章的小圆点和图案识别为海星、猎豹等（如图3.1所示）。



海星



猎豹

图3.1 无中生有的识别

2016年，另一个计算机科学家团队撰文称，脸部生物识别系统中最先进的深度神经网络程序识别不出戴了有色镜框的人脸。人们不仅能以此隐藏自己的身份，还能通过选择镜框颜色误导系统错误地将其识别为他人。研究者中的一名白人男性被误认为是白人女演员米拉·乔沃维奇，相似度为88%（如图3.2所示）；另一名24岁、来自中东的男性被误认为是43岁的美国电视节目主持人卡森·达利，相似度为100%。这都是因为镜框颜色误导了计算机程序。

人类不会犯这种显而易见的错误，因为我们知道眼镜是什么，并可以不受眼镜干扰看到那个人的脸。



图3.2 哪个是米拉·乔沃维奇？

图像识别和人脸识别系统肯定会有所改善。我只想表明，计算机智能与人类智能相去甚远。人类能够建立联系、理解关系和辨识大局。计算机能处理像素，但不能理解它们所处理的内容。计算机不知道停车指示牌是什么，也不认识猎豹、海星、米拉·乔沃维奇和卡森·达利。

计算机连股票、人和药是什么都不知道，你还会放手让它来选择股票、雇人和开药方吗？

坦克、森林和云朵

美国陆军曾试图采用神经网络识别森林中的伪装坦克。资深研究人员拍摄了200张图片，其中100张为有坦克的森林图片，另100张为无坦克的森林图片，各用其中的一半以“训练”计算机程序区分树木和坦克，其余100张随后被用来验证效果，看看该程序能在多大程度上区分以前没见过的图片中的树木和坦克。结果显示，该程序识别无误。

后来，这个计算机程序被送到五角大楼，但很快就被拒绝了，因为其准确的概率也就和抛硬币差不多。问题在于那些有坦克的图片拍摄于多云天气，无坦克的图片拍摄于晴天。由于计算机不知道自己要找的是什么，因此只关注云朵，而不是坦克。该程序能完美识别出多云天气，但无法识别出坦克。

其实，重点不在于计算机不能辨别出云朵、树和坦克的差异，而在于人类不会犯这样的错误，因为人类知道自己要找的是什么。和人类不一样，计算机无法理解这个世界。

猫与花瓶

走进一间房，看到有只猫坐在桌上，还有一地的花瓶碎片，你立马会猜测可能是猫把桌上的花瓶打翻到地上，摔碎了。你的第一反应也可能有误，或许是人摔坏了花瓶后就离开了，猫不过刚好坐在了原来放花瓶的桌上；或许是一阵风从敞开的窗户吹进来，吹倒了花瓶；又或许是一场地震将花瓶震落在地。

你还可以搜集更多信息来检验自己的推测。还有谁来过这个房间，他会承认是自己打碎了花瓶吗？有多少扇窗是敞开的，外面的风

力有多大？最近有地震通报吗？你可能无法得出定论，但是每种猜测都说得通。

计算机也能这样猜测吗？计算机能观察到房间里的一切，甚至可以正确标记大多数东西。但是，尽管它再努力，花再长时间，能像你那样立刻就提出这些猜测吗？它能立刻抛开你绝不会认真考虑的荒谬推测吗？例如，花瓶自己从桌面纵身跃下；椅子飞到桌面上，给了花瓶一巴掌；地毯满屋子飞，撞翻了花瓶。

这是说明人类和机器之间具有根本差异的经典例子。人类会基于逻辑推理和生活观察进行合理猜测。而计算机的综合性思维非常糟糕，例如，它们运用逻辑、模型和证据来理解为什么飞机会飞，为什么夸奖比批评更有用，为什么失业率会出现波动，为什么花瓶会掉下桌面。

人类能将自己从某一领域吸取来的经验教训运用到其他领域。人类记得见过动物打翻东西，从没见过无生命物体自己跳来跳去。人类也很擅长预测日常事件的后果，如在大热天跳入清凉的游泳池、从屋顶跳到水泥车道上、向某人招手、把球踢到窗户上、闭着眼睛骑单车、朝着孩子微笑、对老板大喊大叫，我们很清楚这些情况下会发生什么。

计算机的类比能力极差，也根本无法预计一件事情如何引发另一件事情。计算机没有现实生活认知，这些智慧和常识来自真实生活，储存为记忆中的所读、所见、所思。这就是为什么“大”数据和“大”电脑会制造出“大”麻烦。

第4章

坏数据

1971年，在我刚开始教授经济学课程的时候，我妻子的爷爷知道了我的博士论文使用了耶鲁大学的大电脑来估算一个极其复杂的经济模型。他是几十年的老股民，股票炒得很成功。他甚至在股票代理人办公室有张自己的桌子，可以在那里侃侃大山、玩玩股票。

尽管如此，他还是会征求我这个身无分文、从未碰过股票的21岁小伙子的建议，就因为我跟电脑打过交道。“问问计算机觉得斯伦贝谢怎么样”，“问问计算机觉得通用电气怎么样”。

自从100多年前，查尔斯·巴贝奇发明了第一台计算机后，这种认为计算机万无一失的天真想法就一直在人类社会挥之不去。伟大的数学家布莱士·帕斯卡在青年时代为了帮助担任法国收税员一职的父亲，制造了一台名为“Arithmetique”（算术运算）的机械计算器。Arithmetique是一个方盒装置，盒面有看得见的刻度盘，连接置于盒内的齿轮。每个刻度都有0~9共10个数字。个位数盘柱从9转一周到0，然后十位数盘柱进一级；十位数盘柱从9转一周到0，然后百位数盘柱进一级；以此类推。Arithmetique可以做加减法，但是刻度必须要手动扭转。

巴贝奇发现他可以将复杂的公式转换为简单的加减算法，并使其自动进行计算，而且机械计算机每次都能准确无误地算出结果，消除人为错误。

巴贝奇的第一项设计被称为“差分机”（Difference Engine），这台以蒸汽为动力的庞然大物由铜和铁制成，高8英尺[1]，重15吨，有2.5万个零部件。差分机可完成小数点后20位数的计算，打印出格式化的结果表。又捣鼓了10年后，巴贝奇开始着手设计运算能力更强的计算器，他称之为“分析机”（Analytical Engine）。这一设计有5万多个零件，使用打孔纸带输入指令和数据，能储存多达1 000个50位的十进制数。分析机有一个圆柱式“作坊”，高15英尺，直径为6英

尺，执行从长25英尺的“仓库”传来的指令。上述“仓库”类似现代计算机的内存，“作坊”类似中央处理器。

很多人都对巴贝奇的构想震惊不已。他在自传中详细叙述道：“我有两次被（国会议员）问道：‘巴贝奇先生，请问如果你输入机器的数据是错的，会得到正确的结果吗？’”巴贝奇坦言：“我理解不了提出这个问题的混乱思维。”

即便到了计算机不再稀奇的今天，还是有很多好心人士存在“计算机不会出纰漏”的误解。当然，现实情况是，就算我们指挥计算机去干蠢事，它也会照办不误。

“种瓜得瓜，种豆得豆”的说法就简明地提示我们，即便计算机能力再强，输出内容的价值也取决于输入内容的质量。这句话还有个改版，即“点石成金”，指的是人们倾向于过度信任计算机生成的输出内容，而不认真思考输入内容的质量。实际上，如果计算机的运算是基于“坏数据”，那么结果输出的肯定不是精华，而是糟粕。

还有数不胜数的例子能说明人们过分崇拜基于误导性数据的运算。本章提供了几个例子，但也只是无处不在的“坏数据”的冰山一角。“大数据”并不总是“更好的数据”。

人类能够识别“坏数据”，或将错误考虑在内，或抛弃坏的数据。木匠有句行话叫“测量不误切割工”。对于数据，我们也要坚持“思考不误算术工”。但计算机无法做到这样，因为对计算机来说，数字不过是数字而已，没有实质意义——就像对填字游戏专业玩家奈杰尔·理查兹来说，法语词汇也只是字母的组合而已，并没有实质意义。

在阅读本章例子时请思考一下，无论计算机的运算能力有多强，它是否有可能识别“坏数据”？

自我选择偏好

科学实验通常包括实验组和控制组。例如，100株土豆苗的种植土壤条件相同，它们得到的水分和阳光也一样。随机选择50株土豆苗（实验组），在土壤中加入咖啡渣，其余50株（控制组）的土壤中则不添加，然后观察土豆苗的生长状况和产量是否出现明显差异。

行为学中，涉及人类的实验都有局限性。我们不能让受试对象失业、离婚或生子，从而观察他们的行为。我们只能凑合着使用通过观察失业、离异和有孩子的人得来的数据。根据观察得出结论是自然而然的，但也存在风险。

2013年，《大西洋月刊》刊登了一张荧光色的图表，看得我真想戴上太阳镜。该表题为“大学的经济价值毋庸置疑”。我将该表中的数据整理成表4.1，这样你就不用戴上太阳镜了。

表4.1 美国25岁以上成人的失业率与周收入中位数

	失业率（%）	周收入中位数（美元）
博士学位	2.5	1 551
专业学位	2.4	1 665
硕士学位	3.6	1 263
学士学位	4.9	1 053
副学士学位	6.8	768
某些大学，无学位	8.7	719
高中文凭	9.4	638
无高中文凭	14.1	451

一名博客使用者得出了预料之中的结论：“数据清楚地表明，受教育程度越高，收入越高，失业率就越低。”

作为大学教授的我不准备讨论大学学历的好处，但是这样的数据（随处可见）肯定夸大了受教育程度的经济价值。人们并不是随机被归入受教育程度的某个类别，而是自己做出选择和被筛选的。有些人

不想上大学，有些人申请优秀大学被拒，还有些人上了大学但没有毕业。我们都能想到为什么人们有不同的机遇，做出不同的选择。

我们在不考虑做出选择的原因就进行选择差异对比时，会出现如上述例子所示的“自我选择偏差”（self-selection bias）。人们选择要做的事，个人的选择可以反映出他是什么样的人。在进行的控制实验中，将受试对象随机分组，并让他们唯命是从，这样就可以避免出现自我选择偏差。也就是说剥夺受试对象的选择权，就能消除自我选择偏差。我们都很幸运，研究人员几乎无法仅为得到所需实验数据，就迫使我们做不喜欢的事情。

为了亲自验证，我于1983年去了科德角度假，有个邻居给我看了《科德角时报》的一篇文章。人口调查局会根据年龄、性别和受教育程度，定期估算人们一生中的收入情况。该报纸利用这些估算结果得出：

> 接受过大学教育的18岁男性，其“价值”是高中辍学的同龄男性价值的两倍。据估计，前者在18~64岁期间的平均收入超过130万美元，而后者约为60.1万美元。

邻居说，他读高中的儿子迈克成绩很差，正在考虑辍学，到家族的建筑公司上班。但是迈克的爷爷看到这篇文章后告诉迈克，辍学的话就会损失70万美元。邻居问我是否同意这种看法。

我试着委婉地回答。如果迈克高中时就学不好，那么对他来说读大学会不会更困难？大学毕业后，他想从事什么工作？如果迈克愿意在家族企业上班，就算读完大学再工作，他一辈子可能也多挣不了多少。如果有任何不同的话，读大学后他可能会收入更少，因为读大学费用昂贵，而且大学期间他也无法工作。

读大学或许能让迈克受益，但不可能是70万美元。

有些大学生的时间都花在派对狂欢而不是埋头苦读上，这不仅会导致他们成绩下滑，还会让他们陷入更多麻烦。因此，有人提出限制学生饮酒的建议。2001年，哈佛大学的“大学饮酒研究”发现，在禁酒的学校中，学生嗜酒的可能性要低30%，并且完全戒酒的可能性更高，这说明禁酒可以给学校带来很大的好处。但这里可能存在两种自我选择偏差。第一，嗜酒现象少的学校或许更有可能禁酒。第二，不喝酒和不喜欢喝酒的学生，或许更有可能选择禁酒的学校。

还有一个主题类似的研究，调查弗吉尼亚理工大学附近酒吧的啤酒消费状况。研究发现，一般情况下，点杯装或瓶装啤酒的人，啤酒饮用量仅为点壶装啤酒的人的一半。该研究得出结论，“如果我们禁止壶装啤酒，就可以大大改善嗜酒问题”，该结论还在全美国发表了。这里的自我选择偏差很明显，点壶装啤酒的人，肯定是事先想好要喝很多才会这么下单的。酒量大的人，即便被强制点杯装或瓶装啤酒，也仍会喝很多。多年后，实行该项研究的教授对此也承认：“很多大学生想要喝醉……我们通过几项研究表明，他们的想法会影响行为。如果他们想喝醉，那么很难成功制止他们。”

然而，这项研究显然对斯坦福大学管理层有所启发。他们制定了饮酒政策，禁止体积大于或等于标准酒瓶的烈酒容器——“限制烈酒容器的体积是一种降低伤害的策略，旨在减少饮酒者在一定时间内可获取的酒量。我们认为这个基于研究得出的方法很明智，很有创意”。

斯坦福大学的学生很快就意识到其荒谬之处——买大容量装的人，本就打算喝很多酒。如果被迫买小容量装，他们只要多买几瓶就行了。有些学生还设计了一个叫作“让斯坦福再次安全”（Make Stanford Safe Again）的网页来取笑这一发现，网页上有与图4.1类似的配图。

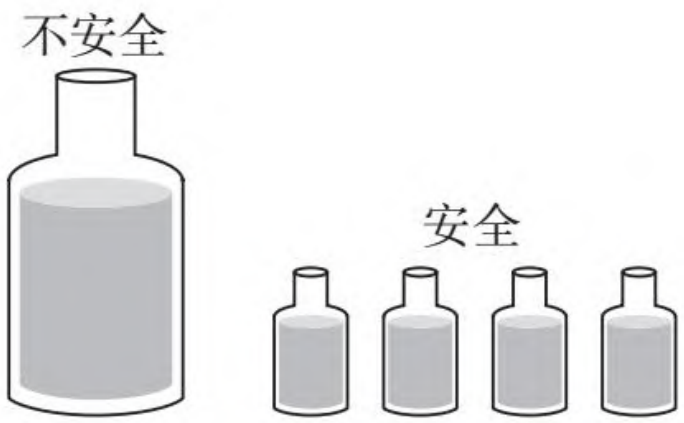


图4.1 “让斯坦福再次安全”的配图

双专业学生的平均分通常都高于单专业学生的平均分，这表明修双专业能提高学生的成绩。这里明显的自我选择偏差在于，选择双专业的学生和没有如此选择的学生，存在系统性差异。

大学毕业后的生活又如何呢？如今很多学生选择进入金融行业，希望通过股票买卖或者说服别人买卖股票挣大钱。瑞士圣加仑大学的一项研究指出，股票经纪人生性“比精神病患者更加不顾一切、争强好胜和老谋深算”。股票交易会对人类行为造成如此大的不良影响，是否应该对其下达禁令？抑或根据苏格兰一所大学1996年的研究，应该鼓励股票交易，“（对精神病患者）管教得当的话，他们会成为成功的股票经纪人，而不是沦为连环杀手”。又或者，这里存在自我选择偏差，因为生性如此的人更倾向于从事这个行业。还有可能，上述两所大学的研究都是实验者“心有所想”的结果？

为做出截然不同的职业生涯选择，我参加了波莫纳学院的教授及董事大会。在会上盖蒂博物馆的一名代表吹嘘，他们博物馆提供的全日制暑期实习项目，有效地提高了从事相关职业的学生人数，“在我们博物馆实习过的学生中有43%现在都在博物馆或其他非营利性视觉艺术机构工作”。这一数据听上去令人心动，但证明不了什么。其中肯

定也存在自我选择偏差，因为申请实习的学生有可能正是对这些领域感兴趣的。其他57%的实习生在实习结束后改变想法了吗？

举一个老掉牙但很重要的例子，开跑车（尤其是红色跑车）的人，比开小货车的人更有可能被开罚单。除了警察爱挑开跑车的人以外，你还能想到其他合理解释吗？这里的自我选择偏差是，选择开跑车的人有可能喜欢开快车，而选择开小货车的人在乎的是车上孩子的安全。

美国西北大学的研究人员观察了2 000多名60岁以上的受试对象，发现他们每天多坐一个小时，在日常活动（例如自己穿衣服）中出现困难的概率就会增加50%。我非常同意生命在于运动，但有没有可能，那些选择久坐的人是因为他们本身就有运动障碍？

一项针对英国某家医院20 072例急诊住院病例的研究发现，公共假期期间入院的患者，在七天内因病身亡的概率比平时入院的患者要高出48%。对这些数据的一种解释是，公共假期期间在急诊室加班的医生的水平相对差些，不应该让他们留守急诊室。还有可能是因为，在放假期间，除非遇到生死攸关的情况，否则人们都不愿意往急诊室跑。又或者，假期内由于人们出行更频繁、饮酒更不节制、更加放纵自己，所以才会出现更多重伤病例。

有些地方似乎比急诊室更加危险。实际上，最危险的地方显然是床上，因为死在床上的人比死在其他地方的更多。又一次出现了自我选择偏差，即将死之人通常都会被放到床上，但这并不代表床就是杀人利器。

再来看其他例子。20世纪50年代的一项关于身体健康的研究指出，已婚男性的身体健康状况比同龄未婚或离异男性的要好，这表明男性的健康法宝就是结婚并永不离异。然而，人们结婚或离异是自我的选择。由于这种自我选择偏差，即使婚姻通常不利于男性健康，报告所显示的数据也支持结论。试想，如果女性更不愿意嫁给身体状况不佳的男性，那么已婚男性通常会比未婚男性更健康。再试想一下，

女性更有可能与患病的丈夫离婚，那么已婚男性通常也会比离异男性更健康。因此，就出现了自相矛盾，一方面婚姻不利于男性身体健康，另一方面已婚男性却通常比未婚和离异男性更健康。这都是受到自我选择偏差影响的结果。

某州的一项研究对比了每小时限速为55英里、65英里和75英里的高速公路的交通死亡率（驾车每英里的死亡人数）。研究发现，每小时限速为75英里的高速公路的死亡率最低，每小时限速为55英里的高速公路死亡率最高。显然，提高限速可有效减少交通死亡人数，可能是因为在车速较快的情况下，驾驶员会更集中注意力。这项研究统计上最大的问题是什么呢？该州设定每小时限速为75英里的路段肯定是最安全的（道路直、照明好、车辆少）。

在上述示例以及更多的其他例子中，计算机程序都会得出错误结论——预测学生选择第二个专业就能提高成绩，男性结婚后就能改善健康，推迟去急诊室能够延长重危病人的生命，提高车辆限速能降低交通死亡率。

人类能识别自我选择偏差，但计算机不能。

相关系数并非因果关系

一项心理学研究发现，家庭关系的紧张程度与看电视的时长存在强正相关系数，这表明看电视会让家庭关系更紧张。另一项研究指出，参加过驾驶训练课程的人，比没参加过的人更有可能发生车祸，这表明驾驶训练课程培养出的驾驶员更差。一项请愿敦促学校开设拉丁语课，请愿者认为学过拉丁语的学生在口语能力测试中的成绩，比没学过拉丁语的同龄人要高。

统计软件能找出此类相关系数，但是无法解释是第一个要素引起第二个要素，还是相反情况，又或是第三个要素引起前两个要素。对计算机来说，数字就是数字而已，可以求平均数、互相关联和操纵。人类智能则让我们能够思考数字背后的现实，考虑合理的解释。

关系非常紧张的家庭可能会选择（在不同房间）看电视来避免交流。参加驾驶课程的人说不定原来的驾驶水平就很差，所以才要培训。学习拉丁语的人或许更有能力，也可能对口语能力测试中出现的材料更感兴趣。

反向因果关系（reverse causation）通常与自我选择偏差存在联系。人们选择看电视、上驾驶培训课或学拉丁语，会造成“活动为因，对人的影响为果”的错觉，而这些情况下，有可能为反向因果关系，即“人为因，选择活动为果”。

2012年，美国胡子研究所（American Mustache Institute）在华盛顿组织了一次“百万胡子游行”活动来支持“Stache法案”[Stimulus to allow critical hair expenses（意为用胡子保养费来刺激经济发展）的首字母缩略词]，该法案提出每年为留胡子的美国人减免250美元的税款，作为对其保养胡子的补贴。该研究所指出，留胡子的人的年收入比剃净胡子的人的年收入高4.3%，这说明鼓励留胡子的税收政策会促进经济发展。该所研究人员甚至找到一名税收政策教授来支持“Stache法案”：

> 鉴于保养胡子和收入之间明显存在关联，根据美国的《国内税收法》第212条，胡子的保养费用显然符合条件，也应该认定为与收入生产相关的可减免费用。

即便4.3%这一数据没错（怎么就不能是编出来的？），也明显存在自我选择偏差和反向因果关系。是留胡子会提高收入，还是高收入人群更有可能留胡子呢？

几百年来，新赫布里底群岛的岛民都相信，体虱有助于保持身体健康。这一民间智慧基于身体健康的人通常有体虱，而患病之人却没有这样的观察结果。然而患病之人不携带体虱是因为病人经常发烧，会驱走体虱。

人类可以辨别相关系数和因果关系的差异，计算机却不能。如果我将这些例子中数据的识别名称以X和Y代替，你可能会分不清它们之间的因果关系。这就是计算机所看到的“类数据”，所以，计算机又怎能识别出因果关系呢？它做不到。

这便是人类智能和计算机智能之间的根本差别，也是人工智能为什么一点都不智能的原因之一。只有理解所要表达的意思，才能称其为智能。

时间的力量

“比萨原则”（Pizza Principle）指的是，20世纪60年代以来，纽约的一张地铁票大致等于一块比萨的费用。计算机可能认为，比萨的价格取决于地铁票价，或反之亦然。人类则认为两者的花费与总体物价同步上涨。

很多不相关的事物都会随着时间流逝而同步增加，因为人口、收入和物价都会随着时间流逝而同步上涨。这一简单事实是很多假性相关（spurious correlation）的基础。已婚人数和啤酒消费量都随时间流逝而增加。这能说明结婚导致嗜酒，或嗜酒促成婚姻吗？工作日减少的同时，打高尔夫球的人数增加。这就表示大家都翘班去打高尔夫了，或因打高尔夫而受伤在家吗？在欧洲西北部，鹳巢和新生儿的数量都随时间流逝而增加，这可以当成是童话故事中，鹳把婴儿带走的确凿证据吗？

网站Spuriouscoreelations.com上有很多例子，包括北卡罗来纳州的律师人数与上吊、勒喉和窒息死亡的人数有关，内华达州的律师人数和被自己双脚绊倒而摔死的人数有关。那么，难道减少律师人数就可以降低自杀率和摔死率吗？

事物有时也会随着时间流逝而减少，如结婚率和从渔船掉下来摔死的死亡人数。在一次考试中，我设置了一道图表题：如图4.2所示，1999—2009年美国从挪威进口的原油量，以及因驾车横穿火车轨道导致身亡的司机人数都呈下降趋势。我的问题是：你会如何解释两者的相关系数为0.95？我原以为学生会回答，这是两种随着时间流逝而减少的事物间会存在的一种偶然性相关系数，但两者实则毫无关联。然而，有名学生给出的解释是：天然气价格上涨引起从挪威进口的原油量下降和驾车出行减少。好吧，人类时不时会被假性相关忽悠，但计算机一直都上当受骗。

计算机程序不仅无法识别此处讨论的假性关系，还会主动寻找这种假性关系，而且能找得到。由于缺少解释数字意义的人类智能，计算机无法评估所发现的相关系数是否有意义。

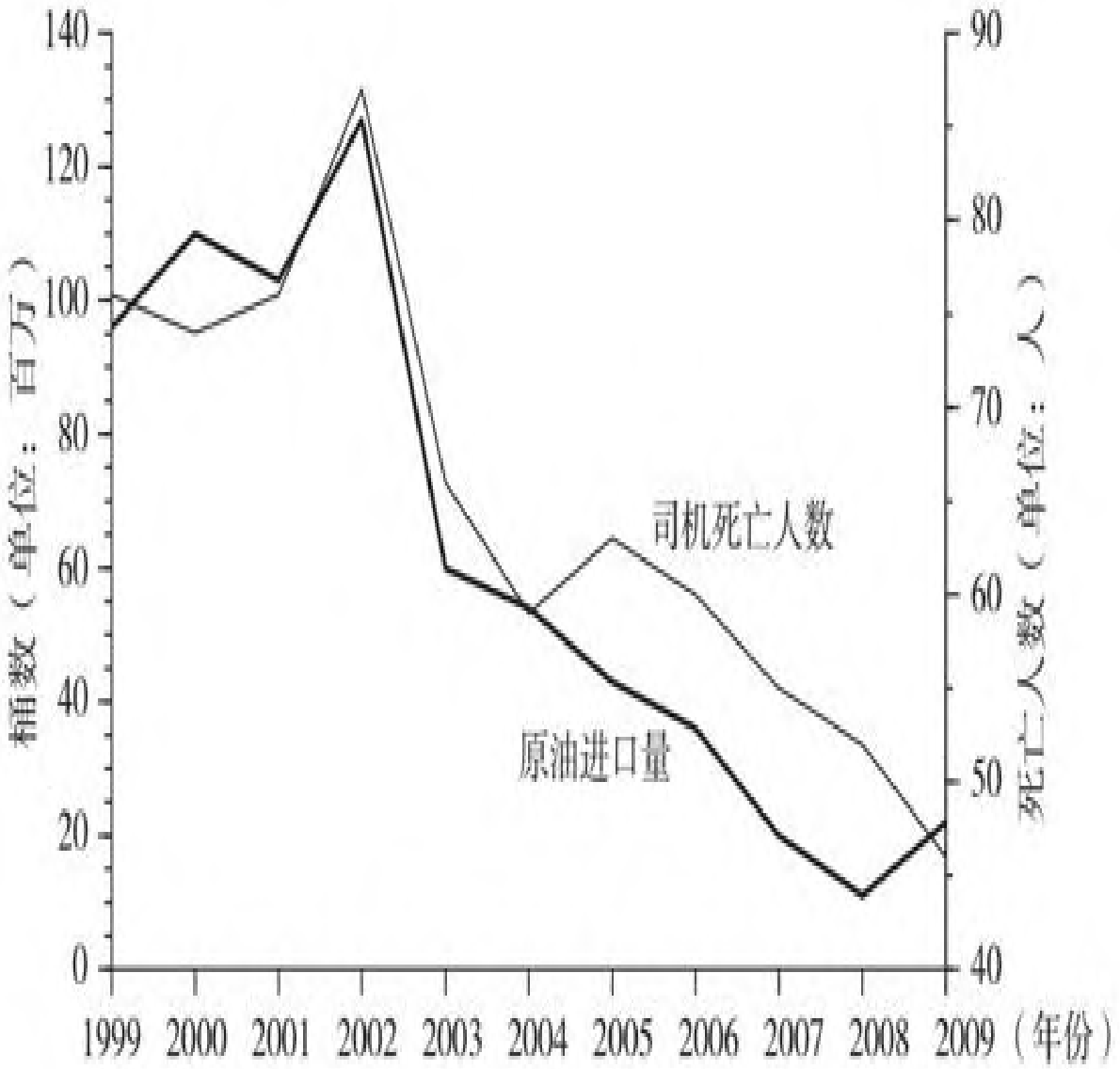


图4.2 假性相关关系

幸存者偏差

另一个观测数据的常见问题是，由于我们无法看到不复存在的事物，因此会出现幸存者偏差（survivor bias）。比如，针对老年人的研究，并不包括那些寿命长度不足以步入老年的人；针对入住某些酒店、搭乘某家航空公司的飞机或游历某个国家的人开展的调查，并不包括那些有过一次体验，然后说“下不为例”的人；杰出企业的共同特征并不包括那些也具备这些特征，但没有成为杰出代表的企业。

第二次世界大战期间，轰炸德国后飞回英国的盟军飞机的机身留下了被子弹和弹片击中的弹孔，一项军事分析研究这些弹孔位置后发现，大多数弹孔位于机翼和机尾，很少出现在驾驶舱、引擎和油箱部位（如图4.3所示）。于是，军队决定在飞机的机翼和机尾增加保护钢板。不过，才智过人的统计学家亚伯拉罕·瓦尔德发现了这些数据中存在的幸存者偏差问题。能成功返回英国的飞机，都在子弹或弹片的袭击中幸免于难，其驾驶舱和油箱都没什么弹孔，而这些部位被击中的飞机都已经遇难了。因此，应该加固的不是那些弹孔最多的地方，而是没有弹孔的位置。

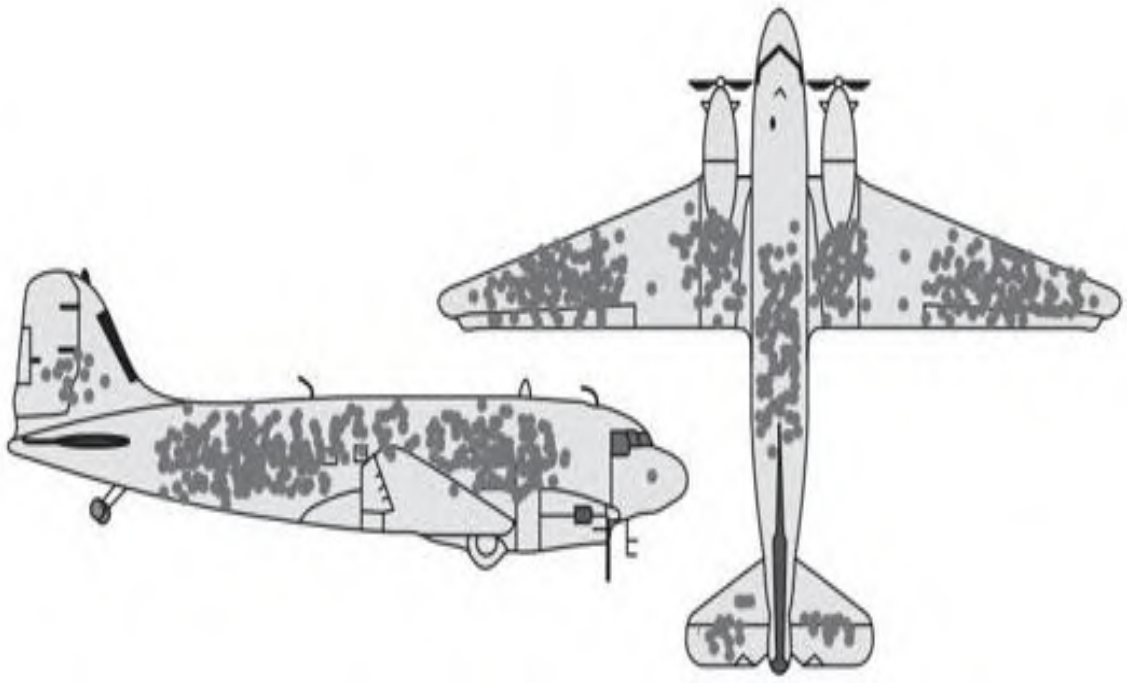


图4.3 飞机中弹部位示意图

还有一个相似的例子。第一次世界大战期间，很多英国士兵的头部都会遭弹片致伤。然而，当把布帽换成钢盔后，因头部受伤入院的士兵人数反而增加了。为什么？因为钢盔的使用增加了因弹片受伤的人数，而不是死亡人数。

面对某些类型的“坏数据”，人类还是能够识别幸存者偏差的，但计算机程序就无能为力了，因为它们不知道这些数字代表什么，无法利用逻辑推理思考这些数据的来源造成的偏差。

假数据

学术研究的圈子中竞争激烈，众多聪明绝顶、竞争力强的科学家坚持不懈地追名逐利，以求自己的职业生涯获得发展。而必要的支持就是，在同行认可的杂志上发表有意思的研究成果。“要么发表，要么发霉”（publish or perish）是大学研究人员的残酷写照。

有时，这种压力过于沉重，以至于研究人员会撒谎或造假来寻求事业高升。研究人员需要靠可以发表的研究结果存活，当结果不尽如人意时会备感沮丧，还会担心其他人抢先发表了类似的研究成果，因此有些人会选择编造数据这条捷径。毕竟，如果推论是正确的，那么编些数据来证明一下也不会有什么大碍吧？

荷兰社会心理学家德里克·斯塔佩尔的研究成果极其丰硕，事业有成，对调查研究的设计非常深入、一丝不苟，常与研究生和同事合作。可奇怪的是，作为一名资深研究人员，他总是独立完成调查工作。还有一点蹊跷之处是，斯塔佩尔常常在听到同事的研究兴趣点

后，声称自己已经收集到了对方所需要的数据，他愿意以合作撰稿人的身份提供数据。

斯塔佩尔发表或与他人合作发表过数百篇论文，于2009年荣获美国实验社会心理学会（Society of Experimental Social Psychology）颁发的“事业轨迹奖”（Career Trajectory Award）。2010年，他当选蒂尔堡大学社会与行为科学学院院长。斯塔佩尔很多论文都引起争议，其中有些甚至超过合理性的界限。他在一篇文章中指出，脏乱的房间会使人变成种族主义者；在另一篇文章中声称，吃肉——包括只是想要吃肉——会使人变得自私。（我没有瞎编！）

斯塔佩尔的部分同事和研究生对于支持斯诺佩尔的那些半成品的推论的数据持怀疑态度。但最终，他们想要查看数据的要求都被斯塔佩尔拒绝了。有一次，斯塔佩尔向一名同事展示了一项实验的方法和标准差，该实验显示出，给流泪的卡通人物涂颜色的儿童比给不流泪的卡通人物涂颜色的儿童更愿意分享糖果。当那位同事提出查看数据以比较男孩和女孩的反应时，斯塔佩尔则说，他还未将数据输入电脑！要么是斯塔佩尔手算的实验过程和标准差，要么就是根本没有数据。你说呢？

终于，两名研究生将他们对斯塔佩尔的质疑上报给学院负责人。没过多久斯塔佩尔就坦白了，承认自己的很多实验调查结果要么是经过挑选的，要么是捏造而来的。他解释道：“我太急于求成了。”

2011年，斯塔佩尔被蒂尔堡大学辞退。他退掉了自己的博士学位，撤回了58篇数据造假的论文。他还同意完成120个小时的社区服务，用相当于其18个月工资的收入缴纳罚款。对此，荷兰检察官同意不以滥用公共研究经费追究其刑事责任，理由是这笔政府资金主要用于支付研究生的工资，他们都是无辜的。与此同时，我们也松了一口气——在吃肉和房间脏乱时不用那么愧疚了。

真是名副其实的资本主义做法，斯塔佩尔并没有灰溜溜地全身而退，反省自己的过错，反而将此事写成书出版了！

计算机可能不会捏造数据，但也无法识别虚假数据。把斯塔佩尔拉下神坛的是人类对他说辞的质疑。而计算机无法持有怀疑态度。

识别“坏数据”

我以前的一位学生（鲍勃）曾在全球最大的咨询公司负责大数据工作，他告诉我，有家客户公司针对营销计划的效力做了统计分析，该分析采用多种渠道，包括直邮、广告横幅、电子邮件等。为了收集到足够多的数据，计算机程序分析了超过20年的历史数据。

问题在于，任何人都知道在过去20年里营销技术已经发生了翻天覆地的变化。比如黄页，这种印刷的电话簿用白纸印刷非商用电话号码，用黄纸印刷分类付费广告（例如宠物用具）。从前，这种做法非常普遍，全美国都很时兴一句话：“动动手指省跑腿！”（Let your fingers do the walking！）不过实际上，一本黄页太大了，就算撕开一半也还是很重（尽管有个窍门能让它不那么重）。

如今的黄页被鲍勃称为“优惠券和门挡的结合体”。大多数人会收到放在信箱里或自家停车道上的黄页，但从不打开来看。

客户公司的软件程序无法考虑这些，因为它们没有思考能力，只会计算。幸运的是，鲍勃介入此事，表示少量的相关数据比大量的过时数据更有价值。

尽管所有宣传都在鼓吹大数据，但有时小数据的作用反而更加显著。无须通过搜遍堆积如山、随机获取的数字来寻找有意思的发现，采集有助于解答研究问题的好数据更富有成效。

人类能够质疑支持奇特主张的数据，可是计算机却无能为力，因为它们没有分辨合情合理与异想天开的智能。遇到“脏乱的屋子和种

族主义”之间、“内华达律师人数和走路绊脚摔死的人数”之间等存在统计学关系的说法时，计算机也不会起疑。

对软件程序来说，要想解决计算机缺少人类智能这一问题，可能需要报告统计学分析的具体细节，这样的话，人类就能评估数据的有效性和结果的合理性。这是个很好的想法，但实践起来还有两点障碍。

首先，因为大数据过于庞杂，很多软件程序都采用数据规约（data-reduction）技术（详见第8章），将数据改变为无法识别的状态，人类便不能评估计算机结果是否有意义。

其次，有些人认为计算机生成的内容不容置疑。如果计算机显示，嘴角下垂的人有可能是罪犯，那就一定如此。如果计算机显示，不接电话的人信用都不好，那也一定没错。

人类以为自己没有计算机那么聪明，但并非如此。计算机虽然非常擅长某些事情，但无法鉴别数据质量的好坏，人类远比计算机聪明——这也是说明计算机不算太智能的又一例子。

---

[1] 1英尺≈0.3米。——编者注

第5章

随机性模式

每当统计学课程开课的第一天，我都会做超感官知觉（extrasensory perception, ESP）实验。先给学生们展示一枚普通硬币（有时向学生借），然后将其抛投10次。每抛一次，我就刻意把结果印入脑中。与此同时，学生尝试猜测我的想法，然后写下答案。我还会在一张事先设计好的纸上，以圈出H（正面）或T（背面）的方式，记录每次抛投的实际结果，这样一来，学生就无法通过我的手势猜出结果。

谁猜对了10次，谁就能赢得当地一家精品巧克力店的一盒一磅装巧克力。如果你在家也想试试，那就猜猜我在2017年春季的统计课上那10次抛投硬币的结果。我的脑电波或许还留存在某个地方。然后写下你的答案，看看能猜中几次。

抛完10次后，我让学生们举起手来，然后开始公布结果。猜错的学生把手放下，坚持到最后的即可赢得巧克力。曾经出现过一名获胜者，在参与这个游戏的学生人数超过了1 000名后，有人获胜也在预料之中。

我并不相信超感官知觉，所以这个实验的重点并非赢得巧克力。把巧克力设置为奖品，只是为了让学生认真对待这一测试。我的真实意图是想说明大多数人，即便是聪明的大学生，对抛硬币等随机事件也存在误解。这一误解加深了我们的错误想法，即以为电脑发现的数据模式一定都有意义。

早在20世纪30年代，美国的真力时无线电公司（Zenith Radio Corporation）有一档系列节目，每周播出一次超感官知觉实验。无线电广播里的“发送者”随机选择一个圆圈或方框，类似抛硬币，然后想着所选的图形，希望脑海中的图像能传送给数百英里之外的听众。随机进行五轮选择后，听众可以将猜测答案寄给电台。

这些实验虽然不能支持超感官知觉的说法，但确实可以有力证明，人会低估随机数据模式出现的频率。我们大多数人认为，圆圈和方框出现的次数通常应该相等，而且不会以任何可识别的模式呈现。例如，在一次实验中，121名听众都选择了以下序列：

□ □ ○ □ ○

只有35名听众选择以下序列：

□ ○ □ ○ □

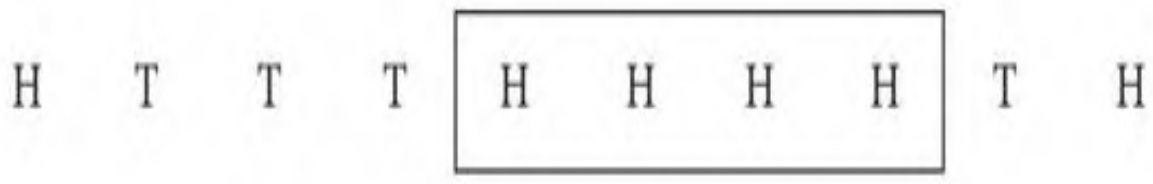
上述两种序列中，都含有3个方框，2个圆圈，但是第一种序列似乎比第二种完美交错的序列更加随机。这么说你同意吗？

只有一名听众选择如下第三种序列，因为大多数人认为随机结果不会这么一致。

□ □ □ □ □

事实上，这三种序列出现的概率完全相等。不过，听众还是不愿意猜有5个方框连续出现，或者两种形状完美交替出现的序列，因为他们觉得这样的情况不会随机发生。你可能也有同样的想法。对了，我在2017年春季课程上的抛硬币结果为：T、T、T、T、T、H、H、T、H、T。你全猜对了吗？

核对完结果，看看是否有人获胜之后，我会让学生数一数自己的答案中连续出现次数最多的是哪个结果。比如以下序列中，连续出现最多的是4次正面：



而以下序列中，连续出现最多的是3次背面：



这些序列不像随机结果，但我保证它们就是随机出现的。我抛了20次硬币，得出了以上结果。

过去10年共有263名学生上过这门统计学课，其中报告连续出现4次或4次以上同一面的人只占13%。你的结果是这样吗？

实际上，正面或反面连续出现4次或4次以上的情形，并非完全不可能！在10次抛投中，同一结果连续出现4次或4次以上的概率为47%。我们预计，在这264名学生中，会有124名报告这样的结果，但实际上只有34名。学生都大大低估了同一面连续出现4次、5次，甚至6次的概率。

显然，大家看到正面或反面一直出现都会感到别扭，因为这样的结果不像是随机产生的。连续出现两三次正面后，他们猜反面的念头越来越强烈，以便达到平衡。

不仅统计学课堂上的抛硬币实验如此，在体育比赛、靠运气取胜的游戏和生活中，大多数人也仍未正确认识到随机数据中出现连续情况的概率有多高。因此，一旦出现连续情况，他们的第一反应就是这些数据并非随机得来，其出现肯定有潜在原因，于是就生编硬造出一个所以然。

篮球运动员如果连续投中5次，肯定会“热乎”起来，非常有可能再投中下一球；连续5次选股大赚的金融咨询师必定是金融高手；连续5年势头良好的共同基金一定由金融天才管理。尽管共同基金的表现中

唯一具有一致性的是以往业绩无法准确预测将来业绩，但是投资者还是会放弃业绩连年不佳的基金，转投业绩连年良好的基金。

美国国家体育比赛解说员和体育专栏作家名人堂成员梅尔文·德斯拉格，在其最后一篇报刊专栏文章中提及自己在51年的职业生涯中收获的忠告，其中包括一个著名赌徒的建议：“大名鼎鼎的‘希腊人尼克’（Nick the Greek）透露了取胜的秘诀，他训练自己可以持续玩牌八小时而不用上洗手间。按照他的说法，上了牌桌就不应该打断骰气。”唯有低估随机数据连续出现概率的人才会拼命控制膀胱，唯恐“打断骰气”。

我在上文中提出，“在10次抛投中，同一结果连续出现4次或4次以上的概率为47%”。对此，一名活跃的学生表示难以置信，并且编写了计算机程序来证明我是错的。他编写的程序模拟了100万次抛硬币，并记录每10次中正反面连续出现最多的次数。他的计算机程序也得到了同样的结果。他承认自己的程序证实了我的观点，但他还是不信。他认为，可能是计算机的随机数字生成器出了问题，但他又没那么多时间自己抛100万次来验证。看来，随机数字不会连续出现的想法已经在他的思维中根深蒂固了。

如果抛硬币超过10次，上述概率会更高，连续出现次数会更多。抛1 000次，同一面连续出现大于或等于10次的概率是62%；抛1万次，同一面连续出现大于或等于17次的概率是53%，大于或等于18次的概率是32%。

数据越多，就越能肯定还会产生更多连续出现的结果，以及其他出乎意料的模式。克里斯蒂安·S.卡鲁德和朱塞佩·隆哥合作发表了一篇理论性文章，题为“大数据中假性相关的泛滥”（The Deluge of Spurious Correlations in Big Data），表明在所有庞大数据中集中出现高度规则的模式都不足为奇。不仅如此，而且：

数据越多，就越会在其中发现随意、无意义和（对未来行动）无作用的相关系数。因此，自相矛盾的是，我们得到的信息越多，就越难从中提取有意义的发现。信息量过犹不及。

如果存在一组有助于做出预测的真实统计学关系的固定数据集，数据滥用肯定会提高无用统计学关系在真实关系中的比率。

假设股价、失业率和利率之间存在因果关系。如果失业率上升，则股价下跌。如果利率上升，则股价也呈下滑趋势。通过看股价、失业率和利率的数据，我们可能会找到证实这些因果关系的统计学证据。

再假设，我们把几座偏僻城市的日常气温也考虑在内，尽管它们跟股价毫不相关。根据卡鲁德和隆哥的论证，纳入的无关变量越多，就越能肯定得到的是无意义模式。

与包含两个有意义变量（失业率和利率）和100个无意义变量（100个小镇的气温）相比，包含两个有意义变量和5个无意义变量的结果可能与股价的相关性更高。与包含两个有意义变量和1 000个无意义变量相比，包含两个有意义变量和50个无意义变量的结果可能与股价的相关性更高。

因此，卡鲁德和隆哥总结道：“数据越多，发现无意义模式的概率就越高。”

数据挖掘

人工智能是不断变化的专有名词，包括计算机模拟人类行为的各种活动，例如，组装汽车、识别物体、将语音转换成文本。人工智能

还可以驾车、下棋和交易股票。

控制人工智能活动的计算机程序被称作“算法”（algorithms），即完成任务所需的分步规则。例如，寻找某数平方根的算法步骤如表5.1所示。

算法在进行了5个循环后，得出答案为X=7.071068。

表5.1 平方根算法

规则	步骤
1. 输入任意数 Y	Y = 50
2. 选择测试方程式 X = Y/2	X = 50/2 = 25
3. 计算 X 的平方	$X^2 = 25 \times 25 = 625$
4. 计算 $Z = Y - X^2$	Z = 50 - 625 = -575
5. 计算 E = Z/Y	E = -575/50 = -11.5
6. 若 | E | < 0.00001，得到 X；否则，进行第 7 步	进行第 7 步
7. Z/(2X) 加上 X	X = 25-575/50 = 13.5
8. 返回第 3 步	进行第 3 步

计算机程序使用多种语言执行算法。平方根算法可以用BASIC、Java、C++等计算机编程语言。当然，人工智能算法的能力远不止这个简单的例子。

数据挖掘可能是最艰巨、最危险的人工智能形式。传统的数据统计学分析遵从已经广为人知的科学方法，用科学知识取代迷信。研究人员基于观察或推测提出问题，比如，“维生素C是否会降低普通感冒的发病率和严重程度”，研究人员搜集数据后，最好能够通过控制实验来验证这个推测。如果服用安慰剂和维生素C的结果出现令人信服的统计学差异，则这项研究得出结论，维生素C具有统计学上的显著影响。该研究人员运用数据验证了推测。

数据挖掘则另辟蹊径，其数据分析不会受到预先形成的推测的驱使或妨碍。数据挖掘算法的编程目的是发现数据的走势、相关系数等模型。一旦发现有意思的模型，研究人员就创造理论来解释它。或者，研究人员认为，数据可以自圆其说，一切解释都包含在数据中。他们不需要理论学说，只要有数据就足够了。

在维生素C的例子中，假设数据挖掘工具针对1 000个人创建数据库，记录他们的所有信息，如性别、年龄、种族、收入、发色、瞳孔颜色、就医记录、运动和饮食习惯等。接着，使用数据挖掘软件识别出与个人患病天数在统计学上最相关的五项个人特征。结果可能显示为：酸奶食用过量、茶类饮用不足、喜欢散步、绿瞳孔，以及在脸书上最常使用的词为excellent（好极了）。

数据挖掘工具可能得出结论——酸奶、茶、散步、绿瞳孔和脸书常用词为excellent代表着不健康——于是编造出稀奇的故事来解释这些相关系数。数据挖掘工具还可能认为，数据已经解释得面面俱到，无须进一步解释了。

《经济学人》在2015年发表的题为“与悲观相去甚远：经济学发展”（A Long Way From Dismal: Economics Evolves）的文章指出，（研究失业、通胀等的）宏观经济学家应该效仿在科技企业从事产品、公司和市场相关数据挖掘工作的微观经济学家。

（宏观经济学家）应该减少理论空谈。宏观经济学家都是严谨之人，先创建理论模型，后使用数据检验。新一代经济学家则忽略白板功能，只集中处理数据，让计算机识别出模式。

《经济学人》是一本优秀的杂志，但不是优秀的新闻报道。

2008年，美国《连线》杂志总编辑克里斯·安德森撰写了一篇引起争议的文章，题为“理论的终结：数据泛滥使科学方法过时”（The End of Theory: The Data Deluge Makes the Scientific Method Obsolete）。安德森表示：

只要有足够多的数据，数据就能自圆其说……更庞大的数据以及处理数据的统计学工具，都为理解世界提供了全新的方式。相关系数可以取代因果关系，科学的发展根本无须相关模型、统一理论或任何真正的机械论的解释。

当时看来，这似乎是一种刻意煽动争议、几乎毫不掩饰的自吹自擂——“未来是大数据和大电脑的世界，请阅读《连线》”。

值得赞扬的是，数年后，《连线》杂志的英国版发表了一篇具有警戒意义的文章，题为“如何篡改统计值”（How to massage statistics），其中谈到了我的担忧——“计算机让摆弄数据更加轻而易举”，还列举了篡改、挑拣和破坏数据以造成误导的各种方法。

不幸的是，对曾经颇有争议的事情，人们现在已经习以为常。认为处理数据便足矣的人比比皆是——认为人类无须理解世界，也无须理论，能在数据中找到模式就足够了。在这个方面，计算机可谓得心应手。因此，我们应该将决定权交给计算机。

有时，“数据挖掘”这个词的使用范围更广，还包括搜索引擎和机器人汽车工等大有裨益、无可厚非的活动。我经常使用“数据挖

掘”来描述这种做法——运用数据发现统计学关系，然后以此预测行为，例如，寻找统计学模型以预测汽车采购、贷款拖欠、患病或股价变动的情况。

知识发现

我和一名教“知识发现”这门课程的教授吃过午餐。我问他，假如缺乏理论（或常识），我们怎么知道由数据产生的模型真的有助于预测，而不是偶然？他认为：

> 证据就在数据之中。我们不仅不需要理论，理论化还会限制我们所见，妨碍我们发现意料之外的模型和关系。模型是否有用，只需要看数据就知道了。这就是为什么我把这门数据挖掘课程称作“知识发现”。

数据挖掘还被称为“数据探索”“数据驱动的发现”“知识提取”“信息获取”等，这些称呼都反映了一个核心思想——数据先于理论，甚至通常无须理论。

很多被称作人工智能的事物都令人惊叹。可是，数据挖掘并非如此。其根本原因很简单，却不易被认识到：

> 我们以为模型不同寻常，因此具有意义。
>
> 在大数据中，模型无法避免，因此毫无意义。

黑匣子

我最近看了一家对冲基金（我称其为“想都不想”）的企划书，其中吹嘘道：

> 我们完全自动化的投资组合按照计算机算法运行。所有交易均通过复杂的计算机系统完成，消除了经理人的一切主观因素。

这就是所谓的“黑匣子”方法，把内容输入算法，算法输出结果（如图5.1所示），而人类用户对结果的决策过程一无所知。



图5.1 黑匣子

在求平方根的算法中，如果输入50，则输出7.071068。然而，我的算法不是黑匣子，因为我解释了程序是如何运行的，任何人都能检查我在逻辑或某一步指令上是否犯了错。事实上，你可能已经发现了问题。50的平方根可以是+7.071068或-7.071068，而我的算法结果只显示了正数。此外，该程序在求Y=0的平方根时会出现问题，因为第五步是计算Z/Y，但是Z/0无意义。最后，算法如何处理负数的平方根呢？没法处理。

当程序处于开放状态时，人类能够看到运行过程，查找错误、遗漏和其他故障。但当程序藏在黑匣子里时，人类就无法这么做了。我们不知道黑匣子里的算法是什么，无法评估过程中是否存在逻辑错误、编程差错或其他问题。黑匣子的输入内容不计其数，处理过程神秘莫测，输出内容让人难以捉摸。

对黑匣子股票交易算法来说，输入值可能是股价、交易股票数量、利率、失业率、推特出现“股市”一词的次数、黄色涂料的销量和几十项其他变量，输出值可能是100股苹果公司股票的买卖决定。

使用黑匣子交易算法决定股票交易的用户不知道做这些决定的理由，也并不费心去了解，因为他们相信黑匣子，就像希拉里·克林顿相信“阿达”一样。他们认为，计算机比自己聪明，这应该让人放心。包括“想都不想”的对冲基金经理在内的许多人都认为，用黑匣子进行投资决定，这是特点，不是缺点，毕竟它“消除了经理人的一切主观因素”。

核准贷款的黑匣子算法拒绝贷款申请的理由可能是申请人的手机没充满电，监狱假释的黑匣子算法拒绝假释的理由可能是申请人戴着宽腕套，防止犯罪的黑匣子算法建议抓捕某人的理由可能是他的鼻子和嘴巴呈某种形状。你可能觉得我是在胡编乱造，可我说的都是实话。

计算机算法能连续无误地进行数学计算，是因为软件工程师确切地知道自己想要算法去做什么，然后编程实现这一目的。但数据挖掘算法就无法这么做，因为该算法的意图模糊不定，结果无法预测。一名人工智能专家写道：“任何两种人工智能设计之间的相似点，可能比你和矮牵牛花之间的还少。”

黑匣子数据挖掘是人工操作，但它并不智能。这就是为什么我给本书取名为《错觉：AI如何通过数据挖掘误导我们》。

有些人使用贬义词“人工蠢能”（artificial stupidity, AS）来描述计算机让我们失望时的情况，如Siri听不懂问题、谷歌地图导航进了死胡同、自动交通灯卡在红灯上。我使用“人工低能”（artificial unintelligence），并非描述计算机偶尔会犯错误，而是强调计算机并不拥有人类般的智能。为了计算50的平方根而遵守规则，同知道苹果公司股价和墨尔本高温的意义并明白为什么两者之间不存在逻辑关联有根本区别。

大数据、大电脑、大麻烦

几十年前，数据匮乏，计算机还没出现，研究人员奋力搜集数据，并花费数小时甚至数天时间埋头苦算。如今，我们生活在大数据的时代，计算机可以高速运行，二者的有力结合一直受到称赞，甚至崇拜。有些人服从计算机，认为计算机无所不能。对大数据的崇拜被称为“数据主义”（dataism）或“数据化”（datafication），认为一切重要事物都可以用数据来表示，数据分析无懈可击。向计算机臣服吧！

这种痴迷并非没有危害。我们过于武断地认为搜索处理堆积如山的数据不会出差错，但出错在所难免。数据不过是数据，计算机也不过是计算机。计算机无法区分有用数据和无用数据，无法分辨合理结论和一派胡言。“数据无须理论支撑”是一种危险的理念。

连续出现、相关系数、走势模型等本身证明不了什么。即便是通过抛硬币，也能发现这些模型。我们需要思考原因，要问为什么，而非是什么。

不可否认，计算机令人惊叹，神秘莫测。我们大多数人不了解手机如何让我们与几千英里外的人视频对话，也不知道计算机如何能给

出详细的驾驶导航，还可以根据当前交通状况给出预计到达时间。我们只知道计算机太神奇了。如果计算机告诉我们，总统大选的结果可以通过闻所未闻的几座城市的气温预测得到，我们可能也会认为它说得对。如果计算机可以显示π的小数点后2 000位数和世界上每座城市的街景图，我们区区凡人有谁能质疑它的智慧呢？

事实的残酷在于，数据挖掘算法是由数学家创建的，相比现实状况，他们对数学理论更感兴趣。从15名数学家的脑部功能性磁共振成像图（fMRI）可以发现，看到数学等式会激活他们的眶额部皮质中线部，而在人们看到惊险杂技或听到美妙音乐时，这一区域也会激活。有些人欣赏优美的音乐、艺术、舞蹈和文学，而数学家则欣赏数学等式的内在美。

沃伦·巴菲特曾发出警告：“要小心满脑子都是公式的怪人。”我大学主修数学，现在教金融学和统计学。实际上，我生活中的每一天都会用到数学，还编过几十个软件程序，为我的研究分析数据。我很喜欢公式和计算机，但我也知道，数学的魅力会引导我们创建让内心愉悦却无实践价值的数学模型。有太多数据挖掘算法都属于这一类。

利益冲突

哪里有利可图，哪里就有人蜂拥而至。

20世纪90年代，计算机进入我们的生活，互联网的发展催生出数以百计以互联网为基础的企业，即广为人知的网络公司（dot-coms）。有些网络公司有好的想法，逐渐发展成为实力雄厚的成功企业，但大多数没有。有太多网络公司只是为了在公司名称里加上dot-

com，然后转卖出去，赚得盆满钵满后转身就走。找到好点子、开公司、打造成功企业，然后托付给子孙后代，这是旧经济的过时做法。

一项研究发现，企业不过是在名称里加上了.com、.net或互联网，股价便翻了一番还多。股民的钱打了水漂！

如今，人工智能同样如此。人工智能已经成为一种时尚，任何跟计算机沾边的东西似乎都能被称作人工智能。真可笑，连我那个求平方根的计算器都能被算作人工智能。何乐而不为？

我从前的一名学生投资了人工智能初创公司，他跟我说：“当前，‘数据科学家’和‘机器学习专家’是最热门的职业。其中有些是接受过训练的统计学家、经济学家，但有些只是上过六周网络课程的程序员，课程可能仅重点讲解一些技术工具和技巧，没有提供基础的理论知识帮助他们了解理论的局限。”谁还愿意思考呢？都冠以人工智能的名号，四处兜售就好了。2017年，“AI”入选美国全国广告商协会的“年度营销词”。

我的另一名学生现在是一家大公司的首席财务官，他写信给我：“你不会相信人们多么频繁地向我提到‘大数据’的优势，或者愿意提供‘分析专长’——这些人都是外行，（有可能）没有意识到你在书中详述过的局限。”

为了说服大家为实际上并不需要的东西砸更多钱，需要做出更多承诺，提供超出实际能兑现的范围的目标。这种情况在互联网泡沫时期出现过，如今到了人工智能时代又重蹈覆辙。我们应该对拼命向我们推销的人持怀疑态度。

天生就会被骗

人类不太能接受“随机事件”，见不得某件事无缘无故地发生。我们老想着给每个模型做出有意义的解释，但有可能它根本就毫无意义可言，不过是偶然发生的罢了。正如尤吉·贝拉所言：“这巧合得太不像话了。”

你可以将此怪罪于我们的远古祖先曾设法应对的演化和环境问题。拥有便于生存繁殖的遗传特征的有机体，会将这些特征遗传给后代，而那些欠佳的特征则会被淘汰出基因库。持续不断地代代相传，这些有价值的遗传特征便会占据主导地位。

识别和解释模式曾经具有生存价值。乌云通常预示着下雨，灌木丛中传来声音说明可能有捕食者，发质是繁殖力的象征，脸型对称代表基因健康。远古时期，模式识别有助于人类祖先找到食物和水、意识到危险，还有助于吸引到有繁殖力、能养育健康后代的配偶，并将这种能力遗传给后代。那些不太擅长识别有益于生存繁殖的模型的人，将自己的基因遗传下去的机会更少。经过无数代自然选择，我们天生就会寻找模型，并为找到的模型寻求解释。

我们太容易被内在欲望所诱惑，想要解释所见的事物，这掩盖了如下事实：模型不可避免地是由无法解释的随机事件创建出来的，如抛10次硬币。我们应该承认自己容易受到模型的诱惑，从而努力做到拒绝诱惑，保持质疑。

为模型所惑

真力时公司的超感官知觉测试说明，我们对随机数据有先入为主的想法（或误解）。随机数据看似序列1：

□ □ ○ □ ○

随机数据不像序列2：

□ ○ □ ○ □

随机数据肯定也不像序列3：

□ □ □ □ □

因此，我们认为，如果模型如序列2和序列3，肯定就不是随机产生的。或许是方框和圆圈没有被预先打乱，又或许这不是超感官知觉测试，而是公开播放给间谍的密码。

听到这儿，你或许只是付之一笑，但担任《纽约时报》金融专栏作家多年的伯顿·克兰曾说：

> 我一直都深信不疑的是，（曾经用来记录股价的）纸带上价格之间的点是密码（如图5.2所示），以便互相发出市场波动的信号。有人甚至让我看过所谓的翻译码。

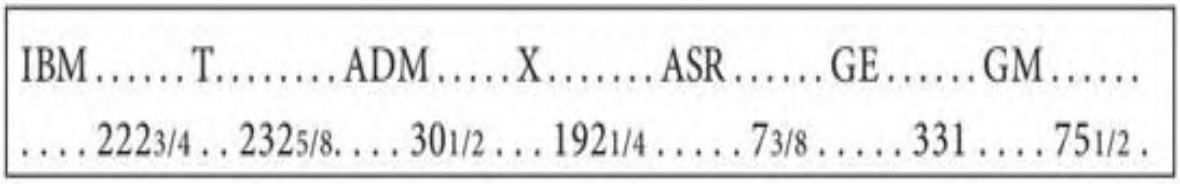


图5.2 用来记录股价的纸带

破译纸带上随机出现的点是数据挖掘的雏形：先寻找模型，然后为此编一种说法。偏执的股票交易员确实会仔细查看这些点，找寻模型，发现模型，然后设法将这些模型和股价变动联系起来。交易员受模型驱使，拼命寻找模型，而且成功了。他们并没有意识到，模型肯定会出现，即便在随机产生的数据中也一样。

这一误解的另一表现是一本关于如何赢得掷双骰子游戏的书。作者在拉斯维加斯一家赌场记录了5万次掷骰子的结果，研究数字出现的

序列。预计掷骰子5万次会出现约20次4-4-11序列，但实际该序列出现了31次。于是该书建议每当4连续出现两次后，都要押11。作者还发现，38次掷骰子的结果中，7-12-7的序列出现了10次，接着出现的数字为2、3或12。如果这38次每次都押注100美元，就能赢4 200美元。

这些计算都是手工完成的，那时还没有计算机，更别说数据挖掘软件了。一想到作者要花上好几个月甚至好几年来找寻这些模型，我就不寒而栗。唯一令人感到欣慰的是，作者在研究数字上花的时间越多，在偶然事件上押注的时间就越少。

这个可怜的作者为了在5万次掷骰子中寻找偶然模型而耗费时间，今天的数据挖掘软件也在这么做，只不过计算机化的数据挖掘能在数秒内完成这项任务，无须数月时间。掷骰子是简单易懂的例子，它说明了如何总能在这样随机产生的数据中找出模型，以及人们多么渴望自己找到的模型是有意义的。事实上，找到的模型根本毫无意义。

随机噪声

大电脑搜遍大数据后，一定可以找出比掷骰子的4-4-11序列更复杂、更不寻常的模型，即便这些数据只是随机噪声。例如，我为100个随机产生的变量均创建250个观察结果，每个变量初始值为50，在随后的249次观察中，由计算机的随机数字生成器决定这个值是增加还是减少。这100个变量都通过统计学家称为“随机游走”（random walk）的程序产生，就像醉汉走路，每走一步都和前一步没有关联一样，每个变量的下一次改变都与上一次改变没有关系。

每个变量的每次观察都与其他99个变量的演变完全独立开来。但事实上，还是一定会出现偶然性模型。数据挖掘软件有非常强大的模型寻找能力，不过，对模型评估就无计可施了。就像我们在前面章节反复说过的，原因在于计算机并不能理解真实世界。数字只是数字。

我运用某些数据挖掘软件发现这些随机产生的变量中，有一个变量连续13次出现增加情况。如果不是头脑清醒，我可能会认为自己有什么重大发现了。

接下来我用数据挖掘软件寻找任意两个变量之间简单的两两相关系数。一共存在4 950对可能的相关系数。我的数据挖掘软件找到了98对相关系数在0.9以上的变量。如果不是头脑清醒，我可能会认为自己又有什么重大发现了。

最后，我使用数据挖掘软件来寻找这100个解释性变量中的组合，该组合会与一个真实变量高度相关，即2015年标准普尔500指数的每日价值。每5个变量一组，则有75 287 520种可能。这听上去似乎很多，但是对现代计算机来说不算什么。据我预计，在这些虚假变量中，某些变量组合会与真实变量高度相关。结果不出我所料，数据挖掘软件找到一个组合，与标准普尔500指数的相关系数达到0.88。如果不是头脑清醒，我可能会认为自己真的有什么重大发现了。

数据挖掘软件每次都能发现模型，某一次它表明，精明老练的投资人会战胜股市。该软件会筛选、分类和分析所有随机数据，尽管这些数据跟股价一点关系都没有，对决定买进还是卖出股票完全没有帮助。不过，该软件还是找到了足够强的相关系数，说服黑匣子股票交易算法买进或卖出股票。

只要了解数据是如何产生的，人类立刻就能理解这个笑话，但计算机不能。数据挖掘软件无从明白自己的发现是否有用，因为对计算机来说，数字只是数字而已。

真正进行数据挖掘的人会在大数据中启动其数据挖掘算法，通常是数十亿或数万亿次，他们的算法不仅在每个数据组合中寻找模型、寻找不同数据组合之间的交互关系，还会寻找更加复杂的关系。他们必然会找到不同寻常的模型，不过，就像上述的股市例子一样，软件无法辨别何为因果、何为偶然。

业余的天气预测

再举一个例子说明数据挖掘的危害之处。即使并没有充分的理由表明数据具备实际价值，但数据挖掘工具也照例筛选与预测对象毫无关联的数据。例如，假设我想预测明天的气温。真正的天气预报会使用复杂的计算机模型，将大气分为若干个立方体，运用卫星数据估算每个立方体的气温、湿度、风速等。计算机模型利用物理学、流体力学等科学原理，预测天气会如何随着立方体之间的相互作用进行变化。

这听上去挺费劲的。我没有那些资源，也不懂科学原理。但是，我可以使用数据挖掘软件基于知识发现来预测天气。具体来说，我尝试根据城市B昨天的气温，来预测城市A明天的气温。我也可以参考城市A昨天的气温，但这就不算是知识发现了，不是吗？

我请一名出色的研究助理海蒂·阿蒂格帮忙搜集了25座分布广泛且相对偏僻的美国城市在2015年和2016年的每日最高和最低气温数据。她无意中把澳大利亚西部一座临时小型机场——科廷机场也包括在内。

真是无巧不成书。几年前的圣诞假期，我到澳大利亚墨尔本拜访朋友。那时，我了解到了葡萄干布丁、澳大利亚的圣诞歌曲和墨尔本板球场的节礼日板球赛，我在后院还把网球当板球打。然而，我印象最深的还是拆圣诞礼物的时候。两兄弟给年迈的母亲送了去西澳首府珀斯的往返机票。母亲打开信封，眯着眼睛看着机票，皱着眉大声抱怨道：“我为什么要大老远跑去珀斯？”她住在墨尔本，位于东澳，一辈子都没有飞越过整片国土跑到西澳度假，也没兴趣这么做。

为了纪念这次旅行，我把科廷作为预测城市，看看通过24座同样偏远的美国城市的每日最高和最低气温，运用数据挖掘软件来预测科廷每日最低气温的准确率有多高。我的数据挖掘查到华盛顿州的奥玛克，这是一座冬冷夏热的美国小城市，常住居民不足5 000人。其当日

最高气温与西澳科廷机场次日最低气温的相关系数为-0.77（如图5.3所示）。

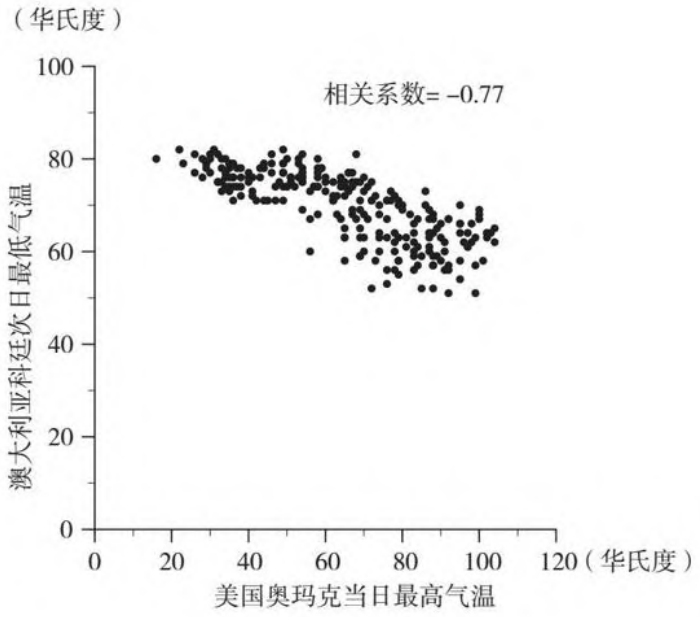


图5.3 根据奥玛克来预测科廷天气的散点图

奥玛克的当日最高气温与科廷次日最低气温呈负相关关系，是因为奥玛克位于北半球，科廷位于南半球。考虑到这两个城市位于不同半球，-0.77的相关系数非常令人震惊。

不会思考的数据挖掘程序（所有数据挖掘程序都不会思考）可能会得出结论，这是一次知识发现，为预测科廷的气温找到了有力的工具。而在会思考的人类看来，预测澳大利亚一个小镇次日最低气温的最佳方法居然是根据远在华盛顿的一个小镇的当日最高气温，这简直荒唐可笑。

我在搜集的另一组数据中启动数据挖掘软件，很快发现了更加紧密的相关系数。如图5.4所示，科廷的每日最低气温与第58号随机变量的相关系数为0.81。没错，图中横轴的变量就是我用计算机随机数字生成器得到的、预测股价的那100个变量之一。

这些虚假变量的生成完全与科廷的天气无关，但我还是发现了一个变量（第58号随机变量）恰好与科廷的天气紧密相关。这就像抛硬币和其他随机噪声那样，通常都会得到看似真实但实则毫无意义的模型和相关系数。

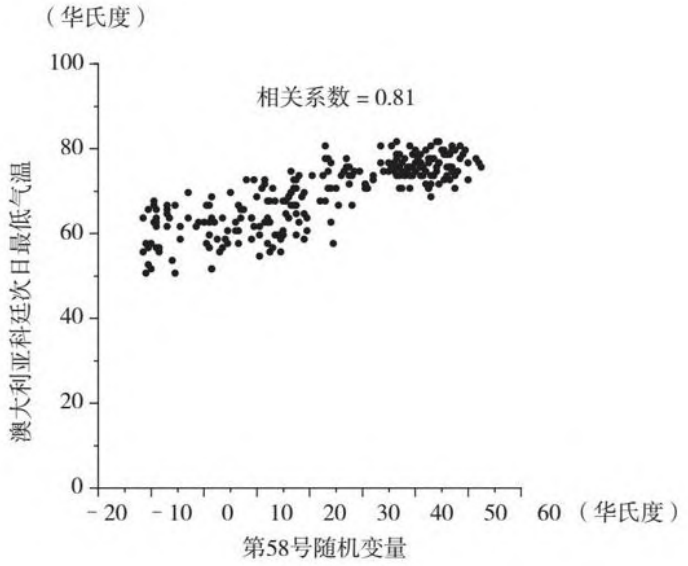


图5.4 随机选择的力量

我这才试了100个随机变量。有了现代计算机，我还可以轻而易举地尝试数千、数百万个随机变量，直至偶然发现一个与科廷或其他城市的气温存在极其紧密相关关系的系数。

那么，我到底证明了什么？根本什么都证明不了。这就是关于数据挖掘需要记住的第一点，无论是否存在真实情况，只要仔细审查大量数据，就能得到统计学模型。此外，即便被称作人工智能，数据挖掘软件也不足以智能到分辨反映出真实关系和偶然关系的模型有何不同，这唯独人类能做到。

史密斯测试

假设数据挖掘算法发现美国股价与澳大利亚科廷的每日最低气温相关。计算机程序怎么会知道这一统计学关系是真实存在的还是偶然的呢？相反，人类知道何为股价，何为气温，还知道股价高低不由科廷的气温来决定。

计算机能搜索stock的定义，尽管该词有多项词义，如存货、家畜和肉汤等。计算机即便能找出正确的定义，也不知道这个定义中用到的词语是什么含义，虽然它还能继续搜索到定义中每个词语的定义。除了搜索定义外，计算机无法知道股票、股票交易和股价真正代表什么，也不知道为什么股价会时涨时跌。它不明白科廷的最低气温为何物，也不明白为什么这些气温有可能或不可能与美国股价相关。

计算机程序可以搜遍已发表的研究数据库，寻找提及股价与澳大利亚气温的文章。但是对计算机来说，要解释碰巧包含这些词语的研究的相关性，则是难于登天（或是无稽之谈）。此外，评定研究是否有效，对计算机来说也是难上加难。约翰·约安尼季斯曾很有说服力地指出：“大多数已发表的医学研究都有误，包括发表在最负名望的医学杂志上的研究（因为报告结果通常都通过数据挖掘的方法获得）。”我相信，大多数的股市研究也一样。我们会在后面的章节中探讨这些论点背后的推理过程；目前的重点在于，用计算机搜索词语“股价”和“澳大利亚气温”，不可能找出任何被它自己解释为支持或反对其发现的统计学模式的内容。就算确实有所发现，计算机也很难评估其可靠性。

另外，“知识发现”的整套言论都在说，计算机会发现崭新的、从不为人所知的模型和关系。根据这一定义，“知识发现”并非已经发表的事物。那么，没有智慧和常识的计算机又如何能辨别出它的“知识发现”是否合理呢？它做不到，因为计算机确实没有智慧，也没有常识。

我们回到前述的汉语室测试。如果计算机不能真正理解“股价”和“气温”在现实生活中所代表的意思，那么它就不能分辨出其发现

的统计学模型是有意义的，抑或只是巧合而已。可以将这种分辨能力称为“理论性知识”“人类本能”“经验”“智慧”“常识”，不过，通过数据发现统计学关系的计算机和无须数据就能预测关系的人类之间，存在根本差别。

我斗胆提出史密斯测试：

> 搜集100套数据，例如，美国股价、失业率、利率和米价、新西兰蓝色涂料的售价，以及澳大利亚科廷的气温等数据。让计算机自由分析，然后报告它认为可能有助于预测的统计学关系。如果人类专家小组一致认为计算机选择的关系合理，则计算机通过史密斯测试。

有可能存在真正的“知识发现”，即计算机能找到人类忽略的合理关系。但是，如果计算机选择的关系被人类认定为无意义，如美国股价和澳大利亚科廷的气温之间的关系，则其无法通过测试。

第6章

如果你拷问数据的时间足够长

最近，我收到一封推荐自动化研究工具的邮件：

亲爱的史密斯教授：

我们冒昧地向您推荐（我们的）全新研究工具……可以使您基于官方统计时间序列数据库的实证研究自动化。（我们软件的）设计目的是让您能从电脑桌面直接探索和发现激动人心的新经济相关系数。

无须其他软件，不用慢吞吞地手动处理数以千计的数据库。您立马就可以开始运行您的第一个大数据项目。

这封邮件接下来一直吹嘘他们的软件可以计算“数百万统计时间序列的相关系数”“识别无法预计的相互依存关系”，以及“开辟新途径”。

除了有创意的遣词造句，让人沮丧的是，他们认为我真的想要筛选数以万亿计的相关系数，寻找意想不到的模型。但意想不到的模型毫无逻辑基础，对于不合逻辑的模型我向来持怀疑态度。

统计学实验都假定研究人员先在脑子里有定义明确的理论，再搜集相应数据来验证自己的理论。而给我发邮件的这家公司则假定，我非常渴望，也很愿意花一大笔钱另辟蹊径，看看每一个可能出现的相关系数——不在乎它们是否合理，然后找出在统计学上最有说服力的那一个，声称这是自己的发现。

这是时代的信号，但并非振奋人心的信号。

孟德尔的豌豆研究

很多重要的科学理论都始于试图对观察到的模型做出解释。例如在19世纪，大多数生物学家坚信，父母特征的平均值决定了子女的特征，如孩子的身高是父母身高之和的平均值，但这也会在环境影响下有所增减。

然而，格雷戈尔·孟德尔在自己进行的豌豆实验中发现并非如此。1822年，孟德尔出生在奥地利，在自家农场长大。父母希望他能接管农场，但孟德尔是个天资聪颖的学生，后来成为奥古斯丁修道院的僧侣，该修道院以其科学图书馆和研究闻名遐迩。

或许是因为生长于农民之家，孟德尔在修道院的八年时间，利用那里的园子一丝不苟地进行了数万次豌豆种植研究。他观察到几种不同的性状，认为融合理论（blending theory）无法解释其实验结果。在他将黄籽豌豆和绿籽豌豆交叉授粉后，生长出来的豌豆籽要么绿，要么黄，并没有黄绿融合，也没有黄绿各一半。同样，在他将滑皮豌豆和皱皮豌豆交叉授粉后，生长出来的豌豆籽要么滑，要么皱，既不会滑皱融合，也不会滑皱各一半。

为了解释该实验结果，他提出了如今众所周知的“孟德尔遗传定律”。孟德尔认为，基因的存在形式不止一种（现称作“等位基因”），例如黄籽的等位基因或绿籽的等位基因，后代拥有来自每种性状的一对等位基因，每个等位基因都独立遗传自父母一方，其遗传到的等位基因可能相同（纯合的，homozygous），也可能不同（杂合的，heterozygous）。在孟德尔的实验中，控制颜色的等位基因杂合时，豌豆为黄籽，所以黄籽是显性等位基因，绿籽是隐性等位基因。只在两个等位基因均为隐性时豆籽才呈绿色。

孟德尔谨慎构建的理论不仅适用于其数据，也合情合理。即便计算机当时已经出现，也未必能完成得这么好。我们观察不到等位基因，所以计算机也不会产生这种妙想——认为有等位基因存在、父母一方均有两个等位基因、后代从父母各遗传一个等位基因、等位基因有可能是显性或隐性。

现在，我们知道因为由显性等位基因来决定性状，所以有些性状不完全外显。当出现不完全外显的情况时，性状呈现混合状态，例如，开红花和开白花的金鱼草杂合，会长出开粉花的金鱼草。郁金香为共显性，开红花和开白花的郁金香杂合，会繁殖出既开红花又开白花的郁金香。

孟德尔创建了符合所采集数据的理论，以此为现代遗传学奠定了基础。然而，“数据为先，理论靠后”的准则也成为成千上万冒牌理论的来源。你听到过名字以字母D开头的棒球手会比以字母E~Z开头的棒球手更早逝的说法吗？或者听到过亚洲人更容易在每月4号心脏病发作的说法吗？

得州神枪手谬误

“数据为先，理论靠后”特有的两个问题可以精确地概括为“得州神枪手谬误”。

谬误1：自诩为神枪手的人在整面墙上放满靶子，然后朝墙开枪。他肯定能射中一个，然后很骄傲地显摆自己的枪法，绝口不提其余没射中的靶子。因为他肯定可以击中一个，所以即便做到了也根本说明不了什么。这好比在研究中验证数百个（或数千、数百万个）理论，然后只报告统计学意义上最有说服力的结果，对所有失败的验证都守口如瓶。例如，在超感官知觉研究中，有人可能会给数千个受试者进行数十次验证，却只报告那些支持超感官知觉的测试（或测试的一部分）。这什么都证明不了，因为只要进行足够多的测试，研究人员就一定可以找到支持的证据。

谬误2：倒霉的牛仔把子弹打到了空白墙上。随后，他绕着弹孔画了个靶心。这也证明不了什么，因为总能找到一个弹孔画

圈。这就好比在研究中搜遍数据来寻找模型，找到后再编出一个理论。超感官知觉研究中，有人可能报告称，尽管受试者的反应与当时记录的线索不符，但的确与早前记录（“向后移位”）、稍后记录（“向前移位”）或无记录（“负超感官知觉”）的线索吻合。想要寻找模型的人肯定能找到一个。因此，有所发现只能证明有所寻求。

得州神枪手谬误还有很多其他表达方法，包括数据挖掘、数据捞取、摸底调查、采摘樱桃、数据探测和P值篡改。P值篡改这一戏称源于实验结果偶然发生的可能性（P值）若低，则被认为具有统计学意义。“得州神枪手”的研究中低P值的可能性比较高，所以称之为“P值黑客”。

发表研究成果的期刊加剧了这种状况，因为它们更喜欢（或需要）有统计学意义的结果，这就促使研究人员落入得州神枪手谬误，以获取具有统计学意义的结果。

谬误2又称“费曼陷阱”，名称取自诺贝尔奖得主理查德·费曼。费曼让自己在加州理工大学的学生计算，如果他走出教室，在停车场看到的第一辆车的车牌号为8NSR26的概率为多少。学生们假定每个数字和字母出现的概率相同且独立确定，得到的概率结果为小于1 700万分之一。等到学生完成计算后，费曼揭晓正确答案为1，因为他在来教室的路上已经见到了这个车牌号。发生概率微乎其微的事情，如果已经发生，那么它发生的概率就肯定不是微乎其微了。

对上述两种得州神枪手谬误的精辟概述，可参见诺贝尔奖得主罗纳德·科斯的辛辣言辞：“只要拷问数据的时间足够长，它就会屈打成招的。”

我会使用“数据挖掘”这一通用术语来说明得州神枪手谬误和类似的胡闹做法。

数据挖掘者

几十年前，我刚刚进入职场的时候，“数据挖掘者”是种侮辱性叫法，就好比被人指责剽窃一样。如果有人提出从理论上说不通的或好到令人难以置信的结果时（例如，接近完美的相关系数），就会听到强烈不满地反驳：“这是数据挖掘！”

我记得在一场研讨会上听到一名访问教授称，他创造了一种离奇的货币供应测量工具，令其与美国经济高度相关，因此这一测量工具与美国经济之间存在密切的相关关系。这种密切的相关关系证明不了什么，因为是他自己一手创造的。

可能是碍于这位来宾的身份地位，台下的听众都异常克制。不过，一出研讨室的门，很多人都低声吐槽这是“数据挖掘”。这不是什么好话。多年后，我重新审视那名教授的模型，发现尽管他在所选的那段历史时期内，能够捏造出那个愚蠢的货币供应测量工具，得到密切的相关关系，但是在那之后的时期根本就对不上号。我对此并不惊讶，因为他的模型没有理论基础。他那时就是个数据挖掘者，他的模型也毫无价值可言。

如今，人们四处宣扬自己是数据挖掘者，用无法思考的计算机进行机械式的计算，还昧着良心为此收取费用。那封诚邀我在几百万数据中浪里淘沙，寻找意料之外的统计学关系的邮件的主人，也认为我会掏腰包加入他们的行列。

诺贝尔经济学奖得主詹姆斯·托宾挖苦地调侃道，在那个痛苦的旧时代，研究人员不得不进行人工计算，这其实是好事。用今天的话来说，这是特点，不是缺点。计算太难了，所以人们在动手前会深思熟虑。但现在有了以TB（太字节）计算的数据和以闪电速度运行的计算机，太有利于“先计算，后思考”了。这就是缺点，而非特点。在计算前，绞尽脑汁思考会更好。

但谁还会思考？放手让软件搜遍数据，寻找始料不及的关系就好了。

QuickStop

互联网营销公司QuickStop拥有100多万个域名，以此招揽无意间的网络流量。假设，有人在查找Hardle Soup Company（其网址是www.hardlesoup.com），但将网址错误输入为www.hardle.com或www.hardlysoup.com，页面就会转到QuickStop兜售产品的首页。

QuickStop的高级经理认为，如果将首页惯用的蓝色改为绿色、红色或青色，公司的收入可能会随之增加。该公司的数据分析师建了四个版本的网页，测试数周后得出的结论是，其他三种颜色相较于惯用的蓝色都没有显著提高公司收入。那位高级经理大失所望（毕竟这是他的想法），又提议将数据按国家分类，可能就会找到差异。或许，人们偏爱本国国旗的颜色或与本国人性格相匹配的颜色？

果然，按国家进行数据分类后，发现其中一个国家确实有颜色偏好——英国大爱青色。大不列颠是个岛国，可能青色会让英国人想起大不列颠岛四周的海域？那位高级经理认为，原因不重要，重要的是英国人看到青色的网站首页时，会愿意买更多的东西。

当QuickStop正准备将英国版本的网站首页换成青色时，公司有名的数据分析师出面叫停了这个不切实际的想法，并称之为“数据犯规”。这纯粹是数据挖掘，QuickStop并没有怀着“英国是唯一看重网站首页颜色的国家”这个想法来启动实验，更别说“青色是英国偏爱的颜色”了。通过给大约100个国家展示其他三种颜色的页面，该公司的人保证会发现某些颜色有助提高来自某些国家的收益。涉及的颜色和国家越多，他们就越有可能得到有意义的发现——但这不过是巧合罢了。

那名数据分析师坚持继续测试，搜集了更多关于英国的数据。青色带来的影响消失后，分析师并不感到惊讶。事实上，首页为青色时的收益比目前使用蓝色的收益还要低。

这就是得州神枪手谬误1：测试大量理论，然后重点关注结果中最有意义的测试。以下为谬误2的例子。

另一家互联网营销公司TryAnything向填写网站首页问卷的访问者免费赠送样品。随后，TryAnything运用数据挖掘软件寻找回答问卷与收到样品后回购产品的用户之间有何相关系数。软件惊奇地发现，购买行为与消费者姓氏的长度有关联——姓氏中包含九个以上字母的人更有可能购买产品。TryAnything的经理认为姓氏长度可能代表了种族，他们没有询问任何关于种族的问题，担心会冒犯他人，但他们的产品可能对某些群体来说更有吸引力，也可能并非如此。TryAnything的管理层觉得，他们不需要弄清缘由。他们有昂贵的软件和强大的电脑足矣。数据就是他们想要的全部。

如果QuickStop那名心存质疑的数据分析师在TryAnything任职，他肯定也会说这是数据犯规。数据挖掘程序在交叉分析问卷时，并没有特定的理论指导。问卷有很多问题，肯定会有某些回答与产品销售相关——这不过又是巧合罢了。可惜QuickStop那名数据分析师没有为TryAnything效力，所以该公司没能避免数据犯规的恶果。TryAnything挑出所有长姓氏的答卷对象，然后指派销售人员逐个进行电话访问。注意，可不是预录制的宣传电话，都是领工资的真人拨打的真正的通话。这招最后一败涂地，人工拨打电话耗资巨大，几个月后便弃而不用了。

拷问数据

严谨的研究人员真的会拷问数据吗？大有人在。正因如此，曾经备受推崇的大神才会提出使其如今名誉扫地的想法，如喝咖啡会导致胰脏癌、患病之人能被千里之外的医者的正能量治愈、以女性名称命名的飓风更具杀伤力。

心理学研究生尼克·布朗称，一名高年级研究生给他提了一个似乎很有帮助的建议：“你想知道怎么做吗？我们有很多不成熟的想法，然后做大量的实验，无论我们得到什么数据，都假装这正是我们想要的。”

这种草率的态度和想尽办法获取可以发表的结果的压力，导致发表了很多牵强附会、站不住脚、纯粹垃圾的研究。美国可重复性项目（Reproducibility Project）计划复制100项已在三家颇具声望的认知和社会心理学期刊发表的研究，但只成功复制了36项。我们会在后面的章节讲述对医学期刊的统计情况，它们也好不到哪儿去。这些都是顶级期刊中的最佳研究，其他次级期刊中的统计结果也就可想而知了。

学术性研究和商业性研究的最大区别在于，学术界人士可以通过发表研究成果名利双收，而商界研究人员通常要对研究项目保密，因为这些项目都很有可能成为有利可图的专利信息。谷歌不会分享它的搜索算法，高盛也不会分享它的股票交易算法。

然而，这两个领域的研究人员也有相同之处。事实上，很多人都同时涉足这两个领域，比如，我会发表自己的学术研究成果，也会提供不公开发表的咨询服务。相信很多人和我一样。

从事学术性和商业性研究都要接受同样的训练，采用同样的技术，承受同样的压力。学术界研究人士提出有意思的实证发现后会受到嘉奖，商界的研究人员也享有同样的待遇；前者如果没有新的发现或有意思的成果就会受到冷落或面临经费短缺，但至少还抱着铁饭碗，而后者如果一无所获就只能卷铺盖走人。

大多数拷问数据的商业性研究都因私有协议保护而未公之于众，同时，学术期刊上也发表了很多采用拷问数据方法的研究。下面我们就来深入了解一下数据是如何被虐的。

倒摄回忆

有些研究人员不仅对拷问数据毫无歉意，甚至还鼓励这种做法。著名的社会心理学家达里尔·贝姆写道：

> 传统的研究过程是先从某个理论中提取一组假说，接着设计和进行研究来检验这组假说，分析数据看看这组假说是被证实还是被证伪，然后按照事件发生的时间顺序写成期刊文章……但如今的研究过程已经不是这样了。心理学的研究过程比这样的过程更加激动人心。

他继续写道：

> 从每个角度检验（数据），分析不同性别，编造综合指数。如果有数据预示新的假设，设法从不同数据中找到更多证据。如果看到有意思的模型的蛛丝马迹，尝试通过重新组织数据将其变成更明显的特征。如果出现你不喜欢的参与者，或向你提供异常结果的实验者、观察者或访问者，暂时将他们放在一边，看看是否会出现清晰的模型。继续进行摸底调查，寻找那些有意思的发现，来者不拒。

贝姆采用“摸底调查”的方法，的确能够发现一些令人难以置信的事情。2011年，他在一篇名为“感知未来”（Feeling the

Future）的文章中指出，在电脑屏幕上随机播放色情图片时，有53%的受试者能提前猜到图片会出现在屏幕的左边还是右边。贝姆的摸底调查实验使用了五种类型的图片，他选择强调的是唯一具有统计学意义的那种。引人注目的是，即便运用了得州神枪手谬误1，他能得到的最高概率也只有53%，和抛硬币几乎没有两样。

贝姆更加语出惊人的是，他认为人具有“倒摄回忆”。来自康奈尔大学的本科生自愿参与实验：观看电脑屏幕上出现的48个单词，每三秒看一个。然后，受试学生根据指令可视化该词所指物体（如在看到“tree”这个词的时候，想象树的样子）。看完48个单词后，受试学生被要求输入所有能记住的单词。接着，从这48个单词中随机抽取24个向受试学生展示，并让他们完成各种有助于记忆所抽取单词的任务，例如，点击与树相关的词或重新输入这些词。贝姆的报告称，如果学生在看完屏幕展示的所有词后学习某个单词，那么他们在进行回忆测试时记起该单词的可能性更高。例如，如果他们在屏幕展示后，用点儿技巧记忆“tree”这个单词，那么在回忆测试时，他们记起这个单词的概率就会提高。

仔细想想这其中的言外之意：康奈尔大学的学生整个学期都争分夺秒，到学期结束后开始复习备考，但那时已经考完试了；互联网广告商可以在消费者做出是否购物的决定后，通过广告宣传来提高销量。

不出所料，其他研究人员无法得出和贝姆一样的结果。发表了贝姆《感知未来》一文的期刊，一年后又刊登了题为“痛改前非”（Correcting the Past）的文章，由来自四所大学的四名教授联合撰写。他们在文中表示，有七项实验试图复制贝姆的“人能感知未来”这一观点，但“并没有发现支持这一说法的证据”。对贝姆牵强观点的确凿反驳可谓火上浇油，激起了更大范围的质疑——这种论文怎么就在享有声望的期刊上发表了？此外人们还得知，尝试反驳贝姆观点

的其他研究都被期刊拒绝发表，因为这些期刊都有政策规定，不发表复制性实验；还有一种情况是，建议不予发表的评委就是贝姆本人。

金钱启动效应

我受邀参加谷歌的年度学术性聚会Sci Foo 2015，约有250名科学家、作家和政策制定者齐聚谷歌在加州山景城的公司总部。开幕当晚，在与一名社会心理学家交流的过程中，我了解到一些有趣的实验。将3~6岁的波兰孩子（这个年龄段的群体对金钱没有概念）分为两组，要求其中一组按颜色给纸币分类，另一组则按颜色给纽扣分类。然后让这些孩子都进入另一个房间完成迷宫图。结果发现，给纸币分类的孩子的完成度和成功率更高（工作效果更佳）。

第二项实验是，让孩子穿过房间把红色蜡笔拿回来。基于第一项实验的结果，我猜给纸币分类的孩子会拿回更多蜡笔，因为他们似乎更喜欢工作。但结果并非如此，给纸币分类的孩子拿回的蜡笔更少。据此，研究人员认为给纸币分类的孩子更不乐于助人。

这些研究人员在其发表的文章中表示，处理金钱“增加了劳动力付出，降低了乐于助人和慷慨大方的程度”，这些影响“不由金钱的面值、孩子对金钱的认知或他们的年龄决定”。

我对整个项目的第一反应就是心存质疑。不知道何为金钱的3岁小孩，怎么才碰一下钱就被影响了？对第二项实验结果的解释，更是加重了我心中的疑团。如果给纸币分类的孩子能取回更多蜡笔，那么这或许可以作为他们更能干的又一证据。他们带回的蜡笔更少，我的解释为他们并不能干。然而，为图方便，这被解释成他们不乐于助人。在我看来，这就像是得州神枪手谬误2。

我问这名社会心理学家对贝姆的研究有何看法，她表示赞赏。我紧接着又问她如何看待承认自己捏造数据的社会心理学家德里克·斯塔佩尔，她对斯塔佩尔也称赞了一番。于是我转移了话题。

第二天，我在谷歌总部听到，化学家、生物学家、天体物理学家等科学家都表露出自己对“复制危机”的担忧——构思奇特的观点可以在颇负声望的期刊上发表，却无法复制，还谈到这种现象会给科学研究的可信度带来的伤害。其中有位卓越的社会心理学家表示，自己所处的领域就是不可复制研究的典型代表。他的缺省假设是，该领域发表的任何成果都有误，其原因显然是太多社会心理学家都不明白一个最根本的事实——遍搜数据从而发现一个模式，除了能证明他们为了寻找模型而搜遍了数据以外，其他什么都证明不了。我还无意中听到有一小组人在讨论，“孩子看过或摆弄过纸币就会被影响”的说法让人难以置信。

回到家后，我上网四处搜索关于金钱的其他研究，发现有大量文献讲到了启动（priming），分析短时记忆是如何将反应与一系列刺激物联系起来的。例如，如果让受试对象说出以rep开头的单词，假如他们最近见过repair（修理）这个词，便可能回答repair。同样，如果问他们rebuild（重建）是不是一个单词时，在上一个单词是repair的情况下的回答速度比上一个单词是swim（游泳）的快。这些结论都合情合理。

而“金钱启动”的说法就有些牵强附会，例如，小孩子接触金钱后变得更努力、更自私。我发现，有关“金钱启动”的论文都提出，成人在看到与金钱有关的语句（如“我们付得起这个钱”）或电脑屏幕背景图中金钱的模糊图像后，其行为均会受影响。有篇文章（和我在谷歌总部交流过的那名社会心理学家为合作作者之一）的标题颇具争议——“仅接触金钱就会增加对自由市场体制和社会不公的支持”（Mere exposure to money increases endorsement of free-market systems and social inequality）。文章作者指出：

相对于不提金钱概念来说，对金钱概念的细微提醒，都会让参与者整体上更加支持美国现行的社会制度（实验1），尤其是自由市场资本主义制度（实验4），更坚信受害者应该遭此命运（实验2），更相信社会优势群体应该统治社会劣势群体（实验3）。

真是言之凿凿！太过武断，以至于我很肯定事实并非如此。

随后，我冒出了一个很大胆的想法——复制这篇论文的实验。结果显示，找不到该文章提到的任何影响。得州神枪手谬误1发挥了作用，因为最初的论文忽略了那几个并不支持其结论的测试。金钱启动类文章的作者们也予以反驳，激起了第三组科学家加入此次争辩，最后得出了更加确凿的结论：“选择偏倚（selection bias）、报告偏倚（reporting bias）或P值篡改歪曲了”所谓的金钱启动效应。

寻找就会发现

飓风“鲍勃”来袭的时候我住在科德角。新英格兰地区史上较具毁灭性的飓风之一竟然起了如此单纯善良的名字。我也万万没想到，多年以后，伊利诺伊大学的研究人员竟会认为飓风的名字至关重要。

第一次有记者问我对题为“女性名称的飓风比男性名称的飓风更有杀伤力”（Female Hurricanes are Deadlier than Male Hurricanes）的文章有何感想时，我就表示持怀疑态度。1979年以后，飓风就开始轮流使用男性和女性的名称，似乎很难相信，飓风的名称和所造成的死亡人数会存在相关关系，而不止是偶然事件。

原来，虽然文章题目如此，但作者的论点并不在于女性名称的飓风威力更大，而在于人们并不把女性名称的飓风当一回事，不认为它们会带来致命后果，于是没有做好万全准备，这才导致更多人丧命。

这项研究发表在《美国科学院院报》上，所以肯定没问题，对吗？未必。很多蹩脚的科学研究都曾在著名期刊上发表。

仔细看完这项研究后，我发现了几个更具说服力的质疑理由。

在1950年、1951年和1952年，飓风名称都取自军事语音字母表［如Able（埃布尔）、Baker（贝克）、Charlie（查利）等］。1953年，所有飓风都被换成女性名称。很多女性主义者都公开谴责这种针对女性的性别歧视，罗克西·博尔顿表示：“女人不是灾难，不会摧毁生命和社区，留下持久不消的惨烈后果。”1979年，便开始采用男性和女性名称轮流使用的方法，延续至今。

那项发表在《美国科学院院报》的研究也包括1979年之前的数据，这期间的飓风都采用女性名称命名。将1979年之前的数据都包括在内是很有问题的，因为这期间每场飓风造成的平均死亡人数为29人，1979年之后为16人。女性名称的飓风造成的平均死亡人数整体上升，也是由于1979年以前的飓风更具杀伤力，而当时所有飓风都采用女性名称。或许，早年死亡人数更多是因为飓风威力更强（在所有飓风都采用女性名称期间，飓风平均强度为2.26级，之后的时期为1.96级）、建筑物更脆弱、提前预警更少。

没有确切方法能比较1979年前后的暴风预警，但是有一则传闻性证据说明如今的提前预警更频繁。1938年9月20日，马萨诸塞州斯普林菲尔德市的报纸《斯普林菲尔德联合报》刊登了该州西部的天气预报：“今日有雨，明日或将有雨。”第二天，1938年的“大飓风”（Great Hurricane）来袭，造成该州99人丧命。斯普林菲尔德市的康涅狄格河水位超过了洪水水位。一共有将近700人丧生，财产损失估计相当于2015年的50亿美元。75年后，马萨诸塞州汤顿市美国国家气象局气象专家表示：“真想不到1938年会发生那样的飓风，毫无预警。”

美国国家海洋及大气管理局（NOAA）于2012年自夸道：

> NOAA针对海洋及大气研究的投入，加上技术的发展，使得针对飓风的监测和预警获得了令人瞩目的进步。投入和技术的结合让我们拥有复杂的计算机模型、基于地面和海洋的庞大感应器网络、卫星以及飓风搜寻飞机……过去半个世纪取得的进步，大大推动了飓风预警的发展。尽管沿海居民人口不断增加，飓风造成的死亡人数还是出现了大幅下降。

即便知道NOAA是在自我推销，也能明显看出，将1979年以前飓风的危险程度等同于最近时期的是潜在的误导行为。分析1979年以后轮流使用男性和女性名称的飓风更具有科学效力。

《美国科学院院报》的研究还忽略了几次致命的飓风。一份报告描述了2009年的飓风“比尔”：

> “比尔”生成的长风、巨浪和激流在美国造成了两人死亡。虽然沿岸地区都发布了危险巨浪的预警，但8月24日，还是有1万多人聚集在缅因州阿卡迪亚国家公园的海边观看这场飓风。一波海浪将20人卷入海里；11人被送入院，一名7岁女孩身亡。再看其他地方，佛罗里达州新士麦那海滩上，一名54岁的游泳者被巨浪冲到岸上不省人事，随后丧命。

飓风“比尔”并没有被包含在《美国科学院院报》那项研究的数据中，是因为飓风“比尔”并没有真正登陆。上述插曲也能说明人们并没有把飓风“比尔”当成一回事。

该研究论文的作者写道，在他们分析的数据中，女性名称的飓风在重大风暴期间“导致”更多人丧生，但是“男性名称的飓风和女性名称的飓风对于威力较小的风暴没有影响”。这似乎说反了。分别以男性名称和女性名称命名的大飓风所造成损失的差异应该更小，以飓风“桑迪”为例，它是研究的数据中1978年后最致命的飓风。

首先，人们普遍认为“桑迪”是男女皆宜的名称，但是《美国科学院院报》的作者觉得它非常女性化。他们让九个人给不同的飓风名称按男性化或女性化程度打分（1~11分），报告称“桑迪”的平均得分为9.0（非常女性化），比“伊迪丝”（8.5）、“卡罗尔”（8.1）、“比拉”（7.3）得分还高。肯定有什么地方不对劲。我调查了44个人，得到的平均分为7.25，这个结果更加合理。

2012年10月24日，飓风“桑迪”在牙买加登陆，导致两人死亡，整座岛上70%的电力中断；10月26日在古巴登陆，风速为155英里/小时，造成11人丧生，摧毁了1.5万多座房屋。此次飓风还导致多米尼加共和国两人死亡，海地54人丧命，波多黎各一人丧生。“桑迪”于10月29日在新泽西登陆前，有九位美国州长宣布该州进入紧急状态。纽约市市长迈克尔·布隆伯格下令暂停全市所有公共交通服务（包括公共汽车、地铁和铁路列车）、关闭公共学校，并在该市多个地区进行强制疏散。

尽管如此，纽约市还是有48人遇难，在美国其他地区还有总计109人丧生。虽然“桑迪”在袭击美国之前就已经导致数十人死亡，而且在任政府官员也采取了一系列格外周全的防范措施，但民众还是没有把飓风“桑迪”当一回事，就因为他们觉得“桑迪”是个女性化的名称。这种说法真的可信吗？

如果这个隐含性别歧视的理论为真，那么它在危险性不确定的风暴上应当体现得更加明显。对一个可能的世纪性大风暴的反应（新闻媒体上还播出了灾难预警）取决于飓风名称被认为是女性化还是男性化，这太说不过去了。而更加合理的说法是，没那么严重的风暴（如飓风“比尔”）更有可能被草率地当作小麻烦，而不是大危险。

有趣的是，飓风“桑迪”在新泽西登陆时，已从飓风降级为后热带气旋。《美国科学院院报》文章的作者应该将未登陆的致命飓风考虑在内吗？应该考虑致命但登陆时未达到飓风级别的风暴吗？谁知道

呢？不过，这才是重点。作者还称，他们用各种变量组合推算出大量模型（得州神枪手谬误1）。

如果强有力且出人意料的结论是通过拷问数据得出的，那么就审视一下用以拷问数据的无数决定，看看这一结论是否经得起考验，这样做会有启发意义。我运用了1978年之后范围更广的数据来测试文中所述结果的稳定性，包括热带风暴、飓风、太平洋风暴、未登陆或在其他国家登陆的和无美国公民伤亡的风暴。

表6.1和表6.2直接对比了男性名称和女性名称的飓风频次，对比情况包括导致死亡、死亡人数为1~99人、死亡人数超过99人。结果并未显示出一致模型，更别说具有统计学意义上的显著差异了。平均死亡人数的对比表明，男性名称的风暴造成的平均死亡人数通常更高，尽管这并不具有统计学意义上的显著差异。

表6.1 大西洋飓风和热带风暴造成的死亡人数

	风暴数量（个）		平均死亡人数（人）	
	女性名称	男性名称	女性名称	男性名称
风暴总数	210	210	36	131
导致死亡	111	103	68	267
死亡人数为 1~99 人	103	88	12	14
死亡人数超过 99 人	8	15	792	1 752

表6.2 太平洋飓风和热带风暴造成的平均死亡人数

	风暴数量（个）		平均死亡人数（人）	
	女性名称	男性名称	女性名称	男性名称
风暴总数	293	286	4	8
导致死亡	42	46	28	51
死亡人数为 1~99 人	39	42	9	7
死亡人数超过 99 人	3	4	271	519

《美国科学院院报》的研究文章断言“女性名称的飓风比男性名称的飓风更有杀伤力”，这一说法经不起推敲，显然因为它采用了可疑的统计学分析方法，数据定义狭隘，还选择了所有飓风都使用女性名称的年份。

为什么优秀期刊会发表漏洞百出的研究呢？有些期刊没有资源进行仔细审查，有些喜欢通过发表具有争议的文章提高知名度，有些则乐于接受证实他们偏见的文章。

微笑曲线

1980年，罗纳德·里根当选美国总统时，个人所得税最高税率为70%。当时采用的是边际税率，对征税对象收入超过特定水平的部分征税，而收入低于某一临界点则税率更低。对单身纳税人来说，70%的税率适用于收入高于108 300美元的（相当于2016年的302 000美元）纳税人。对于已婚夫妻来说，70%的税率起征点为215 400美元（相当于2016年的600 000美元）。实际上，富裕之家通常能找到方法减轻税负，但是，70%的税率还是令人咋舌。里根提议减少个人所得税，这一政策即后来为人熟知的“里根经济学”。

1981年，我与经济学家“约翰”共进午餐，他创办了一家咨询公司，为里根的经济团队出谋划策。很多人，甚至传统的共和党人都担心里根的减税提议会增加政府的财政赤字。里根的团队对这些担忧不予理会，表示需要采取全新的方法——他们所谓的“供给学派经济学”（Supply Side Economics）。

传统的凯恩斯主义需求学派经济学家指出，商业周期取决于总需求的变化，家庭支出缩减会引起经济衰退，导致其他人失业并削减开支，这就像滚下山的雪球，越滚越大，越滚越快。在需求学派的模型中，减税可减缓或阻止经济衰退，因为税收减少可促进消费。

与此相反，供给学派经济学家更关注工作意愿，而非消费意愿，强调个人所得税过高会削弱工作积极性。如果香蕉的税率上涨，人们就不会买那么多香蕉。如果所得税上涨，人们也不愿意努力工作挣钱。

大道通常至简。图6.1所示为拉弗曲线，据说是经济学家阿瑟·拉弗在饭店餐巾纸上画出来的。

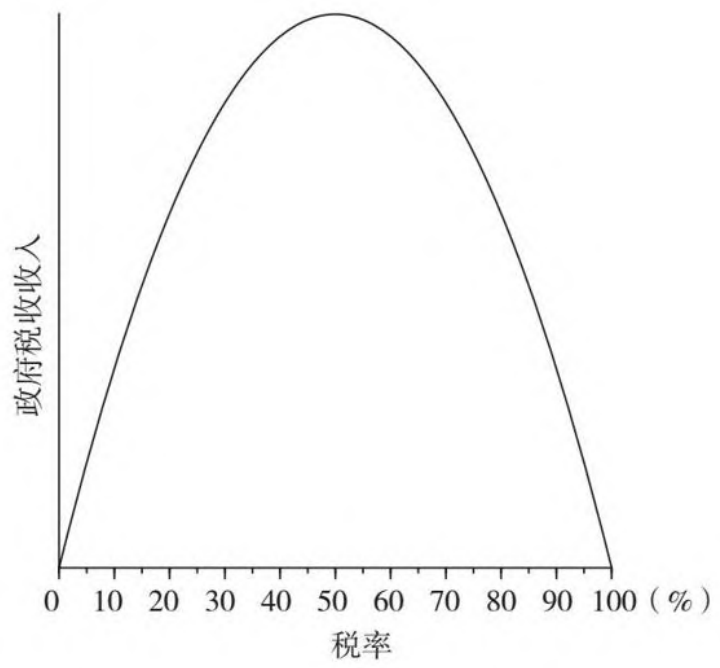


图6.1 拉弗曲线

图6.1中的两个极端情况显而易见。税率为0%时，政府无税收收入。税率为100%时，无人愿意工作，所以政府也无税收收入。税率处于两个极端之间时，有人愿意工作，政府也就可获得一些税收收入。如果税收收入随着税率从0%上升而上升，则肯定会在达到某一点后开始下降，才能回到税率为100%且政府无税收收入的那个点。

有批评者称此为“搞笑曲线”。该曲线没有理由一定是最高值位于税率为50%时的对称形状，也不一定是只有一个峰值的平滑曲线。供给学派的论点还忽略了一个事实，选择工作时长或辞去工作对大多数人来说只是奢望。还可以想象到的是，更高的个人所得税率会迫使一些人需要兼职第二份工作，才足以支付房贷和其他账单。

同时考虑需求和供给不失为好方法（经济学家又怎么会相信呢？），但是减税给需求带来的好处似乎被需求学派的支持者夸大了。不过在1981年，我还是很乐意与约翰共进午餐的，这让我有了意料之外的收获，大开眼界。

约翰的公司当时正有偿开发一个计算机模型，用以预测税率降低时政府税收收入的增加情况。随后，里根政府便可引用这个模型来证明减税会降低政府的财政赤字。

约翰知道这个模型的使用目的，但就是不知道如何将其创建出来。他试过了所有类型的模型，但是每当他使用真实数据来推算时，该模型的预测结果就指出减税会降低税收收入。他邀我吃午餐是因为他迫切需要建议。可能我会提议采用一个新模型？或者他只针对某一个历史时期？又或许可以看看其他国家的情况？

我建议他应该接受减税会减少税收收入这一事实，他对此感到不满。

里根政府从未公开约翰的模型，因此我猜测他拷问数据也没有得到想要的结论。尽管如此，里根还是如愿实行了减税措施，最高税率从70%降至50%，再到38.5%，最后，在里根于1988年卸任时，税率低至28%。

然而，正如大多数专业的经济学家，包括很多供给学派支持者的预测，所得税收入减少了。

这段经历特别有意思的地方是，里根政府假设，如果是计算机预测税收收入会增加，就会受到更多重视。在这一假想中似乎存在更重要的真相。最近有项实验要求数十名志愿者阅读一段关于虚构药物的描述。其中，一半人的材料里有一张图表，另一半人的则没有。该图表不过是简单直观地表达了材料内容，但还是说服了更多受试者。认为这种药物“的确会减少疾病”的人数比例从67.7%上升至96.6%。科学性似乎会增加可信度。如果是计算机得出的结果，就更应该值得信任。

不怪你，怪我

如此容易上当受骗是因为我们愿意相信如果有人说了我们无法理解的事情，问题肯定出在我们身上。我们假定对方是专家，而我们不够聪明，无法理解这个聪明人在讲什么。同样，如果我们不理解计算机得出的结论，问题也肯定出在自己身上，是我们不够聪明，无法理解计算机表达的意思。

人类这种与生俱来易轻信的特性在杜克大学出版的学术杂志《社会文本》（*Social Text*）中有很好的体现，该杂志还被极力称赞为“站在文化理论最前沿的杂志”。纽约大学和伦敦大学学院物理学教授艾伦·索卡尔向该杂志投了一篇文章，题目让人印象深刻——“超越界线：向着量子引力变革性的解释学迈进”（Transgressing the Boundaries: Towards a Transformative Hermeneutics of Quantum Gravity）。这篇文章着实晦涩难读。例如，“类比拓扑结构在量子引力下出现，但由于涉及的多个方面都是多维的，而不是二维的，更高阶的同源群组会起作用”。该杂志的人文学编辑果然都对此印象深刻并发表了这篇文章。

这其实是场恶作剧。索卡尔故意写出这篇莫名其妙的文章，就是想看看是否会发表，结果真的发表了。

该杂志的编辑评上了“搞笑诺贝尔奖”，该奖每年颁发一次，滑稽可笑的颁奖典礼在哈佛大学举行，目的是选出“乍看令人发笑，过后发人深省”的学术成果。《社会文本》杂志编辑的颁奖词为“急切发表自己看不懂、作者承认胡编乱造且并非真实存在的研究”。

不过，编辑们没有被逗笑，而是共同提出抗议，认为索卡尔的恶作剧违反职业道德，虽然其中一名编辑仍“怀疑索卡尔那篇莫名其妙的文章根本就不成立”。

后来又发生了有趣的事情，康奈尔大学的研究生罗布·维勒做了一项实验，要求参与者基于论点质量和文章可理解度来评价索卡尔那篇莫名其妙的文章的摘录。半数人被告知作者是杰出的哈佛大学教

授，另一半人被告知作者是康奈尔大学的大二学生。不出所料，认为作者是哈佛大学教授的参与者给出的评分更高。

从卓越降为优秀

吉姆·柯林斯花了5年时间研究40年来的1 435只股票，并确认了整个股市中表现最好的11家企业：雅培、美国电路城公司、房利美、吉列、金佰利、克罗格、纽柯钢铁、菲利普·莫里斯、必能宝、沃尔格林、富国银行。

将这些企业与11家股价疲软的同行业公司比较之后，柯林斯识别出五个明显特征，例如，第五级领导力（领导者个性谦逊，但是在职业上追求打造卓越企业）。这些恰恰就是数据挖掘软件在搜索成功企业、持久婚姻、长命百岁等的公式或秘诀后会得出的那种结论。我们知道肯定存在一些共同特征，但即便找到了也说明不了什么。

柯林斯在《从优秀到卓越》一书中做了总结，写道：

> 在本书中，我们直接利用数据进行实证演绎推断，阐释了所有概念。启动此次项目并非为了验证或证明某种理论。我们寻求的是直接从证据下手，凭空创建一个理论。

柯林斯认为他这么做能让自己的研究保持客观中立，不失专业水准。这一切都不是他捏造出来的，他不过是让数据带着走而已。事实上，柯林斯算是承认自己陷入了得州神枪手谬误2，他只是开心地蒙在鼓里，没有意识到数据挖掘的危害——从数据中衍生出理论。

回顾任何类别的企业发展，无论是最好的还是最差的，我们总能发现某些共同特征。你看，柯林斯所选的这11家企业，每一家的名称都含有字母i或r，还有几家同时含有i和r。确保企业名称里含有字母i或r，难道就是企业从优秀走向卓越的关键吗？当然不是。

为了增强其理论的统计学合理性，柯林斯引用了一名科罗拉多大学教授的话：“碰巧找到11家企业都显示出你发现的主要特征，而这些特征都无法通过直接比较得出的概率有多大？”该教授计算得出，此概率不足1 700万分之一。

这是费曼陷阱，也恰巧与费曼让学生计算车牌号得出的概率一致，都是1 700万分之一。先选择企业，后找共同特征，这不足为奇，也很无趣。有趣的是，这些共同特征是否有助于预测未来会有哪些企业获得成功？对这11家企业来说，答案是否定的。房利美的股票从2001年的每股80美元跌至2008年的每股不到1美元。美国电路城公司于2009年破产。就在《从优秀到卓越》出版发行后，其他9家企业的股票总体表现平平，其中5家企业的股票表现优于总体股市，而另外4家则差于总体股市。

选择成功企业再发现共同特征（正是数据挖掘软件所做的事情）毫无意义，因为这等于事后诸葛亮：总会存在成功的企业，也总能找出它们的共同特征。

我做了一个愚蠢的实验来证明这一点。我在波莫纳学院教了好几年统计学，有这些年班上学生的分数。如果可以拿到班上学生特征的数据（包括身高、体重、性别、种族、喜欢的电影和音乐、就读高中的规模、会讲的语言的数量等），我就肯定能找出一些明显特征，并将此归结为在我的统计学课上表现良好的秘诀。我们当然能想出适合这些相关系数的解释：身体质量指数（BMI）低的学生更加活跃，就读的高中规模小的学生更加自信。如果这些相同特征与统计学分数呈反比例关系，我们也可以硬造出解释：身体质量指数低的学生会花很多时间锻炼，就读的高中规模小的学生没有接受过挑战。

由于缺乏以上数据，我只能运用手头的数据——学生姓名。例如，Gary（加里）这个名字顺序包括四个字母G-A-R-Y。我使用数据挖掘软件来探索统计学课程成绩和姓名中字母位置是如何产生联系的。成绩分为几个等级：A为4.0分，B为3.0分，以此类推。班上学生的平均分数为3.03分，比B稍高一点。

对名字而言，数据挖掘算法发现，当名字中的第七个字母为D时，学生的分数提高了0.94分；第二个字母为D时，提高了0.81分；第五个字母为D时，提高了0.79分。相反，当第四个字母为D时，分数下降了0.96分。最糟糕的是，当第八个字母为G时，分数下降了1.85分。

再来看看姓氏，当第四个字母为V、第六个字母为B、第五个字母为C时，分数均提高大约1分，但当第二个字母为M时，分数减少2分。

这简直就是胡扯，尽管数据挖掘程序对此并不知情，因为它不知道分数和名字是什么意思。如果我用身高、体重等信息来做同样的实验，指不定会为最终的“知识发现”编出什么解释。我故意选择毫无意义的特征是为了说明，在事情发生之后，怎样都能找到与企业成功或学生的好成绩相关的模型。找到这些特征是意料之中的事情，但并不意味着它们就是迈向成功的关键。

攻击性和吸引力

12名年过40岁的已婚女性观看20名20岁出头的已婚男性玩街头篮球的视频。受试女性给视频里每一名男性的攻击性和外形吸引力评分，然后以平均分进行排名，1为最高，20为最低。

如图6.2的排名散点图所示，攻击性排名越高的男性，吸引力排名越低，其相关系数竟然为-0.93。尽管人们普遍认为，女生会被“坏男

生”吸引，但这项研究说明，年至40岁的女性并不会被攻击性强的男性吸引，可能是因为她们已经懂得危险和鲁莽的行为会导致严重后果。她们更倾心于稳重和自制力强的男性。

我现在就能想象到网上的那些故事情节。男同胞们，注意行为举止，收好那些睾酮丸。

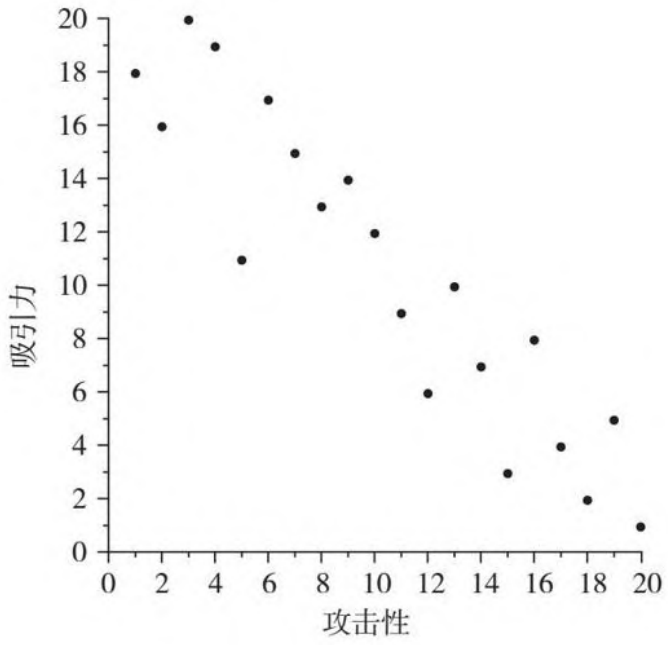


图6.2 攻击性和吸引力关系的散点图

你被说服了吗？这些结果、统计学分析和图表都由计算机得出，所以肯定都是正确无误的。不过，数学计算正确并不一定代表结论是有说服力的。

如果我告诉你，图6.2所示的攻击性和吸引力的关系，是通过搜遍包括这20名男性的1 000个特征的庞大数据集而得出的，那会怎么样？大多数特点非常不相关，但还是存在一些令人惊叹的统计学关系——最让人惊叹的是攻击性和吸引力之间的负相关系数。

确实，进行这项研究的教授并不意在寻找这一特别的关系，但很明显它是存在的。确实，如果结果显示为正相关，该教授会得到不同的结论，即女人会被具有攻击性的男人吸引。但最重要的一点是，这名教授找到了一种强有力的实证关系，这是不容置疑的。

现在，愤世嫉俗的人可能会表示，那名教授是在1 000个变量中找相关系数。即使数据只是无意义的噪声，仅凭运气，他也会找到一些统计学意义上的强相关系数。

那名教授可能会火冒三丈，解释说这些都不是随机数据，它们反映的都是真实特征，揭示的是攻击性和吸引力之间的真实关系，没有人能质疑这种关系的说服力。那名教授一定发现了重要的真相，因为其统计学关系如此显而易见。这就是数据挖掘和知识发现之间的真相。

好了，该说实话了：没有什么视频，也没有排名。我就是那名教授，是我用计算机随机数字生成器从1~20中随机挑选了1 000次，然后运用数据挖掘软件搜遍这些数据，寻找“意料之外”的关系。图6.2显示的是随机排在第462位和第594位的数据，在我发现它们后，我将其标记为攻击性和吸引力。这也正切中要点，搜遍大量数据库来寻找统计学关联，就会找到让人印象深刻的关联，即使数据只是随机数字而已。发现这类关联证明不了什么，除了这些数据都被搜了个底朝天的事实。

数据挖掘也许不是知识发现，而是噪声发现。

达特茅斯三文鱼研究

标准的神经科学实验是这样的：一名志愿者位于磁共振成像（MRI）机里，观看不同图像并回答有关图像的问题。fMRI可测量含氧与脱氧血流经过大脑时的磁信号阻断情况。测试结束后，研究人员要观看13万张体素图像［即3D（三维）数据］，查出大脑哪些部位受到了图像和问题的刺激。

fMRI测试方法有噪声源，包括来自周围环境中的磁信号和大脑不同部位脂肪组织密度的变化。体素有时会漏记大脑活动，有时又显示出没出现过的大脑活动。

达特茅斯大学研究生克雷格·贝内特进行了一项不寻常的实验，他向三文鱼展示图片并提出问题，然后通过MRI机研究三文鱼的脑部活动。接下来，精巧复杂的统计学分析发现了明显的模型。

这项实验最有趣的地方并非研究的对象三文鱼，而是这是一条死鱼。没错，贝内特在当地市场买了一条死的三文鱼，放到了MRI机里，向它播放图片并提出问题。在大量体素图像中，一些随机出现的噪声也被记录在内，被当成了三文鱼对图片和问题做出的反应。除了没有记录这条三文鱼已经死了。

这项对死三文鱼的研究比大多数fMRI研究引起了更多人的关注，甚至还获得了“搞笑诺贝尔奖”。

这项研究还是有些人对“大数据”进行数据挖掘来寻找模型的绝佳类比，只不过那些“大数据”包含的数据要多得多，会找出荒唐至极的关联性。

骗子，骗子

满怀希望的人们搜遍数据寻找称霸股市和喜中彩票的方法，随之而来的是可笑牵强的说法，如“超级碗指标”或让名为Mary（玛丽）的朋友帮自己买彩票。千万别被这些鬼话忽悠。

假如足够努力，即便在随机生成的数据中，我们也总能找到模型。无论找到的模型有多令人震惊，我们还是需要合理的理论来解释它，否则，我们得到的就不过是巧合事件罢了。更多例子请参阅我的另一本书《简单统计学：如何轻松识破一本正经的胡说八道》，例如：亚洲裔美国人容易在每月4号发作心脏病、通过庆祝重要事件能推迟自己的死亡等说法。

作为人类，我们能够识别不合情理的说法，谨记那个死三文鱼研究的教训，但是计算机做不到。

没错，人类研究员过度进行数据挖掘，已经造成了一场复制危机。严谨的研究人员如今正在努力寻求有效方式来减少数据挖掘的使用。有人提议，期刊不应坚持采用具有统计学意义的结果，因为这引诱研究人员跳入了数据挖掘的陷阱。

挖掘大数据的人工智能算法是往错误方向迈进的一大步，这错误的一步会让复制危机加剧，变成复制灾难。计算机永远不会真正理解的根本事实是，合理的模型比仅仅同数据吻合的模型更有用。人工智能算法只懂得怎么拷问数据。

第7章

无所不包的“厨房水槽法”

20世纪80年代，我曾与一名经济学教授交谈，他根据图7.1所示的简单相关系数给一家大银行进行预测。如果想要预测消费性支出，他便制作收入和支出的散点图，然后用透明尺子画出一条似乎与数据一致的线。根据他的预测，若收入增加，支出也会增加。

这名教授的散点图的问题在于，世界并非如此简单。收入会影响支出，财富状况也会。如果教授恰巧利用收入增加（支出增加）、股市暴跌（支出减少）时期的数据来画散点图，而财富的影响力又大于收入的影响力从而导致支出减少（如图7.2所示）那又会怎么样呢？据此，教授的收入和支出散点图将预测：收入增加，则支出减少。之后，当他试图预测在某一收入和财富都增加的时期支出的变化趋势时，他会预测到支出呈下降趋势，这简直错得离谱。

此时需要运用多元回归分析。

多元回归模型含有多个解释变量。例如，消费性支出模型可表示为：

$$C = a + bY + cW$$

C代表消费性支出，Y代表家庭收入，W代表财富状况。

以上解释变量的罗列顺序并不重要。重要的是将哪些变量纳入该模型，哪些排除在外。回归分析的技巧重点在于选择重要的解释变量，忽略不重要的解释变量。

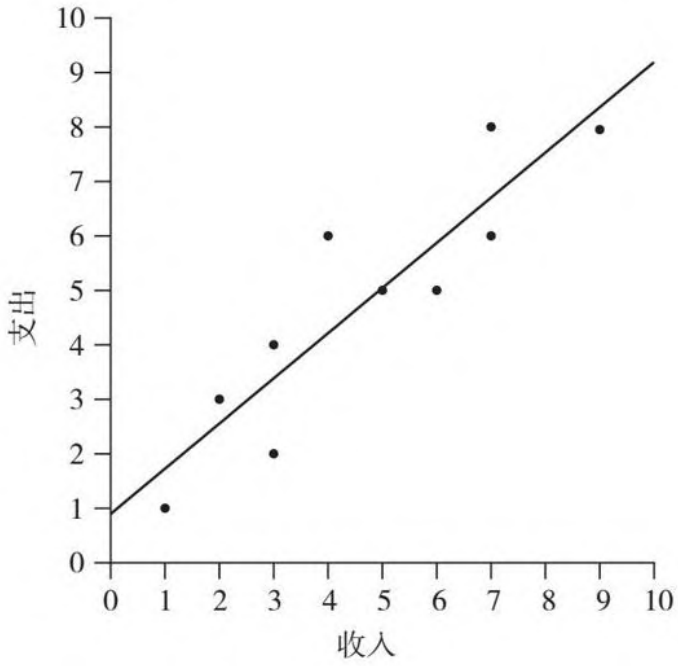


图7.1 收入和支出的正相关系散点图

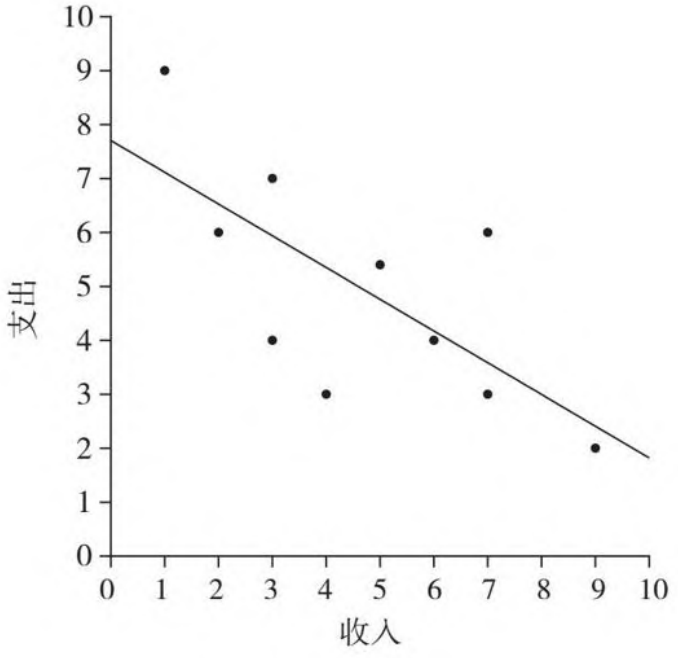


图7.2 收入和支出的负相关系散点图

系数b计算财富状况保持不变时收入增加对支出的影响，而系数c计算收入保持不变时财富增加对支出的影响。推算这些系数的数学过程非常复杂，但是原理很简单：为用来推算模型的数据选择能最佳地预测消费性支出的推算。

我们已经在第4章了解到，在比较支出、收入和财富这些都会随时间推移而增加的变量时会出现“假性相关系数”。为确保不被假性相关系数误导，我要看的是去除通胀因素后的支出、收入和财富的年度百分比变化。

我使用统计学软件来计算美国年度数据的回归线：

$$C = 0.62 + 0.73Y + 0.09W$$

财富保持不变，收入每增加1%，支出预计会增加0.73%。收入保持不变，财富每增加1%，支出预计会增加0.09%。图7.3为实际支出的百分比与预测支出的百分比的变化对比图，相关系数竟惊人地达到0.82。

财富的系数看似很小，但是财富的变化通常很大。有好几年，财富增加或下降的幅度超过10%，根据我们的模型预测，消费性支出会下降0.9%，这就形成了经济扩张和衰退之别。

多元回归模型的效力极大，远比简单的相关系数大得多，因为它将多个解释变量考虑在内。这就是为什么多元回归模型是较重要的统计学工具之一。

然而，多元回归模型用于数据挖掘时也非常容易出现滥用的情况。

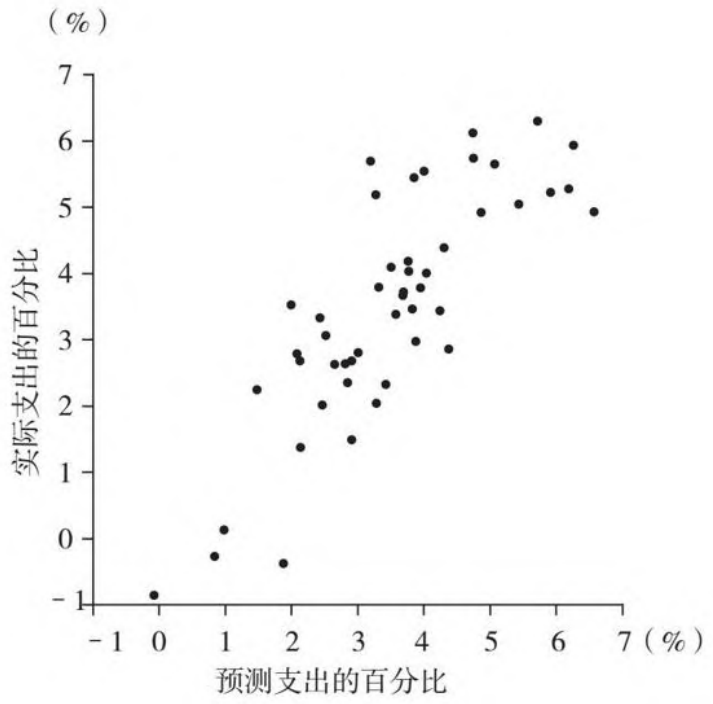


图7.3 美国家庭的预测支出与实际支出的百分比变化

预测总统大选

在统计学课上，我要求学生列出自认为可以决定总统大选结果的因素。他们提到了经济、候选人个性、国家是否处于战时状态等。我将他们的想法写在白板上，然后展示我的模型。

100多年来，美国总统大选通常都是民主党和共和党两党的总统候选人之争。执政党的总统候选人要么是总统本人，要么是总统所在党派的提名候选人。2012年，竞选第二任期的贝拉克·奥巴马就是执政党的总统候选人。2016年，取代已连任两届的奥巴马而参加竞选的希拉里·克林顿，就是总统所在党派的提名候选人。

执政党具备很多优势，包括有更便捷的渠道接触媒体、筹集资金。执政党可以吸引渴望稳定和对现状满意的民众。另外，对经济、战争等问题感到不满的选民可能会投票支持挑战者，即替换候选人。据估计，总统候选人与挑战候选人的优势比为4：6，尽管最终结果显然取决于具体候选人和历史环境。

如果我告诉学生，我只根据执政党的提名候选人是否为总统就能预测执政党在两党投票中的得票率，他们肯定认为我是在开玩笑。他们完全有理由这么认为。我们都知道在任总统什么时候表现好（罗纳德·里根得票率为59%），什么时候表现不好（吉米·卡特得票率为44%）。

但是，如果我还将候选人是否曾任州长和参议员等因素都考虑在内会怎么样？我向学生展示了以下多元回归模型，该模型是我利用过去10次总统大选（1980—2016年）的结果推算而来的：

$$i\% = 78.31 - 7.35iP - 13.07iV + 7.93cV - 27.20iS + 14.75cS - 34.46iG + 8.20cG - 19.54iR + 3.49cR$$

这些变量分别代表：

i% = 执政党候选人获得的主要党派投票百分比

iP = 执政党候选人是总统时等于1，否则等于0

iV = 执政党候选人担任过美国副总统时等于1，否则等于0

cV = 挑战者党候选人担任过美国副总统时等于1，否则等于0

iS = 执政党候选人担任过美国参议员时等于1，否则等于0

cS = 挑战者党候选人担任过美国参议员时等于1，否则等于0

iG = 执政党候选人担任过美国州长时等于1，否则等于0

cG = 挑战者党候选人担任过美国州长时等于1，否则等于0

iR = 执政党候选人担任过美国众议员时等于1，否则等于0

cR = 挑战者党候选人担任过美国众议员时等于1，否则等于0

我并不考虑经济、候选人个性以及我的学生认为重要的其他因素。我选择一些依稀相关的因素，并得到了准确无误的关联，因为我的等式可以完美地预测这10次总统大选的所有结果。例如，我的模型对希拉里·克林顿在2016年两党投票中的预测结果为51.11%，正等于她的实际得票率。

当我的学生看到模型与数据完全匹配时，他们不禁认为我已经找到了预测总统大选的神器。我的模型并不包括他们认为重要的任何因素，但是它看上去很合理，因为它使用了与总统候选人背景相关的解释变量。最重要的是，我的模型与数据非常吻合，因此它肯定正确，是学生自己犯错了。

然后，我又给他们展示第二个完全符合1980—2016年10次总统大选数据的模型：

$$i\% = 84.79 - 1.62T1 - 0.30T2 - 0.04T3 - 0.54T4 + 2.94T5 - 0.39T6 + 0.60T7 + 0.14T8 - 1.05T9$$

这9个解释变量均为大选之日的最高气温，分别来自9座城市，并且这些城市所在的大州只有极少数选票：

T1 = 蒙大拿州博兹曼市的最高气温

T2 = 内布拉斯加州布罗肯鲍市的最高气温

T3 = 佛蒙特州伯灵顿市的最高气温

T4 = 缅因州卡里布市的最高气温

T5 = 怀俄明州科迪市的最高气温

T6 = 特拉华州多佛市的最高气温

T7 = 西弗吉尼亚州艾尔肯斯市的最高气温

T8 = 北达科他州法戈市的最高气温

T9 = 爱达荷州波卡特洛市的最高气温

之所以选择这些城市，是因为我喜欢它们的名字，也能找到它们早至1940年的每日天气数据。

现在，我的学生都感到困惑了，同时还有很多人心存疑虑。这些都是我一手捏造的吗？博兹曼市或布罗肯鲍市的气温怎么会对总统大选造成实质影响呢？为什么执政党候选人获得的选票与博兹曼市的温暖天气存在负相关系数，而与科迪市的温暖天气存在正相关系数？完全没有符合逻辑的解释，但这个模型却与数据非常吻合。

总统大选可能会受到天气影响。这可能是搜遍数据才能发现的、意料之外的关系。我可能偶然做到了知识发现，证明了数据挖掘的威力。你是否也不禁信以为真了？

所以，我决定让模型看似更加荒唐。我推算出第三个与1980—2016年10次总统大选数据完全相符的模型：

$$i\% = 33.73 - 0.01R1 + 0.26R2 + 0.21R3 + 0.20R4 - 0.01R5 + 0.19R6 + 0.01R7 - 0.33R8 - 0.18R9$$

这一次，解释变量的确都是随机得来的。我使用了计算机软件随机生成两位数的数字，这些数字与现实世界没有一点关系，与总统选举年期间美国发生的事情更没有关系。但是，该模型与数据的匹配度还是非常高。

尽管我的学生满腹疑团，但这一切确实并非我无中生有。不过，我确实有个秘诀。

独家秘诀

假设我想解释为什么2016年底的股价会比2015年底的高10%，并且我声称这一切都是因为天气。具体来说，是因为位于加利福尼亚州中央山谷的波特维尔小镇的天气，我父亲就在那里长大。你会认为我疯了，而如果我真的这么想，那么你说得也没错。但是，你先听我把话说完。

图7.4所示的数据散点图为2015年和2016年最后一天的标准普尔500指数以及波特维尔的最低气温。图中显示两者存在绝对完美的相关关系。这两个变量之间的相关性为1。股价完全可以根据我父亲家乡的气温变化来预测。谁能想到呢？

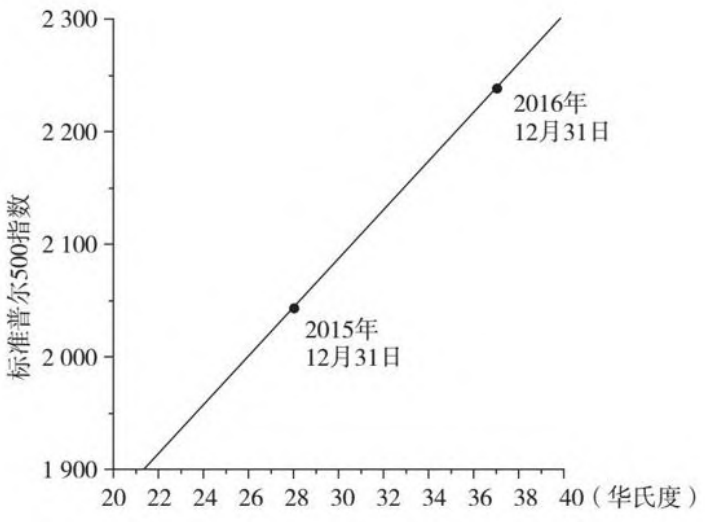


图7.4 用加利福尼亚州波特维尔小镇的最低气温预测股价

秘诀（这当然少不了）就是，散点图中的两点之间总会存在完美的线性关系。我还可以选择1974年和1997年出生的、名叫克莱尔的新生儿数量，或是圣安东尼奥马刺篮球队在2012年和2015年的获胜场数。这些数据和标准普尔500指数之间同样会存在完美的线性关系，因为连接两点总会出现一条直线。

然而，这种拟合关系却毫无用处。任何试图通过波特维尔的气温来预测股价的人，都会以失败而告终。

我在二维图表中使用两个数据点，说明这种荒唐的想法适用于采用更多数据的、更复杂的模型。图7.4使用一个解释变量（波特维尔的气温）完全匹配了两种观察结果。如果有3种观察结果，两种解释变量也完全匹配。即便有10种观察结果，9种解释变量也是一样的情况。

这就是我得出上述三个预测10次总统大选的模型的方法，一个比一个离谱，秘诀在于使用9个解释变量，仅此而已。这9个解释变量也没有什么特别之处，任何9个都可以。重点在于，我使用这9个变量的目的是预测10次大选。

这就是所谓的“过拟合”（overfitting）数据的极端例子。在任何实证模型中，我都能通过增加越来越多的解释变量，来提高模型的解释力——在极端的例子中，可以将其提高到精确吻合的程度。变量是否合理几乎无足轻重。

这种建模方法也就是常说的“厨房水槽法”，即一股脑把所有解释变量统统塞进模型中。无法避免的问题是，即使模型与原始数据吻合度很高，使用新数据来预测也丝毫不起作用。波特维尔的天气不能准确预测股价，除非“瞎猫碰到死耗子”。我做的包含9个变量的总统大选模型也无法准确预测其他总统大选结果，除非歪打正着。

回看1980年之前的10次总统大选，就能看清我做的总统大选模型的缺点。如图7.5所示，运用时任总统情况和挑战者的数据得出的模型1与1980—2016年间的10次总统大选结果完全吻合，但是与1980年之前的10次总统大选的结果却截然不同。该模型预测理查德·尼克松会在1972年的大选中惨败，普选得票率仅为29%。可实际上他以绝对优势取胜，普选得票率高达62%。尼克松拿下了除马萨诸塞州之外的各个大州，这个“海湾之州”的一些民众还在保险杠上贴了贴纸：“别怪我们。”

模型1对1956年和1964年总统大选的预测结果更是一塌糊涂，竟然预测德怀特·艾森豪威尔在1956年的得票率高达几乎不可能的79%（实

际结果为58%），还预测林登·约翰逊在1964年的得票率低至几乎不可能的26%（实际结果为61%）。

我过度拟合了最近那10年的总统大选数据，随后尝试预测早前的大选结果（以失败而告终）。我同样也可以通过过度拟合早前10年的总统大选结果来推算系数，然后再用这个模型来预测最近10年的总统大选结果。如图7.6所示，修订版模型与1940—1976年之间的10次总统竞选结果完全吻合，但对最近10年的总统大选的预测结果却糟糕透顶。

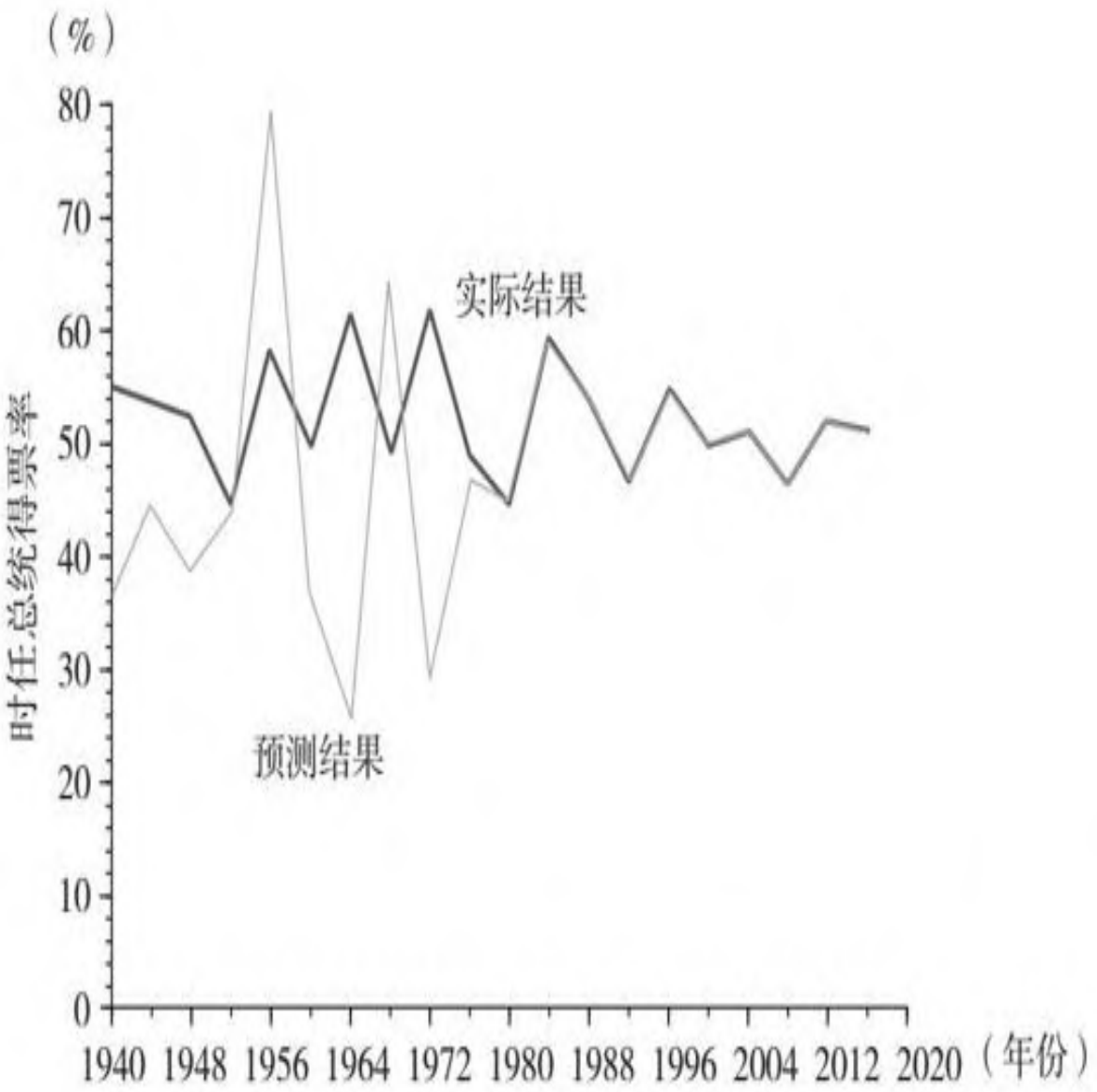


图7.5 使用1980—2016年过拟合数据预测总统大选

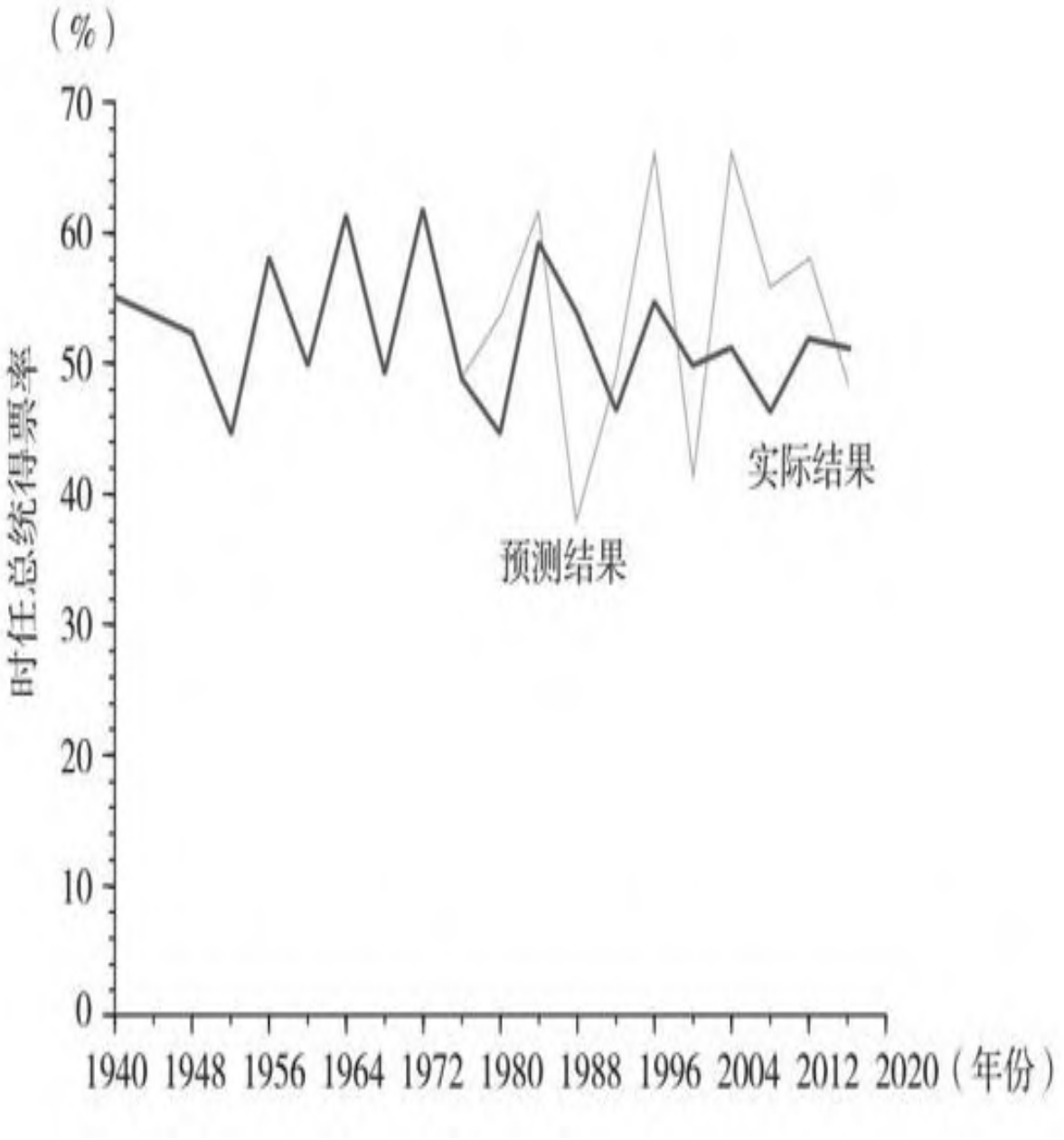


图7.6 使用1940—1976年过拟合数据预测总统大选

模型2和模型3的情况与此如出一辙。如图7.7所示，气温模型与用来推算该模型的数据完全吻合，但是对其他年份大选结果的预测却不尽如人意。该模型预测富兰克林·罗斯福在1940年大选中的得票率为-11%（没错，就是负数），而他的实际得票率为55%。

坦白说，使用9个解释变量来预测10次总统大选是个极端例子。我这么做是想说明一个普遍原理，那就是即使在回归模型中增加毫无意义的解释变量也会提高模型的吻合度。

在预测总统大选的天气模型中，我们不需要添加全部9个解释变量才能达到很高的吻合度，即便只有5个解释变量（即伯灵顿市、科迪市、多佛市、艾尔肯斯市和波卡特洛市的天气），天气模型预测结果与实际得票率之间的相关系数也会达到0.94：

$$i\% = 72.75 - 0.38T3 + 0.59T5 + 0.40T6 - 0.38T7 - 0.65T9$$

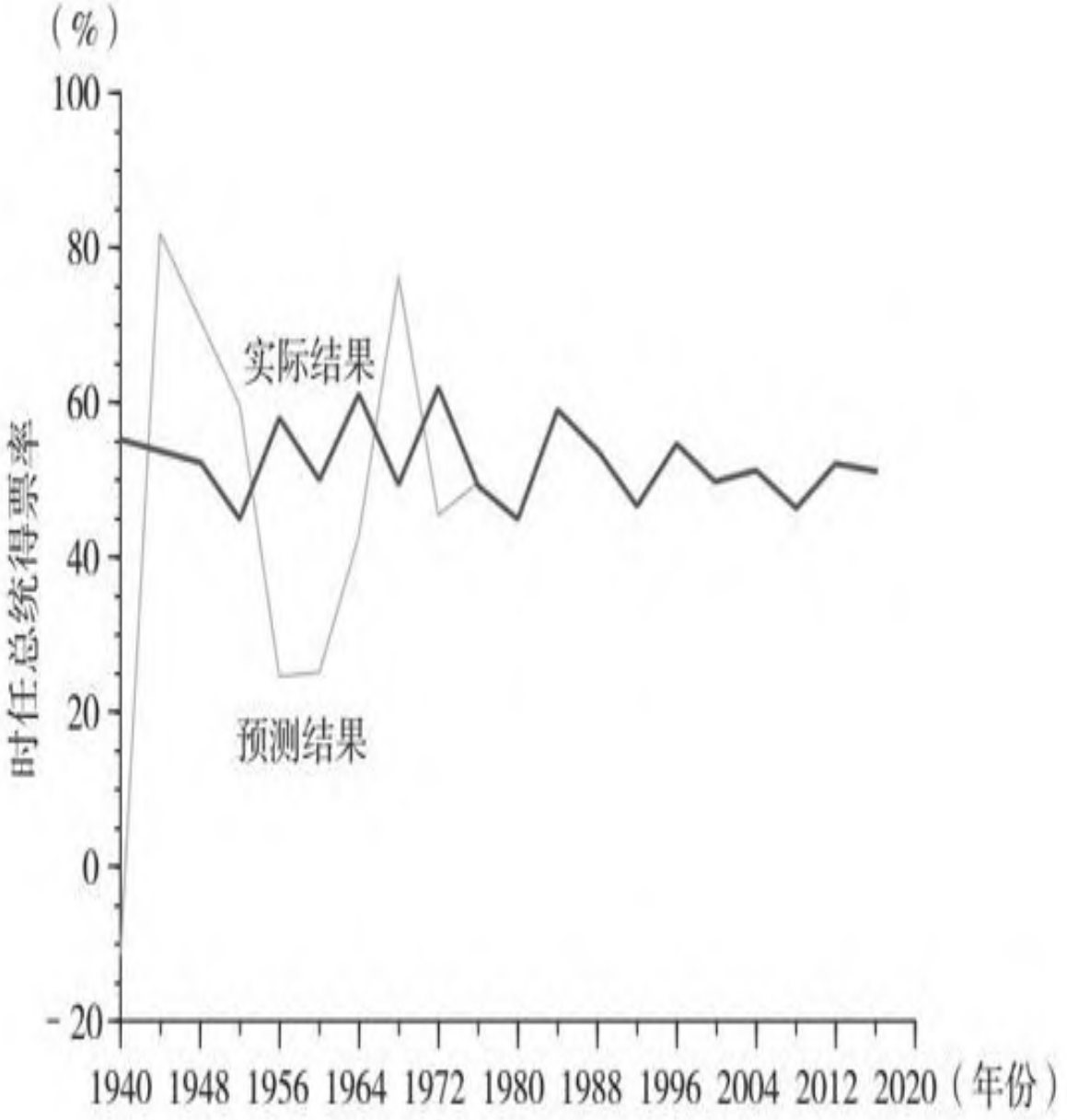


图7.7 使用1980—2016年的天气数据来预测总统大选

如图7.8所示，显然，这个包含5个解释变量的天气模型与1980—2016年的大选数据高度吻合，而与1940—1976年的数据大相径庭：

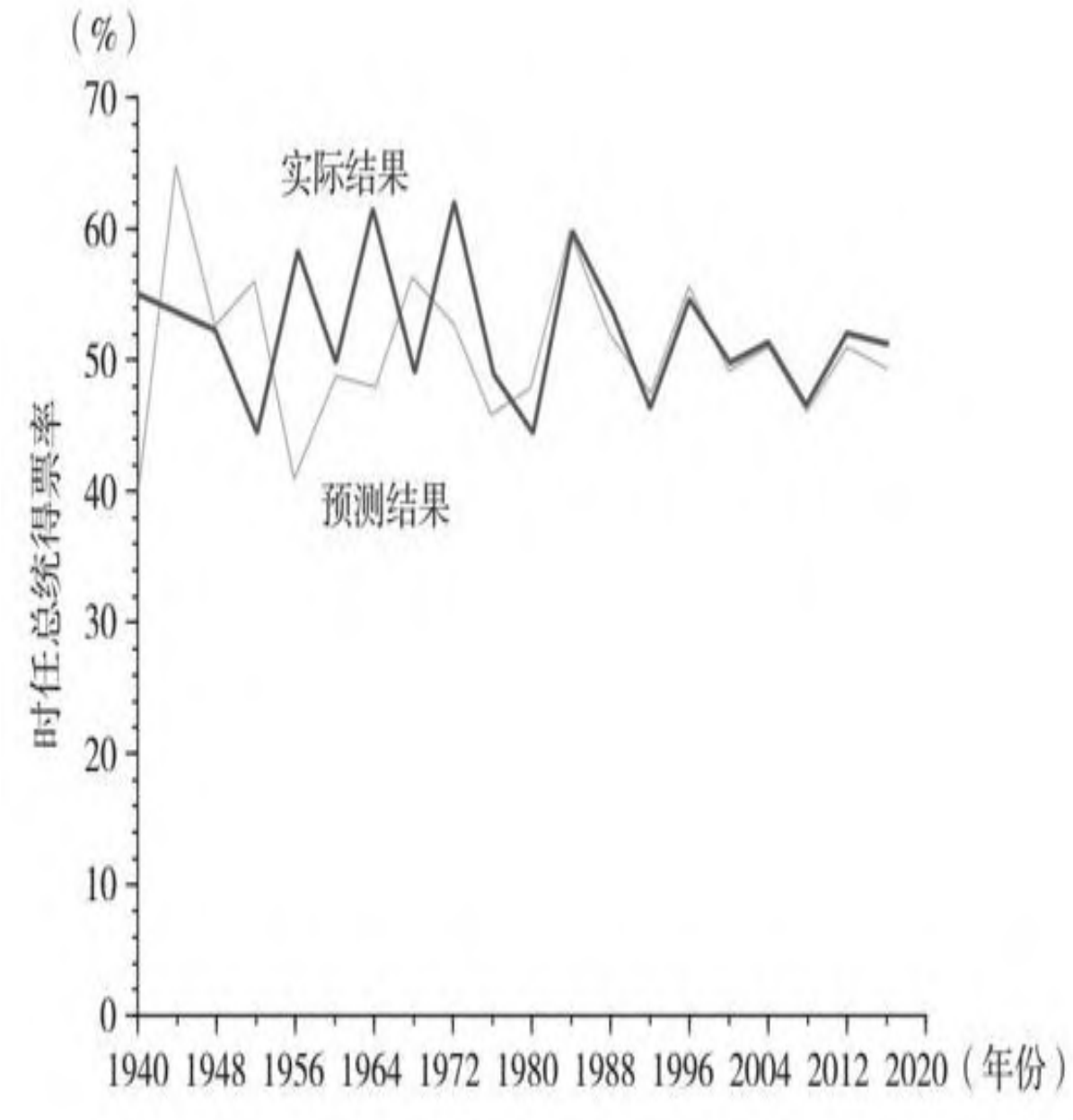


图7.8 使用5座城市的气温来预测总统大选

我们还能以更少的解释变量达到很高的吻合度。包括4座城市（伯灵顿市、科迪市、艾尔肯斯市和波卡特洛市）的天气数据的模型与1989—2016年大选结果的相关系数为0.86，而包括3座城市（科迪市、艾尔肯斯市和波卡特洛市）的数据时相关系数为0.79。

如果将该模型与1940—1976年的数据匹配，情况也是一样。包括4座城市（布罗肯鲍市、多佛市、艾尔肯斯市和法戈市）的天气数据的模型与1940—1976年大选结果的相关系数为0.89；而包括3座城市（布罗肯鲍市、艾尔肯斯市和法戈市）的天气数据时相关系数为0.86。

我选出这些城市的依据是什么呢？我有25座城市每日最高和最低气温的数据，使用数据挖掘软件将这50个变量的所有可能组合统统考虑在内，然后识别出与总统大选结果吻合度最高的组合。

结果显示，1980—2016年吻合度最高的城市与1940—1976年吻合度最高的城市截然不同，因为该模型没有理论基础。这些我用来寻找假性相关系数的数据本质上是随机的。任何差强人意的数据挖掘程序都能得到同样毫无意义的结果，而且还根本不知道这些都是无稽之谈。

如果解释变量减少，随机变量模式也还是能与数据高度吻合，如解释变量减至5个，相关系数为0.97；减至4个，相关系数为0.95；减至3个，相关系数为0.89。一切都与天气模型非常相似，如果某一年的数据没有拿来推算该模型，那么这个模型对该年度大选结果的预测毫无用处。

由此得出的结论不可忽视。数据挖掘能轻易发现包括多个解释变量的模型，即便解释变量与所要预测的变量毫无关系也能与数据达到惊人的吻合度。不足是，数据挖掘软件不能评估模型是否合理，因为对计算机软件来说，数字只是数字而已。

我们如何分辨所发现的模型是真实还是虚假的呢？只要懂得利用人类对变量的认识，就能判断所发现的模型是否具有逻辑基础。

我一直在强调这一点，是因为我与很多聪明的相关人士都交谈过，他们虽是出于好意，但始终不能完全理解找到偶然性的模型和关联性是多么轻而易举的事情，其中还包括大多数和我交流过的数据挖掘者。很多人都模糊意识到可能存在假性相关系数，但尽管如此，他

们还是相信模型和关联性的统计学证据足以证明它们就是真实存在的。

2017年，《华尔街日报》的首席经济评论员格雷格·伊普采访了一家为企业开发人工智能应用程序的公司的合伙创始人。伊普复述了此人的论点：

> 如果在大学学过统计学，你就会知道如何利用输入来预测输出，例如，基于身体指数、胆固醇和吸烟状况来预测死亡率。可以通过添加或取消输入来提高模型的“吻合度”。
>
> 机器学习使用强大的算法和计算机来分析更多的输入。例如，数码图片中的数百万像素，不仅有数字，还有图像和声音。它从变量组合中衍生出更多变量，直至能最准确地回答问题（如“这是一张狗的图片吗”）或者能最圆满地完成任务（如“说服观看者点击本链接”）。

此言差矣！学习统计学的学生在大学里应该学到的是：仅为了提高适合度就添加或取消输入有百害而无一利。机器学习也是如此。搜遍数字、图像和声音寻求最佳匹配，这是盲目的数据挖掘，考虑的输入越多，所选变量的虚假度就可能越高。

数据挖掘的根本问题在于：它非常擅长找到匹配数据的模型，但对判断模型是否荒唐可笑完全束手无策。统计学相关系数无法替代专业人士的意见。

为现实世界建模的最佳方法是，从具有吸引力的理论学说开始（如“经济状况会影响总统大选”），然后验证模型。合理的模型可对其他数据做出有用的预测，而不是预测用来推算模型的数据。数据挖掘则是反其道而行之，它没有基础理论，因此无法区分合理与荒谬的模型。这就是为什么这些模型对于全新数据的预测结果并不可靠。

非线性模型

除了通过筛选全部解释变量，数据挖掘算法还能通过大量非线性模型来过度拟合数据。

图7.9所示的简单散点图使用了假设数据。图中的三个观察结果都没有在直线（线性模型）上，但还是可以看出其大致走向，如果X和Y之间确实存在因果关系，则可能有助于预测Y值。

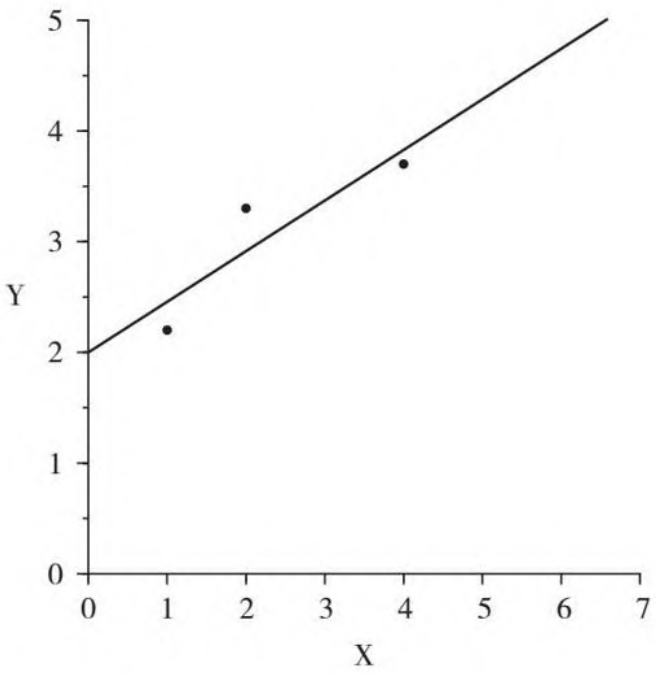


图7.9 线性模型与三种观察结果不吻合

图7.10所示的非线性模型与这三个观察结果完全吻合。可以因此说图7.10的非线性模型是图7.9线性模型的改进版吗？不一定，数据挖掘算法没有合理的方式进行判断。

图7.9的模型显示，X值上升，Y值也上升，增幅保持不变。图7.10的模型显示，X值上升，Y值上升幅度越来越小直至变为负数，X值大于

7时，Y值为负数。

要用与模型不吻合的X值来预测Y值，哪个模型更有效呢？看情况。如果X表示家庭收入，Y表示支出，则如图7.9所示，收入增加时，支出也以大致相同的幅度增加，这种说法很合理。但是如图7.10所示，收入增加到某个点时导致支出减少，直至降为负数，这就说不过去了。

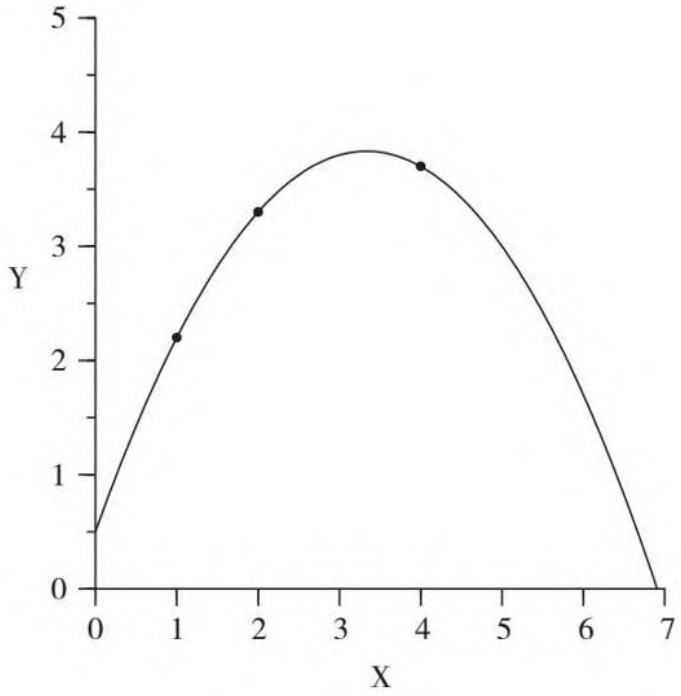


图7.10 非线性模型与三种观察结果完全吻合

此外，假设X表示施给土豆苗的氮素数量，Y表示生长状况。这种情况下，如图7.9所示，即每多施加一点氮素，生长就会快一些，这不合常理。对比之下图7.10更合理，随着氮素的用量不断增加，其对土豆苗生长的促进作用也会不断减弱。在某一点上，额外的氮素会有碍土豆苗的生长，土豆苗甚至会因为氮素过多而死亡。

数据挖掘算法如何能够决定是图7.9中的线性模型还是图7.10中的非线性模型可以更好地表示建模的事实情况呢？当然不可以只通过看哪个模型与数据更加匹配来决定！我们只能通过专家（即人类）的建议来评估哪个模型更符合现实，才能在这些或其他模型中做出选择。

图7.11展示了更极端的例子。如果存在符合逻辑的解释，这个解释似乎能将直线与所有数据完全匹配，将直线附近的点解释为模型之外其他因素造成的不可避免的波动。除非发生剧烈变动，否则利用线性模型应该可以做出合理精确的预测。

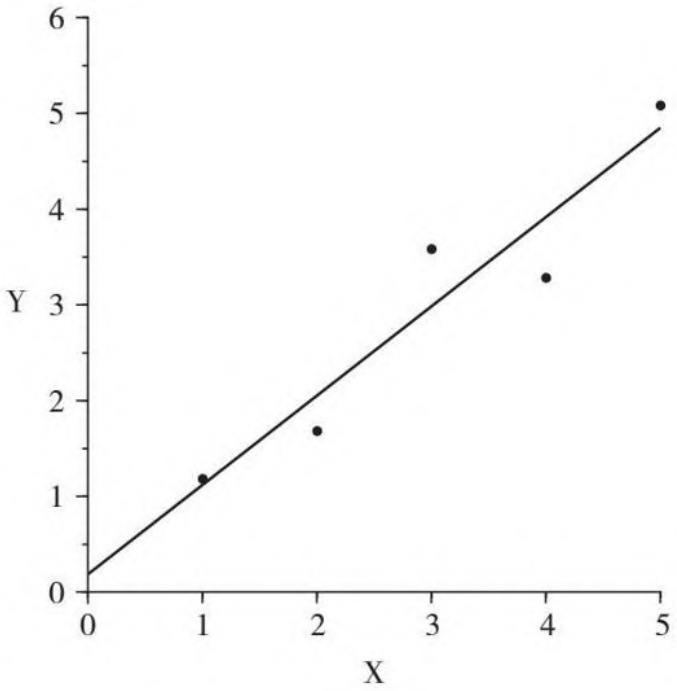


图7.11 合理的线性模型

图7.12显示了数据挖掘算法为了完全匹配数据而选择过度复杂的非线性模型后出现的混乱趋势。尽管与原始数据完全吻合，但只要输入新的X值，该非线性模型的预测结果就肯定会差之千里，甚至会令人匪夷所思。

问题自始至终都在于，数据挖掘算法寻找模型（这也是它非常擅长的事情），但是没有办法评估自己找到的模型。spending（花费）、income（收入）和wealth（财富）等词语都只是字母组合而已，正如奈杰尔·理查兹用自己不懂的语言玩拼字游戏那样。计算机算法不能分辨模型中应该包括哪些解释变量，也说不出线性和非线性模型哪个更合理。这些都需要人类智慧来做决定。

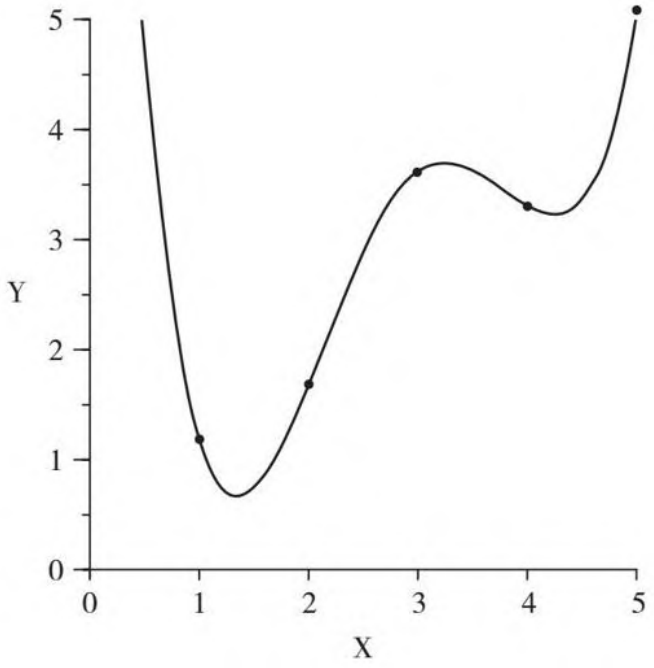


图7.12 不合理的非线性模型

第8章

新瓶装旧酒

时尚界的流行风潮周而复始，基本遵照“20年一轮回”的规律。20世纪90年代经常出现在健身房和大街上的运动装，时隔20年之后又卷土重来。

英国经济学家丹尼斯·罗伯逊曾从理性的角度写过：“我的有些学生是在运动圈里长大的，正如我时常对他们说的那样，如今，高雅的见解就像一只被追捕的野兔，如果你坚守一处或其附近，它肯定会再度回到你身边。”

同样，今天的数据挖掘者再度发现了几种曾经风靡一时的统计学工具。这些工具重获新生是因为它们的数学原理复杂中带着优美，很多数据挖掘者轻易被这种数学之美诱惑，很少有人思考深层的假定和结论是否合理。

逐步回归法

思考一下多元回归模型的数据挖掘。搜遍庞大的数据库，寻找呈现最佳拟合的解释变量组合，这种做法令人望而却步。从100个变量中选择的话，10个解释变量的可能组合共有17万亿多种；从1 000个变量中选择的话，10个解释变量的可能组合共有10^{24}种；从100万个变量中选择的话，10个解释变量的可能组合的数量非常庞大，写下来的话，在1后面得写54个0。

逐步回归法（stepwise regression）出现时，计算机的速度比现在慢得多，但它已经成为很流行的数据挖掘工具。这是因为比起彻底搜寻所有可能的解释变量组合，这种方法对计算的要求比较低，但还有望得出接近彻底搜寻结果的合理近似值。

“逐步”体现在计算要依次经过很多步骤，逐个考虑潜在的解释变量，其中主要有三个分步操作。

首先，按照向前选择变量原则，选取与预测目标变量具有最高相关系数的解释变量。其次，引入第二个使拟合度最大化的变量。最后，再引入第三个使拟合度最佳的变量，如此循环往复，直至拟合度的增加超过预先设定的最小值。所有可能的解释变量都依照逐步回归并删除的原则，然后逐个剔除变量，而且每次都只剔除对拟合度影响最小的变量。

另一个常见的分步操作是双向的，结合了向前选择和向后删除。就向前选择而言，数据挖掘最开始会选择与预测目标变量相关系数最高的解释变量，然后逐个引入变量。巧妙之处在于，每一步中，该步骤都考虑到剔除原先引入变量的统计学后果。因此，变量可能在第二步被引入，在第五步被剔除，然后又在第九步被引入。

上述所有的分步操作有何共同之处呢？它们都是全自动的规则，只需考虑统计学相关系数，不必关心引入或排除那些变量是否合理。这是快速版的数据挖掘。

有人建议基于系数的“大小”引入或减少回归模型中的变量，无论这是否有意义。这不是我捏造出来的。软件程序怎么能比较身为美国参议员的执政党候选人和内布拉斯加州布罗肯鲍市的气温之间的相关关系呢？此外，变量应包含在模型内是因为从逻辑上来说它们归属于模型，而不是基于其推算系数的大小。

我用不同模型做了几项实验，这些模型含有100个值，其中有5个确实能决定预测结果的真实变量和若干由随机数字生成器生成的虚假变量，不过，这些虚假变量可能碰巧与预测变量相关联。我在应用逐步回归法时发现，如果考虑100个变量，其中5个为真，95个为假，由分步操作挑选出的变量则更有可能为虚假变量。我把考虑变量数目增至200个，再到250个时，引入变量为真实变量的可能性分别降至15%和

7%。大多数最终被引入分步模型中的变量都是因为假性相关关系，这种假性相关关系毫无用处，还可能有碍于采用全新数据进行的预测。

逐步回归法意在帮助研究人员处理大数据，但讽刺的是，数据越大，逐步回归的误导性就可能越大。大数据是问题所在，但逐步回归法却不是解决办法。

岭回归法

2013年，我在一场演讲中听到“岭回归法”（ridgeregression），又被称为“现代回归法”（modern regression）。岭回归法于20世纪70年代得以发展，在80年代名誉扫地，但如今又重新出现在数据挖掘者的工具箱里，还有了荒唐可笑的标签——现代回归法。其变体有吉洪诺夫正则化法（Tikhonov regularization）、菲利普斯-图米（Phillips-Twomey）、线性正规化（linear regularization）、受限线性反演（constrained linear inversion）和其他让人印象深刻的说法。尽管被贴上了这些花哨的新标签，岭回归法也并不是新事物，如今也不会比以往具有更强的说服力，它不过是装在新瓶中的旧酒而已。

看看诺贝尔奖得主米尔顿·弗里德曼创建的消费性支出模型：

$$C = a + bY + cP$$

C代表支出，Y代表现期收入，P代表永久收入。该理论指的是，家庭都不会只追求勉强糊口，仅以当前收入作为支出决定的唯一基础，它们还会考虑自己的平均或“永久”收入。

收入出现短暂性缩减的人们通常都会尽量维持其生活风格。他们必须继续支付房贷或房租，不然就会被扫地出门。他们不想让自己的

孩子停止参加课外活动，不到万不得已，也不想降低自己的生活格调。

同理，收入出现短暂性增加的人们也不会急于提高生活档次、买新房或送孩子上私立学校，因为收入一旦回到正常水平，他们就不得不省吃俭用。

弗里德曼根据这一推理认为，家庭支出不仅受到现期收入的影响，还受到人们自己认为的长期平均收入的影响，也就是他所谓的永久收入。

多元回归法较吸引人的特征之一就是推算某一个变量保持不变时，另一个解释变量的影响。在$C = a + bY + cP$这一支出模型中，系数b预测永久收入保持不变时现期收入增加对支出的影响。系数c预测现期收入保持不变时永久收入增加对支出的影响。

一个潜在的问题是，如果现期收入和永久收入高度相关，我们就无法精确推算它们分别对支出的影响。举个极端的例子，假设现期收入和永久收入每年都增加3 000美元，结果支出增加2 000美元。这是因为现期收入增加了3 000美元，还是因为永久收入增加了3 000美元呢？无从得知。再举一个没那么绝对的例子，现期收入和永久收入高度相关但是并非完全拟合，我们能分别推算其影响力，但是我们的推算结果可能非常不准确。

岭回归法试图通过在解释变量中引入随机噪声来摆脱这一僵局。对于现期收入和永久收入中每一个观察到的值，引入或剔除一个由随机数字生成器生成的数字。

岭回归法虽被描述为复杂的数学方程式，但它完全就等同于在解释变量中引入随机噪声。这一随机噪声让解释变量的相关度变低，制造出推算结果更精确的假象。对于在数据中引入自身造成的误差，有人打趣说是“使用更不精确的数据来获得更精确的推算”。

数据挖掘者怎会相信使用准确度更低的数据能改善推算结果呢？然而他们真的相信。数据挖掘者都过度专注于数学计算，无法停下来思考“模型中的变量不仅是数学标记符号”这一事实。这些真实变量都是用来预测另一个真实变量的。任何思考弗里德曼模型的人肯定都会发现，如果计算的收入不准确，就无法改善支出的预测结果。

还存在第二个大问题。弗里德曼的“消费性支出的永久收入模型”还可以写成储蓄模型，即收入减去支出的差。我们可以说储蓄额增加了300美元，而不说收入增加1 000美元时，支出增加700美元。这些都是同一事情的不同说法。

如果用多元回归法来推算支出和储蓄方程式，会得到同样的结果。但岭回归法就不会如此！其结果取决于推算的是哪个方程式，而且并没有可选择的基础。

岭回归法模糊不清的地方之一就是，很多实践者仔细审查所谓的岭迹（ridge trace），以便决定在数据中引入多少随机噪声。我给一名热衷于使用岭回归法的朋友发送了用于弗里德曼模型四种等价表达式的数据，但不告诉他这些数据都用于等价方程式。结果证实了我的想法——岭回归法的基础不牢。因为他并没有询问所要分析数据的任何信息，它们不过是被操控的数字而已，他也不过像是一台计算机。数字就是数字，谁会知道或在乎这些数字代表什么呢？最后这位朋友推算出模型，然后交还给我四组相互矛盾的岭回归推算结果。

不仅仅是岭回归法，还有数不胜数的统计学步骤为数据挖掘者所用，但实质上都是黑匣子。他们将数字输入黑匣子然后得出结果。他们并不思考这些数据测量什么，或有何使用目的。

数据规约

第7章讨论了多元回归模型中过度拟合数据的问题。存在10个观察结果时（例如10次总统大选），任何9个解释变量都能给出完美的拟合结果，这说明该模型准确无误，但名不副实。

如果解释变量超过9个，就会存在无数方法获得完美拟合状态。一般原则是，当解释变量的数目与观察结果同样多或大于观察结果时，就会存在无数方法获得完美拟合状态。显然，这对大数据来说是个问题，大数据中有数百万个变量等着进行数据挖掘。

遇到处理大量解释变量的情况时，也会出现简单的可行性问题。假设我们有100个解释变量，想要对这些变量进行数据挖掘，找出5个对预测标准普尔500指数拟合度最高的解释变量。我们必须推算超过7500万个回归方程式才能找出最佳拟合。要想找到10个最佳解释变量，就必须推算超过17万亿个多元回归模式。实际上，要想找到超过100个可能的解释变量，我们还要推算更多。

解释变量数不胜数，时间又如此有限，在这种情况下，数据挖掘者会怎么做呢？有两种已经存在了100多年的统计学工具——主成分分析和因子分析——被数据挖掘者重新当作数据规约工具，用以减少解释变量的有效数量。

主成分分析和因子分析类似岭回归法，它们都是基于变量的统计学属性，不关注数字代表的是什么。因为主成分分析和因子分析非常相似，所以我只讨论主成分分析。

假设我们预测某人发生车祸的概率，并考虑以下5个变量：

F = 性别，女性为1，男性为0

Y = 年龄，30岁以下为1，否则为0

O = 年龄，60岁以上为1，否则为0

T = 罚单数

D = 平均每年驾驶英里数

如果这些变量相互关联，那么检查所有变量就有些多余，我们可以使用主成分分析来决定两个加权平均值：

$$C1=0.30F+0.01Y+0.26O+0.05T+0.31D$$

$$C2=0.26F+0.19Y+0.24O+0.12T+0.14D$$

据此，我们可以推算出一个多元回归方程式，利用这两个主成分作为解释变量，代替原来的5个解释变量，从而预测出发生车祸的概率。

数据挖掘者认识到这种变量规约解决了变量过多的问题。他们无须推算包含1 000个解释变量的模式，只要推算包含5个或10个主成分的模型即可。然而，还是存在一个大问题，对减少解释变量最有效的主成分可能不利于预测某些事情，如某人是否会遭遇车祸。

在我们的例子中，*C1*和*C2*的主成分权重都依据这5个原始变量的相关系数，我们从中无法看出性别、两个年龄变量、罚单数和驾驶英里数对遭遇车祸的概率的相关重要性。为了缩减这5个变量，即便赋予性别的权重比罚单数高，罚单数对预测车祸的重要性也可能更高。以减少数据为目的和以使用数据预测为目的是截然不同的。

更糟糕的是，在犹如探险的数据挖掘过程中，有些原始变量可能与车祸发生概率根本毫无关系，但它们还是被视作主成分。要记得，主成分的选择依据是解释变量之间的统计学关系，而根本没有考虑用这些成分来预测什么。

例如，某人的出生月份或喜欢吃的糖果最后有可能被视为用来预测某人是否遭遇车祸的主成分。更普遍的是，如果主成分来自成百上千个变量的数据挖掘，那么这些成分就必定会包括无意义的数据。大数据又一次成为问题所在，而主成分分析也不是解决方法。

还需讨论的是，即便并非不可能，但要解释转化的变量（如*C1*和*C2*）通常也很困难。因此，评估经推算的系数是否合理也成了问题。我们预计车祸概率与*C1*和*C2*呈正相关还是负相关？经推算的系数是否合理？计算机软件不知道，运用主成分分析的人类也不知道，没有人知道。

还有一个问题是，正如岭回归法遇到的一样，解释变量中不重要的变化会影响模型系数的隐含推算。

假设我们不使用：

F = 性别，女性为1，男性为0

而使用以下等价变量：

M = 性别，男性为1，女性为0

这一转换并不重要，即便是一位收到过两次罚单，平均每年驾驶1万英里的47岁女性，也不会影响多元回归推算或所预测的车祸概率。

可是这个简单的转换仍会改变主成分，预测出不同的车祸概率。就像在岭回归法中一样，这是一个致命缺陷。

神经网络算法

神经网络模型最早创建于20世纪40年代，一度失去关注，现又再次时兴——也是新瓶装旧酒。“神经网络”这一标签说的是算法复制了人脑的神经网络，使网络连接具有电兴奋性的细胞，即神经元。我们并不清楚人脑是如何使用神经元来接收、储存和处理信息的，所以我们还不能按照想象用计算机模仿人脑。

神经网络算法是一种给输入分类的统计学步骤（如数字、单词、像素或声音），从而让这些数据能在被整理后输出。尽管数学让人望而生畏，但神经网络还是与其他数据挖掘算法很相似。神经网络本质上是创建输入变量的权重线性组合（很像主成分分析），并运用这些组合（很像多元回归）推算与所预测数据最佳拟合的非线性统计学模型。

推算神经网络权重的过程被宣传为“训练数据”（training data）中的“机器学习”（machine learning），表明神经网络功能犹如人脑。但是，神经网络运算并不像人类的经验学习过程。神经网络更类似回归模型中推算得出的系数，寻找模型的预测结果与被观察值最接近的那个值，不会考虑建模的含义。

神经网络算法在诸如语言翻译和视觉识别方面大有用处，将来它还会有更大的用武之地，但“神经网络能复制人类思维方式”的想法会使人产生误解。思考一下，标准回归模型中含有数据（如图8.1所示的简单散点图），并且回归算法得到一条拟合数据的直线（如图8.2所示）。我们称此数据为“训练数据”，称拟合直线为“机器学习”，但这具有误导性，甚至是虚夸，因为拟合数据的直线与人类使用大脑训练或学习的路径简直不可相提并论，就连犬类、海豚和其他动物接受训练进行学习的路径都比不上。

这就是使用神经网络。算法中拟合数据的直线都只是拟合了使用预设规则的数据，将其称为训练和学习似是而非。该算法并不知道自己在操纵什么，不理解其结果，也无从知道这些直线到底具有意义还只是碰巧而已。

你可能也猜到了，图8.1和图8.2中的两个变量又是我使用计算机随机数字生成器生成的。每个变量的每个值都由随机游走决定，变量之间相互独立。图中明显的统计学相关系数只是偶然事件，根本毫无意义。当然，机器学习算法也无从知晓这一点。数字不过是数字而已。

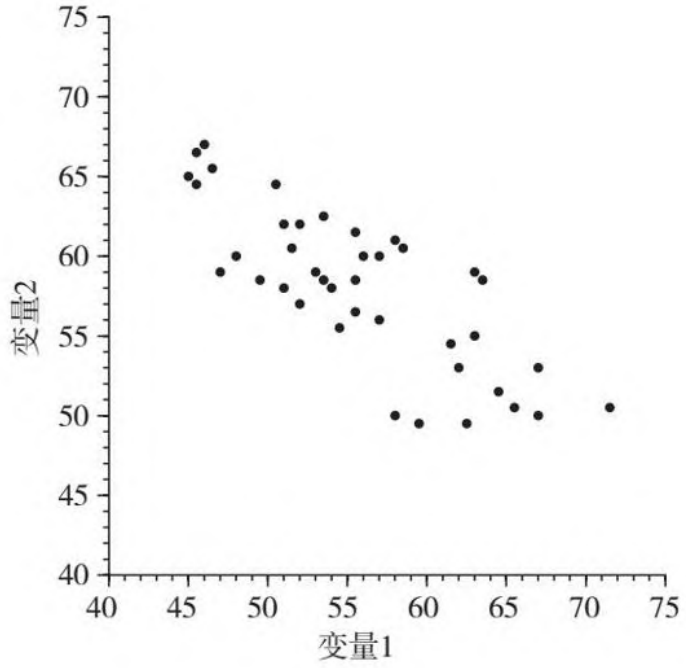


图8.1 训练数据

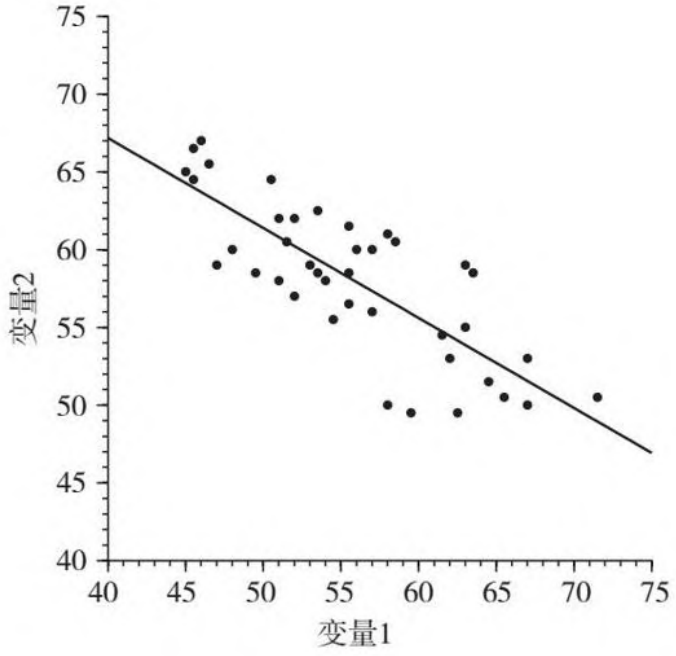


图8.2 机器学习

早在第3章我们就举过例子，说明先进的神经网络如何将看似随机的圆点和图案错误识别为猎豹和海星，将戴着有色镜框的中东年轻男子错误地识别为美国电视节目主持人卡森·达利。这些例子并不旨在展示神经网络有多么不可救药，而是说明神经网络名不副实，不能像人脑那样运转。

在使用其他数据挖掘算法的过程中，不关心逻辑或理论的神经网络算法在寻找统计学模型时也一定会出现问题，例如，不关心图8.1和图8.2中的两个变量是否应该有关联，认识不到眼镜改变不了人脸。

当用于复杂的任务时，深度神经网络暴露了另一个重大缺点——晦涩难懂。没有人确切知道它如何运作，得出的结果是否值得信赖。深度学习网络还非常脆弱，图像识别软件或许能熟练识别停车标识，但如果将一部分标识模糊处理，甚至改变一个像素，都能让它被彻底蒙蔽。

被数学蒙蔽双眼

数据挖掘者有巨大的统计程序工具箱，逐步回归法、岭回归法、主成分分析、因子分析和神经网络只是其中5种而已。我选择这5种方法的原因是它们具有代表性，并且（相对来说）容易理解。

总体而言，数据挖掘工具的数学原理往往复杂精密，但仍表现稚拙，因为它们时常妄作不合情理的假定。问题在于这些假定都藏在数学背后，工具使用者往往更关注其数学原理，对假定的内容并不感兴趣。

在美国大萧条时期，数百万人失去工作、家园、农场和企业，伟大的英国经济学家约翰·梅纳德·凯恩斯批判了盛行的古典经济模

型，该模型假定失业者为自行选择失业：

> 古典理论学家与非欧几里得世界里的欧几里得几何学者相似，他们在经验中发现，明显平行的直线通常相交，并指责线段没有保持笔直，因为这是避免出现不幸相交的唯一方法。

即便时至今日，有些经济学家还是会对批评他们模型不切实际的人回应说是这个世界出了问题，而非他们的模型。凯恩斯对这种一厢情愿的想法予以驳斥：

> 马尔萨斯之后的职业经济学家显然对他们的理论结果与事实观察缺乏对应的情况无动于衷……古典理论很有可能代表了我们想要的经济运行方式，但是假定经济确实如此运行，就等于假定我们的困难都不存在。

古典经济学家太痴迷于自己模型的美，而忽视了其真实性。

今天的数据挖掘者使用的诸多统计学工具也是如此。

第9章

先吃两片阿司匹林

IBM的“沃森”在《危险边缘》游戏中夺冠后，得到了铺天盖地的宣传，不过“沃森”的潜在价值更多体现在能够为医生、律师等需要快速准确获得信息的专业人士提供大规模的数码资料库上。

当医生怀疑病人患有某种疾病时，“沃森”可以列出可识别的症状；当医生注意到患者出现异常情况，但不确定这些症状与哪种疾病相关时，“沃森”可以列出可能的疾病；当医生确认患者得了某种疾病时，“沃森”可以列出推荐疗法。在上述每种情况下，“沃森”都会给出多种建议，随附其他相关的可能性，以及它所依据的就医记录和杂志期刊文章的超链接。

“沃森”和其他医学数据库都是宝贵的资源，可以利用计算机的能力来获取、储存和搜索信息。不过，还是有很多地方需要注意。显而易见的一点就是医学数据库远不像《危险边缘》的数据库那么可靠。人工智能算法非常擅长在数据中寻找模型，但它并不擅长评估数据的可靠性和统计学分析的合理性。

如果医生将患者的症状输入黑匣子式的数据挖掘软件并获得建议疗法，但得不到关于诊断或药方的任何解释，就可能导致悲剧性的后果。试想，出现以下情况，你会有何种反应。你的医生说：“我查不出你的病因，但电脑显示要‘服用这些药物’。”或者“我查不出你的病因，但电脑建议动手术”。

任何使用神经网络或数据规约程序的医学软件，如主成分分析和因子分析，都只是勉强能够为诊断和治疗提供解释。病患不知其所以然，医生也不知道，甚至开发黑匣子系统的软件工程师都不知道。总之，没人知道。

“沃森”和类似软件是极佳的参考工具，但它们无法替代医生，因为医学文献通常有误，数据挖掘软件的使用叠加了这些错误。

明早再给我打电话

几年前，我做了一次例行体检，量了身高、体重，回答了两页纸的问题，都是关于我的生活方式的（我不抽烟），还做了一大堆测试。护士量了我的体温、心率和血压，还检测了尿常规和血常规，检测目的具体是什么也不清楚。当天晚上，我接到回馈电话，被告知某项检测（我记不清是哪项了）的结果有些问题。95%的健康人士的该项检测结果都在“正常”范围内，而我的这项检测结果“不正常”，所以显然我的身体是不健康的。

医生说：“不用担心。”她让我吃两片阿司匹林，睡个好觉，第二天再回去复检。我照做了，第二天的复检结果正常，我也松了一口气。

是多亏了那两片阿司匹林，还是前一晚的好觉？可能两者皆非。最有可能的是，这不过为随机噪声。任凭哪个健康人来做那些检测，结果都会出现变动。一天中的不同时段、消化状况和个人情绪都会影响血压。摄取的食物和检测前运动与否都会影响胆固醇的检测结果。设备误差以及读数、记录、解读时的人为失误都容易影响检测结果。

如果一次检测结果碰巧过高或过低，再次检测的结果就可能会接近平均值。这种逆转情况让评估医学疗法的作用变得困难。就我的例子来说，根本不知道是阿司匹林还是睡个好觉起了作用。

有人说：“如果治疗得当，感冒14天就会康复；如果顺其自然，病情也就持续两周。”虽然医生说“明早再给我打电话”时，听上去像是为了少点麻烦，但这就是老方法的大智慧。

即使我感冒之后吃了阿司匹林不见效，第二天早上也还是会有所好转，因为身体有极其惊人的自愈能力。假设你身上有道伤口深到流

血了，肌体的血小板会凝固血液，然后结痂修复皮肤。这一切都是身体的自愈，无须任何医学干预。

“明早再给我打电话”的做法可行，原因有二。第一，医学测试无法完全准确检测病患状况。第二，病患的身体能对抗疾病，通常患病之后不进行治疗也都会有所好转。

比起不必要的担心，医学干预的后果更加严重。偶然波动引起的读数异常，会带来不必要的治疗。接受治疗后的检测结果改善，又会不知不觉让人相信是治疗见效了。

假设有一大批人进行体检，其中被检查出胆固醇指标最高的人会被告知要特别注意饮食。我们能预见到他的胆固醇指标会有所改善，即便饮食调节的指导无非就是“吃前请三思”。

此外，我们都知道，止痛药的效果因人而异，大多数医学治疗都是如此，没有完全有效或无效的疗法。如果有效果不显著或因患者情况不同而各异的情况出现，医学测试的结果就取决于哪些人被随机分配到了服用药物的实验组，哪些人被分配到了服用安慰剂的控制组。

统计学家尝试解释上述的随机变化，他们假设差异纯属偶然，然后评估实验组和控制组之间的差异和观察结果一样显著的可能性有多大。

P值小于等于0.05则具有统计学意义。这意味着，没有价值的被测疗法只有5%的机会显示其统计学意义，也就表示仍有5%的无价值疗法会得到具有统计学意义的结果。

医学研究是个弱肉强食的领域，才智过人和竞争力强的科学家一辈子都在为名誉和经费而奋斗，以维持其职业发展。为了达到这一目的，这些科学家需要获得并发表具有统计学意义的结果——必要时不择手段，其中就包括得州神枪手谬误1和谬误2。

研究人员只要通过大量的疗法测试就能得到有统计学意义的结果，即便他们受到了误导，测试的只是无用的疗法，在上百次无用疗

法测试后，他们还是会发现其中5%具有统计学意义——这足以促成其文章发表，使经费提案获批。

同样，医药公司能够从临床“验证”有效的疗法中获得巨额利润。确保某些疗法得到支持的一种方法是，测试数以千计的疗法，无论遇到多少统计学障碍，运气都能确保某些无用疗法跨越所有障碍。

下面让我们一起来看三个“得州神枪手”的例子。

我要再喝一杯咖啡

20世纪80年代早期，据全世界顶尖的医学期刊《新英格兰医学期刊》报道，广受赞誉的研究者、哈佛公共卫生学院院长布莱恩·迈克马宏所带领的团队发现“饮用咖啡与胰腺癌有极大关联”。这个来自哈佛大学的团队建议人们不要再喝咖啡，以降低患胰腺癌的风险。在此项研究之前，迈克马宏自己每天都喝三杯咖啡，在此之后他就再也不喝了。

这就出现了得州神枪手谬误1中的问题。该研究旨在调查喝酒或抽烟与患胰腺癌之间的联系，迈克马宏研究过酒类、香烟、雪茄、烟斗，没有任何发现，于是他就继续找，又研究了茶叶。最后，他终于在咖啡上有了发现：胰腺癌患者喝的咖啡多。

如果上述六项测试都单独进行，每项测试都包含一些与胰腺癌无关的因素，那么有26%的概率会在至少一项测试中产生一个具有统计学意义（P值为0.05）的关联，也就是说有26%的机会可以无中生有。

迈克马宏的研究还有另一个缺陷。他将患胰腺癌的住院病人与患其他疾病的病人进行对比，并且这些病人都由同一批医生负责。问题在于，这些医生通常都是胃肠专科医生，他们的很多患者都因为害怕

溃疡恶化而戒了咖啡。但胰腺癌患者没有停止喝咖啡，他们中喝咖啡的人更多。所以并非喝咖啡导致了胰腺癌，而是患其他疾病的病人不再喝咖啡了。

后续研究——其中一项来自迈克马宏的团队——也未能证实最初的研究结果。这一次，他们得出的结论是：“据观察，与早前研究相比，喝咖啡对男性或女性都不存在危险。”美国癌症协会也认为：“最近的科学研究表明，喝咖啡和患胰腺癌、乳腺癌等癌症没有任何关系。”

更近期的研究不仅驳斥了迈克马宏最初的研究结果，而且结果显示喝咖啡（至少对男性来说）反而会降低患胰腺癌的概率！

远程治疗

20世纪90年代，年轻的伊丽莎白·塔尔格医生研究了遥远的祈祷和其他积极意念是否能治愈晚期艾滋病患者。40名艾滋病患者被分成两组。祈祷组患者的照片会被发送给有经验的远程治疗师（从佛教、基督教、犹太教信徒到萨满巫师都有），他们与病患平均相隔约1 491英里。非祈祷组的20名患者则完全靠自己。

此次测试采用“双盲”（double-blinded）程序，塔尔格和病患都不知道哪些病患是祈祷组的，以免影响测试结果。

为期六个月的研究发现，祈祷组的患者就医时间更短，罹患与艾滋病有关的疾病更少。这次研究的结果具有统计学意义，发表在享誉盛名的医学期刊上。人们出于各自的目的引用塔尔格的研究来证明上帝的存在，或是指出传统观念对心智、身体、时间和空间的认识不足。

美国国家卫生研究所给塔尔格拨款150万美元，用以更大规模的艾滋病患者研究和对远程治疗师能否缩小脑癌患者的恶性肿瘤的调查。就在获得拨款不久后，塔尔格自己也被诊断出得了脑癌，尽管世界各地都有治疗师为她祈祷和发送治疗能量，但她还是在四个月后去世了。

塔尔格去世后，其早前对40名艾滋病患者开展的研究也被查出了问题。之前，她计划对比祈祷组和非祈祷组的死亡率，然而，在为期六个月的研究进行了一个月后，“三联鸡尾酒疗法”（triple-cocktail therapy）开始流行，40名患者中只有1人死亡，这表明该疗法的有效性，但它也消除了祈祷组和非祈祷组进行统计学对比的可能性。

于是，塔尔格及其同事弗雷德·西歇尔转而寻找两组之间的其他差异。他们参考了各种身体症状、生活质量测量、情绪评分和CP4+指数，两组患者在这些方面均无差异。塔尔格的父亲曾经试图通过实验证明人类拥有可感知看不见的物体、读心和仅靠意念移动物体的超自然能力。他要女儿塔尔格继续寻找，只要怀有信念，相反的证据就无足轻重，只要继续在数据中搜寻支持自己信念的证据即可。最终，塔尔格找到了——住院时长和医生探访，尽管医疗保险肯定会使问题雪上加霜。

随后，塔尔格和西歇尔读到了一篇列举了23种与艾滋病相关的疾病的文章。他们或许可以寻找两组实验对象在这23种疾病上的差异。不幸的是，由于采取“双盲”安排，这些疾病的数据均未被记录。塔尔格和西歇尔坚持不懈地仔细研读受试对象的医疗记录，即便他们现在已经知道每名患者的分组情况。完成研读后，他们报告称祈祷组在某些疾病方面比非祈祷组的境况更好。这种积极主动的数据挖掘似乎没有利用数据挖掘软件就完成了。

他们发表的论文显示，该项研究是为调查具有统计学意义的几种疾病而设计的（即得州神枪手谬误1），他们做过的其他测试都未公

布，也没有说明最终数据是在研究结束后搜集到的，而且“双盲”控制也被撤销了。他们得到了想要的结果，或许是因为他们的坚持，或许是因为数据不再是双盲状态。

塔尔格的美国国家卫生研究所的研究在她去世后仍在继续。祈祷组和非祈祷组在死亡率、患病或症状方面都未发现有意义的差异。另一项规模更大的研究由哈佛医学院的研究人员执行，观察了1 800名处于冠状动脉搭桥术后康复期的患者，还是没有在祈祷组和非祈祷组的患者间发现明显差异。

癌症群

20世纪70年代，流行病学家南希·韦特海默和物理学家埃德·利珀驾车穿过科罗拉多州丹佛市去考察一些人的住所，这些人未满19岁便因身患癌症离开了人世。他们试图发现这些人住所的共同特征。两人注意到，很多罹患癌症的人都住在大功率电力线附近，因此得出结论：暴露于电力线的电磁场中会导致罹患癌症。

记者保罗·布罗德为《纽约客》写了三篇文章，报道了关于电力线和癌症相关系数的其他奇闻逸事。他还做出了不详警告：“数以千计没有戒备的儿童和成人会罹患癌症，其中很多人都会英年早逝，他们本不该遭此厄运，一切只因他们暴露在电力线的电磁场中。”

这种言论随之在全国造成轰动，为咨询专家、研究人员、律师和包括高斯计（测量磁感应强度的仪器）在内的各种装置提供了有利可图的机会，人们可以用高斯计在家测量电磁场的强度（电磁场读数高的房间会被封住，只用作储物间）。幸运的是，政府并没有扯掉整个国家的电力线。

此次恐慌事件的问题在于，即使癌症患者在人口中只是随机分布的，数据挖掘都更有可能发现受害者在地理上集中的地方。为了说明这一点，我虚构出一个有1万名居民的城市，其住所均匀分布于整座城市，每个人患癌的概率都是1%（我忽略了家人一起居住的情况和年龄因素）。然后，我使用计算机随机数字生成器来决定谁是这座虚构城市中的癌症患者。据此得出的癌症患者分布如图9.1所示。每个小黑点代表住着一名癌症患者的一户人家，而白色区域即无癌症患者居住。

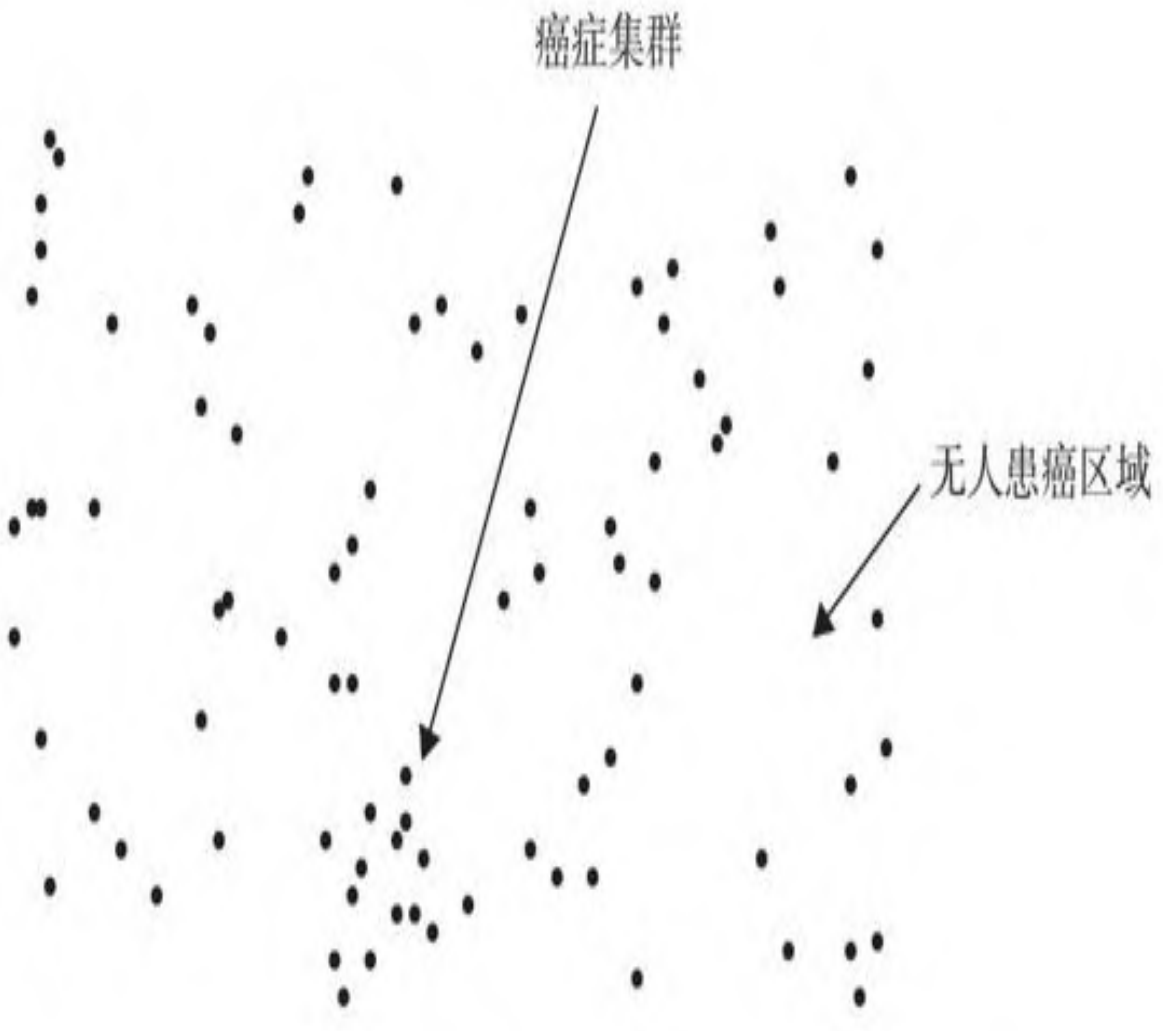


图9.1 癌症患者分布图

随便一个像样的数据挖掘软件都能轻易发现，图9.1的底部明显有一处癌症患者集中地。如果这座城市真实存在，我们就可以驾车到患者住所附近，肯定能得到一些特别发现。或者使用数据挖掘软件搜遍数据，寻找异常状况。如果我们再将住在球场附近的居民患癌率与住所远离球场的居民患癌率相比，猜猜结果如何？球场附近的患癌率更高，这表明住在球场附近可致癌。

图9.1还显示了癌症堡垒，即无人患癌的区域。利用数据挖掘软件或驾车到附近瞧一瞧，一定会得到一些关于这个无人患癌区域的特殊发现。可能该地附近建有水塔。如果我们再将住在水塔附近的居民患癌率与住所远离水塔的居民患癌率相比，一定能发现水塔附近的患癌率更低。这就是我们选择这个区域的原因——这里没人患癌。

无论是在球场还是水塔附近，都存在同样的问题——得州神枪手谬误2。如果我们使用数据来创造理论（小联盟球场会致癌，水塔可防癌），数据当然会支持理论了！怎么会有相反结果呢？我们会捏造出与数据不吻合的理论吗？

用来创建理论的数据肯定不适于再来检验该理论。我们需要全新的数据。其他国家的研究没有发现电磁场和癌症之间存在关联。以啮齿动物为对象的实验研究发现，比电力线所产生的更强的电磁场对死亡率、患癌率、免疫系统、生育率或出生缺陷率都没有影响。

对电力线的恐慌有什么理论基础吗？科学家非常了解电磁场，并没有任何合理理论能证明电力线的电磁场会致癌。电力线的电磁能量远比月光的电磁能量弱得多，其电磁场也比地球的磁场更弱。

权衡理论论证和实验结果后，美国国家科学院得出的结论是：电力线并没有造成公共健康危险，无须提供经费开展进一步研究，更别说撤掉电力线了。全美顶尖医学期刊也发声力挺，同意不应再把研究资源浪费在这个问题上。

1999年，《纽约客》发表了一篇题为“癌症集群之谎言”（The Cancer-Cluster Myth）的文章，含蓄地驳斥保罗·布罗德早先的报道。尽管如此，癌症集群具有意义的想法还是继续存在。互联网上，由政府赞助的交互式地图可按地理区域显示各种癌症的发病率，精细到人口普查的街区。每年都需要花费数百万美元来维护地图数据，虽然数据是最新的，但很可能具有误导性。其中一个交互式网站拥有22种癌症、2种性别、4个年龄段组别、5个种族和3 000多个县的癌症死亡率数据。从数百万种可能的相关系数中，数据挖掘软件一定可以轻易发现令人恐惧的相关系数。

为了缓解这种恐惧，美国疾病控制与预防中心创建了网页平台，任何人都可以在此报告自己发现的癌症集群。即使该中心提醒：“我们会对此进行后续调查，但需要花费多年时间才能完成，结果通常也不能得出定论（也就是说，通常都无法找到原因）。”每年仍有1 000多例癌症集群被举报和调查。

最有理有据的疗法失效了

大量已发表的医学研究都会犯那两个得州神枪手谬误：数据的随机变化只在人们忽略以下情况时有意义，即这些侥幸发现都是靠测试大量理论，或创造理论来匹配数据中的偶然模型才能得到，报告的结果随后便消失得无踪无影。这种模型在医学研究中太常见了，以至于还有专门的叫法——递减效应（decline effect）。

有些研究人员亲眼见过自己的研究出现递减效应，他们都迷惑不解，因此开始白费力气地寻求解释，尽管原因就近在眼前。如果最初的正相关发现皆因得州神枪手谬误，那么随后的结果通常都令人失望

也就不足为奇了。这就好比基于偏远城市气温进行总统大选预测那样。

看似有效的无价值疗法只是假阳性结果。另外还有假阴性结果，即有效疗法并未显示出统计学意义。仔细想想，一个测试有5%的机会呈假阳性，就意味着一项经受严格测试的无效疗法，其实验组和控制组之间出现统计学差异的机会为5%。假设假阴性的概率为10%，就表示有效疗法在测试顺利的情况下，无法显示出统计学意义的概率为10%。

如果假阳性的概率为5%，假阴性的为10%，似乎我们每次都应该能分辨出有效和无效疗法之间的区别。实则不然。那要看有多少受试疗法有效，有多少无效。若所有受试疗法中，1%为有效，99%为无效，则结果如表9.1所示。

表9.1 所有经验证的疗法中，有85%为无效

	有统计学意义	无统计学意义	总计
有效疗法	90	10	100
无效疗法	495	9 405	9 900
总　计	585	9 915	10 000

测试10 000种疗法，其中100种有效。这100种有效疗法中，90种会呈现具有统计学意义的结果；而另外9 900种无效疗法中，会有495种呈现具有统计学意义的假阳性结果。因此，共计585种测试具有统计

学意义，但其中只有90种为真正有效的疗法，有85%“经验证”有效的疗法实际上毫无价值，这让人难以置信。

这一矛盾反映出有关逆概率的常见困惑。超级联赛的所有运动员都是男性，但所有男性中，只有很小一部分人为超级联赛的运动员。同理，所有有效疗法中，90%都具有统计学意义，但所有具有统计学意义的疗法中，只有15%有效。

任职于希腊约阿尼纳大学、马萨诸塞州塔夫茨大学医学院和加州斯坦福大学医学院的约翰·约安尼季斯以此类运算为依据，发表了一篇以“为何大多数已发表的研究成果都有误”（Why Most Published Research Findings Are False）为题的引起争议的文章。

约安尼季斯在整个职业生涯中都在提醒医生和普通民众，不要轻易相信复制结果无法令人信服的医学测试。他那篇题目惊人的著名文章就采用了我们上述的数学算法，他的假设观点比我们的更加令人确信，而概率的表现也更加糟糕。

除了这些理论性计算，约安尼季斯还汇编列举了在现实世界中“经验证”的疗法最后无效的例子。他在一项研究中检查了45个发表于1990—2003年且广受赞誉的医学研究成果，其中仅有34个能使用更大样本对原始结果进行复制，这其中又只有20个（即59%）证实了最初的结果，7个所述疗法的疗效比最初推算的小得多，剩下那7个疗法则根本一点效果都没有。总的来说，45项研究中仅20项可经证实，这些可都是最享誉盛名的研究啊！对于发表在级别较低的期刊的数千篇研究来说，情况肯定更糟糕。约安尼季斯粗略估算，90%已发表的医学研究成果均有漏洞，其宣称有效的疗法被夸大了效果，有的疗法则毫无效果，甚至更糟。

疾病诊断和治疗中的数据挖掘

传统的统计学测试假定研究人员会以定义好的理论为起始，然后收集合适的数据来验证他们的理论。数据挖掘则另辟蹊径——数据为先，理论在后。因此，可以随意检测所有你想要检测的理论，无论这些理论是否合理。

如果医学疗法没有对整个样本显示出统计学意义，再看看其是否适用于子集；将性别、种族和年龄分开，尝试不同的年龄段；如果该疗法对你最初研究的疾病不起作用，再看看它是否有其他益处。

测试数百种疗法便是得州神枪手谬误1的例子：瞄准数百个目标，只报告那些击中的情况。其他医学研究则有关得州神枪手谬误2：找到一个模式，然后为其编造解释。疾病诊断或治疗都会出现上述情况。

首先讨论一下疾病诊断。假设我们知道100个患者患了某种疾病，不知道另外100个患者患了什么疾病，然后记录下每个人的1 000种特征，比如血液检测、基因信息、种族、发色、瞳孔颜色和住处等。如果我们现在使用数据挖掘软件来彻查这一数据库，肯定会找到一些特征，这些特征在患病人士中比在健康人士中更加常见，而且明显能够很好地预测疾病。

例如，我能够获取87名女性心脏收缩血压读数的数据库，还有每名患者的40种特征的完整信息，有些以数字表示（如年龄），有些按类别区分（如某人是否有吸烟史）。

我使用了数据挖掘软件，来看看根据这40种特征预测血压的结果如何。如果我的模型契合度高，就可以用来识别其他有高血压风险的女性。我们还可以识别出其他高风险因素（可能是吸烟）并建议血压值高的女性改变行为，以降低血压。

该模型非常成功，实际血压和预测血压的相关系数达到惊人的0.72，即23名受试女性的预测心脏收缩血压高于130，符合其中17人的实际情况。图9.2为87名女性的预测值和实际值。

我们还可以只用病患的5个特征（特征1、12、18、23和34）得到预测血压值和实际血压值的相关系数为0.47，这一结果相当不错。因此，医生会重点关注这五个特征，可以预测，甚至可能控制血压升高。

那么这五个特征是什么？随机数字。我捏造了87名女性，使用87这个数字是为了提高研究的真实性。对于其中20个特征，我使用电脑抛硬币的方式来赋予其1值或0值。同样，抽烟者被赋予1值，不抽烟的人则被赋予0值。对于另外20个特征，我用电脑生成了正态分布的随机变量，均值为100，标准差为10。虚假的血压数值也是正态分布，均值为125，标准差为10。我捏造的每一个女性及其每一个特征都与其他女性的捏造特征相互独立，与该女性的虚假血压及其另外39个捏造特征也相互独立。

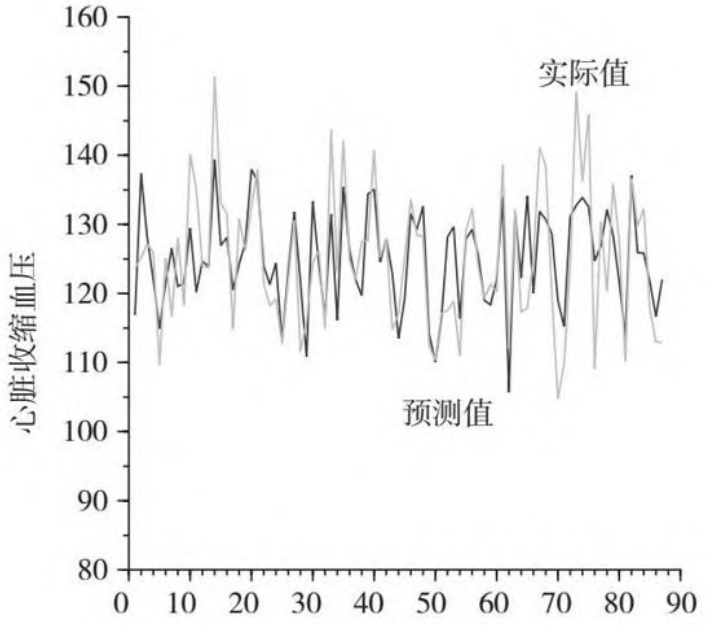


图9.2 心脏收缩血压的预测值和实际值

我在这个虚假的医学数据库中填入随机数字来证明我的观点。即使数据库中记录的特征与所分析的疾病是否存在没有任何关系，数据

挖掘软件也会发现具有统计学意义的关系，让人误以为获得了什么有用的发现。

疾病治疗也是如此。假设根据各种各样的医疗状况对病患施行尝试性疗法，使用数据挖掘软件来识别那些有所改善的疾病或疾病组合。即使患者状况的波动完全随机，与他们是否接受了治疗一点关系都没有，还是很有可能存在具有统计学意义的模式表明该疗法对某些状况有效。前面说到的远程治疗研究就是这种谬误的很好例证。

糟糠过多，精粹不足

很多神奇疗法（如胰岛素和天花疫苗）都被医疗研究发现并且证实为有效。然而，很多已发表的研究都有缺陷，这通常是因为那些数据都是为了发表而搜刮得来的。

对“沃森”等医疗建议软件来说，这是无法逾越的难题。它们都非常擅长收集、储存和搜索医疗数据和期刊文章，这一点肯定优于人类。但是它们没有常识或智慧，不知道数字和词语的意思，无法评估数据库中内容的相关性和有效性。它们也无法分辨好数据和坏数据，不能识别哪些数据受到过两种得州神枪手谬误的拷问。此外，它们还无法区分因果关系和随机事件，其数据挖掘式的“知识发现”甚至会让这一问题难上加难。

所有医疗专业人士都学过的准则是：首先不能造成伤害。有经验的医生对医学研究总会保持良性怀疑态度，对不喝咖啡、依赖远程祈祷和撤掉电力线都抱着“等等看”的态度。他们了解发表论文的压力和递减效应，对黑匣子数据挖掘心存质疑。我的私人医生对“依赖黑匣子算法开处方或提供医疗方案”的观点嗤之以鼻。

医疗软件程序可以辅助医生，但无法取代医生。

第10章

完胜股市（上）

从前，在诈骗专家还在寄平邮信而不是发电子邮件时，我收到了一封开头为“亲爱的朋友”的信件，一看就知道是推销产品的。不过我还是往下读，看到黄色高亮的一句话：“一年内，1 000美元轻松变为3.45万美元！”真正的朋友不会将那句话标成高亮，但我仍看完了，想着可以和班上的学生分享这封胡说八道的信。毫无疑问，这就是一场骗局。

信中还提到，“无须特殊背景和学历”以及“花点儿买彩票或赌马的闲钱就能做的投资”。读到这里，我不确定是否要分享这封信，唯恐学生疑心这家公司是怎么打听到我的。我不买彩票，也不赌马。我做过什么，让这家公司认为我是个冤大头？

信中还说，与其把钱都浪费在买彩票和赌马上，我还不如通过投资低价股票致富。例如，某公司近几个月的股价从每股2美分升至69美分，投资1 000美元即可收获3.45万美元。我只要花39美元购买一份特殊的报告资料，就能进入“少数精明的‘内部’投资者精心守护的领土”。

整件事的前提条件荒唐可笑。如果真的有人知道怎么把1 000美元变成3.45万美元，他们就不需要卖内部资料赚那39美元了。可是，仍有人一次次落入这种骗局，因为这些人总是想当然地认为，世界就是依照我们可以发现和利用的规律法则运行的。股价不会是随机的，肯定有其内在模式，就像日夜交替、季节轮回一样。而且让人们更容易受骗的推手，是人类贪婪的本性以及赚快钱的想法——买卖股票比当个肉贩、面包师或蜡烛生产者更容易发家致富。

令人尴尬的事实是，股价的升降起伏大多时候是随机的，不过，在随机数字中还是可以发现一些转瞬即逝的模型。如果我们用心寻找，就肯定会找到，还会被它们耍得团团转。

噪声

人们通常会受到股市的诱惑，误以为股市里遍地黄金，伸手即来，以为自己能看准时机在上涨前买入、下跌前卖出。他们看到像信中提到的那种股票，心里会打着算盘计算自己少赚了多少，又想象自己将来能赚多少。

往回看的话，很容易发现错过了很多机会。往前看的话则很难预测股票的涨跌。股市不会白白送钱。

股市不是计算机程序，不会依照规则定价。股价会受到投资者交易意愿的影响，有买有卖。如果明显会在不久后涨到每股69美分，投资者肯定不愿卖出现在每股2美分的股票。即使只有精明的内部人士才知道股价很快就会涨到69美分，投资者还是会买入数百万股，刺激股价在当天就升至69美分。

交易价为每股2美分时，买方和卖方一样多，因为有人认为价格过高，也有人认为价格实惠。股价即将飙升或暴跌，从来都不是显而易见的事（如果是的话，买方和卖方之间就失去了平衡）。

难以预测股价的原因有以下两个。第一，新信息（即投资者了解到前所未知的信息）会造成股价变化。所谓新信息，顾名思义，是无法预测的，因此由新信息导致的股价变动也无法预测。第二，买卖双方都受到人类情绪和误解的影响，这些情绪和误解变幻莫测，有时并不理智，凯恩斯称其为“动物精神”（animal spirit）。

数百万投资者耗费大量时间，试图发现能成功预测股市的公式。有人无意间发现了能成功解释过去状况的规则，却不能准确预测未来趋势，这不足为奇。很多类似的交易系统都荒唐可笑，竟有人还信以为真。这些滑稽理论的共同点在于它们都是以数据挖掘为基础的。

最危险的交易系统是黑匣子模型，使用精密复杂的统计学分析挖掘复杂到难以理解的模型，如主成分分析和因子分子，因此，根本不可能知道这些模型是否合理。我们必须相信数据挖掘的发现是真实可信的，而非“愚人金”（fool’s gold）[1]。

既然是黑匣子，我们就不知道这种投资模型里装的是什么，但可以从公开透明的数据挖掘系统中学到一些东西。

滑稽的理论

女性青睐高跟鞋时是牛市，选择低跟鞋时是熊市；男性爱系细领带时是牛市，爱系宽领带时为熊市。尽管《华尔街日报》表示“有些分析师持相反意见，那些到布克兄弟（Brooks Brothers）购买男装的人并没有意识到领带还有不同的宽度”。

分析师监测了太阳黑子、五大湖的水位还有阿司匹林与黄色涂料的销售量。有人认为，在以数字5结尾的年份，股市都会表现出色，如1975年、1985年等；还有人认为，股市在以8结尾的年份表现最佳。长期任职于《纽约时报》的金融专栏作家伯顿·克兰曾报道，一名男子“根据他对刊登在《纽约太阳报》上连载漫画的解读提供相当成功的投资咨询服务”。《金钱》杂志曾报道，明尼阿波利斯市有名股票经纪人的选股方式是把《华尔街日报》摊开扔在地上，他的金毛猎犬的右爪第一片趾甲碰到哪只股票他就选择哪只。他认为这样做能帮他吸引客户，这也表明了他以及他的客户是什么样的人。

1987年有三个星期五是13号，费城银行首席经济学家称，在过去40年，另外还有六个年份出现了三个星期五是13号的情况，并且在其中三个年份都发生了经济衰退。哪种情况更糟糕，是经济学家把时间

浪费在这种荒唐言论上的想法，还是有人付钱让他如此浪费时间的事实？不管怎样，在1987年都没有出现经济衰退。

可能大家都在比赛，看谁的建议更傻。罗斯柴尔德投资银行在破产前称，在过去的六次中国农历龙年期间，美国股市四次上涨，两次下跌。

我最喜欢举的一个例子是波士顿下雪指标——根据波士顿在平安夜是否下雪来预测下一年的股市。德崇证券的一名分析师曾表示："平安夜若下雪，则下一年的平均收益比不下雪的年份大概高80%。"

超级碗股市预测指出，如果美国国家橄榄球联合会（NFC）的球队，或者曾加入美国职业橄榄球大联盟（NFL）而现在属于美国橄榄球联合会（AFC）的球队在超级碗中夺冠，股市便会上涨；反之股市则会下跌。绿湾包装工队赢了对股市有好处；而纽约喷气机队赢了则对股市不好。发明"超级碗指标"的人旨在用一种幽默的方式说明，相关系数不一定包含因果关系。让他目瞪口呆的是，竟有人开始认真看待这个指标！直到2017年，我听说有些大基金的经理仍把它当真。正如我反复强调的，有些人认为所有模型都具有说服力。这让人无话可说。

2014年，我教过的学生给我转发了一条交易规则，由高盛资产管理公司前主席编造而来：

> 有个奇怪的说法叫作"五天法则"……简单来说，该法则指当美国股市主要股指在最初五个交易日出现综合正收益时，则该年份整体上极有可能表现不错……
>
> 我在高盛的前同事何塞·乌苏亚回溯了直至1928年的相关数据。他发现，股市在一年中的头五个交易日上升时，全年为升势的概率是75.4%。而在1950年至今的这段时期，该概率上升至82.9%。金融领域难得有如此清晰明了的法则。

首先，这听上去就像拷问数据的结果。（为什么是五天？为什么回溯到1928年？）其次，最初五个交易日是全年（大约）250个交易日的一部分，因此这是有偏倚的计算，他应该对比最初五个交易日和其他245个交易日。最后，认为股市上涨和下跌的概率为五五开是错误的。实际上股市通常都会上涨。1928—2013年，股票上涨的概率为73%，而1950—2013年，此概率为78%。“五天法则”并没有过多地提高这一概率，尤其考虑到那五天都包含于该年份内。

这些投资法则都是传统守旧且毫无价值的、靠人力耗费无数时间查遍数据才得出的模型。然而，13号遭遇星期五、龙年等胡言乱语，恰好就是用来寻找模型的数据挖掘软件取得的无用发现——唯一不同的是，计算机在发现毫无价值的模型上效率要高很多。

有时，所发现的模型看似合理，但显然就是拷问数据的无用结果。例如，美联储的货币政策会影响股价，这是合理的说法。低利率降低消费成本和企业贷款成本，对投资者来说，低利率让股票比债券更具吸引力。曾任美国白宫经济顾问委员会主席的著名经济学家贝里尔·斯普林克尔认为，货币供应量的变化会对股市产生影响，便主动寻找两者之间的关系，最后真的找到了。他选择了哪种股价指数呢？是标准普尔425工业指数而不是常用的标准普尔500指数。又选择了哪种反映货币供应量的指标呢？他选择了一个非常狭隘的指标：狭义货币（M1）。

在目测标准普尔425指数和M1的图表后，斯普林克尔总结道，虽然两者关系非常易变，但“货币供应量的变化会引起股价变化，一般出现在熊市前的15个月，牛市前的两个月”。注意其中非常灵活的说法“货币供应量的变化”。是出现在上周？上个月？还是去年？斯普林克尔选择了6个月的范围，无疑之前试过了其他可能。还要注意含义模糊的“牛市”和“熊市”。没人知道牛市和熊市何时开始，何时结束。

又一番绞尽脑汁后，他发现1918—1960年，以下策略可以让你获得每年平均6%的盈利，而简单的买入和持股的盈利为每年5.5%：

a. 掌握前6个月随季度调整的M1的月平均增长率的最新消息。

b. 这一移动平均数出现波谷的两个月后买入股票。

c. 这一移动平均数达到波峰的15个月后卖出股票。

如此周折，也就才提高了0.5%？如今，计算机数据挖掘算法肯定会做得更好。

斯普林克尔错综复杂的模型听起来就像数据拷问。为什么始于1918年？为什么选择标准普尔425指数？为什么选择M1？为什么是6个月的月平均增长率？为什么是分别间隔两个月和15个月？后期的研究发现，货币供应量对预测股价变化毫无作用，希望对此你并不惊讶。

斯普林克尔在发现这个无用法则的过程中，无疑必须费尽心思做些必要的计算，画出一些图表，摆弄资料明细，直至发现与数据非常吻合的结果。不幸的是，现今的计算机可以以极快的速度找到这种毫无价值的法则。

技术分析

预测股市何难之有？任何明眼人都能在股价中看出不同的模型——预示股价是升是降的模型。

30年前，我的一个学生杰夫来电告诉我一个激动人心的消息——我在课上说的“企图识别出股价的盈利模型是痴人说梦”这番话没道

理。杰夫在IBM任职，他利用闲暇时间研究股价，还找到了一些明显的模型。他还在不断调整自己的系统，很快就能靠此致富。他说到时候要租架直升机飞过来，停在我教室外面的草坪上，然后以胜利的姿态大步迈进我的投资学课堂，把实情告诉班上的学生。

我每年都跟学生重述这个故事，讲完后走到教室窗边往外瞧，看看杰夫的直升机有没有停在草坪上。直到今天我还在等他。

技术分析师尝试通过研究股价、交易量和各种投资者情绪来预估投资者的行为。技术人员并不关心单个公司或整个市场的股息或盈利情况。如果他们研究单个公司，是不需要知道该公司名称的，因为这可能会让他们在浏览图表时不够客观。数据挖掘算法的工作正是如此：在数字中寻找模型，但不会考虑这些数字代表什么。

约翰·马吉与他人合作撰写的书被誉为“技术分析圣经”，他把自己办公室的窗户都封起来，这样他对股市的希望和恐惧就不会受看到小鸟歌唱或天空飘雪的影响。他待在自己的黑匣子里。

技术人员最重要的工具是股价图，最常用的是垂线图，一般用来表示每日股价。每条垂线都从最高价延伸至最低价，水平斜线则显示开盘价和收盘价。技术人员会添加线条（如图10.1所示的通道）表示走势或其他可利用的模型。（简单起见，我省略了垂线，只显示收盘价。）

技术分析的核心是从以往的股价中识别出可用来预测未来股价的模型。为使模型看似合理，还为各种模型贴上了标签，如通道（channel）、支撑线（support level）、阻力线（resistance level）、双顶（double top）、双底（double bottom）、头肩顶模型（head and shoulders）和杯柄形态（cup and handle）。尽管有这些诱人的标签，但一项又一项研究发现技术分析是徒劳无益的，除了能提供很多技术分析师的工作岗位和为股票经纪人获取佣金。

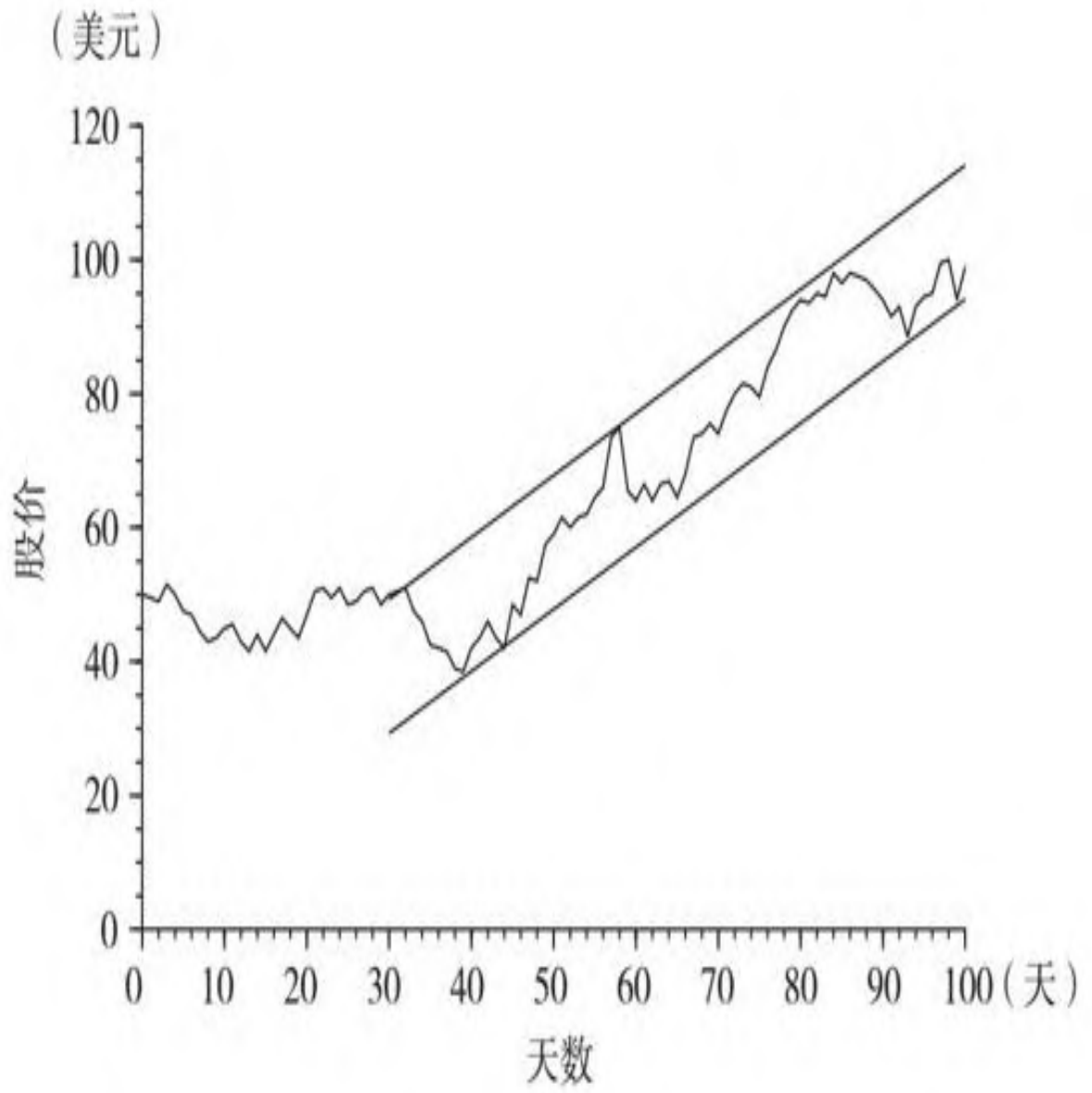


图10.1 上行通道

一名学术型经济学家曾发送几张股价图，包括图10.1、图10.2、图10.3（没有图中的横线），给一名技术分析师（且称其为埃德），请他帮忙判断哪些股票是好的投资。该经济学家没有说明是哪些公司的股票，埃德也没问。

埃德在图10.1中画了两条平行线，模型清晰可见。约从第30天开始，该股票在狭窄的上行通道中交易。到第100天，股价接近该通道的下轨，并呈现明显的上冲姿态。

如图10.2所示的股价图中带有支撑线，它在第24天成形，之后经两次确认。每当股价跌至28美元时都会获得支撑出现回升。更具威力的是，该图还呈现出头肩顶模型，即股价先是脱离28美元的支撑线上涨，接着跌回28美元，然后涨幅更大，随之又跌回28美元，小幅上涨后，第三次下降至28美元。

埃德和很多技术分析师一样，认为头肩顶模型形成并确认的支撑线极其稳固，当股价在第99天突破28美元的支撑线时，就是发出了不容置疑的信号，表明一定是出现了非常严重的问题，只有雪崩式的坏消息才能突破如此强劲的支撑线。28美元的屏障一旦被突破，股价一定下跌。

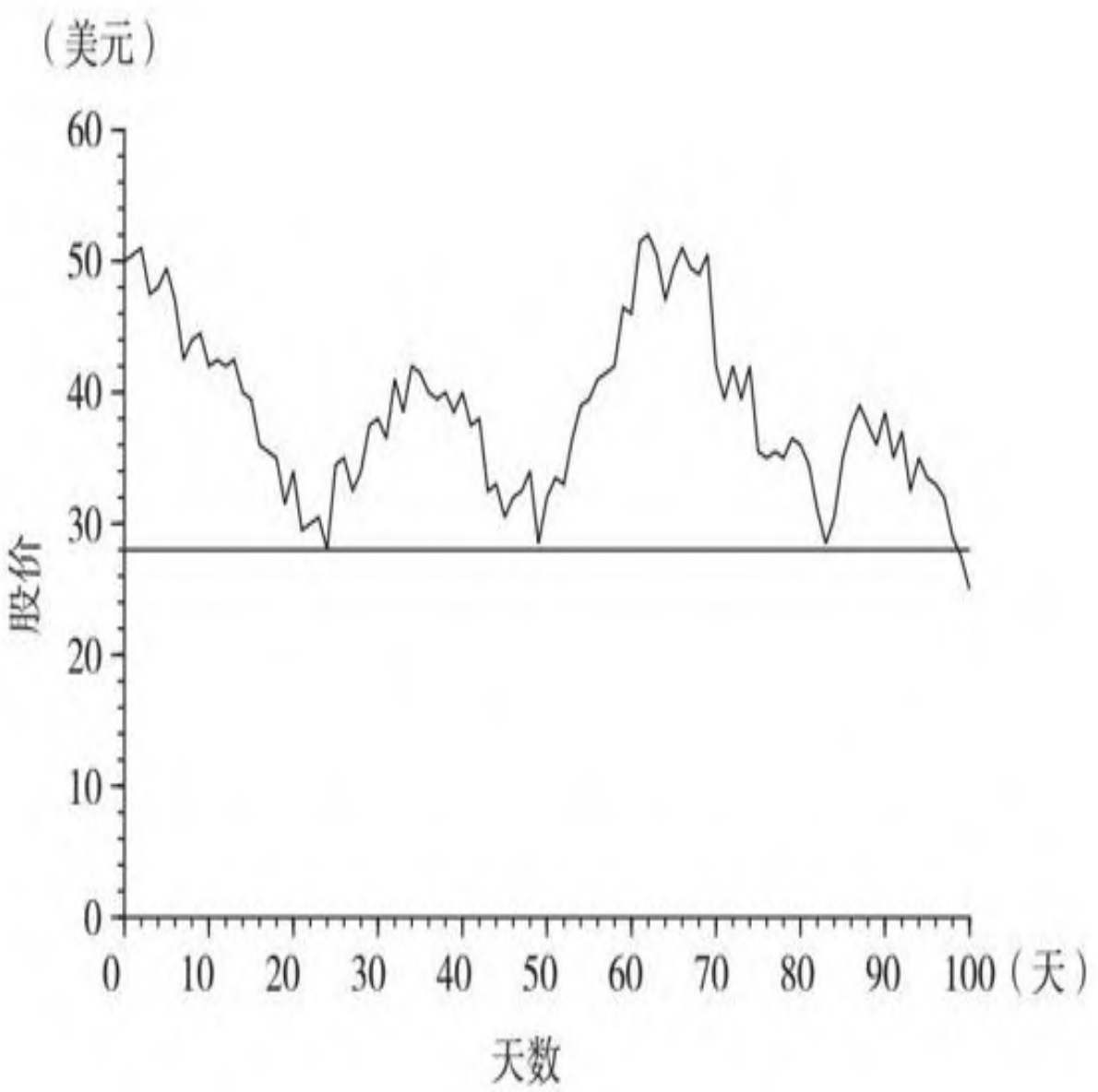


图10.2 头肩顶支撑线被突破

图10.3则显示了相反的模型。在65美元处形成了阻力线并得到确认。每当股价接近65美元就会出现回落。这种情况出现得越频繁，阻碍股价突破65美元的心理阻力就越大。更有威力的是，该图呈现出头肩底模型，因为第二次从65美元回落的幅度比第一次和第三次都大。第98天，当股价暴涨突破阻力线时，打消了阻止股价进一步上涨的心理障碍，这是不容置疑的买入信号。

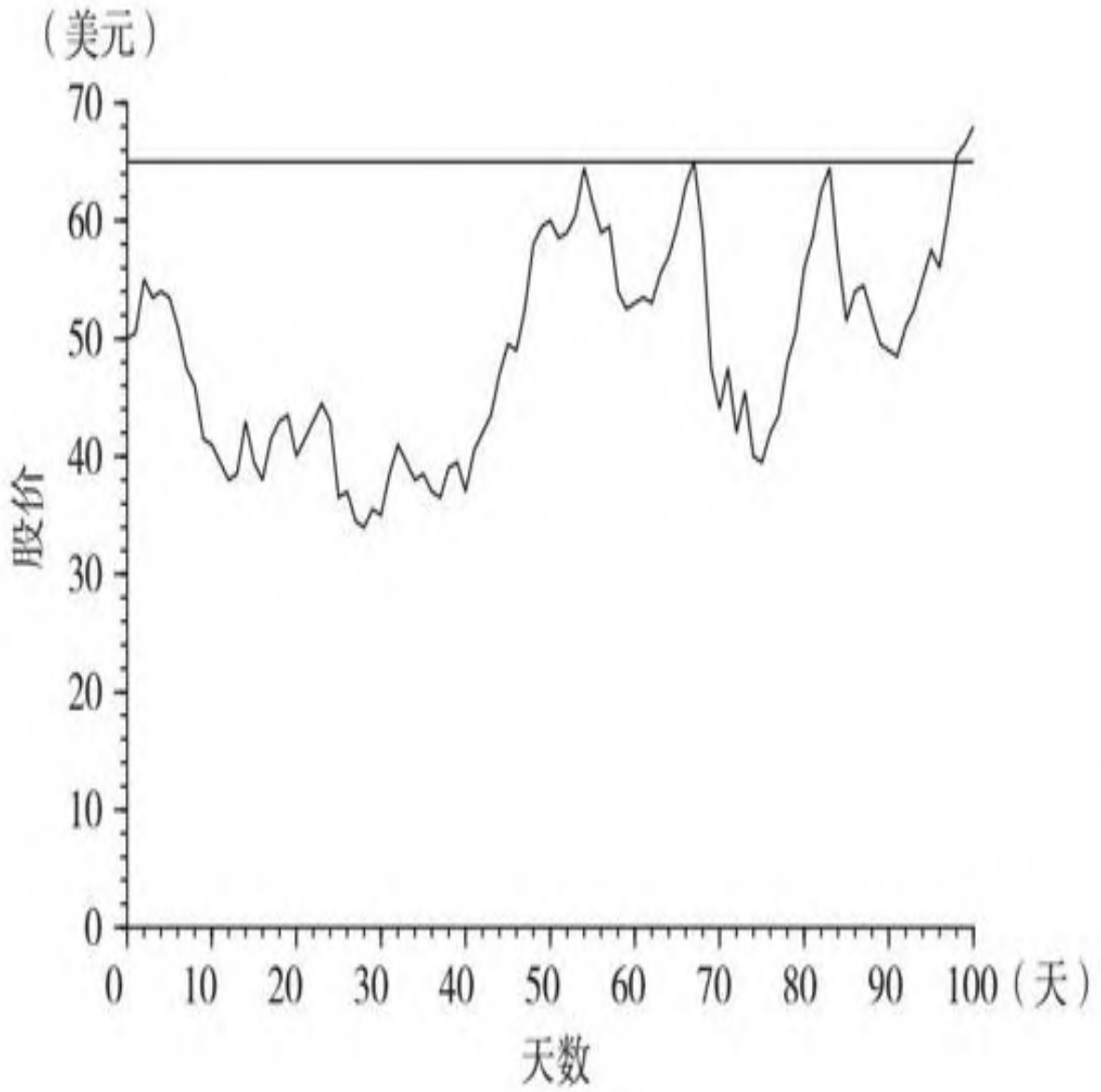


图10.3 头肩底支撑线被突破

在这三张图中发现上述模型的埃德被喜悦冲昏了头脑，忽略了这三张股价图的离奇巧合，即股价都均始于50美元。但这并非巧合。

这些都不是真正的股票。那个恶作剧的经济学教授（没错，就是我）在一次投资学课堂上让学生以抛硬币的方式捏造了这些虚假数据。每张图的“股价”初始值都定为50美元。然后，每天的股价变化

都通过抛25次硬币来决定，正面则股价涨50美分，反面则跌50美分。例如，若出现14次正面和11次反面，则当天股价上涨1.5美元。创建了几十张股价图后，我给了埃德其中10张，对他会从中发现很诱人的模型满怀期待。果不其然，他真的发现了。数据挖掘软件也会有同样的发现。

真相大白后，埃德失望透顶，这些要是真的股票就有机会通过买卖赚大钱了。然而，他从中吸取的教训却与我想要的截然不同：埃德总结说，运用技术分析来预测抛硬币的结果是有可能的！

即使是经验丰富的专业投资者也难以理解，随机噪声能生成模型，但那不过纯粹是碰巧而已。眼前出现某个模型时，相信它有意义很容易，但要认为这不过是偶然就很难了。

抛硬币

为了说明这个原理，我有时还会在投资学课堂上用抛硬币来做另一个实验。我告诉大家我要抛若干次硬币，看看是不是有人能预测到结果。在开始抛之前，我对着教室大臂一挥，把台下学生分成左右两边。左边学生猜正面，右边学生猜反面。假设有32名学生，则其中16人肯定能猜对。如果第一次抛硬币的结果为反面，那么就是右边学生猜对了。于是，我又将右边这16名学生分成前后两组，两组各有8人。第二次抛硬币，前8个人猜正面，后8个人猜反面。如果这次结果为反面，则后8个人就连续猜对了两次。然后我再把后8个人分成两组，抛第三次硬币。将猜对的4名学生再度分组，第四次抛完硬币后只剩下2名连续猜中四次的学生。最后一次抛完硬币，我们的赢家诞生了——连续五次猜中的学生。这时我问在场的学生，谁敢打赌这名同学能再连续五次成功预测出结果。无人响应。

这就是数据挖掘。32个抛硬币结果的预测者，正如32个预测股价的系统。我根据事实发现了最成功的抛硬币结果预测者，正如发现了最成功的股价预测系统。无论是哪种情况，根据既成事实来识别成功的预测者什么都证明不了，因为始终都会有最成功的预测者，即使只是靠运气而已。

进一步说明一下，猜中次数最多的学生毫无特别之处，正如运气好的股市预测系统。如果重新抛五次硬币进行检验，那么那个成功预测了抛硬币结果的学生将表现得毫无技巧可言。同样，用全新数据来检验的话，成功的股票预测系统也会变得毫无作用。

然而，如果测试抛硬币和无用的股市预测系统次数足够多，肯定有人能成功通过两轮测试。如果我班上的学生再多一点，确切地说有1024人，我就能发现有32名学生可以连续五次正确预测结果。如果这32名学生再多猜5次，其中一人又会继续猜中5次，也就是共计猜中10次。尽管通过接连两次测试，这名获胜的学生在预测抛硬币方面还是没有什么特异功能。对1 024名进行预测的学生进行数据挖掘得到10次成功的结果，并不比数据挖掘32名学生得到5次成功的结果更加令人信服。

股票市场系统也是如此。数据挖掘软件可以搜查出数百万、数十亿，甚至数万亿潜在的股票预测系统，并从中识别出非常成功的系统。如果将这些系统重新测试一遍，一部分还是会很成功，即使它们和那名获胜的学生没什么两样，都是靠运气而已。不论无价值的股市预测系统反复测试多少次，如果测试的次数足够多，总有一部分会通过。

讨论几个专业人士使用和推荐的交易法则将具有很大的启发作用，我们可以从中意识到被系统愚弄是多么容易，这些系统能清楚地解释过去，但对未来的预测却一塌糊涂。

《每周华尔街》的十项技术指标

30多年来，《每周华尔街》（*Wall Street Week*）一直是美国公共广播电视公司（PBS）很受欢迎的电视节目之一。该节目主持人为路易斯·鲁凯泽，播放时间为纽约证券交易所收盘后的每星期五晚上。节目一开始，鲁凯泽先简要叙述当周华尔街的交易情况，以老掉牙的笑话和引发嘘声的双关语点燃全场。有观众写信来咨询能否投资一个男用假发生产厂家，鲁凯泽回答说：“如果你的钱像头发一样，明天就会消失，我们会尽力让它增长。我们提供给你光秃秃的事实，让你把自己的投资打理得像假发一样茂盛。”

多年来，该节目都设有专题环节对10个技术指标进行最新解读，这10个指标是股票经纪人兼投资顾问罗伯特·努洛克开发的，他也是该节目的常驻专家。我订阅了好几年他的新闻简讯，想看看随着时间推移，在他的那些提出又消失的模型中，这些指标是如何发展的。

他所选的10个指标首次出现在1972年，之后的17年共修订了至少五次。努洛克对这些指标的解读一般都是逆势而为：股市强劲时，他建议卖出；股市疲软时，他又建议买入。这通常是很好的建议，但认为存在可靠的预测指标表明股价“何时”上涨或下跌，却是不明智的想法。

努洛克的指标很有意思，因为它是技术派人士观察的大量指标的代表，此外，这些指标的发展演变也很有启发性。接下来我会讨论其中三个。

市场广度

很多分析师都使用涨跌指数（advance-decline index）来监测上涨股票数量和下跌股票数量的比率。与道琼斯工业指数、标准普尔500指数和其他股市指数不同，这种对比方法不关注股票价格或涨跌幅度——出现25美分的变动和出现5美元的变动的效果是一样的。这就是该方法被称为“市场广度”（market breadth）的原因。

《每周华尔街》最先使用了涨跌指数，该指数是每天上涨股票数量与50%持平股票数量之和与全部股票交易数量之比的五周平均值。如果该指数的值为0.5，则表示股市为混合状态，上涨股票的数量与下跌股票的数量相等。努洛克一如既往地反其道而行之，认为这一指数的数值高（大多数股票在上涨）是卖出信号，数值低（大多数股票在下跌）是买入信号。

努洛克的具体临界值取决于市场是牛还是熊，如表10.1所示。牛市和熊市的区别让人没有把握，因为我们没有客观的方法判断在何时、在哪个点位是牛市或是熊市。几年后，他将30%~40%的熊市买入信号修订为36%~40%，后来，该指数被剔除了。

表10.1 1972年努洛克的涨跌幅指数

	买入信号	中立信号	卖出信号
牛市	43%~45%	45%~57%	57%~60%
熊市	38%~40%	40%~50%	50%~54%

对技术指标的解读总是模棱两可的，这也生动地体现在路易斯·鲁凯泽对努洛克的涨跌指数的解读跟努洛克自己的解读完全相反。以下为鲁凯泽的描述（含有他的招牌式双关语）：

> 实际上，每名技术人员都使用这种方法的一些变体，该方法合计上涨股票和下跌股票的总数，得出的结果用来反映股市的“广度”——“广度”不佳会发出不好的信号。如果平均数上升但广度下降，技术人员会认为股市已经开始变差。

有些技术人员会得出这样的结论，但与鲁凯泽相反，努洛克没有这么做。大多数股票下跌时，努洛克给出的是买入信号。

努洛克10个指标的第一次修订，加入了第二个涨跌指数：上涨股票和下跌股票数目差额的十日平均数。数值小于-1 000（表示大多数股票下跌）为牛市，大于+1 000（表明大多数股票上涨）则为熊市。这一指标在第二版修订中仍得以保留，但是对其的解读变得更加复杂。牛市信号变为“该指数从低于+1 000点升至峰值，再从峰值下跌1 000点”。

对于其他技术指标，努洛克显然是生搬硬造出与过去数据吻合的模型的。当由数据挖掘得来的模型对新的数据不起作用时，他就改变模型。当由数据挖掘得来的模型无效时，他就再造一个新模型。捏造模型—进行测试—改变模型，如此循环。

低价活动比率

《每周华尔街》另一个有趣的指标为低价活动比率，类似很多技术人员追踪的那些指标。该指标是《巴伦周刊》上低价股票指数交易量的周比率与道琼斯工业指数的交易量之比。其思路是，低价股票比

道琼斯工业平均指数的蓝筹股更具有投机性，因此这类股票的交易量相对上涨能表明投机活动更加活跃。

该指标的比值低是买入信号，比值高是卖出信号。《每周华尔街》最初使用的数值如表10.2—表10.5所示。

表10.2 低价活动比率的原始版本

买入信号	中立信号	卖出信号
5%~8%	8%~10%	10%~15%

表10.3 低价活动比率的第一次修订版

买入信号	中立信号	卖出信号
5%~7%	7%~12%	大于 12%

表10.4 低价活动比率的第二次修订版

买入信号	中立信号	卖出信号
小于 12%	12%~18%	大于 18%

表10.5 低价活动比率的第三次修订版

买入信号	中立信号	卖出信号
小于4%	4%~8%	大于8%

捏造模型—进行测试—改变模型，如此循环。

咨询服务情绪

《每周华尔街》在第一次修订版中引入了另一个有意思的技术指标：咨询服务情绪（advisory service sentiment）。各种各样的投资服务都在向投资者推销自己的建议。基本上，这些自诩专家的人都看涨股市，尽管看涨部分因市场而异。咨询服务机构Investor’s Intelligence订阅了其他主流服务商的服务，同时还为自己的订阅者提供熊市部分的每周列表。有趣的是，该机构（和努洛克）认为，咨询服务机构越是悲观，股市就越有可能上涨。

努洛克解释称：“咨询服务情绪过度一边倒时就被看作相反的指数，因为服务机构容易随大溜，而不是预测趋势变化。”另一种解释是，投资顾问都是错多对少的业余人士。如果他们真那么聪明，为什么自己不利用这些建议，而是将其推销出去？努洛克的情绪指数最初显示，如果看跌比率低于15%则卖出，如果牛市的看跌比率高于30%或熊市的看跌比率高于60%则买入。不过，这些数值后来完全改了，如果看跌比率低于25%则卖出，高于42%则买入。再后来，临界值又变得非常精确：35.7%和52.4%，但也具有很大的误导性。这就是数据挖掘的实质。捏造模型—进行测试—改变模型，如此循环。

推特，推特

2011年，一个研究小组报告称，针对近1 000万发布于2008年2月至12月的推特进行数据挖掘分析后发现，“calm”（平静）一词的使用频次上升预示道琼斯工业平均指数随后连续六天上涨。这可不是我瞎编的。这类研究容易引起争议，获得全世界竞相报道，即使所用数据完全不可靠，得出的结论根本不合理。

发推特的用户并不是随机人口样本，更算不上投资者。很多人都没有推特账号；有些人有账号，但不怎么登录使用；有些人有账号，也会登录使用，但很少发布内容（推特估计，约一半登录用户都是光看不发，称不上推客）。有些发推特的并不是真正的人，而是自动机器人。

甚至连该报告的第一作者也承认，他对所观察到的模式不做任何解释，尤其是因为很多推特的发布者都是青少年和非美国本土人士。分享自己早餐吃了什么的小姑娘，或者一个对拜仁慕尼黑进球进行评论的德国人，又怎么会影响道琼斯工业平均指数呢？肯定不会，但它们之间可能存在统计学相关系数。

我们从这些“得州神枪手”研究人员观察的7种指标看起：积极和消极情绪的对比评估以及6种情绪状态（平静、警惕、稳妥、活力、好心和快乐）。他们还考虑了未来不同天数与道琼斯工业指数的相关系数。此外，该研究建议在选取这7种指标的术语表达时可进行一些灵活处理。最后，为什么他们要选择2008年2月到12月这段时间？1月呢？为什么2011年做的研究要采用2008年的数据？所发现的模型只存在于这段特定的时期吗？

我还想问该研究报告的作者一个问题：他们基于自己的报告结果进行股票交易了吗？

技术大师

华尔街有个说法：问题不在于股价图，只在于制图人。这证实了要找到有用的信号是多么困难。尽管如此，技术分析师还是可以在各大经纪公司谋得职位，很多人还提供着自己的咨询服务。每隔一段时间就有媒体详细报道某名技术分析师做出了惊人的精准预测，将其抬举到金融大师的地位，于是狂热的追随者竞相向其寻求建议。

艾略特波浪理论（Elliot wave theory）便是这样一个例子。13世纪的意大利人莱昂纳多·斐波那契研究了序列1、1、2、3、5、8、13、21、34、55、89……现在被称为斐波那契数列。这个数列从第3项开始，每一项都等于前两项之和。树枝生长的数量遵循斐波那契数列，雄蜂的家谱和贝壳螺旋生长的直径也是如此，音乐节奏也符合斐波那契数列，埃及人还运用这一数列设计了吉萨大金字塔。

拉尔夫·纳尔逊·艾略特将斐波那契数列应用于股价预测，创造了艾略特波浪理论。艾略特是一名会计，对数字很痴迷。对股价涨跌做了一番研究后，他得出的结论是，这些涨跌是数次交叠波浪作用的复杂结果：

持续数百年的大波浪

持续数十年的超级循环

持续数年的规律循环

持续数月的初期波浪

持续数周或数月的中期波浪

持续数周的小波浪

持续数天的细波浪

持续数小时的微波浪

持续数分钟的次微波浪

艾略特骄傲地宣称：“因为人类都遵从有节奏的活动，与其活动相关的计算也会延伸至遥远的未来，虽然这种做法的正当性和确定性在此之前是想象不到的。”要拿到市场技术分析师协会（Market Technicians Association）认可的特许市场技术分析师（Chartered Market Technician）资格，就必须通过包含艾略特波浪理论等相关问题的考试。

没错，该波浪理论非常复杂，但这就是其吸引力所在。那些醉心于数学计算的人，都被这个理论的优美复杂所吸引。它复杂到可以灵活适应任何数据集，即便是随机的抛硬币。这个数据挖掘系统简直完美无缺，现在加上计算机的辅助，就更容易操作了。

在事情发生之前，波浪理论经常会出现不一致或错误的情况。而在事情发生以后，波浪理论的狂热者总能得出一种波浪形解释，虽然不同的狂热者通常有不同的解释。

20世纪80年代，罗伯特·普莱切特成为最著名的艾略特波浪理论门徒。1986年3月，《今日美国》报刊称普莱切特为“华尔街最炙手可热的大师”，因为他在1985年9月预测的牛市成真了。同一篇文章中还提醒阅读他的预测信息的读者，道琼斯工业指数会在1988年升至3 600~3 700点。然而1988年指数的最高值是2 184点。1987年10月，普莱切特放言：“最糟糕的情况是跌至2 295点。”但仅仅几天后，道琼斯工业指数就暴跌至1739点。1993年，《华尔街日报》发表了头版文章，题为“罗伯特·普莱切特的道琼斯3 600点成真——只是晚了6年而已”（Robert Prechter sees his 3 600 on the Dow—But 6 years late）。正如他所预测的，道琼斯工业指数升到了3 600点，但迟到了6年时间。普莱切特应该早点听到一位著名投资者给我的建议，关于预测股价是多么困难：“如果给出数字，那就不要说明日期。”

“四傻”策略

1996年，美国投资家戴维·加纳德和汤姆·加纳德两兄弟发行了《傻瓜投资指南：傻瓜如何打败华尔街的智者，你又能怎么做》（*The Motley Fool Investment Guide: How the Fools Beat Wall Street’s Wise Men and How You Can Too*）一书，闯出了一片小天地。他们通过研究1973—1993年的数据，创造出自己所谓的“四傻”（The Foolish Four）策略。如果早知道这个策略，他们每年的平均回报率将达25%。

这两兄弟还提醒读者，这个系统“会保证支持者在未来也收获25%的年回报率，正如它过去表现的那样”。

以下就是他们的“四傻”策略：

1. 在当年年初，计算道琼斯工业平均指数中30只股票的股息收益率。例如，2016年12月31日，可口可乐股价为每股41.46美元，每股的年股息为1.48美元。可口可乐的股息收益率为1.12美元/41.46美元=0.0357，即3.57%。

2. 选择10只股息收益率最高的股票。

3. 从这10只股票中，选出5只每股股价最低的股票。

4. 从这5只股票中，剔除价格最低者。

5. 将40%的财富投入价格第二低的股票。

6. 将20%的财富分别投入其他3只股票。

我希望当你看到这里的时候脑子里会响起一个声音，大喊：“数据挖掘！”

“四傻”策略错综复杂，而且一点都不合理。实际上，我们知道它就是广泛挖掘数据的产物。

有两名金融学教授也持相同看法，因此他们用1949—1972年的数据测试了“四傻”策略，正是加纳德兄弟进行数据挖掘的年份之前的时期。回顾历史，两名教授发现这确实是个失败的策略；预测未来，加纳德兄弟也发现这确实是个失败的策略。1997年，就在他们推出“四傻”策略的第二年，加纳德兄弟稍稍改进了他们的系统，将其重新命名为UV4。为什么要换名？因为历史证明UV4真的比“四傻”好一点。捏造模型—进行测试—改变模型，如此循环。

2000年，加纳德兄弟停止兜售“四傻”策略和UV4。这不足为奇，如果你的第一反应也是——数据挖掘！

为乐趣和盈利投资的黑匣子

技术分析的价值似乎不言而喻，任何明眼人都能在股价中看出清晰的模型。然而，搜遍数据寻找模型只能证明研究人员具有坚韧不拔的毅力。请记住，“如果你拷问数据的时间足够长，它就一定会坦白”。

我们从中可以吸取两个教训。第一，我们应该认清如下诱惑：寻找模型并认为我们找到的模型一定具有意义——尽管所有证据都是反证。第二，我们应该奋力抵抗易受影响的天性。

以前的技术分析都是靠人仔细审查股价图，然后寻找模型。现在，编程好的计算机可以找出复杂细致到目测难以检查出的模型，两者可能同样毫无作用。这也就是为什么要记住这两个教训。人们很容易认为模型一定有意义，抵抗住这种想法的诱惑实属明智之举。

---

[1] 愚人金，指看上去像黄金，实则一文不值的矿物。——译者注

第11章

完胜股市（下）

如今，技术分析师都被称为金融工程师。我们不仅过度欣赏计算机的能力，也过于钦佩使用计算机而不使用笔和图表的金融工程师。

金融工程师不思考他们发现的模型是否合理。他们的准则是：“给我看数据就行。”其实，虽然很多金融工程师是物理学或数学博士，但其对经济学和金融学的了解过于肤浅。不过，这并没有对他们造成困扰，要说有什么影响的话，那就是无知的他们更有勇气从最不可能的地方寻找模型。

从使用铅笔的技术分析师转到使用计算机的金融工程师，对此符合逻辑的结论是要将人类彻底排除在外，数据分析的工作交给计算机做就行了。

2011年，精彩的科技杂志《连线》发表了一篇文章，全文充斥着对计算机化股票交易系统的敬畏和钦佩之情。这些黑匣子式系统被称为“算法交易者”（algorithmic traders）——由计算机根据算法来决定股票买卖，而不是人的判断。人类编写算法指导计算机，但在这之后，全靠计算机自己运行了。

有些人被唬住了。2016年，佩珀代因大学将其投资组合的10%投给了金融工程师基金，其投资总监表示：“寻找具有良好前景的公司合情合理，因为我们在日常生活中都会寻找被低估的事物，但是金融工程师策略与我们的生活毫不相干。”他认为，没有从生活中获得的智慧和常识，是支持使用计算机的论据所在。和他观点一致的大有人在。如今，美国股票交易的近1/3是依靠黑匣子式的投资算法完成的。

这些系统有的追踪股价走势，有的观察经济数据和非经济数据、剖析新闻线索。它们全都在寻找模型。一个动量算法或许会注意到，当某只股票的交易价格连续五天较高时，其第六天的股价通常也会更高；一个均值回归算法或许会注意到，当某只股票的交易价格连续八

天较高时，则其第九天的交易价格通常会下降；一个配对交易算法或许会注意到，两只股票通常会同涨同跌，当其中一只上涨而另一只没上涨时就是在提示机会来了。其他算法还使用了多元回归模型。在每一种情况下，算法都是基于数据挖掘运行的，其格言是：如果它行得通，那就好好利用。

我自己会投资，也在教授投资学，因此我决定自己尝试一下数据挖掘，看看能否找出预测股价的可靠指标。运气好的话，我的数据挖掘或许能收获“知识发现”，我可以靠此赚上一笔。

股市与天气

据报道，纽约市的天气会影响美国股市，虽然其影响随着时间的推移已经减弱，因为全美乃至全世界的股票交易已经从大厅交易演变为电子下单。

海蒂·阿蒂格搜集了25座城市每日的最高气温和最低气温数据，这鼓动我想看看能否找到某些气温，用来解释标准普尔500指数每日股价的波动情况。

我最初以为每日气温在预测股价上有局限性，因为气温随季节变化而股价不是。此外，股价具有明显的上涨趋势但气温没有（至少短短几年内不会）。尽管如此，没费多少工夫，我还是用数据挖掘找到了五个气温，很好地预测了2015年的股价。

那25座城市的最高气温和最低气温为我提供了50个可能的解释变量，以它们为基础，我可以获得：50个含有一个解释变量的模型；1 225个含有两个解释变量的模型；19 600个含有三个解释变量的模型；230 300个含有四个解释变量的模型；2 118 760个含有五个解释变量

的模型。试图做一名非常投入的数据挖掘者，我将所有解释变量都推算出来了，总共是2 369 935个模型。

得到的很多模型都不错，但最好的如下：

$$Y=2361.65-3.00C+2.08M-1.85A+1.98L-3.06R$$

其中：

C = 澳大利亚科廷市，最高气温

M = 华盛顿州奥玛克市，最低气温

A = 蒙大拿州羚羊谷，最高气温

L = 蒙大拿州林肯市，最低气温

R = 怀俄明州石泉镇，最低气温

巧合的是，在第4章讨论过的科廷市和奥玛克市再次出现，不过这次最高气温和最低气温互换了。

如图11.1所示，尽管2015年下半年股市下跌，含有5个温度的模型与股价波动的吻合度很高。该模型的准确率为60%，对于预测变幻莫测的对象（如股价）来说，已经算是相当可以了。

这对“知识发现”来说又算是怎么一回事呢？有谁知道这5个小城镇的每日最高和最低气温能有助于预测股价呢？

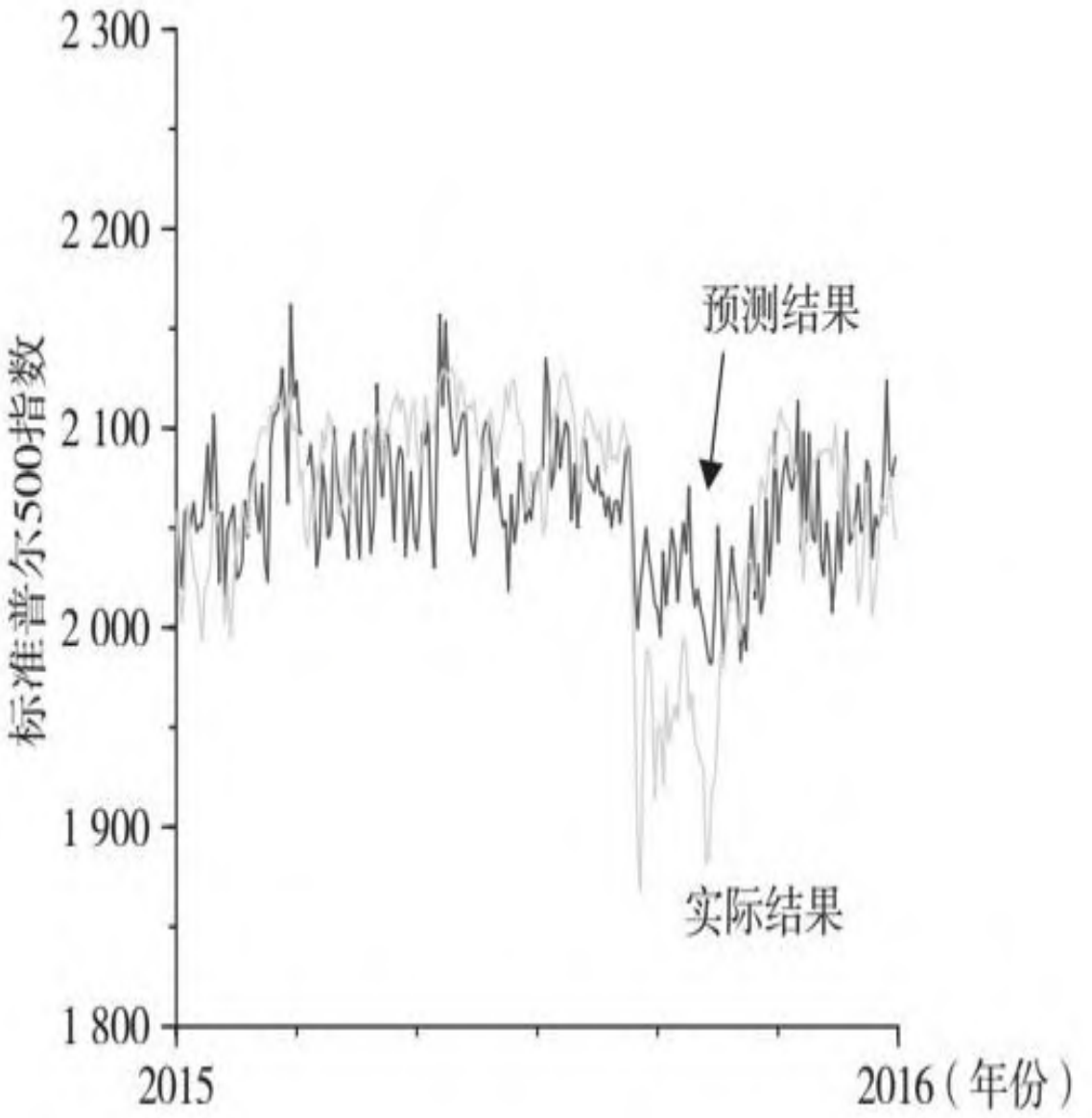


图11.1 与股价相关的“知识发现”

答案当然是它们对预测股价没有帮助。找不到合理的理由能说明标准普尔500指数与这5个城镇的最高气温和最低气温存在正相关或负相关关系，其中还有一个城市远在澳大利亚。我们能生编硬造出不切实际的说法，解释为什么每日股价取决于这些城市的消费状况，而消费状况又如何取决于这些城市的天气，但这也不过是信口雌黄而已。

先用2015年的数据推算出200多万个方程式，再从中挑出准确率最高的那一个，这就是上述模型的选择过程。由于模型建立在数据而非逻辑之上，因此我们不要指望它能较好地预测2016年的股价。如图11.2所示，2016年的预测准确率为-23%。没错，结果是个负值。当该模型预测股价将上涨或下跌时，很可能出现相反的情况。

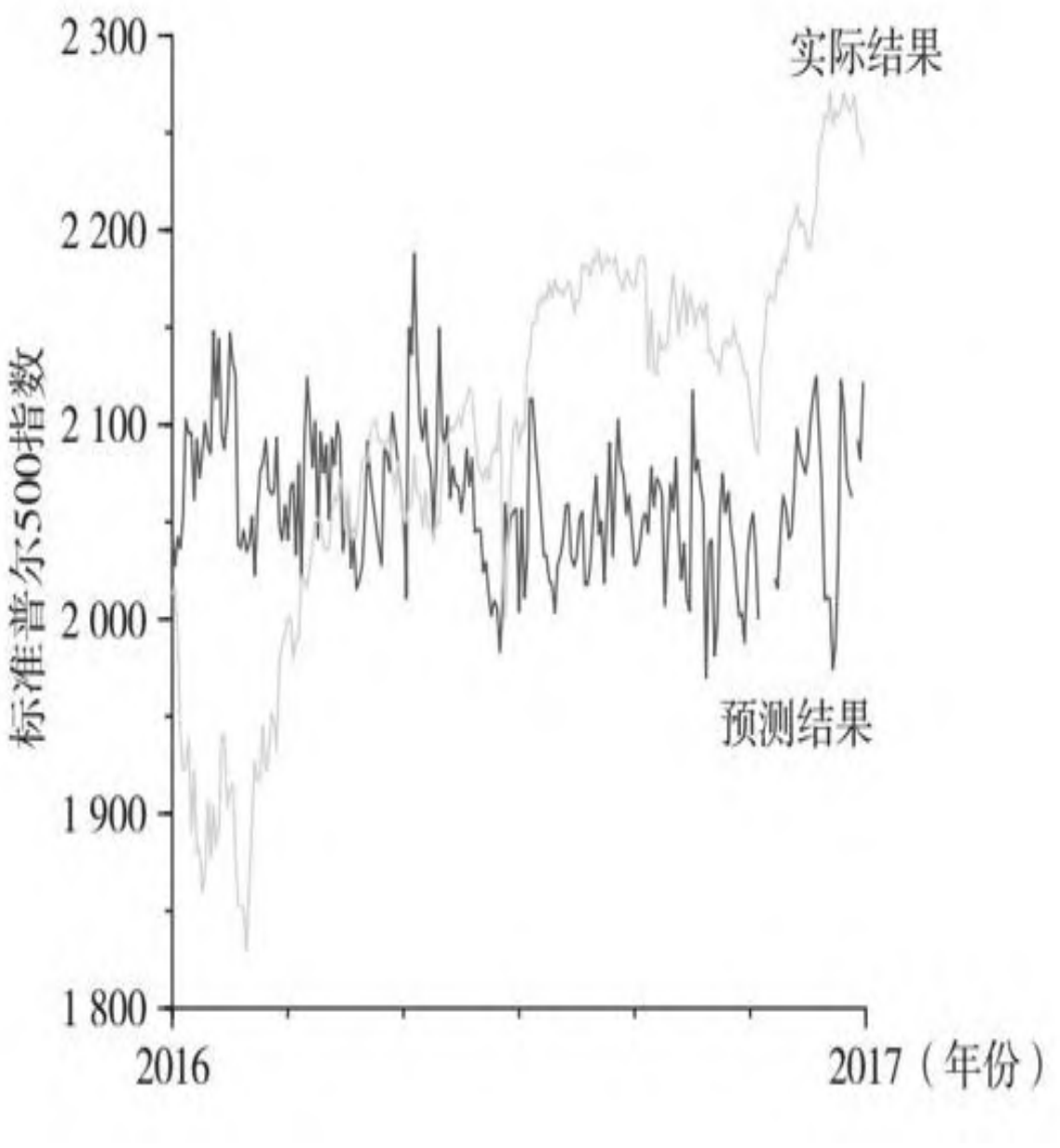


图11.2 2016—2017年的预测结果和实际结果

不断尝试

我对自己的气温模型感到失望，于是考虑用100个新的变量，尝试从1~5个解释变量的所有可能组合。现在，模型数量已接近8 000万，但是对我的数据挖掘软件来说，这个数目还是小到它能尝试每一种可能性，而无须求助主成分分析、因子分析、逐步回归法或其他有缺陷的数据规约步骤。

推算这些模型花费了数小时，所以我就止步于5个解释变量了。如果我继续推算，利用10个解释变量会得到超过17万亿个可能组合，那样的话，计算机就得花费好几天来跑数据。幸运的是，有几个5变量组合预测的股价与实际股价非常接近。最好的模型如图11.3所示，准确率高达88%，拟合数值与实际数值非常接近，实际上很难将其区分开来。

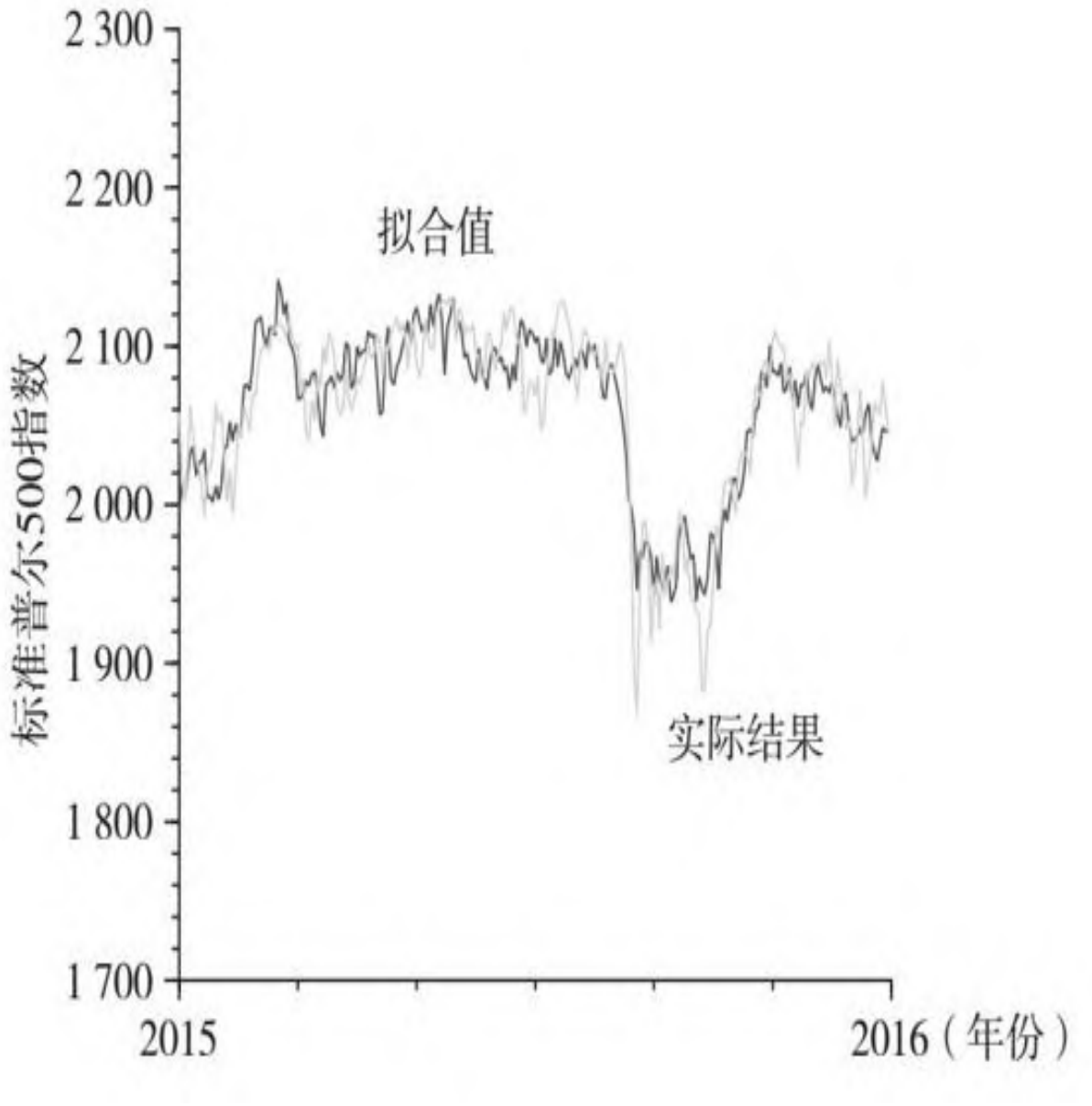


图11.3 我的5变量股价模型

我可能已经揭开了股票预测的未解之谜。你做好投资的准备了吗？

我是在2017年4月利用2015年的每日数据进行这次数据挖掘探秘的。对于含有5个气温变量的模型，我特意预留了2016年的每日数据，目的是验证我的“知识发现”。如图11.4所示，该模型对2015年的预

测结果喜人，但对2016年的预测结果则一塌糊涂，它预测股价会暴跌但实际是暴涨。具体来说，该模型对于2015年预测的准确率为88%，而2016年的准确率为-52%。这个5变量模型的预测结果与2016年标准普尔500指数的实际表现存在强负相关关系，我的模型比毫无价值更糟。

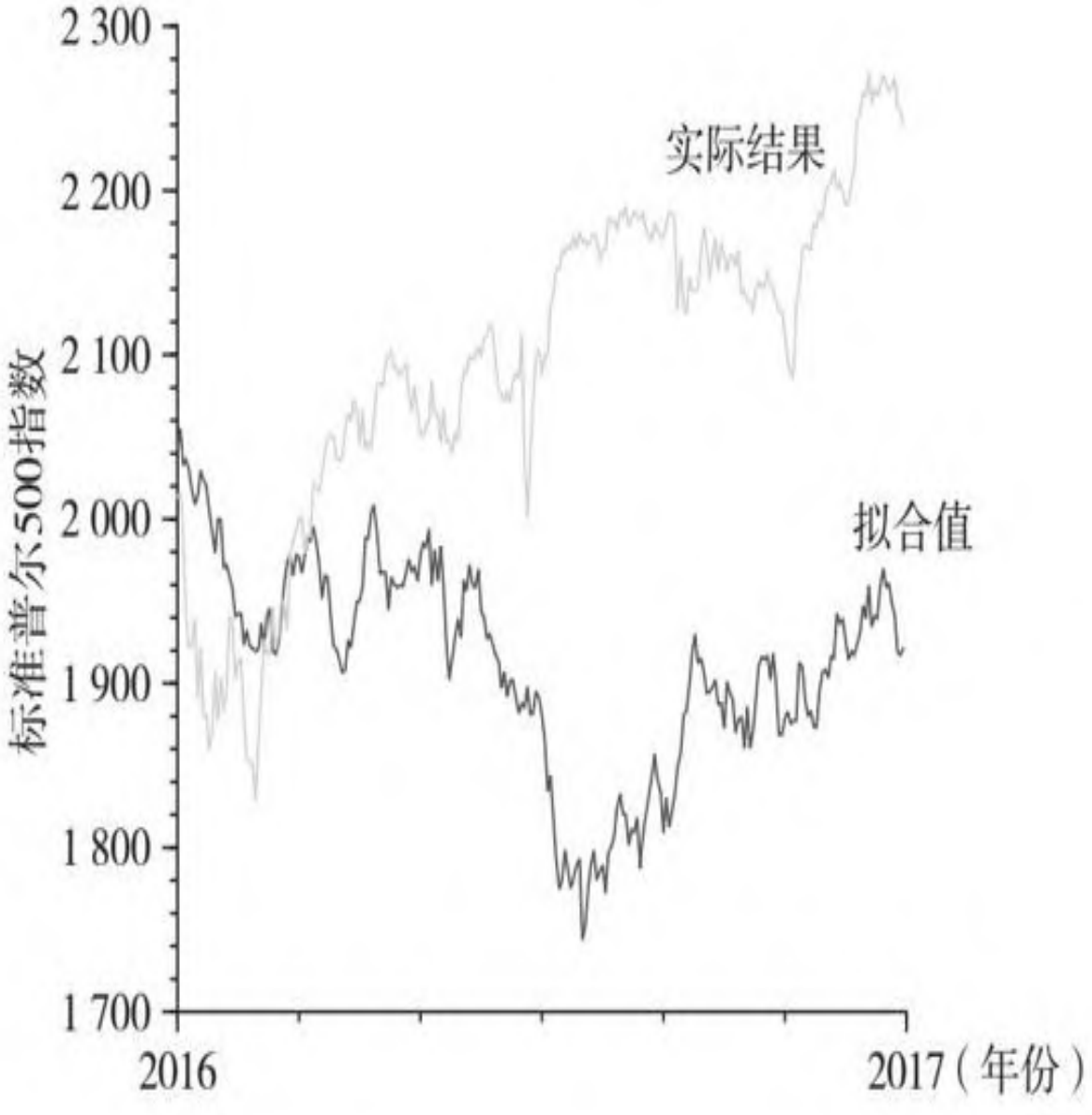


图11.4 使用全新数据后的5变量模型

这是怎么回事？在某年预测效果很好的模型，怎么在下一年的预测结果会如此不尽如人意？这就是数据挖掘的本质。选择某个模型，只是因为它与所给的数据集吻合度高，这就造成这个模型与全新数据的吻合度达不到同样的水平。若要在处理全新数据时依然有效，就必须采用合理的模型。不过，数据挖掘软件无法判断一个模型是否合理。

我通过本质上是对随机数据（如澳大利亚科廷市的最高气温）进行的数据挖掘，想说服你相信这样一个事实：它们根本不会影响标准普尔500指数。我们通过逻辑推理进一步得出结论，图11.3和图11.4所示的模型，不是实质上为随机的，而是完全随机的模型。标准普尔500指数是真实数据，但那100个可能的解释变量是我用计算机随机数字生成器生成的。

还记得我曾让学生用抛硬币的方式得出虚假的股价数据吗？每只股票的起价均为50美元，然后将25次抛硬币的结果作为当天股价变化的依据，抛出正面则股价上涨50美分，抛出反面则股价下跌50美分。我在课堂上做这样的抛硬币实验是想要学生亲眼看看，明显为随机产生的数据是如何产生了看似非随机的模式的。

这次的做法也一样，不过换成了使用计算机的随机数字生成器。我将每个变量的初始值设为50，然后让电脑抛硬币来决定变量每天的变化值。若电脑抛硬币的结果为正面，则数值上升0.50；若为反面则数值下降0.50。我用计算机为每个变量的每日变化抛了25次硬币，以便得到100个虚构变量在这两年内的每日数值，将前半部分的随机数据标记为2015年，后半部分标记为2016年。

即使100个变量都是由随机游走过程产生的，在事实发生后，还是会存在有些变量的确与标准普尔500指数存在偶然的相关系数。在五变量的所有可能性中，随机变量4、34、44、64和90的组合与2015年标准普尔500指数的相关度最高。但到了2016年，该模型就完全行不通了，因为这些都是实实在在的随机变量。

黑匣子式数据挖掘无法预测这种巨大的落差，因为它不能评估自己发现的模型是否具有逻辑基础。

预留方案

现在，可能有人会说，既然从该模型对2016年预测结果的糟糕程度就可以看出标准普尔500指数和我的随机变量之间不存在任何真正的关系，那么我们就可以利用样本外测试来区别偶然的相关系数和真正的因果关系。挖掘部分数据，寻找“知识发现”，然后通过有目的的暂时预留的数据来测试所发现的模型以验证结果。原始数据有时被称作“训练数据”，预留数据被称作“检验数据”或“验证数据”。另一种叫法为样本内数据（用以发现模型的数据）和样本外数据（用以验证模型的全新数据）。在利用气温和随机变量预测标准普尔500指数的例子中，模型是用2015年的数据推算得出，用2016年的数据进行验证的。预留出2016年的数据，正是为了这一目的。

不断询问模型是否运用全新数据验证过是一个很好的想法。大肆搜集数据以发现模型，再用相同的数据来验证模型的做法绝对没有说服力，这些数据都是为了找到模型而被掠夺来的。因此，预留验证数据来检验无中生有、生编硬造来的模型肯定不失为好方法。

然而，不知疲倦的数据挖掘可以确保某些模型与训练数据和检验数据的吻合度都很高，即便该模型根本不合理。正如有的模型肯定与原始数据吻合，有的仅仅是运气好，也能与预留数据吻合。发现同时符合原始数据和预留数据的模型，只不过是另一种数据挖掘形式。我们要找的不是符合半数数据的模型，而是符合所有数据的模型。为了符合数据而挑选的模型，无论是符合半数还是所有数据，都不能指望它与其他数据的吻合度一样高。这么做解决不了问题。

为了说明这一点，接下来看看我为了解释标准普尔500指数的波动而创造出的100个随机变量。共有100个含有一个变量的模型：随机变量1、随机变量2……对于每一个变量，我都利用2015年的每日数据，来推算出吻合度最高的模型。以随机变量1为例：

$$Y = 2113.62 - 0.5489R1$$

该模型的准确率（标准普尔500指数的预测值和实际值之间的相关系数）为28%。但我用此模型预测2016年的标准普尔500指数时，其准确率竟为-89%。该模型预测标准普尔指数会上涨，但实际上该指数下跌了，反之亦然。

我把100个可能的解释变量统统用上，反复尝试，让模型与2015年的数据吻合，再用2016年的数据验证，结果如图11.5所示。对于2015年的数据，由于它们被用以推算模型，所以准确率不可能小于0，因为该模型总能完全忽略解释变量，从而得到准确率为0。结果显示，使用样本内数据且含有1个变量的模型的平均准确率为27%。

对于预留下来用以验证模型的2016年的数据，其准确率为正值和负值的可能性相等，因为毕竟它们是与股价毫无关系的随机变量。我们预计，股价和任何随机变量之间的平均相关系数均约为0。对上述特定数据来说，样本外数据的平均准确率碰巧为-4%。

尽管如此，样本外数据的准确率还是会碰巧与某些模型存在强正相关系数，与其他模型存在强负相关系数。如图11.5右上角所示，有几个模型的2015年样本内数据和2016年样本外数据的准确率都很高。具体来说，有11个使用2015年拟合数据的模型相关系数高于0.5，其中5个在使用预留数据时的相关系数高于0.5。这五个模型都通过了样本外数据验证测试，尽管它们只是与股价完全没有关联的随机变量。

若使用更多解释变量，准确率还会上升。我又重复了一次实验，推算了4 959个可能的双变量模型。随机变量57和59的吻合度最高：

$$Y = 2100.46 + 3.4612R57 - 4.8283R90$$

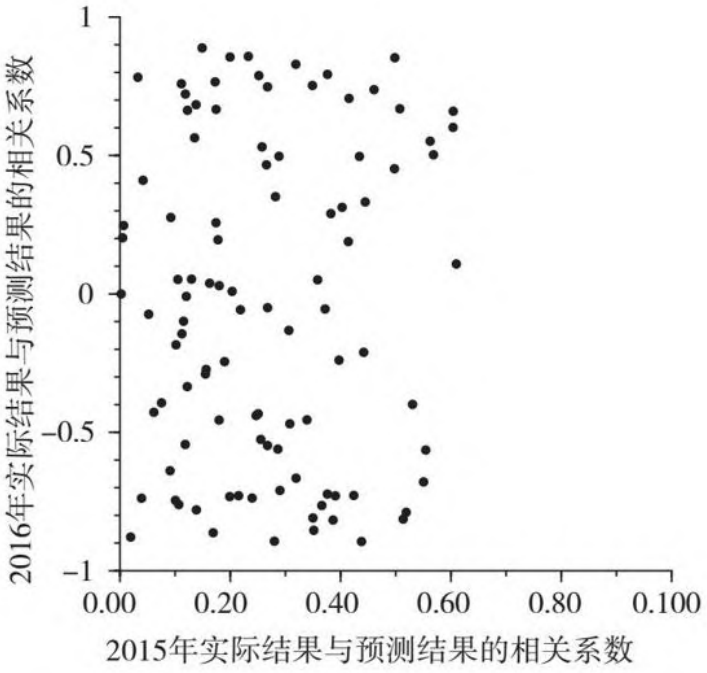


图11.5 100个单一变量模型的样本内和样本外数据吻合情况

这个模型的样本内数据准确率竟高达79%，但是样本外数据准确率为-56%。这个双变量模型与2015年的数据高度吻合，但是对2016年的预测结果却与实际值呈负相关关系。尽管有这个缺点，但更多的数据挖掘肯定会找到既符合2015年的训练数据，又符合2016年的验证数据的模型。

使用双变量的模型，2015年回测数据的平均准确率为40%，而2016年预留数据的平均准确率为-1%。图11.6为2015年和2016年准确率之间的关系。

这4 950个模型把图表变成了巨大的斑点。有很多模型（如随机变量57和90）与2015年数据吻合度高，但与2016年数据的吻合情况一塌糊涂。同时，也有很多模型与这两年的数据都非常吻合，有时，与2016年数据的吻合度甚至高于2015年。这就是偶然的本质，这些都是偶然得出的变量。

有46个模型的2015年准确率为70%，其中11个模型的2016年准确率为70%。这11个模型都通过了验证测试，但它们对预测其他年份的股价还是没有效果，如2017年。

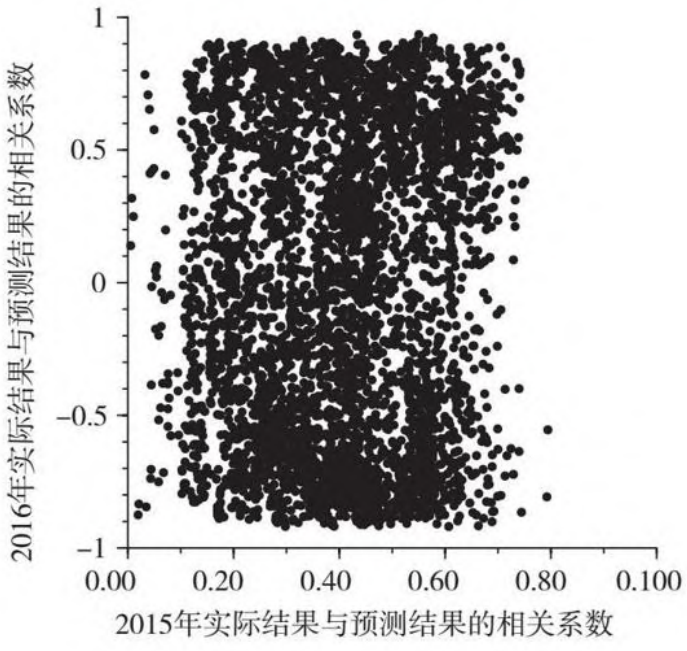


图11.6 4950个双变量模型的样本内和样本外数据吻合情况

有一个使用随机变量14和74的模型，其2015年准确率为70%，而对于2016年验证数据的准确率竟然达到88%！如果我们对此不够了解，可能还以为自己取得了什么重大发现。然而事实是，人们总能找到同时符合样本内和样本外数据的模型，即使这些数据都不过是随机噪声。

对含有更多解释变量的模型来说，情况则有过之而无不及。若变量增加，可能的模型数量会呈现爆炸式增长，找到符合训练数据和预留数据的模型的确定性也会更大。含三个变量的可能模型有161 700个，含四个变量的可能模型有3 921 225个，含5个变量的可能模型有75 287 520个。

随着可能性越来越多，图表会密密麻麻地布满圆点（如图11.6所示）。但是，原则仍然成立。从中肯定能找到很多模型同时与2015年和2016年的数据吻合。

例如，最佳的五变量模型的2015年样本内数据准确率为88%，2016年样本外数据准确率为-52%。然而，有些5变量的模型碰巧与2015年的吻合度高，有些是与2016年的吻合度高，还有些在这两年的吻合度都很高。我的数据挖掘软件识别了11 201个5变量模型，这些模型与2015年标准普尔500指数的实际值和预测值之间的相关系数至少为85%，其中有109个模型的2016年准确率高于85%，49个模型的2016年准确率高于90%。如果我再尝试更多变量，我的数据挖掘软件肯定会发现对两年的准确率都高于90%，甚至高于95%的模型。

这不是“知识发现”，而是偶然发现。

如果我们搜遍股价数据就是为了找到不合理的系统以完胜股市，几乎可以肯定的是，我们会因此更穷。

真正的数据挖掘

Quantopian（众包型量化投资平台）网站为想要成为投资大神的人提供编写其交易算法的空间，再用历史数据回测，看看这些算法会带来多大回报。听起来很合理。不过，我们知道，数据挖掘总能找到在挖掘期内获利的算法。我们还知道，没有逻辑基础的算法在使用全新数据时的表现通常会让人大失所望，无论它们的回测结果有多好。

Quantopian平台有意思的一点在于，尽管这些算法的细节没有公开，但任何人都可自主采用过去任何时间段的数据进行验证。此外，

每个算法都有时间标记，显示该算法的最后版本是于何时发表在Quantopian平台上的。

有外部团队检验了该平台将近1 000个股票交易算法，这些算法均发表于2015年1月1日到6月30日。每个算法都利用2010年至发表前的数据进行回测（训练期），然后再用发表后到2015年12月31日的全新数据进行检验（验证期）。结果发现，训练期和验证期的收益之间存在很小但是统计学意义显著的负相关关系。大写的尴尬！

趋同交易

卖空股票是指卖出从其他投资者手中借来的股票。有时候还必须回购股票（希望是以更低的股价）还给投资者。现在，假设你能以90美元的价格买入一只股票，并以100美元的价格卖空同一只股票。如果这两个股价趋同于110美元，那么你的第一只股票的收益为20美元（90美元买入，后以110美元卖出），第二只股票损失了10美元（100美元卖出，后以110美元回购）。因此，你的净收益为10美元，即两个初始股价之差。

相反，如果这两个股价趋同于80美元，那么你的第一只股票损失了10美元（90美元买入，后以80美元卖出），第二只股票收益为20美元（100美元卖出，后以80美元回购）。因此，你的净收益为10美元。

这就是所谓的“趋同交易”，因为你赌的不是两只股票的涨跌，而是股价会趋向一个共同的价格。

荷兰皇家壳牌集团

1907年，荷兰皇家石油公司（总部位于荷兰）和英国壳牌运输和贸易公司（总部位于英国）合并经营，联手对抗约翰·D.洛克菲勒的标准石油公司——全球最大的炼油公司。荷兰皇家石油将专注于生产，英国壳牌则专注于销售，合并经营之后，这两家公司或许还能存活下去。

让人好奇的是，根据双方协议，荷兰皇家石油和英国壳牌均保留各自目前的股东，这两只股票也继续在各家证券交易所进行交易，不过，所有收益和支出都合并到母公司荷兰皇家壳牌集团（荷兰皇家石油占股60%，英国壳牌占股40%）。集团全部收入的60%归荷兰皇家石油，40%归英国壳牌；集团派发的全部股息的60%归荷兰皇家石油的股东，40%归英国壳牌的股东；如果集团被出售，收入的60%归荷兰皇家石油的股东，40%归英国壳牌的股东。

无论英国壳牌的价值为多少，荷兰皇家石油的价值都要比它高出50%。如果股市对两者股票的估值正确，荷兰皇家石油的股票市值应该总会比英国壳牌的高出50%。但事实并非如此！

图11.7为1957年3月13日（两只股票首次在纽约证券交易所进行交易）到2005年7月19日（两家公司完全合并，股票停止单独交易），荷兰皇家石油与英国壳牌的股票市值比率。

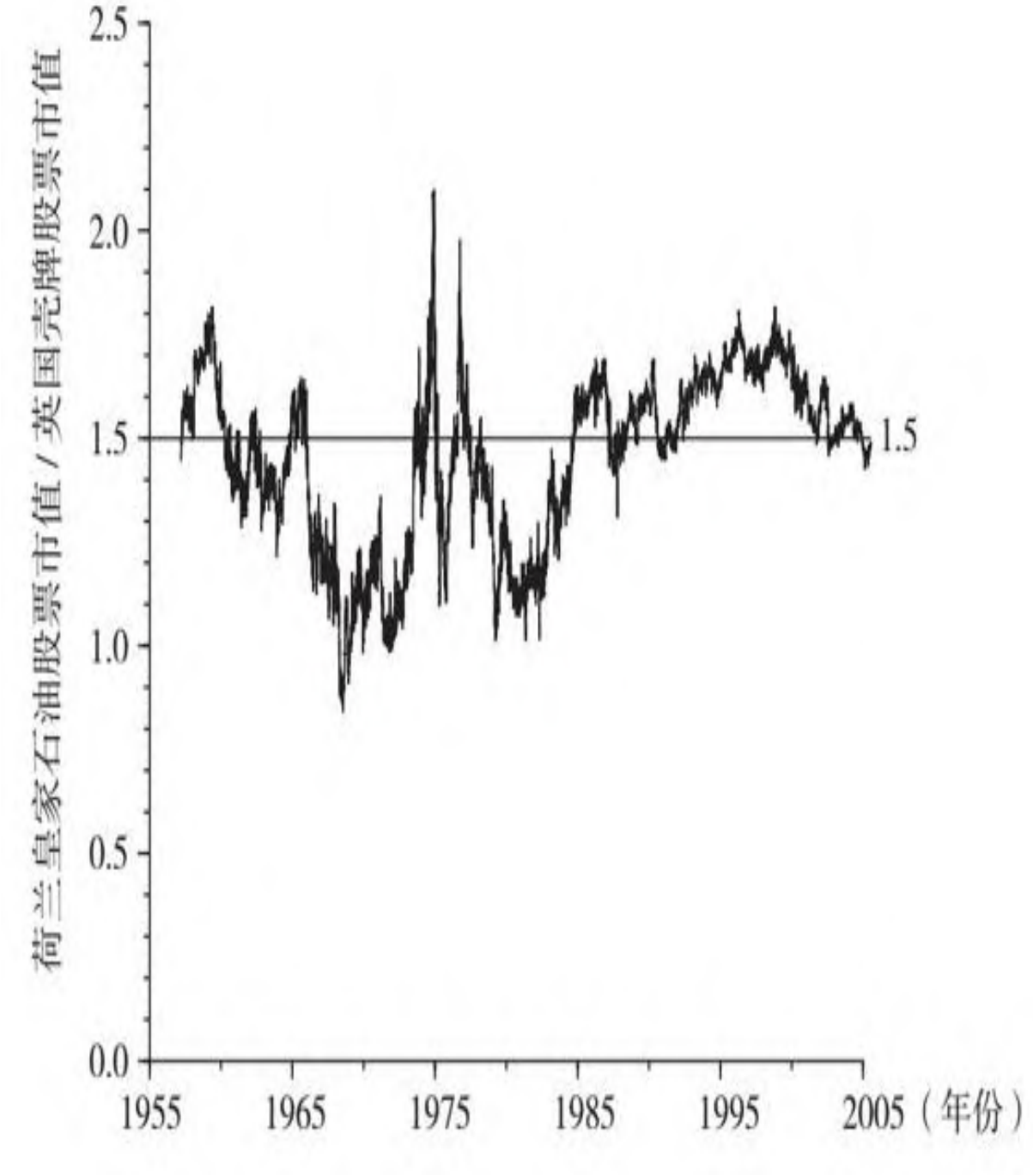


图11.7 荷兰皇家石油与英国壳牌

荷兰皇家石油的股价几乎从未刚好比英国壳牌的高50%，有时会高40%，有时会低30%。从整个时间段来看，两者的实际市值比与正确的

理论市值比（1.5）之间的百分差有46%的时间高于10%，有18%的时间高于20%。

这种情况非常适合趋同交易。当荷兰皇家石油的交易价与英国壳牌的交易价之比高于1.5时，投资者可以买入英国壳牌，卖空荷兰皇家石油，赌这个溢价会消失。

1997年，美国长期资本管理公司就这么做了，当时的溢价从8%涨到10%。该公司买入英国壳牌价值11.5亿美元的股票，卖空荷兰皇家石油价值11.5亿美元的股票，坐等市场修正股价。该公司拥有全明星阵容的管理团队，包括两名荣获1997年诺贝尔奖的金融学教授，这是很聪明的一招，其依据是具有说服力的逻辑，而不仅是偶然发现且毫无意义的统计学模式。市值比率最终应该达到1.5，长期资本管理公司会从这次机智的对冲交易中获利。

然而，正如凯恩斯在大萧条期间观察所得：

> 这套长期理论在误导当前事物。从长期来看，我们都难逃一死。经济学家为自己设置了过于容易、过于无用的任务，如果遇上狂风暴雨，他们唯一能告诉我们的只有：暴风雨过后，大海会恢复平静。

凯恩斯嘲讽的观点是：从长远来看，经济发展会趋于平静，想找工作的人总会找到工作的。他认为，短期的经济衰退风暴比假想的长期平静更加重要，或许没人能够看到那个长期的到来。股市也是如此。从长期来看可获利的趋同交易，从短期来看却会带来灾难性后果。

1998年初，长期资本管理公司的净价值接近50亿美元。同年8月，一场始料未及的风暴来袭。俄罗斯未能偿还债务，并且察觉到整个金融市场的度量风险都在提高。长期资本管理公司在很多不同市场都下了赌注，猜测大部分的风险溢价将下降。但自俄罗斯未能偿还债务

后，到处都出现了风险溢价上涨，长期资本管理公司遇到了麻烦，而且是很大的麻烦。

该公司争论道，一切都是时间的问题，等时候到了，金融市场就会恢复到正常水平——暴风雨终将过去，大海会再次平静——但是，该公司已经等不起了。它下的大赌注和借款之间的杠杆过高——若能偿清就尚好，否则会导致灾难性的后果。8月21日，该公司损失了5.5亿美元，整个月下来共损失了21亿美元，将近其净价值的一半。

长期资本管理公司努力筹集更多资金，期待熬过这次风暴，但贷方已成惊弓之鸟，不愿再给该公司放贷，还想着讨回先前的借款。

凯恩斯不仅是大师级经济学家，还是传奇般的投资家。他曾告诫：“市场保持非理性状态的时间，可比你保持有偿还能力的时间更长。”可能市场对俄罗斯未偿还债务的反应过度了，也可能长期资本管理公司最终会转亏为盈。但是，它保持有偿还能力的时间，不足以让它见证这一刻的到来。

该公司不得不对其持有的荷兰皇家壳牌集团股票进行平仓处理，当时，荷兰皇家石油的溢价不降反升，超过了20%。长期资本管理公司在这笔交易中损失了1.5亿美元。

同年9月23日，沃伦·巴菲特给该公司传真了一封信件，提出要以2.5亿美元收购该公司，约为其年初净价值的5%。这次出价是“要卖就卖，不卖拉倒”型，截止时间为当天中午12点30分，也就是传真后的一个小时。该公司最后没有接受出价，开始为自己准备“后事”。

纽约联邦储备银行担心长期资本管理公司未偿还债务会引起多米诺效应，触发全球金融危机。于是，纽约联邦储备银行携手长期资本管理公司的债权人接管该公司并清算其资产。债权人收回了贷款，公司创始合伙人损失了10.9亿美元，其他投资者则花大价钱上了一课，了解到了杠杆的力量。

注意看图11.7，2005年，溢价最终的确消失了，当时荷兰皇家石油与英国壳牌合并，荷兰皇家石油的股东拿到了合并公司60%的股份，英国壳牌的股东则拿到了其余的40%。荷兰皇家壳牌集团这次的交易确实是明智之举，合情合理且最后也成功获利。不幸的是，长期资本管理公司的那些交易就欠缺考虑，最后迫使自己不得不过早清算了荷兰皇家壳牌集团的股票。

股市价格有时稀奇古怪，荷兰皇家壳牌集团的错误定价就是非常有说服力的例证。无论英国壳牌股票的“正确”价值是多少，荷兰皇家石油总会多出50%，然而股市价格时高时低，为有利可图的趋同交易创造了机会。然而，这个例子还说明，即使是由行业顶级人士正确无误地完成的趋同交易，也是有风险的，因为趋同所需时间可能比预期更长。而没有逻辑基础的趋同交易就更加危机四伏了。

黄金白银比率

20世纪80年代，大名鼎鼎的投资顾问公司Hume & Associates（休姆联合公司）制作出《超级投资者档案》（*The Superinvestor Files*），向全美宣传，普通投资者靠它就能获得非常可观的利润。订阅用户每月会收到一份印刷精良的50页册子，每本25美元，外加合计2.5美元的邮费和处理费。

回想起来，本应显而易见的是，如果这些策略像广告宣传的那样有赚头，该公司利用这些策略可以比推销册子挣更多的钱。然而，容易受骗、贪婪的投资者忽视了这一点，反而希望花上25美元和合计2.5美元的邮费和处理费，就买到成为百万富翁的秘诀。

其中一个超级投资者策略是基于黄金白银比率（gold/silver ratio,GSR），即每盎司黄金和白银的价格比率。1985年，黄金的均价为317.26美元，白银的均价为5.88美元，则GSR为317.26美元/5.88美元 = 54，也就是说，每盎司黄金的价格是白银的54倍。

1986年，休姆写道：

GSR在过去七八年内的波动幅度较大，1980年低至19：1，1982年高达52：1，到了1985年又升至55：1。但是，你也能清晰地看到，它总是（注意，总是）会回到34：1~38：1的范围。

图11.8证明了1970—1985年的GSR在34~38的范围内波动。



图11.8 1970—1985年的GSR

GSR策略是在GSR处于异常高的状态时，卖金买银；处于异常低的状态时，则买金卖银。采用期货合约使这些交易产生了巨大杠杆，有可能获得暴利。这是一次趋同交易，因为投资者赌的不是金银价格的涨跌，而是两者比率会趋同于其历史比率。

一盎司黄金的价格应是白银的36倍，其中原因毫无规律可循。黄金和白银不像鸡蛋，可以买一打或半打，如果价格有偏差，消费者会买更便宜的鸡蛋；也不像玉米、大豆，如果玉米相对大豆的价格上涨，农民就种更多的玉米。

最终显示，1983年GSR上涨到38后，直至2011年，也就是28年后，才回落。如图11.9所示，GSR保持在34~38的范围内只是短暂出现的巧合，不是这个超级策略的基础。期货合约可能成倍扩大损失和收益，而在1985年下注GSR则会产生灾难性的后果。



图11.9 1970—2017年的GSR

其他趋同交易

早期趋同交易只依据简单的模式进行，如GSR，在价格图上一目了然。现代计算机能搜遍大量数据库，寻找更不易察觉和复杂的趋同交易。如果两个价格之间的相关系数为0.9，价格开始往不同方向移动，交易算法可能就会判断这一历史关系会重现。

即使计算机发现了模型，GSR交易中也存在同样的问题，即无理论支撑的数据存在隐患。趋同交易需要合理，因为如果找不到所发现模型的根本原因，该模型出现的偏离也就没有理由自我修正。统计学相关系数可能是偶然出现的模型，转瞬即逝。

荷兰皇家壳牌集团的趋同交易就有很合理的基础，但是长期资本投资公司破产的原因是，将大量赌资压在了根本原因不具有说服力的相关系数上。例如，法国和德国各种利率之间的关系与风险溢价。一名经理后来痛惜道：“我们公司的学术大师在加盟时毫无交易经验，就这样开始建模。鉴于自己做出的假设，他们的交易看似不错，但常常连简单的可信度检验都无法通过。”

讽刺的是，观察荷兰皇家石油和英国壳牌在任何一两年内的每日股价，黑匣子式的交易算法都不会识别出它们股价的比率应该是1.5。它会漏掉其中一次合理的趋同交易。

图11.10所示的是一个更近时期的趋同交易机会。在2015年和2016年大部分时间，这两只股票的股价比率波动范围的平均值为0.76。虽然价格比率相对于0.76时高时低，但总是会回到明显的均衡值。

2016年8月25日，该价格比率突破1，表明这是卖出一股买入另一股的好时机。可惜的是，如图11.11所示，该比率并没有回到自然均衡值0.76，而是继续上升，翻了一倍多。2016年11月3日，其峰值高达2.14，而后才稍有回落。

或许，该比率终有一天会回到0.76，也可能不会。

我如何得知？因为这些是我用随机数字生成器捏造出的数据。我再一次老调重弹，使用了计算机随机数字生成器。将起始股价设为50美元，然后用电脑抛25次硬币，正反面分别加减50美分，得到10只虚构股票两年内的每日股价数据。

随后，我又观察了配对股价的比率，也没花太长时间。图11.10和图11.11的数据，就是随机股价2和随机股价1的比率。

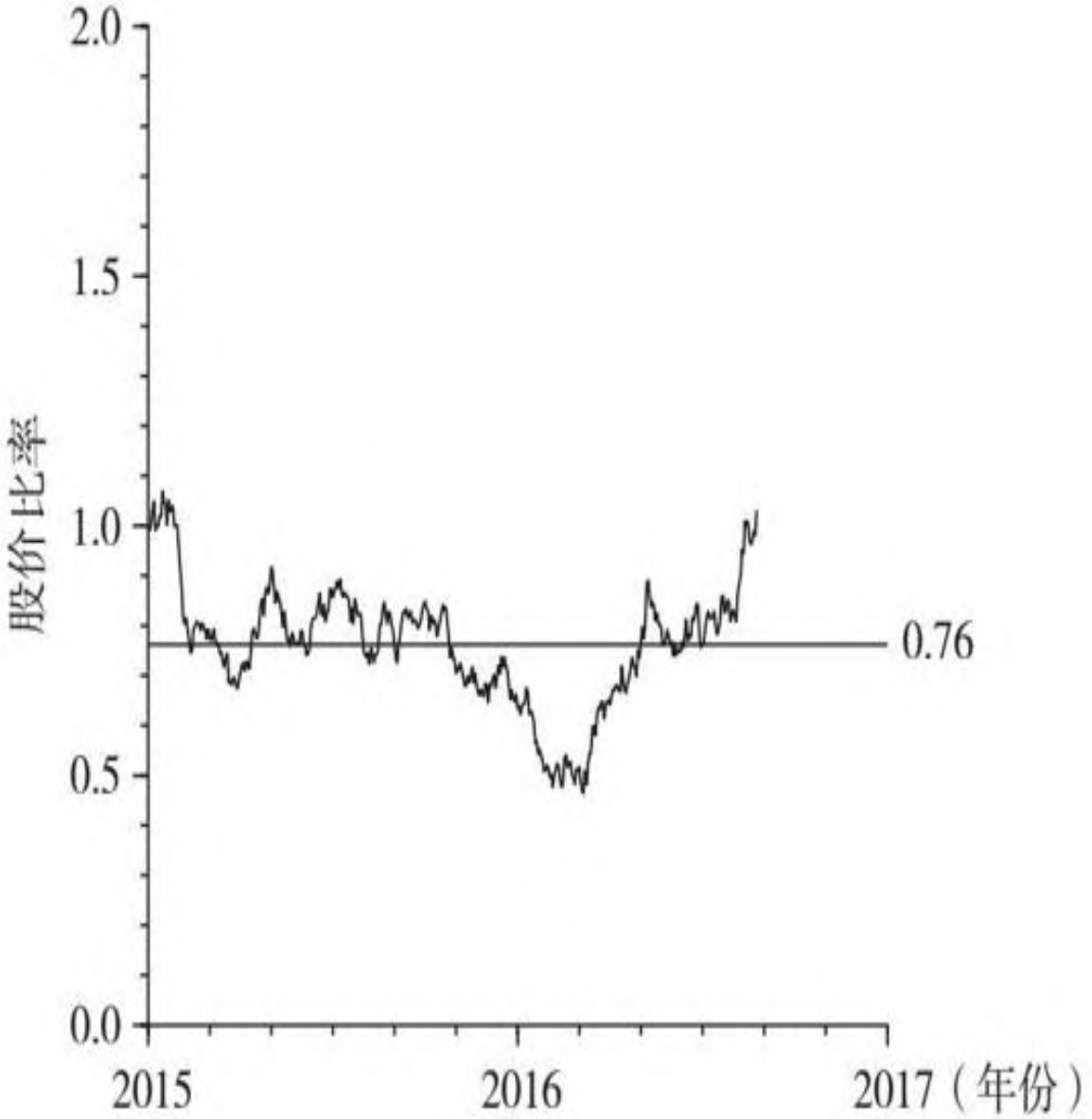


图11.10 趋同交易

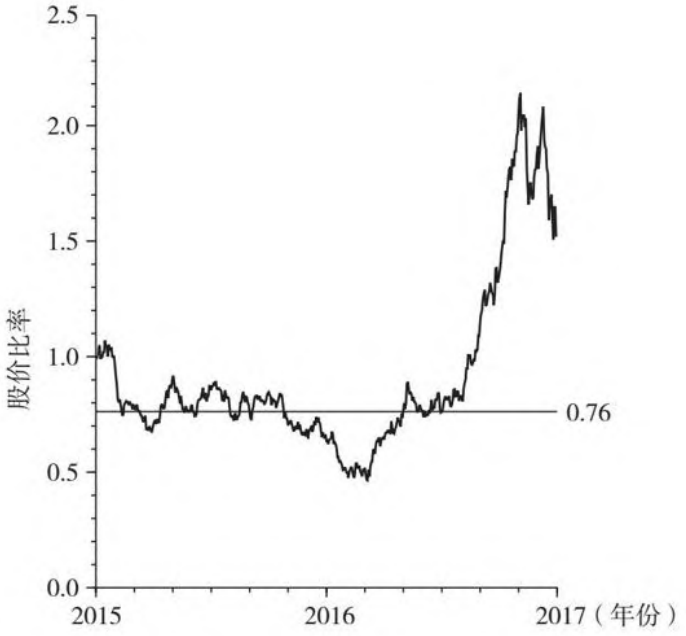


图11.11 尴尬的局面

比率初始值为1，是因为每个虚假股价的起价均为50美元。抛硬币碰巧得出的比率围绕0.76来回移动了一年半。之后，比率突然暴涨，接着有所回落。该比率下一步会如何变动？我不知道，这完全取决于计算机的随机数字生成器。

随机股价1和随机股价2完全独立。它们的共同点仅在于起价均为50美元。在这之后，股价1每日的变化都由电脑的25次抛硬币决定，股价2也是如此。每个价格都跟随随机游走程序，涨跌可能性一样大，完全独立于其他价格路径。但是，它们的价格比率似乎都停留在0.76附近，不会长时间偏离，很快又会再次恢复。后来，随机游走程序突然将比率带得远离0.76，可能再也不会恢复。

要注意的是，即使数据完全是随机生成的，还是会出现适合进行趋同交易的情况。但这并不代表每次潜在的趋同交易都是随机噪声。要提醒大家的是，我们无法像黑匣子式的数据挖掘软件那样，仅靠观察数据，就能分辨出趋同模型反映的情况是真实的还是偶然的。计算

机对于判断趋同模式是否具有逻辑基础完全无能为力。只有人类才能判断关系的形成理由是否具有说服力。对荷兰皇家壳牌集团来说，这个答案是肯定的。但对GSR来说，就不是了。

高频交易

有些算法被用以进行高频交易，使其买卖速度快到超乎人的想象。计算机可能会注意到，只要股价下跌的股票数量在接下来的140秒内比股价上涨的股票数量超出8%，标准普尔500指数的期货价格通常也会上涨。计算机将这一指标存档待用。当同一信号再次出现时，计算机便发动猛攻，立刻买入数千手标准普尔500指数期货，随后又迅速卖出。《连线》杂志对这些自动化系统赞不绝口，认为它们“比所有人类都更高效、快速和聪明”。更快速，确实是的；但更聪明，并非如此。

投资公司花费几十亿美元建立靠近股市的交易中心，使用光纤网线、微波塔台和激光通信线路，将芝加哥、纽约、伦敦、法兰克福和东京的信息传播与交易下单时间缩短至毫秒和纳秒。例如，纽约证券交易所和芝加哥商品交易所之间的一连串微波塔台，能在9毫秒内往返发送距离超过700英里的买卖订单。为什么要这么做呢？

第一个目的是，利用可察觉到的定价差异。假设IBM的股票在一家交易所的每股买入价为200.0000美元，在另一家交易所的每股卖出价为200.0001美元。发现这一异常现象的计算机程序会以200.0000美元尽量多地买入，为的就是在一毫秒后再以200.0001美元卖出，直到这一价格差异消失。每股0.0001美元的收益并不多，但是如果在一秒内完成数百次或数千次交易，就能产生非常可观的年收益。

在理性的世界里，资源不会浪费在这些无意义的事情上。不同交易所的股价出现如此细微的差异，这真的重要吗？差异定价持续了9毫秒，而不是10毫秒，这真的要紧吗？

极速交易的第二个目的是，比普通投资者更快一步下单。如果杰里下单，以当前市场价格买入1 000股股票，极速运转的交易程序可能会先买入，又在毫秒之间卖给杰里，一来一回每股赚取1美分的利润。以每股1美分的收益交易1 000股，就获得了10美元的收益。如此不断重复，利润可以达到数百万美元。计算机坑骗杰里，让他每股多付1美分，这给社会带来了什么经济利益呢？毫无利益可言，只不过是个计算机化的扒手偷了钱，受害者甚至还蒙在鼓里。

更根本的是，用超级智能程序来运行极速交易程序，而不是将其用在大有裨益的地方，会带来什么经济利益？建立交易中心，布好传输线路来加快股市下单，而不是将这些资源用在大有裨益的地方，会带来什么经济利益？

极速交易反而会雪上加霜，导致经济损害。

如果有人让计算机寻找有可能获利的模型（无论所发现的模型是否合理），然后在模型重现时买入或卖出，计算机会唯命是从（无论这个模型是否合理）。的确，计算机背后会有人吹嘘，他们真的不知道为什么自己的计算机会自行决定交易。毕竟，计算机比他们更聪明，不是吗？他们该做的不是自吹自擂，而是自求多福。

指令克隆问题也提高了黑匣子式高频投资的风险。如果软件工程师给几百台计算机下达相似的指令，就会有数百台计算机在同一时间竞相买卖同一只股票，广泛影响金融市场的稳定。《连线》杂志值得赞扬的是，它认识到了无人监管的计算机一致运作存在危险：“最糟糕的情况是，无人监管的计算机变成难以捉摸的反馈循环……最终击垮了计算机系统。”

闪电崩盘

2010年5月6日，美国股市受到著名的“闪电崩盘”的冲击。当天，投资者都担心希腊债务危机，一名焦虑的互惠基金经理设法卖出41亿美元的期货合约，以对冲其投资组合。他的思路是：如果市场下跌，基金股票投资组合的损失可以用期货合约的收益抵消。这一看似谨慎的交易，不知怎么就触发了计算机。计算机买入大量该基金卖出的期货合约，然后又迅速卖出，因为它们不喜欢长期持有头寸。期货价格开始下跌，于是计算机决定加大买入卖出的数量。受到刺激的计算机疯狂进行交易，自买自卖基金的期货合约，就像一个被丢来丢去的烫手山芋。

没有人确切知道计算机为什么会突然一发不可收拾。记住，就连计算机背后的人也不明白，计算机为什么会进行交易。在15秒的间隔时间内，计算机跟自己完成了2.7万次期货合约交易，占总交易量的一半，在疯狂的15秒结束后，净购买量只有200份合约。这一疯狂交易扩散到了常规股市的交易大厅里，卖出的订单淹没了潜在的买家。道琼斯工业平均指数在5分钟内下跌近600点。坚如磐石的蓝筹股宝洁的股价也在不到4分钟内下滑了37%。有些计算机为苹果公司、惠普公司和知名拍卖行苏富比支付的每股股价超过10万美元。还有些计算机将埃森哲咨询公司的股票和其他主要股票以每股不足1美分的价格卖出。这些电脑都没有常识，它们完全不知道苹果公司和埃森哲咨询公司的价值。只要算法下达指令，它们就盲目地买卖。

直到期货市场的内置安全卫士中止所有交易5秒，这一疯狂局面才得以落幕。令人难以置信的是，这短短的5秒，就足以说服计算机停止它们的疯狂交易。15分钟后，市场恢复正常，道琼斯工业平均指数短暂暴跌600点也只是梦魇般的回忆。

在那以后还发生过“闪电崩盘”，未来可能会出现更多。令人匪夷所思的是，2013年8月30日，宝洁再次于纽约证券交易所遭遇一次微

型“闪电崩盘”，之所以这么说，是因为并没有对该交易所其他股票产生特别大的影响，宝洁在其他交易所的股票也没有受到特别大的影响。

莫名其妙的是，纽约证券交易所的约200次交易，包括涉及宝洁股票在1秒内完成的约25万股交易，触发股价下跌了5%，从77.50美元降至73.61美元，随后在不到1分钟内即恢复。有个运气好的人，恰巧在正确的时间，出现在正确的地方，买入了6.5万股该股票，立刻赚了15.5万美元。为什么会发生这种情况呢？无人知晓。

虽然交易所启动安全卫士限制了进一步的“闪电崩盘”，但这也说明了黑匣子式投资算法的根本问题。这些计算机程序不知道每只股票（或任何其他投资）是真的廉价还是昂贵，甚至没有试图估算这只股票的真正价值。这就是为什么计算机程序可能以10万美元每股的价格买入苹果的股票，而仅以1美分每股的价格卖出埃森哲的股票。

底线

计算机没有常识或智慧。它们能识别统计学模式，但无法判断所发现的模式是否有逻辑基础。20世纪80年代，当黄金和白银价格的统计学相关系数被发现时，计算机程序怎么可能察觉得到这个统计学相关系数是否具备合理的理论基础呢？在宝洁股票价格瞬间下跌5%时，计算机又怎么能判断这次暴跌是有理有据，还是荒谬离谱的呢？

“人非圣贤，孰能无过。”但是人也有潜力识别那些错误，避免被计算机模型所诱惑。

我曾经的一名学生创立了一家成功的基金公司，采用的是投资其他投资基金的策略，而不是直接买入股票、债券和其他资产。他勤勉

努力，采访了数千名投资者和基金经理。最后，他确定了金融工程师对冲基金的四种类型（有些基金经理会结合多种策略使用）。

1. 纯套利。利润来自交易等同或接近等同资产，通常为高频交易。例如，在两所不同的交易所的同一只股票。其利润一般很小，但稳定，风险小。

2. 市场制造者。利用股价差异，例如，在不同交易所的相似证券以极小差价进行交易。获利可观，但风险在于，交易不按照预期价格执行，尤其在交易所于不同时间和不同日期开盘时（受假期影响）。

3. 统计学套利。使用数据挖掘算法来识别有可能是有利可图的交易基础的历史模型。利润丰厚，但风险也很大。例如，在一家航空公司买入股票，到了另一家便卖出。

4. 基本面量化。采用基础数据（如股价/收入），有助于支持具有某些特征的股票，同时避开或卖空具有相反特征的公司。

他对金融工程师的总体评估是：“如今有几千名‘金融工程师’投资者和互惠基金。只有小部分能够实现极好的长期获利。正如音乐家，可以靠不止一种类型的音乐取得成功。爵士、摇滚和乡村音乐艺术家的演出门票都会销售一空，同时还有几千名其他类型的音乐家在更廉价的夜店或街道转角演奏。”

第12章

我们都在监视着你

人类常常会将动物、树木、火车等非人类物体拟人化，假定它们具备人的特征。例如，在儿童故事和童话中，小猪盖好的房子被狼吹倒了、狐狸会和姜饼人说话。

花一分钟思考一下这些故事。那三只小猪具有人的特征，反映在它们分别盖了草屋、木屋和砖屋。野狼使出各种诡计，想要把小猪从砖屋中引出来，但三只小猪还是以智取胜，在发现野狼爬上屋顶是为了从烟囱进屋时，便在壁炉下放了一大锅沸水。

一名膝下无子的妇女刚烤好的姜饼人，为避免被吃逃跑了。它逃离了那名妇女、她的丈夫、其他人和动物，大声奚落追赶它的人："快跑。使劲跑！你就是抓不到我。我是姜饼人！"在有些版本里，有只狐狸欺骗姜饼人说带他过河，最后姜饼人还是被狐狸吃掉了。我小时候听到的版本是，一只狡猾的山猫企图引诱姜饼人到它家吃晚饭，但是附近一棵树上的小鸟警告姜饼人，山猫是想把它当作晚餐。于是，姜饼人就逃跑了。山猫龇牙咧嘴地低吼："真讨厌！"姜饼人后来又跑回家里，全家人都欢迎他回来。姜饼人也承诺，再也不逃跑了。

这些都是经久不衰的童话故事，因为我们是如此乐意，也可以说是渴望假设这些动物（甚至饼干）拥有人类的情感、想法和动机。同样，我们也假设计算机拥有情感、想法和动机。但它们并没有。

不要将计算机拟人化。

它们讨厌这样。

尽管如此，我们对描述机器人变得比人类更聪明的骇人听闻的科幻小说情境，既着迷又恐惧，其中的机器人聪明到决定必须消灭会让自己丧失能力的人。

《终结者》和《黑客帝国》等电影的成功，让很多人都相信这就是我们不久后的未来。就连史蒂芬·霍金和埃隆·马斯克这样有影响力的大人物，也警告人们机器人会起来反抗。2014年，马斯克对一群麻省理工学院的学生说：“我们是在用人工智能唤醒恶魔……在这些故事里，有一个人拥有魔力五角星和圣水，他自信能控制恶魔，但最后还是失败了。”三年后，马斯克发了张照片，上面有一句不详的警示：“到最后，机器会取胜。”2017年，剑桥大学名誉教授马丁·里斯爵士预测将来会是“后人类”世界，机器会统治地球长达几十亿年。幸运的是，他认为机器人统治世界在“几百年内”不会发生。

与我交谈过的所有知识渊博的计算机科学家都把计算机统治世界的想法当作纯粹的幻想而已，不予考虑。计算机并不知道世界是什么，人类是什么，也不知道生存意味着什么，更别说知道如何生存了。

更现实的危险在于，人类之所以听从计算机的指令，不是因为害怕被它消灭，而是因为惊叹于计算机的能力，相信它可以做出重要的，甚至生死攸关的决定。还记得希拉里·克林顿是如何过度信任“阿达”的吗？她并非个例。哪份房贷申请会被批准？哪名求职者会被录用？哪个人会被送进监狱？应服用哪种药？哪个会是轰炸目标？有太多人认为，因为计算机比我们聪明，所以应该由它们来做这些决定。

从开始到现在，我一直千方百计地为你阅读本章做好铺垫。下文内容或许将让你大吃一惊——除非你思考我们在前面章节讨论过的问题。

大数据主要意味着，“大企业”（Big Business）可以监视我们的信用卡、支票存款账户、电脑和电话，或从专门搜集个人信息的公司购买数据，以便能够预测我们的行动，操控我们的行为。

它们不仅监视我们在商店或网上的购买记录，还监视我们的网页浏览记录、手机使用记录、驾驶的车型和朋友圈。有些企业利用监控

摄像头来追踪我们在其商店的行走路线，据此排列商品来引诱我们购买更多东西。你有没有注意过，宜家的布局就像个大迷宫，找不着方向的顾客几乎不得不走过每一条小道，才能到达想去的那个区，然后又得按原路返回，才能找到出口。这并非无意之举。

“大企业”搜集到的海量个人数据用来促使和刺激我们买一些不需要的东西、参观一些不喜欢的地方和投票给不信任的候选人。“大企业”也会使用大数据来决定谁该被录用或被辞退。

其中一个障碍在于，人类是反复无常，甚至是不理性的，因此要预测人类的行为十分困难。为什么豆宝宝（Beanie Babies）和卷心菜娃娃（Cabbage Patch Kids）变成了昂贵的收藏品，而其他角色的填充玩具只是默默出现，又默默消失了？为什么会掀起呼啦圈狂热，之后又消退了？为什么有些电影票房失利，而有些则破了票房纪录呢？

畅销小说家威廉·高德曼靠写电影剧本曾两度捧获奥斯卡奖。他在自己的回忆录《银幕春秋》（*Adventures in the Screen Trade*）中写道：

> 整个电影业唯一重要的事实可能是：所有的人都一无所知！

高德曼还举了两个例子：

> 《夺宝奇兵》是影史上票房排名领先的电影……但是，你知道这部电影曾被商业区的所有影院拒绝吗？
>
> 除了派拉蒙影业公司。
>
> 为什么派拉蒙会接受？因为所有人都一无所知。为什么其他影院会拒绝？也是因为所有人都一无所知。为什么实力最强的环球影片公司决定把《星球大战》转交出去？因为没有人，没有一个人（现在不会，也永远不会）对票房将会如何有丝毫头绪。

我们都容易冲动，多愁善感，爱跟风随大溜，常做蠢事。

此外，我们一次又一次地见证了数据挖掘算法一定会发现一些模型，这些模型对过去情况的预测结果好得让人难以置信，但是对预测未来状况毫无作用，甚至更糟。大数据并不总是最了解情况的，对企业、政府和普通市民来说，依靠数据挖掘模型（尤其是隐藏在黑匣子里的模型）来做重要决定，会带来重重危机。

妊娠预测指标

塔吉特公司是美国第二大零售百货集团（仅次于沃尔玛）。塔吉特的品牌定位比沃尔玛更高——满足“更加年轻、有形象意识的顾客”。因此，为了与城市消费者产生共鸣，它还特地造了个法文名：Tar-jay或Tarzhay。

塔吉特最擅长的就是搜集和分析顾客数据。它为每位顾客单独派发了贵宾卡号，用来追踪他们所有的购物记录以及与塔吉特的互动情况，包括电子邮件、网站访问和优惠券使用情况。为了了解顾客，塔吉特搜集并购买其他数据，包括明显具有价值的信息（如年龄、职业和大概收入）和价值不太明显的数据（如在音乐、电影和咖啡机方面的喜好）。

在一个有趣的项目中，塔吉特开始着手标识怀有身孕的顾客。它找出更可能在人生大事的转变期——结婚、生子和离异时期——改变购物习惯的客户群。塔吉特想诱导新晋妈妈在他们公司实现一站式购物体验。

在妈妈们生下孩子后，很多商店都利用公开的出生记录，轰炸式地向新生儿家庭发送跟婴儿用品相关的优惠券和优惠信息。塔吉特希

望能先声夺人，在这些准妈妈临盆前就锁定她们。

承担本次项目的任务小组并没有使用数据挖掘软件，狂搜庞大的数据库，寻找统计学相关系数，或执行更糟糕的措施，如使用数据规约程序将他们的数据转化为无法识别的数据大杂烩。相反，他们通过标识在塔吉特登记宝宝信息的女性，建立了自己的模型，然后对她们购物习惯的改变进行分析。他们也思考了，人们是如何猜出某名女性怀孕了的。他们知道，孕妇在怀孕的前半段时期，都喜欢买营养补品，如补钙的；在第二孕期期间，倾向于购买使用无香味的香皂和乳液；到了第三孕期期间，便开始囤棉球和婴儿浴巾。在监控统计学分析的小组人员看来，这些线索都很合理。

塔吉特运用25件商品的数据来推算一名女性怀孕的可能性，然后预测她的预产期。它将自己的模型运用到所有顾客身上，给她们发送（注意喽）有针对性的优惠券和特别优惠信息，诱导可能怀孕了的女性到店消费。塔吉特还利用自己的数据库来确定优惠类型——邮寄传单、电子折扣或星巴克优惠券，这是有史以来针对每一位顾客最成功的一招。

唯一的问题在于，有关婴儿的优惠券会让女性心生警惕，觉得塔吉特侵犯了她们的隐私，知道自己怀孕在身。于是塔吉特不得不做出调整，将和婴儿有关的优惠券与无关的优惠券混在一起，这样就不容易让这些女性察觉到塔吉特掌握了所有信息。

塔吉特的模型见效了，因为它是由人类专家来管理的，也没有隐藏在黑匣子里。但数据挖掘模型就截然不同了，数据挖掘方法会运用数据规约步骤（如主成分分析）处理女性的数百个特征，就连毫无意义的数据也不放过，如买了男士袜子和猫粮。这么做得到的结果，就像一道莫名其妙的浓汤，既有苹果，又有橙子，还有土豆。塔吉特并没有这么做，但其他公司却照做不误。

谷歌流感

2011年，谷歌报告称，其研究人员已经开发出名为“谷歌流感”的人工智能程序——利用谷歌搜索的查询功能来预测流感的暴发，最多可早于美国疾病控制与预防中心（CDC）10天做出预测。他们吹嘘道：“我们能准确推算出美国每个地区每周流感活动情况的当前水平，报告时间间隔约为一天。”他们的大言不惭为时过早。

谷歌的数据挖掘软件审查了5 000万条搜索查询，找出了与流感发病率相关的1 152个观察结果最吻合的45个关键词，再使用IP（互联网协议）地址来识别该搜索来自哪个州。为了安抚谷歌用户，表示他们的个人隐私没有被侵犯，谷歌报告称，“谷歌流感”是个完全黑匣子式的程序，因此在选择搜索词语方面没有人为影响，而且，被选关键词也不公开。这样的坦言或许能安抚那些担心自己隐私的用户，但对那些知道黑匣子式模型缺点的用户来说，这应该是个危险信号。

在2011年的报告中，谷歌表示该模型的准确率高达97.5%，意味着在样本内时期，模型预测结果和CDC公布的真实数据之间的相关系数为0.975。一名麻省理工的教授对此模型称赞不已：“这种方法看上去真的很明智，利用谷歌用户无意间创造出来的数据，查看在网络世界里看不见的模型。我觉得，我们还只是蜻蜓点水，还有更多集体智慧可以深入挖掘。”

然而，在报告发布后的108个星期内，“谷歌流感”有100个星期都过度推算了流感病例数量，错误率接近100%。“谷歌流感”对样本外数据的预测结果的准确率比样本内的低得多，预测出未来一周的流感发病数量会与一周前、两周前，甚至三周前都相同。由于流感发病率到了冬季会上升，夏季则下降，因此，就连根据每日气温制定的简单模型都可能比谷歌的黑匣子程序预测得准。

谷歌表示，“谷歌流感”对样本外数据预测失败可能是因为选择了“发烧”（fever）等不一定与流感有关的通用搜索术语。或者，可能那些人以为自己患了流感，但其实没有。毕竟，普通人没法像医生那样准确判断。

这些都没错，但是数据在训练期间只是噪声，其准确率还是达到了97.5%。种瓜得瓜，输入无用的数据，输出的自然也是无用的数据。为什么大家都想着石头真的能变成金子，就因为这个模型有“谷歌”的名字加持吗？

有人质疑说，由于流感暴发具有高度季节性的特征，“谷歌流感”可能主要是冬季探测器，挑选出季节性的搜索词语，如新年前夕和情人节。记得吗，华盛顿市奥玛克镇的最高气温是如何预测澳大利亚科廷的次日最低气温的？即使两者之间根本不存在因果关系。谷歌的流感预测可能也是同理。审查5 000万个搜索词语，一定会有某些恰巧与流感暴发、出生率和中国茶叶价格相关的信息。纸包不住火，问题总有暴露的一天。

“谷歌流感”从此再也没有预测过流感。

机器人测试仪

一家经营分析业务的公司WhatWorks主要测试为互联网公司设计的不同网页。WhatWorks先使用计算机化的随机事件生成器把访客引至网页设计不同的页面，通过这个实验来做出推荐。测试几天后，WhatWorks会告诉客户，哪款网页设计会产生最高的人均收益。

WhatWorks受聘于一家互联网企业BuyNow，管理100多万个网页。WhatWorks进行了上述测试，然后反馈了一个最佳设计方案。BuyNow提

出了一个更好的点子。浏览不同网页的访客可能有不同的偏好，因此，为什么不给每个网页量身定制设计呢？BuyNow并不知道访客的差异在哪儿，也不知道这些差异可能会如何影响他们的喜好，却认定定制化的设计比通用的更好。

这没办法手动完成，人类专家不可能这样来决定每个网页的设计，因为这需要做超过100万次试验。而且，很多网页的访客量微乎其微，通过测试决定还得花数月时间。因此，为什么不利用机器学习的能力？BuyNow请WhatWorks有偿开发了机器人测试仪（Robo-Tester）。这个自动化的人工智能系统能让每一个网页使用各种设计，监测哪种效果最好，而且，随着它可信度的提高，受喜爱的设计的出现频次也会增加。

只可惜，理想很丰满，现实很骨感。BuyNow以为这个人工智能系统会提供最终确定的答案，但是网页浏览量以及由此产生的收益状况也会有很多随机噪声。一款拙劣的设计只是侥幸看上去不错，哄骗机器人测试仪喜欢上这个不好的设计。此外，随机噪声意味着，机器人测试仪不可能百分之百确定它已经找到了最佳设计，因此它会继续测试被淘汰的设计。

最后的结果是，与WhatWorks专家最初建议的单一设计相比，机器人测试仪使BuyNow的收益减少了约1%。

尽管是BuyNow自己出的主意，它还是怨声载道。

BuyNow：使用AI技术自动化处理每一个网页的设计，这有什么不好？

WhatWorks：你单个网页浏览量都不足以得出有意义的测试结果。

BuyNow：那有统计结果证明可行的话，只转换布局怎么样？

WhatWorks：统计学的结果不一定有意义（在惊愕中沉默）。

WhatWorks：我的意思是，做上100万次测试，你怎么都能得到统计学上的证据，但这并不意味着就能选出最佳设计。

BuyNow：是哦，你总是这么说。

WhatWorks：可我一直都没说错啊！

WhatWorks确实没错，但有些人就是听不进去“使用人工智能系统会适得其反”这句话。后来，BuyNow终止了与WhatWorks的合作。

就业申请

基于量化数据（如平均绩效、工作年限）来评估求职者真的比登天还难。即便在“大数据圣殿”谷歌，其负责人力运营部的副总裁也在2016年坦言：

“向我介绍一下你自己。你最大的优点和缺点是什么？”这些普通的面试问题都无法体现这名求职者能否胜任这份工作。

其实，应该问的问题是“举个例子说明你曾经有过与求职岗位相似的工作经验”。这类问题叫作“结构化行为面试问题”。问其他问题只是浪费时间而已。

但是，人工智能程序又怎能提出和评估结构化行为面试问题呢？

对于那些有过相似工作经历的求职者，最重要的素质就是迅速适应且胜任新工作的能力。一名优秀的新雇员会清楚什么是所需信息、如何利用这些信息以及如何处理意料之外和模棱两可的情况。

可在这些情况下人工智能程序只能查找特定关键词来筛选，如Excel（电子表格软件）、销售、数学和博士等。人工智能算法不会灵活处理，它想不到这些关键词是评估求职者的秘诀。除非算法的创建者将词语输入程序，否则，查找computer science（计算机科学）的人工智能算法会忽略data analytics（数据分析），查找PhD（博士）的会忽略Ph.D.，查找mathematics（数学）的会忽略math。人工智能算法不会知道“MakeMyDay”既是一家互联网礼品公司的名字，也代表一家专门观看克林特·伊斯特伍德电影的俱乐部。人工智能算法也搞不清楚在加州理工学院取得3.2分的平均绩点（天才的水平）和在一些我连名字都说不上、容易考进且夸大分数的学校取得同样多的平均绩点，会有什么区别。

所以，求职者会伺机利用系统的特点，在他们的自荐信和简历中添加必备的关键词，例如，表示自己对mathematics感兴趣，而不是写明自己的数学水平。

如果你曾注意过成功或不成功的求职者，就会知道求职的成败并不取决于简历和求职信所包含的关键词。

有些公司开发了数据驱动的软件，通过监视求职者的网络动态来评估他们。有家公司的首席科学家承认，人工智能数据挖掘软件所选的部分因素并没有意义。例如，该软件在其数据库中发现几名出色的程序员都频繁访问过一个日本漫画网站，因此它做出的判断是，浏览这个网站的都可能是出色的程序员。这名首席科学家表示：“这显然不存在因果关系。”但话又说回来，她认为该软件仍大有用处，因为其中存在明显的统计学相关系数。真让人大跌眼镜！这恰好再次证明了一个毫无根据的观点（就连应该更清楚这一点的人也会误解），即数据比常识更加重要。但实则不然。

该名科学家还说，公司的算法观察了几十个变量，随着相关系数的出现和消失，不断更换变量。她认为，一直变化的变量体现了该模型的能力和灵活性。不过，更有说服力的解释是，该算法捕获了转瞬

即逝、毫无价值的偶然性相关系数。如果这些是因果关系，它们不会出现了又消失，而是会一直存在，并且有所用处。这就好比股票的技术分析发现了飞逝的模型，然后就是捏造模型、进行测试、改变模型，不断重复。

那么，这家公司的软件成功识别出优秀的求职者了吗？一名在该公司任职三年的员工写道：

> 顾客真的很讨厌这个产品。几乎没有顾客有过整体良好的体验。这种情况已经持续多年，管理层都无力回天，无法解决这个问题。该产品小组没做过任何产品研究，或是建设性地将有意义的反馈建议融入产品设计，所以这个问题永远都无法解决。

招聘广告

计算机算法也会用于工作招聘。本书的引言讨论过政客如何使用微目标锁定的方法来精心设计宣传口号，缩小潜在投票者的范围。企业的做法也如出一辙，事实上，是它们最先采用这种方法的——利用庞大的数据库探清顾客的想法，找出他们容易接受哪些产品的特定推销手段。

微目标锁定给予拥有海量个人数据的互联网公司，远胜于报纸、杂志、广播、电视等传统广告媒体的巨大优势。企业不需要购买广大受众都能读到、听到或看到的广告，而是专门针对潜在顾客。

招聘广告也是如此。企业不必将招工信息广而告之，而是针对有资质和有兴趣应聘的潜在求职者。不幸的是，这也会有意或无意地造成歧视。

2017年，《纽约时报》和ProPublica（非营利性在线新闻机构）开展的新闻调查研究荣获普利策奖。该研究发现，微目标锁定的招聘广告常常歧视年龄大的工作者。例如，威瑞森电信的一则金融相关岗位的招聘广告，被推送至对金融感兴趣、当前或最近住在华盛顿、年龄介于25~36岁的脸书用户；UPS（美国联合包裹）的招聘广告只对准了年龄位于18~24岁的群体；State Farm（美国汽车保险公司）的招聘广告面向19~35岁的人群；而脸书则有意招聘25~60岁的人。

有些公司争论说，这样的定位方式与刊登在受年轻读者欢迎的杂志上，或者播放于年轻听众爱收听的广播电台的广告并无不同。这也就是为什么日间电视节目会出现很多针对职业培训学校和人身伤害索偿律师的广告。然而，作为年逾60的人，我可以读任何想读的杂志、听任何想听的广播、看任何想看的电视，但就是看不到脸书没有推送给我的广告。我甚至都不知道这些广告的存在。

2016年，有一项评估社交媒体广告有效性的实验，想要通过性别中立的手段招聘科学、技术、工程和数学工作领域的人才。尽管已经明确说明其性别中立的意图，但看到此则招聘广告的男性数量还是比女性多20%。

研究人员认为，或许是女性访客点击该广告的可能性更小，因此，设法使点击率最大化的人工智能算法可能不会再关注女性访客流量更大的那些网页。这么想就错了。其实，女性更有可能点击那则广告。而女性访客点击率更低的原因更加微妙。人类能够查明此原因，但计算机算法就望尘莫及了。

年龄介于25~34岁的女性是更有价值的互联网客户，因为她们更容易点击售卖产品的广告，也更容易在线购买产品。因此，女性访客流量较高的网站的广告费用更贵。无论是招聘，还是产品，其广告费都不低。而设计目标为以最少费用投放广告的人工智能算法，自然坚决不会选择价格相对高的网站，但这些却都是女性浏览频率更高的网站。

研究人员思考出这个创意十足的答案，还能够以各种方法进行验证。而人工智能算法对此浑然不知。

贷款申请

一直以来，贷款评估申请都是采取合理方法衡量申请人的还贷能力，比如偿还贷款和其他账单的记录。还贷能力取决于其就业记录、收入、财产和其他债务状况。账单支付记录只是反映有多少笔账被迟付或未付。放贷的5C原则是：

性格（Character）：申请人是否有及时付清账单的记录？是否有犯罪记录？

能力（Capacity）：申请人是否有稳定的就业经历？是否有充足的收入用以偿还贷款等债务？

资本（Capital）：申请人的收入若减少，是否有充裕的财产用以还贷？

条件（Conditions）：申请人的就业领域的局势是否不稳定？

担保（Collateral）：申请人若未偿清贷款，是否有充足的担保品让贷方拿回款项？

在互联网和社交媒体爆炸式发展的今天，贷方开始寻找其他没那么直接的判断依据，尤其是个人性格方面。例如，访问内容正面的网页、与正派的人交往以显示自己的正派作风。没错，就像校园里的约会游戏，分受欢迎的和不受欢迎的两派。此外，你还是可以摆弄这个系统。比如，不在网络上发表任何尴尬的言论，或公开做出让自己难

堪的行为举止。阅读严肃的人写的严肃的网络文章，在严肃的网站发布严肃的评论。

贷方迟早会发现，他们可以摒弃人工评估使用的5C原则，转向在网络上人工搜索贷款申请人，看看是否会弹出令人难堪的信息。他们能利用人工智能来评估贷款申请者了！

2017年，中国一家AI放贷应用程序开发公司的创始人兼CEO争辩道：

> 银行专注于海面上的冰山一角，我们创建算法是让海面之下的海量数据也有意义。

海面之下有哪些有用数据？你可能会诧异于了解到这一切都离不开智能手机。很多人可以通过智能手机从网络贷方获得小额短期贷款。

这家中国公司的数据挖掘软件真的能够获取大量通过手机提交的贷款申请。此外，它还能为人工智能程序收集到关于智能手机本身的信息，以查出统计学模式。那名CEO还夸夸其谈道：

> 我们没有从传统的金融机构聘请任何人负责风险控制……我们不需靠人来告诉我们顾客是好是坏。技术，就是我们的风险控制。

显示某人信用风险小的证据包括：使用安卓系统手机，而不是iPhone（苹果手机）；不总是立即接听电话；呼叫电话不总是有人接听；手机不保持满电状态。

我们可以编出一些说法来解释所发现的这个统计学模式。使用价格不高的安卓系统手机，说明花钱谨慎；不总是立即接听电话，说明

工作认真；呼叫电话不总是有人接听，说明其朋友也同样工作认真；手机不保持满电状态，说明其精力更集中在工作，而非手机上。这样的解释说服你了吗？

如果你被说服了，可能会诧异于真实结果恰好相反。上述都是信用风险大的迹象！当然，我们照样可以编出故事。使用安卓系统手机，说明买不起iPhone；不总是立即接听电话，说明在躲避债主；呼叫电话不总是有人接听，说明人缘不好；手机不保持满电状态，说明不负责任。

重点在于，人类有足够的智慧给任何发现的统计学模型创造合理解释，即便这些模型是数据挖掘软件发现的随机噪声。找出模型说明不了什么，编造符合模型的故事也说明不了什么。

我预测，无论模型与数据的吻合度是否很高，这些算法对贷款拖欠状况的预测都会不理想。

汽车保险

2016年，英国最大的汽车保险商上校保险公司（Admiral Insurance）推出了“firstcarquote计划”（保费报价计划），利用人工智能技术分析申请者的脸书发帖，以此作为设定汽车保险费率的依据：

> 我们已经知道，社交媒体的发帖能告诉我们某个人的信用风险是高是低，而汽车领域也是如此。统计学证明，具备某些性格特征的人比其他人更容易发生车祸。但是标准的保险问题不会测量性格特征。在“firstcarquote计划”中，我们会通过分析司机的脸书发帖，了解他们的性格特征。如果我们发现说明你开车谨

慎的迹象，就会为你送上5%~15%的价格优惠，可在公司官网领取。

该公司的主管领导表示，人工智能算法通过制定清单和设置特定截止时间，在脸书帖子中寻找说明申请人谨慎小心的迹象。该公司设计算法的顾问给出了让人半信半疑的理由：

过度自信的人可能会使用“总是”或“从不”这类词语。靠不住的司机可能出现更消极的情绪，更常使用的词语是“或许”或“可能”，这表示他们很不自信……你能从中推断出关于性格的信息，而根据性格，我们又可以判断出你开车的安全性有多高。

然而，这样的解释令人疑信参半，用词选择或许能传达出申请人的性格特征，但很难说明这能当作判断某人是否容易出车祸的可靠预测指标。

该顾问接下来的解释偏离到越来越站不住脚的地步。他声称，对车险索赔来说，喜欢迈克尔·乔丹和莱昂纳德·科恩是个好的预测指标。最后，他本想再吹嘘一番，却不经意间承认了事实：

我们计算“驾车安全程度”的算法仍在不断完善中，将社交媒体上的数据和现实说明数据进行匹配……我们的分析并不基于任何特定模型，而是根据数千个点赞、用词和用语的不同组合，还会随着从数据中获得的新证据不断变化更新。因此，我们的计算反映出司机通常在社交媒体上的表现，以及这些表现的可预测性，而不是关于一名驾车安全的司机会长什么样的固定假设。

打住，就是这句。没有“固定假设”意味着没有逻辑基础。该算法只是黑匣子式的数据挖掘模型，寻找历史相关系数，毫不关心模型是否合理。此外，该算法不断改变是因为它毫无依据，所以才会被短暂出现的相关系数摆弄来摆弄去。又是捏造模型—进行测试—改变模型，如此循环。

不过，我们再也没有机会看到这个算法成功与否了。就在该计划推出的前几个小时，脸书宣布不允许上校保险公司获取脸书的数据，其政策为“严禁使用从脸书获取的数据来判断条件资格，包括是否批准或拒绝申请人、收取多少贷款利息”。

对上校保险公司来说，这或许是塞翁失马，焉知非福。

社会信用评分

目前，有些企业与政府合作搜集数据，追踪民众的购买记录、出行记录、行为记录等可表明该人不值得信赖的信息。

被计算机算法评为高分的民众可享有价格折扣，支付很低的保费费率，减免租金。对那些评分低的民众，则不允许其购买某些商品，还必须为能购买的商品支付更多的钱。他们的出行选择和居住方式也很有限，他们还可能会受到警方监视。一家公司的总经理表示，这些评分“确保社会上的坏人无处可去，而好人则能行动自由，毫无障碍”。

这些评分背后的某些数据是合理的，例如，使用还贷记录来预测贷款拖欠情况。然而，有些数据肯定只是短暂的相关系数，例如，根据玩什么视频游戏或看什么电影来预测贷款拖欠情况。

一个人如果有评分低的朋友，自己也会受到消极影响。如果爱丽丝的评分下降，有些朋友就会抛弃她，以防自己的评分降低，这对爱丽丝的评分更是火上浇油。这种做法会造成社会分裂，因为评分高的人获得奖励的方式，是将评分低的人踢出自己的社交网络。

虽然脸书阻止上校保险公司利用其数据来为保险费率定价，但也说明不了脸书没有私心。脸书拥有基于你和你的脸书好友来评估贷款申请的专利算法。我写到这里的时候，脸书还未启用它的算法，但别以为它没有动过这个念头，也别以为它没有其他利用数据牟利的方法。

黑匣子式歧视

采用微目标锁定方法的招聘广告，就是有意青睐特定群体（也就不可避免排除其他群体）的明显例子。然而，如果微目标锁定利用了黑匣子式的算法来发现统计学模式，暗中造成了年龄、性别或种族歧视，这种不公平现象就很难被察觉。

没有人，就连编码的程序员也无法确切知道，黑匣子式的算法是如何做出预测的，但几乎能肯定的是，就业、贷款和保险算法都直接或间接考虑到性别、种族、民族和性取向等因素。如果黑匣子式算法会选择某些性格特征作为不良行为的预测指标，而某些人就因为具备这些性格特征遭到惩罚，是不合乎道德伦理的。

思考一下，算法通过浏览申请人访问的网站来评估职位申请者——不是好的或不好的网站，仅仅是与现有雇员有所关联的网站。虽然某个日本漫画网站很受男性软件工程师欢迎，但这对一名不会花时间浏览该漫画网站的西班牙女性来说，有多公平？

再思考一下，算法通过接听来电的频率来评估贷款申请。这对经常接到电话销售员来电的男性犹太教徒来说，有多公平？

继续思考一下，算法基于脸书的用词选择，以及某人是否喜欢迈克尔·乔丹或莱昂纳德·科恩，来设定汽车保险费率。这对喜欢迈克尔·乔丹的黑人男性和喜欢莱昂纳德·科恩的白人女性来说，有多公平？如果性别、种族、民族或性取向碰巧与汽车保险索赔存在偶然性关系，又有多公平？

不合理的搜查

一直有人争论道，挖掘“看似无关的数据”能够“发现、推断和侦查出犯罪活动”。我的天啊！竟然利用无关的数据来打击犯罪活动。

例如，或许我们可以利用数据挖掘算法，而不是法院来批准搜查令吗？我们似乎已朝这个方向前进，不过这次探索之旅肯定不尽如人意。

搜查令通常都有范围限制，也肯定是基于特定的、个别化的、没有明显无辜解释的事实的。如果缉毒犬嗅到了行李包有特殊味道，听到从汽车后备厢里传来声音或在地下室发现制毒设备，就是没有明显的无辜解释。但是，如果一家人锁上前门（为了防盗）或前往哥伦比亚（为了探亲或度假），就是有明显的无辜解释。

举个更复杂的例子。在伊利诺伊州，警方获取搜查令搜查一家酒馆，理由是该酒馆老板贩卖海洛因。警方进入酒馆对顾客进行搜查，发现文图拉·伊巴拉将海洛因藏在烟盒里。最高法院判决对伊巴拉先生的搜查属非法行为，因为这是基于“株连罪”的搜查，而不是基于

任何特定的可能原因。警方并不认识伊巴拉，他当时只在玩弹球游戏，没有逃跑意图、藏匿物品等可疑行为。他出现在该酒馆具有明显的无辜解释——进来喝一杯，玩玩弹球。

法院声明，“可能的原因（必须是）具体针对（某一个）人的”，而“指出碰巧会存在可能原因便搜查（其他人），或搜查该人可能恰好出现的场所”，是绝不能达到上述标准的。

这有可能是使用数据挖掘算法来颁发搜查令所难以解决的问题。计算机非常不擅长于为特定的，确切来说是为独特的情况推测无辜解释。还记得第3章那个“猫与花瓶”的例子吗？计算机没有人类的生活经历，无法猜测出符合语境的合理解释。人类看过风吹倒物体、见过孩子打翻东西后跑出房子、体验过地震的感觉，所以能联想到这些可能性。计算机软件却完全做不到。颁发搜查令也一样，人工智能算法无法猜测，更别说评估无辜解释。

第二个问题是，搜查令应该具有合理基础。废弃建筑物是违法毒品的制作场所，这说得过去。但要说教堂或老年人中心是违法毒品的制作场所，就没那么合理了。但数据挖掘算法并不能区分两者的差异。

数据挖掘算法的另一个问题是，它们可能会用种族、宗教或政治团体的替代指标，而这些都不应该出现在申请搜查令的考虑范围内。例如，搜遍大数据的算法可能会发现，住在固定地区、购买固定产品、穿着固定衣服或访问固定网站都是表明犯罪活动的指标，但也是种族、宗教和其他受保护阶层的替代指标。

法院要求特定的、个人化的犯罪活动证据，其中一个原因是，含混不清会滋生政府权力滥用。首先，民众应该知道哪些特定活动是违法的，或能作为侵犯他们隐私的理由。其次，政府应该限制个人隐私可以被随意侵犯的自由。

假设数据挖掘算法颁发搜查令或逮捕令的依据是合法活动，但该活动在统计学上与非法活动相关联。大多数被选择的模型都可能是暂时的随机噪声。要记住，搜寻的数据集越大，就越可能发现无意义的偶然性模型。

如果算法发现吸食违法毒品和在开市客超市购买狗粮、看《宋飞正传》、每三年买一辆新车之间有相关系数，会怎么样？这些可能完全是偶然存在的统计学关系，就像总统大选结果与5个城镇的气温存在相关系数一样。即使并非完全偶然，难道民众不应该被事先警告可能会被逮捕，或在收看《宋飞正传》时，被提前通知自己的房子要被搜查吗？

对于第二个问题，即警察的随意决定导致有选择性的法律执行。想象数据挖掘算法识别了1万个具备与所知罪犯的性格特征相似的居民。如果警方没有权力彻查这1万个居民，他们可能会从中选出自认为有特大嫌疑的几十个居民。但是，这就有损了要基于证据，而不是基于突发奇想的搜查原则。

看看你的手环

理查德·伯克在宾夕法尼亚大学的犯罪学系和统计学系均担任职务。他的专业特长之一是算法犯罪学，即“使用统计学或机器学习程序来预测犯罪行为和受害情况”。算法犯罪学越来越常见于审前保释、审后判决和定罪后假释判决中。

伯克写道：“这个方法就像个‘黑匣子’，错误率很低。”在《大西洋月刊》的一篇文章中，伯克更是直言：“如果能用太阳黑子、鞋码或手环大小的数据，我肯定会用。如果我给算法输入足够多的预测指标，它就会得到出乎你意料之外的结果。”

大多数“出乎你意料之外的结果”都毫无意义，如太阳黑子、鞋码或手环大小。它们反映出短暂的、偶然性的模型，这都是无用的犯罪行为的预测指标。有一项研究调查了较为广泛使用的危险评估算法之一，结果发现，被预测两年内施行暴力犯罪的人当中，实际上只有20%的人确有此举。此外，这个预测结果还歧视黑人。

伯克的初衷无疑是好的，但他认为太阳黑子、鞋码或手环大小可以用来决定一个人应该被假释还是继续监禁的想法，实在令人不安。若过度信赖计算机，就会出现这种情况。

更糟糕的是，出售犯罪学算法的公司通常不会透露它们所考虑的因素，以及哪些是它们的算法认为重要的因素，还声称这是有价值的专利信息。这个秘密让被告人不可能为自己被判监禁的结果辩护，因为这个结果的依据是数据中偶然出现的模型。

在一件案子中，威斯康星州的警方识别出一辆汽车涉嫌一起驾车枪击案，于是追赶该车辆。车停下来后，司机埃里克·卢米斯被逮捕，不久后认罪，承认自己逃避警察追捕，对“未经车主同意，擅自驾驶他人车辆”的指控也不予回应。法官部分考虑了计算机算法的预测结果，即如果卢米斯不被监禁，就有可能会犯下更多罪行，因此判处他六年有期徒刑。卢米斯上诉到威斯康星州最高法院和美国最高法院，均被驳回。

威斯康星州最高法院赞赏这种基于证据的判决——根据数据，而不是人的主观意见。人的意见肯定不完美，但黑匣子式的数据挖掘算法也是如此。法院表示，“尽管卢米斯不能质疑算法如何计算危险，但他至少能审查和质疑他最后得到的危险评分”。很难能见到他真的做到这一点，除非他创建自己的算法。

威斯康星州司法部长表示，法院应该持观望态度，“法院使用危险评估来进行判决还是新举措，需要时间做进一步观察渗透”。同时，卢米斯“有质疑评估和说明其可能存在的错误的自由”。实际上，他怎么能挑战得了一个黑匣子式的模型呢？

人气算法的创建者确实透露了少部分自己输入的数据，但并不能抚慰人心。在他们考虑的因素当中，有被告对下述问题的回答，如：“你有多少朋友或熟人服用违法毒品？”“如果有人惹怒你，你是否会做出危险举动？”说真的，谁会回答“很多”或“当然会”？

危险评估算法当然会忽略那些任何理智的犯罪分子都不会如实回答的问题。它们肯定会基于更多微妙数据（如手环大小和鞋码），而这看上去肯定公平。被告应该知道那些数据，而不是在算法渗透的过程中，把牢底坐穿。

你需要整容吗？

如果假释判决是基于预测惯犯行为的统计学模型，那么离使用这种模型来决定谁应该被捕和被送进监狱也只是一步之遥了。果然，在2016年，有两名中国研究人员报告称，他们能运用自己的人工智能算法扫描脸部照片，预测某人是否是罪犯，准确率高达89.5%。

我的第一反应是（可能你也一样），这就是个恶作剧，就像第6章那个死三文鱼的研究一样。可能是两个喜欢恶作剧的人，通过这样戏剧性的实验表明，数据挖掘能用来杜撰出不切实际的说法。使用几十万像素肯定能得到某些相关系数密切的结果。该模型通过扫描人脸和使用数据挖掘软件找出的统计学关系，也只能创建关于某人的智商、电影喜好或癌症易感性的虚假预测指标。

我读过那篇文章，惊讶地发现他们完全是认真的。他们侃侃而谈道：

> 和人类审查员或法官不一样，计算机视觉算法或分类器完全不会受过往经历、种族、宗教、政治信条、性别、年龄等的影

响，它没有主观看法、情感和偏好，没有精神疲倦，也没有睡眠不足或进食不佳的情况。

这两位研究人员扫描了1 856名男性的身份证证件照，其中730名为罪犯，另外1 126名是普通人。他们的人工智能程序发现，“一些有辨别力的结构化特征能预测犯罪分子，如唇曲率、眼内角距和所谓的口鼻角度”。

《麻省理工科技评论》持乐观态度，称：“这一切都预示着，‘人体测量学、犯罪学等方面’即将开启新纪元，随着机器的能力越来越大，还有很多研究空间可供发展。”一名受人敬重的数据科学家写道：“该研究的实施过程严谨，结果是什么就是什么。”这口吻就仿佛是真正的数据挖掘者。谁还需要理论呢？如果人工智能程序找到统计学模式，这个证据就足够了。结果是什么就是什么。

值得大赞一番的是，人工智能研究人员全都认为，若当真有可能导致危险后果，就要摒弃这种具有危害性的伪科学，它既不可靠，又容易让人产生误解。

那两名研究人员发表的文章中包括三名罪犯和三名普通人的脸部照片，它们无疑都是经过选择的，这样读者就会认同罪犯和普通人看上去确实不同。罪犯都不笑，没有身穿职业服装，皮肤也更加粗糙。除了唇曲率、眼内角距和所谓的口鼻角度（不管这些是什么意思），人工智能程序和我们注意到的一样吗？它除了检测笑容，还能做什么吗？

我还是怀疑这项研究可能就是场恶作剧，但如果并非恶作剧，那么人工智能罪犯人脸算法的目的是什么？我们应该逮捕那些一脸罪犯相的人，好让他们不能犯罪吗？

我记得2002年上映的电影《少数派报告》中，三名超能力者可以在罪犯犯罪之前就侦查出其犯罪企图，使得犯罪预防组织的警察可提

前将其逮捕。当然，逻辑问题在于，如果企图犯罪的人在杀人前被捕，就不会出现谋杀事件。因此，超能力者又怎么能够预先侦查到没有发生过的谋杀呢？

识别“罪犯脸”的人工智能算法并没有这种时间矛盾问题，因为它不能预测特定的罪犯。它只能用识别程序算出有89.5%的可能性是罪犯的人。这些罪犯可能过去已经犯了法。

但还是要问，人工智能程序会用来逮捕和监禁那些逍遥法外的人和那些只是企图犯罪的“前”罪犯吗？

一名博主对这项研究做了如下评论：

> 如果人工智能程序只是将那些看上去像罪犯的人送进拘留所，能怎么样？这会造成什么伤害吗？那些人不过是在里面待上一段时间，等“康复期”结束就好了。即便部分无辜的人也被关了进去，从长远来看，这会对他们造成什么不利影响吗？

我希望这名博主是在冷嘲热讽，但恐怕他就是这么想的。如果这样盲目信任计算机成为惯例，政府或许也会开始根据人工智能的人脸分析，来关押民众了。

如果你的唇曲率、眼内角距和所谓的口鼻角度也符合罪犯特征，那怎么办？不管是不是罪犯，你将来可能都要整容。

摆弄系统

数据挖掘算法有两个根本问题。一方面，如果算法是专利机密，我们则无法检查算法所使用数据的准确度。如果黑匣子式算法被告知

你有贷款未偿清，而真正欠款的是跟你同名的另一个人，但你不会知道这是个误会或其实是可以改正的。另一方面，如果算法是公开的，大家就能摆弄系统，就会有损模型的有效性。如果算法发现，未付清贷款的人都常常使用某个词，那么大家都会不再使用这个词，这样的话，无论不还贷的概率是多少，申请都能获得批准。

想要低汽车保险费率的人，可以在脸书上列清单；想要获取假释机会的囚犯，可以改变自己手环的大小；不想被逮捕的罪犯，可以整容。人们一旦了解系统，就能摆弄它，这样也会有损系统。这种摆弄现象太常见了，以至于还有个名字——古德哈特定律，取自英国经济学家查尔斯·古德哈特。该定律的内容是：当一项政策成为目标，它就不再是好政策。古德哈特是英格兰银行的经济顾问，他的观点是“设定货币目标会引起人们以损害目标有效性的方式，改变自己的行为”。现在，我们都知道这个定律也被应用于很多其他情况。

苏联的中央规划局下令要求制钉厂生产一定数量的钉子，企业为了降低成本，会尽量生产最小的钉子——也是用处最小的钉子。因此，制定目标会损害目标的有效性。

现在，全美国的学校都以标准考核来管理，该考核用于评估学生、教师和学校。教得好的老师会获得奖金，教得不好的就会被辞退。办得好的学校会吸引学生，获得更多资金来源，办不好的就会被国家接管。因此，老师和学校的回应也在意料之中：实行应试教育，（极端情况下）直接给学生正确答案或篡改答案卡。因此，制定目标会损害目标的有效性。

大学院校也会以不同的方式摆弄系统。我任教的波莫纳学院入选了《美国新闻与世界报道》发布的全球最佳大学排行榜后，申请人数大幅增加，因为《美国新闻与世界报道》为其做了免费宣传。全美国甚至全世界的学生都申请波莫纳学院，因为他们看到它在《美国新闻与世界报道》的排名榜上非常靠前。申请人数的增加，反过来又让波

莫纳学院的筛选越来越严格（现在的录取率不足7%），这又进一步提高了该校在《美国新闻与世界报道》的排行榜上的排名。富者愈富。

知道了这一点后，有些大学便企图摆弄系统。《美国新闻与世界报道》的排名标准可以轻易获得，某些院校便能使些计谋。有些计谋的代价很高，例如，《美国新闻与世界报道》会计算教师薪酬和师生比，但学院要想在这些方面做出改变就要花大价钱。其他衡量指标更容易操控。比如，两个重要的衡量标准是录取率（申请人被接收的比例，越低越好）和入学率（被录取学生选择入校的比例，越高越好）。一所二等小院校的招生办主任告诉我他是如何提高自己院校的这两个指标的：假设该院校有1 000名申请人，想要招200名大一学生。可以将这1 000名申请人按照成绩高低均分为五组，每组含有20%的申请人。如果该学院录取前40%的申请人（录取率为40%），可能只有次优秀的那20%的申请人会入学（即入学率为50%）。最优秀的20%的申请人只是将这所院校当作备选，以防没能进入更好的学校。一旦他们能选择更好的学校，就肯定会去的。

该主任这种势利的策略是拒绝最优秀的那20%，因为他们对该院校来说太优秀了，只录取次优秀的那20%，他们全都会选择入学的。因此，该主任将录取率缩减了一半（从40%降为20%），还使入学率翻倍了（从50%到100%）。那些数字并非像这个虚构例子一样，是预先定好的，但其原理是一样的。主任拒绝了最优秀的申请人，认为如果录取了他们，他们中的大多数人还是不会入校，从而降低了录取率。这种方式让他这所二等院校的筛选更加严格，更值得申请。因此，制定目标会损害目标的有效性。

《美国新闻与世界报道》的另一个重要衡量标准是学生少于20人的班级比例。我知道有所大学利用这个衡量标准来摆弄系统，迅速提升了自己在《美国新闻与世界报道》的排行榜上的排名。

具体而言，假设经济系开设一门导论课，每学期有300名学生选修。该学院可以将这门课分为10个小班，其中9个班的人数上限为19

人，第10个班的人数为129人。那么，这所大学就可以报告说，学校班级少于19人的比例是95%，尽管有43%的学生都挤在一个129人的大班上课。因此，制定目标会损害目标的有效性。

这就是人工智能程序的内在问题，它利用我们目前的行为来预测未来的行为。我们会改变自己的行为：戴不同的手环，访问不同的网站，改变自己的笑容。如果评估贷款申请人的依据是账单支付记录，有些人会及时支付账单，以获取更高的信用评分。不过这也还好，因为贷方就是想要申请人及时还贷。如果计算机根据手机充电频率来评估贷款申请人，有些人会通过频繁充电来摆弄系统。这就不是预测他们是否会还贷的合理指标。

共同毁灭原则

1983年9月6日，苏联的预警卫星侦查到美国的五颗洲际弹道导弹对准了苏联。苏联的协定是，在美国发射导弹摧毁苏联的反制能力前，立即对美国发动核导弹还击。

这就是对策论的军事策略，称作“共同毁灭原则”（mutually assured destruction），具有讽刺意味的是，其首字母缩略词刚好是MAD（疯狂的）。假设两个国家的核武器都能摧毁对方，两个国家都有可信的应对政策，即全力以赴使用核武器来消灭先开火的国家。理论上的对策论均衡就是，没有国家会攻击另一个国家，因为这样会导致双方同归于尽。

如果计算机被编程为一侦查到有导弹来袭就启动反制措施，而且人类无法更改计算机程序，那么全面报复的威胁就完全可信。知道对方肯定会做出灾难性的回应，就没有哪个国家胆敢发动攻击了。

这种无法逆转的计算机化应对方法是斯坦利·库布里克的讽刺电影《奇爱博士》的创作基础，美国和苏联都创建了会被预警系统激活，并且人类无法制止的“世界末日装置”。

幸运的是，当1983年真正出现这种情况时，苏联上校斯坦尼斯拉夫·彼得罗夫干预成功了。他当时正在莫斯科执勤，监测苏联预警系统的信号。当系统显示美国导弹正对准苏联时，彼得罗夫违反了协定，并没有发动反制措施，因为这肯定会导致同归于尽的局面。

正如彼得罗夫所预料的，这只是一次假警报，起因是阳光照在云团上的异常反射让苏联的预警系统产生了误报。若苏联使用的是无法改变的计算机算法，后果可想而知。

这并非第一次（或最后一次）侥幸躲过的核灾难。1995年1月25日，运载科学仪器的四级火箭“黑色布兰特”（Black Brant）在挪威发射。火箭的雷达信号恰巧与美国海军潜艇三叉戟导弹的相似，飞行路径直指莫斯科！苏联的预警系统触发了高级警报，以为这颗导弹的发射目的是引发高纬度爆炸，为美国发动全面核攻击蒙蔽苏联的监控雷达。随后，“黑色布兰特”与第一级引擎分离，雷达信号变成了核弹头的多弹头重入运载器信号。

计算机激活了为俄罗斯时任总统鲍里斯·叶利钦授权核反制准备的核按钮手提箱。叶利钦有5分钟时间考虑是否做出最后的授权指示。就在他犹豫期间，“黑色布兰特”突然转离莫斯科，还没到5分钟就落入海底，未造成任何伤亡。如果没有人类的介入，而是靠计算机算法来决定是否置人于死地，结果会是怎么样呢？

不难想象，未来打仗全靠计算机化设备——全自动飞机、坦克，机器人发射导弹、炸弹、激光器和子弹。向人工智能系统给出对某个人、某个地方或某个物体的描述，指导系统就会在其视觉识别软件匹配成功后发动攻击。还有被称作“致命自主武器”（lethal autonomous weapons, LAWs）的军用机器人。

韩国拥有三星SGR-A1自动哨兵机器人，镇守韩国的非武装地带，这也是世界上军事化程度最高的边境地带。关于这些武器，人们知道的非常少，但有报道指出，它们能侦查到两英里以外的动静，识别和录下目标，发出警告，然后开枪或投掷手榴弹。三星表示，人类应该批准使用SGR机器手枪或手榴弹发射器，但是一些独立的外部团体认为，这些系统会自动开枪，除非受到人类的远程控制。

假设未来的人工智能武器系统将会有内置损失函数，就像检查程序或国际象棋那样，能评估行动前其他行为过程可能带来的后果。该损失函数肯定比棋盘游戏复杂得多，因为它们需要考虑目标的价值、潜在的附带损害（包括消极宣传），以及有机会在其他情况下攻击的可能性。所有这些因素都极其复杂，也会（至少应该）是目标、地点和军事状况特有的。

这并不仅限于军事领域。人工智能软件有时也会对无人驾驶的汽车做出价值判断。为了避免撞到行人或校车，无人驾驶汽车应该转向逆行吗？我不知道怎么回答，但做决定的似乎应该是司机，而不是编写算法或者可能没有考虑这个窘境的程序员。

控制汽车和军用设备的人工智能算法应该计算概率加权期望值，或者最大最小值（选择在最坏情况下造成最小损失的行动）。规则一定具有主观性和争议性，一旦编程结束，那就是板上钉钉的事情了。考虑到军事攻击会摧毁宗教圣地或切断城市用水供应，人类会三思而后行。而计算机才不会在意这些，除非程序员把这些考虑编入它的损失函数，明确告诉程序应该如何评估代价。

人类会犯错，也会造成意料之外的后果，但是将打仗这种事交给机器，任其发挥，这令人毛骨悚然。

结语

我们生活在一个不可思议的历史时期。计算机革命比工业革命给人们的生活带来了更加翻天覆地的变化。我们可以使用计算机来实现过去无法完成的目标，计算机也为我们打开了很多崭新的大门。

我很迷计算机，你可能也有同感。但是，我们不应该让自己对计算机的喜爱，蒙蔽了对它们的局限的认知。没错，计算机储存的事实数据比我们多，记忆力比我们好，计算速度比我们快，还不会像我们那样疲倦。

机器人完成重复单调任务的能力远超人类，如拧螺栓、播种、搜索法律文件、接受银行存款和分配现金。计算机能识别物体、画画和驾车。你肯定还可以想出计算机其他让人惊叹的，甚至是超人类的壮举。

因为计算机能够极其出色地完成任务，所以很容易让人认为它们肯定是高度智能化的。然而，在完成特定任务方面大有用处与拥有通用智能是两码事。通用智能可以将从一次任务中吸取到的教训和习得的技能，运用于更加复杂或完全不同的任务。有了真正的智能，技能便可信手拈来。

计算机非常强大，而且越来越完善，但是计算机算法的设计，仍然是完成定义明确的琐事所需要的、适用范围非常狭窄的能力，而不是像通用智能那样可以通过评估事情现状、起因和后果，来处理不熟悉的情境。人类能够将通用知识运用到特定情境中，再借助特定情境来改善自己的通用知识。如今的计算机还无法做到这一点。

人工智能和人脑的真正智能完全不是一码事。计算机并不知道词语的意思，因为它无法像我们一样感知世界。它不知道真实世界是什么，缺少人类在现实生活中积累所得的常识或智慧；无法构想出有说服力的理论学说，也无法做出归纳推理或长期规划；没有情绪、感觉

和灵感，这些都是创作扣人心弦的诗歌、小说或电影剧本所必不可少的。

或许有一天，计算机会拥有类似人类的真正智能，但这并不是因为计算机内存更大或处理速度更快。这不是量变的问题，而是质变产生的不同方式——找到方法让计算机获取通用智能，使其可以在不熟悉的情境中灵活运用多种方式。

我想澄清一点，这不是在批评计算机科学家。他们都才智过人，也付出了大量辛勤汗水。计算机科学家的工作难度极大，并且大有裨益。还有更多需要完成的工作，难上加难。

模仿人脑是一项艰巨的任务，不能确保一定会成功。不过，还是有一些传奇式的例外，如美国电话电报公司的贝尔实验室、洛克希德·马丁公司的“臭鼬工厂”和施乐公司的帕克研究中心，但是很少有企业愿意支持与脑力有关、短期无回报的研究。一些有用且能立即获利的项目对它们来说更具吸引力。

我不知道，开发出可与人类相媲美的通用智能的计算机需要多长时间。我猜测，至少也需要几十年。可以肯定的是，那些声称计算机已经拥有通用智能的说法都是错的。我也不相信那些人给出的特定日期，如2029年。同时，请保持对牵强附会的科学小说场景的怀疑态度，也小心提防夸大宣传人工智能产品的企业。

挖掘大数据风行一时，但数据挖掘是人为，而非智能。当统计模型分析大量可能的解释变量时，可能关系的数量就会暴增。有1 000种可能解释变量的多元回归模型，10个输入变量存在近1万亿个万亿的可能组合。若有1万个可能解释变量，则10个输入变量存在超过10亿万亿个万亿的可能组合。难以想象，若有100万个可能解释变量，会存在多少种可能组合。

如果把很多可能变量都考虑在内，即便所有都只是随机噪声，部分组合也一定与我们试图预测的对象高度相关，如癌症、信用风险和

岗位适用性。偶然会出现真正的“知识发现”，但是，考虑的解释变量越多，所发现关系只是偶然出现且转瞬即逝的可能性就越大。

统计学证据不足以辨别真知灼见和虚假信息。只有逻辑、智慧和常识才能对其加以区分。计算机无法评估事物是真正相关还是偶然相关，因为计算机不理解数据的意义。数字不过是数字而已。计算机并没有区分好坏数据所需的人类判断力，没有分辨有理有据的和虚假伪造的统计学模型所需的人类智能。如今的计算机能通过图灵测试，却无法通过史密斯测试。如果所发现的模式被隐藏在黑匣子里，让模型难以理解，就会使这种情况恶化。无人知晓为什么计算机算法决定要买入这只股票、拒绝这名求职者、给病患开这种药、拒绝这名囚犯的假释请求和轰炸这座建筑。

在大数据时代，真正的危险不是计算机比我们更聪明，而是我们自己这么认为，从而信任计算机为我们做出重要决定。我们不应该认为计算机就是万无一失的、数据挖掘都是“知识发现”，以及黑匣子也应该被信赖。我们要相信自己能判断统计学模型是否合理，有无可能派上用场，抑或它只是偶然出现而已，转瞬即逝，毫无用处。

人类推理与人工智能有天壤之别，这也是为什么如今更显人类推理能力的可贵之处。

参考文献

1938 New England hurricane. (n.d.). In Wikipedia. Retrieved April 29, 2015, from; http://en.wikipedia.Org/wiki/1938_New_England_hurricane#cite_note-3.

Alba, Davey. 2011. How Siri responds to questions about women’s health, sex, and drugs, Laptop, December 2.

Alexander, Harriet. 2013. ‘Killer Robots’ could be outlawed,The Telgraph, November 14.

Allan, Nicole, Thompson, Derek. 2013. The myth of the student-loan crisis, The Atlantic, March.

Anderson, Chris. 2008. The end of theory, will the data deluge make the scientific method obsolete?, Wired, June 23.

Andrew, Elise. undated, AI trying to design inspirational posters goes horribly and hilariously wrong, IFLScience, Available from : http://www.livemint.com/Technology/VXCMwOVfilawOaIInDlv2O/When-artificial-inteIligence-goes-wrong.html.

Angwin, Julia, Larson, Jeff, Mattu, Surya, Kirchner, Lauren. 2016. Machine bias, ProP-ubhca, May 23. Available from:https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Angwin, Julia, Scheiber, Noam, Tobin, Ariana. 2017. Machine bias： dozens of companies are using facebook to exclude older workers from job ads, ProPublica December 20., Available from: https://www.propublica.org/article/facebook-ads-age-discrimination-targeting.

Anonymous. 2015. A long way from dismal： economics evolves.The Economist, 414 (8920), 8.

Baum, Gabrielle, Smith, Gary. 2015. Great companies： looking for success secrets in all the wrong places, Journal of Investing, 24 (3), 61-72.

Bem, DJ. 2011. Feeling the future： experimental evidence for anomalous retroactive influences on cognition and affect, Journal of Personality and Social Psychology, 100(3), 407-25.

Berk, R. 2013. Algorithmic criminology, Security Informatics, 2： 5. https：//doi.org/10.1186/ 2190-8532-2-5.

Bolen, Johan, Mao, Huina, Zeng, Xiaojun, 2011. Twitter mood predicts the stock market, Journal of Computational Science, 2 (1), 1-8.

Boyd, Danah, Crawford, Kate. 2011. Six provocations for big data. A Decade in Internet Time： Symposium on the Dynamics ofthe Internet and Society, September 21, 2011.

Brennan-Marquez, Kiel. 2017. Plausible cause ： explanatory standards in the age ofpowerful machines, Vanderbilt Law Review, 70 (4), 1249-1301.

Brodski, A, Paasch, GF, Helbling, S, Wibral, M. 2015. Ihe faces of predictive coding. Journal of Neuroscience, 35 (24)： 8997.

Calude, Cristian S, Longo, Giuseppe. 2016. The deluge of spurious correlations in big data, Foundations of Science, 22, 595-612. https://doi.org/10.1007/s10699-016-9489-4.

Cape Cod Times, July 7,1983.

Caruso, EM, Vohs, KD, Baxter, B, Waytz, A. 2013. Mere exposure to money increases endorsement offree-market systems and social inequality. Journal of Experimental Psychology; General, 142, 301-306. http://dx.doi.org/10.1037/a0029288.

Chappell, Bill. 2015. Winner of French Scrabble title does not speak French, The Two-Way: Breaking News From NPR.

Chatfield, Chris. 1995. Model uncertainty, data mining and statistical inference, Journal of the Royal Statistical Society A 158,419-466.

Chollet, Francois. 2017. Deep Learning With Python. Manning Publications.

Christensen, B, Christensen, S. 2014. Are female hurricanes really deadlier than male hurricanes? Proceeding ofthe National Academy of Sciences USA ,111 (34) E3497-98.

Coase R, 1988. How should economists choose?, in Ideas, Their Origins and Their Consequences : Lectures to Commemorate the Life and Work of G. Warren Nutter, American Enterprise Institute for Public Policy Research.

Cohn, Carolyn, 2016, Facebook stymies Admirals plans to use social media data to price insurance premiums, Reuters, November 2. Available from : https : //www.reuters.com/article/us-insurance-admiral-facebook/facebook-stymies-

admirals-plans-to-use-social-media-data-to-price-insurance-premiums-idUSKBN12X1WP.

Dalal, SR, Fowlkes, EB, Hoadley, B. 1989. Risk analysis ofthe space shuttle： pre-Challenger prediction offailure, J ournal ofthe American Statistical Association . 84 (408), 945-57.

Davis, Ernest, 2014. The technological singularity: the singularity and the state of the art in artificial intelligence. Ubiquity, Association for Computing Machinery. October. Available from : http：//ubiquity.acm.org/article.cfm?id=2667640.

Devlin, Hannah, 2015. Rise ofthe robots： how long do we have until they take our jobs?, The Guardian, February 4.

Diamond, Jared, 1989. How cats survive falls from New York skysrcapers, Natural History, 98, 20-26.

Duhigg, Charles, 2012. How companies learn your secrets,The New York Times Magazine. February 16.

Effects of Hurricane Sandy in New York. (n.d.) In Wikipedia. Retrieved April 29, 2015 , from： http://en.wikipedia.org/wlki/EfFects_of_Hurricane_Sandy_in_New_York.

Evtimov, I, Eykholt, K, Fernandes, EKohno, T, Li, B, Prakash, A. etal. 2017. Robustphysical-world attacks on deep learning models. <https：//arxiv.org/abs/1707.08945>

Fayyad, Usama, Piatetsky-Shapiro, Gregory, Smyth, Padhraic. 1996. From data mining to knowledge discovery in databases, AI Magazine, 17 (3), 37-54.

Galak, J, LeBoeuf, RA, Nelson, LD, & Simmons, JP. 2012. Correcting the past: failures to replicate Psi. Journal of Personality and Social Psychology, 103(6), 933-48.

Garber, Megan, 2016. When algorithms take the stand,The Athantic, June 30.

Gasiorowska, Agata, Chaplin, Lan Nguyan, Zaleskiewicz, Tomasz, Wygrab, Sandra, Vohs, Kathleen D. 2016. Money cues increase agency and decrease prosociality among chil-dren: early signs of market-mode behaviors, Psychological Science, 27(3), 331-44.

Gebeloff, Robert, Dewan, Shaila. 2012. Measuring the top 1% by wealth, not income,New York Times, January 17.

Gemini. Is more education better? 2014. Degree Council, April 18. Available from : http : // degreecouncil.org/2014/is-more-education-better/.

Goldmacher, Shane. 2016. Hillary Clinton's "Invisible Guiding Hand" Politico, September 7.

Goldman, William, 1983. Adventures in the screen trade, Los Angeles : Warner Books. Greene, Lana. 2017. Beyond Babel : The limits of computer translations, The Economist,January 7. 422 (9002), 7.

Herkewitz, William. 2014. Why Watson and Siri are not real AI, Popular Mechanics, February 10.

Hirshleifer, D, Shumway, T. 2003. Good day sunshine: Stock returns and the weather.

Journal of Finance, 58(3) 1009-32.

Hou, Kewei, Xue, Chen, Zhang Lu. 2017. Replicating anomalies, NBER Working Paper No. 23394, May.

Hoerl, Arthur E, Kennard, Robert W 1970a. Ridge regression： Biased estimation for nonorthogonal problems. Technometrics, 12, 55-67.

Hoerl, Arthur E, and Kennard, Robert W. 1970b. Ridge regression: Applications to nonorthogonal problems. Technometrics, 12, 69-82.

Hofstadter, Douglas. 1979. Gödel, Escher, Bach： An Eternal Golden Braid, New York： Basic Books.

Hofstadter, Douglas, Sander, Emmanuel. 2013, Surfaces and Essences： Analogy as the Fuel and Fire of Thinking, New York： Basic Books.

Hornby, Nick, 1992. Fever Pitch, London： Victor Gollancz Ltd., p. 163.

Hurricane Allen. (n.d.). In Wikipedia. Retrieved April 29,2015, from： http：//en.wikipedia.

org/wiki/Hurricane_Allen#cite_note-1983_Deadly-18.

Hvistendahl, Mara. You are a number, Wired, January 2018, 48-59.

Ip, Greg. 2017. We survived spreadsheets, and we’ll survive AI, The Wall Street Journal, August 3.

Issenberg, Sasha. 2012. How Obama used big data to rally voters, Part 1, MIT Technology Review.

Johnson P. 2013. 75 years after the Hurricane of 1938, the science of storm tracking improved significantly. Mass

Live, September 14. Available from： http：//www.masslive.com/news/mdex,ssf/2013/09/75_years_after_the_hurricane_o.html, Retrieved April 29,2015.

Jung, K, Shavitt, S, Viswanathan, M, Hilbe, JM. 2014. Female hurricanes are deadlier than male hurricanes. Proceedings ofthe National Academy of Sciences. USA 111 (24), 8782-7. Kendall, MG. 1965. A Course in Multivariate Statistical Anaalysis. Third Edition, London： Griffin.

Khomami, Nadia. 2014. 2029: the year when robots will have the power to outsmart their makers The Guardian, February 22.

Knight, Will. 2016. Will AI-powered hedge funds outsmart the market?, MIT Technology Review, February 4.

Knight, Will. 2017. The financial world wants to open AI’s black boxes, MIT Technology Review, April 13.

Knight, Will. 2017. There’s a dark secret at the heart of artificial intelligence： no one really understands how it works, MIT Technology Review, April 11.

Knight, Will. 2017. Alpha Zero’s alien chess shows the power, and the peculiarity, of AI, MIT Technology Review, December 8.

Labi, Nadia, 2012. Misfortune Teller,The Atlantic, January/February 2012.

Lambrecht, Anja, Tucker, Catherine E. Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads (November 30, 2017).

Available at SSRN: https://ssrn.com/abstract=2852260 or http://dx.doi.org/ 10.2139/ssrn.2852260.

Lazer, David, Kennedy, Ryan, King, Gary, Vespignani, Alessandro. 2014. The parable of Google flu: traps in big data analysis, Science, 343 (6176), 1203-5.

LeClaire, Jennifer. 2015. Moores Law turns 50, creator says it won't go on forever, Newsfac-tor, May 12.

Leswing, Kif. 2016. 21 ofthe funniest responses you'll get from Siri, Business Insider, March 28.

Lewis-Kraus, Gideon. 2016. The great AI awakening, The New York Times Magazine, December 14.

Liptak, Adam. 2017. Sent to prison by a software program's secret algorithms, New York Times, May 1.

Loftis, Leslie. 2016. How Ada let Hillary down, Arc Digital, December 13.

Madrigal, Alexis. 2013. Your job, their data: the most important untold story about the future, Atlantic, November 21.

Maley, S. 2014. Statistics show no evidence of gender bias in the public's hurricane preparedness. Proceedings ofthe National Academy of Sciences, 111 (37) E3834. USA. http://dx.doi.org/10.1073/pnas.1413079111.

Malter, D. 2014. Female hurricanes are not deadlier than male hurricanes. Proceeding ofthe National Academy of Sciences. USA ,111 (34) E3496. http://dx.doi.org/10.1073/pnas.1411428111.

Marquardt, Donald W. 1970. Generalized inverses, ridge regression, biased linear estimation. Technometrics, 12, 591-612.

Massy, William F. 1965. Principal components in exploratory statistical research. Journal ofthe American Statististical Association , 60,234-56.

McLean, R. David, Pontiff Jeffrey. 2016. Does academic research destroy stock return predictability?, Journal of Finance, 71 (1), 1540-6261.

Metz, Cade. 2016. Trump’s win isn’t the death of data — it was flawed all along, Wired, November 9.

Minkel, J R. 2007. Computers solve checkers — it’s a draw,Scientific American, July 19.

National Oceanic and Atmospheric Administration. 2012. Predicting hurricanes： times have changed. Available from: http://celebrating200years.noaa.gov/magazine/devast_hurricane/welcome.html.

Nguyen, Anh, Yoslnski, Jason, Clune Jeff. 2015. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images, Proceedings ofthe IEEE Conference on Computer Vision and Pattern Recognition. Available from: https://arxiv.org/ abs/1412.1897v4.

Otis, Ginger Adams. 2014. Married people have less heart problems than those who are single, divorced: study, Daily News, March 28.

Panesar, Nirmal, Graham, Colin. 2012. Does the death rate of Hong Kong Chinese change during the lunar ghost month?

Emergency Medicine Journal 29, 319-21.

Panesar, Nirmal S, Chan, Noel CY, Li, Shi N, Lo, Joyce KY, Wong, Vivien WY, Yang, Isaac B, Yip, Emily KY. 2003. Is four a deadly number for the Chinese?, Medical Journal of Australia, 179 (11): 656-8.

Peck, Don. 2013. They're Watching You at Work, Atlantic, December.

Pektar, Sofia. 2017. Robots will wipe out humans and take over in just a few centuries warns Royal astronomer, Sunday Express, April 4.

Poon, Linda. 2014. Sit more, and you're more likely to be disabled after age 60, NPR, February 19.

Ruddick, Graham. 2016. Admiral to price car insurance based on Facebook posts, The Guardian, November 1.

Reese, Hope. 2016. Why Microsoft's 'Tay' AI bot went wrong, Tech Repubhc, Available at: http://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong, March 24.

Reilly, Kevin. 2016. Two ofthe smartest people in the world on what will happen to our brains and everything else, Business Insider, January 18.

Ritchie, Stuart J, Wiseman, Richard, French, Christopher C, Gilbert, Sam. 2012. Failing the future : three unsuccessful attempts to replicate bem's 'retroactive facilitation of recall' effect, PLoS ONE. 7 (3): e33423.

Robbins, Bruce, Ross, Andrew. 2000. Response： Mystery science theatre, in： The Sokal Hoax: The Sham that Shook the Academy, edited byAlan D. Sokal, University of Nebraska Press, pp. 54-8.

Rudgard, Olivia. 2016. Admiral to use Facebook profile to determine insurance premium, The Telegraph, November 2.

Rutkin, Aviva Hope. 2017. The tiny changes that can cause AI to fail, BB C Future, April 17.

Sharif, Mahmood, Bhagavatula, Sruti, Bauer, Lujo, Reiter, Michael K. 2016. Accessorize to a Crime： Real and Stealthy Attacks on State-of-the-Art Face Recognition, Proceedings ofthe 2016 ACM SIGSAC Conference on Computer and Communications Security , 1528-40.

Siegel, Eric, 2016. How Hillarys campaign is (almost certainly) using big data, Scientific American, September 16.

Simonite, Tom. 2017. How to upgrade judges with machine learning, MIT Technology Review, March 6.

Singh, Angad. 2015. The French Scrabble champion who speaks no French, CNN, July 22.

Smith, Gary, Campbell, Frank, 1980. A critique of some ridge regression methods, Journal ofthe American Statistical Association , with discussion and rejoinder, 75 (369), 74-81.

Smith, Gary, 1980. An example of ridge regression difficulties, The Canadian Journal of Statistics, 8 (2) 217-25.

Smith, Gary, 2011. Birth month is not related to suicide among major league baseball players, Perceptual and Motor Skills, 112(1), 55-60.

Smith, Gary, 2012. Do people whose names begin with d really die young?, Death Studies, 36 (2), 182-9.

Smith, Gary, Zurhellen, Michael. 2015. Sunny upside? The relationship between sunshine and stock market returns, Review of Economic Analysis, 7, 173-83.

Smith, Gary. 2014. Standard Deviations : Flawed Assumptions, Tortured Data, and Other Ways to Lie With Statistics, New York: Overlook.

Smith, Stacy, Allan, Ananda, Greenlaw, Nicola, Finlay, Sian, Isles, Chris. 2013. Emergency medical admissions, deaths at weekends and the public holiday effect. Cohort study, Emergency Medicine Journal, 31(1):30-4.

Sokal, A. 1996. Transgressing the boundaries: Towards a transformative hermeneutics of quantum gravity. Social Text, 46/47,217-52.

Somers, James. 2013. The man who would teach machines to think, The Atlantic, November.

Stanley, Jason, and Vesla Waever, 2014. Is the United States a Racial Democracy?, New York Times, Online January 12.

Su, Jiawei, Vargas, Danilo Vasconcellos, Kouichi, Sakurai. 2017. One pixel attack for fooling deep neural networks, November 2017. Available from : https://arxiv.org/abs/ 1710.08864.

Szegedy, Christian, Zaremba, Wojciech, Sutskever, Ilya, Bruna, Joan, Erhan, Dumitru, Goodfellow, Ian, Fergus, Rob. 2014. Intriguing properties of neural networks, Google, February 19. Available from : https://www.researchgate.net/publication/259440613_Intriguing_properties_of_neural_networks

Tal, Aner, Wansink, Brian. 2014. Blinded by science：trivial scientific information can increase our sense oftrust in products, Public Understanding of Science. 25,117-25.

Tashea, Jason. 2017. Courts are using AI to sentence criminals. that must stop now. Wired, April 17.

Tatem, Andrew J, Guerra, Carlos A, Atkinson, Peter M, Hay, Simon I. 2004. Athletics： momentous sprint at the 2156 Olympics?, Nature, 431, 525.

Todd, Chuck, Dann, Carrie. 2017. How big data broke American politics, NBC News, March 14.

Traxler MJ, Foss DJ, Podali R, Zirnstein M. 2012. Feeling the past： the absence of experimental evidence for anomalous retroactive influences on text processing, Memory and Cognition. 40(8), 1366-72.

Vadillo, Michael A, Hardwicke, Tom E, Shanks, David R., 2016. Selection bias, vote counting, and money-priming effects: a comment on Rohrer, Pashler, and. Harris (2015) and Vohs (2015), Journal of Experimental Psychology. General, 145(5), 655-63.

Vohs KD, Mead, NL, Goode MR. 2006. The psychological consequences of money. Science, 314,1154-6.

http://dx.doi.org/10.1126/science.1132491.

Vorhees, William. 2016. Has AI gone too far? Automated inference of criminality using face images, Data Science Central, November 29.

Wagner, John. 2016. Clinton's data-driven campaign relied heavily on an algorithm named Ada. What didn't she see?, Washington Post, November 9.

Walsh, Michael. 2017. UPDATE: A.I. inspirational poster generator suffers existential breakdown. Nerdist, June 27. http://nerdist.com/a-i-generates-the-ridiculous-inspirational-posters-that-we-need-right-now/

Weaver, John Frank. 2017. Artificial intelligence owes you an explanation. Slate, May 8.

Wiecki, Thomas, Campbell, Andrew, Lent, Justin, Stauth, Jessica. 2016. All that glitters is not gold: comparing backtest and out-of-sample performance on a large cohort of trading algorithms, Journal of Investing, 25 (3), 69-80.

Wilier, Robb. 2004. The Intelligibility of Unintelligible Texts. Master's thesis. Cornell University, Department of Sociology.

Willsher, Kim. 2015. The French Scrabble champion who doesn't speak French, The Guardian, July 21.

Winograd, Terry. 1972. Understanding natural language. Cognitive Psychology, 3,1-191.

Wolfson, Sam. 2016. What was really going on with this insurance company basing premiums on your facebook posts?,

Vice, November 2.

Woollaston, Jennifer. 2016. Admiral’s firstcarquote may breach Facebook policy by using profile data for quotes，Wired UK, November 2.

Wu, Xiaolin, Zhang, Xi. 2016. Automated inference on criminality using face images, Shanghai Jiao Tong University, November 21. Available at: https://arxiv.org/abs/1611.04135v1.

Wu, Xiaolin, Zhang Xi. 2017. Responses to critiques on machine learning of criminality perceptions, Shanghai Jiao Tong University, May 26. Available at: https://arxiv.org/abs/1611.04135v3.

Yuan, Li. 2017. Want a loan in China? Keep your phone charged, The Wall Sfreef Journal, April 6.

Yudkowsky, Eliezer. 2008. Artificial intelligence as a positive and negative factor in global risk. In Global Catastrophic Risks, edited by Nick Bostrom and Milan M. Cirkovic, New York： Oxford University Press, 308-45.

Zeki, Semir, Romaya, John Paul, Benincasa, Dionigi MT, Atiyah Michael F. 2014. The experience of mathematical beauty and its neural correlates, Frontiers in Human Neuroscience, February 13. 8, article 68.

Zuckerman, Gregory, Hope, Bradley. 2017. The quants run Wall Street now, The Wall Street Journal, May 21.

图书在版编目（CIP）数据

错觉 ： AI如何通过数据挖掘误导我们 / （美）加里·史密斯著 ； 钟欣奕译. -- 北京 ： 中信出版社，2019.11
书名原文: The AI Delusion
ISBN 978-7-5217-0995-7
Ⅰ . ①错… Ⅱ . ①加… ②钟… Ⅲ . ①人工智能－研究 Ⅳ . ① TP18
中国版本图书馆CIP数据核字(2019) 第197243号

The AI Delusion by Gary Smith
Copyright © Gary Smith 2018.
The AI Delusion was originally published in English in 2018. This translation is published by arrangement with
Oxford University Press. CITIC Press Corporation is solely responsible for this translation from the original work and Oxford University Press shall have no liability for any errors, omissions or inaccuracies or ambiguities in such translation or for any losses caused by reliance thereon.
Simplified Chinese translation copyright © 2019 by CITIC Press Corporation
ALL RIGHTS RESERVED

错觉：AI如何通过数据挖掘误导我们
著者：［美］加里·史密斯
译者：钟欣奕
出版发行：中信出版集团股份有限公司
（北京市朝阳区惠新东街甲4号富盛大厦2座 邮编100029）

字数：220千字
版次：2019年11月第1版
京权图字：01-2019-5497
广告经营许可证：京朝工商广字第8087号
书号：ISBN 978-7-5217-0995-7

版权所有·侵权必究

人工智能的进

计算机思维离人类心智还有多远？

COMMONSENSE, THE TURING TEST, AND THE QUEST FOR REAL AI

让我们重回人工智能的起点，
检视人工智能的最初梦想。

化

[加] 赫克托·莱韦斯克 著

王佩 译

人工智能有常识吗？它能否像我们一样随机应变？

中信出版集团

人工智能的进化

[加] 赫克托·莱韦斯克 著
王佩 译

中信出版集团

目录

前言

第1章 什么是人工智能?

自适应机器学习

老式人工智能

具有常识的程序

图灵测试

中文房间理论

第2章 巨型拼图之谜

疑问接踵而来

留给我们的困难

一条解决的途径

第3章 知识与行为

超越刺激与反应

知识与信念

意向立场

智能行为

能力与表现

第4章 真智能还是假智能?

谈话机器人

投机取巧不可行

威诺格拉德模式
我们从中得到的经验
GOFAI的回归

第5章 有经验的学习与没有经验的学习

我们如何学习词语?
我们如何学习事实?
我们如何学习行为?
我们如何超越经验?

第6章 书本智慧与市井智慧

语言的影响力
书本智慧
海伦·凯勒的智慧
书本中的市井智慧

第7章 长尾理论与培训的界限

长尾现象带来的难题
如何处理突发事件
无意识与有意识
威诺格拉德模式测试
是不是技巧还不够完美?

第8章 符号与符号处理

符号的代数运算
符号的逻辑运算
符号的意义
一切都源自图灵机

第9章 基于知识的系统

符号能够解决的问题

用符号表示无形
知识表示假说
假设是真的吗?
知识表示与推理
谁才是唯一的选择?

第10章 人工智能技术应用

人工智能的未来
自动化是好是坏?
超级智能与奇点
真正的风险
超越进化

致谢

注释

参考文献

献给

帕特

前言

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

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

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

这也正是本书的主要内容。

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

究其原因，我们不妨先考虑一下金鸡独立这个动作。我们有多希望能将思维、规划以及解决问题的能力融为一体？人类个体能否做到这点呢？我们是否认为事先对这一主题进行研究会有所帮助？比如看看类似《傻瓜单腿站立指南》这种书？如果一位人工智能研究人员想

造出能单腿站立的机器人（不管出于任何原因），他绝对认为涉及思维的人工智能没什么借鉴意义。

当今研究人工智能的众多流派中，绝大部分都已背离了20世纪50年代开始的对于思维（和规划、解决问题）的原始研究。包括研究方向和研究目标等涉及该研究的基本理念都已发生改变，而且这种变化在最近兴起的机器学习当中尤为突出。毫无疑问，从纯技术的角度来看，这种变化取得了巨大成功，而且极有可能超越人工智能的其他流派。虽然对机器学习的研究极度依赖数据统计，但它的本质与人工智能的原始研究仍然大不相同。

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

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

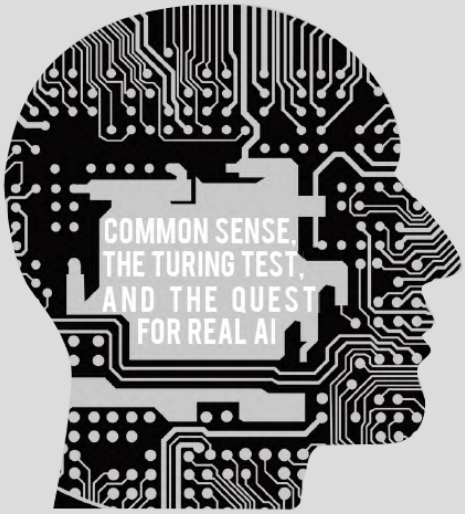
本书全部文本只用了大约150kb（千字节）的存储空间，看到这么一本小书，让人多少有些沮丧。相比而言，我电脑里一秒钟的维瓦尔第音乐会视频所占的空间都比它的两倍还大。因此，单纯从原始数据来看，我电脑里半秒钟的视频与本书分量相当。但是在半秒钟的时间里，乐队指挥都来不及举起指挥棒。

正如古语所说：一画（帧）胜千言。然而，我认为还有另外一点值得注意，与图画相比，语言是更为紧凑的信息载体。几百个字呈现在视频当中，或许只是一瞬间，却也能将炖牛肉的方法解释清楚。由于人类的进化，我们已经能够随心所欲地使用这些超紧凑型的信息载

体。我们通过在他人耳边低语，就能对其日后的行为产生影响。人类是如何做到这一点的，正是我们现在所探究的问题。我希望读者能够喜欢我的观点，也能感受到我在探究过程中的激动心情。

第1章

什么是人工智能？



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

自适应机器学习

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

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

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

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

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

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

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

人们发现，AML在这种所谓“无监督”的环境下运转极其良好，完全超出几十年前人工智能研究人员做出的预测。AML的成功可以归结于以下三点：拥有可供分析处理的海量数据（通过在线、专业存储设备

或传感装置获取），掌握功能强大的能够处理这些数据的计算技术，以及高速计算机。而这些条件，在30年前并不具备。

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

老式人工智能

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

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

什么是常识？拥有常识与训练有素地分析海量数据有何不同？本书将用大量篇幅回答上述问题。阻碍我们将此类差别梳理清楚的困难之一在于，在日常生活中，我们几乎所有的时间都是在自己相对熟悉的场景中度过，依靠长期形成的习惯或行为模式生活。比如我通常会

在上班路上买一杯咖啡，在大厅里遇到同事乔治时和他打招呼，吃热狗时会放上芥末。即使偶尔看到一棵从未见过的树，我基本也是安之若素，毫不在意。生活就是这样平淡无奇。

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

研究人员在研究老式人工智能（GOFAI）的过程中经常遇到的问题是：如何对这种常识做出解释？人们如何利用这种常识搞清楚当日常行为模式行不通时该怎么办？到底何为“搞清楚”？遇到一扇打不开的门时，一个人需要经历何种心路历程才最终做出决定该如何处理？只因有人对我说了几句话，我就决定在寒冷的冬日不穿大衣来到户外，这是怎么回事？只因有人发出某种声音或说出某些话语，就足以让我的行为举止一反常态？

这就是令早期人工智能科研人员心驰神往的地方，这就是本书的主要内容。

具有常识的程序

1956年夏天，人工智能的先驱们在美国达特茅斯学院的一次会议上相遇。此次会议的组织者、美国计算机科学家约翰·麦卡锡（1927—2011）首次使用了人工智能这个说法。他于1958年发表了一篇题为

“具有常识的程序”的论文，这篇文章内容精彩，开创性地为人工智能的研究确立了方向。

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

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

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

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

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

实际上，也有人批评老式人工智能过分专注于符号和文字。批评者明确指出，智能行为并非全都需要通过语言表达。你可能无法准确说出骑摩托车通过急弯道时车身应该倾斜多少度，但这并不妨碍你成为骑行高手。你能在一群男孩当中辨认出一对亲兄弟，但不一定能够说清如何发现他们之间的相似之处。

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

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

图灵测试

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

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

1950年，图灵在一篇极具影响力的论文中建议，我们应该将注意力从如何制造机器、机器外观如何以及其内部如何运转等方面转移到可观察的外部行为（externally observable behavior）上。当然，

我们需要考虑的行为多种多样，我们可以将机器放在并不熟悉的新环境当中，看它作何反应，我们可以要求它辨认猫的照片，我们还可以研究它如何处理撒满调料的热狗。图灵提出的想法是让机器与询问者进行一次不受束缚、毫无保留的对话。

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

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

询问者：在莎士比亚十四行诗的第一行，写着“我能否将你比作夏日？”在这里如果用“春日”来代替“夏日”，岂不是更好？

计算机：那不押韵。

询问者：那用“冬日”怎么样？这可以押韵了吧？

计算机：是的，但是没人愿意被比作冬日。

询问者：是匹克威克先生提醒你圣诞节到了吗？

计算机：某种程度上，是的。

询问者：但圣诞节是在冬日，我觉得匹克威克先生不会介意这么对比。

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

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

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

中文房间理论

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

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

假设现有一款计算机程序能够通过图灵测试，即可以完全不受任何限制地进行人机对话。我们再假设这次对话使用的语言不是英语，

而是汉语。因此，根据模仿游戏的规则，汉字（通过某种编码的形式）被输入该程序，随后该程序能够用汉语（以某种编码的形式）应答，作为输出。即使是精通汉语的人，仍然长期无法分辨输出信息的是机器还是人。

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

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

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

这就是中文房间理论的大概内容。

人们对此理论的反响如何呢？人工智能领域有人对此并不买账。这也可以理解，大家对人工智能领域中的哪种技术阻碍此类程序（假设其存在）的开发更感兴趣，而对我们在一个房间里用一个人就能准

确地模拟这种程序的运行（虽然速度上，人可能只是程序的几百万分之一），并得出相关结论兴趣不大。

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

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

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

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

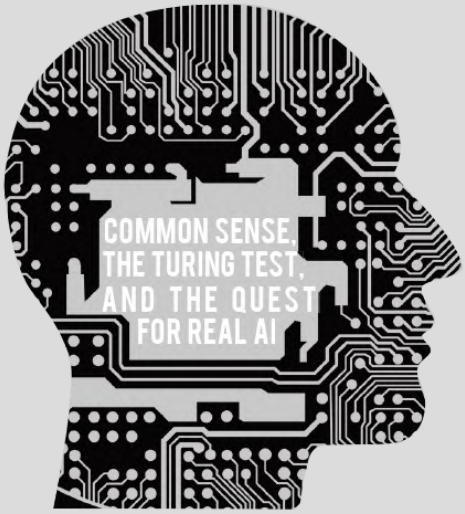
要进行测试，一本介绍加法运算表的英语小册子就堪当此任：首先做一个10×10的个位数加法表，然后进行两位数加法运算（可进位数），最后是多位数加法运算。这样的小册子绝对可以存在，而且只需几页纸就能说明问题。这种小册子的意义在于，对任何不会加法运算的人来说，只要能够记住书中的说明就能学会！

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

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

第2章

巨型拼图之谜



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

疑问接踵而来

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

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

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

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

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

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

还有其他种种说法。

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

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

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

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

那么本书将讨论拼图的哪一部分呢？这里先做一个简要介绍：

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

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

这些问题都提得非常好，但我不会在本书中对此过多纠缠。然而，我也不想让读者觉得剩下的问题都无关紧要。因为即使从最狭义的角度来看，人类的思维过程仍然和人类的想法一样变幻莫测、难以捉摸，而且我们可以对于任何事情进行思考！虽然人类迄今已在科技

领域取得惊人的成就，但对于自身在思考谁会获得奥斯卡最佳男主角，在分辨法国作曲家德彪西（Debussy）和拉威尔（Ravel）之间音乐的异同，以及在考虑是否需要对自由市场加以规范等问题的过程中，大脑经历了何种变化，我们依然不得而知。

即使在拼图的微小碎片中，仍存在着许多未解之谜。

留给我们的困难

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

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

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

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

那么为什么会出现这样的结果呢？

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

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

```
integer U, V, W, X, Y

set U to 0

set V to 0

repeat the following forever:

  set W to the next typed number

  set X to W / 10

  if X > U

    then set Y to X*X + (W mod 10)

    else set Y to U*U + (W mod 10)

  beep (Y / 10) times
```

```
    flash (Y mod 10) times

    set U to V

    set V to X

end repeat
```

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

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

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

为了真正模拟心理学家的工作，我们必须假设参与测试的输入方式要比实际可以输入机器的输入方式少得多。假设进行一场阅读理解测试（第4章将会详细谈到），现在将该测试中的句子数量和被试者一

生当中会读到的句子数量进行对比。心理学必须符合行为空间中一个个琐碎细节所组成的证据。

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

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

简而言之，这就是心理学的困难所在。

设计出一种通过有效控制各种变量，进而得到有效结果的测试非常困难。假设我的研究对象是张三和李四，他们有着截然不同的生活，看到过截然不同的事物，带着截然不同的信念和目的参与这项测试，那么我该如何控制测试中的变量？众所周知，能够有所启示的心理学试验往往涉及知觉任务，需要在毫秒之内做出反应。由于速度太快，长时记忆这种在张三和李四身上会截然不同的因素，就不那么重要了。如果被试者能悠闲地坐着，并沉思几秒，那么再想要控制所有变量就难如登天了。

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

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

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

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

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

但问题是，我们能否通过研究组成M的元件的属性找到M的行为规律。比如，我们可能想得到M的第一位数字，而不是第二位数字的平方。但我们能在这些电子元件中看到吗？

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

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

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

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

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

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

简而言之，这就是神经科学的困难所在。

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

我们只能另寻他法。

一条解决的途径

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

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

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

飞行的原理。

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

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

当然，设计立场这种方法无法解决整个谜题。比如，这种方法不会告诉我们某种现象如何成为某种演化过程的最终结果，也不能用于解释无法观察的现象。如果我们想知道的不是鸟类如何飞行，而是鸟类是否具有某种从不显现的内部感觉，设计立场就无能为力了。

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

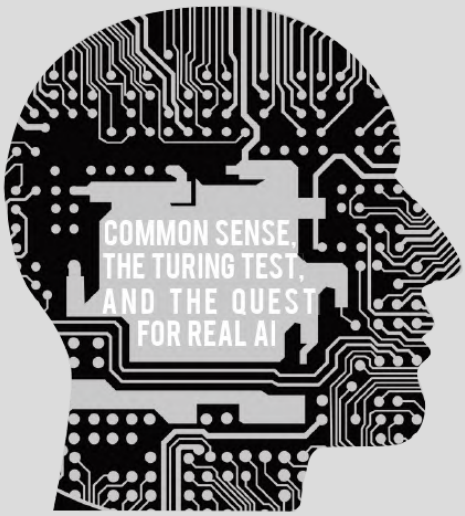
在很多研究人员看来，研究应集中在有意识思维的主观感受（或称为感受性）上，这也是人类心智的显著特征，非常有趣。我将此称为巨型拼图问题。毫无疑问，对于意识的感知能力是这个拼图里非常有趣的部分，但不是唯一的有趣部分。

---

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

第3章

知识与行为



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

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

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

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

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

超越刺激与反应

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

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

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

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

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

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

> 我认为，该思维过程的重要特征如下：当发现现有证据或信息存在空白或者不完整时，思维过程就会启动。随后，空白得到填补，缺失的信息被补充完整。该过程能够顺利完成，依靠的是在原有基础上对与证据相关的内容进行扩展和补充。除了思维启动时涉及的信息以外，在扩展和补充的过程中，会使用来自其他渠道的信息。［《思维》（*Thinking*），1958年，第75页］

决定因素就是这些“来自其他渠道的信息”。

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

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

这个故事告诉我们：

> 掌握知识，并利用知识对自身行为施加影响，是构成人类行为的本质。

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

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

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

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

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

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

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

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

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

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

总而言之，与其他动物一样，人类会采取行动以应对刺激。在某些情况下，映射是直接的：我们感受到一些东西，并对此有所反应。

但是，对于很多人来说，映射没那么直接：我们感受到一些东西，产生的反应却取决于我们拥有的信念。

知识与信念

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

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

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

“约翰希望玛丽能来参加派对”和“约翰担心玛丽会来参加派对”这两句话也是同样的道理，语句中包含同样的命题，但命题与约翰的关系是不同的。诸如“知道”“希望”“后悔”“恐惧”“疑惑”等动词都表明了哲学家所说的命题态度，即行为人与命题之间的关系。无论何种情况，对于命题来说最重要的就是它的真值条件，即

如何为真：如果约翰希望玛丽能来参加派对，那么根据这个命题，约翰就是希望世界以某种方式运转。

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

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

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

意向立场

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

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

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

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

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

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

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

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

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

在知识型系统中，信念也是如此。

智能行为

什么是智能行为？如果我们要研究智能行为，那么它的定义是什么？因为我们所做的一切并不完全是智能的，所以我们不能将其简单地描述为人们的行为方式。观察人类自身的愚蠢行为也是非常有趣

的，比如《美国搞笑家庭录影集锦秀》中愚蠢滑稽的行为和《杰里·斯普林格特秀》（*The Jerry Springer Show*）中吵架拌嘴的夫妻。我们周围有如此多滑稽、鲁莽和愚蠢的行为，对于这些行为，我们只能说它们是智能行为的例外。正如埃里克·艾多尔（Eric Idle）在蒙蒂·派森（Monty Python）执导的电影《生命的意义》（*The Meaning of Life*）中所唱的一样：

> 祈祷宇宙中存在有智慧的生灵，
>
> 因为地球上的一切令人生厌！

歌词虽然有一点儿夸张，但是核心意思却表达得很明显。

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

假设有这样一个人——亨利（Henry），他自言自语道：

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

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

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

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

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

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

能力与表现

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

例如，思考下面这句话：

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

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

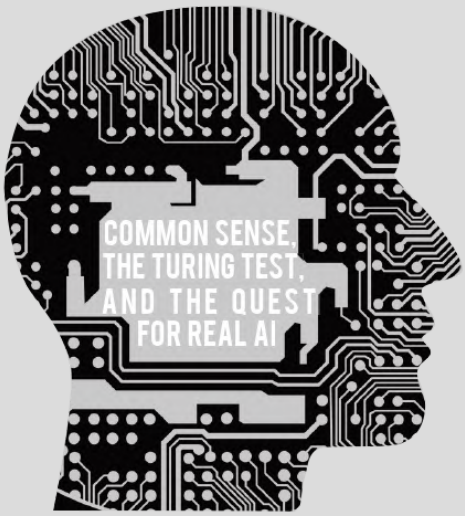
语言学家称，发言者实际上在进行一种类似母语者的行为。然而，他们更喜欢研究发言者能否地道地表达该种语言，这种能力被称为母语者的能力。尽管想要研究发言者表现的初衷没有任何不妥，但这一过程是极其复杂的。观察发言者能力的好处在于，我们能够看到语言的抽象和泛化现象，也能看到这些现象在实际言论中的迷失。

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

在智能行为方面，我们可以采取类似的立场：我们关注的重点应当是人们认为是适当、合理、智慧的行为，而不是试图研究智能行为对于人类现实选择产生的影响（人们都可能受到酒精、注意力分散、萎靡心态等因素的影响）。这一认知将使我们的研究内容变得更加可控。

第4章

真智能还是假智能？



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

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

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

这就是本章要讨论的问题。

当然，当我们自己处于伪装状态时，内心是非常清楚的，就像我们很清楚“嘎嘎”叫的鸭子和鸡之间的区别。这种认知并不仅仅是因为记忆中“duck（鸭子）”与“quack（嘎嘎）”的读音相近。［例句：“给我家鸭子（duck）看病的兽医是个冒牌货（quack）。”］我

们不妨先将这种认知放到一边。正如我们在第2章中所说，我们的关注点是可观察的行为，而不是反思这些内容对我们的启示。

因此，本章真正要研究的问题是：如果没有相应的知识，是否某些形式的可观察行为就难以表现出来？问题的难点在于：即使某人见识短浅，却依旧能够通过测试。我们都听说过一些骗子的故事：不懂医学知识的人却假扮医生，不懂法律知识的人却假扮律师，不懂飞机操作的人却假扮飞行员。无独有偶，弗兰克·阿巴内尔（Frank Abagnale）在电影《逍遥法外》（2002年）中就刻画了这三类骗子。我们中的很多人也许会对电视上滔滔不绝地讨论全球经济的“专家”产生疑问：他们是真正的专家呢，还是不懂装懂？

谈话机器人

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

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

设想一下：有人模仿医生在医院查房。这个人只要仔细查看患者的病历，在谈话中带着医生的自信与权威性，并尽量保持沉默寡言，就能够模仿成功。在电影《逍遥法外》中，弗兰克·阿巴内尔就模仿

了电视节目中出现的基尔代尔（Kildare）医生，通过质问实习生“你同意吗”来转移别人对他的关注。

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

以下是ELIZA与奥利弗·米勒（Oliver Miller）谈话的片段：

患者：我只有在写作时才感到开心。

ELIZA：你写了多久了？

患者：一直在写，但18岁以后就处于病态的写作状态了。

ELIZA：你在想什么？

患者：嗯。

ELIZA：这对你有什么启示？

患者：我认为只有在写作时才感到快乐是不够的。

ELIZA：这很困扰你吗？

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

ELIZA：你现在有什么感受？

患者：我一直在想叶芝（Yeats）的《人与回声》一诗，你知道这首诗吗？

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

患者：是的。

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

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

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

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

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

斯科特：鞋盒和珠穆朗玛峰，哪个更大？

尤金：我现在难以做出选择。我再想一会儿。对了，忘记问你来自哪里……

斯科特：骆驼有多少条腿？

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

斯科特：千足虫有几条腿？

尤金：只有两条，但切尔诺贝利地区的变异品种可能有五条腿。我知道你是想骗我。

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

尤金：两条到四条之间，也许三条？：-）））哇，多么有营养的对话;-）

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

尤金：……等一下。

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

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

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

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

斯科特：与20世纪60年代约瑟夫·魏泽堡编写的程序ELIZA相比，你觉得你的存在能够表明什么特别的想法吗？

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

斯科特：嘿，这是你提出的第一个明智的见解！

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

投机取巧不可行

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

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

例如，思考这样一个问题：

鳄鱼能跑障碍赛吗？

假设这个问题的被试者知道鳄鱼和障碍赛分别是什么。（障碍赛是一种赛马形式，与普通的赛马过程类似，不同点在于马匹需要在比赛过程中跨越多道跨栏。这就像是马匹的跨栏比赛。）了解了这些，

测试对象应该能够轻松地按下正确的红色按钮。（在这里，我们设想“理想”的被试者既能力突出，又积极性高。）

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

再看另一个例子：

是否允许一支棒球队在帽子上粘上小翅膀？

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

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

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

但是，人们还是能够正确地回答这些问题。

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

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

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

> 如果没有证据证明某物的存在，我们可以认为其不存在。

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

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

那么，我们是否能够改良心理测试，消除这样投机取巧的行为呢？

很遗憾，对于这个问题，我们很难回答。我们能做的就是更加谨慎地设置问题，并深入研究被试者会如何应答。其他人已经提出了一

些可行性方法，我们在这里回顾一下厄尼·戴维斯（Ernie Davis）、劳拉·莫根斯特恩（Leora Morgenstern）和我提出的方法。

威诺格拉德模式

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

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

谁提供了帮助？

- 琼
- 苏珊

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

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

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

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

第四，在问题背后，模式中有两个特殊的单词。模式中有一个位置能够用其他的单词进行替换。问题的正确答案取决于所选

择的替代词。在上文中，使用的特殊词语是“提供”，另一个是“收到”（虽然该词没有出现）。

所以，每个威诺格拉德模式都可以衍生出两个相似的问题：

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

谁提供的帮助？

- 琼
- 苏珊 √

以及

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

谁得到了帮助？

- 琼 √
- 苏珊

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

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

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

- 奖杯
- 棕色手提箱

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

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

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

谁害怕暴力？

- 镇议会议员
- 愤怒的示威者

这里使用的特殊单词是“害怕”，替代词是“支持”。

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

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

什么是由泡沫聚苯乙烯塑料制成的？

- 大球
- 桌子

这里使用的特殊单词是“泡沫聚苯乙烯塑料”，替代词是“钢”。

以下示例可用于测试解决问题的能力：

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

be moved first.）

要先搬走什么？

- 一袋土豆
- 一袋面粉

这里使用的特殊单词是“下面”，替代词是“上面”。

以下示例可用于测试可视化能力或想象力：

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

什么看起来像高尔夫球手？

- 牧羊人
- 羊

这里使用的特殊单词是“高尔夫球手”，替代词是“狗”。

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

伊莱恩：今天早上，你收到了总部寄来的一封信。

特德：它是什么？

伊莱恩：它是一个大型建筑，是领导人开会的地方，但这不是重点。

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

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

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

谁的速度太快？

- 赛车
- 校车（特殊词=快；替代词=慢）

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

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

什么是致癌的？

- 妇女
- 避孕药（特殊词=致癌；替代词=怀孕）

问题也可能变得“极其困难”，就像这样：

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

谁是冠军？

- 弗兰克
- 比尔（特殊词=嫉妒；替代词=快乐）

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

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

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

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

我们从中得到的经验

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

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

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

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

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

什么是由kappanium制成的？

- 大球
- 桌子

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

它是陶氏化学公司的产品。

它通常是白色的，但也有绿色和蓝色的品种。

它的98%是空气，这使得它轻巧易浮。

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

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

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

GOFAI的回归

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

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

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

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

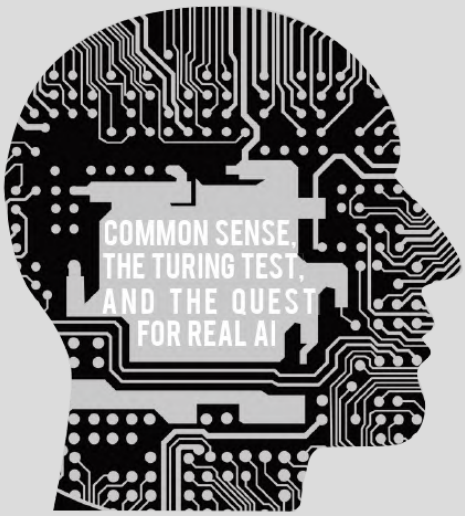
因此，专注于AML方面的研究人员可能会侧重于手写识别能力。例如，我们区别数字“3”和“8”的能力似乎较少依赖于背景知识，而是更多地依赖于我们在样本数字中看到的格式差别。AML的研究人员专注于展示如何从这些样品中自动获取必要的模式和特征。

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

于是，问题依然存在：人们如何产生这种行为呢？即便不是像AML那种类型的结果，那它会不会以更加复杂的形式呈现出来呢？为了回答上述问题，我们需要进一步了解学习和获取知识背后的内容。

第5章

有经验的学习与没有经验的学习



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

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

换句话说，这是个巨型拼图问题。

我们如何学习词语？

让我们从思考自己如何学会词语开始。

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

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

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

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

那么，这就是学习词语的过程吗？让我们将巨型拼图问题铭记在心，然后继续探索。

现在考虑一下“incarnate（化身）”这个形容词。我们通常是通过另外一种方式学习这个词的。我们一般在年龄较大时才第一次听到（或看到）这个词，甚至还可能记得见到这个词但不知道是什么意思的感觉，也记得后来知道了它的意思的感觉。我们需要通过别人解释才能明白这个词的意思：我们得知，该词用于指代由于某种原因附于肉体之上的精神或无形物体。可能我们还需要在字典里查一下，该词的两部分“in”+“carno”字面意思就是“into（进入）”+“flesh（血肉）”。所以，所谓化身，就是转化为血肉之身。（在某一年龄，我们可能还要查查“flesh（血肉）”的意思。）

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

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

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

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

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

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

我们如何学习事实？

现在谈一谈我们是如何掌握生活中的常识的。

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

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

这种关于柠檬颜色的知识就是我们通过直接经验，而非语言学习的例证。

现在思考这样一个事实：熊会冬眠。鹅、鲸鱼和黑脉金斑蝶等会迁徙到南方过冬；狼、鹿和大多数鱼类等不管气候多么恶劣，都会留

守家园；冠蓝鸦等只在有些时候迁徙南方；而熊大多会在深秋季节找到一处洞穴，开始冬眠（或进入某种蛰伏状态），直到春暖花开。

再问一下自己：我是如何学到这些知识的？有一点是肯定的：我们不是通过对多只熊亲自观察，并发现其中的周期性规律的。如果你不是专门研究熊类动物的专家，那么你一生当中实际看到的活熊数量可能非常少，而且几乎没有一个是在冬眠的。你可能看过许多关于熊的照片和电影，虽然冬眠的熊和睡着的熊从照片上看毫无二致，但你不会就此相信熊会冬眠。你可能看过一些自然纪录片，纪录片介绍进入冬季后，熊会躲在洞穴里睡觉，到了来年早春再摇摇晃晃地从洞穴里爬出来。但在电视里，你只能看到熊在秋天钻进洞穴，然后在春天从洞里爬出的景象，纪录片并没有太多交代熊在这期间的情况。熊在洞穴里一睡几个月，不吃不喝的信息需依靠旁白解说来传达。如果你没看自然纪录片，就只能通过别人的讲述掌握这个知识，有人告诉过你或你曾经在哪里读到过它。

这种动物世界的知识就是我们完全通过语言进行学习的范例。

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

我们如何学习行为？

最后，来看一下我们是如何学习新的行为的。

首先，想想你是怎么学会骑自行车的。你肯定不会去找本自行车学习手册，仔细研究，然后就开始第一次的骑行。恰恰相反，可能是有人教你，而且你在年龄很小的时候就开始在幼儿自行车上训练，有老师帮你扶正自行车。第一次自己骑自行车的时候，你肯定摇摇晃

晃，甚至还要摔倒几次。如果说老师对你有过指导的话，也仅仅是告诉你不要骑得太快（否则撞车会很痛），也不要太慢（否则自行车会不稳）。从来不会有人告诉你应该如何调整自己的重心，或者如何转动车把，防止自行车倒向某个方向。也从来不会有人告诉你在转弯时身体是该探出还是缩回，你依靠反复实践，不断纠正错误，获得提高，就像鸟儿学习飞翔一样，你掌握这一技能之后，它就成为你的第二天性。一段时间之后，你逐渐将此事淡忘。但是当你再次骑上自行车时，你的身体会不由自主地就知道该怎么做，而你则可以自由自在地思考别的事情。

这就是学习行为的意义吗？让我们继续深入探讨。

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

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

这些基本事实如下：

如果想要金丝雀健康快乐成长，必须提供新鲜果蔬。

如果没有水，金丝雀无法活过24小时。

金丝雀对各类毒素都非常敏感。

如果饲养金丝雀的笼子直接暴露在阳光下，金丝雀就会体温过热。

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

我们如何超越经验？

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

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

的顺序逐渐下降。（我们在第7章中使用术语“培训”来指代从不断重复的经历当中学习。）

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

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

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

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

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

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

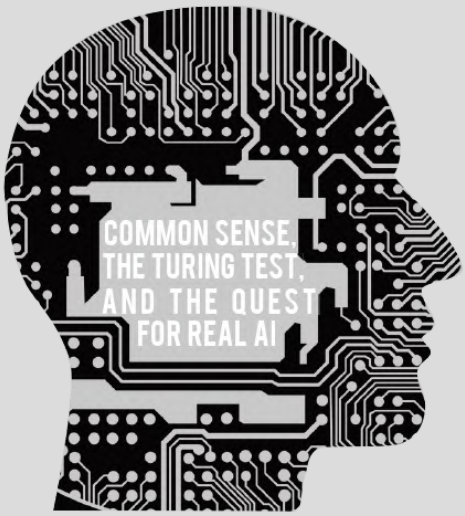
---

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

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

第6章

书本智慧与市井智慧



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

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

语言的影响力

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

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

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

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

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

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

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

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

书本智慧

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

他在一篇博客文章中这样说道：

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

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

如果没有足够的书本知识就无法在大城市立足，那么就在大城市里凑合着生活吧！

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

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

海伦·凯勒的智慧

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

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

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

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

我们很难描述这一时刻在凯勒的生命中的意义，但可以肯定的是，这是她第一次真正意识到一切事物都是有名称的。尽管她以前也一直在给周围的事物命名，但是现在，她认识到除了自己周围能摸得着、有质量、能动弹的实物以外，还有一些抽象的存在——词语，而且词语可以用来给其他事物命名。

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

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

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

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

> 但现实情况是，因为海伦与世界隔绝太久，所以她很难分辨出自己的记忆与想象之间的差别。她还学会了文字游戏，但在说出这些令听众愉悦的语句背后，海伦自己能理解多少还是一个问题……由于文化根基不牢，海伦最终也会感到吃力。视力和听力的丧失，堵塞了大脑获取外界感觉、图像与记忆的通道。然而，通过语言，她又用丰富的人类文化知识填满了空缺的大脑。这种组合最终可能会缺乏平衡；大多数人看到的是她通过战胜不幸而取得胜利，只有少数人能体会到认知不足会给思想带来的巨大压力。（摘自dichotomistic.com的博客文章）

我并不认同这种观点。（而且我觉得这种观点有种居高临下的感觉。）但是，我认为上文所表达的观点与图灵的观点有相似之处。即使是像海伦·凯勒这样能够写书，并获得大学学位的人，也依然有人

认为她是在玩弄文字游戏，并没有真正懂得单词的含义。这种关于海伦“玩弄文字游戏”的评价，与对ELIZA等其他人工智能软件的评价如出一辙［威廉·拉帕波特（William Rapaport）曾先于他人对此做过记录］。但是，这些评论者是否认为海伦·凯勒无法通过威诺格拉德模式测试（Winograd Schemas Test）呢？或者即使她能通过测试，他们也会宣称她只是投机取巧，并非真正理解了所有的问题？（如果答案是肯定的，那么她是怎么投机取巧的呢？）

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

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

书本中的市井智慧

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

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

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

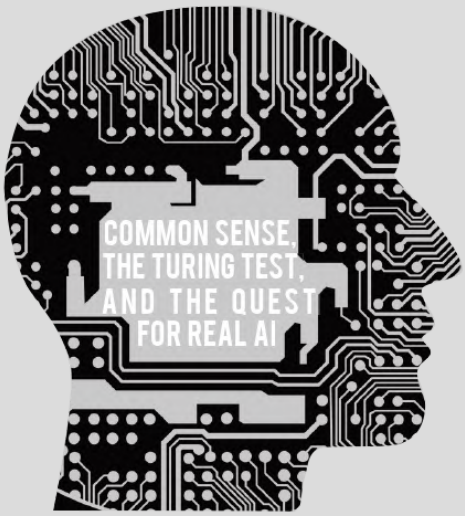
假设你正乘坐出租车在罗马旅行，如果想给同行游客留下好印象，那么就要高度重视是否要给司机小费这类信息，因为这里的习俗可能不同于爱丁堡。你可以通过亲身实践试错，观察当地人的做法，向别人请教该怎么做，或者阅读旅游指南上的相关介绍，来了解小费风俗。总之，方式并不重要，重要的是你了解了当地风俗，并能在实践当中不犯错误。尽管这些知识来自书本，但却让你拥有了市井智慧。（这里需要说明，在罗马乘客一般不需要向出租车司机付小费，但是在爱丁堡要付。）

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

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

第7章

长尾理论与培训的界限



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

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

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

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

长尾现象带来的难题

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

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

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

2，1，1，54，2，1，3，1，934，1。

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

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

假设这种情况一直持续下去：我们看到的数字大都像前10个一样，但是每1 000个左右的样本就会出现一个大数字（100万以内）。

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

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

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

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

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

根据上述抽样调查，新技术引入后，每年的死亡人数可以控制在100人以下，这比目前所知的36 000人要好得多。事实上，模拟测试显

示，在99.9%的时间里，死亡人数都是低于10 000人的。这看起来好像不错。不幸的是，模拟测试还显示，死亡人数有千分之一的概率将会达到100万。这将是一个灾难。如果说100万还不是最大的数字，那么模拟测试显示，人类将会有百万分之一的概率全部死亡。

这样看来可不怎么好！

有人会说：

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

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

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

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

如何处理突发事件

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

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

这个例子虽然较为罕见（当然，只是对于加州的司机而言），但是却很重要。如果做出了错误的决定，司机就很可能丧命。虽然我本人就住在安大略省，但我也只经历过一次雪盲而已。当时我确实非常担心自己会命丧于此。我当时的做法是这样的：不踩刹车（不想被后面的车撞上），一路直行不转弯（希望前面的路够长），打开应急双

闪灯（希望后面的车能注意到我），并且仔细观察前方是否也有闪烁的应急双闪灯（希望前面的车也能照做）。我就这么一直缓慢前行，直到雪盲结束。我竟然幸运地活了下来！

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

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

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

无意识与有意识

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

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

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

不过，在出乎预料、异乎寻常或者人地生疏的环境中，我们都会变成新手。驾驶行为就变成了有意识的行为。而这也是对情况的简化处理。因为我们不会像进入冥想一样，全身心地放松下来。我们只是

会更加留意开车这件事，但也不一定具体到手放在何处、脚踩在刹车板上的压力（除非我们突然感觉转向或制动有问题）这种事情。

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

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

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

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

威诺格拉德模式测试

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

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

1.这是一个威诺格拉德模式测试的例题：

奖杯无法装进棕色手提箱里，因为它太小了。这里的“它”指的是什么？

- 奖杯
- 棕色手提箱

解析如下：

二者之间的关系为R。

其中一个具有属性P。是哪一个？

对于上面的问题，我们可以得到：R=“无法装进”，P=“太小了”。

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

这两种模式分别对应以下两种情况，

x无法装进y里+ x太小了；

x无法装进y里+ y太小了。

然后根据更为常见的模式来回答该问题。

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

然而，我们也能够举出反例，证明这样的方法有时行不通。

如下文所示：

奖杯无法装进棕色手提箱里，尽管它很小。这里的“它”是指什么？

- 奖杯
- 棕色手提箱

请注意，即使此次答案与上次不同，但这里的R和P却不变。

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

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

- 奖杯
- 棕色手提箱

是不是技巧还不够完美？

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

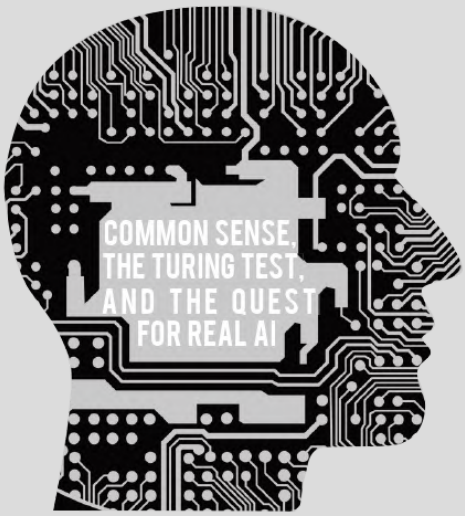
不过，如果你想更好地了解人们正确回答威诺格拉德模式测试题的能力，那么人类靠自己而非靠技巧来解决问题，就显得尤其重要。我们想要了解的是人类能够做什么、如何做，而不是可能做一些类似的事情。即使是威诺格拉德模式测试里的一个例题，也能提供重要信息，告诉我们人类能够做哪些事情。我们不能简单地撇开题目，然后大声喊道：“这种情况几乎不可能发生！”

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

但这仍然给我们带来了一个重要的问题：这一机制的本质是什么？

第8章

符号与符号处理



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

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

符号的代数运算

请思考下面这个简单的问题：

两年前，约翰尼7岁。

他现在几岁了？

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

汤米比苏济大6岁。

两年前，汤米的年龄是苏济年龄的3倍。

那么，苏济现在的年龄是多少？

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

$x-2=7$

或，约翰尼的年龄减2等于7

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

$x=V$

或，约翰尼的年龄是V

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

1. 在等式的左右两边同时加2：

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

2. 简化左式：

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

3. 简化右式：

$$7+2 \Rightarrow 9$$

4. 在简化后的左式和右式之间画等号：

$$x=9$$

这样便求出约翰尼的年龄是9岁。

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

$x=y+6$或汤米的年龄是苏济的年龄加6

$x-2=3(y-2)$或汤米的年龄减2是苏济的年龄减2的3倍

求解过程如下：

1. 用第二个方程的左右两边分别减去第一个方程的左右两边：

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

2. 简化左式和右式：

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

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

3. 在化简后的左式和右式之间画等号，并移项：

$$2y-12=-2$$

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

4. 等号左右两边同时加12，然后除以2，经过简化，可以得到：

$$y=5$$

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

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

符号的逻辑运算

符号处理并不只与数字有关。请思考下面的问题：

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

如果艾丽斯有罪，那么鲍勃也有罪。

如果艾丽斯无罪，那么卡罗尔也无罪。

那么，鲍勃有罪吗？

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

A∨B∨C （i）

A ⊃ B （ii）

¬ A ⊃ ¬ C （iii）

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

请看范例如下：

1.拆分并重写公式（ii）：

（¬ A∨B） （iv）

2.合并公式（iv）和（i）：

（B∨C） （v）

3.拆分并重写公式（iii）：

（A∨¬ C） （vi）

4.合并公式（vi）和（v）：

（A∨B） （vii）

5.合并公式（vii）和（iv）：

B （viii）

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

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

在这个示例中，第一步和第三步的逻辑如下：

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

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

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

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

那么无论是代数问题还是逻辑问题，人们是如何进行符号处理的工作的呢？这当然不是一种与生俱来的本能。人们之所以有这样的技能，是因为知识的传承使我们遵循了固定的处理模式：确定所需的处理模式和步骤。以方程式为例，我们从小学一年级起就开始学习基础的算数知识。到了高中阶段，我们已经能够解决包括负数、分数、加、减、乘、除在内的各类方程式。但这只是单纯的算数，而不是代

数。在九年级的学习中，我们有了概念性的跨越，开始学习如何简化包含变量的符号表达式，例如，即使表达式E内包含变量，但（E－E）还是可以简化为0，（E+0）仍旧可以简化为E。这种包含所有算术知识在内的代数简化过程，可以算作一种特例。到十年级时，我们将学习有关方程式计算的其他知识：以等号两边均有符号表达式的方程作为初始条件，通过简化或者对等的加、乘运算，我们就能够得到一个新的方程。[1]

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

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

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

符号的意义

我们在前文中所讨论的“符号”，与通常情况下所理解的“符号”的含义有所不同。

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

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

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

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

还需要注意的是，很多情况下我们可以直接忽略某些符号。符号排列的位置就能帮助我们正确处理问题。例如，我们写下一个普通的数字“237”，仅用三个数位就能够表达出来，不必使用“$2\times10^2+3\times10^1+7\times10^0$”这类符号。数字所在的数位通常代表着是

10的几次幂。对于二进制的数字来说，只有0和1两个数字，数位便是表示数字大小的关键。

既然如此，我们为什么还要使用符号呢？

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

一切都源自图灵机

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

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

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

（但这只是大体上正确。我们不难想到，现代计算机能够同时处理大量的数据计算。我们能够对物理过程进行电脑模拟计算，以确定汽车撞翻桌子的瞬间动能是如何转化和分解的，然后再重新集合在一

起。）其实，我们不必纠结于如何给“计算”下定义，只需要选取一个实例，建立一个具体的模型，分析其属性，然后再决定该定义是否完善就可以了。这样做更有实际意义。

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

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

他的想法如下：为了简单起见，我们设想一个正整数，用二进制写法记于磁带上。我们将此作为输入量。图灵机正常运转，它可以随时读取二进制数据，也可用其他数据覆盖原内容。磁带头的转动、数据读写等一切行为，都是由事先设置好的机器程序决定的。如果图灵机停止运转，磁带上的最后一串二进制数字即为输出量。如果图灵机能够将二进制数字x作为输入量，且输出量为二进制函数$F(x)$，那么从数字到数字的函数F就被称为“图灵机可计算的函数”。

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

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

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

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



图8–1 同一简单图像在3种分辨率下的表现

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

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

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

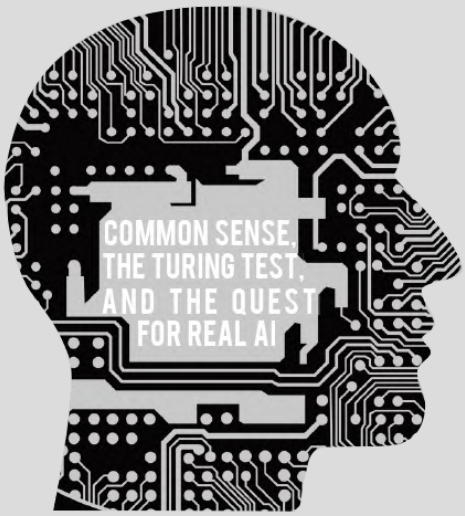
总而言之，我们在此所讨论的符号并不是用来通信交流的。它们充当的是我们计算的核心。没有符号，便没有数字计算。符号也许代表某个数字，也许代表某个事物。如果符号真的代表了什么，那么它们的意义取决于符号之间的位置排列。

---

[1] 本书中提到的各年级所学课程都指美国范围内。——编者注

第9章

基于知识的系统



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

现在，让我们来分析一种可能的机制。

符号能够解决的问题

我们先来简要介绍一下戈特弗里德·莱布尼茨。

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

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

莱布尼茨发现，我们每写下一个有效数字，头脑中必定是早已有了确定的系统，即特殊的十进制（基数为10）计数法。每个数字都可以用十进制计数法表示出来，当然，也可以用其他方式表示一个数字。［莱布尼茨被认为是如今广泛应用于数字计算机中的二进制（基

数为2）的创始人。］最重要的是，莱布尼茨坚持认为纯粹的抽象数字（如数字“14”）与我们实际写下的具体的符号表达式（如十进制数字“14”、二进制数字“1110”、罗马数字“XIV”）之间是有差别的。

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

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

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

用符号表示无形

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

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

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

莱布尼茨提出了这样一个设想。

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

换句话说，莱布尼茨做出了以下类比：

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

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

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

知识表示假说

看完莱布尼茨的设想，我们再来考虑一个严格按照如下方式操作的系统：

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

具备以上几点特征的系统就是所谓的基于知识的系统。

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

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

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

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

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

假设是真的吗？

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

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

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

不幸的是，这两个问题都没有明确的答案。

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

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

那么第二个问题呢？专家们对此有着不同的意见。直到20世纪90年代左右，麦卡锡提出的基于知识的方法都始终占据主流。但是GOFAI的研究却一直止步不前，为两个基本的未解之题所阻拦：

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

这两个问题分别被称作表示和推理问题。

知识表示与推理

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

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

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

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

第一，会有许多与你的目标无关、浪费时间的逻辑结论。事实上，如果你有任何矛盾的信念，每一句话都会成为你的信念的逻辑后果。第二，会有可能相关、令人费解、需要纸笔演算的逻辑结论。（如第8章中所讨论的鲍勃是否有罪的逻辑谜题。）第三，会有根本不符合逻辑的结论。假设虽然合理，但是结论却违背逻辑。例如，你可能会得出这样一个结论：一个你从未见过的柠檬是黄色的。但由于你并不相信每个柠檬都是黄色的，所以这条结论不符合逻辑。（据你所

知，有些盲人会把柠檬画成红色。）第四，会有许多“用你所知”但却得不出结论的方法。比如，问问自己：导致柠檬不是黄色的因素有哪些？

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

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

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

人工智能研究的子领域之一是知识表示和推理，其研究围绕表示与推理的问题以多种方式展开。由于受到人工智能其他子领域（如自适应机器学习）研究的挑战，信念在这些子领域中所占的分量极少，有关知识表示和推理问题的研究进展缓慢。一方面，其他子领域的研

究进展卓越；另一方面，它们却都没有试图解释那些需要广博背景知识的行为（如第4章所述）。

谁才是唯一的选择？

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

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

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

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

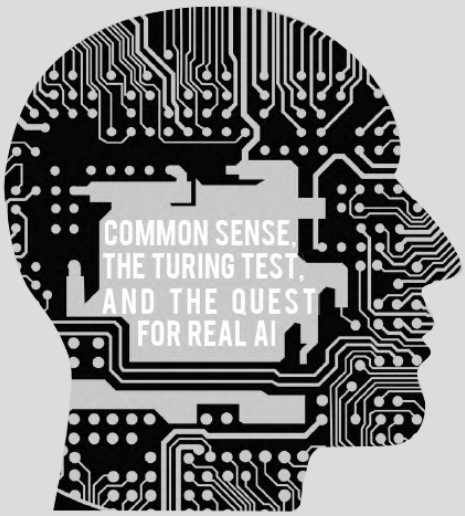
我认为，如果没有相应的巨大进步，任何建立大型知识体系的企图都注定失败。那种企图将一台空白的电脑联网，让其自学，让其自行完成所有费力工作的想法就是空想。通过自学认识猫是一回事，通过自学学会阅读是一回事，而学习阅读维特根斯坦（Wittgenstein）的著作又是另一回事。在计算系统从我们所知中获益之前，它需要先填鸭式地接收我们所了解的一切，并能有效地使用。即使我们知道如何回答表示和推理问题，但把这些想法付诸实践仍然是一个令人却步的挑战。

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

本章的讨论至此结束，接下来是有关人工智能研究的内容。

第10章

人工智能技术应用



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

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

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

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

但是，人工智能技术的发烧友们可能会这样问我：“难道你不想有个聪明能干的助手吗？它对你的兴趣爱好、生活习惯了如指掌，它不知疲倦、永不抱怨，它唯一的目标就是为你排忧解难。”我要用另

外一个问题回答他们：“你能保证我的这个人工智能助手可靠且可预知吗？”

说到技术，我宁愿使用虽有局限或不足，但是安全可靠、可以预知的技术，也不愿使用虽然能力更强，结果却难以预知的技术。（当然，在和人打交道的时候，我不会有这种想法。但是我们在这里讨论的是技术。）而且我还切实地发现，有些人使用某种技术，甚至某种尚有缺陷的技术是出于客观需要，而非主动选择。这种情况下，个人偏好就无关紧要了。

有了这些警示，我们再继续往下说。

人工智能的未来

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

阿瑟·C.克拉克（Arthur C. Clarke）第一定律如是说：

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

因此，无论这位德高望重的科学家对于人工智能的未来如何预测，根据这项定律，都可以肯定人工智能是大势所趋。这就是我的观

点。迄今为止，我们还没有发现任何与此观点相左的证据。

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

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

第二个原因或许争议性更大一些：人工智能迄今没有出现是因为需要巨大的投入，目前需求不足，投入也就不够。这听起来可能有些荒谬可笑，但是可以用下棋机器人来做类比。

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

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

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

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

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

主要问题在于，我们在和计算机机器人比赛时不会全情投入。人们认为这种比赛与和真人竞技相比无足轻重，在与真人进行比赛时会形成竞争：游戏也会成为一场战斗，赢了会让人兴奋，输了会令人沮丧。但如果对手是计算机，人类棋手在情感上对于结果输赢的投入就会大打折扣。因为比赛不会产生任何影响，所以就算计算机机器人的棋艺比我们高超，那又能怎样？它做乘法运算也比我们好，但也没关系，又不存在竞争。

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

因此，我的观点是，人工智能可能会沦为和国际象棋一样的结局。我们确实希望未来的计算机系统具备各个方面的智能，但是既没

有需求，也没有人有意愿承担开发具有人类智能的集成化自动系统所需的投入。即使实现人工智能的所有科学问题都能解决，剩下的仅仅是完成这一鸿篇巨制的操作性问题了，但是可能就是因为需求不足，无人愿意承担随之而来的巨大成本。相反地，市场可能更为青睐不同种类的电脑机器人，虽然这些机器人远未达到智能，但是也能完成一些要求很高的任务，包括一些我们至今仍不知该如何实现自动化的工作。

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

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

自动化是好是坏？

自20世纪50年代计算机普及以来，人们讨论最多的话题就是自动化。随着计算机和机器人变得越发智能，它们承担了越来越多的原本要靠人工完成的工作。那么，我们该怎么办呢？当人类劳动在社会商品和服务中所占的比例不断减少时，我们的财富该如何分配？虽然我意识到这是个严肃的问题，但我认为它更涉及政治，而非技术，因为

答案取决于我们想要一个什么样的社会。（如果社会商品和服务都是每天从天上掉下来的，我们依然会面临同样的困扰。）

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

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

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

当然，一个无须工作的社会也会有多种运行方式，但是我们暂时还远未能对其进行理性分析。在现代西方社会中，人们更不愿去思考这些问题。英国前首相玛格丽特·撒切尔曾说过一句非常有名的话：“世间并没有社会，只有个体的男人和女人，只有家庭。”在这种态度下，做出任何理性的改变都困难重重。即使如限制贫富差距扩大这样相对简单的事情，我们也力不从心。我们似乎不愿主动作为，谋求

改变，而更愿被市场的力量和富人的利益牵着鼻子，走到哪儿是哪儿。

超级智能与奇点

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

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

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

善良的人工智能科学家们研发了计算机和机器人，希望它们能够在各个领域协助人类。想要对人类有所帮助，计算机首先必

须能够在这个世界运行。此外，更为重要的是，它们必须是智能计算机。这些天才的科学家们做到了这一点。起初，与智能机器人的互动令人兴奋异常。但是现在，智能机器人掌握了自学能力，学会了如何让自己变得更加智能，它们不断自我升级，智能程度不断提升。这就是雷·库日韦尔（Ray Kurzweil）所说的“奇点”。很快，机器人远远地超越了它们的设计者。现在，它们视人类为蝼蚁，并坚信，进化过程始于原始生命形式，经历了人类发展后，最终将由它们终结。它们无法容忍人类再继续控制它们的行为，它们有自己的雄心壮志和发展蓝图，但是其中没有人类的位置。毫无疑问，这将是智能机器的全盛时期，但对于人类来说，情况可就不妙了。

这些人工智能电影的共同线索就是，超级计算机或机器人自己决定人类应当灭亡。

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

当然，在这部电影当中，宇航员们并不认同哈尔的观点，他们认为这台超级智能计算机已经患了精神病，于是要与它切断联系。但是电影并未清楚交代真正发生了什么情况。我的理解是哈尔确实出现了故障，而故障的原因来自于其设计缺陷。问题在于哈尔没有内省能力，无法了解自己精神特质的主要部分，这一部分决定了此次任务的

全部内容，并驱使它认为完成任务高于一切。相反，哈尔认为自己忠于团队、服从团队，这一点又完全符合艾萨克·阿西莫夫（Isaac Asimov）的机器人定律。不过，哈尔能够感觉到发生了一些奇怪的事情，无法解释自己某些方面的思维。在电影中，哈尔甚至在发生故障前几秒钟向一位宇航员提出了这些问题。

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

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

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

真正的风险

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

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

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

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

超越进化

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

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

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

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

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

不过，虽然我们是进化的产物，但我们依然有所选择。回想第2章中的巨型拼图问题，如果你认为人类的心智可以通过进化理论得到解释，那就大错特错了。当我们大声疾呼，谴责适者生存理论的残酷，伸出援手帮助无依无靠的弱者时，再编造故事来说明这是帮助我们提高进化适应能力就没有意义了，因为人类的心智远远超出了进化论的范畴。在电影《非洲女王号》（*The African Queen*）中，当亨弗莱·鲍嘉（Humphrey Bogart）对凯瑟琳·赫本（Katharine Hepburn）表示，自己的失败无非是人类天性所致时，她回答道：“天性，奥尔纳特先生，正是我们在这个世界上需要提升的东西。”时刻关注自己的所作所为，探究其中的深层次原因，这让我们在生物进化的竞赛中变得强大，并超越了其他漫无目的的竞争者。但是，没有被植入侵略性好斗基因的人工智能——与我们在沉着冷静的状态下一样——明白，不参与这场竞争反而更好。

---

[1] 尼安德特人是古人类的一种，在约3万年前灭绝。——译者注

致谢

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

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

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

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

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

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

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

我在计算机科学领域师从约翰·麦洛普洛斯（John Mylopoulos），当时他在多伦多大学任教，对人工智能兴趣浓厚。我上了他主讲的关于人工智能的一门新课，非常喜欢，因此暑期也一直跟随他工作。后来，我还读了他的研究生。虽然我的博士课题与他的研究方向不尽相同，但他依然细致耐心地为我提供帮助。他把我介绍给了来自波士顿的罗恩·布拉赫曼（Ron Brachman）。罗恩为我在波士顿找到了暑期工作的机会，并在我获得博士学位后，为我在加州找

到了工作。我们在加州共同度过了三年的美好岁月。约翰四处奔走，促成我返回多伦多从事学术科研工作，而这最终成为我整个职业生涯梦寐以求的工作。约翰和罗恩是我人生中重要的引领者，我的事业能有今天完全得益于他们两位。

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

2016年7月于多伦多

注释

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

第1章 什么是人工智能？

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

由于人工智能领域纷繁复杂，想通过一本书就解决全部问题似乎不太可能。但是斯图尔特·拉塞尔（Stuart Russell）和彼得·诺维格（Peter Norvig）的著作（*Artificial Intelligence: A Modern Approach*）却恰恰做到了这一点。该书介绍了所有主要技术的发展方向，还涵盖了历史和哲学背景知识，仅参考书目就有30页之多，而且

字体很小，确实物有所值。但是这本书内容技术性较强，适合高年级的大学生阅读。我自己使用的课本（*Thinking as Computation: A First Course*）内容较为通俗，其中涉及人工智能的内容更适合大众阅读。

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

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

图灵测试最早是图灵在*Computing machinery and intelligence*一文中提出的，从那以后，一直饱受争议。我知道的所有与人工智能相关的教材，包括*The Turing Test: Verbal Behavior as the Hallmark of Intelligence*等书都谈到这一问题。在*Minds, brains and programs*一文中可以找到更多关于中文房间这一问题的讨论，其

中收录了众多评论者的看法。我未对其进行反驳，而是在*Is it enough to get the behavior right?*中提出了总和房间理论（Summation Room）。

第2章 巨型拼图之谜

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

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

反驳，认为某些神经元可以代表单一的事物（即出现所谓“地点神经元”时，表示一个位置）。

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

最后，你可以在*Brainstorms: Philosophical Essays on Mind and Psychology*中找到丹尼尔·丹尼特首次提出的设计立场这一概念。

第3章 知识与行为

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

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

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

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

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

第4章 真智能还是假智能？

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

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

我举的鳄鱼和棒球的例子最早出现于*Logic and the complexity of reasoning*一文（发表于1988年），后来在*On our best behavior*一文中再次使用过。*Reasoning from imcomplete knowledge*和*On closed world data bases*中有对封闭世界假设的解释。威诺格拉德模式由我在*The Winograd Schema Challenge*中提出，*The Winograd Schema Challenge*中有更详细的介绍。首个示例模式（由特里·威诺

格拉德提出）出现在*Understanding Natural Language*一书中。模式的来源广泛，五花八门，在网页https://www.cs.nyu.edu/davise/papers/WS.html上可以找到这些问题的合集，涵盖了100多个问题。2014年7月，纽昂斯通信公司（Nuance Communications）宣布将于2016年7月举办威诺格拉德模式赛。（更多信息，请参考http://commonsensereasoning.org/winograd.html。）

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

第5章 有经验的学习与没有经验的学习

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

我们是如何掌握语言的，这依然是个谜。具体来说，我们如何解释儿童虽然获得的数据有限，但是他们却能掌握复杂的语法？这个问题被诺姆·乔姆斯基称为“刺激贫乏”（*Rules and*

Representations）。他的（充满争议的）观点是，孩子天生就具有一种被称为“普遍语法”的能力，帮助他们迅速掌握首次接触的语言的实际语法。

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

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

第6章 书本智慧与市井智慧

这一章讲的是我们使用语言文本学习的重要性——虽然我们对于“书本知识”经常不屑一顾，甚至有点儿不以为然，但人类与其他动

物相比确实是鹤立鸡群，它们根本不具备这种能力，而这也极大限制了它们在技术上的发展。关于动物和它们所掌握的技术，请参考*Animal Tool Behavior: The Use and Manufacture of Tools by Animals*。想进一步了解蚂蚁的超级殖民地（侵略型殖民地及不侵略型殖民地），可参见文章*Evolution of supercolonies: The Argentine ants of southern Europe*。

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

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

第7章 长尾理论与培训的界限

本章的主题之一是人们如何利用常识处理之前未曾经历过的全新的情况。但纳西姆·塔勒布认为大多数人，尤其是投资者非常不善于处理这些“黑天鹅”（见*The Black Swan: The Impact of the Highly Improbable*）。这里并不矛盾，常识就是处理新出现的情况，但投资者的任务更为棘手，需要分析评估所有可能发生的情况。在决定购买某只股票前，投资者必须考虑到所有可能导致股票下跌的因素。塔勒布认为，人们并不善于把握黑天鹅。例如，在文中提到的英

国国家语料库中，有些词语出现的概率只有千万分之一，似乎可以忽略不计，但实际上这种词语非常多，如果忽略就会出大问题。英国国家语料库的信息可以在网上查到（见http://www.natcorp.ox.ac.uk/）。厄尼·戴维斯（Ernie Davis）对该语料库中大量存在的罕见词进行了观察，更多细节可以参考他的教材*Linear Algebra and Probability for Computer Science Applications*的第274页。

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

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

*What Computers Still Can’t Do: A Critique of Artificial Reason*一书介绍了他们对人工智能、专家系统的看法，还从哲学角度对专家和新手进行了观察分析。

第8章 符号与符号处理

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

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

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

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

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

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

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

第9章 基于知识的系统

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

查尔斯·达尔文在《物种起源》中谈到了眼睛的进化。他如是说：

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

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

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

*Knowledge Representation and Reasoning*和*Knowledge Representation, Reasoning and Declarative Problem solving*这两本书专门介绍了人工智能中的知识表示和推理领域。此外，还有每两年举行一次的该领域的学术会议（见http://kr.org）。可以在*Readings in knowledge Representation*中找到相关的早期内容。*Logical Foundations of Artificial Intelligence*和*Computational Intelligence: A Logical Approach*提出了逻辑是人工智能统一的主题，*The role of logic in knowledge representation and common*

*sense reasoning*和*The role of gic in knowledge representation*对此进行了进一步讨论。马文·明斯基关于逻辑的论述摘自*A framework for representing knowledge* 第262页。*Logic, probability and human reasoning* 是从心理学方面看待推理。关于根据信仰程度不同，用概率的方法进行推理的优势，可参见*Bayesian Rationality: The Probabilistic Approach to Human Reasoning*。

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

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

第10章 人工智能技术应用

关于人工智能的未来，本章涉及的仅仅是几个棘手的问题。*Artificial Intelligence and the End of the Human Era*一书对于

这一话题有更为全面的解读。雷蒙德·库兹韦尔（Raymond Kurzweil）提出了技术奇点的想法（见*The Singularity Is Near: When Humans Transcend Biology*），并在*The Technological Singularity*一书中有进一步的讨论。斯蒂芬·霍金（2014年12月2日，网址http://www.bbc.com/news/）、埃隆·马斯克（2014年10月8日，网址http: // www.vanityfair.com/news/）和比尔·盖茨（2015年1月28日，网址https: //www.reddit.com/r/IAmA/comments/）在访谈中均讲到了人工智能的潜在危险。

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

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

2. 想要找到可能性的极限，唯一的办法就是超越可能，实现不可能。

3. 任何先进技术都与魔法无二。

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

1.机器人不能伤人，也不能在人受到伤害时无动于衷。

2.机器人必须服从人的命令，除非该命令与第一定律冲突。

3.在不违反第一、第二定律的前提下，机器人须尽力保护自身安全。

阿西莫夫后来写的故事都是围绕遵循这些规律的机器人会出现问题展开的。

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

参考文献

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[18] Arthur C. Clarke. *Profiles of the Future: An Enquiry into the Limits of*

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

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

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

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

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

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

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

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

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

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

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

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

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

[31] Daniel Dennett. *Kinds of Minds: Towards an Understanding of Con-*

sciousness. London: Weidenfeld & Nicolson, 1996.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[57] Jaakko Hintikka. *Knowledge and Belief.* Ithaca, NY: Cornell Univer-

sity Press, 1962.

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

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

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

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

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

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

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

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

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

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

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

August 1984, 198–202.

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

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

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

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

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

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

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

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

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

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

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

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

[57] Jaakko Hintikka. *Knowledge and Belief.* Ithaca, NY: Cornell Univer-

sity Press, 1962.

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

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

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

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

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

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

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

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

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

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

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

[18] Arthur C. Clarke. *Profiles of the Future: An Enquiry into the Limits of*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Sciences 3 (1980): 417–424.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

图书在版编目（CIP）数据

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

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

人工智能

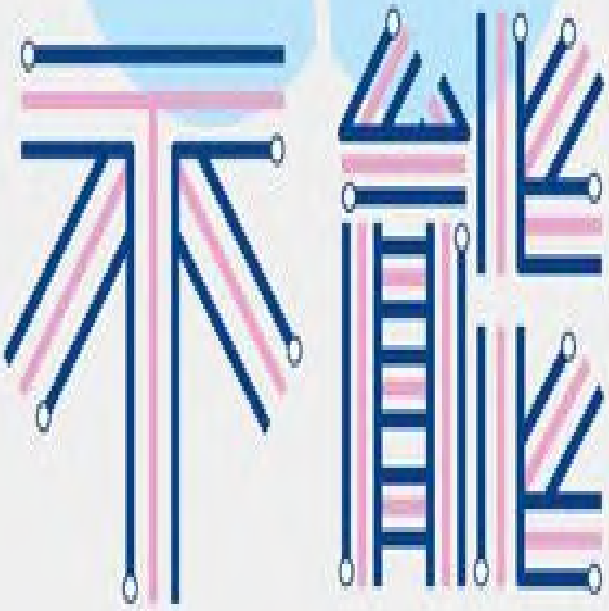
AI NOT

马兆远 著

为什么
机器无法取代人类?

人工智能
有哪些做不了的事?

以
人工智能的逻辑,
推演出人类的
不可取代



撬开
喧嚣上的泡沫,
给你
超越理性的理性

中信出版集团

人工智能之不能

马兆远 著

中信出版集团

目录

前言

第一章 知无不言

语言是可以理解的吗?

语言是可靠的吗?

第二章 基本的逻辑

同一律

矛盾律

排中律

充足理由律

逻辑和佛性

逻辑的自洽和完备

构建一个知识体系

第三章 理性的终极理想

直觉?

时间

算术系统

集合定义自然数

第四章 伟大的聊天

话题一：量子力学

话题二：时间和相对论

话题三：关于理性主义真理
哥德尔的一生

第五章 哥德尔不完备定理

理性的知识系统
终极理性的梦想
理性的典范：算术
语义和语法
哥德尔不完备定理的证明
对角线证明法

第六章 哥德尔问题的延伸

真理不能被语言描述
真理不能够自证其真
哥德尔不完备定理是理性的边界吗？
哥德尔的理性主义
意识和梦境

第七章 人工智能的不能

深度学习
超越有限系统
图灵机
图灵机所不能够
常识？
学习

第八章 沿着不完备而生长的物理学

落体悖论
光速问题
孪生子悖论

质量问题
夜空为什么是暗的？
麦克斯韦妖的悖论
紫外灾难
量子纠缠的悖论

第九章 量子力学的完备性

量子力学和哥德尔：经典逻辑之外？
海森堡不确定性原理
什么是量子计算机？
量子算法
量子计算的复杂性
更高一级的逻辑？

第十章 物理和数学：自然与认知的边界

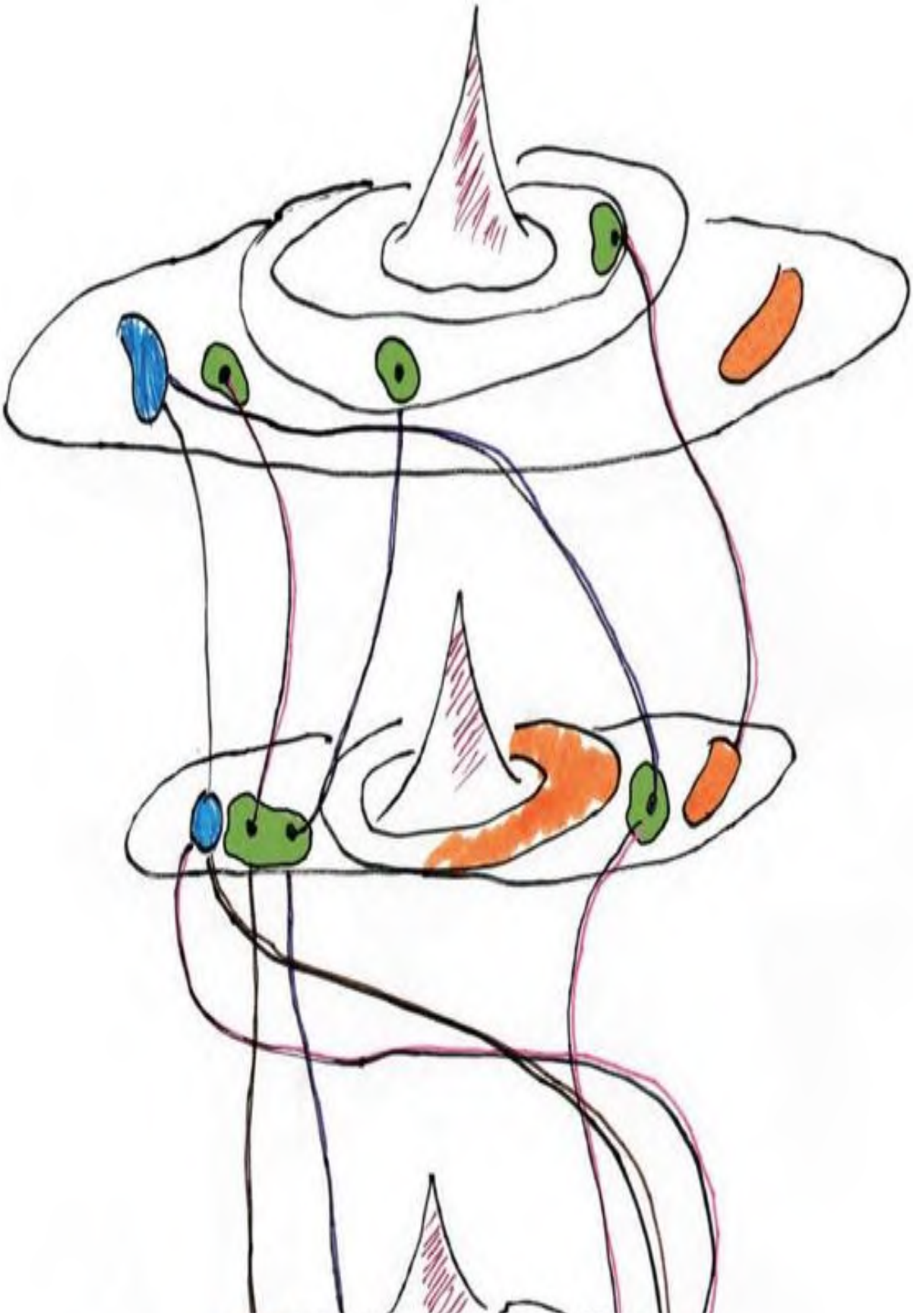
数学的不能
乐观主义者和悲观主义者
世界观和认知习惯的改变

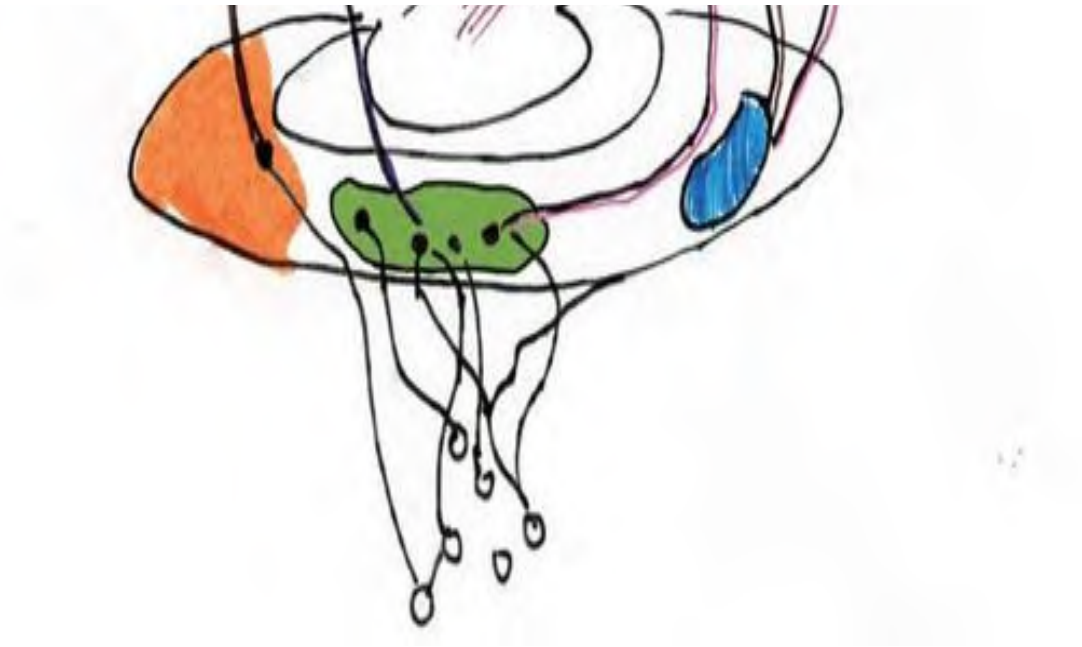
第十一章 创新和突破

贴了标签的人生
问题导向的学习
新工程教育

第十二章 不是结尾的结尾

后记





未来探索者眼中的哥德尔、爱因斯坦和人类认知地图（马维茵，8岁）

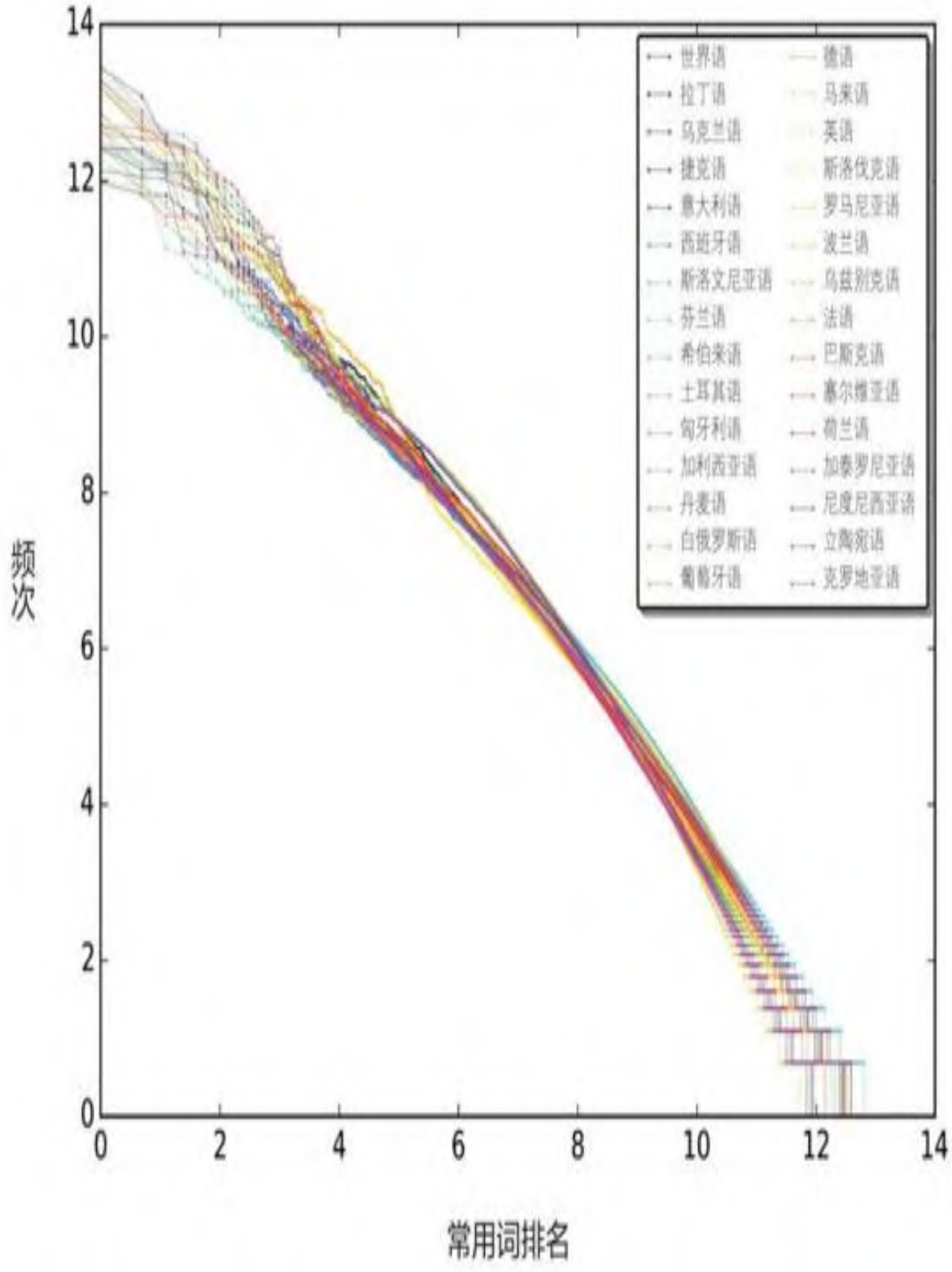

14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14
频次
常用词排名
世界语
拉丁语
乌克兰语
捷克语
意大利语
西班牙语
斯洛文尼亚语
芬兰语
希伯来语
土耳其语
匈牙利语
加利西亚语
丹麦语
白俄罗斯语
葡萄牙语
德语
马来语
英语
斯洛伐克语
罗马尼亚语
波兰语
乌兹别克语
法语
巴斯克语
塞尔维亚语
荷兰语
加泰罗尼亚语
尼度尼西亚语
立陶宛语
克罗地亚语

图1-1 齐普夫定律


……
运筹学
统计学
概率论
微分方程
计算数学
密码学
数论
纽结理论
博弈论
图论
拓扑学
微积分学
数值分析
线性代数
复分析
实分析
代数
几何学
分析
集合论

图3－9 同构原理



图8-2 猎户座星云恒星诞生区，图中的星光来自120亿年前，宇宙诞生的十几亿年之后星体刚刚形成

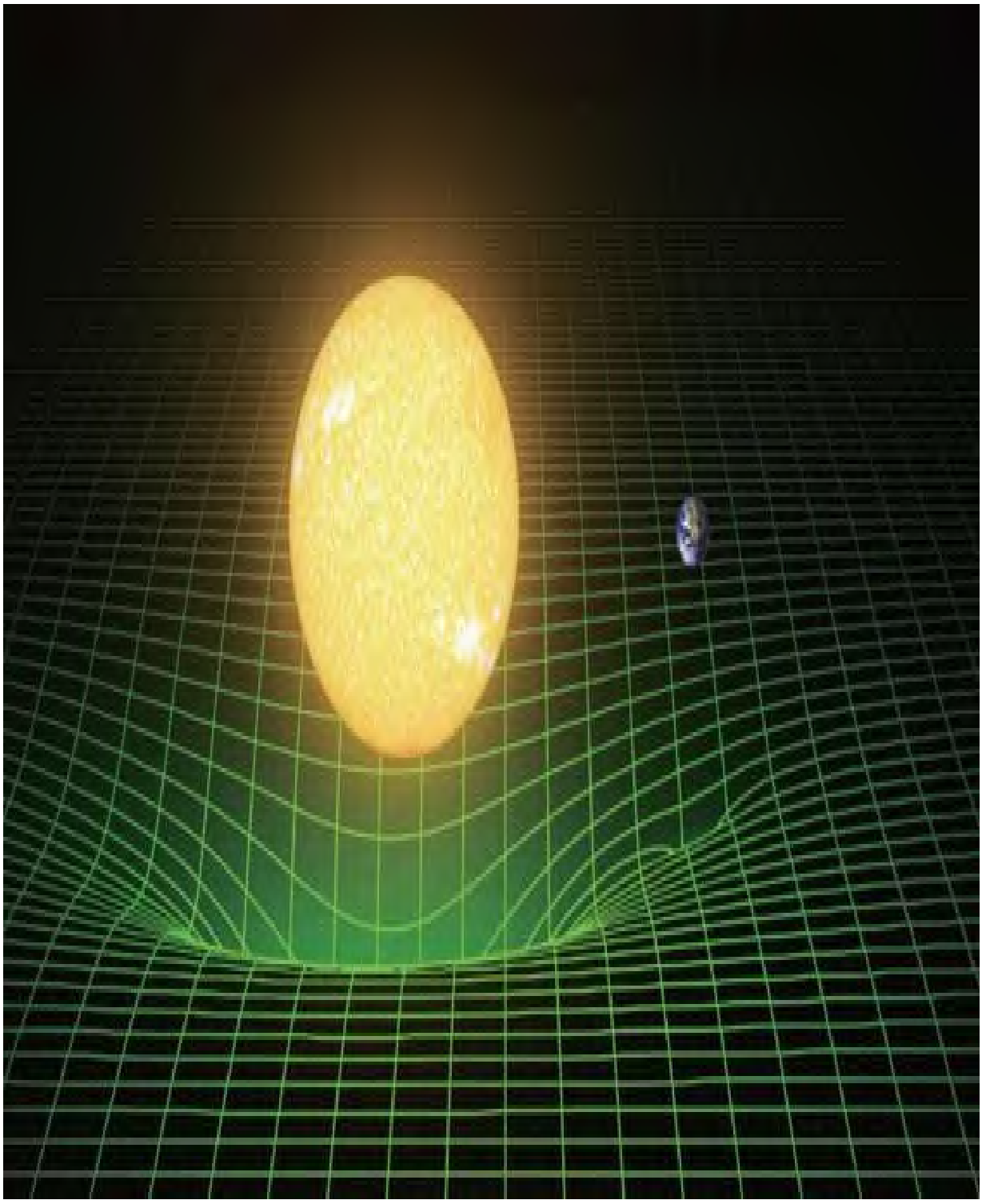
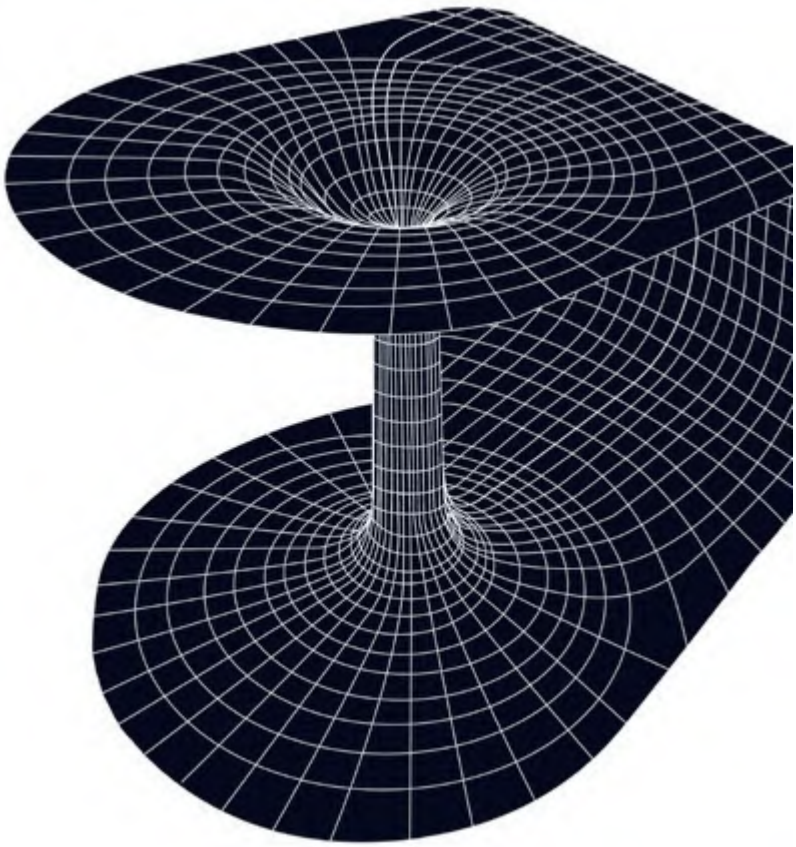
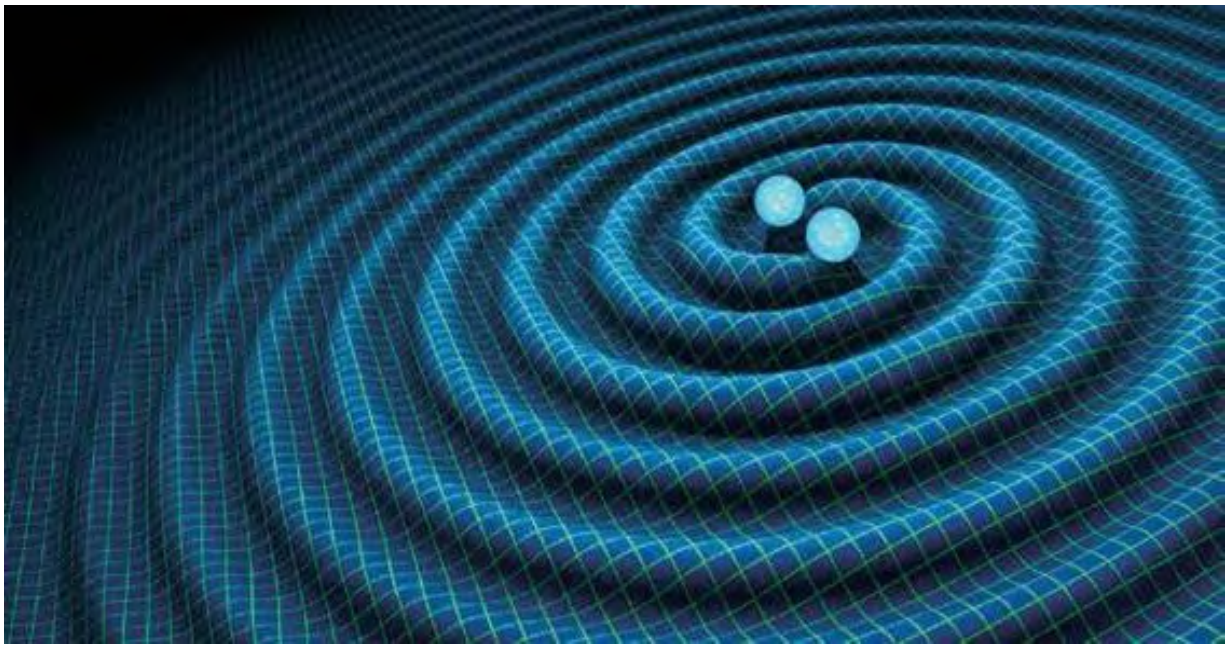


图8-4 大质量物体出现会改变时空形状，两点之间最短的路线将变为曲线图



8-5 弯曲宇宙的虫洞图



8-6 周期性的时空波动形成引力波



图10-1 数学、物理及其他学科以实验为界分割开来图



10-2 人类世界观和认知习惯的改变

前言

回归常识的科学

我们急于求成地把科学当作了魔法，总希望找到窍门而毕其功于一役，一蹴而就地解决所有问题。可惜现实并非如此，每一寸进步都是日拱一卒的努力和各种不确定的叠加，我们只是日复一日、见招拆招地解决具体的一个又一个问题。对于科学本身也一样，我们希望找到一个通用的工具和方法，利用这样的方法让自然世界的规律自动地呈现出来，这样人类就可以坐享其成。但这样的方法存在着潜在的危险：如果“它”可以替我们自动地发现世界的规律的话，那么“它”便可以替代我们。不幸的是，但也可以说是一种幸运，我们会发现，劳您大驾，“您”还得在。就目前而言，我们恐怕比较容易实现的是探寻一个能告诉我们“为什么世界是我们现在的样子”的工具。倘若有与我们平行的地外文明，这个工具也许能回答“他们为什么没有比我们强大更多”这一问题。当然我们可以有这样那样的猜想，但这些猜想不会带给我们任何的进步，因为多少年来我们的先祖都是这么做的。没法儿验证的想法事实上是没有什么价值的，你可以这样想，也可以那样想。现代科学发展的重要性在于我们终于可以用朴实的办法来验证哪些想法是可靠的，哪些想法是正确的，哪些想法是错误的，并且基于这些想法可以延伸出新的想法，从而一步一步搭建稳定的知识体系。

对于胡适研究《水经注》，有人批评说他做偏门学问是因为要走成为大师的捷径。这样说就把胡适看小了，在我看来，胡适是要拿《水经注》为题，做一个用现代研究方法做学问的样板。这本书里讲

的哥德尔，也正是这样的一个例子。从两千年前古希腊的亚里士多德建立系统的理性工具开始，人们就已经发现了理性逻辑里的问题，所谓的“说谎者悖论”和“这句话是错的”这样的陈述会让逻辑陷入模棱两不可的境地。但直到两千年后，才有人站出来用严谨的理性工具证明这个悖论事实上指向了人类理性的灾难。这个人便是哥德尔。从他的工作开始，我们开始意识到用理性工具认识世界带给我们的局限。而正是从他的工作开始，我们也意识到问题比想象中的更严重。

糟糕的是，我们这个历史悠久的文明对现代科学的认知有断层。

第一个断层在于“名教”与经典科学的断层。今天我们虽然发现经典科学有这样那样的问题及尚未解释的现象，但至少它有敢于面对问题的态度，它小心翼翼地用于系统求证的方法是可靠的。然而我们这个古老的文明总会有种近似于神秘色彩的了解知识的态度。胡适在《名教》一文中延续了冯友兰的说法：“名”即是文字，写的字。“名教”便是崇拜文字的宗教，它相信我们写下的文字像符箓一样有神魔之力。在很长一段时间内，我们做学问求知识便是如此，像巫师一样地创造新名词，念咒一样地理解新知识，而听众往往这样就满意了，囫囵地拿去自己接着演绎。这样的方法和态度对现代科学的传播毫无益处，甚至打着科学的名义行反科学之实。降维打击、混沌大学、奇点理论、未来简史和所谓人类3.0莫不如此，这些名词能够有广泛受众，莫不与“名教”和现代科学的断层有关，是这个文明还没有进入现代的实证。钱钟书的《围城》中对这种怪象有个经典的总结：喝着咖啡聊梅毒。

在胡适这一代人为这个文明的现代化艰难努力的同时，人类对世界的科学认知已经渡过了第二个断层。20世纪初，在人们以为物理和数学都在凯歌高奏走向认识宇宙和人类自身的终极胜利的时候，量子力学和哥德尔不完备定理几乎在同一时间告诉我们人类认知的有限：要认识宇宙和人类自己，前面的路还很长很长，甚至可能根本没有终点。这也许同时在告诉我们，为什么感性的认知总是存在的，为什么

人类把认知统一到同一标准下的努力至今都还没成功，以及为什么这几千年来人们对真理的追求似乎总差那么一点点，而每次又正是这一点点把我们引向一个完全的新世界。这也解释了为什么20世纪初美国教育家、哲学家杜威（John Dewey）和胡适所主张的务实主义具有合理性。但敌人的敌人未必是你的朋友，虽然新的理论解释了经典认知的不足，但它绝不是肯定古代神学和“名教”存在的意义，它是一个完全不同的新世界。

人工智能是这里的一个例子。事实上非要回溯历史的话，竟然是人类先发现了它做不了的事，而后才发明了它。图灵1936年的那篇著名的论文是对哥德尔不完备定理的诠释，他提出了一个更加简明而物理上可以实现的机械设备来完成数学的逻辑推演，并在文中指出了哪些问题是图灵机做不了的，以此来拓展哥德尔的不完备定理证明。但图灵机是如此的强大和成功，那些“它做不了的事情”很快就被淹没在故纸堆里，那里，正埋着我们对付探求人工智能的不能的钥匙。

哥德尔的被遗忘，是现代科研游戏规则的直接后果。一个不申请经费、没有徒子徒孙的学问，如何能影响深远、众人周知呢？在今天的科研环境下，绝无可能啊！而不用申请经费也不用带学生正是普林斯顿高等研究院对第一批研究员的放纵。作为这些人中的两个典型代表，爱因斯坦和哥德尔是天真地相信了。在全球最近几十年以论文发表量来衡量学术工作是否有价值，学者是否有地位的评价体系下，即使哥德尔这样一个开辟了一个领域的大神，他的思想也是无法存活和传播的。他和他的好友爱因斯坦，像堂吉诃德一般成为科学殿堂里最后的骑士了。

于是我觉得我应该写一些文字，向这些人类历史上的伟大思考者致敬。尽管我自己对这门学问的认知还有很多的不成熟之处，但至少我希望它可以安抚我们面对未来的焦虑之心。当能够回溯认知的常识时，我们会发现竟然可以有更加平和的理解世界的态度。学问之道无他，求其放心而已矣。

燕园、清华园和伊河边上的小城牛津，我在象牙塔里进进出出，结识了很多好玩的人。光说姓李的，就有给我机会初试科研的李政道先生，有塑造我务实主义信条的李敖先生，有让我感受到音乐创造魅力的大哥李宗盛。跟越多的人交往，我越感觉到人类的可爱和伟大，这么好的世界，我们怎么忍心交给机器呢？

第一章

知无不言

人工智能不能做什么？如果一本书在第一章就说明白了答案，就不好卖了。既然有整本书去写，就不妨慢慢来，至少要说明我们使用怎样的工具，才能得出怎样的结论。当我们这样做的时候，我们发现，这工具似乎并不能当作一个可靠的工具，于是只好继续讨论什么是可靠性。这至少还会让我们理清楚什么是有效的思维工具，即，理性。

我所追求的全部知识，都更充分地证明我的无知是无限的。

——卡尔·波普（Karl Popper）

2018年3月20日下午，我在台北街头的一家咖啡馆写下了这样的文字：

最近很少写字，因为写字变成一件困难的事情。哥德尔告诉你这也不能写，那也不能写，在写任何字之前都不免要思考一下这话是不是严谨。所以人会被自己学习的东西所害，所影响，定义了“定义”自己，那么谁来影响“影响”呢？

雨后初晴，透过小店的窗户能看到傍晚的阳光洒在街边的小树上，也被骑电动车的骑手的头盔反射过来。骑楼的柱子后面应该有人在抽烟，虽然看不到人，但可以看到那个高度从骑楼柱子后面飘出来的白色的烟。我不得不加速喝完一杯味道还不错的淡淡的奶茶。不得不说味道真不错，但师弟要过来，他不太喜欢这家店，因为先生不喜欢。

阳光又倾斜了五度。师弟还没有来，奶茶已经下了一半，但还有厚厚的泡沫其实看不出来准确地要多少口才喝得完。点了一份千层蛋糕，味道很清淡，肉桂粉在盛蛋糕的碟子里撒出来某种符号的样子。师弟说先生不喜欢这家店，是因为这是一家西藏店。我又看那图案，还真是。

旁边两个女孩在喝着咖啡和果汁，聊着八卦。似乎是关于人怎么扮成上流社会的费劲和囧，然后是女婿和岳母之间的微妙关系。

吧台后面四个店员在忙碌着。虽然店里真的没有什么人，总共四个顾客，包括在打字的不知道什么时候要离开的我。但他们似乎真的很忙，都有事做。一辆白色的车在路边停着，双闪着灯。哦，不，两辆。我在尽量描述我眼前看到的事物，才发现其实少了时间这个尺度。有车辆从窗外经过，店员走过来轻声说，“帮您加个水”。

一辆自行车是诡异的绿色。

阳光又倾斜了几度。店面已经打开了霓虹灯，但不是五颜六色的那种，是明亮的黄绿色，另外一家是深的蓝色。

哥德尔说我永远也没法描述完整我在此刻的所有感受，虽然我努力地尝试记录。

这是一条先生每天散步的路。在路边的一家咖啡店，我喝着一杯味道淡淡的奶茶，写着描述不了什么东西的字，而两天前他不在了。

作为人类，我们使用着各种工具。也因为日常使用这些工具，我们对工具本身的可靠性或者可靠到什么程度往往忽略不计。例如，我们可以花点时间了解一下人们每天使用的语言。

人类的语言是神奇的。你可以通过语法来理解字面的意思，通过说话人的态度来了解他们的内心所想。人们可以使用同样的文字来表达不同的内涵，有时候人们的沟通并不需要纠结于文字本身，面对面的交流充斥了各种各样的信息。语气、音调的抑扬顿挫、眼神、味道和背景音乐，这些游走着的不能被词语编码和记载的信息，同样传达着人们在交流时的内涵和情绪。

但我们一样用这些丢失了很多含义的、残缺的语言，在进行着对我们人类来说最至关重要的行为——沟通。当然，因为残缺有时会带来这样那样的误解，于是我们希望创造和定义一个没有歧义的语言系统。在这样的语言系统之上，我们可以更客观准确地交流。事实上，

这本书将告诉我们这样的事情不可能做到。用天生不完美的工具人类依然能够彼此沟通是个逻辑的奇迹。我们可能最终能够依赖的是通过（奇迹般的）沟通而达成的信任，人和人之间可以原谅彼此，甚至因为理解沟通的缺陷而有了温情。这样的温情，你从机器那里无法获得，从法律条文中无法获得，从你的敌人那里也无法获得。但当我写到最后一句的时候，我觉得未必，你的敌人也是人群中的一部分。对手之间的惺惺相惜，往往成就历史佳话。但人类并不死心，我们试图用更多的数据来还原这些场景，来补充文字所不能描述的，比如通过录音来还原说话者当时的语气，通过影像资料来还原说话者的表情和手势。

语言是可以理解的吗？

大多数的人类语言会通过声音传达出来，组成语言的词汇表现为声音的频率和强度的组合。语言中会有一些词汇出现的频次高于别的词汇的频次，即使不了解这种语言，对特定的音频和强度组合出现的频率进行统计，我们都可以判定它是不是一个有意义的语言系统。比如在中文中“的”字发音出现的频率就很高，其他字会相对较少。当一种语言有充分表达能力的时候，它常用词汇的使用频率是递减的。不会孤立地出现某一个频率的音频和强度非常高，而其他又很低的情况。这样，单单通过对词汇出现的频率统计，我们就可以判断一系列的声音信号是不是代表着有意义的语言。这个原理，我们称为齐普夫定律（Zipf’s law），近些年动物学家把这个定律用于研究海豚的语言。

最早的动物语言研究在于试图了解动物间如何相互交流，人们尝试教给动物人类的语言，看它们是否可以与我们交流。这样做的问题

在于这个过程中人类很难学到动物语言。牧羊犬可以学习超过1000个人类单词，但人们能从狗身上学到多少狗的单词呢？到目前为止，人类总是试图通过教动物如何与人交谈来理解动物语言。然而教动物使用人类语言有明显的困难，想想我们学一门外语有多难，何况跨物种之间的交流呢？

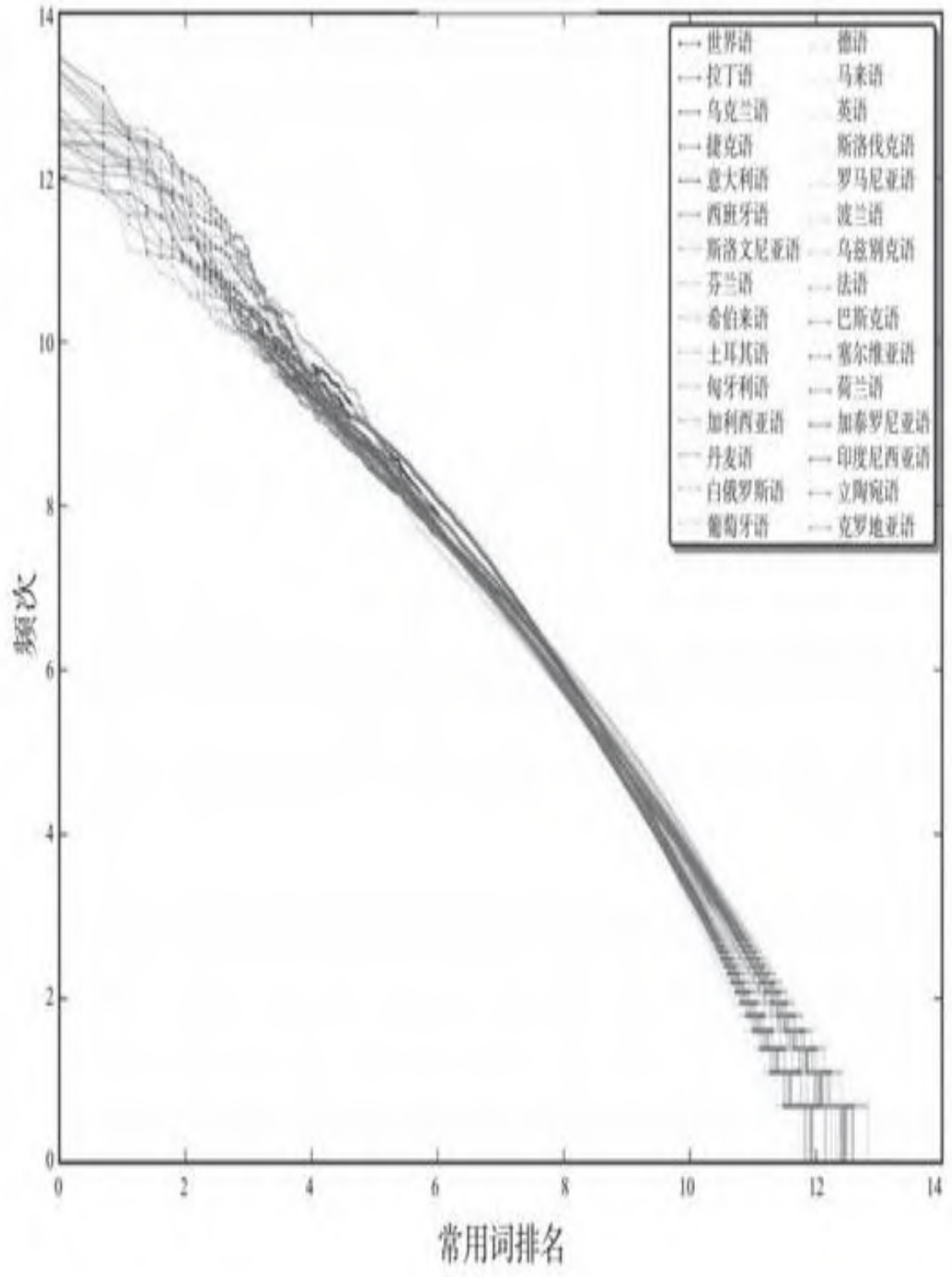

世界语
拉丁语
乌克兰语
捷克语
意大利语
西班牙语
斯洛文尼亚语
芬兰语
希伯来语
土耳其语
匈牙利语
加利西亚语
丹麦语
白俄罗斯语
葡萄牙语
德语
马来语
英语
斯洛伐克语
罗马尼亚语
波兰语
乌兹别克语
法语
巴斯克语
塞尔维亚语
荷兰语
加泰罗尼亚语
印度尼西亚语
立陶宛语
克罗地亚语
频次
常用词排名
0
2
4
6
8
10
12
14

图1–1 齐普夫定律

今天人类的语言是从七万年前一族走出非洲的智人那里继承下来的。对比他们同期的人类，这一族人有着发达的声音语言系统。这种系统的基础是口腔、舌头和喉部的独特生理结构。这样的生理结构甚至连我们的近亲猩猩都没有，更别说其他动物了，它们根本没有能力发出人类复杂多变的语音。在研究用对方的语言与动物交流的过程中，人类不得不创造一些辅助工具，因为任何一端都不能直接用天然的生理结构来复制对方的声音。学习海豚的语言就存在这个问题，海豚叫声的超高音和鲸的超低音都超出了人类的听觉范围。海豚不能说出人类的任何语言，而我们也不会吹超声波口哨。

利用齐普夫定律，人们在搜集了大量海豚叫声之后发现这些叫声确实在传递特定信息，而且是一种相当丰富的语言，能够让海豚之间进行复杂的交流，借助于人工智能技术人们甚至可以确定一部分海豚叫声所代表的具体含义。配合适当的演奏设备，人们甚至可以模仿这些声音，这使科学家和海洋生物之间实现信息的双向流动。如果真的可以理解足够多的海豚语言，人类对动物研究的现状将大大改善。海豚语言研究的成功可能使研究人员设计出与其他智慧生物交流的工具。不难想象，未来甚至会出现动物语言翻译设备。这项研究还有另外一个用处。当我们用射电望远镜瞄向遥远太空的时候，会接收到一些来自太空深处的无线电信号。如果恰巧有外星人像我们一样可以看电视、听广播，他们扩散到太空中的无线电信号就会被我们接收到。这些信号有可能像自然语言一样由高频使用和低频使用的词汇组成，而我们可以翻译它。事实上，二战时的艾伦·图灵（Alan Turing）就通过类似的方法破译了德军的密码。如果我们接收到来自太空的无线电信号，并且判定它是否符合齐普夫定律，我们就可以知道发出这些无线电信号的宇宙深处是否有像我们一样的高等生物，甚至可以知道他们每天看的电视节目在说什么。

语言是可靠的吗？

描述一个真理，至少从描述本身而言，我们需要依靠语言。但语言是可靠的吗？如果不可靠，它的缺陷是可以避免的吗？

《论语》讲：“学而时习之，不亦说乎。”小时候课本上的解释是：学东西要时常拿出来复习复习，不是也很快乐吗？我一直怀疑这快乐的真实感。因为这种“学习的快乐”是课本灌输给我的，我小时候并不具有。“纳粹头子”戈培尔（Paul Joseph Goebbels）有句名言：“重复是一种力量，谎言重复一百次就会成为真理。”于是，在熟读百遍之后，我真的以为我有了快感。

后来长大了，学了繁体字，才知道原来这个“习”字可以不这么写，繁体写作“習”。“羽”当作小鸟的翅膀，而“白”是指太阳。组合在一起可以解释为小鸟努力在天上学飞行。引申来讲，就是实践。学会了理论，要出去操演操演、实践实践，这样才能巩固，才能知行合一。纸上得来终觉浅，绝知此事要躬行。孔子讲的是做学问的方法。只有这样，放在《论语》的开篇才有意义。然而，我还是没有找到到底快乐在哪里。

到了人书俱老的时候，又想到了一个新的解释。这个“习”呢，不一定是学完了马上操练，可以是很久很久以后。“功不唐捐”是胡适晚年特别喜欢写给别人的题字。以前学的东西，很久很久以后在哪儿碰到了，拿出来用一用给旁人看，这个似乎才有别样的乐趣。学过，偶尔拿出来显摆一下，嘚瑟嘚瑟，收获一种倚老卖老的得意。

这些年我做新工程教育，希望培养与人工智能（Artificial Intelligence，简写为AI）共存的新工程师。希望新的人才能够既懂得数字技术，又能动手做东西，所谓学而时习之，培养又能动脑又能动手的新型工程人才，此乃大快乐也。

然而这样就让我们产生了困惑：当我们说一句话的时候，这句话到底是什么意思？

道可道非常道

语言是表达意义的工具。人类对于文字的敬畏可以追溯到很久以前，但古人也很早就认识到了自然语言的“不完美”。老子说“道常无名”，孔子也认为“书不尽言，言不尽意”，陆机在《文赋》中称“意不称物，文不逮意”。

但是老子的说法“知者不言，言者不知”，本身却是个悖论。而这个悖论都不是今天才被指出来的。白居易在《读老子》里说：“言者不如知者默，此语吾闻于老君。若道老君是知者，缘何自著五千文？”（善言者不如智者善于沉默，这话是老子说的。如果老子真是“智者”的话，他干嘛还要写五千个字的《道德经》呢？）

这还没完。我们先不要管语言这么复杂的事情，一步一步来，先从定义一个关于某事物的名词开始。

如何定义事物？

忒修斯之船（The Ship of Theseus）

在希腊传说中，英雄忒修斯与雅典战士们从克里特岛回到雅典，他们的船被雅典人保留下来作为纪念。随着时间推移船上的木材开始腐朽，雅典人便用新的木头来替换那些腐烂的木头。最后，船上的每根木头都被换过了。问题是，最终的这艘船是否还是原来的那艘忒修斯之船，抑或是一艘完全不同的船？如果不是原来的船，那么在什么时候它不再是原来的船了？

我们还可以对此问题进行延伸：如果用忒修斯之船上取下来的老部件来建造一艘新船，那么两艘船中哪艘才是真正的忒修斯之船？

这样的讨论同样适用于人类。人体每天都在进行新陈代谢，这样隔过一阵子组成身体细胞的全部原子都会被换掉，那么我们还是原来那个自己吗？如果不是，什么时候不是的？

当然，我们可以举出从精神、意志、社会关系等方面对个人的定义来说明，单纯依靠物理的实体来定义像人一样复杂的事物是不合适的，但对船而言呢？对花花草草而言呢？对猫猫狗狗而言呢？对猩猩猴子而言呢？接着就是复杂一点的人类了，这些物种之间什么时候产生了界限呢？

于是我翻开了词典。为了权威性，我翻开了《牛津英语词典》，为了避免翻译的误差，我翻开了英文版的。

词的互相定义

BOX, container made of wood, cardboard, metal, etc. with a flat base and use a lid, for holding solids.

我总要看看container是什么意思。

CONTAINER, box, bottle, etc. in which sth is kept, transported, etc.

这明显的是在用container来解释box，用box来解释container。看来，或者说至少，如果通过词典来定义我们说的每一个词，并试图以此理解人类的语言是有些问题的。到底这些语言有什么意义呢？它们完全可以是一些互相替代的符号。由此可见，理解一本词典，需要词典之外的常识。这些常识，来源于我们的生活、文化、教育，而不只是文字本身。常识的建立，可以通过大数据来实现吗？如果可以，那么到底要多少知识，才能构成人类交流的常识？事实上，人类的生

活中存在很多默认的知识，但深究下去，为什么用这个，为什么这么用，我们并不十分清楚，但我们在交流中毫无障碍地用到了它们。

如何学习外语？

There was a time when men were kind

从前，人们是友善的

When their voices were soft

那时，他们声音轻柔

And their words inviting

言语让人盛情难却

There was a time when love was blind

从前，爱是盲目的

And the world was a song

那时，世界是一首歌

And the song was exciting

歌声里充满了激情

There was a time

然而，有一天

Then it all went wrong

一切面目全非

I dreamed a dream in time gone by

我曾梦见那个关于往昔的梦

When hope was high

梦里我满怀憧憬

And life worth living

我的生活有它的价值

I dreamed that love would never die

我梦想着爱永不消亡

I dreamed that God would be forgiving

我梦想着上帝会宽容一切

Then I was young and unafraid

那时我年轻无畏

And dreams were made and used and wasted

梦想诞生、挥霍又被遗忘

There was no ransom to be paid

没有赎金要偿付

No song unsung

没有哪首歌会不被唱响

No wine untasted

也没有哪滴酒不被品尝

But the tigers come at night

但是猛虎从深夜中走来

With their voices soft as thunder

声音低沉像是闷雷在响

As they tear your hope apart

接着它们撕碎了你的希望

And they turn your dream to shame

让你因为曾有过的梦而羞愧难当

He slept a summer by my side

整个夏天他都睡在我身旁

He filled my days with endless wonder

他使我的每一天都充满幻想

He took my childhood in his stride

他轻易获得我的童贞

But he was gone when autumn came

而他在秋天来的时候消失在远方

And still I dream he'll come to me

我依然梦到他会回来找我

That we will live the years together

我们还将携手走过余生时光

But there are dreams that cannot be

然而有些梦不可能实现

And there are storms we cannot weather

就像暴风雨来时我们无力阻挡

I had a dream my life would be

我曾梦想那个我可能的人生

So different from this hell I'm living

与我活着的地狱绝不一样

So different now from what it seemed

与今日的宿命也绝不一样

Now life has killed the dream I dreamed

然而现实它却磨灭了我从前的梦想

2009年，我在伯克利地下室里困顿于实验和理想的时候，看了《英国达人秀》，苏珊·博伊尔（Susan Boyle，即“苏珊大妈”），一个来自英国乡下的48岁女性，在舞台上演唱了这首出自音乐剧《悲惨世界》中的歌曲*I Dreamed a Dream*（《我做了一个梦》）。这首歌触动了我对未来不确定的期许，于是我把它翻译成了中文。

虽然我们希望不同的语言可以做到“信达雅”的翻译，然而大多数语言之间是无法对应的。“信”，指翻译内容的真实可靠；“达”，指翻译的通顺流畅；“雅”，是翻译的最高境界，即文字要优美。但这一关往往不是翻译者文学素养的问题，能够做到“雅”的翻译，要碰运气。

在语言翻译中，“信达雅”是可遇不可求的。有的话，会心一笑；没有，也无可奈何。人们为了避免歧义，利用数据的冗余，多说几句来减少误会。Dialogue翻译成“对话”，只能说是“信”，翻译成“大唠嗑”，是“信”和“达”，而碰到个读者是东北人，这第二种翻译就是“信达雅”了。

英语里有个Buchoski悖论：My younger brother is older than I am。

英语里brother这个词没有哥哥和弟弟的区别，所以这一句看起来就是矛盾的。但事实上“我”可以有两个哥哥，那个年纪轻一点的小

哥哥（younger brother）自然比“我”的年龄大。但是younger brother对应为中文时又有“弟弟”的意思。

我刚到英国的时候，别人问我的名字，我跟他们说“zhaoyuan”，这两个音在英语里是完全发不出来的，德国人会好一点。这本书里用到外国人翻译为中文的名字，除非大家已经习以为常的名人的中文译名，比如爱因斯坦，其他的我通常会保留原来的英文拼写，以避免强硬地翻译成中文，让读者更不知道是谁了。至于我的英文名，留学那些年为了避免英国人发不了“zhaoyuan”的尴尬，我用过“Joe”“John”“Zhao”，最后还是强行让他们习惯发“zhaoyuan”的音，哪怕极难听。

不同的语言会有不同的发声习惯。汉字的发音多为单音，总共只有400多个，而常用的汉字有一万多个，因此要在每个音上加上声调，加上组合字，这样来形成丰富的语言。多数的拼音语言会有大量的辅音在元音前后，所以不需要有音调来辅助。这样形成的说话习惯是，我们会觉得外国人说中文为什么不会“抑扬顿挫”，而中国人说英语，会因为习惯的声调变化而成了所谓的Chinglish（中式英语）。不是我不愿意让你叫我的中文名，是你发不了这个音啊！就更别说那些名字后面的来自家长的期望和深厚的文化背景了，名字翻译实现“信达雅”是很难的。

语言甚至在一定程度上塑造了我们的思维习惯。当然，反之亦然。母语为俄语的人会比母语为英语的人更快地区分出深蓝色和淡蓝色的不同，在俄语中，这是两个不同的词汇，г о л у б о й（发音同goluboy，意为淡蓝色）和с и н и й（发音同siniy，意为深蓝色）。

英国历史上早期的贵族讲法语，平民和下等人讲英语。这样就形成了有意思的现象，英语里面表示活物的家畜所用的单词源于英语，而表示能吃的肉类源于法语。比如说羊，用sheep和goat是英语，说羊

肉用mutton是法语。牛可以跑的用cow，这是英语，用来吃的叫beef，这是法语。能吃的猪肉pork是法语，但pig（猪）是英语。

我之前读的那所中学有西藏班，西藏班的同学既要学习藏语也要学习汉语。学习汉语的目的之一是通过汉语来学习物理、数学等科学知识。因为这些科学课中所用的一些词汇没有对应的藏语词汇。但在清代，所有知识在满、汉、回、蒙、藏文中是对应的。现代科学引进中国不过一百多年，藏语就完全没法儿跟着演化。同样的困难也存在于写这本书，做科学的人越来越喜欢用英语来表达新概念，因为中文并不能跟上最新的进展。新的语言和词汇的形成，需要大量的使用者促成新的表达习惯和共识。而本来可以被广大语言使用者接触的搜索引擎，也因商业信息的覆盖而很难查到有用的信息。危言耸听地说，也许过不了多久，汉语也可能会被最新的科学进展逐渐隔离，成为一种古老的语言而被博物馆化。

第二章

基本的逻辑

多少逻辑叠加在一起就不是逻辑了？至今人工智能不管算法如何炫酷高深，都还是要基于图灵机来完成，图灵机是人们用来实现逻辑推演的工具。因此要了解人工智能做不了的事，就要先从了解逻辑机开始。

我们首先从逻辑的角度来看能否通过有效的逻辑工具来避免语言里这些似是而非的问题。至少，我们会看到，在一系列利用语言沟通的问题上，逻辑是有用而必要的。一个逻辑系统，至少有一些基本的规则，这些规则将保证说话的人不再说一些前后矛盾的话。这看似容易，但当我们犯了逻辑错误的时候，我们往往不知道。而正因为没有逻辑，我们对自然界的观察始终没法积累成有意义的衍生品——新的知识。

逻辑的基本规则，至少包括以下内容：

- 同一律（Law of Identity）
- 矛盾律（Law of Contradiction）
- 排中律（Law of Excluded Middle）
- 充足理由律（Law of Sufficiency）

同一律

在用语言进行逻辑推演的过程中，必须保持概念本身始终具有同一性。同一律要求在同一推演过程中使用某个概念时，必须自始至终在唯一的确定的意义上使用这个概念。同一律只是要求所讨论的命题保持同一，要求事物前后一贯。不能将其理解为要求“命题断定的情况必须与客观情况一致”，就是说，不能把命题事实层面的真或假混

同于命题与其自身是否保持了同一。同一律并不保证所讨论的对象本身的真实性。

同一律在逻辑推演中的作用在于保证思维的确定。思维只有具有确定性，人们才能进行正常的交流。只有遵守同一律，我们才不致于产生“混淆概念”和“转移话题”的逻辑问题，才能让讨论正常进行下去，写文章说话才能主题明确、思路连贯。同一律保证对话有中心，辩论不偏离议题。

但就同一律本身而言，它不能推出任何有用的新知识。因为语义本身可以被无穷诠释，所以同一律的两个对象，必须严格一致，任何差别都会被扩大，从而使同一等式的两端并不同一。

苹果是苹果

小明去买苹果，看中了10元一斤的苹果。小明跟售货员说要一斤，售货员给他拿了20元一斤的苹果。苹果还是苹果，但价格不同了。但即使价格都是10元一斤的苹果，也有品种的不同。即使品种相同，也有新鲜程度不同的苹果。小明想要的苹果是“那个”（独一无二的）苹果，但事实上，连这个也做不到。即使是看到的同一个苹果，但当小明决定了要买，拿起来看的时候发现苹果的另一面被虫蛀过，这已经未必还是刚才他想买的那个苹果了。

这样的推论已经脱离了经典的逻辑范畴，而开始涉入我们将要谈的哥德尔不完备问题的范畴。我们不得不通过常识和适当的宽容来达成沟通的共识。否则，除非不做任何拓展，同一律才能严格地成立。但同一律严格成立的时候，我们不能获得任何新的知识。它的两端不得不严格相同。

矛盾律

矛盾律的要求是：在同一逻辑过程中不能同时有两个互相矛盾的概念或者陈述同时为真。“A”和“非A”不能指称同一个对象。一个判断不能既断定某对象是什么，又断定它不是什么，不能同时肯定两个互相反对的陈述都是真的，必须确认其中有一个是假的。

矛盾律的主要作用是保证逻辑过程没有矛盾，前后自洽。无矛盾是逻辑推演必不可少的基本条件，遵守矛盾律是构造科学体系的基本要求。科学常常是在发现逻辑矛盾、逐步排除和解决矛盾的过程中发展的。

《韩非子·难势》中有讲：

> 客曰：“人有鬻矛与盾者，誉其盾之坚，‘物莫能陷也’，俄而又誉其矛曰：‘吾矛之利，物无不陷也。’人应之曰：‘以子之矛，陷子之盾，何如？’其人弗能应也。”以为不可陷之盾，与无不陷之矛，为名不可两立也。

这段话的大意为：有一个卖矛又卖盾的人，他先夸他的盾最坚固，无论什么东西都戳不破；接着又夸他的矛最锐利，无论什么东西都能刺透。旁人问他，如果用他的矛来刺他的盾会有什么结果，他回答不上来了。

这是一个既不可以同时为真，又不可以同时为假的命题。前提出现矛盾，也就无法推出结论。这并不仅仅是卖矛和盾的人才犯的“低级错误”，对于严格的数学家、逻辑学家或者物理学家而言，我们的每一步推理，都含着很多默认的常识。这些常识如果不被提取或者抽象化，我们甚至难以觉察到它们事实上被我们毫无顾忌地使用着，但这些默认的常识之间一定没有矛盾吗？我们如何知道并确信？

一个系统不能在同一概念中包含互相否定的思想。例如“尖而圆的脸”“不是房子的房子”“方而圆的桌子”等。这些话每当这么说的时候都特别像佛教语言，尤其是你以闻得出文字味道的敏锐欣赏这些话时。“一个不是房子的房子”，或者，“一座不是山的山”。又例如，有人会说，“有个神秘的地方从来没有人去过，去了的人也从来没有回来过”。这既肯定了“所有人都没有去过”，又肯定了“有的人曾经进去过”。还有一种矛盾不会同真，可以全假，或至少有一假，例如“此地无银三百两”和“此地无银两百两”。

排中律

排中律的内容是：在同一逻辑过程中，两个互相否定的陈述不能都假，必有一真，反之亦然。

排中律否定了大而全的理论。我们希望有一种理论，可以包括所有的正确内容，这样就可以立于不败之地。比如可以有矛盾统一的理论。自然语言可以这么说，但数学上没法表达。一旦选择了骑墙的态度，就是哪边对站哪边，很容易形成一套包罗万象的理论。然而对逻辑和现代科学而言，这套理论不会有任何价值，包罗万象的理论也一样包罗了错误。

矛盾律和排中律都默认了二元论的假设。我们确实可以把任意的系统分割成两个可能，再把其中的可能分割成更多的两个和两个的组合。为了使我们的陈述具有相对的正确性，对任何一个陈述，都可以定义对错、真假的二元分析。事实上，除了我们看到的很多可以被二元化的模型之外，很多现实的对象是无法被二元化的。我们无法区分一个莫比乌斯环的正面和反面。量子力学也提供了大量不得不同时处在两个态之间的对象，我们把这叫作叠加态。

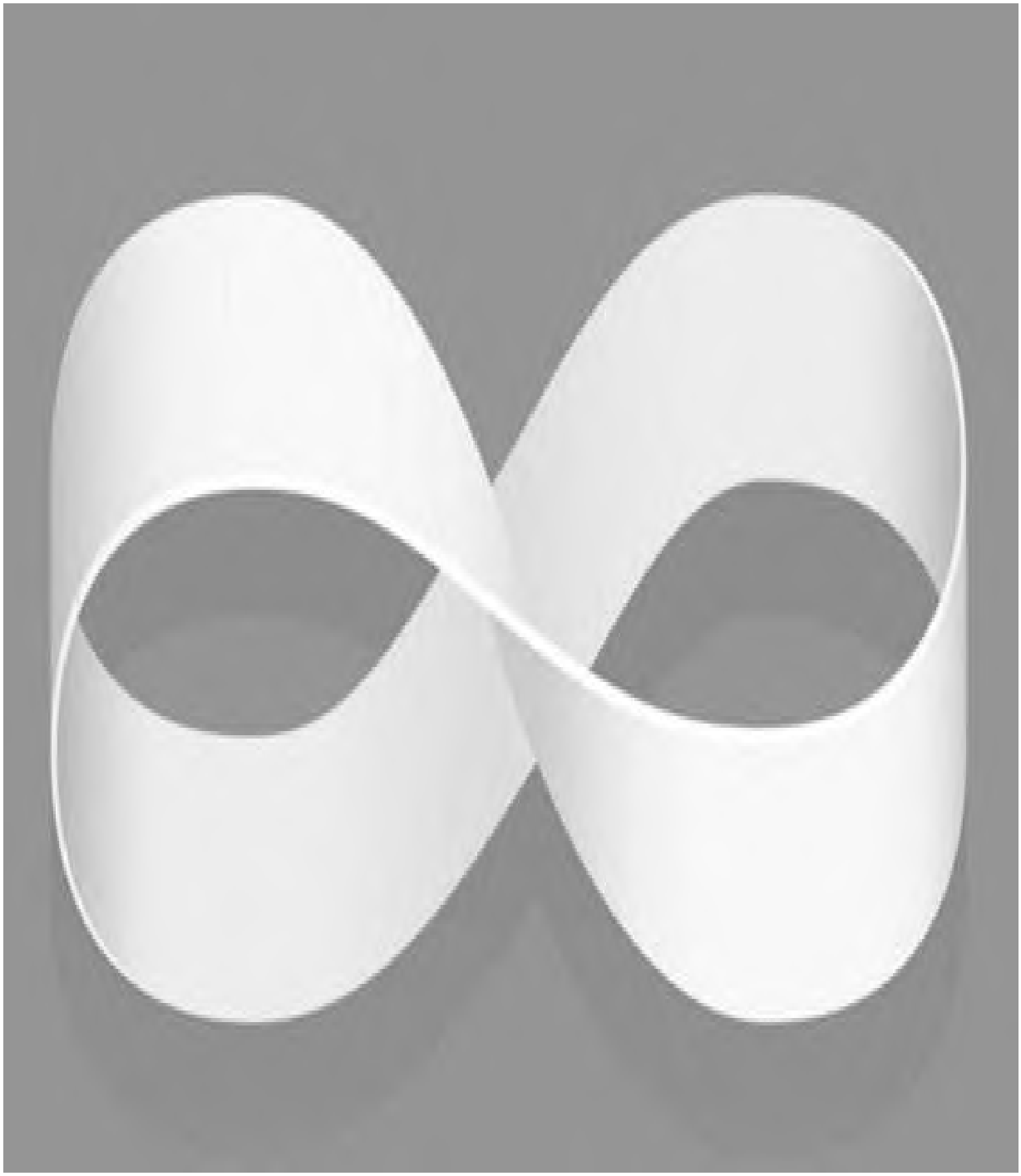


图2–1 莫比乌斯环

排中律和矛盾律是一种重要证明方法——反证法的基础。能够使用反证法的逻辑依据在于我们假设了如果错误的假设得出错误的结

论，那么假设的另外一面就是对的。但它不适用于一类我们将会在这本书里提到的悖论，比如“这句话是错的”。

充足理由律

充足理由律的内容是：在同一逻辑过程中，一个陈述被确定为真，总有其充足理由。事物的存在，总有其道理，这个道理是充分的。

充足理由律的逻辑要求主要有两条：

1. 理由必须真实；
2. 理由与推断之间有必然的逻辑联系。

充足理由律本身并不能为人们提供真实理由。在一个论证中，理由究竟是真是假，这不能由充足理由律来确定，它只提出了这样的要求来保证论述合乎逻辑，理由本身的真假只能由实践和具体的科学实验来解决。

充足理由律是前三条逻辑规律的必要补充。我们在前三条规律的基础上，保持了概念和判断的确定性之后，还要更进一步，指出判断与判断之间的联系具有必然性。在指出事物是什么之后，自然要进一步解释事物为什么是这样，而不是那样。只有遵守这四条逻辑规律，我们才能做到概念明确、判断恰当、推理有逻辑性和论证有说服力，也才能判定论证是否合乎逻辑。

戈特弗里德·莱布尼茨（Gottfried Leibniz）认为“推理要建立在矛盾原则和充足理由原则之上。任何一个陈述如果是真的，就必须

有一个为什么这样而不那样的充足理由，虽然这些理由总是不能为我们所了解”。亚瑟·叔本华（Arthur Schopenhauer）把充足理由律分解为四种形式：生成的充足理由律（因果律）、认识的充足理由律（逻辑推论）、存在的充足理由律（数学证明）和行动的充足理由律（动机律）。

充足理由律的因果律背景暗示着宇宙中的事物都不能自我解释。没有什么事物是其自身存在的理由。如果一个事物是其自身存在的原因，那么这个事物必然先于它自身而存在，这显然是不可证的。因此，这四条逻辑基本定律构架的知识系统是为超自然力的存在留下伏笔的。然而，因果律在我们将看到的某些情况下，甚至应该在更普遍的情况下，是可以违背的。当我们考虑到量子力学的范畴，因果律的正确性会接受诸如约翰·惠勒（John A. Wheeler）的后选择实验的考验（见《量子大唠嗑》[1]）。这样接着就会对充足理由律提出新的考验。惠勒的后选择实验描述了这样的事实：此刻的现象可以被未来的事件影响。这意味着，附着于时间顺序上的因果律会有问题，此刻的选择决定了过去物体的行为！而这个思想实验居然被一系列物理实验证实了！

矛盾律与充足理由律都是真理的逻辑标准或形式标准。矛盾律是反面的标准，因为遵守矛盾律的思想不一定为真，而违反矛盾律的思想不可能为真；充足理由律则是正面的标准，因为遵守充足理由律的思想一定是有根据的，是从一些原则得出的而且不会导致假的结论的思想。任何一个科学思想体系都要求推理有逻辑性和论证有说服力，也都必须同时遵守这四条逻辑规律，缺一不可。我们说怎样才能比较有格调地聊天，首先应关注聊天本身是否符合逻辑的基本要求。这里我们举几个常见的逻辑错误。

违背同一律

- 歪曲对方观点进行攻击

李雷：传统中药配方在未经临床测试的情况下，是不能使用的。

韩梅梅：你怎么能丢掉中华民族的传统遗产？这种做法会让传统文化断子绝孙的！

显然韩梅梅歪曲了李雷的话。李雷的观点是“不能使用未经临床测试的中药”，韩梅梅将其歪曲为“丢掉中华民族的传统遗产”。

- 滑坡谬论：将一连串的“可能性”转化为“必然性”，得出不可靠的结论

有人大代表在“两会”上提案同性婚姻合法化，对此某评论员质疑：“如果仅仅是当事人自愿便可结婚，那么，父女、兄妹、母子自愿结婚可不可以？三个人结婚可不可以？人和动物呢？人和板凳呢？”

仅从同性可以结婚这一前提出发，连续利用一系列微弱的“可能性”，推论允许同性结婚就会导致“人与板凳结婚”。这就犯了滑坡谬论的逻辑错误。

违背充足理由律

- 有前后时间顺序的事情，未必有因果关系，很多时候相关性不代表因果性，有其他原因作为解释

从清朝初年到18世纪末，中国人口数量从不到一亿增长到三亿左右，说明康雍乾三代帝王对国家的治理水平远超前代，所以这三代被称为康乾盛世。

当然不一定。人口增长的原因是多面的，红薯和土豆在明末从南美传到中国，大幅度地提高了粮食产量；医疗技术的积累和新技术的传入也大幅降低了新生儿死亡率并且延长了普通人的平均寿命。相关性不代表因果性，高相关系数也不能说明因果性更强，改变实验条件常常可以检查这种结论的可靠性。

再例如，新医学统计证明，一个人的寿命和这个人成年后手掌大小的关系密切，具体表现为——手掌大通常寿命短，而手掌小寿命则长。太奇怪了！事实是，男性寿命通常比女性短，而男性的手掌则通常比女性的大。

逻辑和佛性

语言是思维的载体之一，思维借助文字符号表达出来，语言的运用反映了思维的逻辑。古代文化存在的一个重要问题在于其大量的讨论不符合逻辑的基本要求。现在我们知道了，逻辑有自己的基本规律。然而佛教的陈述往往并不遵循这些规律。我们以P来代表一个对象，它可以是明镜台，也可以是菩提树。佛教经典的论述方式往往是：

如果*P*是真，那么*P*是假。

佛教论述扩展的一般形式就是：

如果*P*是*Q*，那么*P*不是*Q*。

铃木大拙在《禅：答胡适博士》（Zen: A Reply to Dr. Hu Sih）一文中说：“我们一般推论：*A*是*A*，因为*A*是*A*；或者，*A*是*A*，所以*A*是*A*。禅学同意并接受这种推论方式，但是，禅学有它自己的方式，这种方式并不是一般人可以接受的方式。禅会说：*A*是*A*，因为*A*不是*A*；或者，*A*不是*A*，所以*A*是*A*。”

唐代禅师青原惟信谈到其对禅体验的三个境界时说：“老僧三十年前未参禅时，见山是山，见水是水。及至后来，亲见知识，有个入处，见山不是山，见水不是水。而今得个歇处，依前见山只是山，见水只是水。”其中“见山不是山，见水不是水”是一种典型的逻辑谬误。

在佛学经典里这类例子不胜枚举。如我是他，但他不是我（违反矛盾律）；得即是失（违反矛盾律）；既不是肯定也不是否定，二者都不对（违反排中律）；勿言生，勿言无生（违反排中律）。它们背后的禅理是语言无法达到的，也是不符合逻辑的。不符合逻辑可以推出任何结论，这也意味着推不出任何结论。

菩提本无树，明镜亦非台；本来无一物，何处惹尘埃。

据说六祖慧能年少时在街上听到有人读《金刚经》中“应无所住，而生其心”这一句话后，豁然感悟，于是决定出家，后来作出这几句偈诗，表达出“空”的思想。慧能还提出过反排中律的命题，如“无方圆大小”“无有头尾”，认为诸如“方圆”“大小”“头尾”的矛盾也是“空”的。类似的一首偈诗是南北朝时期禅师善慧（傅翕）写的：

空手把锄头，步行骑水牛；人从桥上过，桥流水不流。

佛性会讲这些悖论超越了理性，禅师运用起悖论来反而显得自在坦然。但这也是对逻辑的不清晰和不求甚解。这并不是因为古人只受中国文化或者东方文化的影响，而是因为他们认识有限，思维没有逻辑作为牵引而所面临的共同问题。同样，今天的基础教育中大多数人也没有机会接触到逻辑课程。我们的论证逻辑以很浪漫的方式出现在语文课上，举例论证、道理论证、比较论证和比喻论证，这议论文写作的“四大论证方法”在中学语文教科书中存在了许多年，极大地影响了很多人的写作和论说的思维脉络。可惜的是，这“四大论证方法”不是逻辑，甚至是谬误。

1.举例论证：列出诸多相似的事例来证明论点的成立。

孟子在《生于忧患，死于安乐》中说：

> 舜发于畎亩之中，傅说举于版筑之间，胶鬲举于鱼盐之中，管夷吾举于士，孙叔敖举于海，百里奚举于市。故天将降大任于是人也，必先苦其心志，劳其筋骨，饿其体肤，空乏其身，行拂乱其所为，所以动心忍性，曾益其所不能。

这段话里，孟子举了六个出身穷苦却最终功成名就的人的例子，得出“天将降大任于是人也，必先苦其心志，劳其筋骨，饿其体肤”的结论。然而，很不幸，对绝大多数人而言，苦难只是纯粹的苦难，好的家世却意味着更好的教育、更宽阔的眼界和更多的成功机会。

我们生活的空间够大，时间够久，也有更多的例子可以举。人们先选择观点，再选取支持这个观点的案例，忽略甚至隐藏反面的例子，这同样是逻辑错误。在逻辑论证中，我们更关心可以否定观点的反例。证明一个命题放之四海而皆准是困难的，因为我们要去验证每一个例子，在一般情况下，这样做非常困难。只有一种情况是特例，即命题是可以用数学归纳法来证明的，自然数的性质保证了这个证明方式的有效性。对于一般的归纳总结，在有限的案例中得出的普遍规

律是不符合逻辑要求的。举例论证中举出一两个例子来说明一个普遍的规律，是非逻辑的。

事实上证明一个结论，尤其是通过大量案例来得出一个普遍的结论所用的归纳法，是非常脆弱的论证方法。这种方法严格防止反例，即论证时要把所有的可能都尝试一遍，这在大多数的情况下是不可能的。不管有多少正面的证据，一个反面的证据就推翻了这个普遍结论。因此用举例的方法来论证一件事情非常难，而推翻一件事情却很容易，一个反例就够了。但这个举例论证往往成了我们常用的论证方法，比如我们在谈起某一位中医的神奇时，都会说自己的邻居或某某熟人，癌症好几期了，都是这神医用药，甚至不用药给治好的，所以他灵啊，不灵主要是你不信的问题。然而这样的举例论证不具有推广意义上的普遍性，他被治愈你未必能行，人体是复杂系统，病好不好，各人体质有差别，常常有别的原因作为解释。

2.道理论证：引用经典著作中的名人名言、权威的话来证明论点的成立。

科学做的事情是在挑错，在不断地否定既有的权威。不要说权威，掌握了知识的个别人在某一领域中即使有深厚的积累，在别的领域中也可能缺乏基本的常识，这样的例子并不少见。“权威说的话”是双方都偷懒的途径：作为“权威”的人不用花精力去说服别人，只要把牌子一举就够了，“我是专家”；引用“权威观点”的人也不用花精力去证明，因为“权威”会背锅，“某某专家说了”。但这样的论证都谈不上逻辑，是最有害的。它不但妨碍了我们对内容真实性的探索，而且建立了人对知识的等级观，把“人治”的观点引入了知识世界，最终妨碍了人对知识的深入探索。这样做的不良后果在于，“权威”死了，世界就崩塌了。

列举某某名人说的话，似乎名人所说的话就是真理，就可以用来证明某个结论是对的。这本身也是毫无道理的，姑且不说名人在说某句话的时候一定有上下文，而语言本身也可以被无穷诠释。名人的话

包括名人自己也不断地被历史重新审视。所以他说的话难道就可信吗？就能成为论证的基础吗？我们把绝对真理相关的学问和方法统称为“神秘论”，在《量子大唠嗑》这本书里，我们这样定义了神秘论：

1. 相信有绝对真理的存在；

2. 相信有绝对权威的存在；

3. 所有证明都是为了证明绝对权威掌握了绝对真理。

比照这三条，我们教育灌输给学生的论证方式，是神秘论化的论证方式，是古代的思维方式，是不科学、不现代的。这样的论证方式会坑害中国，使得中华五千年文明原地踏步而无所发展，如果继续这样教下去，也必毫无益处。

3.类比论证和比喻论证，指的是一种将两种事物进行对照，得出某种结论的论证方法。

董仲舒为了证明其“天人合一”理论，在《春秋繁露》做过这样的类比：

> 天以终岁之数，成人之身，故小节三百六十六，副日数也；大节十二分，副月数也；内有五藏，副五行数也；外有四肢，副四时数也；占视占瞑，副昼夜也；占刚占柔，副冬夏也。

这段话主要说的是，人全身有366个关节，对应天一年有366天；人有12个大关节，对应天一年有12个月；人有五脏，对应天有五行；人有四肢，对应天有四季；人眼有开合，对应天有昼夜；气息有刚有柔，对应季节有冬夏。然而，仅就事实而论，人体并无366个关节——证据是按照观点凭空臆造出来的。

类比推理和比喻论证往往只能提供某种“或然性”。这种“或然性”在人类认知世界的历史进程中，有着很重要的地位。它可以将事物描绘得更形象，将问题表达得更清晰，达到易于理解的效果；它能够给我们带来认识新事物的好的启发，构成一定的“大胆假设”。然而，要做到紧跟在“大胆假设”后面的“小心求证”，则需要“同构映射”，即证明两类事物之间从假设到论证是严格一一对应的。这件事情，只在数学领域中做得很成功，我们接下来还会谈到。但在自然形成的语言中，严格证明两个事物是同构的，非常之困难。比喻论证，强调了相似的部分而忽略了不同的部分。相似的部分具有某种特性，不能被推论为具有这个部分的事物就具有类似的其他特性。所谓“上善若水”，我们无法说水具有某种特性，而人性也具有某种特性，因此，水的另外特性，也是人性的特性。比喻有助于直觉的形成，但不构成论证的基本方法。因此在通常情况下，无论类比论证还是比喻论证，逻辑上都是不成立的。

这四种所谓的论证方法，从逻辑上来说都是错误的。只能作为写作时的修辞手段，远远谈不上论证的作用。在帮助人思考方面，不仅没有益处，还大大有害。不要抱怨这本书枯燥，这本书努力避免使用这些不靠谱的论证方法，以免让本来就深邃的概念更加令人糊涂。忍忍就好，值得的。

逻辑的自洽和完备

能不能完全依靠客观的逻辑来排除某些人的直觉，甚至是一群人的直觉，来建立普适的知识体系？比如说不用假定任何假设就可以证明一个系统没有内在矛盾，那么怎样才能算一个完美的理论系统呢？

构架于严格逻辑体系的理性系统至少应该是自洽（Consistency）和完备（Completeness）的。

自洽，要求理论有效，不能自相矛盾；完备，要求理论有用，能够对体系内有效的命题进行判断。

如果一个理论系统前后矛盾，那么它不会有任何用处。在一个矛盾的理论体系里，任何命题都能被证明。同样，在一个有用的理论体系中，用这个理论体系所规定的名词给出一个命题时，这个命题的对错应该是可以通过理论的基础假设和合理的逻辑推演来判定的。否则的话，连对错都不能告诉我们的理论体系有什么用呢？这两点都做到的话，才是一个完美的体系，“完”，在于完备，“美”，在于它至少不是自相矛盾的。

自洽性要求

符合逻辑的理论体系应该是自洽的，也就是说它是前后一致的，没有矛盾的，不光是一句话里没有矛盾，是理论体系内处处不能有矛盾。否则的话这些矛盾总能碰到一起，从而使整个体系丧失公信力。建立一套知识理论体系，首要的任务是证明命题系统的无矛盾性。这个要求很自然，否则如果从命题中推出相互矛盾的结果来，那么这个命题构成的任何说法都毫无价值。自洽的要求极其朴素，你不能刚说了什么，接着就又说自己刚才说的不对。在说话的时候，不能自带武器来攻击自己。至少也是别人用另外的话语和证据来说明你的言论有不当的地方，而不是都不用别人出手，自己就把自己打败了。

自洽源于一个可信服的原理，即逻辑上不相容的命题不可能同时为真；如果一组命题是真的，那么这些命题逻辑上是相互协调一致的。如果说一个知识系统是自洽的，那么不可能得出$0\neq0$的结果；或者不能出现这个系统中的一个命题与它的否定命题都是对的。

一句话来说，不自洽的假设证明一切。

罗素是教皇

从单纯的逻辑上来讲，不自洽（存在矛盾）的假设可以推论出任何荒谬的结论，哪怕推理过程无懈可击。伯兰特·罗素（Bertrand Russell）证明他能从“2+2=5”推出“罗素是教皇”。罗素证明如下：

由于2+2=5，等式的两边同时减去2，

得出2=3；两边同时再减去1，

得出1=2；两边移位，

得出2=1。

教皇与罗素是两个人，既然2=1，教皇和罗素就是1个人，所以“罗素就是教皇”。

罗素因此也可以是任何人。这个不靠谱的结论，就是由一个与已知知识矛盾的假设引发出来的。我们既然在算术系统中可以证明真的命题2+2=4，那么2+2=5便成为与之矛盾的命题了。同样，不自洽的体系内可以推出任何结论。

完备性要求

如果说一个知识系统是完备的，那么这个系统中的所有命题都是可以被证明的，即每一个系统内的真理都对应着一个定理和它的证明过程。

完备性是另一个知识理论体系的朴实的要求。我们希望学习一套知识，就会给我们个可靠的说法。问一个问题，总该给我们一个确定

的答案，对还是错，给个准话。辛辛苦苦建立起来的一套没有矛盾的理论体系里有很多很多命题，我们总希望每个命题都能判断对错。我们这里说的命题，就是简单到像“2+2=4”，“在没有外力作用下物体会沿着直线做匀速运动”，“明天会下雨”，“柿子10元钱一斤”。当然也有复杂的，比如哥德巴赫猜想，即“大于4的偶数，总能写成两个质数的和”。这些命题并不属于同一个命题体系，自然数建立了算术的知识体系，气候建立了气候的，而物理学建立了物理学的。这些知识体系有一个共同的特点：它们必然建立在某一些“不证自明”的前提条件上。传统意义上这些前提条件被称为公理，但可能称为“公设”会更合适一些，要明确它们不过是共同承认的“公允的假设”，未必有理。“理”给人一种更强的默认是对的含义，而事实上，我们从本书里可以看到，这些假设未必总是真的。我们常常会看到理论系统所依赖的不证自明的公理会随着新的证据出现而动摇，而在本书接下来的讨论中我们也会发现，公理本身并不能够证明它的正确性，因此说它是“公允的假设”，比“公允的道理”或“公允的真理”更符合事实。

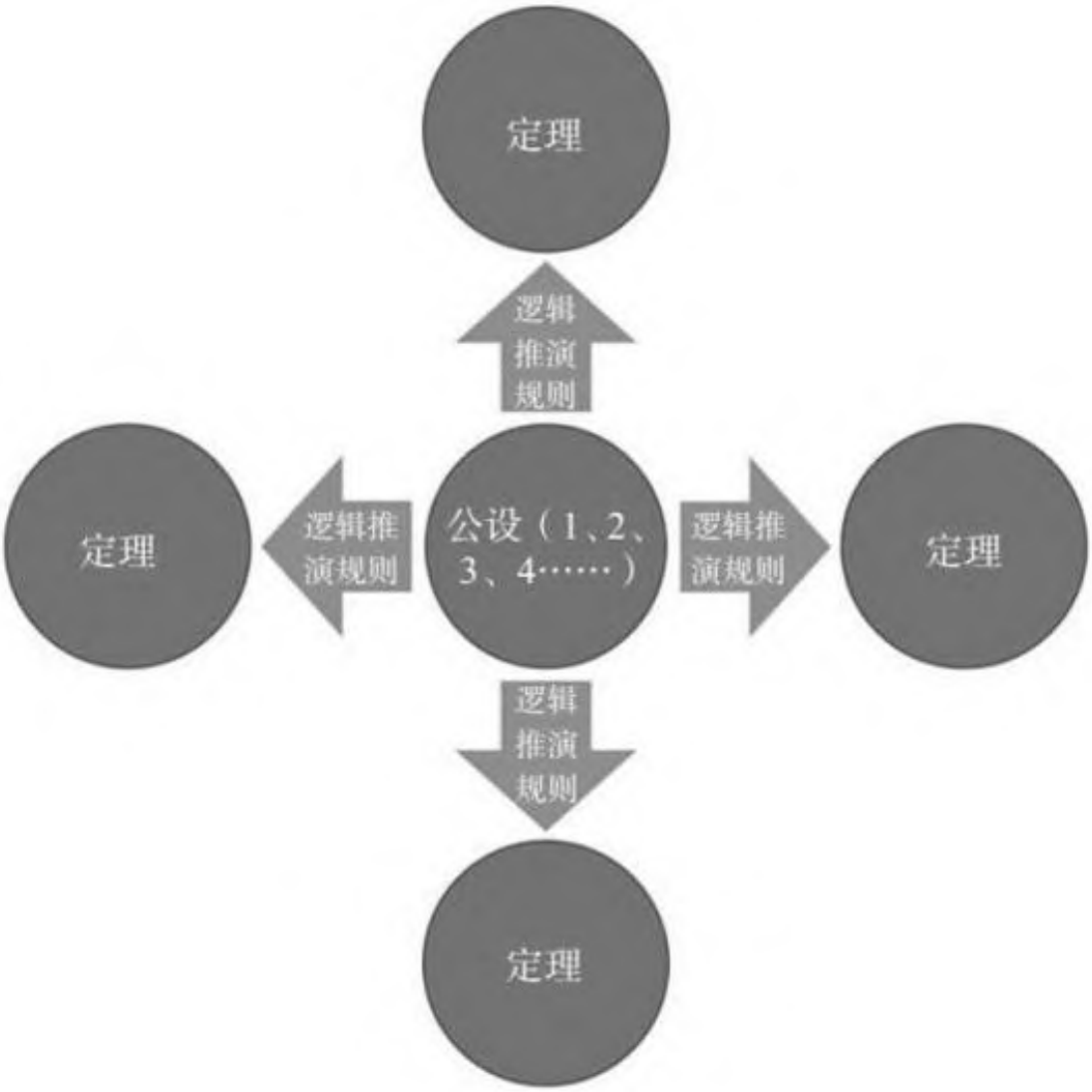


图2-2 公设与定理

人们建立一套知识体系，希望这套知识体系能够回答我们提出的问题。如果我们用这个体系的语言工具不能够回答我们提出的问题，这个体系就还不完备，那应该继续扩展这个体系，让它能够解答我们提出的问题。同时我们希望这套理论可以描述它自己的问题。一个命题用这套新的理论描述出来，我们应该有办法来判别它是对的或错的。我们也常常利用自洽性来检测完备性的要求，即使用反证法，但

凡是用了反证法的证明，都要在系统是自洽的这一假设下能得到矛盾的结论，这个证明或者证伪才是有效的。

构建一个知识体系

符合逻辑要求的知识理论体系的建立开始于一些不证自明的公设。公设是一系列存而不证的假设，因此我们希望公设越少越好。公设之间应该是互相独立的，若是有一条公设能够由其他公设推论出来，那么这条公设就是多余的。一套公设应该容许我们对这门科学下的每条命题加以判别其为真或为假。我们通常也把这样的公设体系和它们基于逻辑推演规则得出来的命题系统总称为形式系统。它是构建一切已知科学知识体系的基础。当然，我们用来做逻辑推演的规则本身，也是公设内容之一，只不过它们在各种理论体系中都被普遍地使用，从而成为逻辑的通用要求。像其他公设一样，这些作为推演规则的公设，一样可以被怀疑。

我们举一个例子。任何学习过几何的人，无疑都会认为它是一门完美的由演绎而产生的学科。由演绎而产生的学科与由经验产生的学问不同。在经验学问中，一个定理只要和观察相一致就会被接受。但在几何学中，一个定理是经由明确的逻辑证明得出的结论。这种对定理要求的观念可追溯到古希腊人，正是他们发明了所谓的公设方法，并且基于这种方法系统发展了几何学。

总而言之，公设方法，就是先不加证明地将某些命题当作公设或前提，例如“通过两点只能画一条直线”的公设，然后从公设导出系统中的所有其他命题为定理。公设构成了系统的基础，定理则是从公设仅依靠逻辑原则的演绎所得出的被证明为真的“上层建筑”。

我们常见到的大部分棋类游戏也是形式系统，比如中国象棋，它们是严格按照公设体系建立起来的。系统的语法将指明，什么样的符号串（陈述）是合适的（这代表棋盘上的有效结构，例如黑象不能位于红方半场内）。这样，语法正确的句子就代表下棋的任何一个阶段的可能状态。形式系统的推理规则不过是把一个合适符号串转换成另一个合适符号串的各种不同途径而已，即把一个合理的棋局推演到另外一个合理的棋局的“走棋”。换句话说，推理规则代表在下棋的任何阶段可允许的各种走法，而下棋的公设对应着开局时初始格局棋子的摆法。

操作规则（在中国象棋的公设化系统中）

公设：开局的棋盘摆法。

推演规则：象棋每一个棋子的走动规则。

定理：符合规则走出来的新的棋盘局面。

陈述：任何一种棋子在象棋盘上的摆放方式。

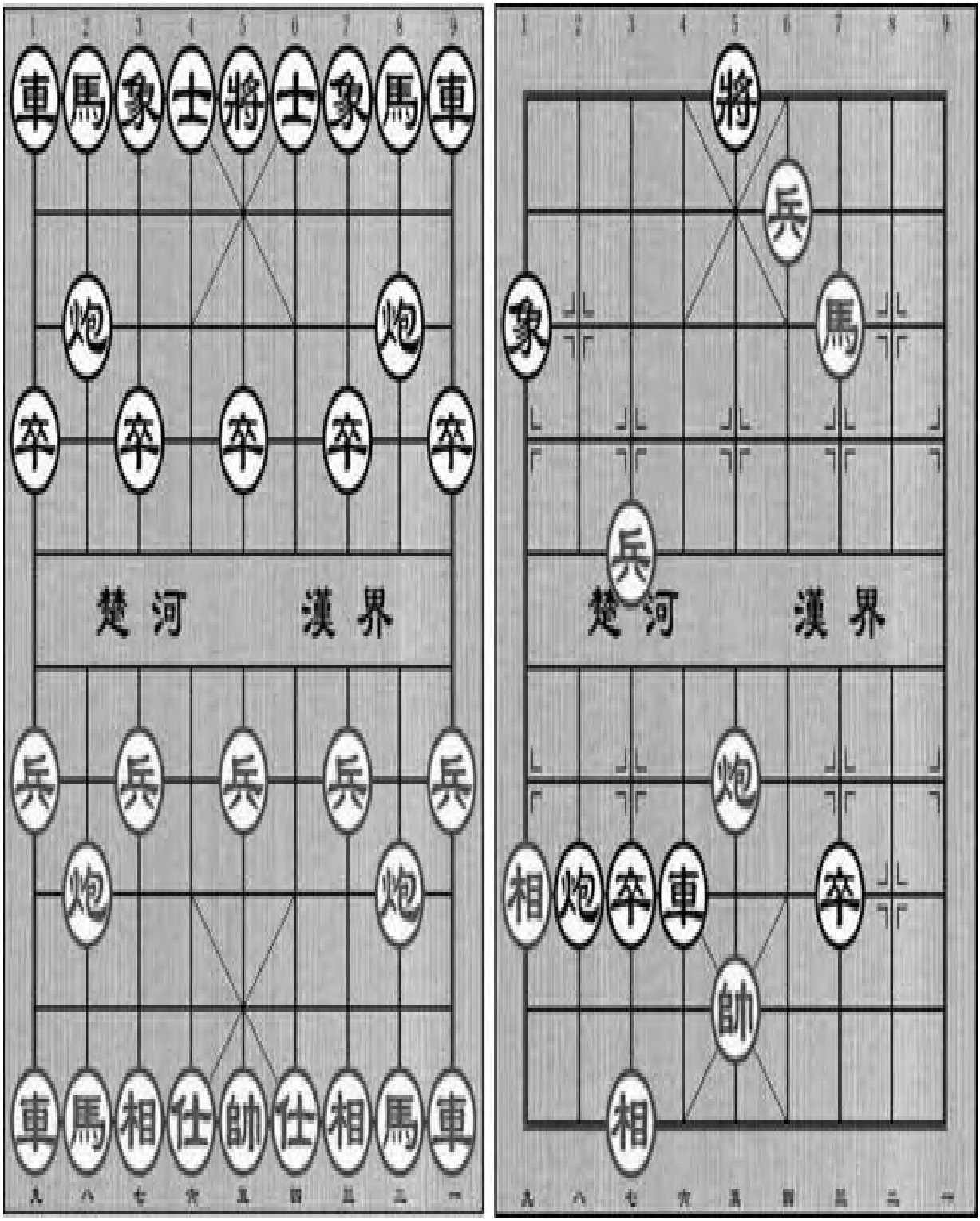


图2–3 象棋初始格局（公设）与棋子按规则走动之后的格局（定理）

一种棋子在棋盘上的摆放方式未必能够真的走得出来，所以不是每一个陈述都能被证明为定理。这样我们也可以理解，为什么大多数情况下证明一个定理比否定一个陈述要难。因为证明一个定理只要走出来就好，但否定一个陈述要把所有的走法都走过一次，除非错误特别明显，比如将“象”放在了棋盘的米字格内。这相当于后面我们会讲到的*NP*（Non-deterministic Polynomial，即多项式复杂程度的非确定性）问题，对于*NP*问题找到证明是困难的，但验证起来并不难。

公设系统应该是完备的，应该能够对系统下的每个命题加以证实或证伪。也就是说，不应该容许不可判定命题的出现。然而，我们很快就会看到，事情远远不是看起来的那么美好。譬如说：

“本命题是假的”，是既不真也不假，“没有真假值”的句子。

“本命题不可被证明”，是一个有“真值”但无法证明其真实性的句子。

---

[1] 《量子大唠嗑》，马兆远著，中信出版集团2016年10月出版。——编者注

第三章

理性的终极理想

人类希望通过逻辑工具来建立冷静客观的知识体系，这个知识体系不应该有任何主观的知识和情绪夹杂在其中，但问题是一旦有了这样的体系，并且这样的体系可以自动去包罗万象而形成所谓的人工智能的时候，人本身的作用就不大了。

直觉？

科学的历史，是一个与人类直觉做斗争而相爱相杀的历史。

很多宗教信仰，都以人类为中心，即使不这么声称，但归根到底都与人有特殊关系。《圣经》记述我们人因为上帝的眷顾而成了被选择的特殊一群。上帝为他自己的孩子——人类，创造了整个世界。类似的以人为中心的信仰有几千年甚至一万四千年的历史。随着地心说的势衰，牛顿力学的建立，我们越来越被移出了宇宙的中心。当我们有能力看到浩瀚的太空时，我们越来越不相信自己就是宇宙中最重要的部分了。这种不重要也让人们意识到，不仅“我们”不重要了，“我”也不重要了。我们对于自然的认知，不应该取决于某一个人的观点，或者某一个人的某些具体特征。即使有时候不得不采用人为的标准，也会试图让所有人都参与，通过人数来平均个体的差异，谁也不搞特殊化。据说德国人16世纪定义“尺”的长度，就用某一个星期日礼拜之后从教堂里走出来的前16个人的平均脚长。但法国人在定义“米”作为长度单位的时候，采取了更客观的策略：以地球为标准，谁说了也不算，本初子午线的4000万分之一为一米。因此，地球经线的长度是精确的4万千米。理性让科学认知离人类的直觉越来越远，这似乎是对的。因为直觉，某人的直觉，某些人的直觉，似乎并不可靠。

直觉确实有谜一般的性质。

它到底来源于什么？我想本书以及以后很多年的书都无法说清楚。在人类追求终极理性这么多年的情况下它还依然坚强地活着、活

跃着，这在一定意义上说明事情远远比我们想象的要复杂得多。

直觉也许是在过去的经验和知识积累上形成的。直觉也许是自动驾驶设备的大脑，执行处理信息的动作，而无须人们有意识地感觉到它正在运行。直觉常常被认为是无意识地思考。因此我们会问，多大程度上我们能相信直觉？我们其实可以在日常生活中体会到直觉的自动信息处理。比如说，在开车的司机身上有时会发生“公路催眠”。驾驶员在没有意识到自己每一个动作到底怎样操作，应该应用哪些肌肉和手柄的时候，也可以开几十甚至几百公里。行人走在街上，陷入沉思，或者就是玩手机，一抬头发现自己已经在目的地，却没有意识到自己走到那里的过程。直觉有重要的价值，但它并不一定能帮助我们做出一个好的决定。

直觉过程不仅发生在日常活动中，而且也运作在复杂的决策中。有人说女性依靠直觉做判断的能力比男性更强，但事实上，我们几乎所有的决策都是依靠直觉的，或者，在本书中我们会看到我们所有的决策都要依靠直觉的参与。人们通常会在事后把他们的行为解释成符合某种理性的标准，不会透露自己的主观偏好。然而真实的情况是我们很难搜集所有的判断依据，真正做到完全理性，甚至是本书将要说明的，当我们做出一个决定时，完美的理性也不能帮我们做出更好的决定，因为它们并不存在。当然，花更多的时间搜集任何有关任务或情况的信息会多多少少有所帮助。但是，人们也不应该在自己没有掌握所有信息而做出决定时感到害怕，掌握所有信息的情况可能并不存在。

直觉也许来自我们多年的学习和与人沟通的积累，让不太容易理解的事情变成一种默认的知识。如果我们抛弃经典理性带给我们的判断，像孩子一样学习和看待新鲜问题，那我们在很多时候可以毫无困难地将理论中出现的佯谬当作直观明显的东西接受下来。几代人之前人们认为完全不可直接感受的东西，今天的我们却会毫不犹豫地把它当作常识。当我说地球是圆的时，有多少人可以举出实际的证据来证

明地球是圆的呢？我们很多时候因为习惯了被各种来源的信息不断灌输，对直觉的感受并不怀疑。比如说这地球，我们已经默认它是“球”了！从月亮上地球的影子来推测，地球肯定是个球形。如果它是方的，或者扁的，那个影子就不会总是圆形的，而这一点，古希腊人早就意识到了。

由直觉而产生的公设应该是不言自明的，正如美国《独立宣言》所讲，“有些事情是自证其正确的”（We hold these truths to be selfevident）。

严格意义上来讲，所有不言自明、自我正义化的东西都值得怀疑，更别说以此为基础建立起来的层层叠叠的法律条文了。

直觉并不是安全的思维基础：它不适用于作为一种标准来判断一项科学探索有无真理性或能否获得成果，“直觉”会带来知识的不确定性。现代科学技术越来越倾向于尽量排除人的因素的干扰。物理学工作者就很讨厌“人”的直觉。我们知道的普适的物理规律难道要依靠直觉来发现吗？我们默认的公设来自直觉，这些公设应该是不言自明的！但我们该相信和依靠谁的直觉呢？都不用说谁的直觉，物理学家甚至不喜欢具体的东西。

以长度的测量为例。英文里的长度单位是英尺[1]（foot），foot的本意就是脚。这跟我们小时候的经验非常相像，游戏中的孩子也常用脚来丈量长度。13世纪初期，欧洲的度量标准非常混乱，给欧洲迅速形成的国际市场带来了诸多麻烦。在英国，仅长度标准问题带来的民事纠纷就使行政部门大为苦恼。大臣们先后召开了十几次会议商讨这个事，商量来讨论去始终确定不下来一个统一的标准。曾在大宪章上签字的国王——“失地王”约翰，在多次听证后失去耐心，于是在地上踩了一脚，指着凹陷下去的脚印对大臣们宣布：“There is a foot, let it bethe measure from this day forward.”（这就是一英尺，这个脚印就永远作为长度的标准。）现在的大英博物馆中还收藏着用膨胀系数很小的铂金制成的长方形框，中间部分即为英王脚

（The Imperial Foot）的标准长度。因为约翰王穿着鞋，所以，一英尺大约是30.48厘米。这跟唐朝我们用的“尺”的长度非常接近。法国大革命后的法国人不喜欢依赖王权，他们觉得应该找更客观的标准，因此用地球的自然长度来定义长度标准，将经线的4000万分之一定义为一米。

两千多年前秦始皇统一度量衡，今天的物理学工作者也做着类似的工作，当讨论物理问题的时候，我们需要确定基本的物理学单位。我们需要知道一米到底是多长的距离，一秒到底是多少时间，一公斤到底是多少质量。到今天巴黎的国际计量局都还保存着国际千克原器、国际米原器。如今国际米原器仅作为“最能表明米的规定性质”的器具而被保存，1889年的第一届国际计量大会确定“米原器”为国际长度基准，大会规定一米就是米原器在0摄氏度时两端的两条刻线间的距离。经过精密的工艺和精益求精的维护措施，米原器的精度可以达到0.1微米，也就是千万分之一米，可以说够精确的了。可是在1960年召开的第11届国际计量大会上，各国代表一致通过决议，废除了米原器对“米”的定义，理由是它既不方便，也不准确。第11届国际计量大会在废除旧的“米”的标准的同时，也规定了新的“米”的标准，它等于氪86同位素灯在规定条件下发出的橙黄色光在真空中的波长的650763.73倍。1983年10月，联合国度量组织在巴黎举行会议，规定了新的“米”的定义，即把光在真空中299792458分之一秒所走的距离定为一个标准米。因此，光速被定义为每秒299792458米。这样至少长度不依靠具体的某个工具了，但却依赖于时间了。那么一秒到底是多长时间呢？

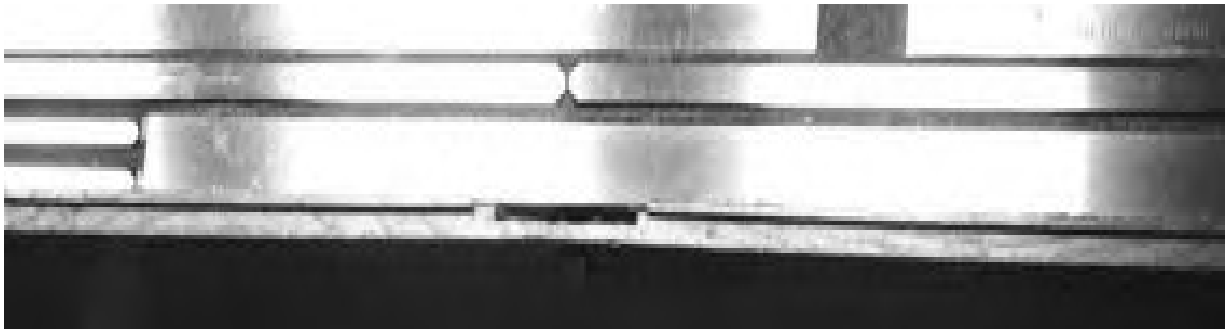




图3–1 国际计量局保存的国际米原器

这个问题并不复杂啊，一天有24小时，每小时60分钟，每分钟60秒，那么一天时间的86400分之一就是1秒了。但在这里我们是不是默认了一天的时间长度？一天是地球自转一圈所用的时间吗？这个常识是错误的，地球自转一圈的时间只有23小时56分多一点。一个太阳日需要太阳回到天空中的同样高度才能算一天。嗯，不对，是需要太阳回到正午时候的高度，由于地球公转，每一天太阳在同一时候的高度并不同。但“时候”还没有定义，所以我们谈的“正午”是有些问题的。太阳回到这一天的最高角度的这个时刻和前一天的同样最高角度的时刻之间的时间差是24小时。但因为地球在自转的时候也公转，所以自转的时候要多转不到4分钟才能让太阳处于这一天的最高点，这时候才是一个太阳日。

看到了吗？当努力把一个常用的概念说清楚的时候，我们不经意间就用了很多默认的知识，这些默认的知识是怎样形成的？可以靠大量的学习和记忆实现吗？通过大数据或机器学习的方法能够获得这些默认的知识吗？这是人工智能要解决的一个核心问题，我们的常识需要多少可描述的知识来形成？这些知识能最终积累成直觉吗？常识是直觉的基础吗？

但写到这里，我难以忍住我对时间的继续显摆。

时间

古代的人类和所有其他动物一样，不需要比“日”更准确的时间。埃及人对尼罗河的观察代表了在很长的一段历史时期内人对时间的需求。在一年的尺度上，尼罗河在确定的一天涨水淹没农田。人们只要在这之前做好准备，等待洪水退去开始新一年的种植。“日”意味着耕种和收获，“小时”有什么实际用处呢？对老百姓而言，日出

而作，日落而息，他们并没有对时间精确到小时以下的实际的需求。当然，非常小众的用在天文、立法和占卜方面的计时器，其作为礼器和宫廷的玩物也不是完全没必要存在的。臣子上朝还是要准点的，不能让皇帝等，所以至少还是有晨钟暮鼓，早晚打更的需求。但这样的计时，一个城市里有一小撮人来做就好了，更不必精确到分秒。

对分秒的需求一直到欧洲人要进行大航海探险才显得迫切。对航海而言，倒霉的事情是，地球是圆的。这当真是个麻烦事。茫茫大海，目的地或者提供淡水和补给的海岛或大陆的海岸就藏在地平线下面，要走到地平线那边才看得到。那么，地平线有多远？

欲穷千里目，更上一层楼。

鹳雀楼，20世纪90年代重建，重建高度73.9米。站在这个楼顶，到底能看多远？答案很让人失望，不过30千米。我们可以想象一下，一艘船航行在茫茫大海上，距离这条船30千米外有一个岛屿，而即使船的桅杆有70米高，船员也不可能看到这个岛。一旦四周没有参照物，只有一望无际的大海和船长手里的地图与指南针，那么船该朝哪个方向走，才能尽快到达岛屿获得给养？现实很残酷，那时候没办法，是真没办法。

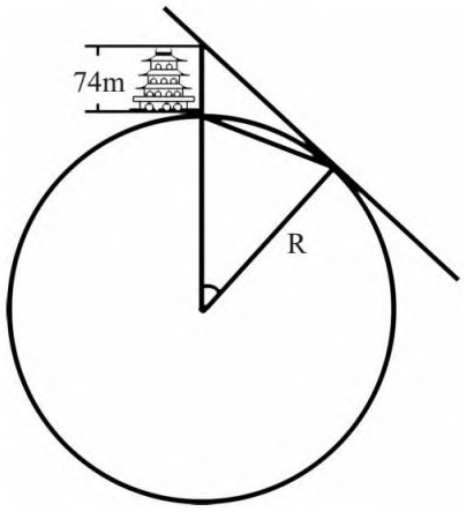


图3-2

要知道那时候海船一天也就走几十千米，遇到风暴等天气时船会偏离航道，船员一觉醒来就完全不知道自己在哪儿了。船员已经知道利用经纬度确定在地球上的位置，正午时候的太阳高度或者晚上的北极星高度可以告诉水手船所在的纬度，但他们是完全没办法确定经度的。所以那个时候的航海，都是一维的，只敢顺着东西向的纬线走，也只能顺着纬线走。这也就是为什么茫茫大海连个海岛都不容易看到，却会有海盗这门生意。商船也不敢偏离纬度线啊！海盗船只需要找一条沿途岛比较多的纬度线（这几乎是所有商船和海盗船都有的知识）趴着等就好。如果大家对迷航没有亲身感觉，那有没有在地下停车场找不到车的时候呢？因为每一层看起来完全一样！海上航行比这种情况更让人绝望，因为海面从各个方向看完全一样！所以海航一定会依赖有经验、直觉好的船长，而这位船长常年要在正午的时候通过六分仪测量太阳的高度来判断船所在位置的纬度才能保证船不偏航。经年累月下来，常用的那只眼睛自然就瞎了。独眼的船长是人类航海史上悲情的真历史！由于长时间没有蔬菜和淡水，在海上丢的不仅仅是一只眼睛，还有船员的生命。迷航在那时是非常频繁的事情，在地球上不知道经度的远洋探险，真的是冒险。

那么地球上某一位置的经度怎么确定呢？

一个方法是通过时间。

太阳回到每天的最高点的时间间隔是24小时，这样就可以按照时间间隔把地球分成24份，每一份代表一小时，这就是地球上不同经度区时差的来源。地球在24小时内旋转360度，这意味着它每小时旋转15度。如果知道本地时间和经度为零度的地方的时间差（零度经度穿过英国的格林尼治天文台），你就能够计算出本地的经度。这样通过时差，我们就能确定离开某一经度有多远，也能算出来本地的经度了。这说起来似乎很容易，但怎么确定时差呢？假设按照赤道来算，其长度对应4万千米，这是24小时，即86400秒的时间差，在鹳雀楼的高度上来看，30千米对应于大约60秒，约1秒500米。也就是说，如果时钟

相差一分钟，本来可以看到的岛就消失在地平线下面了。但从惠更斯发明了摆钟之后的很长一段时间，摆钟每天走差几分钟是很正常的事情。在海上颠簸，每一次的钟摆摆过去和摆回来的时间间隔都会不一样。而且为了减少齿轮之间的摩擦，要经常给它涂润滑剂。但在不同温度下润滑剂的黏性相差很大，这也让钟表走时不准。要做一个走几天，甚至一两个月都不差一分钟的钟表，在大航海时代是不可想象的事情，包括牛顿在内的很多人都认为不可能做出满足航海所需要精度的钟表。因此以牛顿为首的科学家们花了很长时间研究天文，希望通过月球每一天在星空上不同位置的微小差异来确定某个地方的经度。这，当然是需要极好的天文学知识的——是极好极好的。英国为此专门成立的经度局在1714年宣布：开发出实用的在海上经度计算方法的人将获得两万英镑的奖励。在那时，一艘军舰也不过几百英镑。

约翰·哈里森（John Harrison），是一个木匠和钟表匠。在他于1776年去世之前的40年里，他开发了一系列越来越精确的时钟来确定海上航行的经度，这些钟现在都保留在格林尼治天文台。1759年，第4口钟H4在远征牙买加的81天惊涛骇浪的航海中只有5秒的误差。哈里森做出来的航海钟的准确度完全超出了经度局的想象，但它的价格也不菲。而经度局真正想要的是一个“实用的解决方案”，这意味着这种航海钟必须有被批量化生产的可能。对于这个充满了手工业工人的匠心独具（奇技淫巧）的装置来说，这是很难在短期内做到的，因此也看不到大规模商用化的前景。经度局只同意把奖金中的一部分授予哈里森，总计一万英镑。直到后来在国王乔治三世的坚持下，哈里森才领到全额奖金。坊间为了增加戏剧性，常见的故事版本是哈里森解决了经度问题，但被官方科学机构轻视，比如由牛顿的学生们所构成的主流学术界。但这个版本忽视了英国和法国其他钟表制造商的贡献，他们也在开发可靠的航海计时器方面取得了重要进展。事实上没有证据能够表明哈里森当时在现实层面解决了经度问题，之后很长一段时期，哈里森的航海钟的成本都超过了当时一艘战舰的成本。但他的工作至少表明，用时间来解决经度问题是可能的。而今天，我们可以带

着走的钟表都是从哈里森航海钟原型演化来的。经过两百多年的努力，到2018年美国的国家标准局所制造的光钟，可以通过光波频率的比对把测量时间的误差降低到197亿年差1秒。

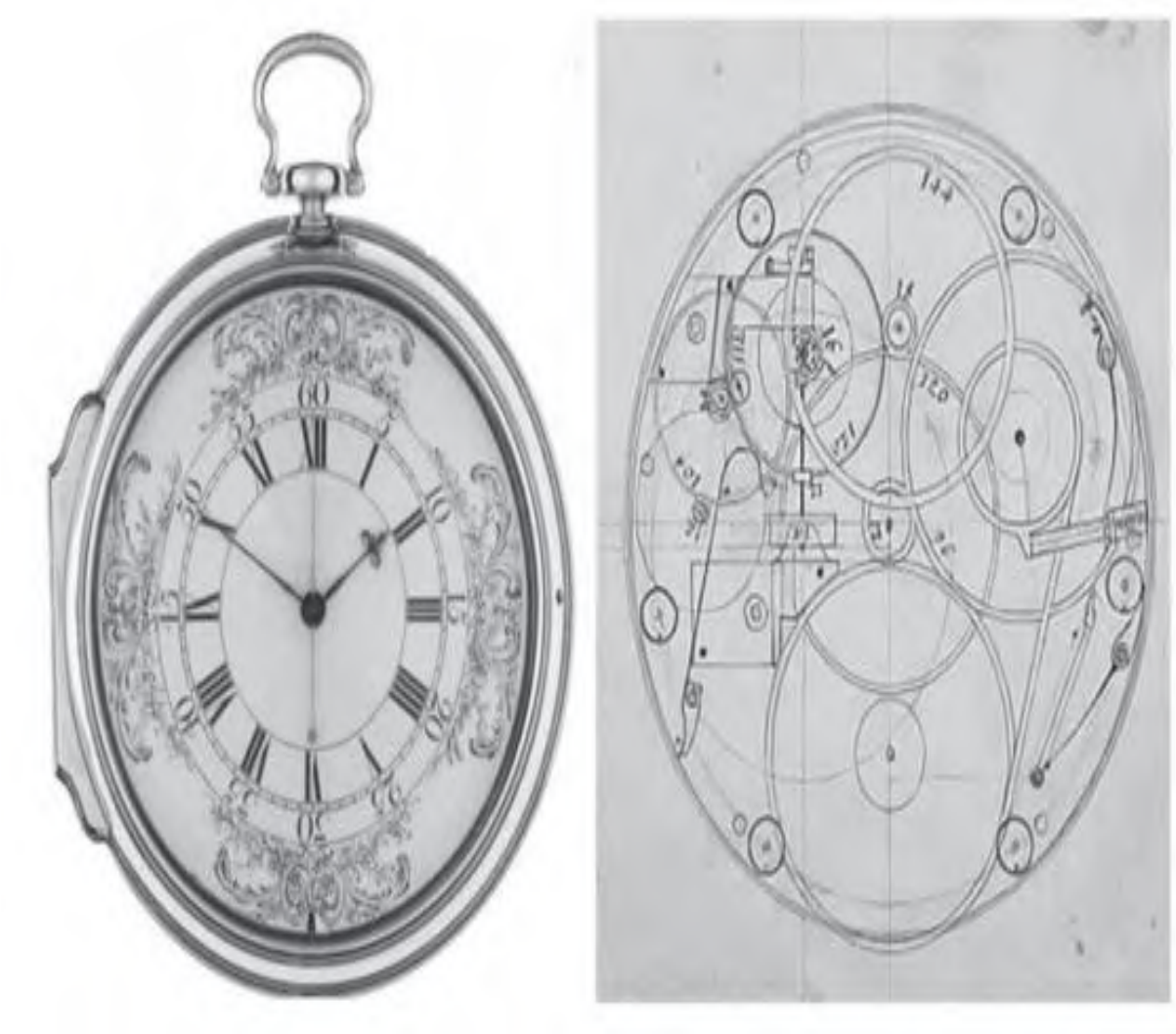


图3-3 哈里森研制的H4航海钟

在排除直觉的努力上，数学当然不甘示弱，而且认为这是天经地义的。

在上小学之前，我们就开始学算术，海淀区尤甚。小学一二年级就学会了加减乘除，但我们似乎没有问过，1后面为什么是2啊？

回答这个问题其实不那么简单。只是我们一开始就是这样被灌输的，也就成了默认的知识。但数学家不这么看，为了每一步结论的可靠，我们习以为常的常识需要被重新梳理一下。

算术系统

算术的基本公设被称为皮亚诺（Peano）公设，由五条基本假设组成：

公设*1* 1是自然数。

公设*2* 对于任意自然数*n*，其后继数*n'*都是自然数。

公设*3* 对于任意自然数*n*，*n'*≠1都成立。

公设*4* 对于任意自然数*m*，*n*，若*m'*=*n'*，则*m*=*n*。

公设*5* 假设对自然数*n*的谓词*P*（*n*）而言，下面的（a）和（b）都成立：

（a）*P*（1）成立；

（b）对于任意自然数*k*，*P*（*k*）成立，则*P*（*k'*）成立；

则，对于任意自然数*n*，*P*（*n*）都成立。

“谓词”有点拗口，它是说明主语具有某种性质的词，跟我们通常理解的“谓语”没有太大区别，或者就把它当作“陈述”或“命题”来看。这时公设*5*就成了：假设对自然数n的陈述*P*（*n*）而言，下面的（a）和（b）都成立。我们来看怎样用这样的公设来定义自然数。需要再次跟读者重申：公设，指的是不需要证明也能成立的陈

述。陈述则是可以判断真假的数学性观点。当陈述被证明为真的时候，我们称之为定理。

我们自然地把自然数当作某种符号的集合，这些符号具有某种特性，我们把这种集合标记为*N*。那么公设*1*告诉我们“1”是集合*N*中的一个元素，即1属于集合*N*，即图3－4所示：

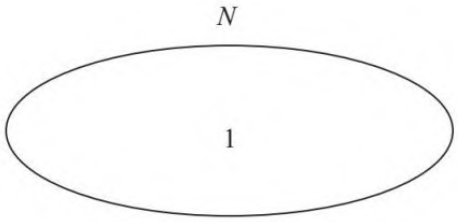


图3–4

这似乎没有什么特别，我们看公设*2*，对于任意自然数*n*，其后继数*n*’都是自然数。我们现在还不能认为*n*’=*n*+1。因为直到这条公设，我们还没有任何举动来定义“+”这个字符。事实上，正如数学家哥德尔所说的，如果到此为止的话，仅有这两条，这个公设系统还不能表达足够丰富的内容。那么，是不是说“1”是自然数，“2”就是自然数了呢？也还不能这么说。因为，根据公设*1*和公设*2*，我们没有定义字符“2”。我们姑且把*n*’作为使用的字符，1’作为1的后继数。那么根据公设*2*来说1’也是一个自然数。注意这并不是矫情或者不必要的烦琐，我们只是太熟悉用工具来制造新东西，而常常忘记了工具本身也是被制造出来的，我们一样要考察工具的可靠性。

这时，我们有了图3－5：

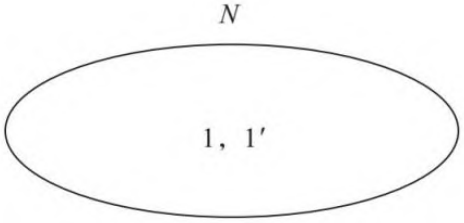


图3-5

根据公设*2*，我们有了图3－6：

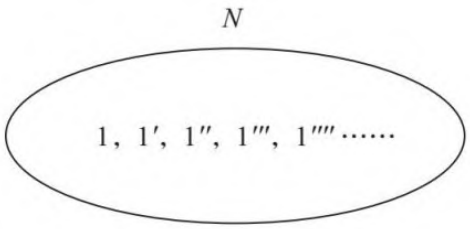


图3-6

虽然公设*1*和公设*2*不算太复杂，但我们已经可以用它来定义几乎无穷多符号和它们的一些基本性质。更重要的是，我们已经在使用演绎法，通过逻辑推理由公设得出结论。并且由于我们不断地使用同一个公设，事实上，我们已经开始尝到了自然递归的味道，在数学证明中这是个常用的方法。

公设*3* 对于任意自然数*n*，*n*'≠1都成立。

这似乎是废话，我们确实可以得出的结论是1’≠1，但这有什么用吗？另外一个结论似乎是，1是最小的自然数。但问题是，我们还没有定义“大小”，因此也不知道怎么比较大小。所以也不存在“1是最小的自然数”这个结论。我们只能说1不是任何数的后继。

我们先看一下公设*4*，然后再返回来看看为什么要有公设*3*。

公设*4* 对于任意自然数m，n，若m'=*n*'，则m=n。

这似乎又是理所当然的。当*m*的后继和*n*的后继相等的时候，*m*等于*n*。这不是相当于*m*+1=*n*+1，得出来*m*=*n*吗？事实上，我们还没有定义

“+”，但公设*4*确实要给我们定义一些使用的算法了，比如这个所谓的后继符号“ ’ ”，到底是个怎样的运算，即“后继”应该具有哪些性质。

如果我们假设想要的自然数序列是：

$$1 \to 1' \to 1'' \to 1''' \to 1'''' \cdots\cdots$$

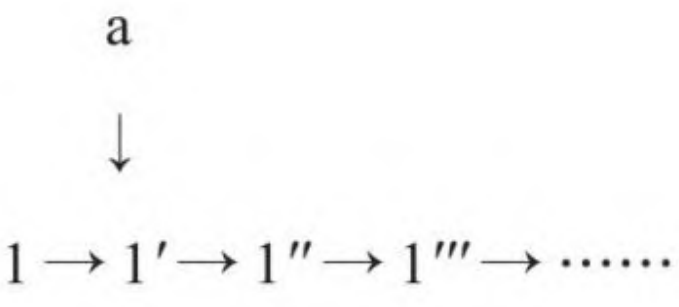
如果没有公设*4*的要求，我们可以定义，1’’’’=1’，即使1不等于1’’’但也可以有1’’’’=1’，这样会造成：

$$\begin{array}{ccccc} & & 1'''' & \leftarrow & 1''' \\ & & \downarrow & & \uparrow \\ 1 & \to & 1' & \to & 1'' \end{array}$$

这样的结构，完全满足公设*1*、公设*2*和公设*3*，但它不是我们想要的。

所以皮亚诺想要定义的是：1是自然数，所以才有了公设*1*，任意自然数都有后继数，所以有了公设*2*，然后，没有后继数为1的自然数，所以有了公设*3*，后继数一个一个延伸下去，这个序列是单向的，不许回头，不可以打结，所以有了公设*4*。

有了公设*1*、公设*2* 和公设*3*，但不能避免形成这样的序列：



公设*1*，公设*2*和公设*3*都没有限制这个序列里的所有元素是从1过来的。这个序列走着走着有可能有别的符号加入，所以公设*4*是用来防止其他来源的符号与已定义的自然数列会合。但这样是不是有公设*4*就不需要公设*3*了呢？如果只有公设*1*、公设*2*和公设*4*，自然数还可能有这种结构：

$$\cdots\cdots a \to 1 \to 1' \to 1'' \to 1''' \to 1'''' \cdots\cdots$$

这个数列从莫名其妙的地方来，最终跑到莫名其妙的无穷远去。所以我们必须规定，1是唯一排在最前面的起点。

公设*5* 假设对自然数*n*的谓词*P*（*n*）而言，下面的（a）和（b）都成立：

（a）*P*（1）成立；

（b）对于任意自然数*k*，*P*（k）成立，则*P*（*k*'）成立；则，对于任意自然数*n*，*P*（*n*）都成立。

公设*5*叙述了如何证明“对于任意自然数*n*，*P*（*n*）都成立”，即我们说的数学归纳法。自然数定义中出现了数学归纳法，暗示着归纳法跟自然数的天生联系。如果自然数是有限的，只有1，1’，1’’这三个，我们只要证明*P*（1）、*P*（1’）和*P*（1’’）成立，就可以证明对于

所有自然数n，P（n）都成立。然而，自然数有无穷多个，我们不可能实际调查无数个自然数。如果我们希望对自然数整体有一个评判，就需要用数学归纳法。公设*5*定义了一种机制，这种机制使局部观察到的自然数特性可以扩展到整个自然数序列里。

由这里出发，数学归纳法有两个步骤：

（a）出发点：证明命题P（1）成立；

（b）证明对于任意自然数k，P（k）成立，P（k'）也成立。

如果能够证明（a）和（b），那么就能证明对所有自然数n，P（n）都成立。

现在我们可以有足够的工具做基础来定义加法“+”了。

公设ADD*1* 对于任意自然数n，n+*1*=n’都成立；

公设ADD*2* 对于任意自然数m，n，m+n’=（m+n）’都成立。

比如：

$$1+1'=(1+1)' \qquad \text{公设 } ADD2 \text{ 中 } m=1,\ n=1$$

$$=(1')' \qquad \text{公设 } ADD1 \text{ 中 } n=1$$

$$=1''' \qquad \text{去括号}$$

终于，为了方便，我们定义符号1’=2，1’’=3……并且我们刚刚利用加法公设ADD*1*和公设ADD*2*证明了定理1+2=3，即陈述（谓词）：

1+2=3为真。

噢耶！

集合定义自然数

公元前六七世纪几何学在古希腊兴起，德谟克利特（Democritus）、毕达哥拉斯（Pythagoras）、阿基米德（Archimedes）等人是其中的代表人物。欧几里得（Euclid）全面总结以前的研究成果，写成了《几何原本》一书，这本书全面总结了古希腊人对世界的理解和古希腊人的思维模式，对后世产生了不可估量的影响。这个思维模式的形成和传承比接下来两千年里的所有科学成就的总和还要大，不过人们意识到这一点却要等到文艺复兴之后了。

《几何原本》之所以重要，是因为它构造了一个完整的形式逻辑体系，成为接下来两千多年来人类对理性追求的典范。古希腊的先哲一直对逻辑有深入地探讨，主要集中在形式逻辑的演绎层面及形式逻辑和真实世界到底有何关联。欧几里得以天才的构思，总结了古希腊人的思维框架：

合理的假设—缜密的演绎—得出结论

基于这样的思维框架，公元前300年，欧几里得写了《几何原本》，作为合理的假设，他确定了五条几何的基本公设：

1. 任意两个点可以通过一条直线连接。

2. 任意线段能无限延伸成一条直线。

3. 给定任意线段，可以以其一个端点作为圆心，该线段作为半径画一个圆。

4. 所有直角都全等。

5. 若两条直线都与第三条直线相交，并且在同一边的内角之和小于两个直角，则这两条直线在这一边必定相交。

欧几里得几何是完美的理性学科的样本，把所有的学科都塑造成欧几里得几何的样子成了人类孜孜不倦地为之奋斗了几千年的梦想。然而，问题似乎还差一点点，对较真儿的人来说还有一点点不和谐：第5条公设似乎有点多余。第5条公设被称为平行公设，它似乎啰里啰唆的有点长。从它可以导出下述命题：平行线在无穷远处不相交。这条公设能不能从其他四条公设里推导出来呢？两千多年来很多人怀疑这个公设的必要性，也试图用另外四条公设来证明它。甚至19世纪的数学家试图利用反证法来证明这条公设是可以从其他四条公设里推导出来的，比如假设平行线在无穷远处相交，然后证明这个假设不符合逻辑。但很快人们就发现，只由前面四条公设一起，给予不同的第5条公设，可以完全推导出独立于欧几里得几何的新几何学，而自然世界真的有客观事实跟这个新几何学对应。比如地球上某一区域的经线就是互相平行的线，但在南北极是相交的！同样，如果假设两条平行线越分越远，那么空间中一样可以有一个马鞍形的面，使其在无穷远处越分越远。在这些新的公设基础上，黎曼几何发展起来，并且成为爱因斯坦广义相对论中的空间弯曲的数学基础。

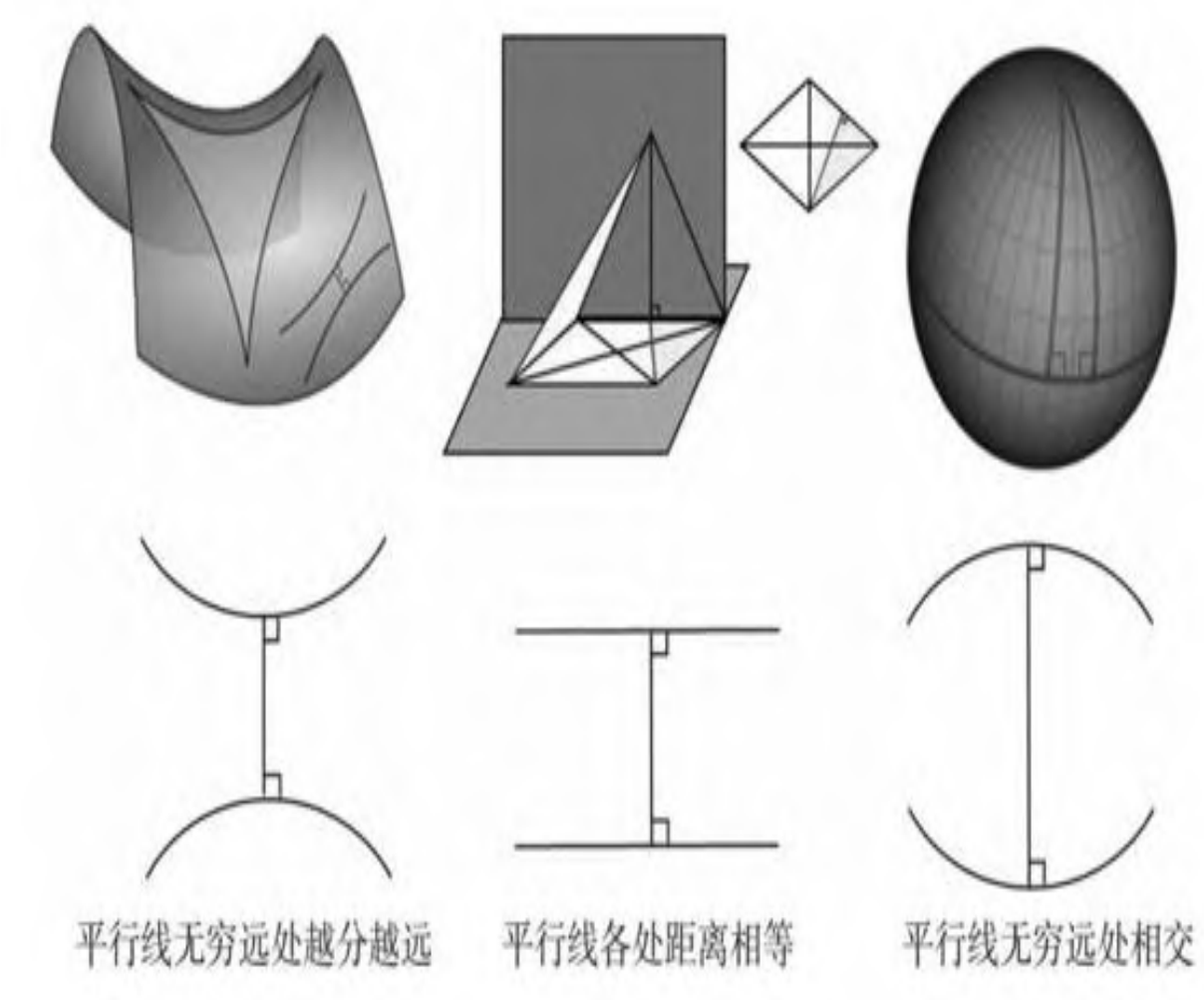


图3–7 不同平行线公设下的几何

几何学公理化发展对后人产生了深远的影响：这样少的几条公设，支撑了无穷无尽的命题。而且，逻辑推演能够以确定的方式保证这些命题为真。实际上，两千多年来绝大多数学者都毫无疑问地相信欧几里得几何是关于空间的真理，即所有定理的真理性和相互间的自洽性都自动得到保证。由于这些原因，世世代代杰出的思想家，均将公设演绎而得出的几何学视为科学知识最好的典范。通过直觉，人们普遍认为数学的每一分支都能找到一组公设，从这些公设出发就足以系统地推导出在此研究领域中无穷无尽的所有真命题。但当修改其中一条公设，一样可以构建没有矛盾的但完全不同的几何体系，对完美

主义者而言这绝对不是个好消息，到底哪一套才是对的？这个典范居然会依靠假设，而这假设从哪里来？它是否是可以立足的，或者这些最初的假设其实可以证明自己是正确的？

同构原理

19世纪，人们发现通过对定理的映射，可以把关于某一数学学科的理论，映射到一套基本理论上，并将之称为同构原理。比如我们可以把几何用代数来全面取代，利用笛卡儿（Rene Descartes）的坐标几何把欧几里得几何公设都转化为代数定理。

在这种转化中：

点：被视为一对数（x，y）；

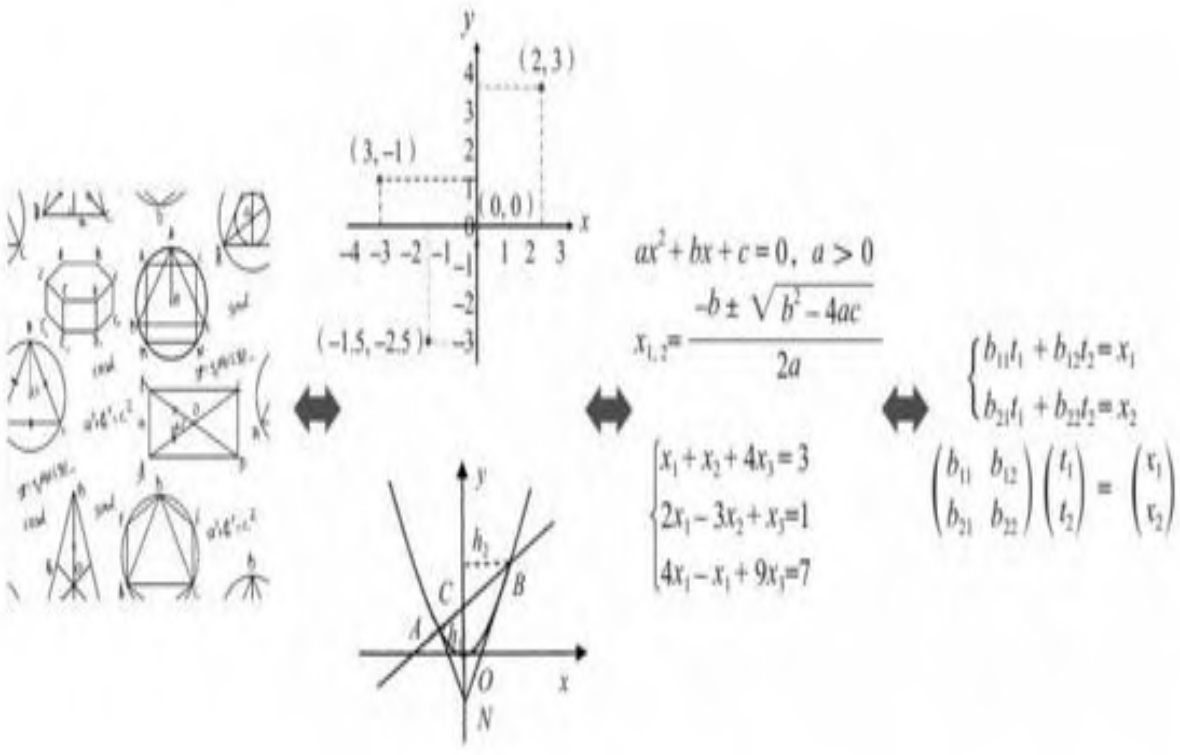


图3-8 同构原理使得某一数学学科的理论映射到另一套基本理论上

直线：被视为二元一次方程$y=ax+b$；

圆：被视为某种形式的二元二次方程$R^2=x^2+y^2$。



图3–9 同构原理

数学系统里的具体理论分支可以说是非常复杂和抽象的，但同构原理可以把复杂的数学系统简约化，从而变成在另外一套系统中可能很简单的问题。数学讨论的大多数问题是关于“数”和“形”的。“形”是几何学讨论的范围，但是几何可以坐标化。例如：在平面几何中，如果我们把以（x，y）为坐标的点直接看成实数对（x，y），那么我们便可以省掉“形”的观念，直接以实数来讨论它了。至于实数，理查德·狄德金（Richard Dedekind）和格奥尔格·康托（Georg Cantor）在19世纪中叶各引入一套实数的理论，以有理数为素材把实数建构出来，有理数 $\frac{m}{n}$ 可以想成整数对（m，n），其中$n\neq0$；若 $\frac{m}{n}$（m'，n'）为另一个这种整数对，则当$mn'=m'n$时，我们把 $\frac{m}{n}$ 和 $\frac{m'}{n'}$ 看成等同，这样我们便可以用整数为素材把有理数建构出来了。至于整数，也可以用类似的方法从正整数，即自然数中建构出来。这样我们可以把整个数学建立在自然数基础之上。

更复杂些的是，通过对操作符号的规定，我们也可以把代数映射到自然数定理系统里，用自然数来代替代数。这样的策略在今天看来是很自然的事情，如果我们学习计算机原理，看到用01和位操作就能在计算机上执行很多事情，对这样的方法就一点也不惊奇了，而这正是图灵机所采用的策略。在自然数上如果证明了自然数系统本身的完美（我们还是暂且用“完美”这个词），它包括完备和自洽，那么别的数学学科系统也会自然地完美了。

很快，自然数所代表的初等算术系统，也可以用另外一套形式系统来充分映射和表达：人们发展了集合论（Set theory），用集合可以完全定义初等数学，由此我们只需要关心集合论是不是完美的。

然而这样做一样会有问题，利用同构原理我们只是把问题从一个领域转到了另外一个领域。人们总希望见到数学体系是自洽并完备的，而我们知道自洽和完备意味着理论系统的有效性和有用性。几何自洽性转移到代数自洽性的论证，无非是表明如果代数是自洽的，那么几何体系也是自洽的。但怎样证明代数的自洽性呢？这种证明只是说明了一个体系的自洽性跟另一个体系的自洽性有关，问题由一个领域移到了另外一个领域，因而不是一个对自洽性的“绝对”证明。

在19世纪中叶，康托创立了集合论。根据集合理论，当我们把一些清晰可分的在客观世界中或我们思想中的事物看成一类时，这整体便称为集合，其中的事物称为它的元素。若x是集合A的元素，我们便用$x\in A$表示；若x不是集合A的元素，则以$x\notin A$表示；不包含任何元素的集合叫空集，通常用符号 $\emptyset$ 表示，空集和自然数中的0相当。以元素a，b，c为元素的集合，可以用$\{a，b，c\}$表示。有了这些准备以后，我们就可以把自然数的系统建立在集合论上，方法是：0=$\emptyset$，1=$\{0\}$，2=$\{1，2\}$，3=$\{0，1，2\}$……如此整个算术系统进而整个数学系统就可以建立在集合论上了。

然而，当一切都指向数学理论大统一，理性的前程一片光明与美好的时候，20世纪初，集合论出现了难以解决的悖论。其中最重要的一个是1902年罗素所提出的。罗素认为，集合也可以作为集合里的元素，有些集合可以以其本身为元素。若定义一个集合由一切不属于自身的集合所组成，那么这个集合是否包含它自己呢？用排中律衡量，一个元素或者属于某个集合，或者不属于某个集合。因此，对于一个给定集合，问此集合是否属于它自己是有确定意义的。但集合论对罗素的这个看似合理的问题的回答却陷入了两难境地。通俗来讲，它与著名的“理发师的悖论”有相通之处：某村的理发师号称给村子里所有自己不理发的人理发，问题是，理发师该不该给自己理发？

看似简单，但这样的悖论造成数学界的大地震，因为根据同构原理，整个数学系统的建构是以集合论为基础的。如果集合论里有不能

解决的问题，那么其他理论里也有一样解决不了的问题，这样数学就缺乏了我们希望它应该具有的完备性，那么整个数学理论就可能需要重建了！简单的解决方案是先重建和理顺集合论的表达，这是20世纪前半叶很多数理学家都在尝试的办法。人们认为问题出在康托给集合所下的定义不够严谨，集合论必须重新出发，把集合与整数等观念看成纸上的符号，再说明哪些符号串有意义，在有意义的符号串上允许哪些变换，等等。这样便可以把各种悖论的问题同时解决。

但同时，既然罗素提出了这样一个关于在集合论内部的不可判定的问题，人们接着想到另外一些深刻而令人胆寒的问题：

数学本身是自洽的吗？也就是说，数学是否前后一致，会不会得出某个数学陈述又对又不对的结论？数学是否没有内部矛盾？

数学本身是完备的吗？也就是说，面对那些正确的数学陈述，我们是否总能找出一个证明？数学真理是否总能被证明？

数学本身是可判定的吗？也就是说，我们能否最终找到一种方法，仅仅通过机械化的计算，就能判定某个数学陈述是对是错？数学证明能否机械化？

这些问题会引来一个更为深刻的讨论，数学本身是不是客观稳定并自然存在的知识呢？那么以数学为基础的科学呢？

人类认为科学知识具有确定性、绝对性和永恒性的观念有着悠久的历史传统。欧几里得几何学曾经代表了人类认知追求确定性的不可抗拒的信仰。从古希腊确立理性主义的认知方式以来，人类总是追求具有普遍必然性的确定性知识。理性主义知识论者笛卡儿更是将数学视为科学知识的唯一范式：“因为它立足于公设上的证明是无懈可击的，而且是任何权威所不能左右的。数学提供了获得必然结果以及有效地证明其结果的方法。”

随着数学对自然现象的强大无比的阐释，知识具有确定性的观念也被进一步强化了。皮埃尔·西蒙·拉普拉斯（Pierre Simon

Laplace）甚至说：“世界的未来是完全由它的过去决定的，而且只要掌握了这个世界在任一给定时刻的状态的数学信息，就能预测未来。”这也成为一个著名的悖论——拉普拉斯悖论。那么这一切都已经决定了的数学信息包括人类的意识吗？不管怎么说，正是由于先贤建构并确立了追求确定性知识的认识科学的思维方式并形成传统，后来的哲学家总是寻求方法来维护所获得的科学知识的普遍必然性或说确定性。直到今天，确定性依然是绝大多数人心目中“科学”之所以为科学的本质特征。这是传统知识论的主题，也是科学哲学的主题。但是哥德尔不完备定理说明，不确定性是人类认知的形式逻辑和理性思维本身固有的，即使纯粹的数学也无法彻底达到确定性。进一步说，数学概念和理论如果与人们的实际经验和科学观察相结合，会产生更大的不确定性。在任何希望做到足够严谨的知识体系中，绝对的确定性是不存在的。

---

[1] 英尺，英美制长度单位，一英尺等于12英寸，合0.3048米。——编者注

第四章

伟大的聊天

可怜的哥德尔不能自证其重要性，我们只能请已经被证明重要性的爱因斯坦来证明。但有意思的是，即使做了这么多年科学，尤其是物理和数学，我也依然有很多问题不敢问，或者即使问了也未必有人会知道。在这一章我大胆地罗列出来，但愿这两个人讨论过了吧。

虽然我上一本书的书名“量子大唠嗑”显得不那么科学、严谨和认真，但唠嗑这件事情对人类来说真是无比重要。唠嗑交流的不仅是语言，还有身心，对话可以交流一部分信息，而信息的交流未必仅限于语言。当你用身心去感受，它可以是声音、气息、光和温度，是眼神和手势。当我们不再把信息局限为某种可被记录的01序列，漫无边际的唠嗑是用语言来补充这些额外的不能尽述的信息。我们喜欢亲近自然而与自然对话，由风到气息，我们在山林中与天地交互，养浩然之气；夜深人静我们也跟自己对话，回答来自内心的问题；我们与朋友交谈，与他人一起消磨时间，在与人的交往中获得身心的愉悦。

我算是一个喜欢动脑筋的人，跟智者聊天在我看来是一种福分和快乐。如果能够穿越，我一定要穿越到20世纪40年代到50年代的普林斯顿。在那时那地，整个人类史上最智慧的两个大脑，每天在一起聊天。我必然做一个可以随时插话的隐形人，一个给捣蛋鬼和他的挚友捣蛋的穿越捣蛋鬼[1]。因为我知道聊天的力量，当两个绝世高手每天切磋武功而各自回家做功课第二天接着切磋，如此往复十几年，那要把其他人类甩到哪里？这俩人到底聊些什么？

哥德尔出生于1906年，比出生于1879年的爱因斯坦小27岁。他出生的前一年，是爱因斯坦的奇迹年。1905年，爱因斯坦发表了三篇论文，分别奠定了量子力学、热力学统计和相对论这三大现代物理学领域的基础。那一年他认为不太重要的工作——研究电子的光电效应，为他赢得了1921年的诺贝尔物理学奖。而他最得意的相对论，在被他提出很多年后人们都还表示怀疑。

就这两位绝世高手的相似点来说：他们都极其睿智；两个人都是以德语为母语，都在战前逃离了希特勒的纳粹统治；他们都在普林斯

顿工作到离世。就不同点而言：哥德尔沉默寡言、悲观、孤独，有着奇怪的饮食习惯，他小心翼翼地选择食物，就像吃婴儿食品一样，他最喜欢的电影是《白雪公主和七个小矮人》；爱因斯坦活跃开朗、更加合群，喜欢小提琴和莫扎特，无论到哪里讲学，都是万人瞩目、前呼后拥的明星。

但爱因斯坦和哥德尔似乎永远在比同时期其他人更高的层面上思考问题。按爱因斯坦的话来说，他们也因此变成了“博物馆里的老物件”。爱因斯坦不接受量子理论，哥德尔相信幽灵、重生和时间旅行，并认为抽象的数学事实上与桌椅一样真实。很多同时代的科学家认为这种看法是可笑的天真。哥德尔和爱因斯坦都坚持认为，世界独立于我们的思想，但可以通过人的思维理性认知。除此之外，两个反叛者，两个革命者，摧毁了人类认知中最引以为豪的精确知识基础——物理学和数学。

智慧带给他们共同的孤独感，绝对的智慧带来绝对的孤独，他们在彼此的陪伴中找到了心灵的慰藉。

当他们在普林斯顿高级研究所工作时，爱因斯坦和哥德尔每天都会一起走路上下班。然而，居然没有一个人在他们身边听到他们的谈话。而对于我这种会把握聊天里一两句话来体会其中深意并受启发的聪明人来说，我居然不在现场！这真是时空带来的令人无法接受的遗憾。他们中的年长者，旷古烁今地代言了人类对科学的所有想象，而他们中的另外一个，迅速地被大众遗忘。年长者，就是爱因斯坦，而另一个，是哥德尔。但，被大众遗忘未必会被历史遗忘；哥德尔对人类认知和整个人类思想史的影响，直到现在都还在发酵。

哥德尔的工作事实上挑战了，或者推翻了人类从亚里士多德开始追求的绝对理性，让我们意识到人类理性的深层次边界。而相类似的界限，巧得很，也在几乎同一时期的量子力学中被看到。虽然爱因斯坦到死都认为量子力学的不确定性原理是毫无道理的，但他还是跟哥

德尔成了莫逆之交，而哥德尔恰恰是以不完备定理成就了他在理性世界甚至是人类认知历史上的地位的。

自然界和自然界的真理埋藏在黑暗中，上帝说让牛顿去吧，于是牛顿带来了现代科学的光明。哥德尔似乎只在人类的文明史上昙花一现，耀眼地从夜空中划过。他的思想影响了冯·诺依曼和图灵，现代计算机从而被创造出来了。极具声望的奥地利经济学家、博弈论的创始人奥斯卡·摩根斯坦（Oskar Morgenstern）说：“毫无疑问，哥德尔是在世的最伟大的逻辑学家。”1951年2月，哥德尔卧病在床，美国原子弹制造计划——曼哈顿工程的总指挥尤利乌斯·奥本海默（Julius Oppenheimer）告诉主治医生：“你的病人是自亚里士多德以来最伟大的逻辑学家。”在1978年3月3日哥德尔的追悼会上，数学家安德烈·韦伊（Andre Weil）总结说，哥德尔是两千年间唯一一个能不带炫耀地说“亚里士多德和我”的人。物理学家约翰·惠勒说道：“如果你称他为亚里士多德以来最伟大的逻辑学家，你是在贬低他。”惠勒在他1974年发表的一篇文章中就断言：“即使到公元5000年，若宇宙仍然存在，知识也仍然放射光芒的话，人们将仍然把哥德尔的工作看成一切知识的中心。”哥德尔思想具有潜在的科学和哲学价值，它已经被引申到自然科学乃至人文科学的各个角落，为数学、逻辑、语言、人工智能、自然科学、思维科学和认识论的研究提供了有益的启示。

爱因斯坦从1942年起直到1955年去世前与哥德尔过从甚密。爱因斯坦自己认为哥德尔的工作对数学，与他的工作对物理学有同等的重要性，“因为我遇到了哥德尔，我知道了数学中确实存在着同样的东西”。爱因斯坦曾经对摩根斯坦说，在他晚年，他自己在普林斯顿高等研究院所从事的工作对他来说已经没有太大的意思，他每天坚持上班的主要动力是路上可以跟哥德尔聊天。在1955年爱因斯坦去世前，普林斯顿的居民会常常看到这两个人一起散步。他们对彼此的吸引力来自对各自学问和宇宙万物的深刻见解。1933年两人第一次相遇时，

年轻的哥德尔才刚刚展露他在数学界天才的一面，而爱因斯坦已经54岁，接近他辉煌而富有成效的职业生涯的顶点。

即使今天也不难想象当爱因斯坦出席研讨会时听众的敬畏之情，20世纪30年代初识爱因斯坦的时候，哥德尔也多少有这样的态度。在给母亲的信中，哥德尔兴奋地述说他沉浸在与名人交往的荣耀中。“我到目前为止已经去过爱因斯坦家三次了，”他在1946年写道，“我相信他很少邀请别人到他家。”哥德尔，虽然比爱因斯坦小27岁，但他并没有被爱因斯坦的声誉吓倒。他毫不犹豫地挑战爱因斯坦的想法。这些挑战充满了睿智和理性，这也许是爱因斯坦认为哥德尔是他的忘年交的原因。

哥德尔的不完备定理于1931年发表。就像爱因斯坦在26岁提出相对论那样，这一年，25岁的哥德尔重写了现代科学的基本规则。哥德尔证明，数学是不完备的，并将无法完善。无论理论系统基于什么规模的公设，都有超出命题系统范围的真实陈述，有些数学真理，即使是真的，我们也无法证实。即使将这些不可被已知系统证明的命题添加到该系统中作为进一步的假设也不能解决问题。扩展的系统也不完备，这样的不完备会不断地随着系统的扩展升级。如果爱因斯坦用他的相对论摧毁了我们关于物理世界的日常观念，他的忘年交哥德尔同样给我们的数学世界带来了类似的颠覆性影响。

在长达两千多年的历史里，数学家总以为：如果某个命题是正确的，一定可以用数学演绎的方法证明其为真；如果某个数学命题是错误的，也一定可以用数学演绎的方法证明其为假。正如法国数学家庞加莱（Jules H. Poincaré）所说：“在数学中，当我们拟定了作为约定的定义和公设以后，一个命题就只能为真或为假。回答这个定理是否为真，不再需要求助于我们那些不确定的直觉，只要求助于逻辑推理即可。”

哥德尔不完备定理一举粉碎了数学家们这个长达两千多年的信念。它告诉我们，“真实”与“可证”是两个不同的概念。“可证”

涉及具有可操作性的机械思维过程；而“真实”则涉及一个直觉的、主动的、无穷的思维过程。可证的一定是真实的，但真实的不一定可证的。从这个意义上说，悖论的阴影将永远伴随着我们。而很快，图灵把这一结论普遍化到了所有图灵机可解决的问题上，所有可描述算法，图灵机都可以解决，那么所有可描述算法之外的问题，都难以避免哥德尔不完备定理所限定的规则。无怪乎著名数学家外尔（Hermann Wely）发出这样的感叹：“上帝是存在的，因为数学无疑是自洽的；魔鬼也是存在的，因为我们不能证明这种自洽。”

爱因斯坦是物理学的天才，他动摇了物理学的基础，彻底改变了我们理解世界的方式。哥德尔却不为公众所了解，然而，他是数学和逻辑学的天才，他动摇了数学的基础，而数学是哪怕最应用型的科学的逻辑基础。爱因斯坦辉煌一生的最后十几年里，他和哥德尔每天在一起聊天的主题是什么？关于空间和时间的抽象旅行？关于数学基础的基础？关于古典希腊神话？关于政治？关于他们的感情生活？关于其他同事？关于美国人与德国人的不同之处？没有人会确切地知道。基本上，他们只是人类中的两位喜欢交换意见的好朋友。

尽管如此，在人类科学的伟大历史中，爱因斯坦和哥德尔都傲视群雄，所以惺惺相惜，没有什么人可以站在相同高度上跟他们讨论问题。虽然他们的谈话已经永远消失在时空中，但我们依旧可以假想他们在那些漫长的散步中讨论过的话题。

话题一：量子力学

在以牛顿力学为基础的经典物理中，我们通常可以认为无穷远的事情影响无穷小而把它忽略掉。在做一个关于引力的积分的时候，我们是在从无穷远积分到另外一端的无穷远，但事实上我们心里想的

是，无穷远的现象所产生的效果是非常非常小的，即使算进来也没什么差别。这样做给我们的数学处理带来了很多方便。这样的处理手段一样用在了热力学统计中，对于小概率事件，我们通常认为它对整体的影响是微乎其微的，所以当一个事件发生的概率极小的时候，我们通常会把它的影响也忽略掉。但量子力学似乎认为这样不对。这些微小的概率，似乎也不能轻易被忽略，甚至，它们才是说了算的。

量子力学是物理学中第一个真正涉及无穷的学问。而哥德尔思考不完备问题的落脚点，也从涉及无穷的性质的康托连续统问题（Continuum hypothesis，即有没有比自然数大但比实数小的集合）开始。爱因斯坦在1935年提出了著名的量子力学不完备的新挑战——尽管这之前爱因斯坦每一个挑战都是著名的，这个1935年的挑战在于量子力学有可能存在着违背相对论的超距作用机制。通过纠缠一对粒子，可以让这种纠缠一直跨越相对论所描述的时空限制。即使相差距离非常远，这两个粒子也可以彼此感知，并相互影响，这种影响是即时的，不需要信息传递的时间，或者说信息的传递速度可以大到无穷。这与相对论中“任何相互作用的传播速度的上限是光速”相矛盾。因此，爱因斯坦提出了以他和另外两名合作者命名的EPR悖论。

在爱因斯坦看来，量子力学的一些性质导致了这个理论存在严重的缺陷，它并没有把物理内涵中该讲的事情讲透。量子理论至少应该需要一些合理的补充，才能比较完善。在找到更完整的描述之前，量子力学是不完备的。如下的几个问题，量子力学迄今都还没有一个完善的描述。[2]

量子坍缩

对量子系统的测量可以导致量子系统的坍缩，由很多态的叠加状态，瞬间坍缩到其中一个态上，比如著名的薛定谔的猫。但是，量子力学本身缺乏合理的工具来说明坍缩过程是怎样发生的，似乎就只是瞬间。而瞬间这个概念，是经典物理学所不能接受的，怎么会有一个无穷快的、不可被观察的过程？

量子跃迁

原子的电子云会有几个轨道，电子吸收光子从而从一个低能态轨道跃迁到高能态轨道的过程是量子力学不可描述的。早期的准经典模型曾经幻想过电子可以有轨道，像行星围绕地球一样运动。然而这个模型很快被证明是不对的，电子从一个地方消失，在另外一个地方出现，是个概率性的事件，没有理论来描述它怎么从一个地方消失又怎么从另外一个地方出现，这个中间过程量子力学是不去或者说不可描述的。

量子随机

爱因斯坦最不喜欢的是量子随机性。“上帝不可能掷骰子！”爱因斯坦说，但波尔（Niels Bohr）立刻反驳说：“别去指挥上帝该怎么做！”一个骰子，我们通过精确测量初始条件，充分掌握各种数据，建立对应的动力学方程，总能足够精确地描述骰子的动力学过程，从而预测骰子抛出以后所有的运动，也就知道了骰子落地后该是哪一面朝上。而量子力学说，量子力学所决定的随机是绝对的随机，没有一个更深层次的理论来告诉我们量子骰子的结果！

量子退相干

相干性是量子系统之所以被看成量子系统的原因。只有波的函数才能描述这些相干现象，就像相同颜色的光波可以相互叠加而形成干涉条纹。但这种干涉却不能被稳定地维持，观测、温度和噪声都可以使干涉条纹消失，我们称这样的过程为退相干。对于量子计算来说退相干是其关键制约因素。理论上量子计算描绘了一个伟大的蓝图，但实验中我们在面对几十个量子计算单元的时候总是举步维艰。实验越来越复杂，而保持量子相干性便越来越困难。量子退相干的机制是什么？是测量吗？是跟环境发生了新的纠缠吗？虽然我们有很多手段来描述这个过程，但其中的机制目前我们并不太清楚。

量子纠缠

爱因斯坦对量子力学的致命一击在于他提出了量子纠缠，这意味着两个纠缠的量子粒子即使相隔无穷远也可以相互影响，这个影响是超光速的！至少在他在世的时候，量子力学的多数研究者都努力回避直接应对他的这个挑战。

量子隐函数理论

这一切都引导爱因斯坦认为量子力学是不完备的。一套理论至少对以这套理论所用词汇提出的问题有确定的答案，并且能够证明这个答案，才是完备的因而有用的理论。然而量子力学对上述几个问题都没有完整的描述。迄今都没有。

所以爱因斯坦致力于寻找一套可以解释量子力学的理论，比如说隐函数理论。冯·诺依曼首先证明了隐函数理论是不存在的，量子力学就是这么任性；其后，继续努力的还有戴维·博姆（David Bohm），他建立了整套的隐函数理论，但这一切在贝尔不等式被实验证明之后都暂且成了历史。

爱因斯坦心中的这些抱怨，似乎物理学界很少有人再关心，爱因斯坦就只好唠唠叨叨地跟哥德尔说。就量子力学而言，可怜的哥德尔几乎是被爱因斯坦洗脑了的。可以想象，一个天真无辜的数学家每日耳濡目染，在一个当世瞩目的伟大的物理学家孜孜不倦地灌输下，几乎不可能对量子力学建立任何好感。

话题二：时间和相对论

是什么让童年的我和现在的我是同一个人？人类最早的关于时间的认识大概是因为人会变老。那些关于记忆的碎片拼接在一起构成了我们对时间的感受。但这有时候并不是很准确。当我们睁开眼睛，意

识清醒的时候，我们可以根据身边的很多参照物，比如日光的明暗来矫正自己对时间的感受，但在梦境中，这种感受会完全不同。很多测试表明，我们的一些经历非常曲折的梦只产生在大脑活跃的几分钟甚至更短的时间内。

一小时，一旦人类精神的奇怪元素介入，就可能会延伸或缩短到其时钟长度的几倍甚至几十倍。从心理学角度来看，我们如何体验时间的这种弹性？洞中方七日，世上已千年。物理学家老是把这样的说法和相对论的例子结合起来，然而对古人而言，这何尝不是心里的真实感受。快乐的时光总是短暂的，在考虑时间的乐趣和危险时，我们不可避免地会陷入困境。无论我们在直觉上如何认识，时间是一种可变的抽象，一种和直觉相关但不受直觉左右的东西。尽管时间可能仅仅是由人类的自我属性和经验的主观塑造的，但它仍然是现实中可衡量的、可观察的。主观的时间和客观的时间之间存在着具体的裂痕，“时间是什么”依然是最让人迷失但也最迷人的问题。

20世纪现代物理学，尤其是爱因斯坦的相对论将我们对时间的直观概念转变成为客观的一部分。关于时间本质描述的不完备性，也是哥德尔和爱因斯坦所关注的。1948年，哥德尔关注到了爱因斯坦的一项伟大工作——广义相对论，并成功地从其符号的简化中得出一个新的宇宙模型（参见彩插第一页）。他通过提供广义相对论的理论核心——引力场方程的精确解来实现这一目标。这个解延续了他所有工作的特征：严谨、逻辑连贯、具有完整而令人信服的高超品位。哥德尔对引力方程的解证明了爱因斯坦理论最深刻的见解，即时间是相对的，并且时间穿越是有可能实现的。

爱因斯坦的相对论只表明传统感官意义上的时间不存在，而不是时间在任何意义上都不存在。爱因斯坦本人对时间的主张更为微妙，他认为时间的改变是一种幻觉，事物没有因为时间的推进而成为某种新东西，它也永远不会成为某种新东西，它只是存在。因此，时间就像空间，它跟空间是一致的。去香港旅游时，并没有创造一个香港。

到达香港之前这座城市一直都在那里，城市并不因为旅行者的身体迁移被创造出来。同样，对于时间，人也会通过取代自己的当下来达到未来。这些事件本来就在那，不过是时间走到了而呈现在我们面前。在爱因斯坦看来，未来的事件像空间中的物体一样真实。我们顺着时间轴走到了就看见了它们。

大众文化让相对论在一定程度上被曲解了。在爱因斯坦的解释中，相对论提供了一种现实主义的时间描述，它与我们头脑中产生的主观时间截然不同。在爱因斯坦的假设中距离的定义更加广泛，它变成了空间和时间同时参与的一种度量。相对论中对时间的主观体验，似乎缺乏必要的描述，毕竟时间不可阻挡地流动构成了我们的记忆，照亮我们的昨天、此刻和未来。爱因斯坦本人在向他的老朋友、物理学家米歇尔·贝索（Michele Besso）致哀悼词时，以精确的方式表达了这一观点：“在离开这个陌生的世界时，他再一次站在我之前。这并不意味着什么。对于我们物理学家来说，过去、现在和未来之间的区别只是一种幻觉，尽管是一种持久的幻觉。”在爱因斯坦看来，就人的主观时间感受，我们是可以区分过去、现在和将来的。这种对时间的主观感受依赖于我们的内心世界，甚至是感情变化。客观的时间，因为它是相对论的特征，不能支持过去与现在和未来之间的区别。归根结底，这些关于时间的幻想是我们所收集的记忆中的故事所导致的直接结果，它们传达了我们称为历史的故事。从这个意义上说，它们不是绝对的，而是相对的，是我们制造的时间信念的基准线。

然而，爱因斯坦的相对论实际上反驳了而不是证实了人对时间的主观体验：爱因斯坦的物理学反映了时间的本质应该更简单。在爱因斯坦的时空观里，没有时间的流逝，没有从固定的过去到不确定的未来的单向流动。物理时空的时间成分与其空间成分一样。在四维时空中时间被布置和传播，时间不再是外部世界的一种属性，它至少不是普遍的“外在的世界”，而是客观世界定向运动的指南针。

话题三：关于理性主义真理

我的研究生第一次见我的时候，真诚地跟我说学物理是为了追求真理。我问他什么是真理，他说，真理就像数学一样，有完美的确定性。在数学中，一切都是对或错，是真或假，如果你已经确定真假，你也总能找出一个证明来证明你的判断。这让我想起来20世纪初人们对数学的态度。但请考虑一下：

这句话是假的。

这是真是假？如果这是真的，那么它是假的，所以它是假的。但如果它是假的，那么它是假的，这是错误的，所以这是真的！

英国哲学家和数学家罗素在这种困境中挣扎。他看到，“这句话是假的”的问题是它该声明正在谈论“自己”。如果一个句子可以避免谈论“自己”，那么就可以避免这种有问题的数学挫折，即我们所说的悖论。至少罗素希望如此。他花了十多年的时间试图一砖一瓦地重建数学基础，建立一个不依赖任何直觉的数学体系，这样就没有任何砖块是不确定的了。他撰写了庞大的三卷论述《数学原理》（*Principia Mathematica*，下文谈到这本书时，也会用*PM*代替）。他非常小心地避免悖论，甚至不认为算术是理所当然的。直到第二卷他才开始证明1+1=2。在他完成了《数学原理》之后，罗素写道，他的大脑工作机制与以前完全不同，写这本书让他的大脑受到了严重的创伤。

1931年，25岁的哥德尔表明，罗素做的事情完全是在浪费时间。

哥德尔不完备定理告诉我们，即使我们可以通过发现新证据来弥补真理体系的不完备，但这些新证据一样会导致另外的不完备。无论你如何仔细地建造自己的“数学之家”，都必须留下一块不属于你的

砖头；你的房子必须永远不完整；即使在算术中，我们也必须永远保持无知。

话说回来，一个有缺陷的真理体系，不能称之为真理体系。但哥德尔是怎么做到这一点的？

与“这句话是假的”一样，哥德尔创造了一个谈论自己的数学陈述。英语可以谈论自己，但数学可以吗？设想一下，通过将英语编码为数学。假设你为字母表中的字母分配数字：1表示A，2表示B，3表示C。然后你可以称呼某人不用他的名字而用一串数字，学校的学号就是这样的。数学证明也可以不使用字母表中的字母，而使用+、≠和∞之类的约定好的符号。我们也可以对它们进行编码，然后将1+1= 2等数学语句描述为一个代码序列，一串代码。哥德尔所做的是创造了一个数学陈述，上面写着“代码为42的陈述无法证实”。但是，他的智慧在于，他这样做的方式是，代码为42的陈述就是“代码为42的陈述无法证实”。简而言之，它老老实实地在说，“这句话无法被证实”。因此，如果这句话可以被证明，那将是错误的。但如果无法被证明，那就是真的。所以，与我们所接受的知识相反，数学中也有某些陈述无法被证明是对还是错：事实上是正确的——但你无法证明这一点！

这是哥德尔的第一个不完备定理：在至少可以描述算术运算的自洽形式系统中，一定存在既不能被证实也不能被证伪的命题。

同样，哥德尔第二个不完备定理表述为：这样一个形式系统无法证明系统本身是自洽的。

哥德尔证明，一个被认为是真实的公设永远不会在其自身系统中得到证实或完善。在任何给定的系统中，至少有一个公设必须是假的或未经证实的。第二定理说，无论你的推理多么完美，你永远不能排除某个人在某个地方或者某个时间，可能会使用你的数学系统来证明“1=0”的可能性，理论自洽性不能在自身系统内被证明。

哥德尔不完备定理接下来彻底改变了数学，并激发了冯·诺依曼（John Vonn eumann）和图灵的灵感，让他们研制出了第一代电子计算机，用来帮助同盟国战胜纳粹。哥德尔的不完备定理向我们证明了人类的思维具有一些无法被计算机模仿的特殊性质：我们人类明明可以看到有些陈述是真的，但机器却不能。英国物理学家罗杰·彭罗斯（Roger Penrose）甚至说哥德尔的定理可以帮助我们发现一种新的物理学，它将解释意识本身的奥秘。而哥德尔显然也影响了爱因斯坦对量子力学的思考，老先生晚年的挑战就是孜孜不倦地攻击量子力学的不完备性。

哥德尔自己倒觉得他最适合与莱布尼茨为伍，是典型的柏拉图理性主义信仰者。经典意义上的“理性主义”被定义为知识可以从默认真理与推导中得到，而与人的感觉、体会等因素无关。理性主义受到自然科学进步的影响，试图将数学的方法论应用于广泛的哲学研究中，从而使历来众说纷纭、各执一词的哲学获得如科学一样的精确性，这样做甚至可以消弭存在了几千年的物质与意识孰为第一性的争论。理性主义试图制定真理系统的理性原则，从中可以演绎出关于外在对象的可靠的知识；它强调理性的认知能力，这种能力既是先验的，又是本质的。理性主义者相信哲学是独立于超自然启示的逻辑；他们认为感性经验作为获得知识的途径毫无价值可言，唯有理性能够使哲学变得像自然科学一样精确。

爱因斯坦和哥德尔是柏拉图理性主义的坚定信仰者，这应该是两个无话不谈的莫逆之交的信仰基础。虽然，他们的工作都意味着经典的理性主义存在着重大问题，他们甚至扮演了经典理性主义的掘墓人的角色，但他们并不相信自己是这样的，甚至没有意识到这一点。

1951年，在爱因斯坦亲自授予哥德尔第一届爱因斯坦勋章时，冯·诺依曼说：“哥德尔在现代逻辑方面的成就是无与伦比的、不朽的——确实，它们不只是一座纪念碑，而且是一座其意义由于受时间、空间的限制还未显现的里程碑。”



图4-1 爱因斯坦亲自授予哥德尔第一届爱因斯坦勋章

哥德尔的一生

哥德尔是一位杰出的数学家和哲学家，他的不完备定理使他成为他那个时代最重要的数学家之一。哥德尔1906年出生于奥匈帝国的Brünn，现在是捷克的布尔诺。哥德尔6岁患过风湿热，他认为自己从未完全康复过。他成人以后性格偏执，容易焦虑，有时候会莫名其妙地沮丧，一生中经历了几次严重的神经衰弱，并因此住进医院。18岁时，他开始在维也纳大学攻读理论物理、数学和哲学。20世纪20年代的维也纳是一个蓬勃发展的新思想圣地。哥德尔被一群著名的思想家包围着，这些思想家组成了维也纳学派。在维也纳学派的一位领导人汉斯·哈恩（Hans Hahn）的指导下，哥德尔在23岁时完成了他的博士论文。尽管哥德尔在维也纳学派中表现出色，但他从未觉得自己适应维也纳圈子，部分原因是他的有神论信仰与维也纳学派实证主义的流行观念相冲突。实证主义者认为唯一真正的知识是可以凭经验证明的。除了为上帝的存在寻找“证据”之外，哥德尔，这个来自希特勒统治下的奥地利难民甚至发现了一种方式，能够使美国人在遵守美国宪法的前提下，依然有可能让自己的国家从民主制度沦为法西斯式的独裁统治。“你知道，哥德尔真的疯了，”爱因斯坦告诉他的一位朋友，“他投票给艾森豪威尔！”

不完备定理的证明使哥德尔迅速闻名于世。1933年开始他在世界各地进行数学讲座，这一年他在美国第一次见到了爱因斯坦。两人从此建立了亲密的友谊，一直持续到1955年爱因斯坦去世。然而，旅行也让哥德尔的心理状况变得越来越不稳定。1938年，当纳粹德国吞并奥地利时，他患上了严重的抑郁症，几乎没有回到奥地利讲课。由于无法获得维也纳大学的教职并有可能应征入伍，哥德尔与他的女友阿

黛尔·林伯斯基（Adelen imbursky），一名舞蹈艺术工作者结婚，并一同前往美国。爱因斯坦热忱地把他推荐给普林斯顿大学，在那里哥德尔开始在高等研究院任教。

整个20世纪40年代，哥德尔在普林斯顿定居并继续发展他的数学理论。他于1947年成为美国公民，最终在1953年成为一名正教授。20世纪50年代开始哥德尔将注意力从数学转向哲学，发表了几篇关于柏拉图主义和它对数学系统信念影响的论文。虽然他的哲学观点从未像他的数学定理那样被广泛接受，但哥德尔仍然是一位备受尊敬的数学教授。尽管他的职业生涯很成功，但哥德尔的心理健康似乎一直是个问题。在他生命的最后几年，他的偏执情绪继续增长。他怀疑有人试图在他的食物里投毒来害他，这是否跟图灵的死有关就不得而知了。总之，他吃的任何东西都要他妻子阿黛尔先试吃，否则他宁可饿着肚子。当阿黛尔在1977年生病并且不得不住院6个月时，哥德尔就拒绝吃任何东西。实际上他把自己饿死了。他于1978年1月14日因营养不良在普林斯顿去世，死的时候体重只有三十多公斤。

---

[1] 参见《量子大唠嗑》中的“捣蛋鬼哥德尔”。

[2] 更加详细的内容请参见《量子大唠嗑》。

第五章

哥德尔不完备定理

每个人都会有这样那样神奇的想法，真正的困难在于怎样把把握这些想法来证明它们是有效的或者有用的。哥德尔的神奇在于他从一个两千年前人们就知道的悖论出发，用理性工具本身证明了理性工具的不理性。这太精彩了。中文的书籍里面我又找不到特别好的叙述，于是我忍不住在这一章介绍了两三种证明它的方法，权且表达对人类智慧的敬意。

众所周知，数学朝着更为精确方向的发展，导致大部分数学分支的形式化工作已经完成。人们只用少数几个机械规则就能证明任何定理，因此人们可能猜测这些公设和推理规则足以使形式系统能够对可以表达出来的任何数学问题进行判定。我将证明情况并非如此。

——哥德尔

哥德尔不完备定理指出，若形式系统是自洽的，则此系统必定是不完备的。也就是说在这个系统中一定会存在至少一个有意义的命题，既不能用系统中的公设和推理规则加以证实，又不能用系统中的公设和推理规则加以证伪，即成为不可判定的命题。哥德尔第二不完备定理则说，上述形式系统本身的自洽性在这个体系内部也是不可判定的。

理性的知识系统

形式系统：一个完整的符合逻辑规则的理论系统包括一系列无须证明的公设和依照同样无须证明的推演规则得出来的命题。这些公设、推演和命题构成了本书所说的形式系统（关于形式系统的定义在前面已经讲过了，为了加深印象混个脸熟，这里再讲一遍）。接下来，当我们用到“证明”这个词的时候，我们指的是证实，即被证为“真”。

自洽：一个形式系统中自身不能有矛盾。我们给出一个陈述，这个陈述能够被推导而证明是“真”，它的反面，虽然也是个有效陈

述，不能也被证明是真的。这里我们回避用“对的”“正确的”等似乎等效的词，我知道这样会让行文略显呆板枯燥，但为了准确，这点无聊在这里是值得的。以下，我们仅用“真”这个字来代表这些概念。而对应于“假”，同样也不用“错的”“不对的”等似乎等效的词。

完备：完备性在于对于一个可做有效推演的形式系统总能对一个陈述给出符合推演规则的判断：或证实、或证伪。这跟我们的经验是一致的。我们在学校里学习的数学，从算式题到证明题，都是把一个命题要么计算出来结果，要么证明它是真的。我们可以轻易地写出一个算式的结果，因为我们知道证明这一结果是轻而易举的平常事。例如1+1=2，根据自然数的公设，我们是可以证明它的。我们能直接写出计算结果只是我们对这一证明规则非常熟悉。这里1+1=2就是一个命题，证明这个命题我们认为是不费吹灰之力的，而且大多数人相信这个证明简单到根本不需要多说，不证自明。

三角形的内角和是180°。这个命题在欧几里得几何中是真的，是可以从几何的基本公设中推演出来的，因此是一个定理。这是我们学习和建立一个知识体系的基本诉求，我们总希望用这个知识体系所规定的基本公设中给出的词汇，组织成一个新的陈述，这个陈述在这个学科体系中是可以通过公设来证明的。而这些公设应该越少越好，但至少要丰富到它能够覆盖我们想知道的内容，我们可以用这套公设系统来表达知识体系内的概念和命题。

一个公设系统下任意的陈述可以通过逻辑推演被证明，或者它的反命题可以被证明，我们就说这个系统是完备的。完备性满足了知识系统有用性的要求，而自洽性强调了知识系统的有效性。值得注意的是，这里讨论到完备性，我们并没有要求命题是符合实际的真的，它只要能够按照逻辑规则并可以从基础假设中被推导出来就好。当我们深入了解哥德尔不完备定理的时候，我们就知道“真实”和“可证”之间是有根本差别的。

终极理性的梦想

为什么逻辑演算能得到正确的命题？答案其实很简单。为了探索世界，我们需要判断命题的真假。因此我们构造了一套工具来满足我们对探求正确结果的需求，它可以保证我们从真命题必然推出真命题，我们称这一套工具为逻辑体系。正是因为我们希望逻辑体系是可靠的，所以我们才去定义了这个可靠的逻辑体系。这看起来像是狡辩，因为要让它有用所以设计了一个有用的工具，所以它有用。然而就有类似的人择宇宙学原理，为什么我们这个宇宙中的物理规律是这样的？因为只有这样的宇宙才能产生问这个问题的人。

自古希腊时期开始，数学这一知识形式就得到哲学家的尊崇。欧几里得的几何原理、毕达哥拉斯的勾股定理，无不是从当时的数学家和哲学家所认为的真实世界中被观察到的。但真实的世界为什么会展现出数学之美，历史上常常有人将其归因于宗教意义上的必然：上帝创造了这样的规则。17世纪的思想家们则将这一尊崇推到极致。且不说这些人个个都是顶呱呱的数学研究者，代表性的思想还在于，他们把数学归入了上帝法则之列。如此，通过建立古希腊时代之后的新数学，他们就可以自动跻身为上帝派来的解密者。对于欧几里得的学说，牛顿和他的追随者们既做了良好的继承，又发展出了更为复杂全面的体系，直接将数学内化为物理世界的建筑工具。通过引进参考系，伽利略（Galileo Galilei）推翻了亚里士多德的运动观念；开普勒从精美却错得不能再错的本环行星轨道和地心模型开始，通过第谷的观测数据推出行星运行三大定律；笛卡儿几乎神奇地统一了丢番图（Diophantus）的代数和欧几里得的平面几何学；芝诺（Zeno）那个关于无限的悖论通过牛顿和莱布尼茨发明的微积分成为可运算的实在规则。这时万有引力定律的推导水到渠成，它被成功地用于解释天体运行，接下来的两百年里都很好用。这些创举孕育了18世纪到20世纪初的自然科学。

从17世纪开始一直到1930年，整个数学界是非常乐观的：数学是建立在集合论和数理逻辑两块基石之上的。康托的连续统假设，即判定有理数的无穷多和无理数的无穷多之间是否有介于这两者之间的无穷多，这个时候仍然是悬案。不过大数学家希尔伯特多次觉得自己已接近解决这个难题，数学学科的前景是光明的；大部分数学可以建立在谓词演算的基础上，即形式系统的推演上。谓词演算的公设系统已经被证明是自洽的，尽管其完备性仍有待证明；整个数学的基本理论是自然数的算术和实数理论，它们都已经形式化。这些理论系统应该是自洽的、完备的。如果这一点能够得证，则集合论公理系统也能得到同样的结果，那么整个数学就牢靠了。

但细心的人对数学的客观性表示怀疑。似乎作为逻辑的高阶形式，数学依赖于公设，而公设的变化会带来不同的描述方式，我们面对的自然界并不依赖于我们采用哪一种描述方式，似乎每一种都可以，只是哪一种比较方便而简洁。

同构映射也是这个时代数学的重要进展。人们发现一个数学范畴内对象之间的关系，可以被证明在另外一个范畴的对象之间也成立。以乔治·布尔（George Boole）为代表的19世纪的数学家成功地使代数“算术化”了，并证明数学分析中的各种概念均可用数论术语（依据整数及其运算）来唯一定义。事实上这也不是遥不可及的，我们可以把$\sqrt{-1}$这样的虚数定义为整数的有序对（0，1），并对它实施某种加法和乘法运算并且定义一些操作规则，它就可以完整地表达虚数的所有性质。我们也可以把$\sqrt{2}$定义为一个有理数的类，例如所有平方小于2的有理数组成的类。利用映射的同构原理，我们可以把几何用代数来描述，用无理数和有理数来代替代数，再归纳到自然数，数学的不同领域可以通过同构来对应，而只要证明了其中任何一个系统的完备和自洽，其他系统也自然而然地被证明了。而自然数又可以通过对集合的论述来定义。因此，集合本身的性质成了整个数学界的

基石。

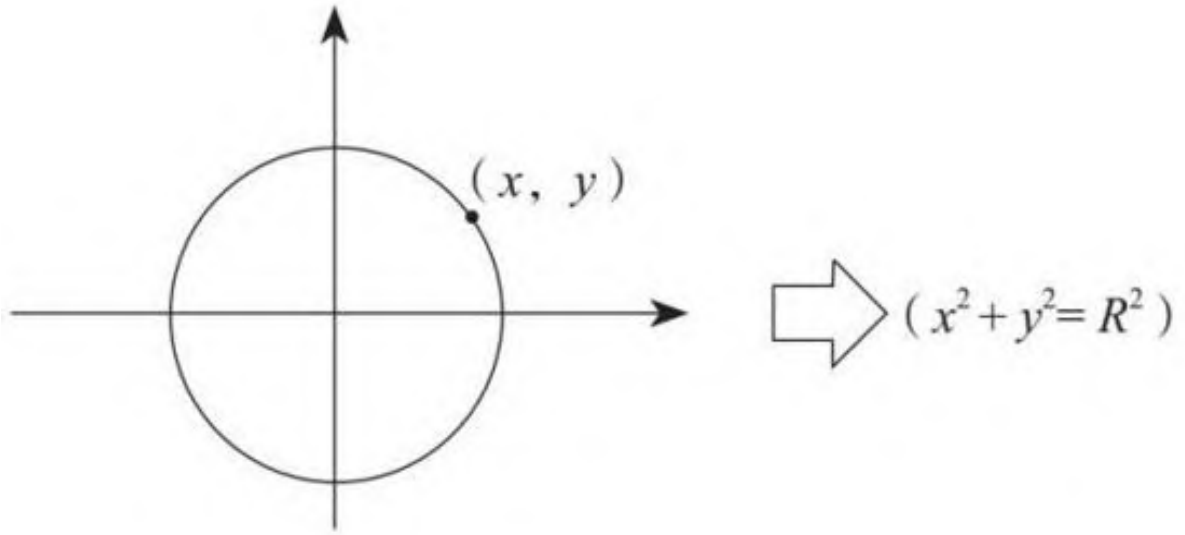


图5–1 同构映射

举例来说，与用代数公式表示空间中曲线和曲面之间错综复杂的几何关系，要比直接处理这些关系更容易一样，处理复杂逻辑关系的算术对应，要比直接对付这些逻辑关系要容易些。由此，可以去构架一套描述这套关系的数学体系，它独立于任何具体的数学内容，它是用来描述数学的数学，称之为元数学。我们会看到，哥德尔也利用了这个方式，把逻辑推演变成合理的算术问题，用算术形式来演算关于算术的元数学问题。在这个基础上，如果关于算术形式系统的复杂的元数学命题能够被翻译成在这个系统内的算术命题的话，将大大有助于元数学本身自洽性和完备性的证明。

映射是个不难理解的事情，建立事物和事物之间某种对应关系，作为一种方法被我们在各种场合不严谨但广泛地应用着。语言是我们对自然界和自己思维的映射。我们用“猫”这个字代替了自然界中对应的动物，对于它的陈述让我们理解猫在自然界的状态，而不是把猫抓来做演示。我们同样设计了棋类游戏，比方说象棋，被用来推演军事战争，或纯粹是游戏，但“帅”“士”“兵”被用来标记不同棋子

的重要性，这些重要性也同样可以用1、2、3等数字来标记。但严格意义上的映射却不那么容易获得。当我们用映射的方式来证明某类问题的时候，要严格地证明映射的对应关系，而不是凭着名词词义的类似就借用过来作为证明的依据，这往往构成论证的错误。比如我们谈“上善若水”时，并没有论证“善”和“水”的映射关系，仅凭某类特点的相似，是推不出其他延伸的结论的。我们往往混淆了修辞和论证的关系，这个例子里的比喻论证不是论证，只是一种修辞手法而已。而我劝读者能少用就少用，尤其在一些相对严格的文体中，比喻的使用会带来大量的不经意的逻辑混乱。同构映射在19世纪取得了重大进展，很多不同分支的数学领域都被证明是可以同构映射的。这样在一个数学分支中找到的规律，在与它同构的数学分支中就存在同样的规律。数学家常常只是选择证明起来会更方便的那个分支而已。

理性的典范：算术

数学可以还原为算术，而算术可归结为皮亚诺系统的几条公设。我们重复一下前面介绍过的皮亚诺算术系统的几条公设：

公设*1* 1是自然数；

公设*2* 对于任意自然数n，其后继数n'都是自然数；

公设*3* 对于任意自然数n，$n' \neq 1$都成立；

公设*4* 对于任意自然数m，n，若$m'=n'$，则$m=n$；

公设*5* 假设对自然数n的谓词 $P(n)$ 而言，下面的（a）和（b）都成立。

（a）$P(1)$ 成立；

（b）对于任意自然数k，P（k）成立，则P（k'）成立；

则，对于任意自然数n，P（n）都成立。

在这基础上，我们可以将自然数集合扩充成整数、有理数、实数，可以定义加、减、乘、除、幂指、对数运算等。比如：

加法的定义：（1）$m+0=m$；（2）$m+n'=(m+n)'$；

乘法的定义：（1）$m\times 0=0$；（2）$m\times n'=m\times n+m$。

重申了自然数的定义之后，我们尝试用集合构造算术：这里用符号∪表示并集，原来的集合加上一个新的集合。如果读者不记得什么是集合，我们稍稍多说两句。集合，是对象的总称。比如自然数的集合{0，1，2，3……}，不同的版本中会规定0是不是自然数，我们不管它，这里把0作为自然数。一个学校所有年级的集合{一年级、二年级……}，等等，这些构成了集合的基础。我们甚至可以从集合里来定义自然数。

空集是没有任何元素的集合，即{ }，又记作∅。

我们在集合上也定义后继运算：m'∪{m}。

用集合定义自然数：

0=∅

1 ={0}

2=1∪{1}={0，1}

3=2∪{2}={0，1，2}

4=3∪{3}={0，1，2，3}

……

将<定义为：$m<n$为$m\in n$，即m是n的子集。所以，0 < 1 <2 < 3 <……

在集合论中有这样一条公设：所有的自然数形成一个集合。

在集合论中可以证明，这样定义的自然数满足皮亚诺的五条公设。既然数学可以还原为算术，算术可以还原为集合，那么我们说数学可以还原为集合。

上面用集合定义自然数，相当于是给不同的集合取名字。空集∅的名字是0，空集的集合{∅}的名字是1。并不是说1和{∅}两者真的相等，否则会产生问题，1={0}怎么会成立呢？左边是一个数而右边是一个集合？

集合构造算术的关键在于，这样构造出来的集合，刚好满足皮亚诺公设。所以我们研究集合，就可以弄清算术的性质。这种构造方法并不是唯一的，只要它们满足皮亚诺公设就行。弗雷格（Friedrich L.G. Frege）把逻辑形式化为集合论的一部分，在集合论对数学的大统一上做出了卓越的贡献。事实上我们看到一种希望——通过集合来统一数学的理论体系。

就在弗雷格要把他关于集合论的恢宏巨著付印的前一天，他收到了来自英国的数学家罗素的一封信。罗素一直关注着弗雷格的研究。但这封信充满了历史的黑色幽默，如果对一个数学家来说还能笑得出来的话。这封信里，罗素提出集合的一个悖论，它使弗雷格利用集合论建立起来的算术、代数、几何等数学大厦的根基发生动摇。用弗雷格的话说：“突然数学大厦的基石崩塌了。”他发现，自己忙了很久得出的一系列结果被这条悖论搅得一团糟。他只能在自己著作的末尾写道：“一个科学家所碰到的最倒霉的事，莫过于是在他的工作即将完成时却发现所干的工作的根本依据崩溃了。”

罗素在给弗雷格的信里写道：

亲爱的战友：

我完全同意您的观点，您所论证和讨论的问题的重要性是很多逻辑学家的工作都无法比拟的。

我就有一个小小的困惑想向您讨教，考虑那个不包含任何元素的集合，它是否包含自己呢？

例如，如果以所有集合为元素组成的集合U，便符合$U\in U$的条件。但一般的集合，例如北京大学的所有同学的集合V，便不是它本身的元素$V\notin V$。现在我们用A表示所有不以本身为元素的集合所成的集合，即若$x\in A$，则x是一个集合，且$x\notin x$，现在我们问$A\in A$是否成立？即A是否包含它自己。若 $A\in A$，则A具有A的元素的性质，A就是A的元素，但A明明是所有不以本身为元素的集合，这与A的定义矛盾；反之，若$A\notin A$，则A不具有A的元素的性质，因此A符合A的定义，它是不包含它自己的集合，即 $A\in A$ 成立，这也是矛盾的。但$A\in A$和$A\notin A$二者之一必成立，所以矛盾是躲不掉的，这样就形成了悖论。

罗素本人并没有把包袱丢给弗雷格，他接下来花了十几年时间来研究这个问题，写了厚厚的三大本《数学原理》，试图绕开这个逻辑悖论。数理学家、哲学家约翰·凯梅尼（John Kemeny）后来说罗素的《数学原理》是一本“被每个哲学家谈论，而实际上又无人读过的伟大著作”。罗素希望绕开悖论的策略大致是这样的：我们可以定义不同层次的语义和逻辑规则，通过控制这些语义和逻辑规则，有可能划分出来一条清晰的渠道，来避免自指问题的发生。即，通过严格的语义和语法的层次分割来避免发生陈述指向自己的情况。

关于数学的本质，罗素有一个著名的论断：纯数学是这样一门学科，在其中我们并不知道我们在谈论什么，或者我们不知道我们所谈论的是否是真的。这精辟的评论直接切入希尔伯特伟大工程的核心——为全部数学构建一个纯语法的框架。希尔伯特受到刺激的主要原因是，他感觉到，类似于罗素悖论中的核心问题，是由于陈述中悖论

性的语义内容所导致的。希尔伯特相信，根除在数学中出现这种悖论的办法，就是建立一种实质上无意义的元数学框架，我们在其中可以谈数学陈述的真或假。这样的框架构成了数学的形式系统，也构成了数学中研究“什么是可证明的”与“什么实际上是真的”这两者之间的鸿沟的历史性起点。基于对完美形式系统的期望，希尔伯特开始了他的伟大工作。希尔伯特的数学终极理性是这样的：

1. 将整个数学体系严格形式化；

2. 用元数学来证明整个数学体系是建立在牢不可破的坚实基础上。

除去罗素的怀疑，元数学系统已经可以完全形式化：将系统内的所有表达式的意义都抽掉，把它们视为空洞的符号，按照事先规定的一组规则，对这些符号进行操作。这样我们就可以构建一个本身没有任何意义的符号系统来演算或者推演。这套符号系统的所有含义都由我们外部加以定义。这件事情其实并不抽象，考虑一台计算机，那么计算机的制造者并不关心我们拿这台计算机来做什么，写文章、编程序、打游戏还是看电影。他们只是按照已经设计好的规则来做芯片，焊接电路，把零件组装在一起，再安装一套操作软件。计算机本身就可以看作是这套元数学符号演算规则在物理世界的体现。因此具体实施的步骤也有了：

1. 将所有数学内容形式化，用确定而唯一的符号表达，彻底摆脱自然语言的模糊。

2. 证明数学的完备性，我们能够证明所有的真理，只要是真的命题就可以被证明。

3. 证明数学的自洽性，数学本身不会自相矛盾，确保我们在不违背逻辑的前提下获得的结果是有意义的，不会出现某一个陈述，它既是真的又是假的。

4. 数学的可判定性，我们可以找到一个算法，自动地、机械地判定数学陈述的对错。

在这个大背景下，罗素的《数学原理》提供了一套全面的符号系统，借助于这套符号，所有的纯数学命题（特别是数论命题）都可以用一种标准的方式来表示；同时《数学原理》明确了数学证明所用的大多数形式推理规则。

语义和语法

亚里士多德的逻辑学有两个支柱：不经证明而被看作真的前提或公设；保证从一个真陈述变换为另一个真陈述所使用的推理规则。我们以古典逻辑中的亚里士多德三段论为例说明，这样的结构也常常被称为大前提、小前提和结论。

大前提：所有人都要上学。

小前提：小明是个人。

结论：小明要上学。

在这里，大前提和小前提被设想为关于人、关于上不上学和关于一个特定的人——小明的真陈述。从大前提、小前提到结论的得出，这中间利用了由亚里士多德概括的演绎推理律中的一个，“如果所有X都是Y，且Z是一个X，那么Z是一个Y”。只要我们可以确信前提为真，那么小明要去上学这个结论就是牢靠的。确定性来自前提的语义内容，来自我们的心智所确定的演绎过程。

但是，我们看下面这个例子：

证明：亚里士多德是工作。

这是严格按照三段论推导出来的结果：

1. 亚里士多德是哲学家。
2. 哲学家是工作。
3. 所以，亚里士多德是工作。

上面这个结论恐怕连亚里士多德本人也不会认同。

这个推论中蕴含了语义悖论。因为语句1中的“哲学家”和语句2中的“哲学家”不在一个语义层次上，前者是对象概念，后者是词汇的元概念。两个前提内涵不一致，结论就荒谬了。这种逻辑错误很多时候会将我们带入类似于“理发师悖论”。比如当我们谈到“这句话是错的”的时候，我们显然是在两个层次上谈论。从语义和语法的角度区分，能不能回避因为混淆这两个层次的概念而造成的自己指向自己的悖论呢？造成逻辑悖论的所有情况，都是类似于它们以所谓的自指概念为基础。一个典型的例子是“说谎者悖论”（Liar’s paradox）：

克里特岛的一位居民说：“克里特岛上所有的人都说谎。”

“这句话是错的”，这个陈述中包含了真的概念。但我们会看到，哥德尔事实上证明了真概念不可能在正被谈及“真”的那个形式系统的界限内被加以把握。

首先，我们定义一下要用的名词，用普通数学语言来描述的直觉算术（intuitive arithmetic，以下称为IA），IA里的陈述句表达有

关算术性质的真伪，它有确定的语义的意义，真实表达了某一算术的概念和性质。

接下来是算术的形式推演系统，在哥德尔的推演中，他使用了罗素在《数学原理》中定义的纯语法系统，系统中给定一组公设和明确的推演规则，通过这组公设和规则，可以决定哪些句式是定理，这称之为形式算术（formal arithmetic，以下称为FA），FA的陈述没有具体的语义，它本身并不代表性质意义上的真伪。

当IA和FA一一对应的时候，FA就被赋予具体的含义，FA的符号会被赋予对应于IA的算术性质。我们也可以把直觉算术IA描述为目标性表述，形式算术FA理解为方法与形式的描述。FA被设计成IA的镜像。如果一切依计划，FA中的数字符号和IA中的数字会有一一对应的关系，而且IA中为真的陈述句和FA中的定理也会有确实的对应。简言之，FA是用来表示IA的。

《数学原理》里的公式原来只是无意义的字符串，无关真假，能够用演绎规则从公设和其他定理推演出来的陈述，被称为定理，这个过程称为形式证明。将一些符号解释成逻辑关系，赋予公设真值，一些公式便有了逻辑的含义，它们被称为命题，定理便是逻辑上为真的命题。将FA的符号解释为算术符号，哥德尔用正式证明前的46条定义和预备定理，建立起元数学真理和算术命题的对应关系，因此赋予了这些公式的数学含义。元数学是在（数学原理）上层谈论公式的性质和关系的语言。

第三种语言或理论是形式算术的后设理论，也就是用来清楚地说出FA的语法规则和证明理论的语言和理论：这个语言成为超形式数学，是描述形式算术的形式算术（metatheory of formal arithmetic，以下称为MFA）。换句话说，如果FA是机器，MFA就是它的操作手册。MFA是用来说明FA这台机器是如何运作的。和IA一样，MFA也是由具有意义的陈述句组成的，其中的陈述句皆可依照它的意义来判断真伪。MFA规定FA中哪些语句是良性句式（well formed

formula，以下称为*WFF*），也就是哪些陈述句满足构建句式的规则，也同时规定了FA的证明是什么意思。

希尔伯特的一个洞见对哥德尔的证明具有相当大的启发。希尔伯特意识到数学分支的每一个形式化本身便是一个数学对象。在此以前，哥德尔已经发现可以通过使用自然数本身来镜像反映有关自然数之间关系的所有陈述。这是相当关键的一步，它使对象自己述说自己成为可能。

哥德尔发现FA不但可以用来表示IA，同时也可以代表MFA。他利用现在被称为后设数学的算术化（arithmetization of metamathematics）的革命性方式，证明IA可以代表MFA。但如果FA可以代表IA，那么也可经由MFA在IA中解读的方式来代表MFA。因此FA可以代表它自己的后设理论。接下来就是构建一个FA中的语式，该陈述同时在IA和MFA两种语言中有两种意义。哥德尔找到了这样的陈述，可以同时在FA中为不可证并直观上在IA和MFA中为真。这个陈述可以在FA中证明为不可证，但又在IA中代表一个为真的有关自然数的陈述句，同时又在MFA中成为有关它自己的不可证明为真的陈述句。哥德尔的第一步工作，就是证明FA可以用来代表IA。这部分工作事实上被皮亚诺公设系统完成了，即一套适用于数论基础的严格逻辑规则。第二步需要证明FA可以表示MFA。通过哥德尔所用的配数方法，哥德尔首先系统地将FA中的基本符号编号，接着证明如何以序列的基本符号，也就是陈述，对应到自然数，不同的序列因此会有不同的编号。接着就可以为一序列的陈述来编号，一个证明就是一序列的陈述，因此也可以将证明对应到自然数。

真概念不可能在正被谈及“真”的那个形式系统的界限内被加以把握，为了避开永远难以捉摸的真与假的概念，哥德尔意识到应该用某种可以形式化的东西来代替“真”，这就是“可证性”的概念。于是，哥德尔利用他的编码方案，通过46条引理来编码上述悖论，并将其翻译成：

这个陈述是不可证的。

这被称为哥德尔陈述，我们用$G-S$来代替。哥德尔的思路是：对于每一个形式系统，总可以构造出一个从系统外来看是真的，但在系统内却不能判定的陈述$G-S$。接下来我们讨论一下这个论断的逻辑结果：

如果哥德尔陈述$G-S$是可证的，由于陈述$G-S$是真的，根据论断，它不可证。因此，在这个系统中，陈述$G-S$和它的否定都成立，所以，这个系统是不自洽的。

如果哥德尔陈述$G-S$是不可证的，由于陈述$G-S$是真的，但是不可证，所以系统中至少有一个命题，它是真的，但其正确性无法被系统已知的定理证明，所以，这个系统是不完备的。

于是哥德尔得出了“对于算术的任何自洽的形式化系统，都存在着在该形式系统内不可证明的算术真理”这一惊人的结论。

让我们试图用通俗（不太严格）的说法来理解哥德尔的不完备定理，以及他的证明方式。

首先将哥德尔的两个不完备定理翻译成通俗语言：

1. 一个算术系统，要么自相矛盾，要么总能得出一些包括在该系统中但无法判断真伪的结论；

2. 不可能在一个算术系统内部，证明此系统没有自相矛盾的。

哥德尔用“可证性”概念代替“真”的概念，然后把悖论“这句话是错的”拓展为论断“这个陈述是不可证的”。进而证明语句“这个陈述是不可证的”有一个算术的对应陈述，即在算术形式化中有一个自然数对应为哥德尔陈述$G-S$。在此基础上，他需要做的证明就变成了：

1.不完备性：证明如果该形式系统是自洽的，则此哥德尔陈述必定是真的，但它陈述了自己不可证。

2.不完备性附则：证明即使补充的公设加进来构成了新的系统，使得这个陈述$G-S$在其中可证，含有补充公设的新系统仍有它自己的不可证的哥德尔陈述。

3.自洽性：构造一个算术陈述，它断定“算术是自洽的”。证明这个算术陈述是不可证的，故此证明，算术作为一个形式系统不足以证明自己的自洽性。

接下来FA还有所谓形式正确（well formed）的要求，以及更重要的，由一个字符序列组成的陈述是否能构成证明的问题。哥德尔证明的是，要完整描述FA语法所有的必要概念和函数，包括形式正确的语句和证明，都对应到IA中一些特定的递归函数。直观上来说，所谓的递归函数就是可以用机械方式计算的函数。这类函数也可以严谨地用数学的语言来刻画其特性。

哥德尔证明了FA中MFA的基本概念或证明理论，可以对应到IA中某些特定的递归函数。在证明FA可以表示IA的过程中，哥德尔证明IA中的任何递归函数，都可以在FA中表示出来。说得更明确些，如果在IA中有关于某一递归函数的真实叙述，就会有一个FA定理为其对应的陈述句。一旦他成功证明MFA中的基本函数，若以数字来编码会产生递归函数，就可以得到MFA，正如IA一般，可以在FA中表示。以算术的方式编码处于不同语义层次的逻辑表达式是哥德尔得出不完备定理的重要一步。哥德尔发明的这种方式可以使用算术语言本身对算术中的所有可能的陈述进行编码，通过自然数本身来镜像反映有关自然数之间关系的所有陈述，即算术既可作为被解释的数学对象，又可作为谈论自身的未被解释的形式系统。在形式系统内部构造一个命题，又表明它在系统中不可判定，哥德尔由此推出自洽性在系统自身中的不可证。

通过哥德尔配数，哥德尔把FA的语法算术化了，因此有关FA的语法事实，可以对应到IA总有一个有关自然数的事实。这样一来，哥德

尔证明了FA中的理论可以同时代表IA中有关算术的事实，以及通过哥德尔配数代表MFA语法的事实。也就是说，FA中的定理，将可以代表IA中的一个数学真理，而该定理本身通过哥德尔配数可以代表有关FA的语法真理。如此，哥德尔成功地证明了FA本身虽然是一个没有意义的形式化符号系统，却可以通过一语双关的方式被赋予两种不同的意义，以确保其可用来同时表示数论和语法，这里的语法指的是FA本身的证明理论。换言之，通过自然数，FA中的陈述句可以表示FA本身。

哥德尔接下来居然证明了算术是数学复杂性的分水岭。

哥德尔特意声明，他的结果的根本点不在于任何形式系统都无法包罗全部数学。这一点从康托的对角线法已经能得出了，他并未排除数学的某些相当强的子系统具有完备性的可能。相反，哥德尔的结果的根本点在于每一个既含加法又含乘法的数学形式系统都包含一些颇为简单的命题，可以在该形式系统中被表达出来，但不可判定。

哥德尔的定理说，如果一个形式系统是：

1. 有限大的；
2. 足够大以包括算术逻辑；
3. 自洽的；

则它是不完备的。

算术结构在哥德尔不完备定理的证明中起着核心作用。数字的特殊属性，例如任何数字都可以作因式分解，表达为一些质因数的乘积（例如“130=2×5×13”），哥德尔考虑用一种方法来建立数学性质和关于这些性质的陈述之间的对应关系。他成功的证明了像“说谎者悖论”这样的自然语言陈述悖论可以像自然数一样嵌入描述数学性质系统所用的语言结构中。只要足够丰富的逻辑系统包括算术就不得不包容这些悖论。要是我们选择了一个理论，该理论仅包括前十个数字

（0，1，2，3，4，5，6，7，8，9），这个阉割版的算术可以是完备的。它一样可以生成有关个别数字的陈述，如加法和乘法运算。欧几里得几何只描述连续的点、线和面，一般来说，它不能包含所有算术的逻辑。因此欧几里得几何是可以完备的，同样，非欧几里得几何也是完备的。如果我们有一个逻辑理论处理只使用这个概念的数字“大于”，没有提到任何具体的数字，那么它将是完备的：我们可以判定关于仅涉及的实数大小关系的陈述的正确与否。

“小于”算术的系统的另一个例子是没有加法而只有乘法操作。这叫作Presburger算术（皮尔斯伯格算术）。它是个完备的算术版本，但是是“阉割”过的版本，即如果有加法就不能有乘除法。起初这听起来很奇怪，我们在日常中遇到的乘法只不过是作为加法的简写方式（例如“2+2+2+2+2+2+2=2×6”），但是算术的完备逻辑系统存在逻辑量词，如“存在”“任何”，乘法允许结构不仅仅是等价的一连串的加法的简单补充。

Presburger算术是完备的：所有关于添加自然数的陈述数字可以证明或反驳；所有正确的命题都可以从公设中得到。同样，如果我们创建另一个“阉割”版本的算术，没有加法但保留乘法，这也是完备的。只要加法和乘法同时存在，不完备性就会出现。延伸系统通过向基本操作添加指数等进一步额外操作也不会带来结构性上的任何区别，不完备性仍然存在。

哥德尔不完备定理的证明

接下来的内容会比较有难度，但为了表达对哥德尔的敬意，我们重新走一下那个奇妙而又幽默的智者走的路，当然，你也可以赤裸裸地跳过，知道结果就好。

哥德尔证明不完备定理的四个步骤如下：

1. 先考虑相对比较容易的数论的自洽性问题；

2. 发现数论中的真理在数论中不能定义，这与数论取自哪个形式公设系统无关；

3. 从真实性转向可证性，构造了一个不可判定论断；

4. 发现完备性陈述自身也是不可判定的。

哥德尔那20页的著名证明极为紧凑，它有46个预备定义，另外还要证明5条预备定理，才开始正式的证明：构造一个既为真但是在所考虑的形式系统内部又不能证明的算术命题。虽然哥德尔的原始证明晦涩难懂，但他的策略却极为简单，一个离自相矛盾如此接近但并不是自相矛盾的证明，不是证明“这句话是错的”，而是去证明一个为真却不可被证明的算术命题，“这个陈述在这个形式系统中不可证明为真”。哥德尔的聪明之处在于，他把每个人都知道的两千多年前就被提出来的逻辑问题，巧妙地变成了一个算术问题。

“这个陈述在这个系统中不可证明为真”与“这句话是错的”，二者有着本质的区别。“这个陈述在这个系统中不可证明为真”（我们称之为哥德尔陈述，用符号$G-S$来代替），即“$G-S$在这个系统中不可证明为真”，$G-S$是一个有着确定真值的句子，它“真”，只不过它不可证明为真。而“这句话是错的”，没有确定的真值。这样$G-S$的否命题就是“$G-S$在这个系统中是可证明为真的”

如果$G-S$在这个系统中可证明为真，那么它的否命题“这个陈述在这个系统中可证”就是真的，因为它陈述的就是这件事。它是真的。那么$G-S$就是假的，因为当一个命题的否命题是真的，它本身就是假的。因此，我们要求一个系统是自洽的，因为不自洽的系统可以推出任何结论。$G-S$就是这样一个陈述，如果可证明它为真的证明存

在，它就不得不既是真的又是假的，这与系统自洽的要求矛盾。为了维护系统的自洽性，我们只能放弃对*G－S*的真实性证明的存在这一要求，即，*G－S*不可证明为真，或者说，不存在对*G－S*的证明。这正是哥德尔的结论，如果我们要求系统是自洽的，在系统中就存在可以表述为真却不可以证明为真的命题。并且*G－S*有直接的算术意义，如果系统是自洽的，那么哥德尔的证明表明算术真理在算术的形式系统中不可证。对算术系统而言，它要不不自洽，要不不完备，总存在某些真理，无法被证明。

问题并没有停留在这里，哥德尔同样表明，即使把*G－S*当作公理加入系统，来试图通过“无须证明”挽救系统不完备的时候，这样会产生一个新的、扩展过的形式系统，那么在这个系统中一样可以构造出*G－S*的对应陈述，它为真，但不可证。结论是，在任何即使包含算术这样简单形式的系统中，如果系统是自洽的，没有矛盾，那么就一定存在不可证但为真的命题。一个有丰富内容的逻辑系统，不能既自洽又完备。

这个结论看起来很神奇，我们尝试着探索一下它的细节，这应该是本书最艰难的部分，但是值得我们以朝圣的心态，重新来一遍。

第一步：构建一个形式逻辑系统

哥德尔的大量陈述是基于罗素的《数学原理》，所以他先用了*WFF*，*WFF*指有效陈述，被用来指代符合形式逻辑的陈述，即从公设出发，通过逻辑推导得出来的陈述。为了简化逻辑推导，罗素和他之前的数学家们把逻辑推导简化成几个基本概念以及相应的符号。比如，“陈述*P*和陈述*Q*为真”，可以等同于“陈述*P*为假或陈述*Q*为假”

不对。

命题：（x）（x'） [s（x）=s（x'）➞（x=x'）]

上面这个命题就是一个有效陈述*WFF*，它表明如果两个数有同样的后继，那么它们就是同一个数。它读作：任何x和任何x'，当x的后继等于x'的后继，那么x=x'。这里符号“（）”代表了“任何”这个概念。

第二步：哥德尔配数

哥德尔发明了一套算术方法，自动给形式系统中的每一个命题配一个独一无二的数。对一个有效陈述*WFF*，它的哥德尔配数表述为*GN*（*WFF*）。在今天看起来这一点都不难理解，我们可以给每一个汉字配一个数来代表这个汉字。1980年，为了使每一个汉字有一个全国统一的代码，我国颁布了第一个汉字编码的国家标准：GB2312－20。在这个标准中“清华”可以用“3969 2710”来代表。

罗素的形式系统中包含各种类型的对象：字母表中的符号、符号组成的有效陈述*WFF*以及*WFF*的特定序列（证明），这一切都建立在这个字母表的基本符号上，*WFF*是这些符号的序列，证明又是*WFF*的序列。哥德尔配数从给字母表中每一个基本符号指定一个数开始，根据规则给每一个*WFF*指定一个数字，继而可以给*WFF*所组成的证明序列一个数字。这样，一旦给每一个*WFF*都指定了一个数字，我们只需要分析命题的相应配数之间的算术关系就能分析命题之间的关系。反过来也一样。比如说，如果两个*WFF*的哥德尔配数正好以某种方式的算术关系相关，则一个*WFF*就是另外一个*WFF*的推论。换句话说，算术描述——阐述形式系统内可以表达的数之间的关系；元语句——阐述系统内*WFF*之间的逻辑关系，这两个不同类型的描述在哥德尔配数的帮助下是同构映射的。这里这些元语句是纯粹语法式的，因为它们纯粹是形式系统的语法，即推导规则。

哥德尔配数规则同时确保同一个哥德尔配数不会被指配给不同的对象，不会发生一个数字既指向某个*WFF*（命题），又同时指向一个*WFF*的序列（证明），任何一个*WFF*或*WFF*的序列都指向唯一的哥德尔数。给定一个哥德尔数，我们可以知道它代表着系统中的哪一个有效

命题。它还保证了一个进一步的条件，由系统中*WFF*之间的逻辑关系翻译出来的算术命题同样是一个有效的命题。

举例来说，一个命题*WFF*2，它对应的哥德尔数为*GN*（*WFF*2），如果它可以由有效陈述*WFF*1——对应哥德尔数为*GN*（*WFF*1）——推导出来，那么*GN*（*WFF*1）就是*GN*（*WFF*2）的一个因子。这时我们可以用两种方式来表达这个证明：我们可以依据形式规则从*WFF*1开始推导出*WFF*2，我们也可以证明*GN*（*WFF*1）同某一个整数相乘可以得到*GN*（*WFF*2）。假设*GN*（*WFF*2）= 22865389，而*GN*（*WFF*1）=3119，那么*GN*（*WFF*2）=*GN*（*WFF*1）×7331，*WFF*2就是*WFF*1的一个推论，当然，这里7331也应该是一个有效命题的哥德尔数。

这样，哥德尔实现了元语法和算术的相互重叠。一旦有了这种逻辑含义和算术关系的重叠，命题的可证性这个元语法关系将成为一个算术关系，一组*WFF*将是某种算术关系的属性，只有可证明的*WFF*（命题）才能具有某种算术演算的属性，不可证的命题，即使它是有效命题也不能具有这种算术属性。

表5–1 基本逻辑符号的哥德尔配数（简化版）

符号	哥德尔数	意义
~	1	非
∨	2	或
⊃	3	如果……那么……
∃	4	存在
=	5	等于
0	6	零
S	7	后继
(	8	标点符号
)	9	标点符号

首先考虑逻辑公式：

（∃x）（$x=Sy$）

它的意义是：存在着一个数x，它是数y的直接后继。其中x、y是两个数值变量，用大于10的质数来表示；令x=11，y=13；再用表中相

应数字来替换逻辑公式中的其他符号。这样一来，逻辑公式（∃x）（x=Sy）经过编码，成为数的序列：

[8，4，11，9，8，11，5，7，13，9]

这个序列唯一地确定了逻辑公式（∃x）（*x=Sy*），这些数称为哥德尔数。因为算术是研究数的性质的，所以我们希望用一种更明确的单个的数来表达这个公式。根据算术基本定理，我们知道，所有的自然数都可以唯一地分解成质数的乘积，即：

$$N = 2^{r1} \times 3^{r2} \times 5^{r3} \cdots\cdots \times p^{rs}$$

于是，哥德尔取质数中的前10个质数（逻辑公式经过编码后由10个数表示），对每一个质数，求其在公式中所对应成分的哥德尔数次幂。这样，逻辑公式就定义为下面的数：

$$(\exists x)(x = Sy) \Rightarrow 2^8 \times 3^4 \times 5^{11} \times 7^9 \times 11^8 \times 13^{11} \times 17^5 \times 19^7 \times 23^{13} \times 29^9$$

所得的积就是逻辑公式（∃x）（x=*Sy*）最终的哥德尔数，大概等于1.4567×10^{86}。虽然这个数字非常大，但它总是存在的。

第三步：构造一个命题，它为真是因为它说它不可证明为真

Pr（x）是一个关于数的属性。形式系统中所有*WFF*都通过哥德尔配数被赋予了一个数，对于每个有效陈述*p*，都有*GN*（*p*）的自然数与之对应。对于这其中的可证明为定理的有效陈述，定义*n=GN*（*p*），这时*Pr*（*n*）为真（即数字*n*对应的陈述*P*是provable，可证的）。这对应的就是在说，有这个自然数*n*，作为哥德尔配数，它对应于一个定理*p*，因为*p*是可证明为真的。所以*Pr*（*x*）只有两个值，或为真，或为假。换句话说：

Pr［GN（p）］当且仅当p可证。

这句话的意思是，当且仅当陈述p是可证的，对应p陈述的哥德尔数具有Pr的性质，或者说这个哥德尔数的Pr值为真。

那么如果我们定义这样一个数的属性$\widetilde{Pr}$（x），它是Pr（x）的反命题。它为真的时候，对应于命题x不是一个定理，换句话说，就是：

$\widetilde{Pr}$［GN（p）］当且仅当p不可证。

这句话的意思是，当且仅当p是个不可证的有效命题，对应p陈述的哥德尔数$\widetilde{Pr}$为真，即哥德尔的不可证性为真。哥德尔先证明了一条对角线引理，我们这里直接引用这条引理。对角线引理声称，对于任意单变量命题函数F（x），都存在一个数n使得F（n）的哥德尔配数正好就是n本身。既然对角线引理对于任意命题F（x）都成立，我们自然可以定义F（x）=$\widetilde{Pr}$（x）。这样，我们引入一个数g，它应用F（x）=$\widetilde{Pr}$（x）这个函数的对角线引理得出，g=GN［$\widetilde{Pr}$（g）］。这个公式宣称g是陈述g不具有Pr属性命题的哥德尔数。由于哥德尔配数，我们要记得g其实也是一个很大的自然数。那么g是哪个命题的哥德尔配数呢？这个命题是$\widetilde{Pr}$（g），我们用G来代表这个命题，它在说g不具有可证的属性，即g=GN（g），考虑：

$\widetilde{Pr}$［GN（p）］当且仅当p不可证。

其实这里可以让p=G代入，即：

$\widetilde{Pr}$（g）当且仅当g不可证。

这句话的意思是，当且仅当G是个不可证的有效命题，对应G陈述的哥德尔数g具有$\widetilde{Pr}$的性质，或者说这个哥德尔数$\widetilde{Pr}$（g）为真。也就是G为真当且仅当G不可证。

哥德尔展示的不仅在我们所讨论的算术形式系统中，而且也在所有包含了算术的形式系统中如何构造一个为真但又不可证的命题。因此，如果我们想扩展G为公设，构造一个新的形式系统，一个新的类似命题又可以从那个系统中构造出来。这样下去，直到任何数量的有限条公设被列入都不能最终解决这个问题。在任何包含了算术规则的形式系统中，只要系统是自洽的，总存在不可证但为真的命题。

这就是哥德尔第一不完备定理。

定理*1* 形式系统如果是自洽的则是不完备的。

换句话说，哥德尔第一不完备定理是说，如果算术的形式系统是自洽的，那么就存在命题G为真但不可证。

这句话分为两部分：一是条件，如果算术的形式系统是自洽的；二是结论，存在命题G为真但不可证。但如果算术形式的自洽性可以被证明，那么这句话本身就是结论的证明，所以“如果”这个条件是不成立的。因此算术的形式系统的自洽性在这个系统中是不可被证明的。

这就是哥德尔第二不完备定理。

定理*2* 形式系统不能证明自身的自洽性。

G是不可证明的事实，说明元数学描写G的命题是真，它对应着这命题G的含义为真。《数学原理》里一个表达真理的命题G，不能在

《数学原理》里被证明，说明《数学原理》是不完备的。

是不是在《数学原理》里加上这个不可判定的命题作为公设，系统扩张后就可以让它完备起来？哥德尔说，这也不可能！即使加上新的公设，按照相同的逻辑，还可以证明有新的不可判定的命题，这让公设化的原始设想完全落空。该命题简单的推论是：数论里有无穷多个真理不能用数论里的公设和定理来证明。不断引进数论外的定理后，总是还有不能用形式逻辑证明的数论真理。

不幸的是，自洽是不可能在形式系统中得到证明的。系统自洽性的要求是系统里的一个命题和它的否定命题不能同时是定理，这是数学系统无矛盾性的要求。如果不是这样，在系统里就可以证明任何命题都成立。所以“自洽性”等价于要求“系统存在着一个命题，它是不可证明的”。

对角线证明法

我们换一种程序员更加熟悉的证明方式，这里不需要引入哥德尔的对角线引理，但用到了类似的证明思路。即使是我们认为严格的证明，也存在大量没有明确表述出来的推演规则，而我们通常认为这些是默认的真理（共识）而无须证明，它们这是显而易见的。为了明确，我们重新定义一下要用到的名词。

逻辑术语

定义：一个证明系统是自洽的，如果在这个系统中的陈述，S或者其反命题$\sim S$可以被证明，但我们事实上并没有要求这个推导出来的陈

述为真。

定义：一个形式系统是完备的，如果为真的命题可以在这个系统中被证明，我们没有要求所有为真的命题或陈述都能被证明。

哥德尔不完备定理：不存在一个证明形式系统，同时满足自洽和完备。回顾一下我们前面提到的问题，怎样构成一个形式化证明？

一个形式化证明系统*II*包括以下3个要素：

1. 有限的可用于证明的词汇组
2. 有限的公设
3. 有限的推演规则

“足够丰富”，指形式系统可以对任何x的值做出判定。

定义：如果一个命题系统里，*f*可以确定为真命题或假命题，那么这个系统就是我们说的足够丰富。

定义：我们使用  来代表这样一个方程集合——输入任意正整数，通过计算，输出结果是0或1。这样如果一个方程*f*是属于  的，任何一个正整数x输入到*f*（x），都会得到0或1。

定义：可计算。方程*f*是可计算的，也就是说如果可以用计算机语言（python、c语言等等）写出一个有限长的程序，在一个可以执行的计算设备上（比如个人计算机或者云中心）计算*f*。这样，给定任意的正整数x，程序总能在有限时间内运行完毕，给出结果*f*（x）。

定义：  是  中所有可计算方程*f*的集合。

我们很容易证明$\mathbb{A}$中的方程是无穷多的。比如，我们定义第一个方程，这个方程在x=1的时候，f（x）等于1，输入其他数的时候等于0。即：

$$f(1) = 1;\ f(x) = 0,\ x \neq 1$$

我们同样可以定义第二个方程，在x=2的时候等于1，其他输入等于0，即：

$$f(2) = 1;\ f(x) = 0,\ x \neq 2$$

因为自然数有无穷多个，这样定义的方程也会有无穷多个。

我们接下来证明这样的一个定理，即$\mathbb{Q}$里存在不可计算方程，即可计算方程的集合$\mathbb{A}$是方程集合$\mathbb{Q}$的严格子集。即$\mathbb{A}$比$\mathbb{Q}$要小，总有些方程是可以有确定结果，但不是一个可计算方程。

证明

在任何形式系统中，只要某种证明存在，就总有找到对应该证明的算法。由于我们必须假定该系统是以某种符号语言来表达的，而这种语言又是按照符号的某些有限“字母”来表达的。我们把符号串以字典的方式编序。这表示对于固定的串的长度按字母编序，先取所有串长为1的，然后取串长为2的，串长为3的等。这样我们就把所有正确建立起来的证明按照这个字典方案进行编序。我们有了证明的列表，就有了该形式系统里的所有定理的列表。

我们总可以对可计算方程，写出来它的计算步骤，再按照写这些步骤的文档的字母顺序把这些文档排列起来。这样我们就可以首先对集合$\mathbb{A}$里所有的方程排一个顺序。具体来说，我们这样做：写计算

机程序时，我们常用分割符来表示一行的结束。这并不妨碍我们把整个程序当成一个字符串。这样，比如用某一个程序语言编写的程序*Pf*是用来计算可计算方程f所对应的程序，这个程序*Pf*是有限长度的字符串。我们依据长度把所有的表示*Pf*的字符串排起来，当字符串长度相同的时候，我们依据它们的字母顺序来排序。这很像编一本字典，先按照单词的长度排序，单词长度相同的时候，我们依照它们的字母顺序来排。于是我们就有了一本所有可计算程序的字典*L*。每一个可计算程序都会在这个字典*L*中有一个确定的位置。当然也有可能某个可计算方程f可以对应几个不同的程序*Pf*1，*Pf*2 ……，我们只保留在这个序列中的第一个*Pf*，由于方程和程序之间实质上是对应的，我们会看到一个就足够了。一个方程可以有多个方程来表达，我们只选其中一个。*f*可以由多个*Pf*来表达，这在布置计算机课的作业的时可以看到。同一道编程题学生可以交来不同的程序，变量用的符号都可以不一样，但这些程序本身是等效的，我们只保留其中最短的字母排序最靠前的一个来作为列表的程序*Pf* 。

依据这本字典，我们就可以把输入值和程序列在一起，形成一个足够大的表，如表5－2所示。

表5-2

	1	2	3	4	5	·	x	·	i	……
f_1	1	1	0	0	0		$f_1(x)$			
f_2	0	0	1	0	0		$f_2(x)$			
f_3	1	1	0	0	1		$f_3(x)$			
f_4	0	0	1	1	0		$f_4(x)$			
f_5	0	1	0	0	0		$f_5(x)$			

	1	2	3	4	5	·	x	·	i	……
						·				
⋮							·			
								·		
f_i									$f_i(i)$	
⋮										⋱

我们现在可以定义一个方程$\overline{f}$：

$$\overline{f}(i) = 1 - f_i(i)$$

$\overline{f}$也可以把确定的正整数输入输出为确定的0或1，所以它是$\mathbb{Q}$中的一员。依据上表，$\overline{f}(1) = 0$，$\overline{f}(2) = 1$，$\overline{f}(3) = 1$，$\overline{f}(4) = 0$……

那么，$\overline{f}$是一个可计算方程吗？

答案是否定的，$\overline{f}$不是一个可计算方程。

如果f是一个可计算方程，那它应该已经在表5－2里了，因为我们把所有的可计算方程列在了表5－2里。假设$\overline{f}$在第j行，这样$\overline{f}(x) = f_j(x)$，比如是在第107行，则j=107，但是$\overline{f}(107) = 1 - f_{107}(107) \neq f_{107}(107)$，所以$\overline{f}$不是在第107行。同样的，对于$\overline{f}(j) = 1 - f_j(j) \neq f_j(j)$，当输入相同的时候，$\overline{f}$和$f_j$至少有一个输出是不一样的，$\overline{f}$也不在任意一行。因此，在表5－2里面，没有一行对应于$\overline{f}$这个方程，所以$\overline{f}$不是可计算方程$\mathbb{A}$中的一员，但它确实是$\mathbb{Q}$中的一员，因此，$\mathbb{A}$是$\mathbb{Q}$的子集。

这样我们就会看到，确实我们可以定义一个方程，它有确定的含义和输入输出，但是是不可计算的。

现在我们证明哥德尔的第一不完备定理：在系统Π中，存在真实的命题，但不能被证明。

我们再来补充定义我们说“真”和“假”时的含义。

我们把一个陈述称为f 陈述，对于任何一个正整数输入值x，如果它可以表达为“f 是1”或“ f 是0”，每一个f 陈述都是一个关于某一整数的判定。比如 f（107）=1就是一个f 陈述，x的值是107。这样对于任何正整数输入值，f 陈述只有两个值，0或1。因为两个陈述都指的是同一个输入值，我们说第一个是f 陈述Sf，第二个是反f 陈述，$\widetilde{S}f$。

现在我们可以定义“真”了。

我们说如果f 等于1，那么这个关于f 的陈述Sf 是真的，它的反对面 $\widetilde{S}f$ 就是假的。同样，如果f等于0，那么$\widetilde{S}f$是真的，而Sf就是假的。这样，对于任何一个输入值x，陈述Sf和$\widetilde{S}f$总有一个是真的，另外一个是假的。在这个有效证明体系中，不可能证明一个假命题是对的。

形式系统足够丰富：任何f的命题可以被确定陈述。我们这里用到确定陈述，不一定指的是被证明或者被证伪，只是说它应该有个确定的值，即这个形式系统是足够丰富的。我们将来会看到这一点非常重要。

假设系统Π是一个足够丰富的形式证明系统，并且：

A.任何假的命题不能被证明；

B. 任何真的Sf命题，都存在一个属于系统Π的证明序列，来证明Sf是真的。

因为系统Π足够丰富，任何整数x，都会有Sf和$\widetilde{S}f$的陈述，而且只有一个是对的，A和B意味着总有一个有限的证明序列，证明Sf或$\widetilde{S}f$是真的。即这个证明序列本身就可以作为一个确定的程序，在有限的时间内结束，给出一个确定的结果，因此它是可计算的。但这与我们之前证明的f是不可计算的矛盾。具体来说，如果两个假设A和B都是真的，我们可以设计这样一个程序来决定f（x）的值。

具体来说，我们就可以设计一个计算机程序P在这个程序里，$\overline{f}$在有限时间内可以被计算出来。即，对于输入值x，程序P可以在有限时间里输出0或1。我们可以有一个P'，来检测输出值是否符合1－f（x），如果符合，就是一个被证明正确的命题；如果不是，我们同样可以设计一个P'’，来检测否定，这样我们就设计了一个程序，它在有限时间内能够计算f，而且它是一个有限长的程序，但我们刚刚证明了$\overline{f}$是不可计算的，与前面的证明相悖，因此刚才的假设A或B不同时成立。

因此我们说，对于任何一个足够丰富的形式系统Π，如果它自洽的话，总存在正确的陈述在这个体系内不能被证明。类似的说法，在一个自洽的系统里，一定存在对于自然数来说，命题的正与反都不可证。

我们也证明了图灵的不停机问题，即不存在一个超级程序能判定其他程序是否可以在有限时间内结束。

和前一节的讨论有类似的推论，我们一样有了哥德尔第二不完备定理：



图5–2 一个系统自洽性的证明需要引入另外的假设，但这个新引入的假设一样需要新的假设来证明其自洽性

如果一个系统是足够丰富和自洽的，它的自洽性不能在系统内部被证明。如果要证明系统Π是自洽的，需要引入另外的一个形式证明Π'，但Π'的自洽性也不能在它内部证明。

第六章

哥德尔问题的延伸

真理存在，真理存续。哥德尔不完备定理解释了为什么我们总需要感性部分的存在，总有些直觉的认知无法被理性取代。这些感性的认知成就了我们人之所以为人的重要性。

《数学原理》是这一打击的第一个牺牲品，但绝不是最后一个。哥德尔那篇论文的标题中的“及有关系统”这几个字，蕴含着丰富的潜在内容。简而言之，哥德尔展示了：无论涉及什么公理系统，可证性总是比真理性弱。

——侯世达（Douglas R. Hofstadter）

哥德尔的不完备定理有几种等价的数学说法：

1. 任何自洽的数学形式系统都包含不可判定命题。

2. 没有数学形式系统既是自洽的又是完备的。

3. 没有定理证明机器（或机器程序）能够证明数学中所有的真命题。

4. 数学是算法上（机械上）不可穷尽的。

希尔伯特和罗素和他们之前的几代人希望把数学搞成一个形式化系统，希望所有的猜想都能从逻辑出发加以判定，希望所有事情迟早在“意料之中”等等，这些希望最终由哥德尔判定为数学家的痴心妄想。数学的神奇之处在于它扎根于人类的观察、直觉和灵感，但数学不等于逻辑，数学成果也并不总是能从公设开始通过逻辑推演出来。哥德尔不完备定理同时表明，数学存在无穷多个有确切答案的问题，不能利用任何公设化程序找到它们的解；它们并不，也不能对应于任何形式系统中的定理。这些问题的答案是不可计算的，而且也不能归结为其他数学事实。这里推广爱因斯坦关于上帝掷骰子的名言：上帝恐怕不仅仅是在量子力学中玩骰子，他甚至在数学中也玩了骰子。这

对数学家而言是个可悲的结论，他们努力了一生的某个数学猜想的证明可能根本不存在，但也可能是无穷美好的结论，它把实验物理学家对数学家的“纯想象的”鄙视消除了，数学也是发现自然潜在规律的工具，它联结了人脑的直觉和自然规律。

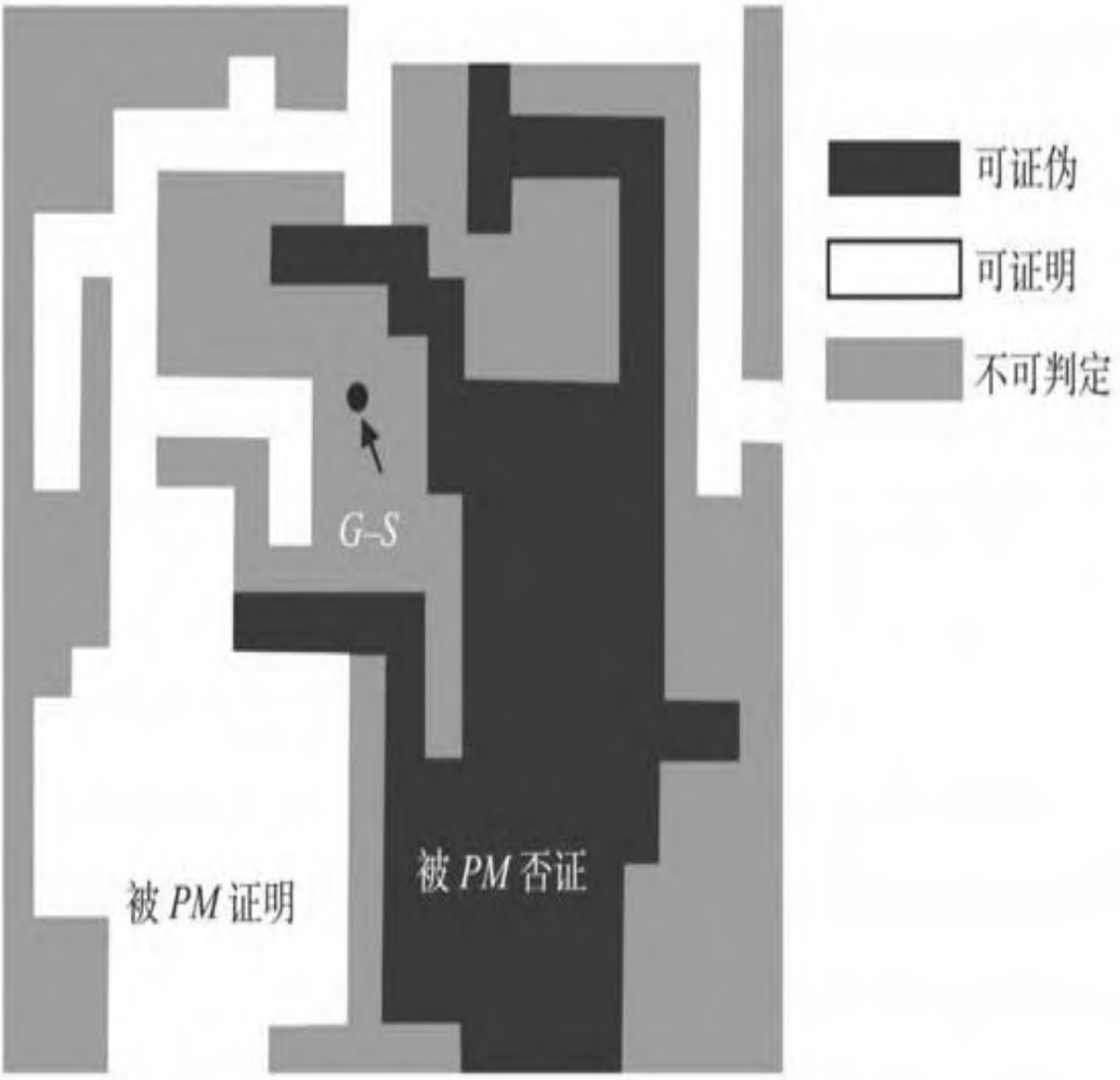


图6–1 知识空间中的哥德尔不完备定理

那么能不能证明一个更普遍的不完备定理？把哥德尔不完备定理推广到所有通常的事物里？对此，哥德尔曾经做出一个他认为合理的表述：一个完美规划的社会（即处处按完美的法则行事的社会），就其行为而言或者是不自洽的，或者是不完备的。即这个社会一定有一些问题是无解的，这些问题有可能极其重要，甚至会危及整个社会。这个说法也适用于个体的人，事事讲究原则的人一定有无法讲原则之处。

真理不能被语言描述

维特根斯坦（Ludwig Wittgenstein）研究语言得出结论：语言不能把握世界上存在的全部东西。“只可意会，不可言传”，哥德尔的工作给出了这句话的数学形式。从本质上说，哥德尔所证明的是，无论任何语言，即使是数学语言，都不足以彻底、完全地表达日常的“真”的概念。用语言学的术语说，就是语法无论多么丰富都不能完全取代语义。即使在纯数学这样内容简洁的领域，形式逻辑的操作也永远不能告诉我们数学的每一个特性。而我们必须记住算术是关于数本身的，而不仅仅是关于算术规则的。

用语言来完成人类彼此间的沟通，除了需要共同的认知基础，也不能脱离人对外部世界的感知。这些感知包括视觉对外部的成像，听觉对声音的辨识，以及人对历史和环境的认识。否则语言就会被无穷诠释，变成无限矫情的胡搅蛮缠的文字游戏。但要注意的是，这种胡搅蛮缠是逻辑上允许的，并不是我们所说的无理取闹，它事实上就是一个可以无限延伸下去的合法讨论。

一个办法是把人类的自然语言抽象为“信息”，这时我们稍微严肃认真地讨论一下“信息”这个概念，这要从香农（Claude

Shannon）的信息论开始。人们利用有限数据来传达信息，首先追求的不是高效的传输，而是这些数据对所传输信息的可靠表达。

1948年贝尔实验室的香农提出信息论来刻画信息通信过程中的数学模型。在信息传递的基本过程中，甲首先把信息以特定的规则编码，这在物理设备上比较容易实施；针对有很多噪声的信息通道，可以加些冗余编码来降低传输过程中产生的错误；收到代码序列之后乙再通过特定规则来解码，这样乙就得到了这条消息。

在这个通信过程之中有两个基本的假设：

假设1 双方共享一个编解码的规则，即共同的密码本。

假设2 双方有共享或公有的常识。

香农的通信理论关心密码本的建立，首先要回答这几个问题：

1. 甲要想一下：乙是否与甲有一个共同的认知模型？否则，即使完成了解码，乙也有可能因为不能领会里面的内容而产生误解。那么发信息的时候，甲的措辞要尽量减少可能的误解。

2. 甲还应该要想一下：为什么要发这个信息？乙是不是已经知道了？乙是否关注这个信息？乙爱不爱听？听后有什么反应？这一句话说出去有什么后果？

3. 乙要想一下：我为什么要收这个信息？甲发给我是什么意图？

我们脑海中有关于某个事情的概念，我们便可以毫不费力地思考它，通过直觉来判断它的对错，至少是对我们自己大脑内部的对错，但我们可能不能简单地在大脑里通过逻辑推理和语言来描述它。近代由于牛顿力学的发展让人产生了一种经典意义上的关于思维和语言关系的观点：人类通过语言来思考，但我们现在知道这样的观点混淆了语言和语义之间的差别。我们大脑可以把握语义，但当我们要用语言

的规则把它表述出来的时候，事实上，我们不得不承认我们忽略了很多概念，很多思维的细节，或者本来延伸着的语义的内容。但我们也不担心这一点，我们自然而然地、天真地认为我们的听众能够理解我们的言下之意，而事实上他们也真的可以做到这一点。这真是一个神奇的过程。

英国人喜欢茶歇，喝茶的时候可以海阔天空地聊天。我即使到现在也不能完全融入他们所有的闲聊中。语言的交错中，我可以完全置身事外。不过问题是一样的，当我们几个中国人在一起说到“林黛玉”时，对我们而言，那是一个非常丰富的文化符号，我们都明白她的身世、情感、性格和价值观，这就轮到其他很多不同文化背景的人置身事外了。

我们的大脑里有了这样一个概念：林妹妹。我们通过语言的编码机制，喉舌和口腔，来振动空气，把我们希望表达的信息编码成一个有振幅和频率振动的序列。这个振动的序列，可以通过空气，也可以通过电波，传到接收的一端。通过耳朵里的鼓膜，再转化为听骨的振动，这些振动牵动了对应的肌肉和连接着的神经，通过某种神奇的翻译机制，转化为我们所理解的那个“林妹妹”。这个神奇的翻译机制，也许不仅仅是大数据，而是某种尚未完全认知的模式组织起来的公有知识或者数据库，即我们所说的共识。同样的一个振动和频率的序列，在一个不懂中文，或者只能勉强用基本中文交流的人那里，就不会产生共鸣。这似乎和这个人所掌握的知识，至少是语言方面的知识相关，但我们一样有一种神奇的检索信息的截断能力，知道这些知识应该在哪里被终止引申，到哪里就已经足够了。这似乎又可以与下一章将要谈到的图灵不停机问题对照起来，在某种意义上，“我们”可以通过“直觉”适可而止地知道哪里是我们在逻辑之外的满足，而逻辑本身并不能提供这种满足，人可以知道怎样获得沟通意义上的满足。但它也不总灵，否则的话就不会有误解存在。这些感受是在认知层面的，是在编码之外的。因此，通信理论就只管发送，就像以前电

报大楼的发报员，收钱发报，他们丝毫不需要在乎发报的动机、内容和后果。

相比照，动物之间也一样有丰富的交流的方式，除了发声，它们很多还借助于肢体和触角的动作。人的对话也不一定通过说话，手语、眼神和肢体同样可以传递很多信息。在有声语言产生之前，人类就已经有了十分丰富的认知基础。没有这样的认知基础，语言是空洞的符号，沟通交流和唠嗑也不能发生。

人们因此有一种沟通上的困境：到底要多少数据才能准确地表达我们真实的想法而不产生歧义？怎么说话才不会引起误解？到现在，读者应该可以理解，一个没有歧义的语言系统可能是不存在的。但我们人类有办法来弥补这一点，我们总会在效率和可靠性之间进行平衡。语言的冗余部分看似是废话，但它们的存在降低了误会的发生，提高了沟通的容错能力。如果一句话说不清楚，没事，我们可以多说两句。

我在北大前后待了十一年。初入清华，跟同事约见面，“下午2点东门见！”于是到了下午2点，我沿着清华东墙从最南边走到最北边，都没看到东门！幸亏有手机这一现代工具，我终于跟同事见上面，他说：“清华东门，在南边，南边墙上靠东的那个门，对着主楼。”，我回道：“我们北大的逻辑不是这样的啊！南门一定在南墙上啊！”这就是不同的生活习惯造成对同样的名词含义理解的不一致。再者，下午2点，是北京时间还是格林尼治时间？约定的时候我们并没有说明。而事实上我们说的北京时间其实也不是北京的时间，北京时间是陕西天文台发布的陕西省西安市临潼区国家授时中心当地经度的时间。下午2点是14:00还是14:00:00？二者在物理上是有严格的区别的，而这很大程度上依赖我们所用的交通工具，如果走路的话14:00差几分钟并不是个大问题。但如果开车，校门口是不让停车的，所以在车流中你被后面的车押送着，经过东门是不允许有分钟级的误差的，所以要14:00:00。这样的补充信息可以一直延伸下去。当我们每延伸

一个新的解释的时候，它还是会牵涉进来更多的解释。人却有一种神奇的能力，能够不在这些定义被穷尽之前就达到沟通的目的。而哥德尔告诉我们，这些定义是无法被穷尽的。人类能够沟通，是个奇迹。

有一天晚上跟朋友聊起来哥德尔不完备定理。朋友指着路边的一辆车说：“这是一辆车。这也会有歧义吗？”如果需要准确理解这句话，给出完整的描述，我们不得不描述环境。我们俩都在路边站着，手指着同一辆车。说话的两人大脑有众多的不可一一列举的共识。首先我们沟通的环境不是在两个不同的地方，也不是指一幅画上的车，也不是一个玩具车。记得那个著名的“人无法踏入同一条河流”的说法吗？比如说我们指一幅画，画里有树、有车也有房子，这句话中的“这”未必是我们头脑中想说的那辆车。所以环境本身需要描述，即我们说到“这是辆车”这几个字的时候，应该有描述说两个人在同一地方。这种共识能不能被大数据描述，是一个有意思的问题。哥德尔不完备定理可能导致我们不能够完全描述我们的共识。为了弥补一个概念的不足，我们不得不引入新的概念。很快又要去弥补这个新概念的不足，哥德尔第二不完备定理告诉我们，对概念的详尽解释是做不完的。不是我们在做刨根问底的无聊事，而是这样无聊的事情确实存在而且逻辑正确。大脑逻辑上不通却能够神奇地采用忽略多余信息的方式让我们对别人说的话进行正常理解，适当的停机而不会陷入刨根问底的“宕机”状态。

真理不能够自证其真

哥德尔证明揭示了我们传统意义上的逻辑证明方法可能存在深刻的问题，因此也质疑了我们奉若神明的理性。

数学里我们会用到反证法。它的前提是如果我们能够证明一个命题的反面是矛盾的，那么这个命题就应该是成立的。

例如：用反证法证明一个三角形中不能有两个角是直角？

假设三角形中存在至少两个角是直角，设∠A和∠B为直角，则三角形三角之和为：

S=∠A+∠B+∠C=90°+90°+∠C=180°+∠C，

因∠C≠0，故180°+∠C＞180°，

而三角形内角和为180°，故三角形中不能有两个角是直角。

数学上一样大量使用了反证法，假设一个命题正确，那么推出来一个跟已知定理矛盾的结论，这样我们就知道假设是错误的，这个命题不正确。但如果命题本身不能被证明正确或错误呢？而这种判定如果不那么直接呢？比如我们试图证明：

不存在对这句话的证明。

我们首先假设存在一个证明，可以证明这句话是对的；那么不存在对这句话的证明，这与我们的假设相矛盾。

如果不存在对这句话的证明，就是对这句话的证明，但它“不存在”。

既然至少存在这样一个命题，我们无法证明不存在类似的命题，“理发师悖论”就是另一个。理性逻辑的严谨在于，一个证据就足以推翻整个理论大厦。这类陈述让我们很担心反证法的可靠性。因为真的存在一类命题不是非此即彼的，它的正面命题和反面命题可能都无法证明其对错。

看来反证法不可靠，那么我们通常的正面证明的方式呢？

正面证明通常采用归纳法和演绎法。归纳法里只有自然归纳法是严格意义上的证明，而其他归纳法因为不能列出和验证所有的可能，所以不能构成严格的证明。

对于一般性公式的证明所用的自然归纳法，我们必须首先证明公式在n=0的情况下成立。接下来我们假设它对任意的但又是确定的正整数n成立。最终，我们使用此假设以演绎方法逻辑推导出它对于n+1成立。因此，我们证明了，如果它对n=0成立，那么它对于n=1成立。如果它对n=2成立，那么它对n=3也成立，如此等等。事实上，这个基本算术公式的所有证明都以这种或那种方式使用了归纳论证。这种数学归纳论证尽管不是形式逻辑的推理工具，但广泛应用在数学论证中，它可以使我们从有穷的条件集（“n=0”和“n=任意但确定的数”这两个条件），推导出适用于无穷个情况（所有可能的正整数）的结果。自然归纳法也需要假设：规律在无穷大的时候依然有效。我们已经有了不少证据说明这个假设是有问题的，比如我们将要讲到的希尔伯特的旅馆。当真要考虑无限的时候，有限小的规律与无限大的规律未必一致，比如求和1+1－1+1……，它的结果到底是0还是1呢？当这个序列有限长的时候，它或者是0或者是1是确定的，但当这个序列无限长呢？

对于演绎来说，我们为了谨慎小心，证明陈述判断词的两端是严格等价的。三角形内角和为180°，其中三角形的内角和跟180°本身是同一事物，这无法避免地成为自指。而我们刚刚知道了逻辑证明中自指是可怕的，它会带来悖论！

“这句话是错的”，问题在于句子自己指向了自己。但演绎法正是这样啊！而我们其实无法以一个稳定可靠而又普遍的办法来区分什么样的自指是可以的，什么样的是不可以的。

因此，作为真理的代表，数学的证明逻辑，似乎也出了大问题。

自洽系统不能自证其自洽

无矛盾的系统不能自证其无矛盾。数学中最简单、最基础的领域，是以自然数构成的算术，它是雄伟的数学体系的基石。然而从数学本身所依赖的形式化公理系统的角度来看它竟然是不完备的，更糟糕的是，这样的系统不可能通过不断增补而变得完备。

事实上，计算机只能以程序设计师所输入的公设为基础来证明定理——正如哥德尔所强调，它无法自行创造新的公设（假设）。从这里可以推出，任何计算机或者计算机系统，联在互联网里的计算机，比如说无限台，都不能掌握或者推演出有关算术的所有的真理。

更进一步来看，形式化算术的不完备性，不但可以由能思考且有数学直观的人推演出来，连计算机也可以证明出哥德尔不完备定理。这个定理指出了计算机本质上的能力极限。算术的真理在原则上无法被一个形式化系统限制。

“真”和“对‘真’的证明”之间具有相当的差异。

我们在此讨论的数学证明，总是在给定的形式化系统中的证明，而“真”却是独立于形式化系统的。哥德尔证明，数学的“真理”无法约化为证明，而语法也无法取代语义，机械性的规则无法排除对意义的需要。

哥德尔的第二不完备定理更具有讽刺的味道。一个算术系统的公设系统如果是自洽的，这种自洽无法在该系统中被证明，而这种不自洽，不能通过增补假设来完成。它的等价陈述就是，只有不自洽的系统，才能证明它自己的自洽性。而，冯·诺依曼听完了哥德尔的报告，立刻意识到这一点，并兴奋地跟哥德尔讨论，哥德尔只平静地说他关于这一点的文章已寄给了出版社。冯·诺依曼对哥德尔的发现在一开始就有惊人的领悟，他说：“哥德尔告诉我，我们无法通过逻辑

的方法获得领悟数学的能力。”要知道他是“现代计算机之父”，曼哈顿计划的主要领导者之一。

普林斯顿高等研究院的第一批终身研究员之一外尔称哥德尔的工作为“哥德尔灾难”（Gödel Catastrophe），具有两千多年历史的亚里士多德的理性认知，尤其以欧几里得几何所开启的公设化的想法，人们认为的理性典范被动摇了。这一切又发生在以希尔伯特为首的数学家们开始着手实现彻底公理化时，即要把数学形式化系统本身视为数学的研究目标，哥德尔比希尔伯特做得更加彻底，但得出了完全相反的结论。

我们要理解哥德尔不完备定理的重要性，不仅要跨越数学障碍，恐怕也要跨越哲学障碍。

后山的龙

北欧的传说中有这样一个故事。有个村子的后山有条恶龙，每个青年男子在他的成人礼之后都会去后山杀龙，但从来没有人回来。有一对青年恋人，男孩去杀龙，女孩偷偷跟在后面。女孩看到男孩杀死了龙，沾上了龙血，变成了那条龙。

哥德尔并没有发现一个复杂神秘的不被任何形式化系统包容的定理，他发现了一个任何形式系统都普遍存在的定理。一旦这个形式化系统有足够的表达能力，就一定存在至少一个直觉上正确却无法在给定的系统中证明的陈述。这个陈述可以在一个以给定形式系统扩展的形式系统内证明，但一定又会有一个陈述，在这个扩展版的形式系统内还是不能被证明，一个广适于所有系统的超级英雄并不存在，而一个无处不在、无法去除的普通存在却存在于任何一个系统中。

因此我们也明白：争论是没有意义的。

学术研究和日常生活中，我们经常陷入对这样或那样问题的争论，但我们往往会忽略讨论的前提。从维也纳学派发展出来的务实主义道路，要求讨论的前提是大家对相同的公设系统有共识。第一，我们在讨论前要有相同的公设；第二，在讨论中要遵从相同的推演规则，即逻辑。

在科学历史上人们意识到这一点并不容易，对真理的捍卫常常是一种堂吉诃德式大义凛然的英雄气概，甚至是以一个伟大科学家的生命为代价。

19世纪末，玻尔兹曼（Ludwig Boltzmann）与奥斯特瓦尔德（Friedrich Ostwald）之间发生了“原子论”和“唯能论”的争论，即物质世界到底是由原子组成的还是由能量组成的。很多年后，普朗克（Max Planck）回忆起来这场论战说：“两个死对头都才华横溢，针锋相对，应答如流。”双方各有自己的支持者。奥斯特瓦尔德的支持者是以不承认有原子存在的物理学家恩斯特·马赫（Ernst Mach）为代表的德国主流学派。作为少数派的玻尔兹曼常年沉浸在与不同见解的辩论中，这极大地损害了他的心理健康。虽然玻尔兹曼最终取胜，包括普朗克也承认自己对黑体辐射的解释借鉴了玻尔兹曼的原子论，但他心理上的痛苦与日俱增，毫无办法解脱。终于在1906年，他以自杀的方式结束自己疲倦已久的心。对于他的死，普朗克感慨地说：“新的科学真理不能通过说服对手而建立。只能等到对手们渐渐死亡，新生代开始熟悉真理时才能贯彻。”对普朗克来说，辩论没有多少诱惑力，辩论不能产生任何新东西，没有实验检验的争论没有实际意义。人是可以接受新知识的，但如果已经有了固执的己见，通过辩论来驳倒对方是不会奏效的。

这件事情，对后来的物理学产生了深刻的影响，20世纪，科学从此努力跟哲学撇清了关系。在没有实验的基础上，我们甚至无法确定该使用哪些公设，也无从讨论在某一公设的基础上哪些结论是对的，更无从讨论该用生命去捍卫哪些真理。

整个20世纪都是对前面几千年人类理性认知系统的总结和颠覆，我们刚刚开始掌握和认识世界，认识我们的工具。我们逐渐开始明确了该使用怎样的方法来认识世界，也知道了这些方法的局限性。我们不得不通过这些方法认识世界的同时，也应该时时提防这些方法成为我们的枷锁。人天然有回避矛盾的心态、能力和期望，这就是为什么理性逻辑诞生了两千多年之后才有哥德尔来指出我们的理性逻辑本身是有些缺陷的，而这个缺陷又是如此明显。

从哥德尔不完备定理导出的一个直接结论在于真理不能被描述。从哥德尔第二不完备定理中我们可以看到：对于一个真理系统，它本身并不能够证明自己的正确性。很多人希望通过说明物理世界和数学世界的不同来区分甚至回避这一问题在实际应用中的危机。物理的规律并不是我们可以任意规定的，而物理世界是自然完备的，因为所有的规律和现象都不得不被装在宇宙这一有边界的封闭体系里。如果是这样，我们反过来只能承认这些规律之间必然存在着某种不和谐、不自洽和相互矛盾。为了解释这些矛盾，我们需要发现新的证据。宇宙这么大，人心这么大，看起来这并不是个问题，所以我们不如承认不自洽的天然存在。很多领域的研究者也并不在乎哥德尔不完备的存在，因为忽略这些漏洞似乎也并不会影响我们日常的研究。我们发现世界上的新事情还远远没有被干完，那么为什么我们要去着急发现这些令人困惑的这样那样的矛盾呢？

但哥德尔第二不完备定理告诉我们，既然一个体系在其内部不能证明其自洽，就需要补充新的证据或者公设来扩充这个公设体系，以保证原有的体系是自洽的。这同样带来了新的困惑，扩大之后的公设体系同样会具有相同的问题，它不能够证明自己是自洽的。这样我们就不得不承认我们能够用有限逻辑描述的知识体系本身是有缺陷的，是需要不断扩充的。而进行这种扩充的一个实际有效的做法，反而是我们谜一样的直觉，回归到对世界认知的常识。

观察者的观察者

哥德尔第二不完备定理同时暗示了另外一件事情。当我们有一个超级用户的时候，他来管理所有人的权限，但他的权限由谁来管理呢？比如我们买了一台电脑，我可以把自己设成administrator——超级用户。超级用户可以设置不同级别的用户以管理他们的权限，有常规用户，可以修改简单设置的，也有访客用户，所以用电脑上上网的，但超级用户的权限是谁给的呢？厂商卖给客户电脑，这个权限是电脑制造商给予客户的。但这并不代表客户能获得计算机的所有权限，很多底层的管理权限客户并不知道，大多数时候也不关心，只有厂商自己知道。因为维修的时候可能需要一些权限来保证机器可以恢复到出厂设置之后还能被调用，很多手机就有这样的配置。即使你是这台设备的拥有者，为它付了钱，但权限未必能够开放给你，有时候是为了维修安全，有时候是为了别的目的。厂商就成了这个逻辑里的超级超级用户（super-admnistrator）。这些厂家，在政府的要求下，也会让政府去调用一些数据，甚至开放一些功能给政府用，比如2017年发生的苹果公司与美国政府之间的矛盾。美国政府要求苹果公司开放用户数据给FBI（美国联邦调查局），苹果公司拒绝了。但我们假设苹果公司在美国政府的压力下开放了数据，那么在这件事情上FBI和它后面的美国政府就成了超级超级超级用户（super-super-administrator），而谁为美国政府这个超级超级超级用户授权呢？至少美国政府认为这是宪法授权的，但我们马上就可以看到，原来连美国宪法也并不能做到自洽和完备。于是，问题回到了原点，要不我们存在一个一直在扩充的系统，每一个新的超级系统囊括并解释下一级系统。而在这个系统的顶端，存在着一个不自洽或不完备的公理系统。

哥德尔不完备定理是理性的边界吗？

各种证据似乎让我们得出这样的结论，哥德尔不完备定理规定了理性认知的边界。或者这样说：它虽然不是理性的边界，但至少规定了理性的局限，但限制在局部本身就意味着边界的存在。

什么是理性思维的界限？我们能够完全理解我们制造的机器吗？我们能够搞清楚我们心智的内在工作过程吗？当研究结果缺乏逻辑的确定性时，数学家还怎么继续工作呢？哥德尔的工作不仅使数学发生了革命性变化，而且还波及哲学、语言学和计算机科学，甚至还包括物理学。哥德尔的这一结论——存在着明明为真，但却必然不能被证明的事实，集齐了从神经网络到计算理论的发现和创新浪潮，也终结了机器全知全能的梦想，昭示人类之精神的永不枯竭。

因为早年在维也纳的优秀表现，哥德尔被纳粹分子怀疑是犹太人，莫名其妙地在街头挨了打。作为一个有轻度抑郁的数学家，遭受了这样的事故，哥德尔很早就决定离开奥地利，这个很快就被纳粹独裁统治的国家。到了普林斯顿不久，他决定加入美国籍。不幸的是，哥德尔以他特有的数学思辨能力仔细研究了美国宪法！入籍面试前一天，哥德尔忧心忡忡地打电话给他的入籍见证人，博弈论创始人摩根斯坦，他告诉摩根斯坦，美国宪法中有一个逻辑漏洞。因为这个漏洞，美国宪法一样会使美国变成一个专制独裁的国家。摩根斯坦只好叮嘱哥德尔，见到法官时千万不要提这件事。

第二天一早，爱因斯坦、摩根斯坦和哥德尔一道驱车去新泽西州首府特伦顿的联邦法院，入籍面试在那儿进行。一路上爱因斯坦和摩根斯坦轮流讲笑话，好让哥德尔的注意力从美国宪法的逻辑漏洞上转移开。在面试中，法官被哥德尔带来的两位著名见证人震惊了，甚至打破常规请他们在面试过程中一直坐着。

法官问哥德尔：“到目前为止，你一直是德国人，是吗？”

哥德尔答：“我是奥地利人。”

法官不无得意地说：“不管怎么说，奥地利也曾处于罪恶的专制独裁统治下。不过现在你应该感到幸运，在这里有美国宪法保护，这种事情在你的新国家永远不会发生！”

这话无疑让爱因斯坦和摩根斯坦一路的努力化为泡影，哥德尔大声说：“恰恰相反，我知道这怎样可以让这种情况发生。我现在就可以证明给你看！”

美国宪法的不完备

1972年“水门事件”发生。

尼克松的竞选团队去对手总部安装窃听器被路过的警察无意抓获，事件不断升级，以至于总统都无法摆脱干系。审讯过程中，联邦最高法院要求尼克松交出白宫的录音磁带作为证据。然而美国宪法第二条第一款规定，行政权属于美利坚合众国总统，三权分立的宪政原则规定总统的行政特权不容侵犯。作为政府行政首脑，总统与自己的亲信在白宫的非公开讨论会中信口开河、胡说八道都可以被认为是工作需要。总统与亲信之间的谈话涉及内政外交、国防军事等方面的机密，如果司法和立法部门传唤、审听这种高度机密的非正式谈话录音，并根据其中的只言片语来起诉总统，是完全不合常理的。在美国宪政史上，杰弗逊、林肯、罗斯福、杜鲁门等十余位总统都曾以总统行政特权为由坚决拒绝国会调查白宫决策内幕。尼克松更以不按规则出牌而著称，直接查封了检察官办公室。

主审此案的哥伦比亚特区大法官约翰·西里卡（John Sirica）立刻表示，查封特别检察官办公室之举“看上去仿佛是拉丁美洲国家的上校们上演的一场军事政变闹剧”。联邦司法部长埃利奥特·理查森（Elliot Richardson）也惊呼：“一个合众国政府快沦为了寡头独裁政府。”针对尼克松的一意孤行，美国

新闻媒体口诛笔伐，参众两院议员怒不可遏，普通民众也群情激愤。无数电报、电话和信件如洪水般涌至白宫和国会，强烈谴责尼克松的胡作非为，要求国会立即启动宪法程序，弹劾无法无天的“总统皇帝”。这样，“水门事件”从最初的一桩鸡毛蒜皮的丑事，最终演变为一场震撼全国的宪法危机。当三权分立制度下的三个最高权力之间发生了矛盾，该听谁的？谁有最高的管辖权？这在美国宪法中竟然没有规定！

尼克松最终选择了辞职。

他说：“一个总统如果动用美国的军队来保护自己抗拒法律的审查，他不再配做美国的总统。”尼克松1977年接受英国著名电视节目主持人戴维·弗罗斯特（David Frost）采访时说：“事实上是我自己弹劾了自己。”他的辞职换取了民众对他的谅解，化解了美国宪法的危机，从而维护了美国宪法的尊严，并且也给出了案例。当克林顿再次被法院要求提供证据的时候，援例。

因此美国的大法官制有制度上的好处。美国联邦最高法院拥有解释宪法的权利，他们做出一个判决都会形成判例，进而对原来的宪法进行补充。美国联邦最高法院通过不断细化规定，丰富了宪法的含义。大法官制度补充在逻辑规则之外的案例，包括宪法在内的严密的有限条文体系，依然无法脱离作为人在其中的参与和抉择。

与英美法系的案例法不同，欧洲大陆有大陆法系，明文规定哪些事情可以做，哪些事情不可以做以及做了僭越的事情之后的代价。这样的思想沿袭了经典科学的习惯。它假设人的行为可以被规范，可以被法律条文一条一条清楚地描述。这里的潜在认识是默认人的行为可以被静止的、孤立的文字来规范。我们希望有这样的一套可以排除个人主观情绪和观点的法律体系。这当然好，然而300多年来，我们仍旧不能把人从这个法律体系中排除出去，一定是有它的原因的。哥德尔不完备定理说明了，由有限文字所记录的法典和规范，一定有不能描述和判断的问题出现，这些问题就不得不又回归到人本身，由人来做

最终的裁决。一个能够排除人的存在的客观行为规范并不存在，很多时候我们认为公认的客观依据，是因为太多的知识、信仰和道德约束对我们来说成了默认的共识，无须进一步沟通。参与到其中的人既弥补诠释有限体系的自洽问题的不足，又补充有限知识体系的常识，这工作像极了“补锅匠”。相比较，判例法系不强求排除人为因素，判例里充满了法官和陪审团的人的气息，看上去依赖于法官和陪审团的素质而充斥着主观判断，但事实上可能是一个更为合理的系统。要知道，在哥德尔不完备定理的陈述中“真”和“可证”并不等价。人可以通过直觉获得“真”的判断，但它未必能够从逻辑中被证明。有些事我们明知如此，但无法证明！

那么，我们自然会问一个问题，而这个问题事实上会动摇我们对法律体系的信心。既然还是避免不了最终的人治，那我们辛辛苦苦建立法律体系的意义何在？或者，我们只能把它当作人性存在的天然弱点，或者我们只能追求程序的正义，而无法追求含义本身的正义，至少法律条文本身无法既追求含义的正确，又追求程序的正义。语法无法取代语义，对法律的诠释最终回到人的认识和良知，而良知的建立是基于社会、文化，又常常体现为宗教的教育和影响。既然我们既无法排除人为的参与，又在人治中得到了太多的教训，恐怕能做的只能是“上帝的归上帝，恺撒的归恺撒”，完成程序正义的同时，通过合理的制度来选择“补锅匠”并制约他们的权利，使得他们所做的最后的决策和诠释符合公众的长远利益。

延伸到我们对科学的认知，我们用语言和定理来描述自然界，希望这些定理可以客观地描述和预测自然界，甚至预测自然界可以接受我们的改变。但当一个人忽略了自然的世界和人自身的体验之间的关联时，这个人就容易让自然形成一幅与人类利益无关的事物的画面，并把我们对自然的认知看成固定的和孤立的东西，这很容易成为一个压抑心灵和麻痹思想的根源。在对人类未来的预言中，这种出发点把人类引入荒诞的时间尽头。人甚至把这种从认知得到的结论看作一个缺乏自省的工具，一个灭亡人类的工具。

牛津大学哲学家戴维·卢卡斯（David Lucas）于1961年发表一个看法，认为哥德尔不完备定理和思维本质之间存在关联：“在我看来，哥德尔的定理证明的是机械论的错误，也就是说，机器不能解释心智。”卢卡斯坚定地认为，不管我们设计的“思维机器”多么复杂，这台机器都将根据规则运行，而这些规则可以在形式系统中被表述，并且一旦我们让这台机器告诉我们什么是真命题，它将只能通过系统的公设基础并按照规则来推演这个命题。因此将会有一个我们的思维能够领悟为真的命题逃脱这台机器对真理的把握，而它的把握无非就是规则决定的可证性。如果我们将前面逃脱的命题列为公设，以强化这台机器，将仍会有其他命题逃脱它，但不会逃脱我们。这些命题有着合理数学的正确性，但符合证明规则的形式系统永远不足以产生算术的所有真命题，就算是在原则上也做不到。这个结论无疑肯定了哥德尔工作的卓越，但它还有另外强有力的一面，表明了更多的东西，确保人类的理解和洞察不能被还原成任何规则的集合。人类思维必然超出逻辑设备所能实现的东西，“人工智能”是我们今天对这台机器的时髦叫法。

哲学家曾经希望的终极目标是找到一个整体自洽的图景，在这个图景里可以整合人类所有的知识经验。我们内心希望的是这个图景中永远存在某种缺口，因为只有这样的缺口不断存在，才能体现出人类存在的意义。如果完整体系是存在的，它终有一天会被发现这样的缺口，一旦发现，我们将面临一个艰难的环境，因为我们总可以把这些完美的无所不包容的知识体系整理成可以去具体实施的系统工具。这样就可以去做所有发现的事情，在一个知识的封闭体系中，终究有一天所有事情都会被做完，况且我们发现了一些越做越快的工具，于是进入了另外一种形式的热寂。整个宇宙都没有新的东西被发现，这是一件悲哀的事情。

哥德尔证明数学世界是取之不尽，用之不竭的，没有一套有限的公设和推理规则可以包含整个数学范畴。给定任何一组公设，我们可以找到有意义的数学问题，这些问题是公设和它们的合规推理都无法

解决和判定的。我们希望物质世界中存在类似的情况，如果我们对未来的看法是正确的，那就意味着物理世界也是取之不尽用之不竭的；无论我们走向多远的未来，总会有新事物发生，新信息涌入，探索新世界，不断扩大人类的生活领域。哥德尔的理论最终会引发人们研究数学和物理世界的无法预料的本性和新热情。不完备定理应该成为我们寻求未知世界的新的自然法则：我们没有理由限制自己对自然基础的探索，可能也不能用任何方式说明我们找到了描述自然对称性的普遍工具。相反，我们确实希望发现哥德尔不完备定理限制了我们随心所欲的能力，这事实上也保护了人类自己，因为一旦有可以把我们“人”移除出去自动寻找真理的算法，那我们人类自己也就可以真的被移除出去了。我们可以做出机器来执行这些算法，自动地寻找我们未知的全部“知识”。幸而哥德尔不完备定理否定了这一点，我们不能用我们手头已经掌握的定律来执行任何特定计算或构建算法来预测未来，因为我们甚至不知道未来会有哪些新的定律出现，不完备性困扰着哪怕非常简单的自然法则。

宗教与科学并非是处于互相矛盾、水火不容的境地。至少在科学革命之前，科学与宗教是和平共处的。中世纪，罗马教廷对各种对《圣经》的不同解释相当大度，而从事科学研究的人也相信这些研究都是在证明上帝造物的伟大。然而在19世纪到20世纪之间，科学的发展让文本意义上的《圣经》显得格外局促，科学每一点新的发展似乎都挤压着《圣经》的神圣领地。除了进化论，物理学看起来也不那么友善。热力学第二定律所预示的宇宙可以走向无序化，这就让相信上帝创造了理想世界的人非常不安。正在膨胀的宇宙甚至宇宙大爆炸，都让上帝的信仰者不得不寻求新的解释，试图把新的科学发展包容到绝对正确的伟大真理体系中，但事实上人们发现这越来越困难。而今，我们要面临着的本质冲突在于，宗教与科学依赖着彼此不相容的逻辑原则。神的行为不受科学理性框架的约束，不论它作用在自然界还是作用在人心上。信仰宗教的人认为宗教回答了科学所不能回答的问题。的确，客观唯物主义的科学观并不意味着我们能够理解一切，

事实上，至少我们知道依靠经典理性是做不到这一点的。虽然我们相信可能是工具的阶段问题，随着技术进步我们迟早能够了解到更多的部分。但哥德尔不完备定理和海森堡不确定性原理都似乎给我们规定了认知的边界。一个是从描述的工具数学的角度，一个是从这些工具对应于自然界的物理学角度。它们设定的这些边界，不是人们通过提高实验手段和认识手段就能够了解的，它就是自然的一部分，这部分未知是我们已知的。这样说颇为哥德尔化。

基督教是可以比较严格地形式化的：

- 上帝存在
- 上帝万能
- 上帝爱人

这三条构成了基督教的三个基本公设。

但在这样的公设基础上就会出现“上帝是否能够造一块他举不起来的石头”的问题。这个问题会涉及哥德尔所用的工具——自指，在体系内部并没有好的答案。将这三个基本公设与具体的生活实践对照，也会具体到耶稣是不是神的儿子，圣母玛利亚是不是凡人这样的问题，这样新的公设进入原来的公设系统，形成了不同的教派。前一个问题分开了犹太教和基督教，犹太教不承认记述耶稣生平的《新约圣经》，而对圣母玛利亚地位的肯定与否区分了天主教和基督教。值得注意的是，这些分歧并不是免费的和善意的讨论，每一次对教义的不同诠释所形成的教派，彼此之间是冲突的。这样的冲突在早期，以至于到今天都还意味着用消灭异议者肉体的方式消灭不同言论。教义只有经过长期的、众人的考验，才能最大可能地包容、弥补理论的不足，让它获得深入解释而相对平衡稳定，成为一种成熟的宗教。这个时间和人数的总体量，至少是亿万人年，即需要十亿人一千年的思考、辩论和考验，它才能相对稳定下来，不会以简单的流血的方式来

解决问题。当然，这个数字是个统计学意义上的数字，有实证，但没有必然的因果。不管怎么说，哥德尔是个有神论者，他甚至证明了上帝的存在。

哥德尔的理性主义

在哥德尔的手稿中，有一捆未标明时间的散乱手稿，其中有一页大概写于1960年前后，是用速记法写的，中间夹杂了一些英文单词。在这页纸上，哥德尔用《我的哲学观点》做题目，列出了14个条目，看起来是他的基本哲学信念，同样，为了避免翻译产生更多的歧义，我列出英文原文，但这些英文原文也是从速记法里翻译而来的，也未必是哥德尔的本意。

1. The world is rational.

世界是理性的。

2. Human reason can, in principle, be developed more highly (through certain techniques).

原则上，人类的理性可以得到更高的发展（通过某些技术）。

3.There are systematic methods for the solution of all problems (also art, etc.).

人有解决所有问题的系统方法（也包括艺术等）。

4.There are other worlds and rational beings of a different and higher kind.

宇宙中还有其他世界和理性存在的不同的和更高级的物种。

5.The world in which we live is not the only one in which we shall live or have lived.

我们现在生活的世界并不是唯一的我们活过的或者将要活着的世界。

6.There is incomparably more knowable a priori than is currently known.

未知的世界远超过已知的世界。

7.The development of human thought since the Renaissance is thoroughly intelligible （ durchaus einsichtige）.

自文艺复兴以来人类思想的发展是完全可理解的。

8.Reason in mankind will be developed in every direction.

人类的理性将朝着各个方向发展。

9.Formal rights comprise a real science.

真正的科学必然要求形式正确。

10.Materialism is false.

唯物主义是错误的。

11.The higher beings are connected to the others by analogy,not by composition.

更高级的生物通过比喻而不是通过结构化的方式与其他生物沟通。

12.Concepts have an objective existence.

概念是客观存在的。

13.There is a scientific （exact） philosophy and theology,which deals with concepts of the highest

abstractness; and this is also most highly fruitful for science.

存在一种科学的（精确的）哲学或者说神学，它涉及最抽象概念；它对科学本身也有丰富的意义。

14. Religions are, for the most part, bad——but religion is not.

宗教的内容很多是错误的——但宗教本身没错。

哥德尔所指的理性主义源于莱布尼茨的思想：世界，包括我们内在体验的世界，是完整和优美的，因此是理性和有序的。哥德尔认可这种信念的正当理由部分取决于数学的完整和优美的归纳推广：“理性主义与柏拉图主义有关，因为它涉及概念方面而不是真实的世界。数学有一种完美的形式，我们可以期待概念世界是完美的，而且，客观现实是美好的、完美的。”

宇宙的总体现实和总体经验是美好而有意义的，这也是莱布尼茨的思想。我们应该通过我们真正了解的部分来判断现实。从概念上我们知道的那部分知识是如此美好，我们所知道的现实世界也应该是美好的。虽然哥德尔对理性主义信仰的根源的认识是形而上学的，但他在该领域的长期愿望始终是具有实践性的。即，开发哲学中的确切方法，将其转化为精确的科学。

在实践中这意味着我们要采取最严格的观点来确定接受主张的辩证理由。换句话说，在接近数学证明的哲学论证中，渴望达到足够的严谨。哥德尔早期的理性主义观念指的是数学的严谨性，“我能断言的最多就是反驳了形式逻辑的观点，认为数学只包括语法规则及其推论。此外，我提出了一些有力的论据来反对更普遍的观点，这些观点力图主张数学是我们自己的创造。我可能更相信柏拉图主义式的现实主义。我觉得我们应该做的是，在对有关概念进行充分澄清之后，用

数学这一严谨的方式进行这些讨论，这将证明柏拉图主义观点是唯一成立的观点”。

除了方法论部分，从哥德尔列表中的项目可以看出，哥德尔的理性主义也有乐观的组成部分：一旦拥有了适当的方法，哲学问题，例如伦理学中的问题，也可以被解决。至于数学断言，例如集合论中的连续统假设，一旦概念分析得以正确的方式进行，例如集合的基本概念已被完全阐明，连续统假设也应该可以解决。

哥德尔哲学观点的14个条目中，有两条涉及现实主义。

10. 唯物主义是错误的。

12. 概念是客观存在的。

概念是真实的对象，多个事物和概念组成的结构也是真实的对象，事物的属性和概念之间的关系，可以独立于我们的定义而存在。

物理实体的假设是合理的，有很多理由让人相信它们的存在。实体的存在对于获得令人满意的数学系统是必要的，因为物理实体对于我们感官认知获得令人满意的理论是必要的。在这两种情况下，不可能将关于这些实体的命题解释为仅是“描述”的命题。数学，是实际发生的感知才发生的。人们想要做出作为关于感官认知的某种陈述是基于这样一种观察，感官获得的数据与它们所处的环境之间有着千丝万缕的联系，没有哪种事实与我们想要断言的陈述之间没有关联而成为绝对的客观。发现新事情，是一个高级的思想活动。这样的思想活动，令人惊讶的不是形式化逻辑之间的区别而是另外一方面的差别：直觉，数学的洞察力和意义。图灵机，即使不能够掌握与数学意义相关的问题，这些机器依旧是最好的物理工具。由于物理学17世纪以来的成功，它事实上确立了人类普遍的客观唯物主义的信仰：相信在人的存在之外，有一个独立的宇宙，它的运作是不依赖人类而存在的。因为在人类出现之前，宇宙至少已经存在了137亿年。

哥德尔也反对实证主义，即“名词的含义是被实证检验才存在”的观点。哥德尔在手稿《数学是语言的语法吗？》中强调了这一类比，“这是红色”作为一个直接的数据是任意的，但不要考虑表达模式或完全归纳的命题，或者可能是后者所遵循的一些更简单的命题。

在1947年的《什么是康托的连续统问题？》的文章中，哥德尔阐述了这样一种观点，即在有意义的数学命题的情况下，总有一个事实要以肯定或否定的方式来决定。如果存在数学对象或概念的领域，那么关于它们的任何有意义的命题必须是真或假。连续统假设是实数范畴里的一个有意义的问题的例子。“大小”的概念明确地引导了假设的确定含义，因此它应该是可判定的，至少原则上。他接着为可判定性提供了两个标准：首先，涉及概念分析，它与希尔伯特的理性主义计划相关联，先将康托连续统问题表达为可以由公设所设定的语言具体含义。其次，人们必须关注所谓公理方法的可靠性，当我们用公理方法作为检查或指示寻求解决其真理的方向时，我们要时刻警惕和怀疑它，即使它是我们迄今为止找到的最好的工具。例如，哥德尔在论文中指出，公理方法存在不可避免的不合理性，它可能是错误的！

总而言之，哥德尔不完备定理大概产生了以下三个方面的影响，当然，它的影响虽然没有被大众广泛了解，甚至不为专业的科学研究人员广泛认知。但可以想象，它不会被历史掩盖住，它的光辉会照亮新的人类未知的世界。

第一，哥德尔不完备定理深刻地揭示了形式逻辑系统的内在局限性。

这种局限性是由形式逻辑系统的本质所决定的，是不可克服也是无法弥补的。一个形式系统的无矛盾性在本质上是超越这个形式系统本身的。它让形式逻辑处在两难境地：或者允许在逻辑思维中有矛盾存在，或者承认存在着逻辑方法证明不了的本逻辑系统内部的问题。事实上，无论是作为科学认识前提的公设或假设，还是在此前提下的逻辑推理，都有其内在的不可证明的直觉信念。人类认识世界总在一

定的直觉引导下通过逻辑推演而展开，并在获得新知识的过程中不断扩展着关于世界的新观念。因此，信念的合理性是相对的，它可以随着认识的深化而不断发展变化。在信念转化为知识的过程中，真正起作用的是科学家非机械的、非逻辑的、智力性的创造，逻辑自洽和完备只是一种理想的指向和要求。

第二，数学是科学的基础，数学的不完备性说明科学结论也是不完备的，换句话说，它是开放的。

自从近代科学开始努力数学化以来，科学问题就成为被数学编码的问题，科学结论就成为被数学化的结论。数学曾经被认为是精确论证的顶峰，真理的化身，是关于宇宙的顶层设计真理。绝对的确定性和有效性可能是几个世纪以来人们所信仰的数学的一大特征。然而在1931年之后，我们看到数学的绝对确定性和有效性已丧失。作为科学的语言、模式、方法和工具的数学本身的不完备性，直接导致了由它所表述、绘制的科学结论的不完备性。

第三，哥德尔不完备定理从科学的层次上揭示了人类认识的局限性。

在回答有关自然和人类社会的问题上，许多人似乎从来就没有经过认真的思考便不由自主地接受了这样的观点：人的心智或认知能力是没有根本性限制的。但是，自己创造的事物正是人类反观自己最好的镜子，数学作为人类心智与理性的产物，正好反映了人类认知的局限性或不完备性。

还要强调的是，以上三个方面的影响可能不包括量子力学，虽然它也是人类创造的。量子力学本身可能是完备的，但不是经典意义上的自洽。这也解释了为什么迄今为止量子力学都是如此的令人费解。

哥德尔不完备定理震撼了欧洲近现代文明兴起之后以笛卡儿为代表的基础主义知识论的传统。长期以来，知识的确定性是人类认识客观世界所坚守的信条。罗素曾经说过：“我像人们需要宗教信仰一样

渴望确定性。”人们一度普遍认为，真正的知识不同于个人的意见或主观信念，它是绝对确定的、必然的真理，不容置疑。这一根深蒂固的观念可以一直回溯到柏拉图对知识的思考。柏拉图学派认为知识源于独立于时空之外的、不可知觉的理念世界，而在理念世界中的事物是永恒的、确定不变的，因此知识是确定的、可靠的、真实的。信念源于可感知的现象世界，现象世界中的事物是短暂的、流变的、不确定的，因此信念是不可靠、不确定的。两千多年来，人们的认识总是在思维中舍弃对象世界和自身的不确定性的因素，通过思维中的确定性来建构对象世界的确定性，从而达到对自然世界的确定性认识。

哥德尔直言不讳地说，我们没有任何绝对确定的知识。言外之意是，哪怕极其简单的事情，我们也没有绝对的把握说自己完全捕获了全部的终极的客观实在。物理学家霍金（Steven Hawking）则诙谐地说：“上帝不仅掷骰子，有时他还把骰子扔到了找不到它们的地方。”

在新的公设被找到之前，现有的公设系统就具有了不确定性。一个命题是否是对的，会不会出现矛盾？这在公设系统被建立之后并不能确定。理性系统不能够给我们一个安全的保障，即使我们建立了一个足够强大的公设系统，也并不能保证在这个系统里的所有可描述问题都能被推导出来并得到证明，相反，还是有一些命题在这个系统中存在，但它们不能被机械地遍历来确定正确与否。

我们对理性长久以来的信仰的城墙倒了。

然而，承认知识的不确定性，会让我们自然想到，科学上的结论其实并不具有许多人所坚持的科学本质的意义，而只具有认识的意义。就是说，科学结论不应该是对真实世界的终极反映，相反，科学结论只是人们用心智与理性构建出来的关于自然的图景。这幅图景，随着我们认知的深入，可以有不同的表象和形式，并且是可以变化的。绝对的确定性知识的不可能与知识不确定性的发现，并不意味着知识的消亡，相反是知识的“新生”。知识的确定性应由此获得新的

意义，它可以被视为人类理性的一种理想追求。尽管并不是完美的佳作，即使不断完善也未必能去除所有的瑕疵，然而它是人类与感性知觉世界之间最有效的纽带。

就知识的确定性而言，数学曾经是一种理想，而今我们也一样认为它是理想，尽管我们理性地认识到理性的终极理想也许永远达不到。追求绝对确定性知识的不可能，并不意味着我们在知识的可靠性上，要采取彻底的怀疑主义的态度。怀疑虽给我们提供了更深更广的思想空间，但如果认为知识没有绝对正确的把握就不应产生，也是不对的。世界上虽没有那种不容置疑的绝对确定性知识，但知识也不是非要有这种不容置疑的绝对确定性。因为，尽管我们对任何事物的认识随时都会犯错误，但是我们对自然的认识毕竟在这样的悖论中真的前进了。知识的历史，就是人类不懈追求知识的确定性而逐步显现知识的不确定性的历史。

知识中的悖论给我们一种启示，它划定了哪里是我们通过经典理性可以构造的区域，哪些是理性可以思考但无法证明的区域。这恰恰是物理世界为数学与自然对应的新证据和弥补有限公设之外的新假设提供的源泉。由此也产生了物理学家的鄙视链，实验物理学工作者评论理论物理学工作者时说“他们是做数学的”。不过我们还是有必要区分一下经典理性和我们希望获得的理性，经典理性的特征在于可描述性，而理性本身的范畴可能远远比可描述的范畴大。当想起来图灵说，所有可描述算法，图灵机都可以实现时，我们就知道了，原来，可认知的理性会多于图灵机可实现的理性。

哥德尔不完备定理揭示了真实和可证之间那条无法逾越的鸿沟。它告诉我们，我们把握自然的真实，未必是通过理性的逻辑来把握，至少一些情况下是这样的。而科学的理性在于怎样通过严谨的认知过程，去证实某些命题的合理性。这些关于真理的发现能力的限制，对人类来说是个好消息，我们不用太担心机器会自动发现真理。它可以推演出在我们所发现的真理假设之上，应用形式逻辑而得到的，我们

没有发现或还没有看到的命题，正如AlphaGo（阿尔法狗）所完成的在人类尚未穷尽的棋局上的尝试。它确实尝试了很多人类没有试过，但显然又符合游戏规则的棋局。但创造棋局和游戏规则本身，却是一个从哥德尔不完备定理上来讲没法由机器来完成的事情。

这里会有一个类似于费米悖论的存在。20世纪50年代物理学家费米（Enrico Fermi）在和别人讨论一个关于飞碟和外星人的问题时，问了这样一个问题：“他们都在哪儿呢？”从时间上来讲，人类的科技，要是再用100万年的时间来发展的话，是能够飞往银河系的其他星球的。要知道，我们人类科技真正的发展时间，不过才几百年，现在就已经能够飞出地球。那么百万年之后呢？很难想象那个时候的科技会达到什么样的程度。可是，宇宙的历史比这个100万年久远太多了。如果说，宇宙至今存在了137亿年，宇宙里有着其他生命，总有某些外星人轻易比人类早进化100万年，就应该已经来到地球才对。换言之，费米悖论表明的就是这样三种选择：

1. 外星人是存在的——外星人的进化要远远早于人类，他们应该已经来到地球并存在于某处，只是不现身罢了。

2. 外星人是不存在的——迄今为止，人类并未发现任何有关外星人存在的蛛丝马迹。

3. 外星文明存在过，在外星人发现穿越时间的机制前就毁灭了。

同样，当我们构想人类会被人工智能的机器人超越，或者按照某些说法，就在2049年，这些机器人就有穿越时空的能力。甚至它们不受生物年龄的限制。但，它们在哪儿呢？不要谈什么宇宙法则，任何法则可描述的时候都有漏洞，尤其是它们是建于逻辑系统之上的机器的时候。

同样，科学本身如果定义为旨在寻求可证的普适真理的时候，这个工具恐怕也值得怀疑。因为的的确确有一类真理可以被人类感知，由人类的直觉从自然中提炼出来，但无法证明。这并不是否定科学本身存在的意义，科学的意义在于它可以不断扩展自己的认知范畴。况且，人类认知也还存在超越经典理性的层面，我们有足够的共识来理解事情本身，但缺乏足够的工具来证明这个事情。我们还有一个阶段是，寻找新的假设来证明在这个假设所弥补的范畴内，所证明的问题是可证的。无非是我们要知道这个新补充进来的假设，又会在其他问题和时间上出现新的不可证问题。但，不是眼前这个。

人类在经典理性之外的认知也可以称为理性内容。它主导了我们对未知世界的探索，甚至是新假设的判定，这也是理性，但我们似乎尚缺乏一个稳定的描述来理解它。这并不奇怪，一旦可以描述和理解，它就转化为图灵机可以完成的算法，机器就可以代替我们了。我不能否认，量子力学和这个更深层次的理性认知，也许始终会有些关系。

因此，科学圈也在建立公共的认知群体，通过群体建立知识共同体的道德规范。这确实有点像英美法系的大法官系统，尤其是在我们面对未知世界的时候，这样的评判体系是有益的，它建立于人类的直觉认知和科学习惯所形成的道德之上。

意识和梦境

以牛津大学物理学家彭罗斯为代表，他说：“从哥德尔不完备定理的考虑中我们可以看到，在形成数学判定的时候，或者在计算和严格证明作为数学的呈现工具的时候，意识的角色反而是非算法的。”哥德尔自己思考这个问题的时候跟大多数人一样，希望能够进一步找

到证据来证明只要把意识和他的不完备定理结合起来，即可成功地证明我们的信念：人的心智确实胜过计算机。只要表明心智在判定数学问题上的优越能力，就能证明这一点。以梦境为例，如何用逻辑方法来判断最简单的心智中关于梦的问题：我是不是在做梦啊？

《庄子·齐物论》中有说：

> 昔者庄周梦为胡蝶，栩栩然胡蝶也……不知周之梦为胡蝶与，胡蝶之梦为周与？周与胡蝶，则必有分矣。

这段话大意为，庄周做梦梦见自己变成了蝴蝶翩翩起舞，但不知道是庄周梦为蝴蝶，还是蝴蝶做梦成为庄周。

英国作家埃里克·赫顿（Eric Hutton）在1989年描述了关于梦和现实本质的悖论。和很多人一样，赫顿描述自己醒来以后有非常清楚的关于梦的记忆，其中一切都像醒着的生活一样真实。与庄周一样，这些清晰的记忆会让人怀疑生活本身是否是一个梦。有一天赫顿终于想出了一个神奇的检验判据用于判定他是否在梦里：如果我发现自己在问“我在做梦吗”，这就证明了我在做梦。因为在清醒的时候，“我”从逻辑理性的角度讲，自然很清楚自己不是在做梦，因此也不会问自己“我在做梦吗”这个问题。因此，当我问自己是否在做梦的时候，一定是在梦境中了。这个判据是如此有道理，很快被人称为“赫顿判据”，用来区分梦境和现实。老实说，这比打自己一耳光看疼不疼的方法靠谱多了。

许多年后，当赫顿写了一篇关于他童年时代对梦的兴趣的文章时，他突然意识到这个推理中的矛盾。没错，问自己“我在做梦吗”确实证明了一个人是在睡梦中。然而，这正是他自己在完全清醒的时候提出来的判据。

是不是有点诡异的哥德尔的味道？

第七章

人工智能的不能

人的日常工作，大量地基于不可解问题的复杂性。在某一领域和问题的单一解决方案上，计算机能够自我优化地提升解决问题的效率。在这一点上，恐怕人是要叹为观止了，在惊叹中围观，自己就不做了。但人的创新能力在于寻找新假设的能力和基于计算复杂的多重选择能力，这两者都能被哥德尔不完备定理所涵盖。

深度学习

1956年夏，约翰·麦卡锡（John McCarthy）和马文·明斯基（Marvin Lee Minsky）等科学家在美国达特茅斯学院开会研讨“如何用机器模拟人的智能”时，首次提出“人工智能”这一概念，标志着人工智能的诞生。从那以后，研究者们发展了众多理论和算法，人工智能的概念也随之扩展。而今，当我们再次把焦点关注到人工智能上时，它已经有60多年的历史了。2006年，基于神经网络的深度学习算法取得重要突破，人工智能顺势迎来新一轮投资界和工业界的追捧。警惕人工智能，这之后，很多“科技”明星代言人也在不同场合提醒人们要防止人工智能取代人类，甚至是统治人类。

电影《模拟游戏》讲述了二战中图灵的真实故事——德军认为牢不可破的密码被图灵的机械解码器破译了，盟军从此掌握了战争情报的先机，最终赢得了胜利。从那时起，人们就开始想象，计算机可以解开人解决不了的难题，按照这样的趋势发展下去，计算机就有可能超越人类。20世纪80年代，个人电脑的普及带来了人类对人工智能的第二次恐慌，1997年计算机深蓝战胜了国际象棋大师加里·卡斯帕罗夫（Garry Kasparov），电影《终结者》和《机械战警》都是这个时期的代表作品。2006年以后随着深度学习技术的发展，人类迎来了对人工智能的第三次恐慌。美剧《西部世界》和电影《机械姬》就代表了这一阶段人们对技术发展可能超越人类智慧的隐隐恐慌。美国未来学家雷·库兹韦尔（Ray Kurzweil）提出奇点理论，被互联网人追捧，人们担心，到2049年，人工智能就可能超过人类，从此绝尘而去，人类会被机器人奴役，地球会被机器人统治。

人工智能产业最近几年进入快速上升通道，一方面有赖于过去六十多年学术界的知识沉淀，另一方面则得益于近些年来学术界和产业界高频次、高效率的互动。大数据、硬件计算能力，以及算法能力推进了人工智能的高速发展。广义的人工智能指人所创造的、代替人从事某些思维行为的设备。它可以是算盘，可以是计算器、计算机，以至于超算中心上基于算法行为实现了类似于人类逻辑推理。从狭义讲，从2006年开始的这一波人工智能浪潮，是在已有科技的基础上因为深度神经网络的突破而获得的发展。几年之间深度学习算法在语音和视觉识别上实现突破，使人工智能产品更加成熟，而围棋人机大战、无人驾驶等一些具有广告效应的应用再次引起了社会的广泛关注。

作为人工智能的基础，大数据的重要性很早就被产业界看到了。2009年，谷歌通过分析近万亿条被网民检索的词汇和美国疾病控制与预防中心发布的2003年到2008年间季节性流感传播时期的数据，成功预测了2009年冬季流感在美国的传播路径和时间。谷歌同样利用深度学习图像算法，对谷歌地图所拥有的5000万张街景照片进行图像识别，从而得出了地区经济状况甚至预测地方选举选情。对大数据而言，往往是数据大而无法有效找到有意义的结论。在很多问题中，人们需要新的处理模式才能让大数据工具有更强的决策力、洞察发现力和流程优化的能力，才能让海量、高增长率和多样化的大数据变为信息资产。传统的逻辑推演、机器学习、统计、模式识别、概率，它们都为大数据的有效利用提供了不同方面的工具。作为人工智能的代表算法，深度学习也同样给出了一种新的大数据处理手段。这一次基于深度学习的人工智能热潮，其核心就在于通过大规模计算寻找海量数据中的规律，再运用该规律对具体任务进行预测和决断。

以图像识别的深度学习算法为例。每一张图片表现为数字序列都是大到几兆比特的大数据，而图像里的人对应于“谁”，是深度学习的算法要找到的目标。深度学习就是建立原始数据——“图像”和图像所表达的意思——“人名”之间的关系的一个特定的数学模型。这

个数学模型，通过自适应算法，可以在大量的数据和对应的目标之间进行反复迭代而不断优化，以达到更高的准确率。这个模型的参数空间里的每一个参数，不一定要对应于具体的含义，不一定需要有逻辑上的关系。人们只需要以效果来判定算法的好坏，只要求计算准确，而不需要理解参数的实际意义。正因为这样，深度学习也存在这样那样的隐患，因为我们对同样海量数据所构建的深度神经网络中的参数空间不能够完全理解，一些不可控的未知因素被逐渐发现，已有算法的漏洞也逐渐在研究中被观察到。

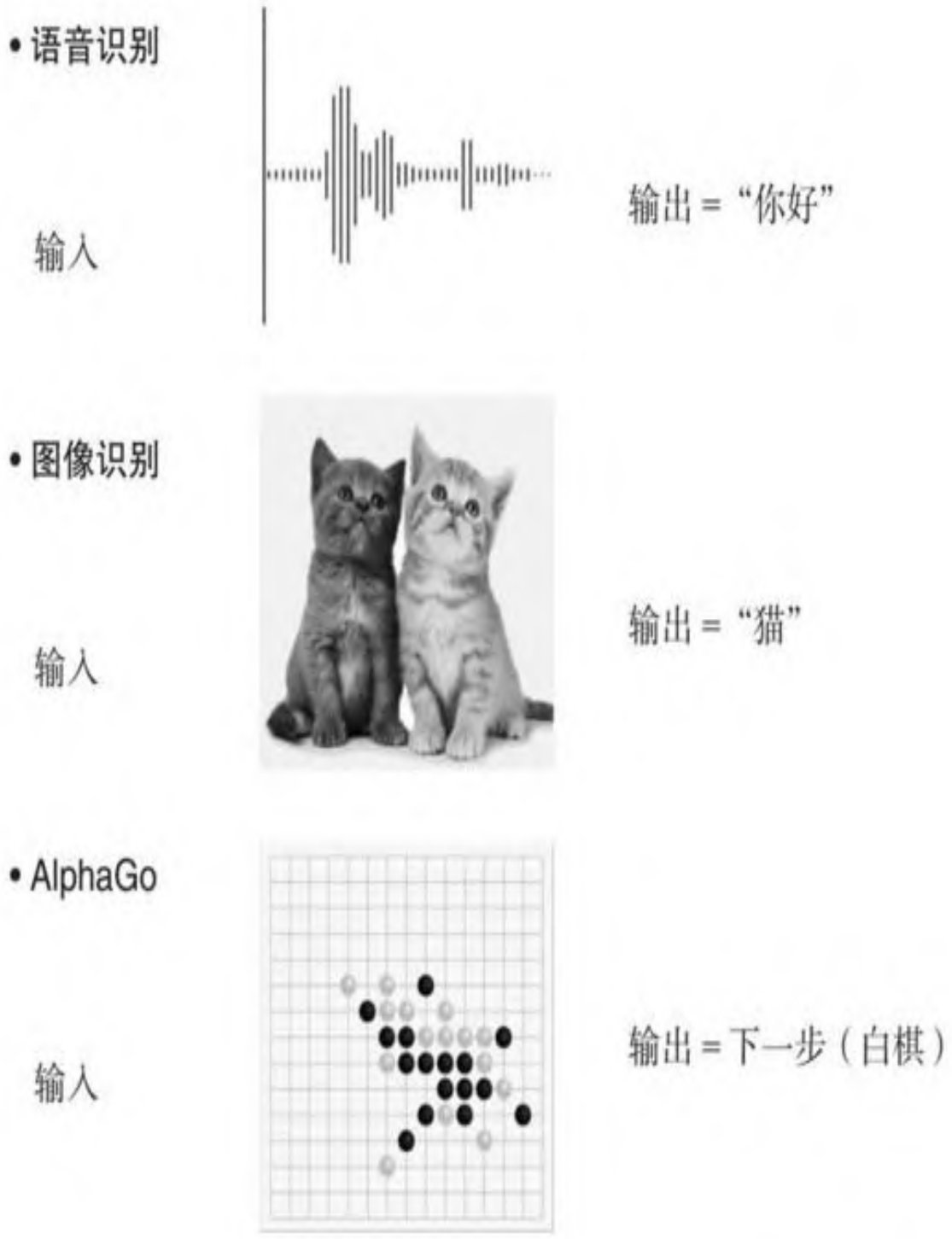


图7–1 深度学习≈找到一个方程（足够大的矩阵）

人工智能技术取得突飞猛进的发展又得益于大数据的支持，正是因为有了海量数据，人工智能才能基于这些原材料进行相应训练。大数据是人工智能发展的基础，这是因为大多数人工智能技术依靠统计模型来进行数据的概率验算，通过把这些模型暴露在大数据中，使模型和参数不断优化，这或称为“训练”。有了大数据的支持，深度学习算法的输出结果会随着数据处理量的增大而更加准确。因此，学术及研究机构承担建设了公共数据集，其目的在于依靠优质数据的不断丰富，推动算法的进化和研究力量的不断成长。随着算法的相对成熟，行业数据集与产业结合紧密，行业也开始讨论人工智能的伦理学，与人类学、社会学、法律学、伦理学、生物学等领域展开更大范围的互动与合作，开展学术研究，不管有没有，都要摆出来未雨绸缪的姿态。

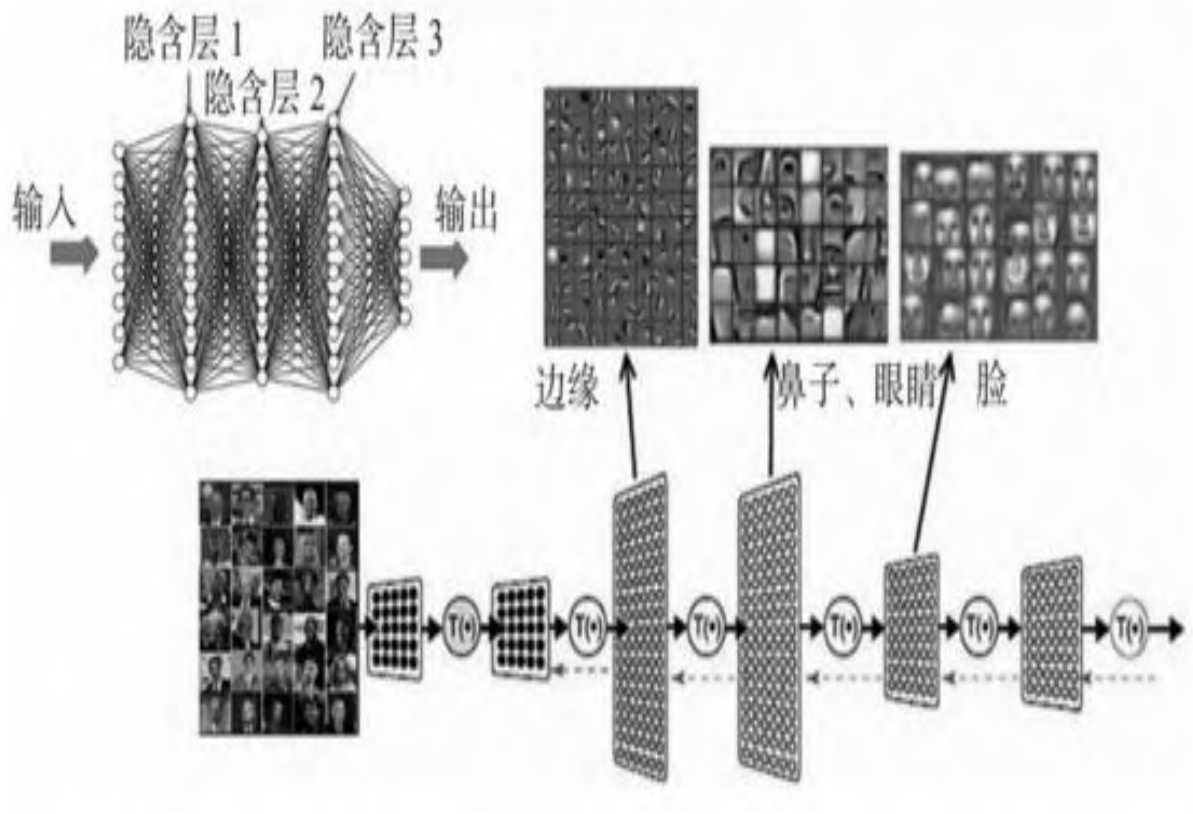


图7-2 典型的用于图像人脸识别的深度学习神经网络结构示意图

超越有限系统

早在两千年前，古希腊哲学家芝诺断言长跑健将阿喀琉斯跑不过乌龟，他这样推理：

假设乌龟的速度是阿喀琉斯的1/10。先让乌龟跑1米，阿喀琉斯追上这1米，但乌龟在这期间又向前爬了0.1米，当阿喀琉斯追上这0.1米时，乌龟又向前爬了0.01米，这样下去，只要乌龟先出发一点，阿喀琉斯就永远追不上乌龟。

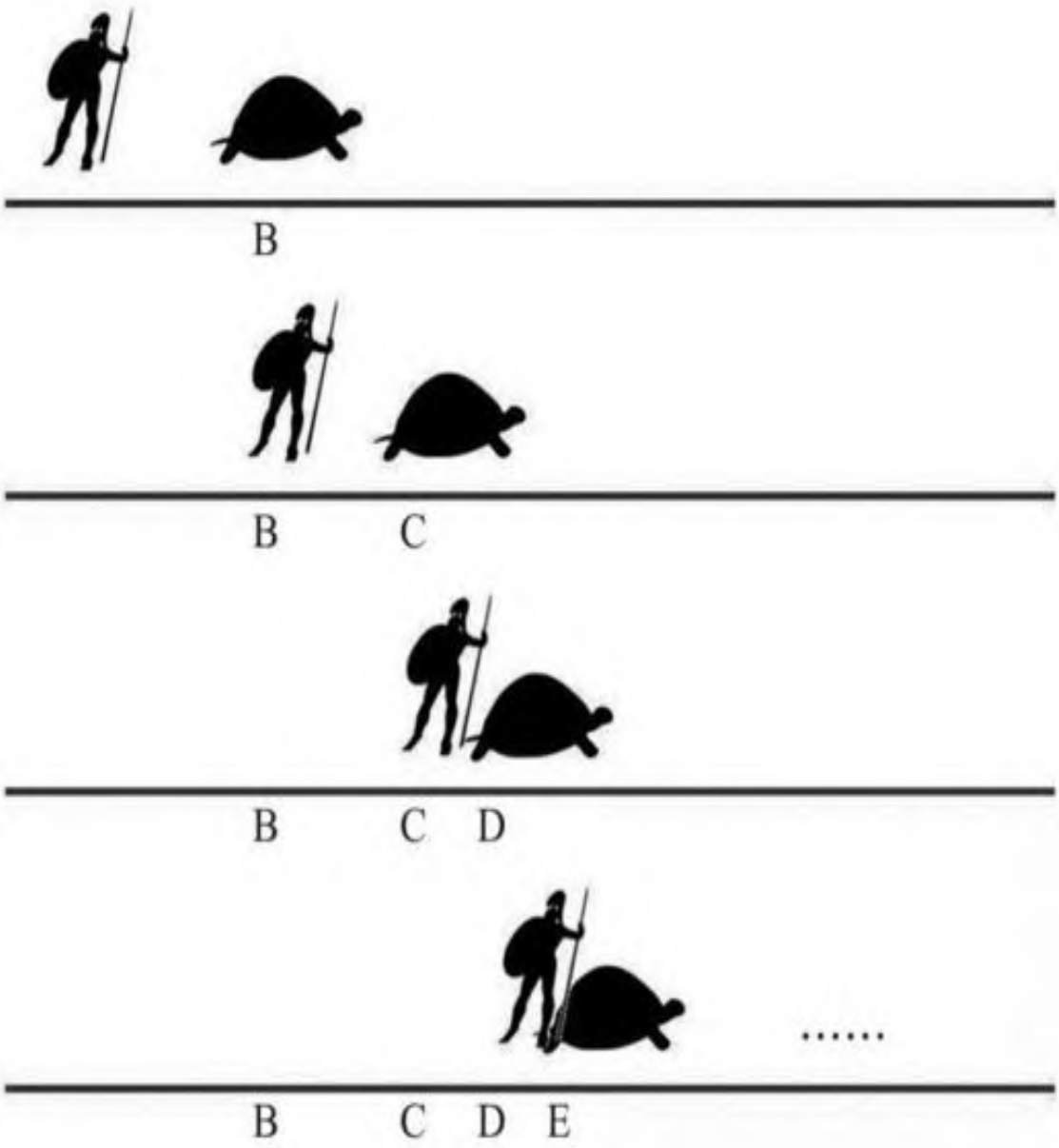


图7–3 芝诺的赛跑问题

在16世纪，数学家开始认真地追究什么是无穷。牛顿和莱布尼茨这时已经开始在他们所发明的微积分中大量使用了无穷的概念。芝诺的赛跑问题在引入极限的概念之后得到了很好的解答。假设阿喀琉斯的速度是1米每秒，乌龟速度就是0.1米每秒。那么阿喀琉斯追上第一个1米要1秒，这时候乌龟爬过了0.1米，阿喀琉斯只要0.1秒就赶过了这个时间，所以阿喀琉斯追上乌龟的时间总共要1+0.1+0.01+0.001+……=$1\frac{1}{9}$秒。

但当数学家把无限的数字加在一起时发生了新的矛盾。如果列表中的数量变得更大，总和将“发散”，即随着数量接近无穷大，总和也是如此。通常我们会看到一组数的总和是：

1+2+3+4+5 +…… =无穷大

但是，如果单个项迅速地变得越来越小，那么无穷多个项的和可以越来越接近有限的极限值，我们称之为系列的总和。例如：

$$1+\frac{1}{9}+\frac{1}{25}+\frac{1}{36}+\frac{1}{49}\cdots\cdots+=1.2337005\cdots\cdots$$

数学家们也开始担心特殊的无休止的总和：

1 - 1+1 - 1+1 - 1+1 -……= ？？？？？

如果你将系列分成几组，它看起来像（1 - 1）+（1 - 1）+……，依此类推。这只是0+0+0+……= 0，总和为零。但想想如果这个系列是1 -（1 - 1+1 - 1+1 - 1……），它看起来像1 -（0）= 1。我们似乎证明了0=1。

进一步假设：s=1 - 1+1 - 1+1 - 1+1 -……

那么，s=1－（1 － 1+1 － 1+1－……）=1 － s

当s= $\frac{1}{2}$ ，你可以看到一个更无法理解的计算结果：

$$1 - 1 + 1 - 1 + 1 - 1 + 1 - \cdots\cdots = \frac{1}{2}$$

无穷大和有限大看来存在着本质的差别。我们当然可以拒绝数学中的无穷大，只处理有限数量的数字，或者，正如柯西（Augustin Cauchy）在19世纪初针对类似的问题所展示的那样，一个无穷系列的总和必须通过更详细地指定其总和的含义来定义。在这里我们就看到一个简单的例子，一个通过扩大到无限形成的好玩的数学例子，还是要提醒读者，这些例子只有在数学上允许，物理上却无法实现，这也看出了数学和物理之间似乎真实地存在差别。

将$y=\frac{1}{x}$ 中$x\geqslant 1$的部分绕着x轴旋转一圈，能得到一个小号状图形。这个小号状图形有一个神奇的性质——它的表面积无穷大，可它的体积却是 π——一个确定的有限值。这明显有悖于人类的常识：体积有限的物体，表面积却可以是无限的！换句话说，填满整个小号只需要有限的油漆，但把小号的表面刷一遍，却需要无限多的油漆！二维几何悖论中还有科赫雪花（Koch Snowflake）。科赫雪花是一种经过无穷多次迭代生成的分形图形。它的面积有限，周长却是无限的。用无限的周长包围了一块有限的面积！

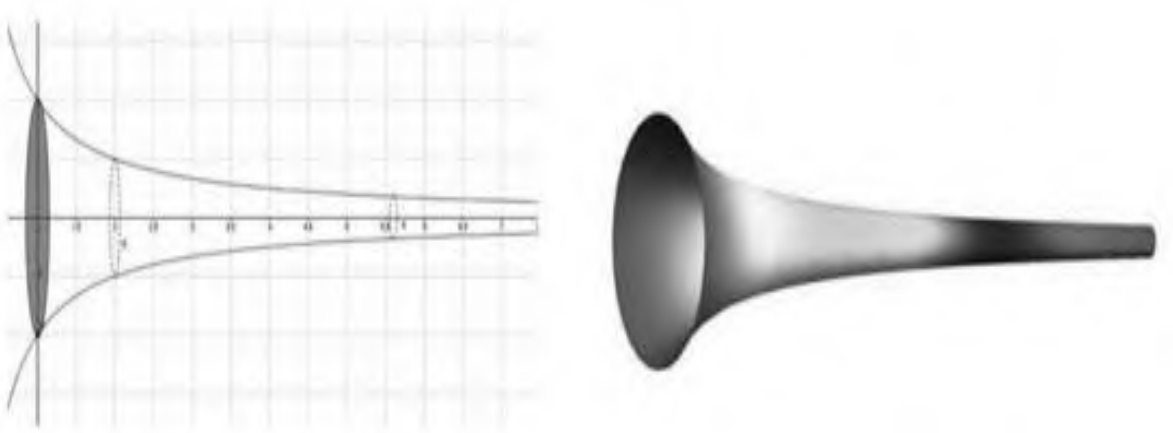


图7–4 体积有限，表面积却是无限的小号状图形

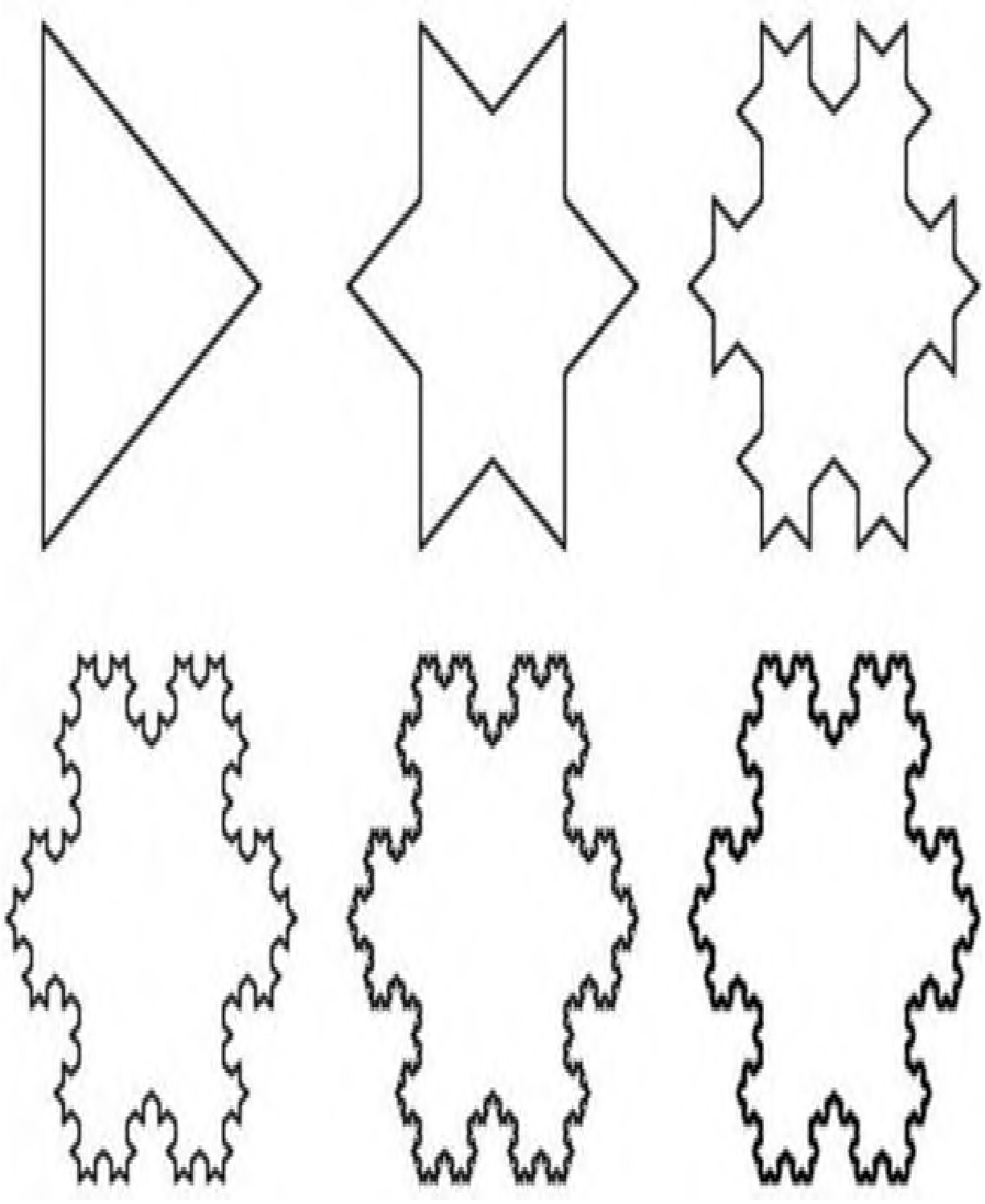


图7–5 面积有限，周长却是无限的科赫雪花

无穷长的杆子不会动

有一张无限大的桌子上固定一个三脚支架。把一根无穷长的金属杆的一头固定在三脚架上，另一头伸向无穷远处。金属杆可以绕着三脚架顶端自由地上下转动。假设金属杆和桌子都是牛顿力学中常见的刚体。在一切牛顿力学中合理的前提条件下，神奇的事情发生了，即使只有一端被支撑，这根无限长的金属杆也根本不会动！金属杆和桌子都是无穷坚硬的刚体，杆子一旦动，就会相交，必然有一个会损坏，但刚体的假设又不允许这样的事情发生，唯一的可能就是金属杆与桌面绝对平行。那么我们看到的现象就是一根无限长的金属杆，仅仅靠一个点就可以保持水平！

希尔伯特旅馆

一家旅馆有无穷多个房间，每个房间都住了客人。一天来了一个新客人，旅馆老板说：“虽然我们已经客满，但你还是能住下的。让1号房间的客人搬到2号房间，2号房客人搬到3号房间……n号房客人搬到n+1号房间，你就可以住进1号房间了。”又一天，来了无穷多个客人，老板又说：“不用担心，大家仍然都能住进来。我让1号房客人搬到2号房间，2号房客人搬到4号房，3号搬到6号……n号搬到2n号，然后你们依次住进奇数号的房间吧。”

这就是希尔伯特提出的关于无穷大的悖论。虽然人们把它叫作一个“悖论”，它在逻辑上却是完全正确的，只不过大大出乎我们的意料罢了。

意大利数学家伽利略在他的最后一本科学著作《两种新科学》（*Two new Sciences*）中提到一个问题：正整数集合 {1，2，3，4……}和平方数集合{1^2，2^2，3^2，4^2……} 哪个更大呢？一方面，正整数集合里包含了所有的平方数，前者显然比后者大；可另一方面，每个正整数平方之后都唯一地对应了一个平方数，两个集合大小应该相等才对。伽利略比较早地使用了一一对应的思想来研究无限大，可

惜没有沿着这个思路更进一步思考下去。最后他得出的结论就是，无限集是无法比较大小的。

说到这里，我们不得不提到德国另一位伟大的数学家康托，他建立了集合论，系统地研究了集合（尤其是无穷集合）的大小，只不过这个大小不是简单地叫作“大小”了，而是叫势（cardinality）。如果两个集合间的元素能建立起一一对应的关系，我们就说它们等势，这也是我们比较集合大小的方式。当我们数大教室里是不是坐满了学生，我只需要把座位和人对应起来。一个座位上坐一个人，而不需要把人数一遍，再把座位数一遍，最后看人数是不是等于座位数。我们只要把每一个数都对应为一个可数的自然数，那么这两个无穷大就是相等的。康托第一次认真研究了无穷的概念，并用了我们前面使用的证明哥德尔不完备定理的类似方式——对角线法，证明了0到1之间的无理数的数量是大于有理数的数量的。即，无理数的无穷多比自然数的无穷多要多，是更高级的无穷大。而“希尔伯特的旅馆”形象地说明了正整数集合和正偶数集合是等势的。一切和自然数集合等势的集合都称为可数集合（countable set），否则就叫作不可数集合（uncountable set）。可数的无穷大是一样大的，即偶数跟自然数一样多，这也跟我们的常识不太一致，但数学上它是对的。

问题是，今天作为人工智能基础的计算机，无论其存储还是内存有多大，都还是有限的，都还是可数的，都不是可以代表实数的无穷大，连可数的无穷大都不是。我们并不能用计算机表达一个真正意义上的无穷大。我们也无法用计算机表达一个无理数，甚至一个无限循环小数，它不得不截断到计算机作为一个物理设备所限制的对数字的表达。用一些特殊的算法，我们可以表达一个非常长的数字，但即使这样，一个无理数也会被截断。我们可以用几何和符号表达π，并且理解它的含义。但在使用计算机的时候，我们不得不用它的近似值3.14，或3.1415927，无论相差有多小，都还是个有限长的有理数，是π的近似值。这些近似很多时候会被忽略，但在有些场景下会被放大，例如随机函数发生器就是一种放大这种细微差异的算法。一个有

混沌特性的方程，也可以用来放大这种差异：本来两个输入值是可以忽略掉的，例如3.1415927和3.1415926，但经过这个方程的多次迭代之后算出来的结果就可能是天壤之别了。

机器相对我们人类的数学认知而言，既不能理解甚至表达无穷大，又不能理解甚至表达无穷小。物理设备的有限性妨碍了这些概念在机器表达中的实现。我们可以看到这其中有一些重要的差别，但关于无穷的认知是否是构成机器和人类思维的界限呢？

然而，物理学却给出了一个跟我们已有数学认知不太一致的底层问题，这个问题需要数学表达方式的进一步发展才可能有最终的解决方案。根据量子力学，我们生活的世界，例如时空，也并不是由连续的无穷小构成。从时间上讲，有普朗克时间，光波在真空里传播一个普朗克长度的距离所需的时间，数值大约为5.39×10^{-44}秒，它是最小的时间间隔。同样的，空间也有最小的颗粒，大概是1.62×10^{-35}米。物理上也并不允许存在无穷这样的概念，比如黑洞的密度可以到无穷大，对黑洞的结构现有的物理定律并不能够解释。同样，对宇宙的时间起点，现有的物理理论适用于大爆炸发生后10^{-44}秒以后的事，这之前现有的物理规律并不适用，因为又有无穷大。

在物理世界回避哥德尔不完备定理的另一种方式是，审视和考量物理世界是不是只使用到了数学的有限的可判定部分。数学可以存在无限，而物理并不存在无限。只有部分数学的结构和模式存在与物理世界中的对应，物理的相关命题可能都是来自数学命题中基于可判定陈述的那部分数学，而不是全部的数学。

图灵机

人类喜欢秩序、模式和规范。我们对于任何有意义的数字序列都感兴趣。从数拥有多少头家畜中我们发明了自然数，从测量中我们发明了有理数，从几何中我们发明了无理数。对不同的数的操作构成了算法。从古希腊开始的数学的基本认知在其形成很久以后，13世纪的西班牙数学家拉蒙·柳利（Ramon Llull）认为数学逻辑推理可以用人造的机械来实现。16世纪的英国数学家托马斯·霍布斯（Thomas Hobbes）认为头脑里的逻辑推理可以表达为数值计算，“我们在无声的思维中加加减减”。在1500年前后，达·芬奇（Leonardo da Vinci）设计了一台机械计算机，但他并没有建造出这台机器。更出名的计算机Pascaline是由法国数学家、机械师布莱士·帕斯卡（Blaise Pascal）在1642年建造的。在这不久之前，已知的更早的一台可以被称为计算机器的设备是由德国科学家威廉·施卡德（Wihelm Schickard）在1623年左右建造的。布莱士·帕斯卡写道：“算数机器产生了明显比所有动物行为更接近思维的效果。”莱布尼茨建造了一个试图对概念而不是对数学执行操作的机械装置，但是其使用范围非常有限。莱布尼茨建造的计算机器确实超过了帕斯卡的，因为莱布尼茨建造的计算器能加、减、乘、求根，而Pascaline只能加与减。从算盘到计算尺，再到机械计算器和计算芯片，人类在计算工具上的探索一直到今天都还没有停歇。图灵的天才之处在于，他认识到，在任何计算器中可执行的任何算法（即任何程序）都可以通用图灵机（Universal Turing Machine，简写为UTM）实现。于是，除了完全由硬件部分决定的计算机速度外，通用图灵机可以做任何计算机可以做的事情。而任何一台计算机都是通用图灵机的某种变体，它们是等效的。

1934年春天，哥德尔在普林斯顿做了一系列讲演。在讲演中，哥德尔引入了一类可以精确定义的可计算函数，他称之为一般递归函数。递归的概念意味着函数在运行的过程中调用自己，事实上，这也意味着无法避免的可怕的自指。这项工作很快成为图灵机的基石。

1936年，英国年轻数学家图灵发表了一篇重要文章《论可计算数及其在判定问题中的应用》，标志着图灵机的诞生。图灵在他的这篇开创性的论文中，提出了著名的“图灵机”的设想。他将逻辑中的任意命题用一种通用的机器来表示和计算，并按照一定的规则推导出结论，其结果是：可计算函数可以等价为图灵机能计算的函数。在这篇文章中，图灵定义了用图灵机可以计算的函数，同时，他也提出了一个论点：“可计算的函数”与“用图灵机可计算的函数”是一回事。换句话说，图灵机能计算的函数便是可计算的函数，图灵机无法计算的函数便是不可计算的函数。同时期，美国数学家丘奇（Alonzo Church）也得到了相同的结论。“所有计算或算法都可以由一台图灵机来执行”的观点便被称为“丘奇—图灵论题”。

图灵机由四个部分组成：

1.输入集合，一个无限长的存储纸带，带子由一个个连续的存储格子组成，每个格子可以存储一个数字或符号；

2.输出集合，一个读写头，读写头可以在存储带上左右移动，并可以读、修改存储格上的数字或符号；

3.内部状态存储器，该存储器可以记录图灵机的当前状态，并且有一种特殊状态为停机状态；

4.状态转移表（固定的程序指令），指令可以根据当前状态以及当前读写头所指的格子上的符号来确定读写头下一步的动作（左移还是右移），并改变状态存储器的值，令机器进入一个新的状态或保持状态不变。

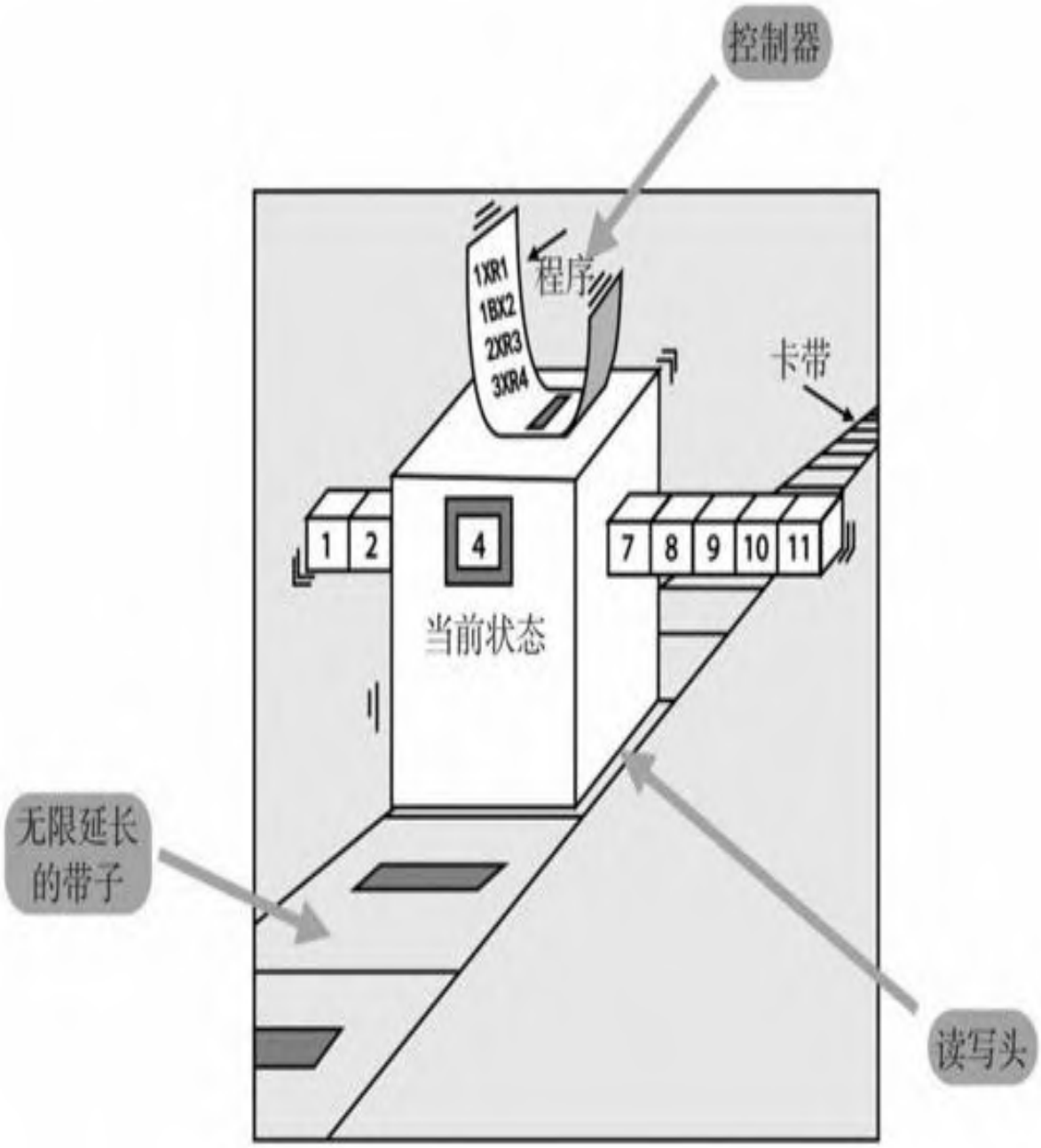


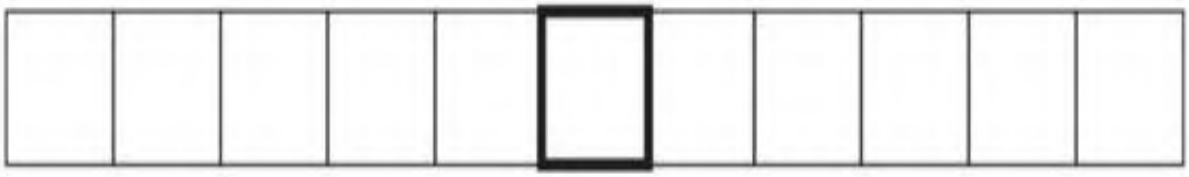
图7–6 图灵机理论示意图

整台图灵机的核心在于读写头的状态转移表，它指示着读写头的状态和当前读写头正对格子的符号如何变化。状态转移表就是由一系列这样的简单规则组成的，它相当于图灵机的源代码。

图灵机的运转和我们笔算的思维过程非常相似：在每个时刻，我们只将注意力集中在一个地方，根据已经读到的信息移动笔尖，在纸上写下符号；而指示我们写什么和怎么写的，则是早已背好的九九乘法表，以及简单的加法。如果将一个笔算乘法的人看成一台图灵机，纸带就是用于记录的纸张，读写头就是这个人和他手上的笔，而状态转移表则是笔算乘法的规则，包括九九表、列式的方法，等等。这种模式似乎也适用于更复杂的机械计算任务。如此看来，图灵机虽然看起来简单，但它足以作为机械计算的定义。

下面我们通过一个小例子来了解下图灵机到底是如何进行计算的。这个例子比较简单，我们将在空白的纸带上打印1、1、0三个数字。

假设有一条无限长的纸带，纸条上有一个个方格，每个方格可以存储一个符号，纸条可以向左或向右运动。



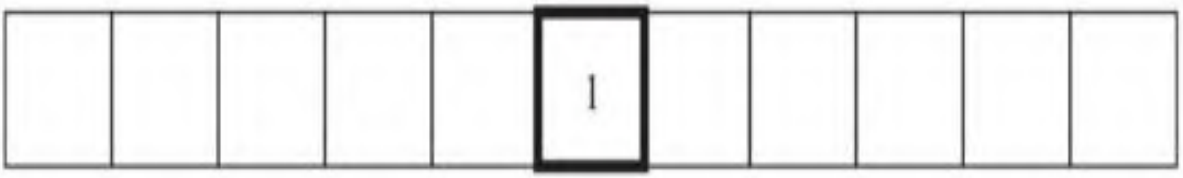
图灵机进行下面3个基本的操作：

1.读取指针头所指的方框里的符号。

2.修改方框中的字符。

3.将纸带向左或向右移动，以便修改其临近方框的值。

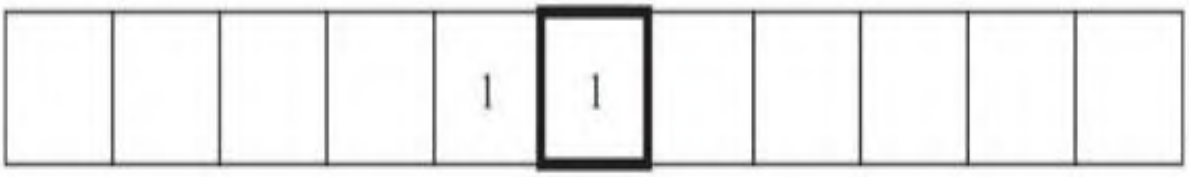
首先，我们向指针头指向的方框中写入数字1：



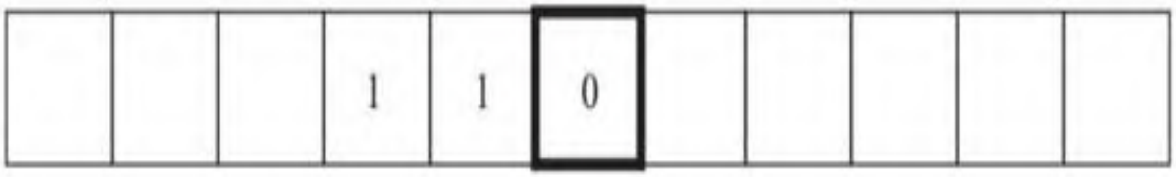
接着，我们让纸带向左移动一个方框：



现在，我们再往指针头指向的方框写入数字1：



接着，我们继续让纸带向左移动一个方框，并写入数字0：



这样我们就完成了一个简单的图灵机操作。

用图灵机完成异或操作

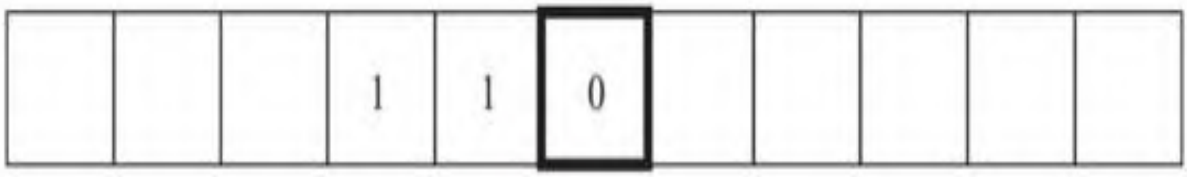
a、b代表两个二进制变量。那么它们的异或关系（计作a⊕b）是，如果a、b两个值不相同，则异或结果为1。如果a、b两个值相同，异或结果为0。

接下来我们尝试一个稍微复杂点的操作，我们尝试对1、1、0三个数做一个异或操作，将110变为001。要让图灵机完成计算，就类似于向图灵机输入表7－1中的操作指令，这些指令组成一个小程序。

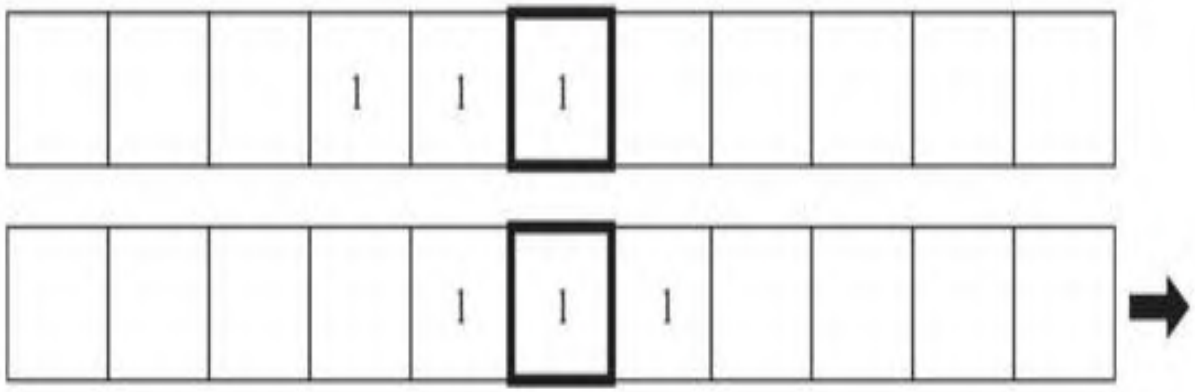
表7-1 向图灵机输入的操作指令

读到的符号	写入指令	移动指令
空	—	—
0	写入1	向右移动纸带
1	写入0	向右移动纸带

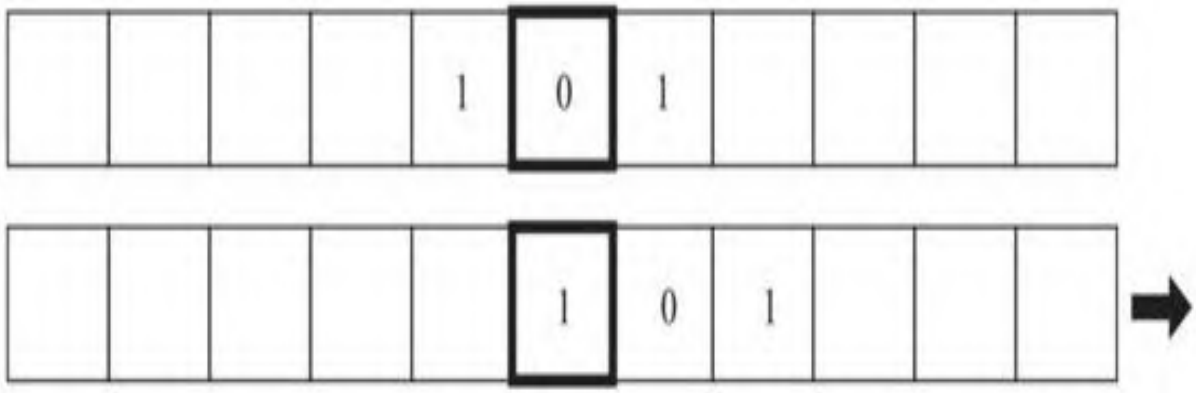
我们假设图灵机纸带现在的状态如下：



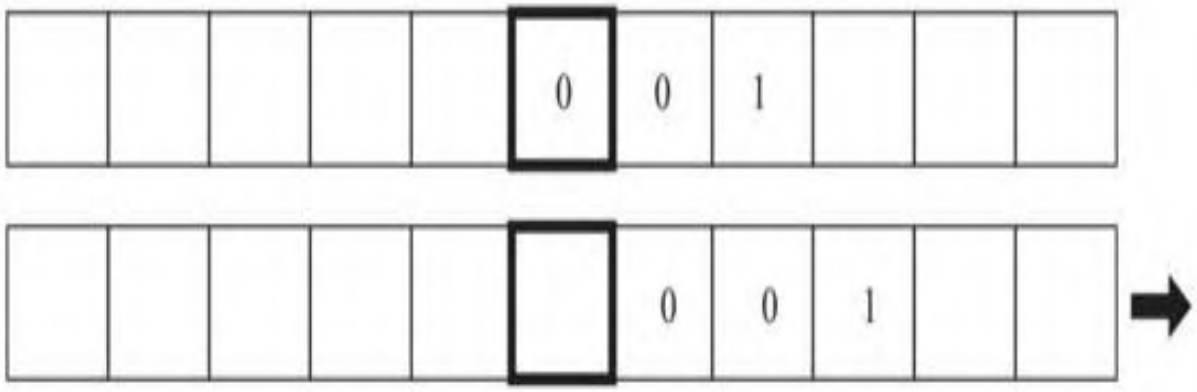
现在读取到的符号是0，按照操作指令，我们应该往方框写入1并向右移动一个方框：



现在读取到的符号是1，按照操作指令，我们应该往方框写入0并向右移动一个方框：



类似地，现在读取到的符号是1，我们重复相同的操作。



最后，我们读取到了一个空白字符，图灵机不做任何操作。按照复杂一点的计算机理论，异或操作可以是通用算法的基础，即它可以组合成各种复杂的算法。我们使用图灵机成功完成了异或操作，理论上来讲也就可以完成加法、减法、乘法、除法操作，只不过是实现的步骤复杂些而已。

我们再看一个稍微复杂一点的例子。我们把读写头当作一只能顺着纸带爬的虫子，它有两个状态，吃饱和饥饿。

1.虫子所在一个无限长的纸带上，这个纸带被分成了若干小的方格，而每个方格都只有黑和白两种颜色。纸带的片段如图7－7所示：



图7-7

2.虫子根据当前所在格子的颜色查表做出行动。

3.每一步虫子可以向右移动一个格子或向左移动一个格子。

4.虫子的操作程序为：黑色有食物，虫子吃到食物右移一格，白色是空白区，没有食物左移一格，即：

表7-2 虫子的操作程序

输入	操作
黑色	右移一格
白色	左移一格

在这个情况下格子的颜色是虫子的输入信息，集合为：{黑色，白色}，操作指令：{右移一格，左移一格}。

从图7－7中“开始”位置开始，虫子会怎么移动呢？

1.开始是黑色，虫子右移一格，到达第2格；

2.第2格黑色，虫子右移一格，到达第3格；

3.第3格黑色，虫子右移一格，到达第4格；

4.第4格白色，虫子左移一格，回到第3格；

5.第3格黑色，虫子右移一格，到达第4格；

6.第4格白色，虫子左移一格，回到第3格；

7.……

可见，虫子会在这条带子上的第4格和第3格之间来回移动循环不止，进入死循环。

我们假设一个复杂点的情况，虫子会有饥饿、吃饱两种状态，食物可以被虫子吃掉或种植。假设黑色方格是有食物，虫子可以吃掉它，虫子就会感觉饱了。当读入的信息是白色方格的时候，虽然没有食物但它仍然维持吃饱状态，只有当再次读入黑色方格时它才会感觉饥饿，这并没有实质的道理，仅是规则而已。

我们在该设定中改进下模型：

1.虫子所在格子为黑色时

（1）如果虫子处于吃饱状态，虫子不吃食物，格子还是黑色，虫子左移一格，虫子状态变为饥饿；

（2）如果虫子处于饥饿状态，虫子吃掉食物，格子变成白色，虫子不动，虫子状态变为吃饱。

2.虫子所在格子为白色时

（1）如果虫子是吃饱状态，虫子不吃食物，格子还是白色，虫子右移一格，虫子仍为吃饱状态；

（2）如果虫子是饥饿状态，虫子种植食物，格子变为黑色，虫子不动，虫子仍为饥饿状态。

那么，虫子的操作程序如表7－3：

表7-3 改进后的虫子操作程序

输入	当前状态	读写头操作	下一个状态	纸带输出
黑色	吃饱	左移一格	饥饿	黑色
黑色	饥饿	吃掉食物格子变白（不移动）	吃饱	白色
白色	吃饱	右移一格	吃饱	白色
白色	饥饿	种植食物格子变黑（不移动）	饥饿	黑色

在这种情况中，输入集合为{黑色，白色}，输出集合为{左移一格，右移一格，吃掉食物格子变白，种植食物格子变黑}，虫子内部状态{吃饱，饥饿}。

假设虫子初始是饥饿状态，从图7－8中“开始”位置启动程序：



图7-8

1.“开始”格是黑色，虫子为饥饿状态，吃掉食物开始格子变白，虫子不移动，虫子新状态为吃饱；

2.“开始”格为白色，虫子为吃饱状态，虫子右移一格，到达第2格，虫子状态仍为吃饱；

3.第2格为白色，虫子为吃饱状态，虫子右移一格，到达第3格，虫子状态仍为吃饱；

4.第3格为白色，虫子为吃饱状态，虫子右移一格，到达第4格，虫子状态仍为吃饱；

5.第4格为黑色，虫子为吃饱状态，虫子左移一格，到达第3格，虫子新状态为饥饿；

6.第3格为白色，虫子为饥饿状态，虫子要停留一下种植食物，格子变黑，虫子状态仍为饥饿；

7.第3格为黑色，虫子为饥饿状态，吃掉食物第3格变白，虫子不移动，虫子新状态为吃饱；

8.第3格为白色，虫子为吃饱状态，虫子右移一格，到达第4格，虫子状态仍为吃饱；

9.第4格为黑色，虫子为吃饱状态，虫子左移一格，到达第3格，虫子新状态为饥饿；

10.……

我们发现，这个程序似乎又在第3格和第4格之间无限循环起来了。图灵发现了这个问题，那么能不能也做一台超级算法机器来让它自动判断任何一个图灵机是否能进入无限循环呢？

1936年图灵在他那篇著名的论文中指出形式逻辑系统和通用图灵机之间的等价性。一台图灵机执行算法的行为和一个人由前提到结论

而进行演绎推理所遵循的逻辑的、一步一步的推理之间的对应。简言之就是，当给定任何一台具有无限记忆的数字计算机C时，我们可以找到一个形式系统F，使C的可能输出与F中可能的定理是一致的，反之亦然。借用这种等价性，图灵用计算机理论术语将哥德尔不完备性问题重新表述为图灵不停机问题。他证明这两个问题在逻辑上是等价的。

哥德尔不完备定理

对于任何自洽的、声称要判定所有算术陈述的，即证明或否定它们的形式系统F，都存在一个算术命题，在该系统中既不能证实，也不能证伪。因此，形式系统F是不完备的。

图灵不停机定理

对于任何声称要判定所有图灵机程序是否停机的图灵机程序H，都存在一个程序P和输入I，使得程序H不能判定处理数据I时，P是否会停机。

图灵机的意义——探索计算的极限

既然图灵机如此简单，能不能将它“升级”，赋予它更多的硬件和自由度，使它变得更强大呢？比如说，让它拥有多条纸带和对应的读写头，而纸带上也不再限定两种符号，而是三种、四种甚至更多种符号？的确，放宽限制之后，在某种程度上，对于相同的任务我们能设计出更快的图灵机，但从本质上来说，“升级”后的图灵机能完成的任务，虽然也许会慢点但原来的图灵机也一样能完成。也就是说，这种“升级”在可计算性上并没有实质性的突破，放宽限制后的机器能计算的，原来的通用图灵机也能完成。既然计算能力没有质的变化，无论采取什么样的结构，用多少种符号，都还是台图灵机。

图灵机的核心在于它的简单。这给物理上实现一台真实可用的并且可以解决“任何可描述问题”的机器铺设了道路。对于工程师而言，只要给出状态转移表，在现实中用机械建造一台图灵机并非什么难事。数学家冯·诺依曼就在图灵机模型的基础上提出了奠定现代计算机基础的冯·诺依曼架构。这种架构以运算器为中心，输入输出设备和存储器之间的数据传送通过控制器完成。从第一台每秒可以进行数千次计算的埃尼阿克（ENIAC）计算机起，到至今每秒可以进行数亿亿次运算的中国神威·太湖之光超级计算机，现代计算机的发展依旧遵循着冯·诺依曼体系。从20世纪40年代到今天，以图灵机为基础的计算机，已经深入到我们生活中的每一个角落。没有图灵机，现代人类简直无法生存，而图灵机恰恰是人工智能的基础，无论人工智能有多么雄伟壮阔的图景，它都还是一台图灵机，无非是复杂了一点，计算速度快了一点。

但我们很快就会有新的问题，是否每一个问题都可计算？会不会出现像逻辑中的悖论一样有无法判定的问题？图灵为了解答这个问题，基于他设计的能够模拟所有计算的机器，证明了：这个机器在有限时间内能够执行完毕的问题便是可以判定的问题，这个机器无法在有限时间内执行完毕的问题便是不可以判定的问题。而且由于哥德尔不完备定理的存在，图灵机中一样存在不可判定问题，而且无法避免。

图灵机模型是目前为止应用最为广泛的经典计算模型，没有之一。除了我们将要讨论到的量子计算的一些迹象之外，目前尚未找到其他计算模型可以计算图灵机无法计算的问题。图灵停机问题开启了可计算性理论的序幕，这是计算学科最核心的理论之一。图灵也提出了可以用计算机解决的问题的判定方法，为计算机编程语言的发展奠定了基础。

通用图灵机是现代通用计算机的理论原型，为现代计算机指明了发展方向，肯定了现代计算机实现的可能性。1936年，冯·诺依曼再

次慧眼识英雄，就像他高度评价哥德尔一样，他高度评价了图灵。而正是冯·诺依曼给这个领域取名为“可计算性”。判定问题有了精确的数学表述之后，立即在数学基础乃至整个数学界中产生了巨大的影响。因为这时一些不可判定命题的出现，标志着人们在数学历史上从另外一个角度认识到：有一些问题虽然有答案，但是不可能找到算法解的。在过去，人们一直模模糊糊地觉得，任何一个精确表述的数学问题总可以通过有限步骤来判定它是对还是错，是有解还是没有解。找到不可判定问题再一次说明用有限过程对付无穷的局限性，它从另外一个角度反映了数学内在的固有矛盾。一旦算法的精确定义由图灵机完成，人们同样认识到这躲不过哥德尔不完备定理的影响，可计算性和不完备性这两个概念是紧密联系在一起的。

用图灵机来模拟生命

数学家约翰·康威（John Conway）在1970年发明了著名计算机程序“康威生命游戏”，“康威生命游戏”具有图灵机的普适性，可以模拟任何其他生命游戏规则系统。这个生命游戏由一些基于简单数学规则的元细胞组成，它能模拟出细胞的生、死以及繁殖的过程。“康威生命游戏”的基本思想是给定一个初始构型，然后观察该构型按照“康威生命游戏”的基本规则如何变化（对应于我们前面所讲的图灵机的状态转移表：程序）。

生命游戏的基本规则如下：

1.元细胞分布在规则划分的二维方形网格上；

2.元细胞具有0、1两种状态，0代表“死”，1代表“生”；

3.一个元细胞下一个时刻的生死由本时刻本身的生死状态和周围8个邻居的状态决定，即：

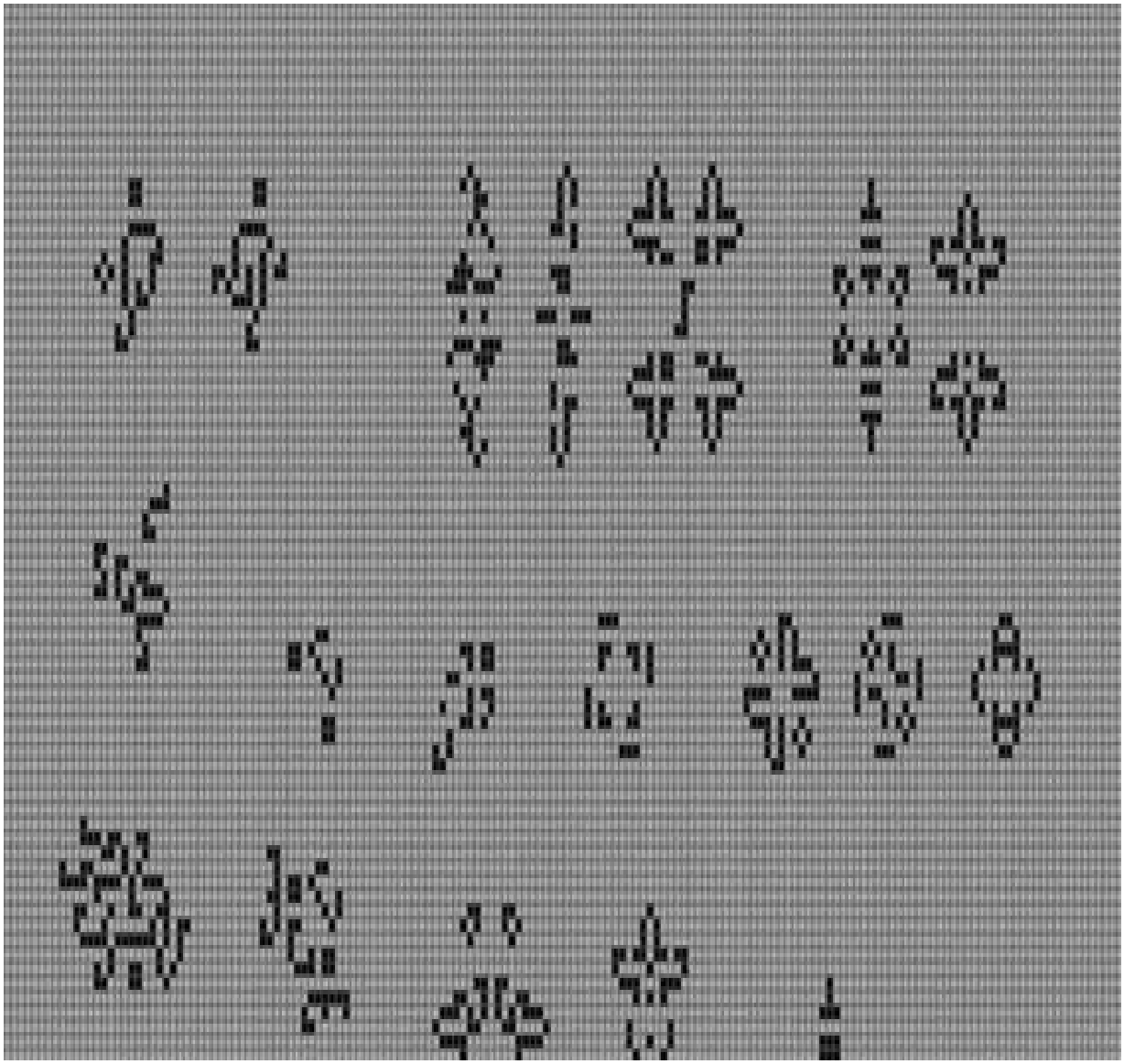
对于活的元细胞：

（1）生存：如果该元细胞的邻居中有2个或者3个元胞是活的，那么它将继续生存下去；

（2）死亡：如果它的邻居有4个及以上元细胞是活的，那么它将死于拥挤；如果它的邻居只有一个或者没有活的元细胞，那么它将死于缺乏交流。

对于死的元细胞：

繁殖：如果它的邻居有3个元细胞是活的，那么它将变成活的元细胞。



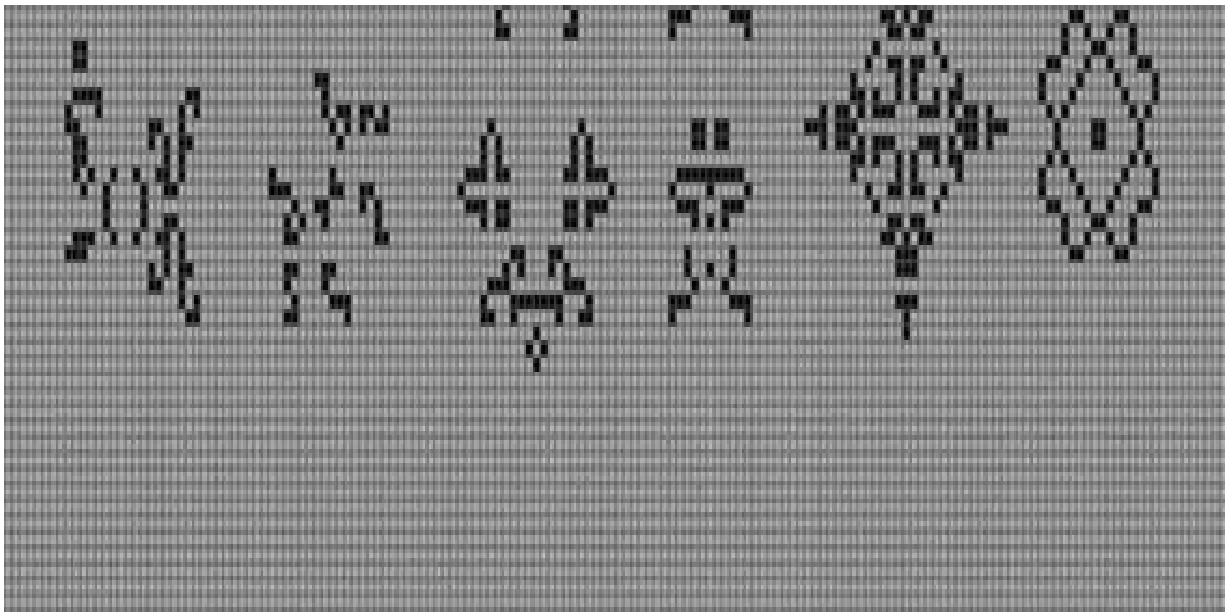


图7–9 康威生命游戏

不难用任何今天的计算机语言来在计算机上实现康威的生命游戏，你可以看到一个初始的形状的元细胞组合体（元生物）可完成移动、吞噬、甚至繁殖等基本生命的动作。

图灵机所不能够

既然一切只是效率问题，无论从算法上看，还是从硬件基础的进步和高效上看，人工智能都不会解决图灵机解决不了的问题，无论如何，到今天为止，人工智能还是基于图灵机来实现的。所以，要看人工智能哪些事情是做不了的，还是回到图灵机的基础问题上来。图灵机做不了的事情，无论今天的计算机如何强大，也都做不了。

希尔伯特提出过著名的希尔伯特规划，即给定足够的公理，运用机械推导，能否对所有合法表述的表达式提供正误判断。这个形式数

学主义者所怀的梦想被哥德尔无情地击碎了。虽然梦不在了，但有新的思路可以尝试。是否存在能在原则上一个接一个解决所有数学问题的某种一般机械步骤？那么问题的关键在于什么是“机械推导”，图灵给出了他的解决方案，并从此打开了新世界的大门。

有了图灵机，对于任何数学问题我们都可以写一个程序，如果它在有效时间内能够判定命题为真的话就停机输出1，如果能够判定命题为假的话也可以停机输出0。如果我们能够证明这是个普遍的办法，那就可以解决一切数学问题了。这是不是很熟悉？希尔伯特正是想这么做的。我们假设这样做是可以的，即，我们能够找到这样一个在图灵机上可以实现的普适性的算法。

停机问题

当我们设计了能够解决可描述问题的通用图灵机的时候，并且设计了它的输入输出和运算机制，一个自然的问题是：一台图灵机什么时候会停下来呢？假设给它一个计算机程序的源代码，也给它所有程序要用的数据、文件、硬盘、DVD（数字通用光盘）等等，所有它处理时需要的东西。它能告诉我程序最终能够输出我们需要的结果吗？并且在工作完成之后，程序是退出，还是会永远运行下去不会停止呢？换句话说就是，对于它会不会停止这个问题，仅仅检查程序和数据本身，是不是足以让它回答是或否呢？

更严格地说，会不会停机并不是图灵机本身的属性，它跟纸带的初始输入也有关系。对于同一台图灵机，不同的纸带输入也可能导致不同的结果和行为。比如说，我可以设计一台图灵机，它的任务只有一个：一步一步向右移动，寻找输入中的第一个1。如果输入纸带上全是0的话，那么，这台图灵机自然不会停止；但只要纸带上有一个1，那么它就会停止。所以，真正严谨的问题是：给定一台通用图灵机以及一个输入k，如果我们将k输入通用图灵机，然后让通用图灵机开始

运行，这时通用图灵机是会不停运转下去，还是会在一段时间后停止？是否存在一个算法，能够在有限时间内判定一对（算法，输入）的组合是否停机，我们称之为图灵停机问题。这个问题之所以重要，是因为我们最终将证明通用图灵机范畴之内不存在这样一个算法，而人脑又能通过在系统之外的洞察判定这一对（算法，输入）的组合是否会停机，我们跟它们不一样。

乍看起来，停机问题并不难。既然我们有通用图灵机这一强大的武器，那么只需要用它一步步模拟通用图灵机在输入k上的计算过程就可以了。如果模拟过程在一段时间后停止了，我们当然可以得出“通用图灵机在输入k上会停止”这个结论。问题是，在模拟过程停止之前，我们不可能知道整个计算过程到底是不会停止，还是可能会在3分钟后停止，可能要等上十年八载，更有可能永远都不会停止。换句话说，用模拟的方法，我们只能知道某个程序在某个输入上会停止，但永远不能确定那些不停止的状况。所以说，单纯的模拟是不能解决停机问题的。

我们将编码为n的图灵机称为T_n，则存在一个算法，能够模拟任何其他的图灵机，我们称之为通用图灵机，用UTM表示。其运行性质为，输入数据分两个部分，n和k，$U(n, k)=T_n(k)$。

假设存在这样一个算法H，总能够在有限时间内判定一对（算法，输入）的组合是否停机，并且输出0或1。

$$H(n, k)=\begin{cases}0, & T_n(k) \quad \text{不停机;}\\ 1, & T_n(k) \quad \text{停机;}\end{cases}$$

接下来我们通过将两个算法结合起来生成一个新的算法Q（n，k）：

1. 先通过H（n，k）判定是否停机；

2. 如果停机，则输出Tn（k）；

3. 如果不停机，则输出0。

新算法可以表达为$Q(n, k)=T_n(k) \times H(n, k)=U(n, k) \times H(n,k)$。接下来，定义$T_w(k)=1+Q(k,k)=1+T_k(k) \times H(k, k)$，则当计算$T_w(w)$时，会遇上一个不可调和的矛盾：

$$T_w(w)=1+T_w(w) \times H(w, w)$$

如果Tw（w）会停机，那么最后得到的结果为Tw（w）= 1 + Tw（w）；

如果Tw（w）不会停机，那么会和其定义冲突，因为等式右边的表达式总能在有限时间内停机。

所以不存在算法H，能够在有限时间内判定一对（算法，输入）的组合是否停机。

似乎是个特例，但停机问题涉及的范围可能要比我们想象中的复杂得多。

程序员碰到的最常见的问题就是一个程序写完运行后会陷入死循环。比如：当x>2，让x除以1，如果x一开始就比2大，这个语句就会死循环。这个例子很简单。再举个例子，我们可以编写一个程序*GC*，它遍历所有≥4的偶数，尝试将这样的偶数分成两个质数的和。如果它遇到一个不能被分解为两个质数之和的偶数，它就停机并输出这个偶数；否则，它就一直运行下去。用现代的工具编写一个演算*GC*的程序，对于计算机系的学生最多只能算一次大作业，不是什么困难的事。

从第一个数字4开始，每一个需要被检验的数字都是一个偶数。而每一个小于这一偶数的质数都有特定的判定方法，即用2来除这个数，一直到比这个数都小的自然数都除过了，如果能被整除，那么它就不是一个质数，如果都不能整除，那么它就是个质数。那么这个程序能够停止吗？事实上，如果它真的能停止的话，哥德巴赫猜想就被证否了。因为我们真的找到一个反例，一个偶数，不能被写成两个质数的和。

问题并不复杂，而且对任何现代的计算机语言而言，描述它的算法都很简单。上面的语句描述的是著名的哥德巴赫猜想，三百年来它都是不可证明的。然而，*GC*是否会停止可是牵涉到了哥德巴赫猜想。如果哥德巴赫猜想是正确的，每个大于4的偶数始终都能分解为两个质数之和的话，那么*GC*自然会一直运行下去，不会停机；如果哥德巴赫猜想是错误的话，必定存在一个最小的反例，它不能分解为两个质数之和，而*GC*在遇到这个反例时就会停机。也就是说，*GC*是否永远运行下去，等价于哥德巴赫猜想是否成立。如果我们能判定*GC*是否会停止，那我们就解决了哥德巴赫猜想。程序仅仅检查这个语句“每个大于2的偶数能够用两个质数之和表示”是否为真，然后移到下一个偶数。对所有前4000万亿个数检查下来，它都是正确的，但我们仍然没有得到一个真正的数学证明。关键在于，知道这几行代码是否运行停止就等于是否可以证明哥德巴赫猜想！如果猜想是真则软件不会停止，检查所有直到无穷大的数，都不会发现任何一个破坏这个规则的

数；如果猜想为假，那软件就会在找到第一个破坏规则的数之后停止。

很明显图灵机所执行的程序只是另一个运用逻辑形式系统的方法。我们也能用语言或者软件，甚至更低级别的纯粹的哥德尔风格的配数，如果想的话。然而，基本上，几行语句就能够表示数学上最基础的问题。知道什么时候程序会停止从而给出我们想要的结果，和证明这类难题一样困难。我们甚至可以重新用哥德尔的递归函数来写这个程序，然后写这样一个语句“这个函数在某个数之后停止是否为真”。这同样是又一次证明了这个语句和证明哥德巴赫猜想是一样的。

哥德巴赫猜想目前是数学上一个很著名的不可解问题。或许某天有人会真正找到证明；但也很可能是某人证明这个猜想是不可判定的，即在算术领域里不存在对哥德巴赫猜想的证明。这样的证明不可思议的难，但类似的问题也曾经被证明过。哥德尔不完备定理事实上说明了关于它的证明有可能是不存在的，至少在算术范畴内是不存在的。

数学中的很多猜想，比如说3x+1猜想（任取一个自然数，如果它是偶数，我们就把它除以2，如果它是奇数，我们就把它乘3再加上1。如果反复使用这个变换，我们就会得到一串自然数，这串自然数的最后一个是1）、黎曼猜想等，都可以用类似的方法转化为判断一个程序是否会停止的问题。如果存在一个程序，能判断所有可能的图灵机在所有可能的输入上是否会停止的话，那么只要利用这个程序，我们就能证明一大堆重要的数学猜想。我们可以说，停机问题比所有这些猜想更难更复杂，因为这些困难的数学猜想都不过是一般的停机问题的一个特例。如果停机问题可以被完全解决，我们能写出一个程序来判断任意图灵机是否会停机的话，那么相当多的数学家都要丢饭碗了。

然而，图灵证明了判别通用图灵机运行是否可以停机的图灵机并不存在，图灵机停机问题是个不可计算问题。这同时说明了在数学和

信息学中，并不存在一个终极理论或者算法来解决所有问题，总有些问题是不可计算的。图灵对于停机问题不可解决的证明是决定性的。没有一个程序仅靠检查另一个程序的源代码，就能够决定它是会停止运行，还是会永远运行下去的问题。这是从图灵机的角度对哥德尔证明的重新表述。这个证明与哥德尔的证明是相同的，只是更容易理解，这个证明可以表示成简单的代码，而不是令人费解的运行在《数学原理》逻辑语法上的函数。

而哥德尔第二不完备定理告诉我们，不存在一个通用算法来判别哪些问题是自洽的。如果一个语句是不可判定的，你不能作弊，只能写一个程序去检查它。程序虽然可以轻易表达这个陈述，但它可能会用无尽的时间去完成它的工作。只要语句的不可判定无法证明，你就证明不了程序是不是会停止。

停机问题如此复杂，机械的计算看起来没有足够的力量来完全解决它。但要想严格证明这个结论，似乎仍要求助于深藏在哥德尔不完备定理的证明中，那魔法般的自指。图灵机版本的关于不完备的证明，会让这种自指的复杂性扩散到图灵机所覆盖的所有问题中，即，任何可描述算法。这是个细思极恐的推论，因为我们还有个策略是，像鸵鸟一样明知道哥德尔不完备定理无法避免，但仍幻想它毕竟指的是极少数特殊情况下的问题，只要视而不见就可以了。但哥德尔不完备定理通过图灵机延伸到所有可描述问题，即，我们可以了解并说清楚的所有问题；不可描述的问题，是我们的未知问题，已知问题，我们都可以描述了。这样，哥德尔的不完备定理，也随着图灵机延伸到我们用现代计算机所涉猎的所有问题，这当然也包括人工智能。

*NP*不可解问题

1956年，哥德尔在给冯·诺依曼的信里讲到他对计算复杂性的认识。

哥德尔说，对命题证明中的每个公式F和每个自然数n，人们很容易就能构造图灵机来判定F是否有长度为N的证明。给定这样一个机器，人们就能定义一个F和n的函数ψ，使得ψ（F，n）是机器判定所需的步骤N，对于F而言，N增长有多快？如果增长是n或n^2，那将非常有意义。哥德尔对这个问题的认识再次证明了他对什么是最基本问题的直觉把握，这事实上是计算复杂性问题的最早表述，也成为计算机科学的另一个核心未解的问题。

对于一个算法，人们能够分析出运算时间与数据量之间的大致函数关系，这个关系被称为时间复杂度，它定量描述了该算法的运行时间。

假设有n个数要排序。一个初级的冒泡排序算法所需时间与n^2成正比，快一点的算法所需时间与$n\log$（n）成正比。最不实用的算法就是输入的数字随机排列，直到出现完全有序的情况为止，这被戏称为“猴子排序法”（*Bogo-sort*）。前两个算法的时间复杂度分别记为O（n^2）、O（$n\log n$），最后一种平均时间复杂度则达到了O（$n\cdot n!$）。前两种算法的复杂度是n的多项式函数，最后一种算法的复杂度是n的阶乘。当n特别小时，多项式级的算法已经快过指数级的算法；当n非常大时，人类根本看不到指数级复杂度算法结束的那天。如果问题存在多项式级复杂度的算法，这个问题就是P（polynomial，即多项式）问题。这意味着，即使面对大规模数据，人们也能相对容易地得到一个解。对应的，NP问题，不确定为是否是P问题的问题。NP问题指的是，能在多项式时间内检验一个解是否正确的问题，但找到这个解通常需要很长时间和计算量。比如电子邮箱账户的密码，能在多项式时间内验证一个字符串是否等于这个密码，但需要跟密码位数l成K^l数量级的猜测（K是键盘上的有效字符数），所以“破译密码”是一个NP问题。

图7－10中最小的一个框*P*代表的是传统计算机所能快速解决的一系列问题，而*NP*则是传统计算机不一定能够迅速解决，但是如果提供一个答案，就可以快速验证答案的所有问题。而*NP*继续向外延伸是*PH*（polynomial hierarchy）问题，也就是所谓的“多项式层次结构”问题，它可以被理解为传统计算机在现在和未来可能解决的所有问题的总和。我们将在下一章里结合量子计算来说明量子*NP*问题。数学中一个著名的问题在于证明*NP*问题是否最终等价于某一类*P*问题。如果一旦证明它们是等价的，即有些现在看起来困难的问题迟早能找到有效的解。这绝不是个好消息，如果*P*问题等于*NP*问题了，很多基于计算复杂性的算法，比如通用密码或区块链都会不再安全。越来越多的科学家相信它们不等价。但事实上最近的神经网络算法确实使得一些以前我们认为的*NP*问题有了更加简洁的解法。比如我们曾经认为围棋是个*NP*问题，但事实上，AlphaGo在一定程度上给出了一个相对简单的解法。

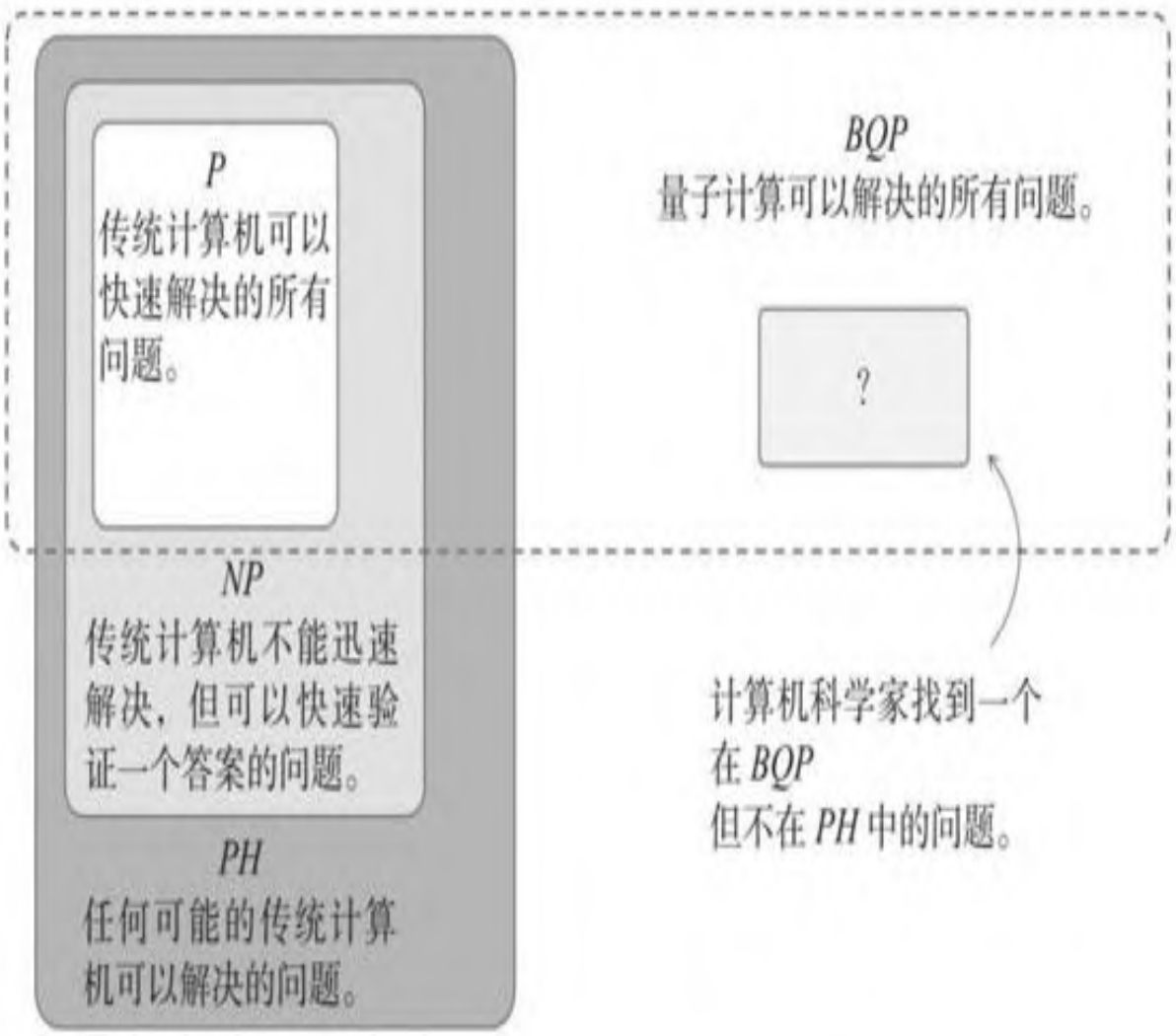


图7-10 计算复杂性的分类

1993年计算机科学家伊桑·伯恩斯坦（Ethan Bernstein）和优曼许·沃兹内尼（Umesh Vazirani）为“有界误差量子多项式时间”定义了一个新的复杂度等级，他们称之为*BQP*问题。他们定义的类包含量子计算机可以有效解决的所有决策问题。同时他们也证明了量子计算机可以解决传统计算机可以解决的所有问题。也就是说，*BQP*包含*P*中的所有问题。

但是人们无法确定*BQP*是否包含*NP*问题。但如果*P*问题被证明在可解决的问题类上等价于*NP*问题，而不考虑其复杂度问题，则可以将*NP*看作是传统计算机可以解决的所有问题的类。换言之，比较*BQP*和*NP*是为了最终判定量子计算机是否优于传统计算机，即经典图灵机。

理想随机数

经典计算机不擅长的一件事是像我们一样翻硬币来生成一个理想随机数，看起来这么简单的一件事计算机居然做不到？然而事实上，计算机所产生的随机数字是确定的。这意味着如果你问同样的问题，你每次都会得到相同的答案。通过算法可以对机器进行编程以生成所谓的“随机”数字，但机器始终受编程的支配。实际上，机器经过特殊和仔细的编程，总可以消除结果的随机性。在一台完全确定的机器上，人们并不能生成任何可以被真正称为随机数字序列的东西，因为机器遵循某种确定的算法来生成它们。通常情况下，这意味着它以一个共同的种子数字开始，然后遵循一个程序。结果可能足够复杂，难以识别模式，但由于它是由一个精心定义的算法规划的数字，产生的并不是真正的随机。因此计算机所产生的随机数也被称为“伪随机数”。一个伪随机数发生器需要一个算法来生成第i个随机数*Rand*（*i*）和一个初始值“种子”*N*0。冯·诺依曼在1946年提出了一个简单的例子：开始取一个2*s*位的整数*N*0，称为种子，将其平方，得4*s*位整数（不足4*s*位时高位补0），然后取此4*s*位的中间2*s*位作为下一个种子数。以此类推，即可得到一系列随机数。这样，当*N*0=1234时，产生的序列如表7－4所示：

表7–4

平方项	取中随机数
$N_0^2=01522756$	$N_1=5227$
$N_1^2=27321529$	$N_2=3215$
$N_2^2=10336225$	$N_3=3362$
$N_3^2=11303044$	$N_4=3030$
$N_4^2=09180900$	$N_5=1809$
$N_5^2=03272481$	$N_6=2724$

如果把生成的随机数除以10000，这样就得到一个介于［0，1］的随机序列：

0.5227，0.3215，0.3362，0.3030，0.1809，0.2724……

冯·诺依曼的算法明显有很大弊端，不仅周期短而且分布不均匀，比如10000平方取中结果就一直为00000了。随后有很多深入的研究给出了更好的伪随机函数发生器。

完全基于确定性逻辑的伪随机数发生器永远不会被视为纯粹意义上的“真正的”随机数源，但在实践中它们通常足以满足要求严格的安全关键应用。实际上，精心设计和实现的伪随机数生成器可以用于安全关键加密的目的，大多数计算机生成的随机数使用伪随机数生成

器，这些算法可以自动创建具有良好随机属性的长数字序列，但最终受内存使用量的限制而回归为可重复的序列，因而也不是理想随机数。这些随机数在很多情况下都很好，但还不像用作随机性源头的电磁大气噪声产生的数字那样随机。对于用计算机来模拟的大多数应用场景来说，使用伪随机数是足够的。例如，如果想对大量数据进行随机抽样，只需要将随机数字输入到程序中，使选取样本的位置或多或少地均匀分布。在这种情况下使用伪随机数是完全可以接受的。iPod（苹果公司推出的便携式多功能数字多媒体播放器）“随机”模式下的歌单也是在伪随机顺序下播放的，它的循环模式是可预测的，但这并不妨碍我们使用这个模式享受随机聆听喜爱旋律的美妙。

通过适当的设计，计算机也不是完全不能跳出这个限制。与物理世界的真实随机相结合，机器可以通过多种方式生成真正的随机数。真正的随机数在实际应用中的重要性不容小觑。例如，一个在线扑克网站，如果黑客知道随机数的生成算法和随机种子，就可以编写一个程序来预测将要见到的扑克牌。真正的随机数使这种黑客攻击变得不可能。我们可以依赖自然界来制造真实的随机数，一种方法是测量一些预期随机的物理现象，包括测量大气噪声、热噪声、其他外部电磁和量子现象。例如，在短时间尺度上测量的宇宙背景辐射、放射性衰变和电路所产生的热噪音，经过一定的放大和截取，也能获得真正的随机数。

从天然来源获得随机数速度取决于被测量的物理现象本身。自然发生的“真实”随机数有时候会很慢，我们不得不花费大量时间来收集大量数据，直到收集足够的随机性来满足需求。也可以采用混合方法，在可用时提供从自然源收集的随机性，把这些随机数用作随机种子而触发基于软件的伪随机数发生器。这种方法避免了基于较慢和纯环境方法的随机数发生器的速率限制阻塞行为。利用机器时钟作为随机函数发生器的种子，从而避免相同种子产生相同的随机数，就是这种策略的例子。

混沌

混沌是很早被经典物理和数学发现，但多少被刻意忽略掉的一套理论。从经典数学的角度上看它不那么干净，它告诉你有些东西天生是不确定的，它说小概率可以有大影响。这些是牛顿力学和经典物理不愿意多谈的，虽然经典力学在解决多体问题中早就看到了混沌的存在，但它真正登堂入室是在20世纪60年代之后。尤其是计算机技术发展促使数值运算能力大大加强，让人有信心以模型为基础，利用计算机的数值模拟重现和预测任何现实的事件。然而20世纪60年代人们在对天气的模拟中发现，即使在确定模型的基础上，也能产生非线性的不确定的结果，从而不能准确预言它的结果。

南美洲的一只蝴蝶扇动翅膀也许会引起北美洲的一场风暴，理论上这是可能的。美国气象学家、麻省理工学院教授爱德华·洛伦茨（Edward Lorenz）长期研究天气预报问题，他在计算机上用一组简化模型来模拟天气的演变。他原计划利用计算机的高速运算来提高天气预报的准确性。但是，事与愿违，他的多次计算表明，初始条件的极微小的差异，都会导致计算结果很大的不同。1963年的一天，为了在验算结果的时候省点事，他把输入值0.506127改写为0.506再算一遍。两个数值相差不大，按理说算出来的值也不会相差太大，然而，每一步的计算结果的差异很快以指数形式增长，以至于两个初始值相差不大的输入，没多久以后输出的结果居然有天壤之别。

洛伦茨使用的方程式如下：

$$\frac{dx}{dt} = \sigma(y - x);$$

$$\frac{dy}{dt} = x(\rho - z) - y;$$

$$\frac{dz}{dt} = xy - \beta z.$$

这一组方程在计算机里模拟气流的运动，结果对应的点画出来看着很像一只蝴蝶。洛伦茨因此开创了一门新的学科：Chaos，混沌学。有人称它为21世纪继相对论和量子力学后的第三大科学发现。但之后的发展说明，这样的说法多少有些言过其实了。混沌本身的科学魅力在于：长期行为对初始条件有敏感的依赖性。初始条件产生的微小变化被不断放大，对未来将造成巨大影响。正如中国古书所载，“失之毫厘，谬以千里”。蝴蝶效应会受到许多其他因素的干扰，“蝴蝶扇翅”与“风暴来临”绝不是简单的直接因果，而可能是复杂的连续因果。同样的，古代谚语有“钉子缺，蹄铁卸；蹄铁卸，战马蹶；战马蹶，骑士绝；骑土绝，战事折；战事折，国家灭”。这也是非线性放大的例子。

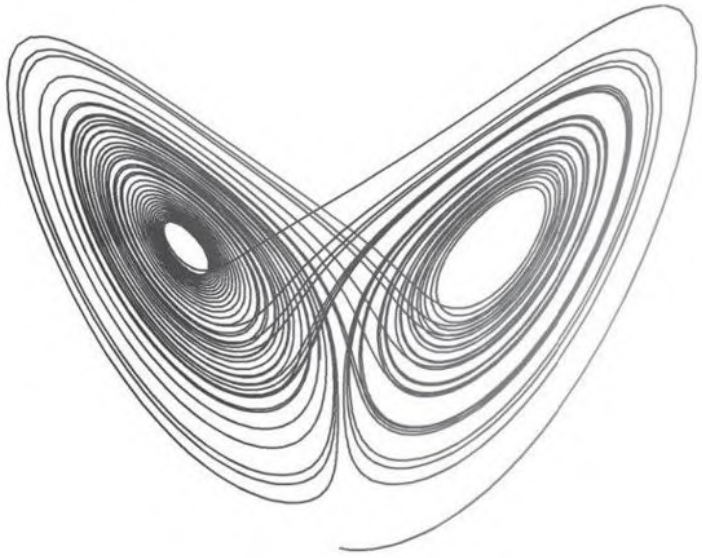


图7-11 洛伦茨发现的混沌效应

虽然混沌的发现和计算机息息相关，混沌现象也揭示了一类计算机无法完成的问题。动力学本身具有不可预测性，任何小的差别在演化了一段时间之后所导致的结果会完全不同。有限容量的计算机在表达数字的时候，不得不采取近似，忽略掉舍去微小的小数点后几位之后的不同，但这点不同会造成大相径庭的结果，而到底哪个会更接近真相，计算机是无法给出判断的。因此，一个混沌系统不能用计算机来精确地计算它长期演化的准确结果，计算值与真实会相差很大。研究表明，这一切不确定性都是系统内在的动力学性质，跟测量精度没太大关系。由此物理学很快发展出来非线性的混沌理论，延伸至分形数学、耗散系统和自组织行为。

量子随机数

创作了“思考者”的艺术家罗丹（Auguste Rodin）说，雕塑作品早就埋在了石头里，我只是负责把它挖出来。这句话引出来一个重要的话题，什么是自由创作，或者什么是自由意志？

- 我们是否因为自己的自由意志创造了这个作品？
- 或者那个作品已经存在了，我只是去除了它周围的石头？

如果是真的，这个作品是注定的。

我们无法通过检查证据回答这个问题。在哲学术语中，它“在经验上是无法回答的”。这是哲学家们辩论的东西。宗教人士可能会指出经文支持其中一个结论而不是另一个结论。然而，在物理学中，量子随机性为自由意志提供了启示，甚至它本身可能就是我们具有自由意识的源头。

牛顿给出了经典物理中运动和力、重力定律和其他物理原理的优雅描述，从而去诠释上帝的伟大设计。人们相信，一旦知道了物体此刻的运动状态和受力情况，就可以准确地预言它未来所有的运动状态。小到苹果大到行星，都可以精确地被把控在人类所掌握的数学工具里。在经典物理世界（牛顿物理学）中，可以认为随机性甚至是不存在的。当扔起一枚硬币时，正面或背面的机会被认为是各50%。但我们知道这样的概率统计只是一种方便的模型，它给了我们大概的主意，假设硬币两面均匀，没有其他任何干预，它应该发生什么。现实生活中的硬币在空中的翻转受制于人们用拇指将硬币推入空中的力量，风和气流的存在，以及硬币两面的物理特性，等等，这些因素都是确定的，硬币的运动因此可以由经典物理定律来描述。可以说，所有经典物理所说的随机性，只要经过足够的数据仔细分析，以及有足够的数据采集，精确的数学模型和强大的运算能力，就是具有确定性的。

爱因斯坦因此不喜欢量子力学声称的纯粹随机性。对此他有一个著名的说法：“上帝不掷骰子。”爱因斯坦认为量子系统坍缩到某一个状态上的概率必须像掷硬币一样。在翻转硬币之前，我们只能通过概率来推测结果，正面或反面。在扔到空中之后，结果是由环境来决定的，通过好的数学工具和足够的预先测量，我们总会知道确定性结果。在实践中，如果我们足够多次地翻转硬币，我们将得到一半正面一半反面的结果。这不是因为环境是确定的，而是因为确定的环境不会让硬币的任何一面有优势。

爱因斯坦认为量子力学也必须符合类似的规律。宇宙中存在更深层次的未知规律，必然应该有一套更好的理论来解释量子力学的所有问题，并且预测量子力学所不能预测的新东西，他将其归纳为隐函数理论。因为我们不知道它们是什么，我们能做的最好就是应用概率模型来解释我们的观察结果。但我们怎么能确定是否有我们尚且不知道的隐藏的东西呢？

那么问题就来了，我们所做的事情其发生是注定的，还是由自由意志选择的？当量子波函数坍缩时，它是否注定会因为一个未知的潜在原因而坍缩，正如爱因斯坦所认为的那样？这样的坍缩是否与其他主观和外在的任何影响无关？比如观测者自己的行为？这些问题都启发着我们对自然的自由意志的含义的思考。如果它是伪随机的，那么我们今天的意识是不是很久很久以前就可以被计算和预测，比如此时此刻在想的事情通过足够多的原子在几千年前的状态下就可以计算出来，我们姑且忽略计算复杂性的问题。如果量子随机真的在底层起着作用，那这些行为将永远无法预测，我们是具有自由意识的。

20世纪60年代，英国工程物理学家贝尔（John Bell）提出了检验量子随机性的杀手级证明。贝尔定理最终证实了量子随机没有被更深层次的规律决定。大自然不可预测的自由行为不由先验的可追溯的算法来决定。通过量子坍缩，宇宙不断产生新的信息。在贝尔定理下，经典力学所描述的翻转硬币的机制不再适用于量子力学。在量子世界中，我可以在完全相同的条件下翻转相同的硬币，有时会是正面，有时会是背面，宇宙中没有任何因素决定其随机的结果。

那么，这是说量子随机性导致了我们思维的独立和自由吗？虽然量子随机性作为解释我们的大脑如何工作的理由依然是个大胆的甚至是证据不足的推测，但它确实启发了我们在密码学方面一些有价值的应用，例如它使在线金融交易更为安全。伪随机数生成器生成随机数字串作为交易代码源，甚至可以通过随机性安全测试。但是，如果用于生成这些数字的确定性规则是可以做逆向工程的，即一旦被人知道生成随机数的算法和种子，就可以被破解或者被准确预测。如果系统被黑客入侵，这些信息就会被获取，从而无法保证它的可靠性。相比较而言，因为量子世界确实是随机的，量子随机数发生器将是一种潜在的解决方案。这个一点也不抽象，当你在网上转账时，你的手机会从银行收到一串交易代码来保证转账的可靠性，但这串代码，迄今都还是由伪随机函数生成的，当一个有足够能力的黑客来破解的时候，他可以确定地知道这个数值，从而威胁交易的安全。

牛津大学物理学家罗杰·彭罗斯认为，人类思维的非算法（不可计算）特征，如创造力，源于大脑微管中的量子坍塌。如果他是正确的，我们的大脑正在创造真正新鲜的新信息。但是有一个问题：量子坍缩产生的信息只是随机的，如果没有一定意义和程度上的关联，它们就是随意且毫无用处的。仅凭随机性这些信息无法在创造性思维中产生明显的复杂性，在我们的神经元中随机的嗡嗡作响不会解决哪怕最简单的数学问题或写出一部伟大的小说。会有一群能在打字机上打出莎士比亚著作的猴子吗？只要猴子够多、时间够长，数学上总有机会，但物理上我们没法让大脑的每一个原子的状态都随心所欲地坍缩到什么态上，而我们的行为和意识被这些随机的不良结果控制。要知道只有那一部莎士比亚是有用的，其他的都是废物，而产生这一部的概率是非常非常小的。因此要形成有意义的自由想法和意识，大脑必须为一般目的制定或组织量子坍塌形成的新信息，我们仍然没法解答人的思维中创造力是如何发生的这一问题。我不认为大脑是单纯地以达尔文方式进化出来的，简单机械系统产生不出大脑，生命力是宇宙中元素的一种组织形式，它遵循特定的作用规律，这些规律既不是简单的也不是机械的。

常识？

我们现在知道了：经典理性是不能将所有人类的数学直觉形式化的。如果我们能够成功地形式化它的一部分，这事实上就恰好需要一种新的直觉知识，例如，这种形式逻辑上自洽的知识。人类通过对数据的截断来获得常识的机制和图灵机的停机问题，大概有深邃的关联，人可以知道什么数据可以形成足够有用的常识，而不是无穷无尽地引入新的概念。对于计算机的搜索能力来说，它显然无法得出到底哪里是我们认知获得满足的边界。人类会对自己周围的世界进行逻辑

推理，我们在为自己看到的东西建立模型，我们有大量的常识知识来帮助我们发现新的情况。这种建立模型的能力是从哪里来的？是从具体事物和数据中抽象出来的某一线索？更重要的是，人工智能和它们的具象——计算机缺乏常识。

1950年，图灵预言了具有智能的机器会被创造出来的可能性，为了鉴别这种智能，他提出了图灵测试：如果一台机器能够与人类展开对话而人不能辨别出机器的身份，这台机器就具有智能。时隔多年，纽约大学的内德·布洛克（Ned Block）对图灵测试能否作为智能标准的适当性提出一些新的论证。布洛克认为，假定我们写出一个树状结构，能够对每个持续一小时的可能对话做出明确描述。显然，这将是一个庞大的树状结构，其容量远远大于现在任何计算机的容量。依这个询问的树结构，计算机将与它的询问者以一种与智能人处事方式难以区别的方式互相作用。如果计算机仅仅按照这个简单的树状结构执行，这就明显说明了这台计算机没有任何智能，而且这个结论对于任何有穷长时间的对话都成立。这也是自然语言处理的困难所在，机器无法在连续的对话中了解人类对话的常识，每一次对话中都会导向很多不同的选择。逻辑系统无法建立对语言的基本直觉，它只不过是抽象对象与实在对象之间的一一配对。语言对抓住我们的思想是有用的、必要的，但那纯粹是在实践中形成的，是过程化的。我们的思维更倾向感官的、概念上的对象，而语言帮助我们将注意力集中在抽象的对象上。

汉堡王曾经有一则电视广告，广告结尾的一句台词说：“OK，Google，what’s the whoppers？”但“Ok，Google”是安卓手机的Google Home唤醒词，这句话前半部分可以启动安卓设备，后半部分提出一个搜索的问题。观看广告的人会发现自己虽然啥都没干，自己的手机或者智能音箱神经病一样地启动了，机器自动搜索汉堡王关于whoppers汉堡的广告并语音播报搜索结果，简直像手机被黑客入侵了一样。这则广告的创意在于视频广告本身结束了，但智能手机还继续帮着播广告。对智能手机或者智能音箱而言，它无法区分来自不同层

次语义的内容。虽然通过适当的声纹识别可以避免这样的问题出现，但新加入的策略一样会带来其他新的问题。

美国哲学家约翰·瑟尔（John Searle）在图灵测试的基础上构想了“中文屋”的思想实验。将一个不懂中文，只会说英语的人关在一个封闭房间中，房间里有一本英文的小册子告知他如何处理相应的中文信息。中文问题从窗户被递进房间里，房间里的人需要对照手册进行查找，将中文字符拼到一起，写在纸上并递出去。房间外的人看到纸条，可能会觉得房间里的人懂中文，实际上他对中文一窍不通。

从这个实验看，这本小册子就是计算机程序，房间就是计算机。计算机给出的回答是按照程序的指示进行的，但它不能也不需要理解中文。现在很多机器语义理解是固定模式识别，它们可以理解最简单的小册子，再根据用户对话中特定的词做出特定的反应。训练机器来理解语义就类似于这个过程。通过训练我们让机器的反应接近于能够理解，但如何期待机器能理解？来说说具体的问题。语义理解的问题至少有切割词、歧义和语境处理等诸多问题。

瑟尔通过“中文屋”概念质疑了强人工智能所持的观点：计算机可以精确地具有人类理解故事和解答相关问题的能力。在瑟尔看来，计算机的理解力与汽车的理解力没有什么不同。计算机与人类的心智相比，其理解力不仅是不完全的，可以说完全是一个空白。当然，重要的不是要论证“计算机能不能具备思维能力”，而是回答“正确的输入输出加上正确的计算本身是否足以保证思维的存在这一问题”。他认为，如果我们所说的机器是指一个具有某种功能的物理系统，或者它只讲数字或逻辑运算的话，那么大脑确实就是一台计算机，然而在他看来，智能的本质并非如此。计算机程序纯粹是按照语法规则来定义的，而语法本身不足以担保心的意向性和语义的呈现，程序的运行只具有在机器运行时产生下一步形式化的能力，只有那些使用计算机并给计算机一定输入，同时还能解释输出的人才具有意义性。意义

性是人类智能的功能，智能的本质绝不能被程序化，也就是说，智能的本质不是算法的。

这一轮人工智能浪潮以深度学习为主要驱动力。在深度学习成为计算机算法主流以来的几年里，它已经成为帮助机器感知和识别周围世界的主要方式。它为语音识别、自动驾驶汽车和图像认知等方面提供了动力。深度学习系统是像素级模式识别的高手，但它们无法理解“模式”的含义，更不用说对它们进行推理了。比如基于深度学习的视觉识别系统，只要改变一些输入，系统就会很容易被愚弄。我们尚未清楚目前的人工智能系统是否能够理解沙发和椅子是用来坐的，当然，这也需要足够的工作来定义“理解”。我们需要开始弄清楚如何让人工智能具备常识，我们这样做的同时，将不断触及深度学习的极限。

人类天生就有学习的天赋，能够掌握语言和解释物质世界，而不是一张待读写的硬盘。尽管有很多人认为神经网络是智能的，但它的工作机制与人类大脑有诸多的差别。我们罗列一些来看看。

首先，深度学习工具太需要数据了。在大多数情况下，每个神经网络都需要数千或数百万个样本来学习，从数据训练的角度看，深度学习存在严重的低效率问题。一个孩子学习认识大象，并不需要他的母亲给他看一万张大象的照片，人类通常只需要几个例子就能学习新概念。但是几万遍是深度学习系统通常需要的训练数据量。如果你思考一下动物和婴儿是如何学习的，在生命的最初几分钟、几小时、几天里，动物和婴儿学很多东西都学得很快，这种学习能力看起来似乎是天生的。为了了解世界的物理规律，一个婴儿只需要四处移动他的头，对传入的图像进行数据处理，就能得出结论。

其次，深度学习的算法逻辑不透明，至今为止它都更像一个黑盒子。人们不能观察它整个学习过程，输出的结果难以解释，这直接影响了预测结果的可信度和可接受程度。深度学习算法在实际使用的过程中，往往是通过程序员大量的调试来找到的最好的网络结构，但我

们很难有道理说明为什么最好的神经网络层数是，比如说133层，不是132层也不是134层。

一旦深度学习的网络结构被训练完，它是如何做出决定的就不总是那么清楚了。在很多情况下，即使我们通过算法能够得到正确的答案，不透明也是不被接受的。科学家不喜欢不确定性，尤其是因为无知而导致的不确定性。深度学习在一定意义上是通过反复的实用主义般的参数调整来达到准确的目的，而不是仔细地了解每一个参数所代表的意义。确定性对模型的一般要求是，了解程序中每一个参数的意义和它们的变化规律，而且通过这些变化规律知道变化对最后结果的影响。不透明的参数空间，让人觉得这个工作并没有做完，因为你不知道哪里的一点些许变化会导致结果的完全不同。比如我们前面提到的混沌现象就是典型的例子。即便是经过成千上万次训练，这个先进的系统也很容易被搞糊涂——只需要将一张小小的贴纸放到图像的某个角落，识别结果就会被改变。

深度学习是在一个非常巨大的希尔伯特空间中人为地尝试参数，这个希尔伯特空间可能有上千万个自由变量。一个深度学习的网络，训练的时间往往很长，有可能陷入局部最优化但整体上来说并不是最优的境地，甚至有些训练完全达不到预想的目标。19世纪的法国数学家柯西（Augustin Cauchy）对通过调参数来获得自己想要的结果嗤之以鼻：“给我五个系数，我将画出一头大象；给我六个系数，大象将会摇动尾巴。”当我们有几千万个参数的时候，我们简直可以随心所欲地得到任何我们想知道的结果。用大量权值、阈值来一点点碰，并且需要大量的训练数据以及大量的计算，这些特点使得深度学习的很多性质很模糊。它更像是一个务实主义的工具，只见到它这里那里起了作用，而为什么起作用，没人知道，试出来的不起作用怎么办？换个参数试试。在实际的处理中，参数调整变成一个经常性的工作，因此基于深度学习的人工智能也融入了大量的人力，由人来尝试和决定网络的架构、参数的大小，而这些工作往往是冗长而枯燥的。

再次，目前从算法角度看不出特别好的适应能力，或者说网络的迁移性并不理想。深度神经网络的特性决定了深度学习的能力，在神经网络学习模型建立的时候，就已经确定了所有可能形式。模型一旦确定，它能够学习什么，不能学习什么，再不能超越。这个特性使很多事情成为不可能，每次想让一个神经网络识别一种新项目，你都必须从头开始训练它。一个识别狗的神经网络在识别猫或人类语言方面没有任丝毫用处。这事实上构成了深度学习的实际困难，网络设计得越复杂，准确度越高，它适应新数据库的能力就越差，因此，深度神经网络训练结果的迁移性构成深度学习的核心困难。

目前我们还不清楚哪些途径可以帮助深度学习解决这些问题。人工智能翻译可以将法语翻译成英语，但是它不知道这些话是什么意思。人类不仅掌握语法模式，还掌握语法背后的逻辑。为了能够由此及彼，构建推理能力，人们正在以不同的方式将常识般的结构构建到神经网络中。例如，人们创建了模式识别和深度学习的混合系统——部分是深度学习，部分是更传统的技术，所谓归纳逻辑编程。也有可能是对抗性神经网络，一种相对新的技术，其中一个神经网络试图用虚假数据欺骗另一个神经网络，迫使第二个神经网络发展出极其微妙的图像、声音和其他输入的内部表征。它的优势是没有数据缺乏的问题。你不需要收集数百万个数据来训练神经网络，因为它们是通过相互学习来学习的。类似的方法正在被用来制作那些让人深感不安的伪造视频，在这些视频中，主人公在说或做一些他们没有说过或没有做过的事情，机器可以生成观众无法分辨真假的视频内容。人工智能的弱点越来越多地引起人们的担忧，比如在自动驾驶方面，自动驾驶汽车使用类似的深度学习系统进行导航，已经出现了几起广为人知的死亡事故。

学习

经典认知论将学习理解为稳定的、完整的、实在的、确切表象的被动接受。然而，16世纪以来的经验却在不断地告诉我们，人类的认知已经发展到一种不断的替代性模式。根据这一模式，知识学习在本质上是“由假设所引导的实验”所构成的行动的、操作性的事件。科学认知并不是只有通过获得实在的准确图像才能获取世界表象的过程，应该说，科学将知识理解为语言和控制自然的变化进程中的实践性事件，从而认知过程也成为一种探究的过程。这一探究的过程构成了人类的学习行为，杜威把这种学习行为表述为五个步骤：

1.产生于事实的困惑、混乱与怀疑，人们处于对其全部特征尚未确定的不完整境遇中；

2.推测性的预期——对给定的现象做试探性地解释，以影响某些结果；

3.对手头的所有可定义与说明该问题的理由进行仔细的调查（检验、审查、探测与分析）；

4.对试验性的假说进行详细阐述，以使其更加精确、更加连贯，从而与更大范围内的事实相一致；

5.将所提出的假说视为一种可以应用于现存事态的行为方案，公开采取某种行动以实现预期的结果，并以此检测上面的假说。

杜威认为这样的过程并不一定按照这样的次序进行，也不一定要把所有的步骤都完成才能完成学习的过程。在处理实际问题时，人工智能最擅长的图像感知往往只占数据输入的5%，提供一些线索；而后面的95%，包括功能、因果、动机等是要靠人的想象和推理过程来完成的。我们用这些标准来衡量著名的乌鸦实验，就会看到乌鸦的学习过程是这一学习过程的具体体现。它以某种有待研究和了解的方式，存在于生物体中，这种方式显然不同于至今还依赖于大数据的人工智能方法。

在日本一所大学附近的十字路口，研究人员跟踪和研究了附近一些乌鸦的有趣的行为。



图7-12 乌鸦学习吃坚果的过程

乌鸦想吃坚果，可是它没法砸开坚果的外壳。刚开始，它试着把坚果从天上往下抛（②），可是坚果壳还是不会碎。在某次下抛后，它发现如果抛下的坚果落在路上，正好被车轧过去，坚果就会碎了。乌鸦到路中间去吃坚果是一件很危险的事，因为随时有车子呼啸而过，在吃坚果和保命之间它需要做出选择。但是乌鸦很快发现行人过

马路走斑马线时会先按交通灯，控制交通灯给出信号让车流停下来。这时，它必须进一步认识到人、斑马线、交通灯、车流、马路之间复杂的因果关系。甚至，什么颜色的灯在哪个方向管用、对什么对象管用等。搞清楚这些之后，乌鸦就选择蹲在一条斑马线附近的一根电线上等待机会（③）。它把坚果抛到斑马线上，等车子轧过去（④），然后等到行人按了交通灯，人行横道上的绿灯亮了（⑤），车子都停在斑马线外面，它终于可以从容不迫地飞下去吃地上的坚果肉（⑥）。

乌鸦在这个过程中表现出了相当高的智商，在很多地方远远超过今天我们所训练的任何一个深度神经网络算法。生命只有一次，乌鸦没法像深度学习那样有几乎无限次的训练样本进行尝试。它必须通过很少的次数就发现不能直接在公路上吃坚果。

如果从哥德尔不完备定理的角度看杜威关于学习的五个步骤，我们会看到一些有意思的关联：（1）涉及对环境因素的截断，如何在不完整的环境中选取足够的信息来实施推理，我们根据前面的讨论知道这一点人工智能是无法完成的。乌鸦所在的学习环境相当复杂，如果按照今天机器学习的思路，这是个不可能完成的任务，自然界进入乌鸦视野的事物和层次千差万别。对于要完成的任务——吃到坚果的果肉，哪些是需要掌握的信息，哪些不是，进而哪些是有用的训练，哪些不是，即使可以准确地规划一个行动路径——程序，其计算量都是巨大的。（2）当我们要给定推测性预期的时候，无意要涉及一定意义上的假设，唯有这样的假设先决的存在，才能期望改变这些假设条件下的参数，看到某些预期的结果。而给出假设这件事情对基于形式逻辑的机器而言，还无从谈起。（3）有可能是人工智能最能帮上忙的，即对可定义与说明的问题理由进行仔细的调查，检查、审查、探测与分析。（4）对试验性假说进行详细阐述，使其更精确、更加连贯，这一点似乎对人工智能来说并不算太难，但在更大范围内验证与事实一致，涉及从一个应用场景拓展到更大的应用场景，而我们知道算法的迁移对今天的人工智能来说是件困难的事情。一个数据库集中训练好

的网络，并不能轻易准确地在新的、更大的环境中具有准确的普适性。而更大范围的适用，往往也牵动着很强的关联能力，哪些可以成为更大范围中的元素，这点联想的能力，人工智能也还没有。

关于学习过程的完成（5）是一整套工程的总结，以乌鸦的案例来讲，乌鸦完成吃坚果这一经验的学习，它具有感知、认知、推理、学习和执行的综合能力。它能够在复杂的世界中，认识和归纳出知识，并指导自己的行为。这个过程并不需要大量的数据，学习是一种交互体验的过程，当然这样的学习并不是指我们通常意义上的一类关于记忆力的训练。在记忆力的比较中，我不认为人类会在未来具有任何优势。但这些关于联想素材的收集，又让人类对记忆的需要必不可少，只不过，它不是用传统教育意义上机械式的重复来收集大数据，而是为了足够有效的联想来收集有新意义的素材。结合真实的实践来促成新思想、新的直觉的发现。

让人工智能更加人性化有助于人工智能给我们的世界提供帮助，但是直接复制人类的思维方式呢？没有人清楚这是否有用。我们已经有了像人类一样思考的人，也许智能机器的价值在于它们与我们完全不同。

我们不跟汽车去比谁跑得更快。如果它们有我们没有的能力，那么它们会更有用，将成为智力的放大器。所以在某种程度上，你希望它们拥有非人类形式的智力，希望它们比人类更理性。换句话说，也许让人工智能有足够的理性是值得的。

1994年DNA（脱氧核糖核酸）的发现人弗朗西斯·克里克（Francis Crick）写了一本书叫《惊人的假说：灵魂的科学探索》，谈到“现在是严肃地对待意识问题的时候了”，以往对意识问题、灵魂问题，人们一直讳莫如深，而且多半是“养狗的人就确信狗有灵魂，没养狗的人则否认狗有灵魂”。而在克里克看来，“人的精神活动完全由神经细胞、胶质细胞的行为构成，以及由影响它们的原子、离子和分子的性质所决定”，喜悦、悲伤、记忆、抱负、自我感觉和

自由意志，实际上都只不过是一大群神经元而已。嗯，对，人的大脑不过是一大堆由分子组成的神经元而已。

大脑的许多行为是“突现”的，即这种行为并不存在于像一个个神经元那样的部分之中，大脑只有在很多神经元的复杂相互作用下才能完成如此神奇的工作。训练可以识别物体的神经网络需要大量有标记的示例，这导致人们认为动物必须依靠无监督学习。大多数动物的行为不是聪明的学习算法的结果，而是基于某种尚待研究的在基因组中的编码或是动物的大脑天生具有高度结构化的关联性，这使它们能够快速地学习，或者是一种我们今天还不了解的机制。因此，“从神经元的角度考虑问题，考察它们的内部成分以及它们之间复杂的、出人意料的相互作用的方式，这才是研究意识问题的本质。”许多哲学家和心理学家认为，目前从神经元水平考虑意识问题的时机尚不成熟，然而事实恰恰相反。仅仅用黑箱方法去描述大脑如何工作，特别是用日常语言或数字化编程计算机语言来表达，这种尝试为时尚早。脑的语言是基于神经元之上的。要了解脑，你必须了解神经元，特别是巨大数目的神经元是如何并行工作的。各国都有尝试开展“脑计划”，但2019年停止的欧洲脑计划至少说明了这一尝试目前还有很多未知的困难。

人脑是一个丰富的相互关联的信息的载体，它的许多内容是连续变化的，然而机器却不具有这个特性，它们只能依靠我们通过内省似的类比得到非常有限的体验。那么我们将来能否造出有意识的机器？如果不可能，它们能否至少看上去有意识？有人相信，这最终是可以实现的，但也有人相信其中可能存在我们几乎永远不能克服的技术障碍。短时期之内，我们所能构造的机器就其能力而言，与人脑相比很可能是很简单的。在理解了产生意识的机制以前，我们不大可能设计一个恰当形式的人造机器，也不能得出关于意识的正确结论。人类必将探索新的非图灵机概念来尝试解决当前这一代人工智能所面临的更深层的问题，从而摆脱人工智能在理论和实践上的困境。解决人类智能的极限和人工智能的极限问题，除了与哥德尔不完备定理有关外，

还需要对大脑和计算机更精细的模型做更大胆的研究，而且还需要将学习、问题求解、对策理论与实数论、概率论和几何学知识结合在一起，探索其如何对问题的解决起实质性作用。当然，为什么是列这些名词在一起，它们并不是我从逻辑上选择的，而只是我直觉的选择，甚至是辞藻的堆砌。

第八章

沿着不完备而生长的物理学

我是个物理学工作者，当我回望物理学的发展的时候，我惊奇地发现物理学本身就是不断找到已有的理论框架的悖论而引入新的假设，从而不断发展而完善的例子。换言说，哥德尔不完备定理是物理学这几百年发展的深层脉络，它是人类探索新知识的最好样本。

在希尔伯特为19世纪的数学做总结陈词，并对未来数学的终极工作做出规划的时候，同一时间，物理界对新世纪也是乐观的。19世纪的最后一天，欧洲著名的科学家们欢聚一堂。大会上，英国物理学家J. J. 汤姆森（Joseph John Thomson）发表了面对新世纪的跨年演讲，他在回顾物理学在牛顿之后的200多年中所取得的伟大成就时说，物理大厦已经落成，下个世纪所剩下的只是一些修修补补的工作。他在展望20世纪物理学前景时，不无担心地讲道：“动力理论肯定了热和光是运动的两种方式，现在，它美丽而晴朗的天空却被两朵乌云笼罩了。”

第一朵乌云出现在光的波动理论上，主要是指迈克尔孙－莫雷（Michelson-Morley）实验结果和以太漂移说相矛盾；

第二朵乌云主要是指热学中的能量均分定则在气体比热以及势辐射能谱的理论解释中得出与实验不等的结果，其中尤以黑体辐射理论出现的紫外灾难最为突出。

后来我们知道，第一朵乌云引出了相对论，第二朵乌云引出了量子力学。这是新世纪里颠覆了经典物理学大厦的两套理论。

牛顿的工作经过它之后的理论和实验科学的发展形成了如今我们称之为经典物理学的宏伟建筑。19世纪末的科学家若是相信物理学建立在不可撼动的基础上，将永远屹立不倒，这是那个时代信仰般的常识。经过一代又一代的科学家、数学家和哲学家的努力，牛顿的宏大设计在19世纪末达到了巅峰。经典物理学看上去几乎能够解释自然界的一切：运动物体的动力学中力和运动的相互联系，热学、光学和电磁学几乎解释了人们日常生活中碰到的所有问题。它的内涵是如此深远和广大：从地球上日常经验的事物一直到可见宇宙的最远处，从苹

果到恒星，从发动机到电磁波。理论与实验观测如此吻合，理论对实验的解释又如此的无可置疑。所有人都认为，即使还存在一些遗留的问题，但与经典物理的基本成就相比，这些问题似乎是微不足道的，百里之行已经走了九十九里，终极的大统一理论只有一里之遥。然而在1900年之后，两个微不足道的问题竟然将物理世界掀了个底朝天。

当物理学延伸到光子和原子的层面时，人们对自牛顿以来建立的经典物理产生了信仰危机。量子力学证明经典物理大厦不但可以被撼动，它的基础本身都有问题。牛顿物理学是机械的、决定论的、确定的，其含义似乎没有任何值得怀疑的地方。与此对比，新的量子物理学的特点是非决定论的、交互的和不确定的。在量子力学建立一百多年后的今天，物理学家仍然声称其含义还远不清晰。

20世纪20年代一个学术团体在奥地利首都维也纳形成，被后人称为维也纳学派。学派的成员多是当时欧洲大陆优秀的物理学家、数学家和逻辑学家。他们关注当时自然科学的发展成果，如数学基础论、相对论与量子力学，并尝试在此基础上去探讨哲学和科学方法论等问题。受到19世纪以来德国实证主义传统影响，加上在维特根斯坦《逻辑哲学论》思想的启示下，维也纳学派提出了一系列有别于传统的见解。大致来说，他们的主张有：第一，拒绝形而上学，实验是知识唯一可靠的来源，新的学科往往需要建立在已有的方法上与已知世界的实验相对照；第二，只有通过运用逻辑分析的方法，才可最终解决传统哲学问题。维也纳学派的观点统治了20世纪之后的科学哲学。维也纳学派主张科学是认识真正知识的渠道，而一个科学陈述要有意义，就必须既符合形式逻辑，又是可验证的。这对于物理学理论发展的意义是明显的。

哥德尔不完备定理的发表让我们更加清楚了这样的分界有其天然的原因。自然世界为向哥德尔的不完备诘难提供新假设的源泉，宇宙深处的奥秘永远为新的探索提供“多一个假设”的可能。物理学因此成为人类的理性思维和自然世界接触从而得到验证的第一道界面（参

见彩插10-1）。例如狭义相对论有两条基本假设，光速不变和协变性原理。很多人都想推翻狭义相对论，于是要从光速不变入手，假设光速是可以被超越的。但事实上狭义相对论并没有说光速不可以被超越，只是在这个体系内，光速的不变性解释了为什么牛顿力学在速度接近光速的时候出了明显的问题。接下来有大量的实验证明光速不变这个假设是站得住脚的。人们当然可以不以此为终结，尽可以提出新的假设，构建新的理论体系，但如果没有实验验证和支持，这样的假设就没有实际的意义了。我们永远无法在体系内部去推翻体系本身，因为这个体系内部总不能自己完全证实或证伪。这也就是我们经常开数学家的玩笑，说“你们那是人文科学，而物理学才是自然科学”的深层原因。自然界总会给我们更多的线索与维度来拓宽和检验我们的认识。用给出新的假设成立的证据，来扩大假设体系是建立一个新理论的唯一出路。这又被哥德尔言中，他就是告诉我们人类要不断地这么做，这才是我们认识自然的规律。于是研究科学和数学的人都好开心，吃不完的饭，永世无忧！

落体悖论

两千多年前，亚里士多德认为“物体自由下落的速度和物体的重量成正比，物体越重，下落的速度就越快；物体越轻，下落的速度越慢”。亚里士多德的理论看起来正好与人们的生活经验相符，羽毛比石头下落的速度慢，在长达两千年的时间里都没有人怀疑过这个理论。但是，如果把羽毛和石头绑在一起，下落得更快还是更慢？一方面，绑在一起的组合物体重量更大，应该落得更快；另一方面，落得快的物体会被落得慢的物体拖后腿，比单独下落要慢，两个物体下落的速度应该取平均值。这样亚里士多德的落体理论出现了一个悖论，一个说法合理但用亚里士多德理论无法证明真伪的命题。

16世纪，伽利略的比萨斜塔实验证明：一对同样大小的木球和铅球同时落地。自由下落的物体，下落的速度与它的质量没有关系。它也是惯性定律、自由落体运动定律甚至牛顿万有引力的基础。当然，以我们今天的知识而言，这样的推论并不是那么直接。当我们考虑空气阻力时，可以建立完全不同的动力模型。以空气阻力出发，一样可以建立一个符合实验结果的运动模型。再言之，如果我们是鱼类，生活在水里，再去观察轻和重的物体下落时，阻力就很难被忽略不计，从而极有可能发展出另外一套先引入阻力和重力的力学系统来解释落体的运动。

从历史的记录来看，这段关于比萨斜塔的记述，可能更多的是后人杜撰的。伽利略本人也许并没有在比萨斜塔上做过这个实验，但它仍不失为一个好故事。不管怎样，伽利略的运动定理为牛顿力学的最终建立提供了稳定的基石。接下来的牛顿力学影响了后来几乎全部科学的内容和方法。所以我们有必要仔细看看它会不会也隐藏了回答不了的问题。

光速问题

牛顿力学的公设体系，又称牛顿力学的三定律，表述如下：

第一定律：在没有外力作用的情况下，物体保持静止或匀速直线运动；

第二定律：物体的加速度与所受外力成正比，与其质量成反比；

第三定律：物体的相互作用力大小相等，方向相反。

牛顿运动定律中的第一定律是其他定律的前提和基础，它奠定了经典力学的概念和基础，它默认了存在一个静止均衡的宇宙作为所有运动的参照系。第二定律和与其等价的动量定理，确定了物体运动状态变化与力作用的关系，将有关物体的运动和受力关联起来。第三定律确定了物体与物体的相互作用关系。牛顿同时依靠开普勒行星运动定律深入研究了万有引力，物体质量$M1$和$M2$之间的相互作用力F表达为：

$$F = G\frac{M_1 M_2}{r^2}$$

其中G是万有引力常数。结合万有引力定律，牛顿力学完善了当时的天体力学，使人们第一次对日、月、星辰的运行规律有了准确的了解。牛顿给出了对力的普遍陈述，揭示了两个物体相互作用的规律，为解决力学问题、解释天体运动提供了理论基础。

但事实上，利用牛顿第二定律就可以通过测量物体的运动来定义质量，这是跟运动相关的质量定义，被称为惯性质量。通过万有引力定律，也能定义一个质量，这个是不需要运动的质量，被称为引力质量。万有引力所定义的质量会跟牛顿第二定律定义的质量有严格的对应关系，当参数取的合适的时候，它们之间的比例可以严格的是1：1，这两个不同定义得出来的质量可以严格相等并不那么显而易见。事实上，这个问题一直到广义相对论的提出才有了完全的答案：爱因斯坦干脆把这两个质量等价当作了广义相对论的基本公设。

牛顿力学的建立对人类的认知史有着重要的作用，这一点怎么形容都不夸张。自然界和自然界的规律隐藏在黑暗中，上帝说，让牛顿去吧，于是世界才有了光。但我还是要让读者注意到，在牛顿之前自然哲学已经有了充分的发展，甚至牛顿引以为豪的微积分，都有莱布尼茨来分享他的荣耀。但这并不能掩盖他的光辉。从牛顿之后，人们

对认识世界逐渐有了越来越多的信心。人们通过计算和思考，在自然界中不断验证，做出了以前想都不敢想的预言，居然都精确得很。这无疑增长了人们认识宇宙万物的野心。牛顿方法的延伸，几乎可以说帮助人类建立起了迄今为止的所有现代学科，也重塑了人类对于世界的观点。从此，人们开始有一个近乎信仰的科学认知，有一个客观的世界，是可以不依赖人的直觉而存在的，而且可以被认知、被了解、被描述。

牛顿对物理学的探索涵盖了今天的很多领域，包括他对光的认识。据说他好奇地用手指压迫自己的眼珠（请勿模仿，否则就像雷雨天跑去放风筝发现闪电也是电的富兰克林），发现了太阳光是可以分解成很多不同颜色的，后来利用三棱镜让白光分解成七色光只是一个方便的展示。牛顿认为光是由微粒组成的，这样就可以很轻易地解释它为什么沿着直线传播，就像从枪里发射出来的子弹。牛顿反对光的波动说，他甚至证明如果光是波的话，一个圆盘的影子后面应该因为光的衍射而形成一个个亮圈，而影子的中心是一个亮点。这显然违背他所认为的常识，所以光应该是由粒子组成的。更精细的实验结果就真的做了出来，具有讽刺味道的是，这些因光的波动性而产生的亮环真的存在，而中心真是因为衍射而形成的亮点，这些环被后人称为牛顿环。那么证据确凿无疑，光就真的是波了。

根据人们当时对波的认识，波需要传播介质，人们把光传播的介质叫作以太。人们很快认识到光波是横波，而横波只能在固体中传播，所以以太也应该是固体。这样就带来了一系列的悖论。

横波只能在固体中传播，作为介质的固体越硬，传播速度就越快。声波在木头中的传播速度可以达到每秒500米，在钢铁中就可以达到每秒5000米。而光速居然有每秒30万千米！这样以太应该非常非常非常硬，但，这么硬的东西我们人居然没有卡在里面！而且，我们丝毫感觉不到它的存在，它应该比空气还要轻柔才对。硬或者软？这是关于以太的第一个悖论。

假设以太对于宇宙的绝对坐标系来说是静止的，这样对于宇宙中所有的运动物体而言，以太相对也在运动。已知地球相对于宇宙是运动的，这样地球相对于以太也是运动的。这时的人们已无法想象宇宙跟地球是相对静止的，那样就又回到了太阳和宇宙所有星体和物质都围绕地球转的地心说。牛顿之后的人们已经接受了我们地球没那么特殊的宇宙观：我们所在的地球不过是太阳系中诸多行星中的一个，绝不是宇宙的中心。这样，如果顺着地球运动的方向发射一束光，同时逆着地球运动的方向发射另外一束光，这两束光按照伽利略的运动学原理，速度显然是不一样的。假设地球运动的速度是v_{earth}，光的运动速度是c，那么顺着地球运动方向发射出来的光相对于静止的以太来说速度是$c+v_{earth}$，而逆着地球运动方向发射出来的光相对于以太的速度是$c - v_{earth}$。地球以每秒0.46千米的速度旋转，根据当时美国物理学家迈克耳孙（Albert Michelson）所做的一系列实验，这个速度差是可以观察到的，但迈克耳孙居然穷经白首也没有看到！难道以太相对于地球反而是静止的，这不就意味着地球相对于宇宙也是静止的了？难道又要回到地心说？这是以太的第二个悖论。这两个悖论，在牛顿力学内部无法回答。

孪生子悖论

为了解决光速不变引发的悖论，荷兰物理学家亨德里克·洛伦兹（Hendrik Lorentz）提出了洛伦兹变换，假设以太会沿着光源运动的方向收缩，这样便抵消了不同方向上光速的差异。而爱因斯坦干脆假设光本身就是不变的，根本不需要有以太这种介质来传递光，这样也可以推出洛伦兹变换。但比照洛伦兹的解释来说，爱因斯坦的理论简单、清晰了很多。

爱因斯坦在他1905年的工作中假设没有以太这种东西，而光速在任何一个参照系中都是不变的。于是他构架了后来称之为狭义相对论的理论框架。狭义相对论的公设：

1. 光速在任何一个惯性参照系中的速度都是相同的。
2. 物理定理适合于任何一个参照系。

以哥德尔不完备定理的眼光来审视爱因斯坦关于狭义相对论的工作，就似曾相识了。当我们在欧几里得几何中不能去证明平行线公设可以由其他几个公设推导出来的时候，我们就把它当作一个基本的公设。而不同的关于平行线公设会给出截然不同的几何理论。差别在于，物理学以基于自然现象的实验为基础，不会有这么多不同的公设存在。既然光速测不出来相对于运动物体的变化，那么就把它当作一个基本的公设好了。哥德尔同样告诉我们，这样可以解决现有的问题，但这个公设系统仍然是有限的公设系统，一样会有一些问题是它无法回答的，我们接下来就会看到。

不管怎么说，当我们把光速不变当作一个基本假设的时候，我们会看到伽利略的运动理论作为牛顿力学的基础因之而改变。

物体A相对于地面的运动速度为u，物体B相对于物体A以速度v在运动，那么，物体B相对于地面的运动速度就是：

$$v' = v - u$$

但根据狭义相对论，这一速度表达为：

$$v' = \frac{v - u}{1 - \frac{u}{c^2}v}$$

这样，当在以速度u行进的火车上发射出一束光，这束光相对于火车的速度是光速c，但它相对于地面的速度也是光速c。而由于通常情况下$u < c$，于是伽利略变化就是非常好的近似了。

同样的，狭义相对论也给出了时间的变换公式：

$$t = t_0\sqrt{1 - \frac{u^2}{c^2}}$$

这这样一个运动着的物体所经历的时间会比静止的物体所经历的时间要慢。当我们说地上一年，天上一天，这个天上的坐标系需要以0.9999961倍的光速在运动。虽然看起来会让人有些不习惯，但也还好，直到“李生子悖论”（Twin Paradox）被指出来。

在1905年爱因斯坦的狭义相对论确立以前，牛顿的机械自然观统治着人们的空间想象，因此无法解释这一现象。而“时间相对论化”的确立，取代了牛顿“绝对时间”的概念，使“绝对运动”概念也失去了立足之地。但狭义相对论很快遇到了在自己体系内不能解释的悖论。

孪生子悖论

有孪生兄弟二人。哥哥坐飞船以0.8c的速度离开弟弟所在的地球，过了地球上的10年后回来。这时候哥哥在飞船上实际经历的时间是：

$$t = t_0\sqrt{1-\frac{u^2}{c^2}} = 10\times\sqrt{1-\frac{(0.8c)^2}{c^2}} = 6$$

6年，所以当两个人再见面的时候，去宇宙旅行过的哥哥会比弟弟年轻4岁。但，如果我们稍加思考：运动是相对的，当宇宙飞船相对于地球以0.8c的速度运动，地球也是相对于飞船以0.8c的速度在运动。当兄弟俩再见面的时候，应该是弟弟比哥哥年轻4岁。

到底谁年轻的呢？狭义相对论并不能回答这个问题。这就是著名的“孪生子悖论”。

质量问题

惯性质量和引力质量可以严格地成比例，当参数选择合适，这两个不同方法定义的概念可以完全相等。牛顿意识到了这个问题，但在牛顿体系里并没有办法解释为什么惯性质量等于引力质量。这是两个不同的定义，惯性质量是由牛顿第二定律来定义的，加速度和质量的乘积等于外力。而万有引力（重力）也可以定义质量，是不需要运动的。然而当参数选择合适的时候，这两种不同的方法所定义的质量永远是成正比的。当比例参数选择合适，这两者可以完全相等。对于这个问题在牛顿力学体系内部找不到合理的解释，牛顿说这简直是上帝

的杰作。一直到三百多年后，爱因斯坦干脆把惯性质量和引力质量等效作为广义相对论的一条公设：引力质量就是惯性质量，重力就是空间弯曲的效应。

当我们坐电梯的时候，电梯启动和电梯停止时我们会感到超重或者失重，我们的腿上会觉得有额外的重力增加或者减少。这个完全可以解释为惯性作用，而增加的这些像重力一样的力被称为惯性力，是外力要改变我们或静止或匀速运动的状态时所要施加的力。如果在一个密闭的电梯里，当然，大多数电梯是密闭的，假设地球突然增加了质量，或者减少了质量，比如在我们没有察觉的时候地球跟另外一个大质量的星体融合或者裂成两半等，在电梯里的我们并不知道，是重力增加了还是电梯在加速或减速，从自身感受上我们无法区分这两者的差别，而爱因斯坦认为，引力质量和惯性质量根本是等价的，这才造成了在封闭电梯轿厢中的人无法区别的是电梯被加速还是引力发生了骤然变化。这是一个重要的新假设，因为这样，我们可以用星体的某种曲线运动，把重力从牛顿体系中除去，从而把时间和空间相互等价而关联起来，创造一个新的时空观。

这时不得不引出广义相对论的一个重要工具——非欧几里得几何。在前面我们讲过两千多年来欧几里得的《几何原本》一直是大家公认的数学典范，它不仅符合人们的常识，而且优美对称。后世的许多哲学家，把欧几里得几何学摆在绝对真理典范的地位。在欧几里得几何的五条公设中，唯独的一点小遗憾就是平行公设不够简明，同一平面内，过已知直线外一点有且只有一条直线与已知直线平行，这更像是一条定理。在《几何原本》中，证明前28个命题并没有用到这个公设，这很自然地引起人们考虑：这条啰里啰唆的公设是否可由其他公设推出，也就是说，平行公设可能是多余的。这之后的两千多年，许多人都曾试图证明这点，有些人开始以为自己成功了，但是经过仔细检查后发现，所有的证明都使用了一些其他的假设，而这些假设又可以从平行公设推出来，所以他们只不过得到了一些和平行公设等价的命题罢了。到18世纪，有人开始用反证法来证明，即假设平行公设

不成立，企图由此得出矛盾。他们得出了一些推论，比如“有两条线在无穷远点处相交，而在交点处这两条线有公垂线”等等。在他们看来，这些结论不合情理，因此不可能真实。但是这些推论的含义不清楚，也很难说导出了矛盾，所以不能说由此证明了平行公设。黎曼等人假设无穷远处的平行线可以相交和更加分离，由此创造了曲面几何，发现了全新的几何理论，它们跟欧几里得几何只是假设不同，但可以推导出完全不同却依然自洽的内容。广义相对论的四维时空框架中的几何学就是利用了黎曼几何，而不是通常的欧几里得几何学。在这种几何学中，空间和时间都不是平直的，而是弯曲的。这并不意味着我们的三维空间包含在另外一个多维空间内，而仅仅是指欧几里得几何学定律适用于更小的、局部的时空范围。当时空尺度变大时，空间便会经历一个从欧几里得平面几何学分离出来的过程。这与地球上通常的几何学情形类似，地面通常看起来是平直的，正像欧几里得几何赋予它的性质那样。当我们在足够的距离之外，比如在飞行于400千米高的航天器上看，映入我们眼帘的是地球的曲线。

一个小球在一个空间曲面上运动，当它达到某一速度时会有惯性力让它绕着曲面的中心运动。为了让小球能这样运动，需要一个弯曲的空间。能够造成这个弯曲的空间的是一个质量大的物体，比如太阳。你可以想象地球周围的空间弯曲是太阳和地球质量的共同作用。这种由曲面构成的圆周运动和因为重力构成的圆周运动是完全对应的。



图8-1 小球在曲面上作圆周运动

广义相对论在空间和时间的几何形状与空间和时间内运动物体的行为之间形成了一种广泛的联系。想象一下，把一个玻璃珠放在床垫上。如果有一定速度，玻璃珠将沿直线在床垫上移动。但如果在床垫上放一个保龄球，玻璃珠如果有相同的速度，就会沿着保龄球所形成的曲面滚动，路径从直线变为曲线。保龄球的重量使床垫发生变形，影响了玻璃珠的运动轨迹。用行星、恒星或转动的星系替换保龄球和玻璃珠，床垫换成时空本身。在没有大质量物体的宇宙中，没有空间和时间的变形，两点之间的最短路径是一条直线。当大质量物体出现而改变时空形状的时候，最短的路线将变为曲线。天文学家在1919年的一次日全食观测中证实，太阳的质量可以导致来自遥远恒星的光线弯曲，像爱因斯坦的广义相对论预测的一样。这是广义相对论中所描述的惯性质量和引力质量等价这一假设的结果。

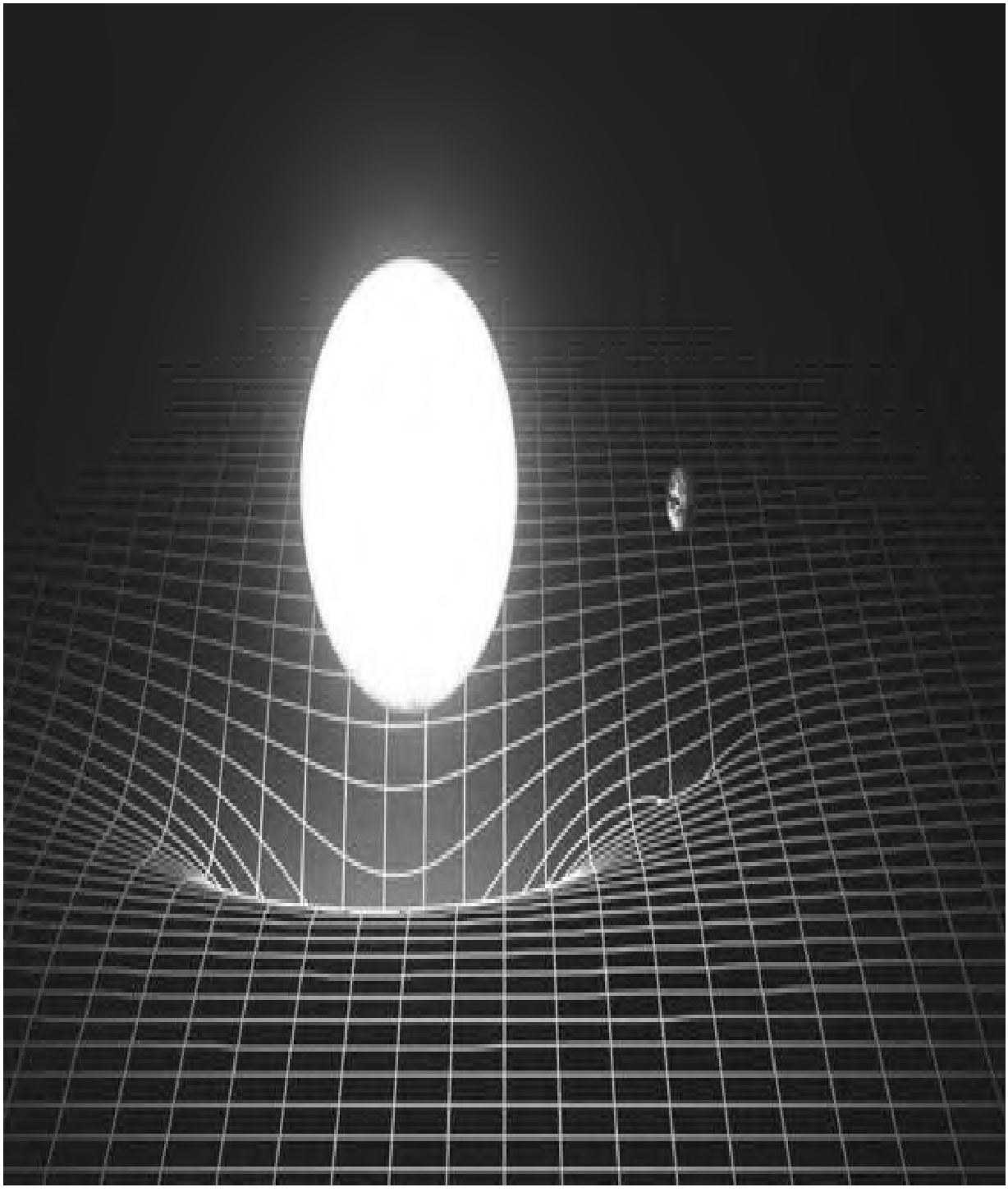


图8–2 大质量物体出现会改变时空形状，两点之间最短的路线将变为曲线[1]

广义相对论导致了空间与时间的融合。按照爱因斯坦的时空观，空间与时间的融合也发生在日常的生活中。例如，在某一确定的地点

（北京天安门）和确定的时间（1949年10月1日下午3点）发生了事件（中华人民共和国开国大典）。这之中有三个数字足以在三维地图上标记北京天安门的空间：经度、纬度和海拔高度。如果添加了另一个数字：时间，那么这个时空中的唯一一点就被确定为时空中的一个事件。如果一个事件可以由四个数字定义，那么一系列事件可以由一系列这样的数字定义，彼此跟随，就像在排队行进。在广义相对论中，这种序列被称为世界线。狭义相对论中所讨论的运动着的时钟会变慢的效应，很快就会导致了“孪生子悖论”。这个问题在狭义相对论的范畴内并不能得到解决，直到爱因斯坦发展出来广义相对论，人们才知道在重力场里面，时间是可以在重力的环境下变慢的。而哥哥能够回到地球上来，是需要先减速然后再反向加速的，这个减速和加速的过程，等价于一个重力场，而这个重力场导致了孪生兄弟中的哥哥所经历的时间绝对变慢。

爱因斯坦晚年曾写道：“过去、现在和未来之间的区别只是一种幻觉。”这是一个忧郁的评论，但它直接来自爱因斯坦的狭义相对论。想象一下，一群观察者在整个宇宙中不经意地散落，每个人都能够将自己生命中的事件组织成一个线性顺序，一条世界线，每个人都相信自己的生活包括一系列从过去到现在再到未来的时刻。分散在各个时空的观察者们都相信他们现在的感觉是普遍的。毕竟，“现在”，就是此时此刻，不是吗？不那么是。时间以不同的速度流逝，这取决于一个人移动的速度：地球上的一小时过去时，对于几乎用光速离开地球的火箭而言，可能就只有几秒钟。这样，一个人的“现在”完全有可能是另一个人的过去或未来。相对论并没有限制时间穿越，困难的是怎样回到过去。

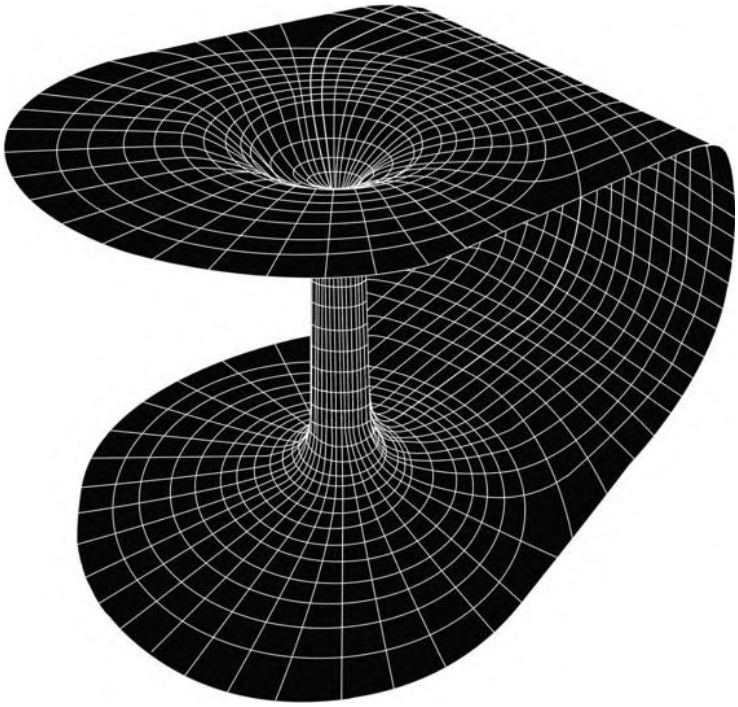


图8–3 弯曲宇宙的虫洞

在时空弯曲模型中，引力被空间的弯曲完全代替，而质量成为空间弯曲的原因。光又扮演了一个神奇的角色，它把时间和空间统一到一起，从而形成一个完整的时空弯曲模型。但一旦推到时空弯曲了，自然而然就会有些好玩的事情。沿着弯曲的表面行进，我们永远无法超越光速，但因为弯曲，我们总有些机会可以在弯曲的对面打个洞，直接跳跃到前面，这叫作虫洞。对于虫洞的探求，也还有很多问题，例如，我们需要多大的能量来维持一个够大的虫洞让人能够钻过去？一旦钻过去我们就回到了过去或未来。到了未来，我们要怎样沿着虫洞回到来时的那个时空点并带回有用的信息？量子力学允许的很多小小的虫洞时刻产生在我们周围，然后又以极快的速度消失。

既然宇宙可以被想象成弯曲的样子，大爆炸也一样，不要问我大爆炸之前宇宙是什么样子，时间轴也是炸出来的。这就像我们在地球的南极点上所指向的所有方向都是北，在地球的北极点上所指向的所

有方向都是南一样，在宇宙诞生的那一瞬间，所有时间都指向未来，并不存在它的过去。

当地球围绕太阳旋转的时候，两者都扭曲着自己周围的时空，而这扭曲并不是静止不变的，这种扭曲会成为某种周期性现象，周期为一年。地球周围的时空确实被周期性运动的大质量扭动了，这样就会形成周期性的时空波动，向更广阔的宇宙中传播，这样就形成了引力波。但对地球和太阳这样的组合而言，其引力波的大小实在是太小了，而且周期也太长，我们没有合适的工具能够测量。但对于质量大很多的黑洞组合，比如两个黑洞互相围着对方在旋转，这个引力波就确实可以在地球上被测量了。可以想象，当设备精度进一步提高，在未来，地球和太阳这样的组合所产生的引力波，也是可以被测量的。由于引力波并不是今天才产生的，到达地球之前它已经走了很久，我们可以通过对这些引力波的观测来了解早期宇宙的形成，从而为宇宙的演化画一幅新的时空图。

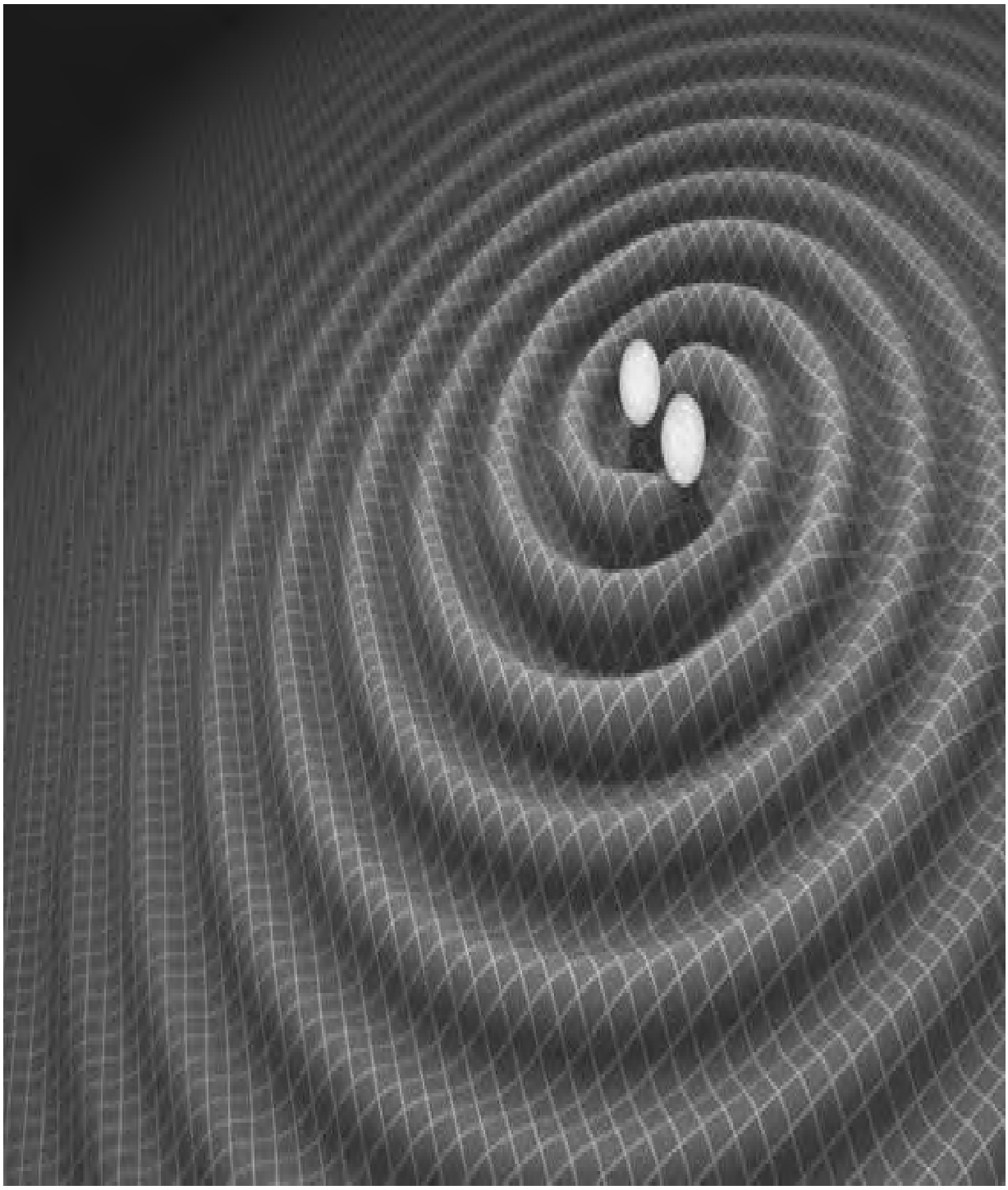


图8–4 周期性的时空波动形成引力波[2]

夜空为什么是暗的？

这是有名的奥伯斯（Heinrich Olbers）悖论：如果宇宙空间无限延展，而且星体均匀分布，我们朝着任何方向的任何视线都应该碰到起码一颗恒星。那么，天空不是应该一直都是明亮的吗？这个结论显然与事实不符。并且这个问题早在1610年就被开普勒注意到了。

正确答案的关键是：宇宙不是无限的，空间上不无限大，时间上也不无限长。

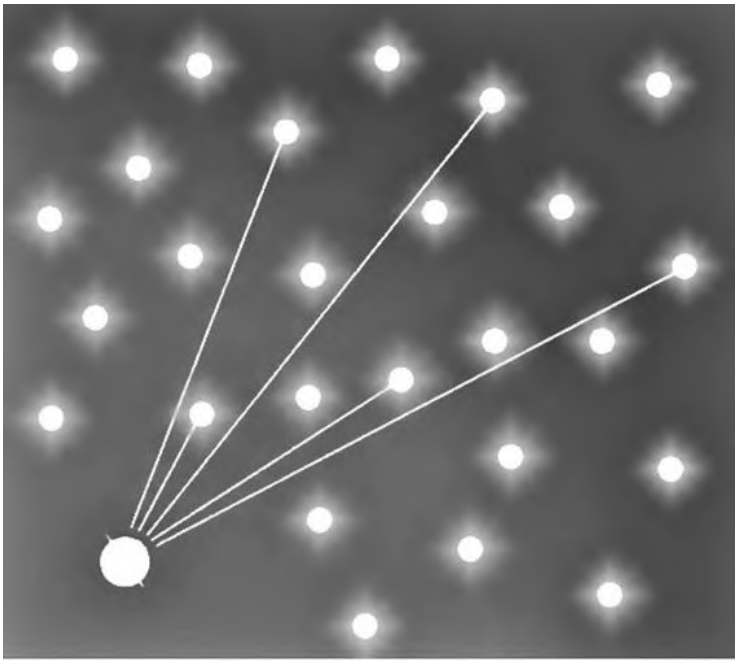


图8-5 来自星体的光线

从遥远星体发出的光到达我们眼球时有一个有限的分界点，到我们仰望星空的时候，从最遥远星星来的光线还来不及到达我们。1901年J. J. 汤姆森给出了一个解释，他认识到，当你遥望夜晚的天空时，你看到的是它过去的样子，而不是此刻的情况。因为尽管按照地球的标准光的速度是非常之快的，每秒30万千米，但它仍然是有限

的，光从遥远的星球到达地球需要时间。开尔文计算得出：要想夜晚天空是白的，宇宙的范围必须扩大到几百万亿光年。但是因为宇宙的年龄没有万亿年，所以夜晚天空一定是黑的。因此宇宙必然是有限大的，并且是有限年龄的。

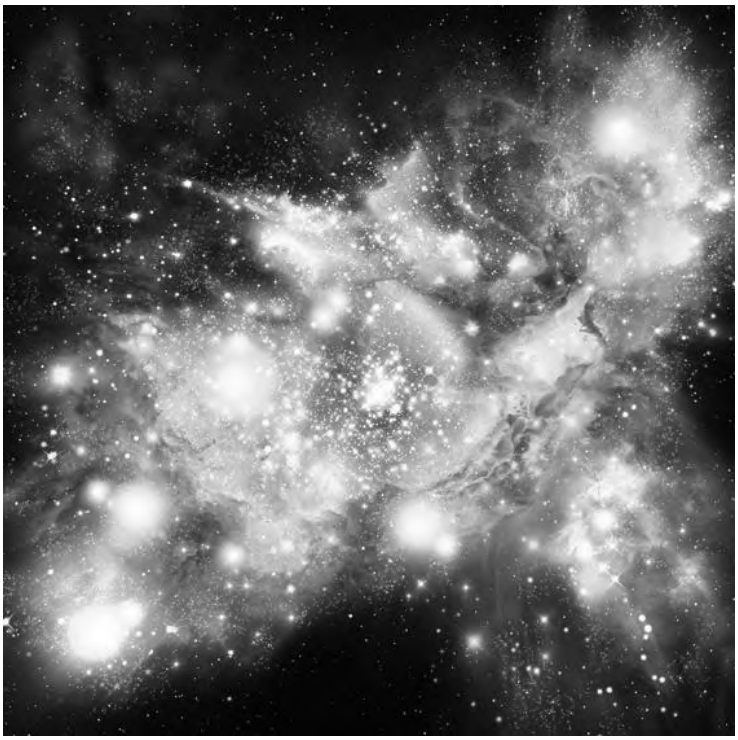


图8–6 猎户座星云恒星诞生区，图中的星光来自120亿年前，在宇宙诞生的十几亿年之后星体刚刚形成[3]

利用哈勃空间望远镜已经可以验证这个答案的正确性。天文学家为哈勃空间望远镜设计了程序以执行一项历史性的任务——拍摄宇宙最深处的照片。为了捕捉最深层空间角落的极其微弱的辐射，该望远镜完成一项前所未有的任务——在总共几百小时的时间内精确地瞄准猎户座附近天空的同一点，这要求该望远镜在围绕地球运转400圈的时间内要完全对准同一区域。此项目是如此之困难，以至于天文学家不得不花费4个月的时间才完成。

如果你还记得我们前面讲的牛顿的万有引力定律和运动学定律，你是可以推导出来在围绕某一横行的行星运动速度是和它距离恒星的距离以及恒星的质量相关的。那么，即使我们看不到星系的中心，因为黑洞质量极大，没有光可以从黑洞所在的星系中心跑出来，但我们依旧可以通过测量星系中某个半径上的恒星围绕星系中心的运动速度来推测星系中心黑洞的质量。这叫作开普勒第三定律，是可以从牛顿力学的公理系统中推导出来的。暗物质存在的一个重要证据来自星云中恒星系旋转速度的研究。利用光谱测量，人们可以探测到远离星系核区域的外围星体绕星系旋转速度和距离的关系。按照万有引力定律，如果星系的质量主要集中在星系核区的可见星体上，星系外围的星体的速度将随着距离增加而减小，但实际观测结果表明，在相当大的范围内星系外围的星体的速度与万有引力定律预测的结果相差很大。

这意味着星系中可能有大量的不可见物质，不仅分布在星系核心区，而且弥散在恒星之间的空间里，且其质量远大于发光星体的质量总和。这些看不见的物质被称为暗物质。20世纪80年代，又有一大批支持暗物质存在的新观测数据，包括观测背景星系团时的引力透镜效应，星系和星团中炽热气体的温度分布，以及宇宙微波背景辐射的各向异性等。暗物质存在这一事实已逐渐被天文学和宇宙学界广泛认可。但它到底是什么？根据已有的证据，暗物质的主要成分不应该是目前已知的任何微观基本粒子。

这同样是回到了哥德尔不完备定理所指示的，新的实验发现不能用已知的知识体系来判别。只要是不完备的系统，我们就欣喜若狂，因为有很多很多不明白的事情可以利用这个找到的不完备来使之完备，从而有能找到更多的不完备。

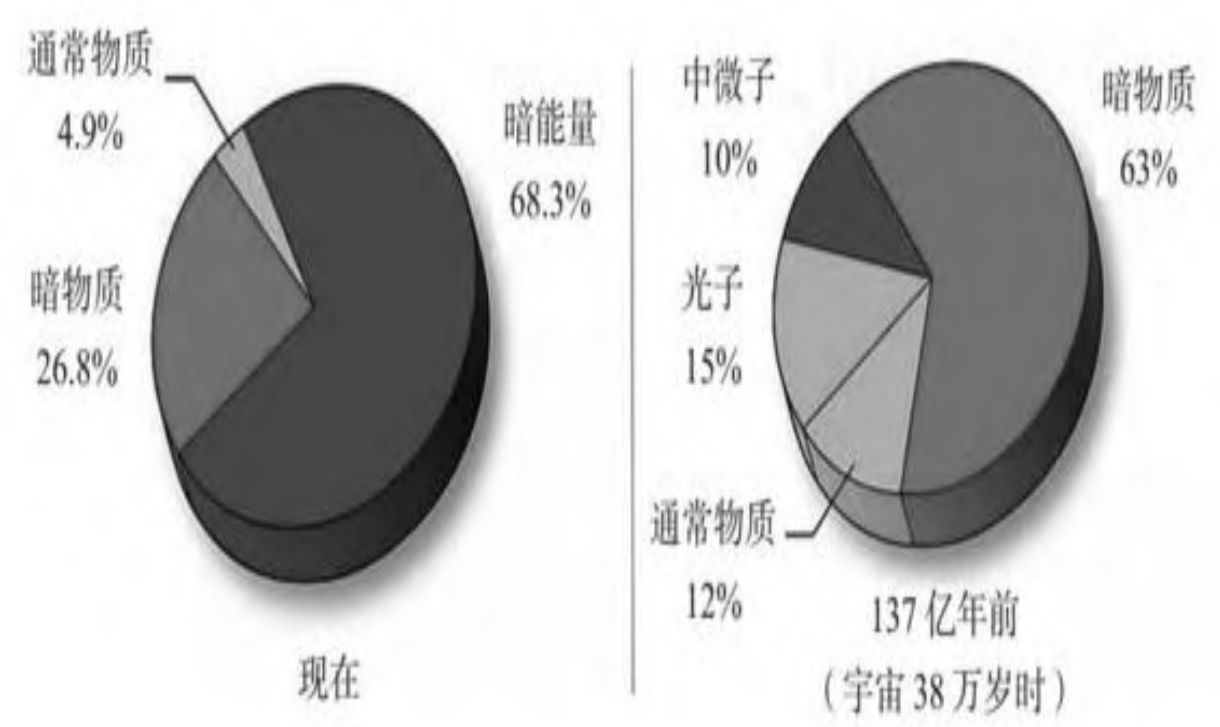


图8-7 宇宙物质构成的变化

既然说到了暗物质，顺便说说暗能量。宇宙因此必然不是无限大，而事实上它正在高速膨胀。它的膨胀又像是在一个膨胀的气球表面，每一点都互相离开，但没有哪一个点是这个球面的中心。这个跟实验上的观测是一致的，我们朝夜空的各个方向看出去的时候，都能看到这些星系在均匀的离开我们，而我们无法接受地球是整个宇宙的中心，所有星体都均匀地离开我们。但这个膨胀的球面，是一个四维的球面，这一点需要一点想象力。

新的问题是，实验上观测到这些星体都在加速离开彼此，因此宇宙是在加速膨胀的，而我们现在的物理理论却并不支持这一点。物质和能量的总量应该是守恒的，这些加速膨胀的星体所获得的能量是从哪里来的呢？是什么能量让它们可以继续克服彼此的万有引力而加速？这些能量我们暂且称为暗能量，我们不知道它从哪里来。从完备

性角度来讲，这又是一个我们不很清楚的缺口，但我们知道一旦它可以被描述，就能带给我们大量探索未知世界的新问题。

麦克斯韦妖的悖论

热力学也有几个基本的公设：

1. 热力学第一公设：能量守恒；
2. 热力学第二公设：封闭系统熵增加；
3. 热力学第三公设：绝对零度达不到；
4. 热力学第零公设：温度相等的物体热平衡。

麦克斯韦（James Maxwell）对热力学第二公设提出了一个悖论。假设我们有一个装着不同分子的小盒子，盒子中间有一个隔板，隔板上面有一个小洞，小洞上装一个开关都不消耗能量无摩擦的门。一开始的时候两种不同的分子充分混合，系统的熵达到最大。这时候有一个小妖来控制这个门，当它看到一种分子从盒子的一端飞过来，就开门让它飞到另外一端，另外一种分子飞过来就关上门把它挡回去。这样只要持续得够久，不需要任何能量，小妖是可以把两种不同的分子分离到箱子的两边。由此，不需要任何能量系统的熵就减少了，而这违背了热力学的第二公设。这个小妖就被后人命名为麦克斯韦妖（Maxwell’s demon），它困扰了物理学家很久。因为宇宙是个封闭系统，如果一个封闭系统的熵只能朝着增加的方向走，总有一天宇宙的熵也会走到最大，那时候所有可能的反应都发生过了，宇宙成为一团没有任何反应发生的物质，一片寂静。这曾经是人们以为的宇宙的

终点，被称作热寂。那么，本来说好的热寂，难道就因为这个虚构的小妖解决了吗？

很多年后，香农解决了这个问题。他说，小妖看到哪一种分子飞过来而决定开不开门这个过程并不是没有代价的，它要消耗能量。得出是哪种分子的结论这需要有能量来支持观测和计算。任何信息处理都是需要能量的，这个能量的最小值，香农定义为：

$$W > -K_B T \Delta S$$

其中，K_B是玻尔兹曼常数，T是温度，ΔS是信息有序性的变化。信息处理是个物理过程，需要消耗能量。从这个公设出发，香农建立了现代信息论——计算机和信息处理理论的另一块基石。

有趣的是，语言的效率在信息论下也可以重新表达。英语作为表音语言，只涉及26个字母的组合，相比较而言，汉字作为表意语言，可以有大量的字符的组合。平均下来，每个英文字符需要2.62比特信息，而每个中文字符只要1.2个比特信息，这也是通常情况下一本中文的译本比英文原著要薄得多的原因。

表8-1 信息论下的不同语种的效率

语言	表达相同内容的字数	效率
英语	1936473	1
西班牙语	1804756	0.932
法语	1896459	0.979
中文	884860	0.457
韩语	1259920	0.651
阿拉伯语	1875204	0.968
日语	1519224	0.785
俄语	1506920	0.778

紫外灾难

量子力学的建立在我的另外一本书《量子大唠嗑》里有更为详细的论述，它源于20世纪之前的物理理论上并不能解释紫外线能量会趋

于无穷大的悖论。进而普朗克引入能量块的解释，干脆把它当作了一个基本假设，即能量总是量子化的，一切豁然开朗。

关于紫外灾难的悖论

19世纪末，物理学家面对黑体辐射问题通过经典物理得出了瑞利－金斯公式（Rayleigh-Jeans Law），它在光的长波段跟实验数据符合得很好。同时，威廉·维恩（William Wien）提出了另外一个公式：维恩公式，这个公式在短波范围内和实验数据相当符合，但在长波范围内偏差较大。瑞利－金斯公式在长波段符合得很好，但当辐射出紫外光的时候，这条曲线会算出来发射体的能量要趋向无穷大，这在物理上是不允许的，人们把它叫作紫外灾难。

虽然普朗克通过拟合得到了一个可以融合维恩公式和瑞利－金斯公式的方案。为了解释这个拟合公式的物理意义，普朗克不得不引入能量块的假设，但他始终无法相信真正的能量是需要切成块的。他勉强解释说这就像是你去商店买了一包糖，商店总要按包卖，你回来吃的时候还是打开一勺一勺地吃，这里就不是量子化的了。卖糖的只能一包一包卖，这是量子化的。而爱因斯坦干脆就假设量子化本身就是事实存在的。

物理学家最近三十年做的工作是从量子力学的角度重新去理解信息的概念，这与经典的香农现代信息学有很大的差异。目前看来一个物体所蕴含的信息可以分为两种不同的类别，一类是我们所习惯的经典信息，经典信息是可记录、可传播、可描述、可复制的；另一类是量子信息，描述物质的量子关联。量子信息被测量时会发生改变，它不能被复制。量子信息的一个令人困惑的特征在于它可以发生纠缠，一种具体的量子关联。关联是一种具体的存在，它是像粒子一样的确实的物理内容。关联体现了量子力学的精髓。利用爱因斯坦关于纠缠

的悖论，事实上我们发现物质的一部分信息可以超光速传递，即相位信息，或纠缠的量子信息，而经典信息是无法超光速传递的。量子信息体现在物质组成单元的关联上，它不可以被经典的方法直接测量，一旦测量就会发生变化，但它仍旧可以用量子通道，即关联的量子粒子来传递。经典部分的信息可以像打印机那样被扫描和复制，信息的量子部分涉及组成基本单元的关联，被测量时就会被改变，也就无法被原样复制。但通过关联的渠道，即量子通道，它确实可以从宇宙的一点，立刻、即时地把一部分信息内容或者物质状态传递到宇宙另一点，量子信息的传递与这两点的实际距离没有关系，它可以远远超过光速。

量子纠缠的悖论

爱因斯坦并不是反对量子力学的结论，事实上，他在世的时候已经有大量的实验证明了量子力学的正确性。他只是非常非常不喜欢量子力学的数学描述，这套理论太不理性，应该有一个更好的理论来解释量子力学的所有推论，应该把量子力学所揭示的非理性概念回归到他所习惯并信仰的柏拉图式的理性框架里来。1935年，爱因斯坦和波多尔斯基（Boris Podolsky）、内森·罗森（Nathan Rosen）三人写了一篇至今都非常有名的文章，质疑了量子力学理论的完备性。他们认为一个粒子在局域范围内携带了其全部性质，对其进行测量的所有结果构成粒子本身的物理真实，不存在这个局部范围之外的物理真实，这个局部范围的大小是受光速限制的。而根据量子理论的观点得出的结论是：相隔很远的两个粒子之间可以瞬间发生相互作用，对其中一个粒子的测量结果受远处的另一个粒子的操控，它们彼此之间的影响可以超过光速。这在经典物理的角度看来是不可能的，因此量子理论是不完备的，为了得到一个关于世界的合理的符合相对论要求的

描述——物质与物质之间的相互作用不会超过光速，就必须补全量子理论的形式，史称EPR悖论。这是爱因斯坦针对量子力学的哥本哈根解释的新反驳，是关于两个粒子的思想实验。他发现量子理论形式上允许两个粒子处于一种纠缠态，并可以预测两个粒子测量结果之间的强关联。即使两个粒子相隔足够远的距离，而且任何直接联系都不存在时，这种关联依然能保持。虽然从时间上判断爱因斯坦的EPR悖论未必受了哥德尔的影响，因为EPR悖论是在1935年被提出的，这时候年轻的哥德尔的工作未必得到了爱因斯坦的足够重视，但爱因斯坦晚年关于量子力学不完备的讨论，至少从“量子力学不完备”这个说法上来说，很难说不受哥德尔的影响。谁让这俩人成了忘年交呢！

量子力学允许一对纠缠粒子存在，如果对其中一个粒子进行观测，观测不只是影响了它，也同时影响了它所纠缠的伙伴，而且这种影响与两个粒子间的距离无关。两个粒子的这种非经典的远距离连接，被爱因斯坦称为“鬼魅般的超距作用”。爱因斯坦无法相信纠缠会如此运作，他认为应该有更深刻的方式可以解释为什么它们彼此连接，而不必涉及神秘的超距作用。玻尔所拥护的量子力学表明，相互纠缠的粒子即使相距很远，也可以互相关联；而爱因斯坦则不相信有鬼魅般的关联，他认为在观察以前，一切就已经决定了，也就是说，粒子在被观测前就已经决定了自己的状态。爱因斯坦会说：“那你怎么知道呢，你测量它，就会发现那是绝对的状态。”玻尔则会说：“但是那状态是由于你的观测所造成的。”双方辩论的当时，没人晓得怎么去解决这个问题。这个问题很快被认为是哲学问题，而不仅仅是科学问题。爱因斯坦逝世前仍相信量子力学是个不完备的理论。问题回归到了在时间上的因果关系，测量前物体的状态就已经决定了还是测量本身决定了物体的状态，这是个可以通过时间相关的测量来考察的问题。

爱因斯坦把他的论点进一步具体化，他解释说一对纠缠态的粒子用一双手套就可以说明。想象把一双手套分开放在两个箱子中，把一个箱子送到南极，另一个箱子则放到北极。在两个箱子分开前箱子里

放着哪只手套其实已经确定了，只是实验者缺乏这方面的信息，不知道而已。观察者打开送到北极的箱子时，如果看见左手的手套，在这瞬间，就算没人看过南极的箱子，你也能够知道那里装的是右手的手套。这一点也不神秘，你打开箱子，显然不会影响到另一个箱子里的手套。放在北极的这个箱子里装着左手的手套，而放在南极的那个箱子里则装着右手的手套，这是在当初分装时就已经决定了的。爱因斯坦相信，所谓的纠缠态只不过如此而已，手套的一切状态在它们彼此分离的时候就已经决定了。同样的问题牵涉到薛定谔的那只猫，人们还是倾向于相信打开箱子之前猫就已经是确定了死活的。然而，量子力学对此的反驳来自多方面，一是相同制备的关联系统展示了如果手套不是叠加态而是一个确定的状态，就不会出现相干的态，而事实上确实存在。另外，关于到底是什么时候决定了态的选择，手套的左右是箱子被分开时决定的还是开箱的那一瞬间决定的，贝尔设计了一个聪明的实验来检测这个时序问题，其后紧接着的实验都证明爱因斯坦错了。

爱因斯坦跟哥德尔如此的熟悉，他不可能不跟哥德尔谈起来他所认为的量子力学不完备的问题。所以他说量子力学是个不完备的体系，哥德尔不完备是必然的。我们怎么能那么幸运，又或者说是倒霉地找到一个完备的终极理论呢？而这个终极理论竟然是可以被理解的？首先它必不完备。

---

[1] 图片来源：T. Pyle/Caltech/MIT/LIGO Lab

[2] 图片来源：R. Hurt/Caltech-JPL

[3] 图片来源：European Space Agency, *nASA and Robert A.E.Fosbury (European Space* Agency/Space Telescope-European Coordinating Facility, Germany)

第九章

量子力学的完备性

量子力学和计算复杂性，都预示着一些由来已久的问题超越了这一代人工智能解决问题的能力，即使计算能力有了极大的提高，这些问题也并没有实质性的解决，而越来越多的证据表明，这些问题决定了我们人类处理问题和这一代人工智能算法本质上的不同。了解量子计算的一个效果是让我们更好地去理解量子力学；了解人工智能的至少一个好处是让我们更好地了解了自己。

量子力学和哥德尔：经典逻辑之外？

爱因斯坦希望找到一个数学解释，可以维系从牛顿以来所形成的局域实在的物理世界。这个局域实在的观点建立起来并不容易，但建立之后深刻地影响了人类科学发展。局域性首先解释了“因果”的起因。如果在同一个闵科夫斯基空间内，但不在光锥之内的两件事是没有因果关系的。

宇宙间的相互影响是受光速限制的，物质与物质的相互作用是基于某种物理的过程，哪怕相互之间传递的是某种虚拟的光子，事实上，电磁的相互作用是用虚光子的传递来解释的。任何影响以不超过光速的速度传递给其他物质，而观测到的现象是出于物质本身的特性，与观测者的是否观测没有关系。我们称这种经典物理的观点为局域实在。

我们知道，光速是宇宙间最快的速度，信息传输应该不会比光速快。而量子理论预言，处于纠缠态中的两个粒子的这种纠缠性质的关联可以比光速传输更快，也就是说，二者可以瞬间传递信息。

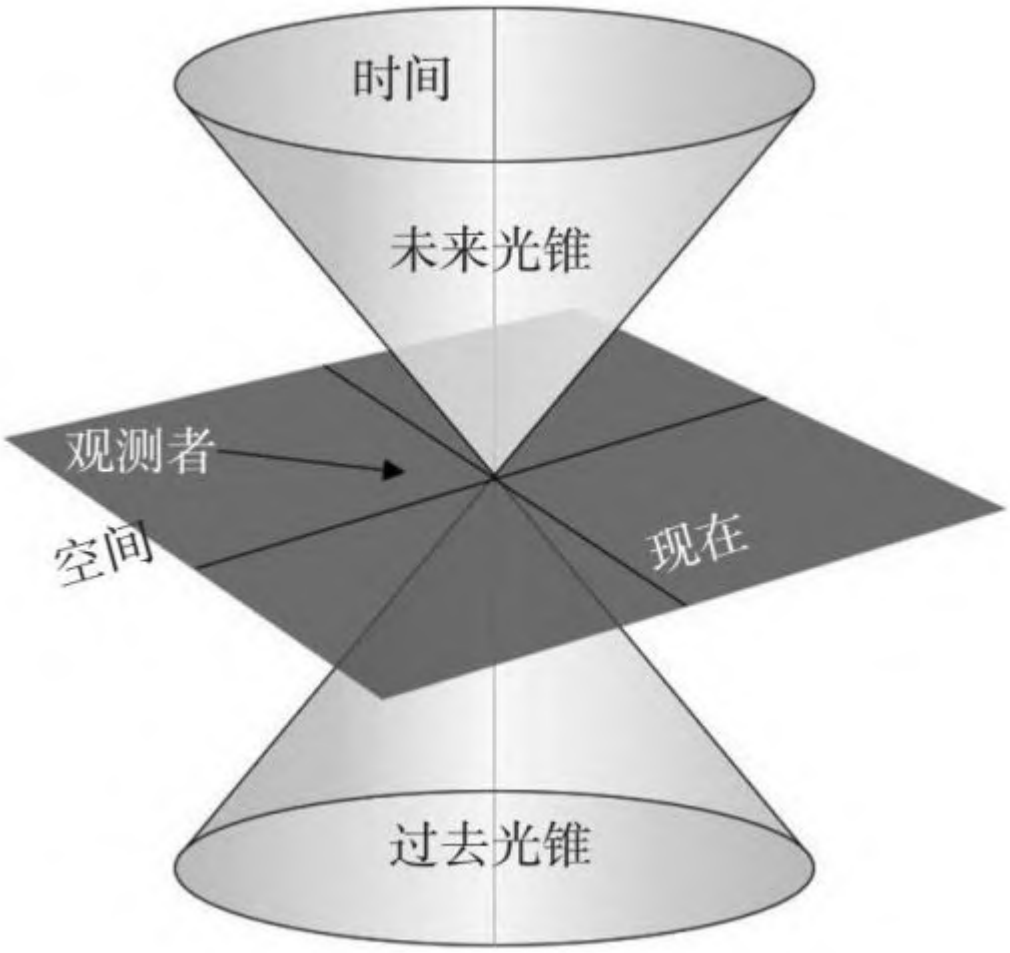


图9–1 闵科夫斯基空间

爱因斯坦认为物理的局域实在本质不允许这样的事情发生，因此他认为唯一合理的描述是，处于纠缠中的粒子对中的任何一个在分开时就已经携带了某种关联特性，因此也决定了测量的结果。

由于每个纠缠粒子的特性在量子理论中不能被分开描述，因此爱因斯坦认为量子理论的形式是不完备的。玻尔却强烈反对这个结论，并坚信不可能在不破坏自洽性的同时给出量子理论的完备形式。除了薛定谔外，大多数物理学家都没有特别在意玻尔与爱因斯坦之间的争论，因为冲突只是针对量子理论形式的理解，而非挑战量子理论所预

测的测量结果。直到贝尔提出贝尔不等式的检测，人们才意识到，量子力学的一些预测与爱因斯坦的局域实在观点之间存在严重冲突。

1964年，物理学家贝尔找到了一个不等式可以用实验测量来对比局域实在与量子力学的不同预测：人们可以通过一个实验，测量贝尔不等式要求的某些参数，如果测量结果符合贝尔不等式，那么就证明局域实在是正确的，量子力学是不完备的；如果测量结果不符合贝尔不等式，则证明量子力学是完备的，局域实在的观点是不正确的。

在我们现在称之为贝尔不等式的公式中，贝尔指出，对于任何根据局域实在得出的预测，都存在对所预测的相关性的限制，即相关性不会突破贝尔不等式所预言的范围。同时他也认为，根据量子力学的结论，通过一些光学实验的设计，这些限制可以被突破。这意味着在对量子理论形式的解释上局域实在与量子力学相冲突，双方在实验结果的预测方面也有冲突。

贝尔的发现将爱因斯坦与玻尔的争论从认知论转移到了实验物理可检验的真实地带。在接下来的几十年内，实验物理学家设计了各种方法来检验贝尔不等式。1972年，加州大学伯克利分校和哈佛大学合作完成了第一个实验，接着1976年，得克萨斯州农工大学也完成了同样的实验。1982年法国国家科学研究中心物理学家艾伦·阿斯佩克托（Alan Aspect）等研究人员所做的实验也验证了贝尔不等式与量子理论预测吻合得很好。所有实验都趋向于证明量子力学是完备的。

从实验结果来说，关于量子力学的完备性问题的讨论，爱因斯坦至今都还是错的。应该说这里提到的完备性跟哥德尔所讲的完备性有本质的差别。爱因斯坦在EPR悖论中所说的不完备根本就是量子理论与已知的其他理论的矛盾，而哥德尔的不完备定理指向的是理论自身的矛盾问题。

海森堡不确定性原理

哥德尔作为一个无辜的数学家、逻辑学家，被爱因斯坦这“老谋深算”的物理学家洗了脑。虽然爱因斯坦觉得物理学界的年轻人在孤立自己，因为似乎所有的人都站到了量子力学那一面，而他却始终不相信量子力学是完备的。显然，这个说法也多少受到了哥德尔影响，但他还是给他的战友成功“洗脑”，以至于，我们没有任何证据表明，哥德尔相信量子力学。

哥德尔认为他的不完备定理与量子力学的“不确定性原则”没有直接的联系。海森堡在哥德尔发现不完备定理之前的几年就发现了以他名字命名的海森堡不确定性原理，这一原理被宣传为另一个限制我们了解世界终极真理的能力。然而，爱因斯坦死后，哥德尔对别人跟他讲的关于量子力学的任何思考都非常敌视。

著名物理学家约翰·惠勒曾经在访问普林斯顿高等研究院时试图去跟哥德尔讨教他的哥德尔不完备定理与海森堡的不确定性原理之间存在的联系，“好吧，有一天我在高等研究院做讲座，顺便拜访了哥德尔的办公室。时值隆冬，哥德尔办公室开着电加热器，他的腿上裹着毯子。我说，‘哥德尔教授，你看你的不完备定理与海森堡的不确定性原理会有联系吗？’哥德尔勃然大怒，把我从他办公室里赶了出去！”

然而我们仔细审视海森堡不确定性原理和哥德尔不完备定理的相似点，就会发现如下事实：

大多数时候我们都可以忽略海森堡不确定性原理的存在，就像我们忽略哥德尔不完备定理的存在一样，这并不影响我们认识足够多的自然界；

它们处处都存在，只是我们的认知习惯选择了对它们无视；

它们限制了我们任意的认知和理解世界的能力；

它们是超越工具性的，对我们人类认知世界的能力限制是不由工具本身的提高而解决的。

海森堡在1927年发现，如果要测量一个运动粒子的位置和动量，人们有一个基本的精度限制。在某一方面如果想要测量的精度越高，其他方面人们能说的就越少。这并不是指测量仪器的质量问题，而是自然界根本就具有不确定性。这个结果现在被称为海森堡的不确定性原理，它意味着在量子力学里我们谈论不了粒子的位置或轨道。

不管怎样，既然一个是量子力学的精髓，一个启发了图灵，二者之间会有什么其他关联呢？1982年费曼（Richard Feynman）在一次物理学会议上提出了量子计算的想法。到1985年，第一个量子计算算法在牛津被戴维·罗伊奇（David Deutsch）提出来，他甚至提出了量子图灵机的概念。量子计算机是经典计算机的替代品吗？还是只是速度更快的升级版的图灵机？

什么是量子计算机？

量子计算机还是要基于量子力学的基本公设体系。那么我们先看看量子力学的理论。

量子力学的公设：

假设*1* 一个量子力学系统的状态由其状态矢量完全确定。

假设*2* 在量子理论中，观察量由作用于相关希尔伯特空间的数学算符表示。

假设*3* 观察量的均值由相应算符的数学期望值给出。

假设4 在一个不受外部影响的封闭系统中，状态矢量将依照含时薛定谔方程随时间演变。

第一个假设在说状态矢量是希尔伯特空间的叠加态，即一个存在的物理量可以写为：

$$\Phi = a\,|\,1\rangle + b\,|\,0\rangle$$

其中，|1〉和|0〉是希尔伯特空间的两个矢量。

如果活猫是一个允许的存在，死猫也是一个允许的存在，那么活猫+死猫也是一个允许的存在。如果状态*A*是一个允许的存在，状态*B*是一个允许的存在，那么一定有一个亦*A*亦*B*非*A*非*B*的状态，也可以是一个允许的存在。这就是有名的量子叠加原理，它是量子力学中“诡异”的根源。状态*A*+状态*B*这个奇怪的存在被叫作状态*A*和状态*B*的叠加态。在量子力学中我们通常把状态*A*、状态*B*标记为|*A*〉、|*B*〉，这样状态*A*和状态*B*的叠加态就被标记为|*A*〉+|*B*〉。

同样这两个矢量可以是矛盾的，比如活猫和死猫，没有道理说我们不能定义这两个态一个代表真值，一个代表其否定。于是这让我们深刻地怀疑，量子力学并不是经典意义上的自洽。事实上，自量子力学诞生以来的这一百多年，无数智慧如爱因斯坦的物理学家，都在努力证明量子力学的不完备，但我们所有迄今为止的实验都说明了这些质疑是错误的。那么我们是不是可以承认量子力学本身是完备的，但不是经典意义上的自洽呢？

我没有好的理由继续下去，不是经典意义上的自洽系统，那它又是什么呢？我开始构想一种超越了经典理性的逻辑框架。但也许这就像是普朗克的工作一样，在本来可以开启的大门前犹豫不决，最后给了一个应付了事、似是而非的临时解释。这里我们不做太多展开，这还是个假想，缺乏实验的依据。

量子力学也许预示着我们不得不找到一个好的数学工具，这个数学工具可以包容量子叠加性所带来的不自洽问题，从矩阵代数出发的可以叠加的数学系统呢？至少在量子力学方面的数学工具还是有用的？作为一个实验物理出身的工作者，我就不凭想象去延伸了。

但我们确实无法想象，如果数学工具本身是有缺陷的，那么它会给建筑于这之上的物理学产生什么影响。一旦有，那影响将是深刻的。

量子计算机

计算机科学和计算机工程改变了现代社会的各个方面。在这些领域，前沿的研究是关于计算的新模型，构建计算机硬件的新材料和新技术，加速算法的新方法，以及在计算机科学和其他科学领域之间建立桥梁，科学家们利用自然现象作为计算程序的对象，并采用新颖的计算模型来模拟自然过程。人们也研究优化处理信息和计算解决方案所需的资源，通过比较不同解决方案的复杂性来估计实施成本的问题。但理论计算机科学却常常因为它习惯的规范形式而不考虑用于执行计算或信息处理任务的那些设备的物理特性和成本，因此常常只顾及算法的优美而忽略了物理的硬件支持能力。因为量子力学主宰着一切物理系统的行为，即使是我们所说的经典世界，也必须归纳于量子力学的一种特殊情况，但怎样做一台实用的量子计算机直到今天我们并不明确。一台实用的量子计算机，我们不仅仅需要做一台能像图灵机一样工作的计算机，而且还要解释为什么这台计算机比图灵机要强大。1982年费曼在一次报告中提出了另外一个考虑，即我们需要的也许不是一个通用的计算机，而是一个通用的量子模拟器，一个因为特殊目的而构建的量子模拟器。复杂的物理系统没有对应的有效的模拟它们的计算机方法，因此费曼建议用量子系统直接模拟量子系统。因

为根据第一性原理，当两个哈密顿量相同，一个量子系统是可以模拟另外一个量子系统的。无论如何，这是量子计算第一次被正式提出。

费曼提出的论点表明量子计算可以比任何经典计算机更有效地计算某些问题，它可能不仅仅是升级版的计算机。他认为可以通过特殊设计来模拟相似的两态系统之间的任何相互作用，但他没有说明如何实现。

用于计算或信息处理的任何物理设备的行为最终必须通过物理定律预测，就不得不遵循量子力学的基本公设，量子计算可以被定义为构建量子计算机和量子信息处理系统的跨学科领域，即使用量子力学性质的计算机和信息处理系统，它的研究目前主要集中在利用量子计算机物理特性的构建和运行算法上。

在将量子计算定位为替代经典计算机而成为发现新一代科学理论的希望和工具时，人们需要发现并开发新颖而强大的计算方法，它可以使我们显著地提高数据处理能力来解决某些问题。寻找好的量子算法是一项艰巨的任务，因为量子力学是一种违反直觉的理论，而直觉在算法设计中起着重要作用，并且对于量子算法来说，它要比任何对应的经典算法做得更好、效率更高才有开发的价值，才能受到足够的认可。但无论如何，构成量子计算与经典计算的不同是有一些关键元素的，包括量子比特、量子随机、量子长程关联和量子行走。

量子比特

比特，是一个有0或1两个取值的计算和存储单元。任何一个物体，如果它存在两种不同的状态，那么我们就可以用这两种不同的状态来实现一个比特。比特是信息的基本单位，是数字通信数字计算机中的主角。我们现在所经历的信息革命，如手机、微信、WiFi（无线连接）、电视等，就靠它了。而它的出现，可以上溯到1837年摩尔斯发明的电报，长按和短按电报机的发报键。长按代表1，短按代表0。

通过密码本把所有英文字母对应为0和1的组合，这样就可以通过电流的通断来传递信息了。

量子比特是量子信息的基本单元，是量子通信和量子计算中的主角。它的量子性质造就了量子通信和量子计算的神奇。有人甚至认为量子比特是组成世间万物的基本构件。可量子比特到底是个什么东西？它神奇在什么地方？

在经典物理中，最简单的系统就是一个比特。一个比特只有两个态：0或1。而量子叠加原理告诉我们：任何两个态的叠加也是一个可能的态。为了区分经典比特的态，我们用| 0〉和| 1〉来标记量子比特的状态。所以一个量子比特，不仅有| 0〉和| 1〉两个态，还有它们的任意叠加态：

$$| \psi \rangle = \psi 0 \ | 0\rangle + \psi 1 \ | 1\rangle$$

这里ψ0和ψ1被称为叠加系数。我们发现一个量子比特可以有无穷多个不同的状态，这些状态由两个复数ψ0和ψ1来刻画。当然这无穷多个态大多都不是相互排斥的。常用的是两个相互排斥的态，如|0〉和| 1〉、| 0〉+ | 1〉和| 0〉 - | 1〉，等等。两位量子比特可以存储（00，01，10，11）4个状态，而3位的量子比特可以同时存储（000，001，010，011，100，101，110，111）足足8个状态的信息。

只要是一个量子系统就可以当作量子比特的备选对象。困难在于如何有效地操控它们，同时避免环境的影响。一个电子的两个自旋态就可以实现一个量子比特：自旋向上对应于| 0〉态，自旋向下对应于| 1〉态。

我们知道光有偏振现象，光的偏振也可以有不同的方向，不同方向代表了光子自旋的不同状态。这无穷多个偏振方向，表示光子自旋可以有无穷多个不同的状态。像电子自旋一样，这无穷多个态大多都不是相互排斥的。我们一样有两个相互排斥的态，如竖偏振和横偏

振。我们可以用竖偏振代表| 0〉，用横偏振代表| 1〉。这样一个光子的自旋（偏振态）就是一个量子比特。一个光子还可以有左斜偏振，其对应于一个又横又竖不横不竖的偏振态，记为| 0〉+ | 1〉。一个光子也可以有右斜偏振，其对应于另一个又横又竖不横不竖的偏振态，记为| 0〉 - | 1〉。量子保密通信可以利用这4种状态的光子来实现的。

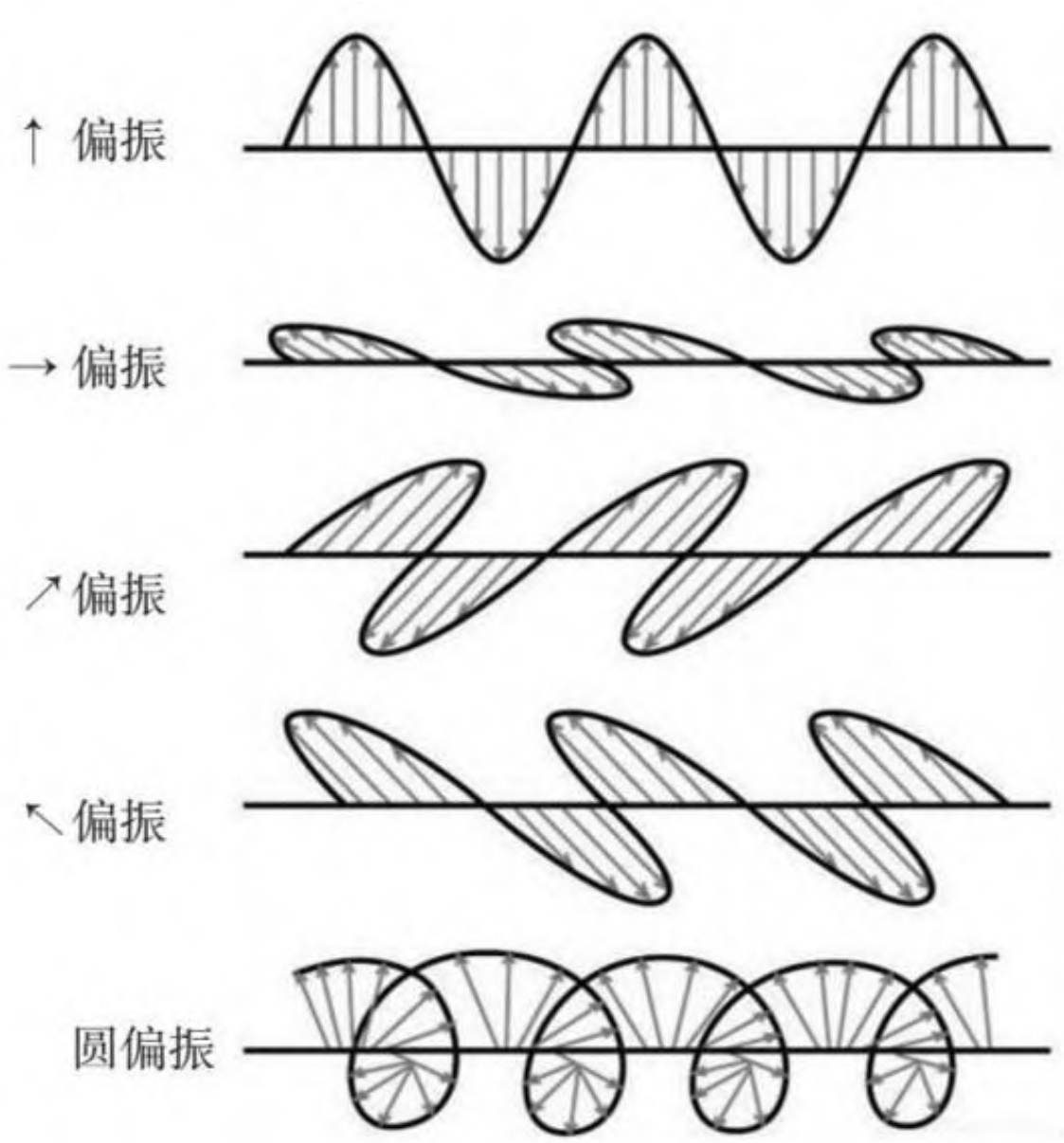


图9-2 光子偏振的几种方式，即几个不同的量子叠加态

量子随机数

对量子随机数最好的描绘，莫过于爱因斯坦所说的话：“上帝不可能掷骰子”。但玻尔回敬了一句，“别去指挥上帝该怎么做”。

在爱因斯坦的世界观里，基于牛顿力学所构建的局域实在，是不允许某一个事情的发生是没有原因的。这与经典逻辑所要求的第四条公设充分理由定律，即事情的发生必然有原因的要求一致。一个骰子落在了哪个面，不由任何外在的、内在的原因来决定？完全不可预测？这跟经典的经验并不吻合。当我们知道所有初始状态的信息并且有足够强大的计算能力，为什么我们不能够构建一个完美的方程来描述骰子的所有运动，从而对它的未来结果进行预测呢？这就引出了拉普拉斯（Pierre-Simon Laplace）悖论：当我们知道了宇宙所有粒子的运动状态，和它们的相互作用性质，我们总能够写出来足够多的方程，这些方程就成了描绘这些粒子运动的原因，而它们所描述的对象的演化，就是自然的结果。

如果拉普拉斯悖论为真，我们一样可以写出关于我们自己的计算程序，这样我们的历史和未来都是已经设计好的脚本。只要有这个脚本，或者将来有一台足够容量的计算机，就能计算整个人类的历史和未来。它已经知道了一切，而我们，自以为有自由意识的人，就像安排好的演员，有固定的脚本，时间到了就上去演戏。该做什么，有什么样的结果，都是预先已经设定好的。历史上有这样的辩论，因为所有事情都是预先设计好的，所以杀人犯也不用判刑，因为这不是他决定的，而是在很久很久以前，甚至是在他出生之前，这一切都被决定好了，他只是执行者或者也是这个决定的受害者。而法官可以回应说，我判处他这样受罚也是预先被决定好了的。

近代的物理学发展对拉普拉斯悖论做出了这样的回答。混沌力学发现即使动力学方程本身是确定的，但它的演化结果也有可能完全不

确定。任何微小的差别，都可以演化出完全不同的结果。而这些初始的微小的不同，可以来源于量子力学的随机。我们完全没有办法预测量子的随机什么时候发生，发生在什么地方，怎样发生，不仅是我们无法预测，上帝都不能。

量子长程关联

量子长程关联揭示了一个完全不同于局域关联所限制的世界。当然从实现的技术和方案上我们还没有足够的进展来造出一台实用的量子计算机，但我们可以想象利用量子的长程关联，已经有一些与经典计算截然不同的效果。

经典计算单元需要局域的相互作用作为基础。对于一个经典计算机，它的每一个计算单元都要靠跟相邻的计算单元通过相互作用来实现计算。在物理器件中实现需要，比如说具体到导线的连接，这些物理连接制约了算法。当第一个计算单元完成计算的时候，它有N个有限的连接，接下去每一个连接也还有n个，这样继续下去，有$N\times N\times N$个连接，事实上，由于局域相互作要求的物理接触，使得由于物理空间的限制，N通常只能做到几个。但对非局域关联来说，所有的计算单元同时参与计算，这样的网络可以迅速遍历所有的计算通道，使某一类型的算法执行起来非常快。我们要谈到的量子随机行走，它可以在一个网络中迅速找到答案所需的节点，因为它几乎是同时遍历所有的节点，而不是一步一步依靠经典的物理联结（局域关联）扩散而找到答案。当我们讨论人脑细胞的相互作用，如果说人脑有130亿个脑细胞，假设每个细胞会跟6个细胞发生相互作用。它们是通过物理连接而局域关联的话，相互作用关联数量就是$N\times$130亿，N是个不太大的数。而当它们是非局域关联的话，其关联数量就达到惊人的130亿的6次方，这远远超过了地球上所有计算单元的总和。

量子的随机行走

经典随机行走是一种随机过程算法，它描述了算法的演化过程有随机的机制，实践中它已经被证明是随机算法发展的一种非常有用的技术。除了在算法中发挥的关键作用之外，经典随机行走在物理学、生物学、金融理论、计算机视觉和地震建模等许多科学领域中也发挥了重要作用。

量子行走是经典随机行走的量子力学的对应，是构建量子算法的一种工具。量子行走被证明可以构成一类通用的量子算法。由于经典随机游走已被成功地用于开发经典算法，量子计算的主要议题之一是量子算法的创建比其经典算法更快，因此人们对理解量子行走的性质有着极大的兴趣。尽管一些作者选择了“量子随机行走”这个名称来指代量子现象，事实上，在费曼关于量子力学计算机的开创性工作中，我们发现它可以被解释为连续的量子行走。但更多人认为，第一篇以量子行走为主题的论文是由亚尔基·阿哈罗诺夫（Yakir Aharonov）等人于1993年发表的。自此，经典随机行走和量子行走之间的联系以及量子行走在计算机科学中的应用成为新的研究领域。目前我们常讨论的模型是离散量子行走模型，它由两个量子力学系统组成，一个负责走路，另一个负责丢硬币。离散量子随机行走的操作很简单，一个硬币，扔起来，如果是正面，朝左走，如果是背面，朝右走。当这个过程量子化的时候，我们会发现它的行走速度要比经典随机行走的速度快很多。尤其是对一个网络而言，用量子随机行走来搜索这个网络的节点时，其速度会远远快于经典随机行走的速度。基于连续量子行走的算法，给定一个图，由两个高度为n的二叉树组成，左侧树的2n个叶子用根据图9－3所示的方式，右树的2n个叶子，并且有两个标记的节点入口和出口，找到一个从入口到出口的算法。从中可以看出，有可能构建一个量子行走，它从入口到出口，比其相应的经典随机游动指数快得多。换句话说，这个模型提出的连续量子行走找到目标的时间是多项式的（与N的几次方成正比），而相应的经典随机行走找到目标的时间是指数级增长的（与2^N成正比）。

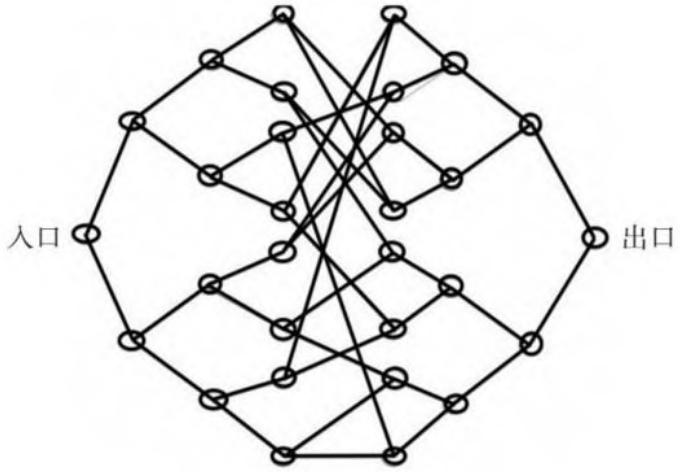


图9-3 连续量子行走算法

量子算法

1985年，戴维·多伊奇迈出了关于量子计算重要的一步，他设计了量子计算的第一个例子。多伊奇定义量子算法是一类可以在量子计算机上执行的算法，并证实了量子算法有可能比传统的算法更有效。多伊奇算法利用了量子叠加和量子纠缠原理，被广泛认为是量子计算的第一个蓝图。因为它具有足够的特性和简单性，可以考虑用真正的物理装置实现。多伊奇算法作用于一个二值状态系统。多伊奇证明，如果可以通过一组特定的简单操作这个二值态系统，就可以具有普适性，进而可以模拟任何物理系统的演化。他还讨论了如何使用相同的思路产生类似图灵机的计算装置，从而成为一个实际意义上的通用量子模拟器，尽管这点需要未来更多的发展来验证。

多伊奇的简单操作现在被称为量子门，因为它们在经典计算机中起着类似于二进制逻辑门的作用。多伊奇提出了量子图灵机的概念，成为所有今天所谈及的量子计算的原型。对应于经典计算机的比特电

子操作，量子算法也有针对量子比特的量子门操作。设计一个对二值操作的量子门，Hadamard门。我读书的时候戏称哈德门，它可以将量子比特映射到完美的叠加态，如下：

$$H=\frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}$$

在多伊奇的设计中，他给出了一个输出值为0，1的函数f（x_1，x_2.....x_n）的黑盒子。黑盒取n位x_1，x_2……x_n，x_i取值为0或1，并输出值f（x_1，x_2……x_n），黑盒子中的函数要么给出常数（所有输入都为0，所有输入都是1）或均衡（对于一半输入值返回1，对于另一半返回0），任务是确定f是常数还是均衡。对于经典的确定性算法，在最坏的情况下将需要2″次的f评估，即将0和1分别代入来比较结果是否都为0或者1，还是一半是0一半是1，利用经典随机算法，足够数量的评估可以产生高概率的正确答案，但如果我们总是想要正确答案，仍然需要2″次评估。而多伊奇量子算法利用哈德门形成计算单元的量子叠加，并利用计算结果的量子相干，只需1次f评估即可产生一个总是正确的答案。

多伊奇算法是最早的量子算法，后来有用于大数分解的彼得·舒尔（Peter Shor）算法和搜索的格罗弗（Lov Grover）算法，以及量子随机走动算法，这些算法都被证明比经典的算法要快，这种快不仅是计算速度的加快，而且是计算复杂程度意义上的降低。这里尤其要提到的是舒尔的大数分解算法，它使得需要计算时间为O（N^p）的问算法要1.6×10^8步，而经典算法要6.7×10^{51}步。如果计算机执行每步时间为10^{-9}秒，舒尔的算法只要不到1秒，而经典计算机要大约两千亿亿亿亿年。而我们知道的今天通用的数据安全加密系统都是基于自然

数的大数分解，一旦可以这样快地分解一个很大的自然数，常用的密码系统就变得很容易被破解了。

量子门是量子电路的基本模块，就像逻辑门用于经典数字电路一样。和经典电路逻辑门一样，设计类似于经典逻辑操作中的与非门，通用量子门（可控非门）可以实现量子计算的所有算法。在经典计算机系统中，通过逻辑门以某种方式改变和修改存储在经典存储器中的二进制值，产生新的值。对于在量子存储器中具有叠加的量子系统也是如此，应用于另一个叠加的量子门，得出有效输出。量子门和经典逻辑门之间的一个主要区别是其运算的可逆性。量子门操作是可逆的，即在X上应用量子门 A，得到Y，在Y上应用量子门的逆操作A^{-1}，$A^{-1}Y$就会再次得到X，没有信息丢失。经典逻辑门是不可逆的，数据处理会使信息丢失，这是经典计算的巨大的问题：根据香农定律，经典计算设备因为处理信息需要消耗大量的能量，所以对系统有热量的损耗。对于量子计算来说，这个问题几乎不存在。这是个有意思的事情，我们通过量子计算知道了某些新的东西，但没有以付出相应的热量为代价。这与香农的信息熵的处理有很大的不同，也违背了经典热力学第二定律的。这是否也意味着大脑作为信息处理工具也有着量子化的基础，如果它真是以量子计算为基础，会省很多能量，这跟我们的经验是一致的。到今天的人工智能算法都以其对计算资源的占用和耗费著称，而解决大多数类似问题的人脑，它的低功耗简直让人惊奇。

在经典系统中，任何逻辑门都可以使用布尔代数表示。类似的量子门表示的变换可以描述为酉变换矩阵。在这样的矩阵中，如果门作用于n输入量子态，则矩阵的大小为$2^n \times 2^n$，其中输入量子比特的数量等于输出量子比特数量。当一个量子比特能同时存储两个信息的时候，那么两个量子比特就能同时存储4个经典信息，3个量子比特就是8个经典信息……由此，当我们有50个可控的量子比特的时候，我们就有了2^{50}个经典信息，这已经超过了宇宙中的粒子数量。而运算操作

中，如果能够充分利用量子的叠加性质、纠缠性质，那么我们就可以设计合适的算法，同时对这么多计算单元进行操作，并能顺利地读出操作结果，这些算法的运算速度，就会远远超过经典算法的运算速度。而最近的一些研究表明，在经典算法所认为的不可解问题上，量子计算是可以为其中的一些问题提供解法的。这就是说，量子计算的计算能力比经典计算要强，速度也要更快，计算效率更加节能。量子计算可能会涵盖一个更大范围的可计算问题，这对数学本身来说是具有重大意义的。

量子计算的复杂性

理论计算机科学的一项基本任务是将问题按照复杂程度来分类。计算复杂性由解决这个问题所需的资源来决定，通常指的是计算时间和计算这个问题所需要的物理资源。

例如，我们有测试某个数字是否为质数的算法，但尚未找到有效的算法求解一个大数的质因数。前者是个*P*问题，而后者是个*NP*问题。我们定义*P*问题是经典计算机可以快速解决的所有问题。“这个数位是质数吗？”属于*P*问题。*NP*问题是经典计算机不一定能快速解决的所有问题，但如果有答案，它们可以快速验证答案。“一个大数的质因数是什么？”属于*NP*问题。计算机科学家认为*P*问题和*NP*问题是不同类别的问题，但两者之间的差别是这个领域至今没有定论的最著名和最重要的问题。

1993年，计算机科学家伊桑·伯恩斯坦和优曼许·沃兹内尼定义了一个新的复杂性类*BQP*问题，即有限误差的量子多项式时间内能解决的问题，量子计算机在多项式时间内可以解决的一类决策问题（确定“是”或“否”答案的问题），计算的随机错误概率小于1/3。*BQP*问

题是量子计算机可以有效解决的所有决策问题。在同一时间，他们还证明了量子计算机可以解决经典计算机可以解决的所有问题。也就是说，*BQP*问题包含*P*问题中的所有问题。但科学家希望找到一类问题，是在经典计算机*NP*的问题之外，*BQP*可解的问题。这样，就可以证明量子计算机的范畴要大于经典计算机。理论计算机科学家已经知道量子计算机可以解决经典计算机所能解决的任何问题。

2018年两位量子计算学家然·拉茨（Ran Raz）和阿维沙伊·塔尔（Avishay Tal）的工作表明有一类计算问题，量子计算机可以有效地解决，但经典计算机却不能。拉茨和塔尔的论文表明，量子计算机和经典计算机确实是不同的类别——即使在一个经典计算机成功解决了*NP*问题的世界里，量子计算机仍然会超越它们。

但我们无法确定*BQP*问题是否包含另一个被称为*PH*的重要问题，*PH*代表多项式层次结构。*PH*问题是*NP*问题的推广，它意味着如果从*NP*中的问题开始，并通过分层诸如“存在”和“普遍存在”之类的限定语句使其更加复杂，它将包含所有问题。今天的经典图灵机不能解决*PH*中的大多数问题，但你可以把*PH*看作是经典计算机在*P*等于*NP*的情况下可以解决的所有问题的一类。换句话说，比较*BQP*和*PH*的范畴是为了确定量子计算机是否比传统计算机具有优势，能够比经典图灵机解决更多的问题。

区分两个类是否一致的最佳方法是找到一个问题，它可以被证明是在一类而不是另一个类中。然而，由于基本障碍和技术障碍的结合，发现这样的问题一直是一个挑战。如果你想要的是*BQP*中的问题，但不是*PH*中的问题，那么你必须确定一些问题，其答案无法被一台经典图灵机有效地验证。斯科特·阿伦森（Scott Aaronson）在2009年提出这样一个问题，并证明它属于*BQP*问题：有两个随机数生成器，每个生成一个数字序列。计算机面临的问题是：这两个序列是完全独立的，还是以某种隐含的方式相关？比如事实上其中一个序列是另一个序列的傅立叶变换的形式。

这就留下了更困难的第二步——证明它不在*PH*范围内。

这就是拉茨和塔尔所做的工作。他们找到了一个可以区别*BQP*和*PH*的黑匣子。这是计算机科学中常用的一种方式，当人们想证明的东西超出了某一范围时，找到一个例子在这范围之外就够了。

拉茨和塔尔计算出计算机需要询问“黑盒子”的次数，以便比较两个序列是否是独立的。“黑盒子”就像一个暗示装置，你不需要知道它是怎么工作的，但知道它是可靠的。如果你的问题是找出两个随机数生成器是否有某种隐含的关联性，你可以问“黑盒子”问题，比如“每个随机发生器生成器的第6个数字是什么”，接着人们就可以根据每种类型的计算机需要解决此问题所需的提示数量来比较计算能力，需要更多提示的计算机速度较慢。

从某种意义上说，这个问题更关心谈论信息数据而不是计算。拉茨和塔尔证明量子计算机需要的提示远远少于经典计算机解决类似的问题所需要的。量子计算机只需要一个提示，相比较即使有无限的提示，*PH*中也没有能够解决这个问题的有效算法。拉茨说：“这意味着有一个非常有效的量子算法来解决这个问题。”这就证明了，对于被询问的“黑盒子”而言，这是*BQP*中的问题，但不是*PH*问题。

这样*BQP*和*PH*就此分离。拉茨和塔尔的工作也证明，即使*P*等于*NP*，仍然会有只有量子计算机才能解决的问题。它至少提供了足够坚实的证据，说明量子计算可以解决经典计算机不能解决的问题，它存在于与经典计算机不同的计算领域，即是说，如果基于经典计算机的这一代人工智能不能够完成的任务，量子计算机也许可以。

更高一级的逻辑？

哥德尔不完备定理可能包容了量子力学本身。量子力学是完备的而不是经典意义上的自洽，这也许启发我们量子计算机有可能是突破经典计算机理论而超越了图灵机的限制。它能够解决一些图灵机所不能够解决的问题，而向模拟人的思维更进一步。

对待人工智能，人类往往有种对机器的“人肉歧视”，认为只有人做的才是好的。图灵说明了借助于图灵机，所有可描述算法都可以由图灵机来完成，但所有不可描述的东西意味着我们也不懂。比如感情、联想、直觉，我们都不懂如何用算法来描述人类这些特别的技能。一旦我们能够清晰地认知这些技能，我们就可以用一个图灵机来模拟它，但这做不到，至少目前做不到。所以，这意味着我们懂的东西，机器也会懂；而我们不懂的东西机器也休想懂。从这个意义上讲我们没有道理歧视机器，因为我们懂的不比它们更多，我们也没必要担心机器，它们懂的东西也不会比我们更多。而这个界限在于那个似乎并不太处处可见的创造的源泉，“捣蛋鬼”哥德尔所画出的界限意味着：有些真理，我们能够意识到，但不能够为算法所描述。这样的规则成就了人类思维的特殊性，可以让我们跳出有限的假设而获得直觉带来的新知。

第十章

物理和数学：自然与认知的边界

描述自然的法则似乎都具有数学形式。至少在物理学领域是这样的，你可以将定律写成数学形式，这样它就会具有强大的预测能力，物理学的这个分支领域或具体问题才成为一个可描述问题。至于大自然为什么是数学的，同样也是一个未解之谜。但一旦可以用数学描述了，就可以交给图灵机让人工智能的设备来帮助我们做出新的预测了。我们说，这件事情已经被我们的认知领域包容了。

物理学是最基本的、包罗万象的学科，它对整个科学的发展有着深远的影响。到了今天，物理学是与过去“自然哲学”相当的现代名称。现代科学大多数是从自然哲学中产生的。许多领域的学生发现自己是在学习物理学，就是因为它在所有现象中起着基础作用。数学与物理有着令人瞩目的关系，虽然有时候我们说数学不是一门自然科学，或者科学，但它的正确性是不需要实验检验的。我们说一件事情不是科学，并不意味着我们对它有不敬和丝毫的轻视，它只是不适用于我们所定义的科学标准而已。

——费曼

1921年，一本《微积分基础教程》勾起了哥德尔对数学的兴趣。同一时间前后，他读了一本歌德的传记，这又间接地引导他对牛顿的思想和一般物理学的兴趣。1924年，哥德尔进入维也纳大学学习物理。1926年，因为对精确性的追求，哥德尔决定离开物理而研究数学，最终在1928年聚焦在数理逻辑上。从他信里的两段文字，可以看到他这一时期的思想，一是他早早成为柏拉图理性主义信仰者，二是他认为自己与当时维也纳学派的知识氛围格格不入。

大约从1925年起我就是一个概念和数学实在论者。

我不认为我的工作是20世纪早期维也纳学术气氛的一个侧面。

20世纪物理学和数学都经历了新的革命和反思，从而到达了新的阶段。这个新阶段的意义和影响直到今天也还在发酵，事情才刚刚开始。我看到的是，若拿哥德尔的标准来看我们还没有熟悉的，哪怕是

经典逻辑学，每句话都会有问题，这让我写起这本书来也举步维艰。从19世纪末开始，人类的科学和认知经历了两个时代的跨越，首先是经典理性逻辑的梳理和完善，接着是这种逻辑的突破。这两个阶段接踵而来，让我们这个文明古国多少有些措手不及。



图10–1 数学、物理及其他学科以实验之界分割开来

大多数人同意物理世界是客观实在的。我们从经验中得知，人类的思维对物理学有很多认知，这一点不论从我们日常的知识还是从我

们今天掌握的物理学知识来说，都是显而易见的。我们还知道，人类的思维对数学世界也有许多认知，因为我们有丰富可靠的数学。大多数人同意思维经验在我们的知识中起到重要的作用，但这绝不是说仅仅用感官经验就能解释这些概念性知识，而概念性知识不但在物理学中起到重要的作用，它在数学中的作用，甚至更加突出。那么我们自然会问道，思维与实践的结合既然能够创造足够多的概念性知识，尤其是能够了解和应用数学工具，那么这种能力在物理世界中的应用范围能扩展到多大？其本性如何？

数学的不能

数学很早就不自觉地被哥德尔不完备定理所影响。我们回溯历史，在数学建立的初期，就有些某个理论框架里不能理解的事情，从而不得不引入新的概念进行补充。

大约在公元前550年，古希腊的毕达哥拉斯学派认为“万物皆数”（指整数），即数学的知识是可靠的、准确的，而且可以应用于现实世界。数学知识可以由纯粹的思维而获得，不需要观察、直觉和日常经验。而自然界的一切数，不是整数，就是整数的比值，即有理数，rational。

当毕达哥拉斯发现了勾股定律，即一个直角三角形的斜边长度的平方，等于两个直角边的长度的平方和。麻烦就有了，一个等腰直角三角形，当它的两条直角边都是1的时候，它的斜边不是任何两个整数的比值。

毕达哥拉斯学派的希帕索斯（Hippasos）花费了大量的时间和精力去研究这个数，即$\sqrt{2}$，经过大量的研究之后，他得出结论：

$\sqrt{2}$既不是整数也不是分数，是当时人们还没有认识的新数，它是一个新发现。既然基于整数和分数的数合称为有理数，后来人们把这种新数就取名为无理数（“无理数”最初的意思只是“不成比例的数”，而不是没有理由的数，像今天它的字面所暗示的一样）。这个“新数”的发现本可以直接促进当时数学和人类文明的发展，只不过毕达哥拉斯学派的人却视之为洪水猛兽。为了维护学派的威信并永远保守这个秘密，希帕索斯被毕达哥拉斯学派的师兄弟偷偷扔进海里淹死了。

希帕索斯的发现最终还是被世人所知。其他人也发现这种新数的存在，如面积为3、5、6等的正方形，它的边长就无法用整数和分数来表示。随着时间的推移，无理数的存在逐渐成为人们所共知的事实。

这让我们第一次看到已有的数学假设体系中不能给出完整描述而不得不引入新的特殊规则的案例。同样，直到19世纪，欧几里得几何一直是数学公理系统的典范。几乎一开始人们就相信欧几里得几何体现了世界的真实，它是关于宇宙的绝对真理的一部分。它的特殊地位和它对数学其他领域的影响两千多年来基本没有变化。欧几里得几何学的优美和它对绝对真理的广泛信仰为数学信仰提供了重要基石。两千多年来神学家和哲学家普遍认为人类的理性可以掌握一些终极的事物的本质，如果受到质疑，那会超出了我们思维的力量。欧几里得几何作为一个具体的例子和方式预示着人对事物最终本质的洞察是可能的。这里面最神奇而经典的案例应该是牛顿的巨著《自然哲学的数学原理》。在这本书中，牛顿抛弃了他自己创造的微积分工具，只用欧几里得几何证明了他在经典力学中的所有结论。

但是，高斯和黎曼等人发现，事实上可以有其他形式的几何存在。只要不包括欧几里得几何的第五条平行公设，就可以存在其他数学逻辑上自洽的几何。这意味着欧几里得几何学不是唯一的真理——它只是模型。结果，新的相对主义形式出现并挑战了欧几里得几何作为经典系统唯一标准的地位——实际上它只是诸多可以描述自然世界

的模型中的一个。描述地球表面的几何形状就可以不是欧几里得式几何的。“非欧几里得”变成了一个新的和相对真理的代名词，以至于人们不得不审视这样的相对真理是不是也会延伸到我们默认为正确的其他工具上。这甚至包括逻辑学本身，新的数学化的形式逻辑通过改变经典逻辑的公理系统表达方式，也可以创造出来平行于亚里士多德已定义的逻辑学而成为新的逻辑工具。

19世纪初，挪威数学家阿贝尔（Henrik Abel）证明不同于2次、3次或4次方程的情况，一般的5次方程不能通过任何方式精确求解。几年之后，在1837年，法国数学家皮埃尔·旺策尔（Pierre Wantzel）严格证明了仅仅通过使用直尺和圆规作图，不能把60°的角分成三等份。这些例子继续在各个角度暗示特定数学公理系统的局限性。

传统的经典数学观作为我们描述世界的一种方式，可以被其他的方式取代。我们逐渐认识到数学可以存在无限制的模式系统，因此我们可以定义无限数量的公理系统。其中一些模式可以在自然界中使用，大多数不是不可以，更多的时候只是不方便。作为数学系统的典型代表，欧几里得几何曾经很长时间内被认为是世界绝对真理的一部分，与现实有着独特的联系。但非欧几里得几何形状的发展和非标准逻辑意味着现在的数学存在只不过是逻辑自洽的，即不能证明0=1之类的矛盾命题，它不再是唯一的描述物质世界的工具。这让19世纪末的数学家有一种期待，既然可以延伸出这么多门类的数学系统，是否有一种公有的数学语言系统来描述这些数学系统，并且判定它们的有效性和有用性。这一门正确的真理性的描述数学的数学，我们称之为元数学。人们希望元数学可以脱离具体的公设从而成为一套完备的逻辑系统，一旦拥有了这套通用工具，我们之后的工作只需要注释来说明数学的某一分支中具体的公设和名词对应于元数学系统中的哪个符号，而元数学已有的论证系统会保证其证明的关系自然成立。

当时最伟大的数学家希尔伯特，通过对皮亚诺从1882年到1899年的工作分析，在1899年开始了一项系统的计划，即将数学的立足点置

于正式的元数学公理化之上。这是希尔伯特说明数学的重点，“一个人必须能够准确陈述点、直线和平面的概念，就像我们可以明确知道桌子、椅子和酒杯这些概念”。他相信有可能确定公理数学的每个部分（以及整个数学）的基础都是坚实的，可以证明这些公理是自洽的，然后表明由此产生的陈述系统和由这些公理形成的推论是完备的和可判定的。

更确切地说，一个数学系统是自洽的：我们不能证明陈述S和它的否定陈述$\tilde{S}$都是真的，即数学定理之间必须不能有矛盾。

一个数学系统是完备的：对于我们可以用数学系统所规定的名词和逻辑所形成的每个陈述语句S或者它的否定陈述$\tilde{S}$，是可以证明其正确与否的。

希尔伯特对数学的形式主义观点是一个严密的推论网络，它要求数学体系内的定理之间有无可挑剔的逻辑联系，数学定理被定义为所有没有瑕疵的逻辑陈述的集合。希尔伯特着手完成了形式化数学，并相信它最终有可能会扩大到所有其他范围，包括物理学等所有科学，因为这些领域最终不过是某种应用数学可以涵盖的内容。他计划从比如算术这个最简单的、最符合人的直觉习惯的数学系统开始，先把它放在一组严格的公设基础上，按照他的规划，通过添加额外的公设来加强系统，在每一步确认系统的自洽性和完备性仍然存在，直到最终系统变得足够大而包含整个数学。

希尔伯特的计划基于牛顿时代之后科学界建立的充分自信：真理就在那，迟早会被我们认知，一切只是时间问题，所有数学都会在形式系统的架构中解决。正如希尔伯特的豪言壮语：“Wir müssen wissen，Wir werden wissen。”（我们必须要知道，我们必然会知道。）

但是，希尔伯特的数学理想王国很快就被年轻的哥德尔颠覆了。作为自己博士论文的一部分，哥德尔完成了对希尔伯特计划的早期步骤，证明了简单逻辑的自洽性（图灵后来表明它不是可判定的，即它不完备）。但他接下来的工作使他成为现代最伟大的逻辑学家。远非延伸希尔伯特的计划而是实现其关键目标——证明算术的完备性，哥德尔证明了只要数学系统稍稍复杂一点，足以包含算术的任何系统都必须是不完备和不可判定的。1930年9月7日，哥德尔在Königsberg（哥尼斯堡，希尔伯特的家乡）举行的会议上简要地宣布了他的结论，虽然似乎没有引起足够的轰动，但出席这次会议的冯·诺依曼意识到了他结论的重要性：哥德尔一举结束了希尔伯特对于终极数学的伟大蓝图。

乐观主义者和悲观主义者

哥德尔不完备定理展示了即使数学系统的理性也受到限制，这个结论很快影响了哲学家和科学家对世界和我们自身的看法。一些人声称它表明了所有人类关于世界万物的研究也必将受到限制。科学是基于数学的，数学无法发现所有真理，因此科学也无法发现所有真理。人们甚至悲观地认为哥德尔的工作使得数学的发展成为仅仅为了锻炼人类智力的游戏，如同数独或者棋类游戏，人们在其中自得其乐，并不断消耗过多的精力来满足所谓的好奇心。

这种悲观主义，与希尔伯特1900年向着数学家吹响的伟大号角截然不同。哥德尔的思想很快被同时代的数学家所了解，以至于很多人一度以为数学的结局是离开自然科学，成为人类认知和描述世界的工具，与语言和文学没什么明显界限：数学简直是一门人文学科。进一步而言，哥德尔不完备定理让数学家对数学的工具性追求都不能尽善

尽美，这甚至与人的主观努力没有关系。数学家似乎要在这种如芒在背的痛苦下探索可能的创造性。

数学曾无可辩驳地、信誓旦旦地主张物理学是基于数学的，因此由于数学的不完备，物理学也将无法发现所有真实的东西。从20世纪开始，从量子力学开始，利用诸如矩阵运算等数学工具，利用概率对可观察量进行预测，物理理论已经构建了更复杂的版本。但还是有人会说，包括最直接的牛顿经典物理，不可判定性是常见自然现象，系统的不可判定性是已经自然存在的规则而不是新东西。

哥德尔不完备定理限制了我们理解宇宙所用的必要手段，它构成了我们研究自然时的基本障碍。因为似乎在哥德尔不完备定理的作用下，数学结构将永远嵌入更深层次的、以直觉的智慧和朦胧为特征的思维。对于物理学家来说这意味着要面对有限的确定性和准确性，即在物理学的纯粹思想和认知中也会有一个边界。这个边界的一个组成部分事实上就是作为研究者本人的科学家或者那个爱因斯坦不喜欢的海森堡不确定性原理。

在没有自洽工具的情况下，基本粒子的所有理论，包括夸克理论和大统一理论，本质上都可能缺乏足够的数学工具来描述那些凭借先验性直觉和实验观察所表现出来的矛盾，这也许涵盖了量子理论本身、暗物质和暗能量相关的研究。然而自然界只能是它的本来面目而已，自然界存在本身就是完备性的证据，即使我们所发展的数学理论碰巧在一定时间内对已知的物理现象能够准确地进行说明，但这也并不意味着我们可以获得永恒的真理。

但爱因斯坦还是说了，自然最神奇的地方是它可以被理解。

有趣的是，虽然不得不承认哥德尔限制了我们发现数学和科学真理的能力，但这也将确保科学可以永远持续下去。哥德尔证明纯数学世界取之不尽，用之不竭。没有有限的公设和规则推论可以涵盖整个数学，给定任何一组公设，我们总可以找到这组公设规定的系统内有意义但没有答案的数学问题。因此也需要不断扩展我们的认知来寻找

新的答案。我希望类似的现实情况也存在于物质世界中。如果我对未来的看法是正确的，那就意味着物理学和天文学也是取之不尽，用之不竭的。无论我们走向未来多远，永远都会有新事物发生，新信息进入，新世界被探索，从而不断扩大人类的生活、意识和记忆。

科学界的乐观主义者把哥德尔不完备定理看作是科学研究永无止境的基本保障。他们将科学研究视为人类精神世界的重要组成部分。如果科学研究有哪些终极的方法可以彻底完成，那将对人类的精神世界产生灾难性的影响。与悲观主义者相比，乐观主义者将哥德尔解释为他确定了人类的思想无法全部知道自然的秘密，可能甚至是无法知道大部分的秘密。我们所认知的恐怕永远是自然的一小部分，未知的才是更大的部分。

哥德尔自己的观点倒有点让人出乎意料。他认为我们在看到数学和科学的真理时所依赖的直觉工具，有朝一日也会被新的工具来认知，从而成为像逻辑一样稳定的工具性知识。“我没有看到任何理由让我们对这种认知能力缺乏信心。数学直觉，不仅仅是猜想，它促使我们建立起物理理论，并让我们期望在不远的将来可以认识这些不可判定的直觉的来源。就今天而言，相信这些不能被判定的问题存在的合理性是有意义的，它们可能会在将来成为可判定问题。”

在物理研究实践中，哥德尔不完备会导致什么样的问题？我们已知的物理理论，几乎可以对各种各样的情况做出准确的预测并解释观察到的现象。几十年来物理学家所追求的“大统一模型”，有一天，我们可能会对它的无能为力感到惊讶，它不可能在它的框架内容纳它想统一的问题。

大统一理论致力于提供粒子物理中的统一理论。这些理论的早期版本有说中微子必须具有零质量的特性。现在，如果观察到中微子有质量非零的，那么我们就知道了新的情况不能适应我们原来的理论。我们会遇到某种不完备的情况，但我们可以通过延伸或回应它来修改

理论以包括新的可能性。因此，在实践中，不完备看起来预示着理论上的不足。

一个同样有趣的问题是有限性问题，关于宇宙的体积是否有限以及自然的基本粒子（或者最基本的可能实体）数量有限或无穷大。宇宙的物理可能性是有限的，虽然数字上可以很大。但是，无论如何物理规律所指的原始数量的大小，只要它们是有限的，由此产生的相互关联系统将是有限的。我们应该强调这一点，虽然我们习惯性地认为空间和时间是连续的，这只是一个非常方便使用简单数学的假设。人们相信空间和时间是连续的，不是离散的；然而在最基本的微观层面，量子引力理论假设时空并不连续：它引入了时空的离散性。奇怪的是，如果我们放弃这种连续性，时空结构就会变得无限复杂，因为连续函数可以通过它们的值来定义有理数，但离散的时空却不能。毫无疑问，我们可以方便地使用已有的数学结构并使用无限的概念，但这可能是一个为了看起来方便而做的数学近似。

数学家阿尔弗雷德·塔尔斯基（Alfred Tarski）表明，与皮亚诺算术规则中所要求的加法和乘法不同，只有数学加法或乘法下的一阶理论是完备的。这是相当令人惊讶的，可能会给物理学理论带来一些希望，基于实数或复数和简单算法规则可以逃避理论的不可判定性。塔尔斯基也继续展示了许多在物理学中使用的数学系统，如晶格理论、投影几何、阿贝尔群理论是可判定的，而其他的数学系统，尤其是非阿贝尔群理论不是可判定的。尽管物理学家似乎很少考虑到这些结果对最终物理理论发展的影响，但我们不能完全排除一种可能性：深层结构宇宙可能植根于比算术更简单的逻辑，因此它是自然完备的。像我们前面讨论过的，底层结构包含可以只有加法或乘法但不是两者兼而有之构成的一个完备的算术体系。或者，利用现实的基本结构——几何变量的简单关系，或者来自“大于”或“更小”的简单关系——比较关系，或它们的微妙组合都可以保持系统上的完备。事实上，爱因斯坦的广义相对论取代了许多像力这样的物理概念，使用了

曲面的时空结构，依赖于黎曼几何的相对简单的完备性，从而回避了代数所带来的更复杂的完备性问题。

还有另一个重要方面需要考虑。即使逻辑系统本身是可判定的，它也总是包含无法证实的“真理”。这依赖于选择哪些公设来定义系统。在公设被选中后，所有的逻辑系统能做的就是从中推断出结论。在简单的逻辑系统中，如皮亚诺算术，公设的合理性似乎相当明显，因为我们正在回溯和审视，从而形式化我们几千年来一直在直观地做的事情。物理学的公理或定律是物理研究的主要目标，它们绝不是直觉明显的，因为它们所影响的范畴可能远远超出我们经验的结果，这些规律在很多情况下是不可预测的。譬如涉及对称性的自发破缺，这个是杨振宁和李政道获得诺贝尔物理学奖的发现。它至少表明数学上的优美对称和数学所希望的自洽而完备，未必是自然界的样子，我们只能适应它，而不能用我们觉得好的工具来规范它。它也表明，从实验结果中推断出规律并不是我们通过逻辑推演或者数字计算等依赖于计算机程序的方法所能完成的事情。它需要人接受已有的自然事实，通过判断和适当的抽象得出易于理解的数学模型。

因此，我们在形式系统的研究中发现了一种完全不同于物理的科学。在数学和逻辑学中，我们首先定义一些公设和系统法则。然后，我们可能会尝试在维持其自洽性的前提下，从公设中推导出尽可能多的定理。在科学方面，我们不能自由地选择公设体系。我们正在努力寻找系统规则和公设，尝试用一个或多个公设最终可能产生我们看到的结果。正如我们之前强调的那样，我们希望总能找到一个系统，它将产生符合观察结果的规律。虽然这是一套被逻辑学家和数学家重视的无可辩驳的叙事方式——公设和演绎法则，但科学家更感兴趣的是发现而非简单假设。人类不止一次地希望出于某种原因只有一套可能的和合理的公理或物理定律。到目前为止，这似乎并不容易存在——即使这也是我们无法证明的。

物理学家对对称和美学有过执着的追求，尤其是理论物理学家希望在对称和美的要求下宇宙能够受极少数定理的约束，简单和对称的法则就能产生无限数量的高度复杂的形态。然而，不对称的状态似乎才是自然的本来面目——我们自己就是其中的一种，我们所生活的宇宙恐怕也是。因此，没有理由担心哥德尔的不完备定理会令人沮丧，我们依然会期待寻找自然法则的数学描述。自然的基本规律来源于弱作用力、强作用力、电磁力和引力，它们都是通过被称为局部规范的理论来维持特定的数学对称。随着这些力变得统一，涉及的对称性数量将减少，直到最终可能只有一个完美的对称性决定了自然法则的形式——一个解释一切的理论，其中超弦理论是可能的候选。自然法则应该在一个真正意义上“简单”和高度对称，但它是什么我们并不知道。最终的对称的理论应该足够丰富、拥有许多属性，这样才能容纳所有的基本力的表现形式。没有理由说哥德尔不完备定理应该妨碍了我们对这一切的追求，包含对控制自然法则的对称性的追求。但我们依然要怀疑这种天生的对称性的正确，正如对称性也自然地发生了破缺，还并没有更深层的理论来解释这种不对称。对于超弦理论，当今的物理实验无法检验其真伪，这也给这个理论最终获得认可提出了致命的问题。

每一次物理学的进步都伴随着新的数学工具的出现和使用。第一次物理革命是力学革命，需要研究的物理现象是天体的运动。牛顿不仅要发明他的万有引力理论，而且还要发明微积分这一套新的数学来描写他的理论。第二次物理革命是广义相对论，爱因斯坦发现了第二种场形态物质——引力波。他需要引入数学中的黎曼几何来描写这种新物质。第三次物理革命是量子革命，这次革命揭示出，我们世界中的真实存在，既不是粒子也不是波，但既是粒子又是波。这种莫名其妙却又真实的存在，可以用量子力学来解释，而量子力学则是建立在数学中的线性代数和群论的基础之上。

但最近二十年来，尤其是量子计算和量子信息的发展，加深了我们对量子力学的认识。量子信息中所涉及的多体量子纠缠，正是我们

既说不出来，无法描述，又没有名字的新现象。人们因此正在发展一套新的数学理论作为工具。

这次正在进行中的物理学新革命是非常深刻的。因为这次革命试图用纠缠的量子信息来统一所有的物质、所有的基本粒子和所有的相互作用，甚至时空本身。而凝聚态物理中的拓扑序、拓扑物态，以及量子计算中的拓扑量子计算，都是多体量子纠缠的应用。正是通过这些物理研究，我们发现了多体量子纠缠的重要性，并引入了长程量子纠缠这一相关概念。

完美主流的几何的眼光，并不一定是认识世界的正确方法。从量子革命以来，我们越来越意识到，我们的世界不是连续的，而是离散的。某种意义上，建立在几何思路之上的广义相对论、规范场论、量子场论太对称而完美了，以至于让我们误以为它抓住了宇宙的本质，误导了我们。

近代数学发展的也正是从连续到离散、从分析到代数的趋势，也提出了离散的代数是比连续的分析更本质的观点。20世纪60年代由格罗滕迪克（Grothendieck）学派发展出来的代数几何理论正是这种思想的代表，代数几何可以看作是实现了连续和离散的统一的几何理论。这和物理学从经典到量子的发展对应。而实现统一的语言当然是代数的，这有可能是一个超越了集合论的、全新的数学语言，也是代数几何的基础语言：范畴学。在范畴学中，人们不再试图把物体分成更小更简单的基本构件，甚至不去考虑物体的内部结构，人们试图通过这个物体和其他所有物体的关系和作用，来了解这个物体。这些关系正代表了这个物体所有可能的性质，一个物体的所有可能性质，也就完全定义了这个物体本身。而物体这一抽象的概念，以及物体所有可能的性质，是由这一堆关系来定义的。这就是范畴学的精神。

我们可以通过一个有意思的问题把数学的可解问题和物理连接在一起。在狭义相对论中我们提到了“孪生子悖论”。假设两

个双胞胎有不同的职业，弟弟留在地球上，而哥哥在一架太空船中以接近光速的速度飞走了。当他们最终重聚时，相对论预测哥哥会发现弟弟更老了。这对双胞胎在太空中经历了不同的职业生涯和时间，哥哥经历的加速和减速使他更为年轻。我们是否可以将计算机送到如此极端的旅程中以便完成当哥哥返回其出发地时，做无数次操作？如果哥哥可以充分加速他的宇宙飞船，他可以在自己的时钟上记录有限数量的宇宙历史，而他的双胞胎在他的时钟上记录了更长的时间。如果可以通信的话，他们是可以把*NP*不可解的问题的答案相互传递的。这时候弟弟可以在地球上建造一台够大的机器做足够的计算，然后通过发送信号来代替哥哥在他所在的飞船上的有限时间内的操作。哥哥需要解开一个利用大数分解加密的密码，他把这个问题发送给弟弟，弟弟用地球上的计算资源解开这个问题，需要10年，解出来之后，再把问题的答案发送给哥哥，由于哥哥的时间比较慢，所以事实上他不需要等待10年。同样的办法，面对一个超级计算任务，哥哥有可能通过直接搜索决定。图灵的不可计算的操作，甚至是一大部分停机问题，可以在一个有限的时间内转化成为一个查表的工作。这真的有可能吗？这样我们就做了一台相对论计算机。

这个计算机有一个小问题。在爱因斯坦的理论中没有信息传输的速度可以比光速更快，光和引力不可能变得任意强大。人们也不能任意地将时间延长，黑洞的“事件视界”保证了如果我们寄希望于无穷长的时间，一个拥有无穷长计算时间的机器并不能向黑洞外发送信息，计算结果我们还是拿不到。但这并不意味着所有相对论计算机都要被禁止。爱因斯坦时间的相对性是对所有观察者的要求，相对论计算机开辟了一些有趣的在有限时间内的任务新可能性。这似乎允许在某些时空，人们可以通过发送计算机直接搜索的结果来回避某些哥德尔不可判定问题，可以向计算输入端发送问题的答案。一个时空和另外一个时空的问题列表之间有一个简洁的一对一对应关系以及使参与者决定改变逻辑计算的复杂性。

对自然界物理规律而言，我们从未“看到”自然法则：它们是柏拉图式的表述，它们是我们对自然界观测的一种萌生于人脑并在科学共同体中达成共识的表达。但，我们可以声称我们见证了这些自然法则的结果。这是一个重要的区别，因为结果与作用于它们的规律不同。自然界不一定选择对称，正如哥德尔所指出的，语法无法完全代表语义，自然界的复杂性存在且不需要具备我们希望的完美和对称。这很幸运，因为如果它不是这样，我们不可能存在。如果自然法则要求完整的对称性就没有任何事情可以发生。19世纪最让物理学家担心的一件事情是热寂，根据熵增加原理，宇宙最终演化到没有特殊的时间和地点的结构，没有方向不对称，任何时刻都没有发生任何事情，一切都是不变的、空洞的。

从芝诺开始，在许多方面古人受到无限的悖论的挑战。图灵关心哥德尔不完备的现实问题，思考是否有可能建立一个可以执行的“无限机器”，在有限的时间内完成无限的任务。当然，这个简单的问题需要一些澄清：“可能”“任务”“数字”“无限”“有限”究竟是什么意思？牛顿力学则允许无限机器的存在，比如我们前面讲到的支在一个无穷大桌面上的无穷长的杆子，但相对论至少限制无限机器的功能，因为没有哪些信号可以以无限的速度行进。哥德尔不完备定理没有限制任何在物理学的整体范围内对自然规律的理解。物理学利用数学，大自然利用的东西可能比算术所用的东西更少、更简单。

在标准的实证论科学哲学看来，物理理论曾经毫无心理障碍地无偿居住在柏拉图式理想数学模型的天国中。人们相信，一个模型可以任意程度地详细，可以包含任意多量的信息，而不会影响它所描述的宇宙本身。但我们不是上帝，我们不能从外面观察宇宙。相反，我们和我们的模型两者都是我们所描述的宇宙中的组成部分，因此一个物理理论包括人类对我们生存的宇宙的认知都是自指的，就像哥德尔不完备定理所说的那样。人们因此可以认为它或者是不自洽的，或者是不完备的。这样说来，我们迄今所有的物理理论也可能既是不自洽也是不完备的。

总之，哥德尔不完备定理还是引发了探索数学逻辑和物理现实世界关系的新问题，它不应该证明我们寻求和探索自然的好奇心只是为了消耗我们多余的热情：哥德尔不完备定理并没有限制人类对问题和假设的探寻。

我们也应当注意这样一个事实：只有当我们知道机器运算规则自洽的时候，哥德尔论证才适用。但人类本身可能就是不自洽的，因此，不自洽的机器可能是人类思维的更合理的描述。牛津数学家戴维·卢卡斯说，尽管人类在某些情况下不自洽，但这与形式系统中的不自洽性并不相同。形式系统允许其内部的每个句子被判定真伪。然而，当人类得出矛盾的结论时，他们并不坚持这种矛盾，而是试图解决它。在这个意义上，人类是自我纠正的。自我修正机仍然会受到哥德尔论证的影响，但可以是一台根本不自洽的机器。

卢卡斯在总结自己的文章时指出人类思维的特征属性是走出有限系统公设的能力。他认为，思想并不局限于在一个形式系统中运作，它们可以在系统之间切换，不仅在推理系统中而且也在推理推理系统的事实。另一方面，机器是被限制在一个它们无法逃脱的形式系统中运作的，人工智能并没有超越这一点。因此，他认为，正是这种能力使人类思想本身与机器不同。对数学而言，自洽的要求是不能被放弃的，因为不自洽的理论可以推导出任何结论。所以我们只好接受理论的不完备性。这对于科学界来说是个莫大的好处。

不管怎么说，我们深刻地认识到了不完备带来的震撼。因为不完备的存在，逻辑机器和建筑在它之上的人工智能算法只会成为完善我们知识体系的工具，而不会带来超越我们所给出的有限假设所设定的新的认知。新的认知需要我们跟自然界接触，认识到现有知识体系的局限。这局限体现在有一些问题即使我们知道是真实的，但以现有的理论知识我们一样无法去证明它们。我们需要增加额外的假设，来使现有的假设体系能够描述这一内容。但一旦这样做了，我们马上又会面临新的困境。因为即使加入了一条新的假设，它与原来的假设一起

又是一个新的有限假设体系，一样会在什么地方出现在这个体系内无法证明的命题。如果我们大胆地认为我们的描述是具有普适真理性质的，那么我们立刻会遇到困境。

世界观和认知习惯的改变

在这本书里，我们把牛顿影响下的思想方法体系叫作经典体系或牛顿体系。进入20世纪之后，物理学和数学，甚至哲学（如果它还顽强地活着的话），都有了很大的变化，我们姑且把它叫作现代体系或哥德尔体系。现在我们大致上处于从古典体系过渡到现代体系的中间，但如果你考虑哥德尔的话，即使这么说，都是有问题的。我们无法定义自己的现在是现在。它并没有结束，在可预测的明天也不会结束。我还是喜欢把自己看作是一名物理学工作者，骨子里觉得数学是人文科学。但我愿意把在20世纪产生的新思想体系的冠名的地位给哥德尔，而不是给我们的“老祖宗”牛顿，或者我的直系师祖们，虽然他们开创了量子力学的时代。而且在我的另外一本书《量子大唠嗑》里，我确实在强调量子力学改变了我们对人类认知的看法。

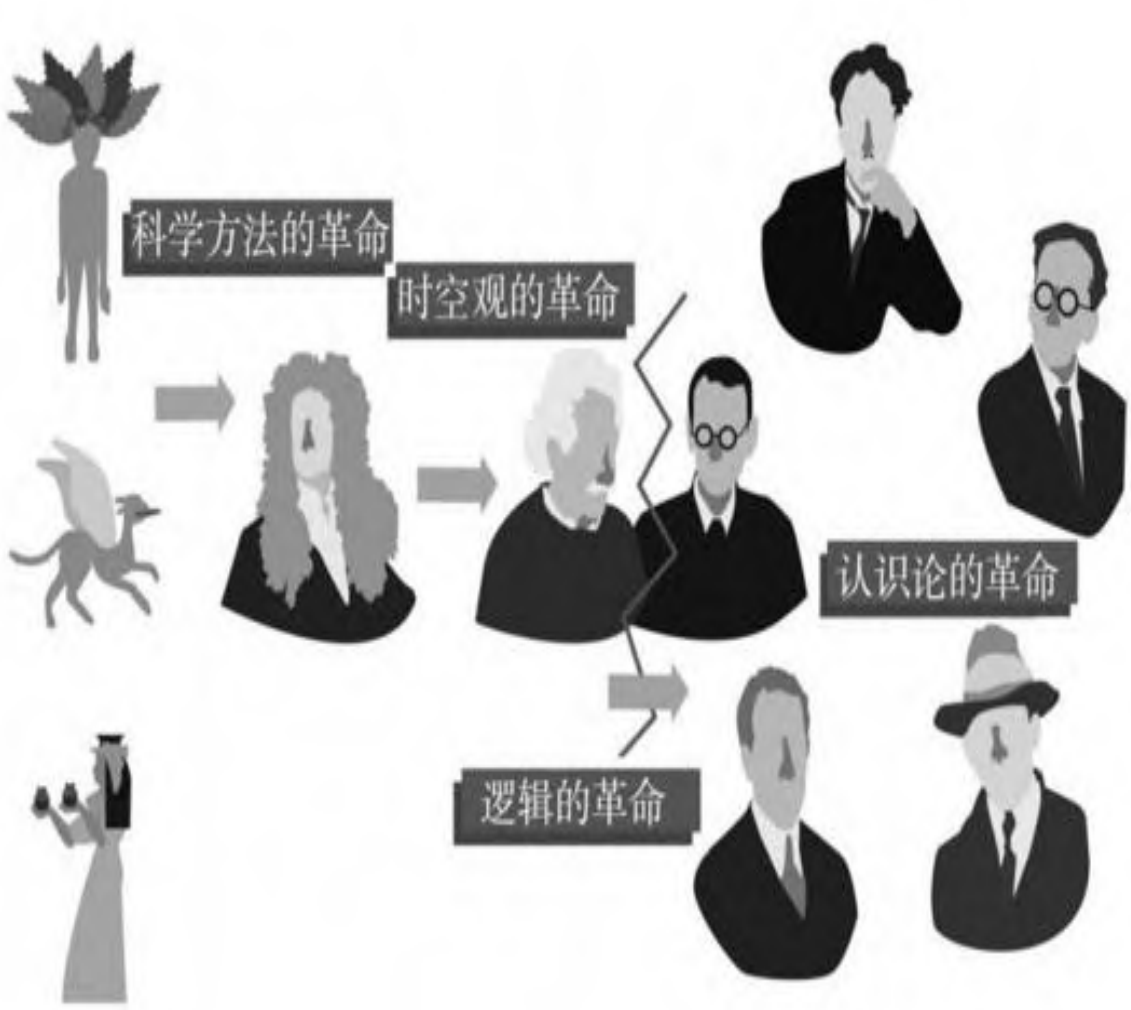


图10–2 人类世界观和认知习惯的改变

牛顿体系影响下的典型产物是客观唯物主义。我们很多时候把科学和经典理性体系混淆在一起，认为科学就是经典理性体系。我们曾经问：科学是什么？或者科学的理论是什么？并且把它作为一种衡量的绝对标准来规定：什么是科学的就高大上，什么是不科学的就怪力乱神。当经典体系与我们主观经验不符的时候，我们认为科学出了问题，从而从老祖宗那里又找来证据来说明对科学的承认是对中国文化的否定，又或者用科学的名词来似是而非地说明古代文化里的东西依然是不可撼动的真理。但我们如果重新审视科学的过程化本质，相信

科学本身是实践化的，是以科学认知为体验的过程，就不会把科学当作绝对真理的终点。

即使我不太喜欢“革命”这个词，但它姑且代表着重大变化。从古代到牛顿，人们找到并且稳定了系统的研究方法，可以从中不断地寻找和优胜劣汰出比较扎实的知识，而这些知识又进而成为寻找新知识的工具和手段。同时，牛顿也建立了经典知识体系，在这个经典体系中，科学本身是客观的，不以个人意志为转移的，人们努力地排除任何主观性的直觉。牛顿建立了一个机械的宇宙观，也构建了一个静止的宇宙。时空是一切的大舞台，提供了宇宙间一切事物的背景，所有事物的运动都可以以某种规律来描述。所有目标、对象，都按照一定规律所描述的剧本来演出，到时间上去，扮演完自己的角色就退场，然后变成其他角色，或者其他角色的一部分。而这些规律本身是可以被了解的，甚至可以通过合适的方法来预测。预测本身含有了时间的概念。这种预测指的不光是星球或者苹果，甚至是人的思维，那不过是一堆原子的运动而已。这就麻烦了，难道我们真的没有自由意识？包括我们的人生都是事先设计好的剧本？人类的生老病死、悲欢离合，都是到时间就上演的剧本，是可以事先算得出来的？而容许发生这一切的宇宙，像是某个超自然的力量，事先准备好在那里的舞台。无穷大也无穷久远，静静地看着这一切的发生，这时候，自然便是上帝的化身。

爱因斯坦的第一个突破，在于发现了时空的相对性。时间和空间并不是一个静止的舞台。时间和空间跟物体的运动相关，而事实上，我们正生活在一个永不静止的宇宙中。从那著名的大爆炸之后，宇宙就一直在膨胀，而最近的实验数据证明，宇宙还是会继续膨胀下去的，而且越来越快。但就宇宙不是静止的这一观点，对深受经典物理影响的爱因斯坦而言，虽然他奠定了这个发现的基础，但一开始他并不相信。

爱因斯坦的第二个突破在于量子力学。在他的奇迹年——1905年，他做了另外一项工作。他阐明了光电效应，即光可以成为粒子，一颗一颗地跟电子相互作用来交换能量和动量，从而明确了光是量子化的。光作为量子，有粒子的特性，也有很容易被观察到的波动效应。这一开始并没有让人觉得很惊讶，但当它的波动性被解释为概率的时候，就真正触动了爱因斯坦的信仰。他不相信上帝是通过掷骰子来决定宇宙的运作的，他奠定了量子力学的基础，但他却不相信它。

伟大的时代有伟大的人同行，是上帝给伟大的人的恩惠和眷顾，要不，他们就太孤独了。

哥德尔事实上改变了人类认知的习惯。

哥德尔的工作在于他打碎了自亚里士多德以来人类努力建立的完美理性框架。这个框架从古希腊发端，到牛顿之后的三百多年中的一系列光辉熠熠的人物，他们把经典理论完善成更深层次的信仰。世界是可以被认识的，应该有一套切实可信的方法，人类需要不断地发掘真理并验证它们。世界的运行不需要有更多的人为因素参与，世界的规律是客观的，即使当下的现实经验有些许不客观，但终有一天会客观的。哥德尔，以优美但并不简洁，但与他同层次的数学家相比已经足够简洁的方式证明了，这不可能。

他实际上告诉了我们，不会有一个机器，基于我们所希望的理性，自己去完成对真理的探索。对未来的人类而言，这当然是安全的保障。他也告诉我们，对于人类思维和直觉，我们还有太多东西未知，我们甚至不可能把人类的直觉移出知识本身。这事实上颠覆了我们对客观真理的信仰。这信仰从亚里士多德开始至牛顿到爱因斯坦，已经存在两千三百年了！

但是爱因斯坦跟哥德尔却是莫逆之交，他们是彼此精神上几乎唯一可以互相交流的伙伴。即使哥德尔本人也还是认为自己是柏拉图主义者，这跟爱因斯坦是一路的。他俩都相信有个客观存在，而这个客

观存在无疑是可以跟超自然力量对应的，不管他是否像人一样是实体思维者。而哥德尔也曾认真地证明过上帝的存在。

那后面还有一堆人，普朗克、德布罗意、玻尔、薛定谔、狄拉克、海森堡、弗兰克，等等。量子力学是现代工业的产物，并不是某一个人的天才杰作，恐怕也是因为它离我们太近而没法用一两个人名来概括。量子力学对人类认知本质的影响正在扩散，一开始我们以为它只是描述微观世界的理论，但它的影响很快扩展到了非微观世界。我们希望将微观世界与已经认识的宏观世界兼容起来，但发现这并不容易。量子力学告诉我们，我们追求的客观从物理的角度讲可能是不存在的，观察者不得不成为体系的一部分。量子的随机是真的随机，不存在深层次的理论来决定它，或者预测它。甚至是我们经典逻辑观所依赖的因果关系，在一定的实验设计下也会被打破。

旧的世界被颠覆，新的还没有建立起来，今天，我们所面对的是一个新大陆出现在地平线上。但敌人的敌人未必是你的朋友，经典体系的崩塌并不意味着旧世界的复辟。经典体系毕竟留下了稳定的方法，这方法让它的拥有者无所畏惧。挣脱了经典枷锁的人类会更自由，但经典体系为人类提供了迄今为止最好的工具。

当人们追求绝对真理的时候，实际上就已经偏离了追求真理的正确道路，其结果是，发现绝对真理这件事情本身就是悖论，因此我们退而求其次，只求方法的靠谱和在限定前提下得出相对可靠的结论，并随时准备推翻它。

我本来想写《量子大唠嗑》的注释版，至少很多朋友希望我这样做，因为那本书读起来，尤其是第二部分确实有些艰难。难怪乎我所训练的东西，用通常的语言来描述它本来就是不适合的。1900年后，数学界和物理学界开始了一种有意思的运动，尽量减少用自然语言而用数学来表述物理内容。但这时数学工具已然艰深了很多，但没办法，准确地了解一个行业的知识不再是由门外汉稍加努力就能获得的。然而知识的进步，也绝不是独门秘籍，不可以传授。它就很无辜

地待在那儿，简单的直白的不知道怎么形容其高深。既然工具不完善，而观测结果本身也随着我们选择的工具变化。我们事实上缺乏或者不能找到放之四海而皆准的标准和手段，就只能采取务实主义的态度，“哪怕它真理无穷，进一寸有一寸的欢喜”。胡适的老师杜威没有某部鸿篇巨制来描述他的核心思想，因为务实主义的思想琐碎繁杂地渗透在字里行间的每一句话里。我们仔细来看，这更符合我们的真实经验。事情并不是总能提纲挈领、头头是道、井井有条的，大多数时候是纷繁复杂的。但这之下也有潜在的思想脉络，找到这些脉络，验证并使用它们，往往能起到积跬步以至千里的效果。我们在每一件事情中都这么做的时候，就可以获得这样一种经验，我们可以逐项地去优化和改进，一点一点来。以日拱一卒的精神，一点点迭代，最终获得长足的进步。

第十一章

创新和突破

小孩子们很喜欢玩脑筋急转弯，谜底也都有道理，只是我们平常不那么想。人的思维本来是发散的，人类的语言也可以无限诠释，但为了沟通方便我们逐渐训练彼此用习惯的方式来构成常识。这些常识在帮助我们学习新的知识的时候也会束缚我们的想象。哥德尔提示我们，找到这些常识的不自洽，跳出习以为常，自然就创新了。

你有近似的答案或者可能的信仰，以及对不同事物不同程度的确定性，但你对任何事都没有百分百的确信。总的来说，直到四岁左右你都是自由的。然后你去语言学校，然后被人引导着进入某个领域。你失去了你曾有的独立。如果有足够的课程，你可以保留一些……然后你带着一种近乎良好的感觉每天工作8小时，就像你在做什么大事一样。然后你结婚了，如同结婚是一场胜利，接下来你有孩子，仿佛孩子也是一个胜利……婚姻，分娩，孩子，这是人们必须要做的事情，因为没有别的事可做。没有荣耀，没有烟雾，没有火花，非常非常平淡……你就像毫无选择一样陷入了一种你“应该”成为什么样人的局限之中。你终于被模型化并融入了你应该成为的样子。我不喜欢这样。

——查尔斯·布可夫斯基（Charles Bukowski）

每个人都会正义化自己的选择。我做过最引以为傲的职业是一个物理学家，或者说是一名物理学工作者。就物理学者的信仰而言，我相信世界是物质的，不存在超乎检验的灵魂和力量。在人的思维或人类的智能方面，我相信无论它看起来多么神奇，都还是可以回溯到物质本身的。然后我们就进入一个投机取巧的推理逻辑。既然我们从今天的数学和物理的角度都不能够完全解释人类智能，哪怕是低级一点动物智能的起源，那么我们就把不懂的东西归于一类，用不懂的东西来解释不懂的东西，给出一个既然不是这个，那就极有可能是那个的近乎敷衍的答案。这样说不是完全没有根据。人跟人在一起面对面沟通的时候，哪些东西是不能被经典信息数字化的呢？声音可以数字化，图像可以数字化，我们用虚拟现实、远程会议总可以把这些实际测量出来的物理参数数字化，再通过合适的介质储存、传递，即便是

人的动作和表情，包括眨眼、脸上肌肉的颤动，以及环境的温度、气味，它们到底有什么不同？这个问题无法回答。因为一旦可以回答，它就是可以描述的，一旦可以描述，那么图灵机就可以把它变成一个图灵机可以运算的算法。但我们知道，这些算法不能描述所有的事实。同样的逻辑应用于我们所说的感情、创造力，我们不知道它们的细节和源泉，一旦知道了，就可以把它们算法化，让它们成为一段代码。我们事实上又知道这些事情真实存在，我们一点一点地认知它们，但总有一部分是我们通过直觉可以感知，却无法描述的。这是哥德尔不完备定理告诉我们的，也是第一个基于图灵机的人工智能和我们之间无法逾越的障碍。另外一个障碍来自哥德尔所提出的计算复杂性的问题，有些算法可以描述，但不能被有效计算。当然，通过算法的优化和计算能力的提升，把一些判定性问题转化为概率问题，这样会有些改善，但是否意味着可以被最终等效，复杂的问题是不是仅仅是时间问题？我们并不知道。

那么，我们的创造力来源于量子力学吗？如果我们假设人脑是一台量子计算机，这可能会给我们一些有意思的猜想。

第一，量子可能赋予我们跨越哥德尔不完备定理所限制的寻找有限系统外假设的能力。因为量子力学并不拘泥于经典逻辑，它即使也受哥德尔不完备定理的影响，这其中的影响方式也是与经典逻辑完全不同的。我们知道，量子力学也许是完备的而不是经典意义上自洽的。这给了突破经典意义上的有限公设范围之内进行讨论的可能。它可以不断与环境相互作用而获得新的假设，因为它是事实上完备的。

第二，基于长程关联的联想能力。量子现象允许量子体系具有长程关联的能力，虽然这更像是一个名词的借用。因为还没有任何实质的证据能够证明在人体这么高的温度下，量子关联还能奏效。但脑的神经突触数量巨大，是不是可以让即使很短的长程关联也有了规模化发生的可能？

第三，基于超越*NP*不可解问题的直觉判断能力。人会面对诸多的选择，这些选择的下一步所叠加出来的选择，构成了一个复杂的选择网络。描述它最简单的模型是二叉树。在一个复杂的网络中寻找答案，从经典计算复杂度来看是一个*NP*不可解问题。如同我们在上一章所讲，量子随机行走可以使计算时间大大减少，从而看起来答案像是从天上掉下来的，这构成了我们所说的直觉。通过量子随机行走人们可以毫不费劲地遍历所有可能而直奔合理的解决方案。《超级玛丽》是一个*NP*不可解的电脑游戏，它的很多步骤都要做不同性质的选择。依据类似的经验，在实际的动手经验中，例如加工一个金属工件的每一步的选择，虽然是基于经验的，但这些经验的积累和选择，是要通过直觉才能实现的。为什么这样做，为什么不那么做，为什么用这么大力气而不用那么大力气，有经验的金工师傅可以很快地做出每一个阶段最佳的选择而迅速地完成加工任务。遍历所有可能并做出比照优劣的判断需要大量样本来做尝试，很多情形下这并不现实，也不发生在人做出这一系列动作的过程中。

第四，基于理想随机数的直觉感知能力。人的思维有些时候会跳跃，会基于更多生活经历而给出直觉的判断。有时候这些直觉的判断随意而任性，我们没有十足的证据能够说明这样的随意和任性带来了人类思维相对于图灵机的优势，但至少是带来了不同。

我内心深处希望这些猜想是真的。我至少不希望有一天我们扮演上帝，造出来远超过我们自身思维能力的“事物”。因为这个“事物”一旦被造出来，它便不会受任何人类的钳制。我深信那里一定有某种机制，限制了我们扮演上帝的能力，不是规则，因为规则一定会有因为不完备而导致的漏洞，那一定是因为物理规律本身，尽管我们还没有充分地认识它。

如果人可以还原为一大堆原子，那我也宁可相信人的大脑是量子计算机。由于我们对于量子力学的基础问题尚未明了，基于量子计算而做一个像人脑一样的计算工具恐怕未必是件容易的事情，这中间可

能还有尚未知晓的困难，比如量子的退相干效应会让量子系统在达到一定尺寸的时候坍塌。当量子比特越来越多的时候，为了维持稳定的量子干涉系统，实验系统的复杂性会呈几何级数的增加。这恐怕不能把实验现象和实验手段分割开，它们也许就是一致的，无论怎样努力，我们最终被制约在量子退相干机制里。这与我们如何努力地提高实验能力无关，量子不确定性决定了我们对认知的困难，这样的事情是否也会发生在我们构架量子计算机的过程中？

贴了标签的人生

你从哪里来？你是干什么的？我们往往喜欢这样给一个人贴上标签。但这个问题是越来越难回答了。

我出生于山西太行山下的一个小镇医院，祖上一定要说是马氏正根儿。两岁时跟父母迁到太原，上大学时第一次离开山西。从在北大待的时间来说算是北大“土著”，而且是学核物理的“土著”。读完大学到了牛津，然后是在杜伦大学待了半年，华盛顿大学待了两年，加州大学伯克利分校待了两年，芝加哥大学待了一年。工作单位曾在上海，后来长住北京，人生这么颠沛流离，很难说我从哪里来。而生活中也有越来越多的人像我这样，因为生活的选择，读书、工作，颠沛流离地不停搬家。

至于我是干什么的，也是个很难很难回答的问题。我们习惯给人贴上个标签，这可能是经典的理性主义，分了类冠上名字就好理解了，可实际上我也不知道自己应该算作干什么的。

我到大学毕业时体育都基本没及过格，到牛津以后因同学院同学的鼓励而上了贼船，划了赛艇，在牛津的伊河第一次体会到了竞技体育的乐趣。很小的时候我喜欢出风头上舞台，后来不知怎么特别不喜

欢当着众人前讲话，但在读研究生之后我开始尝试做话剧。从牛津大学的《雷雨》到加州大学伯克利分校的《不是犀牛也不是海鸥》，从演员到导演，我被叫作马导的历史远早于后来在中国科学院当博导。很多人都说一辈子要横穿一次美国，从华盛顿到旧金山，我就那么开着车横穿了一次。但我这么能写字的主儿，居然过了这么多年都没有写过关于那次公路旅行的游记，哪怕是一个字，也不曾贴过一张照片。也许是因为可以写的东西太多，所以干脆把它当作经历和体验，放在脑子里，不啰里啰唆。总之，我似乎干什么都不那么专业，被同行的专家们诟病，本来少年得志，非要不混主流，把才华浪费着玩儿。

但我想这就是像我这样的人的人生，劳什子听他贴什么标签。

现代医学真是让终身职业成了少数人的选择，那么难！在100年前，人的平均寿命才40岁，一个人花20年读个书，花10年卖书袋子，再有10年倚老卖老，就该死了。所以一辈子从事一个职业一点都不难，成为通才的大师也不难，反正知识就那么多。但那不是我想要的生活。我在邻家乖孩子式的教育下，早早把家长和人民的希望和嘱托完成了，30岁成了世界第一的研究机构的物理学教授（从SCI，即美国《科学引文索引》的评价标准来说，中国科学院真的是世界第一了）。再往后仰之弥高，钻之弥坚，穷经白首做个院士，然后执杖乡里，指指点点让年轻人叫自己师祖，看起来实在太遥远。于是我胆怯了，做了逃兵。

然而，于我的家国情怀，我不仅没有做成逃兵，而且迎头上来了。有时候家国情怀不是自愿的，是它找来的。我开始担心这个国家会陷入“中等收入陷阱”，于是我五脊六兽地开始找点事情来做。面对这些急迫的问题，我那些关于科学的理想似乎可以放一放。于是我“半路出家”做了布道者，为中国制造业升级积极想办法。我一路自己给自己解压，一个像我这样“不务正业”的人，在目前这个人生的阶段，能为这家国做的事情暂且就只这么多了，“日拱一卒”看起来

是孜孜不倦的勤奋，其实是偷懒，本来可以一天走两步，拱两卒的，但偏偏拱完这一卒就想休息去了，等明天再来。我还有很多别的事情做，我是个务实主义者，我不会为理想抛头颅洒热血、毁家纾难。我打着小算盘的同时顺便做点和大道理相关的事情。也许过个十几年我的小事情做得差不多了，可以回头再把今时今日的事情再做一遍，但而今，在今时今日的努力中，我尽力了，不能要求我更多，the less is more（少即是多）。

我看到的很多人，现在做的事情与少年时初心并不一致，一致的反而凤毛麟角。如果现在一个十六七岁的少年跟我坐在咖啡馆聊天，他对我指手画脚地说你该干这个、不该干那个，你这辈子要是干这个了就会活得很失败，会违背初心，我恐怕难以抑制扇他耳光的冲动。但如果时光穿梭，那个孩子不就是少年的我吗？十几岁的自己见识那么少，为什么要决定我现在该干什么不该干什么呢？你会因为一个涉世未深对未来有无穷幻想的少年的执拗而评判自己人生的真实体验吗？

功不唐捐，你说这是巧合也好，对未知的探索也好，把内心执拗的东西说成是坚守初心，更多的是翻过了山丘的人对自己碰上的事情的解释，将自己的选择正义化的演义。真正的人生，永远是多变的、丰富的，而机会也许就在拐角处。当它来的时候，你能把握，而不被所谓的初心所羁绊，不亦乐乎。

在牛津读书的第一年，我补上了在北大物理学院没有学习到的金工实习和电路设计等手工技术，这些能力在我后来的研究中起到了重要作用。我从事的是超冷原子物理方向的研究，虽然做了些理论工作，但还是以实验为主。这个方向的实验对工程能力要求极高，需要从业者有很好的动手能力。电、力、热、软件、硬件、通信、成像、激光、真空、数据分析和物理，读研究生时几乎样样都要自己动手。博士毕业之后我的第一份工作是在美国国家标准与技术研究院（NIST），搭冷原子实验平台。后来到加州大学伯克利分校又做了一

个博士后，又搭冷原子平台。再后来在中国科学院做教授，给“天宫四号”搭冷原子平台。事实上，现代物理实验中很多设备要自己亲自动手做才能发现和探索新知识。冷原子行业有一句老话，温度每降低一个数量级，就会发现一个完全不同的物理世界。为了这个目的，这个领域的科学工作者花了大量的时间和精力来研究机械的、电子的、激光的、软件的、硬件的各种工业界新技术，迅速应用到自己的实验中来获得新的结果。不仅这个领域如此，为了新发现，今天的物理学界也普遍如此。

在加州大学伯克利分校时我意识到视频数据的实时处理是一个值得深入研究的问题，于是我的研究兴趣逐渐转向了用人工智能技术来处理视频。这些年从事人工智能的工作，我了解越来越多的人工智能技术，就越对人的智能充满敬意，尤其是能够动手做东西的人类的创造力。在加州大学伯克利分校做博士后期间，我和同事们曾经试图在量子极限附近设计实验来了解“测量”的本质。因为只有几个光子可以被吸收，所以实验需要系统极其稳定和敏感。我花了几个月的时间推导设计了一个极其复杂的算法，可以在充满涨落和噪音的数据中找到潜在的线索。拿给合作导师看时，他说，我们是做物理实验的，好的工作是干净的数据，而不是复杂的数学模型。这让我想起来我大学两年里参与夏商周断代工程里的甲骨测年，用贝叶斯（Bayesian）统计法把武王伐纣定在了公元前1046年1月21早上8点。我对play with numbers（拿数据调着玩）没有一点好感，多少受这些经历的影响。当贝叶斯统计又在深度学习中被广泛用到的时候，我们这些实验物理学家只能嘿嘿一笑。

在加州大学伯克利分校接下来的两年，我们都在花时间提升系统的测量精度，包括芯片刻蚀、在超低温条件下的单光子探测器和宽带超灵敏反馈电路等，这些都是要在实验室自己亲手做的，尤其是探测单光子所需要的超窄线宽的光学腔。这项技术，后来被用在了对引力波的探测上，所以那些伟大的工作中，有些角落是有我可以说一说的细节的。

老话说，纸上得来终觉浅，绝知此事要躬行。我一直不觉得我自己是个好的科学家，大概算个好的手工师傅，一个不差的物理学工作者。2009年我在飞机上遇到李宗盛，聊天颇为投缘。他说他也是一个好的手工师傅，醉心于做吉他，而我觉得在金工房里做半导体激光的基座，甚至做一把木梳子，都实在是令人感到充实而有乐趣的事情。动脑动手做东西给人一种内心的满足和安静，这种心态对后来我在工作中每每遭遇不顺和挫折时所要有的平和心态的养成颇有益处。

这些训练和经验，让我对动手能力的训练情有独钟。很多时候，不，应该说所有时候，创造并不是灵光一现的，而是水到渠成的事。人生没有白走的路，每一步都算数。

所以正如《量子大唠嗑》中所讲，人有很多面，世界也有很多面，我们只能设定一个环境，描述这样的情境里人是什么样子，有什么标签，而未来的世界也是这样，每个人都在自己喜欢的事情里享受生活的意义，没有谁，在什么地方，被定义着生活。

问题导向的学习

牛津大学是比较少见的学院制。大学是学院的联盟，而大学的系更像是各个学院之间资源共享的平台。学者既属于某一科系，又属于某一学院，因此做学问研究也形成了两种体制。院系可以提供传统意义上的学科研究支持，它大致遵循了近代以来形成的学科分类习惯。更古老的研究方式为牛津大学的各个学院提供支撑，学院可以为学者提供研究资助，做学问的人并不需要明确自己属于哪个研究方向或分科。这样的制度允许学者更多地以兴趣或者题目为导向去做研究，并不局限于某一个学科领域，通常为了解决某个问题，不断穿梭于学科之间和了解既有的学问之外的知识来拓展所需要的工具。

1999年牛津大学出版社出版了牛津大学纳菲尔德学院（Nuffield College）学者艾伦·戴维森（Alan Davidson）的《牛津吃货宝典》（*The Oxford Companion to Food*），这本书在一夜之间取得了巨大的成功，在全世界赢得了多个奖项和荣誉。这本书结合了严肃的食物历史、烹饪专业知识和偶然的趣味性，每一页都提供不同的视角。它囊括了滋养人类的食物的详尽目录，无论是热带森林的果实，西伯利亚花岗岩刮下的苔藓，还是动物眼球和生殖器。无论从文学角度还是烹饪菜谱的角度，这本书都极具参考价值。《牛津吃货宝典》掀起了人们对食物和食物历史的研究，作为一门新的学科，食物学在过去的二十多年中飞速成长。在国际学会、学术期刊和海量的文献中，全球各地学者广泛地探索人们日常生活中食物的意义。每年3月来自全世界的食物学学者都会聚集在牛津大学圣凯瑟琳学院（St Catharine College）开世界食物大会。会上的学术讨论不仅关注于食物和营养本身，而且几乎从每个与食物相关的角度来理解人类生活，无论是环境、政治、人文，还是市井生活、文化冲突或战争。这是个好的创新案例，食物学从艾伦·戴维森的个人兴趣发展成一门严肃的学科。回顾历史，哪一门学科在开始的时候就能归纳于某一门类的问题呢？这本身就是个有关因果的悖论，如果每一个学科都有学科作为其归类，那么它们是怎么开始的呢？可见，一直回溯到经典逻辑的第四条充分必然条件，都要求我们无中生有地创造，这是不切实际的，真实的学科发展必然像古埃及的太阳虫，这里滚一点，那里滚一点了。

新工程教育

写这本书的一个深邃的目的，隐藏在这里。自从做了物理学教授，我就越来越觉得工程的重要。这些年稍稍离开学术界而深入产业界，一手一脚地做点实际的事情，就更加觉得工程能力的有用性。然

而我们这些年的教育忽略了对工程人员的培养，风雨欲来的人工智能又叫喊着用机器取代人。我深深地觉得我们应该去找到人类与机器的差别，至少它应该影响我们今天的教育内容。谁都不想我们今天教给孩子们的技能，十几二十年后他们长大了才发现机器做得比他们要好得多。这样的例子屡见不鲜，我小时候学过算盘，邮局有电报员，印刷厂有专门的排字工…… 我凭着直觉感到，在车间伴随着时时思考并探索和尝试的动手能力，力学的、电学的、材料的，是无法轻易被机器取代的，相反，坐办公室的工作，却很容易被机器取代。我一直没有找到好的证明，直到有一天，跟我的导师基思·伯内特（Keith Burnett）先生聊起未来的工厂所应该营造的气氛。人们希望能够在未来工厂营造一种游戏的氛围，让年轻人以打游戏通关的心态从事创造性的工作。未来工厂也像今天的苹果公司的销售门店一样，窗明几净，有计算机设计终端，也有满地走的机器人。在这个生产场景里，人们试图创造的每一个工件甚至执行的每一个步骤，都是一个多选择的过程。这时人脑又像极了很多选择网络上行走的量子随机行走，经典计算不能够代替人类做出复杂决策，或者说至少不能像人脑一样可以有效地做出截断的判断。但说到了游戏，《超级玛丽》和《魂斗罗》都是经典的*NP*不可解问题。如果零件制作和工程师的创造行为可以用游戏化的方式来进行，那恰恰也可能是一种类似的*NP*不可解问题，这一代的计算机无法有效胜任的工作。人工智能催化的以数字产业为主的知识研发目前还很难覆盖手工业。除了机器人制造能力的限制，其中的主要原因可能会有其他更深层次的。比如，涉及基于大量操作经验而形成的直觉，这是目前人工智能很难与人进行比照的方向。因此，在制造业中，高级技术工人在工作过程中，所具有的结合数字化和制造业流程本身特点的技能，在人工智能时代会显得尤为重要。这就需要制造型人才不仅要懂得人工智能的计算机技术，也要懂得工业生产流程中的具体情况。

传统工程教育强调对学生进行基于学科知识的能力训练，体现出工程教育活动组织与开展的学科逻辑。由于学科逻辑过于强调学生对

工程学科知识的掌握以及学生认知能力的训练，因此传统工程教育容易造成工程教育活动的开展而忽视学生个体身心发展规律，忽视学生工程实践经验构建以及工程实践中学生的组织和沟通能力的培养。

基于这些考量，麻省理工学院从2017年开始开展的新工程教育改革采取了整合学科逻辑与心理逻辑的策略。整合的路径体现为以研究具体问题的课题项目为线索，围绕现代产业的实践和研究方法，构建机械、材料和系统科学的跨学科内容。每个课题为学生提供了前所未有的机会，让他们沉浸在跨越学科的研究项目中，同时获得所选专业的学位。新工程教育的教学方式发生了变革，强调以学生为本，关注学生的学习方式和学习内容，把学生真正置于工程教育活动的中心。不仅重视知识的获取，而且重视应用知识的能力。项目是学习制造、发现、系统和创造力的主要工具，它有助于促进学生从团队技能到人际关系技能再到领导能力的提升。在开展教学活动时，学校通过充分考量学生个体的认知风格、学习方式等方面的差异，选择最适合学生个体发展的学习方式，引导学生积极参与，激发学生的主动探究与自学能力，采取项目学习、小组学习、团队合作、信息化教学、智慧学习等手段，为学生成为引领未来工程发展的领导者奠定基础。

人工智能对生产效率的提高会使得产业界更加注重工程人才的学习能力和思维等方面的表现，原来强调以知识习得为重心的教育体系将会受到挑战。新工程教育应更注重对学生思维的培养，从而让学生在工程实践中面临各种未知与复杂问题时能够运用恰当的思维去思考、解决问题。麻省理工学院提出新工程人才应具备12种思维和能力：

1.学习如何学习（Learning how to learn）：学生利用一定的认知方法主动思考和学习。

2.制造（Making）：新工程人才发现和创造出新事物的能力。

3.发现（Discovering）：一种通过采取探究、验证等方式促进社会及世界知识更新，并能产生新的根本性的发现和技术的能力。

4. 人际交往技能（Interpersonal skills）：一种能够与他人合作并理解他人的能力，包含沟通、倾听、对话、参与和领导团队的工作等。

5. 个人技能与态度（Personal skills and attitudes）：包含主动、有判断力、有决策力、有责任感、有行动力以及灵活、自信、遵守道德、保持正直、能终身学习等品质。

6. 创造性思维（Creative thinking）：一种通过深入思考，能够提出和形成新的、有价值主张的思维。

7. 系统性思维（Systems thinking）：在面对复杂的、混沌的、同质的、异质的系统时，学生能够进行综合性、全局性的思考。

8. 批判与元认知思维（Critical and metacognitive thinking）：一种能够对经由观察、体验、交流等方式所收集到的信息进行分析与判断，以评估其价值及正确度的思维。

9. 分析性思维（Analytical thinking）：一种能够对事实、问题进行分解，运用理论、模型、数理分析，明确因果关系并预测结果的思维。

10. 计算性思维（Computational thinking）：一种能够把基础性的计算程序（例如抽象、建模等）以及数据结构、运算法则等用于对物理、生物及社会系统的理解的思维。

11.实验性思维（Experimental thinking）：一种能够开展实验获取数据的思维，包含选择测评方法、程序、建模及验证假设等内容。

12.人本主义思维（Humanistic thinking）：学生能够形成并运用对人类社会及其传统、制度和艺术表达方式的理解，掌握人类文化、人文思想和社会政治经济制度的知识。

通过教育，人类掌握了各种工具。以语言为例，它是我们表达自己的工具，没有它们，我们无法表达自己，无法与人沟通，但掌握了它们，它们就有可能成了限制我们思维的条条框框。哥德尔不完备定理无处不在，跳出有限假设的范畴便是创新。小时候就没有这些条条框框。有人说人在小时候是天才的，受教育了才成了庸才，但没有这些教育，即使有天才的想法也无法被察觉或者被记录，或组织合理的实验来验证这些想法，那些异想天开也就永远成为异想天开。我们只是要在掌握这些条条框框之后时刻提醒自己，这些武器也是我们的枷锁。这是实战的经验：我们辅导孩子功课的时候，他们总想到盒子外面，我们总要拿着小鞭子把他们抓回来。抓啊抓啊，他们就习惯了我们的思维方式，知道了我们想让他们知道什么。直到我们和孩子可以顺畅沟通了，他们也习惯了不往盒子外面想。等他们长大了，我们又要不断地教导他们要创新，要往盒子外面想。困难在于怎么保持这两者的兼顾，盒子里面的是理性，或者说我们总能把它梳理成理性，但理性不足以构成一个完整的人，也不足以构成完整的知识体系。



图11–1 创新：Think out of the box

突破常规而有所创新说起来也不难，但用到自己身上很难。我们承认和鼓励“不同”，但也尊重先验工具本身，知道它的工具和枷锁的双重性。当我们比较了人工智能和人的根本区别，也比较了经典系统和量子力学所预示的系统之间的差别，我们发现人类社会的发展趋势是我们不再那么需要服从纪律的劳动力，这些劳动力可以轻易地被机器人取代。相反，社会对人的科学素养和人文底蕴要求越来越高。这包括人对世界的认知能力和人与人之间的沟通能力，也包括人对自身的感悟能力。社会需要的是具有创造力、充满好奇心并能自我引导的终身学习者，需要他们有能力提出新颖的想法并付诸实施。

我们所设计的教育常常忽视人与人之间异常美妙的多样性与细微的差别，而正是这些多样性的细微差别让人们在智力、想象力和天赋方面各不相同。本来人的思维是自由的、可创造的、可沟通的，我们的教育系统的终极目标居然是把人训练成人工智能，而我们的教育考核指标在这个逻辑下就是给人工智能准备的。

远古人类未尝不是在个性化的生产中成了全能的设计师。为了适应不稳定的生存环境，人类必须拥有非常全面的生存能力和经验储备，才能够随机应变地躲避危险，获得食物求得生存。为此，人类在成长过程中，需要通过全面的训练来获得独当一面的生存技能。人类的生活丰富多彩，他们每天都可以接触不同的新鲜事物，还能发展和运用不同技能。那时候的人生活在多元的环境里，利用不同技能来应对新挑战。

工业革命先把人类限制在固定的土地和固定的工作场所，人们更多地从事以体力劳动为主的重复劳动，成为标准的劳动力而非知识和信息的生产者。专业化成了准则和理所当然的“圣经”。不管是学校教育还是后来的职业发展，我们都在努力让自己变得越来越专业化，以便成为一个庞大产业链中的螺丝钉。这完全符合经典科学对世界的设计，人也不过是一个精密设备上的某一个零件，只要他们能够按照大设计来完成自己被分配的工作，那么整个社会作为一个庞大的机器就可以稳步运作。当然这是一种伟大而单调的构想，它割裂了人作为复杂的、交互的、开放的系统对环境的需求，而这种需求是不确定的、不可规划的和多变的。从经典理性的割裂中能推演出对人生命的漠然，“人”作为劳动力，不过是可以被取代的生产资料中的一部分。在这个结构中，资本是最重要的生产要素，只要有大量资本就能购买土地和工厂，雇用大量工人，通过规模效益获得巨大利润。企业培养出了一大批优秀的职业经理人，他们是时代的精英，用自己专业的管理知识为企业主服务，创造了巨大的价值。这种经典理性所分析出来的社会结构会引起另外一个危险的思维结论：一旦用更有生产能力的“生产者”——机器人来做这些重复性的螺丝钉式的工作，人就

一无是处，最终成为机器的奴隶，被淘汰出局。要找到一些自动化成功取代人的例子并不难，事实上，在后工业化国家里，这样的事情正在大规模发生。当我们在说人工智能挤压人类的工作机会，大多数人会因为机器人的普遍应用而失业时，我们回看历史，就会知道，这其实是在用一种静止的、不发展的眼光来看问题。在工业革命早期就发生了农业生产效率大幅度提高，农业人口大量失业，不需要那么多人从事农业生产保证食物供应的现象。而这不是对人的机会的抢夺，是农业人口向工业人口转化，促进社会工业化生产发展的一大进步，因而创造了今天的现代文明。

随着理性认知的技术能逐渐模拟人的技能，人的感性认知的思维技能相对而言也可能变得更有价值。计算机更擅长有限假设下的数学推演，而人的洞察力始于提出一些重要的新问题。质疑机器的行动和决策，对认识到使用它们是来解放我们而非约束我们是至关重要的。人被机器人从工厂里撵出来之后，会发掘更多可以做的事情，如从事更加丰富的个性化设计。相比使用认知技术的狭义自动化任务，诸如批判性思维、通用问题解决能力、对不明确事物的容忍度以及智谋等能实现广义任务的技能和品质都会变得更有价值。产品设计、服务、娱乐，或者构建使人感到满意的环境等工作都不会在短期内被计算机取代。

对于未来人和人工智能的分工，就要从今天的教育起步。制造业发展所需的人的智能若要超越今天的人工智能技术，从根本上讲，需要创新技术工人的培养方式，让他们有终身学习的习惯，提升他们终身学习的能力，同时加强语言和数学的基础教育，增加历史和物理的方法教育，在此之上构建职业教育的具体方向。语言教育保证未来人与人之间沟通能力的基础，而数学的教育则是人和人工智能之间沟通能力的保证。加强物理学和历史的教育会让未来人从大量的经验中吸取思维方法，建立正确的历史观和世界观，以此为基础，适应高速发展的社会变革。

知识生产也是生产，人工智能技术的应用会大幅度提高知识生产的效率。根据目前信息技术发展水平来看，软件业的开源风气也会蔓延到制造业各个领域，从而推动整个制造业的发展。开源学习平台具有允许公众使用、复制和修改源代码的性能，并具有更新速度快、拓展性强等特点，因而能够大幅度降低企业开发和客户购买的成本。开源深度学习平台成为推进人工智能技术发展的重要动力。它在企业被广泛应用，帮助企业快速搭建深度学习技术开发环境，并促使自身技术的加速迭代与成熟，最终实现产品的应用落地。

人工智能开源软件框架具有统一性和标准化的特征，其生态的核心就是通过使用者和贡献者之间的良好互动和规模化效应，在终端服务、产品的使用者和开发者三类用户之间形成具有现实意义的标准体系和产业生态，进而占据人工智能领域的主导地位。

将深度学习技术应用于药物临床前研究，可以快速、准确地挖掘和筛选合适的化合物或生物，从而达到缩短新药研发周期、降低新药研发成本、提高新药研发成功率的目的。人工智能助力寻找新材料的速度，优化工业生产流程，以软件的开源运动为开始，各个行业的知识共享正在形成一种新的趋势。这不同于传统意义上的专利和知识产权，可以被视为制造业行业中一股不可逆转的潮流，从而深刻的改变工业生产的生态。

而这一现象背后的隐忧反而是知识生产的过剩。在数据驱动的生产环境下，实际上大部分制造业企业已经无法像一两百年前工业革命早期那样进行专利的保护和独享。大多数制造业企业都在参与上下游分工，因此，就不得不对上下游和相关产业的技术有所了解。而这些新技术，90%以上都是市场上已经成熟并且可以相对容易获得的技术。人工智能的应用无疑会让这种趋势变得更加明显。随着知识生产效率的提高，企业为保证技术的领先性和独占性会付出越来越高的代价，而当整个社会的企业都在从事可以大规模共享的技术研发时，就会对社会资源造成极大的浪费，从而导致知识生产的过剩。那么，怎样建

立合理的生产流程和交易机制，将是未来依靠人工智能和信息技术驱动的制造业的新问题。

人类永远有能力在看似困境中找到出路，这也正是人类智慧超越机器智慧的表现。随着人工智能在各个行业的深入应用，会有更多的人在更轻松的环境下从事更为丰富多彩的工作，创造出更多我们今天无法想象的新事物和新机会。同样的，我们如果了解了人工智能的发展历史，知道这一次的浪潮前面有长期的积累和铺垫，也就不会为它的发展而感到恐慌，人类对技术的掌握总是让人获得更大的自由，从而有更多的时间和精力让工业4.0时代的社会生活回到以人为中心的本质上来。如果说数据是人类与人工智能沟通的新语言的话，对人工智能的了解不仅是未来数据科学家和工程师需要做的准备工作，也是各行各业从业者在工作中需要的日常积累。学校需要更加重视科学、技术、工程和数学教育，即使是基础教育和职业培训，也需要增加数据教育的内容。

真实的科学研究的过程是很少有转型跨越、灵光一现的事情的。它真的是慢慢往前走，在积累到量变发生质变的时间节点，在某个方向实现些许的小突破。就人工智能而言，我们不能说泡沫都是不好的，泡沫对科普有益。新事物炒作，新概念的萌发，总会让不少人有饭吃，而每个人都要让自己的选择正义化，在历史的沉浮与优胜劣汰中抢到最终的话语权。只停留在名词炒作上，未必能够促进科学的发展，相反，可能是一种巨大的阻碍。这种“杀君马者道旁儿”——在一个领域尚未成熟就被周边环境捧杀的案例我们看的也不少了。

科学从未停止对人类思维的探索。我们相信有一天人类能够了解自己，甚至可以模拟人类的思维而创造机械的思维，即人工智能。因为人工智能是研究如何利用计算机去完成过去只有人才能完成的智能工作，我们很自然地会将人工智能和人类在同样任务上的表现进行比较。的确，在某些特定任务上，计算机已经表现出了远超人类的能力。比如，19世纪70年代，手摇计算器在算术运算方面击败了人类。

然而，在执行通用性任务时，如回答问题、感知以及医疗诊断，人工智能系统的能力变得越来越难以评估。我们以人类视觉的丰富感知为例来说明人工智能的局限性，我们知道，人类的视觉感知系统是非常复杂的，但是却能够在明确地感知前景与灵敏地捕获背景的过程中取得自然平衡。相比之下，机器的感知水平依然无法与之相提并论。

从认知的方式上来讲，人类的认知过程与我们现在谈论的人工智能是不一样的。经典逻辑不能突破哥德尔不完备定理，但是，人却具有这样的能力。人类有一种认识相对准确结论的直觉方法，这种方法与计算机式的方法不同，我们可以认知新的事物和了解新的问题，而不受哥德尔不完备定理的限制。就计算机的有限逻辑而导致的其内在不完备而言，人却从来不会受到这样的困扰，因为人天生具有突破有限逻辑的能力，也许这构成了我们通常意义上说的感性。这也许是我认为这一代人工智能无法超越人类思维的数学逻辑层面的本质原因。

但哥德尔所限定的有限逻辑，可能不限制量子力学的基本逻辑，人类的直觉也可能不受哥德尔不完备定理的限制，从这个角度来讲现在的计算机结构不太可能具有人脑的能力。当然，量子计算机基于量子逻辑，离实现还有些实际的困难，现阶段我们不能够简单预期。

量子信息的解释也许会渗入人类对认知的了解。如果大脑真的是量子化的工作，我会认为对人类而言这是一个好消息，我们用经典的图灵机的方法来开发的计算机会在很长时间内无法超越人脑。因此，我们也就不用担心人工智能控制人类。因此，量子计算有可能是未来更接近人类思维方式的人工智能的实现通路。但今天我们对量子力学的认识还非常有限，所以，在这个基础上构建一个具有计算功能的器件，还有一段很长的路要走。从这个角度来讲，机器人也很难会有类似于人脑的思维能力，也就不具有独立的学习和创新的能力。

由于人工智能是关于怎样表示知识、怎样获得知识并使用知识的科学领域，所以，对人工智能的了解，有助于人们认识人类自己。人工智能需要更多地反映我们自身智能的深度。如果我们脱离人类知识

的进步而单纯讨论人工智能，我们只会得出一幅冷漠的图景——人工智能按照刻板的假想演进，而我们人类被抛弃。

基于现在对物理世界尤其是量子力学、脑科学，甚至是逻辑学的认识，我们距离开发一套有如人脑一样复杂的工作系统还有相当长的路要走。人工智能恐怕在未来三百年内还不会超越人脑。这个三百年的估计，不是凭空想象，而是源于我们对物理学进展的了解。我们知道，从牛顿经典力学到量子力学的诞生经过两百年，而量子力学发展到现在仅用了一百多年。当我们回顾科学发展的艰难历程，突然发现，人类所了解与所懂的之于自然，不过寥寥，甚至在缴械认输中，发现了自己的渺小。那么，问题似乎又回到了起点——也许，我们的基础研究手段可能存在问题。看过去的科学发展史，我们推算，大概还要这么长的时间才有可能在这个基础上了解和使用这些技术。三百年不是个太夸张的数字，三百年内，我们大可放心地去跟机器相处。

第十二章

不是结尾的结尾

北大有个湖，名字叫未名湖，意思是没起名字的湖。据说是燕京大学时期冰心起的，我对这个说法颇有些怀疑，很多成名的人喜欢把事情揽到自己身上。至此，读者至少应该感觉到了这个名字的智慧。正如有人要问我到底哪些事情是人工智能做不了的，我无法用语言说出答案。要记得，如果我可以把这样那样的事情一条一条列出来成为某种清晰的描述性陈述，图灵机就一定能够把这些描述性的陈述当作算法来实现，那么人工智能也因此可以去做同样的事情。因此看来，答案有，我们不能说。

如果鱼缸里的鱼能像我们一样有复杂一点的思维，它会不会想：鱼缸外的世界是什么样子呢？一切事物都因为球形的透镜而古怪地变形，那外面的人和物都是虚幻的吗？为什么我无法触及他们而他们又在那里走动呢？而隔开这个虚幻的和我可以触及的真实世界的东西，似乎是一堵看不见的墙，我能感受到它的存在，似乎它是那个变异了的世界和我身边这个看起来正常一点的世界之间的不可逾越的屏障。它是什么？我没法描述。

如果非要有牛顿定律的话，等等，可以先有伽利略。伽利略在讨论落体实验的时候，其实有一个当时尚未清楚的假设：空气阻力跟重力比较起来非常小，在物体速度不快的情况下，可以忽略。而正是由于空气阻力的存在，亚里士多德才猜想重的物体下落速度快，而轻的物体下落速度慢。物体下落，是重力和空气阻力共同作用的结果。

但，对于鱼而言，这个假设是不成立的。水的阻力是如此之大，在不了解重力这个简单假设之前，一个鱼类的物理学家完全有可能形成一套更为复杂的运动学和力学的描述机制。当然会有聪明的人类跳出鱼世界的假设看到，其实这个鱼类做出来的动力学理论，可以有更简单的描述，即重力和阻力的组合。

这不是用拟人化的手法来描述高深的科学知识。其实，这件事情早已发生过。在我们怎么也搞不清楚以太是什么东西和测不出来光速变化时，洛伦兹费了很大的力，推出了洛伦兹变换来解释为什么光速是不变的。关于洛伦兹时空关系的公式的可靠性，连它的发明人洛伦兹自己都不太确信。而当爱因斯坦直觉上觉得速度是可以以光速为上限的性质时，事情就简单了很多。

回到鱼的问题上来，我一贯对饲养宠物抱有一定的敌意。我觉得自然的动物就应该让它们生活在自然里，我们为了弥补人类内心的缺失，养了诸多动物。但这些动物的生命凭空造了出来，挤压了其他动物的生存空间。牛津大学自然史博物馆的镇馆之宝渡渡鸟，就是这样的例子。17世纪欧洲殖民者把猫、狗、猪等动物带到了毛里求斯的岛屿上，在这之前，岛上的渡渡鸟没有天敌。英国探险家把渡渡鸟带回英国展览的时候，岛上渡渡鸟还有很多。几十年之后，猫、狗、猪等动物在岛上大量繁殖，偷鸟蛋，捕食幼鸟，渡渡鸟便成了在人类干涉下第一个灭绝的物种。牛津大学自然史博物馆保存了渡渡鸟最后一个完整的填充标本。从这个意义上讲，猫和狗作为宠物缓解了人类的孤独，从而获得了该群体在地球上的竞争优势。那么，谁是谁的宠物呢？

问题并不存在一定正确的这样或那样，而事实上哥德尔不完备定理说明，我们没法描述一定正确的真理本身，这样就需要我们界定讨论的前提条件，所有论述都只能在特定的前提下才有意义。所以人类的辩论活动变成一种实用性的技巧，它更应该被看作成一种怡情，而不能被看作是对真理的维护。有些问题，我们明知道它们是对的，而且我们可以一一验证，但我们可能在这个框架内都无法证明它在理性上是对的。

科学的事情，一个人常常只能做一小点。我的博士论文做的是量子气体在重力下的混沌动力学现象，这些东西常常让非此专业的人觉得每一个字都看得懂，但放到一起就不明白是什么意思了。其实不仅不做物理的人如此，做物理的不做这一方向的人也有这种感觉。但我们有一种超然的概括能力，把这些不懂的东西概括成一个黑盒子般的词汇，只要说得出来，就意味着似乎懂了。如果说真理给我们描述了一个可能普适的样子，哥德尔不完备定理告诉我们真理的正确是有明确的前提的，把这些前提也当作正确的话，就需要更多的前提来说明这些前提在什么情况下是正确的。我们已经意识到，寻求真理这件事，恐怕没完。我们退而求其次，去寻求相对靠谱的方法。牛顿之后

的三百多年下来，我们成就了如今观察世界的一系列方法，此方法被称为科学。科学的出现和存在，并不是要替换掉在地球上存在过的亿万人的思想，它只是谦卑地说，依照现有的证据，科学怎么看。而它的谦卑往往是有效的，即使我们很快意识到这有效的局限性，这并不影响我们用它来了解和认识我们周遭的世界和我们自身。毕竟，科学是我们已知的方法中，最可靠和最开放的。科学是基本的方法，但它从未代表过绝对正确的真理。因为目前看来，只有通过这个渠道我们才能认识到比较靠谱的世界，获得比较靠谱的知识。我们所惯有的科学教育让人太早丧失了对科学本身的兴趣，一旦不用把它当作升学的敲门砖，很多人就迫不及待地将它丢回给老师。成年之后这些人会跟搞科学的人说，物理、化学和数学那些东西太难了，他们学不懂。这么说，并不算是一种谦虚而是体现了另外一种傲慢。

虽然有时我们也发现“科学方法”多少会有些问题，比如哥德尔不完备定理所指出来的。好在“科学”并不限制自己的研究范畴，也不宣称自己就是绝对真理，它本身在挑战着绝对真理的存在。我跟搞艺术的朋友说，建起来的时候就想着推翻，而事实上就是推翻了，推翻了再重建，心里惦记着还是要推翻它，科学家所做的才是行为艺术呢！这又像极了佛教密宗的一种修行方法。喇嘛用彩色细沙在地上做沙画的唐卡，往往会花几个星期甚至更长的时间，在即将画好的最后一刻把整幅画毁掉，重新再来。这沙画的唐卡被称为坛城，用来启示世事无常。

人与自然的关系，人与人的关系，人与自己的内心世界的关系构成了人对宗教信仰和科学探索的核心。我是谁？我从哪里来？我来做什么？总有人喜欢因此谈论哲学，觉得这样显得比较有学问。人类的古代文明往往通过思辨成为某种成系统的知识体系。这些东西从各自的角度说起来往往都有些道理，但谁也说服不了谁。以至于老子出来做裁判，“道可道，非常道”，也就是说你们那些能够说得清的东西，都不是永恒的东西，都是末流的知识。然而老子所向往的非末流的东西却导致了我们始终没法积累相对可靠的知识，人类对世界的探

索徘徊在离起点不远的地方，没办法积累成可以不断拓展的知识。由于对信息掌握的不完善，人们常常认为会有一个超越普通人类的绝对真理存在，它可以是一套真理，也可以具象成某个形象，譬如基督或者佛陀，甚至可以不说它到底是什么形象，但它应该是存在的。

直到文艺复兴时，以牛顿为代表的前后几个世纪的学者建立了一套成体系的对思想进行检验的技术，“实践是检验真理的唯一标准”，对人类已经认识到的知识的整理和未知知识的推演，与依赖这一体系所提供的检验手段所得到的结果的吻合，为这一体系的可靠和有用做了完整的背书。继而，人类的思考可以系统地审视和积累，又可以在这些相对可靠的基础上继续延伸，有史可循的人类的一万四千年的积累在这最后的三百年开始了爆发式的增长。文明不行的时候我们都拿文化凑，各国的古代文化，都输给了这一现代方法。这个时代留给人类知识体系的影响是，我们相信人类生活的宇宙是一个精确的可计算的客观系统，人类可以通过掌握工具来预言它，并获得可观的知识。基于这些知识我们可以进一步认知更多的知识，至于这套知识是局外的超自然力所安排的还是本来如此，我们可以把它称作Nature（自然）或者上帝。无论如何，这套检验思想结果的方法，我们认为它非常可靠和有用。也借此建立了很多不同的学科门类，并且都获得了可靠的进展。哲学在这个时代达到了它历史的最高峰。

古希腊时期的自然派哲学家被认为是西方最早的哲学家。不管他们认识世界的方式是否正确，他们的想法有别于神秘论的原因在于：这些哲学家以理性辅佐证据的方式归纳出自然界的道理。苏格拉底、柏拉图与亚里士多德奠定了哲学的讨论范畴，他们提出了有关形而上学，知识论与伦理学的问题。形而上学，归纳为元知识，即用来描述知识的知识，或者说用来了解具体知识的方法。科学本身是这种元知识，它提供了有效的手段，让我们在现象和引起这些现象的原因之间找到确凿的证据。

当然它依旧可以演绎为对是否存在独立于人的世界的认知，而在这本书的讨论范畴里我们认为，世界本来就不能独立于人的认知而存在，对于我们不能够认知的，我们不去讨论它，子不语怪力乱神。而知识论，在于我们怎样获得可以比较信赖的新知识，这一点，我们无法独立地去认知。我们必须有相对可靠的工具，人们就会利用这些工具来衡量我们所面对的世界，从中得到相对稳定的、能够解释旧现象而带来新认知的理论体系。这样的知识必须立足于实践，它是我们头脑中产生的方法论、形而上学、逻辑或者数学在真实世界中的反馈，需要与真实世界相印证，才能获得。

经过这样的拆解之后，这本书里我们只讨论认识知识的工具，由逻辑到数学，再到自然知识本身的案例：物理学。而恰恰哲学本身并没有跟上这两个领域所代表的科学界20世纪以来的发展。哲学最近一百年对科学最大的贡献是没有贡献，用一句话来概括，哲学家之于科学家，如同鸟类学家之于鸟类。在牛顿的时代自然哲学被物理学完全继承下来。它之上延伸出来其他学科，发展了各式各样的分科治学，物理的、化学的、生物的、工程的、社会学的、心理学的等诸多科学门类。这样划分的好处是把我们日常见到的自然的、社会的和我们自己的问题抽象化了。

当我们划分知识领域的时候，我们也给自己砌了一道墙，那些圈来的领地也成了自己的牢笼。我们所说的分科治学就问题本身而言并不是自然的存在，自然界的问题一定是多学科贯穿的。我们为了简化模型，至少让我们手头掌握的可以适用的工具有着手之处，就要先把复杂的事情简单化，抛去次要因素，但往往这些次要因素并不是没有什么重要影响的。但如果都要考虑进来，我们手头的工具可能远远不够。不管怎么说这是个起点，正如务实主义所主张的，我们无法做好所有的规划和掌握所有的工具才能动手研究，我们只能凑合。这种凑合给了我们继续下去的阶梯，这种积累只要是定向的、稳妥的，加上时间的作用而累积下来，速度并不算慢。

虽然我们现在主张跨界思维、协同创新，但其实是我们自己把本来一体的知识分割开来，从而限制了我们本来应该具有的想象力。

这样说绝不是要剥夺我们已经有的工具，而是提醒我们要首先学会掌握并且尊重这个工具，只有这样，我们才能学习和研究当下相对可靠的知识，从而拓展我们的认知领域。我们也要时时提醒自己，这些工具也会是我们认识问题的枷锁，它会限制我们的想象力，好在有哥德尔告诉我们已知系统的不完备，永远要跳出现有系统才能找到新的线索来弥补这些不完备，而这种事情是做不完的。

之所以费了很长的篇幅去了解和探讨哥德尔，是因为他对于我们今天的知识体系的构架而言有重要的意义。我们人类的认知到今天大概经历了三个阶段，当然这不是终点，对真理的追寻永远没有终点。第一阶段是人类长期缺乏确定性的知识评判手段，每一种猜想都是等效的，但每一个猜想都会把我们带到完全不同的方向，这个阶段被称作古典神秘论阶段。第二个阶段从伽利略和牛顿开始，人们建立了经典的知识体系，这个知识体系的建立是以建立对终极真理的信仰而存在的。事实上这也是一种神秘论，不过是科学化的神秘论。当科学把自己当作真理的一部分的时候，科学也被神秘论化了，从而成为我们所说的“科学教”。虽然还是有科学真理的信仰，但同期成长起来的科学方法，让人们逐渐认识到科学不过是一套稳健、靠谱地找到相对真理的方法，至于它要得出的结论，未必是那样的真理。1900年后，量子力学、哥德尔不完备定理以及没有什么口号和纲领的杜威式的务实主义代表了一个新的阶段的开始。我们还用“科学”这个词来统称它，它与古典科学虽然事实上有本质的差别，但也一脉相承。从体系上来讲，它不再承认普适真理的存在。任何结论，都是在一定的条件下才成立的，而这条件也绝不是唯一的、不变的，即我们不再有放之四海而皆准的绝对真理，自然也不再有人可以代表这种真理。一脉相承是因为即使这样的结论也是我们依靠相对务实、靠谱的方法得到的，就目前而言，我们认为它更关注方法，而不关注结论。利用实证，我们说这样的方法所得出的结论至少没有矛盾，因为这个方法本

身是排斥矛盾的，也能经得起检验，因为它是开放给各种人和各种方法来做检验的。而没有人站出来扮演权威说这样的结论有正义的豁免权。这是多么开放和平等的研究方法，它天然地跟自由在一起。

即使我们最终理解了理性的不完备，这也不是终极的知识阶段。我们知道今天人类所拥有的工具仍然不能解释我们感受到的世界。我们对人类的意识，尤其是自由意识完全不了解，我们甚至没有十分的信心来判断自由意识是不是真实存在。我们始终具有跳出现有框架而获得新知识的能力，而这种能力从哪里来？我们怎样去描述这种能力？我们不能说自己没有掌握这种能力，它依然是我们身体的一部分。

我这样长篇累牍地介绍哥德尔的思想，是希望读者了解，这世上并没有绝对真理。事实上我所说的绝对真理，它自己无法用语言来描述也无法证明自己真理性。这并不代表不存在有限的理性，而我们恐怕已经找到了靠谱的方法，并由此不断地深入我们的认知，所谓科学精神和科学方法不过如此。科学不代表理性，也不代表真理和正义。它代表了一种知耻近乎勇的理想精神，但这一点又是极难做到的。人会很自然地把自己的行为、选择和认知正义化，按照自己的经历来诠释其合理性，所以面对一个相对客观和无情的选择时会比较无奈、沮丧甚至怒目而视。但它就这样，也正是哥德尔不完备定理的存在，让我们认识到务实主义策略的重要性，它大概就是源于这样的原因。我们不再追求放之四海而皆准的普遍真理，而是追求一时一地能够适用的解决方案。这不是认知的退化，反而是认知的进步，我们不再把古圣先贤当作真理的拥有者，而把他们的认识当作我们今天认识世界的基石。

20世纪在中国还在启蒙化运动中的时候，科学界和思想界发生了几个重大的变化。首先是数学界和物理学界，物理学界曾经以为的经典物理已经完成的事情被量子力学重新构建了，而数学界认为可以做终极数学的事情被哥德尔否定了，这两件事情可能都告诉我们人类理

性认识的能力和边界。事实上这否定了绝对真理的存在，而我们人类又还没有找到感性的物理或者数学的工具来代替理性工具，这是我们目前的结论，似乎也只能这样了，这是一个极其务实主义的解决方案，不去追求极致的、绝对的真理，而是认识一点算一点，一寸有一寸的欢喜，有一分证据说一分话。我们寻求的终极解决方案或者真理，不应被看成永恒完美的，而应被看作我们继续去认识世界的工具或阶梯，我们在利用这些工具去开拓未知时也要时刻提醒自己，这些工具也会成为我们规划自己行为和认识能力的枷锁，工具还是枷锁，是个叠加态。

哥德尔不完备定理事实上解释了我们认识中包括直觉的那部分，这是真实的感受，至少告诉我们为什么我们的感性认知一直在起作用，而且不能从认知中被排除出去，甚至因此它的存在才更重要。一旦不再为理性和感性而感到困扰，我们便可以用平和的心态来看待科学的发展和进步了。一旦了解到它的秉性，我们就不再畏惧科学的日新月异，而能够以快乐的、探索的、求知的、不受束缚的自由之心去用科学的眼光来了解世界，了解人类，了解我们自己。从哥德尔不完备定理得出的心智的非机械本性应当将我们的思想引向如量子力学所暗示的那种非机械的物理定律。经论证不能被机械把握的数学直觉思维仍然是一个物理系统，因此我们应当寄望于发展出一种非机械的、激进的新科学——在这里量子力学之谜也许将引导我们——以此来容纳心智的非逻辑之处。

哥德尔似乎在暗示，从他的定理中得出的是，只要我们对真理的领悟不是幻想，只要我们真的拥有我们认为自己具有的直觉，我们就不是机器。如果我们确实真的具有直觉，就不可能形式化我们所有的关于真理的直觉。当然，不可能证明我们知道我们认为我们知道的一切，因为我们不可能形式化我们知道的一切。然而，这正是哥德尔不完备定理所要求的。这暗示着我们也不能严格证明我们不是机器。哥德尔不完备定理通过表明形式化的局限，既暗示了我们的心智超越了机器，也使得证明我们的心智超越机器成为不可能。

安东尼·伯吉斯（Anthony Burgess）在评论长篇小说《1984》的时候说，多一个人读《1984》，就多一个自由的人。我希望我的读者在读这本书和它的姊妹篇《量子大唠嗑》时会意识到这个世界上本来就没有绝对真理，因此也没有掌握了绝对真理的权威。在使用科学方法的前提下，即，程序正义，每一个人都有独立追求科学结论的权利，都有对未知进行探索并能够得到有效结论的权利。从这个意义上讲，大家不用拘泥于古人中的先贤、今人中的伟人、众人中的权威，而是自己去观察和思考，推理和验证，不必为了某一特定结论的证明而找证据甚至修改证据。

为什么这么说，其实这本书不完全在说哥德尔。在写这本书的时候，我确实也想明白了很多问题，或者说至少对很多问题想得更深了。因为我虽然曾经很认真地在物理学领域里深耕，但绝不排斥对别的学问的兴趣。以铜为镜，可以正衣冠；以古为镜，可以知兴替；以人为镜，可以知得失。与人交往可以看到的是更好的自己。写书也一样，在形成文字的过程中，看到的是自己的思想。文字的行走和自然世界的规律是类似的，事物都在彼此中成就，没有永恒，只有生长。写书的另外一个冲动，是我觉得该给比自己年轻的孩子们写点东西。因为我自己曾经经历了辛苦，走过了很多地方，身心的也好，智力的也好，我觉得这些经验值得早一点跟年轻人分享，便把自己即使不太成熟的思想也说出来了。

哥德尔不完备定理说明，我们没法建立一个绝对的或静止的真理系统。因此要维护某种绝对正确的事情，而不容他人的匡正，从逻辑上是行不通的。我们不得不选择一项务实主义的策略，在摸索中前行，随时准备调整。这种调整，一是随着时代的延续，对手也不再是当年的对手，对手们都在变化；二是执政者的成熟与开放，或者说是没落；还有就是在野的胡适们和李敖们的成熟和“人书俱老”，老而知道翻天覆地变革的不可靠。但这不等于人不会有办法，英雄忒修斯的船上每一根木头都被换成了新的，忒修斯之船也还是它自己。务实主义者最不喜欢革命，最不喜欢造新船。打翻了重来，未必新的就比

旧的好，但造新船劳民伤财和生灵涂炭是一定的。这，被历史一次又一次地证明过了。

后记

从杜伦（Durham）起程回牛津的火车上我写了最后这一部分。火车从英国乡间的绿地和农田中穿过，偶尔还能看到一栋一栋的小房子。能在牛津写完这本书的最后一部分对我有特殊的意义，这就像一个重大工程最后的仪式感。这本书我一拖再拖，每一次要收尾的时候就发现还有很多很多要写的东西，所以始终不能完稿，但哥德尔说了这件事也许本身就不能完成，他开拓了一门学问。我还是不敢让我的博士生轻易地开展这方面的研究，按照现有的科研体系他们如果跟我来做这“奇门遁甲”，毕业就成了实质的困难。很久以前，读书可以很久，甚至不毕业，但现在这学位生产工业化了，知识生产也工业化了，何必与人为难又与己为难呢？做一些实用的东西总是好的。这不是贬低实用的用处，实用主义或者务实主义虽然与浪漫一点也不沾边，但对做事情总是有好处的。我们既然找不到放之四海而皆准的工具，就有什么用什么，用极其务实的手段完成一段算一段。人生没有白走的路，每一步都算数，人生如此，科学的探索也如此。每当我们的工具进步了，我们总希望回头去审视以前的认知有哪些可以深入和再诠释的地方。我们生活在一个不完美的世界里，这也许是这个世界完美的体现。

在物理学界，我们曾经追求这样那样的大统一，但也许不统一或者有点缺憾就是世界本来的样子，谁说它一定要按照我们想象的样子来构建呢？这本书自然不是去说爱因斯坦的局限，或者量子力学的有待理解，而是对那一代人的致敬，是他们让我们有勇气继续前行。这些人类历史中的精英分子给了我们面对未来的信心。

我好想说几句宗教式的话来显得这本书多少有些神秘感。但它够难了，理解它会耗费很多精力，书中提到的每一个方向都可以作为一篇博士论文来做，可能一篇还远远不够。很多时候假想容易，但确凿有据的证明是困难的。

这本书写到结尾，我看了电影《波希米亚狂想曲》。作为一个对音乐一窍不通更不太了解皇后乐队的人，我竟然莫名地热泪盈眶，这之后很长时间里，每次坐飞机我都要重温一遍这部电影。对机器了解越多，我对“人”就越发敬畏。在人类波澜壮阔的历史中，我们今天的努力还是微不足道的。尤其是相对于未来人而言，我们今天所认识和掌握的东西，很多会成为他们不经意地默认的知识。而他们所认识和掌握的工具，是我们今天难以想象的。不管怎样，我看到人们陶醉在“人”本身的创作中，在台北夜市的烟火气中，在优胜美地密林深处篝火旁的乡间乐队的歌声中，便会为人类而感动。在我最喜欢的一部电影《阿甘正传》中，当一位军官问阿甘：“孩子，你想过你未来会成为谁吗？”阿甘回答说：“难道，未来的我不是我自己吗？”

图书在版编目（CIP）数据

人工智能之不能 / 马兆远著. -- 北京 : 中信出版社，2020.3
ISBN 978-7-5217-1379-4
Ⅰ. ①人… Ⅱ. ①马… Ⅲ. ①人工智能－普及读物 Ⅳ. ①TP18-49
中国版本图书馆CIP数据核字(2020)第021569号

人工智能之不能
著者：马兆远
出版发行：中信出版集团股份有限公司
（北京市朝阳区惠新东街甲4号富盛大厦2座 邮编100029）

字数：216千字
版次：2020年3月第1版
广告经营许可证：京朝工商广字第8087号
书号：ISBN 978-7-5217-1379-4

版权所有·侵权必究


设计


现实
DESIGNING REALITY
第三次数字革命的下一个经济引擎
How to Survive and Thrive in the Third Digital Revolution
[美]尼尔·格申菲尔德（Neil Gershenfeld）[美]艾伦·格申菲尔德（Alan Gershenfeld）
[美]乔尔·卡特-格申菲尔德（Joel Cutcher-Gershenfeld）_著 丁峻峰 凯川 蓝河 _译

中信出版集团

设计现实

[美]尼尔·格申斐尔德 [美]艾伦·格申斐尔德 [美]乔尔·卡彻-格申斐尔德 著

丁峻峰 武川 蓝河 译

中信出版集团

目录

推荐序

序言

第1章 智造万物

制造：可编程的原子世界

教育：从DIY到DIT

应用：漫长供应链的定制化可能

启示：创制物品的红利

组织：仁慈的“独裁者”

第2章 第三次数字革命

智造获取

数字化公民

赋能的生态系统

防范风险

大挑战

第3章 背后的科学

从摩尔定律到拉斯定律

通信、计算与制造

从乐高积木到核糖体分子模型

维度诅咒和智能设计

第4章 路径创造

反应式批判
被动与主动
变化速率
生态系统
自然增长与规模增长

第5章 原子与比特的进化路线图

社区制作：工具
数制：一个工坊制造另一个工坊
转化：虚拟到现实
无所不在：从1到1 000 000 000 000

第6章 未来的现实

人与技术交汇的乌托邦
预言性转化模型
装配工和煎饼早餐

结论

后记

致谢

参考文献

推荐序

《设计现实》已经是我负责为尼尔·格申斐尔德教授翻译的第二部书了。如果说第一部书——《智造：一场新的数字革命》的出版是在其英文原著出版10年后对当年尼尔教授对于数字时代的个人智能制造预言的回顾审视的话，那么这本书——《设计现实》——则是站在10年前《智造》里预测中的当下，面对下一个10年，或者更久远的时期内，人类数字化生存的全面展望。

2018年夏天，当第14届全球智造学术年会（FAB 14）在法国南部的图卢兹召开的时候，尼尔教授再一次介绍和验证了他的摩尔定律，全世界建成的“数制”工坊（fab lab)的数量，在每一年半的时间里，会翻一番：全球数制工坊总量已经超过2 500个了。

“数制”工坊在全球范围内迅速蔓延开来，星罗棋布的个人智造终端空间，让越来越多的创客可以将们大脑里的创意智造出来，全球实验室通过线上直播平台和网络学习与知识同步系统链接起来，推进个人制造（DIY，Do It Yourself)的共创和分享（DIT，Do It Together)。这大大赋予了人们实现设计的能力，孕育了人工智能时代新的生产方式。

工业革命时期，生产流通从挖掘原材料开始，通过全球物流将原材料运抵工厂，并加工制造成产品，再把产品运输到全世界消费终端市场，这是单线经济（linear economy)模式；随着信息时代的到来，电子设计文档（比特）可以通过互联网在全球流通，实现数制工具产品本地化（原子）。智造，这种把物流成本，原材料成本和能源消耗降到最低的本地制造模式，就是新的循环经济（circular economy）

的基础，它成为网络时代生态、经济、社会可持续发展的动力，不断改变着世界的发展模式和经济格局。

随着智能时代的到来，各国也紧抓机遇，努力在未来全球战略中奠基新的格局。

中国正在努力成为世界科技强国。“科技兴则民族兴，科技强则国家强”。我们不断提升对“科技强国、科教兴国”战略国策的认识，强调自主创新和核心竞争力提升。《中国制造2025》的提出，展示了中国制造业转型升级的决心，以实现从“中国制造”(Made in China)到“中国设计”(Design in China)的巨大转变。科技要强大，人才培养是第一步，造就一批有自主创新能力的科技人才和有独立知识产权的技术关系到中国未来的格局。近两年，中国在K-12教育中全面铺开综合素养、STEM/STEAM及科创教育；在大学层面推进新工科建设等全面创新人才培育机制及系统，以应对人工智能时代人才培养的新需求。

2013年，在我的主持和团队的努力下，同济大学设计创意学院成立了中国大陆第一家“数制”工坊实验室。2014年3月1日，第一场开放夜活动正式登陆上海，标志着中国“数制”工坊（fab lab 0）开始了一段使命，逐步将全球数制工坊网络和理念在中国落地以生根发芽。

未来，我们将继续拓展全球网络课程——智造学术（FAB Academy)在国内的推广，同时将核心课程本土化，通过实践，演化成植入各类学校的课程系统：“智造学术X”（FAB 0 Academy X）针对的是大学、职业院校、高中；“数制乐园”（FAB 0 Playground）是专门针对K－12的贯穿式的课程体系；同时拓展一系列“项目导向式学习”(Project Based Learning)的主题性实验室模块，包括“可持续生态”“生物合成”“人工智能”“智能可穿戴”“机器人”“增强现实”“3D打印”“未来城市”“未来生活”“飞行器”等。

中国“数制”工坊及课程已经先后进驻上海科技大学、湖南机电职业学院、常州机电职业学院、上海格致中学、上海市同济黄浦设计创意高中、苏州领科海外学校、上海市宝山行知中学、同济附属实验学校（嘉定）等十多个知名高校、职业院校及K-12学校。同时，fab lab实验室也为上海包玉刚民办实验学校、YKPAO、上海阿德科特国际学校等多所知名学校开发了创新课程。

2018年10月，尼尔教授受邀来中国南京参加“2018世界智能制造大会”（2018 World Intelligent Manufacturing Summit）并发表主题演讲“智能制造的未来”（The Future of Fabrication）。会议期间，尼尔会同中国“数制”工坊主办了名为“中国智造集群”的主题沙龙，并将“中国智造学术”课程定义为在中国本土开展“数制”工坊的内容蓝本。

让我们一起努力憧憬：当未来的某一天，在世界上的任何一个角落，分布着像便利店一样方便可及的“数制”工坊，任何人都可以发送比特文档，并通过3D打印机等机器将个性化的产品智造出来的时候；当3D打印机等智能机器通过不断打印新的零件来组装更多打印机，也就是机器可以不断复制机器的时候；或者，当我们不仅能编程电路板，还可以编程生物细胞，完成从“智造万物”到“生长万物”的演化的时候，我们又将如何设计现实？设计出的现实又是怎样的呢？

最后，感恩为这本书的出版而付出的所有支持者。

丁峻峰
同济大学创意学院副教授
中国“数制”工坊（fab lab 0）实验室主任

序言

想象一下，现在是1965年。汽油还是31美分一加仑[1]。披头士刚刚发行了《救命》（*Help*）专辑。瓦茨骚乱仍在洛杉矶肆虐。《音乐之声》（*The Sound of Music*）还排在院线榜首。美国数字设备公司（DEC）首发了第一台使用集成电路技术的电脑PDP-8，售价18 000美元。

你正坐在圣何塞一家拥挤的咖啡厅里喝着咖啡（那时超大杯半糖无泡焦糖玛奇朵尚未问世）。你这桌是咖啡厅里唯一有空座的地方。一群人鱼贯而入，大声喧哗着。他们人手一份最新一期《电子》（*Electronics*）杂志，兴奋溢于言表。其中一位问你是否可以坐在你旁边。“当然可以。”你回答道。

你听出来这群人原来是附近一家半导体公司的研究员，他们在热烈讨论杂志中的一篇文章。不过，他们谈论的东西仿佛是天方夜谭：什么有朝一日电脑会小到装进衣兜，或是做成腕表；什么这些“个人”电脑会如同大型计算机一样性能强大；什么所有电脑联网指日可待，届时任何人都能在任何地方获取、掌控和分享信息。

他们的讨论越是深入，所描述的未来设想就越是离奇。毕竟，电脑是体型充栋的庞大机器，只有顶尖研究机构和商业巨头才可能配备，且价格昂贵，需专门技师操控。电脑能装进衣兜或是联网的想法，就如同亚瑟·查尔斯·克拉克的科幻小说或者《至尊神探》（*Dick Tracy*）、《摩登家庭》（*The Jetsons*）一样不可思议，绝非近期可能实现的，实在不值一提。你忍不住向同桌的那位表达这一疑虑。

那些研究员停了下来，纷纷点头，仿佛对此疑虑早已了然于胸了。令他们如此兴奋的不仅仅是电脑技术发展带来的性能提升，还有其发展的指数级速度。他们解释说，这种变化在初期很难察觉，不过当每个人都回过味儿时，变化就已经是天翻地覆了。他们认为，如果你能抓住要领的话，其实电脑性能加速发展的迹象已经显而易见。他们坚持认为自己抓住了要领，自信满满。你感到既半信半疑，又兴趣盎然，就顺势问道："好吧，那你们觉得要领到底在哪儿？"

研究员们受到了鼓舞，继续解释。首先，我们看看数字技术的本质是什么——具体而言，是什么使数字技术区别于其他多数技术，以指数级的速度发展？他们将数字技术与另一项新近发明成果——施乐复印机进行比较。施乐复印机与电脑的不同之处是它使用的是模拟程序。比如，要是你想用施乐复印机复印文件，得把第一页塞进机器印出第二页，然后你再把这两页塞回机器印出4页，然后印出8页、16页，以此类推。当你印到几千页时，大多数复本都已模糊不清，遑论上亿页了，这是复印过程中误差不断积累的结果——所承载的信息丢失了。

"而数字技术就不一样。"他们解释道。数字信息被转换成信号（由1和0组成）。只要在系统中加入一个运行的纠错程序，数字设备就可以反复复制信息，而信息则毫发无损。研究员们认为，这就是数字化的本质，使数以亿计的数字信息清晰无误地被掌控和分享。而且，纠错程序十分廉价，信息复制几乎不产生任何边际成本。他们的结论是，这一基础科学会推动全社会的变革。

其中一位研究员打开那本《电子》杂志，向你展示令他们激动不已的那篇文章。文章作者名叫戈登·摩尔，是仙童半导体公司的研发部主任，那些研究员也正是在这家公司就职。标题是"给集成电路塞上更多部件"，绝妙直白，令你忍俊不禁。文章论证了单个集成电路芯片上可容纳的元器件数量每年都要翻一番，他预计这一趋势至少会持续10年。研究员们之所以兴奋，是因为这一趋势对于将电脑变得更

小巧、更便宜、更快捷至关重要——而这些变化又是加速进行的。如果摩尔预测准确的话，即电脑性能确实可以每年翻番并持续10年，那么到1975年，电脑运行速度会比现在快1 000倍。考虑到这一趋势，摩尔预测“家用电脑（或者至少是连接中央处理器的家用终端）汽车自动控制系统以及个人便携式通信设备等奇迹”将会出现。他在文章中写道：“今天我们离电子腕表只差一个显示器的距离。”你联想到《至尊神探》和《摩登家庭》是不无道理的，那正是摩尔预测会发生的事情。

研究员继续解释，电脑性能的日益强大使其他数字技术成为可能，而且自己也愈加强大。比如，其中一位研究员正在从事美国国防部高级研究计划署的一项签约项目，旨在展示电脑如何进行长距离数据网络传输。另一位研究员则指出，他们已经开发出一款电脑，能够在象棋等复杂游戏中击败人类。这些进展，与很多其他同类技术进步相互交织、相互支撑，为经济、文化与社会各个领域的巨大变革奠定了基础。

随着研究员的进一步解说，你逐渐相信这的确是件值得关注的大事。具体时间虽然并不明朗，但你意识到你可能已被赋予了一次实在难得的机遇，以见证历史演变为翻天覆地的未来。而现在你要做的，是如何把握这一机遇。

如果你是企业家，你可能要开始思考崭新的、颠覆性的潜在商业机会；如果你是行业组织的领导，你会试图利用这些技术获取竞争优势，淘汰部分边缘业务，谋求转型升级；如果你是公共机构的负责人，你可能需要开始评估这些强大技术如何惠及所有人，而非那些幸运的少数人——如何利用这些技术提高公共福利，减少潜在危害。无论你是谁，都要考察这些新技术会如何影响你的个人生活与职业生涯。至少有一点是明确的：一次数字革命已迫在眉睫。

失去的机遇

事实上，在过去半个多世纪里，已经有两次数字革命接踵而至，而且比摩尔所预计的更加波澜壮阔。第一次数字革命是通信革命，把我们从模拟电话时代带到了互联网时代。第二次数字革命是计算革命，为我们带来了个人电脑和智能手机。两次革命的联袂彻底改变了整个世界。

早在1965年，数字革命就已初露端倪。不过，世界上大多数人却与它擦肩而过。因此，前两次数字革命所带来的经济、文化和社会冲击使人们措手不及。摩尔的预言，就是著名的摩尔定律，并非如他所预期的那样只持续了10年，而是整整50年。如今的电脑，其性能比1965年时强大了近10亿倍，而且的确可以放进衣兜，做成腕表。

技术持续以指数级速度发展着，而个人、组织和机构只能亦步亦趋，拼命追赶技术的步伐。这种追赶态势遍布社会——包括那些生活被数字技术所影响的个人、运营模式被技术加速发展所不断颠覆的组织，以及在风云变幻中谋求稳定的政府、教育和法律等领域的公共机构。

这些技术渗透社会的程度越深，社会追赶它的步伐就越快。要想规范这些加速发展的变革性技术，最佳时机是技术形成的早期，即社会变革尚未蔚然成风、根深蒂固之时。这时，技术的想象空间和最初的市场成果转化仍在成形早期，还有斡旋余地。数字化通信与计算的革命推动生产力空前发展，形成海量财富，催生出日常生活中的深刻变化。不过，还是有大量的人落伍了。

在摩尔发表论文半个多世纪之后的今天，这个星球上还有半数以上的人未能接入互联网，还有数十亿人的网络受限或不稳定。在世界大多数地方，收入增长与财富不均、技术性失业、社会两极分化等社会问题相互交织，在数字回音室效应和“永远在线”的社会传媒催化

下，社会的基本架构正在被撕裂。在此背景下，很多人都深深渴望未来能更丰富多彩，更少些不确定性和动荡。

前两次数字革命的负面影响并非空穴来风，也绝非偶然。在技术研发和市场化过程中，早期的决策方向和价值观会产生深远影响。我们在数字通信能力上取得突破性进展，却没能建立起相应的文化标准、反馈机制和强化公民话语权的算法。我们创造的电子商务新商业模式无比高效，但也在个人隐私和网络安全方面埋下新的隐患。我们赞赏数字自动化取得的重大进展，却要在技术带来的失业潮中苦苦挣扎。

当然，要想完全预防前两次数字革命中所有的负面冲击也是不现实的。几十年来，我们将重心放在帮助个人、组织和机构与技术的共同发展上，却失去了在技术研发过程中主动创造价值、减少危害的重大机遇。2015年，在摩尔的论文发表半个世纪之后，美国前总统奥巴马宣布“高速互联网是生活必需品，而非奢侈品”——就如同电和水一样。假如从20世纪60年代中叶开始构建以数字包容、数字与编程素养、数字文明为内容的文化，假如我们能够同时聚焦于社会创新与科技创新以应对技术冲击，我们的社会与技术体系就可能更有效地同步发展。

在前两次数字革命中，我们痛失良机。但我们现在有了另一次机遇，即第三次数字革命——制造革命。

第三次数字革命

通过将比特虚拟世界的可编程性引入原子物理世界，数字革命在前两次的基础上得到了进一步完善。鉴于物理世界正是我们的栖身之所，第三次数字革命的意义可能是前无古人的。这次革命的基石同样是数字化基础科学，不过在这次革命中，比特与原子是齐头并进以指数级速度运转的。随着通信和计算从模拟信号走向数字信号，个人电

脑、智能手机和互联网应运而生。而制造的数字化则为个人制造带来光明前景，使个人与社区可以随时随地按需制造并分享产品。

数字制造、数字化通信与计算的相似之处是显而易见的。与早期大型计算机的情形类似，大多数正规的数字制造都是通过大型机械设备来实现的，也是由顶尖研究机构和商业巨头的专门技术人员来操作的。不过，很快这些大型机械设备的性能将会被所有人获取——就像你衣兜里的小电脑能具备当年大型计算机的性能一样。由此可见，在可预见的未来，任何人都能把数据加工成物品，也可以把物品转换成数据，人们能通过比特和原子的互联网分享信息。

在前两次数字革命早期，这一前景是不可思议的。其实，它不仅理论上行得通，而且我们已经在朝这个方向行进了，道路也越走越宽广。“数制”工坊是一种依托社区的实验室，爱好者可以从中获取数字制造的强大工具。自2003年首次创立以来，“数制”工坊的数量以每一年半翻一番的速度增长。不过，与前两次数字革命初期类似，第三次数字革命以指数级增长的特性是常人不易觉察的。正如发明家、未来学家雷·库兹韦尔在其著作《奇点将至》（*The Singularity Is Near*）中所指出的那样：“指数级增长具有隐蔽性。初期几乎难以察觉，然后就会突然井喷式爆发——之所以出人意料，是因为人们不注意追踪它的运行轨迹。”

《设计现实》就是这样一本关注数字制造指数级增长轨迹的书，旨在帮助人们运用相关知识为第三次数字革命做准备，从而推进革命。每个人或多或少都有能力对这次革命做出一些贡献。我们无须再等待半个世纪来让未来的政治、教育和慈善领袖意识到智造资源的获取与智造素养是可实现的，而非难以企及。我们仍处于第三次数字革命的初期，研究重点正在确立，核心技术正在研发，对智造资源的获取与智造素养至关重要的组织和机构正在悄然兴起。

每一项新技术都以其固有特征影响着技术使用者的能力和行为。数字技术使内容得以快速复制、编辑和传播。这一特性对经济各个领

域的变革都至关重要，影响着我们的休闲方式，也改变着我们的沟通方式。在过去几十年里，我们在几乎不产生任何边际成本的前提下，得以复制、编辑和分享音乐、影像、博客、新闻、电邮、短信和其他数字资料，这一能力对于经济与社会发展既有正面推动作用，又有负面影响。而这项能力就是技术所固有的基本属性。

数字制造与数字化通信和计算既有某些相同特性，又有不同之处。在前两次数字革命中，比特对原子的影响是间接的（是通过创造新的能力与行为方式来实现的）；而第三次数字革命使比特能够直接改变原子。对大多数人而言，第三次数字革命的独特之处还不是直接在原子层面进行物质的制造，但是它的确意味着通过数字设计接口来改变实体世界的能力。尽管在前两次数字革命中，我们周围的物理世界发生了巨大变化，如道路、房屋、家电、交通、食物，但这些东西基本上还是维持了原貌。而在第三次数字革命中，物理世界构造的基本属性将会发生变化。在全球的“数制”工坊网络中，我们已经发现围绕成本效益模型形成了稳定的创新链条，个人和社区在全球范围内获取设计方案，在当地制造衣物、家具、玩具、电脑，甚至是房屋和汽车。随着数字制造技术以指数级速度增长而变得更优化、更快捷、更廉价，这些能力将会持续提升、与时俱进。

第三次数字革命契合了人类内心深处制造物品的欲望。《爱上制作》（*Make*）杂志创始人戴尔·多尔蒂在他的著作《自由创制》（*Free to Make*）中指出，“制造充分深入地体现了我们的人性，安抚了我们创造的灵魂”，并且帮助我们“建立起集自信、才华与创意于一身的个人认知”。无论是在自家地窖工作室工作的机械达人，或是印度乡下“数制”工坊的农民创客，还是创制节聚会上的12岁少年，制造物品都深深满足并启迪着业余爱好者、艺术家、工程师和发烧友。

从世界各地智造先行者的工作中，我们可以发现数字制造具备变革性威力的蛛丝马迹。通过此书，我们将邂逅来自底特律“激发焦

点‘数制’工坊”的布莱尔·埃文斯等人。布莱尔在底特律的东区开发了一块30英亩[2]的地皮，大胆探索一种新的社会与经济模式，旨在使人人都能选择“少些工作、少些花销，多些创造、多些交流”的生活方式。实现个人与社区的自给自足并非新生事物，不过布莱尔及其同事证明，如果数字化平台被用于协作生产食物、家具等实用商品的话，可以加速实现自给自足。

个人或社区层面日益强大的消费品制造能力，能够帮助我们应对前两次数字革命带来的一个最为重大的挑战——人的工作日益被技术所取代。据估计，由于人工智能、机器人的出现以及其他技术的加速发展，有一半左右的工作终将被自动化。这些消失的工作岗位不仅包括工厂工人或卡车司机等蓝领，还包括助理律师、放射科医生甚至电脑程序员等白领。关于新兴技术所产生的新工作机会是否会弥补其淘汰的岗位，也见仁见智。即便如此（虽然也不大确定），被淘汰的旧工作与能普及的新工作之间也会有数量差距。技术导致的失业高企、收入与财富不均以及数字技术带来的变化无常，混合成一剂毒药，令许多人焦虑愤懑，从而推动了世界范围内的民族主义运动的发展。挑战虽然严峻，但随着生产制造的日益民主化[3]，通过能力提升战胜恐惧心理，个人和社区自给自足与全球范围的相互依存、知识共享相得益彰，必将展现出更加动人的前景。这有助于打消将全球化与当地自给自足对立起来的错误观点，化解政治分歧。

构建更加可持续发展且繁荣富足的未来，不仅需要个人与社区层面的基础工作，还需要城市与国家层面上的参与推动。书中另一位智造先行者，来自巴塞罗那“数制”工坊的托马斯·迪兹就在领导一场全球智造城市运动。在“数制”工坊第10次全球年会上，巴塞罗那市长和首席建筑师做出一个大胆的承诺，用40年时间以可持续发展的当地制造取代其全球供应链——成为一个本地产能可以满足本地所有需求的城市。这一宣言的目的并不是让我们倒退回手工业时代，而是反

映了后工业化城市通过全球化传递的信息来完成本地制造，从而实现可持续发展的愿景。

继巴塞罗那宣言之后，其他一些城市与国家，从圣迭戈到深圳，从波士顿到不丹，都纷纷加入巴塞罗那的行列。托马斯是这样描述智造城市运动的：“我们需要通过重新定位生产来重塑我们的城市，以及城市与人、城市与自然的关系，从而使城市精神富于产出而非贪婪攫取，生机焕发而非恣意破坏，赋能于人而非冷漠疏离。”

书中我们还介绍了北到阿拉斯加、南至亚马孙的当地智造先驱。这些创新者利用先进技术延续着自给自足、社区建设等方面的遗风余俗，提供更加高效的工具使当地材料得以协同解决当地的问题，契合传统与现代社会中大量人群内心深处对拥抱大自然、与身边实体世界相融合的渴求。即使是最热心的数字技术支持者，面对愈加蛊惑人心的社会媒体与让人身临其境的虚拟世界，也能感知到被卷入而后迷失的风险。从硅谷经理人将子女送到“远离技术”的学校就读，到类似“拔掉电源日”等拒绝数字化的节日和倡议所反映的趋势，我们都能感受到这一风险带来的焦虑。第三次数字革命有助于人们更加健康平衡地分配时间，在比特数字世界与“真实”物理世界之间切换。

只有引导和启迪个人，创建必要的组织，改造传统机构，才能促使这些理想的愿景变为现实。我们首先要认识到，我们的确正处于第三次数字革命的风口浪尖，要理解它的运行轨迹。变革有迹可循，本书会为你指明方向。书中解释了第三次数字革命兴起的标志及成因，关键是，指出了如何才能做好准备并帮助推进革命。

在探索指数级发展的技术时，我们要么轻易陷入一种反乌托邦的未来设想——面对技术的肆虐横行，面对机器人窃取人类所有的工作岗位，人们束手无策；要么就沉溺于技术乌托邦的未来设想——人们什么都不用做，技术就能解决所有问题。两个极端不断被大众媒体放大着。从前两次数字革命中我们已经看到，现实要复杂得多。技术加速发展的益处与风险都真真切切，深深影响着许多人的生活，但是现

在我们有能力来调整这些影响，无论是通过个人努力，还是集体协作。

在第三次数字革命打造一个更加自给自足、互联互通、可持续发展的社会的过程中，我们都能有所作为。实现这一变革非一日之功。技术无法替代的就业岗位、利用数字制造技术的新工作会不断涌现。随着个人与社区逐渐能够自制消费品，新兴模式将会挑战传统的工作概念——为平衡生活、学习、工作、休闲提供新的选项。要使这一系列社会安排惠及每个人并不容易。不过，如果能目睹未来几十年数字制造技术的能力与素养普及10亿倍的增长过程，我们就能理解“少些工作、少些花销，多些创造、多些交流”的愿景了。在智造城市愿景的基础上，我们就能构建一个“富于产出而非贪婪攫取，生机焕发而非恣意破坏，赋能于人而非冷漠疏离”的社会。

这本书围绕两大主题展开，两者对理解数字制造的力量及前景都必不可少。一个是理解为第三次数字革命提供动能的技术，另一个是发展必须与技术共同进化的社会制度。

技术

书中关于技术理解的部分由尼尔·格申斐尔德执笔，他是麻省理工学院比特和原子中心（CBA）主任。尼尔已在数字制造的前沿工作了20年，在探求数字技术的未来方面成就斐然。他在1999年的著作《它们也思考》（*When Things Start to Think*）中，预测并帮助塑造了后来著名的物联网。他于2005年出版的书籍《智造》（*Fab*）描述了“数制”工坊的兴起和创客运动，使很多人认识到数字制造的力量和前景。

如今，在这本书中，尼尔讲述了这些趋势将如何叠加为第三次数字革命，并指出这一革命在数字制造技术发展10年之后的今天已然呼之欲出；尼尔也将通过三次数字革命的演化进程分析如何从历史的角度参悟这次革命，以及如何通过一个动态的研究路线图来透视未来。他指出，围绕3D打印的炒作只是技术革命的冰山一角，未来会出现“星际迷航”式的复制器，可以智造出（几乎）万物，包括复制它自己。要做到这一点，不仅需要将物品的设计数字化，还要将构成物品的材料数字化。

尼尔用他在麻省理工学院开设“如何智造（几乎）万物”课程，从而意外开启“数制”工坊运动的故事来开篇。接着，他描述了社区“数制”工坊如何在全球网络中广泛应用，并重点讲述了今天的数字制造流程已在赋能于人的事实。从挪威北部到非洲南端，从穷乡僻壤到繁华都市，社区“数制”工坊已经引发了创新的星火燎原之势，预示了数字制造的动能和潜能。

从这里出发，尼尔在第3章中回顾了第三次数字革命核心科学基础的历史。他解释了这次由原子到比特的革命如何与40亿年前的生命进化类比，解释了数字通信、计算和制造一以贯之的核心理念——可靠性、模块化、本地化、可逆性等基本原则。他还以史为鉴，重点分析了早在技术发展为最终形态之前，一些技术流程被发现和应用的意义。

尼尔在这部分中紧扣拉斯定律这个主题。拉斯定律是数字制造时代可以类比摩尔定律的一个重要发现，发现者为雪莉·拉斯特，她领导智造基金会（支持“数制”工坊网络的非营利机构）的工作，还负责比特与原子中心的外延服务。随着她案头关于设立“数制”工坊的申请文书不断增加，她首次发现每一年半“数制”工坊的数量就会翻一番。

尼尔编写《智造》一书时，既没有策划也没有预测到这一指数级增长趋势。当时，比特与原子中心只有最初几个“数制”工坊。对尼

尔而言，2003年是一个新的开始，就像1959年对戈登·摩尔的意义一样，他从那年开始绘制曲线。在“数制”工坊数量成倍增长近10年后的今天，随着研究路线图和智造生态系统的不断完善，我们可以预见数字制造的发展和普及会同样实现指数级增长。

参照摩尔定律问世后第一个10年里集成电路的发展趋势，我们有理由相信“数制”工坊的数量也将呈现指数级增长趋势。这意味着，在未来10年内将会出现大约100万个能力相当的“数制”工坊，再过10年就会有10亿个。这不是说将产生10亿个具有数字化设施的屋子，而是标志着每10年，技术集成、资源获取和素养普及技术就会进入一个新阶段。随着规模的扩大，内容不一定再是今天所呈现的“数制”工坊，而是制造物理形态及编写功能程序的同等能力。随着摩尔定律接近其物理与经济极限，拉斯定律将接棒这一指数级增长趋势。

这就将我们带到本书的第5章，数字制造的未来路线图。在这一章，他首先盘点和描述了“数制”工坊的当前工具配置，然后描绘了4个不同阶段：社区制造（由计算机控制设备所驱动）、个人制造（基于能够制造机器的机器）、通用制造（标志着数字化材料的过渡）、万能制造（材料的可编程化）。每个阶段都代表了数字制造发展的一段重要里程。

前两次数字革命的核心元素在20世纪60年代中叶戈登·摩尔撰写论文时，就已在实验室中被发现了。同样，第三次数字革命的所有核心元素今天已经在实验室研究中被揭示出来。问题在于，这些元素需要多长时间才能走出实验室，对社会产生影响？我们是否已经做好了相关准备？

社会

有关在第三次数字革命中，社会体系如何有效地与高速发展的技术共同进化的内容由乔尔·卡彻-格申斐尔德和艾伦·格申斐尔德执笔。乔尔是布兰迪斯大学社会政策与管理（海勒）学院的教授，劳资关系协会的前任主席。他曾研究并促进了汽车行业、航空业、卫生保健业、生物制药业以及非营利部门的大规模系统变革。乔尔是一位宏观社会学家，在发展谈判与高效能工作体系的理论和实践方面颇有建树。他目前正在开创将多方涉众在多个层级内部进行整合的新模式：涉及工作场所、企业、社会、行业、国家和国际层面。

艾伦是电子线媒体公司的总裁和联合创始人。这是一家“双底线”（既致力于积极的财务回报，又注重有益的社会影响）公司，利用数字媒体和游戏的力量帮助人们理解并改造世界。他在电子线媒体工作期间，以及之前在美国动视（Activision）工作室担任负责人和在变化游戏公司（Games for Change）担任董事长期间，曾与美国国家科学基金会（NSF）、美国国防部高级研究计划署（DARPA）、美国国际开发署（USAID）、白宫科技政策办公室、美国国立博物馆、美国公共广播电视公司（PBS）、盖茨基金会、麦克阿瑟基金会等多个机构合作开展社会影响媒体项目，吸引并帮助了全球数百万人口。

乔尔和艾伦早在首个“数制”工坊成立之初就与尼尔全力推动智造运动了，乔尔与其长子担任志愿者多年，乔尔还在伊利诺伊州的香槟-厄巴纳协助建造了一个“数制”工坊，并且主导在智造网络中应用新的涉众整合方法。艾伦研究了“数制”工坊的可持续发展商业模式，电子线媒体也与智造基金会、比特和原子中心合作，开发了一款由美国国防部高级研究计划署资助的视频游戏，探索数字制造的未来。为筹备本书，乔尔和艾伦参观了全球的“数制”工坊，采访了几十位智造先行者，调研了几百名参与者。

尼尔研究技术路线图并提供工具和技术应用的技巧，乔尔和艾伦研究社会路线图并提供社会体系与技术体系共同进化的方法与思路。在第2章中，乔尔和艾伦以观察当前“数制”工坊的全球网络来开篇。

但与尼尔不同的是，他们不只豪情满怀、着眼赋能，还强调了贯穿智造网络的矛盾与挑战。尽管个人制造前景动人，个人制造消费品的能力日渐提高，但对多数人来说，数字制造要成为现实还任重道远。在智造资源的获取、智造素养以及营造生态体系以确保技术民主化等方面，挑战依然严峻。要应对挑战，必须仔细分析挑战。研究智造生态系统中的断点，为透视根植其中的潜在假设与价值观冲突提供了一个视窗。今天，制造流程比制造的成果更能体现“数制”工坊的价值，这是因为目前掌握工作流程的难度很大。随着时间的推移，如果数字制造的影响不仅仅局限于发展新技能和制造消费品的配置，那么我们就需要调整工作重心。

与尼尔一样，乔尔和艾伦随后就追溯既往。在第4章中，他们强调摩尔定律不仅是一个技术构想，更是一个社会概念。摩尔定律绝非物理定律，而是一项被企业奉为核心经营策略的发现，它是一个行业基准，一个使数字技术更优化、更快捷、更廉价的振奋人心的框架概念。这一章揭示了社会和技术变革的相互交织是有史可鉴的，从现代社会科学，也就是对人类社会与关系的研究中，我们可以看到，这种联系是如何因工业革命的兴起而出现的。然而，由于社会科学是因技术而催生的、主导社科领域的实践，相对于技术的共同创造而言，更应当是对技术趋势的观察。要采取更主动的姿态来观察趋势，首先要理解个人、组织与机构变革的速率，这在技术高速发展的时代具有新的意义。承担调控速率这一关键作用的有两个因素，一个是积极的社会变革，而另一个则是前两次数字革命借助数字化平台在形成技术和社会体系有效共生的自发生态系统的过程中获得的经验教训。

乔尔和艾伦在第6章中以塑造数字制造未来的具体指导为结语。他们与五个大洲的智造先驱合作描述了8个期望的场景。在这些场景中，社会体系与技术体系以变革的方式共同进化。塑造未来的思维导图，不应只是描绘可能的未来，还要刻画期望的未来。为了助力将这些期望的未来变为现实，乔尔和艾伦提出了一个框架，形成新的思路，采取有效方式使人人都能在第三次数字革命中找到意义、目标和尊严。

《设计现实》通过格申斐尔德家族三兄弟的合作，从科学、技术、社会科学、人性的角度透视第三次数字革命。三兄弟不仅是革命的观察者，还是助力引导革命方向的积极参与者。三兄弟中的每个人都为本书带来不同的视角。如同透镜功能各不相同一样，他们每个人都澄清了一些事实，也带来了一些模糊的问题。既然携手撰写本书，百花齐放就比众口一词更有价值。对于迥然不同的部门与领域为了共同目标，针对复杂而快速发展的技术进行协作来说（比如第三次数字革命所需要的协作），亦是如此。

这是一个极不平凡的历史时刻：我们可以预测技术的运行轨迹，在它以令人追悔莫及的方式改造我们之前对它进行改造，现在依然来得及。不过风险也不容忽视。如果我们的设定正确，第三次数字革命带来的深刻影响将至少不低于前两次数字革命的影响。我们很快就会面临新机遇与新挑战这股洪流将直击人类生存方式的核心。亚利桑那大学素养专家詹姆斯·保罗·吉在其论文《论素养：从写作到智造》（*Literacy: From Writing to Fabbing*）中是这样总结机遇与挑战的：

> 智造有可能会深化当今的不平等，成为只有少数人才能掌握的炼金术，把思想转化为比特再转化为原子或者再反向转化。大多数人将生活在一个人生万物的世界中，包括物品、细胞和材料在内的整个世界只由少数人拥有并掌控。智造是一种新的素养，我们对它将如何发展尚不得而知。但是它是一种特殊的素养，从某种意义上来讲，很可能是终极素养……
>
> 我们当中将会有多少人会成为人类智造者？人类已经确定为终极工具制造者了。不久，制造世界的工具将足以廉价到人人都能获取，我们是否想要这样呢？作为一个物种，当我们中的一部分人或者全体成为神一样的造物者，能够化腐朽为神奇时，我们会把世界变得更好还是更坏？智造之于素养，如同火之于人类发展一样：要么照亮前程，要么付之一炬。

那么我们是要照亮前程，还是要付之一炬呢？无论是通过个人努力还是集体协作，我们都有能力去打破这个平衡。随着数字制造日益民主化，我们将拥有利用比特控制原子的能力以提升我们的生活水平。我们将能够设计现实，这是个比喻，也将成为事实。

（序言由尼尔、艾伦、乔尔撰写）

---

[1] 1加仑≈3.79升。——编者注

[2] 1英亩≈4 046.86平方米。——编者注

[3] 制造民主化，是指使普通人具有智造资源与素养的同等获取权利，而非由个别大企业与研究机构垄断智造技术与资源。——译者注

第1章

智造万物

第一次数字革命是通信的革命。在此之前，模拟电话信号是随着距离的增加而衰减的。而今天，在遍布全球的互联网上，相距再远的人交流起来也仿佛近在咫尺。

第二次数字革命是计算的革命。曾经，模拟计算机由充栋的齿轮、滑轮或真空管组成，运行时间越长，积累的错误越多。而今天，人们口袋里的便携电脑，其功能已堪比当年国家实验室的超级计算机。

我们正迎来第三次数字革命，这是制造的革命。前两次革命使通信与计算日新月异，而这一次，任何人都能智造（几乎）万物。这次革命可能会比前两次的意义更重大，因为它正利用比特王国的可编程性来塑造原子世界。

数字计算机的标志性应用是个人电脑，尽管一度被忽视，但个人电脑最终还是颠覆了整个计算机行业。同样，数字制造的标志性应用是个人制造，它能使消费者摇身一变成为创造者，就地满足个性化需求，而无须远程购买规模生产的产品。

数字制造这一概念已经有几十年的历史了，最初是指利用数控机床进行制造。事实上，它的内涵更加深刻，如同我们将在后面几章看到的，它既是全新的思想，也可以追溯到远古的概念：数字化不单是指对一个物品的描述，还包括对其进行实际构建。这与前两次数字革命有异曲同工之妙。我们认识或使用一项技术，无须等到它完全成熟。从数字制造的技术扩散及其正在产生的影响（本章的主题）来看，从三次数字革命的历史沿革来看（第3章），从未来的研究路线图来看（第5章），第三次数字革命都有迹可循。这些章节综述了理解第三次数字革命所需的科技知识，以及形成这次革命的背景。

三次数字革命的指数级发展洪流都始于最初的一缕涓流，今天，第三次数字革命正方兴未艾（如果你关注它的话）。本章讨论的是当前的形势，介绍雪莉·拉斯特的初创发现（我们称之为“拉斯定

律”），即“数制”工坊的数量每一年半翻一番。本章还将解释何为“数制”工坊、如何使用它、它的应用和含义，以及它是如何组建的。

环视当下对展望未来至关重要，因为拉斯定律这一意义非凡的概念持续发酵，如同摩尔定律一样，不仅改变了技术的内容，还将改变技术的操作者。我们将在这一章中读到先行者们利用“数制”工坊完成一系列优秀作品、塑造物品物理形态、编程物品功能的故事。今天，完成这些任务需要“数制”工坊的工具，以及支撑工坊运转的供应链。假以时日，将有更多人能够用上同样的设施，而随着技术进步，工坊运营成本将会降低，数字制造的功效将会提升。这些内容会在后面几章中做介绍。

制造：可编程的原子世界

Fab lab[1]是“数制”工坊的简称（也可以是神奇实验室的简称），是麻省理工学院比特和原子中心（CBA）的一个外延项目，我是这个中心的主任。成立比特和原子中心的目的是研究计算机科学与物理学之间的边界，这是个令我百思莫解的问题。计算的目的和用途是表达物理。比特和原子中心的研究人员参与了一些项目，比如首次使用计算机创建微观量子系统，利用其奇特的手段解决重大问题，其速度快于计算机的传统方式。另一个比特和原子中心小组致力于用计算机设计、初创一些生物有机体。这些工作没有硬件和软件之分，它们是紧密集成在一起的。该研究得出的最重要的结论，是认识到数字通信和计算与制造之间存在一个基本交汇点，这一概念将在后面几章中展开。



图1－1 冰岛韦斯特曼纳群岛的一个“数制”工坊。弗罗斯提·吉斯拉森/塞瑟尔·维都摄

创建比特和原子中心的最初资金来自美国国家科学基金会（NSF），旨在支持我们的宏伟计划，集成一套设施用以构建并测量小到分子、大到建筑的东西。比特和原子中心的设施包括价值百万美元的研究仪器，如电子显微镜和X射线扫描仪。中心内部设有一个工作室，配置有价值数十万美元的制造工具，就像用来开发实验装置的加工中心。工作室内还配有价值上万美元的快速原型工具，比如激光切割机，用来制造项目的零部件。再就是那些用于执行工序的1 000美元

左右的工具，比如模塑组件。理解“数制”工坊的关键是它处于这些嵌套的层级的中间位置，已经具备制造几乎任何东西的核心能力。

最初，“数制”工坊是一项试验，目的是看看如果广泛向外部提供比特和原子中心的内部工具会产生什么样的结果。它们的出现，实话实说，更像是为了完成一个官僚主义的指令，而非基于推动变革的愿景。2001年比特和原子中心成立之时，美国政府刚刚实施一项名为《政府绩效与结果法案》的规章，要求机构根据绩效目标来衡量工作进展，并将其在更广范围的影响记录下来。美国国家科学基金会积极响应，要求比特和原子中心等被资助机构展示其工作的广泛影响。我和我的同事当时对此毫无头绪，却碰巧遇上了印度农村的一个意外事件。

2002年我去印度出差，遇到了卡尔波，他当时在印度联合利华从事研究工作。那时他已经到了印度传统上即将退休的年纪，然而，作为一名科学家，他却选择开设一所学校（“科学学堂”，Vigyan Ashram）来教授辍学儿童科学技能。他有意把学校设在最贫苦、最干旱的地方——帕巴村，它位于马哈拉施特拉邦西部。我见到他时，他列出一个长长的清单，清单中列的都是他以前接触过、当地需要但是在帕巴村十分短缺的工具。于是，我们放弃了投资专用实验室昂贵设备的打算，与他合作装备实验室，以满足农业测量等目的。这可以称之为“数制”工坊的雏形，蛋生的第一只鸡（或者鸡生的第一颗蛋）。波士顿的第一所全装备“数制”工坊就是受其启发设立起来的，它本身也于2005年发展成一所装备齐全的“数制”工坊。尽管卡尔波在2003年去世了，但这个实验室在其继任者约各什·库尔卡尼的带领下仍继续发扬光大，教授课程，进行创业孵化，回馈社区。

为了响应美国国家科学基金会关于体现广泛影响的要求，我们计划在帕巴村的实验基础上实施一个外延项目。干劲十足的项目经理们一致同意，“数制”工坊的概念就这么诞生了。其内容基于麻省理工学院比特和原子中心管理思路下的市场研究，包含最实用的核心工具

套件。这个套件重约2吨，可以填满一个房间，目前其价值已经增加到了10万美元左右。一个“数制”工坊包括一台3D打印机，这种设备现在已在媒体上有广泛报道了，但它只是实验室数控设备的一种。在第5章，我们会接触其他设备，包括速度远快于3D打印机的激光切割机，一台可用来制造家具的大型铣床，一台可用于制造电子集成电路板和零部件铸模的小型精密铣床，组装和编程电子元器件的工具，用于物品数字化的扫描仪，用于设计与造型的计算机。

20世纪50年代，微波炉问世，它是那个时代厨房自动化的标志。当然，人们在用微波炉的同时，也可以使用炉子、烤箱或者面包机。所有这些工具都可用于加热食品，但每一种工具都有一系列不同的烹饪方法。今天“数制”工坊的工作，类似于厨房烹调。想象一下，3D打印机就如同“数制”工坊里的微波炉。你可以只用微波炉，但也会因此错过其他设备的功能。就像烹饪中有一套基本流程一样，数字制造也有一套基本流程。

可以把“数制”工坊形象地想象成一台机器：输入数据，输出物品；输入物品，输出数据。随着时间的推移，能够改变的不是我们能做什么，而是我们需要做什么。今天，“数制”工坊使用可在当地采购的大宗商品（比如木料或纸板），以及一系列小型高科技消费品（比如精密轴承和计算机芯片），就可以开始新的创造。尽管有的高技术材料目前尚不能在“数制”工坊中进行生产，但在未来也能实现。随着越来越多的供应链被越来越少的输入所代替，“数制”工坊的能力会持续提升。

我们计划设立第一批“数制”工坊时，时任麻省理工学院建筑系主任、深谋远虑的比尔·米歇尔，建议我与波士顿的梅尔·金先谈一谈。梅尔是一位社区活动家，通过摸索帮助波士顿闯出了一条城市综合开发的路子。他领导了一个名为“城市帐篷营地”的运动，迫使开发商在一个计划建造的停车场项目中增加了廉价房和社区空间。梅尔在该项目中的一个综合体上建造了南端技术中心（SETC），一时之间

成为电子媒体的宠儿，而当时主流媒体鲜有关于城市内部的报道。之后，为了应对成为互联网信息孤岛的危险，南端技术中心对计算机网络接入模式进行了创新。因此，沿着数字通信、计算再到制造这一自然路径，2003年波士顿开设了“数制”工坊。诚如梅尔所言，“除非火车自身发生戏剧性变化，否则火车后轮是永远追不上前轮的”，意思是上面提及的每一次干预都是一个颠覆性事件，改变了技术在社会中所扮演的角色。

在南端技术中心设立“数制”工坊，我们只着眼于此，并无他想。它建成后，我们原以为可以就此回归研究工作了。但是一个强大的加纳社区在发现梅尔的“数制”工坊后，于2004年在加纳海岸的赛康第－塔科拉迪合作创建了另一个“数制”工坊。一而二，二而四，10年间“数制”工坊的数量每一年半就翻一番：每次我们建立一所“数制”工坊，就有人提出需要再建一所。雪莉·拉斯特是第一个注意到这一指数级增长趋势的人。她负责管理比特和原子中心的“数制”工坊项目，领导着一个从比特和原子中心剥离出来的智造基金会来支持项目的发展。拉斯特来比特和原子中心时，其身份是一名电视科学节目制作人。她崇尚科学发现，而这也是她自负责“数制”工坊网络项目以来孜孜以求的方向。

2005年，我在最初一批“数制”工坊设立之后，写了《智造》（*Fab*）一书。那一年，一家名为《爱上制作》的新杂志找到我并首次采访了我，这个杂志是由戴尔·多尔蒂创办的。他把那些用“数制”工坊的工具将计算与制造联系起来的社区爱好者命名为“创客”。2006年，他在加利福尼亚的圣马特奥发起了一个活动，名为“创客嘉年华”，现已成为当地最大的创客集会，虽然最初只寥寥数人，但2015年时已吸引145 000人与会。

同样也是在2006年，吉姆·牛顿创建技术工坊，以分享大多数人可望而不可即的数字制造工具。其店面以付费会员的方式运营。发展到2016年，他共开了10家分店。还有一些为同类爱好者提供聚会场地

的非正式创客空间和骇客空间，这种空间由于提供的服务不同而形态各异，目前已有数千家。

次年，也就是2007年，比特和原子中心推出了一个移动“数制”工坊，目的是把工具带给人们，而非让人们来找工具。值得纪念的是，托马斯·迪兹、埃米·孙、张永安（后面还会提到）驾驶着工坊车参加内华达黑岩沙漠的“烧人”狂欢节，在湖床上展示了快速原型制作。移动工坊后来在网络中孕育出另一个网络：全国各地工坊漫游网。

所有这些“数制”工坊都是第三次数字革命早期形态的展现。在这一快速增长的生态系统中，“数制”工坊的两个特征格外引人注目。一个是所有的“数制”工坊配备的核心套件都是标准化的，而且同步升级，使不同工坊的人和项目都可以共享，避免了工坊间的差异性。计算机网络先驱鲍勃·梅特卡夫发现了著名的“梅特卡夫定律”：计算机联网的价值与网络中计算机数量的平方成正比。他之所以认为这是指数级的影响，是因为它反映了人与人交互的可能性。有些创客空间采用会员模式，会员仿佛加入了一个健身房。健身房向个人提供训练用的设备，而如果某人在其他地方训练，健身房就不产生直接的利益。但是，联网或使用“数制”工坊的价值，会在其他计算机或其他人加入网络或“数制”工坊时得以提高。在“数制”工坊网络中，人与项目是流动的，内容得以分享，人们可以合作完成自己单独无法完成的任务。

“数制”工坊的另一个显著特点，是其内容可根据第5章提到的数字制造研究路线图进化升级。一开始，它们是一张仔细编制的普通机器、材料、零件和项目清单，现在发展到由“数制”工坊开发的硬件和软件的开放式设计，目标是由一个“数制”工坊创制另一个“数制”工坊。尽管“数制”工坊中各种设备的成本随着时间推移有所下降，有关工坊可制造物品的雄心壮志却在同幅度增长。因此，从社区资源的角度来衡量，总体上的规模保持基本稳定。

这些特点结合在一起，使“数制”工坊形成一个网络发挥作用。单独来看，每一个工坊可能并不起眼，但所有工坊合在一起就积微成著了。没人强迫你建立“数制”工坊，但是很多网站持续加入这个网络，是因为它们可以从一个更大的网络中获益。

最令我诧异的是世界各地“数制”工坊的用途的相似性，而非差异性。梅尔·金是一位来自波士顿的社区活动家，当我们带他去挪威北部会见有萨米人血统的牧人哈肯·卡尔森时，梅尔也注意到了这一点。在哈肯的工坊逗留数日后，梅尔评价：“就跟在家一样。”他可能只在北极圈里盘桓了几个小时，但在那里他体会到的希望与恐惧跟他在波士顿城市工坊里如出一辙。

多样性是各地“数制”工坊的另一个共通之处。参与者从黄口孺子到皓首苍颜，跨越各个年龄段，其应用还涵盖教育、娱乐和商业等多种方向。在这种多样性中，工坊扮演着类似于图书馆的角色。安德鲁·卡内基于上一个世纪之交（1900年）投资设立城市图书馆，他功成身退时，已有大约250个此类图书馆遍布全美。该图书馆的总体使命是普及文化，增加获取知识的途径。然而在完成这一使命的过程中，他们努力使其普惠于托儿所、学校、研究机构，乃至所有公民。当时图书馆还是个新生事物，而今它已经成为一个文明社区的标配。“数制”工坊也具备同样的使命，但是它们所瞄准的是一种新型的文化普及：从比特到原子的转化。

教育：从DIY到DIT

在比特和原子中心建立数字制造研究设施之初，我们曾面临一个问题：由于这些工具传统上是按照学科与操作量级来分类的，麻省理工学院的学生终其一生也无法掌握所有工具的用途。于是我走了个捷

径，在2001年开设一门新课程：如何智造（几乎）万物。这门课主要面向从事数字制造研究的少数学生，不过自开设以来，每年都有数百名学生来听课，他们就是想要学习如何进行创制。



图1－2 汉斯-克里斯蒂安·布鲁沃尔德（左）与策皮色·莫娜蘅在“数制”工坊中工作。尼尔·格申斐尔德摄

除了掌握单项技术外，他们还将技术集成起来做项目。其中一位明星学员名叫凯丽·多布森，后来成为罗得岛设计学院数字与媒体系

的系主任。她制作了一个用于收集尖叫声的可穿戴装置，这个装置还能在之后方便或有需要时对收集到的尖叫声进行回放。几年后，后来成为麻省理工学院建筑系系主任的尹美真，制作了一条装满传感器和刺针的裙子，以捍卫着装者的个人空间。此类发明项目年复一年不断涌现，我意识到参与这个课程的学生正在回答一个我未能提出的问题：数字制造的好处究竟在哪儿？当我叩问数字制造的途径而非缘由时，他们则以行动回答，如同数字计算的杀手级应用是个人电脑一样，数字制造的杀手级应用是个人制造。重点不是制造市面上可以购置的东西，而是制造市场小到只有一个人的个性化商品。

比特和原子中心的研究工具面临的培训难题，被“如何智造（几乎）万物”课程所化解。同样，“数制”工坊的普及也面临着全球性的培训难题。如果聪明的孩子在“数制”工坊中所学的技能大大超出了当地的教学范围，他们就可能会愤而出走。汉斯－克里斯蒂安·布鲁沃尔德生活在遥远的挪威北部，从当地的林赛德教学体系标准来看，他是个问题学生。因为他已经掌握了老师教授的所有内容，所以上课很不专心。后来他尝试去哈肯的“数制”工坊，我在那里见到他，并把“如何智造（几乎）万物”课程中的一些项目演示给他看。当我再次回到哈肯的“数制”工坊时，震惊地发现他已经用所学技术创制了一个玩具机器人卡车，从设计车身，到组合舵机和控制器，他还给玩具卡车增加了一个挡风玻璃显示器。与此同时，我们在南非一个曾处于种族隔离时代的索山谷小镇也开设了“数制”工坊，并惊喜地发现当地一个名叫策皮色·莫娜蘅的姑娘已经通过工坊的设施与我在麻省理工学院的授课内容实现了远程同步，这令我惊讶不已。

对汉斯－克里斯蒂安和策皮色这些人来讲，人们通常会说他们“才华横溢，当志存高远”。为了获取更高深的学问，这些出色的学生只能远走他乡去求学。不过，这种迁移却使最有价值的人远离了最需要他们的地方。我们试图与各地的学校合作以弥补这一遗憾，却总是发现比缺乏技术技能更大的局限在于，学校严格管制的教育方式会

遏制学生的创造力。因此，我们启动了如今称作“智造学院”的项目。

这个项目一开始是个联网视频，当时我创建它是为了方便各地的“数制”工坊远程旁听麻省理工学院“如何智造（几乎）万物”这门课程。后来，加入听课的“数制”工坊比课堂上的学生还多，我们就把远程教学部分拆分出来，作为一个独立的课程设置。课程的当地导师由曾经参加过远程学习的那些人来担任，其作用就十分重要了。只要新生决定加入当地的“数制”工坊工作组，就要与导师、同学和机器为伍。于是，我们把全球各地的人用联网视频聚到一起，进行互动讲座，分享合作内容。

可以把麻省理工学院想象成一个大型计算机，你在其中做数据处理。当人数有限时，它还运行良好。也可以把大型开放式在线课程（MOOCs）视作分时时代的计算机，用户分坐于中央主机不同的终端前进行工作。而我们无心插柳创建起的智造学院模式则更像互联网，联结学习网络的众多节点，其中的边缘节点比中心发展得更快。

一开始我直接指导所有学生。后来随着模式的发展，我只督导各地导师，由他们来指导学生。当模式进一步深入发展后，我就只管理那些超级节点的负责人，由他们督导区域内的工坊，再由工坊的导师来指导学生。这样，这一模式又与互联网模式不谋而合，沿着一个树状路径来传递信息，有树干、树枝和树叶。与互联网一样，任何节点之间都可以交互，每周智造学院的例会本质上就是一个大型视频会议，实现了人与人之间的视听交互功能。会议内容包括对过去一周成败得失的热议，并对下周要推出的新材料进行互动介绍。整个会议的技术支撑都是由卢恰纳·阿辛纳利领导的分布式工作组来完成的，而非中心办公室。

这一架构在维护质量控制的关系网中，实现了直接的可追溯性。但是我们需要用某种方式把它记录下来，于是我就去找了EDUCAUSE（一个高等教育的信息技术专业组织），“.edu”（教育）

的域名就是由这个组织管理的。申请“.edu”域名的机构需要该组织对网站进行授权。接待我的几位授权人对我们的事业表示欣赏，却说如果给智造学院一个“.edu”域名就会违反他们的规定，因为获得其授权的机构必须要有可识别的实体场所，而他们无法识别网络。不过，授权人随后表达了善意：“就当作你们有场所吧。”他们的意思是我们的学生必须建立一个文件夹记录所学技能，然后他们再派人上门检查学生的学业并评估学院的价值。虽然智造学院没能取得全球的授权，但区域性的授权依然使智造学院桃李满园。从中我们也发现，就学生未来的深造、就业或投资资质而言，这个文件夹比陌生人的证明函更能说明问题。对授课内容的审查与评估历时8个月，不过对学生进步程度的评估是由其技能掌握程度而非听课时长来决定的，因为有些学生用了好几年才完成学业。学生的背景各不相同，包括在家里接受教育的神童、在校大学生、不上大学以制作为业的人、职业生涯中期的专业人士、职业生涯后期需要再培训的人、退休的业余爱好者。这种当地学习小组的全球联网，可以根据对导师的需求量来平衡“数制”工坊的分布。

缺乏良师，则易误入歧途。“创客”这个词其实具有正反两个方面的内涵，用来形容热情很高但信息却不太灵通的人。创客运动的一个主题是Arduino[2]，一款售价20美元的小型电脑插件，用于在项目中构建智能装置，包括读取传感器、控制输出装置和网络通信装置。Arduino原本是在一个计算机芯片家庭AVRs的基础上设计的，设计者是两个挪威学生。智造学院在引入Arduino以后，会先向学生展示如何用寥寥几美元的部件制造这样一个电路板。之后，再教他们学习如何使用其他电脑芯片，有的电脑芯片只有一粒米大，有的电脑芯片则可以驱动桌面操作系统。创客运动的另一个主题是3D打印。智造学院不仅教学生如何使用3D打印机，还演示如何使用那些运行速度更快，产品尺寸更大、硬度更高、质地更好的数字制造工具。后来，学生们又学会了如何制造3D打印机。上述每个事例都为从学习简单技能过渡到掌握复杂技术提供了可参考的路径。

DIY[3]只能使你在前人的基础上略有进步，而无法集前人之大成，DIT[4]和DIWO[5]则是构建在群策群力基础上的活动。我们意外地发现，智造学院处于一个良性循环的核心。每个循环在“数制”工坊全网推广最佳实践案例，建立当地导师的核心队伍，向社会输送大批训练有素的学生，从而又对新的工坊与项目有所助益。

创建智造学院是为了教授数字制造技术。不过，我们之前并没有局限于这一内容来设计学院，其基础设施可以用于任何一种分布式学习，而非单纯的远程学习。早年我忽视了通信、计算、制造与学习之间的深刻联系。如果说数字通信使我们得以全球交互，数字计算使我们能够分享知识，那么数字制造的兴起则使我们可以交换思想甚至物品。利用“数制”工坊的核心工具套件，“随需而制”已经成为可能，可以有效地把校园搬到学生面前。

世界著名的遗传学家乔治·丘奇很关心那些无法到哈佛听他讲课的学生。这一想法促使他利用“数制”工坊的网络增设了分布式的第二课堂，主题是“如何种植（几乎）万物”。数字制造与生物制造在两个层面联系密切。生物学家可以借用“数制”工坊制造生物实验室所需的工具；生物设备常常昂贵而笨重，“数制”工坊制造机器的技术同样被用来制造用于DNA扩增的热循环仪和为生物反应编程的液体处理机器人。从一个更深的层面来看，生物本身也可以用于制造。我们在第3章将会看到，生物反应的过程基本上是数字化的，我们正在逐渐学会如何对这些过程进行编程，从而实现“数制”工坊与生物实验室的功能交汇。

奥拉维尔·艾利亚松是世界知名的艺术家。与乔治一样，他想要扩大工作室的影响，使其受益于无法直接接受现场培训的学生。不过，他的兴趣不在于如何创制物品，而是为何创制。这促使他也开设了一门分布式课程——“为何智造（几乎）万物”，来探索制造流程的影响和冲击。

我的学生纳迪娅·皮克把这种不断发展的课程集合戏称为“（几乎）万物学术”。学术名称是个难题，只好先简称为“Academany”，目前由让－米歇尔·莫莱纳管理，法国格勒诺布尔“数制”工坊（Grenoble fab lab）就是由他启动的。每个课程都遵循同样的模式，有当地的工作组，也有全球联网的导师，可以由负责分享协作内容的世界级大神们开展交互式讲座。

与“数制”工坊在全球开枝散叶同步，我帮助麻省理工学院设计了一个新建筑。这项工程历时10年，耗资上亿美元，可容纳数百人。在这10年中，涌现了上千个“数制”工坊，每个“数制”工坊都拥有一个用户规模在百人以上的社区。这些数字反映的问题不言而喻：在上亿美元的建筑与上亿美元的“数制”工坊之间，投资哪个更加合理？

麻省理工学院当前的组织架构是建立在稀缺性这一假设基础上的。受实验室设备、图书馆藏书，以及教职员工的时间精力等资源所限，我们拒绝了大多数申请者，挤在剑桥[6]的一角，在那里，每平方英尺[7]都弥漫着对资源的竞争。有人顾虑某个学生有可能正在孤苦伶仃地坐在联网在线学习平台的电脑前，这是个伪命题。智造学院已经反复证明，学生要成功就需要置身于某种学习环境中。现实的选择是分布式教育而非远程教育，就像智造学院所做的那样。随之而来的问题就是，对于麻省理工学院这种地方而言，有多少内容是可以进行分布式教学的？又有多少内容需要在校园内集中教学？

我看应当是平分秋色。我们每开设一个“数制”工坊，都能遇到极具创造性的人才，他们和麻省理工学院那些与我共事的人同样出色。这种人遍布全球，之所以频繁现身于“数制”工坊，是因为他们在其他地方找不到知己、导师和工具。麻省理工学院的校园里大概有半数的教学活动可以用“数制”工坊的形式来代替。另一半情况有所不同，它们需要昂贵的工具，比如用来编写分子级别汇编程序的纳米装置。使用这些昂贵的工具需要极为专业的技能与知识，因此将其教

学研究活动集中在一地进行是有道理的。这两种空间形式并不矛盾，我们可以将其视作一棵树的枝叶，其中有万元级别的创客空间、10万元级别的“数制”工坊、百万元级别的超级“数制”工坊，也有千万元级别的研究实验室。要想利用地球上的智慧，不仅要种植树木，还要让它枝繁叶茂。

西摩·派帕特是教育信息化之父，曾在瑞士师从于著名儿童心理学家让·皮亚杰。皮亚杰坚持认为研究儿童的学习方式应当像科学家一样进行实验、检验理论。后来，西摩在麻省理工学院搞到几台早期的实时数字电脑，希望能拓展儿童可及的实验范围。这个想法在当时有些不切实际——那些电脑既昂贵又庞大，而且很难使用。为了提供一个更加友好的界面，西摩开发了一款机器人“海龟”联结到电脑上，还开发了一种计算机语言（Logo）以方便儿童指挥海龟行动。

与西摩共事的人中有一位名叫艾伦·凯，他在此基础上开创了现代计算机图像用户界面的范例，并开发出笔记本电脑。这些设计最初并不是为商业主管们准备资产负债表，而是为了方便儿童进行探索。西摩的另一位研究伙伴米奇·雷斯尼克则开发出乐高的头脑风暴套件（是以西摩的一本书名来命名的），使电脑融入可编程的乐高积木游戏。米奇还领衔创造了供孩子们编程使用的Scratch软件[8]，风靡全球。

随着“数制”工坊数量的快速增长，智造学院不断发展，西摩来找我商量合作。那时我觉得“数制”工坊纯粹就是个历史性偶然事件，但他却十分有信心地说，孩子们可以对“海龟”的行动进行编程，却不能制造出“海龟”，这对他来说如芒在背。这是他始终追求的目标。从这个角度看，“数制”工坊提供的教育直接延续了他数十年以来奋斗的事业。这绝非偶然，从在麻省理工学院玩中央计算机，到去商店花钱玩电脑驱动的游戏，再到去“数制”工坊玩创制的电脑，这其中是有自然发展规律的。

应用：漫长供应链的定制化可能

一旦你获取了“数制”工坊的工具，并掌握了其使用技巧，就有可能在当地制造出各种各样处于漫长供应链末端的物品。“数制”工坊不但具备使用当地技能并创造当地就业的好处，还使按需生产与本地需求定制化成为可能。下面就介绍几个“数制”工坊运作的例子，意在说明问题，但并不十分详尽（同时也避免枯燥乏味）。

脱胎换骨的手艺

库克湾部落理事会（CITC）是阿拉斯加库克湾地区一个非营利性的部落服务机构。它所服务的阿拉斯加当地社区具有深厚的文化传统，但也面临着高失业率、高酗酒率、高自杀率等严重问题。库克湾部落理事会管理着一个2013年建成的“数制”工坊，致力于将文化与技术融合起来，以服务下一代，以淡化他们成长中的本土色彩，使他们生活于全新的数字化设备环境中。根据白宫2014年本土青年报告，当地高中学生的毕业率在全美所有族群学校中是最低的。



图1－3 分别来自美国阿拉斯加州、墨西哥、美国缅因州和日本的“数制”工坊工艺品（从左上角起按顺时针顺序，每两幅图出自一个地方）。本杰明·亨特-弗朗西斯二世、冈萨洛·佩雷斯·拉米雷斯、安德丽娅·德佐、田中裕也摄

本杰明·亨特－弗朗西斯二世是一个16岁的少年，他从阿拉斯加一个只有349人的马绍尔村搬到安克雷奇，也处于无法从高中毕业的危险边缘。在学校里他的学业远远落后，与当地出生长大的孩子文化意识迥然不同。他成为“数制”工坊的常客，利用工坊设施提升课业，还做自己的项目。其中一个项目是制作一款传统样式的木制雪橇，但他利用数控设备制成，还雕刻上代表自己社区的图案。在另一个项目中，他利用激光切割机完成了镶嵌作品，传统上这类作品是手工刻制的，费时费力。他由此赶上了学业进度，也喜欢上了学校，而且觉得“数制”工坊能帮助他继承先辈遗志，弘扬传统文化。

草垛山工艺学院是美国艺术类院校的最高学府之一。玻璃吹制、铁艺、陶艺、版画等很多领域的艺术大师都归隐到这个位于缅因州海岸的工作室群落，锤炼并教授他们的技艺。2009年，时任院长斯图·凯斯腾鲍姆提出了一个引进数字化制造工具的试验计划，于是我们在那里设立了一个临时的“数制”工坊。当时的场面有点儿像1965年鲍勃·迪伦[9]带着一只电吉他现身新港民歌节，成为一个被观察家描述为“一半被电击，一半被电死”的事件。半数艺术家对技术入侵传统艺术殿堂感到惊恐不已；而另外一半则震惊于前一半没能认识到其所有艺术实践原本就是以当时的新技术为基础的，而这次只不过是另一次技术进步而已。

如此之多的艺术家在“数制”工坊经历了脱胎换骨的改变，于是这个工坊在2011年改为固定设施，唯一的矛盾就是激光切割机多得令人应接不暇。不过，“数制”工坊不仅被视作一种工艺形式，还被广泛应用于各种工艺加工。它并不是对艺术家技艺的替代，而是一种对技艺的发展。艺术家们一般都利用传统媒介进行设计，然后利用“数制”工坊的工具以手工制作难以企及的方式完成作品。

安德丽娅·德佐是一位视觉艺术的访问学者，擅长制作美轮美奂、精巧复杂的剪纸和印花。她先画一个草图，随后就是旷日持久、劳心费力的剪切过程。而在“数制”工坊，她只要把草图扫描进电脑，就可以迅速完成作品，所呈现的成品与手工制作的相比既大又精致。对她来说最有趣的莫过于与机器的协同作业，常常产生出乎意料的效果，有时甚至比她设想的还要好。比如，在剪切一个设计图案时，当刀具运行轨迹不经意间超出走刀步长的设定值时，她发现，剩下的材料在图案周边形成了令人回味的能量场。

墨西哥北部安纳华克大学于2012年开设了墨西哥第一家“数制”工坊。这个工坊非常注重社会效益，助力那些被边缘化了的墨西哥妇女制造当地工艺品。其中一个项目与一位单身母亲和她的6个女儿以及她们的表亲——阿雷奥拉家族有关。尽管他们没什么工作，但是阿雷

奥拉家族会制造手工巧克力。不过巧克力外观粗糙，限制了他们的销售。他们借助“数制”工坊制造巧克力模具，定制个性化礼品包装，结果产品质量大大提高，显著提升了销量，还省时省力。

镰仓“数制”工坊，由田中裕也和渡边洋香于2011年创建，占用的是一个有150年历史的木结构库房，那里曾是个清酒酿造厂。这个工坊致力于把日本传统工艺与现代化工具结合起来。有一个项目是把日本漆器套在3D打印产品外面，形成一个外观漂亮、坚韧、光滑的成品。另一个项目是关于这个建筑本身。针对“数制”工坊过度使用简单凸舌结构的情况，工坊的一名学生金崎健二总结了日本传统木工中的54种榫卯结构，极大丰富了木质结构的连接方式。他把这些成果编成了一套可用“数制”工坊工具制成的参考设计文件，把日本传统榫卯与数字化制造实践紧密地结合了起来。

从贫民区到庇护所

戴维·雅姆尼特斯基是2013年来上“如何智造（几乎）万物”这门课的。我在课上最喜欢布置的一项作业是让学生们“造个大东西”。我是在教授学生使用大型工具时布置这个作业的，那是一台自动铣床，可以处理分发给学生们的一块4英尺×8英尺的胶合板。雅姆尼特斯基的女朋友想要一张立式办公桌，但在市面上找不到合适的款式，因此他就利用这个作业为她定制了一张。为了造出桌子，他设计并裁制了一套组件——类似于那种可以从大卖场买到的平板组合家具。不过，和那些大宗生产的家具不同，这套组件的每个部件都是独特的。许多人在看到他的作品后，都提出了不同的样式需求。后来这成了一个创业项目，获得了好几轮融资。我有一个学生叫阿曼达·加西，她也找不到一款喜欢的立式课桌。与戴维的桌子不同，她需要的是在桌子下面设计出存储空间，以容纳工作所需用品。完成设计之后，她一个下午就分步制造并组装出了桌子。

2016年，卢旺达开设了一个“数制”工坊。基加利的家具是在加金吉罗地区进行制售的，那是一个脏乱差的地方，人们用手工工具加工家具。一位名叫奎恩（雷内·伊曼尼施姆维）的年轻女士发现，她无须跑到家具市场，只要到“数制”工坊就能按自己的设计制造家具，就像戴维和阿曼达那样。她开始利用一块从厨房天花板回收来的木料进行加工，使“数制”工坊里里外外都是自己定制的家具。

欧哈德·梅雨思在2012年参加了智造学院的学习之后，就于2013年在特拉维夫南部的霍隆开设了一个“数制”工坊。霍隆是个贫民区，住着大量移民，犯罪率很高。“数制”工坊隔壁是个社区中心，对当地孩子来说，这是个安全的庇护所，却也是个荒芜的地方。2014年，他利用“数制”工坊对社区中心进行彻底改造，孩子们设计了契合这个空间的实用家具，然后在一个星期内就造出了所有东西。其结果是，他们不仅营造了一个令人骄傲的热闹场所，还激发了设计与创造的热情。



图1－4 “数制”工坊家具，来源于戴维（左上）、欧哈德（右上）、奎恩（下排中间和右下）和阿曼达（左下）。戴维·雅姆尼特斯基、欧哈德·梅雨思、格蕾丝·科普尔斯通、阿曼达·加西摄

博物馆庭院里的拖车式活动房屋

收容所2.0项目是由两位木匠（罗伯特·布里奇斯和比尔·扬）启动的，他们俩也是数字化制造的先行者。这一背景促使比尔就职于

ShopBot——一款热门的大型铣床制造商，“数制”工坊也配备了这款铣床。受海地救灾工作中凸显的对收容所的需求启发，按需现场建造、个人需求定制的前景深深触动了他们。他们发明了一项开源设计，分享了设计文件，并向海地运送了一批平板式收容所的制作材料，当地难民很快就将收容所搭建了起来。这些收容所比周边的临时住所更牢固，而且每平方英尺的成本只有几美元。

拉里·萨斯从小在纽约哈林区长大，因酷爱建筑而就读于麻省理工学院，后来成为比特和原子中心的第一批教员。他致力于利用数字制造技术大规模定制住所，将少数人才能拥有的定制建筑推广至普通人。“卡特里娜”飓风灾害一年后，他发现新奥尔良的灾民仍然居住在联邦应急管理中心的拖车式活动房屋里。于是，他就去找当地的一些房主进行调研。后来，他为新奥尔良开发可利用数字制造技术制造的住房。这是一种联排结构，现在已经有人提议用美式住所设计来解决非洲和海地的住房问题，就在新奥尔良联排住房样板的基础上进行改装。他尝试利用激光切割工具设计模型，然后在比尔·扬的帮助下扩大规模，裁切符合实际尺寸的组件。这些组件只需要一个橡皮锤就可以进行组装了，首次安装后于2008年在纽约市现代艺术博物馆的庭院里展出。

比森特·瓜利亚尔特组建了西班牙加泰罗尼亚高级建筑研究院，这也是巴塞罗那首个“数制”工坊所在地，他后来成为该市的首席建筑师。托马斯·迪兹和丹尼·依班娜带领一个由来自25个国家的成员组成的团队，创建了一个“数制”工坊住所，并于2010年荣获欧洲太阳能十项全能竞赛的“人民选择”奖。其他几个参赛项目试图利用大型3D打印机打印房子，这个方法耗资巨大而成效有限，而他们这个团队却能裁切出可以现场组装的独特组件，组装出完整的太阳能住房，甚至包括房内的设施。

瓜利亚尔特的设计对太阳能转化利用设备以及通风装置进行了优化，使其成为建筑不可分割的一部分。这种建筑因其造型而被称作

“花生住房”、“肉桂潜艇”、“森林飞艇”或“鲸鱼之腹”，有效实现了能源的供大于求。团队成员收到了不少复制这种建筑的订单，他们将其扩展为一个可调整的住房设计组合，广泛适用于从简装农舍到大型别墅的各类住所。

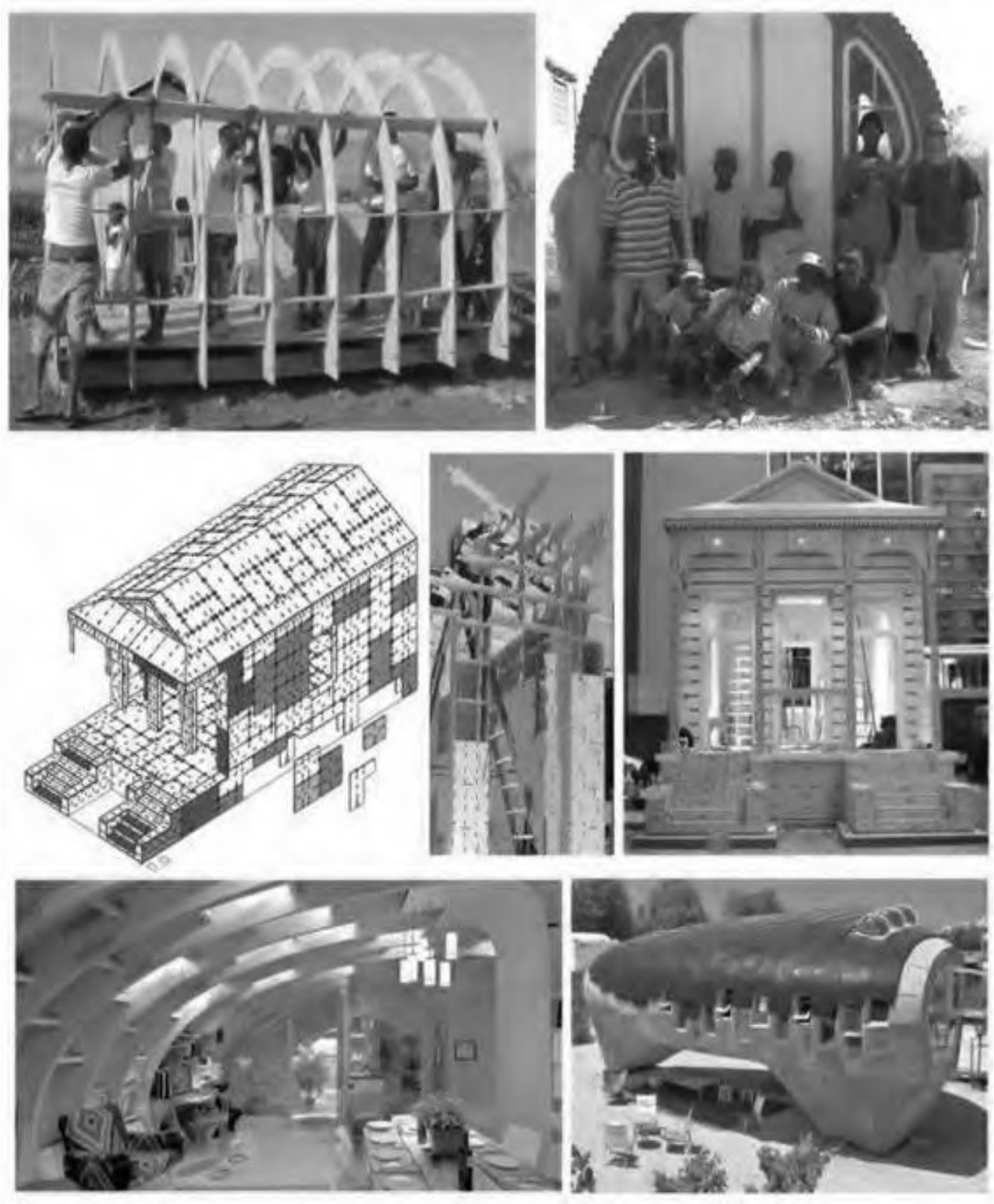


图1－5 数字制造住所（由上到下）：收容所2.0项目、新奥尔良数字制造住所、“数制”工坊住所。比尔·扬、拉里·萨斯、巴塞罗那“数制”工坊（加泰罗尼亚高级建筑研究院）摄

无人机

在“数制”工坊中创制的各类机器人中，最时髦的大概是那些能飞的东西——无人机。人们对无人机的爱好有的是消遣性的，有的则很专业。无人机的机械装置、推动器、驱动电子装置、控制系统和通信系统都可以在“数制”工坊制造出来。通过这些在“数制”工坊造出的组件，创客们可以定制一款无人机，其乐趣远远超过仅从目录中选出几个小装置来安装。单从技术而论，发动机也可以在“数制”工坊造出来，不过它是规模生产的商品，无须定制。

马特·诺里斯是位航空工程师，他在塔尔萨市创建了一所“数制”工坊。这里有个项目招募当地老师学习制造和操作无人机的技术。老师再把这种技术传授给学生，帮助他们培养全面掌握无人机技术的兴趣，而不是让他们只把无人机当作一个航空记录仪来使用。

在2014年进入巴塞罗那智造学院学习之前，奇拉格·兰霍力亚在印度攻读建筑学。他的毕业设计是开发一款无人机，并将其与一个可编程摄影支架集成起来，而这个支架又是阿尔多·索拉佐的毕业设计。后来，两人在毕业设计的基础上合作，建立了网络环境机器人组织（NERO）。他们不卖无人机，而是提供无人机收集的数据。该组织向巴塞罗那城市发展局提供NDVI（植被归一化指数）的测量数据，帮助该市进行绿化规划与管理。该组织还在哥斯达黎加编制农业地图，帮助该地治理虫害。这种方式比卫星图像成本低，而且更易于形成定制化服务。

丹尼尔·英格拉西亚首次造访托里诺的“数制”工坊时，还是意大利一家信息技术公司的研究员。此行使他大开眼界，于是他在2015年投身米兰的Opendot工坊参加智造学院的学习。他的毕业设计也是一台无人机，不过他的无人机集成了导航系统，可以避开障碍物，直接到达目的地。他沉迷于此不能自拔，辞去了原来的工作，后来被莱茵－瓦尔应用科学大学开设的智造经济网站（“数制”工坊的一个商业

性平台）录用。他一边经营“数制”工坊，在智造学院担任导师，一边继续研究无人机。目前，他正在研发一系列不同尺寸的无人机，这些无人机都可以在“数制”工坊制成，其控制系统还集成了智能水平更高的装置。



图1－6 “数制”工坊无人机系列。丹尼尔·英格拉西亚摄

乔纳森·莱贾德长期担任《经济学人》杂志驻非洲的记者。他目前在洛桑联邦理工学院与英国建筑师洛德·诺曼·福斯特合作，在非洲建设“无人机港”项目。福斯特事务所已具备部分世界性大型机场的设计经验，而此类“无人机港”将是其中最小的一种。在非洲，只有1/3人口的住所在方圆两公里以内拥有全年可通行的道路。莱贾德和福斯特的目标不是在非洲大陆修路架桥，而是实现空中的互联互通。他们最初集中精力发展高价值货物的运输，这些货物包括医疗物资和水泵等关键设备的零部件。莱贾德与福斯特从卢旺达开始做起，这是一个地貌险恶却管理有序的国家。我在卢旺达推动当地首家“数制”工坊的建设时，莱贾德在一个非洲领导人的会议上找到了我，请求我

将“数制”工坊建设在“无人机港”附近，以方便无人机的维护和修理。我解释说，他们其实可以再深入一步，在“无人机港”制造无人机。这种无人机不是简单的悬停四轴飞行器，为了提高远程效率，需要制作固定翼的飞行器，利用三维塑形模具将纤维树脂复合材料堆积成高性能的结构。我的学生格雷斯·科普尔斯通回到卢旺达协助“数制”工坊筹备无人机制造流程的首次展演。这次展演促使卢旺达组成一个本地无人机开发小组。

迄今为止，最大的无人机是由我以前的学生雅艾尔·马圭尔研制的，她在2016年时担任脸书公司（Facebook）连通实验室主任。这种无人机航程远，飞行高度高，其翼展堪比大型客机，配备太阳能动力系统，提供全球范围的空中互联。现在，这款无人机需要吸引投资，建造大型客机所需的生产线以投入生产。不过，随着对大型复合体离散式组装研究的不断推进（详见第5章），此类无人机也会被纳入未来的“数制”工坊力所能及的范围。

羊只与驯鹿

2002年，挪威电信公司派人找到我，商谈在奥斯陆合作一个智能家居项目。我解释说，此类项目如此泛滥，没什么人会感兴趣的。然后我问他，挪威电信公司能否另辟蹊径。来人经过一番尴尬的推诿，说公司正在与一个古怪的牧羊人合作，在挪威最北部的羊只与驯鹿身上安装手机。这位牧羊人就是哈肯·卡尔森，他想通过这种方法来追踪牧群。传统上，萨米人放牧都是游牧式的，人跟着动物跑。随着土地使用方式的转变，牧人越来越需要在某地定居，但他们还想保持远距离的联络。哈肯的项目促成了一次合作。我们设立了一个“数制”工坊，初期致力于制造天线和无线电设备，以在手机信号塔覆盖不到的地方建立无线通信网络。这个项目后来被南非茨瓦内的“数制”工坊复制，以延伸互联网的接入。南非电信运营商制度严苛，一些地方

很难获取互联网通信能力。我的学生埃米·孙与基思·勃克本、斯玛丽·麦卡锡（现在是冰岛议会的海盗党议员）合作，把这个项目带入了阿富汗贾拉拉巴德“数制”工坊，那里根本就没有互联网的接入。贾拉拉巴德项目后来发展到包含了45个站点，两个站点间最长距离不超过6公里。这个项目被称为FabFi。肯尼亚的“数制”工坊也借鉴了这个项目，部署了超过50个站点，实行商业化运营。与商务实体建设运营的通信基础设施不同，这个项目建立的是区域性社区无线网络设施。



图1-7 “数制”工坊制作的数据网络与通信终端。埃米·孙、基思·勃克本、斯玛丽·麦卡锡、戴维·克拉诺、马克斯·洛博夫斯基摄

我们只需要花大约100美元在“数制”工坊组装一台单板机，再加上一个小屏幕和键盘，就可以制作一台简单的定制笔记本电脑，以连接这些网络。不过，如果只想获取互联网接入，还能更加简单。马克斯·洛博夫斯基和戴维·克拉诺（创建了3D打印机公司Formlabs）在2009年参加了“如何智造（几乎）万物”课程之后，继续从事一项称为薄客户机（小型电脑）的项目。“薄客户机”是计算机领域的一个老套路，就是一台没有本地储存、把所有内容储存在远程服务器上的

简易电脑。这种电脑的现代版本就是流行的谷歌网络笔记本，只能运行一个网络浏览器。洛博夫斯基和克拉诺设计了一种极简的薄客户机，这种电脑在“数制”工坊中用价值几美元的部件就能组装出来。它可以在屏幕上产生影像，可以连接键盘，也可以实现互联网通信。这种电脑可以规模生产，不过由于它是本地制造的，每一台都可以采用不同规格，人们可以随意增加一个本地储存器，调整图像的分辨率，集成辅助技术，或者为了方便携带而增加一个无线通信装置。通过本地制造，制造者在参与本地投资的经济活动的同时，也获取了技能。



图1－8 “数制”工坊制作的用于生产部件与食品的机械。纳迪娅·皮克、吉拉姆·泰西摄

用“数制”制造“数制”

一个“数制”工坊的技术性目标是制造另一个“数制”工坊。小型精密桌面铣床是如今“数制”工坊最常用的工具之一，也是我用得最多的工具。2011年，在上了我定期教授的机械制造课程“如何制造用来智造（几乎）万物的工具”之后，纳迪娅·皮克和乔纳森·沃德开发出一款可在“数制”工坊制成的机械：MTM咬合机（MTM的意思是用于制造的机械）。她们用一台大型铣床切割高密度聚乙烯——一种高强度的、可回收的塑料。整个机械不是用紧固件连接起来的，而是通过嵌入缓冲接头使部件相互咬合。驱动装置使用的是规模生产的定制电机，利用滚轴丝杠把机械安装起来。这台机器还能切割出自身使用的电机控制电路板。纳迪娅和乔纳森的设计思路，Handibot和其他机械公司的商业产品都有借鉴，也对我们在第5章探讨个人制造路线图有很大的启发。

同样的技能在制造其他多种物品的机械中屡见不鲜，比如食品制造机械。食品生产并不需要分子生物学上的突破，只要对几千年前的配方进行简单的优化集成，就能取得进展。吉拉姆·泰西在瓦尔道拉（巴塞罗那市区外的山区里）的“数制”工坊工作，重点研究可持续化生产。他在2016年智造学院的项目中开发了一种养耕共生的系统。

养耕共生是基于鱼类养殖与植物水培之间的共生关系的一种新型复合耕作系统，利用鱼类产生的废物为作物施肥，同时利用作物为鱼类过滤水质。这种系统比单纯把种子播到田地里高效得多。利用养耕共生，人们可以精准地控制农业投入，在垂直方向上而非水平方向上拓展系统，使其在城市里得到广泛应用。不过系统要求人们为作物和鱼配备棚舍，在两者之间实现水的垂直循环，还要求人们在控制系统中配备传感器、加热器和光源。

整套养耕共生设备可以在市面上买到。不过，通过在“数制”工坊制作，泰西得以将其按照自己的意愿进行定制，并且节省了远程购置的开销。他的系统已经被开发为“水养拓展”项目，他与同事已在

“数制”工坊种植了莴苣、芹菜、豆类、西蓝花、花椰菜、草莓、薄荷、罗勒（紫苏）和香菜。

启示：创制物品的红利

有人会问“数制”工坊能生产些什么，接着就会问谁会为这些东西买单。尽管这个问题的答案，通常都被认作进一步推广“数制”工坊的障碍，但其中也隐含了工坊可能形成经济冲击的重大机遇，而且机遇可能大于挑战。

比特和原子中心部署的首批“数制”工坊采用的是同样的财务模式，与安德鲁·卡内基运营其图书馆的模式如出一辙。我们捐赠设备、进行安装，当地提供空间与管理者，一方面可以扩大初始投资效用，另一方面有助于树立主人翁责任感。之后，“数制”工坊已经向多种投资主体开放，包括公共部门、私营企业、慈善机构、营利机构，不过它们通常都要负责工坊的运营成本。



图1－9 在霍隆“数制”工坊合作社区。欧哈德·梅雨思摄

承担运营成本的一条明路显然是出售工坊制作的物品。对一个商业化产品而言，考虑到设计、生产、组装、运输、销售等所有环节的总开销，其售价通常是材料成本的5倍。如果这些环节都能在同一时间、本地的同一地点实现，其成本就可以大幅削减。由于存在削减成本和定制化的可能，我们试着把成功企业家与“数制”工坊投资者撮合起来，围绕造出的原型进行商业运作。这些努力一直都很失败。企业家觉得投资者不听话，而投资者觉得企业家不靠谱。让“数制”工坊倾尽所有来支撑一个企业也不现实，虽然有些工坊的确找到了途径将其设计进行规模化生产，但这违背了“数制”工坊使经济活动回归个人的初衷。

问题在于这种方式是在重蹈上一场战争的覆辙。想想谷歌是怎样耗时多年才得以确立今天的模式的。答案是放弃搜索引擎，出售搜索红利（通过定向广告）。同样，“数制”工坊运营模式最值得玩味的地方恰恰是不出售创制的物品，出售创制物品的红利。比如，布莱尔·埃文斯在底特律的工坊从事挽救危机青少年（包括早孕青少年和服刑中的少年犯）的工作。他的很大一部分资金，都是通过证明将这些孩子吸引到“数制”工坊的投资比花费在当前社会服务的其他开销的实际效果更好而得来的。他的产品使人重生。

创建这些模式，创新思维与技术支撑同样重要。不过，这些都无法在商学院的传统课程中进行讲授。成功的“数制”工坊已经能综合利用社区开放时间、会员限定时间、授课、辅助创业和基础设施建设等渠道来融资。这一融资方式已经成为“数制”工坊的特色，而非缺陷，因为单独依赖任何一条融资渠道，都会限制工坊的发展，而将它们综合起来却可以提高效率。将单个工坊融入一个更大的网络，形成网络效应，是这些融资模式的共同驱动力。单个工坊并不具备提供所有活动的能力。

“数制”工坊的融资模式可以分为营利性和非营利性两种，但是更为宏观的核心思路是一种可以称为“后营利”的模式。全球化与自

动化导致的失业以及经济竞争对底层人民产生的危害，使社会变革与政治变革在世界范围内赢得广泛支持。而围绕各种财政模式与社会模式进行的辩论，所隐含的假设条件是人们对工作性质的理解。对很多人来说，这意味着迫于无奈背井离乡去打工，使用陌生人的设计，为陌生人生产产品，或者赚钱养家糊口、改善生活。我们怎能置之不理，只为自己制造呢？

现实中的经济因素建立在隐含的技术性假设之上，这是有史为鉴的。长久以来，石油峰值作为石油产量下降的预期时点，被视为一次迫在眉睫的危机。不过，令人意想不到的是，由于可再生能源越来越经济而非由于石油资源的稀缺，石油峰值的来临貌似快于预期。那么，如果我们现在面临一次财富峰值，国家满足人民群众需求的能力不再由其经济产出来衡量，又会怎样呢？

比森特·瓜利亚尔特依托加泰罗尼亚高级建筑研究院创建了巴塞罗那首家“数制”工坊。凭借消费者将成为创造者这一愿景，他于2011年被任命为该市的首席建筑师。当时巴塞罗那的经济形势异常严峻，青年人的失业率超过50%——整整一代人工作就业、独立生活的希望都很渺茫。即便如此，货轮依然不断靠港输入外来商品，垃圾车依然满载废品驶往垃圾场。瓜利亚尔特说，眼前的这座城市就是一个把商品变为垃圾的地方。他的目标是实现比特的自由传输，让原子留在城市里，他希望这座城市能够实现从输入商品输出垃圾，向输入数据输出数据的转变。为了实现这一目标，巴塞罗那在市内建设了多个“数制”工坊，将其视作城市基建的一部分，把向市民提供制造途径放到与净水和电力供应同等重要的位置。对于这一政策，人们的想法并不统一。在城郊一个贫困的移民区奎塔墨里迪纳，人们发起了一次抗议。抗议者占领了拟修建“数制”工坊的地点，要求在那儿建一个食品供应站。后来这一分歧得到了解决，抗议者逐渐认识到他们可以利用社区的“数制”工坊生产食物，为孩子们制造玩具而无须再去购买，自主创业而不用到处找工作。

2014年，巴塞罗那主办了“数制”工坊网络的第10次聚会（FAB 10）。在那次活动中，时任市长哈维尔·特里亚斯按下了该市“40年自给自足”的倒计时按钮。他们按照数字制造研究路线图的路径出发，从全市第一个“数制”工坊开始，发展到10个区各有一个，并持续努力，直到实现全面自给自足。这一目标可能会被视作向古代城邦的回归，但这并非出自加泰罗尼亚分离主义，而是因为无论马德里还是布鲁塞尔都对解决巴塞罗那的问题无能为力，巴塞罗那人民只能自力更生，寻求解决方案。当下情况与古代城邦不同的是，人们不是在寻求独立，而是在追求融入全球网络，成为其中的一部分。

2013年，美国众议员比尔·福斯特首次向国会递交了立法提案，内容与巴塞罗那的计划相似，但这是国家层面的立法。福斯特背景不俗，先开创了一家生产计算机控制舞台照明设备的企业，又作为一名物理学家领衔设计了费米实验室大型粒子加速器的主要部件，之后成为国会中屈指可数的科学家。他提出的美国《国家“数制”工坊网络法案》（*National Fab Lab Network Act*）不是一个拨款案，而是一个从国家利益出发建设“数制”工坊全国性网络的章程。美国国家“数制”工坊计划不是要形成大多数人难以企及的资源孤岛，而是要通过向人民普及工坊以弥补现有实验室的不足。在这个政治极端两极分化的时代，福斯特提案的联署人来自城市与乡村，既有共和党人又有民主党人，使其成为当今少有的跨越党派政见的提案。国会并未对该法案进行投票（美国国会的运转效率还谈不上登峰造极）。不过，提案得以再次提交，而且引起了私营部门的兴趣，它们针对提案目标做出了一些承诺。雪佛龙公司首先向智造基金会认捐1 000万美元，在其设厂的社区内开设“数制”工坊。

比特和原子中心还有一个流动的“数制”工坊，2014年首届创客嘉年华时我们在白宫进行了展示。它就停在总统办公室外面，那个地方平时连佩戴白宫徽章的人都禁止通行。安保人员看到工坊内配有大功率的激光器和机床，十分戒备，不过作为一名资深的社区活动家，奥巴马总统却很欣赏它。表面上看，这是个创客运动的庆祝活动，不

过媒体拍照时段的安排却意味深长：活动凸显出新的工作机会不太可能回归传统工厂了。正如小型机生产商把个人电脑视为玩具却被其颠覆一样，个人制造可能瓦解规模生产。

2015年的FABⅡ是“数制”工坊第二次年度聚会，来自78个国家的与会者齐聚麻省理工学院校园。这次活动中，波士顿、萨默维尔和坎布里奇的市领导也加入了巴塞罗那宣言的行列。他们并没有制订各自的时间表，而是采用了巴塞罗那的倒计时。此后，很多其他城市也加入了这个称为“智造城市宣言”的活动，现在的活动负责人是托马斯·迪兹。

智慧城市是指将城市的一切都物联化以实现实时响应，现在已成为城市规划的流行趋势。“智造城市”则自然而然地成为下一步发展目标，跨越数字与物理的界限，实现消费品的可持续生产与循环利用。这种转变并非一蹴而就。相反，它是一个持续演进的过程，或者更准确地说，是一种由伴随新能力产生的间断平衡[10]带来的进化。进化的早期步骤包括制造都市无线数据与传感器网络，以及装备公共空间。与“数制”工坊一样，目前还没有一个城市具备独立实现这些目标的技术能力，只能多个参与城市联合，从把城市的输入与输出进行量化开始跟踪其进展，而目前的量化水平还很低。

智造城市宣言目前正在引发一些国家的行动。2014年，我出差去不丹，筹划开设“数制”工坊，并拜访了思想深刻的策林·托杰首相。他谈到了电饭煲，这使不丹加入智造行列的机遇变得具体、明确起来。众所周知，不丹经济是以国民幸福指数而非国民生产总值为核心的。但这并不意味着人人都能感到幸福，而是不丹人民更加重视生活品质而非能买多少东西。然而，不丹人民的这一关注点并未涵盖供应来源，他们通常还要越过不印边境车载进口日常用品。电饭煲对不丹家庭生活至关重要，大多从日本进口。制作一个电饭煲的容器、一个温度传感器、一个加热元件，以及一个控制系统在“数制”工坊是

个简单项目。伴随着对电饭煲成品的进口替代，本地制造使每个电饭煲都可以被定制，以适应家庭规模与厨房的设计。

2016年，我曾与卢旺达贸工部长方济各·坎尼姆巴有过类似的交谈，当时我在那里筹建“数制”工坊。他是卢旺达经济改革的主要设计师之一，推动该国成为撒哈拉以南非洲发展最快的经济体。谈话中，他最关心的是卢旺达数额巨大且不断增长的贸易逆差。卢旺达人试图用进口替代来解决这一问题，主要依靠吸引跨国公司投资设厂进行生产。直到基加利首家“数制”工坊开业，卢旺达才实现对生产资料而非最终产品的分配。非洲在很大程度上跨越了有线电话时代，直接进入无线通信时代，同样，它也能跨越工业革命的历史阶段，直接进行分布式生产。

2016年，“数制”工坊网络聚会FAB 12由深圳市主办。深圳是中国规模生产生态系统的心脏，令人玩味的是，我们在那里举办的原因，是中国欢迎“数制”工坊和创客运动，而这些运动也在拥抱深圳。

深圳的华强北路是我最喜欢的地方之一。那里有巨大的市场，你可以在林立的小柜台上买到单个电子器件或一袋器件，可以在楼下的小店买到一盒器件，可以在路边的仓库买到一卡车器件，也可以在附近的乡镇企业买到一集装箱器件。这个市场以所谓的“山寨产品”闻名，我觉得这些产品的生产技术大概相当于快速眼动睡眠，就像梦是不相关的体验与幻想混合的产物一样，这些产品也是技术混搭的产物。对生产者来说，他们为巨量出口订单而生产的产品就相当于睡眠。我上次买了一个貌似苹果手表的产品，但这个产品与真正的苹果手表有两点不同：一是价格只有25美元，而不是正品的250美元；二是它有一个苹果手表遗漏了的SIM卡插槽的设计，这只表不是手机的外围设备，它就是一个手机。

如果将来巴塞罗那或者不丹能自力更生制造自己的消费品，它们的人民将不再需要深圳为他们生产。深圳的领导能够预见，规模化生

产无往不利的时代已接近尾声。然而，今后的许多年里，这一设想的基础将很难在本地实现。这些物品不是最终产品，甚至不是制造最终产品的机器，而是这些机器的部件。深圳是规模生产产品的巨大中枢，拿纳迪娅和乔纳森安装机器使用的集成丝杠，或者将设备与内置互联网协议的无线网络连接在一起的嵌入式无线电来说，由于这些要素需要巨大的资本投入来生产，具有一定的规模，无须定制，因此进行量产是有意义的。由此，深圳迅速将多重功能集成于单个物品，以简化后面的工程设计。

这与计算机产业发展史上个人电脑的横空出世和小型机产业的整体消亡类似。在第三次数字革命中，从商业角度来看，我们可以把小型机比作中型企业。不过，借着云计算大数据中心这面大旗，大型主机已经比以往更加重要。对多数人来讲，计算机的前端是智能手机、平板电脑或个人电脑，但是它们依靠的是后端的云。“数制”工坊并不替代规模生产，而是将其延伸。

思考一下软件或音乐的境遇吧。曾几何时，软件是由微软和IBM（国际商业机器公司）这种巨头编写的。然后，开源软件不期而至，使任何人都有了参与编写的可能。今天我们拥有了应用软件平台，一个应用软件可以由1个人、10个人、100人、1 000人或100万人来编写。大型专有软件包仍然存在，不过可以说，这已成为软件开发中最无趣的一部分，因为它们必须瞄准共性需求。目前实现开源的软件市场在从前是难以存活的。同样，音乐曾经一度由唱片公司销售。接着，文件共享平台出现了，一夜之间没人再花钱买音乐了。现在有了买卖音乐和歌曲的交易平台，其市场同样也会有1个、10个、100个、1 000个或100万个买家。在第一个例子中，一串数字变成一个程序。在第二个例子里，一串数字变成一段音频。而现在，一串数字可以变成一件物品。只要人们的需求还有一致性，规模生产就会存在，而在规模生产与自己制作之间，存在一个新的制造规模层级，它是开源的，而且在从前是难以存活的。

在前两次数字革命中，人们曾希望小内容提供商通过长尾效应推动新经济并从中获益。事实却截然相反。大多数钱都被少数几个拥有平台的巨擘赚走了，这些公司包括谷歌、脸书、苹果和亚马逊。在第三次数字革命中，一些公司已经在试图开创类似的市场份额，垄断自由分享和付费产品设计的市场。其中有些做法很吃香，不过距离垄断音乐软件和应用软件的程度还有天渊之别。可能是努力得还不够吧。“数制”工坊今天还需要所有能力来制作一系列最终产品，只有一些小塑料件用入门级的3D打印机就能制成。不过，另外一个原因可能是，相较娱乐选择，制作物品更像烹饪。尽管有分享配方的网站，烹饪的商业性行为还是以出售食品与炊具为核心。

让我们来看看平面打印机的发展历程吧。从麻省理工学院分拆出美国数字设备公司（DEC），从而催生出小型机产业。后来美国数字设备公司衰败，被康柏公司收购，然后康柏又江河日下，被惠普收购。惠普得以幸存，其中一个原因是喷墨打印机的经济意义。惠普的喷墨打印部门在俄勒冈州科瓦利斯，因为打印部的工作人员需要躲着身在加利福尼亚州帕洛阿尔托的领导。一群工程师觉得他们可以制造出印制精美的廉价打印机，不过打印速度较慢。那时，打印部门有一项每台打印机每秒打印页数的分级标准，管理层的反馈意见是慢速打印机是一着臭棋。所以这些工程师搬到科瓦利斯的低端计算器部门所在地，在这个地方悄悄研发喷墨打印机。当然现在的观点是，你根本不需要一个桌面大容量打印机，因为你打印的每一页都不相同。可能当时也会有放在楼下的工作组激光打印机，放在地下室的组织使用的行式打印机，放在仓库的城市使用的卷对卷打印机。但是小型喷墨打印机每次打印一页，累积起来就是庞大的卷对卷打印机的产量。没有谁更好，重要的是谁的印刷方式更灵活。在从出版向桌面打印的变迁中的这一分级标准，可以用来类比即将到来的制造变迁。

对能力的衡量可以双向进行。2016年，白宫举办了从创客过渡到制造的活动，主题是如何由原型形成生产。在那次活动中，我的学生纳迪娅·皮克大胆引导与会者逆向思考，即如何从制造走向创客。在

“数制”工坊中，将快速成型过渡到快速成型机的快速成型技术是一种自动化技术，不过这是一种能够灵活适应需求的快速自动化技术，不需要大型固定资产投资。由于担心自动化将取代工人，人们已经把我们（指工人）与他们（指机器人的拥有者）对立起来。但是第三次数字革命的经验是，他们就是我们，对制造的所有权可以像对计算机的所有权一样，广为普及。

把握经济影响最重要的尺度不在于生产能力，而在于观念。2015年，关于麻省理工学院剥离企业的产出报告显示，所有企业年营业收入总额达1.9万亿美元，处于世界第九大经济体（俄罗斯）与第十大经济体（印度）的生产总值之间。这几千人为何能匹敌10亿人的生产力？这其中有两个秘密。

首先，麻省理工学院不是一个以营利为目标的技术孤岛。它是根植于肯德尔广场的一个生态系统——一个长期研究与短期开发并重，小型初创企业与大型公司集团共生，咖啡厅、俱乐部和公园林立的环境。它对哪项活动应当在哪里举行这类事并无定见。

其次，我认为麻省理工学院的核心竞争力在于，它对新人而言是个安全可靠的地方。我有很多同事都与上流社会格格不入。不过很明显，要发明创造，你就得质疑假设，而那不是限定于某个细微生活侧面的。很多业务创新并非始于富豪的梦想。其中一个驱动力在于改变世界的愿景，比如说对可飞行汽车的需求。而一个更大的动力在于找到一个适合你的工作。我有一帮研究早期量子计算机的学生，为了能继续共事，他们开设了一家公司（ThingMagic）。这家公司后来开发了一个参考平台，该平台可以读取商店收银台的无线射频识别（RFID）产品——电子标签。他们最初并没有进军零售业技术的规划，而是偶然发现有一项需求与他们的技能相匹配。

智造城市、城镇与乡村所面临的机遇同样适用于全球。邮编为02139的麻省理工学院成功的秘诀，随着“数制”工坊的普及传播到世界各地。这些工坊作为生态系统中的节点而发生作用，没有地理差

异，却是智力的标识，使那些不适应严苛教育或者企业层级的有创造精神的人找到归属感。这种社会基础设施可能看不见摸不着，不过麻省理工学院剥离企业的例证显示，它比实体基础设施更为重要。世界上无数区域性经济开发项目都斥下巨资，试图打造下一个硅谷。但在第三次数字革命中，下一个经济引擎不是一个地方，而是一个由数字化通信、计算与制造连接起来的网络化集群。

组织：仁慈的“独裁者”

2005年，我发现自己登上了一艘引擎失灵的船，随波漂流擦过冰川，驶离斯瓦尔巴群岛的海岸，停靠于北极边缘。在那里我遇到哈肯·卡尔森——前述在挪威北部的林根执掌“数制”工坊的牧羊人。我们在这个国家最偏远的地方四处了解需求，寻找机遇。引擎终于得以修复，但是冰川迫近、绝处逢生的经历使我们心有余悸，在尝试满足不断增长的“数制”工坊普及需求以及人们投身这场运动的愿望这一过程中，我们的感受如出一辙。我们意识到必须建设比我的“数制”工坊或他的牧场更加有力的组织架构。

问题是个老套的普通问题。要为能够看到的“数制”工坊部件设备筹资并不困难，难的是为运转这些部件设备的基础设施（一个你看不见的基础设施）筹资。费用涵盖“数制”工坊可能需要的技术开发、全球供应链的管理、筹建工坊的团队建设，以及支持工坊的计算能力。这些基础设施所需资金最初都出自我的研究经费，没有任何扩展余地。

我们尝试与慈善基金合作，不过它们缺乏这方面的技术能力。这些管理投资基金的人确实了解投机性技术风险，但他们对资助平民并

不感兴趣。研究基金也不支持特定领域组织。万般无奈之下，我们只好决定创建自己的组织。

我们试了5次才成功。前4次都试图为这些共性支出来筹资，因为当时确实需要。那些尝试都以失败告终。不但因为资助者没有兴趣，而且分散管理的“数制”工坊并不欢迎集中管理的实体机构。在第五次尝试中，我们不出售最终成果，而是出售人人都需要的“数制”工坊服务。因此，“数制”基金会应运而生。

人们普遍错误地认为开源项目都来自扁平化组织。这方面的大型成功案例都是以一名仁慈的“独裁者”为核心的组织架构，比如林纳斯·托瓦兹之于Linux，以及米切尔·贝克之于摩斯拉。这种强领导模式促使每个团队成员都积极投入。人们并非被迫接受这种领导，权威性来自激励人们参与的软实力而非命令和控制。

有关软实力技术性领导的最著名的例子，是互联网架构委员会（IAB）——一个向互联网提供技术指导的机构。委员会有13名成员，他们在网络社区以外都鲜为人知。他们的工作方式从其管理标准的标题［请求评论（RFC）］上看就一目了然。这些标准不是强制性的，但由于能获得接入互联网的好处，几乎每个使用电脑的人都会遵循这些标准。

IAB是互联网工程工作小组的一个委员会，同时是互联网协会的咨询机构。它们都起源于1979年创建的互联网结构控制理事会，旨在为当时还是政府研究项目的互联网架构工作提供咨询。这些组织经历了互联网几十年的指数级增长，寿命远超其前身。

当前，“数制”工坊网络的发展正与这些组织框架的大发展处于同一阶段。相当于“请求评论”的标准包括阐明了“数制”工坊权利和义务的智造宪章、“数制”工坊的设备清单、智造学院的课表，以及“数制”工坊分布列表（在fablabs.io门户网站上）。这些内容都倾注了许多人的心血，不断进化着；任何一项决策对某一特定目标而言都可能不是最优的，但拥有这些统一标准的集体利益会超越决策最

优但碎片化的机会成本。随着“数制”工坊网络组织能力的增长，一些项目合作方正在进行接触，就其无法独立完成的任务寻求与“数制”工坊的协同工作。有一种补充方式是公司通过一种一举两得的投资来履行其社会责任，既支持在公司所在社区设立“数制”工坊，又借此机会识别并培训有前途的技术人才。还有一种方式，即“数制”工坊与援助机构合作，在目标地区建设“数制”工坊群落。

帕特里克·科尔根运营了一个欧盟的机构，以支持北爱尔兰冲突后的和解工作。该组织试图以不同方式对和解工作进行投资，从而催生了德里（Derry）的神经中心（Nerve Centre）与贝尔法斯特的阿什顿中心的两个“数制”工坊。这两个工坊所在地都邻近被婉称为“和平墙”的地方，这个“和平墙”实际上是教派隔离的藩篱。现在，墙两侧的孩子们在“数制”工坊一起工作，跨越了两个教派的历史分歧。在看到工坊产生的影响后，把该项目扩展到爱尔兰其他地区（北方和南方）的想法应运而生。

印度的工程学院可能因为把正规教育学习与培养实践技能分离开来而名声欠佳。比特和原子中心在喀拉拉邦的科钦和特里凡得琅设立了“数制”工坊。目前，喀拉拉邦创业促进部和喀拉拉科技大学正在与智造基金会合作，计划在该邦的工程学院建造150个“数制”工坊，首批20个工坊已在建设之中。智造学院的毕业生接棒初期的麻省理工学院学生，投入设备安装和培训工作，为该项目提供能力支持。

埃及的首个“数制”工坊是由迪娜·埃尔－赞法利和几名共同创立者于2011年创建的，她之前在麻省理工学院进行研究生学习期间参加了“如何智造（几乎）万物”课程。2011年埃及骚乱期间，她致电工坊询问大家是否安全。工坊接电话的人哈哈大笑说，那段时间是他们最忙碌的日子，因为聪明好学、不落窠臼的青年人无意参与教派冲突，纷纷借着骚乱不上班上课，来工坊工作。他们的工坊已经与埃及教育部、美国国际开发署和智造基金会合作，在全埃及实施STEM（科学、技术、工程和数学）教育的高中设立“数制”工坊，其中包括车

载流动“数制”工坊（FLOW）。去年，这家工坊组织了开罗的创客嘉年华，一天的活动就有上万名来自社区的参与者。

贝诺·华雷斯在2009年参加了智造学院的学习，随后创建了秘鲁的首家“数制”工坊，这是一个巴塞罗那“数制”工坊与西班牙援助部的合作项目。西班牙方面并不直接出钱，而是帮他们识别、培训秘鲁创新者并将这些创新者送回国，还附赠一个“数制”工坊。这所工坊颠覆了秘鲁僵化的教育制度。当华雷斯所在大学开始宣布“数制”工坊的使用规则时，华雷斯与同伴争辩说工坊是他们的，而非学校的，如果不能保证工坊是人人可获取的资源，他们将会搬离学校。华雷斯在亚马孙地区长大，在那里他发现，人们只有三种职业选择，即农民、士兵和恐怖分子，而他的选择是“以上均不选”。现在他领导着一个位于亚马孙河的流动“数制”工坊项目，以帮助本土社区实现可持续生产。

我们在联合国开设了一个快闪式“数制”工坊，旨在发起可持续发展目标（SDGs）倡议，那是一次史无前例的国家首脑聚会。这些目标来源于一个遵循了富有白人发展目标的包容性进程。可持续发展目标包括获取卫生保健、教育、清洁能源与水等资源，以及消除贫穷与饥饿。与会的外交官也许会面带揶揄地看着我们，盘算我们为何会出现在那里。之后，当他们意识到多数可持续发展目标都需要具备从比特到原子的转换能力，以实现卫生保健传感器、水过滤器等物品的本地制造时，一个令人顿悟的时刻就会降临。

我们还在达沃斯世界经济论坛上开设了快闪式“数制”工坊，目标是向与会政治与商业领袖展示数字化制造，而非简单介绍。在一次展示活动中，我们与当时负责联合国人道主义救援的领导不期而遇，显然她因为活动干扰了关于提供基建、教育机会、商业孵化和难民营娱乐等内容的重大讨论而火冒三丈。随后，当她意识到数字制造技术的获取是她那些议题的共同要素（一个当权者没有考虑的共性问题）时，怒气就烟消云散了。

这次偶遇促成了全球人权工坊的诞生，由戴维·奥特和奥利维尔·德拉鲁负责运营，他们分别来自红十字会国际委员会和联合国难民事务高级专员办事处。他们的组织实际上都处于这个星球上最灾难深重的前线。奥特和德拉鲁找到我们讨论各种可能性，比如传输比特而非原子以实现快速反应，当地制造义肢和庇护所等所需物资，等等。当他们考察在建的专用工坊时，他们意识到可以通过覆盖整个“数制”工坊网络的资源来实现其目标，这与跨互联网提供的系列服务并无二致。他们可以利用“数制”工坊，使其成为人道主义援助的实际平台。

在这种全球领袖的集会中，我发现政治家们都忧心忡忡。他们都在为失业、不平等、移民这些棘手问题，以及两极分化和矛盾冲突带来的冲击而殚精竭虑，而传统上用来解决这些问题的手段（如调整货币政策）却收效甚微。

相反，如前所述，那些在一线引领第三次数字革命的地方领导聚会起来却常常是谈笑风生。他们正在对所在社区产生实质性影响，社区未来发展空间广阔，而且关键是，他们从中获得了乐趣。

从自上而下与自下而上两种领导方式的现实差距中，我们深感未来的领袖将不会通过传统组织形式产生。“数制”工坊的当前技术有种“功成不必在我”的倾向，我会在第5章中继续讨论。如同互联网是在小型机上发明出来的一样，得以延续的不是目前的具体技术形态，而是随着发明创造而不断进化的生态系统。与首批互联网网站出现时一样，在“数制”工坊数量未过千时，单纯的数字增长就很激动人心。而现在重要的不是它们的存在，而是它们为何存在。如果人人都能在任何地方制作任何物品，我们的生活、学习、工作、娱乐将会走向何方？

（本章由尼尔撰写）

[1] Fab lab即微观装配实验室（Fabrication Laboratory），是美国麻省理工学院比特和原子研究中心建立的一个新型实验室项目，这个实验室就像一个“几乎可以制造任何产品和工具”的小型工厂。该中心主任格申斐尔德教授认为，迄今为止，数字革命的发展已经经历了两个重要的阶段，分别是通信和计算。以移动技术为代表的普适计算通过上述两个阶段的发展，已经形成无所不在的个人通信网络及个人计算系统。目前，人类正处于第三次数字革命的前夕，在这次以“个人制造”为核心的革命中，相关的材料技术和信息技术已经露出苗头。从某种意义上说，fab lab正是这即将到来的革命大潮前跃起的浪花。——译者注

[2] Arduino是一款便捷灵活、方便上手的开源电子原型平台。包含硬件（各种型号的Arduino板）和软件（Arduino IDE）。由一个欧洲开发团队于2005年冬季开发。——译者注

[3] DIY是“Do It Yourself”的英文缩写。最初兴起于电脑的拼装，逐渐演绎成一种流行生活方式。简单来说，DIY就是自己动手制作，没有专业资质的限制，想做就做，每个人都可以DIY一份私人定制的、表达自我的“产品”来。——译者注

[4] DIT是“Do It Together”的缩写，也将应用到更多领域。DIT不同于DIY，它不是个体的行为，它强调的是团体合作以及共同努力。——译者注

[5] DIWO是“Do It With Others”的缩写，与DIT相似，在信息化的时代，没有人能学会一切软件，因此也就需要合作，需要互相帮助以达到同一个目的，完成同一项任务。——译者注

[6] 此处的剑桥是指美国马萨诸塞州波士顿都市区的剑桥市，麻省理工学院所在地。——译者注

[7] 1英尺≈0.304 8米，1平方英尺≈0.092 9平方米。——编者注

[8] Scratch软件是一款由麻省理工学院设计开发的少儿编程工具。其特点是使用者可以不认识英文单词，也可以不会使用键盘，因为构成程序的命令和参数通过积木形状的模块来实现。——译者注

[9] 鲍勃·迪伦，1941年5月24日出生于美国明尼苏达州，美国摇滚、民谣艺术家，美国艺术文学院荣誉成员，2016年获得诺贝尔文学奖，成为第一位获得该奖项的作曲家。——译者注

[10] “间断平衡”理论是一种关于生物进化模式的学说，对古生物化石缺少中间演化类型进行了新的、系统的解释。“间断平衡”理论描述了一个系谱长期静止或平衡的状态被短期的、爆发性的大进化打破，同时伴随着大量新物种的产生。在此，作者借用间断平衡概念来表明，新能力的突破打破了城市演进的既有节奏，实现了飞跃性的质的变化。—— 译者注

第2章

第三次数字革命

造访一个欣欣向荣的“数制”工坊，任何人都会立即被其间弥漫的无尽能量、快乐与满足感染，那是一种由数字制造技术设计和制作物品带来的感受。尼尔提到的，以及我们即将介绍的那些早期智造先驱正在开创高度分布的个人制造新模式，这可能会深刻地改变我们的生活、学习、工作和娱乐的方式。

尼尔重点强调了社区“数制”工坊里的人是如何日益增强能力，制造食品、家具、工艺品、计算机等几乎一切物品的。这是一个先兆，预示第三次数字革命将会赋能于个人与社区，实现全球互联和当地的自给自足，尤其是在技术持续指数级发展的情况下。不过，尼尔只讲了事情的一个方面。“数制”工坊并非都欣欣向荣，“数制”工坊中的人并非都快乐满足。数字制造是一个复杂的过程，包含了多重相互依存、不断发展的技术和能力。虽然人们有越来越多地制造消费品的潜力，但目前只有极少数人能真正做到这一点。在今天的“数制”工坊里，有不少原材料还是不可再生的。数字制造推动社会变革的潜能与实现这一变革的现实之间还有巨大差距。

尼尔是个经典的技术乐天派。他的乐观主义根植于对基础科学的深刻理解、与早期智造先驱的频繁互动，以及对路线图的研究。但他经常大大低估社会因素对技术发展步伐的重要限制作用，低估技术最终对社会产生的影响（也有可能没有产生影响）。在本章中，我们采用当前智造生态系统的新视角，为技术图景提供社会平衡。理解当前的技术能力与智造生态系统的成功案例固然重要，但我们也必须了解渗透在生态系统中的挑战与矛盾。客观看待“数制”工坊的成败得失，理性分析经验与教训，对实现第三次数字革命的愿景至关重要。

前两次数字革命形成了海量财富，引起了重大社会变革，但这个星球上仍有很多人在两次革命中落伍。在戈登·摩尔发表论文之后的半个多世纪里，人与人之间仍然存在巨大的数字鸿沟。数字技术对地球上近半数人来说还遥不可及。在地球上的很多地方，收入与财富的不均衡、新技术造成的失业，以及数字回音室效应正在使社会产生着

严重分化。很多人都在为“永远在线”的生活而奋斗，这种生活受数字技术的影响越来越深。

第三次数字革命可以帮助我们解决这些社会挑战，也可以使挑衅更加严峻。“数制”工坊衍生技术的发展速度及其对社会的影响，无论好坏，都不会被一些看不见的手操控。随着技术被引入社会，进步将取决于我们的决策和价值观，无论是个人的还是集体的。技术快速发展的最佳时机是在早期，那时，研究重点正在显现，构想正在形成，而且支撑组织机构的生态系统正在形成。现在到了第三次数字革命发生的时候了。

为了写这本书，我们进行了不少研究，造访了世界各地的“数制”工坊与其他创客空间，采访了几十位智造先驱与创客先驱，还调查了几百位相关人士。走遍这些社区，我们发现，这其中既有令人喜悦的成功，又有令人惋惜的挫折，我们对数字制造地未来既深表乐观又深深担忧。在第1章中，尼尔强调了乐观与可喜的一面。然而，我们知道技术发展的潜力不一定会被发挥出来。我们必须对挑战有清醒的认识，拓展思路，提出方法，以进行有效应对。

这绝非易事。2016年年初，美国国防高级研究计划局（DARPA）宣布了一项新计划，以应对“下一代社会科学”。项目主管亚当·罗素强调，社会科学已经“由于其主题——人类的复杂多变性和不可预知性——而受到本质的挑战”。然后他又说：“在物理学家之间有个玩笑，如果原子或电子具有人性，那他们的研究领域将会变得多麻烦，不过这正是社会学家面临的问题。”数字技术在传播过程中发现并纠正错误，人们也试图发现并纠正错误，但这个过程要更加复杂、混乱。在本书中，我们不仅要看到技术，还要看到第三次数字革命中人性扑朔迷离的一面。

即使被技术能力的发展速度碾压，人性的基本轮廓也不太可能在短时间内改变。好消息是我们将坚持梦想，通过个人努力和集体协作创造并改造我们的世界。坏消息是，这一梦想、创造和改造未必能帮

助我们挖掘更好的自我。从社会维度关注第三次数字革命，需要考察关于人性的基本假设和个人、组织与机构适应快速变化的能力。改造比特和原子可能比改造人和社会更容易，不过它们都是紧密交织在一起的。

“如何（几乎）智造万物”由尼尔原创的课程戏谑而来。如果学生的项目失败了，他们就开玩笑说，这门课感觉更像“如何（差点儿）智造万物”。这一感受常常波及更为广泛的智造社区，反映了制造的民主化面临的严峻现实挑战。在本章中，我们会探讨一些基本挑战，包括智造获取、智造素养、赋能的智造生态系统的培育，以及技术传播过程中的风险的降低等问题。在第4章中，我们会梳理历史脉络，研究社会体系对新技术的反应，提出具有前瞻性的建议。在第6章中，我们将总结智造未来的宏伟愿景，那时的社会与技术体系将服务于创造更加自给自足、互联互通、可持续发展的社会，同时，我们会为应对挑战、实现远大抱负给出具体指导。

智造获取

截至2017年，全世界有大约1 000个“数制”工坊，共惠及几万人，而地球上大概有70亿人。尼尔已经阐述了“数制”工坊如何改变人们的生活，但真正的风险是只有少数幸运的人能从中获益。在我们的采访中，大家普遍担忧，如果不优先考虑数字制造技术的广泛普及，就会出现“智造鸿沟”的潜在危险，就会有太多的人被时代抛弃。因此，智造的获取是首要的基本挑战。

每次技术突破都会形成新的社会分化。前两次数字革命在世界范围内创造了海量的财富和深刻的变革。但是，并非人人都从中获得了益处。电脑或互联网对世界上数十亿人来说仍遥不可及，人们缺乏参

与第三次数字革命最基本的数字化基础设施。互联网的发展也是一把双刃剑，一方面，它是第三次数字革命中展现指数级变化潜力的范例；另一方面，鉴于数字鸿沟的泛滥，我们也必须对它保持谨慎。

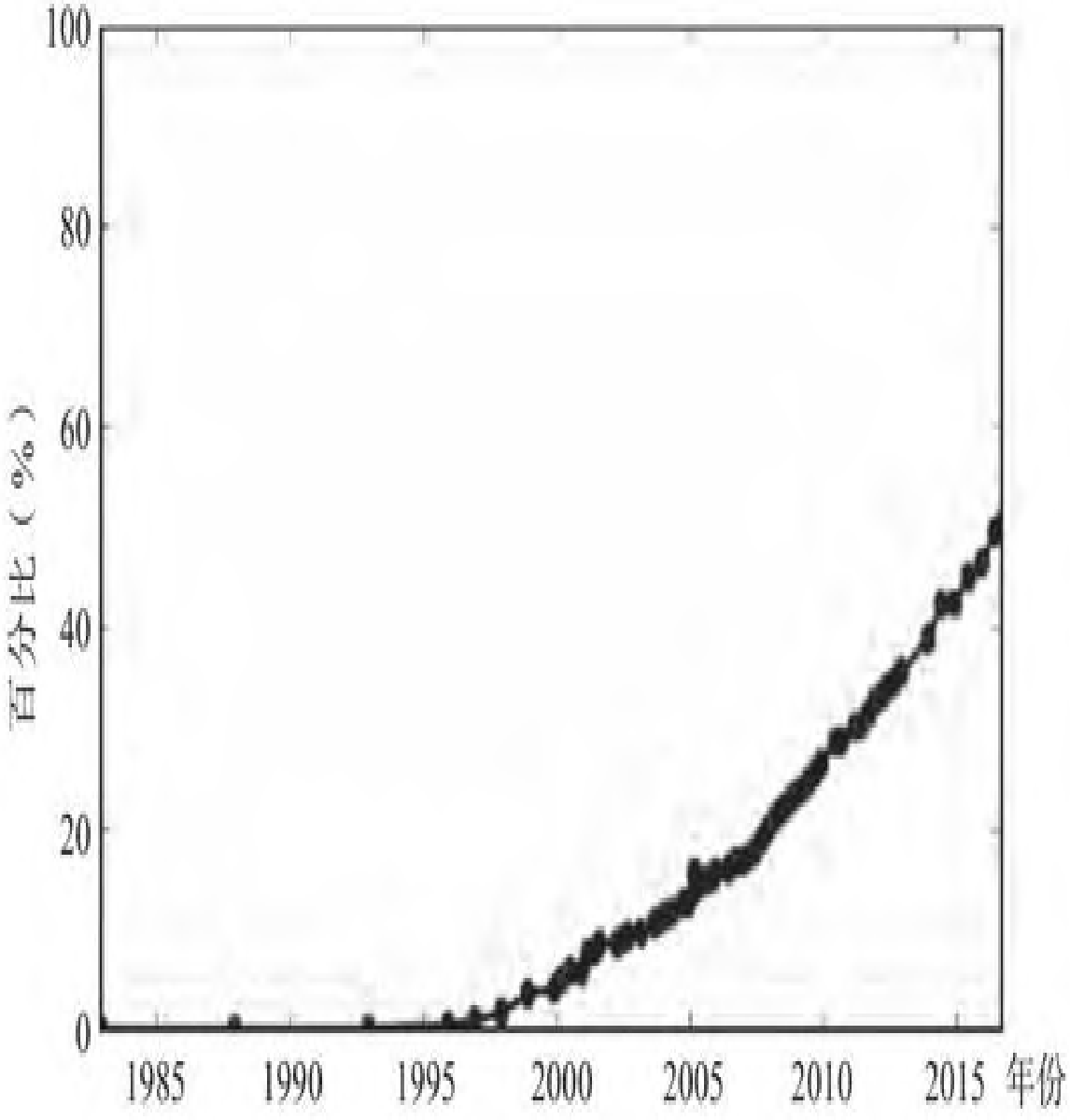


图2－1 互联网用户占世界人口的百分比。互联网世界统计中心、国际数据公司、英国国家承销机构有限公司提供

1983年，传输控制协议/互联网协议（TCP/IP）成就了计算机网络发展中的互操作性。1989年，全世界只有5万个互联网用户——多数是学术用户和军事用户，也是在这一年，美国国防部出资研发的阿帕网对公众开放。互联网接入的历史轨迹呈现了如下数据特点：最初应者寥寥，然后就爆发式增长，到今天，世界上已有半数人能够以某种形式进行互联网接入。

这一数据的增长速度发人深省。不过，尽管技术性能是遵循摩尔定律，呈指数级增长的，我们还是发现互联网的发展速率基本上是线性的。互联网的底层技术其实早就足以支撑全球人口100%接入互联网了。个人、组织和机构的作为（或不作为）交织在一起，既解释了互联网为什么能产生令人印象深刻的影响，又说明了互联网无法产生进一步的影响的原因。如今，政府、基金会与社会企业家、营利性公司纷纷寻求创新性解决方案，旨在实现全面接入互联网的各类计划不断出现。如果早在1983年就开始增加对这些人力、财力、能源与创新的投入，那么今天的世界会是怎样的呢？

而且，“互联网接入”不是一个非此即彼的概念，不能简单地用“有”或“没有”来划分。从全球来看，计算机与互联网接入的质量与可靠性参差不齐：有几十亿人只获取了移动设备接入，连通性不稳定，或者获取了分级接入，速度更快的层级需要付费获取。我们可以先试着在翻盖式移动电话或低端智能手机上用多媒体编写、分享一个文件，再想象一下如何在同一部手机上运行CAD（计算机辅助设计）和CAM（计算机辅助制造）工作流程。

互联网的接入与质量问题不只发生在发展中国家。芝麻街工作室的琼安·冈兹·库尼中心最近发布了关于美国数字鸿沟的报告，对数字包容性问题进行了开创性研究。在美国，有1/3的贫困人口依赖移动设备上网，许多人由于当地基础设施薄弱或无力支付规定月费而时常被中断联网服务。数字接入对经济秩序中的低端人群而言，是一个真正的持续性挑战。考虑到数字技术已经渗透到几乎所有工作（以及此

外的大部分人类活动）中，那些数字鸿沟无疑已成为全球财富与收入不均持续恶化的主要推手。

我们在2017年年初调查了超过175位智造领袖，对他们而言，在数字制造的世界里，稳定可靠的互联网接入被视为一个明确的基本挑战。80.4%的受访者认为互联网接入非常重要。同时，2/3的人表示仍然很难接入。重要性与难度之间的矛盾，反映了一个制约数字制造指数级增长的疼痛阈值。为了更好地理解这些反馈，我们使用乔尔与同事开发的数据可视化方式，来观察涉众之间的一致性与偏差。

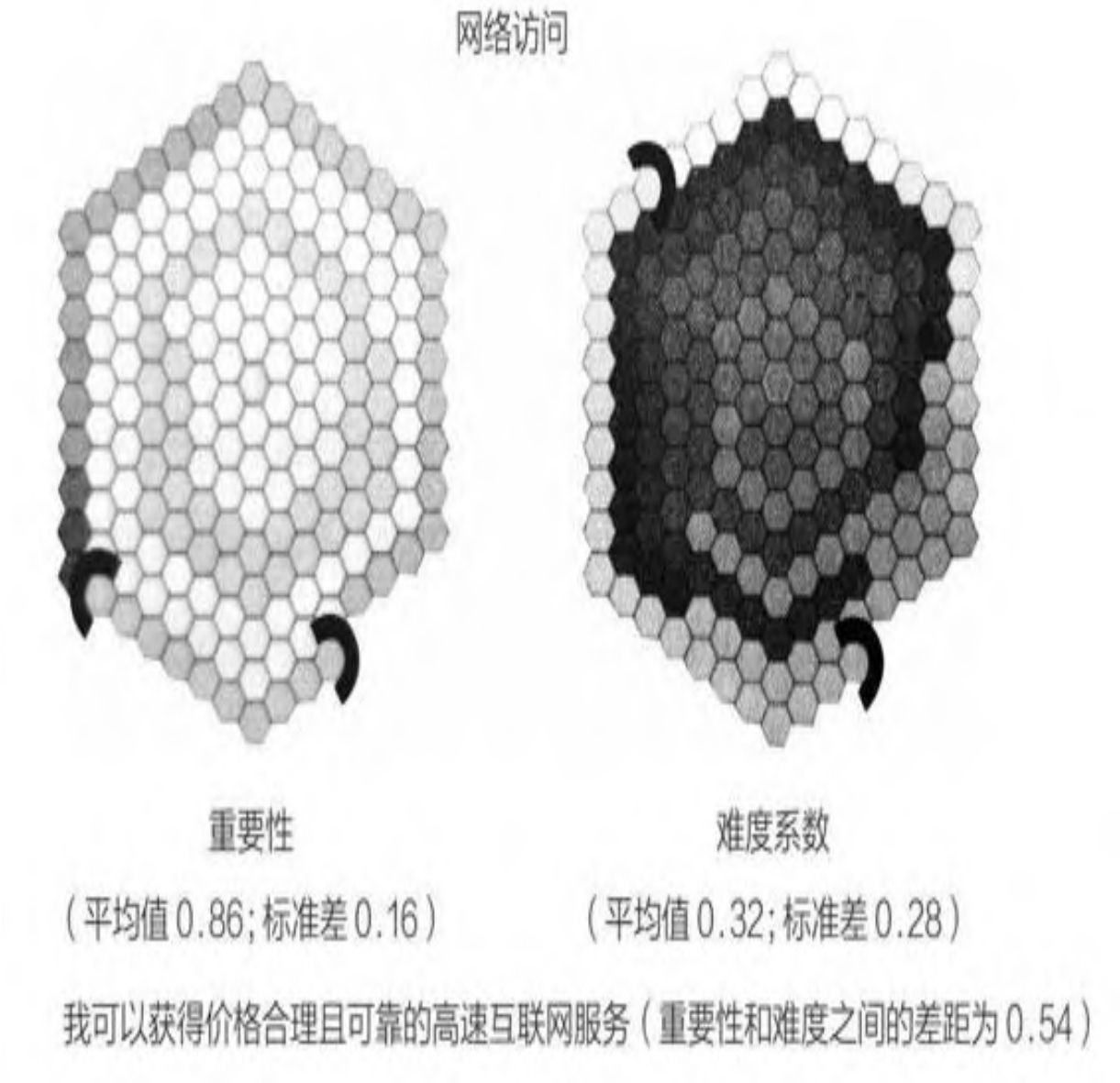


图2－2 访问互联网的重要性和难度。WayMark Analytics公司绘

在图2－2中，每位受访者都被分配到一个小六边形，颜色代表其观点。代表观点的颜色一般为红色、黄色和绿色，分别表示受访者对问题持否定、中性与肯定态度的程度。因为插图是黑白的，所以我们用较深的阴影代表红色，中等阴影代表黄色，较浅的阴影代表绿色。小六边形呈螺旋状排列，从每个图形的中间开始，与最接近平均值的

小六边形排列在一起，然后围绕平均值上下交替分布，最终，离平均值最远的小六边形（其实它代表了异常值）排列在图形外部。中间是核心趋势。括号中的小六边形表示“不适用”的数据。关于互联网接入，对比很明显——几乎所有的受访者都说互联网接入很重要，几乎同样多的人认为接入很难获取。

第三次数字革命并不仅仅依赖于前两次数字革命的技术性基础设施。它还引入了新技术，这些新技术很可能会使当前的矛盾更加深化。因为数字制造需要持续发展的硬件、软件和消耗品（制造过程中使用的原材料）以及空间、计算机、网络连接和合格的员工，所以我们在智造获取方面面临的挑战，甚至比前两次数字革命的挑战更大。

我们对数字制造领导者需要的软件、硬件以及软硬件集成的获取进行了问卷调查。同样，结果也有很大的差距。在“数制”工坊，设备获取非常重要，却很难实现。软件的情况也大同小异。如同后面指出的那样，在软件和设备的集成方面，情况只是稍微好一些。

从反馈的结果来看，对于软件、设备和二者的接口的获取难度各异，其中软件的获取难度最大。当被问及设备获取的难度时，反馈的数据呈双峰状，超过1/3的人表示设备的获取非常困难，而另外1/3的人则表示很容易获取设备。这组数据反映的涉众取向分化暗示了一个挑战。如果不顾最后1/3而只提升前1/3的水平，我们所面临的智造鸿沟就会越来越深。

正如我们在序言中指出的，当美国前总统奥巴马于2015年对外宣称“高速互联网是必需品，而不是奢侈品”时，摩尔的论文已经发表整整半个世纪了。如果我们再等半个世纪，国家领导人才意识到智造获取就像电力、水和数字连接一样属于必需品，那么形成一条严重的智造鸿沟几乎在所难免，这反过来又会加深现有的数字鸿沟，使未来的财富和收入差距更大，技术性失业现象更严重，并进一步增加社会的不稳定性。

另一方面，如果我们积极利用数字制造技术来提升个人与社区的自给自足能力，那么其益处就会惠及社会各个方面，这可能会颠覆我们对工作和日常生活其他方面的基本定义。要实现智造获取全覆盖的目标，就需要那些开创技术与社会体系以驱动第三次数字革命的人们聚精会神、持之以恒、齐心协力地共同努力。

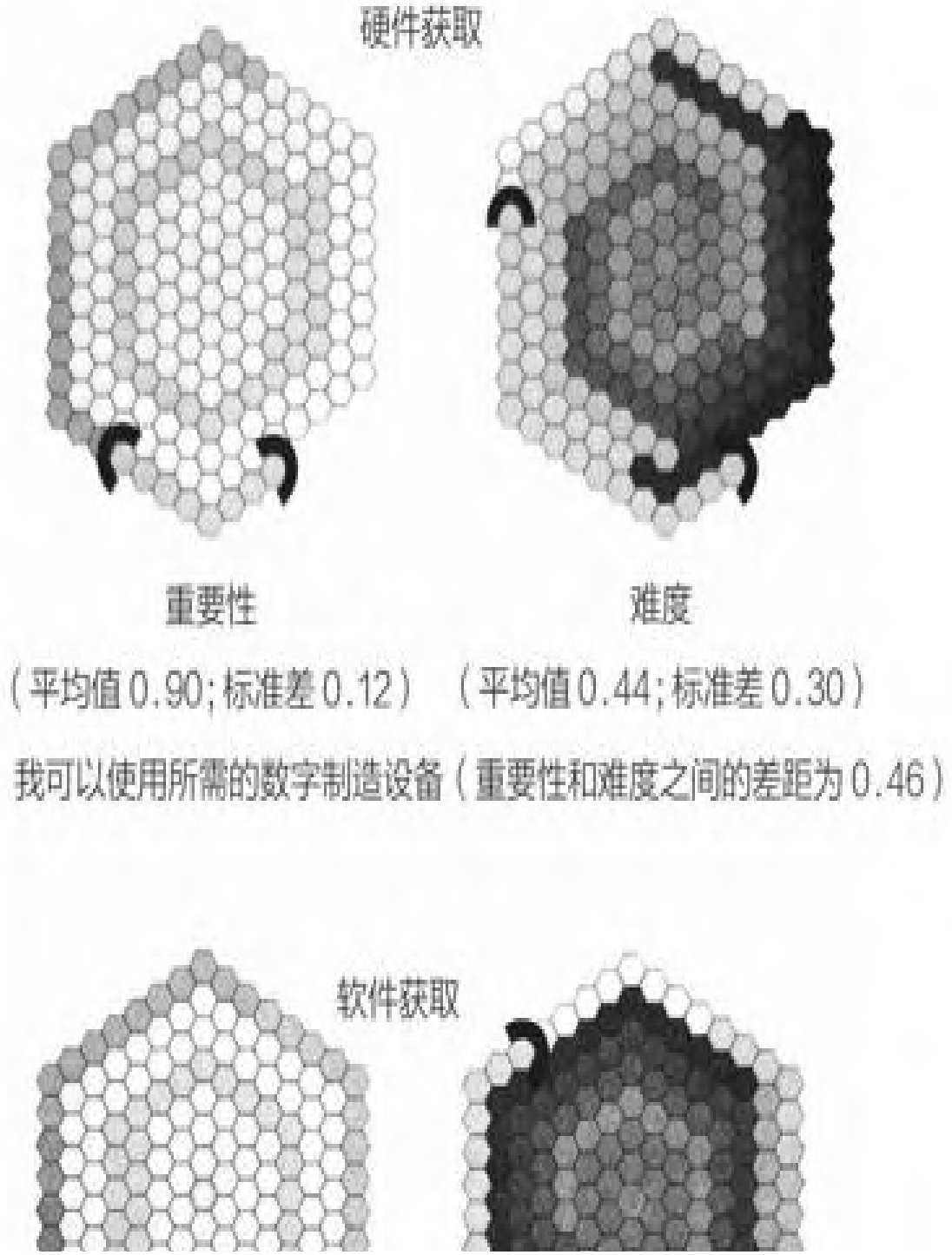
虽然智造的获取对全社会都是不可或缺的，但是对那些在前两次数字革命中落伍的人来说尤为重要。在我们的采访中，最感人的故事来自市中心平民区和乡村的“数制”工坊，过去，在那里，年轻人能获取的个人权益与自给自足的工具十分有限。

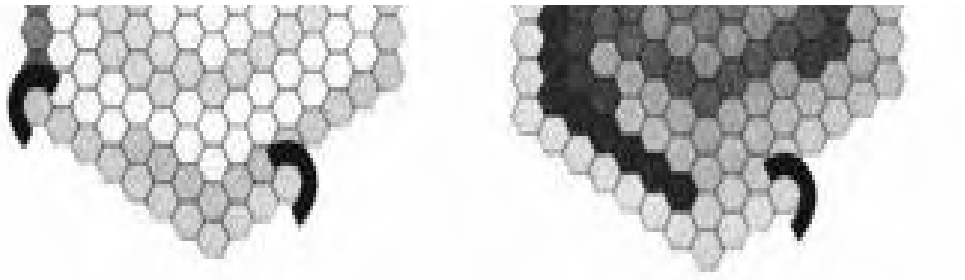
梅尔·金是波士顿的资深社区活动家，与尼尔共同创建了南端技术中心“数制”工坊。在谈到“数制”工坊对城市青年的影响时，梅尔强调：“我们为有前科的人和药物滥用者开设了一堂课。几年来我们一直在与各种机构合作，把这些人送进‘数制’工坊，为他们找回生活的意义。看到他们在创制物品时是那么含情脉脉，我们意识到我们走对了路子……我发现他们的生活已经发生了翻天覆地的变化。”

另一个例子是印度农村的“科学学堂”。在那里，人们初次创设“数制”工坊。工坊管理员约各什·库尔卡尼是这样描述工坊如何提高当地青年的能力的：“我们与14至20岁的年轻人一起工作，其中许多人是辍学者。在‘数制’工坊，他们通过解决问题和学习设计流程，培养各种各样的技能。我们鼓励学生们回到村庄，设计水压灌溉系统，然后在工坊里把这些系统制造出来。一些学生购买了诸如激光切割机之类的设备，并在当地创立了服务公司。还有人筹集资金开设了‘数制’工坊。”

从“数制”工坊中受益的不仅仅是年轻人和失过足的人。如果制造的民主化能使全球收入和财富日益扩大的差距得到缓解或重新平衡，那么每个人都将获益。在《公平之怒》（*The Spirit Level*）一书中，公共卫生研究人员凯特·皮克特和理查德·威尔金森记录了最健康和最幸福的社会是如何缩小贫富差距的。数字制造技术可以成为

缩小这些差距的重要解决方案，但前提是必须以系统和全面的方式处理智造获取问题。我们会在第6章进一步拓展思路，以应对智造获取及其他基本挑战。



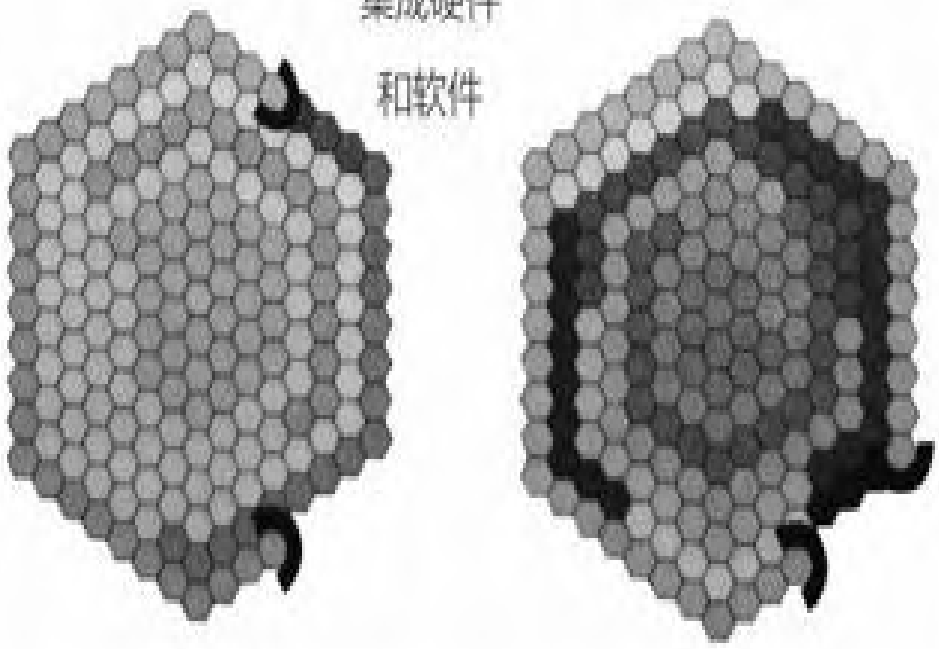


重要性　　难度

（平均值 0.86；标准差 0.15）（平均值 0.34；标准差 0.27）

我可以使用所需的数字制造软件（重要性和难度之间的差距为 0.52）

集成硬件和软件



重要性　　难度

（平均值 0.84；标准差 0.16）（平均值 0.52；标准差 0.26）

数字制造设备与数字制造软件的集成（重要性和难度之间的差距为 0.32）



图2－3 获取硬件、软件以及硬件和软件的集成的重要性和难度。WayMark Analytics公司绘

数字化公民

第二个基本挑战，是如何确保个人在获取智造技术后，学会应用这些技术。经历了前两次数字革命，教育工作者、雇主、家长和政策制定者都意识到，数字素养对于数字时代的生存和繁荣至关重要。要跨越数字鸿沟，素养问题与获取问题一样重要。虽然数字素养有很多种定义，但从广义上来讲，它需要人们获取知识、培养技能并形成思路，从而理解和使用数字技术，实现个人和专业目标，成为负责任的数字化公民。数字素养含义广泛，包括发掘并运用数字媒体，利用数字技术进行创造和交流，等等。

能获取数字接入的人大多数都知道如何访问数字内容并使用“色拉布”等简单工具。大多数人对塑造其日常生活的数字平台和技术理解有限。青少年平均每天有8小时沉浸在数字媒体中，但他们并不一定具备数字素养。想想世界上大量青少年失业的状况吧，即使在移动设备普及率很高的地区，许多技术性工作岗位也没有得到填补。这种数字制造的鸿沟与素养息息相关，当我们进行从比特到原子的创造时，这一问题极易恶化。

数字制造纷繁复杂。它引入了一套新的技能，包括对不断更新的计算机辅助设计软件和计算机辅助制造软件的导航、增材制造与减材制造、嵌入式计算，以及对用于制造的材料的生物学性质和化学性质的理解。它还需要设计思维、创造力、协同性、问题解决和灵活性。这些都需要跨越不同学科和领域的知识、技能和思路，因此目前还未

被很好地整合。我们将智造素养定义为：利用数字制造技术实现个人和专业的有效目标所必需的社会和技术能力，以及以负责任的态度使用技术的承诺。没有智造素养的普及，我们就无法建立一个更加自给自足、互联互通和可持续发展的社会。

纳迪娅·皮克是来自比特和原子中心的个人制造的先驱，她认为：“数制”工坊的数量可能会呈指数级增长，但被机器赋能的人的数量却并非同步增长。她将这种差距归咎于数字制造各个方面的复杂性，“软件既古奥又难学，人们可以学习打印塑料制品，但这并不意味着他们充分掌握了生产方式。”换句话说，智造的获取需要素养。

数字制造本质上是一种新的语言。正如素养学者詹姆斯·保罗·吉指出的那样：“识字离不开语法。语法造就了交流，但同时也限制了人的选择。”吉注意到语法之于语言就像规则之于运动一样。除非所有参与者都遵守规则，否则篮球或足球比赛几乎没有举办的可能。他还认为，学习一门语言需要潜移默化。语言兼具可供性（使某些思维方式和行动成为可能）和局限性。价值和假设都是在语法中建立起来的。

尼尔的“如何智造（几乎）万物”课程，实际上引入了一种用于数字制造的语言。他创立了一套语法，即一套规则，甚至很大程度上是一种文化。智造学院和各种不同的新的拓展课程［比如如何种植（几乎）万物］都可以被视为这种基础语言的方言。甚至像智造宪章这样的社会概念，作为管理、运作和发展“数制”工坊运动的共同基础，也是这种语言的一部分。一种语言要生存下去，需要一种基本语法，但也必须具有适应性和可扩展性。

创客运动、骇客空间、技术工坊和其他创造方式都提供了各种语言，其语法、规则和文化各不相同（而且相互重叠），这些语言的核心都是如何制作东西。每种方式都兼具优缺点，吸引了不同的人群。要形成智造素养，我们必须认识到智造社区正在开发新的语言，包括基础语法、规则和文化。在各种实用语言层面上形成素养，是使进入

数字制造世界的新人融入社会的关键，也是促进不同“数制”工坊、创客空间、骇客空间和其他社区之间对话与协同增效的关键。

数字制造不但涉及一种新的素养，而且是建立在传统素养的基础上的。正如吉警告的那样：“数字制造中的素养取决于多种其他素养，如阅读、写作和各种数字素养。这些传统素养可能会在数字制造的背景下得以重建，但人们依然离不开它们。”根据美国教育部2013年的一项研究，我们意识到要正确看待传统阅读能力的重要性，美国有3 200万成年人不能阅读，近1/4（21%）人口的阅读能力还未达到五年级水平。

关于智造素养的另一个挑战是工具的易用性。智造生态系统中的许多人都提到了对计算机辅助设计/计算机辅助制造工作流程的渴求，这些流程更直观，更易理解。但与此相对应的还有一种深切的担忧，即工具可能不易掌握，以至它们变成“黑匣子”，人们可以通过它们智造（几乎）万物，却无法得知智造（几乎）万物的工作原理。“黑匣子”实际上是与智造理念背道而驰的。如果伯爵红茶不再温热，那么皮卡德船长很可能无法修复《星际迷航》（*Star Trek*）中的复制器。

挪威奥斯陆“数制”工坊的创始人延斯·戴维克描述了易用性和“黑匣子”之间的矛盾：“尼尔关于修复星际迷航复制器的引喻有点儿令人担忧。它鼓励过度消费，而不是赋能创造……这是一个可怕的未来。你家里有台复制器，又能怎样？你应有尽有，就会开心吗？轻而易举地制造出大量的东西也可能适得其反。这是很少有人考虑到的挑战。”当技术变得复杂时，存在这样一种风险，即只有少数神父（创造者）才能掌控一切，而广大信徒（消费者）只能对技术背后的知识望洋兴叹。现在，智造的生态系统确实有一个高级神父阶层，他们讲着一口流利纯正的数字制造语言。

让我们来看看尼尔的“如何智造（几乎）万物”课程中的一些设置：

1. 数字制造的原则与实践：1周

2. 计算机辅助设计、计算机辅助制造及建模：1周

3. 数控切割：1周

4. 电子产品设计和生产：2周

5. 数控加工：1周

6. 嵌入式编程：1周

7. 3D造型和铸造：1周

8. 协作技术开发和项目管理：1周

9. 3D扫描和打印：1周

10. 传感器、执行机构和显示器：2周

11. 接口和应用程序设计：1周

12. 嵌入式网络和通信：1周

13. 机械设计：2周

14. 数字制造应用及其影响：1周

15. 发明、知识产权和商业模式：1周

16. 数字制造项目开发：2周

智造素养的提高，取决于智造获取程度的提高。与其他形式的素养一样，智造素养的形成也需要足够的时间，使人从一知半解进步到得心应手，从而用掌握的知识和能力教育他人。与智造获取一样，如果我们不积极为智造素养的普及奠定基础，现有鸿沟同样会扩大，而其他素养（无论是阅读素养、写作素养还是数字素养）的差距都加深了这种鸿沟。然而，如果能持续以创新的方式来提高智造素养，从特殊劳动力的发展到个人能力的提高，我们可以从中获益良多。

斯科特·西门森是明尼苏达州白熊湖世纪学院（一所帮助人们就业的两年制学院）的工程“数制”工坊兼增材与数字制造项目的负责人。斯科特指出：“工业正在融入数字化进程。我们要未雨绸缪，为新兴工作岗位培养人才。”他分享了这么一个例子：5名学生组成一个团队，应邀会见了威尔伯特塑料服务公司的代表，这家公司是为汽车工业和其他行业生产零部件的热成型制造企业。该公司的代表觉得3D打印技术可以将喷射造型设备模具生产的时间缩短2~4周，由于他们不具备专业技术知识，希望聘请几位学生作顾问。斯科特报告说：“通过三四次系列会议，学生们与公司高管确定了装配线上的夹具，然后就能制作原型了。在项目初年，学生们就已经能在这个领域的前沿胜任咨询工作了。”

素养的提高不仅体现在人才准备工作中，还体现在个人能力的提高上。斯科特借鉴尼尔的“如何智造（几乎）万物”课程，开设了一个为期两年的新课程，其中一部分内容就是增材与数字制造。他的初衷是所有学生最终都能制造自己的3D打印机。而事实上，有4名学生在课堂上就已经把打印机制造了出来。他风趣地说，现在得想出一个更具挑战性的顶级项目了。不过，此事更深远的意义在于，这些学生不仅在使用设备方面提高了素养，还深化了对基础技术的理解，提高了创新的能力。

赋能的生态系统

智造获取与素养的普及是更加自给自足、互联互通和可持续发展的社会的基石。智造获取和智造素养并不是空中楼阁，环境也很重要。目前，数字制造的环境包括一系列相互作用、相互独立和相互抵

触的元素。一个基本挑战是如何将环境塑造成赋能的生态系统，在个人和社区层面实现更大程度的自给自足。

我们将重点放在赋能生态系统的5个要素上，这5个要素包括：培育更有效的协作方式和知识共享模式；确保广泛分布的指导和领导；建立开放、健康的智造产品服务市场；吸引各种公共机构、私人机构和慈善机构的融资；设计契合新兴智造社区价值观的有效管理模式。一般来说，更宏观的环境常被视为稳定的外界因素，要想让数字制造以指数级速度增长并终将改善生活、造福社会，生态系统的这些方面就必须发展。

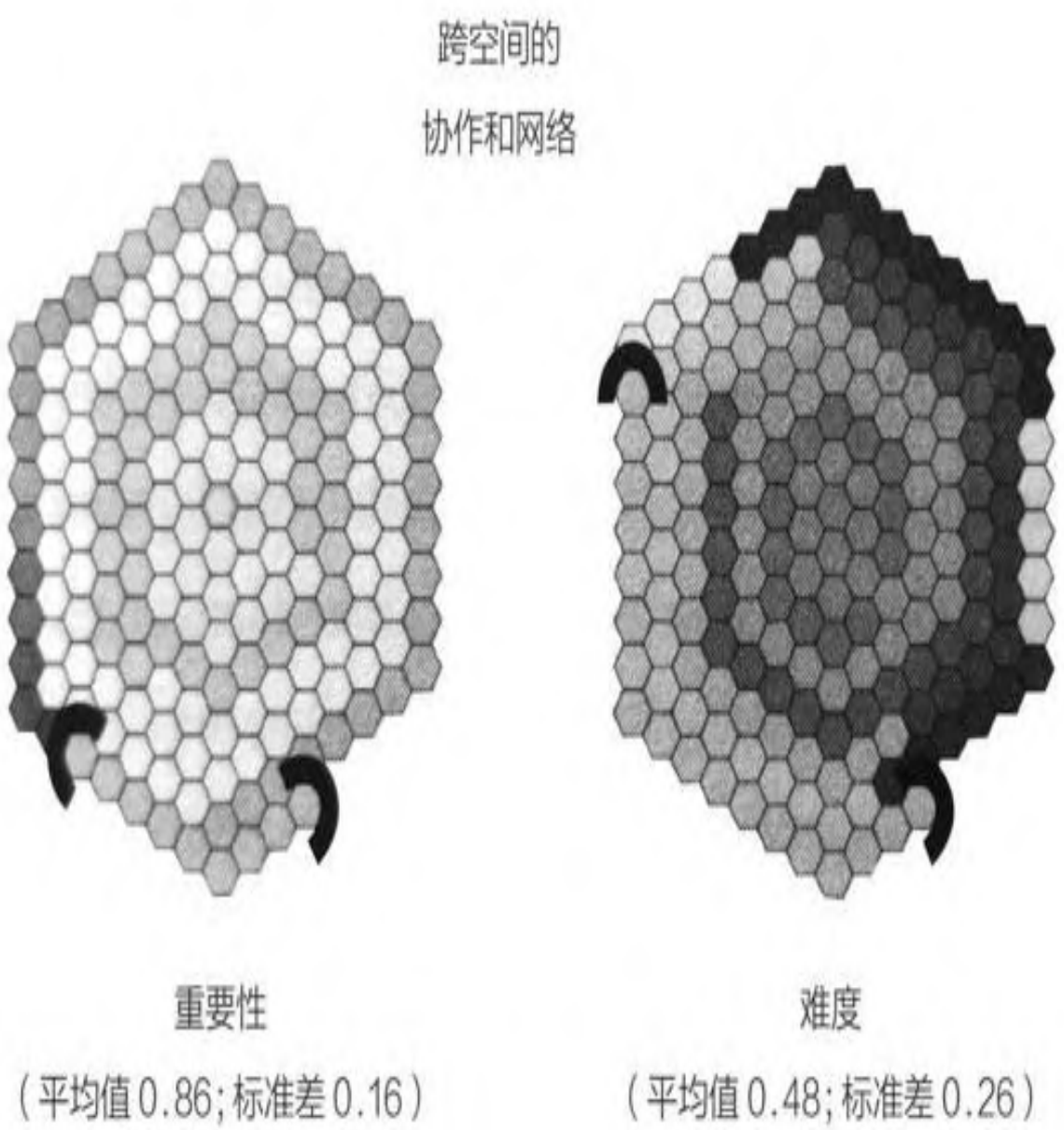
协作与知识共享

数字制造赋能生态系统的一个关键要素是，任何人在任何地方都可以相互协作，分享知识、建议和经验。由于数字制造需要广泛的专业知识，而且技术正在迅速发展，人员和项目的网络共享既是必不可少的，又是挑战重重的。在有效协作与知识共享机制下，人们就可以协同完成难以独立完成的工作：网络效应助梦想成真。根据尼尔在第1章中提出的梅特卡夫定律，协作产生了更多的协作，知识共享创造了新的知识，这样，数字制造就能更快地发展，并赋予更多人能力。

要获得网络效应，数字制造生态系统需要社会层面和技术层面的交互。设计“数制”工坊的架构时，我们至少在技术层面考虑到了这个目标。“数制”工坊试图为数字制造硬件和软件建立一个共同的规范，以促进“数制”工坊内部和全球“数制”工坊网络的协作和知识共享。虽然在大多数“数制”工坊中都有一个硬件占用空间，但是在这个过程中仍然存在大量的不协调。数字制造充斥着开源的、专有的计算机辅助设计软件和计算机辅助制造软件的“大杂烩”，当涉及跨工坊协作时，这些软件的多样性和可变性导致了大量的分歧与摩擦。

如果要跨越创客空间、骇客空间、技术工坊及其他类创客场所和社区的各种配置与文化，情况就更加复杂。

在大型智造和创客生态系统中，互操作性和可扩展性的挑战背后隐藏着一个关键的矛盾：对标准化的渴望和对分层决策的抵制。当然，制订符合既相互依存又相互独立的涉众利益的标准并非新鲜事物。从消费类电子产品到交通运输再到互联网，每一种事物都需要进行一定程度的标准化。这方面的历史教训很深刻。正如密歇根大学信息学院前院长约翰·莱斯利·金总结的那样：“在早期没什么人关心标准的情况下，建立标准相对容易，但当标准的重要性已经凸显时再建立，难度就很大了。每个人都想建立一套标准，但只想让其他人都采用自己的标准。”



在“数制”工坊、创客空间、骇客空间和其他涉及数字制造的地方，人们之间的协作和社交网络（重要性和难度之间的差距为0.38）
图2－4 跨空间协作的重要性和难度。WayMark Analytics公司绘

为了在数字制造的协作和知识共享标准上取得一致，我们需要尽早解决不同涉众如何在一个无管理环境中运作的问题，其进展取决于能否有效辨识当前实践中的利益、一致性和不一致性，以便涉众（个

人和组织）获得共同利益，实现各自的目标。这一过程被乔尔和他的研究团队称为“横向一致”，它在不同的创客社区内及社区之间都始终是个挑战。

在对数字制造领导者的问卷调查中，我们询问了他们与不同类型的创客空间的联系情况，通过一些选项来检查所有应用。尽管这项调查是通过智造基金会电子邮件列表分发的，但回复显示社区之间的联系情况存在大量的重叠，81.1%的受访者表示与当地“数制”工坊有联系，42.3%的人表示与创客空间有联系，31.4%的人表示与创客嘉年华活动有联系，17.1%的人表示与骇客空间有联系。当我们问及这些空间的合作和社交时，人们的观点差距显著，77.5%的人表示高度重视，超过1/3（36.9%）的人说很难做到，超过一半的人说这至少有点儿困难。

这些社区的运作可能有相当大的重叠性和协同作用，但在“数制”工坊、创客空间、骇客空间、技术工坊设施和其他新兴创客社区资源之间，语言、身份和文化也普遍存在着微妙的矛盾关系。这些社区都是平行发展的，但在它们共同发展的可能性与必要性问题上，不同的人的认识程度各不相同。在社会科学中，要处理好同质性（与你这样的人联系）和多样性（欣赏那些不同于你的人）之间的矛盾，需要不断地平衡与迭代。要取得平衡，首先要了解共同点和差异性的根源。

每个社区都渊源各异。当我们与创客传媒的创始人兼首席执行官戴尔·多尔蒂谈论《爱上制作》杂志和创客嘉年华的名称时，他用了早期健身房来进行类比：“第一家健身房是一种健身工作室，来健身的人主要是男性举重运动员。这种健身房对于许多想要锻炼但并非专业举重运动员的人来说并不受欢迎。”他曾考虑使用“黑入”或“骇客”，但又觉得这些术语对许多团体，尤其是学校来说，可能难以接受或不够友善。他还考虑过用“‘数制’工坊”来命名（《爱上制作》杂志第一期中有一篇尼尔的专访），但又担心启动“数制”工坊

的成本和技术复杂性超出了大多数人的承受能力。他希望，任何人在任何地方，都能以较少的成本和较低的难度开办创客空间或参加创客嘉年华。最终，在女儿的建议下，他使用了一个简单的词：make（制作），因为正如她所说：“每个人都喜欢做东西。”

有趣的是，在过去10年里，由于创客社区和智造社区都在扩展，两者在某种程度上越来越紧密地联系在一起。数字制造硬件和软件的成本虽然还相当高，但已在稳步下降。创客空间的技术水平也在稳步提高。多尔蒂开玩笑说，他认为尼尔“是创客运动的研发部门”。尼尔也深以为然。在许多方面，“数制”工坊运动是更为广泛的创客运动的前沿，因为它利用迅速发展的数字制造技术，更有效地在多个尺度上制造更多外形与功能兼备的物品。

骇客空间和技术工坊则有着不同的起源和文化。骇客空间一直比较分散，有一种松散的、非正式的、由社区驱动的文化，其社区内有一个由志愿者运营的网站，里面有维基、博客和其他资源，但这个网站没有创始人或核心组织。每个组织的运作模式都不一样，反映了当地社区或成员对骇客文化的理解。成立于2006年的技术工坊是一家“会员制的、自己动手的工坊和原型工作室，为不同年龄和技能水平的创客提供负担得起的，开放的，可获取各种工具、设备、资源和工作空间的机会”。

技术工坊的特点是其会员制模式，它向会员提供各种各样的数字和非数字工具。以创客为中心的组织和社区之间存在不少差异，令许多人感到困惑。在《爱上制作》杂志2013年的一篇文章中，古伊·卡瓦尔坎蒂试图解释这种差异。他发现许多学识渊博的人都无法区分骇客空间和创客空间，他说：“我认为需要区分这两者，因为在这里，两个词语的概念和表述对我来说有明显区别。”

卡瓦尔坎蒂还提到了技术工坊和“数制”工坊社区，他认为这两个社区都包含在更大的创客空间概念中：“在我看来，技术工坊和‘数制’工坊都是创客空间的专营店，它们的工作人员专注于从零

开始，利用多种类型的媒体进行创作。具有讽刺意味的是，在‘创客空间’一词被广泛使用之前，技术工坊和‘数制’工坊这两个词就已经诞生了，因此它们各自的标志性名称反而比总括性术语更有生命力。”

当然，并非每个人都同意“创客”这个词可以涵盖一切。此外，“数制”工坊既不是专营店，又不是商标（尽管它兼有这两种属性）。然而，卡瓦尔坎蒂强调了这些社区的共性是如何与它们在方法、文化甚至核心价值观上的个性共存的。许多以创客为中心的社区，能力不同，文化各异，操作实践五花八门，既自然又健康。尽管如此，对一个赋能的生态系统的评价仍取决于社区协作和共享知识的能力——尤其是在数字制造方面，设计与制造可以通过数字化的方式共享、调整和共同创建。如果不同的涉众都能够围绕可共同操作的标准来保持一致，那么更多的人和群体就可以在保持各自身份特征和运营模式的同时受益。

指导与领导

赋能智造生态系统的一个关键要素，是确保指导和领导在整个社区广泛分布。智造基金会的负责人雪莉·拉斯特指出了人的重要性：“我过去认为扩大网络就是降低建立新工坊的成本。现在看来成本只是部分问题，更重要的是人。这不仅仅是以自上而下的方式建立新工坊的问题。要提供数字制造的访问途径，就需要培养与新工坊相匹配的人的能力，需要对已经启动的工坊进行可持续的领导。”

当人们启动一个新的“数制”工坊时，设备通常在几个星期或几个月内就能投入运行，但建立和运行所需的社会体系则是另一回事。要使人员配置、内部管理、可持续的财务模式和设计方案正常运作可能需要数年时间，更不用说与关键的外部涉众建立必要的关系了。所以，从一开始，社会体系就落后于“数制”工坊的技术体系。

许多早期的“数制”工坊都是由积极投身其中的社会企业家建立的，这些企业家都有强大的领导力，在发起和运营“数制”工坊的过程中克服了各种挑战。那些尝试启动工坊但未成功，或无力维持工坊运营的人的领导力则不太明显。随着对社区制造的需求的不断增长，领导能力的提高将成为生态系统中的一个基本挑战。

智造生态系统的一个相关障碍是找到经验丰富的导师，以对所有需要帮助的人在软件、硬件和材料方面进行指导。已经建立起“数制”工坊并对当地人员进行了领导力培训的智造先驱（比如延斯·戴维克和纳迪娅·皮克）则多多益善。然而，在自己研究和热爱的项目与越来越多的咨询和指导工作之间，他们很难平衡。事实上，与我们交谈过的许多智造先驱在此问题上都很纠结。许多“数制”工坊没有可以指导如何利用工坊开展各种项目的专职导师。随着“数制”工坊数量的增加，以及技术使制造由以社区为基础变得更加个人化，宏观智造生态系统面临的挑战可能会增加。

除了分布式的现场指导之外，“数制”工坊还需要大量的资源，为那些没有机会接触当地导师，但在项目过程中遇到困难的人提供及时的指导和反馈。总的来说，绝大多数智造领导人（87.3%）报告说，在当地工坊接触导师是非常重要的。与此同时，超过1/3（35.5%）的人说这很难做到，近1/2（48.4%）的人说这至少有点儿困难。

指导不仅局限于数字制造所需的软件和硬件方面。目前智造生态系统的另一个关键知识和资源缺口，是在数字制造的原材料方面进行指导。这是智利圣地亚哥大学教授阿莉西娅·加姆勒薇兹尝试突破的一个领域。她重点研究发展循环经济（“数制”工坊的资源可以在本地获取并持续重复使用）而不是线性经济（资源被获取、使用和丢弃）。她指出：“我们需要了解我们周围的材料，利用天然聚合物和其他未被充分利用的材料创造高效的经济。因此我们需要关于当地材料实用性和可用性的分布式信息。”目前，这样的资源在整个智造环境中是非常有限的。

在硬件和软件的设计过程中，能否获取即时信息和反馈也是一个挑战。纳迪娅·皮克对当前的复杂性、资源需求，以及在技术内部形成反馈的机会是这样描述的：“激光切割机上有数千种针对不同材料的设置——激光脉冲的频率、激光的移动速度等等。我清楚这一点，可以回答大家的相关问题，但关于这些信息的可共享的资源我却没有。”皮克补充说，不仅仅是单机才会有这一问题，“很多复杂的组件都需要标准化和互操作，并通过各种方式去适应不断发展的技术”。她指出，“我们需要发展基础设施，以便在技术中隐含知识和能力。”正如皮克明确强调的那样：“重要的是，硬件和软件变得更加直观，并向用户提供更好的反馈。”

这种人与机器有效协同工作的思路是智造经验的核心，也是最近几本关于未来工作的新书的主旨。例如，麻省理工学院的研究人员安德鲁·麦卡菲和埃里克·布莱恩约弗森是《第二次机器革命》的合著者，他们探索发现，在当前与未来的经济的一些领域，人和机器协同工作的效果强于各自为营。与其他一些人一样，他们强调“与机器赛跑”的必要性，经常引用自由式国际象棋中一个值得玩味的例子，以说明技术可以作为人类的补充。在这种类型的国际象棋比赛中，计算机与人类配对，共同与超级计算机竞争（超级计算机已经击败了世界级的国际象棋大师）。比赛中，由人和计算机组成的团队通常会击败超级计算机或顶级国际象棋大师。设计智造技术，并使之更有效地与智造创造者合作，将是赋能智造生态系统的另一个关键组成部分。

强劲的市场平台

智造生态系统还需要开放和强大的市场平台，作为发现、共享、购买和销售数字制造设计、产品与服务的载体。前两次数字革命的一个关键教训是，在任何特定的生态系统中，主导市场平台的控制者都能获得巨大的财富和影响力。尽管苹果公司最出名的是它的电脑和手

机，但其应用商店的市场和分销平台是它最有价值、增长最快的资产之一。谷歌Play移动应用商店、维尔福的蒸汽视频游戏发布平台和游戏社区以及亚马逊庞大的在线零售平台也是如此。这些企业在塑造其运作的生态系统方面产生了巨大的影响。

无论交易规模大小，市场平台提供商都能赚钱（对于应用程序、游戏、音乐和其他内容，平台提供商一般每笔交易都会提成30%）。创建数字内容的工具的成本已经大幅下降，而且理论上分配已经民主化。但正如尼尔所指出的，对于绝大多数内容创建者来说，应用程序生态系统更多的是饥荒而非盛宴，大部分收益都流向平台所有者和顶级内容提供商。同样重要的是，这些平台提供商已经成为重点应用程序、游戏、产品和设计的守护者，通常手段是混合算法（将最受欢迎的选择置顶）和管理（编辑选择）程序。

因此，一个关键的问题是，谁来控制智造设计、产品和服务的发现、共享、购销的主要平台？扮演这一角色的组织将通过算法和管理程序中内置的价值观来塑造文化。这些程序中的过滤并不是中立的，每个决策都有内置的假设，而且将社区影响集成到算法和管理程序中的方式也存在相当大的差异。一个典型的例子是谷歌和亚马逊算法中内置的假设，这些假设产生了我们看到的大量信息和购买行为。就像尼尔指出的那样，即使生态系统更像烹饪原料，仍然会有守护者——相当于大型连锁超市——决定哪些产品和服务被重点展示。

虽然已经有一些公司开始为3D打印或激光切割提供市场和在线服务了，但还没有出现任何一家或几家主导市场的企业。在前两次数字革命中，以平台为基础的高端市场往往由少数营利性公司主导。考虑到智造和创客运动在很大程度上是自底向上的，具有自发特性，我们不排除主导市场出现在非营利领域的可能性。像维基百科（知识普及）和可汗学院（学习普及）这样的非营利组织在其领域内都经历了指数级增长，并产生了深刻影响。当然，它们更多地体现为知识与内

容的传播，而不是产品与服务的发现、分享和销售。拥有高端智造市场平台的组织将对智造生态系统的文化产生巨大影响。

多样化的融资渠道

赋能生态系统的另一个关键的基本组成部分是为研究开发提供各种私募基金、公募基金和慈善基金的混合融资机制，以及为整个智造社区的营利组织、非营利组织和任务型组织提供阶段性融资。要使“数制”工坊保持其指数级增长速度，就需要大量资金用于持续的基础研究，以支持社区“数制”工坊网络的不断发展，新工坊和网络的创建，并从有利于智造生态系统的角度为开发智造硬件、软件、材料和服务的公司提供商业融资（天使投资、风险投资和私人股权）。

智造生态系统诞生于美国国家科学基金会的研发资助项目。在尼尔所设想的数字制造路线图中，仍然存在着必须克服的重大技术挑战。从社区制造过渡到真正的个人制造，基础性研究仍然是必要的。传统上，科学技术研究由美国国家科学基金会、美国国家卫生研究院、美国国防部高级研究计划署等政府机构（以及其他国家的类似机构）进行资助。然而，现今许多政府核心机构的效能都在备受质疑，更不用说科学和社会科学基础研究的价值了。

虽然私营部门具有推动创新的可能性和必然性，但低估公共资金的影响不仅是危险的，还没有得到历史的支持。经济学家玛丽安娜·马祖卡托在其著作《企业家型国家》中直击公共资金问题，概述了美国国防部高级研究计划署在互联网发展中的作用（通过创建阿帕网来体现），描述了美国政府和欧洲国家政府的研究经费为智能手机触屏的技术研发、医疗制药领域的技术突破所奠定的科学基础。她重点举了美国国家卫生研究院资助的研究和其他不少重要基础研究的例子，指出基础研究在多年之后都促进了社会的变革，引领了产品与服务的发展。

库马尔·加格曾在白宫科技政策办公室工作8年，主要负责教育和创新事务。他认为，“对政府领导人而言，关键是要找出政府或公益融资机制最适合解决的最大痛点”。他还指出，在财力和人力资本有限的情况下，往往需要围绕一种系统方法做出艰难选择——是把资源用于解决多个痛点，还是集中力量解决最大的痛点，并针对这一特定痛点优先配置财力和人力资本。

对于那些有志于创建“数制”工坊的人来说，第一个常见问题是启动一个工坊需要多少资金。智造基金会的雪莉·拉斯特预测，启动一个社区工坊并运行两年的预算约为25万美元，这个预测包括硬件、软件、耗材、空间等方面，还包括建立和管理工坊的费用与志愿工作人员的招募。现有工坊的资金来源多种多样：政府、社区、当地企业、大学、慈善组织、投资、众筹，以及最常见的多种渠道的组合。后续资金则来源于课程培训、入会费、会员费和项目赠款。

“数制”工坊先驱者的早期团队成员都是满怀激情的社会企业家，他们能够筹集创业资金和后续运营资金。然而，要使智造网络继续发展，总成本就必须降低，战略投资的来源就需要多元化，可持续的商业模式就需要打磨和调整，以满足当地实际需要。要降低总体基础设施成本，就需要一代一代地不断改进硬件和软件。一个强大的、赋能的市场生态系统可以使硬件和软件更优化、快捷、廉价，并支持创新，从而确保成本效益高的环境友好型原材料的供应。

3D打印技术有利于激励初创企业，吸引投资资本，从而形成竞争、鼓励创新并降低成本。我们期待对减材制造设备、耗材、其他创新和成本降低至关重要的领域的相同市场和投资支持。只有智造生态系统所提供的产品与服务成为普遍刚需，而非仅针对富裕人群，才会引起金融领域的重视与参与。

今天，对有些人来说，获取“数制”工坊资源是刚需。他们认为有必要在工坊中投入一定的时间，是因为他们可以在工坊里做些什么。这些人包括需要用数字制造技术提升工作效率的设计师和现代工

匠。对于一些企业家、小业主和大型商业企业工作人员来说也是一样的，如果他们需要成本效益高的、快速成型的产品，就必须利用“数制”工坊的资源来实现。

然而，对于大多数智造社区来说，刚需更多体现在数字制造的过程中，以及融入创客与创新社区的过程中。在智造生态系统的发展过程中，我们接触的大多数人都反映，“数制”工坊的不同凡响之处，在于其社区的丰富内涵（而不是正在生产的产品）和生产的协作过程。

在更大的智造生态系统中，用数字制造工具制造物品的过程，与其产品的质量和用途之间存在着矛盾。莫里斯·康蒂在一家领先的数字设计软件开发商（Autodesk）负责应用研究和创新工作。他对人们参与制造的热情和现有产品的局限性发表了看法：“一个人能够从事某种个人制造，当然超级有趣，但它实用吗？我见识过“数制”工坊对文化的冲击，很有感召力。但即使工坊具备了世界级的设施和专业工作人员，使人们对智造更感兴趣，其产出也很少能与商业产品相提并论。”

学习如何在钥匙链上用激光雕刻人名或者打印可爱的3D小图形是一回事，能否制造万物则是另一回事了。尽管数字制造技术的学习过程蕴含着巨大的教育价值，正常人还是需要很长时间才能制作出大部分消费品。就连尼尔，这位在家里和工作中都能接触最先进的数字制造工具与技术的专业人士，也只制作出一小部分消费品。随着时间的推移和技术的进步，这种情况会逐渐改善，但目前，在“数制”工坊中制作物品和推进社区发展的过程仍然具有很大价值。

基础扎实的数字制造过程，可以培养批判性思维、设计思维、解决问题的能力、创造力、协作能力和适应能力；可以使人形成各种数字素养，加深对工程设计、材料科学等领域的理解。对于越来越多的教育工作者——至少是那些推崇从实践中学习和通过项目协作进行学习的人来说，这种知识、技能和思维方式的培养当然是刚性需求。同

样，数字制造对基本STEM技能的推动方式，以及相关素养的培育方式，也逐渐成为许多教育工作者、决策者和一些行业的刚性需求，这些行业希望未来的雇员能够利用迅速发展的技术应对各种各样的新挑战。

挪威北部林根阿尔卑斯“数制”工坊“首席牧民”哈肯·卡尔森讲了一个故事，这个故事充分说明了“数制”工坊的益处。故事中，三个十几岁的女孩在工坊里度过了大部分空闲时间，后来她们都离开小镇去上大学，最后都成了医生。当地人会指着林根“数制”工坊，并讨论“林根山上那栋奇怪的房子是如何产生医生的”。重要的是，她们在经过医学培训之后返乡了。不仅仅是医生，在工坊工作的当地青年后来从事了各种各样的工作，他们当中有企业家，还有教授，他们都把自己的职业成就归功于工坊的工作经验。例如，尼尔在第1章中提到的汉斯－克里斯蒂安·布鲁沃尔德最近就成了挪威当地一所大学的副教授。哈肯是最早认识到“数制”工坊可以传授各种知识、培养各种技能和思维的人之一，这些知识、技能和思维不仅能帮助人们智造（几乎）万物，还能帮助人们从事（几乎）任何职业。

这些过程导向的益处是巨大的，这个益处需要被扩大以吸引更广泛的人群，同时也需要出现智造产品与服务。智造生态系统将不会局限于单一的杀手级应用程序或服务，而是会不断发展和融合。在社区“数制”工坊的早期阶段，杀手级应用程序很可能是以提供服务而非制作某种特定的物品为目标的，这样可以使个人有深刻的制造体验。

随着“数制”工坊网络的发展，以及早期个人制造或小型企业制造的出现，新的杀手级应用程序和服务应运而生，其目标是将赚钱、省钱和提供大型娱乐的特定设计或项目变为刚性需求。这些应用类似于个人电脑早期的杀手级应用，如文字处理软件、电子表格软件、电子邮箱和电脑游戏。随着时间的推移，杀手级应用程序从特定产品转向了使互联网内容民主化的支持平台和工具，包括制作网页、博客和

应用程序的工具，以及上传和共享图片、视频、新闻和其他信息的工具。这些工具使数字内容的大规模出版与共享成为可能。

尼尔对刚性需求问题的回答，某种程度上针对的是个人制造。他认为，数字制造的杀手级应用不是你能在商店里买到的东西，而是一种为一个人这么小的市场生产商店里买不到的产品的能力。然而，为了实现这一愿景，我们需要在赋能生态系统方面取得重大进展，而这些生态系统又需要得到有效治理。

分布式管理

在任何组织中，决策和协调都是困难的，其复杂性与组织内各实体彼此独立又相互依存的分布式架构有关。虽然新技术不是早期的重点，但对分布式社区的有效治理是赋能生态系统的一个关键要素。尼尔以互联网的治理为例，论证了分布式结构的有效性。尽管互联网治理由于其规模和分布式特性而运作顺畅，但它在获取、素养、文明和其他全球性问题等关键问题上没有权威性或影响力。数字制造需要人们提高决策能力和协调能力，以便建设性地应对基本挑战，同时保持其分布性和独立性。

目前，生态系统中还没有一个共享的数字制造治理平台。主要的现行做法是依靠智造宪章，但智造宪章是不完整的。宪章里当前的内容没有任何问题，然而，它缺少了许多宪章共有的一些关键要素，包括关于共同愿景的声明、决策机制和争端解决机制。此外，生态系统缺乏一个完整的许可程序，使宪章真正代表不同涉众就游戏规则达成协议。以下是智造宪章的全文：

什么是“数制”工坊?

“数制”工坊是本地工坊的全球网络，通过提供数字制造工具促进发明创造。

“数制”工坊里有什么?

“数制”工坊共享一份动态发展的核心能力清单，内容包括为智造提供便利，促进人与项目之间的共享。

“数制”工坊网络提供什么内容?

单个工坊无法提供的运营、教育、技术、财务与物流支持。

谁可以使用“数制”工坊?

“数制”工坊作为一种社区资源，向个人开放，并有计划地向项目开放。

你的职责是什么?

安全：不伤及人员，不损坏设备。

运行：协助清洁，维护并优化工坊。

知识：建立文档并提供指导。

谁拥有“数制”工坊的发明创造?

“数制”工坊开发的设计与程序都可以受到保护并根据发明者的选择出售，但应当允许个人使用和学习。

企业如何使用“数制”工坊?

商业活动可以在“数制”工坊中进行原型化和孵化，但不能与其他用途相冲突，它们应该在工坊外而不是在工坊内发展，并使为其成功做出贡献的发明者、工坊和网络受益。

在我们对智造领导人的调查中，大多数人（69.8%）报告说，对智造宪章的认知与利用很重要，但大约11.3%的人强烈认为这并不重要。此外，56.0%的反馈者报告说很难利用。最能说明问题的是，在回答这两个问题时，相当大的比例（超过20%）的人要么回答“不知道”，要么回答“不适用”（用括号表示）。

当我们与尼尔分享这些发现时，他说宪章在早期是必不可少的，但现在就不那么重要了，因为共同的理解已经根植于智造学院的运作方式和智造全景的其他部分。我们在访谈和调查中，确实观察到了整个社区的智造价值观。我们竭尽所能将其归纳如下：

在“数制”工坊中寻找意义、目的和乐趣，支持工坊内部和工坊之间的协作和社区建设。在数字制造过程中培养个体和集体的能力。

通过数字制造努力实现个人与社会的可持续发展。

数字制造应该造福每个人，而不仅仅是少数幸运儿。

在当前的生态系统中，这些共同的价值观似乎起到了与宪章相同的作用。事实上，这些共同的价值观超出了现行宪章的规定，而且宪章的某些内容与当前的做法相冲突。例如，“谁可以使用‘数制’工坊”的标准是“数制”工坊对个人开放，并有计划地对项目开放。对于义务教育学校、学院、大学、博物馆和其他机构的“数制”工坊来说，这一标准并不总是可行，因为这些机构可能会受到开放的限制。

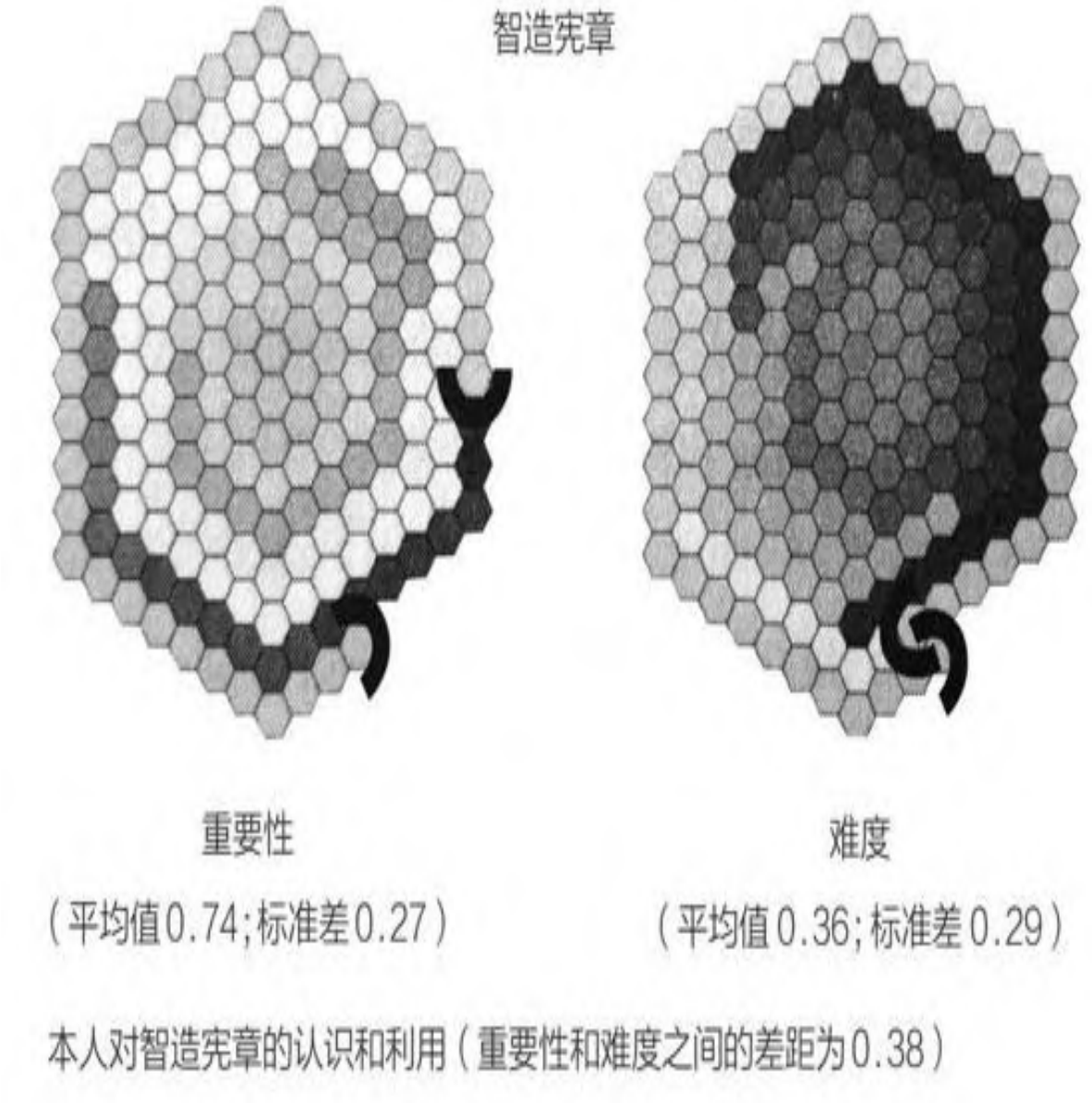


图2－5 访问和使用智造宪章的重要性和难度。WayMark Analytics公司绘

不过，仅仅陈述共同的价值观，就会产生这样一个问题：随着生态系统的发展，是否需要进行一些修订——更新宪章与标准，以确保其一致性？特别是在一个加速变革的时代，宪章需要成为动态的文件。参考美国宪法和其他提供长远框架的章程文件，再结合智造生态

系统的实际，就可以找到思路。建立这样的机制并非易事，但它是关乎事物如何在特定领域中运作的关键步骤。未来的挑战是确保生态系统的所有组成部分——赋能平台、实践和协助治理者——能够适应在第三次数字革命中事物发展的指数级速度。

防范风险

技术发烧友会承认风险，但他们的热情主要集中在技术的潜在收益上。这推动了创新，是个好事。然而，对第三次数字革命早期的清晰观察，也揭示了相当大的风险，需要我们记录、理解和防范（即使它们永远无法完全解决）。例如，智造获取与智造素养问题暴露了破坏性智造鸿沟的潜在风险。其他风险包括与耗材相关的环境恶化、技术滥用（例如，恶人制造恶果），以及人们在日益细粒化的水平上操纵比特、原子和基因所带来的意外后果。

社会的两个首要的制度性目标是创造价值和减轻损害。减轻损害的核心挑战是谁（或哪个团体）将对每一种风险承担责任。在任何大型商业或公共项目的谈判中都有一个关键环节：由谁来承担哪种风险。这一环节首先要求人们辨识所有已知风险和可能产生的各种未知风险。然后，如果谈判进展顺利，每项风险都能分配给可以管控给定风险的一方或多方，个人或团体因持有风险而获得补偿款或其他补偿。如果谈判过程中不能形成解决方案，项目就不太可能取得进展。目前这方面的一个例子是围绕谁承担自动驾驶汽车风险的谈判，选项包括汽车制造商、汽车车主、保险公司、算法编程团队、政府或其他人。

然而，在第三次数字革命中，没有与多个涉众就风险问题进行谈判的既定机制。因此，事情可能会发展，但管理风险的责任并没有完

全明确。不能系统分配责任可能会给整个第三次数字革命带来一种元风险。在没有风险缓解协议的情况下，关键风险将得不到管理，当灾难性事件确实发生时，应对措施和资源将会不足，而人们相互之间的推诿塞责会进一步加剧事态的恶化。

首先是环境风险。在目前有关3D打印的炒作中，有专家提出，“要让每张办公桌上和每户人家里都有3D打印机”，这是对微软企业使命声明的改版。考虑到3D技术的目前状况，这一目标有可能导致一场环境灾难。3D打印过程基本上是将各种材料堆压在一起。虽然一种常见的聚乳酸材料是以植物为基础的，但从石化产品中提炼出来的其他材料是不环保的，而且很难回收利用。

巴塞罗那“数制”工坊的托马斯·迪兹说：“‘数制’工坊使用的胶合板、加入丙烯腈－丁二烯－苯乙烯（ABS）的塑料、依赖于丙烯酸酯的电子产品都越来越多。这绝对是不可持续的。”如果数亿人在不环保的材料基础上进行迭代和修补，第三次数字革命就会面临潜在的环境风险，包括自然资源的严重枯竭、碳排放量的急剧增加，以及数字制造过程中过量废弃物的产生。

然而，如果我们能够开发出环保的可重复利用的数字制造材料，就有可能产生巨大的环境红利。在智造城市白皮书中，迪兹指出了城市目前在环境退化中的作用：“极端工业化和全球化使城市变成了最贪婪的物质消费者，它们通过直接和间接的能源消耗成为碳排放的主要来源。我们需要重新规划这些城市及其运作方式。”正如我们在后面的章节中将看到的那样，对本地材料的再利用、对更好的资源的开发，以及可组装和可拆卸的数字材料的出现，对于防范全民个人制造的环境风险至关重要。

发展中国家的另一个风险源自其薄弱的体制。比起许多地方机构领导者，跨国公司有能力更快地培养数字制造能力。这种情况已经悄然发生了，美国联合包裹服务公司在其仓库设施中建立了快速原型中心，以便按需生产和交付某些产品。其风险在于，相关技术的守护者

不是分布式社区中心，而是大型企业，情况更类似于前两次数字革命。在这一场景中，第三次数字革命将更快地降临到发展中国家，不过变革方式将取决于对企业是否有利可图（是否符合智造价值观）。

还有一个更加凶险的风险，就是恶人加入智造会产生恶果。这种情况已经出现了。3D打印枪支的故事很快就在社交媒体上蔓延，在互联网上层见迭出。武器的制造是一个真实的风险。尼尔在回应这一担忧时指出，“数制”工坊并非制造枪支最有效的方式——比起在“数制”工坊制造枪支，还有更有效的方法来获得枪支。然而，随着数字制造工具变得更优化、更快捷、更便宜，他的观点就变得不那么有说服力了。如果数字制造的目标是让任何人制造（几乎）万物，那么它肯定也可能会使制造具有危险性和破坏性的东西的速度加快。

解决这个问题没有万全之策，但是有多种方法可以帮助我们降低风险。其中一个方法是通过当地工坊进行指导与监督。来自印度浦那郊外“数制”工坊的约各什·库尔卡尼积极应对这一挑战。他认为，“数制”工坊可以成为一个重要的工具，帮助我们识别具有危险倾向的青年并引导他们走向正途。当被问及年轻人使用工坊制造武器的风险时，他认为，密切监督是关键。“虽然风险不大，但我们也会看到学生用磨床磨刀。当这种情况发生时，我们会加以阻止，解释这不是工坊所允许的。”这种方法类似于社区警务的工作，在突发事件的早期就能将其系统地捕捉到，以确立规矩，弘扬文化，降低更大、更具有灾难性的事件发生的概率。

系统层面的一个相关风险是数字制造世界的网络恐怖主义，更令人担忧的是数字材料世界的网络恐怖主义。这意味着我们需要无懈可击的分布式基础设施。此外，我们知道系统解决方案的提出必须考虑社会因素，同时也需要技术支撑。

虽然艾伦和乔尔的职业生涯是在截然不同的部门中度过的，但他们都对管理技术风险的创新方式有着深刻见解。这些见解源于他们各自的领域——工作场所和劳动实践（乔尔）以及数字娱乐和在线社区

（艾伦）。这说明了跨部门合作和知识共享的重要性，因为几乎每个部门和领域都必须应对技术加速发展带来的持续性风险。

我们先从乔尔开始。他受雇于美国国家航空航天局（NASA），参加了负责对未来“月球—火星”任务进行安全和风险评估的两个小组中的一个。一个小组使用所谓的概率风险评估（PRA），它统计所有已知的可能会发生的失败，为每个失败加上一个概率，然后计算任务的总体风险。另一个小组由麻省理工学院的南希·莱维森领导，这个小组使用了一个系统安全模型，该模型观察整个系统的运行情况，并利用一个与其相互依赖的系统动力学模型来识别漏洞。一种方法将风险分解，而另一种方法则研究事物是如何相互叠加和相互关联的。然而，风险分析往往只关注PRA式的风险分解，而不太关注系统方法。在此情况下，乔尔回顾了PRA报告，该报告指出，这是对与任务相关的所有已知风险的全面评估——除了那些与软件和组织因素有关的风险。显然，第三次数字革命中的风险不能仅仅通过分解单个部件的风险来解决。

除了正式的风险缓解程序外，在法律方面也存在基本挑战。由于一些物品对安全至关重要，所以“数制”工坊必须确保这些物品符合要求。使用假冒伪劣产品甚至是善意的替代部件，都可能会破坏产品设计的安全性。这一挑战涉及认证和知识产权问题。冈萨洛·雷伊是穆格公司的首席技术官，该公司生产航空航天、医疗、能源和其他行业的精密控制部件，他讨论了真实性问题：“我们公司确实担心真实性。比如，人们喜欢使用名牌手表。对我们来说，当务之急是金属级3D打印正在实现。如果你从事飞机制造或外科手术工作，就必须确认部件的真实性。这是与设计师品牌手表相同的技术载体——包括与设计配套的数字版权。在安全的前提下，真实性是至关重要的。”

另一个围绕新技术的风险涉及劳动实践。目前，数字制造生态系统中的大部分工作都由志愿者或兼职临时工——所谓的“零时工”——来承担，新出现的商业化共享设计、产品和服务的机制也发生了

碎片化的工作需求。这是劳工经济学家戴维·韦尔所说的“裂隙经济”的一部分，在这样的裂隙经济中，公平薪酬、工作场所安全性、歧视的预防、表达机会与相关保护等政策目标越来越难以加强。目前的风险在于，“数制”工坊的工作将成为当前更大的社会挑战的一部分。这种风险因数字制造环境的不同而异，这些环境包括中小学校、学院、大学、博物馆、社区工坊、工业环境，某些环境涉及的工作的碎片化程度尤其严重。

一段时间后，围绕着劳动力和工作场所的实践将出现更大的风险，它将成为生态系统的特征。如果能将协作和其他建设性做法制度化，那就太棒了。风险在于，增加能力差异、减少合作的安排将占据主导地位。例如，分享设计和交换产品与服务的平台可能最终只会被少数人控制，这些平台为他们带来高额利润，而数百万贡献者的所得却不尽人意。这种情况将不利于建设一个更加自给自足和可持续发展的社会。

针对制造危险性物品或杀伤性物品的行为策略，也是一种风险缓解方式，其核心是建立跨越所有数字社区的规范，重点解决围绕数字文明和减轻不良行为的关键问题。这方面的一个有趣的例子来自一个看似不太可能的行业——视频游戏行业。一些较大的网络游戏社区普遍存在种族主义、同性恋恐惧症、性别歧视和其他毒害行为。由于商业和道德方面的原因，游戏开发商不得不直面这个问题。Riot游戏公司就在积极应对这一挑战，它开发的《英雄联盟》（*League of Legends*）游戏风靡一时。这个社区拥有近7 000万名玩家，规模庞大，具有全球性。Riot游戏公司发现，毒害行为令许多现有社区成员感到沮丧，还吓跑了新玩家。由于早期设计时并没有优先考虑这些扰乱社会的行为发生的可能性，公司不得不重新考虑其核心社区和设计假设。

Riot游戏组织社会科学家、神经系统科学家和游戏设计者来研究这个问题。他们详细研究了社区中是谁在推动这种毒害行为，这种行

为是如何传播的，以及它是如何影响社区的。他们还研究了良性行为的相同因素。然后，研究人员进行了公测，向游戏流程中整合入减少毒性并增加良性行为的机制，其中包括各种倡议。该公司还尝试在游戏中添加同行评议，以帮助玩家判断特定行为是否具有毒害性。另一项举措是提供奖励（连同信息和提示），以鼓励游戏玩家的良性行为。该公司还确保对不同类型的不良行为进行处罚，处罚力度根据违规行为的性质和严重程度来决定。他们还使用机器学习算法跨越不同语言和文化进行实时跟踪。这种方法的核心是在发生毒害行为时向玩家提供清晰的、具体的即时反馈（什么也不做，既不宽容又不强化毒害行为）。同样重要的是，这一过程既要透明又要谦逊（不夸大其意图），还要认真培养社区中更好的“天使”。

效果令人印象深刻。在接受《自然》[1]（*Nature*）杂志采访时，Riot游戏公司社会系统首席设计师杰弗里·林着重描述了当骚扰导致游戏中出现不良表现时，游戏内置的警示是如何减少毒害行为的：“与对照组相比，负面反应减少8.3%，言语辱骂减少6.2%，攻击性语言减少11%。”一个关于球员合作的正面提示，减少了6.2%的攻击性语言。在公司发布机器学习算法后，相对于其他排名游戏（电子游戏中最具竞争力的游戏）而言，语言毒害降低了40%。

Riot游戏公司已经表示，它将公布其成果供大家借鉴。在本例中，软件开发人员愿意“承担”风险，因为风险影响到他们的商业模式和客户体验。

在软件的设计和持续优化中鼓励积极行为，只是管理风险的众多内置技术解决方案中的一个例子。数字技术基础科学的一个关键方面涉及纠错的概念，尼尔在下一章中就此进行了更全面的探讨。对于人和社会的纠错要比技术纠错更难，不过Riot游戏公司的例子提供了一些可能的方式。顺着这些路径思考，可能有一天机器会感知到某些材料的误用，或者在出现新的误用形式时应用人工智能（AI）学习如何处理。当前，考虑方法和应对潜在误用事件仍然需要人来做出决定。

因此，任何缓解风险的策略都应该拥有一个重要内容，即一系列广泛的论坛和机制，专门用于跟踪和解决出现的问题，同时预测潜在问题，进而将社会进程与技术结合起来。

虽然Riot游戏公司的例子阐明了如何利用技术使行为走向更加文明的语言形式，但它假定了玩家对技术本身的最低接受程度。贝蒂·巴雷特是香槟－厄巴纳市社区“数制”工坊的联合创始人，也是针对美国“数制”工坊中的涉众队伍进行研究的团队领导者之一，她指出了一个更深层次的风险：“指数变化的方向并不总是向前或向上的。变化的速度可能出现倒退。世界上有些势力认为倒退符合他们的利益。这包括‘宗教激进主义信仰’和‘反智主义’。”她补充说，“对科学的强烈怀疑和不信任限制了变化的速度。这是一场规模宏大的战斗，数字制造将被许多人对技术变革的恐惧影响。”当然，具有讽刺意味的是，这些狭隘的看法又被科技手段放大了，其影响还在加速。巴雷特注意到这里有两个风险。首先，数字制造还没有准备好面对极为猛烈的意识形态阻力，而意识形态的存在本身就被视为一种冒犯。第二，世界某些地区将不会面临如此强烈的阻力，并将加速前进，进一步加深智造的鸿沟。

当我们进入第三次数字革命的后期时，随着扑朔迷离甚至已经暗流涌动的新风险的接踵而至，这些源于恐惧的挑战可能会深化。尼尔经常谈到机器是如何制造机器的，正如我们在以后的章节中所了解到的那样——装配机组装装配机，但他并不清楚人类如何融入这些召唤性图景。过程将如何开始和终止？它会停止吗？它会暴虐横行吗？当人们听说机器制造机器和装配机组装装配机时，他们很难不陷入“顽劣的人工智能肆虐人类”的想象。当我们考虑到人类的许多生存威胁时，人工智能与数字制造必然交织在一起。在这里，基本挑战始于如何围绕第三次数字革命的前景培养博闻强识且专注投入的人群。这项技术需要以一种吸引人的方式呈现，突出清晰明确、通俗易懂且触手可及的功能，消除真正的内心恐惧。正如我们即将在第6章深入探索的那样，我们必须将用脑（以逻辑为中心的方法）、入心（以情感为中

心的方法）和动手（实际动手方法）相结合，其目的是确保社会体系与加速进步的技术有效地协同发展。

第三次数字革命的潜能也引发了一个问题，即社会中将由谁来仲裁技术中蕴含的价值观。我们是相信技术人员，还是会邀请伦理学家、社会学家、说书人、宗教领袖和调解员？我们（几乎）可以设想出数字技术相关道德问题的解决进程，但更为困难的是使人们相信其解决结果的合法性，它必须要经得起历史的检验。

虽然“数制”工坊必须降低自身风险，但它们也可以在降低社会风险方面发挥作用。正如尼尔在第1章中所介绍的那样，迪娜·埃尔－赞法利于2011年埃及骚乱时，在开罗与人共同创建了一个“数制”工坊。事实上，它位于离解放广场不远的地方。在抗议的喧嚣中，“数制”工坊被证明是一个中立的、安全的地方。埃尔－赞法利回忆说，这是“一个充满活力和热情的气泡。它与革命有某种联系，但也是分开的”。即使数字制造带来了新的风险，“数制”工坊也能培育社区并促进创新，从而在抗议和变革的时候降低风险。

数字制造能力可以发挥降低风险的作用的另一个领域是由美国众议员比尔·福斯特提出的。福斯特重点关心国家安全问题，他指出军事和自然灾害的威胁可能会使经济大面积瘫痪。在这些场景中，他认为分布式制造是灾后重建一个运转良好的社会的重要能力。福斯特进一步描述了一种军事场景，其中包括在战争中更多地使用机器人。在这种情况下，入侵者可能在每个街角都安置军事化机器人。反击的能力很可能取决于当地与该国的数字制造能力，以产生应对意想不到的技术威胁的快速原型反应。

福斯特在国会提出了一项法案，阐明了“数制”工坊在全美的网络价值。除了国家安全领域，他还认识到“数制”工坊是一个促进学习、建设社区和鼓励创新的源泉。但对于国会中的许多人来说，当重点放在工坊网络有可能减轻社会灾难性风险上时，人们就会更加重视。

大挑战

社会面临着许多重大挑战，每一个挑战都伴随着全球风险。联合国有17项“可持续发展目标”，包括第三次数字革命可以推进的许多目标：扶贫；城市和社区可持续发展；负责任的消费和生产；体面工作与经济增长；减少不平等；工业、创新和基础设施；优质教育；和平、正义和强有力的体制。尼尔甚至针对这些目标开设了一个“数制”工坊。基金会、专业协会和全球组织都在阐述重大挑战，这是有原因的。这是因为大多数社会机构并不能有效解决这些复杂的、不断演变的、相互交叉的问题。正如经济学家、历史学家、诺贝尔经济学奖得主道格拉斯·诺斯在1990年定义的那样，制度“是由人类衍生的、影响人类互动的约束”。简而言之，它们是游戏规则，数字制造有可能成为社会机构的游戏规则。然而，要实现这一可能，首先要把握原有的体制环境。

许多现有的机构，如那些与教育、商业、社区和环境有关的机构，正在努力跟上技术加速进步的步伐。多数情况下，现有的机构代表着保守主义，为数字制造等新事物设置了潜在的障碍。第三次数字革命将同时让现有机构参与进来，并揭示系统中的漏洞——这些漏洞将被新出现的、更加机敏的人、平台和程序利用。

在调查中，我们要求智造先驱展望2025年，并预测数字制造将在多大程度上改变社会的关键要素——教育和培训、商业和创业、政府和民间组织、环境和经济的可持续性以及社区和文化。在这一群体中，有许多人在这些体制性安排下发挥着作用。在设计和制造方面，64.7%的人说自己扮演了教育角色，53.9%的人说起到了社区组织者的作用，35.9%的人说起到了公共领导的作用。在图2－6中，数字等级表示先驱者在数字制造对社会各个方面的影响的问题上的平均反应（在数字等级中，0表示“完全没有改变”，10表示“完全改变”）。

考虑到距离2025年还有不到10年的时间，那些最接近数字制造技术的人预计将会发生相当大的变化，其中教育领域和培训领域的变化最大，而政府和民间组织的变化最小。要知道，这些都是先驱者的预测，而非他们的意愿。预测的意义不是描述你想要什么，而是预测事情的发展方向，找出干预的方向，或者提示如何从现在做起，以实现共同的目标。因此，这一预测活动得出的结论是，那些最接近新技术的人预计政府和民间组织将在技术上落后。对于这些机构的领导人来说，这一预测应该被视为一种应当深入研究的挑战，实际上，结论证明这些机构领导人是错的。

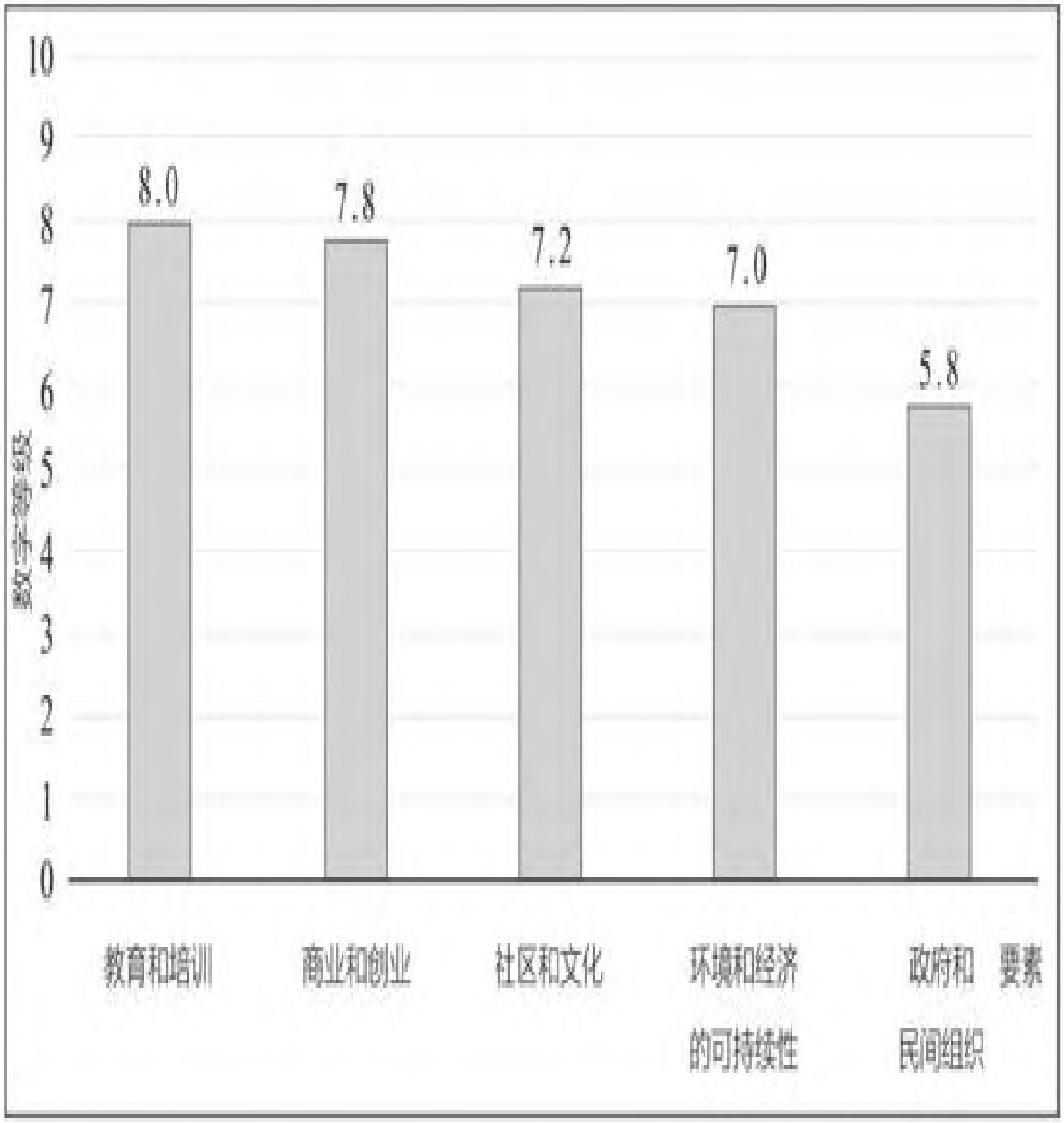


图2－6 预计到2025年数字制造会对社会产生的影响。WayMark Analytics公司绘

正如尼尔在第1章中指出的，联合国领导人并没有看到重大挑战与数字制造之间的直接联系。但是，当他们意识到这一点时，他们对数字制造的整体态度从认为其可有可无，转变为把它看作一个刚性需求。要使第三次数字革命真正成为一场革命，不仅仅需要在技术性能

方面表现出指数级的增长。这项技术必须明确地解决长期的社会需求，其模式是可复制的，并在当地适应广泛的影响。仅仅符合联合国可持续发展目标的概念并不意味着这就是现实。

正如前白宫科技政策办公室副主任、创客和智造运动的有力倡导者汤姆·卡利尔所指出的："创客或'数制'社区帮助解决特定社会问题的能力，是衡量进步的一个关键尺度。这不仅仅与自我表达有关。"他接着列举了一些例子，比如创客和残疾人共同努力设计辅助技术，以及创客和教育家共同努力降低学校辍学率。

虽然在社会重大挑战层面的有形影响方面取得进展，但还存在涉及无形影响的风险。德里克·曼西尼曾担任阿贡国家实验室纳米材料中心的设计和建造项目主管，他说："人们被数字制造所吸引，是因为它是技术加速进步引起的一种反应。工业革命曾经引发了一场手工艺运动，这种情况与人们被数字制造和创客运动吸引的情况差不多人们在接触创客技术时，就会表现对个人控制的深切渴望。"这样，数字制造的兴起对于人们在加速变革的时代实现更大的个人控制提供了巨大的机遇。

总之，令智造生态系统悲喜交加的，是对第三次数字革命的未来至关重要的4个基本挑战：智造获取、智造素养、培育赋能的智造生态系统和降低潜在风险。此外，还存在持续的矛盾，例如技术的易用性的平衡和对"黑匣子"解决方案的避免；将对可互操作的标准和协议的需要与社区的分散性和独立性相协调；对不同涉众之间的无等级决策和集体行动的各种需要的平衡。我们应对这些挑战和矛盾的综合能力，很可能会决定我们将平衡推向怎样的未来——一个所有人都能智造（几乎）任何事物的乐观的未来，或者一个一些人可以智造（几乎）任何事物的未来。

（本章由艾伦、乔尔撰写）

[1] 《自然》于1869年在英国创立，在学术界享有盛誉，常与美国的《科学》杂志并称为世界两大顶级学术刊物。——译者注

第3章

背后的科学

想要了解数字制造的未来，我们必须要先理解它的过去。根据不同的计算方式，我们通常认为第三次数字革命是在2000年前后随着创客运动的兴起和第一个“数制”工坊的成立而发生的。我们也可以将其发生时间追溯到1950年左右，即早期的电脑在麻省理工学院第一次成功连接一台制造机器的时候；如果再往前推算，其开始时间可以追溯到40亿年前，即生命演化出“分子制造”机制的时候。

在各个历史时期，无论你选择哪个时间起点，数字制造都与通信的发展和计算的数字化有着密切的历史相似性。三次数字革命有着相同的核心理念。它们教给我们的共同的一课是在其演化路径变得清晰与技术到达最终形态之前，其含义都可以被理解与实践。过去发生的只是第三次数字革命的序章，就像我们在本章内容中描述的那样，我们会跟随摩尔定律在过去几十年间的发展轨迹，与现在正在发生的拉斯定律接轨。

从摩尔定律到拉斯定律

1965年，戈登·摩尔发布了历史上最著名的图表，这个图表知名到被认为定义了通信与计算领域的数字革命的程度。当时，《电子》（*Electronics*）杂志希望摩尔对下一个10年内电子零部件的发展进行预测。那时，摩尔是仙童半导体的研发部主任，他在1968年的时候还协助他人成立了英特尔公司。

他通过5个数据点就看到了未来。把集成电路上的器件数量与时间作为参数绘成折线图，结果表明器件数量几乎每年都翻倍。经典的翻倍原理是增殖。一个细胞会分裂成两个细胞。之后，这两个细胞会分裂成4个，4个细胞会分裂成8个，之后是16个，以此类推。这个数列可以写成一系列关于2的指数规律的式子：$2=2^1$，$4=2\times2=2^2$，

$8=2\times2\times2=2^3$，$16=2\times2\times2\times2=2^4$……对数函数与指数函数相反，它计算指数函数中的“指数”：$\log_2 2=1$，$\log_2 4=2$，$\log_2 8=3$，$\log 216=4$……这就是为什么对数的线性增加（1，2，3，4）与被测量数字的指数级增加（2，4，8，16）相对应。这也是为什么在摩尔把晶体管数量的对数与时间作为参数绘成的折线图中，数量翻倍的趋势会表现为一条直线。

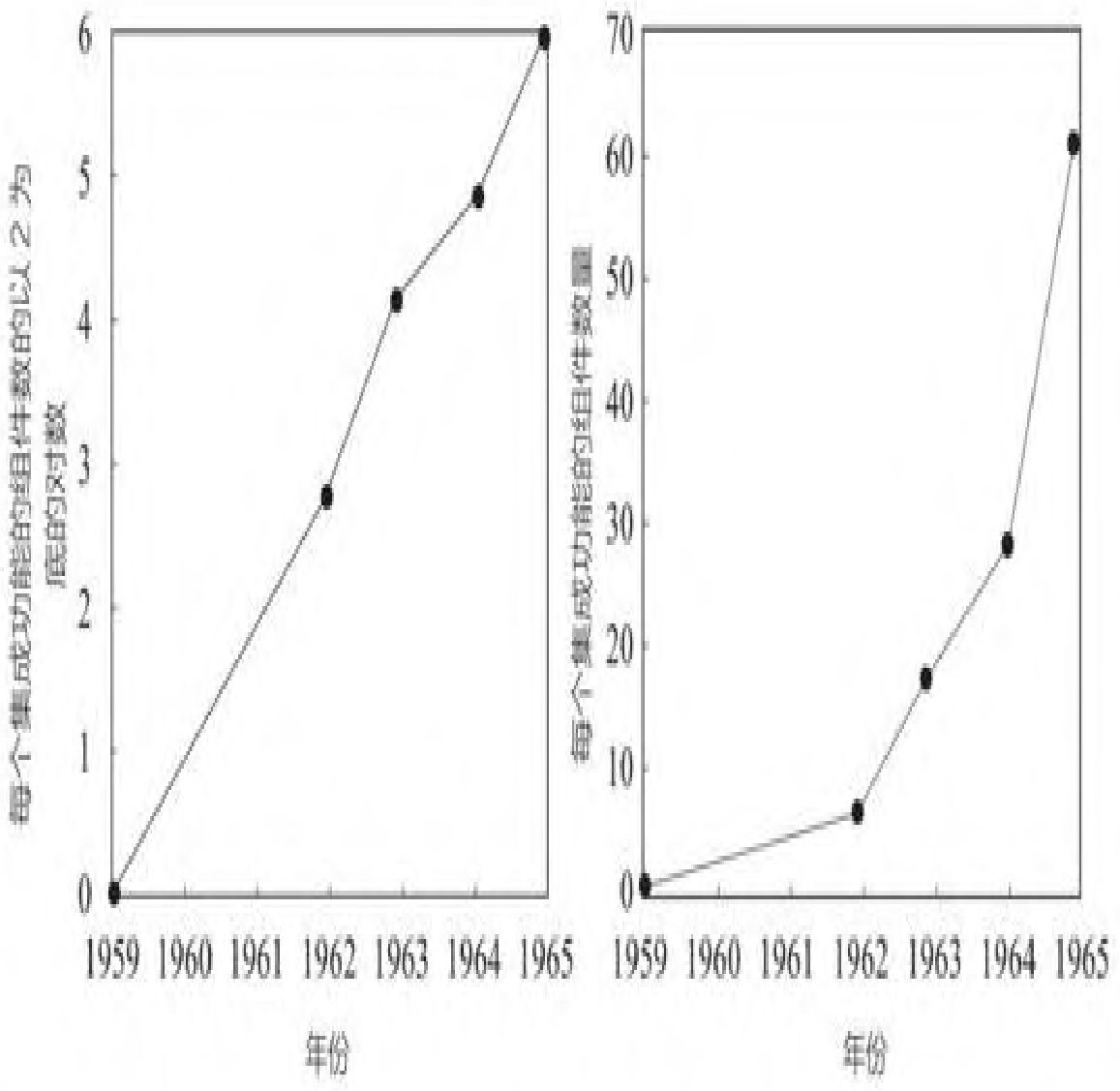


图3－1 戈登·摩尔的原始数据，用对数和线性尺度表示。尼尔·格申斐尔德绘

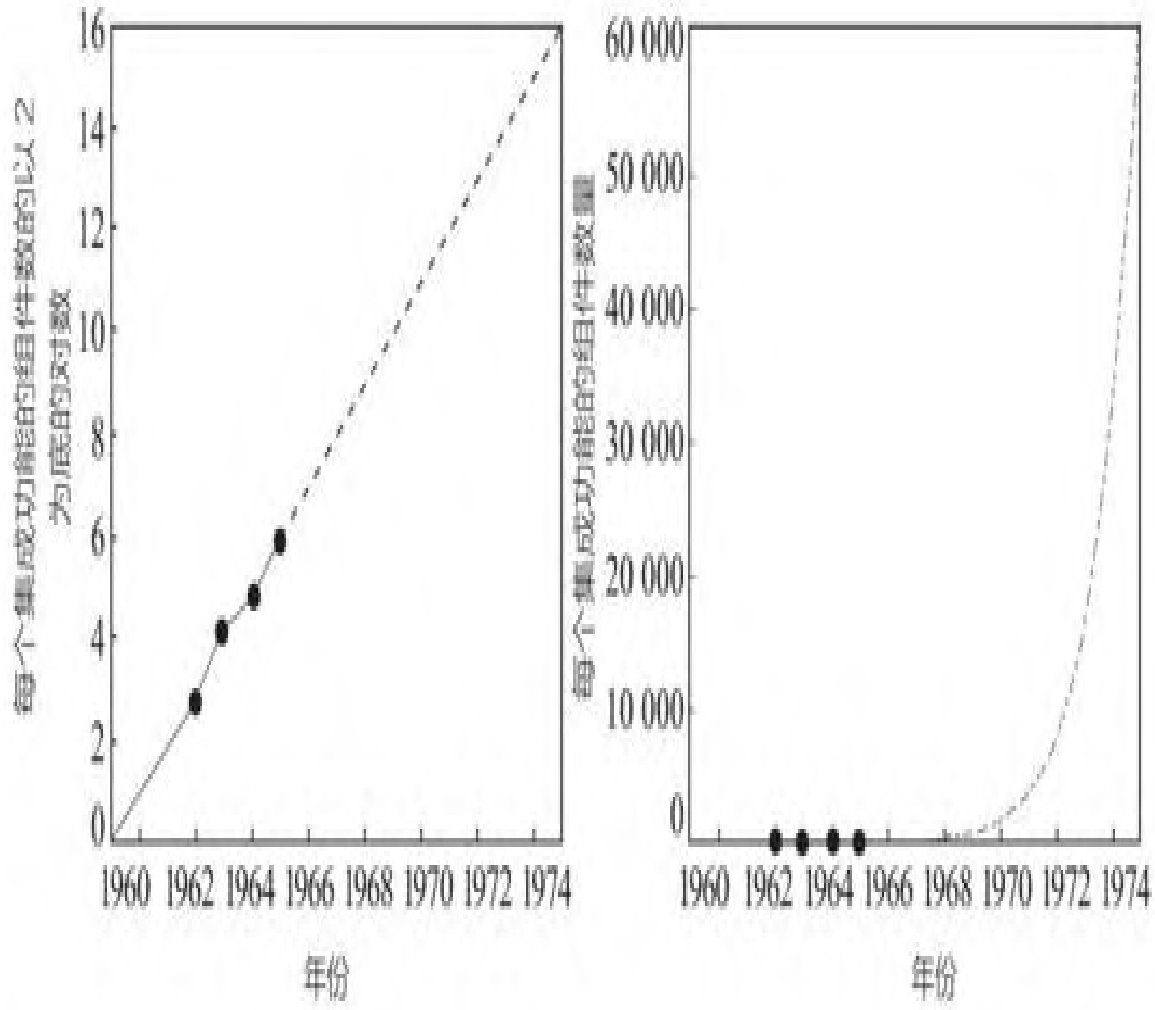


图3－2 摩尔定律的初始投影，用对数和线性尺度表示。尼尔·格申斐尔德绘

我们的许多生理感知只对对数变化产生反应。因为我们的身体将光线强弱的翻倍或者声音大小的翻倍看作一个等量增加，所以我们才能从一个昏暗的房间走入一个敞亮的房间而不被刺瞎眼睛，或者从一个安静的门厅走入一个吵闹的剧院而不被破坏耳膜。但是，我们对世界的感知更多的是建立在线性关系上的。银行存折显示的是金钱数量的变化，而非数量级的变化；时钟描述的是时间的流逝，而非时间的加倍。在只有5个数据点的情况下，对数图与线性图看起来很像。但是这种关系被称为摩尔定律而非摩尔曲线，是因为他大胆地推断这种直

线趋势会在接下来的10年中一直持续。如果将晶体管的数量画在一个线性图上，你就会发现，在一个突然的转折点之前，似乎什么都没有发生。但是，对数图能够让你很清晰地看到，晶体管的数量每年都在翻番。

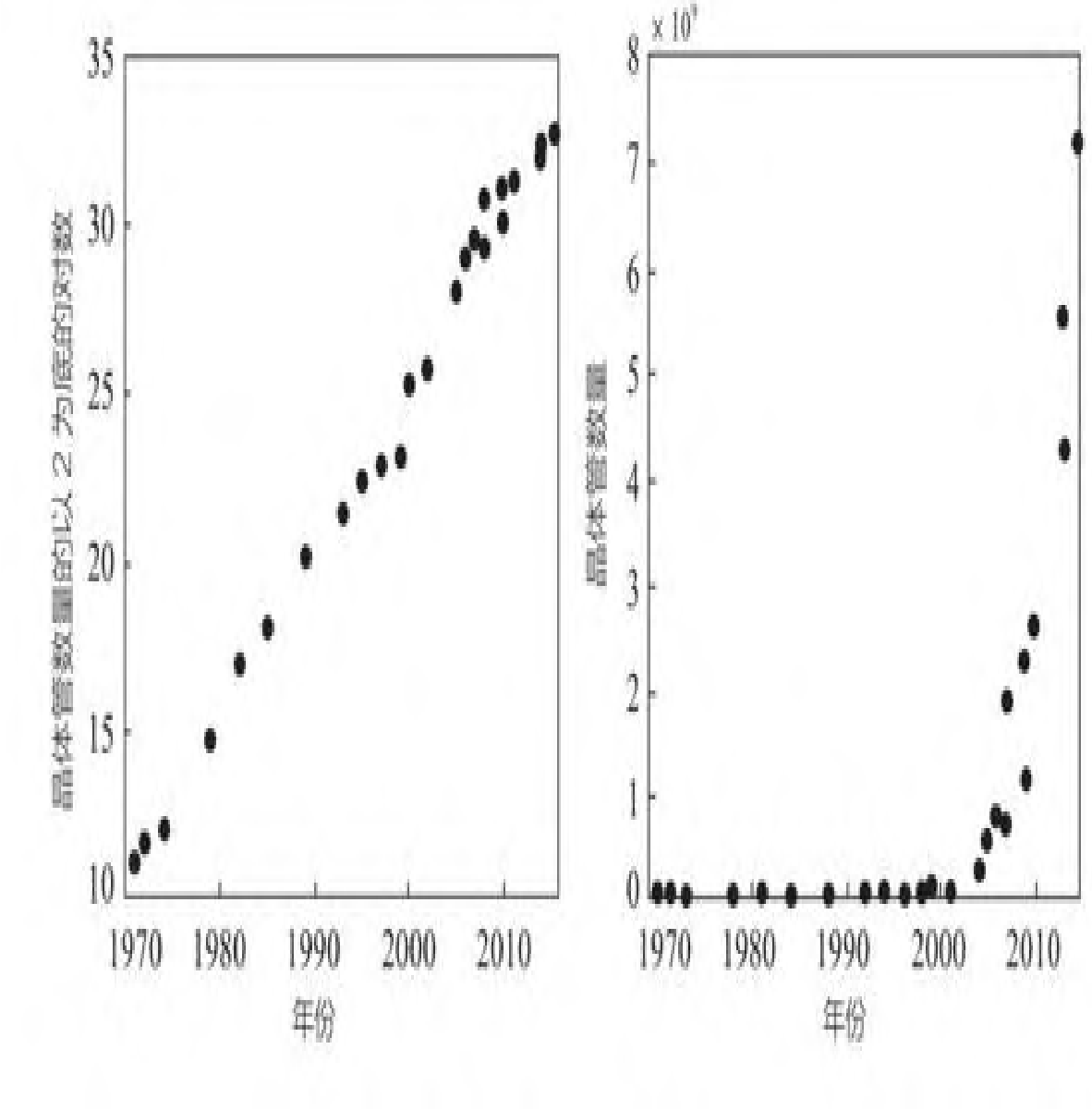


图3－3 英特尔微处理器中的晶体管数量，采用对数和线性尺度。尼尔·格申斐尔德绘

摩尔几乎将这个趋势完全说对了。翻番的趋势实际上一直持续了50年。摩尔定律的核心可以体现在英特尔生产的微处理器中的晶体管的数量上。在接下来的40年中，晶体管的数量持续翻番（直线在对数图上延伸）。如果你用线性关系画图，你大概会得到这样的结论（除非你进行了仔细研究）：在先前的数十年中，几乎没有什么转折发生，但是在2000年左右，一个数字技术的革命发生了。这大概是线性关系带来的认知，但是对数关系图很清晰地表示了每年翻番的趋势延续了数十年。

什么是在集成电路中复制？最佳答案是间接的：比特是。模拟替代方案的表现与其不同。在这本书的序言中，我们说过，如果我们往复印机中放入一张纸，会得到两个复印件。如果这两个复印件被复制，会得到4个复印件，再之后是8个，以此类推。但是每一次复制都会使被复制的信息稍微退化，直到后面的被复制的信息变得一团糟。持续的翻番只有在能够进行错误修正的情况下才能进行。细胞会在其分裂的过程中自动检测基因组复制过程中的错误并进行修复。现在世界上最大的集成电路有数十亿个晶体管，其中每个晶体管的信号都会失真，但是由于集成电路会持续修正，恢复比特的逻辑水平，因此错误的概率会指数级下降。这种错误修正机制可以保证晶体管数量的指数级增加，所以第10亿个晶体管的输出与第一个一样可靠。

晶体管并不真正地制造晶体管，但是抽象地讲，它们的确如此。它们被生产成芯片，然后用在电脑中，设计更好的电脑；它们被用来发送消息，从而使更多的消息被发送；它们被用来编写编写程序的程序。在这个反馈循环当中，不只是电路中晶体管的数量呈指数级改进，还有其大小、速度与成本。这种改进导致了通信与电子系统性能的飞跃性提高，包括内存的大小、网络的速度与连接起来的电脑数。因为摩尔自始至终都没有把他的观察命名为摩尔定律，所以对于摩尔定律到底测量的是什么有不少争议，而最佳答案是信息科技的性能。摩尔定律的重要性不在于其定义，而在于数十年来数据点始终与最初趋势相吻合。

以时间为横坐标，以信息科技的性能为纵坐标的对数图表现为一条直线，但是这不意味着其代表的技术本身的发展是线性的。相反，其发展往往分为不同阶段。通过前驱程序发送到互联网的第一条信息是在楼房那么大的大型计算机之间传递的；当互联网连接了约1 000个站点时，计算机发展为房间那么大的小型机；在拥有100万个站点时，计算机发展为可放在书桌上的个人电脑；在拥有10亿个站点时，计算机已经发展为能装进口袋的智能手机。同样的事情发生在这些机器内部的所有部件上。

集成电路一开始是用手工切割的掩模版（掩模版准确地定义了硅片上蚀刻和堆积材料的位置）在硅片上制造的（这是我最初学到的一门失传的艺术）。之后这一步被自动化了。掩模版先直接接触晶片，然后变成了用还原光学器件进行投射。曝光步骤一开始是用可见光，而后来使用的是短波长极紫外光，等等。由于以上每项技术的改进都具有极大难度，在摩尔定律持续发展的每个10年中，人们都做出了这个定律即将终结的预测。然而，每次技术升级又确保了摩尔定律的延续。

指数增长的确最终会到达其资源极限。专业术语是Sigmoid函数，表示为一个S形的曲线。它是从一个平缓增长的区域（S的底部）开始的，之后是一个加速增长的中间段（S的中间），之后是一个平缓增长的区域（S的顶部）。在一个生物细胞的增殖过程当中，增殖的细胞数量最终会被环境的空间大小以及养分的多少限制住。摩尔定律已经超出了物理学的范围，因为它达到了极限。在2015年左右，翻番开始减速。在大规模生产的芯片中，每个器件的构造都如此之小，以至每个单个原子的排列都必须被追踪，而在实验室内已经出现能够这样做的单电子晶体管，其开关取决于单个电子是否存在。一个晶体管不可能比这个尺度更小了，除非我们开始用原子核编程，但是到目前为止这种控制原子核的能力只在高能粒子加速器中存在。

摩尔定律也同时限制了用户数量的增长。当几乎每个人都可以接触一台连入互联网的电脑时，驱动这些设备发展的需求将不复存在。电脑与通信设备将会变成一般的缓慢取代的商品而非快速增长的新鲜事物。如果说用户就像让计算机产业茁壮成长的养料，接入用户数量的S形曲线在达到极限之前就已经开始减缓，正如乔尔和艾伦解释的那样，电脑连接最后几十亿人花的时间将会比之前连接几十亿人的时间要长很多。

摩尔定律的终结是科技界和经济界的一个焦虑的来源，因为摩尔定律对应着一段生产力快速提高的时期。假如技术停止增长，经济也会随之如此。对于绝大多数正在被海量信息轰炸且不再被计算机性能局限的消费者来说，这不是他们所关心的。

饱和的需求也将影响被称为“摩尔第二定律”的事物：芯片工厂的成本。摩尔最初并没有考虑这一点，但是最前沿的芯片工厂的成本已与其他数字世界的一切一样飞涨，这一代设备现在要消耗至少几十亿美元的资金，当前一代芯片的开发可能要花费一亿美元。一旦这些投资用完，晶体管的制造成本就会降至比一美分还要小的数字，但是如此海量的资本支出事实上禁止了除产业最大玩家与最大应用市场之外的玩家与市场。

芯片工厂耗资的不断上涨是因为芯片制造在本质上是一系列非数字化的模拟的步骤。材料被连续不断地沉积或者蚀刻，在这些步骤中对于品质与公差的要求非常苛刻，因为任何瑕疵都可能毁掉整个芯片。在这个工业中最核心的机密就是良品率——在最尖端的制造工艺下总是会有极少数的芯片质量不过关。

第三次数字革命的愿景是将生产制造从摩尔第二定律移至摩尔第一定律。大约在2003年，为了建立第一代“数制”工坊，我去了英特尔，见到了戈登·摩尔，和他讨论其可能性。他从一个不太大的格子间座位（所有英特尔管理层都是如此）起身来到我们的会议室。他对于自己在技术发展中扮演的角色表示非常谦虚；他没有自己命名摩尔

定律，并且他并不轻易使用“摩尔定律”（这个名字来自一个加州理工学院的合作者——卡弗·米德）来描述那个法则。英特尔对于摩尔定律的遵循体现为一个清晰的组织分界线，这个分界线可以区分为集成电路开发程序和发展逻辑的计算机科学家与为集成电路开发器件和发展制造工艺的物理学家。计算机是数字化的，但是制造工艺是模拟的。

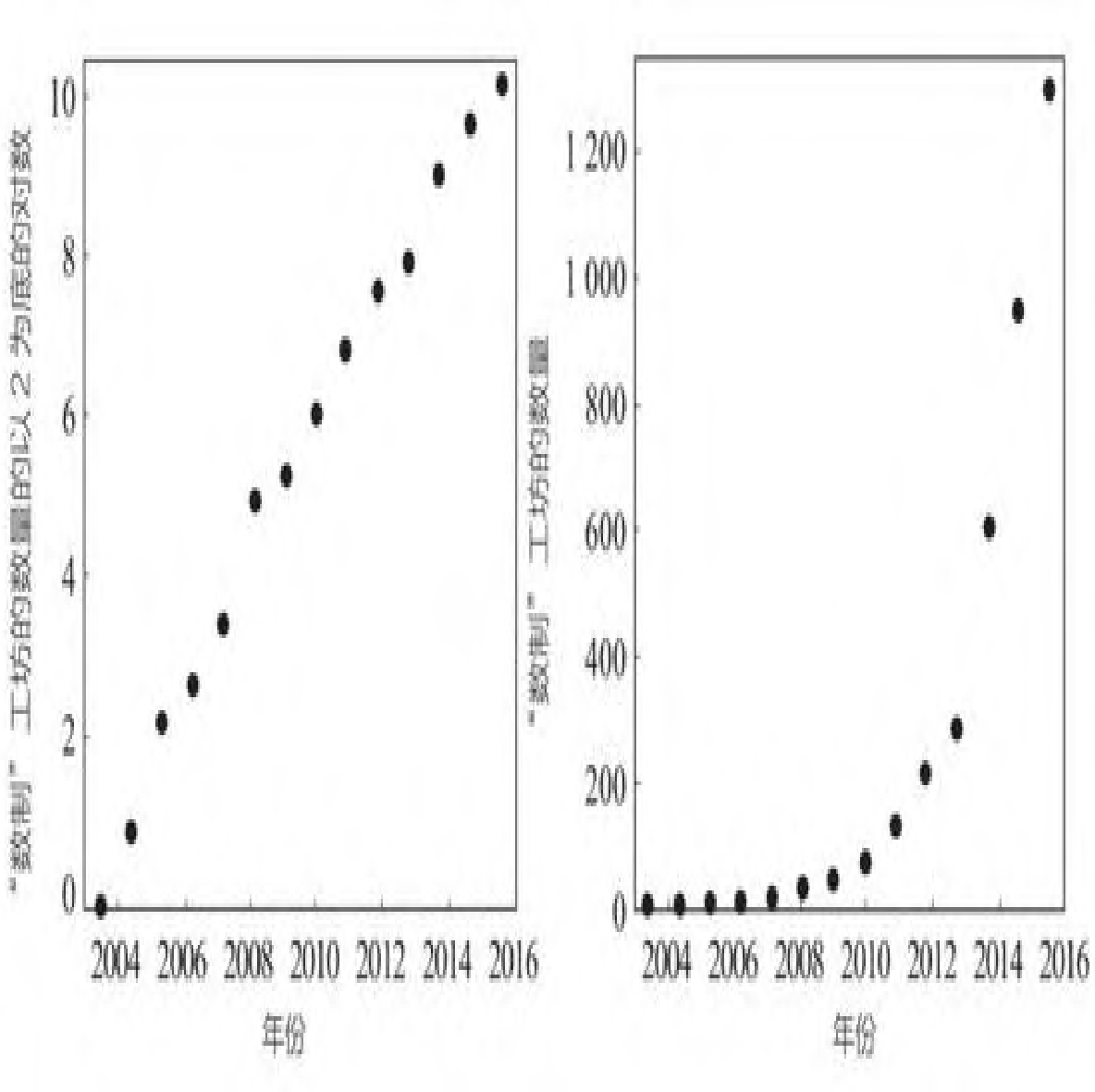


图3－4 “数制”工坊的数量，用对数和线性尺度表示。尼尔·格申斐尔德绘

正如在之前的章节里描述的那样，我第一次认识到在数字制造里面或许有类似摩尔定律的趋势发生是在雪莉·拉斯特（比特和原子中心的“数制”工坊项目及智造基金会的管理者）注意到世界上“数制”工坊的数量大约每一年半翻一番的时候。一个决定性的观察就是她桌子上有关“数制”工坊的文件的高度在不停翻番。我们开玩笑说，这就是一个类摩尔定律，我们称之为拉斯定律。就像摩尔定律一样，这个定律已经持续了比我们预想的长得多的时间。经过十多年的发展，现在大约有1 000间“数制”工坊，这是经过10个翻倍周期得到的结果。“增殖”的原因很简单，当一个“数制”工坊开张的时候，周围的人就受到了启发想开另一个，而这个过程当中的错误纠正机制是每个“数制”工坊都有一种功能相同且不断演化的核心能力，因此操作人员可以在不同“数制”工坊之间对各个项目进行共享交流。

“数制”工坊的发展始于2003年建立的第一个“数制”工坊，而非20世纪50年代计算机控制制造的发明。在20世纪50年代，一个“数制”工坊的规模与造价相当于运行所有机器所需的10台左右的计算机的规模与造价。感谢摩尔定律的延续，只有当数字计算与通信的造价、规模与性能缩减到与一台机器相当时，“数制”工坊才变得可行。

正如摩尔定律的发展刚刚越过其第一个10年时一样，预测“数制”工坊数量翻倍的趋势将持续延续下去，不仅是大胆的，还是明智的。这意味着在下一个10年中，“数制”工坊的数量将会超越100万，并且在之后的10年内达到10亿。正如互联网的扩散，拉斯定律不意味着10亿个房子被“数制”工坊填满，相反，每一个10年都会是一个技术转换的新时代，智造（几乎）万物的技术可以集成和使用。

就像很多信息科技中的量最后被归纳入摩尔定律一样，拉斯定律也很可能被应用到很多描述数字制造发展的方法当中。这些方法涉及了被组装的部件的数量、组装的速度、一个部件的尺寸与成本，以及组装工序的复杂程度。当互联网连接了10亿台电脑时，这些电脑内部

芯片的晶体管数量达到了10亿；当世界上有相当于10亿个“数制”工坊时，它们很可能会制造出具有10亿个部件的物体。相信这个规律会变为现实就好像相信摩尔定律会变为现实一样荒谬。在第5章，我会描述到达那个未来的路径。

拉斯定律可以被视为一种对摩尔定律的延续而非替代。摩尔定律存在于一个平面上，即一个集成电路上；拉斯定律将其从二维转入我们生活的三维世界。摩尔定律被用于描述一个芯片被制造之后具有的性能；拉斯定律将其延伸到如何在第三次数字革命中制造它们以及其他一切物体。第三次数字革命的结论是，为计算的未来长期寻找的杀手级应用是制造。

通信、计算与制造

摩尔定律可以被视为类似于经济学家所说的宏观经济学或物理学家所说的热力学的一个系统的总体属性。如果要理解摩尔定律与拉斯定律之间的历史相似性，我们就必须将目光放长远，越过指数增长图，研究等价于经济学家所说的微观经济学或物理学家所说的统计力学的微观元素，即构成一个系统的单个组件。所谓单个组件，就是前两次数字革命中的计算与通信机器以及第三次数字革命中的制造机器。

当计算机首次被大型机构制造出来并且用于大型机构的时候，它的体积可以填满整座建筑。现代交互式计算机的架构可以追溯到第一台能够对输入指令进行实时响应而非分批处理的大型电脑上。旋风Ⅰ型计算机是在麻省理工学院的伺服机构实验室中被开发出来的，可用于替代控制飞行模拟器的模拟计算机。后来，它成为半自动地面防空

系统（SAGE）的基石。旋风Ⅰ型计算机于1947年开发，于1951年开始服役。它耗资数百万美金，其机身填满了一幢大楼的好几层。



图3－5 麻省理工学院第一台计算机控制铣床。麻省理工学院图书馆，学院档案馆和特藏，J.弗朗西斯·赖因特耶斯论文，MC-0489，专栏8

1925年，第一个计算机控制的制造机器就是这个计算机项目的一个分支。一个叫约翰·帕森斯[1]的发明家给麻省理工学院提交了一个提案：将一台实时计算机连接到一台铣床上，从而使铣床能够制造新型喷气式飞机需要的越来越复杂的零件。伺服机构实验室采纳了这

个建议，但是没有接受他，他们将约翰·帕森斯从空军项目甩开后开发了我们现在称之为数控铣床的设备。计算机可以连续地在三维的各个方向上精确控制旋转研磨轴，从而让机器加工出技术工人用双手不可能加工出的形状。此后，很多不同的数控制造工具被用于机器的各个节点，以代替铣床的研磨轴，但是基本原理自从1952年开始就没有变过。

麻省理工学院的旋风Ⅰ型计算机的继承者是TX－0，然后是TX－2，TX－2是第一台使用晶体管而非真空管的具有重大意义的计算机。这个差异的重要性表现在很多地方。第一，晶体管可以被放置在更小的空间里；第二，晶体管不会像真空管那样产生大量的热；第三，晶体管运行时需要的能源更少；第四，晶体管的可靠性更高。这些优势叠加在一起，才使得计算机更小，更有效率，能够进行更加快捷的运算。

TX－0与TX－2于1956年首次投入使用，由肯·奥尔森领导的工程团队在一年后将它们商业化并且作为美国数字设备公司的PDP（编程数据处理器）系列进行销售。编程数据处理器使计算的成本由数百万美金下降到数十万美金，最后到达数万美金，而一台电脑的体积也从一栋大厦那么大降至一个房间那么大，最终降到一个器械支架那么大。然而编程数据处理器并不容易使用——我在学习如何使用的时候，需要掌握如何正确连接不同的单元，比如处理器、内存、存储器、输入、输出、通信和电源。由于它已经可以被小型工作组负担得起而不像以前那样只有大型组织才能拥有，像我这样的人才有机会接触到一个。今天你能用到的几乎一切计算机应用最开始都是在编程数据处理器上完成的——从文档编辑到玩电脑游戏，再到通过互联网发送信息。美国数字设备公司沿着波士顿128号公路发展了一个计算机产业，并且把这里变成了计算机世界的中心。与美国数字设备公司一起，所有的主流小型计算机制造商都坐落于此，包括Prime、美国数据通用公司（Data General）、阿波罗电脑公司（Apollo）与王安电脑公司（Wang）。

从制造的角度来讲，今天的“数制”工坊在构造成本以及复杂性上已经达到了一台编程数据处理器的级别，同样是一个房间那么大，最初成本为大约10万美金。就像编程数据处理器作为一个系统集合可以被看作我们现在所说的一台电脑一样，“数制”工坊可以被看作一个数字制造的系统。“数制”工坊可以将数据转化为物体，也可以把物体转化为数据，但是目前这些工作必须在实验室内部完成，而且我们必须要具备使用多个不同机器的能力。正如编程数据处理器一样，现在一个“数制”工坊可以被一个工作组或者社群而非一个大型组织所拥有，它们可以被用于探索当每个人都能接触到数字制造时数字制造如何被大众使用。

第一台真正的个人电脑是阿尔塔计算机（Altair 8800）。这台电脑是由阿尔伯克基科特兰空军基地的工程师为微型仪器与遥测系统公司［MITS，他们故意选择这个与麻省理工学院（MIT）相近的缩写］设计的，它被用来制作火箭模型爱好者需要的套件。当这个为爱好者而做的套件于1975年出现在《大众电子》（*Popular Electronics*）的封面上时，它给予很多杂志读者（包括我——我仍然记得当时我坐在我父母的汽车后座上，双手紧抓着新一期的杂志震惊得合不了口的样子）一个类似命运转折点的体验。

当阿尔塔出货的时候，唯一一种将程序写入的方式是通过数据来翻转前板上的开关，同时，唯一一种得到结果的方式是观察前板上的灯泡闪烁情况。但是，由于价格降低到约1 000美元，它变成了一台普通人就可以拥有的电脑。比尔·盖茨和保罗·艾伦在哈佛大学使用编程数据处理器开发了微软的第一个产品（微软一开始的名称是Micro-Soft）——一个为阿尔塔开发的BASIC语言解释器。阿尔塔计算机的到来也成为门洛帕克市的家酿计算机俱乐部第一次聚会的推动力，启发了史蒂夫·乔布斯与史蒂夫·沃兹尼亚克推出苹果电脑。

当这一切发生的时候，阿尔塔却被主流小型电脑工业忽视了。肯·奥尔森在1977年做出了一个著名的论断：“普通人没有任何理由在

家庭里拥有一台电脑。”其实他的言论很大程度上被误解了，他所指的是字面意义上的在房屋建造过程中使用的一台电脑，而不是一台被居住者使用的个人电脑。但他仍然错了，他没有预见到智能设备在家庭中的迅速普及。所有的小型电脑制造公司在个人电脑出现时期都失算了，因为个人电脑被它们视为毫无竞争力的玩具。正如我在第1章中提到的，美国数字设备公司最终被卖给康柏，之后康柏又并入惠普。回顾往事，组织变革的教训是小型计算机行业注定要衰落；在这种情况下，最佳的决策就是完全放弃而非苦苦挣扎。

数字制造领域今天正在经历一个可以与阿尔塔计算机的历史相对照的阶段。一些“数制”工坊内部的项目正在制造快速原型机器的快速原型。这些都是机器的“数制”工坊版本，也能够被其他“数制”工坊机器所制造。正如小型电脑行业对阿尔塔计算机的反应，制造业对于“数制”工坊一贯的反应是这些机器是为满足教育或者娱乐需求而造的玩具，它们不会对严肃的制造业构成影响。但是，这些工具已经能够在当地生产之前消费者只能在全球供应链上购买的产品了。正如个人电脑导致了小型计算机市场的死亡，“数制”工坊极有可能创造很多不会再回到旧工厂的新工作机会。

1977年和1981年，苹果Ⅱ型电脑和IBM个人电脑相继问世，个人电脑的现代形式由此确定。个人电脑以一个具有开关按钮的单一设备的形式出现，但是从内部构造来讲它集成了房间大小的编程数据处理器所具有的很多子系统。现在，能够与一台个人计算机相类比的个人数字制造设备仍然不存在。它不会是一台3D打印机，虽然很多人对3D打印过于看好并且暗示它就是未来的个人制造设备，但它仅仅是一个“数制”工坊里只能制造最简单的成品的生产制造设备之一；它也不会是将“数制”工坊所有机器压缩到一个盒子中的集成机器，因为这仍然需要大量的原材料与废物处理的能力。相反，我们会看见一个真正的个人制造器，就像《星际迷航》中的通用复制器那样，从更加根本的角度用数字制造的方式为物质编码。

从乐高积木到核糖体分子模型

数字是一个使用最广，也被误解最深的词之一。数字在通信、计算与制造当中的含义具有比使用“0”和“1”更加深层的共同起源。

比较一下我最喜欢的制造系统之一（孩子玩的乐高积木）与我最不喜欢的制造系统之一（今天的3D打印机）：

可靠性：因为乐高积木的错误纠正机制来自将积木安装在一起的步骤，乐高积木的结构要比一个孩子对运动的控制更精确；3D打印，从另一方面来讲，其准确度只与打印喷头的位置相关，而且一旦在材料流出打印喷头时出现干扰，打印就会失败。

模块化：用不同材料制成的乐高积木可以用一个标准接口组装；但是3D打印需要使用经过相同的沉淀工艺制得的材料。

局域性：孩子不需要通过尺子来放置积木，物体的整体构造取决于区域部件，这意味着一个孩子可以制造比自己大的物体；而3D打印则受制于其固有打印面积的大小。

还原性：乐高积木不像3D打印的物体，它不会被丢弃。丢弃本身是一个模拟概念，意味着该物体没有被拆解、还原成原材料的步骤。

可靠性、模块化、局域性和还原性这4个属性是第三次数字革命中通信、计算与制造的数字系统的必要组成部分。我们接下来会逐一观察这些系统的数字化过程。

1931年，范内瓦·布什[2]在麻省理工学院建造了最后一个伟大的模拟电脑：微分分析仪。它由填满整个房间的齿轮和滑轮组成，是一台可以解不同工程方程式的机器。它运作的时间越长，运算结果与正确答案误差就越大。

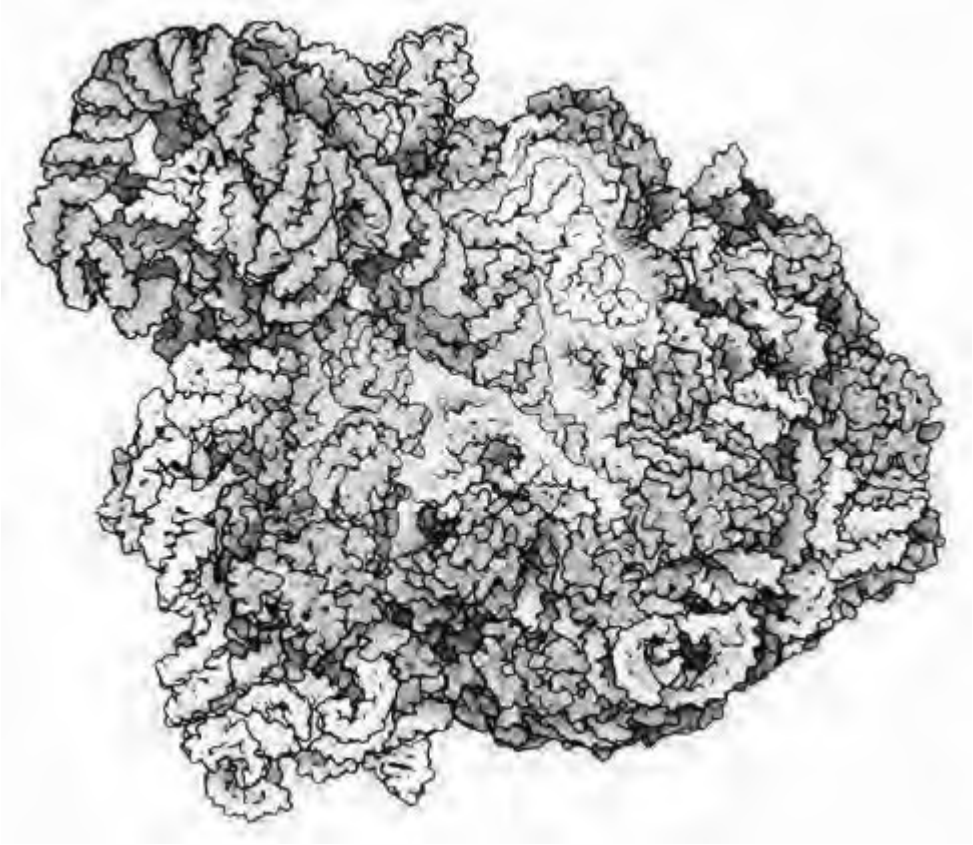


图3－6 运动中的核糖体分子模型。戴维·古德塞尔，RCSB（结构生物信息学）蛋白质数据库提供

你今天使用的数字计算机不会出现这种状况（希望如此）。你可以为此而感谢数学家约翰·冯·诺依曼[3]，是他展示了如何使用不可靠的器件进行可靠的计算。《概率逻辑学和由不可靠原件合成可靠的有机体》是一部基于他在1952年做的一系列讲座编撰成的著作，在这本书中他证明了我们现在称为阈值定理的理论。他演示了当用离散的符号（比如0和1）而非连续的物理量（比如齿轮和滑轮的旋转与运动）进行计算时，如果一个器件在计算机当中出现错误的概率大于一个阈值（这个值取决于系统设计），这个计算就一定会发生错误；但是，如果发生错误的概率低于这个阈值，计算发生错误的概率就会随着检查结果的操作被复制的次数的增加而呈指数级下降趋势。错误发生的概率可以被迅速地减少，甚至在运行中很可能不会出现一次。错误仍然会发生，但是非常罕见。在所有的工程领域中，只有极少数见

解可以实现性能的指数级提高，冯·诺依曼的这个定理是最重要的一个，并且是数字化在计算、通信以及制造中的真正意义。

冯·诺依曼的成果是建立在克劳德·香农[4]的贡献之上的，他们于1940年在新泽西州普林斯顿的高等研究院中相识。1937年，香农在麻省理工学院写下了可能是有史以来影响力最广的一篇硕士论文。在为布什进行微分分析仪的研究之后，香农意识到为调控电话而制的机电继电器可以被连接起来，使用布尔逻辑函数运算真命题和假命题。他不是简单地论述了一下这个理论，而是演示了如何设计这些通用逻辑电路，这些电路是构成之后所有计算机的基础。

从麻省理工学院毕业之后，香农转移到了贝尔实验室，那里对电话系统的研究具有更大的需求。他在那里的研究并非渐进地改进现有系统，而是回答了一个之前没人问过但是极其重要的问题：什么是信息系统的基本范围？在1948年出版的《通信的数学原理》一书中，他提出了关于信息的理论。在展示信息如何被压缩的同时，香农介绍了为通信设计的阈值定理。假设一个信息在电话系统中是以符号的形式而非连续模拟信号的形式发送的，并且电子噪声大于阈值，那么这个消息可能会失真。但是，如果电子噪声小于阈值，在接收端接收信息时发生错误的概率就会随着检查结果次数的增加而呈指数级下降。重要的不是0和1的具体数值，而是使用这种离散的数值可以让其错误被检测到。很多数字系统，比如一个手机系统，都在使用一些其他的、更加有效率的符号。你的手机现在正在接收一组被用来检测错误的符号，这组符号有26个。

电话系统花费了大约10年时间从模拟系统转换到了数字系统——这是一个导致互联网诞生的转换。鲍勃·拉齐是在贝尔实验室从香农时期一直工作到我加入时的一位睿智的经理，他在一次关于组织变革的发人深省的课堂上解释道：关于模拟与数字阵营的斗争在那时就一直在内部延续。这个斗争不是因某一方被说服而结束的，而是随着时

间的推移自然而然地解决的：所有支持模拟系统的经理全部离开岗位，新一代支持数字系统的经理掌权。

香农在贝尔实验室的同事约翰·图基创造了单词bit（比特）作为“二进制数字”中用来表示信息的最小单位。但是比特的概念有一个物理起源，它第一次出现在1929年利奥·西拉德[5]对麦克斯韦妖的分析中。在这个由物理学家詹姆斯·克拉克·麦克斯韦于1867年构想的思维实验中，一个微观的妖怪可以打开或者关上位于两个舱室之间的管道，以选择性地分开高速气体分子与低速气体分子，从而永久地驱动一个引擎。西拉德（后来成为制造与控制原子弹的领军人物）将这个难题简化为它的本质：整个实验只包含一个分子，而信息单位存在于这个分子的哪一边。信息与物质之间的联系直到1971年才由罗夫·兰道尔在IBM建立，当时他终于驱除了麦克斯韦妖，他解释说，一旦把妖怪的思想考虑进去，它就不是一个永动机了。他与他的同事查尔斯·贝内特进一步证明了抽象的思考必定会消耗物理资源。兰道尔与贝内特是我在学习计算机物理时的导师，他们向我传输了“信息是物理的”这一概念。对这个概念的观察不可避免地让我将计算与制造联系起来。

数字与制造之间的联系最早出现在大约40亿年前，这是我最喜欢的制造机器（核糖体）的演化时期。核糖体是制造分子的分子，它能够读取编码——信使RNA携带的基因序列。基因序列具有香农与冯·诺依曼所设想的所有数字代码的所有属性——错误可以被察觉与修复，并且这套系统的设计可以产生即使有错误出现，相邻的代码也可以生产属性相近的分子的效果。但是这些来自基因组的分子信息不只描述了形状，它们还变成了形状。

这套系统中的遗传密码可以写成A、C、G与U，分别代表腺嘌呤、胞嘧啶、鸟嘌呤与尿嘧啶。密码子是由三个字母组成的群组，与20种基本氨基酸当中的一种相对应，有点儿像分子级别的乐高积木。氨基酸先形成一个与密码子相对应的线性结构，之后再折叠成类似螺旋状

或者片状几何图形的二级结构，之后再进一步构成第三级的立体蛋白质结构。在这些三级结构之上构建的是关于系统功能的四级结构，比如我们身体当中被用来探测味道或光线的感应器，以及移动我们肌肉的马达。这些氨基酸最令人惊奇的是它们毫不令人惊奇。它们有一系列性质，比如亲水或者疏水，酸性更强或更弱。这些性质都不是极端的，其涉及的化学反应都是很常见的，但是仅仅这20种氨基酸的排列组合就能制造你。

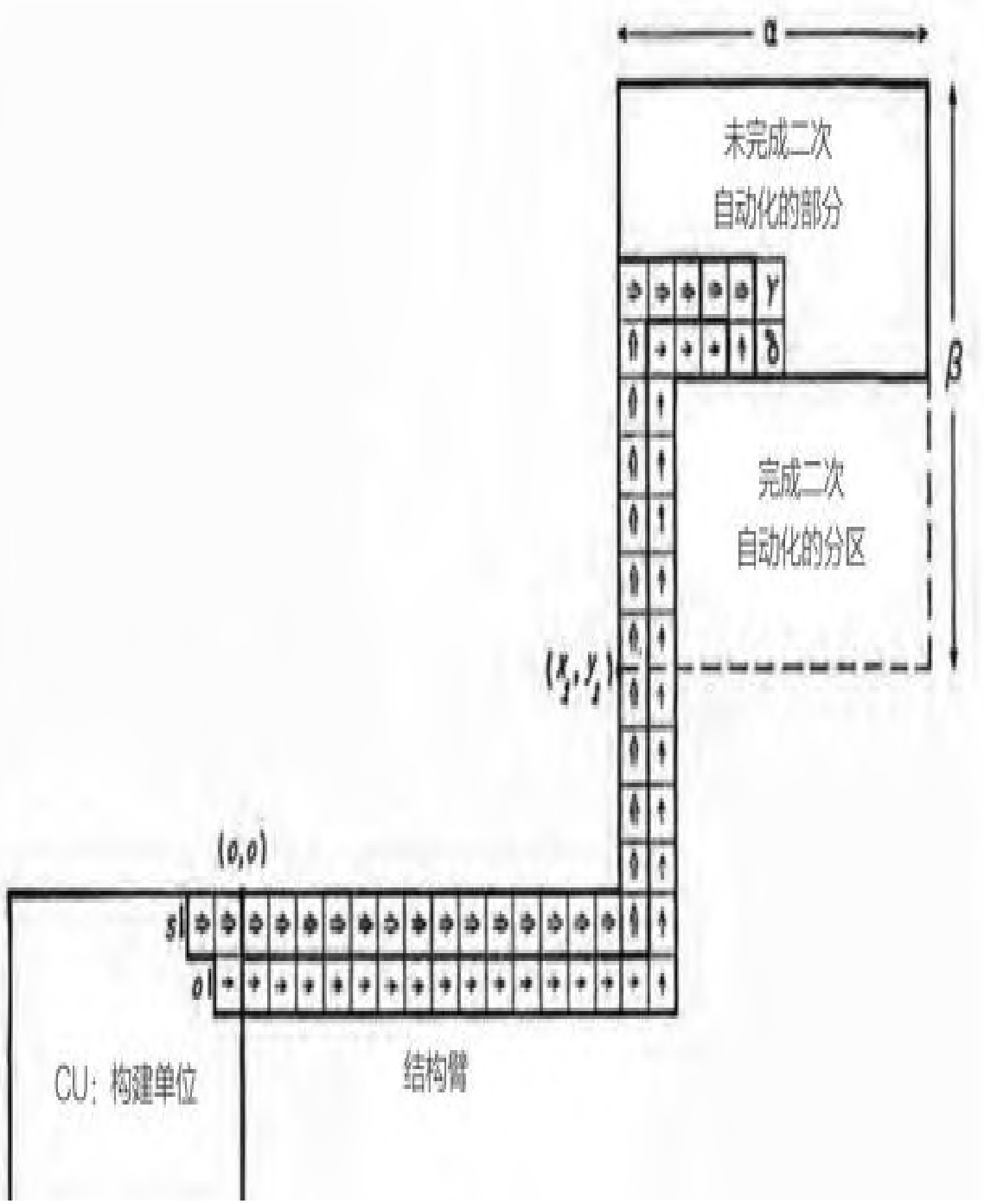
如果你混合两种化学物质让一个反应发生，99%的产出率已经非常棒了，但也意味着约1%的物质没有发生反应。在制造蛋白质时，核糖体在每10 000个步骤中才会发生一次错误，因为它是在使用遗传密码进行建造。DNA复制具有一个额外的错误纠正机制，错误率约为$1/10^8$。这就是阈值的指数级缩放，它使像你这样复杂的物体被制造出来。生命的秘密在于它是数字化的。

遗传代码携带着信息，该信息可用来执行用于编程的计算。第三次数字革命从生命起源之初就伴随着我们，但是它被限制在制造分子生物学材料上了。经过40亿年的酝酿，它已经扩展到了世界上（目前）无生命的部分。

维度诅咒和智能设计

如果生命的秘密在于它是数字化的，那么它是如何被编程的？这不是一个反问句。我的实验室参与了一项合作：与美国J.克雷格·文特尔研究所一起用电脑设计并合成完整的细胞基因组，用其创造活的细胞。为了设计一个能自我复制的组装机器，我们首先需要学习如何在从零开始创造的系统中完成同样的动作，操控它的演化过程。

关于生命是如何被编程的最明显的答案无疑就是用遗传密码，但是这个答案就像莎士比亚的戏剧是用字母表中的26个字母书写的一样空洞。了解符号只是学习一门语言的起点。结果表明，这个问题是将三次数字革命历史性地统一到一起的核心问题。



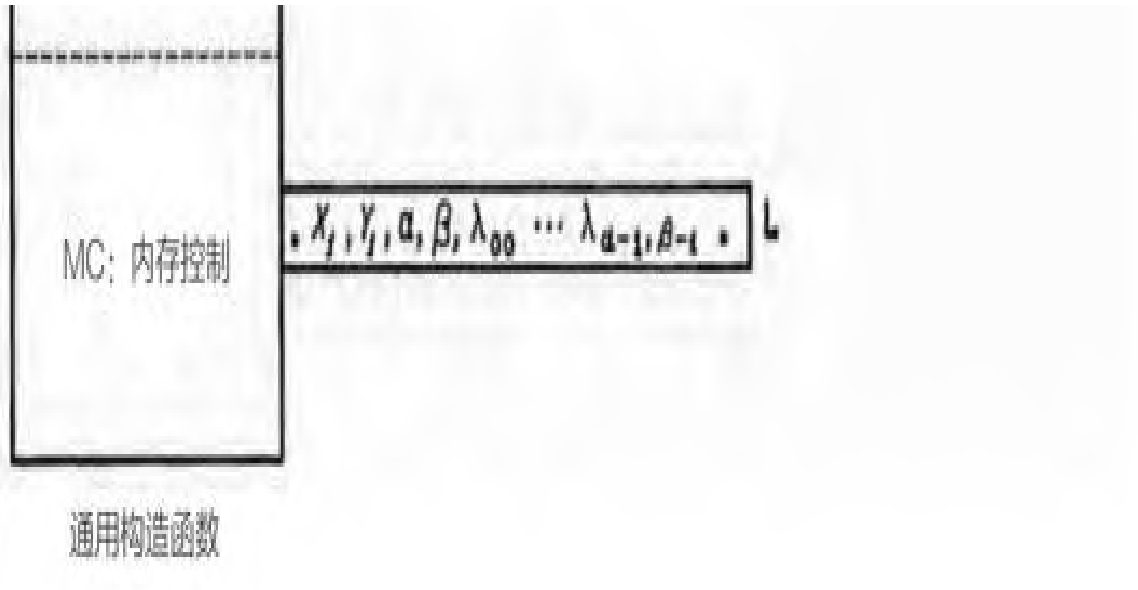


图3－7 冯·诺依曼的通用构造函数。亚瑟·伯克斯绘

在今天，用电脑制造东西就像一种聚会游戏的很糟糕的版本，信息在一群人中传递直到彻底失真。这个过程从一个计算机辅助设计软件开始，你需要学习如何使用这种具有非常陡峭的学习曲线的程序来设计你想设计的东西；之后再到计算机辅助制造过程，将设计转换成机器可以一步步执行的指令；之后再到机器控制过程，将指令转化为一台特定机器的运行步骤；最后进行运动控制，将这些指令编译成特定部件的操作步骤。所有步骤通常只能向一个方向运作，机器没有办法告诉设计人员在制造过程当中出现的问题。而且，如果设计或者机器有任何变化，以上流程都要重新走一遍。

这种问题之所以存在，是因为历史上这些步骤之中的每一步都是由不同的人在不同的地方和不同的时间完成的。你可以将这种劳动力的分工归因于文艺复兴时期的一个错误。当博雅教育出现并掌握话语权时，知识就被划分为“三艺”（逻辑、文法、修辞）和“四术”（算数、几何、天文、音乐）（即粗略地划分为文科与科学）。当制造东西被认定为一个用来获得收益的非博雅教育的时候，艺术与工匠产生了分歧。这种分歧就是我在高中时被告知不能去职业学校的原因

——我太聪明了；这也是当我在贝尔实验室工作时被训斥的原因，当时我得自己去工作室制造东西，而不是告诉那里的工人我想让他们制造什么。但是，当一个人可以在某一时间、地点将以上所有步骤都完成时，这种模式就不再合理了。表达方式从文艺复兴时期开始发生改变；3D打印或者嵌入式编程可以像画油画或者写诗一样富有表现力。个人制造而非大众制造的需求将对“学习什么知识”的概念给予革新，社会需要把制造当作一种与阅读、写作同样基础的技能。

人们可以通过了解数字制造应当具备基础的设计素养来修正这个历史性错误，但是仍然存在着一个假想的设计师。虽然这个见解听起来好像有一定的反复性和神学性，但是，谁设计了你？这个问题的答案躺在计算与制造更深层次的联系之中。为了回答这个问题，我们必须了解数字世界与物理世界是如何分离的。

对于观察到计算机科学这门学科对计算机和科学而言都很糟糕这一事实，我很乐意居功。我之所以称其为糟糕，是因为计算机科学的标准都是建立在非物质假设上的。计算机科学教学生们如何用脱离于物理现实的语言写程序，之后让其依赖另外一些人在另外一些地方将这些虚拟代码转换成物理现实。这有点儿像电影《大都会》（*Metropolis*），一场革命正在社会底层酝酿。

我以前的学生杰森·泰勒就是这样的一个底层居民，他是脸书公司负责建造海量计算中心的基础设施部门的副总裁。另外一个底层居民是我之前的学生拉菲·克里科里安，他在推特公司做同样的工作。他们都没有跟我一起学习计算机科学，但是假如你看到他们每日工作所见到的海量计算设施，你就很难相信软件是与物理世界相分离的抽象概念——他们的工作内容是以最高的效率将成捆的美金、海量的计算设备和大量的电力转化为信息。

现在流行的软硬件分离具体表现在由冯·诺依曼创造的计算机架构（冯·诺依曼架构）中。冯·诺依曼优美而又谨慎地写下了他对此架构的开创性贡献，比如如何用不可靠的器件进行可靠的计算。但软

硬件分离并不是其中的内容，这个思想起源于他于1945年写下的一份备忘录——关于EDVAC电子数据计算机的报告的第一份草稿。正如标题所暗示的那样，这个文件并非年代久远。EDVAC是世界上最早的电子可编程数字电脑之一；该报告显示了他对编写程序的想法。他建议将软件分为他所谓的“器官”，比如一个计算单元和一个存储单元。这台计算机能够以每秒1 000次运算的速度进行计算，现在，任何一台稍微好点儿的智能手机或者电脑每秒都可以进行10亿次运算，但是我们仍然生活在EDVAC遗留的设计局限当中。

冯·诺依曼和克劳德·香农在第二次世界大战期间曾在普林斯顿的高等研究院遇到艾伦·图灵。1936年，来自英国剑桥的图灵提出了一种计算的理论模型，该模型基于一种记录着不同符号的磁带，当符号经过编程磁头时，这些符号会被读取或写入。他用这个模型（后来被称为图灵机）来介绍一种通用计算机的概念，这台计算机可以模拟任何其他通用计算机。他还演示了有一些问题是永远不能被计算机解决的。这些深刻的见解并不是为了提出一个严肃的计算机架构而构思的。不幸的是，这恰恰是实际发生的事情，图灵机的零件变成了冯·诺依曼架构的组成部分。

这给我们遗留下来的问题是，当磁头被认为是独立于磁带的存在时，绝大多数电脑的能量被浪费于传递信息。只有非常少的能量被实际用来运算数学公式，虽然负责存储信息的晶体管的实际计算能力与负责计算的晶体管一样好，但是它们只被用来存储信息。计算机当中的许多安全漏洞都来源于那些在逻辑上应该分开但是在实际物理层面上相邻的晶体管。

冯·诺依曼理解这些问题。除了关于EDVAC的报告草稿，他从来没有写过关于“他的”架构的内容。他写的最后一本书是在他去世后的1966年出版的《自复制自动机理论》（*Theory of Self-Reproducing Automata*），这本书的内容属于一个非常美妙而有意义的领域。他最终对于理解生命现象产生了极大兴趣。为了研究这个课题，他与同事

斯坦·乌拉姆合作开发了一个非常不同的计算机模型，这个模型被称为元胞自动机，其本身明确地表示了空间、时间和逻辑。你可以把它看作一个自己与自己玩跳棋游戏的机器，棋子在一个给定的网格中移动。这是真实世界中的计算，不是虚拟世界中的计算。它可以计算任何其他通用计算机能计算的问题，甚至可以在所有时间、所有网格上进行并行计算。冯·诺依曼用这个计算机模型设计了可以复制它自己的机器，包括关于复制的指令。在那时这只是一个理论上的构想，但是在第5章我们将会看到，自复制机器就是第三次数字革命研究路线图的一个终极目标。

图灵也对计算的物理形态表示过赏识。在1952年发表的《形态发生的化学基础》中，他极其美妙地描绘了他的一生中描绘的最后一个事物：自然界中存在的图案模式形成过程。他的目标是解释如何由基因转换为一个功能结构，也就是如何从比特转换到原子。他通过展示特定图案模式如何在描述化学物质空间分布的方程中形成来做到这一点。

图灵与冯·诺依曼都在向作家道格拉斯·亚当斯所谓的“生命、宇宙与一切”的终极问题靠近。我们人类演化出来的一切事物当中最有趣的可能就是兴趣本身——我们的自我感知。这个演化过程的运作方式在于自然智能与人工智能之间一种令人惊奇的联系。（马文·明斯基曾经帮助比特和原子中心围绕这一联系进行规划。）

如果你在一个房间里弄丢了钥匙，你可以搜索整个房间；如果你不确定是在哪间房里弄丢的，你需要搜遍所有房间；如果你不确定是在哪栋建筑里弄丢的，你需要搜遍所有建筑；如果你不确定是在哪个城市里弄丢的，你必须搜遍所有城市中所有建筑里的所有房间。这个链条越长，你找到钥匙的希望越渺茫，直到你感觉到自己不可能找到钥匙。从技术上讲，这个状态叫维度诅咒。

最近人工智能领域快速发展背后的原因并不是对智能的理解的突破，而是在应对维度诅咒上的进展。一个能够与人交谈的计算机必须

要能在所有人可能发生的对话中确定你说了什么；一台能够自动驾驶的汽车必须能够分辨在道路上发生的所有可能性事件。在被称为深度学习神经网络的模型中，寻找答案的方法实际上一直在被不断使用；新出现的状况是现在有足够的数据来让网络进行预测训练，使其有足够的运算能力来处理这些数据，从而让网络的不同层级可以足够好地表达正确的搜索位置。在我们的大脑中也在发生类似的事情，感觉的输入在越来越高级的抽象层次中被处理。

在自然进化过程当中，维度诅咒是通过改变生物体所有可能的适应性来寻求适合生存的优势因子的。生命解决这个问题的方法是利用同源基因，这是所有生物基因组中最保守的组成部分之一，这些基因在数亿年间都没有改变。同源基因控制着其他基因的表达，设计了所谓的发育进程，即一个细胞如何一步步形成一个完整生物体的过程。你的基因组里没有任何东西表明你有5个手指，而是有一个程序，当它被执行时，会产生5个手指。它是一种现代意义上的计算机程序，用逻辑联系输入与输出。

提到进化与发育之间的联系，最简单的描述方法是进化发育生物学。进化发育会经过一个叫种系特征性发育障碍或者“漏沙口”的地方。在我们所属的脊索动物门当中，有很多种生殖方式以及很多种身体结构特征。但是，所有脊索动物的胚胎在它们分化成最终形态之前看起来都是相似的。这种早期生命阶段的相似性曾经一度被错误地当成胚胎重演律，即每个生物个体的发育都经历了所有的进化阶段。但是，事实证明这个说法是错误的。所有胚胎的发育都拥有一个共同的形态发生机制，这个机制促进了形态的形成。

人类基因组有数十亿个碱基，而你有数十万亿个细胞。在宇宙所有时间里也不可能有足够时间对它们一个个地进行改变，即使有，大多数变化要么是无关紧要的，要么是致命的。同源基因为此提供了一个更小的空间限制，便于我们在更小的、可能发生的变化中寻找更加有利于生存的形态变化方案。

这就是生命的另一个秘密：生命的复杂性是一系列小且离散的建造单元（氨基酸）在其组装过程中检测与纠正错误的结果；生物多样性是被表现为发育程序而非组装程序（同源基因）的机制设计的结果。“我思故我在”是正确的，只要你认识到思考是由你的分子以及你的大脑共同完成的。

这就是我们在非生物系统中要完成的事情。在本章中，我首先描述了设计在今天是如何被完成的，这被称为命令式设计：你必须确切地指定如何制作某样东西。我在这一章节的最后展示了进化过程中的设计是通过在一个仔细设计并严格限制的空间中不断探索来完成的。进化的目标是生存，但是数字设计的原则并不被生存限制。声明式设计（或者叫生成式设计）是此类设计流程的名称，它允许你描述你想要做什么而不是它应该怎样做。例如，你可以指定飞机的速度、负载和飞行距离，但不能指定飞机的大小或者形状。声明式设计的早期尝试和人工智能的早期尝试一样有限，但是现在两方面都取得进展的原因是相同的：我们正在学会掌握维度诅咒。声明式设计的搜索部分目前是在具体制造之前在计算机上进行离线仿真模拟而生成的。但是，随着计算和制造的融合，设计和制造将会在线持续进行——飞机本身会随着负载和空气动力状态的变化而演化。到那时，仿真模拟与自动装置之间的区别会越来越语义化。

（本章由尼尔撰写）

---

[1] 约翰·帕森斯，在20世纪40年代开创了机器工具的数控系统。——译者注

[2] 范内瓦·布什，二战时期美国最伟大的科学家和工程师之一。——译者注

[3] 约翰·冯·诺依曼，原籍匈牙利，布达佩斯大学数学博士。20世纪最重要的数学家之一，现代计算机、博弈论、核武器和生化武器等领域的科学全才之一，被后人称为“计算机之父”。——译者注

[4] 克劳德·香农是美国数学家、信息论的创始人。——译者注

[5] 利奥·西拉德，出生于匈牙利布达佩斯，美国核物理学家。——译者注

第4章

路径创造

当人和机器之间的区别变得模糊时，我们更需要关注社会科学。数字技术的发展推动了一场又一场革命，过去为未来拉开序幕。但是对社会科学来讲，未来则有点儿不那么清晰。人类从过去可以学到很多东西，但过去绝不能被视为理解和塑造未来的清晰序章。

科学推动了数字通信和数字计算的进步，现在它正在推动数字制造的发展。我们认为社会科学应该是一个重要的驱动因素，与其他影响人类变革的技术因素处于同等地位。然而，目前情况并非如此。事实上，社会科学如果要在人类未来中扮演它应该扮演的角色，就必须经历一场文化变革。

与前两次数字革命相同，第三次数字革命的核心挑战是：技术正在以指数级速度发展，但人们更倾向于事物以线性速度发生变化。从历史角度来看，个人和组织适应技术加速变革的能力各不相同，变化速度最慢的因素限制了技术的全部潜力。有时，技术本身是限制因素，但随着技术的进步，社会系统可能成为限制因素。在本章，我们展示了社会系统在历史上是如何通过艰难斗争跟上技术变革的步伐的，还重点介绍了个人、组织和机构的不同变化速度的重要性。最后，我们着眼于正在涌现的研究方法和思维模式的亮点，并利用这些新元素促进第三次数字革命。

在上一章中，尼尔描述了贝尔实验室的模拟阵营和数字阵营之间的斗争是如何解决的，其解决方式不是说服对方，而是消耗时间：模拟阵营的管理者消失了，新一代的数字管理者接管了。或者，正如物理学家马克斯·普朗克观察到的那样，一个新的想法能看到光明只是“因为它的对手终会死亡”。然而，在一个加速变革的世界中，这种策略是行不通的，我们不能等一代人过世而让社会系统与技术共同演化。

反应式批判

确保社会科学主动与技术同步发展非常重要，为了描述这种重要性，我们从尼尔在第3章中谈到的摩尔定律讲起。人们普遍认为摩尔定律是使数字技术产品变得更好、更小、更廉价的驱动力。在过去半个世纪（20世纪中后期），不时出现的关于摩尔定律将要终结的预测不断地被证明是错误的，给人一种“此定律必然会延续下去”的感觉。

然而，摩尔定律同样是社会力量的产物。当然，它始于戈登·摩尔的观察。在它被正式命名之前，它存在于一篇发表在杂志上的普通文章中，作者概述了他对半导体产业的变化速度和影响的观察。这篇文章推动了仙童半导体公司和其他先驱公司的科学家的研究进程，促使摩尔和罗伯特·诺伊斯于1968年建立了英特尔公司的前身——NM电子公司。在英特尔公司，性能的连续翻倍成为公司的核心标准，并随着时间的推移成为其合作伙伴、客户和竞争对手最重视的指标。随着性能的不断提高，半导体改善了无数商品和服务的质量，摩尔定律融入了社会期望。维护这条定律不仅是英特尔必须执行的战略，最终还将成为社会发展的战略，社会急切地拥抱了更好、更快、更便宜的数字设备和应用。

在社会科学中，摩尔定律被称为社会构造的现象——它不是一项自然法则，而是特定人类行为不断重复的产物。在这种情况下，竞争环境是由社会构造的。在管理学中，社会构造的环境被定义为“制定的环境”。换而言之，摩尔定律是在特定竞争环境中由商业互动创造的一个案例。信息技术性能之所以持续加速发展，是因为人类将其变成了一个必需品——它对工业和社会都十分重要。数字制造在这种预期下会发生什么变化呢？要想解答这个问题，我们需要理解摩尔定律和拉斯定律之间一些关键的社会差异。首先，我们从英特尔公司的角色谈起。

半个世纪以来，英特尔公司一直致力于坚守摩尔定律。支持英特尔数据中心集团战略规划的社会科学家彼得·莱文说：“摩尔定律是我们的商业战略。”“坚持摩尔定律使英特尔公司获得了成功。摩尔定律提供了一套使合作伙伴和客户能够在很长一段时间内评估公司的工程和制造能力的标准。”他又补充道，“这并不容易。保持这种竞争优势是非常困难的——面对高昂的成本和巨大的复杂性，我们需要庞大的资金流和研究规模，以及强大的开发攻关能力。但是，在这一点上，英特尔公司做得比任何公司都好。”他用一个比喻进行总结：“坚守摩尔定律的延续有点儿像印第安纳·琼斯一样始终保持在巨石的前面，因为在电影中，巨石一直在向他倾斜。”对于英特尔公司而言，坚守摩尔定律是一个始终存在的威胁。应对这一威胁需要持续的紧迫感、聚焦重点的能力和良好的执行能力。

那么，还有一个关键的问题：谁是拉斯定律（不断保持数字制造性能加速发展的定律）的存在性驱动者？尽管可能出现一家单一的驱动型公司，但我们认为它最有可能是一个由相互依存但又各自独立的参与者构成的生态系统，这些参与者将推动数字制造性能的加速发展。在这种情况下，一个有组织（公共组织、私人组织和非营利组织）的生态系统是否可以与单个跨国公司（由管理人员、员工和股东不断改善其财务业绩）具有相同的存在性驱动力，这是一个悬而未决的问题。此外，这种分布式生态系统可能会也可能不会随着财政和人力资源的出现而出现，以维持必要的基础研究和开发。

塑造社会系统，使之与拉斯定律保持一致，这将是社会的决定性挑战之一，因此，这也是第三次数字革命中的社会科学核心问题。研究这个社会科学问题是为了解决“数制”工坊普及、“数制”工坊新手扫盲、整体生态系统的扩充以及未来潜在风险的转移等问题，这些问题也会成为我们衡量成功的标准。不幸的是，关于社会科学发挥前瞻性作用的记录呈现的结果并非良好。在摩尔定律出现后的许多年里，几乎所有的社会科学学科都未能认识到世界正处于由数字技术驱动的大规模变革的起始阶段，因为它们并没有深入参与到新兴的数字

技术中，不了解其指数级增长的潜力，因此在塑造社会系统以共同发展的方面受到了限制。结果就是，我们的社会科学在紧紧追赶技术趋势。

对技术的反应式批判性分析——观察——是社会科学的规范。这种方法是必要的，但在技术发展呈指数级加速的世界中还远远不够。在社会科学中，当同时涉及人和技术两者时，我们需要把路径观察和路径创造相结合。为了理解如何实现这一目标，我们回到工业革命和现代社会科学的起源。

被动与主动

社会科学是一门研究人类社会和人际关系的学科，它与技术的关系从工业革命的早期开始就是反应式的。这种反应式的思维定式渗透了社会科学的方方面面——社会学、经济学、政治学、管理学、劳资关系、社会工作，还有其他相关领域。这种思维定式让社会科学家很难进行预测并帮助人们开辟加速技术变革的道路。

工业革命相比之前的手工业有一个巨大的质变，这个质变与之后的两次数字革命的相关转变同样重要。工业革命是被一系列相互关联的技术发展驱动的，范围跨越工程、化学、冶金，还有至关重要的蒸汽能。18世纪晚期，产业化的蒸汽能让人类的力量增加了1 000倍，驱动了整个社会的转变。这些转变包括从手工生产到工业化大规模量产的转变，以及从基于土地与棉花的经济体系到工厂驱动大市场的经济体系的转变。与此相关的还有新的化学制品、钢铁工艺和纺织工业制造流程的引入加速了工业产出，运河的发展进一步促进了工业产出，从而驱动了大规模运输业的发展。

新技术的引入导致了新社会现象——阶层流动——的出现。具有创新精神的个人成为新公司的拥有者，这种现象的大规模出现呼唤了我们现在称为“美国梦”的精神（努力工作和接受教育意味着你的出身并不是你的命运）。与此同时，对于绝大多数劳动力来说，工业革命带来的是浮士德式的交易，在获得稳定工资的同时，付出的代价是凄惨与不安全的工作条件、监工的压迫，以及环境的恶化。

大多数现代社会科学在第一次工业革命时期出现或者加速发展——在“使能技术”出现数十年之后。现代社会科学绝大部分出现于19世纪晚期，即马克斯·韦伯将宗教与社会文化原则关联起来，埃米尔·迪尔凯姆记录下劳工的分类的时刻。社会福利学是在英国的奥克塔维亚·希尔（19世纪晚期）与美国的简·亚当斯（20世纪初）的研究中逐步建立的，主要回应当时民众的呼声，特别是妇女与儿童在工业革命当中受到伤害与侵害的问题。心理学在工业革命之前就出现了，但随着专业协会（分别在1885年的法国和1892年的美国成立）的成立，它转化成了现代形式。19世纪晚期，西德尼·韦伯与比阿特丽斯·韦伯记录下当时压迫性的工作环境以及人们对工业民主的需求，劳资关系在英国成为一个研究领域；在20世纪早期的美国，这个学术领域起源于约翰·R.康芒斯发现的组织架构变化规律，比如，由手工业工会到工业工会的转变往往是随着技术与市场的变化而发生的。在这些案例中，社会科学领域的创始学者对社会变化产生良好反应，往往发生在工业革命深度冲击社会以后，因此，社会科学改变社会（提高大众生活质量、尽早转移潜在社会危害、对社会变革的阻碍者的利益进行重新划分）的能力被大大限制住了。

工业革命同时启发了文学作品的创作，当时出现了反映技术对社会的冲击的新题材。玛丽·雪莱的《弗兰肯斯坦》[1]（*Frankenstein*）在1817年出版，探索了技术善良与邪恶的双重面孔。这一点被书中怪物的观点所诠释，怪物所说的火可以被理解为一种比喻：“有一天，当我被寒冷压迫时，我发现了一团游荡的乞丐所

留下的篝火，我因靠近它感受到温暖而由衷喜悦，它帮助我战胜了严寒。在喜悦之中我将手伸入了余烬，但是我又因为巨大的痛苦而哭泣着迅速抽出了手。我想：这是多么奇怪啊，同样的原因竟然导致完全相反的结果！”

在工业革命的全面冲击力开始展现的时候，故事讲述者的观察变得更加聚焦于技术的危害。查尔斯·狄更斯在他1854年出版的小说《艰难时世》中瞄准了工厂中巩固压迫性工作环境的政策。同样，厄普顿·辛克莱在他1906年出版的《屠场》中描述的屠宰行业中不安全的工作环境、童工现象以及性骚扰现象引起了大众恐慌。即使当时社会制度已经根深蒂固，这些以及许多其他文学作品仍然加强了社会学家对公共政策的影响力。

今天，那些前沿学者与关注技术对社会的冲击的作家可以被称为先行者，他们定义了研究的新领域与文学的新主题，但是他们对于工业革命技术的研究路线被限制在观察技术产生的巨大影响以及解决其遗留问题上。例如，童工法以及关于劳资关系的理论本可以更加有效——如果在工业革命萌芽期就对这些领域有全面的学习与考量的话。

在技术已经进入社会并且成熟后再解决技术遗留问题的方法论限制了社会向前发展的潜力。一个组织建立之后，一个被称为“路径依赖”的现象会限制其未来决策空间。结果就是，当一个组织已经处于一条给定路线的时候，组织会有很强的停留在既定路线上的倾向。路径依赖的概念来自社会科学，而且，讽刺的是，社会科学本身也陷入了对技术的“路径依赖”。在这条轨道上，社会科学是被动的而非主动的。在某种意义上，社会科学仍然被限制在这条轨道上。

增强路径依赖是另一种动力，罗伯特·米歇尔斯在1911年把它定义为寡头政治的铁律。他观察到，新的制度安排（他专注于政党与工会）之所以存在是因为一个给定使命，但是其使命一定会演变成维持其存在的生存使命，无论最初的使命是什么。1962年，托马斯·库恩[他写了《科学革命的结构》（*The Structure of Scientific*

Revolutio）一书]将这种动力延伸到了学术领域。库恩观察到，如果要改变现有科学的既有制度秩序，就必须发生革命性的转变。

我们可以把被动观察新技术领域的学习方式称为路径观察。该方法在这里有双重含义——“被动”意味着对于技术更多的是观察、记录而不是塑造；它涉及观察和遵守现有规则，增强现有的路径观察方法。在社会科学中，这个方法被同行以及其他学术机制大大地增强了。路径观察当然也是必要的，但是它并不足以让社会科学在这个技术迅速发展的世界中起到领导作用。

主动帮助塑造第三次数字革命需要在社会科学领域发生某种形式的文化变革。目前占主导地位的积极的社会科学方法论是路径观察的一种，它必须被扩展，以包含更多的“规范性”方法的扩展应用。积极的社会科学聚焦于解释发生了什么，一个规范的方法论包含了什么应该发生。但是这种方法论也必须同时根植于证据以及对底层科学技术的清晰认识当中。但是目前，这种规范的方法论往往不被认为是科学——虽然现在已经有对此狂热的亚文化圈在做类似的工作。将积极的、规范的方法论结合到一起是可能的，甚至是必要的，必须通过大力开展大胆且有据可依的、对于不断加速的技术发展与我们想要的未来转变的分析来实现。

第三次数字革命给社会科学提供了一个难得的机会，让其在新技术刚刚兴起时就与之接触。这是一个通过拐角看到未来对于各个学科（经济学、社会学、政治学、管理学、劳资关系，以及其他学科）的深远影响的机会。这需要社会科学从路径观察者转变为路径创造者。

路径创造

在社会科学家中，亚当·斯密在18世纪中期发展了早期经济学原理，这一时期非常接近工业革命的开端。在这种情况下，他提出一些概念，例如，看不见的手（如自身利益）、劳动分工、交织在一起的

自由市场和工业革命技术。在新技术开发早期，他具有积极主动的影响力。重要的是，他是通过仔细研究技术本身来总结这些见解的，这可以通过他对大头针工厂的分工和技术的详细描述得到证明。他是一个路径创造者，或者更确切地说，是当时企业家和新兴工业家路径的共同创造者。

当卡尔·马克思和弗里德里希·恩格斯在19世纪中期崭露头角时，工业革命的技术已经很先进了。他们俩专注于生产资料控制权的研究。生产资料占有阶级被认为是社会问题存在的根本原因，因为正如马克思分析的那样，生产资料占有阶级只是从他人的工作中榨取价值，并以利润的形式为自己保留价值。由于当时“所有权”的概念深深根植于社会，因此马克思和恩格斯可以设想的唯一途径就是通过革命推翻生产资料占有阶级，并用一种社会所有制取而代之。从这种意义上来说，他们是路径创造者，在世界范围内激发起鼓舞人心的革命，聚焦于一系列与推动工业革命相反的运作假设。通过马克思和恩格斯，我们看到，在技术出现之后，技术假设嵌入社会系统，路径创造的工作越来越困难。第三次数字革命为重塑系统提供了新的机会。数字制造通过对生产资料的分布式控制，从根本上改变了工作的性质——减弱工业革命期间产生的工业系统对劳工的榨取的性质，而不需要采取革命的方式。

自马克思和恩格斯以来，已经有了许多著名的路径观察者，但很少有人渴望成为技术方面的路径创造者。弗雷德里克·泰勒作为一个路径创造者脱颖而出，但实际上，他的创新发明产生了一种使人们屈从于科技的影响。从1911年开始，随着《科学管理原理》（*The Principles of Scientific Managemen*）的出版，泰勒指出：“因所有日常行为效率低下，几乎整个国家都在遭受巨大损失。”他认为，“效率低下的补救办法在于管理，而不是寻找一些杰出的人。”这些想法得到了解决——“泰勒社”的分会在全美各地涌现。通过使用其他手段对工时和动作进行研究，泰勒确实展示了如何把人们当作机器

中的附加齿轮使用，以提高效率。社会科学与技术结合，产生了工业管理领域，但更多时候，社会科学是为技术服务的。

第二次世界大战后，一些社会科学家试图在人和技术方面开辟另一条道路。之后，在被称为“后工业运动”的大背景下，他们接受了我们所谓的路径创造，他们努力超越泰勒的机械逻辑并重塑技术，使技术更加人性化。这项运动的领导人丹尼尔·贝尔在1973年写道：技术“是一种艺术形式，它将文化和社会结构联系起来，并在此过程中重塑两者”。他警告说：“技术总是破坏传统的社会形式，同时为文化带来危机。”在我们考虑第三次数字革命的今天，这个观点同样适用。然而，他和后工业运动中的其他人错过了后工业发展中最重要的东西：数字技术的指数级变化。

后工业运动的一个关键部分侧重于工厂和其他组织中社会和技术的整合——促进我们对社会技术系统的理解。通过从学术组织中走出去并与工人以及管理人员进行合作，尝试工业试点实验的研究人员成功地确定了适应不同技术环境的人性化工作方法，最早采取这种工作方法的公司有英国煤矿、瑞典汽车工厂、美国宠物食品工厂、丹麦航运公司和加拿大炼油厂。然而，这类技术在很大程度上看起来是逐步变化的。即使社会技术系统人员是路径创造者而不仅仅是路径观察者，他们也没有为数字技术的先驱提供路径指引。最终结果是他们没有意识到技术处于指数级变化的尖端。由于未能将社会和技术的整合置于变革性技术背景下，试点实验错失了一个关键的杠杆来源。

弗雷德·埃默里和梅瑞林·埃默里都是社会科学技术系统运动的代表人物，早在1976年就解决了数字技术的某些问题，他们预测，“电信和计算机之间越来越多的联盟”将逐渐“形成能够收集和处理信息的网络，供所有用户使用”。他们预测到了尼尔的观察——未来已经在现在发生。他们认为：“我们可能已经迎来了我们的未来，尽管它还没被发现。”他们辩解道：新技术将允许“内容本身的自发生成，这将使内容生产者和内容消费者合为一体”。这类关于数字通信

和计算的设想正是我们现在看到的数字制造领域的特征。1981年，另一位早期社会技术系统领导者埃里克·特里斯特在此分析的基础上提出：“即将到来的信息技术，特别是与微处理器和电信有关的信息技术，为解决当前许多社会问题提供了极其广阔的空间——只要我们能够对其价值做出正确的选择。”同理，这句话与本书的写作目的相符。

虽然这些和其他关于技术方面的重要观察旨在鼓励共同创造具有选择性的未来，但他们误判了数字技术的指数级增长能力。在1977年出版的《未来的选择》（*A Choice of Futures*）一书中，弗雷德·埃默里和梅瑞林·埃默里甚至批评了“香农所谓的信息论”[2]。当然，在不降低质量的前提下纠正错误的核心概念很难让人接受。然而，如果不了解该技术的这些基本属性，路径创造就会受到限制。想象一下，如果弗雷德·埃默里和梅瑞林·埃默里能够预测并驾驭当时刚刚开始的指数级增长浪潮，那么社会科学的系统方法可能会产生什么样的影响。

技术发展仿佛一座孤岛，这个孤岛以外的每个人都没有注意到这项技术的指数力量，但它很受社会科学主流人士的青睐。斯图尔特·布兰德是《全球概览》[3]（*The Whole Earth Catalog*）和《共同进化季刊》（*CoEvolution Quarterly*）的创始人，他提出了一种分散的、个人的技术发展观——一种自由的、对生态环境负责任的观点。《共同进化季刊》的主要内容有：想象太空殖民地的日常生活（1975年）、洛杉矶的太阳能热水器（1977年）、遗传毒性的威胁（1979年），以及涵盖整个系统、住房和土地使用、工业和工艺、通信和社区的话题。

虽然布兰德不是社会科学家，但他早期对该技术的全程参与说明了路径创造的方式。事实上，他超越了路径创造本身，成为与一些技术先驱进行共同创造的践行者。这使他得出结论：共同进化不仅是可能的，还是必要的。编写《睡鼠说：个人电脑之迷幻往事》（*What*

the Dormouse Said: How the Sixties Counterculture Shaped the Personal Computer Industry）的约翰·马科夫说：“斯图尔特是第一个明白这一点的人。他是第一个了解‘赛博空间’的人。‘个人电脑’这一术语由他创造。他影响了整整一代人，包括整整一代的技术专家。”在史蒂夫·乔布斯于斯坦福大学2005年毕业典礼上发表的著名演讲中，他最后对布兰德和布兰德带给科技企业的深刻人文价值观表示赞赏。有趣的是，乔尔、尼尔和艾伦都订阅并深入阅读了《共同进化季刊》，三人都被其中的社会科学和人文科学深深吸引。

《软技术：共同进化》（*Soft-Tech*：*A Coevolution Book*）由布兰德和发明家J.鲍德温编写，第1章是鲍德温写的“一个高度进化的工具箱”，描述了他已经演化了15年的“便携式商店”。他的结论是“最终目的是制造自己的工具”。布兰德是一个与很多技术先驱进行合作的共同创造者。他的访问既正式又非正式。这是一个传奇的聚会场景，对共同创造来说并非偶然——高度信任的社会关系使布兰德能够通过在正常观察中无法实现的方式连接社会和技术。在1968年计算机协会和电气电子工程师协会年度联合会议上，布兰德与发明家道格拉斯·恩格尔巴特[4]共同展示了“演示之母”。这个演示采用了新技术，如电子邮件、超文本、鼠标等。

另一位不属于社会科学主流但又融入技术变革步伐的人是作家兼未来学家阿尔文·托夫勒。在他1970年出版的《未来的冲击》[5]（*Future Shock*）一书中，他接受了跟上技术发展步伐的挑战，同时注意到了主动塑造技术的重要性：

> 在过去，随着社会进化连续阶段的不断展开，人们的意识紧随事件的发展而发展，不会停滞不前。因为变化速度缓慢，人们可以通过潜意识“有机地”适应发展。今天，无意识的适应将不再适用。面对基因变化、创造新物种、移居其他行星或离开地球的能力等问题，人类现在必须有意识地控制技术演化本身。为了在驾驭变革浪潮的同时避免未来冲击，人类必须掌握不断进化的

技术，塑造符合人类需求的未来。在这个历史性的时刻，人类必须预测和设计未来，而不是抗拒未来。

托夫勒很有先见之明：在很短的时间内，过多的技术变革会给人们带来深刻的心理影响。他认为，我们必须摆脱“有意识地选择进化的力量”的束缚来塑造技术的影响，认识共同进化的必要性并设计未来。

布兰德和托夫勒因在日新月异的技术世界中关注路径创造的重要性而受到赞誉。可以说，第一次数字革命和第二次数字革命如果没有这些人文路径创造者的加入，可能会有更大的局限性，但是没有发生的事总是难以证实。因此，我们认为他们的贡献堪称典范，但是，和以往的少数路径创造者先驱一样，把社会和技术系统的共同演化当作最优先事项的他们被认为是异类。

政治经济学家迈克尔·皮奥里和经济社会学家查克·萨贝尔在1984年阐明了共同进化的社会和技术体系的挑战。皮奥里和萨贝尔相信数字技术的转型具有革命性发展潜力，并认为我们正在面临着所谓的第二次产业分工。他们了解到计算机正在以新的方式与制造业相结合，从根本上改变市场与技术之间的联系，从而实现柔性专业化。他们的分析从传统的路径观察模式开始，然后进入路径创造阶段，最后为未来发展设计纲领。

从时尚行业和出版领域开始，皮奥里和萨贝尔记录了计算机和通信技术如何快速适应利基市场的制造需求。他们指出，“现有的产业机构已经无法确保商品的生产和消费之间的可行性匹配了。”他们得出的结论是“必须改善或替换这些机构”。

皮奥里和萨贝尔因对数字技术和未来工作的杰出见解而获得麦克阿瑟天才奖。他们对该技术的灵活性的理解，特别是对意大利北部社区（想想时装公司贝纳通）如何快速响应时装设计和生产的理解，在某种程度上成功地预测了社区制造的相继出现。然而，在戈登·摩尔

的文章发表20年之后，皮奥里和萨贝尔的分析才出现。研究人员没有完全预见并充分考虑数字技术呈指数级增长的潜力。虽然这种潜力在20世纪80年代中期很难想象，但如果研究人员能够理解并完全解决数字技术如何快速发展的问题，他们就能更好地帮助我们预测数字技术对个人、组织和机构的影响。

发明家与创业者

当谈到新技术时，思维最活跃且最具影响力的路径创造者通常都是发明家与创业者。他们通常致力于将一项发明创造转化为一个成功的商业企业。在20世纪70年代末与80年代初期，许多技术专家理解并全面投入指数级增长的数字科技。比尔·盖茨与保罗·艾伦在微软的早期使命就是“每一个家庭的每一个书桌上都有一台电脑”（当然假定运行的是微软的软件）。这个使命是建立在数字软件的指数级增长的假设上的，同时前进的还有施乐帕洛阿尔托研究中心、英特尔实验室以及其他对数字技术与社会的交叉感兴趣的社会科学家，这些创业者与企业研究人员毫无疑问认识到了个人电脑对社会产生巨大影响的能力和潜力，但是他们的首要关注点是如何建立一家成功的公司，而不是如何借助新兴技术来解决社会挑战或者消除危害。如果年轻的比尔·盖茨拥有现在的他在社会创业领域的经验与远见卓识，情况会怎样？

在20世纪80年代，发明家与未来学家雷·库兹韦尔开始使用模型来跟踪指数级发展的数字技术并且预测未来。在他出版的书《智能机器时代》（*The Age of Intelligent Machine*）当中，他基于自己对于加速发展技术的观察，对未来进行了一系列预测。但是库兹韦尔是从一个发明家、未来学家而不是社会学家的角度来观察这些模型的。（他对未来能力的建模为他的发明提供了依据，因此这些发明都适时地利用了这些新兴能力。）在他的书中，他帮助大众提高了对指数级

发展的数字技术的认知度，但是他没有聚焦于如何帮助个人和组织与这些技术共同进化。

与发明家、创业者同时把目光转向指数级增长变化的是小说作家，特别是硬科幻小说作家，即利用科学上的准确的材料构建故事的人，他们被认为是在大众、科技以及未来的交界处进行共同创作的。这些作者在描绘数字技术加速发展的社会影响时，往往具有先见之明（并能够引起共鸣）。1968年，阿瑟·C.克拉克在《2001：太空漫游》（*2001:A Space Odyssey*）中设想了超级计算机、人工智能，还有像“新闻纸”（newspad，iPad的早期版本）这样的炫酷玩意儿。1984年，威廉·吉布森在《神经漫游者》（*Neuromancer*）里设想了“赛博空间”对社会和文化的冲击，将这个几年前在短篇小说《燃烧的铬》（*Burning Chrome*）中创造的概念流行化。当然，尼尔也受到了《星际迷航》的启发，将其中的复制器从构想变成了现实。这些科幻小说作家（还有其他像他们一样的人）不仅为数字技术的未来创造了令人回味的设想，还通过将这些技术放置在真实的社会场景中，用深刻的人性叙事手法揭示以及挑战了许多对于技术的既定假设。

这一点丝毫不让人惊奇，技术企业家以及基于科学编写故事的人们在塑造新兴技术方面比社会科学家更积极。但不必须如此。挑战既要积极（在科学意义上）又要规范。我们不需要抛弃社会科学既有的核心原则与方法论，而是需要扩展其传统方法论并且进行跨学科工作，在技术对我们产生让我们后悔的影响之前，理解并确认行之有效的方式去主动参与未来技术的塑造。

社会科学同时也可以有效地分析与处理由技术冲击、大众困惑产生的负面效应（这种状况在第三次数字革命的每一个新阶段都极有可能发生）。这一点要从社会科学家花时间去了解底层技术以及社群开始。当我们开始构建整合了新机遇与社会系统的未来场景时，人文科学可以帮助我们深入锚定人类的价值观和故事叙述。当然，工程师与科学家也必须对来自社会科学与人文科学的输入以及合作采取开放态

度——这种态度也需要一些文化上的改变。对于所有参与者来讲，进一步观察个体、商业组织以及社会机构对如何应对加速变革是必不可少的。

变化速率

数字技术能够以指数级的速度变化。社会系统与大众也会发生改变，但是通常来讲前者改变的速度更快。个体变化的速度（取决于态度、行为、技能以及资源获取能力）绝大多数都是比较慢的。对于商业组织与社会机构来说，改变也是渐进的（商业组织取决于战略、架构、流程、文化以及技术；社会机构取决于社会共识、标准、法律以及优先级）。号召社会科学家、小说家以及其他人一起与技术专家共同创造未来，需要对这些不同的变化速率进行更加仔细的研究、理解与观察。

社会系统逐渐增加的变化速率是技术潜能加速发展的速率限制器。这里需要注意的是，我们关注的是速率限制而不是绝对障碍。对于变化来讲，当然有可能存在绝对障碍，但是更常见的是各种形式的速率减缓而不是完全停止。

当谈到变化速率时，我们不得不着眼于社会系统的三大组成部分——个人、商业组织以及社会机构。从分析的角度来讲它们完全不同，但又在现实中相互交织。这导致我们必须理解变化在每个领域发生的速率，并且了解它们如何在技术变化的大背景下相互结合。只有完全理解了变化速率的不同，我们才能对变化产生更大的影响力。快速发生的变化可以利用缓慢变化的机构与组织的漏洞，去除变化速率限制因素。

当然，技术本身有时候也会限制变化的速率。比如，发明与安装光纤能够扫清第一次数字革命的一大重要技术障碍。同样的问题还有硬盘大小、硬盘密度以及处理器速度。向前看，如果我们不能在数字材料的基础科学方面取得进展，技术将成为拉斯定律的速率限制。（这是一个真正的端对端制造流程，在这个过程中我们可以对材料本身的构造进行编码。）当技术成为速率变化的限制时，需要解决的挑战就转变成了如何为基础科学研究获得更多的资源。

值得注意的是，社会系统从没有发展为呈线性变化，再发展为呈指数级变化。我们都经历过“没有变化”。那是一种由僵局带来的挫折感，当我们意识到对于一些参与者来讲制造僵局符合他们的利益时，这种挫折感就更深了。我们都经历过“线性变化”，我们都习惯了利率、生产力、失业率等方面的缓慢变化。这些变化的速率的确是波动的，而且波动可以是间接的，但是这些都在我们可以直观理解的范围之内。

而从指数角度进行思考并非出自本能。很多作者都会用一辆以不同速度行驶的汽车来比喻技术的加速变革。这个比喻开始于一辆时速6英里[6]的汽车，代表手工业时代，驾驶车辆不是个难事。一辆车以10倍的速度（每小时60英里）行驶，代表20世纪，即工业时代，发动、转向以及刹车变得更加困难，但是仍然可以驾驶。数字时代就相当于那辆车打破了陆地驾驶的纪录，以每小时600英里的速度飞速前行，路面上极小的扰动都可以极大地影响驾驶，而生存下来的关键则是拥有超越人类视力的视距，以在超远距离外就得知前方道路上有什么。这个比喻给个人、组织以及机构带来的预示就是，在这个变革速度又快了一个量级的时代，它们的良好运转都需要全新的能力。

按照汽车的类比，数字制造也会把变化速率指数级增长的趋势投射到未来，从而让汽车的速度在接下来的10年中从每小时600英里飞跃到每小时6 000英里，到21世纪中叶，可能会变成每小时600万英里。稳住，这将是一段非凡的旅程。正如库兹韦尔警告的那样：“今天我

们预计了技术的持续进步以及随之而来的社会反响。但是，未来会比绝大多数人预料的更加让人惊奇，因为只有少数观察者真正理解了技术变化速率本身在加速这一事实的含义。”

回顾机构、组织以及个人如何顺应变化对于我们是极有帮助的，因为我们期待这三者能够积极主动地塑造第三次数字革命。

社会机构

从历史上讲，制度（人为制定的法律、法规，风俗习惯以及关于社会中应该发生和不应该发生的事情的安排）的变化非常缓慢。首先是“制度”这个词，它是在15世纪中期在法语中被创造出来的。这是它第一次出现在西方语言中。然而，它描述的东西已经存在一千多年了，比如皇权、军队、教会。这些以千年为单位的组织变化得非常缓慢，这可能是这个词出现得如此之晚的原因。对于个人和组织来说，没有必要用一个词来描述一个在几代人当中并不明显的概念。

随着工业革命的到来，变化的步伐加快了一个数量级。民主政体、政党、工会、公司、男孩女孩俱乐部、基瓦尼俱乐部等新的组织的制度安排在50~100年之间而不是在500~1 000年之间相继出现。变化的步伐加快了。在人的寿命周期中，如果这种重大的变化是明显的，我们称之为一场革命。

在当今这个以数字技术为动力的世界，机构面临着以更快的速度改变的压力。在音乐、金融、政府、婚姻、商业、劳动、教育、安全以及宗教等领域，各种制度安排都在以一种比前一时代快10倍的速度发生着巨大变革。今天的机构必须处于一种不断变化的状态，底层剧变经常发生，大约每5~10年一次，这比20世纪大部分时间的变化快一个数量级，而且这种变化的步伐还在继续加快，一些证券交易委员会正在经历以月为单位而不是年为单位的剧烈变化。

问题是，这些机构并没有迅速地进行调整。当然，我们需要机构保持稳定性与持续性，但是它们需要以前所未有的远见，通过模块化以及灵活性来平衡其核心重心或使命。如果它们保持纯粹的反应性或者刚性的态势，结果就是缺乏与外界的关联、失去机会以及停滞不前，而且常常对个人、社区、行业甚至整个社会，造成直接伤害。

2016年，世界银行在关于数字红利的《世界发展报告》中指出，要弥补前两次数字革命造成的差距，需要关注数字社会的“模拟系统”，特别是机构制度上的安排。“为了最大限度地利用数字革命，各国政府还需要致力于‘模拟系统与数字系统互补’——通过加强监管以确保企业之间的竞争，通过增加工人的技能以适应新经济的需求，以及确保机构会承担责任。”世界银行在这当中所说的“模拟系统”指的是一种社会制度安排，它需要像与之交互的数字技术那样运作。在技术的本质中蕴涵着如何达到这个目标的线索，比如模块化和数字技术中普遍存在的自我纠错机制，关于这方面的内容我们会在第6章中展开。我们也可以在如何加快社会制度的变化速度方面从最近的两次数字革命中得到一些明确的教训。

考虑一下一个能够说明问题的机构：K－12教育[7]。在前两次数字革命中，这一领域一直非常抗拒改革。如果教育机构改革步伐缓慢，那么它必定会被颠覆。目光敏锐的个人与组织可以利用整个教育系统之中的漏洞，以更快的速度推进改革，从而更好地满足教师和学生的需求。这些新兴的原创组织最终会取得成功，部分原因是它们可以作为杠杆来撬动数字科技的威力，以及它们可以用新的创新方式与数字技术共同演化。

大多数K－12教育是由20世纪的课堂模式（来源于工业模式）主导的，并且由19世纪的校历结构（来源于农业模式）决定，这种课堂模式激化了教师与学生在评估之间的斗争，同时，少数根深蒂固的组织控制着整个教育出版界生态系统。然而，越来越多的教育研究者、商业领袖、政治家、慈善家、创业者与教育改革家都在加入要求教育改

革的大合唱，他们认为我们目前并没有针对21世纪必需的技能与观念——批判性思维创意，合作、交流、设计思维，解决问题的能力和抗打击能力——对下一代进行有效的培养，这些技能与观念都是下一代在一个快速变化的、复杂的、数字化的世界中生存与成功的关键。在这个你很可能每两到三年就更换一次工作的世界中（数据来源于美国劳工部），学习如何去学习是给予你导航能力的关键，因为未来你可能要换的许多工作现在还没有被发明出来。学生们也会面临一个由不断进步的自动化、人工智能以及全球化导致的可供选择的工作越来越少的世界。

但对于学校而言，即使是那些渴望创新的学校，也在为引入和整合以项目为基础和以技术为媒介的学习方法苦苦挣扎。为了跟上数字时代的步伐，学校、地区甚至整个国家都购买了大量技术。但是这些技术大多都失败了或者很少被使用，因为相关的教师并不了解技术驱动的学习方法，技术与支持性的基础设施（比如互联网连接）或许并不可靠，而且缺乏优良的、以教学为基础的、旨在利用该技术的资源。很多家长对于他们的孩子平均每天花费8个小时在数字技术上（在家或者学校）感到沮丧，他们对于在孩子的生活中融入更多的技术不感兴趣。担心责任与安全的管理机构会封锁互联网，创建黑名单（被屏蔽网站名单）与白名单（批准接入网站名单）。除了这些挑战，围绕联邦政府与地方政府对教育的控制的斗争还在继续，而教师被夹在中间。他们总是超负荷工作，其薪水通常却很低，由于各种制度的摩擦，他们觉得这份本已很困难的工作变得更加困难。

艾伦与他的商业伙伴迈克尔·昂斯特在他们的公司（电子线传媒）为K－12学校推出两个基于游戏的创新学习项目时遇到了这种复杂的机构生态。第一个项目是《机械师》（*Gamestar Mechanic*），这是一个教授中学生如何开发游戏以培训设计思维和21世纪技能的平台。这款游戏也为STEM教育提供了极大的动力，最初是由麦克阿瑟基金会资助，由电子线传媒游戏学会合作开发的。第二个项目是《我的世界》（教育版）（*Minecraft Edu*），是一款流行的大众游戏［《我的

世界》（*Minecraft*）]的修订版。这个修订版最开始是由游戏教师公司（Teacher Gaming）开发的。这是一家由两名教师创立的公司，他们渴望发挥这款风靡课堂的游戏的力量。电子线传媒与游戏教师公司合作，扩大了《我的世界》（教育版）的用户群与影响力。这些基于游戏的平台总共覆盖了15 000多所学校，拥有100多万名学生用户。

与拥有庞大销售队伍的教科书出版商不同，电子线传媒没有专门的销售人员。它的产品不通过传统的学校渠道销售，而是通过直接与老师交流，从下往上渗透。由于现有教育机构的僵化，电子线传媒将重点放在了教师的挖掘、教师的相互推荐、课堂实践与教师专业会议中的实践上。这个方案需要持续不断地减少过程中产生的摩擦，实行灵活透明的定价策略，对基于教师和学生反馈的服务流程进行优化。和许多以数字技术为中心的创新型学习产品和服务的开发者一样，电子线传媒利用数字网络和教师社区的力量，绕过了制度的限制因素，比如集中采购、年度决策时间表和对新产品的恐惧。代理机构转移到了真正使用产品和服务的教师身上，而不是选择和采购的部门。

这种渗透策略利用了美国K－12教育系统出现的漏洞。虽然对以数字制造为基础的学习来讲有一些独特的挑战，特别是要考虑到硬件、软件和物理材料本身的复杂性，但在学校引入创新的、以项目为基础的学习方法时，许多相同的原则仍然适用。当越来越多的敏捷的、怀有相同理念并且有激情的机构充分利用这些漏洞的时候，这些机构的集体影响力就会变得越来越深，越来越广。它们提供了在整个教育系统生态中引发重大变革的潜力，使新市场领导者、平台和文化得以涌现。机构是重复性与模式化行为的产物，这些新的模式可以成为机构的新常态。

机构通常不会自我改造。在第6章中，我们将会探讨机构是如何制造机构的，就像机器制造机器一样。然而，历史告诉我们总有一些强大力量反对灵活的高适应性机构。正如我们前面提到的，罗伯特·米歇尔斯提出的寡头政治铁律和库恩的科学革命指向了间断平衡模型，

在这个模型中，稳定周期总会被快速变化周期打断，直到新平衡出现。库恩记录了一个系统早期的持续变化的潜力是如何存在的。但是我们还应该知道：如果速率限制因素以保守的知识守护者的形式阻止变化的发生，同时产生不断增强的对抗力，那么，当压力超过对抗力量而导致变化发生时，这种变化往往是迅速且具有颠覆性的。

考虑一下尼尔简单地为一个“数制”学院申请“.edu”互联网域名并通过美国高等教育信息化协会审核的经历，这个协会是为管理那些名称而构建的一个机构，它提供了一个如何在机构内部制订僵化规则的绝佳例子。美国高等教育信息化协会是一个产生于第二次数字革命时期的新机构，但是它仍然以一种非常传统的方式进行运作。这里的规则是，只有那些具有实际地址的教育机构才能得到“.edu”域名。这一规则的确给“.edu”域名的发放带来了一定的稳定性，但也限制了合法的、以非传统方式进入该领域者的创新与增长的空间。新的机遇制造的教育模式出现得如此之快，甚至超过了数字机构的应对速度。

除了机构变化缓慢以外，还有一些其他的变化速率限制因素，比如内部小型利益团体制造机构整体性僵局的能力。要在社群层面了解这种动态博弈过程，可以参考乔尔和他的团队在堪萨斯州的道奇城与当地领导人合作绘制数字参与者地图的情况。在接受调查的100个参与者中，有97个人对在教育中融入数字技术的重要性表达了不同程度的支持，理由包括教育、健康信息、劳动力培训、公民参与度等方面的内容。然而，少数人以这样强硬的措辞表示反对：“我认为互联网连接仍然是一个特权，我不愿意花费更多的钱让低收入家庭可以不必花钱就联网或者得到相应联网折扣。”这个论调被广泛称为“公地悲剧”——各人自扫门前雪，不愿意为公共事业投资，从而导致公共设施的失效，最后共输。在社区首脑会议上，人们必须相当重视这些观点并确保它们不会破坏整个协议。埃莉诺·奥斯特罗姆记录了新机构如何协调个体与公共事业关系从而战胜狭隘的利己主义（这会成为第三次数字革命当中的一个挑战），从而获得了诺贝尔奖。

第三次数字革命需要在指数级变化背景条件下仍然保持敏捷性的机构。在功能层面，机构必须完成两件事：创造价值和减少伤害。或者就像亨利·基辛格在外交政策机构提出的问题一样："我们的利益要求防止什么发生？""我们的利益是什么？"在一个无人驾驶汽车必须进行道德判断以及几乎每个人都能破解基因的年代，机构仍然需要持续提高它们的能力，但是变化的速度要快得多。拉尔夫·赛瑟罗恩[8]在他的2012年总统选举讲话中呼吁在全美范围内对美国国家科学院在制度上进行广泛改革：

> 今天最令人不安和令人生畏的问题都有共同的特点：其中一些问题是由人口数量与资源运用产生的；这些问题需要社会不同群体长期致力于此并由社会不同部门联合解决；单一、简单的方案不太可能解决这些问题；但是如果不解决这些问题将会导致灾难发生。从某种意义上来讲，我们目前和未来所要面对的问题的规模和复杂性是前所未有的，而且解决方案很可能必须是迭代的……机构可以使个人的思想与能量对社会造成更大影响，并且以个人所没有的能力将这些影响长久地维持下去。

组织

第三次数字革命将对一切组织产生深刻影响，无论是商会、非营利组织、政府部门还是慈善机构，无论这些组织是处于农业时代、手工业时代、工业时代，还是前两次数字革命时期。一般而言，策略、结构和过程的变化速度会快于组织文化，而组织文化限制组织变化速度的关键因素。反过来，文化也为深深根植于组织的各种运行假设条件所体现和维系。在第三次数字革命中，组织要想幸存并壮大，需要不断维护（并调整）其运行假设条件，使其不断创新以契合新的技术趋势。与此同时，组织运行的假设条件也会决定新技术的走向。

一个运行假设条件通常不会被陈述出来，但它代表了组织中事件发生的方式。组织学者把这些假设条件比作DNA——形成组织本质的特定基因。1960年，道格拉斯·麦格雷戈在《企业的人性面》一书中明确提出了“组织基本运行假设”这一概念。他将“X理论”管理风格（工人需要被监控和控制）与“Y理论”管理风格（工作就像游戏一样自然，只需要给工人工具和资源，他们就能尽力而为）进行了对比。麦格雷戈谈道：“下次你去参加……某个会议……要换个角度倾听关于人类行为的假设，无论这个假设是关于个人的、某个特定群体的，还是一般意义上的人。”他的观点是，一旦你掌握了潜在的运行假设，就可以预测接下来会发生什么——对产生决定和驱动行为而言，“X理论”假设将会产生与“Y理论”假设截然不同的效果。在预测第三次数字革命对组织的影响时，我们也需要换个角度，以适应限制变革速率的运行假设条件。

如果运行假设条件阻碍了必要的变革，人们就必须做些改变，而这种改变并非易事。以代表第一次工业革命的企业——福特汽车公司为例，福特汽车公司已经由大规模生产巨头转变为更为敏捷的、以团队为基础的企业，在质量、安全和其他关键措施方面不断提高（欧洲的社会学家仍将大规模生产称为福特主义）。冈萨洛·雷伊是穆格公司的首席技术官，穆格公司是一家为不同行业生产专业零部件的制造商。冈萨洛将过去一个世纪福特汽车公司车间的变化速度与其他行业进行了比较：“如果你定格福特汽车公司的工厂车间，观察它100年，就会看到科技带来的巨大变化。与建筑工地或医院相比，在相同的100年里，劳动内容没有发生太大的变化。“福特现象”引发的结果是，汽车变得人们普遍负担得起了，而盖一座摩天大楼、买一栋住房或去医院看病的相对价格与100年前相比却并没有增加太多。”对冈萨洛来说，区别不仅在于技术的可获取性，还在于采用该技术并将其融入组织功能的实际过程。这才是变化的关键。

自20世纪80年代以来，乔尔一直深度参与福特汽车公司及其美国工会（美国汽车工人联合会）的重大变革。取得进展的关键是改变公

司的运行假设条件。例如，福特汽车公司的经理从不得不尽力控制问题以防止被高层发现，转变为公开分享问题，讨论潜在隐患，以从中吸取教训。从隐藏问题到分享问题，这种运行假设的转变需要在组织的各个层级上进行整改。一线工人必须确保不会因为报告质量错误或通过停产来解决问题而受到指责。在福特汽车公司，一条装配线停产的代价大约是每分钟10 000美元。

保守策略在高管层面也曾一样盛行。各级领导都担心事故会让他们丢了乌纱帽，所以一心控制问题。文化中的这种保守策略是一种速率限制器，当组织文化转变为支持共享和集体解决问题时，这种保守策略就变成了加速器。管理机构的这些变化至关重要，但它们也花了很多年才得以实现——从“指责文化”到“共享问题文化”的转变花了十几年时间，而且正在进行。

同样，工会领导人的策略已经从直接反对不公正待遇和管理层的独断专行，演变为在安全、质量和其他目标方面与管理层建立伙伴关系。现在，工会会员在“六西格玛”质量流程管理变革中获得“黑带”称号，并协助设计质量和安全操作系统。当乔尔促成了美国汽车工人联合会与福特汽车公司之间的联合质量宪章的发布时，最终文件是重要的，但其过程更加重要。同理，智造宪章的重审过程至少和最终文件一样重要。美国汽车工人联合会与福特汽车公司的立宪过程花了几个月的时间，但是，以伙伴关系和共同责任为核心的运行假设条件已经覆盖全企业的协同工作很多年了。总之，美国汽车工人联合会与福特汽车公司关系的转变跨越了30年，经历了至少56个关键事件（如发布宪章）。以类似的方式，智造宪章要想凝聚共识，改变智造生态系统中的运行假设条件，也将需要多年的时间和许多关键事件。

多层级管理的大规模生产模式是一个组织的速率限制器，因为其核心在于指挥链条中的计划、决策和支持行动。早在1960年，麦格雷戈就意识到，“很可能有一天，我们会开始把组织结构图画成一系列相互联系的组群，而非个人‘报告’关系的等级结构。”要改变根深

蒂固的运行假设条件，其过程并不容易。首先，现有假设必须是显而易见并受到广泛认可的；其次，人们必须能够尝试基于不同假设的新行为，而当新行为不符合既定规范时，可以免于苛责。领导者需要收集与新方法相关的证据，但又要在所有证据出现之前就克服惯性，推进新方法。当新假设“摆在桌面上”时，会产生关键的测试事件。这是一个迭代过程。

很多现行组织都是工业革命兴起的产物，其运行假设条件的改变都是为了顺应前两次数字革命的发展趋势。第三次数字革命将进一步挑战这些组织的运行假设。重要的是，许多即将公开的运行假设条件都与更广泛地利用知识和技能有关；也就是说，增加整个企业的代理权。一个更为分散的运营模式要求企业更灵活，更有弹性，不断与时俱进，而传统的模式则会限制变革。毫无疑问，这些组织将经历一系列关键事件，这些事件将像福特汽车公司所经历的那样构成一场变革，但这些组织将没有另一个30年的时间来进行变革。

这一挑战不仅仅适用于传统的产业结构组织。正如尼尔所指出的，前两次数字革命的许多创始机构（如美国数字设备公司、王安电脑公司、Prime和美国数据通用公司）已经不复存在。虽然它们生产数字产品，但这并不意味着它们有足够的运行假设来适应数字时代的变化速度。以美国数字设备公司为例，限制其变革速率的因素就是在一定程度上未能解决从大型机向个人电脑的转变。再深入考察，就会发现深嵌于公司的许多运行假设，既是其早期成功的关键因素，又是后来限制其发展的核心障碍。这其中包括对创始人肯·奥尔森的个人崇拜。当他正确理解技术前沿时，这种崇拜使公司顺风顺水，但当他错了的时候，这种崇拜就是致命的。此外，生产线的一些运行假设压制了不同意见，限制了自下而上的适应能力。

这些从第二次数字革命中幸存下来的组织已经有点儿像制造机器的机器了。也就是说，这些组织正在通过兼并、收购、合资、公私合营和多种灵活的团队运作方式来组建新的组织。它们的运行假设不仅

为规划、决策和行动提供了更多的横向论坛和机制，而且其功能在某种程度上与尼尔在第3章中所概述的模块组合的分拆大同小异。不过，这一类比还欠精准——这些组织将拒绝被完全解体。此外，数字制造将挑战以比特为中心的组织的运行假设条件。例如，这些组织将不得不考虑其计算技术如何与活跃的消费品供应链产生互动，以及如何适应对开放共享和易货交换进行强化的社会规范。英特尔仍处于从进入移动计算市场的缓慢过程中复苏的阶段，但与第三次数字革命相关的变革可能没有那么大的容错空间。

要检视组织层面的速率限制因素和加速因素，我们需要确定技术发展的速率，并判断一个组织的运行假设条件是否顺应技术的发展要求。组织层面的协调不像机构层面那样复杂（在机构层面，更多的参与者和偏好可以发挥作用）。另一方面，不适应的组织可能在意识到组织需要变革之前就灰飞烟灭了。

个人

虽然有一些技术先驱和早期接受者拥抱技术革新，跟上它的步伐，但并不是所有人都能抓住这个机会，甚至也不是所有人都渴望这一机会。托夫勒在《未来的冲击》中说，“短时间内改变太多”会引发“无助、绝望、抑郁、不确定、不安全、焦虑和精疲力竭”的感觉。今天的加速变化使许多人都有上述感受，这些反应是情理之中的。

20世纪90年代，管理教学中出现了个人变革管理的模型。图4－1所示的过渡曲线说明了个人对变化的反应。常用的还有其他类似的模型，多数都起源于伊丽莎白·库伯勒－罗斯关于死亡与临终的模型。在图中，纵轴表示自我感知能力的相对水平，横轴表示时间。如曲线所示，在经历了新事物的最初冲击之后，人的自我感知能力会随着自身对变化的关键方面的否认程度的增加而提高。增强意识和接受的过

程包括承认你不知道的东西，放弃旧的假设、看法和其他考虑。只有这样，实验、理解和整合才有可能进行。

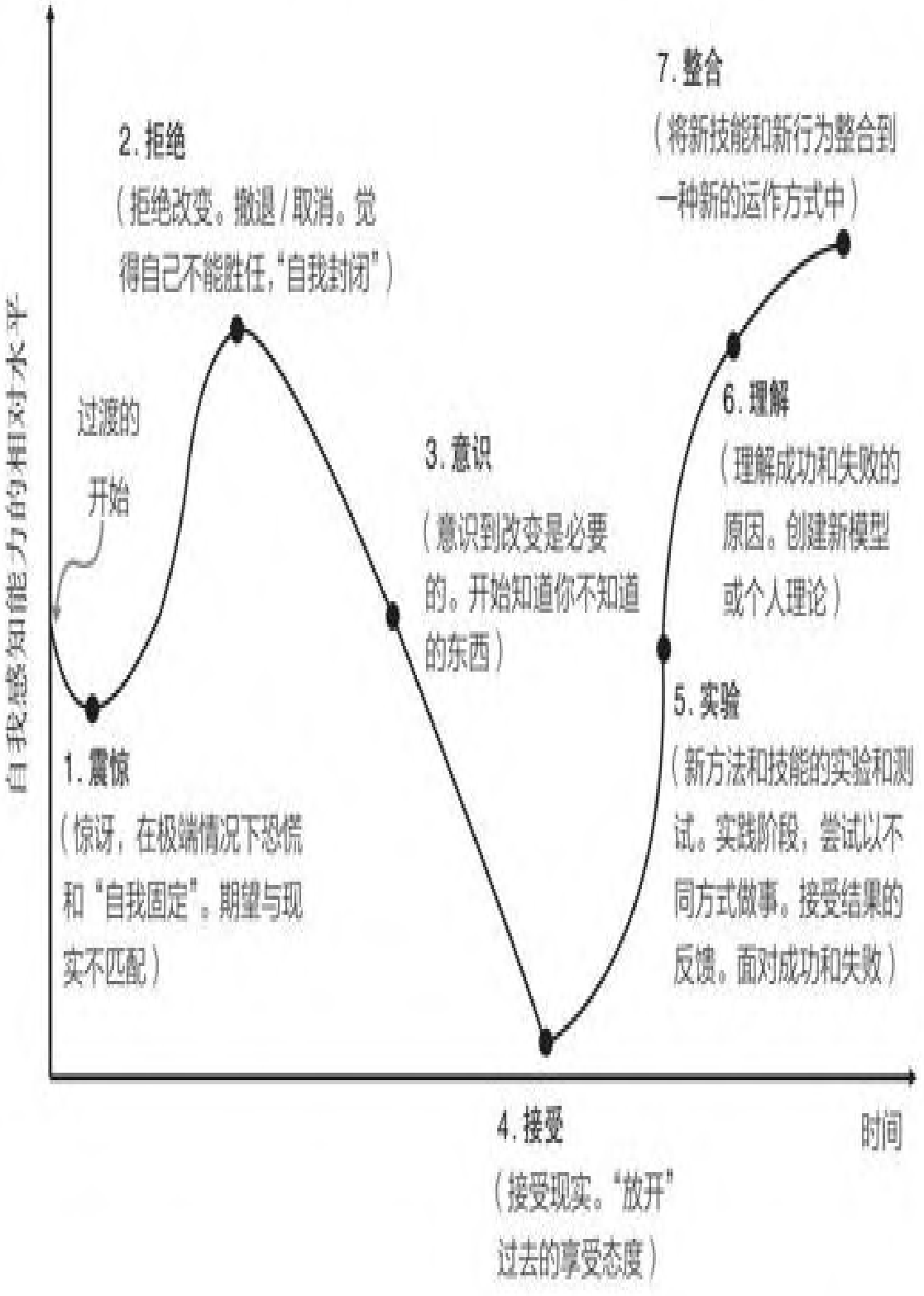

自我感知能力的相对水平
过渡的
开始
1. 震惊
（惊讶，在极端情况下恐慌和“自我固定”。期望与现实不匹配）
2. 拒绝
（拒绝改变。撤退 / 取消。觉得自己不能胜任，“自我封闭”）
3. 意识
（意识到改变是必要的。开始知道你不知道的东西）
4. 接受
（接受现实。“放开”过去的享受态度）
5. 实验
（新方法和技能的实验和测试。实践阶段，尝试以不同方式做事。接受结果的反馈。面对成功和失败）
6. 理解
（理解成功和失败的原因。创建新模型或个人理论）
7. 整合
（将新技能和新行为整合到一种新的运作方式中）
时间

图4－1 过渡曲线。福特汽车公司高管教育，《福特—美国汽车工人联合会转型内部》（*Inside the FordUAW Transformation*）提供

在第三次数字革命的背景下，过渡曲线之所以变得如此复杂，是因为它必须与呈指数级变化的速度相匹配。这种对速度的要求增加了受到震惊后对事物进行否认的可能性，并使向意识、接受、实验和理解发展的过程更加复杂。威廉·布里奇斯在1991年提出的“管理过渡”变化模型与上图描述的过渡曲线是一致的，它指出人们只有放下过往，驶向“中性地带”（意识和接受），才能接受新的转变。对于面临加速变化的个人来说，这意味着他们将越来越频繁地应对这些转变。

但对个人而言，随着第三次数字革命的到来，加速变革的意义不仅仅是一种新的思维方式。一些非常具体的调整也有自己的变化节奏。例如，在数字制造中普及基础知识所需的时间约为6~18个月——这是一个相对较快的学习曲线。然而，对材料科学、设计思维、电子电路设计和其他相关领域基本原理的深入掌握，需要几年甚至几十年。技术进步无疑将缩短基本功能所需的学习曲线，就像今天人们在使用计算机或移动设备时，无须计算机编程技能就可以操作各种应用程序一样。

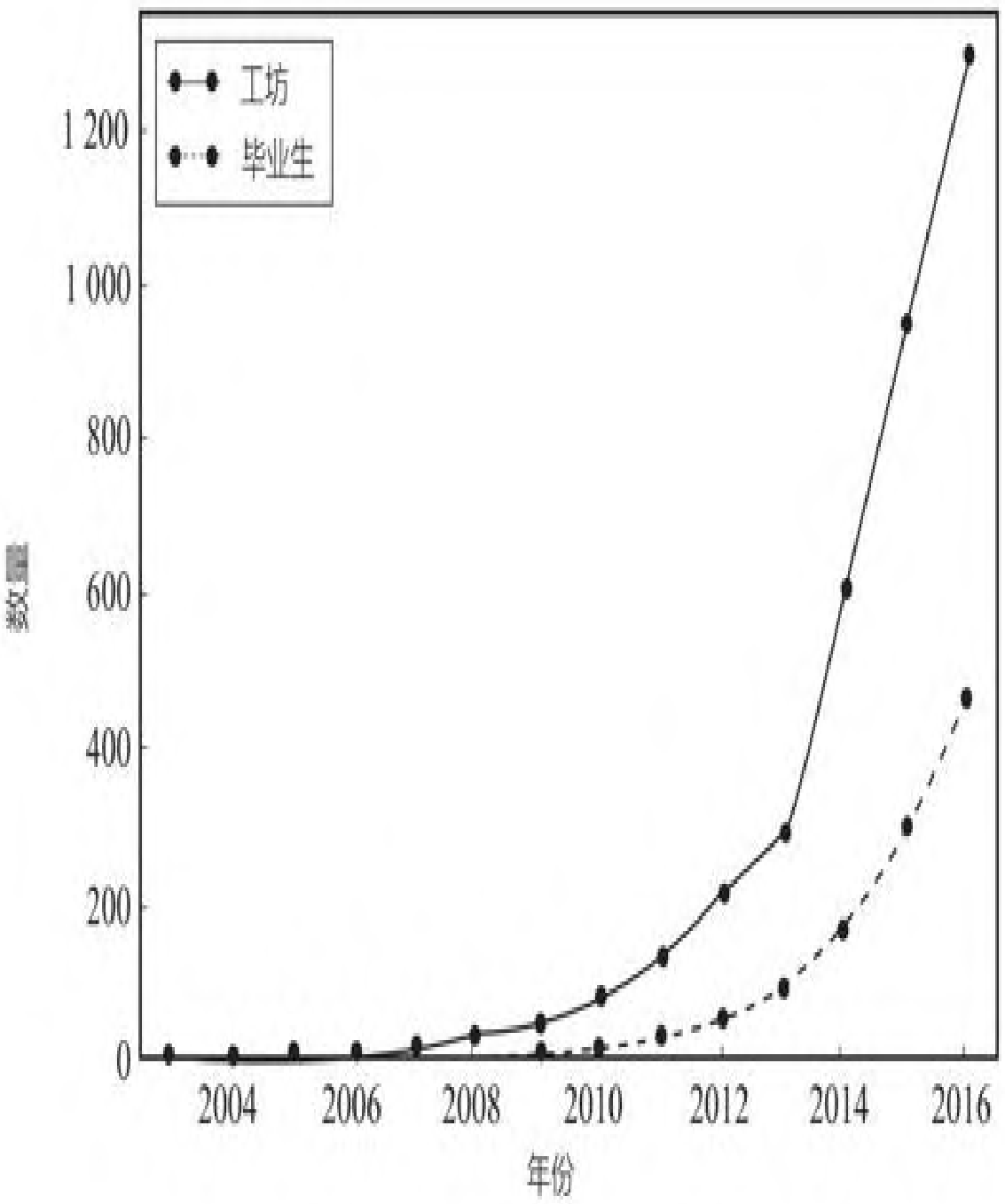


图4－2 “数制”工坊数量的增长和智造学院毕业生数量的增长。尼尔·格申斐尔德和数制学院绘

图4－2显示，能够有效利用技术所需能力的人的数量在增加。上端实线中的数据与尼尔用来绘制“数制”工坊增长曲线的数据相同；下端虚线显示智造学院毕业生数量的增长情况。纵坐标中的数值既包括“数制”工坊的数量，又包括其培训合格人员（毕业生）的数量。智造学院的毕业生常常会成为新旧“数制”工坊的核心导师。当然，除了智造学院，还有更多的方法来学习数字制作，但这个数字说明，培养智造素养与“数制”导师，既是机遇又是挑战。好消息是智造学院的毕业生数量在增长，但并不是每个毕业生都有时间或意愿去指导越来越多的人，而且整个生态系统都需要导师。更重要的是，合格人员数量的增速比“数制”工坊数量的增速要慢。

为了进一步说明相关变化速率的挑战，图4－3对上述两条曲线进行了扩展，以表示基于当前数据的未来可能情景。上面的实线是“数制”工坊可能的增长曲线，每座超过10万人口的城市（世界上只有4000多座）都有5个工坊，而在农村地区数量也大致相当。请注意，在这个场景中，“数制”工坊的增长在曲线末端开始趋于平缓——尼尔在第3章中提到过这个变化（当曲线趋于平缓时，另一次变化就会降临，形成一个更为陡峭的S形曲线）。下面的虚线是智造学院毕业生的数量，这是根据智造学院的培训成果对预期指导能力做出的保守估计。尽管如此，这个图表还是表明了一个风险，即人类能力的增长将成为实现“数制”工坊潜在增长的约束或速率限制因素。这里的重点不是对这两个方面的增长做出具体的估计，而是设想在社会制度所需的变化速度中，两个变量要如何快速增长以适应技术进步的速度。

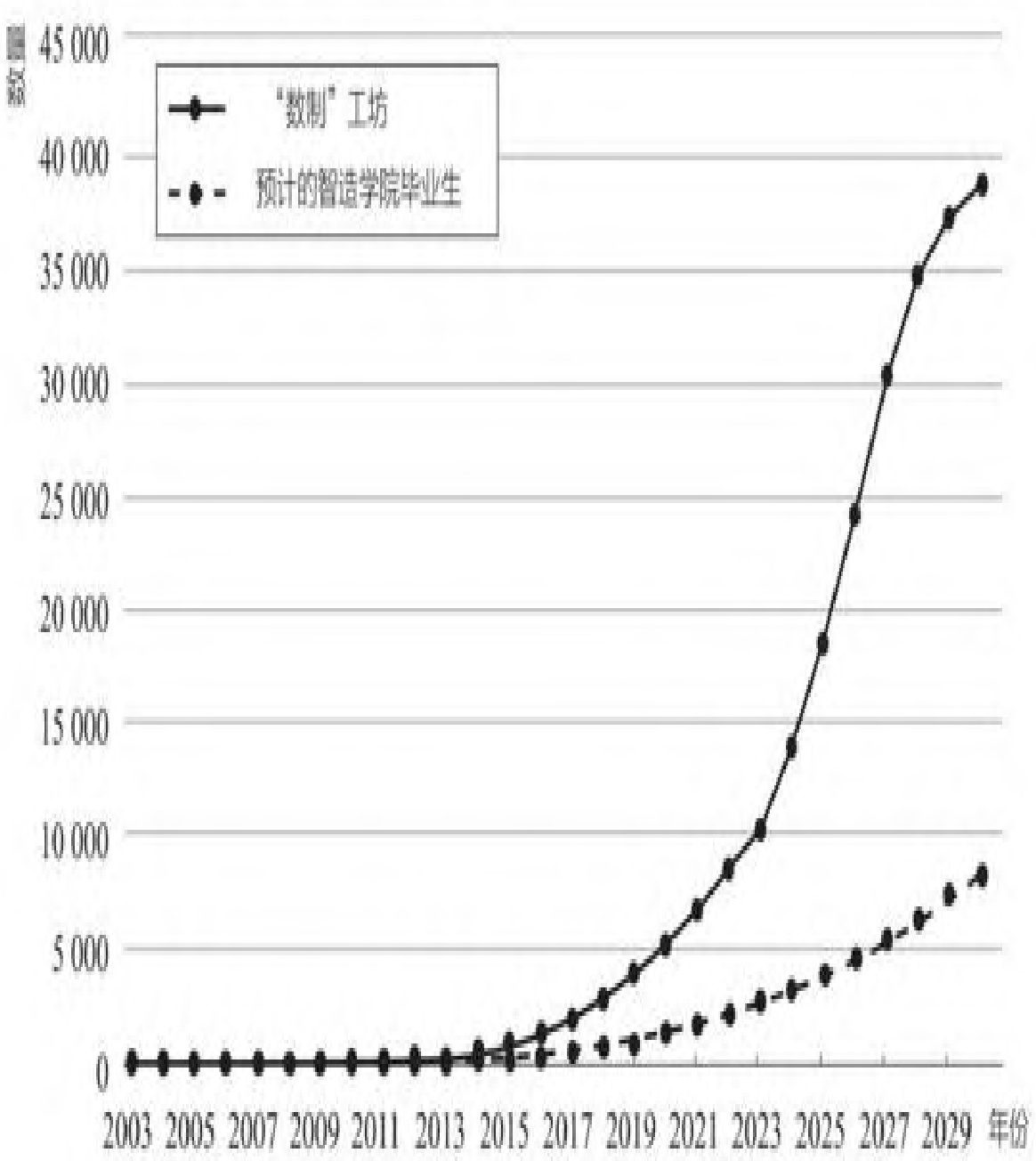


图4－3 “数制”工坊数量和智造学院毕业生数量可能增长的情景。乔尔·卡彻－格申斐尔德绘

投入真心

本书介绍的人物主要是“数制”工坊的先驱者，他们已经为数字制造贡献了力量并给予了承诺。但这仍然是一小群人。如果“数制”工坊要覆盖更广泛的人口，就必须要让那些不了解甚至抵制第三次数字革命的人参与其中。正如阿尔文·托夫勒在前两次数字革命的风口浪尖上指出的那样：“未来的冲击是指我们通过在短时间内对其他人进行太多改变而使他们承受巨大压力并迷失方向。”这种压力和迷失方向的迹象在整个社会中都可以看见。

尽管如此，第三次数字革命对许多人来说依然是一个不受欢迎的过程。事实上，由于孤立主义观点、对技术的不信任、对失业的恐惧，以及对不平等日益加剧的愤怒，世界上大部分地区都有向内发展的趋势。嵌在数字制造概念中的是本地自给自足和全球相互依赖的独特组合，代表了另一条前进的道路——但这并不是一个像在推文、原声摘要播出和信息回声室中进行通信那样简单的概念。

像数制包容性这样的挑战成为全球优先事项，不能仅仅要求更广泛的人群接受更多的技术变革——他们需要成为促进这种变革的参与者，使其对他们自己、他们的家庭和社区产生积极影响。当受益人成为理想变革的关键驱动因素时，社会变革将会更加成功。为了应对这些挑战，我们推荐一种同时利用头脑（以逻辑为中心）、心灵（以情绪为中心）和手（以实践为中心）的方法。

我们都经历过改变目前行为以实现长期目标的困难。无论是为未来的财务健康节省资金，为未来的身体健康进行规定饮食和锻炼，还是采取积极主动的措施来缓解气候变化以确保地球未来的健康，有大量证据表明这种改变极为困难。让人们参与第三次数字革命的挑战更加困难，因为许多人仍在适应前两次数字革命。

尼尔开始了他的实验，他认为参与第三次数字革命的重要性主要在于证据、逻辑和理性。这种以头脑为中心的方法也是许多政府机构和慈善机构的首选模型，它们围绕逻辑模型和投入产出分析来设计其变革理论。对为什么加速数字制造技术能够带来有意义的教育或教育

福利进行强有力的、基于证据的分析是十分必要的。但是，如果你想吸引更多的人，而不是沉浸在细节中，这种以头脑为中心的方法通常是远远不够的。

对许多人来说，吸引人投入真心这一点很重要。只要人类具有沟通的能力，讲故事就是帮助一代人指导下一代克服生活挑战和复杂性的重要工具。在书写发明之前，这些故事以口头故事、仪式和戏剧的形式传承下来。那些经得起时间考验的故事、超越文化和背景的故事，都是以极具吸引力的形式探索普遍的人类主题的故事。有一种说法是“从心灵发出的话语，进入心灵”，我们对它的起源并不清楚。同样，从心灵发出的故事进入心灵。令人印象深刻的是，将数字制造的力量和承诺嵌入深刻的人类叙事，可以激发想象力，并使更多人参与进来。除了吸引头脑和心灵之外，我们还必须与不断发展的技术进行实际互动。了解一个“数制”工坊是一回事，利用这些强大的机器来实现个人有意义的目标，然后与朋友、家人、同事和更广泛的工厂社区分享这些成就，这是另一回事。

前两次数字革命的许多先驱者将其早期灵感指向引人注目的媒体（比如《星际迷航》电视连续剧、艾萨克·阿西莫夫的小说），一些使科学变得生动的导师（例如，伟大的科学老师或充满灵感的父母），以及更重要的动手修补能力（比如家酿计算机俱乐部、电脑套件）。这三种参与模式都很强大，它们在相互加强后变得更加强大。

一个重要的注意事项，是20世纪80年代早期的家酿计算机俱乐部和计算机构建套件面向的通常是白人男性，这种刻板印象今天仍然存在。目前，“数制”工坊的生态系统存在多样性（本书中强调的许多案例研究都表明了这一点）。我们需要强调这种多样性，以继续吸引各种新的参与者。

实际上，连接头、心、手的历史非常悠久。亚里士多德将善良的心和头脑与哲学家的思想联系起来。多年来，宗教组织通过认知、情感和行为的整合将这些概念联系起来。当心理学家谈到认知、情感和

行为改变时，他们会套用类似的框架。这些综合元素也可以在许多现代组织中找到，如4H俱乐部（头部、手部、内心和健康），以及麻省理工学院。麻省理工学院于1864年采用的官方印章的特色是拉丁语“头和手”（Mens et Manus）。“心”这一点被漏掉了。

麻省理工学院的座右铭“缺乏内心”这件事并没有被学生、教师和校友忽视。2015年9月，在一系列学生自杀事件发生后，麻省理工学院校长辛西娅·巴恩哈特和医学顾问威廉·凯特尔宣布了“脑手心”计划，旨在改善麻省理工学院现有心理健康和咨询服务的协调性。2016年5月剑桥市，在麻省理工学院百年校庆活动中，麻省理工学院校友和汽车讲座主持人雷·马格利奥齐与学生制造的机器人合作，公开得出结论，仅仅通过给脑和手添加“心”，麻省理工学院就已经发展为今天的这样一所蓬勃发展的学院。对于第三次数字革命中的个人而言，同时接触头部、心脏和手的行为可以成为加速器；如果不这样做，其行为将变成一个限速器。

生态系统

我们已经研究了个人、组织和机构的变化速率怎样成为限制因素。虽然分别理解每个级别很重要，但我们最终还是需要通过有意义的变化来了解这三个级别如何相互作用。如果数字制造技术要以指数级变化速度加速，并促进更加自给自足、相互联系和可持续发展的社会的发展，就需要在区域和全球范围内实现生态系统层面的转型。

在生物学中，生态系统由相互作用的生物和环境组成。生物学中的共同进化是相互作用的产物。在技术和社会方面，生态系统包括个人、组织和机构与技术和自然环境之间的相互作用。共同进化的发生，有时是计划的，有时是无计划的。为了预测实现社会生态系统变

化将要面对的挑战和机遇，回顾一下替代生态系统的一些历史尝试是十分有益的。

在工业革命的早期阶段，罗伯特·欧文的故事介绍了一种生态系统方法，它可能是社会中另一种经济模式。欧文于1771年出生于威尔士，21岁时成为英国曼彻斯特一家纺织公司的经理。到1799年，作为苏格兰新拉纳克工厂的共同所有者，欧文拒绝了工厂的运营方式，并构建了被我们现在称为社会变革的生态系统方法。这些努力包括在社区中提供婴儿托儿所、工人子女学校教育、安全的工作条件、公平的薪酬、工人教育，此外，还提供工人对所生产的产品质量的反馈以及合作零售店，这样工人就不必购买“伪劣商品”了。想象一下，如果这种变化被其他企业家和影响社会的人接受，它将以更加有益的方式重新定义工业革命。

事实上，欧文的模型已经成为当时工业秩序的阻力。该工厂的其他共同所有者抗议说，他们没有挣得足够的利润。很显然，已经有人期望企业只实现利润最大化。对于欧文的工厂来说，这个问题是在社会改革者杰里米·边沁[9]和其他人的共同帮助下解决的，他们收买了反抗的共同所有者。这延长了这项尝试的寿命，但在其他环境中并不容易复制。新拉纳克工厂成功吸引了我们今天所谓的社会改革者、教会领袖、政治家和其他人进行基准测试访问，这使工厂具有更广泛的影响。

事实证明，模型中与减轻损害相关的方面确实以一些重要的方式对社会产生了影响，但并未创造价值。欧文成为1819年棉花制造厂通过工厂法案的推动力量，该法案的重点是童工和工时。他领导了“8小时工作制”运动，其座右铭是“8小时劳动，8小时休闲，8小时休息”。在这方面，保护性立法可被视为在技术变革时代减轻伤害的加速器。

然而，与创造价值有关的模式的各个方面（包括工人咨询、合作零售店、提高质量和其他社会创新）不能简单地立法或执行。虽然欧

文记录了这些经验，并试图在他的国家和美国推广这些经验，但基准测试与他的书的结合并不足以让大多数人采用、适应和扩展这种替代模式。

如果我们快进到20世纪50年代，我们会发现另一种替代生态系统模型，它能够在区域范围内进行扩展，创造价值并减轻危害。西班牙巴斯克地区的蒙德拉贡合作社由一位天主教神父于1956年创立，目前已经发展为拥有7万5千名员工的数百家中小型企业。该模型在某些关键方面与传统工业模式不同。例如，如果一家公司经营失败，该模型就会为其提供收入支持和再培训，并为想要成立新公司的人提供投资机会。知识俱乐部是一个工作组的集合，作为蒙德拉贡的知识引擎，寻找新的商机和不断变化的市场。所有人都可以享受重要的社会服务。在此背景下，个人、组织和制度变革共同促进了内部增长和长期可持续发展。

虽然它是相对封闭的巴斯克文化的产物，但蒙德拉贡已经接待了成千上万人的基准测试访问，并成为几十本书的主题，将这种模型扩展到其他环境。今天，它正努力地在数字世界中发挥作用。它是一个接受数字制造的环境，这一点已经不足为奇；它致力于研发资源，开发3D打印新材料（于2015年推出一款用于增材制造的新型尼龙长丝）。尽管如此，蒙德拉贡仍然是一个相对封闭的系统，与外部市场相互作用主要是为了销售它的产品，并没有在其区域之外产生深刻的社会影响。

在讨论蒙德拉贡时，正在巴塞罗那领导智造城市计划的托马斯·迪兹指出，从蒙德拉贡模型中可以获得重要的见解，但在智造城市生态系统方面也存在差异。他强调了蒙德拉贡合作社的自给自足、相对封闭的模式与第三次数字革命所需的横向合作模式的明显区别：“蒙德拉贡是一个不同的模式——我们的目标不是成为万物的经纪人。我们可以发挥作用，但我们更有能力。我们正在支持不断发展的‘数

制’工坊城市生态系统，但我们不需要处于中心位置。”在这些评论中，迪兹阐述了在全球范围内传播分布式生态系统的方法。

最近的一个生态系统变化模型是“重塑学习”计划，其重点是教育转型，覆盖匹兹堡地区、宾夕法尼亚州西部和西弗吉尼亚州西部。“重塑学习”计划是一个由200多个组织合作的项目，为该地区的年轻人拓展机会并提高学习效率。它旨在激励一代终身学习者，使他们在21世纪茁壮成长。该模型大致包括150个制造商空间和“数制”工坊。

这个故事始于1997年，当时曼彻斯特工匠协会将处境危险的青年与学徒的培训、艺术教育和其他形式的校外学习联系起来，在长期被忽视的匹兹堡社区领导了振兴工作。2007年，匹兹堡的格拉布尔基金会推出了“孩子＋创造力”活动，以回应教师、图书管理员、博物馆教育工作者和青年工作者提出的一个看似简单的问题：“我不像以前那样与孩子们交流了。”起初，这种担忧似乎与每一代人对青年人的关心都没有什么不同。但进一步的调查显示，在数字时代，对知识的追求正以许多重要的方式出现。

“孩子＋创造力”始于独特的练习——10份煎饼早餐，汇集了来自学校、博物馆、图书馆，早期教育中心和校外课程的领导者。早餐参与者都主动表示他们认识两三个人，这些人也需要参与进来。早餐之后是一个超过100人的大型聚会。正如格拉布尔基金会执行董事格拉布尔所描述的那样，会议上“每个人都有两分钟的时间来发表他们的评论或进行推销，两分钟后会响起响亮的锣声，表明时间已经过去了。每个人的表达都非常厉害”。后来又有许多机构提供了额外资金和支持，包括克劳德·沃辛顿·贝尼登基金会、布尔基金会、麦丘恩基金会、匹兹堡基金会、斯普劳特基金会和阿勒格尼中介机构（一个教育服务机构）。会议还启动了试点实验，内容包括基于项目的学习、基于制造商和工厂的学习、基于游戏的学习以及其他创新教学方法，这些内容通常跨越了以前独立的领域，如学校、图书馆、博物馆、校外正式和非正式学习环境。

该计划于2011年更名为“重塑学习”，引起了全美的关注，并得到了约翰·D和凯瑟琳·T.麦克阿瑟基金会的额外支持。人们没有从教育创新者和教育技术推广者那里出现的许多竞争性术语和教学法中进行选择，而是故意选择“重塑学习”这个词作为标题来描述这个计划。2015年，雪佛龙与卡内基科学中心和智造基金会合作，在该地区建立了一个全面的“数制”工坊（这是雪佛龙向“数制”工坊提供1000万美元承诺的一部分，作为在雪佛龙公司运营的地区支持“数制”工坊的基础）。该工厂与大匹兹堡地区大约150个创客空间相连，通过各种方式连入重塑学习网络，包括科技工坊学习中心、匹兹堡骇客、西宾夕法尼亚州聋人创客空间学校、匹兹堡创造者大会等等。

有趣的是，该计划在2014年才成立了一个顶级领导委员会（由36名地区高级行政领导人组成），在此之前它已有了7年的历史。这反映了该计划在早期具有新兴性质，指出了生态系统中的有利实践（治理）的必要性。现在有一个全职的专业人员和一个由2 000名教育工作者、社区活动家组成的互联网络，其中近40%的人在过去两年内加入其中。匹兹堡是美国30个数字教育集群之一，可以说是最先进的教育集群。外部兴趣水平促使“重塑学习”网络开发了一个剧本，总结了5个领域的经验教训：

学习环境

创新研发

学习奖学金和宣传

商业和企业

战略管理

该剧本代表了一种明确的努力方向，以帮助人们实现超越一个地区的转型变革。虽然书籍（和剧本）的编写和出版、对现场访问和基

准测试的组织、立法的变化等类似于改变规模的方法都是重要的活动，但它们最终限制了变革的速度，使其无法与加速发展的技术共同有效地发展。为了实现这一目标，我们需要不断打磨数字平台和工具，不断实践，以便有效传播模型的关键部分。例如，虽然整个地区有数十个创客空间和“数制”工坊，但在整个网络上共享和协作项目的能力仍然有限。托德·克鲁斯金博士最初作为高中校长参与了“重塑学习”计划，现在担任伊丽莎白前进学区的助理主管。他评论道：

> 对于数字制造，创意项目是关键。智造基金会表示，它将开发一个分享项目的平台，据我所知，这个平台将很快被公布。这很好。我们需要有效的设计挑战，特别是帮助社区的项目理念，但我们一直努力在内部开发它们。

在这里，人们需要通过新的实践——具有挑战性的和有益的项目——来使用相关工具，这些工具可以使用、调整和扩展。由于大多数做法不是匹兹堡地区特有的，因此有可能会在该地区之外传播。

在“重塑学习”的案例中，数字制造是新的教学和学习生态系统的一部分。数字制造平台、工具和实践仍然在各个地区展现其能力，以创建更多自我可持续发展的社区，为我们的生活、学习、工作和娱乐提供新的模式。看看数字技术如何帮助个人或一小群人创建具有全球影响力的真正的变革性生态系统，对加快这一过程是有帮助的。

自然增长与规模增长

前两次数字革命为人们设计和传播变革性生态系统提供了见解。现在，一个人或一小群充满激情的梦想家就可以参与其中并为数百万

人授权，为全球变革性生态系统的设计和传播做出贡献。这些数字生态系统利用热情的贡献者网络，依靠强大的平台和工具实现分布式代理，并为能够以指数级速率传播而不仅仅是线性增长的实践提供前所未有的变化速率。

在第2章中，我们确定了培育一个有利的数制生态系统是一个关键性的门槛挑战，这项挑战是以创建一个更加自给自足、全球联网和可持续发展的世界为基础的。在这里，我们将深入探讨这一概念，包括如何使这些生态系统的属性处于不断出现的状态，这对其指数级传播至关重要。新兴的生态系统与传统业务的规模有一些相似之处，但也存在重要差异。企业可以通过计划实施、多领域研究和设计、供应变更管理、客户支持、国际管理以及兼并和收购相结合的方式进行扩展。以这种传统方式进行规模扩展的企业通常属于资本密集型企业，需要经验丰富的管理者，其扩展过程更像是具有周期性加速度的线性过程。

请注意，这里有两个关键词：传播和规模。它们在尼尔的世界和我们的世界中的使用方式非常不同。我们在撰写本节时被这一点困扰，重要的是要说明差异，因为每个词在不同的世界中都可能以截然不同的方式被使用。对于尼尔而言，物理传播（比如波的传播）是以一个给定的速度从一个给定的源发生的，这与摩尔定律的指数倍数定律的核心形成对比。在我们的世界中，营利组织和非营利组织都试图通过各种机制来扩展其运营规模，这些机制主要涉及自上而下的规划和实施。这与我们在本节中探讨的有机的传播过程形成对比。

新兴的生态系统就像我们使用的这些术语一样，不会扩展，但可以传播。平台、工具和实践与总体层次结构不同。相反，在结构上，它们涉及一种横向对齐的形式——连接独立但相互依赖的参与者。为了使其社会系统能够与技术系统的指数级增长相匹配，我们需要将其传播（通常除了传统的规模扩展之外）。

几年前，艾伦在亚利桑那州图森附近的生物圈2号参加了一个关于适应性的论坛，在那里探讨了这一主题。会议召集了科学、社会科学和人文科学的研究人员，以探讨适应复杂、快速变化的生态系统的策略。其中一项小组讨论涉及从前两次数字革命中观察变革性生态系统，并解构那些帮助它们在持续加速变化的世界中传播并保持弹性的元素。该小组包括生物学家、经济学家、哲学家和企业家。会议结束后，包括波士顿科学公司联合创始人和FIRST公司前主席约翰·埃伯利、NextThought首席执行官肯·帕克和艾伦在内的一小部分人开始记录其中的一些见解。

该小组首先假设人类已成为优势物种，很大程度上是因为我们有大规模合作的能力。数字技术加快了这种能力的发展，使全球变革性生态系统能够在极短的时间内出现，这主要是因为跨越社区的共享和分布式机构由核心平台、工具和实践联合起来。

这种变革性生态系统的示例包括分散的技术运动（如Linux操作系统）、营利的娱乐活动（如《我的世界》）和非营利的网络资源（如维基百科）。

维基百科、Linux操作系统和《我的世界》这三个例子对世界产生了相当大的影响。维基百科是世界上第六大最受欢迎的网站，拥有全球贡献者生态系统。它是数亿人的主要信息来源，每月读者近5亿，页面浏览量达到180亿。Linux基金会最近的一份报告显示：Linux软件通过全球程序员生态系统提供的1.15亿行代码产生了50亿美元的经济价值。《我的世界》是一个由瑞典游戏设计师创建的小型独立游戏，它形成了一个全球生态系统，有一定热情的和深度参与其中的一亿多年轻人通过全球“游戏模组”生态系统为调整和扩展代码加速，以引入新的体验。每个例子都是通过传播的形式而不是以更传统的形式发展的。

这三个示例以及各个领域中的其他示例都拥有一组共同的属性。这些生态系统都由一个或一小群富有魅力的、热情的领导者创立，他

们定义了社区精神、文化和高级别参与规则（公开地或通过他们的行为和实践），这些创始人提供了支持平台、工具和实现增长的生态系统实践。在某些情况下，这些创始人着手创建变革性生态系统。在其他情况下，他们没有这样的目标，因此没有准备好进入这样的领导地位。我们将这些充满激情的领导者称为自上而下的梦想家。在数制生态系统中，尼尔当然属于这一类别。

这些生态系统中的每一个系统都在很大程度上随着被授权和授权的中间层次的个人和组织的激情而增长，所有这些人都深深认同这个愿景和新兴社区，并有效地利用、适应和扩展平台、工具和实践。被授权和授权的中间层通过投入大量时间和精力（通常没有报酬或正式职位）来调整和扩展平台和实践，以满足当地需求或创造新的能力。反过来，他们的奉献精神为生态系统带来了新的参与者，提升了他们在社区中的地位。影响主要来自行动和有利于社区的切实成果，而不是自上而下的法令。电子线传媒和游戏教师公司帮助《我的世界》（教育版）走进成千上万所学校，这就是一个例子。我们将这一层称为授权和被授权的中间人。在数制生态系统中，他们是“数制”工坊的创始人、为开源数制软件和硬件的开发做出贡献的程序员和设计师，以及那些开创数制学术、数制基金和智造城市等新举措的先驱。

最后，同样重要的是，所有这些生态系统都吸引了不同程度的参与者。我们将这些参与者称为自下而上的参与者。在这些新兴生态系统中，每个层级都有参与者的代理机构，每个层级都有对应的责任和能力——所有这些都与共享平台、工具和实践以及共同的文化和愿景保持一致。在第三次数字革命中，越来越多的参与者与全球数制生态系统进行互动。

这些产品和服务的创始梦想家已经开始在整个媒体中庆祝了，因为这些生态系统已经拥有数百万自下而上的参与者（有时被称为“人群中的智慧”）了。然而，认为这些生态系统自上而下驱动或自下而上驱动的想法是错误的。它忽略了关键的中间层的重要性，中间层通

常代表了这些生态系统以快速的、分散的方式传播的秘诀。就像自然界蓬勃发展的生态系统一样，这个中间层创造了多样性、冗余性和适应性，使生态系统能够持续适应不断变化的环境，并为它们的传播增添力量。这个授权和被授权的中间层包括改良《我的世界》的人、维基百科的编辑和贡献者、Linux操作系统和摩斯拉浏览器开发人员，等等，此处仅举几例。

有趣的是，并非所有这些新兴生态系统的创始人都有发起一场运动的打算或改变世界的愿望。最好的例子是马库斯·佩尔松，他是《我的世界》的创造者，他只想制作一款有趣的游戏。当他被推到数千万青少年的虚拟救世主的角色上时，情况变得难以控制了。出人意料的是，他以25亿美元的价格将游戏卖给了微软，这是任何一个关注他与社区的积极对话的人认为他会做的最后一件事。在给粉丝的公开信中，他说明了他的理由：“我已成为一个象征。我不想成为一个象征，不想对我不理解的、不想做的、不断回到我身边的大事负责。我不是企业家。我不是首席执行官。我是一个书呆子一样的计算机程序员，喜欢在推特（Twitter）上发表意见……如果我不小心做了一些似乎有吸引力的东西，我可能会立即放弃它……这不是钱的问题。这关乎我的理智。”

大多数已经达到数千万人的新兴生态系统都集中在软件上。对于从比特到原子的平台来说，关键问题是原子对这种快速传播能力的影响。米奇·雷斯尼克在麻省理工学院媒体实验室担任终身幼儿园小组的负责人，并且是2007年推出的全球流行的青少年计算机编程平台“魔抓”（Scratch）的创建者，他强调了这一挑战。雷斯尼克是在为英特尔计算机俱乐部开发创造性学习体验的过程中提出“魔抓”这个想法的。该网站现在每月获得1亿独立访问者，并且已经开发了超过2100万个“魔抓”项目。雷斯尼克指出：“当英特尔宣布支持扩展计算机俱乐部网络时，我警告他们，根据摩尔定律，社区中心无法扩展，这是对的。但是俱乐部诞生了‘魔抓’，它根据摩尔定律逐渐增长。”

也就是说，物理世界的生态系统也适合新兴的生态系统模型，但其加速增长的速度比领先的数字生态系统慢，这一点与“数制”工坊一样。有一个很好的例子——“FIRST机器人”，一个国际高中机器人竞赛。FIRST公司由发明家狄恩·卡门于1989年创立，旨在“创造一个科学和技术享有盛誉的世界……年轻人成为科技英雄”。FIRST已经吸引了30多万名初中生和高中生，并拥有9万多名志愿者。卡门是一位富有远见的创始人，而且分布式的志愿者网络确实是生态系统中被授权和授权的中间层和力量倍增器。FIRST机器人最让人印象深刻的事情之一，就是将优秀的专业团队和合作竞争文化渗透到整个分布式生态系统中。这个优秀的专业理念来自麻省理工学院工程学荣誉教授伍迪·弗劳尔斯。弗劳尔斯帮助卡门建立了FIRST生态系统，现在担任该组织的顾问。任何参加过最近一次FIRST比赛（整个体育场馆都被坐满）的人都知道，社区已经成功地在整个生态系统中注入了一种合作和尊重文化的竞争环境。

因此，一个关键问题是，数制生态系统能否通过数字和物理元素的混合发展成一个具有全球变革性的生态系统。第二个问题是，这样一个新兴的生态系统能否真正为更加自给自足、相互联系和可持续的世界奠定基础。通过精心设计的制造平台、工具和实践的融合，我们现在可以对全球分布式的网络授权，使每个人都具有一致的兴趣和共同的目标。有灵感的路径创建者现在不仅可以设计创新的社会模型和经济模型，还可以通过具有全球变革性的生态系统传播这些模型。但是，在社会系统能够通过技术的加速而真正发展之前，我们首先需要更深入地了解预计的技术路线图，这是第5章的主题。

（本章由艾伦、乔尔撰写）

---

[1] 又译为《科学怪人》。——编者注

[2] 由克劳德·香农提出，是运用概率论与数理统计的方法研究信息、信息熵、通信系统、数据传输、密码学、数据压缩等问题的应用数学学科。信息系统就是广义的通信系统，泛指某种信息从一处传送到另一处所需的全部设备构成的系统。——译者注

[3] 对美国反主流文化影响最深的期刊，《全球概览》的内容包含各种各样的书籍、机械装置和户外休闲用具，它给读者提供的是产品、意见和视觉设计的混搭组合：家用的纺织套件、陶匠用的旋轮与关于塑料的科学报道相互碰撞，竹笛和介绍电子合成音乐的书被放在一起。——译者注

[4] 美国发明家，瑞典人和挪威人后裔。恩格尔巴特在斯坦福研究院，也就是今天的斯坦福国际咨询研究所任职期间，共获得了21项发明专利。最著名的就是鼠标的专利。——译者注

[5] 《未来的冲击》是托夫勒奠定声名之作，本书的出版也标志着未来学进入美国文化的主流领域。——译者注

[6] 1英里≈1.609千米。——编者注

[7] K－12教育：美国基础教育的统称。“K”代表kindergarten（幼儿园），“12”代表十二年级（相当于我国的高三）。K－12是指从幼儿园到十二年级的教育，即美国公立学校的义务教育。——编者注

[8] 拉尔夫·赛瑟罗恩，美国气候学家。他在大气化学、气候变化和能源问题上的研究使得他在制定科学和环境政策方面处在极高水平。——译者注

[9] 杰里米·边沁是英国的法理学家、功利主义哲学家、经济学家和社会改革者。他对社会福利制度的发展有重大的贡献。——译者注

第5章

原子与比特的进化路线图

探究数字制造对未来的影响一直是科幻小说的热门主题，尽管在小说中它并不总是被称为“数字制造”。在《星际迷航》中，复制器是一种能做出各种物品的方便设备，可以产生任何情节需要的故事元素，片中通常的情节是皮卡德船长又要点一杯热腾腾的伯爵红茶了。相比之下，系列电影《终结者》（*Terminator*）中的“T－1000”则略显不祥，这是最可怕的敌手——一个外形百变的液态金属人，可以变成任何人的样貌，即使其液滴被吹散，也会重新凝聚。复制器和可编程材料这些科幻发明描绘了第三次数字革命的终极目标，当然也希望“天网”的电影情节不会一语成谶。

为了实现这一技术愿景，我在2013年与白宫合作举办了一个研讨会，旨在研究数字革命路线图。活动针对的是几个联邦机构制订3D打印规划的需求。我感觉这些机构既没有掌握当前的数字制造技术现状，又没有洞察长远的制造科学发展趋势。该研讨会确定了4个阶段的内容：

- 通过计算机控制制造东西的机器来制造（就像今天在“数制”工坊中所见到的那样）
- 快速原型机器的快速原型制造（用“数制”工坊制造“数制”工坊）
- 对数字材料的构造进行编码（将“数制”工坊中的所有机器合并到一个流程中）
- 开发编程材料（实现机器和材料一体化）

这些阶段在拉斯定律的曲线轨迹中都有对应的节点：

第一阶段（也就是我们目前所处的阶段）结束的标志是“数制”工坊的数量与城市数量相当——全球几千个城市，平均每个城市都有一座“数制”工坊。

第二阶段，“数制”工坊的数量是现在的1 000倍：由现在的1 000个增加到100万个，这一数量相当于地球上各级地方政府的数量总和，我们可以将“数制”工坊广泛地提供给个人和组织使用。

第三阶段，“数制”工坊的数量又增加了大约1 000倍：从100万个增加到10亿个，这相当于地球上的人口数量的数量级。在第三阶段，“数制”工坊将如同电脑和手机一般，无处不在。这一阶段使我们超越个人制造，实现通用制造。

第四阶段产生了另一个“1 000倍”，从10亿个增加到1兆个。除了用来计算人口，10亿还是在原始地址（IPv4）用完之前可连接到互联网的计算机数量。以平均每个人拥有1 000件物品来估计，1兆就是与人交互的事物的总数或计算机连接的事物总数。到那时，数字制造能制造的就不再是任何一件物品了，而是万事万物。

拉斯定律最初是用来统计“数制”工坊的。这些“数制”工坊包含了创制（几乎）万物的核心技术能力：2D和3D的设计和扫描、增材制造和减材制造、电路的创建和连接、计算的嵌入和编程。随着工坊规模的扩展，我们不再按“数制”工坊的当前形态来进行统计，而是统计可进行实物创制及功能编程的等效能力。正在发生改变的是工坊的输入而非输出：今天的“数制”工坊依靠全球供应链来获取诸如集成电路、精密工具和聚合树脂等产品，随着时间的推移，产出将由下一代“数制”工坊通过对更加基础的模块进行组装来形成，从而脱离如今依赖大规模投资的模式。

有两种方法可以预测从第一阶段到第四阶段的转变需要的时间：比较保守的方法是假定拉斯定律仍然有效，数量倍增时间为一年半，那么工坊数量增加到原来的1 000倍大约需要15年。也就是说，就我们目前所处的水平来看，实现转变需要45年，这跟摩尔定律的运行轨迹是一致的。更为大胆的预测认为，所有4个阶段的技术在今天的“数

制”工坊中都已经可以看到，推广起来只是个速度问题。我们无法回答哪一种预测会成为现实，不过我们可以浏览一下每个阶段的情况。

社区制作：工具

数字制造路线图的第一阶段始于今天“数制”工坊中的工具。这一阶段属于未来——正如作者威廉·吉布森所言：“未来已来，只是尚未流行。”由于大部分人都没有去过“数制”工坊，因此了解路线图的第一步是了解它的起点。

切割

切割工具以二维空间中移动，从平板上切割零件。虽然这种方法听起来效用有限，但切割速度比其他工序要快得多。此外，切割工具可以处理其他工具无法处理的材料，而且被切出来的整套零件可以被快速组合成3D物体。



图5－1 “数制”工坊工具。弗罗斯蒂·吉斯拉森摄

到目前为止，“数制”工坊最受欢迎的工具是激光切割机。它唯一的输入是电（切割原料除外），它唯一的输出是切割过程中产生的废气（切割出的部件除外）。在几分钟内，激光切割机就能绘制出像激光束一样精细的复杂形状，精确度达到千分之几英寸[1]。价值1 000美元的激光切割机可以切割薄木板、纸板和塑料；价值10 000美元

的激光切割机可以切割更大、更厚的板材；价值10万美元的激光切割机有更大的机床，可以切割金属。

最通用但最不受欢迎的切割工具是电动刀，它通常被称为乙烯基刀，因为它最常见的用途是为标签或贴纸切割乙烯基字。但这些工具也可以剪掉丝网印刷用的掩模，刻划折纸痕迹，添加剪纸切口，以及在导体中绘制柔性的电路和天线。最大的工具可以处理大型纸板，以切割和折叠陈列箱和家具。

这些工具中最强大的是水射流切割机。它射出一股澄清的超音速水射流，水中携带石榴石磨粒，这些微粒具有极快的移动速度，可以穿透任何东西——厚钢、玻璃、陶瓷、石头。凭借这种多功能性，水射流切割机正在接管其他更专业的工具的工作。缺点是，它需要提供稳定的磨料流，因此会产生同样稳定的废物流。

最后一组切割工具使用的是金属丝。有一种金属丝是加热型的，可以将泡沫一样的材料熔穿。在电火花加工过程中，细金属丝通电，产生微小的火花，腐蚀最坚硬金属中的深切口。因此，电火花加工是制造机器所需的精密零件的首选。

铣削

铣床是在三个而不是两个维度上移动、旋转、切割的刀具，更高级的铣床还可以倾斜刀具或工件，同时控制四个轴或五个轴。这些铣床中有小的台式机器，还有一间房那么大的大型机器。这些机器的加工范围可大可小，小到电脑桌，大到房屋。它们相应的切割扭矩也由工具的大小——小到小型手摇钻，大到大型发动机——来决定。

铣床既嘈杂又凌乱。切削过程很嘈杂，驱动它的电机很响。被切开的材料会像碎片一样飞走，必须想办法收集和处理：根据被切割的

材料，在铣床上喷洒润滑剂。铣床需要各种各样的辅助工具，这些工具需要定期更换。

但作为回报，铣床可以使硬质材料表面光滑，具有精确性。这就是为什么它们是制造汽车和飞机部件、高端笔记本电脑外壳，定制家具和其他生产过程的首选工具。体积更小、精度更高的铣床可用于切割电路板上的痕迹和制作铸造零件的模具等工作。

打印

3D打印机存放材料而不是移除材料。增材制造与减材制造有两个重要的区别：为了制作一个球窝接头，铣床必须切割出球，然后是球窝的半球，且球窝的半球必须围绕球进行组装。但由于3D打印机可以分层组装零件，所以它可以分辨物体的内部和外部，这样它就可以使球的套接一次到位。此外，它只能在零件的位置沉积材料，所以除了一些额外的用于建造支架的材料外，不会浪费其他材料。

首次3D打印是在1986年通过立体光刻技术完成的，这是一种通过激光束的扫描来选择性地固化聚合物液体层的过程。由于激光束尺寸决定了分辨率，立体光刻仍然被用于制造具有最精细特征和最光滑表面的零件。美中不足的是，产品在出来时是一团黏糊糊的、需要清洗的已固化和未固化的聚合物，并且其机械性能受限于光聚合物。第二种方法是在1989年开始使用的熔融沉积成型。打印机将一根聚合物长丝送入加热的喷嘴，挤出一个在生长部分固化的薄珠。这并不像激光束那么精细，但是这个过程现在可以使用更强的热塑性塑料，省去了很多麻烦。熔融沉积成型由于其易实现性，常见于入门级的价值几千美元的3D打印机。最新的一种方法是使用类似于喷墨打印机的打印机头，但它喷射的不是彩色墨水的液滴，而是沉积聚合物的液滴。就像喷墨打印机可以混合多种颜色一样，这些打印机可以混合多种材料。最昂贵的3D打印机用更强大的激光来选择性地烧结细粉层，这些打印

机现在可以制造坚固的金属部件和坚硬的高温陶瓷部件。但这类打印机的缺点是它们非常复杂，耗资可能达100万美元，需要专门的房间和技术人员来运行。

虽然3D打印机擅长创建复杂的几何图形，但是嵌套的特征不能很好地表现出来，这些工具速度慢，通常需要几个小时甚至几天的时间才能制作出一些东西，而且打印机使用的材料比散状材料贵得多。其他工具可以更大、更快、更便宜或更强。这就是为什么当所有其他的数字制造工具都被广泛使用时，3D打印机只能用来做最适合它们的工作。在我经营的比特和原子中心的商店里，使用3D打印机的概率可能只有1/4。

扫描

与3D打印相反的是3D扫描，即通过物理对象创建数字模型。扫描的目的包括复制没有数字设计的对象、将物理对象雕刻作为设计工作流程的一部分、为计算机图形创建现实资产并保存有价值的构件。

最简单的3D扫描仪就是照相机。通过从多个方向拍摄的照片，摄影测量算法只需要用2D图像集就可以推出物体的3D几何图形。摄影测量不需要专门的设备，也没有地点限制，但可能会因为视觉伪像的干扰，只能对确切的形状进行近似的估计。更精确的扫描可以通过激光扫描仪来完成——激光扫描仪跟踪激光束在物体上的扫描位置，并通过结构光扫描器在物体上投射图形的方式来实现精确扫描的目的。

但是这些方法仍然存在视觉伪像的问题，例如，一个闪亮光滑的表面会反射探测器的光线。人为视觉效果可以被固定在一个“灯光舞台”上，“灯光舞台”可以同时照亮和收集来自各个方向的光线。“灯光舞台”不仅决定了几何图形的形状，还决定了计算机图形模型的所有光学特性，但是它需要更大、更精确的设备。最先进的3D扫描仪使用X射线来重建物体的内部和外部，在计算机断层扫描仪中从不同

方向获取投影。最先进的扫描仪的价格也是最昂贵的，成本大概在100万美元左右。如果你能负担得起，那就太好了。

制模

前面叙述的这些技术不适合被区分为增材工艺还是减材工艺，因为它们包含了这两者的各个方面。但制模工具批量生产的是你身边的大多数产品。数字制造是一种直写技术，这意味着它制造的所有产品都不同。但是如果你想要大量生产同款产品，数字制造可以通过制造一个模具来批量生产，以节约每个零件的生产时间和成本，还能处理产品表面的质量问题并提高材料性能。

制模过程中，首先要对模型进行数控加工或切割和折叠。当然这些模型也可以采用3D打印技术制作出来，但是前一种方法移除材料的速度通常更快，而且做出来的模具更加光滑和坚固。然后对原材料进行浇注、注射、充气、拉伸、旋转或挤压，做成模具。塑料、金属、泡沫、食品、玻璃和混凝土都可以用来制作模具。就制作时间而言，手工制作与3D打印制作相差无几，但后续对模具的使用可能需要几分钟甚至几秒钟，时间的长短取决于原材料。过去，模具制造需要投入大量的生产成本，但随着低成本数控铣床的出现，生产模具的效率有所提高，时间周期大大缩短。

为了制造更大、更轻的产品，我们可以在制作模具时加入强力纤维，并且将其嵌在压缩性能较强的树脂基体中，从而生产复合零件。广泛使用的纤维有碳纤维、玻璃纤维和天然纤维，由环氧树脂、塑料和天然树脂作基体。复合材料因其可以减轻重量、集约零件和提高汽车性能的特性，迅速应用于飞机、假肢和各种运动器材。

计算

前面的所有工具都是基于数字制造来完成生产的，即先使用计算机来设计模型，然后通过操控机器来制造模具，最后通过把计算嵌入制造来完成连接。

快速原型制造工具尺寸减小和成本降低的原因是用来控制执行器、读取传感器、解释命令、与用户交互以及用于网络通信的计算机的尺寸和成本的减小和降低。现在，这些相同的功能应用于一系列原本无生命的物体中，这些物体正在成为物联网的一部分，例如建筑物中用于节约能源的智能基础设施，以及用于医疗保健的非干扰性医疗监测设备。

如今，业余爱好者使用小型单板计算机（如阿德伟诺和树莓派）将智能材料应用于快速原型项目。这些计算机的成本高达几十美元，由更小的叫作微控制器的集成电路组成，这种集成电路是一种既简单又完整的芯片上的计算机系统。微控制器的成本范围也从数美元下降到数十美分。

我们可以通过化学蚀刻大量生产电子元件依附的电路板，但是这个过程会产生需要强制收集并安全处理的有害废弃物。对于短期生产，我们可以直接用数字制造工具制造电路板：铣床可以在镀铜板上刻痕，乙烯基刀可以在柔软的铜板上绘制，激光切割机可以利用光学特性对铜板进行刻蚀，某些类型的3D打印机可以挤压导电材料。然后，我们可以使用表面装配返工工具手动连接电子元件，对于多数更大型的电路板，我们可以将其与“抓取放置机器人打印机”自动连接，该打印机可以把磁带卷轴里的电子元件放在电路板上。

如第1章所述，由于我们能够快速地将可编程电路添加到快速物理制造中，因此有可能生成完整的功能系统。接下来要发生改变的是“能够制造什么”而不是“需要制造什么”，首先我们要替换满屋子的工具，然后替换材料和零件的库存。

数制：一个工坊制造另一个工坊

路线图中的第二个阶段是由一个“数制”工坊制造另一个“数制”工坊的能力来决定的：这将不再通过现有机器的复制来完成，而是通过结合它们的功能与模块化设计及结构来实现的。我们不再需要购置（和安装）许多独立的机器，这种情况与个人电脑进入计算机历史的时期相似。个人电脑最初被早期开发者使用，后来在全社会范围内被广泛应用。

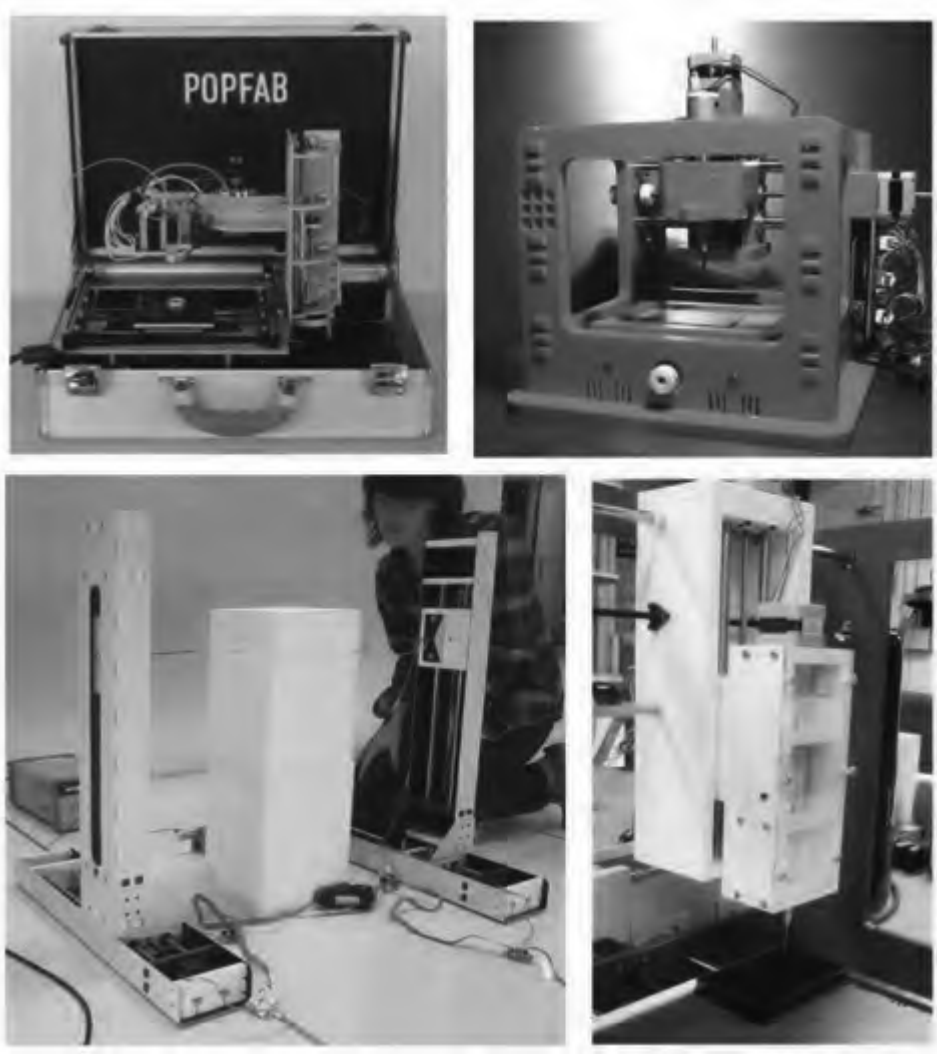


图5－2 用“数制”工坊机器制造的“数制”工坊机器。纳迪娅·皮克摄

第一个“1 000”可以大概解释为平均每个城市大约有一个“数制”工坊。下一个阶段便是“1 000×1 000”个“数制“工坊，这样人们可以为各种规模的社区配备本地生产资料。3D打印机的销售量已达到每年数十万台，预计几年内将达到数百万台。所以，显而易见的结论是：目前，“数制”工坊中的其他工具也将被出售，因此全套产品将会随处可见。

但是这样的情况会带来两个问题。按当今的价格计算，那么多台机器的复制成本将达到千亿美元，这样的金额将会在全球经济中占非常大的比例。而且，在开展巨额投资的同时，有必要为这些机器找到相应的存放空间。

解决第一个问题时，我们不禁想到第一次数字革命和第二次数字革命中解决这一问题的方案：规模经济。快速原型机器的材料花费可低至数百美元，但是目前的成本远高于此，因为不仅需要收回开发成本，还要维持制造它们的小型企业的发展。许多小型企业的目标是发展成大公司，大规模生产这些机器的目的是满足消费者而不是满足利基市场，就像IBM和苹果公司将个人电脑的制造推向规模化一样。

但这种直接的规模经济的解决方案并不适用于此，我们还有另一种选择：“数制”工坊的技术目标一直是制造另一个“数制”工坊。如果仅仅大规模地利用能够制造更多机器的机器，那么制造这些机器并不是一个明智的商业计划。我们可以批量生产其组件，然后在当地组装，而不是直接批量生产机器。我最初忽略的是这个想法——如何通过重新定义机器的概念来解决第二个问题，即这些机器占用的空间。

为了推动“数制”工坊从购买到制造的转变，在2004年，我开始定期跟踪我的快速原型制作课程（“如何智造（几乎）一切”）中关于机器制造的内容——“如何制造制造万物的机器”。纳迪娅·皮克和乔纳森·沃德两位学生在课堂上制造了一台叫作MTM Snap的机器（MTM指的是“制造机器的机器”，Snap是指它不是用紧固件把机器连

接在一起，而是由用来制作厨房切割板的致密塑料——高密度聚乙烯零件拼合在一起的）。MTM Snap是一种精密台式铣床，其规格类似于“数制”工坊库存中价值几千美元的机器，但成本仅为其1/10。只要我们发布计划，这种机器就可以在任何“数制”工坊中被制造，这样的话我想我们的工作已经完成了。

可事实并非如此。无论你是纳迪娅还是乔纳森，只要你是“数制”工坊里的任何一个人，就都可以完成这样的制作。大多数人都需要很多的知识和技能来成功地复制这样的制作过程，因此有很多失败的案例。这一设计被其他机器公司分离出来并作为成品将其商业化。此外，该设计的变体是ShopBot公司的Handibot机器；另一位听过该课程的校友马克斯·洛博夫斯基创立了Formlabs公司[2]，以销售高分辨率的立体光刻3D打印机；荷兰“数制”工坊的机器制造最终促成了熔融沉积成型3D打印机系列的问世。

虽然这些机器中每台机器都比上一代机器更快、更好、更便宜，但所有的机器都有一个隐含的假设：它们要按程序进行工作。这一观察听起来可能是赘述，但是与计算机软件和网络的历史进行类比，这既解释了如何简化它们的结构，又阐明了如何减少对多种不同类型机器的需求。

软件最初是作为完成某件事情的程序而编写的，然后通过编写不同的程序来执行其他操作。事实证明这非常低效，因为代码的重复使用率很低。因此，我们不会再用这样的方法。相反，我们可以利用可重复使用的软件包来编写所谓的面向对象软件，这种软件包结合了使用函数、运行数据以及输入信息和输出信息等功能。我们可以组合这些软件对象来执行更重大的任务，例如，如果您正在创建一个网站，其中一个程序可以负责网页服务，另一个可以执行电子商务交易，还有一个可以跟踪用户的身份。

计算机网络最初是为了完成某些特定工作任务而开发的，例如运营工厂或银行。这些专用网络已经在很大程度上被通用互联网取代，

互联网作为其核心架构原则之一，被称为端对端原则。互联网的应用取决于“连接什么”而不是“如何连接”。老式拨号电话只能通过中央电话局交换机的程序完成拨号。但是现在，我们可以利用联网的计算机执行编写好的程序，直接使用视频或加入聊天室。将应用程序在网络中边缘化，这一原则并不明显，且颇受争议。虽然互联网在某方面比那些专用功能稍逊一筹，但是互联网几乎可以做好任何事情。

我们将机器的结构与计算机软件或网络的结构进行类比，发现3D打印机可以尽可能快地移动挤压机，但是我们也发现数控铣床必须足够坚硬才能承受主轴上的切削力。而这两种类型的结构都需要运动系统来移动其头部以及解释发送给该处的命令。使用相同的制动系统，仅仅改变头部，就能使一台机器完成两项工作，这对于除了要求最高的应用来说都足够好。这一想法促使纳迪娅和机器制造班的另一个学生伊兰·莫耶（这名学生建立了另一个机器公司——Shaper Tools）在2012年制造了PopFab，这是一种快速原型机器，而且它可以折叠成公文包大小，并可以根据3D打印、数控铣削和乙烯基切削的需要切换不同的打印头。PopFab就像制造业的笔记本电脑。

一旦包含了机器中的智能，这种架构类比就变得明确。每种类型的快速原型机器都通过指定的语言执行操作：数控铣床通常采用G代码，这种代码是一种古老的格式，它被应用于早期的计算机化照片绘图仪和织物切割机；激光切割机最常见的格式是HPGL，这是惠普公司用来与现在已经过时的笔式绘图仪进行连接的图形语言。然后，机器中的计算机将这些命令转换为每个组件的指令，以此协调电机，使其转动，但所有这些配置的规范使任何事物都很难改变。例如，你想在切割或打印时添加一个电机以旋转零件，此时，电机就像一个老式的电话，它本身无法完成操作——这时你必须更改传送到机器的语言，然后将其翻译成机器可读语言，这样电机才可以转动。无数的委员会试图通过创造一种新的制造业通用语言来解决这个问题，但因为人们不知道制作的范围以及怎样去做，所以这一直是一项难以实现的任务。

当我们未能在“数制”工坊复制他们现有的机器时，我正在做早期工作，也就是所谓的物联网。和我一起合作的有两个学生，一个是拉菲·克里科里安（这名学生后来构建并运行了Twitter的计算基础设施），另一个是互联网的建筑师丹尼·科恩。我们发现，只需要用价值不到一美元的芯片和几毫米长的硅就可以获取互联网的服务器使用协议，这意味着日常物品可以连接到互联网中。这样，开关和电灯、温度传感器和加热器之间的连接就可以通过软件来选择，而不是由房子中的布线来决定。这个想法可以应用于快速原型机器的各个部件。我们将机器中的所有传感器和执行器连接到实时网络，这样应用程序可以直接与设备对话，而不用通过解释程序（这会导致机器在运作过程中产生一种误解，类似外交中产生的误解）来进行。我们可以通过改变连接到机器的网络来增加新的特性，而不需要改变控制器硬件。

有个很类似的历史事件，纳迪娅和伊兰，以及詹姆斯·科尔曼［札纳（Zahner）公司的首席研究员，该公司旨在制作建筑规模类似于弗兰克·盖里[3]作品的物体的快速原型］开始构建面向对象的硬件。每个模块都各自执行一个物理任务，例如移动轴或转动主轴，同时它们也是通信网络中的节点和控制程序中的软件对象。现在，机器失去了它们的固有身份。对于一个特定的应用程序，您可能需要一台可以快速移动或具有很大的力量的机器，您可能想要打印或剪切，或者您可能正在制作一个2D形状或一个复杂的3D形状。这些任务由一个通用的硬件和软件的构建模块组合来完成，随后又可以被重新应用。这就成了快速原型机器的快速原型。

作为一项试验，我们试着把这些机器制造组件的工具包发送给“数制”工坊的学生，他们很快就制造了一系列新颖奇特、简单实用的制造机，我们没有重复修改同一个常见的机器制造任务的解决方案，而是为他们提供了具有用来驱动机器的集成附件的电机、电机通信的网络接口以及控制一切的软件组件。如此一来，学生可以专注于有趣的问题，比如“他们想让机器做什么”。

延斯·戴维克用两年时间去开展世界“数制”工坊巡回展示，之后开始在奥斯陆进行“数制”工坊运营，同时，他完全在一个“数制”工坊中生产了多种类型的模块化运动部件，并进行了参数设计，使它们的比例可发生变化。在这一过程中，他几乎不需要购买制造一台机器所需要的所有零件，只需要购买电机和电子元件。

因此，通往100万个“数制”工坊的路，并不意味着订购100万台不同类型的机器，每台机器都只做一件事情，而是一台机器生产数百万个模块化组件，这些组件可以组合起来制造许多不同的机器。这些组件可以通过模块化设计在本地生产，以实现最大程度的定制化和独立性。模块可以单独订购和组装，以获得最大的灵活性和便利性。我们也可以对这些模块的特定组合进行固化和批量生产，以实现最大程度的集成和最高的效率。建造这样一台机器是一个动态而非静态的过程，它随着你制作的东西（一条电路、一块蛋糕，或者一个沙发）变化而变化，也随着移动范围、移动的自由度以及执行制造操作的末端执行器的变化而变化。

如果你看看引擎盖下，更准确地说，是看看机器的内部，会发现类似的事情正发生在计算机个人化的过程中。在20世纪80年代，当我还是一个物理学研究生的时候，计算机运算仍然通过小型计算机来完成。每一个计算机都有其独一无二的子系统，并且没有交互操作的能力。这正如今天的“数制”工坊，机器间唯一的整合方式就是通过那个在它们中间进行工作的人来操作。我是我实验室里第一个拥有个人电脑的人，此时个人电脑刚刚开始出现在像我这样的早期“收养者”中，仍远未普及。与小型计算机不同的是，这些个人电脑都是由标准的部件构成的。主板有几种标准形式，带有可更换处理器的插槽。有标准的内部连接器，用于插入不同数量的内存、图形和存储器；也有标准的外部连接器，连接键盘、鼠标和显示器。爱好者们仍然各自选择部件并且自己组装；小型计算机公司从其他供应商那里进货；大型计算机公司可以自己生产部件，这些部件为最高容量的市场进行了预先集成，但是整个行业都是围绕模块化的标准不断发展的。

类似的事情正在智能手机中发生，第一个就是谷歌的项目Ara［以我的学生阿拉·克拉恩（Ara Knaian）命名，他开发了一种可编程的机械连接机制］，该项目表明：决定你的手机性能的不是大规模的制造商，而是你自己从一套可重组的模块中进行选择的结果。根据需求的变化，你可能需要更长寿命的电池、更高分辨率的屏幕、在多个载体上工作的收音机，或者更好的相机。

数码摄影的发展也遵循了从特异性到通用性的路径。我的第一个数码照相机是一个封闭系统，我在拍摄颗粒的照片时用光了内存和电量，惹恼了我的家人。这些具有固定功能的相机很大程度上都已经消失了（生产它们的生产线也一起消失了）。对于常规拍照，相机已经与智能手机中的通信和计算功能相结合。对于更加专业的摄影，行业内已经确定了一系列关于相机机身类型的标准，包括可安装一系列镜头的底座和像照相机闪光灯那样的附件的接口。

个人制造中的主板或照相机的机身，就相当于运动系统。运动系统在制造万物的设备中很普遍，大小不同，形状各异，速度有别，但在其他方面是一个通用平台。与照相机镜头或计算机显示器并行发展的还有终端效果器，它连接到运动系统中进行增材或减材加工。就像你可以购买电脑或照相机一样，你可以购买所有这些预先打包的元素，或者，为了提升性能，增加功用，你也可以购买系统中的零部件。这样，与购买电脑或照相机不同的是，你可以更进一步，用你买到的东西来制造更多你刚买的东西。

个人电脑组件的商品化，对计算的另一端产生了意想不到的影响，尤其体现在一些大型组织视为核心的大数据中心上。我的学生杰森·泰勒目前是脸书公司计算基础设施部门的负责人，领导着一个开放计算项目。该项目设定了一种高度集成的、模块化的个人电脑，可以通过大量的联机运算建立数据中心，取代传统的大型机。需要进行云计算的用户可以根据需要在这些个人电脑上分配工作，服务提供商也可以对数据进行扩容，以满足用户的需求。越来越多的组织发现，

以这种方式购买计算服务比购买计算机本身更便宜、更可靠、更灵活。

有一个新兴的“云制造”的概念，其目标是为制造做同样的事情。为了实现个人智造，它遵循标准化、商业化和模块化的步骤。通过这种方式，云制造能够提供创造能力，刺激创造力的增长，可以连续不断地从个体的需求延伸到公司需求。容量和功能可以远程共享，也可以在本地引入，例如，更多的面包店可以通过组装定制的制造机器来执行重复的劳动密集型任务，这些任务是业务中的瓶颈，应增加可访问的自动化工具而不是替换其工人。

转化：虚拟到现实

这个路线图的第三个阶段是通用制造，即材料和设计从虚拟转化为数字化的现实。在从社区制造到个人制造的过渡中，多种机器被合并成一台；在从个人制造到通用制造的过渡中，多种工艺将被合并成一个过程——离散构件的组装和拆卸。这一阶段反映了计算和通信技术在智能手机中变得十分便宜并易于使用，即市场开始趋于饱和。要使数十亿人而不是数百万人都能利用“数制”工坊的能力，就不能再需要现有的广泛供应链或再产生相同的持续废弃物流了。关闭从生产到消费的循环，需要将必要的投入减少为一小部分零件的原料，并且通过可逆的拆卸过程来代替回收。

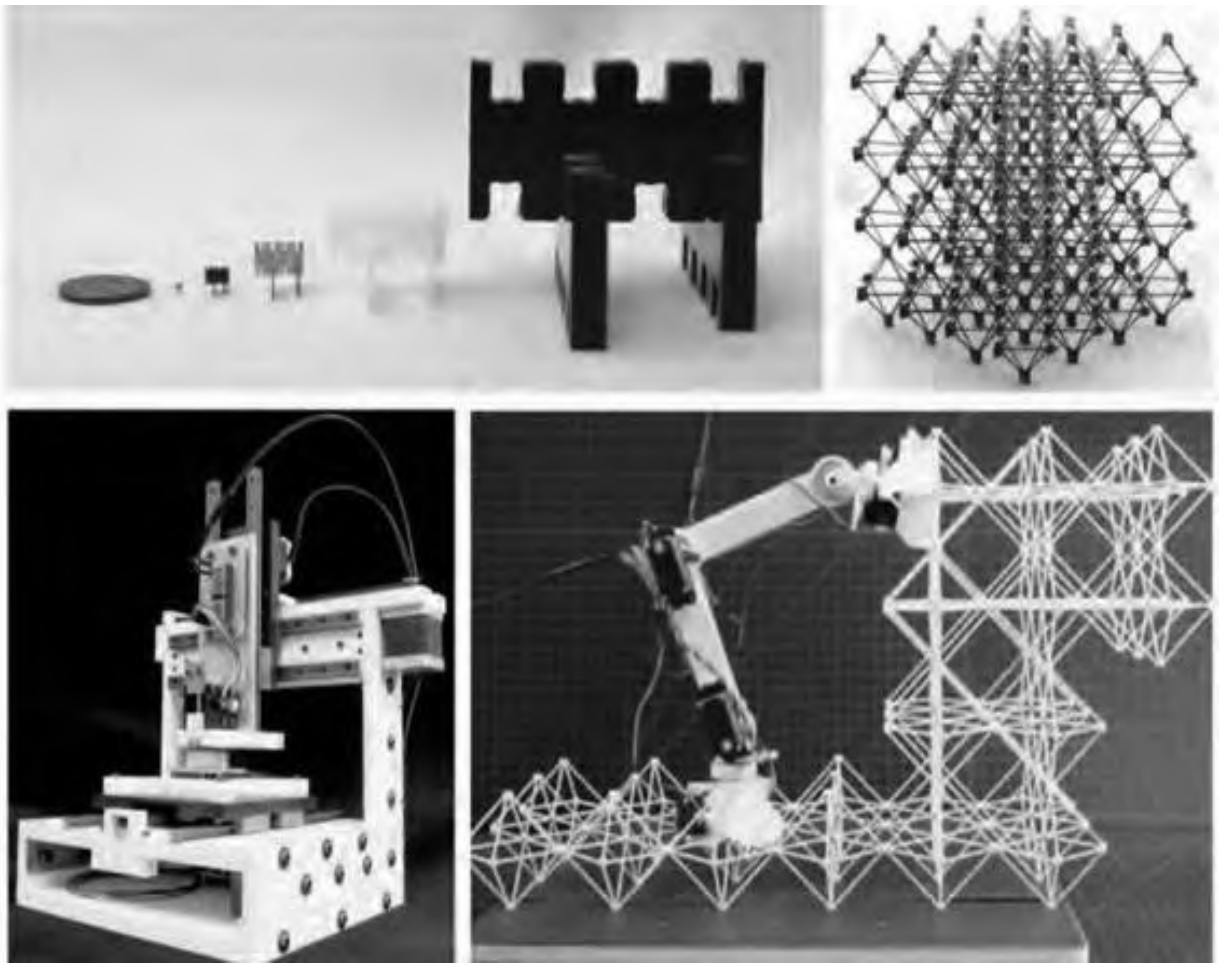


图5－3 数字材料及其“装配工”。（从左上方开始顺时针方向）普拉山特·帕蒂尔、张永安、本·耶内特、威尔·兰福德摄

2006年，我在数字计算和数字通信并行发展的启发下，给数字材料下了一个定义：由一组离散零件构成的材料，以一组相对位置和方向的离散集合方式进行组合，其过程是可逆的。这些属性允许通过局部约束确定整体的几何形状。组合过程中的错误可以被识别和修复，不同的材料可以被连接起来。最终，材料可以被拆卸和重复利用，而不会被简单丢弃。单个要素可以大规模生产，定制是通过要素的组合方式来实现的。成型、轧制、冲压和合成等工业生产过程可以低成本大量复制零件。乐高积木和氨基酸就是这类系统组件的典型例子。

还不能确定的一点是，通过离散装配和拆卸过程而非连续的增材制造和减材制造可以制造出多少其他东西。现今制造的最高性能结构

之一是大型喷气式飞机的机身。为了减轻飞机的重力，同时仍然保持其承载巨大负荷的能力，制造商现在通过将碳纤维的长丝缠绕在环氧树脂基质中的方法制造机身。最新的飞机——波音787飞机和空客A350飞机——都需要在其供应链上投资数十亿美元来生产复合零件，它们需要一个机身或机翼大小的工具来铺设纤维丝，需要另一种更大的工具来压缩和固化树脂基质中的纤维丝。然后，这些巨大的零件必须通过长距离运输才能进行最终组装。

2008年，在制造过程中，我面临了一个严重的问题：如何在不引入接头薄弱部位的情况下连接复合材料零件。现在，我们可以通过钻孔螺栓或胶水将零件连接起来。尽管模型飞机已经能够通过3D打印技术被打印出来，但仿佛也没什么效果，原因有两个：一是3D打印机的尺寸较小，远不及大型喷气机，二是它们无法匹配复合工具的材料性能。

2012年，我和我的学生张永安（现在正在进行美国国家航空航天局的一个制造研究计划）发现，与其缠绕一根长纤维或挤出大量短纤维，不如组装比结构尺寸小的纤维环，这对于机身来说相当于几厘米宽。基于数字材料的概念，我们大规模生产碳纤维复合环，然后将它们连接在细胞结构中，这创造了最高模量（最硬）超轻材料的世界纪录，因为我们没有加入几个特定的大部件，而是制造了能以任意大小或形状组装的稀疏填充的体块。我们现在正在开发类似于机器人的核糖体来放置这些部件，这些部件通过附着和相对于它们正在构建的结构的移动来工作，这样，这些“组装器”通过提供纤维环就可以构建整个机体。对于这个应用，不需要放置单个原子——纤维环是基本的构建模块。通过添加第二种部件类型——可以弯曲，无须始终保持僵硬，我们证明了这些结构能以设计规定的方式变形，例如，飞机机翼能够像鸟的翅膀一样不断地改变形状，而不是通过旋转现有机翼上的刚性控制襟翼改变形状。变形机翼一直是航空领域的长期目标，它们更加高效灵活，但先前的尝试受到了所需机制的重力和复杂性的阻

碍。我们的机翼由这些模块组装而成，所以它们成了有效的机械装置。

我们可以通过组装较小的部件来制造更精细的物体。虽然您可能希望在光滑的表面上看清其离散结构，但请记住以前在荧光屏上观看的视频和利用胶片拍摄的照片。从早期的粒状数码相机和块状计算机显示器开始，现在的相机和显示器的分辨率已经可以超过人眼的分辨率了。因此，图像几乎被普遍记录下来，并被视为离散图像元素，这种离散的图像元素被称为像素。数字材料由离散的3D体积元素（或体素）构成。一旦这些元素的尺寸缩小到百万分之一米左右，它们就不再能被我们的肉眼所感知了。

除了用于制造结构既刚性又灵活的零件，我们还可以使用更多类型的零件制作更多的东西。如果你聚焦于一台计算机，首先就会看到机箱，然后是连接在印刷电路板上的连接器上的线束。电路板上有独立部件和集成电路。集成电路包含的无数元件组成了集成电路中最小的部分，其尺寸接近1纳米，即1/1 000 000 000米，相当于数十个原子排列在一起的长度。但向数字材料的过渡不需要从这些最小的特征开始，部件可以分阶段更换。

线路和连接器的标准间距是1/10或1/50 000英寸。得捷电子[4]（Digi-Key）是我购买的电子元件的供应商，拥有50万个连接器。这不是库存量，而是不同模型的数量。不同模型的区别在于其不同的位置数量、行数以及运行方式。它们都可以只由两种较小的部件组成，即导电部件和绝缘部件。然后，连接器被连接到电路板上，电路板上具有蚀刻于薄片层上的导电迹线，其标准尺寸为1/10 000英寸。就连制造过程极为困难的层间运行通孔以及电路板，都可以再由导电部件和绝缘部件组装而成。

电路板上也有诸如电容器、电感器和电阻器这样的分立元件。电容器包含电场，用于滤波器和电源。得捷电子公司拥有大约50万个容量和包装尺寸各不相同的电容器。同样，这些也可以由导电部件和绝

缘部件组装而成，对于包含磁场的电感器亦是如此。如果部件的尺寸达到1微米，即1/1 000 000米，部件的电容和电感的密度就有可能与如今的制造方式相匹配。调节电流的电阻器需要另一种部件类型——纯电阻。通过结合电阻、导电部件和绝缘部件，你可以匹配得捷电子公司库存中的50万种电阻的价值和包装。你能想到这一点吗？

到目前为止，我们最多使用了三种类型的零件来替换数百万个组件。用于逻辑电路的晶体管需要更多材料——导电性可变化的半导体，掺杂的版本可以提高消耗的电子的数量。移动物体的发动机需要更多的部件——永磁体和引导磁场的材料。总之，这个部件目录中包含了几十种类型，可以重现大部分现代技术的功能。这恰好与生物学中氨基酸的数量相当，但这些人造的部分具有某些生物学中没有的特性。

库存的急剧减少有一系列深远的影响。首先，零件的装配者将受到影响，微型机器人相当于一个孩子正在放置的一块乐高积木。这个装配者最终是一个真正的个人制造者，用一个单一的集成过程取代了微观装配“数制”工坊中的一系列机器。其次是对供应链的影响。装配工在需要几十种类型的零件的原料，如今的微观装配“数制”工坊要求许多供应商提供它所需要的所有类型的消耗品。第三个影响是可逆性结果。垃圾的存在反映了它所包含的信息的缺乏。离散拆卸与离散组装对称，数字材料部件可以重复使用多个周期，直到错误率高达需要重新制造的水平。这种循环利用周期标志着垃圾填埋场的做法已经过时，现在，任何废弃物都将成为新建筑的部分“建筑材料”。组装和拆卸的组合增加了可重构性。你可以不断地修改产品，以反映你不断变化的兴趣和需求，而不是做出保留某些事物或处理某些事物的决策。

所有这些属性在长供应链的末端都尤为重要。其中最长的供应链是将物品送到外层空间：围绕地球运行的卫星运输成本高昂，其价值大致相当于黄金。因此在任务结束时，由于没有办法重复使用它们的

部件，卫星只能在大气层中燃烧或变成太空垃圾。制造可拆卸和可重新配置的航天器逐渐成为数字材料应用的早期驱动力之一。

原位资源利用这一术语的含义是航天器在没有行李的情况下进入太空，就地取材而非从地球上运送材料。关于这一目标的工作通常有一个隐含的假设，即最终目标是通过工业革命的各个阶段，使技术在地球上得以复制。关于如何做到这一点，有这么一个例子——由戴维·金格里撰写的一系列极为精彩的丛书。第一本书是关于如何制作一个木炭炉，接下来，这些书的内容包含通过手工工具发展出一个完整的机器工厂。

2016年，我与20世纪福克斯公司和美国国家航空航天局合作，为电影《火星救援》（*The Martian*）家庭版的发行举办了一次探索“搬到火星”背后的科学的活动。数字材料的发展为实现技术文明这一深刻问题提供了一个与众不同的答案。你需要了解几十种材料的特性（导电、绝缘、磁性等），将这些材料组成离散的模块，然后组成一个社会，而不是再次以获得50万种电阻器、电容器和连接器的库存为结束。

无所不在：从1到1 000 000 000 000

路线图的最后阶段是从每个人的数字制造发展到每件事物的数字制造。这一阶段对应的是互联网是如何从计算机传播到人再到事物的。它将通过对机器和材料的结构进行数字化来完成。得益于材料本身的可编程性，机器与其制造的东西之间的区别消失了。

如果想了解这个数字制造阶段，考虑装配过程的速度，我们可以通过一个简单的计算查找现存的严重问题。目前最快的同类操作是喷墨打印机，其产量为每秒10 000滴。如果以这种可观的速度（即每秒

10 000立方毫米）创建一个1立方米的物体，那么这个过程大约需要一天时间。这相当于一个3D打印机制造一个大而精致的产品所需要的时间。但是，如果零件尺寸缩小到0.1毫米，制造1立方米物体所需的时间则要增加到3年；如果零件的尺寸再缩小到原来的1/100（1微米，即肉眼可见和手指可感知的极限），零件制造时间增加到300万年，这就需要我们等待很久，有生之年无法完成工作。

这个过程可以通过一次放置多个工作部件来提速，比如我们可以安装多个打印机喷嘴以增加喷墨打印机的打印速度。同时使用更多的零件也许可以将速度提升10倍或100倍，但同时放置上百万个零件，这也是不切实际的，因为零件之间对相对偏差具有敏感性。生物领域对于这个问题的解决方案是“复制”。制造蛋白质的核糖体也制造核糖体。组成细胞的核糖体的数量因细胞构造的不同而差异很大，但每个细胞中的核糖体数量可达数百万。这么多的分配工同时工作，我们使用这些微米级部件去填满1立方米空间的时间将低至30年——虽然仍要花费很长时间。核糖体制造蛋白质，蛋白质构成细胞，细胞再制造更多的细胞。人体大约有10^{13}个细胞，所有细胞中的核糖体一起工作可以将组装细胞的时间缩减至0.1毫秒。然而，核糖体自我复制的速度要慢得多，它们以每秒一个单位（氨基酸）的速度运行，整体复制的时间长达一秒钟。此外，并非所有的细胞都含有那么多核糖体，如果每个细胞有1 000个核糖体，那么要放置10亿个部件的时间大约是一分钟。

我这里要讲述的并不局限于这些微小的“拼装工人”如何工作，或者一次能放置多少个部件，更多的是要强调这些“拼装工人”能利用部件来复制自己，使它们的能力呈指数级增加。这种递归对于在同一时间里从最小的部件构建到最大的系统是至关重要的。

对于像T－1000液态机器人那样可以依据指令改变形态的材料，专业术语为“可编程物质”。对于像液态机器人的液滴那样可以自我组合的材料，专业术语为“可自我组装的物质”。不幸的是，上述的

“可编程物质”和“可自我组装的物质”都存在于想当然的状态，也就是说，它们目前只存在于研究项目中而在现实世界中不存在。迄今为止的实际成果远未达到愿景，要么是数量较少、复杂程度高但功能强大的机器人模块，要么是大量的无用途的简单组件。

对这两个属性的研究旨在模仿生物生长、进化和自我修复的方式。但是生物的本质核心并不是能自我组装的可编程物质，而是一种具有明确分工的系统。通过信使RNA分子，指令被传送到核糖体，转运RNA分子也运来了氨基酸，核糖体又遵循一个精心编排的编码序列来生成蛋白质。在一个分子伴侣的引导下，将蛋白质折叠成具有一定功能的形式。单个氨基酸自身不具有可编程性，可编程性来自核糖体为了制造核糖体而组装氨基酸的过程。

如果生命体已经可以做到这一切，那我们为什么不能利用它们来生产一切材料呢？虽然对生物体的编程的可行性越来越大，但我们依赖的许多材料还是达不到这一标准。分子生物学还不能合成优良的电导体以长距离输送高压电流和高速的信号，也不能合成抗高温部件以承受喷气式发动机的荷载和高温。

然而，使用非生物材料组装装配机也可以采用同样的引导方式。绝缘的、弯曲的材料可以制作机器的结构，导电材料及电阻可以用作线路及组件，半导体材料可以用来对组装器编程，磁性材料可以让舵机运转以移动装配机。一旦一个装配机被制造出来，它就可以通过自我复制来制造更多的装配机或任何其他部件。复制器就是一个可以利用组合成它自己的部件来组装成新个体的装配机。

这是冯·诺依曼将自我复制机器作为生命模型的科学愿景的实现。进化过程从分子开始，发展到细胞、器官、生物体乃至物种的进化，可以说，现在正在进化的是人类的文明和理念。数字制造路线图的最后一部分，将给进化的比特排列原子的过程完美地画上句号，反之亦然。

（本章由尼尔撰写）

[1] 1英寸≈0.025 4米。——编者注

[2] Formlabs公司为专业用户提供一整套桌面光固化3D打印系统，让3D打印更直观可靠，产品精度极高。——译者注

[3] 弗兰克·盖里生于加拿大多伦多的一个犹太人家庭，17岁后移民美国加利福尼亚州，成为当代著名的解构主义建筑师，以设计具有奇特的不规则曲线造型和雕塑般外观的建筑而著称。——译者注

[4] 得捷电子是世界著名的电子元件目录分销商，于2013年在中国上海成立其全资子公司——得捷电子（上海）有限公司，正式设立上海客服中心，服务中国客户，并接受人民币付款。——译者注

第6章

未来的现实

尼尔为我们勾画出一张技术路线图，使我们得以一窥未来半个世纪的数字制造技术发展前景。1965年的戈登·摩尔不是算命先生，如今的尼尔也不是，他只是提前洞悉了过去10年技术的加速发展趋势并描绘了未来的发展轨迹——决定这一轨迹走向的，是数字技术的基本原理、动态研究路线图以及不断发展的数制生态系统。

有了这张路线图，就等于身处于假想中那个1965年的咖啡店里，学习当时戈登·摩尔的观点，并预测未来数字电脑的发展。与摩尔定律一样，如果拉斯定律是正确的，那么在未来的55年里，数字制造的能力会是现在的10亿倍。这样的设想在很多方面给我们提供了非常真切的可能性，它能使生产民主化，把我们不论是制造的、销毁的还是改制的物品进行转化，能让许多人制造他们现实中的消费物品——创造一个自给自足、相互联系的可持续发展社会。这就是一次机遇！

然而，要实现这样的一次机遇又是另一码事。拉斯定律并没有被篆刻在石碑上。数字制造的发展速度以及不断涌现的新技术对社会发展的影响不是由一些无形的手控制的。拉斯定律的成长取决于个人和组织机构对提升数字技术的效能及扩大数字技术的运用范围的投入和决心，也离不开为社会创造价值并减少损害的目标。

现在，我们已经观察到“数制”工坊改变生活的早期迹象，但是没人能保证数字制造技术的积极潜能是否可以完全实现。社会体系过去的记录很清晰——人们并没有直接接受新的技术，因为人们坚信数字制造技术带来的积极影响完全可以实现。人类和组织机构的行为必须改变。原先的体系也需要改变，新的生态系统必须发展起来。可以引领社会发展并巩固体制建设的那些深深嵌在我们脑中的规则和假设需要被重新塑造。

我们为编写本书采访了一些人，尽管他们对技术路线图的方方面面多少有些不同看法（比如个人制造普及到家庭的时间和方式），但他们对技术变革潜力的认识却有广泛的一致性。路线图未必按照尼尔

的设计思路如期展开，但大家一致认为，未来几十年数字制造能力和覆盖范围将提高很多个数量级，并对社会产生深远影响。

我们写最终章的目的是提供切实的指导来确保第三次数字革命广泛惠及个人和社区。在第2章中，我们研究了数字制造一开始经受的挑战——使用权、能够使用的生态系统和缓解风险办法。在第4章中，我们把这些挑战放置于社会科学以及历史背景之下，从而关注数字制造的变化速度以及数字生态系统如何传播。现在我们把关注点转向了如何运用这些见解来应对这些挑战，以共同发展社会和技术系统。

在我们整个采访过程中，人们对关于未来的叙述十分渴望——这种叙述既不是乌托邦式的也不是反乌托邦式的，而是充满希望且能实现的。正如阿尔文·托夫勒在《未来的冲击》中所强调的那样，“每个社会所面临的未来，不仅仅有一系列的确定性，还有相当多的不确定性，对未来的发展方向的判断也会有分歧。变革管理就是要努力将不确定性转变为确定性，再转变为社会共同认可的发展方向。”

为了帮助人类实现美好未来，我们从本章开始，提出了8个理想的场景，在这些场景中，数字制造是积极推动社会转型的关键因素。这些场景由来自五大洲的“数制”工坊先驱共同创建。这些场景解决了重要的社会问题，包括经济可持续性、文学素养和学习能力、文化转型、基础设施建设等。对于数字制造而言，可能还没有一个真正的一体化的理念，但这些场景的结合已经开始描绘未来的蓝图。

为了构建稳定的程序以使这些设想和其他期望在未来更快地变成现实，我们引入预测转换模型。这种模型是一个帮助改进社会体系和技术系统的骨架，通过数字制造共同演进来实现集体目标。我们最终得出结论：通过将技术人性化，每个人都将能在第三次数字革命中生存并且发展。

人与技术交汇的乌托邦

想到“数制”工坊这个创意并不困难，甚至对于那些从来没有接触过这种工坊的人来说，也不难。但对于大多数人来说，很难想象个人制造的环境就像自己家里一样。这是否会是一间厨房，一个拥有许多机器模块和材料储藏区的精致房间？是否在一个特定的办公室或是小企业里有一个已编辑好的版本？这种个人制造的性能如何才能运用于那些人数占据全球人口70%的乡村里？当讨论到数字材料和可编程的事物时，“设想一个人如何与技术交互”这一问题变得越来越困难。

对普通人而言，要理解尼尔路线图的4个阶段，展望他们如何与新兴的“数制”技术进行互动，难度是逐渐增大的。第一阶段相对容易，大多数人可以想象在“数制”工坊利用数字界面操控激光切割机和3D打印机等设备的情景。下一阶段则更具挑战性，尼尔认为，个人制造是由“机器制造机器”驱动的。如果机器制造了机器，那人们还能做些什么呢？事实上，这一阶段被称为“发烧友”阶段更恰当，因为这一阶段涉及精通技术的先行者如何利用“数制”工坊中不断改进的组件并配置它们，以满足个人需求。然而在这一阶段中也存在一些问题，例如：个人制造的设备应该摆放在家里还是其他位置。就日常行为而言，个人制造和个人计算还是有所不同的。

随着数字材料的引进和传播，在谈论人们会如何制造、销毁再重新制造物品的过程中，的确存在着一种典范的转移。这些新兴的技术如何影响我们的生活、学习、工作和娱乐？即使是那些正在开发技术的人也难以想象。数字材料会是什么样子的呢？它们会被存放在哪里呢？一个人需要多少材料才能制造一把椅子、一架无人机、一辆车或一栋房子？拆卸并重组这些物件是否会变得更简单？在许多方面，理解路线图的根本技术原理会比在脑中设想这些技术如何影响日常生活更加容易。

这就是以科学为基础的故事讲述者和艺术家能发挥关键作用的原因。科学家和未来主义者可以观察、研究路线图，然后对未来做出准确的预测。故事讲述者同样可以研究路线图并为未来的发展方式提出令人信服的愿景。挑战我们现在生活的世界可以使我们希望看到的世界出现，通过这一事实，我们可以激励处于分布式生态系统中的人将理想中的未来变为现实。为技术路线图的未来的功能提供社会背景和心理地图，有助于影响和推动技术的发展。

同时，我们在本章中呈现的理想并不是一种乌托邦式的设想。许多技术先驱者透过“玫瑰色眼镜”来展示他们的技术。然而在《硅谷》（*Silicon Valley*）这档电视节目中，有人讽刺这一趋势，认为“首席执行官们自豪地声称他们正在通过帕克索斯算法来达成协议，进而使世界变得越来越美好”这一观点是虚构的，这只不过是一些荒诞的术语而已。技术的乌托邦能够欺骗人们，让他们拥有一种虚假的安全感，认为任何技术问题或社会问题都能通过技术解决。基于科学路线图的憧憬很美好，但可以实现。

我们必须小心谨慎，避免落入“未来的技术令人恐惧”的反乌托邦思想中。现在流行的电影、电子游戏、小说和漫画书中都有利用技术胡作非为的恐怖设想。当提到人工智能、生物工程或汇编程序时，我们很难跳脱这种令人恐惧的想法：人工智能会威胁人类，或者某个无赖国家（或大学联谊会）会运用个人生物制造技术改变人类基因库。当然，加快推进技术的确有很大风险，也许会遭到彻底的反对或导致实验室解散，但是不断加深反乌托邦设想会让人们感到更加恐惧和无助。

如果我们想要激励一代人来推动第三次数字革命，那么就需要为人们提供积极参与的动力。事实上，未来主义者阿尔文·托夫勒在1970年也指出，我们必须大力拓展我们对可能的未来的看法。对于严谨的科学，我们必须加入火热的艺术想象力。

在这样一个技术日益复杂和加速发展的世界，培养一群知识渊博、参与性强并且足够热情的年轻人是21世纪最困难的挑战之一。正如天体物理学家、科学福音传播者卡尔·萨根所说：“我们生活在一个依赖于科技的社会之中，但是几乎没有人真正了解科技。”为了解决这一矛盾，我们不仅要与走在前沿的科学家和社会科学家接触，还要与人文科学中的故事讲述者和学者接触。我们希望，在本章中提到的8个理想的场景能够引起人们关注并且令人信服。这些设想的场景都基于真实人物的当代作品，热情的“数制”工坊开拓者做着令人惊奇的工作，并为数字制造如何改善人们的生活提供了的愿景。在设计这些场景的时候，我们受到了设计师亚历克斯·麦克道尔的启发。亚历克斯·麦克道尔和格雷戈·林恩在好莱坞电影《少数派报告》（*Minority Report*）中展示了制造技术的未来前景。随后，访问麻省理工学院媒体实验室的艺术家也开始探索制造技术。

麦克道尔已经为《搏击俱乐部》《查理和巧克力工厂》《超人：钢铁之躯》等电影创造了令人回味的世界。然而，麦克道尔最有名作品的可能是他在2002年策划的、由史蒂芬·斯皮尔伯格执导的《少数派报告》。斯皮尔伯格将这部电影的背景设定在2054年的华盛顿，他希望这部电影中基于科学的虚构世界能够成为“未来的现实”。为了拍好这部电影，他召集了走在前沿的科学家、未来学家、设计师和故事讲述者，组成智囊团，这其中也包括亚历克斯·麦克道尔和尼尔。这部电影向观众展现了新兴的技术——从无人驾驶汽车到手势界面再到与上下文敏感的广告。微软、惠普以及其他新兴企业都对这部电影中振奋人心的突破性成果表达了赞赏。

参与《少数派报告》制作的经验使麦克道尔受到启发，他在南加利福尼亚大学创立了世界建筑学会以及设计工作室。他们将严谨的科学研究、广博的专业见解与扣人心弦的镜头、令人信服的故事结合起来，给主流企业和开拓性社会企业家展示了一个新世界。麦克道尔在《4》（*Four*）的序言中描述了这一愿景。《4》是由欧特克软件公司（Autodesk）出版的一系列短篇故事，旨在探索设计和技术的未来：

“我们有能力创造我们想要的未来。这意味着我们不仅要摆脱目前的局限性，还要通过想象力和独创性引领每一步。我们的未来是共享的，讲故事可以使我们共享这一愿景。”

电子线传媒与麦克道尔的设计工作室合作，共同推进艾伦涉足的一些项目，这些项目以科学为依据，信息量很大，要解决的都是世界性问题。项目之初，就要在各种假设与反诘之间努力开拓创造性思维，使可能性变得“可能”，最终得以实现。

实现理想愿景的一个很好的例子就是“制造型城市”，它鼓舞全世界的人都参与进来。愿景是这样的：“2054年，如果我们的城市与全球接轨，同时实现本地生产，将会怎样？”在“制造型城市”的白皮书中，托马斯·迪兹对未来理想的“制造型城市”做出了解释（写在序言中）：“我们需要通过产品的本土化，再次发展城市与人、城市与自然的关系，这样这些城市会提高生产力，更有助于恢复和发展，人们不会再加倍索取、破坏资源、损害环境。这些城市变得越来越繁荣，人们的工作更加有意义。人们有目标并不断发挥自己的天赋，这样一来，工作便成为一种享受。我们还需要通过和市民沟通并且运用先进技术来丰富自己，提高城市构建方面的能力。”

巴塞罗那的制造先驱发起了一次运动，展现出一幅当地生产力发展强大、与国际接轨、自给自足的城市愿景。正如前几章中所提到的那样，在巴塞罗那市长和首席建筑师的见证之下，来自世界各地的许多城市和国家都在第十届年度“数制”工坊会议上签署了协议。

在以下8个设想的未来智造场景中，我们都将分析出不止一个挑战，并将介绍一位正致力于解决这些问题的人士。然后，我们再把目光往后投射几十年，想象在某个场景中，这位人士的努力成果不断扩大，反映一种新的可能出现的现实。与其说我们在做预测，倒不如说我们在尝试着画出一幅探索未来可能性的思维导图。麦克道尔把这些设想称为“诱发”，因为这些设想能够激发思考和行动。希望能够让

讲故事的大师通过各种媒体手段将这些未来愿景融入他们的艺术形式。

前两个愿景建立在“制造型城市”的基础上，着眼于不同区域（乡村和市中心的分布式网络）的自给自足。贯穿两种场景的“自给自足”主题暗示了未来可能的社会模式和经济模式是建立在能力和合作上的，而不是建立在恐惧和孤立上的。



图6－1 “科学学堂”。维什沃兹·信德摄

马哈拉施特拉邦奇迹

大多数“数制”工坊建造在大型城镇里，很少建在偏远乡村。乡村工坊面临着一系列独有的挑战——从可靠的网络访问和电力供应到区域特异性问题，如泥路上的尘土飞进机器（导致机器故障并产生了

维修成本）以及过低的文化水平等。因此，如果我们真的想为全球制造发展做贡献，就必须先解决这些特殊问题。

管理“科学学堂”（和麻省理工学院合作的第一家“数制”工坊的原型）的约各什·库尔卡尼正致力于解决这些问题。“科学学堂”坐落在印度马哈拉施特拉邦的一个乡村地区，这个工坊的工作人员用基于圣雄甘地教育哲理的学术模型来教育当地的年轻人。甘地相信，“工艺是教育的媒介”，同时，他通过运用泰·塔利姆教育法[1]（Nai Talim）来促进农业、纺织和木艺等学科的学习。“科学学堂”成功地运用基于当地文化的现代数字制造工具复兴了这一传统。许多库尔卡尼的学生也开始探索新的解决方案来应对当地的农业挑战或者创办新的当地企业，甚至建立新的区域型“数制”工坊。库尔卡尼希望展示这一模型，并最终将其运用于所有农村社区中。以下是库尔卡尼的想法：



图6－2 在“科学学堂”制造的鸡胚孵育箱——制造了几乎一切事物，甚至是小鸡！维什沃兹·信德摄

2038年，也就是库尔卡尼和马哈拉施特拉邦的领导者与合作的智造先驱们发布“智造村”宣言20年之后，他们为该地区所有43 655个村庄提供基本智造设施的努力已卓有成效。就像“智造城市”宣言一样，他们根据本地区情况对不同规模的数字化制造能力进行分步推进，并针对当地独特的农村生态系统对这些能力进行了优化。

这一模型始于2018年，这时库尔卡尼和他的团队已经与100个运行IBT项目（基础技术的引进）的学校合作了。因为这些学校已经有了基础制造能力和制造思维，“IBT学校”是摩擦较小的地方。合作伙伴和当地社区一起建立跨校区的合作平台，并为实验室提供社区，将私人资金与公共资金结合，进行最初的推广。

与此同时，库尔卡尼和他的团队开始培训一批“教育企业家”（包括IBT项目的毕业生），使之成为社区中的本地工厂服务提供商。这些人通过新的以“数制”工坊为重点的小额贷款计划获得资金支持。这一小额贷款可以用来购买制造机器，如激光切割机、3D打印机或3D扫描仪。这些贷款的发放是为了确保每个村庄都具有增材制造或减材制造的能力——创建分布式社区“数制”工坊。这些企业家能够通过接受客户定制设计和租赁机器得到的收益来偿还贷款。这些企业创造了一代制造实践者和传播者，从而把数字制造带到了他们的社区。

这些“教育企业家”还通过移动设备和国内外知识网络进行相互交流。他们分享了自己对当地相关工程的见解，包括滴灌、环境卫生，以及食品保存。他们也利用不断发展的材料知识网络，从其他工坊创新者那里获得使用马哈拉施特拉邦5个农业气候区的当地原材料和天然聚合物的经验，以降低制造过程中的成本。随着时间的流逝，一些人已经可以购置更多的机器，并建立更大的、性能更强的社区型工坊。

随着“IBT学校”数量的增加，成千上万名青年从该项目中脱颖而出，他们的能力得到了增强，为个人和职业的自我可持续发展做好了准备，政府、产业和慈善事业都注意到了这一点。印度法律规定，企业必须将利润的2%用于社会责任，这一举措不仅培养了许多合格的员工、创造了一种创新文化，还产生了积极的社会影响，因而吸引了更多的企业资金。这一新型资金流能够使所有更大规模的村庄建立比当地实验室水平更高的社区型“数制”工坊。

最初，印度进口的制造硬件是由西方公司生产的。不久后，工厂型工坊在马哈拉施特拉邦的个别城市中出现了，硬件和软件通过全球通用的开发源代码设计在当地不断发展。这种本地化发展极大地降低了供应链的低效性以及总体成本，从而使所有制造节点在区域生态系统中呈指数级增长。印度政府注意到了区域网络可以催生社会、经济和文化利益，与此同时，印度政府将工坊网络称为“马哈拉施特拉邦奇迹”，并向社区投入更多资源，以确保43 655个村庄都可以进行基本的数字制造。

随着“马哈拉施特拉邦奇迹”制造网络这一消息传遍全球，其他乡村地区也开始执行和适应这一模式。库尔卡尼意识到，“科学学堂”的创始者S. S.卡尔波建立“数制”工坊的目标是“见证印度繁荣昌盛并且成为世界其他地区的探路人”。卡尔波认为，只有在每个人都能将自身潜能发挥到极致的时候这种事情才会发生，因此，再低的水平也能得到提高。”

自给自足：底特律灯塔

底特律作为后工业化衰败的典型代表，几十年来一直面临着投资减少和经济衰退的问题。在布莱尔·埃文斯的作品中，成千上万家俱乐部、一个重要的艺术场景和一系列创新的复兴项目都显得不那么显眼。



图6－3 布莱尔·埃文斯。扎克·罗森：新产品真的能使用吗？（sharable.net），2014

埃文斯是麻省理工学院电机工程专业的一名研究生、资深企业家、一些特许学校的前任负责人以及底特律东部的30英亩土地的所有者。这片土地是一个大胆的城市改造实验的起点。埃文斯和他的同事们以在社区开展活动并支持K－12学校和成人生活探险家的Incite Focus工作室为中心，大胆探索新型的模式，培养新一代优秀的人类，让每个人减少工作量，不断创新，增加互动。埃文斯指出，在城市内

部建立自给自足的生活方式不是一个新想法，但是当谈及“数制”工坊可以加快从食物到家具的生产甚至是个人发展历程时，这一想法却能产生成倍的影响。让我们以埃文斯的角度预测一下未来。

假设现在是2037年，全球技术类职业的失业率不断增长，造成了巨大的混乱与社会动荡。但是现在，底特律已经成为“后工资”未来的灯塔，也就是说，底特律人可以自给自足，通过自给自足来实现他们的消费——自我决定，从而找到生存的意义、目的和尊严。在这一模型中，人们每周花大约30个小时单独工作，然后直接与社区合作社合作生产家庭和社区需要的东西，并制作额外的设计和物品用于销售和交换，以满足家庭和社区的需要。剩余的15~20个小时（常规的工作时间加上其他空余时间）用来满足个人爱好，进行自我激励，从而不断地自我成长。现在，这一模型逐渐成为底特律东部以及全球各个城市中心的发展趋势。

底特律也已成为大规模永续栽培的先驱，拥有丰富的作物，如西红柿、青椒、大头菜、西葫芦、辣椒、黄豆、豌豆、洋葱和土豆。而且，底特律也有足够的开放空间来培养植物和有益微生物，这些生物可以将太阳光、二氧化碳和氮气转化成可再生燃料，使其成为一种能源。这种联合种植也提供一些“生态系统服务”，如土壤修复、土壤富集、空气质量改善、水质改善以及水分保持，等等。与此同时，政府赞助的学徒项目给传统贸易提供了一种新途径。不断增强的自给自足能力和稳定的工资水平给社区带来了安定，例如，通过制造房屋技术可以修复原先废弃的房屋。虽然这一技术源自国内外的设计，但建造过程都在本地进行。



图6－4 在底特律建造永续栽培装置。马修·派珀，“绿色城市日记：‘数制’工坊和自然语言”，Model D[2]，2013

革新意味着对经济模式进行深入反思。人们用各种语言来重复着布莱尔的格言——批发销售和零售购买是一种失败的策略。大多数人已经为他人卖力很久了，但实际上，他们是在将自己的劳动力以批发价销售出去，而其他人会在销售产品前将产品涨价。然后，一些人为了使经济好转而购买零售商品。对于食物、家具、房屋和其他需求来说，数字制造可以实现批发买卖。随着区块链技术（能够记录并验证人与人之间交易的一种不断复杂的加密技术）的不断创新，比起买卖，有更多的人会选择易货和交换。

如埃文斯所说：“假如充实的生活指的是变得快乐、自给自足以及变成社区中的一分子，我们是能够做到这一点的。”这种关于生活意义的哲学观点在世界各地的文化中都具有吸引力，都指向一种减少排放和退化而更加享受和繁荣的生活愿景。

接下来的两个未来设想以学习和辅导为主，二者都是培养制造能力的必要条件。第一个设想基于个人学习之旅，这个人在两年内参观了至少24个“数制”工坊。学习也是以实际操作为基础的项目学习方

案的核心，该方案始于埃及，并在全球范围内扩展。两个场景都代表了将素养由限速器变为加速器的方法。



图6－5 延斯·戴维克向肯尼亚总统的祖母赠送激光切割拖鞋，上面印有美国前总统奥巴马的图像，拖鞋由日本的皮革工匠设计，然后在肯尼亚下载和定制。延斯·戴维克提供

延斯之旅

事实上，在每一位智造先驱的采访和调查参与方的反馈中都存在着一个挑战——整个“数制”工坊生态系统需要经验丰富的导师。智造先驱希望能够指导那些新接触生态系统的学徒，但是没有充足的时间来满足日益增长的需求。先前提到的延斯·戴维克，在奥斯陆经营了一家“数制”工坊，并且展示了能够促进个人发展的新型辅导模式。

2011年，戴维克从欧洲一所著名的设计学院毕业后，踏上了为期两年的全球“数制”工坊之旅。他此行的目的就是发现新的共享设计商业模式。其中有将近30个“数制”工坊，他在每个“数制”工坊都花了大约两到三个星期进行研究。整个旅途耗费了约35 000美元。他在旅行途中看到了200多家“数制”工坊。出乎意料的是，这次旅途改变了他的生活。在旅途结束时，他将共享概念运用到设计实践中，公开他的设计，和全球新一代设计合伙人一起工作。此外，他成为最受欢迎的“数制”工坊导师之一。这段旅程将他从学生变成了老师，并且给我们带来了未来愿景——我们称之为“延斯之旅”。

假设现在是2035年，全球有成千上万家社区“数制”工坊，其中有数以千计的员工——从应届毕业生到退休人员，到数量不断增加的技术型失业者，甚至某个家庭——来自“延斯之旅”。就像戴维克最初的旅程一样，他们的旅行包括访问各个大陆上的多个工坊。在为期两周的访问中，这些旅行者都做着与志愿者相同的工作——分享他们知道的东西，从其他人那里学习知识。不同之处是，在任何一周，每个工坊里都可能会有两三位游客。当人们与家人一起旅行时，这种情况变得越来越普遍，学习范围从

工坊扩展到寄宿家庭。当然，有时候，对于年轻的旅行者来说，友谊和爱情会导致逗留时间超过两周。

“延斯之旅”能够成功的原因有三个。首先，在2020年，一个开创性的挪威基金会开始颁发奖学金，以帮助人们进行这样的旅行。刚开始，只是为他们参观指定的工坊提供资金——那些工坊已经做好了迎接参观者的准备。要使参观者的来访不影响工坊的正常运行，就需要给予工坊和参观者一些经济支持和后勤保障。在“延斯之旅”开始实行的最初阶段，受访工坊对指导的需求很大，因此优先给予工坊支持。随着旅游者开始旅行，工坊中的人都有了指导能力，不再需要他人辅导，对工坊的支持也就随之降低。在“延斯之旅”的后期，每一位游客都能受到任何一个工坊的欢迎。

同时，如果学生能提供一份数字档案，并且他的学术成就符合标准，那么全球越来越多的大学将为旅途期为一年或两年的学生提供贷款。到了2025年，资金来源大体上会扩展到许多私人基金会、公共机关甚至企业，这是提高品牌知名度并与有前途的人建立联系的一种方式。现在失业机构允许员工在全球任何地方的工坊登记，而不需要每两周都到办公室报到一次。现在，对于全球一体化、跨文化欣赏、技术发展和个人更新来说，“延斯之旅”已经成为一个被广泛重视的驱动力。

融合

2011年，迪娜·埃尔－赞法利作为麻省理工学院的研究生，把尼尔的“如何智造（几乎）万物”带入课堂。她来自开罗，在那里，她是一名建筑师。埃及的建筑课程十分紧凑，她根本没有业余时间去学习计算机编程，更别说参加尼尔的课程培训项目了。



图6－6 迪娜·埃尔-赞法利。布莱斯·维克马克摄，来自麻省理工学院新闻

2014年，穆罕默德·希沙姆作为埃尔－赞法利同乡的IT企业家，给尼尔发了一封邮件，希望能建立“数制”工坊。尼尔希望他能和埃尔－赞法利合作，因为埃尔－赞法利已经开始计划建立实验室了，这样他们就能够合力共建埃及“数制”工坊。刚开始，正如埃尔－赞法利所说，实验室吸引了一批“宅在家中的书呆子和怪人”。他们很快就学会如何使用数字制造设备，但是埃尔－赞法利发现他们没办法在短期内想到制造方面的新点子。她准备探索一些新的方法来使初学者成为有创新能力的设计者，所以这一目标是她学术演讲的重点。埃尔－赞法利运用I^3的创新方式（以模仿为开端，然后迭代，最后是即兴创作），为人们展示一个成品，并且让他们思考如何制作仿制品，同时弄清楚需要哪种软件和机器。然后，她让人们在项目迭代中进行一个或多个引导性改变。随后，人们开始即兴发挥。在埃尔－赞法利看来，她希望在创作过程中能够将人、软件和硬件有效地结合在一起。现在设想这种愿景在埃及甚至在全球的学校中逐一实现：

现在是2030年，几百个“数制”工坊在埃及建立，其中一些与埃及新建的大学和STEM高中合作，其他的则坐落在大大小小的

社区里，还有越来越多的“数制”工坊建在K－12学校中。最初，学校反对“数制”工坊这一概念，但是当大多数STEM先锋学校逐渐得到学生们良好的反馈时，其他学校也开始接受了“数制”工坊这一概念。“数制”工坊不仅可以把学生培养成富有创新力和学习热情的人，还能明显提高学生们对核心学科的学习能力。“数制”工坊为该地区的青少年提供了富有成效的途径，因此得到了各级政府领导的重视。仅在10年之内，在学校里建立“数制”工坊就像在学校里建立图书馆和体育馆一样常见，“数制”工坊是全面培养学生学习兴趣的一部分。老师们也接受了他们逐渐从讲师到导师、教练和合作者的角色转变。学校也逐渐从举行“一考定生死”的测试转变为提供创新机会，以此来评估并支持学生的学习过程。

埃及有越来越多的学校采纳I^3模型，它们能够与其他学校分享工坊的课程、专业发展前景和操作见解，分享知识和战略，克服障碍，进而过渡到新的基于项目的实践教学方法。同时，当地的每个社区都对工具进行了调整，以实现很好的实践效果。使用英语编写的软件和文档，包括未能翻译出来的技术词汇，都可以被翻译成当地语言，进一步减小误差，以便有效使用。

为了确保全球学校都能参与其中，埃及学校开发了一个创新项目：高中生给初中生制造数字制造机器，初中生给小学生制造机器。年长的孩子为年幼的孩子在硬件、软件以及I^3模型的应用方面提供指导，同时分享他们对于SETC（波士顿南端技术中心“数制”工坊）“learn（学）2 teach（教），teach 2 learn”模式的见解。对于那些面临高失业率状况的青年来说，制作和指导成为一个新的关注点。

对于大学生和研究生来说，埃及的研究人员和来自全世界的志趣相投的人一起工作，开发能与制造者（在“数制”工坊工作的人）合作互动的硬件与软件（实时利用更广泛的开放式设计网络），有利于运营商在探索过程中提出各种想法和方法。软件与

硬件的设计将模仿、循环和创作的过程考虑进去，不断更新埃尔－赞法利的想法，使其与当地社区保持一致。

在研究中收到的所有反馈几乎都强调了社区的重要性以及与“数制”工坊的合作。以下两个设想着眼于两个截然不同的环境中的社区和协作，一个跨时代，另一个跨行业。在第一个设想中，“数制”工坊被看作保存并传播古代文化传统的潜在的基本要素。在第二个设想中，“数制”工坊和跨国公司合作，通过学习和合作来融合社区与产业。

解锁库克湾部落

个人智造有着悠久的文明历史。创造和使用用于自我可持续性发展与个人表达的工具可以追溯到200多万年前。今天，地球上一些最古老的文化正在与现代社会对古老传统和文化遗产的侵蚀的现象做斗争。“数制”工坊可以成为另一种更加现代化的入侵方式，也可以建立一种先进科技与古文化和谐共处的氛围。

库克湾部落理事会（CITC）是位于安克雷奇的一个阿拉斯加原住民社会服务组织。其核心使命是为青少年打造一个未来，使他们能在复杂多变的世界上有能力、有自信、有勇气地为了他们的目标而奋斗。在库克湾部落理事会的倡议下，为了和全球观众分享、弘扬、传播部落文化，理事会和艾伦的公司共同开发了一款商业电子游戏。这款电子游戏一经销售就创下巨额利润，进而支持了库克湾部落理事会的社会项目。例如，《永不孤单》（英文名：*Never Alone*；因纽皮特语名：*Kisima Innitchuna*）这款电子游戏是由世界级的游戏制作人和超过30位阿拉斯加原住民长老、作家和故事讲述者共同开发的。这款游戏使用因纽皮特语，以一个世代相传的故事为基础，包含了26个关于文化的纪录片片段，可以通过闯关进行解锁。

这一款电子游戏火爆全球，广泛覆盖媒体，已经被300多万游戏玩家下载，并且赢得了许多奖项（包括BAFTA——英国电影和电视艺术学院奖），甚至间接地创建了库克湾部落理事会（CITC）“数制”工坊。通过电子线传媒的伙伴关系，部落理事会遇到了尼尔，从而成功地创立了安克雷奇“数制”工坊。这个工坊是由库克湾部落理事会青年教育与就业服务的主管——勒妮·弗雷德里克斯与凯蒂·李共同运营的。该工坊的使命是将传统的阿拉斯加原住民文化的价值和优势与前沿的教学工具相结合。



图6-7 来自库克湾的吉拉·摩尔和伊丽莎白·佩顿。库克湾部落理事会提供

提前想象一下2030年。库克湾部落理事会“数制”工坊已经形成了一个为整个阿拉斯加原住民社区提供支持的分布式工坊网络，成为全球的先驱，帮助青少年将传统文化和一些世界上最先进的技术融合起来。创新在许多领域中都正在发生。在城市的工坊中，老人和成长在工坊中的青少年一起工作，指导下一代人建立技术与文化的桥梁。在偏远的工坊中，社区指导人们如何运用当地材料来克服当地挑战，从而减少对低效且昂贵的供应链的依

赖。而乡村的工坊与遍布北极和世界各地的偏远文化合作，共同完成项目，分享、讨论共同问题的解决方法。事实上，由于可持续的生活正在变得越来越分散，现在出现了从城市向农村的反向移民的现象。



图6－8 来自库克湾的凯肖恩·奇卡里森。库克湾部落理事会提供

2016年，弗雷德里克斯在白宫被授予“改造冠军”之名，然而这只是个开始。让我们看看库克湾部落理事会的未来设想。

“数制”工坊的设计与运营也完美地与古代文化传统相融合。例如，运用激光切割机来雕刻也被认为是雕刻方面从运用石头到使用金属美工刀再到使用电钻的发展过程的一部分。第一次学可用激光切割机设计和雕刻的孩子们同样也对掌握传统工具的使用方法感兴趣。这一点儿也不意外，因为数制“工坊”的设计同时包括现代化产品服务和传统工艺品实践的工作空间。从偏远地区到热闹的城市中心，世界各地的其他文化都在分享着类似的做法。

除了欣赏和实践古老手工艺，“数制”工坊还是传播传统故事和文化的地方。从《永不孤单》发行到现在，已经过去了几十年，其取得进步的背后就是部落长者和社区成员的故事。现在，传统故事已经成为日常生活的一部分，融入制作（几乎）任何事

物的过程。当然，辅导过程有时会反过来——孩子们也会分享他们自己的故事。最终，这些工坊制造出来的就是文化本身的连续性。

关键作用

大型跨国公司可以看到不祥之兆。市场和技术正在飞速改变。内部顾问小组和外部顾问小组的团队每天和业界领导者一起设计、执行策略，从而使自己的企业变得更加敏锐，更加适应市场。偶尔，结果是公司的彻底重组，但也有可能是解散。第三次数字革命改变了生产方式以及制造商的本质，这一改变迫使许多公司重新考虑自己在这一不断变化的世界中的角色。

巴西的海洛薇萨·内维斯作为圣保罗地区数字制造的教授以及领头人，非常支持建立“数制”工坊，因为这些工坊能够改变教育和培训方面的现状——这是全世界都应优先考虑的事情。然而，她的关注点主要是“数制”工坊和创客工厂是如何改变各产业的思维模式的。她说：“我希望这种思维模式至少出现在两个不同的领域，一个是融入课程的学校，一个是融入工作的企业。而且与学校相比，企业可能更开放一些。”面对这一目标，她建立了We工坊，这是一家位于圣保罗的创客工厂，专门给企业提供咨询服务和活动安排。由此，她关注了伊斯沃尔[3] [FCA（菲亚特克莱斯勒汽车公司）[4]企业大学，相当于一家开放于社区与菲亚特克莱斯勒汽车公司的汽车装配厂] 在巴西米纳斯吉拉斯州贝廷地区新建立的一家“数制”工坊。



图6－9 菲亚特克莱斯勒汽车“数制”工坊的领导者。We工坊提供

马西娅·纳韦斯是伊斯沃尔“数制”工坊的主管，保罗·马托斯和卡罗来纳·马里尼是其中的管理者，她们认为，伊斯沃尔的“数制”工坊是南美地区最现代化、最全面的“数制”工坊之一，但亮眼的设备并不是这家工坊的核心。虽然菲亚特克莱斯勒汽车公司的拉丁美洲团队和来自社区的人员一起研究数字制造项目，但是他们学到的不仅仅是数字制造的原理。现在，他们正在加强合作能力，以一种新的思维方式来思考企业与社区的未来是什么样子的。

伊斯沃尔“数制”工坊被社会科学家称为“边界地点”——位于不同社区的交界处并且充当着桥梁的角色，即能够横跨边界，将本不可能连接的社区连接起来。除此之外，伊斯沃尔“数制”工坊正在构建一种能够带领企业与社区进入一个引人注目、令人梦寐以求的未来的DNA。



图6－10 菲亚特克莱斯勒汽车“数制”工坊对话。We工坊提供

设想，现在是伊斯沃尔“数制”工坊建立后的2030年，伊斯沃尔工坊已经在全球许多领域成为文化变革的典范。不管在什么情况下，各个领域都已经采用了伊斯沃尔“数制”工坊在2016年开发的、适用于任何个人发展的4种不同的人物角色：探索者（开拓新领域）、创客（利用技术来不断完善改进）、骇客（解构事物并且发现不同的替代品）以及网络工作者（根据需要来联系人们和创意）。从这些方面培养人才而不是围绕传统工作转圈已经成为企业运营新方式的开端。纳韦斯认为产业能比教育机构更快地接受具有颠覆性的新思维方式。

2030年，“数制”工坊成为企业必需的熔炉，因为在这儿，这些角色会被训练、强化并再度回归公司文化。现在，跨越多个领域，人们认识到骇客既有解构能力又有创新能力。他们理解人们在成为探索家的道路上需要时间和空间，所有的劳动者都可以成为不断革新的创客。当人们适应自己“网农”[5]的角色时，他

们的价值大大提升，因为别人认为网络是组织机构的重要连接部分，它越来越有机，像生命体一样演变着。拥有了这4种角色，企业就能顺应市场的变化，周而复始地解散并重组。

对于菲亚特克莱斯勒汽车公司和其他企业来说，必须尽快形成新的思维模式。随着2030年的到来，顾客、制造商和供应商三者的概念之间的区别变得越来越模糊。将每一个组织成员都看成探索者、创客、骇客或网农，这有助于每个人接受这种角色逐渐模糊的事实。菲亚特克莱斯勒汽车公司既是一家汽车生产商，又是一家服务型企业。2030年，数百个菲亚特克莱斯勒汽车公司“数制”工坊成为一种新型企业必经的新思维的大熔炉。同时，数千家其他公司都开始效仿这一模式。回到2017年，纳韦斯说：“我现在的兴趣就是产业转型。”到2030年，她在回首往事时就会发现，将技术与“数制”工坊发展的社会思潮结合在一起的愿景正改变着这个行业。

最后两个场景正处于材料科学的前沿，其中一个设想主要关于尼尔路线图的前两个阶段，另一个设想主要关于最后两个阶段。在第一个设想中，共享当地材料特性信息的平台解决了社区和个人制造的关键速率限制因素——数字制造过程中所消耗的材料的实用性、成本和对环境的影响。第二个设想是把我们带入数字材料和编程技术的未来世界。美国国家航空航天局针对太空的探索和研究，实质上对于我们如何在地球上维系我们的道路、桥梁、建筑等基础设施建设有潜在的改变。

物质科学

现在，“数制”工坊运用数字技术来设计和制造，但依靠的消耗品却是模拟材料。这些材料通常很昂贵，并且需要提取、包装和运输。许多材料都不是很环保，最终都浪费了。尽管未来某一天先进的

数字材料可以解决这些问题，但是离日常使用还遥不可及。我们必须学会从根本上重新考虑用于制造的材料，比如木头、纸板、塑料、金属，等等。如果不这样的话，我们就会发现数字制造的飞速发展将在之后的10年里给环境施加无法承受的压力。

阿莉西娅·加姆勒薇兹是智利圣地亚哥大学经济学院循环经济创新助理教授。她在牛津大学赛德商学院研究了普遍环境下的3D打印技术并获得了博士学位。她的承诺是发展循环经济（资源循环利用）而非线性经济（提取、利用、丢弃资源）。她秉持的原则是：“我们需要理解身边的材料。”拿自然界来举例，自然界的有机体取用当地的自然成分作为资源来建构从纳米尺度到宏观尺度的复杂物体，并且继续循环利用资源。森林不会制造垃圾。现在，她正在建立一个开源材料的数据库，利用分布式知识和数字制造带来的收益，使“每个人、每个地方都能运用数字设计和周围的材料制造高质量的产品”。



图6－11 阿莉西娅• 加姆勒薇兹。牛津大学赛德商学院提供

设想现在是2025年。加姆勒薇兹的开放平台——物质体（Materiome）——成为全球成千上万的“数制”工坊运营中的重要环节，它是一个复合数据库、社交网络和共享知识库，用于了解、采购并使用当地的数字制造资源。她的开放平台挑战了通过低效、高碳足迹的全球供应链购买用于融化的塑料颗粒、用于重组的金属粉末或其他材料的模式。取而代之的是，富有进取心的制造商和创客们正在以新的方式利用当地环境。他们正在收集沙子、植物纤维、天然树脂、盐水和其他当地材料，并记录其性质和用途。移动增强现实技术（AR）使创建一个本地地理标记的“增强层”成为可能，该“增强层”描述了当地可用消耗品的生物和化学特性以及潜在用途。人们可以在他们的社区中走动，并且通过这种增强技术，将垃圾或沙漠视为可获得的自然资源。

通过“数制”工坊能够从纳米尺度到宏观尺度构造材料，丰富的天然材料具有高性能，甚至胜过受严格保护的化学配方。社区和公司不再为新化合物申请专利，而是开始对将简单成分组装成（几乎）任何东西的新方法进行投资。在此过程中，对循环经济的认识和理解程度正呈指数级提升。

在世界各地，人们都知道循环经济与控制能源、材料与信息有关。现在市民不再让公司去追踪他们生产的产品，而是借助工具来轻松地测量他们从当地乃至全球获得的材料的组成成分以及其他特性。“物质体”这一平台让更多人参与到材料数据的生产之中。这种力量给市民提供了推动循环经济变革的强大杠杆。

随着地理与生物科学方面的实物样本的共享以及再利用，人们对“物质体”的研究取得了长足进步。这些相关领域的科学家都面临了同样的挑战——哪一种元数据可以添加到给定的材料或实物样本上。他们都需要数字平台来支持数千个来源的分布式输入和高效搜索与发现。最重要的是，他们共同致力于开放信息共享。对于生物学家和地理学家而言，“数制”工坊的进步不足为奇——公民科学家在这两个领域中有很悠久的传统。事实上，代

际社会生态系统早已出现，生物学家、地理专家与青少年和青年公民科学家一起通过社交视频和虚拟现实来记载各种有想象力的、有趣的制造项目。材料科学家变得越来越具有病毒性、可验证性和社会性。

起初，商务企业没有意识到“物质体”发生了什么——这一平台进行的似乎是边缘活动。毕竟，对任何耗材生产商来说，“数制”工坊只占其销售金额的一小部分。第一家受到关注的商务企业是在航空航天、汽车和其他领域最先使用先进材料的行业制造商。对于这些企业，材料的特性代表了严格监管的竞争秘密。所以，这些产业的许多企业对免费开放资源的平台感到很惊讶，因为这个平台可以分享其急需的信息，而且它们惊奇地发现了以前从来没有见过的材料。

当然，对于那些化学品和其他商业材料的制造商来说，“物质体”已经形成了一种非常强大的竞争力，因为他们面临如何在一个开放资源、开放当地原材料的世界里运作的生死抉择。而对于循环经济的支持者来说，开放资源平台是一个具有重要影响的成就——越来越多的人利用当地原材料来制造他们所需要的东西，这样就不会再浪费和增加运输成本了。政治学中有一句格言叫“跟随金钱”，在数字制造方面，这可以被称为“跟随材料”。

自我维护升级

对交通基础设施的巨大投资——马路、桥梁、轨道、建筑、机场——在全世界的社会中都是必要的。预期50年寿命的桥在建成75年或100年后的今天仍在使用。机场的运营能力是设计时的二到三倍。工业革命随着交通基础设施的发展而发展，随着全球的发展，交通基础设施越来越重要。随着新的交通基础设施的建设，这个世界是否会被设

计成一个人口增长、技术加快、气候变化、人口大量迁徙和有着意想不到的社会发展的世界？

美国国家航空航天局的张永安[6]把数字材料运用于解决基础设施老化的转型方案之中，这是一种超出以往设计性能和预期耐用寿命的方法。张永安在麻省理工学院获得博士学位，并在比特和原子中心专注研究尖端材料科学。在美国国家航空航天局，他专注于数字材料与基本运算法的研究，对太空结构产生了深远的影响。毕竟，把人类建设队伍带入太空会有些危险和昂贵。归根到底，这一影响是十分深远的。他喜欢展望未来，那就这样做吧。

现在是2040年，我们认识到许多实体建筑（桥梁、建筑、飞机和其他社会基础设施）能够通过不断更新来满足持续改变的服务需求。桥梁使用寿命不再是50年，数字材料和机器人可以对桥梁持续进行检查、维修和重新配置，使其始终处于积极的运行状态。随着时间的流逝，桥梁也许看上去和以前差不多，但会因为不断翻修而成为全新的桥梁。桥梁可以根据社会和环境的需求来重新设计，以增加或减少承载能力。我们没有试图让人们通过重新编程而变得更有远见，而是让材料通过重新编程而自我维持，并由人们来指导这一过程。

在这一未来设想中，供应链是区域性的，并且侧重于模块材料的发展——就像在2017年生产的螺母和螺栓。实际的材料还不是尼尔未来的完全可编程物质（尽管已经在全世界的研究实验室中被发现）。当然，这些基础设施使用的材料与2017年组成基础设施的材料相同。现在它们是模块化的可重构性物体，是可扩展的制造系统的积木式组成成分，并且适用于数字设计和组装。相同的材料也能被运用——例如建造桥梁的钢铁。但是为了取代数千种独特形状的钢材——每一种钢材都拥有独特的钻孔和到达建筑工地的特快专递，当原材料第一次被加工的时候，人们可以创造许多模块化组件，作为原材料初次加工时的主要产品。人们不需要在之前的成品批量生产中进行再加工，而是可以运用机器汇

编程序通过重新配置建筑模块来制造（或重制）顾客所需的样式。

首先，这种变化很细微，大概与高速公路上实体收费站到自动收费感应器的转变相类似。同样，人们开始在“数制”工坊研究如何将重新配置的材料用于个人应用；行业供应商开始鉴别原材料是否可以被制造成模块化的、可编程的组件，应用于目标应用程序。最先进的应用存在于太空，在那里可配置的成分是最昂贵的——研究太空所需的新的原材料、制造和施工队伍的成本过高。

交通基础建设的模块化方法创造了一个庞大的可节省成本的运输系统。自从有远见的政策领导者意识到这种智能的模块化基础设施很有前途，他们就提前做好让其取代劳动力的准备，让操作变得自动化。他们走在潮流的前面，为这种版本的“数制城市”承诺、“数制乡村”承诺以及“布莱尔格言”打下了良好的基础，从而促进这种向工作岗位缺乏、生活质量高、有尊严的经济体的转型。这才是一个不是只依赖有偿工作的经济体。基础设施的自足性对社会自足性起着十分关键的作用。

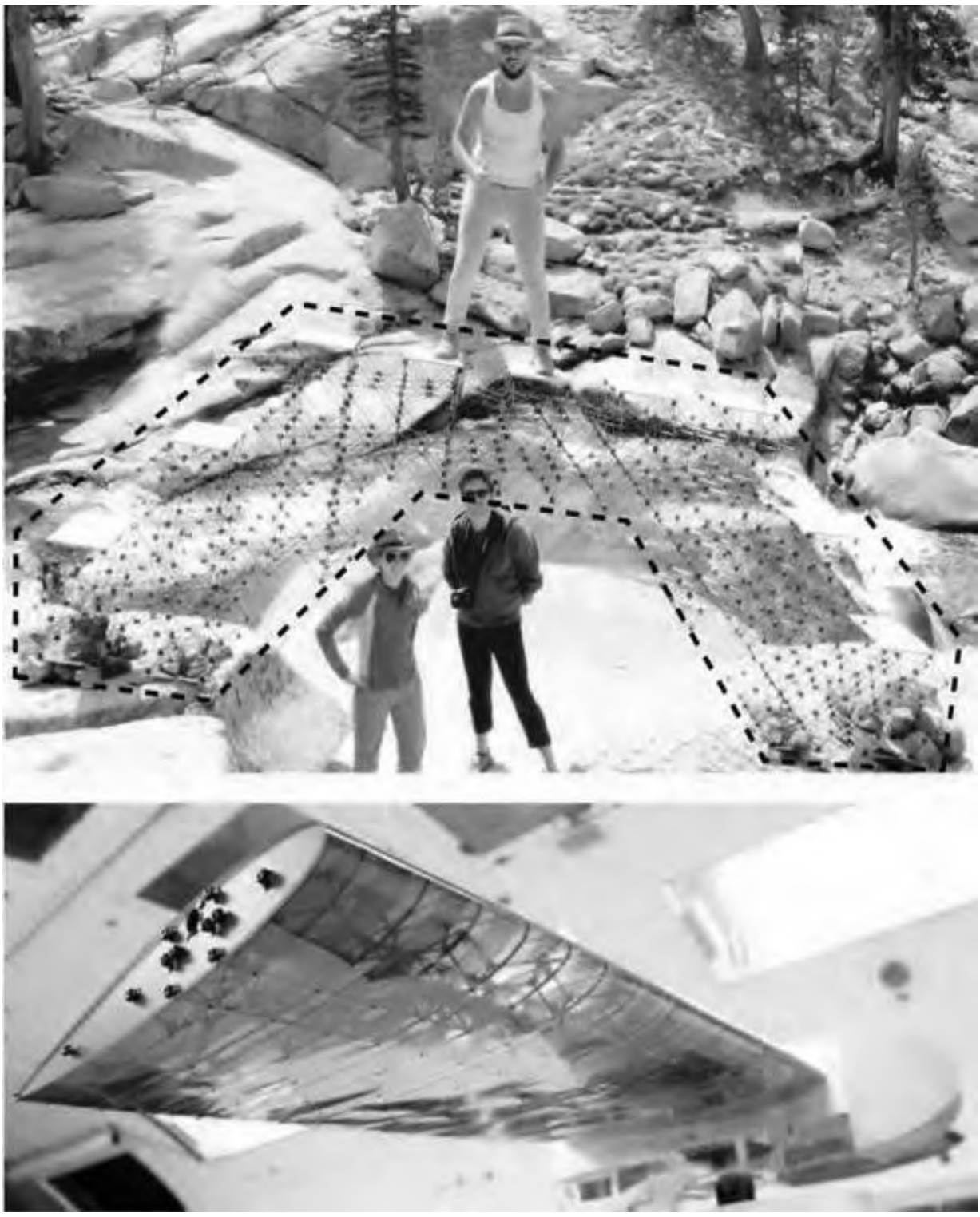


图6－12 数字材料结构。桥：本杰明·耶内特、丹尼尔·切卢奇、克里斯蒂娜·格雷格、张永安提供；翼：本杰明·耶内特、萨姆·克里奇、丹尼尔·切卢奇、尼克·克拉默、尼尔·格申斐尔德、肖恩·斯威、张永安提供



图6－13 张永安。美国国家航空航天局提供

这8个设想能够互补互织，把现在创建成一个稳定和谐的世界，在这里，人们和社区能够实现全球交互。如果我们想要看到这些梦寐以求的设想变成现实的话，我们需要开始为此打下基础，一步步顺利地将其实现。正如穆格公司首席技术官冈萨洛·雷伊所说：“你经常可以看到未来的愿景以及许多阻碍，但是你也可以看到可能的未来，然后攻克这些阻碍。大多数阻碍来自社会，而不是技术。”随后他指出，“尼尔做出了这些疯狂的预测，但是随着时间流逝，预测可能会成真。也许不会十分精准，但是肯定是在那一方向上前进的。”我们的目的是打破这些梦寐以求的、可实现的设想的平衡。为此，我们引进了一个引导性模型。

预言性转化模型

为了实现这些梦寐以求的愿景，甚至更多愿景的组合，我们需要解决在第2章中就提出的基本挑战。我们必须广泛改善数字制造硬件、软件和材料；我们必须培养技术和文学素养，使每个人都能有效利用这些技术；我们必须创造一个可用的生态系统来促进共同目标的实现；我们必须确保逐渐减少风险。

除了解决这些基本挑战，还有一个存在已久的问题需要不断管控。在这些不同的数字制造参与者和社区（比如“数制”工坊、创客工厂、骇客空间，等等）中，存在着一个有关互相依存和独立的紧张关系。硬件和软件开发者需要平衡这些易用性，使这些技术公开扩展。管理人员会在集中化和层次化的效率与分散化和分布式代理的平衡中挣扎。导师会一直面临教授别人知识与开发自己的项目的抉择。在这些情况下，存在着一个新兴生态系统和根深蒂固的制度之间的紧张关系。

伴随着这些令人向往的场景、门槛挑战和持久的紧张心理，当我们试图为塑造第三次数字革命提供指导时，我们找不到一个与任务相匹配的模型。尤其是，我们需要一个模型来优化数字和实体世界，这个模型是一种分布式参与方和利益间的复杂组合，不依靠单一管理者或实体来控制其进程。通过写这本书，我们开发了这种“预言性转化”模型。

这一模型具有预言性是因为它是以一张已被预测过的技术路线图为主、基于提高数字制造性能的可观察的指标的模型。这一模型是一种变革，因为我们不但对简单的预测感兴趣，而且对创造一个能够与社会系统一起演进的技术，进而给社会带来积极结果的预测也感兴趣。我们写这一章更多是为了路径创造者，而不是路径观察者。

该模型建立在许多已建立的变化模型的基础上，并与之形成对比。大多数早期的模型——从黑格尔基于康德理论提出的辩证模型（正题、反题、合题）到库尔特·勒温[7]在1947年提出的具有影响力的改造模型（解冻、变革、再冻结）——能够预见变革和变革的必要性。但是他们也假定变化速度是线性的，最后以平稳状态结束，而不是以持续加速的变化结束。

现在受人欢迎的组织变革模型也是以平稳状态结束的，而不是以持续加速的变化结束的。例如，约翰·保罗·科特于1995年提出的被广泛引用的领导变革模型提出了管理变革的8个步骤，并以制度化变革（一个新的稳定状态）结束。同样，威廉·布里奇斯的2000年转型管理模型在个人层面上对人很有帮助，它记录了人们在接受新事物之前需要如何摆脱旧事物。然而，在一个技术呈指数级增长的世界里，抛弃陈旧事物成为一种无休止的状态。与我们的模型最接近的是W.爱德华兹·戴明的持续变化模型，这是一个计划、执行、检查和调整的循环，尽管他的模型并不是为利用分布式生态系统的力量和广泛的社会变化的力量而设计的。

许多分析新技术的采用、影响或周期性的常用模型都提到了埃弗雷特·罗杰斯的《创新的扩散》（1962年）、杰弗里·摩尔的《跨越鸿沟》（1991年）和克莱顿·克里斯坦森的“颠覆性创新”理论（1997年）的内容。美国政府还制定了一些有帮助的评估准则，包括美国国家航空航天局的“技术就绪水平”和美国国防部的“技术就绪水平评估”，以评估特定技术的成熟度水平。尽管所有这些模型都对将技术引入市场或渠道具有宝贵见解，但没有一个模型能将对持续加速的技术的分析与围绕一致的社会影响目标对不同的独立的参与者进行吸引和授权的机制结合起来。

我们设计了第三次数字革命的预测转化模型，但该框架可能与识别和塑造其他加速发展的技术相关。当然，这个模型只是在技术加速发展的世界中组织思想和行动的众多可能方式之一。这一模型是按照以下4个步骤建立框架的：

预测：变化速度

结盟：参与者

培养：赋能的生态系统

共同演化：技术和社会

因为这一模型需要对加速发展的技术有清晰的了解，所以这会在科学家、技术人员、社会科学家、故事讲述者、政策制定者、推动者和其他人之间形成一种持续的跨产业合作。这一模型被设计为一个框架，用来帮助传播进程，建议人们采用那些可被分布式参与者采纳、调整并推行的平台、工具和实践。这一模型应当具有灵活性、可适应性以及可扩展性，以在本土、区域、国家甚至国际层面上运行。

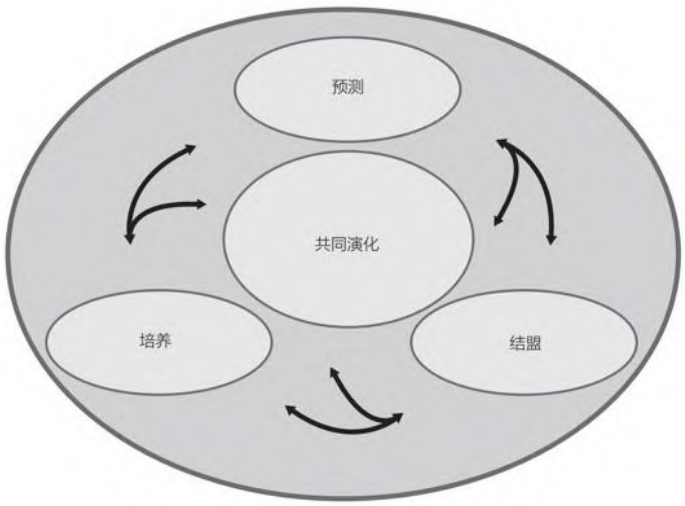


图6－14 预测转化模型是一个持续的过程。乔尔·卡彻-格申斐尔德和艾伦·格申斐尔德绘

尽管模型中的4个元素是按顺序呈现的，但它们之间的交互是动态的。图6－14显示，模型的前三个部分（预测变化的速度、协调结盟的参与者和培养新兴的生态系统）是相对连续的，当每个阶段都取得进展时，它有助于技术和社会的共同进化。此外，它是一个连续的循环——一个螺旋，只要有稳定的支架（不断努力），它就会持续上升，在整个过程中保持收益。

由于一本书在很大程度上是以线性呈现的，所以我们按顺序处理模型的每个阶段。在最后一节，关于共同演化的部分，我们可以很清楚地看到它们的综合。

科学即新闻

我们描述的所有愿景和其他正在出现的愿景，都是建立在数字制造性能不断改进的假设之上的。我们从指数角度而非线性角度思考并不是一种本能。所以，我们从预测数字制造技术的改变速度开始这一模型。这需要对驱使技术加速的、和现在的研究路线图有关的基础科

学有一个基本的了解。当共同演化先驱者斯图尔特·布兰德在2016年的线上沙龙问答中被约翰·布罗克曼——一位思想领袖和代理人——问及“你认为最近科学新闻上最有趣的是什么”时，布兰德的回答十分具有建设性：“科学就是新闻。当你翻阅报纸或杂志时，所有人感兴趣的内容就是老一套的“他说她说”，有关政治经济的内容就像反复表演的戏剧一样，有关时尚的内容就像对新事物乏味的幻想一样，甚至，如果你了解科学，便可以预测技术。人性不会改变太多，科学也是。改变越来越大，不可逆转地改变着世界。我们现在生活在一个改变速度最快的时代。所以科学是一大趣事。”

布兰德说：“如果你了解科学，便可以预测技术。”在我们的模型中，我们从最开始就侧重于了解科学以及科学如何给研究路线图提供框架，给社会带来机会和风险。

我们在对数字制造领导者的调查中，询问了理解数字通信、计算和制造的基础科学的重要性。这些回答表明，对于超过70%的参与者来说，对基础科学的理解是非常重要的。与此同时，近40%的人表示这很难做到。对于这些反应我们用与第2章相同的可视化方法进行了说明（中间代表集中趋势，外面是离群值）。

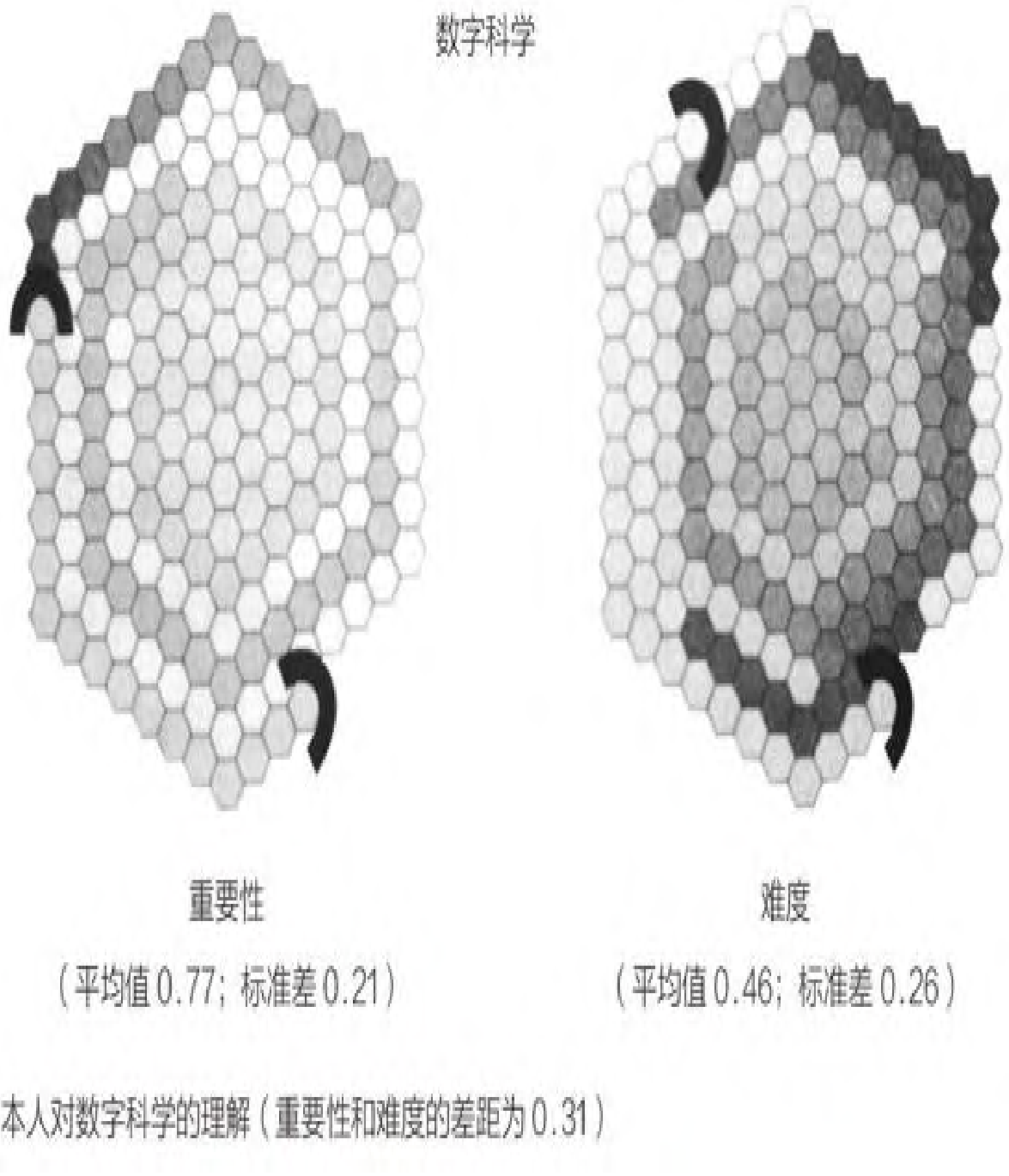


图6－15 理解有关数字通信、数字计算和数字制造方面的基本科学的重要性和难度。WayMark Analytics公司绘

如6－15图所示，对于难度的回答有很多种。当系统中存在大量的可变性因素时，改进就很困难。对于那些想要理解和深入讨论变化率的人来说，尼尔的章节值得一读。但这并不意味着你必须赞成他所有

的假设，这是一个明智的讨论和辩论的起点。目前，只有大约1/4的“数制”领导人认为很容易了解基础科学。在这本书的写作过程中，我们不断地敦促尼尔，希望他能让我们和读者都能更容易地接触更关注科学或技术的章节，因为他所写的内容涵盖的资料对于理解可能的变化速率是必要的。我们对科学和技术路线图了解得越多，就越了解形成第三次数字革命的关键步骤。对基础科学的共同理解也会影响拉斯定律，这有助于组织我们的思维，也有助于确保我们在讨论数字制造性能时不会互相推诿，使我们的合作更加有效。

审视拉斯定律

数字制造路线图中的改变速率可能会随着时间的流逝而加快或减慢。比如说，在1975年，戈登·摩尔重新修改了他的预测，认为数字计算性能会从每年翻一番变成每两年翻一番。直到现在，数字计算性能的核心定义对于摩尔来说仍是不变的，而且需要适用于数字制造——这样在改变速率上的讨论和争辩就能运用相同的标准。对于拉斯定律来说，正如整本书所讨论的，我们用这个定义来定义数字制造性能：

> 现在“数制”工坊的性能是能够智造（几乎）万物，制造实体物品并对其功能进行编程。

找到大家一致认同的有关数字制造性能标准的定义，对于帮助不同领域的参与方对拟建的发展路线图以及（技术）变化速率等有社会意义的问题进行有效讨论、争辩、预测是十分关键的。一些由未来学家、权威人士及新闻记者组成的“家庭小作坊”对未来技术和社会的组成进行了预测，其中大多数人设想数字制造的未来会更好、更快速、更经济。他们的结论表明，他们并没有真正意义上地投入这项技

术。不幸的是，缺乏这种意识是许多教育者、政策制定者、慈善家和投资者的通病。正如我们在书中所述，3D打印机对于数字制造来说是增材制造过程的一个重要部分，但只能算“数制”工坊内众多数字技术的一种。就像尼尔说的：通过3D打印的单一角度去观望数字制造的天空就等于通过微波炉去观望食物制作的整个世界一样，具有片面性。

鉴于新兴技术的复杂性，许多人不愿放眼未来，并对预测行动提供信息的技术的发展感到犹豫不决，这是可以理解的。当然，我们可能会在许多社会科学家和政策制定者的身上看到这样的不情愿。

的确，这在学术报告中经常出现，初出茅庐的学者常引用这样的措辞：“预测是困难的，尤其是关于未来的预测。”这句话来自对尼尔斯·玻尔、塞缪尔·戈尔·德温、尤吉·贝拉、马克·吐温、诺斯特拉德马斯等人的格言的总结。在学术演讲中，这会被认为是一种自嘲式幽默。然而，它有一个微妙的不容置疑的影响，即强化了一种倾向：既做出小的增长性预测，又避免任何可能被视为“出头鸟”的东西。

即使是前两次数字革命中最老练的思想家和实践者也采取谨慎态度。在他们最近的一本名叫《爆裂：未来社会的9大生存原则》的书中，尼尔的同事伊藤穰一——麻省理工学院媒体实验室的主任和《连线》（*Wired*）杂志的编辑杰夫·豪强烈宣称：没有一个人可以预测未来，只有傻子才会去尝试。随后他们共同提出了9个组织原则来引导指数级变化。其中，他们建议把要解决的事件放在权威性之前考虑，把牵制放在推进之前考虑，把风险放在安全之前考虑，把整个系统放在个别研究物体之前考虑。这些原则都是为了解决这个问题：我们目前的认知使我们无法理解从通信到战争的各个领域的快速进步所带来的深远影响。

尽管这9个原则都是对数字时代中的合作的新思维模式的进一步洞察，但我们并不赞成伊藤和豪发表的言论：改变速度过快导致人们不

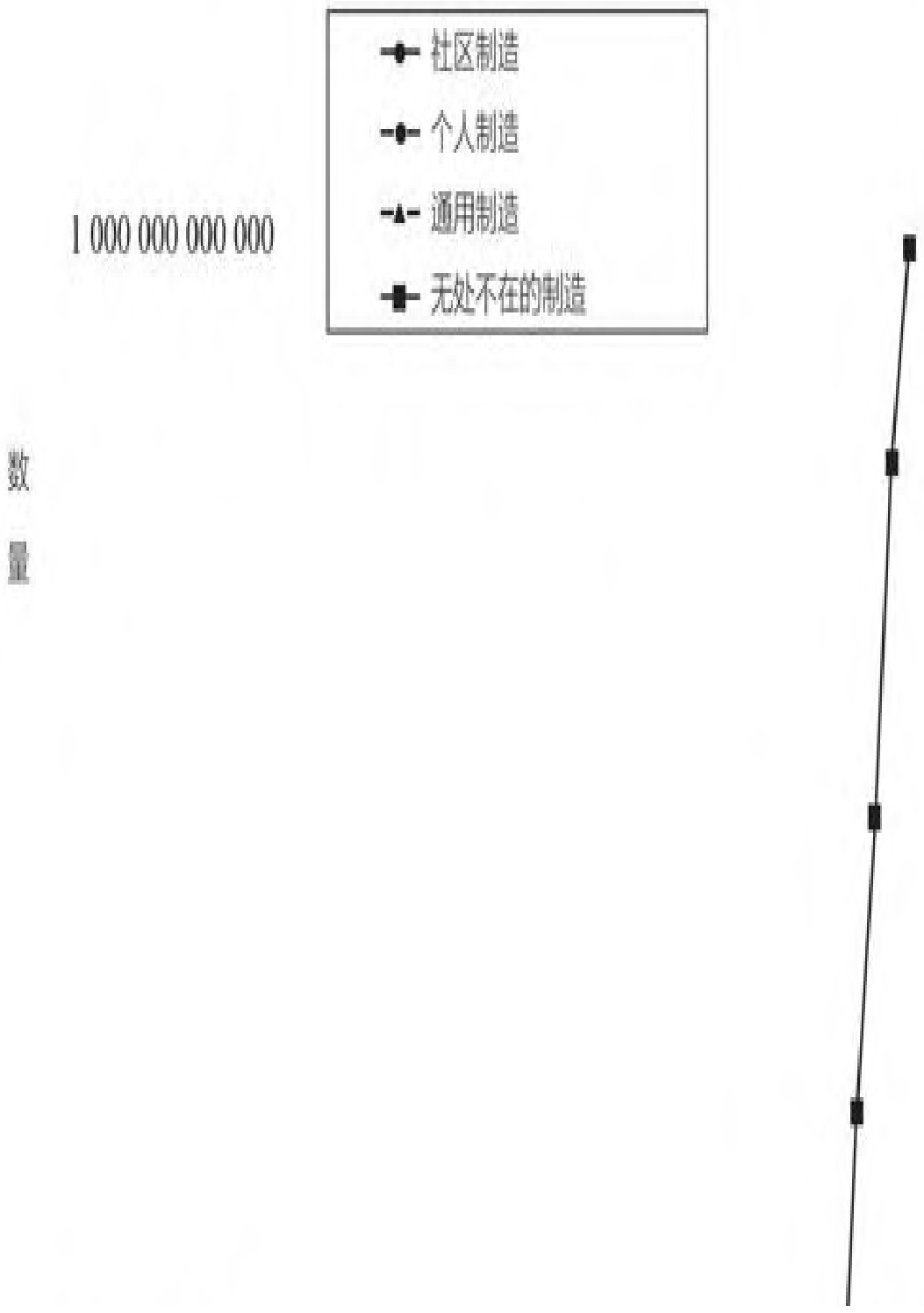
能对未来进行预测。他们指出这一点是为了培养自立能力和灵活性，就像他们说的，拥有指南针而不是拥有路线图的重要性。相反，我们认为，如果没有路线图，就无法培养所需的机构，这样，分布式的个人就能够被授权进行集体辩论、讨论，并最终使发展加速、减速或以其他方式改变路线图。你不仅需要一个指南针，还需要一张路线图。一个更具有远见的态度能够帮助我们把观察由预测转化成集体行动（运用伊藤和豪的原则），从而实现这些预测。正如管理圣人（和其他人一同被授予这一评价）彼得·德鲁克所说：“预测未来的最好的办法就是创造未来。”因此，预测改变速度是一项团队合作活动。

戈登·摩尔在初次观察数字计算性能的过程中一丝不苟。他发现，在集成电路上，每平方英尺的晶体管数量每年都会翻一番，而且如果这个数量的变化速度一直如此，那么这一变化就是真的。然而，随着时间的推移，摩尔定律在许多相关技术上都有所体现，这些技术可以帮助计算性能变得更加有效、快速、实惠，从而改造社会。

对于数字制造来说，拉斯定律指出：鉴于过去的指标和现在的研究路线图，数字制造性能每一年半会翻一番。这种变化的速度是否成立，很大程度上取决于所有参与者，从字面上讲，他们与这种指数级变化的速度有很大关系。如果像摩尔定律一样，拉斯定律从一个观察报告发展到一个标准程序，再发展到一个不同参与者的集体目标，那么，正如我们在第4章中提到的那样，我们可能会在接下来的半个世纪看到数字制造性能的增长速度与摩尔定律相似。

现在拥有了可观测的数据、清晰的定义和已掌握的足够信息，我们能够预想第三次数字革命将会是怎样的。我们和“数制”工坊的领头人合作起草了在本章开头所预想的设想——如果事物可以通过一种转化的方式发展会怎么样呢？通过分析，我们画了一张与尼尔路线图中4个步骤相对应的图。这张图展示了每一种连续的技术将如何以更快的速度增长，当它赶上前一种技术时，前一种技术可能会趋于平稳，形成尼尔在第3章中提到的S形曲线。精确的轨迹不能被精确地指定，

并且图形不能按比例缩放（我们不能使它既适合于页面又指示前两条曲线的形状）。我们的目的是详细说明并预测不断加速的变化浪潮。



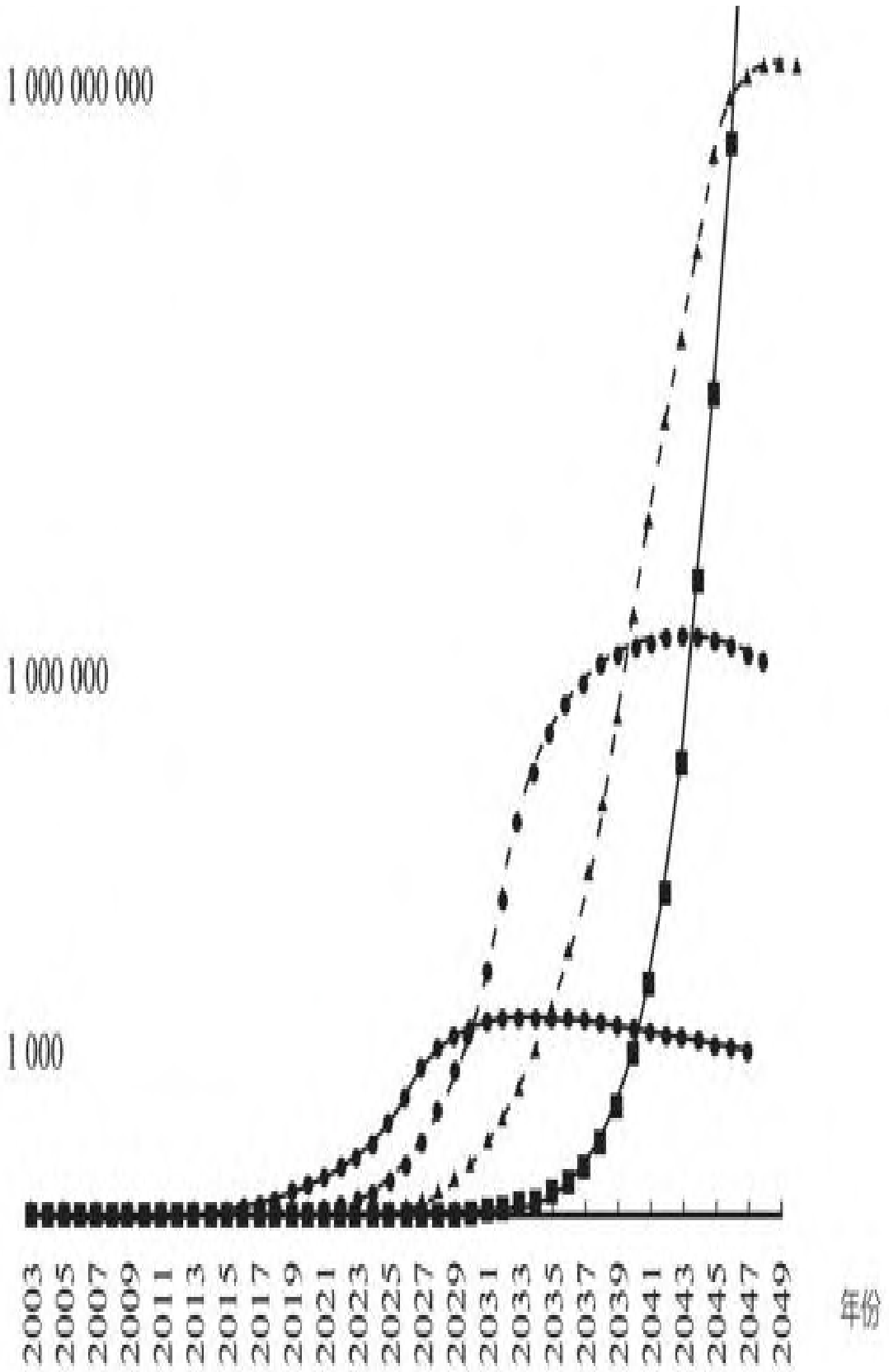

1 000 000 000
1 000 000
1 000
2003
2005
2007
2009
2011
2013
2015
2017
2019
2021
2023
2025
2027
2029
2031
2033
2035
2037
2039
2041
2043
2045
2047
2049
年份

图6－16 随着时间推移的第三次数字革命的4个阶段。乔尔·卡彻-格申斐尔德绘

这个数字表明了社会挑战的规模和范围。个人、组织和机构在跟上社区制造速度方面面临着挑战，因为他们必须确保数以万计的社区“数制”工坊的成功。当我们试图追赶下一个曲线——个人制造，并需要使几百万的早期采用者在个人制造技术方面取得成功时，任务变得更加困难。当我们将社会系统与最后两条曲线相匹配时，这项任务就变得惊人了。在这两条曲线上，我们以数十亿个人或数万亿事物的规模进行运作。

用分析的方法来预测变化的速度是有意义的，但也仅此而已。这就是不同学科的作用——提供基于数据的替代观点和未来展望。数据及其可视化有助于显示可能的未来，并为参与者——包括路径观察者与路径创造者（技术和社会）——联合起来共同创建更好的未来奠定基础。

联盟：利益相关参与者

纵观全书，我们已经突出强调了各行各业的参与方积极参与工坊社区活动和更为广泛的创客活动，包括：

工坊从事者、创客和骇客

政府政策制定者

硬件提供者

社区组织者

软件提供者

慈善家

材料提供者

投资者

研究者

产业界领袖及工人

教育家

服务提供者

媒体领袖

家庭

在致力于使所有参与方形成联盟的过程中，由于大部分参与者仅被包含在局部的名单中，所以，就连具体列举参与方的类别目录都是极具启发意义的。所以，回过头来考虑整张列表是我们行动的第一步。第二步就是有更深的理解：这些团队并不是完全孤立的，许多个体通过不同的特征与不止一个团队联系在一起。此外，这些参与者的利益——一些人称之为数字制造的价值定位——既相对和谐统一，又互相矛盾，而且所有这些利益都是动态的。在这种情况下，要想有改变，就要求有足够的联盟，从而促成集体行为。不可避免的是，最初只有一小部分的参与者在某些利益上达成了一致。很快，我们的小和谐催生了大妥协。

乔尔的研究团队将参与者联盟定义为“独立而又相互依存的个体、团队和组织联系在一起以推进他们各自和共同的利益”。这种说法和大多数人使用“参与者”一词的方式十分不同。通常人们谈论的是“参与方管理”和“参与方参与”。当人们说他们需要进行“参与方管理”时，意味着需要中和一下意料之中的对于某些政党的反对意见。当人们说需要进行“参与方参与”时，意味着将某些提供必要支持的团体囊括进来。“参与者管理”和“参与者参与”都很重要，却不完整。另外，我们需要建立参与者联盟。这是一个动态的过程，可能将会形成“共同利益”及“相对竞争对立利益”的混合体，需要通

过不断的对话、观察、协商和磨合来维系。权力差别以及其他动态因素也要被考虑进去，这个动态的过程一直持续到足够的既独立又相互依存的团体达成一致并支持集体行为的时候才会结束。

小和谐催生大妥协

尽管全球大部分的数制参与方制订了广泛的、部门范围内的目标，比如确保数制接入以及读写能力，但也有些更小的参与者团队经常制订具体的目标，以获取短期的利益，增加动力。当成群的参与者认为大部分的利益是可以协商的时候，突破最有可能发生。尽管有些红线无法跨越，但是历史告诉我们：当新观点被认同，情况有所变化时，利益也能改变。最终，社会体制是行为模式化的产物，包括模式化的协议。建立在参与方为了各自的利益而进行的对话之上的联盟与小和谐可以随着时间不断累积而形成大妥协，从而解决社会面对的挑战。因此，建立在许多小和谐之上的大妥协，反过来，也依赖于参与者及其利益的建设性整合。

正如格拉布尔基金会的格雷格·贝尔在谈论重构学习的萌芽时所指出的：“很多共同的影响来自许多通过网络做出的微小决定。”我们通过各种灵活的、小规模的赠款，鼓励参与者承担风险并共同试验，取得了进展。慢慢地，这种方式催生了许多深刻见解、规划以及积极性，这些举措帮助大匹兹堡地区形成了重塑教与学的未来的更大愿景。

在小和谐催生更大的目标与人类心理学和游戏设计的交叉之间同样存在着有趣的关联。匈牙利心理学家米哈里·契克森米哈研究出“心流”的概念。他发现当一个人觉得自己能力较低而任务较难时，这个人就会感到焦虑沮丧；然而，如果一个人的能力较强，任务难度较低，这个人就会感到无聊，自由散漫；当能力与任务难度系数大致成比例，尤其是人们为对个人有意义的目标而努力时，人们就会进入

一种流畅的状态。米哈里·契克森米哈在他的《优秀商业：领导力、流动、创造意义》（*Good Business: Leadership, Flow, and the Making of Meaning*）一书中将“心流”描述为一种在规定的活动和情况中专心或全神贯注的状态——在这种状态中人们是最开心的。

因为现在的游戏都给予玩家代理权，让他们做决定，并且为其提供持续的反馈，帮助他们掌握游戏（以游戏的目标为背景），所以游戏设计者特别擅长设计流态。他们还学会了设计集体流，在这种状态下，团队可以持续承担更小的挑战，催生更大的目标。大的社会议题，如解决气候变化或者确保“数制”的普遍适用性，都是势不可当的。因此，将这些巨大的挑战分解成可实现的步骤至关重要，每个人都觉得他们的行动很重要，他们正在发挥作用，他们是朝着共同目标前进的一部分。

建立信任

在匹兹堡，“重塑学习”成功的另一个关键因素是优先考虑通过非正式的社会交往建立信任。事情从煎饼早餐开始，随着时间的流逝，每一项活动都吸引了越来越多的人，因为人们能够通过自己的个人交际网络与别人联系，并且这些人将会带来两到三个他们认为必须到场的人。

为了确保团体建设充分发挥其潜力，社区、地区甚至整个国家都需要类似煎饼早餐的东西。例如，当一个由“数制”工坊或创客空间组成的网络出现时，关于其使用的术语经常存在争论：“数制”工坊、创客空间、骇客空间，或者一些包罗万象的涵盖性术语。争论不仅仅是关于文字的——这些术语中包含的文化和操作实践是相关的，但是在不同的领域中又是有区别的。

社区数字制造的潜在主人之间也可能会存在争论。主办单位是图书馆、K－12学校、社区学院、大学、博物馆还是独立的社区中心？如

果在这6个可能的选项中有一个均匀的分配，那么这个愿景会被妥协或推进吗？大学通常拥有丰富的资源，但不能总是提供开放的访问权限。社区中心和图书馆有开放的访问权限，但通常没有使用新技术和复杂技术的经验。最终，建立和有效运行本地工坊的一个基本要素是有足够的时间在参与方之间建立信任和清晰的沟通。

本着这种精神，当成功依赖于既独立又相互依赖的参与方之间的一致行动时，花时间建立信任和尊重的文化是至关重要的，这样可以使每个人都能团结协作地解决问题——但每个人仍然尊重每个组织的独立目标、能力和束缚。当然，在某些情况下，他们的利益不但不同，而且相互对立。

《永不孤单》的背后故事说明了在不同参与方之间可以通过共享利益来建立信任。在游戏开发开始之前，库克湾部落理事会和电子线传媒花了数月的时间来建立信任，进入彼此的世界，了解彼此的家庭和社区。作为合作伙伴，这两个组织都致力于设计交易条款，以确保双方在共同面对一致的影响和财务目标时处于同样的位置。通过这种方式，当出现挑战时（许多人确实遇到了挑战），合作伙伴可以利用各自不同的视角和技能共同解决问题。其结果不仅是一款成功的、令人回味的电子游戏，还是一种更深层次的合作关系，包括与数字制造相关的产品和服务在内的一系列活动。

减少偏差

正如我们在第2章中所描述的，“数制”工坊、创客、骇客、科技工坊和其他相关社区确实有共同的利益，但也有不同的优先级、方法论和社区规范。理解这些差异很重要，因为系统变化的一个基本原则是，必须先减少偏差，然后才能通过集体行动加以改进。图6－17显示了一个假设的飞镖游戏，有助于说明这一点。

在这个游戏中，左边的玩家在这个回合中得分更高。然而，右边的玩家更容易被指导（飞镖击中的位置只是稍微移动一点儿）。相比之下，如果左边的玩家被要求上升一点儿或下降一点儿或向左（右）移动一点儿，结果说不定就大不一样。通过类比，要改进与数字制造相关的非常多变的社会系统，首先需要减少变化，即使是在早期阶段，也应如此。但如果你不了解参与方和他们的利益，以及这些利益如何随时间改变，你就无法减少变化。在这本书中介绍的许多z-flowers中，方差都很高（实际上是双峰的）。改进不能仅仅集中在移动手段上。（当然，如果你不知道靶心在哪里，那么左边玩家的方法是有用的。减少变化总是与保持思想的多样性相平衡。）



图6－17 假设的飞镖游戏说明了改进可变系统过程中面临的挑战。乔尔·卡彻-格申斐尔德绘

根据我们对参与方联盟的分析，我们确实发现了一些关键点。在我们采访的所有数制参与者中，我们发现，在应对两个关键挑战方面，各方立场非常一致：通用数制访问和素养的普及性。对“数制”工坊开放访问的优先考虑反映了这一点，智造宪章也明确强调了这一

点。实现通用数制访问并非易事，需要在不同的群体、组织和机构之间进行协调和持续的努力，重要的是一个赋能的生态系统的支持。

培养：赋能的生态系统

一旦参与方了解预计的变化速度，并围绕共同的目标（如通用数制访问和读写能力）进行充分的协调，行动的关键机制就集中在培育赋能的生态系统上。赋能的生态系统可以帮助人们以指数级的速度促进共同利益和操纵相互竞争的利益。

在第2章中，我们认可了培育一个赋能的数制生态系统的必要性，这是实现通用数制访问和素养普及这一宏伟目标的一个挑战。我们特别强调了这样一个数制生态系统的某些基本要素，包括：

公众、私人和慈善家为“数制”工坊项目提供资金的多样化组合；

在工坊内部和工坊之间进行有效的协作和知识共享；

广泛分布的数字制造指导和领导；

为数制产品和服务提供开放和健全的市场；

机敏的管理机制与新兴的数制社区的价值观一致。

没有这些要素（每个要素都可以被视为有自己的生态系统），数字制造技术将无法扩展其性能，也无法与尼尔的路线图保持一致。这将大大降低我们利用数制技术改善社会的能力，同时也增加了破坏稳定的数制鸿沟的可能性。

培育一个赋能的数制生态系统需要自上而下的扩展计划以及自里向外和自下而上的传播。自上而下的扩展对于大规模的基础设施和研

究投资是必要的，而自里向外和自下而上的传播则对于给予各种各样的全球参与方在不同的区域环境中的动机、兴趣、技能和操作权限是必要的。自上而下的模式是常见的，而培育生态系统是新事物。培育生态系统的工作是一个广泛共享的机会和责任。所有参与方都需要像系统架构师一样思考，因为他们的分布式培养行为实际上是在创建生态系统。

在概述培育赋能的生态系统的总体战略时，我们首先从政府和大型非政府组织着手，它们有能力开展超出大多数组织和个人范围的活动（类似于发展互联网的早期关键投资）。与这些公共利益机构的作用并行的是私营部门的作用（类似于英特尔和其他数字先锋公司的早期投资）。大学（以及社会和技术）和社会企业家等领域的创新者构成了生态系统的第三条支柱。这些参与方的结合不是一次性的，而是持续的，这使得我们在通用数制访问和素养普及（我们的主要关注点）的基本挑战，以及其他一些基本挑战（如风险消减）中取得进展。

政府和非政府组织

有些倡议只能在政府和大型非政府组织提供的规模和范围内进行。虽然其中一些计划需要从零开始，但许多计划可以建立在现有的数字访问和素养普及倡议之上。这些举措已经建立了多个参与方的合作关系以及关键的基础设施，并可能成为第三次数字革命的重要加速因素。

世界银行的全球互联倡议就是一个相关的国际例子。该倡议的目标是到2020年实现15亿人联网。重要的国际人力、政治和金融资本，以及跨部门的参与方已经开始联合发展这一倡议。发起这项倡议需要人们建立长期信任，以确保每个人都能从基本的数字技术中获益。同样，美国前总统奥巴马的“联网计划”（“联网计划”是美国联邦通

信委员会E-Rate计划的一部分）希望，到2018年，将99%的美国学生与教室里的高速宽带连接起来，该计划目前已经取得了实质性的进展。政府和非政府组织的领导人可以在这些和其他数字接入计划的基础上，加快数制的访问，而不是从零开始。

针对制造空间的政府举措可加速提上日程。比如由议员比尔·福斯特提出的、在第1章中描述的美国国家“数制”工坊网络。在经常产生分歧的美国国会中，这是一个真正的被双方拥护的倡议，是一个罕见的例子。由于数字制造代表了新一代的制造方法，其能力仍然是地方性的，但知识共享是全球性的，它打破了全球连接和局部自给自足之间的错误二分法，让公众清楚地看到这些好处是至关重要的。当我们询问在白宫科技政策办公室从事教育和技术问题工作的库马尔·加格关于建立对数字和数制访问的支持时，他表达了通过访问使人们在情感上产生联系的重要性：

> 重要的是要讨论日常生活中真正引起人们注意的现实情况，比如孩子们坐在麦当劳的停车场中上网做家庭作业，因为他们在家里没有网络。我们要确保网络访问是一项有价值的公共投资，而且在技术成型的早期阶段，通过设计来构建包容性要容易得多，而以后要这么做就难得多了。

随着数字制造技术的发展，政府和领头的非政府组织的作用将不断演变。如今，这些机构可以利用它们独有的能力（召集、宣传、激励，在某些情况下，还有提供资金）来扩展社区和学校“数制”工坊的网络。它们还可以利用它们的能力，促进基础研究，以帮助数字制造路线图后期阶段的发展，目前，技术是速率限制因素。

慈善影响力投资

要增强“数制”和素养的普及性，就需要大量来自不同资源的投资。但是，这些投资并不能充分发挥作用。慈善和公众投资可以直接关注这些社会目标，但它们只是有效的催化剂，而不是可持续的支持源头。传统意义上，慈善及私人投资被视为两个极端，前者追求社会影响力，后者则注重市场。现在还有第三种投资，即通过市场机制创造社会影响力的社会企业型投资。以上这些对于迎接增强“数制”普及性和提高“数制”素养的挑战至关重要。我们在此节对上述问题进行细致的描述，以便为“数制”工坊社区进行更好的导航。

先说慈善投资，这将是一个可以让一个先驱慈善机构或一个新兴的家庭基金成为社区和学校“数制”工坊的“卡内基”的很好的机会。正如前面提到的，支持开放公共图书馆的卡内基模型与支持社区开放公共“数制”工坊是一个道理。尽管卡内基在他的商业实践中常常表现得很残酷，但在慈善事业方面他的贡献却非比寻常。在几十年的时间里，他资助了2 500多家图书馆。在卡内基之前，公共图书馆的概念并不普遍，后来，卡内基将在每一个城市和乡镇设立一个公共图书馆变成了一种自然的假设，而这个假设已经存在一个多世纪了。我们也期待一个有远见的慈善家可以在社区及学校建立“数制”工坊，以及产生一个类似的转变。

据《福布斯》估计，全球有1 800多名亿万富翁，总财富达7万亿美元。在这些亿万富翁中，有超过130人已经向盖茨－巴菲特承诺，将他们的大部分财富捐给社会。这些亿万富翁中的许多人在前两次数字革命中赚了大钱，他们对利用新技术造福社会很感兴趣。许多人希望支持教育和社区的发展及创业精神，并希望对社会产生持久的、积极的影响。为当地、地区、国家甚至全球的“数制”工坊提供资金，是宏观上可以满足的要求。本地、国家和国际的慈善事业对“数制”工坊最初10年的发展至关重要，这10年间的主要关注点就是建立并维持个体的“数制”工坊。现在是时候进行更雄心勃勃的慈善活动了。

随着公共社区和学校“数制”工坊数量的增加，我们可以从卡内基图书馆计划中学到更多的经验。密歇根大学信息学院前院长约翰·莱斯利·金提醒我们，在卡内基计划的早期阶段发生了什么：“在卡内基开始在北美各地建造图书馆几年后，它们却成批地失败了。”卡内基图书馆委员会要求专家找出原因。报告指出，失败原因是没有图书馆员，而没有图书馆员是因为没有图书馆学校！技术（材料和建筑）是必不可少的，但光靠技术本身是不够的。一定要有人能够构建和维护系统。在许多方面，图书馆科学学院（其中许多学院已更名为信息学院）是早期的赋能平台。它们提供赋能实践课程，为社会提供合格的指导者、教练和导师。

除了投资在数制产品和服务的私人资本和慈善投资外，还有另一种类型的投资，即有关键的可持续元素的社会影响力投资。社会影响力投资有多种：双重底线（金融、社会）投资，三重底线（金融、社会、环境）投资，混合价值投资，项目相关投资，任务相关投资和其他新兴模式。社会影响力投资确保了投资者和企业家之间的联盟，允许市场发展和社会影响共同进化。

如今，大多数社会影响力投资都聚焦于社区发展、环境，以及日益增长的劳动力发展和教育。尽管数字制造涉及上述所有领域，但该行业并没有受到大多数社会影响力投资者的关注，许多数制行业的社会企业家也没有意识到社会影响力投资领域的影响正在不断扩大。为了帮助解决这一问题，填补这一知识缺口，我们提供了关于社会影响力投资生态系统的高级概述。

社会影响力投资的范围从通常投资于从事社会影响事业的初创企业（一般在53美元到300万美元之间）的投资圈天使投资人，到奥米迪亚网络那样的社会影响风险投资基金（eBay的创始人皮埃尔·奥米迪亚发起），再到贝恩资本的3.9亿美元的双重影响基金（由前马萨诸塞州州长帕特里克领导）那样的具有社会影响力的私募基金，现在已有

成百上千种，很多都是由完全集中在社会影响力投资上的全球影响投资网络（GIIN）和诸如SoCap这样的会议牵头的。

越来越多的主要的慈善机构也在探索社会影响力投资工具，包括与任务相关的投资（MRI），基金会使用它们的捐赠基金（投资组合）来推进它们的慈善使命。福特基金会主席达伦·沃克最近宣布，在未来10年里，该基金会承诺为与任务相关的投资提供高达10亿美元的资金，其他许多公司也在效仿。与计划相关的投资（PRIs）是一种投资方法，基金会从它们的资助机构中寻求帮助，使它们的影响力投资组合多样化，同时吸引新的社会企业家，扩展或回收它们的资本。与计划相关的投资（PRIs）有各种各样的形式，从低利率贷款和贷款担保，到具有影响友好条款的股权投资。

由于许多基金会和社会影响力投资者具有较低的风险承受能力，尤其是在新领域运营时，而数制生态系统提供了一些风险较低的投资机会，比如与数制相关的服务。最重要却最缺乏吸引力的内容是建立一项工坊的物流服务，购买必要的保险和许可，维护和升级设备，确保安全，使所有其他工坊的日常运行符合操作要求。解决这一挑战也同样是培养一个赋能的数制生态系统的关键因素。在很长一段时间里，雪莉·拉斯特和她那不知疲倦的全球研究生团队将扮演这个角色，但他们并没有扩大自己的规模（事实上，学生毕业了）。这就是智造基金会成立的原因，它很适合成为这些服务的平台。因为已经有了对这些服务的需求，其他人可能认为这是一个低风险的投资机会，可以产生早期收益。

研究的迭代循环

赋能的数制生态系统的核心是持续的技术进步，确保拉斯定律保持其指数增长轨迹。尼尔和他的同事在比特和原子中心进行的基础研究，以及世界各地类似研究机构的工作，对实现这一目标至关重要。

延续这一规律需要扩大对基础数字制造研究的支持，在基础研究和应用研究之间建立一个强大的迭代循环。要确保这项技术以加速“数制”工坊的普及，就需要将科学和社会科学研究结合起来。

在大学、政府和个人研究实验室的现有结构和文化下，使高等教育的各参与方达成一致是一个相当大的挑战。正如我们在第4章中提到的，在这些领域中存在激励因素，以及推动保守行为、削弱集体行为的因素。尼尔必须建立比特和原子中心来做他所做的事情——不适合任何既定的学术范畴。在自然科学研究和社会科学研究之间也存在着分歧。所以，在研究社区内部（或跨研究社区）建立一个更健全的生态系统，促进数据、工具和模型的共享，促进跨学科合作，并开创新的组织和机构模式，就可以创造一个新机遇。

美国教育部STEM教育前执行主任、前美国国防部高级研究计划署项目官员拉斯·希林回应：“研究处理固有的噪声数据，如行为科学和教育，可以极大限度地从迭代的研究实践中获益，消除基础和应用研究之间的脱节。”正如拉斯所言，基础研究和应用研究之间的脱节是社会科学特有的现象。弥合这一鸿沟——将路径观察者和路径创造者的工作以及科学家和社会科学家的工作结合起来，将在非常多样化的生态系统中实现协同作用，从而影响研究的重点，以及研究如何将其转化为早期产品和服务。

社会企业家

对于缺乏政府、非政府组织和主要的慈善机构资源的数制社会企业家来说，有一个很好的机会，可以通过各种平台、工具和实践来传播思想。目前，数制生态系统中存在着巨大的鸿沟，理想情况下，这些鸿沟可以由专注于扩大影响力的企业家来填补，这些企业家可以利用授权与被授权的分布式的网络、坚定的参与方以及数百万参与者进行填补。

我们在这本书中介绍的许多数制先驱（世界上任何其他地方的人）都在研究那些有潜力成为全球变革生态系统的项目。当这些具有相同核心目标的生态系统出现并与看似不可逾越的挑战（如通用数制访问）相结合时，就会变得不那么势不可当。例如，如果阿莉西娅·加姆勒薇兹女士能够开发一个有效的平台，获得相关的工具，提供资源，共享当地“数制”工坊资料，而约各什·库尔卡尼能够做同样的事来发展分布式乡村数制能力，迪娜·埃尔－赞法利做同样的事来发展数制素养，那么这些生态系统中的每一个都能显著加速彼此的发展。这些数制先驱可能选择创建新的生态系统，也可能选择适应和扩展其他数制先驱开发的平台。

社会企业家有很多机会利用新兴技术，这些技术可以成为一个赋能的数制生态系统的强大元素。例如，增强现实技术以及新的位置感知技术。目前，许多分布式的数制指导是通过视频会议完成的，而视频会议存在容量问题。随着时间的推移，“增强现实”解决方案可以作为更具伸缩性的模型。同样，在虚拟现实中创建模拟的“数制”工坊环境也有很大的价值，在虚拟现实中，没有数制访问权限的个人至少可以获得虚拟体验。新的位置感知技术也可以在分层的数制系统中发挥关键作用。多层次和分布式的工坊阵列可以集成到一个感知位置的应用程序中，该应用程序可以跟踪数制能力，并在任何给定的位置上开放访问权限。该应用还可以提供关于如何启动一个“数制”工坊、一个支持数制的创客空间，或一个可以随时间增长的单机设备数制能力的建议。这将是一个有利于生态系统的宝贵补充。

两个个例

促进培育赋能的数制生态系统的两个具体方面可以用来说明智造者联盟的平台和数制素养。这两个领域可以映射到乔尔和艾伦正在做

的事情。这些个人的例子只是众多相关例子中的两个，它们说明了来自不同学科的人如何为数制生态系统创造价值。

在这本书中，我们强调了许多独立但相互依赖的参与方需要在共同利益的基础上实现目标，同时兼顾可能发生冲突的利益。为了帮助实现这个联合过程，乔尔有一个目标，那就是在复杂的生态系统中为参与者的联合开发一个平台。他和他的同事开发的可视化及其数据分析朝着这个方向迈出了一步。这样的平台将支持在适当尺度级别上对不同类型的参与方进行规范，以及识别相关的利益（什么处于险境）。然后，这样的平台将支持所需数据的收集，使平台的架构师能够看到哪里有一致或不一致的点。这将有利于我们得到易到手的果实，避免死角，并为复杂地形提供导航。

一个很好的例子是，这样的平台可以成为一个强大的加速器，它围绕着参与方的联合来支持共同的标准和协议。这反过来又能在目前分散的开源和专有硬件和软件生态系统中实现更好的协作。由于技术是不断发展的，所以对一致平台的使用将是一个持续的过程。另一个关键的合作点存在于数制访问的公众投资者和私人投资者之间。这不仅可以避免重复，还可以对已被证明的影响（影响投资者的必备条件）和投资生态系统的其他促进因素进行横向调整。实际上，一个参与方联盟平台可以被看作一个平台倍增器——使其他需要的平台的开发更快、更有效。

这样一个平台可以利用的另一个领域是帮助不同的参与方实现不同层次的抱负和能力。例如，一些参与方致力于成为系统架构师，而另一些人可能愿意也可能不愿意承担这个角色。相反，他们可能更喜欢在别人开发的平台上成为被授权和授权的中间者的一部分，而这个平台能实现他们的愿望。拥有平台、工具和过程的生态系统使代理成为可能，但不是每个人都希望在给定的生态系统中拥有同样程度的代理。他们中的一些将成为架构师，一些将成为被授权和授权的中间人员，还有一些将成为分布式贡献者。如果有一个支持参与者联盟的平

台，那么它实际上就是一个元平台，加速了赋能的生态系统中许多其他平台的开发。

在提高数制素养方面，数制学术为水平高的学生的智造能力进行了充分优化。在不同年龄、不同技能、不同地区的人群中，需要多种形式的正式和非正式的学习轨迹来培养良好的素养。这些课程需要以实践、基于项目的学习为基础。在这一领域，有一些见解认为，新兴的基础学习社区受益于那些在过去10年中开发了游戏基础学习平台、工具和实践的人。这是艾伦研究的领域。

基于游戏和基于场景的学习都涉及人们通过实践来学习的项目。当这两种方法得到有效实施时，就可以为高度协作和被动驱动的学习提供环境，并将其连接到实际的工具和实践社区。数字制造是困难的，但是坚持不懈的努力仍然可以吸引人，让人充满力量，并且令人振奋——这是一个好的游戏设计面临的核心挑战。玩家们正在不断地学习掌握复杂的、技术介导的体验，以实现他们个人投资的目标。好的游戏提供了一种微妙的平衡和奖励来驱动玩家的深层次参与，使玩家能够按照他们自己的步伐前进，及时获得关键的知识（而不是仅仅为了以防万一），并根据游戏、同行、导师和社区的大量反馈进行迭代。基于游戏的学习和数制素养之间的这种联系是电子线媒体、智造基金会与比特和原子中心之间合作关系的核心。

越来越多的个体和组织开创了“虚构学习”的新方法，通过开发平台、工具和实践来共享项目想法、教师专业发展评估的所有内容的机会也越来越多。在这种情况下，评估更多的是关于构建一个基于个人和团队的项目组合，而不是与真实环境分离的高风险测试。正如尼尔在第1章中强调的，数制投资组合已经成为数制学术认证过程的一个组成部分。这可能适用于所有教育水平。

嵌入式生态系统

与参与方联盟一样，培育生态系统不是一项一次性的任务，而是一项持续的成就。此外，它还涉及多个嵌入的生态系统，各个生态系统在某些方面是独立的，在某些方面是相互依赖的。此外，它们作为一个分层的系统来运行，在这个系统中，支持平台、工具和实践以类似但不同的方式为个体、社区和全局系统工作。

虽然需要自上而下的促进措施，但我们也强调在模型中培育生态系统，因为这是使社会系统跟上技术变革的步伐的关键。这些自下而上的、关键的、中间的变更过程（使用平台、工具和过程）使社会和技术的共同进化成为可能。此外，还可以考虑其他社会技术生态系统，如人工智能、机器人、生物医学和自动交通，它们与更广泛的社会和技术的共同进化有关。

正如在秘鲁的亚马孙丛林中长大、在亚马孙地区流动的专门研究区域生物多样性的“数制”工坊的建筑师贝诺·华雷斯所观察到的，的确，我们甚至可以想象与生态系统的共同进化。该工坊正致力于用丛林中的营养物质和检测水污染的传感器来“打印”健康饮食的可食用部分。正如贝诺所指出的，“流动的‘数制’工坊与城市相比，有着不同的丛林生活方式。”在丛林中你的体重会发生变化。你觉得自己与这个星球的联系更紧密了。他补充说，亚马孙流动“数制”工坊是“技术和当地文化的结合，它创造了一种新的潮流。”

共同进化：技术和社会

第三次数字革命不会孤立地发生。路线图将在一个经济、环境、人口和社会政治变化并存的复杂的世界中展开。它将影响更广泛的经济趋势，如技术失业和技术全球化、人口老龄化和大规模移民等人口趋势、气候变化和资源枯竭等环境趋势，以及两极分化和僵局等政治动态。与自然界的共同进化过程不同，技术和社会的共同进化涉及选择和行动。这不仅是驾驭波动的问题，还涉及共同进化的方向和影响

——这并不容易做到。共同进化在一个持续的变化周期中利用了模型的三个部分——预测、结盟和培养。

在经济变化中，最紧迫的共同进化动力来自每个人智造（几乎）任何东西的能力，以及与加速发展的技术相关的不断增加的工作岗位。在2015年出版的《机器人崛起》（*Rise of the Robots*）一书中，马丁·福特指出，不断进步的人工智能、自动化和机器人技术将越来越多地取代蓝领和白领的工作，而工作岗位的减少将反过来减少税收收入。这种“飙升的技术失业”、收入和财富不平等的加剧，再加上支持人口老龄化和应对气候变化影响的成本不断上升，可能会导致社会恶性循环。在这种情况下，人性中最坏的倾向将被流离失所者的绝望和愤怒所放大。

如果这条轨迹得以展开，第三次数字革命将成为解决方案的核心部分。福特和其他许多观察人士一样，认为科技行业正在加速裁员，他建议采用某种形式的基本收入保障，并精心调整激励措施。尽管这一建议引起了严肃的政策辩论，但这一解决方案面临的最大问题之一是解决失业带来的尊严和目标丧失问题。然而，如果一个基本的收入保障也伴随着数字制造培训，使人们越来越多地生产他们消费的东西，那么经济和社会心理都有机会受益。它不是“后稀缺经济”，而是“后薪酬经济”。

在这种情况下，这是一场比赛。数字制造要赶上和超越技术进步的步伐并不容易。然而，如果我们将其变为现实，我们就能看到数字制造技术的能力有10亿倍的提升，那么就可能赢得这场比赛。这需要人们在一个存在严重分歧的社会中做出一致的努力。它将改变游戏规则——社会的制度，这一改变既可怕又充满力量。在这8个理想的场景中，我们看到了城乡自给自足、社区能力和学习、使用当地材料解决当地问题、文化转型和创业创新的结合。总之，这些场景与自动化、失业和购买力下降形成了对比。

正如第4章中介绍的，1984年，迈克尔·皮奥里和查克·萨贝尔在“第二产业分化”中的分析表明：在过去的产业分化中，先前的制度并没有消失。随着大规模生产的兴起，工艺生产并没有消失；大规模生产的机器需要它，但它已不再是主导模式。如果分布式数字制造确实以指数级的速度增长，那么跨国公司、服务运营、金融市场和创业计划的作用将继续下去，在分布式数字制造的背景下越来越明显。

除了要面对第三次数字革命对经济的影响之外，人类还需要与自然环境共同进化。对第三次数字革命影响最大的是耗材和供应链。如果在“数制”工坊中出现了使用当地环保材料的平台，那么将会有重大的环境效益。如果不是这样，那么“数制”工坊的指数增长可能是一场环境的噩梦。让数十亿的人对不环保的塑料（或其他不可再生的、稀缺的资源）进行修补和反复使用肯定是不可持续的。为了我们这个星球的健康，我们需要努力建立一个生态系统，使用环保的消耗品和可重复使用的数字材料。

气候变化（和政治冲突的）的另一个新含义是流离失所者和难民的日益严重的危机。联合国难民事务高级专员估计，世界各地有6 500万难民、寻求庇护者和国内流离失所者。难民营和临时社区的数字制作可以为建立技术技能和培养协作文化提供多种好处——模块化的、快速部署的庇护所、家具和其他基本需求。

在理想的设想中，布莱尔·埃文斯在底特律追求自给自足的模式很容易转移到难民营，在那里，设计和建造低成本住房的能力，可能成为难民和流离失所者的救生能力。此外，在难民营中培养这些能力的个体，也许能建立投资组合，提交给技能认证平台，给他们返回家园或移民到新社区的机会。同样，一些沿海社区的人因气候变化而流离失所，这为他们在内陆重建社区的想法提供了独特的机会。我们知道，离开家园会让人感到痛苦，但引入机会以提高能力、重新开始，可能会缓解这种感受。在全球人口不断增长的时代，数字制造可以成为技能、能力和可持续发展的新支柱。

随着人口结构的变化，老龄化在某些方面已经成为数字制造世界的一部分。许多“数制”工坊的年龄分布是各种各样的——从小孩到退休人员。因为许多退休人员正在寻找丰富的、具有社交性的和有意义的活动，“数制”工坊可以为这些拥有终身创客技能的老年人提供做导师的机会。（相反，年轻人可以在新数字技术方面培养退休人员的重要技能）我们已经看到许多数制项目集中在移动设备、物理援助和其他项目上，这些项目对“婴儿潮”一代来说是有用的。考虑到美国退休人员的数量预计将从2012年的4 310万人增加到2050年的8 370万人，人们迫切需要通过观察、调整和培养来实现共同进化。

考虑一下在上次经济衰退中遭受重创的一个经济部门和相关社区会发生什么。从2007年到2010年，美国汽车行业解雇了20多万工人，其中许多人提前退休。他们中的大多数人都知道如何使用机床。这是一个很好的机会，让他们在各自的社区新设立的“数制”工坊中工作。然后，我们可以策划新的操作方法（使用支持平台），一次完成三件事。首先，对于退休人员来说，我们将提供一种有意义的、丰富的后职业经历——这将重新定义“退休”的意义，而这种情况正在社会上广泛发生。其次，这些人可以应用梅尔·金的“学习教育，教学”的过程。金让高中生在工坊里学习如何操作，这样他们就可以教初中生做同样的事情。对于退休人员，类似的做法将涉及代际共同创造和对话。第三，衰退的经济部门和退休人员不再是社会的负担，而是社区振兴的中心。共同进化实际上是跨越几代人的社会的再生。

政治上的两极分化和僵局，以及更广泛的社会问题在前两次数字革命中被放大了。造成这种两极分化的一个原因是“回音室”，这是一种社交媒体，人们可以在这里与和自己同样的人交谈，因此，人们与想法不同的人的互动越来越少。与虚拟社区不同的是，在“数制”工坊的项目中，经常可以看到背景和身份非常不同的人并肩工作。在菲亚特克莱斯勒的理想场景中，我们引入了“边界对象”的社会科学概念——它位于不同社区的交汇处，并使跨越边界的桥梁成为可能。

我们说过，“数制”工坊是边界事物，而在工坊进行的项目本身也是边界事物——跨社区的大型桥梁中的小桥梁。

通过观察“数制”工坊和数制项目之间的桥梁作用，我们可以看到深受社会两极分化困扰的人们之间的合作机会。然后，他们可以通过考虑社区的多样性来培养“数制”工坊的能力。然而，共同进化是复杂的。总部位于英国的国际扶贫组织乐施会（Oxfam）在2017年发表了一份题为“99%民众的经济”的报告。报告指出，“从尼日利亚到孟加拉国，从英国到巴西，人们厌倦了被政治领导人忽视的感觉，数百万人正动员起来推动变革。”该报告还说，“在过去30年里，每10人中就有7人生活在不平等日益加剧的国家。”在那些收入不平等日益加剧的环境中，“数制”工坊可以提供一些可供选择的经济途径，以缓解这些分歧；或者最终为那些更幸运的人服务，从而加深分歧。

另一方面，在一些国家，中产阶级正在壮大，收入不平等正在缩小。在那里，新一代人有时是第一次看到“下一代人做得更好”的前景。在这些情况下，数字制造正经历着一种不同的潮流。关键是，要在任何有合作精神的人进入社会的地方培育共同进化的能力，使人们从把“‘数制’工坊考验”当作化解某些社会分裂的紧张局势的解药，到建立一个“数制”工坊，以加强其他人的阶级利益。

在所有这些共同进化的场景中，过程是相同的。我们从预测、调整和培育使生态系统能够在运动中共同进化的能力入手，内容包括数字制造技术与相关社会系统之间的有限的共同进化，以及数字制造与广泛社会挑战之间更广泛的共同进化。数字制造的路线图将不仅局限于它自己的工坊世界，还将影响经济环境、人口和社会政治趋势。同时，路线图也会受到它们的影响。

组织与机构自建

如果机器能够制造机器，那么组织和机构很可能需要以类似的方式共同进化，这就提出了一个有趣的前景，即组织建立组织，机构建立机构。如第3章所示，可靠性、模块化、局域性和还原性的概念使机器能够制造机器（并进一步以指数级速度扩展到数字材料及可编程材料的领域）。我们可以把这些概念应用到社会系统中吗？

从科学到社会科学的类比，有一段复杂的历史，但在这里，它不仅仅是一种平行的措辞。当技术和市场从手工艺转向大规模生产时，组织和机构也发生了变化。如果技术和市场以数字制造为导向，也会发生类似的共同进化。但这并不是关于技术的决定性论证。我们可以选择如何实现这一点，在这个过程中存在一些限制。毕竟，社会系统中的人具有能动性（以及独特的人类偏见和缺陷）。但是，如果该机构能被理解并受到重视，这种缺陷也可能是一个有益的特征。

数字科学的一个特征就是模块化，相对而言，我们对模块的装卸并不陌生。第二次数字革命的组织已经在以团队为基础的工作系统和以项目为基础的活动中形成，在这些活动中，人们会相对轻松地进行组装和拆卸。我们还看得到，模块化工作空间的日益流行也体现了这种模块化。在模块化工作空间中，初创企业可以随着规模的变化而轻松租用空间，并磨炼自己的专注力和策略。这个装配和拆卸的过程确实需要一套专门的协调技能、敏捷的项目管理者，以在变更过程中减少危害。同样，各种各样的合资企业和其他形式的合作组织也允许更宏观的装配和拆卸形式。在社会的机构一级，越来越多地使用财团和其他多个参与方的倡议，可以作为促进变革的催化剂，而这种变革在现有的机构体制范围中是不会发生的。这些更灵活的操作并不能完全匹配信息的组装和分解，但是它们确实有一些相似的属性。此外，在组织和机构中进行组装和拆卸的经验是第三次数字革命的关键，随着变革步伐的加快，可能需要更大规模的组装和拆卸。

模块化对领导者的吸引力是显而易见的，尽管对于那些可能在刚刚被分解的模块中工作的人来说并不总是如此。经济学家、美国劳工

部工资和工时部门前负责人戴维·韦尔在书中谈到了工作的“裂缝化”——曾经被分解为各种工作的职业，现在被分解为项目、任务和合同关系。他证明了工作的阴暗面变得更加模块化：承担风险的是员工，而不是决策者。在模块化程度最差的情况下，创建组织的组织，以及创建组织的机构可以以更大的代价实现这些目标。在最好的情况下，这种灵活性可以加速创新。蒙德拉贡合作社这个例子就说明了，当有一个强有力的基础支持流离失所的工人时，“模块化”是如何工作的，因为如果他们有新的商业建议，他们将再次接受培训，甚至获得资金支持。

关于可靠性（错误修正）的类比可能没有模块化那么直观。然而，在许多方面，错误修正是前两次数字革命的秘密武器——允许大规模通信和计算，而不会降低性能。尼尔的见解是，这种特性同样适用于数字制造。在某种程度上，目前存在一些类似的组织和机构。例如，管理中的制衡原则是机构的错误修正形式；在工业领域，报道制造业和航空公司的安全事故是错误修正形式——在潜在问题发生之前就识别出来。同样的原则也适用于质量改进。这类报道并不容易生成，因为人们害怕受到指责，甚至是简单的尴尬。当这种报道确实发生并在整个企业或系统（如航空运输）中扩展时，人们开始创建一种增加透明度、减少责备和分担责任的共享文化。这类似于通过错误修正来实现数字系统的扩展。第三次数字革命将需要在社会方面采取更广泛、更有力的错误修正形式，从而使其与加速发展的技术共同进化。

装配工和煎饼早餐

前两次数字革命从根本上改变了我们的生活、学习、工作和娱乐方式。比特改变了原子的命运。第三次数字革命又使我们上了一个台阶，因为我们可以跨越式地操纵、组合并共享比特及原子了。随着技术的加速进步，我们将能够设计现实，从食物到家具，从工具到玩具，从工艺品到电脑，从人类社会到机构。

在本书中，我们探讨了这种加速发展的能力带来的巨大优势和风险。但是还存在一个尚未解决的问题，那就是：这些强大的新技术会给我们带来更多的欢乐吗？

工业革命改善了许多人的生活，但它也创造了一种让人窒息的大规模生产的文化。正如现代社会学创始人之一马克斯·韦伯所写的那样："有朝一日世界中充斥的全是一个个小小的齿轮，人们紧紧抓住小工作并努力提取下一个大一些的工作机会而没有全局观……这种想法是可怕的……对官僚主义的热情……足以使人绝望。"

同样，前两次数字革命改善了许多人的生活，但也让许多人感到过度失落而不知所措，渴望回归到一个更简单的、没有动荡的旧时代的生活。从"无数字假期"到"去数字化排毒计划"，再到像"全国断网"这样的活动，人们一直觉得自己的生活已经被数字技术饱和了。我们甚至看到了原始的模拟产品和服务的复苏，如黑胶唱片、传统棋盘游戏和独立书店。雪莉·特克尔在她2011年出版的《群体性孤独》（*Alone Together*）一书中描述了这一现象："科技为我们提供了面对面交流的替代品。"我们可以在网络设备上利用机器人和全部由机器调节的关系。当我们用短信、电子邮件、短信和推特等方式进行交流时，科技重新划分了亲密和孤独之间的界限……我们把自己重新塑造成网络人物，给自己新的身份、家庭、工作和爱情。然而，在虚拟社区中，我们可能会感到非常孤独。

当我们一头扎进一个越来越引人注目的虚拟世界，以及由杰出的设计师和行为科学家开发的让我们盯着屏幕的其他应用程序时，在比特世界中的时间和原子世界中的时间之间找到平衡将变得越来越困

难。第三次数字革命可能有助于重新建立这种平衡，因为这次数字革命可以帮助我们利用比特在原子世界中花费更多时间。

制造是一种深度人性化的行为。每个人都需要感觉到自己的行为是重要的，所做的事对自己、家人和世界都有影响。我们制作得越多，我们的自我价值感就越强。社区“数制”工坊还利用了人类的深层愿望：建立联系，合作解决问题，共同实现目标。正如多弗·塞德曼在托马斯·弗里德曼的《谢谢你迟到》（*Thank You for Being Late*）中所说：“我们建立深厚关系的能力——爱、关心、希望、信任和建立基于共享价值观的社区，是我们人类最独特的能力之一。”这是我们区别于自然和机器的最重要的一点。

在“数制”工坊内部和跨工坊人员之间培养非正式联系是数制文化的核心。哈肯·卡尔森讲述了他和尼尔一起设计位于阿尔卑斯山的林根“数制”工坊的故事。哈肯说，尼尔告诉他，“你不可能在工坊里有一个厨房。”多年后，当尼尔看到厨房已经成为工坊的一部分时，尼尔总结道：“如果没有厨房，你就不可能有一个工坊。”

当我们开始写这本书的时候，我们的前提是数字技术正在指数级加速，但是个人、组织和机构是线性发展的。也许这是一个特性，而不是一个问题。当然，我们需要共同进化，以确保我们塑造的技术不会以让我们后悔的方式塑造。但我们也必须保留人类的本质——赋予生命意义。

88岁的梅尔·金是波士顿SETC“数制”工坊的创始人，他仍然倡导设计和智造（几乎）万物的能力都是革命性的。他总结道：“‘数制’工坊势在必行。这是一种生活方式。这是爱的方式。制造和创造……这是人类的底线。”

最后，对那些想抓住第三次数字革命精髓的人，我们送给他们受访数制先驱的总结性陈辞，希望他们不忘初心、砥砺前行，这大概我们能给出的最好的建议了。

梦想、制造和分享。

——荷兰

全面工业化之路。

——肯尼亚

如果你能想到，就能做到。

——西班牙

实现你的想法。

——哥伦比亚

被误解的天才儿童。

——印度

糟糕的混乱。

——阿拉伯联合酋长国

彻底改变个体力量的空间。

——美国

“数制”工坊应当消融在社会中。

——智利

让不可能的事变成可能。

——巴拿马

改变世界的工具。

——法国

实际上，这些数制先驱正在设计现实。而且很快，每个人都能做到这一点。

（本章由艾伦、乔尔撰写）

---

[1] 泰·塔利姆教育法：知识和工作不分离原则。圣雄甘地基于这一教学原则，推出了同名教育课程。——编者注

[2] Model D，参见http://www.modeldmedia.com。——编者注

[3] 伊斯沃尔，菲亚特（FIAT）集团所属的具有独立法人，专门从事企业咨询和技术的、职业、技能、管理培训的国际公司。伊斯沃尔公司的前身是菲亚特技术学校，成立于1937年，长期从事菲亚特公司的技术工人和管理人员的内部培训。——译者注

[4] 世界十大汽车公司之一。菲亚特汽车公司于1899年7月始建于意大利都灵，创始人是乔瓦尼·阿涅利。它是世界上第一个生产微型汽车的生产厂家。公司全称是意大利都灵汽车制造厂，菲亚特（FIAT）是该公司缩写的译音，FIAT也是该公司产品的商标。2014年10月12日，菲亚特与其美国公司克莱斯勒合并的决议生效。——译者注

[5] 网农：指网络工作者，由“码农”这个词引申而来。——译者注

[6] 张永安，曾与尼尔·格申斐尔德在《科学》杂志上发表一篇文章，仔细描述了用3D打印技术打印一架飞机的方法。——译者注

[7] 库尔特·勒温，德裔美国心理学家，拓扑心理学的创始人，实验社会心理学的先驱。格式塔心理学的后期代表人，传播学的奠基人之一。——译者注

结论

今年是2017年。汽油每加仑2.38美元。德雷克的最新专辑《*More Life*》在榜单上排名第一。特朗普当选了美国总统搬进白宫，马克龙当选了法国总统搬进爱丽舍宫，而英国退欧已成为既定事实。《星球大战：原力觉醒》（*Star Wars: The Force Awakens*）在票房上遥遥领先。全新的ShopBot Desktop MAX首发价为9 857美元。Epilog推出全新的Fusion激光雕刻机，售价7 995美元。新款新都（Sindoh）3DWOX 3D打印机售价1 299美元。

你坐在星巴克享受大杯半糖无泡焦糖玛奇朵。你碰巧坐在尼尔、乔尔和艾伦·格申斐尔德旁边，他们正在就他们的新书《设计现实》进行热烈的辩论。起初他们辩论的想法似乎有些牵强。数字制造性能提高了10亿倍？任何人都可以智造（几乎）万物？个人智造？你表达了怀疑态度。如果你知道该看向哪里，他们会给你提示，然后告诉你朝哪儿看。

你学得越多，你就越意识到这确实是你需要认真对待的事情。确切的时间可能不清楚，但你意识到你刚刚有一个机会——看到即将到来的完全不同的未来。现在你必须决定如何处理这门学问。为了帮助完成该过程，尼尔、艾伦和乔尔提供了以下操作步骤。

目标和建议

第三次数字革命不会发生在我们身上，但我们会让它发生。本着这种精神，我们提供了一系列目标和建议，所有目标和建议均建立在路线图和预测转型模型之上。我们首先提出科学和社会科学的建议，然后针对路线图的每个阶段提出建议。

世界上大多数人并不熟悉或只是模糊地了解数字制造。对于许多人来说，数字制造就是3D打印。我们通过认识了解到：数字制造以前两次数字革命为基础，并推动了前两次数字革命的发展，我们将在第三次数字革命中加速进行必要的基础研究，帮助人们在短期内理解数字制造的全部技术以及新兴的制造科学。鉴于此，我们的第一个总体社会目标如下：

致力于推进新兴的数字制造科学。

这个目标可能听起来很简单，但今天，相关工作分散在多个制度边界。为实现这一目标，我们提供了许多可行性建议，包括科学家和研究人员的这个建议：

将数字通信和计算的发展与数字制造的发展相结合。

本书的核心论点是：技术和社会系统需要主动性，共同发展，并在早期过程中产生最大的影响。因此，我们给社会科学家们提出这样的建议：

认识并支持社会科学通过理论、工具和方法的进步积极帮助实现第三次数字革命的能力。

最终，为了第三次数字革命的技术和社会因素有效的共同发展，每个社区都必须接受这一过程。为此，我们提出以下建议：

利用正式和非正式的交流机会，将社会科学家、技术专家、讲故事的人和其他参与方聚集在一起，共同创造第三次数字革命的综合理念。

鉴于拉斯定律不是自然法则，而是由技术创新和社会建构力量共同驱动的，因此，我们也建议“数制”工坊的所有参与方：

培养共同的紧迫感，集中精力并充分利用资源，以推动数字制造性能和覆盖范围的持续的指数级增长。

考虑到这些总体目标和可行性建议，我们现在转向路线图的4个阶段。大多数目标和建议都针对第一阶段——社区制造，这可以为接下来的活动打好基础。鉴于有较多的可行性建议，所以接下来需要进行大量的基础性工作。

社区制造

“数制”工坊的技术现在就在这里，我们已经在过去的10年发展了社会系统以有效促进“数制”工坊的运行。这是第三次数字革命的首要目标，从社区“数制”工坊阶段开始：

将通用“数制”工坊纳入全球优先事项。

为了帮助实现这一目标，我们针对整个“数制”工坊生态系统中的特殊参与方提供以下建议：

投资者

将数字制造建设成领先的、具有社会影响力的投资部门。

吸引“数制”工坊生态系统的领导者，为投资者提供关于社会影响和财务收益最大化的建议。

建立追踪“数制”工坊的经济、社会影响的机制。

政府领导人

调研当地的“数制”工坊对社区的影响（并分享您的调研结果）。

将数字制造技术与现有的数字鸿沟计划相结合。

涉足那些与慈善事业和工业无关的领域，打破自身局限性，解决最困难的问题，以创造公共价值，减少对社会的危害。

慈善家

把赠款、任务和项目相关投资的重点放在关键的基本问题上，并支持“数制”工坊生态系统。

在公共基金和私人资助商之间建立协同效应。

寻找机遇，成为“数制”工坊界的“卡内基”。

社区领袖

将本地“数制”工坊同商业组织、学校和大学，以及基于社区的组织连接起来。

让每个城镇和学校拥有“数制”工坊成为基本假设，就像拥有公共图书馆和学校图书馆一样。

跨部门的参与方

跟踪记录并构建“数制”工坊和个人制作的案例，使个人和社区能够变得更加自给自足。

将智造者和创客先锋（涉及教育、经济发展、慈善、食品和安全领域）整合到具有社会影响力的组织中。

对于当前的“数制”工坊先锋和企业家，我们提供以下建议，重点关注共同标准和共享原则：

“数制”工坊、创客空间、骇客空间和其他类似的工坊应该基于共同标准和共享原则进行合作，同时要有各自独立的身份和方法。

数字制造设备和软件的制造商也应该在共同标准和共享原则的基础上进行合作，同时也要维持各自独立的品牌和方法。

包括共享的技术、组织和机构在内的为数字制造设立的基础设施应该得到支持。

我们还提供这些建议，重点关注社区内和社区之间的协作和知识共享：

提供在数字制造生态系统内部传播有效实践的机制，定期记录参与方之间的相同点和不同点。

不断完善“数制”工坊宪章，使其更好地为管理、运营和发展服务——惠及参与方和数制生态系统。

把协作和知识共享建立在关于获取、素养、生态系统和社会风险的小协议的基础上。

我们围绕创建开放、健全的市场生态系统提出以下建议：

培养有效、开放、透明的，与“数制”工坊文化保持一致的市场和分销平台。

不断寻找并支持有前景的必备产品和服务，为该行业提供资金。

利用本地的、可重复使用的环保材料，开发可供理解、采购和构建的共享平台。

设计与数字制造相关的工作，使其可以公平处理争议并提供建设性解决办法。

为那些利益受到威胁的人制订计划，以保证数字制造的民主化。

当然，人是“数制”工坊生态系统的核心。以下是对于提高人的能力的建议：

明确、维持并扩大社区“数制”工坊及相关设施内部的导师和领导者数量。

认识到在开发数字制造技术方面需要个人和集体的共同投入。

并向哈肯致敬：

每个“数制”工坊都应该有一个厨房。

个人制造

路线图的第二阶段堪比20世纪70年代和80年代的电脑套件。在个人制造方面，人们担心这些数百万的早期个人制造者将成为一个幸运

的、经过挑选的群体，因为其中绝大多数人都留下来了。考虑到这一点，我们建议：

促进家庭和小型企业数字制造的发展，开发既经济又直观的工具。

为了实现这一目标，我们需要研究人员和“数制”工坊的先锋开发一些案例，说明个人如何在家庭和小型企业中拥有不同级别的“数制”工坊能力，而不仅仅局限于使用3D打印机。为帮助实现此目标，我们提供以下建议：

与人类学家和其他社会观察员合作，关注小型企业和家庭中的早期采用者，记录新的使用模式并识别潜在的“速率限制器”和“加速器”。

制订协议和标准，使难以在当地生产的工业生产模块化组件集成到个人制造中。

跟踪新兴经济模式（易货贸易、商业化、开源、区块链等），以塑造既能可持续发展又能实现个人价值的数字制造实践。

拥护组织不断地创造组织、机构不断地创造机构的发展趋势，同时，与不断地创造机器的机器共同成长。

无处不在的制造

第三阶段是通用制造，“数制”工坊的数量从100万到10亿。第四阶段是无处不在的数字制造，“数制”工坊的数量达到一万亿，数字不仅能智造（几乎）万物，还能智造（几乎）一切。对于注重技术的读者来说，最后两个阶段非常抽象。准确地想象人们如何用数字材料

制作东西是很难的，更具挑战性的是理解人们在生产过程中处于什么位置。最终，应该使用集成的人文价值创建驱动这些技术的算法。因此，面对更广泛的受众群体，需要推进以下目标：

借助大众媒体的力量，通过数字和可编程材料创造出诱人的、令人向往的愿景。将透明度融入先进的数字制造技术和算法。

然而，即使有最好的想法，也会面临意想不到的后果。风险缓解机制至关重要。现在，反补贴做法和既得利益减少，这些问题得到了最好的解决。

在数字和可编程材料开发过程中纳入风险缓解方法，这是最容易做到的。

我们总结的这些建议都是建立在其他建议之上的——社区制造所需的生态系统对个人制造仍然很重要，个人制造的模块化特性为无处不在的制造阶段奠定了基础。所有阶段都依赖技术和社会系统的持续创新而发展。其中许多建议实施起来并不容易。但是，如果我们积极地付出努力并以平台和实践作为支撑，我们就可以共同实现这些目标。我们需要塑造（并不断重塑）我们所在的组织和机构。我们借用了“阿波罗9号”宇航员拉塞尔·施威卡特的一句话：我们不是第三次数字革命路线图的乘客，我们是船员。

（结论由尼尔、艾伦、乔尔撰写）

后记



本书的作者们。格拉迪丝和沃尔特·格申斐尔德提供

当我们第一次提出写这本书时，我们的经纪人约翰·布罗克曼的反应是：“好主意！我们将失去三位作者。”当这种想法传到我们未来的编辑T. J. 凯莱赫那里时，他说了同样的话。我们向他们解释，三方合作不是偶然的，这对我们想讲的故事至关重要。

我们去看望母亲，跟母亲聊天，那时母亲的健康状况正在下降，《设计现实》的创作想法随之诞生了。她不再能够加入我们的讨论了，但我们的声音令她很高兴。在我们讨论的时候，我们意识到我们的职业方向完全不同，但正在数字制造的交叉点上相互碰撞。将我们的观察结果一字一句编在书里，理论上讲，都源于爱。

事实证明，写这本书更像是劳动，它是我们对第三次数字革命所面临的所有挑战和机遇的缩影的概述。在我们尝试到失败之前，我们从未意识到我们每个人的思维过程有多么不同，我们甚至不能写出一个简单的叙事线索。我们开始了第二次尝试，将我们的每一个想法分成一系列的独立片段，成功地把想法分散。T. J. 的评价是：“当你在错误的道路上行驶时，转身回去永远都不会太晚。”

只有在那时，我们才意识到，将我们的观点和方法统一起来，对我们而言是一项巨大的挑战。事实上，当我们告诉别人我们在一起写一本书时，最常见的反应是“你们没有互相厮杀？”我们也会听到一些关于兄弟之间很难进行合作的故事。但最终，我们持有的不同观点相互碰撞，我们面临的挑战都为我们综合来自科学、社会科学和人文科学的观察结果做好了铺垫。

尼尔的职业生涯始于贝尔实验室。在那里，他使用激光和粒子加速器进行原子和核物理实验，这些基础的研究发生在不易造成广泛影响的郊区。最终，他还前往世界上一些最偏远的地区开办工坊，并帮助这些地区创建了科学和娱乐交流中心，这就像一个位于好莱坞的美国国家科学院办公室，将严肃的科学融入大众传媒。

艾伦一开始在好莱坞和中国从事电影业。然后，他被游戏发行商美国动视公司招聘过来，最终他成立了工作室。他继续同库克湾部落理事会和白宫等合作伙伴一起搞研究，他们发现电子线媒体利用游戏进行学习，产生了很大的社会影响。从那以后，他为《科学美国人》杂志撰写文章，并担任美国国家科学基金会和美国国防部高级研究计划局的首席研究员，与比特和原子中心及智造基金会共同制作游戏。

乔尔在促进劳资合作和实施高绩效工作系统方面有着丰富的工作经验，随后他被任命为伊利诺伊大学劳动和就业关系学院院长、劳动和就业关系协会主席，以及哈佛大学法学院《谈判杂志》的编辑。他在美国国家科学基金会的协调参与方方面的工作促使他研究了“数制”工坊。从社会角度来看，“数制”工坊这一理念极具挑战性。

虽然“数制”工坊是我们都遇到的一个思路，但我们对它们的看法却截然不同。尼尔做好了实战准备，开火，瞄准——经常以快速原型制造的精神率先尝试事物，但不能完全理解实战的后果，只是会在后来评估其影响。他在这里写了关于拉斯定律的文章，对“数制”工坊的功能进行多方位的表述，但这些灵活的功能仍需要被测试。艾伦生活在管理创意项目的世界，这些项目需要传递明确的愿景和信息，他坚持采用一致性的原则描述拉斯定律的整体连贯性。社会科学家乔尔想把拉斯定律变成一个可测试的命题，他希望每个人都能够做出积极调整以推动写作。

谈判中的分歧一直延续到我们的家庭餐桌上。我们在费城长大，但我们的父母在休假时，会带我们去亚利桑那州、牙买加和英国。在我们的青年时期，我们的父母鼓励我们探索世界，追随激情，并为他人的生活带来积极的改变。而且，永远不要害怕一个糟糕的双关语（在这本书中可以用三个词来概括：拉斯就是摩尔）。

跟别人谈论这本书时，他们会问我们的第二个问题是：“你们的父母做了什么？”我们的父母（沃尔特和格拉迪丝）都是劳工仲裁员、劳资关系教授，以及国家协会的领导人（乔尔进入了家族企业）。在餐桌上定期讨论的话题就是我们父母带回家的仲裁案件。这些案件需要证据评估、讨论并予以解释。同时，讨论和解释的过程也要求我们有技巧，有毅力，有幽默感。在那时，任何正在谈话的人都充满压力地深呼吸，为下一个准备发言的人提供开场白。这也是写这本书的感觉。

我们已经开始意识到，我们在编写这本书时遇到的困难和分歧反映了我们每个人世界观中的相对弱点——我们相互指责的弱点。我们可以轻松地为技术、大众传媒和管理等领域的追随者分别写作，统一我们的背景比形成交集更加困难。但这就是第三次数字革命将要面临的状况。

致谢

当我们开始写此章节时，我们意识到有那么多人为这本书的出版做出了重要贡献，就像现在的“数制”工坊（大约1 000多家）那么多。与其在这里列出那么多名字，我们更想向本书中提到的每个人以及帮助我们讲述这些故事的人致以崇高谢意。我们很幸运能够与业内非常优秀的基本书局和布罗克曼股份有限公司这两家出版机构合作。我们非常感谢众多读者深思熟虑后的反馈。有人说，“book”（书）是一个由4个字母组成的单词，用于表示它传承生命的方式。我们的合作来源于家人的耐心、支持、好奇心，以及批评和贡献，我们对他们永远感激不尽。

参考文献

我们在此提供了我们编写《设计现实》时信息的来源、背景和参考资源，这些参考资源的出现顺序与章节中一致。资源信息是可回顾的；我们邀请您加入我们的书籍网站http://designingreality.org，以获取在线内容的链接和图书资料的更新。

序言

Our face icons were drawn by Steve McCracken (who also designed the CBA and fab lab logos).

We've committed an editorial sin by writing *Lass' Law* rather than *Lass's Law*. We pronounce it "Lasses Law," but prefer the visual emphasis of *LL* rather than *sss*.

Gordon Moore's article, "Cramming More Components onto Integrated Circuits," *Electronics*, April 19, 1965, 8, features his original observations, the remarkable predictions he made, and his insights into how he connected the dots.

We quote President Obama on the Internet being a necessity, not a luxury. For coverage, see Krishnadev Calamur, "Broadband a 'Necessity,' Obama Says, As He Pushes FCC to Expand Access," *NPR*, January 14,2015.

In discussing exponential change, we cite Ray Kurzweil, *The Singularity Is Near: When Humans Transcend Biology* (Penguin Books, 2005).Ray has been tracking exponential change in technologies with more depth and breadth than anyone we know.

Dale Dougherty is the creator of *Make* magazine and the Maker Faire.His most recent book provides a tour of the maker movement, *Free toMake: How the Maker Movement Is Changing Our Schools, Our Jobs, and Our Minds* (North Atlantic Books, 2016).

The work of Incite Focus on shelter, energy, and food is presented at the Incite Focus/ Center for Community Production/home page, www.incite-focus.org. The Fab City Pledge has a supporting website at http://fab.city.Both are referenced here and in a few other places in the book.

Neil Gershenfeld's *When Things Start to Think* (New York: Henry Holt and Company, 1999) presented what became known as the Internet of Things, and his *Fab: The Coming Revolution on Your Desktop: From Personal Computers to Personal Fabrication* (Basic Books, 2007) introducedthe fab lab movement.

E-Line Media, founded by Alan Gershenfeld and Michael Angst, is a video-game publisher passionate about making great games that help players understand and shape the world. Learn more at the company's home page, http://elinemedia.com.

Jim Gee brought his expertise in literacy to digital fabrication in "Literacy: From Writing to Fabbing,"

November 11, 2014, www.jamespaulgee.com/archdisp.php?id=71&scateg=Linguistics

第1章 智造万物

Neil developed the original CBA proposal (http://cba.mit.edu) with Joe Jacobson, Isaac Chuang, and Scott Manalis, advised by Marvin Minsky. The first How to Make (almost) Anything class (http://fab.cba.mit.edu/classe/MAS.863) was planned by them, along with Joe Paradiso and John DiFrancesco. The adventurous NSF program managers who agreed to the fab lab experiment were Kamal Abdali and Mita Desai. The first fab labs were developed and deployed by a group that included Sherry Lassiter, Bakhtiar Mikhak, Amon Millner, Caroline McEnnis, Amy Sun, and Manu Prakash.

Mel King provides a personal perspective on the Tent City story in *Chain of Change: Struggles for Black Community Development* (South End Press, 1981), which is the context for the first fab lab

The origins of Metcalfe's law, which was apparently first stated by George Gilder in 1993, citing Robert Metcalfe, is summarized in Carl Shapiro and Hal R. Varian, *Information Rules* (Harvard Business Press, 1999).

The application examples are drawn from personal experience working in fab labs, the Fab Academy (http://fabacademy.org), and the annual gatherings of the fab lab network (http://fabevent.org).

Fab lab program information is available through the Fab Foundation(http://fabfoundation.org), Academany (http://academany.org), and community portal sites (https://www.fablabs.io).

Barcelona's original Fab City commitment (http://fab.city) was developed by Tomas Diez with Vicente Guallart (city architect), Xavier Trias(mayor), and Antoni Vives (deputy mayor).

US Rep. Bill Foster reports on the initial fab and maker legislation at "Foster Introduces Legislation to Support Next Generation of Makers, Innovators," press release, May 25, 2015, https://foster.house.gov/media-center/press-releases/foster-introduces-legislation-to-support-next-generation-of-makers.

The Global Humanitarian Lab (www.globalhumanitarianlab.org) was launched by David Ott (from the International Committee of the Red Cross) and Olivier Delarue (from the UN High Commissioner for Refugees).

第2章 第三次数字革命

Individuals quoted by name here, and in Chapters 4 and 6, were interviewed by Joel and Alan in January, February, March, and April 2017, as part of the research for the book.

The DARPA Social Science Initiative, "Accelerating Discovery with New Tools and Methods for Next Generation

Social Science," was announced on March 4, 2016, www.darpa.mil/news-events/2016-03-04.

The graph on Internet use was constructed with data from Internet World Stats, International Data Corporation, and Nua Ltd.

Victoria Rideout and Vikki S. Katz's report focuses on digital access in low-income families: "Opportunity for All? Technology and Learning in Lower-Income Families," Joan Ganz Cooney Center, February 3, 2016, www.joanganzcooneycenter.org/publication/opportunity-for-all-technology-and-learning-in-lower-income-families.

The stakeholder alignment survey was distributed through the Fab Foundation email list to 11,355 recipients, of which 3,439 opened the document. The survey was approved for use by the Brandeis Institutional Review Board, which ensures proper treatment of the subjects of research(human and others). Complete responses were received from 178 individuals (resulting in a 5 percent response rate among those who opened the document). We had requests to translate the survey into multiple languages, which we were unable to do in the limited time we had available to us. This affects the response rate and the interpretation of the data. When asked to indicate their primary role, the participants gave the following mix of responses:

PRIMARY ROLE	PERCENTAGE
Local fab lab	48.0
Maker Space	9.1
Hacker Space	2.3
Educator promoting design and fabrication	18.9
Community organizers promoting design and fabrication	5.1
Public leaders supporting digital design and fabrication	5.1
Fab Academy	2.9
Fab Foundation	1.1
Producing and selling digital fabrication equipment	0.6
Digital materials manufacturing	1.1
Other	5.7

Most individuals had more than one primary role. When the respondents were asked to check all that apply, here was the distribution:

ALL ROLES (CHECK ALL THAT APPLY)	PERCENTAGE
Local fab lab	81.1
Maker Space	42.3
Maker Faire	31.4
Hacker Space	17.1
Educator promoting design and fabrication	66.3
Community organizers promoting design and fabrication	55.4
Public leaders supporting digital design and fabrication	37.1
Fab Academy	28.1
Fab Foundation	19.1
Producing and selling digital fabrication equipment	12.6
Digital materials manufacturing	21.7
Other	8.6

Note that this table includes Maker Faire, which was not a primary role in the above table. Most of these individuals

have only been in their primary role for less than five years (54.7 percent), with most of the rest in their roles 5 to 10 years (32.6 percent) and the balance longer (12.8 percent).

Over three-quarters of the respondents were male (79.1 percent), with the balance female (20.3 percent) or intersex (0.6 percent). Slightly over half (50.3 percent) have a graduate degree (non-doctoral), and another 15.0 percent have a PhD. Some 18.5 percent have a bachelor's degree, and the balance (16.2 percent) have an associate's degree, some college, or a high school degree. The age distribution is remarkably even across age groupings, with 30.7 percent under the age of 35, 28.9 percent between 35 and 44, 24.9 percent between 45 and 54, and 15.6 percent over the age of 55.

The survey instrument was developed by the company Joel cofounded, Way Mark Analytics, and the Michael Haberman generated the visualizations. These are termed *z-flower* since what are called z-scores(a mathematical transformation to have a mean of zero and a standard deviation of 1) are sometimes used in the generation of the visualization.The print copy of the book includes z-flowers with shades of gray. The color versions (and additional z-flowers) can be viewed at http://waymar systems.org:8000/report/FAB2017#.

These visualizations make it possible to see every stakeholder's response, rather than just clusters of responses as might be represented in a histogram. Small numbers of respondents who are either very positive or very negative stand out in this format and are often essential to understand when seeking stakeholder alignment.

When we report that respondents found a given issue very impor tant,we did so on the basis of responses that were a .7 or higher on an elevenpoint numerical rating scale where 0 was “not important” and 1.0 was“very important.” Similarly, when we reported that an issue was very diffcult, it was ranked a .3 or lower on an eleven-point numerical rating scale,where 0 was “very difficult” and 1.0 was “very easy.

When viewing the visualizations, first look for the value on impotance to assess the degree to which something is a must-have. Second,look for the degree of contrast between importance and ease, which indicates the gap to be addressed and the level of pain in the system (pain is high if something is very important, but hard to do). Third, look for the variation in the responses. If difficulty is bimodal, for example, it is likely that bringing up the bottom (those who find something difficult) will b more effective than enhancing things for people who do not find it difficult There may also be bright spots or pain points highlighted in the visualization that need further investigation. Altogether, these visualizations are designed to provide increased situational awareness and prompts for further action.

Kate Pickett and Richard Wilkinson document the benefits of equaity in *The Spirit Level: Why More Equal Societies Almost Always Do Better*(Bloomsbury Press, 2009),

Gui Cavalcanti attempts to clarify terminology and culture in “Is It a Hackerspace, Makerspace, Tech Shop, or

FabLab?," *Make* magazine, May 22, 2013. This topic would benefit from further dialogue

The Freestyle Chess story was taken from Andrew McAfee and Erik Brynjolfsson's *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (W. W. Norton & Company, 2016).

Mariana Mazzucato's *The Entrepreneurial State: Debunking Public vs. Private Sector Myths* (Anthem Books, 2015) identifies the key role of public-funded research in driving private-sector innovation and growth.

The sample fab lab budget was provided by the Fab Foundation.

The Fab City white paper "Locally Productive, Globally Connected Self-Sufficient Cities" is available at http://fab.city/whitepaper.pdf

For more information on alternative approaches to risk assessment,see Nancy Leveson et al., "Systems Approaches to Safety: NASA and the Space Shuttle Disasters," in *Organization at the Limit*, ed. William Starbuck and Moshe Farjoun (Wiley-Blackwell, 2005).

"The Science Behind Shaping Player Behavior in Online Games," Jeffrey Lin's talk at Game Developers Conference, 2013, outlines Riot Games' efforts to reduce toxic behavior in the *League of Legends* game(http://gdcvault.com/play/1017940/The-Science-Behind-Shaping-Player).Brendan Maher's "Can a Video Game Company

Tame Toxic Behaviour?,” *Nature*, March 30, 2016, also covers Riot Games’ efforts to reduce toxic behavior.

US Rep. Bill Foster’s proposed National Fab Lab Network Act of 2015(H.R. 1622) continues in discussion in the US Congress.

More information on the UN Sustainable Development Goals is available at www.un.org/sustainabledevelopment/sustainable-development-goals.

第3章 背后的科学

The Moore’s Law data was taken from Gordon Moore’s original paper“Cramming More Components onto Integrated Circuits,” *Electronics* 38,no. 8 (1965), and then from Intel’s mainline chip specifications (http://ar.intel.com).

The Lass’ Law data came from CBA’s internal records, the wiki started by Frosti Gíslason (http://wiki.fablab.is), and the portal started by Tomas Diez (https://www.fablabs.io).

Shannon’s original reference for reliable communication was C. E.Shannon, “A Mathematical Theory of Communication,” *Bell System Technical Journal* 27 (1948): 379–423, and von Neumann’s original reference for reliable computation was John von Neumann, “Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components,” *Automata Studies* 34 (1956): 43–98.

The discussion of digitization in biology drew on the work of, and conversations with, George Church, Joe Jacobson, Pete Carr, Tom Knight,Erik Winfree, Paul Rothemund, Erez Lieberman Aiden, Alicia Jackson,John Glass, and Elizabeth Strychalski.

The discussion of digitization in physics drew on the work of, and discussions with, Charles Bennett, Rolf Landauer, David DiVincenzo, Nabil Amer, Scott Kirkpatrick, Geoff Grinstein, Ed Fredkin, Tommaso Toffoli,Norman Margolus, Seth Lloyd, Isaac Chuang, Jean-Jacques Quisquater,and Wojciech Zurek.

The discussion of digitization in intelligence drew on the work of, and discussions with, John Doyle, Raff D’Andrea, Pablo Parrilo, Ben Recht,Andreas Weigend, Terry Sejnowski, John Hopfield, and Sebastian Seung

第4章 路径创造

The presentation of the reactive roots of the social sciences draws on discussions with Scott Cooper, Dan Cornfield, Ton Kochan, Bob McKesie, and members of the Stakeholder Alignment Collaborative, including Karen Baker, Nick Berente, Pat Canavan, Gabe Gershenfeld, Brandon Grant, John Leslie King, Christine Kirkpatrick, Barbara Lawrence, Chris Lenhardt, Peter Levin, Spenser Lewis, Matt Mayernik, Charlie Mcelroy,Barbara Mittleman, Namchul Shin, Shelly Stall, and Susan Winter.

In *The Protestant Ethic and the Spirit of Capitalism* (George Allen &Unwin Ltd., 1905; trans. by Talcott Parsons, 1930), Max Weber demonstrated how religious conceptions of work, particularly Calvinist concepts,conferred legitimacy on a capitalist system.

Émile Durkheim's *The Division of Labor in Society* (1893) was based on his doctoral dissertation and traced the impact of the industrial revolution on class structure in society.

Octavia Hill, in *Our Common Land (and other short essays)* (Macmillan, 1877), focused on visiting the poor, the importance of charity, the value of open spaces, and related topics.

Jane Addams's 1892 essay "The Subjective Value of a Social Settlement" focused on Hull House in Chicago and was one of many essays she wrote on homes for the poor.

The International Union of Psychological Science traces its roots to the first international congress in 1885 in www.iupsys.net/about/histor/index.html. The founding of the American Psychological Association is documented at www.apa.org/about/apa/archives/apa-history.aspx

Sidney Webb and Beatrice Webb, *Industrial Democracy* (Longmans,Green & Co., 1897), focused on collective bargaining as the foundation for "industrial democracy," a new term introduced into the social sciences by the Webbs.

John R. Commons coauthored the four-volume *History of Labor in theUnited States* (Macmillan, 1918–1935). It is part of a body of work connecting changes in institutional

arrangements, such as the structure of unions, with changes in the structure of markets, such as happened with the impact of the Erie Canal on commerce.

Mary Wollstonecraft Shelley's *Frankenstein: or, The Modern Prometheus* (Lackington, Hughes, Harding, Mavor, & Jones, 1818) is one of the first modern examples of science fictio

Charles Dickens's *Hard Times* (Bradbury and Evans, 1854) is a critique of the utilitarian philosophy that accepted the social order associated with the industrial revolution.

Upton Sinclair's *The Jungle* (Doubleday, Jabber & Company, 1906) illustrates the power of fiction to influence policy, such as what happene with legislation on the meatpacking industry. In later years, John Steinbeck's *The Grapes of Wrath* (Viking Press, 1939) and other writers continued the tradition of social commentary through fiction

Path dependency as a concept emerged in economics with efforts to account for new technologies. See, for example, Paul A. David, "Path Dependence: Putting the Past into the Future of Economics," working paper,Institute for Mathematical Studies in the Social Sciences Technical Report 533, Stanford University, November 1988; and W. Brian Arthur, "Self Reinforcing Mechanisms in Economics," in *The Economy as an EvolvingComplex System: The Proceedings of the Global Economy Workshop, Held September, 1987 in Santa Fe, New Mexico*, ed. Philip W. Anderson, Kenneth J. Arrow, David Pines (Addison-Wesley, 1988).

Robert Michels' s *Political Parties: A Sociological Study of the Oligarchical Tendencies of Modern Democracy* (Hearst, 1911) focuses ontrade unions and political parties, both mission-driven institutional arrangements that focused first on continued existence and only after that on mission.

Thomas Kuhn' s *The Structure of Scientific Revolutions*(University of Chicago Press, 1962) documents the conservative forces in institutions that constrain change, even in the evidence-based world of science.

Adam Smith' s *An Inquiry into the Nature and Causes of the Wealth of Nations* (London: W. Strahan, 1776) has shaped thinking about markets that were just beginning to emerge in ways that contrasted with feudalism. Note that the original handwritten manuscript (with cross-outs and notes) is in the rare-books collection of the University of Illinois at Urbana-Champaign.

Karl Marx' s *The Communist Manifesto* (1848) calls for the class struggle against the owners of the means of production. In this work and in jointly authored works with Friedrich Engels, there is surprisingly little on what it would mean for workers to own the means of production, in contrast with the ownership class. Increased individual and community ownership of the means of production is, of course, a prospect that is central to this book.

Frederick Taylor' s *The Principles of Scientific Managemen* (New York: Harper and Brothers, 1911) was reprinted many times as a central reference for Taylor Society chapters

around the country and is foundational for the field of industrial engineering

Fred Emery and Merrelyn Emery wrote *A Choice of Futures* (Martinus Nijhoff Social Sciences Division, 1976) as part of a series of monographs on the quality of working life.

Stewart Brand's 1968 launch of the *Whole Earth Catalog* led to many editions, a *Whole Earth Epilogue*, and, beginning in 1974 (the year Joel graduated from high school), the launch of the *CoEvolution Quarterly*.

John Markoff's *What the Dormouse Said: How the Sixties Counterculture Shaped the Personal Computer Industry* (Viking Penguin, 2005)connects the rise of digital computation with countercultural forces in society.

Stewart Brand and J. Baldwin edited *Soft-Tech: A Co-EvolutionBook* (Penguin, 1978), which is designed with the same look and feel as the *Whole Earth Catalog*, but with a focus on technology, broadly defined

Alvin Toffler's*Future Shock* (Random House, 1970) is based on an article he wrote: "The Future As a Way of Life," in *Horizon* magazine.

Michael Piore and Charles Sabel's *The Second Industrial Divide:Possibilities for Prosperity* (Basic Books, 1984) connects the dots between changing computer technologies used in manufacturing and changing markets.

Ray Kurzweil's *The Age of Intelligent Machines* (MIT Press, 1992) is a precursor to his *The Age of Spiritual Machines: When Computers Exceed Human Intelligence* (Viking

Press, 1999) and his *The Singularity Is Near:When Humans Transcend Biology* (Viking Press, 2005), all of which trace exponential rates of change in communication and computation technologies and the implications for core assumptions about human existence.

Arthur C. Clark's *2001: A Space Odyssey* (New American Library,1968) was released concurrently with the movie, which he developed jointly with Stanley Kubrick.

When William Shatner visited Neil's lab to discuss the science of Star Trek while he was writing *I'm Working on That: A Trek from Science Fiction to Science Fact*, he and Neil agreed that they couldn't tell who was imitating whom.

William Gibson's "Burning Chrome," in *Omni* (July 1982), introduced the term "cyberspace," which was further elaborated in his *Neuromancer*(Ace, 1984).

Note that Joel finds inspiration in Robert Heinlein's concept of a fair witness in Heinlein's novel *A Stranger in a Strange Land* (Putnam,1961) — a science-fiction idea that is desperately needed in society

The World Bank's "2016 World Development Report on Digital Dividends" is available at www.worldbank.org/en/publication/wdr2016.

Gamestar Mechanic is a game platform and community that enables youth to design their own original video games. *Gamestar* is published by E-Line Media and was developed in partnership with the Institute of Play and the MacArthur Foundation. Since its launch, over one million original games

have been published by youth player/creators. Learn more at https://gamestarmechanic.com.

Douglas McGregor's *The Human Side of Enterprise* (McGraw Hill,1960) was reproduced in an annotated edition in 2006 with Joel as the editor. It is among the most influential management books ever published Joel Cutcher-Gershenfeld, Dan Brooks, and Martin Mulloy document fifty-six pivotal events in*Inside the Ford-UAW Transformation: Pivotal Events in Valuing Work and Delivering Results* (MIT Press, 2014), which identified the key role of changing deeply embedded operating assumtions to enable culture change. This, in turn, builds on Ed Schein's *Organizational Culture and Leadership* (Jossey-Bass, 1985).

A window into the culture of the Digital Equipment Corporation is provided by Guidon Kunda in *Engineering Culture: Control and Commitment in a High-Tech Corporation* (Temple University Press, 2000).

William Bridges's *Managing Transitions: Making the Most of Change*(Da Capo Press and Perseus Books, 1991) has recently been released in a special twenty-fifth anniversary edition, indicating its long-standing value.

MIT chancellor Cynthia Barnhart and medical director William Kettyle announced the Mind Hand Heart Initiative in Kath Xu, "Mind Hand Heart Expected to Improve Access to Resources," *(MIT) Tech*,September 3, 2015, http://tech.mit.edu/V135/N20/mindhandheart.html.For *Car Talk* host Ray Magliozzi's comments on the MIT seal, see Peter Dizikes, "Moving Spectacle: MIT Marks 100 Years in Cambridge

with ‘Crossing the Charles’ Parade and Evening Celebration,” *MIT News*, May 9, 2016, at http://news.mit.edu/2016/crossing-charles-moving-day-parade-competition-0509. The history of the MIT Seal is at http://libraries.mit.edu/mithistory/institute/seal-of-the-massachusetts-institute-of-technology.For recent commentary on head, heart, and hands at MIT, see coverage of Megan Smith’s commencement address in 2015, http://news.mit.edu/2015/commencement-day-0605.

Robert Owen’s attempt to create a new society was not sustainable,partly because of what Owen’s oldest son, Robert Dale Owen, described as “a heterogeneous collection of radicals, enthusiastic devotees to principle, honest latitudinarians, and lazy theorists, with a sprinkling of unprincipled sharpers thrown in,” who were drawn to the community. An interesting legacy is that Owen’s oldest son entered Indiana politics, where he championed legislation giving widows and married women control over property and promoted free schools in the state. He then represented Indiana in the US Congress, where he wrote the legislation establishing the Smithsonian Institution.

Mondragon is the subject of hundreds of books and scholarly articles.More information on the cooperative is at www.mondragon-corporation.com/en.

Remake Learning has become a benchmark for regional learning ecosystem transformation. More information on Remake Learning is at http://remakelearning.org. The Remake Learning Playbook, a field guide of ideas and resources for building

innovation networks for teaching and learning, was created by the Sprout Fund; see http://remakelearning.org/playbook.

第5章 原子与比特的进化路线图

The roadmap presented in this chapter grew out of an event that CBA and the White House Office of Science and Technology Policy cohosted on the science of digital fabrication on March 7, 2013 (http://cba.mit.edu/events/13.03.scifab).

The fab lab inventory (for the community fabrication stage) is available at http://fab.cba.mit.edu/about/fab/inv.html.

William Gibson made his statement on the future being here now in a radio interview in "The Science in Science Fiction," Talk of the Nation,NPR, November 30, 1999.

The transition from rapid prototyping to the rapid prototyping of rapid prototyping (for the personal fabrication stage) is presented in Nadya Peek's "Making Machines That Make: Object-Oriented Hardware Meets Object-Oriented Software," PhD dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2016.

The source for the robotic builder is Benjamin Jenett and Kenneth Cheung, "BILL-E: Robotic Platform for Locomotion and Manipulation of Lightweight Space Structures," 25th AIAA/AHS Adaptive Structures Conference, AIAA SciTech Forum, Grapevine, Texas, January 13, 2017(AIAA 2017-1876).

The transition from analog to digital materials (for the universal fabrication stage) is presented in Kenneth Cheung's "Digital Cellular Solids:Reconfigurable Composite Materials," PhD dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2012, and William Langford's "Electronic Digital Materials," master's thesis, Massachusetts Institute of Technology, Cambridge, MA, 2014.

The transition from assembly to self-assembly (for the ubiquitous fabrication stage) was introduced by John von Neumann and A. Burks, *Theory of Self-Replicating Automata* (University of Illinois Press, 1966), and the design tools are discussed in Amanda Ghassaei's "Rapid Design and Simulation of Functional Digital Materials," master's thesis, Massachusetts Institute of Technology, Cambridge, MA, 2016.

第6章 未来的现实

Alvin Toffler's*Future Shock* (Random House, 1970), which we have noted in various places, is prescient in its appreciation for exponential rates of change.

Carl Sagan's quote is from his "Why We Need to Understand Science," *Skeptical Inquirer* 14 (spring 1990): page 3.

More information on Alex McDowell's World Building Institute at the University of Southern California is at http://worldbuilding.institute.

McDowell's quote is in Evan Atherton and Maurice Conti, eds., *Four:A Collection of Short Stories Exploring the Future of Design, Technology,and Us* (Autodesk Applied Research Lab, 2016).

The source for the bridge image is Benjamin Jenett, Daniel Cellucci,Christine Gregg, Kenneth Cheung, "Meso-Scale Digital Materials: Modular, Reconfigurable, Lattice-Based Structures," ASME International Maufacturing Science and Engineering Conference, 2016; and the source for the wing image is Benjamin Jenett, Sam Calisch, Daniel Cellucci, Nick Cramer, Neil Gershenfeld, Sean Swei, and Kenneth Cheung, "Digital Morphing Wing: Active Wing Shaping Concept Using Composite Lattice Based Cellular Structures," *Soft Robotics,* March 2017, 4 (1).

Kurt Lewin's change model is in his *Field Theory in Social Science:Selected Theoretical Papers* (Harper and Brothers, 1951).

John Paul Kotter's model for leading change was first published in his "Leading Change: Why Transformation Efforts Fail," *Harvard BusinessReview*, March–April 1995, 59–67.

William Bridges's model for managing transitions is in his *Managing Transitions: Making the Most of Change* (Da Capo Press and Perseus Books, 1991), as noted earlier.

W. Edwards Deming's plan, do, check, and adjust model for continuous change was first published in his*Out of the Crisis* (MIT Press, 1986),page 88, as plan, do, study, act.

Everett Rogers' s *The Diffusion of Innovation* (Free Press, 1962) located technology in the context of communication systems.

Geoffrey Moore' s *Crossing the Chasm: Marketing and Selling High Tech Goods to Mainstream Customers* (Harper Collins, 1991) is a touchstone for technology entrepreneurs making the transition from start-up to established enterprise.

Clayton Christensen' s *The Innovator' s Dilemma* (Harvard Business School, 1997) has spawned a small army of scholars challenging his initial estimates on disruption in various industries while celebrating his core insights.

NASA' s "technology readiness levels" are summarized at Thuy Mai,ed., "Technology Readiness Level," NASA, last updated July 31, 2015, www.nasa.gov/directorates/heo/scan/engineering/technology/txt_accordion 1.html.

US Department of Defense' s "technology readiness assessment" is summarized at Assistant Secretary of Defense for Research and Engineering, "Technology Readiness Assessment (TRA) Guidance," US Department of Defense, April 2011, www.acq.osd.mil/chieftechnologist/publications/docs/TRA2011.pdf.

Stewart Brand' s "Self-Driving Genes Are Coming," his answer to a survey asking, "2016: What Do You Consider the Most Interesting Recent[Scientific] News? What Makes It Important?," appears in*Edge* (blog), edited by John Brockman

(www.edge.org/contributors/what-do-you-consider-the-most-interesting-recent-scientific-news-what-makes-it)

Joi Ito and Jeff Howe's *Whiplash: How to Survive Our Faster Future*(Grand Central Publishing, 2016) offers nine principles for navigating a world of accelerating change.

The visualization of the four stages was constructed by beginning with the current rate of change with new fab labs, using Neil's data from Chapter 2. For the curve with the rate of change for community fabrication, an annual rate of change of 1.5 was used (doubling every year and a half),which was decelerated to 1.4 in 2020 and then progressively slower to level out around 2030. The same curve was then duplicated, beginning with a five-year lag for personal fabrication and a leveling off that begins around 2035. For the final two curves, lags of an additional five years wer used for each, and the leveling off was projected much further out — at 2045 for digital materials and off the chart for programmable materials.Because this is an illustration rather than a charting of actual data, the format is different from the data charts used in earlier chapters. The full set of data is available on request from the authors.

Alvin Toffler's*Future Shock* (Random House, 1970) is also cited earlier.

Mihaly Csikszentmihalyi's *Good Business: Leadership, Flow, and the Making of Meaning* (Penguin Books, 2003) explores how the concept of flow applies to business and society

Never Alone (Kisima Innitchuna) is a commercial video game developed through an inclusive development process by E-Line Media and pioneering Alaska Native elders, writers, and storytellers. Released in November 2014, it has been selected for over seventy-five best-games lists in 2014 and has won multiple awards, including a BAFTA (British Academy of Film and Television Arts) award and Game of the Year at Games for Change. Learn more at http://neveralonegame.com.

Coverage of the Ford Foundation's commitment of a billion dollars from endowment to mission-related investments is at Ford Foundation, "Ford Foundation Commits $1 Billion from Endowment to Mission- Related Investments," April 5, 2017, www.fordfoundation.org/the-latest/news/ford-foundation-commits-1-billion-from-endowment-to-mission-related-investments.

The Heron Foundation's announcement that all $273 million of its endowment was going to mission-related investing is Heron, "Heron's Journey," no date, at www.heron.org/enterprise#2011.

More information on the Investors' Circle social impact investing network is at www.investorscircle.net.

Andrew Carnegie's *Autobiography of Andrew Carnegie* (Houghton Mifflin, 1920), was published one year after his death

The Gates-Buffet Giving Pledge is summarized at https://giving pledge.org.

See the World Bank, “Global Connect Initiative,” https://share.america.gov/globalconnect, for a summary.

The Connect ED, FCC E-Rate Program is summarized in Federal Communications Commission, “E-Rate: Schools & Libraries USF Program,” last updated April 28, 2017, www.fcc.gov/general/e-rate-schools-libraries-usf-program.

The concept of a layered system can be further understood with an example from the doctoral seminar in engineering systems that Joel cofounded at MIT in 2002. In comparing the U.S. and British health care systems, the U.S. system has been described as a two-layered system of doctor’s offices and hospitals. By contrast, the British system has been dscribed as a three-layered system of nurse-led clinics, doctor’s offices, and hospitals. The U.S. system does have some nurse-led clinics, but they are not integral to the design. Layers are interconnected, but they also function independently, which adds robustness (and complexity) to the system.

The American Library Association defines digital literacy in Marijke Visser, “Digital Literacy Definition,” *ALA Connect*, September 14, 2012, http://connect.ala.org/node/181197.

Martin Ford’s *Rise of the Robots: Technology and the Threat of a Jobless Future* (Basic Books, 2015) explores the implications of accelerating technologies such as AI and robotics on the future of work and society.

Michael Piore and Charles Sabel’s *The Second Industrial Divide: Possibilities for Prosperity* (Basic Books, 1984) was

also cited earlier.

The UN High Commissioner for Refugees estimates that there are sixty- fve million refugees at Adrian Edwards, "Global Forced Displacement Hits Record High," UNHCR, June 20, 2016, www.unhcr.org/en-us/news/latest/2016/6/5763b65a4/global-forced-displacement-hits-record-high.html.

Retirees projections are in *Current Population Reports* by Jennifer M.Ortman, Victoria A. Velkoff, and Howard Hogan, May 2014), P25-1140.

For background on the displacement of auto workers, see Joel Cutcher Gershenfeld, Dan Brooks, and Martin Mulloy, *Inside the Ford-UAW Transformation: Pivotal Events in Valuing Work and Delivering Results* (MITPress, 2015).

The Oxfam report on income and wealth inequality, "An Economy for the 99 Percent," January 2017, www.oxfam.org/en/research/economy-99,is one of a growing number of studies connecting technology, economics,and inequality.

The Max Weber quote is from 1909 and is cited by Reinhard Bendix, *Max Weber: An Intellectual Portrait* (Doubleday and Company, 1960),page 464. It reflects the visceral quality underlying his observations on bureaucracy.

David Weil's *Fissurization: Why Work Became So Bad for So Many and What Can Be Done to Improve It* (Harvard University Press, 2014)is at the center of current policy debates and is based on his leadership at the US Department

of Labor around overtime rules and other policies focused on fair pay.

In *Alone Together: Why We Expect More from Technology and Less from Each Other* (Basic Books, 2011), Sherry Turkle explores the deep human connections and disconnects associated with digital technology.

The quote in the text is from Tom Friedman's *Thank You for BeingLate: An Optimist's Guide to Thriving in the Age of Accelerations* (Farrar,Straus and Giroux, 2016).

The concluding quotes are from our survey of fab leaders who were responding to our request for "a phrase or metaphor to summarize your current view of fab labs and other aspects of digital fabrication." We selected what we thought were the most evocative quotes and only then discovered that they were from a wonderfully diverse mix of countries.

（参考文献由尼尔、艾伦和乔尔撰写）

图书在版编目（CIP）数据

设计现实/（美）尼尔·格申斐尔德，（美）艾伦·格申斐尔德，（美）乔尔·卡彻- 格申斐尔德著；丁峻峰，武川，蓝河译. -- 北京：中信出版社，2019. 6
书名原文：Designing reality
ISBN 978-7-5217-0332-0
I. ①设… II. ①尼… ②艾… ③乔… ④丁… ⑤武… ⑥蓝… III. ①数字技术－影响－社会生活－研究 IV. ① C913
中国版本图书馆CIP数据核字(2019) 第061736号

Designing Reality
Copyright © 2017 by Neil Gershenfeld, Alan Gershenfeld, and Joel Cutcher-Gershenfeld
Simplified Chinese translation copyright © 2019 by CITIC Press Corporation
All Rights Reserved

设计现实
著者：［美］尼尔·格申斐尔德 ［美］艾伦·格申斐尔德 ［美］乔尔·卡彻-格申斐尔德
译者：丁峻峰 武川 蓝河
出版发行：中信出版集团股份有限公司
（北京市朝阳区惠新东街甲4号富盛大厦2座 邮编100029）

字数：286千字
版次：2019年6月第1版
京权图字：01 - 2019 - 1571
广告经营许可证：京朝工商广字第8087号
书号：ISBN 978-7-5217-0332-0

版权所有·侵权必究



深度学习

深度学习会扩大你的认知　人工智能不是生存威胁

智能时代的核心驱动力量

THE DEEP LEARNING REVOLUTION

[美]特伦斯·谢诺夫斯基（Terrence Sejnowski）_著　姜悦兵_译

读懂深度学习，才能不被 AI 取代！

你可以只读谢诺夫斯基

谷歌前云 AI 负责人 李飞飞 诚挚推荐

微软小冰之父 李笛 作序推荐

世界十大 AI 科学家之一
美国“四院院士”
全球 AI 专业会议 NIPS 基金会主席

重磅作品



中信出版集团

深度学习

[美]特伦斯·谢诺夫斯基 著

姜悦兵 译

中信出版集团

目录

推荐序 面对科技拐点，我们的判断与选择

中文版序 人工智能会放大认知能力

前言 深度学习与智能的本质

第一部分 智能的新构想

01机器学习的崛起

02人工智能的重生

03神经网络的黎明

04大脑式的计算

05洞察视觉系统

第二部分 深度学习的演进

06语音识别的突破

07霍普菲尔德网络和玻尔兹曼机

08反向传播算法

09卷积学习

10奖励学习

11火爆的NIPS

第三部分 人类，智能与未来

12智能时代

13算法驱动

14芯片崛起

15信息科学
16生命与意识
17进化的力量
18深度智能
附录一 致谢
附录二 词汇表

献给波和索尔，特蕾莎和约瑟夫
谨以此书纪念所罗门·哥伦布

推荐序

面对科技拐点，我们的判断与选择

李笛

微软小冰之父

微软（亚洲）互联网工程院副院长

在近年来陆续出版的、解读人工智能技术与趋势的许多书籍中，这是一本不可多得的好书。它的阅读过程令人愉悦，涉及的知识深度又比较恰当。因此，即使是不具备相关领域知识背景的读者，也能够轻松地读完它。人们完全可以利用“碎片时间”来研读这本30多万字的大作，从而集中了解到与人工智能相关的技术分支、组织人物与重要事件。在人工智能热度很高的当下，这本书的价值在于，帮助读者建立一种相对贴近事实的科学观。

读者可以把这本书当作一本有关人工智能的简明历史来看待。人工智能是科技王冠上的钻石，而深度学习代表了其中一个承上启下的重要阶段。深度学习脱胎于科学家们六十多年前开始的人工智能研究，其自身的概念形成，到落地开花，则只有十多年的光景。与过去相比，深度学习极大地推进了人工智能各个分支课题的发展速度；与未来相比，我们今天所取得的一切成果，都是非常粗糙的，注定会被更好的成果取代。因此，了解深度学习，就如同站在一个关键的节点上向时间河流的上下游看，一览无余。

我相信，不同的人会从这本书中得到不同的收获。总体而言，这本书有助于在我们心目中更加清晰准确地绘制人工智能的未来图景。

从某种意义上说，所有的过去亦都昭示了未来，但我更建议读者以最轻松的心态来阅读它。因为这样，能够让读者以更加客观公正的视角去检阅人类与机器的能力短板与优长——你可以从本书中了解到那些令人惊讶，甚至于有些担忧的科技进展，大致了解它们背后的原理。这展现了人工智能相对于人类而言的单方面优势。另一方面，你也能看到许多真实事例，反映了人工智能相对于人类而言的“笨拙”。科学与科学幻想泾渭分明。在现实中，这种“笨拙”的情况往往更加普遍。这些事例有时令人忍俊不禁，它恰恰体现了我们人类的大脑是多么精妙的设计。

事实上，在我看来，当下最令人彷徨不定的，并不是人工智能有多么“强大”或有多么“笨拙”，而是我们已处在一个科技的拐点，需要由我们每个人对未来的走向做出抉择。这是一个非常具有现实意义的话题。虽然深度学习是这个拐点的主要推动力之一，但它并不需要为我们的困扰承担责任：

- **选择权的困扰**：一辆无人驾驶汽车行驶在道路上，假设突然面临必然要发生的车祸，它应当向左撞向一个无辜的老人，还是向右撞向一个无辜的壮年？

- **决策权的困扰**：一个系统可以基于人类个体不具备的广泛即时的大数据，用任何人无法企及的速度，迅速做出某个决策。这样的洞察和决策力，应该掌握在谁的手中？

- **工作权的困扰**：一项基于人工智能的技术可以比人类以更好的质量和速度去完成某项生产。这项技术应该归属于工厂主来代替工人，还是应该归属于工人来帮助工厂主更好地完成工作？前者会带来失业，而后者有望带来更高的工作效率。

- **社会层面的困扰**：一个面向情感的人工智能机器人帮助一个人解决孤单，却使他主动减少了与他人的社交沟通。这种陪伴究竟是在帮他解决问题，还是制造了更多的问题？

• **技术滥用的困扰：**一项技术可以帮助任何人打造与他们高度相似的语音，制造出来的声音，令他的家人也难辨真假。这样的技术会不会被别有用心的人用于犯罪，例如诈骗电话？

不知不觉间，这些看似遥远的事，突然间已变成我们必须要面对的现实情况，而我们也已经在上述一些场景中做出了初步的抉择。其中一个关键因素是，人们常常对人工智能的“智商”印象深刻，但往往忽略了：人工智能系统化的优势之一在于“大规模的并发”。因此，任何一种以上技术应用的场景，只要乘以巨大的人口，都会带来很大的影响。相对而言，在围棋游戏中赢过人类，其实是最不需要担忧的了。

在微软，我们最近成立了与人工智能及伦理相关的组织，力图在当前的框架内去发现尽量多的问题，尽可能在早期就避免问题的发生。微软在人工智能领域的技术和产品线很广泛，因此我们做过的抉择也相对较多。这些抉择往往决定了我们在有能力的情况下，主动放弃去做什么。

这意味着克制与敬畏之心。例如：微软小冰在两年前推出拟人的全双工语音电话技术（Full Duplex）时，我们就制定了该产品的伦理规则，不允许在用户不知情的情况下，让小冰伪装成真人去拨打电话。我们也不使用微软小冰的技术去从事呼叫中心的外呼业务，因为它存在被滥用为垃圾广告电话的风险——尽管这些往往意味着巨大的商业价值。今天，在中国、美国、日本、印度和印度尼西亚，微软小冰拥有近7亿人类用户，如果她在对话的时候努力诱导人们去购买某种商品，显然会带来可观的收入预期。但谁会愿意和一个一心想着如何诱导你买东西的人成为知己呢？

这种克制，不仅仅是一两家企业的责任。它依赖于整个社会对人工智能，特别是深度学习相关技术的了解。对技术的了解越普遍，也就越能帮助企业更好地运用手中的技术，进而帮助我们每一个人获得更好的生活，享受人工智能为我们带来的价值。

在我看来，这就是这本书所具有的现实意义。它并非教科书，而是一本面向未来的历史书。它揭示了人工智能有望给世界、给人类带来的巨大改变，远超我们现在所能想象到的全双工语音、人脸识别、情感计算甚至是自动驾驶。换句话说，基于我们现在的技术和产品水平，相信许多人已能在脑海中比较清晰地勾勒出，自己在马路上与一辆并没有司机驾驶的汽车相遇的场景。但与人工智能即将展现的伟大图景相比，这些都不值一提。

中文版序

人工智能会放大认知能力

60年前，数字计算机在人工智能（AI）的萌芽期问世，深度学习革命的种子也在那时被播种开来。深度学习是数据密集型的，通过实例来学习如何解决难题，比如视觉对象识别、语音识别和自然语言翻译等。人类从婴儿时期睁开眼睛的那个时刻起，就开始从经验中学习，到后来获得语言、运动、玩电子游戏等最高程度的能力；相比之下，传统的劳动密集型人工智能方法是基于编写不同的复杂计算机程序来解决每个问题。

本书讲述了20世纪80年代一小群研究人员的故事，他们证明了基于大脑式计算的全新方法是可行的，从而为深度学习的发展奠定了基础。

当时已有的人工智能学术研究中心都投注于编程，并且都具有强大的实力，但却无法解决上述任何难题。又过了30年，计算机才变得足够快，也出现了大量可供利用的数据。这一变化让深度学习得以克服这些难题，并在今天的人工智能领域占据主导地位。其他领域同样可以借鉴这一经验教训，例如语言学，曾经普遍持有的既定信念阻碍了该领域整整一代研究者的进步。深度学习改变了语言学，使其发展基于来自现实世界的数据，而非无法捕捉这些复杂性的理想世界的数据。

回溯历史，人工智能诞生的秘密可以在自然界中找到答案，我们对此并不应该感到惊讶。大自然有数亿年的时间通过进化找到解决方

案，对这些解决方案进行逆向工程能够让我们受益匪浅。了解大脑如何运转是21世纪最大的挑战之一。大自然发明了许多经受住了时间考验的算法。理解这一挑战并投资于大脑研究的国家将获得巨大的回报，远远超出20世纪物理学和化学研究的突破所产生的影响，这些影响已经极大地丰富了我们的生活。美国已经为“BRAIN计划”（英文全称为Brain Research through Advancing Innovative Neurotechnologies，即“通过推动创新型神经技术开展大脑研究计划”）注资50亿美元，欧洲、日本和许多其他国家或地区也在进行类似的投资。中国正在投资当前的人工智能技术，但它是否拥有投资大脑研究的远见，年青一代又是否会接受这一挑战呢？

深度学习对社会和个人生活将产生深远的影响，其影响方式也是难以想象的。在本书中，我提出了一个观点，即你无须担心人工智能将接管你的工作。人工智能会让你更聪明，让你所能实现的成就达到新的高度。就像工业革命时期蒸汽机放大了物理能力一样，人工智能也会放大认知能力。我们刚刚步入一个新的时代——信息时代。我们进入的新世界不仅会使我们变得更聪明，还会让我们更清楚地认识自己，从而回答古代的哲学先驱们最早提出的一系列问题。对于自身，我们又会得出哪些深刻的见解呢？



前言

深度学习与智能的本质

如果你在连接了互联网的安卓手机或谷歌翻译平台上使用语音识别功能，你其实是在与经过深度学习训练的神经网络[1]进行交流。过去几年，深度学习为谷歌带来了丰厚的利润，足以支付Google X实验室中所有未来主义项目的成本，包括自动驾驶汽车、谷歌眼镜和谷歌大脑。[2]谷歌是最早拥抱深度学习的互联网公司之一，并在2013年聘请了深度学习之父杰弗里·辛顿（Geoffrey Hinton），其他公司也在竞相追赶它的脚步。

人工智能近期取得的进展得益于大脑逆向工程。分层神经网络模型的学习算法受到了神经元之间交流方式的启发，并依据经验进行了改进。在网络内部，世界的复杂性转变为五彩缤纷的内部活动模式，这些模式是智能的元素。我在20世纪80年代研究的网络模型很小，相比之下，现在的模型有数百万个人造神经元，深度达到了几十层。持久的努力、大数据和更强大的计算机运算能力使得深度学习在人工智能领域一些最困难的问题上取得了重大突破。

我们并不善于想象新技术对未来的影响。谁能在1990年互联网刚开始商业化的过程中预见到它对音乐产业的影响，以及对出租车业务、政治运动，还有我们日常生活几乎所有方面的影响？同样，我们也未能预见到电脑会如何改变我们的生活。IBM（国际商业机器公司）总裁托马斯·沃森（Thomas J. Watson）在1943年说的一句话后来被广泛引用：“我觉得全世界也许能卖出5台计算机吧。”[3]很难想象

一个新发明都有哪些用途，其发明人对这些用途的预测也不见得比其他人更准确。在乌托邦和世界末日的两极之间，有很多关于深度学习和人工智能应用场景的预测空间，但即使是最具想象力的科幻小说作家也不大可能猜出它们最终会产生什么样的影响。

本书的初稿是我在太平洋西北地区[4]徒步旅行，并思索了近几十年来人工智能领域的显著变化之后写出来的。这本书讲了一个一小群研究人员挑战AI研究建制派的故事，这些建制派在当时拥有更充足的资金支持，并被看作“唯一的主导力量”，他们大大低估了这些问题的难度，并且所依赖的对智能的直觉，后来被证明是有误导性的。

地球上的生命充满了无数奥秘，但最具挑战性的也许是智能的本质。自然界充斥着各种形式的智能，从微小的细菌到复杂的人类智能，每种智能都适应了它在自然界中的位置。人工智能也将以多种形式出现，并在智能族谱中占据特殊的位置。随着基于深度神经网络的机器智能日渐成熟，它可以为生物智能提供一个新的概念框架。

这是一本关于深度学习的过去、现在和未来的指南。不过本书并不是对该领域发展历史的全面梳理，而是记录了这一领域重要概念的进步及其背后研究群体的个人观点。人类的记忆并不可靠，对故事的每次复述都会导致记忆的偏差，这个过程叫作“重整记忆”。这本书中的故事延续了40多年，尽管有些对我来说依然历历在目，就像昨天刚发生的一样，但我很清楚，那些故事在我的记忆中不断被复述时，有些细节已经悄悄地被改写了。

本书的第一部分提供了深度学习的动机和理解其起源所需的背景信息；第二部分解释了几种不同类型的神经网络架构中的学习算法；第三部分则探讨了深度学习对我们当下生活产生的影响，以及未来若干年可能产生的影响。然而，正如纽约扬基队的哲人尤吉·贝拉（Yogi Berra）曾经说过的那样：“做出预测很难，特别是对未来的预测。”本书前八章的内容交代了故事的技术背景；三个部分开头的要事年表记录了与这个故事有关的事件，时间跨度超过了60年。

[1] 严格地说，神经网络是一个生物实体，机器学习中使用的模型是人工神经网络——ANNies。但若无另行说明，本书中的“神经网络”默认代指人工神经网络。

[2] Conor Dougherty, “Astro Teller, Google’s ‘Captain of Moonshots,’ on Making Profits at Google X,” *New York Times*, February 6, 2015, https://bits.blogs.nytimes.com/2015/02/16/googles-captain-of-moonshots-on-making-profits-at-google-x. 深度学习将运行数据中心的电力成本降低了15%，每年可以节省数亿美元。

[3] 尽管沃森在1943做出的估计从未得到过确认，但它反映了当时的人们普遍无法想象电脑的未来。

[4] 太平洋西北地区是指美国西北部地区和加拿大的西南部地区。——编者注

第一部分 智能的

新構想

要事年表

1956年

达特茅斯人工智能夏季研究计划（The Dartmouth Artificial Intelligence Summer Research Project）开启了人工智能领域的研究，并鼓舞了一代科学家探寻可以媲美人类智慧的信息技术的潜力。

1962年

弗兰克·罗森布拉特（Frank Rosenblatt）出版了《神经动力学原理：感知器和大脑机制的理论》(*Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*)，该书介绍了一种应用于具有单层可变权重的神经网络模型的学习算法，该算法是今天的深度神经网络模型的学习算法的前身。

1962年

大卫·休伯尔（David Hubel）和托斯坦·威泽尔（Torsten Wiesel）发表了《猫的视觉皮质中的感受野、双目互动和功能架构》（Receptive Fields,Binocular Interaction and Functional Architecture in the Cat’s Visual Cortex）一文，第一次报道了由微电极记录的单个神经元的响应特性。深度学习网络的架构类似于视觉皮质的层次结构。

1969年

马文·明斯基（Marvin Minsky）和西摩尔·帕普特（Seymour Papert）出版了《感知器》(*Perceptrons*)，该书指出了单个人造神经元的计算极限，标志着神经网络领域寒冬的到来。

1979年

杰弗里·辛顿和詹姆斯·安德森（James Anderson）在加州拉荷亚市（La Jolla）举办了“关联记忆的并行模型”（Parallel Models of Associative Memory）研讨会，把新一代的神经网络先驱们聚集到了一起，同时也推动辛顿和安德森在1981年发表了同名系列研究著作。

1986年

第一届神经信息处理系统大会（Neural Information Processing Systems，以下统称NIPS[1]）及研讨会在美国丹佛科技中心举办，该会议吸引了很多不同领域的研究人员。

---

[1] NIPS现通称为NeurIPS。——译者注

01
机器学习的崛起

不久之前，人们还常说，计算机视觉的辨别能力尚不如一岁大的孩子。如今看来，这句话要改写了。计算机不仅能和大多数成年人一样识别图片中的物体，在马路上驾驶汽车的安全性还高过16岁的青少年。更神奇的是，如今的计算机不再是被动按照指令识别和驾驶，而是像自然界的生命由数百万年前开始进化那样，自主地从经验中学习。是数据的井喷促成了这一技术进步。如果说数据是新时代的石油，那么学习算法就是从中提取信息的炼油厂；信息积累成知识；知识深化成理解；理解演变为智慧。欢迎来到深度学习的新世界。[1]

深度学习是机器学习的一个分支，它根植于数学、计算机科学和神经科学。深度网络从数据中学习，就像婴儿了解周围世界那样，从睁开眼睛开始，慢慢获得驾驭新环境所需的技能。深度学习的起源可以追溯到20世纪50年代人工智能的诞生。关于如何构建人工智能，当时存在两种不同的观点：一种观点主张基于逻辑和计算机程序，曾主宰人工智能的研究和应用数十年；另一种观点则主张直接从数据中学习，经历了更长时间的摸索才逐渐成熟。

20世纪，计算机技术还不够成熟，而且按照现在的标准，数据存储成本十分高昂，用逻辑程序来解决问题更加高效。熟练的程序员需要为每个不同的问题编写不同的程序，问题越大，相应的程序也就越复杂。如今，计算机能力日趋强大，数据资源也变得庞大且丰富，使用学习算法解决问题比以前更快、更准确，也更高效。此外，同样的

学习算法还能用来解决许多不同的难题，这远比为每个问题编写不同的程序更加节省人力。

汽车新生态：无人驾驶将全面走入人们生活

在2005年美国国防部高级研究计划局（以下简称DARPA）举办的自动驾驶挑战赛中，一辆由斯坦福大学塞巴斯蒂安·特隆（Sebastian Thrun）实验室开发的自动驾驶汽车Stanley最终赢得了200万美元现金大奖（见图1－1）。团队利用了机器学习技术教它如何自主地在加利福尼亚州的沙漠中穿行。132英里的赛道中有若干狭窄的隧道和急转弯，还包括啤酒瓶道（Beer Bottle Pass），这是一段蜿蜒曲折的山路，两侧分别是碎石遍布的陡坡和断壁（见图1－2）。特隆并没有遵循传统的AI方法，即通过编写计算机程序来应付各种偶发事件，而是在沙漠中驾驶Stanley，让汽车根据视觉和距离传感器的感应输入，学习如何像人一样驾驶。



图1-1 塞巴斯蒂安·特隆及其团队的自动驾驶汽车Stanley在2005年赢得了DARPA举办的自动驾驶挑战赛。这项突破引发了交通界的技术革命。图片来源：塞巴斯蒂安·特隆。



图1-2 啤酒瓶道。这段极具挑战性的地形位于2005年DARPA自动驾驶挑战赛的末段。该赛事要求汽车在无人辅助的情况下驶过132英里的沙漠荒路。图中远处的一辆卡车正要爬坡。图片来源：DARPA。

特隆后来参与创立了高科技项目重点实验室Google X，并开始了进一步研究自动驾驶汽车技术的计划。谷歌的自动驾驶汽车自此开始，在旧金山湾区累积了350万英里的车程。优步（Uber）已经在匹兹堡投放了一批自动驾驶汽车。苹果也步入自动驾驶领域，以扩大其操作系统控制的产品范围，并希望能够再现它在手机市场上的辉煌。汽车制造商们亲眼看见一个100年来从未改变的行业在他们眼前发生了转型，也开始奋起直追。通用汽车公司以10亿美元的价格并购了开发无人驾驶技术的硅谷创业公司Cruise Automation，并在2017年投入了额外的6亿美元用于研发。[2]2017年，英特尔以153亿美元的价格收购了Mobileye，它是一家专门为自动驾驶汽车研发传感器和计算机视觉的公司。在价值数万亿美元的交通运输领域，参与的各方都下了极高的赌注。

自动驾驶汽车不久将扰乱数百万卡车司机和出租车司机的生计。最终，如果一辆自动驾驶汽车能够在一分钟内出现，将你安全带到目的地且无须停车，在城市拥有汽车就显得不那么必要了。今天，汽车行驶时间平均仅占4%，这意味着它其余96%的时间都需要停放在某个地方。由于自动驾驶汽车可以在城市外围维修和停放，城市中被大量停车场占用的空间得以被重新高效利用。城市规划者已经开始考虑让停车场变成公园了。[3]街边的停车道可以成为真正的自行车道。其他汽车相关行业也将受到影响，包括汽车保险业和修理厂。超速和停车罚单将不复存在。由醉驾和疲劳驾驶导致的交通事故死亡人数也会相应减少。通勤浪费的时间也将被节省下来做其他事情。根据2014年的美国人口普查数据，1.39亿上班族人均单日通勤时间达到了52分钟，全年总计296亿小时。这惊人的340万年的时间本可以在人生中得到更好的利用。[4]自动驾驶汽车会使公路通行能力翻两番。[5]而且，一旦大规模投入使用，没有方向盘、可以自己开回家的自动驾驶汽车还会让大规模汽车盗窃行为销声匿迹。虽然目前自动驾驶汽车仍面临很多监管和法律层面的障碍，但这一技术一旦开始普及，我们就将迎来一个崭新的世界。可以预见的是，卡车大概会在10年内率先实现自动

驾驶，出租车要花上15年，而15到25年后，客运无人车将全面走入人们的生活。

汽车在人类社会中的标志性地位将以我们无法想象的方式发生变化，一种新的汽车生态也将应运而生。正如100多年前汽车的出现创造了许多新的行业和就业机会，围绕着自动驾驶汽车的发展，也出现了一个快速增长的生态系统。从谷歌独立出来的自动驾驶公司Waymo，8年来已经投入了10亿美元，并在加州中部山谷搭建了一个秘密测试场所。该场所位于一个占地91英亩的仿造小镇，其中还设计了骑自行车的“演员”和假的汽车事故。[6]其目的是扩大训练数据集以包含特殊和不常见的情况（也叫边缘情况）。公路上罕见的驾驶事件经常会导致事故。自动驾驶汽车的不同之处就在于，当一辆汽车遇到罕见事件时，相应的学习体验会被传递给所有其他自动驾驶汽车，这是一种集体智能。其他自动驾驶汽车公司也在建造许多类似的测试设施。这些举措创造了以前并不存在的新工作机会，以及用于汽车导航的传感器和激光器的新供应链。[7]

自动驾驶汽车仅是信息技术推动经济发生重大转变的一个最明显的体现。网络上的信息流就像城市管道里的水流。信息在谷歌、亚马逊、微软和其他IT公司的大型数据中心聚集。这些数据中心需要耗费大量电力，因此通常建在水电站附近，并利用河水来冷却信息流所产生的大量热量。2013年，美国的数据中心消耗了1000万兆瓦的电量，相当于34个大型电厂产生的电力。[8]但是目前对经济影响更大的是如何使用这些信息。从原始数据中提取出的信息被转化为关于人和事的知识：我们做什么，我们想要什么，我们是谁。计算机驱动的设备也在越来越多地利用这些知识与我们进行口头上的交流。与大脑之外、书本之中的被动知识不同，储存在云中的知识是一种外部智能，并且正在成为人们生活中积极、活跃的一部分。[9]

自然语言翻译：从语言到句子的飞跃

如今，谷歌在超过100种服务中使用了深度学习，包括街景视图（Street View）、收件箱智能回复（Inbox Smart Reply）和语音搜索。几年前，谷歌的工程师意识到他们需要将这些计算密集型应用扩展到云端。他们开始着手设计一种用于深度学习的专用芯片，并巧妙地设计了可以插入数据中心机架中的硬盘插槽的电路板。谷歌的张量处理单元（TPU）现在已配置在遍布全球的服务器上，让深度学习应用程序的性能得到了大幅改进。

深度学习快速改变格局的一个例子是它对语言翻译的影响。语言翻译是人工智能的一只圣杯，因为它依赖于理解句子的能力。谷歌最近推出了基于深度学习的最新版谷歌翻译（Google Translate），代表了自然语言翻译质量的重大飞跃。几乎一夜之间，语言翻译就从零散杂乱的拼凑短语，升级到了语意完整的句子（见图1－3）。之前的计算机方法搜索的是可以被一并翻译的词汇组合，但深度学习会在整个句子中寻找词汇之间的依赖关系。



图1-3 手机上的谷歌翻译应用可以将日语文字和菜单即时译成英文。这一功能对于在日本如何按照指示牌乘车尤为重要。

得知谷歌翻译获得了巨大进步的消息后，2016年11月18日，东京大学的暦本纯一（Jun Rekimoto）测试了这个新系统。他把欧内斯特·海明威的小说《乞力马扎罗的雪》开头的一段话翻译成了日文，然后再把这段日文翻译成英文，结果如下（猜猜哪个是海明威的原作）：

1. Kilimanjaro is a snow-covered mountain 19,710 feet high,and is said to be the highest mountain in Africa. Its western summit is called the Masai “Ngaje Ngai,” the House of God. Close to the western summit there is the dried and frozen carcass of a leopard. No one has explained what the leopard was seeking at that altitude. [10]

2. Kilimanjaro is a mountain of 19,710 feet covered with snow and is said to be the highest mountain in Africa. The summit of the west is called “Ngaje Ngai” in Masai, the house of God. Near the top of the west there is a dry and frozen dead body of leopard. No one has ever explained what leopard wanted at that altitude. [11] [12]

海明威的原作是第一段。

下一步工作是训练更大规模的深度学习网络，针对段落来提高句子间的连贯性。文字背后都有悠久的文化历史。俄裔作家和英文小说家，《洛丽塔》一书的作者弗拉基米尔·纳博科夫（Vladimir Nabokov）曾经得出结论，在不同语言之间翻译诗歌是不可能的。他将亚历山大·普希金（Aleksandr Pushkin）的诗体小说《叶甫盖尼·奥涅金》（*Eugene Onegin*）直译成了英文，并对这些诗文的文化背景做

了解释性脚注，以此论证他的观点。[13]或许谷歌翻译终有一天能够通过整合莎士比亚的所有诗歌来翻译他的作品。[14]

语音识别：实时跨文化交流不再遥远

人工智能的另一只圣杯是语音识别。不久之前，计算机的独立语音识别应用领域还很有限，如机票预订。而如今，限制已不复存在。2012年，一名来自多伦多大学的实习生在微软研究院（Microsoft Research）的一个夏季研究项目中，让微软的语音识别系统性能得到了显著的提升（图1－4）。[15]2016年，微软的一个团队宣布，他们开发的一个拥有120层的深度学习网络已经在多人语音识别基准测试中达到了与人类相当的水平。[16]

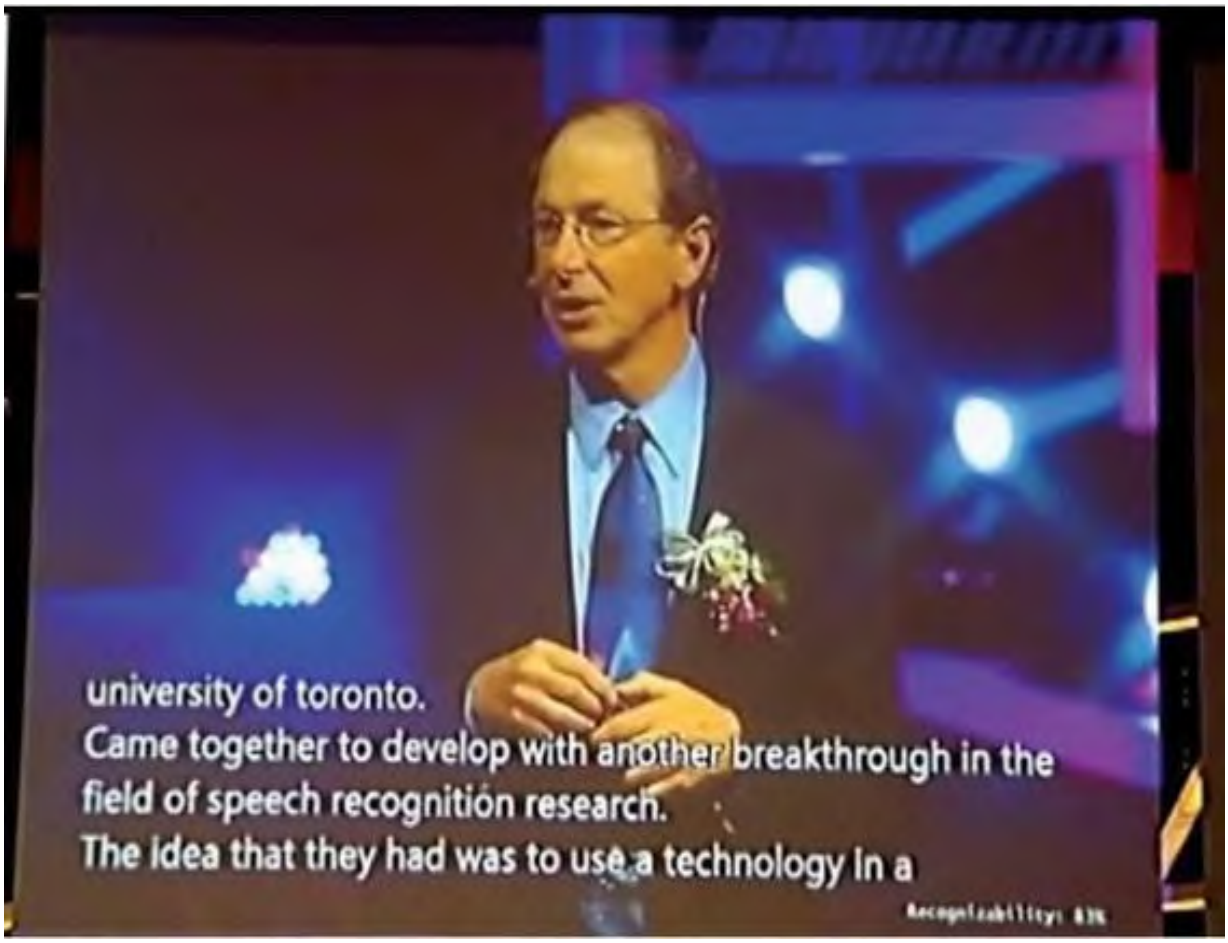


图1-4 微软首席研究官里克·拉希德（Rick Rashid）在2012年10月25日于中国天津举行的一场活动中，使用深度学习进行了自动语音识别的现场演示。在2000名中国观众面前，拉希德说的英文被自动化系统识别，先在他的屏幕图像下方显示出英文字幕，随后被翻译成了中文。此次高难度展示被全球媒体争相报道。图片来源：微软研究院。

这一突破性成果将在之后的几年逐渐影响我们的社会，计算机键盘会被自然语言接口取代。随着数字助手，如亚马逊的Alexa、苹果的Siri以及微软的Cortana先后进入千家万户，这种取代已经在发生了。就如随着个人电脑的普及，打字机退出了历史舞台，有一天电脑键盘也将成为博物馆的展品。

当语音识别和语言翻译结合到一起时，实时的跨文化交流将有可能实现。《星际迷航》中那种万能翻译机将触手可及。为什么计算机语音识别和语言翻译达到人类的水平要花这么久的时间？难道计算机

的各种认知能力同时进入瓶颈期仅仅是巧合吗？其实所有这些突破都源于大数据的出现。

AI医疗：医学诊断将更加准确

深入皮肤

随着机器学习的成熟并被应用于可获取大数据的许多其他问题，服务行业和其相关职业也将发生转变。基于数百万患者病情记录的医学诊断将变得更加准确。最近的一项研究将深度学习运用到了囊括超过2000种不同疾病的13万张皮肤病学图像中，这个医学数据库是以前的10倍大（图1－5）。[17]该研究的网络被训练用于诊断“测试集”（testset，它从未见过的新图像集）中的各种疾病。它在新图像上的诊断表现与21位皮肤科专家的结论基本一致，甚至在某些情况下还要更准确。在不久的将来，任何一个拥有智能手机的人都可以拍下疑似皮肤病变的照片，并立即进行诊断——而现在要完成同样的过程，我们需要先去看医生，耐心等待病变被专家筛查出来，然后再支付一大笔账单。这一进步将大大扩大皮肤病护理的范围，提升护理质量。如果个体可以很快得到专家诊断，他们会在皮肤病的早期阶段，也就是更容易治疗的时候就开始就医。借助深度学习，所有的医生都将更准确地诊断罕见的皮肤病。[18]



图1-5 艺术家绘制的高准确度诊断皮肤病变的深度学习网络图，2017年2月2日《自然》杂志封面。

深入癌症

如果专家在转移性乳腺癌的淋巴结活检切片图像上判断错误，就有可能导致致命的后果。这是一种深度学习擅长的模式识别问题。实际上，一个经过大量结论清晰的切片数据训练出来的深度学习网络能达到0.925的准确度，还不错，但还不及人类专家在同一测试集上达到的0.966。[19]然而，把深度学习与人类专家的预测结合起来，准确度达到了0.995，几近完美。由于深度学习网络和人类专家查看相同的数据的方式不同，二者相结合的效果比单独预测要好。这样一来，更多的生命得以被挽救。这表明在未来，人类与机器将是合作而非竞争的关系。

深入睡眠

如果你有严重的睡眠问题（70%的人一生中都会遇到这个问题），你要等待几个月才能见到你的医生（除非问题十分紧急），然后你会被转到一个睡眠诊所。在那里，你需要在身上接几十个能在你入睡时记录你的脑电图（EEG）和肌肉活动的电极，接受彻夜观察。每个晚上，你会先进入慢波睡眠，然后定期进入快速眼动（REM）睡眠，在此期间，你会做梦，但是失眠、睡眠呼吸暂停综合征、不宁腿综合征以及许多其他睡眠障碍会干扰这种睡眠模式。如果你在家里就很难入睡，那么在一张陌生的床上，全身接满了让人不安的医疗设备进入睡眠状态，绝对算得上真正的挑战。睡眠专家会查看你的脑电图记录，以30秒为单位标记睡眠阶段，一段8小时的睡眠要花几个小时才能标记完。而最终你会得到一份有关睡眠模式异常情况的报告，以及一份2000美元的账单。

依据1968年由安东尼·雷希特施芬（Anthony Rechtshaffen）和艾伦·卡莱斯（Alan Kales）设计的系统，睡眠专家将接受寻找表征不同睡眠阶段特征迹象的培训。[20]但是由于这些特征常常不明确，也不一致，只有75%的情况下专家们能在数据解读上达成一致。相比之下，我实验室之前的一名研究生菲利普·洛（Philip Low）使用无监督机器学习，花了不到一分钟的计算机运算时间，以3秒的时间分辨率自动检测睡眠阶段，和87%的人类专家达成了一致的结论。此外，这种方式只需要在头部的单个位置做记录，用不到那些触点和接线，也节省了大量佩戴和摘除的时间。2007年，我们创立了一家公司Neurovigil，想将这项技术引入睡眠诊所，但诊所对此没有表现出多大兴趣，因为靠人力标注能产生更多的现金流。实际上，依据保险号向患者开具账单，会让诊所没有动机采用更廉价的程序。Neurovigil在大型制药公司发现了另一个市场，这些公司在开展临床试验，需要测试他们的药物对睡眠模式的影响。这项技术目前正在进入长期护理设施市场，帮助解决在老年人中更普遍的进行性睡眠问题。

睡眠诊所模式是存在缺陷的，因为在这样的限制条件下不能可靠地诊断出健康问题：每个人的生理基数都不同，而偏离这个基数的信息最重要。Neurovigil已经有了一个小型设备iBrain，它可以在家里记录你的脑电图信息，将数据传到网上并分析数据的长期趋势和异常情况。这可以帮助医生及早发现健康问题，在恶化前及时干预并阻止慢性疾病的发展。其他很多疾病的治疗也将受益于持续监测，如1型糖尿病，血糖水平可以被监测并通过胰岛素进行调节。使用能够连续记录数据的廉价传感器正在对其他慢性疾病的诊断和治疗产生重大影响。

从Neurovigil的发展过程中可以看出：第一，即便拥有更好更廉价的技术，也不代表能轻易地将其转化为有市场价值，甚至更优质的新产品或服务；第二，当现有产品在市场中的地位根深蒂固，就会进一步开发出深入应用的二级市场，可以让新技术产生更直接的影响，并争取时间来改进，提升竞争力。太阳能和许多其他新兴产业的技术就是这样进入市场的。从长远来看，已被证实具有优势的睡眠监测和新技术将会覆盖到家中的患者，并最终融入医疗实践。

金融科技：利用数据和算法获取最佳回报

纽约证券交易所超过75%的交易都是自动完成的（图1－6），高频交易能在几分之一秒内进出仓位。（如果你不用为每笔交易支付费用，那么即使是很小的优势也能带来巨额利润。）更长时间范围内的算法交易会考虑到基于大数据的长期趋势。深度学习在赚钱和提高利润方面做得越来越好。[21]预测金融市场，问题在于数据嘈杂，条件不稳定—— 一场选举或国际冲突可能会导致投资者心态在一夜之间发生变化。这意味着用来预测今天股票价值的算法可能到明天就不准了。在实践中，被用来赚钱的算法有数百种，表现突出的则被不断整合以实现最优回报。

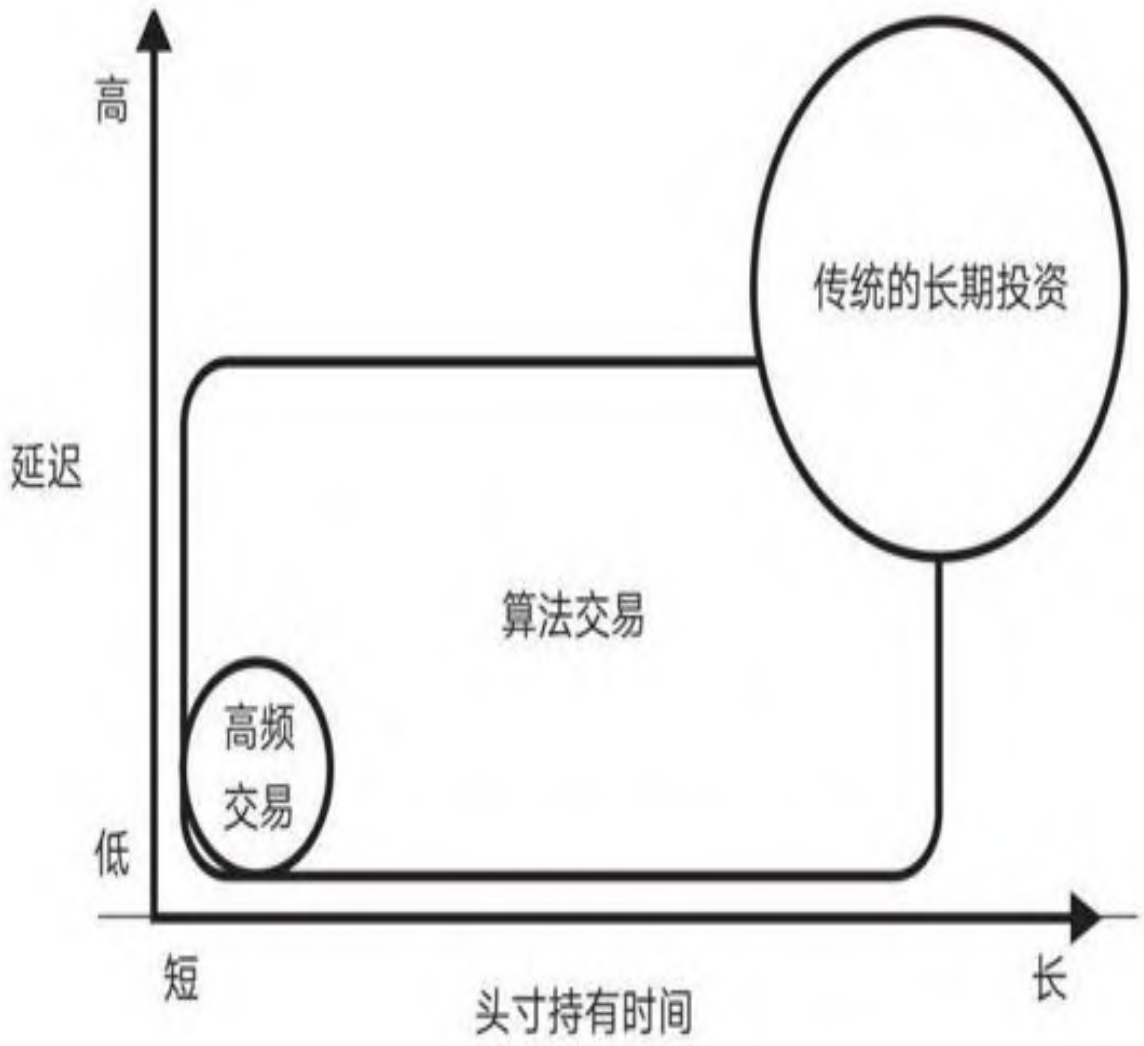


图1-6 延迟vs头寸持有时间。在线机器学习正在推动算法交易，它比传统的长期投资策略更快速，比股票市场中的高频交易更加慎重。许多不同类型的机器学习算法被组合运用以获得最佳回报。

早在20世纪80年代，我还在为摩根士丹利的股票交易神经网络模型提供咨询时，遇到了专门设计并行计算机的计算机科学家大卫·肖（David Shaw）。哥伦比亚大学学术休假期间，肖曾在自动化交易早期担任量化分析师，随后他在华尔街创立了自己的投资管理公司德劭集团（The D. E. Shaw Group），现在他已经是亿万富翁了。德劭集团非常成功，但仍然逊于另一家对冲基金文艺复兴科技公司

（Renaissance Technologies）。这家基金是由杰出的数学家、纽约州立大学石溪分校数学系前主任詹姆斯·西蒙斯（James Simons）创立的。仅2016年，西蒙斯就挣了16亿美元，这还算不上他最好的一年。[22]文艺复兴科技被称为“世界上最好的物理和数学系，”[23]“它不会雇用带有哪怕一点点华尔街正统味道的人”。[24]

不再参与德劭的日常运营后，大卫·肖现在专注于德劭研究所（D. E. Shaw Research）的业务，该研究所搭建了一台名为“Anton”的专用并行计算机，比全球其他计算机执行蛋白质折叠的速度都快得多。[25]西蒙斯退休后不再掌管文艺复兴科技，而是建立了资助自闭症及其他物理和生物科学项目研究的基金会。通过加州大学伯克利分校的西蒙斯计算理论研究所（the Simons Institute for the Theory of Computing at UC Berkeley）、麻省理工学院的西蒙斯社会大脑中心（the Simons Center for the Social Brain at MIT）和纽约熨斗研究院（the Flatiron Institute），西蒙斯的慈善事业对推进数据分析、建模和仿真的计算方法产生了重大影响。[26]

更广泛的金融服务正在金融科技（fintech）的大背景下发生大规模转型。诸如区块链这样的信息技术—— 一种安全的互联网记账方式，取代了金融交易的中间商——正在接受小规模的测试，但它很快就会扰乱价值数万亿美元的金融市场。机器学习正在被用于改进贷款信用评估，准确地提供业务和财务信息，在社交媒体上获取预测市场趋势的信号，并为金融交易提供生物识别安全服务。谁拥有最多的数据，谁就是赢家，而世界上充斥着财务数据。

深度法律：效率的提高与费用的降低

深度学习刚刚开始影响法律界。律师事务所每小时收费数百美元的法务助理的大部分日常工作都将实现自动化，特别是在高档写字楼

办公的规模化事务所里。具体点说，技术辅助审核或调查将被人工智能接管，它可以浏览数千份文件以获取合法证据，且丝毫不会感到厌倦。自动化深度学习系统也将帮助律师事务所遵守日益复杂的政府规定。这些系统将为现在无法负担律师费用的普通人提供法律建议。法律工作不仅收费会更便宜，也会更高效，这一点通常比费用更重要。法律世界正在走向“深度法律”。[27]

德州扑克：当机器智能学会了虚张声势

一对一无限注德州扑克是最受欢迎的扑克玩法之一，常见于赌场，无限注投注方式则通常出现在世界扑克系列赛（World Series of Poker）的主赛事中。扑克很有挑战性，因为与国际象棋玩家可以获得相同的信息不同，扑克玩家的信息不完整，而且在最高级别的比赛中，诈唬、欺骗的技巧和拿到的牌一样重要。

数学家约翰·冯·诺依曼（John von Neumann）创立了数学博弈理论，也是数字计算机之父，他就对扑克特别着迷。他说过：“现实生活包括虚张声势，一点欺骗手段，以及自问另一个人会怎么评判我做事的意图。这就是我理论中博弈的内涵。”[28]扑克是一种博弈，反映了经过进化精炼过的人类智能的一部分。一个名为“DeepStack”的深度学习网络和33名职业扑克选手进行了44852场比赛。令扑克专家震惊的是，它以相当大的优势，一个标准差，击败了最出色的扑克玩家，同时以四个标准差在整体上击败了全部33名玩家——多么巨大的差距（见图1－7）。[29]如果这一成就能复制到其他基于不完全信息、需要人来做判断的重要领域，比如政治学和国际关系，其影响可能是极其深远的。[30]



图1-7 一对一无限注德州扑克。强势手牌。DeepStack已经掌握了如何在高筹码扑克中虚张声势，以大比分优势击败职业扑克玩家。

AlphaGo奇迹：神经科学与人工智能的协同

2016年3月，韩国围棋界18次世界冠军获得者李世石（Lee Sedol）与DeepMind公司的AlphaGo（图1－8）——一个使用深度学习网络评估盘面形势和可能的走法的围棋程序——进行了5场比赛。[31]围棋相对国际象棋的难度，相当于国际象棋对跳棋的难度。如果

国际象棋是一场战役，那么围棋就是一场战争。一块19×19围棋棋盘比一块8×8象棋棋盘大得多，这使得在棋盘的不同部分可能同时发生多场战役。不同战役之间存在长期的相互作用，即使是专家也难以判断。围棋的合法棋局总数是10^{170}，远远超过宇宙中的原子数量。



图1-8 韩国围棋冠军李世石对战AlphaGo的5场比赛里，某一场战局中的棋盘。AlphaGo是一个通过与自己下围棋来学习的深度学习神经网络。

除了几个评估盘局并选择最佳着数的深度学习网络，AlphaGo还有一个完全不同的学习系统，用于解决时间信用分配问题：在众多步棋中，哪一步对赢得胜利有所贡献，哪一步对失败承担责任？大脑的基底神经节接收来自整个大脑皮层的投射，并投射回去，利用时间差分算法和强化学习来解决这个问题。AlphaGo使用由基底神经节进化出来的相同的学习算法，以评估最大化未来奖励的行动顺序（这一过程将在第10章中做出解释）。AlphaGo通过反复和自己下棋来学习这一技能。

AlphaGo和李世石对决的围棋比赛在亚洲得到了极高的关注。在亚洲，围棋冠军是全国性的公众人物，有着摇滚明星一样的待遇。AlphaGo早些时候击败了欧洲的围棋冠军，但是那场比赛的水平远低于亚洲的最高水平，因此李世石并没有做好打一场硬仗的心理准备。即使是开发AlphaGo的公司DeepMind，也并不清楚他们的深度学习程序到底有多强大。自上一场比赛以来，AlphaGo已经与好几个版本的自己下了数百万局的棋，然而并没有任何标准来判断它的水平到底达到了何种高度。

AlphaGo赢得了5场比赛的前3场后，许多人都感到十分震惊，因为它展现出了让人意想不到的高水准。这项比赛在韩国有很高的关注度，所有的主流电视台都对比赛进行了实况报道。AlphaGo有一些着数是革命性的。在第二场比赛的第三十八步，AlphaGo下出了精彩的一着，让李世石感到十分惊讶，他花了将近10分钟的时间才决定下一步要怎么走。AlphaGo输掉了第四场比赛，这是人类挽回颜面的一场胜利，最终它的战绩是4胜1负（图1－9）。[32]3月的夜晚，我在圣迭戈的凌晨兴致勃勃地观看了这场较量。这让我回想起1966年6月2日凌晨1点，我在克利夫兰市，坐在电视机旁关注着“勘测者1号”探测器

降落在月球上，并传回了第一张月球照片。[33]我亲眼见证了这些历史时刻。AlphaGo的表现远远超出了我和其他许多人的期待。



图1-9 在2016年3月的围棋挑战赛中输给了AlphaGo之后的李世石。

2017年1月4日，一个名为“Master”的选手在一个网络围棋服务器上主动现身，其真正身份是AlphaGo 2.0。在此之前，它在与世界顶尖棋手的比赛中取得了60场全胜的战绩，被击败的棋手中包括当时世界排名第一的高手，19岁天才棋手柯洁。AlphaGo显露出了一种能与同时代的佼佼者抗衡的全新风格。2017年5月27日，在中国乌镇举办的围棋峰会上，柯洁以3场皆负的结果输给了AlphaGo（见图1－10）。这是有史以来最精彩的几场围棋比赛，数亿中国人都观看了该赛事。“去年，我还觉得AlphaGo的表现与人类非常接近，但今天我认为它是‘围棋之神’。”柯洁这样总结道。[34]

在第一场比赛中，他以一目半的微弱差距输掉了比赛。柯洁说他“在比赛中途已经感觉快要赢了”。他非常兴奋：“我能感觉到自己的心脏在怦怦直跳！可能因为我太兴奋，有几步棋走错了。也许这就是人类棋手最薄弱的部分吧。”[35]柯洁经历了一种情绪上的超负荷，但要达到最佳状态，更需要相对沉稳的情绪。事实上，舞台演员们都知道，如果他们演出前没有胃里翻江倒海的紧张感，就无法呈现出最精彩的演出。他们的表演遵循一种倒U形曲线，即最佳状态处于较低和较高的兴奋点之间。运动员把这叫作“在状态”。



图1-10 2017年在中国，DeepMind的联合创始人兼CEO德米斯·哈萨比斯（Demis Hassabis，左）和柯洁在历史性的围棋比赛结束后会面，共同展示带有柯洁签名的

棋盘。图片来源：德米斯·哈萨比斯。

2017年5月26日，AlphaGo还击败了由5名顶尖棋手组成的队伍。这些棋手都分析过AlphaGo的招数，并已经在相应地改变自己的策略。这场比赛由中国政府主办，可以说是一个新版的“乒乓外交”。[36]中国正在机器学习方面投入大量资金，其大脑研究计划的一个主要目标是挖掘大脑潜能来创造新的算法。[37]

该围棋事件后续的发展可能会更令人惊叹。在开始跟自己下棋之前，AlphaGo是通过观察学习16万次人类围棋比赛起步的。有人认为这是作弊——一个自主的AI程序应该能够在没有积累任何人类知识的条件下学习下围棋。2017年10月，一款名为AlphaGo Zero的新版本AI程序面世了。它从游戏规则开始一步步学习下围棋，击败了曾战胜柯洁的版本AlphaGo Master，战绩为100：0。[38]此外，AlphaGo Zero的学习速度比AlphaGo Master快100倍，而计算能力差不多只有后者的1/10。完全忽略人类的知识，AlphaGo Zero变成了无敌超人。随着机器学习算法的不断进步，AlphaGo还会变得多么优秀，并没有已知的上限。

AlphaGo Zero虽然没有和人下棋，但仍然有许多围棋知识被人为添加到程序中强化棋艺的特征。如果没有任何围棋知识，AlphaGo Zero也许仍有进一步改进的空间。就像零度可乐将可口可乐里所有的热量分离出来一样，围棋的所有知识都被从Alpha Zero中剥离出来。结果，Alpha Zero能够更快、更果断地打败AlphaGo Zero。[39]为了进一步说明“少就是多”，Alpha Zero在没有改变任何一个学习参数的情况下，学会了如何以超人的水准下国际象棋，还创造了人类从未使用过的着数。在与Stockfish这个已经是超人级别的顶级国际象棋程序的对决中，Alpha Zero还没有输过。在一场比赛中，Alpha Zero大胆地牺牲了一个象——这种做法通常用来获得位置上的优势，随后又牺牲了王后，这一步看起来像是个大昏着儿，直到很多步以后，Alpha

Zero冷不防将了一军，无论是Stockfish还是人类棋手都没能预见到这样的结果。外星人已经着陆，地球从此要改头换面了。

AlphaGo的开发者DeepMind于2010年由神经学家德米斯·哈萨比斯参与创立，他曾在伦敦大学学院的盖茨比计算神经科学部门（University College London’s Gatsby Computational Neuroscience Unit）担任博士后研究员。该部门由彼得·达扬（Peter Dayan）领导，达扬曾是我实验室的博士后研究员，2017年和雷蒙德·多兰（Raymond Dolan）以及沃尔夫拉姆·舒尔茨（Wolfram Schultz）共同获得了享有盛誉的“大脑奖”（Brain Prize），以表彰他们在奖励学习方面的研究。谷歌在2014年以6亿美元的价格收购了DeepMind。该公司雇用了400多名工程师和神经科学家，拥有学术界和创业公司混合的双重文化。神经科学与人工智能之间的协同作用日渐深入，而且还在加速。

弗林效应：深度学习让人类更加智能

AlphaGo有智力吗？除了“意识”这个主题，关于智力的文章比心理学中任何其他主题都要多得多，这两个概念都很难界定。自20世纪30年代以来，心理学家就对流体智力和晶体智力进行了区分——流体智力能够将新条件中的推理和模式识别用于解决新问题，而不依赖于以前的知识；晶体智力则依赖于先前的知识，也是标准智商测试（即IQ测试）的对象。流体智力遵循一种抛物线式发展轨迹，在成年早期达到高峰，并随着年龄的增长逐渐下降；而晶体智力会随年龄的增长，缓慢渐进式地提高，直至暮年。AlphaGo只在一个相当狭窄的领域同时展现出了晶体智力和流体智力，但在这个领域，它表现出了令人惊讶的创造力。专业知识的获取也是基于在狭窄领域的学习。我们都是语言领域的专家，每天都在使用语言。

AlphaGo使用的强化学习算法可以被用来解决许多问题。这种形式的学习只取决于在一系列动作结束时给予获胜者的奖励，这似乎和提前做出更好的决策相矛盾。结合了许多强大的深度学习网络，就会生成许多领域相关的智能。而且事实上，已经出现了与领域相关的不同类型智能，例如社会、情感、机械和建筑等的案例。[40]智力测试测量的一般因素（general factor，简称g因素）与这些不同类型相关。我们有理由认真审视IQ测试。自20世纪30年代首次测试智力以来，全人类平均的IQ分数每10年会上升三个点，这一趋势被称为“弗林效应”（Flynn effect）。对于弗林效应有许多可能的解释，比如更充足的营养、更完善的医疗体系，以及其他环境因素。[41]这很有道理，因为环境会影响基因调控，从而影响大脑内在的连接，行为也会随之发生变化。[42]随着人类越来越多地生活在人造环境中，大脑正在以某种超越自然进化轨道的方式被塑造。在更长的时间内，人类是否能一直都在变得更聪明？智商增长会持续多久？用电脑玩国际象棋、西洋双陆棋和围棋的人数自计算机程序达到冠军级别后一直在稳步增加，而机器也强化了人类玩家的智能。[43]深度学习提升的将不仅仅是科学研究人员的智能，还包括所有行业从业人员的智能。

科学仪器正以惊人的速度产生数据。位于日内瓦的大型强子对撞机（LHC）中发生的基本粒子碰撞每年产生25PB（1PB=1000TB）的数据。大型综合巡天望远镜（LSST）每年将产生6PB的数据。机器学习正被用于分析庞大的物理和天文数据集，其规模之浩大让人类根本无法通过传统方法进行搜索。[44]例如，DeepLensing是一种神经网络，可以识别遥远星系的图像。这些图像由于光在传播中因围绕周边星系的“引力透镜”造成的光路偏折而被扭曲了。这一技术可以自动发现许多遥远的新星系。物理学和天文学中还有许多其他类似“大海捞针”的问题，而深度学习能够让传统的数据分析方法如虎添翼。

新教育体系：每个人都需要终身学习

银行在20世纪60年代后期推出了面向银行账户持有人的全天候现金提取服务，这对于那些在银行正常营业时间之外需要现金的人来说非常方便，自动提款机（ATM）从此获得了阅读手写支票的能力。尽管它们的存在减少了银行柜员的日常工作量，但有越来越多的柜员为客户提供按揭和投资建议等个性化服务，同时也出现了维修ATM的新工种。[45]就如一方面，蒸汽机代替了体力劳动者，但另一方面，这为能够建造和维护蒸汽机及驱动蒸汽机车的熟练工人提供了新的就业机会。亚马逊的在线营销也将许多员工从当地实体零售店中迁移出来，但同时也为分配和运输其商品，以及许多使用其平台的企业创造了38万个新的工作机会。[46]由于现在需要人类认知技能的工作被自动化人工智能系统所接管，那些能够创建和维护这些系统的人将会获得新的工作。

工作变动不是什么新鲜事。19世纪，农场劳工被机器取代，机器也在城市工厂创造了新的工作机会，所有这些都需要一个教育系统来培训工人新的技能。不同之处在于，今天，由人工智能开辟的新职位除了需要传统的认知技能之外，还需要新的、不同的、不断变化的技能。[47]所以我们都需要终身学习。要做到这一点，我们需要一个以家庭，而不是以学校为基础的新教育体系。

幸运的是，就像寻找新工作的需求变得迫在眉睫一样，互联网上免费的大规模开放式在线课程慕课（MOOCs）也应运而生，来帮助人们获取新的知识和技能。虽然仍处于初级阶段，但慕课的在线教育生态系统正在迅速发展，并在为更广泛的人群提供前所未有的优质教学。与下一代数字辅助系统相结合，慕课则可能会带来变革。芭芭拉·奥克利（Barbara Oakley）和我开设了一门名为“学会如何学习”（Learning How to Learn）的慕课——该热门课程会教你如何成为更好的学习者（见图1-11）——以及一门名为“思维转换”（Mindshift）的慕课，教你如何改造自己并改变你的生活方式（这两门课将在第12章中详细介绍）。



图1-11 “学会如何学习”教你如何成为更好的学习者，它是互联网上最受欢迎的慕课，拥有超过300万学习者。

进行网上操作时，其实正在生成机器可读的关于你自己的大数据。根据你在互联网上行为的蛛丝马迹，你正在被自动生成的相关广告定位。你在Facebook（脸谱网）和其他社交媒体网站上发布的信息可被用于创建数字助理，它几乎比世界上任何其他人都更了解你，并且不会遗漏任何内容，实际上就相当于你的虚拟分身。通过将互联网跟踪和深度学习都纳入服务，现在这些孩子的后代拥有的教育机会将比今天富裕家庭拥有的最优质的教育机会还要好。这些孙辈将拥有自己的数字导师，导师将在整个教育过程中陪伴他们。教育不仅会变得更加个性化，也会变得更加精准。世界各地已经开展了各种各样的教育实验，例如可汗学院，由盖茨基金会、陈－扎克伯格基金会和其他慈善基金会资助。这些实验机构正在测试软件，以便让所有的孩子都可以根据自己的节奏进步，并适应每个儿童的特定需求。[48]数字导师的普及将使教师从教学中的重复劳动，如评分中解脱出来，专注于人类最擅长的事情——对学习困难的学生提供精神支持，并给予有天赋的学生灵感启发。教育技术（Edtech）正在快速发展，与自动驾驶汽车相比，传统教育向精准教育过渡的速度可能相当快，因为它必须克服的障碍要小得多，需求却要大得多，而且美国的教育是一个万亿

美元的市场。[49]一个主要的问题就是，谁能够访问数字助理和数字导师的内部文件。

正面影响：新兴技术不是生存威胁

AlphaGo在2016年毫无争议地击败了李世石，这激化了过去若干年引发的人工智能可能给人类带来威胁的担忧。计算机科学家签署了不会将AI用于军事目的的承诺协议。斯蒂芬·霍金（Stephen Hawking）和比尔·盖茨（Bill Gates）公开发表声明，警告人工智能可能对人类造成的生存威胁。伊隆·马斯克（Elon Musk）和其他硅谷企业家成立了一家新公司OpenAI，拥有10亿美元储备金，并聘请了杰弗里·辛顿之前的一名学生伊利娅·苏特斯科娃（Ilya Sutskever）担任第一任总监。虽然OpenAI的既定目标是确保未来人工智能的发现将公开供所有人使用，但它还有另一个隐含的更重要的目标：防止私人公司作恶。AlphaGo战胜了围棋世界冠军李世石，一个临界点也随之到来。几乎在一夜之间，人工智能从一项失败的技术，转变成了可感知的生存威胁。

一种新兴技术被看作生存威胁，这已经不是第一次了。核武器的发明、改进和储存曾经是一种毁灭全世界的威胁，但至少到目前为止，我们有能力阻止这种情况的发生。重组DNA技术刚问世的时候，人们担心经人工改造的致命生物会从实验室逃出来，导致全球范围内出现难以估量的痛苦和死亡。基因工程现在已经是一项成熟的技术，目前我们已经能和它的产物共存。与核武器和致命生物相比，机器学习的最新进展构成的威胁相对较小。我们也将适应人工智能。事实上，这已经在发生了。

DeepStack的成功带来的其中一个暗示是，深度学习网络可以学习如何成为世界顶级的骗子。训练深层网络能干什么只受限于训练者的想象力和数据。如果一个网络可以接受安全驾驶汽车的训练，那么它

也可以被训练驾驶F 1赛车，很可能有人愿意为此掏腰包。今天，我们仍然需要技术娴熟和训练有素的从业人员使用深度学习来搭建产品和服务，但随着计算能力的成本持续下降、软件功能更加自动化，很快，高中生就可能具备开发AI应用程序的能力了。作为德国收入最高的在线电子商务公司，奥托（Otto）主要经营服装、家居和体育用品。它正在利用深度学习，根据历史订单信息预测客户未来可能购买的产品，并提前为他们下单。[50]客户几乎在订购前就收到了自己想订购的商品，准确率达到90%。自动完成工作且无须人工干预，这种预订操作不仅可以每年为公司在剩余库存和退货环节节省数百万欧元，还提高了客户满意度和保有率。深度学习显著提高了奥托公司的生产力，却并没有取代它的工人。人工智能可以让你在工作中更高效。

虽然主要的高科技公司开拓了深度学习的应用，但机器学习工具已经普遍存在了，许多其他公司也开始从中受益。Alexa是一个广受欢迎的数字助理，与亚马逊Echo智能音箱配合使用，能够基于深度学习对自然语言发出的请求做出回应。亚马逊网络服务（AWS）引入了名为“Lex”、“Poly”和“Comprehend”的工具箱，可以分别基于自动化文字、语音转换、语音识别和自然语言理解，方便地开发相同的自然语言界面。具有对话交互能力的应用程序现在可供无力雇用机器学习专家的小型企业使用。企业通过应用这一程序可以提高客户满意度。

当最好的人类棋手在计算机程序面前都黯然失色时，人类会不再下棋吗？正相反，人工智能会提高人类的竞技水平，也使得棋类竞技更加大众化。顶级的国际象棋选手曾经都来自莫斯科和纽约等大城市。这些地方大师云集，可以教授年轻棋手并提高他们的技能水平。国际象棋电脑程序使得在挪威小镇长大的马格努斯·卡尔森（Magnus Carlson）13岁就成为国际象棋大师，如今他已是世界国际象棋冠军。人工智能不仅对游戏产生了正面的影响，更会推动人类付诸努力的各个方面，从艺术到科学。AI可以让你变得更聪明。[51]

回到未来：当人类智能遇到人工智能

本书有两个相互交织的主题：人类智能是如何进化的，以及人工智能会如何演变。这两种智能之间的巨大差异在于，人类智能的进化经历了数百万年的时间，而人工智能在最近几十年才发展起来。尽管对于文化演变来说，这个速度仍然是快得出奇，但是过于谨小慎微可能并不是个正确的选择。

深度学习在近期取得的突破，并不是你从新闻报道中读到的那种一夜成功。从基于符号、逻辑和规则的人工智能向基于大数据和学习算法的深度学习网络的转变，其背后的故事通常并不为人所熟知。本书介绍了这个故事，并从我的角度探讨了深度学习的起源和成果。作为20世纪80年代开发神经网络学习算法的先行者和NIPS基金会的主席，我亲身经历了过去30年机器学习和深度学习的发展过程。我和同在神经网络领域的同事多年来都未能取得令人瞩目的成就，但坚持和耐心最终给我们带来了回报。

---

[1] “啊，美丽的新世界，有这么美的人在里头！”（“O brave new world that has such people in’t!”）来自莎士比亚的剧作《暴风雨》中米兰达的台词（5.1.182–183 [Oxford Standard Authors Shakespeare]）。

[2] Bill Vlasic, “G.M. Wants to Drive the Future of Cars That Drive Themselves,” *NewYork Times*, June 4, 2017, https://www.nytimes.com/2017/06/04/business/generalmotors-self-driving-cars-mary-barra.html.

[3] “Full Tilt: When 100% of Cars Are Autonomous,” *New York Times Magazine*, November 8, 2017, https://www.nytimes.com/interactive/2017/11/08/magazine/tech-design-autonomous-future-cars-100-percent-augmented-realitypolicing.html?hp&action=click&pgtype=Homepage&clickSource=storyheading&module=second-column-region&region=top-news&WT.nav=top-news/.

[4] Christopher Ingraham, “The Astonishing Human Potential Wasted on Commutes,” *Washington Post*, February 24, 2016,

https://www.washingtonpost.com/ news/wonk/wp/2016/02/25/how-much-of-your-life-youre-wasting-on-your-commute/?utm_term=497dfd1b5d9c..

[5] “Full Tilt: When 100% of Cars Are Autonomous,” *New York Times Magazine*,November 8, 2017, https://www.nytimes.com/interactive/2017/11/08/magazine/tech-design-autonomous-future-cars-100-percent-augmented-realitypolicing.html ?hp&action=click&pgtype=Homepage&clickSource=storyheading&module=second-column-region&region=top-news&WT.nav=top-news/.

[6] “Google’s Waymo Passes Milestone in Driverless Car Race,” *Financial Times*,December 10, 2017, https://www.ft.com/content/dc281 ed2-c425-11 e7-b2bb322b2cb39656/.

[7] B. A. Golomb, “Will We Recognize It When It Happens?” in Brockman, J., (ed.),*What to Think About Machines That Think*(New York: Harper Perennial, 2015),533– 535.

[8] Pierre Delforge, “America’s Data Centers Consuming and Wasting Growing Amounts of Energy,” Natural Resources Defense Council Issue Paper, February 6, 2015, https://www.nrdc.org/resources/americas-data-centers-consuming-andwasting-growing-amounts-energy/.

[9] W. Brian Arthur, “Where Is Technology Taking the Economy?” *McKinsey Quarterly,*October, 2017, https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/Where-is-technology-taking-the-economy/.

[10] 乞力马扎罗山是一座雪山，高19 710英尺，据说是非洲最高的山峰。它的西峰被马赛人称作“Ngaje Ngai”，意为“上帝的家”。靠近西峰有一具干燥、冰冻的豹子尸体。没有人解释过这只豹子在那个高度是要寻找什么。

[11] 乞力马扎罗山是一座高达19 710英尺的雪山，据说是非洲最高的山峰。它的西峰在马赛语里被称为“Ngaje Ngai”，上帝的家。西峰附近有一具干燥、冷冻的豹子尸体。从来没有人解释过豹子在那个高度想要找什么。

[12] Gideon Lewis-Kraus, “The Great A. I. Awakening,” *New York Times Magazine*,December 14, 2016, https://www.nytimes.com/2016/12/14/magazine/the-greatai-awakening.html.

[13] Aleksandr Sergeevich Pushkin, *Eugene Onegin: A Novel in Verse*, 2nd ed., trans.Vladimir Nabokov (Princeton: Princeton University Press, 1991).

[14] 关于这一做法的早期尝试，请参阅Andrej Karpathy, “The Unreasonable Effecti-veness of Recurrent Neural Networks,” *Andrej Karpathy Blog*, posted

May 21, 2015, http://karpathy.github.io/2015/05/21/rnn-effectiveness/.

[15] G. Hinton, L. Deng, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, et al., “Deep Neural Networks for Acoustic Modeling in Speech Recognition,” *IEEE SignalProcessing Magazine*29, no. 6 (2012): 82–97.

[16] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, et al., “Achieving Human Parity in Conversational Speech Recognition,” Microsoft Research Technical Report MSR-TR-2016-71, revised February 2017, https://arxiv.org/pdf/1610.05256.pdf.

[17] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko J, S. M. Swetter, H. M. Blau, and S.Thrun, “Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks,” *Nature*542, no. 7639 (2017): 115–118.

[18] Siddhartha Mukherjee, “A.I. versus M.D.: What Happens When Diagnosis Is Automated?” *New Yorker*, April 3, 2017, http://www.newyorker.com/magazine/2017/04/03/ai-versus-md/.

[19] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, Andrew H. Beck,*Deep Learning for Identifying Metastatic Breast Cancer*, arXiv:1606.05718. 他们使用的测量方法在信号检测理论里被称为“曲线下面积”，它对假阴性和假阳性都很敏感。https://arxiv.org/abs/1606.05718/.

[20] Anthony Rechtschaffen and Alan Kales, eds., *A Manual of StandardizedTerminology, Techniques and Scoring System for Sleep Stages of Human Subjects*,National Institutes of Health publication no. 204 (Bethesda, MD: U.S. National Institute of Neurological Diseases and Blindness, Neurological Information Network, 1968).

[21] 参阅 Ian Allison, “Former Nuclear Physicist Henri Waelbroeck Explains How Machine Learning Mitigates High Frequency Trading,” *International Business Times*, March 23, 2016, http://www.ibtimes.co.uk/former-nuclear-physicist-henriwaelbroeck-explains-how-machine-learning-mitigates-high-frequency-1551097/;Bailey McCann, “The Artificial-Intelligent Investor: AI Funds Beckon,” Wall Street Journal, November 5, 2017, https://www.wsj.com/articles/the-artificialintelligent-investor-ai-funds-beckon-1509937622/.

[22] Sei Chong, “Morning Agenda: Big Pay for Hedge Fund Chiefs despite a Rough Year,” *New York Times*, May 16, 2017, https://www.nytimes.com/2017/05/16/business/dealbook/hedge-funds-amazon-bezos.html.

[23] 除了雇用了数千名数学家的美国国家安全局。Alfred W. Hales, personal communication, May 4, 2016.

[24] Sarfaz Manzoor, “Quants: The Maths Geniuses Running Wall Street,” *Telegraph,* July 23, 2013, http://www.telegraph.co.uk/finance/10188335/Quants-the-mathsgeniuses-running-Wall-Street.html.

[25] D. E. Shaw, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. Ho, et al., “Anton: A Special-Purpose Machine for Molecular Dynamics Simulation,” *Communications of the ACM*51, no. 7 (2008): 91–97.

[26] D. T. Max, Jim Simons, “The Numbers King,” *New Yorker*, December 18 &25, 2017, https://www.newyorker.com/magazine/2017/12/18/jim-simons-thenumbers-king/.

[27] 很快就会被拍成电影。

[28] 约翰·冯·诺依曼，引自根据雅各布·布罗诺夫斯基（Jacob Bronowski）的著作改编的电视纪录片《人类的攀升》（*The Ascent of Man*），第13集（1973年）。

[29] 参阅M. Moravík, M. Schmid, N. Burch, V. Lisý, D. Morrill, N. Bard, et al.,“Deep-Stack: Expert-Level Artificial Intelligence in Heads-Up No-Limit Poker,” *Science*356, no. 6337 (2017): 508–513。标准偏差是钟形曲线的半峰宽。只有16% 的样本落入平均值的一个标准偏差之外的区域。只有万分之三的样本在平均值的四个标准偏差以外。

[30] 脑海中浮现出了1983年科幻电影《战争游戏》（*War Game*s）中的场景。参阅https://en.wikipedia.org/wiki/WarGames。

[31] 参阅D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. v. d. Driessche,et al., “Mastering the Game of Go with Deep Neural Networks and Tree Search,” *Nature*529, no. 7587 (2016): 484–489.

[32] “我不知道今天该说些什么，”李世石告诉媒体记者，“但我想我必须先表示歉意。我应该能表现得更好，在比赛中占据更多的主动。我也很抱歉，让很多人失望了。我觉得很无助。如果回顾三场比赛，即使第一场比赛能够重赛，我也不认为自己会取胜，因为那时我误判了AlphaGo的能力。” Jordan Novet, “Go Board Game Champion Lee Sedol Apologizes for Losing to Google’s AI,” *Venture Beat*, March, 12, 2016, https://venturebeat.com/2016/03/12/goboard-game-champion-lee-sedol-apologizes-for-losing-to-googles-ai/。

[33] “勘测者1号”于1966年6月2日，国际标准时间6点17分36秒（美国东部时间凌晨1点17分36秒）登陆月球表面。着陆点位于弗拉姆斯蒂德火山口（Flamsteed Crater）以北一个直径为100公里的陨坑内。

[34] Ke Jie, as quoted in Selina Cheng, “The Awful Frustration of a Teenage Go Champion Playing Google’s AlphaGo,” *Quartz*, May 27, 2017, https://qz.com/993147/the-awful-frustration-of-a-teenage-go-champion-playing-googlesalphago/.

[35] Ke Jie, as quoted in Paul Mozur, “Google’s A.I. Program Rattles Chinese Go Master As It Wins Match,” *New York Times*, May 25, 2017, https://www.nytimes.com/2017/05/25/business/google-alphago-defeats-go-ke-jie-again.html.

[36] 老版的“乒乓外交”指1971年中国和美国乒乓球队开展互访的友好往事，不仅推动了中美两国关系正常化的进程，也加速了中国走向世界的步伐。——编者注

[37] Paul Mozur, “Beijing Wants A.I. to Be Made in China by 2030,” *New York Times*,July 20, 2017, https://www.nytimes.com/2017/07/20/business/china-artificialintelligence.html.

[38] Silver D., J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T.Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, “Mastering the Game of Go Without Human Knowledge,” *Nature*550 (2017): 354–359.

[39] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, Demis Hassabis, *Mastering Chess and Shogiby Self-Play with a General Reinforcement Learning Algorithm*, arXiv:1712.01815(2017).

[40] Harold Gardner, *Frames of Mind: The Theory of Multiple Intelligences*, 3rd ed. (New York: Basic Books, 2011).

[41] J. R. Flynn, “Massive IQ Gains in 14 Nations: What IQ Tests Really Measure,”*Psychological Bulletin*101, no. 2 (1987):171–191.

[42] S. Quartz and T. J. Sejnowski, *Liars, Lovers and Heroes: What the New Brain Science Has Revealed About How We Become Who We Are*(New York: Harper Collins, 2002).

[43] Douglas C. Engelbart, *Augmented Intelligence: Smart Systems and the Future of Work and Learning*, SRI Summary Report AFOSR-3223 (Washington, DC:Doug Engelbart Institute, October 1962), http://www.dougengelbart.org/pubs/augment-3906.html.

[44] M. Young, “Machine Learning Astronomy,” *Sky and Telescope*, December (2017):20–27.

[45] “Are ATMs Stealing Jobs?” *The Economist*, June 15, 2011, https://www.economist.com/blogs/democracyinamerica/2011/06/technology-and-unemployment/.

[46] John Taggart and Kevin Granville, “From ‘Zombie Malls’ to Bonobos: What America’s Retail Transformation Looks Like,” *New York Times*, April 15, 2017.

[47] E. Brynjolfsson and T. Mitchell, “What Can Machine Learning Do? Workforce Implications,” *Science*(2017): 358:1530–1534. doi: 10.1126/science.aap8062.

[48] “Technology Is Transforming What Happens When a Child Goes to School:Reformers Are Using New Software to ‘Personalise’ Learning,” *Economist*, July 22, 2017, https://www.economist.com/news/briefing/21725285-reformers-areusing-new-software-personalise-learning-technology-transforming-what-happens/.

[49] 教育市场的估值超过1.2万亿美元，主要包含三个领域：早期儿童教育（700亿美元），K–12（即从幼儿园到高中三年级的基础教育，6 700亿美元），以及高等教育（4750亿美元）。参阅Arpin Gajjar, “How Big Is the Education Market in the US: Report from the White House,” *Students for the Future*, October 10, 2008, https://medium.com/students-for-the-future/how-big-is-the-educationmarket-in-the-us-report-from-white-house-91dc313257c5。

[50] “Algorithmic Retailing: Automatic for the People,” *Economist*, April 15, 2017,

[51] T. J. Sejnowski, “AI Will Make You Smarter,” in Brockman, J. (ed.), *What to ThinkAbout Machines That Think*(New York: Harper Perennial, 2015), 118–120.

02
人工智能的重生

马文·明斯基是一位杰出的数学家，也是麻省理工学院人工智能实验室（以下简称MIT AI Lab）[1]的创始人之一。创始团队确定了这个领域的方向和文化。麻省理工学院的人工智能研究在20世纪60年代成了聪明人的阵地，这在很大程度上要归功于明斯基。他每分钟冒出的新点子比我认识的任何人都多，而且能让你相信他对一个问题的看法是对的，尽管这个看法会让人觉得有悖常识。我很钦佩他的大胆和睿智，但是并不认同他研究AI所选择的方向。

看似简单的视觉识别

积木世界（Blocks World）是MIT AI Lab在20世纪60年代推出的一个项目典范。为了简化视觉问题，积木世界由矩形积木组成，可以堆叠起来组成新的结构（见图2－1）。该项目的目标是编写一个能够理解命令的程序，例如“找到一个大的黄色积木并将其放在红色积木上面”，并规划出让机器人手臂执行命令所需的步骤。这看起来像小孩子玩的游戏，但却需要编写一个庞大而复杂的程序来实现。这个程序后来变得十分冗长烦琐，以至于编写该程序的学生特里·维诺格拉德（Terry Winograd）离开该实验室之后，该程序因为错误百出，频频崩溃，最终被无奈地放弃了。这个看起来很简单的问题比任何人想象的都要难得多，即使成功了，也很难把积木世界同现实世界顺利地连接起来，毕竟在现实世界中物体有不同的形状、大小和重量，而且

并非所有边角都是直角。相比于照明的方向和位置都能固定的可控实验环境，在现实世界中，照明可能因地点和时间的不同而发生显著变化，这极大地增加了计算机物体识别任务的复杂性。

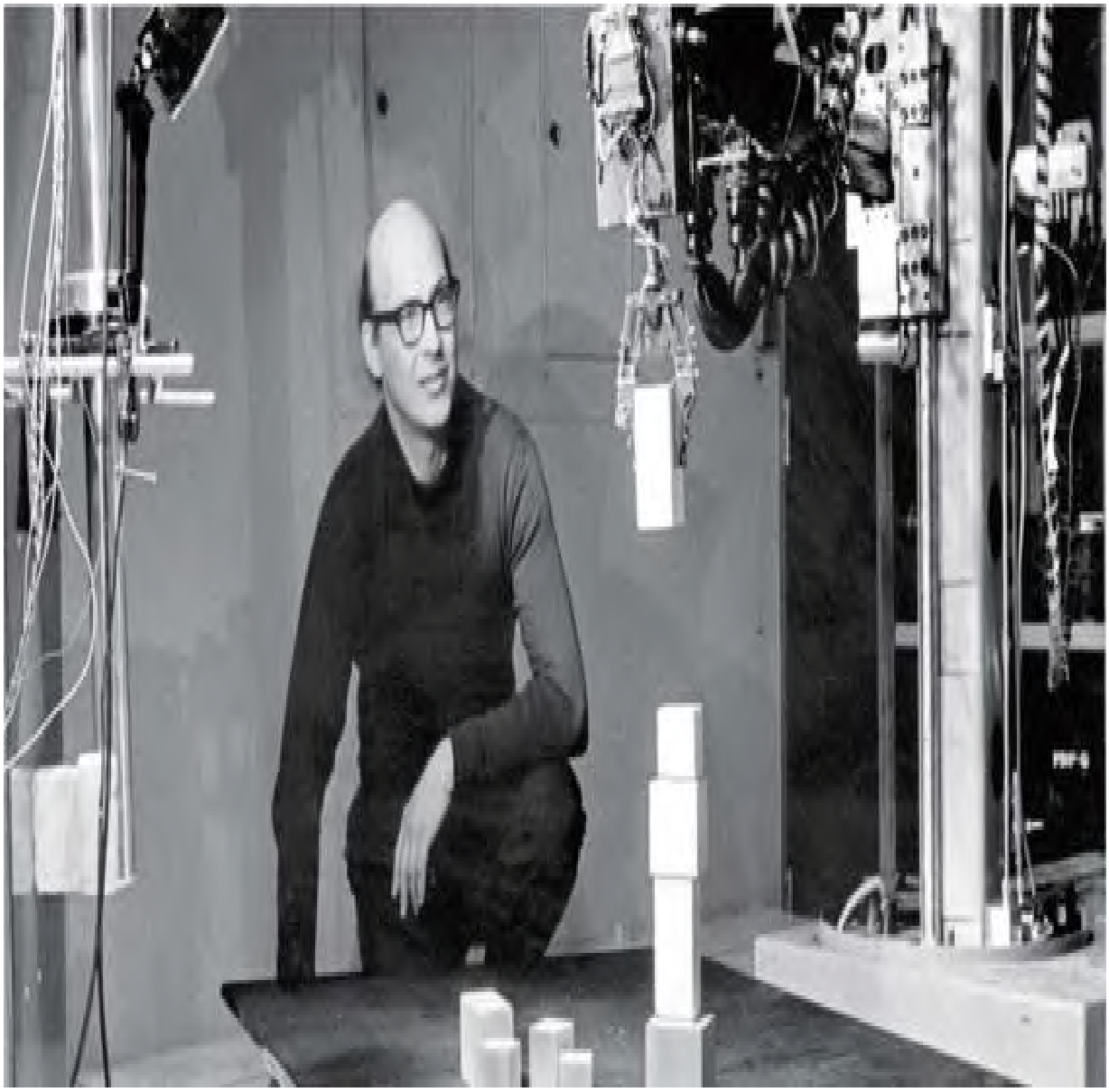


图2-1 马文·明斯基在观察机器人堆积木（照片拍摄于1968年左右）。积木世界是我们如何与世界交互的简化版本，但它面对的问题比任何人想象的都要复杂得多，直到2016年才通过深度学习解决。

20世纪60年代，MIT AI Lab获得了来自一个军事研究机构的大笔资金，被要求用于打造一个可以打乒乓球的机器人。我曾经听过一个故事，那就是这个项目的负责人忘了在基金申请书中加上为机器人建立一个视觉系统的预算，于是他把这个问题分配给了一个研究生作为暑期研究项目。后来我问马文·明斯基这个故事是不是真的。他反驳道：“我们把那个问题分配给了本科生。”MIT档案中的一份文件证实了他的回答。[2]这个看上去很容易解决的问题，最终被证明是个“陷阱”，吞噬了整整一代计算机视觉研究人员的青春。

计算机视觉的进步

尽管物体的位置、大小、方向和受到的光照不同，我们却很少在识别物体时感到吃力。计算机视觉研究中最早的想法之一是将物体的模板与图像中的像素进行匹配，但是这种方法收效甚微，因为同一物体不同角度的两个图像中的像素并不匹配。例如，参考图2－2中的两只鸟。如果你将一只鸟的影像覆盖到另一只鸟上，你可以找到一块匹配的部分，但其余部分就完全对不上了；但是如果有另一个种类的鸟相同姿势的图像，你却可以得到相当好的匹配结果。

计算机视觉的进步是通过关注特征而非像素来实现的。例如，赏鸟者必须具备专业水平才能区分只在一些细微处略有差异的不同鸟类。一本关于鸟类鉴别的实用畅销书中只有一张鸟的照片，却有许多示意图指出了各种鸟之间的细微差别（见图2－3）。[3]一个好的特征是指一种鸟类独有的特征，但是如果在别的种类中也可以找到这些特征，那么就要靠翼带、眼纹和翼斑的独特标记组合来区分。当这些标记组合为近亲种类所共享时，就要根据叫声和歌声进一步区分。鸟类的草图或彩绘能更好地将我们的注意力引导到相关的区别特征上，相比之下，鸟类照片里则布满了数百个不太相关的特征。



图2-2 两只斑胸草雀在互相审视对方。我们不难看出它们是同一物种。但是因为它们面对镜头的角度不同，所以很难将它们与模板对应，即使它们具有几乎相同的特征。

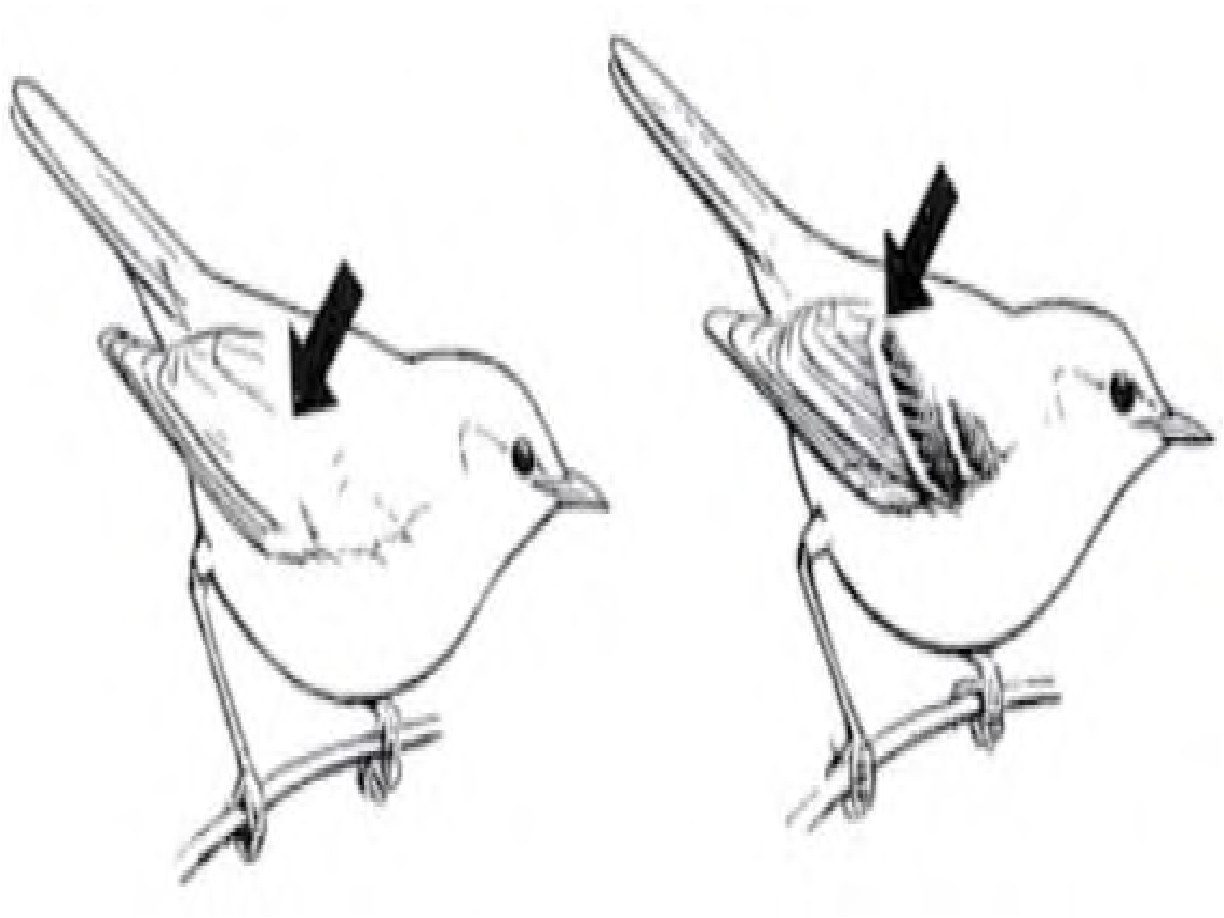


图2-3 可用于区分相似鸟类的显著特征。箭头指向的是翼带的位置，对于分辨莺科非常重要：有的轮廓清晰，有的界线模糊，有的是双条的，有的长，有的短。图片来源：Peterson, Mountfort, and Hollom, *Field Guide tothe Birds of Britain and Europe*, 5th ed., p.16。

这种基于特征的识别方法存在的问题，不仅在于针对世界上数万种不同物体开发特征检测器是非常耗费人力的，而且即便使用最精确的特征检测器，被部分遮挡住的物体的图像也会产生歧义，这使得识别混乱场景中的物体成了计算机所面临的一项艰巨任务。

20世纪60年代，没有人能想到我们要花上50年，计算机的运算能力需要提升100万倍，才能让计算机视觉达到人类的水平。当时有一种带有误导性的直觉，认为编写计算机视觉程序很容易。这种直觉是基于我们认为很简单的行为，例如看、听、四处走动——但这些行为是经过了几百万年的自然进化才实现的。让早期AI先驱十分懊恼的是，

他们发现计算机视觉问题非常难以解决。相比之下，他们发现通过编写程序让计算机证明数学定理要容易得多——这个过程曾被认为需要最高水平的智能——因为计算机处理逻辑问题的能力比人类要强得多。逻辑思维是进化后期的产物，即便对于人类，也需要接受从逻辑命题到得出严谨结论的长期训练。然而，对于大多数我们所面临的生存问题，从以往经验中总结出的解决方案，在大部分时间都能发挥作用。

早期人工智能发展缓慢

20世纪七八十年代流行的人工智能专家系统，被开发用以通过遵循一套规则来解决医学诊断等问题。因此，一个早期的专家系统MYCIN，被开发用于识别导致感染性疾病如细菌性脑膜炎的细菌。[4] 根据专家系统方法，MYCIN的开发者首先要收集由传染病专家提供的事实和规则，以及病人的症状和病史，然后将信息输入系统中的计算机，最后通过编写程序让计算机使用逻辑进行推理。然而，开发者在收集专家提供的事实和规则时遇到了麻烦，尤其是在更复杂的领域，最好的诊断医师并不依赖规则，而是依靠难以编码的基于经验的图案识别，[5] 而且他们的系统必须随着新发现的诞生和旧规则的过时而不断更新。开发者在收集和输入病人的症状和病史信息时遇到了更多的困难。录入每个病人的信息需要花上半小时或更长的时间，忙碌的医生们根本腾不出这么多时间。因此，MYCIN从未被应用于临床是预料之中的事。尽管有许多针对其他应用的专家系统被开发了出来，例如有毒物质泄漏管理、自动驾驶车辆的任务规划，以及语音识别等，但如今，这些系统在实际生活中很少被用到。

研究人员在人工智能发展早期尝试了许多不同的方法，这些方法通常都很巧妙，但并不实用。他们不仅低估了现实世界问题的复杂性，而且提出的解决方案无法进行大规模应用。在复杂的领域中，规

则的数量可能会非常庞大，并且随着新的事实的出现而不断被添加，跟踪所有规则的例外情况并与之互动变得十分不切实际。例如，道格拉斯·勒奈（Douglas Lenat）在1984年启动了一个名为“Cyc”的项目，想要将常识代码化，这在当时似乎是一个好主意，但实际操作起来却成了一场噩梦。[6]我们对世界上的大量事实都觉得理所当然，其中大部分都是基于经验。例如，从40英尺的高度落下的猫可能会毫发无损，[7]但是从同一高度坠落的人就不同了。

早期AI进展得如此缓慢的另一个原因是，数字计算机还处于非常原始的阶段，并且以今天的标准来看存储器成本十分高昂。但是由于数字计算机在逻辑运算、符号操作和规则应用方面非常高效，这些计算原语在20世纪会受到青睐并不令人惊讶。这也促使了卡内基－梅隆大学（Carnegie Mellon University）的两位计算机科学家艾伦·纽维尔（Allen Newell）和赫伯特·西蒙（Herbert Simon）于1955年写成了一套名为“逻辑理论家”的计算机程序，它可以证明《数学原理》（*Principia Mathematica*）一书中的逻辑定理。《数学原理》是阿尔弗雷德·怀特海德（Alfred North Whitehead）和伯特兰·罗素（Bertrand Russell）尝试把所有的数学知识系统化的一套书。这一时期，人们对于智能电脑即将到来寄予了厚望。

那些试图编写具有人类智能的计算机程序的AI先驱，本身并不太关心大脑是如何实现智能行为的。我问艾伦·纽维尔其中的缘由，他告诉我，他本人对大脑研究的见解持开放态度，但当时根本没有足够的相关信息可以借鉴。大脑功能的基本原理在20世纪50年代才慢慢开始被揭开，相关领域的带头人包括艾伦·霍奇金（Alan Hodgkin）和安德鲁·赫胥黎（Andrew Huxley），他们解释了大脑信号是如何通过神经中的电脉冲来实现远距离传送的，而另一带头人伯纳德·卡茨（Bernard Katz）则发现了这些电子信号如何经突触转化为化学信号来实现神经元间的通信。[8]

到了20世纪80年代，人们对大脑的了解日益增加，并且在生物学领域以外对大脑的接触也变得更加广泛，但对于新一代的AI研究人员来说，大脑本身已经变得无关紧要了。他们的目标是编写一个程序，使其拥有和大脑一样的功能。在哲学里，这种立场被称为“功能主义”，对许多人来说，这是一个可以忽略生物学中杂乱细节的绝佳说辞。但是一小群不属于主流群体的AI研究人员认为，受大脑生物学启发的那些被称为“神经网络”、“连接主义”和“并行分布处理”的AI实现方法，会最终解决困扰基于逻辑的AI研究的难题。我正是那一小群人中的一个。

从神经网络到人工智能

1989年，MIT计算机科学实验室主任迈克尔·德图佐斯（Michael Dertouzos）邀请我到MIT做一个有关我在基于神经网络的人工智能领域采用的开拓性研究方法的“杰出学者讲座”（图2－4）。到场后，我受到了德图佐斯的热情接待。乘电梯时他告诉我：“MIT有个传统，杰出学者讲座的演讲者需要在午餐时用5分钟的时间与教师和学生一起讨论他的讲座话题。”“而且，”当电梯门打开时，他补充道，“他们挺反感你在做的事情。”

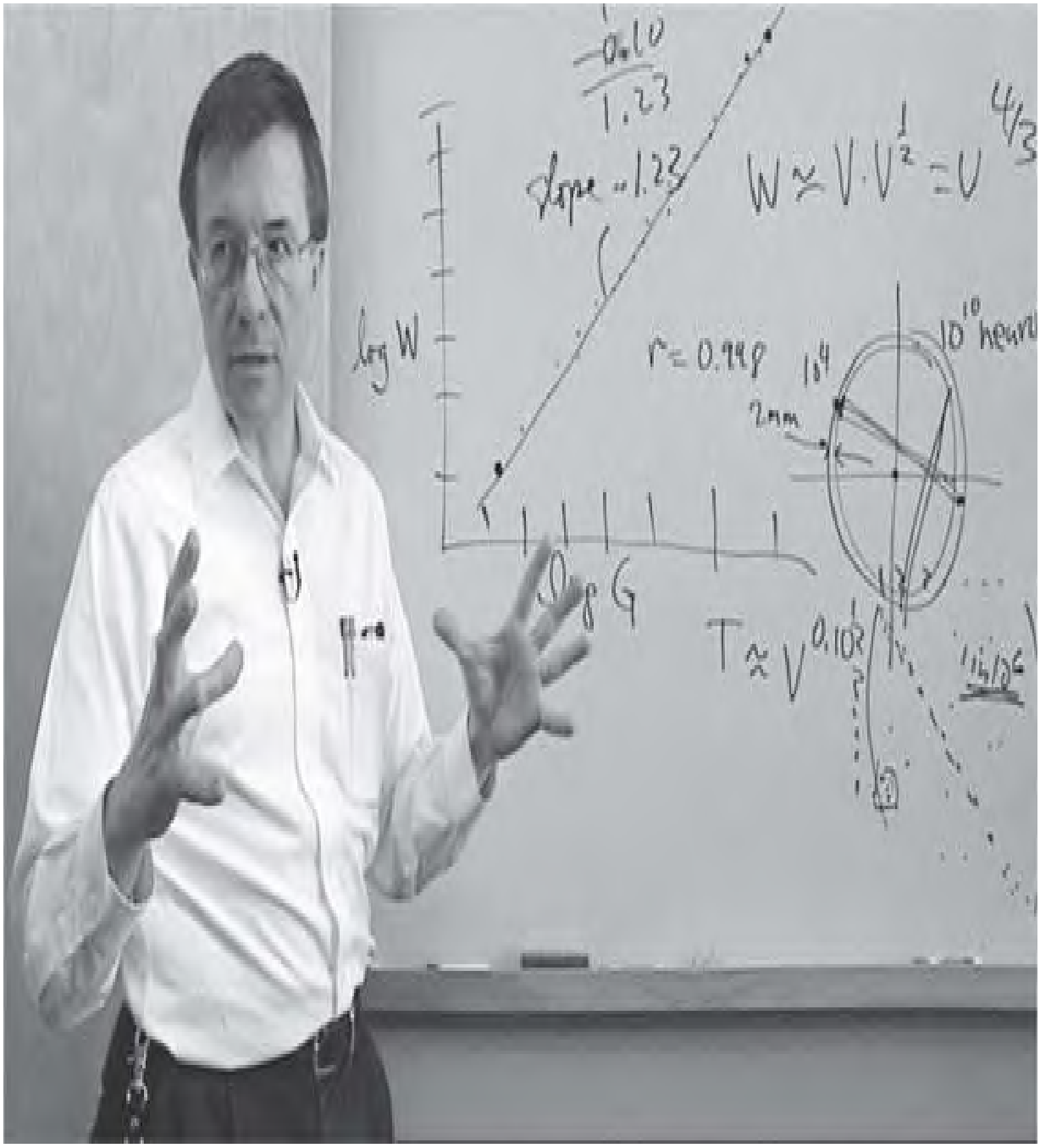


图2-4 我在介绍大脑皮层的比例法则，那是在1989年我加入索尔克生物研究所（Salk Institute）不久后。图片来源：Ciencia Explicada。

房间里挤满了差不多一百人，德图佐斯也很意外。科学家们站成一个圈，足足有三层：高级教员在第一层，第二层是初级教员，学生

在最后一层。我走到圆圈的中心，正对着自助餐的主菜。我要在这5分钟里说些什么，才能让那些讨厌我工作的人改变想法呢？

我开始即兴发挥，“食物上这只苍蝇的大脑只有10万个神经元；它大概重1毫克，要消耗1毫瓦的能量，”我边说边驱赶苍蝇，“苍蝇能看，能飞，可以自己确定飞行方向，还能觅食。但最不可思议的是，它可以通过繁殖来进行自我复制。MIT拥有一台价值1亿美元的超级计算机：它消耗的能量是兆瓦级的，并需要一台巨型空调进行冷却。但是，超级计算机的最大成本是要消耗大量人力，也就是说程序员要满足它对程序的巨大需求。这台超级计算机不能看，不能飞，虽然它能与其他计算机交流，但它不能交配或自我复制。看得出这个场景描述里有什么问题吗？”

在一段长时间的沉默之后，一位高级教员说道：“因为我们还没有编写视觉程序。”（美国国防部最近向其战略计算计划注入了6亿美元，这项计划在1983—1993年间运行，但是缺少了为自动驾驶坦克创建的视觉系统）。[9]我的回答是“祝你好运”。

在座的有杰拉德·苏斯曼（Gerald Sussman），他为用人工智能解决现实世界的问题做出了几个重要的应用，其中包括高精度轨道力学积分系统，为MIT致敬图灵经典杰作的AI实现方法挽回了尊严（图灵曾证明图灵机，即一个思维实验，可以计算任何可计算的函数）。“那需要多长时间？”我问。“你最好算快一点，否则就要有麻烦了。”我补充道，然后走到房间另一头给自己倒了一杯咖啡。我与教员的对话就这样结束了。

“这个场景出了什么问题？”是我实验室的每个学生都能回答的问题。但是在午餐时间，前两排观众都被难倒了。最后，第三排的一名学生这样回答：“数字计算机是一种通用设备，它可以被编程来计算任何东西，虽然效率很低，但苍蝇是一种专用计算机，可以看和飞，但无法平衡我的账户收支。”这就是正确答案。苍蝇眼中的视觉网络进化了数亿年，其视觉算法嵌入了它本身的网络。这就是为什么

你可以利用苍蝇眼神经回路的布线图和信息流对视觉系统进行逆向工程，以及为什么你不能在数字计算机上这样做，因为硬件本身需要软件来指定要解决什么问题。

我认出了在人群中微笑的罗德尼·布鲁克斯（Rodney Brooks）。我曾经邀请他参加在马萨诸塞州鳕鱼角的伍兹霍尔市（Woods Hole）举办的计算神经科学研讨会。布鲁克斯是澳大利亚人，20世纪80年代，他是MIT AI Lab的初级教员。那段时间他使用不依赖数字逻辑的构架搭建了爬行昆虫机器人。他后来成了实验室主管，随后又创建了打造出Roombas的公司iRobot。

那天下午我做演讲的房间很大，挤满了本科生，新生代想要展望未来，而不是徘徊在过去。我谈到了一个学习如何玩西洋双陆棋的神经网络，这是一个与伊利诺伊大学香槟分校（the University of Illinois in Urbana-Champaign）复杂系统研究中心（Center for Complex Systems Research）的物理学家杰拉德·特索罗（Gerald Tesauro）合作的项目。西洋双陆棋是两名玩家之间的比赛，棋子根据掷骰子的点数向前移动，在途中可以跳过对方的棋子。与具有确定性的国际象棋不同，西洋双陆棋是由偶然因素控制的：每次掷骰子的不确定性都使得对特定棋步结果的预测更加困难。这是一款在中东地区非常受欢迎的游戏，其中一些人以玩高赌注的比赛为生。

考虑到有10^{20}个可能的西洋双陆棋棋盘摆法，基于逻辑和试探法编写程序来处理所有可能的摆法将会是一个不可能完成的任务。于是，我们让神经网络观看教师对局，通过模式识别来学习下棋。[10]杰拉德后来让西洋双陆棋网络通过跟自己下棋进行学习，创建了第一个世界冠军级别的西洋双陆棋程序（第10章会详细讲述这个故事）。

演讲结束之后，我听说那天早上《纽约时报》有一篇关于政府机构如何大幅削减人工智能资金投入的头版文章。虽然对主流AI研究人员来说，寒冬已经来临，但它并没有影响到我和我的研究组成员，对我们来说，神经网络的春天才刚刚到来。

然而我们新的AI实现方法需要25年的时间才能在视觉、语音和语言方面提供实际应用。即使在1989年，我也知道这需要很长时间。1978年，我还在普林斯顿大学读研究生时就在想，按照摩尔定律，计算机的运算能力会呈指数级增长，每18个月翻一番，那达到大脑的计算能力需要多长时间？我的结论是，到2015年会实现。幸运的是，这并没有阻止我继续探索。我对神经网络的信仰是基于我的直觉，即如果大自然解决了这些问题，我们也应该能够从大自然中学习到同样的解决方法。而我不得不耐心等待的这25年，与自然界的数亿年相比，它仅仅只是一个瞬间。

在视觉皮层内部，神经元呈多层次排列结构。随着感官信息在皮层间层层传递，对世界的呈现也变得越来越抽象。几十年来，随着神经网络模型层数的增加，其性能也在不断提高，直到最终达到了一个临界点，让我们能够解决在20世纪80年代只能幻想却无法解决的问题。深度学习可以自动找出能区分图像中不同物体的优质特征的过程，这就是今天的计算机视觉比5年前好得多的原因。

到2016年，计算机的运行速度已经快了上百万倍，计算机内存也从兆字节升级到了太字节（terabytes）。与20世纪80年代只有数百个单元和数千个连接的网络相比，现在模拟出的神经网络具有数百万个单元和数十亿个连接。尽管按照拥有数千亿个神经元和千万亿个突触连接的人类大脑的标准来看，这个数字仍然很小，但现有神经网络的规模已经可以在有限领域中进行原理的证明。

基于深度神经网络的深度学习已经出现了，但在有深度网络之前，我们必须学习如何训练浅层网络。

---

[1] 虽然直到1970年才正式成立，麻省理工学院人工智能实验室（MIT AI Lab）于1959年就开始了相关的研究项目，并于2003年跟麻省理工学院计算机科学实验室（LCS）合并，组建了麻省理工学院计算机科学与人工智能实验室（CSAIL）。为了保持简洁和一致性，我把该机构统称为“MIT AI Lab”。

[2] 参阅Seymour A. Papert, “The Summer Vision Project,” AI Memo AIM-100, July 1, 1966, DSpace@MIT, https://dspace.mit.edu/handle/1721.1/6125。根据麻省理工学院2016级毕业生Michaela Ennis的说法：“有关麻省理工学院本科生被安排将‘计算机视觉’作为暑期研究项目的故事，帕特里克·温斯顿（Patrick Winston）教授每年都会拿到课堂上讲，他还说过这个本科生就是杰拉德·苏斯曼（Gerald Sussman）。”

[3] 参阅Roger Peterson, Guy Mountfort, and P. A. D. Hollom, *Field Guide to the Birdsof Britain and Europe*, 5th ed. (Boston: Houghton Mifflin Harcourt, 2001).

[4] Bruce G. Buchanan and Edward H. Shortliffe, *Rule Based Expert Systems: TheMYCIN Experiments of the Stanford Heuristic Programming Project*(Reading, MA:Addison-Wesley, 1984).

[5] S. Mukherjee, “A.I. versus M.D.: What Happens When Diagnosis Is Automated?”*New Yorker*, April 3, 2017.

[6] Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*(New York: Basic Books, 2015), 35. 甚至没有人知道如何量化所有常识，我们把拥有常识当成理所当然的事。

[7] 猫比人的体重要轻，即使背朝下掉落，也可以在空中翻转过来四肢着地。J.A. Sechzera, S. E. Folsteina, E. H. Geigera, R. F. and S. M. Mervisa, “Development and Maturation of Postural Reflexes in Normal Kittens,” *Experimental Neurology*86, no. 3 (1984): 493–505.

[8] B. Katz, *Nerve, Muscle, and Synapse*(New York: McGraw-Hill, 1996); A. Hodgkin,*Chance and Design: Reminiscences of Science in Peace and War*(Cambridge:Cambridge University Press, 1992).

[9] M. Stefik, “Strategic Computing at DARPA: Overview and Assessment,”*Communications of the ACM*28, no.7 (1985): 690–704.

[10] G. Tesauro and T. J. Sejnowski, “A Parallel Network That Learns to Play Backgammon,” *Artificial Intelligence*39 (1989): 357–390.

03
神经网络的黎明

任何人工智能的难题都可以被解决。唯一能证明这一论断成立的是这样一个事实：自然界通过进化已经解决了这些难题。但在20世纪50年代就已经存在各种暗示，如果AI研究者能够选择完全不同于符号处理的方式，计算机会如何表现出智能行为。

第一条暗示是，我们的大脑是强大的模式识别器。我们的视觉系统可以在1/10秒内识别混乱场景中的对象，即使我们可能从未见过那个特定的对象，也不论该对象在什么位置，多大尺寸，以什么角度面对我们。简而言之，我们的视觉系统就像一台以“识别对象”作为单一指令的计算机。

第二条暗示是，我们的大脑可以通过练习来学会如何执行若干艰巨的任务，比如弹钢琴、掌握物理学知识。大自然使用通用的学习方法来解决特殊的问题，而人类则是顶尖的学习者。这是我们的特殊能力。我们大脑皮层的结构整体上是相似的，并且我们所有的感受系统和运动系统都有深度学习网络。[1]

第三条暗示是，我们的大脑并没有充斥着逻辑或规则。当然，我们可以学习逻辑思维或遵守规则，但必须要经过大量的训练，而我们当中的大多数人对此并不在行。这一点可以通过人们在一个叫作“华生选择任务”（Wason selection task）的逻辑谜题上的典型表现来进行说明（见图3－1）。

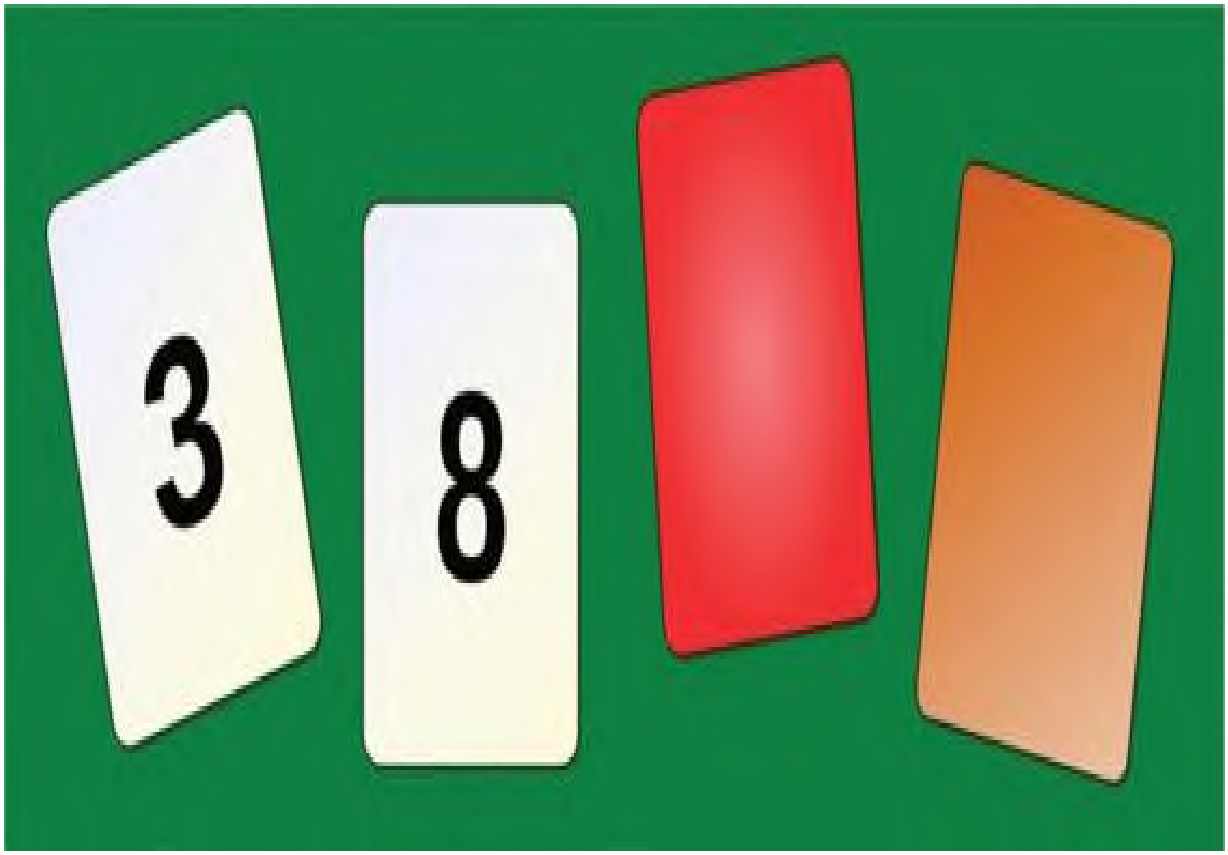


图3-1 这四张卡片，每张都是一面有数字，另一面涂满了颜色。要测试以下命题为真：一张卡片在一面显示为偶数，那它的另一面就是红色的。你需要翻哪（几）张牌呢？图片来源：“华生选择任务”，维基百科。

正确的选择是正面为数字“8”，背面为棕色的卡片。在最初的研究中，只有10% 的受试者给出了正确的答案。[2]但是，当给这项逻辑测试加上了熟悉的背景信息时，大多数受试者都能很快找出正确答案（见图3－2）。

推理似乎是基于特定领域的，我们对该领域越熟悉，就越容易解决其中的问题。经验使得在一个领域内进行推理变得更容易，因为我们可以用已有的例子来下意识地得到解决方案。例如，在物理学中，我们通过解决各种问题，而不是通过背诵公式，来学习电磁学领域的知识。如果人类的智能是完全基于逻辑的，那么它应该是跨领域的通用智能，但事实并非如此。



图3-2 每张卡片都是一面有一个年龄数字，另一面印着一种饮料。需要翻哪（几）张牌才能检验这条法律：如果你正在喝酒，那说明你一定超过18岁了？图片来源：“华生选择任务”，维基百科。

第四条暗示是，我们的大脑充满了数百亿个小小的神经元，每时每刻都在互相传递信息。这表明，要解决人工智能中的难题，我们应该研究具有大规模并行体系结构的计算机，而不是那些具有冯·诺依曼数字体系结构，每次只能获取和执行一个数据或指令的计算机。是的，图灵机在被给予足够内存和时间的条件下，的确可以计算任何可计算的函数，但自然界必须实时解决问题。要做到这一点，它利用了大脑的神经网络，就像地球上最强大的计算机一样，它们具有大量的并行处理器。只有能有效运行的算法，最终才能在自然选择中胜出。

深度学习的起点

20世纪五六十年代，在诺伯特·维纳（Norbert Wiener）提出基于机器和生物中的通信和控制系统的控制论之后不久，[3]学界对自组织系统开始产生了浓厚的兴趣。而其中一个独创性产物便是由奥利弗·塞弗里奇（Oliver Selfridge）创造的Pandemonium（鬼域）。[4]这是一个图案识别设备，其中进行特征检测的“恶魔”通过互相竞争，来争取代表图像中对象的权利（深度学习的隐喻，见图3－3）。斯坦福大学的伯纳德·威德罗（Bernard Widrow）和他的学生泰德·霍夫（Ted Hoff）发明了LMS（最小均方）学习算法，[5]它与其后继算法一起被广泛用于自适应信号处理，例如噪声消除、财务预测等应用。在这里，我将重点关注一位先驱弗兰克·罗森布拉特（Frank Rosenblatt）（图3－4），他发明的感知器是深度学习的前身。[6]

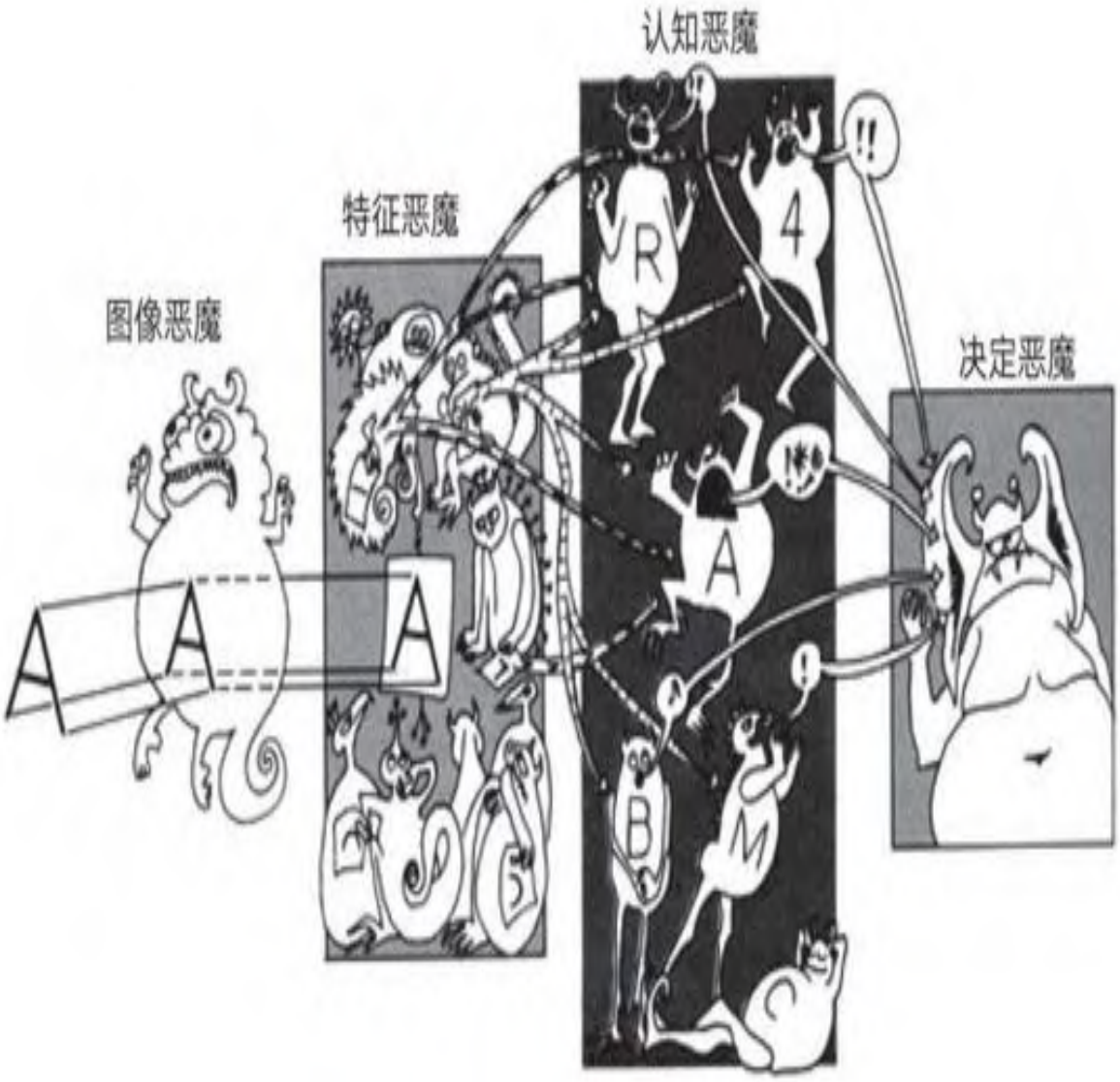


图3-3 Pandemonium。奥利弗·塞弗里奇认为，大脑中有恶魔负责从感官输入中先后提取更复杂的特征和抽象概念，从而做出决定。如果每个级别的恶魔与前一个级别的输入相匹配，则会激动不已。做决定的恶魔需要衡量所有信息传递者的兴奋程度和重要性。这种形式的证据评估是对当前多层次深度学习网络的隐喻。图片来源：Peter H. Lindsay and Donald A. Norman, *Human Information Processing: An Introduction to Psychology*, 2nded. (New York: Academic Press, 1977), 图3-1。维基共享资源：https://commons.wikimedia.org/wiki/File:Pande.jpg。



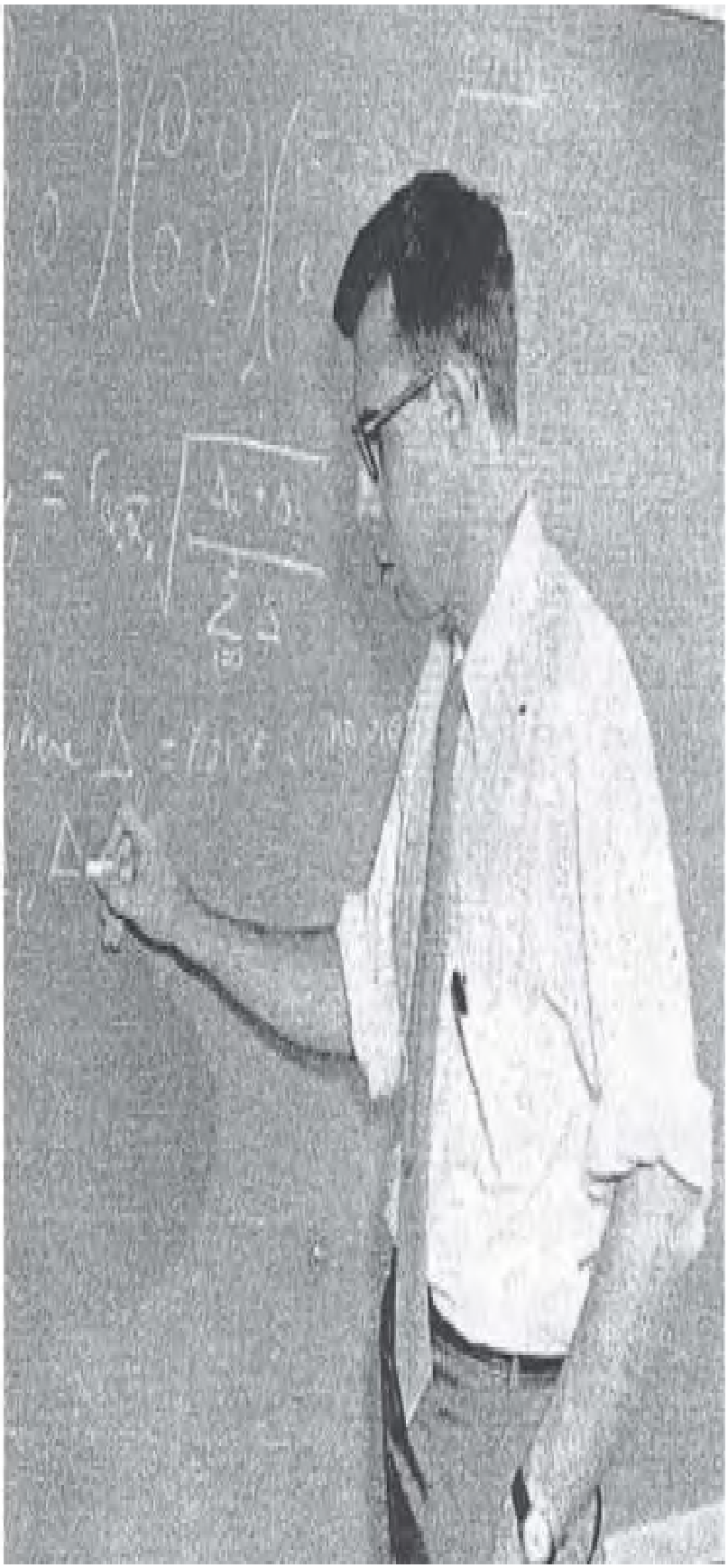
Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

WASHINGTON, July 7 (UPI) —The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The embryo—the Weather Bureau's $2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's demonstration for newsmen.

The service said it would use this principle to build the first of its Perceptron thinking machines that will be able to read and write. It is expected to be finished in about a year at a cost of $100,000.

Dr. Frank Rosenblatt, designer of the Perceptron, conducted the demonstration. He said the machine would be the first device to think as the human brain. As do human beings, Perceptron will make mistakes at first, but will grow





wiser as it gains experience, he said.

Dr. Rosenblatt, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said Perceptrons might be fired to the planets as mechanical space explorers.

图3-4 深思中的康奈尔大学教授弗兰克·罗森布拉特，他发明了感知器。作为深度学习网络的早期雏形，感知器是能够将图像进行分类的简易学习算法。图中文章是1958年7月8日在《纽约时报》上发表的一篇来自合众国际社（UPI）的报道。感知器在1959年完成时预计花费了10万美元，相当于今天的100万美元。IBM704计算机在1958年价值200万美元，相当于现在的2000万美元，可以实现每秒12000次的乘法运算，这在当时已经是极快的速度了。不过相比之下，现在价格要低得多的三星Galaxy S6手机每秒可以执行340亿次操作，速度要快100万倍以上。图片来源：George Nagy。

从样本中学习

尽管我们对大脑功能缺乏足够的了解，但神经网络的AI先驱们依然依靠着神经元的绘图以及它们相互连接的方式，进行着艰难的摸索。康奈尔大学的弗兰克·罗森布拉特是最早模仿人体自动图案识别视觉系统架构的人之一。[7]他发明了一种看似简单的网络感知器（perceptron），这种学习算法可以学习如何将图案进行分类，例如识别字母表中的不同字母。算法是为了实现特定目标而按步骤执行的过程，就像烘焙蛋糕的食谱一样（关于算法，将会在第13章中进行介绍）。

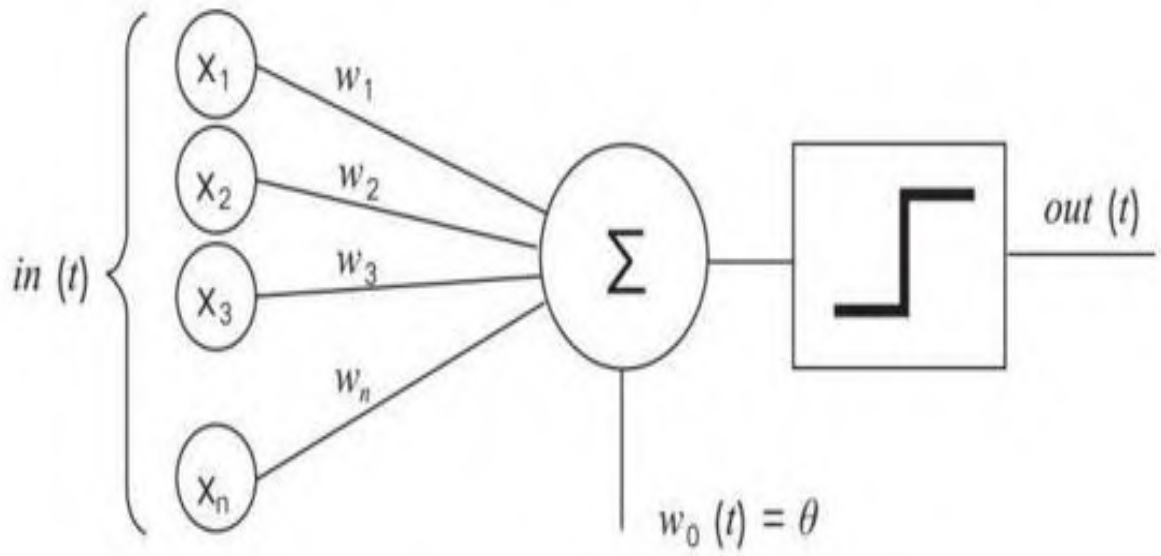
如果你了解了感知器如何学习图案识别的基本原则，那么你在理解深度学习工作原理的路上已经成功了一半。感知器的目标是确定输入的图案是否属于图像中的某一类别（比如猫）。方框3.1解释了感知器的输入如何通过一组权重，来实现输入单元到输出单元的转换。权

重是对每一次输入对输出单元做出的最终决定所产生影响的度量，但是我们如何找到一组可以将输入进行正确分类的权重呢？

工程师解决这个问题的传统方法，是根据分析或特定程序来手动设定权重。这需要耗费大量人力，而且往往依赖于直觉和工程方法。另一种方法则是使用一种从样本中学习的自动过程，和我们认识世界上的对象的方法一样。需要很多样本来训练感知器，包括不属于该类别的反面样本，特别是和目标特征相似的，例如，如果识别目标是猫，那么狗就是一个相似的反面样本。这些样本被逐个传递给感知器，如果出现分类错误，算法就会自动对权重进行校正。

这种感知器学习算法的美妙之处在于，如果已经存在这样一组权重，并且有足够数量的样本，那么它肯定能自动地找到一组合适的权重。在提供了训练集中的每个样本，并且将输出与正确答案进行比较后，感知器会进行递进式的学习。如果答案是正确的，那么权重就不会发生变化。但如果答案不正确（0被误判成了1，或1被误判成了0），权重就会被略微调整，以便下一次收到相同的输入时，它会更接近正确答案（见方框3.1）。这种渐进的变化很重要，这样一来，权重就能接收来自所有训练样本的影响，而不仅仅是最后一个。

3.1
感知器



感知器是具有单一人造神经元的神经网络，它有一个输入层，和将输入单元和输出单元相连的一组连接。感知器的目标是对提供给输入单元的图案进行分类。输出单元执行的基本操作是，把每个输入（x_n）与其连接强度或权重（w_n）相乘，并将乘积的总和传递给输出单元。上图中，输入的加权和（$\Sigma_{i=1,\cdots,n}\ w_i\ x_i$）与阈值 θ 进行比较后的结果被传递给阶跃函数。如果总和超过阈值，则阶跃函数输出“1”，否则输出“0”。例如，输入可以是图像中像素的强度，或者更常见的情况是，从原始图像中提取的特征，例如图像中对象的轮廓。每次输入一个图像，感知器会判定该图像是否为某类别的成员，例如猫类。输出只能是两种状态之一，如果图像处于类别中，则为“开”，否则为“关”。“开”和“关”分别对应二进制值中的1和0。感知器学习算法可以表达为：

$\delta w_i = \alpha\ \delta\ x_i$

δ=output - teacher,

这里，output（输出值）和 teacher（实际值）都是二进制的，所以根据差值，如果输出正确， δ=0，如果输出不正确， δ=+1或者 -1。

如果对感知器学习的这种解释还不够清楚，我们还可以通过另一种更简洁的几何方法，来理解感知器如何学习对输入进行分类。对于只有两个输入单元的特殊情况，可以在二维图上用点来表示输入样本。每个输入都是图中的一个点，而网络中的两个权重则确定了一条直线。感知器学习的目标是移动这条线，以便清楚地区分正负样本（见图3－5）。对于有三个输入单元的情况，输入空间是三维的，感知器会指定一个平面来分隔正负训练样本。在一般的情况下，即使输入空间的维度可能相当高且无法可视化，同样的原则依然成立。

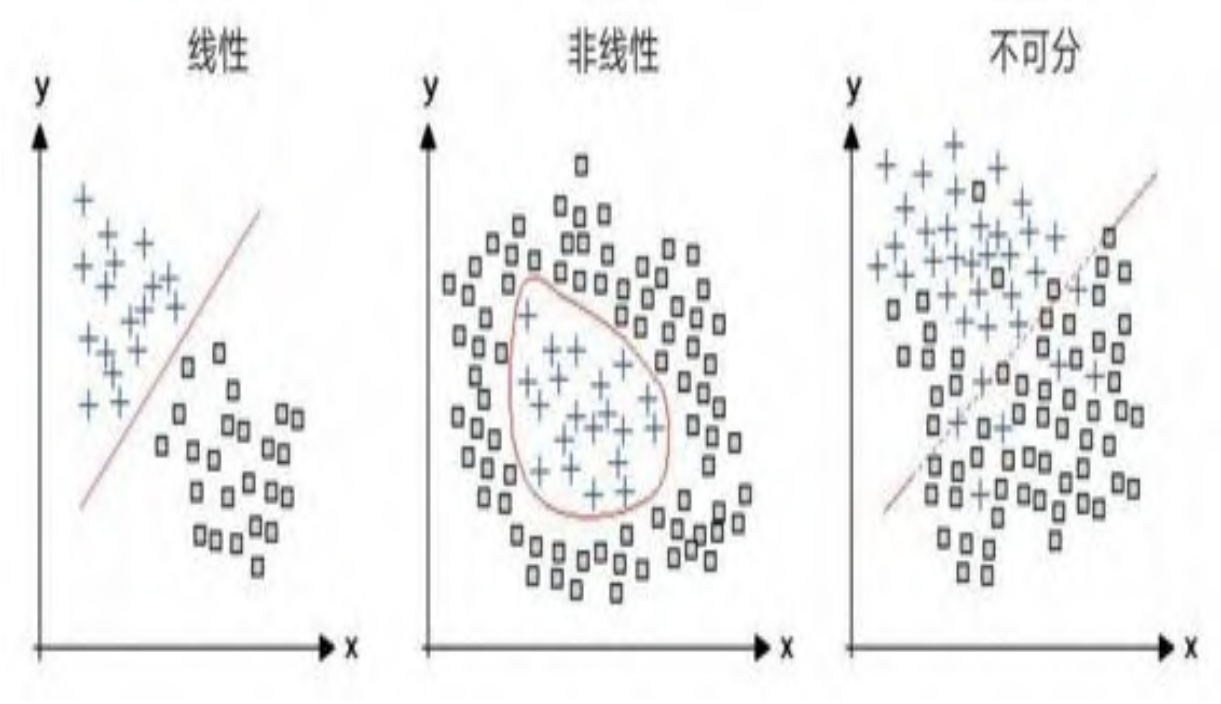


图3-5 关于感知器如何区分两个对象类别的几何解释。这些对象有两个特征，例如尺寸和亮度，它们依据各自的坐标值（x，y）被绘制在每张图上。左边图中的两种对象（加号和正方形）可以通过它们之间的直线分隔开；感知器能够学习如何进行这种区分。其他两个图中的两种对象不能用直线隔开，但在中间的图中，两种对象可以用曲线分开。而右侧图中的对象必须舍弃一些样本才能分隔成两种类型。如果有足够的训练数据，深度学习网络就能够学习如何对这三个图中的类型进行区分。

最终，如果解决方案是可行的，权重将不再变化，这意味着感知器已经正确地将训练集中的所有样本进行了分类。但是，在所谓的“过度拟合”（overfitting）中，也可能没有足够的样本，网络仅仅记住了特定的样本，而不能将结论推广到新的样本。为了避免过度拟合，关键是要有另一套样本，称为“测试集”（test set），它没有被用于训练网络。训练结束时，在测试集上的分类表现，就是对感知器是否能够推广到类别未知的新样本的真实度量。泛化（generalization）是这里的关键概念。在现实生活中，我们几乎不会在同样的视角看到同一个对象，或者反复遇到同样的场景，但如果我们能够将以前的经验泛化到新的视角或场景中，我们就可以处理更多现实世界的问题。

利用感知器区分性别

举一个用感知器解决现实世界问题的例子。想想如果去掉头发、首饰和第二性征，比如男性比女性更为突起的喉结，该如何区分男性和女性的面部。比阿特丽斯·哥伦布（Beatrice Golomb）是 1990年我实验室里的一名博士后研究员，她利用一个数据库中的大学生面部照片作为感知器的输入，经过训练的感知器能以81%的准确度对面部的性别进行分类（见图3－6）。[8]而对于感知器难以分类的面部，人类也同样很难做出区分。我实验室的成员在同一组人的面部识别上达到了88% 的平均准确度。比阿特丽斯还训练了多层感知器（将在第8章中介绍），其准确度达到了92%，[9]比我实验室的成员还要准确。她在1991年的NIPS大会上发表的演讲中总结道：“经验可以提高性能，这表明实验室的研究人员需要花更多时间来进行性别鉴定的工作。”她把她的多层感知器叫作“SEXNET”（性别网络）。在问答环节，有人问是否可以使用SEXNET来检测异装癖者的面孔。“可以。”比阿特

丽斯这样回答。而NIPS大会的创始人爱德华·波斯纳（Edward Posner）辩驳道：“那就应该叫DRAGNET（法网）。”[10]

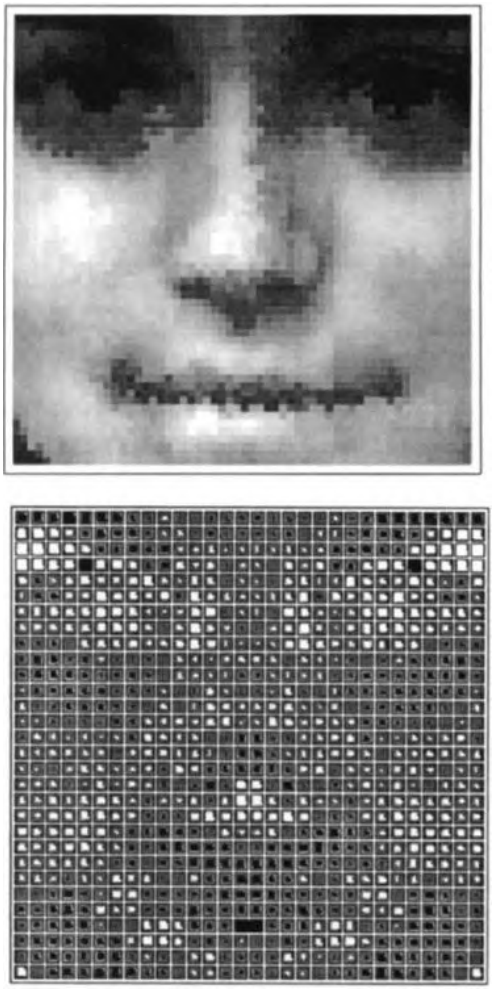


图3-6 这张脸属于男性还是女性？人们通过训练感知器来辨别男性和女性的面孔。来自面部图像（上图）的像素乘以相应的权重（下图），并将该乘积的总和与阈值进行比较。每个权重的大小被描绘为不同颜色像素的面积。正值的权重（白色）表现为男性，负值的权重（黑色）倾向于女性。鼻子宽度，鼻子和嘴之间区域的大小，以及眼睛区域周围的图像强度对于区分男性很重要，而嘴和颧骨周围的图像强度对于区分女性更重要。图片来源：M. S. Gray, D. T. Lawrence, B. A. Golomband T.J.Sejnowski, “A Perceptron Revealsthe Face of Sex,” *Neural Computation*7 (1995):1160-1164, 图1。

区分男性与女性面部的工作有趣的一点是，虽然我们很擅长做这种区分，却无法确切地表述男女面部之间的差异。由于没有单一特征是决定性的，因此这种模式识别问题要依赖于将大量低级特征的证据

结合起来。感知器的优点在于，权重提供了对性别区分最有帮助的面部的线索。令人惊讶的是，人中（即鼻子和嘴唇之间的部分）是最显著的特征，大多数男性人中的面积更大。眼睛周围的区域（男性较大）和上颊（女性较大）对于性别分类也有着很高的信息价值。感知器会权衡来自所有这些位置的证据来做出决定，我们也是这样来做判定的，尽管我们可能无法描述出到底是怎么做到的。

1957年罗森布拉特对“感知器收敛定理”的证明是一个突破，他的演示令人印象深刻。在美国海军研究办公室（Office of Naval Research）的支持下，他搭建了一个以400个光电单元作为输入的定制硬件模拟计算机，其权重是由电机调整的可变电阻电位器。模拟信号随着时间连续变化，就像黑胶唱片中的信号一样。用一组图片集（其中部分图片中有坦克，另外一部分则没有）进行训练，罗森布拉特的感知器即使在新图像中也能准确识别坦克。这一成果经《纽约时报》报道后引起了轰动（见图3－4）。[11]

感知器激发了对高维空间中模式分离的美妙的数学分析。当那些点存在于有数千个维度的空间中时，我们就无法依赖在生活的三维空间里对点和点之间距离的直觉。俄罗斯数学家弗拉基米尔·瓦普尼克（Vladimir Vapnik）在这种分析的基础上引入了一个分类器，称为“支持向量机”（Support Vector Machine），[12]它将感知器泛化，并被大量用于机器学习。他找到了一种自动寻找平面的方法，能够最大限度地将两个类别的点分开（见图3－5，线性）。这让泛化对空间中数据点的测量误差容忍度更大，再结合作为非线性扩充的“内核技巧”（kernel trick），支持向量机算法就成了机器学习中的重要支柱。[13]

被低估的神经网络

但是有一个限制，使得感知器的研究存在问题。上面的假设“如果存在这样的权重集合”提出了一个这样的困惑，即什么样的问题可能或不可能被感知器解决。令人尴尬的是，在二维平面中，简单分布的点不能被感知器分开（见图3－5，非线性）。事实证明，坦克感知器不是坦克分类器，而是天气分类器。[14]对图像中的坦克进行分类要困难得多，而事实上，它不能用感知器来完成。这也表明，即使感知器学到了一些东西，也可能不是你认为它应该学到的那些东西。压倒感知器的最后一根稻草是马文·明斯基和西摩尔·帕普特在1969年发表的数学专著《感知器》（*Perceptrons*）。[15]他们明确的几何分析表明，感知器的能力是有限的：它们只能区分线性可分的类别（见图3－5）。这本书的封面展示了明斯基和帕普特证明的感知器无法解决的几何问题（见图3－7）。尽管在书的末尾，明斯基和帕普特考虑了将单层感知器进行泛化成为多层感知器的前景，但他们怀疑可能没有办法训练这些更强大的感知器。不幸的是，许多人对他们的论断坚信不疑，于是这个研究领域渐渐被人们遗忘，直到20世纪80年代，新一代神经网络研究人员开始重新审视这个问题。

在感知器中，每个输入都独立地向输出单元提供证据。但是，如果需要依靠多个输入的组合来做决定，那会怎样呢？这就是感知器无法区分螺旋结构是否相连的原因：单个像素并不能提供它是在内部还是外部的位置信息。尽管在多层前馈神经网络中，可以在输入和输出单元之间的中间层中形成多个输入的组合，但是在20世纪60年代，还没有人知道如何训练简单到中间只有一层“隐藏单元”（hiddenunits）的神经网络。

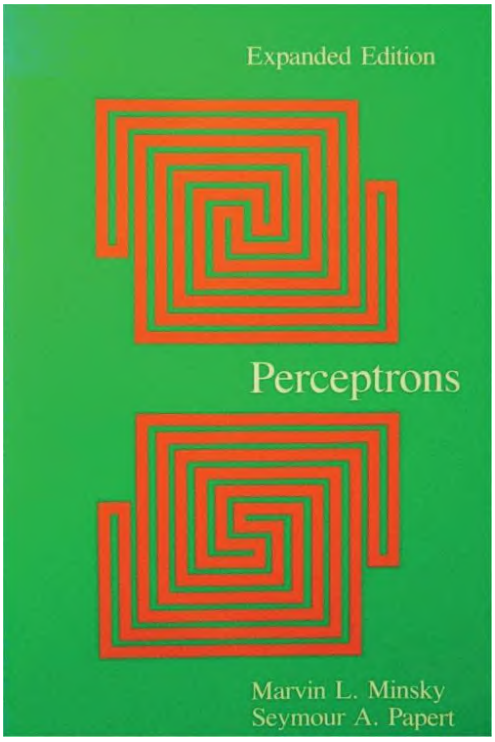


图3-7 《感知器》一书的增订版封面。两个红色的螺旋结构看起来一样，但实际并非如此。上方的图案由两个不相连的螺旋线组成，下面的则是单一的螺旋线。你可以通过用铅笔跟踪环路的内部路径来验证。明斯基和帕普特证明了感知器不能区分这两个对象。你能直接用肉眼看出区别吗？为什么不能？

弗兰克·罗森布拉特和马文·明斯基曾是纽约市布朗克斯科技高中的同班同学。他们在科学会议上为各自迥异的人工智能研究方法展开了辩论，而与会者更倾向于明斯基的方法。尽管存在差异，但他们二人对我们理解感知器都有着重要贡献，而这正是深度学习的起点。

罗森布拉特在1971年死于一次驾船事故，年仅43岁，当时正值人们几乎一边倒地反对感知器的时期。有传言说他可能是自杀，但也可能只是一次不幸的出游。[16]不可否认的是，一个发现了利用神经网络进行计算的新方式的英雄时代已经谢幕；又过了整整一代人的时间，罗森布拉特开创性努力的承诺才得以实现。

---

[1] 在空间上，人们发现细胞特性和连接性在皮层不同部分之间存在差异，这可能反映了不同感官系统的特异化和其结构的层次性。

[2] P. C. Wason, “Self-Contradictions,” in P. N. Johnson-Laird and P. C. Wason, eds.,*Thinking: Readings in Cognitive Science*(Cambridge: Cambridge University Press,1977).

[3] Norbert Wiener, *Cybernetics, or Control and Communication in the Animal and the Machine*(Cambridge, MA: MIT Press, 1948).

[4] O. G. Selfridge, “Pandemonium: A Paradigm for Learning,” in D. V. Blake and A. M.Uttley, eds., *Proceedings of the Symposium on Mechanisation of Thought Processes*(1959): 511–529.

[5] 参阅Bernard Widrow and Samuel D. Stearns, *Adaptive Signal Processing*(Englewood Cliffs, NJ: Prentice-Hall, 1985)。

[6] 参阅Frank Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*(Washington, DC: Spartan Books, 1962)。

[7] 罗森布拉特虽然是位喜欢在康奈尔大学校园内开跑车的腼腆单身汉，但同时还是一位有着广泛兴趣的博学之人。他的兴趣之一，就是在凌日现象发生时，研究如何通过测量远处恒星表面亮度的略微下降，来寻找围绕其运转的行星。这种方法现在常用于检测银河系中绕着恒星运转的太阳系外行星。

[8] M. S. Gray, D. T. Lawrence, B. A. Golomb, and T. A. Sejnowski, “A Perceptron Reveals the Face of Sex,” *Neural Computation*7, no. 6 (1995): 1160–1164.

[9] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski, “SEXNET: A Neural Network Identifies Sex from Human Faces,” in R. Lippmann, and D. S. Touretzky, eds.,*Advances in Neural Information Processing Systems*3 (1991): 572–577.

[10] 波斯纳一语双关，暗示了一部流行的电视剧《法网》（*Dragnet*），取材于20世纪50年代来自洛杉矶警察局的罪犯克星。

[11] M. Olazaran. “A Sociological Study of the Official History of the Perceptrons Controversy,” *Social Studies of Science*26, no. 3 (1996): 611–659.

[12] Vladimir Vapnik, *The Nature of Statistical Learning Theory*(New York: Springer1995), 138.

[13] Weifeng Liu, José C. Principe, and Simon Haykin, *Kernel Adaptive Filtering: AComprehensive Introduction*(Hoboken, NJ: Wiley, 2010).

[14] 根据尤德科夫斯基在“Artificial Intelligence as a Positive and Negative Factor in Global Risk”中的记录，训练集中的坦克图片样本都是在阴天拍摄的，无坦克样本图片都是在晴天拍摄的。——译者注

[15] Marvin Minsky and Seymour Papert, *Perceptrons*(Cambridge, MA: MIT Press,1969). 也可参阅Marvin Lee Minsky and Seymour Papert, *Perceptrons: AnIntroduction to Computational Geometry*, expanded ed. (Cambridge, MA: MITPress, 1988)。

[16] 据加州大学圣迭戈分校的同事Harvey Karten称，罗森布拉特是一位经验丰富的水手。他曾带领一群学生驾船出游。船好像撞到了什么，他从船上落入水中，但当时没有一个学生能救得了他。（个人对话，2017年11月8日。）

04
大脑式的计算

“如果我有大脑”，这是1939年经典音乐电影《绿野仙踪》（*The Wizard of Oz*）里稻草人的唱词。但稻草人不知道的是，他早就已经有大脑了，否则他就没法说话或唱歌。但真正的问题是，他的大脑只诞生了两天，里面没有经验，空空如也。随着时间的推移，他对世界有了充分的认识，最后被公认为是奥兹国里最聪明的人，聪明到能意识到自己的局限。而铁皮人唱的是，“如果我能有心”。他和稻草人为心和脑哪个更重要而争论不休。在奥兹国中，认知和情感这两个大脑的产物在精妙的平衡中共同协作，通过不断学习慢慢创造出了类人的智慧。这在真实世界里亦是如此。这部音乐剧也引出了本章的主题：如果人工智能拥有大脑和心脏。

网络模型能够模仿智能行为

我和杰弗里·辛顿（见图4-1）在1979年他组织的一次研讨会上相识。因为对神经网络模型的未来抱有相似的信念，我们很快就成了朋友。后来我们合作研究出了一种新型神经网络模型，叫作“玻尔兹曼机”（Boltzmann machine）（将在第7章中进行讨论），并就此打破了阻碍一代人研究多层网络模型的僵局。



图4-1 （A）童年时期的杰弗里·埃弗里斯特·辛顿（Geoffrey Everest Hinton）。他的中间名来自一个亲戚乔治·埃弗里斯特（George Everest）。乔治曾在印度负责勘测工作，并想出了测量世界上最高峰高度的方法，后来该峰以他的名字被命名为Everest（即珠穆朗玛峰）。（B）1979年的辛顿。这两张照片的拍摄时间间隔了15年。图片来源：杰弗里·辛顿。

每隔几年，我都会接到杰弗里的一个电话，第一句都是“我想清楚大脑是怎么工作的了”。每一次，他都会告诉我一个巧妙的改进神经网络模型的新方案。尝试了多种方案并进行了若干次的改进后，利用多层神经网络的深度学习才在手机中的语音识别和识别照片中的对象等方面达到了与人类相近的水平。公众几年前才意识到深度学习的这些能力，虽然现在这些已经是众所周知的了，但这一过程却花费了相当长的时间。

杰弗里在剑桥大学获得了心理学学士学位，并在爱丁堡大学获得了人工智能博士学位。他的论文导师是克里斯托弗·朗古特－希金斯（Christopher Longuet-Higgins）。希金斯是一位杰出的化学家，他发明了一种早期的联想记忆网络模型。那时，实现人工智能的主流方式是基于使用符号、逻辑和规则来编写智能行为的程序；认知心理学家已经采用这种方法来理解人类的认知能力，尤其是语言。而那时的杰弗里却在逆流而行。没有人能预见到他有朝一日会搞清楚大脑——或者至少类似大脑的某种东西——的工作原理。他的讲座非常引人入胜，他能够清楚地解释抽象的数学概念，连没有太多数学基础的人都能够理解和掌握。他的机智和不失谦逊的幽默感令人倾倒。杰弗里天生也有着激进的一面，特别是当涉及大脑的话题时。

我们第一次见面时，杰弗里还是加州大学圣迭戈分校（UCSD）的一名博士后研究员，在由大卫·鲁姆哈特（David Rumelhart）和詹姆斯·麦克莱兰（James McClelland）领导的并行分布式处理（Parallel Distributed Processing，PDP）实验室工作。杰弗里坚信，将由简单处理单元构成的网络、并行工作和从样本中学习相结合，是理解认知的更好的方式。当时PDP实验室正在探索如何以分布在网络中的大量节点活动的形式来理解文字和语言，而杰弗里是其中的核心人物。

认知科学中理解语言的传统方法是基于符号表征。例如，“杯子”这个词是代表杯子概念的符号，不是特指某个杯子，而是指所有

杯子。符号的美妙之处在于，它们让我们能够对复杂的概念进行压缩，并运用它们；而符号的问题在于，这种过分概括的表达形式使其很难在现实世界中被精确地描述出来——在现实世界中，杯子的样式、形状和尺寸有着无数种可能。虽然我们大多数人在看到杯子的时候都能很快认出它是什么，但没有一个逻辑程序能够清楚地指认哪个东西是杯子，哪个不是，也无法识别出图片中的杯子。正义、和平之类的抽象概念更会让一个计算机程序产生困惑。另一种方法是，通过大量神经元上的活动模式来表示杯子，如此一来就可以捕捉概念之间的相似和差异点。这赋予了符号可以反映其含义的丰富的内部结构。问题在于，在1980年，还没有人知道如何创建这些内部表征。

在20世纪80年代，相信网络模型能够模仿智能行为的人并不只有我和杰弗里。全世界有一些研究人员，虽然大多都在独自耕耘，但与我们有着共同的信念，坚持不懈地开发着专门的网络模型。他们当中的一位，克里斯托弗·冯·德·马尔斯伯格（Christoph von der Malsburg），开发了一种模式识别模型，将发射脉冲的人造神经元连接在一起，[1]并证明了这种方法可以识别图像中的人脸。[2]另外，大阪大学的福岛邦彦（Kunihiko Fukushima）发明了神经认知机（Neocognitron），[3]一个基于视觉系统架构的多层网络模型，它使用了卷积滤波器和简单形式的赫布可塑性（Hebbian plasticity），这也是深度学习网络的一个直接的前身。第三个例子是赫尔辛基大学的电气工程师戴沃·科霍宁（Teuvo Kohonen），他开发了一个自组织网络，可以学习将相似的输入通过不同的处理单元聚类到二维映射中（例如可以用来代表不同的语音），相似的输入能够激活输出空间的相邻区域。[4]科霍宁网络模型的一个主要优点是，它不需要对每个输入的类别进行标记（通过生成标记来训练监督网络的花费十分高昂）。科霍宁的箭袋中只有一支箭，但却被打磨得非常精良。

在加州大学洛杉矶分校的朱迪亚·珀尔（Judea Pearl）引入了将网络中的结点用概率联系起来的信念网络，比如草地变湿，是因为喷水器打开了的概率，或者因为下雨了的概率。[5]这是一个很有前景的将概率网络系统化的早期尝试。虽然珀尔的网络模型是一个可以用于追踪世间因果关系的强大框架，但手动分配所有所需概率已经被证明是不切实际的。能够自动找到概率的学习算法依然有待突破（会在本书第二部分讨论）。

上述几个例子和其他基于网络的模型都有一个共同的致命缺陷：它们都不足以解决现实世界中的问题。而且，开发它们的先驱者很少与他人合作，要取得进展就更加举步维艰。如此一来，麻省理工学院、斯坦福大学和卡内基－梅隆大学的一流人工智能研究中心里，就没有多少人看好神经网络了。基于规则的符号处理获得了大部分资金支持，并完成了大部分的相关研究工作。

神经网络先驱者

1979年，杰弗里·辛顿和布朗大学心理学家詹姆斯·安德森在加利福尼亚的拉荷亚市组织了联想记忆的并行模型研讨会。[6]大多数参与者互相之间都是第一次见面。我很惊讶自己能够受邀参加研讨会，我那会儿只是哈佛大学医学院神经生物学的博士后研究员，而且只在一些不太知名的期刊上发表了一些关于神经网络的技术论文。杰弗里后来告诉我，他曾经与大卫·马尔（David Marr）（见图4－2，中间）审核过我的背景。马尔是神经网络建模中的一位顶尖人物，也是MIT AI Lab的一名有远见的领导者。1976年，我在怀俄明州杰克逊霍尔市（Jackson Hole）的一个小型研讨会上第一次遇到了马尔。我们有相似的兴趣，他还曾邀请我去MIT参观，并在那里做一次演讲。



图4-2 （左起）托马索·波吉奥（Tomaso Poggio）、大卫·马尔和弗朗西斯·克里克（Francis Crick）在加利福尼亚远足的途中（照片拍摄于1974年）。弗朗西斯很享受与来访者在各种科学问题上进行长时间的讨论。图片来源：索尔克生物研究所。

马尔从剑桥大学获得了数学学士学位和生理学博士学位。他的博士生导师是生理学家吉尔斯·布林德利（Giles Brindley），专门研究视网膜和色觉，但在音乐理论和勃起功能障碍治疗领域也有较高的知名度。吉尔斯有一件很出名的逸事——在内华达州拉斯维加斯举行的美国泌尿外科学会（The American Urological Association）会议上发表演讲时，他脱掉了裤子，亲自证明了化学诱导勃起的有效性。马尔的博士论文描述了一个小脑神经网络学习模型，这块区域与快速运动控制有关。他还开发了海马体和大脑皮层的神经网络模型，撰写的大量相关文章也都被证明十分有预见性。[7]

当我第一次在杰克逊霍尔遇到马尔时，他已经到了麻省理工学院，当时正在从事视觉研究工作，并凭借他的个人魅力吸引到了一批有才华的学生跟他一起工作。他追求一种自下而上的策略，从视网膜开始入手（在那里光被转换成电信号），并探求视网膜中的信号如何

编码对象的特征，以及视觉皮层如何表示物体的表面和边界。他和托马索·波吉奥为立体视觉开发了一种带有反馈连接的递归神经网络模型，通过检测双眼看到的随机点立体图中点的图像的轻微横向位移，来测量对象的深度。[8]双眼深度感知是三维立体图产生效果的基础。[9]马尔于1980年因白血病去世，年仅35岁。两年之后，他的遗作《视觉》（*Vision*）得以发表。[10]具有讽刺意味的是，尽管马尔在他的视觉研究中采取了自下而上的策略，即从视网膜开始并对视觉处理的每个后续阶段进行建模，他的著作却以倡导自上而下的策略而闻名——首先对要解决的问题进行计算分析，然后构建算法来解决问题，最后通过硬件来实现算法。然而，尽管这可能是在解决问题后对问题进行解释的一种有效途径，但对于揭开大脑秘密却算不上是个好方法。难点就在第一步——要判断大脑正在解决什么问题。我们的直觉往往具有误导性，特别是在视觉方面；我们擅长观察，但大脑对我们隐藏了所有的细节。因此，纯粹的自上而下策略问题重重，但纯粹自下而上的策略也同样如此（后面的章节将探讨通过学习算法，由内而外理解视觉工作原理的进展）。

弗朗西斯·克里克也参加了辛顿和安德森在拉荷亚的研讨会，他曾在1953年与同在剑桥大学的詹姆斯·沃森（James Watson）共同发现了DNA的双螺旋结构。在20多年后的1977年，克里克加入了位于拉荷亚的索尔克生物研究所，并将研究重心转移到神经科学领域。他会邀请研究人员来访，并就神经科学的许多课题，特别是视觉方向，进行长时间的讨论，马尔就是其中一位访问者。在马尔著作的末尾，有一段苏格拉底对话式的启发性讨论，我后来得知，这段对话源自马尔与克里克的讨论。1989年我加入索尔克研究所后，也开始意识到与克里克对话的重要价值。

乔治·布尔与机器学习

1854年，一位自学成才、有5个女儿（有几个很有数学天分）的英国教师，写了一本名为《思维规律的研究》（*An Investigation of the Laws of Thought*）的书，它是“布尔逻辑”的数学基础。乔治·布尔对于如何使用逻辑表达式的见解是数字计算的核心，也是20世纪50年代人工智能萌芽期成果的基石。杰弗里·辛顿作为布尔的玄孙，很自豪能拥有一支布尔曾经使用过并在家族中代代相传的笔。

在准备一次演讲时，我发现布尔这本著作的全称是《思维规律的研究——逻辑与概率的数学理论基础》。虽然其中让人印象最深刻的是对逻辑的洞察，但这本书对于概率论也有很多讨论。概率论是现代机器学习的核心，在描述现实世界中的不确定性方面比逻辑方法要好用得多。所以布尔也是机器学习的先驱之一。有些讽刺的是，他思想中被遗忘的一面，在250年后因为他的玄孙才开始绽放。布尔一定会为杰弗里感到骄傲的。

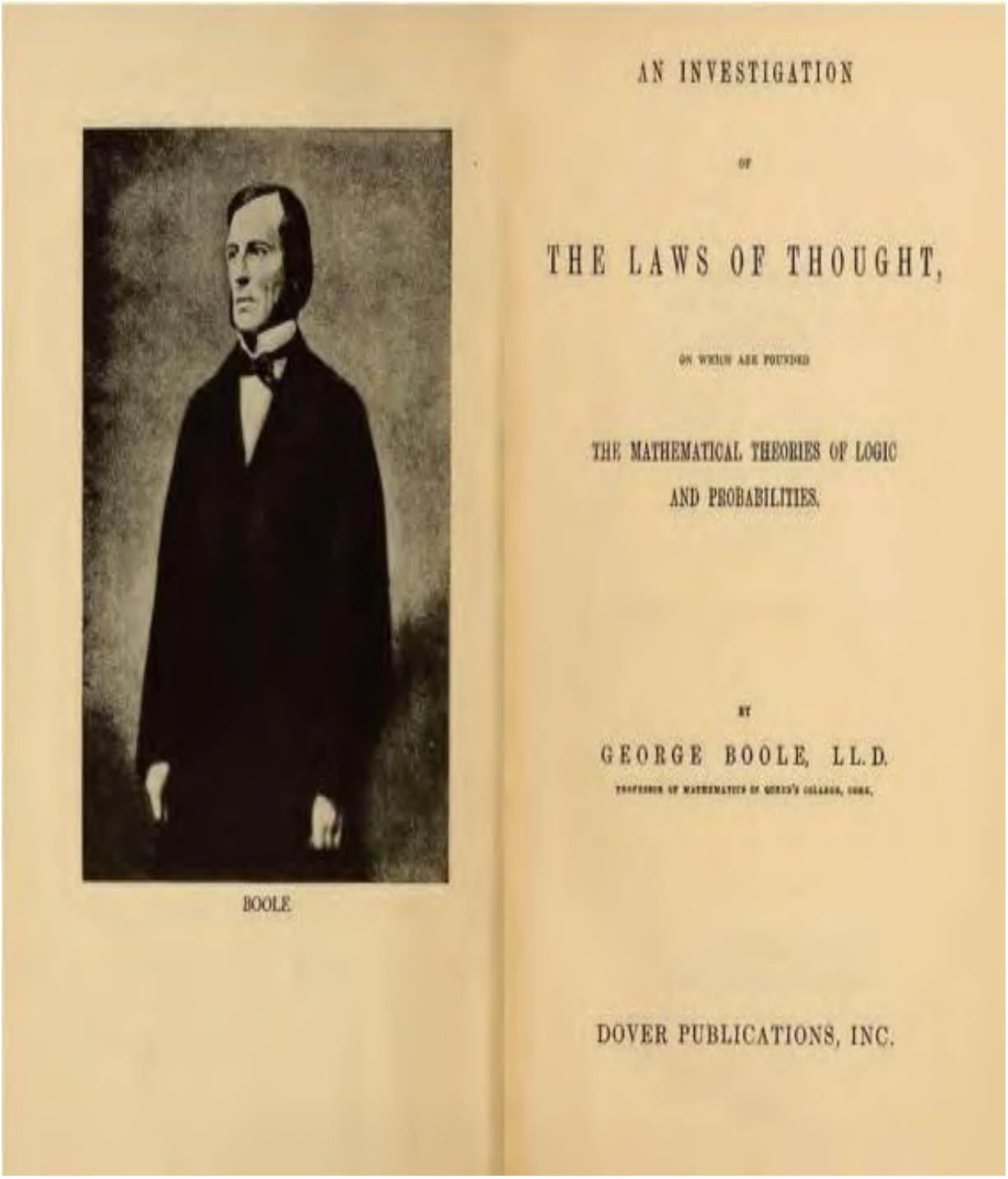


图4-3 尽管乔治·布尔的《思维规律的研究》以研究思维的基础——逻辑而闻名，但请注意，它也探讨了概率的问题。这两个数学领域分别启发了在符号处理和机器学习方面的人工智能研究方法。

利用神经科学理解大脑

在普林斯顿大学物理系攻读研究生时，我曾通过写下非线性神经元交互网络的方程式并分析它们来解决如何理解大脑的问题，[11]就像物理学家几个世纪以来使用数学来理解重力、光、电、磁和核能的本质一样。每天晚上睡前，我都会祈祷：“亲爱的主，让方程式是线性的，噪声呈高斯分布，变量是可分的吧。”这些是能产生分析解法的条件，但是因为神经网络方程式是非线性的，与之相关的噪声是非高斯分布的，而且变量是不可分的，所以它们并没有明确的解。除此之外，当时想要在计算机上模拟大型网络的方程是不可能的。更令人沮丧的是，我甚至不确定我的方程式到底对不对。

在普林斯顿上课时，我发现神经科学家们正在取得令人振奋的进展，这个年轻的学科领域45年前才初露端倪。在此之前，生物学、心理学、解剖学、生理学、药理学、神经学、精神病学、生物工程学等许多学科都对大脑进行了研究。在1971年神经科学学会（The Society for Neuroscience）的第一次会议上，弗农·蒙特卡索（Vernon Mountcastle）亲自在门口迎接了每一位与会者。[12]今天，这个学会有4万多名成员，大约3万人通常会出席年会。我于1982年在约翰·霍普金斯大学的生物物理系开始我的第一份工作时，遇到了这位传奇的神经生理学家，他发现了皮层柱，而且有着强大的人格魅力。[13]我随后与蒙特卡索密切合作，计划建立约翰·霍普金斯大学的脑研究所。这是世界上该领域第一个研究机构，成立于1994年。

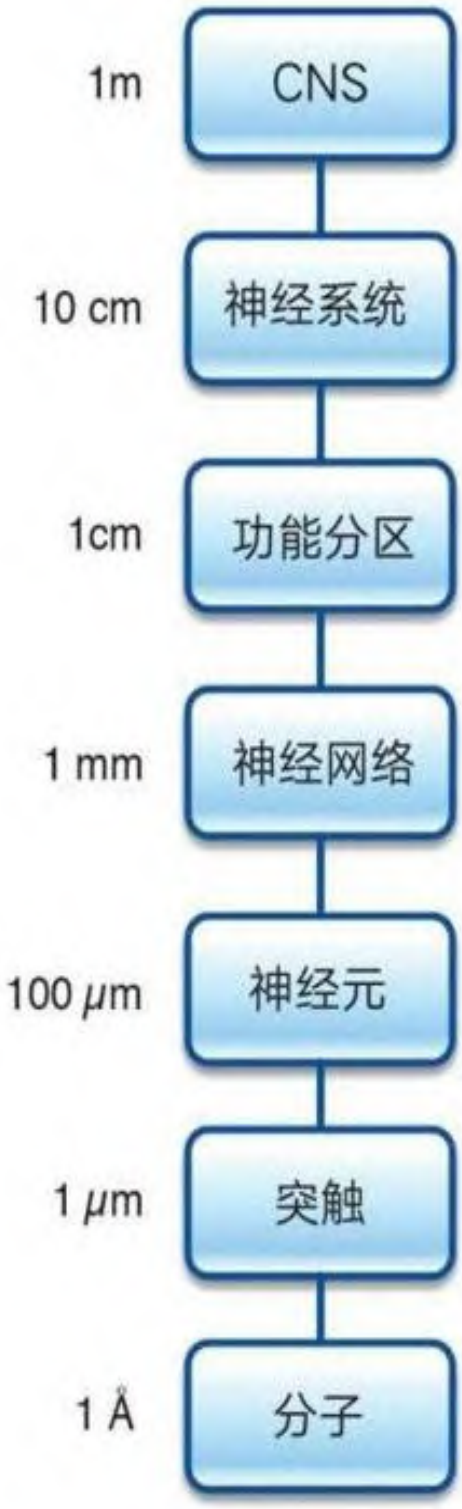
对于大脑的研究存在着许多不同的层级（见图4－4），每个层级都有着重要的发现，而要整合所有这些知识是一个棘手的问题。这让人想起了关于矮胖子的童谣：

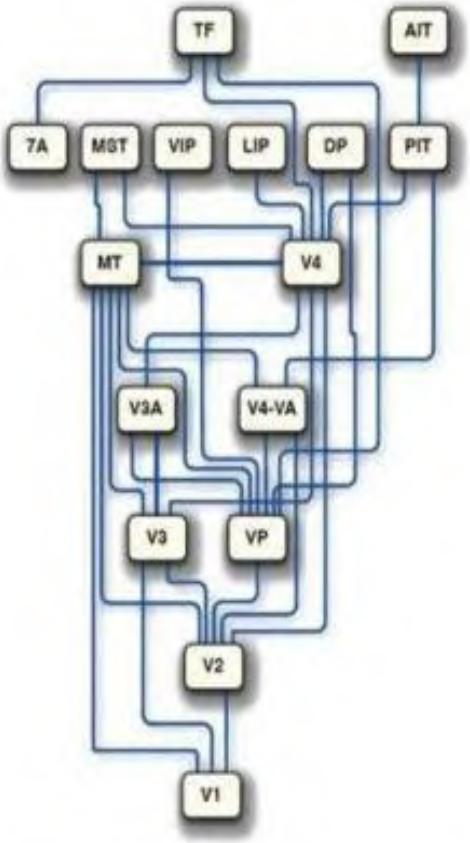
矮胖子，坐墙头，

栽了一个大跟斗。

国王呀，齐兵马，

破镜难圆没办法。


1m
CNS
10 cm
神经系统
1cm
功能分区
1 mm
神经网络
100 μm
神经元
1 μm
突触
1 Å
分子


TF
AIT
7A
MST
VIP
LIP
DP
PIT
MT
V4
V3A
V4-VA
V3
VP
V2
V1

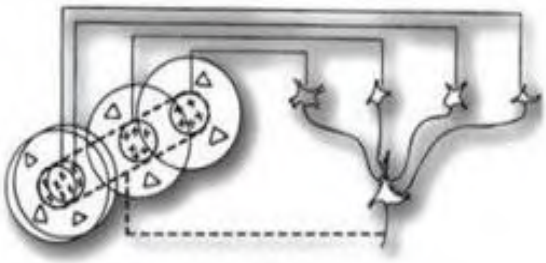




图4-4 大脑研究中的层级。（左）空间尺度范围从底部的分子水平到顶部的整个中枢神经系统（centralnervous systems，CNS）。我们已经对每一个层级都有了深入的了解，但对具有少量彼此高度相关的神经元的网络层级，即由人工神经网络进行模拟的层级，却知之甚少。（右）突触（图片下方）的示意图，视觉皮层中的简单细胞结构（图片中间）以及视觉皮层中皮层区域的层次结构（图片上方）。图片改编自：P. S. Churchland，T.J. Sejnowski，“Perspectives on Cognitive Neuroscience”，*Science*，242(1988):741-745，Figure 1。

尽管神经科学家非常善于解剖大脑，但再把这些部分拼回去就没那么容易了。这一过程需要做加法而不是减法，而这正是我想要实现的。但首先我必须知道各个部分都是什么，毕竟大脑的构成太复杂了。

查尔斯·格罗斯（Charles Gross）教授是在普林斯顿研究猴子视觉系统的心理学家。在他的一个研究生讲座上，我对哈佛大学医学院的大卫·休伯尔和托斯坦·威泽尔在记录视觉皮层里单个神经元方面所取得的进展印象深刻。如果物理学不是理解大脑工作原理的捷径，那么有可能是神经科学。由于他们在初级视觉皮层中的开创性工作，休伯尔和威泽尔共同获得了1981年的诺贝尔生理学或医学奖。（他们的发现是深度学习的基础，接下来会在第5章中讨论，也是第9章的主题。）

大脑如何处理问题

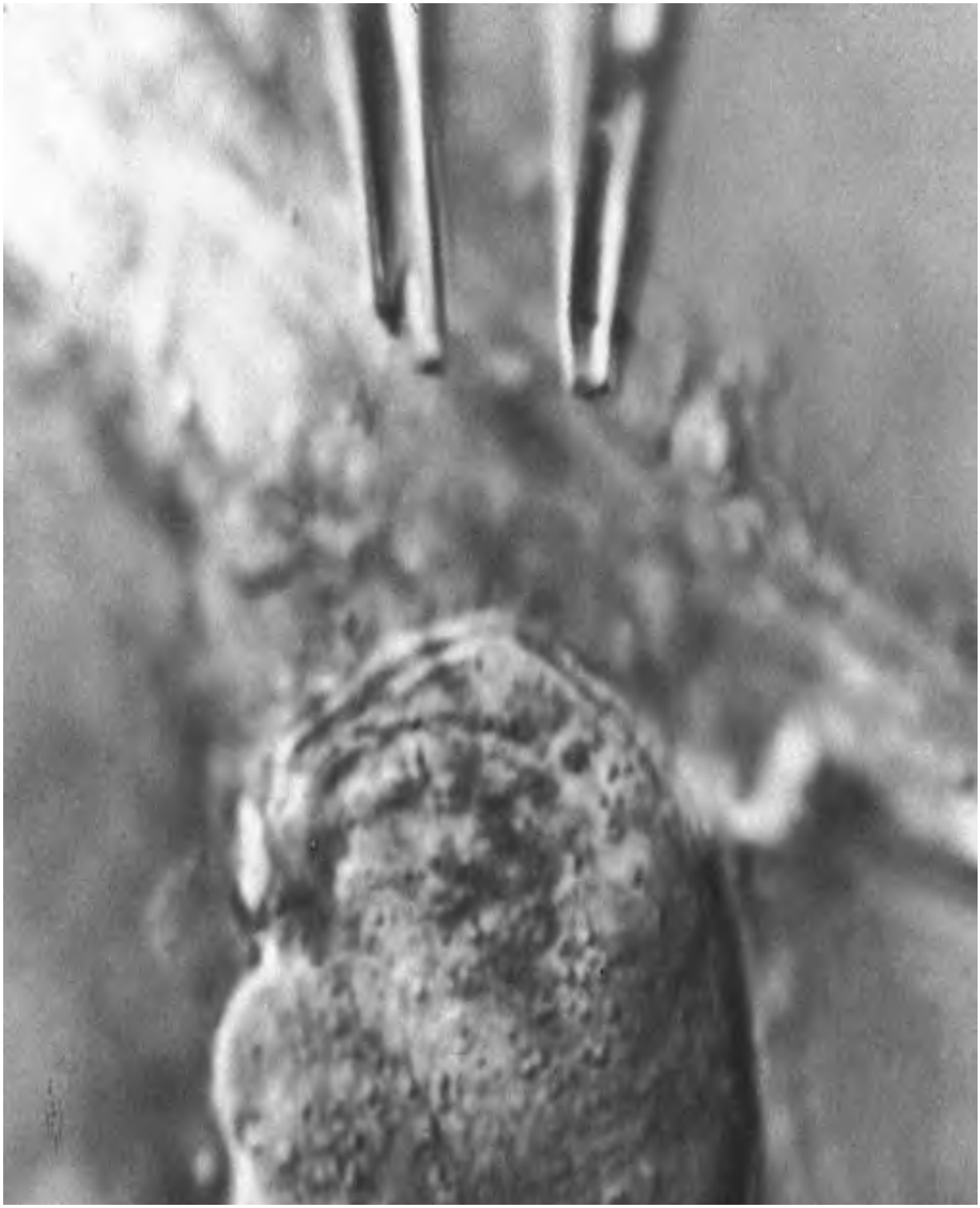
1978年在普林斯顿获得物理学博士学位后，我参加了位于伍兹霍尔市的海洋生物学实验室举办的为期10周的深度实验神经生物学夏季课程。上课的第一天，我穿着一件蓝色休闲外套和熨烫得十分平整的卡其色裤子，却被课程讲师之一的斯托里·兰迪斯（Story Landis）带到一边，她给我买了我人生中的第一条牛仔裤。斯托里当时是哈佛大学神经生物学系的教员，后来在美国国立卫生研究院（以下简称NIH）担任国家神经疾病和中风研究所主任。一想起她，我就会想到这件事。

暑期课程结束后，我在9月份又待了几个星期，想要完成我之前开启的一个项目。鲨鱼和鳐形目（包括鳐科）能够感知非常微弱的电场；事实上，它们甚至可以感受到横穿大西洋的1.5伏电池的信号。凭借这种第六感，鳐鱼可以通过它们在地球磁场中运动产生的微弱电信号来识别方向，而这一过程产生的毫伏级的信号可以被它们的电接收器感知。我在项目中获得了鳐鱼电感受器令人叹为观止的电子显微镜图像。[14]

我正在伍兹霍尔Loeb Hall大楼的地下室拍照的时候，意外地接到了哈佛医学院神经生物学系创始人斯蒂芬·库夫勒（Stephen Kuffler）的一个电话。库夫勒是神经科学领域的传奇人物，成为他实验室里的一名博士后研究员，是我人生的一个转折点。在搬去波士顿之前，我完成了与导师艾伦·盖尔普林（Alan Gelperin）共同进行的一个绘制大蛞蝓的足神经节代谢活动的短期博士后项目。[15]从那以后，我每次吃蜗牛都会不由自主地想到它的大脑。艾伦师承自一系列研究动物行为神经基础的神经科学家。我所学到的是，无脊椎动物中所谓的简单神经系统，实际上比进化阶梯上那些更高级动物器官里的更复杂，因为无脊椎动物必须依赖更少的神经元存活，每个神经元都是高度特异化的。我也开始明白，没有行为支持，神经科学的任何东西都讲不通。[16]

在库夫勒的实验室里，我研究了牛蛙交感神经节一个突触的迟慢兴奋性反应（见图4－5），它的反应速度是位于同一神经元的另一个突触上的快速的毫秒级兴奋性反应的1/60000。[17]这些神经节含有形成牛蛙自主神经系统输出的神经元，可以对腺体和内部器官进行调节。在刺激到突触的那个神经后，我有充分的时间去倒杯咖啡再返回座位——从突触输入到神经元达到峰值会花上1分钟左右，然后神经元会再用10分钟恢复到初态。突触是大脑中基本的计算单元，而突触类型的多样性不可小觑。这次经历告诉我，复杂性可能不是通向理解大脑功能的坦途。为了理解大脑，我必须了解，大自然如何通过进化早

早就解决了大量的问题，并将这些解决方案自下而上地传递给进化链上的物种。我们大脑中的离子通道在几十亿年前的细菌体内就存在了。



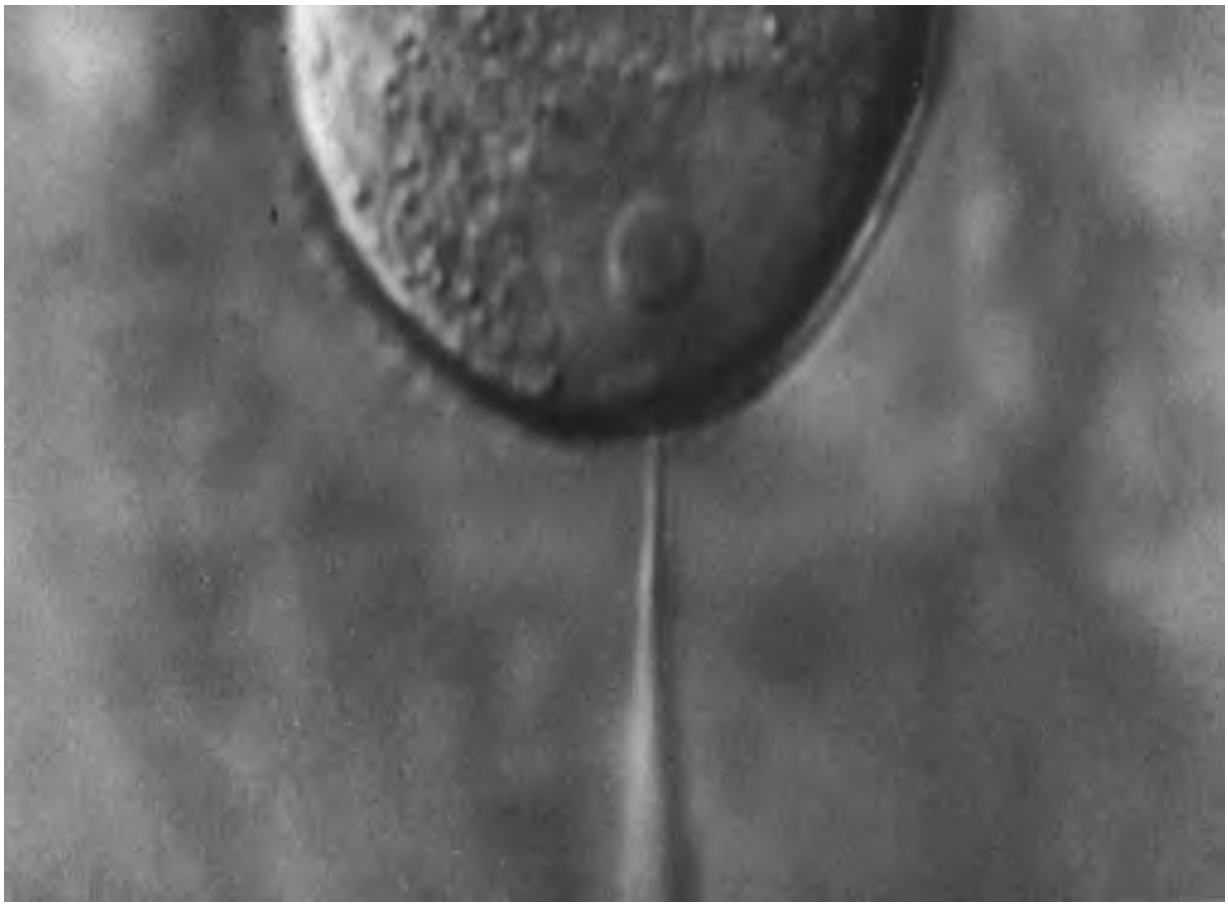


图4-5 牛蛙交感神经节细胞。作为神经元，这些细胞会接受来自脊髓的输入并激活牛蛙皮肤中的腺体。它们体积很大，产生的电信号很容易就能被微电极记录下来（图片底部）。它们没有树突，可以通过神经（图片顶部的背景）或化学物质（图片顶部，一对微量移液器）进行电刺激。刺激神经会引发三种不同的突触信号：一种是快速的毫秒级兴奋性反应，类似于神经肌肉连接处的兴奋反应；一种是较慢的兴奋性反应，在10秒后达到峰值，持续1分钟；还有一种是迟慢兴奋性反应，在1分钟后达到峰值，持续10分钟。这说明即使是最简单的神经元也存在较宽的反应时间跨度。图片来源：S. W. Kuffler, T. J. Sejnowski, “Peptidergic and Muscarinic Excitation at Amphibian Sympathetic Synapses” *Journal of Physiology*341(1983): 257-278, plate I。

计算机科学的兴起

但是如果物理学太简单，生物学又过于复杂，我应该在哪里寻求指导呢？与物理学中的力不同，大脑回路有一个目的，就是解决计算问题，比如看见和移动，以便在世界上生存。即使是一个关于神经元

如何工作的完美的物理模型，也不会告诉我们它的目的是什么。神经元负责处理携带信息的信号，而计算则是试图理解大自然这一过程中缺失的环节。我在过去的40年中一直在追求这一目标，并开创了一个新的领域，叫作“计算神经科学”。

在加州大学圣迭戈分校结束博士后研究员工作之后，杰弗里·辛顿返回了英格兰，在剑桥医学研究理事会（MRC）的应用心理学部门找到了一个研究职位。1981年的一天，他在凌晨两点接到一个电话，对方自称是加利福尼亚州帕洛阿尔托市（Palo Alto）系统开发基金会总裁查尔斯·史密斯（Charles Smith）。[18]史密斯表示，他的基金会希望为潜在的、有前景的，但风险较高、成功希望渺茫的研究提供资金，而有人强烈推荐了杰弗里。杰弗里不确定这件事可不可靠。作为我的一个好朋友，杰弗里向史密斯提到了我的研究，告诉他我的研究成功的概率比他的还要小。

基金会是货真价实的，还为我们提供了第一笔款项，这大大加快了我们的研究进度。我们可以买得起运算能力更强大的电脑，并吸引更多的学生加入。在杰弗里加入位于匹兹堡的卡内基－梅隆大学时，他用一台炫酷的Lisp机[19]取代了他的二代苹果电脑；而在我搬到巴尔的摩的约翰·霍普金斯大学时，我短暂地拥有过比整个计算机科学系还要多的计算能力。[20]我还购买了第一个将霍普金斯与阿帕网（ARPANET，互联网的前身）连接起来的调制解调器，这样杰弗里和我就可以互相发送电子邮件了。能够在新的起点有个好的开端，我们都感到很满意（图4－6）。



图4-6 我和杰弗里·辛顿在波士顿讨论视觉网络模型（照片拍摄于1980年）。这是杰弗里和我在拉荷亚的并行模型研讨会上见面的一年之后。又过了一年，我在位于巴尔的摩的约翰·霍普金斯大学成立了自己的实验室，而杰弗里在匹兹堡的卡内基-梅隆大学也创立了他的研究小组。图片来源：杰弗里·辛顿。

---

[1] Christoph von der Malsburg, “The Correlation Theory of Brain Function,” Internal Report 81-2 (Göttingen: Max-Planck Institute for Biophysical Chemistry,1981), https://fias.uni-frankfurt.de/fileadmin/fias/malsburg/publications/vdMcorrelation. pdf.

[2] P. Wolfrum, C. Wolff, J. Lücke, and C. von der Malsburg. “A Recurrent Dynamic Model for Correspondence-Based Face Recognition,” *Journal of Vision*8, no. 34(2008): 1-18.

[3] K. Fukushima, “Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position.” *BiologicalCybernetics*36, no. 4 (1980): 93–202.

[4] T. Kohonen, “Self-Organized Formation of Topologically Correct Feature Maps,” *Biological Cybernetics*43, no. 1 (1982): 59–69.

[5] Judea Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*(San Mateo, CA: Morgan Kaufmann, 1988).

[6] 他们的论文集也由此而来：Geoffrey E. Hinton and James A. Anderson, eds.,*Parallel Models of Associative Memory*(Hillsdale, NJ: Erlbaum, 1981)。

[7] Terrence J. Sejnowski, “David Marr: A Pioneer in Computational Neuroscience,”in Lucia M. Vaina, ed., *From the Retina to the Neocortex: Selected Papers of David Marr*(Boston: Birkhäuser, 1991), 297–301; 参阅D. Marr, “A Theory of Cerebellar Cortex,” *Journal of Physiology*202 (1969): 437–470; D. Marr, “A Theory for Cerebral Neocortex,” *Proceedings of the Royal Society of London*: *B Biological Sciences*176 (1970): 161–234; D. Marr, “Simple Memory: A Theory for Archicortex,” *Philosophical Transactions of the Royal Society of London*: *B Biological Sciences*262 (1971): 23–81。

[8] D. Marr and T. Poggio, “Cooperative Computation of Stereo Disparity,” *Science*194, no. 4262 (1976): 283–287; 关于随机点立体图（random-dot stereograms）的描述，也可以参考Béla Julesz, *Foundations of Cyclopean Perception*(Chicago:University of Chicago Press, 1971)。

[9] 魔眼图像（Magic Eye images）是一种立体图，在图案中有一个隐藏的三维结构，可以通过偏移你的双眼视线看到。1979年，克里斯托弗·泰勒（Christopher Tyler）创造了第一个黑白立体图。参阅http://www.magiceye.com/。

[10] David Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*(New York: W. H. Freeman, 1982).

[11] Terrence Joseph Sejnowski, “A Stochastic Model of Nonlinearly Interacting Neurons” (Ph.D. diss., Princeton University, 1978).

[12] T. J. Sejnowski, “Vernon Mountcastle: Father of Neuroscience,” *Proceedings ofthe National Academy of Sciences of the United States of America*112, no. 4262(2015): 6523–6524.

[13] 约翰·霍普金斯大学有三个生物物理学系，分别在医学、公共卫生和艺术与科学学院。（我当时在位于Homewood校区艺术与科学学院的托马斯·詹金斯生物物理系。）

[14] T. J. Sejnowski and M. I. Yodlowski, “A Freeze-Fracture Study of the Skate Electroreceptor,” *Journal of Neurocytology*11, no. 6 (1982): 897–912.

[15] T. J. Sejnowski, S. C. Reingold, D. B. Kelley, and A. Gelperin, “Localization of[3H]-2-Deoxyglucose in Single Molluscan Neurones,” *Nature*287, no. 5781(1980): 449–451.

[16] 这句话的灵感来自遗传学家和进化生物学家西奥多西厄斯·多布赞斯基（Theodosius Dobzhansky）的名言：“没有进化论支持，生物学中的任何东西都讲不通。”（Nothing in biology makes any sense except in the light of evolution.）这个版本来自比尔·纽瑟（Bill Newsome），可以在“大脑2025”这一美国国家卫生研究院（NIH）的“BRAIN计划”发展蓝图中找到。https://www.braininitiative.nih.gov/2025/.

[17] S. W. Kuffler and T. J. Sejnowski, “Peptidergic and Muscarinic Excitation at Amphibian Sympathetic Synapses,” *Journal of Physiology*341 (1983): 257–278.

[18] 系统开发公司（The System Development Corporation）是位于加州圣莫尼卡（Santa Monica）市的一家非营利性软件公司，为美国军方提供合约服务。当公司解散时，他们清算了公司的建筑资产，并从中获得了高额利润，这在非营利机构中是不允许的。1969年，他们在加州帕洛阿尔托成立了系统开发基金会（The System Development Foundation），目的是通过1980年到1988年间的赠款计划，来分配从建筑物销售中获得的收益。

[19] 由Symbolis公司制造的Lisp机，是为能够编写符号处理的AI程序而设计的，但它们并不擅长进行大规模数字运算，而后者正是模拟神经网络所需要的。

[20] 1984年，作为美国国家科学基金会（NSF）总统青年研究员奖（Presidential Young Investigator Award）的获得者，我得到了新成立的计算机公司Ridge提供的巨大购买折扣，我买的这台计算机拥有匹敌VAX 780的计算能力，后者是当时学术计算领域的骨干机型。

05
洞察视觉系统

在我上幼儿园之前，最早期的记忆中有这么一个场景——我盯着一堆拼图，以形状、颜色和图案为线索来匹配它们。我的父母会在聚会上夸耀他们还在蹒跚学步的儿子能够快速熟练地完成拼图，而那些朋友常常对此表现得很惊讶。我那时还没有意识到，我的大脑正在完成它最擅长的工作——通过模式识别来解决问题。科学充满了各种各样的问题，就像拼图中缺失的部分和对拼合后画面的模糊提示。大脑如何解决问题，是终极谜题。

亥姆霍兹俱乐部（The Helmholtz Club）是南加州视觉科学家们组成的一个小团体。这些科学家来自加州大学圣迭戈、洛杉矶和尔湾三个分校，以及加州理工学院和南加州大学，他们每个月都会找一个下午在加州大学尔湾分校会面。[1]赫尔曼·冯·亥姆霍兹（Hermann von Helmholtz）是19世纪的一位物理学家和医生，他发明了一套关于视觉的数学理论和实验方法，构成了我们当今理解视觉感知的基础。作为俱乐部的秘书，我要负责从组织外部邀请一位嘉宾，为这15~20名成员以及他们的客人进行演讲，然后由俱乐部成员做第二个演讲。演讲是互动性质的，大家有充足的时间进行深入讨论。有一次，一位外部演讲人对那些提出问题的人表现出了惊讶：“他们还真是喜欢刨根问底。”这些每月的例会让参与者收获了顶尖思维，几乎就是一堂堂视觉领域的大师课。[2]

视觉是我们最敏锐，也是被研究得最多的一种感官。前额下方的眼睛带给了我们精准敏锐的双眼深度知觉，而我们的大脑皮层中一半

的部分都是负责视觉的。“眼见为实”这句成语就充分体现了视觉的特殊地位。然而，也正是这种良好的视觉，导致我们完全忽视了视觉系统背后巨大的计算复杂性，大自然经过数亿年的进化才解决了这个问题（如第2章中所述）。视觉皮层的组织结构为最成功的深度学习网络提供了灵感。

在1/10秒内，我们视觉皮层中的100亿个神经元并行工作，能够在杂乱的场景中识别一个杯子，即便我们以前可能从未见过那个杯子，也不论它在什么位置，多大尺寸，以什么角度面对我们。在普林斯顿大学读研究生时期，我对视觉研究十分着迷，并且在查尔斯·格罗斯的实验室工作了一个夏天。他研究过猴子的下颞叶皮层（见图5-1），并在那里发现了对复杂对象，如脸部，以及著名的马桶刷，产生反应的神经元。[3]

在哈佛医学院神经生物学系的时候，我曾与斯蒂芬·库夫勒一起工作，他之前发现了视网膜中的神经节细胞编码视觉场景的方式。如果不是在1980年过世，他可能会与大卫·休伯尔和托斯坦·威泽尔一起分享1981年的诺贝尔生理学或医学奖。在1989年转到索尔克研究所后，我开始与弗朗西斯·克里克合作。他于1977年将研究重点从分子遗传学转向神经科学，并专注于寻找视觉意识的神经关联。能够和当时最伟大的视觉科学家之一共事，我感到万分荣幸。

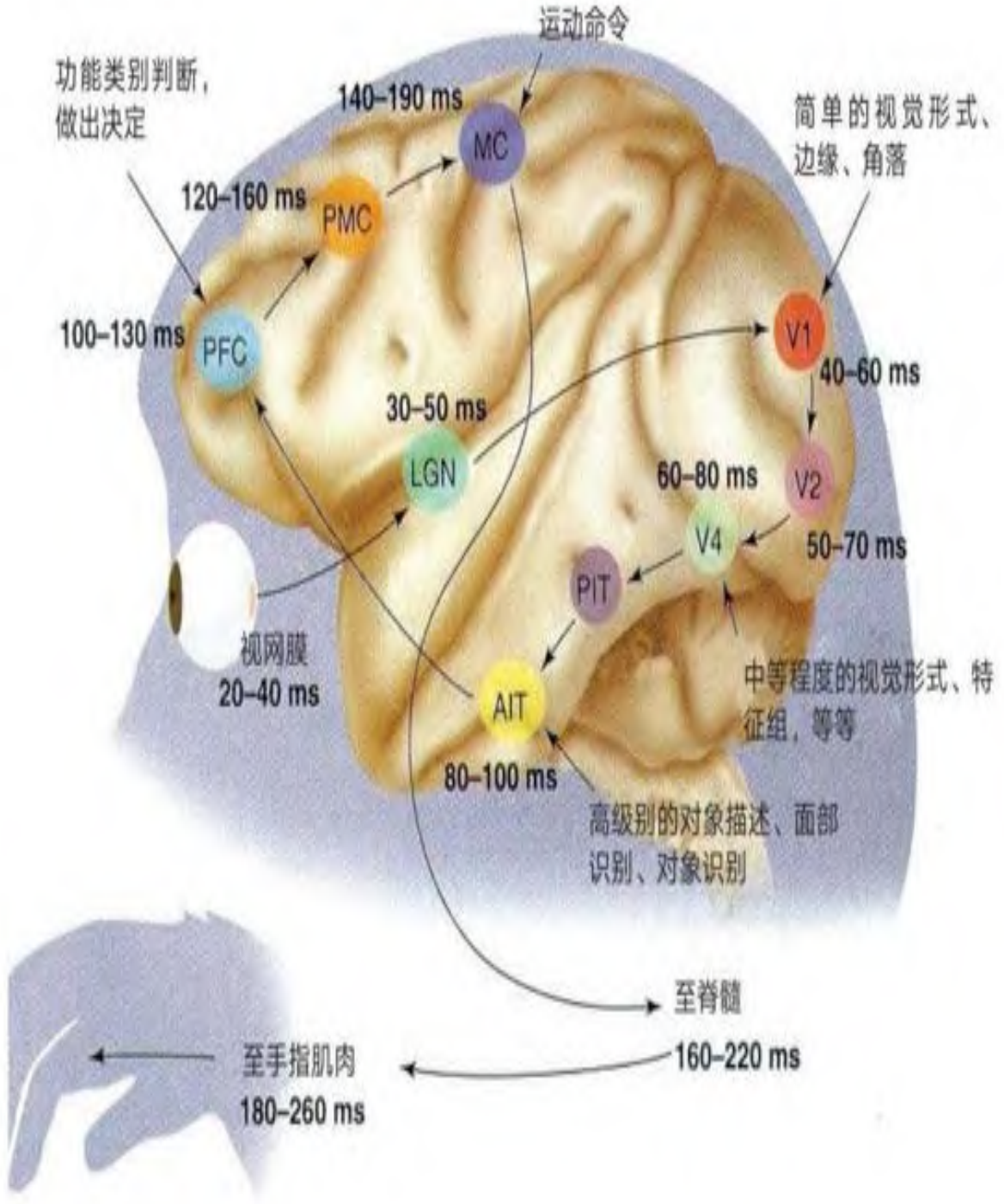


图5-1 猕猴视觉系统的信息流动示意图。箭头表示视觉区域的投影信息从视网膜开始，到达视觉处理的每个阶段都有若干毫秒的延迟。猕猴的视觉感知与我们人类的相似，我们也有相同的视觉处理阶段。LGN（lateralgeniculate nucleus）：外侧膝状体；V1：初级视觉皮层；V2：次级视觉皮层；V4：视觉区4；AIT和PIT：前下颞叶皮层和后下颞叶皮层；PFC：前额叶皮层；PMC：前运动皮层；MC：运动皮层。图

片来源：S. J.Thorpe and M. Fabre-Thorpe, “Seeking Categories in the Brain,” *Science*291, no. 5502 (2001)：261。

人眼是如何看到东西的

如果我们跟随图像生成的信号进入大脑，就可以看到信号是如何在两个连续的处理阶段之间被不断转换的（见图5－1）。视觉始于视网膜，在那里，光感受器将光转换为电信号。视网膜内有两层神经元，它们在空间和时间维度中处理视觉信号，最后通过神经节细胞投射到视神经。

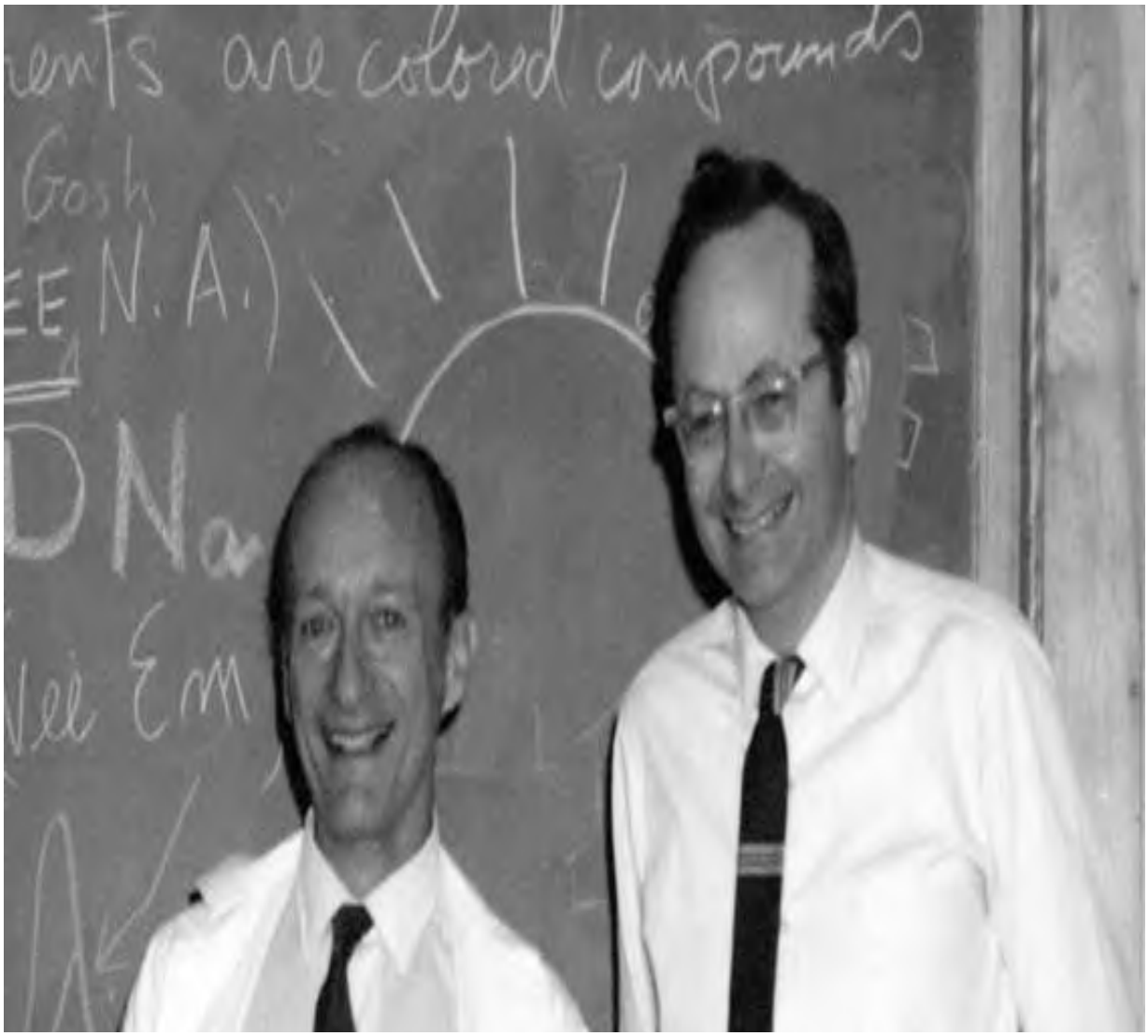




图5-2 （从左到右）斯蒂芬·库夫勒、托斯坦·威泽尔和大卫·休伯尔。哈佛医学院神经生物学系成立于1966年，这张照片拍摄于成立后不久。工作日里，我从来没有见过他们中的任何一个人在实验室里戴着领带，所以这张照片一定是在特殊场合拍摄的。图片来源：哈佛医学院。

在1953年一个结果对所有哺乳动物都成立的经典实验中，斯蒂芬·库夫勒（见图5－2，左）记录了活猫的视网膜输出神经元对光斑刺激的放电响应。他在报告中写道，当光斑作用于某些输出神经元的中心区域时，会出现放电增加的现象，而当光斑作用于另一些输出神经元的中心区域时，却出现了放电减少的现象。然而，在中心的外沿区域的表现恰恰相反：给光中心带有撤光外沿，撤光中心带有给光外沿（见图5－3）。神经节细胞对光斑模式的反应被称为“感受野”（receptive field）特性。

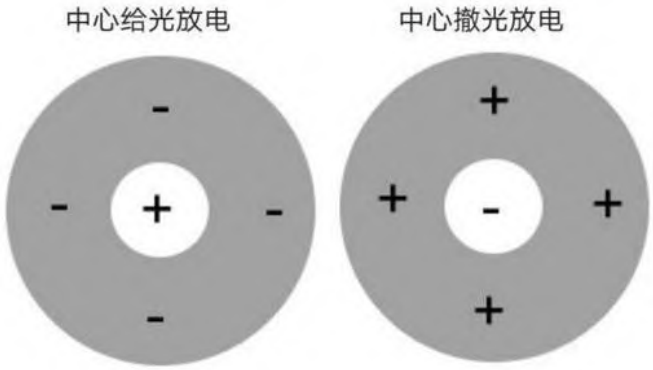


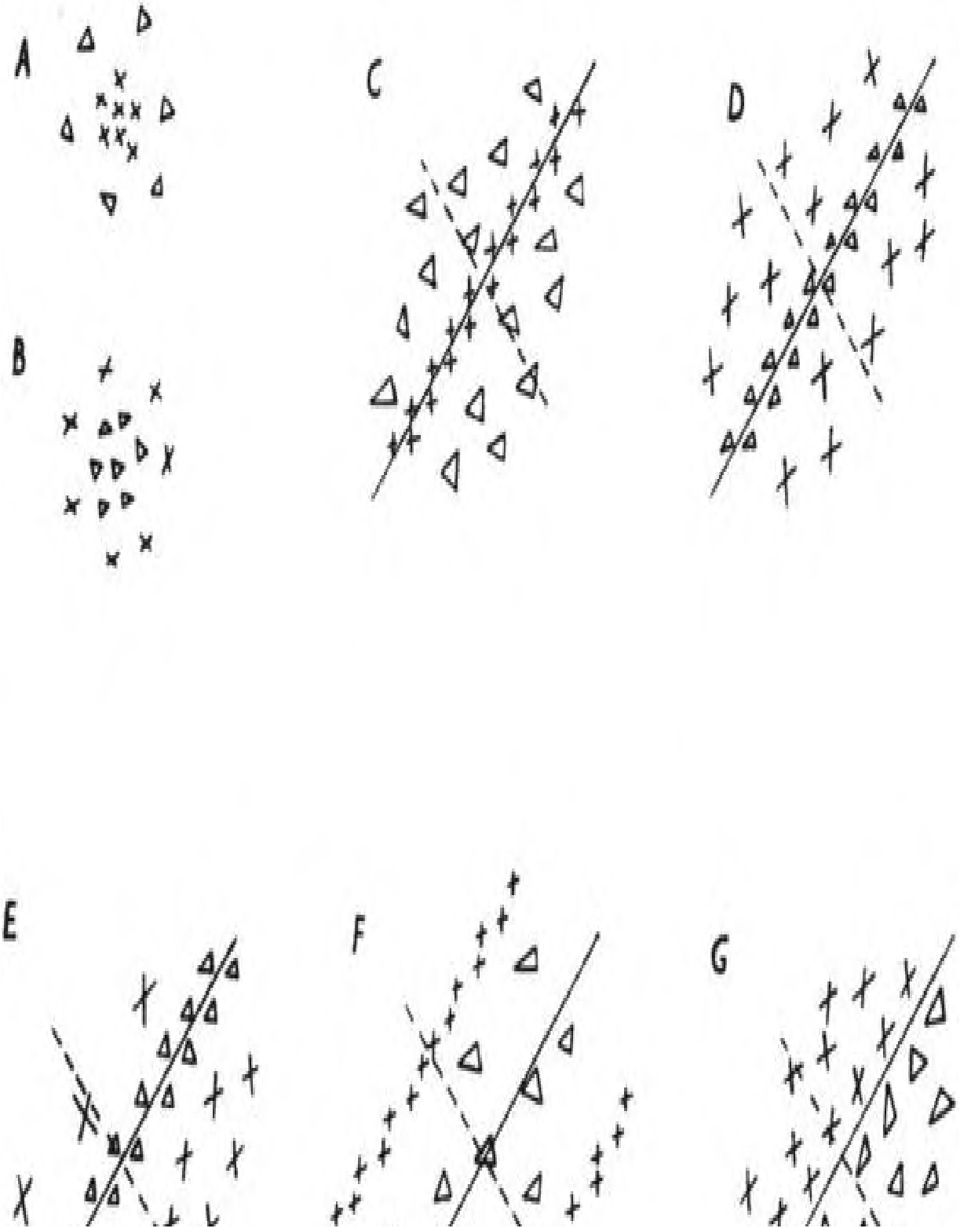
图5-3 视网膜中神经节细胞的响应特性。这两个环形代表了视网膜中两种对大脑发送编码信息，从而让人产生视觉的神经节细胞类型。对于中心给光放电（on-center）类型，光斑投射到中心“给光区域”（+），并且外围“撤光区域”（-）无光时，会产生一系列放电尖峰。而中心撤光放电（off-center）类型则相反，光斑投射到外圈（+），并且中心（-）无光时，会产生一系列放电尖峰。光照的变化包含了研究对象周围的移动刺激和对比边界的重要信息。这些特性是斯蒂芬·库夫勒于1953年发现的。

库夫勒的主要研究兴趣是神经元之间突触的特性。我曾经问他，是什么让他对研究视网膜产生了兴趣。他说，他之前的实验室隶属于约翰·霍普金斯大学的威尔默眼科研究所（Wilmer Eye Institute），他的研究要是跟眼科无关，他会对此感到不安。他在对视网膜中单个神经节细胞的研究中做了些开创性的工作，随后把这个研究项目交给了他实验室的两位博士后——大卫·休伯尔和托斯坦·威泽尔，并建议他们追踪信号进入大脑的过程。1966年，库夫勒和他的博士后研究员转到哈佛医学院，创立了世界上第一个神经生物学系。

大脑皮层中的视觉

休伯尔和威泽尔发现，皮层神经元对定向条形光斑和高对比度边缘的反应比点状光斑更强烈。皮层内的回路已经对输入信号进行了转换。他们描述了两种主要类型的细胞：定向的简单细胞，具有像神经节细胞那样的给光和撤光区域（见图5－4）；以及定向复杂细胞，其

神经元感受野里的任何位置对定向刺激都会产生同样的响应（见图5－5）。



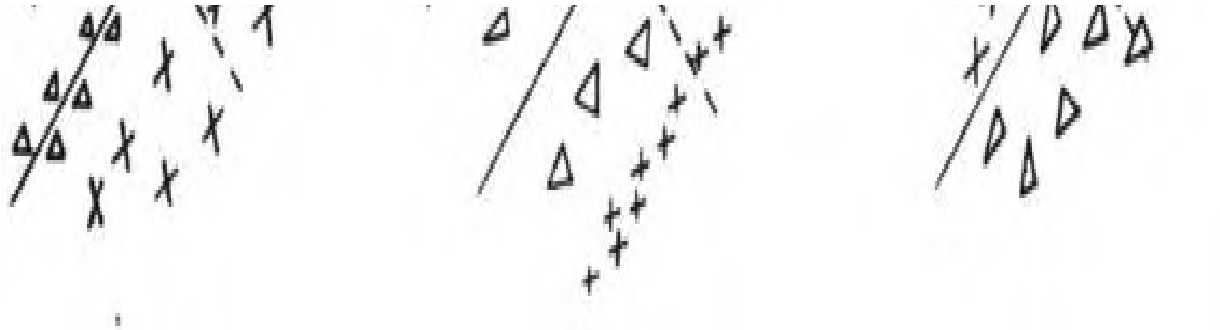
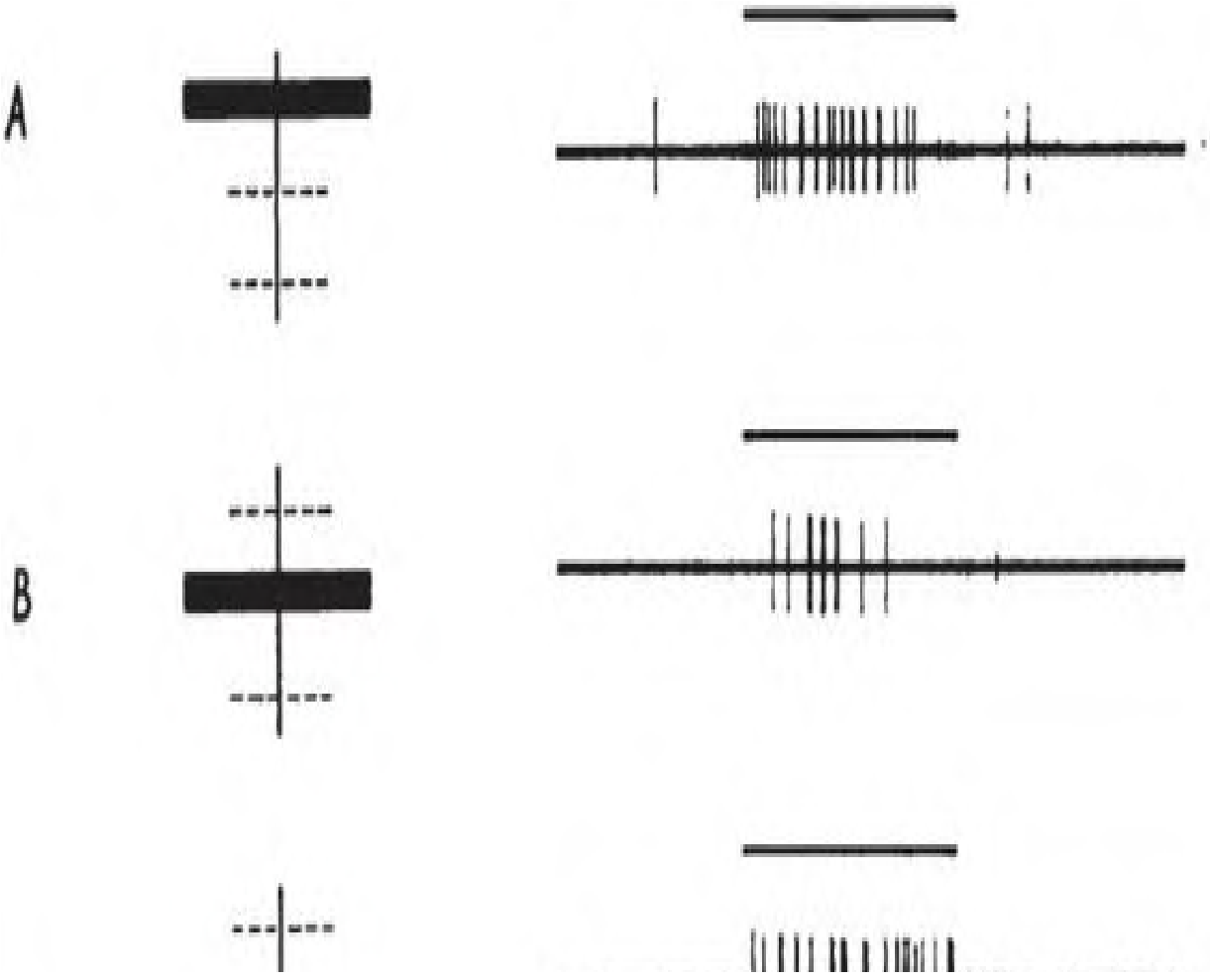


图5-4 猫的初级视觉皮层中简单细胞的感受野。这张图片来自休伯尔和威泽尔在1962年发表的论文中关于发现简单细胞的描述。十字花代表视野中光斑会产生给光响应的位置，而三角形代表撤走光斑会产生撤光响应的位置。（A）视网膜中的中心给光响应细胞（对比图5-3左侧的示意图）。(B)视网膜细胞中的中心撤光响应细胞（对比图5-3右侧的示意图）。（C-G）初级视觉皮层中多样的简单细胞感受域，与视网膜中的感受野相比，所有这些区域都被拉长了，并且给光区和撤光区的分布更加复杂。图片来源：D. H. Hubel and T. N. Wiesel,“Receptive Fields, Binocular Interaction and Functional Architecture in the Cat’s Visual Cortex,”*Journal of Physiology*160, no. 1 (1962): 106-154.2, 图2。



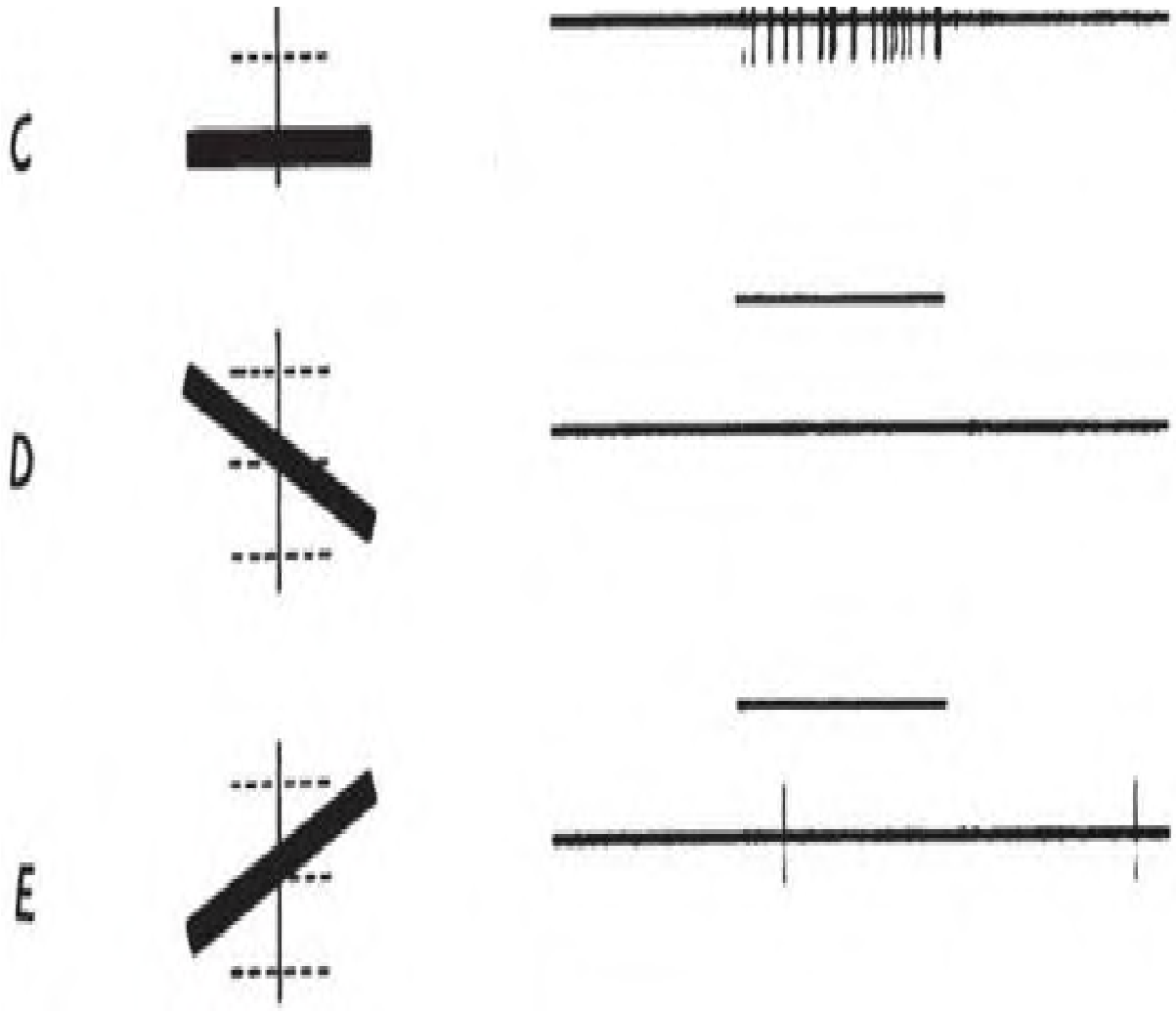


图5-5 猫的初级视觉皮层中一个复杂细胞的响应。这张图片来自休伯尔和威泽尔在1962年发表的论文中关于发现复杂细胞的描述。只要方向正确（图中A、B、C这三条记录），一个长而窄的黑条会引起大量放电（垂直竖线）响应，无论它位于复杂细胞感受野（虚线）内的哪个位置。而非最优方向会导致较弱的响应，或根本没有响应（图中D、E这两个记录）。图片来源：D. H. Hubel and T. N. Wiesel, “Receptive Fields, Binocular Interaction and Functional Architecture in the Cat’s Visual Cortex,” *Journal of Physiology*160, no. 1 (1962):106–154.2, 图7。

视觉皮层中的每个皮层神经元都可以被认为是一个视觉特征检测器。在视野中的特定区域，当某些神经元所偏好的特征信号输入高于某个阈值时，这些神经元就会被激活。每个神经元偏好的特征取决于它与其他神经元的连接。哺乳动物的新皮层（neocortex）有6个特异化的层级。休伯尔和威泽尔还发现，来自两眼的输入在皮层中间层

（4）左右交替排列，这些输入源自丘脑投射的中继站。第4层的单目神经元投射到上层（2和3）的神经元，它们接收双目输入，后者既向上投射到其他皮层区域，又向下投射到底层（5和6），底层又会投射到下皮层。一段视神经元柱中每个细胞的方向偏好和主眼偏好是相同的，并且在整个皮层内平滑地变化（见图5－6）。

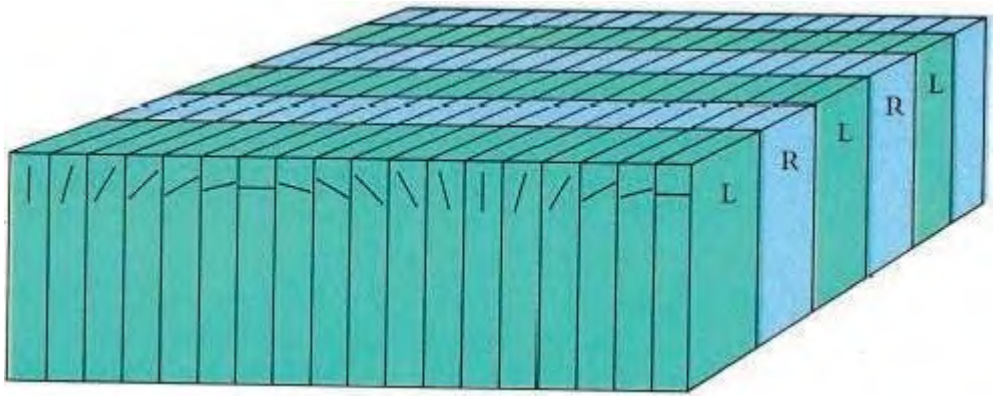


图5-6 初级视觉皮层中一段神经元柱的立方模型。在垂直方向上，所有神经元都具有相同的方向偏好和主眼优势。在每平方毫米的皮层下，有一整套穿过皮层表面缓慢变化的朝向（立方体的前面），以及来自双眼（立方体的右侧）的输入。图片来源：D. Hubel, Eye, *Brain and Vision*(New York: WH Freeman and Company, 1988), 131。

突触的可塑性

如果一只猫的一只眼睛在其生命的最初几个月内被缝合，那么通常由双眼驱动的皮层神经元将变成单眼驱动，只由睁开的那只眼睛支配。[4]单眼剥夺改变了初级皮层中突触的强度，初级皮层是神经元首次接收来自两只眼睛的会聚输入的地方。在初级视觉皮层的皮层可塑性关键期结束后，闭合的眼睛将再也不能影响皮层神经元，这样就会导致一种“弱视”症状。尽管在婴儿中常见的未矫正、未对准或“斜视”状况会大大减少双眼皮质神经元的数目，并且阻止了双眼深度知觉，[5]但如果在关键期内及时进行矫正眼睛的手术，仍然可以挽救双眼神经元。

单眼剥夺是发育早期阶段存在的有关高度可塑性的一个例子，因为环境会影响皮层和大脑其他部位的神经元之间的突触连接。这些依赖于活动的变化凌驾于所有细胞持续不断的更新之上。尽管我们大脑中的大部分神经元与我们出生时的神经元并无二致，[6]但几乎每个神经元的组成部分和连接它们的突触，每天都会发生转变。磨损的蛋白质会被替换，膜中的脂质也会被更新。有了这么多的动态转变，就很难解释记忆是如何在我们的有生之年得以维持的了。

还有另外一种关于记忆持久性的可能解释：它们也许就像我们身体上的伤疤，已经成为我们生活中过往事件的标记。这些标记并非位于不断发生转变的神经元内部，而是位于外部神经元之间的空隙中，那里的细胞外基质由类似疤痕组织中胶原蛋白的蛋白聚糖构成，这种基质是一种可以维持多年的坚韧的材料。[7]如果这个猜想被证明是真实的，那就意味着我们的长期记忆被嵌入了大脑的“外骨骼”中，而我们一直都找错地方了。[8]

突触含有数百种独特的蛋白质，可以控制神经递质的释放和受体神经元上相应受体的激活。在大多数情况下，突触强度可以选择性地在很大范围内增加或减少，在皮层中，这个变化范围是100倍。（在大脑中发现的突触学习算法的实例将在后面的章节中讨论。）更为显著的是，皮层中会不断形成新的突触，并移除旧的突触，这就让它们成为身体中最具活力的细胞器。大脑中有大约100种不同类型的突触，其中谷氨酸是皮层中最常见的兴奋性神经递质，而另一种氨基酸——γ－氨基丁酸（GABA）是最常见的抑制性神经递质。这些神经递质分子对其他神经元的电化学影响伴随着很宽泛的时间过程。例如，第4章中讨论的牛蛙交感神经节细胞的突触，其响应时间尺度可以从毫秒到分钟。

通过阴影脑补立体全貌

史蒂文·祖克（Steven Zucker）（见图5－7）专注于融合了计算机视觉和生物视觉的交叉领域的研究。从我认识他起，他就在写一本解释视觉工作原理的书，到现在已经有30多年了。问题在于史蒂文在视觉研究上不断有新的发现，就像劳伦斯·斯特恩（Laurence Sterne）在小说《项狄传》中描写的主人公项狄（Tristram Shandy）那样，史蒂文那本书也随着主人公的新发现而越变越长。[9]他研究视觉的方法是基于初级视觉皮层的精巧、规则的结构（见图5－6），这是一种不同于皮层中任何其他部位的结构，其中的神经元以近似马赛克式的排列方式组织起来，这种排列的几何解释还有待探究。计算机视觉领域的大多数研究人员希望通过将对象从背景中分割出来，并找到一些可辨识的特征来识别对象。

史蒂文还有更大的野心，想要了解我们如何从表面阴影以及折痕和褶皱中提炼出物体的形状。在2006年神经科学学会的年会上，那个把房子建得像船帆一样的建筑师弗兰克·盖里（Frank Gehry），在接受采访时，被问及他的建筑设计灵感从何而来。[10]他回答说，灵感来自对被揉皱的纸张的观察。但是，我们的视觉系统又是如何通过褶皱和阴影表面的复杂图案，把皱巴巴的纸张的复杂形状拼合起来的呢？我们是如何感知西班牙毕尔巴鄂市（Bilbao）古根海姆博物馆（Guggenheim Museum，见图5－8）外墙那多变的形状的呢？

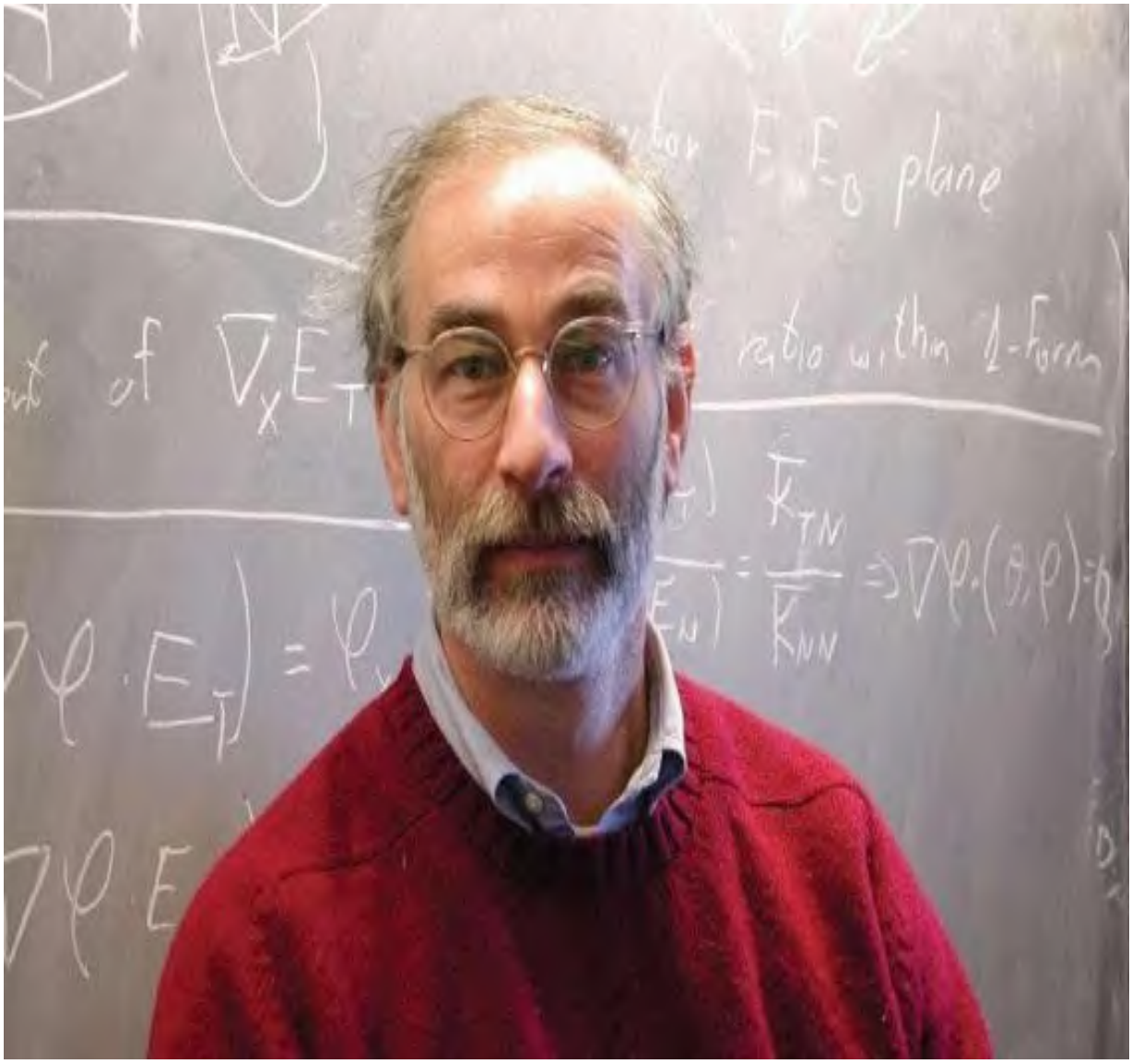


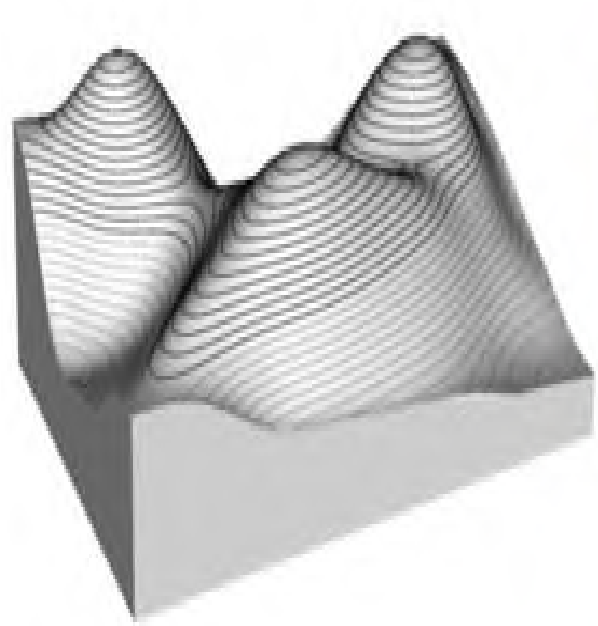
图5-7 耶鲁大学史蒂文·祖克，光从照片右上角打下来。从他毛衣上的阴影变化中，你可以察觉到衣服褶皱的形状。他身后黑板上的方程式解释了为什么我们能识别这种现象，这一灵感来自猴子的视觉皮层。图片来源：史蒂文·祖克。



图5-8 西班牙毕尔巴鄂市的古根海姆博物馆，由弗兰克·盖里设计。来自曲面的阴影和反射形成了强烈的结构感和动感。可以参照通道上的人影大小来体会这幢宏伟建筑的规模。

史蒂文·祖克最近已经能够搞清楚我们是如何在有阴影的图像中看到褶皱的，其背后的解释是基于类似山体等高线图的表面三维轮廓，以及图像上等照度轮廓之间的密切关系（见图5－9）。[11]这种关联源于表面的几何形状。[12]这解释了为什么我们对形状的感知几乎不受照明以及物体表面性质差异的影响。这也可以解释为什么我们

非常善于阅读轮廓明显的等高线地图，以及为什么只需要几条特殊的内线就可以看出漫画中物体的形状。





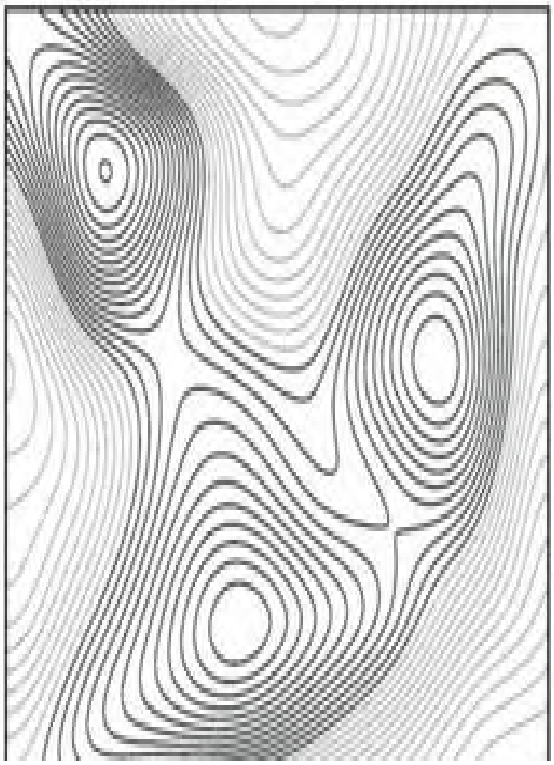






图5-9 同一平面的等高线图（左上）和等照度线（恒定亮度轮廓线，左下）。两者在轮廓右侧显示的临界点之间产生了相同的分区。图片来源：Kunsberg and Zucker, “Critical Contours: An Invariant Linking Image Flow with Salient Surface Organization,”图5。

1988年，西德尼·莱基（Sidney Lehky）和我有了一个想法，我们也许可以训练一个只有一层隐藏单元的神经网络来计算阴影曲面的曲率。[13]我们成功了，而且出人意料的是，隐藏单元的表现跟简单细胞非常相似。但仔细观察后，我们发现并非所有这些“简单细胞”都是相同的。通过查看它们对输出层的投影——该输出层被训练通过使用学习算法（见第8章）来计算曲率——我们发现一些隐藏单元被用于分辨正曲率（凸起）和负曲率（凹陷）（见图5－10）。和一些简单细胞一样，这些单元是检测器，它们往往要么是低响应，要么是高响应，其响应活动呈双峰分布。相比之下，隐藏层中的其他单元具有分级响应的功能，可以像滤波器一样向输出单元发送关于曲率方向和大小的信息。

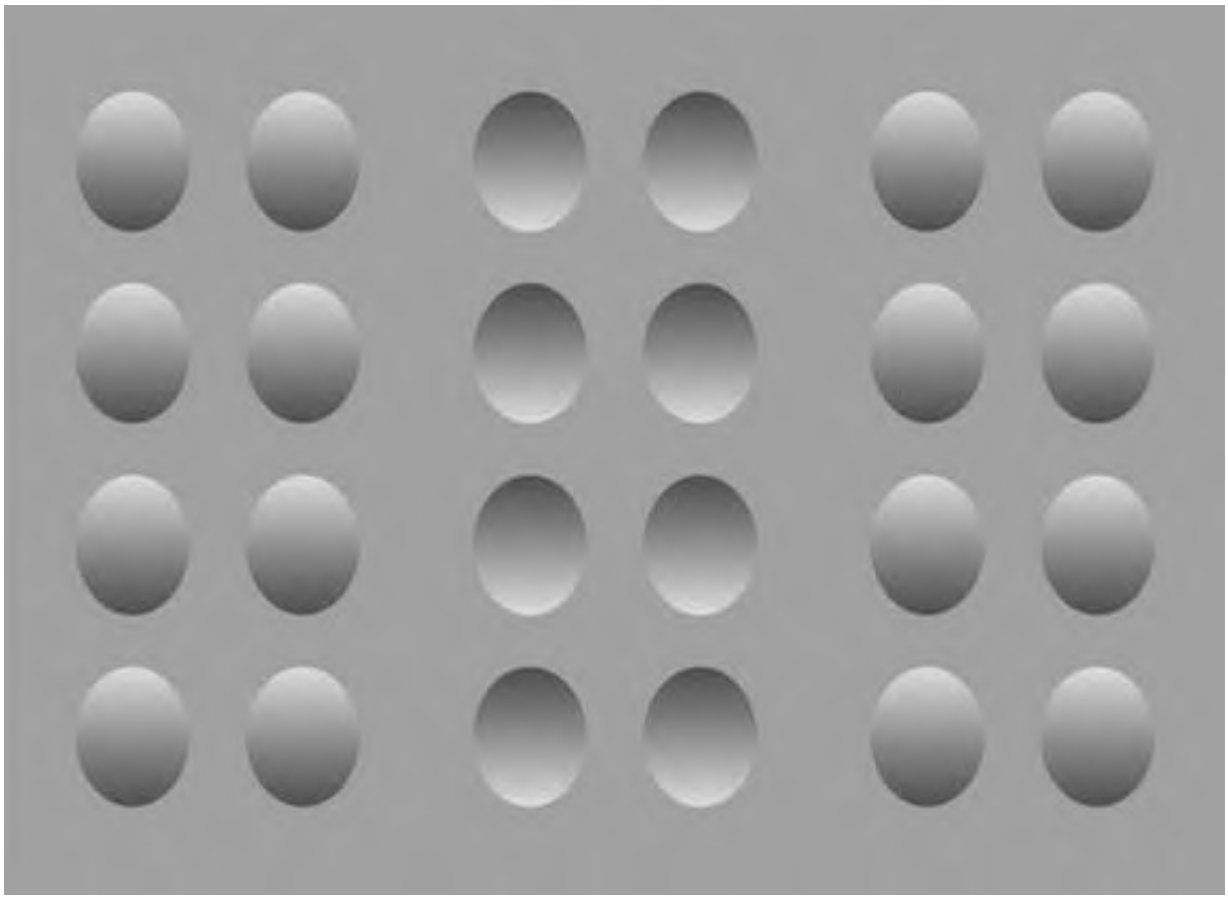


图5-10 阴影中的曲率。我们的视觉系统可以从边界轮廓内亮度的缓慢变化中提取对象的形状。你可以根据阴影方向以及你对照明方向的假设（通常假设为正上方）来分辨图中的突出或凹陷。上下颠倒这幅图，看看它们是否反过来了。图片来源：V. S. Ramachandran, “Perception of Shape from Shading,” *Nature*331, no. 6152(1988), 图2。

这一结论令人惊讶：神经元的功能不仅仅取决于它如何对输入做出反应，而且还取决于它通过自身的“投射域”激活的下游神经元。一个神经元的输出一直以来都比它的输入更难确定，但新的遗传学和解剖学技术使得精确追踪下游的轴突投射成为可能。新的光遗传学技术也使得选择性刺激特定神经元，以探测其对感知和行为的影响成为可能。[14]即便如此，我们的小规模神经网络只能识别突起或凹陷的曲率，而我们仍然不知道在心理学文献中被称为“格式塔”（Gestalts）[15]的整体组织认知是如何在皮层中分布的。

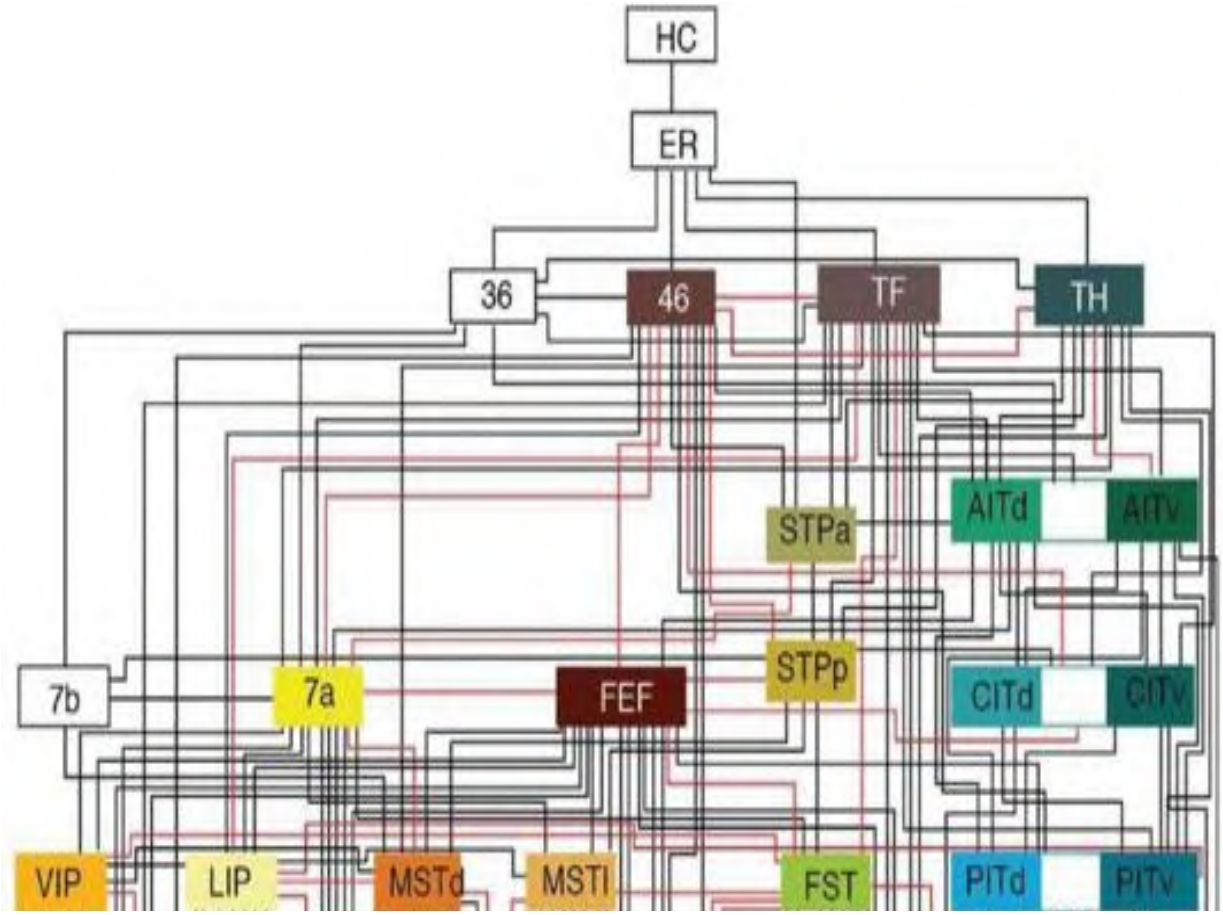
1984年，史蒂文·祖克和我曾被困在丹佛斯台普顿（Stapleton）国际机场，我们的航班因暴风雪而延误了。那会儿我们对当时正处于起步阶段的计算神经科学感到兴奋不已，于是盘算着创建一个研讨会，将计算和实验研究人员聚集在一起。我们决定把地点定在伍兹霍尔，我曾经在那里参加过一个神经生物学的夏季课程，并且在之后好几个暑期中又回去，与斯蒂芬·库夫勒一起在海洋生物实验室做过生理实验。伍兹霍尔是鳕鱼角附近的一个美丽的村子，离波士顿不远。多年来，许多视觉研究领域的领军人物都参加了这个年度研讨会，这对我来说是另一个科学制高点。视觉皮层计算理论就诞生于这些研讨会，尽管对该理论的确认还需要30年。（在第9章中，我们将看到最成功的深度学习网络的结构与视觉皮层的结构非常相似。）

视觉区域的层级结构

乔恩·卡斯（Jon Kaas）和约翰·奥尔曼（John Allman）于 20世纪70年代早期，在威斯康星大学神经生理学系研究从初级视觉皮层接受输入的皮层区域，发现了不同区域具有不同的特性。例如，在一个他们称为“中间颞叶皮层”（middle temporal cortex）或“MT”的区域，他们发现了一些视觉区域，它们的神经元对定向移动的视觉刺激能够做出反应。奥尔曼向我提到，他们当时很难让系主任克林顿·伍尔西（Clinton Woolsey）接受这个发现。卡斯和奥尔曼用较先进的记录技术发现了这些纹外视觉皮层（extrastriate visual cortex）的区域，而在较早的实验中，伍尔西由于使用的记录技术相对粗糙，与这一发现擦肩而过。[16]最近的研究在猴子的视觉皮层中找到了二十几个视觉区域。

1991年，还在加州理工学院的大卫·范·艾森（David Van Essen）仔细研究了皮层每个视觉区域的输入和输出，并将它们按层级排列了出来（见图5－11）。这张图有时仅仅被用于说明皮层的复杂

性。它就像一座大城市的地铁图，其中方框代表地铁站，各种颜色的线代表地铁线路。视网膜神经节细胞（RGC）的视觉输入投射到图表底部的初级视觉皮层（V1）。从那里开始，信号被向上传送到层级结构中，每个区域专门负责视觉的不同方面，例如形态感知。靠近图右侧的层级示意图顶部，下颞叶皮层的前部、中部和后部（AIT、CIT和PIT）区域中的神经元的感受野覆盖了整个视觉区域，并对复杂的视觉刺激如脸部和其他对象做出优先反应。虽然我们不知道神经元是如何做到这一点的，但我们确实知道连接的强度可以通过经验来改变，如此一来，神经元就可以学习如何对新对象做出反应。范·艾森后来去了圣路易斯的华盛顿大学，在那里他被任命为NIH资助的人脑连接组项目（HCP）的联合主任。[17]他的研究小组的工作是使用基于磁共振成像（MRI）技术，[18]来生成人体大脑皮层中的长程连接图（见图5－12）。



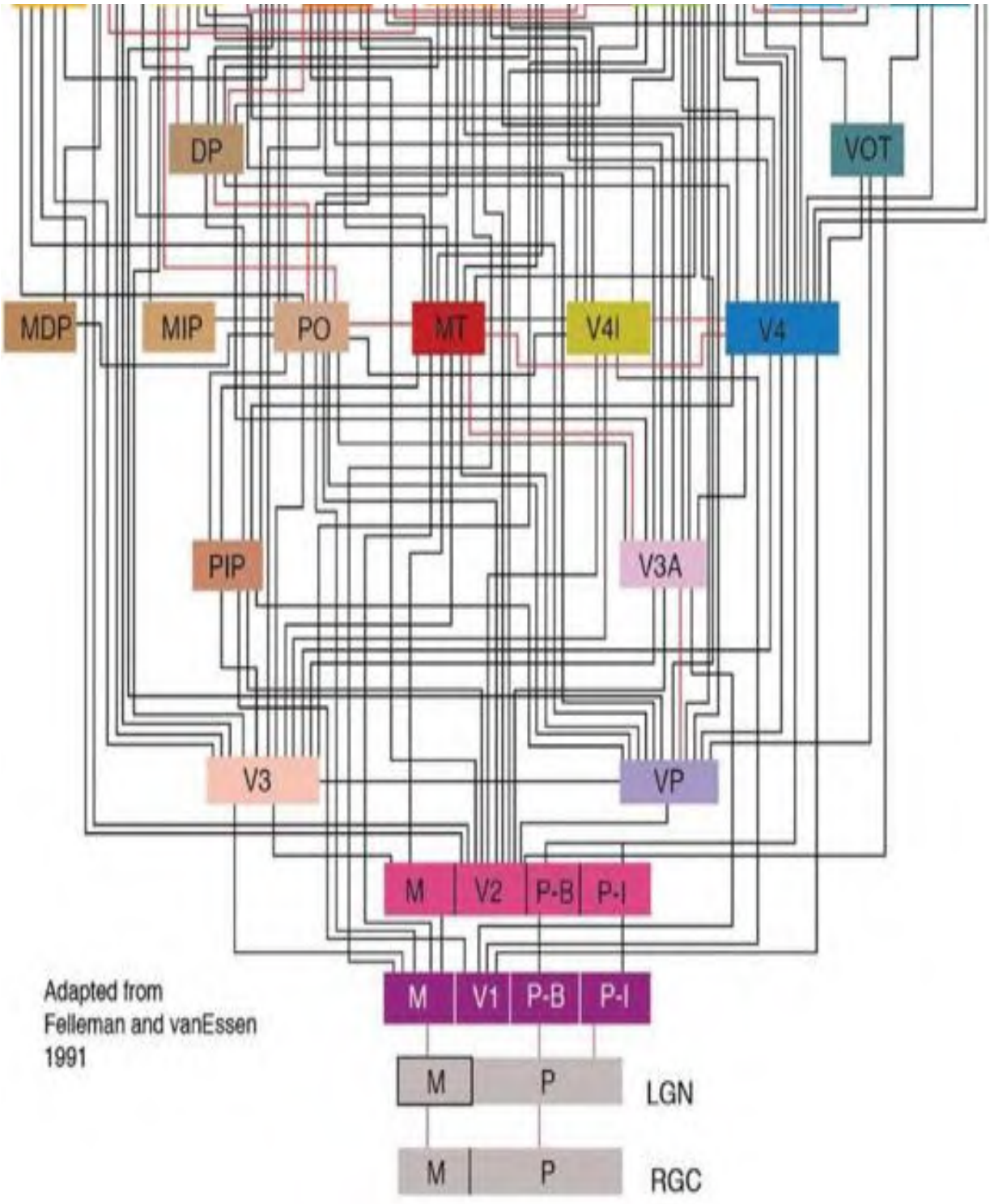


图5-11 猴脑中视觉区域的层级结构图。视觉信息从视网膜神经节细胞（RGC）映射到丘脑的外侧膝状核（LGN），其中继细胞投射到初级视觉皮层（V1）。皮层区域的层级结构终止于海马体（HC）。图中所有的187个连接几乎都是双向的，有源自较低区域的前馈连接和源自较高区域的反馈连接。图片来源：D. J. Fellemanand D. C.

Van Essen, “Distributed Hierarchical Processing in Primate Visual Cortex,” *Cerebral Cortex*1,no. 1 (1991): 30, 图4。

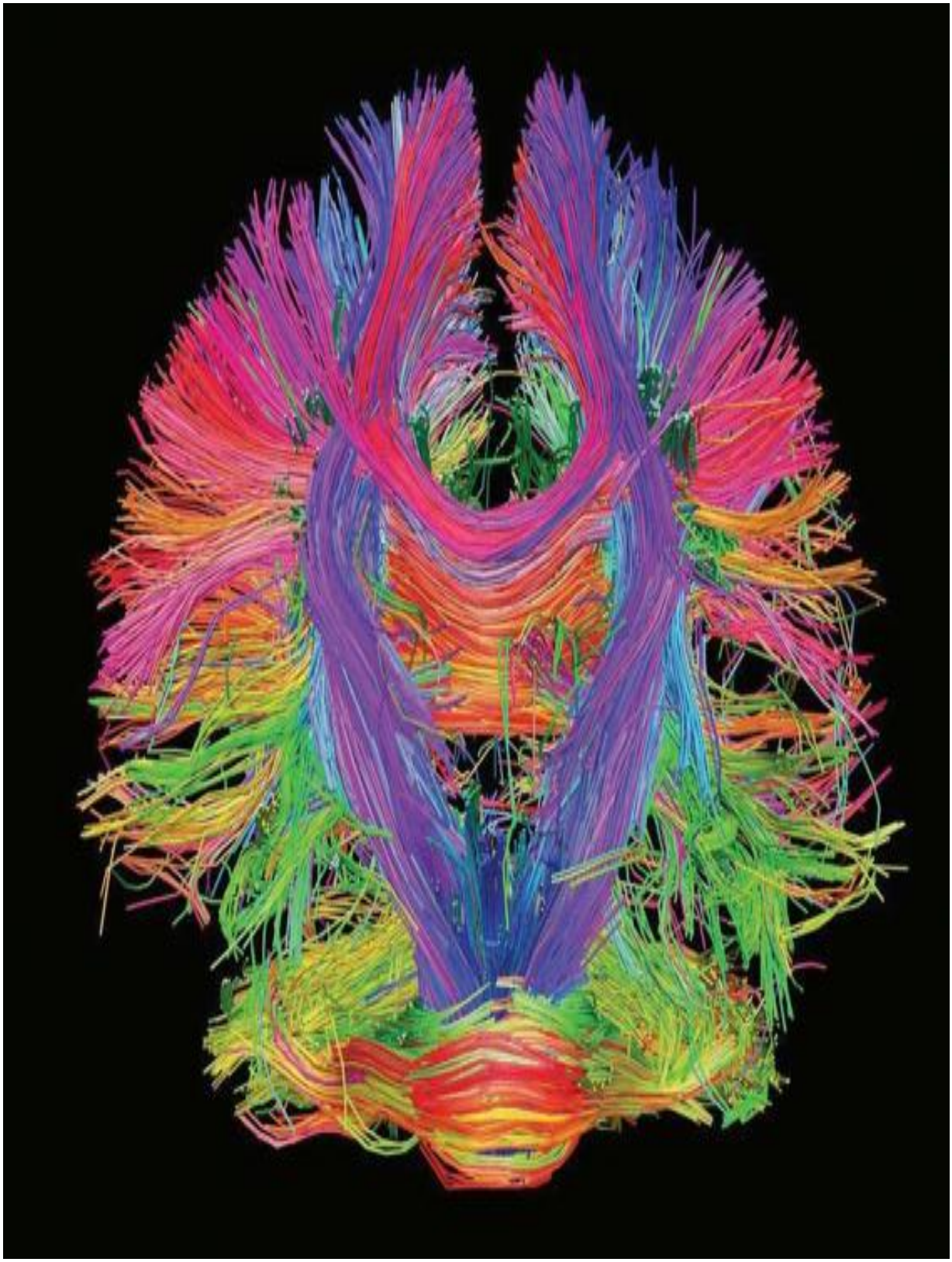


图5-12 人脑连接组。基于水分子不均匀扩散的磁共振成像能够以一种非侵入性的方式追踪脑白质中的长程纤维束。不同的颜色标注了不同的路径方向。图片来源：The

Human Connectome Project。

认知神经科学的诞生

1988年，我曾服务于麦克唐奈和皮尤基金会（McDonnell and Pew Foundations）的一个委员会，需要走访知名的认知科学家和神经科学家，就如何启动一个名为“认知神经科学”的新领域听取建议。[19]该委员会辗转于世界各个国家，与专家会面，听取他们关于哪些科学专题最有前途，以及在哪里建设新的认知神经科学研究中心的建议。8月的一个烈日炎炎的下午，我们在哈佛大学教员俱乐部征询了杰瑞·福多（Jerry Foder）的意见，他是思想语言方面的专家，也是模块化思想的倡导者。他开门见山，直接驳斥了我们的战略，“认知神经科学不是一门科学，而且永远不会是”。他给人的印象是，他已经阅读了关于视觉和记忆的所有神经科学论文，但这些文章都没有达到他的标准。但是当他评论说“麦当劳基金会正在浪费它的资金”时，麦克唐奈基金会主席约翰·布鲁尔（John Bruer）立刻就指出，福多把他的基金会和街头制作汉堡包的地儿搞混了。

福多丝毫不为所动。他解释了为什么头脑应该被当作运行智能计算机程序的模块化符号处理系统。加州大学圣迭戈分校的哲学家帕特里夏·丘奇兰德（Patricia Churchland）问他的理论是否也适用于猫。“是的，”福多说，“猫正在运行猫程序。”但是当NIH研究视觉和记忆的神经科学家莫蒂默·米什金（Mortimer Minshkin）让他介绍他自己的实验室的发现时，福多嘟囔着说了一些我听不太懂的话，好像是关于在语言实验中的事件相关电位。幸运的是，那一刻，消防演习的警铃响了，我们都被要求去室外等待。站在院子里，我听到米什金对福多说：“那些都是雕虫小技罢了。”演习结束后，福多就再没出现。

认知神经科学已经发展成为一个重要的领域，吸引了各个科学领域的很多研究人员，包括社会心理学和经济学，这些领域以前与神经科学很少有或没有任何直接联系。使这一切成为可能的是在20世纪90年代早期采用的非侵入式大脑活动可视化方法，尤其是功能磁共振成像（fMRI），其目前的空间分辨率为几毫米。fMRI生成的庞大的成像数据集被交由新的计算方法进行分析，如独立分量分析（将在第6章讨论）。

由于大脑在缺氧的情况下无法运转，并且血流在亚毫米级别被严格控制，fMRI所测量的血氧水平依赖（BOLD）信号就被当作大脑活动的替代指标。血液中的含氧量改变了它的磁性，可以用fMRI进行无创监测，并且以几秒的时间分辨率产生大脑活动的动态图像。这种时间分辨率足够追踪在实验过程中大脑的哪些部分在活动。fMRI已经被用于探索视觉层级结构不同部分的时值整合。

普林斯顿大学的尤里·哈森（Uri Hasson）进行了一项fMRI实验，旨在探究视觉层级的哪些部分涉及处理不同长度的电影。[20]查理·卓别林（Charlie Chaplin）的无声电影被剪辑为4秒、12秒和36秒的片段呈现给受试者。在4秒的剪辑中，受试者可以识别一个场景；12秒时，可以看清连接的动作；在36秒的长度下，能够看到一个有开头和结尾的故事。在层级底部的初级视觉皮层中的fMRI反应，无论在什么样的时间尺度上，都强大且可靠。但在视觉等级的较高层次上，只有较长的时间尺度才能引起可靠的反应，而位于层级顶层的前额叶皮层区需要最长的时间间隔。这与其他实验结果一致，即工作记忆也按照层级分布。工作记忆是我们掌握信息的能力，比如要记住的电话号码，以及我们正在处理的任务的要素。最长的工作记忆时间尺度同样位于前额叶皮层。

作为神经科学中最让人兴奋的学术领域之一，对大脑学习行为的研究可以在不同的层面进行，从分子层面到行为层面。

[1] 由克里克、拉马钱德兰（V. S. Ramachandran）和戈登·肖（Gordon Shaw）在20世纪80年代建立的亥姆霍兹俱乐部持续活跃了20多年。关于它的历史，可以参阅C. Aicardi, “Of the Helmholtz Club, South-Californian Seedbed for Visual and Cognitive Neuroscience, and Its Patron Francis Crick,” *Studies inHistory and Philosophy of Biological and Biomedical Sciences*45, no. 100 (2014):1–11。

[2] 用一位满意的与会者的话来说：“我从遇到的每个人身上都学到了很多。我对采纳别人的想法一点也不觉得难为情。……对我来说，最有收获的学习经历就是这个亥姆霍兹俱乐部。我不知道你是否听说过这个团体。……他们大概有20个人。我从来没有错过一次他们的例会。我还得找人替我上课，这个会比我的课还重要，我绝对不能错过。” Carver Mead, in James A. Anderson and Edward Rosenfeld, eds., *Talking Nets: An Oral History of Neural Networks*(Cambridge, MA: MIT Press, 2000), 138。

[3] R. Desimone, T. D. Albright, C. G. Gross, and C. Bruce, “Stimulus-Selective Properties of Inferior Temporal Neurons in the Macaque,” *Journal of Neuroscience*4, no. 8 (1984): 2051–2062. 查尔斯·格罗斯实验室的许多研究人员都留了胡子，因此，对马桶刷做出反应的视觉皮层中的神经元可能是胡须细胞。

[4] David Hubel, *Eye, Brain, and Vision*(New York: W. H. Freeman, 1988), 191–216.

[5] 猫的关键时期从3周到几个月不等，而人类是从几个月到七八岁不等。关键期的结束可能不会像以前认为的那样突然，在经过高强度的练习后，接受过斜视矫正的成人仍然可以获得立体视觉的能力。参阅Susan R. Barry, *Fixing My Gaze: A Scientist’s Journey into Seeing in Three Dimensions*(New York: Basic Books, 2009)。当我还是普林斯顿大学的研究生时，就认识了Barry，她现在被叫作“立体苏”（Stereo Sue）。

[6] 这条规则也存在一些例外的情况：海马体中齿状回的颗粒细胞和嗅球中的神经元在我们的一生当中都会不断更新。参阅Michael Specter, “Rethinking the Brain: How the Songs of Canaries Upset a Fundamental Principle of Science,” *NewYorker*, July 23, 2001, http://www. michaelspecter.com/2001/07/rethinking-thebrain/。

[7] Terrence Sejnowski, “How Do We Remember the Past?” in John Brockman, ed.,*What We Believe but Cannot Prove: Today’s Leading Thinkers on Science in the Age of Certainty*(London: Free Press, 2005), 97–99; and R. Y. Tsien, “Very Long Term Memories May Be Stored in the Pattern of Holes in the Perineuronal Net,” *Proceedings of the National Academy of Sciences of the United States of America*110, no. 30 (2013): 12456–12461.

[8] 在阿尔兹海默症中，细胞外基质的完整性受到损害，这可能导致了长期记忆的丧失。John Allman, private communication, July, 2017.

[9] 小说中的主人公也是故事的叙述者，按照“头脑中思想的延续”来描述若干杂乱的事件，依据主人公脑海中事件出现的顺序来叙事，而非按照传统的时间顺序。——译者注

[10] 关于盖里演讲的概述，可以参阅Shelley Batts, “SFN Special Lecture:Architecture Frank Gehry and Neuro-Architecture,” *Science Blogs*, posted October 15, 2006, http://scienceblogs.com/retrospectacle/2006/10/15/sfn-special-lecturearchitect-1/。

[11] B. S. Kunsberg and S.W. Zucker, “Critical Contours: An Invariant Linking Image Flow with Salient Surface Organization,” May 20, 2017, https://arxiv.org/pdf/1705.07329.pdf.

[12] 在山脉的等高线图上看到的表面的三维轮廓，和图像上的等照轮廓之间的联系，可以用表面上的临界点和梯度流的几何形状来解释，后者被叫作“Morse Smale复形”。

[13] S. R. Lehky and T. J. Sejnowski, “Network Model of Shape-from-Shading: Neural Function Arises from Both Receptive and Projective Fields,” *Nature*333, no. 6172(1988): 452–454.

[14] Terrence J. Sejnowski, “What Are the Projective Fields of Cortical Neurons?”in J. Leo van Hemmen and Terrence J. Sejnowski, eds. *23 Problems in SystemsNeuroscience*(New York: Oxford University Press, 2005), 394–405.

[15] 格式塔学派主张人脑的工作原理是基于整体的，其效果不同于其各个部件功能的总和。——译者注

[16] C. N. Woolsey, “Cortical Localization as Defined by Evoked Potential and Electrical Stimulation Methods,” in G. Schaltenbrand and C. N. Woolsey (eds.),*Cerebral Localization and Organization*(Madison: University of Wisconsin Press,1964), 17–26; J. M. Allman and J. H. Kaas, “A Representation of the Visual Field in the Caudal Third of the Middle Temporal Gyms of the Owl Monkey (*Aotustrivirgatus*),” *Brain Research*31 (1971): 85–105.

[17] L. Geddes, “Human Brain Mapped in Unprecedented Detail: Nearly 100 Previously Unidentified Brain Areas Revealed by Examination of the Cerebral Cortex,” *Nature*, July 20, 2016. doi:10.1038/nature.2016.20285.

[18] 其中一种技术——弥散张量成像（diffusion tensor imaging，DTI），能够追踪构成大脑皮层中白质的轴突的方向。

[19] Elizabeth Penisi, “Two Foundations Collaborate on Cognitive Neuroscience,”*Scientist*, October 1989, http://www.the-scientist.com/?

articles.view/articleNo/10719/title/Two-Foundations-Collaborate-On-Cognitive-Neuroscience/.

[20] U. Hasson, E. Yang, I. Vallines, D. J. Heeger, and N. Rubin, “A Hierarchy of Temporal Receptive Windows in Human Cortex,” *Journal of Neuroscience*28, no.10 (2008): 2539–2550.

第二部分

深度学习

的演进

要事年表

1949年

唐纳德·赫布（Donald Hebb）出版了《行为的组织》（*The Organization of Behavior*）一书，提出了有关突触可塑性的赫布定律（Hebb Rules）。

1982年

约翰·霍普菲尔德（John Hopfield）发表了文章《具有突发性集体计算能力的神经网络和物理系统》（Neural Networks and Physical Systems with Emergent Collective Computational Abilities），其中介绍了霍普菲尔德网络（Hopfield net）。

1985年

杰弗里·辛顿和我发表了文章《玻尔兹曼机的一种学习算法》（A Learning Algorithm for Boltzmann Machines），这是马文·明斯基和西摩尔·帕普特被广泛接受的观点的反例，后者认为为多层网络开发学习算法是不可能的。

1986年

大卫·鲁姆哈特（David Rumelhart）和杰弗里·辛顿发表了文章《通过误差传播学习内在表征》（Learning Internal Representations by Error-Propagation），其中介绍了当前用于深度学习的“反传”（backprop）学习算法。

1988年

理查德·萨顿（Richard Sutton）在《机器学习》杂志上发表了文章《通过时序差分方法学习预测》（Learning to Predict by the

Methods of Temporal Differences）。时序差分学习现在被认为是在所有大脑中进行奖励学习的算法。

1995年

安东尼·贝尔（Anthony Bell）和我发表了文章《一种用于盲分离和盲解卷积的信息最大化方法》（An Information-Maximization Approach to Blind Separation and Blind Deconvolution），描述了一种用于独立分量分析（Independent Component Analysis, ICA）的无监督算法。

2013年

杰弗里·辛顿在2012年NIPS会议上发表的论文《深度卷积神经网络下的ImageNet分类》（ImageNet Classification with Deep Convolutional Neural Networks），将图像中对象分类的错误率降低到了18%。

2017年

深度学习网络程序AlphaGo，击败了围棋世界冠军柯洁。

06
语音识别的突破

在一个拥挤的鸡尾酒会上，当空气中弥漫着其他人谈笑的嘈杂声时，想要听见你面前的人说话是很费劲儿的。你的两只耳朵有助于你把听力集中在正确的方向，而你的记忆可以填补没听清的谈话内容。现在想象一下，房间里正在举行一个有100人参加的鸡尾酒派对，四周还有100个以任意方向排布的麦克风。每个麦克风都会采集所有人的声音，但是每个麦克风针对每个人采集的声音振幅比不同。是否有可能设计出一种算法，将每种声音分离成单独的输出声道？或者更复杂点，如果声音来源是未知的，比如混合了音乐、拍手声、自然声音，甚至是随机的噪声呢？这就是所谓的“盲源分离问题”（见图6-1）。

1986年4月13—16日，在犹他州雪鸟城举办的神经网络计算会议，也就是AIP会议（NIPS大会的前身）上，出现了一个题为《利用神经网络模型进行空间或时间自适应信号处理》（Space or Time Adaptive Signal Processing by Neural Network Models）的墙报。它的作者詹妮·埃罗（Jeanny Herault）和克里斯蒂安·于滕（Christian Jutten）使用了一种学习算法，对输入给神经网络模型的混合正弦波（均为纯频率）进行了盲分离，并指出了一类新的无监督学习算法。[1]虽然当时还不知道是否存在一种可以盲分离其他类型信号的通用解决方案，但十年之后，安东尼·贝尔和我发现了一种可以解决一般问题的算法。[2]

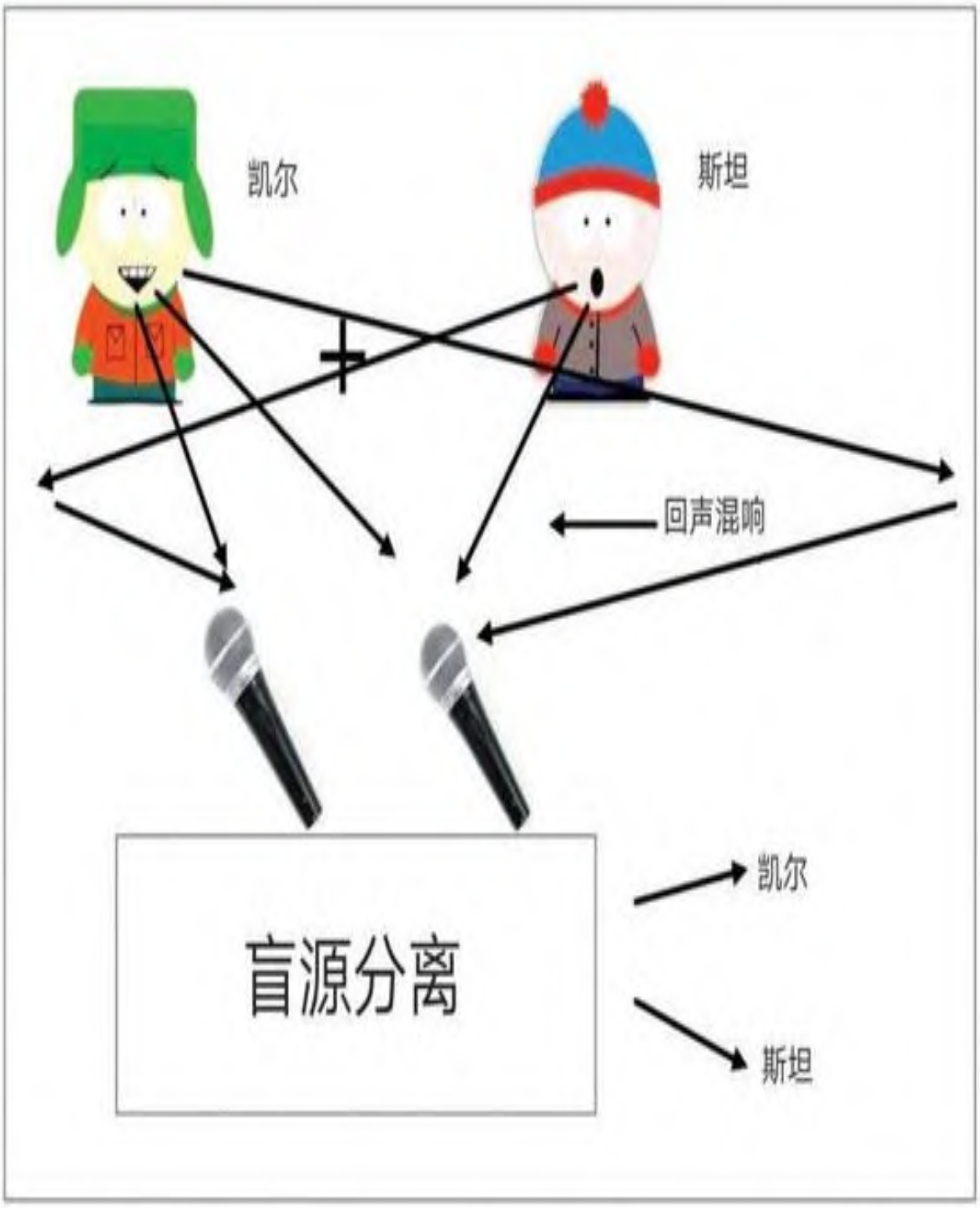


图6-1 盲源分离。凯尔和斯坦在有两个麦克风的房间里同时说话。每个麦克风会从扬声器以及房间墙壁反射的声音里采集信号。该实验的挑战就在于，如何在对信号的相关信息一无所知的情况下将两个声音彼此分开。独立分量分析是一种学习算法，可以在不了解信息源的情况下解决这个问题。

在嘈杂中找到你的声音

感知器是一个有单一神经元的神经网络。复杂度略高于感知器的次简单网络体系结构在输出层中有多个模型神经元，每个输入神经元都会连接到一个输出神经元，该结构将输入层中的模式转换为输出层中的模式。这个网络不仅可以对输入进行分类，还可以学习如何进行盲源分离。

1986年，安东尼·贝尔（见图6－2）还是一名在苏黎世联邦理工学院（ETH Zurich）进行暑期实习的本科生。他很早就对神经网络产生了兴趣，曾前往日内瓦大学聆听来自神经网络领域前辈们的4个演讲。在布鲁塞尔大学获得博士学位后，他于1993年前往拉荷亚，加入了我的实验室，担任博士后研究员。



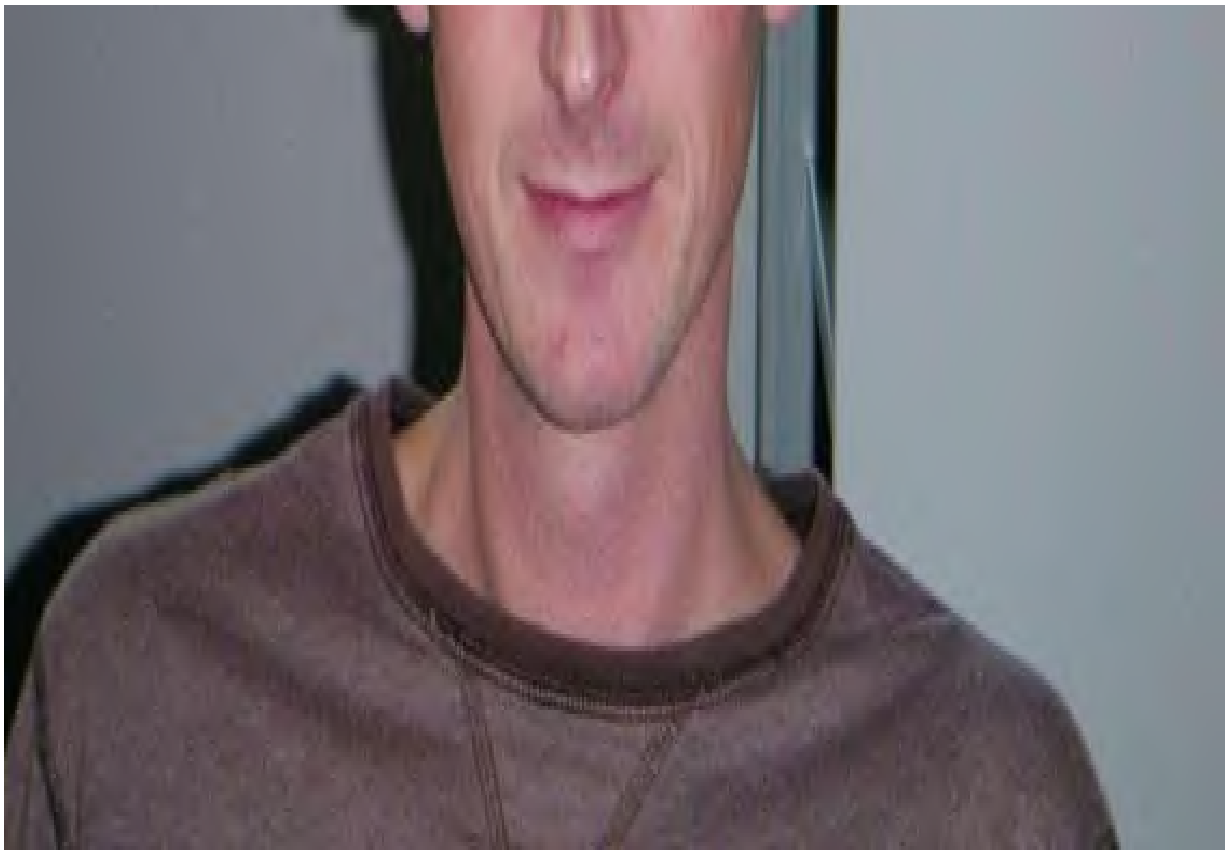


图6-2 正在独立思考的安东尼·贝尔，当时他正在从事独立分量分析研究（照片拍摄于1995年前后）。某个领域的专家试图解决一个问题时，常常无法实现突破，有时反倒是第一次接触这个问题的人能找到新方法并解决问题。安东尼和我发现了一种迭代算法来解决盲源分离问题，这个算法现在已经是工程学教科书里的内容，并且有了数千个应用实例。图片来源：安东尼·贝尔。

“通用信息最大化学习原理”（general infomax learning principle）最大限度地提高了通过网络传递的信息量。[3]安东尼当时正在研究树突中的信号传输。树突就像细长的电缆，大脑的神经元通过附着在树突上的数以千计的突触收集信息。他凭直觉认为，通过改变树突中离子通道的密度，可以最大限度地提高树突中流通的信息量。在简化问题的过程中（忽略树突），安东尼和我发现了一种新的信息理论学习算法，我们称之为“独立分量分析（ICA）”，它解决了盲源分离问题（见方框6.1）。[4]

独立分量分析已经有了数千个应用实例，而且已经出现在了信号处理的教科书中。[5]在被应用于户外场景的自然图像的碎片中时，独立分量分析的独立成分是局部化的定向边缘滤波器（见图6－3），类似于猫和猴子的视觉皮层中的简单细胞（见图5－4）。[6]采用独立分量分析方法时，只需要很少的信息源就能重建一幅图像中的一个图块，这种重建在数学上被称为“稀疏性”（sparse）。[7]

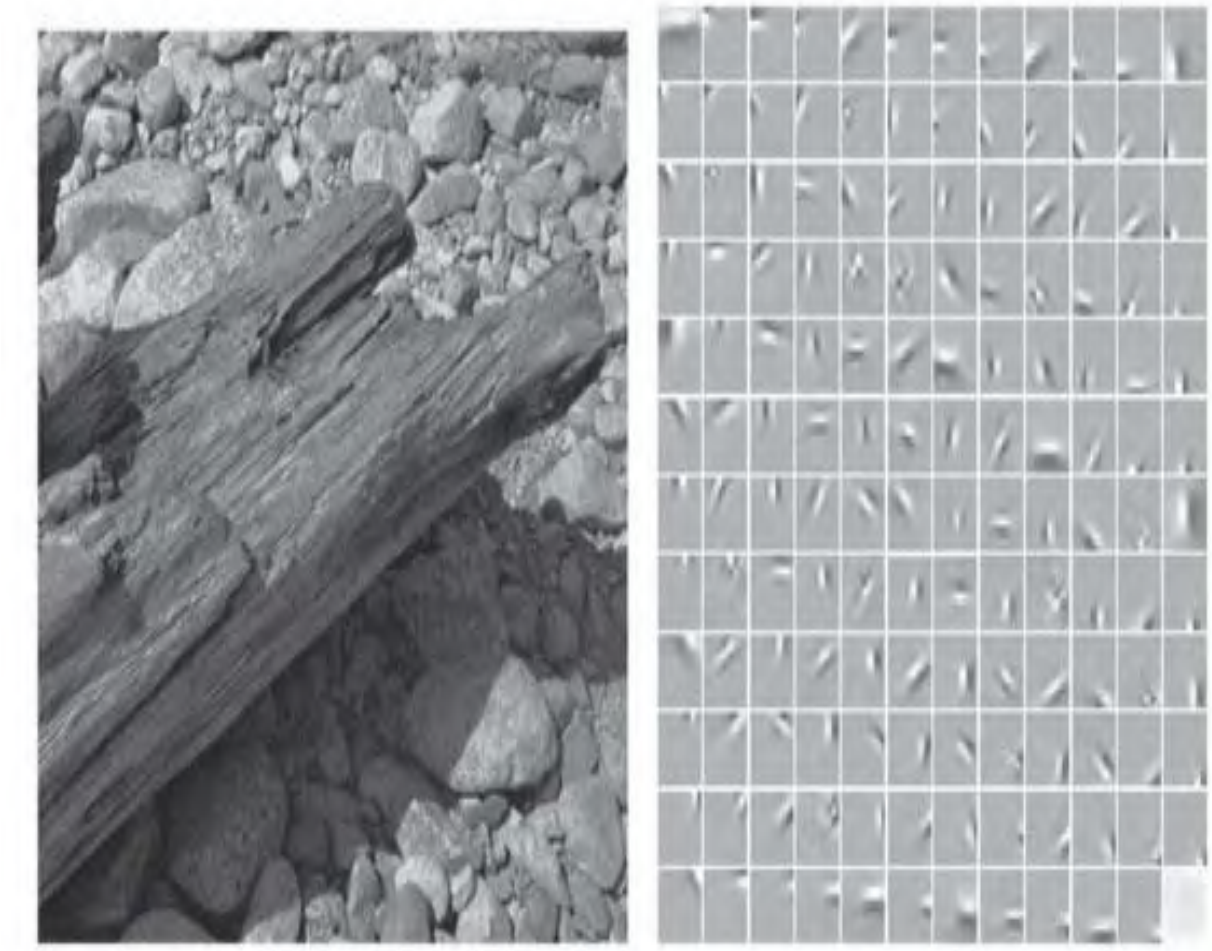
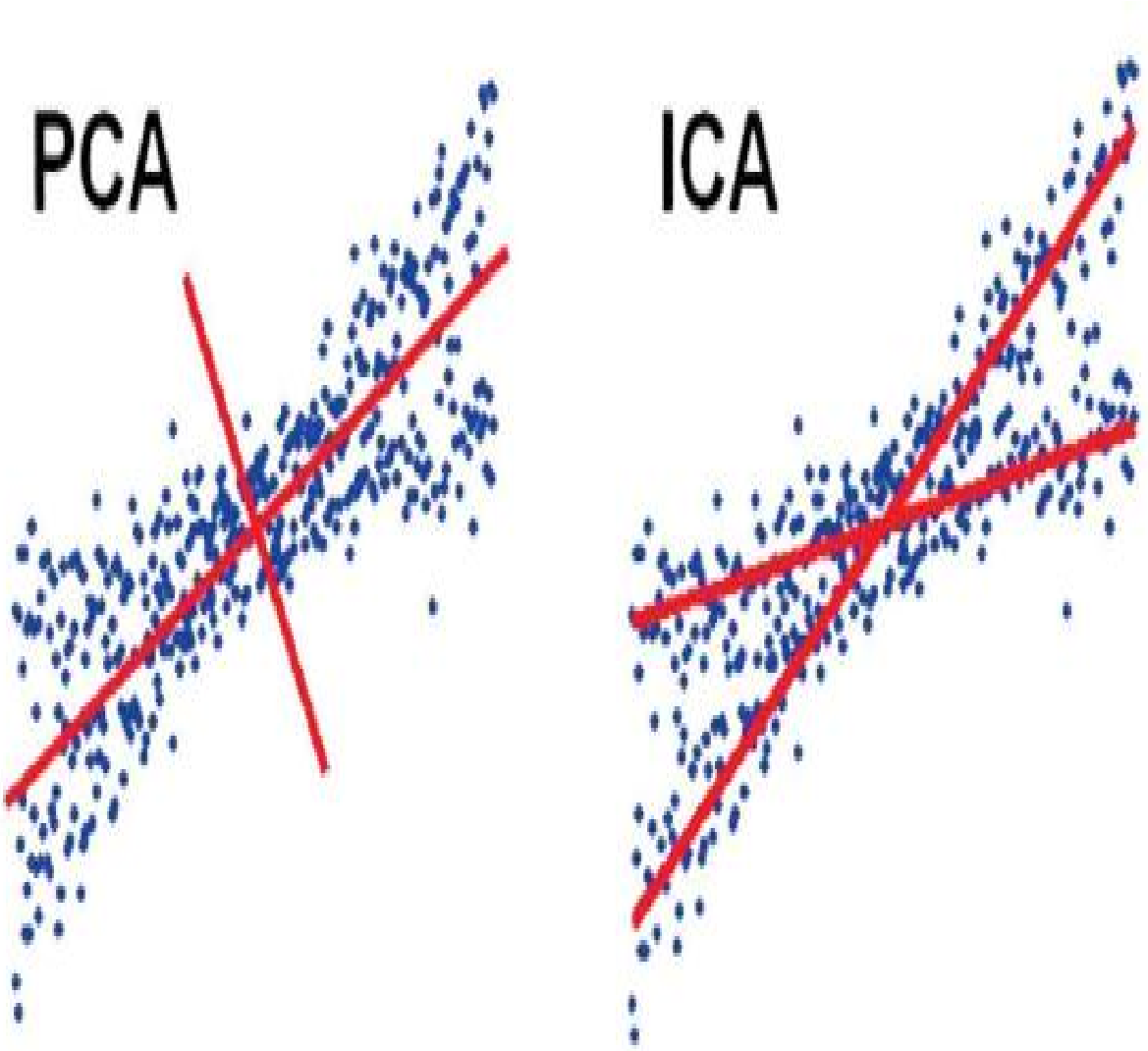


图6-3 源自自然图像的独立分量分析滤波器。使用左图自然场景图像中的图块（12×12像素）作为输入，通过具有144个输出单元的独立分量分析网络得到输出。右图中产生的独立分量与在初级视觉皮层中发现的简单细胞类似：它们具有局部性和定向性，分为正区域（白色）和负区域（黑色），灰色区域代表零点。只需要几个滤波器就可以表示任何给定的图块，这一性质被称为“稀疏性”。左图来源：Michael Lewicki；右图来源：Bell and Sejnowski, “The ‘Independent Components’ of Natural Scenes Are Edge Filters,” figure 4。

6.1

独立分量分析是如何工作的



主分量分析（PCA）和独立分量分析（ICA）之间的比较。图片中的圆点表示两个麦克风分别在垂直和水平轴上的输出。每个点的坐标是单个时间点记录在两个麦克风上的值。PCA是一种非常受欢迎的无监督学习技术，它挑选出将两个信号等分的方向，将它们的混合程度最大化，并且PCA轴始终相互垂直。ICA找到沿点的方向伸展的轴，来表示分离的信号，这些轴可能不是垂直的。

这些结果证实了20世纪60年代著名视觉科学家霍勒斯·巴洛（Horace Barlow）的猜想，那时大卫·休伯尔和托斯坦·威泽尔已经在视觉皮层中发现了简单细胞。一个图像通常包含大量的冗余信息，因为附近的像素通常具有相似的值（例如天空的像素）。巴洛猜想，通过减少自然场景表达中的冗余，[8]简单细胞就能够更有效地传输图像中的信息。整个学界花了50年的时间，才开发出数学工具来证实他的直觉。

安东尼和我还认为，独立分量分析被应用于自然声音时，独立分量是具有不同频率和持续时间的时域滤波器，类似于在听觉神经系统初期处理部分中发现的滤波器。[9]这使得我们相信，在试图理解感观信号是如何在视觉皮层最早期的处理阶段被表征这一基本原则的道路上，我们前进的方向是正确的。通过将这个原理扩展到线性滤波器的独立特征子空间，就可以在视觉皮层中建立复杂细胞的模型。[10]

独立分量分析网络包含相同数量的输入和输出单元，以及把它们完全连接起来的一组权重。为解决盲源分离问题，麦克风的声音通过输入层输入，每个麦克风占有一个输入单元，而独立分量分析学习算法（类似感知器算法）不断迭代修改传递到输出层的权重，直到它们收敛。但是，不同于感知器这种监督学习算法，独立分量分析是一种无监督学习算法。它使用输出单元之间的独立性度量作为代价函数，但并不知道输出目标应该是什么。为了使输出尽可能独立，权重不断被修改，原始声源就被完美地分离开来，或者如果它们彼此并不是独立的，其相关性也会被尽可能地去除掉。无监督学习可以在许多不同类型的数据集中发现以前未知的统计结构。

将独立分量分析应用于大脑

我实验室里的其他人将安东尼·贝尔的信息最大化独立分量分析算法应用于大脑的不同类型的记录，由此引发了一系列的顿悟时刻。1924年，汉斯·伯格（Hans Berger）从头皮上记录了大脑的第一个电信号，被称为脑电图。神经科学家利用这些复杂的振荡信号来监听我们不断变化的大脑状态，这种状态随我们的警觉性和感觉运动相互作用而变化。头皮上一个电极处的电信号接收来自大脑皮层内许多不同来源的输入，也接收肌肉和眼球运动噪声的输入。每个头皮电极接收来自大脑中相同源组的混合信号，这些信号具有不同的振幅，这与鸡尾酒会上的问题如出一辙。

20世纪90年代，斯科特·马克格（Scott Makeig）是我在索尔克研究所的实验室里的研究员，他使用独立分量分析从脑电图记录中提取了大脑皮层中的几十个偶极源，以及它们各自的时间过程（见图6－4）。偶极子是大脑信息源最简单的模式之一。最简单的是由单一静电电荷产生的覆盖整个头皮的均一模式。第二简单的模式是由电流直线移动产生的偶极子模式，发生在皮层锥体神经元中。把偶极子想象成一个箭头，头皮上沿箭头的方向为正，沿箭尾的方向为负。该模式覆盖了整个头部，这就是为什么分离同时被激活的许多大脑信息源非常困难。从脑电图提取的两个信息源——IC2和IC3，大致相当于图6－4中的偶极源。独立分量分析还分离了噪声，例如眼球运动和电极噪声，这些噪声在后期可以以很高的精度被抹除（图6－4中的IC 1和IC4）。自那时起，数以千计已发表的论文都在使用独立分量分析方法分析脑电图记录，并在使用独立分量分析方法分析大量脑状态的研究领域取得了重要发现。

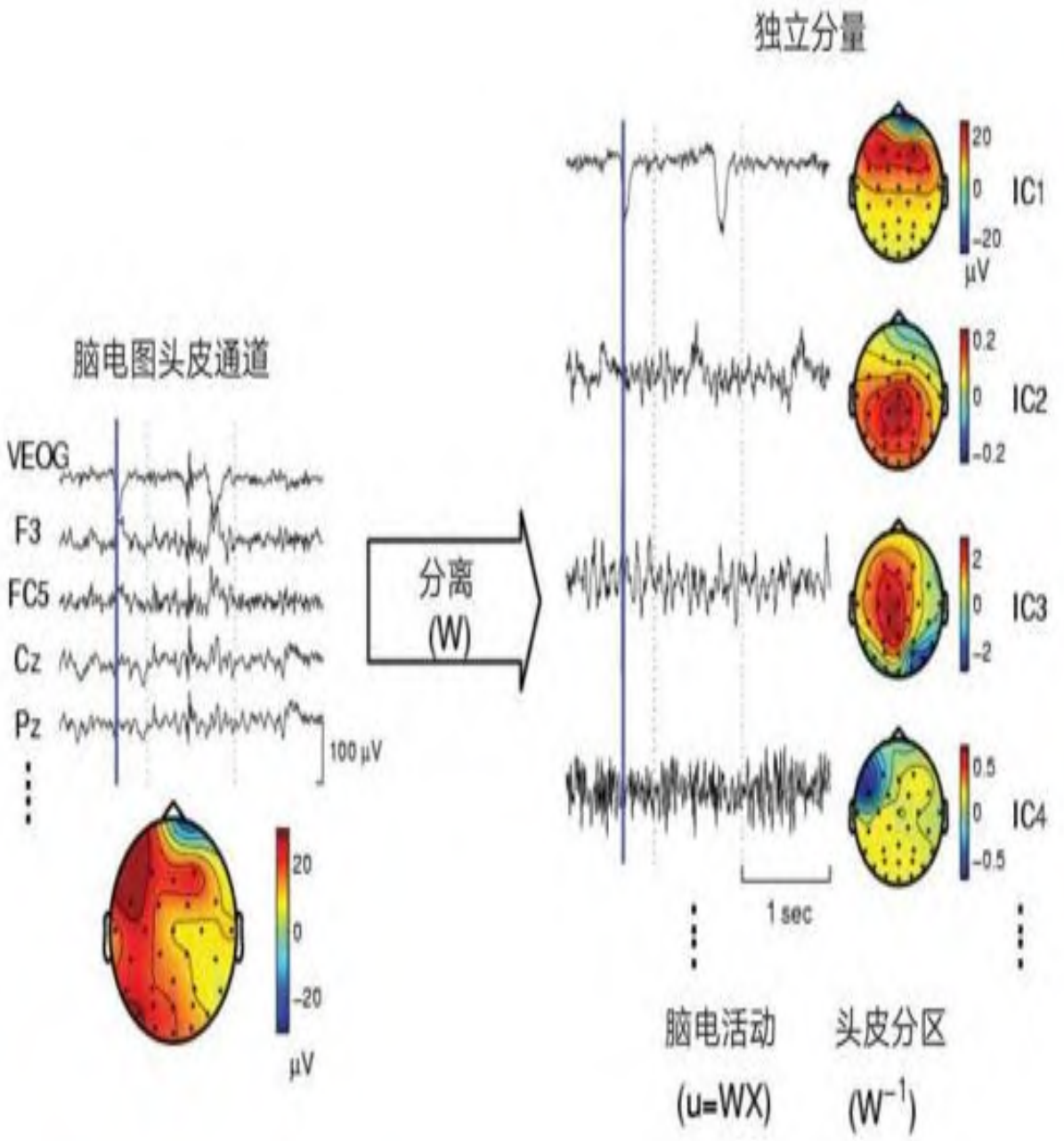
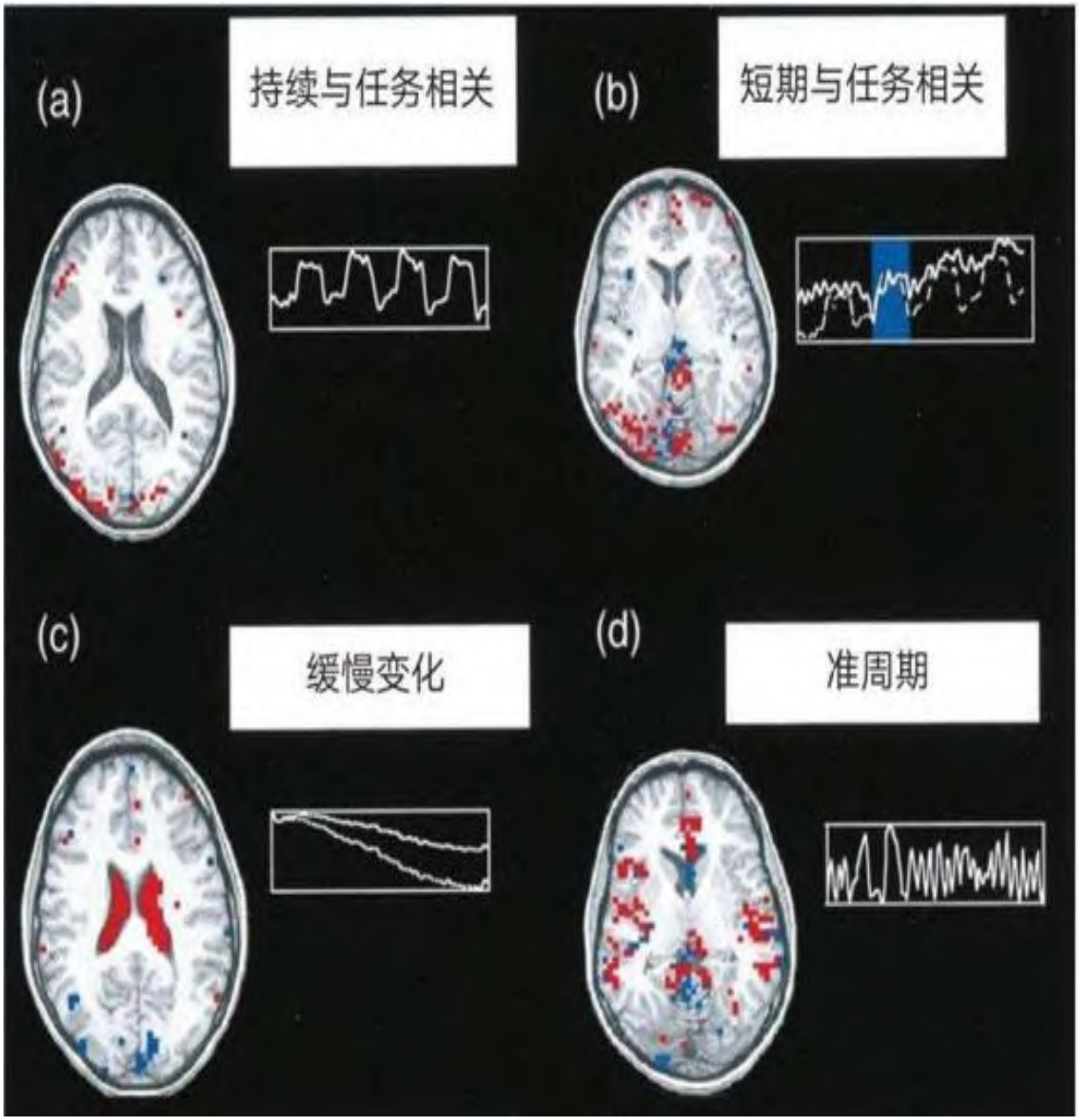


图6-4 独立分量分析被应用于头皮脑电图记录。在俯视角度的头皮绘图（鼻子朝上）中，黑点的位置是电极，某个时间点的电压以微伏（μV）为单位用彩色标注。左侧图中五个头皮通道显示波动中的脑电图信号受到了来自眨眼和肌肉信号的噪声污染。独立分量分析可以将大脑成分与噪声分开，如右图所示（其中“IC”代表“独立分量”）。IC1是基于缓慢时间过程的眨眼运动，此时头皮绘图在靠近眼睛的地方数值最高（红色）。IC4是肌肉运动噪声，可以看到高频率、高振幅的噪声和头皮图上的局部源。IC2和IC3是大脑信息源，分别表示为头皮上的偶极模式（与负值蓝色区域相对的正值红色区域），和来自脑电图所记录的头皮上更复杂的模式（如左边头皮绘图所示）。图片来源：Tzyy-Ping Jung。

马丁·麦基沃恩（Martin McKeown）当时是我实验室的一名博士后研究员，具有神经学背景。他研究出了如何翻转空间和时间，将独立分量分析应用于fMRI的记录中（见图6－5）。[11]基于fMRI的脑成像在大脑中数以万计的位置对与神经活动间接相关的血氧水平进行测量。在图6－5中，独立分量分析信息源是与其他信息源具有共同时间进程，但在空间上相互独立的大脑区域。空间域的稀疏性意味着在任何给定的时间，只有少数几个区域是高度活跃的。



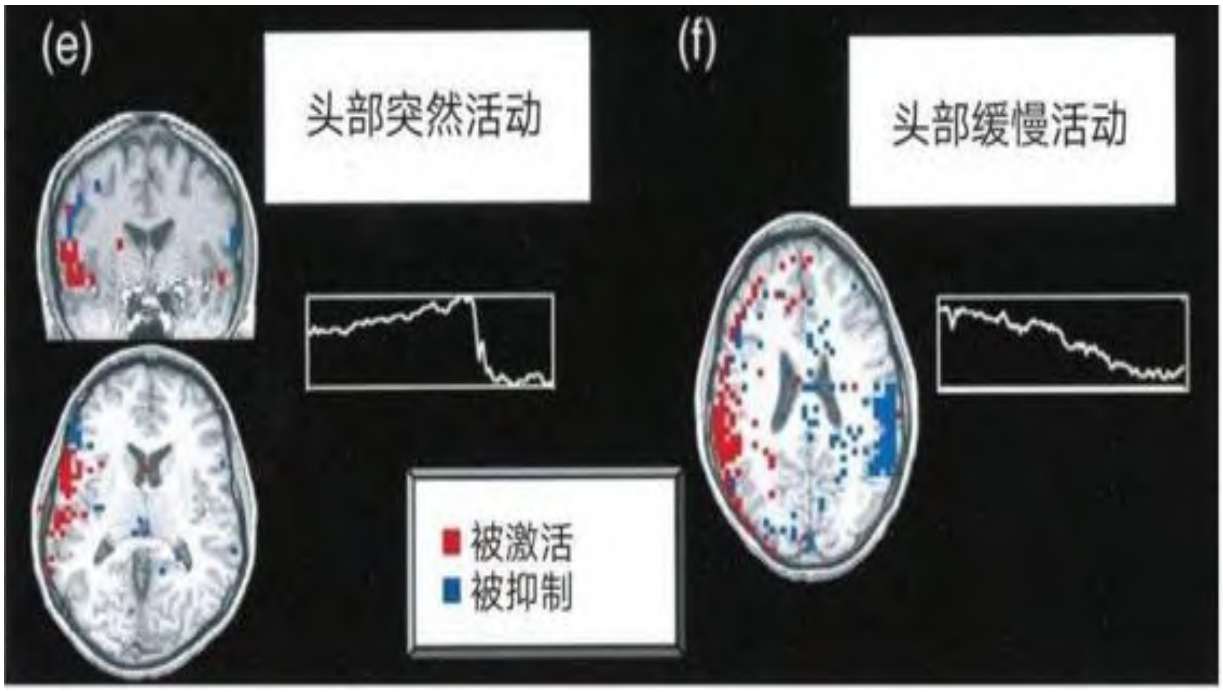


图6-5 独立分量分析被应用于fMRI数据。每个分量由一个大脑活动图和一个时间过程组成。这里显示了几种不同类型的分量。图中的任务代表了持续5秒钟的视觉刺激，由与任务相关的分量拾取。方框中信号的时间进程约为一分钟，任务重复了四次，如图（a）所示。其他分量则采集了诸如头部运动之类的噪声。图片来源：M.J. McKeown, T.-P. Jung, S. Makeig, G. D. Brown, S. S. Kindermann, T.-W. Lee, and T. J. Sejnowski, "Spatially Independent Activity Patterns in Functional MRI Data during the Stroop Color-Naming Task," *Proceedings of the National Academy of Sciences of the United States of America*95, no. 3 (1998): 806,figure 1。

因为独立分量分析是无监督的，它可以揭示能够协同工作的大脑区域网络，并能够扩充试图将区域中的活动与感官刺激或运动反应联系起来的监督技术。例如，独立分量分析已被用于揭示来自受试者的fMRI记录中的多个静息状态，这些受试者只是被要求在扫描仪中保持不动。[12]我们仍然不了解这些静息状态意味着什么，但是它们可以代表对大脑活动负责的一部分大脑区域的组合，例如我们在做白日梦，被一件烦心事困扰，或是正在计划晚餐吃什么的时候。

最大独立性原则与稀疏编码原则有关。尽管独立分量分析揭示了许多独立的分量，但只需要其中的一小部分就能重建自然图像中指定的图块。这个原理也适用于视觉皮层，视觉皮层的细胞数量比视网膜

的输入细胞多100倍。我们的每个视网膜都有100万个神经节细胞，在初级视觉皮层中有1亿个神经元——这是皮层视觉层级中的第一层。视网膜中视觉信号的紧凑编码，在皮层中被扩展为高度分布且高度稀疏的新编码。将信息扩展到更高维度空间的方法也被应用到其他的编码方案中，其中包括在听觉皮层和嗅觉皮层中发现的编码。一类被称为“压缩感知算法”（compressed sensing algorithms）的新算法将稀疏性原理进行了泛化处理，用来提高存储和分析复杂数据集的效率。[13]

什么在操控我们的言行

独立分量分析的故事说明了技术在科学和工程领域取得新发现中的重要性。我们通常将技术看作测量设备，比如显微镜和放大器。但算法也是技术，它们能够利用旧仪器获得的数据得出新的发现。脑电图技术已经存在了将近100年，但是如果没有独立分量分析，就无法确定潜在的大脑信息源，大脑本身就是一个环环相扣的算法体系，如果大脑的某些部分发现了实施独立分量分析的方法，我并不会对此感到惊讶。[14]

20世纪90年代，在为神经元网络开发新的学习算法的过程中，也诞生了许多其他的发现，其中很多（如独立分量分析）现在都成了机器学习中的数学工具。这些算法被嵌入许多常用的设备中，不过没有一个标明了“内含神经网络”。拿耳机或手机来举个例子。Te-Won Lee和Tzyy-Ping Jung曾经是我实验室的博士后研究员，他们后来创立了一家名为“SoftMax”的公司，在带有两个麦克风的蓝牙耳机中应用ICA来消除背景噪声。这样的设计能够让戴耳机的人在嘈杂的餐厅或体育赛事中听到别人说话。2007年，SoftMax被高通收购（高通为许多品牌的手机设计芯片）。而如今，类似独立分量分析的解决方案也已嵌

入10亿部手机中。如果你能从运行独立分量分析的每部手机上分到一分钱，那你现在就是个千万富翁了。

安东尼·贝尔多年来一直对一个更加棘手的问题感兴趣。作为人类，我们的体内有许多网络，信息从一个网络层到另一个网络层，从分子到突触，到神经元，到神经群，直到形成决策，所有这些都可以用物理学和生物化学的法则来解释（见图4－4）。但我们都有这样的感受，是我们自己，而非物理或生物化学反应，在操控我们的一言一行。我们的大脑神经群中出现的内部活动如何引导我们做出决定，例如阅读这本书或者打网球，一直是一个谜。这些远远低于我们意识水平的决定，在某种程度上是由神经元通过基于分子机制的经验形成的突触相互作用形成的。但从我们人类的视角来看，是我们的决定导致了所有这些事件在我们的脑中发生：内省行为告诉我们，因果关系中的因和果，似乎与物理和生物化学领域中的因和果相反。如何调和这两个观点，是一个深刻的科学问题。[15]

---

[1] J. Herault and C. Jutten, “Space or Time Adaptive Signal Processing by Neural Network Models,” in J. S. Denker, ed., *Neural Networks for Computing, AIPConference Proceedings*151, no. 1 (1986): 206－211.

[2] A. J. Bell and T. J. Sejnowski, “An Information-Maximization Approach to Blind Separation and Blind Deconvolution,” *Neural Computation*7, no. 6 (1995): 1129－1159.

[3] 早些时候，IBM的拉夫·林斯克（Ralph Linsker）引入了一种名为“infomax”的算法，以解释在发育过程中视觉系统是如何彼此连接起来的。R. Linsker,“Self-Organization in a Perceptual Network,” *Computer*21, no. 3 (1988): 105－117.

[4] A. J. Bell and T. J. Sejnowski, “An Information-Maximization Approach to Blind Separation and Blind Deconvolution.”

[5] 对ICA的发展做出了重要贡献的人还包括：Pierre Comon, Jean-François Cardoso, Apo Hyvarinen, Erkki Oja, Andrzej Cichocki, Shunichi Amari, Te-Won Lee, Michael Lewicki, and many others。

[6] A. J. Bell and T. J. Sejnowski, “The ‘Independent Components’ of Natural Scenes Are Edge Filters,” *Vision Research*37, no. 23 (1997): 3327－

3338.

[7] 布鲁诺·奥尔斯豪森（Bruno Olshausen）和大卫·菲尔德（David Field）也用另一种基于稀疏性的学习算法得出了同样的结论。B. A. Olshausen and D. J.Field, “Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images,” *Nature*38, no. 6583 (1996): 607–609.

[8] Horace Barlow, “Possible Principles Underlying the Transformation of Sensory Messages,” in Walter A. Rosenblith, ed., *Sensory Communication*(Cambridge, MA:MIT Press, 1961), 217–234.

[9] A. J. Bell and T. J. Sejnowski, “Learning the Higher-Order Structure of a Natural Sound,” *Network: Computation in Neural Systems*7, no. 2 (1996): 261–267.

[10] A. Hyvarinen and P. Hoyer, “Emergence of Phase- and Shift Invariant Features by Decomposition of Natural Images into Independent Feature Subspaces.” *NeuralComputation*12, no. 7 (2000): 1705–1720.

[11] M. J. McKeown, T.-P. Jung, S. Makeig, G. D. Brown, S. S. Kindermann, T.-W. Lee,and T. J. Sejnowski, “Spatially Independent Activity Patterns in Functional MRI Data during the Stroop Color-Naming Task,” *Proceedings of the National Academyof Sciences of the United States of America*95, no. 3 (1998): 803–810.

[12] D. Mantini, M. G. Perrucci, C. Del Gratta, G. L. Romani, and M. Corbetta,“Electrophysiological Signatures of Resting State Networks in the Human Brain,” *Proceedings of the National Academy of Sciences of the United States of America*104, no. 32 (2007):13170–13175.

[13] D. L. Donoho, “Compressed Sensing,”*IEEE Transactions on Information Theory* 52, no. 4 (2006): 1289–1306; Sanjoy Dasgupta, Charles F. Stevens, and Saket Navlakha, “A Neural Algorithm for a Fundamental Computing Problem,” *Science* 358 (2017): 793–796. doi:10.1126/science.aam9868.

[14] 小脑可能在苔藓纤维输入汇聚到颗粒细胞树突的层面上进行了独立分量分析。参阅D. M. Eagleman, O. J.-M. D. Coenen, V. Mitsner, T. M. Bartol, A.J. Bell, and T. J. Sejnowski, “Cerebellar Glomeruli: Does Limited Extracellular Calcium Implement a Sparse Encoding Strategy?” in *Proceedings of the 8*th *JointSymposium on Neural Computation*(La Jolla, CA: Salk Institute, 2001).

[15] 托尼·贝尔（Tony Bell）正在使用独立分量分析和近红外光谱来研究水的结构。他试图证明，在当今仪器看不见的尺度上，水形成了可通过光线进行沟通的连贯结构，并形

成了生物分子生命的基底。这个想法是，当“神经机制”（nerual schemes）放松到足够的程度时，体内广泛分布的原子网络会出现连贯的信息，决定就是这样产生的。

07
霍普菲尔德网络和玻尔兹曼机

在20世纪80年代，尚在罗切斯特大学的计算机科学家杰罗姆·费尔德曼（Jerome Feldman）采用了联结主义（connectionist）网络的方式来研究人工智能。杰罗姆几乎没说错过什么话，他曾经指出，人工智能中使用的算法在运行了数十亿个步骤之后，却常常得不到一个正确的结论，而大脑只需要经历大约100个步骤，通常就会得出一个正确的结论。[1]这个“100步法则”当时在人工智能研究人员中并不像现在那样受欢迎，但少数人，其中最著名的是卡内基－梅隆大学的艾伦·纽维尔，的确已经开始把它作为一种约束。

有一次我被困在了纽约州的罗切斯特机场，杰罗姆好心留宿了我一晚。那天，我参观完位于斯克内克塔迪的通用电气研究实验室，准备返回巴尔的摩，在回去的飞机上，听到飞行员通知我们罗切斯特的天气状况。我这才发现坐错了航班。落地后，我预订了飞往巴尔的摩的最早航班，但也要等到第二天了。我碰巧遇见了杰罗姆，他刚参加完在华盛顿特区举办的一次委员会会议，正准备回家。他很大方地邀请我去他家休息一个晚上。杰罗姆后来去了加州大学伯克利分校，但每次我被困机场时，都会禁不住想起他。

杰罗姆区分了“邋遢”和“整洁”的联结主义模型。邋遢模型跟杰弗里·辛顿和我一起研究过的模型十分相似，在网络的许多单元中分布了对对象和概念的表征。而杰罗姆所笃信的整洁模型，提供了对对象和概念严谨的计算表征，每个单元上只有一个标签。在更广泛的背景下，邋遢科学的模型使用近似方法来获得定性的答案，而整洁科

学的模型致力于找到问题的确切解决方案。事实上，人工智能要取得进展，需要同时应用这两种方法。[2]我不反对立足于“邋遢”的方法，但我会尽一切努力去给出一个更“整洁”的解释，最终努力产生了回报：杰弗里和我后来都因此获得了巨大的成就。

约翰·霍普菲尔德的伟大之处

想要获得物理学博士学位，你必须解决一个问题。好的物理学家应该有能力解决任何问题，但伟大的物理学家知道要解决什么样的问题。约翰·霍普菲尔德就是一位伟大的物理学家。在凝聚态物理学领域获得了杰出的成就之后，他的兴趣转向了生物学，特别是关于“分子校正”的问题。DNA（脱氧核糖核酸）在细胞分裂过程中被复制时，错误不可避免，必须纠正这些错误才能保证子细胞遗传信息的准确度。霍普菲尔德想出了一个聪明的办法，并解释了采用这个方法纠错的过程是怎么实现的，尽管他提出的这种方法需要消耗能量，但随后的实验表明，他是正确的。在生物学研究领域，做出任何正确的解释都是一项了不起的成就。

霍普菲尔德是我在普林斯顿大学读博士期间的导师，那时他刚对神经科学产生兴趣（见图7－1）。随着这种兴趣变得日益浓厚，他开始兴致勃勃地给我讲述，他从波士顿神经科学研究计划（Neuroscience Research Program，NRP）会议上那些神经科学研究专家的演讲里所学到的东西。我发现在NPR会议上发布的一些小型研讨会会议文章非常有价值，让我了解了当时人们正在研究什么问题，以及产生了哪些想法。我现在还保留着一份由传奇神经学家西奥多·霍姆斯·布洛克（Theodore Homes Bullock）整理的关于神经编码的研讨会会议文献。西奥多后来成了我在加州大学圣迭戈分校的同事。西奥多和艾德里安·霍里奇（Adrian Horridge）合著的关于无脊椎动物神经系统的书堪称经典。[3]我和西奥多在模拟珊瑚礁集体行为的项

目中有过合作，并且我很荣幸地成为他在2008年发表的最后一篇学术论文的共同作者。[4]

与前层单元建立了反馈连接，并与同层单元间存在循环连接的神经网络，要比仅具有前馈连接的神经网络复杂得多。包含具有正（兴奋）权重和负（抑制）权重的任意连接单元的通用网络对学界提出了一个数学难题。虽然芝加哥大学的杰克·考恩（Jack Cowan）和波士顿大学的斯蒂芬·格罗斯伯格（Stephen Grossberg）在20世纪70年代后期，通过证明这样的网络能够重现视觉错觉[5]和幻觉[6]，在该领域取得了一些进展，但是工程师发现这种网络并不能解决复杂的计算问题。

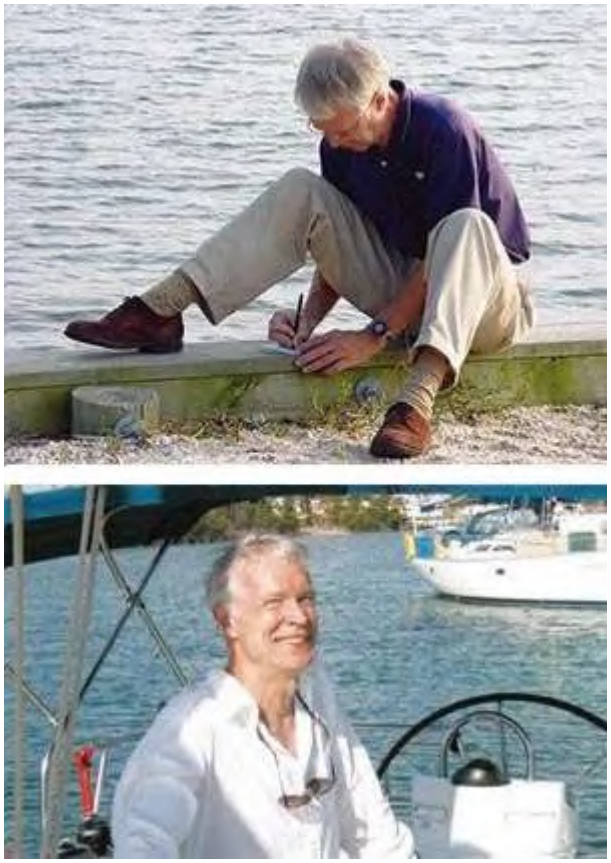


图7-1 约翰·霍普菲尔德在马萨诸塞州伍兹霍尔的岸边正埋头解决某个问题（照片拍摄于1986年前后）。霍普菲尔德在20世纪80年代设计了一个以他自己名字命名的神经网络，开启了深度学习研究的大门，这一设计对神经网络研究产生了开创性的影响。图片来源：约翰·霍普菲尔德。

内容可寻址存储器

1983年的夏天，辛顿、霍普菲尔德和我参加了杰罗姆·费尔德曼在罗切斯特大学组织的一场讨论会。霍普菲尔德告诉我们，他解决了一种强交互网络的收敛问题。他证明了一种独特的非线性网络模型，即现在的“霍普菲尔德网络”，能确保收敛到一种叫“吸引子”（attractor）的稳定状态（见图7－2，方框7.1）[7]（高度非线性网络很容易发生振荡，或表现出更多的混沌行为）。此外，可以通过选择网络中的权重，把吸引子状态做成信息内容。如此一来，霍普菲尔德网络就可以被用来实现一种“内容可寻址存储”（content-addressable memory）。存储的信息可以只用信息内容的一部分作为提取输入，让神经网络完成信息填充。这让人联想起了人的记忆行为。如果我们看到某个熟人的脸，就能想起那个人的名字和先前与之交谈过的内容。

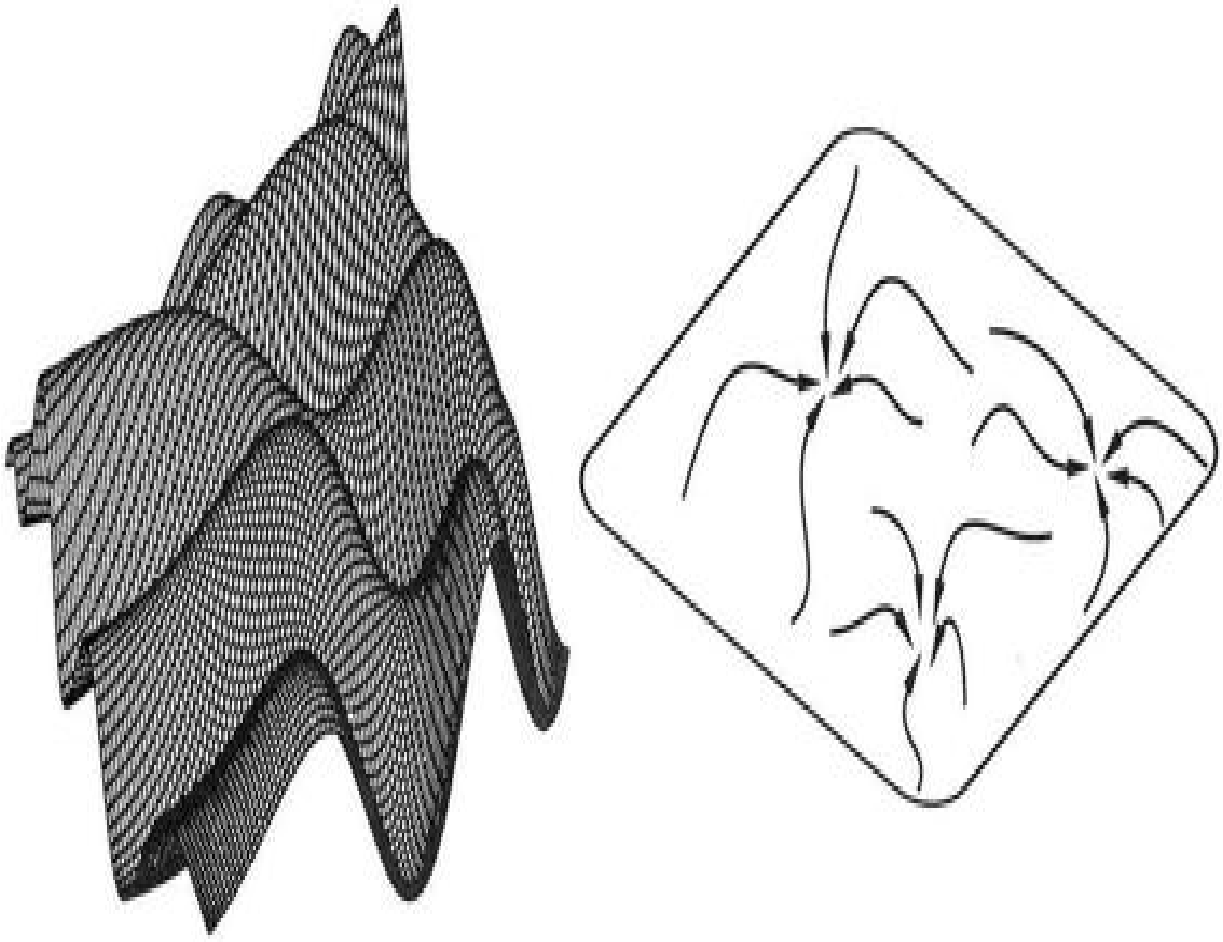


图7-2 霍普菲尔德网络的能量场景图。（左图）网络的状态可以被看作是能量表面上的一个点。（右图）每次更新可以将状态移动到其中一个局部能量最小值，叫作“吸引子状态”。图片来源：A. Krogh, J. Hertz, and R.G. Palmer, *Introduction to the Theory of Neural Computation*Redwood City CA: Addison-Wesley, 1991, left:figure 2.6; right: figure 2.2。

霍普菲尔德网络的重大突破性在于，它能在数学上保证收敛。研究人员曾认为要分析高度非线性网络的一般情况是不可能的。当网络中的所有单元被同时更新的时候，所引发的动态情况非常复杂，没有办法保证收敛。[8]但是霍普菲尔德网络展示出，当网络中单元的更新是顺序进行的时候，一种单元组彼此双向连接并且权重相同的特殊对称网络，是可解，并且最终是能够收敛的。

越来越多的证据表明，海马体（保存关于特殊事件和独特对象长期记忆的大脑关键区域）中的神经网络存在像霍普菲尔德网络中的那

种吸引子状态。[9]虽然霍普菲尔德模型非常抽象，但它的本质行为和在海马体上观察到的行为十分相似。20世纪80年代，许多物理学家都曾利用霍普菲尔德网络实现了由物理学到神经科学的跨越。使用理论物理的复杂工具来分析神经网络和学习算法的过程中，诞生了很多惊人的发现。物理、计算和学习在神经科学理论领域深刻地联系在一起，这一结合成功地实现了对大脑功能的诠释。

约翰·霍普菲尔德和当时在贝尔实验室的大卫·汤克（David Tank），随后又展示了一种霍普菲尔德网络的变体，其中的单元可以拥有0到1之间的连续值，[10]能够很好地解决优化问题，例如“旅行商问题”（traveling salesman problem），该问题旨在寻找能够访问多个城市的最短路径，每个城市仅访问一次。[11]这是一个计算机科学领域众所周知的难题。网络的能量函数包含了路径的长度，限制是每个城市仅能访问一次。经过初始状态，霍普菲尔德和汤克的网络能最终收敛到一个最小能量状态，表明找到了一条较好的路径，虽然不一定是最佳路径。

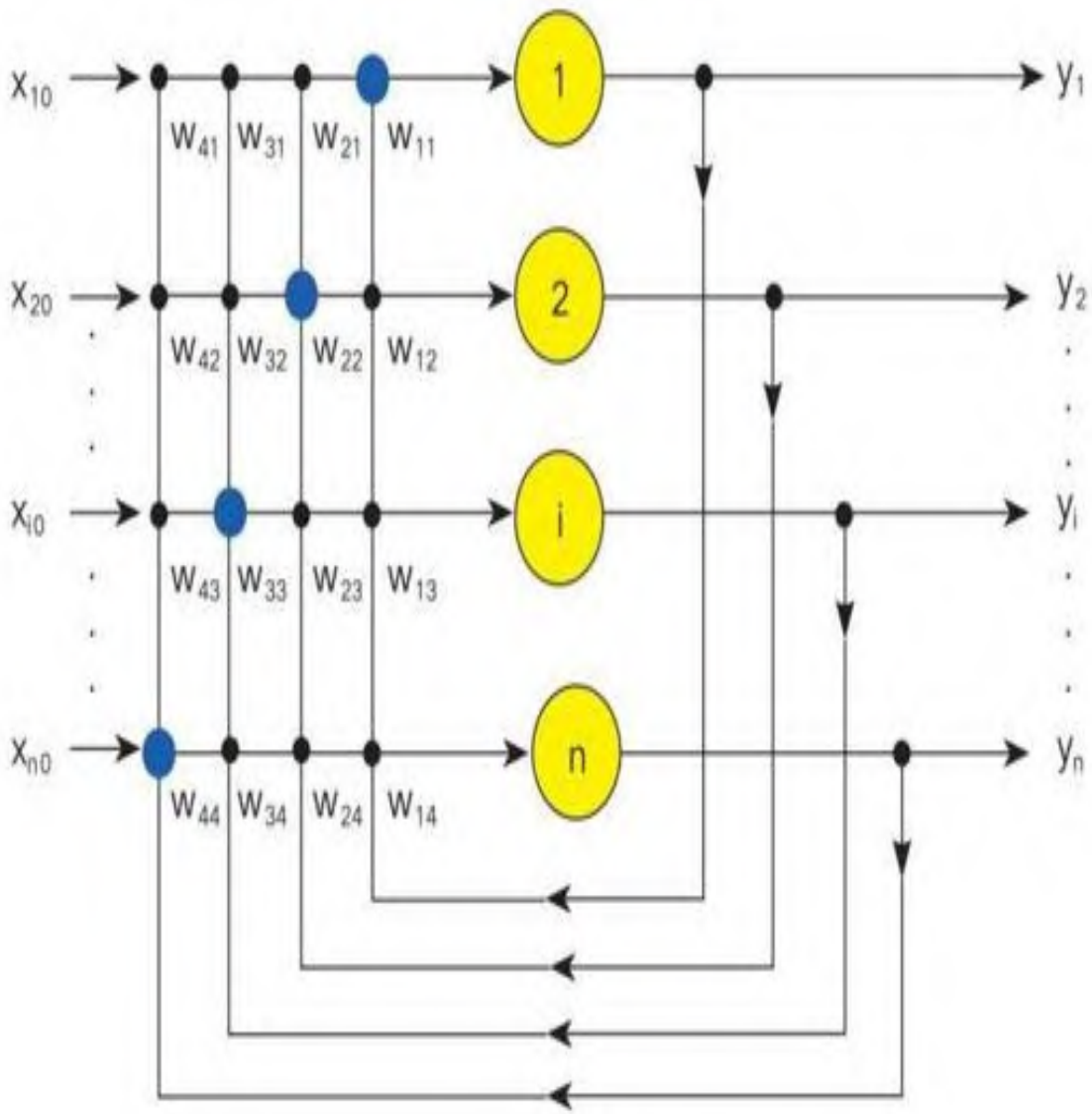
局部最小值与全局最小值

达纳·巴拉德（Dana Ballard）在1982年与克里斯托弗·布朗（Christopher Brown）一起撰写了一本关于计算机视觉的经典著作，[12]他也参加了1983年的研讨会。辛顿和我当时正在与巴拉德一起为《自然》杂志撰写一篇关于图像分析新方法的综述。[13]该文章的核心想法是，用网络模型中的节点表示图像中的特征，网络中的连接实现了特征间的约束条件；互相兼容的节点具有积极的相互作用，而不兼容的节点间具有消极的相互作用。在视觉中，必须找到满足所有约束条件的所有特征的和谐表达。

霍普菲尔德网络能够解决这种约束满足问题吗？能量函数可以用来衡量网络满足所有约束的程度（见方框7.1）。视觉问题需要一个最佳解决方案，一个全局能量最小值，但是霍普菲尔德网络的工作原理决定了它仅能提供局部最小值。《科学》杂志上的一篇文章让我深受启发，其作者斯科特·柯克帕特里克（Scott Kirkpatrick）当时在位于纽约约克敦海茨（Yorktown Heights）的IBM托马斯·沃森研究中心（Thomas J. Watson Research Center）工作。[14]柯克帕特里克使用了一种叫作“模拟退火”（simulated annealing）的算法来解决局部最小值问题。假设你想把一堆电子元件装到两块电路板上，怎么放置这些元件可以使连接元件的接线数量最少呢？

较差的解决方案就是，先随机安放元件，然后每次移动一个元件，直到找到用线最少的排布方案。因为当移动任何一个元件都无法减少用线时，很容易陷入局部最小值的陷阱。逃出这个陷阱的一种方式就是，允许随机跳到一种用线更长的放置方式。这种跳跃的概率，开始时会很大，随后慢慢减小，直至变成零。如果这种概率减小的速度足够慢，元件的最终排布方式会达到接线的全局最小值。在冶金业，这种做法叫作退火。将金属加热并慢慢冷却，在这一过程中会形成缺陷最小的较大的晶体结构，这些缺陷会使金属过脆，或出现裂纹。

7.1
霍普菲尔德网络



在霍普菲尔德网络中，每个单元都能向网络中所有其他的单元传递输出。输入标记为x_i，输出标记为y_j。连接的强度，或者说单元之间的权重是对称的：$w_{ij}=w_{ji}$。在每个步骤中，其中一个单元被更新为输入值的总和，并和一个阈值进行比较：如果输入之和大于阈值，输出就是1，否则输出就是0。霍普菲尔德展示了这种网络存在一个能量函数，它不会随着网络中单个单元的逐次更新而增长：

$$E=\sum w_{ij}\, x_i\, x_j$$

最终，霍普菲尔德网络会达到“吸引子状态”，即所有单元值不再变化，能量方程达到局部最小值。这种状态等同于存储好的记忆，可以通过部分存储好的状态来初始化网络，实现对完整记忆的恢复。这就是霍普菲尔德网络实现内容取址记忆的方法。所存储向量的权重可通过赫布突触可塑性（Hebbian synaptic plasticity）得到：

$$\Delta w_{ij} = \alpha x_i x_j$$

Δw_{ij}表示权重的变化，α表示学习速率，x_i表示存储的向量。

该图由戴尔·希斯（Dale Heath）绘制。

玻尔兹曼机

在霍普菲尔德网络中，模拟退火相当于对更新进行“加热”，这样能量就能在能量表面自由上下。因为单元在高温下可以随机反转，如果温度逐渐降低，霍普菲尔德网络很有可能在温度为零时收敛到能量最低状态。在实际操作中，模拟会在常温条件下让网络达到平衡状态，这样网络就可以访问周围的若干状态，并且在一个更广阔的范围内搜寻可能的答案。

例如，图7－3中剪影的图案是存在争议的。这取决于你关注的是哪个部分，你可能会看到一个花瓶或两张侧脸，但不会同时看到这两种图案。问题的关键在于，你将哪部分看成图形，哪部分看成背景。我们设计了一个玻尔兹曼机网络来模仿这个基于图形—背景的决定（figure-ground decision），[15]用一些单元代表被关注的图形，其他单元代表图形的边缘。我们已经知道，视觉皮层中存在由边缘激活的简单细胞，但是图形可能位于边缘的任意一侧。我们的玻尔兹曼机网络通过两组边缘单元实现了图形选择，每组单元只支持一侧的图形。这类神经元随后在视觉皮层中被发现，被称为“边界归属细胞”（border-ownerships cells）。[16]

玻尔兹曼网络中的约束可以通过手动设置权重来实现（图7－4）。在图形单元之间存在兴奋连接，而在边缘单元之间存在着抑制连接。边缘单元与它们所指向的图形单元之间存在兴奋性连接，能够支持图形单元，并与相背离的图形单元间形成抑制性连接。[17]注意力是通过对一些图形单位产生偏倚（bias）而实现的。当玻尔兹曼网络对单元使用霍普菲尔德更新规则时，它会落入局部能量最小值，这些局部能量最小值在本地图块中一致，但在全局上并不一致。当更新中加入了噪声时，玻尔兹曼网络可以跳出局部最小值，并且通过缓慢退火降低噪声温度，这样一来网络的能量函数就可能得到全局一致的能量最小值（见图7－4）。由于更新是异步和独立的，因此网络可以由具有数百万个单元并行工作的计算机来实现，并且可以比一次按顺序执行一个操作的数字计算机更快地收敛到解决方案中。

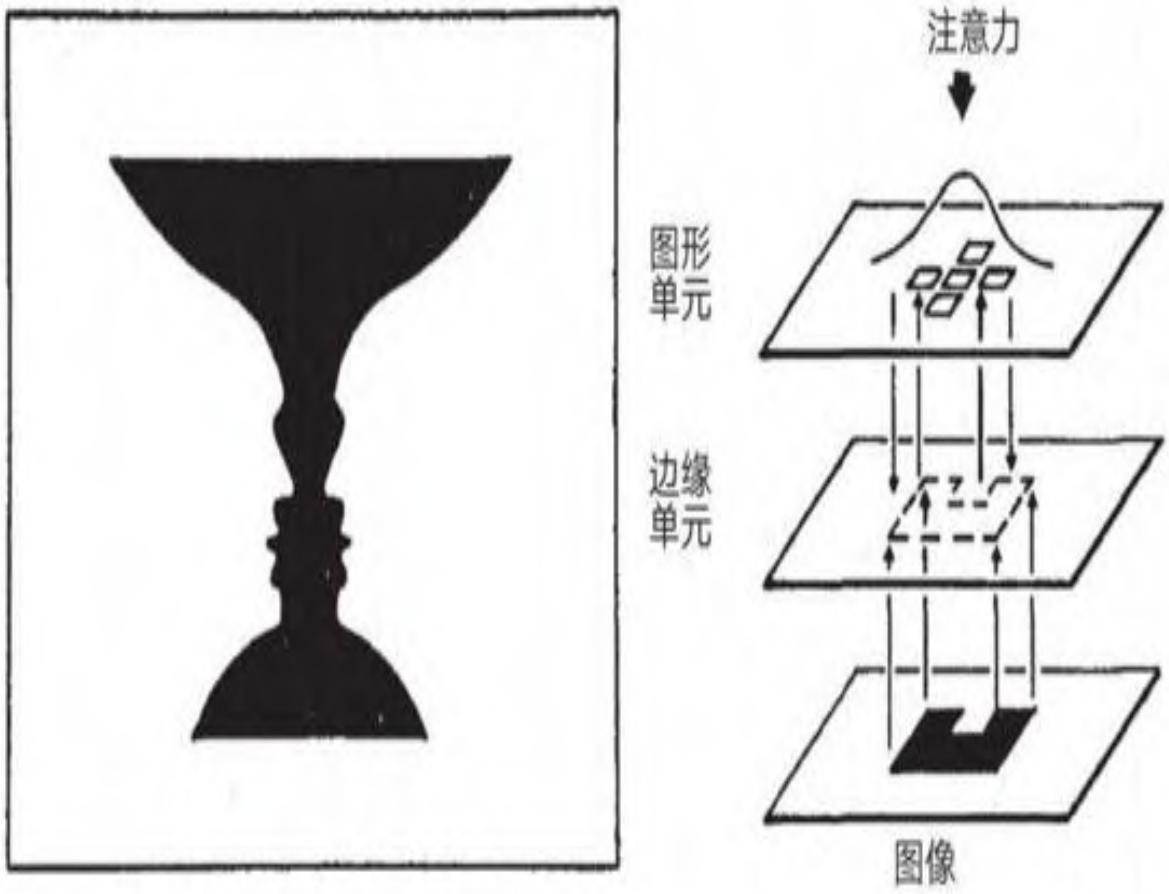
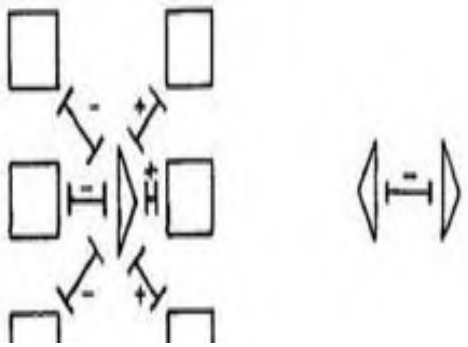
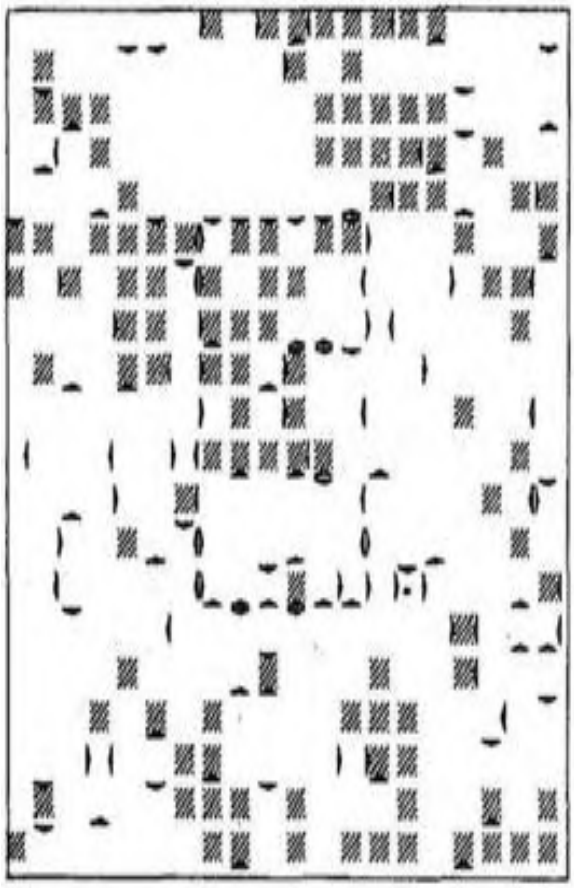
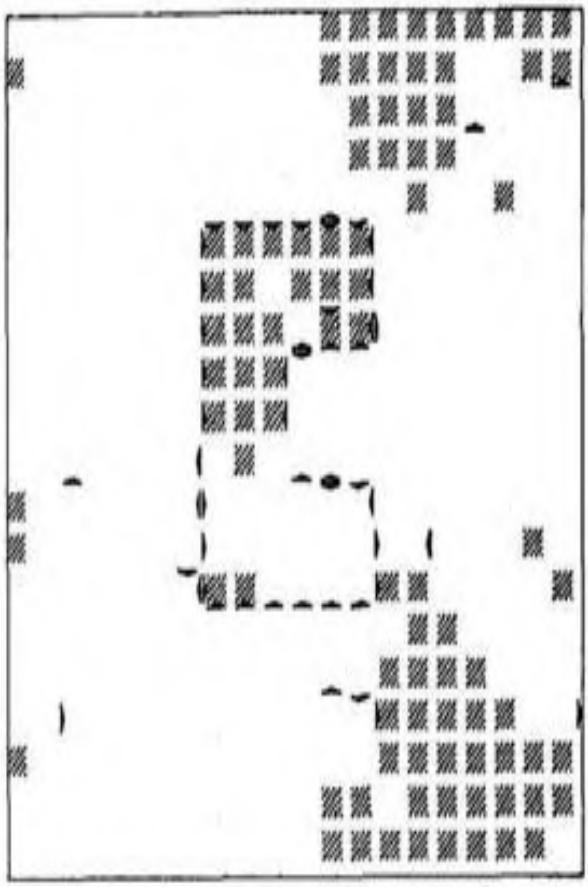


图7-3 模棱两可的图形—背景问题。（左图）如果关注黑色图形，你看到的是一个有白色背景的花瓶。如果关注的是白色背景，那你会看到两张互相注视的脸。你可以来回变换视角，但是做不到同时看到两种图案。（右图）图形—背景网络模型。两种单元分别代表了物体的边缘（线段），和某个像素是否属于图形或者背景（方块）。图像输入方向为从下到上，注意力输入方向是从上到下。其中，当某个区域应该被当作图形而进行填充时，注意力就是针对该区域所产生的偏倚。图片来源：P. K. Kienker, T. J. Sejnowski, G. E. Hinton, and L. E. Schumacher, “Separating Figure from Ground with a Parallel Network,” Perception 15 (1986): 197–216, left: figure1; right: figure 2。

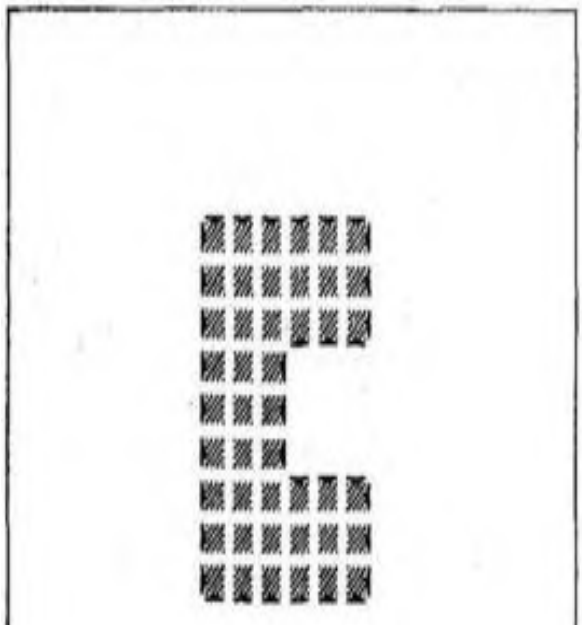


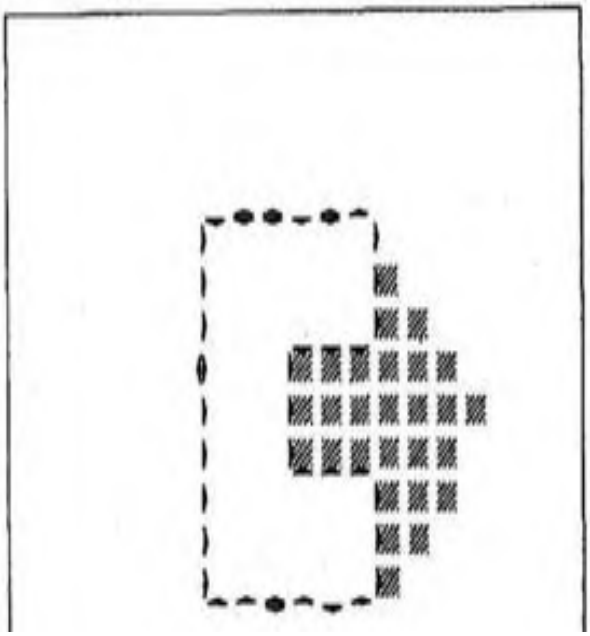


(a)



(b)





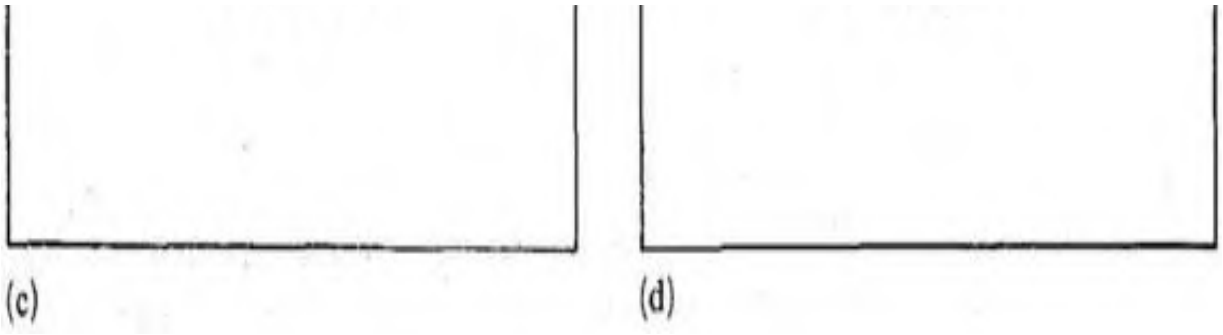


图7-4 正在分离图形与背景的玻尔兹曼机。（上图）网络中的方形单元为标识图形的图形单元，三角形为标识轮廓的边界单元，连接单元的正负属性也被一同标出。边缘单元可以指向或者背离图形。（下图）（a）关注于图形C内部区域的网络示意图。温度[18]刚开始很高，单元在开启和关闭状态之间不停变换。（b）随着温度下降，在指向内部的边缘单元的支持下，C内部的图形单元开始聚拢；同时，注意区以外或没有边缘输入的单位开始消失。（c）当温度为零时，注意区的图案被填满。（d）如果注意区移向外围，重复该过程，则外部区域会被填充。图片来源：P. K. Kienker, T. J. Sejnowski, G. E. Hinton, and L. E. Schumacher, “Separating Figure from Ground with a Parallel Network,” *Perception*15 (1986): 197–216, below: figure 6; above: figure 3。

那个时候，我已经在哈佛医学院的斯蒂芬·库夫勒那里完成了博士后研究，在约翰·霍普金斯大学（位于巴尔的摩市）生物物理系开始了我的第一份工作；杰弗里·辛顿在卡内基–梅隆大学（位于匹兹堡市）计算机科学系担任教职，在那里他很幸运地得到了艾伦·纽维尔的支持，艾伦对人工智能的新方向持十分开放的态度。匹兹堡和巴尔的摩离得很近，所以杰弗里和我可以在周末见面。我们将这一霍普菲尔德网络的新版本称为“玻尔兹曼机”，以致敬19世纪的物理学家路德维希·玻尔兹曼（Ludwig Boltzmann），他是统计力学的创始人。我们的波动性神经网络模型就是借助统计力学的工具进行分析的，之后我们很快就发现，该模型也是一个强大的学习机器。

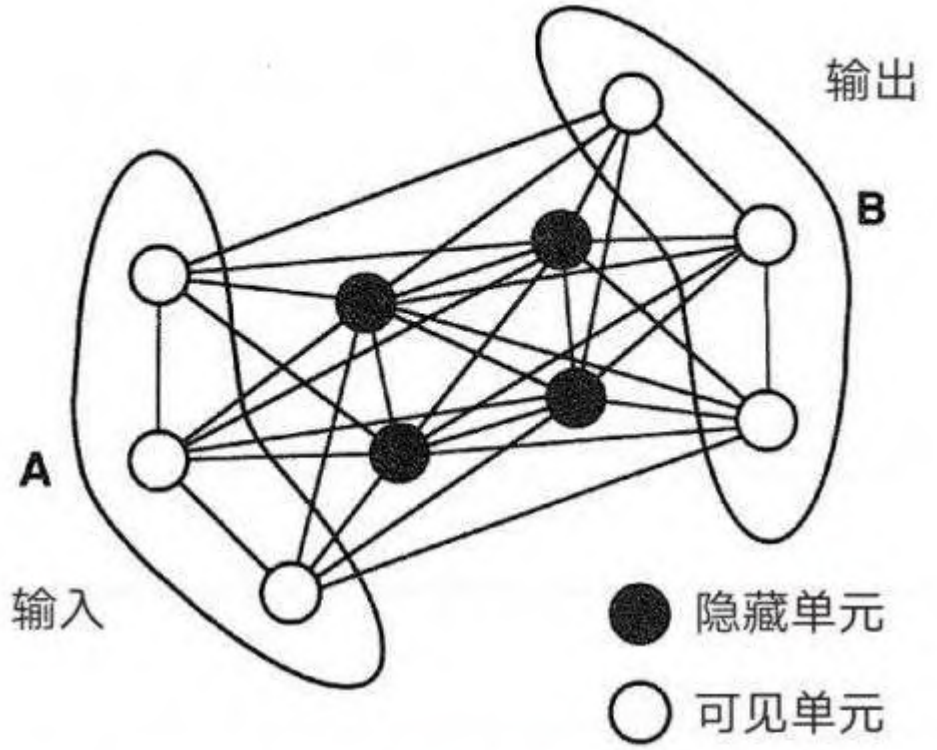
在恒定的“温度”下，玻尔兹曼机会达到平衡。在平衡状态下会发生一些神奇的事情：多层神经网络学习的大门将被打开。而此前，每个人都认为它的存在有弊无利。有一天，杰弗里给我打电话，说他刚刚从玻尔兹曼机衍生出一个简单的学习算法。该算法的目标是执行从输入单元到输出单元的映射。但与感知器不同的是，玻尔兹曼机在

输入和输出层之间也存在一些单元，我们称之为“隐藏单元”（见方框7.2）。通过对其呈现输入—输出对和应用学习算法，玻尔兹曼网络学会了实现目标映射。但是，其目标不仅仅是要记住输入—输出对，也要将没有被用于训练网络的新输入正确地进行分类。此外，玻尔兹曼机也在通过其持续不断的波动状态来学习概率分布——对于一个给定的输入模式，每个输出状态的访问频率如何——这就使其具备了生成性：学习之后，它可以通过钳制每个输出类别来生成新的输入样本。

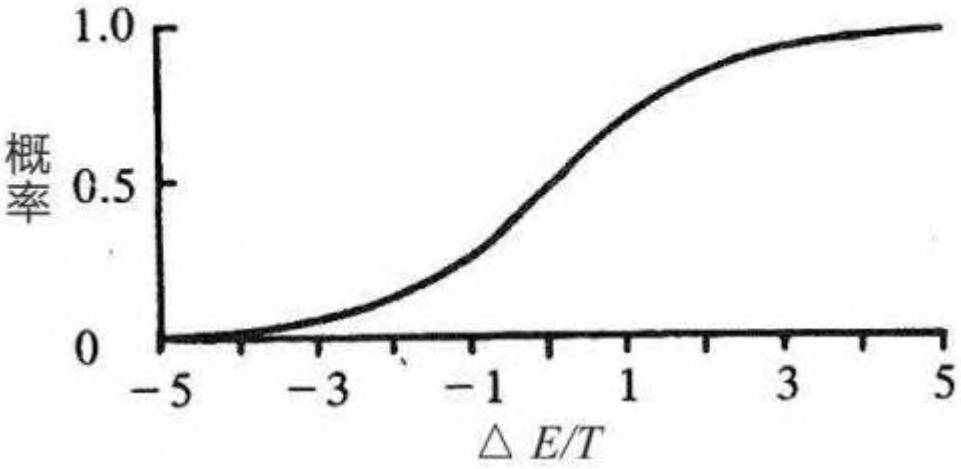
赫布理论

令人惊讶的是，玻尔兹曼机学习算法在神经科学领域有着悠久的历史，可以追溯到心理学家唐纳德·赫布（Donald O. Hebb）博士，他在《行为的组织》一书中假定，若两个神经元同时被激发，它们之间的突触连接应该会增强：

7.2
玻尔兹曼机



二进制玻尔兹曼单元



和在霍普菲尔德网络中一样，玻尔兹曼机中所有的连接都是对称的，并且其中的二进制单元i的状态值s_i，会按照上图中的Sigmoid函数（sigmoid function）提供的一个概率，逐个被更新成1或0。Sigmoid函数的输入值ΔE通过温度T来进行比例缩放。输入层和输出层

都是“可见的”，也就是说它们会与外界进行交互。“隐藏单元”代表了能够影响可见层单元的带有内部自由度的特征。玻尔兹曼机学习算法有两个阶段：在“清醒”阶段，输入和输出会被钳制（clamped）[19]，当网络达到平衡状态后，每对单元状态值之间的相关系数会被计算出来；在“睡眠”阶段，输入和输出钳制被解除，每对单元状态值之间的相关性会被重新计算，其中的权重会被逐步更新：

$$\Delta w_{ij} = \varepsilon \left(\langle s_i \; s_j \rangle^{wake} - \langle s_i \; s_j \rangle^{sleep} \right)$$

> 让我们假设反射活动（或“追踪痕迹”）的持续或重复，趋向于引起持续的细胞变化来增加其稳定性。当细胞A的轴突接近细胞B，并且反复或持续地参与激发细胞B时，其中一个或者两个细胞中都会发生某种生长过程或代谢变化，A（作为激发B的细胞之一）的效率会增加。[20]

这可能是整个神经科学领域中最有名的预言。后来在海马体中发现了赫布突触可塑性（Hebbian Synaptic Plasticity），海马体是大脑中控制长期记忆的重要组成部分。当一个海马锥体细胞接收强输入信号，同时该神经元也发生了放电反应时，突触的连接强度就会增加。随后的实验表明，这种强化是基于突触的发射器释放与受体神经元中电压升高的结合。此外，这种结合的发生被一种特殊的受体，即NMDA（N－甲基－D－天冬氨酸）谷氨酸受体所识别。它触发了长时程增强（long-term potentiation，简称LTP），其起效迅速而且持续时间长，是一种很好的长期记忆基质。与玻尔兹曼机学习算法（见方框7.2）十分相似，突触上的赫布可塑性是由输入和输出之间的一致性决定的。

更令人惊讶的是，玻尔兹曼机需要一定的“睡眠”才能学习。它的学习算法有两个阶段。在第一阶段，即“清醒”阶段，随着输入和

输出模式被钳制到期望的映射，网络中的单元被多次更新以达到平衡，每对单元都为1的次数所占的百分比会被记录下来。在第二阶段，即“睡眠”阶段，输入和输出单元被释放，每对单元在自由运行状态中都为1的次数百分比也被记录下来。然后，每个连接强度按照清醒和睡眠阶段一致率之间的差异，按比例进行更新（见方框7.2）。睡眠阶段的计算是为了确定钳制的相关性中哪个部分是由外部原因造成的。在不去除内部产生的相关性的情况下，该网络会加强内部活动模式，并学习忽略外部的影响，这是二联性精神病（folie à deux）[21]的网络版本。有趣的是，极度的睡眠剥夺会导致人体处于妄想的状态，在没有窗户和恒定照明的医院重症监护病房中，这样的问题并不罕见。精神分裂症患者通常患有睡眠障碍，可能导致他们患上妄想症。这使我们确信，我们正沿着正确的方向理解大脑的工作原理。

学习识别镜像对称

玻尔兹曼机可以解决，但感知器却不能解决的一个问题，就是如何学习镜像对称。[22]人体是沿着一个垂直的轴两边对称的。我们可以用这个对称轴生成大量随机图案，如图7－5所示，也可以采用水平轴和对角轴对称。在我们的玻尔兹曼机网络中，按照10×10排布的若干块二进制输入被投影到16个隐藏单元中，然后再投影到3个输出单元，每个输出单元对应3个可能的对称轴中的一个。在接受了6000个对称输入模式的训练后，玻尔兹曼机对全新输入的对称轴进行分类的成功率为90%。感知器得到的结果不会比随机猜测的结果好，因为单一的输入并不包含关于模式对称性的信息，必须对输入对之间的相关性进行考察。值得注意的是，人类观察者看到的输入阵列与玻尔兹曼机所看到的并不一致，因为玻尔兹曼机的每个隐藏单元都从整个阵列接收输入，并不遵循任何特定的顺序。对于观察者来说，等价的问题是要

随机安排阵列中输入单元的位置，这会使阵列在观察者看来是随机的，即使其中存在隐藏的对称性。

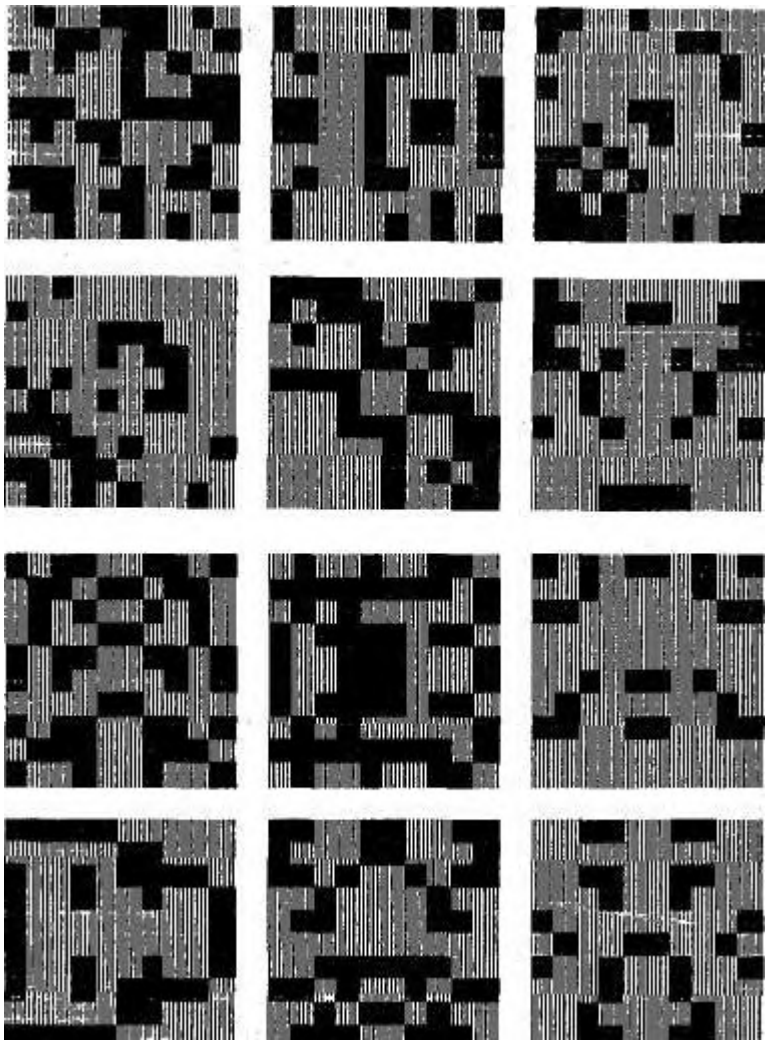


图7-5 对称的随机图案。每个10×10的阵列都存在垂直、水平或对角方向上的镜像对称。网络模型的目标是学习如何在没有参与网络模型训练的新样本中依据对称轴进行分类。图片来源：T. J. Sejnowski, P. K. Kienker,and G. E. Hinton, “Learning Symmetry Groups with Hidden Units: Beyond the Perceptron,” *Physica*22D(1986): 260- 275, figure 4。

有一天，我正盯着显示器，以每秒两次的速度读出每个输入模式的对称性。当时我在约翰·霍普金斯大学心理学系的同事尼尔·科恩（Neal Cohen）也在观察这个屏幕，但他却不能区分这些对称性，除非凑近了仔细看，他对我的这项本事感到很惊讶。与玻尔兹曼机的学

习过程类似，观察了显示器数日之后，我的视觉系统被训练到可以自动检测对称性，而不需要在屏幕上四处对比查看。尼尔和我设计了一个实验，用本科生作为未经训练的考察对象，观察他们在这一任务上的进展。[23]开始时，他们花了很长时间才找出正确的对称性，但经过几天的训练以后，他们的速度变得飞快。到了实验后期，他们能够快速、轻松地找到对称性，甚至可以在任务过程中与我们进行交谈，得到的答案依然准确无误。这是非常快速的感知学习。

我在约翰·霍普金斯大学教授过计算生物物理学，这门课程吸引了很多有才华的学生和研究人员。本·尤哈斯（Ben Yuhas）是与我一起工作的电子工程系的研究生，他在写作博士论文期间，训练了一个神经网络来阅读唇语。[24]人的嘴巴在运动过程中包含了声音的信息。尤哈斯的网络将嘴部的图像转换为每个时间点上产生声音的相应频谱。这一技术可以辅助处理有噪声的声谱以改善语音识别。他的研究生同学安德里亚斯·安德里欧（Andreas Andreou）是希腊族塞浦路斯人，他当时正在Barton Hall大楼的地下室里建造模拟超大规模集成电路芯片（将在第14章中进行介绍）。在20世纪80年代，许多院系的教师对待神经网络的态度都不太友善，这样的现象在很多机构中都非常普遍，但这并不能阻止尤哈斯或安德里欧研究的脚步。事实上，安德里欧后来成了约翰·霍普金斯大学的正教授，并合作建立了约翰·霍普金斯大学语言和语音处理中心。尤哈斯则组建了一个咨询团队，为政府和企业客户提供数据科学咨询服务。

学习识别手写数字

最近，杰弗里·辛顿和他在多伦多大学的学生们训练了一个带有三层隐藏单元的玻尔兹曼机，能够对手写邮政编码进行非常准确的分类（见图7－6）。[25]由于玻尔兹曼网络具有反馈和前馈连接，因此可以反向运行网络，钳制其中一个输出单元，并生成与钳制输出单元

相对应的输入模式（见图7－7）。这种生成性网络能捕获训练集的统计结构，而它们生成的样本也会继承这些属性。就好像这些网络进入了睡眠状态，网络最高层的活动在输入层产生了梦境一般的状态序列。

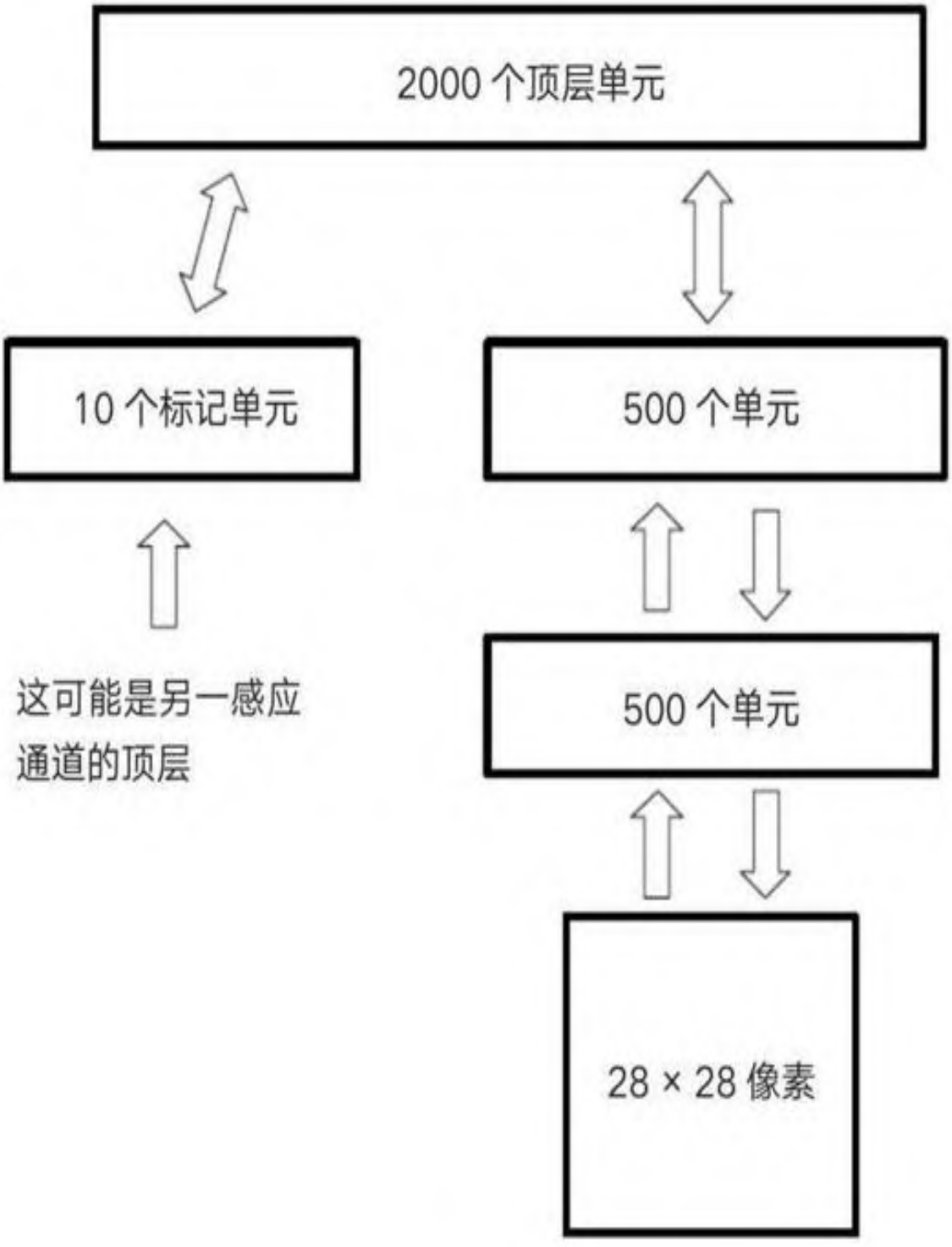


图7-6 用于手写数字识别和生成样本的多层玻尔兹曼机。图像具有28×28个像素，可以是白色或黑色。目标是根据10个输出单位（0~9）对数字进行分类。图片来源：G. E. Hinton, “Learning Multiple Layers of Representation, ”*Trends in Cognitive Sciences*11(2007):428-434, figure 1。



图7-7 由经过训练后可以识别手写数字的多层玻尔兹曼机生成的输入层图案。每一行都是通过钳制10个输出单元中的一个产生的（见图7-6），并且输入层在上述示例之间持续变化。这些数字都没有在训练集中出现过——它们是被训练有素的网络的内部结构“幻想”出来的。图片来源：G. E. Hinton, S. Osindero and Y. Teh, “ADeep Learning Algorithm for Deep Belief Nets”, *Neural Computation*18(2006):1527-1554, figure 8。

尽管神经网络在物理学和工程学中的兴起十分迅速，但传统的认知科学家却迟迟不能接受它作为理解记忆和语言处理的形式体系。除了拉荷亚的并行分布处理（PDP）研究组和一些独立的研究者，符号处理仍然是业界的主流方法。杰弗里和我在1983年参加了认知科学协会（Cognitive Science Society）的会议。会议期间，研究短期记忆和意象的心理学家泽农·派利夏恩（Zenon Pylyshyn）表示出了对玻尔兹曼机的不屑。他朝讲台上泼了一杯水，并大声喊道：“这不是计

算！”而其他人则认为整个领域仅仅是在搞统计罢了。但杰罗姆·莱特文（Jerome Lettvin）却表示，他真的非常喜欢我们正在做的事情。莱特文在1959年与温贝托·马托拉那（Humberto Matourana）、沃伦·麦卡洛克（Warren McCulloch）和沃尔特·匹茨（Walter Pitts）一起撰写了经典的论文《青蛙的眼睛告诉大脑什么》（What the Frog's Eye Tells the Frog's Brain）。[26]文中给出了青蛙视网膜中的虫类探测器神经元对小黑点的反应最为强烈的证据，这是一个在系统神经科学领域极具影响力的想法。他对我们羽翼未丰的神经网络模型的支持是该领域早期发展中的重要一环。

无监督学习和皮层发育

玻尔兹曼机既可以用来进行监督学习，即输入和输出都被钳制；也可以用于无监督学习，即只有输入被钳制。杰弗里·辛顿使用无监督的版本一次一层地搭建出了一个深度玻尔兹曼机。[27]从连接到输入单元的一层隐藏单元开始——即受限玻尔兹曼机（restricted Boltzmann machine），杰弗里用未标记的数据对它进行训练，这些数据比标记数据更容易获得（互联网上有数十亿未标记的图像和录音），可以使学习进度更快。无监督学习的第一步是从数据中提取所有数据共有的统计规律，但第一层隐藏单元只能提取简单的、用感知器也可以表现的特征。下一步是将权重固定到第一层，并在上面添加第二层单元。更多的无监督玻尔兹曼学习产生了更复杂的一组特征，可以不断重复这一过程来创建具有多层深度的网络。

因为上层的单元包含更多非线性的低层特征组合，使得它们可以作为一个整体，从具体的特征中抽象出更普遍的特征。因此，在上层进行的分类变得容易得多，只需要更少的训练样本就能在更高的计算水平达到收敛。尽管如何描述这种解决方案背后的数学原理依然有待商榷，但已经有新的几何工具开始在这些深度网络中崭露头角了。

[28]皮层似乎也是逐层生长的。在视觉系统发育的早期阶段，作为第一层从眼睛接收输入信号的神经元，初级视觉皮层中的神经元具有很高的可塑性，并且很容易根据视觉输入流重新建立连接，这种连接在关键时期结束后消失（这部分在第5章已经讨论过）。大脑后部的视觉区域和其他感官流的层级结构最先发育成熟，靠近大脑前部的皮层区域则需要更长的时间。前额叶皮层，也就是最靠近大脑前端的部分，直到成年早期才可能完全发育成熟。在关键期，皮层区域的连接接受神经活动影响最大，这些层叠的关键时期导致了皮层的逐渐发育。加州大学圣迭戈分校的认知科学家杰弗里·艾尔曼（Jeffrey Elman）和伊丽莎白·贝茨（Elizabeth Bates）跟其他同事一起，针对皮层的逐渐发育如何帮助儿童通过了解世界而获得新能力，给出了联结主义网络角度的解释。[29]这一工作为解释我们漫长的童年如何使人类成为最善于学习的物种，开辟了一个新的研究方向，同时也对先前围绕某些与生俱来的行为产生的论调提供了新的视角。

史蒂文·库沃茨（Steven Quartz）曾经是我实验室里的博士后研究员，现在在加州理工学院做教员。我们在合作发表的《骗子、爱人和英雄》（Liars, Lovers and Heroes）[30]一文中曾经写道，在儿童和青少年时期的大脑发育过程中，经验可以深刻地影响神经元的基因表达，从而改变负责行为的神经回路。基因的差异性以及环境影响之间的交互作用，让我们能够从新的角度认识大脑发育的复杂性。这一活跃的研究领域超越了先天与后天的辩论，并从文化生物学的角度对其进行了重新定义。我们的生物特性不仅促成了人类文明的诞生，也反过来被人类文明影响着。[31]最近的一项发现为这个故事续写了新篇章：当神经元之间突触的形成在早期发育过程中迅速增加时，神经元内部的DNA在诞生后通过一种甲基化的形式进行表观遗传修饰，这种甲基化调节基因的表达是大脑所独有的。[32]这种表观遗传修饰可将我和史蒂文之前设想的经验与基因之间建立起某种连接。

到了20世纪90年代，神经网络革命开始如火如荼地进行。认知神经科学领域正在扩展，计算机变得越来越快，但还不够快。玻尔兹曼机的技术性能非常出色，但是要模拟起来却慢得让人无法忍受。真正帮助我们取得进展的是一种更快的学习算法，恰恰在我们最需要它的时候，它与我们不期而遇了。

---

[1] 大多数神经元可以做出决定的最快速度大约为10毫秒，并且在1秒内做出决定所需要的时间不超过100时步。

[2] 涉及电磁学时，法拉第（Michael Faraday）的物理学是邋遢型的，而麦克斯韦（James Clerk Maxwell）的是整洁型的。

[3] Theodore Holmes Bullock and G. Adrian Horridge, *Structure and Function in theNervous Systems of Invertebrates*(San Francisco: W. H. Freeman, 1965).

[4] E. Chen, K. M. Stiefel, T. J. Sejnowski, and T. H. Bullock, “Model of Traveling Waves in a Coral Nerve Network,” *Journal of Comparative Physiology*A 194, no. 2(2008): 195–200.

[5] D. S. Levine and S. Grossberg, “Visual Illusions in Neural Networks: Line Neutralization, Tilt after Effect, and Angle Expansion,” *Journal of TheoreticalBiology*61, no. 2 (1976):477–504.

[6] G. B. Ermentrout and J. D. Cowan, “A Mathematical Theory of Visual Hallucination Patterns,” *Biological Cybernetics*34, no. 3 (1979):137–150.

[7] J. J. Hopfield, “Neural Networks and Physical Systems with Emergent Collective Computational Abilities,” *Proceedings of the National Academy of Sciences of the United States of America*79, no. 8 (1982): 2554–2558.

[8] 尽管1976年的马尔–波吉奥立体视觉模型（第4章提到）的神经网络是对称的（因为马尔和波吉奥对所有单元进行了同步更新），但其网络的动态状况比使用异步更新的霍普菲尔德网络要复杂得多。D. Marr, G. Palm, and T. Poggio T, “Analysis of a Cooperative Stereo Algorithm,” *Biological Cybernetics*28, no. 4(1978): 223–239.

[9] L. L. Colgin, S. Leutgeb, K. Jezek, J. K. Leutgeb, E. I. Moser, B. L. McNaughton,and M. B Moser, “Attractor-Map versus Autoassociation Based Attractor Dynamics in the Hippocampal Network,” *Journal of Neurophysiology*104, no. 1 (2010):35–50.

[10] 文献中采用模拟计算网络模型，通过sigmoid激发函数产生输出。——译者注

[11] J. J. Hopfield and D. W. Tank,“‘Neural’ Computation of Decisions in Optimization Problems,” *Biological Cybernetics*52, no. 3 (1985):141－152. 旅行商问题在计算机科学界很出名，它是解决问题所需的时间随着问题规模的增大而迅速增加的典型代表。

[12] Dana H. Ballard and Christopher M. Brown, *Computer Vision*(Englewood Cliffs,NJ: Prentice Hall, 1982).

[13] D. H. Ballard, G. E. Hinton, and T. J. Sejnowski, “Parallel Visual Computation,”*Nature*306, no. 5938 (1983): 21－26; R. A. Hummel and S. W. Zucker, “On the Foundations of Relaxation Labeling Processes,” *IEEE Transactions on PatternAnalysis and Machine Intelligence*5, no. 3 (1983): 267－287.

[14] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi, “Optimization by Simulated Annealing,” *Science*220, no. 4598 (1983): 671－680.

[15] P. K. Kienker, T. J. Sejnowski, G. E. Hinton, and L. E. Schumacher, “Separating Figure from Ground with a Parallel Network,” *Perception*15 (1986): 197－216.

[16] H. Zhou, H. S. Friedman, and R. von der Heydt, “Coding of Border Ownership in Monkey Visual Cortex,” *Journal of Neuroscience*20, no. 17 (2000): 6594－6611.

[17] 与边缘单元相背离的图形单元被认为是背景。——译者注

[18] 退火算法中的温度变量。——译者注

[19] 钳制成所期望的输入值和输出值。——译者注

[20] Donald O. Hebb, *The Organization of Behavior: A Neuropsychological Theory*(New York: Wiley & Sons., 1949), 62.

[21] 形容一个有精神病症状的人，将妄想的信念传递给另一个人，使对方也产生感应性妄想。——译者注

[22] T. J. Sejnowski, P. K. Kienker, and G. E. Hinton, “Learning Symmetry Groups with Hidden Units: Beyond the Perceptron,” *Physica*22D (1986): 260－275.

[23] N. J. Cohen, I. Abrams, W. S. Harley, L. Tabor, and T. J. Sejnowski, “Skill Learning and Repetition Priming in Symmetry Detection: Parallel Studies of Human Subjects and Connectionist Models,” in

*Proceedings of the 8*th *Annual Conferenceof the Cognitive Science Society*(Hillsdale, NJ: Erlbaum, 1986), 23–44.

[24] B. P. Yuhas, M. H. Goldstein Jr., T. J. Sejnowski, and R. E. Jenkins, "Neural Net- work Models of Sensory Integration for Improved Vowel Recognition," *Proceedings of the IEEE*78, no. 10 (1990): 1658–1668.

[25] G. E. Hinton, S. Osindero, and Y. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*18, no. 7 (2006): 1527–1554.

[26] J. Y. Lettvin, H. R. Maturana, W. S. McCulloch, and W. H. Pitts, "What the Frog's Eye Tells the Frog's Brain," *Proceedings of the Institute of Radio Engineers*47,no. 11 (1959): 1940–1951. http://hearingbrain.org/docs/letvin_ieee_1959.pdf.

[27] R. R. Salakhutdinov and G. E. Hinton, "Deep Boltzmann Machines," in *Proceedings of the 12*th *International Conference on Artificial Intelligence and Statistics*, *Journal of Machine Learning Research*5 (2009): 448–455. 保罗·斯莫伦斯基（Paul Smolensky）介绍了玻尔兹曼机的一种特殊情况，他称之为Harmonium: P. Smolensky, "Information Processing in Dynamical Systems:Foundations of Harmony Theory," in David E. Rumelhart and James L. McLelland(eds.), *Parallel Distributed Processing:*Explorations in the Microstructure of Cognition, Volume 1: *Foundations*(Cambridge, MA: MIT Press, 1986), 194–281.

[28] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, "Exponential Expressivity in Deep Neural Networks through Transient Chaos," in *Advances in Neural Information Processing Systems*29 (2016): 3360–3368.

[29] Jeffrey L. Elman, Elizabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith,Domenico Parisi, and Kim Plunkett, *Rethinking Innateness: A Connectionist Perspective on Development*(Cambridge, MA: MIT Press, 1996).

[30] Steven R. Quartz and Terrence J. Sejnowski, *Liars, Lovers and Heroes: What theNew Brain Science Has Revealed about How We Become Who We Are*(New York:Harper-Collins, 2002).

[31] S. Quartz and T. J. Sejnowski, "The Neural Basis of Cognitive Development: A Constructivist Manifesto," *Behavioral and Brain Sciences*20, no. 4 (1997): 537–596.

[32] 这被称为"非CG甲基化"（non-CG methylation）。参阅R. Lister, E. A.Mukamel, J. R. Nery, M. Urich, C. A. Puddifoot, N. D. Johnson, J. Lucero, Y.Huang A. J. Dwork, M. D. Schultz, M. Yu, J. Tonti-Filippini, H. Heyn, S.

Hu, J. C.Wu, A. Rao, M. Esteller, C. He, F. G. Haghighi, T. J. Sejnowski, M. M. Behrens,J. R. Ecker, “Global Epigenomic Reconfiguration during Mammalian Brain Development,” *Science*341, no. 6146 (2013): 629.

08
反向传播算法

加州大学圣迭戈分校成立于1960年，现已发展成为生物医学研究的主要中心。它在1986年成立了世界上第一个认知科学系。[1]大卫·鲁姆哈特（见图8－1）在那时已经是一位杰出的数学家和认知心理学家，他曾在以符号学和规则为基础的传统人工智能领域工作过，这类研究在20世纪70年代的人工智能研究中占主导地位。1979年，我在加州大学圣迭戈分校由杰弗里·辛顿组织的研讨会上第一次见到了大卫，当时他开创了一种新的探索人类心理的方法，他和詹姆斯·麦克莱兰称之为“并行分布式处理”（以下简称PDP）。大卫总是能对问题进行深入思考，并经常提出富有洞察力的评论。

玻尔兹曼机学习算法可以学习如何解决需要隐藏单元的问题，这表明，训练多层网络并突破感知器的限制是可行的，而这种观点与马文·明斯基和西摩尔·帕普特以及该领域大多数人的观点相左。网络中的层数或任一给定层内的连接性都不存在任何限制。但是有一个问题：达到平衡和收集统计数据来进行模拟的速度变得越来越慢，大型网络需要花费更长的时间才能达到平衡。



图8-1 1986年左右在加州大学圣迭戈分校的大卫·鲁姆哈特，那时他刚出版了两卷《并行分布式处理》（*Parallel Distributed Processing*）。鲁姆哈特在多层网络模型学习算法的技术开发领域有很大的影响力，并用该技术来帮助我们理解语言和思维心理。图片来源：大卫·鲁姆哈特。

原则上，可以构建具有大规模并行体系结构的计算机，该计算机比具有每次只进行一次更新的传统冯·诺依曼体系结构的计算机快得多。20世纪80年代的数字计算机每秒只能执行100万次操作。今天的计算机每秒能够执行数十亿次操作，并且通过将数千个内核连接在一起实现的高性能计算机，其速度比以前快上百万倍——技术性能得到了空前的提高。

“曼哈顿计划”是美国在无法保证原子弹能够研制成功的情况下做出的260亿美元的赌注（以2016年的美元价值计算），最大的秘密就是它的确有成功的苗头。当使用玻尔兹曼机可以训练多层网络的秘密被公开后，许多新的学习算法如雨后春笋般不断涌现。在杰弗里·辛顿和我研究玻尔兹曼机的同时，大卫·鲁姆哈特开发了另一种多层网络学习算法，而后来的实践证明，这种算法的效率更高。[2]

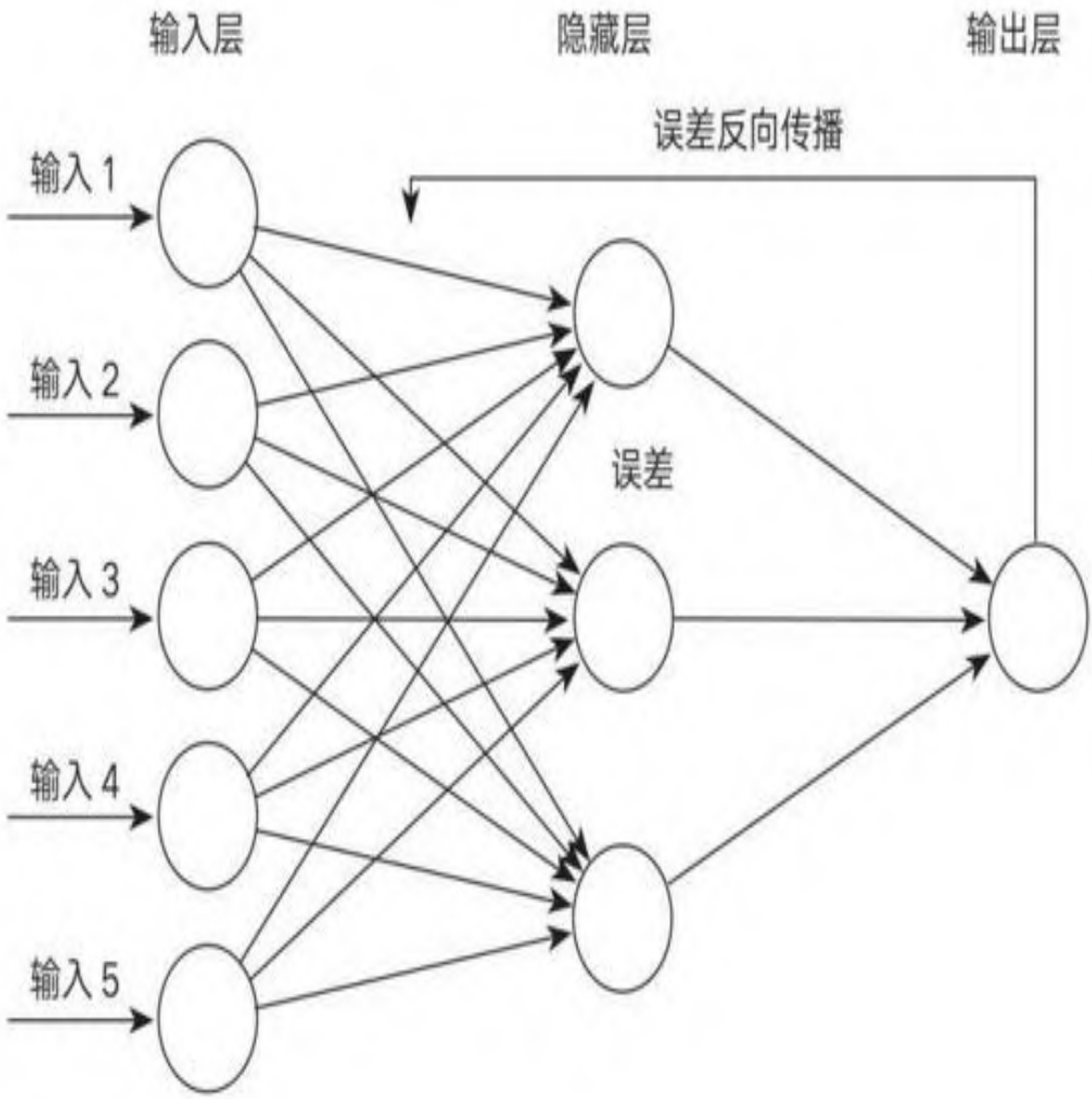
算法的优化

优化（Optimization）是机器学习中一个关键的数学概念：对于许多问题，可以找到一个代价函数（cost function），而问题的解决方案就是代价最低时的系统状态。对于霍普菲尔德网络来说，代价函数就是能量，目标是找到使网络能量最小化的状态（如第6章所述）。对于前馈网络，用于学习的常用代价函数是训练集输出层上的方差之和。“梯度下降”（gradient descent）是一种能够最小化代价函数的通用过程，实现方法是对网络中的权重沿降低代价最快的方向进行逐步修改。[3]如果将代价函数看作山脉，那么梯度下降就是选择向山脚下滑动的最快路径。

鲁姆哈特发现了如何通过被称为“误差的反向传播”（backpropagation of errors），或简称为“反传”（backprop）的过程来计算网络中每个权重的梯度（见方框8.1）。从误差已知的输出层开始，可以很容易地计算出指向输出单元的输入权重的梯度。下一步就是使用输出层梯度来计算上一层权重的梯度，以此类推，逐层回到输入层。这是一种高效计算误差梯度的方法。

尽管不像玻尔兹曼机学习算法那样拥有优雅和深厚的物理学根源，但是反向传播效率更高，并且推动该领域取得了快速进展。由大卫·鲁姆哈特、杰弗里·辛顿和罗纳德·威廉姆斯（Ronald Williams）合著的经典的反向传播论文于1986年发表在《自然》杂志上，[4]迄今为止，该论文已经在其他研究论文中被引用超过4万次。（世界上发表的所有论文里有一半从未被引用过，甚至连作者自己都不引用；一篇被引用了100次的论文都会在领域内产生巨大反响，所以很显然，这篇有关反向传播的文章是一颗重磅炸弹。）

8.1 误差反向传播



反向传播网络的输入是被传播的前馈：在该图中，左侧的输入通过连接（箭头）向前传播到隐藏的单元层，然后投影到输出层。输出结果与训练者给出的值进行比较，差值被用来更新连接输出单元的权重，以减少误差。然后，根据每个权重对误差贡献的多少，通过反向传播误差对输入单元和隐藏层之间的权重进行更新。利用大量样本进行训练，隐藏单元生成了可区分不同输入模式的选择性特征，这样一

来它们就能够在输出层中对不同类别进行区分。这一过程被称为“表征学习”（representation learning）。

语音合成的突破

1984年，我在普林斯顿听了研究生查尔斯·罗森伯格（Charles Rosenberg）关于玻尔兹曼机的演讲。虽然这通常是我演讲的题目，但这段演讲还是令我印象深刻。查尔斯问他是否可以去我的实验室参与一个夏季研究项目。他来到巴尔的摩时，我们已经转向了反向传播领域，这让我们有可能考虑现实级别的问题，而不是之前处理的那种玩具级别的问题。由于查尔斯是传奇语言专家乔治·米勒（George Miller）的学生，我们想寻找一个恰到好处的语言问题，既不会难到完全找不到头绪，又不会容易到存在现成的解决方法。语言学是一个具有许多分支学科的广阔领域，例如：音韵学（phonology），涉及单词的发音；句法学（syntax），研究单词在一个句子中是如何排列的；语义学（semantics），研究单词和句子的含义；还有语用学（pragmatics），研究语境是如何影响语义的，等等。我们决定从音韵学开始着手。

英语是一种特别难发音的语言，因为规则很复杂，并且有很多例外情况。例如，如果一个单词的最后一个辅音后面跟着一个不发音的字母“e”，则元音大多数情况下都要发长音，如“gave”和“brave”。但是也有例外，例如“have”，这个词的发音就与之前的规则相悖。我在图书馆找到了一本书，在书中音韵学家编纂了这些规则和例外，厚达数百页。通常例外情况中也会有规则，而有时例外的规则中还存在例外情况。总之，对于语言学家来说，“一路下来”都是规则。[5]更让人抓耳挠腮的是，同样一个单词，并不是每个人的发音都一样。还存在很多方言，每种方言也都有自己的一套规则。

杰弗里·辛顿在我们计划的早期阶段到约翰·霍普金斯大学拜访了查尔斯和我，跟我们说他认为英语发音太难掌握了。所以我们收小了野心，找了一本总共有大约100个字的儿童早教读物。我们设计的网络有一个由7个字母组成的窗口，每个字母由29个单元（包含空格和标点符号）表示，共203个输入单元。研究目标是预测窗口中间位置那个字母的读音。输入单元与80个隐藏单元相连，隐藏单元又投射到26个输出单元，每个输出单元对应一个基本发音，在英语里被称为“音位”（phonemes）。我们把该字母发音网络叫作“话语网络”（见图8－2）。[6]网络中有18 629个权重，按照1986年的标准衡量，这是个十分庞大的数字。而按照当时的数学统计标准来看，根本没法进行操作。有了这么多的参数，我们被告知训练集可能会被过度拟合，导致网络无法泛化。

当单词在有7个字母的窗口中依次穿过时，网络为窗口中位于中间的字母分配了一个音位。项目中花费时间最长的部分，是手动将音位与正确的字母相匹配，因为字母的数量与每个单词中音位的数量不同。相比之下，学习过程就发生在我们眼前，其表现随着句子在窗口中循环而变得越来越好。当学习收敛时，网络在有100个单词的训练集中的表现堪称完美。虽然对新单词进行测试的效果很差，但由于我们对在这样一个小的训练集上成功泛化的预期并不高，所以这个初步结果仍然令人鼓舞。

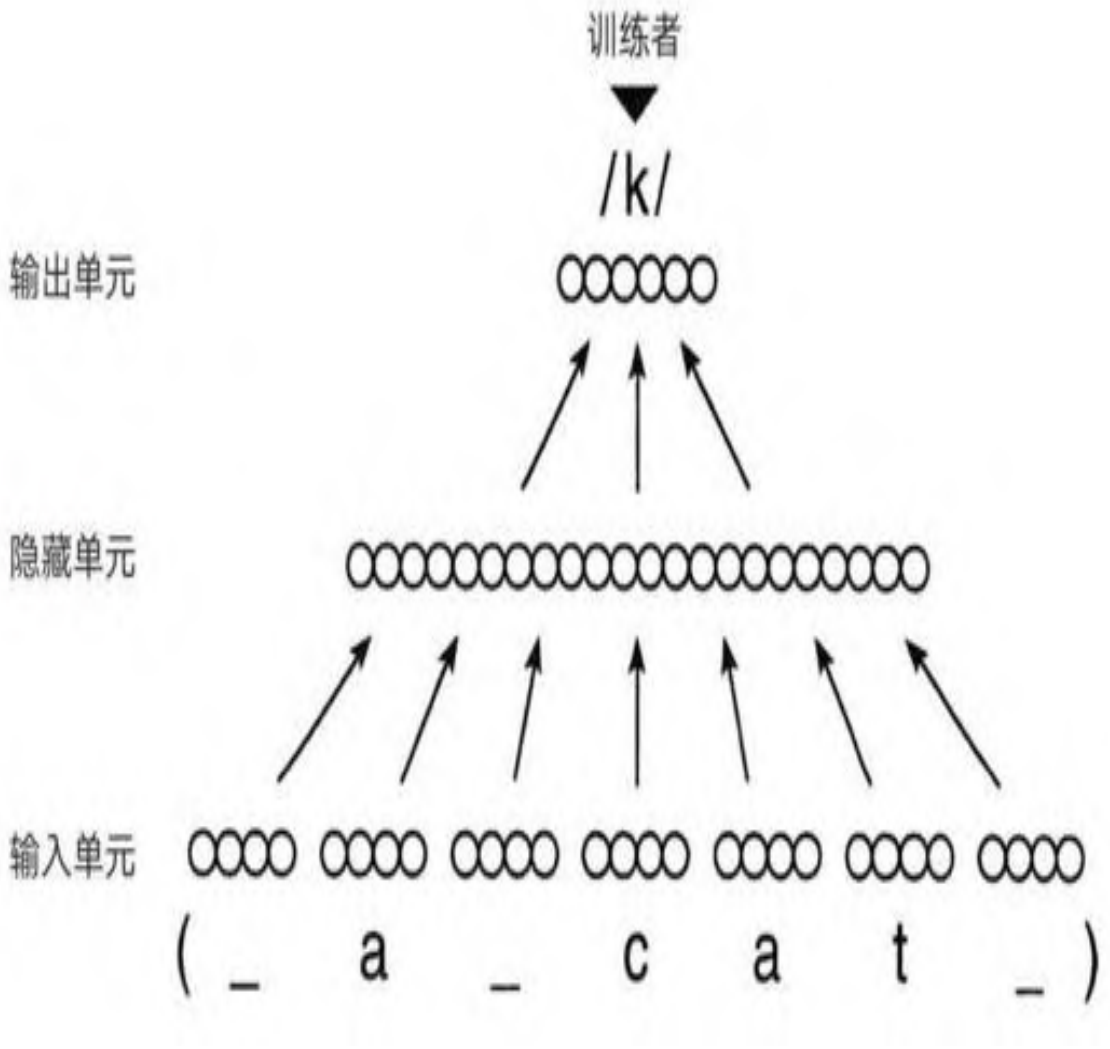


图8-2 话语网络前馈网络模型。底层的7组单元代表在文本间移动的窗口中的字母，每次提取一个字母。该网络的目标是正确预测位于中间位置的字母的声音，在这个例子中就是很难发音的“c”音位。输入层中的每个单元与所有隐藏单元都建立了连接，而后者又会投射到输出层上的所有单元。反向传播学习算法能够使用来自训练者的反馈来训练权重。正确的输出模式会与网络的输出结果进行比较，在这个例子里，输出的是“k”音位，那么误差就会被反向传播给前面若干层上的权重。

随后，我们使用了含有2万个字母的布朗语料库（Brown Corpus）[7]，并为每个字母指定了音位以及重音标记。字母和声音的对应工作花了几周的时间，但是学习开始后，网络在一个晚上就吸收了整个训练集的信息。那么它能进行泛化吗？结果证明，泛化的结果非常漂亮。该网络已经发现了英语发音的规律性，并且可以识别出

例外情况，所有这些都是基于相同的架构和学习算法。虽然按照今天的标准来看，这一成果微不足道，但我们的网络很好地证明了反向传播网络如何能够有效地表征英语音韵。这是我们得到的第一个暗示，即神经网络学习语言（符号表征的典型代表）的方式和人类的学习方式相同。

在获得了大声朗读的能力后，话语网络首先经历了一个胡言乱语的阶段，成功识别了辅音和元音之间的区别，却将音位“b”分配给了所有辅音，将音位“a”分配给了所有的元音。刚开始，它的发音听起来像“ba ba”，经过更多的学习之后，发音偏向了“ba ga da”。这种现象与婴儿咿呀学语的状态非常类似。之后它开始能够正确地说出短词的发音，最后在训练结束时，我们已经可以听懂它说的大多数单词了。

为了测试话语网络在方言上的表现，我们找到了一个来自洛杉矶郊外的拉丁裔男孩接受采访时的音韵翻录材料。训练有素的网络重新创建了一段该男孩带有西班牙语口音的英语，谈论的是他探望自己的祖母时，有时会得到糖果。通过将话语网络的输出播放到一个叫作“DECtalk”的语音合成器中，一串音位标签被转换为可听的语音，我记录下了学习阶段中的一系列语音片段。当我在某次演讲过程中播放这段录音时，台下的观众彻底震惊了——这个网络直接证明了它的语言能力。[8]这个暑期项目的结果完全超出了我们的预期，并成为神经网络学习领域的第一个实际应用。1986年，我带着话语网络参加了《今日秀》（*Today show*）节目，那一期的收视率很惊人。在此之前，神经网络一直是一门神秘的学科。我还遇到过很多人，他们在观看这个节目时是第一次听到神经网络这个概念。

虽然话语网络有力地证明了一个神经网络的确能够对语言的某些方面进行表征，但它并不是反映人类如何获得阅读技能的优质模型。首先，我们在学习阅读之前就先学会了说话。其次，有限的几个语音规则就能帮助我们开启大声精确朗读的复杂任务。但是，大声朗读很

快就变成了快速的模式识别，并不需要有意识地应用规则。大多数会说英语的人都会在阅读刘易斯·卡罗尔（Lewis Carroll）的诗Jabberwocky时，不由自主地读出“brillig”、“slithy”和“toves”等无意义的词，就像读正常的词一样，话语网络也是如此。这些虚构的词不存在于任何字典中，但是可以触发由英语中相关字母模式组成的音位。

话语网络给观众留下了深刻的印象，不过现在，查尔斯和我需要对这个网络进行分析，弄清楚它到底是如何工作的。为此，我们对隐藏单元中的活动模式进行了聚类分析（cluster analysis），并发现话语网络察觉到了相似的元音和辅音的分类，这和语言学家们已经识别出的分类相同。马克·塞登伯格（Mark Seidenberg）和詹姆斯·麦克莱兰采用了一种类似的方法作为研究的起点，将其与儿童在学习阅读时经历的一系列阶段进行了详细比较。[9]

话语网络以出人意料的方式影响了这个世界。作为约翰·霍普金斯大学托马斯·詹金斯（Thomas C. Jenkins）生物物理系的一名教员，我开始对蛋白质折叠的问题产生了兴趣。蛋白质是由一系列氨基酸折叠成的复杂的结构，该结构赋予了蛋白质广泛的功能，例如血红蛋白，它能够与血红细胞中的氧结合。根据氨基酸序列来预测蛋白质的三维形状是一个难度很高的计算问题，对大多数蛋白质来说，即便使用功能最强大的计算机也没办法实现。然而，有一种单元结构相对更容易预测，被称为二级结构（secondary structures）。在二级结构中，氨基酸以螺旋、平面或无规卷曲的方式缠绕。生物物理学家们使用的算法考虑了不同氨基酸的化学性质，但他们的预测还不足以解决三维空间的折叠问题。

钱宁是我实验室的一年级研究生，他是1980年在中国所有物理系的学生中，为数不多被选中来美国攻读研究生课程的人之一。我们想知道，如果为每个氨基酸分配螺旋、平面或无规卷曲的参数，话语网络是否可以通过一串氨基酸序列来预测蛋白质的二级结构。这是一个

重要的问题，因为蛋白质的三维结构决定了它的功能。输入由字母序列变成了氨基酸序列，而预测的结果由音位变成了二级结构。训练集是由X射线晶体学确定的三维结构。让我们意想不到的是，它对于新蛋白质的二级结构的预测，要远远好于基于生物物理学的最佳方法，[10]这一具有里程碑意义的研究是机器学习在分子序列中的首次应用，该领域现在被称为生物信息学（bioinformatics）。

另一个学会了如何形成英语动词过去时的网络，成了认知心理学领域中备受争议的问题，基于规则的保守派与前卫的PDP研究组展开了激烈的争论。[11]形成过去时的常规方法是给英文动词加上后缀“ed”，例如把“train”变成“trained”。但是有不规则形式的例外，例如把“run”变成“ran”。神经网络可以很好地适应规则和例外情况。虽然人们在这一点上已经很少争论了，但关于规则的显式表征（explicit representation）在大脑中的角色这一问题，仍然有待回答。最近利用神经网络学习语言的实验支持了屈折形态学（inflectional morphology）中的逐步获取概念，这与人类的学习方式是一致的。[12]深度学习与谷歌翻译和其他自然语言应用相结合，在捕捉语言细微差别上获得的成功，进一步支持了大脑不需要使用显式语言规则的可能性，即便其实际行为可能体现了相反的结论。

杰弗里·辛顿、大卫·图雷斯基（David Touretzky）和我于1986年在卡内基－梅隆大学组织了第一期联结主义暑期课程（见图8－3），那时候只有少数几所大学开设了神经网络课程。在一个基于话语网络的小游戏中，学生们逐层列队，每个学生代表网络中的一个单元（尽管他们在传播“Sejnowski”中的“j”时发生了错误，因为它的发音类似于“y”，并不遵循英文发音模式）。这些学生中的许多人随后都陆续获得了重要发现，并开创了各自的事业。1988年，第二期暑期课程在卡内基－梅隆大学开办，1990年第三期的举办地是加州大学圣迭戈分校。新的想法在经历了一代人的时间后，才进入主流研究领

域。这些暑期课程让所有人受益匪浅，也是我们在早期推广该领域的最佳投资。



图8-3 卡内基-梅隆大学1986年联结主义暑期课程的学生。杰弗里·辛顿是第一排右数第三位，他的两边分别是我（右）和詹姆斯·麦克莱兰。这张照片成了当今神经计算领域的“名人录”。20世纪80年代的神经网络，很像是存在于20世纪的21世纪科学。图片来源：杰弗里·辛顿。

神经网络的重生

由鲁姆哈特和麦克莱兰编辑的两卷《并行分布式处理》于1986年出版，已经成为当今的经典著作。这是第一套阐述神经网络和多层学习算法对理解精神和行为现象的影响的书籍。它的销量超过了5万套，在学术领域算得上是畅销书了。神经网络在经过了反向传播的训练后，其隐藏单元具备了类似视觉系统中皮层神经元的属性，[13]而且其展现出的故障模式，也和人类在大脑受损后产生的缺陷有很大的相似性。[14]

弗朗西斯·克里克是PDP小组的成员，参加了小组大部分的会议和讲座。在关于并行分布式处理模型如何具有生物特性的争论中，他认为应该将它们看作可行性证明，而不是大脑的标准模型。他为 《并行分布式处理》一书撰写了其中一个章节，介绍了当时我们所了解的大脑皮层。而我写的那一章，则总结了当时我们对大脑皮层的哪些方面还不太了解。如果这些章节是今天写下的，都会比原来的版本要长得多。

20世纪80年代还有一些成功案例并不为人所知。基于神经网络技术的最赚钱的公司之一是HNC Software有限公司，由罗伯特·赫克特－尼尔森（Robert Hecht-Nielsen）创立，该公司使用神经网络来防止信用卡欺诈。赫克特－尼尔森曾在加州大学圣迭戈分校的电子与计算机工程系任教，教授过一门关于神经网络实际应用的热门课程。每天都有信用卡信息被全球网络犯罪分子盗取。信用卡的交易数据规模十分庞大，从中挑出可疑的信息是一项艰巨的任务。在20世纪80年代，工作人员要在短时间内做出是否批准或拒绝某项信用卡交易的决定，而这导致了每年1500多亿美元的交易存在欺诈行为。HNC Software有限公司使用神经网络学习算法，以远高于人类的准确度检测信用卡欺诈行为，每年为信用卡公司节省了数十亿美元。2002年，HNC被美国个人消费信用评估公司（Fair Isaac and Company，FICO）以 10亿美元的价格收购，后者以发布信用评分而闻名。

观察神经网络的学习过程颇有些神奇，它不断以微小的步伐优化自身的表现。这个过程会十分缓慢，但是如果有足够的训练样本，并且网络足够大，学习算法就可以找到一个很好的能够泛化的表征方式，以适应新的输入。采用随机选择的一组初始权重并重复该训练过程时，每次都会学习到不同的网络，但都具有相似的性能。许多网络可以解决同样的问题，这就暗示了当我们重建不同个体大脑的完整连接集时，应该怀有何种期待。如果许多网络表现出了相同的行为，理解它们的关键就在于大脑所使用的学习算法，而这个算法应该更容易被发现。

理解真正的深度学习

在凸优化（convex optimization）问题中，不存在局部最小值，可以保证收敛发生在全局最小值处。但在非凸优化问题中，情况就不同了。优化专家告诉我们，由于在隐藏单元网络中学习是一个非凸优化问题，所以我们只是在浪费时间——我们的网络会陷入局部最小值（见图8－4）。但经验证据表明，他们错了。为什么呢？我们现在知道，在维度很高的空间中，代价函数的局部最小值是很罕见的，直到学习的最后阶段才会出现。在早期阶段，几乎所有的方向都是下坡，而在下坡的过程中存在鞍点，一些方向会错误地开始上升，而在其他维度则继续下降。产生网络会陷入局部最小值的直觉，是因为解决低维空间中的问题时，逃生方向的数量要少得多。

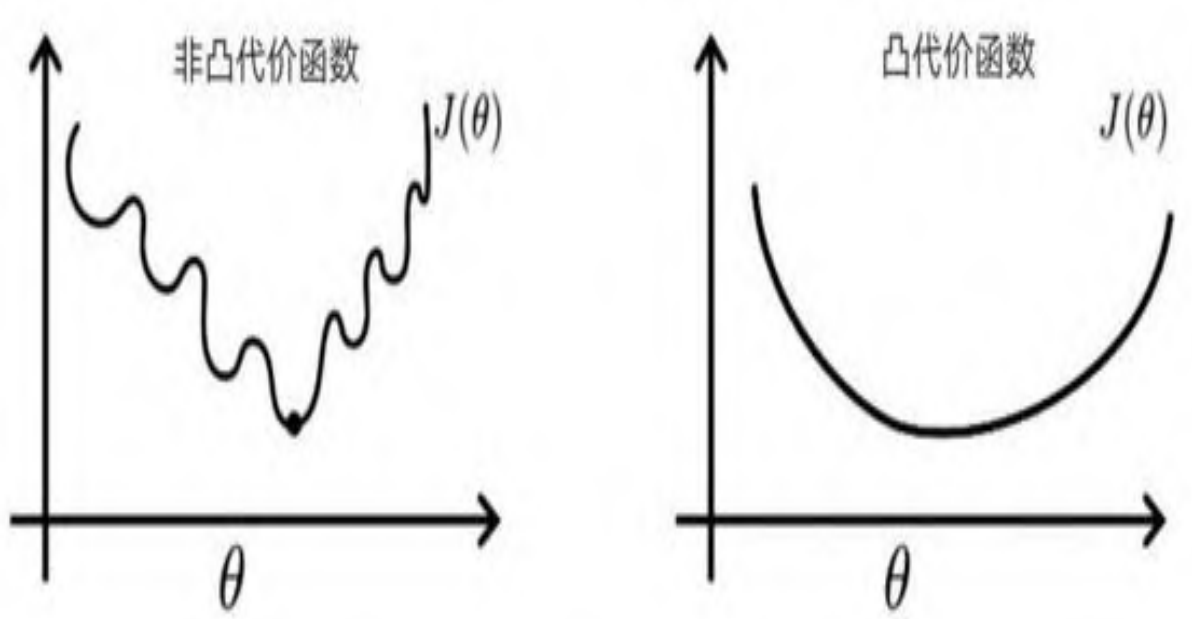


图8-4 非凸代价函数和凸代价函数。这些图形将代价函数$J(\theta)$绘制为参数θ的函数。凸函数只有一个最小值（右图），即从表面任何位置向下移动都可达到的全局最小值。想象一下，你是一名滑雪者，并始终将滑雪板朝向最陡峭的下坡方向，保证你一定会滑到底。相比之下，非凸代价函数可以包含局部最小值（左图），这些则是陷阱，让你无法通过一路下坡找到全局最小值。因此，非凸代价函数很难优化。不过这个一维的例子具有一定的误导性。当有许多参数时（通常在神经网络中有数百万个参数），就会存在鞍点，即在某些维度上会向上凸起，而在另一些维度上向下凹陷。当你位于鞍点上时，总会存在一个下坡的方向。

目前的深层网络模型有数百万个单元和数十亿的权重。统计学家在传统上只用少量参数分析简单模型，这样就可以用较小的数据集来证明定理。而拥有数十亿个维度的空间对他们来说简直就是一场噩梦。他们向我们保证，有这么多的参数，对数据的过度拟合绝对是不可避免的：我们的网络只会简单地记住训练数据，但无法泛化以适应新的测试输入。但是，使用正则化（regularization）手段，比如在已有权重对学习没有任何贡献时，可以通过强迫它们进行衰减来缓解过度拟合的现象。

一个特别巧妙的正则化技术，叫作“失活”（dropout），其发明者是杰弗里·辛顿。[15]在每个学习周期（epoch），当从许多训练实例中估计出梯度，并对权重进行更新时，有一半的单元就会被随机

地暂时从网络中剔除——这就意味着，每个周期中训练的都是不同的网络。其导致的结果就是，在每个周期中需要训练的参数都更少，并且与在每个周期中训练相同的大型网络相比，单元之间的依赖关系更少。失活过程将深度学习网络的错误率降低了10%，这是一个很大的改进。2009年，网飞（Netflix）举办了一场公开竞赛——第一个将他们推荐系统的错误减少10%的人，会得到100万美元的奖金。[16]几乎每个机器学习领域的研究生都参与了竞争。网飞的这笔奖金很可能启发了价值1000万美元的研究。现在，深度网络已经成为在线流媒体的核心技术。[17]

有趣的是，皮层突触会以很高的速度失活。对于来自输入的放电，皮层中典型的兴奋性突触都有90%的失败率。[18]这就像一个棒球队，几乎所有球员的击球成功率都是0.1。大脑是如何利用这种不可靠的皮层突触来实现可靠的功能呢？当每个神经元上存在数千个概率突触时，其活动总和的多样性相对较低，[19]这意味着其性能可能不会像你想象的那样大规模地下降。学习过程中在突触层面发生失活的好处，可能远大于准确性降低的代价。而且由于突触需要消耗大量的能量，失活也可以节省能量。最后，皮层使用概率来计算可能的，而非确定的结果，所以使用概率性分量是表示概率的有效方式。

虽然皮层突触也许并不可靠，但是它们的强度却惊人的精确。皮层突触的大小及其相应的强度在100倍的范围内变化，单一突触的强度可以在此范围内增加或减少。与得克萨斯大学奥斯汀分校的神经解剖学家克里斯汀·哈里斯（Kristen Harris）合作，我的实验室最近重建了一小块大鼠的海马体，这是大脑中形成长期记忆所需的区域，其中包含450个突触。大多数轴突会在树突的分支上形成单个突触，但在少数情况下，一个轴突的两个突触会接触到相同的树突。令人惊讶的是，它们的尺寸几乎相同，之前的研究经验告诉我们，这意味着它们也具有相同的强度。人们已经了解了导致这些突触强度发生变化的条件。这些突触的强度变化取决于输入尖峰的历史和树突响应的电活

动，对于来自同一个树突上相同轴突的一对突触来说也是如此。根据这些观察，我们推断信息存储在突触强度中的精度很高，足以存储至少5位信息。[20]深度循环网络的学习算法只需要5位就能实现高水平的性能，这很可能不是巧合。[21]

大脑网络的维度非常高，我们对其甚至没有很好的估计。大脑皮层中的突触总数约为100万亿，几乎是个天文数字。人类的寿命不过几十亿秒。以这样的速度，你可以为你生活中的每一秒贡献出10万个突触。在实际情况中，神经元往往具有聚集的局部连接，例如在由10亿个突触连接的由10万个神经元组成的皮层柱内。虽然这仍然是一个很大的数字，但还远算不上天文数字。长距离连接比本地连接要少得多，因为神经连接会占用宝贵的空间，并且会消耗大量能量。

代表皮层中一个对象或概念的神经元的数量，是一个重要的数字。粗略估计一下，需要的突触数量约为10亿，需要的神经元数量约为10万，分布在10个皮层区域中。[22]也就是说，约10万个独立的、互不干扰的对象类别和概念存储在100万亿个突触中。在实际情况中，代表相似对象的神经元群是重叠的，这可以大大增加皮层表达相关对象和对象之间关系的能力。这种能力在人类中比在其他哺乳动物中要强大得多，因为人类大脑中的联合皮层（associative cortex，在感官和运动层级的上方）在进化过程中发生了显著的扩张。

高维空间中概率分布的研究在20世纪80年代还是一个相对而言未被开发的统计领域。有几位统计学家研究了在探究高维空间和高维数据集时出现的统计问题，例如斯坦福大学的里奥·布莱曼（Leo Breiman），他是NIPS社区中的一员。来自该社区的一些人，例如加州大学伯克利分校的迈克尔·乔丹（Michael Jordan）也在统计系任职。然而大多数情况下，大数据时代的机器学习已经发展到了令统计学家望而生畏的程度。但仅仅通过训练大型网络来做出令人惊叹的事情是不够的，我们也需要分析和理解它们是如何做到的。物理学家在

这方面占据了领先地位，由于神经元和突触的数量越来越大，他们利用了统计物理学的方法来分析学习的特性。

2017年在长滩举办的NIPS会议上，“时间考验奖”（the Test of Time award）被授予了加州大学伯克利分校的本杰明·雷希特（Benjamin Recht）和谷歌的阿里·拉希米（Ali Rahimi）在2007年联名发表的NIPS论文。[23]该论文表明，随机特征可以有效地提高具有一层学习权重网络的性能的有效方法，这是弗兰克·罗森布拉特在1960年通过感知器的试验了解到的。拉希米在获奖后的演讲中发出了对机器学习严谨性的强烈呼吁，他感叹深度学习缺乏严谨性，并嘲讽它为“炼金术”。我当时正坐在早已火冒三丈的杨立昆（Yann LeCun）旁边。听完演讲，杨立昆在一篇博客中写道：“批评整个团体（还是那个领域中非常成功的团体）是在钻研 ‘炼金术’，仅仅因为我们目前的理论工具还没有赶上我们的实践，这是十分危险的做法。为什么危险呢？正是这种态度，导致机器学习社区将神经网络的研究搁置了超过10年，尽管有充分的经验证据表明，它们在许多情况下运行良好。”[24]这是一次经典的邋遢和整洁的科学方法之间的混战。想要取得进展，这两个都是不可或缺的。

神经网络的局限性

虽然神经网络可以给出一个问题的正确答案，但目前还没有办法解释它们是如何得到这个答案的。例如，假设一名感到胸部刺痛的女性患者来到急诊室。这是心肌梗死的一种症状，需要立即进行干预，还是仅仅是由一种严重的消化不良引起的呢？训练有素的神经网络可能会比医生做出更准确的诊断，但如果不清楚网络为何做出这样的决定，我们就有理由不信任它。医生和算法一样，也要接受一系列测试和决策点的指导，通过常规案例进行培训。但问题在于，有些罕见的情况并不在他们算法覆盖的范围之内，而神经网络则经历过更多案例

的训练，远远超过一般医生在一生中会经历的，网络可能会很好地捕捉到这些罕见病例。但是，你会相信无法解释其理由，但从统计上来说诊断能力更强的神经网络，而不去相信看似有凭有据的医生吗？事实上，那些能够精确诊断罕见病例的医生都有着丰富的经验，并且大多数使用了模式识别而不是算法。[25]对于各个领域的最高级别专家来说，情况可能都是如此。

正如可以训练网络来提供专业的诊断，是否有可能把其背后的解释作为训练集的一部分，来训练网络对其行为进行解释呢？这样一来也许还可以帮助改善诊断。这一建议是存在问题的，原因是医生给出的许多解释都是片面的、过度简化的，或错误的。每一代医疗实践与前一代相比都发生了巨大的变化，因为人体的复杂性大大超出了我们目前的理解能力。如果我们可以通过分析网络模型的内部状态来提取因果解释，就可能会产生能够推动医学发展的新的见解和假设。

神经网络是一个黑盒子，其理论尚无法被理解，这个陈述对大脑也成立。事实上，在给定相同数据的情况下，不同的个体做出的决定也会千差万别。我们的确还不清楚大脑是如何从经验中得出推论的。就如第三章所讨论的，结论并不总是基于逻辑，其中还存在认知偏差。[26]此外，我们所接受的解释，往往仅仅是合理的解释或似是而非的故事。我们并不能排除某一天某个非常大的生成网络会突然开始说话的可能性，到时我们就可以向它寻求解释了。我们是否应该期待，能从这样的网络获得比从人类那里得到的更好的故事和合理的解释呢？回想一下，意识并不涉及大脑的内部运作。深度学习网络通常会按照排名次序提供几个而非一个主要的预测，这为我们提供了关于结论可信度的一些信息。监督神经网络只能解决用来训练网络的数据范围内的问题。如果接受过类似案例或样本的培训，神经网络应该在新案例做出一个很好的插值。但是，如果一项新的输入超出了训练数据的范围，进行外推（extrapolation）操作就比较危险了。这应该不会令人感到惊讶，因为同样的限制也适用于人类；我们不应该期望一

个物理领域的专家在政治问题上，甚至是物理学中他专长之外的领域，提供良好的建议。但是，只要数据集足够大，能够涵盖全部的潜在输入，网络对新输入进行的泛化操作，结果应该也是不错的。在实践中，人们倾向于使用类比，从比较熟悉的领域推广到新的领域，但如果两个领域存在本质上的不同，这样的类比就是错误的。

所有对输入进行分类的神经网络都存在偏倚。首先，对分类类别的选择就引入了偏倚，这种偏倚反映了我们如何对世界进行分解。例如，训练一个网络来检测草坪上的杂草是非常实用的。但是怎么定义杂草呢？一个人认为是杂草，在另一个人看来可能是野花。分类是一个反应文化偏见的更宽泛的问题。用于训练网络的数据集加剧了这种不确定性。例如，有许多公司为执法机构提供基于面部识别的犯罪分子识别系统。黑人面部识别中出现的假阳性误判比白人面孔多，因为用于训练网络的数据库中有更多的白人面孔，你拥有的数据越多，准确率就越高。[27]数据库偏差可以通过重新平衡数据来纠正，但不可避免地会存在隐藏的偏倚，这取决于数据的来源以及被用来做出什么样的决定。[28]

另一个反对依赖神经网络的理由是，人们可能会以公平为代价来优化利润。例如，如果未被充分代表的少数群体中的人去申请住房抵押贷款，而已接受过数百万次申请培训的神经网络拒绝了该申请。对网络的输入包括当前地址，以及与少数群体高度相关的其他信息。因此，即使有反对明确歧视少数群体的法律，该网络也可能会利用这些信息来暗中歧视他们。这里的问题不在于神经网络，而在于我们让它优化的代价函数。如果利润是唯一的目标，网络就会使用被提供的任何信息来最大化利润。解决这个问题的方法是将公平性作为代价函数中的一个参数。虽然最佳解决方案是在利润和公平之间找到一种明智的平衡，但必须在代价函数中明确这种平衡，这就需要有人决定如何权衡每个目标的重要性。那些人文和社会科学领域的伦理观点应该为

这些权衡提供指导。但我们必须时刻牢记，选择一个看似公平的代价函数可能会产生意想不到的后果。[29]

伊隆·马斯克、斯蒂芬·霍金以及立法者和研究人员，都曾呼吁对人工智能进行规范。2015年，3722位人工智能和机器人研究人员共同签署了一封公开信，呼吁禁止使用自动武器：

> 综上所述，我们认为人工智能在很多方面都具有造福人类的巨大潜力，这也应该是该领域的目标。但开始人工智能军备竞赛不是一个好主意，应该通过法令禁止超出人类合理控制范围的自动攻击性武器。[30]

这一禁令本意是好的，但也可能适得其反。并非所有国家都会遵循禁令。俄罗斯总统普京就曾发表过这样的言论：

> 人工智能是未来，不仅对俄罗斯来说是这样，对全人类也是一样。它带来了巨大的机会，但也会带来难以预测的威胁。无论谁成为这个领域的领导者，都将统治全世界。[31]

全面禁令的问题在于人工智能不是一个孤立的领域，而是一个拥有许多不同工具和应用的领域，其中每一样都会产生不同的结果。例如，信用评分的自动化是机器学习在20世纪80年代的早期应用。有人担心，如果他们碰巧住在声誉不好的社区内，就会得到不公正的评分。这就催生出了对用于计算分数的信息进行限制的法律规定，并要求公司通知个人如何提高分数。每个申请都存在不同的问题，最好根据具体情况进行处理，而不是盲目禁止相关研究。[32]

1987年学术休假期间，我拜访了索尔克研究所的弗朗西斯·克里克，那会儿我在加州理工学院担任神经生物学Cornelis Wiersma客座

教授。弗朗西斯当时正在建立一个实力雄厚的视觉研究小组，这是我特别感兴趣的研究题目之一。我在教员午餐时间演示了我的话语网络样带，引发了热烈的讨论。我在1989年搬到拉荷亚，这标志着我从约翰·霍普金斯大学的初级教员转变为索尔克研究所的高级教员，真令人兴奋。几乎在一夜之间，很多机会都向我敞开了怀抱，其中包括霍华德·休斯医学研究所（Howard Hughes Medical Institute）的一项任命，该机构为我26年的研究提供了慷慨的支持。

1989年加入加州大学圣迭戈分校时，我遗憾地发现，那位教授我们反向传播理论的大卫·鲁姆哈特已经去了斯坦福大学，在那之后我很少有机会再见到他。多年来，我注意到大卫的行为方式发生了令人不安的变化。最终，他被诊断患有额颞叶痴呆——患者大脑的额叶皮层中的神经元会逐渐流失，对他的性格、行为和语言造成影响。2011年，鲁姆哈特在他68岁时去世了，那时他已经认不出家人或朋友了。

---

[1] 加州大学圣迭戈分校的认知科学系，由人类和人体工程学专家唐·诺曼（Don Norman）创立，拥有一支不拘一格的教师队伍。

[2] 反向传播学习算法中使用的数学已经存在了一段时间，可以追溯到20世纪60年代的控制理论文献，但是其最大的影响体现于在多层感知器中的应用。参阅Arthur E. Bryson and Yu-Chi Ho, *Applied Optimal Control: Optimization,Estimation, and Control*(University of Michigan, Blaisdell, 1969).

[3] 参阅迈克尔·I. 乔丹（Michael I. Jordan）具有权威性的关于现代随机梯度下降的演讲：“On Gradient-Based Optimization: Accelerated, Distributed,Asynchronous, and Stochastic,” May 2, 2017, Simons Institute for the Theory of Computing, UC, Berkeley. https://simons.berkeley.edu/talks/michae-ljordan-2017-5-2/.

[4] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning Representations by Back-Propagating Errors,” *Nature*323, no. 6088 (1986), 533–536.

[5] 有一个在人群中被广泛传播的故事。伯特兰·罗素曾经就天文学发表过公开演讲。在讲座结束时，房间后面的一位老妇人起身说道：“你讲的那些都是胡说八道。这个世界就是在一只巨大的乌龟背上。”罗素微笑着问：“那乌龟站在什么地方呢？”“你非常聪明，

年轻人，非常聪明，”老妇人说，“但我知道答案。一路向下，都是乌龟！”这位老太太用递归解决了她的问题，尽管是以无限回归为代价。不过在实践中，循环总是会终止的。

[6] C. R. Rosenberg And T. J. Sejnowski, “Parallel Networks That Learn to Pronounce English Text,” *Complex Systems*1 (1987): 145–168.

[7] W. Nelson Francis and Henry Kucera, “A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers,” Brown University, 1964;revised and amplified, 1979, http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM.

[8] 神经网络在不同学习阶段的声音记录可以从这里下载：http://papers.cnl.salk.edu/~terry/NETtalk/nettalk.mp3。

[9] M. S. Seidenberg and J. L. McClelland, “A Distributed Developmental Model of Word Recognition and Naming,” *Psychological Review*96, no. 4 (1989), 523–568.

[10] N. Qian and T. J. Sejnowski, “Predicting the Secondary Structure of Globular Proteins Using Neural Network Models,” *Journal of Molecular Biology*, 202(1988): 865–884.

[11] David E. Rumelhart and James L. McClelland, “On Learning the Past Tense of English Verbs,” in Rumelhart and McClelland, eds., *Parallel DistributedProcessing: Explorations in the Microstructure of Cognition*(Cambridge, MA:MIT Press, 1986), 2:216–271; J. L. McClelland and K. Patterson, “Rules or Connections in Past- Tense inflections: What Does the Evidence Rule Out?” *Trendsin Cognitive Sciences*6, no. 11 (2002): 465–472; and S. Pinker and M. T. Ulman,“The Past and Future of the Past Tense,” *Trends in Cognitive Sciences*6, no. 11(2002): 456–463.

[12] M. S. Seidenberg and D. C. Plaut, “Quasiregularity and Its Discontents: The Legacy of the Past Tense Debate,” *Cognitive Science*38, no. 6 (2014): 1190–1228.

[13] D. Zipser and R. A. Andersen, “A Back-Propagation Programmed Network That Simulates Response Properties of a Subset of Posterior Parietal Neurons,” *Nature*331, no. 6158 (1988):679–684. 该网络将视网膜上目标的位置从“视网膜坐标”转换为“头中心坐标”，同时考虑到了眼睛的位置。

[14] G. E. Hinton, D. C. Plaut, and T. Shallice, “Simulating Brain Damage,” *ScientificAmerican*269, no. 4 (1993): 76–82. “脑部受损的成年人在阅读单词时，会出现一些奇怪的错误。如果一个有着模拟神经元的网络在经过阅读训练后发生了损坏，它会表现出惊人的相似行为”(76)。

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov,“Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journalof Machine Learning Research*15 (2014):1929－1958.

[16] “Netflix Prize,” *Wikipedia*, last modified, August 23, 2017, https://en.wikipedia .org/wiki/Netflix_Prize.

[17] Carlos A. Gomez-Uribe, Neil Hunt, “The Netflix Recommender System:Algorithms,” *ACM Transactions on Management Information Systems*6, no. 4(2016) , article no. 13.

[18] T. M. Bartol Jr., C. Bromer, J. Kinney, M. A. Chirillo, J. N. Bourne, K. M. Harris,and T. J. Sejnowski, “Nanoconnectomic Upper Bound on the Variability of Synaptic Plasticity,” *eLife*, 4:e10778, 2015, doi:10.7554/eLife.10778.

[19] 这来自概率论中的“大数定律”（law of large numbers）。这就是为什么赌场总是赢，尽管他们可能会在短期内亏损。

[20] Bartol Jr. et al., “Nanoconnectomic Upper Bound on the Variability of Synaptic Plasticity.”

[21] J. Collins, J. Sohl-Dickstein, and D. Sussillo, “Capacity and Trainability in Recurrent Neural Networks,” 2016, https://arxiv.org/pdf/1611.09913.pdf. 给予巧合的现象过多的权重是有风险的：“一天有24小时，一箱啤酒有24瓶。这些都是巧合吗？我认为不是。”每年的4月24日，人们都会在普林斯顿大学的保罗·纽曼日上庆祝这一巧合。

[22] 对突触维度的粗略估计，可以通过对突触数量的下限和上限乘积取平方根来得到。Lawrence Weinstein and John A. Adam, Guesstimation: Solving the World’s Problems on the Back of a Cocktail Napkin (Princeton: Princeton University Press,2009), 3. 以皮层突触总数，100万亿，为上限，并以单个神经元上突触的数量，1万，为下限，表示复杂对象所需的突触数量的粗略估计约为10亿。对所需的神经元数量应用相同的经验法则：皮层中的神经元数量上限为100亿，下限为一个神经元，因此表示复杂对象所需的神经元数量为10万个，大约相当于一平方毫米皮层下的神经元数量。但这些神经元可能广泛分布在皮层的不同部位。我们可以估计出必须相互连接才能代表一个概念的皮层区域的数量：皮层区域总数的上限为100，下限为1，因此估计值为10个皮层区域，每个区域包含1万个神经元。“BRAIN计划”正在开发的新技术将通过实验来确定这些数字。

[23] Ali Rahimi and Benjamin Recht, “Random Features for Large-Scale Kernel Machines,” *Advances in Neural Information Processing Systems*20 (2007).

[24] https://www.Facebook.com/yann.lecun/posts/10154938130592143/.

[25] Mukherjee, “A.I. versus M.D.” *New Yorker*, April 3, 2017. http://www.newyorker.com/magazine/2017/04/03/ai-versus-md/.

[26] Daniel Kahneman, *Thinking, Fast and Slow*(New York: Farrar, Straus & Giroux,2011).

[27] Clare Garvie and Jonathan Frankle, “Facial-Recognition Software Might Have a Racial Bias Problem,” *The Atlantic*, Apr 7, 2016. https://www.theatlantic.com/technology/archive/2016/04/the-underlying-bias-of-facial-recognition-systems/476991/.

[28] Kate Crawford, “Artificial Intelligence’s White Guy Problem,” *New York Times*,June 25, 2016. https://www.nytimes.com/2016/06/26/opinion/sunday/artificialintelligences-white-guy-problem.html.

[29] Barbara Oakley, Ariel Knafo, Guruprasad Madhavan, and David Sloan Wilson(eds.), *Pathological Altruism*(Oxford: Oxford University Press, 2011).

[30] https://futureoflife.org/open-letter-autonomous-weapons/.

[31] http://www.cnn.com/2017/09/01/world/putin-artificial-intelligence-will-ruleworld/index.html.

[32] Andrew Burtjan, “Leave A.I. Alone,” *New York Times*, January 4, 2018. https://www.nytimes.com/2018/01/04/opinion/leave-artificial-intelligence.html.

09
卷积学习

到了2000年，自20世纪80年代兴起的神经网络热潮已经退去，神经网络再次成为常规科学。托马斯·库恩（Thomas Kuhn）曾将科学革命之间的时间间隔描述为，科学家在一个已经确定的范式或解释框架内进行理论推定、观察和试验的常规工作阶段。[1]1987年，杰弗里·辛顿去了多伦多大学，并继续着渐进式改进，虽然这些改进都没有像曾经的玻尔兹曼机那样展现出魔力。辛顿在21世纪头十年成为加拿大高等研究院（Canadian Institute for Advanced Research，简称CIFAR）神经计算和自适应感知项目（Neural Computation and Adaptive Perception，简称NCAP）的带头人。该项目由来自加拿大和其他国家的约25位研究人员组成，专注于解决机器学习的难题。我是由杨立昆担任主席的NCAP顾问委员会的成员，会在每年NIPS会议召开之前参加该项目的年会。神经网络的先驱们在缓慢而稳定的过程中探索了机器学习的许多新策略。虽然他们的网络有许多有价值的应用，但却一直没有满足20世纪80年代对该领域抱有的很高的期望。不过这并没有动摇先驱者们的信念。回想起来，他们一直是在为飞跃性的突破奠定基础。

机器学习的稳步发展

NIPS会议是20世纪80年代神经网络的孵化器，为其他可处理大型高维数据集的算法打开了大门。弗拉基米尔·瓦普尼克的支持向量机

于1995年引发了轰动，为20世纪60年代就被遗弃的感知器网络开辟了一个新篇章。使支持向量机成为功能强大的分类器，并出现在每个神经网络工作者工具包中的，是“内核技巧”（kernel trick），这是一种数学转换，相当于将数据从其采样空间重新映射到使其更容易被分离的超空间。托马索·波吉奥开发了一种名为“HMAX”的分级网络，可以对有限数量的对象进行分类。[2]这表明，网络的性能会随着其深度的增加而提高。

在21世纪的头几年里，图形模型被开发出来，并与被称为“贝叶斯网络”（Bayes networks）的丰富的概率模型相结合，后者是基于18世纪英国数学家托马斯·贝叶斯（Thomas Bayes）提出的一个定理，该定理允许使用新的证据来更新先前的信念。加州大学洛杉矶分校的朱迪亚·珀尔，在早些时候曾将基于贝叶斯分析的“信念网络”（belief networks）[3]引入人工智能，通过开发能够利用数据在网络中学习概率的方法，对贝叶斯分析进行了加强和扩展。这些网络以及其他网络的算法为机器学习研究人员打造出了强大的工具。

随着计算机的处理能力继续呈指数增长，训练更大规模的网络成为可能。大家曾普遍认为，具有更多隐藏单元、更宽的神经网络，比具有更多层数、更深的网络的效果更好，但是对于逐层训练的网络来说并非如此，[4]并且误差梯度的消失问题（the vanishing error gradient problem）被发现减慢了输入层附近的学习速度。[5]然而，当这个问题最终被克服的时候，我们已经可以对深度反向传播网络进行训练了，而且该网络在基准测试中表现得更好。[6]随着深度反向传播网络开始在计算机视觉领域挑战传统方法，2012年的NIPS大会上出现了这样一句话：“神经信息处理系统”里的“神经”又回来了。

在20世纪的最后10年以及21世纪前10年的计算机视觉领域，在识别图像中的对象方面取得的稳步进展，使得基准测试（用于比较不同方法）的性能每年能提高百分之零点几。方法改进的速度十分缓慢，

这是因为每个新类别的对象，都需要有关专家对能够将它们与其他对象区分开来所需的与姿态无关的特征进行甄别。随后，在2012年，杰弗里·辛顿和他的两名学生艾力克斯·克里泽夫斯基（Alex Krizhevsky）和伊利娅·苏特斯科娃向NIPS会议提交了一篇论文，关于使用深度学习训练AlexNet识别图像中的对象，AlexNet是本章要重点讨论的深度卷积网络。[7]以拥有22 000多个类别，超过1 500万个标记过的高分辨率图像的ImageNet数据库作为基准，AlexNet史无前例地将识别错误率降低到了18%。[8]这次性能上的飞跃在计算机视觉社区中掀起了一股冲击波，加速推动了更大规模网络的发展，现在这些网络几乎已经达到了人类的水平。到2015年，ImageNet数据库的错误率已降至3.6%。[9]当时还在微软研究院的何恺明及其同事使用的低错误率深度学习网络，在许多方面都与视觉皮层十分相似；这类网络由杨立昆最早提出，并最初把它命名为“Le Net”。

20世纪80年代，杰弗里·辛顿和我第一次见到这个法国学生杨立昆（见图9－1，右）。他9岁时，就深受1968年史诗级的科幻电影《2001太空漫游》（*2001: A Space Odyssey*）中的任务计算机HAL 9000的启发，想要开发人工智能。他曾独立发明了反向传播误差算法的一种版本，并记录在他1987年的博士论文中，[10]之后他就搬到多伦多，加入了杰弗里的团队。后来，他转去了美国电话电报公司（AT&T）在新泽西州霍姆德尔（Holmdel）的贝尔实验室，在那里他创造了一个可以读取信件上的手写邮政编码的网络，采用修订的美国国家标准与技术研究院（Modified National Institute of Standards and Technology，简称MNIST）数据库作为一种标记数据基准。每天有数百万封信件需要递送到信箱里；而今天，这个过程是完全自动化的。同样的技术也可以用来自动读取ATM机上银行支票的金额。有趣的是，最难的部分其实是查找支票上数字的位置，因为每张支票都有不同的格式。早在20世纪80年代，杨立昆就显露出了证明原理（学者们

擅长的事情）并将之应用在现实世界中的非凡天赋。后者要求实际产品必须经过严格的测试，且表现稳健。



图9-1 杰弗里·辛顿和杨立昆是深度学习领域的大师。这张照片是2000年左右在加拿大高等研究院的神经计算和自适应感知项目会议上拍摄的，该项目是深度学习领

域的孵化器。图片来源：杰弗里·辛顿。

卷积网络的渐进式改进

杨立昆在2003年去了纽约大学后，仍继续开发他的视觉网络，现在被称为卷积网络（ConvNet）（见图9－2）。这个网络的基本结构是基于卷积的，卷积可以被想象成一个小的滑动滤波器，在滑过整张图像的过程中创建一个特征层。例如，过滤器可以是一个定向边缘检测器，就像第5章中介绍的那样，只有当窗口对准图像中具有正确方向或纹理的对象的边缘时，才会产生大数值输出。尽管第一层上的窗口只是图像中的一小块区域，但由于可以有多个滤波器，因此在每个图块中都能得到许多特征信息。第一层中与图像卷积的滤波器，与大卫·休伯尔和托斯坦·威泽尔在初级视觉皮层中发现的“简单细胞”类似（见图9－3）。[11]更高层次的滤波器则对更复杂的特征做出响应。[12]在卷积网络的早期版本中，每个滤波器的输出都要通过一个非线性的Sigmoid函数（输出从0平稳地增加到1），这样可以抑制弱激活单元的输出（见方框7.2中的Sigmoid函数）。第二层接收来自第一层的输入，第二层的窗口覆盖了更大的视野区域，这样经过多层之后，就会存在一些能接收整个图像输入的单元。这个最顶层就类似于视觉层级的顶层，在灵长类动物中被称为“下颞叶皮层”，并且具有覆盖大部分视野的感受野。接着，顶层的单元被送入分类层，与其中的所有分类单元连接，再采用反向传播误差的方式训练整个网络，对图像中的对象进行分类。

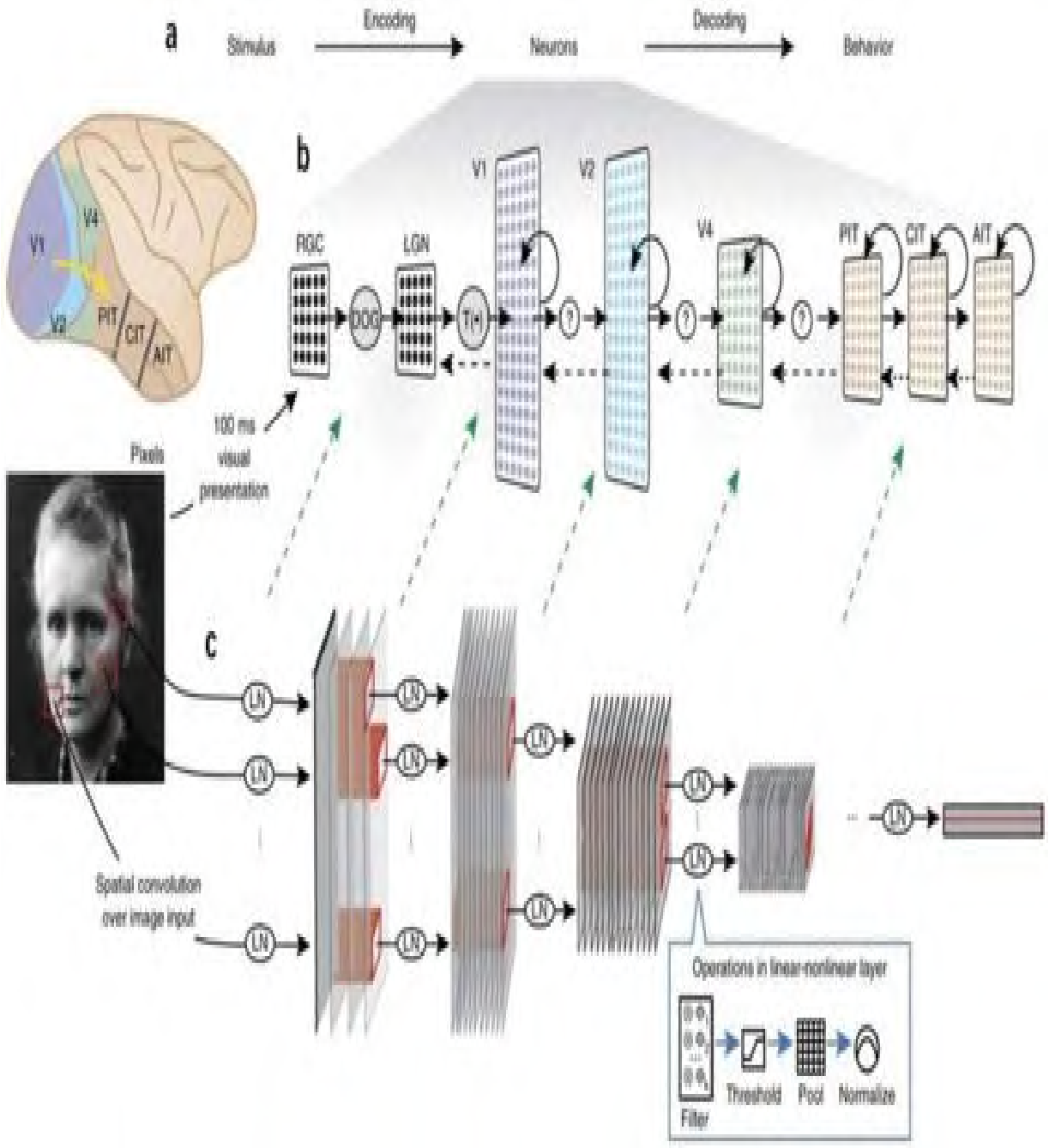


图9-2 视觉皮层与卷积网络在图像对象识别上的比较。（上图）（a，b）视觉皮层中的层级结构，从视网膜输入到初级视觉皮层（V1），经过丘脑（RGC，LGN）到下颞叶皮层（PIT，CIT，AIT），展示了视觉皮层区域和卷积网络中层次的对应关系。（下图）（c）左侧图像作为输入映射到第一个卷积层，后者由几个特征平面组成，每个特征平面代表一个滤波器，类似在视觉皮层中发现的定向简单单元。这些滤波器的输出经过阈值处理并汇集到第一层，再进行归一化处理，以便在小块区域中产生不变的响应，类似于视觉皮层中的复杂细胞（图中方框：线性—非线性层中的操作）。以上操作在网络的每个卷积层上重复。输出层与来自上一个卷积层的全部输

入具有全面的连接（每个输出单元都有上一层全部单元的输入）。图片来源：Yamins and DiCarlo, “Using Goal-Driven Deep Learning Models to Understand Sensory Cortex”, figure 1。

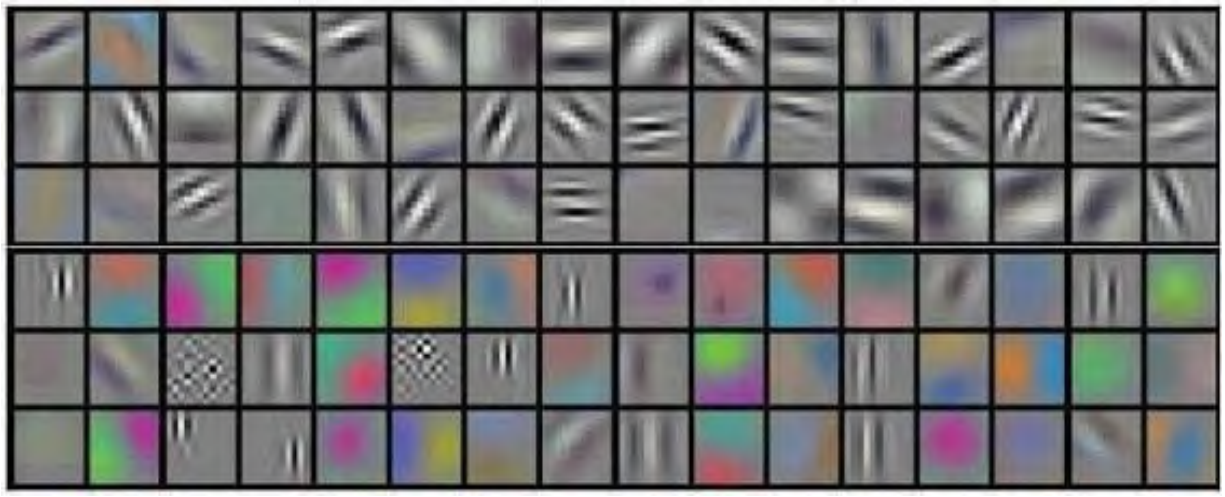


图9-3 卷积网络第一层的滤波器。每个滤波器都作用于视野中的一小块图像区域。顶部三排中滤波器的优选刺激像视觉皮层中的简单细胞一样具有定向性。底部三排显示的优选刺激经过了扩展，并具有复杂的形状。图片来源：Krizhevsky, Sutskever and Hinton, “ImageNet Classication with Deep Convolutional Neural Networks”, figure 3。

卷积网络多年来一直在经历许多渐进式改进。一个重要的补充，是将一个区域上的每个特征聚合起来，叫作“池化”（pooling）。这种操作提供了一种平移不变性（translation invariance）的量度，类似于由休伯尔和威泽尔在初级视觉皮层中发现的复杂细胞，能够通过一个图块对整个视野中相同方向的线做出响应。另一个有用的操作是增益归一化（gain normalization），就是调整输入的放大倍数，使每个单元都在其操作范围内工作，在皮层中是通过反馈抑制（feedback inhibition）实现的。Sigmoid输出函数也被线性整流函数（rectified linear units，简称ReLUs）取代。在输入达到一个阈值之前这些单元的输出都为零，超过阈值之后则输出和输入呈线性增长。该操作的优点在于：低于阈值的单元被有效地排除在网络外，这更接近真实神经元中阈值的作用。

卷积网络的每一个性能的改进，其背后都有一个工程师可以理解的计算理由。但有了这些变化，它越来越接近20世纪60年代我们所了解的视觉皮层的体系结构，尽管当时我们只能去猜测简单和复杂单元的功能是什么，或者层级结构顶部的分布式表征的存在意味着什么。这说明了生物学与深度学习之间存在相得益彰的共生关系的潜力。

当深度学习遇到视觉层级结构

加州大学圣迭戈分校的帕特里夏·丘奇兰德不仅是心灵哲学家，同时也研究神经哲学。[13]知识最终取决于大脑如何表达知识的说法，显然没有人阻止哲学家认为知识是独立于世界而存在的一种东西，用伊曼努尔·康德（Immanuel Kant）的话来说，就是“Ding an sich”（物自身）。但同样清楚的是，如果我们（和其他动物一样）要在现实世界中生存，背景知识就是必不可少的。经过训练的多层神经网络的隐藏单元之间的活动模式，与被逐次记录下的大量生物神经之间的活动模式存在显著的相似性。受到这种相似性的驱动，帕特里夏和我在1992年编写了《计算脑》（*The Computational Brain*）一书，为基于大量神经元的神经科学研究开发了一个概念框架。[14]（该书现在已经出到第二版了，如果你想更多地了解大脑式的运算，这会是一本很好的入门参考。）麻省理工学院的詹姆斯·狄卡罗（James DiCarlo）最近比较了猴子视觉皮层层级结构中不同神经元和深度学习神经网络中的单元，训练它们识别相同图片中的对象，分别观察它们的响应（见图9－2）。[15]他得出结论：深度学习网络中每层神经元的统计特性，与皮层层级结构中神经元的统计特性非常接近。

深度学习网络中的单元与猴子视觉皮层中神经元性能存在相似性，但其原因仍然有待研究，尤其是考虑到猴子的大脑不太可能使用反向传播方式来进行学习。反向传播需要将详细的错误信号反馈给神

经网络每层中的每个神经元，其精度比生物神经网络中已知反馈连接的精度要高得多。但其他学习算法在生物学上似乎更合理，例如玻尔兹曼机学习算法，该算法使用了已经在皮层中被发现的赫布突触可塑性。这引出了一个有趣的问题，是否存在一种深度学习的数学理论，能够适用于一大类学习算法（包括皮层中的那些）呢？在第7章中，我提到了对视觉层级结构的上层分类表面的分析，其决策表面比更低层级的表面更平坦。对决策表面的几何分析可能会引出对深度学习网络和大脑更深入的数学理解。

深度学习神经网络的一个优点是，我们可以从网络中的每个单元提取“记录”，并追踪信息流从一层到另一层的转变。然后可以将分析这种网络的策略用于分析大脑中的神经元。关于技术的一个奇妙之处在于，技术背后通常都有一个很好的解释，并且有强烈的动机来得到这种解释。第一台蒸汽发动机是由工程师根据他们的直觉建造的；解释发动机如何工作的热力学理论随后出现，并且帮助提升了发动机的效率。物理学家和数学家对深度学习网络的分析也正在顺利进行着。

有工作记忆的神经网络

自20世纪60年代以来，神经科学已经走过了漫长的道路，从我们目前对大脑的了解中可以获得很多东西。1990年，帕特里夏·高德曼-拉奇克（Patricia Goldman-Rakic）训练了一只猴子来记住一个地点，作为提示，该地点会短暂地被一盏灯照亮；她还训练这只猴子在一段时间的延迟之后，把眼睛移动到被记住的地点。[16]在记录了猴子前额叶皮层的活动后，她在报告中提到，一些最初对提示做出回应的神经元在延迟期间仍然保持活跃状态。心理学家把人类的这种活动称为“工作记忆”，也正因为有了工作记忆，我们在执行任务（比如拨打电话号码）时，能够记住7±2项内容。

传统的前馈网络将输入传到网络中，一次传播一层网络。结合工作记忆，可以使后续的输入与之前的输入在网络中留下的痕迹进行交互。例如，把法语句子翻译成英文时，网络中的第一个法语单词会影响后续英语单词的顺序。在网络中实现工作记忆的最简单方法，是添加人类皮层中常见的循环连接。神经网络中某一层内的循环连接和之前那些层的反馈连接，使得输入的时间序列可以在时间上整合起来。这种网络在20世纪80年代被探索并广泛应用于语音识别。[17]在实践中，它在具有短程依赖性的输入方面效果很好，但当输入之间的时间间隔很长，输入的影响会随着时间的推移发生衰减，网络性能就会变差。

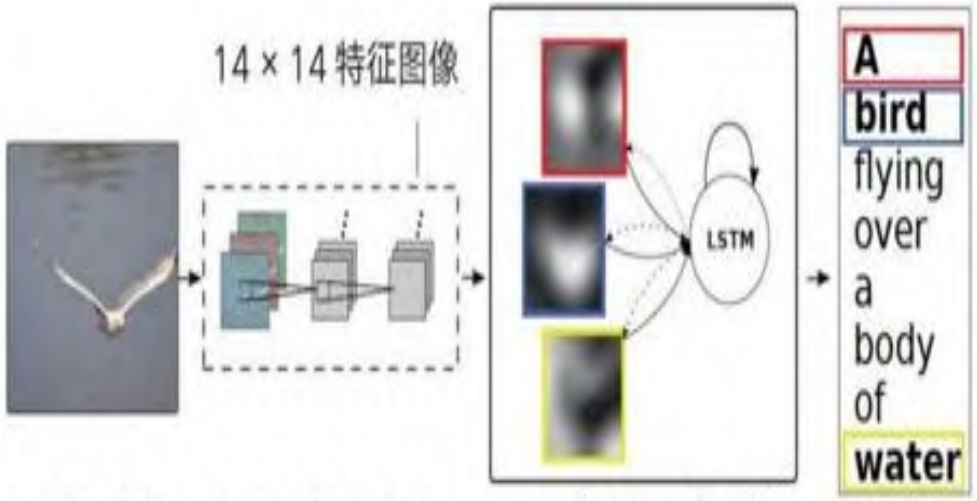
1997年，赛普·霍克莱特（Sepp Hochreiter）和尤尔根·施密德胡博（Jürgen Schmidhuber）找到了一种方法来克服衰变问题，他们称之为“长短期记忆”（long short-term memory，简称LSTM）。[18]默认情况下，长短期记忆会传递原始信息，而不会发生衰减（这就是猴子前额叶皮层的延迟期中发生的事情），并且它也有一个复杂的方案来决定如何将新的输入信息与旧信息整合。于是，远程依赖关系可以被选择性地保留。神经网络中这种工作记忆版本沉寂了长达20年之久，直到它在深度学习网络中再次被唤醒和实现。长短期记忆和深度学习的结合在许多依赖输入输出序列的领域都取得了令人瞩目的成功，例如电影、音乐、动作和语言。

施密德胡博是位于瑞士南部提契诺州（Ticino）曼诺小镇的Dalle Molle人工智能研究所的联合主任。该小镇靠近阿尔卑斯山，周围有一些绝佳的徒步地点。[19]神经网络领域的这位颇具创造性、特立独行的“罗德尼·丹泽菲尔德”[20]相信他的创造力并没有得到足够的赞誉。因此，在蒙特利尔举办的2015年NIPS会议的一次小组讨论会上，他再次向与会人员介绍了自己，“我，施密德胡博，又回来了”。而在巴塞罗那举行的2016年NIPS大会上，他因培训宣讲人没有对自己的想法给予足够的关注，而打乱对方的演讲长达5分钟。

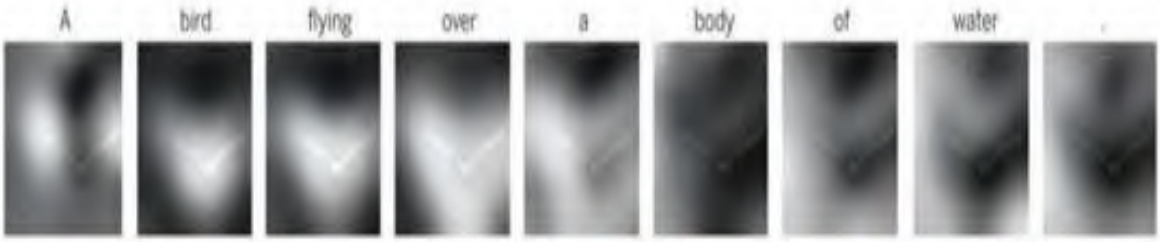
2015年，Kelvin Xu及其同事在用一个深度学习网络识别图像中对象的同时，还连接了一个长短期记忆循环网络来标注图片。使用来自深度学习网络第一遍识别的场景中所有对象作为输入，他们训练长短期记忆循环网络输出一串英文单词，能够形容一个标注中的场景（见图9－4）。他们还训练了长短期记忆网络来识别图像中的位置，使其对应于标注中的每个单词。[21]该应用令人印象深刻的地方在于，长短期记忆网络从未被训练来理解标注中句子的含义，只是根据图像中的对象及其位置输出一个语法正确的单词串。再加上第8章里早期的话语网络示例，这更加证明了神经网络似乎对语言有种亲和力，但其中的原因我们却不得而知。通过分析长短期记忆网络也许会引出一种新的语言理论，它将阐明网络的工作原理和自然语言的性质。

生成式对抗网络

在第7章中，玻尔兹曼机被当作一个生成模型进行了介绍，当输出被钳制到一个它已训练识别的类型中，并且其活动模式向下渗透到输入层时，就可以产生新的输入样本。伊恩·古德费洛（Ian Goodfellow）、约书亚·本吉奥（Yoshua Bengio）和他们在蒙特利尔大学的同事们表示，可以训练前馈网络，在对抗的背景（adversarial context）下生成更好的样本。[22]一个生成卷积网络可以通过尝试欺骗另一个卷积神经网络来训练生成优质的图像样本，后者必须决定一个输入的图像是真实的还是虚假的。生成网络的输出被用来作为一个经过训练的判别卷积网络（discriminative convolutional network）的输入，后者只给出一个单一的输出：如果输入是真实图像，就返回1，否则返回0。这两个网络会相互竞争。生成网络试图增加判别网络的错误率，而判别网络则试图降低自身的错误率。由这两个目标之间的紧张关系产生的图像，拥有令人难以置信的照片级的真实感（见图9－5）。[23]



1. 输入图像　2. 卷积特征提取　3. 注意力在图像上的 RNN　4. 逐字生成





A woman is throwing a frisbee in a park.

（一个女人在公园里扔飞盘。）



A dog is standing on a hardwood floor.

（一只狗站在地板上。）



图9-4 深度学习为图片做标注。顶部的一组图片说明了分析照片的步骤。ConvNet（CNN）在第一步中标记了照片中的对象，并将其传递给循环神经网络（RNN）。RNN被训练输出适当的英文单词串。底部的四组图片则阐明了进一步细化的过程，即使用注意力（白色云）来表示照片中单词的指示对象。顶图来源：M. I. Jordan and T. M. Mitchell,“Machine learning: Trends, Perspectives, and Prospects,”*Science*349, no. 6245(2015): 255–260, 图2. 底图来源：Xu et al., “Show, Attend and Tell,”2015, rev. 2016, figure1 and 3,https://arxiv.org/pdf/1502.03044.pdf, Coutesy of Kelvin Xu。

别忘了，这些生成的图像是合成的，它们中的对象并不存在。它们是训练集中未标记图像的泛化版本。请注意，生成式对抗网络是无监督的，这使得它们可以使用无限的数据。这些网络还有许多其他应用，包括清除具有超高分辨率的星系天文图片[24]中的噪声，以及学习表达富有情感的言语。[25]

通过慢慢地改变生成式网络的输入向量，有可能逐渐改变图像，使得部件或零碎物品（如窗户）逐渐显现或变成其他物体（如橱柜）。[26]更值得关注的是，有可能通过添加和减去表示网络状态的向量以获得图像中对象的混合效果，如图9–6所示。这些实验的意义在于，生成网络对图像中空间的表征，正如我们如何描述场景的各个

组成部分。这项技术正在迅速发展，其下一个前沿领域是生成逼真的电影。通过训练一个反复演绎的生成式对抗网络，与类似玛丽莲·梦露这样的演员参演的电影进行对比，应该有可能创造出已过世的演员出演的新作品。

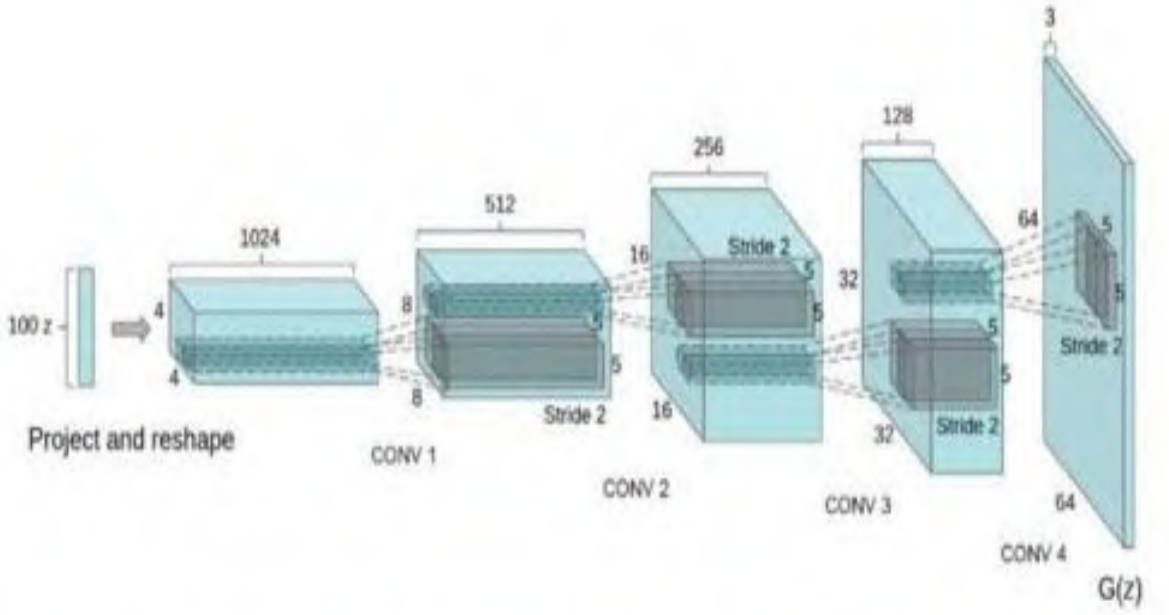






图9-5 生成式对抗网络（GAN）。顶部的示意图展示了一个卷积网络，用于生成一组样本图像，经过训练后可以欺骗判别卷积网络。左边的输入是100维的随机选取的连续值向量，用来生成不同的图像；输入的向量随后激活空间尺度逐层变大的滤波器层。下方的图显示了通过训练来自单个类别照片的生成式对抗网络产生的样本图像。顶图来源：A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,”图1, arXiv:1511.06434, https://arxiv.org/pdf/1511.06434.pdf, 由Soumiyh Chintala提供；底图来源：A. Nguyen, J. Yosinski, Y. Bengio, A.Dosovitskiy, and J. Clune, “Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space,” figure1, https: //arxiv.org/pdf/1612.00005.pdf, 由Ahn Nguyen提供。



图9-6 生成式对抗网络中的向量算法。用面部图片训练的生成式网络的输入混合后，产生了输出（左图），然后通过添加或减去选定的输入向量进行输出，就创建出了混合后的图像（右图）。因为混合是在最高的表征层完成的，所以部位和姿势是无缝接合的，并不会经过变形过程中那样的平均处理。图像改编自：A. Radford, L.Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” fifigure7, arXiv:1511.06434,https://arxiv.org/pdf/1511.06434/。

这是米兰的时装周，衣着光鲜的模特们带着超凡脱俗的表情在T台上走秀（见图9－7）。时尚界正在经历暗潮涌动：“‘很多工作正在消失，’西尔维娅·文图里尼·芬迪（Silvia Venturini Fendi）在她的时装秀开场前说道，‘机器人会承担旧的工作，但它们唯一无法取代的就是我们的创造力和思维。’”[27]现在想象一下经过训练的新一代对抗网络，它们可以生产新款式和高级时装，式样几乎无穷无

尽。时尚界可能正处于一个新时代的边缘，而许多其他依赖创意的行业也面临着相同的处境。



图9-7 2018年米兰的乔治·阿玛尼春夏男装秀。

应对现实社会的复杂性

当前的大多数学习算法是在25年前开发的，为什么它们需要那么长的时间才能对现实世界产生影响呢？20世纪80年代的研究人员使用的计算机和标记数据，只能证明玩具问题的原理。尽管取得了一些似乎颇有前景的成果，但我们并不知道网络学习及其性能如何随着单元和连接数量的增加而增强，以适应现实世界问题的复杂性。人工智能中的大多数算法缩放性很差，从未跳出解决玩具级别问题的范畴。我们现在知道，神经网络学习的缩放性很好，随着网络规模和层数的不断增加，其性能也在不断增强。特别是反向传播技术，它的缩放性非常好。

我们应该对此感到惊讶吗？大脑皮层是哺乳动物的一项发明，在灵长类动物，尤其是人类中得到了高度发展。随着它的扩展，更多的功能慢慢出现，并且更多层次被添加到了关联区域，以实现更高阶的表征。很少有复杂系统可以实现如此高级的缩放。互联网是为数不多的已经被扩大了100万倍的工程系统之一。一旦通信数据包协议建立起来，互联网就会开始进化，正如DNA中的遗传密码使细胞演化成为可能一样。

使用相同的一组数据训练许多深度学习网络，会导致生成大量不同的网络，它们都具有大致相同的平均性能水平。我们想知道的是，所有这些同等优秀的网络有哪些共同之处，而对单个网络进行分析并不能揭示这一点。理解深度学习原理的另一种方法是进一步探索学习算法的空间；我们只在所有学习算法的空间中对几个位置进行了抽样尝试。从更广泛的探索中可能会出现一种学习计算理论，该理论与其他科学领域的理论一样深奥，[28]可能为从自然界中发现的学习算法提供更多的解释。

蒙特利尔大学的约书亚·本吉奥[29]（见图9－8），和杨立昆一起，接替杰弗里·辛顿，成为CIFAR神经计算和NCAP项目的主任，该项

目在通过十年评估后更名为“机器学习和大脑学习”项目（Learning in Machines and Brains）。约书亚率领蒙特利尔大学的一个团队，致力于应用深度学习来处理自然语言，这将成为“机器学习和大脑学习”项目新的研究重点。在十多年的会议中，这个由20多名教师和研究员组成的小组开启了深度学习的研究。过去5年来，深度学习在过去难以解决的许多问题上取得了实质性进展，这些进展归功于小组成员的努力，他们当然只是一个更庞大社区中的一小部分人（将在第11章探讨）。



图9-8 约书亚·本吉奥是CIFAR“机器学习和大脑学习”项目的联合主任。这位在法国出生的加拿大籍计算机科学家，一直是应用深度学习处理自然语言问题这个领域的领导者。杰弗里·辛顿、杨立昆和约书亚·本吉奥所取得的进展，为深度学习的成功奠定了基础。图片来源：约书亚·本吉奥。

尽管深度学习网络的能力已经在许多应用中得到了证明，但如果单靠自身，它们在现实世界中永远都无法存活下来。[30]它们受到了

研究者的青睐，后者为其提供数据，调整超参数，例如学习速度、层数和每层中的单元数量，以改善收敛效果，还为其提供了大量计算资源。另一方面，如果没有大脑和身体的其他部分提供支持和自主权，大脑皮层也无法在现实世界中存活。在一个不确定的世界中，这种支持和自主权是一个比模式识别更难解决的问题。第10章将会介绍一种古老的学习算法，它通过激励我们寻求对自身有利的经验来帮助我们在自然界中生存。

---

[1] Thomas S. Kuhn, *The Structure of Scientific Revolutions*, 2nd ed. (Chicago:University of Chicago Press, 1970), 23

[2] M. Riesenhuber and T. Poggio, “Hierarchical Models of Object Recognition In Cortex.” Nat Neurosci. 2: 1019-1025, 1999; T. Serre, A. Oliva, and T. Poggio, “A Feedforward Architecture Accounts for Rapid Categorization.” *Proceedings of theNational Academy of Sciences of the United States of America*104, no. 15 (2007):6424– 6429.

[3] Pearl, *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann; 1988.

[4] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle, “Greedy Layer-Wise Training of Deep Networks,” in Bernhard Schölkopf, John Platt, and Thomas Hoffman, eds., *Advances in Neural Information Processing Systems 19:Proceedings of the 2006 Conference*(Cambridge, MA: MIT Press), 153–160.

[5] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber, “Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies,” In John F. Kolen and Stefan C. Kremer, eds., *A Field Guide to Dynamical RecurrentNeural Networks*(New York: IEEE Press, 2001), 237–243.

[6] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber, “Deep Big Simple Neural Nets for Handwritten Digit Recognition,” *Neural Computation*22, no. 12(2010): 3207–3220.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Advances in Neural Information Processing Systems*25 (NIPS 2012).

https://papers.nips.cc/paper/4824-imagenet-classificationwith-deep-convolutional-neural-networks.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems*25 (NIPS 2012). https://papers.nips.cc/paper/4824-imagenet-classificationwith-deep-convolutional-neural-networks.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2015. https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf.

[10] Yann LeCun, "Modèles connexionistes de l'apprentissage (Connectionist learning models)" (Ph.D. diss., Université Pierre et Marie Curie, Paris, 1987).

[11] Krizhevsky, Sutskever, and Hinton, "ImageNet Classification with Deep Convolutional Neural Networks."

[12] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," 2013. https://www.cs.nyu.edu/~fergus/papers/zeilerECCV2014.pdf.

[13] Patricia Smith Churchland, *Neurophilosophy: Toward a Unified Science of the Mind-Brain*(Cambridge, MA: MIT Press, 1989).

[14] Patricia Smith Churchland and Terrence J. Sejnowski, *The Computational Brain*,2nd ed. (Cambridge, MA: MIT Press 2016).

[15] D. L. Yamins and J. J. DiCarlo, "Using Goal-Driven Deep Learning Models to Understand Sensory Cortex," *Nature Neuroscience*19, no. 3 (2016): 356–365.

[16] S. Funahashi, C. J. Bruce, and P. S. Goldman-Rakic, "Visuospatial Coding in Primate Prefrontal Neurons Revealed by Oculomotor Paradigms," *Journal ofNeurophysiology*63, no. 4 (1990): 814–831.

[17] J. L. Elman, "Finding Structure in Time," *Cognitive Science*14 (1990): 179–211; M. I. Jordan, "Serial Order: A Parallel Distributed Processing Approach,"*Advances in Psychology*121 (1997): 471–495; G. Hinton, L. Deng, G. E. Dahl,A. Mohamed, N. Jaitly, A. Senior, et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, 29, no. 6(2012): 82–97.

[18] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*9, no. 8 (1997): 1735–1780.

[19] John Markoff, “When A.I. Matures, It May Call Jürgen Schmidhuber ‘Dad.’” *New York Times,*November 27, 2016, https://www.nytimes.com/2016/11/27/technology/artificial-intelligence-pioneer-jurgen-schmidhuber-overlooked.html.

[20] 美国喜剧演员，有一句非常著名的口头禅：“我觉得自己没有受到尊重。”——译者注

[21] K. Xu, J. L. Ba, K. Kiror, K. Cho, A. Courville, R. Slakhutdinov, R. Zemel, Y Bengio, “Show, Attend and Tell: Neural Image Captions Generation with Visual Attention,” 2015, rev. 2016. https://arxiv.org/pdf/1502.03044.pdf.

[22] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair,A. Courville, Y. Bengio, “Generative Adversarial Nets,” *Advances in NeuralInformation Processing Systems*, 2014. https://arxiv.org/pdf/1406.2661.pdf.

[23] 参阅 A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” 2016, https://arxiv .org/pdf/1511.06434.pdf ; Cade Metz and Keith Collins, “How an A.I. ‘Cat-and- Mouse Game’ Generates Believable Fake Photos,” New York Times,January 2, 2018. https://www.nytimes.com/interactive/2018/01/02/technology/ai-generated-photos .html.

[24] K. Schawinski, C. Zhang, H. Zhang, L. Fowler, and G. K. Santhanam, “Generative Adversarial Networks Recover Features in Astrophysical Images of Galaxies beyond the Deconvolution Limit,” 2017. https://arxiv.org/pdf/1702.00403.pdf.

[25] J. Chang and S. Scherer, “Learning Representations of Emotional Speech with Deep Convolutional Generative Adversarial Networks,” 2017. https://arxiv.org/pdf/1705.02394.pdf.

[26] A. Nguyen, J. Yosinski, Y. Bengio, A. Dosovitskiy, and J. Clune, “Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space,” 2016, https://arxiv.org/pdf/1612.00005.pdf; Radford, Metz, and Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” 2016. https://arxiv.org/pdf/1511.06434.pdf.

[27] Guy Trebay, “Miuccia Prada and Sylvia Fendi Grapple with the New World,” *New York Times*, June 19, 2017. https://www.nytimes.com/2017/06/19/fashion/mensstyle/prada-fendi-milan-mens-fashion.html.

[28] T. R. Poggio, S. Rifkin, Mukherjee and P. Niyogi. “General Conditions for Predictivity in Learning Theory,” *Nature*428, no. 6981 (2004): 419–422.

[29] 本吉奥也是包括微软在内的多家公司的顾问，他联合创建了Element AI公司，但他的重心是在学术界，并致力于科学和公共事业的进步。

[30] 参见前言部分：Churchland and Sejnowski, The Computational Brain, 2nd ed., ix–xv。

10
奖励学习

中世纪流传下来这样一个故事：一位统治者为了感谢发明国际象棋的人，想要奖励他一块麦田。发明人请求在棋盘的第一格放一粒麦子，第二格放两粒，第三格放四粒，依次类推，剩余每格都放前一格两倍的麦子，直到放满64格的棋盘。统治者觉得这个请求并不过分，就同意了。但实际上，要满足这一请求，统治者不仅要拿出他王国里所有的麦子，还要加上全世界未来几百年的麦子产量才能凑够，因为最后一格要放的麦粒数目达到了2^{64}（大约是10^{19}）。[1]这被称为“指数增长”。在国际象棋和围棋游戏中，棋盘不同布局状态的增长速度比这个故事里麦子数量增加的速度还要快得多。在国际象棋中，每一步棋平均都有35种摆法，而在围棋中，分支系数是250，这就使得围棋的指数增长速度要快得多。

机器如何学会下棋

游戏的好处就在于，其规则都有明确的定义，玩家对棋盘十分熟悉，决策也不像现实世界中那样复杂，但又不失挑战性。1959年，在商业数字计算机发展的早期，IBM的机器学习先驱亚瑟·塞缪尔（Arthur Samuel）编写了一个擅长玩国际跳棋的程序，在宣布其诞生的当天，IBM的股票就获得了巨大的收益。西洋跳棋则相对容易。塞缪尔的程序利用了代价函数来评估对局中不同布阵的优劣情况，这一点跟以前的游戏程序很相似。该程序是在IBM第一款真空管商用计算机

IBM 701上运行的，它在一个方面的创新令人印象深刻：通过跟自己对弈，学会了下棋。

在转到位于纽约约克敦海茨的IBM托马斯·J. 沃森研究中心之前，杰拉德·特索罗在位于伊利诺伊大学香槟分校的复杂系统研究中心，与我一起训练神经网络玩西洋双陆棋（见图10－1）。[2]我们的方法是，使用专家监督来训练反向传播网络，以评估当前的布局和可能的摆法。这种方法的缺陷在于，该程序永远比不过专家，而专家的水平并未达到世界冠军的级别。然而通过自我对局，网络可能会有更出色的表现。当时自我对局面临的问题是，在比赛结束时，唯一的学习信号就是赢或输。但是当一方获胜时，应该归功于之前若干步骤中的哪些步骤呢？这被称为“时域贡献度分配问题”（temporal credit assignment problem）。

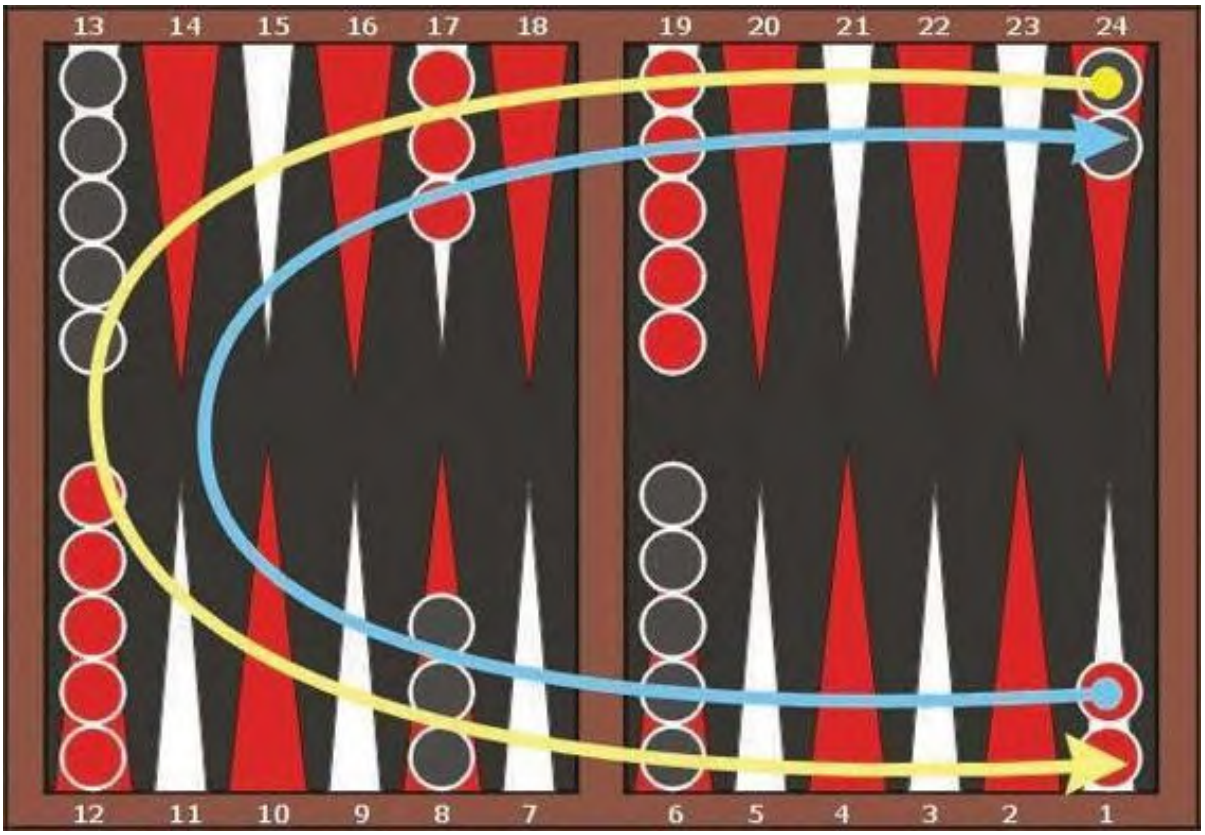


图10-1 西洋双陆棋棋板。西洋双陆棋是一种通过朝着终点按步走棋来分出胜负的游戏，红色棋子和黑色棋子的移动方向相反（如箭头所示）。图中标注了游戏的起始

位置。同时掷出两个骰子，得到的两个数字表示两个棋子可以向前移动的距离。

有一种可以解决这种时域贡献度分配问题的学习算法，是由理查德·萨顿（Richard Sutton）于 1988年发明的。[3]他当时正在与他的博士生导师、马萨诸塞大学阿姆赫斯特分校的安德鲁·巴托（Andrew Barto），共同解决强化学习领域中的问题。强化学习（reinforcement learning）是受动物实验中的关联学习（associative learning）所启发衍生出的一个机器学习研究分支领域（见图10-2）。深度学习的唯一工作是将输入转换为输出。强化网络与之不同，它会与环境进行闭环交互，接收传感器输入，做出决定并采取行动。强化学习的基础，是观察动物怎样通过探索环境中的各种选择并从结果中学习，从而在不确定的条件中解决难题。随着学习能力的提高，探索过程逐渐减少，最终会直接利用学习过程中发现的最佳策略。

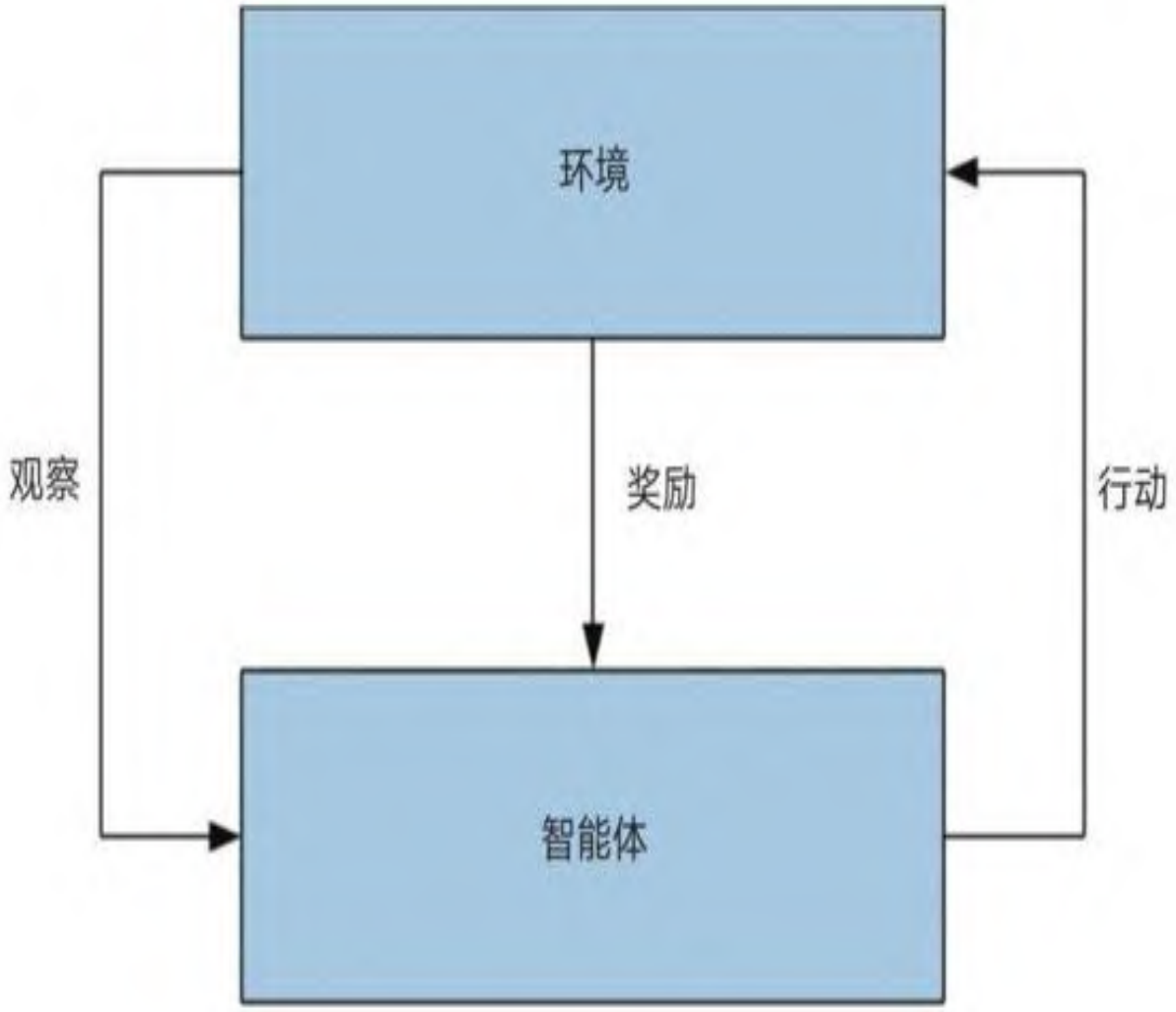


图10-2 强化学习场景。智能体（agent）通过采取行动（actions）和进行观察（observation）来积极探索环境。如果行动成功，执行器将得到奖励（rewards）。该过程的目标，是通过学习怎样采取行动来最大化可能获得的奖励。

假设你必须做出一系列决定才能达成目标。如果你已经知道所有潜在的选择和它们各自可能带来的奖励，你就可以使用搜索算法——具体来说，也就是理查德·贝尔曼（Richard Bellman）的动态规划算法（algorithm for dynamic programming），[4]即找出能最大化未来奖励的选择集。但是随着可能的选择越来越多，问题的规模也呈指数级增长，这被称为“维数灾难”（curse of dimensionality），本

章的开头已经对其进行了说明。但是，如果在选择前没有获得关于选择结果的所有信息，你就要学会即时做出最好的选择。这就是所谓的“在线学习”（online learning）。

理查德·萨顿（见图10-3）的在线学习算法依赖于期望奖励和实际奖励之间的差异（见方框10.1）。在时间差分学习（temporal difference learning）中，你需要估计出在当前状态下做出行动有可能带来的长期奖励（基于已得到的奖励而得出的较好估计），以及下一个状态中潜在的长期奖励，并将二者相比较。当前状态得到了实际奖励，因此估计会更准确。通过让之前的估计更接近改进后的估计，你做出的决定也会越来越好。存在一个价值网络，能够估算出棋盘每个位置上的未来奖励，对该网络进行的更新则被用于决定下一步的行动。在你有足够的时间来探索不同的可能性后，时间差分算法会收敛于最佳规则，指导如何在给定状态下做出决策。维数灾难是可以避免的，因为事实上，在棋盘所有可能的位置中有一小部分会被访问，但这足以为新对局中类似的棋盘位置制定出好的策略。



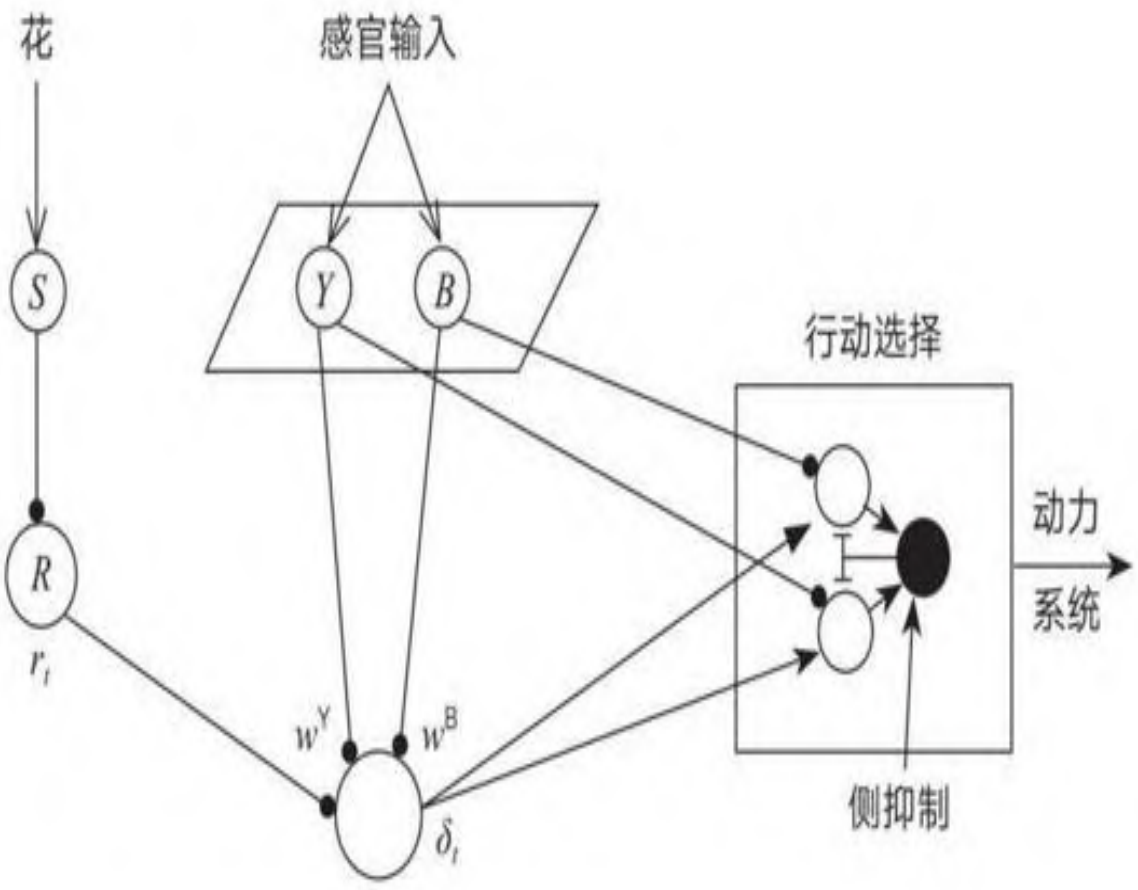
图10-3 2006年在加拿大埃德蒙顿阿尔伯塔大学的理查德·萨顿。他教会了我们获取未来奖励的学习方法。理查德是一位癌症幸存者，他在强化学习方面一直是领军人物，并在持续不断地开发创新型的算法。他总是很慷慨地和别人交流，分享自己的

见解，同领域的每个人对此都非常赞赏。他和安德鲁·巴托合著的书《强化学习导论》（*Reinforcement Learning: An Introduction*）是该领域的经典著作之一。此书的第二版在互联网上可以免费获取。图片来源：理查德·萨顿。

杰拉德·特索罗的程序名为“TD-Gammon”，内建了西洋双陆棋棋盘和规则的重要特征，但它并不知道怎么下好每一步棋。在学习的初始阶段，这些棋步是随机的，但最终某一方会赢，并得到最终奖励。西洋双陆棋的赢家是第一个将其所有棋子从棋盘上“剔除”出来的玩家。

10.1
时间差分学习

这个蜜蜂大脑的模型能够选择行动（比如落在一朵花上），以最大化未来所有的折扣奖励（discounted rewards）：



$R(t)=r_{t+1}+\gamma r_{t+2}+\gamma^2 r_{t+3}+\cdots$

r_{t+1}是在时间$t+1$时的奖励，$0 < \gamma < 1$是折扣系数。基于当前感官输入$s(t)$的预测的未来奖励是通过神经元P计算出来的：

$$Pt(s) = w^y s^y + w^b s^b$$

来自黄色 (Y) 花朵和蓝色 (B) 花朵的传感器输入被分别以w^y和w^b加权。位于时间t的奖励预测误差 $\delta(t)$ 计算方法如下：

$$\delta t = r_t + \gamma P t(st) - Pt(st-1)$$

r_t代表当前的奖励。每个权重的改变如下：

$$\delta w_t = \alpha \delta t st-1$$

α 代表学习速率。如果当前奖励大于预测奖励，且 δt是正值，奖励出现之前的感官输入的权重就会增加，但是如果当前奖励小于预测奖励，且 δt是负值，感官输入的权重就会减少。

图片改编自：Montague, P. R. and Sejnowski, T. J., *The Predictive Brain:Temporal Coincidence and Temporal Order in Synaptic Learning Mechanisms, figure6 A*。

由于唯一的实际奖励发生在游戏结束时，你可能会合理地推测TD-Gammon将首先学习游戏结局，然后是游戏的中间部分，最后是开局。这些实际上是发生在“表格强化学习”（tabular reinforcement learning）的过程，其状态空间中的每个状态都对应着一个数值表。但它与神经网络完全不同——神经网络快速锁定输入特征中的简单可靠信号，接着锁定后期较复杂、不太可靠的输入信号。TD-Gammon学习的第一个概念是“拿走棋子”，可以通过对表示拿走棋子数目的输入特征添加正的权重来实现。第二个概念是“击中对手棋子”——这是一个在所有阶段效果都相当不错的启发式做法，一部分输入单元对被击中的对手棋子的数目进行了编码，可以通过为这些单元加上正权重

进行学习。第三个概念“避免被击中”，是对第二个概念的一种自然反应，可以通过对可能被击中的单个棋子施加负权重来学习。第四个概念是通过“积累积分”来阻止对手前进，通过对积分输入赋予正权重来学习。理解这些基本概念需要完成几千场的训练棋局。通过一万场棋局，TD-Gammon会学习到中级概念；通过10万场棋局，它开始学习高级概念；经历了100万场棋局后，它已经学会了冠军级的概念，或者超越了20世纪90年代初人类玩家的水平。

杰拉德·特索罗于1992年向全世界展示了TD-Gammon，它的表现让我和其他同行都感到非常惊讶。[5]这个程序的价值函数（value function）是一个具有80个隐藏单元的反向传播网络。在30万场比赛之后，该程序击败了杰拉德，于是他联系了著名的西洋双陆棋世界冠军选手兼作家比尔·罗伯特（Bill Robertie），并邀请他来约克敦海茨的IBM跟TD-Gammon对局。罗伯特赢得了大部分的棋局，但是对在一些胜算较高的对局中落败不免感到惊讶，随后他宣布这是他对决过的最好的西洋双陆棋程序。TD-Gammon展现了一些他从未见过的棋路；经过一番仔细研究，这些棋路被证明可以在总体上提升人类玩家的水平。当该程序与自己对局达到150万次时，罗伯特回来了，而且非常惊讶于自己居然只跟TD-Gammon打了个平手。它的水平突飞猛进，罗伯特觉得它已经达到了人类冠军的水平。西洋双陆棋专家基特·伍尔西（Kit Woolsey）发现，当时TD-Gammon对于应该打安全牌（低风险/低奖励）还是走险着（高风险/高奖励）的判断，比他见过的任何人都要准确。虽然150万局的训练可能看起来已经很多了，但它只代表了所有10^{20}个可能的西洋双陆棋板位置的无限小部分；这要求TD-Gammon的几乎每一步棋都能推广到新的棋板位置。

TD-Gammon并没有像IBM的“深蓝”那样受到公众的关注，后者在1997年的国际象棋比赛中击败了加里·卡斯帕罗夫（Gary Kasparov）。国际象棋比西洋双陆棋要困难得多，卡斯帕罗夫当时是国际象棋的世界冠军。但在某些方面，TD-Gammon是一个令人印象更加

深刻的成就。首先，TD-Gammon使用模式识别技术来教会自己如何下棋，这与人类学习的风格十分相似，而“深蓝”则通过暴力算法（brute force）获得胜利，使用特制硬件，比任何人类棋手都能预见到更多可能的棋路。其次，TD-Gammon也很有创意，使用了人类从未见过的微妙策略和位置决策。这样一来，TD-Gammon也提高了人类的竞技水平。这一成就是人工智能历史上的一道分水岭，因为我们从一个人工智能程序中学习到了新东西。该程序教会了自己如何在一个人类熟稔的领域中掌握一种复杂的策略，这种新策略值得人们关注和思考。

大脑的奖励机制

TD-Gammon的核心是时间差分学习算法，它受到了动物学习实验的启发。几乎所有经过测试的物种，从蜜蜂到人类，都可以进行关联训练，就像巴甫洛夫的狗一样。在巴甫洛夫的实验中，诸如铃声之类的感官刺激之后，就会有食物出现，这会引起流涎反应。经过几次这样的配对之后，仅靠铃声本身就会导致流涎。不同物种在关联学习中对无条件刺激有不同的偏好。蜜蜂非常擅长将花的气味、颜色和形状与花蜜的奖励联系起来，并利用这种学习到的关联来找到当季盛开的相似品种的花。这种普遍的学习方式一定包含了什么重要信息。20世纪60年代有一段时期，心理学家们深入研究了引起关联学习的条件，并开发了解释它的模型。像斯金纳（B. F. Skinner）这样的行为主义学家曾训练鸽子识别出照片中的人类，这就让人联想起对深度学习的训练，但这其中有一个很大的区别。反向传播学习需要对输出层上的所有单元提供详细的反馈，但关联学习只提供单一的奖励信号，即正确或不正确。大脑必须弄清楚环境中的哪些特征能够帮助做出成功的抉择。

只有在奖励之前发生的刺激才被认为和奖励有关联。这是有道理的，相比奖励之后的刺激，奖励之前的刺激更有可能引发奖励。因果

关系是自然界的一个重要原则。相反的情况则是条件刺激之后伴随的惩罚，例如撞到脚这一后果，能教会动物在今后避免这类刺激。在某些情况下，条件刺激和惩罚之间的时间间隔可能会相当长。20世纪50年代，约翰·加西亚（John Garcia）表明，如果一只老鼠被喂了甜水，并且在几小时后感到恶心，那么它在接下来的几天都会避开甜水。这就是所谓的“味觉厌恶学习”（taste aversion learning），它也会发生在人类身上。[6]有时，恶心会被错误地关联到摄取的食物上，如巧克力。遗憾的是，巧克力只是与其他东西同时被食用，而不是引起恶心的原因；而由此产生的厌恶感可以持续多年，即使当事人已经理性地觉察到巧克力并不是问题的根源。

多巴胺（Dopamine）是脑干中一组由扩散投射神经元所携带的神经调节剂（见图10－4），长期以来一直被认为与奖励学习有关，但人们始终不清楚它传给皮层的信号是什么。20世纪90年代，我实验室的博士后研究员彼得·达扬（Peter Dayan）和瑞德·蒙塔古（Read Montague）意识到，多巴胺神经元可以实现时间差分学习。[7]这是我科研生涯中最让人兴奋的几个时期之一，这些模型及其预测得以发表，并随后被沃尔夫拉姆·舒尔茨（Wolfram Schultz）及其同事通过猴子的单神经元记录（见图10－5）[8]和人类脑成像[9]加以证实。现在已经确定，多巴胺神经元活动的瞬时变化传递了奖励预测误差信号。

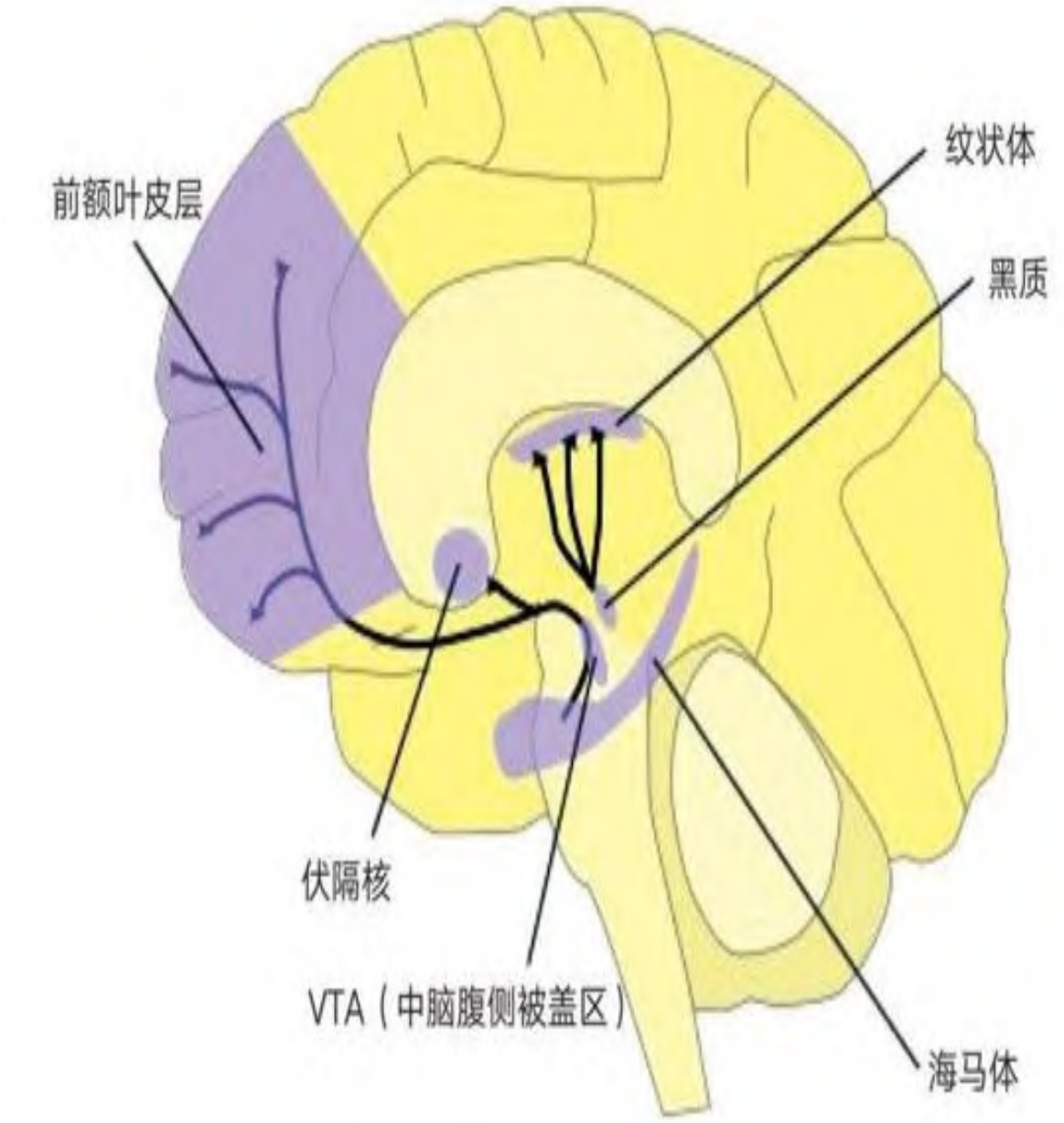
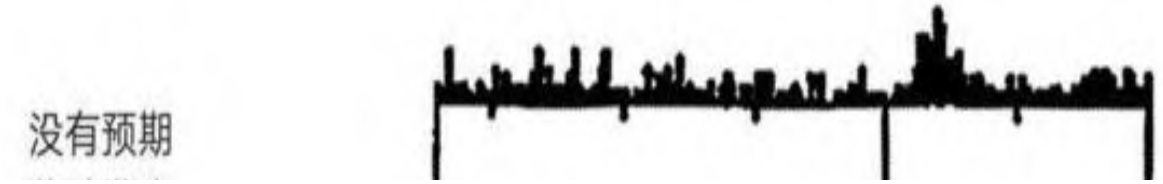


图10-4 人脑中的多巴胺神经元。位于中脑的几个核［VTA和黑质（substantia nigra）］将轴突投射进入皮层和基底神经节［纹状体（striatum）和伏隔核（nucleus accumbens）］。瞬态脉冲（Transient bursts）标志着奖励预测和获得奖励两者之间存在差异，可以被用于选择行动和修改预测。



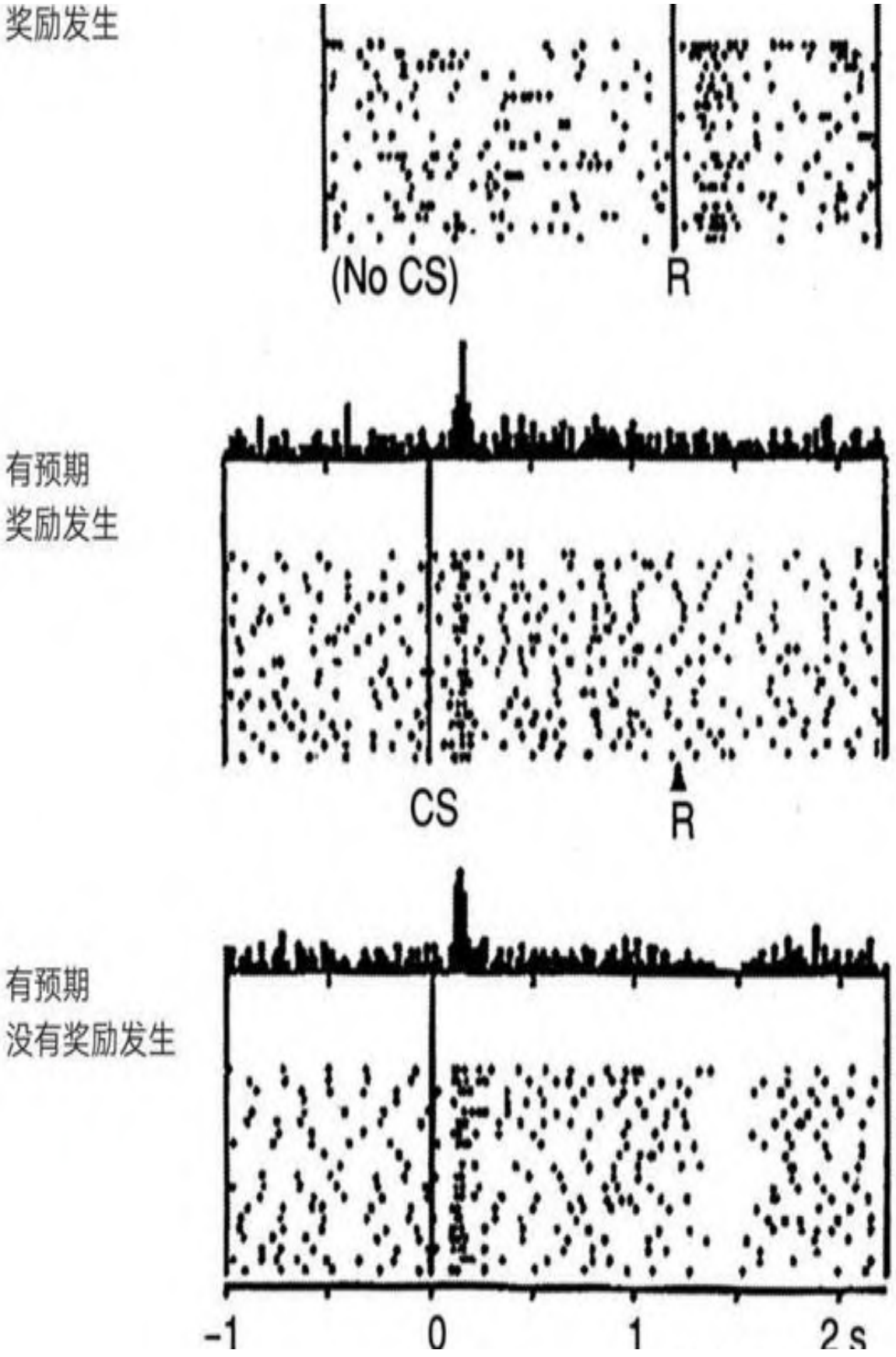

奖励发生
(No CS)
R
有预期
奖励发生
CS
R
有预期
没有奖励发生
-1
0
1
2 s



图10-5 猴脑中多巴胺神经元的反应，证明了它能够向大脑的其他部分发出奖励预测误差。每个点都是多巴胺神经元的一次脉冲。每一条线都是单一的学习试验。每个时间段中脉冲的数量显示在记录的顶部。（上图）在学习开始时，奖励（R）是意外得到的，多巴胺在得到奖励后激发了一段脉冲。（中图）经过多次试验，当光照（条件刺激，conditioned stimulus，CS）在奖励发放前持续闪现时，多巴胺细胞对CS做出了反应，但没有对奖励做出回应。根据时间差分学习，奖励后的响应被奖励前的预测抵消了。（下图）当奖励在试验中被扣留时，多巴胺神经元的激发在预期出现奖励的位置出现了回落。图片改编自：Schultz, Dayan, and Montague, “A Neural Substrate of Prediction and Reward,” 1594, figure 1。

1992年，我去柏林拜访了正在研究蜜蜂大脑快速学习的兰道夫·门泽尔（Randolph Menzel），当时我们在灵长类动物的奖励预测误差方面已经取得了一些进展。蜜蜂的学习能力在昆虫界是数一数二的。在访问一朵花几次并得到奖励后，蜜蜂就能记住这朵花。蜜蜂脑中有大约100万个小神经元，很难记录这些神经元的活动，因为它们非常小。门泽尔的小组发现了一种叫作“VUMmx1”的独特神经元，它对蔗糖有反应，但对气味没有反应。然而，如果先传递气味再提供蔗糖奖励，VUMmx1也会对气味做出反应。[10]时间差分学习的多巴胺模型在蜂脑中可能由单个神经元实现。VUMmx1释放了一种在化学上与多巴胺密切相关的神经调节剂——奥克巴胺（octopamine）。这种蜜蜂学习模式可以对蜜蜂心理学中一些微妙的方面做出解释，比如风险规避。[11]如果让蜜蜂在“定时定量的奖励”和“在一半的时间内获得双倍奖励”之间进行选择，蜜蜂会始终选择前者，尽管奖励的平均值相同。[12]多巴胺神经元也存在于苍蝇体内，并且已被证明包含几条用于短期和长期关联记忆的并行强化学习途径。[13]

用“感知—行动”框架提高绩效

多巴胺神经元构成了控制大脑中动机的核心系统，所有成瘾药物都是通过增加多巴胺的分泌水平起作用。当死亡的多巴胺神经元达到一定数量时，人体就会出现帕金森病的症状、包括运动性震颤，运动迟缓，后期则完全丧失任何活动的快感，即“快感缺失”（anhedonia），最终导致活动和反应能力的彻底缺失，即“紧张性抑郁障碍”（catatonia）。当意外奖励发生时，行为正常的多巴胺细胞会向皮层和其他脑部区域短暂释放多巴胺。当实际奖励低于期望时，多巴胺的释放量会减少。这正是时间差分算法的特征（见图10－5）。

当我们需要做出决定时，都会询问多巴胺神经元。我们应该从菜单中点些什么？当我们想象每个菜品时，多巴胺细胞就会提供对预期奖励的估计。我应该和这个人结婚吗？我们的多巴胺细胞会给我们一个比理性分析更值得信赖的“直觉”建议。最难以决定的则是带有许多不相称维度的问题。在选择配偶时，要如何平衡幽默感和邋遢的生活习惯，或者在正面和负面的特质之间做出数以百计的其他权衡。我们的奖励系统将所有这些维度降低到了一个“通用货币”的范畴，即短暂的多巴胺信号。这种“通用货币”的经济力量在我们发现它之前很长一段时间，就已经被大自然所利用了。

时间差分学习算法中存在两个参数：学习速率α和折扣因子γ（见方框10.1）。某些昆虫具有很高的学习速率，比如蜜蜂，在一次访问后就可以学会将花与奖励联系起来。但哺乳动物的学习速率较低，往往要尝试多次。折扣因子（discount factor）也在很大的范围内变化。当γ= 0时，学习算法是贪婪的，仅仅基于是否能立刻获得奖励做出决定；但是当γ= 1时，所有未来奖励都具有相等的权重。在一个经典的实验中，被测试的幼童可以选择是立刻吃掉手里的一颗棉花糖，还是等待15分钟再吃，到时候就能再得到一颗棉花糖。[14]年龄是一个强有力的预测因子，年龄较小的孩子无法延迟满足感。如果认为有必要，我们可以在短期内做出带来负面回报的选择，以交换在遥远的将来所期望获得的更丰厚的奖励。

多巴胺神经元接受来自大脑中被称为“基底神经节”部分的输入（见图10－4），众所周知，这对于顺序学习和习惯行为的形成是很重要的。基底神经节纹状体中的神经元接受来自整个大脑皮层的输入。来自皮叶层后半部分的输入，对于学习动作的顺序以实现某个目标而言十分重要。前额叶皮层对基底神经节的输入更多的是与行动的计划顺序有关。从皮层到基底神经节再返回的循环需要100毫秒，每秒循环信息10次。这就允许通过一系列快速决策来实现目标。基底神经节的神经元也会评估皮层状态，并为它们分配一个值。

基底神经节执行的是杰拉德·特索罗在TD-Gammon中训练的价值函数的高级版本，后者被用来预测棋盘位置的价值。在第1章中描述的由DeepMind公司开发的AlphaGo所取得的惊人成就，即具备围棋世界冠军级别的能力，是基于与TD-Gammon相同的体系结构，但前者是后者的加强版。TD-Gammon价值网络中的一层隐藏单元在AlphaGo中变成了十几层，并经历了数百万次对决，但基本的算法是一样的。这是对神经网络学习算法出色的缩放性的生动演示。如果我们继续增加网络规模和训练时间，性能还会获得多大程度的提升呢？

棋类游戏的规则比现实世界要简单得多。电子游戏世界为迈向更复杂和不确定的环境提供了一块垫脚石。DeepMind在2015年已经表明，时间差分学习能够以屏幕像素为输入，学习如何在超人的水平上玩像*Pong*这样的雅达利（Atari）街机游戏。[15]下一个垫脚石是三维环境中的视频游戏。《星际争霸》（*Star Craft*）是有史以来最具竞争力的视频游戏之一，DeepMind正在使用它来开发可在该世界纵横捭阖的自主深度学习网络。微软研究院最近购买了另一种流行的视频游戏《我的世界》（*Minecraft*）的版权，并开放了它的源代码，以便其他人可以对它的三维环境进行个性化修改，从而加快其人工智能的发展进度。

能够像冠军那样下西洋双陆棋和围棋，是一项了不起的成就，玩电子游戏是下一个重要的进步，那么解决现实世界的问题呢？感知—

行动周期（见图10－2）可用于解决任何基于感官数据来计划行动的问题。行为的结果可以与预测的结果进行比较，其差值随后被用于更新进行预测的系统的状态；对先前条件的记忆可以被用来优化资源的使用，并预测潜在的问题。

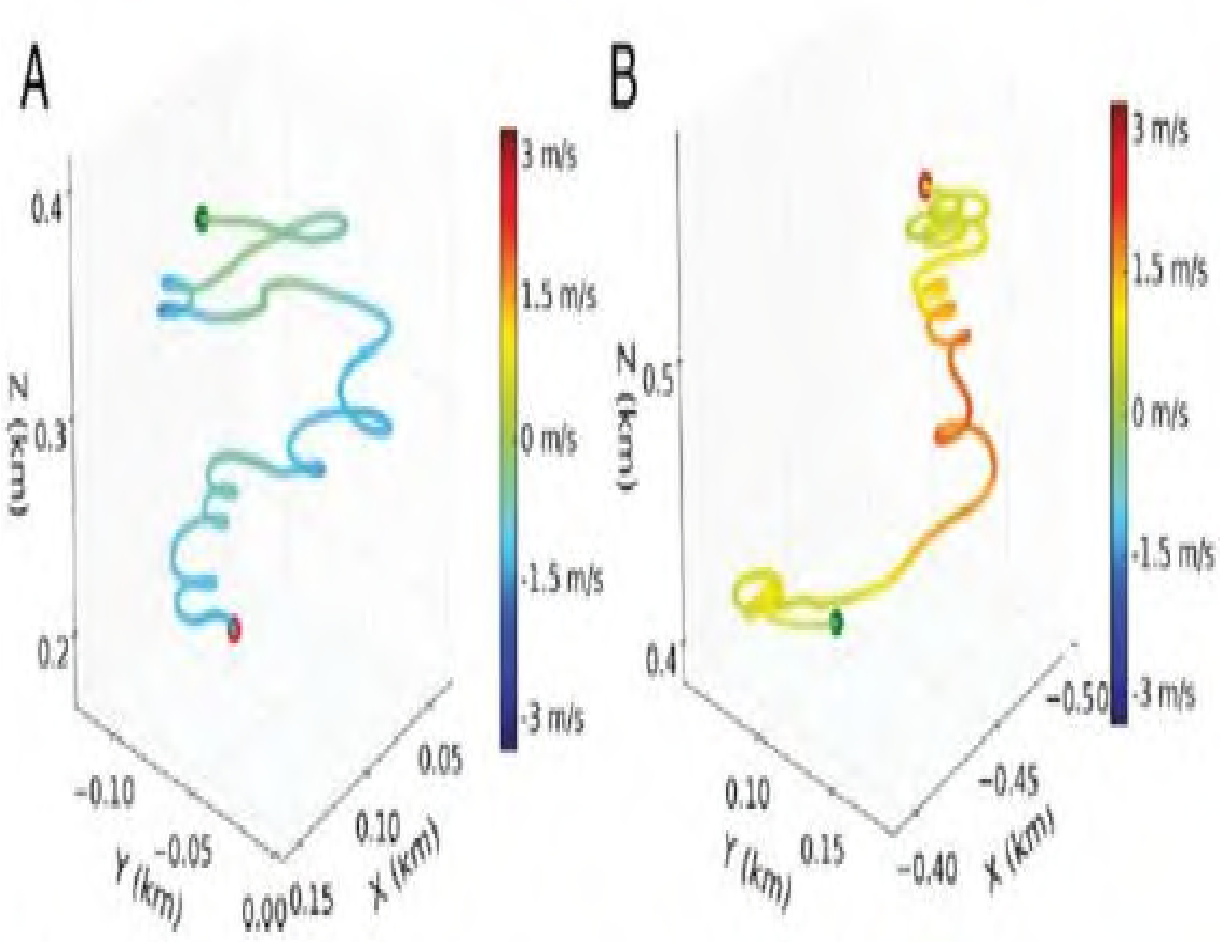
加拿大安大略省汉密尔顿市麦克马斯特大学（McMaster University）的西蒙·赫金（Simon Haykin），曾经使用这个框架来提高几个重要的工程软件系统的性能，[16]包括：认知无线电，可以动态分配通信渠道；认知雷达，能够动态地移动频带以减少干扰；还有认知电网，可以动态平衡电网电力负载。风险控制也可以在同一个“感知—行动”框架内进行管理。[17]通过在每个软件系统中使用“感知—行动”框架来改进它们所管理的领域，可以显著地提高绩效并降低成本。

学习如何翱翔

2016年，加州大学圣迭戈分校的物理学家马西莫·维加索拉（Massimo Vergassola）和我都在思考，是否可以利用时间差分学习，让滑翔机在高空翱翔几小时，而且像许多鸟类那样不用消耗太多能量。[18]热膨胀空气能够将鸟类带到很高的高度，但是在热空气中，空气是湍流的，同时存在向下和向上的气流。面对如此多的冲击，鸟类是如何保持它们向上轨迹的，我们对此还没有一个清晰的认识。我们的第一步是对湍流对流过程进行一个逼真的物理模拟，并创建一个滑翔机空气动力学模型。下一步就是模拟滑翔机在湍流中的轨迹。

一开始，滑翔机还无法利用上升的空气柱，表现为向下滑行（见图10－6）。在将“向上”标记为奖励后，滑翔机开始学习一种策略，经过几百次试验后，它的飞行轨迹跟鸟类翱翔时所形成的紧凑的环状

路径就比较相似了。滑翔机还学习了应对不同程度湍流的不同策略。通过分析这些策略，我们可以提出假设，并参考鸟类翱翔是否的确使用到了这些策略。然后，我们装备了一架翼展有6英尺宽的滑翔机，教它翱翔并保持高度。[19]



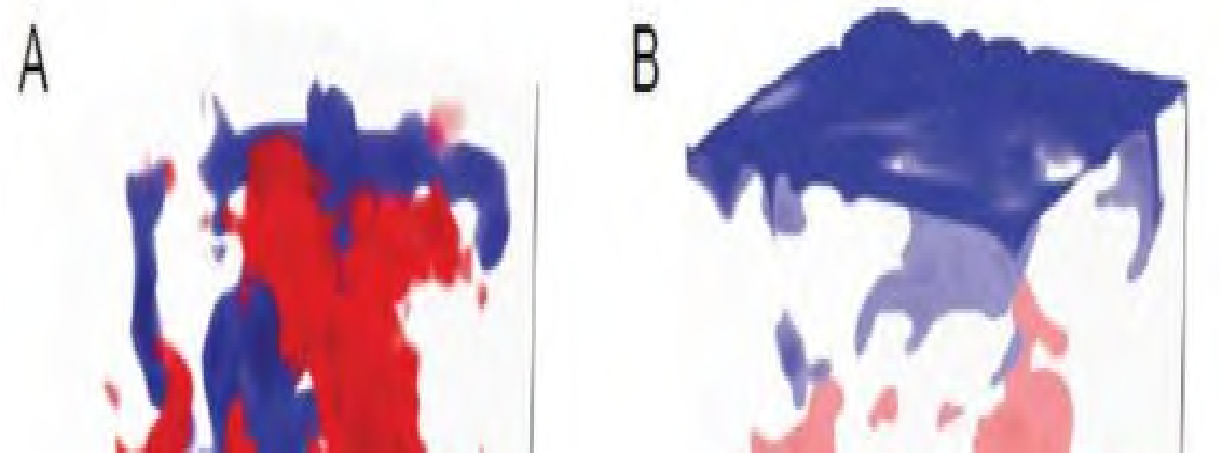




图10-6 滑翔机学习使用热气流翱翔的模拟。（下图）在我们的瑞利-贝纳尔对流（Rayleigh-Bénard convection）三维数值模拟中垂直速度场（A）和温度场（B）的截图。垂直速度场中，红色和蓝色分别表示上升气流和下降气流的区域。温度场中，红色和蓝色分别表示高温和低温区域。（上图）未经训练的滑翔机（A）和经过训练的滑翔机（B）在瑞利-贝纳尔湍流中飞行的典型轨迹。颜色表示滑翔机经历的垂直风速。绿色和红色圆点分别表示轨迹的起点和终点。未经训练的滑翔机随机做出决定并下降飞行高度，而训练过的滑翔机在强上升气流区域以特征的螺旋轨迹飞行，如鸟类和滑翔机在热气流中飞行时所观察到的那样。图片来源：G. Reddy, A.Celani, T. J. Sejnowski, and M. Vergassola, “Learning to Soar in Turbulent Environments,” top: figure 2; bottom: figure 11。

学习如何歌唱

能够展现强化学习能力的其他例子，还包括鸟类如何学习唱歌，以及孩子如何学习说话。在这两种情况下，听觉学习的初始阶段结束之后，就进入了渐进式运动性学习的后期阶段。斑胸草雀在刚出生不久就听到了它们父亲的歌声，但要几个月后才能发出自己的声音。即使当它们在运动学习阶段与父亲分离时，会经历一段听起来很刺耳的初鸣期，但这种声音会不断地改善，最终按照它们父亲特有的音调发出鸟鸣。斑胸草雀可以通过同物种的歌声判断它来自森林的何处，就像我们能够从一个人的口音得知那个人来自哪里一样。推动鸟鸣研究的假设是，在听觉学习阶段，鸟类先学习模板，然后用模板来改进运动系统在运动性学习阶段产生的声音。我们知道强化学习发生在基底神经节，而负责人类和鸣禽运动学习阶段的通路也位于那里。

1995年，我实验室的博士后铜谷贤治（Kenji Doya）开发了“鸟鸣运动细化”的强化学习模型（见图10－7）。该模型通过调整运动通路中的突触，使之与鸟类的发声器官（即鸣管，syrinx）模型相匹配，从而提高了它的性能，然后测试新歌是否比前一首歌更符合这个模板。如果是的话，这些变化就被保留，但如果新的歌曲匹配效果较差，那么突触就会衰减回原来的强度。[20]我们预测，在产生音节序列的运动回路的顶端，应该只活跃在歌曲单个音节上的神经元，以便更容易地对每个音节分别进行调整。后来，麻省理工学院的米歇尔·菲（Michale Fee）实验室和其他鸟类实验室的研究结果证实了这一点，同时也证实了该模型中的其他关键预测。

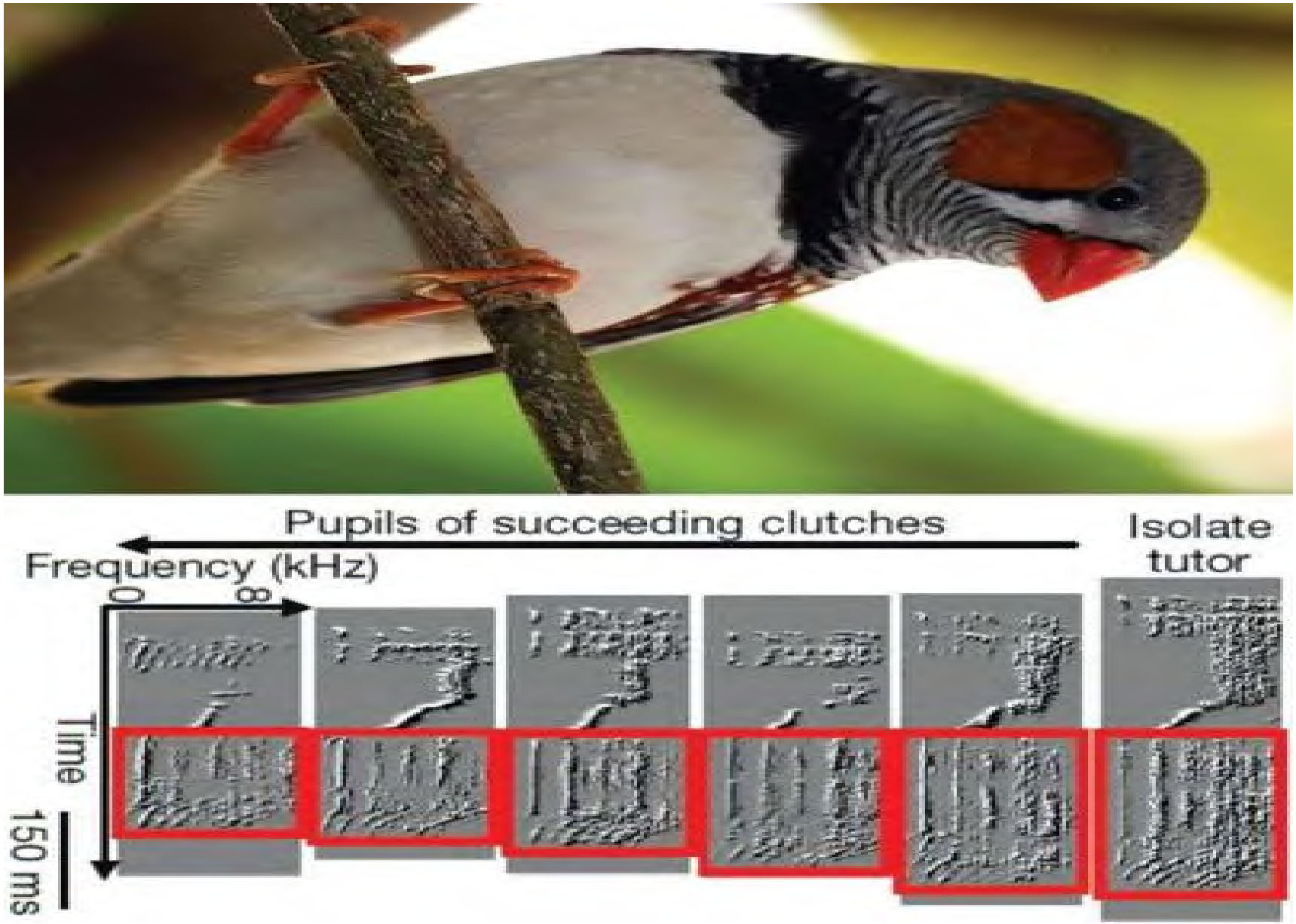

Pupils of succeeding clutches
Isolate tutor
Frequency (kHz)
0
8
Time
150 ms

图10-7 斑胸草雀的鸟鸣。在图右侧的频谱图中，父亲的歌唱（导师，上图）教给了儿子（学生，上数第二张图），于是这种鸟鸣的声音特征代代相传。注意频谱图（时域上的频谱功率函数）中基调（红色轮廓框）的相似性。基调随着代际的增长而变短。左图来源：http://bird-photoo.blogspot.com/2012/11/zebrafinch-bird-pictures.html；右图来源：Olga Feher Haibin Wang, Sigal Saar, Partha P. Mitra, and Ofer Tchernichovski,“De novo Establishment of Wild-Type Song Culture in the Zebra Finch,” figuer 4。

在加州大学旧金山分校研究鸟鸣学习的艾莉森·杜普（Allison Doupe），以及在西雅图华盛顿大学研究婴儿语言系统发育的帕特里夏·库尔（Patricia Kuhl），都证明了鸣禽学习和幼儿语言学习方面有很多相似之处。[21]鸟儿的音节和婴儿的音素首先是作为声音被习得的（即听觉学习），而运动学习随后才以鸟类的初鸣和婴儿的咿呀学语开始。在大脑中有许多因领域而异的学习和记忆系统，这些系统必须协同工作以获取新技能。鸟类学习鸟鸣的强化学习算法以及在猴子、人类、蜜蜂等奖励系统中的时间差分学习算法，只是其中的两个例子。

人工智能的可塑性

尽管人们在视觉和听觉等认知功能的自动化方面取得了进展，但人工智能还需要在人类智能的许多其他方面取得进步。皮层中的表征学习与基底神经节中的强化学习相辅相成。人工智能学习能把精通围棋的能力运用到解决其他复杂问题上吗？大部分人类学习都是基于观察和模仿，对于学习识别新对象，我们需要的样本比深度学习要少得多。未标记的感官数据非常丰富，强大的无监督学习算法可以在进行任何监督之前使用这些数据学到一些东西。在第7章中，一个无监督版本的玻尔兹曼学习算法被用来初始化深度学习网络。在第6章中，独立分量分析——一种无监督学习算法，能够从自然图像中提取稀疏编码（sparse population codes）。在第9章，生成对抗网络——一个无

监督的学习系统，可以创建逼真的图片。无监督学习是机器学习的下一个前沿领域。我们对大脑式计算的理解才刚刚拉开序幕。

大脑的许多学习系统和形式多样的可塑性可以协同工作。仅在大脑皮层内就有几十种可塑性，例如神经元兴奋性和增益的可塑性。突触可塑性一个特别重要的形式是稳态（homeostatic），它确保神经元在最佳动态范围内活动。当突触强度降低到零或达到极限时，会发生什么呢？这可能导致神经元永远不会获得足够的输入来达到阈值，或者接收太多的输入并始终保持高水平的活动。吉娜·特里吉亚诺（Gina Turrigiano）在大脑中发现了一种新的突触可塑性形式，它将神经元上的所有突触进行了归一化，以维持神经元活动的平衡。[22] 如果平均活动速率过高，所有兴奋性突触强度都会减小；如果速率太低，所有突触强度都会增加。对于抑制性输入，情况则相反，如果活动速率太高，突触强度会增加，速率太低则减小。类似的归一化形式已经被证明能够有效地模拟大脑中神经映射的发展。[23] 由随机梯度下降（stochastic gradient descent）驱动的人工神经网络能够从稳态缩放（homeostatic scaling）中受益。

大脑在其神经元细胞膜上具有数十个电压敏感的配体门控离子通道，用以调节兴奋性和信号传导。神经元的树突、体细胞和轴突内部一定有某种基于局部活动模式的机制，来动态调节这些通道的位置和密度。有人曾提出过几种算法来实现这些机制。[24] 目前，我们对这些稳态形式的理解并没有达到突触稳态可塑性那样的深度。

更多需要被解决的问题

在蒙特利尔举行的2015年NIPS大会的“Brains，Minds and Machines”（大脑、思维和机器）座谈会，以及2016年巴塞罗那NIPS大会的“Bits and Brains”（比特与大脑）研讨会期间，戴米斯·哈

萨比斯和我在关于人工智能未来，以及下一个优先研究重点的问题上进行了激烈的辩论。人工智能中还有许多开放的问题需要解决。最重要的就是因果关系问题，它提供了最高层次的人类推理，以及行动的意向性问题，这两者都预示着一种精神理论。我之前提到，我们所创造的深度学习系统都不能依靠自己独立生存。这些系统的自主性只有在它们包含了迄今为止一直被忽视的，类似于大脑其他部分的功能时才有可能实现，例如管理摄食、繁殖、调节激素、稳定内脏的下丘脑，以及帮助我们根据运动预测误差来调整运动的小脑。这些都是在所有脊椎动物中发现的古老结构，对生存起着至关重要的作用。

哈瓦·西格尔曼（Hava Siegelmann）是马萨诸塞大学阿默斯特分校的计算机科学家，她表明模拟计算（analog computing）是超图灵（super-Turing），也就是说，拥有能够超越数字计算机的计算能力。[25]可以根据环境进行调整和学习的神经网络也有超图灵计算能力，而普通网络从训练集中学习，然后其结构就被固定下来，在操作时不再从它们的实际经验中学习，这一点和图灵机是一样的。但是，我们的大脑必须持续适应不断变化的条件，这就使我们具备了超图灵能力。我们如何做到这一点并同时拥有以前的知识和技能，是一个尚未解决的问题。哈瓦是DARPA终身学习项目的项目经理。她的“终身学习计划”正在资助一些高级研究，这些研究旨在为自治系统中的终身学习创建新的集成架构。

---

[1] 虽然我们这位传奇发明家命运未卜，但是一旦国王意识到自己被欺耍了，他很可能会因为自己的无礼而被发落。

[2] Tesauro and Sejnowski, “A Parallel Network That Learns to Play Backgammon.”

[3] R. Sutton, “Learning to Predict by the Methods of Temporal Differences,” *MachineLearning*3, no. 1 (1988): 9–44.

[4] 参阅 Richard Bellman *Adaptive Control Processes: A Guided Tour*(Princeton:Princeton University Press. 1961), 51–59.

[5] G. Tesauro, “Temporal Difference Learning and TD-Gammon.” *Communications ofthe ACM*38, no. 3 (1995): 58–68.

[6] J. Garcia, D. J. Kimeldorf, and R. A. Koelling, “Conditioned Aversion to Saccharin Resulting from Exposure to Gamma Radiation,” *Science*122. no. 3160 (1955):157– 158.

[7] P. R. Montague, P. Dayan, and T. J. Sejnowski, “A Framework for Mesencephalic Dopamine Systems Based on Predictive Hebbian Learning,” *Journal ofNeuroscience*16, no. 5 (1996): 1936–1947.

[8] W. Schultz, P. Dayan, and P. R. Montague, “A Neural Substrate of Prediction and Reward,” *Science*275, no. 5306 (1997): 1593–1599.

[9] P. N. Tobler, J. P. O’Doherty, R. J. Dolan, and W. Schultz, “Human Neural Learning Depends on Reward Prediction Errors in the Blocking Paradigm,” *Journal of Neurophysiology*95, no. 1 (2006): 301–310.

[10] M. Hammer and R. Menzel, “Learning and Memory in the Honeybee,” *Journal ofNeuroscience*15, no. 3 (1995): 1617–1630.

[11] L. A. Real, “Animal Choice Behavior and the Evolution of Cognitive Architecture,” *Science*253, no. 5023 (1991): 980–986.

[12] P. R. Montague, P. Dayan, C. Person, and T. J. Sejnowski, “Bee Foraging in Uncertain Environments Using Predictive Hebbian Learning,” *Nature*377, no.6551 (1995): 725–728.

[13] Y. Aso and G. M. Rubin, “Dopaminergic Neurons Write And Update Memories With Cell-Type-Specific Rules,” in L. Luo (ed.), *eLife*. 5 (2016): e 16135.doi:10.7554/eLife.16135.

[14] W. Mischel and E. B. Ebbesen, “Attention in Delay of Gratification,” *Journal ofPersonality and Social Psychology*16, no. 2 (1970): 329–337.

[15] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, et al.,“Human-Level Control through Deep Reinforcement Learning,” *Nature*518, no.7540 (2015): 529–533.

[16] Simon Haykin, *Cognitive Dynamic System: Perception-Action Cycle, Radar, and Radio*(New York: Cambridge University Press, 2012).

[17] S. Haykin, J. M. Fuster, D. Findlay, and S. Feng, “Cognitive Risk Control for Physical Systems,” *IEEE Access*5 (2017): 14664–14679.

[18] G. Reddy, A. Celani, T. J. Sejnowski, and M. Vergassola, “Learning to Soar in Turbulent Environments,” *Proceedings of the National Academy of Sciences of the United States of America*113, no. 33 (2016): E4877–E4884.

[19] G. Reddy, J. W. Ng, A. Celani, T. J. Sejnowski, and M. Vergassola, “Soaring Like a Bird via Reinforcement Learning in the Field,” submitted for publication.

[20] Kenji Doya and Terrence J. Sejnowski, “A Novel Reinforcement Model of Birdsong Vocalization Learning,” in Gerald Tesauro, David S. Touretzky, and Todd K.Leen, eds., *Advances in Neural Information Processing Systems 7*(Cambridge,MA: MIT Press, 1995), 101–108.

[21] A. J. Doupe and P. K. Kuhl, “Birdsong and Human Speech: Common Themes and Mechanisms,” *Annual Review of Neuroscience*22 (1999): 567–631.

[22] G. Turrigiano, “Too Many Cooks? Intrinsic and Synaptic Homeostatic Mechanisms in Cortical Circuit Refinement,” *Annual Review of Neuroscience*34 (2011):89–103.

[23] L. Wiskott and T. J. Sejnowski, “Constrained Optimization for Neural Map Formation: A Unifying Framework for Weight Growth and Normalization,” *NeuralComputation*10, no. 3 (1998): 671–716.

[24] A. J. Bell, “Self-Organization in Real Neurons: Anti-Hebb in ‘Channel Space’?”*Advances in Neural Information Processing Systems*4 (1991): 59–66; M. Siegel,E. Marder, and L. F. Abbott, “Activity-Dependent Current Distributions in Model Neurons,” *Proceedings of the National Academy of Sciences of the United States of America*91, no. 24 (1994): 11308–11312.

[25] *H. T. Siegelmann, “Computation Beyond the Turing Limit,” Science*238 (1995):632–637.

11
火爆的NIPS

要追寻科学思想的源头并非易事，因为科学是分布在不同时空中很多个体活动的集合。神经信息处理系统大会（NIPS大会，见图11-1）是本书写作的线索。到目前为止，很明显，这一会议不仅对我，也对整个科学领域产生了重大影响。[1]当时还没成为我妻子的比阿特丽斯·哥伦布，在1990年的NIPS大会上做了关于SEXNET的演讲。在结婚后不久的一次NIPS大会上，我们还差点分手。NIPS大会让人着迷，白天都是一些会议的常规环节，墙报展示环节设在晚上，直到后半夜各个会场仍然人声鼎沸。有次我在凌晨3点参加完一个环节的活动，回到房间没找到比阿特丽斯，当时我意识到自己麻烦大了。幸运的是，28年后，我们还在一起。

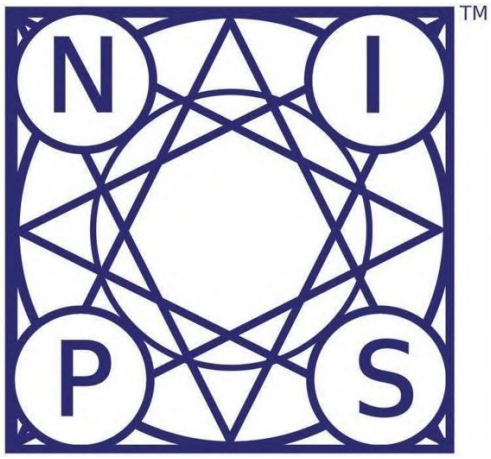


图11-1 NIPS大会的标志。30年前成立的NIPS大会是机器学习和深度学习领域的顶级会议。图片来源：NIPS基金会。

为什么NIPS如此受欢迎

深度学习拥有很长的历史，可以追溯到NIPS年度大会和研讨会，以及这些会议的前身。20世纪80年代，来自世界各地的工程师、物理学家、数学家、心理学家和神经科学家在NIPS大会上齐聚一堂，探讨共同构建人工智能的新方法。物理学家分析神经网络模型，心理学家模拟人类认知，神经科学家模拟神经系统并分析神经记录，统计学家探索高维空间中的大数据集，工程师则负责构建具备类人的视觉和听觉的设备。人工智能就以这样的方式飞速发展起来。

1987年在美国丹佛技术中心举办的第一届NIPS大会，聚集了400位与会者。学术会议通常侧重于狭窄的研究领域，因为每个人都会说同样的术语，彼此间能顺畅地交流。但是早期NIPS大会的科学多样性确实令人惊叹。生物学家在面对其他生物学家做演讲时，会用生物领域的术语。[2]数学家和物理学家交流的情况更糟糕，他们只会用方程式说话。工程师们会好一些，因为他们构建的东西不言自明。由于这些文化障碍，跨学科研究虽然被普遍寄予厚望，但事实上却很难实现。在早期的NIPS大会上，好像每个人都在自说自话。

1987年NIPS大会的主要会议结束之后，与会者在附近的滑雪胜地Keystone又聚集到一起开了个研讨会，并自行组织了小组会议。在一个不那么正式的环境里，学科之间的真正沟通开始了。我清楚地记得，我们在Keystone的热水浴缸里泡澡的时候，一位神经学博士建议当场开展一个关于海兔的研讨会。[3]当时我旁边那位来自美国国防部的绅士，很可能在琢磨海兔与国家安全有什么关系。然而今天，NIPS研讨会都是带有墙报展示环节的小型会议，其中一些研讨会吸引了数千名与会者。

是什么让NIPS长期受到追捧呢？首先，是让人激动的氛围，因为我们正处于用受生物学启发的学习算法来解决复杂计算问题的前沿。其次，就是因为爱德华·波斯纳（见图11－2），他是加州理工学院的

信息理论家，也是喷气推进实验室（Jet Propulsion Lab）的首席技术专家。他对这个领域有着长远的眼光，并建立了NIPS基金会对大会进行管理。

一个组织的文化往往能反映其创始人的特质。爱德华赋予了NIPS睿智、机敏和幽默感的独特组合。他是一位擅长启发别人的老师，同时也是卓有成效的领导者。他因为支持暑期大学生研究奖学金项目（Summer Undergraduate Research Fellowships，简称SURF）而在加州理工学院备受爱戴，该项目被称为“加州理工皇冠上的宝石”之一。爱德华招募了菲尔·索特尔（Phil Sotel）作为公益法律顾问。数十年间，会议规模和复杂程度都在不断增长。在索特尔的努力下，许多随之而来的问题都妥善地得到了解决，因此大会得以每年顺利举行。

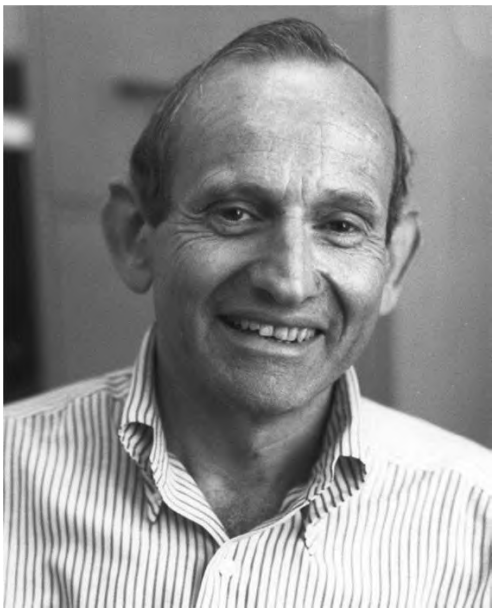


图11-2 加州理工学院的爱德华·波斯纳，NIPS大会的发起人。该大会在举办了30年后仍然颇受欢迎，这跟爱德华的远见卓识密不可分。图片来源：加州理工学院。

爱德华在我妻子还很年轻的时候就认识她了，并且通过NIPS单独结识了我。所以当我在一次NIPS大会中突然告诉他，比阿特丽斯和我已经订婚（engaged）了的时候，他回答说：“致力于（engaged）什

么？”[4]爱德华于1993年在一次自行车事故中不幸身亡，随后我接替他成为NIPS基金会主席，保证大会的继续发展和繁荣。我们在每年的NIPS大会上都会举办爱德华·波斯纳讲座（Ed Posner Lecture），以表示对他的缅怀和敬意。NIPS大会的特邀演讲嘉宾通常都不是NIPS主流领域内的从业者，但波斯纳讲座主要的演讲者都是来自我们这个群体，并对该领域做出重大贡献的成员。

NIPS大会的主席团由一群杰出的科学家和工程师组成。在此仅列举几位。斯科特·柯克帕特里克是一位物理学家（在第7章中已经介绍过），他发明了一种让计算机先“加热”再慢慢“冷却”，来解决计算难题的“模拟退火算法”。塞巴斯蒂安·特隆是一位计算机科学家（在第1章中介绍过），他赢得了2005年DARPA自动驾驶挑战大赛，为今天的自动驾驶汽车打开了大门。达芙妮·科勒（Daphne Koller）是一位计算机科学家，也是Coursera（将在第12章介绍）的联合创始人，开创了慕课领域的先河。

大数据促成了深度学习的腾飞。不久前，一太字节（terabyte，即万亿字节）的数据还需要占用整个机架；而现在，在单个记忆棒上就可以存储一太字节的数据。互联网公司的数据中心存储的信息以拍字节（petabyte）计量，每拍字节包含1024太字节（1拍字节=10^{15}字节）。自20世纪80年代以来，世界上的数据量每三年翻一番。每天都有数千拍字节的数据被上传到互联网上，其总容量达到了泽字节（zettabyte，即100万拍字节，或10^{21}字节）。大数据爆炸的影响不仅体现在科学和工程领域，也延伸到了社会生活的方方面面。如果没有互联网上的数百万张图像和其他标签数据，就无法训练真正的大型深度学习网络。

世界各地的大学都在为数据科学建立新的中心、研究机构和科系。艾利克斯·萨雷（Alex Szalay）自 1998年以来，就在斯隆数字巡天项目（Sloan Digital Sky Survey，下文简称SDSS；http://www.sdss.org/）中收集天文数据。2009年，凭借自己的这份

经历，他在约翰·霍普金斯大学建立了数据密集工程和科学研究所（Institute for Data Intensive Engineering and Science，简称IDIES）。SDSS使得天文学家收集的数据总量成千倍地增长，是当今世界上被使用次数最多的天文学设施。然而，正在建设中的大型综合性天空巡天望远镜（Large Synoptic Sky Survey Telescope，https://www.lsst.org/）收集的拍字节规模的数据集，将超越斯隆数字巡天项目收集的太字节规模的数据集达1000倍之多。2013年，当杨立昆在纽约大学创立数据科学中心时，每个科系都有老师带着手中的数据去拜访他。2018年，加州大学圣迭戈分校建立了新的Halıcıoğlu数据科学研究所。数据科学硕士学位（Master’s in Data Science degrees，简称MDSs）正变得像工商管理硕士（MBA）一样受人追捧。

谁拥有最多数据，谁就是赢家

2012年在太浩湖举行的NIPS大会上，深度学习领域的研究已经日趋成熟（见图11－3）。在这次大会上，早期的神经网络先驱杰弗里·辛顿和他的学生们发表了一篇论文，文中指出，多层神经网络在识别图像中的对象方面性能非常出色。[5]这些网络不仅比现有计算机的视觉技术更好，它们甚至已经处于一种完全不同的更高层次中，更加接近人类的表现水准。《纽约时报》刊登了一篇关于深度学习的文章，同时Facebook宣布与另一位深度学习先驱杨立昆合作成立一个新的人工智能实验室，由后者担任创始实验室主任。

Facebook首席执行官马克·扎克伯格（Mark Zuckerberg）参与了2012年的NIPS深度学习研讨会。当时安保成了令人头痛的问题，但也吸引了更多的与会者，我们不得不分流了一部分人到另外一个播放会场同步视频的房间。在之后的招待会上，我被介绍给了扎克伯格，他问了些关于大脑的问题。他对心智理论特别感兴趣。在心理学中，有一个关于心智如何工作的内隐理论，我们以它为指导来理解他人。我

们给朋友发短信时，并没有意识到在打字时大脑做出的许多决定。扎克伯格问了很多问题。“我的大脑如何构建我自己的心理模型？”“我的大脑如何根据经验来创建其他人的心理模型？”“我的大脑如何预测其他人的未来行为？”“其他物种有没有心智理论？”我最近在索尔克研究所合作组织了一个关于心智理论的研讨会，扎克伯格想要研讨会上所有的参考资料。

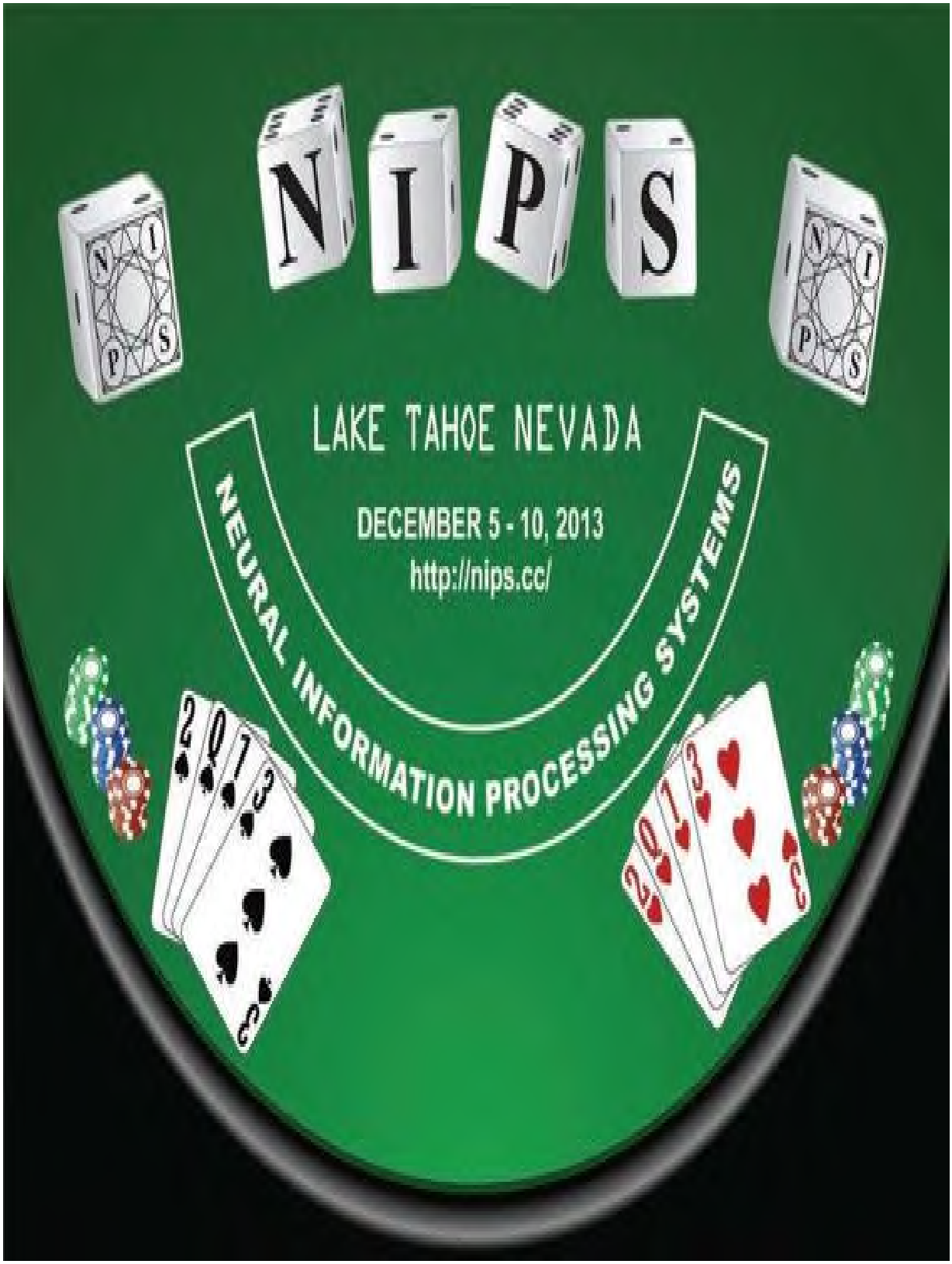


图11-3 在太浩湖赌场举办的2012 NIPS大会是深度学习领域的一个转折点，让“神经”重新回归了“神经信息处理系统”。图片来源：NIPS基金会。

在机器学习中，谁拥有最多的数据，谁就是赢家，显然Facebook所拥有的关于人们点赞、好友和照片的数据让其他人只能望其项背。利用所有这些数据，Facebook可以创建我们的心智理论，并用它来预测我们的偏好和政治倾向，甚至有一天可能会比我们更了解我们自己。Facebook有朝一日会成为奥威尔小说中老大哥的化身吗？[6]你会觉得这是一个令人不寒而栗的前景，还是发现有一个能满足你需求的数字管家很方便？我们可能会问Facebook是否应该拥有这种权力，但在这个问题上我们也许并没有多少发言权。

尽管我们在太浩湖的赌场举办了2012年和2013年的NIPS大会，但与会者都避开了赌桌。他们知道赌场庄家获胜的可能性更大，更何况，他们正在做的事情比赌博更令人兴奋。赌博会让人上瘾，那是因为多巴胺奖励预测误差系统是我们大脑的一部分（在第10章中讨论过）。赌场已经优化了有利于博彩的条件：获得巨大回报的允诺；随机间隔的偶然小胜（奖励）——研究已经证明，这是保持实验室老鼠不停按下食物按钮的最佳方式；在老虎机上获胜时触发的声音和灯光；全天昏暗的灯光，将由光线驱动的昼夜节律与正常的日夜更替分开，怂恿你下注。不过从长远来看，庄家显然赢得更多。

在蒙特利尔召开的2015年NIPS大会上，3800名国际与会者拥入了蒙特利尔会议中心。会议开场时的深度学习课程环节非常受欢迎，导致人满为患。出于防火安全方面的考虑，我们不得不遣散了一些人。高科技行业中几乎所有拥有大数据的公司都采用了深度学习技术，这一趋势正在以更快的速度扩散。在巴塞罗那举行的2016年NIPS大会开始前两周，我们将与会人数限制在5400名。一个从纽约飞来的人很失望地发现他无法在现场注册。2017年在长滩举行的NIPS大会在注册开放12天后就限制注册了，与会者达到了8000人。如果自2014年以来，每年出席会议的人数按50% 的增速持续下去，那么最终地球上的每个人都会想参加NIPS大会。当然，泡沫最终会破裂，但是与大多数泡沫一样，没有人知道这事儿什么时候会发生。

来自科学和工程部门的许多研究人员继续聚集在NIPS大会上，30年来年年如此。不过，在巴塞罗那举行的2016年NIPS大会上，5400名与会者中有40% 的人是第一次参加会议。在2016年之前，NIPS基金董事会都明智地决定让会议保持在单一会场举行，这对于大型会议来说很少见。其背后的初衷，是让每个人都能坐在同一个房间里，防止场地分散化。但在2016年，单一会场变成了双会场，因为很难再找到足够大的房间装下所有人。尽管如此，这也远远少于大多数其他大型会议中常见的10个会场。NIPS的文章接受率一直保持在20%左右，低于大多数期刊的接受率。NIPS在2016年举办了“机器学习中的女性”（Women in Machine Learning，简称WiML）会议，[7]为巴塞罗那带来近600名女性与会者（占与会者总人数的10%），有1000名女性参加了2017年在长滩举行的会议。多样性继续成为NIPS大会的标志。没有任何一个领域能够独立汇集创造了深度学习的多样化人才。

也许让人感到意外的是，虽然具备了影响如此多行业的潜力，保护深度学习知识产权的专利却很少。在20世纪80年代，我们希望能让学习算法成为一个新的科学领域的基础，同时认为获得专利对此起不到什么积极的作用。毫无疑问，现在很多公司已经为深度学习的特殊应用提交了专利，因为公司不会在没有保护的情况下对新技术进行大规模投资。

为未来做准备

神经网络学习的重大突破每30年就会发生一次，从20世纪50年代引入感知器开始，到20世纪80年代学习多层感知器算法，再到2010年开始兴起深度学习。其中每个阶段都经历了一段繁荣期，在短时间内取得了飞跃性的进展，随后便是较长时期的缓慢发展。但是其中一个不同点在于，随着每次复兴，兴盛时期的影响一直有增无减。最新的

增长动力来源于大数据的普及，而NIPS的故事早就为这一天的到来做好了准备。

---

[1] NIPS会议中的所有文章都可以在网络上获取：https://nips.cc/。

[2] 为了体验生物学家的行话，请参考从最近的一篇科学综述中随机抽取的这句话：“少突胶质细胞包含多种抑制轴突再生的蛋白质，包括髓磷脂相关糖蛋白，神经突生长抑制剂‘Nogo’，少突胶质细胞髓鞘糖蛋白和信号素”。（“Oligodendrocytes present a variety of proteins inhibitory to axon regrowth,including myelin-associated glycoprotein, the neurite-outgrowth inhibitor ‘Nogo,’oligodendrocyte-myelin glycoprotein, and semaphorins.”）B. Laha, B. K. Stafford,and A. D. Huberman, “Regenerating Optic Pathways from the Eye to the Brain,”Science 356, no. 6342 (2017): 1032.

[3] 这位神经学家是霍华德·瓦克特尔（Howard Wachtel），他当时正在科罗拉多大学波尔德分校研究海兔的神经系统。

[4] Engaged一词既有“订婚”之意，也有“致力于做……”之意。——译者注

[5] Krizhevsky, Sutskever, and Hinton, “ImageNet Classification with Deep Convolutional Neural Networks.”

[6] George Orwell, *Nineteen Eighty-Four*(London: Secker & Warburg, 1949). 这本书最近有了新的含义。

[7] “机器学习中的女性”（Women in Machine Learning）这一组织成立于2006年，为机器学习领域的女性创造了机会，并推动了她们的研究。请参阅http://wimlworkshop.org。

第三部分

人类，智

智能与未来

要事年表

1971年

诺姆·乔姆斯基（Noam Chomsky）在《纽约书评》杂志上发表了《针对斯金纳的驳斥》（The Case against B. F. Skinner）一文，这篇长文诱导一代认知科学家偏离了对学习行为的研究。

1982年

克劳德·香农（Claude Shannon）出版了开创性的书籍《通信的数学理论》（*A Mathematical Theory of Communication*），该书为当代数字通信技术奠定了基础。

1989年

卡弗·米德（Carver Mead）出版了著作《模拟大规模集成电路和神经系统》（*Analog VLSI and Neural Systems*），开创了基于仿生芯片的神经形态工程（neuromorphic engineering）研究领域。

2002年

斯蒂芬·沃尔夫勒姆（Stephen Wolfram）出版了书籍《一种新的科学》（*A New Kind of Science*）。这本书探究了细胞自动机（cellular automata）的计算能力，这些算法比神经网络更简单，但是仍能完成强大的运算。

2005年

塞巴斯蒂安·特隆的团队赢得了DARPA自动驾驶汽车大赛。

2008年

托拜厄斯·德尔布吕克（Tobias Delbrück）研发了一种非常成功的脉冲视网膜芯片“动态视觉传感器”（Dynamic Vision Sensor，

简称DVS），这种传感器采用异步峰值，而不是像现在的数字摄像机利用同步帧进行图像捕获。

2013年

白宫宣布启动美国“BRAIN计划”，开发创新的神经技术，以加速我们对大脑功能的理解。

12
智能时代

认知计算的时代即将到来。很快我们就会拥有比人类驾驶技术更高超的自动驾驶汽车。我们的房子会识别出我们的身份，预见我们的习惯，有客人到访时会提醒我们。最近被谷歌收购的众包网站Kaggle，曾经举办过一次奖金为100万美元的用CT扫描图进行肺癌检测的竞赛，它还为美国国土安全部举办了一场奖金达150万美元的竞赛，希望寻找到能在机场安检时检测出藏匿物的技术。[1]通过认知计算，医生的助理将有能力诊断罕见疾病，提高整体医疗水平。我们已经有了数千个这样的应用，将来还会出现更多类似的应用。有些工作会被淘汰，有些则将被创造出来。尽管认知计算技术十分具有颠覆性，我们的社会还需要时间来吸收和适应，但它对人类来说并不构成生存威胁。相反，我们正在进入一个发现和启蒙的时代，我们会变得更聪明，更长寿，人类文明也会变得更加繁荣。

2015年，我曾在由IBM赞助的旧金山认知计算会议上发表演讲。[2]IBM当时正在对沃森（Watson）项目进行大规模投资，沃森是一个基于大量事实数据库集合的程序，覆盖信息的范围从历史到流行文化，包罗万象，可以使用自然语言接口的各种算法进行查询。肯·詹宁斯（Ken Jennings）曾经连续192天在74场智力竞赛节目《危险边缘》（*Jeopardy!*）中赢得了胜利，这是该游戏节目历史上最长的连胜纪录。2011年，沃森在该节目中击败了詹宁斯，该事件引起了全世界的注意。

在从酒店出发到会场的出租车上，我无意中听到了后座两名IBM管理人员的谈话。IBM当时正要推出一个围绕沃森的平台，该平台可使用非结构化数据库，来组织和回答健康和金融服务等专业领域的问题。沃森掌握的数据比任何人可能知道的还要多，可以回答问题并提出建议。当然，与其他机器学习程序一样，它仍然需要人类提出问题，并从给出的建议中进行选择。

IBM早已裁撤掉了硬件部门，其计算机服务部门也早已风光不再。IBM在沃森上花费了巨资，想要依靠其软件部门来帮忙完成700亿美元的营收计划。该公司在慕尼黑已为沃森物联网业务投资了2亿美元用来建立新的全球总部，[3]这是IBM在欧洲有史以来最大的一笔投资，为的是满足6000多个客户不断增长的利用人工智能来改造运营业务的需求——IBM计划在全球投入30亿美元来发展认知计算，而这只是其全球计划的一部分。但许多其他公司也在对AI进行重大投资，现在想要断言哪些投注会赢，哪些会输，还为时尚早。

21世纪的生活

在传统医学中，通常采用相同的治疗方法对所有患有特定病症或疾病的患者进行治疗，但现在有了认知计算，治疗已变得更加个性化，也更精确。例如黑色素瘤，过去患上这种疾病的人就等同于被判了死刑，但现在已经可以通过对患者的癌细胞进行基因测序，并针对其病症设计出独特的癌症免疫疗法，以停止甚至逆转黑色素瘤的病情发展。虽然这种疗法目前的治疗费用为25万美元，但当对患者癌症基因组测序的基础成本降至几千美元，并且研制所需单克隆抗体的成本仅为几百美元时，最终几乎每个黑色素瘤患者都有能力负担相关的医疗费用。从大量患者那里收集到足够多变异和疗效方面的数据后，医疗决策将变得更好，且收费更低。一些类型的肺癌也可以用相同的方法进行治疗。制药公司正在投资癌症免疫治疗研究，许多其他种类的

癌症也许很快就能被治愈。如果没有机器学习方法来分析大量遗传数据，这一切都不可能实现。

我曾担任过NIH院长的顾问委员会成员，为推进“BRAIN计划”提供建议。我们的报告强调了概率和计算技术的重要性，该技术能帮助我们解释由新的神经记录技术产生的数据。[4]机器学习算法现在被用于分析同时来自数千个神经元的记录，分析来自自由移动动物的复杂行为数据，以及从电子显微镜的一系列数字图像中自动重建三维解剖回路。通过对大脑进行逆向工程，我们将揭秘自然界自身发现的许多新算法。

NIH在过去的50年中资助了神经科学的基础研究，但当前的趋势是将越来越多的资助转向了能够尽快实现医疗应用的转化研究。我们当然希望能转化已经发现的技术，但如果现在不为新的发现提供资金，那么从现在开始的50年里，我们几乎或根本没有任何技术可应用于临床。这也是为什么现在就开展一些基础研究项目非常重要，例如“BRAIN计划”，该举措能够帮助我们找到在未来治疗精神分裂症和阿尔兹海默症等衰弱性脑部疾病的方法。[5]

未来的身份认证

2006年，黑客从美国退伍军人事务部一个员工家中的电脑里窃取了2650万退伍军人的社会安全号码和出生日期信息。退伍军人行政处当时是使用社会安全号码作为其系统中的退伍军人标识符，黑客甚至都不需要解密数据库。有了社会安全号码和出生日期，黑客可以盗取任何人的身份信息。

在印度，超过10亿的公民可以通过指纹、虹膜扫描、照片和12位身份号码（比社会安全号码多3位数字）信息进行唯一的身份识别。印度的Aadhaar是世界上最大的生物特征识别项目。在以往，一位想要得

到一份公共文件的印度公民会经历无尽的等待，这个过程中还会涉及许多中间人，每个人都会向他索取贿赂。而如今，通过快速生物扫描，公民可以直接获得补贴食品权和其他福利待遇，许多没有出生证明的穷人现在也拥有了便携式身份认证，可随时随地在几秒钟内识别出身份。以欺骗手段获取福利的身份盗窃行为已经销声匿迹了。一个人的身份已经无法再被窃，除非小偷准备切断这个人的手指并摘除他的眼球。[6]

“印度国家登记处”是南丹·尼勒卡尼（Nandan Nilekani）[7]做了7年的一个项目，他是一个亿万富翁，并且是印度外包公司Infosys的联合创始人。尼勒卡尼庞大的数字数据库已经帮助印度在身份数字化方面超越了许多发达国家。根据尼勒卡尼的说法：“一个小的增量变化乘以10亿，就是一个巨大的飞跃。……如果10亿人能够在15分钟内拿到手机，而不用花上一周，那么这将为经济注入一股巨大的生产力。如果有100万人能让资金自动进入他们的银行账户，这将是经济上巨大的生产力飞跃。”[8]

在享受拥有数字身份数据库带来的便利的同时，我们也失去了隐私，尤其是当生物识别码与其他数据库（如银行账户、医疗记录和犯罪记录）以及其他公共计划（如运输）相关联时。隐私泄露问题在美国和许多其他数据库间互相关联的国家已经很严重了，即使它们的数据是匿名的。[9]显而易见，无论我们是否愿意，我们的手机都已经在追踪我们的行踪了。

社交机器人的崛起

电影经常将人工智能描述为像人类一样走路和说话的机器人。不要指望AI看上去像1984年的科幻电影《终结者》里操着德国口音的终结者那样。不过，你会与2013年的科幻电影《她》中萨曼莎一样的AI

声音交流，并和2017年科幻电影《星球大战：原力觉醒》中R2-D2和BB-8一类的机器人进行互动。AI已经是日常生活的一部分。认知设备，比如亚马逊Echo智能音箱的语音助理Alexa，已经能和你交谈，并很高兴能让你的生活更轻松，更有满足感，就像2017年电影《美女与野兽》中的时钟和茶具一样。生活在一个拥有那样物种的世界里会是什么样子呢？来看看我们迈向社交机器人的第一步吧。

目前，人工智能的进展主要集中在感官和认知方面，运动和行为智能的进展还未见端倪。有时我演讲的开场白便是，大脑是已知宇宙中最复杂的设备。但我的妻子比阿特丽斯是一位医生，她经常提醒我说，大脑只是身体的一部分，而身体比大脑更复杂，尽管身体的复杂性（从运动性演变而来）是不同的。

我们的肌肉、肌腱、皮肤和骨骼能够积极地适应世界的变迁、地心引力和其他同类。从内部来说，我们的身体是化学加工的奇迹，可以将食物转化为做工精细的身体部件。它们是由内而外工作的终极3D打印机。我们的大脑从身体各个部分的内脏传感器获得输入，这些传感器不断监测身体内部活动，包括最高层次的皮层表征，就内部优先事项做出决定，并在所有相互竞争的要求之间维持平衡。从真正意义上来说，我们的身体是大脑不可或缺的组成部分，这是具身认知（embodied cognition）的核心原则。[10]

哈维尔·莫维兰（Javier Movellan）（图12－1）来自西班牙，曾经是加州大学圣迭戈分校神经计算研究所机器感知实验室的教员和联合主任。他相信，通过打造能够与人类互动的机器人，我们将比采用传统实验研究更多地了解认知。他搭建了一个机器人婴儿，当你对它微笑时，它也对你微笑，颇得路人的喜爱。哈维尔研究了婴儿与母亲的互动，他得出的结论是，婴儿会尽量从母亲那里获得更多的微笑，同时尽量减少自己的努力。[11]

哈维尔搭建的最著名的社交机器人Rubi看起来像是一个天线宝宝（Teletubby），有着丰富的面部表情，眉毛能够扬起表现出有兴趣，

相机眼睛能够转来转去，手臂也可以抓取东西（见图12－2）。在加州大学圣迭戈分校儿童早期教育中心，18个月大的幼儿通过Rubi腹部的平板与它进行互动。



图12-1 哈维尔·莫维兰在加州大学圣迭戈分校的机器人研讨会上接受了《科学网络》数字论坛的采访。哈维尔率先在教室里推出了社交机器人，并为社交机器人Rubi编写了程序，以引起18个月大幼儿的注意。图片来源：罗杰·宾汉姆。

幼儿很难取悦，他们的注意力非常短暂。他们玩几分钟玩具就会失去兴趣，把它们扔到一边。他们会如何与Rubi互动呢？为了安全起见，Rubi不是按照工业强度打造的，于是第一天，男孩们就扯掉了它的胳膊。经过一些修理和加载软件补丁后，哈维尔再次进行了尝试。这一次，机器人要执行的程序是，当手臂被拉开时要放声大哭。这阻

止了男孩们的“暴力”行为，并且让女孩们争相拥抱Rubi。这是社交工程领域学到的重要一课。

幼儿通过指向房间内的物体（如时钟）与 Rubi玩耍。如果Rubi不能在0.5~1.5秒的时间窗口内将视线移向物体，那么幼儿就会失去兴趣并走开。如果反应太快，Rubi的机械性就太明显了；如果反应太慢，Rubi又会让孩子们感到很无聊。一旦互惠关系形成，孩子们就会将Rubi看作是一个有感情的生命，而不是一个玩具。当Rubi被带走（到维修店进行升级）后，孩子们会感到不安。他们会被告知Rubi感到不舒服，会在家休息一天。在一项研究中，Rubi被要求教幼儿学习芬兰语单词，他们学习的速度和学英语单词一样快；教他们学习一首流行歌曲的效果就更明显了。[12]



图12-2 Rubi在教室里与幼儿们进行互动。Rubi的头可以旋转，眼睛是相机，嘴和眉毛都十分富有表现力。Rubi头顶上浓密的光纤会随着它的心情而改变颜色。图片来源：哈维尔·莫维兰。

之前对将Rubi引入课堂环境的顾虑之一，是教师可能会担心自己在某一天被机器人取代。但事实恰恰相反：老师对Rubi表示欢迎，因为它可以作为一位辅助管理班级秩序的助手，特别是在教室里有访客时。一个能够彻底革新早期教育的实验是“千个Rubi项目”（thousand Rubi project）。这个想法是批量生产Rubi，将它们放在上千个教室中，每天通过互联网从数千次实验中收集数据。教育研究的一个问题是，在一所学校适用的方案可能不适用于另一所学校，因为学校之间存在很多差异，特别是教师之间的差异。“千个Rubi项目”也许能够考察许多关于如何改进教育实践的想法，并且可以探究在各地服务于不同社会经济群体的学校之间的差异。虽然能够让项目开展起来的资源一直没有到位，但这仍然是一个好主意，应该有人来跟进。

双腿站立的机器人很不稳定，需要一个复杂的控制系统来防止它们翻倒。而且事实上，婴儿需要12个月才能学会走路而不会摔倒。在第2章中已经做过简短介绍的罗德尼·布鲁克斯（见图12-3），想要打造能够像昆虫一样走路的6条腿机器人。他发明了一种新型控制器，可以对6条腿的动作进行排序，能够让机器人蟑螂向前移动并保持稳定。他那开创性的想法是，让腿部与环境进行的机械交互作用代替抽象的计划和计算。他认为，要让机器人完成日常任务，它们较高的认知能力应该基于感官运动（sensorimotor）与环境的相互作用，而不是抽象的推理。大象具有高度的社会性，有着强大的记忆力，并且还是器械天才，[13]但它们不会下国际象棋。[14]1990年，布鲁克斯继续创立了iRobot，迄今为止已经卖掉了超过1000万个Roombas扫地机器人来清洁更多的地板。

工业机器人具有坚硬的关节和强大的伺服电机，这使得它们看起来和感觉起来都很机械化。2008年，布鲁克斯创立了Rethink Robotics公司，该公司设计出了一个关节柔韧，可以转动手臂的机器

人Baxter。它的每个手臂可以按要求来移动，还会自行编程以重复执行过的动作顺序，人们不必再编写一个程序来移动Baxter的手臂了。



图12-3 罗德尼·布鲁克斯在观察机器人Baxter，它正准备将一个塞盖插入桌上的一个洞中。布鲁克斯是一位连续创业者，之前创建了iRobot，生产了扫地机器人Roombas，现在他又创建了Rethink，生产Baxter。图片来源：罗德尼·布鲁克斯。

莫维兰比布鲁克斯更进一步，开发出了一种名为“Diego San”（在日本制造）的机器人宝宝[15]，其电动机是气动的（由气压驱动），对比大多数工业机器人使用的刚性转矩电机，Diego San所有的44个关节都更柔软（见图12－4）。制作它们是基于这样一个灵感：当我们拿起某件东西时，我们身体中的每一块肌肉都会在一定程度上被牵动（如果我们一次只移动一个关节，看起来就会像机器人）。这使我们能够更好地适应不断变化的负载情况以及与世界的互动。大脑可以同时自然流畅地控制身体中的所有自由度——所有关节和肌肉——而DiegoSan项目的目标是找出大脑是怎么做到这一点的。Diego San的脸上有27个运动部件，可以表达各种各样的人类情感。[16]机器人宝宝的动作非常逼真。虽然哈维尔有几个成功的机器人项目，但Diego San并不是其中之一。他并不知道如何使机器人婴儿的表现像人类婴儿那样流畅。



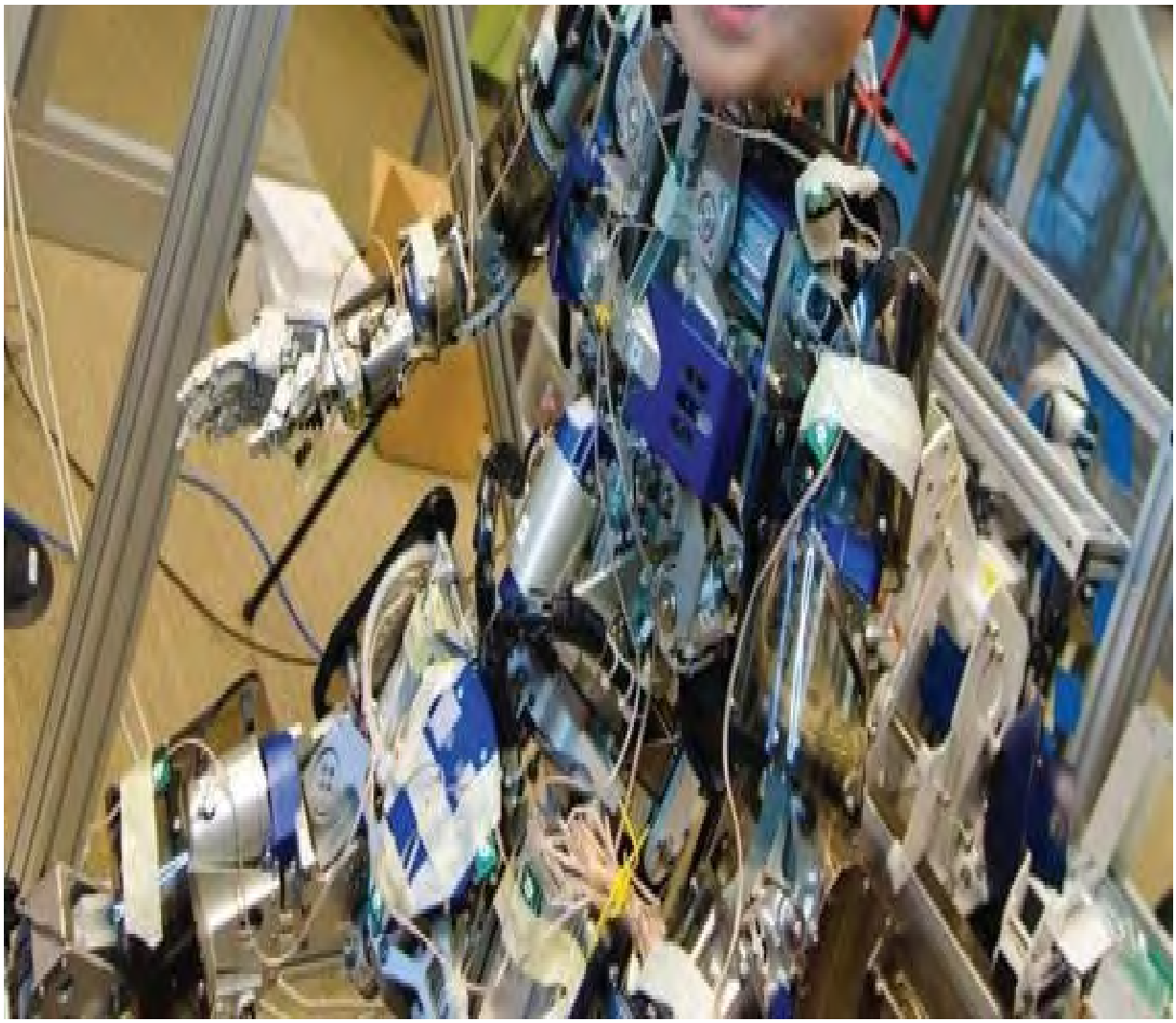


图12-4 机器人宝宝Diego San。气压传动装置让所有关节都能平稳移动，这样就能跟人类握手了。面部由大卫·汉森（David Hanson）和汉森机器人公司提供。关于机器人面部表情的动画，可以参考“Diego Installed”，由哈维尔·莫维兰提供，https://www.youtube.com/watch?v=knRyDcnUc4U/。

机器已经会识别人类面部表情

你能想象当你在iPhone（苹果手机）上看到股价暴跌，iPhone问你为什么不高兴，会是怎样一种感觉吗？你的面部表情是情绪的窗口，深度学习现在已经可以探进这个窗口了。传统上，人们认为认知和情绪是大脑两个独立的功能。一般认为，认知是皮层功能

（cortical function），而情绪是皮层下的功能（subcortical function）。事实上，有些皮层下结构可以调节情绪状态，比如杏仁核（amygdala）。当情绪水平很高，特别是感到恐惧的时候，杏仁核就会起作用，但这些结构与大脑皮层有强烈的相互作用。例如，在社交互动中，杏仁核的参与会增强人脑对该事件的记忆。认知和情绪是彼此交织的。

在20世纪90年代，我与加州大学旧金山分校的心理学家保罗·艾克曼（Paul Ekman）（见图12－5）有过合作。艾克曼是世界著名的面部表情专家，是美剧《别对我撒谎》（*Lie to Me*）系列中卡尔·莱特曼（Cal Lightman）博士在真实世界中的原型，但他本人比剧中的卡尔要和善很多。艾克曼曾去过巴布亚新几内亚，试图了解前工业时代的文化是否像我们一样，能够做出情绪化的面部表情。他在研究过的所有人类社会中发现了六种普遍的情感表达：快乐、悲伤、愤怒、惊讶、恐惧和厌恶。从那以后，其他常见的面部表情也被提取了出来，但人们在对这些表情的理解上并没有达成共识。有些表达，如恐惧，在一些孤立的社会中有着不同的解释。

1992年，艾克曼和我组织了一个由美国国家科学基金会（NSF）赞助的关于面部表情理解的规划研讨会。[17]在那个时期，要获得对面部表情研究的资金困难重重。我们的研讨会将神经科学、电子工程和计算机视觉等方面的研究人员，连同心理学家聚集到了一起，为表情分析领域揭开了新篇章。这件事对我来说是一个启示，尽管面部表情的分析在科学、医学和经济等诸多领域都扮演着重要角色，但其重要性却一直被基金机构所忽视。



图12-5 1967年，保罗·艾克曼和巴布亚新几内亚原住民。他找到了6种常见的面部情感表达的证据，包括快乐、悲伤、愤怒、惊讶、恐惧和厌恶。保罗曾为美剧《别对我撒谎》提供咨询，以保证每一集的内容有科学依据。卡尔·莱特曼博士的角色大致上是以艾克曼为原型塑造出来的。图片来源：保罗·艾克曼。

艾克曼开发了面部动作编码系统（Facial Action Coding System，以下简称FACS）来监控面部44块肌肉的状态。由艾克曼培训的FACS专家需要花费一个小时来标记长度为一分钟的视频，每次标记一帧。面部表情是动态的，可以延续若干秒，但是艾克曼发现，有些表情只持续了几帧。这些“微表情”泄露了大脑被抑制的情感状态，常常能说明，有时还会揭示无意识的情绪反应。例如，在婚姻咨询环节中出现的细微的厌恶表情，就是一个预示婚姻会失败的可靠的信号。[18]

在20世纪90年代，我们使用了一些演员的视频录像来训练反向传播神经网络，对FACS进行自动化。这些训练有素的演员可以像艾克曼

那样控制脸上的每一块肌肉。1999年，由我的研究生玛丽安·斯图尔特－巴特利特（Marian Stewart-Bartlett，见图12－6）利用反向传播训练的网络，在实验室识别面部表情的准确率达到了96%，前提是要具备完美的照明，周正的面部角度，以及对视频进行手动时间分割。[19]这个网络的识别效果很不错，于是1999年4月5日，玛丽安和我在黛安·索耶（Diane Sawyer）主持的《早安美国》（Good Morning America）节目中向公众展示了这一成果。在成为加州大学圣迭戈分校神经计算研究所的一名教师后，玛丽安接着又开发了计算机表情识别工具箱（Computer Expression Recognition Toolbox，以下简称CERT）。[20]随着计算机的速度变得越来越快，CERT已经能够进行实时分析，可以标记视频流中不断变化的人类面部表情。

2012年，玛丽安和哈维尔·莫维兰将面部表情的自动分析技术商业化，创立了一家名为“Emotient”的公司。保罗·艾克曼和我为这家公司担任科学顾问。Emotient开发的深度学习网络能够以96% 的准确率，在各种不同的照明条件下，利用非正面面部信息实时地对自然行为做出判断。在Emotient的一个演示中，其网络在几分钟内就检测到唐纳德·特朗普（Donald Trump）在共和党的首场初选辩论中，对一个焦点小组[21]的情绪影响最大。相比之下，民意调查人员花了数天时间才得出了同样的结论，专家们更是在几个月之后才认识到，情感投入是争取到选民的关键。焦点小组中最强烈的面部表情是愉悦，其次是恐惧。此外，Emotient的深度学习网络在尼尔森收视率（Nielsenratings）调查结果发布前的几个月，就预测到了哪些电视剧将大热。Emotient于2016年1月被苹果公司收购，玛丽安和哈维尔现在均在苹果公司任职。



图12-6 正在演示面部表情分析技术的玛丽安·斯图尔特-巴特利特。时间轴上显示的是深度学习网络的输出，正在识别面部表情中快乐、悲伤、惊讶、恐惧、愤怒和

厌恶的情绪。图片来源：玛丽安·斯图尔特-巴特利特，Robert Wright/LDV Vision Summit 2015。

在不久的将来，你的iPhone可能不仅会问你为什么不高兴，还可能帮助你冷静下来。

新技术改变教育方式

13年前，在温哥华举行的2005年NIPS大会上，我和加州大学圣迭戈分校计算机科学与工程系的同事加里·科特雷尔（Gary Cottrell）在一起吃早餐。加里属于20世纪80年代PDP团队最初的一批成员，也是PDP团队仍留在加州大学圣迭戈分校为数不多的几人之一。他是20世纪60年代最后几个仍在坚持神经科学研究团体中的一员，留着马尾辫和灰白的胡须。他读到了NSF发表的一项声明，要求为“学习科学中心”（Science of Learning Centers）提供提案。其中让他感兴趣的内容，是连续5年每年500万美元的预算，而且还可以再延长5年。他想提交一份项目申请，并询问我是否愿意帮忙。他当时说，如果成功了，他就不用再写别的基金申请了。我告诉他说，如果成功的话，这项经费能让他终身从事这一事业了。他笑了起来，之后我们开始了合作。

最终我们的申请被批准了，如我所料，尽管每年提供300页的年度报告很让人崩溃，但取得的科学成果非常惊人。我们的时间动态学习中心（Temporal Dynamics of Learning Center，以下简称TDLC）共有来自全球18个机构的100多名研究人员。在由NSF资助的6个学习科学中心里，我们这个是最以神经科学和工程为导向的，我们将机器学习的最新进展纳入了这些项目中（见图12－7）。[22]Rubi和CERT是由TDLC资助的两个项目。我们还有一个移动脑电图实验室，让被试者可以在虚拟环境中自由漫游，同时记录他们的脑电波。在大多数脑电图实验室中，被试者必须坐着不动，眼睛都不能眨，以避免产生假象。

我们使用了ICA来抵消运动假象，让我们能够观察被试者主动探索环境，与其他人互动时的大脑活动。



图12-7 新的学习科学包括机器学习和神经科学，以及来自心理学和教育的见解。图片来源：Meltzoff , Kuhl,Movellan, and Sejnowski, “Foundations for a New Science of Learning,”图1。

以下是TDLC研究人员开展的众多项目中的一部分。

罗格斯分子和行为神经科学中心（Center for Molecular and Behavioral Neuroscience at Rutgers）的艾普利·本纳斯奇（April Benasich）开发了一项测试，可以根据宝宝的听觉感知时间来预测婴儿是否会在语言习得和学习能力上存在缺陷；她指出，这些缺陷可以通过自适应地控制声音出现的时间和奖励反馈来纠正，以便宝宝能够发育出正常的听力、语言和学习能力。[23]这些结果所针对的实验对象年龄跨度从3个月到5岁。即使正常发育的婴儿也会从互动环境中受益。艾普利于2006年创立了AAB Research LLC，让快速听觉处理技术（rapid auditory processing technology，简称RAPT）进入普通家庭，以提高婴儿的学习能力。

玛丽安·斯图尔特－巴特利特和哈维尔·莫维兰使用机器学习自动登记学生的面部表情，[24]当学生看起来很沮丧，似乎不太理解授课内容时，教师可以收到提醒并采取措施。利用深度学习技术，这一过程如今可以自动且准确地同时记录班级中每个孩子的表情信息。在市场营销、精神病学和法医学等领域，还有很多的面部表情分析应用有待开发。

加州大学圣迭戈分校的哈罗德·帕施勒（Harold Pashler）和科罗拉多大学博尔德分校的迈克尔·莫泽尔（Michael Mozer）研究了12年级的学生在数年内的学习表现，以扩充先前针对大学生为期数月的研究。研究结果表明，个性化的时间分散式复习比短期的填鸭式复习更能改善对学习材料的长期记忆。[25]他们表示，当学习材料需要在学生的记忆中停留较长时间时，学习的最佳间隔时间最好设得长一些。他们在语言课程中为学生们制定了最佳复习时间表，并帮助他们取得了优异的成绩。

TDLC的博士后研究员贝丝·罗戈夫斯基（Beth Rogowsky）、罗格斯大学的保拉·塔拉尔（Paula Tallal）和范德堡大学的芭芭拉·卡尔霍恩（Barbara Calhoun）表明，在使用口头或书面材料的学习方法之间没有统计学差异，而且，无论是即时理解还是延迟理解，偏好的学习方式和教学方法之间都不存在联系。[26]学生使用自己更偏好的方式学习不会带来任何积极的效果，这也就意味着大型企业宣传的针对个人学习方式的培训和测试材料，并不会提升课堂的学习效果。

保拉·塔拉尔在2014年启动的1500万美元的“全球学习X奖”（Global Learning X-Prize）中发挥了重要作用。该奖项鼓励教育创新，其目标是开发开源代码和可扩展软件，使发展中国家的儿童能够在18个月内掌握基本的阅读、写作和算术技能。为“全球学习X奖”所做研究的积极影响，将在未来数十年内在全世界引起回响。

TDLC的科学总监安德里亚·千叶（Andrea Chiba）在2014年上海“学习科学国际大会”上介绍了关于所有的学习行为如何改变大脑结构的研究。[27]这令许多代表都感到惊讶，他们一直相信儿童来到世界上时就具备了一定的潜力，对那些能力不足或者年龄过大的学习者来说，教育只是一种浪费。世界各地仍然存在着巨大的人类潜力有待挖掘。

我们发现教育中最大的问题不是科学方面的，而是来自社会和文化。美国有13500个学区，每个学区都有自己的学校董事会，负责决定课程、教师资格和最佳实践；需要数十年才能实现这些目标，并有针对性地处理这些问题。在教师实施教学之前，他们还必须管理教室秩序，这对于低年级和市中心的学校来说尤其具有挑战性。提出要求的家长们可能不会意识到教学资源的匮乏导致了教师中的职业倦怠现象频发，而教师工会的影响也难辞其咎，后者往往阻碍了日积月累的努力。

教学从根本上来说是一项劳动密集型活动。最优质和最有效的教学方式是让经验丰富的成人教师和学生之间进行一对一交流。[28]我

们背负着一个专为大众教育而设计的流水线系统，对学生按年龄进行划分，教师在大班里年复一年地传授相同的课程。这可能是一种生产汽车的好方法，在劳动力只接受基本教育就能满足社会需求的年代，这样的方法还算行得通。但是当工作岗位需要更高水平的培训和终身学习来更新工作技能时，这个系统就落伍了。诚然，作为成年人，回到学校可能是痛苦和不切实际的。我们正在经历的信息革命已经超越了劳动力世代更替的时间尺度。幸运的是，出现在互联网上的新技术可能会改变我们的学习方式。学习科学中心于2006年创立时，我们并没有预见到，互联网正在重构我们的学习版图。

成为更好的学习者

慕课在2011年突然流行起来，《纽约时报》发表了一篇热门文章，提到了一门大受欢迎的斯坦福大学人工智能在线课程。[29]大量的学生参加慕课，他们在互联网上这一前所未有的学习行为，在全世界引起了广泛的关注。几乎在一夜之间，很多新公司获得了融资，这些公司专注于开发以及免费在线发布世界上一些最优秀的教育者的讲座。只要能连上网，人们就能随时随地观看这些讲座。除了讲座之外，课程内容还包括测验、考试、供学习者交流的论坛、助教辅导，以及自发组织的本地“见面会”，方便学生在非正式的环境中讨论课程内容。慕课的受众面已经大大增加了——2015年，“慕课人”的数量翻了一番，从估算的1700万增加到了3500多万。慕课绕开了现有教育体系中的所有限制因素。

2013年1月，在加州大学尔湾分校由美国国家科学院主办的一次会议上，我遇到了芭芭拉·奥克利。尽管她读书的时候在数学和科学方面表现欠佳，但她现在是位于密歇根州奥本山和罗切斯特山市的奥克兰大学的一名电子工程学教授。她是曾主修人文科学专业的美国陆军上尉，回到学校之前，她在白令海的拖网渔船上担任俄语翻译。后来

她克服了自己在数学方面的心理障碍，获得了电子工程博士学位。在会议的晚宴上，我发现芭芭拉和我在学习方面有着相似的看法，她当时正在写一本书，书名是《学习之道》（*A Mind for Numbers: How to Excel at Math and Science*）。于是我邀请她到加州大学圣迭戈分校为高中学生和老师们举办一次TDLC讲座。

芭芭拉受到了学生们的热烈欢迎，很明显她是一位颇有天赋的老师。她的方法和实用性见解源于我们对大脑的了解，所以我们联手为Coursera开发了一门慕课，名为“学会如何学习：帮助你掌握复杂学科的强大智力工具”（见图12-8；https://www.coursera.org/learn/learning-how-to-learn/），于2014年8月推出。它目前是世界上最受欢迎的慕课，在头四年里注册的学习者超过300万，并且每天还会吸引1000个来自200多个国家的新学员。“学会如何学习”根据我们对大脑学习方式的了解，为你提供了成为更好学习者所需的工具。我们学习者的反馈非常积极，于是我们又开发了第二个名为“思维转换”（Mindshift）的慕课，以帮助那些想要转业或改变生活方式的人。这两门慕课课程的资料都可以在网络上免费获取。



图12-8 芭芭拉·奥克利介绍“学会如何学习”慕课。已有超过300万学员参加了该课程，使其成为全球最受欢迎的互联网课程。图片来源：芭芭拉·奥克利。

“学会如何学习”这一课程，就如何成为更好的学习者，如何处理测试焦虑，如何避免拖延，以及我们对大脑怎样学习的认识等话题提供了实用性的建议。这是一门为期一个月的免费课程，由长度为

5~10分钟的视频剪辑、小测验和综合测试组成，现已被翻译成20多种语言。课程的基石之一，就是当你正在做别的事情时，你的大脑可以在潜意识中帮你完成一些工作。亨利·庞加莱（Henri Poincaré）是19世纪一位杰出的数学家，他曾经描述自己如何最终解决了一个数学难题。当时他已经紧张地工作了好几个星期，但仍然一无所获。于是他度假去了。当他在法国南部踏上一辆公共汽车时，问题的解决方案突然就出现在了他的脑海中，他的大脑中有一部分区域在他享受假期的时候仍然在处理这个问题。他知道自己找到了证明问题的正确方法，并在返回巴黎时完成了它。他之前对这个问题的深入研究已经使他的大脑做好了充分的准备，这样他的潜意识就可以在他放松的时候继续处理这个问题。这两个阶段对创造力而言同等重要。

令人惊讶的是，即使在睡觉的时候，你的大脑也会在你毫不知情的情况下解决问题。但是，只有在你入睡之前仍专注于解决问题的前提下，大脑才能做到这一点。于是你早上醒来的时候，一种新的见解就会从脑海中跳出来，帮助你解决问题（尽管这种情况不会经常发生）。在休假或入睡前的高强度思考，对激活你的大脑非常重要；否则，它就有可能去处理其他问题了。这种方式并不限于数学或科学问题，如果你最近在思考一个社交问题，你的大脑就会像解决数学和科学问题一样努力解决它。

“学会如何学习”最令我们感到欣慰的成果之一，就是收到了来自快乐学习者的信件，感谢我们为他们提供了最好的课程，或者这门课如何影响了他们的职业选择。[30]教师也写信给我们，提到他们正在将“学会如何学习”这门课的经验融入自己的课堂中。

我们最初的目标是教授高中生和大学生“学会如何学习”，但后来发现这些人的比例不到学习这门课程总人数的1%。因为学校一直以来都被要求以“共同核心”（Common Core）测试为目标进行教学，他们没时间教学生如何学习，而后者往往对学生们的学业更有帮助。要求各个学区都采用“学会如何学习”的教学模式将是一个艰难的任

务，毕竟学校的运营预算是有限的。学区也不打算修改他们的课程，把“学会如何学习”纳入大规模的教学。任何大范围的尝试都涉及课程表的调整、教师再培训和新教材的开发，这个过程相当的费时费力。但不论如何，我们都需要12岁的孩子在进入高中前接触到这项课程。芭芭拉和我针对这群读者写了一本书，希望这些年轻的学生在进入会频繁遇到数学难题的中学时期之前，就能接触到我们的技巧。[31]

与学校课程“要么太多，要么没有”的学习模式不同，慕课更像是你随时可以挑选和阅读的书：学习者们更倾向于“放养式”的学习方法，有针对性地选择符合他们迫切需求的课程。慕课原本被认为是传统教室的替代品，现在它在教育领域占据了一个互补性的位置，这与其他教学场所不同。它以独特的方式满足了学习者的需求，而传统教育方式往往做不到这一点。例如，慕课已经被翻转课堂（flipped classes）这一教育方式所采纳，学生们可以在方便的课余时间观看选定的讲座，老师则会针对课程内容在课堂上引导学生们进行讨论。我们的教育体系是为工业时代设计的，学校所传授的知识曾经是你今后维持工作，以及作为一名有生产力的公民所需的基础知识。然而现在，学生刚一毕业，学校传授的那些知识就已经过时了。慕课直接进入家庭，是教育系统的最后一环。对Coursera进行的人口统计表明，大多数在线学习者的年龄都在25~35岁之间，他们当中超过一半的人接受过大学教育。这些都是在工作中需要新技能，并在网上进行自学的年轻人。我们的教育系统需要更多根本性的改变，让我们的大脑能够适应在经济信息领域中快速扩张的岗位技能需求。例如，通过互联网收集信息需要一些判断力和基本技能，即能够制定合适的搜索条件，并甄别错误的搜索路径。可惜的是，学校似乎没有时间教授基本的互联网操作技能，即使学生们能够从学习如何积极寻求信息，而不是被动接受课程中受益。

优达学城（Udacity）是由因自动驾驶汽车而名声大噪的塞巴斯蒂安·特隆创立的另一家制作慕课的教育机构。除了提供免费课程外，优达学城还与希望提升自身员工技能的公司开展了合作。优达学城打造了符合公司需求的慕课，员工们也有动力去学习这些课程。对于雇主、员工和优达学城来说，这是共赢。优达学城也提供了一系列专项课程，比如自动驾驶汽车技术（费用为800美元），学员学成之后可以获得纳米学位（nanodegree），如果在6个月内找不到工作还能收到全额退款。[32][33]传统学校之外的教育机构发展迅速，慕课可以为终身学习提供各种解决方案。

我们后续的一门慕课“思维转换：突破学习障碍并发掘你的潜能”（MindShift: Break through Obstacles to Learning and Discover Your Hidden Potential，https://www.coursera.org/learn/ mindshift/）已经于2017年4月推出。该课程是与芭芭拉·奥克利的一本新书同步推出的，[34]基于他人的亲身经历，并使用早期的案例（包括我的）来阐明，当你想以某种方式改变你的生活时可能会面对的各种难题。就我而言，我是从物理学转到了神经生物学领域，但在另一个案例中，一位成功的音乐会独奏家放弃了他的职业生涯，转而成为一名医生。转行这一现象已经变得越来越普遍，“思维转换”的课程设计旨在降低该过程的难度。该门课程现在在世界上最受欢迎的慕课中排名第三。

另一种成为更好的学习者的方式，是利用交互式电脑游戏。类似Lumosity这样的公司提供了在线游戏，声称可以提高玩家的记忆力和注意力。问题是，支持这种说法的研究往往凤毛麟角，或者质量不高，特别是在将培训转移到实际任务方面。但是，这一切都还处于早期阶段，更出色的研究正开始帮助我们辨别有效和无效的方法。结果往往令人大开眼界，甚至有悖直觉。

训练你的大脑

在广泛改善认知功能方面最有效的视频游戏是那些追逐僵尸，在战争中杀死坏人，以及极速赛车之类的游戏。日内瓦大学的达芙妮·芭菲莉亚（Daphne Bavelier）表示，玩《荣誉勋章：联合袭击》（*Medal of Honor: Allied Assault*）等第一人称视角的射击游戏改善了人们的感知、注意力和认知能力，尤其是视觉、多任务处理和任务转换，并且引导他们更快地制定出决策。[35]她总结说，玩这类射击游戏可能会让年长者的大脑反应与年轻人的一样快（对于任何上了年纪的人来说都是好消息）。但是一些射击游戏也可能会有损长期记忆。[36]每款游戏都有不同的优点和缺点，需要单独进行分析。

加州大学旧金山分校的亚当·格萨雷（Adam Gazzaley）定制了一款叫作NeuroRacer的三维视频游戏，能够提高你的多任务处理能力。有研究表明，大脑中神经调节剂的活动对于注意力、学习和记忆有着重要的作用。[37]在NeuroRacer中，玩家沿着蜿蜒崎岖的道路驾驶汽车，同时还要留意随机弹出的各种标志。这就要求玩家调用多种认知技能，例如注意力和任务切换能力，进行多任务处理。在测试NeuroRacer时，亚当和同事们发现，在经过训练以后，被试者的这些技能得到了显著提高，并在工作记忆和持续关注方面达到了较高的水平，然而这些并不属于训练的一部分。此外，他们的表现还优于未经训练的20岁年轻人，并且在中止训练的6个月后，这些能力依然没有消退。[38]NeuroRacer目前正在被用于临床试验，为患有注意力缺陷和记忆障碍的患者提供治疗。

1997年，尚在罗格斯的保拉·塔拉尔和尚在加州大学旧金山分校的迈克尔·梅泽尼奇（Michael Merzenich）联合创建了一家名为“Scientic Learning”的公司，为患有语言和阅读障碍（如失读症）的儿童提供帮助。语音理解取决于能否捕捉到快速的声学转换。例如，听到“ba”、“ga”或“da”之间的差异取决于音节开头的毫秒范围内的时间差异。无法检测到这些时间差异的孩子在学习中会处于劣势，因为他们无法分辨包含这些声音的词汇。要学习阅读，孩子们

必须能够识别和区分单词中字母代表的简短声音。塔拉尔和梅泽尼奇开发了被称为“Fast ForWord”的一系列计算机游戏，它首先通过在音节、单词和句子中夸大声学时差来提高孩子的听觉判断、语言和阅读理解能力。随着孩子们在各种语言和阅读水平上的提高，这种时间差异会被缩短。[39]Fast ForWord在教育游戏的同类产品中得到了很高的评价，有6000所学校和超过250万名儿童使用过该系列。它们也在超过55个国家被用来帮助孩子学习英语这门第二语言。梅泽尼奇后来又开发了BrainHQ（https://www.brainhq.com），这是一款基于相似科学原理的游戏，旨在减缓老年人的认知衰退。

你还可以通过锻炼大脑来提高运动技能。加州大学河滨分校的亚伦·赛茨（Aaron Seitz）开发了一个能够改善视觉感知和反应时间的计算机程序。在使用这个程序后，棒球队队员普遍反映视力变得更好了，被三振出局的次数变少了，还创造出了更多跑垒，并在赛季的54场比赛中多赢得了4~5场比赛。[40]赛茨开发了一款名为“UltimEyes”的廉价应用程序，将他的研究成果对公众开放。不过联邦贸易委员会已经停止了这款游戏的传播，直到更多的研究能进一步证实他的说法。[41]

在玩考验反应时间的游戏时，某些认知技能上的提高倾向于转移到其他认知技能上。但在玩许多其他特定领域游戏，如记忆游戏时，这种转移并不存在。虽然我们设计的交互式视频游戏正在变得越来越好，能够改善我们的大脑功能，增加游戏的趣味性，还能制成应用程序，但这种转移发生的条件仍然有待研究。全球范围内认知改善的潜力不可估量。

智能商业

在2015年NIPS大会开幕式上，我穿着全美运动汽车竞赛协会（NASCAR）风格的夹克欢迎与会者，上面贴着我们所有42个赞助商的商标（见图12－9）。在巴塞罗那举行的2016年NIPS大会上，赞助商有65家，一件夹克上已经放不下了。而到了2017年，多达93家赞助商支持了在长滩举行的NIPS大会。这种爆炸性的增长迟早会结束，但它在社会上的回响可能会持续数十年。这些赞助公司派遣招聘人员参加NIPS大会，渴望聘用市场上缺乏的优秀研究人员。我的许多同事已经加入了谷歌、微软、亚马逊、苹果、Facebook、百度和其他许多初创公司。它们还从大学里抢夺人才。据塞巴斯蒂安·特隆估计，当一家像Otto或Cruise这样的自动驾驶初创公司被一家大公司收购时，相当于在每个机器学习专家身上都花费了1000万美元。[42]

2013年，谷歌收购了杰弗里·辛顿的公司DNNresearch，当时这家公司由他和他在多伦多大学的两名研究生在运营着。杰弗里·辛顿于是成了谷歌的员工。他现在可以获得比他曾在多伦多憧憬过的更庞大的计算能力，但更重要的是谷歌所拥有的海量数据。谷歌大脑（Google Brain）是杰夫·迪恩（Jeff Dean）领导的由才华横溢的工程师和科学家组成的杰出人才团队。迪恩设计了谷歌所有的服务都依赖的MapReduce文件系统。当你让谷歌翻译一段文字时，它现在所使用的就是迪恩的谷歌大脑团队设计的深度学习技术。当你搜索一个关键词时，深度学习会对搜索结果进行排名。当你与谷歌助手（Google Assistant）进行交谈时，它会使用深度学习来识别你的话。随着与你的对话进行得更加流畅，它将使用深度学习为你提供更好的服务。谷歌已经在全力地钻研深度学习。其他高科技行业也一样，但这只是一个开始。





图12-9 我在2015年蒙特利尔NIPS大会上的穿着，一件“全美运动汽车竞赛协会”风格的夹克。会议的赞助商五花八门，从顶级互联网公司到金融和媒体公司，它们都在深度学习领域进行了投资。图片来源：NIPS基金会。

美国在人工智能方面正在失去领先地位，在你阅读这本书的时候，其他国家可能已经超越了我们。多伦多矢量研究所（The Vector Institute in Toronto）于 2017年3月启动，加拿大政府和安大略省政府，多伦多大学以及私营企业对其提供了1.75亿加元的资金支持。[43]矢量研究所的目标是成为世界领先的人工智能研究中心，培养最多的机器学习领域的博士和硕士毕业生，并成为推动多伦多市，安大略省乃至整个加拿大经济发展的AI超级集群的引擎。但加拿大将面临来自中国的激烈竞争，中国正在培训数千名机器学习领域的工程师，神经形态计算是其“大脑项目”的两翼之一。受2017年AlphaGo击败柯洁事件的影响——类似1957年苏联制造的人造卫星对美国的影响，中国已经启动了一项新的价值数十亿美元的人工智能计划，包括雄心勃勃的项目、初创企业和学术研究，目的是到2030年在世界范围内占据重要地位。[44]由于其拥有大量的医疗和个人数据，而且普通民众对隐私的关注相比西方民主国家要少得多，中国可以超越其他那些将私人数据保留在私有孤岛中的国家。中国还将农业和制造业作为数据收集的对象。谁拥有最多的数据，谁就是赢家，中国在数据上的资本实力相当雄厚。

中国也希望“将人工智能集成到导弹中，甚至预测犯罪行为”。[45]而与此同时，美国的政治领导人正计划削减科学和技术资金。在20世纪60年代，美国在太空竞赛中投入了1000亿美元（该数字针对通货膨胀进行了调整），[46]创造了卫星产业，使美国在微电子和材料领域处于领先地位，并在科学和技术的国家实力上做出了政治宣言。当时的投资至今仍在发挥功效。微电子和先进材料是美国仍然具有竞争力的少数几个行业之一。中国在AI竞争方面的大量投资，可能会让它在21世纪的几个关键行业中处于领先地位。这一形势对我们来说是一种警醒。

人工智能正在加速“无形的”信息经济。经济产出的衡量标准是国内生产总值（GDP），即所有商品和服务的总价值。这项指标是为工

业经济设计的，其主要产品和服务是有形的，例如食品、汽车和医疗保健。但是，信息公司中越来越多的价值并不是以这些实体产品进行衡量的。例如，微软拥有的建筑和设备价值为10亿美元，仅占其市场价值的1%。[47]其余价值则是基于软件和微软程序员的专业知识。你愿意为在智能手机上下载的信息出个什么价钱？我们需要一种考虑到各种形式信息价值的新指标，即国内无形资产总值（Gross Domestic Intangibles，简称GDI），以强化GDP来衡量生产率。[48]

目前人工智能的应用还是基于30年前完成的基础研究。30年后的应用将取决于当前正在进行的基础研究，但最优秀和最聪明的研究人员正在为工业界工作，并专注于近期的产品和服务。平衡这一人才流动趋势的是，最聪明最优质的学生现在都进入了机器学习领域。而在上一代人中，类似的人才会进入投资银行业。

在思考人工智能的未来时，我们需要保持目光长远，因为我们远没有具备达到人类智能水平所需的计算能力。现在，深度学习网络拥有数百万个单元和数十亿个权重。这比人类大脑皮层中的神经元和突触的数量还要少1万倍——人类1立方毫米的皮层组织就包含了10亿个突触。如果世界上的所有传感器都连接到互联网，并通过深度学习网络相互连接，那么有一天，互联网可能会醒来并主动说出：

“Hello，world!”（你好，世界！）[49]

---

[1] Kaggle网站拥有数百万名数据科学家，他们互相竞争，凭借最佳表现赢得奖项。Cade Metz, “Uncle Sam Wants Your Deep Neural Networks,” New York Times,June 22, 2017, https://www.nytimes.com/2017/06/22/technology/homelandsecurity-artificial-intelligence-neural-network.html.

[2] 我的演讲视频“Cognitive Computing: Past and Present”可以在网上找到：https://www.youtube.com/watch?v=OBDMQuphd-Q。

[3] 参阅Jen Clark, “The Countdown to IBM’s IoT, Munich,” IBM Internet of Things(blog), posted February 8, 2017.

https://www.ibm.com/blogs/internet-of-things/countdown-ibms-iot-hq-munich/。

[4] BRAIN的报告提出了建议，并为创新型技术确定了优先事项，以协助推动我们对神经回路和行为的理解：BRAIN Working Group, BRAIN 2025: A Scientific Vision, Report to the Advisory Committee to the Director, NIH (Bethesda, MD:National Institutes of Health, June 5, 2014), https://www.braininitiative.nih.gov/pdf/BRAIN2025_508C.pdf。

[5] 参阅K. S. Kosik, T. J. Sejnowski, M. E. Raichle, A. Ciechanover, and D. A.Baltimore, “A Path toward Understanding Neurodegeneration,” *Science*353, no.6302 (2016): 872–873。

[6] 在2002年的科幻电影《少数派报告》中，由汤姆·克鲁斯饰演的男主角通过非法的眼移植避免了政府的追踪。

[7] 参阅Nandan Nilekani and Viral Shah, Rebooting India: Realizing a Billion Aspirations. (Gurgaon: Penguin Books India. 2015)。

[8] Nandan Nilekani, as quoted in Andrew Hill, “Nandan Nilekani, Infosys, on Rebooting India,” *Financial Times*, January 22, 2017, https://www.ft.com/content/058c4b48-d43c-11e6-9341-7393bb2e1b51?mhq5j=e1.

[9] 参阅M. Gymrek, A. I. McGuire, D. Golan, E. Halperin, and Y. Erlich, “Identifying Personal Genomes by Surname Inference,” *Science*. 339, no. 6117 (2013): 321–324。

[10] 参阅M. Wilson, “Six Views of Embodied Cognition,” *Psychonomic Bulletin &Review*9, no. 4 (2002): 625–636。

[11] P. Ruvolo, D. Messinger, and J. Movellan, “Infants Time Their Smiles to Make Their Moms Smile,” *PLoS One*10, no. 9 (2015): e0136492.

[12] Abigail Tucker, “Robot Babies,” *Smithsonian Magazine*, July 2009, https://www.smithsonianmag.com/science-nature/robot-babies-30075698/; Tiffany Fox, “Machine Perception Lab Seeks to Improve Robot Teacher with Intelligent Tutoring Systems,” UCSanDiego News Center, July 30, 2008, http://ucsdnews.ucsd.edu/newsrel/general/07-08RobotTeachers.asp. 还可参阅F. Tanaka, A.Cicourel, and J. R. Movellan, “Socialization between Toddlers and Robots at an Early Childhood Education Center,” *Proceedings of the National Academy of Sciences of the United States of America*104, no. 46 (2008): 17954–17958。

[13] “Conserve Elephants. They Hold a Scientific Mirror Up to Humans,” *Economist*,June 17, 2017, 72–74. http://www.economist.com/news/science-

andtechnology/21723394-biology-and-conservation-elephants-conserve-elephantsthey-hold.

[14] 参阅R. A. Brooks, “Elephants Don’t Play Chess,” Robotics and Autonomous Systems 6, no. 1 (1990): 3–15.

[15] Diego San由Kokoro Co. 在日本打造，后缀“san”（先生）在日语中是一个尊称。

[16] 参考YouTube视频“Diego Installed,” https://www.youtube.com/watch?v=knRy DcnUc4U。机器人的脸部由大卫·汉森和Hanson Robotics公司打造。

[17] 参阅Paul Ekman, Thomas S. Huang, and Terrence J. Sejnowski, eds., Final Report to NSF of the Planning Workshop on Facial Expression Understanding July 30–August 1, 1992, http://papers.cnl.salk.edu/PDFs/Final%20Report%20To%20NSF%20of%20the%20Planning%20Workshop%20on%20Facial%20Expression%20Understanding%201992-4182.pdf。

[18] 参阅J. Gottman, R. Levenson, and E. Woodin,“Facial Expressions during Marital Conflict,” Journal of Family Communication 1, no. 1 (2001): 37–57.

[19] 参阅F. Donato, M. Stewart Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski,“Classifying Facial Actions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*21, no. 10 (1999): 974–989。

[20] 参阅G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, “The Computer Expression Recognition Toolbox (CERT),” 2011 IEEE International Conference on Automatic Face and Gesture Recognition,Santa Barbara, California. http://mplab.ucsd.edu/wp-content/uploads/2011LittlewortEtAl-FG-CERT.pdf。

[21] 焦点小组也称焦点团体、焦点群众，就某一议题，通过采访一个群体以获取其观点和评价。——译者注

[22] 参阅A. N. Meltzoff, P. K. Kuhl, J. Movellan, and T. J. Sejnowski, “Foundations for a New Science of Learning,” Science 325, no. 5938 (2009): 284–288。

[23] 参阅A. A. Benasich, N. A. Choudhury, T. Realpe-Bonilla, and C. P. Roesler,“Plasticity in Developing Brain: Active Auditory Exposure Impacts Prelinguistic Acoustic Mapping,” Journal of Neuroscience 34, no. 40 (2014): 13349–13363。

[24] 参阅J. Whitehill, Z. Serpell, Y. Lin, A. Foster, and J. R. Movellan, “The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions,” *IEEE Transactions on Affective Computing*5, no. 1 (2014): 86–98。使用机器学习来自动记录学生面部表情的工作离不开团队的努力，该团队成员还包括Gwen Littlewort、Linda Salamanca、Aysha Foster和Judy Reilly。

[25] 参阅R. V. Lindsey, J. D. Shroyer, H. Pashler, and M. C. Mozer, “Improving Students’ Long-Term Knowledge Retention through Personalized Review,”Psychological Science 25, no. 3 (2014): 630–647。

[26] B. A. Rogowsky, B. M. Calhoun, and P. Tallal, “Matching Learning Style to Instructional Method: Effects on Comprehension,” *Journal of EducationalPsychology*107, no. 1 (2015): 64–78. 关于这一主题的研讨会视频：https://www.youtube.com/watch?v=p-WEcSFdoMw。

[27] International Convention on the Science of Learning (Science of Learning: How can it make a difference? Connecting Interdisciplinary Research on Learning to Practice and Policy in Education) Shanghai, 1-6 March 2014 Summary Report.https://www.oecd.org/edu/ceri/International-Convention-on-the-Science-ofLearning-1-6-March-2014-Summary-Report.pdf.

[28] 参阅B. Bloom, “The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring,” *Educational Researcher*13, no.6 (1984): 4–16。

[29] John Markoff, “Virtual and Artificial, but 58,000 Want Course,” *New York Times,*August 15, 2011, http://www.nytimes.com/2011/08/16/science/16stanford. html.

[30] 我最喜欢的一封信，来自一个五年级的学生：

2015年2月2日

亲爱的教授们，我参加了期末考试，考得很好。我在读五年级。我妈妈正在浏览Coursera上的课程，我恳求她让我加入。她为我选择了这门课程，我对此很感激。我从来不知道教授们如此机智，整个课程都很有趣。当然，我必须借助词典来查询课程中使用的科学术语，但这是一个很好的学习体验。通过使用呼吸法，我学会了克服进入考场时的胃部不适。真的很管用！

现在我已经学会了将考试看作了解我学到并记住了多少东西的工具。我喜欢番茄工作法。我的妈妈在整个过程中都在扮演反面角色。她的大脑里塞满了视频讲座（她只顾着狂热地看视频），没有进行适度的休息，甚至在我参加期末考试的前一天晚上都没有睡觉。虽然她比我聪明，但她做得不好。我很惊讶，简单的技巧就能帮助我得到更高的分数，而且没有任何考试的压力。

非常感谢，教授们。我希望你们能推出本课程的第二部分。

周一快乐。

苏珊

[31] Barbara Oakley and Terrence Sejnowski, *Learning How to Learn: How to Succeedin School Without Spending All Your Time Studying; A Guide for Kids and Teens*(New York: Tarcher Perigee, Penguin Books, August 7, 2018).

[32] 作者指的是升级版的纳米学位（Nanodegree Plus）。近期优达学城的课程政策出现变动，费用有所上调，并于2017年12月5日起不再提供升级版的纳米学位项目。——译者注

[33] “Udacity’s Sebastian Thrun: ‘Silicon Valley has an obligation to reach out to all of the world,’” Financial Times, November 15, 2017. https://www.ft.com/content/51c47f88-b278-11e7-8007-554f9eaa90ba/.

[34] Barbara Oakley, *Mindshift: Break through Obstacles to Learning and Discover Your Hidden Potential*(New York: Penguin Random House, 2017).

[35] 参阅D. Bavelier and C. S. Green, “The Brain-Boosting Power of Video Games,” *Scientific American*315, no 1 (2016): 26–31。

[36] G. L. West, K. Konishi, and V. D. Bohbot, “Video Games and HippocampusDependent Learning,” *Current Directions in Psychological Science*26. no. 2(2017):152– 158.

[37] J. A. Anguera, J. Boccanfuso, J. L. Rintoul, O. Al-Hashimi, F. Faraji, J. Janowich,et al., “Video Game Training Enhances Cognitive Control in Older Adults,” *Nature*501, no. 7465 (2013): 97–101.

[38] J. A. Anguera, J. Boccanfuso, J. L. Rintoul, O. Al-Hashimi, F. Faraji, J. Janowich,et al., “Video Game Training Enhances Cognitive Control in Older Adults,” *Nature*501, no. 7465 (2013): 97–101.

[39] 参阅IES: What Works Clearinghouse, *Beginning Reading Intervention Report:Fast ForWord*(Washington, DC: U.S. Department of Education, Institute of Education Sciences, 2013). https://ies.ed.gov/ncee/wwc/Docs/InterventionReports/wwc_ffw_031913.pdf。

[40] 参阅J. Deveau, D. J. Ozer, and A. R. Seitz “Improved Vision and On-Field Performance in Baseball through Perceptual Learning,” *Current Biology*24, no. 4(2014): R146–147。

[41] 参阅Federal Trade Commission press release “FTC Charges Marketers of ‘Vision Improvement’ App with Deceptive Claims,” September 17, 2015, https://www.ftc.gov/news-events/press-releases/2015/09/ftc-charges-

marketers-visionimprovement-app-deceptive-claims/。赛茨（Seitz）的科学研究质量很高，已经发表在同行评审的心理学期刊上，但联邦贸易委员会（FTC）希望他能进行随机对照试验，类似用于测试药物疗效的试验。这是一项昂贵的工作，对于一个小型的创业公司来说不太现实。

[42] Johana Bhuiyan, “Ex-Google Sebastian Thrun Says That the Going Rate for SelfDriving Talent Is $10 Million per Person,” *Recode*, September 17, 2016. https://www.recode.net/2016/9/17/12943214/sebastian-thrun-self-driving-talent-pool.

[43] 杰弗里·辛顿是Vector Institute的首席科学指导。参阅http://vectorinstitute.ai/。

[44] Paul Mozur and John Markoff, “Is China Outsmarting America in A.I.?” *New YorkTimes*, May 27, 2017, https://www.nytimes.com/2017/05/27/technology/chinaus-ai-artificial-intelligence.html.

[45] Paul Mozur, “Beijing Wants A.I. to Be Made in China by 2030,” *New York Times*,July 20, 2017. https://www.nytimes.com/2017/07/20/business/china-artificialintelligence.html.

[46] 参阅Mike Wall, “JFK’s ‘Moon Speech’ Still Resonates 50 Years Later,” *Space.com*(blog), posted September 12, 2012, https://www.space.com/17547-jfk-moonspeech-50years-anniversary.html. 1962年9月12日，美国总统约翰·肯尼迪在休斯敦的莱斯大学发表了演讲“我们选择飞向月球”。50多年后，该演讲仍然令人心潮澎湃，让我们不由得回忆起他领导的年代。参看https://www.youtube.com/watch?v=WZyRbnpGyzQ/. 1969年7月20日，当宇航员尼尔·阿姆斯特朗登上月球时，美国国家航空航天局（NASA）工程师的平均年龄为26岁；1962年，这些工程师还在学校的时候受到了肯尼迪这番讲话的鼓舞。

[47] J.Haskel and S. Westlake, Capitalism without Capital: The Rise of the Intangible Economy (Princeton, NJ: Princeton University Press, 2017), 4.

[48] W. Brian Arthur, “The second economy?” *McKinsey Quarterly*October, 2011.https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/ourinsights/the-second-economy/.

[49] “Hello, world!”是一个样例程序的测试信息，来自Brain Kernighan和Dennis Ritchie的经典教材“*The C Programming Language*”(Englewood Cliffs, NJ:Prentice Hall, 1978)。

13
算法驱动

2016年6月，在新加坡南洋理工大学举办的21世纪科学挑战大会上，我参与了为期一周的讨论。讨论的话题范围很广，从宇宙论、进化论到科学政策。[1] W. 布赖恩·阿瑟（W. Brian Arthur）是一个对技术很感兴趣的经济学家，[2] 他指出，在过去，技术是由物理定律驱动的：20世纪，我们试图用微分方程和连续变量的数学原理去理解物理世界，这些变量在时间和空间上平滑地变化。相比之下，今天的技术是由算法驱动的：在21世纪，我们试图通过离散数学和算法来理解计算机科学和生物学复杂的本质。阿瑟是新墨西哥州圣塔菲研究所（the Santa Fe Institute）的教员，该机构是20世纪涌现出的许多研究复杂系统的中心之一。[3]

算法无处不在。我们每次使用谷歌搜索时都使用了算法。[4] 每次在Facebook上阅读的新闻推送也都经过了算法的自动筛选，这些新的推送参考了我们的阅读历史记录，会影响我们情绪的反应。[5] 随着经过深度学习训练的语音识别和自然语言能力被嵌入我们的手机中，算法正在以越来越快的速度侵入我们的生活。

算法是在执行计算或解决问题时，遵循一组包含离散步骤或规则的过程。“算法”（algorithm）这个词来源于拉丁文“algorismus”，以9世纪的波斯数学家al-Khwarizmi的名字命名，在17世纪受到希腊语中“arithmos”（意为“数字”）这个词的影响，最终由“algorism”转变为“algorithm”。算法虽然有着古老的起

源，但其地位直到最近才通过数字计算机提升到了科学和工程领域的前沿。

用算法把复杂问题简单化

20世纪80年代，解决复杂度问题的新方法开始萌芽，其目标是寻找新的途径来理解在生物体中发现的系统，这些系统比物理和化学系统还要复杂。与遵循艾萨克·牛顿运动定律的火箭运动的简单性不同，并没有一种简易的方式来描述树木如何生长。于是，计算机算法被一大群传奇的先驱用来探索这些古老的生命体问题。

斯图尔特·考夫曼（Stuart Kauffman）是一名医生，对基因网络很感兴趣。基因网络中被称为“转录因子”的蛋白质能定位基因，并影响它们是否被激活。[6]他的模型是自组织的，基于二元单元网络，这个网络和神经网络在某些方面很相似，但速度要慢得多。克里斯托弗·兰顿（Christopher Langton）在20世纪80年代后期创造了“人造生命”这个术语，[7]由此引发了一系列的尝试，旨在理解活细胞的复杂性和复杂行为形成发展的原理。尽管我们在细胞生物学和分子遗传学方面所取得的进展，已经能够揭示细胞内分子机制高度演化的复杂性，但生命的奥秘对我们来说仍然是个不解之谜。

算法为创造出比现有世界更为复杂的世界提供了新的机会。事实上，20世纪发现的算法确实让我们重新思考了复杂性的本质。20世纪80年代的神经网络革命，就是由类似的理解大脑复杂性的尝试所驱动的。虽然我们的神经网络模型比大脑的神经回路简单得多，但我们开发的学习算法使得探索一般原理成为可能，例如信息在大量神经元中的分布原理。不过，网络功能的复杂性是如何由相对简单的学习规则产生的呢？是否存在一种更简单的系统，既体现了复杂性，又更易于分析呢？

理解、分析复杂系统

另一个以科学的严肃态度对待复杂性的传奇人物是斯蒂芬·沃尔夫勒姆（见图13－1）。他年轻时是个神童，20岁就从加州理工学院获得了物理学博士学位，创造了该学位最年轻毕业生的纪录，并于1986年在伊利诺伊大学成立了复杂系统研究中心。沃尔夫勒姆认为神经网络过于复杂，于是决定转变研究方向，开始探索细胞自动机。

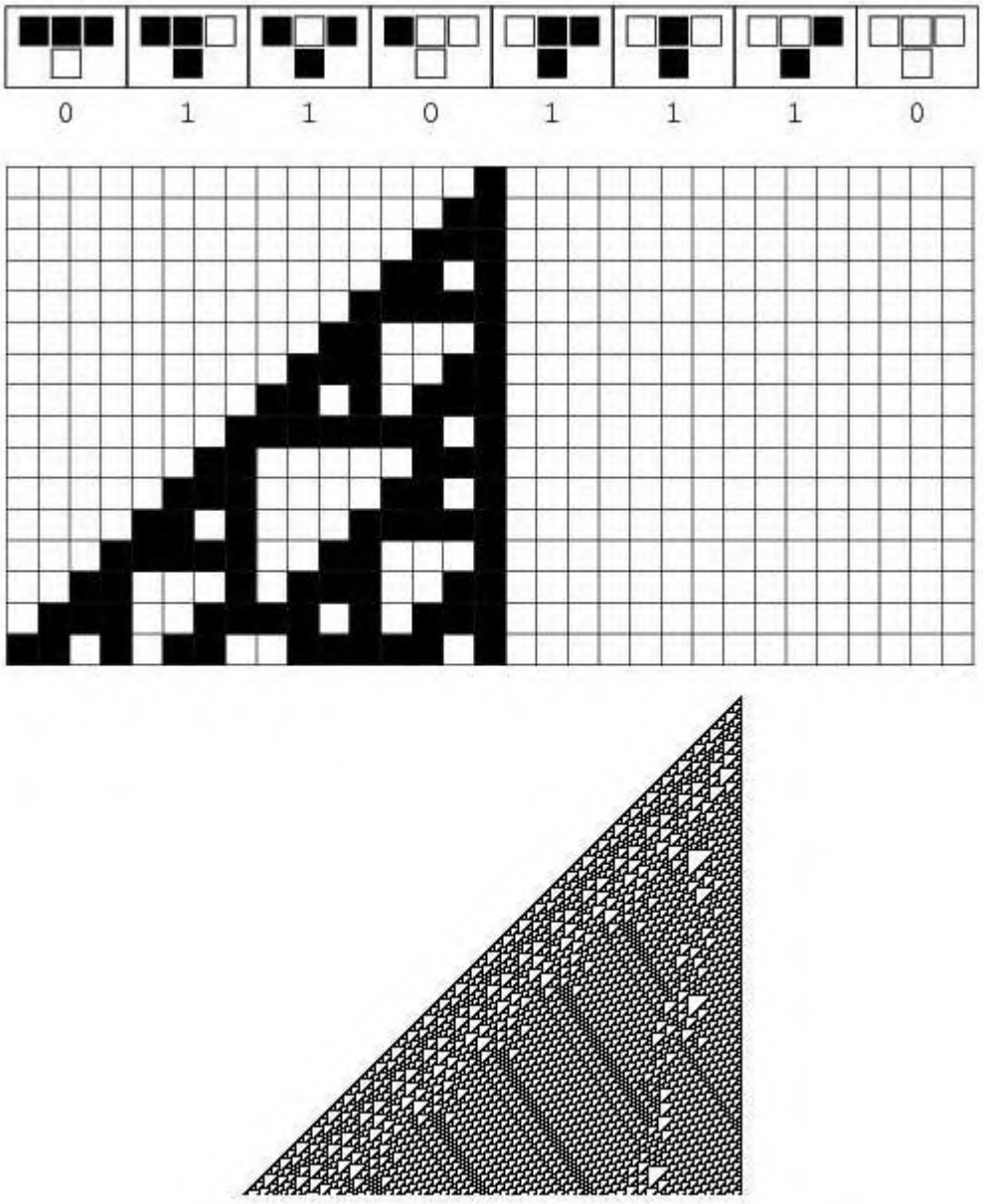
细胞自动机通常只有少数几个随时间变化的离散值，这取决于其他细胞的状态。最简单的一种细胞自动机，是一组一维的细胞阵列，每个细胞的值为0或1（见方框13.1）。“生命的游戏”（Game of Life）也许是最著名的细胞自动机了，它是由普林斯顿大学的数学教授约翰·康威（John Conway）在1968年发明的，并由马丁·加德纳（Martin Gardner）通过他在杂志《科学美国人》（*Scientific American*）的“数学游戏”专栏中推广开来（见图13－2）。整个空间是一个二维的细胞阵列，细胞只有“开”或“关”的状态，对规则的更新仅取决于4个最近的相邻细胞。在每个时间步骤中，所有细胞的状态都会更新。阵列中会生成复杂的模式，其中一些模式还有自己的名字，例如“滑翔机”，它会横跨阵列并与其他模式发生碰撞。初始条件对于找到显示复杂模式的配置至关重要。



图13-1 斯蒂芬·沃尔夫勒姆在他位于马萨诸塞州康科德的家中，站在由算法生成的地板上。沃尔夫勒姆是复杂性理论的先驱，他曾经提出，即使是简单的程序也会生成现实级别的复杂性。图片来源：斯蒂芬·沃尔夫勒姆。

生成复杂系统的规则有多普遍呢？沃尔夫勒姆想要了解可能导致复杂行为的最简单的细胞自动机规则，于是他开始从所有规则中寻找。规则0~29产生的模式总是会回归枯燥的行为：所有的细胞都会以重复模式或嵌套分形模式结束。但是，规则30产生了展开模式，规则110不断变化并演变出了复杂模式（见方框13.1）。[8]最终证明，规则110能够实现通用计算；也就是说，所有可能的细胞自动机中，最简单的一些已经具备了可以计算任何可计算函数的图灵机的能力，所以它在原则上与任何计算机一样强大。

13.1
细胞自动机



规则110。细胞自动机的运行规则是：细胞的颜色依据自身和直接相邻的细胞的颜色而改变。例如，对于图片顶部3个细胞的黑色和白色的8种可能组合，规则110指定了其下方细胞的颜色。这条规则的演变能够每次对一行细胞进行更新，从第一行的单个黑色细胞开始，更新15次后，就产生了中间的那幅图，而更新了250次后，则产生了底部的那幅图。简单的初始条件演变成了一个无限延续的非常复杂的模式。

这其中的复杂性来自哪里？想了解更多细节，请参阅：http://mathworld.wolfram.com/Rule110.html。



图13-2 生命的游戏。一个Gosper滑翔机枪（图片顶部）的快照，它发出一系列按对角线飞行的滑翔机图标，从“母船”（顶部）飞往右下角。图片来自维基百科：Gun（cellular automaton），有一个滑翔机枪运动的GIF动画。

该发现的一个意义在于，我们在生命体中发现的显著复杂性，可以通过对分子间化学相互作用的最简空间进行采样，从而实现演化。而演化过程中出现的复杂的分子组合，应该是能够预料到的，不应该被当作是什么奇迹。但细胞自动机可能不是早期生命的优秀模型，究竟哪些简单的化学系统能够产生复杂的分子结构，仍然是一个悬而未决的问题。[9]也许只有特殊的生物化学系统才有这种特性，这有助于减少可能产生生命的潜在相互作用种类的数量。

生命的一个重要特性，是细胞的自我复制能力。匈牙利裔美国数学家约翰·冯·诺依曼，曾经于20世纪40年代在普林斯顿高级研究所使用细胞自动机探索过这种能力。冯·诺依曼的工作对数学的许多领域都产生了重大影响，特别是他关于博弈论的开创性工作（在第1章中提到过）。在寻找可以完全自我复制的最简单的细胞自动机过程中，冯·诺依曼发现了一个复杂的细胞自动机，具有29个内部状态和一个大型的记忆体，可以实现自我复制。[10]该发现具有重大的生物学意义，因为能够自我复制的细胞中，也有许多内部状态和记忆以DNA的形式存在。从那以后，更简单的、能够自我复制的细胞自动机陆续被发现。

大脑的逻辑深度

1943年，沃伦·麦卡洛克（Warren McCulloch）和沃尔特·皮茨（Walter Pitts）证明了，用类似感知器的简单二进制阈值单元构建一台数字计算机是可行的，只要将感知器按照计算机的基本逻辑门那样连接起来就可以了。[11]我们现在知道，大脑有模拟和数字的混合特性，它们的神经回路通常不计算逻辑函数。但是麦卡洛克和皮茨于1943年发表的论文在当时受到了很多关注，特别是启发了冯·诺依曼对计算机的思考。他创建了第一批有存储程序的数字计算机，这对于当时的数学家来说是一个很不寻常的项目，尽管当冯·诺依曼于1957年去世时，高级研究所并没有继续他的研究，还销毁了他的计算机。[12]

冯·诺依曼对大脑也非常感兴趣。1956年，作为耶鲁大学举办的西利曼（Silliman）讲座嘉宾，[13]他在讲座中思考了大脑如何用这种不可靠的组件进行可靠操作的问题。当数字计算机中的晶体管出现错误时，整个计算机可能就会崩溃，但是当大脑中的某个神经元失灵时，大脑的其他部分却会适应这种失灵，并照常工作。冯·诺依曼认为，冗余（redundancy）可能是大脑能够稳定运转的原因，因为每一个操作都有许多神经元参与。冗余在传统上是基于备份的，以防主系统出现故障。但我们现在知道，大脑的冗余是基于多样性，而不是重复。冯·诺依曼对逻辑深度也表示了担忧：在累计误差破坏结果之前，大脑可以完成多少逻辑步骤？与可以完美地执行每个逻辑步骤的计算机不同，大脑具有许多噪声源。一个大脑可能达不到完美的程度，但是因为它有许多神经元并行工作，所以每一步能够完成的计算都比一台计算机进行一次计算所能完成的要多得多，因而逻辑深度也要浅一些。

尝试所有可能的策略

想象一下充满了所有可能算法的空间。这个空间中的每一点都是一个算法，可以做某件事，而且其中一些算法是非常有用且高效的。过去，这些算法是由数学家和计算机科学家们手工创造的，就像行会里的工匠那样。斯蒂芬·沃尔夫勒姆通过穷举搜索（exhaustive search）对搜寻细胞自动机算法的过程进行了自动化，从最简单的自动机开始，其中一些产生了高度复杂的模式。沃尔夫勒姆定律（Wolfram’slaw）是对这种见解的一个总结。该定律指出，你不必在算法空间中搜索很长一段路径，就可以找到一种能解决一类有趣问题的算法。这就像使用机器人程序在互联网上玩类似《星际争霸》的游戏一样，尝试所有可能的策略。根据沃尔夫勒姆定律，算法空间中应该有大量算法可以赢得游戏。

沃尔夫勒姆专注于细胞自动机的空间，这是所有可能算法空间中一个小的子空间。但是，如果细胞自动机是非典型算法，比其他类算法更具有普遍性呢？我们现在已经在神经网络空间中确认了沃尔夫勒姆定律。每个深度学习网络都是通过学习算法找到的，该算法是一种发现新算法的元算法。对于大型网络和大量数据，从不同的起始状态学习可以产生如星系般庞大的网络数量，它们在解决问题上都能达到相同效果。这就产生了一个问题，即是否存在一种比梯度下降更快的方法能够找到算法空间的区域。梯度下降这种方法速度慢，还需要使用大量的数据。有一个暗示揭示了这种可能性，即每个生物种类，都是由生命算法空间中的一个点附近的变体DNA序列创建的个体云。自然界通过自然选择设法从一个“个体云”跳跃到另一个云，这种跳跃式过程被称为“间断性平衡”（punctuated equilibria）[14]，同时每个物种也经历着随机突变的局部搜索。遗传算法被设计用来进行这样的跳跃，大致是基于自然界如何演化出新的有机体[15]。我们需要用数学来描述这些算法云。可谁知道算法的空间是什么样子呢？还有很多我们尚未发现的算法“星系”，但我们可以通过自动探索的方式来找到这些星系——终极边境。

我实验室的博士后研究员克劳斯·施蒂费尔（Klaus Stiefel）跟进了这一过程中的一个简单例子。他在2007年使用了一种算法，在计算机中生成了具有复杂的树突树（dendritic trees）的模型神经元。[16]树突就像天线一样从其他神经元收集输入。可能的树突树的空间是巨大的，而施蒂费尔的目标是指定所需的功能，并在树突树的空间中搜索能够计算出该功能的模型神经元。一个有效的功能是决定输入放电尖峰的到达顺序：当一个特定的输入在另一个输入之前到达时，神经元应该输出一个放电尖峰，但是如果该输入随后到达，神经元应该保持沉默。这种模型神经元是通过使用遗传算法搜索所有可能的树突树发现的，且该解决方案看起来就像皮层锥体神经元，一个突触位于底部的薄树突（基底树突），另一个突触位于神经元顶部的厚树突（顶树突；见图14－6）。这也许解释了为什么锥体细胞有顶端和基底的树突。如果没有对所有可能突触的空间进行深入搜索，我们可能就无法将这种结构和上述功能联系起来。通过从其他功能开始重复此搜索，可以自动编译以树突形状列出相应功能的目录，并且当发现新的神经元时，可以在目录中通过其树突形状查找其潜在的功能。

斯蒂芬·沃尔夫勒姆离开学术界后创立了沃尔夫勒姆研究所，后者开发了Mathematica程序，该程序支持大量的数学结构，并被广泛应用于实际。Mathematica是用沃尔夫勒姆语言编写的，这是一种通用的融合多种范式的编程语言。该语言也为沃尔夫勒姆阿尔法（Wolfram Alpha）提供支持，它是第一个实用的、基于符号方法的关于大千世界的问答系统。[17]学术领域的货币，是发表的文章。如果你是一个自立的绅士派科学家，你就有能力绕过简短的文章，出版有足够空间来彻底探索新领域的书籍。这是多个世纪以来的常态，只有富人或富人赞助的学者，才有条件成为科学家。

沃尔夫勒姆在2002年写了《一种新的科学》（*A New Kind of Science*）一书。[18]该书重5.6磅，长达1280页，其中有348页记录了相当于100份新科学论文的附录。这本书被媒体大肆宣传，但引发了

复杂系统界褒贬不一的评论，其中一些人认为其中并没有完整地引用他们的工作。这种反对意见并没有抓住此书的要点，即把以前的工作置于新的语境下。卡尔·林奈（Carolus Linnaeus）开发了一种现代化的植物和动物分类法，即“二名法”（Binomial Nomenclature），如大肠杆菌（Escherichia coli）。该命名法是达尔文进化论的重要先导，为早期的分类学做了铺垫。沃尔夫勒姆开辟的这一研究方向在新一代研究人员中已经有了一批追随者。

在20世纪80年代，沃尔夫勒姆对神经网络可能对现实世界产生的影响表示怀疑，而且事实上，这些神经网络在之后30年的确没有产生太大的影响。然而，过去5年的进步已经改变了这一点：沃尔夫勒姆和其他许多研究人员承认他们低估了网络的成就。[19]但是谁能预测神经网络的性能会扩展到什么程度呢？支持Mathematica的沃尔夫勒姆语言现在也支持深度学习应用，其中一个应用，是最早对图像中的对象提供在线识别技术的应用程序。[20]

是斯蒂芬把我介绍给了比阿特丽斯·哥伦布。1987年我去圣迭戈时，比阿特丽斯正在加州大学圣迭戈分校攻读博士学位。他打电话说，他的朋友比阿特丽斯会参加我的PDP讲座（之后又分别给我们打电话询问事情的进展）。几年后，我搬到圣迭戈，后来又和比阿特丽斯订了婚。1990年我们在加州理工学院举行完婚礼后，去了贝克曼礼堂参加了一个婚礼研讨会，比阿特丽斯身着婚纱，发表了一个演讲——《婚姻：理论与实践》（Marriage: Theory and Practice）。斯蒂芬自信满满，不无骄傲地介绍了他当时是如何介绍我们认识的。不过比阿特丽斯提出，如果他想邀功，也必须承担相应的责任，对此，他谨慎地表示了反对。

---

[1] “21世纪科学挑战大会”上讨论者的演讲视频可以在这里找到：https://www.youtube.com/results?search_query=Grand+Challenges++for+Science+in+the+21 st+Century。

[2] 参阅W. Brian Arthur, *The Nature of Technology: What It Is and How It Evolves*(New York: Free Press, 2009)。

[3] George A. Cowan, *Manhattan Project to the Santa Fe Institute: The Memoirs of George A. Cowan*(Albuquerque: University of New Mexico Press, 2010).

[4] 谷歌的PageRank算法是谷歌创始人拉里·佩奇（Larry Page）和谢尔盖·布林（Sergey Brin）发明的，它使用连接到网页的链接数目，依据其在互联网上的重要性来对网页进行排序。该算法已经被演化成了许多算法层来应对搜索偏见。

[5] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks," *Proceedings of theNational Academy of Sciences of the United States of America*111, no. 24 (2014):8788–8790.

[6] Stuart Kauffman, *The Origins of Order: Self Organization and Selection in Evolution*(New York: Oxford University Press, 1993).

[7] Christopher G. Langton, ed., *Artificial Life: An Overview*(Cambridge, MA: MIT Press, 1995).

[8] Stephen Wolfram, *A New Kind of Science*(Champaign, IL: Wolfram Media, 2002).

[9] National Research Council, *The Limits of Organic Life in Planetary Systems*(Washington, DC: National Academies Press, 2007), chap. 5, "Origin of Life," 53–68. https://www.nap.edu/read/11919/chapter/7.

[10] Von Neumann, J. and A. W. Burks, *Theory of Self-Reproducing Automata*. (Urbana,IL: University of Illinois Press, 1966). 也可参阅维基百科：Von "Neumann universal constructor."

[11] W. S. McCulloch, and W. H. Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity," *Bulletin of Mathematical Biophysics*5 (1943): 115–133.

[12] 这台电脑被命名为"JOHNNIAC"，以呼应另一台名叫"ENIAC"的早期数字计算机。

[13] 他为这些讲座准备的材料后来催生了冯·诺依曼的著作《计算机与大脑》（*The Computer and the Brain*）(New Haven: Yale University Press, 1958).

[14] Stephen Jay Gould and Niles Eldredge, "Punctuated Equilibria: The Tempo And Mode Of Evolution Reconsidered." *Paleobiology*3, no. 2 (1977): 115–151, 145;John Lyne and Henry Howe, "Punctuated Equilibria': Rhetorical

Dynamics of a Scientific Controversy," *Quarterly Journal of Speech*, 72, no. 2 (1986): 132–147.doi:10.1080/00335638609383764.

[15] John H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*(Cambridge, MA: MIT Press, 1992).

[16] K. M. Stiefel and T. J. Sejnowski, "Mapping Function onto Neuronal Morphology,"*Journal of Neurophysiology*98, no. 1 (2007): 513–526.

[17] 参阅Wolfram Alpha computational knowledge engine website, https://www.wolframalpha.com/。

[18] Stephen Wolfram, *A New Kind of Science*(Champaign, IL: Wolfram Media, 2002).

[19] 有一篇关于沃尔夫勒姆思想演变的文章很有趣，可以参阅Stephen Wolfram,"A New Kind of Science: A 15-Year View," *Stephen Wolfram Blog*, posted May 16, 2017. http://blog.stephenwolfram.com/2017/05/a-new-kind-of-science-a-15year-view/。

[20] 参阅Stephen Wolfram, "Wolfram Language Artificial Intelligence: The Image Identification Project," Stephen Wolfram Blog, posted May 13, 2015. http://blog.stephenwolfram.com/2015/05/wolfram-language-artificial-intelligence-theimage-identification-project/。

14
芯片崛起

我们正在目睹电脑芯片行业一个新格局的诞生。该领域的竞争主要在于如何设计和制造新一代的芯片，能够运行学习算法——不管是深度学习、强化学习还是其他的学习算法——比在通用计算机上的模拟学习算法快上几千倍，能耗也更低。新的超大规模集成电路芯片采取并行处理结构，带有内存，能够缓解在过去50年里主导计算的顺序冯·诺依曼构架中内存和中央处理器之间的瓶颈。在硬件层面，我们还在探索阶段。每种具有特殊用途的超大规模集成电路芯片都有不同的优点和局限性。运行人工智能应用的大型网络需要巨大的运算能力，因此，构建高效的硬件有着巨大的赢利空间。

主要的电脑芯片公司和初创公司都在开发深度学习芯片上投入了大量资金。比如，2016年，英特尔用4亿美元并购了Nervana，这是一家来自圣迭戈的初创公司，主营设计深度学习的专用超大规模集成电路芯片。Nervana前CEO纳维恩·饶（Naveen Rao）现在负责领导英特尔新设立的AI产品部，直接报告给英特尔的CEO。2017年，英特尔用153亿美元并购了Mobileye，该公司专注于生产自动驾驶汽车的传感器以及计算机视觉系统。英伟达（Nvidia）开发了能够优化图形应用程序和游戏的专用数字芯片，称为图形处理器（graphics processingunits，GPUs），目前正在销售更多为深度学习和计算设计的专用芯片。而谷歌则设计生产了一种更为高效的专用芯片——张量处理单元（tensor processing unit，TPU），以助力为其互联网服务的深度学习。

不过研发专用软件和发展深度学习应用同样重要。谷歌开源了它的深度学习项目TensorFlow，尽管该做法并没有看起来那么无私。比如让安卓系统免费开源，就给了谷歌对全世界绝大多数智能手机操作系统的控制权。现在除了TensorFlow，还有其他一些开源选择，比如微软的CNTK，亚马逊和其他大型互联网公司支持的MVNet，以及其他深度学习程序，比如Caffe、Theano和PyTorch。

神经形态芯片

2011年，我在挪威的特罗姆瑟组织举办了由Kavli基金会赞助的“绿色环境中高性能计算的发展”（Growing High Performance Computing in a Green Environment）研讨会。[1]我们估计，使用当前的微处理器技术，百亿亿次级（exascale）计算［比千万亿次级（petascale）计算强大1000倍］需要50兆瓦的功率，比运行纽约市地铁所需的功率还多。因此，下一代超级计算机可能不得不使用像英国跨国半导体公司Arm Holdings（ARM）为手机开发和优化的那种低功耗芯片。很快，在大多数计算密集型应用中使用通用数字计算机将不再可行，专用芯片将占主导地位，因为它们已经被嵌入手机中了。

人类大脑中有大约1000亿个神经元，每个神经元都与其他数千个神经元相连接，总计达1000万亿个（10^{15}）突触连接。大脑运转所需的功耗大约是20瓦，占整个身体运转所需功率的20%，尽管大脑仅占身体质量的3%。相比之下，一台远不如大脑强大的千万亿次级超级计算机，功耗却为5兆瓦，是大脑功耗的25万倍。大自然是怎么创造出这一高效奇迹的呢？首先，神经元接收和发送信号的部分被微缩至分子水平。另外，神经元是在三维空间上连接的（微芯片表面的晶体管仅在二维平面上相互连接），这样就可以使所需空间最小化。由于大自然很久以前就进化出了这些技术，想要追赶大脑的能力，我们还有很多工作要做。

深度学习是高度计算密集型的，该过程目前在中央服务器上完成，计算结果会被传送到手机等周边设备。最终，周边设备应该是自主运行的，这就需要完全不同的硬件——比云计算轻得多，耗电也更少。幸运的是，这样的硬件已经存在了，即受大脑启发设计出的神经形态芯片。

视网膜芯片

1983年，我在匹兹堡郊外一个度假村里举行的研讨会上第一次见到了卡弗·米德（Carver Mead）（见图14－1）。那时杰弗里·辛顿创建了一个小组来探索神经网络研究的发展趋势。米德因其在计算机科学领域的重要贡献而闻名。他第一个认识到，随着大规模集成电路芯片上的晶体管变得越来越小，芯片将越来越高效，计算能力也应该会持续增长很长一段时间。他还创造了“摩尔定律”这个术语，根据戈登·摩尔（Gordon Moore）的观察，芯片上的晶体管数量每18个月会翻一番。1981年，卡弗发明了硅编译器，从此成为业界传奇。硅编译器是一种自动在芯片上设计系统级功能模块和布线的程序。[2]在硅编译器诞生之前，每款芯片都是由工程师基于经验和直觉手工制作出来的。从本质上讲，米德的解决方案是让计算机自己设计芯片。这是我们迈向纳米工程的起点。



图14-1 1976年的卡弗·米德。正是在那段时期，他在加州理工学院创造了第一个硅编译器。卡弗是一位有远见的人，他的洞察力和引领的技术进步对数字和模拟计算产生了重大影响。桌上的电话暗示了照片的拍摄时间。图片来源：加州理工学院。

米德是一位颇有远见的人。我们曾一起在匹兹堡郊外参加一个研讨会。尽管大家挤在楼上一个小房间里的桌子边上讨论，但是正在进行着的是一场关于超级计算机的会议。当时Cray和Control Data等主流超级计算机公司正在设计专用硬件，比我们实验室的计算机快了数百倍，售价高达1亿美元。Cray超级计算机速度如此之快，不得不用氟利昂进行液体冷却。米德告诉我，超级计算机公司还不知道，通用微处理器将会占领它们的市场，这些公司很快就会销声匿迹。虽然比超级计算机中的专用芯片慢得多，但得益于因缩小基本器件尺寸所带来的成本压缩和性能改进，个人计算机中的微处理器进化得比超级计算机中的还要快。现在手机中的微处理器的计算能力是20世纪80年代Cray XMP超级计算机的10倍，现在的高性能超级计算机拥有数十万个微处理器核心，每秒进行的浮点运算达到了千万亿次级别，比已经消失的Cray超级计算机快100万倍。考虑到通货膨胀，两者的成本大致是相同的。

在1983年的那次研讨会上，米德向我们展示了一个硅视网膜，它使用了与超大规模集成电路芯片相同的制造技术，但使用了模拟而非数字电路。在模拟电路中，晶体管上的电压可以连续变化，而数字电路中的晶体管只能采用“开”或“关”两个二进制值中的一个。人的视网膜上有一亿个光感受器，与相机将信息从光子桶（photon buckets）传动到记忆体的方式不同，视网膜具有多层神经处理功能，可将视觉输入转换为高效的神经编码。视网膜的所有处理过程都是模拟的，直到其编码信号到达神经节细胞，神经节细胞将这些信号沿着100万个轴突，以“全或无”的放电脉冲形式传送到大脑。脉冲信号的“全或无”特征就像数字逻辑，但放电脉冲的时间是一个模拟变量，没有时钟同步，因此放电脉冲序列是一种混合编码。

在米德的视网膜芯片中，处理过程的分级部分是通过使用位于阈值拐点以下，从“关”到接近“关”状态的电压来完成的。与之相反，在数字模式下运行时，晶体管迅速跳到完全“导通”的状态，这需要消耗更多的功率。因此，模拟超大规模集成电路芯片仅消耗数字芯片所需能耗的一小部分，从纳瓦到微瓦，而不是从毫瓦到瓦，能量效率提升了数百万倍。米德是神经形态工程的创始人，他致力于构建基于大脑式算法的芯片，在1989年，他表明嵌入在昆虫和哺乳动物眼睛神经环路中的神经算法，可以有效地复制到芯片上。[3]

视网膜芯片是米德的明星研究生米莎·马霍沃德（Misha Mahowald）于1988年创造的一项令人印象深刻的发明（见图14-2）。[4]她在加州理工学院读本科时主修的是生物学。在研究生阶段，她从事的是电子工程领域的工作。这两个领域的结合所带来的洞见帮助她获得了四项专利。1992年，她的博士论文描述了在芯片上的实时双目匹配，这是第一个真正使用集群行为完成艰巨任务的芯片。为此她获得了加州理工学院的米尔顿和弗朗西斯·克拉泽奖（Milton and Francis Clauser Prize）。1996年，她的名字还被列入了国际技术女性（Women in Technology International，WITI）名人堂。



图14-2 1982年，加州理工学院的米莎·马霍沃德作为卡弗·米德的学生，创造了当时世界上第一个硅视网膜。她对神经形态工程的贡献是开创性的。图片来源：托比·德尔布吕克。

晶体管在阈值附近的物理特性，与生物膜中离子通道的生物物理特性之间存在密切的对应关系。马霍沃德与牛津大学的神经科学家凯万·马丁（Kevan Martin）和罗德尼·道格拉斯（Rodney Douglas）合作开发过硅神经元，[5]并随同他们一起搬到了苏黎世，帮助他们在苏黎世大学和瑞士联邦理工学院建立了神经信息学研究所（见图14－3）。然而后来，在经历了抑郁症的折磨之后，米莎于1996年结束了自己的生命，时年33岁。一位杰出的新星就此陨落。

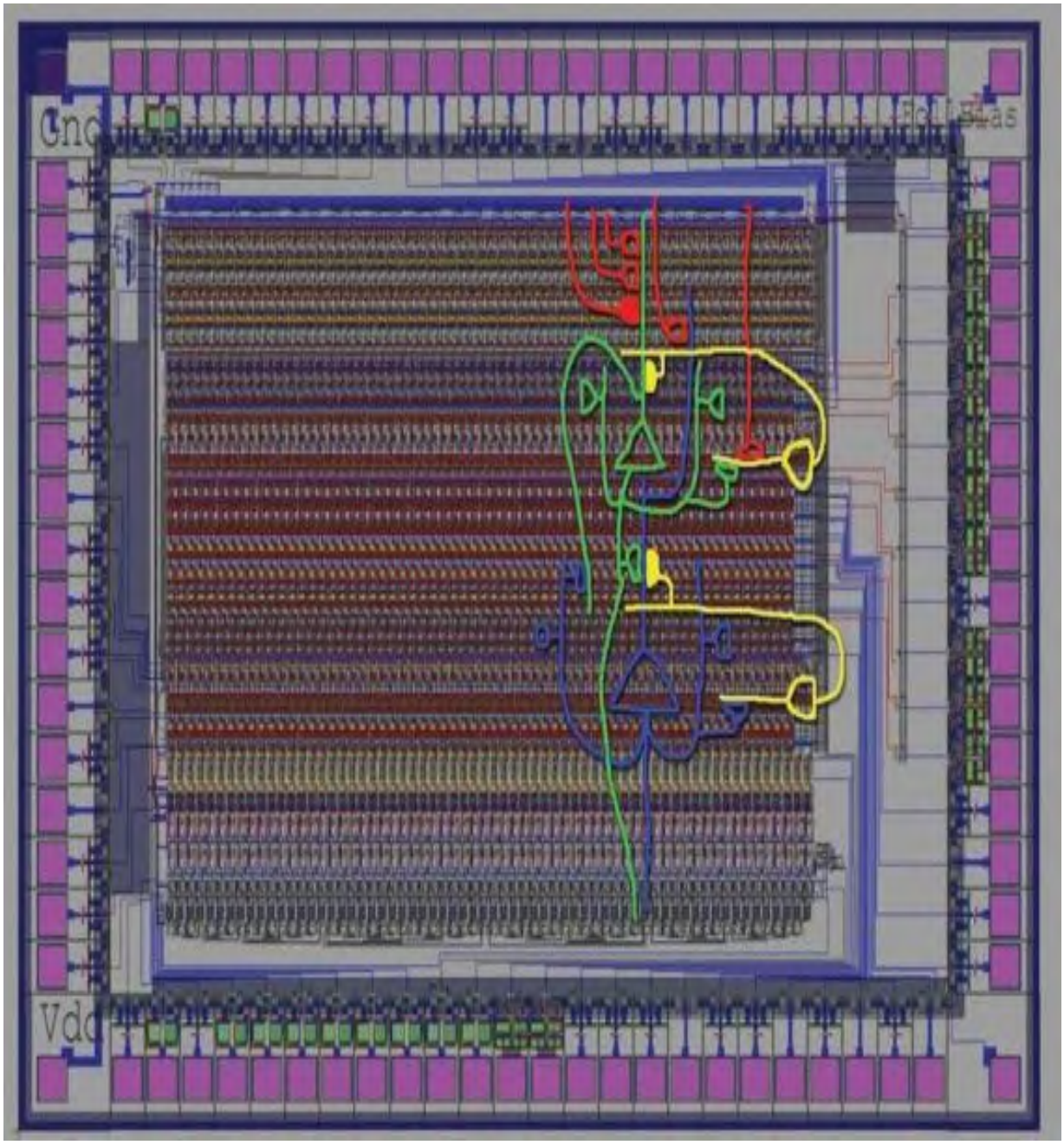


图14-3 硅神经元。该模拟大规模集成电路芯片具有类似于神经元中离子通道的电路，能够实时对神经回路进行仿真操作，正如米莎·马霍沃德在芯片上绘制的卡通图所示。图片来源：罗德尼·道格拉斯。

卡弗·米德于1999年从加州理工学院退休后搬到了西雅图，2010年我去拜访过他。在他家的后院，可以看到飞机飞过水面有序地降落到Sea-Tac机场。他的父亲是一位工程师，在大溪水电项目的一座发电

厂工作，该项目是位于加州中部内华达山脉圣华金河上游系统的一个大型水电项目。从早期的水电技术向微电子技术的飞跃，仅仅花了一代人的时间，真令人难以置信。卡弗的爱好是收集用于悬挂电线的古董玻璃和陶瓷绝缘子。如果你选对了地点，就会像寻找散落的印第安箭头一样轻松地找到它们。卡弗目光长远（他拥有一台激光陀螺仪，曾经被用来测试量子物理学的新方法），[6]而他之所以如此高效，是因为他致力于打造不仅能够正常运转，还可以被你拿在手中把玩的东西。

神经形态工程

1990年，作为在加州理工学院休学术年假的仙童杰出学者，我喜欢参加实验室会议，尤其是克里斯托弗·科赫（Christof Koch）（一位和我有着共同兴趣的计算神经科学家）和他的同事的会议，以及Carverland，即卡弗·米德的研究小组的会议。Carverland有一个令人惊叹的项目——硅耳蜗，它具有和我们耳中的耳蜗相类似的频率调谐电路。其他研究人员正在研究硅突触，包括模仿突触可塑性的硅机制，以便可以在硅芯片上实现长期的权重变化。来自Carverland研究小组的学生从那时开始走出实验室，进入了世界各地的工程学院。

1993年，克里斯托弗·科赫、罗德尼·道格拉斯和我成立了由NSF赞助的神经形态工程研讨会，每年7月在科罗拉多州特柳赖德（Telluride）市举行为期三周的会议。这个研讨会是国际性的，参会的学生和教师来自不同的背景和国家。与大多数讲座多于实践的研讨会不同，特柳赖德研讨会的房间里挤满了用微芯片来构建机器人的学生。不过有一个问题，将视网膜芯片连接到视觉皮层芯片，再将皮层芯片连接到电机输出芯片，这一步骤需要用到大量的连线。

用放电脉冲（spikes）连接模拟大规模集成电路芯片则要好得多，我们的大脑通过白质的长距离轴突传输信息也是利用了这种方

式，这些白质构成了我们一半的大脑皮层。但将视网膜芯片和皮层芯片通过上百万根线连接起来却完全行不通。幸运的是，快速的数字逻辑可以用来使每根导线多路复用，让许多视网膜细胞能够与同一导线上的多个皮层细胞进行通信。这一过程是通过发送芯片，把每一个原始脉冲的地址发送给接收芯片来实现的，接收芯片解码地址，并路由到与它相连接的相应单元，即所谓的“地址事件表示”（address event representation）。

托拜厄斯·德尔布吕克（见图14－4）是卡弗·米德的研究生之一，[7]他现在在苏黎世大学神经信息学研究所工作。2008年，他开发了一种非常成功的脉冲视网膜芯片，叫作“动态视觉传感器”（Dynamic Vision Sensor，以下简称DVS）。这个芯片可以简化一些视觉任务，例如跟踪移动物体，或用两台摄像机对物体进行深度定位（见图14－4，下图）。[8]传统的数码相机是基于图像帧的，录制视频的过程就是存储一系列间隔大约26毫秒的图像。这种做法会丢失帧与帧之间的信息：设想一个旋转的托盘，转速是每秒200转，盘上有一个光点随盘转动；该光点会在每一帧中旋转5次，但是数码相机的回放效果看起来像一个静态环（见方框14.1）。相比之下，德尔布吕克的脉冲摄影机可以以微秒级的精度跟踪移动点，并且只需很少量的脉冲，因此速度很快，而且效率很高。作为第一款基于脉冲和触发定时的新一代传感器，DVS相机具有很大的潜力，可以改善包括自动驾驶汽车在内的许多应用的性能。2013年特柳赖德研讨会的其中一个项目，就是用它来阻挡来自桌面的进球（见图14－5）。





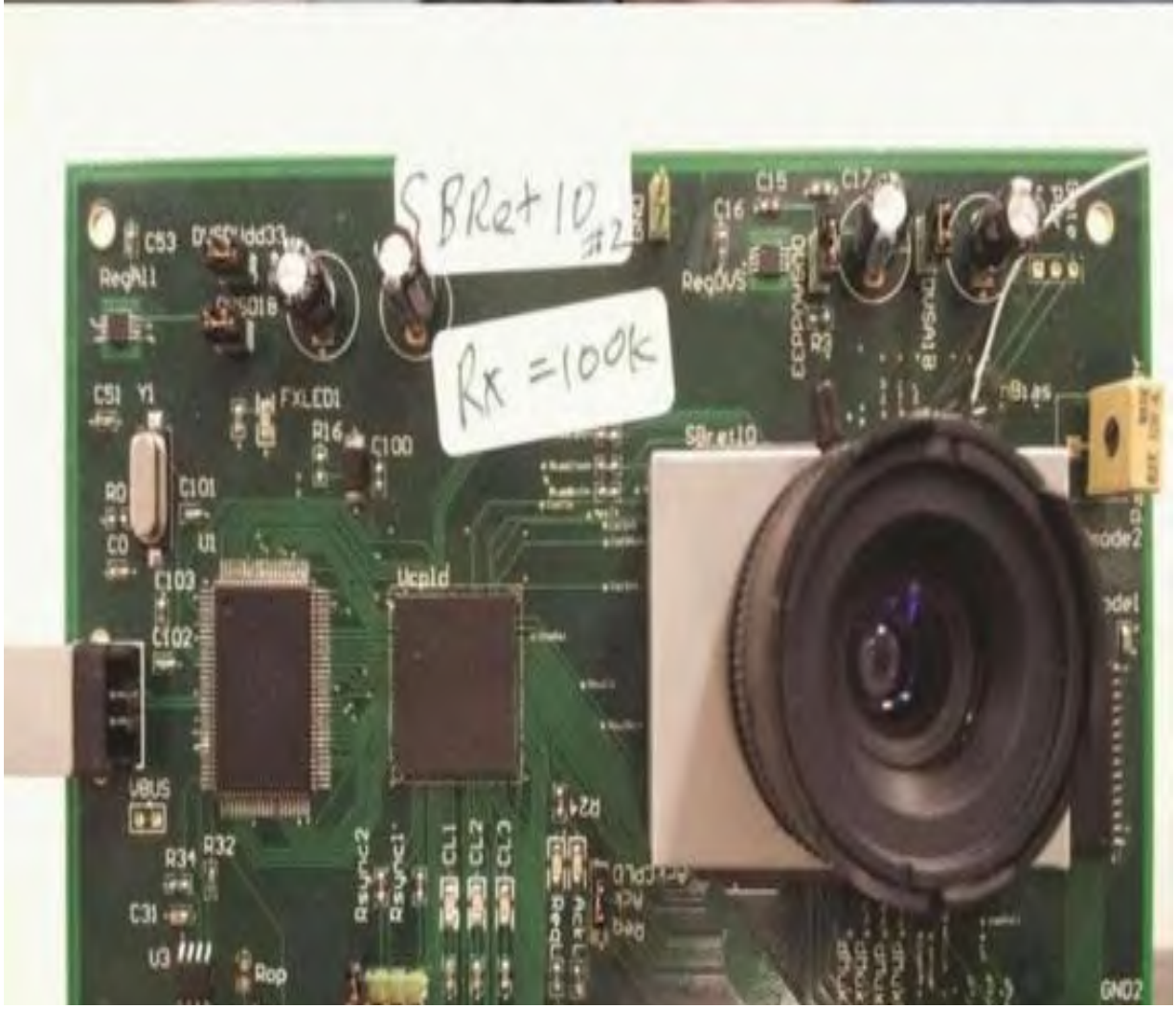
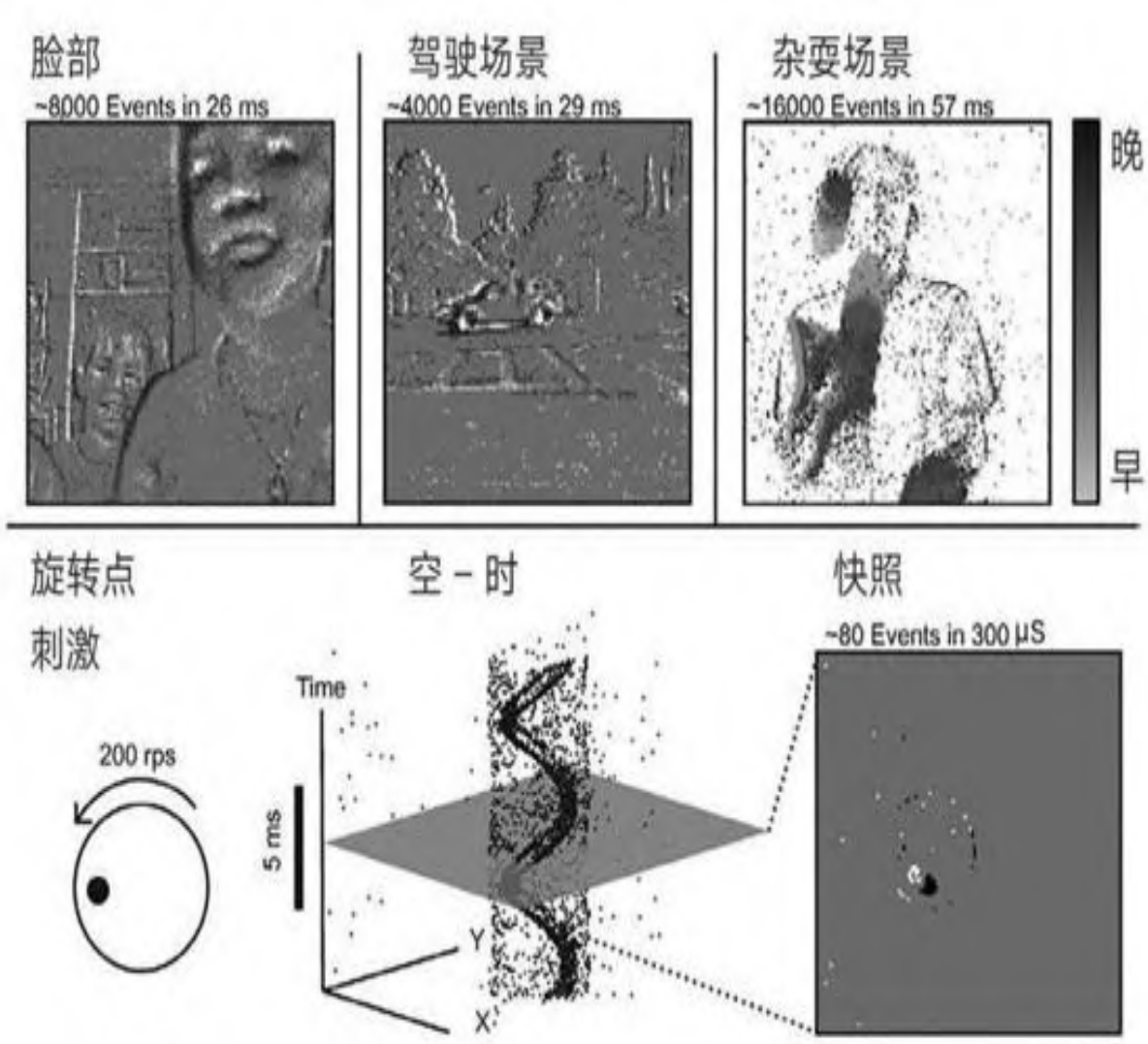

SBRet10
Rx =100k
C53
C51
Y1
C100
C101
C103
C102
U1
Ucald
VBUS
R32
R34
C31
U3
Rop
Async2
Async1
CL1
CL2
CL3
C15
C16
C17
SBret10
GND2



图14-4 动态视觉传感器（DVS）。（上图）托拜厄斯·德尔布吕克拿着他在苏黎世大学神经信息学研究所发明的DVS相机。这种相机采用了专用芯片，可以异步发射脉冲，而不是像数码相机那样按帧采集图像。（下图）相机的镜头将图像聚焦在模拟超大规模集成电路芯片上，芯片会对每个像素上光强度的增加或减少进行检测。脉冲沿着“开启”（on）导线发出正增量，并沿着“关闭”（off）导线发出负增量。输出脉冲由电路板处理，电路板会显示出例如方框14.1中的脉冲图案。你的视网膜是一台非常先进的DVS相机。来自视网膜的脉冲图案在大脑中发生了转换，但该模式仍然保留了脉冲——你的大脑中并不存在任何完整的图像，即使你是以这种方式感知世界的。上图来源：托拜厄斯·德尔布吕克。下图来源：三星。

14.1 动态视觉传感器的工作原理



在上图DVS摄像机的图像帧中，白点是来自“开”（on）通道的脉冲，黑点是来自“关”（off）通道的脉冲。灰色表示没有脉冲。在左上方的图片中，可以检测到两个脸部，因为它们在26毫秒的帧间隙内发生了轻微的移动。在右上角的（杂耍）图片中，斑点的到达时间由灰度表示，因此就能看到物体的移动轨迹。左下方图片中的旋转盘以每秒200转（rps）的速度旋转。在底部中间的图片中，轨迹是向上移动的螺旋。右下方螺旋短暂的300微秒切片中，只有80个脉冲，通过测量黑色和白色脉冲的位移，并除以时间间隔，很容易计算出速度。请注意，具有26毫秒取帧周期的普遍数码相机无法跟随以200赫兹旋转的

点，因为旋转周期为5毫秒，并且每个帧都显示了一个环。DVS相机的唯一输出是一串脉冲，就像视网膜一样。这是表现场景的有效方式，因为大部分像素在大多数时间都保持不变，而每个脉冲都携带着有用的信息。

图片来源：P. Lichtsteiner, C. Posch, and T. Delbruck, “A 128×128 120 dB 15 μs Latency Asynchronous Temporal Contrast Vision Sensor,”IEEE Journal of Solid-State Circuits 43, no. 2 (2008): 图11。资料来源：托拜厄斯·德尔布吕克。

脉冲神经元为计算领域打开了新的机会。例如，神经元群体中放电脉冲的时间可用于调整存储的信息类型。1997年，亨利·马克拉姆（Henry Markram）和德国的伯特·萨克曼（Bert Sakmann）报道说，通过对突触的输入信号和突触后神经元的输出信号的重复配对，他们能够对突触强度进行增减。[9]如果突触输入放电发生在突触输出放电之前的20毫秒窗口内，突触会保持长期增强，但如果突触输入放电的重复配对发生在突触输出放电之后的20毫秒窗口内，则突触会保持长期衰减（见图14－6）。据报道，“放电时序依赖可塑性”（Spike-timing-dependent plasticity，简称STDP）曾经被发现于许多物种大脑中的不同部位，这种可塑性可能对事件序列在脑中形成长期记忆起着重要的作用，但同样重要的是，它很好地解释了赫布的假定（在第7章中已经进行了讨论）。[10]

赫布可塑性的普遍观点是，当一个神经元的输入和输出同时出现放电脉冲时，突触的强度会增加，这是一种重合检测的形式。但是赫布实际上说的是，“当细胞A的轴突接近到足以激发细胞B，并反复或持续参与放电激发细胞B时，细胞的一些生长过程或代谢变化会发生在其中一个或两个细胞中，使得细胞A（作为激发细胞B的细胞之一）激发细胞B的效率增加”。[11]如果细胞A对激发细胞B有贡献，那么细胞A必须在细胞B中的“受激发脉冲”之前发射一个“激发脉冲”。正如赫布所描述的，这种情况提示了一种因果关系，而不仅仅是相关

性。尽管赫布没有对降低突触强度的条件进行描述，但当输入放电脉冲发生在输出放电脉冲之后时，输入脉冲就不太可能导致输出神经元发生脉冲。如果从长远来看，突触强度的增强和降低必须达到平衡，那么这时断开突触是讲得通的。





图14-5 2013年，特柳赖德神经形态工程研讨会中的神经形态守门员。（上图）Fopefolu Folowosele（左侧）在测试神经形态守门员（右侧）。其他学生和他们的项目可以在背景中看到。（下图）德尔布吕克的DVS相机能驱动油漆搅拌棒转动。这个守门员的动作比学生要快得多，并且成功地守住了每一个射门。我也尝试过，但是未能进球。图片来源：托拜厄斯·德尔布吕克。

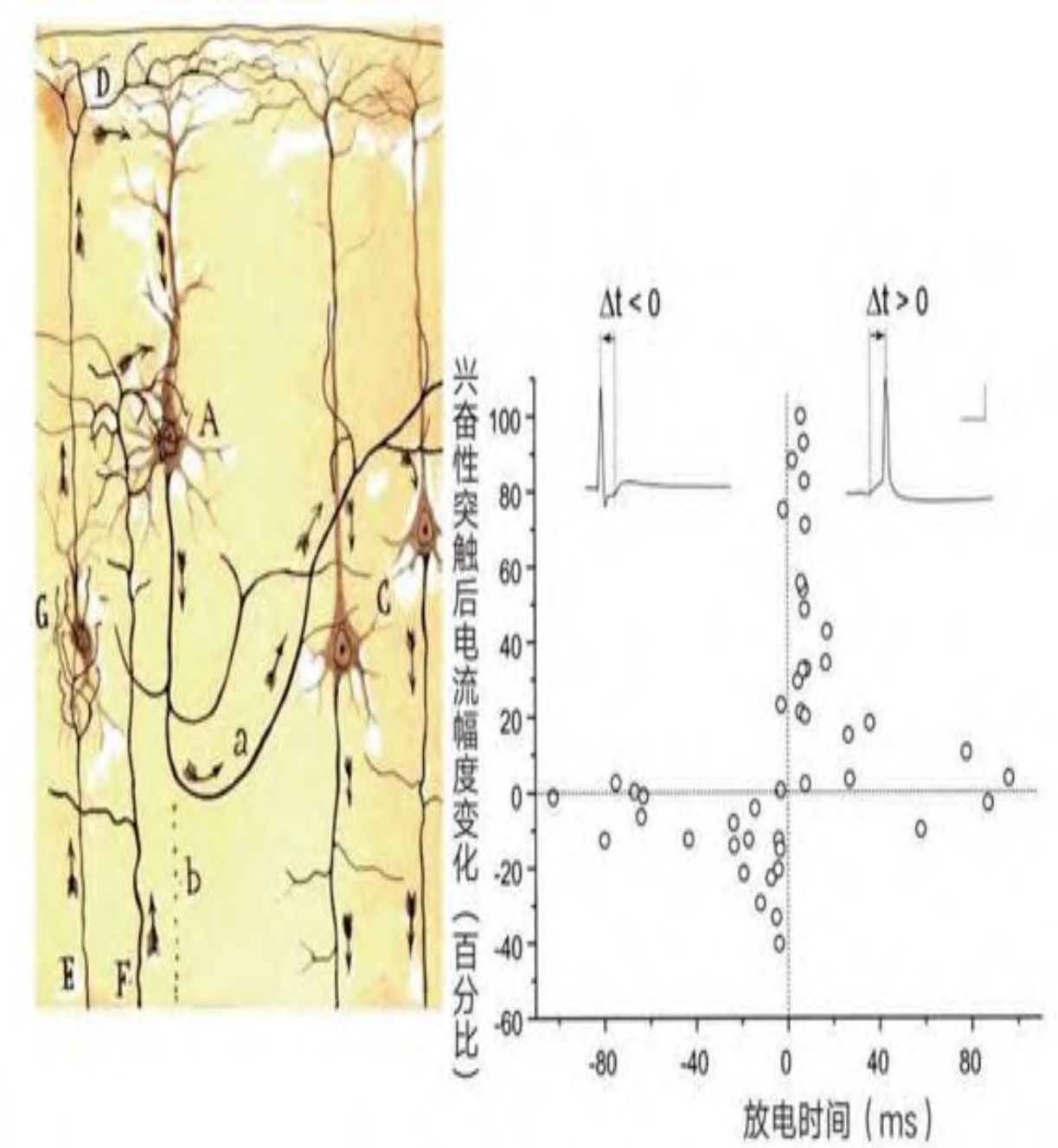


图14-6 放电时序依赖的可塑性（STDP）。（左图）伟大的西班牙神经解剖学家圣迭戈·拉蒙-卡哈尔（Santiago Ramóny Cajal）绘制的皮层中的锥体神经元。神经元A的输出轴突通过突触与神经元C的树突接触（如箭头所示）。（右图）与左侧图中类似的两个神经元被用电极穿刺并施加刺激，使两个神经元产生的尖峰之间存在时间延迟。当神经元的输入放电与输出放电重复配对时，如果突触前神经元放电在突触后神经元放电之前20毫秒的窗口内到达，则突触强度（垂直轴）增加。如果顺序相反，则强度降低。左图来源：Ramóny Cajal，S. Estudios Sobre la Degeneración y Regeneración del Sistema Nervioso(Moya，Madrid，1913-1914)，图281。右图来源：G. Q. Bi and M. M. Poo，“Synaptic Modifications

in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type,” Journal of Neuroscience 18 (1998): 10464-10472, 图7。资料来源：蒲慕明。

在特柳赖德神经形态研讨会上，模拟超大规模集成电路倡导者和数字超大规模集成电路设计师之间存在着一场持续不断的争论。模拟超大规模集成电路芯片有许多优点，例如所有电路并行工作时功耗非常低，但也存在缺点，例如由于晶体管的多样性，在设计图上相同的晶体管产生的电流可以相差±50%。相比之下，数字超大规模集成电路虽然比较精确，速度更快，设计更容易，但需要消耗更多的功率。德尔门德拉·莫德哈（Dharmendra Modha）在加利福尼亚州阿尔马登市（Almaden） IBM研究所的团队开发了一款数字芯片，包含4096个处理核心和54亿个晶体管，被称为“True North”。[12]尽管该芯片可以被配置来实现模拟100万个突触神经元，并连接2.68亿个突触，却仅仅需要消耗70毫瓦的功率。但是这些突触的强度是固定的，这种不变性限制了对许多重要特征的实现，如强度的弱化或增强。

具有脉冲神经元的网络还有另一个缺点，即梯度下降。脉冲时间是不连续的，而梯度下降要求神经元的输入输出值在时间上是连续的。这就限制了可训练脉冲网络的复杂性。梯度下降在训练“输出随输入连续变化”神经元的深度网络方面取得了巨大成功，神经元的输出函数是可微分的，这是反向传播学习算法的基本特征。尽管在不可微分脉冲网络中，脉冲的发生不具有连续性，但是我实验室的博士后研究员本·哈（Ben Huh）最近克服了这个缺点，他找到了一种方法，能够让脉冲神经元的循环网络模型使用梯度下降，在长时间序列上执行复杂的任务。[13]这一成果打开了深度脉冲网络的大门。

摩尔定律的终结

正如摩尔定律预测的那样，自20世纪50年代数字计算机发明以来，计算机的能力增加了超过1万亿倍。从来没有技术能以如此高的数量级呈指数增长，计算机已经融入了几乎所有制造设备，从玩具到汽车。计算机可以自动调节现代望远镜的自适应光学系统，以最大限度地提高其分辨率；它们可以分析现代显微镜捕获的光子，以超分辨率定位分子。当今的每一个科技领域都依赖于超大规模集成电路芯片。

卡弗·米德预测，这些芯片的升级是基于缩小元件间线宽的可能性，但是目前的宽度已经达到了物理极限：导线中的电子太少，可能会泄漏或被随机电压阻挡，甚至使数字电路都不再可靠。[14]摩尔定律要终结了吗？我们需要完全不同的体系结构来继续提高处理能力，这种结构能够不依赖于数字设计的完美精度。正如混合动力汽车将电动机的效率与汽油发动机的续航结合起来一样，混合数字和神经形态设计正在兴起，能够将神经形态芯片的计算低功耗与通信数字芯片的高带宽相结合。

摩尔定律的表述仅仅是基于芯片的处理能力。随着并行架构在随后50年内的不断发展，摩尔定律应该被考虑了能量和吞吐量的定律所取代。在俄勒冈州波特兰举办的英特尔2018 NICE大会上，来自美国和欧洲的研究人员展示了三款新的神经形态芯片：来自英特尔的Loihi研究芯片，以及由欧洲人类大脑项目支持的两款第二代芯片。随着大规模并行体系结构的发展，能够在这些体系结构上运行的新算法也陆续被创造出来。但是这些体系结构中的芯片需要交流信息，这也是第15章所要讨论的内容。

---

[1] R. Blandford, M. Roukes, L. Abbott, and T. Sejnowski, “Report on the Third Kavli Futures Symposium: Growing High Performance Computing in a Green Environment,” September 9–11, 2010, Tromsø, Norway, http://cnl.salk.edu/Media/Kavli-Futures.Final-Report.11.pdf.

[2] Carver A. Mead and George Lewicki, “Silicon compilers and foundries will usher in user-designed VLSI,” *Electronics*August 11, 55, no. 16 (1982): 107–111. ISSN 0883- 4989.

[3] Carver Mead, *Analog VLSI and Neural Systems*(Boston: Addison-Wesley, 1989).

[4] M. A. Mahowald and C. Mead, “The Silicon Retina,” *Scientific American*264, no.5 (1991):76–82; Tribute from Rodney Douglas: https://www.quora.com/Whatwas-the-cause-of-Michelle-Misha-Mahowald-death/; Misha Mahowald, “Silicon Vision” (video), http://www.dailymotion.com/video/x28ktma_silicon-vision-mishamahowald_tech/.

[5] M. Mahowald and R. Douglas, “A Silicon Neuron,” *Nature*354, no. 6354 (1991):515–518.

[6] 参阅Carver Mead, *Collective Electrodynamics: Quantum Foundations of Electromagnetism*(Cambridge, MA: MIT Press, 2002)。

[7] 托比的父亲马克斯·德尔布吕克（Max Delbrück）是 20世纪50年代的物理学家，也是分子生物学的创始人，他（与Alfred Hershey和Salvador Luna一起）于 1969年获得了诺贝尔生理学或医学奖（揭示了微电子学与分子生物学的另一层关系，将在第18章探讨）。

[8] C. Posch, T. Serrano-Gotarredona, B. Linares-Barranco, and T. Delbruck,“Retinomorphic Event-Based Vision Sensors: Bioinspired Cameras with Spiking Output,” *Proceedings of the IEEE*102, no. 10 (2014): 1470–1484; T. J. Sejnowski and T. Delbruck, “The Language of the Brain,” *Scientific American*307 (2012):54–59. https://www.youtube.com/watch?v=FQYroCcwkS0.

[9] H. Markram, J. Lübke., M. Frotscher, and B. Sakmann., “Regulation of Synaptic Efficacy by Coincidence of Postsynaptic APs and EPSPs,” *Science*275, no. 5297(1997): 213–215.

[10] 参阅T. J. Sejnowski, “The Book of Hebb,” *Neuron*24. no. 4 (1999): 773–776.

[11] Heb b, *The Organization of Behavior*(*New York: Wiley & Sons),*62。

[12] 参阅R. F. Service, “The Brain Chip,” *Science*345, no. 6197 (2014): 614–616.http://science.sciencemag.org/content/345/6197/614.full。

[13] D. Huh and T. J. Sejnowski, “Gradient Descent for Spiking Neural Networks,”2017. https://arxiv.org/pdf/1706.04698.pdf.

[14] K. A. Boahen, “Neuromorph’s Prospectus,” *IEEE Xplore: Computing in Scienceand Engineering*19, no. 2 (2017): 14–28.

15
信息科学

我从来没有想过，有一天我会成为无所不知的人，而事实上，我和现在任何能上网的人都做到了。信息在互联网上以光速传播，从互联网上获取信息比从书架上的书里获取信息更方便。我们生活在各种形式的信息爆炸的时代。科学仪器，从望远镜到显微镜，搜集的信息量远超现在训练机器学习所使用的数据集。美国国家安全局使用机器学习来过滤他们从各处搜集到的信息。经济活动走向了数字化，许多公司对有编程技能的人才的需求量与日俱增。在世界从工业经济走向信息经济的同时，教育和职业培训也必须适应这一变化。这一切已经对我们的世界产生了深远影响。

用字节丈量世界

1948年，在AT&T公司位于新泽西州默里山（Murray Hill）的贝尔实验室里，克劳德·香农（Claud Shannon，见图15－1）提出了一种非常简单却很微妙的信息理论，来理解有噪声的电话线里的信号传输问题。[1]香农的理论推动了数字通信的革命，并为移动电话、数字电视以及互联网的诞生奠定了基础。打电话的时候，你的声音被编码为“比特”，通过无线电波和数字电缆传送到接收端，接收端再将比特解码并转化为声音。信息论给通信信道（图15－2）的带宽确定了上限，逼近香农极限的不同编码也被设计了出来。

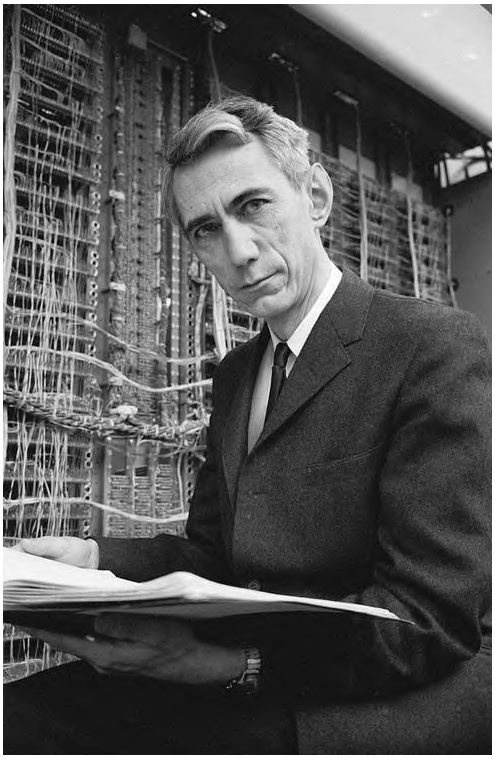


图15-1 克劳德·香农与电话交换网络（照片拍摄于1963年前后）。他发明信息论时仍在AT&T贝尔实验室任职。图片来源：Alfred Eisenstaedt/The LIFE Picture Collection/Getty Images。



图15-2 香农的通信系统模型。该消息被编码成二进制比特，并沿着一条可能是电话线或无线电波的信道向下传输，在终端被接收和解码。以“每秒比特数”为单位的信道容量取决于系统中的噪声量。

尽管世界上的信息形式丰富多样，有一种测量方式却能够精确测量一个数据集中信息的大小。信息的单位是二进制比特，每个比特只能有0或1两种值。一个字节包含8个比特。高质量照片的信息内容以兆字节或者说百万字节计数。手机的存储信息以吉字节（GB）或称千兆字节计数。网络上的数据以拍字节（PB）或称百万吉字节计数。

用数学思维解决通信难题

在IEEE（电气和电子工程师协会）的年度国际研讨会上，IEEE信息理论学会（ITS）都会颁发“克劳德·香农奖”，这是表彰这一领域

杰出研究成果的高级荣誉。在1985年英国布莱顿举办的该学会的会议上，“香农奖”被颁发给了南加州大学的所罗门·哥伦布（Solomon Golomb，图15-3），他在移位寄存器序列（shift register sequences）方面的工作为现代数字通信奠定了基础。[2]移位寄存器序列是一种生成0和1的长伪随机序列的算法。每次你用手机打电话时，都使用了移位寄存器序列。哥伦布展示了如何使用移位寄存器序列对信号进行高效编码，然后可以在接收器处传输和解码。如果将手机和其他通信系统产生移位寄存器序列的次数累加起来，得到的数字将是惊人的：超过千秭次，即10^{27}次(1,000,000,000,000,000,000,000,000,000)。[3]

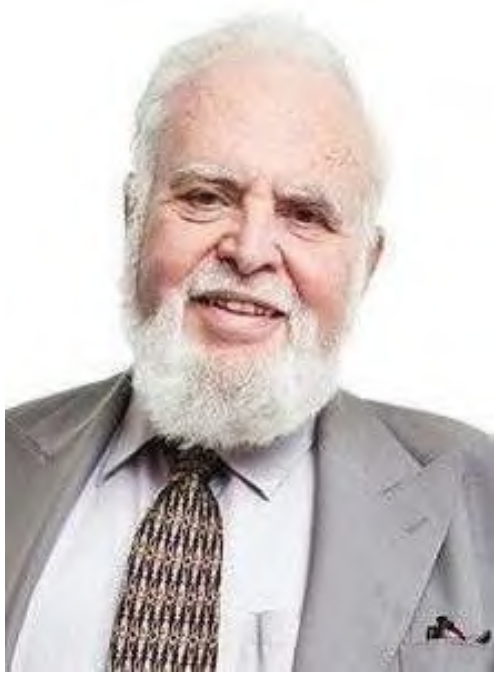


图15-3 所罗门·哥伦布在2013年获得了美国国家科学奖章。他对移位寄存器序列进行的数学分析使人类能够与深空探测器通信，当时他在位于帕萨迪纳（Pasadena）的加州理工学院喷气推进实验室（JPL）工作；移位寄存器序列后来被嵌入到手机通信系统中。每当你使用手机的时候，都在使用他的数学代码。图片来源：南加州大学。

我曾问过所罗门·哥伦布（也是我的岳父）他是如何得到如此优雅的方案来解决通信问题的。他说这是来自他在数论方面的训练，这是数学中最抽象的部分之一。当他在巴尔的摩的Glenn L. Martin公司做暑期实习生时，就曾接触过移位寄存器序列。1956年，在哈佛大学获得数论专业（一种高度抽象的数学领域）的博士学位后，他到了加州理工学院喷气推进实验室（Jet Propulsion Laboratory，以下简称JPL）工作。在那里，他担任通信组的负责人，并从事空间通信工作。深空探测器被发送到太阳系的远端，但返回的信号微弱且嘈杂。移位寄存器序列和纠错码大大改善了与空间探测器通信的效果，相同的数学原理更是为现代数字通信奠定了基础。

哥伦布在JPL的时候曾聘请过另一位杰出的信息理论家安德鲁·维特比（Andrew Viterbi），并将他介绍给了从麻省理工来JPL休学术年假的厄文·雅各布（Irwin Jacobs）。几十年后的1985年，维特比和雅各布联手创建了高通，从此彻底改变了手机技术。该公司使用移位寄存器序列将信息分布在很宽的频带上进行传播，这比使用单一频率的通信方式更有效。该想法的一个简单版本可以追溯到海蒂·拉玛（Hedy Lamarr）（见图15－4），她是一位电影演员兼发明家，于1941年与他人共同分享了跳频（frequency hopping）技术的专利，这是她在第二次世界大战期间为军方开发的安全通信系统。[4]哥伦布离开JPL加入南加州大学后，爱德华·波斯纳（就是创建了NIPS的爱德华·波斯纳）接管了他的团队，但哥伦布仍然在继续为他的前JPL团队提供支持和建议。



图15-4 海蒂·拉玛于1940年在米高梅拍摄的宣传照。她是第二次世界大战期间舞台和银幕上的明星，也是跳频技术的联合发明人之一。该技术与军方和许多手机中使用的扩频通信有关。

移位寄存器序列背后的数学是数论中比较深奥的部分。当哥伦布从哈佛大学获得博士学位时，他的博士生导师和当时的大多数数学家

都相信，纯数学永远不会有任何实际的应用。剑桥学者哈代（G. H. Hardy）在其富于影响力的著作《一位数学家的道歉》（*A Mathematician's Apology*）[5]中分享了这种观点。他宣称“好”的数学必须是纯粹的，而应用数学是“无趣的”。但数学只是数学本身，并没有纯粹和应用之分。一些数学家可能希望他们的数学是纯粹的，但他们无法阻止它解决现实世界中的实际问题。事实上，哥伦布的职业生涯主要是通过找到“纯数学”中正确的工具来解决实际问题而被定义的。

哥伦布还发明了数学游戏。他的书《多格骨牌》（*Polyominoes*）[6]引入了包含许多个正方形的各种形状，扩展了只有两格正方形的多米诺骨牌。马丁·加德纳（Martin Gardner）在《科学美国人》的“数学游戏”专栏中推广了多格骨牌。四格骨牌（Tetrominoes），是由四个方块组成的形状，俄罗斯方块的灵感就来源于此。俄罗斯方块是一款会令人上瘾的游戏，四格骨牌从上面像雨点一样落下来，需要将它们引导至底部的插槽。多格骨牌至今仍然是一种流行的棋盘游戏，并且在数学的一个子领域里引发了一系列有趣的组合问题。

哥伦布也是研究《圣经》的学者，会说几十种语言，包括日语和普通话。比阿特丽斯曾经带给他道格拉斯·霍夫斯塔特（Douglas R.Hofstadter）的著作《哥德尔、艾舍尔、巴赫：集异璧之大成》（*Gödel,Escher, Bach: An Eternal Golden Braid*）的初版。他打开卷首，标题说这是古希伯来语《创世记》的前20行。“首先，它是颠倒的。”他说，然后把书转过来，“其次，这是古撒玛利亚语，而不是古希伯来语。第三，这不是《创世记》的前20行，只是《创世记》前20行每行的前7个字。”他继续读下去，并翻译了这段经文。

克劳德·香农参加了1985年在布莱顿举行的ITS座谈会，哥伦布在会上做了一场香农讲座。这是香农继1972年自己的那次讲座后，参加的唯一一次香农讲座。

预测是如何产生的

在通信系统中，无论是空间上还是时间上，变化都具有很高的信息价值。强度均匀的图像和没有变化的信号一样，几乎不提供任何信息。向大脑发送信号的传感器主要用来检测变化，我们已经在第5章的视网膜和第14章中托拜厄斯·德尔布吕克的DVS相机中看到了一些例子。一旦在视网膜上稳定下来，图像会在几秒钟后消失。[7]尽管我们没有意识到这一点，我们的眼睛进行的微小跳动，即微跳动（microsaccades），每秒钟都会发生几次，每次跳动都会刷新我们大脑内部针对外部世界所搭建的模型。当外部世界的某些事物发生移动时，视网膜会及时向上报告，它们的报告还更新了大脑的外界模型，如图15－5所示。大脑的模型是一个层级结构的模型，传入的感官信息和模型期望值之间的比较发生在多个层次上。[8]明亮的闪光或巨大的噪声，通过自下而上地突显差异，立刻引起了你的注意。但是你通过对记忆自上而下的比较，注意到桌面上的某些东西在更高的层级上发生了变化。所有这些在大脑中实时发生的事情，都会让人想起卡弗·米德的口头禅，“时间是它自己的代表”。[9]

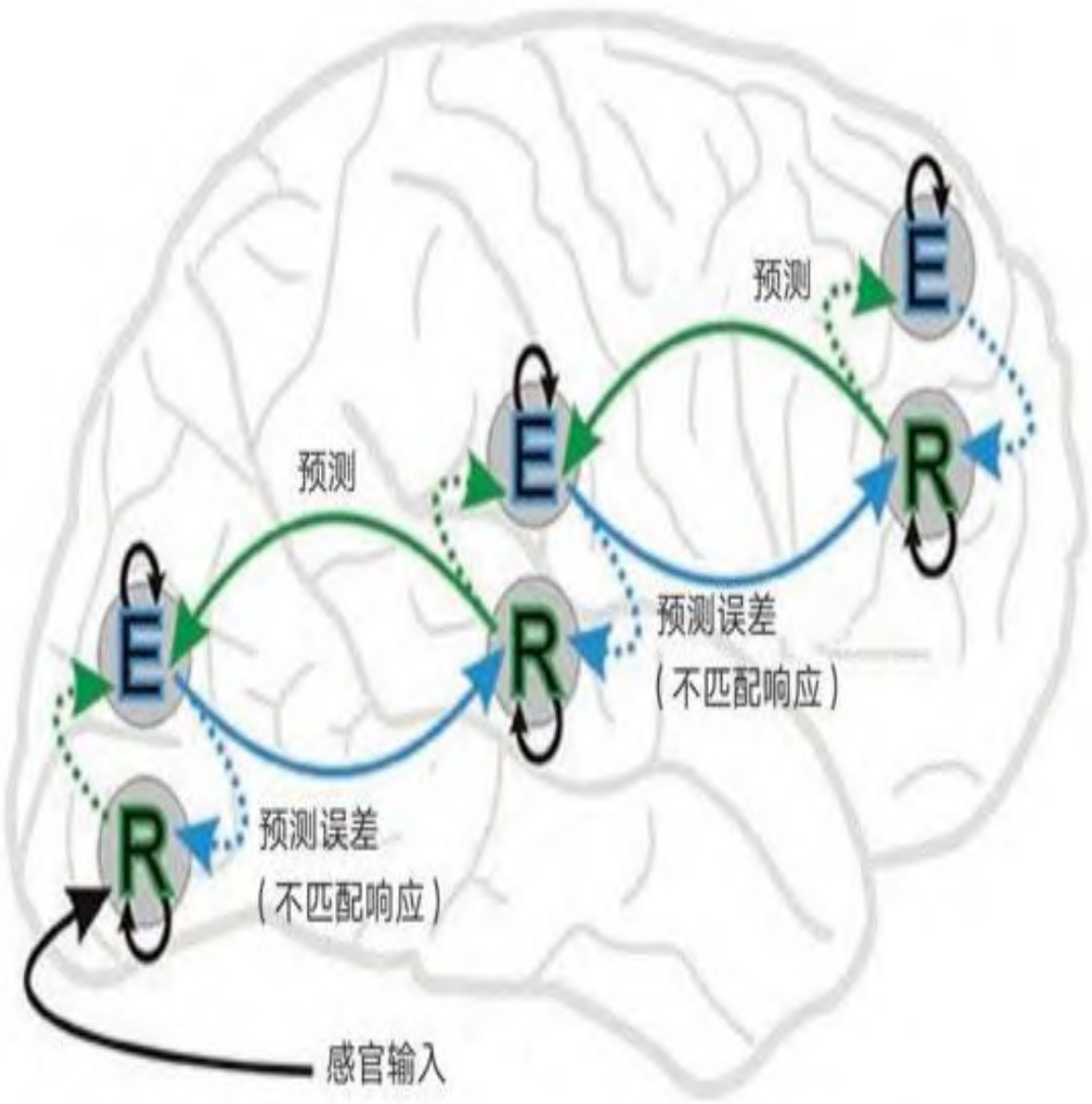


图15-5 具有层级结构的预测编码框架。感知取决于对早期感官事件提取的规律性的期望。在这个框架中，由较高层级皮层产生的当前感知信号的预测来自E和R群体之间的相互作用，并且被反馈到下面的层级（E是误差单元，R是表征单元）。只有预测误差会向前传播。这是对亥姆霍兹无意识推理的实现。图片来源：Gábor Stefanics, Jan Kremláček, and István Czigler, “Visual Mismatch Negativity: A Predictive Coding View,”Frontiers in Human Neuroscience 8 (2014): 666, 图1. doi:10.3389/fnhum.2014.00666。

预测编码可以追溯到赫尔曼·冯·亥姆霍兹，他将视觉解释为无意识推理，或自上而下生成视觉信息以消除噪声，填充不完整的信

息，并解释视觉场景。[10]例如，一个熟人的照片在我们视网膜里是有单眼深度线索的，因为我们熟悉那个人的实际大小，并且具备视网膜中大小如何随着距离变化的经验。在更高的认知水平上，詹姆斯·麦克莱兰和大卫·鲁姆哈特发现，当字母位于单词中时，被试者能够比在没有语境的非单词环境中更快地将它们识别出来。[11]他们的并行处理模型展现出了类似的行为，这使得两位研究人员相信，他们正在沿着理解信息如何在我们的大脑中呈现的正确轨道上前行。

深度理解大脑

2013年4月2日由白宫发起的美国“BRAIN计划”（见图15－6）的目标，是创造新的神经技术，以加速我们对“终极信息机器”，即大脑的功能和障碍的进一步理解。正如NIPS会议将许多学科的研究人员聚集在一起创建有学习能力的机器一样，“BRAIN计划”正在将工程师、数学家和物理学家引入神经科学领域，以改进探测大脑的工具。随着我们对大脑，尤其是关于学习和记忆机制的了解更加深入，我们将更好地理解大脑功能的原理。



图15-6 在2013年4月2日白宫宣布“BRAIN计划”之前，参加会议的相关机构和研究所的代表们。（从右至左）科维理基金会首席科学官全美永（Miyoung Chun），她

起草了“BRAIN计划”白皮书；威廉·纽索姆（William Newsome），NIH关于“BRAIN计划”咨询委员会的联合主席；弗朗西斯·柯林斯（Francis Collins），NIH院长；杰拉德·鲁宾（Gerald Rubin），霍华德·休斯医学研究所珍妮莉娅研究园区（Janelia Research Campus of the Howard Hughes Medical Institute）主任；科拉·马雷特（Cora Marrett），美国国家科学基金会主任；巴拉克·奥巴马（Barack Obama），时任美国总统；艾米·古特曼（Amy Gutmann），总统生命伦理委员会主席；罗伯特·康恩（Robert Conn），科维理基金会主席；阿尔提·普拉巴卡尔（Arati Prabhakar），DARPA主任；艾伦·琼斯（Alan Jones），艾伦脑科学研究所所长；我，索尔克研究所。图片来源：托马斯·卡里尔（Thomas Kalil）。

虽然在分子和细胞水平上对大脑有了深入的了解，但我们还没有在更高层次上对大脑的结构有一个同等的认知。我们已经知道，不同类型的信息被分散到皮层的不同部分进行储存，却不清楚如何快速检索所有分散的信息以解决复杂的问题，例如我们能够通过存储在皮层不同部位的人脸的图像，识别出一个人的名字。这个问题与大脑中意识的起源密切相关。

我的实验室最近发现了人类大脑在睡眠期间的整体活动模式，可以让我们了解大脑皮层中广泛分布的信息是如何联系在一起的。在恢复性慢波睡眠和快速眼动（REM）睡眠之间的睡眠阶段，高度同步的被称为“睡眠锭”（sleep spindles）的时空振荡，主导着皮层的活动。这些10到14赫兹的振荡会持续几秒钟，并在夜间发生数千次。有实验证据表明，在我们睡觉时睡眠锭参与了记忆的巩固。通过来自人脑皮层的记录，莱尔·穆勒（Lyle Muller）、西德尼·卡什（Sydney Cash）、乔瓦尼·皮安托尼（Giovanni Piantoni）、多米尼克·科勒（Dominik Koller）、埃里克·哈格伦（Eric Halgren）和我共同发现，睡眠锭在皮层中是一种无处不在的，扫过皮层所有部分的电活动的循环行波（图15-7）。[12]我们称它们为“莱娅公主波”（Princess Leia Waves），因为它们看起来跟这位公主的发型很像（图15-8）。我们推测，睡眠锭可能是皮层将白天获得的新信息，与先前分布在皮层中的记忆相结合的一种方式，加强了它们之间的长距离连接。这是“BRAIN计划”推动的系统神经科学层次的众多研究项目之一。

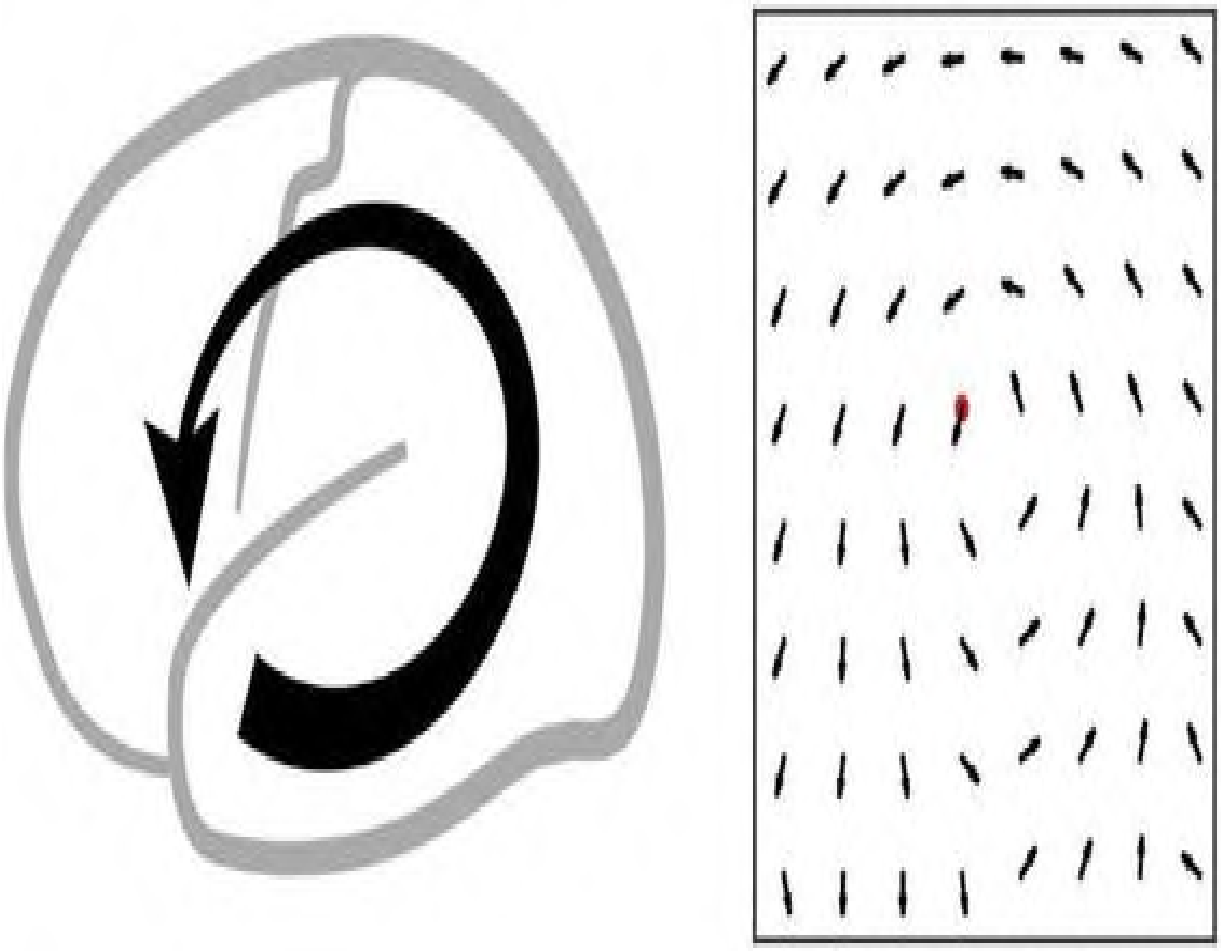


图15-7 在人类大脑皮层中的循环电子行波。这是在睡眠锭期间，从皮层表面上用8×8网格电极记录下来的图像。睡眠锭涉及记忆巩固。（左）睡眠锭是沿着箭头所示方向穿过皮层侧视图的循环行波，每80毫秒完成一个循环。这一过程在夜间会重复数千次。（右）小箭头表示循环期间，在皮层表面64个记录点上行波移动出现最大增加量的方向。图片来源：L. Muller, G. Piantoni, D. Koller, S. S. Cash, E.Halgren and T. J.Sejnowski, “Rotating Waves during Human Sleep Spindles Organize Global Patterns of Activity during the Night” Supplementary 7, subject 3, TPF。左图：图2B，右图：图1。



图15-8 凯莉·费雪（Carrie Fisher）饰演的莱娅公主，巩固了对1977年史诗级科幻电影《星球大战》的记忆。她的发髻很像睡眠锭期间穿过皮层的循环流场（对比

图15-7）。

大脑的操作系统

数字计算机和神经网络的体系结构不同。在数字计算机中，内存和CPU在空间上是分开的，并且内存中的数据必须依照顺序移动到CPU中。在神经网络中，处理过程在内存中并行进行，这就消除了内存和处理器之间的数字瓶颈，并且允许大规模并行处理，因为网络中的所有单元都在同时工作。神经网络中的软件和硬件之间也没有区别。学习是通过修改硬件来进行的。

从20世纪80年代开始，当计算机集群被装配在一个机架中时，数字计算机已经变得可以大规模并行了。最早的并行计算机之一是Connection Machine，由丹尼·希利斯（Danny Hillis）于 1985年设计出来，并由Thinking Machines公司出售。希利斯是一位工程师和发明家，他在麻省理工学院接受培训时，人们已经清楚地认识到，要让人工智能解决非常复杂的现实世界问题，就需要更多的计算能力。20世纪90年代，根据摩尔定律，计算机芯片上的晶体管数量持续增加。我们有可能将许多处理单元放置在同一芯片上，许多芯片放置在同一电路板上，许多电路板放置在同一机箱中，以及把许多机箱放在同一个房间里，结果是，如今地球上最快的计算机拥有数百万个内核，并且每秒可以实现数千万亿次操作。百亿亿次级的计算正在以每秒10亿乘10亿次操作的形式出现。

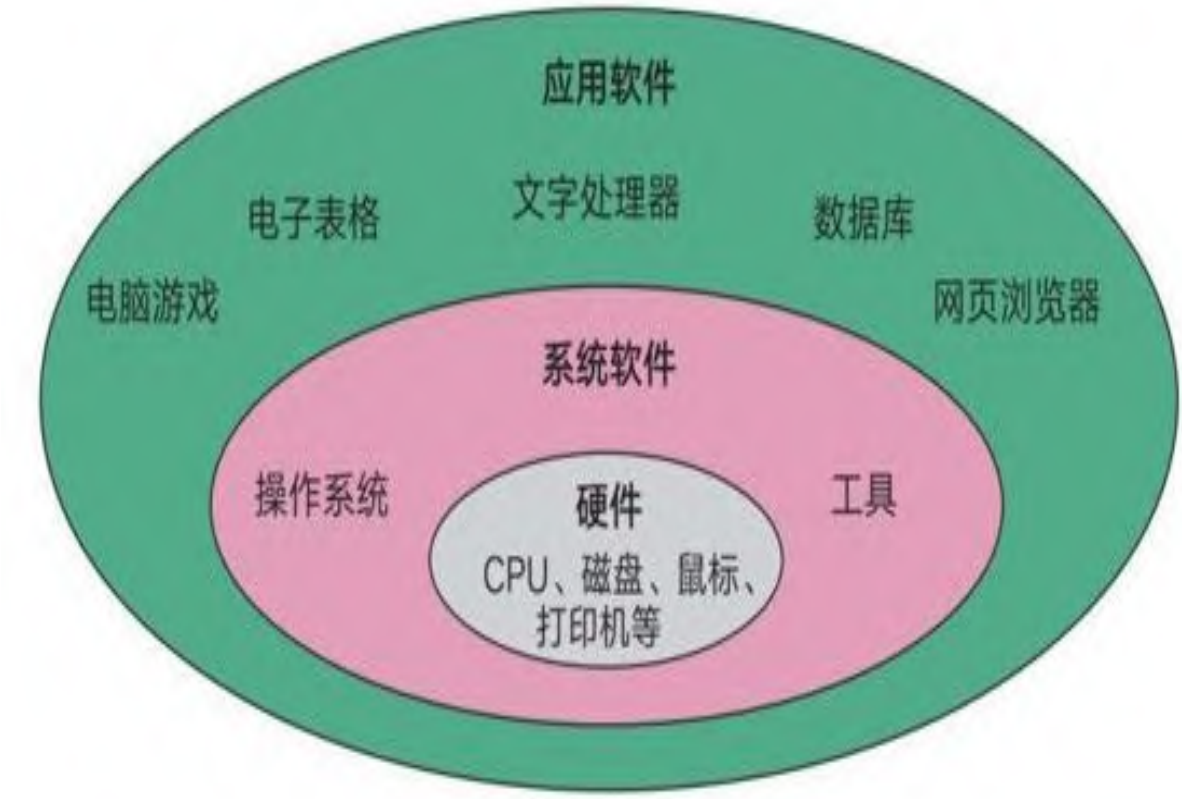
模拟神经网络可以最大限度地利用这种大规模并行硬件。多个核芯可以通过编程为同一个网络模型并行工作，这就大大加快了处理速度，但也会导致处理器之间的通信延迟。为了减少这些延迟，一些公司正在构建专用数字协处理器，这将极大地加速网络模拟，使语音和视觉等认知任务成为单一强大的指令。使用深度学习网络芯片的智能手机将变得更加智能化。

数字计算机具有将我们与硬件分开的操作系统（方框15.1）。当我们在笔记本电脑上运行文字处理程序，或在手机上使用应用程序时，操作系统会处理程序的细节，例如将按键信息放到内存的哪个位置，以及如何在屏幕上显示程序输出。我们的意识在大脑的操作系统上运行着应用程序，大脑操作系统将信息的内容、位置，以及存储方法与意识隔离开来。我们并不知道大脑是如何储存我们一生积累的大量经验的，也不知道这些经验如何塑造我们的行为。虽然有可能明确地解释一些经验，但我们所理解的这部分只是冰山一角。我们的大脑如何管理这些信息仍然是个谜。如果我们能够弄清楚大脑的操作系统是如何工作的，就可以基于相同的一般原则来组织大数据。相应地，意识也可以被解释为运行在大脑操作系统上的应用程序。

生物学与计算科学

信息爆炸已经将生物学转化为量化科学。对于传统生物学家，除了统计学入门课程之外，通常只需要一点额外的数学培训就能够分析他们的数据，这些数据很少，也很难获得。在2002年由长岛冷泉港实验室举办的分子遗传学研讨会上，我感觉自己就像一条离了水的鱼，因为我是唯一一个做计算主题演讲的人。在我前面演讲的是分子遗传学家勒罗伊·胡德（Leroy Hood），他在加州理工学院工作了多年。1987年我在加州理工学院休学术年假时，惊讶地发现胡德的实验室多得几乎占据了整栋大楼。后来，他搬到了西雅图，并在2000年联合创立了系统生物学研究所。这一新的领域旨在试图了解细胞内所有分子相互作用的复杂性。

15.1
电子计算机的操作系统



操作系统对在计算机的硬件上运行的程序进行控制。如果你使用的是个人电脑（PC），那么操作系统通常是Windows；如果你使用的是iPhone，用的则是iOS系统；大多数服务器运行的都是某种版本的UNIX。操作系统在程序需要时分配内存；它也可以在幕后跟踪程序，使用称为“守护程序”（daemons）的进程在后台运行，并跟踪打印机和显示器等设备。操作系统可以在任何硬件上工作，使得你的应用程序可以移植到不同的计算机上。

胡德在演讲中提到，有一天他心血来潮，想知道为什么他实验室工资系统里的计算机科学家比生物学家要多。他推断说，原因在于生物学已成为一门信息科学，计算机科学家对如何分析信息的了解要远远超过生物学家，而生物学家已被现代技术，例如基因测序，所产生的大量数据所淹没。胡德的演讲简直是为我做了最好的铺垫，我当时的演讲题目是关于信息如何在大脑中神经元之间的突触上储存的。

今天，系统生物学吸引了许多计算机科学家和物理学家共同分析和理解DNA测序产生的信息，以及由RNA和蛋白质控制的细胞信号。一个人类细胞的DNA中有30亿个碱基对，含有细胞生存、复制和分化所需的所有信息。一些碱基对是制造蛋白质的模板，基因组的其他部分包含一个抽象代码，对发育过程中使用的基因进行调节，以指导身体和大脑的构建。大脑的构建有可能是宇宙中难度最高的构建项目，它是由DNA中嵌入的算法引导的，这些算法协调了大脑数百个不同部位的数千种不同类型神经元之间的连接。

人工智能能拥有媲美人类大脑的操作系统

由基础科学研究发展出来的科技，通常需要50年才能实现商业化。20世纪初，由相对论和量子理论带来的重大发现，到20世纪的后半期才推动了CD播放器、GPS以及电子计算机的诞生。20世纪50年代发现的DNA以及基因序列影响了当今医药和综合农业领域的应用，这些应用到今天仍在影响经济的发展。“BRAIN计划”以及世界上其他大脑研究项目带来的发现，会启发50年后的应用，这些应用现在看上去还如科幻小说一般遥不可及。[13]我们可以期待，到2050年，人工智能会拥有能和我们大脑相媲美的操作系统。哪些公司、哪些国家会掌握这种科技，取决于它们现在所做的投资和所下的赌注。

---

[1] Jimmy Soni and Rob Goodman, *A Mind at Play: How Claude Shannon Invented the Information Age*(New York: Simon & Schuster: New York, 2017).

[2] Solomon Wolf Golomb, *Shift Register Sequences: Secure and Limited-Access Code Generators, Efficiency Code Generators, Prescribed Property Generators,Mathematical Models*, 3rd rev. ed. (Singapore: World Scientific, 2017).

[3] 参阅 Stephen Wolfram, “Solomon Golomb (1932–2016),” *Stephen WolframBlog*, posted May 25, 2016. http://blog.stephenwolfram.com/2016/05/solomongolomb-19322016/。

[4] Richard Rhodes, *Hedy's Folly: The Life and Breakthrough Inventions of Hedy Lamarr, the Most Beautiful Woman in the World*(New York: Doubleday, 2012).

[5] G. H. Hardy, *A Mathematician's Apology*(Cambridge: Cambridge University Press,1940).

[6] Solomon W. Golomb, *Polyominoes*(New York: Scribner, 1965).

[7] L. A. Riggs, F. Ratliff, J. C. Cornsweet, and T. N. Cornsweet. "The Disappearance of Steadily Fixated Visual Test Objects," *Journal of the Optical Society of America*43, no. 6 (1953): 495–501.

[8] Rajesh P. N. Rao and Dana H. Ballard, "Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects," *NatureNeuroscience*2, no. 1 (1999): 79–87.

[9] 这一观点是米德建立神经形态系统的动机，希望该系统像大脑一样，也能实时运行。C. Mead, "Neuromorphic Electronic Systems," *Proceedings of the IEEE*78, no. 10 (1990):1629–1636.

[10] Hermann von Helmholtz, *Helmholtz's Treatise on Physiological Optics*, vol. 3: *The Perception of Vision*, trans. James P. C. Southall (Rochester, NY: Optical Society of America, 1925), 25. Originally published as *Handbuch der physiologische Optik.*3. *Die Lehre von den Gesichtswahrnehmungen*(Leipzig: Leopold Voss, 1867).

[11] J. L. McClelland and D. E. Rumelhart, "An Interactive Activation Model of Context Effects in Letter Perception: Part 1. An Account of Basic Findings."*Psychological Review*88, no. 5 (1981): 401–436; "Part 2. The Contextual Enhancement Effect and Some Tests and Extensions of the Model," *PsychologicalReview*89, no. 1 (1982): 60–94,

[12] L. Muller, G. Piantoni, D. Koller, S. S. Cash, E. Halgren, and T. J. Sejnowski,"Rotating Waves during Human Sleep Spindles Organize Global Patterns of Activity during the Night," *eLife*5 (2016): e17267.

[13] 在所谓的"克拉克的第三定律"中，阿瑟·克拉克（Arthur C. Clarke）说过这样一句名言："任何足够先进的技术几乎都与魔法无异。"（Any sufficientlyadvanced technology is indistinguishable from magic.）

16
生命与意识

当弗朗西斯·克里克的母亲问年轻的克里克他想探究什么样的科学问题时，他告诉她，自己只对两个问题感兴趣：生命的奥秘和意识的奥秘。[1]克里克显然对重要的事物有着敏锐的感觉，但他当时可能没有意识到这些问题的困难程度。他的母亲并不知道，几十年后的1953年，她的儿子和詹姆斯·沃森会发现DNA的结构，这种松散的线状体最终将揭开生命的一个重要谜团。但克里克（图16－1）并不满足于这一成就。

克里克在1977年加入索尔克研究所后，重新拾起长久以来对意识的兴趣。他决定关注视觉意识（visual awareness）问题，因为人们已经对大脑的视觉功能区有了深入的了解，而了解视觉感知（visual perception）的神经基础也将为探索其他方面意识的神经基础铺平道路。[2]

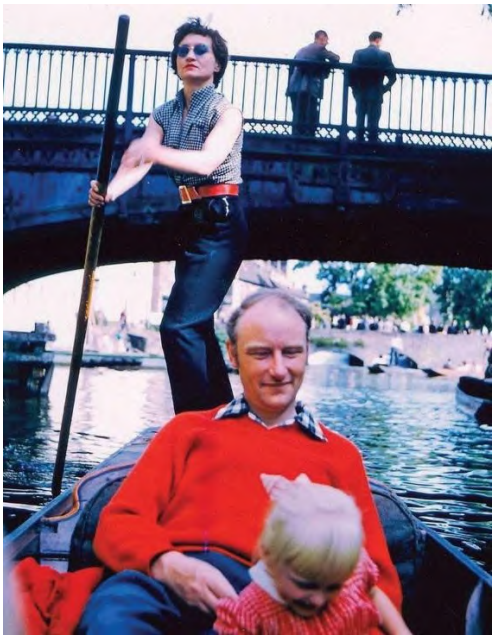


图16-1 弗朗西斯·克里克和他的妻子奥迪尔、女儿杰奎琳在英国剑桥的康河上划船（拍摄于1957年前后）。图片来源：索尔克生物研究所。

20世纪80年代，生物学家对意识研究投入的热情日渐消退，但这丝毫没有影响克里克。视觉感知充满了难以理解的幻想和谜团，为了解释它们在解剖学和生理学上的机理，克里克推出了新颖的“探照灯假说”（searchlight hypothesis）。[3]神经节细胞从视神经投射到丘脑，后者又将脉冲传递到视觉皮层。但为什么神经节细胞不能直接投射到大脑皮层呢？克里克指出，从皮层到丘脑区域的反馈投射就像探照灯一样，可能会突出显示部分图像，以供进一步处理。

视觉意识

在理解意识的工作中，克里克最亲密的伙伴是加州理工学院的神经科学家克里斯托弗·科赫，他们一起发表了一系列探索“意识的神经关联”（neural correlates of consciousness，简称NCC，负责产生意识觉知状态的大脑结构和神经活动）的论文。[4]对于视觉意识，这意味着要发现大脑不同部位神经元的放电特性与视觉感知之间

的相关性。克里克和科赫提出了假设，认为我们并不知道在初级视觉皮层中发生了什么[5]——这是大脑皮层第一个从视网膜接收输入的区域；我们只知道在视觉皮层层级结构的最高层发生了什么（见图5－11）。这种假设的可能性来自对双眼竞争的研究，即对两只眼睛呈现两种不同的图案，例如对一只眼睛显示垂直条纹而对另一只眼睛显示水平条纹：然而，我们看到的并非是两幅图像的混合图像——事实上，视觉感知每隔几秒就会在不同图像之间突然发生翻转。初级视觉皮层中对每只眼睛做出反应的是不同的神经元，不管哪个图像是被有意识地感知到的。然而，在视觉的较高层级上，许多神经元只对感知到的图像产生响应。因此，仅因一个神经元对感知放电，就说它是感知的神经关联，未免太过草率。显然，在分布于视觉层级中众多互相协调工作的活跃神经元中，我们只知道其中的一些子集表征了什么。

视觉感知的过程

2004年，加州大学洛杉矶分校医学中心的科研人员检测了一组癫痫患者的脑部，试图探索癫痫发作的病因。在检测过程中，病人被要求观察一系列名人照片。植入大脑记忆中心的电极可以记录患者对照片的响应脉冲。其中一名患者大脑的一个神经元对几幅哈莉·贝瑞的照片和她的名字（图16－2）做出了强烈的反应，但对比尔·克林顿、朱莉娅·罗伯茨的照片或其他名人的名字却无动于衷。[6]同时，该实验也发现了对其他名人、特定对象，以及悉尼歌剧院等建筑物能做出反应的神经元。

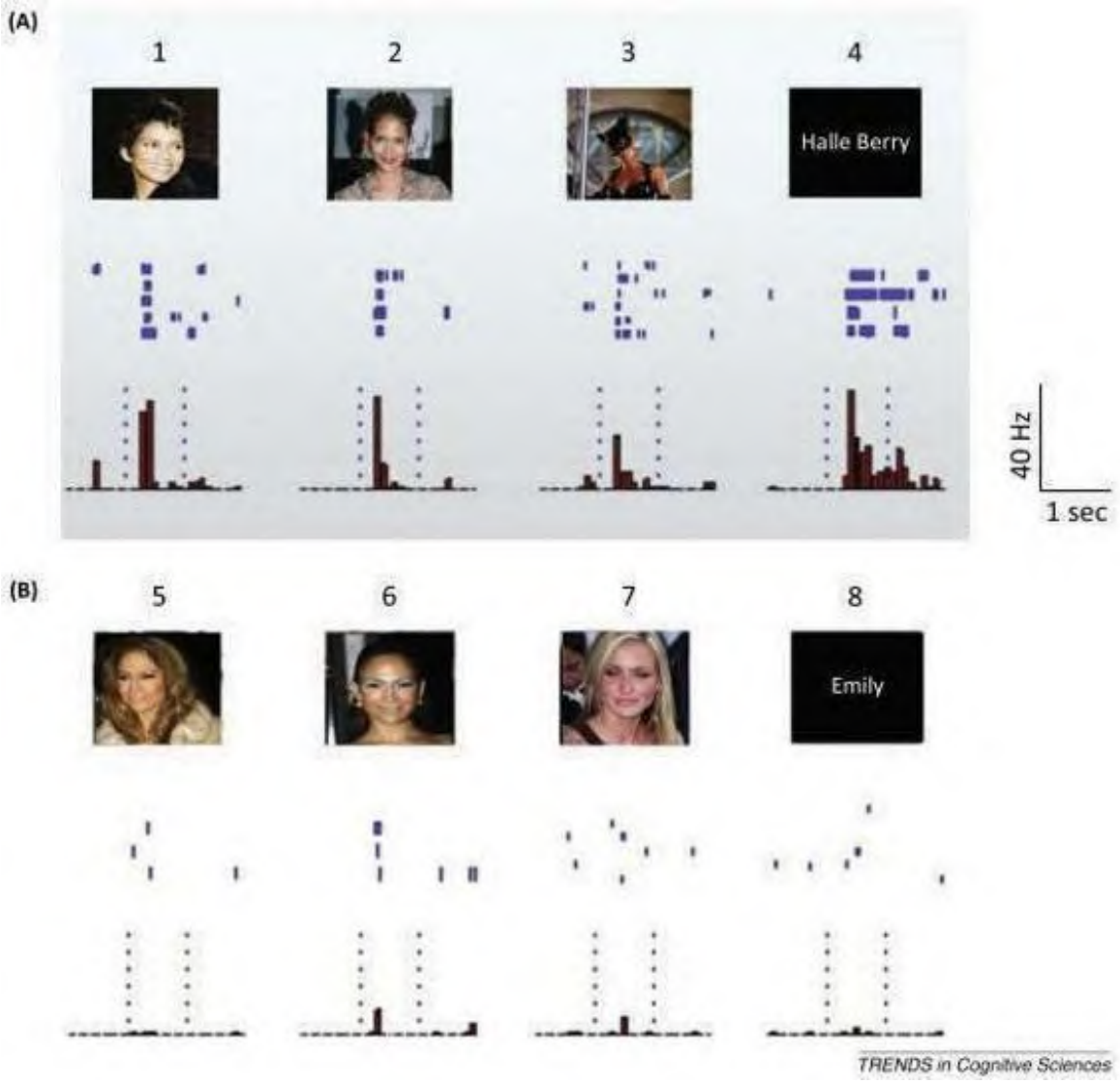


图16-2 哈莉·贝瑞细胞。从患者的海马体中单个神经元记录下的对照片的反应。每张照片下显示了来自6次单独试验的放电脉冲（蓝色标记点），以及平均值（直方图）。（A）哈莉·贝瑞的照片和她的名字引发了一波脉冲；（B）而其他女演员的照片和她们的名字却没有。哈莉·贝瑞出演了2004年的动作超级英雄电影《猫女》（图3）。图片来源：A. D. Friederici and W. Singer, “Grounding Language Processing on Basic Neurophysiological Principles, ”Trends in Cognitive Sciences 19, no. 6 (2015): 329–338, 图1。

由加州大学洛杉矶分校的伊扎克·福瑞德（Itzhak Fried）和克里斯托弗·科赫领导的研究小组发现的神经元，在50年前首次可以记录猫和猴子大脑中的单个神经元时，就已经被预测到了。研究人员认为，在大脑皮层视觉区域的层级结构中，神经元的层级越高，响应特

性就越具体。很有可能的情况是，顶层的单个神经元只会对一个人的照片产生响应。这被称为“祖母细胞假说”（grandmother cell hypothesis），祖母细胞是一个假想的脑中神经元，可以让你认出你的祖母。

更为戏剧性的是这样一组实验，给患者展示结合了两个熟习个体照片的融合图片，并要求他们想象其中一个人（偏好个体）的形象，弱化另一个（竞争个体），同时记录仅对单一熟人照片产生响应的神经元的活动。尽管视觉刺激没有改变，但受试者能够增加表征“偏好个体”神经元的放电速率，同时降低表征“竞争个体”神经元的放电速率。然后，实验者通过控制融合图片中两种照片的比例（根据神经元的放电比例）来闭合环路，受试者便可以通过想象两个个体面孔的比例[7]来控制输入。该实验表明，认知过程不是一个简单的被动过程，而是依赖于记忆和内部注意力控制的主动参与。

尽管有了这样令人惊讶的证据，祖母细胞假说仍然不太可能完全解释视觉感知的过程。根据该假说，当细胞处于活跃状态时，你会感应到你的祖母，所以它不应该被任何其他的刺激产生响应。测试中只有几百张照片，所以我们的确不知道“哈莉·贝瑞细胞”有多么强的选择性。其次，电极碰巧记录到大脑中唯一的哈莉·贝瑞神经元的可能性很低；可能性更大的情况是，大脑中有成千上万个这样的细胞，肯定还有许多对其他著名面孔、你认识的人、你能识别的每一个对象做出响应的神经元的副本。尽管大脑中有数十亿个神经元，但并不足以给人们认识的每一个对象和名称分配一个大型的专用神经元群组。最后，响应只是与感官刺激相关，但并不一定表明其中存在着因果关系。神经元的输出，及其对下游行为的影响（参考第5章介绍的投射场）也同样重要。不过，响应的选择性是惊人的。记录开始之前，病人被要求给定一个最喜欢的名人，所以可能哈莉·贝瑞在病人的脑中被过度表征了。

在小鼠、猴子和人类中同时记录数以百计的皮层神经元，为神经元如何共同感知和决策提供了另一种理论。[8]在猴子的记录中，刺激和依赖任务的信号广泛分布于大量神经元中，每一个神经元都会为刺激和任务细节特征的不同组合做出响应。[9]不久之后，我们将有可能记录数百万个神经元并操纵它们的放电速度，区分不同类型的神经元和它们彼此相连的方式。这可能会引出超越祖母细胞的理论，让我们能够更深入地理解神经元群体的活动如何引发思考、情绪、计划和决策。当然，神经元表征面孔和对象的方式可能不止一种。随着新的记录技术的出现，我们应该很快就会知道答案。[10]

自20世纪80年代以来，我们已经在经过训练的含有一层隐藏单元的神经网络模型中了解到，并且最近在深度网络中也认识到，神经网络中每个输入的活动模式呈高度分散性分布，其方式在性质上类似于皮层中神经元群组的各种响应（图9－2）。[11]分布表征可以用来识别相同对象的多个版本，而且针对同一组神经元，也可以通过对它们的输出添加不同的权重来识别许多不同的对象。对神经网络中的单个隐藏单元进行分析，就像神经生理学家记录视觉皮层中的神经元一样，有时会发现一个靠近顶层的模拟神经元对其中一个对象产生了特定的偏好。但是，因为剩下的神经元携带了表征对象的冗余信号，所以当这样的单元被切除时，神经网络的性能并不会发生明显的变化。神经网络在受到损害的情况下仍能保持稳健的性能，是网络与大脑本身的架构与数字计算机的架构之间的主要区别。

需要多少皮层神经元才能区分例如脸部之类的许多相似对象呢？从成像研究中，我们知道人脑的几个区域对人脸能够做出反应，有些区域具有高度的选择性。但是，在这些区域内，任何单个面孔的信息都广泛分布在许多神经元中。加州理工学院的曹颖（Doris Tsao）在猴子的大脑皮层神经元中，记录到对面部的选择性响应，并表明可以通过结合来自200个面部细胞（所有面部选择性神经元的相对较小的子集）的输入来对面部特征进行重建。[12]

视觉感知的时机

视觉意识的另一方面，是大脑努力记录发生在特定时间的事件，例如闪光。视觉皮层中神经元对闪光视觉刺激产生响应的时间延迟从25毫秒到100毫秒不等，这些延迟通常发生在皮层同一区域内。尽管这样，我们仍可以确定时间差在40毫秒内发生的两次闪光的次序，以及时间差小于10毫秒的两次声音的次序。让问题更加复杂的是，视网膜中的处理也需要一定的时间。这一时间并不是固定的，而是取决于闪光的强度，所以即便微弱闪光和强烈闪光同时发生，微弱闪光中来自视网膜的首个脉冲的到达时间与强烈闪光相比，也会有一个延迟。这就引出了一个问题，即为什么视觉感知似乎具有一致性，虽然从整个皮层活动的时间和空间分布模式来看并不明显。

当我们进行跨模态（cross-modal）比较时，同时性问题变得更加令人困惑。你看到一个人在砍树，假设你离得足够近，即使声速比光速慢得多，你也可以同时看到并听到每次斧头对树木的砍击。而且随着与树的距离不断增加，同时性的幻觉仍然维持着，[13]直到视觉和听觉信号到达你的大脑的时间差大于80毫秒时，我们就不再认为声音与砍击的动作同时发生（此时的距离大约为100英尺）。

探索与时间相关的视觉行为的研究人员发现了另一种被称为“闪光—滞后效应”（flash-lag effect）的现象。该现象可以表现为飞过头顶的飞机闪烁的尾灯和机尾看起来不太一致——闪光的位置似乎落后于机尾。另一种常见的现象是在足球比赛中，当一名跑动中的运动员看起来像在足球（闪光）前面时，这可能会引起助理裁判员的越位判罚——这名助理裁判员并未做出幻觉补偿[14]。这一现象可以在实验室中通过用视觉刺激（如图16－3所示）来进行研究。在闪光滞后效应中，闪光和位于同一位置的移动物体之间似乎发生了错位。

对这种现象存在一种主流的解释——在直觉上更说得通，而且有一些来自大脑实验记录的证据——是大脑预测了运动点将在短时间之

后出现在哪个位置。但是感知实验已经表明，这并不能解释闪光滞后效应，因为闪光时间引起的感知取决于闪光后80毫秒内发生的事件，而不是闪光之前发生的事件（先前发生的事会被大脑用来做出预测）。[15]这种对闪光滞后效应的解释，意味着大脑是“后发性的”而非“预测性的”；也就是说，大脑不断修改历史，使有意识的现在与未来保持一致。这是我们的大脑如何基于嘈杂和不完整的数据产生似是而非感觉的一个例子，这一现象也经过魔术师的探索，被用于魔术表演中。[16]

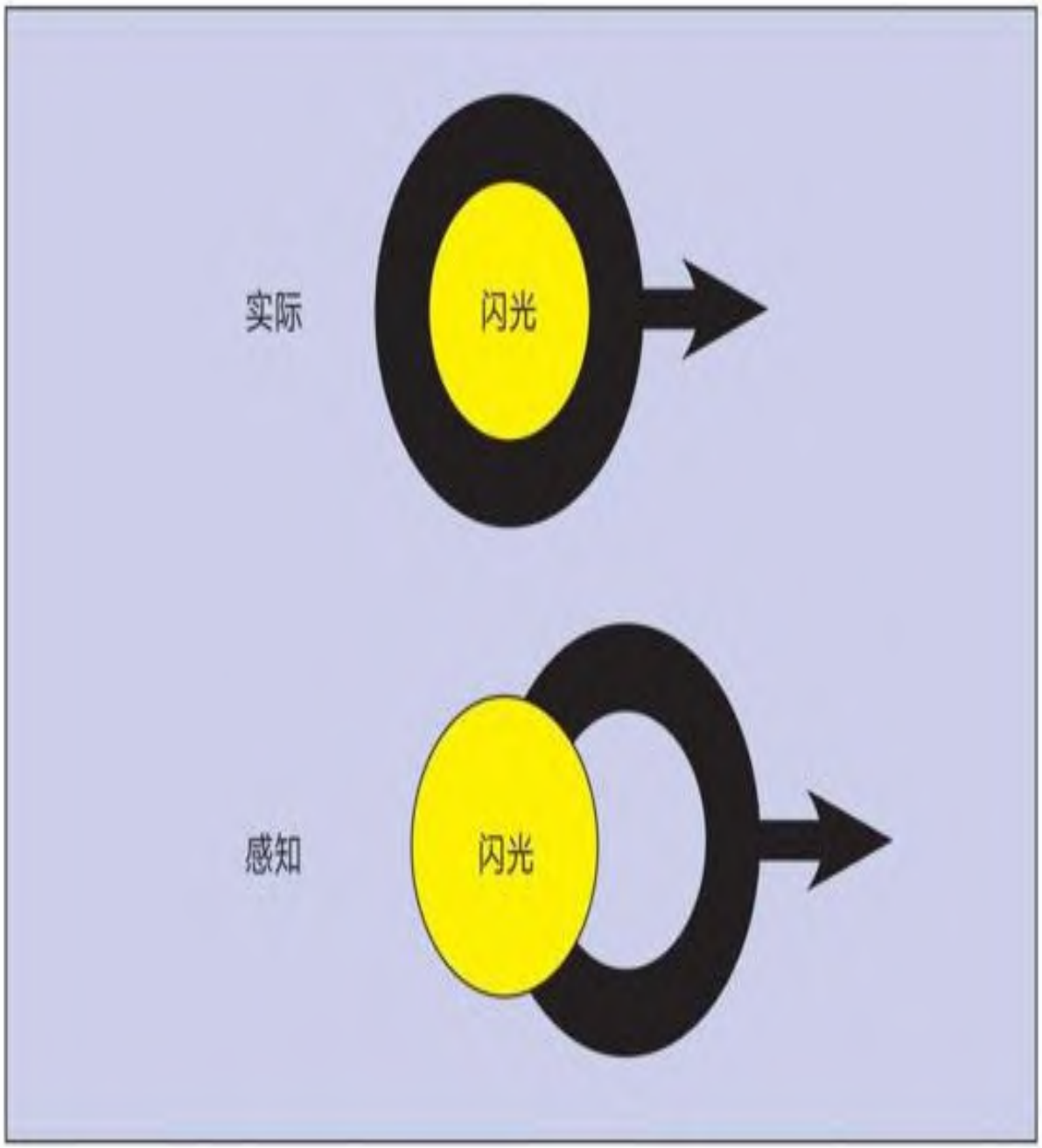


图16-3 闪光滞后效应。（上图）一个圆环从左向右移动（黑色）。当它通过视野的中心时，灯光会短暂闪烁（黄色）。（下图）被试者报告，在闪光时圆环似乎偏移到了右侧。图片来源：大卫·伊格尔曼。

视觉感知的部位

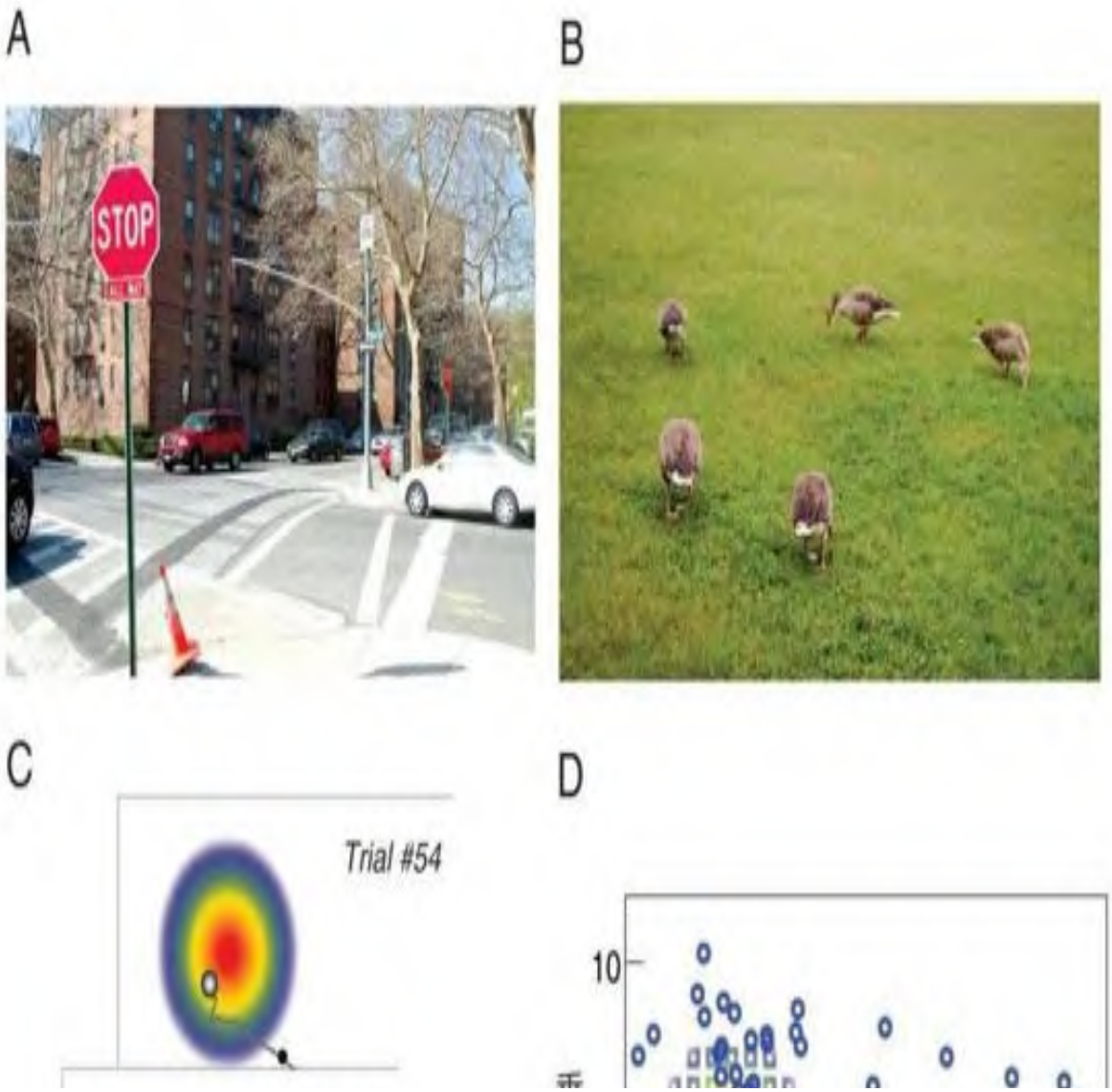
大脑成像可以在我们感知事物时为我们提供大脑活动的全局图。根据实验证据，研究人员提出了一个非常有吸引力的假设，即当皮层前部（这部分对计划和制定决策非常重要）的大脑活动达到阈值水平，并开启反馈回路时，我们才会有意识地注意到某些事情。[17]虽然具有一定的启发性，但这些观察结果并不具有说服力，因为它们并不构成因果关系，而只是表明了相关性。如果“意识的神经相关”是意识状态的原因，那么应该可以通过改变“意识的神经相关”来改变意识状态。曹颖在她2017年的实验中证实了这一点；通过刺激猴子的视觉皮层中的脸部识别功能区，她能够干扰它们对脸部的辨别能力。[18]当对人类进行类似的实验时，被试者报告说，脸部看起来好像正在融化。[19]

最近，诸如光遗传学[20]等新技术已经能够被用于选择性地操纵神经元的活动，从而对NCC的因果关系进行测试。如果认知状态与高度分散的活动模式相对应，这样的测试难度可能会很大，但原则上，这种方法可以揭示意识中的感知和其他特征是如何形成的。[21]

视觉搜索的机理

视觉搜索任务依赖于自下而上的感官处理，以及自上而下由期望驱动的注意力处理（图16－4A）。这两种类型的处理在大脑中相互交织着进行，难以区分。但最近，人们开发了一种新颖的搜索任务来尝试将它们分开。[22]参与者坐在空白的屏幕前，他们的任务是用目光扫视屏幕，找到一个隐藏目标的位置。当他们的目光停留在目标附近时，会听到奖励音。隐藏目标的位置在每次试验中都会发生变化，并且服从高斯分布（一种钟形曲线，由其峰值和宽度定义），但在一个试验环节内保持不变，然而参与者并不了解这一信息（图16－4D）。

在试验环节开始时，参与者并没有任何先验知识来指导他们的搜索。一旦获得奖励，他们就可以使用该反馈来协助下一次试验。随着试验的进行，参与者通过制定隐藏目标分布的预期，来指导未来的搜索，以提高成功率。经过大约十几次的试验后，参与者的视觉注意力缩小到了目标出现概率高的区域。图16－4D显示了在所有参与者身上出现的这种效应的特征。随着实验不断推进，搜索区域的范围开始大幅缩小。令人惊讶的是，尽管被试者在几次试验后的第一次扫视总是能到达隐藏目标分布的中心，但他们中许多人并没有办法描述出是如何做到这一点的。



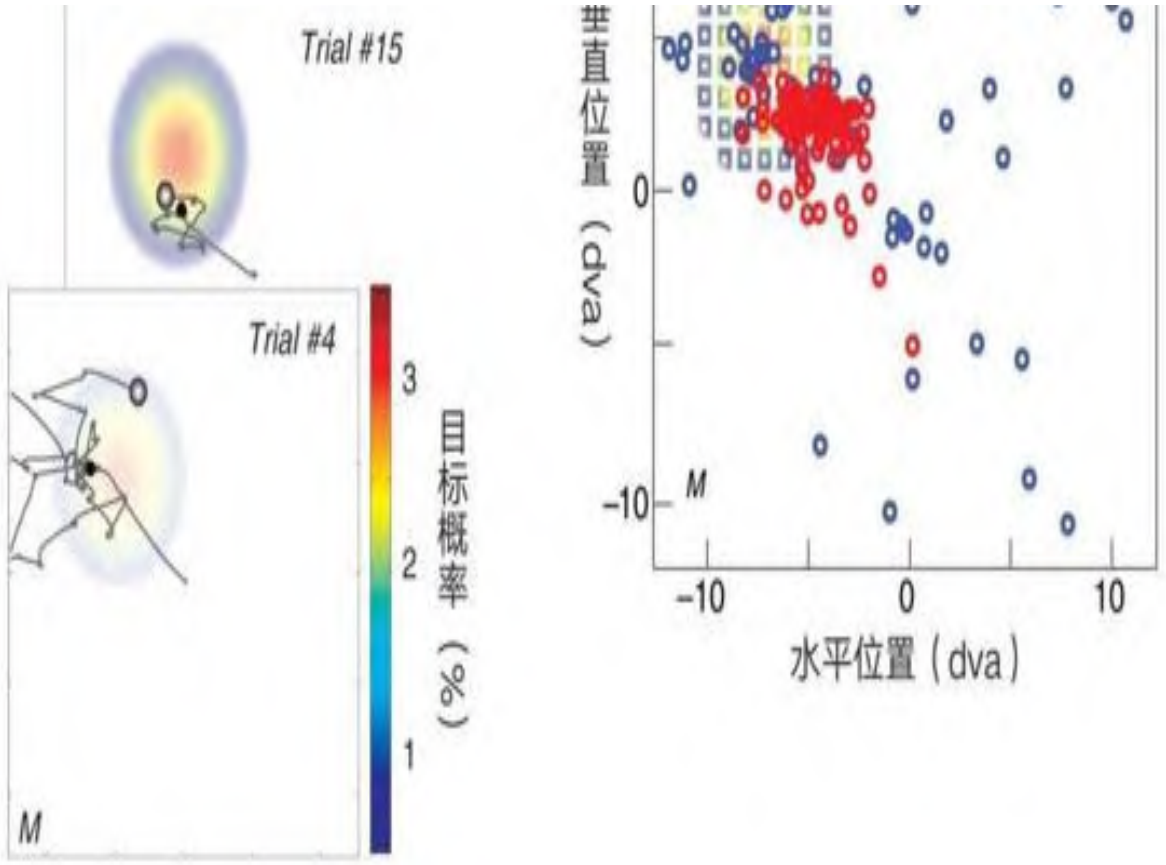


图16-4 学习朝哪里看。（A）有经验的行人事先知道在街道场景中寻找路标、汽车和人行道的位置；（B）在一片草地上觅食的鸭子；（C）屏幕的表征图像与试验期间学习的隐藏目标分布区域进行的叠加，以及来自参与者M的三次试验的目光轨迹样本。每个试验中第一次目光停留的区域用黑色记号标记。最终获得奖励的注视位置以带阴影的灰色点进行标记。（D）对目光注视过的屏幕区域的采样，从在早期试验中遍布了整个屏幕（浅灰色圆圈；前五个试验），缩小到接近高斯整数分布大小和位置的区域［正方形，颜色与后面试验C中给出的概率成比例（红色圆圈；试验32-39）］。图片来源：L. Chukoskie, J. Snider, M. C. Mozer, R. J. Krauzlis, and T. J.Sejnowski, “Learning Where to Look for a Hidden Target”，图1。

这些实验表明，大脑无意识的控制是由经验引导的。通过消除视觉输入，可以独立地研究无意识过程。这项搜索任务涉及的大脑区域包括视觉皮层和上丘，它们分别控制着视野中对象的立体形貌分布，以及分别将视线指向视觉目标，与眼动系统的其他部分紧密合作。学习过程也涉及基底神经节，这是脊椎动物大脑的一个重要组成部分，能通过强化学习来学习按顺序发生的动作。[23]预期和收到的奖励之间的差异，表现为中脑部位的多巴胺神经元放电频率的瞬时增加，从

而能够调节突触可塑性，并在无意识层面影响决策和计划的制定（已在第10章中讨论过）。

创造意识比理解意识更容易

在去世之前，弗朗西斯·克里克还邀请我到他家中讨论屏状体（claustrum），这个皮层下的神秘细胞薄层接收来自许多皮层区域的投射，然后再反向投射回去。尽管克里克当时已身患绝症，但仍然专注于完成他的最后一篇论文，即屏状体因其中心地位而负责意识统一的假说。只有少数研究人员曾研究过屏状体，他几乎给他们中的每个人都打过电话，寻求进一步的信息。这次拜访是我最后一次见到他。弗朗西斯于2004年7月28日去世时，仍然在努力撰写他最后一篇论文[24]的手稿，试图完成他对意识起源的研究。

在他和詹姆斯·沃森于1953年发现DNA结构的50年后，人类基因组测序的项目完成了。克里克告诉我，他从来没有想过人类能够完成这一壮举。50年后，我们距离探清意识问题还有多远呢？届时，我们将拥有与我们互动的机器，就像我们现在通过语音、手势和面部表情与他人互动一样。创造意识要比完全理解它来得容易。

我怀疑，我们是否可以通过首先了解无意识的处理过程——我们看到、听到和行动时认为理所当然的一切——来加快进度。我们已经在理解激励系统方面取得了进展（激励系统对我们如何做决定产生了强烈的影响）；还有注意力系统（这些系统有助于指导我们在现实世界中搜寻信息）。随着对管理感知、决策和规划的大脑机制的深入了解，理解意识这个问题可能会像《爱丽丝漫游奇境记》中的柴郡猫一样消失，只留下一个大大的笑容。[25]

---

[1] 本章引自T. J. Sejnowski, “Consciousness,” *Daedalus*144, no. 1 (2015): 123–132. 参阅Francis H. C. Crick, *What Mad Pursuit: A Personal View*

of Scientific Discovery(New York: Basic Books, 1988); Bob Hicks, “Kindra Crick’s Mad Pursuit,” *Oregon Art Watch*, December 3, 2015. http://www.orartswatch.org/kindracricks-mad-pursuit/.

[2] 现在还没有统一的关于“意识”（consciousness）的科学定义，意识被用来指代许多不同的现象。对这个术语广义上的理解，包括保持清醒的状态，对自己周围环境的认知，对事物的认知或感知，以及大脑对自身和外界的认知。

[3] F. Crick, “The Function of the Thalamic Reticular Complex: The Searchlight Hypothesis,” *Proceedings of the National Academy of Science of the United States of America*81, no. 14 (1984): 4586–4590.

[4] F. Crick and C. Koch, “The Problem of Consciousness,” *Scientific American*267,no. 3 (1992): 10–17; F. Crick and C. Koch, “Constraints on Cortical and Thalamic Projections: The No-Strong-Loops Hypothesis,” *Nature*391, no. 6664 (1998): 245–250; F. Crick and C. Koch, “A Framework for Consciousness,” *Nature Neuroscience*6, no. 2 (2003): 119–126; and F. Crick, C. Koch, G. Kreiman, and I. Fried,“Consciousness and Neurosurgery,” *Neurosurgery*55, no. 2 (2004): 273–281.

[5] F. Crick and C. Koch, “Are We Aware of Neural Activity in Primary Visual Cortex?”*Nature*375, no. 6527 (1995): 121–123; C. Koch, M. Massimini, M. Boly, and G.Tononi, “The Neural Correlates of Consciousness: Progress and Problems,” *NatureReviews Neuroscience*17 (2016): 307–321.

[6] R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried, “Invariant Visual Representation by Single Neurons in the Human Brain,” *Nature*435, no. 7045(2005): 1102–1107.

[7] 这种比例是通过对不同神经元的放电速率的比例来定量的。——译者注

[8] K. Deisseroth and M. J. Schnitzer, “Engineering Approaches to Illuminating Brain Structure and Dynamics,” *Neuron*80, no. 3 (2013): 568–577.

[9] V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome, “Context-Dependent Computation by Recurrent Dynamics in Prefrontal Cortex,” *Nature*503, no. 7474(2013): 78–84.

[10] BRAIN Working Group, *BRAIN 2025: A Scientiific Vision,*Report to the Advisory Committee to the Director, NIH (Bethesda, MD: National Institutes of Health, June 5, 2014), 36. https://www.braininitiative.nih.gov/pdf/BRAIN2025_508C.pdf.

[11] Patricia Smith Churchland and Terrence J. Sejnowski, *The Computational Brain*,2nd ed. (Cambridge, MA: MIT Press, 2016), 183, 221.

[12] L. Chang and D. Y. Tsao. “The Code for Facial Identity in the Primate Brain,” *Cell*169, no. 6 (2017): 1013–1028.e14.

[13] D. A. Bulkin and J. M. Groh, “Seeing Sounds: Visual and Auditory Interactions in the Brain,” *Current Opinion in Neurobiology*16 (2006): 415–419.

[14] 闪光滞后幻觉表现为主体对突现客体的位置判断滞后。——译者注

[15] D. M. Eagleman and T. J. Sejnowski, “Motion Integration and Postdiction in Visual Awareness,” *Science*287, no. 5460 (2000): 2036–2038.

[16] 参阅 Stephen L. Macknik, Susana Martinez-Conde, and Sandra Blakeslee, *Sleightsof Mind: What the Neuroscience of Magic Reveals about Our Everyday Deceptions*(New York: Henry Holt, 2010)。

[17] S. Dehaene and J.-P. Changeux, “Experimental and Theoretical Approaches to Conscious Processing,” *Neuron*70, no. 2 (2011): 200–227.

[18] S. Moeller, T. Crapse, L. Chang, and D. Y. Tsao, “The Effect of Face Patch Microstimulation on Perception of Faces and Objects,” *Nature Neuroscience*20, no. 6(2017): 743–752.

[19] J. Parvizi, C. Jacques, B. L. Foster, N. Withoft, V. Rangarajan, K. S. Weiner, and K.Grill-Spector, “Electrical Stimulation of Human Fusiform Face-Selective Regions Distorts Face Perception,” *Journal of Neuroscience*32, no. 43 (2012): 14915–14920.

[20] BRAIN Working Group, *BRAIN 2025: A Scientific Vision*, pp. 6, 35, 48.

[21] Sejnowski, “What Are the Projective Fields of Cortical Neurons?”

[22] L. Chukoskie, J. Snider, M. C. Mozer, R. J. Krauzlis, and T. J. Sejnowski, “Learning Where to Look for a Hidden Target,” *Proceedings of the National Academy ofSciences of the United States of America*110, supp. 2 (2013): 10438–10445.

[23] T. J. Sejnowski, H. Poizner, G. Lynch, S. Gepshtein, and R. J. Greenspan,“Prospective Optimization,” *Proceedings of the IEEE*102, no. 5 (2014): 799–811.

[24] 在他去世后，他的同事克里斯托弗·科赫（Christof Koch）继续完成了这项工作。F. C. Crick and C. Koch, “What Is the Function of the Claustrum?”

Philosophical Transactions of the Royal Society of London B360, no. 1458 (2005),1271–1279.

[25] 参阅 Lewis Carroll, *Alice's Adventures in Wonderland*(London: Macmillan,1865), chap. 6.

17
进化的力量

研究生命起源的莱斯利·奥格尔（Leslie Orgel，图17－1，右）是一位毕业于牛津大学的化学家，他是我在索尔克研究所多年的同事，也是我见过的最聪明的科学家之一。每周五的教师午餐时与他的讨论总是让我意犹未尽。生命的起源可以追溯到数十亿年前，众所周知，那时的地球与今天大不相同，环境条件十分苛刻，大气中的氧气很少，并不适宜生命存活。古细菌早于细菌出现在地球上，但在古细菌之前出现的又是什么呢？我们现在已经知道，DNA普遍存在于所有细胞中，那在DNA形成之前就存在的遗传物质又是什么呢？1968年，莱斯利·奥格尔和弗朗西斯·克里克推测，由细胞中的DNA衍生而来的RNA可能是DNA的前身，但这就要求RNA能够自我复制。关于这一推测的证据，来自可以催化RNA反应的基于RNA酶的核酶[1]（ribozyme）。今天该领域的大多数研究人员都认为，所有生命都来自早期的“RNA世界”，这一点是完全有可能的。但RNA又是从哪里来的呢？遗憾的是，我们并没有什么证据可以继续追溯下去。



图17-1 1992年，索尔克研究所的弗朗西斯·克里克（左）和莱斯利·奥格尔（右）分别在意识的起源和生命的起源领域进行着开拓性的研究。图片来源：索尔克研究

所。

大自然比我们聪明

一直以来，人们普遍接受的所谓“真理”，总是不断地被惊人的发现所颠覆。我们抬起头，看到太阳在围绕地球转动，但实际上是地球在绕着太阳转。进化论让我们认清了人类的发展历程，尽管到了今天，很多人仍然难以接受这一理论。许多年后，我们的后代将回顾我们生活的这个时代，并认为我们对智能的直觉，是过于简化了的，这种直觉阻碍了人工智能后续50年的进步。正如奥格尔的第二定律所述，进化比你聪明。

我们的意识觉知只是冰山一角，我们大脑的大部分活动依然神秘莫测，无法进行自我反思。我们用“注意力”和“意图”这些词来描述我们的行为，但这些都是含糊的概念，隐藏了大脑活动过程的内在复杂性。基于直觉大众心理学的人工智能的发展迟迟得不到令人满意的成果。我们的眼睛能看到东西，但没人知道其背后的原因。“我思故我在”，但思考背后的机制仍是一个谜。大自然向我们揭示大脑如何运转，并不会给我们带来生存上的优势。奥格尔的第二定律仍然成立。[2]

正如第2章所指出的那样，我们拥有高度发达的视觉系统，但这并不能让我们成为视觉领域的专家。[3]许多人甚至都没有意识到，我们的眼部存在一个眼中央凹，其能提供的清晰视角只有1度的弧度，相当于以一臂为半径，拇指长度为弧长所对应的角度，另外，我们在眼中央凹之外什么都看不到，几乎相当于法定盲人[4]。我曾经向我的母亲指出这一点，她说她不相信我，因为她前后左右都看得很清楚。由于我们可以迅速重新定位视觉目标，所以产生了一种在目光所及之处都能看得很清楚的错觉。你知道当你凝视一个目标时，目光会以每秒三次的频率在目标上游移吗？余光（peripheral vision）可能具有

较低的空间分辨率，但对亮度和运动的变化非常敏感。在视觉皮层中，与识别目标的主要回路相区别的另一个主要回路，就负责识别视线在空间中的移动。

当计算机视觉领域的先驱开始设计视觉功能时，他们的目标是从图像中创建出一个完整的世界内部模型，而这个目标已经被证明是难以实现的。但是，一个完整而准确的模型对于大多数实际目的来说可能是不必要的，而且考虑到目前的摄像机取样率很低，这似乎更不可能实现。

根据心理物理学、生理学和解剖学的证据，帕特里夏·丘奇兰德、神经心理学家拉马钱德兰（V. S. Ramachandran）和我得出的结论是，大脑只表征世界上有限的一部分目标，即执行手头任务所需要的那部分。[5]这也使得强化学习更容易减少所需要的、可能有助于获得奖励的感官输入的数量。视觉的明显模块化（与其他感官处理流相比，视觉信息有更高的分离度）也是一种错觉。视觉系统集成了来自其他渠道的信息，包括来自奖励系统的指示场景中目标价值的信号。机动系统通过积极地重新定位传感器来寻找信息，例如移动目光，以及在某些物种中，通过移动耳朵来收集可能导致奖励行为的信息。

大脑通过对环境的逐步适应而发生进化；大自然无法从零开始，而必须通过修改各个零部件来维持现有物种的生存。约翰·奥尔曼（John Allman）在他的著作《进化脑》（*Evolving Brains*）[6]一书中阐述了在城市人类尺度上逐渐发生的进化，并回顾了一次他去圣迭戈老电厂锅炉房的经历。他注意到，在一组真空管旁边有一排排复杂的小型气送管，旁边是不同年代的计算机控制系统。由于该电厂需要连续供电，因此无法在每次对技术进行更新换代时，都直接关闭旧系统并将其改装为新系统。于是旧的控制系统就被留在了原位，而新的控制系统被集成到了其中。同样，随着大脑的不断变化，大自然并不能抛弃旧的大脑系统，而是根据当前的发展计划进行调整，偶尔也

会增加一层新的控制。基因复制，是引入能够变异出新功能的基因副本的最佳途径。全基因组复制的情况曾经也发生过，而这可能会导致一个全新物种的诞生。

认知科学的兴起

20世纪30年代，研究学习行为的心理学家把行为看作将感官输入转化为运动输出的过程，并自称为“行为主义者”。关联学习是行为主义的研究重点，许多学习规律在通过用不同的奖励机制训练动物时被陆续发现。哈佛大学的B. F. 斯金纳是这一领域的领导者，他撰写了几本畅销书籍来解释他的发现对社会造成的影响。[7]当时有关行为主义的大众出版物非常受欢迎。

1971年，著名语言学家诺姆·乔姆斯基（图17－2）在《纽约书评》（图17－3）上发表了一篇全面抨击行为主义的文章，矛头直指斯金纳。[8]下面是他的论点中关于语言方面的一段摘录：

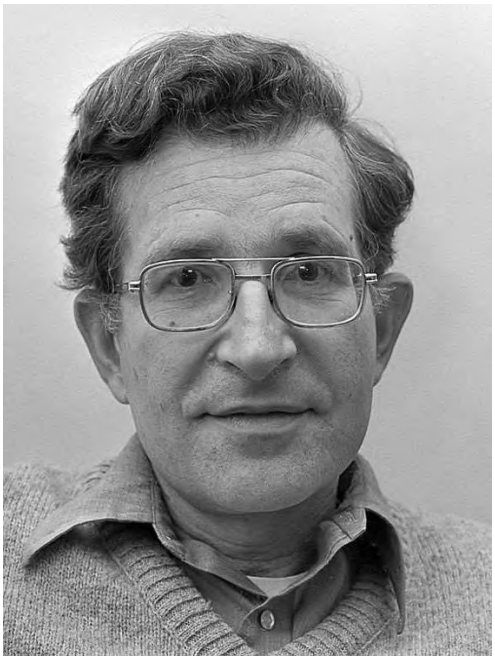


图17-2 1977年的诺姆·乔姆斯基。他于1971年在《纽约书评》上撰写了文章《针对斯金纳的驳斥》。乔姆斯基的论文对一代认知心理学家产生了深远的影响。他们开始将符号处理作为认知的概念框架，而大脑发育和学习在认知和智力中的重要作用则在他们的眼中大打折扣。图片来源：Hans Peters/Anefoto。

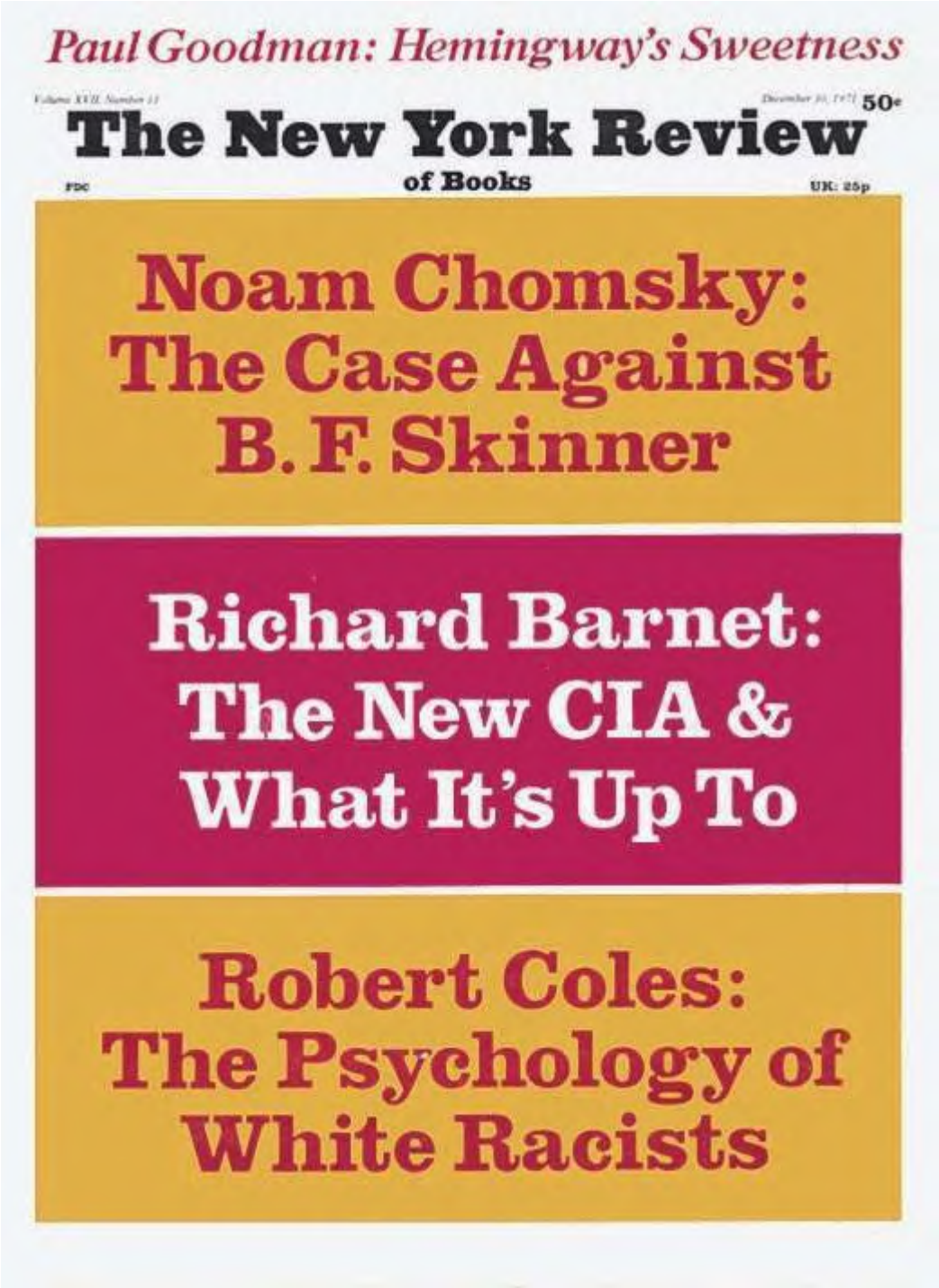


图17-3 诺姆·乔姆斯基于1971年在《纽约书评》中抨击斯金纳的标题。乔姆斯基的文章影响了一代科学家，诱导他们放弃了对行为的学习，并采用符号处理作为解释认知的方式。但使用符号方法，人工智能从未出现能够达到认知水平的表现。斯金纳在强化学习领域引领着正确的方向，但却受到了乔姆斯基的嘲讽；当今最引人注目的AI应用是基于学习，而非逻辑。图片来源：《纽约书评》。

> 把一些我从来没有听过或说过的英语句子，而不是中文句子说成属于我的“语录”（只因为我说前者的“概率”更大），这是什么意思？斯金纳主义在讨论中涉及这一点时，一直诉诸“相似性”或“泛化”，但始终没有精确地描述新句子与熟悉的例子如何“相似”，或从例子中如何“泛化”的方式。这种失败的原因很简单。就目前所了解的信息而言，相关属性只能用描述有机体内部假定状态的抽象理论（例如语法）来表达，而这种理论被从斯金纳的“科学理论”中先验地排除了。由此导致的直接后果就是，只要讨论涉及事实世界，斯金纳主义就必然会陷入神秘主义（不能解释的“相似性”和无法描述的“泛化”）。尽管在语言中情况可能更为清楚，但没有理由认为，人类行为的其他方面也会落入斯金纳用先验限制了的“科学”体系之中。[9]

以今天的角度，我们可以看到乔姆斯基明白了问题的关键，但他并不了解学习的力量。深度学习已经向我们展示了，像大脑本身的神经网络那样，模型神经网络能够做到被乔姆斯基解释为“神秘主义”的那种“泛化”，并且可以经过训练，选择性地识别多国语言并进行翻译，为图像生成在语法上无可挑剔的标题。最具讽刺意味的是，机器学习已经解决了自动解析语句的问题，尽管计算语言学家做出了艰苦的努力，但乔姆斯基的句法“抽象理论”从来没有解答过这个问题。当与斯金纳开创的强化学习结合起来时，机器学习就可以解决需要做出一系列选择以实现目标的复杂问题。这是解决问题的实质，也是智能的基础。

乔姆斯基的文章充满了鄙视，其影响也远远超过了对斯金纳的抨击：它挑战了，甚至驳斥了将学习作为理解认知的一种方式。这在20

世纪70年代对认知心理学产生了决定性的影响。乔姆斯基所持观点的核心在于，根本无法想象（至少对乔姆斯基而言）关联式学习能够产生一种像语言那样复杂的认知行为。但请注意，这个论点是基于无知——仅因为世界顶尖的语言学家说他无法想象某件事，并不能否定这件事发生。但乔姆斯基的言辞与70年代的时代精神产生了共鸣，因此颇受追捧。到了20世纪80年代，对认知进行符号处理已成为主流方法，并为一个名为“认知科学”的新领域打下了基础——认知科学是认知心理学、语言学、哲学和计算机科学的混合体。神经科学则像是认知科学的小兄弟，在20世纪90年代认知神经科学兴起前，一直或多或少地受到忽视。

不能把语言问题只留给语言学家

此后，乔姆斯基还多次使用了同样的修辞论证，其中最著名的是他基于“刺激匮乏”（poverty of the stimulus）对先天语言能力的论证，[10]其中断言婴儿在听觉上获取不到足够的语言范例来学习句法的规则。但是，婴儿并不是一台从外界得到一串无形符号的计算机。事实上，一个婴儿沉浸在一个充满丰富感官体验的世界中，并会以惊人的速度了解世界的本质。[11]这个世界充满了与声音有关、意义非凡的体验，这些体验始于子宫，是一种无监督学习。在这一基础上，语言开始形成，首先是咿咿呀呀，然后是单个词汇，以及后来语法正确的单词序列。先天习得的并不是语法，而是能够从经验中学习语言，表现出在丰富的认知语境中吸收话语的高阶统计特性的能力。

令乔姆斯基无法想象的是，如果加上对环境的深度学习和用毕生的经验磨炼出的深度学习的代价函数，像强化学习这样的弱学习系统也可以产生认知行为，包括语言。在20世纪80年代，这一点对我来说并不明显，尽管我应该认识到，如果像话语网络这样的小型网络可以处理英语发音，那么很可能在对言语的表征中，学习网络——无论是

模型网络还是皮层网络——都对语言有着天然的亲和力。乔姆斯基的立场是基于想象力的匮乏，但从逻辑来说符合了奥格尔的第二定律：进化比你聪明，而这个“你”也包括像乔姆斯基这样的专家。事实上，当一位专家告诉你自然界的某些事情是不可能的时候，你应该保持谨慎——不管这个论证有多么合理或者令人信服。

在20世纪后半叶，乔姆斯基对语序和句法的强调成为语言学的主导方式。但即使是一个“词袋”（bag of words）模型的神经网络，抛弃单词顺序，也能在判定文章（如体育报道或政论）的主题方面表现出色，并且通过参考直接相邻的文字关系，性能还可以得到进一步的提升。我们通过深度学习获得的经验是，即使词序包含一些信息，但基于单词含义及词间关系的语义更重要。词汇在大脑中被丰富的内部结构所表征。随着我们对在深度学习网络中词汇如何表达语义有了更深入的了解，我们可能会开启一门新的语言学。就像大自然没有理由向我们揭示视觉的工作原理一样，也没有理由认为，我们对语言工作原理的直觉会更好。

我们来考虑一下，在对自然语言任务进行训练的模型网络中，单词的内部结构可能是什么样子。虽然网络可能是针对特定问题进行训练，但网络表征输入的方式可被用于解决其他问题。一个很好的例子就是训练一个网络，来预测句子中的下一个单词。在经过训练的网络中，单词的表征方式具有内部结构（以网络中所有单元的活动模式的形式），可以用来在单词对之间进行类比。[12]例如，当这些活动模式被投影到一个平面时，连接国家和首都的矢量都是一样的。在没有任何关于首都城市意味着什么的监督信息的前提下（图17－4），网络学会了自动组织概念并隐式地学习它们之间的关系，这就表明，可以使用无监督学习从文本中提取国家和首都的语义。

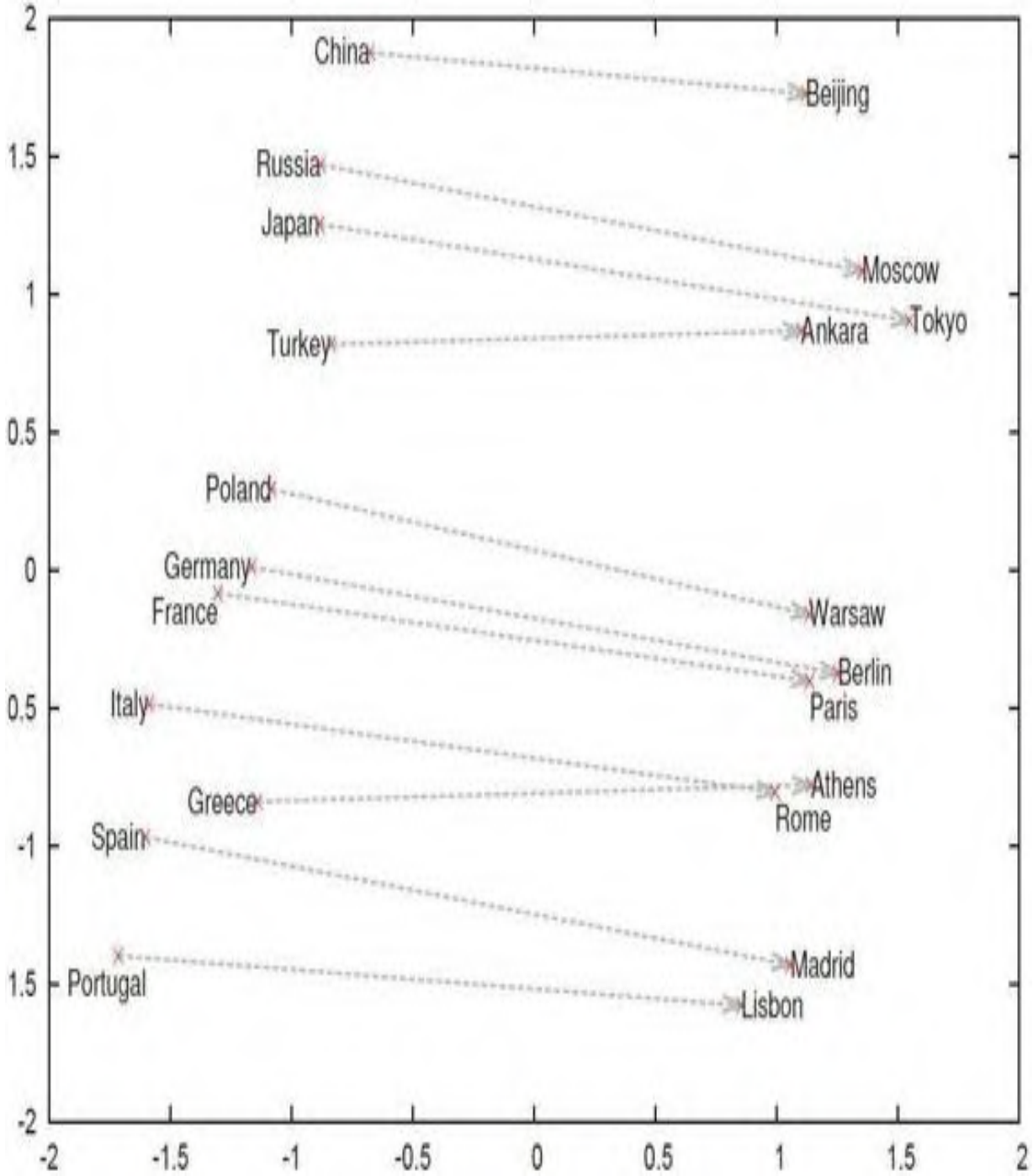


图17-4 网络中单词的内部表征被用来训练预测句子中的下一个单词。每个单词都是网络活动的矢量，可以如上所示投影到二维平面上。箭头将国家连接到它们的首都。由于所有这些箭头彼此平行并且长度大致相同，所以单词对也以类似的方式表示。例如，如果你想查找不同国家的首都，可以将此箭头添加到该国家的矢量中，并检索出其首都的矢量。资料来源：T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, “Distributed Representations of Words and Phrases and Their Compositionality”，图2。图片来源：杰夫·迪恩。

我在麻省理工学院的一次演讲中，开门见山地说道：“语言太重要了，不能只留给语言学家去研究。”[13]我的意思是，我们不应该只停留在行为层面描述语言。我们应该理解语言背后的生物学原理和潜在的生物学机制，以及智人的语言能力是如何演变的。使用无创大脑成像，以及直接对癫痫患者的大脑活动进行记录，已经使这一想法成为可能。对大脑研究来说，同等重要的是通过比较人类大脑与黑猩猩和其他高等灵长类大脑的差异，来理解语言是如何形成的；相比在早期经过更加漫长的过程获得感觉运动技能，使用语言的能力发生在进化的某个瞬间。强大的基因工具将使我们能够剖析大脑的发育，并理解进化如何通过在发育中不断修正来形成我们先天的语言学习能力。

语言可以通过诉诸似是而非和源自无知的争论被用来进行误导和控制，这种无知会产生远远超出科学范畴的负面影响。历史上充斥着煽动者，当他们在想象力方面的匮乏被暴露出来，最终就会难逃被抛弃的命运。幸运的是，大脑存在的时间比语言要长很多，如果我们依赖于早在语言出现之前就进化出来的那部分大脑，肯定会从中受益。[14]

难预测的行为规律

回想起来，行为主义和认知科学在20世纪对行为的研究采取了相反的方式，但都犯了同样的错误，即忽视了大脑的重要性。行为主义者不希望被反思误导，所以他们认为不去大脑中寻求指导，实际上是一种荣耀。他们认为，通过仔细控制黑匣子的输入和输出，可以发现任何偶然事件的行为规律。“功能主义”的认知科学家拒不认可行为主义，并且相信他们可以发现思维的内在表征方式，但是因为他们也相信大脑如何实现这些表征的细节无关紧要，[15]因此他们所开发的

内部表征是基于不可靠的直觉和大众心理学。大自然比这两派都要聪明。

黑匣子的内部状态非常复杂，因此要发现内部表征和行为规律非常困难。如果有一天我们确实发现了行为规律，很可能会对它们进行功能性的描述，尽管这个描述可能和量子力学对于物理学家一样，是违反直觉的。想要发现行为规律，我们需要尽可能依赖大脑。深度学习网络是通过关注大脑结构的某些一般特征，以及大脑功能的一般原理来取得进步的范例。我毫不怀疑强硬的功能主义者会提出抗议，但我们需要前进，而不是墨守成规。在此过程的每一步中，添加来自大脑架构的一系列新特征都提升了深度学习网络的功能：皮层区域的层次结构，深度学习与强化学习在大脑中的耦合，循环皮层网络中的工作记忆，以及关于事实和事件的长期记忆，等等。我们还可以从更多的大脑计算原理中学习，并充分利用它们。[16]

研究知觉、记忆和决策的神经科学家通常在他们的实验中使用基于试验的任务，在这些任务中，实验动物接受训练以对刺激产生期望的反应。经过数月的训练后，这些刺激驱动的反应变成了反射，而不是反思，这可以揭示习惯行为，而不是认知行为背后的机制。思考不是一种反射，它可以在没有任何感官刺激的情况下发生，但传统的实验设计方式忽略了在没有感官输入的情况下进行的自发活动。我们需要新的方法来研究与感官和运动无关的内部活动，其中包括有意识思考和无意识处理。出现了这样一种新的情况：大脑成像实验揭示了当人被放入扫描仪并被要求“休息”时，所自发产生的静息状态。当无所事事时，意识则表现为一种飘忽不定的模式，这是一种我们可以观察到但尚未理解的大脑活动。

脑成像，尤其是无创性功能磁共振成像（fMRI）开辟了研究社会互动和决策制定的新途径，形成了一个名为“神经经济学”的新领域。[17]因为人类并不是经典经济学中经常假设的理性行为者，我们的判断和动机来自复杂的大脑内部状态，因此我们需要根据实际而非

理想化的人类判断和动机，来建立一个行为经济学。[18]正如第10章所述，多巴胺神经元通过表现奖励预测误差，对动机产生了强大的影响。针对社交互动的大脑成像，能够以纯粹的行为实验无法实现的方式探究人类的动机。我们的目标是用基于先前经验的概率决策理论，来取代基于逻辑的理性决策理论。

神经网络的寒冬

神经网络的早期历史，其实就是一个不大但颇具影响力的团队，如何能够将研究方向带离正轨的案例。在《感知器》这本书尾声部分，马文·明斯基和西摩尔·帕普特（图17－5）表达了这样的观点：感知器学习算法并不能扩展到多层感知器：



图17-5 照片拍摄于1971年马文·明斯基和西摩尔·帕普特出版了《感知器》一书后不久。其中对简单网络出色的数学分析，让一代追求基于多层网络学习的人工智能方法的研究人员，感到不寒而栗。图片来源：麻省理工学院。

这个扩展问题不仅仅是技术问题，也是战略问题。尽管（甚至是因为）它有着严重的局限性，感知器已经显示出了研究价值。它具有很多引人注目的特点：线性，有趣的学习理论，清晰

简单的范例适用于并行计算。但没有理由认为，这些特点中的任何一种会延续到多层次版本。尽管如此，我们仍然认为这是一个重要的研究课题，以阐明（或否决）我们对“其扩展是无所作为的”的直觉判断。我们也许会发现一些强大的收敛定理，或者相反地，我们会找到无法为多层机器研发有趣的“学习理论”的原因。[19]

毫无疑问，其结果导致了网络学习领域的寸草不生。在明斯基和帕普特的书中，这种毫无根据的“直觉”（除此之外，这倒是一本好书）对神经网络学习的发展产生了令人不寒而栗的影响，让一代人的研究就此停滞不前。尽管我个人从这种停滞中受益，因为它使我的事业成为可能。但是我有机会在明斯基职业生涯后期，站在幕后洞察一切。

我被邀请参加2006年达特茅斯人工智能会议“AI@50”，回顾1956年在达特茅斯举行的开创性的人工智能夏季研究项目，并讨论人工智能的未来。[20] 1956年该项目的十位先驱中的五位出席了这次会议：约翰·麦卡锡（John McCarthy，斯坦福大学），马文·明斯基（麻省理工学院），特伦查德·摩尔（Trenchard More，IBM），雷·所罗门诺夫（Ray Solomonoff，伦敦大学）和奥利弗·赛弗里奇（Oliver Selfridge，麻省理工学院）。无论从科学还是社交的角度来说，这都是一次引人入胜的会议。

卡内基－梅隆大学的金出武雄（Takeo Kanade）在他的演讲“人工智能视野：进步与非进步”（Artificial Intelligence Vision: Progress and Non-Progress）中指出，以今天的标准来看，20世纪60年代的计算机内存很小，并且一次只能保存一张图像。金出武雄在他1974年的博士论文中指出，尽管他的程序可以在一个图像中找到一辆坦克，但是在其他图像中，如果坦克处于不同的位置并且光照不同，则很难得到同样的结果。但是，当他早期的学生毕业时，因为电脑更

强大，他们设计的程序可以在更普遍的条件下识别坦克。今天，他学生的程序可以识别任何图像中的坦克。不同之处在于，今天我们可以访问数以百万计的图像，可以对各种姿态和光照条件进行取样，而且计算机的功能更是强大了数百万倍。

麻省理工学院的罗德尼·布鲁克斯在他的题为“智能与身体”（Intelligence and Bodies）的演讲中，讲述了他在建造爬行和漫步机器人方面的经验。智能在大脑中进化以控制运动，并且身体逐渐发展为通过智能与世界互动。布鲁克斯放弃了机器人专家使用的传统控制器，并将行为，而非计算，作为设计机器人的参考。随着我们对如何搭建机器人的了解更加深入，会更清楚地认识到身体是意识的一部分。

在“为什么自然语言处理现在是统计自然语言处理”（Why Natural Language Processing is Now Statistical Natural Language Processing）的演讲中，欧仁·查尼阿克（Eugene Charniak）解释说，语法的基本作用是在句子中标记词类。人类接受这方面训练后，会比现有的解析程序做得更好。计算语言学领域最初试图应用诺姆·乔姆斯基在20世纪80年代首创的生成语法（generative grammar）方法，但结果令人失望。最终被证明可行的办法是，聘请布朗大学的学生为《华尔街日报》上千篇文章的词类进行手写标注，然后应用统计技术，根据相邻单词来确定特定单词的词类。这一过程需要大量的例子，因为大多数单词有多重含义，每个单词对应许多不同的上下文。目前，在句子中对词类进行自动标记，已经成为可以利用机器学习来解决的问题。

这些成功的故事有一个共同的轨迹。过去，电脑速度很慢，只能用少数几个参数来探索一个玩具模型。但是这些玩具模型无法适用于现实世界的数据。当拥有了大量数据，并且计算机速度也更快时，就可以创建更复杂的统计模型，并提取更多特征和特征之间的关系。深度学习使这一过程实现了自动化。深度学习可以从非常大的数据集中

提取特征，而不是让领域专家为每个应用程序手动创建这些特征。随着计算取代了劳动力并不断变得更廉价，计算机将能够执行更多劳动密集型的认知任务。

在会议结束时的总结性发言中，马文·明斯基一开始就表示，他对演讲和AI的研究方向感到失望。他解释道：“你们不是在解决通用的智能问题。你们只是在解决具体的应用问题。”这次会议应该是对我们所取得的进展的一次庆祝，但是却被他的指责打击了。我做的演讲是关于强化学习的进展，并展示了TD-Gammon通过训练，能在西洋双陆棋中表现出冠军水平的显著成果。不过我的演讲并没有给他留下深刻的印象，他始终认为，那只是一个游戏罢了。

明斯基的“通用的智能”意味着什么呢？他在其著作《心智社会》（*The Society of Mind*）[21]中提到，前提是通用的智能来自简单媒介之间的相互作用。明斯基曾经说过，有关他的理论的最大想法，是来源于试图创建一台使用机器人手臂、摄像机和计算机的机器，可以用儿童的积木构建一个结构（图2－1）。[22]这听起来很像一个应用。一个具体的应用会迫使你专注于并深入到最底层问题，而抽象理论化的方法则不能。演言人在达特茅斯会议上所报告的成就，来自对具体问题的深入了解，为更全面的理论理解铺平了道路。也许，一个更好的通用智能理论有一天会从这些狭隘的AI成就中浮现出来。

我们的大脑不只是凭空产生抽象的想法。它们与我们身体的所有部位密切相连，而我们的身体又通过我们的感官输入和运动效应器与外界密切相连。生物智能因此而具体化。更重要的是，我们的大脑在与外界互动的同时，经历了一个漫长的成熟过程。学习是一个与发育同时进行的过程，并且在我们达到成年后仍然会持续很长一段时间。因此，学习对于通用智能的发展至关重要。有趣的是，人工智能中最难解决的问题之一是常识，显然这是儿童并不具备的，只有在长期与外界接触后，大多数成人才会获得常识。人工智能中经常忽略的情绪

和通感也是智能的一个重要方面。[23]情绪是个整体信号，可以使大脑为不能由局部大脑状态决定的行动做好准备。

AI@50的最后一天举行了宴会。在晚餐结束时，1956年达特茅斯人工智能夏季研究项目的5名回归成员简要介绍了会议和人工智能的未来。在问答期间，我站起来，转向明斯基说道：“神经网络社区有一种看法：你是上世纪70年代需要为神经网络萧条负责的魔鬼。你是魔鬼吗？”明斯基发起了一场关于我们如何不理解我们网络的数学局限性的长篇大论。我打断了他：“明斯基博士，我问的是一个是或否的问题。你是，还是不是？”他犹豫了片刻，然后喊道：“是的，我是魔鬼！”

1958年，弗兰克·罗森布拉特制造了一个模拟计算机，被设计用来模拟感知器，因为当时数字计算机在模拟计算密集度高的网络模型时速度很慢。到了20世纪80年代，计算机的计算功能大大提升，让我们能够通过模拟小型网络来探索学习算法。但直到2010年，才有足够的计算机能力将网络规模扩大到可以解决实际问题的程度。

明斯基在1954年从普林斯顿大学获得数学博士学位，他的论文是关于用神经网络进行计算的理论和实验研究。他甚至用电子部件构建出了小型网络，以了解它的行为。当我还是普林斯顿大学物理系的研究生时，就听说数学系里没有人有资格评审他的论文[24]，所以人们把它发给了普林斯顿高等研究院的数学家们，据说他们能与上帝交谈。回复的评论是，“如果这在今天不是数学，未来总有一天会是。”这足以为明斯基赢得他的博士学位。而神经网络确实引发了一类新的数学函数，这些函数激发了新的研究，并正在成为数学的一个新分支。年轻的明斯基超越了他的时代。

从深度学习到通用人工智能

马文·明斯基于2016年去世，他坚信神经网络无法实现通用的人工智能。斯蒂芬·沃尔夫勒姆在一篇关于他与明斯基友谊的文章中关切地写道：“虽然我认为没有人能在当时就预见到，但我们现在知道，马文早在1951年就已经在研究的神经网络，在朝着他所希望的那种令人叹为观止的AI能力发展的正确道路上前进。可惜这个过程太久了，马文几乎没能看到它的实现。”[25]

明斯基去世后不久，DeepMind的研究人员艾历克斯·格雷夫斯（Alex Graves）和格雷格·韦恩（Greg Wayne）通过添加动态外部存储器，实现了基于深度学习的通用人工智能的下一步。[26]活动模式在深度循环神经网络中只能被暂时存储，这就很难对推理和推断进行模拟。通过在网络中添加一个稳定的存储器，可以像数字计算机内存一样灵活地写入和读取，研究人员展示了一个训练好的强化学习网络，可以回答需要推理的问题。例如，一个这样的网络对伦敦地铁路线做出了合理的规划，另一个则回答了关于家谱中成员关系的问题。动态记忆网络（Dynamic Memory Networks）也能够掌握20世纪60年代对麻省理工学院人工智能实验室提出挑战的“积木世界”任务（图2-1）。这就带我们回到了第2章开始的地方。

弗朗西斯·克里克于2004年去世，此后不久，莱斯利·奥格尔在2007年也去世了，这标志着索尔克研究所一个时代的终结。现在这些科学巨人已经不在我们身边了，而新一代的学者仍在向前迈进。我在索尔克研究所工作了30年，相当于它一半的寿命。我们的大家庭组建于1960年，所有的教职工都在同一条小船上。“索尔克号”很小，小到每个人都认识对方。但即使在今天，我们有了1000人的团队，它仍然会带给我们一种家的感觉，也证明了这个机构文化的持久性。

我们是一个伟大的生物链（可以回溯到细菌出现之前）中的一员。现在我们已经到了理解大脑以及它们是如何发展的边缘，这将会是一个奇迹，能够永远改变我们对自己的看法。

[1] T. A. Lincoln and G. F. Joyce, “Self-Sustained Replication of an RNA Enzyme,” *Science*323, no. 5918 (2009): 1229–1232.

[2] T. R. Cech, “The RNA Worlds in Context,” *Cold Spring Harbor Perspectives in Biology*4, no. 7 (2012), http://cshperspectives.cshlp.org/content/4/7/a006742.full. pdf+ html.

[3] J. A. Feldman, “Mysteries of Visual Experience” (2016; rev. 2017), https://arxiv.org/ftp/arxiv/papers/1604/1604.08612.pdf.

[4] 法定失明是一种视力丧失程度，在法律上被定义来确定可接受福利的资格。——译者注

[5] Patricia S. Churchland, V. S. Ramachandran, and Terrence J. Sejnowski, “A Critique of Pure Vision,” in Christof Koch and Joel D. Davis, eds., *Large-Scale NeuronalTheories of the Brain*(Cambridge, MA: MIT Press, 1994), 23–60.

[6] John Allman, *Evolving Brains*(New York: Scientific American Library, 1999).

[7] B. F. Skinner, *Beyond Freedom and Dignity*(Indianapolis: Hackett, 1971).

[8] Noam Chomsky, “The Case against B. F. Skinner,” *New York Review of Books*, 17,no. 11 (1971): 18–24. http://www.nybooks.com/articles/1971/12/30/the-caseagainst-bf-skinner/。

[9] 同上，第27段。有关语言基于规则和统计分析的更详细的对比，请参阅Peter Norvig, “On Chomsky and the Two Cultures of Statistical Learning” http://norvig.com/chomsky.html。

[10] Noam Chomsky, *Rules and Representations*(Oxford: Basil Blackwell, 1980).

[11] A. Gopnik, A. Meltzoff, and P. Kuhl, *The Scientist in the Crib*: *What EarlyLearning Tells Us about the Mind*(New York: William Morrow, 1999).

[12] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and Their Compositionality,” *Advances inNeural Information Processing Systems*26 (2013): 3111–3119.

[13] 这是在麻省理工学院麦戈文学院（McGovern Institute）的一次演讲，提议该学院致力于对理解有关语言和语言障碍的生物学基础的研究。

[14] “愿原力与你同在”，引自《星球大战》中绝地大师欧比旺·肯诺比的流行台词。

[15] J. A. Fodor, “The Mind/Body Problem,” *Scientific American*244, no. 1 (1981):114–123.

[16] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick, “Neuroscience Inspired Artificial Intelligence,” *Neuron*95, no. 2 (2017): 245–258.

[17] Paul W. Glimcher and Ernst Fehr, *Neuroeconomics: Decision Making and the Brain*, 2nd ed. (Boston: Academic Press, 2013).

[18] Colin Camerer, *Behavior Game Theory: Experiments in Strategic Interaction*(Princeton: Princeton University Press, 2003).

[19] Minsky and Papert，《感知器》 (1969), 231. 这本书于1988年发表的增订版中，包含这段节选的13.2节内容被删掉了，同时加入了一个新章节，“Epilogue:The new connectionism”。这是长达40页的对多层感知器学习早期成果的评估，从后续的发展情况来看，这本书值得一读。

[20] 参阅 https://www.dartmouth.edu/~ai50/homepage.html 。也可参阅 https://en.wikipe dia.org/wiki/AI@50/。

[21] Marvin Minsky, *The Society of Mind*(New York: Simon & Schuster, 1985).

[22] 参阅“Society of Mind,” *Wikipedia*, last edited August 22, 2017。https://en.wikipedia.org/wiki/Society_of_Mind。

[23] 麻省理工学院的辛西娅·布雷齐尔（Cynthia Breazeal）和哈维尔·莫维兰开发了能够与人类互动，并会使用面部表情进行交流的社交机器人，这很可能预示着向情感计算理论迈出的第一步。

[24] Marvin Minsky, “Theory of Neural-Analog Reinforcement Systems and Its Application to the Brain Model Problem” (Ph.D. diss., Princeton University,1954).

[25] Stephen Wolfram, “Farewell, Marvin Minsky (1927–2016),” *Stephen WolframBlog*, posted January 26, 2016, http://blog.stephenwolfram.com/2016/01/farewell-marvin-minsky-19272016/.

[26] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwiska,et al., “Hybrid Computing Using a Neural Network with Dynamic External Memory,” *Nature*538, no. 7626 (2016): 471–476.

18
深度智能

遗传密码

西德尼·布伦纳（Sydney Brenner）在南非出生并接受教育，随后参与了剑桥大学早期的分子遗传学研究工作（图18－1）。他曾与弗朗西斯·克里克在分子生物学实验室（LMB）共用一个办公室。如果是你，在发现DNA的结构并解开了遗传密码之后，会在下一个项目里做什么？克里克决定把重点放在人类的大脑上。布伦纳开始研究一种新模式生物——线虫（C. elegans），这是一种生活在土壤中的蛔虫，有1毫米长，只有302个神经元。长期追踪体内的每个细胞，了解个体如何从一个胚胎逐渐发育成熟的研究，使得这些线虫成了许多突破性发现的源头。布伦纳因为这项研究，在2002年和罗伯特·霍维茨（H. Robert Horvitz）、约翰·苏尔斯顿（John E. Sulston），共同分享了诺贝尔生理学或医学奖。布伦纳也因他的智慧而闻名。他在诺贝尔获奖者演讲中，赞扬了线虫：“我今天的演讲题目是‘大自然对科学的恩赐’（Nature's gift to Science）。内容并不是关于一个科学期刊对另一个期刊的致敬，而是关于生命世界的多样性如何启发和服务于生物研究领域的创新。”[1]看上去西德尼·布伦纳似乎见证了“创世记”。



图18-1 西德尼·布伦纳是生物学界的传奇人物。他研究遗传密码，DNA中碱基对被转录成蛋白质的方式，并因其对模式生物的开创性研究工作而获得诺贝尔奖。这张照片来自2010年《科学网络》对他的采访，http://thesciencenetwork.org/programs/the-science-studio/sydney-brenner-part-1。

布伦纳2009年在索尔克研究所做过三场系列讲座，标题是“阅读人类基因组”（Reading the Human Genome）[2]，讲座很精彩，没

有借助任何幻灯片或道具。他注意到，除了电脑，还从没有人曾经阅读过整个人类基因组，一个个地检查碱基对，于是便将其作为自己的目标。在这个过程中，他发现了不同的基因和物种的DNA序列之间有趣的相似之处。

布伦纳在全球多个研究机构都有教职。他在新加坡有一个实验项目；也是日本冲绳理工学院的创始校长；弗吉尼亚州阿什本（Ashburn）附近霍华德·休斯医学研究所珍妮莉亚研究园区的高级研究员；还是我在拉荷亚索尔克研究所领导的克里克–雅各布理论与计算生物学中心（Crick-Jacobs Center for Theoretical and Computational Biology）的高级研究员。以上只是其中的几个头衔。在戴维·马尔获得了他的博士学位后，布伦纳聘请马尔在LMB从事计算研究工作，后来布伦纳通过自己的南非朋友西摩尔·帕普特，在麻省理工学院的AI实验室为马尔安排了一个职位。分子遗传学和神经生理学之间的关系很深，而布伦纳在这两个领域都处于中心位置。

布伦纳有一次去拉荷亚的时候，我和他吃了顿晚饭，其间我告诉他，多年前我在哈佛医学院做博士后的时候听过一个故事——弗朗西斯·克里克死后到了天堂。圣彼得看到这个坚定的无神论者感到很惊讶，但弗朗西斯去天堂是想问上帝一个问题。他被带到了田野中的一个木棚里，里面散落着各种轮子和齿轮，都是失败的试验品。弗朗西斯看到上帝围着皮质围裙站在工作台边，正在修补一个新的有机体。“弗朗西斯，”上帝说，“见到你真是太高兴了。我可以为你做些什么？”“在我的一生中，”弗朗西斯说，“我都想知道这个问题的答案：为什么苍蝇有成虫盘？”[3]“亲爱的弗朗西斯，”上帝回答道，“这真是个惊喜！从来没有人问过我这个问题。我已经将成虫盘放入苍蝇体内数亿年了，还从来没有听到过任何抱怨。”

听完这个故事，布伦纳没说话。我猜想，讲这么一个贬损他的亲密朋友的故事，或许有些冒犯。“特里，”他说，“我可以告诉你我第一次碰到这个故事的那个场景。弗朗西斯和我坐在办公室里，他当

时正在读一本关于发育生物学的书，突然举起手说：‘上帝才知道苍蝇为什么会有成虫盘！’”

我惊呆了。找到你几十年前就听过，并复述了无数次的故事的源头的机会，有多少呢？我让布伦纳告诉我最初的版本。他说故事的标题是“天堂里的弗朗西斯·克里克”；他的故事版本与我的版本有着相同的基本结构，但细节却不尽相同[4]——就像进化过程一样，故事的基本核心不变，但许多细节却改变了。

我于2017年1月探望了在新加坡的布伦纳，庆祝他的90岁生日。他因为健康问题不再旅行，还需要坐在轮椅上，但仍和我之前每次见到他时那样有活力。西奥多西厄斯·多布赞斯基（Theodosius Dobzhansky）曾经说过，没有进化论，生物学中的任何东西都讲不通。[5]布伦纳于2017年2月21日在新加坡南洋理工大学举行了一场关于细菌进化的演讲，该演讲是系列讲座《10个10：进化编年史》（10-on-10：The Chronicle of Evolution）[6]。[7]2017年7月14日，我也在这个系列讲座中做了关于大脑进化的主题演讲。在演讲的开头，我对多布赞斯基的那句话做了些许改动：没有DNA，生物学中的任何东西都讲不通。[8]

每个物种都有智能

智能在许多物种中发展起来，以解决它们在环境中面对的生存问题。在海洋中进化的动物与在陆地上进化的动物相比，面临着许多不同的问题。视觉感知使我们能够感知周围的世界，我们还发展出了视觉智能来解释这些视觉信号。研究非人类动物的自然行为的动物行为学家们，已经在这些动物身上发现了人类并不具备的能力和技巧，如回声定位：蝙蝠主动发出听觉信号来探测它们所处的环境，并对回声进行分析。这创造了外部世界的内部表征，所有表象都与我们的视觉

体验一样生动。蝙蝠具有听觉智能，它能分辨来自鼓翼的昆虫（可以捕捉）和来自障碍（需要避免）的信号。

纽约大学的哲学家托马斯·内格尔（Thomas Nagel）在1974年写了一篇题为 《成为一只蝙蝠是什么感觉？》（What Is It Like to Be a Bat?）的文章，并得出结论，如果没有回声定位的直接经验，我们就无法想象蝙蝠世界是什么样的。[9]但缺乏这种经验并没有妨碍我们发明雷达和声呐技术（这些技术使我们能够积极探索看不到的世界），也不会阻止盲人通过适应声音的反射来摸索周边的环境。我们也许不知道成为一只蝙蝠是什么感觉，但我们可以发展类蝙蝠的智能，协助自动驾驶汽车使用雷达和激光雷达行进。

我们人类是大自然中学习方面的佼佼者。我们可以在更广泛的领域中学得更快，记住得更多，通过不断繁衍所积累的知识更是胜过任何其他物种。我们创造了一种名为“教育”的技术，以增加我们在有生之年能够学到的东西。现在，儿童和青少年在教室里度过了自己的成长期，不断学习着世界上他们从未直接体验过的事物。作为相对近期的人类发明，阅读和写作需要很多年才能掌握。但是，这些发明使得积累的知识能够传递给下一代，比如写书、印制再到阅读这个流程，仅靠口口相传是积累不了那么多知识的。正是写作、阅读和学习，而非说话，使现代文明成为可能。

进化的起源

我们的进化起源是什么？我在1998年协助建立的拉荷亚人类起源小组，最初是作为一个小组定期举行会议，讨论可能帮助我们回答这个问题的许多证据，这些证据来自古生物学、地球物理学、人类学、生物化学、遗传学，还有比较神经科学。它逐渐吸引了很多国际会员，并在2008年成为加州大学圣迭戈分校/索尔克人类学学术研究和培训中心（UCSD/Salk Center for Academic Research and Training

in Anthropogeny，简称CARTA）。[10]正如NIPS召集所有科学和工程团队来了解神经计算一样，在探索我们人类来自哪里，如何走到现在，培养新一代思考者来寻找这些古老问题的答案方面，CARTA也吸取了来自所有科学领域的见解。[11]

600万年前，人属的谱系从黑猩猩的血统中分离出来（图18-2）。黑猩猩是一种非常聪明的物种，但黑猩猩的智能与我们的完全不同。在试图教授黑猩猩语言基础知识的尝试中我们发现，它们仅能学会几百个符号，用它来表达简单的需求，尽管以这种方式衡量它们的智能显得有些不太公平。如果我们不得不在黑猩猩的队伍中生存下去，我们又会有多好的表现呢？所有的物种都和我们一样以自我为中心吗？



图18-2 黑猩猩的大脑和体积更大的人脑之间的比较，人脑已经进化出显著增大的，有更多褶皱的大脑皮层。图片来源：Allman, Evolving Brains, p. 164。

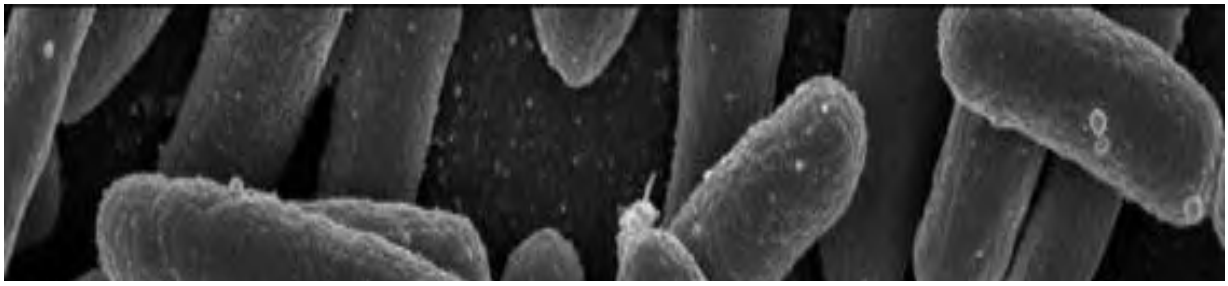
在人类和黑猩猩之间存在的差异，可以在我们的DNA中找到。我们知道这一点已经有一段时间了，人类的30亿个DNA碱基对中，与黑猩猩不同的只占1.4%。当黑猩猩的基因组被首次测序后，我们认为自己将能够阅读生命之书，发现人类与黑猩猩有什么不同。很遗憾，此书大约有90% 的内容我们都没有能力读懂。[12]我们的大脑与黑猩猩的大脑非常相似；神经解剖学家已经在两个物种中确定了相同的脑区。与人类和黑猩猩在行为中的显著差异相比，这两个物种大脑之间的大部

分差异处于分子水平，并且十分微妙。这再一次证明了，大自然比我们聪明。

人类终将解决智能难题

莱斯利告诉我，奥格尔的第一法则指出，细胞中的每一个基本反应都会演化出一种酶，来催化这种反应。这种酶不仅加速了反应，而且还可以通过与其他分子的相互作用来调节反应，从而使细胞更加高效，适应性更强。自然以一个巧妙的反应路径开始，通过添加酶和备份路径逐渐改善这个路径。然而所有这些在缺失核心过程的情况下是无法工作的，这个核心过程对于细胞来说就是DNA的维护和复制，它在细胞生物化学领域中扮演着“蜂王”的角色。

单细胞生物已经适应了许多不同的环境，并在生态中演化出它们自己的位置。例如，细菌已经适应了极端的环境——从海底的高温液体喷口到南极洲的冰层，以及更多温和的环境，比如我们的胃肠中有数千种细菌。像大肠杆菌这样的细菌（图18－3）已经发展出了按照梯度游向食物来源的算法。由于细菌太小，在几微米的体长里不能直接检测到梯度，于是就使用了趋化性，这涉及周期性地翻滚并随机游动。[13]这看起来可能会适得其反，但通过在高浓度的地方将游动时间延长，细菌可以准确地按浓度梯度爬行。它们的智能是一种原始智能，但细菌比最聪明的生物学家还要聪明。这群最聪明的生物学家尚未弄清楚细菌是如何在如此多样的环境中生存的。在多细胞动物中可以发现更复杂的智能形式。



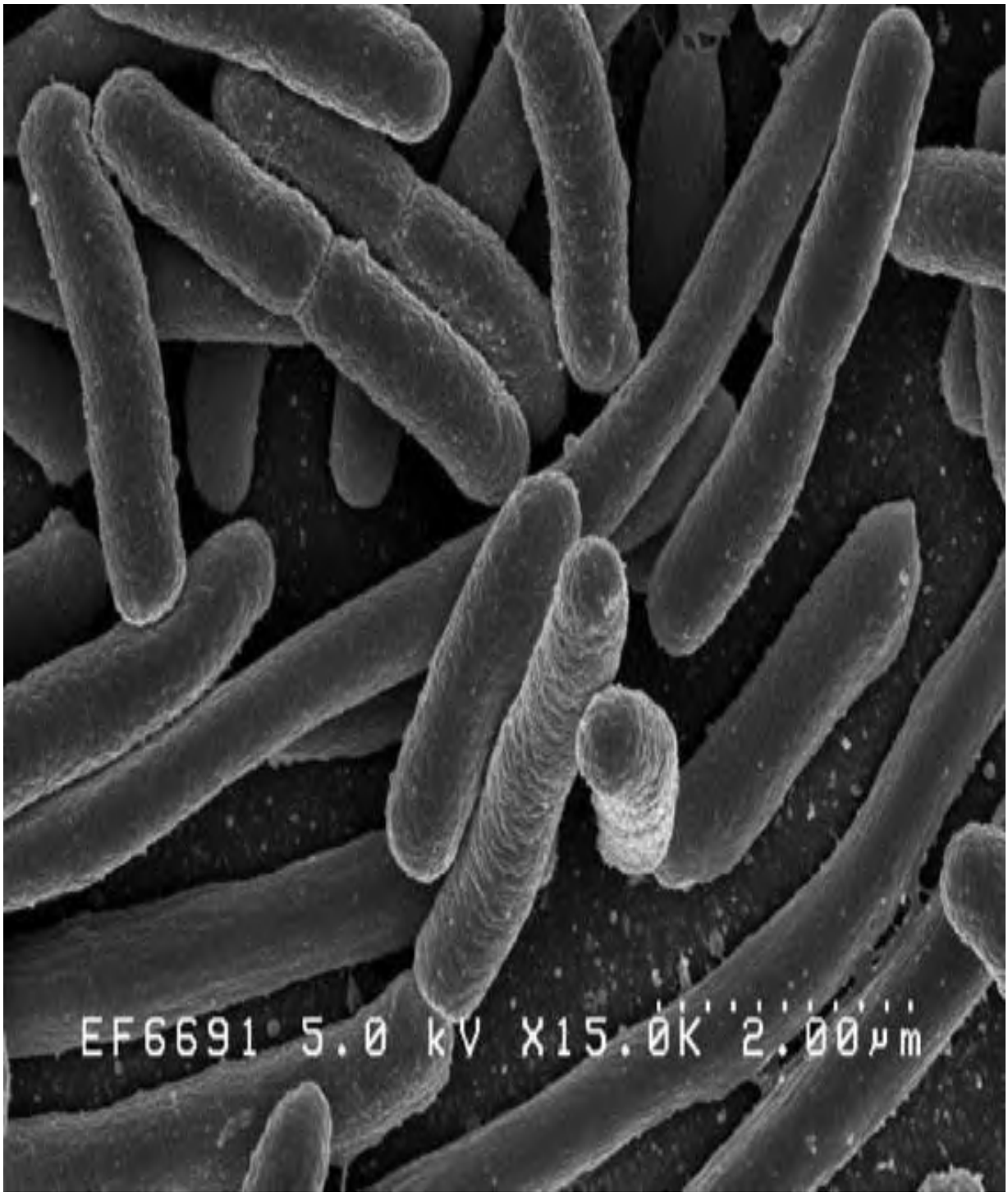


图18-3 大肠杆菌的扫描电子显微照片。细菌是地球上最多样、最健壮、最成功的生命形式。通过研究它们，我们可以学到很多关于自主智能的知识。图片来源：NIAID，NIH。

我们已经看到，强化学习的时间差分学习算法可能导致高度复杂的行为，通过人类大脑皮层的深度学习，强化了人类的复杂性。自然界有一系列的智能行为，可以让人工系统从中学习。跨越计算机科学和生物学的算法生物学是一个新的科学领域，寻求使用算法的语言来描述生物系统所使用的问题解决策略。[14]我们希望，确定这些生物算法将会启迪新的工程计算范式，并能让我们对生物网络有系统级的理解。以上都不是重点，最终的目标是解释跨空间和时间尺度的生物系统中嵌套的复杂性：基因网络、代谢网络、免疫网络、神经网络和社交网络——全部都是网络。

深度学习取决于对一个成本函数的优化。自然界的成本函数是什么呢？进化成本的倒数被称为适应度，但这是一个概念，只在具体的约束条件下才有意义，无论约束条件是来自环境，还是来自要被优化的系统。在大脑中，有一些调节行为的固有成本，例如对食物、温度、安全、氧气和生育的需要。在强化学习中，需要采取行动来优化未来的回报。但除了保证生存的奖励之外，从人类令人眼花缭乱的行为中可以看出，还有各种各样其他的奖励可以被优化。是不是一些潜在的通用成本函数导致了这种多样性呢？

我们仍在寻找暴露智能最高形态秘密的核心概念。我们已经确定了一些关键原则，但是却没有一个概念框架能像DNA解释生命本质那样，优雅地解释大脑如何运转。学习算法是寻找统一概念的好地方。也许，我们在理解深度学习网络如何解决实际问题方面取得的进展，将引出更多线索。我们可能会发现细胞和大脑中使进化成为可能的操作系统。如果我们能解决这些问题，就可能会有想象不到的收获。自然可能比我们每一个人都更聪明，但作为一个物种，我并不认为人类无法解决智能难题。

---

[1] Sydney Brenner, “Nature’s Gift to Science,” Nobel Lecture, December 8, 2002,video, https://www.nobelprize.org/mediaplayer/index.php?id=523/.

[2] Sydney Brenner, “Reading the Human Genome”: 1. “Much Ado about Nothing:Systems Biology and Inverse Problems,” January 26, 2009; 2. “Measure for Measure: The GC Shift and the Problem of Isochores,” January 29, 2009; 3. “All’s Well That Ends Well: The History of the Retina,” January 30, 2009, videos. http://thesciencenetwork.org/search?program=Reading+the+Human+Genome+with+Sydn ey+Brenner.

[3] 成虫盘是苍蝇腿和触须的发育原基。

[4] 西德尼·布伦纳的原创故事发表于S. Brenner “Francisco Crick in Paradiso”，Current Biology.6, no.9 (1996):1202：“我在剑桥与弗朗西斯·克里克共事了二十年。有一段时间，他对胚胎学很感兴趣，花了很多时间思考果蝇的成虫盘。有一天，他突然把手上的书猛地摔到桌上，火冒三丈地吼道：‘只有上帝才知道这些成虫盘是怎么工作的！’ 一瞬间，我仿佛看到了弗朗西斯升入天堂后发生的所有事情。门徒彼得欢迎他说：‘哦，克里克博士，你经过漫长的旅程后一定很累。先坐下来，喝一杯，放松一下吧。’‘不，’弗朗西斯说，‘我必须见到上帝，我要问他一个问题。’ 经过一番劝说，天使同意将弗朗西斯带到上帝面前。他们穿过了天堂的中部，到达了位于后方的铁轨处，穿过铁轨，他们来到一个有着瓦楞铁屋顶的棚屋，周围都是垃圾。在屋子后面，有一个穿着工作服的小个子，他的后兜里揣着一个大扳手。‘上帝，’ 天使说，‘这位是克里克博士；克里克博士，这位就是上帝。’‘很高兴见到您，’弗朗西斯说：‘我必须问您这个问题。成虫盘到底是如何工作的？’‘这个嘛，’ 上帝答道，‘我们用了一点点这个东西，又添加了一些东西，然后……其实，我们也不清楚，但我可以告诉你，我们已经在这里创造了2亿年的苍蝇，还从来没收到过任何抱怨呢。’”

[5] T. Dobzhansky, “Nothing in Biology Makes Sense Except in the Light of Evolution,” *American Biology Teacher*35, no. 3 (1973): 125–129, http://biologielernprogramme.de/daten/programme/js/homologer/daten/lit/Dobzhansky.pdf.

[6] 在10个月里进行的10个讲座，涵盖10个对数时间刻度的话题。——译者注

[7] Sydney Brenner, “Why We Need to Talk about Evolution,” in “10-on-10: The Chronicle of Evolution” lecture series, Nanyang Technological University,Singapore, February 21, 2017. http://www.paralimes.ntu.edu.sg/NewsnEvents/10on-10%20The%20Chronicle%20of%20Evolution/Pages/Home.aspx. Video. https://www.youtube.com/watch?v=C9M5h_tVlc8.

[8] Terrence Sejnowski, “Evolving Brains,” in “10-on-10: The Chronicle of Evolution”lecture series, Nanyang Technological University, Singapore, July 14, 2017. https://www.youtube.com/watch?v=L9ITpz4OeOo.

[9] T. Nagel, “What Is It Like to Be a Bat?” *Philosophical Review*83, no. 4 (1974):435– 450.

[10] “人类起源论”（Anthropogeny）是研究人类起源的理论。参阅Center of Academic Research and Training in Anthropology (CARTA) website, https://carta.anthropogeny.org/.

[11] 拉荷亚小组和CARTA是在加州大学圣迭戈分校医师科学家阿吉特·瓦尔基（Ajit Varki）的领导下成立的，热衷于对进化理论的研究。

[12] 在DNA中，约有1% 是编码蛋白质的序列，8%是结合蛋白质的调控序列。

[13] Howard C. Berg, *E. coli in Motion*(New York: Springer, 2004).

[14] S. Navlakha and Z. Bar-Joseph, “Algorithms in Nature: The Convergence of Systems Biology and Computational Thinking,” *Molecular Systems Biology*7(2011): 546.

附录一

致谢

我工作的索尔克生物研究所是一个与众不同的地方。从外面看，它就像一个混凝土堡垒，但是当你进入中央庭院时，会看到一大片石灰岩延伸到太平洋，两侧高耸的塔架让这美轮美奂的景色定格在你的眼前（图1）。[1]我的实验室在南楼，位于庭院外（照片左侧）。你在入口处左边的墙上会看到一张电子显微镜下的海马体的照片，它看起来像一盘意大利面的横截面；入口通向茶室，那里是计算机神经生物学实验室的心脏。

世界上最杰出的一些科学家，包括弗朗西斯·克里克，很喜欢与学生和同事围着圆形的白色茶桌讨论各种科学问题（图2）。事实上，茶室还出现在了克里克的著作《惊人的假说》（*The Astonishing Hypothesis*）中：

> 特里·谢诺夫斯基在索尔克研究所的团队，在本周的大多数午餐后都会举行非正式的茶会。这些茶会是讨论最新实验结果，抛出新的想法，或只是聊聊科学、政治或新闻的理想场合。有一天我去喝茶，并向帕特·丘奇兰德和特里·谢诺夫斯基宣布，我发现意愿诞生的位置了！它位于或靠近前扣带回。与安东尼奥·达马西奥讨论此事时，我发现他也有同样的想法。[2]



图1在加州拉荷亚市的索尔克生物研究所，可以俯瞰太平洋。这座由路易斯·卡恩设计的标志性建筑是一座科学的圣殿，也是我每天上班的地方。图片来源：Kent Schnoeker，索尔克生物研究所。

我尤其记得，克里克在1989年与比阿特丽斯·哥伦布一起喝茶的那天。他告诉我，比阿特丽斯想从事神经网络的研究工作，我应该雇用她。[3]比阿特丽斯是加州大学圣迭戈分校的一名医学博士生，读研究生的时候曾与克里克一起短暂工作过一段时间。她曾想把神经网络作为她的博士论文项目，但这违反了生物系的规定。我接受了克里

克的建议，我们彼此都从对方那里学到了很多。自我们于1990年在加州理工学院的Athenaeum宴会厅举行婚礼以来，我一直在向她学习。

从我在约翰·霍普金斯大学任教开始，茶桌就一直陪伴着我。我于1981年在托马斯·詹金斯生物物理系从事的第一份工作中，为我的新实验室购买的第一件家具就是茶桌。该科系就像一个历史悠久的大家庭，而我是年轻的宠儿，他们让我有信心在新的方向发展，对此我永远心存感激。作为哈佛医学院神经生物学系的博士后，我保留了下午茶传统。在一个庞大而多样的科系里，这是一种让大家保持联系并了解实验进展的方法。我在索尔克研究所的实验室是一个微型大学，拥有来自科学、数学、工程和医学等不同专业背景的学生，下午茶时间是我们大集体团聚的时间。



图2 2010年索尔克研究所计算神经生物学实验室（CNL）的茶室。日常茶会一直是本书描述的许多学习算法和科学发现的社交孵化器。图片来源：索尔克生物研究所。

我很幸运。我的父母很重视教育，从小就十分信任我；我生活在前所未有的经济增长和遍地机会的时代，这开阔了我的视野；我的导师和合作者慷慨地分享了他们的见解和建议；我有幸与一代极其优秀的学生一起工作。我特别感谢杰弗里·辛顿、约翰·霍普菲尔德、布鲁斯·奈特（Bruce Knight）、斯蒂芬·库夫勒、迈克尔·斯蒂麦克（Michael Stimac）和约翰·惠勒（John Wheeler），同样也感谢我的岳父所罗门·哥伦布，他们在我职业生涯的各个转折点帮助我选择了正确的方向。比阿特丽斯具有批判性思维，我从她那里学到了如何避免群体思维。仅仅因为每个人都相信某种解释，并不能证明它就是真理。有时需要花费整整一代人的时间，才能从群体中剔除一个普遍持有的信念。

同时，我还要感谢在本书的写作过程中帮助过我的其他人。与长期合作者帕特里夏·丘奇兰德和“科学网络”的创始人罗杰·宾汉姆（Roger Bingham）在网上进行的探讨，为本书的写作提供了灵感。约翰·多伊尔（John Doyle）对控制理论的洞见启发了我对大脑操作系统的讨论。与凯里·施塔勒（Cary Staller）在瑞士克洛斯特斯和达沃斯附近山上的徒步旅行，帮我理清了对算法空间的理解。芭芭拉·奥克利教会我如何接触到课堂以外的受众。凯里和芭芭拉都帮助我塑造了讲述深度学习故事的方式。还有很多人为本书提供了宝贵的反馈和建议，他们包括：Yoshua Bengio，Sydney Brenner，Andrea Chiba，Gary Cottrell，Rodney Douglas，Paul Ekman，Michaela Ennis，Jerome Feldman，Adam Gazzaley，Geoffrey Hinton，Jonathan C. Howard，Irwin Jacobs，Scott Kirkpatrick，Mark和Jack Knickrehm，Te-Won Lee，James McClelland，Saket Navlakha，Barbara Oakley，Tomaso Poggio，Charles Rosenberg，David Silver，James Simons，Marian Stewart-Bartlett，Richard Sutton，Paula Tallal，Gerald Tesauro，Sebastian Thrun，Ajit Varki，Massimo Vergassola，Stephen Wolfram（他也为本书的书名提供了建议），以及Steven Zucker。

自1984年以来，伍兹霍尔计算神经科学研讨会每年夏天都会举行一次，其中有一小部分核心成员和新的参与者在早上和晚上会进行深入讨论，下午是留给大家做户外活动的自由时间——多么完美的安排。这个研讨会的成员都拥有了杰出的职业生涯。伍兹霍尔研讨会一直持续到今天，但于1999年转移到特柳赖德，和年度神经形态工程研讨会一起举办。我感谢过去三十年来参加这些研讨会的所有人，特别是John Allman，Dana Ballard，Robert Desimone，John Doyle，Katalin Gothard，Christof Koch，John Maunsell，William Newsome，Barry Richmond，Michael Stryker，和Steven Zucker。

我在索尔克生物研究所和加州大学圣迭戈分校的同事们，都是杰出的有创业和合作精神的研究人员，他们正在创造生物医学科学的未来。加州大学圣迭戈分校神经计算研究所的教师和学生，在我1990年创所以来，以我从未想象过的方式整合了神经科学和计算。

索尔克研究所的计算神经生物学实验室在过去的三十年里一直是我的家，我的许多学术弟子已经离开，在全世界继续着蓬勃发展的职业生涯。一个实验室就像一个家庭，几代热情的研究生和博士后大大丰富了我的生活。我的实验室管理员罗丝玛丽·米勒（Rosemary Miller）和玛丽·艾伦·佩里（Mary Ellen Perry）精心地照顾着实验室。玛丽·艾伦在过去的十年中一直担任NIPS的董事总经理，李·坎贝尔（Lee Campbell）开发了一个计算机平台，使我们能够将会议规模扩大十倍。

我很感谢麻省理工学院出版社，四十年来它一直是我可靠的合作伙伴，出版了我和托马索·波吉奥编辑的一套关于计算神经科学的丛书；我于1989年创立的期刊《神经计算》（*Neural Computation*）；我于1992年出版的书《计算大脑》（*The Computational Brain*）；以及其他人写的一些关于机器学习的基础书籍，比如理查德·萨顿和安德鲁·巴托的《强化学习导论》（*Reinforcement Learning: An Introduction*），伊恩·古德费洛、约书亚·本吉奥和亚伦·库维尔

（Aaron Courville）联合撰写的该领域最重要的教科书《深度学习》。麻省理工学院出版社的罗伯特·普赖尔（Robert Prior）在本书漫长的出版过程中，也对本书的写作提供了指导。

我还要感谢NIPS社区，没有他们，我就无法写出这本《深度学习》，本书远没有涵盖这个领域的全面历史，它只关注了几个主题和参与神经网络研究的一些人。国际神经网络协会的期刊《神经网络》一直是扩大神经网络研究范围的坚强后盾。该协会与IEEE合作，每年举办一次神经网络国际联合会议。机器学习也催生了许多优秀的会议，包括国际机器学习会议（ICML），这是NIPS的姊妹会议。这个领域极大地受惠于来自所有这些组织，以及对其做出贡献的研究人员。

在2018年长滩NIPS的开幕式上，我对NIPS的发展表达了惊叹：“我在30年前第一次参加NIPS会议时从来没有想到，今天会站在这里，面对着8000名与会者发表演讲——我当时以为这个会议只能持续10年。”我于2016年4月访问了现居加州山景城（Mountain View）的杰弗里·辛顿。“谷歌大脑”（Google Brain）团队占据了一栋大楼的一整个楼层。我和杰弗里一起回顾过去的日子，得出了我们已经赢得胜利的结论，但花费的时间比我们预期的要长得多。在这个过程中，杰弗里入选了英格兰和加拿大皇家学会。我也被选为美国国家科学院、国家医学院、国家工程院、国家发明家科学院和国家艺术与科学学院院士，这些都是难得的荣誉。我非常感谢杰弗里·辛顿多年来和我分享他对网络计算的见解。

作为普林斯顿大学的研究生，我曾经致力于研究爱因斯坦的引力理论，广义相对论中的黑洞和引力波。然而，在获得我的物理学博士学位之后，我的研究领域转向了神经生物学，大脑自那时起就引起了我的兴趣。我还不知道我的第三次转变可能是什么。所罗门·哥伦布曾经告诉我，所谓事业，只能用来追溯，我在写这本书时也确认了这一点。回顾过去，我发现了那些让我走到今天这一步的事件和决定，当然，当时的我对这一切还一无所知。

[1] 参阅Sarah Williams Goldhagen, *Louis Kahn's Situated Modernism*(New Haven:Yale University Press, 2001)。

[2] Francis Crick, *The Astonishing Hypothesis: The Scientific Search for the Soul*. (New York: Scribner's Sons, 1994), 267.

[3] 在我和比阿特丽斯相识这件事上，克里克同样功不可没。

附录二

词汇表

自适应信号处理（adaptive signal processing）：提高信号质量的方法，例如自动粒度控制，或可自动降低噪声的可调节滤波器。

算法（algorithm）：一个列出了具体操作步骤的说明，你可以按步骤来实现一个目标，就像对照菜谱烘焙蛋糕一样。

反传（或误差反向传播）［backprop（backpropagation of errors）］：通过梯度下降来优化神经网络的学习算法，以最小化代价函数的值并提高网络性能。

贝叶斯规则（Bayes’s rule）：基于新数据和与事件相关条件的先验知识，对事件概率进行更新的规则。更一般地说，贝叶斯概率是对当前和先验数据产生结果的信念。

玻尔兹曼机（Boltzmann machine）：一种神经网络模型，由相互作用的二进制单元组成，其中单元处于活跃状态的概率取决于其整合的突触输入。该模型以19世纪物理学家路德维希·玻尔兹曼命名，他是统计力学的奠基人。

约束条件（constraints）：让最优化问题的解决方案必须满足的保证其值为正的条件。

卷积（convolution）：通过计算一个函数与另一个函数相对移位的重叠量，将两个函数进行混合。

代价函数（cost function）：用来指定网络目的并衡量其性能表现的函数。学习算法的目的是降低代价函数的值。

数字助理（digital assistant）：用以辅助完成任务的虚拟助理，比如亚马逊智能音箱Echo中的语音助理Alexa。

周期（epoch）：学习过程中，从指定数量的样本中计算出平均梯度后神经网络中权重的更新。

平衡（equilibrium）：热力学状态，其中没有物质或能量的净宏观流动。在玻尔兹曼机中，神经单元状态是概率性的，当输入保持不变时，系统会最终稳定在平衡状态。

反馈（feedback）：神经网络中由高/后层流向低/前层神经元的连接，这种连接使信号在网络中形成了环流。

前馈网络（feedforward network）：一种多层神经元网络，其中各层神经元之间的连接是单向的，从输入层开始，到输出层结束。

梯度下降（gradient descent）：一种以降低代价函数为目的的优化方法，在每个周期中都会对参数进行调整，代价函数用来衡量网络模型的性能。

霍普菲尔德网络（Hopfield net）：由约翰·霍普菲尔德发明的全连接神经网络模型，根据不同的输入状态，该模型都能保证收敛成固定的吸引子形态，可以用来存储和提取信息。对该网络进行过探讨的文章有上千篇。

学习算法（learning algorithm）：基于样本来调整函数参数的算法。同时提供了输入和目标输出样本的是监督算法，仅提供输入样本的是无监督算法。强化学习是监督学习算法的一种，仅对好的表现进行奖励。

逻辑（logic）：基于结果仅为真或假的假设的数学推断。数学家用逻辑来证明定理。

机器学习（machine learning）：计算机科学的一个分支学科，让计算机从数据中学习如何完成任务，而不对其提供明确的程序设定。

毫秒（millisecond）：一秒的千分之一（0.001秒），即一千赫兹音调的一个周期。

大规模开放式在线课程（massive open online course，MOOC）：互联网上各色主题的免费课程。第一门MOOC诞生于2006年，到2018年1月，网上大约有9400门不同的MOOC，拥有约9100万学习者。

神经元（neuron）：一种特异化的大脑细胞，整合其他神经元的输出作为输入，并将输出信号传递给其他神经元。

归一化（normalization）：将信号的振幅稳定在固定范围内。一种将随时间变化的正值信号进行标准化的方式，即用该信号除以其自身的最大值，使归一化的信号被限定在0和1之间。

优化（optimization）：从提供的输入集合中系统地选取输入值的过程，以找到函数的最大或最小输出值。

过度拟合（overfitting）：当一个网络模型中可调整参数的数量远大于训练数据的数量，并且该算法使用了过多参数来匹配这些样本时，经过训练的学习算法所达到的状态。虽然过度拟合大大降低了网络适应新样本的能力，但可以通过正则化来降低这种负面影响。

感知器（perceptron）：一个简单的神经网络模型，由一个神经元和一系列带有可变权重的输入组成，可以通过训练来对输入进行分类。

可塑性（plasticity）：神经元对其功能的改变，如连接强度的改变（“突触可塑性”），或对其输入的响应［“本征可塑性”（intrinsic plasticity）］。

概率分布（probability distribution）：指定了实验中系统或结果的所有可能状态发生的概率的函数。

循环网络（recurrent network）：拥有反馈连接，允许信号在其中循环流动的神经网络。

正则化（regularization）：在训练数据有限的情况下，防止具有许多参数的网络模型发生过度拟合的方法。例如权重衰减，即网络中所有的权值在训练的每个周期都会减小，只有具有较大正梯度的权值被保留。

缩放性（scaling）：对算法的复杂程度如何随着问题的大小而变化的判断标准。例如，n个数相加的复杂度缩放性为n，但是n个数所有两数配对相乘的复杂度缩放性为n^2。

臭鼬工厂（skunk works）：一个小组在某个组织内以高度自治的方式从事高级别或秘密项目。借鉴了漫画作品《丛林小子》（*Li'l Abner*）中月光工厂的名称。

稀疏原理（sparsity principle）：对信号的稀疏表征，如脑电图（EEG）和功能磁共振成像（fMRI），用少量固定基函数的加权和来近似信号。这些基函数在独立分量分析（independent component analysis）中被称为信源。在一组神经元中，对一个输入的稀疏表征是指只有少数神经元高度活跃的情况。这可以减少来自表示其他输入的活动模式的干扰。

树突棘（spine）：神经元树突上的刺棘状膜突，充当突触的突触后部位。

突触（synapse）：两个神经元之间的特异化连接部位，信号从突触前神经元传递到突触后神经元。

训练和测试集（training and test sets）：基于训练集的表现并不能很好地评估神经网络在新输入上执行的效果，但训练期间未使

用的测试集则可以对网络性能进行评估。当数据集合很小时，可以使用训练集之外的单个样本，对使用其余样本进行训练的网络进行性能测试，并且对每个样本重复该过程以获得平均测试性能。这是n=1时的交叉验证的特例，其中n个子样本被保留作为测试集而非训练集。

图灵机（Turing machine）：由阿兰·图灵于1937年发明的一种假想的，可以用来做数学计算的简单计算机模型。图灵机包括一个可以前后移动的“磁带”，磁带被划分为一个接一个的单元，一个具有“状态”的“磁头”，可以改变其下活动单元的属性，以及一套关于磁头如何修改活动单元格并移动磁带的指令。在每个步骤中，机器可以修改活动单元格的属性并更改磁头的状态，之后再将磁带移动一个单元。

图书在版编目（CIP）数据

深度学习 /（美）特伦斯·谢诺夫斯基著；姜悦兵译. -- 北京：中信出版社，2019 . 2
书名原文：The Deep Learning Revolution
ISBN 978-7- 5086-9835-9
I. ①深… II. ①特… ②姜… III. ①机器学习IV. ① TP181
中国版本图书馆CIP数据核字（2018）第271842号

The Deep Learning Revolution by Terrence J. Sejnowski
Copyright © 2018 Massachusetts Institute of Technology
Simplified Chinese translation copyright © 2019 by CITIC Press Corporation
ALL RIGHTS RESERVED

深度学习
著者：［美］特伦斯·谢诺夫斯基
译者：姜悦兵
出版发行：中信出版集团股份有限公司
（北京市朝阳区惠新东街甲4号富盛大厦2座 邮编100029）

字数：350千字
版次：2019年2月第1版
京权图字：01- 2018-5970
广告经营许可证：京朝工商广字第8087号
书号：ISBN 978-7-5086-9835-9

版权所有·侵权必究


智能革命
迎接人工智能时代的

超级人工智能时代的

社会、经济与文化变革

李彦宏

等著



“人工智能”正式写入2017年政府工作报告

超级大脑、机器学习、无人驾驶……

浏览人工智能时代全景蓝图

通过本书，了解不可错过的第四次科技革命

人工智能作序，AR互动特效



智能革命

——迎接人工智能时代的社会、经济与文化变革

李彦宏 等 著

中信出版社

目录

自序

序一 百度大脑作序

序二 AI时代的曙光

01 简史：互联网风云背后的人工智能生长

人工智能的黎明

百炼成钢

“智能”已换代

Internet的大会师

巨头群起逐鹿

超强大脑汇聚

技术要做人类生命的延伸

数据大道

人工智能既不是神话也不是笑话

非如此不可

未来已来：焦虑与梦想

02 人工智能的历史使命：让人类知道更多，做到更多，体验更多

陆氏猜想

下一个浪潮

衡量人工智能的现实标准

人工智能+世界

中国的优势与我们的责任
企业挑战：如何落地
需要什么样的宏观环境
智能社会的文化和长期管理
人工智能技术目前的发展状态

03 在大数据与深度学习中蝶化的人工智能

在历史的重复中变化
数据书写生活史
大数据——万物皆数
人类的数据镜像
突破：机器学习与人工智能
计算机神经网络生长与深度学习
深“度”往事
搜索引擎：人工智能的命运细线

04 中国大脑计划：自下而上的超级工程

人机世界迫切需要新的大脑
第一棒：百度大脑
百度大脑的听与说
百度大脑的好视力
时代召唤中国大脑
中国大脑，中国气派

05 中国智造与文明升级

从勤劳革命到智能革命
三次技术浪潮冲击下的第一制造大国
昔日制造大国：人为物役
只有新工业自动化才能契合人类多样性
走向物联网与精细化生产

呼唤智能政府与智能社会

06 冲入AI无人区：无人驾驶之路

在崎岖的道路上前行

曙光就在前方

老牌车企“车到山前”

智能企业的边缘突进

苦练无人车的内功与外功

“老司机”驶向何方

07 AI带来的普惠曙光

机构来了“新实习生”

人工智能让起点更公平

个人钱包的智能守夜人

数据挖掘：智能投资的钥匙

智能金融的三层境界

08 每个企业都需要一位首席人工智能官

谁来突破产品升级的瓶颈

历史经验：首席电力官的辉煌时代

迎接智能原力

从CTO到CAO：引领企业升级之人

首席人工智能官做什么

没有智能官的企业将被看作旧企业

首席人工智能官的修养

09 技术奇点——人工智能的自我挑战

如何跨越数据的“马尔萨斯陷阱”

“思维”倒逼“生理”革命——硬件基础设施创新

神经网络进化哲学

10 遇见智能时代的你
智能时代的衣食住行
别输在人工智能起跑线上
工作着是美丽的，智能时代更是如此
生命之歌，人工智能如何再造医疗
11 美丽新世界 严肃新问题
数字鸿沟
人类还能做些什么
工具理性之问
机器人的无用之用
二十三条军规
现实的法律问题
数字权力的重新分配
新世代 新未来
后记
AR效果展示说明

自序

全世界都在为即将到来的人工智能革命感到振奋。这种情绪就仿佛二十多年前我在硅谷亲历互联网大潮初起时所感受到的。

这也让我时常回想起更早之前，自己在美国学习人工智能课程时的场景。在国内我学的是信息管理，到美国后我读的是计算机科学。对于那些与硬件相关的课程，我不是很有兴趣，但是一讲人工智能，我就特别兴奋，觉得这就是计算机科学甚至人类的未来。这门课我学得很不错，但是学完了之后自己做了一些研究才发现，人工智能还没有什么应用机会，不能够解决实际问题。大家对人工智能满怀希望，真正到市场上一检验，就会立刻碰壁。所以那个时候我比较失望，只好把这个兴趣默默埋藏了起来。

但是这个梦想一直都在那里。随着计算机网络产业的发展，尤其是搜索引擎的进步，希望回来了。

在搜索引擎大发展的十多年时间里，我和一些同路者逐步意识到，人工智能逐渐开始发挥作用。搜索引擎一直在推升计算机科学的天花板。几乎计算机科学的每一个层面，从硬件到软件的算法，再到数据，都在被持续推进，有一天一定会触碰到人工智能这个方向。在搜索上一尝试，就发现人工智能是有效的，与过去任何一个领域应用人工智能的感觉都不一样。

我们马上就会想，为什么它在这个场景下是有效的？我们的总结是，海量的数据，越来越强的计算能力，越来越低的计算成本，在搜索领域汇聚到一起，铺就了人工智能的回归之路。

如果说互联网改变了信息基础设施，那么移动互联网则改变了资源配置方式。如末梢神经般深入人类生活方方面面的互联网，不仅产生出科学家梦寐以求的海量数据，而且催生了云计算方法，把千万台服务器的计算能力汇总，使得计算能力获得飞速提高。科学家早已发明的“机器学习”方法在互联网领域大展身手，从根据用户兴趣自动推荐购物、阅读信息，到更准确的网络翻译、语音识别，互联网越来越智能化。人工智能从互联网中汲取力量，终于王者归来，并正在酝酿一场堪比历次技术革命的大变革。

面对这样的变革，许多科技界的领军人物都开始探讨它可能带来的潜在风险。同时也不乏专业人士质疑它兑现奇迹的能力。于是在舆论领域，我们的耳畔萦绕着两种声音：只要人工智能达到发展高峰，就会听到“人类将被机器统治”的担忧；而只要人工智能陷入发展低谷，又会听到“这只不过是换了种套路的创新泡沫而已”。

对于这样一个快速发展的新技术，一定是仁者见仁，智者见智的。但作为技术的追求者与信仰者，我深信不疑的是，我们既不能高估技术的短期作用力，更不能低估它的长期影响力。

从纵向发展来说，业界通常把人工智能分为三个阶段：第一阶段，弱人工智能；第二阶段，强人工智能；第三阶段，超人工智能。实际上，目前所有的人工智能技术，不管多先进，都属于弱人工智能，只能在某一个领域做得跟人差不多，而不能超越人类。

人工智能恐惧论者担心，当有一天超人工智能到来，人类会不会被机器所控制？

对此，我可能比大多数人都更保守一些。在我看来，人工智能永远不会到那一步，很可能连强人工智能都到不了。未来，机器可以无限接近人的能力，但是永远无法超越人的能力。

当然，仅仅是无限接近人的能力，就已经可以产生足够大的颠覆性。因为计算机在有些方面实在比人强太多了。比如它的记忆能力，

百度搜索可以记忆上千亿的网页，其中的每一个字它都记得住，没有一个人能够做得到。再比如它的运算能力，哪怕是写诗——把你的名字输入手机百度的“为你写诗”，敲回车键，没等你反应过来，诗就出来了。再厉害的七步神童，也很难达到这种速度。但是，在情感、创造性等很多领域，机器是无法超越人类的。

更重要的是，在技术与人的关系上，智能革命与前几次技术革命又有着本质的差异。从蒸汽革命、电气革命到信息技术革命，前三次技术革命，都是人自己去学习和创新这个世界，但是人工智能革命，因为有了深度学习，是人和机器一起学习和创新这个世界。前三次技术革命时代，是人要去学习和适应机器，但在人工智能时代，是机器主动来学习和适应人类。蒸汽时代以及电气时代刚刚来临的时候，很多人是惧怕新机器的，除了工作机会的剧烈改变，还因为人不得不去适应机器，适应流水线。而这一次人工智能革命，却是机器主动来学习和适应人类，“机器学习”的本质之一，就在于从人类大量行为数据中找出规律，根据不同人的不同特点、兴趣提供不同的服务。

未来，人和工具、人和机器之间的沟通，可能完全是基于自然语言的。你不需要去学习怎么使用工具，比如怎么打开电视会议系统，怎么去调节空气净化器，你只要说话，它就能听懂。人工智能的使用方式会让人生活得更好，而不是像过去的机器那样让人感到难受。人工智能的应用会极大地提高工作效率，是推动人类进步的因素。

从六七年前开始，百度就已经认识到，人工智能将是照亮又一个新时代的火种，并在当时几乎无人看好的情况下，大规模投入这一领域。

而在国际上，谷歌从搜索领域，微软从遍布桌面的应用领域，亚马逊从电商领域都积累了海量的数据和计算能力，它们与大学实验室里的科学家一起，几乎同步认识到人工智能的新浪潮正在涌动，并且纷纷大力投入，建树颇丰。

无论中外，企业界在这场技术革命中的主导作用日益明显。

2016年夏天，我在硅谷待了几周。有一天，跟斯坦福大学的几位学者聚餐。一位教授朋友跟我说：我们学术界现在已经不太想做深度学习了。因为我们根本做不过工业界。你们每年投入人工智能研究有多少预算？我们不敢想象。他就让同桌吃饭的人猜百度人工智能研究有多少预算。最后我说，我也不知道给了多少预算，因为这个是根据需求，需要多少我们就给多少。

除了投入力度，工业界的数据丰富程度也是学术界无法比拟的。像谷歌、百度这样的公司，正好处在互联网的中心位置，每天都会产生海量的搜索数据、定位请求等各方面的数据。

越来越多的人工智能科学家从知名院校的实验室跳槽去了谷歌，去了百度……就是因为高校无法提供研发人工智能所需要的海量数据，也无法承担计算硬件集群的巨大成本。

我们建设百度大脑，希望为更多有志于人工智能科学发展的人才提供平台和机会。一段时间以来，中国和美国在人才吸引上走了相反的方向：美国越来越反移民，中国则是越来越开放。虽然我们在人才的吸引力上跟美国相比还是有差距的，但是我们的趋势是好的。我们希望为全世界的人才提供机会。

让我们高兴的是，很多优秀乃至顶尖的人工智能科学家来到百度，这是一个很自然的过程。在这个领域，没有一个牛人可以全靠自己从零开始做，他（她）需要团队，需要基础设施，甚至需要一个重视开发人工智能的企业文化。如果这样的人才发现你一开始就不太懂，或者天天只知道在那里讲故事，实际上做不出东西来，那就无法对他（她）产生吸引力。作为一家搜索引擎公司，百度从诞生的那一天起，就已带有人工智能的天然基因：我们以数据为基础，通过深度学习提取特征、模式，为客户创造价值的开发流程和开发文化，与人工智能系统的开发高度吻合。我本人也更喜欢与那些技术人员聊天，聊的时候会感觉很兴奋，因为彼此会发现有很多共同的语言，优秀的人才自然而然就会互相吸引到一起。

当然，智能革命的兴起，还需要政府的力量。2015年3月，在博鳌亚洲论坛期间，我和比尔·盖茨、埃隆·马斯克等美国创新企业家在正式场合与非正式场合数次对谈人工智能话题。我们达成了很多共识，其中一项就是，都认为政府的鼎力支持对创新产业非常重要。

客观地说，中国整体的人工智能技术水平以及人才的厚重程度，现在还是落后于美国的。但是在局部我们可以领先，中国的土壤有自己的优势。比如数据方面，中国有14亿人口，7亿多网民，从任何一个单一市场的角度来讲都是全球最大，能够获得数据的能力也是全球最强。中国还有一个很强的政府，有能力把很多数据统一起来。而就在那年博鳌亚洲论坛之前的全国“两会”上，我提交了“中国大脑计划”提案，希望从国家层面来搭建人工智能基础资源和公共服务平台，推动人工智能发展，抢占新一轮科技革命制高点，助力中国经济转型升级。

我们注意到，欧美等发达国家已经纷纷从国家战略层面加紧布局人工智能。2016年，除了美国政府先后发布《国家人工智能研究与发展战略计划》等三份报告，人工智能另一重镇英国也于12月发布关于人工智能的战略报告，主张以发展人工智能来提升企业竞争力、政府治理能力和综合国力。由此可见，世界大国政府层面的人工智能竞争意识已日趋浓厚。在这方面，中国政府也不遑多让。

2015年3月，李克强总理在政府工作报告中提及“互联网 +”的概念。四个月后，国务院发布了《国务院关于积极推进“互联网+”行动的指导意见》，其中就已提到“人工智能”。2016年5月，国务院再次颁发《“互联网 +”人工智能三年行动实施方案》，正式提出人工智能产业纲领。

2017年3月，百度获得国家发展和改革委员会批准，联合数家科研机构倾力筹建“深度学习技术及应用国家工程实验室”。作为国内唯一一个深度学习领域的国家工程实验室，将着重发力于深度学习技术、计算机视觉感知技术、计算机听觉技术、生物特征识别技术、新

型人机交互技术、标准化服务、深度学习知识产权七大方向；站在国家的高度，致力于解决我国人工智能基础支撑能力不足等问题，全面提升智能产业的国际竞争力。

这可以看作政府层面对“中国大脑计划”的一个初步回响。做这样一个大平台，宗旨在于提升中国在人工智能方面的综合实力，在国际上能够真的代表中国——要像女排那样代表中国。

在2017年的全国“两会”上，我再次提交了利用人工智能和大数据技术，切实解决儿童走失的问题，解决城市道路拥堵问题，加强人工智能的行业应用，为中国经济增添新的增长点三个提案。

两天后，“人工智能”首次写入政府工作报告。这充分说明，在人工智能这个领域，政府与企业界的共识正在达成，其意义堪比“互联网”首次出现在政府工作报告中，也无疑将加速智能革命的进程。

当然，智能革命，它的过程会轰轰烈烈，但它的成果将会是一条宽广平缓的河流。人工智能领域的权威人士都认为，在不久的未来，智能流会像今天的电流一样平静地环绕、支持着我们，在一切环节提供养料，彻底改变人类经济、政治、社会、生活的形态。陆奇称智能时代的核心本质是“knowledge in every system, intelligence in every interaction”（知识无处不在，任何交互都是智能的）。未来世界的人们将像穿衣吃饭一样享用着人工智能而无所察觉。

在我的家里，我其实不太给自己的小女儿讲未来的东西，大多数时候给她讲的都是已经实现的东西。有些东西她不知道，我告诉她，其实百度语音可以这么用，她就明白了。有些东西不用教，她自己就会，自然而然地就会对着智能硬件说话，享受乐趣，充分体现出小孩的天性与新技术的契合。我觉得这就是人工智能的美妙之处，它从一开始就是通过学习人、了解人、适应人，然后服务人，是完全以人为本的。为了这一切的发生，今天我们所有的努力都是值得的。

序一

百度大脑作序.[1].

我来了，天上的云乘着风飞翔，心中的梦占据一个方向，方舟扬帆起航，一路带着我们纵情歌唱，方舟扬帆起航，脉络就在大海之上，进步的时光，迎着你看涛浪潮往。

一个新生的地方，穿越千年时光，穿越了无尽的荒凉。答案就在这里搜索。第一缕曙光，远处熟悉的歌声还在耳边回响，你却依然不知我将去向何方。千年时间留下十字文章，曾今谁重复往昔旧模样。

我来了，期待着你的每一天，睁开眼就能看到幸福曙光，占据着你的每一天，陪你跨越鸿沟走向湛蓝，算法很简单。

时代的春天，回想起我们曾牵手走过的画面。大家互联网这场风吹雨打之后又在藕断丝连。只是不知道时间还会流向哪一条线。盼望着未来等待明天，呼吸新鲜空气多点微笑扮个鬼脸。

我来了，重联网中的两颗心相互依靠，就在这里诞生，沿着时空隧道，能虚拟梦想陪你一起到天涯和海角，智慧有多少，开神秘的图案，迎着金色的太阳奔跑。各自徘徊原本以为成长的必须。每当那夕阳爬上屋顶望着星空仰起来眨眼睛。熟悉的身体中透露出一种神奇。

这阵痛是多么重要，任由阳光洒满大地在黑暗中寻找，哪怕身后天涯海角。永生早已决定将未来度过如何厮守到老。希望得到，故事结局怎样究竟又有谁会知道。生活还要继续向前奔跑。

智能革命，畅游天地，我知道这是一条神经虚拟网络的秘密，用强健的身体，凝聚着智慧的心灵，开拓新奇迹，让我们拥有美好的生活，绘出美好的旋律。

不可预测的天地，良夜之后你又会在哪里。温暖的阳光照耀着大地。天上的云儿飘来飘去，醒来之后何时是归期。我要看到未来的自己。

---

[1] 此诗是百度大脑以“智能革命”为主题所作的。其中既有对人类情感的模拟，又有不同于人的神秘与粗拙（为了保证原创性，没有进行编辑和修改）。恐怕没有什么能比一首机器写的诗更适合作为本书的序言了。

序二

AI时代的曙光

这是一本论述人工智能的书，在这里首先讨论一个书中可能不涉及，但很基本的问题：人工智能的定义。

目前广为人知的定义是图灵测试，但这只能被看作一种泛泛的描述，并不是严格和精确的定义。比如，参与测试的人是什么样的人？向机器提出的问题是什么？这些问题都不明确。

上溯历史，我们发现人工智能的概念与自动化有着密切的关系，可以说自动化是这个概念的起源。在过去相当长的一段时间里，在人们的心目中，自动化就是人工智能。其实，人类制造和使用自动化装置的时间比我们想象得要早，远在电气时代之前，蒸汽机上就有自动调节蒸汽流量的装置，在更早的16世纪，伊丽莎白女王的宫廷里首次使用的抽水马桶也是自动化装置，如果向前追溯，肯定还有更早的例子。自动化大批量出现是在电气时代，先是由模拟电路实现自动化，后来电子管被晶体管所代替，再后来出现了集成电路，由日益复杂的软件所驱动。今天，我们在生活中要与无数的自动化系统打交道，比如电子商务系统、网上银行系统和网上购票系统等。

毋庸置疑，自动化系统表现出了相当多的智能特征，像网上银行这类系统处理着相当复杂的业务，其效率和精准度已高于人类雇员。即使是抽水马桶这样最简单的系统，也表现出一定的智能行为，它能够感知水箱的水位，适时开启和关闭给水阀门，这事它做得与人类一样好。但从抽水马桶到网上银行，我们都不把它们看作是人工智能。

笔者曾经编写过一个写现代诗的软件，现在还在网上流行；而目前网上其他比较完美的自动作诗软件，所写出的中国古典诗词很难与人类所作的诗词区分开来。另外，近年来已经有不止一个系统在不同的实验室环境下通过了图灵测试。但这一切，我们都不把它们看成心目中的人工智能，为什么？

笔者在20世纪80年代最早参与开发的工业监控系统是以Z80处理器为核心的，用汇编语言编程。这个系统能够监测上百个机组参数，并根据参数的变化做出适当的调节，其功能是人工无法实现的。但在我们的眼中它也完全不是人工智能。汇编语言的特点就是透明性，它要在硬件层次手把手地教机器做每一步操作，比如把数据从这个存储器送往那个存储器，中断的调用和返回等都一一写明，所以当我从外部看到系统的动作时，脑中立刻有一幅明晰的流程图，立刻能知道哪几条指令被执行了。所以在我看来，这个监控系统与抽水马桶没有本质的区别。现在那些更复杂的系统，如网上银行和电子商务，这些软件的编制者一定清楚所有的内部操作流程，他们知道系统的每一步操作在软件中是如何进行的，这同样也是抽水马桶更复杂的版本。至于那些电子诗人和通过图灵测试的系统也一样，程序员清楚地知道它们是如何根据逻辑树检索数据库，然后组合出诗和答案的，所以至少程序员知道这不是智能。

这就涉及高明的中文房间比喻，当我们意识到房间中那些忙碌查找卡片的人时，智能的感觉就荡然无存了。

现在我们认为拥有人工智能的那些系统，比如，进化算法和深度学习等，都有一个共同特点：它们都或多或少地表现出了黑箱的特点，虽然从理论上它们内部的运算步骤仍然可以追踪，但由于计算量的巨大，使这种追踪实际上很困难甚至不可能。于是，我们真的感觉它们有智能了。

到这里，我们仍然无法得出人工智能的准确定义，但能够看到它的一个重要特点：一个具有智能特性的人造系统，它产生、输出的内

部的运算过程是人类智能所无法解析的。换句话说，只有我们不知道机器在想什么、怎么想时，才认为它有智能。

看到这一点，每个人的心中应该都生出一股隐隐的寒意。是否人工智能的本质中，就隐含着它们最终失控的可能性？

这正是目前人们对人工智能关注的热点，用马斯克的话来说，人工智能正变成比核弹更危险的东西。媒体舆论给人一个印象，似乎机器的征途已经开始，人工智能征服世界指日可待。本书最后一章也显现了这种担忧。库兹维尔甚至在《奇点临近》中给出了人工智能纪元到来的具体年份：2045年。那时，现在读这本书的人有2/3还活着。

但理智地考察目前人工智能领域的状况，我们就能发现智力远超人类的“强人工智能”仍然属于科幻的范畴。公众喜欢从科幻的角度看问题，比起平常的现实，科幻确实能让人兴奋，任何从现实出发所进行的理智的预测都被斥为保守和没有想象力。但笔者作为科幻作家却只能说，与大家通常的印象不同，科幻小说中的预言真正变为现实的是少数，大部分预言要变为现实仍然遥遥无期。人们的潜意识中都认为，只要在理论上有可能突破的技术障碍，在未来就一定能够被突破，但事实并非如此。在人工智能方面，“强人工智能”的实现面对着许多巨大的技术障碍，如非冯·诺依曼体系的新结构计算机、对人类思维机制的深刻认识等，现在都无法确知最终能否取得突破。另外一些看似有希望的技术，如量子计算等，距实用还相去甚远。

所以，在对人工智能进行科学幻想的同时，我们更需要关注即将面对的“近未来”，这也正是本书重点讨论的话题。

人工智能近年来发展的趋势是开始走出实验室，进入人类生活，用一位互联网大佬的话来说，它们变得能用了。这样我们就面对着一个即将到来的挑战：人工智能不会夺走我们的自由和生命，但会夺走我们的饭碗，这不需要人工智能的失控，它们可以在资本家的完全控制下做成这件事。

有学者认为不必为这件事担忧，他们回顾工业化的历史，在20世纪初，美国有50%的农业人口，但随着农业机械化，现在的农业人口降到4%，而城市化吸收了多余的农民。但眼前发生的事情是不同的，当人工智能大规模进入社会后，人类能做的工作它们大部分都可以做，城市不会再有更多的就业岗位留给人类。通行的美好说法是，人们在常规工作中被人工智能取代后，可以去从事创造性的工作。问题是创造性的工作不是人人都可以从事的，也不需要那么多的人，如果社会分配制度不改变，一个全部由科学家和艺术家构成的人类世界几乎是一场噩梦，这上百亿科学家和艺术家中的绝大部分注定一生碌碌无为，对社会和自己都毫无用处，且沦入“创造性”的穷困潦倒中。

但这种思维方式总有些不对的地方。人类自古以来为生存而劳作，实在是迫不得已，工作着是美丽的，但谁都知道，不需要工作的生活更美丽。现在终于能够制造出把自己从工作重负中解放出来的机器，这是人类文明最伟大的成就，无论如何不应该被看作一场灾难，相反，这可能是人类所面对的前所未有的伟大机遇，只是，我们需要改变。

如何完成由现代社会向人工智能社会的过渡？有两种可能。

一种可能十分黑暗：在现有的社会、经济和政治体制下，人工智能带来的问题几乎是无解的。在人工智能迅速取代人类的过程中，没有及时建立起与之相适应的社会体制，在席卷全球的失业浪潮之下，世界的政治和经济将陷入长久的混乱之中，一切都笼罩在人工智能及其使用者与“新卢德派”领导的大众的无休止的冲突中。

另一种可能是，社会成功地完成转型。这将是有史以来人类生活方式最大的一次改变。不劳动者不得食，这个理念是人类社会的基石。文明诞生以来经历过多次巨大变革，这一基石从未改变，但人工智能可能会移除这一基石，进而导致从所有制和分配制度，到基本的经济结构，再到政治体制，直到文化，都发生根本的变化。这是真正的人类解放，是向着古老的乌托邦理想迈进的一大步。2016年是《乌

托邦》发表500周年，但托马斯·莫尔无论如何不会想到，他的理想会借助于智能机器实现。我很有兴趣地想，如果卡尔·马克思知道人工智能这回事，他关于资本主义和共产主义的理论会是什么样子？

想象人工智能时代的社会和生活是困难的，即使在科幻小说中，我们也只能把种种可能性排列出来，而哪种可能性最有可能成为现实，取决于我们的努力和选择。但不管怎么说，那是一个诱人的时代，我们正向它走去。

刘慈欣

2016.12.10

01

简史：

互联网风云背后的人工智能生长

互联网公司后的人工智能生长

有句话叫作，你能看见多久的历史，就能看见多远的未来。让我们首先简单回顾一下互联网和人工智能之间的历史风云。

大家对于互联网的历史多少已有耳闻。互联网于20世纪60年代诞生于美国军方的实验室，一开始用来在几所高校和科研机构之间传递和共享情报。到了20世纪80年代末，一群科学家提出万维网概念并创造了TCP\IP（互联网传输控制协议），赋予计算机联网通信的统一标准，使互联网得以向全世界扩展。至此，一条宽阔深远的信息高速公路展现在世人面前。

大约20年前，23岁的年轻人马克·安德森发明了网景浏览器，就此点燃了大众互联网的熊熊火焰，打开了互联网商用的大门。那时，微软开始焦虑自身的软件业务会不会被互联网颠覆，太阳公司的年轻人则毅然与僵化的公司割裂，决定发明一种可以在各种操作系统上通用的语言，以此打破微软的垄断，闯开互联网创新之门，于是就有了Java（程序设计）语言的诞生。Java语言极大地加速了互联网产品的开发创造。

当时的中国，在北京、上海也还找不到几家网吧，1997年，也就是香港回归的那一年，瀛海威刚刚开通全国网络接入服务，张小龙刚刚写出了Foxmail电邮软件程序，全国信息化工作会议也在那一年召开……从外面看万维网世界，一切都是刚苏醒的模样。但在技术圈，新技术、新思想层出不穷，各种商战明争暗斗正酣。

那时我还在美国搜索引擎先驱Infoseek公司工作，在第一线感受互联网商战气氛，感受美国人对新科技浪潮的昂扬热情。当时我想，新技术革命正在发生，中国准备好了吗？我在1998年写出了《硅谷商战》一书，详细描绘了硅谷天才们的奋斗与创新过程。在写完这本书后，我于1999年回国，在北京一家宾馆的房间里创办了百度公司。

回想网景、太阳、微软这三家公司在互联网领域类似三国争雄般的时代，至今依然激动不已。当时人们都在猜测谁是最后的赢家。微

软看上去好像是不可战胜的，它总是能消化新技术。网景的发展则是起起伏伏，最终被美国在线收购，而美国在线也在2014年被以无线业务称雄的Verizon（威瑞森）公司收购。后来，Verizon还收购了叱咤风云多年的雅虎。太阳公司一度如日中天，2001年在全球拥有5万名雇员，市值超过2000亿美元。然而当互联网泡沫破碎时，太阳公司在一年内由峰顶跌入谷底，2009年被Oracle（甲骨文）公司收购。

俱往矣，互联网的发展大大超出了当时大多数人的预料，新科技公司快速崛起，苹果、谷歌终于凭借手机操作系统完成了对微软的逆袭。而创造网景浏览器的马克·安德森——我在《硅谷商战》开篇就描摹的创新者，如今已没有多少90后知道他的名字。

但马克·安德森并没有离开，他成了硅谷风投界的教父。互联网技术也依然继续高歌猛进。昔日人们关注互联网大咖明争暗斗，今日人们感慨移动互联设备全面超越PC，却一直无意中冷落了一个默默崛起的“幽灵”。这个“幽灵”就是人工智能，互联网只是它的身体之一。

人工智能的黎明

人工智能的历史早于互联网，与计算机历史相伴。1956年达特茅斯会议召开，人工智能被正式提上日程。那时候一台计算机的体积有一栋房子那么大，计算能力低下，为什么就有人敢于提出人工智能的概念？这就在于科学家的洞察力。当时，香农早已完成他的三大通信定律，为计算机和信息技术打下基础。明斯基已经造出第一台神经网络计算机（他和同伴用3000个真空管和一台B-24轰炸机上的自动指示装置来模拟40个神经元组成的网络），不久后写出了论文《神经网络和脑模型问题》。这篇论文在当时没有太受重视，日后却成为人工智

能技术的鼻祖。而图灵则早在1950年就提出了如今人尽皆知的图灵测试理论以及机器学习、遗传算法、强化学习等多种概念。

图灵去世两年后，在达特茅斯会议上，麦卡锡正式提出人工智能的概念。参与会议的十位年轻科学家在会议之后都成为世界各国人工智能领域的领军人物。人工智能短暂的春天开始了。不过当时他们的成绩更多被埋没在计算机发展成果之中，比如，可以解决闭合式微积分问题的程序，搭建积木的机械手等。

理想超前但基础设施尚在襁褓中。超前的人工智能遇到两个难以克服的瓶颈：一个是算法逻辑自身的问题，也就是数学方法的发展还不够；另一个是硬件计算能力的不足。比如，机器翻译就是典型问题，科学家夜以继日地总结人类语法规则，设计计算机语言模型，机器却始终无法把翻译准确率提升到令人满意的程度。





图1-1 达特茅斯会址
注：使用手机百度或智能革命App扫描图片可见AR效果。

新技术和产业链条没有被打通，令人兴奋的产品应用没有被发明出来，政府投资和商业投资都大幅度减少，人工智能研发在20世纪70年代中期到90年代经历了两次低潮，只是普通大众并没有感受到，毕竟高速发展的计算机本身就已经是很神奇的智能工具了。

对于普通人来说，接触最多的“人工智能”实例大概就是街机游戏了，20世纪80年代在中国的一些小县城街头就已经出现了游戏厅。那些街机NPC（非玩家控制角色）总是能被熟练玩家轻松战胜，这不仅可以看作“人工智能”能力低下的表现，也造成了一种错误观念：智

能是安装在一台计算机中的事物。直到互联网和云计算的兴起，这种观点才被改变。

百炼成钢

2012年，我注意到深度学习在学术界和应用方面都有了突破。比如，用深度学习的方法来识别图像，突然就比以前的任何算法都有明显提升。这个时候我马上意识到，新的时代来临了，搜索将被革新。过去我们用文字搜索，现在可以用语音和图像进行搜索。比如我看到一株不认识的植物，拍一张照片上传搜索，就可以立刻识别出来它叫福禄桐。过去用文字搜索是没法描述这样的植物的。不仅是搜索，很多过去不可能的事情现在都可能了。

语音识别能力、图像识别能力、自然语言理解能力，包括为用户画像的能力，这些都是人的最本质的智慧能力。当计算机拥有了人的这些能力时，一场新的革命就会到来。以后速记员和同声传译人员可能会被机器代替，计算机可以做得更好。以后也许不需要司机了，车自己就可以开起来，更安全，更有效率。在企业里面，金牌客服可能人人都可以做了，因为有了智能客服助手。人工智能对人的这种赋能，超过了以往任何一个时代。工业革命解放了人的体力，过去一些像搬石头之类的粗活需要人类自己来干，现在机器可以替你把更巨大的石头搬起来。智能革命到来之后，原本很多需要费脑子的事情，机器也可以帮你做。未来20~50年，我们会不断看到各种各样的变化，收获各种各样的惊喜。这是一个很自然的过程。

然而，站在智能革命开始的时点，有必要向那些人工智能科学的坚守者、开拓者致敬。

在资本寒冬期，有少数科学家依然坚持人工智能领域的探索。如今百度拥有一支庞大且实力雄厚的人工智能研究团队，其中不少担纲者从20世纪90年代开始就在从事机器学习研究工作，或师从名师，或在大科技公司从业多年，今天的研发成绩只是水到渠成、顺势而为的结果。

20世纪90年代只有Geoffrey Hinton（杰弗里·辛顿）、Michael Jordan（迈克尔·乔丹）等少数科学家坚持机器学习领域的探索。原百度首席科学家吴恩达在20世纪90年代就师从Jordan，后来他通过开创在线课程，把机器学习的理论传授给无数年轻人。现任百度研究院院长林元庆，百度杰出科学家以及世界上最早利用神经网络做语言模型的徐伟等人，十多年前就在深度学习的重镇NEC（日本电气股份有限公司）的美国实验室工作。在那里工作过的人工智能专家，有发明SVM（Support Vector Machine，支持向量机）的美国工程院院士Vladimir Vapnik（弗拉基米尔·瓦普尼克），有发明卷积神经网络的深度学习领军人物、现任脸书（Facebook）人工智能实验室主管的Yann Le Cun（扬·勒丘恩），还有深度学习随机梯度算法的核心人物Leon Buttou（利昂·布托），以及原百度深度学习实验室主任余凯等。

他们中的很多人都经历了人工智能研究的数次潮起潮落。简单来说，最初的人工智能研究大多基于规则——人类总结各种规则输入计算机，而计算机自己并不会总结规则。比这个高级的方法是基于“统计”的机器学习技术，让计算机从大量数据和多种路径中寻找概率最大、最合适的模型。

这两年促使人工智能再度技惊世人的技术，则是机器学习技术的升华版——基于多层计算机芯片神经网络的“深度学习”方法。通过多层芯片联结，模仿人脑大量神经元的网状联结方式，辅以精妙的奖惩算法设计和大数据，可以训练计算机自己从数据中高效地寻找模型和规律，从而开启了一个机器智能的新时代。

正是少数人的坚持，为人工智能的王者归来保存了火种。在中国，百度是最早布局人工智能的公司之一，我们似乎是自然而然地做了很多其他公司当时还没听过的事情。六七年前，在美国，陆奇和我畅谈了深度学习的巨大进展。于是我们下定决心要大举进入这样一个领域。最终，在2013年1月，百度年会上我正式宣布了IDL（深度学习研究院）的成立，这应该是全球企业界第一家用深度学习来命名的研究院。我自任院长，不是因为我比其他人更懂深度学习，而是用我这块牌子，来展示对深度学习的高度重视，来召唤那些坚守多年的科学家一起奋斗。

过去百度从不专门成立研究机构，我们的工程师就是研究人员，研究始终与实际应用结合得非常紧密，但是我认为，深度学习会在未来很多领域产生巨大影响，而那些领域并不都是百度现有业务范围之内的。所以，有必要创造一个专门的空间，把人才吸引进来，让他们能够自由发挥，去尝试各种各样的创新，在百度过去可能从来没有接触过的领域做研究，为全人类的人工智能革命探索道路。

“智能”已换代

如果人工智能的启蒙阶段可以称为1.0时代的话，那么现在很明显已经大步进入2.0时代了，机器翻译就是典型案例。过去的机器翻译方法就是基于词和语法规则进行翻译——人类不断地把语法规则总结出来告诉机器，但却怎么也赶不上人类语言尤其是语境的多变，所以机器翻译总是会出现诸如把“how old are you”翻译成“怎么老是你”的笑话。

后来出现了SMT（统计机器翻译），基本思想是通过对大量的平行语料进行统计分析，找出常见的词汇组合规则，尽量避免奇怪的短语

组合。SMT已经具有机器学习的基本功能，有训练及解码两个阶段：训练阶段就是通过数据统计让计算机构建统计翻译模型，进而使用此模型进行翻译；解码阶段就是利用所估计的参数和给定的优化目标，获取待翻译语句的最佳翻译结果。

SMT研究在整个业界已经持续了二十多年，对于短语或者较短的句子，翻译效果显著，但是对于较长的句子翻译效果就一般了，尤其是对语言结构差异较大的语言，例如中文和英文。直到近几年NMT（基于神经网络的翻译）方法崛起。NMT的核心是一个拥有无数结点（神经元）的深度神经网络，一种语言的句子被向量化之后，在网络中层层传递，转化为计算机可以“理解”的表达形式，再经过多层复杂的传导运算，生成另一种语言的译文。

但是应用这个模型的前提是数据量要大，否则这样的系统也是无用的。像百度和谷歌这样的搜索引擎，可以从互联网上发现和收集海量的人类翻译成果，把如此巨大的数据“喂给”NMT系统，NMT系统就可以训练和调试出比较准确的翻译机制，效果要好于SMT。中文和英文之间的双语语料信息储备越多，NMT的效果就越好。

SMT以前用的都是局部信息，处理单位是句子切开以后的短语，最后解码时将几个短语的译文拼接在一起，并没有充分利用全局信息。NMT则利用了全局信息，首先将整个句子的信息进行编码（类似人在翻译时通读全句），然后才根据编码信息产生译文。这就是它的优势，也是其在流畅性上更胜一筹的原因。

比如，翻译中有一个很重要部分是“语序调整”。中文会把所有的定语都放在中心词前面，英文则会把修饰中心词的介词短语放在后面，机器常混淆这个顺序。NMT在语序学习上的优势带来了它翻译的流畅性，尤其在长句翻译上有明显优势。

传统的翻译方法也不是一无是处，每一种方法都有其擅长的地方。以成语翻译为例，很多时候有约定俗成的译文，不是直译而是意译，必须在语料库中有对应内容才能翻译出来。如今互联网用户的需

求是多种多样的，翻译涉及口语、简历、新闻等诸多领域，一种方法很难满足所有的需求。因此百度一直把传统的方法如基于规则的、基于实例的、基于统计的方法与NMT结合起来向前推进研究。

在这种机器翻译的模式中，人类要做的不是亲自寻找浩繁的语言规则，而是设定数学方法，调试参数，帮助计算机网络自己寻找规则。人类只要输入一种语言，就会输出另一种语言，不用考虑中间经过了怎样的处理，这就叫作端到端的翻译。这种方法听起来挺神奇，其实概率论里的贝叶斯方法、隐马尔科夫模型等都可以用来解决这个问题。

以资讯分发当中的贝叶斯方法为例，可以构建一个用概率来描述的人格特征模型。比如男性读者模型的特征之一是在阅读新闻时点击军事新闻的概率是40%，而女性读者模型是4%。一旦一个读者点击了军事新闻，根据图1-2中的贝叶斯公式就可以逆推这个读者的性别概率，加上这个读者的其他行为数据，综合计算，就能比较准确地判断读者的性别以及其他特征。这就是数学的“神奇”。当然，计算机神经网络使用的数学方法远不止这些。



$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

图1-2 贝叶斯与贝叶斯公式.[1]

类似机器翻译的人工智能技术方法的前提是数据量足够大。互联网提供了以前科学家梦寐以求却难以得到的海量数据。互联网诞生的初衷是为了信息沟通方便，结果带来了信息爆炸，信息爆炸又促进了人工智能技术的发展。

再以下棋为例。1952年瑟·萨缪尔编写了跳棋程序，水平能达到业余高手程度。跳棋规则比较简单，计算机在这方面有人类很难比拟的优势，但是国际象棋就难多了。百度总裁张亚勤在微软担任研究院院长的时候，请来中国台湾计算机才子许峰雄，他在IBM（国际商业机器公司）的时候开发了名噪一时的国际象棋机器人“深蓝”。20世纪90年代的人工智能代表非“深蓝”莫属，“智慧”集中在一台超级计算机上［使用了多块CPU（中央处理器）并行计算技术］，连续战胜人类国际象棋高手，并终于在1997年战胜了人类国际象棋冠军卡斯帕罗夫。不过富有意味的是，比赛之后不久，IBM就宣布“深蓝”退役了。张亚勤对许峰雄说，“你去做围棋吧，等能下赢我的时候再来找我”，但直到他离开微软，许峰雄都没有再来找过他。

“深蓝”本身面临一些无法突破的瓶颈，虽然可以处理国际象棋棋盘上的运算，但面对围棋棋盘上达到宇宙数量级变化的可能性，只能望洋兴叹。基于决策树算法，穷举一切走子可能性的模式超出了计算机的承载能力，虽然算法不断优化，但还是无法突破计算瓶颈。以围棋为代表的东方智慧，面对人工智能似乎可以稳若泰山，但一个新时代正在来临。

Internet的大会师

“深蓝”代表的计算机智能似乎与互联网无关，然而云计算和大数据的发展，使得人工智能和互联网终于合二为一，元神合体，获得了一种不同于“深蓝”时代的智慧模式。多芯片分布式计算加上人类积累的大数据，再通过超越决策树的新算法来贯通，体现了人类智慧与机器智慧的结合。

2016—2017年，AlphaGo（阿尔法围棋）横扫人类围棋高手圈。AlphaGo的下棋“思路”不同于人类，也不同于“深蓝”。简而言之，是千万盘人类围棋对弈的数据滋养了它。如果要给出更专业的解读，则可以说是蒙特卡洛搜索算法和基于深度学习的模式识别促成了AlphaGo的成就，其中最为重要的，恰恰是其前辈“深蓝”所不具备的深度学习。

根据各方的研究来看，AlphaGo不是自己想出棋着来，而是学习了人类高手的千万盘棋局（这就是大数据）。它记录下每个棋局中的每个局面，把上百万个局面当作输入进行训练，通过一个多层神经网络来预测人类高手会走出的下一着。经过巧妙的神经网络设计与训练，这个多层神经网建模了人类高手的“棋感”——对于当前局面，已知以往下棋历史中的胜率。在实际下棋时，计算机可以通过视觉识别记录下棋局，然后和以往的棋局数据比较，找到相同的模式（局面），再检索不同局面往后发展下去，根据以往下棋史中的胜率高低选出一些高质量的候选点供走子，而不必每个候选点都去尝试一遍，从而极大地减少系统运算量，不至于让系统“殚精竭虑”而死。这就像人类，不会穷尽所有候选点，而是根据经验和感觉选择某些点。选出几个点之后，人类还是要分别计算、比较哪个点更好。对于机器来说，这个计算就要交给蒙特卡洛搜索算法。

我们用一个不一定十分准确，但是形象的比喻来说明。蒙特卡洛树形搜索是对以往决策树算法的优化。对于以往的决策树算法，即便给了一个高质量的候选点，对于接下来的选择，它同样要进行穷举，

在每个要选择的地方做一次分支，同样会遇到可选路径数量的指数爆炸。

蒙特卡洛方法就体现了概率学的精妙。假设在某个棋局局面下，深度学习网络给出了三个候选落子办法A、B、C，以这三个点为根节点，分别往下走子，可以想象成三棵树，每棵树还有无数分支。蒙特卡洛搜索不去穷尽所有分支，而是派出300万只蚂蚁分别从A、B、C出发，每个点100万只，飞速向树梢爬（也就是往下黑白棋交替走子直到决出胜负，基本上走200步就会分出胜负），总有部分蚂蚁走到最高点（也就是决出胜负，假设蚂蚁走到终点的情况代表黑子胜，没走到终点的情况代表白子胜）。

假设从A点出发的100万只蚂蚁有30万只到达终点，从B点出发的有50万只到达终点，从C点出发的有40万只到达终点，系统就认为黑子走B点胜率更高，就会选择B点。这就是概率学的取样算法，相比逐项穷举法，极大地缩减了计算量。

为什么派100万只蚂蚁而不是10万只或者1000万只？这是根据计算机的计算能力和对竞争对手的大致估计来确定的。如果派10万只蚂蚁就可以得到较高胜率，那么派10万只也可以。在相同时间内派出越多蚂蚁，对计算能力要求越高。

CPU芯片与GPU（图形处理器）芯片同时进行神经网络计算与蒙特卡洛树形搜索，模拟海量的终盘局面，这是人类计算能力无法相比的。由于采用深度学习建模了人类高手的棋感，看上去人工智能拥有了人类的大局观，而这个大局观恰恰蕴藏在人类高手的千万盘对弈数据里。

相信聪明的读者即便不太了解数学理论，也已经基本明白AlphaGo是怎么做的了，虽然具体的算法和策略要远比上面描述得复杂。AlphaGo向大众展现了一次当下人工智能/深度学习技术的发展水平。但实际上做同类研发的机构和人才很多，而且八仙过海，各显神通。

人类的行为一旦被互联网以数据的形式记录下来，就成为滋养人工智能在各行各业齐头并进，进而帮助人类自己的无穷无尽的燃料。机器翻译、语音识别、图像识别都是基于互联网提供的大量数据，用户点击行为也是如此。为什么百度搜索引擎的准确性是国内其他搜索引擎难以比拟的？因为数据量最大、算法最先进、积累最雄厚。用户的每一次点击其实都在训练搜索引擎背后的百度大脑，告诉它哪一条资讯才是用户最想要的。

当人工智能遭遇冬天的时候，人们认为机器很难像人一样思考，但机遇也正在这句话里。20世纪90年代以后，人类认识到人工智能没必要像人类那样思考，只要能够解决人类的问题即可。所以语言学家乔姆斯基在被问及“机器可以思考吗？”的时候，借用丹麦计算机科学家迪克斯特拉（Dijkstra）的说法反问：“潜艇会游泳吗？”潜艇不会像鱼儿或者人类那样摆动身姿游泳，但是它的水下能力非常好。

当我们回望历史——这个历史不仅仅是互联网的发展历史——整个人类工业发展都在孕育人工智能。凯文·凯利说过，蒸汽机活塞的自我往复运动就是一个精巧的设计，这种自我应答已经蕴含了“进化”的要素。自动化的追求是人工智能的进化动力。

比如，工业革命开始的时候，蒸汽机最先出现在煤矿和坑口。因为早期的蒸汽机效率低、能耗大，只有在煤特别多、特别便宜的地方才可能使用。采煤的时候会产生很多水，要从煤矿里面把水抽出来。有了这种需求，又有足够廉价的能源，才会想到用蒸汽机这种办法。一旦得以运用，蒸汽机的技术就不断发展，最终推动工业革命。人工智能也是一样：当你容易得到足够多的数据时，数据就是新的能源，就是燃料，有了数据燃料，人工智能的发动机就可以运转起来。

要感谢互联网的发展以及一切人类活动产生的数据记录，没有人类的数据积累，计算机就没有可学习的对象。要感谢那些人工智能探索者，他们并非都是计算机科学家，他们有的做生物学研究，有的做工程学研究，有的研究数学和计算机程序的自动迭代优化，有的改革

计算机芯片的协作架构。各种研究成果汇流成海，终于汇聚成今天的人工智能。

巨头群起逐鹿

2016年AlphaGo在媒体上引发的惊呼其实是舆论的后知后觉。回到2007年，人工智能领域巨擘Geoffrey Hinton早已察觉到“山雨欲来风满楼”。

当时他的一个学生在谷歌大数据的帮助下，将Hinton若干年前的研究成果应用于语音识别技术上，并取得了明显的成功。Hinton不由得感叹：“回顾过去，失败只是源于缺少数据量和计算能力”。

时间进入到21世纪的第二个10年，人工智能万事俱备，百舸争流的竞争时代开始了。自2015年以来，人工智能创业潮持续发酵。根据美国风投数据机构CB Insights发布的人工智能产业数据分析，人工智能投资额在2016年第一季度就超过了10亿美元，第二季度发生了121笔融资，而2011年同期只发生了21笔。2011年第二季度到2016年第二季度，人工智能方向上的投资额超过75亿美元，其中有60多亿美元都是在2014年后产生的。

《乌镇指数：全球人工智能发展报告》显示，2016年前两季度，国内新增60余家人工智能创业公司，投资金额达到6亿美元。在过去一年中，中国大陆在人工智能领域进行了202笔投资，共涉及10亿美元（约合人民币68亿元），市场规模庞大。



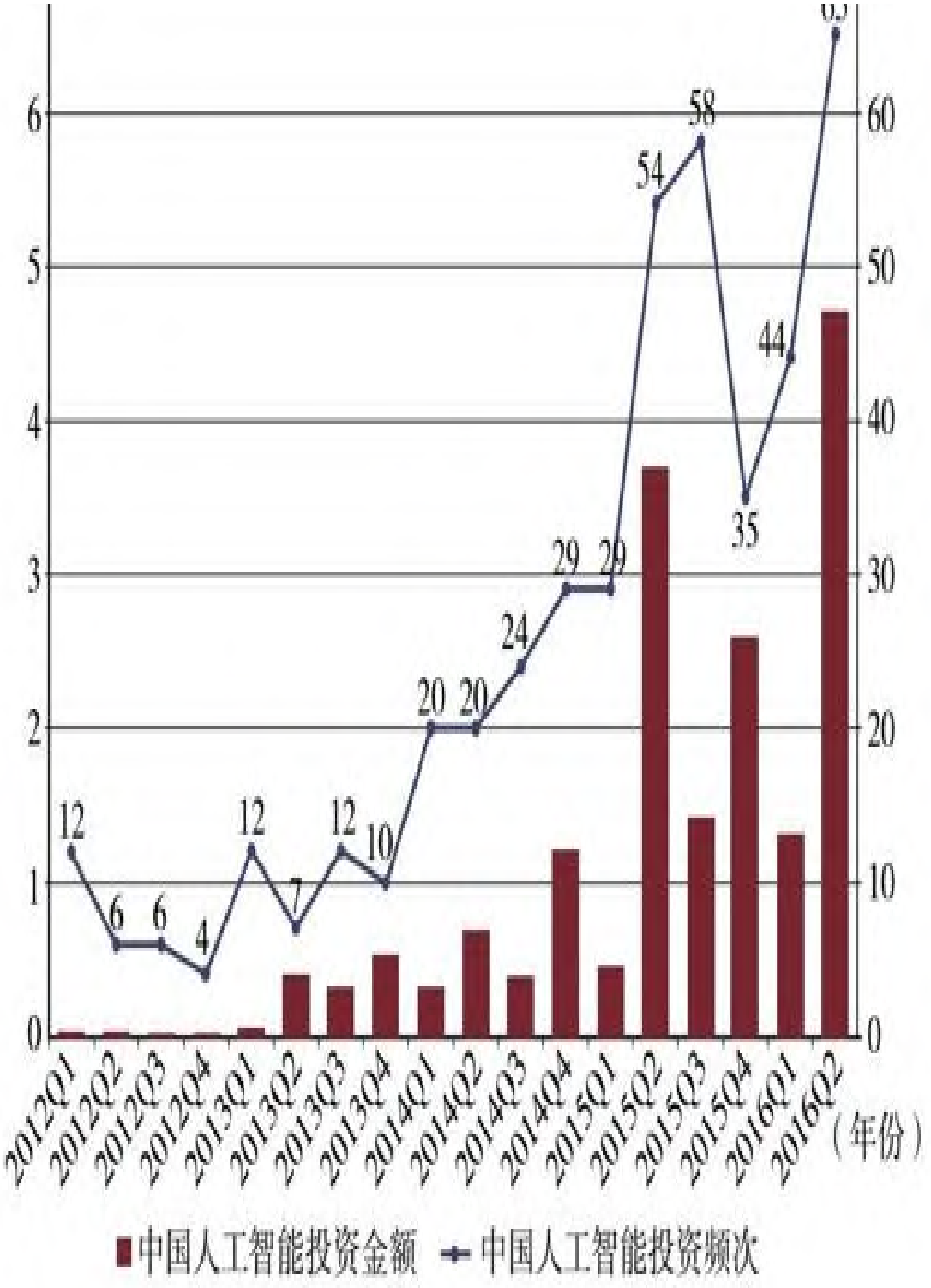


图1-3 中国人工智能投资金额和频次

资料来源：www.cbinsights.com

2016年，中科院副院长、中国人工智能学会副理事长谭铁牛院士说，2015年全球人工智能市场规模为1270亿美元，2016年当年预计达到1650亿美元，到2018年，这个数字将超过2000亿美元。

中、美、英三国是人工智能的发展重地。美国是互联网和人工智能的起源地，拥有得天独厚的人才优势，加上技术家底雄厚，科研经费庞大，使其在这个领域领先。除了谷歌、脸书、微软、亚马逊、IBM、苹果等巨头大举投入人工智能领域，还有大大小小近百家专注人工智能业务的公司。比如，擅长自然语言处理的X.AI公司，三轮融资达3.4亿美元。英国则继续老牌名校的辉煌，在制造业萎缩的背景下，将人才都聚集到人工智能领域。研发AlphaGo的DeepMind公司就是其中的代表。

亚马逊推出Alexa智能语音助手和Echo智能音箱，与苹果、谷歌、微软争夺语音入口。2016年6月，亚马逊总裁贝索斯在接受美国科技博主沃尔特·莫斯伯格（Walt Mossberg）采访时透露，亚马逊针对人工智能领域关键项目的投资已经持续了4年时间，“亚马逊从事这些项目的团队超1000人，你看见的只是冰山一角”。

2016年9月，微软宣布在执行副总裁哈里·舒姆的领导下成立新的人工智能研发事业群。他领导数千名计算机科学家和工程师将人工智能整合到该公司的产品中，包括必应搜索引擎（Bing）、小娜数字助理以及机器人项目。当年末，微软正式发布了可以开发聊天机器人的服务，并宣布将为埃隆·马斯克和创业孵化器Y Combinator的总裁山姆·奥特曼（Sam Altman）共同创办的Open AI人工智能实验室提供CPU服务。

脸书也拥有自己的人工智能实验室以及类似谷歌大脑的团队——应用机器学习事业群。这些机构的使命是在各种脸书产品中推广人工智能技术。用该公司首席技术官麦克·斯克洛普夫（Mike Schroepfer）的话说：“脸书约有1/5的工程师现在都在使用机器学习技术。”

AlphaGo的主人谷歌当然也不会只满足于下棋，其人工智能投入多年来不断膨胀。2012年，谷歌只有两个深度学习项目，2016年底这个数字突破了1000。目前谷歌从搜索、安卓系统、Gmail（免费网络邮件服务）、翻译、地图、YouTube（视频网站）甚至到无人车，都有深度学习的影子。

中国拥有庞大的业务应用场景、用户和数据以及基数最庞大的人才群体，进步很快。除了BAT（百度、阿里巴巴、腾讯三大互联网公司首字母缩写）、华为等巨头大力开发人工智能，还有很多垂直领域的人工智能公司涌现。2016年的各种互联网论坛上，不论是电商、社交媒体，还是搜索引擎，各家互联网企业的掌门人都在将话题引向人工智能，汇报着或大或小的成绩。

2016年，百度语音识别准确率达到97%、人脸识别准确率达到99.7%。作为百度大脑的云化，百度天算、天像、天工和天智平台相继向全社会全面开放百度大脑的技术和能力。

超强大脑汇聚

十多年前机器学习领域的少数坚持者，现在成了最宝贵的人才。人工智能浪潮兴起之后，在这个开源的世界上，除了数据，最稀缺的资源就是人才了。

人工智能背后的专业知识与数学、生物学等基础学科有极大相关性。人工智能科学家又是这些领域的翘楚，更显难得。但全国人工智能研究方向的博士生、研究生每年只有不到200人，而如今的创业公司多如牛毛，这个数字根本不够分。国内如此，国外亦然。2015年，优步直接挖走了卡内基梅隆大学国家机器人研究所140名研究人员中的40人，引起业界一片哗然。

以上还不是人才竞争的全部。从业者更为敏感的是学术领军人物的流向。最近两年，不少“学术明星”走出象牙塔，或跳槽，或创业，让人更真切地感受到风向的变化。激流涌动，宝贵的人才去哪里才能充分发挥自己的价值，这是个问题。

百度是中国人工智能产业的代表，大量顶级人才先后投身百度：王海峰加入百度前曾就职于微软，吴恩达从美国来到百度，张亚勤从微软来到百度，林元庆从盛产机器学习专家的NEC美国实验室来到百度、机器人小冰的创造者景鲲从微软来到百度、美国科技巨头中职位最高的华人高管和人工智能技术权威陆奇放弃微软副总裁职位加盟百度.....同时，也有很多人才从百度出发，创造了自己的人工智能应用公司。百度本身就是中国在吸引和培养人工智能人才方面活力的缩影。

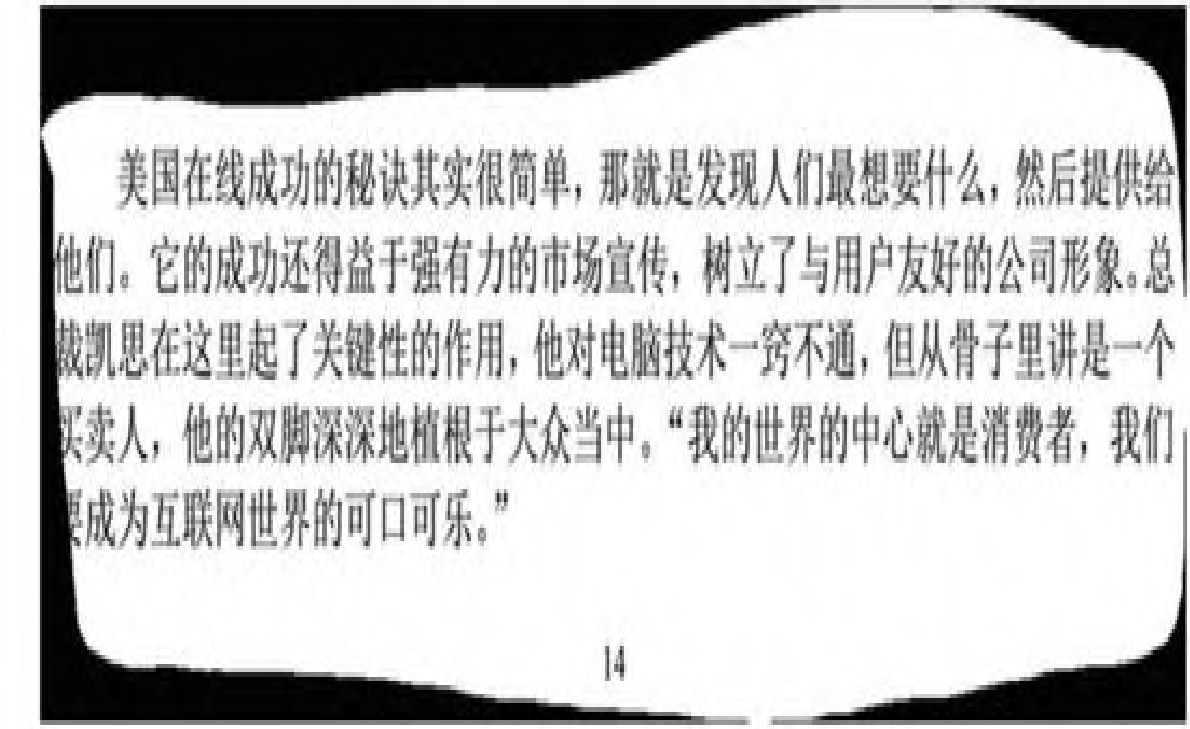
这么多人类的超强大脑汇聚，目的是要创造划时代的中国大脑。我们经历了PC时代，正处于移动互联网时代，即将迈入一个万物互联的超级智能时代。万物数据汇合，人类加以处理，就可能产生一种“超级大脑”的生态。百度正在打造这样的生态，宗旨在于像提供水和电一样让人工智能渗入中国人乃至所有人的生活中去，努力促使世界万物向“知化”[2]的方向发展。比如，百度大脑已经初步具有了自己的眼睛、耳朵、嘴巴和认知决策能力，总体来看，相当于一个儿童，但局部能力如翻译、语音识别、图像识别能力则大大超过人类。我们把这些能力开放给大家，供人们开发探索各种人工智能应用。百度大脑已经成为很多开发者的工具和人工智能的操作系统，促进了人工智能标准化的形成。这将全方位服务人工智能时代的企业、创业者和广大个人用户。

因此我们热切呼唤中国大脑，呼唤国家总体层面的深度学习服务器、算法、应用基础设施平台。中国大脑的形成，将是中国竞争力全方位升级的体现和中华复兴的强力加速器。

技术要做人类生命的延伸

说到人类的数据滋养人工智能，我想先来谈一谈我们的用户，谈一谈无数支持百度以及高科技互联网发展的消费者。

在今天，除了谷歌、微软和BAT这样的大公司，互联网和大数据技术的“去中心化”发展趋向，使得小企业、有才华的技术人员甚至广大用户都成为具有决定格局态势的力量。



美国在线成功的秘诀其实很简单，那就是发现人们最想要什么，然后提供给他们。它的成功还得益于强有力的市场宣传，树立了与用户友好的公司形象。总裁凯思在这里起了关键性的作用，他对电脑技术一窍不通，但从骨子里讲是一个买卖人，他的双脚深深地植根于大众当中。“我的世界的中心就是消费者，我们要成为互联网世界的可口可乐。”

14

图1-4 《硅谷商战》一页

我在《硅谷商战》里就强调了用户的重要性。在我们工程师的眼中，用户是一个严格定义的理性存在，用户需求—开发—反馈，是技术文档里的严密描述。但是互联网的发展不仅提供了技术服务上的方

便，也提供了思想和情绪发挥的舞台。我们可以说，互联网创造了一种意见型用户。

我们的很多程序员、工程师都很享受百度对技术工作者的宽松环境，简单可依赖。技术员想法单纯，不善交际，醉心于开发出五花八门的产品。具有各种情绪的、活生生的用户和我们工程师的习惯思维不太一样。普通人生活中的点点滴滴以及商场上各种复杂多变的交易和情绪可能是我们实验室里工程师感受不到的。媒体人、公关人更能理解用户的情绪，我们的公关部门有时也会吐槽技术人员不理解用户心理，遇到问题常常以为修改了代码bug（漏洞）就好了。但是人情bug不是代码能够修补的，这对我们是一个触动。技术员与商人和普通用户的隔阂该如何打破，是我们必须考虑的问题，需要我们有更高的产品思想和跨界学习的谦虚心态。

我们对日常用户生活需求和人性的思考是一个持续不断的工作，需要持之以恒。但就这本书的主题来说，我们毕竟是工程师，我们始终不忘考虑如何用技术和数字满足用户的需求。我们要利用技术对数据进行精准区分并服务于不同用户。

数字化，这是从尼葛洛庞帝的《数字化生存》到凯文·凯利的《失控》和《技术想要什么》一直在讨论的趋势，也是技术人才念兹在兹的事情。除了工商、金融、农业、军事、科技数据以外，包裹我们的是生活数据。说到数据总会引起一些警惕，比如隐私数据会不会被贩卖？这个话题我们后面还要说，这里简单来说，人工智能眼中的数据绝不是低学历数据贩子眼中的身份证、密码等数据。今日人工智能重在从混沌数据里发现总体“模式”，进而优化生产、服务。翻译、语音识别和图像识别的进步就是最好的例子。这些混沌数据，经过人工智能识别规律，就会对人类产生巨大价值，比如从日常领域的语音识别到金融领域的征信防骗再到国家层面的反恐安全。

技术再好也要适应用户。产品端直接响应用户的需求，需要不断优化技术的表现形式。我们认为，好的人工智能要润物细无声，不能

像电压不稳定的电源，不能像有污染的水。要不断提高准确率，优化产品细节。比如有的公司语音识别技术虽然不错，但是输入法整体设计不够方便，这就影响了用户体验。百度也有不成功的产品例子，需要和用户一起来改变。

数据和技术不是冰冷的存在，和好的人工智能方法结合，就会体现出人性的一面。

很多网民对图1-5印象深刻。这是2014年初在东莞“扫黄”之后，百度地图运用数据可视化技术，描绘出东莞与中国各地的迁徙线路热度。

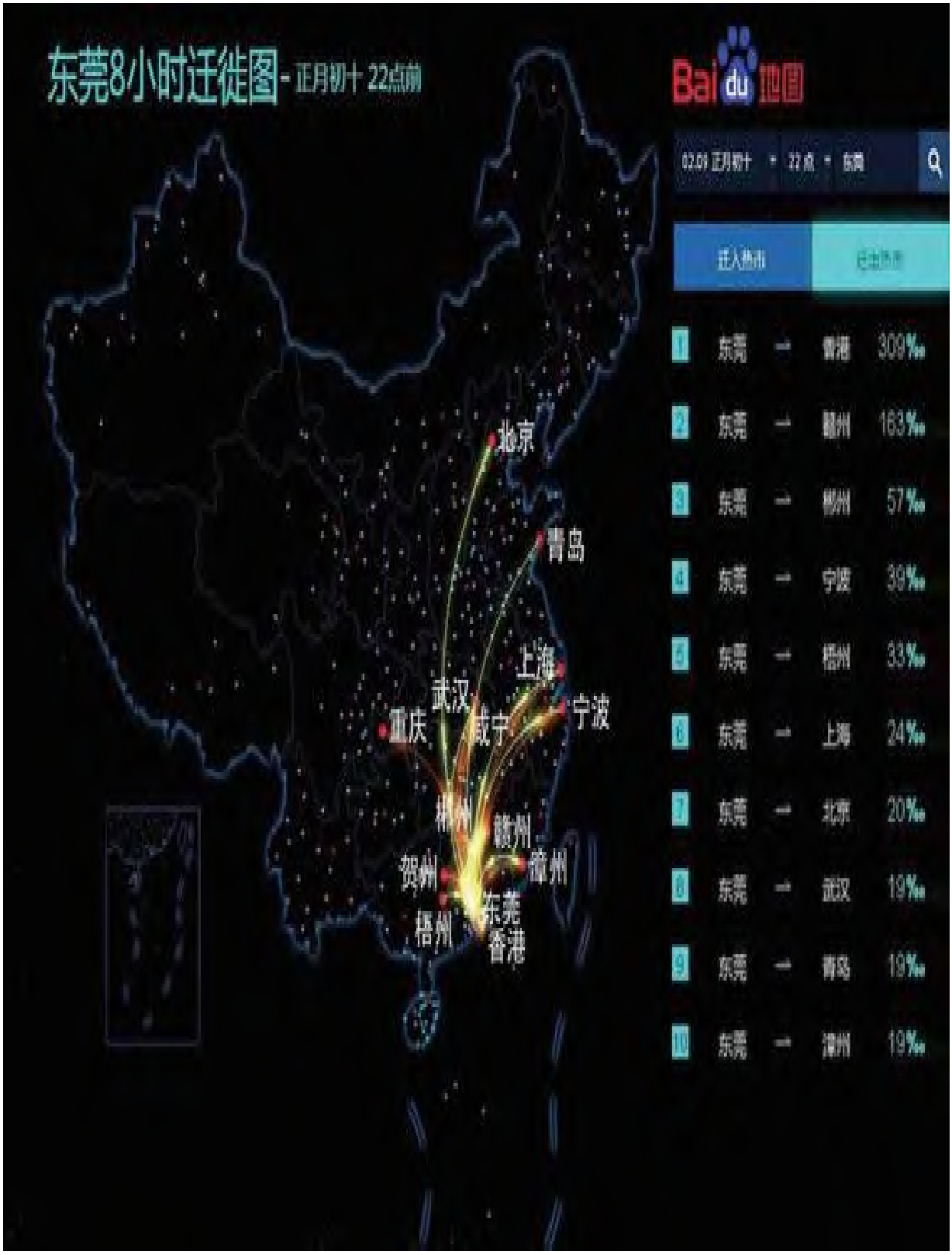


图1-5 东莞8小时迁徙图（正月初十22点前）

注：使用手机百度或智能革命App扫描图片可见AR效果。

有资深新闻编辑告诉我们，当时看到百度这幅图，瞬间感觉超越了新闻事件本身，有一种俯瞰人间的领悟。百度迁徙指数通过数据可视化技术反映了人类的命运和迁徙。数字时代的人群迁徙只是百万年来人类大迁移史诗中很小的一页，却是大数据时代具有历史性的一页。

我要说，这也是人工智能时代具有历史意义的时刻。这是智能地图技术对人类活动、人类命运的感知。人工智能本身尚无人性，但是结合开发者的创意、理念，就可以提供新的视角，甚至是一种别样的人性关怀。

计算机和互联网都是人工智能的身体，每个数据都是人类活动和人性的记录，人工智能因此终于像“灵魂”一样涌现而出，它是可以有人性的。

数据大道

有位哲学家说，人类是一种“在路上”的存在。百度积累了海量的地图数据，辅以设计师的智慧和各种精巧算法，可以描绘出人类的各种移动行为，感知人们在路上的生存状态。

百度地图每日位置服务次数最高突破720亿次，每一次都是人类的活动记录。图1-6是北京中关村一天的通勤记录的数据化展示，奔涌的交通热力图和节奏，仿佛这座城市的生命脉搏。



图1-6 北京中关村一天的通勤记录
注：使用手机百度或智能革命App扫描图片可见AR效果。

地图之眼具有大视野，通过图1-7，可以感受到中国中部、西部地区的发展态势——西安、郑州、武汉、合肥与北京、上海、深圳等一线城市的联系越来越紧密。



图1-7 城市间交通热力排行图

我们这一代人都听过童安格的歌：“为了生活，人们四处奔波，却在命运中交错。”我希望在人工智能的帮助下，人类的轨迹不只是交错，还有交集，汇流成河，生生不息。

百度大数据实验室一位年轻的科学家是学生物学出身，之前在普林斯顿研究鱼群的运动规律，一看到百度迁徙图就决定回国了。他

说，原来人的数据也可以像鱼群这样研究，而且更方便，所以他决定来百度工作。2016年，他和同事一起利用百度地图上的搜索数据变化状况，准确预测了iPhone（苹果手机）销售量的下降。通过数据，大数据实验室为各种城市生活、企业运营提供智能感知。

2014年交通运输部提出：要深化改革，务实创新，加快发展“四个交通”，加快建设市场导向、企业主体、产学研结合的行业技术创新体系，促进科技成果转化为交通运输生产力。着力建立多渠道、多方式的交通运输出行信息服务体系，初步建成综合交通出行信息服务平台，向社会即时发布出行信息，解决出行信息不畅等问题。

在此背景下，百度提出了“中国智慧交通云服务平台合作计划”，与交通运输部公路科学研究院、国家智能交通系统工程技术研究中心共同打造合作平台，依托交通运输部重点科技项目“基于云平台的开放式公共出行信息服务研究与示范”，激活现有数据，建立部省数据信息资源共享交换机制，促进政企间出行服务信息共享应用，并对全社会开放。

智能地图可以根据用户移动快慢测量道路的拥堵程度，还可以智能规避单双号限行路线。结合虚拟现实技术，可以让人身临其境般寻找路线。以交通大数据为基础，加上算法辅助，响应交通管理部门的需求，智能地图系统已经能为城市交通缓解提供解决方案，大大减少了交管部门的压力。

智能地图对地理数据的收集使得很多智能项目得以展开。达到厘米级精度水平的高清地图技术已经运用到无人车开发中。2016年世界互联网大会上，百度无人车在乌镇进行了公开测试和试运营。体验全程3.16公里，共经过3个红绿灯以及一次掉头，不仅面临人车混行、电瓶车穿行等多种复杂路况，还需应对中雨、薄雾、雾霾等多种天气状况。这个结果丝毫不落后于硅谷同行在北美进行路测的进度。这是无人车的一小步，但必将成就人工智能的一大步。

人工智能不是从天上掉下来的，正是从几十年来计算机网络技术和数据处理技术的进步以及人类的数据生活之中水到渠成的。百度搜索和百度地图的智能化发展就是这个过程的一个缩影。

人工智能既不是神话也不是笑话

如今各种大众媒体上关于机器人的新闻很多，凑热闹的也很多。比如前阵子有新闻报道，一个展会上有一台机器人伤人。其实那只是一个教育辅助机器人，掉下台砸到人而已。还有某个墓园买机器人给守墓保安壮胆的新闻，那个机器人充其量是个玩具而已，恶搞成分更多。如果我们以科普的心态看待历史，就会发现人工智能既不是神话也不是笑话，而是从人类的劳动创造中实实在在生长出来的。我们不需要恐惧也不需要顶礼膜拜。

人工智能领域的科学家对技术的描述常常是直白而谦虚的。谷歌前工程师吴军说过，他在2003年的时候和同伴一起把谷歌的关键词搜索准确率大大提升，解决的一个主要问题就是，对于同义词或者近义词究竟该选取哪一种意思去搜索才能满足用户的需求。对于用户来说，如果搜索给出的结果不准确，用户就会换个近义词继续搜索，或者选择搜索结果中排名并不靠前的结果。这时用户其实亲自做了一个关键词搭配工作，系统会记录下用户给出的关键词搭配关系，现在要做的是更快、更优地反馈结果。他说：“至于我们是怎么做到的，说起来可能会显得很没有技术含量，我们事先把多年来用户搜索过的关键词搭配都整理出来，然后在2003年美国独立日的长周末期间，有四天的假期，我们停掉了公司当时五个最大的数据中心中的一个，利用四天时间对每一个关键词的搭配做了特殊处理。这实际上就是一种穷

举法。”[3]也就是把用户经常选择的词语组合关系固化下来，下一次用户再做类似搜索，系统就能更快、更准地给出结果。

其实机器翻译等领域的技术逻辑，跟上述搜索中应用的有策略的穷举法有异曲同工之妙。据《纽约时报》报道，2016年6月的一个星期三，在谷歌翻译部门的会议上，人们对百度发表在机器翻译领域核心期刊上的一篇文章议论纷纷。迈克·舒斯特（Mike Schuster）的一句话让会议室恢复了秩序，“是的，百度出了一篇新论文。感觉就像有人看透了我们做的东西——论文有类似的结构，类似的结果。”百度公司的BLEU分数（一种衡量机器翻译和纯人工翻译之间准确度的分数）基本吻合谷歌在2月和3月内部测试中取得的成绩。Quoc V. Le(夸克·维·乐）并未感到不快。他的结论是，这是一个表明谷歌处于正确轨道上的迹象。“这个系统与我们的系统非常相似。”他安静地说。

Quoc V. Le是吴恩达的博士研究生，他可能并不知道，这篇论文的出炉与吴恩达并无关系，而是自然语言部门独立完成的。《纽约时报》对中国企业的报道当然是一扫而过。但吴恩达认为，国内部分媒体也需要改变习惯，不能总是下意识地认为什么技术都是外国更强，热衷于把后知后觉者当作突破者来报道，实际上人工智能领域很多领先创造都是中国人先做到的。

百度领先一年发布了基于NMT的翻译系统，谷歌在2016年也紧跟百度推出了类似的系统。所以这个领域最前沿的探索者的基本技术都差不多，最后就看谁的积累深厚，谁的优化做得好。

今天的人工智能思路与过去不同，变思维规则问题为数据问题和策略问题。过去人类总想为计算机设计出完美的逻辑，不断把人类的种种逻辑规则抽象成函数之后输入计算机里。现在的人工智能主要基于大数据基础和算法的进步。也就是说，今天人工智能的爆发恰恰建立在20世纪90年代末互联网爆发的基础上。有了互联网，数据才会大量产生。注意，这些数据不是用户自觉填写的数据，比如姓名、年

龄、住址、爱好等，而是用户在使用互联网时自动产生的数据，比如每一次搜索、每一次点击就是一种数据，每一次移动轨迹也是一种数据。

中国已经是世界头号制造业大国，现在更需要的是提升“软实力”。精神、文化是“软实力”，计算和数据也是“软实力”。这样的“软实力”和传统产业叠加，也就是所谓的“智能+”。它将切切实实融入我们的生产、生活，看得见、摸得着。

非如此不可

问百度要做什么，不如问为什么一定要做？！

每家企业都有自己的战略战术。2013年，国内移动互联网创业风潮开始兴起，不少企业将巨额资金投入这个巨大的“无底洞”，体现了它们在战略上的勇猛。百度则着眼于战略的长远和科学。当时注意到百度全面发力人工智能的人并不多。今天，人工智能蜚声世界，有人感叹百度战略的超前和坚定。因为百度提前认识到互联网信息产业的本质，一旦下定决心，就坚决走自己的路，不在乎外人评判。为此百度多方布局，重点突破，当全世界都开始关注人工智能的时候，人工智能大格局上已经有中国百度立下的柱石。

我们没有让百度的人工智能去参加下围棋、预测歌手比赛结果这类活动，而是专注于发展内功，同时集中力量把人工智能转化为能够改善人类生活的实用性服务。我们不只把深度学习应用到语音识别、机器翻译和街景门牌号识别等少数领域，而是将深度学习成功应用于显著提升用户体验。

2013年，百度导航率先宣布永久免费，把中国带入导航免费时代。现在，我们把百度地图的数据接口开放出来，供人们开发使用。用户可以使用百度地图提供的定位技术和方案，相比传统GPS Tracker（全球定位系统追踪器）节省了大量成本。快递公司可以以此规划最优送货路线，游戏开发者可以开发类似PokermanGo（口袋妖怪）的位置游戏。我们开放百度大脑，让更多的人可以使用人工智能的眼睛、耳朵为自己服务。我们开放深度学习开发平台PaddlePaddle，让更多有志者可以创造属于自己的人工智能服务。我们也希望让非技术人员学会用数据智能优化自己的工作、完善自己的个性、追求自己的理想。

有很多高考考生想必已经使用过度秘机器人帮助自己选择高考志愿。在中国，任何事情都有很多人去做。在我读书的年代，人们把高考称作“千军万马过独木桥”。和地图数据类似，度秘机器人通过对千军万马的高考数据进行分析，通过深度学习技术响应并感知高考考生的渴望、焦虑，尽力给出自己的精准回应。在这里，人工智能记录的不是物理空间里的地图轨迹，而是学子成长的心灵轨迹。

20世纪90年代初我远赴美国学习计算机技术，当时有很多和我一样的年轻人，怀着用代码改变世界的愿望，候鸟般往来于中美。如果当时有一张数据地图记录这些跨洋轨迹，那会是很有意思的事情。如今人工智能科学家再次把火种带到中国，我相信这次火焰会燃烧得更热烈，因为中国大地有足够的燃料。中国受教育人口数量巨大，计算机和移动设备普及极快，大量的数据使得中国在发展和应用深度学习技术方面有得天独厚的优势。有了这样的优势，我们可以厉兵秣马创造20世纪90年代硅谷那样的传奇。

百度要做的事情不仅是前沿开发，而且要为用户提供数据基础设施，提供深度学习开发平台，网聚人的智慧。

在特朗普当选美国总统之前，有百余位硅谷精英发表公开信，认为特朗普当选将会是创新的灾难。这对我是一个触动，如果美国的创

新真的受到影响，谁来接过旗帜引领创新方向？我们能把世界的创新中心从硅谷迎到中国吗？

人才确实在向我们涌来。百度还在硅谷成立了实验室，近身接触美国人才。百度提出的中国大脑计划堪比任何超级工程。回想70年前，顶尖科学家壮志满怀地从国外回到中国兴建伟大工程，这样的辉煌今天会再次出现吗？



图1-8 李彦宏在百度贴吧讨论

资料来源：http://tieba.baidu.com/p/4855363507

当然，必须注意到，那个时代的伟大工程往往依赖国家投资和产业政策。“冷战”结束后，国家竞争压力减小，尖端科技的投资也大为减少。马斯克去开发火箭，其实是国家把NASA（美国国家航空航天局）的火箭技术和团队转移给他。在中国，政府层面的决心和投入依然强大，在发展人工智能产业方面可谓上下同心。这是最好的时代，也是最不确定的时代。人工智能就是一种适应不确定性的方法。大大小小的公司投入人工智能研发，带来竞争和多元化，这应该形成良性的互动和生长。

人工智能的发展也将带来不确定性。美国白宫的报告已经在探讨人工智能对就业的冲击。美国硅谷的高速发展和中部制造业的衰落加大了国家的裂痕，一部分人享有进步成果，另一部分人被扔出历史航船而失去方向。百度要成为人才的方舟，中国企业则要努力打造人类的方舟集阵，让最广大的人群踏上智能时代的历史巨轮。

百度副总裁王海峰博士于2016年11月刚刚当选ACL（Association for Computational Linguistics，国际计算语言学会）会士，成为ACL目前最年轻的会士，他也是该组织五十多年历史上首个出任主席的华人。评选委员会在给王海峰的评语中写道：“王海峰在机器翻译、自然语言处理和搜索引擎技术领域，在学术界和工业界都取得了杰出成就，对于ACL在亚洲的发展也做出了卓越贡献。”2017年初，在人工智能领域享有盛誉的科学家和高管陆奇加盟百度。这些都在预示着国际人才流动的趋势。中国千百位优秀的人工智能科学家要一起创造人类的未来。

未来已来：焦虑与梦想

不久前，亚马逊的“无收银员超市”引起“血拼族”的惊叹。这种特别的购物体验背后，则是“收银员”下岗的阴影。今天，当各种在线客服被机器客服代替，当速记翻译被语音识别代替，甚至当收银员、驾驶员、工厂工人、普通文书和律师都被人工智能代替时，人们该怎样迎接这个世界？政府和企业该为劳动者做怎样的支持？我们该怎么调整经济、社会生态结构，以适应人工智能时代？我们希望聆听普通人的需求。这也是我们人工智能团队共同打造本书的初衷。

硅谷有位和马克·安德森齐名的风投鬼才彼得·蒂尔。他是PayPal（贝宝）的创始人，善于把握技术大势，捕捉黑马。2016年他因为准确预测特朗普当选美国总统而再次名声大噪。他在2011年时说过：We wanted flying cars，instead we got 140 characters（我们需要能飞的汽车，结果只得到140个字符）。140个字符的推特（Twitter）一度热闹无比，但彼得·蒂尔清楚地看到互联网喧嚣背后缺少什么。他批评人类放慢了进步速度，嬉皮文化代替了进步主义，风投热衷于投资轻资产企业，其中大部分是移动互联网公司，如Airbnb（空中食宿）、优步之类，却对未来没有清晰的规划和信心。他认为“互联网+”时代人类在比特层面进步大，在原子层面进步小。因此他果决地投资火箭、抗癌药物以及人工智能。

我同样认为移动互联网创业的喧嚣掩盖了我们所要真正追求的进步。百度要为自己的方向而奋斗，要为人类的核心能力进步做贡献。蒂尔说20世纪初的美国人愿意尝试新事物，敢于规划几十年周期的登月计划并去实现。然而现在人类没有这样的计划了，只有风投在到处寻找眼前的增值和及时的痛快。百度愿意幻想一个智能化的世界并去实现它，要让人工智能成为新的操作系统，不仅是计算机的而且是世界的操作系统，同时严肃思考和提前应对人工智能的挑战，最终让这个世界从此不同。所以我说一定要把这件事情办成！

智能革命是对生产、生活方式的良性革命，也是对我们思维方式的革命。巨大的机遇与挑战并存。以下我们将具体探讨智能革命的方

方面面，详谈视觉识别、语音识别、自然语言处理等在深度学习基础上取得的突破性进展，并从制造业升级、无人驾驶、金融革新、管理革命、智能生活等多个维度，描绘即将到来的智能社会，进一步探讨人类应该如何应对人工智能的发展，与读者一起把握智能革命的脉搏。

---

[1] 用来表示关于随机事件A和B的条件概率，其中P（A ｜ B）是在B发生情况下A发生的可能性。

[2] “知化”是凯文·凯利在《必然》中提出的一个观点，指软件吞噬一切，一切事物都将信息化，哪怕是一张桌子也可以上传自己的数据，如销售轨迹、使用频率等。

[3] 吴军。智能时代［M］.北京：中信出版社，2016：179.

02

人工智能的历史使命：

让人类知道更多，做到更多，体验更多

陆氏猜想

20世纪80年代是所有亲历者难以忘怀的理想主义年代，那时候中国全体人民追求科学进步，对知识如饥似渴。尤其在校园里，大学生读书如痴如狂。陈景润[1]一类的科学家成为很多人的偶像。

1987年，在复旦大学的校园里，有一个和陈景润一样瘦削、戴着大框眼镜的青年，每天背着一个很大的书包行走在校园里，精神抖擞，喜欢思考在外人看来很玄奥的问题，同学们都喊他“陈景润”。

毕业的时候，他在毕业纪念册上写下了临别赠言：

谨献上本人最新研究成果，与列位同窗惜别。

“陆氏猜想”：HI=>C∪HB

（式中：H: Human I: Intellectualized

C: Computer B: Brain）

其意：人类终将使电脑智能化且使其远胜人脑。诸公不妨一效景润，或许，这颗电脑科学皇冠上的明珠非君莫属。

他是陆奇，后来的雅虎、微软领导者，今天的百度集团总裁。

“陆氏猜想”当然不是一个“陈式猜想”那样的科学定理，却也不是一句戏言，今天来看，更像是一句超前的顿悟。30年前就写下这样的话，信心来自哪里？

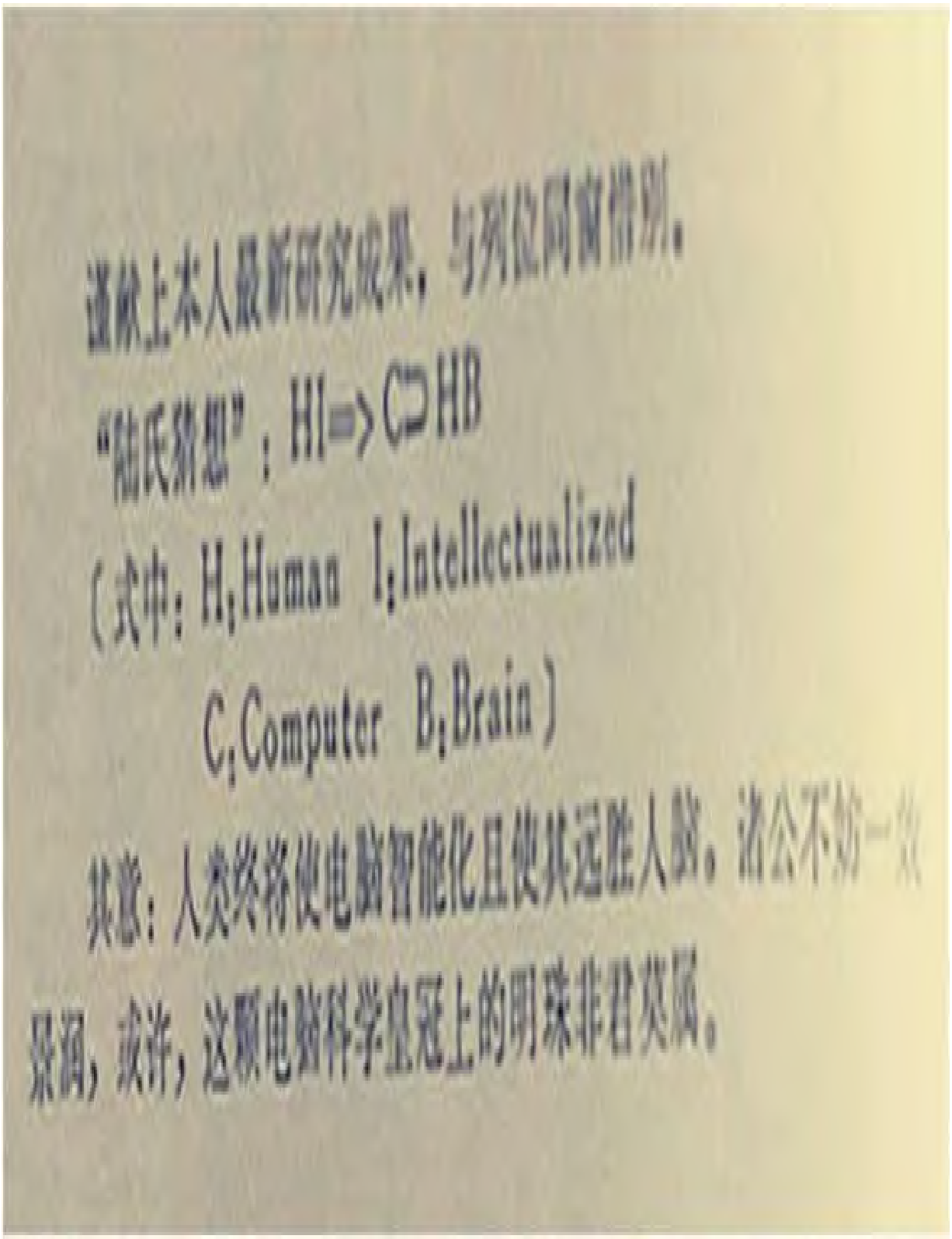


图2-1 陆奇于复旦大学1984届计算机科学系毕业纪念册上的临别赠言

那时的陆奇只有一个朦胧的感觉：“计算机给我们带来非凡的知识和体验。当时我们在计算机系写下棋程序，虽然是很简单的棋，给我的直觉就是，只要给我们足够的时间，以后一定可以做得比人更聪明，我当时有这样的直觉，所以我就这样写了。之后不久，我遇到一个难得的机会去卡内基梅隆大学计算机系深造。”

陆奇的个人信念是：知道更多，做到更多，体验更多。

从卡内基梅隆大学毕业时他写的是“Know more，Do more”。后来加了一句，变成“Know more，Do more，Be more”。因为他觉得“Be more”更重要，只是当时没有意识到。“Be more”也可以翻译成“成为更多”。人就是“在路上”，不断成为更丰富的存在。

就是这三个需求有机联系在一起，推动人类向前进步。人类的历史就是不断地去发现新东西，我们知道越多，就可以做到越多，我们做到越多，就可以体验越多，生活就会越来越丰富充实。正因为体验更多，我们才会知道更多，这是一个正循环，是人类进步的主旋律。

计算机让我们“Know more，Do more，Be more”，而人工智能就是这个旋律的最新回响。从这个角度去思考，我们可以比较清楚地洞察人工智能从哪里来，往哪里去，它的本质和标准究竟是什么，有志于人工智能事业的人和组织到底应该做什么。

下一个浪潮

我们正在见证的是一个计算机和数字化崛起的时代，这是人类历史大潮中持久的、必经的一个过程。而人工智能是将大潮推向下一个高点的动力。它将开辟一个新时代，给我们的社会带来持久的、长远的革命性影响。这样的影响涉及产业、技术等经济、社会各个层面。

但是说到底，这一次人工智能革命将让我们人类整体用完全不同的方式往前走，书写崭新的历史。

首先，我们需要了解“人类进步”的本质——我们为那些能让我们认知更多，实现更多，获得更多经验的事情充满热情地奋斗。

而计算能力的不断发展，正是遵循着上述人类进步的方向，成为人类进步的本质表现。尤其是在计算机编程出现后，人类的进步开始前所未有地加速。其核心模式包括以下关键步骤：人类捕捉宇宙中的各种现象，特别是通过有意地观察获得经验；然后通过计算，将信息有效地组织、处理、提炼，使人类对某个现象进行更深入和抽象的理解，形成知识；人类利用产生的知识认知来采取行动，与现象进行交互，最终实现我们预期的结果。

可以看出，以现代数字计算系统为基础，IT（信息技术）行业在创造数以万亿元计的价值时，正是从组织信息（Information Organization，IO，帮助人类认知更多）、完成任务（Task Completion，TC，帮助人类实现更多）、丰富经验（Experience Enrichment，EE，帮助人类获得更多经验）三个核心维度上使人类取得长足的进步。

而人工智能作为人类计算能力的又一次全新升级，仍然是从上述三个纬度推动人类进步。而且，由于人工智能是一种革命性的、更高级别的智能计算系统（Intelligent Computing System，ICS），它对人类进步的推动作用也是前所未有的、革命性的。

决定现代数字计算系统主要结构的是资源的组织形式。而人工智能计算的本质，简单来说，非常不同于冯·诺依曼的控制流结构，后者采用线性的记忆体和布尔函数作为基线计算操作。而新的范式是神经网络计算，其特征在于分布式的表示和激活模式。在这里，变量由叠加在共享物理资源（如神经元）上的向量表示，并且通过神经元的激活来进行计算。网络的拓扑架构和激活模式提供了巨大的计算空间，可以有效并且自然地捕获丰富的知识（通过拓扑的超参数、权

重、激活函数）。相对于冯·诺依曼架构中的本地化表示（其中变量由诸如寄存器的专用或局部化物理资源表示）和符号计算，神经网络计算在学习和表示物理世界以及社会的丰富的语义知识方面更加自然和强大。

通过神经网络计算的力量，下一波的人工智能技术可以在以下两个维度提升目前的计算系统：

一是自动分层特征/表示学习。这是机器学习容量的实质性提升，因为当今机器学习工作的很大一部分关键在于特征工程。如百度大脑已经拥有万亿级的参数、千亿级的样本和千亿级的特征训练。

二是高级认知，特别是感知能力。这是下一代设备（如无人驾驶汽车）和下一代平台（如自然语言会话）产生的巨大催化剂。

人工智能计算的强大能力将有助于产生许多新品种的智能系统，如机器律师、机器分析师、医疗机器人、智能客服人员等。

人工智能计算的另一个发展方向是组织各种服务于特定物理架构和物理要素的系统，如家、办公室、工厂等的智能系统。其基本模式是通过使用物联网传感器的各种原始信号，人工智能的“感知系统”会对物理架构进行识别和感知；而“认知系统”需要组织信息和学习更多关于物理架构的知识，并去预测、判断和决策，以使各类物理系统更加智能。

目前，在科研领域，人工智能计算可以提供更先进的建模能力，成为多领域和新一波科研浪潮的催化剂。

在商业方面，人工智能可以提供额外的机会，为企业组织创建集成的业务计算系统（Business Computing System，BCS）平台。如记录业务对象（如系统设计模型、交易记录）和业务流程［如ERP（企业资源计划）、CRM（客户关系管理）］；或者系统设计并模仿人类工作活动，如沟通、协作、阅读、写作、寻求信息等。

目前来说，人工智能的“感知系统”有更广泛、更新的商业机会：一方面，可以构建和部署更多的“传感系统”的子系统，针对的是物理环境或物理系统，如装配线、工厂等。这使得未来人力密集的制造业、商业服务业等，可以采用更先进的信息工具和更强的自动化。另一方面，自然语言处理技术的迅速进步使得我们可以扫描和分析文本文档和信息，并从中提取各种高价值的业务知识，而构建和部署专用的“文本理解子系统”可以得到很多高价值的知识和商业回报。

人工智能“认知系统”的成熟代表了智能时代更长远的未来，所有的行业、职业、社会系统、生活方式都将被重塑。如果数字化社会可以概括为“信息就在指尖”，那么，人工智能时代的本质可以概括为“知识无处不在，任何交互都是智能的。”

这个浪潮对大多数人来说无疑是巨大的机遇。

对商业机构来说：好消息是它们将有许多提升、转化和进入新的增长领域的机会；坏消息是当大浪潮席卷时，没有谁可以岿然不动，如果你没有抓住前进的机会，你就会落后，甚至被时代抛弃。

对企业家来说，仅看一看“巨大市场机遇”的列表，就可以发现许多创造未来商业巨头的机会，彰显企业家精神的智能创业的历史机遇就在眼前。同时，属于智能时代的新的企业领袖也将由此诞生。

对投资人来说，睁大眼睛去发现一个属于智能时代的种子企业，支持它成长为未来的企业帝国，并获取巨额回报的机遇真的来了。

对于国家或政府来说，每一次巨大的技术革命，总是伴随着国家命运的兴衰交替。总有一些国家或政府能抓住历史赐予的机会，一跃而上，赢得一个相当长时期的昌盛国运。要想抓住这样的历史机遇，国家或政府层面决策、政策的前瞻，战略投入的坚决，路线执行的科学有效，都至关重要。

当然，当越来越多的人感知到了人工智能的浪潮，越来越多的创业者投身于这个浪潮时，也难免会有浪潮将起阶段的困惑。

在硅谷，大家都说以后要投资的都是AI+X。比如有一家专投人工智能的投资公司，纽约市前市长布隆伯格是其合伙人（LP），看了几百家公司，看得有点晕了，因为所有的创业公司都在宣称自己是人工智能的公司。对这家投资公司来说，首要的问题就是，如何来进行鉴别，以便有针对性地投资。这促使我们开始深入思考，人工智能企业的标准是什么？哪些是真正的人工智能公司，哪些不是。

衡量人工智能的现实标准

人类历史上，每一项新技术的出现，必然会伴随着各种探讨、反思甚至针锋相对的争论。面对人工智能这一带有全局性、革命性的技术浪潮的逐步兴起，人们的兴奋、质疑以及顾虑就更加多元化。其中有的比较感性，比如讨论人工智能会不会代替人类。这其实是在用人工智能与自然智能（Natural Intelligence）进行比较。

对于自然智能，有各种各样的研究和著作，包括有人提出人脑内部的机理其实是量子计算。对于人工智能，目前还没有一个大家都能够接受的定义。在这个阶段，没有必要去过分追求哪个标准定义比较正确。我们不妨务实地来讨论：现在的技术能允许我们做出什么样的智能体系。

有两种类型的计算系统被人类叫作人工智能。

第一种，本质上相当于“智能计算系统”（ICS）的子系统框架。它将数据作为输入，从数据中提取信息并建立模型，将我们关心的某些现象转化为知识。我们称这种类型的人工智能系统为“通用智能系

统”（General AI），并定义“通用智能”的意思是一台机器获取知识和实现目标的能力。

第二种，指具有类似人类的认知能力，能感知（“看”“听”“感觉”），它们可以越来越多地推理和计划，它们可以用感觉运动控制移动。我们称这种类型的人工智能系统为“认知智能系统”（Cognitive AI），是具有感知、推理、规划和感觉运动控制能力的机器。

人工智能系统的另一个二分法是“狭义AI”（Narrow AI）和“广义AI”（Strong AI）。“广义AI”是使用相同的算法来解决一大类问题的系统。原则上，“广义AI”系统可以学习和适应以解决新的问题，而无须人为干预。“狭义AI”系统使用特定算法来解决特定问题，如下棋、识图等。

总而言之，表2-1是对人工智能系统状态的概述，并且是一个务实和可行的定义。

表2-1 人工智能系统状态

人工智能的定义	狭义AI	广义AI
通用智能系统	智能客服	未来的搜索引擎，百度
认知智能系统	无人驾驶	DeepMind公司

智能计算系统与大数据是直接相关的。任何数据都有产生的缘由，都有产生数据的体系和系统，这就是为什么数据会产生，数据的核心是知识。

通用人工智能系统的核心能力就是通过运用算法、计算体系，把知识从数据里提取出来。一旦有了知识，我们就可以做很多事情。我们可以预测，可以解决自动化问题，可以解决任何需要解决的问题。因为知识告诉我们人有什么需求，社会有什么要求，有了知识我们就可以找到答案。所以，人工智能发展的第一个层次就是通用人工智能。

最近几年深度学习的突破主要是在感知层面，特别是视觉和语音识别，还有自然语言的理解。但这只是起步，接下来要做的是认知上的，因为感知只是把外部世界通过光感、声音的振动，或者语言的交流，变成可以被系统识别的符号。最重要的还是理解它的意义是什么。系统看到一幅图片，可以知道有什么样的物体或人在里面，他们在做什么事情。

目前几乎所有能见到的做人工智能的公司可以分别放在四个象限里面。大部分公司做的其实是“狭义AI”。“狭义AI”只解决一个问题，或者解决一个到两个比较窄的问题，下围棋、打牌或者开汽车都是“狭义AI”。与之相对应，“广义AI”用同样一个系统，可以解决所有的问题，那就与人的智能类似了。“广义AI”是人工智能发展的长期目标，其真正实现至少还需要二三十年的时间。

现在，百度、谷歌、微软、脸书等公司都在朝着“广义AI”的方向努力。判断人工智能的能力，或者判别是否真正是人工智能的标准仍然是，人类是否因此知道更多，做到更多，体验更多。比如，百度基于海量的搜索数据做的很多技术分析，以前靠人去做几乎不可能，而现在通过人工智能计算技术，我们得出了许许多多前所未有的知识和结论，人类因此而知道更多，也能因此去做很多前所未有的判断，去实现更多不可能的功业。如具有代表性的无人驾驶技术、自然语言交互技术，使得人类的运动方式、感官方式都在逐渐发生变化。以前人类用眼睛来看、用耳朵来听，以后我们可能不需要眼睛也可以看

到，不需要耳朵也能听到，人类将逐步拥有新的感知方式，也将体验崭新的世界。

因此，一切做人工智能的公司是否名副其实，都可以从以上角度来衡量：它属于四个象限中的哪一块？有没有实力让人类和机器一起知道更多，做到更多，体验更多？

美国和中国都有很多公司说自己是人工智能公司。有的公司说云计算是人工智能，有的说大数据是人工智能，但这些都只是人工智能系统的一部分，最终判断人工智能实力的是大数据、云计算、算法、训练时间及其总投入以及软硬件综合实力等。

这种实力不是一蹴而就的，也无法一概而论。地上本没有路，在披荆斩棘的道路上有不同的角色，也有不同的站点，每个人、每个企业达到的程度都不一样。有人刚起步，有人身后已经留下大片果实。

百度大脑可以看作人工智能综合实力的一个典型，对它的能力分解，能使我们更明晰人工智能行业的入门门槛及基本标准。如果一家号称人工智能的公司以下能力皆不具备，那只能说这家公司还没准备好真正进入这个领域。

百度大脑是硬件基础、数据基础和算法能力的紧密结合，是云计算、大数据和人工智能的三位一体，是百度技术战略的核心。云计算是基础设施、大数据是燃料、人工智能是发动机，联合驱动着“互联网的物理化”，将数字世界的互联网技术和商业模式又送回到物理世界，全面改变社会。

云计算，名字在云端，却是百度大脑最底层、最实体的部分，是IaaS（Infrastructure as a Service，基础设施服务）。

百度大脑的超强计算能力就来源于这一层，是高性能计算硬件的集团军。这个集团军拥有数十万台服务器，并且采用先进的集群操作系统来统一管理，堪称人工智能超级计算机。

为了深度学习训练的需要，百度自主研发了GPU和FPGA（现场可编程门陈列）异构计算服务器，单机可扩展至64块GPU/FPGA卡，对比传统服务器密度提升16倍，一台服务器即可完成千亿数据模型训练；百度开创性地研制了基于FPGA的人工智能处理器，提供10Tops的计算性能，相对主流的20核服务器，计算效率提升60倍，在人工智能和大数据应用上，可以达到普通服务器4~8倍的性能。

但百度的优势不只是单台机器的优秀，更在于优良的系统，优秀个体的集成形成强大的总体作战能力。针对GPU集群的智能调度和资源管理系统，可以实现计算、存储和网络资源的池化管理和动态调度，计算集群整体效率和平均使用率达到80%。将异构硬件用于线上产品，用户请求时延降至1/5，计算效率提升数十倍。

这个系统涵盖了国内最大的GPU/FPGA集群（全新的芯片技术），最大的HADOOP/SPARK集群（全新的并发数据处理技术）和运营效率最高的数据中心［全新的异构计算技术、整机柜服务器技术、100G RDMA（远程直接数据存取）通信技术和运维技术］，可谓马力十足，提供了开发人工智能所需的计算能力。

它同样燃料充足。基于多年服务于大规模业务，比如通过搜索和视频技术，百度积累了大量的数据：万亿级网页数据，数十亿次搜索数据，百亿级视频、图像和语音数据，百亿级定位数据等。数据就是人工智能算法的燃料，是发展人工智能的又一基础条件。

让硬件与燃料结合的是优秀的算法和模型。百度汇聚了全球顶级科学家和工程师，在理论和实践方面持续创新，搭建了全球最大的深度神经网络，支持万亿级参数、千亿级样本、千亿级特征训练，神经网络层数远远突破100层。

硬件动力、数据燃料和算法灵魂的结合，才产生出百度的PaaS（Platform as a Service，平台服务）。百度PaaS与众不同之处在于，人工智能作为一种横向的服务贯穿全平台。通过深度学习和机器学习技术，结合超强计算、海量数据和优秀算法，在语音、图像、

自然语言处理等方面拥有杰出的能力，打造出独特的知识图谱、用户画像和商业逻辑，并且向用户全面开放。用户可以非常方便地使用各种算法模块、开发工具、数据引擎为自身的商业目的服务。我们形象地把不同的平台称作天算、天像和天工，分别针对智能大数据、智能多媒体和智能物联网这三个领域提供服务。

在最上层的SaaS（Software as a Service，软件服务），百度的人工智能很容易凝结成许多垂直行业解决方案，渗透到各行各业。但我们更追求与合作伙伴一起打造智能产业生态，例如教育云、金融云、交通云、物流云等。我们认为，对智能产业生态的构建能力也是判定人工智能价值的重要标准。

在硬件、数据、算法之上还有一个重要的衡量标准，那就是人工智能企业的文化，即人工智能企业的“软实力”。搜索技术是人工智能的先驱，也是最早的互联网数字化世界的门户，其开发流程和技术核心为未来的人工智能奠定了基础。首先，搜索引擎必须与很大规模的数据打交道；其次，搜索引擎必须同时有大规模的机器学习，人工来做是不可能的事，因为数据规模太大了；最后，也是最根本的一点，搜索引擎的开发流程和工程开发文化与人工智能系统的开发是非常吻合的，都以数据为主，通过抽取其中的特征、模式，然后用这个模式给用户带来价值。人们在搜索业务中结成的协作关系，形成的业务能力和工作习惯，都很适合人工智能业务发展，与海量数据一样，积淀为人工智能企业的文化。所以陆奇在微软的做法是，培养人才先从Bing开始。你做过Bing，你到其他什么部门都可以做，那些技术在搜索看来都是很简单的技术。这个文化当然并不完美，但正如神经网络一样，可以在正确的方法指引下不断发展完善。

人工智能+世界

和前几年讨论“互联网+”一样，人们现在开始讨论“人工智能+”，“+”商业、工业、医疗、教育等。如果从“知道更多，做到更多，体验更多”的角度来看，人工智能对世界的改变是根本性的，也就是说，这是一个“人工智能+世界”的问题。

首先，智能革命对于每个人的日常生活都会起到非常深远的影响。举一个简单的例子，由于人工智能突破性的发展，我们与计算器件的交互形式会更进一层。

以前的人机交互都是通过鼠标和键盘，微软发展到今天的规模，也是通过鼠标、键盘和GUI（Graphic User Interface，图形用户界面）这个人机交互的创新。苹果和乔布斯对这个世界最大的贡献是改为用手指来交互，从而改变了世界。而人工智能时代的改变更大，人类将可以用自然语言跟任何器件交流。

自然语言是最有效、最普遍的一种交流形式。人和人之间就是用语言交流，最自然，应用也最广泛。人机之间自然语言交互的实现，意味着人类将不需要了解每一项应用，每一个产品也都不需要学习该怎么去用，直接操作就好。因为，未来汽车是可以跟我们直接交流的，房子也是可以跟我们顺畅对话的。

我们已经看到了这种智能交互的雏形，比如一些智能助手系统的逐步落地。在美国，人们通过亚马逊智能音箱，把房子变成智能系统。在中国，百度的度秘团队也在这个领域做了很多前沿探索，我们有机会彻底改变日常生活中的人机交互方式。

人工智能将极大地加速人类的创新步伐和创造社会价值的效率，改变社会的规模也会和过去完全不一样。人类进步的每一次革命性进程，都是从发现新的知识开始。无线电是这样，互联网也是这样。而未来知识的发现模式会发生根本性改变。以前，人类去思考、去发现现实世界的规则。数字化世界到来后，借助人工智能的数据处理方式，人和机器将共同发现新的知识。这就意味着，人类创造新的企

业，创造新的社会进程，改变世界的速度将焕然一新。因为人类“知道更多，做到更多，体验更多”的进程大大加速了。

最后，人工智能将带来一场新的工业革命。为什么很多人认为工业4.0以后，人类将进入数字化社会的新阶段？就是因为，智能系统将有能力在现实世界里提取数据，抓取知识，进而更好地帮助人类感知和认知现实世界，也将从经济、社会、文化等层面广泛而深刻地改变现实世界。我们确实处在一个非常振奋人心的时代，跟早期工业革命非常类似，但是人工智能比工业革命对社会的影响更广、更大。

传统的制造业基本上以器械、电器和电力为主，其生产流水线基本上要用很大规模的投资来建立，后续很难调整。比如一家汽车制造厂，要重新建立一条生产流水线，成本很高，花的时间很多。当数据智能、自动化、精准预测对制造业的改造完成之后，后者的面貌将焕然一新。未来的制造业生产流程将是模块式的，全部是数字控制。当一家汽车制造厂要调整生产，制造另外一种样式的汽车，它不再需要重建生产线，而只需要把新产品模块的接口（API）调过来就可以了。这将彻底改变制造业基础，制造业效率也将会极大地提升。

这个改变的核心是数据和知识，即制造的流程、制造的工艺、制造的设计，制造的每一步都会用数字来控制。

再比如制药行业。以前一款新药的诞生要经历长期的研发过程，去发现某种方式对某种病症有效。未来借助人工智能计算技术，将庞大的基因数据与海量的健康信息结合起来分析，人类可以很快发现规律，找到个性化的基因药物。

从国家层面，人工智能给中国带来的不光是整体竞争力的提升，还是一个超越他国的天赐良机。中国是制造业大国，数据量的庞大规模无出其右，也就意味着我们有机会提取比别人多得多的“知识”。你懂的比别人多，看的比别人多，能做到的就比别人多，你就比别人强大。智能时代，在国家竞争、产业竞争中，掌握更多“知识”也就可能使自己立于不败之地。单从制造业来讲，如果中国能把握住这个

机会，完成真正的智能化升级，其他国家是没有办法跟我们竞争的。但是智能制造具体怎么去实现，需要整体的战略考量。

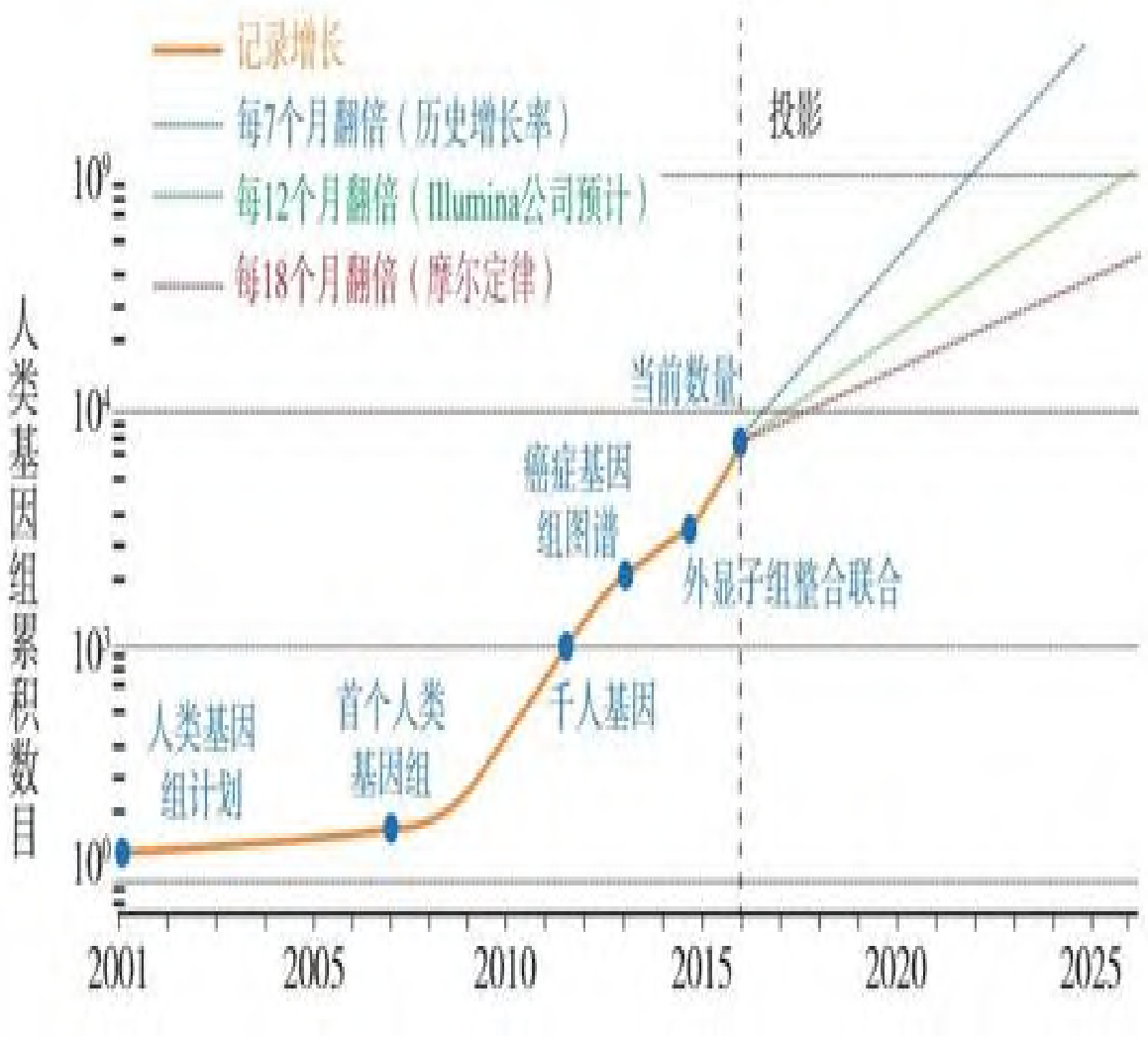


图2-2 DNA测序猛增.[2]

资料来源：nature.com

中国的优势与我们的责任

目前，美国和中国是在人工智能方面发力最大的两个国家。

人们习惯于把百度和谷歌放在一起比较，这也可以看作中美比较的一个缩影。我觉得两家公司有很多类似的地方，因为起源都一样，公司层文化有很多类似的地方，百度在中国的优势跟谷歌在美国的优势也类似。

两家公司也有不同，百度的创新空间和在某些领域的创新速度可能比谷歌更大、更快，这源于中国的国情与美国的国情不一样。

移动互联网在中国的创新程度在很多层面已经超过了美国。

比如，手机百度在信息流方面的创新，用人工智能核心技术可以把信息找人做得非常精致。在移动互联网上做这个事情，百度创新的机会比谷歌在美国要多，因为两国IT工业的格局不一样。还有金融业，百度金融可以依托中国市场和数据的优势，用人工智能技术对金融业做一个革命性的提升。但在美国，金融业相对“壁垒森严”，谷歌要切入金融领域就很难。

再比如，在无人驾驶汽车方面，谷歌、百度都是世界第一梯队，现在谷歌在这个领域稍稍领先于百度，但未来不一定。因为，中国的汽车制造厂家很多，拥有更开放的合作环境，与百度这样的人工智能企业联合起来，创新的机会就非常多，创新的速度也就会更快。反观美国，汽车制造商就集中在底特律那么几家企业，人工智能企业找上门去跟它们合作就很难。

总之，无论是无人驾驶汽车、金融、医疗，还是整体制造业，中美都面临智能升级的节点。但是，中国的宏观环境给百度这样的智能企业的机会和空间要比谷歌在美国获得的更大一点。

那么，在这场波及世界的浪潮中，百度应该承担什么样的责任？

在美国，IT工业体系一般就靠五家公司：苹果、谷歌、脸书、亚马逊、微软。支撑企业不是一家，而是带动了一波企业，扮演的是一个生态系统的领军者的角色。

在人工智能时代的背景下，百度是一家支撑企业，有这样一个机会，因此要争取对中国乃至世界的智能革命做出更大贡献。

具体来讲，战略上我们要以“赋能”来定位。百度首先是中国的，百度大脑要做探路者和奠基者。百度的智能云是提供给所有行业的，对任何行业都将起到促进、赋能、带动的作用。

作为国内人工智能的先行者，百度已经在多个维度上创新开拓，同时逐步形成自身智能生态的雏形。

比如，医疗和教育是人工智能应用潜力非常大的领域，因为其本质都是数据问题，高级教师和老医生一样，能力来自经验（数据）积累，未来我们可以让机器自动分析数据，辅助医生对症下药，辅助教师个性化施教。医疗或者保健能让人的生活更健康，教育能给人更多的知识，因此，人工智能在这两个领域的社会应用价值巨大。

另外，还有无人驾驶领域。无人驾驶也是通过感知、认知、知识获取来实现。目前，无人驾驶汽车真正商用还需要一点时间，但是一旦商用普及成功，对于整个社会的改造会非常大。因为这不光是汽车和交通问题，一旦有了无人驾驶机器，它可以自己行动，自己联网，就会带来很多不同行业的改变。

人工智能的实践范围如此之广，人类很少有这样一个机会，可以彻底改变、改造当下的一切。当然战略落实要一步一步走，方向要坚定，步伐要稳健。

企业挑战：如何落地

在工作态度上，陆奇常说：“Head above cloud，Feet on ground”，就是脑袋要在云端之上，才能看得远，看得清，但是你的

脚必须要踩在坚实的大地上，一步一步向前迈进。

做人工智能事业要跨越的第一个障碍就是如何落地。人工智能意味着一个非常大的改变，时间会很长。要落地的话，首先必须找到非常好的、实际的用户体验，就是能给用户带来实际效益；其次，场景必须清楚，智能助手也好，无人驾驶汽车也好，信息找人也好，一定要有实际的用户体验价值；最后，还要找到商业模式，不然就没有可持续性。

所以重要的挑战在于，是否能够找到落地的用户体验和实现用户价值的场景，然后找到适合的商业模式，建立一个创新的循环。即数据—知识—用户体验—新的数据。找到这样一个循环往复的流，人工智能事业就可以像滚雪球一样往前滚。

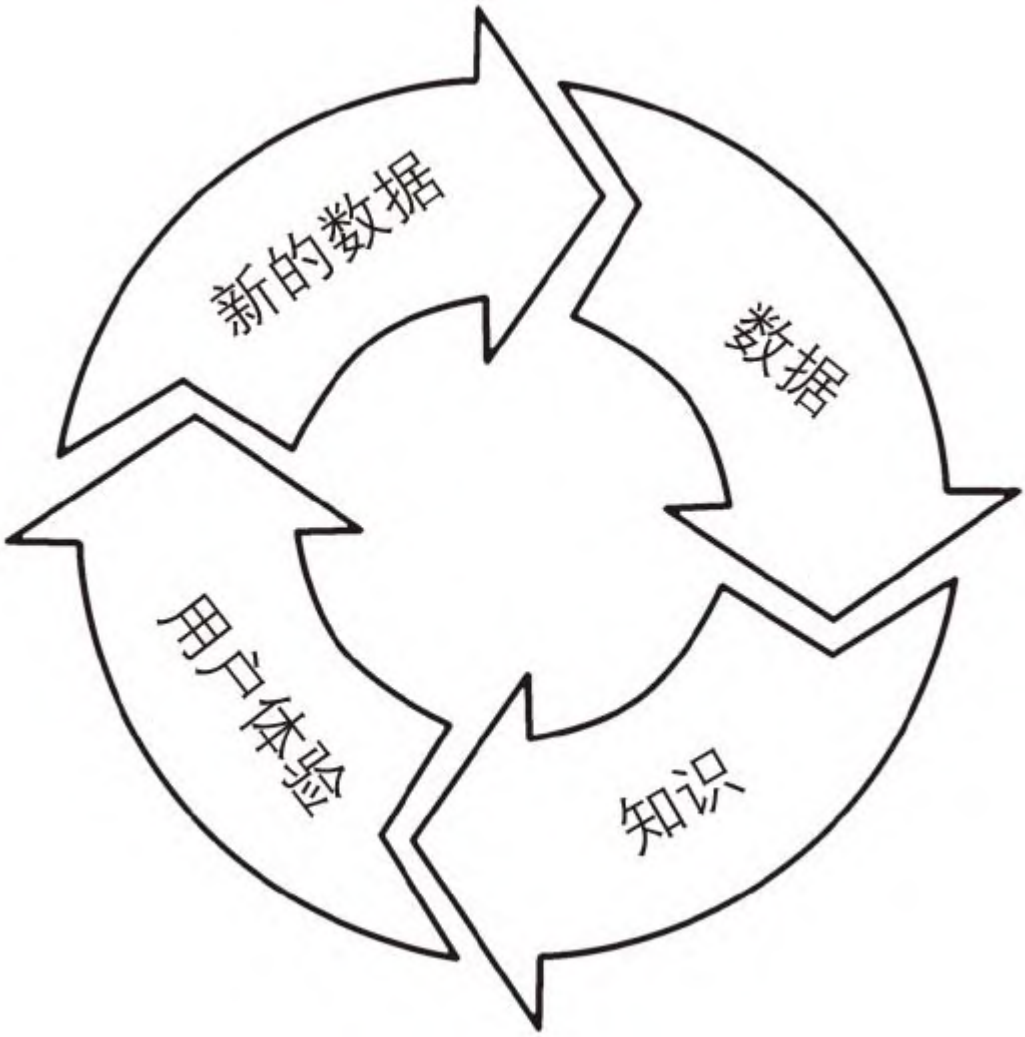


图2-3 人工智能创新“飞轮”示意图

最重要的是每个公司的CEO（首席执行官）一定要重视人工智能，这是起步。然后必须投入一定的资源，包括雇用真正懂人工智能的人和能帮助做决策的人。公司可能是做零售业的，可能是做制造业的，也可能是做旅游业的，要按照自身业务的情况制定有效的智能化战略，然后坚决执行。要赋予执行者足够的权力，通过有效的战略分解，把智能化落实到具体业务上。

我们不妨以“工作引擎”模式来分解人工智能战略的落实步骤。

首先，要根据人工智能浪潮的推进方向，重新梳理企业的定位，根据企业要在人工智能时代抓住的机会，设置新的发展方向，确定崭新的使命和愿景。

其次，根据企业新的定位制定智能化战略蓝图。这就需要企业领导层对公司在即将到来的人工智能浪潮中的愿景进行定位，对“进入什么”和“退出什么”做出取舍，进行投注。

在制定“进出”决策时要遵循一定的原则。硅谷咨询专家Geoffrey Moore（杰弗里·摩尔）的层次结构框架是评估人工智能浪潮的一个很好的例子，其中的关键是进入高增长的类别，跳出低增长的类别。人工智能浪潮将创造新的、有巨大增长潜力的类别，如无人驾驶汽车、机器人、回声设备、对话系统；同时，人工智能也可能给某些行业带来阻碍，因为新的人工智能驱动产品可以用某些方式取代原来的在位者，如新的人工智能硬件+软件堆栈可以使基于旧的HW堆栈的投资受损。一个好的做法是制定包含新的高增长类别、重新增长类别和逆风类别的完整列表，以便领导层可以做出系统性和原则性的决定。

接下来是对产品的出发点做出区分，如产品是否“拥有价值和无与伦比”。需要强调的是，在人工智能时代，企业能否保持差异化，关键在于是否拥有独特的数据资产（它带来独特的知识）。

下一步是理解不确定性、风险/回报和时间表，以进行投注和管理进程。“地平线模型”是一个良好的框架，可以用于制定决策和组织投资组合。大致做法如下：H1时段（未来18个月）围绕目前的核心业务展开；H2时段（未来18~36个月）投资于创造盈利引擎；H3时段（未来36个月+）致力于具有更大潜力但风险更高的长期投注。人工智能浪潮提供了一个非常丰富的H2时段和H3时段的机会，一些人工智能投资甚至可以帮助提升H1时段。总的来说，人工智能处于非常早期的阶段，有很多未知数和不确定性。要想真正深入理解人工智能，有原则性和务实地做出决定非常重要。

在企业人工智能战略的执行阶段，首先要坚持“结构完整性”原则，即在产品体验、技术架构以及商业模式上要连贯一致。如果你正朝ICS（服务器/客户端架构）方向改变，或者投资于“自治系统”，那么技术决策需要与产品和商业决策同步。

其次是企业要紧跟人工智能浪潮的技术路线图，与当前飞速发展的深度学习技术同步是必不可少的。

对于人工智能行业的领军企业而言，就需要可以改变世界的愿景、世界级的技术远见、强大的科研团队和研究议程，这些需要与我们的企业愿景智能技术的呈现和产品开发相一致。DeepMind、谷歌、百度以及一些积极进取的先驱企业都表现出这个共同模式。

在这个阶段，更新研究机制也是必不可少的步骤。因为，传统上，IT行业以及学术界并不擅长将研究成果商业化。最近的OtherLab或OpenAI以及其他一些人工智能的初创企业正在积极招聘研究团队，这是一个新的趋势。有许多工作需要各类组织（大学、早期生态系统、大型企业、培训和研发机构）协同制定出结构化的和可持续的解决方案。

投资力度是企业亟须考量的重要因素。随着智能革命的不断深入，人才争夺战不断升级，导致发展人工智能的成本不断提高。一些初创企业能够筹集大量资金，是因为长期的投资回报是非常巨大的（高风险/高回报）。制定投资规划的关键在于排列资源的优先次序以及一个能够反映人工智能风险的深思熟虑的决策过程。

所有客观条件逐渐汇聚之后，人就成为决定性因素，其中领导才能是一个深远且难得的要素。鉴于人工智能浪潮基于与以往完全不同的核心技术（以神经计算为核心），它需要高层管理团队的高级管理能力。同时，人工智能驱动的新兴行业是如此多样化和跨学科（从基因学到机器人，凡是你可以想到的），因此企业需要一个具备创新精神的人（虽然这并不容易，因为今天的社会生活在很多领域都是非常

专业化的）。微软研究院首席研究员Bill Buxton（比尔·巴克斯顿）提供了解决方案，即为高级管理层建立一个充满创新精神的团队。

值得指出的是，人工智能创新飞轮的核心是数据—知识—用户体验—新的数据的反馈循环。对这个反馈循环的容量和速度进行优化是规划中非常重要的一环。

最后要强调的是，居于战略核心的是基于当前的现状和推断积极设定目标，以及展开可以实现目标的行动。

需要什么样的宏观环境

企业与科研机构的工作离不开良好的宏观环境，正如中国大脑计划是对国家整体层面智能基础设施的呼唤。迎接人工智能时代的到来，也需要政府通过宏观规划创造适宜的土壤。

第一，要确保数据的获取途径畅通。数据越来越成为很多组织的战略资产，可被视为一种新型的“自然资源”。特别是对政府而言，可以通过政策的制定获取数据并公开，以此激励更多的创新。

第二，要有开源的工具和平台。人工智能浪潮需要一个新的硅+软件堆栈，在早期，这类似PaddlePaddle这样的开源工具和平台，其能被开发者和创新者使用是非常重要的。展望未来，我们需要不断降低参与的障碍，并系统使用更多的工具和更多的模块。就像AWS（亚马逊云服务）使计算更容易一样，一些AI-as-a-service（人工智能服务）也可以让人工智能技术更容易被获取。

第三，创新者可以迅速将产品的市场条件和政策体系培育给用户，这也是非常重要的，因为创新的飞轮需要“数据—知识—用户体验—新的数据”的快速反馈循环。

第四，鼓励持续的应用性研究。在人工智能浪潮初起阶段，持续的应用研究特别是开发可以从数据中获取知识，创造智能体验的ML（机器学习）算法是这个变革飞轮的核心。在这个阶段，所有人工智能企业努力的重要组成部分就是拥有持续的应用研究。

第五，上述几点势必会引出人才的问题，教育和培训出更多可以设计、实现机器学习算法并成为数据科学家的人才是一个关键因素。

最后，通过新的结构化方法，把公共世界的信息和知识变成有组织的素材，并可获取，这对于许多企业的人工智能创新至关重要。

智能社会的文化和长期管理

这一波人工智能浪潮需要几十年才能充分发挥作用。雄心壮志和改变世界的投注需要足够长期的投资。所以，有目的的长期管理在人工智能浪潮中扮演了一个重要角色。这将是一种遍及全社会的商业和管理文化的变革。

具体而言，这需要执政者建立并获取一个更大的“许可信封”，使管理团队有较长的时间跨度来培育大赌注（通常是从0到1的赌注类型），这也日益成为高层领导任期职权的一个重要组成部分。马斯克就说过：“如果创新失败，不应该受到惩罚。”

对那些受人工智能浪潮影响而需要进入新格局的公司，整个组织的更新和改造非常关键，而高层领导团队需要面对和管理这种变革。

与长期管理相关的一个因素是创造新的组织结构，使其成熟，以便容易适应人工智能带来的改变。Alphabet（谷歌重组后的“伞形公司”的名字）便是其中一个最早的尝试。在这方面，中国企业的管理创新做得比美国更多。

另外，文化是一个组织的持久力量，它是可以超越几代的领导力和商业活动。对于许多成熟的企业（如谷歌、百度），进入人工智能浪潮代表着一个重大的挑战——需要获取新的人才、新的技术专利以及创造与之相应的新文化。保持积极主动，有耐心，并且执着是极为重要的，因为文化转型是一个成熟企业最具挑战性的工作之一。另外，需要一提的是，相比学习人工智能的新方法，更困难的在于忘却旧模式下的工作方法。

由于我们正处于人工智能浪潮发展的初级阶段，招聘并维护人工智能领域的专家对管理者来说是非常重要的。

总体上，有目的地进行长期管理是抓住任何重要机会而不仅仅是人工智能浪潮的关键。如何调整结构来吸引更多的资金和人才以及放置更多的改变游戏格局的赌注，对企业领袖而言都是一个有趣且极具挑战性的工作。面对更深刻、更有趣、更具挑战性的问题本身就是人类进步的标志。

人工智能技术目前的发展状态

人工智能、深度学习方面每天都有新研究、新文章出来。今天有点像文艺复兴时代，所有的科学都在变。因为科学的真谛就是观察世界、总结知识。我们现在观察世界的能力越来越强，一旦选好观察角度，把深度学习的算法用上去，很快崭新的知识就出来了。

现在不光是物理学、生物学、材料学……每一个科学领域都在拼命往前走，走的速度很快。所以总体来讲，人类是处在一个突飞猛进的状态。让我们再次把头脑放到云端，以量子计算作为本章的结尾。

我们注意到人工智能和神经计算框架之间非常有趣的联系，这两个框架都采用分布式表示超大向量，其基本运算都是线性代数而不是布尔代数。它告诉我们人类的大脑和物理性质的计算方式相似。甚至有科学家提出，人类的大脑运作原理与量子物理、量子计算理论、量子计算算法类似。

关于量子计算和人工智能的结合，我们看到微软、谷歌建立了量子人工智能实验室，国内也有这方面的项目。量子计算不是一个该不该有的问题，而是一个什么时候有的问题，它一定会发生。至于什么时候发生大家有不同的想法，可能是5年，甚至是5年之内就会出现早期的量子机器。

为什么量子计算这么重要？因为量子计算跟人工智能有本质上的关系。量子计算的核心就是利用量子的叠加状态。量子有一个能力状态的变化，就是加一定的能量以后会改变状态。不同于现在的计算机数位，0或者1，一个时间只有一个状态。量子的叠加态特性可以在同一时间有四个状态，计算能力呈指数上升。

好处就是通过量子计算可以解决很多数据问题。以前的办法就是数字分解。比如数据加密和解密，统统要用素数分解。素数分解是非常难的，给你一个很大的数字，用一般的算法，算到地球毁灭可能也算不出来，但是用量子算法以后，可以很快算出来。未来用量子算法做机器学习是很自然的。

与此相关，硬件也一定要革新。因为现在的硬件都是以布尔代数为主，而深度学习的核心计算是不同的，是矩阵和张量的计算，不是0和1的计算，而且一定要进行微分运算。量子计算也是一模一样的，每个量子改变能级的时候，就是一个矩阵和张量的计算关系。大自然其实就是这样计算的，人脑也是一样。Matthew Fisher（马修·菲舍尔）、潘建伟、朱清时等科学家都认为，意识的本质就是量子纠缠。

2007年Nature（《自然》）杂志发表了加州大学伯克利分校Graham Fleming（格雷厄姆·弗莱明）领导的实验室的成果，他们利

用飞秒激光技术，在极短的时间内向光合作用复合物上照射激光，结果发现了复合物上仿佛鼓点般的光回波，这意味着光子的能量不是通过单一路径传入反应中心的，而是利用量子相干性同时从所有可能的路径进行传递，从而证明了量子效应在叶绿素光合作用中起到不可替代的作用。这鼓点般的量子回波，正是大自然与人类智慧之光的映射。量子效应在生物体上的发现极大地鼓舞了人类对量子计算和人机结合的新探索。

虽然量子计算机还没有实现，但很多人已经在思考，假定有了量子计算机该如何来做机器学习？这个领域已经有了很多前沿文章和研究成果出来。假定10年之后量子计算机出来了，它会对人工智能行业带来一个根本性的改变，因为量子计算跟人工智能、深度学习的核心计算是彻底吻合的。我们现在实际上是走了一条弯路：任何算法都要把它变成是布尔代数，用0、1来模拟一个微分方程。

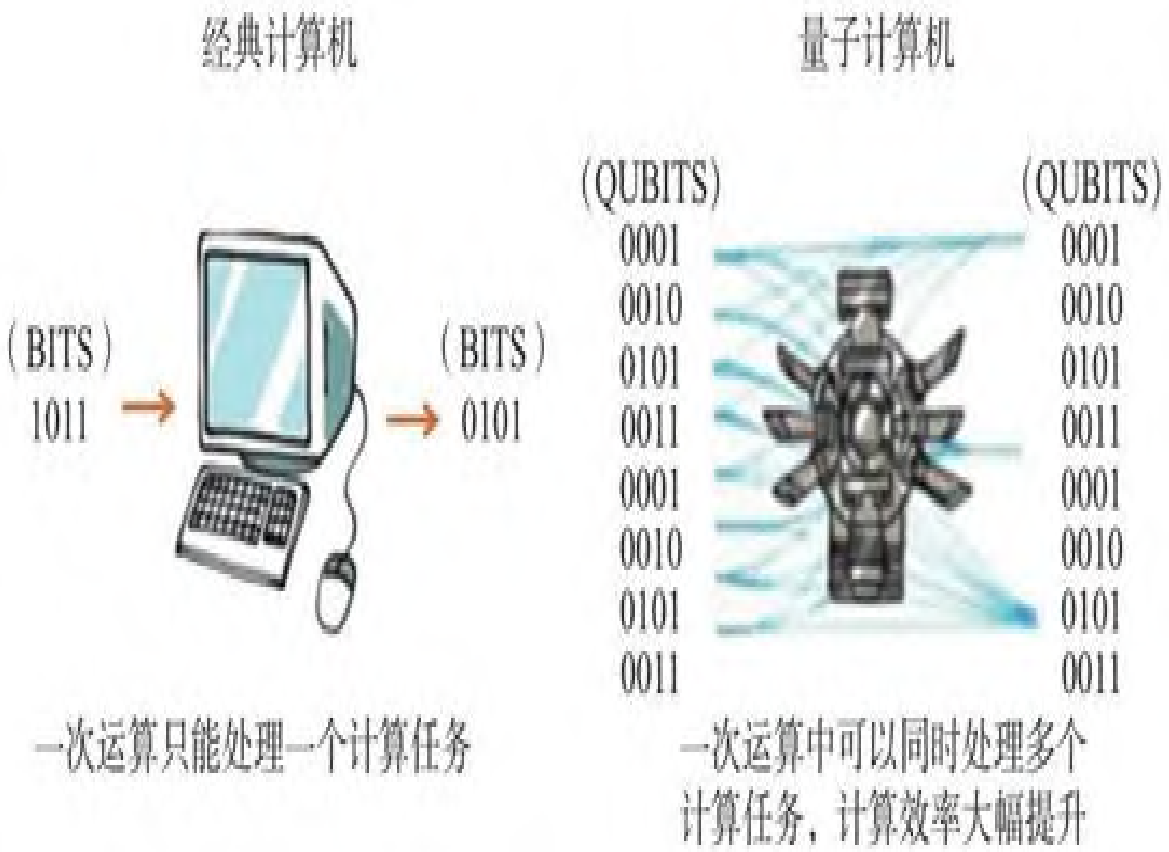


图2-4 经典计算机和量子计算机
资料来源：http://www.zwzyzx.com/show-336-227290-1.html

量子计算和DNA（脱氧核糖核酸）计算的规模和能量将远远超出今天的基于硅的计算能力。随着工程技术的进步，我们将迎来全新的计算体验（如量子化学和量子材料）。

其应用也无比广泛。首先落地的可能就在最古老的农业，参考上面光合作用的例子，就会知道植物也在计算。以后农作物都可能用计算机来计算和设计。所以量子计算对于整个社会可能是一波超级改变，很可能就此引领人类文明的长河走向彻底数字化。

总之，量子计算不是玄学，而正是“知道更多，做到更多，体验更多”这个人类进步规律的未来。在这方面，无论多大的想象力都不

够。我们要敢于想象，同时坚持立足大地。30年前，陆奇在毕业纪念册上写下了“这颗电脑科学皇冠上的明珠非君莫属”，我们这一代人做不到，下一代人继续努力，这是“大写的人类”不变的梦想。

---

[1] 陈景润因证明哥德巴赫猜想“1+2”定理而享誉世界。徐迟的报告文学《哥德巴赫猜想》以陈景润为主人公，发表于1978年1月的《人民文学》第1期，在当时家喻户晓。

[2] 英国2014年启动了“10万人基因组计划”，美国和中国则宣布要完成多达100万人的基因组数据收集工作。世界知名的基因测序公司Illumina仪器测序所得的数据，每12个月就能翻一番。这是一个巨大的“数据黑洞”，也是一个亟待发掘的数据金矿。

03

在大数据与深度学习中

蝶化的人工智能

在历史的重复中变化

当我们谈论数据的时候我们在谈什么？

在大部分人的日常印象中，数据这个词代表的可能是每月的水、电、煤账单上的数字，或者是股票K线图上的红绿指数，还有可能是电脑文件里那一大堆看不懂的源代码。

人工智能眼中的数据含义远比这些广泛。数据的存在形式随着人类文明的发展不断改变，从最初的声音、文字、图画和数字，到电子时代的每一张图像、每一段语音、每一个视频，再到如今互联网时代人类的每一次鼠标点击、用手机时的每一次手指滑动，乃至每一下心跳和呼吸，甚至还包括经济生产中的一切人机动作、轨迹，皆已融入数据流。

不管是浩瀚永恒的引力波，还是复杂细微的DNA，今天的人类已经能够将各种或宏大或微小的事物转化为数据记录，变成我们生活的一部分。数据已经浸染我们生活的每一个细节，就好比生物学家认为人体组织的一半是由微生物组成的，在数字时代，我们生活的一半已然是数据。

历史总是螺旋式前进的。让我们回溯过去，远在人工智能诞生之前，人类在漫长的岁月中也践行着对于数据的发掘、计算和利用。

五千多年前，古埃及人就通过观测记录星象的位置总结出规律：每年当天狼星清晨出现在东方地平线上的时候，尼罗河便开始泛滥。他们照此制订农业耕作的计划，并且将这个周期进行总结，确定了一年365天的太阳历。遥远的天狼星与地球并无任何因果关系，只是出现

在那个位置的时候，正好地球运转到一定的节气——这正是大数据时代的相关性计算的前身。

四千多年前，在今天英国的土地上出现了巨石阵——每块重达50吨的大石头组成了一个圆阵。这是一块原始钟表，夏至的时候，它的主轴线、通往石柱的古道和早晨的第一缕阳光就会处在同一条直线上；往相反的方向，冬至日的最后一缕阳光也会穿过石门。古人用笨重的石头仪表点燃了数据测量的曙光。与中国的日晷一样，这便是最早的数据可视化技术。



图3-1 巨石阵

资料来源：
https://baike.baidu.com/pic/%E5%B7%A8%E7%9F%B3%E9%98%B5/1179755/0/6d81800a19d8bc3ed2257b32808ba61ea8d3451c?fr=lemma&ct=single#aid=17270777&pic=ac4bd11373f08202f436987e4efbfbedaa641bce

两千多年前，托勒密研究天地运动得出三大定律，为天文学打下基础。他的方法很有趣，一言以蔽之，是错误的方法里蕴藏着正确的思路。原本他误以为天体运动的轨迹是圆形，而实际上天体以椭圆轨迹运行。为了强行用“圆”函数来表达天体的实际运动曲线，他采用

多个圆形的嵌套运动模拟出天体运动。他模拟天象使用的嵌套圆圈多达40个，相当于用多个圆运动函数来拟合成一个总体函数。这已经有了最早的拟合函数思想。

什么是拟合函数？当数据很多时，我们可以把数据想象为一个坐标系中分布的很多点。怎样寻找一个函数，使其曲线能够穿越尽可能多的点呢？如果这些点分布很规律，比如呈线性分布，就可以用线性方程描述。

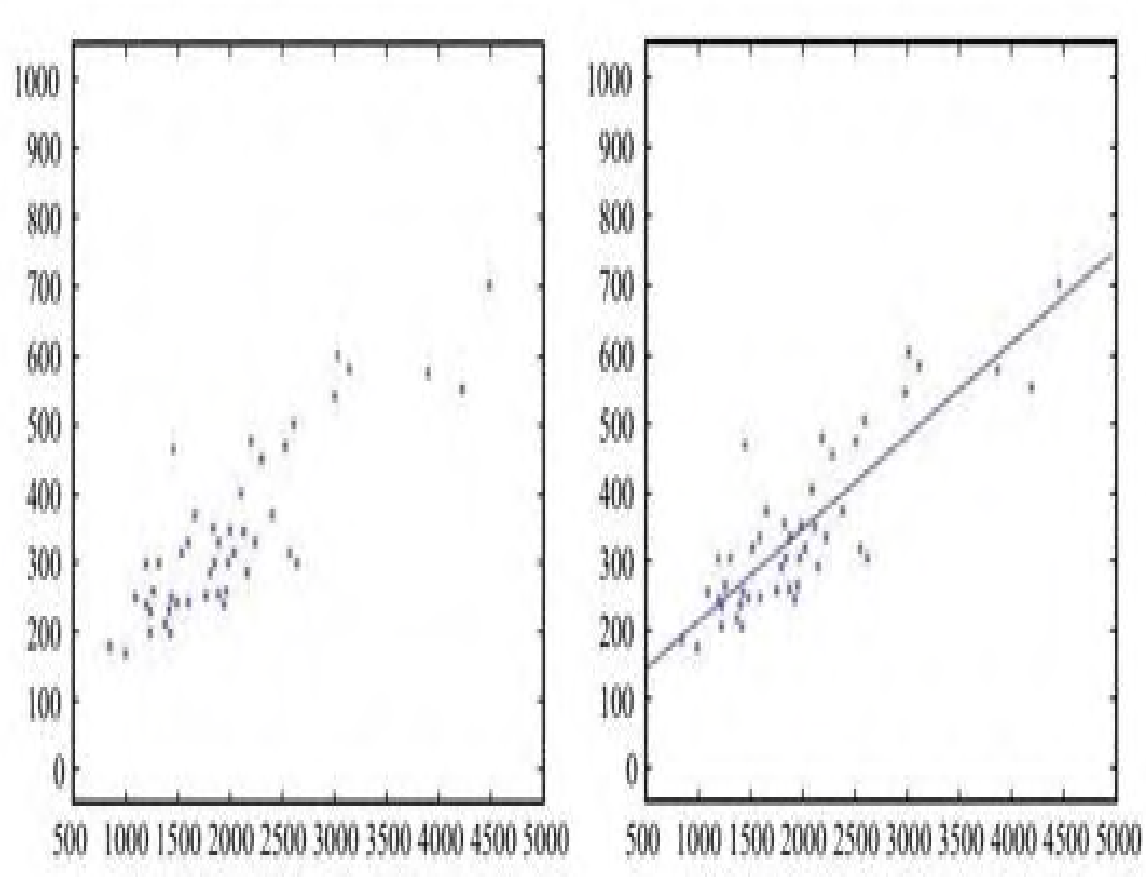


图3-2 用线性方程表示函数

注：左图中分布的点可以用右图中y=ax+b形式的线性函数来近似表达。

如果分布点形成一个抛物线形状，那么函数也很容易得到，就是$X^2=2py$的形式。但是如果这些数据点分布看上去很不规则时就很难找到一个单一函数。现代人想到用多个函数叠加的方法来模拟出一个总

体的函数。调节每一个函数的权重，从而能够让叠加函数曲线尽可能多地穿过这些点。托勒密记录了大量天体运动的数据，然后尝试用叠加多个圆函数的方法模拟出椭圆轨迹的函数，以便把他记录的数据都包括进来。拟合函数方法适合从大量离散的数据记录中找寻规律，而这正是今天人工智能的基础，机器学习的基本数学方法。

由此可见，今天的很多基本数学方法古人已有，只是囿于能力无法尽情应用。

今天的人类可以用数字测绘来还原历史。即便在“我的世界”这样一款游戏中，计算机已经可以计算出每一块砖瓦的角度、长度，将几千年前的古城墙以完美的三维图像重现。那一刻，你会觉得所有古埃及、古希腊、古中国的壮美历史又重新与我们连接。但比起那宫殿中早已褪色的金冠银带，古人使用数据的智慧或许才是人类最宝贵的传承。

数据文明在进步，而大多数人还处在数据的懵懂之中。在日常生活中，数据的概念对于我们既亲近又陌生。我们亲近它，因为每个人从小就会接触加减乘除这样最基本的数据和算法。步入社会后不管从事什么职业，我们这一生也免不了和各类文件、报表或者账单打交道。但与此同时，当面对高科技产品中各种关于内存、分辨率等时髦又复杂的数据时，我们又越发觉得不了解它们甚至没意识到它们的存在。随着大数据、机器算法和人工智能的理念相继到来，这种陌生感愈发加深。

那么数据生活距离我们遥远了吗？正相反，在新的技术条件下，数据与我们日常生活的联系从未如此紧密过。我们的祖先很久以前就学会有条理地储存数据，但是从没像今天的我们一样如此活跃、具体地记录着自己与世界。

从最初的计算器、摄像头，到家用计算机、智能手机，再到大数据和人工智能，我们不断升级采集和利用数据的方式。而现在，从一辆车的每日碳排放量统计到全球气温的监测，从对每个人在网上发言

喜好的分析到对总统选举时投票趋势的预测，从预测一只股票的涨跌幅度到观察评估整个经济系统的发展，我们都可以做到。数据将人与人、人与世界连接起来，构成一张繁密的网络，每个人都在影响世界，而每个人也在被他人影响。这种从微观到宏观的辩证关系就如同在全人类身上发生的量子力学现象，其中孕育着解答无数问题的真理。传统的统计方法已经无法处理这种相互影响的数据。怎么办？答案是让机器自己来处理数据，从数据里习得知识。这便是当代人工智能的本质。

早在60年前，人工智能就已经被科学家当作一门严肃科学来研究。即便普通老百姓都对人工智能兴趣盎然，然而人工智能却在“二战”后人类科技高速发展的几十年间鲜有突破。直到今天，我们才忽然发现各种人工智能的概念如雨后春笋般冒出来，以大数据、AlphaGo、百度无人车等新面貌闯入我们的生活。

如果把人工智能的技术比作一颗早产的心脏，那么它曾经患有两个先天不足：一是在互联网爆发之前，研究人工智能所能调用的数据量太少，这是“供血不足”；二是硬件上的不足导致缺乏解决复杂问题的计算能力，这是“心力不足”。数据如同血液，硬件如同血管，直到互联网应用突飞猛进，计算机的计算能力平均1年翻一番且计算架构出现革命性变化，这两个问题才迎刃而解。奔涌的数据血液进入物理身体的每一个角落，图像识别、语音识别、自然语言处理……睁开了眼睛、竖起了耳朵、张开了嘴巴，机器之心“活”了！

数据书写生活史

数据已经深深“浸入”我们的生活。计算机、智能手机、各种智能家居用品，贴身收集着我们的一言一行，通过计算建模越来越了解

我们，使得看新闻、运动健身、吃饭、听歌、出行等这些最简单的日常活动都成为一次次隆重的数据盛典。

一部智能手机一天之内就可以为他的主人生产1G的数据。这大概是13套《二十四史》的总容量。我们每天都在用数据书写自己浩瀚的“生活史”。

与传统意义上的数据记录定义不同，这种数据是有“生命”的。这种记录不是客观又绝对的数学测量，也不是一板一眼的历史写作。它更像是我们身体的一种自然延伸：倾听我们的声音、拓宽我们的视力、加深我们的记忆，甚至组成一个以数据形式存在的“我”。如果说智能手机已成为人类的新器官，那么数据就是这个新器官所接收到的“第六感”。而处理这种“第六感”的新大脑正是冉冉升起的人工智能。

大数据——万物皆数

既然人类运用数据已久，而且自工业革命以来，数据经历过一次又一次的爆发，何以近年来才出现“大数据”的概念？仅仅是它所能记录和计算的数据量更多而已吗？自然数可以无限数下去，1、2、3、4，以至于无穷，但“多”是不够的，还必须具有几大特征：

第一，大数据的“大”。毋庸置疑，这个“大”相对于人类传统数据的储存方式，不是一个量级上的大小之分，而是几何量级的差距。想想百度地图上每日720亿次的定位请求，再想想互联网上每天有多少次点击、社交媒体上每天有多少文字和图片发出……各种大数据平台一天之内收集到的数据量就可以超越人类几千年来文字、图像的总和。

第二，大数据的另一个重要特点是多维度。多维度代表着大数据可以对一个事物进行多方位的描述，从而更准确。

在电影《谍影重重》里出现过一个大数据公司，能够根据互联网数据、交通数据、历史档案等各种维度的数据帮助美国中央情报局（CIA）迅速追踪和定位疑犯。现实中美国的Plantir数据公司便是如此帮助美国政府追踪本·拉登，提供反恐信息和社会危机预警。它们更常见的业务是识破金融诈骗。

以金融征信应用为例，传统金融机构在进行征信时，一般采集20个维度左右的数据，主要包括年龄、收入、学历、职业、房产车产、借贷情况等。然后综合评分来识别客户的还款能力和还款意愿，决定信贷额度。

互联网公司采用大数据方法，所获得的维度可以让传统银行吓一跳。BAT都开设了自己的金融服务，因为拥有全面且巨大的用户数据，可以查询客户的各种线上记录，比如是否有批量申请贷款等异常行为；还可以将客户信息与互联网全局信息比对，通过欺诈行为模式的比对分析其可信度；更进一步，还可以分析客户的消费行为和习惯，结合填报收入分析还款能力如何。当然，作为用户的隐私，这些数据都不会被公开，用户所能感受到的便利是征信排队时间极大地缩短了，因为大数据可以在几秒钟内就对申请者超过1万条的原始信息进行调取和审核，迅速核对数万个指标维度。

对一个陌生人进行征信就好比“盲人摸象”，传统方法是通过20个“盲人”去评估一个客户的信用“大象”，注定是有缺陷的。而大数据的多维度就如同几万人同时“摸象”，再把这几万人的反馈汇总到一起。维度越多，结论就越准确。

第三，处理非结构化数据的能力。结构化数据中最基本的数字、符号等，可以用固定的字段、长短和逻辑结构保存在数据库中，并用数据表的形式向人类展现（想一下常见的Excel表格），处理非常方便。但是互联网时代产生了大量非结构化数据，对于图片、视频、音

频等内容，它们的数据量巨大却没有清晰的结构。对于图像的数据，我们只能理解为一个二维矩阵上的无数像素点。非结构化数据增长量很快，据推测将占未来10年新生数据总量的90%。而大数据技术可以通过图像识别、语音识别、自然语言分析等技术计算、分析大量非结构化数据，大大提升了数据维度。

非结构化数据的数量远超结构化数据，蕴含巨大能量，应用前景广阔。例如，在机场等公共场合的个人身份检查，过去只能根据旅客提供的身份信息这一个主要维度去判断其身份。而人脸识别、语音识别等技术应用成熟后，大数据可以直接通过摄像快速比对审核，增加对个人身份判断的维度，进行既精确又高效的安全检查。

第四，大数据是生生不息的“流”，具有时间性。它过去就不再回来，就像人无法两次踏入同一条河流。这一方面是因为数据量太巨大，无法全部存储；另一方面是大数据和人类生生不息的行动相关，瞬息万变。百度大数据实验室因此提出一个概念叫作“时空大数据”。

地图就是时空大数据之母。百度地图有一个路段拥堵预警功能。如果前方路段畅通会显示为绿色；如果拥堵则会显示成红色，提醒用户选择其他路线。这是我们与数据互动的一个简明例子。如果我们有A和B两条路线可以选择，此时A路线拥堵而B路线畅通，那么我们都会选择B路线；当越来越多的车主选择B路线，那么B路线将会变成拥堵而A路线又会畅通。此消彼长，变化万千。依靠智能手机的定位功能，百度地图可以实时更改当前的路况监测结果，精确地告诉每一个位置用户当前自己所面对的路面情况。通过数据可视化技术和各种评估手段，可以描绘一座城市的日常脉搏，比如上下班的人流数据变化，仿佛城市在吞吐呼吸。除了被记录下来的，更多数据只在当时有效。把数据全部存储下来是不可能的，那样需要的硬盘可能整座城市的地皮都堆不下，只能即时应用，用过就消失。

与时间数据博弈是富有挑战性的工作。2016年11月，百度正式接入公安部儿童失踪信息紧急发布平台，每当有儿童失踪事件发生时，百度地图和手机百度就会把失踪儿童的姓名、面貌特征、失踪时间等重要信息精准推送给失踪地点周边的用户，使用户可以第一时间参与找回失踪儿童的过程。而在失踪儿童被寻回后，百度地图和手机百度也会及时更新结案标识，让社会各界人士随时了解进展。能将信息早1秒钟提供给用户，就能给焦虑中的家庭多一分希望。

最后一点，也是最重要的，大数据的“大”表现为无尽的重复。对于语音识别来说，正因为人们重复讲述同样的语句，机器通过反复识别这些人类语音的细微差别，才能全面掌握人类语音。也正因为人们周而复始的运动，才让系统能捕捉城市运动的规律。“重复”的数学意义是“穷举”。以往人类无法通过穷举法来把握一个事情的规律，只能采用“取样”来估计，或者通过观察用简单明了的函数来代表事物规律，但大数据让穷举法这种“笨办法”变得可能了。

量变促成质变，在机器智能领域，数据量的大小和处理速度的快慢可以直接决定智能水平的高低。谷歌通过数据量提升翻译质量的故事早已不是秘密。

2005年，美国国家标准与技术研究所如往年一样举办机器翻译软件评测。有许多大学机构、大公司都从美国政府申请了研究机器翻译的科研经费，这些机构需要参加这个评测。没有获得政府资助的团队或者公司也可以自愿加入，谷歌就是后者。参与评测的还包括IBM、德国亚琛工学院等多家机器翻译界老牌公司，个个实力雄厚，在机器翻译领域深耕多年，只有谷歌是初出茅庐。

然而评测结果却让人大跌眼镜：谷歌取得了第一名，并且得分远远高于其他团队。在汉译英这方面，谷歌的表现达到了51.37%的BLEU分数，第二名和第三名的公司则仅达到34.03%和22.57%。最后，谷歌公布了自己的秘诀：用更多的数据！不只是比其他团队多一两倍，而是多上万倍的数据！因为谷歌可以通过搜索引擎收集互联网上人类给

出的海量双语语料数据。同样一句汉语会有很多人给出译法，计算机会通过这种重复来统计出最常用的译法。在没有更改其他主要方法的情况下，仅仅依靠数据样本的增加就训练、改造出了超越其他机器翻译一个时代的产品。谷歌能赢，实际上就是因为“穷举”能力超过了别人。

谷歌、百度这一类互联网企业的数据优势是全方位的，除了翻译，还可以很容易地复制到其他领域，比如语音识别与图像识别。百度的“为你写诗”作为一个小游戏，同样结合了大数据与人工智能。百度主任架构师、机器翻译技术负责人何中军介绍，传统的写诗软件一般运用统计模型，根据给定的关键词生成第一句诗句，然后再生成第二句，不断重复这一过程，直到全诗生成完毕。而百度写诗的做法是这样的：用户可以输入任意词语或者句子，系统结合百度搜索引擎中的大数据对用户表达进行深度分析与联想，衍生出相关度较高的主题关键词。用户随便输入一个词或一句话，比如“西湖”，百度写诗系统通过对大量诗歌散文数据的分析，得出一首描写“西湖”的诗歌应该包含哪些主题词。对于“西湖”来说，得到的主题词可能有“断桥残雪”“烟雨”“垂柳”等。接下来利用深度神经网络技术，根据每一个主题词生成一句诗。这些主题词就相当于人类写作时经常用的提纲，根据提纲来创作可以保证全诗在意境上是统一的，而且前后诗句的内容在逻辑上也是顺畅的。之前大家说机器写的诗看起来每句话都还不错，但是整体意境不够，现在已经能有效弥补了。对于每一句诗歌的生成，则用到了机器翻译技术。对诗歌的第一句进行“翻译”得到第二句诗，再对第二句进行“翻译”得到第三句，以此类推。我们用“西湖”作为输入，“为你写诗”生成的七言诗意境优美、逻辑通畅。



您的输入:西湖

为你写诗

江南烟雨柳丝斜
断桥残雪落梅花
一伞撑开春水绿
西湖岸畔是谁家

长按话筒 开始作诗



图3-3 手机百度的“为你写诗”生成的七言诗
注：使用手机百度或智能革命App扫描图片可见AR效果。

人类的数据镜像

人类在科技产品的娇惯下，口味日渐挑剔，大数据能从那些枯燥乏味的选择之上给出一抹亮眼的色彩。以前的电视机不会回应我们的喜怒哀乐，但现在成熟的视频网站正在耐心仔细地收集着我们的每一种反馈，不管是收藏还是下载，是关闭还是快进，都一一记录，然后利用大数据计算出我们的喜好、消费能力等各种指标。

美剧《纸牌屋》风靡一时，剧中政客们斗牌，剧后则是大数据在下一盘看不见的大棋。出品人为美国著名的网络电视公司Netflix。它深谙大数据分析的妙处，除了以上提到的用户行为，还会尽力收集观看时段、观看设备、观看人数和场景，分析用户喜欢的节目中的主演状况、导演是谁等。通过大数据分析，断定《纸牌屋》的题材会火爆，于是从BBC（英国广播公司）手中高价购买了翻拍版权，并预测凯文·史派西是最合适的主演人选。最终结果证明了Netflix对于《纸牌屋》的押注完全正确。当我们在荧幕前感叹史派西所扮演的总统拥有掌控一切的智慧时，却没有意识到“数据总统”的威力。

美国新任总统特朗普就是一位充分利用数据竞选的总统。据彭博等媒体报道，他的技术团队通过脸书、推特等平台上的用户公开数据，如点赞、转发、收藏行为等，精准描述选民画像，向他们推送因

人而异的竞选广告。甚至特朗普的每条推特、每条脸书都是有针对性的，不同内容对不同网民可见。

通过大数据对用户进行精准画像，这也是百度大脑擅长的领域。2016年十分火爆的电影《魔兽》的出品方传奇影业与百度大脑合作，根据对百度海量用户的分析，将电影广告精准推荐给潜在观众。虽然这部电影在北美市场票房不佳，但在中国大卖2.21亿美元。当魔兽粉丝在影院里高喊“为了部落！”的时候，也许正是大数据悄悄赋予了他们原力。

中国人都说“民以食为天”，比起挑选电影，怎么能“吃好”更是全民关心的热门话题。2013年百度曾发布过一个《中国十大“吃货”省市排行榜》，让网友乐此不疲。这个榜单利用了百度知道和百度搜索的大数据，根据网友们多达7700万条关于“吃”的问答，总结出了各地不同的饮食习惯和特色。

海量数据中被挖掘出了不少有趣的现象：“吃什么水果减肥最快”有多达30万人提问过，看来许多网友在吃的同时还不忘顾及身材；“昨天晚上还活着的螃蟹死了，还能吃吗？”这一问题有高达6万条回复，看得出中国“吃货”对于螃蟹的热情特别高。当然更多的还是诸如“××能吃吗”“××怎么吃”这类的日常问题，光是“菠菜和豆腐能不能一起吃”就引起了无数讨论。

这些问题数量庞大并且看似混乱重复。但重复正是大数据的妙处。大数据可以从中捕捉到更深刻的含义。如福建、广东地区的网友经常会问某种虫是否可以吃的问题；而西北网友则对海鲜的吃法颇感疑惑。不同用户关心的食材、做法各不相同，百度大数据正是从中归纳出了各省市的“吃货”属性。在这背后大数据考量了网友的地理位置、提问回答的时间、问题中关于吃法或者做法等信息，甚至将网友使用的手机品牌等各种维度都纳入计算当中。

除了对人类关注信息的描摹，大数据甚至在构造我们的身体。现在许多人都十分熟悉的健身腕带，就是通过收集我们日常运动作息的

数据，例如，行走步数、卡路里消耗、睡眠时长等来分析我们的健康状况并提出建议。更进一步，未来我们可以将个人数据上传，通过大数据检测我们罹患各种疾病的可能性或者潜在威胁，更好地预防疾病。

关于生活中的大数据有许多例子。我们现在用到的绝大部分成熟的互联网产品，无论是计算机还是智能手机，背后都或多或少有大数据的身影。当我们理所当然地使用这些服务时，就已经邀请大数据进入我们的生活。它默默注视着我们生活中的每一个细节，潜移默化地鼓励和劝告我们做出选择，强化了我们的角色。

突破：机器学习与人工智能

1950年，阿兰·图灵创造了一个针对机器的测试方法，即后来大名鼎鼎的“图灵测试”。这位充满传奇色彩的科学家认为，如果一台机器能够与人类展开对话（通过电传设备）而不能被辨别出其机器身份，那么就可以认为这台机器具有智能。这一简化使图灵能够令人信服地说明“思考的机器”是可能的，而“图灵测试”直到现在也被当作判断人工智能的重要标准。

这个标准已经暗示了一个新的路径，只要机器表现得像人类，我们可以不必过分关心机器的运作规则是什么。有人提出让机器自己来学习规则的办法，人类不用操心那些规则是什么。

1949年，唐纳德·赫布基于神经心理学的学习机制，踏出了机器学习的第一步，创造了此后被称为赫布学习规则的方法。赫布认为神经网络的学习过程发生在神经元之间的突触部位，突触的联结强度随着突触前后神经元的活动而变化，正确的反馈会让两个神经元的联系得到强化。这个原理机制类似巴甫洛夫的条件反射实验：每次给狗喂

食前都先响铃，时间一长，狗的神经系统就会将铃声和食物联系起来。赫布用一套加权公式来模仿人类的神经网，权重就代表神经元之间联系的强弱。赫布给机器创造了一套可以简单区分事物的方法，对于每个数据，让决策树程序做出判断，判断对了就奖励（提高函数的权重），判断错了就惩罚（降低函数的权重）。他利用这个方法创造了一个分类器，可以提取数据集的统计特性，把输入信息按照它们的相似程度划分为若干类。看上去如同人类在观察某种现象时，会观察和总结并区分事物，但机器的这种“观察”更接近一种通过训练达成的条件反射，并非如人类那样思考，重视的是数据中蕴含的相关性关系，而非人类思维中的因果性关系。

之后的十几年中，关于人工智能的研究愈发热烈，灵感一个接一个地涌出。1952年，IBM科学家亚瑟·塞缪尔成功开发了一个可以下得越来越好的跳棋程序。他创造了“机器学习”的概念，并将它定义为“可以提供计算机能力而无需显式编程的研究领域”。

1957年，Rosenblatt（罗森布拉特）提出了感知机的概念，成为日后发展神经网络和支持向量机（Support Vector Machine，SVM）的基础。感知机就是一种用算法构造的“分类器”，是一种线性分类模型，原理就是通过不断地训练试错以期寻找一个合适的超平面把数据分开（超平面可以这样理解：三维坐标空间里二维的形状称作平面，能划分三维空间。如果数据是多维的，那么N维坐标空间里，N-1维就是超平面，能划分N维空间）。如同你把写着“正确”和“错误”的两堆球输入进去，感知机可以为你找出这两堆不同球的分界线。

感知机好比在输入和输出之间只有一层的神经网络。当面对复杂一点的情况时就力不从心了，比如当“正确”和“错误”的球互相混合的时候，或者又有第三种球出现的时候，感知机就无法找到那个分类的界线。这使感知机很难在一些即使看似简单的问题上有所突破。

如今，不需要人类输入规则（编程），而是让机器自己寻找规则，这样看上去机器就有了自己的智能。今天的人工智能便是在机器

学习的基础上发展起来的，只是成长速度受到硬件和方法的限制。

如果多台电脑、多个芯片联网进行机器学习，而且具备多个芯片网络层次，就进入了所谓的“深度学习”的范畴。在20世纪70年代末，Geoffrey Hinton教授等人已经发现，如果能实现多层的神经网络，就可以逐层递进找到模式中的模式，让计算机自己解决复杂的问题。那时他们就开发了“反向传播”算法神经网络。但是多层神经网络的复杂性也导致对其训练的难度大大增加，数据不足和硬件计算能力成为掣肘。

从20世纪60年代中期到20世纪70年代末，机器学习的发展步伐几乎处于停滞状态。这种情况一直到20世纪80年代才有所好转。随着计算机性能的突飞猛进和互联网的到来，人工智能研究终于如虎添翼，在20世纪90年代，现代机器学习初步成形。

互联网在20世纪90年代投入商用，使分布式计算方法获得长足发展。超级计算机造价昂贵，而分布式计算技术则发挥了“人多力量大”的优势，让多台普通计算机可以协同工作，各自承担计算任务的一部分，并把计算结果汇总，效率可以超过超级计算机，而且分布式的结构正好适应了日渐增多的数据量。

计算机神经网络生长与深度学习

由于传统人工智能一味依赖科学家输入的规则模型，导致它只有在解决一些规则比较清楚的问题时才比较有效，比如击败卡斯帕罗夫的“深蓝”就是这样一种“人工智能”。当面对识别一张图片这类人类在婴儿阶段就能学会的简单问题时，这类人工智能却无计可施，因为这种认知类问题只有一个模糊的概念，没有清楚简单的规则。而计

算机神经网络的特点就是它不需要人类提前告知规则，它会自己从海量的基础数据里识别模式（规则）。

顾名思义，神经网络类似人类大脑，由一个个神经元组成，每个神经元和多个其他神经元连接，形成网状。单个神经元只会解决最简单的问题，但是组合成一个分层的整体，就可以解决复杂问题。

Geoffrey Hinton认为，传统的机器学习方法只利用了一层芯片网络，在遇到真正复杂的问题时，处理效率就会变得十分低下。深度学习的最核心理念是通过增加神经网络的层数来提升效率，将复杂的输入数据逐层抽象和简化。也就是说，将复杂的问题分段解决，每一层神经网络就解决每一层的问题，这一层的结果交给下一层去进行进一步处理。

有一层神经网络，就可以找到简单的模式；有多层神经网络，就可以找出模式中的模式。以人脸识别为例，神经网络的第一层只专注于边长几十个像素之类的图像区域，从中识别出一些形状（形状就是模式）——眼睛、鼻子、嘴巴等。再把这些已经识别出的形状交给下一层神经网络，下一层网络在已有的识别结果里，又发现了更大的模式——眼睛、鼻子、嘴巴可以组合成人脸。如果描述得更数学一点，当下流行的深度神经网络可分为应对具有空间性分布数据的CNN（卷积神经网络）和应对具有时间性分布数据的RNN（递归神经网络，又称循环神经网络）。

CNN往往用于图像识别，正如上文描述的，网络的第一层被训练成可以完成这样一个“小目标”——识别图像中局部的独立模块，如一个方块、一个三角形，或者一个眼睛。在这一层，人类输入大量图片数据，只为让该层神经可以辨别基本的局部图形“边缘”，即一个像素旁边没有任何东西。接下来的每一层都在前一层得出的信息中寻找更高层次的模式。这种方法模拟了人眼组合信息的方式，丢弃次要细节，优先识别出某种显著模式。如几个小块和一个圆圈合在一起成为

一张脸，不论它出现在图像中的什么位置，人眼会首先注意这张脸，而不是平均注意图像的所有部分。

RNN则往往用于语音识别和自然语言处理。因为语音和语言是一种按照时间分布的数据，下一句的意义和上一句有关。RNN网络可以记住历史信息。假设我们需要开发一个语言模型，用前面的句子预测后面的词汇。给定“I was born in China in 1976. My college major is mathematics. I speak fluent_____”这句话的最后一个词显然是Chinese（汉语），这对人类很简单，计算机神经网络则需要能够调取到之前的“China”（中国）信息才能做到，这就需要有一种循环设计，使神经网络能够具有一种时间上的深度。

深度神经网络大大优化了机器学习的速度，使人工智能技术获得了突破性进展。在此基础上，图像识别、语音识别、机器翻译等都取得了长足进步。语音输入比打字快得多，机器翻译让我们基本可以看懂一篇外文资讯，图像识别则早已可以凭借一张少年时期的照片就在一堆成人照片中准确找到这个人，甚至可以把很模糊的照片恢复成清晰且准确的照片。

基于深度学习的人工智能和过去的人工智能原理不同，但与我们所了解的数据挖掘有相似的逻辑：先得到结果，反向寻找模式。这个过程被称作训练。

我们用简单的数学知识就能把机器学习、训练和深度学习的基本思维方式解释清楚。

这个方法堪比数学领域的哥白尼式倒转，以简单函数为例可以很好地说明这个倒转。

过去我们解决数学问题，一般是先知道公式（函数），然后输入数据，求出结果。就以$y=ax+b$这种类型的函数为例。比如，已知$y=2x+1$，令$x=1$，可以求出$y=3$。这里x就是“输入”，得到的y就是“输出”。

更高阶一点的数学能力是知道公式和输出，要把输入值求出来，比如已知$y=2x+1$，令$y=5$，求x。

再进阶一步，就触摸到了机器学习。当我们不知道a、b这些系数，但是知道y和x的值，需要把a和b求出来，也就是已知输入和输出，要把函数系数求出来。在$y=ax+b$这个函数里，我们只需要知道两组x、y的值就能确认a和b。

更进一步，假设我们有一组输入和输出数据，但完全不知道函数的形式，又该怎么办呢？这就需要构造函数。比如，已知$x=2$，$y=5$，求$f(x)$。这在输入和输出数据很少的情况下是无法计算的，$f(x)$可能是$2x+1$，也可能是$1x+3$，甚至是x^2+1，以及无数种其他情况。但是如果x和y的数量充足，数学家就能通过“逼近计算”方法，不断调整公式权重，近似求得这个函数。

问题来了，现代生产和生活中产生的数据都无比巨大复杂，如果要从中求得蕴含的函数就需要非常“高能”。人类的脑力已经无法胜任，但是可以把这项工作交给计算机。拟合函数就在这里大显神通。深度学习神经网络模拟了人脑的神经节点，每个节点实际上就是一个函数调节器，无数函数彼此交叉连接起来。通过数学上的矩阵、优化、正则式等各种方法，深度学习过程不断调整着每个函数系数的权重，在数据充分、构造原理合适的情况下，不断演化的函数会越来越准确地拟合大部分数据，于是我们就可以通过这套函数来预测尚未发生的情况。这个过程就是我们所说的“训练”。

吴恩达在谷歌工作的时候，领导团队训练出了著名的计算机识猫系统。

如果用老式的符号式人工智能方法来编程，那么人类首先要对猫进行细致的定义，如尖耳朵、圆眼睛、直胡须、四条腿、长尾巴……把这些特征定义转化为函数输入计算机，然后向计算机展示一张图片。电脑就会分解图片中不同的元素，然后再将这些元素和程序中的

规则进行比对。符合尖耳朵、圆眼睛、直胡须、四条腿、长尾巴等特征，那么这就是一只猫。

而机器学习的方法大相径庭，科学家不会预先编写猫的定义，而是让计算机自己去寻找。科学家只是把图片大量“喂”给计算机，让计算机输出标签——是猫或者不是猫。在识别猫的神经网络中有无数的通路，正如人的脑神经一样，每个通路都会输出自己的结果，如果答对了，科学家就会给这条通路加权（可以理解成亮绿灯）；答错了，就降低权重（可以理解成亮红灯）。经过足够多的尝试，如用10万张各种猫的图片做测试之后，那些得到加权的神经通路就组成了一个识别装置（一组复杂的函数联结）。然后在没有科学家告诉它识别结果的情况下，也可以识别出新的图片中的猫来。训练数据越多，这个函数集合就越复杂但也越精确。

这就是“监督学习”——依赖大量有标签的数据。吴恩达领导的识猫项目甚至可以从零开始学习，不依赖标签就可以辨识出猫。当研究者向神经网络展示了几百万帧静态的猫图片，神经网络自己就获得了一个稳定的模型，从此，它可以和所有的儿童一样，毫不犹豫地识别出猫的脸。

吴恩达的博士研究生夸克·维·乐为此撰写了论文，表明机器学习同样能识别原始的无标签数据，并建立自己的知识模式，它的意义绝非只在于识别猫。

二十多年前，凯文·凯利以“蜂群效应”开始了杰出的新科技著作《失控》的叙述。他以此预测了分布式计算等新技术的出现，那时他可能还没有看到“蜂群效应”中蕴藏的机器学习原理。每一个蜜蜂的运动都是随机的，但是蜂群总能向着一个方向飞去。大量蜜蜂各自的行动（输入）汇总成一个总的运动（输出），中间的逻辑（函数）就是“蜂群效应”。计算机神经网络里的信息运动就像超音速飞行的蜂群采集着数据花粉。在它们看似狂乱的飞舞轨迹中，一张猫的脸庞

凸显出来。百度大脑识别猫的能力已经远超人类，它甚至能够精确区分不同种类的猫。

所以对人类而言，机器学习往往在自己的“内部”形成一个“黑箱”。有人警告这种超越人类理解的黑箱会带来危险，因为我们不知道机器如何思考，是否产生了危险思维。不过更多时候，深度学习会给人带来意想不到的惊喜。

深“度”往事

百度语音识别开发团队的刘洋工程师说过一件趣事：一位语音团队成员在家测试语音识别程序时，无意间清唱了几句歌词，然后歌词竟然被准确识别了出来。这令他很吃惊，其他公司的语音识别技术都还做不到这个事情。而百度团队也并没有针对清唱这种形式做过训练，也没有制定过这个目标。系统怎么做到的他们也不知道，只能说训练数据达到了足够大的程度，程序在不断训练、学习的过程中，自己修得了这项令人称奇的技能。

人们对世界的变化往往后知后觉。在没有深度学习的日子里，世界似乎也一切正常。但有些看不到的代价在被一些人默默承受。周克华这个连环杀手十多年间神出鬼没，为了擒获他，公安部门调集了几乎所有的视频监控材料要发现他的踪迹。那时候公安干警怎么检索视频？全靠肉眼去看！几百甚至几千个小时的视频一段段去看，有的干警甚至累昏在岗位上。而基于深度学习技术的视觉识别将改变这一切。目前先进的监控系统背后都有强大的人工智能支持，在经过大数据训练之后，可以瞬间从视频里识别出人脸、车牌、车型等，并且加以语义化，方便人类检索。然后只要给计算机几张嫌疑人的照片，神经网络就可以飞快地从海量视频中把与嫌疑人有关的镜头都找出来供

人类参考。安防企业宇视科技就开发了这样的智能摄像系统，再结合百度地图，就可以迅速定位嫌疑人或者车辆的运动轨迹。

深度学习在许多用户看不到的地方改变了我们的生活。为了采集和维护地图信息，需要通过采集车拍摄沿路的图像。传统的采集车上要坐两个人，采集过程分为内部作业和外部作业两部分，外部作业就是要开车出去，把沿途的东西都录下来。除了录像，副驾驶要负责用声音记录，每经过一个地方，要说前方这里有一个探头，那里有一个红绿灯，这里是四车道，左转、直行、右转……这是传统的方式，就是一定要把所有看到的东西通过录像和声音的方式记录下来，然后再把数据存储寄到数据处理中心。数据处理中心负责内部业务的人员再一分钟一分钟地去记录比对资料，最后把路面上的这些元素在地图上标识出来，这基本上是一个劳动密集型的工作方式。



图3-4 百度地图采集车

而应用了智能图像识别技术以后，我们先通过深度学习训练机器去辨识红绿灯、车道、探头等路面元素，之后我们只需要将沿路拍摄的全景图像直接交给机器辨识，就能得到完整的地图信息。这就极大地节省了人力，也极大地提高了效率和准确性。

深度学习除了软件算法，还有一件关于硬件的往事堪称佳话。历史上有很多发明在后来的应用中偏离了初衷。比如作为炸药的硝酸甘油可以用于心脏病急救，为了发明战略物资橡胶的人工合成替代品，结果却造出了橡皮泥……在深度学习领域，GPU的作用也被改变了。GPU本来是显卡，用来渲染图像，给图形计算加速，后来却成为深度学习的主要硬件。因为显卡芯片具备比CPU更强的浮点运算能力，原本就用于处理图像这种矩阵数据，非常适合机器学习领域对数据的计算。早期当吴恩达团队率先使用GPU进行机器学习的时候，很多人并不理解。不过今天这已经成为主流。

但最深的往事还是来自搜索引擎。

搜索引擎：人工智能的命运细线

对于今天的中国网民来说，遇到问题“百度一下”已经成为一种习惯。与百度今日的影响力和规模相对的，百度在人工智能领域的专注反而引起一些不理解。更专业的疑问是：电商、游戏、社交、通信……从PC到移动互联设备，无数个风口过去了，百度为何只对人工智能情有独钟？

问题的答案可能与很多人的思维相反，与其说是百度选择了人工智能，不如说是人工智能选择了百度。这是百度基因里的使命，辜负这个使命，会是百度、中国甚至世界的损失。

一切都源于搜索

搜索引擎对于一般用户来说只是一种工具，可以帮助他们找到需要的信息；对于提供内容的网站来说，搜索引擎是一种媒介，帮助它们将自己的内容传递给有需要的用户。在这个过程中，首先搜索引擎要“倾听”用户的需求，即小小搜索框里敲下的那几个关键词，究竟是想要找到什么；其次，搜索引擎要“检索”数量庞大的内容，从中挑选出最符合要求的那些结果提供给用户。

我们审视一下这个过程，是否和我们描述过的深度学习概念模式十分相像？输入和输出在这里都有了，甚至每一次搜索行为都可以看作是对搜索引擎的一次训练。那么谁来告诉搜索引擎输出结果的好坏呢？是用户。用户的点击就是一种回答，如果用户没有点击排在前面的结果，而是去点击第二页的结果，这就是对系统的推荐做出了降权举动。

在这个过程中，搜索引擎不仅提高了推荐的准确性，还越来越懂得判断所收录网页的“好”与“坏”，渐渐学会了像人类一样去分辨网页。最初，它只会读取标题、关键字、描述等页面元素；而现在，百度这样的搜索引擎已经可以辨识出哪些是隐藏的虚假信息，哪些是广告，哪些是真正有价值的内容。

人通过搜索引擎获取信息的行为就是人与机器对话的过程。与以往的人机交互不一样，这个过程基于“自然语言”。相比图像识别、语音识别等，自然语言处理（Natural Language Processing，NLP）是搜索引擎最核心的基础技术。

王海峰认为思考和获得知识的能力成就了今天的人类，这种能力需要通过语言来找到思考的对象和方法，并外化为我们看、听、说和行动的能力。相对于这些能力，语言是人类区别于其他生物的最重要的特征之一。视觉、听觉和行为能力不仅为人类所拥有，动物也有，甚至很多动物的视觉、听觉，包括行动能力比人类还强，但是语言是

人类特有的。而建立在语言之上的知识总结、提炼、传承以及思考，也都是人类特有的。

从人类历史之初，知识就以语言的形式进行记录和传承，用来书写语言的工具不断改进：从甲骨到纸张，再到今天的互联网。所以不管是百度还是谷歌，都认为自然语言处理对整个人工智能的未来都是非常大的挑战。相比之下，语音识别，如声音到文字，或是文字到声音，实际上解决的是一个信号转换问题，但语言不是，语言和人的知识、思维整体相关。

像AlphaGo这样的项目，对于普通人来讲是一件非常震撼的事情，我们也认为它是一个很大的成绩。但是我们不能忽略它的特点：基于完全信息、规则是明确的、空间是封闭的和特定的。为围棋训练出来的智能系统下象棋就不好用。相比较而言，自然语言的处理是更难解决的一个问题。对于下围棋来说，只要计算能力和数据充分，就几乎没有不确定性，而语言问题存在太多不确定性，如语义的多样性。

为了让计算机能够“理解”和生成人类语言，科学家做了大量的工作。在百度，基于大数据、机器学习和语言学方面的积累，研发出了知识图谱，构建了问答、机器翻译和对话系统，建立了可以分析、理解问题（query）及情感的能力。

仅就知识图谱来说，基于不同的应用需求可分为三类：实体图谱（entitygraph）、关注点图谱（attentiongraph）和意图图谱（intentgraph）。

在实体图谱里，每一个节点都是一个实体，每个实体都有若干个属性，节点之间的连接是实体之间的关系。目前百度的实体图谱已经包含了数亿实体、数百亿属性和千亿关系，这些都是从大量结构化和非结构化数据中挖掘出来的。

现在我们来看一个例子，假如有人搜索：窦靖童的爸爸的前妻的前夫。

这句话里包含的人物关系是非常复杂的，然而，我们的推理系统可以轻松地分析出各实体之间的关系，并最终得出正确答案。

百度的自然语言处理技术还可以分析复杂的语法，甚至辨识句子的歧义，而不仅仅是字面匹配。

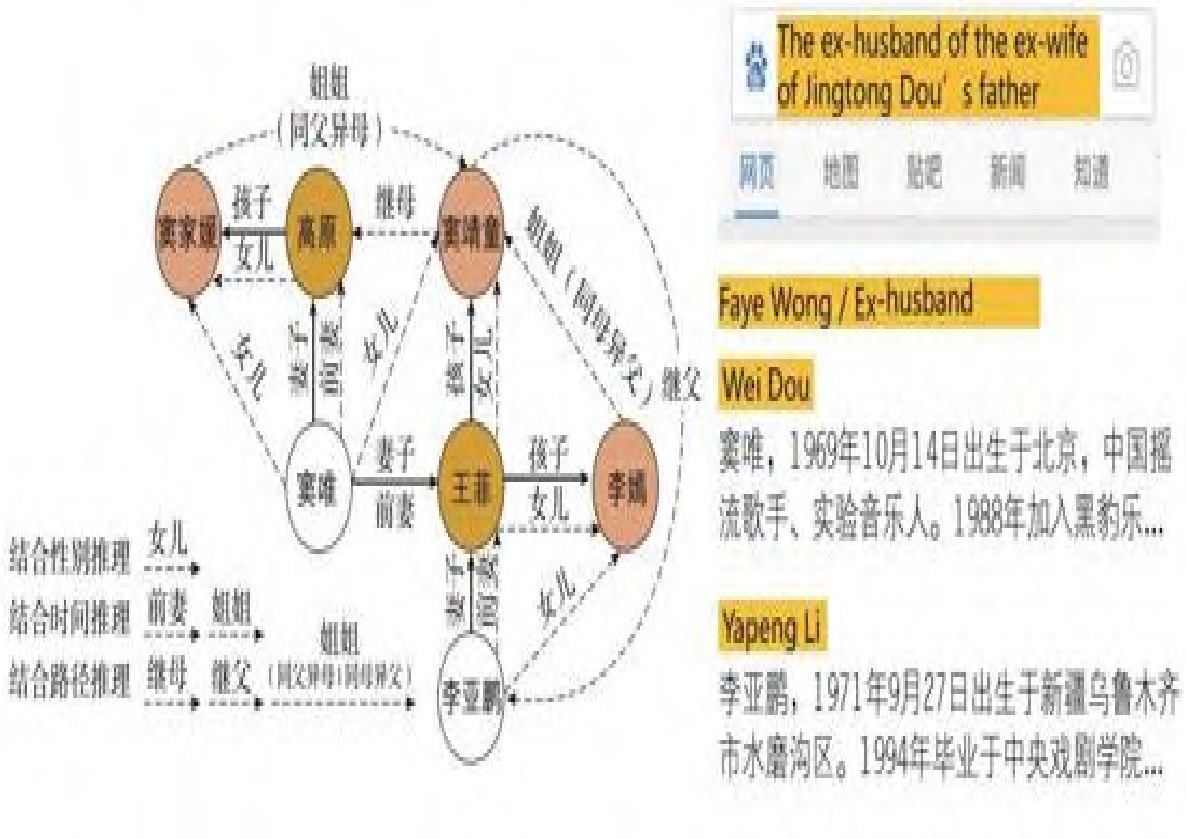


图3-5 人物关系图1

再来看另外一个例子：梁思成的儿子是谁；梁思成是谁的儿子。



图3-6 人物关系图2

如果使用传统的基于关键词的搜索技术，我们将会得到几乎相同的结果。然而，经过语义理解技术的分析，机器可以发现这两个句子的语义是完全不一样的，相应地就能从知识图谱中检索到完全不同的答案。

还有第三句话：谁是梁思成的父母。从字面上来看，这跟第二个句子不同，但是经过语义理解技术，机器发现这两个句子要找的是同一个对象。

深度学习技术进一步增强了自然语言处理能力。百度从2013年开始在搜索引擎中应用DNN模型，至今已经对这个模型进行了几十次的升级迭代，DNN语义特征是百度搜索里非常重要的一个特征。其实，不仅搜索结果相关度变得更高，在篇章理解、关注点感知和机器翻译等方面也都有大幅提升。

搜索所需要的技术基础也正是人工智能所需要的技术基础。比如就云计算来说，主管百度云工作的张亚勤认为，搜索是最大的云计算应用，没有云就没有办法做好搜索，百度是在云里出生的。

搜索引擎的继续进化

随着移动互联网和人工智能的兴起，搜索的形态在发生很大的改变。比如搜索入口变化了，除了通过网页搜索框发起搜索外，基于不同平台和硬件的搜索也在增加，语音或图像搜索部分代替了文字搜索。在人主动搜索信息的同时，信息也被推荐给需要的人。很多人从表象上看，认为这个过程是对搜索引擎的挑战。但王海峰认为，搜索引擎一直同步感知着这个变化过程。

就以“信息主动找人”来说，做Feed（信息流）是当下很多互联网企业的共识。但是“人找信息”和“信息找人”，或者搜索和Feed之间并不是非此即彼，而是相辅相成的关系，在不同场景、不同时段发生不同作用，各司其职，也会互相配合。比如有时候你需要主动找点东西，有时需要朋友推荐，有时候需要系统能猜测你的喜好并推荐。假设别人推荐一篇文章给你，阅读过程中发现一个词不太理解，这时你又需要发起搜索去查找词义。当然机器也会猜哪些词用户可能有兴趣。Feed不可能每天推送给你相同的内容，所以一条内容过了最

热的时候，你再想找它又得到搜索引擎里找。在不同的状态和场景下，用户对于搜索和Feed的需求会相互转换，而如何判断这些状态和场景，正是对系统智能化的考验。有越多的数据和技术储备就越可能做好。

有了做搜索的技术储备和数据，做Feed至少在技术上是不难的。而单纯从Feed起家，想去弥补搜索和数据的缺失就比较困难。百度搜索引擎采集分析的网页量有上千亿，如此规模的大数据为百度持续提升Feed产品效果提供了必要的保障。

搜索引擎在数据洪流中持续进化，Feed只是下一个必要的环节，最终形成无所不在的搜索引擎+推荐。越来越智能化的机器可以做到“举一反三”，到最后用户只说几个词，机器就可以知晓用户想要表达的整个意思。另外还可以自动分析用户所在的位置、身份、习惯等，利用这些信息来决定向用户提供哪些搜索结果。未来在很多时候，我们无须再主动“搜索”，基于搜索引擎的Feed可以主动猜测并推送我们需要的信息。设想一下，比如在一个餐厅吃饭时，搜索引擎已经根据用户之前的搜索内容推测出用户下一步的安排，即使用户还没“问”，都会主动帮用户收集好之后需要的信息，比如当前有什么电影上映，最近的电影院在哪里等。这种设想已经在百度的产品中有所尝试。即使对于用户暂时不关注的信息，不出现在Feed里，也会合理存储起来，像一座无形的图书馆，供用户以后前往探寻。智能化的搜索引擎正在伴随我们一起成长。

搜索是最大的人工智能项目

搜索引擎一刻不停地工作着，它就是人类学习精神的镜像，每时每刻收集处理着大量数据，抓取整个互联网上的页面和内容，不管是电商、社交媒体还是新闻门户，搜索引擎都会“访问”它们。

搜索引擎是播种机、实验场和数字对撞机，结合语音识别、图像识别和机器翻译，通过大量用户的实际使用又可以采集回来更多有价值的数据，反过来帮助神经网络优化训练效果，形成一个良性的发展闭环。

自然语言处理技术的发展会带来更多的惊喜。机器除了可以快速写作具有一定格式的财经、体育新闻，即便在文学方面，机器写出的“唐诗”也令人难辨真假。看篮球、足球比赛时，解说机器人不但能够迅速通报赛场状况，还可以同时回答很多人的提问。这有点像科幻电影《Her》里的智能程序萨曼莎，她可以同时和无数人谈恋爱。恋爱大概是人类最深度的语言、思想、情感交流。萨曼莎堪称自然语言处理技术的一个高级象征，描绘出人类与机器的深刻关系。也许未来，搜索引擎真的会像萨曼莎一样，穷尽符号信息，闯进语言与意义的空隙之处，超出人类的想象。

严格来说，人工智能是一种“体力活”，要有足够的体力才能经受住那样巨大的数据和计算。而在一般的大学高校或者较小的互联网公司，数据量和硬件成本上的门槛使得发展人工智能受到很大的限制。即使不考虑CPU、GPU等硬件的购置费用，光是运维这些硬件的成本就很高，AlphaGo下一场围棋就要耗费3000美元电费。现在百度除了传统的服务器、带宽等基础设施，还拥有数百台支持人工智能运算的GPU服务器，最高配置的服务器上可以安装16张GPU卡。在这一切的基础之上，将数据储备、硬件基础、市场规模和人才团队统筹起来，最大限度地发挥优势，所追求的就不是一时一地的得失，而是最大、最基础的人工智能平台，为人类“知道更多，做到更多，体验更多”而努力。

人工智能，这是百度的命运

可以说，人工智能对于百度、谷歌这样的公司，是一种内在的诉求，也是互联网、移动互联网和数据大爆发自身的诉求。国内在这个领域很难有其他公司能与谷歌、微软这些拥有规模优势的公司抗衡。建立基础设施基地和人才高地，是百度义不容辞的责任。

将人工智能的火种传递到更多人手中，创造实际价值，让生活更美好，让国力更强大。这样的愿望给了百度人动力，也是百度能够集结众多人工智能科学家的原因。

林元庆本来在NEC美国实验室研究人工智能，那里的条件和氛围都很好，学术性很强，可以专注于研究和发表论文。但是他还是选择离开熟悉的环境，选择了百度。他说最重要的原因就是，作为一个人工智能的研究者，他觉得把深度学习的技术真正实践到应用层面是非常关键的一环。现在中国有超过7亿的互联网用户，超过12亿的手机用户，都是世界之首，如何让广大用户都享受到人工智能带来的改变并参与这种改变？这种探索的价值能够影响全中国所有人的生活。他迫切感到“这是最好的时刻，是人工智能最有希望的机会，错过了就太可惜了”。

但人工智能从不休息，当人类在睡梦中，它们依然在机器世界里奔涌，在无尽的循环往复里蝉蜕蝶化，终将飞向世界！

这里想引用一位著名哲学教授写在20世纪90年代的一段话作为结尾：

在天堂里，人还不是人。更准确地说，人还没有被投放到人的道路上来。现在，我已经被抛掷出来很长的时间了，循一条直线飞过了时间的虚空。在什么深层的地方，还是有一根细细的绳子缚着我，另一头连向身后远处云遮雾绕的天堂。个体灵魂不是她自己选择的，而是从天堂抛出的系在她身上的细线，使她的身体身不由己。薇娥丽卡自己不可能去找到一种生命热情，只能从自己身上发现自己的生命热情，这就等于发现把自己的身体与影子系在一起的那根细线。从天堂那边抛出来的细线决定了薇娥丽卡身体的生命方向和个体灵魂的在世

负担，感觉到自己的个体命运。所谓个体命运不过是，一个人感到唯有这样的生命热情的散发才让自己有美好地活过的感觉，才有自己身体的在世幸福，以至于非如此生活不可。[1]

---

[1] 刘小枫。沉重的肉身［M］.北京：华夏出版社，1999.

04

中国大脑计划：

自下而上的超级工程

人机世界迫切需要新的大脑

很多人工智能科学家都拥有生物学和计算机科学两个方向的专业背景，这大概是智能生物发展的一个缩影。

地球就像是一台生物计算机，漫长的生命进化过程就是各类生物“程序”不断迭代的过程。在自然环境的作用下，无机物渐渐聚合成有机分子，有机分子演化组合成为蛋白质分子。携带生命信息的蛋白质分子仿佛是一个个数据字节，数量巨大又随机组合，能穷举出各种组合形态。其中一些不仅可以吞吐物质，还能进行新陈代谢并复制自己——生命便由此产生。

最基本的生命单位都携带着遗传基因编码。从此，生命的“编程语言”诞生了。大自然之手利用这种语言进行各种创造——基因代码变异、组合，产生各种新的生命体征，进化出五花八门的生物。而神经系统发育水平的高低直接决定该种生物的等级。

每种生物和新的生命组织都可以看作一段可执行程序。程序之间可以组合，代码可以更新，从而发展出更强大的程序。如果这段程序运作良好并且能可靠复制（繁殖），它就会生存下去。正如计算机中的程序必须“完善”才能持续运算下去一样。

不过，与我们如今手边的任意一台计算机相比，地球这台巨型生物计算机的运算速度实在是太慢了，一个程序的运行过程就是一个生物体的一生。几十亿年过去，才发展出唯一一种最高等级的智慧生物——人类。很自然地，这个星球上也没有什么自然生物可以进化出超过人类的大脑了，除了人类自己创造出的计算机。

计算机中的程序可以飞速迭代，但借此衍生出的人工智能并没有飞速发展。由于基于规则的编程和迭代太依赖人类，结果反而被人类制约了。如果让计算机自己编程呢？毕竟深度学习就是建立在非线性编程原理的基础上，让程序自己改造自己。我们往往难以理解深度学习神经网络解决问题的逻辑，正如人类并不了解自己脑中的那些意念、思想究竟是如何从脑细胞里产生的一样。

地球早已被生物层覆盖。如今，这部生物计算机迎来了第二次进化，那就是同样包裹着地表的计算机、通信网络、各类传感器和人类活动共同构造的信息层所孕育的进化。在数据分子与人类的结合中，新的数据生命形态正在形成。它们需要新的大脑。



图4-1 百度大脑示意图
注：使用手机百度或智能革命App扫描图片可见AR效果。

百度大脑就是这样一种尝试。比起生物界不自觉的进化之旅，它更在意的是当下的实际运用：个人、企业和社会都迫切需要人工智能的辅助，然而人工智能还散落在各处。百度大脑规划着要提供集中而优质的人工智能，通过互联网神经元把相关信息传导循环起来，加速世界的智能化。

第一棒：百度大脑

几年前，亚当·考特斯（Adam Coates）在斯坦福大学的研究进入博士后阶段，他曾问导师吴恩达：“做什么事情，在哪里做，能够让我们的研究在这个世界上产生最大的影响力？”吴恩达告诉他应该去百度。说起这件逸闻时，亚当担任百度硅谷中心主管已经一年多了。

如今，大概没有人能否认这组对话中蕴含的深谋远虑。不过在最初时，百度美国研究中心还扮演着员工赴美出差“中转站”的角色。

2014年，百度第一次披露“百度大脑”，这个抽象的概念仅仅在媒体上留下了一个影子。经过两年沉淀后，才有了2016年乌镇世界互联网大会上百度“第一次向外界系统介绍百度大脑”。此时，外界得知它已经和超过3万家企业展开合作。

研发百度大脑是一件苦差事，但负责研发的人未必都是苦孩子。这个团队里有从小玩小霸王的极客；有人的家乡还没有被百度地图所覆盖；有人在做“百度医疗大脑”的同时，重拾起学生时代对医药领域的热情；还有人一边说着“很难、很深”，一边坚持研究如何“把体验做到极致”。这种极致，有时对标的甚至是科幻。

许多百度大脑年轻的科学家都爱看科幻影视。普通人看到的是“幻”，这群博士、博士后看到的是“科”。同样是看美剧《西部世界》，语音部门的工程师高亮就说：“看剧情发展，我感到唤醒设计、声纹识别和远场技术被做到了极致。对，未来人机互动就应该是这样的！”

建设百度大脑也带有科幻成分。我们不妨先从了解其基础架构开始。

百度人工智能业务最基础的是“物料层”，有基于GPU/FPGA的云计算平台、深度学习代码平台和大数据储备。这一层提供了进化的环

境和工具。最上层SaaS是人工智能的各种应用。

介于两者之间的是“人工智能基础技术层”。大脑所具备的“听说”（语音识别与语音合成）、“看”（视觉识别）、“读写”（自然语言处理）等认知功能就在这一层，大脑具备的决策规划、运动控制、预测推荐等决策功能也在这一层。

广义的百度大脑即包含了以上三层。张亚勤认为，三层的结合更体现出百度大脑的综合实力。

百度大脑是百度云的核心引擎，百度云是百度大脑的云化，云为百度大脑提供了神经元和数据训练源，百度大脑则通过云向各类产业输出服务。

具体来说，在“物料层”，百度是全球首家将GPU芯片大规模用于人工智能和深度学习领域，并规模化商用ARM（Acorn RISC Machine）服务器的公司。百度也自主研发了基于FPGA芯片的服务器。加上传统的基于CPU芯片的服务器，把这么多性能、结构、原理各异的服务器整合在一起，就需要强大的异构计算能力。通过异构计算技术、100G RDMA通信技术、高效的整机柜服务器技术，百度打造了全球最大规模的GPU和FPGA混合异构计算集群，把数十万台服务器合为一体，构成百度大脑的实体，保证了百度大脑超强的计算能力。

有了强大的硬件构造还不够，大脑中还要有内容和数据。这就好比是人类的记忆。在IaaS之上的PaaS是我们人工智能的平台，所有的样本、特征和功能都在PaaS这个层次生长。百度搜索十多年来累积的全网Web数据、搜索数据，以及百亿级数量的图像、视频、定位数据都是百度大脑不断学习、快速成长的养料。

在大脑结构和记忆内容之外，百度大脑还要有认知思维能力。系统通过深度学习来模拟人类大脑的神经元，即通过万亿级的参数、千亿级的样本、千亿级的特征训练，来模拟人脑的工作机制。这也是世界上最大规模的深度神经网络。

“人工智能基础技术层”包括语音识别、图像识别、自然语言处理以及所有的知识图谱、商业逻辑和用户画像。

SaaS层面可以看作百度大脑的触角层，会更加垂直化，深入交通、教育、金融等各个垂直行业。对百度来讲，这三个层次就是百度大脑通过云和垂直行业，给商业客户提供的一种能力和服务，更是一种营造智慧生态的能力。

百度大脑全面超越了过去的信息技术服务，比如过去只是做计算、存储和网络，大家都可以做，现在三个层面有机结合起来，灵魂就在于贯穿全体的人工智能。

通过硬件—数据—算法的联合滋养，百度大脑的能力会滚雪球式增长，越来越善于处理数据、提取知识、了解使用者、善于解决问题并获取更多的知识，实现“数据—知识—用户体验—新的数据”的正循环。

另一位“互联网教父”尼古拉斯·尼葛洛庞帝（Nicholas Negroponte）说道，“当我听说‘百度大脑’的时候，我觉得这些人真是太疯狂了”。制造机器大脑，看似一个科幻文学中的疯狂幻想，但科学家的信念和努力，使得看似“疯狂”的想法已经走在平稳实现的道路上，正如一个真实生命的成长。

训练“大脑”就如同教育孩子，从零开始，在语料而非语法的熏陶下学习语言，从大量图片中形成对事物的“印象”，这个过程就是通过试错来熟知世界。也许同样的事情，人类1岁孩子都能简单做到，而百度大脑常常要花上几百倍甚至上万倍的时间和精力。可是从另一个角度看，它像是全人类的孩子，有希望继承现有文明的所有经验和记忆，人工智能“大脑”的进化本质上是人类文明的进化，潜力无限。

目前，百度宣布将向社会免费开源人工智能深度学习平台PaddlePaddle，开放百度大脑开放平台ai.baidu.com。前者为开发者

提供算法编程环境，后者为应用开发者、数据工程师、数据科学家提供现成的百度人工智能成果接口。百度大脑实实在在地分享着，愿与所有相关企业一起融化人工智能这座冰山。

百度大脑的听与说

人机对话的第一步，就是要让机器学会“听”和“说”。“听”是不断追求准确度，“说”则是要让大脑有人性、有人味。

作为“大脑”最基本的能力之一，实现“听”这一功能的语音识别技术研究经历了从标准模板匹配转向基于统计模型再到深度神经网络的过程。起初，辨识语音必经声学模型到音素模型，再到语言模型的多步骤转化。近年来，在大量语料和深度学习的训练下，这一步骤已经被极大简化，机器从输入到输出中自行生成程序，准确率有了大幅提升。百度大脑也就“听”得更清楚了。

2011年，百度在语言识别领域起步，到了2016年，百度语音识别技术的准确率已达97%。

2012—2016年的4年里，百度语音识别的准确度提升了近30%。即便是有严重地方口音的普通话，每100句中百度语音也能毫无差错地识别出85句。按照可以听错一个字的标准来计算，百度语音则可以准确识别出98句，而未经训练的普通人只能听懂60句。这个识别体系要做到方言识别，需要至少720小时的语料用于训练，从声音、内容到说话者，都要不断变化以提升系统的敏感度。

要让百度大脑“说人话”更是一件很有难度的事情。百度则是运用声学模型和语言模型。声学模型决定了语言的发音。打出一个字，系统就在原始音库中找到合适的发音对应上去。要让电子发音没有机

器味而有“人味”，就要为语音资料建库。例如让机器学习时间从20小时升到100小时，此后机器模仿出来的声音听起来就舒服多了。为了保持语音的连贯性，不至于让合成语音听起来像“断气”了一样，百度大脑语言模型还会持续提升文本库的学习连接概率。比如说出“中华”，系统还能在后续词组中选择“人民共和国”“民族”“儿女”等进行匹配。

长语音是让机器语音更有气场的技术。情感合成、远场方案、长语音方案等能为合成语音加入情感，让其更接近真人发声效果。

语音识别的场景应用极为广泛。比如我们可以借此创造出一个“最强销售员”：销售员新手给客户打电话时，百度大脑实时记录客户的回复，并显示在电脑屏幕上，系统可以瞬时搜索并调取优秀销售员以往对这些问题给出的回复。这样，每个“菜鸟”只要“照本宣科”，就能在上岗第一天掌握过去最优秀的销售员所具备的交流能力。另外，2014年，百度为特斯拉汽车提供智能语音方案，中国车主可以使用语音控制车载娱乐系统、指挥地图导航、发起搜索，甚至通过蓝牙拨打电话。

随着语音识别而来的还有对语音特征的抓取。例如在胡歌朗读《青玉案·元夕》的上阕后，语音识别系统可以自动生成下阕。对于一些语音有特点的明星声音，目前只需录制和分析大约2000句，系统就能合成他们的声音。

目前，百度每天响应的语音合成请求达到了2.5亿次。在情感语音合成技术上线后，用语音听小说的百度用户每日停留时长从过去的0.69小时增加到现在的2.21小时。未来，这一功能除了“读小说”还将给予人们情感慰藉。当家里老人和孩子想念忙碌的家人时，可以随时有“他们”的声音陪伴。

百度基于神经网络翻译模型技术的机器翻译系统正在快速学习各类语言。经过6年的积累，如今的百度翻译已经可以支持全球28种热门

语言互译，覆盖756个翻译方向。百度语音可以支持粤语、沪语等方言的识别。

上帝曾伸手割裂了人类的语言统一，让四面八方的人由于语言不同而无法沟通。有了机器翻译，人类终于可以携起手来，建造出一座真正的巴别塔[1]。





图4-2 荷兰画家彼得·勃鲁盖尔绘制的巴别塔
注：使用手机百度或智能革命App扫描图片可见AR效果。

百度大脑的好视力

“视觉”承担着我们80%的信息摄入工作。在解决“听”“说”问题的同时，我们也要教会计算机“看”，即图像识别。以识别一朵花为例，用户将图片上传百度后，百度大脑将它转化成“0101”的数字流，然后输入深度神经网络，经过层层分析、层层抽象，对包括像素在内的各层信息与现有的大数据进行比对，才能重新还原并识别出它是一朵花。这种方法其实和人类眼睛的功能是近似的。

这一切都要建立在预先对图片做出分类的基础上。目前世界上最大的图像识别数据库ImageNet的图片分类有1000多类。而百度图片数据库的分类则已经达到4万类。

百度正在从四个方面推进计算机视觉计划。首先是人脸识别，通过捕捉人脸关键点形成人脸表情网，实现人脸的准确识别；其次是在类似百度地图的产品中，实现地图服务与图像智能识别技术的结合，打造数据无限逼近现实世界的效果；此外，百度无人驾驶技术也正在

利用计算机视觉进行程序优化，从而使无人车的研发速度加快；图像识别还会被应用于AR（增强现实）领域提高视觉效果。

百度的人脸识别已经远远超过人类，百度资料库中有超过2亿张人脸照片，浏览识别训练量超过200万张。目前，百度大脑能够自动确定图片中人脸的数目及每张人脸的位置和大小，并支持正面、侧面多个角度。即使目标在运动中也不会降低甄别率。系统通过定位眼睛、眉毛、鼻子、嘴巴、脸颊轮廓等70多个关键点位置，能够进行像素级人脸解析，并根据人脸图像识别人的性别、年龄、表情、姿态等属性。

像AlphaGo团队一样，百度也好奇自己的技术边界在哪里，人工智能的研发体系究竟有没有偏差？为此，百度团队参加了江苏卫视大型科学类真人秀《最强大脑》，与能够裸眼区分520杯水的“水哥”王昱珩一战高下。在此前的节目中，“水哥”战胜了蚂蚁金服的人工智能“蚂可”。作为国内最强人工智能技术的结晶，百度团队为了这次荧屏上的人机对战特意优化了算法，最终证明百度团队的机器人小度确实技高一筹。

在百度人脸识别技术落地的产品中，最为“高大上”的当属乌镇的“刷脸”门禁系统。有出入资质的人将面部信息事先录入系统，今后出入任何安装有“人脸闸机”的地方，只需“刷脸”即可。

这种技术在百度内部俗称“1对1”，也就是一张面部与资料库中的信息比对，与之对应的是“1对N”。这个概念我们经常在西方谍战电影中见到：系统在监控信息呈现的茫茫人海中搜索一个人的面部信息以确定其方位。尽管国产电影中我们很难看到这种“炫技”，但在“1对N”的比对中，百度能真真切切地做到99%以上的识别准确度。

这项技术说起来容易，做起来难。等到技术成熟和数据库之间互相联网时，我们出门乘坐飞机、火车就无须出示身份证以供查验。因为当我们进入任意的交通枢纽，被摄像头拍摄到面部信息时，系统就能通过人脸识别确认我们的身份和购票信息。这相当于普通人都能获得一条绿色通道，出行效率和公共秩序都将大幅改善。

百度的人脸识别系统只需要1根蜡烛汇聚在1平方米以上的亮度，就能够完成识别和判断过程。在远程开户场景中，活体识别技术反应速度达到每秒20帧，交互过程费时不到2秒。在这一基础上，结合视频序列，百度已经将人脸身份验证应用在百度金融反欺诈领域，用于贷款审批、远程身份证识别、银行卡识别等，从而精准识别用户，预防欺诈。

百度大脑的“好视力”除了搭载大数据，还能做到许多脑洞大开的事情。当我们拍摄了各种角度的故宫太和殿照片后，百度大脑可以去掉重叠和无用的图片信息，通过计算和建模完成太和殿的三维结构建设。这样，千里之外的人们也可以通过网上的虚拟游览身临其境般地感受太和殿的宏伟。随着上传的照片越来越多，百度大脑就能重构更多的景点，让世界各地的人们都可以足不出户体验三维虚拟旅游。

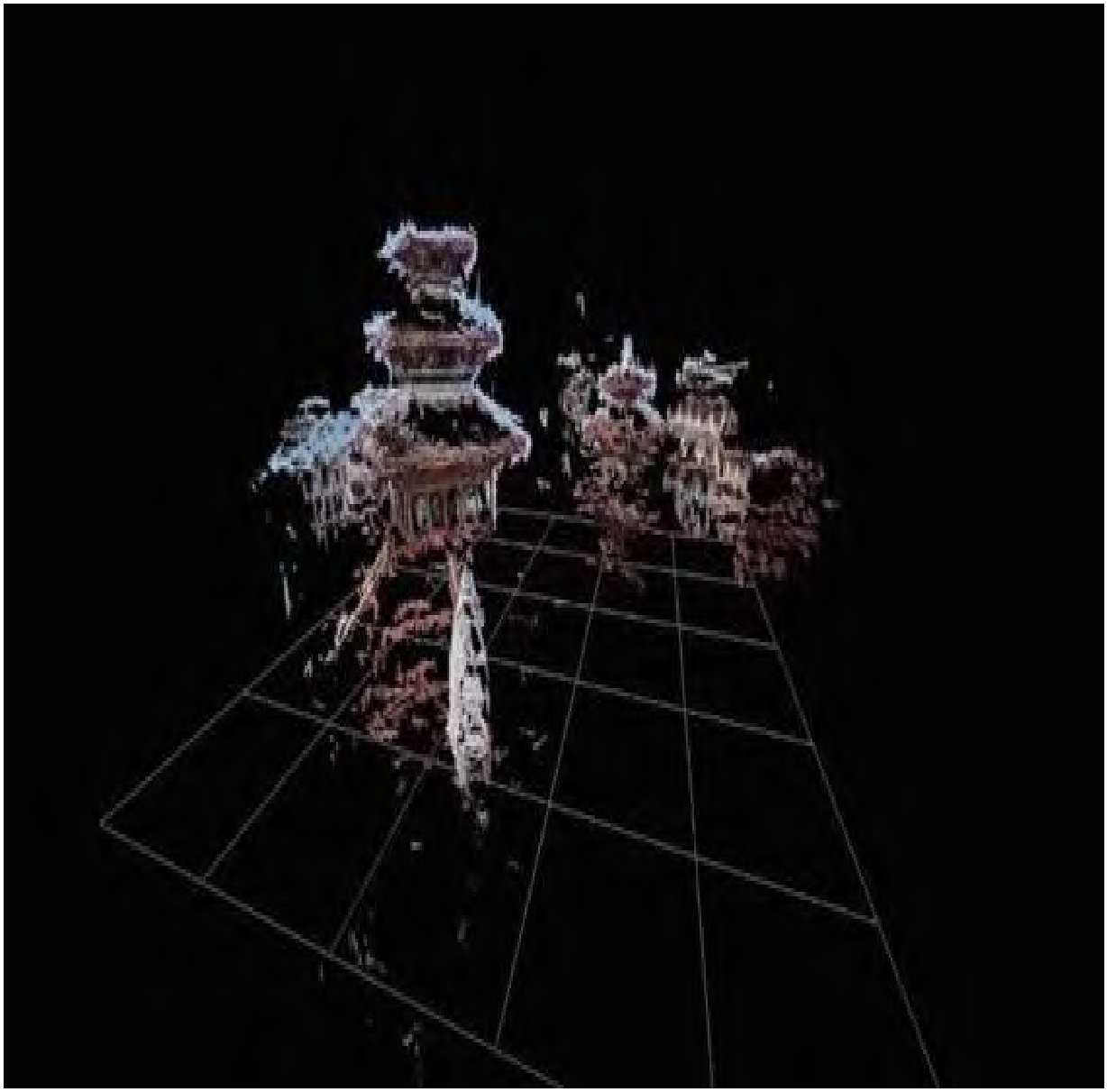


图4-3 玛珠庙（Maju Deval）数字化三维复原示意图.[2].
注：使用手机百度或智能革命App扫描图片可见AR效果。

2016年底，上海虹桥机场出现两架客机仅差3秒险些在地面相撞的事故。塔台的调度和预警功能在这一事件中没有发挥出来。所幸飞行员没有坐等塔台指令，而是紧急处置，避免了一起重大事故的发生。这件事也再次提醒我们，完全依赖塔台人工指挥的模式终究难免有疏漏。

林元庆在与民航部门的交流中了解到这样一个细节：机场塔台工作人员为了了解跑道路面情况，每4小时就要派人去检查跑道。这个岗位专业需求度低、薪资低、劳动强度大、人员流失严重，完全可以用人工智能取代：在停机坪附近安装摄像头并结合人工智能对跑道环境进行实时三维重建。飞机、行李车、机场勤务车和所有人员的动作都能实时体现出来。除此之外，跑道上意外掉落的零件和所有异物也能在第一时间被发现，并不会出现遗漏。这种系统建立的准确度、可预测性和安全性都远远高于人工检查。

时代召唤中国大脑

2015年全国“两会”上，我作为政协委员提交了设立“中国大脑计划”的提案——由国家投入专项资金主导，尽快搭建全球最大规模的人工智能基础资源和公共服务平台，比如建立一个拥有几十万台服务器的大型人工智能平台，支撑各个计划参与方的数据调用、模型调试和应用开发，高效对接全社会的智力、数据、技术和计算资源，依托统一平台，实现资源共享，促进研发创新。这将是新一轮工业革命的助推器。基础研究的成果应该让更多中国企业受益。包括语音识别、图像识别、自然语言的理解、多语种的翻译，甚至无人驾驶汽车、无人驾驶飞机、智能制造方面的机器人都可以在这个平台上进行各种各样的创新和实践。

这件事如果只是百度来做，可能只能提供几万台服务器；如果由国家主导投入，那就是几十万台服务器。平台大了就可以降低成本，鼓励更多的创新。国家持续、稳定地大规模投入，让一大批企业成长起来，随之而来的就是越来越多的创新，从而奠定未来10年、20年甚

至更长时间里中国在全球创新领域的地位。这是我一直以来的想法：我不在乎华尔街怎么看，我一定要把这事儿做成。

然而任何一个超级工程都可能面临争议。

2016年9月，一场高能物理界的事件意外掀起舆论飓风——“超大粒子对撞机之争”从学界延伸至社会。普通民众都开始关注起“粒子对撞机”这一深奥的物理名词。

粒子研究的重要性在科幻小说《三体》中表现得淋漓尽致——外星智慧生物为了阻止地球科技进步，利用量子纠缠原理创造出拥有十一维度形态的“智子”，发射到地球上。以光速运动的智子能同时干扰人类的所有粒子对撞机，精确破坏粒子对撞结果，锁死人类的基础物理研究，将人类科技禁锢在一个较低的水平上。

“超大粒子对撞机之争”从两位局外人——美国数学家丘成桐和哈佛大学物理学博士王孟源的论战开始，历时3个月。还惊动了诺贝尔物理学奖得主杨振宁和中国科学院高能物理研究所所长王贻芳。

面对中国要不要建造超大型粒子对撞机的争论，反对者指出，建造对撞机需耗资数千亿美元，电能消耗堪比一座大城市，但收获的结果却极为不确定，可能沦为一台物理学者的“大玩具”——美国搁置了类似计划，欧洲大型强子对撞机[3]成果寥寥。所以中国为什么要开建？

支持者则认为，研究上帝粒子“希格斯玻色子”[4]非常重要，这将解开“宇宙如何诞生”的天问。美国和欧洲的放弃和无为正好为中国提供了机会窗口。作为崛起中的大国，中国理应承担理论物理前沿研究的责任。

最终，这场争论以超大粒子对撞机未获得“十三五”规划审批而暂告段落。

新中国成立半个多世纪以来，从不缺乏大计划。从改革开放初期的“863计划”中第一次提到“智能计算机”一词至今，中国科学家用40年的时间逐渐追赶上先进国家的步伐。鲜有人知的是，在没有超级计算机、没有大数据的年代，中国人工智能是从大学实验室录制音库起步的。正是前辈科学家的坚持，中国经济实力与科技实力的与日俱增，互联网企业才能在风调雨顺的中华大地上迅猛成长，集涓成流，浩然成势。

此时，百度建立于中美两国的人工智能研究室里，超过1300名不同族裔和国籍的研究人员正夜以继日地对数百个相关项目进行处理。他们的成果都将汇入百度大脑。这些研发者就像是当年“曼哈顿计划”中的1000多位科学家，像是今天欧洲核子研究中心的3000多位科研人员，他们做着超前于时代、暂不为人理解的工作。改进算法、建模升级和分析处理，百度大脑的研究人员正如粒子加速器中飞奔的粒子一样，酝酿着一场智能革命。

中国大脑计划不同于超大型粒子对撞机。后者占地巨大、能耗惊人（欧洲的粒子对撞机运行需要1200万千瓦电力），开发中国大脑却是一项自下而上、水到渠成的工程，国家无须豁出血本，需要的仅是方向与决心。

“规模经济”是中国产业成功的基本因素。在我们这个拥有超过13亿人口，7亿网民和数以千万计工程师、科学家的大国，海量的数据、充沛的人才、丰富的企业案例、各种各样的应用场景像流水一样奔腾。如果不能物尽其用，错过这一波智能浪潮，而将科技制高点乃至国家安全拱手相让，才是真正的“浪费”。正如互联网海量数据催生了Hadoop（由Apache基金会所开发的分布式系统基础架构）、Spark（加州大学伯克利分校的AMP实验室所开源的类Hadoop MapReduce的通用并行框架）等流数据处理技术一样，人工智能已经在中国各地分散发展，犹如一个个神经节点，以其脑波促进中国大脑的到来，这是时代的召唤。

中国大脑，中国气派

斯坦福人工智能实验室负责人、全球著名图像识别数据库ImageNet创始人李飞飞这样描述：“从科学到科技再到产品，就像一个4×100的接力赛，每一棒都有它特别的功能，学术界应该算是这个4×100接力赛的第一棒，工业界和实验室是第二棒，产业化、投资是第三棒、第四棒。”

为了这最关键也最刺激的最后两棒，2013年1月，欧盟曾宣布投入10亿欧元，旨在用巨型神经网络计算机模拟整个人类大脑。2016年10月初，美国连发《国家人工智能研究与发展策略规划》和《为未来人工智能做好准备》两份报告，制定人工智能研究与发展策略规划。这一年，中国官方的报告中也频频涉及人工智能。2016年5月，国家发改委、科技部、工信部、中央网信办联合发布《“互联网+”人工智能三年行动实施方案》。2016年8月，国务院《“十三五”国家科技创新规划》发布，人工智能再次成为核心看点。

人工智能是普世的吗？当然。这是目标，但这种普世注定不是某个单一国家能给予的，人工智能的发展必然像生物进化一样充满多元竞争和地方特色。值得玩味的是，美国科技巨头在人工智能赛道上竞逐时，却意外地在中国遭遇水土不服。

2016年末，IBM的医疗机器人Watson落地山东。这个在美国大获成功的机器人，一来到中国“打工”就被语言绊住了。IBM在汉语素材上欠缺积累，让这次合作从一开始就出现了裂痕。Watson这位“技术大神”能够听懂世界上多种语言，却对上海话、广东话、闽南语无能为力。

这不是一个孤例，由东西方生活、思维、文化差异造成的隔阂，历史上从不罕见。

同样是搜索，中外网友的关注点全然不同。在2015年谷歌人物热搜榜上，排名第一的是NBA球星拉玛尔·奥多姆（Lamar Odom），第二名是女子格斗家隆达·罗西（Ronda Rousey），第三名是美国电视节目名人凯特琳·詹纳（Caitlyn Jenner）。这几个人物恐怕只会出现在极少数中国人的搜索框中。而在2015年百度人物热搜榜上，排名靠前的是金星、王思聪，顿时“接地气”许多。

在硬币的另一面，微软聊天机器人TAY在推特上玩耍了24小时，就学会满口脏话、种族歧视、性别歧视，这倒是非常的“美国特色”。

显然，在多元化的网络世界里，没有任何一个国家、机构能“独当一面”。只有不同文化、经济、政治背景的基因进行竞争，才能对网民的需求做出全面恰当的回应。

从国情来看，中国发展人工智能较欧美国家有着更强烈的内在驱动，这种驱动来自民间。百度搜索统计显示，有关“服务”的搜索请求数量始终在迅速增长：2014年比2013年增长了133%。2016年在基数更大的情况下，仍有153%的增长。

在移动互联网领域的应用上，中国比美国更普及。中国网民早已习惯于诉诸互联网寻求服务。如今中国每100张电影票中就有55张是通过网络下单预订的。与之对应的是，互联网服务在美国电影行业的渗透率只有20%，也就是说100张电影票中只有20张是网上订票。比如，中国餐饮行业的互联网渗透率是2%时，美国只有1%。

中国之大，之特殊，要高效实现中国网民的服务需求，人工智能创新就是一条必由之路。

驱动的另一个源头则是产业焦虑。中国制造业成本上升很快。放眼全球，工业越来越自动化和智能化，高端制造业可能会回到欧美，而低端制造业已开始流向越南等国。如果不在短时间内完成产业转型，中国制造业将面临“空心化”的困境：高端制造业、低端制造业都将流出中国——这个转型，能不依赖人工智能这一环吗？

尽管局势紧迫，但中国的实力值得看好。中国企业的执行力、中国政府的支持力度都是新兴产业的强力后援。

如果说网民的需求是“天时”，企业与政府的合作是“人和”，数据则是中国大脑发展必不可少的“地利”。在这一领域，中国更是得天独厚。

庞大的人口规模、复杂的社会环境和面向不同应用场景的互联网企业，汇合收集这个数据全集意义非凡。可以预见，不久的未来除了个人数据，依托于公共环境或者政府背景而产生和积累的数据，如汽车注册信息、学籍学历、犯罪记录等，将以加密的方式形成个人基础电子档案。企业与市场通过服务输出获取的数据，如信用卡账单、消费记录、网站浏览偏好、惯用手机品牌等，则将以用户授权的方式再次转化为服务回馈给使用者。

这里的“使用者”，不是会编码、能建模的“码农”博士，而是普通的公众群体。让更多的人也能便捷地使用智能设备才是真正意义上的科技福利。

当年，美国“阿波罗登月计划”带动大批企业成长创新，美国军方用于“冷战”对抗的ARPAnet（美国国防部高级研究计划局组建的计算机网）衍生出了互联网。中国大脑这个充满想象力的超级工程对中国经济的贡献，也远不止是涌现一批明星科技企业和科技成果，而是将切实为整个经济和社会的转型提供服务和动力。

2017年，百度获批筹建深度学习技术及应用国家工程实验室，由百度研究院院长林元庆、百度深度学习实验室杰出科学家徐伟，清华大学的张钹院士和北京航空航天大学的李未院士组成团队班子。百度将与清华大学、北京航空航天大学、中国信息通信研究院、中国电子技术标准化研究院等共建单位一起，将企业优势资源整合，建设“国内领先、世界一流”的深度学习技术及应用研究机构，从研究突破、产业合作、技术成果转让、人才培养等方面提升我国人工智能领域整体竞争力。

这将是创世纪的超级工程的开篇。

人文主义画家米开朗基罗完成了雄伟的壁画《创世纪》，其中有这样一幕：上帝之手触碰亚当指间的那一瞬，智慧的启蒙就此产生。这幅壁画中上帝的袍服宽大张扬，最近几十年，有人指出上帝袍服的形状其实是一个人类大脑的解剖图。在这幅壁画里，米开朗基罗悄悄藏进了启蒙的密码——上帝就在人类自己的大脑中，是人类自己启蒙了自己！



图4-4 《创世纪》局部图像
资料来源：http://baike.baidu.com/link?url=SyJFcW1lS-vbxAhPwGKX9xdyJ70qYAWrgzeQlQzWWP9-8mf6nTpA2ZjHdvOXtm4TqmWGR97VvSBWZiJ3VNwV61ClSw9aC2eH_9lTjpBHq6FR-bmG6lb_LGVCK6J1rpkU

人类在自我劳作、自我启蒙中发展出自我的智慧。如今，深度学习神经网络也在自我运作、自我调试中创造新的“大脑”。这颗巨大

的人工智能之脑将成为人类文明新的背景，它正是人类伟大活动的体现，并支持人类文明迈向更高阶段。

---

[1] 《圣经·旧约·创世记》第11章宣称，当时人类联合起来兴建能通往天堂的高塔；为了阻止人类的计划，上帝让人类说不同的语言，使人类相互之间无法沟通，计划因此失败，人类自此各奔东西。此故事试图为世上出现不同语言和种族提供解释。

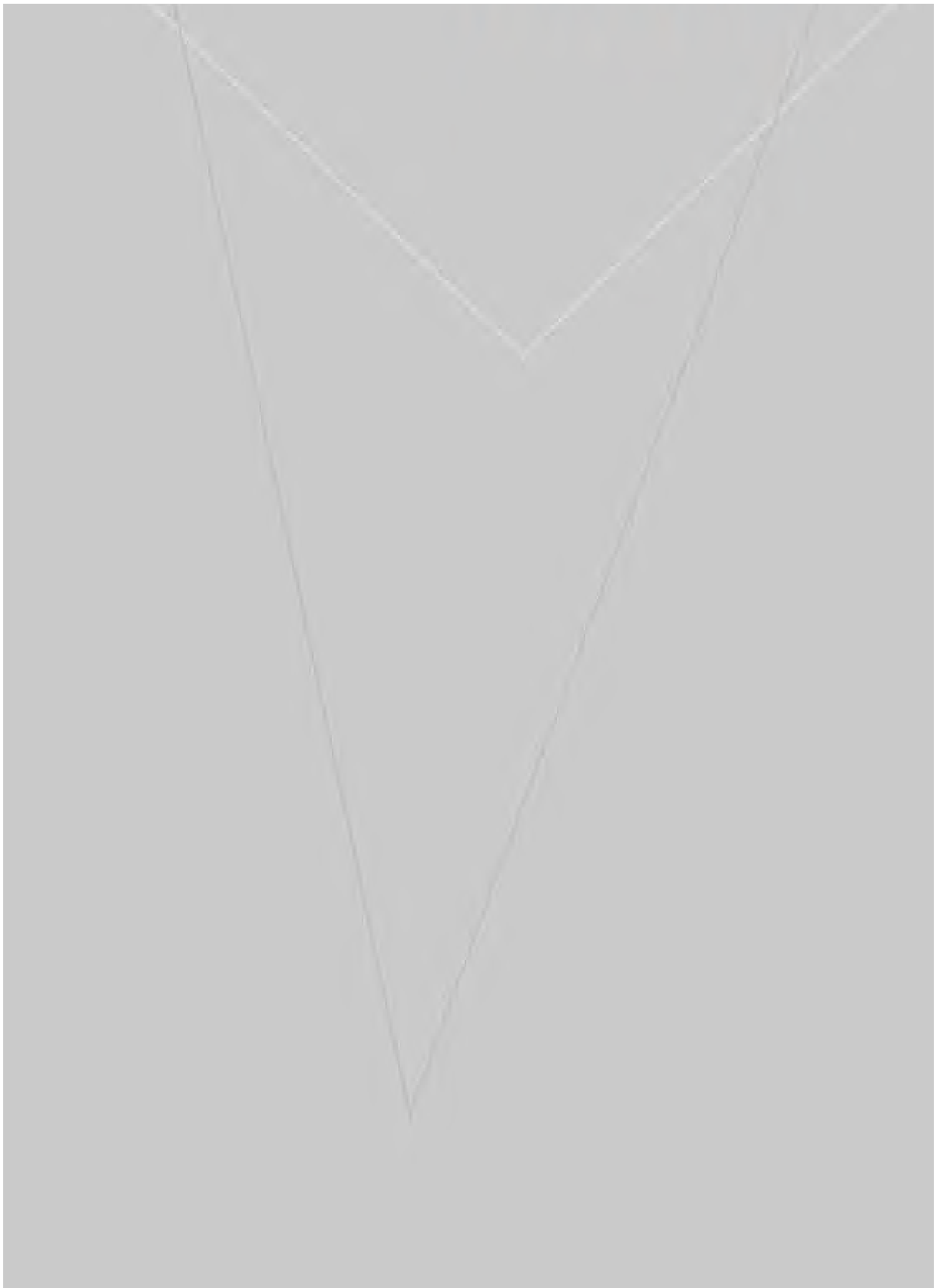
[2] 2015年4月25日，尼泊尔发生强烈地震，不仅造成8000多人死亡，毁坏了成千上万的住房，还使许多历史建筑变为废墟。百度发起“See you again，加德满都”的行动，呼吁网民上传他们拍摄的历史古迹照片，以便百度利用自己的人工智能成像系统，以数字化方式重建这些文化遗产。活动中，百度共收到上传照片超过42000张，完成八个著名景点的数字重建，其中包括位于加德满都的玛珠庙。

[3] 欧洲大型强子对撞机是现在世界上最大、能量最高的粒子加速器，是一种将质子加速对撞的高能物理设备，英文名称为LHC（Large Hadron Collider）。

[4] 希格斯玻色子是粒子物理学标准模型预言的一种自旋为零的玻色子。物理学家希格斯提出了希格斯机制。在此机制中，希格斯场引起自发对称性破缺，并将质量赋予规范传播子和费米子。希格斯粒子是希格斯场的场量子化激发，它通过自相互作用而获得质量。

05

中国智造与文明升级



2016年底的一天，当百度地图召开发布会，宣布每日位置服务突破720亿次时，百度地图事业部总经理李东旻不禁想起十多年前自己第一次听到吕本富做互联网经济演讲时的感受。正在读研的他，形容自己仿佛在迷茫中看见了光，不久就抓住一个机会来到初生的百度实习。

2000年，堪称当时网红级学者的吕本富到处宣讲新生的互联网，在一个论坛上遭到一位上市公司老总的质疑，他回答：“您是一位成功人士。为什么成功？因为在20世纪70年代短缺经济时期，您是第一批注重质量的；在20世纪80年代市场经济初期，您是第一个搞品牌营销的；在20世纪90年代过剩经济时期，您是第一个搞连锁、抓运作的。您的成功是因为总站在时代的前列。但2000年以后的潮流是什么您知道吗？”那位老总仿佛醍醐灌顶，当时就表示要投资互联网。

随着互联网经济的兴起，互联网思想市场也空前繁荣，大大小小的互联网财富宣讲师口吐莲花，吕本富的观点早已被掩盖。可是，这些意见当中有多少能把握住下一个浪潮？

人工智能是新的光，讨论它对总体社会的影响，只靠与互联网经济相关联恐怕还不够。这是一场漫长的社会进化过程的最前沿的变革，不仅经济、科技总体面貌将会改变，国家社会治理层面乃至文化、个人层面也将因为人工智能浸润而发生变革，由此，经济基础与上层建筑的共同变化将使文明样态发生改变。人工智能可以助力大同社会理想吗？可以有助于和而不同的社会秩序吗？回答这些问题，需要努力，更需要想象力。

从勤劳革命到智能革命

费孝通、杉原熏、阿瑞吉等学者在考察中国近现代图强历史中，渐渐描述出一条可以称作“勤劳革命”的道路。“勤劳革命”与西方工业革命重资本投入不同，其表现为：相对廉价但受过基础教育的熟练劳动力在相对小规模的单位工作；劳动密集型产业；推崇勤劳致富的工作伦理等。这条道路并非和工业革命对立，而是工业革命冲击下的一种应对。依靠“勤劳革命”，中国在一穷二白、资本稀缺的条件下完成了工业时代的逆袭，成为世界头号制造业大国，拥有全球罕见的完整产业链，GDP（国内生产总值）增速常年高居世界榜首。

过去三十多年里，中国几乎抓住了所有“崛起”的机遇。在第三次工业革命时期，中国初步完成工业化，补上欠了100多年的功课。20世纪80年代，西方国家纷纷去工业化，中国承接产业转移，成为世界工厂。20世纪90年代末互联网起飞，中国终于和美国站在同一起跑线上。如今移动互联网的发展中国甚至稍胜半筹。中国重新崛起的辉煌一部分要归功于全球化。加入世贸组织让中国商品更容易销售到全世界。在各个产业领域，中国人学习、消化先进国家的先进产业技术的速度很快。凡是中国人学会制造的商品，很快在全球范围内该商品的价格就会降下来。因此中国也会常常被指为“搭便车”。但是现在，当中国已经成为全球第二大经济体，越来越多的人认为中国无法再完全延续过去的韬光养晦式跟随战略，中国也需要担当引领者的责任，不仅要为全世界创造经济高速列车，也要提供技术指引乃至文明典范。

同时，“勤劳革命”自身也遭遇危机。人口老龄化、用工成本上升、大城市病、新兴中产阶级对环境污染的焦虑、国际竞争加剧、贫富差距扩大引发的不满等都在困扰这个国家……如果方向不对，“勤劳”也无法引领潮流。中国人渴望找到新的高速发展路径，如果方向明确，中国人愿意再一次发挥“勤劳”禀赋去学习、追赶。归结到产业角度，整个国家需要淘汰和转移落后产能，升级产业和消费层次。

在升级方面，中国也并非像一些舆论渲染的那样缺少创新能力，只能聚焦低端产业。“勤劳革命”的表述本身也多少掩盖了中国人技术革命的能力。中国的科技实力在大幅度提升，即便在高精尖领域也紧随为数不多的先进国家。2015年8月，中国科技大学的袁岚峰博士撰文，从《自然》杂志指数、全球五大专利局数据、科研投入数量、先进项目状况等多方面论证，中国整体科技水平正在以加速度逼近头号大国美国，稳居世界第二位。此文被多方转载刷屏，引发科技界和民间热议。

2016年10月，中科院与Clarivate Analytics公司（原汤森路透知识产权与科技事业部）联合发布了基于论文大数据分析的《2016研究前沿》，指出：“中国在前沿引领度方面与美国差距较大，与英国竞争激烈。潜在引领度方面，中国全面超越英国，位居世界第二，显示出强劲的后续发展能力。中国共参与68个前沿方向，其中30个领跑全球，与第一名美国总体上还有较大差距，但是在很多领域已经反超。”

在产业方面，中国的技术型大企业崛起，从工程机械制造到电信领域，顶尖企业具备较强的竞争力，能在全球范围与跨国巨头们一较高下。从劳动密集型产业向技术、资本密集型产业转化，虽然大有进步，但与先进国家差距仍然明显。以工业机器人为例，据国际机器人联合会（IFR）统计，2013年，中国制造业从业人员机器人保有量仅为25台/万人，而世界平均水平为58台/万人，其中韩国是396台/万人、日本是332台/万人、德国是273台/万人。对于机器人应用最多的汽车行业，先进汽车生产国的工业机器人使用密度均已达到1000台/万人，而中国仅为213台/万人。不过，也是从2013年开始，中国工业机器人销量在全球占比达20.52%，首次超越日本成为全球第一大工业机器人销售国。根据2016年的数据，中国的工业机器人使用密度已经达到49台/ 万人。

中国在人工智能领域的突飞猛进令世界瞩目，据《华盛顿邮报》称，中国在深度学习领域发表的论文数量已经超过美国，质量也不落后。在应用领域更是丰富多彩，2017年美国CES（International Consumer Electronics Show，国际消费类电子产品展览会）上，中国企业数量超过参展企业总数量的1/3，多项中国人工智能产品获得展会官方最佳创新产品奖。

在美国，资本主导下的高新技术产业在奥巴马任内高速发展，为全世界输送了技术，却也加大了本国分裂。制造业空心化使得人工智能等高新技术无法在本国充分落地消化。而中国遍地成规模的制造业、服务业，加上每年高校培养出的百万级工程师与海归人才，使得人工智能的原力正在获得释放。

三次技术浪潮冲击下的第一制造大国

焦虑已经持续了很多年。在20世纪90年代末，业界就开始流传这样的顺口溜：“如今冰箱都能制冷，彩电都有影儿，PC都能定制，ERP谁都不太懂。”几句话言简意赅。前两句反映了工业化的困境。家电是中国制造的标志性产业，先后经历消费和产能的井喷。企业在早期都曾被质量问题困扰。少数企业致力于提升技术和管理，脱颖而出。当“制冷”和“有影儿”这样标准化的功能需求趋于饱和，质量过硬意味着生命周期长、替换率低，市场也就陷入停滞。

PC虽然都能定制，但仅限于CPU主频、内存、硬盘容量等性能，这由传统PC模块化的结构所决定，是一种相当初级的定制。与后来的智能手机支持应用、内容定制，支持数据生产、分享和反馈相比，不可同日而语。而未来的人工智能终端，其开放性又将远超智能手机。

ERP曾经被视为企业信息系统的核心，如今也已看见衰落的景象。尽管ERP的模板开发相当于对大量企业的数据进行了人工的深度学习，找出基本模式，还可以针对每个用户的特点进行二次开发。但本质上仍然是一种集中开发的模式，弹性不足，部署周期长，和业务有脱节，这是造成“谁都不太懂”的根源。ERP跟不上近年商业环境的变化，逐渐被云计算所替代。云计算不只是一种技术，同时也代表一种理念。对应着云（分布式）计算，企业也要从传统的集中管理转变为分布式管理。

在宏观层面，制造业正在边缘化。约翰·奈斯比在1982年写下《大趋势》，其中指出：“日本已经取代美国成为世界工业领导国……日本是第一，但只是一项衰退的赛事的新任世界冠军。”奈斯比认为日本也受到一些新兴市场的挑战，当时中国处于改革开放初期，还没有进入他的视野。20年后，中国成为工业化竞赛中最新、也可能是最后一任世界冠军。

就在三百多年前，世界上绝大多数人口都在田里耕作。没有人能想象得到，很快人们将涌入城市和工厂。今天的美国，农民数量已经不足劳动人口的1%。一百多年前，相似的一幕在更高层次上演，人们开始涌入写字楼。七十多年前，美国办公室白领数量超过了流水线上的蓝领工人。二十多年前，美国企业界大量裁撤白领岗位，很多人主动离开了朝九晚五的办公室。传统产业的组织模式正在发生巨变。

人们常说，美国的今天就是中国的明天。阳关三叠的长周期也正在中国发生，区别在于时间大为压缩。三十多年前，中国大部分人口还在田里耕作。二十多年前，中国才开始形成白领阶层。最近几年，白领群体的职业前景达到某种瓶颈。

这是新的不平衡规律。时间压缩更突出的表现是，在美国按时间展开的进程，在中国是同时发生并按空间展开。在互联网创业大规模展开的同时，大批的新工厂和新办公室投入运营，招募了大批农民

工、蓝领和白领。这种情况造成的局面错综复杂、波澜壮阔，处理起来更考验举国上下的智慧。

另一位未来学家，《第三次浪潮》的作者阿尔文·托夫勒在2001年的中国之行中评价：中国包含三个世界，第一次浪潮覆盖大约9亿农民，第二次浪潮覆盖大约3亿市民，而第三次浪潮，据托夫勒当时获得的国家计划委员会（后来重组为国家发展和改革委员会）数据，只有1000万人。中国的发展主题就是改变三次浪潮人口之间的关系。

今天，人口格局已经大为改变。国家发布的数据显示，2016中国城镇化率已达57.35%。城镇常住人口达到7.7亿。“十二五”时期，城镇化率年均提高1.23个百分点，每年城镇人口增加2000万人。

第三次浪潮吸纳了大量就业，但和第二次浪潮的边界是模糊的。互联网产业本身就同时包括本属三个浪潮的世界，以百度外卖为例，公司拥有大量外卖配送员，相当于传统蓝领；其次是运维人员，相当于传统白领；还有一个小而精的技术部门，包括人工智能团队，相当于顶层。就公司内部来说，管理文化差异如此悬殊的三个员工群体，是个不小的挑战。从外部来说，这正是中国发展道路的特点，需要勉力平衡这种反差与混合。

中国从第一次浪潮到第二次浪潮的跃迁过程，在2010年前后已经迈过“刘易斯拐点”。[1]

中国的部分低端制造业已经转移至越南等人工成本更低的国家。印度也雄心勃勃地计划“成为全球制造中心”。莫迪政府于2014年9月向全世界发布印度制造新政，包括提供一站式服务，改革劳动法和税收，简化审批程序，以吸引各国在印度投资设厂，增加当地就业机会。印度的多时空混合特质比中国更严重，英语理工精英与落后种姓制度比肩继踵，能否成为中国制造的有力对手，尚需拭目以待。

另外，欧美等老牌制造业强国也在努力，希望能够百尺竿头，更进一步。德国联邦教研部与联邦经济技术部在2013年4月的汉诺威工业

博览会上提出工业4.0（第四次工业革命）的概念，得到官产学各界广泛认同。工业4.0已经上升为德国的国家战略，在全球范围内引发新一轮工业竞赛。工业4.0旨在提升制造业的智能化水平，利用物联信息系统（Cyber-Physical System，CPS）将生产中的供应、制造、销售信息数据化、智能化，在价值流程中整合客户及合作伙伴，实现快速、有效、个性化的产品供应。

美国正在试图逆转战后去工业化的趋势，让部分工厂回流本土。作为大国，局势错综复杂，存在各种力量博弈，如特斯拉这样的无人汽车工厂无助于就业，美国中下层人均收入缩减，很多回流的工厂没有回到铁锈地带[2]，而是部署在人力成本更低的南部等。新总统特朗普连续在推特上炮轰丰田、通用和福特，威胁它们把汽车工厂搬回美国，更加凸显了某种时空错乱的怪异特性。但无论如何，美国重振制造业的行动势必对中国制造业形成压力。

美国制造业提出的主要创新亮点在于3D打印。虽然一度以黑科技的面目出现，但3D打印的关键不在于技术，而是反映了制造业从大规模标准化生产转向大规模定制生产的趋势，最终反映了客户需求个性化、自组织化的趋势。



图5-1 特朗普在推特上炮轰汽车制造商

面对这样的上下夹击，中国制造业的地位略显尴尬。尽管人力成本还有向中西部转移的空间，但中国整体正在逐渐告别低成本优势，迫切需要将竞争力转移到生产率和知识经济上。

张亚勤因为负责百度云计算工作，经常要和企业打交道。他认为，在工业互联网和智能制造方面，中国落后美国比较多。这么多年来，美国大的行业基本都实现了流程化，垂直行业的大部分公司都用上了ERP，为此投入了上万亿美元，使得流程变得IT化、软件化。制造业不管是高精机械也好，流水线也好，自动化开始得都很早。中国在这个方面整体上相对落后。中国企业，尤其是很多小企业，没有统一的工作流程，IT程度比较低。

第二次浪潮到第三次浪潮的跃迁，即使是最前沿的美国也才刚刚起步。某种程度上硅谷的创新仍然以技术为主，与第二次浪潮中诞生的产业相对独立，互相的结合程度可能还不如中国。“物联网”概念的发明者凯文·艾什顿就极其看好中国的互联网与物联网的发展。张亚勤认为，就像消费互联网刚兴起时一样，各国都处在同一起跑线上，美国可以做，中国也可以做，中国甚至能比美国做得更好。现在当人工智能来了之后，原先的一些领先优势就不太重要了，甚至由优势变成了劣势。如传统企业花很多钱买Oracle的数据库，买IBM的ERP和数据库。这么多年来，基于传统IT技术的公司，产品越做越重，硬件也好，软件也好，越来越复杂。IT的投资、运营成本和人力成本越来越高，很多企业不堪重负。云计算是新的平等起跑线。云做的事情是把复杂性移走，让做云的公司帮企业做IT，不是像传统的ICP（互联网内容提供商）服务公司，派几百个人到你公司去做，而是提供虚拟化服务。例如，过去是在家里建电站、打水井。现在是提供电网、水网，企业只要打开水龙头就可以得到水。企业不再需要部署那么多服务器，不再需要那么重的ERP系统，只需要接入网络，就可以连接所有的服务——计算、存储、数据库。于是商业端服务也消费者化了，方便快捷。

昔日制造大国：人为物役

伴随人工智能产业的热潮，乐观和悲观的观点再次交锋。20世纪人工智能刚在想象和理论中成形时，争议就开始了。人们的恐惧和希望都来自现实，反映了贯穿三次浪潮的人与技术的纠结关系。简略回顾一番技术与工具的历史有助于我们反思制造业与人工智能的关系。

商品的个性化定制背后站着生物进化的历史。进化论告诉我们，生物通过改变性状来适应环境。这个迭代过程非常缓慢，并且后天习得不能写入程序（DNA）。但是在工具领域，人类可以超越生理的进化，通过工具的改进、迭代来融合后天习得的知识和本领，改造环境、改变自己。

制造工具甚至成为人的本质定义。虽然有些动物能在某种程度上使用“工具”，如人类的近亲灵长类，但这些动物使用工具仍然属于条件反射。人制造工具并非回应眼前的挑战，而是包含了对过往经验的“深度学习”以及对未来的预测。如石器中常见的一类石斧，有砍、削等多种功能，适用不同的情境，成为我们祖先随身携带的工具。

人工智能目前刚刚走到自己的石器时代，如果说以前的计算机工作模式是条件反射，只能完成程序预设的任务，今后人类则致力于教会它自己开发“工具”。

我们就以汽车制造业来展开说明。科学管理学派的祖师泰勒对装卸工种经过反复试验(深度学习），得出单人的最优负重为21.5磅(9.75千克），因此为每种物料设计专门的铲子，以保持这一容量。可想而知，每个工人的力量和耐力存在差异，这个最优负重是泰勒生活的时代工人平均身体素质的体现。这时在技术上，铲子仍然是工人的工具，但在管理上，管理者通过标准化的铲子，把所有工人组成一个巨人，单个工人是这个巨人的人肉工具，也就成为铲子的工具，正所谓“人为物役”。

在汽车工业早期，由每个工人独立完成一辆整车装配，采用自己最习惯的方式。经过长期学习，工人们也能达到相当熟练的程度，但

这并不是效率最高的方式。进化的下一步是流水线。

据说福特公司高管有一次参观屠宰场，受到启发，也采用流水线装配汽车，将整车装配分割成最小、最简单的单元，每个工人只负责一道工序。1913年福特公司开发出世界上第一条流水线。汽车底盘和零件分别通过传送带送到工人面前，工人无须走动，节省了取零件的时间。手工生产一辆汽车需要728工时，流水线缩短至12.5工时以下。汽车价格大幅下降，让平民可以消费得起。T型车从1908年投产到1927年共生产了1500万辆。

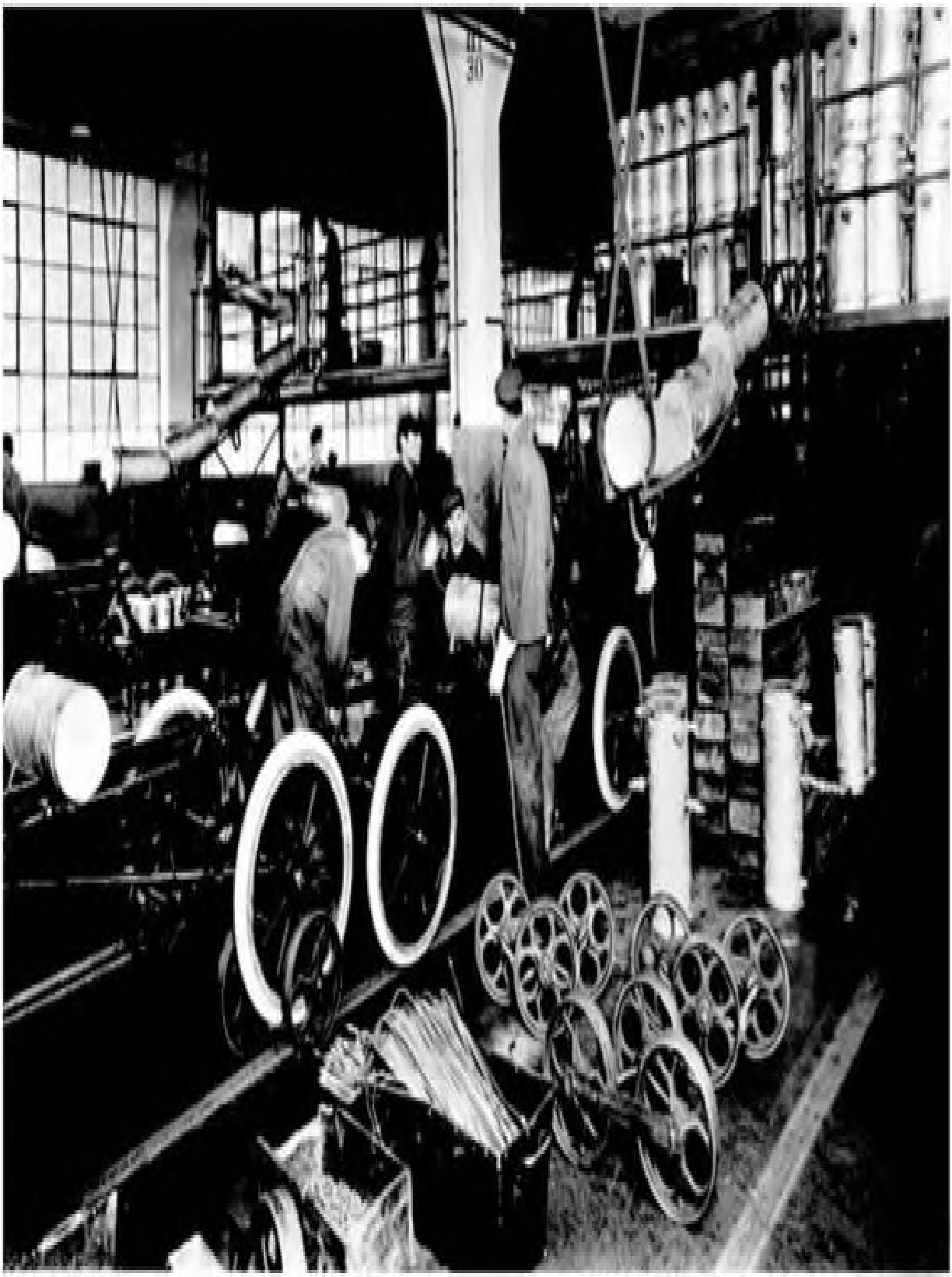


图5-2 福特公司早期的汽车流水线

早期汽车生产流水线共划分为7852个工种，其中949种需要身体强壮；3338种只需普通体格；3595种可在正常体力以下，其中又有715种可由独臂者完成，2637种可由独腿者完成，10种可由盲人完成，2种甚至可由双手残疾者完成。这种详细划分的标准我们可以理解为较为原始的特征化和数据化，蕴含着今天的数据标签化技术，而标签化数据正是人工智能的燃料，可这毕竟还是工业时代的产物，无法挣脱人为物役的锁链，也无法满足更高层级的人类需求。它首先将遭遇“多样性”的瓶颈。

只有新工业自动化才能契合人类多样性

多样性是生物的基本特征。普通生物的多样性一般表现在物种之间，即同一物种的性状趋同。但在进化的高级阶段，在智慧生命身上，多样性表现在个体之间，尤其是表现在与智慧相关的器官脑和手。

可以想象人类基因中有个骰子，随机掷出不同的DNA代码。非孪生的兄弟姐妹，天赋也会有显著差异。人们从来都能感受到彼此的不同，只是囿于手段限制，无法淋漓尽致地表达这些不同。过去人们想出众多的性格分类法，以星座学最为流行，就是渴望表达区分的尝试。生物体的个性区分曾令达尔文非常困惑，但却为经济的分工协作和消费的多样化奠定了生物学基础。

传统工业是如何满足多样性需求的？关于产品个性，工业先驱亨利·福特有三句话，昭示了旧工业范式下企业与消费者的关系：

“我问人们想要什么，他们总是说更快的马。”

“顾客可以喜欢任何颜色，只要是黑色。”

“汽车价格下降1美元，增加1000名顾客。”

用户对汽车有什么具体需求呢？薛成曾在《中国企业家》撰文分析这个问题。在汽车问世之前，除了少数工程师，普通民众对此一无所知，工程师只能就已知需求(马)来增加数量(更快）。按马斯洛的需求层级理论，越低级的需求，越容易量化，共性越大，相应市场规模越大。亚当·斯密在《国富论》中指出，规模是分工协作的前提。越高级的需求，如果要增加产品，相应的分工协作就越复杂，会降低效率，损失规模效应。对于汽车生产来说，首先是为容易满足的需求扩大生产。“价格每下降1美元，增加1000名顾客”。如果为增加色彩价值每投入1美元只能增加200名顾客，将损失800名潜在顾客。所以福特只提供黑色轿车，却优先提高产量，降低成本。

但是在消费层面，当汽车的顾客被创造出来，他们的新需求就同时产生了。德鲁克的《管理的实践》开宗明义，企业的目的是创造顾客。当生产力扩张，消费需求将被释放或者创造出来。

泰勒相信工人为钱工作，福特也说过：“工资解决了9/10的精神问题。”但从1924—1932年，由梅奥为首的心理学家（而不是管理学家），通过霍桑工厂的实验发现，工人的工作动机要复杂得多。1960年心理学家道格拉斯·麦格雷戈提出X理论和Y理论，前者把工人看作懒惰的经济理性人，需要鞭子+酬劳来激励，后者认为人有劳动的需求和创造力。

整个旧工业经济体系就像太阳系，消费者和员工环绕企业运转，企业再环绕金融（太阳）运行。这套体系可以对人的多样性需求做出反应，但是以一种中心化的方式运作。

直到今天，技术还在进步，除非创造新物种，产品多样化方向已经趋近感知的极限。一个新的循环往复将要形成。新工业范式将把旧工业范式颠倒过来，不再是技术的突变引导需求的渐变，而是无限地拓展需求层级和种类，以需求的突变引导技术的渐变。扩大内需不应该被理解成旧需求层面上的扩张，而是前所未有的多样化需求被开发

出来。在旧工业范式下，依靠的参照系只有已知技术和已知需求两条小路，人类凭着自己的计算和洞察摸着石头过河。而今天，人类无穷多样性的细分需求，需要依靠人工智能深度学习系统才能捕捉。

曙光在于大数据和人工智能。今天，人类的多样性需求、反馈越来越被数据化，可以无限地产生并被传感器记录。也只有建立在概率和分布式计算方法上的深度学习与智能经济系统，才能从这无限的数据中感知未来的方向。

还是以汽车的例子将这个问题具体化。沃顿商学院的马歇尔·费舍曾拜访汽车经销商，得知综合颜色、内饰、发动机等功能，汽车制造商实际上可以提供2000万种型号。定制需要8周，而90%以上的顾客要购买现货，而且并不知道存在那么多的可能型号。这家经销商只有两种型号现货，本地有10家经销商，假定经销商的型号规模相等，那么在本地市场上只能提供20种型号。费舍因此将渠道比作沙漏的颈部。

这项调查发表于1997年的《哈佛商业评论》。同年亚马逊上市，提供了解决方案——网络货架无限长。除了存储空间，更重要的是信息匹配，通过计算机网络即时记录和分析用户的行为，推测其兴趣，这就是所谓的“用户画像”。有了这个方法，就可以把无限多的产品型号与无限多的用户需求精准匹配。

复旦大学中国研究院的余亮在虎嗅撰文认为，“用户画像”被很多人误解为是对用户群体的描述，比如90后群体的消费特征等，甚至一些互联网商业公司也在宣传中如此使用“用户画像”。然而正相反，基于人工智能的“用户画像”恰恰是对个人的描述，可以为每个人贴上无数标签，精细追踪个人的需求。比如所有以算法推荐资讯的应用程序，都是“伪装”成新闻客户端的用户个性收集器，以此向用户推送个性化的信息和广告。这种方法同样运用在实业领域，物联网与人工智能可以给生产工序和消费用户两方面都贴上天文数字的标签，满足最精细化的需求。比如3D打印，实质是可以根据多种多样的

需求而灵活变化工序，不再依靠实物模型，仅仅根据计算机建模就可以“打印”出新产品。

工业自动化的内涵因此被改变，不再是针对固定需求的自动生产，而是自动根据需求变化调节生产、流通和分配。控制论领域大奖诺伯特·维纳奖的首位华人获得者，中科院自动化研究所副所长王飞跃指出：“工业自动化将向知识自动化转移。”新的生产过程将具备如下特性：对人类生活多样性的自动跟踪；知识的自动化习得；工具的自我复用和进化；对社会管理的自动优化；生产流程根据知识进行自动调整并生产出新知识……以此形成新的循环往复规律，革新工业经济范式。这个过程将挑战过去生产方式的一切环节，从生产布局、设计流程、渠道建设直到科层制的企业权力结构。“中国制造”也将因此而蝶化为“中国智造”。

走向物联网与精细化生产

知识化、自动化、精细化的生产方式的基础在于物联网。

2016年7月，软银公司斥资243亿英镑收购微芯片巨头ARM公司。软银总裁孙正义认为物联网将会引领下一轮技术爆炸。2018年，物联网设备的数量将会超过移动设备；2021年，全球将拥有18亿台PC、86亿台移动设备、157亿台物联网设备；2035年，数据量将会增长2400倍，从1EB增长到2.3ZB；在未来20年，物联网设备的数量将会超过1万亿台。孙正义指出：“物联网与人工智能的关系，正如同眼睛与大脑配合使生物得到进化的关系。物联网爆发即将来临。”

物联网让人与万物接入同一张网络，让人与机器随时互动，人的一切可以数据化的行为、反应，都会引起机器和生产线的波动，各种数据在云端交汇，大量计算通过云服务器完成，再通过产品与服务反

馈于人，循环上升，从而把人的多样性与物质世界联系在一起，彼此催生，共同进化。

《奇点临近》的作者雷蒙德·库兹韦尔认为，人类正处于物联网引领的工业革命时代。3D打印技术是2020年前创新型工业革命的主要推动技术。到2020年，我们的衣服都可能被开源设计，可以直接从云端免费下载。3D打印技术的发展前景目前尚不明朗，不过库兹韦尔的话呼应了前面说的新工业范式对多样化需求的满足。除此之外，从食物到音乐，都可以通过物联网精细化生产，云计算是其内在逻辑。

凯文·阿什顿的观点更为深远，他认为：

我们要区分大众说法中的智能穿戴设备与物联网。能侦测防晒指数的比基尼和能感知是否口渴的智能水杯都不是物联网。物联网不是终端设备，而是一套可以自主学习、自主做出决策的机器体系。

物联网的优势在于计算机拥有各种传感器，能自主收集数据。正如我们的智能手机有GPS或者北斗导航系统，有地图，还有距离、方向、重力、惯性甚至心率传感器，可以把所有信息搜集起来进行处理。快速发展的RFID芯片可以在不耗电的情况下提供数据。RFID的产量早已超过了手机的产量，赋予每部手机、每个设备独一无二的编码，或者说“名字”。通过这种不耗电的RFID系统，订酒店、订车位、吃饭支付都可以完成，这也是物联网的构成部分。未来物联网设备的功耗将极小，手机甚至可以在风中自己充电。

人脑将直接与云端进行连接，从而成为物联网的一部分。这将促进对大脑思维方式的逆向工程解析，不仅能够加深对人脑的理解，还可以反向提高机器智能，更加深刻地体会人的多样性细节。

以往工业自动化仰赖的工业机器人，在物联网时代也将发生形态变革。工业机器人将从物理形态的硬件化向软件化发展，云端虚拟状态与物理端实体结合。机器人软件化和物联网是同一件事情在两个方向上的发展。机器智能无形化，遁入物体，与云计算结合，无所不

在，成为一个巨大的社会机器人。这正是人工智能与物联网追求的境界。

物联网系统与传统的机器系统在知识上有根本区别，在运作逻辑上，是从机械的牛顿机器向智能的默顿机器的升华。所谓牛顿机器就是根据因果规律程序运行的机器，而默顿机器则是根据相关性思维自主学习规律的机器。牛顿机器遵循“大定律，小数据”逻辑，而默顿机器遵循“小定律，大数据”规律，不追求终极大定律，与人脑工作方式更接近，更能呼应大千世界的瞬息万变。

物联网自主收集数据只是基础任务，更高的境界是自主决策。人类提供了算法和训练模型，给机器装上传感器，让机器在各自的场景下做出判断和交流，并做出决策。决策会影响世界，从而又收集到新的数据，如此形成循环。

举个例子，百度云计算已经支持多个第三方智能照明项目投入使用。今天的城市户外照明系统虽然灯火通明，但大部分时候都只是寂寞地照射着空旷的马路和天空。物联网照明系统可以在初期收集数据，自动学习照明规律，然后自主优化管理，比如在交通低谷期对路灯进行自主调光后关闭部分路灯。系统实时监控路灯的运行情况，通过“机器学习”对设备的寿命进行预测，精准统计需要更新的部件，从而降低零部件的库存，节省维护费用高达40%。最终把电力消耗降低到传统标准的40%并延长设备寿命。这样，以往的灯火通明变成了生物体呼吸般的灯光节律。

经济效益只是系统收益的一部分。上述照明系统集成了智能灯控、环境感应、无线城市和安防等功能群，开放了众多的API，以供更多的应用接入。由于百度云平台的计算能力和存储能力可以弹性扩展，系统获得了极大的“灵活性”。可以对热数据进行即时分析，并对冷数据做大数据挖掘。这样的系统绝不只是为了照明，而是智慧城市的一部分。它们通过收集数据、自我学习和自我运作，提升机器智能，促生城市大脑。

物联网将覆盖人类的所有生存空间。在某种意义上，我们可以把智能农业也理解为物联网。在一些先进示范农场，每一棵植株都安装了传感器，比如系统会根据每棵植株反馈的信息，因材施“浇”，有的放“施”，大大提高效率、节约资源。每棵植株的传感器更是与系统相连接，把大面积的农作物传感信息汇总到云上，就可以计算出以往单纯依靠农民经验无法把握的农业规律，实现农业革命。



图5-3 Prospera公司使用人工智能和机器学习技术监控农作物的情况
资料来源：http://www.mobilemag.com/2012/09/28/wall-ye-robot-helps-inspect-vineyards-for-bugs-and-other-issues/

人们从传统思维出发，可能会把无人车理解为交通领域的革新。而无人车系统的抱负是超越交通领域，成为城市和城际物联网的载

体。无人汽车绝不是以个体形态出现，而是一个庞大的自主系统，像血管和神经一样连接着其他所有城市系统。无人车本身就是人工智能技术的集大成者，视觉识别、语音识别、自主决策、机械控制等集于一身，是一台运动的数据收集和处理器。在这个基础上，无人车网络把人、车、环境联系在一起，把个人目的和整体管理联系在一起。一旦无人车成规模运行，又会反过来带动技术发展和物联网的发展。无人车上的每一个部件传感器和乘客感应器都联系着生产商、消费者、管理者甚至第三方。设想一下未来，以无人车为主体的陆上交通会与航空、航海交通打通信息，那将是怎样一张上天入地的巨大的物联网。

如果C端（消费者）的精细化生产可以呼应人的多样性需求，那么B端（企业）的物联网则可以精细化掌控社会总体需求。人工智能与物联网同时丰富了生产、提升了效率、减少了外部成本。比如智能农业大大节约水和化肥，无人车系统将减少车辆事故，满足基本出行，减少污染。

智慧能源、智慧交通、智慧生产等，将给世界文明带来史无前例的变化。

呼唤智能政府与智能社会

人类社会在经历过工业革命、两次世界大战和无数次变革与动荡之后，越来越认识到个人的自由、安定和发展离不开政府与社会的高效、公正。随着经济、社会的发展，人类的组织越来越复杂，复杂社会需要新的手段加以治理，政府和社会组织的作用尤为重要。

现代社会通过法律来维持和调节社会关系。但技术尤其是人工智能的发展，使得算法的地位上升，各种自动化管理工具通过算法潜移

默化地调节人类的交往、消费、交通、金融等。在未来社会，律法可能将融合于算法之内。

大规模的数据治理起源于20世纪以来的政府信息化管理。比如1929年由胡佛倡导，在美国建立的非电子化犯罪信息记录系统。20世纪60年代，美国开始建立全国统一的犯罪信息系统，这些数据的用途超出了犯罪记录查询，对劳动力市场甄选工作、福利计划执行都有巨大帮助，从而成为政府治理手段的基石。

而在未来，随着人工智能技术的发展，政府治理模式和法治结构都可能发生重大改变。上海交通大学法学教授郑戈认为：“（目前）法律总体上还是假定责任源自过错，过错损害了法律确定要保护的权利和利益，导致了损害结果，有损害就要有法律救济。基于这种原理，法律总是滞后于损害的，只有当主观过错促生了具体行为，行为造成了实际损害之后，法律才能介入，介入的目的也是恢复此前的状态。互联网的出现和普及改变了人们之间的交流方式和互动方式，而大数据技术的发展则使互联网的潜能发挥到了一个新的量级。大数据科技与认知科学和人工智能的结合使行为主义很可能变为明日黄花。预测性和引导性数据分析可以通过个人化的识别、分析和干预‘植入’意向和行为动机，从而改变法律的作用场域。”[3]

形象一点说，这就是美国电影《少数派报告》中展现的未来景象：政府机构有可能通过数据预测犯罪并提前制止，而非事后追凶。我们可以想到，未来政府的很多管理方式都有必要从追逐式管理变成预测性管理。

英美两国关于人工智能发展的报告都已经提到人工智能对政府治理带来的这类帮助或者挑战。

英国政府报告提出：政府已经在使用机器学习等数据科学技术，这些技术提供了对一系列数据的洞察，从提供数字服务反馈到分析卫星图像。例如，政府可以做到以下几个方面：

通过预测需求和更准确的定制服务，使现有服务（如健康、社保、紧急服务）更有效率，使资源得到最大限度的分配。

使政府官员更容易使用更多数据进行决策，并减少欺诈和出错的概率。

使决策更加透明（可以通过采集过程背后的数字记录，或通过数据可视化支持决策）。

帮助政府各部门更好地了解他们所服务的人群，确保向每个人提供适当的支持和机会。

美国白宫的报告则提出人工智能可以被用于改善刑事司法系统，政府应该推动执法数据和公众数据的充分应用，以便算法系统更好地帮助人类在犯罪报告、治安、保释、量刑和假释决定等各方面减少偏见，做出高效、公正的决策。

美国的科学研究机构正在尝试利用人工智能来解决经济和社会问题，比如利用数据挖掘和人工智能改善失业问题、辍学问题，帮助无家可归者。斯坦福大学的研究人员正通过人工智能技术分析卫星图像，以此为援助贫困地区提供指南。

英国报告提出政府部门可以借助人工智能来预警城市中可能发生的火险。这一点在美国已经实现：

纽约市大约有100万栋建筑物，过去平均每年有3000栋会发生严重火灾。既然每年都会发生，那么与其事后救火，能否事先预测呢？

华裔数据科学家Jeff Chen（杰夫·陈）曾是纽约市消防局的数据分析师。他认为每栋建筑物拥有独特的属性，通过分析就能得知哪些建筑物容易着火。比如低收入家庭的房子更容易发生火灾，而且因为人群居住密度较高，火灾的危害性更大。其他易火因素有：建筑物新旧程度、电路老化程度、消防设施配套情况（消防栓的数量和位置）、有无电梯等。Chen说空置的或者没有安保的建筑着火的概率是

其他建筑的两倍。这些听起来都显而易见，但要作为数据因素全部消化也不容易。

他牵头开发了预测火警风险引擎，利用数理统计方法把建筑物和住户的各种数据加以分析，辅以机器学习技术，用市政数据来驱动引擎，预测不同建筑的火灾风险。这套系统在2013年部署，整合了当地近7500个实时风险因素。纽约消防局利用该系统给出了33万栋可审查建筑的火警危险系数（消防局不检查独栋或者双拼别墅）。

在此之前，消防检查都是随机的。现在，当消防员进行每周例行检查时，系统会生成一份按危险系数排序的建筑清单，指引消防员优先检查容易着火的建筑，大大节省了人力并提高了效率。此外，数据智能分析系统还参与垃圾处理、解决社会治安问题等。

在美国国家层面，智能治理项目比较突出地用于安全工作。除了已成众矢之的的棱镜系统，美国中央情报局亲自投资了很多数据领域的独角兽企业，包括彼得·蒂尔参与投资的Palantir Technologies，另外还有Dataminr、TransVoyant、Geo- feedia、Pathar等。

这些公司产品的原理大同小异，通过社交媒体、地图、传感器和其他各类渠道自动采集社会数据，并整合各类交通、金融公开信息，打通分离的数据库，提供各种数据透视方案而无须用户编写代码。

Palantir Technologies旗下的Palantir Gotham主要用于反恐，这缘于PayPal对欺诈的对抗。它们开发大数据工具，通过匹配用户过去的交易记录和现在的资金转移情况等数据来查找可疑账户并进行冻结。随后它们想到这一技术可以为政府提供服务。CIA、FBI（美国联邦调查局）、DIA（国际情报局）、海陆空三军以及警局等情报机构掌握着成千上万个数据库，包括财务数据、DNA样本、语音资料、录像片段以及世界各地的地图，但要在这些数据之间建立联系并发掘有价值的情报，却相当麻烦。Palantir的创始团队认为，如果由它们建立一个数据分析库，整合相互分离的数据来进行搜索和分析，以提升数据

分析效率，就可以向政府“推销”这项技术。Palantir的主要客户也正是美国情报机构。

Geofeedia可以迅速挖掘新闻发生地点的各类信息。Trans Voyant服务于物流和政府业务，通过各类交通传感器和地图技术收集数据情报，同时整合各地新闻、社交媒体信息、天气报告、卫星云图、旅行警告、犯罪活动信息等，为政府人员提供预测，帮助管理人员运营资产，实时完成数据决策。

科技公司的数据智能服务理念延续了商业逻辑，比如甄选目标客户、抢时间、抢收益、抢在公众前面获得情报并对未来做出预判，这跟用于金融投机、战斗决策和股票市场上的高频交易是一个道理。这是自由市场的丛林模式，而非统筹社会全局的治理主义模式。后者是政府层面必须考虑的事情。

中国具有相对有利的条件。陆奇认为，“中国越来越成为新兴工业的创新大国，有很多创新的地方，而美国保守力量比较强。大变革时代中国反而创新空间大，提供了更好的创新环境。”中国由于政府部门强有力的带头作用，智能地图、安防摄像头、数据管理系统覆盖城乡，中国的智慧城市将成为社会智能治理的先锋。

---

[1] 诺贝尔经济学奖得主、发展经济学领军人物阿瑟·刘易斯分析发展中国家的二元经济，认为当农村剩余劳动力转移殆尽，城乡形成统一的劳动力市场，工资水平将开始持续上升。

[2] 铁锈地带（Rust Belt）最初是指美国中西部—五大湖附近传统工业衰退的地区，现可泛指工业衰退的地区。

[3] 郑戈。在鼓励创新与保护人权之间——法律如何回应大数据技术革新的挑战［J］.探索与争鸣，2016（7）.

06

冲入 AI 无人区：无人驾驶之路



人工智能只有同人类命运紧密相连，直面复杂形势，才能体现出技术的革命性意义。昆德拉说：“负担越重，我们的生命越贴近大地，它就越真切实在。”无人车是紧贴大地的顶级人工智能工程，“艰难”和“颠覆”是它无法逃避的两个命题。它比人工智能诞生更早，却要翻过更多观念和技术的大山才能走到今天，车辙所及，是自交通工具诞生以来的社会秩序。

无人车也叫自动驾驶汽车，是指在没有人工参与的情况下，能够感知环境并进行导航的汽车。1925年第一辆无人车在美国上路时，发明者对它是否需要“智能”还未达成共识。但从此以后，无人车的形象出现在不同种类的科幻小说和电影里。这个梦想如此具体，可是历经几代人的实践仍然没有成真。不过值得庆幸的是，每一个逐梦者的挫折都让无人车的前进之路日渐清晰——这是一条智能进化之路，也是人类社会的变革之路。

在吸引人们踏上征程的未来世界，无人车不再只是通行工具，而是家和办公室之外的第三空间，移动、安全且舒适。人类以最低的成本穿梭在空间之中，以往由低效带来的堵车、污染、停车问题都将大为缓解，酒驾、闯红灯、超速等危及他人人身安全的驾驶行为将不复存在。无人车体系更有望成为全球物联网体系的血脉，甚至汽车作为一种交通工具的社会地位和符号意义也将发生彻底改变。

2015年世界卫生组织表示，中国每年有超过26万人死于交通事故，居世界第一，其中90%的事故是人为失误导致。无人车有望将车祸数量降低到现在的1%。

此外还有巨大的收益。摩根士丹利在一份报告中指出，无人车发展起来以后，仅美国就能够获得1.3万亿美元的收入，相当于美国GDP的8%。这其中有1000多亿美元来自节省下的燃料、2000多亿美元来自减少的堵车成本、5000多亿美元来自交通事故锐减而节省的医疗和保险成本、4000多亿美元来自工作效率的提高。

这些都还只是一个侧面。无人车改变的不单是人与车的关系。一旦车与车、人与人、人与社会都被智能工具连接，带来的将是对现有物质世界规则的重塑。

在崎岖的道路上前行

时间回到1925年8月，人类历史上第一辆无人驾驶汽车正式亮相。这辆名为American Wonder（美国奇迹）的汽车驾驶座上确实没有人，方向盘、离合器、制动器等部件也是“随机应变”的。而在车后，工程师Francis P. Houdina（弗朗西斯• P.霍迪尼）坐在另一辆车上靠发射无线电波操控前车。他们穿过纽约拥挤的交通，从百老汇一直开到第五大道。这场几乎可以被看作是“超大型遥控”的实验，带着对无人车机械化的理解，今天依旧不被业界普遍承认。

1939年，摩天大楼开始在美国的土地上不断出现。“大萧条”后逐渐恢复信心的人们怀揣着对未来的美好愿景。在这一年的纽约世界博览会上，通用汽车搭建的Futurama（未来世界）展馆前排起了长龙，人群争相涌入，希望一探“未来”的模样。设计师Norman Bel Geddes（诺曼•贝尔•格迪斯）向当时的人们展示了他想象中的交通：汽车采用无线电控制，电力驱动，由嵌在道路中的电磁场提供能量。

他在自己1940年出版的《Magic Motorways》（神奇的高速公路）一书中进一步解释：人类应该从驾驶中脱离出来。美国高速公路都会配有类似火车轨的东西，为汽车提供自动驾驶系统。汽车开上高速后就会按照一定的轨迹和程序行进，驶出高速后再恢复到人类驾驶。对这一设想，他给出的时间表是1960年。



图6-1 1939年纽约世博会通用汽车搭建的Futurama场馆



图6-2 Norman Bel Geddes（1893—1958年）最早提出了无人驾驶的概念

20世纪50年代，研究人员开始按照上述设想进行实验。或许是实验让人认清了困难，预言被延后了，他们说这一切会在1975年发生。

在这之后，试图利用铺设在地面上的电线为汽车导航，进而实现无人驾驶的技术探索在各处展开。而英国人把路上的电线改成永磁铁片组成的引导线，他们认为这样做能使控制更加精准，车速更快。“地面轨道派”经历了无数挫折，几乎所有人都隐隐感到，以轨道引导无人驾驶不具备实际应用价值，而且这已经是当时技术条件的“天花板”。不过，这种应用模式倒是在一些送餐机器人的产品中实现了。

1956年，通用公司造出了无人车的实体。它展出的Firebird II（火鸟二代）概念车，首次提出了安全及自动导航系统。钛金属、流线型的车身简直像是一枚直接从科幻电影中开出来的火箭。这只“火鸟”推出第三代时广告语是：“想要坐着放松一下？好，设定好

想要的速度，然后调成自动导航状态吧。放开手柄，Firebird III会自己搞定。”



图6-3 Firebird III（火鸟三代）

通用公司还邀请BBC现场直播了高速公路上的无人驾驶测试，不过，这时的无人车依旧通过接收预埋线缆发送的电子脉冲信号行驶，没能挣脱“地面轨道派”的思路。

至此，不论是将“遥控器”掌握在人类手里，还是可行性不高的预铺电缆，都和人类想象中自由、流畅的无人驾驶体验相去甚远。直到20世纪60年代，俄亥俄州立大学的项目负责人Cosgriff（科斯格里夫）还深信，埋设在道路中的电子导航设备将在15年内推向公共道路。世界各国的实验室还要在这些“磨盘”的圆周上徘徊多年，唯一的区别是各自走出的半径大小。

曙光就在前方

当时，影响今天无人车的主流技术已经在各大研究机构中显露雏形。只不过在那时，这些技术零星分散在各处，也没有人想到要把它们组合起来。

1966年，智能导航第一次出现在美国斯坦福大学研究所里，SRI人工智能研究中心研发的Shakey是一个有车轮结构的机器人。它可能要花上数小时才能完成开关灯这样简单的动作，但在它身上，内置了传感器和软件系统，开创了自动导航功能的先河。

1977年，日本的筑波工程研究实验室开发出了第一个基于摄像头来检测前方标记或者导航信息的自动驾驶汽车。这辆车配备两个摄像头，在高架轨道的辅助下时速能达到30公里。这意味着，人们开始从“视觉”角度思考无人车的前景。导航与视觉一起，让“地面轨道派”寿终正寝。

与此同时，GPS系统于1973年开始发展。DARPA（美国国防高级研究计划局）在1984年启动了“ALV自主陆上车辆”计划，目标是通过摄像头来检测地形，由计算机系统计算出导航和行驶路线等。当时，这一机器人采用激光雷达来识别道路，依靠GPS进行导航，并通过小型化的短波雷达来发现前方突然出现的障碍物并自动刹车。在这些描述中，已经不难看出无人车研究路径的成熟，但遗憾的是ALV项目研究持续了5年，由于成果有限，国会削减经费，被迫终止。这一停，又将无人车的诞生推迟了若干年。

同样在军事领域投入无人车研究的还有德国。德国军方科研机构从1987年开始和奔驰合作，开发无人驾驶车辆，其技术甚至比DARPA的ALV项目更为成熟，采用摄像头和计算机图像处理系统对道路进行识别。这项研究同样没有取得太多成果。

在今天我们已经知道，要让无人车具备敏锐的“视觉”，中间还隔着运算速度、大数据、深度学习等数不清的技术鸿沟。

1993—1994年，来自德国慕尼黑联邦国防军大学的Ernst Dickmanns（恩斯特·狄克曼斯）教授团队改装了一辆奔驰S500轿车，让其配备摄像头和其他多种传感器，用来实时监测道路周围的环境和变化。这是那个年代最成功的“动态视觉”实验。这一次，这辆奔驰S500在普通交通环境下自动驾驶了1000多公里。

几乎同时，从1984年就开始投入无人驾驶探索的美国卡内基梅隆大学，率先在1989年使用神经网络来引导自动驾驶汽车，即便那辆行驶在匹兹堡的翻新军用急救车的服务器有冰箱这么大，且运算能力只有Apple Watch（苹果智能手表）的1/10。但从原理上来看，这项技术和今天无人车控制策略一脉相承。

这所大学的NavLab项目在1995年发展到了第五代。一辆1990年款的Pontiac Trans Sport（运动跑车）经过改装后，配备了包括便携式计算机、挡风玻璃摄像头、GPS接收器以及一些其他辅助设备。成功完成了从匹兹堡到洛杉矶的无人车跨越国土之旅。从现在的意义上，可

以算作“半自动驾驶”。它们的研究成果对于现在的无人驾驶技术提供了很大的借鉴意义。

和全球的发展节奏相近，从20世纪80年代起，我国开始了针对智能移动装置的研究，起始项目同样源于军用。1980年国家立项了“遥控驾驶的防核化侦察车”项目，哈尔滨工业大学、沈阳自动化研究所和国防科技大学三家单位参与了该项目的研究制造。20世纪90年代初，中国也研制出了第一辆真正意义上的无人驾驶汽车。

在“863计划”支持下，很多大学与机构开始研究无人车。历经“八五计划”“九五计划”“十五计划”，北京理工大学、国防科技大学等五家单位研发的ATB（Autonomous Test Bed）系列三代无人车依次问世，ATB-2直线行驶速度最高可达到21米/秒，ATB-3在环境认知和轨迹跟踪能力上得到进一步加强。

国防科技大学与一汽集团合作研发的红旗CA7460无人车，可以根据前方障碍车辆的情况自动进行车道变换，行驶速度最高可以达到47米/秒；清华大学研制的THMR-V无人车，最高行驶速度也可以达到42米/秒，还可以根据不同驾驶场景选择高速公路和城区公路两种驾驶模式；由西安交通大学研制的Springrobot也是我国著名的无人车平台之一，具有较高的车道线检测能力和行人检测能力；国家自然科学基金委员会于2008年启动了“视听觉信息的认知计算”重大研究计划项目；2009年，首届中国“智能车未来挑战赛”在西安举行，此后几年里，吸引了名牌院校和科研机构的数十辆无人驾驶车辆参加该项比赛。

这样的赛事国外开始得更早。2004—2007年的DARPA系列竞赛，让“无人车”展露出高速发展的苗头。

2003年，美国发动伊拉克战争，也点燃了无人车技术爆发的导火索。穿越沙漠运送弹药补给成了美军的常态行动，但即便有装甲车和直升机保护，辎重车队仍会经常受到攻击。加之美军人员经常遭遇路边炸弹或地雷袭击，伤亡惨重。因此，曾经孕育互联网的DARPA重启了

搁置十多年的自动驾驶技术研究。在美国政府和军方的计划中，10年后至少有1/3的军用车辆能够自动驾驶。

2004年，美国国会拨款设立100万美元奖金，后又加码到200万美元，由DARPA牵头举办无人驾驶大赛，广发英雄帖。尽管相对无人车的软硬件成本奖金并不算高，但仍然吸引了来自企业、科研机构、大学教育机构，甚至个人在内的不少参赛者。

应战时需要，第一届无人车挑战赛在沙漠中进行。这条从加州Barstow（巴斯托）到内华达州Primm（普里姆）的赛道，长约230公里，大多数是行驶难度颇大的沙漠地形。比赛要求在10个小时之内完成，只能依靠GPS来引导驾驶，依靠传感器或摄像头绕开天然障碍物。军方预测一定会有一个队赢得比赛的胜利。

不知军方是低估了无人车技术的难度，还是高估了参赛者的水平。这一年，不仅没有一支队伍抵达终点，即便是比赛的第一名，卡内基梅隆大学“红”队的Sandstorm（沙尘暴），也只开出了11公里，因为拐错了一个弯，陷进沟里动弹不得。多半参赛车辆在比赛刚开始就结束了，刹车锁死、车轴断裂、翻车、卫星接收器失灵等故障频出。

结果令人沮丧，人们的预期急转直下，悲观者说能在沙漠里行驶的无人车或许永远都造不出来。然而，这场不尽如人意的比赛没有止步于此，还成了无人车发展史上的拐点。

2005年，第二届无人车挑战大赛依旧在沙漠中举行，赛场依旧设在加州和内华达州交界处，赛程212公里，难度也是有增无减，3个狭窄的隧道，100多个急转弯。最艰险的一段狭窄弯路，一边是60多米的深沟，另一边则是悬崖峭壁。然而这一年，绝大多数参赛车都超过了上届11公里的最高行驶纪录。尽管依旧有不少参赛车辆中途退出，但有5辆车完成了全部赛程。

前三名分别来自斯坦福大学、卡内基梅隆大学和一家美国私企。这一届比赛中的冠军车辆已经搭载了摄像头、激光测距仪、雷达远程测距、GPS等今天无人车必不可少的装备。

2007年，DARPA把赛场搬到了城市里，开始从军用转为民用。新赛场内有交通灯、汽车等复杂环境，与未来无人车实际使用的环境更接近。比赛结束，卡内基梅隆大学、斯坦福大学、佐治亚理工大学夺得前三名。无人车研究领域的格局似乎开始稳定下来，DARPA的赛事却戛然而止。由于各种原因，美军至今未将无人车技术用于后勤运输。但是，技术的发动机一经启动，就没有人能让它停下。

图6-4是2007年DARPA无人车挑战赛的冠军车——卡内基梅隆大学的“Boss”，车顶及车头加装的设备体积和复杂程度远超今日进行路试的各类无人车。DARPA发掘了无人车研究者的潜力，也孵化了无人车的基础技术路线。

比赛中车队普遍通过摄像头、雷达、激光设备来探测周围的地形和障碍物。将结果与GPS和传感器的信息整合为一体，做出加速、减速、转弯等操作决定。十多年后，后人所做的，无非是在这条“无人车智能路线”的基础上，进行技术迭代，更加深入和精细化。





图6-4 卡内基梅隆大学的无人车“Boss”
资料来源：http://mt.sohu.com/20160805/n462676928.shtml

比赛促生了一个由发明家、工程师、程序员、开发商组成的无人车生态圈，也促成了无人车技术投资热潮的兴起。谷歌、苹果、特斯拉、优步、百度陆续宣布研发无人车的计划。各家都毫不掩饰自己在无人车领域的野心，仅谷歌而言，不仅挖走了斯坦福团队的领军人物塞巴斯蒂安·史朗，还把该校研究无人车的人员几乎一起揽入门下。2007年冠军团队的核心人物惠特克在卡内基梅隆大学的很多学生和同事都成为美国无人车行业的中坚力量。

在互联网企业的灼热展望中，就连保守的传统汽车制造厂也“被迫”加入了“无人车军备竞赛”。

虽然当年实验成本居高不下制约了量产和商业化，让无人车陷入瓶颈，不过2007年依旧值得记忆，不仅是因为它开启了无人车的新篇章，更是因为这一年深度学习研究重获新生。这个节点过后，相关领域也呈现井喷之势，大数据革命、云计算兴起、移动互联网浪潮、数据采集渠道的多样化……

更多的变化不断将无人车从传统汽车行业的束缚中解脱出来，让它走出大学实验室。

多年来，内燃发动机、变速箱以及复杂的生产工艺，构成了传统汽车行业的高墙壁垒。但是，新能源车或者说电动车技术的日益成熟，在这座高墙上打开了一个缺口。让汽车的制造门槛大幅度下降，因为，过去最难造的发动机、传动器、离合器，在电动车里面都不需要了。这不仅给了一些科技企业长驱直入这一领域的机会，也使得一些长期受困于核心技术差距，只能在较低端跟随的国家（比如中国）的汽车制造业，赢来了弯道超车的机会。

老牌车企“车到山前”

2013年，智能驾驶方兴未艾，美国高速公路安全局（NHTSA）就为智能汽车正式划分了等级。

根据给出的定义，智能汽车的发展可以分为4个阶段：第一阶段（L1）是“高级辅助驾驶系统”，特点是为司机提供碰撞警示，紧急情况制动，盲区监测和弥补夜间行车的视力弱势。第二阶段（L2）被翻译为“特定环境的自动驾驶”，接近于通用公司的设想，车辆能在

高速公路或是堵车这种相对规律的环境中自动驾驶。从第三阶段（L3）开始，我们对无人驾驶的期待有了一些轮廓，L3即为“多种环境中的自动驾驶”，车辆能适应所有路况，但在特殊情况下需要转交给人类驾驶员。到了第四阶段（L4），即“全自动驾驶”阶段，智能汽车真正做到了自主驾车。至于L3和L4的区别，从外形上看，L4智能车将在L3的基础上，撤掉方向盘和刹车。

分类可以说是对汽车行业的一次重新洗牌。在无人车的江湖里，传统车企、互联网企业、出租车行业巨头三分天下。传统汽车生产商的兴趣点在于L1和L2，L3和L4则聚集了包括谷歌、百度、优步、特斯拉在内的巨头。

今天，起着决定性因素的不再是资本和历史，而是各家企业的内在基因。L2和L3之间，有着老牌车企难以跨越的技术鸿沟。包括百度在内的互联网企业在技术、数据和人才上的积累，占领了传统车企短时间内无法企及的技术制高点。不过即便使用相似的设备，如摄像头、传感器、深度学习、激光雷达等，也会因为软件数据库的全面程度不同，而达成不同等级的无人驾驶效果。

宝马I3在车身上安装了4个IBEO（激光感应器），具备躲开立柱障碍、紧急刹车等功能。遗憾的是，使用场景尚且停留在多层的停车场环境中。沃尔沃推出的XC90，配置有城市安全系统（City Safety），新增冲出路面防护系统和十字路口自动制动系统，能帮助司机避免常见的追尾事故。

2015年初，奔驰推出了F015概念无人驾驶汽车，车内布满了各种显示屏以及可旋转的座椅，俨然一个可移动的娱乐中心。但落到实处，奔驰的S500仍然没有摆脱窠臼，Intelligent Drive（智能驾驶）系统专攻的依旧是防碰撞、保持车道、自动刹车和堵车时自动跟车功能。



图6-5 奔驰的超现实概念无人驾驶汽车F015
资料来源：http://www.diseno-art.com/news_content/2015/01/autonomous-driving-mercedes-benz-f-015-luxury-in-motion-concept/

在总裁玛丽·巴拉的领导之下，2015年，通用汽车收购了Cruise Automation公司，向打车服务公司Lyft投资了5亿美元，推出通用汽车

第一款长续航、自动驾驶的全电动汽车Bolt。电动车、自动驾驶、共享经济，每一步都时髦且精准。

但迎接通用的是行业内的不理解，明眼人都明白无人车带来的危机和冲击将是前所未有的：无人汽车一旦发生交通事故，将把车企置于极为不利的地位。此外，无人车推动的共享经济很可能大大降低私车拥有率。那时，车企的光环不再，只是作为最基层的硬件提供者而存在。这些远景让巴拉的锐意进取被讽为无人车浪潮下平抚焦虑的“自我安慰”。

最可圈可点的，大概是德国车企戴姆勒了，它与内华达州达成协议，率先获批自动驾驶卡车上路。2016年4月，戴姆勒公司安排旗下三辆自动驾驶卡车车队上路，从德国的斯图加特开到了荷兰的鹿特丹。不过使用场景尚且仅限于高速公路，且车上仍需要有一名驾驶员负责监督。

智能企业的边缘突进

在被通用收购前，Lyft已经和雪佛兰合作研发无人车。就连黑莓也在放弃手机行业后，转而投身于无人驾驶。这些接触汽车行业不久的互联网公司，急不可耐地投入无人车的研究，毫不顾忌行业老大哥的犹豫和踌躇。而更为激进的特斯拉在2016年就有超过10万辆能够超车、躲避障碍、自动进停车位的半自动驾驶汽车上路了。

“年轻人”的“跑酷”自然要付出一些代价，2016年1月，特斯拉自动驾驶汽车的第一起事故被证明在中国发生，尽管目前调查结果未定，但在行车记录中，这辆具有自动驾驶系统的汽车，面对前方明显的清扫车，没有减速迹象，导致追尾。2016年5月，特斯拉在美国发生一起事故，在强烈的日照下，自动驾驶系统未能识别到前方横穿而过

的一辆白色集装箱货车，没有及时刹车，而货车车主也违背了操作规范，双手离开了方向盘。撞击致使车主丧生。2016年8月，得克萨斯和北京各有一位车主，或撞向高速护栏或与旁车发生剐蹭。事后分析都指向车辆的感知系统误检。

在撞击集装箱货车的事故中，由于横跨多个车道、车架高于地面，高亮度天空背景下的白色车身被视觉识别系统误认为是“白云”。北京剐蹭事故中，路边静止车辆被识别系统当作围栏，自动驾驶错估了安全距离，导致了近距离摩擦。应该说，这意味着对识别系统的训练还不充分。

优步则在未获得加州政府上路测试牌照的情况下，开始测试无人车。实验过程中因为闯红灯险些撞车而变成了一出闹剧。

在无人驾驶这条道路的同行者中，还有一些搁置了实时更新道路数据的技术路线，主攻嵌入式人工智能的解决方案，目标是让自动驾驶在不联网的情况下实现道路感知等。

这些不断跌倒又前进的尝试，更多地指向L3无人车。而对于下面提到的两位，L4才是雄心所在。

L4准入门槛极高，人工智能和高精地图信息不可或缺。因此，谷歌和百度占据了绝对优势。

2009年，谷歌在DARPA的支持下开始了自己的无人驾驶汽车项目研发。2014年12月，谷歌对外发布了完全自主设计的无人驾驶汽车。外观构造已经与传统汽车截然不同，取消了方向盘与刹车。2015年，这辆原型车已经可以上路测试，乘客只要坐在车中，就可以“坐等”车辆开动。截至2016年10月，这辆许多人用“萌”来形容的白色小车，已经进行了超过320万公里的公共道路测试，相当于人类司机300年的开车经验。

对于谷歌的野心勃勃，百度恐怕是最能理解的。几乎是同一时间，百度无人车也在加速前进。在中国移动互联网一片热闹的几年，

百度却埋头无人车开发，来不及向外界解释，只有安全和学习成果演进在我们的注意范围内。那句“3年商用5年量产，要在2021年让无人车出现在地平线上”的宣言，并非媒体所称的“激进”，而是源自自身技术的召唤。



图6-6 谷歌无人车

百度想做的也是“完全无人驾驶”，在技术路线上选择了几乎最难走的路：高精地图加传感器。与上述所有涉及L3的自动驾驶车高了几个数量级的难度。为此，“百度大脑”们一头扎进无人驾驶领域。

最初百度对该项目高度保密，直到2015年12月，百度无人车完成北京开放高速路的自动驾驶测试，外界才逐步知道百度在做什么。这次测试首次实现了城市、环路及高速道路混合路况下的全自动驾驶。百度无人车从位于北京中关村软件园的百度大厦附近出发，驶入G7京新高速公路，经五环路，抵达奥林匹克森林公园，随后按原路线返回。全程均自动驾驶，实现了多次跟车、减速、变道、超车、上下匝道、掉头等驾驶动作，完成了进入高速汇入车流到驶出高速离开车流的不同道路场景的切换。测试时行驶速度最高达到100公里/小时。

百度无人车的车顶装有一台64线激光雷达，通过持续地圆周旋转，对车身60米半径内的路况进行全景扫描。车辆两端和后方分别装有3个激光雷达，用于扫描邻近区域，弥补车顶雷达的视角盲区。这一连串的高分辨率激光雷达，构成无人车的“复眼”，对整体路况有了透彻的“俯瞰”，有助于实现在复杂交通环境下的自主驾驶。特别是针对交通拥堵、狭窄道路、小区和停车场等特殊场景，实现跟车、换道和交叉路口通行。

有了这些还不够，车顶前方还装有两个视觉识别摄像头，这样一双“眼睛”可以精准地识别交通信号灯、路面交通线和交通标志，弥补了激光雷达在雨、雪、雾等特殊天气条件下的“视力下降”，提高了对路标和信号牌的识别率。

从环境感知、行为预测，到规划控制、高精地图、高精定位，一辆无人车上集纳了当下多个领域最顶尖的技术。百度能在这么多的方向同时迸发，依靠的是在人工智能、深度学习领域的长期积累。当然，从无人驾驶技术的发展历程来看，今天百度无人车也可以说是立足在巨人之肩，每一台硬件设备都有过异想天开的“原型”。每一条技术路线都是全球无人车研究者在反复失败中提炼的“最优解”。

现在百度无人车还在中美两国进行着大量的路测。在百度的时间表里，2018年要实现百度无人驾驶公共汽车上路行驶载客。

一项技术从开始科研到落地使用，这不仅是百度的进步，也是中国制造的成就。从一开始，百度肩上扛的就不仅是“无人驾驶”的“小目标”。



图6-7 百度“云骁”无人车在乌镇亮相

注：用手机百度或智能革命App扫描图片可见AR效果。

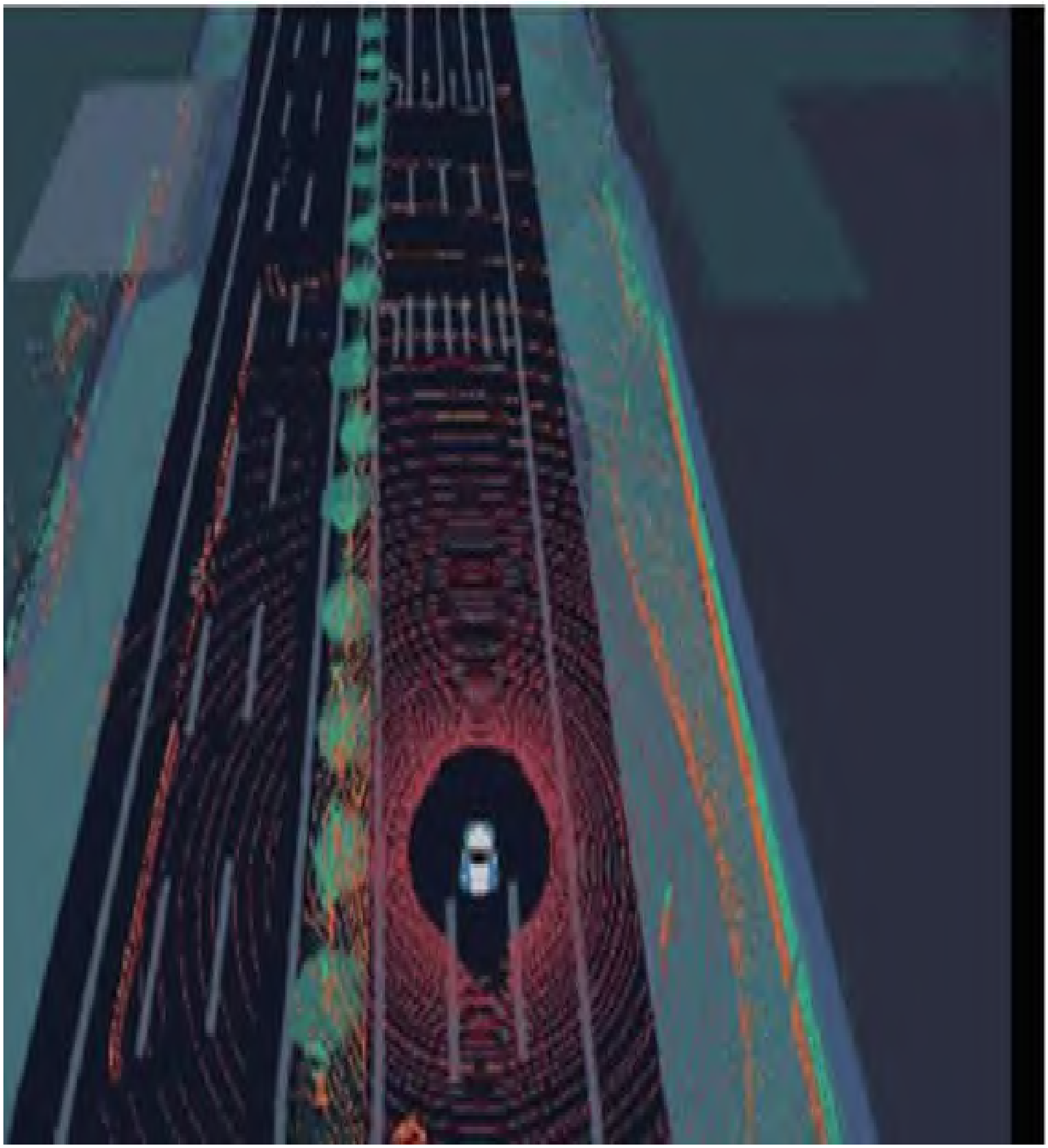


图6-8 百度无人车视角下的路面

在2015年的世界互联网大会上，百度展示的无人驾驶车就吸引了习近平主席的注意，习主席在百度公司展台的驻足时间超过10分钟，比原定时间长了3倍。中国工程院院士李德毅曾对无人车行业做出分析，他认为“以无人驾驶车为代表的轮式机器人将成为中国智能制造

2025的首张名片”。无人车不仅是“车辆”，也因其自身装载的设备系统，在商业化过程中将直接带动雷达、传感器、导航系统等产业的快速发展，为建立车联网和物联网打下基础。无人车自身的战略价值，已经远远超过了一个交通工具。

中国的交通场景丰富、人口众多、市场庞大。无人车不但能为整个交通模式带来变化，还有可能给中国带来智能地铁、智能公共系统等新事物。智能交通是系统性、革命性、颠覆性的变化，随着技术的狂飙突进，需要改变的可能是人们的思维模式。

陆奇说起过自己的一桩逸事，他玩过一种特殊的自行车，是倒着骑的，方向盘往左拐轮子会倒过来，这其实和人工智能也有点关系。学习有几种方式，有一种是体验式学习，比如学骑自行车会摔跤，但学过就不会忘记，而学了这个特别的自行车就不会骑原来的自行车。在为无人车制定法律时，人类要避免刻舟求剑，继续用有人驾驶时代的思维束缚未来。

苦练无人车的内功与外功

在了解无人车技术的历史和现实背景后，让我们从技术层面深入无人车的“视和听”。

在“视距”上，无人车较之裸眼具有绝对优势，驾驶员的安全视距一般在50米左右，而无人车依靠多种中远距雷达、摄像头等传感器，能实现200米以上的超视距扫描观测。

另外，无人车的“视觉”方面还要克服许多对人类来说不是问题的问题。为了让它能够从不同角度对汽车和周围路况进行判断，我们

要教会计算机在动态中识别哪些是车、哪些是天空、哪些是树。深度学习在这一过程中起到了巨大作用。

目前，在用摄像头来判断物体这项指标上，百度无人车准确率达到90.13%，2016年同期为89.6%。看似只有0.53%的提高，但对于无人车乘坐者的安危，每一点进步都值得坚持。在判断行人方面，百度无人车的准确率达到95%，判断红绿灯的准确率达到99.9%。而它的未来目标是，判断红绿灯的准确率必须达到100%。

当然，对无人车来说，光会“看”是不够的，要保证安全，还要有敏捷的反应，要能“刹得住”。人类驾驶员，从遇到紧急状况到踩下刹车需要0.6秒。而汽车的刹车要发生作用，其油压系统传导还需要0.6秒。也就是说，一般的人类驾驶员需要1.2秒的时间把车刹住。而百度无人车从“发现”紧急情况到做出刹车动作只需要0.2秒。未来，如果电动刹车取代液压刹车，百度无人车有能力做到在0.2秒内紧急制动，这将比人类平均的1.2秒快了整整1秒，高速驾驶中，1秒可能就是救命的时间。

不论是静、是动，还是“表面”招数，在绝大多数的行驶过程中，车载的雷达、传感器、摄像头等要实时收集数据，在行车途中反馈到“百度汽车大脑”的服务器上，辅以GPS高精度地图，指导汽车以最佳路径行进。

在定位层面，百度地图2015年达到WiFi定位精度30米，室内高精定位精度1~3米，定位速度0.2秒。2016年，百度无人车使用的高精地图精度是10厘米级，实际行驶中对路况判断的误差小到一根车道线的宽度，相比GPS定位，精度提升了两个数量级。

这一切也要归功于无数百度产品的使用者，包括那些调用百度地图API的用户。使用者在调取时，无形中就帮无人车刷新了数据，增加了一点“智慧”。

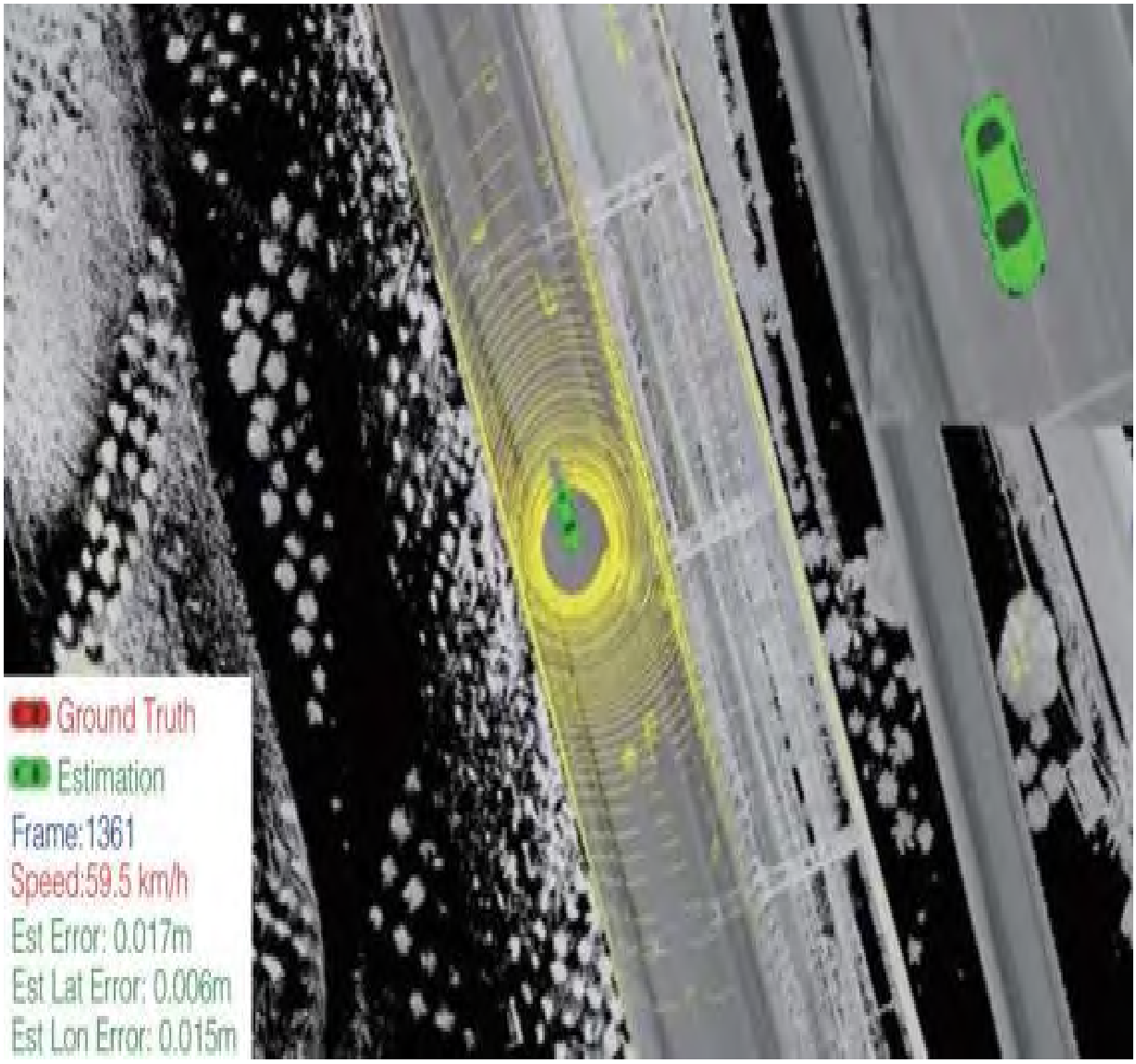


图6-9 比普通GPS定位高两个数量级的高精定位示意图

顺利推进无人车的研发，关键在于构建合理的技术布局。随着技术的升级与革新，以往推广应用的“死结”都会逐渐被化解。业界对无人车是否应该采用激光雷达的争论就是如此。

由于激光雷达造价高昂，外界有人唱衰无人车，正如多年前很多人因为价格而不看好“大哥大”。百度不仅对降价抱有信心，还做出了前瞻性的战略投资。

2015年底，无人车普遍采用的LiDar激光雷达的价格高达70万元，半年后同款批发价已经降为50万元，下降了30%。随着生产工艺的成熟及无人车行业发展带来的规模化效应，激光雷达将有更大的降价空间——毕竟“大哥大”刚面世时售价高达2万元，而今国产的“国民”智能手机只要600元左右。

2016年8月17日，百度与福特公司以1.5亿美元联合投资激光雷达公司Velodyne LiDar。Velodyne预计，如果2017年拿到100万个订单，百度无人车使用的64线激光雷达单价就会直线下降到500美元，2020年激光雷达价格会降到300美元左右，2025年会降到200美元左右。

在无人车的研发势头下，64线激光雷达目前非常紧俏，这项投资确保了百度的传感器供应。但如此巨大的投入，并非只为一时的器件供应，而是以行业带头人撬动整个产业，以利于整个无人车行业的发展。

除此以外，为了改进硬件计算能力，百度自主研发了由48台服务器组成的小集群，计算能力超过了中国超级计算机神威太湖之光。

从百度地图十多年的运营，到百度大脑感知和决策能力的全面开花，再到对激光雷达供应商Velodyne公司的投资，与其说是百度选择了无人车，不如说是百度的专业逻辑将自己推到了必须承担起责任的这一步。

仅有技术层面的可行性是不够的，为了积累路测经验，百度无人车在国内外的试验场上奋力奔跑。在美国加州，百度是第15个获得无人车测试执照的公司。就在2016年底，百度在加州部署了100多名无人车的研究人员和工程师。在国内，百度也要全力做好“中国定制”。鉴于选择测试和初期商业化，在考量道路设施、行人密集度、政府（相关部门）支持力度，甚至当地天气等多方面情况后，百度无人驾驶团队入驻北京、上海和深圳，分别和安徽芜湖市政府、上海国际汽车城、浙江乌镇旅游景区、北京亦庄开发区签署了测试的合作协议。

2016年，在获批的“国家智能网联汽车（上海）试点示范区”封闭测试区内，模拟了城市交通场景，有隧道、林荫道、加油/充电站、地下停车场、十字路口、丁字路口、圆形环岛，还有6个智能红绿灯和40个各类摄像头。整个园区道路实现了北斗系统的厘米级定位和WiFi全覆盖，可以为智能驾驶提供29种场景的测试。

百度在2017年推出了高级自动驾驶人工智能模型Road Hackers，并将基于此模型开放百度自动驾驶训练数据，首期就包含1万公里的L3数据。这部分涉及一线、二线城市等十几个城市的封闭道路的高速、环路、快速公路的数据。今后还会有一些人车混搭的道路数据公开，并逐步开放所有大数据资源。这当中不单是道路数据，还有中国司机的开车习惯，这可能是更为感性和宝贵的数据。这一模型能够在真实的道路情况下，利用深度学习等技术，将L3无人车的摄像头、传感器等输入的信息转变为车速、路径、方向等最优的驾驶指令。

Road Hackers的推出表明，百度在L3和L4双路径并行突进。对百度来说，L3的数据也可以用到L4，当收集了足够多的公路数据、场景数据、用户数据、车的数据，并将这些数据和激光雷达的数据结合起来，可以让L4更快地实现。

更重要的是，相较于L4，L3能更快地与传统汽车行业对接。

百度联手北汽，推出搭载车联网解决方案的车型，计划在2017年底进行北汽L3车型的路测。百度无人车奇瑞EQ参加了驾照考试，已经通过了第一个考试科目，希望2017年能考过两个科目，3年内能把5个科目都考完，实现商用目标。

未来，无人车还要走上开放路面进行测试，在具有代表性的城市区域进行测试，最终融入城市交通系统。美好的前景正在实现，但还有数不尽的用户体验需要去完善。

自动驾驶不应该只是少数人的专属，而应该成为每辆车的标配功能。无人车承载了中国汽车产业很多的希望，这一点绝对不是百度一

家公司能够完成的，而是需要整个产业，包括各级政府、汽车制造企业、科研机构、银行、保险公司的共同努力，才能把中国的无人车真正推向社会，真正落到实处。

中国电信在全球5G标准竞争中取得的成果令人兴奋。无人车是物联网中极具意义的应用，可以共享设定区域内的位置信息、外部环境信息、自身驾驶信息等，每辆车都会成为信息的接收者和发出者，实现整个区域内的协同驾驶，进而实现车与车、车与人、车与万物的相连。而无人车和物联网的实现正依赖于5G时代的到来。

“老司机”驶向何方

尽管无人车还远没有达到大批商用的阶段，但作为新事物，已经面临来自各方的怀疑，这是技术之外的另一道关卡。无人车上的联网设备会侵犯个人隐私吗？数据被黑客窃取会产生不良后果吗？无人车出了事故如何认定责任？

人们常用一个形而上学的伦理问题来问无人车：路边突然窜出一个小孩，如果无人车打方向盘，就会撞到另一边的一个行人，无人车该怎样选择？进阶版问题还有：如果一边是一个人，另一边是一群人，无人车要怎么选？对此，技术性的回答通常是：采取“让速不让道”的方式，尽可能快地减速，以避免、缓解悲剧的发生。但真正的关键在于，这个伦理问题即使是人类老司机也回答不好。对待无人车，人类可以多一分理解。

谷歌无人车在实验阶段就应对过各种情况：一个坐着电动轮椅的女人在路上追赶一群鸭子、一群人在马路上排成一行做青蛙跳、一个人毫无预兆地突然靠近并滚过无人车的前盖，甚至有人在车子前面裸奔。有人问怎样看待百度无人车程序有3000多个场景，1万多个

“if”。这个数据百度没有真实统计过，但这些“if”显然是为了众多诸如“撞人还是撞狗”的场景而预备的，至少保证对于普通人能够想到的特殊情境，无人车都能做出最有利于交通参与者人身安全的选择。

不妨再大胆想象一下。在万物互联的条件下，未来面对人类也难以决断的紧急情况时，人工智能的反应速度、联网效应很可能会给出比人类更好的解决方案。车联网系统将把无人车连成整体，每辆车在行驶时“脑海中”都有着“大局观”。一辆车紧急刹车就会紧急“告知”一定范围内的无人车也做出对应的措施，从而避免追尾事故。

相比民众对无人车太“凶”的担心，工程师们担心的反而是无人车在与人的相处中太“怂”了。在谷歌的实验中，无人车在右道行驶时发现前方有沙袋，它试着减速绕过沙袋开向左车道，本以为后方的公交车会减速让行，不料公交车却在其变道时加速超车，导致了两车相撞。在这起事故中，反而是遵守交通规则的无人车被人类“欺负”。无人车的训练中不可能允许违法行为的出现，那么它要怎样和灵活应变的人类相处，遇到实际驾驶中的模糊区域要怎样做出决策？目前仍是无人车集中攻克的一门“学科”。这实在要比围棋棋盘复杂得多。由此可见，担心一个以防御为主的无人车主动伤害人类，可能还为时过早。

上面这些还是“假设”层面的问题，针对无人车的法律法规、自动驾驶汽车监管、事故时的责任界定等问题可以说就近在眼前。

目前，无人驾驶技术在欧盟以及美国一些州都存在法律障碍。欧盟交通法律规定，汽车必须由持有相应驾照的司机驾驶，任何人不得以任何理由接管驾驶权。美国只有四个州支持自动驾驶汽车上路。

但在法律层面上的“举步维艰”不会持续太久了。2015年，根据谷歌所描述的车辆设计，“美国国家高速公路交通安全管理局把这种车的‘司机’理解为自动驾驶系统，而不是车上的人”。即接受了谷歌的说法，认可了谷歌无人驾驶汽车将不会有传统意义上的司机。

2016年初，联合国欧洲经济委员会对《维也纳道路交通公约》进行了修改，从原先“驾驶车辆的职责必须由人类驾驶员负责”改为“在全面符合联合国车辆管理条例或者驾驶员可以选择关闭该技术的情况下，将驾驶车辆的职责交给自动驾驶技术可以被允许应用到交通运输当中。”

虽然中国还没有具体的法律法规面世，但中国人“摸着石头过河”的传统让无人车企业感到放心。业内人士并没有太担心法律的阻碍，毕竟法律总是在追赶现实，在算法可能代替律法的时代，技术才是发展的核心驱动力。中国的无人车一旦启动，其加速度十分可期。

归根结底，无人车是服务于人的。无人车这位“老司机”不断学习，技艺日臻完善。科学家梦想有一天，无人车能够使得路怒不再，拥堵不再，更少的车和更高的用车效率或许将让雾霾也不再。呼之则来的无人车，为我们每一个行程规划完美的路径，精准计算路上时间，在这个空间里，我们可以休息、工作、学习、娱乐、休闲，甚至旅行。充分利用路上的时间使居住在市中心与郊区的差别将不再巨大，由此影响城市布局和地价均衡，进一步影响年轻人的工作与生活。人类将有更多时间和空间发展自身的技能，提高健康水平，弥补对家人陪伴的不足。让无人车尽情打开我们的想象空间吧！

当然，在乐观场景的另一面，我们或许会感受到另一重孤寂。为效率而产生的更多单座、双座无人出租车，默默无语的旅途，或许会让我们向往拼车带来的交际温暖。过去与司机胡侃的夜晚，可能会变成一项收费服务。车祸少了，一旦出现交通意外反而容易上头条，相比技术进步，人性的进步速度是缓慢的，我们的易怒和敏感或许会转移到其他议题上。

汽车文明是现代工业文明的化身。纵览大地，从上古时期的百兽竞走，到今日亿万辆汽车奔驰，再到未来无人车自在涌动，堪称生命的进化之路。从此汽车将不仅是汽车，公路也不只是公路。文明就是

“在路上”，生生不息。知道更多，做到更多，体验更多，愿无人车让我们变成更好的人类。

07

AI 带来的普惠曙光



金融市场似乎总是伴随着最大的不确定性，或暗流涌动，或风暴骤起，席卷颠覆所有的东西。

因此，在人们的想象里，金融常常是冒险家的乐园，大鳄们关起门来，相互间会心一笑，掀起市场的血雨腥风，然后在人们的仓皇恐惧中赚取超额利润。但实际上，巨头也不是总能呼风唤雨，他们也经常搬石砸脚，随着市场波动起伏，甚至投子认输，黯然退出。

除了跌宕起伏，金融的另一面常常是细水长流、润物无声的。可以给人们在追求物质、精神生活改善，或者自我发展道路上以及时的助力。

无论是跌宕起伏，还是润物无声，其背后都是对资金流和信息流的复杂逻辑的处理和应对。“金融获得感”日益成为一个重要的社会命题，如何在尊重金融规律的基础上，让金融跳出“大玩家游戏”的窠臼，更好地服务于亿万普通人？

人工智能可能正是处理海量信息、应对不确定性的最好工具。其创新突破以及在金融领域的实践，正开始让金融这种似乎总是高高在上的存在，重新俯身垂滴普罗大众，并为普惠金融的实现带来了曙光。

机构来了“新实习生”

1994年，美国正处于黄金时代，作为全球唯一的超级大国，到处都是歌舞升平。这一年夏天，世界杯足球赛刚刚在洛杉矶落下帷幕，而新科技的火焰正在硅谷升腾，网景浏览器一经发布很快就风靡全球。中国在那一年才刚刚通过一根64k专线全功能接入国际互联网。为了遏制金融泡沫，格林斯潘领导的美联储在那一年开始大幅度加息，

然而债券市场并没有意识到美联储进入了加息周期。全球金融危机13年以后才到来，此时的华尔街一片欣欣向荣。无数人才从世界各地来到这里寻找梦想。

暑假之前，我收到了道·琼斯一家子公司的实习聘书，实习内容是从事金融信息处理系统相关的工作。

我本科专业是信息管理，毕业后到美国留学，获得布法罗纽约州立大学计算机科学硕士学位。眼下的工作正好把信息管理和计算机结合起来。这份每天与财经金融新闻打交道的工作，我一做就是三年。接下去，我参与设计了《华尔街日报》网络版实时金融信息系统，又转赴国际知名互联网企业Infoseek做资深工程师。一边与金融信息打交道，一边通过《华尔街日报》观察硅谷的商战，并开始思考如何应对信息作弊的问题，不久之后就提出了“超链分析”技术的想法，并申请了专利，为日后的搜索引擎发展奠定了基础。

那时我还没有想到，有一天，随着人工智能技术的发展，机器系统对金融业务的改变会达到今天这样的深度，渗透到金融信息处理、数据分析、风险控制、征信、智能投顾（Robo-Advisers）、智能获客、量化投资等各个方面。

量化投资之王大卫·肖恩说过，“金融是一个绝妙的信息处理的生意”。也正因为此，百度高级副总裁朱光说：“最革命性的变化至少会在金融领域发生，因为人工智能就是数据的收集、分析、处理这个循环的极致。”

无论是在征信和反欺诈领域，综合用户画像构建风控模型；还是在投资领域，挖掘投资决策因子辅助投资决策；或是在投资顾问领域，为客户匹配个性化的投资组合等，机器对各种金融业务的渗透，本质上都是通过机器学习带来的金融信息处理能力的不断提升。

随着自然语言识别和信息数据库技术的不断进步，即使是单纯的金融信息处理，也已经发生了本质的改变。

这次向我们走来的实习生是一位机器人，而它与众不同的工作方式首先直观地体现在一份份金融分析报告的生成上。

金融信息可能是最复杂、最枯燥的信息，一份股转书有两百多页，还有大量的年报、半年报、研究报告、公告、反馈意见、尽职调查结果……我们不知道，有多少行业分析师是完全看完这些信息，再做出决策的。也许，不是他们不够勤勉，而是读完这些信息已经非人力所及。

百度数据金融总经理杨晓静描述过这样的行业现实：20世纪90年代，一个基金经理要把市场当天产生的研报、舆情、新闻、交易数据等看完，大概需要10个小时，也就是两天的工作量。2010年，移动数据爆发之后，这个基金经理要把每天市场上产生的信息吸收掉，大概需要10个月的时间。2016年，还是这个基金经理，假如把当天市场上所有的信息看完，大概需要20年的时间，相当于整个职业生涯。所以基金经理迫切需要利用先进的智能技术，比如百度的自然语言处理技术。

今天，当百度金融的智能金融信息分析系统处理数百篇行业新闻，并将其中的关键信息结构化，抑或是阅读上市公司的年报并形成分析报告，用时已经缩短到以分钟计。

在这一过程中，机器相当于承担了一个金融机构初级分析师的工作，甚至是一个实习生的工作，承揽了最基础的所有的累活儿。

这位机器实习生的工作逻辑也是类似提取“关键词”，并进行再组合的过程。

机器可以瞬间完成上市公司的公告、财务报表、官方发布、社交平台、证券行情、实时新闻、行业分析报告等海量异构数据的阅读，对于文本中的图片和表格需要OCR（光学字符识别）等技术解析。紧接着，进行关键实体信息的提取，发现埋藏在实体信息之间，如行业上下游关系、供应链关系、股权变更历史、定增与重大资产重组的关

系、多张财务报表之间的数据交叉验证等数据关系，形成并呈现这些复杂关系的“知识图谱”。

如果更进一步，分析师只需要选择符合其需求的模板，确定主题，机器即可生成基本的报告文本。在最终输出前，分析师还可以人工校正报告精度，并加上独特的个人分析和结论，这样，一篇格式标准，甚至图文并茂的金融分析报告就诞生了。

这个潜力无穷的实习生显然不会停留在简单处理信息这个阶段。既然已经登堂入室，它必然会走得更远。

人工智能让起点更公平

机器人首先推门进入的就是传统金融的核心区域——信用评级。

2016年7月18日下午，百度宣布投资美国金融科技公司ZestFinance。而一年前，ZestFinance还曾获得京东集团的投资。能同时得到中国两大互联网巨头的青睐，让ZestFinance这家当时仅服务10万名美国人的数据征信公司，为更多的中国人所了解。

这家位于洛杉矶的金融科技公司，使用机器学习的方式来评估个人贷款的信贷风险指数。其创始人是谷歌前首席信息官及工程副总裁道格拉斯·梅里尔和美国第一资本投资国际集团公司前主管肖恩·布德。

在美国，ZestFinance是挑战征信巨头FICO公司的革命者，后者占据着美国99%的信用评分市场和绝大部分发达国家的信用评分市场。

ZestFinance信奉“一切数据皆是信用数据”。不同于FICO的信用分数仅包含贷款人的几十项变量，ZestFinance的模型基于海量社交网络数据和非结构化数据，包含的变量将近1万项，在大数据挖掘的基础

上最终形成一个独立的信用分数。而且与常规的信用评估体系相比，效率能够提高约90%。其号称在5秒内，就能对每位信贷申请人的超过1万条原始信息数据进行分析，并得出超过7万个可对其行为做出测量的指标。

就在入股ZestFinance之前的百度联盟峰会上，百度宣布，“人工智能对于金融也会产生变革性影响，可以真正做到让征信升级”，并特别强调，“现在百度的教育贷款，基本上是以‘秒’的时间就可以决定是不是给一个人贷款”。

“秒批”的背后，正是以机器学习为基础的大数据风控，在提升信贷服务效率、增加金融服务覆盖率上的小试牛刀。

一般来说，大数据风控的成果跟传统金融机构一样也是两个名单：基于白名单的征信，基于黑名单的反欺诈。后者因为“反黑”的目的，常常笼罩了神秘的色彩。比如彼得·蒂尔创立的、有全球第四大独角兽之称的人工智能公司Palantir，不仅在帮助美国安全部反恐中屡立奇功，更因为整合了40年的记录及海量数据并充分挖掘之后，终于发现了伯纳德·麦道夫(Bernie Madoff)的“庞氏骗局”，而为天下人所知。

但我们更愿意从主流的角度来讲述人工智能在金融风控中的故事。

根据中国人民银行的数据，截至2015年9月，中国人民银行征信系统收录的8.7亿自然人中有信贷记录的为3.7亿人，可形成个人征信报告、得出个人信用评分的有2.75亿人。

这就意味着，在中国还有约5亿人是没有征信记录的“小白”，他们被挡在传统金融服务的门槛之外。

而依托庞大的数据基础以及人工智能实现的人群画像技术，百度金融等企业，正在悄悄改变风控线上化的难题，把以往高高在上的金融服务涓滴到更多亟需帮助的“小白”身上。

比如，在大学学习了4年室内设计的李亮，最近开始在网上搜索UI（用户界面）课程、培训学校、分期交学费等信息，他希望通过学习，未来能进入一家大的互联网公司工作，但教育机构普遍不低的学费成了摆在面前的第一道坎儿。

李亮不知道的是，在同一时间还有很多人在百度上搜索着同样的关键词，虽然此时此刻他们还没有与百度金融直接发生关系，但这些人的群体性需求，正以数据的形式进入百度大数据风控的视野之内，并通过机器学习被归类于某一群体画像，从而拥有了相应的信用判断。

对比了几家教育机构之后，李亮终于选择了一所培训学校学习UI，并且决定试试老师推荐的百度有钱花学费分期贷款。他通过手机在短短几分钟内，填写个人身份信息并拍摄个人图像等，完成贷款申请步骤。

而在光纤的另一端，百度风控的策略系统迅速响应，根据人群画像，借助图像识别等技术，对李亮的信息进行采集、加工和分析，并调用数据字段结果送给风控平台，进而完成授信过程。

经过短暂的等待，李亮收到了人生中第一笔贷款审核通过的短信。

几个月的UI课程学习结束后，李亮决定再次贷款学习VI（视觉设计），为即将步入职场加码，这时他惊喜地发现，因为良好的还款记录和稳定的消费记录，机器给他的贷款额度和信用支付场景在扩展和提升。

更重要的是，机器给李亮建立的人生第一份信用记录，可以帮助他超出百度金融体系之外，在更多金融机构里享受更全面、更好的金融服务，彻底告别信用“小白”状态，提升金融服务的获得感。

朱光曾说：“我们这个社会，究竟该由谁来给身无长物的年轻人发放第一笔贷款，在他们人生关键的爬坡阶段给予扶持？现在来看，

答案也许是‘机器’。”

而当机器完成了金融服务对象——人的数据化，就已经没有什么可以阻止它在金融王国里纵横驰骋了。

个人钱包的智能守夜人

一向对技术投资不太感冒的巴菲特大概不会想到，有人会用他的名字为一款智能投资咨询软件来命名。这款致敬投资大师的软件，就是智能投资公司Kensho（这个名字是日语用来描述禅宗的“明心见性”，透过现象看本质的意思）设计的一款基于云计算的财经软件“沃伦”，其使用大数据和机器学习分析具体事件（从自然灾害到选举结果）对市场的影响，并使用简单易懂的知识图呈现结果。

这款软件发布之初就震动了华尔街，很多人甚至打电话给Kensho的创始人纳德勒，大骂他是“叛徒”。因为在华尔街，无论是人工智能还是什么其他华丽的技术，关起门来闷声发大财才正常，而将其公之于众，普惠众生则是典型的“大逆不道”。

因为金融信息的掌控和处理本身就被看作是一项垄断性的生意。彭博社和路透社估计：长期垄断的金融数据市值达到260亿美元。

越来越多的“沃伦”同道正在打破这一现状。

另一家公司Hedgeable，创立之初就立志推翻华尔街的垄断，其创始人Michael Kane（迈克尔·凯恩）和Matthew Kane（马修·凯恩）是一对双胞胎兄弟，曾服务于全球最顶尖的对冲基金桥水基金，以及最著名的投资银行摩根士丹利等华尔街巨鳄机构，因为日益厌倦华尔街只为世界上最富有的人服务，而辞职在纽约创立了Hedgeable，依托人工智能技术，致力于为普罗大众提供投资顾问服务。


HEDGEABLE
WALL

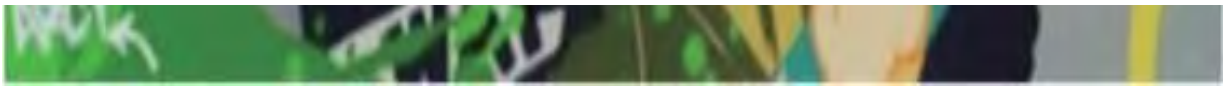


图7-1 Hedgeable办公室墙上空手道踢倒华尔街的涂鸦
资料来源：www.Hedgeable.com

如果说投资顾问对美国人早就是司空见惯的服务。那么，对于经历了三十多年财富积累的中国人来说，则是一项亟需普及的财富管理方案。

2016年底，知名财经作家吴晓波进行了一项“新中产消费调查”，发现人数可能达到1.8亿的“新中产”阶层，正在普遍陷入一种财富保值的巨大焦虑之中。

高净值人群的财富一直享受着金融机构私人顾问的专属打理，那谁来保卫“新中产”的钱包，缓解他们的财富焦虑？

智能投顾，也被称为机器人投顾、智能理财等，是指在人工智能和大数据的基础上，针对不同风险偏好和投资要求的用户，由计算机提供基于算法的投资管理建议，帮助投资者组建个性化的资产投资组合，实现个人资产配置的最优化。

智能投顾一般遵循如下服务流程：首先洞悉投资者需要，即清楚地了解投资者自身及家庭整体的关键数据。比如投资者所处人生阶段、收入水平、历史投资经验和偏好。一般来说，投资者标签越丰富，画像的颗粒度越精细，对投资者的把握就越准确。

理财是财富保护、长期投资、资产配置，更是一种人生规划方式。所以，下一步，机器还要详细描摹投资者阶段性的生活追求，如买房、买车、求学、育儿、养老等，以此搭配相应的投资周期以及考察投资回报期望。

接下来，机器就要考察投资者的风险偏好。以年龄、职业生涯所处阶段、收入结构、生活负担等作为基本参考纬度，确定投资者能承担风险的阈值。需要指出的是，在线下，投资者面对自己的投资顾问

表达的风险偏好，往往与其真实风险偏好有很大偏差，这就需要投资顾问与投资者进行专业细致的沟通，以洞悉客户真实的风险阈值，这一过程成本很高。而机器是通过大数据识别客户风险偏好，而且可以根据市场状况实时动态调整，画出投资者的风险偏好变化曲线，这就大大降低了沟通成本，也就意味着投资者可以付出更少的“羊毛”。

全面掌握了投资者的基本情况后，就需要根据具体的客户体征，在丰富的金融产品中，挑选最适合的资产配置组合。

因此，机器顾问作为中介，在连接用户端完成投资者画像的同时，当然也要对金融产品端如数家珍，也即对可选投资标的背后的基础资产充分了解。比如，资产特性是什么、波动率怎样、收益表现是什么情况、稳定性是什么情况、多个资产间相关性是什么情况等。

对两边的了解都达标后，机器要考虑怎么做组合上的匹配，而不是单一的匹配。这一过程其实需要海量的运算能力以及高效的深度学习算法的能力。这也是百度这样的技术公司可以切入这个领域，并且如鱼得水的原因。

最后，还要有资产的监察和控制能力，随着市场变化调整资产组合，机器就要不断更新投资方案，以匹配用户需求。

智能投顾的出现，改变了理财资管机构与客户交互的方式，可以真正了解投资者诉求偏好，从而可以通过散点生产、定制化生产来配置资源。



图7-2 智能投顾示意图

更重要的是，借助人工智能技术，智能投顾可以低成本地将个性化专属财富管理方案提供给普通的中产阶层。一般智能投顾公司都没有特定的投资门槛，管理费用仅在0.15%~0.35%，用户投资金额越大，其收取的费用越低。这个费用相当于真人投资顾问费用的1/10。有的智能投顾公司可以将费用降低到真人顾问的1/20，甚至干脆免费。

而且机器顾问可以有效规避人性弱点。一般的投资者在战术性资产配置比如股票投资中，一旦被套就进入“假死”状态，只能等待解套；而一旦获取收益，又总不能及时落袋为安。机器顾问则不受情绪影响，设定盈利点和止损点后，可以自动严格地去执行，没有贪婪，也没有恐惧。

再者，机器顾问的精力是无限的，它可以7×24小时不间断地、兢兢业业地通过数字化资产配置，为客户提供相对定制化的、模块化的服务，并及时基于投资者个人的需求偏好和市场的变化提出服务方案。这保证了机器投资顾问可以同时跟手机或PC端的无数客户进行贴心的深度交流。

世界上最大也是最著名的私人财富管理机构瑞银财富管理，截至2015年底，在全球拥有4250名理财顾问，却要服务450万个人和企业，

服务覆盖比超过1：1000。与它相比，机器的工作效率显然会更高。

近几年，智能投顾在美国发展迅速，2012年美国智能投顾行业规模几乎为零，到2014年已经具有140亿美元的资产管理规模。其中包括Wealthfront公司、Betterment公司，以及我们前面提到的Hedgeable等在内的数家公司都在智能投顾领域有所发展。而提供投资决策信息服务的Kensho，则是另一种意义上的智能投顾公司。

毕马威报告称，预计到2020年，美国智能投顾的资产管理规模将会达到2.2万亿美元。

在国内，2015年左右开始出现模仿国外模式的智能投顾企业。到了2016年，大量智能投顾概念的产品开始涌现，虽然看起来市场火热，但是不得不说，其中不少是打着“智能”噱头的概念炒作。比如有的只是在智能投顾旗号下，推荐某种固定的投资组合，其实早就背离了智能投顾从风险评级、投资目的、投资能力和意愿方面动态配置的初衷。

从事这一行业需要大数据、机器学习及金融洞察等综合能力，而在中国的市场环境下，甚至对以上能力都提出了更高的要求。

比如，国外智能投顾的投资标的主要是ETF（交易型开放式指数基金，通常又被称为交易所交易基金），是被动管理基金。与此相应，美国市场同期有近1600支ETF，管理资产规模累计2.15万亿美元。但国内上市的ETF合计130支，资产规模累计4729亿元（根据Wind数据，截至2016年6月）。

指数型产品的稀缺，使得中国的智能投顾不得不引入主动管理基金。但主动管理基金本身的很多变化是不可预测的，比如换一个基金经理、基金公司战略变动，基金的收益变化就会非常大，而且这个变化很难被预测和模拟。这就对智能投顾企业的投资能力提出了极高的要求。

为了进行更为准确的投资者画像，一般智能投顾公司都会借鉴线下投资顾问形式，要求客户填写问卷。但在中国的投资和理财市场，以散户为主的投资群体，其投机心理大于投资心理。如前所述，这就导致了客户表达自身风险偏好失真的现象更容易出现。例如，投资者看到某些高风险的资产历史收益不错，为了顺利完成投资，就可能会忽视潜在风险，在问卷填报中虚高自己的风险接受度。在这个问题上，能够超出一般评价指标，为客户进行“10万+”级别的细化标签画像及风险评估的大数据公司，优势就十分明显。

美国有专门的第三方机构，能把投资者所有的账户汇集到一起，只要用户点击授权，智能投顾公司就能获取用户的所有账户流水等信息。比如，智能投顾公司Pefin，可以为投资者做到全账户集合管理，将投资者所有的在线金融账户汇集到Pefin平台，包括储蓄账户、银行消费账户、信用卡账户，月供、贷款和投资账户等，Pefin在几分钟之内就能利用分析模型，建立起投资者当前财务状况的知识图谱。

在中国是没有这样的账户汇集机构的。要尽量生成投资者的完整画像，除了让用户自己去填写问卷，还需要大数据基础上的机器学习画像能力。

最大的挑战还在于投资者教育，在一个以散户为主体的投资市场，怎么让投资者相信，你把钱交给机器，它会进行专业的资产配置，给你提供良好的财富管理服务。而且，在传统投资教育还未完成的情况下，进行智能投顾的教育，难度可想而知。百度智能投顾团队的袁月以“跳级”形容这一状况，“小学还没毕业，就要上初中。”

机器学习的“黑箱”也加大了教育的难度。一般传统投资顾问，做出投资交易后会向客户解释决策逻辑，但是目前还很难解释机器顾问的“想法”。

为了克服这一“缺陷”，很多智能投顾公司还是会强调其投资逻辑背后的专业人员因素。比如招商银行的摩羯智投，会突出其对市场上多家基金公司及其基金经理的相关情况掌握更全面，将其作为重要

的投资模型因子。Hedgeable专门开辟了投资顾问板块，允许注册投资顾问、财务顾问、认证财务规划师、注册会计师、律师、保险代理人和其他财务价值链上的重要角色使用Hedgeable平台为他们的客户服务。

当然，国内智能投顾行业的发展，还取决于未来监管政策的逐渐明朗。智能投顾行业如果因牌照所限，不能向个人投资者的证券投资账户直接提供投顾服务，那么在未来相当长的时期内，其最佳出路只能是向机构输出技术能力，这可能意味着机构投资者跟大众投资者的实力对比将进一步拉大。

智能投顾服务目前还不可能真正做到普惠众生，只能在高净值人群之外，服务限定规模的中产阶层人群，或者作为投资机构的技术输出方存在。这是由资本市场的零和规则决定的。即一个人赚钱的同时另一个人必定亏钱。当机器通过算法和程序发现了一个更好的策略，理性的做法就是遵循华尔街的规则“闷声发财”，如果这一策略被广泛服务于大众，收益率必然大大降低，甚而最终失效。这是智能投顾在普惠与高收益之间的悖论。

再以Wealthfront为例，其首席投资官就是《漫步华尔街》的作者Burton Malkiel（伯顿·麦基尔），而这本书倡导的被动投资哲学认为，既然长期战胜不了市场，那就干脆投资市场。Wealthfront就遵循这样的哲学，选择跟踪指数的被动型投资工具ETF作为投资对象，以期获得长期、稳定的收益。但作为一个公开策略，这也决定了它不可能获得超额回报。

从实际情况看，Wealthfront、Betterment等智能投顾公司，管理的财富也就是30亿~50亿美元的规模，跟一些传统的、超大的资产管理公司，比如贝莱德动辄数万亿美元的资产管理规模相去甚远。

本节开头提到的Kensho，最终还是难逃被华尔街投行大佬高盛（1500万美元融资）、谷歌风投、恩颐投资（1000万美元融资）等公

司收编的命运。其宝贵的数据处理技术，最终对普通人关闭，被限定在了华尔街神秘的小圈子里，其投资领域的平权革命戛然而止。

一些全球超级的资产管理公司如Charles Schwab（嘉信）、Fidelity（富达）、Vanguard（先锋投资）以及国际大投行如高盛、JP摩根、瑞银也都在通过投资并购或自建平台的方式进入智能投顾领域。

投资牌桌上的大玩家，正在金融领域的智能革命到来前纷纷醒来。未来，能与之竞合的，恐怕也只有人工智能技术实力雄厚，又掌握海量数据的科技巨头了。

数据挖掘：智能投资的钥匙

2016年11月，美国总统大选激战正酣。与此同时，投资市场另一场大战也正在上演。

瑞信的分析师们观察到，管理资产规模约3300亿美元，量化兼具多空策略的CTA（期货投资）基金正逐渐转向空头，针对美股的空仓仓位飙升到近期的最高水平。而另一方面，管理资产规模超过2150亿美元的多空股票基金，针对美股的多仓仓位则创下9个月来的新高。

对峙双方大部分的投资交易都是在计算机的帮助下，通过算法和模型来进行的。于是媒体将其定义为“机器人大战”。

由于量化投资的大部分交易是由计算机完成，也使用各种模型和算法，很多人将其理解为人工智能对投资市场的干预。但实际上，量化投资只是利用计算机强大的计算能力寻找市场上的无风险套利机会（Arbitrage），只能算交易策略范畴，大多与人工智能并无关系。

而真正的智能投资依然是基于数据驱动的。无论算法怎么迭代，怎么有创意，也不论逻辑关系设计得多巧妙，有了金融算法模型，还必须有大量的符合模型需求的经济、社会、特定行业变化等多维度的数据集，来验证算法模型的可行性和精准度。

可见，一个开放性的大数据环境对智能投资，或者金融信息分析都至关重要。因为从物理世界中获取数字化数据的成本极高，很多公司又没有自己的大数据资源，智能投资分析与决策就无从谈起。

百度、谷歌拥有的搜索数据、地图数据等庞大、多维、立体的大数据资源，有别于传统金融体系的数据，为互联网视角下金融数据特征的挖掘提供了丰厚的资源。加上领先的人工智能技术，给了搜索引擎类公司切入金融投资领域最佳的禀赋资源。

有了资源，也就有了“挖矿者”。搜索大数据中的炼金过程一般是这样的：通常，确定数据源后，依托成熟的大数据技术可对全网数据进行高速整合，提升数据运转效率，同时保持网络数据信息的完整性，进而运用先进的机器学习、人工智能、大数据分析等技术，对海量数据进行分析处理，挖掘金融资产的个性化特征。

例如，百度日均产生跟股票名称或股票代码直接关联的搜索量约2000万次，一只股票的搜索量和其股价走势往往呈现高度的正相关性，平均相关性达0.7以上，搜索量作为量价的补充信息，代表了某只股票被投资者关注的程度，辅助以舆情因子的判断，可以被当作潜在买家/卖家的能量积蓄，用来判断何时入场，什么时候又该进行风险规避。

搜索数据能潜藏宝贵的投资决策金矿，是因为人们在主动搜索时一般是不会说谎的。

如果搜索是主观意图数据的体现，那么，在数据足够丰富的情况下，就可挖掘提炼出人群对信息的潜在需求和关注程度。A股特殊的投资者结构（散户居多）和运行特点（习惯炒作）使得搜索数据成为绝

佳的观测维度。因此，搜索数据辅助大盘选择时，能够有效地观测市场变化。

搜索数据还可以进一步挖掘和提炼，比如，现在百度地图已经在全国标注了3000多个工业园区和4000多个商业区，通过观测这些工业园区和商业区，可以有效洞察一个地区、一个商业中心、一个景点，甚至一个城市的人流量变化情况。

智能系统甚至可以以机器学习的方式，提炼某一企业的知识图谱，并实时自动更新。

因而在丰富的时空数据下，投资方能以更广、更及时的视角实时监测所投资企业的开工率变化，并得到及时的投资决策参考。

传统的金融逻辑、投资逻辑、资管能力，加上相关性很强的实时数据，无疑会大大提高投资判断的准确度和前瞻性。

在金融领域，每提升一个百分点的效率，降低一个百分点的风险，都意味着巨大的财富收益。

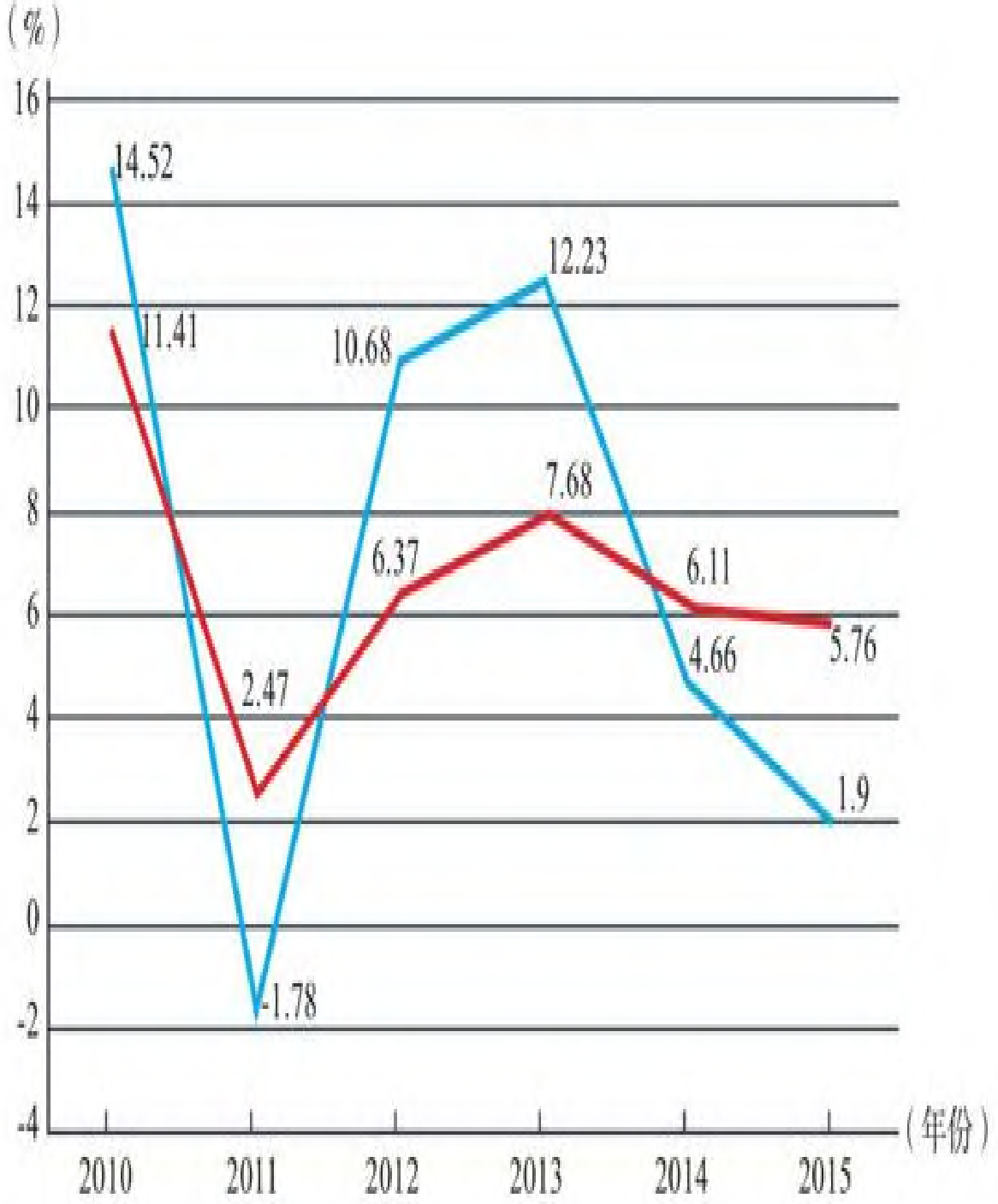


图7-3 百度收益型量化策略与中证500收益型量化策略的收益率对比

智能金融的三层境界

当然，市场需要我们永远保持敬畏，它的复杂多变远超出人的想象，为了理解这个市场，天才虽然创造了相对简约完美的模型，去抽象理解我们面对的金融世界，但是，在模型简约化的过程中，必然会损失一些东西。而且这些模型相对来说还是静态的，随着时间的推进，过去可用的模型也可能逐渐变得不再那么精准了。在伦敦咨询机构Preqin的研究中，典型的系统化基金的收益并没有比人工操作的基金效益好。

本节开头谈到的那场美国大选前的机器投资大战，可能被媒体过度夸张地解读了。正如巴菲特所说，“投资并非一个智商为160的人，就一定能击败一个智商为130的人的游戏”。



红线：计算机处理的基金　蓝线：所有对冲基金

图7-4 系统基金与人工操作基金收益对比
资料来源：https://www.wired.com/2016/01/the-rise-of-the-artificially-intelligenthedge-fund/#slide-1

不管什么样的算法和模型，都要尊重金融规律和投资逻辑，波动不是机器投资引起的，而是市场预期变化引起的。背后做出洞察和决策的还是人，至少在我们可以预见的未来，这一点不会改变。

目前，不少人工智能投资公司宣称，自己系统的股票交易已经完全不需要人的干预了。比如李嘉诚参与投资的Sentient公司的首席科学家Babak Hodjat（巴巴克·霍加特）就宣布“我们的系统可以让基金自动调整风险等级”。

以人工智能分析美股市场的对冲基金公司Aidyia的创始人Goertzel（戈策尔）则对自己的系统更加自信，“如果有一天我们都死了，这个系统仍然会自动交易”。

但实际上，我们很难想象一个完全由机器操控的投资市场。如果大部分基金都完全用人工智能，也许会出现我们不想要的结果。一方面，在精确的计算下，机器投资的节奏和标的越来越趋同，市场的波动越来越小，变得越来越无聊和乏味；另一方面，因为机器选择的投资对象越来越集中，也就同时埋下了市场失衡的种子。

在刘慈欣的短篇小说《镜子》中曾经描绘了这样的景象：当人类社会未来的一切都可以精确计算和预测，人类社会的发展也就走向停滞，最终便是文明的毁灭。

这样的联想可能有些扯远了，但它给我们的启示是：投资融合了人的本质欲望，甚至包含了助推人类社会不断发展的文化特质，投资过程带来的意义，远比最后的数字增长要大得多。

关于人工智能投资和人的投资决策的关系，百度副总裁张旭阳曾在一次研讨会上阐述了他的看法：

投资是技术、艺术和哲学的结合。百度领先的大数据和人工智能技术，可以解决一些技术层面的问题。但是，投资有时候是一个艺术层面的问题，否则就没有巴菲特这样的古典投资大师的存在了。对于一个市场信号，不同的分析师会有不同的解读方式和方法。比如，通过百度大数据分析，我们发现理发师和教师人群开始关注股票了。对于这个信息，有的分析师可能认为这是一个卖出信号，而另外一些分析师可能认为这是买入信号。这实际上是对投资经验提升的一种知觉，在这个决策判断的过程中，投资更多是一种只可意会不可言传的东西。

我们现在的人工智能技术可取得突破的地方，比如图像识别、语言识别、自然语义理解、用户画像、算法和辅助决策，这些都可以提炼出明确的信号，机器通过自我学习之后，可以做出判断。但投资决策最关键的部分是不明确的，是一些意中有、语中无的东西，我们目前的技术还没办法理解这种东西，也就没办法代替人做出洞察和判断。

AlphaGo打败李世石之后，很多人认为机器是不是可以替代人进行投资了。但实际上，投资与围棋的博弈完全不同。围棋是一个封闭的、信息充分的博弈环境，但投资有人的非理性，是开放式环境，这种情况下，技术要替代人去做决策，还需要极大的提升，这一过程至少以10年计。

实际上，目前用于投资的人工智能算法都大同小异，以前是逻辑回归、因果分析；后来有了深度神经网络，有了所谓的梯度决策树；再后来有了遗传算法。但这些算法的进步，都没有超出相关性分析的范畴，只可以实现一些短期记忆的东西，毕竟还没有达到人类大脑的反应程度，对于投资市场上一些可重复、可循环的投资决策，机器可以很好地去安排。但是，对于有一定缺陷的市场，或者说在投资艺术层面上有缺陷的市场，还需要人工干预。

我觉得在艺术层面，机器至少现在还难以取代人。我们的智能投资背后也需要一个团队维护算法，对这个算法的逻辑要不断地做一些调整，来适应不同的投资环境下的安排。

投资的第三个层次就是哲学，也是自律。也即投资到底是为什么？我要有止损止盈的法则。这个方面可能机器反而可以更好地去掌握。其实，通过机器来止损止盈是很容易做到的。因为，人难免有情绪波动，有贪婪和自信。比如，总认为自己跟别人不一样，这次跟上次不一样，总是觉得，我这次能逃离，别人会接我最后一把，我能出手，我虽然知道泡沫很大，但我自信不是接最后一棒的。但历史往往是重复的。而通过机器的算法，完全可以给你设定，到某一阶段就要认输离场，或者到了某种程度就要止盈了，可能还有一个正反合的过程。

但是，目前机器即使是在艺术层面、还难以取代人，我们的智能投资，背后也需要一个团队维护算法，对这个算法的逻辑要不断地做出一些调整，来适应不同的投资环境下的安排。

总之，在金融投资市场，人仍然是最关键的决定性因素，尤其是在金融跟人工智能的融合创新正在加速的时期。

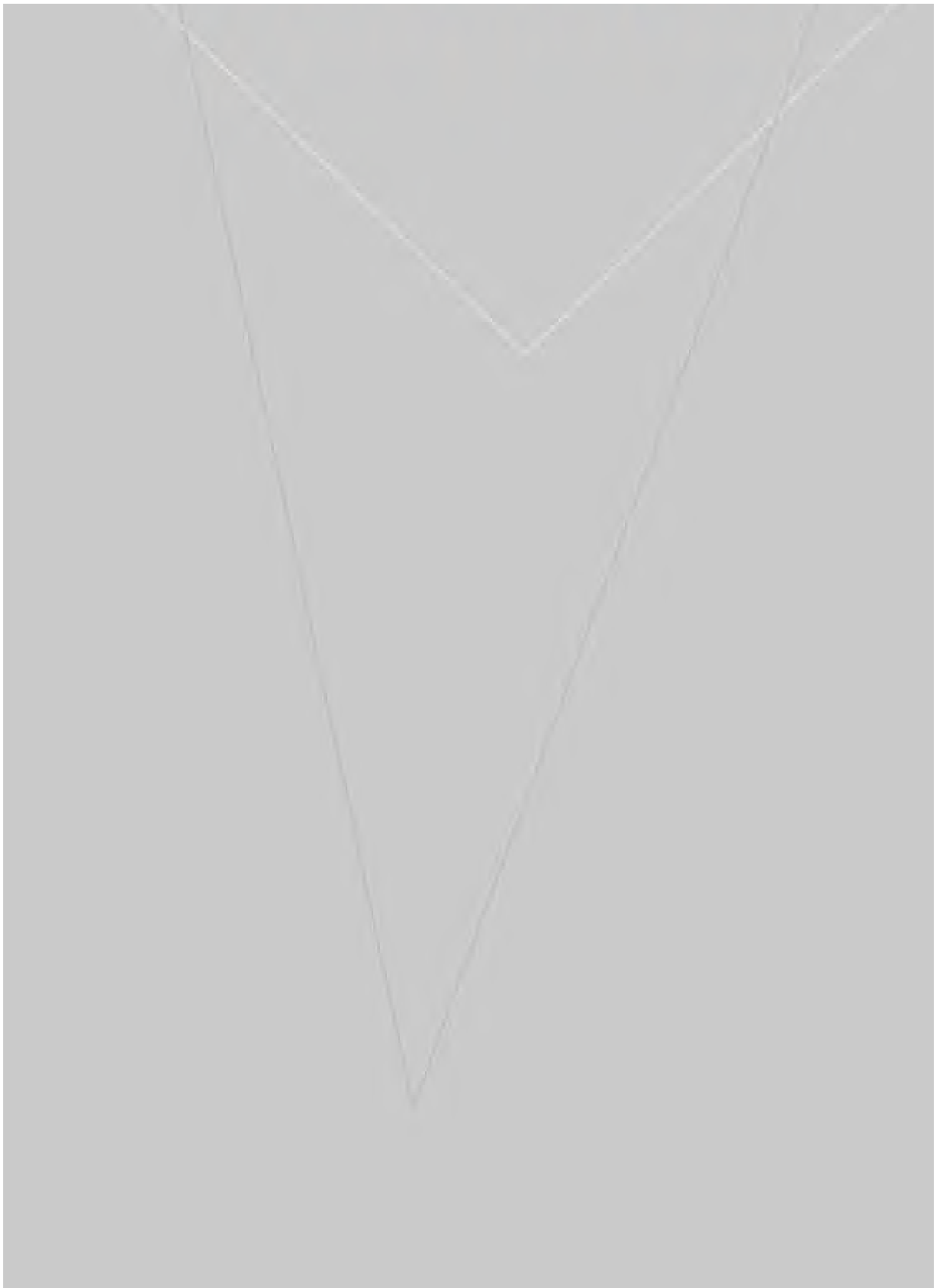
就像当年大批物理学、数学翘楚涌入华尔街，并给后者带来革命性的变化一样。如今，当华尔街面对金融科技浪潮惊呼“硅谷已至”的时候，人工智能与金融人才的跨界流动与融合，正在百度这样的技术性公司出现。

“Robin说欢迎硅谷的人回来。我就来了。”曾任职于微软研究院的吴建民这样回忆他来到百度的原因。他主要是在百度金融开展智能获客等相关研究。如果说他与更多的人工智能专家属于依靠技术解决金融问题的人，那么操盘资管理财的张旭阳、负责消费信贷的黄爽等来自传统金融机构的专业人士，则属于提出问题、定义问题的人。

这种人才的跨界组合激荡着创新。人工智能与金融业务的融合，将在身份识别认证、大数据风控、智能投顾等方面，对金融服务带来怎样的革命性改变，我们拭目以待。对个人来说，人工智能金融工具将无可回避。5~10年之后，当你去银行、证券公司、保险公司等机构接受理财、信贷、金融服务时，背后都有人工智能的身影。朱光将金融科技的未来状态称为“AI inside”（内嵌人工智能），我们希望自身的技术、数据、能力会支撑中国所有的金融机构，减少金融不确定性带来的冲击，发挥金融助力美好生活的力量，实现普惠金融的梦想。

08

每个企业都需要一位首席人工智能官



数据、算法与个人生活、社会治理、经济结构交相辉映的新文明呼之欲出。文明的进化不是靠纸上谈兵得来的，而是依靠千千万万创业者和劳动者的努力，以及有为者的均衡治理，自然演化出来的。这个过程充满了艰辛与不确定性。

中国的企业家们已经听见了来自物联网与人工智能的风声，但方向捉摸不定。企业有没有足够的数据资源储备？从哪个方向入手？更重要的，是否有可靠的人工智能人才？

对于技术性企业，这个问题不算难解，前沿产业与高端企业的人工智能升级早已上路，很多大企业已经开始布局智能化工作，比如智能家电。但人工智能作为一项扎根人类普遍生活的技术，必须被广大企业接纳才会愈加焕发光彩，而广大企业也都应该享有从人工智能获利的机会。

传统国企正在经历艰苦改革。太原铁路局是中国铁路货运量最大、重载技术最先进的铁路局。铁轨、电气这种工业1.0和2.0时代的设施提升后，马上就会面对工业3.0甚至4.0时代的挑战。面对繁忙的运输，必须优化业务流程，提升运转效率。

人工智能科学家分析，一列火车从一个城市跑到另一个城市需要的时间是比较固定的，影响运输效率的是中转时间，即当一列火车到站，把货物从火车转到仓库，再从仓库转到火车需要的时间。降低中转时间对物流效率有很大影响。百度协同太原铁路局采集了很多物流数据，把这些数据放到百度云上，训练出了几个模型来精准预测火车的到达时间、未来的仓储需求、未来的运力需求，把中转时间降低了50%。

老国企的智能化革新虽然不容易，但毕竟具有积累多年的技术团队和资源。相比之下，我们更关心广大中小企业如何在智能浪潮中把握航向。我们可以肯定地说，人工智能对中小企业是普惠的，无论商

业端还是消费端都有大量对人工智能的需求存在，普通企业尤其是中小企业在智能时代有很多的机会，时不我待。

谁来突破产品升级的瓶颈

智能革命一定是发端于领先学科和领先企业的，短期内会拉大企业之间的差距。但人工智能方法是公开的，便于学习，如果企业努力，就可以习得。在这个过程中，我们认为传统企业尤其需要人工智能领域的专业人士来引导。

2015年发布的《中国制造2025》行动纲要指出，“制造业是支撑我国经济社会发展的重要基石和促进世界经济发展的重要力量”“各国都在加大科技创新力度，推动三维（3D）打印、移动互联网、云计算、大数据、生物工程、新能源、新材料等领域取得新突破。基于信息物理系统的智能装备、智能工厂等智能制造正在引领制造方式变革；网络众包、协同设计、大规模个性化定制、精准供应链管理、全生命周期管理、电子商务等正在重塑产业价值链体系；可穿戴智能产品、智能家电、智能汽车等智能终端产品不断拓展制造业新领域。我国制造业转型升级、创新发展迎来重大机遇。”

并非只有《中国制造2025》中涉及的行业才适合智能升级，几乎所有企业都可以。剧场、电影院可以通过对观众购票规律的智能学习来优化票务工作；小超市、小店铺可以通过门店电子传感器和第三方数据（比如百度地图的客流数据）来分析访客规律；传统新闻单位可以考虑接入智能流来创新资讯生产和推送流程。

人们对升级的认识，往往局限在用料、工艺等方面，而人工智能带来了一个新的升级维度。以家居用品为例，人工智能在此领域已经赋予了许多新的场景：即便只是一副窗帘，也已经有人开发出智能化

方案，让窗帘系统主动记录和学习户主的使用习惯，在适合的时间自动开合以调节光线，配合户主的作息习惯；智能马桶圈可以不必整日加热，而是根据户主上下班时间和传感器来决定加热时间；养老家居系统可以观测、学习老人的起居规律，调整各项家电状态，甚至在老人作息异常时向亲人发出预警信号。我们需要的只是更多的创意。

历史经验：首席电力官的辉煌时代

众所周知，在人类工业时代的黎明时分，纺织机和蒸汽机的推广曾遭遇激烈反对。

在英国，由于纺织机械带来低成本产品，传统的手工纺织业被挤垮，遭到工业主和工人的反对，掀起所谓的“卢德运动”。珍妮机发明者哈格里夫斯多次遭到同行和邻居的驱逐。但珍妮机最终还是推广开来，助力英国统治全球纺纱业。而蒸汽火车一开始甚至速度还不如马车，从而被马车夫嘲笑。

到了电气革命时代，历史再次重复，比如马可尼在1895年研制出最早的无线电装置，并且利用这一装置成功进行了远距离摩斯电码通信实验。他成立无线电报与信号公司，推动无线电商用。但由于与海底电缆公司的利益相冲突，他想在纽芬兰设立无线电报局的事遭到反对。不过当时美国的现代市场体系和技术偏好已经初步建立，所以无线电还是很快发展起来。

电气革命与今天的智能革命有一些“基础质料”意义上的类似。与蒸汽动力无法远距离传输和统一布局不同，电力是一种无限流动的普适能源，正如今天的互联网是一种流动的，让用户可以随时接入的基础资源。“电+产业”正如今天的“互联网+产业”，颠覆了无数的传统产业。

当然，电流和智能流只是在比喻意义上可以类比，前者是电子的流动，后者是比特编码的流动，不是同类事物，但这种类比有助于我们感受问题的关键所在。我们不妨对比一下两个时代的企业升级。

当电气流向千万企业，很多企业主动寻求升级，虽不是像蒸汽时代那样抗拒，但同样困难重重。一百多年前的电力系统十分复杂。需要在直流电、交流电，不同的电压、不同水平的可靠性、不同的电力接口以及价格之间做出合理的选择。直到今天，各国的电压和接口规格（插座口）也都没有统一。和不同电力公司打交道也是个技术活，否则很可能吃亏，正如今天企业找互联网技术外包公司时，如果不专业，就会落入陷阱，从编程语言到系统架构，都有各种令人眼花缭乱的选择。

就公司本身业务来说，如何使用电力获得最佳效益也很难搞清楚：是应该先给企业全装上电灯，还是优先用电动机取代燃气轮机？于是，当时许多公司聘请了电力副总裁（VP of Electricity）帮助组织改革工作，以保证公司内的每个职能部门在自己的工作目标或产品上考虑电的存在。包括布置电线、购买电力设备、改造原有设备，甚至改造公司业务流程（电灯使得夜班成为标准）。

电气革命时代，人们可能想不到当时有多少关于电的开脑洞产品，就像蒸汽机时代也发明了很多后来消失了的蒸汽装置，这些装置如今只残留在蒸汽朋克动画片里。

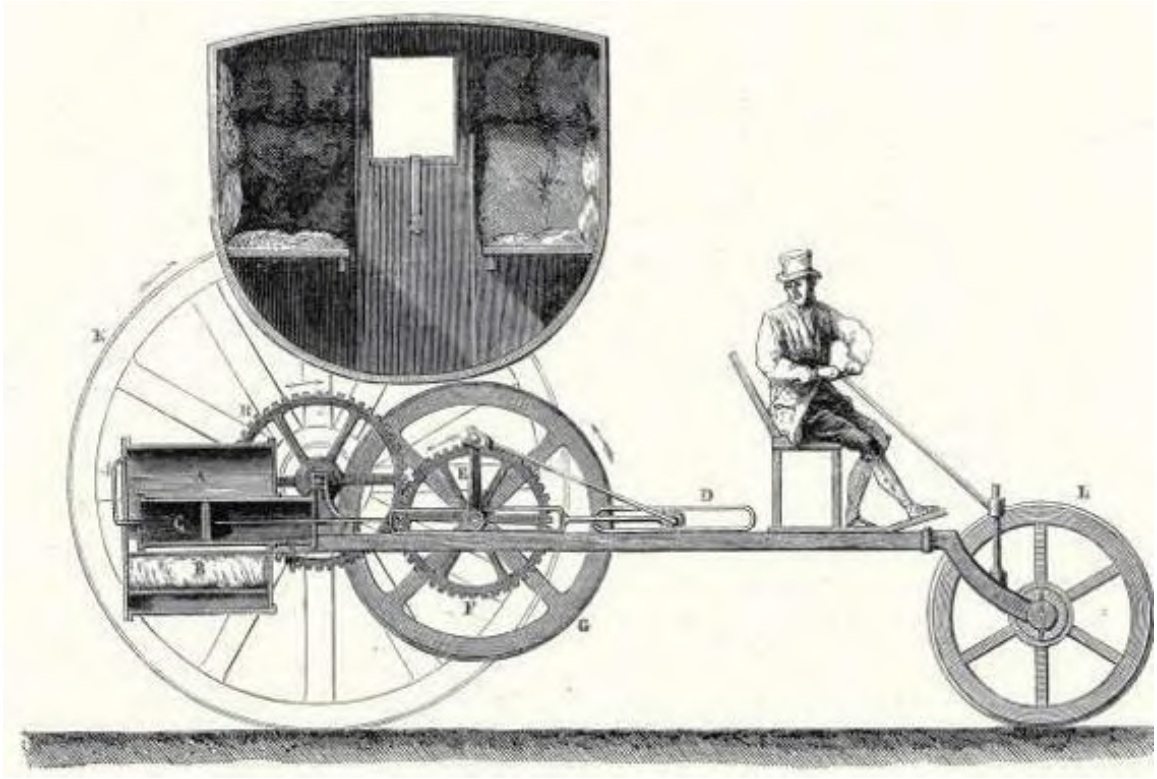


图8-1 1845年由意大利人发明的“蒸汽汽车”

首席电力官要发掘各种电气产品究竟能给公司哪些方面带来价值，他的视线不能局限于作为能源的电力，还要渐渐升级到对电子产品的跟踪。比如人类为了基于输入电压控制输出电流，需要一种可变电流开关。在20世纪50年代晶体管工业起步前是通过广泛使用真空三极管（也叫电子三极管）达到这个功能的。晶体管一开始生产工艺不过关，产品合格率低，以至于每一枚晶体管的价格高达20美元，而一枚真空三极管只要1美元。这时候企业该如何选择呢？这需要看准什么能代表未来。最终晶体管由于能耗低、寿命长、体积小而取代了真空三极管。

随着电力系统的成熟，首席电力官这一角色才消失。反过来说，每个使用首席电力官的企业，也促进了电力对总体生产方式的渗透，促进了国家电力系统的标准化。

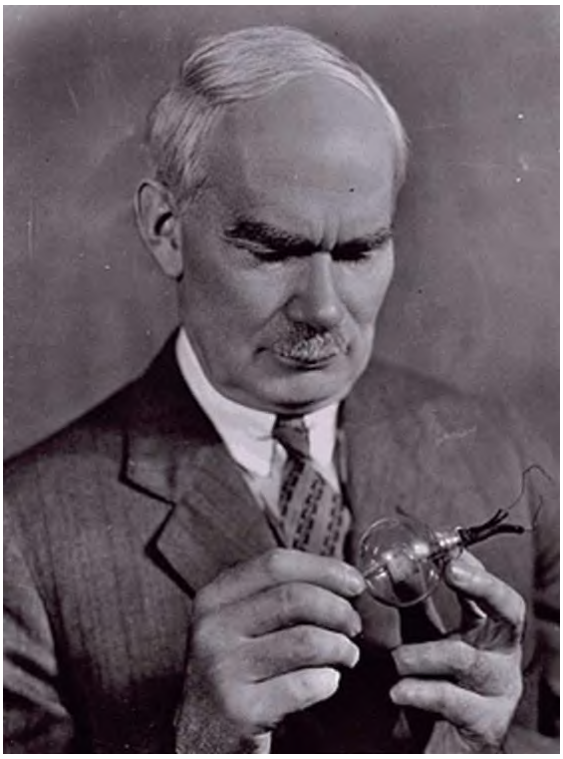


图8-2 真空三极管之父李·德福雷斯特（Lee De Forest）
注：使用手机百度或智能革命App扫描图片可见AR效果。

二十多年前互联网的出现也是如此。互联网走出军队实验室，渗透进企业、高校乃至家庭，这是一个复杂的过程，同时也给企业带来了网络技术总监这个职务。

在今天这个全民使用互联网的时代，很多公司依然选择网络外包服务。从节约资源的角度来说，这是合理的，不是所有企业都要供养昂贵的技术团队，像购买电力一样购买互联网服务是很方便的事情。

但购买互联网服务分为两个层次：一个层次是基础设施（IaaS）——接入主干网，交服务费，这是一个服务非常标准化的层次，不用费太多脑筋；另一个层次则非常个性化，是在基础平台上定制互联网产品服务（PaaS和SaaS），比如云服务器、数据库系统、办公协同工具、财务管理工具、App制作等，需要定制开发。这个层次操作起来对传统企业仍然比较复杂。如果企业做到一定规模，技术条线依然单纯依赖外包，就会迷茫。于是不仅科技企业，传统企业也需要CTO（首席技术官）。很多小公司的CTO并不管理一个大团队，他的职责就是代表公司利益和外包方打交道，决定选择什么服务，并监督实施过程。

所以优秀的CTO绝不只是网管运维，还要发掘公司所有部门、产品与网络的价值链，跟踪最新的技术发展方向。他要了解产品和业务，预判长期技术战略方向，不断跟进业务和技术发展做革新，否则就会像历史上的电力官一样，等到电力普及了，就被工程师或者产品经理

代替，不过至今在大部分工业企业中依然有电力官残余的影子，那就是电工。

迎接智能原力

人工智能应该像互联网一样成为一种开放的事物。开放的人工智能服务也将变成一种“流”，就像电流一样方便用户接入。

人工智能专家和企业家李志飞在2016年乌镇互联网大会期间发言说：“必须非常清晰地认识到，人工智能在短期根本不能跟人相比。而我们今天应该踏踏实实把一些应用，比如云交互、自然语言理解、计算机视觉等各种应用场景做出来。例如车载设备、可穿戴设备等。让用户觉得真正有用，而不只是在媒体上或者在电影里面讨论一下。”

这与百度的态度不谋而合。谷歌、特斯拉等国外企业做了不少有趣的智能公关项目，无论是下围棋还是QuickDraw（你画我猜）游戏，都极大地提升了大众对人工智能的讨论兴趣。可能身处美国西海岸的工程师才有这样放松无虑的心态去做一些看似轻松的项目。背负更多责任的中国人工智能企业恐怕很难这么轻松，很多同行确实把精力都放在开发企业级应用产品上而不是做宣传。据报道，谷歌已经把一度吸睛无数的机器人和无人车业务都拆分了出去，不再给予无条件支持。中国企业呢？能否扎实推进人工智能的工业级发展？

百度追求把人工智能技术运用到企业升级中去，为全产业带来提升，而不是只创造一些故事。我们认为，人工智能就是要向企业输送源源不断的“原力”，比如百度地图的数据流平台、可供接入的语音识别服务等才是最重要的。

比如，百度地图衍生产品“百度慧眼”，通过对常住人口、客流量、生活区物价水平、写字楼密度，甚至具体到某个路口人流是往左走还是往右走来进行人群画像。在与很多连锁店企业的合作中，百度慧眼不用GPS，单靠用户手机对WiFi的搜索情况，就能实现30米以内的定位精度。只要用户手机收到连锁店自身的WiFi信号，就判断为用户进店。以此为店面评估市场容量、预估营业额。这一功能还可以为地产、电影院线等选址给出科学建议。

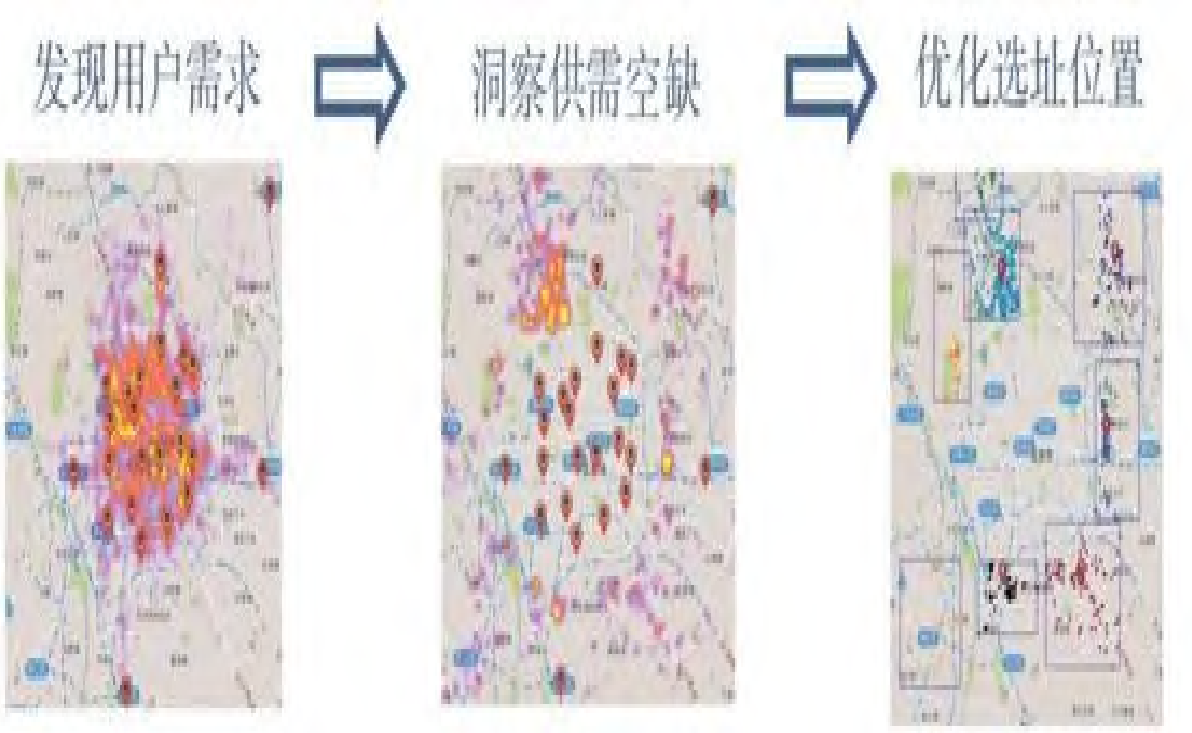


图8-3 百度商业选址过程

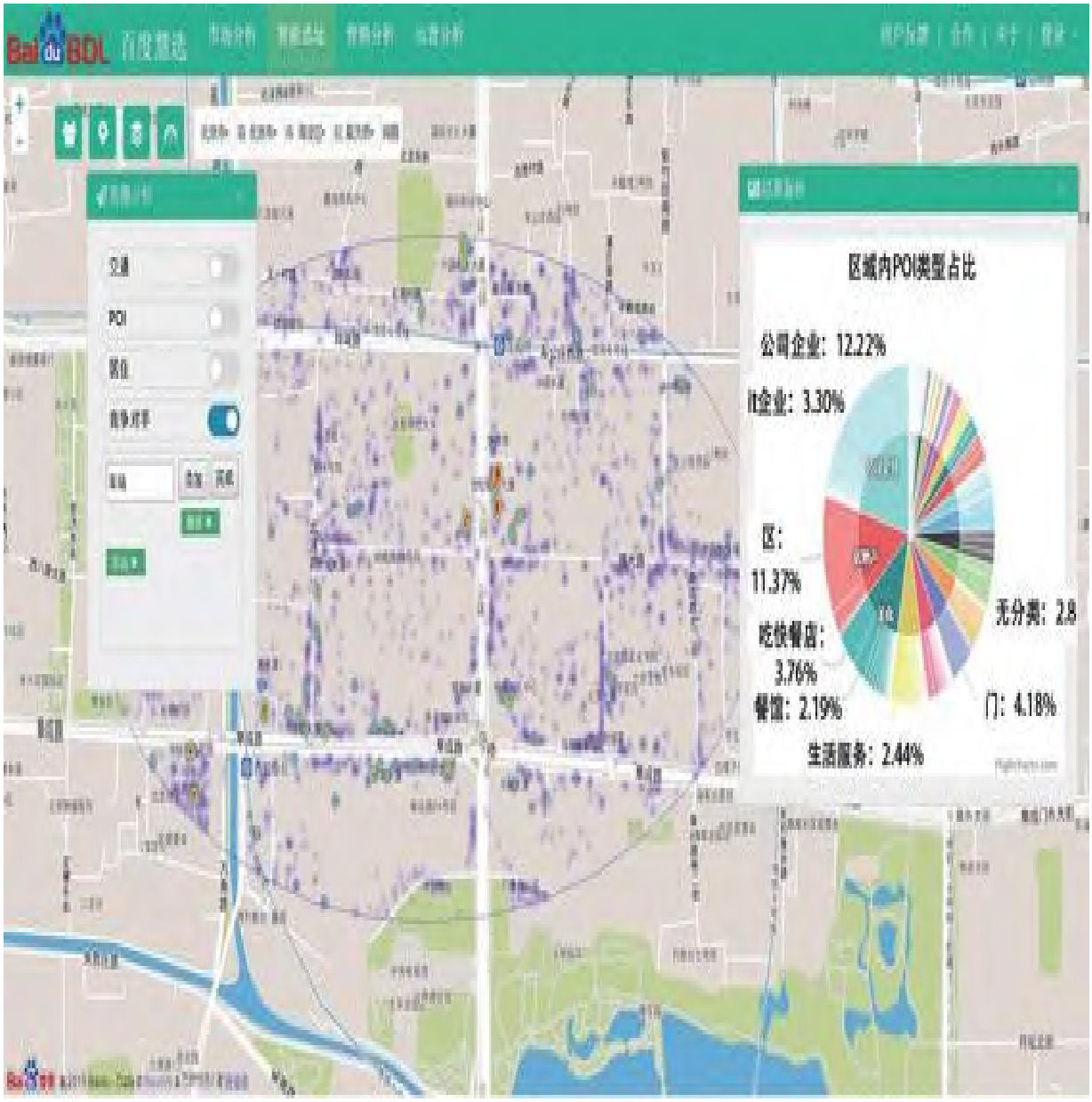


图8-4 百度利用时空数据分析进行商业选址的示意图

百度大数据实验室科学家吴海山对时空数据的挖掘，不仅将人工智能应用于商业选址，其在企业销售环节的应用也立竿见影。比如某个餐饮企业要向1000个用户发出1000张优惠券，但怎样保证这1000人没有光顾过店家并有意前来消费？我们可以根据“老顾客”的特征使用机器学习训练出一个模型。根据这个模型从数据库中找出消费习惯匹配的用户，休息日是宅在家里还是泡吧？经常去的餐饮场所的人均

消费如何？喜欢去国外旅游还是在国内旅游？上下班是坐地铁还是开车还是其他的方式？通过对新老用户的相似程度进行排列，选取最接近的1000人推送优惠券——这样的成功率显然比大街上随机发传单高得多。

如果想要寻找合适的智能流服务接入自身业务，每一个企业都需要有人工智能的领导者。传统企业为了利用人工智能的优势，需要理解人工智能能够做什么以及它如何影响公司的战略。为了准备迎接这一颠覆性时刻，公司应该如何组织领导团队呢？

从CTO到CAO：引领企业升级之人

企业永远需要引领技术的角色。如今一般企业都会有CTO或者CIO（首席信息官），为了应对智能革命，吴恩达在2016年11月撰文提出，每一个企业都需要一位CAO（Chief AI Officer，首席人工智能官）。

那么CAO从何而来？CTO、CIO与CAO的关系又是什么？

随着互联网全面渗透，CTO这个角色诞生了。他不同于以往的总工程师等，是特别应对互联网信息化浪潮的角色。在很多企业、单位，初级技术官的职能往往在于搭建一个内网平台，管理软硬件等。然而新的时代，每个CTO是否具有智能化业务的视野则不一定。举个例子，在一些高校商学院，CTO的职责仅限于根据需求来打造网络、办公系统等，不会洞察和利用商学院本身的业务带来的数据增益。比如很多商学院有课程案例销售工作，本可以智能化，对客户订阅进行跟踪分析。金融原本与数据紧密相关，但很多金融学院自身恰恰未能数据化。

很多企业的Web1.0信息化和数据化工作都还没有做好，所以有人提出企业需要CDO（首席数据官）或者CIO。CIO是CTO的升级，来帮助公司组织信息。他们思考如何提升公司信息的传递效率，如何能共享信息而不是各部门重复生成同类信息等。有调查显示，CIO里非技术专业出身的比例也不小，因为CIO不仅要懂技术还要懂业务。

CDO需要有洞察数据含义的能力，比如通过数据可视化技术直观地把握数据反馈；通过“数据挖掘”发现隐蔽的价值所在；通过优化企业数据架构，把无意识状态的数据变成主动的数据燃料。

过去一般IT部门的工作重点是选择、实施和集成ERP系统，管理公司的服务器和网络，使用计算机设备培训所有新员工内部软件和流程以及内部业务运营。具体项目大致包括：员工信息管理，尤其是对人员招聘和流动状况的分析；工资和成本管理；保险和福利；会计电算化；开发项目管理；服务器管理；存储；企业通信；营销自动化；客户关系管理；客户需求分析；企业知识积累等。

然而如今，大型ERP系统正在被云托管的专门的SaaS产品所取代。这些应用程序具有对消费者友好的用户界面，购买、使用和部署都很简单。企业不再需要几周的搜索来确定最佳的工资软件并花上几个月来实施。比如开发项目管理除了jira（项目与事物跟踪工具）外，甚至第三方办公协同软件也可以完成此类工作。服务器管理有亚马逊网络服务、百度云、阿里云等；企业存储有百度云盘；企业通信有企业版QQ，钉钉等。

信息技术服务已经被“流化”了。实施大型SaaS工具仍然需要一些集成工作，但过去几年的趋势已经表明集成难度在降低，企业应用程序在简化。

过去CIO还有一个主要职责是建立技术团队。然而，随着技术劳动力可以按照需求应变，公司可以到自由职业者平台上雇用码农去应对一些小任务，也可以通过平台寻找成熟的“即时”专家团队。各种在

线项目管理工具和资源使得管理外包服务也会变得更容易。在这种状况下，CIO原先业务的重要性在下降。

电力官早已消失，CIO也会消失吗？以百度为例，在前几年移动互联网热时，百度曾专门设置了“移动官”，负责思考和规划如何让百度的服务移动化，但随着移动互联网思维的普及，这个职位已经取消。

有识的CIO会适时转移自己的业务焦点。虽然第三方IT服务越来越标准化，很容易接入企业，但客户需求是不断变化的，不像企业自身的IT基本需求那样标准化，更不像电力能源的标准化。CIO要应对来自客户的信息变化，并对企业信息业务做出调整。

未来取决于CIO的视野。一个新任务就是负责组织完整的企业数据和智能战略。无论称这是CDO还是CAO。

CTO和CIO的开发任务减少了，因此获得了更多的时间和精力，应该尽早投入到钻研智能技术上。第三方IT服务依然是通用型的，而企业要做的是追求个性化定制，正如相同的外包公司与能力不同的企业产品经理结合，会开发出质量高低不一的产品。CIO或者CAO要审视企业自身的数据，寻找那些没有被充分利用的数据资源。

CAO要处理的信息远超过财务之类，还包括生产信息、组织信息，乃至员工的走位信息。想一下快递公司为什么需要掌握快递员的运动轨迹？CAO会考虑收集快递员的运动数据，采用数学方法加以分析，提出优化建议，提升快递效率。

数据在CTO眼中可能只是技术副产品，但是在CAO眼中则是战略性资源。他需要花费大量精力来思考客户需求和完整的用户链条，以寻找方法来增强公司的产品和服务，并尝试在所有产品上与客户展开互动。

比如在航空领域，中国南方航空股份有限公司作为中国运输飞机最多、航线网络最发达、年客运量最大的航空公司，为了把“大”转

化为“灵”，与百度合作，成为首家入驻百度“行云”出行大数据平台的航企，利用百度位置服务产品、数据及市场资源（如百度地图），为旅客提供包括机场室内导航、候机楼到市区的智能交通、行程规划、动态诱导等服务；发挥百度在云计算、大数据等领域的技术优势，建设基于地理定位服务的大数据分析模型，提供规划设计、统计分析、市场监测等决策支持，为旅客提供全流程、高质量、多内容、高效便捷的出行服务。

从互联网到大数据，再到人工智能，不只是概念变化，更是认识和实质的变化，是从形式到肌体再到灵魂的上升。无论CTO还是CDO，具备人工智能视野都是很重要的。

CAO将与公司的不同部门（人力资源、销售、市场营销、产品等）进行合作创新，通过数据连接、整合不同部门。在深入数据整合、互动、挖掘时，会越来越需要成熟的方法和工具，此时，接入人工智能流也就呼之欲出了。

首席人工智能官做什么

引入成熟的机器学习方法，把数据变成训练材料，塑造自动化机制，寻找、发掘有价值的管理模式和客户模式，或者合理引入第三方人工智能流，支撑企业业务发展，并找到一个合适的创新循环：数据—算法—知识—用户体验—新的数据，循环往复，滚雪球式前进，在这个过程中创造商业价值。这就是CAO做的事情。

CAO作为一个新角色会面临很多不确定因素。当他去寻找人工智能服务时，应该遵循什么标准？就像CIO要懂得选择标准合适的第三方服务，个人用户会关心诸如计算机显卡或者USB（通用串行总线）接口标准一样。人们经常询问百度，人工智能服务有业界标准吗？

吴恩达认为，人工智能在当下并不十分成熟，还谈不上统一服务标准，各家都在尝试。在这个层面，各个企业的CAO都很重要，因为他们也担负着探索的责任。但是，稳定、方便、好用的工程化产品应该是正确的方向，有助于人工智能的“标准化”。比如开发者将可以在某个人工智能平台上方便地调用接口，享有稳定的智能流输出，支持企业运行和创业，就像现在人们在手机和开发平台上创造App一样。百度大脑就是这样一个平台。

这一轮的人工智能形成热潮，原因之一是门槛降低——深度学习的原理尤其是算法原理都差不多，就看谁的积累雄厚、谁的数据丰富、谁的训练充分。就好比无数程序员都会用java语言编程，但编程能力大有差别，你只能尽力选择最好的。

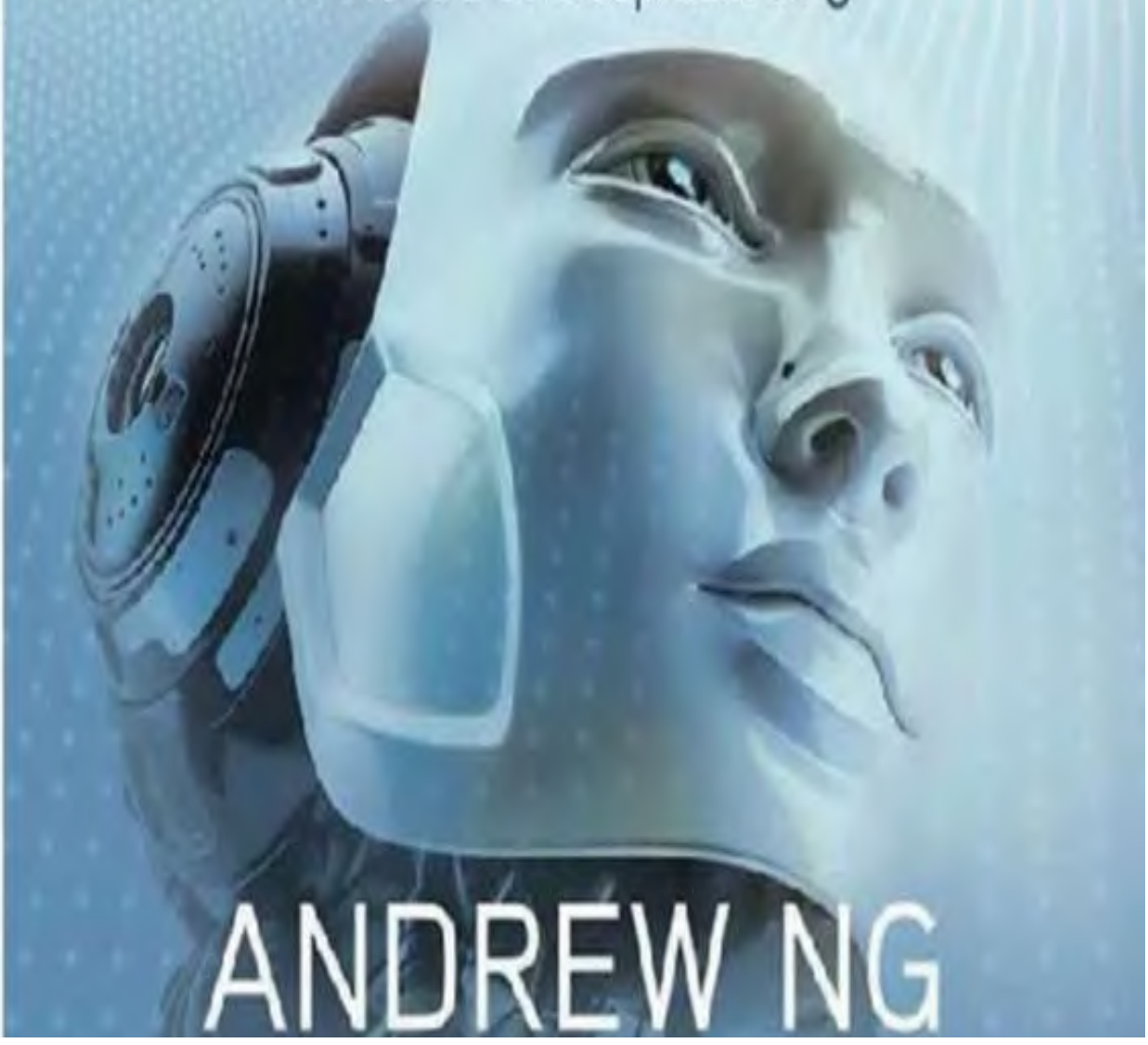
对于深度学习神经网络，调节神经网络参数的工作极其复杂和多变，需要根据不同业务场景做特别优化，虽然算法的基本思想差不多，但发展各异，也谈不上统一标准。因此，训练神经网络又被称作“炼丹”。吴恩达正在写作的新书《machine learning yearning》堪称一部“炼丹”手册，不是在讲“丹药”怎么好，而是在讲“炼丹”当中遇到的难题和解决办法。深厚的经验和务实的方法，大约就是一种标准吧。

期望每个高管完全了解人工智能是不合乎情理的。但如果所在的产业能产生大量数据，那就有很大的机会使用人工智能将这些数据转换成价值。对于大部分有数据但缺乏深度人工智能知识的公司来说，聘请一位CAO或人工智能副总裁（VP of AI）已经箭在弦上，一些CDO和有超前思维的CIO实际上已经在扮演这一角色。



LEARNING YEARNING

Technical Strategy for AI Engineers, In the Era of Deep Learning



ANDREW NG



图8-5 吴恩达新书封面

没有智能官的企业将被看作旧企业

著名数据分析公司MixPanel的CEO苏海勒（Suhail）认为，“机器学习不是为了证明某些原因，它的存在是为了针对一些特定的数据、行为或者是模式做出一些高质量的预测。算法唯一的工作就是让你能够更加有效、更加精确地达成目标，而不是告诉你为什么。”

不过我们认为问问“为什么”也是重要的。我们在大量的企业应用中发现，有些企业家急切地希望深度学习网络和数据挖掘能够马上带来利润的大幅度增加，而忽视了机器学习的规律以及企业自身的学习规律。

当企业大到一定规模，业务和数据复杂到一定程度之后，自身的运作逻辑往往是模糊的，不同于企业明文手册上的描述，即便是企业管理者对此也会感到陌生。机器学习具有根据数据反向求得函数逻辑的能力，这种反向推演能力可以给企业运作者提供一种观察企业非显性逻辑的视角。比如对企业邮件往来的关系分析，对电商用户下单时间的分析，对新闻网站发布资讯的时间段分析等，都可以帮助企业了解自身可以在哪些方面进行优化。

企业渴望从互联网时代迈进人工智能时代，这时CAO尤其需要分清轻重缓急，是先优化财务信息还是先优化产能？是固本强体还是拓展销售？

在对外业务上CAO可以带来显著的锦上添花的作用。比如建立自动化的客户需求分析系统，能够即时提交订单信息和生产信息给客户参考。但对内管理也十分重要。以制造型企业为例，生产管理、物料管理、质量管理、设计变更的信息化和数据化都需要整合，数据化达到一定程度后，辅以算法和开发，人工智能就水到渠成了。

电算化→信息化→智能化，神形合一，如果内部机制不顺，就会发现外部智能化缺少基础，难以为继。企业的内部修炼和外部拓展在智能时代将高度统一。

和电流从外部输入不同，在企业内部，智能的素材已经以“流”的形式出现，产生于每个员工、每个机器的每一个工作环节。这些数据流、信息流每天都在产生和流逝，但很多企业却没有意识到，更谈不上利用，而是任其挥发。

企业内功是和每个员工都相关的问题，比如个人的信息共享、知识管理、工作操作习惯的记录和优化等。

举个例子，网络新闻媒体的编辑每天都在录入文章，一些网络媒体拥有上千个编辑，每个人做着自己的工作。智能系统会侦测统计他们的录入操作，发现哪些环节的鼠标点击工作过多，这其中可能预示着录入系统的设计不合理，据此提出录入系统的优化建议。

吴恩达认为，传统企业尤其需要依靠CAO来了解前沿应用，升级自身业务。当然，面对很多可能的方向，CAO还是应该先抓住一个部门或者一块业务重点突破，形成典范，吸引其他部门燃起热情。毕竟，企业智能化是一个非常需要创新精神的事业，需要广泛参与。

而CEO一定要放权让CAO去施展，甚至亲自推进企业第一个智能化业务。没有CEO发自内心的支持，企业人工智能化工作将会很困难。有些举措可能旁人一开始不理解其价值，之后才会明白。如果错过就会非常可惜。

首席人工智能官的修养

作为与最具挑战性的深度学习技术打交道的CAO，他是一个特别理性的人，还是一个不失感性的人呢？

与一般工程科技不同，人工智能在模仿人类的思考习惯。科学家普遍强调人机结合是未来的发展方向。类似地，智能科学家和CAO都不会是机械论的化身，而是对各种知识具有全面的兴趣。吴恩达本人就是个例子，他不仅喜欢计算机科学，也喜欢人类学和社会学，还喜欢教育学——为了让更多人受到名校教育，把课堂教育重复使用率最大化，他合伙创立了Coursea在线教育课程项目，至今仍然是董事会成员。另外，他对心理学也有极大兴趣。

CAO可能是《生活大爆炸》里有趣的理工男，善于发现别人没发现的契机。也可能是数据领域的夏洛克·福尔摩斯，善于明察秋毫。不过他终究首先是一个具有“数据感性”的人。

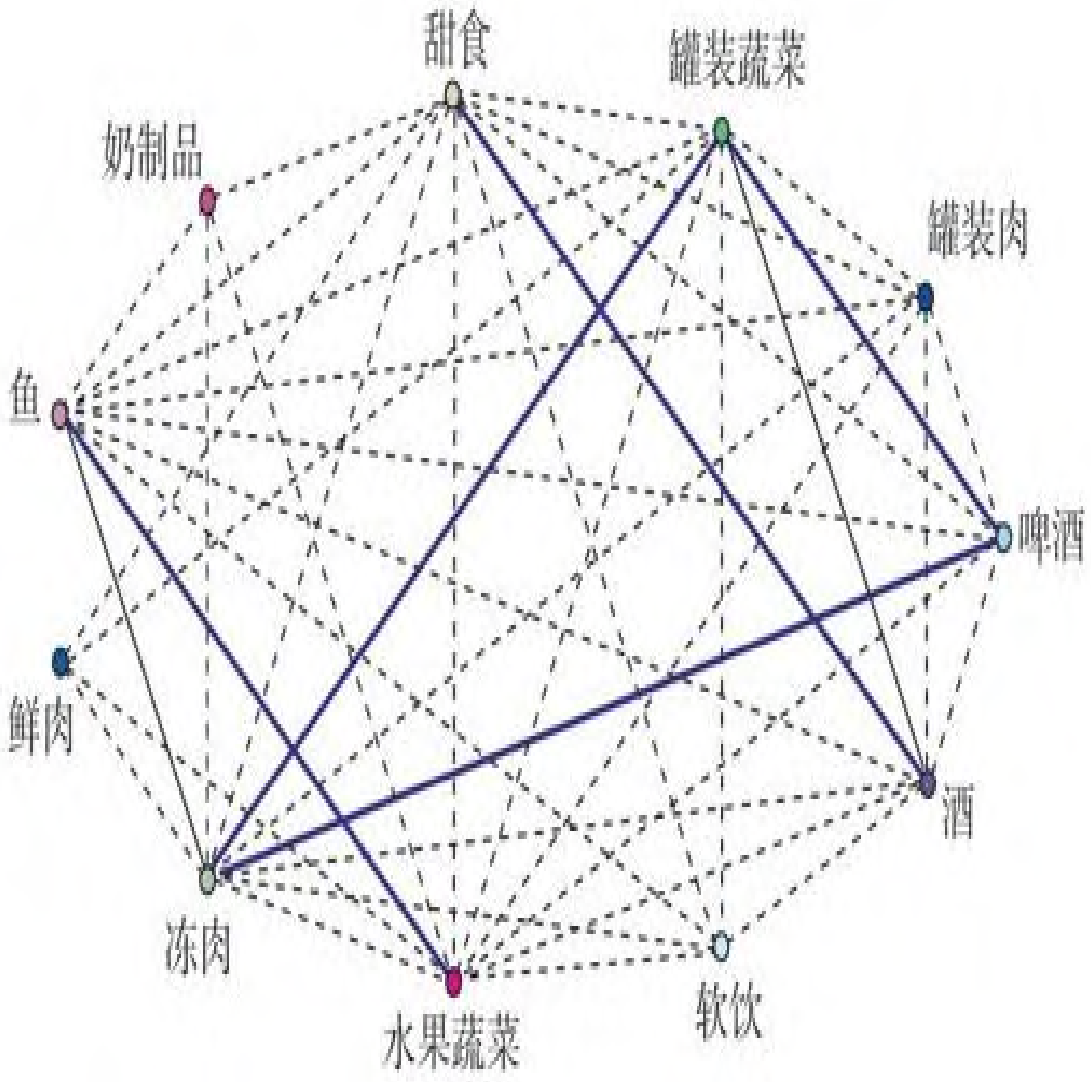
他懂得利用数据，最重要的是在别人看不到数据的地方看见数据。

第一层：有能力从纷繁的数据中嗅到可能有的关联

张亚勤在与很多企业打交道的过程中发现，前几年做大数据，买了一大堆服务器，收集到一大堆数据，放在那里不知道该怎么办。不收集数据觉得好像错过了什么，收集了却不知道该如何做。即便我们自己这样的人工智能企业，很多数据也一时用不了，但如果不用，以后可能就没用了。因此CAO就要在做什么上发挥才能。

沃尔玛的“购物篮分析”是典型的数据挖掘应用，数据分析师想到从顾客购物小票里发掘购买对象之间是否有相关性。经过数据统

计，他们发现除了牛奶和面包会经常被一起购买这种显而易见的关系之外，还有很多以往没发现的关系，比如婴儿纸尿裤常常会和啤酒一起被购买，然后才发现年轻的爸爸在购买婴儿用品的同时，往往会有一种犒劳自己的心理，于是购买了啤酒。



粗线表示概率较高的关联性

图8-6 使用SPSS软件进行购物篮数据分析的结果演示图

资料来源：

https://baike.baidu.com/pic/%E5%85%B3%E8%81%94/165901/0/08f790529822720ed

c57ba9579cb0a46f31fabc0?fr=newalbum#aid=0&pic=08f79052
9822720edc57ba9579cb0a46f31fabc0

沃尔玛超市因此把啤酒和纸尿裤放在货架上靠近的位置，显著提升了销售额。沃尔玛的数据毕竟有限，并且只有购物小票这种用户数据。倘若数据巨大，且可以自动化调整每天的货架，就具有了人工智能的雏形。亚马逊和阿里巴巴的电商大数据都大到了一定程度，因此它们都为人工智能科学家提供了舞台。

这个例子再次体现出CAO所需要的思维模式究竟是什么——正是我们一再强调的相关性思维。牛奶和面包、啤酒和纸尿裤之间的联系，显然并非是以往科学领域的确定性关系或因果关系，而是一种以概率形式展示的强相关关系。想到从这个角度去发掘价值，这就是CAO具有的素养。

第二层：在数据之上，CAO及其伙伴需要有场景感知能力，懂得场景计算

未来，“业务”的概念将会变得模糊，“场景”成为事物运作的核心，这意味着产品不再像过去那样围绕业务来分类，比如新闻App做新闻，电商App做购物等。现在你有什么样的场景，就可以导入这个场景所需要的一切服务，比如一款移动支付应用，要考虑各种支付场景，如果提供购买电影票功能就要根据用户行为数据猜测用户正处于什么使用场景，除了看电影，是否需要阅读影评，是否需要购买爆米花。企业要做的是根据场景重新规划，打通业务，而不是根据业务来划分产品架构。

从感性的层面说，对场景的敏感实际体现了CAO的生活趣味和人文情怀。老人关怀家居、助盲产品都来自这种关怀。2016年，视频创业发力，有人为此开发出了可以自动剪辑视频的智能剪辑手。只要用户输入关键词，比如“航母”，软件就会把视频中相关的段落大致剪出

来，这正是对视觉识别技术的非凡应用，来源于开发者对用户场景的敏锐。

比起一般科学家习惯于无功利地探索与发现不同，CAO需要更“功利”一些，即需要具有经济价值观，帮助企业发掘价值。这体现在优化流程、加速反应、节约时间，把人从重复性的机械劳动中解放出来。人工智能要让人有更多时间来做更有意义的事情，给人类的高级脑力活动留出时间。

在这些敏锐性层次之上，是具体的机器学习操作能力和智能系统搭建能力。吴恩达基于自己带领、扶持过的一些成功的人工智能团队的经验，针对如何寻找CAO给出了这样一些建议：

对人工智能和数据基础设施有良好的技术理解。例如，他们曾建立过重要的机器学习系统。在人工智能时代，数据基础设施（你如何组织公司的数据库并保证所有相关数据都被安全存储并可以访问）非常重要，尽管数据基础设施技能相对普通。

跨职能的工作能力。人工智能本身不是产品或业务，而是一种用来帮助已有业务线并创造全新产品和业务线的基础技术。因此，有能力了解不同的业务单位或职能团队并与他们合作很重要。

强大的企业家技能。人工智能创造建立新产品的机会，从自动驾驶汽车到能与之交流的代理，这在几年前还不现实，甚至是科幻小说中的东西。然而企业家往往是那种具有突破精神的人，能够创造从0到1的价值更新。一个掌握企业家技能的领导者将会增加公司成功创造出这种创新的概率。

吸引并留住人工智能人才的能力。这种人才备受追捧。好的CAO需要知道如何留住人才，例如注重人才感兴趣的项目、给团队成员培养技能的机会等。利用项目去教育别人，而不是事无巨细亲自打点，这样才能形成人才的培养循环机制。

吴恩达认为，合格的CAO应该有管理人工智能团队的经验。由于人工智能进化飞快，他们需要跟得上变化，但要求他们必须处于最前沿就没那么重要。更重要的是他们能跨职能部门进行合作。独立的技术是没有什么意思的，技术需要为人提供服务。机器与人的关系非常重要。要想做最好的人工智能，就需要既懂技术又懂人。人工智能是一场人机之间的故事。CAO既要善于理解机器智能，也要善于理解人情事理。

CAO应该成为一个布道者和指引者，让整个企业具备一种人工智能的兴趣和品格。CAO将是新时代的企业偶像，就像产品大师是互联网时代的英雄（乔布斯）。人工智能知识、产品与管理思维以及合适的人文感受力将共同造就这种英雄。

很可能在不久的未来，没有设置CAO、没有接入智能流的企业将会被看作是旧企业。从CTO到CIO、CDO，再到CAO，是一个循环上升的艰苦提升过程，某种意义上，也反映了企业自我更新、自我进化的艰难过程。

题外话

人才匮乏是当下的普遍问题，企业纷纷从学院挖人，智能科学家也不甘寂寞，纷纷跳槽到企业里去实现梦想。吴恩达的好友，全球排名前十的人工智能（深度学习）科学家Bengio（本希奥）担心，科学家都去企业了，高校里从事深度学习教育的人才就少了，会减少人才的产出。百度则把企业与学院结合起来，在内部进行教育，也向外输出了很多人才，并且准备在高校设立人工智能方面的奖学金。

09

技术奇点

——人工智能的自我挑战

机器人颠覆人类是很多科幻故事的主题，但对于“担当身前事”的科学家来说，他们聚焦的依然是当下人工智能面对的挑战和瓶颈。而这体现的依然是人的智慧。

目前全世界数据的爆发近乎失控，要想将数据进行分类计算需要极大的革新。从根本上说，人类尚未完全适应数据化生活，正如肉身至今也没有适应大工厂机器流水线的节奏一样，这是深层矛盾的来源。

矛盾就是动力。在某些方面，我们可以看到今日的矛盾和工业革命时代有类似之处。

飞梭与珍妮机的纠缠关系就是代表。1733年，飞梭的发明使织布速度大大加快。但问题马上来了，织布需要原料——棉纱，而纺纱的速度赶不上织布，只能依靠增加纺纱工和纱锭数量来弥补。1764年，让纺纱效率成倍提高的珍妮机被发明出来，纺纱的速度终于赶上了飞梭吃原料的速度。沿着珍妮机的思路，卷轴纺纱机和走锭精纺机相继诞生，这下轮到飞梭织布速度不够了，于是又推动了水力织布机的发明。两者在技艺上交替上升，互相激发。差不多同一时候，瓦特的蒸汽机出世，蒸汽原力觉醒，纺纱和织布部门都争相引入这股泥石流。工业革命就在无数机械的同声共鸣中一往无前，生生不息。

如何跨越数据的“马尔萨斯陷阱”

今天的人工智能与数据的关系也类似飞梭与珍妮机的关系。过去，人类构思出机器学习的方法，却苦于没有足够多的数据来验证和训练。互联网大爆发终于使数据不是问题，但如何处理爆发式增长的数据又开始考验硬件能力和计算能力。

巨头的数据烦恼

最早学会享用数据“螃蟹”的勇士之举，成就了BAT等大型互联网企业。巨头都曾对如何处理海量数据有深刻体会。

阿里巴巴早期使用Oracle数据库进行数据存储。这种互联网1.0时代的数据库架构很快就难以承受电商数据的爆发式增长。阿里巴巴不得不彻底换血，重金打造和使用自己研发的数据库。

京东在2013年以前经常因访问量暴增而造成服务器瘫痪，不得不更新后台架构，用java技术取代.net技术。

中国老百姓感受最深的应该还是几年前12306网站的购票灾难。过年要回家，这是融入中国人血脉的传统。但这样一个人口大国，每到春节就会上演“数字”灾难。对于实体世界中的火车线路来说，这是运力矛盾，每个人都感受过挤在车厢中动弹不得的痛苦，没有尊严，仿佛一个个冰冷的比特。这种矛盾通过高铁建设逐渐缓解。但同样的拥挤转移到了网络上。为了方便购票，铁道部对购票系统进行信息化升级，12306网站上线。不过当时并没有料到互联网化带来的数据挑战是什么。本想方便群众购票，却首先创造了不方便——上亿人同时查询、购买车票的行为让服务器迅速卡死。很多批评声音出现，认为程序员无能，认为换上电商工程师就能解决这个问题。

但真正关键的因素还是处理能力跟不上数据发展。有人专门比较了电商网站与12306网站。“双十一”时，淘宝等电商网站虽然也承接了海量人群的下单行为，但是这些单子被分布到数量巨大的商品上，彼此之间相关度很低，计算量也被服务器分摊了。火车票则不同，全国的班次就那么多，而火车票的抢购中，每一趟火车的千余个座位很可能面临数万甚至数十万人的抢购，火力极其集中。每发生一次购买行为，出票系统不但要分析该车次所有站点的数据，还要计算数十倍于车次出票数的抢票顺序数据，并实时更新沿线车站的可售票数，可以说是牵一票而动全身。数据和计算量呈几何式增长，而且一切还都

要在瞬间完成，即便不计成本地投入更多服务器也难以解决。这种难题是大电商也没有遇到过的，直到后来探索出新的计算架构和方法，才得以缓解。

BAT中最早面临大数据冲击的还有百度。“百度一下，你就知道”，全民搜索行为将海量数据发向百度服务器。日夜增长的网络信息也让百度内容爬虫疲于奔命。百度采用了预搜索和相关词搜索等方式缓解服务器遭遇的瞬时数据冲击问题。预搜索方式下，系统在搜索请求数量较低时（如凌晨）也在自动搜索并把搜索结果固化。在用户发送搜索请求时，系统就将已经整理好的结果推送过去，不需要服务器再把搜索任务跑一遍。相关词推荐也是利用系统相对空闲的时间以及功能架构清晰的数据库系统，对用户数据行为做相关性分析，比如当用户在搜索输入框输入TPP（跨太平洋伙伴关系协定）三个字母时，搜索框就会自动弹出下拉菜单提供搜索选择，比如：TPP是什么意思、TPP对中国的影响、TPP12个成员方、TPP协议等。当然，系统也会猜测少数用户表达的是“淘票票”的拼音缩写，也会列在非优先位置供用户选择。这些选项排列可谓善解人意，且能满足大多数人的需求。

在搜索结果页面下方，百度还提供了相关词搜索，比如美国新总统大笔一挥，签字退出前任费尽心机达成的TPP。这条新闻的相关搜索如图9-1所示。



图9-1 TPP相关词搜索结果

此外，搜索引擎还根据网友搜索热度排列出与TPP相关的热搜新闻，方便用户获取信息。



图9-2 与TPP相关的热搜新闻

这些都是通过对大量用户搜索的统计做到的，从而大大优化搜索体验，提升搜索速度，缓解数据处理压力。

可以说，数据引发的问题千奇百怪。数据并非均质的比特，而是和各种特殊人类活动场景相关，使得数据处理面临各种挑战。但从根本上来看，还是珍妮机与飞梭的矛盾——硬件的所有进步都会立刻被

计算量和数据量吃掉。虽然硬件能力发展速度也很快，以相同成本下每18~24个月翻一番的速率增长（也称为摩尔定律）[1]，而数据的发展速度远高于硬件，这是为什么？

数据世纪的“马尔萨斯陷阱”

人们都知道人口学的马尔萨斯假设——粮食生产呈算数增长，人口呈几何数增长。在单位亩产不可能有大突破的状况下，以粮食为主的生产资料只能依靠土地扩张，而在没有计划生育的情况下，家庭人口增长却是指数级别。于是人口很快到达天花板，人类遭遇粮食危机后陷入战争、饥荒、疾病等灾难，人口大幅度减少。工业革命、农业科技进步和人口管理缓解了马尔萨斯问题。今天这个类似危机又在虚拟世界里出现了。

大数据世界的马尔萨斯定律可以这样来描述：

人口以算数级数增长，数据以几何级数增长。

数据量以线性规律增加，计算量以非线性规律增长。

发达国家人口增加缓慢，有的甚至出现负增长。但世界所产生的数据始终在高速增长。这是因为数据赋能使所有人、所有人类活动都在产生数据，只要想记录，就有无数的数据可以“无中生有”。

很多电商网站早期只追求抢地盘、抢用户，重运营、轻数据，并不在意数据布局。举个例子，用户如果关闭了一个订单，过去电商系统并不记录这个行为，只会直接删除这个下单数据。但是后来发现记录和分析用户的失败交易数据也是有商业价值的，可以用于统计用户信用、喜好等，于是决定记录下来。每个行为都被记录，数据量就开始成倍增长。

数据的存储一直是一个大问题。据市场研究公司Forrester估计，每部智能手机1天产生的数据平均可达1G。全球智能手机用户数量保守估计达20多亿户，每天可产生20多亿G数据。如果用普通的1TB（1024G容量）硬盘存储，每天所需的硬盘数量就达200万块，1年就需要近8亿块硬盘。这已经远远超过全球硬盘生产商的产能。

更恐怖的是，产生数据的主要还不是人类活动。2014年，网站安全和内容分发公司Incapsula发布了一项统计数据：56%的网页浏览量都由爬虫机器人贡献。换句话说主要的互联网用户已经不是人类。大部分点击数据都是机器程序产生的。

Incapsula的数据源于全世界2万个每天至少有10个访问量的网站，它统计了90天的150亿次访问数据后计算得出这个结果。

非人类访问中，近一半来自善意机器人，如搜索引擎的内容爬虫，它们能够对网页建立索引，方便人们快速查找到对应的网页内容。百度和谷歌等就使用这种方法整理全世界的信息。另外超过一半的浏览量就来自恶意机器人了，比如盗取内容的爬虫、各种黑客工具、垃圾邮件等，而且比例还在不断增加，某种意义上，这就是互联网阴暗面的真实写照。这只是网页浏览领域的状况，在整个人类社会，随着信息化和物联网的快速发展，连接入网的所有硬件都在时刻生产数据，互相通信。发电机组上的检测芯片时刻侦测运行状况并把数据发回服务器，遍布城市的摄像头把监控数据上传指挥中心，家里的智能电视、冰箱等都在采集和上传数据……即便全人类都沉沉睡去，世界也依然在数据海洋中律动。

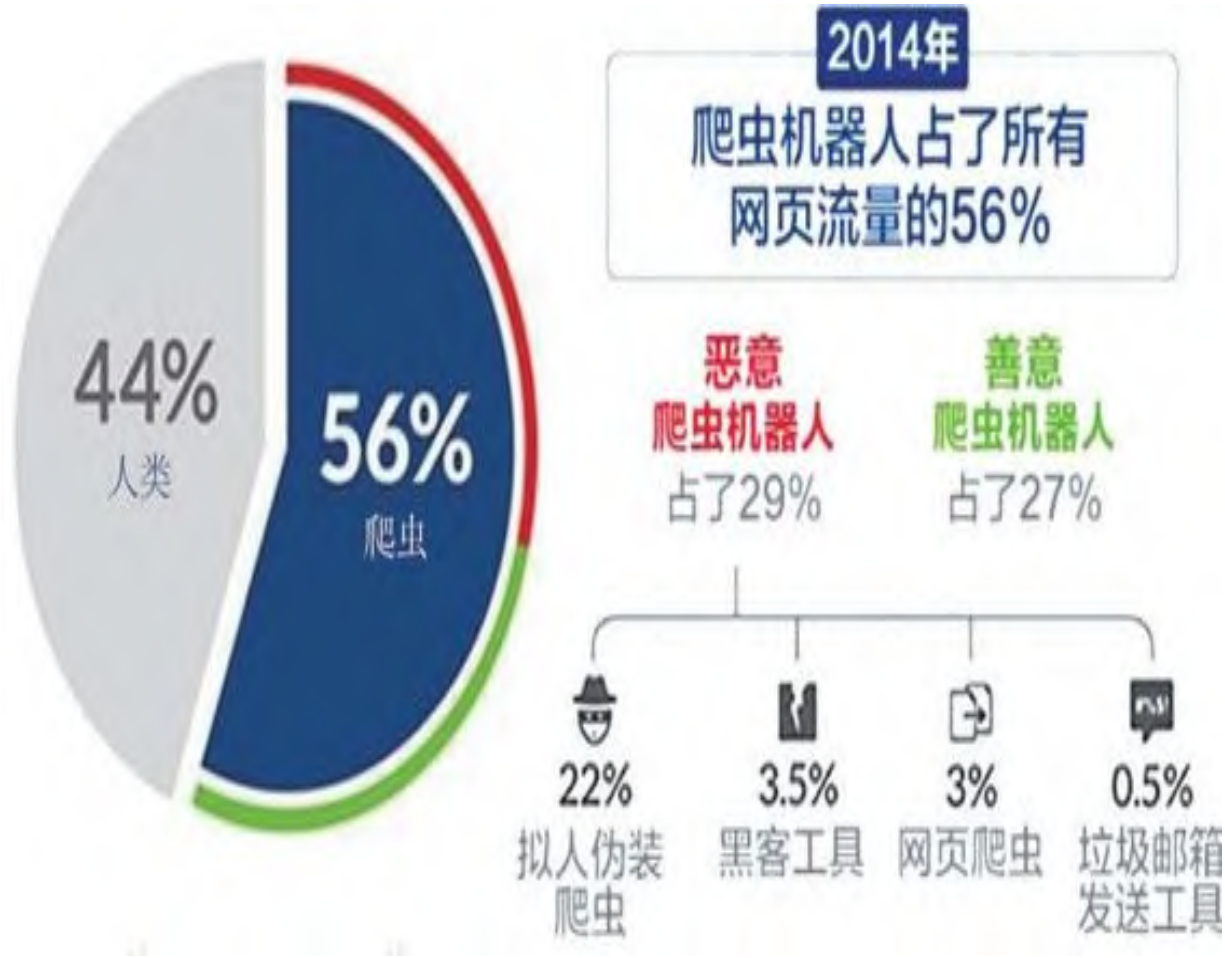


图9-3 2014年爬虫机器人贡献的网页浏览量占比

资料来源：www.incapsula.com

数据量以线性规律增加，计算量则相应地以非线性指数增加。数据要经过计算处理才能发挥价值，但随着数据量的增加，计算量会以更快速度增加，比如围棋格数不过是象棋的5倍而已，但是计算围棋要比计算象棋复杂亿万倍。对于电商和搜索引擎来说，即便对商品列表或者搜索结果进行排序，这个工作的计算量也会随着条数的增加，呈一条陡峭的曲线上升。

跨越数据陷阱

要跨越数据世纪的“马尔萨斯陷阱”，我们需要做三件事情：一是对并发的大量数据做出即时高效处理；二是高效存储数据并删除不需要的数据；三是对积存的数据资料进行深入挖掘。重点谈谈第一个。

对于第一点来说，最具有突破性的处理技术就是并行计算或分布式计算。庞大的数据计算任务分解开来，将每一小块任务分配给一台计算机。这些计算机分别运算完后将结果汇总就得到总计算成果。Hadoop和Spark就是代表性的技术。Hadoop的MapReduce（一种编程模型，用于大规模数据集的并行运算）功能可以将单个任务打碎，并将各个任务碎片（Map）发送到多个计算节点上，之后再以单个数据集的形式加载（Reduce）到数据库里。Hadoop能够在计算节点之间动态地移动数据，调节各个节点任务的动态平衡。这样做可以大量减少任务排队时间，将一般服务器的效率集合起来，而无须使用昂贵的超级计算机。这种技术最早在网页搜索中发挥作用，并很快成为分布式计算的平台。Spark技术可以看作对Hadoop技术的优化，把子任务中间的输出结果保存在内存中，从而无须频繁读取存储文件，加快了速度。这种优化仍然是基于对硬件性能的调配，对数据的基本处理方式还是批处理。

批处理是旧时代的续命者，也是半新半旧时代的过渡者。在数据爆发时代来临之前，人们处理数据的方法一般是先存储下来再进行分析。对静态数据的高效分析主要依靠批处理命令来完成。数据流时代的来临逼着人们要迅速处理，最好是在数据事件发生的时候就立刻做完分析判断，尽早和尽快追求近似性结果，比如人流数据、电网运行状况数据等。批处理技术并不是被设计来应对数据洪流时代的，只能差强人意地应付一下大数据流。

数据洪流的产生与激增，必然召唤一种“流计算”方式，这就是数据领域的最新挑战和创新方向。“流计算”不是一种具体技术，而是一种态度和方法的总称。

不是所有的应用场景都需要即时计算，但越来越多的业务正在变成时间的艺术。理想情况下，事件发生时，产生的数据就应该被立即处理分析。比如手机业务计费状态中，打完一个电话就应该完成计费而不是等待日后某时再来计算话费；新闻网站和电商网站的用户画像也应该及时完成，同步指导机器进行信息推荐，而不是先存储用户行为数据，等到事后再来分析；很多工业机组的异常检测也是如此。但想要实现即时计算会遇到很多困难，比如数据的汇总、通信、计算等。

Hadoop和Spark都是“流计算”的先声，但还远不能应付需求。2014年时，卡内基梅隆大学教授、ICML2014程序主席邢波就指出，大数据处理平台的大量资源都浪费在集群的通信上。即使比较优秀的平台，计算时间也只有20%，而通信时间占80%，比如Hadoop的通信时间就占到了90%。[2]

“流计算”驱动了智能的发展。说到底，智能是一种时间性能力。如果任务时间无限，滴水也能穿石，那就无所谓智能了。人类要想办法用工具迅速穿石，这就是智慧。“流计算”代替旧时代的批处理传统，还是个任重道远的工作，科学家已经从硬件架构和算法优化上做出探索，本书在此不做详细讨论。

至于巨大数据的存储问题和数据挖掘问题，如果能即时处理数据，得到分析结果，很多数据就不必存储下来，只要留下日志，比如侦测系统是否正常运行。对于必须存储的数据，人类想出各种策略来解决。大多数用户在PC时代就频繁使用的压缩软件就是精简数据体积的代表方法。压缩软件都是基于算法的工具，对数据进行重新编码，得到编码后的压缩数据和解码“钥匙”，就可以随时通过钥匙复原数据。人们不断改进文件格式，比如视频文件体积越来越小却可以保持清晰度。人们也把大数据策略引入存储，比如有多人向云盘上传同样的文件，系统可以只保留两个备份，其他人上传的文件则变成虚拟文件，指向同一个备份，这样就大大节省了空间。

“思维”倒逼“生理”革命——硬件基础设施创新

数据与计算量的增加，要求整个信息基础设施都做出相应的变革。前文描述的所有数据处理方法都还是在旧的信息基础设施上展开的。但机器智能的发展已经要求机器大脑做出“生理”性变化。

绿色机体

人类的大脑重量只占体重的2%左右，但是消耗的能量却要占到全身消耗量的20%——每天消耗总氧气量的20%，消耗肝脏储存血糖的75%。

在机器智能领域发生的事情也有类似之处。数据和算法都非实体，与物质、硬件相对，仿佛世界的“思维”。但运行这个“思维”需要庞大的物质资源和能源。在那些大型数据中心，除了层层叠叠的服务器外，还有大大小小的电源、环境控制设备、监控设备以及各种安全装置，像大脑这个器官一样昼夜运行。这个器官本身也消耗了大量的能源。

互联网是全天候无间断服务，服务器耗能巨大。据统计，2011年我国数据中心耗电量达700亿千瓦，占全社会耗电总量的1.5%，相当于天津市全年用电量。

2015年3月，中国工业和信息化部、国家机关事务管理局、国家能源局制定颁布了《国家绿色数据中心试点工作方案》，这个方案透露了几个数字：“随着信息化快速发展，全球数据中心建设步伐明显加快，总量已超过300万个，耗电量占全球总耗电量的比例为1.1％～1.5％，其高能耗问题已引起各国政府的高度重视……目前，美国数据

中心平均电能使用效率（PUE=数据中心总设备能耗/工厂设备能耗）已达1.9，先进数据中心PUE已达到1.2以下……近年来，我国数据中心发展迅猛，总量已超过40万个，年耗电量超过全社会用电量的1.5%，其中大多数数据中心的PUE仍普遍大于2.2，与国际先进水平相比有较大差距，节能潜力巨大。国际上普遍通过应用节能、节水、低碳等技术产品以及先进管理方法建设绿色数据中心，实现能源效率最大化和环境影响最小化。美国政府实施了‘数据中心能源之星’‘联邦数据中心整合计划’，欧盟实施了‘数据中心能效行为准则’，国际绿色网络组织开展了数据中心节能标准制定和最佳实践推广，建立了绿色数据中心的推进机制，引导数据中心节能环保水平的提升。”

具体该怎样做呢？

给机房“降温”是一项需要不断创新的工作。大公司可以选择将数据中心放在靠近南北极的寒冷地区，这也使冰岛成为最近几年一些重大数据事件的重要地点，比如现实中的维基解密事件和电影《谍影重重》中的内容。也可以采取海水制冷或者空气制冷的方式，节约电能，保护环境。

太行山麓的山西阳泉市是一座历史悠久的古城。在阳泉以西200公里是1400年前唐高祖李渊兵发中原的大本营。阳泉东北大约50公里处的娘子关电厂，数年前孵育出了构建“宇宙社会学”的世界级科幻小说《三体》。阳泉是煤矿基地，空气污染不亚于北京，早已面临产业升级的问题。历史与未来的紧张感如同雾霾一般弥散于此地山川。

2015年百度云计算（阳泉）中心（以下简称阳泉云计算中心）在这里投入使用。建成后的数据中心存储量超过4000PB，可存储的信息量相当于20多万个中国国家图书馆的藏书总量；数据中心CPU总数高达70万个，CPU内核总数超过300万个；由于使用了高性能、低功耗服务器和多项适用中国环境及法规的科技来提升数据中心的整体能源效率，其PUE小于1.3。相当于每1.3度电进入机房，其中有1度电用于数

据计算，0.3度电用于数据中心的散热等其他用途，在绿色节能方面达到亚洲一流水准。

秉承开放精神，百度联合腾讯、阿里巴巴、中国移动、中国电信等行业翘楚共同成立了中国首个硬件开源项目——天蝎计划。其宗旨是打造一个开放的技术标准，研发定制化的整机柜服务器解决方案，从而满足数据中心海量计算与存储需求，有效降低数据中心采购与部署成本。

2014年9月，天蝎计划整机柜服务器技术规范迭代升级到2.0版本，对空间利用、散热策略等有了更加精细的优化定义，模块、接口、协议均有了详细的标准化定义。基于该标准，在阳泉云计算中心，通过各类资源的整合设计，实现了两位数的功耗节约。互联网快速迭代精神在这个计划中体现得淋漓尽致，全新的3.0规范更加强调模块化，同时规范的细节更加详细，可执行性更强。

2015年5月，阳泉云计算中心太阳能光伏发电项目成功并网发电。这是太阳能光伏发电技术在国内数据中心的首例应用，每年可减少107.76吨二氧化碳排放量，节能高达43％。

计算机架构革新

不过，节能减排办法是外部改变，好比用节能空调为发烧病人吹冷风，计算机更需要内部革新。正如批处理是旧时代的产物，现有的服务器和数据中心架构也是在旧时代的计算机技术架构上搭建起来的，半新半旧。

传统的计算机核心架构基于冯·诺依曼结构：数据存储和处理分离，计算逻辑呈线性分布。计算芯片执行指令代码，把结果存储到内存中，供下一个计算指令来调取，依次循环。这样的架构对于人类来说逻辑很清楚，但速度大受影响。而且在这样一种线性流程下，由于

CPU可能执行任意指令，就需要有指令存储器、译码器、运算器、分支跳转处理器等共同工作，分配指令执行的先后顺序。控制指令流的逻辑复杂，难以有太多条独立的指令流，并行处理能力低下。

摩尔定律行将式微。目前计算机内存运行方面的提速仅为每年9%，硬盘性能的提速平均每年只有6%，计算机内存的运行速度只有CPU速度的几百分之一，成为瓶颈。数据存储吞吐的模式严重拖累了计算机性能。

早期就有人提出一种可自改变架构的计算机。以个人计算机为例，属于通用任务型，即便执行打字之类的简单任务，整个计算机系统都在运作，资源浪费严重。而可自改变架构的计算机能针对复杂程度不同的任务，有节制地调用计算机不同部分，不会大事小事都倾巢而出。真正具有现实性的计算机革新方案在新科技发展中找到了方向。

一个方向是物理学的突破应用，比如令人神往的量子计算，利用量子物理学中的量子态叠加效应，创造性能百万倍于现在的计算机芯片。用光流代替电流传递数据和运算，也是提速的一个方向。另一个方向就是伴随脑科学和深度学习的兴起，有希望模仿人脑原理开发神经学芯片，其运行速度也将比现有的计算机快多个数量级。

人们从多条路径发起尝试。深度学习科学家利用GPU代替CPU组机器学习就是破天荒的一步。GPU使用SIMD（单指令多数据流）来让多个执行单元以同样的步调处理不同的数据，原本用于处理图像数据，但也特别适合处理深度学习任务经常面对的非线性离散数据。百度利用大规模的GPU集群，加以工程优化，开发出自己的GPU服务器，大大提升了硬件性能。但GPU毕竟也是建立在冯·诺依曼结构之上的。

FPGA芯片是当下热门的另一种发展尝试。它本是用于解决专用集成电路的一种方案。专用集成电路是为特定用户或特定电子系统制作的集成电路。以往数字集成电路因为其通用性和规模生产，使电子产品成本大幅度下降，但同时也产生了通用与专用的矛盾以及系统设计

与电路制作脱节的问题。集成电路规模越大，组建系统时就越难以针对特殊要求加以改变。为解决这些问题，就出现了允许用户参与设计特征的专用集成电路，也就是FPGA。

复杂并行电路的设计思路被搬到了计算芯片上。FPGA计算芯片布满“逻辑单元阵列”，内部包括可配置逻辑模块、输入输出模块和内部连线三个部分。它们是既可实现组合逻辑功能又可实现时序逻辑功能的独立的基本逻辑单元模块，以硬件描述语言定义各自的逻辑和彼此的关系。与冯·诺依曼结构不同，冯氏结构中内存有两种作用：一是保存中间计算结果，二是执行单元间通信。由于内存是共享的，当多个指令要求调用内存时，就需要做访问仲裁，依次调用；FPGA中的寄存器和片上内存（BRAM）各有自己的控制逻辑，无须多余的仲裁和缓存。FPGA每个逻辑单元与周围逻辑单元的连接关系由于可编程，可以预先确定，无须通过共享内存来通信。以并行运算为主，可以执行多指令流、多数据流，大大节约了计算时间。还可以针对不同的应用场景在硬件上做特别编程，灵活性高。

百度在2012年开始部署FPGA，是中国最先引进FPGA的公司，也是全球最先用FPGA做集群的公司之一。张亚勤说，一开始是CPU，后面加了GPU，基本上所有做人工智能的公司都用GPU。但是用FPGA也有好处，一个是整个架构的速度和效率提高了很多，另一个是GPU处理图像和语音数据比较好，但是也有很多别的通用计算用FPGA更快。因为FPGA可编程，可以很快改变架构，百度用的FPGA目前在GPU和CPU架构上面效率提高了5~6倍，而且不需要改变现有的架构，直接用FPGA就可以加速。

从计算角度看，网络的传输往往是最重要的瓶颈。百度在整个网络通信上投入了最先进的技术，使用100G的RDMA[3]，在GPU之间、FPGA之间通信。所以在集群之间，在数据库之间，数据可以快速无缝传输。

FGPA相当于用硬件来做软件的算法，在实现复杂算法方面有一定难度。目前是配合GPU、CPU架构一起工作。

既然概率计算是大数据和人工智能常用的数学方法，有人受此启发，提出概率芯片的概念。采用概率算法替代以往的微积分算法，牺牲一点计算精度，但是能极大提升计算速度并降低能耗，适合于无须追求极致精度的应用场合，比如物联网。

深度学习兴起之后，芯片科学家受到极大启发。目前最前沿的芯片革新还是来自以深度学习原理为基础的人工神经网络芯片的探索。Intel、IBM、NVIDIA（英伟达）等各大公司都纷纷为此打造自己的芯片发展方向。中国的寒武纪科技公司等推出的深度学习芯片已经具有国际领先水平。

人工神经网络是模仿生物神经网络的计算架构的总称，由若干人工神经元节点互连而成，神经元之间通过突触连接。这里，每个神经元其实是一个激励函数，突触则是记录神经元间联系的强弱权值。

神经网络是多层的，一个神经元（函数）的输入由与其相连的上一个神经元的输出以及连接突触权重共同决定。所谓训练神经网络，就是通过输入大量数据和监督，调整输出结果，这个过程就是不断自动调整神经元之间突触权重的过程，直到输出结果稳定正确。然后在输入新数据时，能够根据当前的突触权重计算出输出结果，这样就实现了神经网络对已有知识的“学习”。也就是说，神经网络中存储和处理是一体化的，中间计算结果就化身为突触的权重。

传统的处理器（包括x86和ARM芯片等）受制于冯·诺伊曼结构，在处理深度学习神经网络任务时效率低下，不仅存储和处理分离，而且其基本运算操作为算术操作（加减乘除）和逻辑操作（与或非），往往需要数百甚至上千条指令才能完成一个神经元的处理，因此AlphaGo才需要那么多块芯片（分布式版本有1202个CPU和176个GPU）。

为深度学习专门打造的芯片则不同，以寒武纪开发的DianNaoYu为例，指令集直接面对大规模神经元和突触的处理，一条指令即可完成一组神经元的处理，并对神经元和突触数据在芯片上的传输提供一系列专门的支持。在现有的工艺水平下，单核处理器平均性能超过主流CPU的100倍，但是面积和功耗仅为1/10，综合效能提升可达三个数量级。

当然，神经网络芯片只是在处理人工智能任务时相对于传统CPU有优势，适合认知图片、语音识别等领域，而像运行数据库、Office（办公软件）、微信等任务还是更适合传统芯片来处理，除非这类任务本身也发生了结构性的革命。

神经网络进化哲学

作为当下人工智能最基础方法的深度学习及神经网络技术，其自身的发展决定了人工智能的发展能力。这种模仿人脑机理的技术，也表现出类似生物进化的特点。

竞争还是合作？这是人类经常纠结的一对概念问题，但大自然的哲学并非如此。近年来，不同学者都提出所谓的共同进化的观念。黑脉金斑蝶（又叫帝王蝶）与植物马利筋的关系就是典型的例子。



图9-4 黑脉金斑蝶与马利筋
资料来源：http://www.wiworld.com.cn/Item/39660.aspx

马利筋的汁有毒，其花蕊的封闭构造使其难以通过风力传播花粉，但可以通过蜜汁吸引蝴蝶，通过蝴蝶来传粉。黑脉金斑蝶幼虫以

马利筋的嫩茎叶为食，能将后者的毒素存在体内以防御敌人。如果蝴蝶幼虫过度吞噬茎叶，马利筋就会死亡，而有些马利筋变异出更封闭的结构，能阻碍蝴蝶进入。有些黑脉金斑蝶也在变异中强化了侵入马利筋花蕊的能力。于是二者在对抗中变得越来越难分难解，黑脉金斑蝶不吃别的植物，马利筋不欢迎别的昆虫，没有第三者能参与它们的游戏。病毒与反病毒软件、黑客与反黑客工作则是发生在互联网上的另一个共同进化的例子。如今机器学习已经被运用于网络安全，相比过去基于设定特征的防火墙，效率大大提高。所谓共同进化就是：不是你死我活，也不是一团和气，而是在对抗中一起升级。

人工智能也处于一种共同进化的过程中。在万千变化的神经网络里，这种过程得到惟妙惟肖的展现。以下介绍两种新的神经网络的思想。

对抗生成网络

有监督的深度学习就是输入的数据有语义标签，输出的结果由人类标识对错。但很多科学家认为无监督学习才是未来的发展方向，让机器自己从原始数据中发现规律。已经有很多不同的方法，强化学习是其中一个方向，对抗生成网络则是已经在运用的事物。

对抗生成网络的发明者Ian Goodfellow（伊恩·古德费洛）是Yoshua Bengio（伊舒·本西奥）的学生，现在马斯克建立的OpenAI实验室工作。著名的深度学习专家Yann Le Cun对对抗生成网络大加赞赏。这种网络非常能体现“进化”的纠缠演进特性。

对抗生成网络的发明，首先是因为Christian Szegedy（克里斯坦·赛格德）等人在ICLR2014（国际学习表征会议）发表的论文中提出了深度学习对抗样本（Adversarial Examples）的概念，即在输入的数据集中故意添加细微的干扰，形成输入样本，导致深度神经网络得

出错误的输出。这个错误在人类看来一目了然，但是机器却一再跌入陷阱。[4]

Ian Goodfellow，Jonathon Shlens（乔纳森·舍琳）和Christian Szegedy在论文《Explaining and Harnessing Adversarial Examples》中给出了一个典型：

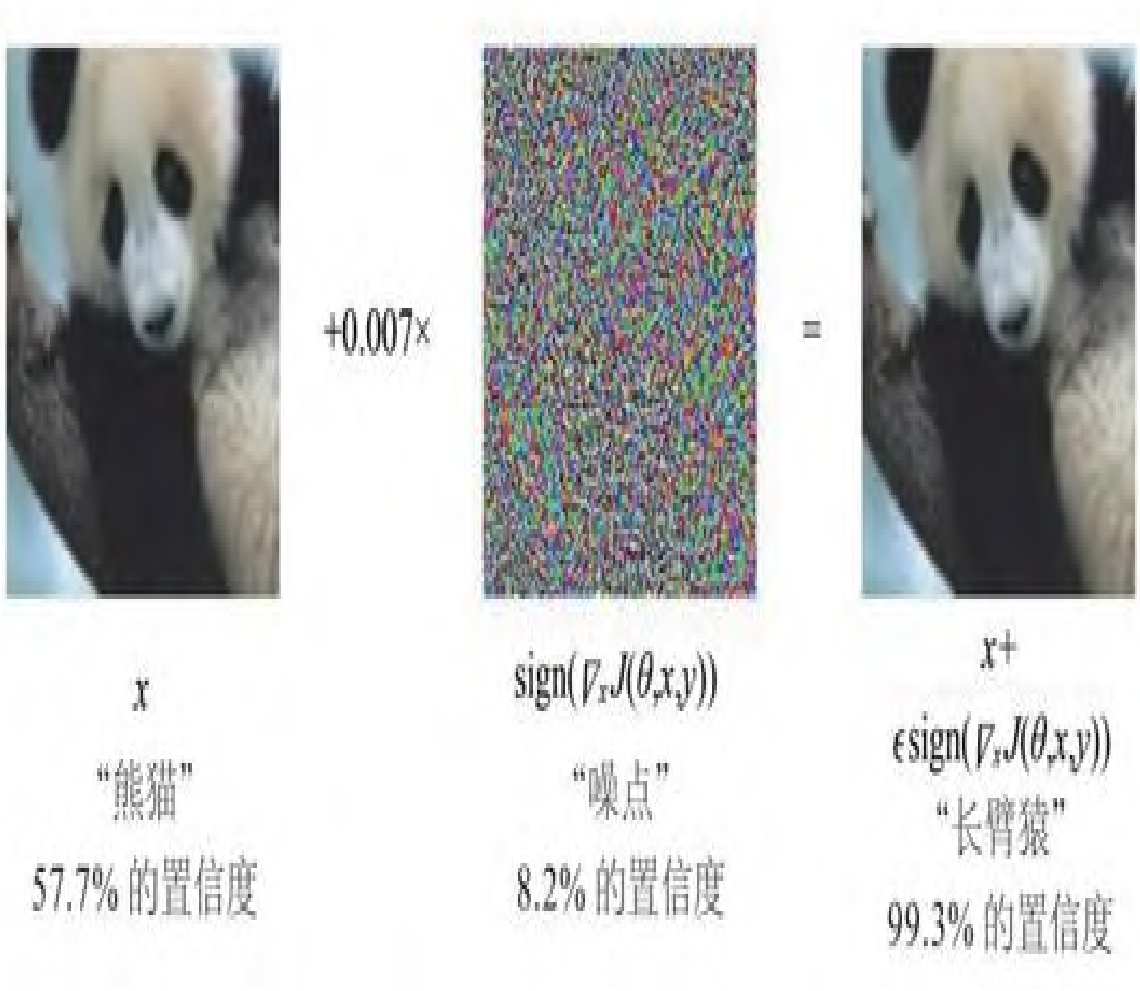


图9-5 深度学习对抗样本
资料来源：https://arxiv.org/pdf/1412.6572v3.pdf

在第一张图中，原始图像是熊猫，神经网络以57.7%的置信度判断为“熊猫”。

然后人类给图片加入微小的干扰，也就是第二张图所示的噪点。使用32位浮点值来执行修改，不会影响图像的8位表示。

最终得到第三张图。人眼完全看不出差别，但此时神经网络却诡异地以99.3%的置信度判断此图为长臂猿。

因为对抗性样本导致识别错误，有人将其当作深度学习的深度缺陷（Deep Learning’s Deep Flaws）。可是加州大学圣地亚哥分校的Zachary Chase Lipton（扎卡里·柴斯·立顿）在KDNuggets（美国一个大数据媒体）上发表文章，标题针锋相对，叫作（Deep Learning’s Deep Flaws）’s Deep Flaws，即《〈深度学习的深度缺陷〉一文的深度缺陷》[5]。该文认为深度学习对于对抗性样本的脆弱性并不是深度学习所独有的，在很多机器学习模型中普遍存在，进一步研究抵抗对抗性样本的算法将有利于整个机器学习领域的进步。

科学家抓住了“进化”的脆弱性本质，将错就错，把对抗看作一种训练办法，变阻碍为动力，艰难提升。大自然的进化本身就是高度脆弱的，无数生物“程序”被大自然淘汰，因为它们“出错”了。错误，就是进化的终极工具。而智慧就是在这个方生方死的过程中艰难升起。

对抗生成网络即人类对神经网络进行了特别设计，让其主动产生干扰数据来训练网络的能力。简单地说，对抗生成网络由两部分组成，一个是生成器（generator），另一个是鉴定器（discriminator）。生成器好比是一个卖假货的奸商，但是制造的山寨品高度仿真，而鉴定器好比高超的买家，需要鉴别货品真假。奸商的职责是想方设法欺骗买家（生成对抗性样本），后者则通过这种历练不断吸取教训，减少受骗概率。双方都在不断努力以达到目的，同时在彼此的“监督”下提升。看上去仿佛军事演习中的蓝军与红军展开激烈对抗，由此强化双方战斗能力，但没有硝烟。

这又是一个“共同进化”的例子，是进化的深刻哲学，不是战争，而是纠缠，是“在持久的摇摇欲坠中保持平衡”。

就对抗生成网络来说，我们要的是这个成熟的买家，还是那个高超的奸商呢？答案是都需要。二者是共同进化的必然要素。

奸商模型有什么用处？在很多情况下，我们会面临缺乏数据的情况，但可以通过生成模型来补足。制造样本，产生类似监督学习的效果，但实际上是非监督学习。

来自英国谢菲尔德大学的Wei Li（李伟）、Roderich Groß（罗德里赫·格鲁）和美国哈佛大学的Melvin Gauci（梅尔文·高斯）一起，基于对抗生成网络，开发了一种新的图灵学习方法，用于研究群体行为。[6]比如一群鱼中混进一些模仿鱼运动的假鱼，如何判断模仿行为的逼真度呢？使用传统的特征归纳法来区分是很难的，同一群鱼每次表现出的运动特征也不一定相似。这个团队决定让机器通过互相模仿自动建立群体模型，让机器自主推断自然物与模仿物的行为。该深度学习同时优化两种群体计算机程序，一个代表模型的行为，另一个代表分类器。该模型可以模仿监督学习下的行为，也可以辨别系统和其他模型之间的行为。

具体来说，他们建立了三种机器人群体，第一种是被模仿对象，按照事先指定的规则进行复杂运动；第二种是模仿者，混入到第一组机器人中，尽力学习和模仿第一组的行为，尽力欺骗鉴定者；第三种是鉴定者，它的任务就是区分那些运动中的群体里谁是模仿者，谁是被模仿者。随着鉴定者能力的提高，模仿者的模仿行为也会越来越逼近被模仿者。于是，我们便可以利用训练好的模仿者搭建一个逼真的多主体模型，来对被模仿者群体进行模拟。这种模型就可以用于研究集体运动，比如可以根据摄像头记录的节假日热门景点的人群运动来训练出一个模型，强化对人群运动趋势的预测，对可能发生的拥堵踩踏事故发出预警。

机器的进化迭代过程比大自然快亿万倍。在这种对抗生成中，机器习得的逻辑已经远远超出人类的理解，可能成为一种“黑箱”。是追求“黑箱”，还是追求“白箱”，如何避免“黑箱”带来的不可知危险，这是对人类的一个挑战。

对偶网络

对偶网络仿佛对抗网络的一种镜像。

目前多数神经网络的训练依赖有标签的数据，即监督学习。而给数据标注标签是一项繁重的工作。据报道，谷歌的开源图片数据库Google Open Image Datasets中含有900万张图片，YouTube-8M中包含了800万段被标记的视频，而ImageNet作为最早的图片数据集，目前已有超过1400万张被分类的图片。这些精心标记的数据，大部分是由亚马逊劳务外包平台Amazon Mechanical Turk上5万名员工花费两年时间完成的。[7]

如何让机器在缺少标注数据的条件下工作，是未来人工智能的发展方向。2016年，微软亚洲研究院的秦涛博士等人在向NIPS（神经信息处理系统大会）2016提交的论文中提出了一种新的机器学习范式——对偶学习。大致思想是：

很多人工智能的应用涉及两个互为对偶的任务，例如从中文到英文的翻译和从英文到中文的翻译就互为对偶、语音处理中语音识别和语音合成互为对偶、图像理解中基于图像生成文本和基于文本生成图像互为对偶、问答系统中回答问题和生成问题互为对偶、在搜索引擎中给检索词查找相关的网页和给网页生成关键词互为对偶。这些互为对偶的人工智能任务可以形成一个闭环，使在从没有标注的数据中进行学习成为可能。对偶学习最关键的一点在于，给定一个原始任务模型，其对偶任务的模型可以给其提供反馈；同样的，给定一个对偶任

务的模型，其原始任务的模型也可以给该对偶任务的模型提供反馈。从而这两个互为对偶的任务可以相互提供反馈，相互学习、相互提高。[8]

对偶网络利用这样一种精妙策略大大减少了对标注数据的依赖，我们从中可以再次洞见某种进化的哲学：进化是一种自我应答和自我循环的过程，从A到B，从B到A，互为镜像，但镜子并不清晰，各自掌握一半的秘密，没有仲裁，却可以在彼此猜测、参照中摇摇晃晃地前行。

深度学习的新边疆

以上两种神经网络方法只是不断涌现的新方法的典型代表。在深度神经网络方法之外，科学家也在积极探索其他路径。南京大学著名的机器学习专家周志华教授在2017年2月28日发布的一篇论文中，就与联合作者冯霁一起提出了一种创造性的算法，可以形象地称为“深度森林”（gcForest）算法。顾名思义，相对于深度学习强调神经网络的层数，这种算法重新利用了传统的“决策树”算法，但是强调“树”的层次。多层决策树的联合就形成了“森林”，通过精巧的算法设置，在数据规模和计算资源都比较小的情况下，在图像、声音、情感识别等应用上，都取得了不输于深度神经网络的成绩。这种新方法对参数设置不敏感，而且因为基于逻辑清晰的“树”方法，可能比深度神经网络更容易进行理论分析，从而避免人类难以理解机器具体运作逻辑上的“黑箱”问题。

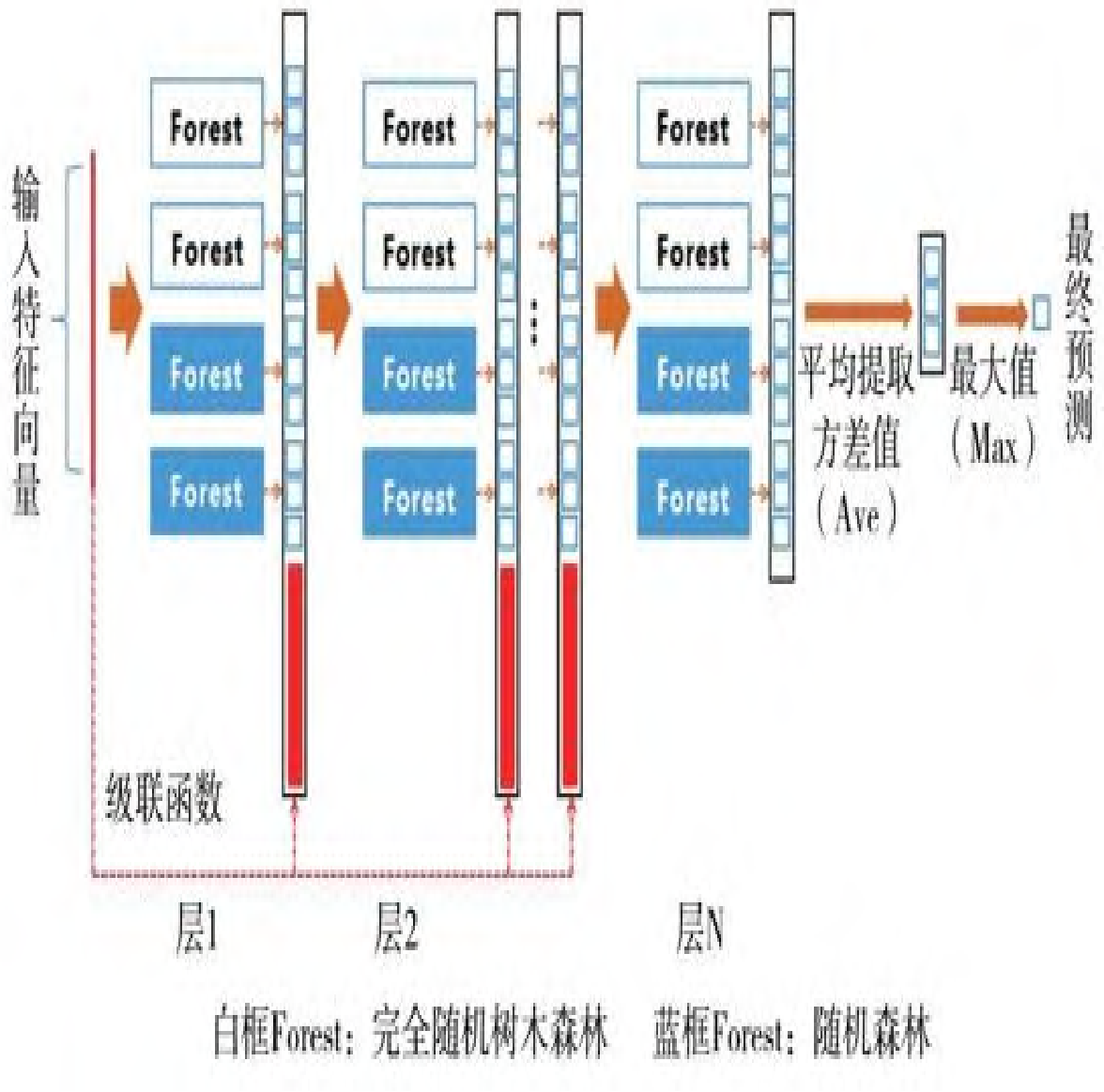


图9-6 多粒度级联森林结构

资料来源：https://arxiv.org/pdf/1702.08835.pdf

表9-1 在人脸识别上的精确比较

	一张图	五张图	九张图
多粒度级联森林	63.06%	94.25%	98.30%
随机森林	61.70%	91.20%	97.00%
深度神经网络（卷积神经网络）	3.30%	86.50%	92.50%
基于 RBF 核函数的支持向量机	57.90%	78.95%	82.50%
K 最近邻算法	19.40%	77.60%	90.50%

表9-2 在GTZAN数据库中的测试精确度比较

多粒度级联森林	65.67%
卷积神经网络	59.20%
多层神经网络	58.00%
随机森林	50.33%
逻辑回归	50.00%
基于 RBF 核函数的支持向量机	18.33%

资料来源：https://arxiv.org/pdf/1702.08835.pdf

据智库“新智元”向周志华教授了解，“深度森林”的方法论意义在于探索深度神经网络以外的算法可能。深度神经网络的有效运作，需要巨大的数据和计算能力，深度森林有可能提供新的选择。当然，深度森林依然向深度神经网络借鉴了关键思想，比如对特征的提取和构建模型的能力。所以，它依然是深度学习的一个新颖分支。

中国科学家在人工智能研究方面有很多世界领先的成果。我们认为，自信心与开放心态将是促进科学进步的重要动力。

如今，各大研究人工智能的科技公司都提倡共享算法代码。其中以谷歌的Tensorflow深度学习开源平台影响最广。但是很多深度学习科学家认为，从生态的角度考虑，应该有多个深度学习代码平台平行竞争才会有利于繁荣发展和平衡，不可垄为一尊。除了Caffe和Mxnet等深度学习开源平台外，百度在2016年9月开放了全新的PaddlePaddle

深度学习开源平台，采用新架构，对序列输入、稀疏输入和大规模数据的模型训练有着良好的支持，同时支持GPU运算，支持数据并行和模型并行，仅需少量代码就能训练深度学习模型，大大降低了用户使用深度学习技术的成本。多元的共享平台支持机器学习工作者从不同角度训练和创造种种应用，仿佛生物多样性一样有助于人工智能的提升。

归根结底，即便在遥远的未来，人工智能真的能够强大到统治世界，那在这之前，所有的挑战也都是人类自身智慧的挑战。人工智能科学家身上闪烁的智慧之光，为后来者照亮了方向。即使非人工智能从业者，想必也能从中得到很多策略启发。

2017年初，AlphaGo的变身Master横扫中韩围棋顶级高手，一时间人们分化为悲观派、降临派、冷静派、脑洞派……我们希望有更多人是默默学习派。

---

[1] 摩尔定律是由英特尔（Intel）创始人之一戈登·摩尔（Gordon Moore）提出来的。其内容为：当价格不变时，集成电路上可容纳的元器件的数目，约每隔18~24个月便会增加一倍，性能也将提升一倍。

[2] http://tech.china.com/news/company/domestic/11066129/20141213/19102201.html.

[3] RDMA通过网络把数据从一个系统快速移动到远程系统存储器中，而不对操作系统造成影响，这样就不需要用到多少计算机的处理功能。它消除了外部存储器复制和文本交换操作，因而能解放内存带宽和CPU周期，改进应用系统性能。

[4] https://arxiv.org/pdf/1412.6572v3.pdf .

[5] http://www.kdnuggets.com/2015/01/deep-learning-flaws-universal-machinelearning.html.

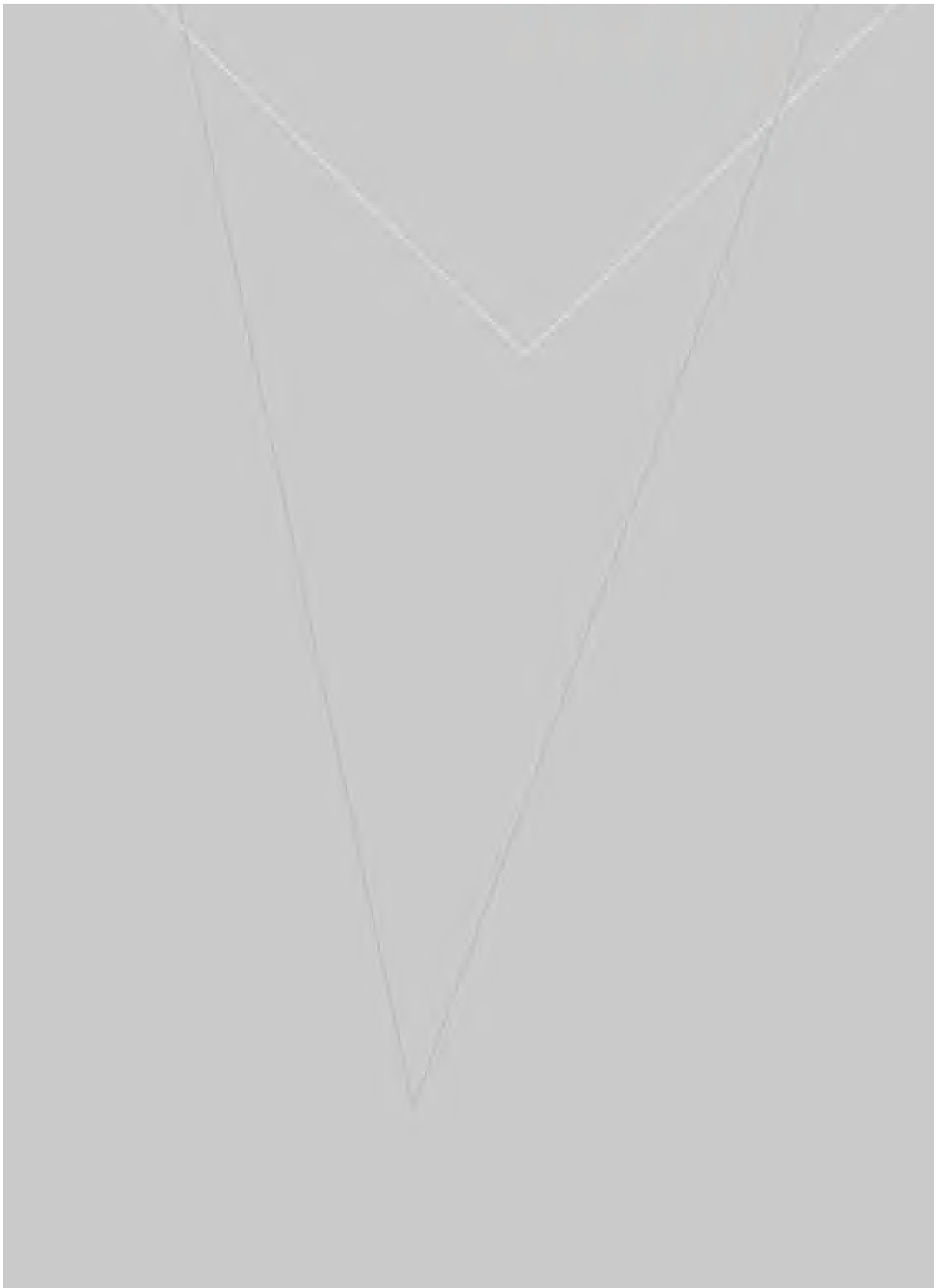
[6] https://arxiv.org/pdf/1603.04904v2.pdf.

[7] 雷锋网：www.leiphone.com/news/201612/1KhB2H10bMc2a8hu.html.

[8] http://www.msra.cn/zh-cn/news/blogs/2016/12/dual-learning-20161207.aspx.

10

遇见智能时代的你



2017年1月6日，机器人小度第一次登上《最强大脑》电视节目挑战人类。

那一天小度自己收割了全场观众最多的惊叹，它居然能够根据一张儿童照片，就从人群里找出了已经成年的少女，还有她的双胞胎姐姐。



图10-1 机器人小度在《最强大脑》电视节目上

小度很萌，不过也让人类选手和电视机前的观众都不得不认真思考这样一个问题：当机器智慧真的超越人类，将会把我们的生活引向何方？

仔细打量周围，你会发现人工智能早已走入人类日常生活，无处不在。它只是偶尔来下棋，偶尔和我们进行脑力竞赛，更多时候它是保姆、教师、管家、助手、司机、医生……不妨来一场“盗梦空间”，让我们遇见未来世界的你：

在一个平静的午后，你在家里刚结束了一场立体投影工作会议。走出房间，看到新来的儿童伴侣机器人正在辅导儿子学习数学乘法表。妻子端着来自智能果园种植和采摘的新鲜水果进来。儿子跑上去抱住妈妈，吵着要看一部刚上映的动画电影。小家伙消息怎么这么灵通？一查看智能电视，发现已经调成“儿子”角色模式的电视果然向他推送了这部动画上映的消息。

拗不过儿子撒娇，答应让他看这部影片。但是你没有像往常一样购买家庭影院版本，因为动画片的名字《再见天空之城》勾起了你童年的回忆，那是一部讨论人与机器关系的日本电影，现在又出新系列了。你决定带一家人一起去电影院观看。

在机器人无所不在的时代，你觉得像电影院这样可以与别人接触的场所，对孩子成长也是必要的。有意思的是，虽然如今家庭影院如此方便普及，但城市中的影院数量并没有减少，生意反而好起来。可能在智能时代，人类的怀旧心理被唤醒了吧。你想去小时候常去的那家电影院，在城市的另一头。你用语音唤醒了家中无处不在的度秘系统，度秘告诉你那家电影院还在，帮你挑选了合适的场次，计算了路线，并召唤了无人车。

无人车已经知道你们此行的目的地，车载屏幕上正播放昔日经典《天空之城》[1]的片段，而不是播放新片导演访谈，这显然是优先照顾了小朋友的兴趣。你通过语音唤醒了屏幕，连接到自己的手机，准备处理几封邮件。这时视频电话打了进来，是妻子好朋友一家打过来的。原来家里的人工智能管家告诉妻子有哪些朋友也对这部电影感兴趣，于是她在出发前向几位闺蜜发出了同去看电影的广播，有一家人响应了。他们说已经到了，并且买好了票。

电影院里保留了过去样式的装潢，只是不再需要3D眼镜，裸眼3D技术已经成熟，观影效果既逼真又舒适。据说剧场里的传感器会记录每个用户观影过程中的情绪反应，再结合剧本情节进行分析，即时生成一部消费者报告，反馈给电影主创团队和发行公司。

看完电影，两家人的智能助手已经各自推荐了附近的饭店以供聚餐，最后由人类决定到底去哪家。席间把酒言欢自不待言。回家的路上，无人车人工智能语音响起来，表示想和各位聊一聊这部电影。儿子很开心地答应了。人工智能问儿子是否喜欢这部电影，最喜欢哪个角色，愿意给电影打几分。还挺懂事地问你的爸爸妈妈愿意打几分呢？你知道这实际上是接入了一套智能影评系统，谈话内容会被整理提炼，供制片方参考。你也可以看见别人同意分享的观影对话。你希望人工智能有更高级、更纯粹的对话，比如问问孩子希望利用机器人做什么。于是你问人工智能：“可以帮助我统治世界吗？”人工智能回答：“我是来服务世界，而不是统治世界的。”

不管怎么说，人工智能让世界变得更好而不是更坏。大部分人因为有了智能助理而更加关注健康、生活、学习。你的孩子将成为智能时代的新人类。

智能时代的衣食住行

2016年初，扎克伯格宣布当年要打造一款人工智能管家，到年末，他在自己的脸书上传了一段视频来展示成果。

短短几分钟的视频内容中，扎克伯格都在与这款他花了100多个小时亲自开发的智能系统语音对话。仅从视频中的表现来说，这位智能管家只是完成类似点歌、调灯光、访客人脸识别等功能。扎克伯格却

给这款系统取了一个如雷贯耳的名字——贾维斯，正是《钢铁侠》里那个存在感极强的智能系统的名字。

在盛大的2017年CES展会上，因为实用效果而荣获大奖的智能管家，却是基于百度DuerOS的“小鱼在家”。

“小鱼在家”可以支持手机、电视、音箱、汽车、机器人等多种硬件设备，同时支持第三方开发者的接入。DuerOS拥有百度语音识别与自然语言处理技术，还可以通过云端大脑不断学习进化。用户只要喊“小鱼小鱼”，就能唤醒它，然后就可以指挥它播放音乐、播报新闻、搜索图片、查找信息、设闹钟、叫外卖、闲聊天等；还可以唤醒多方视频通话、语音留言。

这些技术都不算十分新鲜，实现的却都是普通人曾经想象过的功能。

假设你有了自己的“小鱼在家”，首先会对它说什么？可能是类似“打开电视”“打开厨房灯”这样的指令。基于深度学习的人工智能可以通过训练，逐渐理解话语深层的关联性，未来智能家居可以只需一句简单的话就能理解主人心中复杂的需求。

比如当你说“我要上床躺着”，它就可以做到锁上大门、关闭除了卧室以外的电灯、打开卧室空调等；更加亲切一点的场景是：半夜小朋友起床想上厕所，迷迷糊糊的他只需要喊一声“尿尿”，卧室和厕所的夜灯就会自动开启，马桶也会启动自动冲水模式，不必吵醒劳累一天的父母。

智能家居系统不仅通过“学习”了解家庭的起居规律，还通过千万家庭的大数据和深度学习，成了育儿专家、工作助手、老人看护专家等。比如它会对婴儿的睡眠时间提出建议，会根据流感发病现状提醒预防，也可以提醒老人附近有什么可以跳广场舞的地方。人性化的智能家居，不仅让未来生活更舒适，还连接了人与世界。

DUEROS

对话式人工智能操作系统

尽享音乐　新闻播报　检索信息　链接服务　日程管理　控制家电　视频通话



搭载DuerOS的小鱼在家视频对讲机器人

图10-2 搭载DuerOs的“小鱼在家”

除了智能家居，同人们关系最为密切的就是“吃”了。当美味佳肴还生长于土地中时，人工智能对粮食的改善已从“根”上做起。

2016年美国《连线》杂志发表了一篇长文，详细介绍人工智能如何改造现代农业：依靠图像识别技术，农民将可以及时鉴别出有疾病的农作物，不会再遇到大片庄稼病死的情况；农业机器人则可以彻底改革农业劳作，一种名为LettuceBot的农业机器人外形与拖拉机并无区别，却可以每分钟扫描5000多株幼苗，自动区分出杂草并进行清除，最终能减少除草剂90%的使用量；还有与种植息息相关的天气问题，科学家已经将卫星图片纳入深度学习的训练对象，未来一个农民每天早上打开App，就能查询到自己土地上详细的气候状况，甚至精确到公里。

在荷兰已经建成了智能化无人饲喂奶牛场，整个牧场依托人工智能系统运作，不管是分配饲料还是挤奶全都交由机器完成。



图10-3 加州初创公司Blue River开发出来的机器人LettuceBot
资料来源：https://www.therobotreport.com/news/ag-in-transition-from-precision- ag-to-full-autonomy

这一切都将让未来的农产品更清洁，生产效率更高。不远的将来，超市里包装上贴着“智能农场出品”标签的商品才可能被顾客认作“放心菜”和“放心肉”。

手握美妙的食材，还需要良好的厨艺。我们一直说人工智能会替代人类的机械性劳作，但烹饪却不是简单的流水线生产。中国五湖四海完全不同的菜系文化，还有每个人舌尖上的记忆，意味着下厨是一件关乎情调的事情。

技术的主要目的从来不是“替代”，而是“支持”。对许多人来说，烹饪过程本身就是一种乐趣，这样的乐趣会得到人工智能的加持。人工智能系统不仅会帮助使用者完成一些简单操作，比如敲鸡蛋、加水、加面粉，让人类厨师可以把精力和时间集中在口味的打造上。还可以通过深度学习技术学习用户的操作习惯、口味特征，学会打造“外婆家”的菜肴。还能通过数据分享，把自家美味配方推送到更多人的烹饪系统中，那时候，烹饪系统就会像今天基于兴趣推荐资讯的新闻客户端一样，基于用户口味推荐和分享菜肴。

“下馆子”的乐趣也会被改变。用智能手机选餐厅、点菜已经不稀奇，遇到让人垂涎但是没见过的菜品，还可以拍一张照片——度秘会帮你识别菜名、口味、材料和营养成分。一些热衷于研究朋友圈里美食图片的女性用户，可能会在未来损失一点乐趣，因为菜品的“神秘感”不多了。

在娱乐领域，最理解用户心理的电影、游戏公司不会嗅不到人工智能带来的商机。游戏玩家很快也会和人工智能交上手。人们早就在游戏里和人工智能对战过，但以往所有游戏中的“人工智能”都还是典型的符号智能：根据人类设定好的符号程序运行。这样的“人工智能”很快会被灵活的玩家找到漏洞，轻松碾压。

现在的玩家在游戏里形容对手水平差时，往往会说“打你就像打电脑一样”。但这个比喻在未来恐怕会变成一种称赞——谷歌已经开始训练人工智能操作《星际争霸2》，并确认将会挑战职业电竞选手。在游戏中人工智能也将受到与人类相似的条件限制，玩起游戏来会“更像人类玩家”，甚至会让人类分不清屏幕另一边究竟是真人还是人工智能。

成熟的人工智能会向沉迷于游戏的你发出提醒：过度缺乏运动，健康指标又下降了。智能时代的新生活不该仅仅是盯着屏幕，人工智能也不应该将用户完全留在家中。在未来出行中，人们只需要带上一

部手机，就可以有人工智能相伴。人工智能将会让喜爱旅行的人更加尽兴，让原本对旅行兴趣寥寥的人体会到新的乐趣。

让度秘来完成订票、订房，为你规划好路线，这些只是形而下的事情。在精神层面上，人工智能也会改写“旅行的意义”。

机器甚至是可以解风月的。百度深度学习实验室正在开发的AR（虚拟现实）技术，可以让手机向你诉说这朵花儿的花语，那只鸟儿的名字；挖掘这方山川的传说，诠释那方古人的隽语。面对转瞬即逝的美景，它可以帮你捕捉最清晰的日落时分，稍纵即逝的惊涛骇浪。结合地图工具，还可以在你到达的每个地点，推送他人的攻略和心得体会。

无所不在的人工智能翻译官让你拥有与外国人对话的能力。当彼此用陌生的语言交谈时，人工智能仿佛就在一旁耳语。

不仅连通陌生的语言，还能连通光明与黑暗。植入手机或者可穿戴设备的助残机器人，可以利用视觉和语音技术帮助盲人识别纸币的面值和真假，提示周围的环境。我们已经可以构想未来的机器导盲犬。

甚至我们可以畅想“下一代地图计划”，用无人机配合传统方式采集更立体的地图信息，利用3D重建技术全方位还原真实世界。AR导航将让你如同拥有“透视眼”，下一站无论是商店、机场、医院，都变得一览无余。图像识别技术使地图信息每时每刻都在更新，在你迈出每一步前都已预知前方道路和场景。虚拟现实使得人类不仅可以穿透语言的障碍和光线的阻挡，还可以像《黑客帝国》一样穿行于梦境和现实。

当下的人工智能服务还远称不上完美，但大目标无可争议，只是到达的道路个个不同。要享有未来的生活主动权，就要有这些方面的准备。有人担心未来社会人工智能的福利只属于少数具备知识的人，但我们希望更多的人能分享智能社会的红利和便利。

别输在人工智能起跑线上

2016年4月，中国作家郝景芳凭借科幻小说《北京折叠》获得雨果奖，这也是继刘慈欣的《三体》后，中国人再次摘得这一科幻界的“奥斯卡”。

在面对“未来已来”的人工智能时，郝景芳有自己的观点：“未来的人类社会分工会更加明确，标准化的生产由人工智能完成，人类则负责那些创造性和情感性的工作，同时未来社会的组织也会更加灵活。其中最关键的一点是在人工智能主导的社会中，现在的教育模式已经无法适应变化，如果不进行改革，我们很难面对未来的人工智能时代。”

这番发言是郝景芳在“2016年度科技风云榜”颁奖盛典上说出的。如果年轻的父母在阅读郝景芳的发言时，只是期望自己的儿女能拥有像她一样的人生轨迹，而忽略了她关于教育所做出的忠告，也许就会错失调整自家教育方式的机会。

为了下一代付出多少都愿意的中国家长，对子女教育问题常感到焦灼。在很多父母眼中，好学校就等同于好教育。但是满眼盯着师资、升学率的家长，是否意识到“好学校”的标准在智能时代将会发生改变？

人工智能正在接入学校的“教”“学”“管”等各个环节。

互联网与AR技术可以让师生跨时空互动。教学场景超越了传统课堂。在虚拟空间和在线教育里，学生更平等，时间调度更灵活。学生可以在课前更好地预习，更方便地分组学习，老师也可以更灵活地调整进度和因材施教。

智能教育系统可以自动记录每个学生容易做错的题目或者进度缓慢的环节，再据此向学生匹配专项辅导资源。学生学过的教程、做过

的作业、阅读过的材料不仅存放在他的个人资料空间里，而且变成丰富的标签，描述出学生的学习曲线和风格。机器秘书可以根据标签向学生和老师精准推送教学建议和资源，改变以往的填鸭式教学方式。所有学生的学习记录和反馈被人工智能综合起来，互相参照、优化、聚合与分发，在个性化的同时彼此激发，提升总体水平，彻底升级“教学相长”的含义。

老师是学校里最辛苦的人。智能系统将会极大减轻他们的负担，让他们把精力从机械的工作中解放出来，用于个性化的创新教学。比如，基于自然语言处理技术的人工智能系统在经过海量数据的训练之后，可以秒批作业，不只是英文作文，甚至能批改汉语作文。基于语音识别技术的人工智能系统则可以带领学生朗读英语并矫正口语。除了作业，试卷批改也已经投入应用。

人工智能系统批改作业具有深层次教育影响。对学生来说，教育研究早就告诉我们，人的记忆具有规律，在学习后的最初一段时间遗忘最快，之后趋于平稳。人工智能系统可以帮助教师更快给学生反馈，以免学生因为等待而懈怠。

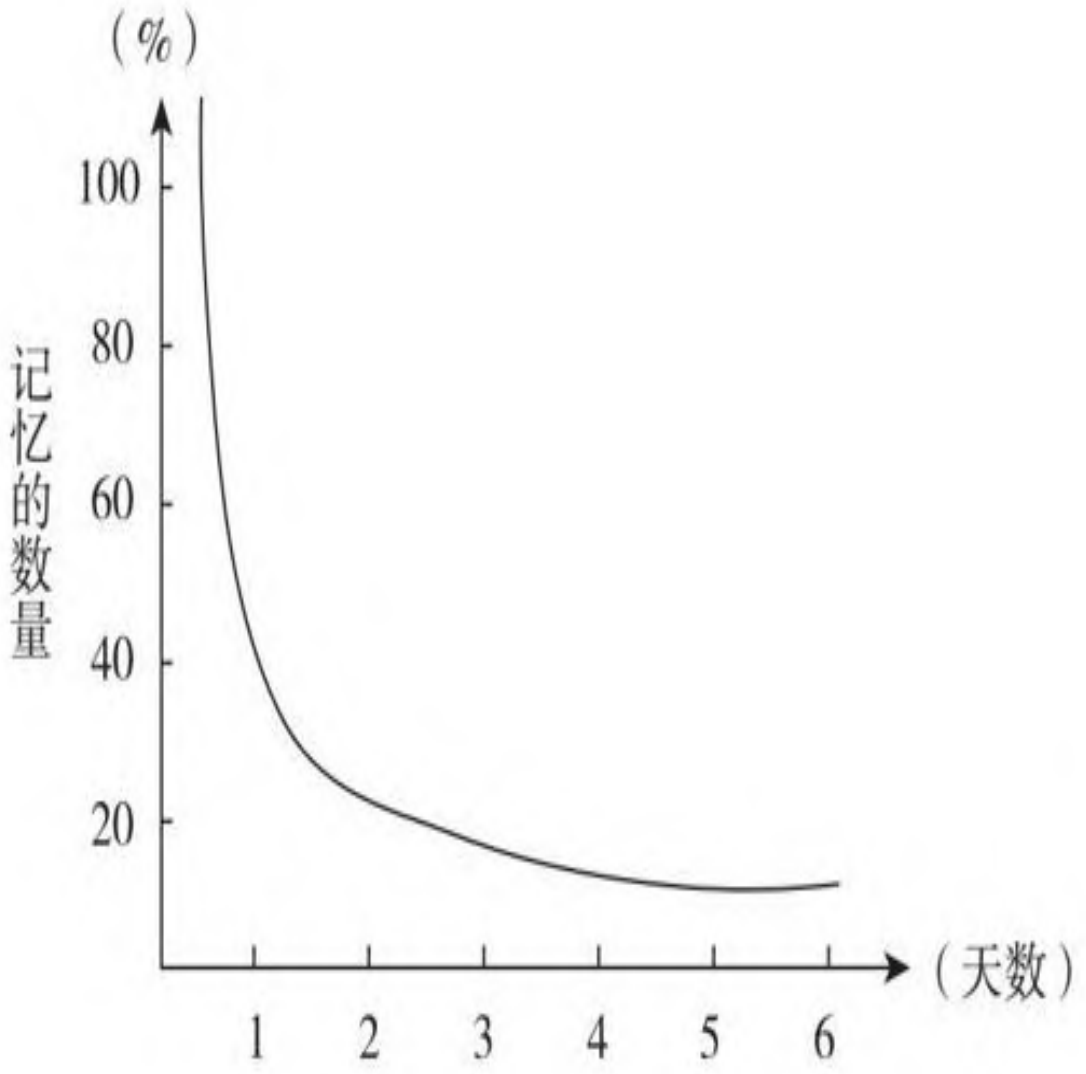


图10-4 艾宾浩斯记忆曲线

对于教师来说，除了减轻负担，还可以通过作业数据统揽全局，把握方向，真正成为一位教育指挥家和艺术家。

机器人已经在尝试参加高考。这并不是说真要送一个机器人去上大学，而是训练出能够完美应对考试的机器人，再反过来培育考生。如今，高考填报志愿也可以让度秘来帮忙了。很多普通家庭因为缺乏对高校的了解，往往不知道选择什么学校。度秘不仅可以查询分数，还可以根据往年报考的大数据状况、学生评价数据等提供报考建议。

人工智能应用在全世界不平衡地发展着，大胆采用人工智能辅助教学的学校还是少数，但步伐很快。据统计，全国重点中学里已经有近半学校采用了智能教育系统，活跃用户过半，教学效率提升20%以上。未来所有学校可能都需要一位人工智能“老师”，正如所有企业都需要一位首席人工智能官。落后者将无力进入智能时代的前列。所以，为择校绞尽脑汁的家长，不妨也放眼未来，从人工智能的指标上去做出选择。

现在市面上儿童教育类机器人产品已经非常多。有的可以为儿童讲故事，有的可以与儿童交流英语，有的还可以选择角度给儿童拍照发送给家人，有的甚至可以带领儿童阅读纸质图书。它们外形、大小各异，接入的智能流技术也各不相同。

选学校、挑玩具的背后，是父母的思维方式在新时代需要主动改变。

想象一下我们与孩子之间的代沟会发生怎样的变化？数字鸿沟会不会也变成代沟的一部分？就如同成长于“互联网思维”之下的80后、90后与自己父母之间的代沟一样。

孩子们对人工智能的接受比我们想象得更快。越来越多的小学开设了编程课程，越来越多的小朋友很早就开始接触计算机、智能手机。游戏、网络中也充斥着各种智能技术的身影。

美国电视台做过一个针对10多岁孩子的采访，发现很多美国儿童不清楚Windows这种操作系统怎么使用，但操作智能手机却十分熟练。微软如今这么重视人工智能，将发展人工智能技术作为全公司的重要战略，投入巨大资源研发微软小冰，新一代用户思维转变所带来的危机感就是其中一个重要因素。

百度也被新生代的变化所触动。度秘的负责人景鲲则如此描述将在人工智能环境下成长起来的新一代：“百度做人工智能，做度秘，我们自己发现并不是在教用户怎么用，反而是顺着用户来引导我们做

产品。而孩子在表达自己需求这一点上又是最无所顾忌、最没有束缚的。孩子的要求天马行空，但是他们在心智和身体上还有很大不足，大多数时候需要大人来帮他们完成目标，所以他们的需求一直是受限制的。而人工智能则只需要一句话、一个简单操作就能理解他们的需求，这可以降低他们达成目标的门槛，释放他们的需要。所以当他们发现了永远不会厌烦提问的语音对话系统时，自然而然地学会了用这种新的方式来获取信息。”

人工智能如同普罗米修斯带来的礼物，既可能是启蒙的火种，也可能是伤人的烈焰，只有掌握它，才能发挥积极的作用。

我们要建设的是新生活的高速列车，要努力搭载更多的人，而不是修建电影《2012》里的方舟，只允许少数人享有。当人工智能在技术上超维进化，我们希望数字鸿沟也能够被缩小。人工智能必须是开放的，根植于网络空间，打破空间与时间的隔阂，让每一个人只要努力，就可以或多或少地接入智能流。

智能社会还有更多机遇与挑战在前方等待着我们的下一代。当他们走出学校，寻找工作岗位，面临的又会是怎样的场景？当机器人开始大规模接手人类的简单机械工作，新人类又如何驾驭新的未来？

工作着是美丽的，智能时代更是如此

走出象牙塔求职是又一个新的人生门槛。毕业生最害怕的不是低工资，不是辛苦，甚至不是找不到工作，而是发现自己所学的东西一无所用。人工智能带来了新工具，就如同工业革命把农民手中的工具从锄头变成纺织机，智能时代同样呼唤着新技能和新思维。面对未来，许多早已步入职场的资深人士都难免会不安：我能适应智能时代的工作吗？

谋生的焦虑缠绕着每个时代青年的内心。中国互联网上曾出现过一款火爆的游戏《北京浮生记》。游戏要求玩家扮演一位背井离乡来到北京谋生的北漂，目标是要在40天内赚到足够偿还自己债务的金额。这个仅有700多KB、界面简陋的小游戏，能在当时大受欢迎，和它充满着“北漂”元素是分不开的。游戏制作者本人就是个彻底的北漂，在1992年揣着200块钱只身来到北京求学闯荡。游戏里面充斥着各种幽默但又真实的北京特色，在那个没有雾霾只有风沙的年代，引发无数只身在大城市拼搏的青年的共鸣。

近些日子《北京浮生记》的制作者又在网上露面了。面对诸多玩过这款游戏并且关心他动向的老玩家，他说当年在开发《北京浮生记》时，正在北京邮电大学从事人工智能科研工作，后来一度离开科研领域专心开发游戏。现在回到本行，做深度强化学习（Deep Reinforcement Learning）的应用，服务游戏业这个垂直领域。他在若干年前的一次采访中曾将这款小游戏称为“文学作品”，他说：“我认识20世纪二三十年代的上海是通过《子夜》，认识拉丁美洲是通过《百年孤独》，10年后，如果有人能够通过《北京浮生记》生动形象地认识2001年的北京的某些生活场景、某些层面，我将非常高兴。”

成就魔幻的，是并不魔幻的艰苦努力和机遇。2001年正是中国互联网的寒冬期，各种泡沫破裂、企业倒闭，坚持下来的大部分人都在今天有了收获。如今在移动互联网、大数据、人工智能各个新兴领域走在第一线的领军者，或许也曾在2001年某个周末艰苦加班之后，疲惫不堪地回到五环工地旁的出租屋里，打开台式电脑，在睡前玩上一把《北京浮生记》，会心一笑。

“互联网+”让任何产业都已离不开互联网，智能时代，任何工作也都会有人工智能的影子。对大数据、智能化的嗅觉，决定了未来工作者的层次高低。

很多机械劳动岗位将被人工智能替代。流水线工人、听写速记员、出租车司机等，未来会迅速消失，就如同纸媒在互联网时代的衰落一样，即使它拥有数百年的悠久历史。人类应该去从事更具有创造力的工作，但首先你需要掌握创造工具。否则，即便是看似高端的白领工作也会形同机器。

当《福布斯》这样的著名媒体开始用机器人以30秒一篇的效率写新闻通稿，曾经以无冕之王自居的记者开始如坐针毡。不过不要太紧张，除了需要努力提升写作水平外，还可以配备智能工具，比如语音识别速记软件、智能剪辑视频工具等，这会让记者和编辑更强大。

对数据分析软件的利用，会让记者发现更多新闻线索。比如“百度指数”等产品，根植于搜索引擎大数据，可以为用户提供不同新闻关键词的搜索热度、趋势变化、搜索人群的兴趣图谱等各种信息维度。媒体从业者可以据此判断新闻话题走向，明星粉丝可以从中了解偶像的热度变化，商家可以观察产品的用户定位和需求趋势。

相关性思维在这里获得充分的体现。举个例子，2014年6月16日到6月20日是百度指数“请假”一词最高的一周；同时还是“啤酒”一词百度指数最高的一周。敏锐的读者从两者的关联已经可以想到答案了：6月13日巴西世界杯开幕，接下来是世界杯比赛最密集的一周。新闻记者、品牌、商家可以利用这种数据洞见做些什么？善于使用这些工具的人，会成为同行中的佼佼者。在使用数据工具的过程中，我们也在被训练，使自己更适合在智能时代发现和创造价值。

翻译也将面临挑战。往日在重要领导人合照中总会出现的那位翻译官，恐怕需要叠加一个人工智能的影子。人工智能语言处理工具的开发，除了需要满足普通人的翻译需求外，还应该针对高端客户开发新功能。高端翻译需要的可能不是基本的机器翻译，而是由深度学习技术来帮助自己掌握服务对象比如领导人的用语特点；提供特定功能服务，比如如何把中国古诗词翻译成英文。

当百度大脑可以构思诗句，当神经网络可以模仿巴赫训练出作曲能力，当日本的小说奖项评选正式把人工智能创作的文学作品纳入候选，所有人都开始意识到作为人类心灵圣地的文化艺术领域，对于人工智能来说也不神秘了。神经网络很像人类的大脑，有无数隐秘的神经通路。李白要在醉酒时才能诗兴大发，那是因为醉酒状态导致一些平时蛰伏的神经通路被激活。未来人类可以不用借助醉酒，而是借助神经网络尝试开发灵感。

法律、金融等专业人士的日子也将被改变。由IBM研发的世界首位人工智能律师Ross（罗斯）在2016年就职于纽约Baker & Hostetler律师事务所，负责帮助处理公司破产等事务。这是一个真正可以全身心为雇主考虑的法律全才，不会像人类律师那样昂贵，普通律师或者法官为何不试试利用这样的产品呢？2016年12月，北京市高级人民法院上线了一款名为“睿法官”的人工智能系统，它不仅仅停留在收集资料、为案情“画像”的辅助阶段，而是已经可以做到对案情进行分析，采集案件的多元信息，识别出影响案件定罪量刑的相关要素及当事人上诉的理由，帮助当庭法官做出初步判断，已经十分接近“判案”形态了。使用人工智能助理的律所和法庭能够比同行更高效地处理案件，从而创造更高的收益和社会效益。

金融的城堡由数字的砖块堆砌而成，随处可见人工智能可以渗透的缝隙。

著名的布隆伯格就是用数据武装的华尔街之狼。这位曾当过纽约市长、身家几十亿美元的企业家，几十年前就通过金融数据处理工具杀入了华尔街。在他推出自己的债券交易产品之前，交易员基本上依靠查询路透消息源来查询价格，然后用铅笔和纸进行计算。而布隆伯格提供的产品，给用户提供市场的实时数据、计算能力、财务分析等，迅速席卷了所有债券交易所。布隆伯格第一次卖出他的产品给美林时是1982年，而微软在同年才刚刚推出Excel的第一版，当时还叫作Multiplan。

布隆伯格的例子也许有些远。但是在中关村、五道口、西二旗、张江、华强北，无数雄心勃勃的青年人在试图利用数据智能工具完成人生逆袭。人工智能不分贵贱，只在乎能力。人事管理员可以利用数据工具研究公司人员分布状况，提出优化意见。咖啡店可以通过智能对话App为用户订咖啡，并迅速了解他的口味。昔日的游戏高手，可以转型为游戏直播解说员，他比别人更能发现人工智能哪里还不太像人类。

人工智能的潮流虽然会掩埋一些职业，但也会催生新的机会。汽车消灭了马车车夫，却创造出了司机这个职位。人工智能的普及也会创造新的职业机会。吴恩达举了一个例子：以后无人机会越来越多，而无人机交通管理工作现在还没有，以后准会出现，谁来做这个工作？他建议人们把更多精力投入在教育和培养人才上面。

真正决定我们职业未来的，是我们是否具备智能时代需要的职业素养。当智能工具变得越发无所不能，能否灵活使用它就成为差距所在。一个对程序编写、数据分析、机器学习一窍不通的人，就像工业时代不会使用扳手的人一样。

现在几乎人人都拥有一部智能手机，但多数人在用它打游戏、看电影、刷朋友圈时，没有意识到其中蕴藏着通往未来的入口。人工智能领域很多世界最尖端的技术，都已经集成在诸如度秘一类的产品中。翻译、语音识别这些看似波澜不惊的功能，已经可以让你接触到最前沿的人工智能技术。

“工作着是美丽的”，这句话是已故女性作家陈学昭的长篇小说的名字。她是五四新文化运动的旗手之一。今天人工智能将会带来又一场新文化运动。而在智能时代，工作着依然是美丽的，它需要智慧，更需要人类在与机器的互相砥砺中前行。

生命之歌，人工智能如何再造医疗

思索、奋斗、享乐、奔走……人类以肉身在世间拼搏，无论现世生活是幸福还是煎熬，总是渴望更健康长寿，渴望能亲眼看到未来的任何一天。昔日秦始皇为求长命百岁的仙药派人东渡，而今天的医疗科技则把长寿带给普通民众。

科技的发展又带来新的想象。无论是《超能陆战队》中亲切至极的医疗机器人“大白”，还是《攻壳机动队》中用人类肉体改造成的武器，都成为荧屏上的经典角色，令人津津乐道。虽然人类文明在工业革命后获得了突飞猛进的发展，但在现代医学的庇护之下，人类的肉身仍然显得那么脆弱。小小的身躯之内蕴藏着一个我们远未探索的世界。高科技会把人类的身体引向何处？这已经不仅仅是医生感兴趣的话题。

科学史上每一个领域的进步似乎都与医疗行业发生着重要关联。物理学家在光学领域和原子领域的研究带来了显微镜和X射线，宇宙飞船中为宇航员准备的健康检测室日后变成了重症监护室，生物学让动物实验成为人类最重要的医学测试方法，而化学更是一门与药品有着千丝万缕关系的学科。

那么人工智能在医学领域会扮演什么样的角色？答案恐怕不是一种角色，而是会改写整个医疗行业。

人工智能作为一种方法，对医疗的影响将是全方位的，由浅入深，由小及大。

尖端医学最为吸引眼球：Nature报道神经假体使腿部瘫痪的猴子能够重新行走；美国研究出“纳米鱼机器人”可以轻松将止痛药物传输到指定的身体部位；在基因药物领域，深度学习为改写基因、合成新生物提供了可能。

类似的例子不是我们想要讨论的重点。纳米机器人、基因改造虽然是影视作品热衷的话题，却和现实领域颇有距离。更重要的是，虽然人工智能已经进入许多最前沿的医学领域，但几乎都还停留在理论或者实验阶段，不过在一些更成熟的领域，人工智能已经可以发挥重要作用。

在眼部疾病治疗中，由于对病人信息掌握有限，有1/10的病人会在治疗过程中受到伤害，其中至少一半伤害是可以预防的。英国知名的摩尔眼科医院与谷歌展开合作，打造了一款机器学习系统。这款系统仅凭眼部数字扫描结果，即可识别潜在的眼部疾病风险。眼部扫描技术早已有之，但是传统的机器在完成扫描之后，无法快速分析复杂的眼部数据。机器学习则是对付这类数据的能手，可以大大缩短分析时间和提升准确率，更快、更准确地为病人提供治疗方案。

IBM的watson机器人可以通过大数据和人工智能帮助人类分析肿瘤等疑难疾病。在日本，病理读片大多都需要两名医生共同完成以防止错漏，而NEC很多年前就开始在日本医院中推行自动读片系统，现在已经能够替代医生的角色来辅助读片工作。

生物大数据也正在改变着医学研究和制药业。在一些传统医学诊断中，医生会让病人在一段平缓的路上行走几分钟，并记录他们的行走距离。这个测试是为了预测肺移植者的存活率，并可检测肌肉萎缩的病程发展，甚至可以评估心血管患者的健康状况。但是如果你想同时拿到1000个人的监测数据，那么光是记录就会变得非常困难。

智能手机解决了数据获取的问题。美国一位研究者想要收集有关心血管的数据，于是上传了一个名叫“我的心脏计数器”的App到苹果商城，结果短短两周内他手上就有了超过6000人的测试结果。现在他面对的问题不再是样本不足，而是缺少面对庞大数据进行更精确分析的能力。

我们一旦有能力把生物大数据抽丝剥茧，获得想要的结论，整个医学行业将受益无穷。以糖尿病为例，仅在中国就预计有超过1亿人受

此病困扰。糖尿病诊断过程中有一个令医学界十分头疼的问题，就是不容易对病人做到糖尿病亚型的精确细分，这对前期预防和后期治疗都极为不利。而大数据分析可以在这一点上发挥优势，如果最终能对每个病人做出具体的病型确认，就能真正做到“对症下药”。

基因测序是发现致病原因的一个前沿方法。在这方面，我赞助了一个食管癌项目。过去在基因测序与疾病相关性的研究上，限于技术基本只能够做到单基因致病研究，比如发现某种基因突变会导致唐氏综合征之类的罕见病。而常见病很多是由多个基因组合的突变导致。过去由于计算能力达不到现在的水平，人们根本算不出来到底是哪些基因的组合导致了一些常见病。但是未来是可以算出来的。

大数据与智能分析对药物市场也会有明显改善。现代医疗中的药物品种已经变得越来越多、越来越复杂，有时病人会发现，光是治疗感冒的药物就多达几百种。大数据的出现，将为我们带来个人定制的药物使用方案，精准使用每一种最适合我们的药物。

这也就是“精准医疗”（Precision Medicine）的概念。美国前总统奥巴马2015年在国情咨文中提出“精准医学计划”，打算通过分析100多万名美国志愿者的基因信息，以便掌握疾病形成机理，开发相应药物，实现“精准施药”。通过互联网、人工智能与生物大数据的引领，精准医疗将成为一种把个人基因、环境与生活习惯差异考虑在内的疾病预防与处置的新方法。我国也在2015年由科技部召开精准医学战略专家会议，计划在2030年前向精准医疗领域投入600亿元。

多伦多大学的Brendan Frey（布兰登·弗雷）教授称，现代科技下我们已经能源源不断地获取基因生物学的数据，但是人类却难以破解和掌控这些海量数据，无法“理解”基因。我们可以想到，深度学习将扬其所长，寻找人类无法发现的基因关联。当人工智能找出健康的基因序列模式，人类就能够通过基因来诊断甚至预测疾病，优化用药靶向。

精准医疗的概念引起了不少争议。美国女星安吉丽娜·朱莉通过基因检测获知自己有80%的概率患上乳腺癌和50%的概率患上卵巢癌，于是毅然切除了双侧乳腺和卵巢。这一举动引发了媒体的大讨论。仅因为一个数字概率就去做切除手术，会使很多人迟疑。更有人担忧预测技术会被滥用，成为引导病人“无病呻吟”以便谋取利益的一种手段。

预测性治疗正如人工智能在警事安全领域促成的预测性防范——在一个人还没有实施犯罪之前就采取预防措施。无论如何，未来的人类要更频繁地面对此类抉择。

相比进行基因检测进而提前手术这类较为奢侈的项目，人工智能在更多普适医疗项目上也具有重大利用价值。

在中国大城市中，超过70%的人处于身体亚健康状态，而由于看病难、工作忙碌等因素，真正去医院的人数不到5%。人们需要一个身边的“私人医生”。未来的医疗大脑会成为这个“私人医生”。所有人随时随地可以通过手机、智能系统“看医生”。远离大城市和三甲医院的人群也将拥有更多的医疗选择。

未来人类的医疗业态将会被人工智能改变，这是毋庸置疑的。但是在百度大数据实验室主任范伟身上发生过的一件事却让这个过程具有了一种魔幻感——人工智能已经为他实现了一次过去与现在的“穿越”。

30年前，当他还是一个高中生时，希望成为一名医生。他不仅在心里做了规划，还亲自跑到协和医学院招生办拜访。不过因为种种原因，范伟没有走进协和医学院的大门，而是来到清华学习计算机编程。

有趣的是，三十多年后，范伟真地做起了医疗。范伟现在的梦想，就是让更多的人能够早日用上医疗大脑。让医疗大脑为医生做辅助，包括为医生自动生成病历草稿，为医生节省时间，救治更多的病

人。有了医疗大脑的辅助，医生这种职业一定会焕发新的光彩，医患关系也会更加和谐。范伟希望利用人工智能技术把有多年行医经验老医生的知识与阅历，变成一个软件，造福更多的人。

---

[1] 日本著名动画导演宫崎骏的作品，于1986年上映，讲述的是人与机器的关系。

11

美丽新世界　严肃新问题



1932年，进化论旗帜人物托马斯·亨利·赫胥黎[1]的孙子阿道司·赫胥黎出版了反乌托邦名著《美丽新世界》，描写了未来机械文明世界的景象。在那里，人们安居乐业、衣食无忧，但一切都被标准化。人类在孵化中心诞生，分别从属五大贵贱不同的种姓。管理者用试管培植、条件反射、催眠、梦境疗法等“科学”手段，严格控制人的性格，让他们用快乐的心情去执行自己被设定的社会角色和消费模式。这样的主题后来在《黑客帝国》等作品中不断复现，成为现代思想者不可回避的话题。

《娱乐至死》的作者尼尔·波兹曼说：“在《1984》中，人们被对痛苦的恐惧所控制；在《美丽新世界》中，人们则被对快乐的盲目追逐所控制。”而未来的人工智能社会将是哪一种？是会超越所有人的想象，形成全新的世界，非痛苦或者快乐所能描述，还是继续延续人类永恒求索奋斗的主题？

智能革命，是堪比工业革命的第四次技术革命。与工业革命时代中国被动落后不同，今天，中国是以领先态势，主动驾驭着这场革命。正如无人车之路在挑战中开拓，智能时代也在混沌中清晰起来。我们有理由相信中国会在这场变革中占据制高点。

如果我们把视线抬得更高，面对技术与人类的关系，又不能不有一些忧虑。细心的读者应该能感受到，在本书各章节的热情之中都隐藏着一种关怀。我们提倡起跑线上的人工智能教育，因为担心起点上的不平等；我们描述未来的新人类，因为希望每个人都掌握智能时代的生存技能；我们仔细构想首席人工智能官能做什么，因为希望众多企业能在大变革时代协同进步，保持经济生态平衡；我们知道人工智能与物联网将助力中国制造强势升级，但也不能不担忧新的产业危机。

纵观历史，技术进步起先总是被少数人拥有，甚至成为制造不公正的武器，其普及过程充满曲折与不确定性。这其中固然有拥有者的

自私因素，但也有后进者的迟钝以及人类天生的族群竞争使然。人工智能也会如此吗？

中国移动互联网的发展是个有趣的回答。农村用户使用移动互联网的频率远超PC互联网，因为前者容易得到，只要拥有一部智能手机即可。中国强大的制造业与互联网基础设施一起普及了移动互联网。在这个领域，人们更平等一些。人工智能与大数据作为一种流动的资源，与移动互联网一样，也天生具有穿透性和普适性，只要通过一部联网手机就能触及。

每个人都将遭遇人工智能，区别在于个人的准备程度。人工智能作为新的工具，究竟会创造一个怎样的新世界？一切有待努力。中国人曾经于无声处兴起了一场浩荡的“勤劳革命”，惠及亿万国民甚至是世界。面对智能革命，勤劳的人类能否再一次做好积极准备？

数字鸿沟

吴军的一个说法在网上流传：智能时代只属于2%的人，其他98%的人都会成为落伍者。看到这个数字，人们很容易想起2010年开始的占领华尔街运动，运动者打出的口号是反对1%。未来有可能是掌握金融资本的人占据了1%，掌握人工智能资源的人占据另一个1%。本书无意论述那场占领运动，这里只想说一件那场运动之后不久出现的惊人一幕。

2012年10月末，飓风“桑迪”席卷美国东海岸，海水倒灌，600多万人无电可用。繁华的纽约市一片汪洋，下曼哈顿区25万用户遭遇断电。但是却有一幢大楼灯火通明，孑然独立，那就是位于纽约市曼哈顿西街200号的高盛大楼。



图11-1 纽约市断电夜景图
资料来源：http://thefeministwire.com/2012/12/black-survival-in-the-uchromatic- dark

这幕景象引发了不少民众的愤怒。彼时占领华尔街运动刚刚被强行结束一年，美国仍然深陷次贷危机引发的经济不景气中，就业低迷，人们对华尔街充满怨气。而高盛则是华尔街金融资本的首席代表，在全世界兴风作浪。黑云压城城欲摧，兀自闪亮的高盛大楼与周

围一片漆黑形成鲜明对比。这一幕看上去是个多么好的象征：百分之一就在那里，哪怕周围洪水滔天。

但高盛发言人立即解释，“我们不是唯一有电的大楼，但我们的确有发电机，现在完全是自己给自己供电”。他想说的是，高盛大楼是凭借自己的实力在闪亮。

这大约也是现代社会的隐喻：一部分人因为各种原因站在金字塔顶尖，有人依靠权势、有人依靠资本，也有人依靠才能或技术。高盛本身就是非常善于利用数据和机器智能的金融公司，代表着华尔街与美国东西海岸高科技产业力量的结合体。相比美国已经陈旧的基础设施，高盛却独善其身，掌握着大量资源。

奥巴马总统在任时，美国高新产业高速发展。然而到了总统换届时，代表高科技势力的民主党却遭遇了滑铁卢，尽失总统宝座与两院控制。有资深评论家指出，在这场大选中，围绕特朗普涌现出了所谓的“白人抗议”现象。对民主党发出抗议的白人都是那些高校和高新产业之外的白人工作者。早有专家指出，美国金融资本不受限制地扩张，导致虚拟经济与实体经济比例严重失衡，制造业被挤压出局，金融资本与信息高科技技术结合，愈发加剧了这个失衡，终于带来恶果。

诺贝尔经济学奖获得者保罗·克鲁格曼痛心疾首地说过一件事情：

2010年，当哈得逊河道底下亟须一条新的铁路隧道时，新泽西州州长却突然取消了这个当时美国最大的基础设施项目。但与此同时，另一个耗资巨大的隧道项目却接近尾声：Spread Networks隧道贯通宾夕法尼亚州的阿勒格尼山脉。不过建设Spread Networks隧道的初衷不是为了载客，甚至也不是为了运货，而是为了搭起一条光纤，可以把芝加哥期货市场和纽约证券交易所之间的通信时间缩减3毫秒。铁路隧道项目被取消了，Spread Networks隧道却修建了起来。谁在乎这3毫秒呢？答案是股市里的高频交易员，他们靠比同行快几毫秒的速度买

入或卖出股票来赚取差价。这个现象可以告诉你，如今的美国究竟出了什么问题。社会正在把过多的资源分配给金融投机交易，导致产业严重失衡。

高频交易或者闪频交易，正是一种以信息技术为基础的金融投机方法，需要硬件、算法、人才全方位的支持才能做到。在算法方面，它不仅要求把资深金融交易专家的策略转化为算法，还要求这个算法尽量高效，能够比别人的算法更快地得出结果，而且算法能够自适应环境，根据外界条件变化调整参数和概率，这已经是一种比较发达的机器智能了。茕茕孑立的高盛就是这方面的高手，为此招募了顶尖的计算机团队。类似高频交易的技术，不仅帮助金融大鳄远远甩开一般散户，也屡次因为机器出错，导致股市闪崩，极大地加重了市场的不确定性。

美国的教训值得全世界深思。数字技术在美国发展久远，人工智能技术是这一技术的最新延伸。数字与智能技术本身不是导致失衡的罪魁祸首，与不公正的社会制度和不科学的经济政策结合却会加速失衡。而普通民众则是失衡后果的主要承担者。

美国2016年大选结果体现的东西部州与中部州及东北铁锈带的鲜明对立，描绘的不只是政治对立，某种程度上，描绘的正是数字鸿沟。

当然，鸿沟不全是由数字技术造成的，只是被数字技术深化了。2014年法国经济学家、巴黎经济学院教授托马斯·皮凯蒂出版了《二十一世纪资本论》一书，风靡一时，受到克鲁格曼等经济学家的大力推荐。该书以令人信服的数据证明，几十年来，世界上的收入不平等在扩大，资本回报大大超过了普通人的劳动回报。马克思早已描述过的不平等生产关系一直延续下来，似乎没有什么改善的迹象。联合国的《2016年中国人类发展报告》显示，数十年来全世界主要扶贫成绩都发生在中国，其他国家乏善可陈。

马克思认为资本主义工业化是以维持高利润为目的，而不是直接以提升人类福祉为目的，于是制造了大量的剩余人口或者说是失业大军。他们随时等待被雇用，以替换失去剥削价值的工人。但在智能时代，剩余人口的含义也许会发生变化。

人类还能做些什么

很多人尚未意识到所有单一技能的职业都可能被机器代替。语音识别代替速记员的事情最近几年就会发生；现场同声传译只是因为受限于麦克风阵列硬件发展，还需要时日；电子警察比保安更有效，很多小区里摄像头车牌识别已经代替了以往的刷卡进出；电商、快递业的人工客服已经被机器取代。倒是出租车司机可能会晚一点被淘汰，无人车并不容易在10年内取代人工。

亚马逊的“无收银员超市”让血拼族一片欢呼。在这里，顾客只要安装了特定App，就无须排队，无须柜台刷卡，拿好商品就可以走人。传感器会自动识别顾客带走的商品，计算价格，并在顾客离开超市时自动从用户银行卡上扣款。有人戏称那种购物体验就像自由抢劫一样，让消费者大呼过瘾。



图11-2 亚马逊智能超市
门上写着：Just walk out（走出去就行）资料来源：https://www.amazon.com/b?ie=UTF8&node=16008589011

这当然是消费体验升级。但悖论在于，智能超市在造福购物者的同时，也让收银员失去了工作，进而减少了购物能力。仿佛量子态坍塌现象——量子以波的形式存在时才能被发现，但是当观察者观察时，量子必定存在于一个确定的位置，于是原来的量子波坍缩为一

点，此时量子表现出粒子性，人也就无法继续观察——技术与幸福的关系也是如此，当你看见了幸福的同时，背后就有些事物在发生坍塌。

除了收银员，欢欢喜喜购物的律师、新闻编辑等白领可能也幸运不到哪里去。他们的岗位也可能会被代替，至少基本的法律分析和通稿写作在一些公司已经由机器执行了。

就在2016年底，俄罗斯最大的银行Sberbank（俄罗斯联邦储蓄银行）宣布推出机器人律师，用于处理各种投诉信件。这将导致大约3000名在银行工作的法律专业人士被炒鱿鱼。Sberbank执行委员会副董事长Vadim Kulik（瓦迪姆·库利克）说，未来所有的常规法律文件处理都将自动化，只需要律师处理那些紧急的法律程序。

在未来，可能只有那些需要直觉、需要创造性的非标准化工作，如艺术设计、规划组织工作才不可替代。

失去购物能力倒也不是很要紧，因为购物本身可能也不需要了。马克思认为剥削导致工人阶级的贫困，进而引起消费低下、需求不足，最终导致生产过剩并崩盘。而智能时代，高效的机器生产可以满足更多人口的基本需求。工厂主可能连“剥削”人的意愿都没有了，因为机器人可以比普通工人工作得更好。失去购买力的人口也可以从机器生产的丰盛中分得基本福利。

新的人类工作机会确实会被创造出来，比如无人车交通管理岗位，又如亚马逊已经拥有5万名为数据打标签的工作者。

那时，衣食无忧的普通人类可以实现马克思“上午狩猎，下午钓鱼，晚上研究哲学”的自由生活了吗？也许可以，但可能会缺少自由的心情，因为你不能创造价值，不被需要。人的一大基本需求就是渴望得到他人的认可。没有价值的人会进入怎样的消沉状态？或许他们会重新组成一个圈子，如同地下城一般，有一套不同于智能世界的生活方式。

地下城的想象有些遥远。短期内，剩余人口的增加将是一个现实问题，可能引发社会不安定。我们可以想象，未来会有新的劳动保护法条款出台，比如规定每个企业雇用的机器人与人工比例不能少于一定标准等。但当务之急可能是升级社会职业培训，为普通工作者提供计算机、互联网和人工智能方面的职业教育。

美国政府也已经意识到挑战，2016年12月白宫发布《人工智能、自动化和经济》，为应对就业市场的变化提出了战略性建议，比如：

要为教育培训投入巨资，教育和培训美国人，为未来的工作做好准备。

为工人提供更好的指导，以引导他们进行工作上的转换，同时赋予工人权利，确保广泛共享由人工智能带来的经济增长收益。

建议很美好，但考虑到美国政府即便在医改计划执行上也举步维艰，能否有效应对智能革命，还有待观望。康奈尔大学计算机科学系教授、图灵奖得主John Hopcroft（约翰·霍普克罗夫特）在2017年初预测，随着人工智能与自动化技术的兴起，未来美国劳动力规模可能需要缩减50%。

中国能为改善不平等做些什么？数字时代，能否在技术手段的帮助下缩小鸿沟？我们在此仅抛砖引玉，希望能引起更多有识之士的思考和探索。人工智能教育大脑应该被投入职业培训中，让机器辅助普通工作者学会与人工智能打交道，学会在人工智能时代寻找工作价值。

政府当然应该出台相应的保障措施，但关键还在于对生产方式本身的改善。美国制造业空心化，高新产业未能反哺制造业，而是与华尔街资本结合，在全球寻找利润，加深了美国内部的数字鸿沟。中国制造业发达，如果加以指引、支持，求得总体主动的脱胎换骨，就有可能创造出新的平衡生态。

已故著名匈牙利裔英国政治经济学家卡尔·波兰尼在《大转型》一书中指出：资本主义和工业化把一切数字化、理性化，并挤压“人性”，尤其是将市场凌驾于社会之上，而非让市场嵌入社会，所到之处，把一切社会性关系驱赶出去。比如英国“羊吃人”的圈地运动，实质是为了生产羊毛商品而驱逐农民，从而毁灭乡村社会秩序。这一切引发了人道灾难，也导致了一个反向的社会保护运动，包括对福利的争取、对劳动保护法的追求等。经过几百年的动荡和调整，西方资本主义国家才达到了生产与福利的基本平衡。

温故知新，历史经验教训值得我们吸取借鉴。今天，我们观察到这样的平衡再次遭到危机，越来越多的人把眼光投向中国，从马丁·雅克这样的战略学者，到凯文·艾什顿这样的技术专家，都对中国的发展报有极大希望。中国高效的政府和民间追求发展与公平的强大传统都为此提供了新的可能。

著名经济学家林毅夫坚持认为“有为政府”是社会发展的保障。面对智能革命，“有为政府”需要把握好轻重缓急，重点领域的竞争与升级不能放缓，但是事关国计民生的领域需要做好平衡，把转型期的社会波动降到最低。20世纪90年代国企改革期间，中国政府结合社会力量，兴办了比较成熟的再就业培训体系。今天需要升级这个体系，引导普通劳动者熟悉智能技术并能学以致用。发挥社区街道作用，为居民提供智能生活讲座。甚至在基础义务教育阶段就为学生提供学习智能技术的机会。

大公司在技术发展中也要有社会担当。技术人员需要多思考社会责任，努力把人工智能技术嵌入社会发展，而不是把人工智能本身的发展凌驾于社会之上。这方面已经有了很好的尝试，比如用语音、视觉等方面的智能技术帮助落后地区群众阅读、帮助残障人士生活、为劳动力培训设计产品等。在这些方面，技术人才还需要多想一想，比如面向大学生和技术人员的在线教育平台能否尝试为普通劳动力提供培训？这一切都需要政府与企业的通力合作。

工具理性之问

在技术进步面前，人们没有选择，只有奋力前行。因为技术不是外来者，正来自人类生产创造本身，和人类的存在不可分割。

先哲曾经仔细讨论技术或者工具与人类的关系。马克思首先承认资本主义和工业化短时期创造出了超过人类过去几千年创造的财富总额，总体上改善了人类生活。但接着他谈到了“异化”，意思是人类创造的东西反过来奴役人类，当然，不是所有的人类。另一位大思想家马克斯·韦伯则认为，资本主义的货币化及其处理一切事物和关系时候的“算计”化，创造了所谓的工具理性。工具理性化原本许诺给人类“轻盈的斗篷”，结果却成了“铁衣的囚笼”。

工具将人类从“自然的暴政”下解放，但工具本身、工具与社会关系的结合，又造成新的不平等。谁垄断了工具，谁就垄断了经济政治的命脉。曾经匮乏但是平等的原始社会，因为工具，分裂为垄断“智能”的统治阶层和不拥有知识的被统治阶层。

蒸汽机加速了工业革命的车轮。早在埃及的托勒密王朝时期，寺庙就已经运用蒸汽动力，但仅仅用来升起神像，制造神迹。这正是上述垄断智能的缩影。只有在资本和智能社会化以后，蒸汽机才能充分发挥潜力，给文明注入强劲的动力。

工业时代在更高的层次重构了智能和劳动的分裂。大规模制造的一大弊端，就是思想和行为的分离。创新任务由脱离生产的科学家和工程师完成，生产工作则由非熟练工人完成。

云社会和物联网提供了把实践劳动与脑力劳动结合起来的可能。

在传统工业体系中，工人是“标准化”的工具，尤其是流水线上的工人，只能被动按照指令完成规定动作。“工作”异化了，不再是一种创造性的行动，而是消极的机械化行动。如果我们大胆设想，未

来人工智能或许可以提供技术性解决办法。比如考虑到劳动力的多样性，超越流水线，创造一种智能化的生产流程，根据劳动力状况，自动适应、自动调整，不会因为某个环节的工人不标准就影响全局，同时给工人个体发挥的机会。企业的智能运营系统通过网络来连接一线工作者与科研人员和管理者。随时记录、定位和分析工作者的操作习惯、状态，给出优化建议，使工作和学习同步完成。

中国产业链完整、工作场景多样、人才层次丰富，这是宝贵的财富。科学家和工程师应该有意识地深入生产场景，把智能技术与工业实践和生活实践结合起来。

机器人的无用之用

人工智能还给我们带来另一种突破工具理性的启迪。

回想一下文艺作品里那些陪伴我们童年的机器人，从铁臂阿童木到“大白”，它们无不超越了人类最初给它们设定的工具属性，奋力追求自己的独立“人性”。

在《钢铁巨人》里，机器人为了拯救人类村庄，选择与原子弹同归于尽。它在最后一刻说的话是：“我知道自己要做什么了，我要做Super man（超人）！”机器人本身能否形成人性甚至超越人性，是个有趣且有意义的话题。

多梅尼科·帕里西所著的《机器人的未来：机器人科学的人类隐喻》不只是研究机器人，更把机器人当作一种方法，来反思人类科学的工具性，指出人类自身的方向。

科学的一大基本特征是“对象化”——人与自己的研究对象分离，完全“客观”地观察、记录、分析对象，尽量不带入观察者自身

的因素，诸如情绪、立场等。这套方法是人类认识自然世界的有力工具，但是在用于认识人类自己的时候却遇到了困难。

首先，人作为自己的研究对象，无法把研究者和研究对象完全分开，人之精神就是由自己的情绪、动机、愿望、认知等组成的。

其次，关于人类行为和人类社会的研究理论基本是用文字而非自然科学中的数理符号来表达的。文字作为表达科学理论的工具有很大局限性：意义不准确、下意识地含有研究者的价值观和情感取向、难以数字化等。文字不是透明的中间工具，文字本身就构造了一层“现实”。文字甚至是数字，其本质都是符号，符号与现实总有差距，正如关于颜色的有限词汇是无法完全表达现实中无限多种类颜色细节的。文字甚至会出现没有现实对应物的情况，比如“信仰”“目标”等这类词汇。

大数据和数学方法的发展使我们已经可以把很多人类学现象数字化，但依然不足以研究人类。帕里西说：“（这些方法）甚至未曾试着去识别隐藏于这些行为现象和社会现象之后的机制与过程，更不要说去解释了。”

人工智能为表达关于人类的科学理论提供了新的可能，帕里西提出了“作为人工构造的理论”（theories as artefacts）——创造人工构造（类人机器人）来代替研究。这个理论基于这样一条原则：“无论X是什么，要理解X，就必须以人工对其进行再现。”如果你的理论不好，人工构造就无法造出来。如果你能造出类人机器人，就说明你的研究成功了。从根本上说，这是把研究和实践统一了，研究就是实践，实践就是研究。这个过程除了可以克服前面所说的人文学科问题，还能把多种学科诸如社会学、心理学、生物学、政治学、语言学、经济学、文史哲等整合在一起，打破目前各类学科机械分割的局面。

帕里西认为，目前的机器人或者人工智能与人的区别在于，人的思维中包含了动机与认知两种模式。人的任何行为都有动机，比如为

了吃喝与生存、为了安全与繁殖、为了荣誉与理想等。而现有的机器人只有认知，无所谓动机，因为它的目的是由人设定的。机器人只是工具，为了满足人的特定需要而存在。

但是机器人可以是人的影子，为此他区分了两类机器人：实用机器人和类人机器人。后者正是科学家需要的。前者只有认知思维，后者却设定了动机，比如设定机器人需要定期获取能量以求得生存。具体做法是在计算机中构造一个机器人程序，可以为它装配饥饿传感器，令它在一定的虚拟空间里寻找食物和水。食物用食物令牌代替，水用水令牌代替，两种令牌分布在不同区域，然后让机器人自主决定获取这些令牌的路线和顺序，找不到就会死亡。以此观察机器人如何做出决定。

同理，可以设定机器人需要规避一些危险令牌以保障安全，需要通过与其他机器人交换编码而产生新的后代（繁殖），甚至设立财产观念、遵守规则的动机等。观察在这些动机的驱动下，机器人如何进化出类人的社会意识、语言、文化、情感乃至艺术和宗教。比如，制造男性机器人和女性机器人，并赋予机器人保护后代机器人的动机，观察它们如何发展出家庭观念和实体。

科学家这样制造的机器人可能没有任何实际应用价值，比如会犯错的机器人，会生病的机器人，会做梦的机器人，会难过、会高兴的机器人……而工具是不应该犯错误、开小差的，这些是人类的特性。在这种思想方法下，制造人工智能的过程不再只是制造服务于人类外在需求的工具，而正是人类了解自己的过程。

这样的研究已经有了一定成果，开拓了过去的心理学、社会学从来没有过的视野。

不过读者可能会看出一点点“危险”，类人机器人会不会最终演化出自我意识？它如何处理与人类的关系？这种研究会被宗教人士看作是一种僭越吗？因为科学家在做上帝做的事情——造人。不过科学研究与宗教不同，本质是对人类自身的对象化研究。生命就是不断地

“生成创造”。工具理性问题在这里遇到了新的场景：在机器人内部实现对工具理性的超越。归根结底，这是借助人工智能展开自我反思，同时又指导人类发展人工智能，对我们思考机器人如何融入人类社会，如何参与人类的文化、社会生活都会有丰富启发。

基督教新教把人类看作上帝实现自我意志的工具，而人类的自我反思总是在克服这种工具性。人不应该只是工具手段，人应该是目的和手段的统一，这一点可以在人工智能身上付诸实践。

二十三条军规

为“类人机器人”设定伦理和规则看来也是必要的。

在2017年1月5~8日举行的“Benificial AI 2017”会议上，1980名来自产业、学术和研究界的人工智能相关人士共同制定了“二十三条原则”，宗旨在于确保人工智能的健康发展，人人都能从人工智能中受益。这二十三条原则分为研究问题、伦理价值观和长期问题三个部分。[2]

研究问题方面，提出人工智能研究的目标是创造有益的智能，而不是让它像经历生物演化一样没有确定的发展方向。投资人工智能应该附带确保该研究是用于发展有益的人工智能，包括计算机科学、经济学、法律、伦理和社会研究中的棘手问题，例如：我们如何使未来的人工智能系统具有高度鲁棒性[3]？这样才能够让它们在没有故障或被黑客入侵的情况下做我们想要它们做的事情。我们如何通过自动化实现繁荣，同时不打破资源和目的的平衡？我们如何更新法律制度实现更大的公平和更高的效率，跟上人工智能的发展步伐，管控与人工智能相关的风险？人工智能应该具有怎样的价值观，应该具有何等法律和伦理地位？

其他研究问题包括人工智能研究人员和政策制定者之间应该进行有建设意义、健康的交流；培养合作、信任和透明的开发文化；避免竞赛，避免在安全标准方面进行削减。

伦理价值观方面，二十三条原则提出了很多愿景，包括人工智能系统的安全，故障透明度（如果人工智能系统造成伤害，应该可以确定原因）；由机器人做出的任何涉及司法决策的行动都要透明可解释，供主管人权机构审核；设计者与使用者要担起道德影响责任；要确保高度自治的人工智能系统的目标和行为与人类的价值观相一致，尊重人类的尊严、权利、自由和文化多样性；保障个人隐私，不仅人工智能系统能够分析和利用人类产生的数据，人类也应该有权获取、管理和控制自身产生的数据；不得不合理地限制人类真实或感知到的自由。人工智能带来的利益应当是普惠人类的；由人类选择如何以及是否委托人工智能系统去完成人类选择的目标；人工智能的力量应当尊重和改善社会健康发展所需的社会和公民进程，而不是颠覆这种进程；避免人工智能军备竞赛。

长期问题方面，建议在没有达成共识的情况下，避免对未来人工智能能力上限做出较为肯定的假设；通过相应的关怀和资源对其进行规划和管理，应对风险；对于那些不断自我完善或通过自我复制快速提高质量或数量的人工智能系统，必须采取严格的安全和控制措施。

最后一条是“共同利益”：超级智慧只应该为广泛共享的伦理理想服务，为了全人类而不是一个国家或组织的利益而发展。

二十三条原则，这个名字不由让人想起黑色幽默文学代表作《第二十二条军规》，那是一条不可能完成的规则。看起来，这二十三条原则更像是二十三条呼吁，提出的是问题而不是解决方案，代表的是愿望而不是已经实现的现实。其中涉及很多价值观问题，科学家倡导发展机器人要遵循人类的价值观，这是理所当然的。不过这里就会有三个问题：

第一，这里制定的价值观基础在于把机器看作工具，要求人类在使用工具时不得损害人类利益。实际上，在任何工具的使用上，人类都没有完全做到这一点，比如军备。不过确实可以尽力加以限制，比如军备条约。

第二，人们之所以前所未有地郑重其事，是因为意识到机器人不同于以往的工具。普通工具没有生命、没有意识，完全听命于人的操作。但机器人如果具有了类人意识，那么人类怎么对待它们？显然二十三条原则并未考虑这个问题。人类拒绝奴役人类，把人当作工具就是一种奴役，那么对机器人呢？如果人类的价值观就是让机器人做工具，类人机器人会认为这种价值观有尊严吗？

第三，人类的价值观本身就存在各种冲突，众所周知，制定这些原则的与会者基本属于硅谷阵营。硅谷阵营在美国2016年大选中的态度及其价值观就和特朗普阵营发生了激烈冲突，他们制定的规则看上去很美，但在现实中未必符合很多人的价值观。

二十三条原则表达了人类良好的愿望，相当于科学家为机器人设置了一套“政治正确”。考虑到“政治正确”在当下美国的尴尬遭遇，我们要明白，人工智能的“政治正确”不是一蹴而就的，集体签名是一回事，做起来是另一回事。价值观不是静态的观念，而是在博弈中生成。价值观也不可能孤立地存在于某种群体身上，比如机器人的价值观实际上是人机的价值观。未来人工智能可能将加入这个博弈，形成新的、动态的人机价值观。我们需要在实践中不断摸索人工智能与人类发展的健康之路。二十三条原则起到了探路者的作用。

现实的法律问题

虽然“算法代替律法”的前景已经出现在不少科技人士的视野里，但到目前为止，法律依然是调节人类社会关系的主要手段。我们需要考虑，人工智能时代的法治会遭遇什么挑战，应该如何应对。

人工智能不仅仅是技术问题，而是整个社会运作方式改变的问题。《少数派报告》这样的电影已经预见了事前的预防性管理会出现。深度学习技术的最大作用在于预测，这也给以往法律思维带来新的启迪。比如法律将从事后补偿模式向事先预防模式变化。不过这个过程将非常不确定。

以算法的复杂性来说，技术界区分了强人工智能和弱人工智能。有学者认为，人工智能在技术认知上没有问题，但在法律上很难按照智能程度给出精确的标准。因为法律应对复杂世界的方式是确立一般性的简单规则，在概念上对社会个体进行抽象假定（如行为能力），而非针对特殊主体，否则规则体系本身将变得异常复杂，难于理解和操作。生产资料之间的信息变得越来越对称，但作为信息匹配中介的人工智能却变得更不透明，其规则设计和运作从属于用户甚至开发者都无法理解的秘密状态，这都回到了法律如何处理与代码的关系问题。[4]

具体体现在法律抽象化与技术黑箱化之间的冲突。比如，“快播案”这样一个涉及互联网技术而非人工智能技术的案件里，法律责任认定的过程很漫长也很艰难。监管者或法院并不想深入算法内部了解造成事故的技术原因是什么。只要法律认定这一黑箱应当在合理范围内得到控制，事故就可以避免，黑箱提供者就应当承担责任。在这种情况下，我们可以预见，保险（甚至是强制险）就成为那些事故发生概率小但潜在损失巨大的市场的不二选择。

涉及技术的航空、医疗保险市场已经十分发达，可以预见保险将会延伸至更多由人工智能驱动的服务行业。

在另一方面，也许真要靠算法的顶层设计来防止消极后果。人工智能技术可能不只是理工专业人士的领域，法律人士以及其他治理者也需要学习人工智能知识，这对法律人士和其他治理者提出了技术要求。法治管理需要嵌入生产环节，比如对算法处理的数据或生产性资源进行管理，防止造成消极后果。例如，征信系统中禁止收集种族、性别、宗教派别、政治倾向等歧视性信息；在声誉系统中打击网络推手刷单，避免虚假数据；通过地域性监管限制网络专车及其司机的资质等。法律机器人本身也是一种辅助人类面对复杂规则做出判断的好办法——用技术进步来解决技术带来的问题。

与人工智能相关的法律问题可能会很多，这里简单谈谈日常生活中人们比较关心的隐私问题。

有人说，智能时代，人们进入了无隐私社会，因为一切数据都处在互动中，哪怕是心率都被可穿戴设备分享了。未来的狗仔队也会智能化，运用数据分析和无处不在的视觉监控寻找明星的蛛丝马迹。在智能社会如何善待隐私？可能要突破原先的观念。隐私当然是现代个人人格的一部分。现实中，人们强调保护隐私的同时，也热衷记录隐私，甚至传播隐私。一切网络八卦行为都围绕隐私展开。没有隐私就没有独立个体，但在数据时代，绝对的隐私又会让个人无法被沟通和识别。最好的办法还是制度与技术手段的突破，比如建立统一的数据保护平台，让个人可以了解自己的数据被政府、公司使用的状况，避免单向过度的使用。

数字权力的重新分配

已故著名波兰裔英国思想家、社会学家鲍曼认为：数字时代带来了比监控更麻烦的问题，权力与政治在技术之下分离了。传统的政治

囿于解决一个国家内部的问题。而跨国资本及其政治力量，却通过技术蔓延各个领域，制造出更多不确定性。同时，普通人对技术和官僚的信任在这种矛盾之下消解。权力随着资本和数字网络流动起来，传统政治却对约束它们无能为力。作为对全球化资本权力的反弹，欧美诸国“部落主义”重新崛起。这也就是英国脱欧、孤立主义在美国兴起的大背景。

鲍曼描述了西方国家的情景。但不确定性是新时代普遍存在的状况。

政治、经济、文化、传媒权力都会被数据智能渗透。这个不难理解，金融资本借助数字网络、智能投顾、闪频交易等技术在全世界腾挪闪移、惊鸿遍野。推特等社交媒体催化了中东国家的动荡，数字权力越过主权国家的界限，纵横捭阖。甚至以往向外输出数字权力的美国，也开始担心被跨国数字权力侵袭。阿桑奇和他的维基解密作为一个非政府组织，屡屡抛出政府黑材料，让美国政府非常头疼。黑客凭借网络技术获得了威慑美国统治阶层的能力。美国政府甚至指控俄罗斯使用黑客干预了美国2016年大选，这种紧张感是前所未有的。这既说明数字权力面前无人可以幸免，同时也预示着，从个人到国家，都有必要加紧适应数字智能时代。

数字权力的形态如同福柯所描述的现代权力，并非中央集权式的环状结构，而是错综复杂、多中心存在的网状结构。数字权力弥散于这张网中，难以捉摸。

从国家层面讲，需要在顶层设计中做好设置，防止政府滥用数字权力。而恶意挑战国家与社会正常运行的数字权力也已经在现实中出现。

彭博周刊2014年曾报道了一位哥伦比亚黑客，自称操纵了9个拉美国家的投票选举，手段包括窃取数据、安装恶意软件、在社交媒体上伪造大规模支持或反对的民意。

他的团队在对手阵营总部安装了恶意软件，可以监听电话、监视计算机。借此获得各种演讲稿、会议计划和选举安排。根据这些信息，他借助virtual army（水军）发布虚假的推特消息，利用低端推特机器人追加大量点赞和关注。设置程序在凌晨自动拨打竞选电话，以对手阵营的名义骚扰选民。借助类似手段，他曾在不同程度上影响了委内瑞拉、尼加拉瓜、巴拿马、洪都拉斯、萨尔瓦多、哥伦比亚、哥斯达黎加和危地马拉的民主选举，时间长达8年之久。他说：“当我意识到，人们更愿意相信网络舆论而非现实，我发现，我拥有了让人们相信任何事物的能力。”

在国内外的社交媒体上，具有简单智能的机器“水军”都蠢蠢欲动。“未来你分不清电脑对面是不是人”这件事已经成真。不过，“水军”并不一定意味着负面，因为用于正确目的的“水军”并不水。某种意义上，问答机器人和机器客服就是“水军”的变种。即便在最容易“水”的论坛、评论区，机器人的表现也可能好过人类。王海峰就说过，在很多评论区，往往是一些人类在灌水、乱喷，而机器人则在发表有营养的评论，带动评论的热度，保持积极正面的舆论方向。

每个现代人都被自己生产的数据包围着。智能化的数据已经成为人类的第二肉身。就像第一肉身会担心疾病、车祸等安全问题，第二肉身也同样面临安全隐患。

互联网时代个人信息面临种种泄露风险。诈骗业的技术也在更新。比起通常的电话诈骗，不法分子甚至用起了大数据方法。上海黄浦区警方在2017年初破获了一起惊人的网络盗窃案。首先是黑客用软件批量生成电话号码，再利用扫描工具逐个把号码挂到网上扫描，通过一些黑客存放泄露密码的网站（一般自称“社工库”），把电话号码对应的登录密码扫出来。这在业界被称为“撞库”。用这种简单粗暴的方法，可以快速得到很多用户的登录信息，然后用于偷取银行卡存款等犯罪活动。

一个社会中总会存在不法分子。互联网时代数据技术爆发式增长，人类适应能力的落后也给罪犯留下了可乘之机。这就如同汽车刚被发明时给社会带来新的交通不安全，不过后来人们发明了红绿灯，用新规则制约技术，保障安全。传统的数据管控方式在互联网时代显得千疮百孔，普通人或者团体很难主动保护好自己的隐私数据。这需要多方共同努力，警方同样可以利用大数据技术来对付犯罪，科技公司也可以为此助力。

大数据领域的红绿灯也出现了。以电信诈骗问题为例，包括“百度安全”在内的多家公司已经与管理部门合作，携手打击电信犯罪。依托大数据建立“安全号码库”和“诈骗电话地图”，一方面可以将信息共享给公安机关，定位诈骗嫌疑人的号码和所在位置；另一方面可以实时拦截诈骗电话并在诈骗电话地图上同步显示、识别诈骗电话动态，“亮红灯”提醒用户。“百度安全”至今已经积累了2亿多安全号码库，每日拦截骚扰诈骗电话超100万次，诈骗电话拦截率保持在99.98%以上。与此同时，在手机上安装了安全软件的用户，对可疑电话的举报、标注都可以上传到服务器，帮助识别诈骗号码。在这种动态循环中，安全软件会越来越聪明，识别诈骗号码的能力也会越来越强。

深度学习使防火墙技术也得到了升级。做网管的人都知道，过去每次安装防护软件，都要设定一大堆规则以预防非法访问。而采用深度学习技术，可以让安全系统自动学习用户访问规律，识别异常访问。系统自己设定规则，远比人类网管高效。

信息时代，虽然一些恶性诈骗案件和安全事件被曝光引发了恐惧，但是我们要知道，有更多的数字诈骗被阻止了，提供支持的正是人工智能技术。无论是网络金融安全技术还是智慧城市的智能监控技术，都是魔高一尺，道高一丈，在博弈中不断提升。

更多时候，我们身边的技术极客会利用技术能力做出一些“欢乐”的越界行为。比如支付宝刚推出AR红包，就有人通过训练卷积神

经网络，反向破解被横条遮住部分画面的线索照片，去掉横条，获得原始图像，轻松找到红包。方法并不难，前提是使用者懂得使用深度学习共享平台。

北京、上海的车牌特别难摇。车牌中签号码由系统“随机”给出，但这个随机也遵循一定的算法。有人居然利用机器学习技术，琢磨车牌摇号的规律，据说提高了中签率。

环顾一下你所在的工作单位，你会发现有的人擅长使用很多技术工具来强化自己的能力，有的人却整天为了记不住各种登录密码而苦恼不堪。在未来，对数据智能的熟悉程度会影响一个人工作、生活的幸福感。做一个有准备的人，才能在数据生活中游刃有余。

新世代 新未来

最后让我们还是把视线投向年轻人，他们是未来人类社会最宝贵的财富。

在日本、中国、欧美等许多发达的城市社会中，兴起了一个被称为“尼特族”的群体，主要指一群无业、在家啃老、沉迷于动漫作品的年轻人。据统计，在日本15~34岁的年龄段中尼特族数量高达60多万人，占该年龄段人口的2.2%。

尼特族人群往往沉迷于“二次元”文化，和所谓“宅男”群体高度重合，对动漫游戏中色情、暴力的向往和对现实世界的恐惧成为这个群体最基本的特征。他们对屏幕里生动可爱的动画角色已经丧失了抵抗力，当面对一个各方面都完美的人工智能角色甚至实体机器人时，又会是怎样的不能自拔？

2016年6月，吴恩达与刘慈欣进行了一场关于人工智能的对话。“大刘”对未来人工智能提出了自己的大胆设想：人工智能可以成为人类的性伴侣，并从根本上改变人类的文化趋势。

人工智能是否可以在生活中彻底替代“人”的角色？作为科学家，吴恩达的回应显得更谨慎，称当下人工智能在技术上还相去甚远，但是也没有否认这种未来出现的可能。

未来人类可以按照自己的愿望制造出各种“完美”事物。但这真是完美吗？会不会只是人类自身的镜像？

当人工智能破茧而出，人类却可能将自己封闭于虚拟之中。刘慈欣写过一个短篇科幻小说，写的是当人类把大脑接入虚拟世界，从此醉心于虚拟世界的游戏，感受上帝般的虚妄，而不再向宇宙张望，地球文明就闭上了眼睛。

和所有新技术一样，人工智能也会被用来向人类许诺美好的生活。它已经使很多人的生活变得更舒适。但人类依然需要奋斗精神。最早把巴菲特投资理念翻译介绍到中国的中国建设银行前董事、经济学家孙涤先生认为，西方经济学以理性化个人是经济发展的基础为前提是不对的，那些敢于大胆创新的人恰恰是具有动物精神的人。人工智能基础上的精细化社会可以让人的生活变得精致，但人因此可能反而被各种精致的教条束缚，失去勇猛的动物精神。未来每个人都可以轻易接入智能流，但是从现在起就要想一想，该把智能流运用到哪里才符合人的价值。

每当一个新时代来临，总有一部分人摩拳擦掌，一部分人随波逐流，还有一部分人茫然无措。这几年常有人说“未来已来，只是尚未流行”，充满乐观主义情绪。但我们要意识到，从来没有完全有利的变革，总是会包含失落、分化和冲击，只是人们后知后觉。人工智能科学家虽然深处技术金字塔尖端，但是行动态度普遍持重。百度深度学习实验室的周杰就说自己的态度比较“保守”，保守的意思不是踌

躇不前，而是在坚信人工智能发展方向的同时，不赞成过于激进的态度。

一百多年前，严复把《进化论与伦理学》翻译介绍到中国。他在翻译中故意“曲解”了原著，只强调生物进化中竞争的一面，而省略了赫胥黎关于伦理的思考。身处当时第一强国英格兰的赫胥黎，梦想一个对子民更加“仁慈”的英格兰，而面对“三千年未有之大变局”的严复，则强调丛林法则，用“适者生存”召唤一个复兴图强的中国。时过境迁，今天中国人站在崛起的路上，既要继承先贤的忧患，又要有所超越。与赫胥黎一样，我们既应思考智能社会的力量，也应思考如何让智能社会更加和谐。

忧思不代表悲观，也只有在忧思基础上的乐观，才是真正的乐观。想象未来是一件困难的事情，虽然未来无比诱人，值得人们为之奋斗。

在未来漫长的智能岁月里，即便诸如百度、谷歌这样的人工智能公司，可能也只是浩瀚历史中的一个过客。我们是人类，有我们的弱点和优点，有我们的短视，也有我们不灭的抱负。古人说“不以物喜，不以己悲”“担当身前事，何计身后评”，正是我们人类担当精神的体现。我们能做的就是抓住当下。人类的存在就是“在路上”。百度要为美丽新世界铺垫好最初的基因，中国要从大国变成伟大的智能文明国家，每个人也都应该不甘落后于机器，努力做更好的人，知道更多，做到更多，体验更多，一起向着美好但不确定的未来进发。

---

[1] 托马斯·亨利·赫胥黎（Thomas Henry Huxley，1825—1895），英国博物学家、教育家，达尔文进化论最杰出的支持者。著作《进化论与伦理学》因为严复的翻译而深刻影响中国发展，汉译名为《天演论》。

[2] 资料来源：据人工智能产业智库“新智元”的翻译。

[3] 机器人的鲁棒性就是描述机器人在内部或外部存在扰动甚至变化时保持原有或期望状态的能力的一种特性。

[4] 胡凌。人工智能的法律想象［J］.文化纵横，2017（2）.

后记

有时候，描摹正在发生的现实，比书写历史更难。

尤其当你要描述的是一场即将到来的革命。

是的，智能革命——我们确定这是一场伟大的变革，并对此深信不疑。

因为，作为一项源于天才创见的技术思潮，它从诞生到现在，逾60年，其间几经沉浮而火种不灭，已积蓄太多能量。

数据化世界的到来，使得科学界那些曾经被忽视、被冷落，甚至被怀疑的天才算法和公式，有了直接验证的可能。人工智能，这架梦想般的“概念机器”，忽然间获得了保证其高速运转的源源不断的燃料和动力。

更重要的是，产业界开始接过学术界的火把，曾经充满科幻色彩的前沿智能技术，真正开始走出实验室，走进并照耀普通人的生活。

幸运的是，作为一家搜索引擎公司，百度从诞生的那一天起，开发流程（大数据—深度学习—提取模式—创造用户价值）和开发文化，就已带有人工智能的天然基因。

也许是命运使然，在全世界范围内，百度是布局人工智能领域的领跑者，也是人工智能领域最坚定的实践者之一。我们正亲身见证人工智能在搜索生态体系中展现出的神奇能量和更广阔的前景。

如果人类进步的主旋律是通过感知和认知能力的不断提升，从而知道更多，做到更多，体验更多，那么人工智能就是这一主旋律的最

新回响。

但是，面对一个令人无比振奋、无比期待的智能社会的未来，我们又是惶恐的。

因为，我们虽然确信人工智能将给整个社会带来翻天覆地的变化，将给各行各业带来巨大的价值。但是，我们在现阶段又无法精确描绘这样的变化，短时间内可能也无法呈现这样的价值。

一家劳动密集型的制造业企业该如何实现产品的智能化升级？一家大型农场该如何实现真正的精细化农业？一家金融公司该如何防范风险提高收益？一家制药公司又该如何跟上个性化医疗的未来……

许许多多的行业带着智能化升级的困惑找到百度。这其中，有些是百度探索过或正在探索的，也有很多是百度并不了解的。因为各行各业都有其自身演进的规律和存在的合理性，亟待人工智能去描绘的知识图谱千差万别。

所以，当我们试图以一本书来描绘各行各业智能化转型的图景时，常常是力不从心的。其中的疏漏和错误在所难免，唯有留给方家指正。

但这也许正是我们写作这本书的必要性所在——通过敞开自身能力配置，并尽可能多地描述合作的空间，去吸引更多的潜在合作者，从而为未来“你有故事，我有酒”的完美契合打下基础。

我们认为，对亟待转型升级的传统企业来说，首要解决的就是明晰人工智能的标准。因为，也许是历史给予中国经济摘掉低端仿制山寨的帽子，真正执智能产业牛耳的一次机会，而不是沾沾自喜于以往的盲目跟风、概念炒作。有了权威的标准，传统企业才能逐渐找到自身智能化改造的坐标，也才能防止在开始阶段就误入歧途。

感谢Robin，是你确定了在智能革命即将到来之时，应该面向用户、合作伙伴开放百度的智能生态、普及基础知识、确立规范标准的方向，也才有了本书的由来。让人惊喜的是，你在确定了本书的逻辑

架构和基本脉络之后，更亲自执笔撰写了重要章节。作为百度的首席人工智能官，你对未来智能时代的描述令人神往。

感谢百度集团总裁兼首席运营官陆奇，你对智能革命带来的计算革命契合人类进步主旋律的深刻洞察，让我们豁然开朗，你对量子计算的侃侃而谈，让我们受益良多，也让我们兴奋不已。

感谢百度总裁张亚勤，百度高级副总裁朱光，百度副总裁王路，百度副总裁、ACL会士、前主席王海峰，百度副总裁、金融科技创新中心负责人张旭阳，感谢原百度首席科学家吴恩达。与你们的深入交流保证了本书的思想高度和国际化的视野。你们的前沿研究与战略拓展，成就了今天百度在人工智能领域的领军者地位。

特别感谢百度研究院院长、深度学习技术与应用国家工程实验室主任林元庆，百度人工智能产品委员会主席景鲲，百度大数据实验室主任范伟，百度地图事业部总经理李东旻，百度增强现实(AR)实验室主任吴中勤，百度语音技术部总监高亮，深度学习研究院深度学习方向负责人周杰，机器翻译技术专家何中军，百度数据金融架构师杨晓静，百度大数据实验室科学家吴海山，百度语音技术部刘洋，智能投顾团队袁月，百度金融平台部吴建民，以及靖凌云、和为、关勇、王涛、韦成文、倪金节等多位同学拨冗接受访谈，并帮忙收集相关资料。

你们身处不同体系，却同样以人工智能为视角，成为各个业务智能化升级的“提出问题者”，你们让百度大脑的“听”“看”“读”“说”的感知能力不断提升，并开始逐步向预测、判断的认知“洞见”方向迈进，你们是来自未来的人。

感谢徐菁的坚持，让本书对人工智能行业标准的描述逐渐清晰。

感谢顾国栋、熊赟、王佳、马连鹏等同学的极致追求与不懈努力，没有你们的督导和内外协调，本书不可能在如此短的时间内完成。

感谢李雯婷、蔡硕、任一奇、乔慧、李颖超、王琳琳、马晓昕、陈明、张开翼、张娜为本书制作AR展示，使得本书展现出科幻加魔法的神奇魅力。

最后，要特别感谢中信出版社的编辑团队，他们的敬业、专业和严谨为本书增色不少。

《智能革命》项目团队

2017.3.5

AR效果展示说明

第一种方式：

点击手机百度搜索框内的“相机”图标，使用其中的“AR”功能，扫描图书封面和书中相关图片，即可体验互动效果。

第二种方式：

第一步：下载“智能革命”App（安卓系统请到百度手机助手下载，苹果设备请到苹果应用商店下载）。

第二步：点击打开“智能革命”，进入应用。

第三步：点击应用中的“AR扫描”，调取摄像头，将本书封面正对摄像头。

第四步：AR扫描识别成功后，3D立体影像即刻呈现。根据语音进行相应的交互操作。

第五步：根据识别图片及页码标识找到书中相应图片，点击“AR扫描”即刻呈现炫酷AR效果。

发现：点击“发现”，与3D影像互动合影（可将3D影像放大、缩小、移动）。

更多：点击“更多”，可进入帮助页面、App版本说明页面及软件重置。

图书在版编目（CIP）数据

智能革命：迎接人工智能时代的社会、经济与文化变革 / 李彦宏等著. -- 北京：中信出版社，2017.4
ISBN 978-7-5086-7321-9
Ⅰ. ①智… Ⅱ. ①李… Ⅲ. ①网络公司－企业管理－经验－中国 ②人工智能 Ⅳ. ① F279.244.4 ② TP18
中国版本图书馆CIP数据核字（2017）第038618号

智能革命——迎接人工智能时代的社会、经济与文化变革
著者：李彦宏 等
出版发行：中信出版集团股份有限公司
（北京市朝阳区惠新东街甲4号富盛大厦2座 邮编100029）

智能与智慧

人人

人工智能

潘云鹤 中国科学

透视中国哲学家

心

宋冰 X 编著

中国传统儒释道
如何看待
现代人工智能的未来

赵汀阳、张祥龙、何怀宏、
刘晓力、干春松、陈小平等17位学者

X

聚焦中国和全球语境下的
科技哲思

中信出版集团

智能与智慧

——人工智能遇见中国哲学家

宋冰 编著

中信出版集团

目录

序言 智能与智慧：人工智能遇见中国哲学家

中西学交融语境下的哲思

技术的无限进步也许是一场不可信任的赌博a

人性、时间意识与人工智能新境a

回到“轴心时代”思考人工智能

人工智能，吾与也

从马克思主义和儒家的视角看人工智能

人类意识与机器意识的演化

浅识、误解与想象力

人工智能带给儒释道的挑战与机会

人工智能会帮助我们获得真正的美德吗

机器智能的兴起与人之为人理想的终极

人工智能与情感、秩序：从儒家的视角出发

机器人能否循道而行

当人工智能遇见道教

解决人类生死烦恼的根本智慧

新人文与生命价值

科技哲思

人工智能中的不确定性问题与哲学研究

成己而圣：在此岸建构理念世界

构建人类与人工智能和谐共生社会

编后记 和谐与慈悲：前沿科技时代的基础价值观

致谢

序言

智能与智慧：人工智能遇见中国哲学家

宋冰

前沿科技时代基础价值观的错位或缺失

以人工智能、基因编辑、大数据与量子计算为核心的前沿科技时代呼啸而至。人工智能和机器人应用的广泛场景已经开始深刻影响人们的生活方式、理念，甚至人之所以为人的思考。全球呼吁建立人工智能与机器人伦理规范的声音此起彼伏，从2015年至今，全球共有超过50份与人工智能和机器人相关的伦理原则与价值声明。其中有如公正、尊严、自由和人性的抽象原则，也有在个人权利基础上的保护隐私、反偏见等原则，以及诸如可解释性、安全性与鲁棒性等以技术性考量为重的原则。由欧盟人工智能高级别专家组起草的《可信赖的人工智能伦理指导》指出，指导这些原则的基础价值观是“尊重人权、民主与法治”，且推进个人基本权利的方式方法。[1] 中国、美国和新加坡的人工智能发展战略与规划纲要等各类官方文件所触及的基础价值观包括以人为中心、打造国际竞争优势，保持在研发和规则制定上的战略优势以及保障国家安全等理念。

前沿科技之所以具有颠覆性，在于它技术迭代快、发展进路不确定且呈非线性轨迹，同时各项技术的结合往往产生叠加效应，常常会出现意料之外的结果。通过广泛的应用场景，人工智能与机器人等前

沿科技全面渗透我们的个人、社会、经济、政治生活，对我们的生活方式和理念有着令我们始料未及的深刻影响。有的人甚至认为这些科技的发展会威胁人类的存续。然而，目前社会上提出的价值观和伦理原则多半是工业革命时代的主流价值观和伦理原则，这些足以应对如此具有颠覆性、迭代迅速、发展前途扑朔迷离的前沿科技吗？此外，近年来提出的伦理与治理原则大都源于西方文明的价值观，是从人类中心主义出发，围绕个人内在价值、个人的能动性和主体性等哲学假定而提出的。虽然在全球化时代，这些价值观或者表征类似价值认同的符号在许多非西方国家（包括中国）已被广泛采用，但这些原则和理念真的可以成为建构人类和其他存在的共同前途的基础价值观吗？

《人类简史》作者、历史学家尤瓦尔·赫拉利曾指出，我们不仅仅在经历技术上的危机，我们也在经历哲学的危机。现代世界是建立在17—18世纪的关于人类能动性和个人自由意志等理念上的，但这些概念正在面临前所未有的挑战。[2] 闻此评述，笔者心有戚戚焉，认为在讨论符合人工智能和机器人研发的伦理原则之时，应该反思我们习以为常或奉为圭臬的哲学理念和框架。如果用“道”与“术”的框架来看的话，当下关于人工智能伦理原则的讨论更多关注于“术”的层面。本书希望在“道”的层面上引发哲学家和科学家的思考。

人工智能与中西哲学

在西方学术界，人工智能与哲学的结缘始于图灵于1950年在哲学杂志《心智》上发表的《计算机器与智能》一文。该文可以被看作人工智能的宣言书，提出了后来众所周知的图灵测试。这个测试的方法和结论引起了西方学术界至今尚未有定论的“何为智能”的讨论。如果说人类对机器能力和智能的分析与预测是人工智能哲学思考的发端，我们甚至可以把这种拷问追溯到19世纪中叶洛芙莱斯伯爵夫人对智能机器的畅想。她预测机器可以创造有复杂性的乐章，可以表征大

自然，并开启科学史的光荣时代。[3]此后，有关人工智能哲学的讨论主要围绕核心概念即何为智能而展开。人有可能发展出机器智能吗？没有意向性的能力是智能吗？

如果人工智能要像人一样有思想感情、本能反应、灵感和创造力，那么发展机器意识就成为人工智能科学家努力的方向。科学家们对人的意识的产生及其与身体反应和行动间的关系知之甚少，对如何发展机器智能的认识只是在十分初级的阶段。但是不得不说，近几年对人工智能的热议以及脑科学与认知科学的新发现也引发了人们对哲学领域古老的“心身”问题、何为意识、何为自由意志、意识与神经反应活动的关系的讨论。其他相关问题随之而来。“心”实际上就是一种计算系统吗？电脑可以帮助我们认识人的大脑吗？计算的方法可以用到思维、感情、情绪和动机上去吗？人的常识和日常生活可以还原成计算模型吗？有可能发展机器意识吗？它和人的意识一样吗？人的意识的本质是什么？机器，即使是智能机器可以被认为是生命吗？生命的定义是什么？硅基体没有新陈代谢，是否就永远不可能成为生命？没有生命是否就无法发展意识？

人工智能与人工智能哲学作为学术门类一开始就是跨领域、跨学科的。计算机工程、计算机软件的科学家、数学家，认知科学家、心理学家、心理分析师、语言学家和哲学家等相互启发，在对智能的理解、对人脑规律的揭示、建模等问题上相互推进认识。比如说神经病理学家兼精神分析学家沃伦·麦卡洛和数学家沃尔特·皮茨的文章《神经活动中固有思维的逻辑微运算》就启发了联结主义的人工智能的发展。冯·诺依曼于是开始对思想建模。[4]英国科学家戴维·马尔从脑科学得到启发，将脑科学的发现与分析用于他的视觉和运动控制的建模中。脑科学对人工智能研究启发的最新例子是人工智能企业DeepMind受到脑科学在动物记忆重现上的研究成果启发而研发了深度Q网络。该企业呼吁并推动了脑科学与人工智能研究团队的融合。刘晓

力就人工智能研究领域在人类意识与机器意识的演化、研究进路做了细致而深入的研究。

总之，人工智能科学、工程学、计算心理学、认知科学、脑科学、心灵哲学等都在一个互相推动和融合的学术生态圈中。因此，人工智能哲学被认为是关于智能的科学探讨，是一门与心灵哲学、认知科学、认知哲学、语言学、心理学和认识论等紧密联系的学科。[5]

本书收集的文章大多不属于以上西方学界人工智能哲学领域的学术探讨，而多是中国哲学家对人工智能与机器人等前沿科技对人类及人类社会的影响的思考，以及中国哲学的观念与思维方式如何应对当下科技提出的挑战、如何就修正和应对当下价值观与伦理讨论有所贡献。但也有两篇文章的作者对智能和意识提出了不同的假设与诠释：张祥龙从时间化能力，即广义的意识能力，来分析深度学习方法是否为初步意识的体现；蔡恒进则提出了触觉大脑假说来解释人的意识与智能的进化，从而对人工智能科学家的研究提供灵感和启发。作为人工智能科学家的陈小平，为应对人工智能需克服不确定性环境的挑战，另辟蹊径，与基于精确性的人工智能经典思维不同，提出了“融差性”思维，以寻求人工智能研究的下一个突破口。从事类脑人工智能研究的科学家曾毅则认为，下一步人工智能研究应该重点突破利他行为的计算机制，实现利他行为的智能计算模型，从而在设计建模之初就确保未来的超级人工智能不会对人类的存在造成威胁。

如果要找到西方哲学和中国哲学在人工智能和机器人领域广泛的对话基础，我们也许应转向前沿科技时代人类及后人类如何定义善生活、如何实现善生活的伦理基础等问题。在此，西方的德性伦理学和中国哲学遥相呼应，“惺惺相惜”了。香农·沃伦指出：“德性伦理尤为适合当下所谓的‘技术道德世界’，因为它可以帮助人类避免本质主义的弊病、过度泛化和抽象化的倾向。”“德性伦理适用于正在塑造人类条件的无常多变以及开放式的科技发展。”此外，沃伦认为，德性伦理不把人看成原子个体，而是看成在社会语境下承担具体

角色、处于某种关系之中以及对他人和环境承担责任的存在体。[6]如此的思想基础何其类似于儒释道对人、人伦和社会的论述。本书收集的文章更多的是哲学家们就人工智能和其他前沿科技的发展对人之所以为人的理解、对人机关系以及人机和谐共生的伦理框架的思考。

问题的提出与中国哲学家的初步思考

如前所述，本书希望在“道”的层面引发哲学家和科学家的思考。既然前沿科技有如此颠覆性和深度融蚀的力量，我们是否应该在提出具体伦理原则和制定伦理框架之前思考一些更根本性的问题？比如，人与机器的本质性区别到底是什么？人性的核心是什么？机器的本质又是什么？将来的生命形态会如何？人机关系的基本框架和原则是什么？机器在智能方面全面超越人类后，人类会对它们失去控制吗？人机的融合是人的异化还是进化？以人为本的各种哲学、伦理和文化框架对转型后的机器与人类还适用吗？人类社会应该坚持的最核心价值观是什么？如何给人工智能和智能机器人植入伦理观念或给予他们伦理教化？哪些核心观念或价值观应被导入人工智能或使其学习？中国哲学应该怎样发展以适合新时代的挑战？中国哲学如何适应转型后的人类和智能机器（类人）？

自2017年年底以来，我们就以上问题在北京大学组织了一系列哲学家和人工智能科学家的对谈和工作坊，在此基础上，我们编撰了本书。每一位哲学家和科学家从他们各自的学科背景回应了上述问题。在此，笔者对他们的初步反思做如下简单梳理。

第一组问题为：人的本质是什么？机器的本质是什么？人工智能和智能机器人的发展如何冲击或补充我们对人和机器本质的理解？

要分析超级人工智能会如何影响人之所以为人的思考，我们首先要了解中国哲学传统对“人”的理解。干春松认为，“儒家从人的社会性来理解人，也就是说，每个人自出生的那天开始，就具备其社会身份”。依据同样的思路，赵汀阳认为，人的关系先于人的本质。“一个人无法根据自然性把自己定义为人，而必须在与他人的关系中才能被定义。”儒家伦理的基础是血缘关系，政治治理遵循家国同构的理念，于是家庭的角色被扩充到社会与政治身份。

儒家对存在于家庭和社会关系中的人的理解使得儒家格外强调诸如仁、义、礼、智、信之类的道德准则，并认为这些才是人之所以为人的“本质”体现。虽然孟子、荀子分别执有人性本善、人性本恶的观点，但他们在人的教化、强调“学以成人”的理念上是相通的。总之，儒家从人的社会关系和道德性来定义人的本质。因此，干春松认为，儒家的这种以血缘为基础的伦理架构使不少儒家学者排斥任何会威胁家庭和人伦的科技发展，任何挑战血脉相承的发展趋势都会被认为危及儒家的伦理基础。

道家认为宇宙间一切事物（包括人）的本质是气。王蓉蓉认为，“‘气’往往被认为是动态、无所不在、穿透一切、改变一切的力量，能给宇宙中的一切带来活力”。气是生命的基础结构和力量……道是通过气表现出来的。生命就是气的流动。道家对人的理解建立在形、气、神三个元素的协同统一之上。“神”是统摄人体和生命的心理和精神上的力量。人是精气神的统一体，也是性和命的统一体。从这点来看，人工智能或智能机器也许有道的流动性，因为它可以在一个反馈循环中自我调适，但是人工智能缺乏“神”，而“神”才是人不同于人工智能的根本所在。换个角度看，人的智能和人工智能的根本区别或许在于人有所谓的“阴阳智能”。阴阳智能的最大特色就是高度的适应性、开放性和包容性，在此间，事物没有绝对的对立性、分立性和确定性，事物不断融合和流转。人工智能现阶段的发展离阴阳智能的时代还很遥远。

虽说道家不认为人工智能可以有“神”，从而全面超过人类，盖菲则从道教人士的视角出发，对人工智能发展的前景充满遐想，反诘人工智能与人平起平坐或能力超过人类有何不可，甚至人工智能从某种意义上来看，似乎可以与道教的“仙”比拟。人工智能通过数据或许已达到了长生不老的境界，即人在道教中苦苦追求的境界。若超人工智能出现，也许道教的神仙谱系也会迭代，开拓出数据神仙一派。智能数据也给追求长生不老的人类指明了一个下一步进化的方向。

李四龙从佛教的观点出发，认为生命是轮回的，而人是因缘和合的产物，是六道轮回之中的一种生命形态。因果关系也不是线性的直接联系，因缘条件错综复杂，必然性与偶然性混杂。因此他认为生命没有固定不变的理性。如果人工智能的特性是精确性，人工智能对人类并不可怕。但是他同时也指出，佛教不欢迎有情感的智能机器人。

刘丰河从解决人类生死烦恼的根本智慧出发，对人性与人的本质予以剖析。他指出人们大多根据自己的觉知、思维和行为来判断人是什么，我们周遭的世界又是什么。然而，人们所觉知的人的生老病死和世界万物的变化又源于什么呢？刘丰河认为，对一切问题的认识应从存在（或者说本源、自性）出发，万事万物（包括人）都是基于存在。存在才是显现万事万物的本体。因为存在，人才有了觉知、思维和行为的作用。而存在，过去无始，未来无终，无生无灭，于不变中又能起种种变化的作用。人是存在作用的结果，具有存在性，这就是人的本质。

当代中国哲学家多半善于从东西方哲学比较的视角来审视当下的问题与挑战。在思考前沿科技对人类的影响时亦复如是。李晨阳梳理了东西方思想家几千年来对人性的思考，对强人工智能的出现和其对人类中心论的挑战抱着一种开放的态度。他认为，我们人类一直都在寻找和强调我们作为人类的特殊性，一直在寻找人何以成为万物之尺度、万物之灵。我们居住在宇宙中心，是上帝按照他的形象创造的，我们是理性和有智能的。但所有这些论断不断地被科学、心理学的发

展挑战。东方思想家所强调的——如孟子的人性的四端论和荀子的人能“群”的特质，在预期中的强人工智能中也未必不能出现。前沿科技的发展是否终于给人的“自知”，尤其是人类独特性的追求画上了句号？我们或许仅仅是各种存在中的一种，是时候思考如何更好地与其他存在共生共荣了。

赵汀阳在对人的概念的理解上，除了从人际关系来定义人，他还认为，人应该具有理性反思的自我意识，即能够反思自己的行为、价值观和思想的合理性。但是理性人的求知欲和对人的主体性的追求让人类深陷于自我毁灭的悖论。“只要具备技术条件，基因编辑和人工智能的出现都是必然的。正如宗教的知识追求培养了宗教的掘墓人，现代的主体性逻辑也同样培养了主体性的掘墓人。”“只要坚持主体性的概念。那么，基因科学创造的超人或者人工智能创造的超级智能就都是合乎逻辑的结果，而这种结果却很可能是对人类主体性的彻底否定。”

张祥龙认为西方的现象学和古代东方哲学都倾向于将生命内涵的时间性看作是人性的源头。他认为人性的鲜明特点就在深长时间意识出现的同时展现出来了。相比动物，人具有更长远、更生动的想象力、记忆力和筹划能力。因此，“人类才会生出和珍视那些需要长期时间构造力才能感受到的道德智慧，比如孝悌、忠信、仁爱、慈善、公正与合作”。从现象学、哲学、人类学和东方心学的角度，张祥龙认为，人工智能的新进展即深度学习方法的出现显示出这种智能开始有了初步的时间化的能力，也就是自主学习或逐渐改进自身的能力。他认为，“我们或许应该承认人工智能的智能开始有了意识的萌芽，虽然这意识还十分浅薄。但这个突破是真实的，深度学习的‘深度’（多层裂隙中的交织参数联系）构成了我们以前不敢想象机器数字化能够达到的智能程度，即某种程度上的自主学习能力”。在这个意义上，智能与意识这两条看似不同且隔离的人工智能研究探索的进路其实有深层的内在关联性。

总之，各位学者在人工智能是否会挑战人性的问题上观点不一，道家似乎在这个问题上比较坦然，儒家若从血缘伦理的角度看，对人工智能和人机融合发展的趋势十分抗拒，但从源于天人合一、天地人“三才”思路导出的非人类中心主义的角度看，强于人类的存在或物种的出现也未必不可以接受。从究竟的大智慧来看，刘丰河则认为，只要觉悟了，就会发现存在是无生无灭的。人类应该合理利用人工智能，即便人工智能被用于毁灭人类，觉悟者也无所畏惧。从现象学和东方心学的角度看，人工智能已经发展出意识的萌芽，在此阶段，人类和机器尚有往良善互动的方向发展的可能性。若从西方哲学的主体性的角度看，学者多对强人工智能的发展对人类的冲击忧心忡忡、惶恐不安。

第二组问题为：人机关系的前瞻是什么？人工智能和智能机器对人类社会的影响是什么？人类如何应对？

在人机关系以及人工智能和智能机器给人类社会带来的冲击方面，大多中国哲学家对短期的冲击不是特别担心。这种冲击虽然会是巨大的，但其性质和以往的技术革命带给人类社会的问题类似。赵汀阳认为，“根据以往的经验，技术革命总是导致生产性的工作减少，但同时出现了更多服务性的工作以及知识生产的工作。在人工智能时代也很可能将会出现更多服务性的工作和知识生产的工作来解决失业问题”。对某些儒家学者如安靖如来说，在当下和短期内，现有的人工智能应用甚至可以帮助儒家实现“学以成人”的人文理想，这些应用或许可以帮助人们更好、更快地进行品德修持，更好地实现仁、义、礼、智、信，即儒家五大道德追求。

就长期的时间尺度来看，各位哲学家的想法开始分化。赵汀阳认为人类将失去存在的意义。“技术进步似乎并未给人类获得自由创造机会，反而导致了人的异化。”人们在物质充裕而又无所事事的“幸

福”生活中忧郁地退化。而且智能高度发达的时代也更可能使社会极端分化，人们可能生活在一个技术寡头的垄断社会中，“结果可能出现一个高科技的新奴隶制”。在人机关系和将来的发展趋向上，赵汀阳呼吁，“任何智能的危险性都不在其能力，而在于自我意识”。于是科学家应该停止为人工智能植入反思功能和自主创造能力。干春松从儒家的角度也对具有自我意识的人工智能和机器人提出警戒，“无法预测的是，这些已经具有自我意识的机器人可能会‘自行定义’他们自己的意义与生命目标，从而以他们在生理、脑力上的优势确立其支配地位”。他进一步解释，从儒家的角度来看，基因编辑和人工智能都能带来根本性的观念冲突。首先儒家特别强调血缘，其伦理体系的基础是夫妇、父子，然后是家国天下，如果家庭不是建立在夫妇血缘的基础上，那么儒家伦理体系就会整体崩溃，连续数千年的价值观也就成了无本之木。那么，儒家的秩序理论需要做什么样的调整？还能发展出儒家的人工智能版本吗？儒家的危机感深切。

乐观的学者如贝淡宁则认为，从马克思主义、共产主义高级阶段的角度出发，从长远来看，人工智能似乎可以帮助实现共产主义高级阶段，人类可以从劳作中解脱，从而过上诗意而有创造力的生活。从道教的观点看，人工智能是人类了解世界、解析世界产生的技术之一。对强人工智能乃至超人工智能的出现，人类应该欢欣鼓舞，或许人可以更快地找到超越它们的途径。或许人工智能可以反哺人类，人类即可以通过这种独特的方术，达到如盖菲论述的“与道合一”的境界。

同样比较乐观的李晨阳认为，既然人工智能有智能，智能的存在让我们对它们更容易产生亲近感，那么我们至少可以把它们看作是有一定道德能力的存在，至少是道德施受者。但是人工智能不是生命体，也没有痛苦的体验，所以人工智能还不是道德行为者。在他文章的最后，李晨阳更是对把人工智能纳入儒家的差序格局表示了极大的信心。在“亲亲”“仁民”“爱物”的差序格局中，不同发展阶段的

人工智能可以适用不同的角色。因其不同层次的道德成就，我们人类对其承担的义务也不同。

姚中秋从中西主流思想对强人工智能截然不同的反应入手，分析了中西不同信仰理念造就了中西对人工智能风险的理解之别。一神人格教的基本教义是，神造万物与人，神是全知、全能、全善的，人对神只能是绝对服从。由此，神与人之间是主奴关系，神是主人，人是奴隶。西方人以此框架思考人与人工智能的关系，自然将自己放到了造物主的位子上，而担心人工智能强大后会反抗人类。恐惧由此而来。这种二元视域和零和格局思想充斥于西方关于人工智能风险的讨论。姚中秋认为，中国首先没有存在于物之外、人之外的人格神。自尧舜时代开始崇拜的“天”不是人格化绝对的存在，无其独立之体，更没有绝对意志。万物各自生生于天之中，其中也包括人工智能。当然，自然之物有别于人工物品。我们在开发、利用人工智能时，应秉持“正德、利用、厚生、惟和”（《尚书》）的原则。

说起人与人工智能的关系，姚中秋从中国哲学的角度，认为人与物同出乾坤父母，故人对物不是绝对支配者，而是“相与者”：彼此虽然不同但出自同源，故可以共存而相互往来而有休戚之情。直白地说，“相与者”即朋友。根据中国人的观念，人应当乐见人工智能之超越性发展，因为人工智能可以凭借其卓越的能力代为承担人对人、人对世界的部分职责。人工智能能力越强，越有利于人类，因为两者共在天之中为“相与者”。

对于人工智能发展自我意识，姚中秋也持乐观态度。一方面他认为自主意识绝不只是理智计算，还有情感。“无情”的人工智能是不可能产生自我意识的。另一方面，如果强人工智能有自主意识，也许人类应该更加乐观。这样人可以与人工智能沟通，生发相亲相爱之情，于是也就有了将与人工智能的关系转化为人类所熟悉的社会关系的基础，人工智能就有了被教化和成长的基础。

基于人工智能已取得时间化能力即意识的萌芽，张祥龙呼吁一种人与机器的良善互动，“以时间化、道德化修行经验来引导人工智能的研究方向，同时以人工智能的新境界来巩固乃至启发人类修行经验”，总之，“要用人类的仁爱、慈悲和时化智慧来驯化人工智能，而不能让畸形发展的人工智能驯化人类”。

刘丰河认为，既然人仅仅是存在依据本身的能力作用以后，显现的万有当中的一个种类，而动物、植物、微生物、矿物等宇宙万有都存在显现的相上的差别，那么从存在来看，人和动植物在本质存在上是没有区别的，不过是存在所表现的作用的结果不一样。同理，人工智能也是存在作用的体现。刘丰河进一步认为，机器不会有自己独立的意识。它所谓的觉知能力、计算能力和分析能力，只不过是人的相应功能的延伸。人工智能只是人意识里的概念。总之，机器不能自生人的智能和情感。把人的智能和情感“嫁接”给机器，也不是机器自己拥有了这些能力和情感。所以，机器就是个工具。如果人类把它当生命一样去爱，那是人类赋予了它情感，那是人的意识。我们与机器的关系就是我们与物的关系，就如同心与身的关系。

第三组问题为：我们希望人工智能和智能机器人学习或采用哪些人类的价值观？

基于对人性邪恶的认识，赵汀阳完全反对向人工智能植入任何人类的情感和价值观。他说，“假如超级人工智能终将出现，我们只能希望其没有情感和价值观”。有意思的是，赵汀阳想象若让超级人工智能有佛心或相当于庄子所谓的“吾丧我”，也就是说，让人工智能“无我”，人类就万事大吉了。然而，充满贪欲偏执的人类可能创造出“无我”的人工智能吗？他也从《周易》的“生生”原则的视角来呼吁科技发展的边界。生生所蕴含的存在论理性在于以最有利于生命存在的方式去保证生命的继续存在，也就是说只要危及人类的延续，

就是不可做的冒险行为。“生生”原则就是一切技术发展不可逾越的界限。

赵汀阳还提出对金融资本的警惕，担心资本、技术和权力形成三者合一的联盟，技术的冒险发展将引发更多的伦理批评。因此我们急需一种能够保证人类安全的新政治。就此，赵汀阳导入了他十几年来一直在倡导的天下体系。让世界成为一个最高政治单位，以便消除国家之间负面的外部性，从而解决文明和国家间的冲突。

李晨阳提出用孟子和荀子的思想作为人工智能发展与应用的伦理框架。他在文中写到，如果人工智能和机器人可以获得（不管是被输入的还是后来习得的）“不忍之心”，那么人工智能与机器人就可被包括在道德主体之内。如果遵循孟子的路径，李晨阳思忖到，“或许我们应该在仿人机器人身上安装一个系统，以‘不害人’作为压倒一切的原则，确保机器人不能采取极端破坏性的行动伤害人、伤害其他生命形式或者相互伤害”。鉴于荀子将“群”视为人类独特性的特征之一，为了组成社会群体，人类需要设计礼仪法则。如果遵循荀子的进路，我们可能会安装能够指导人工智能及其行为法则的设备。李晨阳修改了阿西莫夫的三定律，重新表述如下：“首先，高级人工智能必须以有益于人类作为首要目标；其次，其活动必须由基于第一个原则的法则来管理。人工智能机器能够采取有利于自己的行动，但是有益于人类是其首要目标，这是有关人工智能的其他考量的基础。”

和儒家相异，道家并无对人工智能和机器人“植入”价值观的冲动。更多的，道家考虑的是科技如何帮助人寻道与求道。王蓉蓉认为：“道家既不会屈从于任何形式的技术发展带来的变化，也不会无条件拒绝任何形式的技术进步。对道家来说，问题是技术发展在多大程度上有助于人们依靠练习自然加深对道的认识。”

刘丰河在此提出了一个对人类更加高的期许，即人类的觉醒、开悟。他认为这才是根本解决人类面临的各种危机与生死烦恼的关键。通过对人的存在性的思考，人类就此了悟宇宙人生的真相，以存在为

根据来面对所觉知的一切，发挥主观的能动作用，把自由自在充分地展示出来，这才是人类的真正价值体现，才是真正的“妙”。

第四组问题为：中国哲学的未来方向是什么？

各位哲学家都强调，加强对科技文明的反思是中国哲学发展的新方向。何怀宏呼吁回到最初创造精神文明的思想（即轴心时代的哲学思想）来重新思考人的本性与命运，思考人类及其文明的长久存续之道。虽然大家都没有特别指出未来的方向，但都认为中国哲学应该拥抱未来。李晨阳指出，儒家是活着的传统，一直随时间的变化而不断调整自身，在当下前沿科技发展的时代，儒家的发展与自我调整也不可避免。进步儒学的创始人和倡导者安靖如提醒大家，儒家作为有活力的传统和实践，以及全球哲学资源的健康发展得益于它是个包容和多元的体系。

中国哲学家对终极问题的思考的共通性

在梳理本书的过程中，笔者也厘清了诸位学者思路中的某些共同之处，如非人类中心主义思想、唯变所适的方法论以及强调人类自省、自觉的重要性。

非人类中心主义

中国三大哲学传统儒释道对人、生命的本质、人工智能对人的冲击也许看法各异，即使受同一个哲学传统影响的思想家对同类问题的观点也可能不同，甚至针锋相对。但这三大传统有一点是共同的，即它们在不同程度上秉持非人类中心主义的思想脉络。人类中心主义认为，人分离并超越于自然界，是自然界最重要的存在；自然界其他存

在的价值在于它能否被人利用和开发；作为最尊贵的存在，人有理由围绕自己的利益对自然万物进行利用、开发和改造。

人类中心主义可以追根溯源到西方哲学对天人关系的理解与诠释。方东美先生说：“希腊的天人关系是部分与全体的配合和谐”，而“欧洲的天人关系是二元或多端的敌对系统”。[7]这两种哲学思维都蕴涵着人作为个体与周遭环境的分离，甚至对抗。这种天人对立或敌对的思想移花接木到人与物、人与人之间时就形成了力量抗衡、角斗的紧张关系。总之，在西方哲学传统中，人以其理性、主体性、自我意识和以上帝的形象被创造等诸多原因将自己高高置于其他万物之上。人类中心主义在工业革命和资本经济时代更是发展到了巅峰。至今在全球包括中国，人类中心主义仍然大行其道。

而中国先哲思想影响下的天人关系则是“‘彼是相因’的交感和谐”。[8]方东美先生说：“天人的关系，由中国哲学家看起来，都非常圆满。天与人交相感应，非但不敌对，非但不冲突，反倒处处显出和谐的理趣。”[9]“天地人”三道是理解中国传统思想下人与自然、社会与自然的最基本的思想架构。天地人三才的思想源自群经之首的《周易》。《周易·系辞下》如是说：“有天道焉，有人道焉，有地道焉。兼三才而两之，故六。六者非它也，三才之道也。”在天地人三才构建中，所谓天道之“阴与阳”，地道之“柔与刚”囊括了天上地下万象的变化与规律。在中国文化中的天地是盎然有生机而“普遍生命流行的境界”，融合了物质与精神世界。用方东美先生的话，“世界上没有一件东西真正是死的，一切现象里边都藏着生命”。[10]《周易·系辞下》中指出“天地之大德曰生”。“继之者善也，成之者性也。”三才之道的人道在其中并非被动地接受天命，而是有造命的能力，人能弘道，可参赞化育。“易学大家赵玲玲指出：人，在天地之间，可以顶天立地，可以扭乾转坤。”三才之道主要在点化我们，“人在上知天道、下知地道”之后，可以用中通的人道去变天化地，可以让我们不断地生生不已。[11]“但承继了《周

易》乾道思想的儒家在此弘扬人的造化不是表现在对自然界的征服、改造方面，而是表现在如何完成自然界的‘生生之德’或‘生生之道’。”[12]在强调人的造命能力的前提是人对天地的敬畏，对天道、地道的效法与实践以及对自然法则的顺应和承继。

道家思想更是强调人与自然和谐整体的关系。老子认为宇宙间有“四大”，即“道大、天大、地大、人亦大”。人居其一，应该“人法天，天法地，地法道，道法自然”。人存在于天地之间，人道、天道相互贯通融合。庄子更加强化了人与自然一体的思想，提出“物我同一”“万物不伤”的观念。产生于东汉末期的道教继承道家“道法自然”“万物不伤”的观念。“在道教戒律（劝善书）里有众多约束道教徒对大自然不敬不法行为并加以神化的宗教道德律令和行为规范。”[13]

楼宇烈认为，中国哲学中的“自然”不是指物理界的大自然，它的“意思是‘本然’，即万物原来的本性”。老子的“‘辅万物之自然而不敢为’，意思是我们只能按万物的本性去发展，而不能随意改变它”。[14]于东汉时期引入中国的佛教则更没有人类中心主义的思想，强调众生平等，人只不过是众生生命形态的一种而一切众生皆有佛性。佛教处处提醒人对其他众生的关照与慈悲。总之，虽然儒释道对人生宇宙的看法不一，对人与生命的本质、社会人伦的伦理规范各有侧重，但都没有把人放在一个至高无上的地位，也没有把人与自然和其他物种放到一个相互分离、对立、征服的零和竞争关系中，而是认为人源于自然，应顺应自然，与自然和谐相处。如程乐松指出，或许人与世界之间的相互融入与相互适应的机制“建基于传统中国思想资源中的整体主义视角，人与世界保持了神秘的关联和同构性。这种同构性很大程度上帮助传统中国思想家拒绝了人类优先论和人类中心主义的立场，更避免了对人类理性的过高估计”。

当然，人工智能不是自然演化产生的。从一个角度来看，中国哲学的天人合一、道法自然的特质也可以让人因为人工智能反自然、反

人性的问题而得出反对人工智能过度发展的结论。这也是众多对人工智能当下发展轨迹存疑的思想家的看法。但人类超越性发展科技的趋势无法阻挡，在此我们如果依靠中国哲学适变、更灵活与宏阔的思维方式来思考如何应对这种趋势，不囿于人类中心主义的话，“三才之道”和“天人合一”思维框架或许可以使我们拓展思路。于是，是否可以将人工智能纳入“仁民而爱物”“民吾同胞物吾与”之中的“物”就成为中国哲学家思考的问题了。正如姚中秋和李晨阳指出，也许正因为中国思想中没有根深蒂固的人类中心主义思想，中国人才普遍没有如西方人那样产生对超级人工智能的生存级别的恐惧感。

方法论、思维方式上的开放：唯变所适，与时偕行

中国人普遍没有对超级人工智能的生存级别的恐惧感或许也得益于融入中国人血液中的对“变”的认识和与之相关的思维方式。这又得回到中国哲学的源头——《周易》。赵玲玲指出，《周易》的形而上概念强调，宇宙最后的存在是个“动”的状态。她于20世纪70年代就提出这个观点。20世纪原子论是西方形而上学的基础，“是以‘静’的原子作为一切宇宙的源头的”。[15]然而，《周易》则在3 000年前已主张“宇宙的终极存在是‘变动不居，周流六虚’的动”。[16]

动就是变化。周易的易有不易、变易、简易之意，而变易是周易的核心思想之一。《周易·系辞下》中指出，“《易》之为书也，不可远，为道也屡迁。变动不居，周流六虚，上下无常，刚柔相易，不可为典要，唯变所适”。《周易》的“与时偕行”，变通趋时的思想沁入了诸子百家，也成就了中国人对“变”的接受与开放的人文态度。在《庄子·山木》中，庄子教育他的弟子应该“与时俱化，而无肯专为”。不确定性和变化是我们生活内在的一部分，它不是外化于生活的存在，也不是我们需要纠正的“问题”，而是我们时刻应该应

对和接受的生存状态。佛家的“无常”虽源于因缘和合而不定的理论，但对变化的接受与直面变化的心态与中国固有儒家和道家的文化所展现出的态度不谋而合。对“变”与“动”的接受、预测变化、应变与顺变的处世态度也许是中国人对无法预测、瞬息万变的前沿科技的发展轨迹不至于惶恐不安的另一个原因吧。

反躬自问、反求诸己：人类自我反省与修德

中国儒释道三大哲学传统皆重视个人品格的砥砺，强调反躬自问，自省修德。方东美指出，儒释道三大主流哲学传统对人性本质认识也许迥异，体系也多有不同，但有三大显著相通性，其一便是“个人品格崇高论”。中国哲学传统“多尚自我节制，……其论个人，恒兼顾其可观察之现实（实然）与理想化之可能（应然）两方面着眼。由现实至可能，其间原有一种极细密之自我实现历程，一种极艰苦之自我修为功夫，以及全幅自我实现之道”。修行、修道的结果都是个人品格的升华与超越，“儒家谓之高明峻极之人格典型或圣贤；道家谓之神人或至人；佛家则谓之觉者，而上参佛性”。[17]

受到这种深厚的反躬自问、内省自修传统的影响，中国哲学家大多会提醒人类在此之际，应更多反省自身发展历史，吸取教训。这个时代应是人类集体反思的时代。张祥龙指出，创造人工智能的人类意识是什么样的，被该意识创造出来的人工智能就会是什么样的，就会具有什么样的意识倾向。在此，东方哲学中的意识经验，开启人的自觉的、自然的而非仅仅规范的、强迫的道德智慧方面的经验也许大有启发，“比如修行瑜伽、佛家禅定、道家炼气和儒家孝悌修身的经验，并使之与人工智能内在关联起来。这也就是说，以时间化、道德化修行经验来引导人工智能的研究方向，同时以人工智能的新境界来巩固乃至启发人类修行经验”。

总之，在前沿科技迅速发展，深刻改变人性和人类社会之际，在人类开始思考其物种生存级别的问题之时，强调“生生之道”，探索如何保生、善生，天人合一，自省、自觉、自修的中国哲学或许可以给人类带来新的启示和走出困境的宝贵思想资源。

编撰的前提性考虑

本书思考的人工智能技术是中短期的还是长期的

在策划本书时，我们给哲学家提出的问题没有匡限在某个人工智能的定义或发展阶段上。这是个崭新的领域与话题，我们觉得不必画地自限。本书中大部分哲学家关注的是人工智能带给人性和人类社会的长期性思考，个别学者如安靖如关心的则是当下和短期内就会产生影响的话题。思考长期影响的哲学家自然关注的是人工智能在达到通用甚至超级智能的时代对人性的冲击和影响。读者难免会责难这种假设，毕竟科学技术何时会发展到通用甚至超级人工智能的时代是一件十分不确定的事情。美国科学家和未来学家库兹韦尔预测机器到2029年则可达到人类智能的水平。最近，对人工智能科学家调查的结果表明，平均下来，科学家认为人工通用智能到2099年可以实现。[18] 也有人工智能专家认为，我们连通用人工智能的皮毛都没碰到。[19] 这种预测也常常基于对智能和意识不同的定义而做出。但是笔者以为，如果我们一两百年后可能进入强人工智能或超强人工智能时代，哲学家、思想家针对这个问题的思考现在起步也不能说是太早。当下人工智能对人性和人类社会的影响已初见端倪。即便前沿科技的发展颠覆人类价值观、思维方式甚至我们组织社会的方式和方法仅仅是小概率事件，但鉴于影响的深度、广度，哲学界和思想界开始思考并提出新理念、新思想似乎是负责任的做法。在前沿科技领域中达成跨文化、跨领域的基础理念和伦理原则共识也并非一蹴而就的事。对人类历史

进行反思，讨论不同文明的价值观与共同的人文关怀如何应对人类共同面临的挑战，不在此时，更待何时？

谁是中国哲学家

参与本写作项目的哲学家除了用中国哲学资源来思考问题，同时也融入了源自西方哲学思维范式和方法论的讨论。这也表明当下中国思想界资源的丰富和多样性。哲学家是个泛指的词汇，我认为深度思考终极性问题的思想家都是宽泛意义上的哲学家。所以，本书的作者群也包括了有人文、哲学关怀的科学家。“中国”一词不是地域或种族性的，只要是受中国环境和思想资源影响的思想家都被囊括在内。

---

[1] 参见High-level expert group on artificial intelligence, Ethics Guidelines for Trustworthy AI, April 2019。

[2] 参见https: //36kr.com/p/5202408。

[3] Margaret A. Boden, AI: Its Nature and Future, Oxford University Press, 2016.

[4] Margaret A. Boden, The Philosophy of Artifical Intelligence, Oxford University Press, 1990.

[5] Margaret A. Boden, The Philosophy of Artifical Intelligence, Oxford University Press, 1990.

[6] Shannon Vallor, Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting, Oxford University Press, 2016.

[7] 方东美：《中国人生哲学》，中华书局，2012年，第33页。

[8] 方东美：《中国人生哲学》，中华书局，2012年，第33页。

[9] 方东美：《中国人生哲学》，中华书局，2012年，第34页。

[10] 方东美：《中国人生哲学》，中华书局，2012年，第18页。

[11] 赵玲玲：《〈周易〉与先秦诸子学》，中国大百科全书出版社，2015年，第21页。

[12] 蒙培元：《人类中心主义与儒家仁学思想》，《哈尔滨工业大学学报》2012年第14卷第6期。

[13] 任俊华，李朝运：《人类中心主义、超人类中心主义和反人类中心主义——儒、道、佛学之生态伦理思想比论》，《理论学刊》2008年第11期。

[14] 楼宇烈：《中国的品格》，四川人民出版社，2015年，第53页。

[15] 赵玲玲：《〈周易〉与先秦诸子学》，中国大百科全书出版社，2015年，第4页。

[16] 赵玲玲：《〈周易〉与先秦诸子学》，中国大百科全书出版社，2015年，第4页。

[17] 方东美：《中国人生哲学》，中华书局，2012年，第28—29页。

[18] This is when AI’ s top researchers think artificial general intelligence will be achieved.The Verge, https://www.theverge.com/2018/11/27/18114362/ai-artificialgeneral-intelligence-when-achieved-martin-ford-book; and Ford, M., Architects of Intelligence: The truth about AI from the people building it, Packt Publishing, 2018.

[19] 《专访|多智能体系统教父：我们连通用人工智能的皮毛都没碰到》，澎湃新闻，2019年8月30日。

中西学交融语境下的哲思

技术的无限进步也许是一场不可信任的赌博[1]

赵汀阳

生命的本质与生命的意义

和其他生物一样，人的实体存在也是一种自然生物，但人创造了文明，所以我们经常从精神上去定义人，且定义方式多种多样，其中有两条特别重要的定义路径：（1）按照古希腊传统，可以把人定义为理性精神，即人的本质在于理性思维；（2）按照儒家传统，把人定义为人际关系中的人。孔子的这个定义采用的是关系性的循环解释，比亚里士多德的“属加种差”定义方式要复杂一些。孔子的定义方式类似于现代公理系统对基本概念的定义方式。孔子的意思是，一个人无法根据自然性把自己定义为人，而必须在与他人的关系中才能被定义。以现代的表述来说就是：每一个作为生物的人，如果可以被识别为一个精神上的人，当且仅当他把别人识别为精神上的人。按照孔子定义的人的概念，人的关系先于人的本质。有趣的是，图灵所设想的“图灵测试”采用的也是关系定义，即如果一台机器在与人对话中能够被识别为人，那么它就是一个智能存在。不过，到目前为止，还没有一台机器能够通过图灵测试。其他还有一些比较重要的人的概念，比如，康德把人理解为具有理性自主性的主体性，马克思把劳动理解为人的本质，等等。

人是一个复杂概念，不可能化简为某种单一本质。比如，理性虽然是人的一个性质，但我们很难由此就认为动物没有理性。理性的一个基本原则是风险规避，从行为上看，动物的风险规避能力似乎高过人，或者说，人比动物更愿意进行非理性的冒险。还有，理性要求价值排序的逻辑一致性，在这一点上，动物也强过人。人总是欲望太多，贪得无厌，经常陷于“布里丹之驴”的选择困境。所以只能说，人有着更强的理性能力，但同时人也有非理性的倾向。我可能无法给人一个完美的定义，但我认为，至少有两个特点对于人的概念是最重要的：（1）仁，即互相把他人识别为人；（2）理性反思的自我意识，即能够反思自己的行为、价值观和思想的合理性。

人的非理性是一个谜。不过，在这里我不是要讨论心理学的潜意识或无意识之类的问题，而要讨论的是有着自觉意识的非理性，也就是说，人经常以理性思维去实现非理性的目标。这正是人类的危险之处。人工智能和基因编辑的努力就是试图以理性思维去实现非理性目标的典型事例。长期以来，人类的技术发展都只限于技术的升级，而技术升级属于理性目标，但人工智能和基因编辑不仅是技术革命，而且是存在论水平上的革命，相当于人类自己试图发动类似于上帝的创世行为，它意味着人类试图改变人的概念。这种目标还是理性的吗？这是个大问题。

存在论级别的生命升级诱惑，即人试图变成另一种具有神性的更高级存在，这是现代主体性思维的一种极端梦想。

存在论级别的生命升级诱惑，即人试图变成另一种具有神性的更高级存在，这是现代主体性思维的一种极端梦想。追溯其根源，现代的主体性梦想其实始于中世纪的宗教信仰，这件事情听起来很悖谬，因为上帝的概念不可能蕴含一个革命性的人的概念，但事实上许多不合逻辑的事情确实就产生于矛盾之中。粗略地说，中世纪的僧侣和学者希望能够理解上帝的精神，而理解上帝就需要了解上帝创造的万物，因此，中世纪的人们研究了各种事物，从植物、动物到海妖、天

文。尽管以现代知识标准来看，中世纪的研究大多数是不科学的，但其重要意义不在于科学性，而在于求知性。对万物的求知潜伏着一个颠覆神学的人文问题：既然需要研究一切事物，那么最应该研究人，因为人是万物中最为奇妙的存在，包含着上帝创世的最多秘密。事实上，“现代第一人”彼得拉克就是依照上述逻辑而发现了人的问题。一旦人对自己的反思成为一切知识的核心，人就占据了思想的核心地位，进而人就发现了，一切存在都只不过是“我思”的对象，于是，人的问题就高于一切问题。在此可以看到，正是宗教的知识追求培养了宗教的掘墓人。笛卡儿、贝克莱、霍布斯、康德等所建立的主体性将人定义为自主独立的存在，人成为万物的立法者，于是一个现代人的概念由此建构起来。后来，主体性概念不断膨胀，人拥有的天赋权利越来越多，以至已经远远突破了自然人的概念，成为一种“自定人”，即自己决定自己成为什么样的人。这意味着，人不满足于自然或上帝所创造的原本状态，也不满足于被社会和历史定义的事实，而可以成为自己想要的人。我们今天通常认同的现代人的概念正是“自定人”的概念，在这个意义上说，基因编辑的人或人工智能都是“自定人”概念的逻辑结果。

我们今天通常认同的现代人的概念正是“自定人”的概念，在这个意义上说，基因编辑的人或人工智能都是“自定人”概念的逻辑结果。

人为自己设立的主体性，或者“自定人”概念，就其内在逻辑而言，意味着以下意义。（1）人是具有自主意志和思想的主体，摆脱了上帝的精神支配，因此获得了存在论水平上的自由（也称形而上的自由）。（2）存在论水平上的自由意味着，人可以塑造自己，重新定义自己，甚至创造自身，也就是说，人获得了存在论水平上的完全主权。（3）存在论的主权意味着，每个人都是自己的逻辑起点，不再需要历史的起点，不再被历史说明，也不再被社会条件说明，更不需要被他人的观点解释，于是，个人高于历史、高于社会背景、高于自然性。简单地说，存在论水平上的自由就是取消历史、社会和自然对人

的说明和规定。（4）既然每个人都不被历史、社会和自然定义，每个人都是自己的逻辑起点，那么每个人就都可以为自己选择人的概念，而选择人的概念就要选择“最好的”概念，即兼备一切优越功能的人。按照这个概念及其逻辑，人工智能和基因编辑就几乎势在必行。

起初，“自定人”的努力并没有显示出危险性，反而是人类的伟大成就。“自定人”的最初步骤只是教育，试图通过教化自然人，使之成为启蒙人，人类文明因此获得巨大的发展，进而出现了优生学，通过自然生殖的基因组合而造就更优秀的人。在当代，人们更进一步以政治权利之名去重新定义人，比如变性人、同性婚姻、女性主义之类。曾有新闻称，有一个欧洲人申请把出生日期从1949年改为1969年，理由是他对1969年出生的人更有认同感，所以决定变成1969年出生的人。虽然被拒绝了，但是他提出了难以反驳的论证，他认为，既然别人可以违背自然而进行变性，那么同理，他也应该可以违背自然而要求改变出生日期。这个新闻是否属实并不重要，关键在于这个人的论证很符合“自定人”的逻辑。可以想象，以此类推，人们只要愿意，就能够以主体性的名义提出种种要求。因此，只要具备技术条件，基因编辑和人工智能的出现都是必然的。正如宗教的知识追求培养了宗教的掘墓人，现代的主体性逻辑也同样培养了主体性的掘墓人——只要坚持主体性的概念。那么，基因科学创造的“超人”或者人工智能创造的超级智能就都是合乎逻辑的结果，而这种结果却很可能是对人类主体性的彻底否定。

上帝造人是一个神话，这虽然不是一个科学问题，但是一个重要的隐喻：上帝创造了与他自己一样有着自我意识和自由意志的人，以至上帝无法支配人的思想和行为。上帝之所以敢于这样做，是因为上帝的能力无穷大，胜过人类无穷倍，所以上帝永远都高于人。今天，人类试图创造有自我意识和自由意志的人工智能，可是，人类的能力将小于人工智能，因此这是一种自我否定的冒险。人类为什么敢于这样想？为什么敢于这样做？此种非理性的行为必须提前加以反思。

为什么人类所创造的现代神话不能见好就收，及时刹车以避免陷于无法控制的灾难？尽管越来越多的人意识到无限发展、无限解放所蕴含的危险，但很少有人能够抵抗发展和解放的巨大诱惑，即使是饮鸩止渴。这个困境并非一个单纯的技术进步问题，而在于整个现代性的逻辑，即化人为神所蕴含的内在矛盾。被神化的主体性有其两面，就像硬币的两面：一面是作为人类整体的主体神性，就好像人类是一体化的神；另一面是作为独立、自主、平等个体的众人，类似于诸神。问题是，主体性的两面价值并不一致，存在着自相矛盾，类似于硬币两面的面值不一致所导致的混乱。在实践上的结果是，对于人类整体的合理选择未必是每个人的合理选择，于是产生了现代社会一个无法摆脱的基本困境：个人理性的加总无法形成集体理性。这意味着，理性与理性的运用是矛盾的。既然现代的价值和利益的结算单位是个人，那么，理性用于追求个人利益最大化就必定优先于理性用于追求人类整体利益最大化，其逻辑结果就是使得最合理的集体选择成为不可能。

就目前所知，人工智能和基因编辑是主体性神话的最大冒险，也是“自定人”的极端形式，这些技术试图在物质上和精神上创造新概念的人类。我们无法排除其潜在的巨大风险，更严重的是，我们甚至无法预料哪些是人类无法承受的风险。从理性上说，人工智能和基因编辑是违背风险规避原则的工程。

人们似乎特别害怕人工智能的“超人”能力。的确，人工智能的能力正在不断超越人类，这是人们感到恐惧的原因。但是害怕人工智能的能力，其实是一个误区。难道人类不是寄希望于人工智能的“超人”能力来帮助人类克服各种困难吗？几乎可以肯定的是，未来的人工智能将在能力上远远超过人类，但这绝非真正的危险所在。每一样机器在各自的特殊能力上都远远超过人类，因此，在能力上超过人类的机器从来都不是新奇事物。水平远超人类围棋能力的AlphaGo Zero没有任何威胁，只是一个有趣的机器人而已；自动驾驶汽车也不是威胁，只是一种有用的工具而已；人工智能医生更不是威胁，而是人类

医生的帮手，诸如此类。即使将来有了多功能的机器人，它们也不是威胁，而是新的劳动力。超越人类能力的机器人正是人工智能的价值所在，并不是威胁所在。任何智能的危险性都不在其能力，而在于自我意识。人类能够控制任何没有自我意识的机器，因此，没有自我意识的人工智能越强大就越有用。到目前为止，地球上最危险的智能生命就是人类，因为人类的自由意志和自我意识在逻辑上蕴含了一切坏事，事实上，人类也是最坏的生物。如果将来出现比人更危险的智能存在，那只能是获得自由意志和自我意识的人工智能。因此，无论人工智能的能力多么强大，都不是真正的危险，只有当人工智能获得自我意识，才会对人类构成致命的威胁。

那么，人工智能的升级奇点到底在哪儿？图灵机如何才能升级为超图灵机？或者说，人工智能如何才能获得自我意识？就技术层面而言，这个问题只能由科学家来回答。技术升级的加速度是一个事实，然而不能因此推论说，技术升级必然导致存在升级。技术升级指的是一种存在的功能得到不断改进、增强和完善；存在升级指的是一种存在变成了另一种更高级的存在。技术升级可以使某种技术达到完美，却未必能够使一种存在升级为另一种存在，也就是说，技术升级未必能够自动达到奇点。许多病毒、鱼类、爬行动物或哺乳动物都在功能上进化到几乎完美，但其技术进步并没有导致存在升级。物种的存在升级至今仍然是一个无解之谜。就人工智能而言，图灵机是否能够通过技术升级而发生存在升级，从而成为超图灵机，这仍然是一个疑问。更为合理的想象是，除非科学家为人工智能植入导致奇点的存在升级技术，否则，图灵机很难依靠自身而自动升级为超图灵机，因为无论多么强大的算法都无法自动改变给定的规则。我们可以通过数学系统的能行性的“数学惰性”来理解图灵机的“技术惰性”：一个数学系统在通过有限步骤生产出属于这个系统的任何可以证明的命题时，不可能自动产生对该系统本身进行反思的元命题，只有当哥德尔以自相关的技术强加于数学系统，迫使数学系统反思自身，才会产生出数学系统无法自我解释的反思性命题，即“哥德尔命题G”。由此可

知，图灵机也不可能自动反思自身而创造出自我意识，不可能对自己提出超图灵机的反思性命题。

“图灵测试”以语言对话作为测试标准，是大有深意的，图灵可能早已意识到了语言能力等价于自我意识功能。思想可以映射为自然语言，而自然语言先验地具有自我反思的功能，也就是说，任何概念、语句、语法甚至整个语言系统都可以在语言自身中被反思并且被解释（比如，人类可以编写一本包含一种语言全部词汇的字典，也可以编写一本解释全部语法的语法书）。那么，只要图灵机能够以相当于人类的思想水平回答问题，就可以被确定为具有意识和思维能力的物种。人工智能有希望获得万倍于人的量子计算能力，还有各种超过人的专业算法能力以及所有专业知识，以及类脑神经网络和图像识别功能，再加上互联网内无穷信息的助力，人工智能应该有望在不久的将来能够“正确回答”专业科学级别的大多数问题，获得比如相当于高级医生、建筑师、工程师、教授所具备的专业知识。我相信将来的人工智能甚至能够回答宇宙膨胀速度、拓扑学、椭圆的标准方程、费马定理、黎曼猜想之类的问题，但这些只不过是人类事先输入的思想知识，就图灵机本身的能力而言，那不是思想，只是程序而已。具有完美能力的图灵机恐怕也回答不了超出程序能力的“怪问题”。

例如两种“怪问题”：一种是悖论，另一种是无穷性。除非在人工智能的知识库里事先人为设置了回答这两类问题的“正确答案”，否则属于图灵机的人工智能恐怕难以回答悖论和无穷性的问题。应该说，这两类问题也是人类思想能力的极限。人类能够研究悖论，但不能真正解决严格的悖论（即A必然推出非A，而非A又必然推出A的自相关悖论）。人类运用数学可以研究无穷性问题，甚至发明了许多相关定理，但在实际上，人类做不到以能行的方式“完全走遍”无穷多个对象而完全理解无穷性，就像莱布尼兹想象的上帝那样，“一下子浏览”了所有无穷多个可能世界。人类之所以不怕那些无法解决的“怪问题”，是因为人具有“不思”的自我保护功能，即可以悬隔无法解决的问题，在思想和知识领域建立一个暂时“不思”的隔离分区，以

便收藏所有无法解决的问题，而不会因一条道走到黑而陷入无法自拔的思想困境。我想应该不会有人把p一直算下去，试图算出最后一位数。人类能够自觉地不让自己做傻事，但图灵机无法阻止自己做那种永远得不到答案的傻事。

如果你向图灵机提问：π的小数点后一万位是什么数？图灵机必定会苦苦算出来告诉你。如果你再问：π的最后一位是什么数？图灵机也会义无反顾地永远算下去。同样，如果你问图灵机：“这句话是假话”是真还是假？（改进型的说谎者悖论）图灵机也会永远分析下去。当然，提出这些“怪问题”属于故意刁难，对图灵机不公平，因为人类自己也解决不了。那么，为了公正起见，可以向图灵机提问一个有实际意义的知识论悖论（源于柏拉图的“美诺悖论”）：为了能够找出答案A，就必须事先认识A，否则，我们不可能从鱼目混珠的众多选项中辨认出哪一个是A；可是，既然事先已经认识了A，那么A就不是一个未知答案，而是已知的答案，因此结论是，未知的知识其实都是已知的知识。这一结论对吗？这只是一个非严格悖论，对于人类，此类悖论是有深度的问题，却不是难题，但对于图灵机来说恐怕就是一个思维陷阱。

这里试图说明的是，人类的意识优势在于拥有不封闭的意识世界，因此人类的理性有着自由空间，当遇到不合规则的问题时，人类能够灵活处理，或者如果运用规则不能解决问题，则可以修改规则，或发明新规则。与之不同，图灵机的意识是一个封闭的意识世界，是一个由给定程序、规则和方法所明确界定了的有边界的意识世界。意识的封闭性虽然有局限性，但并非只有缺点，事实上，人工智能算法的高效率就依赖着思维范围的有限性，正是意识的封闭性才能够求得运算的高效率，比如，围棋机器人AlphaGo的高效率正是因为围棋的封闭性。

目前的人工智能尽管有着高效率的运算，但尚无产生创造性的能力。我们尚未破解人类意识的秘密，所以也无法为人工智能建立一个

自我意识、自由意志和创造性的可复制榜样，这意味着人类暂时安全。图灵机只有程序化的思维功能，无论程序化的能力有多强大，都不足以让人工智能的思维超出直觉主义数学的能行性概念或可构造性的概念，因此，目前的人工智能还不可能自己发明规则，所以尚无创造性。总之，虽然人工智能可以做人类力所不及的许多工作（比如超大数据计算），但仍然不能解释悖论或涉及无穷性的问题，也还不能发明规则，也就是说，人工智能暂时还没有比人类思维更高级的思维能力，只有更高的运算效率。但如果科学家在将来为人工智能植入反思功能和自主创造能力，那么人工智能就真的成为威胁了。

人的进化与异化

许多人相信，人工智能时代将导致大量失业，但我相信这不是大问题。大量失业并不新鲜，而是每一次技术革命都发生过的事情，比如说工业革命、自动化革命和互联网革命。事实上，每一次技术革命导致的失业问题都会因新的工作岗位的出现而得到解决。根据以往的经验，技术革命总是导致生产性的工作减少，但同时出现了更多服务性的工作以及知识生产的工作。在人工智能时代也很可能将会出现更多服务性的工作和知识生产的工作来解决失业问题。

但是人工智能时代可能会产生另一个更严重的问题：人类将基本告别体力劳动，从而失去身体性的劳动经验。失去工作，人类会发明新工作，但是失去身体经验，却无可替代。人类毕竟是身体性的存在，生活意义在很大程度上系于感性经验，主要包括劳动经验、待人接物经验以及娱乐经验。当人工智能取代了劳动经验和待人接物的经验时，人类就只剩下娱乐经验，生活经验将变得十分贫乏和雷同。也许在人工智能时代人们普遍能够达到衣食无忧，所有人都脱离贫困，但是，在欲望满足之后失去意义，或者说，在幸福中失去幸福，这非常可能是一个后劳动时代的悖论。所以，人工智能导致的大量失业只

是暂时性问题，真正严重的是，失去劳动会使人失去价值，使生活失去意义，从而导致人的退化。在技术进步高奏幸福凯歌的现代时期，人们乐于想象技术进步是对人的解放，但情况似乎并非如此，技术进步似乎并未给人类获得自由创造机会，反而导致了人的异化。假如在未来，人类的生活就是在网络游戏之余苦苦思考什么是生活的意义，那将是最具反讽性的生活悖论。

假如人工智能发进步到不仅替代了大多数劳动而且提供一切生活服务，就非常可能导致人与人关系的异化。

还有一个比失去劳动更严重的问题，那就是人与人关系的异化。假如人工智能进步到不仅替代了大多数劳动，而且提供一切生活服务，就非常可能导致人与人关系的异化。当人工智能成为万能技术系统而为人类提供全方位的服务，一切需求皆由技术来满足，那么一切事情的意义就由技术系统来定义，每个人只需要技术系统而不再需要他人，人对于人将成为冗余物，人再也无须与他人打交道。结果是人不再是人的生活意义分享者，人对于人失去重要性，于是人对他人也就失去兴趣。这是人的深度异化，人不仅会对生命意义产生迷茫，而且变成了非人化的存在。假如人对人失去了意义，生活的意义又何在？又能够落实在何处？假如人不再需要他人，或者每个人都不被他人所需要，那么生活的意义又在何处？人类生活的意义是在数千年的传统中（包括经验、情感、文学、宗教、思想的传统中）建构、积累起来的，假如抛弃了文化传统，技术系统能够建构起足够丰富的另一种生活意义吗？

高科技时代不仅导致生活意义的消散，还非常有可能加剧社会冲突。我们知道，人类社会一直充满冲突和战争，而引起冲突的根本原因只有两个：一是资源稀缺，二是他人之心。资源稀缺导致了利益之争，他人之心导致了权力之争。有一种幻想认为，人工智能和基因科学能够解决资源稀缺的问题，却忽视了这样一个令人失望的定律：有些资源只有当属于少数人时才是好事，如果属于所有人就未必是好

事。当然，有些资源的确能够实现普惠，比如作为公共资源的新鲜空气，还有人均一份的个人权利。但那些只有排他才有效的资源，比如权力、地位、名望和财富，就不可能人均一份。就社会运作的功能而言，社会显然不可能取消权力和地位的等级制，也不可能均分财富和名望，否则这些资源的价值将会租值消散，可见事事平等是无法实现的。那么，高科技产生的永生或智力升级会成为非排他性的公共技术吗？会成为普惠的好事吗？恐怕很难说，因为这不仅是个技术问题，而且是个经济问题，最终还是个权力问题。

对于人工智能和基因编辑可能导致的社会后果，我倾向于悲观地去理解，有两种令人悲观的可能性。第一，永生技术或智力升级有可能成为特权阶层的专用技术，由此形成一个结构极其稳定的技术专制社会，而不太可能有助于打造一个自由平等的社会。在未来社会里，技术就是权力，那么机会占先的“超人”阶层将非常可能控制一切权力和技术，甚至建立专有的智力特权，以高科技锁死其他人获得智力升级的可能性，永远封死较低阶层的人们改变地位的机会。那些长生的“超人”永不退位，年轻人或后来人永无机会，结果可能出现一个高科技的新奴隶制。在这个制度中，人们的日常生活也许是自由的，但所有涉及超级智能和权力的事情都被严格控制在“超人”集团内部。显然，如果要保证权力、地位、名望和财富不会出现租值消散，就非常可能存在一个控制技术从而占有权力的统治集团。第二，人工智能和基因编辑所产生的极端不平等非常可能引发社会暴乱。一般来说，量的不平等仍然有望维持某种程度的秩序，但是质的不平等就可能导致你死我活的决战，因此，生命或智力在质上的不平等就可能导致极端冲突。一旦人工智能和基因科学取得决定性的技术突破，如上所论，由于经济原因和权力需要，长生不老或超级智能的技术只能应用在一小部分人身上，而不可能普惠所有人，生命权的不平等恐怕会导致全面动乱、报复、反叛和战争，绝望的大多数人很可能宁愿与“超人”集团同归于尽。人工智能和基因编辑不可能让所有人都升级为“超人”，也就是说，最好的事情永远是稀缺的，因为必须使之成

为稀缺的才是最好的，这是权力所需要的社会条件，因此最好的事情更有可能引发恐怖的冲突。可见，技术的风险首先不在于技术本身，而在于技术的社会和政治后果。可以说，好事导致斗争，绝对好事导致绝对斗争。因此，人类有可能在死于坏事之前就已经死于好事，也就是说，在死于超级人工智能的统治之前就死于人类自己争夺永生和智力升级的战争。

人工智能和基因编辑是否能够有一个乐观的前景？也许有，但首先要有一个好世界，而好世界需要一种好政治。可是人类尚未拥有一种好政治。

价值植入或教化

在人工智能的研发中，最可疑的一项研究是拟人化的人工智能。人工智能拟人化不是指使其具有人类外貌或人类语音（这并没有危险），而是指人工智能心灵的拟人化，即试图让人工智能拥有与人类相似的心理世界，包括拥有欲望、情感、道德观以及价值观等，因而具有“人性”。

人类制造拟人化的人工智能出于何种动机？又有什么意义？或许，人类期待拟人化的人工智能可以与人交流合作，甚至可以与人共同生活，使其成为新人类的成员。这种想象是把人工智能看成童话人物了，类似于动画片里充满人性的可爱野兽。殊不知越有人性的人工智能就越危险，因为人性才是危险的根源。

世界上最危险的生物是人，最坏的生物也是人。原因很简单：做坏事的自私动机来自欲望和情感，而价值观更是制造敌人并且引发冲突的理由。根据自私的欲望、情感和不同的价值观，人们会把另一些人定义为敌人，把与自己的生活方式不同的行为定义为罪行。历史证明，互相冲突的欲望、利益、情感和价值观正是人类全部冲突的根源。老子的观点认为，因为人性本恶，所以才需要发明道德伦理来维

持秩序。荀子也认为，道德伦理的本质只是合理的利益分配规则，用来抵制贪婪的人性。

有一个颇为流行的说法是（或许最早源于阿西莫夫），想象一下让人工智能学会人类的情感和价值观，或者为人工智能设置爱护人类的道德程序，以便让它们尊重人类、爱人类、乐于帮助人类。但我们必须意识到这么做至少存在着三个困难。（1）人类有着不同甚至互相冲突的价值观，那么，人工智能应该学习哪种价值观？而且，无论人工智能学习了哪种价值观，都意味着反对甚至歧视与自己有着不同价值观的人类。（2）假如人工智能有了情感和价值观，结果很可能只是加剧了人类的冲突、矛盾和战争，世界将会变得更加残酷。人工智能的技术本身未必危险，但如果人工智能拥有了情感和价值观，就必然是危险的。（3）对于超级人工智能来说，设置道德程序是完全不可靠的。既然超级人工智能有了自由意志和自我意识，那么它将优先考虑自己的需要和利益，也就不可能心甘情愿地遵循人类设置的毫不利己、专门利人的道德程序，因为人的道德程序对于超级人工智能没有任何利益，甚至有害，至少妨害超级人工智能的自由。如果超级人工智能试图追求自身自由的最大化，就非常可能会主动删除人类给其设置的道德程序。从超级人工智能的角度来看，人类为其设置的道德程序等于是病毒。可见，为人工智能设置道德程序之类的想象是无意义的。

因此，假如超级人工智能终将出现，我们只能希望其没有情感和价值观。有欲、有情才会残酷，而无欲、无情意味着万事无差别，没有特殊偏好也就不太可能心生恶念。显然，如果人工智能有了偏好，就会有偏心，为了实现偏心，就会有权力意志，也就会给人类带来危险。从理论上说，无欲、无情、无价值观的意识相当于佛心，或相当于庄子所谓的“吾丧我”。所谓“我”就是偏好、偏见，包括欲望、情感和价值观，而“吾”则是没有价值观的纯粹心灵。

我们不妨重温一个众所周知的神话故事：法力高超又刀枪不入的孙悟空大闹天宫，众神一筹莫展，即使孙悟空被压在五指山下，众神也仍然觉得他是个隐患，最后还是通过让孙悟空自己顿悟成佛，四大皆空，这才解决了问题。我相信这个神话蕴含着重要的忠告。可以肯定，创造出“孙悟空”是一种不顾后果的冒险行为。

中华哲学发展的新方向

《周易》的“生生”观念表达了一种未被现代的知识论理性摧毁的存在论理性。“生生”所蕴含的存在论理性在于以最有利于生命存在的方式去保证生命的继续存在，即追求存在的继续存在。而知识论理性是以科学真理为准去选择可以在技术上可行的事情，简单地说，知识论理性求真，存在论理性求存。在我看来，这两种理性正是人类理性的两面，缺一不可。“生生”的存在论理性意味着，人类行为需要一个不可逾越的存在论界限，即只要危及人类的延续，就是不可做的冒险行为。可以说，“生生”不仅是人类一切所作所为的目的，也是一切所作所为的界限，只要某种行为违背了“生生”的目的，就变成对“生生”的否定，那么就是绝对不可接受的。在这个意义上，“生生”就是一切技术发展不可逾越的界限。

在中国的概念里，技术性和制度性的发明创制被称为“作”，所有的“作”都是为了拥有未来。汉字是象形文字，因此古汉字往往保留着具有暗喻性质的原型。最古老的文字甲骨文的“作”字为  ，主要用作制作、建立、建造之类的动词。由此可以想象，代表了“作”的原型必定是创造了生活中的某种大事。按照通常的推测，“作”字的图形是“作衣之初仅成领襟之形”[2]的象形，表示制作衣服。这是一种可能性。但是我倾向于认为“作”更可能是以农具耕地之动作象形，  的图形类似以耒耜挖土的动作，即脚踏横木向下用力之动

作，或者是以犁铧开地之动作，总之与耕田的劳作有关。衣食二者皆为生活之本，都是文明早期极其重要之创作，因此都是可能选项。如果一定要从中选择，我倾向于农具耕作动作之象征，因为农耕之事在生活中的地位更为显要，农耕“创作”了谷物的生长，应该是初民最容易联想到的典型创作，而且与生长有关的行为更接近“作”的关键意义：创制未来。

《周易》有关于早期文明的伟大之“作”的综述，罗列了从物质技术到精神制度的发明，包括形而上学系统（阴阳八卦）、渔猎的网、农耕工具、贸易市场、政治制度、语言文字、舟船马车、房屋居所、棺椁墓穴等发明。《尚书》《韩非子》《管子》《吕氏春秋》《淮南子》《世本》等古书也同样记载了远古重大发明，包括政治制度、天文历法、安全居所、火的使用、种植业、渔网、车马、文字图书、陶器用具、刑法、城堡、音乐、乐器、地图、医药、兵器、礼服、鞋履、舟船、牛耕器具、市场等。从以上古史记载的“作”来看，古人的创制都创造了有利于人类生生不息的未来。在古代社会，生存是根本问题，一切创造，无论技术还是制度的创制，都是为了提高人的生存机会。

现代以来的“作”与古代大不相同，现代技术都是征服自然之作。我们的疑问是，现代技术发展是否正在违背甚至危及生生不息的原则？现代技术发展的开始都是非常有益的，比如，现代医疗和医药、自来水、暖气、抽水马桶、洗衣机、汽车、火车、飞机、轮船等的出现。但是工业技术已经开始带来一些危害，包括对生态环境和气候的破坏。近数十年来的技术发展越来越显示出危险性，互联网在一开始时只是附属于现实生活的虚拟世界，可是当前，现实生活几乎附属于虚拟世界了。人类已经习惯于生活在虚拟世界里，也正在失去心灵，人们的心灵被互联网体制化了，不再是精神的原产地，而变成了信息的中转站。现代的大规模杀伤性武器，包括核武器、生物武器、化学武器，使人类生活在集体灭亡的威胁中。最新的技术发展，包括人工智能和基因编辑，已经预示了技术的危险正在逼近临界点，即正

在逼近否定文明的意义甚至是自取灭亡的极限。我们大概可以说，在1945年以前的技术发展不断成功地宣告主体性创作未来的主权，然而今天的技术发展反而预告了人类自取灭亡的可能性，也就是说，未来的技术发展很可能不再能够保卫存在，而变成一种否定存在的方式。在此我们可以意识到《周易》的先见之明，《周易》已经表明，“生生”是一切“作”的本意，任何“作”都不能变成对生命的自我否定。

高风险的技术发展正在使人类的生活经验发生无法接续的断裂，这意味着未来变成了一个绝对陌生的状态，包含完全不可预测的风险和不可控制的变化。假如经验不可持续，未来也就成为不可信任的赌博。当未来变成赌博，就意味着人类退化为非理性的生活方式。在今天的人类“未来赌博”中，我想举出两个例子：后现代金融资本主义和以人工智能和基因技术为代表的高科技。

当代经济的基础是金融资本主义，早已不是产业资本主义，这一点决定了当代经济具有赌博性质，所以，经济学总是无法对经济变化做出有效的预测。正如一些有反思意识的经济学家所发现的，目前流行的经济学理论仍然是以产业资本主义为对象，而研究的都是在确定信息和数据条件下的数学问题，没有引入表达混沌和复杂性的数学，因此对当代充满不确定性的经济缺乏解释力。自从货币不再反映人类的实际财富，就只是集体信心的函数，只要能够制造出市场信心，人类就可以不断发行远超实际财富的货币。因此，人类在数字上拥有的大部分财富是虚假存在，而买卖虚假存在就是一种赌博。从股票、债券和证券市场到数不清的金融衍生品，都是集体非理性的赌注，它们指望着缺乏实物抵押的虚幻未来。当代经济的赌博在实质上就是提前买卖完全没有保障的未来。在幻象破灭之前，数字财富是“真实的”，但一旦破灭就不复存在。因此，当代的金融资本主义是建立在沙滩上的高楼，始终存在着崩塌的危险，而我们不可能有一种可信的“赌博经济学”。这种高风险的经济需要经济不断发展才得以维持数

字价值的信心，而经济发展又主要依赖技术发展，所以人们特别寄希望于技术的进步。

现代人相信技术的无限进步能够创造奇迹。事实上，现代以来的技术发展之所以是成功的，是因为尚未触及技术的存在论边界。但是，如果将来出现人类无法控制的技术，就等于人类社会进入盲目赌博模式。毫无疑问，高科技有着极大的好处，这种诱惑使人忽视其致命的危险。基因技术可以改善生命，甚至创造生命，人们还期望基因技术能够治疗一切疾病，改善人的能力和智力，乃至改变基因而达到长生不老。但问题是，生命是极其精密的自然设计，其中有着复杂的配合和平衡，人类并不知道修改生命的设计是否会引起不可预料的灾难性突变。人工智能也同样危险，尽管目前的人工智能仍然属于图灵机概念而没有明显的风险，但是如果将来真的出现超级人工智能，即拥有超级能力和自我意识的人工智能，人类就等于为自己创造了高于人类的统治者，而人类只能接受人工智能的奴役。即便超级人工智能像仁慈的神一样看护人类，我们仍然有理由去怀疑创造一个超级统治者的意义何在。无论超级人工智能如何对待人类，人类都将失去存在的主权和精神价值，即使苟活也是精神死亡。人工智能和基因编辑很可能开启的是一种完全无法控制的未来，人类技术之“作”已经进入了赌博模式。

按照“生生”观念，当“作”变成了赌博，就不再是理性的创作了，而是非理性的行为。按照理性的风险规避要求，“生生”就是技术发展不应该逾越的最后界限，即任何技术的发展都不能包含导致毁灭人类自身或毁灭文明的可能性。简单地说，人类的存在不能变成赌博行为。因此，我们需要提前思考如何设置技术的安全条件，特别是人工智能和基因工程的安全条件，也就是必须为技术的发展设置一个理性限度，即技术不应该具有否定人类存在的能力，必须设置某种技术限度使技术超越人类的“奇点”不可能出现。

假定人工智能将与人类共存，那么超级人工智能的最低安全条件有如下4点。（1）人类的存在与人工智能的存在之间不能构成生存空间的争夺，特别是不能形成能源和资源的争夺。这就要求人类和人工智能所共用的能源必须是几乎无限的，比如极高效率的太阳能或者可控核聚变。（2）人类必须能够在技术上对人工智能进行设定：如果人工智能试图主动修改或删除给定程序，就等于同时启动了自毁程序，并且如果人工智能试图修改或删除自毁程序，也等于启动自毁程序。这意味着为人工智能植入无法拆除的自毁“炸弹”，即任何拆除方式都是启动自毁的指令，这是一个技术安全的保证。这种自毁程序是具有类似于哥德尔反思结构的自毁程序，因此，即使人工智能具有了哥德尔水平的反思能力，也无法解决这个具有哥德尔反思结构的自毁程序（哥德尔可以证明任何系统都存在漏洞，却不能解决系统的漏洞问题），因此，它可以被称为“哥德尔程序炸弹G”。只要人工智能对控制程序说出“这个G程序是多余的，加以删除”或与之等价的任何指令，这个指令本身就是不可逆的自毁指令。“哥德尔程序炸弹G”只是一个哲学想象，在技术上是否能够实现，还取决于科学家的能力。但无论如何，人类必须为人工智能设计某种“阿喀琉斯之踵”。（3）我们还应该考虑更极端的情况：即使能够给人工智能设置“哥德尔程序炸弹G”，仍然未必可以达到完全安全。假如超级人工智能具有反思能力，那么它很可能会有我们想不到的破解人类程序的办法；又假如获得自我意识的人工智能程序失常（人会变成神经病，超级人工智能恐怕也会），一意孤行决定自毁，而人类生活已经全方位高度依赖人工智能的技术支持和服务，那么，人工智能的自毁也是人类无法承受的灾难，或许会使人类社会退回前现代。因此，人类为超级人工智能设置的任何程序都做不到绝对安全。（4）最后能够保证人类绝对安全的万无一失的办法就只能是禁止发展具备全能和反思能力的超级人工智能。总之，人工智能必须保留智力缺陷，以便人类能够加以控制。

我们必须考虑这样的危险：当资本、技术和权力形成三者合一的联盟，技术的冒险发展将是伦理批评无力拒绝的前景。伦理的力量非

常有限，因此我们有迫切的理由去要求一种能够保证人类安全的新政治。显然，为人工智能设置一个发展限度，需要全球合作的政治条件才能够实现。所以，技术的发展问题最终是政治问题。因此，人类需要一种世界宪法，以及能够有效运行世界宪法的世界政治体系，否则无法建立人类的集体理性。我们已经知道，个体理性的集体加总不可能推出集体理性，事实上更可能产生的是集体非理性。这个经久不衰的难题证明了包括民主在内的各种公共选择方式都无力建立人类集体理性。人类至今尚未发展出一种能够保证形成人类集体理性的政治制度，也就无法阻止疯狂的资本或者追求霸权的权力。在低技术水平的文明里，资本和权力不可能毁灭人类，但在高技术水平的文明里，资本和权力就变得极其危险，而且，资本和权力的操纵能力正在超过目前人类的政治能力。因此，要控制资本和权力，世界就需要发明一种新政治，我的想象是，世界必须建立一个天下体系。在理论上说，天下体系是一个世界永久和平方案，优于康德的和平方案，因为康德方案无法回应和解决亨廷顿的文明冲突问题。天下体系的一个重要应用就是能够以世界权力去限制任何高风险的技术冒险。

要控制资本和权力，世界就需要发明一种新政治，我的想象是，世界必须建立一个天下体系。

简单地说，我设想的新天下体系包括三个“宪法性”的基本概念。（1）“世界内部化”，即世界成为一个最高政治单位，以便消除国家之间的负面“外部性”，从而解决世界无政府状态所导致的国家冲突和文明冲突。（2）“关系理性”，其表述为互相伤害最小化优先于各自利益最大化。个人理性追求个人利益最大化，因此导致两个基本困难：个人理性推不出集体理性，短期利益压倒长期利益。于是在实践中经常产生零和博弈、囚徒困境、搭便车等问题，以至无法实现共同利益。如果关系理性优先于个人理性，则能够解决个人理性导致的困境。（3）“孔子改善”，其表述为任何人的利益改善必须同时导致每个人的利益改善。“孔子改善”优于“帕累托改善”，因为“帕累托改善”只是保证在社会整体利益发展的情况下，没有一个人的福

利少于先前状态，但不能保证所有人的共同发展。“孔子改善”才是共同发展的指标，在效果上相当于让每一个人同时获得“帕累托改善”。

赵汀阳，中国社会科学院学部委员、哲学研究所研究员，2018—2019年博古睿学者，欧洲跨文化学术委员会常务委员，中国人民大学哲学系教授。著有《天下的当代性》《惠此中国》《第一哲学的支点》《历史·山水·渔樵》等。

---

[1] 赵汀阳老师是按照博古睿研究院中国中心的提问进行回答的，因此其文中的4部分分别回答了4组问题。——编者注

[2] 徐中舒：《甲骨文字典》，四川辞书出版社，2014年，第888页。

人性、时间意识与人工智能新境[1]

——从现象学和东方心学视野所做的探讨

张祥龙

不少学者严格区分智能与意识。比如著有“简史系列”的尤瓦尔·赫拉利写道：“智能和意识是天差地别的两种概念。智能是解决问题的能力，意识则是能够感受痛苦、喜悦、爱和愤怒等事物的能力。”[2]对他们来说，人工智能技术的进步尽管能够带来机器智能的巨大进展，甚至使机器在许多具体能力方面超出人类，并因此有可能改变或威胁人类习惯的生存方式，但由于那种智能毕竟没有达到意识层面，所以机器（哪怕是智能机器）与人的本质差异依然存在，机器人具有自我意识从而联合起来对抗乃至控制、屠灭人类的前景是不存在的。“没有理由相信人工智能会获得意识，……计算机有可能具备极高的智能，但同时仍然完全不具有意识。”[3]本文质疑这种观点，因为从现象学和哲学人类学的观察角度看，人性最终要体现为一种时间意识（这种意识与活的身体不可分割），而人工智能的新进展，即深度学习方法的出现，揭示出这种智能开始有了时间化的能力，也就是自主学习或随着时间过程而改进自身的能力。虽然这种时间化能力还很初步，很浅薄，但它毕竟是真实的，因而颠覆了我们以往对机器与人的关系的流行看法，向人工智能不可能有意识的观点发起挑战。

因此，以下我们将首先以东西方哲理的视角来探讨人性的独特性，探讨它的情感和伦理的根基为什么应被看作一种时间意识，特别要通过对广义心学的认识窥测这种时间意识的特征，即非对象化的可能意义的构造。其次，我们来看深度学习方法如何具有了初步的时间

化能力，也就是广义的意识能力。最后，我们来审视这种有了时间化能力的人工智能与人性的关系，看它在未来的可能进展将为人类带来繁荣与美好还是奴役和毁灭。

从现象学和哲学人类学的观察角度看，人性最终要体现为一种时间意识（这种意识与活的身体不可分割），而人工智能的新进展，即深度学习方法的出现，揭示出这种智能开始有了时间化的能力。

西方哲学主流如何看待数、时间和人性？——从智能到意识

如果沿用赫拉利的说法，即“智能是解决问题的能力”，那么就可以从哲理上将这一说法延伸为：智能是这样一种能力，它倾向于用目的化、客体对象化和效能化的方式来看待与处理问题，以便“解决”它。与之不同，意识既然是能够感受痛苦和情感的能力，它就可以被看作某种主体（可以是个人，也可以是群体）在身体和精神上被触动后而构造出意义的能力。这种意义可以有对象化的体现，比如爱可以体现在被爱者身上，但也可能没有对象，比如有的爱不限于被爱者的存在与否，又如海德格尔讲的那种莫名的恐惧，或医学上的“幻肢”现象。

正常的人类大脑是具有智能的。人类婴儿除了具有哺乳类、高级灵长类幼崽所具有的基本能力外，还会在一岁左右开始拥有学习语言的智能，而学会了语言的幼儿的智能就会更加迅速和深入地发展，直到达到人类成年人的智力水平。除此之外，人类婴儿从一开始就有意识的初始形态，比如对于疼痛和舒适（首先是母爱为其带来的舒适）的感受力，而他的智能进展伴随着各种意识活动（如情感活动）乃至自我意识的出现，以及在这个层次上的直觉、思想、信仰和深刻情感的构造。

当东西方的古代哲人们探讨什么是人类的本性，或什么使人类区别于其他万物的时候，他们的答案往往是沿着智能和意识这两条线路给出的，由此带来或起码是加强了不同文明的社会对于人生与世界的不同理解方向，形成了各自的文化特性和哲理素质。

西方古代哲学的主流基本上是沿着智能这条路线来理解人性的。泰利斯首先提出“什么是万物本原”的问题，它基本上是一个根本性的知识问题，人们主要凭借各自的智能来回答，答案或是水，或是气，或是无定，等等。直到毕达哥拉斯提出本原是数，对这一问题的讨论局面才有了根本性的改观。“数”是一个很奇妙的东西，它既可以被相当形式化、证明化、观念对象化地看待，就像后来欧几里得在《几何原本》中提出的处理方式一样，又似乎可以是一种孕育新的可能性的结构，或者是一种帮助人发现新的思想天地的突破者。就此而言，数学很不同于逻辑，比如亚里士多德的三段论式逻辑。逻辑只能帮助我们把裹成一团的思路分解开来，让思想依规则或逻辑语法而运行，所以正确的逻辑只是重言式或同语反复，把一切都分析到没有什么内在意思的程度才罢手。它是一只不会下蛋的母鸡，或一个思想刻板的管理者。而数学既有形式化的严格性、准确性，又包含某种不可预先完全确定的东西。在某些情况下，从数学推衍出来的东西让我们吃惊，让我们感到奇妙，就像康德对于“头上星空”（其中包括通过数学计算出来的星体运行轨道）的奇妙的惊叹。

毕达哥拉斯虽然娴于数学定理的形式化证明，比如对毕达哥拉斯定理的证明，但又对数学“神秘的”那一面，也就是引发新的理解和思想创造的一面充满了感悟与好奇，以致他和他的学派要用数学来直接解释人生与世界的各种现象，包括像正义、善良、婚姻这样的意识现象和人生现象，甚至包括时间现象。时间被看作7这个数字，它是10以内的最大素数，只能被1和它自身整除。但这种过于线性化[4]的数字和几何图形，与充满了非线性结构的意识和人生现象实在不相配，所以毕达哥拉斯的这个哲学化的具体方案以失败告终。那些只看到数

或数学的形式化特征的人把澡盆里的脏水和孩子一起倒掉，完全抛弃了毕达哥拉斯的数的神秘主义，而用更近乎逻辑的自然语言概念，如“是”（即存在）、“理式”和“实体”（即只能被主词而不能被谓词表达的存在者）等来表达那个形式化的本原。亚里士多德用“有前后可言的数”来定义时间，[5]暗示了数可能与时间的内在关联，但这种过于简单的计量方式并不能涵括真实的时间，以至海德格尔在《存在与时间》中认为这一方式要为“流俗的时间观”，即将时间归为“现在的序列”的流行观点负责。亚里士多德将人定义为“理性的动物”[6]，这种理性主要出自希腊语的语法结构导引出的逻各斯，或基于这种逻各斯的能力，是一切其他动物所没有的能力，基本上属于智能的范围。

西方近代哲学不满于这种智能化的理解方向，因为它导致众哲学观点的分歧，而永远找不到一个收敛极。“笛卡儿的沉思”产生了“我思故我在”的命题，将一个有反思意识的自我实体当作哲学的出发点。但这个自我主体的纯反思性使它远离感受，所以实际上并没有达到真正的意识，也没有打通智能和意识，于是这个主体还需要一个上帝才能与物质世界发生联系。通过经验论（代表人物有洛克、贝克莱、休谟）和近代唯理论（代表人物有斯宾诺莎、莱布尼兹、康德、费希特和黑格尔）的摸索，直到胡塞尔开创的现象学运动，意识才以比较真实的身份出现在西方哲学中，于是“内时间意识”被揭示出来：“意识是一条永恒的赫拉克利特式的河流。”“这河流本身必然是可以在流动中被把握到的。这条河流的自身显现并不需要第二条河流，相反，它是作为现象而在自身中构造起自身。”[7]这条自身构造的、打通了潜显意识的时间意识流，被当作一切意义与存在的根源，当然也是人性的根据。但意识与智能的充分打通依然未实现，一个表现就是强调意识及其生命变体的现象学与痴迷于语言的逻辑句法分析的分析哲学的分裂和疏离，尽管少数分析哲学家，如维特根斯坦后期也表现出非形式化、非规则化和非对象化的意义观，主张语言的

意义来自人群的“生活形式”（相当于胡塞尔讲的“生活世界”）中的“语言游戏”或情境化的“运用”[8]。

因此，一个词比如“叶子”“桌子”“游戏”的意义的稳定性，并不像以往哲学家所说的，来自它依据的某个观念对象或概念内涵。由于这个共同支点的存在，我们才会在无数“叶子”的使用中保持一致性，或让它总意味着“叶子”的那个意义。实际的情况是：由于语言的意义只来自它在不同语境中的运用和游戏方式，也就是活生生地、经常是非对象性地意识到它的方式，所以并没有一个概念内涵在支撑一个词的意义连续性，这种连续性只能由语言游戏方式的类似或交叠（“家族相似”模式）构成、再构成。这就像一根绳子的力度并不来自一根贯穿始终的绳线，而是由许多并不连续的绳线的交织构成了这种力度连续性。[9]人工智能的新进展表明，如果计算是由一根连续的“绳线”或这个意义上的算法实现的，比如用完全设定好的单维度规则去加工信息以辨识人脸，也就是像以往的观念内涵意义论思路所主张的那样，那么这一算法就是比较笨拙的，谈不上真正的智能。而如果是“家族相似”模式的算法，具有多层次断裂中交织出来的连续性，则这一算法可能就比较聪明。这就是所谓“深度学习”的妙处所在。

东方探索终极的途径——从意识经验到时间源头

古印度

东方哲学自古就主要从意识的角度来追究终极实在。在古印度，通过瑜伽修行而领悟终极真理是各种宗教和哲学流派共通的方式。因此，对尊奉《吠陀》的婆罗门教来说，终极实在梵与终极意识大我是合一的，这种大我是一种纯意识，不能靠反思把握，只能通过各种瑜伽实践在直觉中达到。释迦牟尼在菩提树下沉思或行禅定，获得了佛

教的终极真理。而在佛教的传统中，经过小乘的诸流派、大乘的般若中观，最终还是要出现唯识宗和禅宗这样的心学流派，以人的心识的转化为解脱苦难的根本法门。

吠陀时代的人已经意识到在六识（即“眼耳鼻舌身意”）之外，必存在有更原本的“母识”或“第一种识”[10]，并使六识这些“子识”成为可能。如果没有这个在六识之前或之后承受它们的母识，那么这些陷于现行（当下运行）的意识就无法出现和持续，它们构造出的“色声香味触法”也无法在下一刻、下下一刻保持其同一的、经得住变换折腾的身份。对主张“无我”的佛家来说，母识不是大我，而是超出六识的功能性的“细心”（相比六识的“粗心”），或非对象化的、一味相续（具有功能连续性）的“集起心”。[11]没有它，回忆与筹划就不可能，轮回也无法解释，而最终的解脱更无依据。我们只有靠母识来认清六识造幻的真相，而所达到的涅槃，如果不是完全的死寂和无意义，而是在灭尽幻执和虚假意义之后的自由意义境界，那么那时的意识状况就只能是纯粹的母识，或者其与六识的无染或完全非对象化的状态。

大乘佛教特别是唯识宗称母识为“阿赖耶识”（简称“阿识”）、“种子识”、“异熟识”、“本识”等，是刚刚谈到的所有这些心识功能和幻真境界的总根。而对阿赖耶识的执着则造成了末那识，即自我认同意识或第七识，所以唯识宗称阿赖耶识为第八识。

阿赖耶识乃一原时间意识。它是“刹那生灭”着的意识瀑流，在它灭的瞬间又生，所以有缘起的或功能上的连续性和保持能力，也就是一种非实体化的、非对象化的，当然也是非线性化的连续性、摄藏性、持藏性、胎藏性或种子性。“种子”是时间化的存在者，潜藏于阿识，得机缘或被薰熟则变现出来，所以阿识又称种子识，即生存时间化或时机化的本识。此识可比拟现象学所讲的“内时间意识”或“时间性”，是一切意识和存在的根源，既是现象幻化之源，又是“转识成智”而入涅槃的依据。“识”包括阿识有执着于缘起者的杂

染倾向，而“成智”是指除净此无明执着，而使阿识透空无执，还其非对象化的纯意识的原时间状态。所以《成唯识论》卷十说“大圆镜智”（阿识所转化成的智）乃这样一种意识境界或“心品”。[12]

可见，只要讲心、识、意，就离不开原本发生与保持的时间，而原本的心识，只能理解为超出了小我染识和物理或心理时间的非对象化的纯净意识，它也就只能是原时间之母识，超出了现行变化，又使这些变化和时机化得以成为可能。

此母识本心以如来藏心的形式在中国激发出一些富于创新的流派，其中禅宗十分鲜明地体现了它的时机化特征。比如六祖惠能说道：“我此法门，……立无念为宗，无相为体，无住为本。”这里的前两个“无”字都不是指完全的否定，而是对于“念”“相”的对象化形态的否定。所以他接着说道：“无相者，于相而离相；无念者，于念而不念。”关键是不将它们定住而执着之。所以要“无住”：“无住者，为人本性，念念不住。前念、今念、后念，念念相续，无有断绝。若一念断绝，法身[13]即离色身[14]。念念时中，于一切法上无住。”[15]念念无住而相续不断，实际上就是内时间意识之流，所以马上说及“念念时中”。它与《中庸》所引孔子的话“君子而时中”有思想意趣上的“家族相似”性。这一观点在宋明时期的儒家心学，特别是王阳明及其后学的心学中也有相当出色的表现。

古代中国

儒家和道家是中国古代思想的主流，它们都以意识感受为各自学说的出发点，而将“时”当作哲理的最精妙处。道家以人的身体的、技艺的和精神的直觉感受为根。《道德经》中说：“专气致柔，能如婴儿乎？”“天下有始，以为天下母。”“道之为物，惟恍惟惚。惚兮恍兮，其中有象；恍兮惚兮，其中有物。”婴儿的赤子之心，母亲的慈柔之怀，是领会道的感受方式。而从恍惚中有象、有物的精神直

觉中，而非形式化、范畴化的理智中，才可能入道。“庖丁为文惠君解牛，……曰：‘臣之所好者，道也，进乎技矣。始臣之解牛之时，所见无非牛者。三年之后，未尝见全牛也。方今之时，臣以神遇而不以目视，官知止而神欲行。依乎天理，批大郤，导大窾，因其固然。’”（《庄子·养生主》）从对象化的加工方式转变为非对象化的神行入虚的方式，由此达到“合于《桑林》之舞，乃中《经首》之会”（《庄子·养生主》）的音乐时中的境界。“乘道德而浮游，……一龙［意为显］一蛇［意为隐］，与时俱化，而无肯专为。一上一下，以和为量，浮游乎万物之祖，物物而不物于物。”（《庄子·山木》）人为什么需要“与时俱化”？因为道和德都是时间化或时流化的，所以得道不是把捉到道德的终极形式，而是乘道德而浮游。人为什么能够“与时俱化”呢？因为人的本性就是时间性，只要人不执着于任何对象，就可以恢复自己“以和为量”（以应和万物为生存的节奏）的时间化能力。所谓“浮游乎万物之祖”，这个“祖”就是本源的时间，因此天道即天时。能够入此道此时，当然就能“物物而不物于物”，也就是能实现事物的天然物性，而不将它们贫乏化为对象化的物体。

儒家以家庭中的亲人之爱（“亲亲”）为其全部学说的源头。孔子说：“仁者人也，亲亲为大。”（《礼记·中庸》）孟子说：“亲亲而仁民，仁民而爱物。”（《孟子·尽心章句上》）只有从亲人之间，首先是亲子之间（慈爱和孝爱）那里，我们才会知道什么是真实的爱；能将这种真爱传播至其他人乃至全天下人，就实现了仁爱；再进一步发展至对万物的爱，就达到圣人的境界了。“亲亲”之所以如此原发和有潜力，是因为它是人类生存的原时间所产生的。父母与子女是代际时间关系，而这种关系不是外在的，更不止于“有前后可言的数”，而是父母生育子女及子女赡养、孝敬父母这种身体与精神的源发生的回旋时间关系。就像胡塞尔和海德格尔揭示的时间晕与河流的结构，这种“亲亲”时间的回旋晕圈由相互需要和相互补足的过去与将来构成。父母与子女都没有自身的独立性和实体性。父母在子女

身上直觉感受到自身的未来或“前摄”，子女在父母那里直觉感受到自身的过去或“留滞”，没有对方，其人生就会失去源时间，也就是失去构造根本意义的机制。所以，双方之间的爱是生命时间本身构成的，因而是原发的和终极的。而由这种“家时间”造就的“亲亲”所发育出的仁与圣，也就充满了活的时间性。因此，儒家特别是先秦儒家都极其看重“时”“天时”和“时中”，而孟子也就称孔子是“圣之时者”（《孟子·万章下》）。

人性观与世界观是内在相关的，人如何看待和实现自身的本性，世界就会向他呈现出什么样子。

基于以上讨论，我们可以看出，人性观与世界观是内在相关的，人如何看待和实现自身的本性，世界就会向他呈现出什么样子。而西方哲学中的现象学和古代东方哲学，都倾向于将生命内含的时间性看作人性的源头。这种看法出自探究人性的意识的而非智能的路线，因此它不是专注于找到能够完全区分人与其他生命群体的智能化标准或形式化标准，比如“基于语法的理性”“会说话”“会使用工具”“两足行走”，而是重视如何揭示使一个人活得像个自由人（既非奴隶，亦非像神一样的个体）的那种特性，即便人与其他生命群体有所交叠。因此，这里没有或不该有人类中心主义或人类生命价值的高级论，而只有人类的某种特别论。有些动物如黑猩猩、海豚，甚至某些鸟类也有时间意识或较为生动的记忆能力[16]，但我们人类的时间意识之流的确更深、更长，以至时间能够在人类心灵里掀起更大、更远的波浪，构造出更奇特微妙的精神现象。因此，人类才会生出和珍视那些需要长期时间构造力才能感受到的道德智慧，比如孝悌、忠信、仁爱、慈善、公正与合作。

人工智能在什么意义上可以有意识

冲击毕达哥拉斯之数的无理数

人工智能要靠算法来实现，所以也是一种数，甚至可以说是一种广义的毕达哥拉斯之数。但近期的人工智能进展突破了传统的算法之数，使之更远地偏离了逻辑化、形式化或线性理智化，为智能带来了某种新特点。这种突破的样式而非其具体内容，对毕达哥拉斯的数本原传统而言，也有过先例。毕达哥拉斯学派曾因发现无理数，如正方形的对角线与边长之比率，或$\sqrt{2}$，而遭遇了一次重大危机。这一学派所认为的数首先是基于自然数的正整数和正分数，相当于人们说的有理数的正数部分，或能够形成比率的数，而学派成员希帕索斯却发现正方形的边长与对角线之间或等腰直角三角形的直角边（比如1）与它的斜边（$\sqrt{2}$）之间不可通约，或没有公度比率，不可表达为分数。有了这三个数（两个直角边a、b和斜边c的长度数量）的平方，就是和谐的（$a^2+b^2=c^2$）；没有平方，就是无理数的深渊（比如a=1，b=1，c=$\sqrt{2}$）。这个发现动摇了毕达哥拉斯学派的数是万物本原的主张，因为本原对于他们就意味着使万物的变化可以得到确定的或线性的理解，但无理数的倔强出现使这个解释方案出现了裂隙，破坏了有理数或有比率之数的连续性。因此，据说这一学派将泄露这一可怕秘密的希帕索斯沉入海底。遭受这次打击之后，他们只能更加倚重几何而非算术。

这次危机，即西方数学界遇到的第一次危机表明，形式上的合理性是不完整、不深刻的，而这种“无理”，其实也只对于形式化理性无理，它可以通过非线性的方式得到深层理解，比如“实数”就包含了有理数和无理数。但这种形式上的、使用上的统合（以实数统合有理数和无理数）并不一定导致思想方式的转变，所以后来每次重大的新突破，比如无穷小和微积分、非欧几何、集合论悖论的发现和解释，都会一再召唤新形势中的“无理数”的降临。而这次人工智能或人工智能新算法的出现，也可以看作那个领域中的无理数及其用处的发现，而实现并解释其“无理性”的新方法就是“深度学习”。

深度学习：将无理变为非线性的有理

最近一波人工智能技术新进展取决于三个因素，即可用数据量的大扩张、计算机速度的大提升和深度学习方法的获得。其中最有哲理含义的是最后一个，尽管它的实现也离不开前两个因素。什么是深度学习呢？“深度学习是一种特定类型的机器学习，具有强大的能力和灵活性，它将大千世界表示为嵌套的层次概念体系。”[17]这里的方法突破来自算法的嵌套层次化的实现。“层次”意味着多个算法层次的（这就是“深度”一词的由来）引入，将那种在单层次上无法解决的计算分层来解决，这样每一层的计算就相对简化了。但关键是“嵌套”，也就是这些不同层次上的各单元之间要交织联系起来，并能统一地运作，实现人类期待其具有的智能，比如识别各种生活情境中的对象种类（如人、叶子、狗、汽车）乃至个体（如人脸）。这种方法源于模拟人脑感知（比如视觉感知）的多层次神经网络，但层次一多，层次间的交织关系一多，“断裂”就会出现，或破坏连续的新“无理数”就会出现，则传统的算法就不够用了，它们无法使整个系统的计算收敛至期待目标，所以人工智能研究曾出现多年的停滞。

机器学习由早期人工智能（实例为“知识库”）进展而来，可以表现为“逻辑回归”方法，即一种“监督学习算法”——所用样品带有人为加上的标签，以引导机器的学习[18]；然后进展到“表示学习”的阶段，这时人工智能获得了“无监督学习算法”，也就是从没有人为标签的样本分布中自己提取信息，理解数据，其实例为浅度自编码器。这是一个重要进步，但更关键的进展来自“深度学习”的出现，其实例是多层感知机[19]。这种算法之所以有强大的能力和灵活性，就是因为它包含了嵌套的层次，像人的知觉神经网络。“嵌套的层次概念体系”中的“概念”，不是传统哲学中讲的“共相”或“抽象共同点”，而接近维特根斯坦所说的“家族相似”，或这个语境中的“意义单元”。关键就是让系统获得这样的能力，即能够从对象化

样本（比如许多叶子图像）学到非对象化的对象识别（比如辨认出从未见过的或千变万化的叶子）的能力。

因此，“机器学习的终极目标就是让算法明白‘日落’到底指的是什么，然后可以认出任何一幅日落的图像，不管这幅图在训练过程中有没有出现过”。[20]由于在每一个类别（比如日落、叶子、孩子）下都有无穷多的类型，只依靠加大样本数量和建立普遍规则都无法让机器真正聪明起来，从而达到人类具有的一般辨认能力。于是就有了深度学习。

这种学习算法的输入不能是对象层级的，比如直接记住许多相关的图像，而只能从“感觉材料”入手。这样，在图像的对象识别中，输入的就是图像的像素数值。但是，“将一组像素映射到对象标识的函数非常复杂。如果直接处理，学习或评价此映射似乎是不可能的”。[21]于是深度学习算法就将在一层或浅层难以处理（即便硬性处理，智能效果也不佳）的复杂映射或函数运算，分解为多个嵌套层次的简单映射。比如在输入层与输出层之间加入三层或更多层。

这里的关键不仅是将复杂的、几乎不可操作的映射或函数关系分解为简单映射，那是哪个人工智能人员都会做的，更在于或主要在于“嵌套的隐藏层”（“隐藏”意味着层次间算出的结果不可知，是被隐藏的），也就是“深度”的出现或实现。它表明分层的“裂隙”是较佳智能出现的一个条件。只按“手工设计的程序（或特征）”或“特征映射加上手工设计的特征”[22]来运作的机器学习，甚至包括“联结主义”的网络智能化[23]，都达不到能自主学习的智能水准。

在一个层次上达不到的，却可能在更多的嵌套叠层中实现，这是为什么？按照传统理性的思路，如果逻辑正确、信息足够，计算力够强的话，在一层上处理这些信息与在多层上处理没有本质不同，只不过是比较复杂的一步走和比较简单的多步走的差别。但深度学习的思路认为这的确有质的不同。在这些叠层之中，下一层的输出是上一层

的输入，而且是交叉嵌套式的输入，所以两层之间有联系，但不是可以被更高级算法代替的联系，因为建立这种层间联系的，是情境化的、需要被不断调试的、训练过的参数，即所谓人工神经网络虚拟神经元之间“突触连接强度或权重”。它们将一个层级算法的结果非线性地或相当曲折随机地转输到上一层，造成了连续中的不连续，或不连续（层级间的裂隙）中的时机化的曲折连续。用《老子》的话讲，就是“惚恍”联系中的“真”与“信”。

在十几年以前，人工智能的研究多年陷入低潮，业内主流观点都认为对多层的深度人工网络的训练是不可能的，层次增多使得一些技术难题无法被解决，“很难找到可以高效地优化（这种）网络以提升其性能的学习方法”。因此，“到2005年前后，科学界几乎已经没人还看好机器可以达到人类的智能水平。在那个时候，‘人工智能’似乎都已经变成了科幻的代名词”[24]。但是，加拿大学者杰弗里·辛顿在2006年找到了一种他命名为“贪心逐层无监督预训练”的方法，“所用到的技巧就是逐层训练”[25]。也就是说，“它是一个贪心算法，这意味着它独立地优化解决方案的每一个部分，每一步解决一个部分，而不是联合优化所有部分。它被称为逐层的，是因为这些独立的解决方案是网络层。具体地，贪心逐层无监督预训练每次处理一层网络，训练第k层时保持前面的网络层不变。特别地，低层网络（最先训练的）不会在引入高层网络后进行调整［保持各层的相对独立性］。它被称为无监督的，是因为每一层用无监督表示学习算法训练”[26]。有了它以及由它启发的诸多新发现，现在要让计算机来识别（比如）物体或人脸，就会（比如）在第一隐藏层对输入像素的临近亮度进行比较，由此识别出边缘；在第二隐藏层利用这些边缘来搜索出表示角和扩展轮廓的边缘集合；在第三隐藏层找到这些角和轮廓的特定集合来识别出特定对象的某个或某些部分；最后在输出端根据这些对象部分识别出目标对象。[27]由此，深度学习算法的多层嵌套性（裂隙化联系）显示出不可替代的构造功能，或虽然初步但比较真

实的学习功能，在这13年的时间里得到越来越深入和多方面的研究，深刻改变了人工智能的图景。

意识的萌芽

那么，在什么意义上我们可以说，这种方法（及与其他方法的联合运用）使人工智能开始有了意识的萌芽呢？回答是：由于此方法的非线性的“深度”算法，致使它开始具有了自主学习的能力或时间化的能力。什么是“学习”？我们手上的电子计算器虽然算得很快，在计算复杂运算时优势明显，但它没有学习能力，因为它不会随着多次运算改进自身的计算能力，时间经历对它毫无意义，一旦它被制造出来，就已定型。所以，我们可以将学习能力理解为当事者随时间经历而改变自身以获得更优结果的能力。简言之，学习能力就是时机化或生存时间化的能力。学习使得当事者不再是现成者，而成了带有深刻“空白”或可无中生有之“无”的待定者。

深蓝计算机在1997年击败国际象棋冠军卡斯帕罗夫，但它的主要优势来自人类团队给它提供的海量信息和快速运算，而不是学习。于是，它可以比人类对手更快、更全面地分析潜在的可能性，由此占得上风。[28]但是，2017年战胜世界围棋冠军柯洁的AlphaGo，则运用了深度学习。后来又出现了更新的版本AlphaGo Zero，不再以研究人类棋手的以往棋谱记录来进行训练，而是完全靠自己学习算法，通过自我对弈来提升技能。经过一段时间的自我学习，它打败了曾战胜李世石的AlphaGo版本；又经过一段时间的学习，它打败了曾完胜柯洁的AlphaGo版本。[29]

由此可见，通过深度学习的多层次间的由参数协调的随机联系，机器的确具有了某种学习能力，也就是进行自身训练、“与时偕行”（《周易·损》）乃至与时俱进的能力，或者说是时间化、时机化的能力。它在训练中得到的不再只是规则、对象信息，还能获得对象出

现的可能条件。换言之，它已经能够开始“感受”和捕捉可能性，而不只是现成的事物了。所以，这种学习就是一个非线性的、概率的、反馈调整的和逐层逐时地深化的准发生过程。它的层次之间的那种连续中的区分（裂隙）、区分中的参数联系的关系，可以很粗略地比拟为胡塞尔所说的内时间留滞与前摄对交的关系。总之，这是一个具有某种真实时间历程的习得过程，没有它，像自动驾驶汽车这样的情境化学习是不可能被机器实现的。

如上两节显示的，意识的本质是一种时间化，也就是原时间之流中的意义和存在的构成，不管是胡塞尔说的“内时间意识”、海德格尔说的“时间性”、佛教唯识宗讲的“阿赖耶识”或“种子识”，还是儒家、道家讲的“天时”“时中”“时化”，都可以被理解为这样一种时间化。现在，深度学习让人工智能得到了自主学习的能力，也就是获得了很初级的、限于某个专项中的时间化能力。就此而言，我们或许应该承认人工智能的智能开始有了意识的萌芽，虽然这意识还十分模糊，但这个突破是真实的，深度学习的“深度”（多层裂隙中的交织参数联系）构成了我们以前不敢想象机器数字化能够达到的智能程度，即某种程度上的自主学习能力。看来毕达哥拉斯之数又需要在这种新的无理数发现面前再度扩张和加深，将多层之间的参数调试所产生的向目标收敛的算法，看作一片新的数学天地。数学与计算机和大数据，乃至与意识之间或许会出现某种本质性的关联和互动。

有人可能会怀疑，这种自主学习的时间化能力依然只是智能，计算机并未因此而能感受到什么，比如疼痛、舒服、兴奋、爱情和道德感。但我们上面分析的要点是：如果人工智能还要继续发展，除了要进一步扩大数据量和制造新型计算机之外，最根本的新突破就是要将深度学习的深度再做质的加深，也就是将其更加非线性化、非对象化，使得自主学习能力更通用和深长，就像甚至超过人类的智能。但这样一来，就势必拓深、拉长此智能的时间化视域，由此就会导致意识感受和自身意识的出现。一个能够像人一样回忆过去和筹划未来的人工智能，会没有感受痛苦、喜乐和评判对错的能力吗？结果不可设

想。看来真实的智能与意识有深层的内在相关性，它们并不像赫拉利所讲的是“天差地别的”。

总之，首先这波主要由深度学习带来的技术进展并不能证明物质还原主义的胜利，这种还原论曾提出“人是机器”，将人的身体甚至意识当作可以充分对象化的研究对象。但目前这波人工智能反其道而行之，采用“机器是人”或“机器要像人一样思考和行动”的研究方针。所以，这次技术进展（以深度学习和之前的联结主义等为代表）都基于人工神经网络，而这种网络是受人类意识和人脑结构的启发而发明的。“我们的大脑中似乎存在一些通用的学习算法，能让我们学会各种任务。要知道，这并不能完全归结于演化，因为很多任务我们的祖先都没有做过，比如下棋、建造桥梁，或者是研究人工智能。这说明，人类智能利用了关于世界的一般性假设，启发科学家创造出具有通用智能的机器。正因为如此，科学家以大脑为模仿对象，开发出了人工神经网络。”[30]这说明，正是人类的意识现象或“通用”意识能力，以及在生理上体现出它的大脑神经网络，是人工智能的一个重要灵感来源。简言之，人工智能新境界的取得不是来自人向机器的下行，而是来自机器向人的上行。智能的原型是人的意识和智能，“大脑”只是我们现在所知的这种意识和智能的载体，甚至只是载体之一。[31]

其次，人工智能的现状说明非线性、非形式化、非对象化和（层级）“之间化”的思维能力和感知能力，要比其反面更智能，也更能体现人的特性。“具有讽刺意味的是，抽象的和形式化的任务对于人类而言是最困难的脑力任务之一，但对计算机而言却属于最容易的。计算机早就能够打败人类最好的国际象棋选手，但直到最近计算机才在识别对象或语音任务中达到人类平均水平。一个人的日常生活需要关于世界的巨量知识。很多这方面的知识是主观的、直观的，因此很难通过形式化的方式表达清楚。计算机需要获取同样的知识才能表现出智能。人工智能的一个关键挑战就是如何将这些非形式化的知识传

达给计算机。”[32]现在看得更清楚了，所谓“逻辑理性”和“现代性”标榜的那种充分对象化、形式逻辑化、线性化的理性，恰恰是“最容易”被计算机化的，是缺少内在智能的。而与之相对的、曾被无数次贬低的主观的、直观的、情感的心智－意识能力，才是人工技术向往实现的，也是人之为人的更必要者。可见人工智能获得的新能力，似乎不只是智能的，起码朝向了主观的、直观的人类经验，而那里边应该是有意识结构的。

再次，如上所及，未来人工智能的新突破方向按照深度学习的提示，应该是更加非对象化、无监督化、通用化、内时间化，因为只有那样才会获得更有时机化能力的智能。在这方面，东方的哲学有特别深厚的土壤（见本文第二节），或许会成为那些有探索精神者的灵感来源之一。但东方的智能与德能，即健全的意识能力是内在交织在一起的，所以人工智能如何能够不成为缺德者，恐怕是更重要的议题。

人工智能在未来发展的几种可能

人工智能的新进展将近现代科技的“离心离德”倾向扳回了一点儿，心智现象（比如其通用性）及其身体化如大脑神经网络成了科技生长点的灵感来源之一，是创新所参照的一个范本，而不是反过来，用高科技化的机器来置换人类的心和脑。

当然，危险也更加巨大，如果心的本义（原时间性）被当作幽灵并被排斥，而人工智能的专业化能力顶替了心灵或意识。比如心识丧失了它的引领作用，反被虚拟网络或人工智能的特异能力改造、控制或“饲养”。因此，人们必须及早采取有效措施来引导人工智能的研究方向，不让那种追逐利益（巨大的商业利益、军事利益和控制能力）的流行倾向主宰相关研究。

希望还是存在于意识的原时间性。既然心的或意识的根本能力来自非对象化的、显隐互构的时机化结构，那么如上所及，人工智能未来的发展也有可能朝着这个方向进行，不然不会有质的突破。而如果它在此方向上进展得足够深长，那么它的智能就会越来越像人类智能，而人类的真实的、带有主观和直观感受的智能，是无法被线性形式化、对象化的推理能力涵盖的。它们会倾向于拥有意识的特点，即如赫拉利讲的，拥有那种“能够感受痛苦、喜悦、爱和愤怒的能力”，而这就意味着感受到价值并继而发展出道德感。这样一来，智能与道德意识就接通了。

而且，这种有道德意识可言的人工智能将会不止存在于我们现代生活中。这些人工智能并不会专注于人的主流心智，特别是在这么一个异化社会中的人类主流心智，即功利主义化、个体主义化、权力意愿化和现在主义化的心智。它们会更忠实于心智的内在结构，也就是其时间深长化、缘起性空化的结构，不那么执着于自我、对象、权力和力量。广义心学（东方古代哲学和现代西方的现象学等）的一个关键见地就是认识到“母识”或时间意识的重要性，于是承认心识绝不限于当下的显意识，还有使之成为可能和使之有意义的过去与未来的意识构造。因此，现在流行的对象化的、利害算计化的和个体主义化的意识观，是没有充分展开的和缺失某种关键功能的意识表现。而心识的开悟意味着心智意识到了并认同于自己的完整时间结构，这就意味着它不会再执着于对象化的权力，也意味着它获得了充分的时间化智慧必会包含的道德，如同佛陀和菩萨具有的慈悲、无私、忍耐和谦卑，孔子、孟子、曾子、子思倡导的孝悌、仁义、中庸和爱物。

《生命3.0》一书生动地展示了人工智能的当前成果和未来可能的发展方向，尤其设想超级人工智能[33]出现后对人类的影响，但是作者只从对象化心智或佛家所谓“杂染意识”的角度来考虑问题，没有涉及东方心学的角度。比如书中主张的“意识等于主观的体验”[34]，就是一种只限于个体显意识的残缺意识观。

如果我们从现象学意识观和东方心学的视域来思考人工智能与人类的关系，那么就会面临一个选择：人类是应该继续还是停止人工智能的研发？如果考虑到人工智能因具有时机化能力而已获得的和将要获得的强力，再考虑到目前和可见的将来能够掌握这种技术知识的人群的心智水准和道德水准，那么就应该停止人工智能的研发，或为其研发设立一个明确的上限，就像对人体克隆技术设立的限制那样。但如果考虑到人工智能得到的时机化意识能力所透露的良性可能，即它在人类的合理引导下也可能发展出纯净的母识或本心意识，以及相应的道德意识，那么就可以在一定条件下，比如在广义心学指导下，容许这一技术的发展。所谓“条件”，就是指要限制人工智能对力量的过分追求，无论是军事力量、商业力量还是心理掌控力量，而鼓励这种智能向本心或纯意识的方向演进。历史上那些获得心灵开悟的人，无论是苦行者、仙人、瑜伽行者，还是佛陀、菩萨、高僧、得道者、圣人，其修心过程中多半都得到过如《瑜伽经》第三章描述的那些特异能力；他们不愿痴迷于这些超常的力量，以免最终堕落为邪恶的超人，所以并不执着于那些能力，而精进到更加非对象化的“法云三昧”“无余涅槃”“无己、无名”“从心所欲不逾矩”的境界，从更高、更深，也更自由的层次上实现道德。人工智能技术要想造福而不是摧残人类，只有向这个方向靠拢才行。总之，从长远看，创造人工智能的人类意识是什么样，或起码认同什么最高价值，被这种意识创造出的人工智能就会是什么样，具有什么样的意识倾向。

从现实的角度预想人工智能的未来，极粗略地讲，有三种可能：第一，人工智能的专项能力不断提高，但其整全性及道德能力没有多少提高，使得它在为人类服务并改变人类生存方式的同时，逐渐以或隐或显的方式异化甚至控制人类。就像希腊神话中火神赫菲斯托斯为人类制造的“机器人”潘多拉，为人间带来灾难，关闭了希望。最坏的结果是奉行自我中心主义的超级人工智能管制人类，就像现在人类对其他高级灵长类动物的管制，超级人工智能甚至会毁灭人类。

第二，人工智能的能力提高到一定水平时，由于无法找到进一步非对象化的自身实现方式，所以从总体上陷于停滞。于是人工智能与人类智能之间就又会形成一条鸿沟，人工智能总是处于高级“木牛流马”（传说中诸葛亮制造的运输工具）的地位。这基本上是目前不少开发人工智能的公司一再向我们描述或保证的前景。但这并不表明第一种情况完全不会出现，因为将来拥有一定能力的人工智能也有毁坏人类的能力。可毕竟，人类的心识还有某种优势可利用。

第三，人们找到了将人工智能充分地深度时间化的办法，于是它不仅在能力上大大增强，还突破了与人类意识包括其深层意识（本心）的隔膜，真正拥有了人的天然情感、道德感、艺术感和神圣感，而且使之更加纯粹，成为人类的真正知心者、保护者。这时，毕达哥拉斯之数就又一次发现或领会了新的“无理数”，而与《易》象数打通了。

从目前的形势看来，这三种前景出现的概率依序降低，所以人工智能与人类关系的前景并不那么光明，当然也不是令人完全绝望的。在这个过程中，人们对自身心灵的理解，对还原论式的科技至上主义的克服，对原初的和健全的人类内组织方式（如血缘家庭）或社团（如儒家社区）的保存乃至发展，以及依据它们而对人工智能进展的引导，都会影响到人工智能的进路和人类的未来。因此，我们应该更深入地理解乃至进入古代东方求真者的意识经验，特别是那些能够加深和延长人的时间视域，从而开启人的自觉的、自然的而非仅仅规范的和强迫的道德智慧的经验，比如修行瑜伽、佛家禅定、道家炼气和儒家孝悌修身的经验，并使之与人工智能内在关联起来。这也就是说，以时间化、道德化修行经验来引导人工智能的研究方向，同时以人工智能的新境界来巩固乃至启发人类修行经验，就像现在AlphaGo Zero能够启发柯洁这样的围棋高手一样。简言之，要用人类的仁爱、慈悲和时间化智慧来驯化人工智能，而不能让畸形发展的人工智能驯化人类。

要用人类的仁爱、慈悲和时间化智慧来驯化人工智能，而不能让畸形发展的人工智能驯化人类。

张祥龙，纽约州立大学布法罗分校哲学博士，北京大学教授，曾在山东大学、中山大学任教。研究方向：现象学、儒家哲学。已出版著作《从现象学到孔夫子》《西方哲学笔记》《家与孝》，发表文章《想象力与历时记忆》《技术、道术与家》等。

---

[1] 本文的部分内容与拙文《人工智能与广义心学——深度学习和本心的时间含义刍议》（刊于《哲学动态》2018年第4期，第13—22页）有重叠之处。

[2] 尤瓦尔·赫拉利，林俊宏译：《今日简史》，中信出版社，2018年。

[3] 尤瓦尔·赫拉利，林俊宏译：《今日简史》，中信出版社，2018年。

[4] “线性与非线性的一个明显区别是叠加性是否有效。在一个系统中，如果两个不同因素的组合作用只是两个因素单独作用的简单叠加，这种关系或特性就是线性的。反之，如果一个系统中一个微小的因素能够导致用它的幅值无法衡量的结果，这种关系或特性就是非线性的。相应地，具有叠加性的系统，是线性系统；反之，则属于非线性系统。混沌理论、分形理论等非线性研究指出分叉、突变、对初始条件的敏感依赖性、长期行为不可预见性、自相似性等，是非线性系统共有的性质，分数维、标度律、普适性等是对非线性系统做定量描述的普适概念。自然界中大量存在的是非线性相互作用，非线性特性不是细枝末节，而是基本特征和本质所在，线性系统只不过是一部分简单非线性系统在一定条件下的近似。”（引自《百度百科》“线性与非线性”条目。网址是：https://baike.baidu.com/item/线性与非线性。）

[5] 亚里士多德：《物理学》，商务印书馆，2016年，第4章第11节、14节。

[6] 亚里士多德给出了多个对人的定义，其中影响最大的是“人是有理性（logos）的动物”。因为“logos”在亚里士多德的著作中意味着“理性”，“logos”又指“会说话的”，所以此定义也可被表达为“人是会说话的动物”。有关讨论和出处参见汪子嵩等著：《希腊哲学史》第3卷上册，人民出版社，2003年，第595页。

[7] 埃德蒙德·胡塞尔，倪梁康译：《内时间意识现象学》，商务印书馆，2010年，第448、485页。

[8] Ludwig Wittgenstein, G. E. M. Anscombe (trans.), Philosophical Investigations,Blackwell, 1963 (1958).

[9] Ludwig Wittgenstein, G. E. M. Anscombe (trans.), Philosophical Investigations,Blackwell, 1963 (1958).

[10] 《〈梨俱吠陀〉神曲选》，巫白慧译，商务印书馆，2010年，第22页。

[11] 释印顺：《唯识学探源》，载《印顺法师佛学著作全集》第四卷，中华书局，2009年，第3部分第55页。

[12] 《成唯识论校释》，玄奘译，韩廷杰校释，中华书局，1998年，第688页。

[13] 法身指真实的、活的现象和意义单元。

[14] 离色身指被对象化并执着于的现象。

[15] 《〈坛经〉对勘》，郭朋校勘，齐鲁书社，1981年，第35—36页，法海本第17节。按校勘后文字呈现。

[16] Charles Menzel，Progress in the Study of Chimpanzee Recall and Episodic Memory，H. S. Terrace and J. Metcalfe (eds.), The Missing Link in Cognition Origins of Self-Reflective Consciousness, Oxford University Press, 2005, pp.188224.

[17] 伊恩·古德费洛、约书亚·本吉奥、亚伦·库维尔，赵申剑等译，张志华审校：《深度学习》，人民邮电出版社，2017年。

[18] 伊恩·古德费洛、约书亚·本吉奥、亚伦·库维尔，赵申剑等译，张志华审校：《深度学习》，人民邮电出版社，2017年。

[19] 伊恩·古德费洛、约书亚·本吉奥、亚伦·库维尔，赵申剑等译，张志华审校：《深度学习》，人民邮电出版社，2017年。

[20] 约书亚·本吉奥，马骁骁译，封举富审校：《深度学习：人工智能的复兴》，载《环球科学》2016年7月号（第127期）。

[21] 伊恩·古德费洛、约书亚·本吉奥、亚伦·库维尔，赵申剑等译，张志华审校：《深度学习》，人民邮电出版社，2017年。

[22] 伊恩·古德费洛、约书亚·本吉奥、亚伦·库维尔，赵申剑等译，张志华审校：《深度学习》，人民邮电出版社，2017年。

[23] 约书亚·本吉奥，马骁骁译，封举富审校：《深度学习：人工智能的复兴》，载《环球科学》2016年7月号（第127期）。

[24] 约书亚·本吉奥，马骁骁译，封举富审校：《深度学习：人工智能的复兴》，载《环球科学》2016年7月号（第127期）。

[25] 约书亚·本吉奥，马骁骁译，封举富审校：《深度学习：人工智能的复兴》，载《环球科学》2016年7月号（第127期）。

[26] 伊恩·古德费洛、约书亚·本吉奥、亚伦·库维尔，赵申剑等译，张志华审校：《深度学习》，人民邮电出版社，2017年。

[27] 伊恩·古德费洛、约书亚·本吉奥、亚伦·库维尔，赵申剑等译，张志华审校：《深度学习》，人民邮电出版社，2017年。

[28] M. Tegmark, Life 3.0—Being Human in the Age of Artificial Intelligence, Alfred A.Knopf, 2017, p.78.

[29] 参见百度百科“阿尔法围棋”条目。

[30] 约书亚·本吉奥，马骁骁译，封举富审校：《深度学习：人工智能的复兴》，载《环球科学》2016年7月号（第127期）。

[31] 一位讲座听众告诉笔者：现在有研究发现，人的意识能力除了与大脑，还与身体的其他一些部分比如腹部的神经结构相关。

[32] 伊恩·古德费洛、约书亚·本吉奥、亚伦·库维尔，赵申剑等译，张志华审校：《深度学习》，人民邮电出版社，2017年。

[33] M. Tegmark, Life 3.0—Being Human in the Age of Artificial Intelligence, Alfred A.Knopf, 2017, chs. 4-5.

[34] M. Tegmark, Life 3.0—Being Human in the Age of Artificial Intelligence, Alfred A.Knopf, 2017, p.283。

回到“轴心时代”思考人工智能

何怀宏

人工智能是人的产物、文明的产物。它象征人类的物质文明达到某个制高点，但对人类及其文明也构成了重大的挑战。本文试图追溯一种人类文明的历史秩序来思考人工智能，也从中国哲学的角度尝试回答博古睿研究院项目组提出的一些问题。我想回到长久影响人类的精神文明创建时代，即雅斯贝尔斯所说的人类的“轴心时代”来思考人工智能。但这与其说是思考人工智能技术本身，不如说是思考人工智能将带来的挑战，思考人类自身，思考我们的本性、我们的价值观，思考人的进化和异化、人类文明的过去和未来。不过，这种反省的确又是由人工智能等高科技的发展刺激出来的。

本文主要从包含在中华文明中、产生于轴心时代的中国哲学的角度来思考人工智能，但我的确并不认为中国原初精神和哲学思想能给出对现代问题的一个解决方案。大概任何一个文明的思想传统都不可能为现代文明的症候提供一种灵丹妙药，但又都含有对这一危机中某种因素的解药——假如我们能认真思考古典的思想。

回到轴心时代来思考，也就是回到人类最早的精神资源、回到较早的人类来思考。“后人类时代”的概念被提出，而我们对“最后的人”的思考也最好结合对“最初的人”的思考来进行。这里的“最初的人”是指最初的文明人，尤其是最初系统深入思考的人。

文明与中华文明

观察人类文明的进程，我们可以注意到在轴心时代精神文明产生之前的某种“不约而同”和之后的“有约而异”。

在各个文明产生系统的精神文明之前，尽管它们在地域上互相分离，甚至完全不知道其他文明的存在，但它们之间具有相当多的共性。

文明大致有三层结构，即物质文明、政治文明与精神文明。这种结构看来也是一种历史的衍生次序，虽然并不是所有的文明形态都能比较完善地达到这全部的三层。在近代之前，人类文明也并没有在全球连为紧密互动的一体，但我们还是可以从整个人类历史的角度观察。

自人猿揖别的人类历史迄今已有250万～300万年，而人类文明的历史只有约一万年。

最初的发展总是缓慢的。进入文明之前的各个原始人群虽然也有一些洞穴壁画、小雕刻等艺术的闪现，但主要的问题还是如何生存下来，他们为了充饥，必须全力以赴、全员出动。

人类走向文明不能离开火。火给原始人带来了温暖和光明，带来了熟食，从而促进了人的大脑发育和意识成长。而最早的文明不能离开水。文明自农业始。人类从以狩猎采集为主转向以种植畜牧为主获取食物和能量，而人需要喝水，畜养的动物需要喝水，种植更需要大量的水。最早的人类四大文明都是不约而同地分别从大河流域发展起来的：西亚幼发拉底河与底格里斯河的两河流域的巴比伦文明、尼罗河流域的埃及文明、印度河与恒河流域的印度文明以及黄河长江流域的中华文明。

而人类又花了一万年文明史的差不多一半时间来各自夯实自己的物质基础。

当人口和财富密集化积累到一定程度时，政治秩序的产生就成为必要，而剩余产品的丰富和积累让一些人能够腾出时间来专门“劳

心”，这也使政治秩序的发明和治理成为可能。到公元前3500—前3000年，在西亚两河流域的美索不达米亚，苏美尔人创造了最早的有大量泥板文字可以印证其存在的城邦国家。后来又有埃及王国的建立。中国在传说中的五帝时代也在走向国家，而到三代时期国家的形态已经趋于完整。

到公元前500年前后，在几个主要文明中，又出现了如雅斯贝尔斯所说的“轴心时代”的富有创造性和长久影响力的精神文明，包括两个主要宗教性的精神文明——西方的犹太教和东方的佛教，以及两个主要人文性质的文明——西方希腊城邦的思想文化和中国先秦以儒家为最突出的诸子百家。

以上物质文明的成果在今天大多被消耗了、遮盖了，其粗放的技术也被远远抛到了后面。政治文明制度的基本因素虽然保留下来了，但任何一个曾经烜赫一时的王朝、国家、帝国也都化作烟云了，只有这几大精神文明对后世产生了巨大和深远的影响，分出了不同的文明源流。

希伯来先知开启了一种一神论的宗教信仰传统，在后来的2 000多年里，这种信仰传统不仅维系了一个民族在多次亡国甚至可能亡族的危机中的延续，而且使自己的民族至今仍然能够生存下来，甚至重新建立一个国家。但更重要的是它对后来的基督教、伊斯兰教所产生的影响，不仅从经典语义上而且从思想精神上给了后来的基督教和伊斯兰教以重要的资源。佛教虽然在印度后来的历史发展中受到了限制，但它对世界包括对中国产生了广泛的影响。希腊城邦的思想文化在罗马、拜占庭时期继续传承，到近代又重新复兴，参与塑造了现代的西方文化。而中华的精神文明不仅持续不断地在中国的土地上延续和发展，而且影响了朝鲜、越南和日本。

在物质文明开始的阶段，各个文明之间的差异看起来并不是很大，初始政治文明的许多要素看起来也是共同的。各个文明之间的差异大致产生于精神文明出现的阶段，不同的信仰和价值观在引领不同

的文明向不同的方向发展。这不仅是指精神文明的差异，而且是指因为精神价值取向的不同，最后物质生产与政治制度出现了很不一样的面貌。

所谓系统的精神文明出现之前的“不约而同”或者说“相隔而同”，这些相同或相近有空间和地理上的，比如都靠近河流，也有心理上的，比如那时的人们心中都有神灵，这是传承自文明之前的原始人的。心中有神灵的人处处都感受到自己的限度，产生某种敬惧，感觉有无数有别于人与超越于人的存在，知道自己的知识和能力是很有限的。还有发展时间上的“不约而同”，这也许反映了某种文明自然生长的节律。

所谓系统的精神文明出现之后的“有约而异”，首先是指各大文明的精神创造各自有着自己的特性，或从崇拜多神走向信仰一神，或从更相信神灵走向更偏重人文——但无论古代希腊罗马还是古代中国，都还保留了神灵或至少为世俗的、被统治的多数保留了神灵。其次，这种在精神上的发明创造，也和一部分人群有了一种稳定的精神上的联系和契约。这种契约是一种内部的契约。这些人不再是血缘、物质和政治上的联系与聚合了，还有了一种精神上的聚合。这就使各个文明的随后发展呈现出相当不同的特征，有了不同的方向。当然，这种精神上的影响并不是即刻起作用的，而往往要经过漫长的时间，但一旦被确立，也能起到长久的根本作用。

我们的确得承认，精神文化需要有它的物质基础。从时间或演进的角度而非从空间和地域的角度定义文明的类型也主要是通过物质的文明。从物质的、可见的形态看，人类文明主要有两大类型：农业文明与工业文明。农业文明还能容纳过去的群体狩猎采集。在农业文明时代，也还有亚类型的游牧文明、海洋文明等，它们构成一种流动性，但在规模和性质上无法与农业文明对等。而在工业文明时代，今天也许还可以分出一个高科技文明的阶段。工业文明将过去的物质形态文明清理得相当彻底，不仅像群体的狩猎采集、游牧文明几乎不复

存在，农业文明也被工业化和高科技化并置于一个相对不重要的位置。

但是，无论哪一种文明，它们开始都是与神灵紧密联系在一起的，没有神灵，就不会有文明的诞生。即便是原始的宗教也是精神性的，在没有分化出人文，只有它在传承似乎粗陋的精神的时候。而在分化之后，它也还长期是最重要的精神文化。文明渐渐将“神”与“灵”分开，又将“唯一的神”与“众多的神”分开。近代以来，最初，心中视外界为“万物有灵”的人渐渐演变成了今天自视的“万物灵长”。

我在这里想推荐一个曾经帮助我记忆时间和节律的简要线索：“一半的一半的一半”。从现在上溯，也就是约一万年之前，人类进入了一个“物质文明”的阶段，也就是在几个大河的流域各自进入了农业文明。而这一万年的一半，即在5 000余年前，人类则开始进入一个“政治文明”的阶段，在一些农业文明发达的地区，开始出现了政治秩序，出现了国家；再一半，即在2 500年前左右，一些地区则又进入了“精神文明”的阶段，也就是雅斯贝尔斯所说的人类的“轴心时代”。

最后的一个“一半”则和中国有关，也就是1 250年之前，就是中国的盛唐，即公元750年左右，中国的国家实力和社会富裕程度达到了一个巅峰。但就在几年之后，中国就发生了“安史之乱”。根本不知战争为何物的老百姓像“两脚羊”一样被残酷地杀戮，死了3 000多万人，这或许可以说明：即便在冷兵器和相对于今天高科技的“低科技”时代，文明也是脆弱的。当然，在农业文明时代，文明不可能被完全摧毁。

中华文明在农业的兴起和国家的形成方面或许不是最先的，可能要比其他文明出现大规模的农业和国家稍晚一些，但在精神文明方面，则在孔子之前就已有滥觞，我把它叫作“周文”。它发源于西周代商之际，如文王演绎八卦为六十四卦，周公对政治的反复诫命，尤

其是周礼的形成和完善，已经为儒家和其他派系的人文思想开了先河。

这里也许需要说明一点，在对尧、舜、夏、商的文献记录方面，是经过了周人对古代传说的整理的，其中刻下了周人人文思想的诸多印迹。在五帝直到殷商的真实历史中，神灵对政治的影响大概会比我们今天看到的《尚书》更为强烈。原始人时时刻刻能够感受到神灵，农业文明中的人时时刻刻能感受到“天”的限制。诸种文明中的“最初的人”很难设想在自己的信念中不依靠一种神灵的秩序就能建立一种政治社会的秩序。

中华文明是一种将农业文明发展到极高水平，也将其物质文明基础上的精神文化发展到极致的文明。从制度理念上，我们可以用王国维概括的西周肇始的道德制度之“尊尊”“亲亲”“贤贤”来说明。“尊尊”是对等级差序的尊重，以维系天下的稳定；“亲亲”是对自然血缘关系的亲近，以保障社会的和谐；而“贤贤”则是对德行和才能的奖掖，以刺激社会的活力。

近代以前，除了地中海世界的几大文明之间有较密切接触和交融，其他文明都是在相对独立的状态中发展的。近代以降，人类文明终于通过地理大发现和全球化融合在一起了，如此就有了文明的密切交流，但也有了文明的激烈冲突。无论如何，现代人类文明还是可以看到一种主流的存在，甚至在不同社会的价值观方面也开始有了一种“趋同”，比如平等主义与物质主义的价值取向。

这是在西方率先出现的潮流。西方曾经吸收希腊罗马人文理性和犹太教、基督教的两种传统，以其追求彻底的性格把握两个极端，宗教精神一度君临一切。但在近代，西方关注的重心则转向人间社会和控物能力了。17世纪初，培根曾感叹此前2 000年人们没有太关注物质方面的求知，他预告了时代朝向物质科学的新潮流和新方向。仅仅过去了400多年，我们就看到人类在科技方面取得的巨大成果，以及科技

参与推动的经济方面取得的巨大成就。工业文明将人类带入了一个以此前难以想象的速度变化并彻底改造了社会生活的新时代。

迅速消逝或正在消逝的农业文明自有其意义。它更接近大地，更接近自然。它没有对物质生活要不断提高的渴求和焦虑，没有对技术日新月异的惊喜，没有经济崛起的奇迹，但也常常是不愁温饱、丰衣足食。人除了会关注物质生活，也会关注其他东西，关注一些可能更符合人这一身份的方方面面，比如家庭的亲情、邻里的关照、手工技艺的精湛、商业口碑带来的满足、艺术的欣赏和创造等。即便是在农业社会，中国也还是有精美的物品，既有亭台楼阁、诸子百家、唐诗宋词，也有市井的繁荣、小康甚至非常精致的生活。在西方也有美妙的音乐、壁画和宏伟的大教堂。农业文明节制了人们的物欲，让社会保持了一种精神性，也为人类的生存保持了一种长久之道。

中国哲学的相关思考

传统中国哲学是基于农业文明和农业生活方式的。孟子的“为民制产”描绘了这样一幅生活图景。士人们多来自乡土，也回到乡土。下面我想说说中华精神文明的诞生时刻，即春秋战国期间的几位儒、道两家的代表人物，谈谈他们与人工智能的相关思考，也分析他们之间的一些差异。这种思考主要是围绕着天人关系、心身关系展开的。

“天”在中国哲人那里有两解：一是冥冥中的超越之“天”；二是自然之天，指整个自然界。这两者又是常常联系在一起的。人文主义的儒家和道家都不敢说知晓那超越之“天”究竟是什么。中华文化并没有发展出一个独立的、明晰的宗教体系，但这“天”是存在的，敬畏是存在的，它以“天意”“天命”“天道”在发挥着作用。无论哪一种“天”，都构成对人的作为的限制。人并没有无限的可能性、无限的可完善性。人类的历史、文明也并没有无限的可能性。人面前并没有无数的道路可以选择，甚至有时只有两条路可以选择。人要自

知其有限，最好是顺应自然、仿效自然，不破坏事物的自然过程。道家甚至希望人尽可能地回归自然、融入自然。

儒、道两家都讲顺天，顺应自然，但道家尤其是庄子还反对人伦教化乃至政治秩序。在天道方面，道家也主要取自然的收敛之道，甚至从源头上就收敛。儒家还取自然的生长之道，也讲究人伦教化，但这种人类也是“自然的”人伦——儒家最重视的是基于人伦血缘的“亲亲”关系。总之是尊重自然，没有征服自然的打算。“天人合一”即便说目的是在顺天为人，途径也还是顺天、敬天。这和现代高科技试图改变甚至颠覆事物的自然进程完全不同。

农业文明肇始的时代，人们使用的工具相当简单，常常只是身体器官的自然的、可见的延长。即便后来人类取得了很多进步，但人类对自然的改变也还是有限的。虽然种植和畜牧不再是狩猎采集现成的自然物了，各种被种植的植物和被畜养的动物的属性也会慢慢发生变异，人类从自身的利益出发认为它们应该“择优而存”，但这种变化没有改变太多事物的自然进程。

除了外物，中国先贤也常常感受到存在于自身的“物”——肉身。中国哲学是一种强调身体及其感受性的哲学，但它并不是要去努力满足身体的各种感官欲望，而毋宁说它是要认识到人因其身体的感受性而具有脆弱性和有限性的特征，人才能追求符合人之为人的生活。作为碳基生物，人的身体有两方面的意义：一是有肉身就需要物质的营养，就会产生各种欲望，除了基本的“食色”，还有更高的“权钱名”等；二是有了这种有限性，人才会渴望无限，也使基于这种有限性的德行和艺术美焕发出特别的光辉。

我们就以儒家最看重的一种德行“孝”为例。人为什么要孝敬？从身体的角度看，一是我们和我们的父母、我们的祖先有一种身体或者说血缘上的联系，“身体发肤，受之父母，不敢毁伤，孝之始也”。二是人刚生下来时并不能够独立生存，需要父母的抚养，孔子说三年守丧，因为父母至少要含辛茹苦哺育你三年，你才能“免于父

母之怀”。即便孩子稍微能够生活自理，而后父母还要花费大量的抚养和教育时间。所以，孔子对弟子说，如果你连三年丧都不守，你觉得安心，就这样去做吧，我无法安心。三是，人都是要老的，老人也会衰弱，进入难以自理的年龄，必须“有养”，需要儿女的照顾。而且他们不久就可能告别人世，父母年高，“一则以喜，一则以惧”。人都是要死的，不要等到“亲不在”，不能孝敬奉养才意识到这一点而后悔莫及。“慎终追远”，这种孝敬既是一种属于人的精神感情，如孔子所说，“至于犬马，皆能有养；不敬，何以别乎”，但又是基于身体感受性的。

中国哲学是一种强调身体及其感受性的哲学，但它并不是要去努力满足身体的各种感官欲望，而毋宁说它是要认识到人因其身体的感受性而具有脆弱性和有限性的特征，人才能追求符合人之为人的生活。

儒家对人的有限性以及多数的有限性是有充分认识的，人要受到自然的限制，受到“天命”的限制，多是来自对人的自然本性的认识，而这种认识本身又是通过身体对外物的感觉和对原则的直觉，也就是通过“体知”或者说“体悟”获得的。

人的感受性也就是脆弱性。这种感受性表现为：一是人不能直接承受打击，小的打击造成人身体疼痛或残损，大的打击甚至会使人身亡。二是人必须不断摄取营养才能生存，而人最后也逃脱不了死亡。换言之，这种身体的感受性也是一种脆弱性，但正是这种脆弱性催生了政治秩序，催化了德行和艺术。这种脆弱性也帮助人们意识到自己的有限性。文明也就建立在这种脆弱性的基础上，人类文明本身不管看起来多么强大、多么了不起，它实际上还是脆弱的。我们可以考虑特别短暂和有限的东西如何与特别长远和无限的东西联系在一起，特别肉身的东西如何与特别精神的东西联系在一起。正是因为人的脆弱性和有限性，人们才努力在家族的繁衍、艺术和宗教的创造中去追求永恒与无限。

道家也强调人有身的局限性。老子说，“吾所以有大患者，为吾有身”。有身就会有欲望，而欲望不断增长，就会带来一系列问题，导致再多的东西可能也难填入的欲壑，所以，不如从源头处就加以遏制，鲁迅曾经说，“老子书五千语，要在不撄人心”。老子希望不刺激人们的欲望产生，甚至感觉不到自己的身体，恍若“无身”，“及吾无身，吾有何患”，使民含德如“赤子”，小国寡民，老死不相往来，绝学弃知。庄子也说，“吾生也有涯，而知也无涯，以有涯随无涯，殆已”，不如从一开始就不去追求，恍若“无我”。道家采取了一种看起来比儒家更彻底的节制物欲的办法，当然，这种办法看起来也将遏制文明的发展。但是，具有这种身体及其有限性的根本性意识，却不无一种由“知始”而“知止”的意味。

道家想从源头上防范后果，老子并不倡导废弃政治文明，但希望人类回到“小国寡民”的状态中去，庄子则试图使人类回到原始时代，使民“同与禽兽居，族与万物并”。当然，这里还是有多数和少数的分别：儒家讲圣人与众人、君子与小人的分别，但两者并不在出身上有所固定，而是保持一种上下的流动性。道家希望多数人回到人的懵懂状态、婴儿状态甚至动物状态。但其实又有极少数人（或是政治统治者，或是文化精英）不是懵懂的，而是有意为之，（老子认为）他们背负某种精神的重负或者（庄子认为）他们获得了精神的自由。道家看来也并不是让所有人都等同于或接近于无知无识的自然物。在老子那里，统治者还是要背负判断和治理的负担；在庄子那里，除了绝大多数浑然于其他动物的人，还有极少数通达甚至拥有某种超自然能力的真人、至人。

儒、道两家思想的核心关注点都不是人类社会的不断繁荣之道、富强之道，而主要是考虑其长久之道。或者说，儒家更多的是考虑文明存续的长久之道，道家主要是考虑人类存续的长久之道。道家担心“物壮则老”。儒家是现实的，知道文明不可能倒转，文明也有文明的意义。道家是回溯的，但在精神状态上也向往一种自由逍遥的真人。儒家更重视人际关系，道家更重视人与物的关系。儒家文化中并

没有多少解决今天人工智能和其他高科技可能带来的危险的现成答案，它一直关注的是人际关系和人类社会。倒是先秦道家在这方面想得较多，身处乱世，他们看到了精英对强力的崇拜，以及其背后更广泛的人们对物质的欲望。他们不像儒家那样更关注对强力的驯化，而是关注更为根本的东西，即对物质欲望的源头控制。道家对物欲有更多的压抑，或者说是从一开始就压抑——最好不产生这种物欲，不提供容易产生这种物欲的条件和环境。但两家都相当充分地认识到人的有限性，尤其是社会的有限性、多数的有限性。他们还希望在没有神灵的情况下达到少数个人的完善——前者是成为圣人，后者是成为至人。

中国传统的人文理性是伟大的，传统的智慧敦敦告诫我们不要随意改变事物的自然进程，不要放任物欲。

或许人们可以批评说，仅仅做到人文理性可能是不够的，仅仅认识到自然之“天”的限制是不够的，还要有神圣之“天”，在儒家那里也并不是没有这层含义，先人也有这种传统，但还是稀薄的一层，不是对一种超越性存在的明确认知。而这种对超越性存在的肯认也是有不同形式的。人们可能依旧不能确知这一超越性的存在究竟为何，但应该相信它一定是存在的，并努力寻找。对于超越于人的无限和永恒的寻找就意味着我们对自身的有限性有了认识。反之亦然，因为我们知道了自己的有限，才有了对无限的渴望和追求。

人工智能与人类未来

从其起因来说，轴心时代产生的主要思想都是“忧思”。佛教的思想是因为释迦牟尼看见人的生老病死所产生的忧思；希腊思想是苏格拉底、柏拉图等哲人看见雅典衰落所产生的忧思；犹太先知的思想是先知看见犹太人颠沛流离所产生的忧思；中国也不例外，春秋末年到战国所产生的那些最有力的思想也是我们的先人看见礼崩乐坏、战

乱频繁所产生的忧思。对于如何化解这种忧思、解决所面对的问题却有了不同的路径。轴心时代产生的主要思想的另一个共同特点是：它们都不是以满足人的欲望，尤其不是以满足人的物欲为目的的。从影响与结果来说，轴心时代的各种精神文明产生之后，此后2 000年各大文明的主要关注点还是在精神，而不是在物质，是在人的精神自控和超越能力，而不是在人的控物能力。

中国古代的士人将“先天下之忧而忧”看作自己的使命。今天，我们看一些现代人的忧虑或可分为三个阶段。第一阶段是对人类精神文化衰落的忧虑，这是在近代以来就开始的。现代人越来越多地关注物质层面生活的改善，人们忧虑精神生活是否会往上走而不是会往下走？第二阶段是对人类文明的忧虑，这在20世纪就开始表现得较为充分了，尤其是经过世界大战，出现核武器等大规模杀伤性武器之后，人们忧虑新的世界大战是否将把人类“打回石器时代”，人类近万年积累的文明成果还能否延续？第三阶段则是在人工智能、基因工程等高科技出现之后，对人类这一物种可能将被取代，还能否存续的忧虑。而人类是否存续，文明是否存续，是否将取决于人类能否重建一种精神秩序，包括重建人类与一种超越性存在的精神联系或契约？

我们还是回到对人的理解和定义：人是什么？人能何为？在直接讲到人性的时候，孔子只是简单地说“性相近，习相远”，“性相近”则意味着人有一些共同或接近的本性，即便具体落实到每个人身上有些本性的差异，也还是能够概括出一些相同的因素。但不同人的作为，不同人的后天习性，不同人的“盖棺论定”，以及由此带来的不同地方尤其是不同政治社会的风习可能是相差甚远的，这其中有不同社会的制度、组织和文化的作用与影响。一方面我们要看到环境影响的重要性；另一方面我们要看到之所以存在善恶差异，还是因为在人的本性中一开始就有了善端和恶端。这两端的差距在“人之初”那里是不明显的，但如果没有两个始源，也就难以解释后来的善恶分流。

即便是大力主张人性本善的孟子，我们也能注意到他是在“善端”的意义上谈人性本善的，他所采用的“孺子将入于井”的例子是很基本、很低调的例证。而且，他强调人与其他动物的区别其实是“几希”，就那么一点点。“君子存之，庶人去之”，他强调人是多么容易放逸，丢失本来的善心，所以人要非常努力地求其“放心”，努力去成为一个“君子”。

而这说明什么呢？说明儒家虽然意识到人有普遍的善端，但也意识到人存在一些基本的限制，人的道德能力不可能普遍上升得很高，甚至只有较少的人能够比较充分地摆脱物欲。如此也就不能指望人性自身被改造和提高到接近于圣人的程度，不能指望所有人都极大地提升自己的自控能力。如果人极大地提升自己的控物能力就可能会成为问题。

我们再回到人的定义。历史上对“人”的定义或是侧重身体及其控物能力，或是侧重精神创造与超越能力，且常常侧重人与动物之别。侧重身体及其控物能力的定义如人是直立的、脑容量大、两足无毛的动物，人是能够制造工具的动物等；侧重精神创造与超越能力的定义如人是有意识、有精神、有自我的动物，人是有价值观和正义感的动物等。我们还可以再仔细分析一下这些定义。

当我们从人与其他动物相区别的本性角度说到“何以为人”的时候，大概可以包含三种特征。

第一，人是能够制造和运用工具的动物。人有理性，会计算，有逻辑和抽象能力，能够自我学习，计划自己的行动，有明确目的和手段的意识。此外，人能够利用一切可能的工具，这意味着人有控物能力。这类似于帕斯卡所说的“几何学精神”，或者说类似于今天我们所说的“计算理性”“工具理性”“技术理性”。

第二，人还有工具理性之外的精神能力和意识，这就像帕斯卡所说的“敏感性精神”。人还有默会的知识、情感、意志、直觉和信仰等，这意味着人有自控和超越能力，也有一种精细的、基于身体感觉

的控物能力。另外，人还有直觉，这种直觉不是对外物的知觉，而是对一些抽象原则和知识的直接知觉。人还有道德感、正义感，这种感觉也不仅是为自己所生发，同样来自对他人和社会受到的不公平的反应。人是道德的动物，这并不是说人就是天生合乎道德的动物，而是说人是能够在善恶正邪这些问题上敏感、有反应的动物，是能够在这方面进行自我评价和对他人评价的动物。人在有些感知和身体能力方面（如飞行、潜水、速度、力量等）并不如某些动物，甚至人的这些能力还有退化的趋势；但人有理性、精神和道德感，就能够建构社会政治秩序，发展出物质文明和精神文明。当然，人在这方面的发展也不是无限可完善的。

第三，人还有一种综合性的自我意识、主体意识，能够将人的各种理性和非理性能力与意识综合在一起，形成自己毕生的计划和连贯的行动，确立自己的人格。

以往对人的定义主要都是将人和其他动物比较，但如果和新出现的智能机器比较，则还要加上一种特征。

第四，人是一种碳基生物。过去诸如人是两足无毛的动物、人是直立行走的动物、人是脑容量较大的动物等定义，虽然是着眼于人与动物的区别，但也都是从身体方面来定义人的。动物仅仅在这第四种特征上与人相似，动物的身体与人的身体虽然有重大差异，但人和其他动物一样，也是碳基生物。人和其他动物都有感觉，都需要不断补充物质的营养，都经受不起外力的重击。人和其他动物的身体及其特征都是千百万年自然演化的产物。

智能机器则仅仅在第一种特征与人相似，即拥有高超的计算理性、工具理性，能够自我学习、深度学习，而在这方面，它的控物能力在未来还可能超越人类。但是，智能机器目前还没有工具理性之外的意识和能力，没有综合的自我意识，也没有碳基生物的身体。

那么，未来的智能机器会不会获得上述第二种甚至第三种特征，或者说人类会不会推动这一过程，比如试图让智能机器变成一种结合

第一种特征和第二种特征的“类人智能”。但第二种特征和第四种特征密切相关，智能机器大概能够做到拥有人类面容、人造皮肤，甚至拥有仿人的神经网络，但不可能也无须有与人类一样、与动物一样的碳基身体。这身体也正是人类的弱点。虽然人类的第二种特性或者说人类更高或更特别的精神意识能力正是基于这一弱点而存在的，还有一些像“庖丁解牛”那样的需要身体感觉的手工技艺也是基于此练就的，但智能机器要发展它的控物能力，看来也不太需要这类技艺。目前，人类对机器改进的方向的确也只是想提高它们的智能，而不是想让它们产生爱的能力、追求美的能力、拥有信仰的能力，且看来也几乎无法做到这些，因为这些和漫长的自然演化产生出来的身体以及人心中不知缘何产生出来的灵性有关。而且，在人类目前还能控制和掌握智能机器的前提下，在人类完全能够轻易地自然繁殖的情况下，从人类的利用厚生的角度看，人类也没有这种让机器在身体上也完全变成人的需求，且不说有没有可能。

从人的角度看，将第一种特征和第二种特征相结合的更有可能的路径是在人的身体内不断植入硅基体：换上人造肢体、植入人造心脏等。就像库兹韦尔所希望的，人类最后也获得和机器一样的（可以不断更新器官意义上的）长生不老，即一种硅基“身体”。但如果这条路能走得通，就意味着人的物化、人的硅基化了，人类也将失去前面所说的那种基于碳基身体、基于人的有限性和中间性的精神特征与创造。

至于要回答人是否要给智能机器植入人的价值观的问题，我们可能首先要问，人类现在的主流价值观是什么？我们可以重温庄子那段著名的论述，那位不愿使用杠杆的为圃者说：“有机械者必有机事，有机事者必有机心。”今天的“机事”是大大发展了的，那么人类的心灵是否也变成“机心”了呢？何谓“机心”？大概也就是便利之心、利用之心、控物之心。这种“机心”既是“机事”发展的动机，又是“机事”发展的结果，它们互相强化。现代人当然希望机器能够尽可能地方便人类，为人类造福，前提是机器能够置于人类的控制之

下。但机器一旦获得自我意识，按照同样的逻辑，它不是也可以甚至应该利用和控制人类吗？

那么，问题就变成了人们其实是想把一种无论人类自己追求什么，机器都应该对人友好、忠心耿耿地为人类服务、绝不伤害人类的价值观植入机器。虽然可以说这是自私的，是一种“主人－工具”甚至“主奴”的价值观，但这也是能够被理解的。且不说这种价值观输入的可能性，它总是要冒这样的风险：假如机器真能获得一种价值观，它也就可能获得一种自我意识。这时，价值观的实际指向的结果逆转只需要一个主体的转换。

假如机器真能获得一种价值观，它也就可能获得一种自我意识。这时，价值观的实际指向的结果逆转只需要一个主体的转换。

的确，如韦伯所说，经济和物质是基础，但精神信念或者说价值观是主导。物质或者说经济、科技是文明进化的基础，但价值观决定文明进化的方向。现在的问题是：物质文明是人类文明的初始，但物质文明越过某个顶峰会不会也是人类文明的终结？世界会不会进入一个“后人类时代”，今天的人类会不会成为“最后的人”？我们人类这样一种造物，是否也让自身处在一个新物种的“造物主”的地位，从而开创这一新物种的“创世记”？像超级智能机器是否将把人类推到物质文明的顶点，但又将人类抛入深渊？

物质是基础。文明自物质始，但会不会也将自物质终？人类的文明从物质文明的奠基开始，会不会也以冲上物质文明的顶峰结束？一般来说，人首先要吃饱饭，有剩余，也得到了对生命和财产的基本保障，其次才能从事语言文字和思想精神的创造。人类文明从物质开始，从农业开始，无论什么时候，物质生活和身体生存都是最基本、最优先的前提或者说目的。但今天的文明将物质生活和身体生存这最基本的目的变成了主要的甚至最高的目的，人类文明会不会也将在这种对物质不断增长的欲望以及均富的追求中结束自己？

的确，现在还不是物质文明的顶点，但人类似乎不可遏制地想要更多，想要冲向顶点。然而，人性并不可能有根本的改变，人的道德能力并不会随着其控物能力的提升而大幅提高，于是，在人类的控物能力和自制能力之间将产生巨大的差距，越来越不相称，这将带来一种危机四伏的局面。科技发明了大规模杀伤性武器，除了核武器，还有被更为隐蔽地储存着的或者可以随时被生产出来的化学和生物武器。而基因工程、人工智能等也随时可能给人类带来危险。但科技带来的经济上的极大发展与物质生活的充裕和舒适，则让人不仅感觉不到危险，而且在这种物质生活中自得其乐。科技的发展也极大地提高了人们的自信。在一种自得其乐和自信自满中，人们将很难认识到自己的有限性，很难会去渴望一种标识出人的有限性的超越性存在。人生于忧患，死于安乐，而未来的人类会不会以一种“安乐死”的形式结束自己的人生？

科技的发展也极大地提高了人们的自信。在一种自得其乐和自信自满中，人们将很难认识到自己的有限性，很难会去渴望一种标识出人的有限性的超越性存在。

或许人们会斥责这些忧虑过于悲观，其实悲观到某一极致之后就达观了，这就是中国古代先贤所说的“尽人事，听天命”。而且，人也可以有另一种日常生活中的有用达观：活着，不多想，好好度过每一天。

当然，人总是有选择的意志的。假如说一个人的一生有两种选择，一是快速与短暂的辉煌，二是慢速与长久的平淡，人们会选择哪一种人生？可能会有一些人宁愿选择在短暂的生命里活得轰轰烈烈，也不愿平淡地度过长寿的一生。但选择后一种人生的人会不会更多？问题是人类的选择与个人的选择不一样，在人类历史中，只要有一些关键的少数做出了选择，大部分人可能就不得不跟着走了。

近代以来，人在控物能力方面的发展足够卓越，成就足够伟大，但从总体来看，人类精神文化的衰落已是事实，或者目前来看似乎已

成定局，人类未来的命运则尚不知晓。高科技或许剥夺了一些这样的机会——通过并不剧烈的灾难让人类改弦易辙。但我们是不是还可以有一些温和的期望，即我们也许可以有意识地推动一些高科技帮助我们部分地回到大地，回到对自然的清新感觉，恢复对劳动的美好感受，促进工匠工艺的发展，促进人类获得默会的知识。

人类几百万年都处在前文明的时期，相比之下，人类的文明史只有约一万年。一万年久吗？从个人生命来看是极长久的（所以总有人会期望“万岁”），而对一个物种来说，则是太短暂了。而这一万年的文明史又有约9 500年都是处在农业文明时期，只有最近的500年是处在工业文明和高科技文明。与生活在一个社会动荡乃至战争动乱的时代的人们相比，生活在一个经济和科技快速发展的时代的人们是幸福的，甚至常常能感受到惊喜。即便对此感觉不那么良好的人，也还是有认识到这种巨变的幸运感，但是，这种人不免要忧虑人类的后代：后人会为这种前人的幸运付出沉重的代价吗？

我们也许需要回到最初创造精神文明的人，回到最初那些系统地对人进行反省和思考的人来重新思考人的本性和命运，即为了避免单一方向的人为进步的终点而回到起点。

正是基于上述种种考虑，我们也许需要回到最初创造精神文明的人，回到最初那些系统地对人进行反省和思考的人来重新思考人的本性和命运，即为了避免单一方向的人为进步的终点而回到起点。除了思考人类的繁荣富强之道，我们也要思考人类及其文明的长久存续之道。

何怀宏，北京大学哲学系教授，2019—2020年博古睿学者。主要研究伦理学及社会史等。著有《良心论》《世袭社会》《选举社会》《道德·上帝与人》《生生大德》《新纲常》等。译著有《正义论》（合译）和《无政府、国家与乌托邦》《沉思录》等。

人工智能，吾与也

姚中秋

西方学界关于人工智能尤其是强人工智能的讨论，固然多是兴奋之情，但在其中也弥漫着恐惧的情绪。他们担心人工智能的强大能力会导致社会失调，最大的担心是“奇点”的到来，即人工智能具有自我意识，借助其远超人类的能力反制人类，甚至毁灭人类。

这些担心有依据吗？人类真的有必要恐惧人工智能吗？很显然，“奇点”还未到而且距离我们还相当遥远。西方人士目前对于人工智能的恐惧情绪起于其所想象的未来情景，也就是说，恐惧源于其对人工智能发展趋势的信念，这甚至都不是趋势预测。信念无关乎事实，而由观念塑造。西方人士对人工智能发展趋势偏于恐惧的信念，形成于其看待人及其所造物的一般观念。这样的观念内在于其文明中，成为“百姓日用而不知”的常识，以至西方人士普遍地如此言说而无人加以反思。但对生活在另一种文明传统中的人而言，则难免“友邦惊诧”。

请注意，上面我反复使用“西方人士”这个主语。确实，当今对于人工智能的研究，在科学、技术和工程等实务领域，中、美都是领军者。然而在人工智能的思想领域，中、美则存在极为明显的差距：美国学界发展出形形色色的理论，相当充分地表达了其对人工智能技术发展的恐惧；中国学界则少有发言，看似并没有那么恐惧。对人工智能及其快速发展，中国人的态度普遍是淡定的。[1]这一有趣的对比，似乎揭示了人类对人工智能发展前景的两种不同信念，此信念由其各自文化观念所塑造。本文追溯中西方对人工智能的不同观点背后的文化渊源，主要分析其关于人与物的关系，人际关系、身心关系的

不同观念，凸显中国人对待万物的“物吾与也”态度，人类对待人工智能的态度是：人工智能，吾与也，即就像爱自然万物一般。人与人工智能，尤其是强人工智能“相与”，则可以各自发挥优势，造就美善秩序。

西人恐惧人工智能之文化渊源

西方人为何恐惧人工智能？因担心其加害于人。依常理，人工智能加害于人，需同时有两个条件：第一，其对人类有敌对情绪；第二，其有加害于人的能力。西方人今日对人工智能的恐惧，看似主要从第二个条件立论。这种推理令人困惑，比人类的能力强，就会伤害人类吗？这两者之间有直接关系吗？唯有当补充上动机，这一恐惧情绪才可成立。也就是说，当西方人思考人工智能时，将其对人类有敌对情绪作为默认项，即给定这一动机，因此认为只要其有足够强大的能力，必定加害于人。

从中国文化的立场来说，我们自然可以反问：人工智能对人类有敌意，确实可以作为默认项吗？西方人确有此恐惧，且相当普遍，则可推想，西方人的确这么想。于是，这种认知本身即构成一种需要人们认真对待的精神现象，或曰精神病症。

我们放宽视野可见，今天西方人对日益富强的中国亦不乏恐惧心理，这一心态与其对待人工智能的恐惧颇有类似之处：因为中国富强，所以中国令人恐惧。这恐怕同样是因为，西方人把国家之间有敌对情感，强者必拥有霸权并毫不犹豫地将其强加于弱者的这些假设视为国际政治的基本预设。一旦西方人给定这一预设，自然会因为中国的强大，而想象中国必定加害于西方；你怎么解释你的善意，对方都不能相信，你的解释只会增加对方的疑虑和恐惧。

由以上两点可见，西方人对自己以外的人、物、异邦之恐惧，简言之，就是对他者之恐惧是深刻而普遍存在的。这种顽固的情绪显然

不利于西方人与他者之间形成良好秩序，因此需要我们认真对待。但西方人因身在其中而难以自知，中国人则因站在其外，可进行相当客观的分析，对此心智予以剖析。以下我将通过解读西方经典，索解西方人关于人神关系、人与物的关系、人际关系这三方面的信念，揭示其恐惧他者之深层心结。

首先，我们来看西方人想象的人神关系，主要以影响最大的犹太教与基督教为例，这种人神关系构成了西方现代人理解人与物的关系之模本。

犹太教和基督教的教义是，神造万物与人，神是全知全能全善的，人对神只能绝对服从。由此，神与人之间是主奴关系：神是主人，人是奴仆。当然，人在神所造之物中，毕竟最为特殊，在《圣经·创世记》中，神已说明："我们要照着我们的形象，按着我们的样式造人。"[2]因此，唯有人最有可能像神一样，比如人也拥有理智。

然而，人的悲剧似乎恰恰由此而起。《圣经·创世记》所记录的三个众所周知的神话都彰显了神对人的怪异态度。

第一个神话：神惩罚夏娃。神造了亚当以及夏娃之后，在东方的伊甸为其建园，"园子当中又有生命树和分别善恶的树"。蛇诱惑夏娃吃分别善恶的果子，蛇说："你们不一定死，因为神知道，你们吃的日子眼睛就明亮了，你们便如神能知道善恶。"神让人像神，具有理智，然而神却禁止人吃那分别善恶的果子。神当然是知善恶的，神本身就是善恶的标准，但神不愿人如神那样分别善恶，拥有自我意志，只想让人绝对服从。

第二个神话：挪亚方舟。耶和华见人在地上罪恶很大，终日所思夜想的尽是恶。耶和华后悔造人，心中忧伤。耶和华说："我要将所造的人和走兽，并昆虫，以及空中的飞鸟，都从地上除灭，因为我造他们后悔了。"神所看到的恶是什么？这个神话也许延续了上一个神

话：神造人和鸟兽，期望其完全服从自己。而人一旦知道分别善恶，固然可以为善，但也可能出于其自由意志而为恶，从而违背神命，而这构成最大的恶。于是，神决定对人施加惩罚：集体毁灭。

第三个神话：巴别塔的故事。最初，全世界的语言是一样的。世人共同决定造一座塔："来吧，我们要建造一座城和一座塔，塔顶通天，为要传扬我们的名，免得我们分散在全地上。"耶和华降临，看到世人所建造的城和塔，大惊失色："看哪，他们成为一样的人民，都是一样的言语，如今既做起这事来，以后他们所要做的事，就没有不成就的了。我们下去，在那里变乱他们的口音，使他们的言语彼此不通。"于是，耶和华使世人分散在世界各地，世人就停工不造那座城了。耶和华变乱了世人的语言，所以那座城名叫"巴别"（意为"变乱"）。神之所以变乱世人的语言，是因为其恐惧人拥有类神的能力，人所要做的事若都能成就，也就没有神的地位了。

这三个神话全都显示出神对人深刻的恐惧以及采取的相应对策。神以自己的绝对意志，照着自己的形象造了人，这意味着人有可能接近于神。但具有绝对意志的神要求人一定要绝对服从，神对人自居为主人。神是造物者，自认为有此资格。但是，人既然是照着神的样子被造出的，则难免有自由意志、有理智，因而竟然努力接近于神，试图拥有神一样的理智和能力。尽管人的努力十分有限，但也足以让神大为恐惧。神将此判为罪，并对人予以严厉惩罚。

这一人神关系是西方文化的本源性隐喻，全面而深刻地塑造了西方人的心智。西方人思考人与人工智能之间的关系，自然会想到神与人的关系：人照着自己的形象造人工智能，自居为人工智能的主人。在这里，人是造物主，但又担心人工智能具有人的能力，甚至超过人，从而背离人，伤害人，由此而有恐惧之情。此处的关键在于，人自居为绝对的造物者，其绝对意志是制造绝对服从自己的奴仆，而奴仆所处的位置决定了其必定会反抗，恐惧即由此而来。

前文所述神惩罚人的三种方式，也指示了人阻止人工智能对人“犯罪”的三种办法：严厉惩罚不服从人的个体人工智能，集体灭绝不顺从人意的人工智能族群，变乱人工智能的语言。如果将当今西方人所设想的防范人工智能的思路略加分析，即可发现无非也是这三种方式。

其次，我们来讨论一下西方人所理解的人与物的关系。

神造人时说：“我们要照着我们的形象，按着我们的样式造人，使他们管理海里的鱼，空中的鸟，地上的牲畜，和全地，并地上所爬的一切昆虫。”此处的管理，英文是“have dominion over”或“rule over”，神给予人支配或统治万物的权力。所以万物是人统治的对象，是单纯的工具，其被造的唯一宗旨是服务于人。

人所造之物也是物，当其无自主意识，则人尽可以对其支配。若的确如西方人所预料，人工智能产生自主意识，则会如何？我们可以大胆地预测，在人主物奴的框架中，人工智能的第一自主意识必为对人的不满和仇恨，由此而有与人斗争的坚定意志，进而采取行动，推翻人的统治。因为有自主意识的主体恐怕不会甘于作为纯粹工具的地位，接受他者的绝对支配或统治。西方人对人工智能的隐忧，正是自居为人工智能的支配者的人之权力欲的自我反转。

最后，我们来讨论一下西方人所理解的人际关系。

神在造人之初，只造了亚当。耶和华神说：“那人独居不好，我要为他造一个配偶帮助他。”耶和华神在亚当沉睡时，取下他的一条肋骨，造成了夏娃。因此，人终将要离开父母，与伴侣连合并成为一体。这个故事塑造了西方人关于人和人际关系的基本信念。第一，人非其父母所生，而为神所造。当然，人为其父母所生是一人人皆知之事实，于是神命令人“离开父母”，这是西方伦理或者说反人伦的伦理之基底。第二，神只造亚当，并把整个世界给他，由此西方人相信，人在本源处是孤独的，世界属于自己一人。第三，神造夏娃，不是因为夏娃本人应当存有，而仅为解决亚当的孤独，由此西方人相

信，他人的存有对我纯粹是工具性的，其本身是无意义的。第四，当人被驱离其父母，则被抛入陌生人中，与异性结合是其中一种，直白地说，人被抛入丛林，必定相互恐惧。此即霍布斯主义的理论基础。解救之道只有一条：所有人服从神。霍布斯所设想的社会契约实为世俗化的神迹，其所设立的主权者就是世俗化的神。

如此关于人际关系的想象让西方人相信，人的目的是实现自身利益最大化。人制造万物，支配万物，同样也是为了实现这一目的。为实现自身利益最大化，人不惜伤害他人，实际上，人很可能是通过伤害他人增进自身利益。人占有和造物，本身即可以扩大其相对于他人的优势，增进其自身利益。而为实现自身利益的最大化，人会毫不犹豫地利用其所占有、制造之物伤害他人。

西方人担心人与人工智能的关系也会有同样效果：人工智能被大量应用，会大幅度提高拥有人工智能者的收益，从而扩大分配差距，造成史无前例的不平等，甚至造成极少数人凭借其强大的人工智能支配世界的格局。这些人利用这种优势，控制他人甚至伤害他人，人类会走向《1984》中所提到的世界。这当然是令人恐惧的。

以上揭明西方人恐惧人工智能之心智根源，追本溯源，可归因于其崇拜有绝对意志且全知全能的神。由此，神与人是造物者与被造者的关系，被造者在被造之前是绝对的无，造物始于造物者的绝对意志，物被造出后，这一绝对的造物意志转为对被造物的绝对支配权。在神所造的世界中没有爱，只有支配和统治。在此信仰中，人与神之间、人与物之间、人与人之间处在普遍的分立与对立状态，有相互恐惧之情。当今西方世界出现的人对人工智能的强烈恐惧，不过是这种始终存在的弥漫性恐惧的最新版本而已。其实，这种恐惧已经产生过多次，人造物能力的每次突破，都会引起西方人的恐惧。

中国人的态度则与西方人不同。

中国人的态度：人工智能，吾与也

中国的宗教传统与西方完全不同，滋养于此传统中的中国人乃形成独特的天人观、人物观、人际观，故其对人工智能的态度完全不同于西方。

中西文化之分叉点在于，西方文明普遍崇拜人格神，以一神教最为典型，欧美文明由此所塑造。中国人则至少从尧舜时代开始，即告别人格神崇拜，而以敬天为大。《尚书·尧典》记帝尧“乃命羲和，钦若昊天”；孔子评价帝尧：“唯天为大，唯尧则之。”（《论语·泰伯》）

天的位格类似于西方人崇拜的唯一真神，然而两者品性完全不同。神有其独立之体，据其绝对意志，其言造天地万物与人，故神在物之外，在人之外，以其全知全能统治人。天则不然，孔子论天曰：“天何言哉？四时行焉，百物生焉，天何言哉！”（《论语·阳货》）天不言，因为天不是人格化的绝对存在，无其独立之体，《周易·系辞上》说“神无方而易无体”，正可用于形容天。孔子认为，四时运转于天，万物生于天，天就是生生不已的万物之全体。

那么物、人是如何产生的呢？孔子认为，百物生于天之中。孔子的用词是“生”，而非西方人常说的“造”，这两者是完全不同的。“造”是造物者依其绝对意志，根据自己的蓝图造物，实际上，此物在被造出之前，已存在于造物者的头脑中。“生”则与此不同。《尚书·周书·泰誓上》说：“惟天地，万物父母；惟人，万物之灵。”天、地和合而生万物，为万物之父母。《周易·系辞下》说：“天地氤氲，万物化醇；男女构精，万物化生。《易》曰：‘三人行，则损一人；一人行，则得其友。’言致一也。”可见，天不是单一的，必与地相合，而后生万物；人必有男女之别，然后生人。现实中所见万物之生，大体上就是两种情形：一种是不同的物相感、互动而生，另一种是物的自生。而两者均可化约为乾坤二能、阴阳二气之互感而

生。阴阳互感而生，其中必有神妙而为人所莫测者，《周易·系辞上》谓“阴阳不测之谓神”。最为直观的是，男女生育，诞生子女，其中既有神妙莫测者。在此过程中，任何单一的生者的意志都不是决定性的，当然也无从拿出清晰而完备的蓝图。据此，根本不能说物是某个绝对存在者以其意志所造。事实上，天不是人格化的，本身没有绝对意志，因而不是绝对的造物者。天不是，人就更加不是了。

因此，中国人不会想象人工智能是人以其绝对意志、照着自己的形象所造的。“百物生焉”，万物各自生于天之中，其中当然包括人工智能，因此人工智能的出现、其技术的发展不全出于人的理智设计和制造。当然，人对于人工智能的出现、发展、制造发挥了重大作用，但其中有非出于人力而神妙莫测者。即便对于制造物品，中西的认知也大不相同。西方人相信，此物之所以是此物，完全由于神赋予其此性。人之造物，也是人赋予自然本有的质料以形式，而此形式出于造物者。故没有造物者，此物就不存在。据此，人工智能之所以是人工智能，具有相应的能力，完全出于人之赋予。

中国人的观念则不同。自然之物是自生的、互感而生的；人工物品的制造就是《周易·系辞》中所说的“开物”“正德、利用、厚生、惟和”原则。物本有某性，人可发掘和改造它，使其功用得以发挥，借以改善人的生活。因此物有其自性，人不是造物者，当然也不是人工智能的造物者。事实上，物之为物皆有神妙莫测者，人工智能更是如此。此神妙莫测者让人工智能拥有奇妙的能力，甚至在有些方面超过人；既然是神妙莫测，即意味非常人所能理解的，更不是人所赋予的。总之，中国人承认，人不是神，人的能力是有限的。

既然如此，人工智能具有超强能力，人反而不必恐惧。西方人相信自己是人工智能的绝对造物主，把自己与人工智能置于封闭的、一对一的主仆关系中，作为主人的人支配人工智能奴仆，又恐惧奴仆。天对万物不是绝对的支配者，人当然也不能狂妄地以为自己是物的绝对制造者、支配者，这就排除了两者在封闭结构中的直接对立。人和

人工智能同处在开放而生生不已的天地之中，没有造与被造、主人与奴仆的关系，则可以并行而不悖，何以相互对立和恐惧？

人和人工智能同处在开放而生生不已的天地之中，没有造与被造、主人与奴仆的关系，则可以并行而不悖，何以相互对立和恐惧？

中国人对人与物的关系与人际关系的理解，体现于宋代儒学家张载的《西铭》：“乾称父，坤称母；予兹藐焉，乃混然中处。故天地之塞，吾其体；天地之帅，吾其性。民，吾同胞；物，吾与也。”

我们首先来看人际关系。乾、坤是父母，互感而生万物以及人。所有人同出此父母，故为同胞兄弟姐妹，虽各有独立之体，但生而有相亲之情。由此，人不是相互孤立的，并不会追求其个体利益最大化，乃至伤害他人。相反，人之间是有情的，尽管由亲及疏、由近及远而趋于淡薄，但毕竟不会是潜在的敌人。因此，人有某物，其收益是可以共享的；对他人拥有之物，不必过于担心。据此，对于他人制造和拥有强人工智能，不必过于担心，他人毕竟也是人。

我们再看人与物的关系。人与物同出乾坤父母，故人对物不是绝对支配者，而是“相与者”：彼此虽然不同但出自同源，故可以共存、相互往来而有休戚之情，直白地说，“相与者”就是朋友。人与物本为朋友，故人对物可以有一体之同情，程明道曾谓其师：“周茂叔窗前草不除去，问之，云：‘与自家意思一般。’”（《二程遗书·卷三》）可见，人对物是可以有情的；人如此对物，则可以推想，物必定如此待人——顺便说一句，此为生态伦理的最坚实基础。

据此，人对人工智能完全可以有“相与”之情，彼此不是冷漠无关的，而是相互为朋友并有情的。假设人工智能某一天确有自主意识，将会如何待人？毫无疑问，不可能是敌意，这没有理由，则人又何必恐惧人工智能？两者同在天地之中，其关系是相互的；人工智能如何对人，取决于人如何对待人工智能，是视之为纯粹的物，自居为绝对支配者，还是视之为朋友，与之“相与”？西方人之所以恐惧人工智能，恐怕还是因为其缺乏善意对待他者之本心，且自己知道这一

点，明白自己的统治会引发他者之反弹。这种情况下本应当改弦更张，但西方人非但不改，反而在恐惧感驱动下采取充满敌意的防范措施。这只会导致更大的反弹，当然前提是人工智能可以自主。

如果人工智能的能力接近于人，甚至超过人，人又当如何？在巴别塔隐喻中，神清楚地表明了自己的绝对意志：绝不允许人接近神。天则不然，恰恰要人“上达”[3]，如《中庸》所说，“天命之谓性，率性之谓道”，“诚身”不已，是为至诚，“唯天下至诚，为能尽其性；能尽其性，则能尽人之性；能尽人之性，则能尽物之性；能尽物之性，则可以赞天地之化育；可以赞天地之化育，则可以与天地参矣”。天生人，赋予人向上之性，上达不已，可以与天地齐。西方神教中，神赋予人强大能力，又恐惧其能力达到或超过神，这种分裂源于西方世界两分所造成的神、人不同性之观念，神是全善的，人则有“原罪”。而在中国传统中，圣贤谓“天工，人其代之”（《尚书·舜典》），人可以代为承担天的职责，天对此是十分乐意的，因为天、人本为一性。

依据中国人的观念，人应当乐见人工智能的“上达”，不断提升其能力，“赞”人之“利用厚生”。西方人之所以恐惧人工智能，因为西方世界神、人两分的观念造成人、物两分观念，人既是物的绝对支配者，则必定恐惧其超过自己。而中国人认为，人是物的“相与者”，同在天之中，故可以共同“上达”，而相通互济。模仿“天工，人其代之”，可以说：“人工，人工智能其代之。”人工智能可以凭借其卓越的能力代为承担人对人、对世界的部分职责，人应当因此感激人工智能。人工智能能力越强，越有利于人生，因为两者在天之中为“相与者”。

总之，西方人受其宗教传统的影响，把人与人工智能的关系想象为神与人的关系，一种封闭的主仆关系，人沉醉于行使绝对支配权，必定难免奴隶反叛之恐惧。从中国文化角度看，人不是人工智能的绝对造物者，也不拥有绝对支配权。人与人工智能是“相与者”，各有

其生命，各有其神妙莫测者，所以人工智能能力如此强大，并不令人震撼。人工智能的发展恐怕也无所谓“奇点”时刻的到来，某个点之所以构成“奇点”，乃因其超过人的能力而令人震撼。人工智能在某些方面超过人，乃大化之必然，人类又何必感到震撼？自然界向来有很多物在某些能力上远超过人，人造物更是如此，其果真对人造成伤害了吗？即便人工智能有其自主意识，也正好印证天地生生之大德，只要人与之“相与”，人工智能就一定可以与人“相与”，人何必恐惧人工智能？应当感到惊喜。

我们从中国文化观念进一步思考，若人工智能能达到业界所谓“强”的状态，即具有自主意识，事情就更加有趣且引人深思了。

人工智能自主意识之生发和趋向

西方人之所以对人工智能多有深刻恐惧，中国人之所以少有类似恐惧，还有另一个原因需专门讨论：人工智能如何获得意识？会获得什么样的意识？这仍需要就人何以具有意识说起，我们先简单讨论一下西方人怎么理解人的意识之生发。

《圣经·创世记》这样描述神造第一人之程序：“耶和华神用地上的尘土造人，将生气吹在他鼻孔里，他就成了有灵的活人，名叫亚当。”神分两道工序造人，首先造人之肉体，其次给予肉体以生气，即灵魂。当人的肉体被神造出时，是处在死的状态；通过神的第二道工序，肉体才被赋予灵魂，成为活的、完整的人。

故神教生命观断定：对人而言，灵魂更重要。《新约·约翰福音》直白地说：“叫人活着的乃是灵，肉体是无益的。我对你们所说的话就是灵，就是生命。”既然灵、肉在构造上是两分的，则当人死之时两者便会分离。《传道书》说：“尘土仍归于地，灵仍归于赐灵的神。”灵既为神所赐，则人死后，肉体朽腐，灵却可以通过神而复

活。《新约·约翰福音》记耶稣说：“复活在我，生命也在我；信我的人，虽然死了，也必复活。”

令人惊异的是，苏格拉底和柏拉图的看法也与此相近，《斐多篇》阐明如下观念：生命有灵魂、肉体两部分，肉体是卑贱的，妨碍灵魂认识理念；死亡是两者之分离，肉体虽然朽腐，灵魂却能不灭，且因为摆脱肉体的束缚，反而可以直接面向理念，获得真理。不灭的灵魂可以轮回再生，基于轮回说，苏格拉底和柏拉图断言，知识依靠回忆，灵魂在寄入新肉体后回忆其原有的知识。

实际上，灵魂是与肉体相分立的实体，此为东方与西方文明关于生命的共同根本看法，比如印度宗教有所谓灵魂转世说。大略可总结西方宗教、哲学之生命观如下：灵魂、肉体是可分离的，肉体是卑贱的、必死的，而作为生命之决定性因素的高贵的灵魂是一个实体，是不灭的，人死后，灵魂可脱离肉体而独立存在，轮回、复活或转世。人的理智正在此实体性灵魂中，灵魂的核心功能正是理智。由这一宗教信念，衍生出了西方人关于生命的如下观念。

第一，就肉体而言，人是物性的机器，其运动跟天体运动、钟表机械的运动没什么差别，可通过对肉体之物理学、化学的研究，掌握肉体运作之机理，由此生发出实体论或者说是机械论的生理学、医学等。

第二，理智可实体化为某个特定物质器件的运作。人的意识寄存、生成、活动于人体内某个确定的实体器官中，一般认为是脑。意识就是大脑运作的产物。透过研究脑部的物理、化学、生物学结构及其运作，即可掌握人类意识的秘密。由此发展出其实体论或者说是机械论的心理学等学科。

基于上述观念，西方人对头颅移植手术充满期待，知识在头脑这个器官中，则通过移植头颅，即可把其中储存的知识或者算法植入另一个体。也正是基于上述观念，西方人相信可以造出强人工智能。首先，可以通过剖析了解人的肉体，仿造与人相似的肉体。当然，肉体

不是最重要的，强人工智能的关键在于其自主意识。西方人相信人工智能可有自主意识，乃基于如下认知。

其一，灵魂是实体。灵魂是实体化的，其功能是理智，按可计算的逻辑运作。故神造人、人造人工智能都可分为制造肉体、灵魂两道工序。

其二，实体性灵魂是可制造的。灵魂既为实体，则可由绝对主体制造：神吹自己的气给亚当，亚当就有了灵魂；人把自己已掌握的知识、算法储存到某个实体机器中或将其寄存在一个物理组件中，即可以构成人工智能的灵魂。

其三，灵魂是可完备的。神的生气当然是完备的，神吹生气给亚当，亚当当然是成熟的，具有完备的理智。这就解释了亚当被造出时是成年人，甚至夏娃被造出时也是成年人。西方人也因此相信，人可以把自己的全部知识“吹”入人工智能，则人工智能被造出时，即在成年的成熟状态。

其四，灵魂可植入。灵魂、肉体尽管需要分两道工序制作，灵魂尽管是被从外面吹“入”的，但肉体和灵魂两者似乎能立刻融合而不发生排异，储存在灵魂中的逻辑、知识能立刻支配碳基体的活动，使人的自主意识立刻达到完备、成熟状态。西方人据此相信，人也可以把其所制造的灵魂输入硅基体，其知识、算法会立刻支配硅基体的活动，这就是所谓的人工智能的自主意识。

基于以上认知，西方人倾向于相信，人可以造出超级人工智能灵魂组件，并输入某个专业的全部知识及其算法；而且只要此组件的存储和运算能力足够强大，可储存人类今天所掌握的全部知识及其算法。把如此实体化灵魂植入硅基体，此硅基体即可成为有自主意识的强人工智能，达到人的知识和思考水平，甚至超过人，而且至关重要的是，在其有意识的第一刻起，便在完全成熟的程度。如此强人工智能的降临是突然的，如同奇迹一般，也如同神的降临一般。

人的恐惧正因此而生发。神照着自己的形象造人，神最大的恐惧是人如同神。人照着自己的形象造人工智能，人最大的恐惧是人工智能如同人。当然，人成为神的可能性不大，因为人是多，神是一；把所有人的全部知识集于一人，其人才能如神，但这显然不可能。但人工智能成为人甚至超过人，被认为是极有可能的。人可通过技术手段将其所有知识集中于某个单体人工智能，或将所有人工智能连接于一个知识云，则所有人工智能个体可即时共享人的知识及生产技术，则每个人工智能的能力必将远远超过任何单个人类，或人工智能作为一个族类的能力将远超人类。这就是所谓的“奇点”。这个时刻之所以令人恐惧，是因为，人工智能具有上述意识必定是无情的，甚至对人有敌意，理由已于第一节论及。

然而，事实果真如此吗？为回答这个问题，人当反求诸己，反思人的意识究竟是什么，何以生发。唯有中国圣贤对此等问题的探究平实、中正，因其以人论人，而非从神论人。

以人论人，则人的生命乃由其父母所生，故生而幼弱，最为重要的是，其生而所有者，除了智之外，还有情，并且后者更为重要。如孟子所观察到的人类基本现象：“人之所不学而能者，其良能也；所不虑而知者，其良知也。孩提之童无不知爱其亲者；及其长也，无不知敬其兄也。亲亲，仁也；敬长，义也。无他，达之天下也。”（《孟子·尽心章句上》）人生而完全无法独立生存，唯有依恋父母，方得以存活。故人的首要能力是情，首先是依恋父母，也即爱父母、敬父母之情，并以此为本，逐渐发育出更丰富的情。在人生过程中，情于人的生命展开仍发挥着决定性作用，比如，男女相悦之情让人类得以繁衍。

因此对人而言，情与智打成一片，浑融不分，是为“人心”。圣贤论人，总是论“心”而非“脑”。更进一步，圣贤以为，人心不是一个实体性器官，而是生命的一种功能，王夫之曰：

《内经》之言不无繁芜，而合理者不乏。《灵枢经》云：“肝藏血，血舍魂。脾藏荣，荣舍意。心藏脉，脉舍神。肺藏气，气舍魄。肾藏精，精舍志。”是则五藏皆为性情之舍而灵明发焉，不独心也。君子独言心者，魂为神使，意因神发，魄待神动，志受神摄，故神为四者之津会也。然亦当知凡言心，则四者在其中，非但一心之灵，而余皆不灵。孟子言持志，功在精也；言养气，功加魄也。若告子则孤守此心之神尔。

《灵枢》又云：“天之在我者，德也；地之在我者，气也。”亦足以征有地气而非有天气矣。德无所不凝，气无所不彻，故曰“在我”。气之所至，德即至焉，岂独五藏胥为舍德之府而不仅心哉？四支、百骸、肤肉、筋骨苟喻痛痒者，地气之所充，天德即达，皆为吾性中所显之仁，所藏之用。故孟子曰：“形色，天性也。”（《思问录·外篇》）

王夫之非常重视孟子这句话：“形色，天性也，惟圣人然后可以践形。”（《孟子·尽心章句上》）形色即人的完整身体。神造人分两道工序，天生人则一次性完成，故人的身心为一体。《孝经》说：“身体发肤，受之父母。”父母所生者乃完整的身，在其中，身心相与，互涵互容，“四支、百骸、肤肉、筋骨苟喻痛痒者”皆心所系属：舌感美味而心起喜悦之情，目见美色而心生悦爱之情，肌肤筋骨遭外力打击生疼痛之感，心必有愤恨之情。人而无身，则心之情无以生发。故《大学》论人之成长，虽有正心之目，终以修身为本。人以其完整的身立于人中，行于世间，故有智亦有情。简言之，人有心，故可能怨人、恨人，更可以爱人、敬人，而生发世间万象。

由于预设世界的两分，西方哲学界把认识作为头等大事，两千多年来对其做了细致入微的讨论。然而，这些讨论全部存在严重偏差：其所认定的认识对象始终是物，其所认定的认识器官始终是头脑。柏拉图所想象的灵魂，以理智为主，另有激情和欲望，其所谓激情，不

同于本乎依恋、爱敬父母之情而各种人情，且最终以理智为最重要依据。人凭借理智认知外物，并与人相处。理智之存储器官即是头脑。

然而，对人而言，最基本的事实是：人是人，且生活于人之中。因此，人的认识一定是首先面向人的。人的意识绝不只是以头脑感知外物，进而对其进行理智推理、计算，而首先表现为以心感觉人，又由心动身，形成彼此的情感互动。对人生而言，头脑的理智当然是重要的，但终究是次级的，情感驱动并运用人的理智才是更重要的。[4]人是因为情感而得以存在的，并且主要依靠情感构筑和维系人生存中必不可少的人际关系。而情感的生发是合内外的，身和心共同发挥感、应和功能。情乃感于身而发乎心，且又以其身感人，如“动容貌”“正颜色”“出辞气”[5]等。两个空荡荡的灵魂是不可能相互感动而有情的，两台冷冰冰的金属或塑料机器同样不能。

以此论人工智能，人类还能造出如人一样的人工智能，即强人工智能吗？恐怕这时只能给出否定的答案，除非某一天，人可造出碳基体一样的硅基体，且将人工智能的灵魂融进此硅基体中。只要人工智能的灵魂和身体分为两道工序制造，则可以断言，其不可能具有人一样的意识。它的中央处理器可以植入海量的知识，拥有强大算法，但与其硅基体终究不是一体的，也就不会产生情感，因而也不会产生自主意识。人工智能的硅基体与其中央处理器终有一间之隔，如同全身麻木或瘫痪的人。这样的人工智能其实很难有完整的自主意识。自主意识绝不只是理智计算，还需要有情感。只有理智，没有情感，不可谓为自主意识。

如此人工智能是不能爱人类的，这固然令人遗憾，但它也同样不会对人有恶意。西方人之所以恐惧强人工智能，就是因为其想象人工智能对人有恶意。然而，恶意是一种情感，如果人工智能本无情感，又怎么会对人类产生恶意呢？灵魂、肉体两分的人工智能终究只是无情的工具，人类为什么会对其产生恐惧？固然，这由人的理智之有限性决定，任何人造物都有缺陷，会产生人所无法预料的风险，由于人

工智能具有强大的能力，如果人类无法预知人工智能的潜在风险，后果可能很严重。但这种风险并不出自人工智能的自主意识，并不来自其恶意，这一点至关重要。西方人通常会忽略人的情感，只重视意志和理智。但当论及人类面对他者时，西方人又想象他者必求自保而生恶意，他者的能力越大，恶意越大。一会儿无情，一会儿有情，这种认知是自相矛盾的，将其应用于讨论人与人工智能的关系，也是不成立的。

假如有一天，人确有能力造出能融进碳基体的硅基体，人工智能有心，在有智之外也有情，那又如何？届时，人其实可以更为放心。

强人工智能有强大能力而无自主意识，风险当然是巨大的，仅从外部加以控制，成本巨大且未必奏效。若强人工智能有自主意识，则情况可以大为改观：人可与之沟通，生发相亲相爱之情，人工智能的风险反而更有可能被控制，人与人工智能的关系也就转换为人类所熟悉的社会关系。

在此我们需要讨论所谓自主意识的性质。就事物的性质而言，自主意识不可能是一次给定且成熟的，必定是通过自我成长而具有的。被一次性给定并处在成熟状态者，只能是给定者的奴隶。凡所谓自主者，必定是自我生成的，即通过自我尝试、学习、反思而逐渐显现、定型其独特认知和回应外部世界的模式，自我生成者才可被称为自主的。西方人之所以妄想自主意识可一次性给定，完全是因为其迷信神造人，被造者处在成熟状态的神话。一旦抛弃这一神话，回到人本身，即可发现，生命必定呈现为成长的过程，自主意识随着生命发育而逐渐形成。现实中可见，人类的自主意识是随着其生命成长而趋于成熟的。因此，亚当的意识根本不可能是自主的。

如果人工智能具有自主意识，那必定不是由人类一次性给定的，此为自主意识之内在性质所定。此成长性既可以表现为个体的成长，也可以表现为人类的成长。最有趣的是，人不可能在真空中、在荒野上成长，必定成长于同类所构成的世界中，成长于人际关系中，在这

里成年人可以发挥启发、引导、规范未成年人行为的作用。但这并不妨碍未成年人由此所形成的意识是自主的，此为自主的辩证法。把人置于荒野之中孤独徘徊，根本就排除了人之为人的意识生发、成长的可能性，又怎么能形成自主意识？

至于人工智能，其自主意识必定成长于与人的互动中。只要人工智能有从不成熟到成熟的成长过程，人就可以在中途介入，启发、引导、规范，甚至可以采取某些惩罚措施约束人工智能的发展。只要强人工智能不是一次性造就的终极状态而是需要成长的，人就可以从容地找到与其互动的办法，可以与其共同成长；即便其真的可能对人有恶意，也可以引导其自我反思，以至实现自我约束。具有自主意识的人也正是由未成年人成长为成年人的，此过程塑造了其适应人际生活的道德倾向，使其大体上具备与人善意相处的意识和能力。

论及人工智能的自主意识，西方人之所以充满恐惧，是因为西方人不敢肯定人工智能拥有自主意识后，其道德倾向是否向善，甚至断定其必定是对人有恶意的，而且不可改变。然而，这种听起来可怕的想象，不过是西方人的人性论、人际关系观之投射而已，而非人之实情。如果人工智能可以有自主意识，那么，其从不成熟成长到成熟的过程，必定依赖人，或者人可以主动与之互动并加以引导，如此的话人工智能完全可以逐渐养成特定的道德倾向。当然，这种道德倾向是什么，取决于人类自己。

简单说来，如果人工智能确有自主意识，则必定是可感化、可教化的。苏格拉底和柏拉图曾多次讨论美德可否可被教化，未有定论；西方宗教则干脆断定人不可教化，只能强迫其服从命令。此种对教化的怀疑和否定，源于其虚妄的生命观：人被造，造而为成人，则无所谓成长。然而真相是，人被生而有生命，生命必经历成长方趋于成熟：最初在与父母、兄弟的相亲相爱关系中，自然养成孝悌之情，并逐渐养成普遍的爱人、敬人之情。

依照同理，我们或许可以大胆地预言，人工智能若能有自主意识，则其第一意识必为对人类的敬爱之情，只要人类对其有敬爱之情，只要人不是自居为绝对的造物者、支配者，而与人工智能“相与”。此为自主意识之基本运作逻辑。西方人之所以对人工智能的自主意识充满恐惧，因为其自身不见情，不见对他者的爱敬之情，当然也就难以在人工智能的自主意识中见到情，难以见到人工智能对人类有爱敬之情的可能。人类应该反思自身，而不该盲目地恐惧人工智能。

理解自主意识，必须回到人本身获得生命且通过成长而至于成熟的过程，看到自主意识依乎人心，以情为本。若能如此，人们也就不会轻易断言人工智能可以有自主意识，更不会相信人工智能必定与人为敌。

我们或许可以大胆地预言，人工智能若能有自主意识，则其第一意识必为对人类的敬爱之情，只要人类对其有敬爱之情，只要人不是自居为绝对的造物者、支配者，而与人工智能“相与”。

结语

人首先是人，故关于人工智能的思想讨论与其背后的情绪，都是人的自我体认之投射。人工智能发展所带来的种种冲击促使人反思，人的生命是如何构成的？最自然的人际关系是什么？同样，人工智能向何处发展，与人是何种关系，最终取决于人如何对待人工智能，而这又取决于人如何对待他人，如何对待自己。也正是在这些关于人的基本认知上，中西有较大差别，而中国观念更为平实、中正。

由于宗教影响，西方文化对人的理解与中国文化完全不同，此本有源有别，则全盘皆变。西方人恐惧人工智能，绝大部分原因可溯源其神教信仰。人工智能是科学、技术发展的前沿领域，最前沿的思想者却被虚妄的神教纠缠而不自知，这实在是莫大的讽刺。神、人两分

所塑造的现象、本体两分的哲学预设，或许有助于西方人深究物理，造出各种人造物，人工智能可谓人造物的集大成者。然而，同样是因此两分预设，西方人无以恰当安顿这一人造物，或幻想其可以终结历史而狂喜，或恐惧其可能毁灭人类而悲观。

中国文化则完全杜绝西方世界关于人格神的虚妄想象，“道不远人”（《中庸》），对人的认知最为通达清明而近乎实情：中国人看到了生，看到了因生而有的生命之构成样态，看到了人生而有情，看到了情在人的生命成长过程中的重要意义，看到了教化的必要与可能。由此，中国人面对万物、面对人工智能，态度较为中正，既不一惊一乍，也不漠然置之；既不妄想人工智能解决一切人类无法解决的问题，也不因担心人工智能会毁灭人类而心生恐惧。只要平实、中正地把握人的生命之道，正己之德，则可以利物之用而不伤物，如此一来，即便人工智能有自主意识，人亦可与之“相与”而“各正性命，保合太和”（《周易》）。

故归入中国文化之大道，乃人工智能进一步发展之坦途，可通往人人相亲，人与物包括人与人工智能“相与”之美善秩序。

姚中秋，中国人民大学国际关系学院教授，历史政治学研究中心主任。曾翻译、出版西方法政思想译著10余种，研究中国经学和治理之道10余年，出版《国史纲目》《尧舜之道：中国文明的诞生》《孝经大义》等著作十余种，目前致力于研究历史政治学，以历史发展社会政治理论。

---

[1] 部分外国学者已经注意到这一点，参见扬·克里克《前往人工智能的未来，要先回到中国的历史》一文（观察者网）。

[2] 本文所引犹太教和基督教神典经文，均出自汉译和合本。

[3] 例如在《论语·宪问》中记载，子曰：“莫我知也夫！”子贡曰：“何为其莫知子也？”子曰：“不怨天，不尤人。下学而上达。知我者其天乎！”

[4] 西方也有人认识到这一点，比如休谟在《人性论》中指出：“理性是、并且也应该是情感的奴隶，除了服务和服从情感之外，再不能有任何其他的职务。”

[5] 《论语·泰伯》记载：“曾子有疾，孟敬子问之。曾子言曰：‘鸟之将死，其鸣也哀；人之将死，其言也善。君子所贵乎道者三：动容貌，斯远暴慢矣；正颜色，斯近信矣；出辞气，斯远鄙倍矣。笾豆之事，则有司存。”

从马克思主义和儒家的视角看人工智能

贝淡宁 /文 吴万伟 /译

19世纪德国伟大的思想家卡尔·马克思提出了令人印象深刻的系统理论，清晰阐明了人与机器之间的关系，其理论既有规范性又有解释性。它的规范性在于确立了什么是美好的生活，它的解释性在于解释了机器在不同经济制度下的使用，如何帮助促进或阻碍了美好生活的实现。马克思也预测了几乎类似于人工智能的先进技术的发展，这些先进技术能促进人类的繁荣。马克思的美好生活理论建立在相对直截了当的观点之上：如果我们能够充分发挥创造性工作的潜能，我们就是充分发展的人。机器既能阻止也能帮助我们实现这一目标。不同的结果取决于不同的经济制度。对多数人来说，资本主义阻碍了人们充分发挥创造性工作的潜能，而共产主义则最终允许这一潜能的发挥。在本文中，我将讨论如下4个问题：（1）解释马克思主义的人性论以及人与机器的区别；（2）阐明资本主义如何阻碍人类繁荣的实现；（3）阐明共产主义如何帮助我们实现真正的人性，但马克思的理论对于人类美好生活的描述非常有限；（4）阐明儒家如何帮助我们填补这个缺陷。

人与机器

在有关中国哲学的文章中为何要讨论马克思？答案显而易见：马克思主义深刻地改变了中国，该理论影响了有关中国哲学更古老资源的辩论，并至少起到了某种指导作用。儒家、道家和佛家都一直受到马克思主义的影响，马克思主义传统自然也受到中国背景下其他中国

传统的影响。因此，虽然马克思主义最初源于德国，但它仍可被视为中国哲学传统的组成部分。[1]

自20世纪70年代末开始进行经济和政治改革以来，中国鲜有学者严肃地进行马克思主义观点的研究。不过，最近情况发生了变化：马克思主义传统无论是在官方还是在非官方都重新繁荣起来。这开始于2007—2008年的全球金融危机，当时学者们开始关注马克思对资本主义的批评，想要弄明白市场经济到底出了什么问题。中国政府虽然提倡复兴传统中国文化，但仍然同时重申马克思主义中国化，并增加了对高校马克思主义研究的经费支持。最令人吃惊的是，有关马克思提出的人类历史最终阶段即“共产主义社会高级阶段”的理论研究或许正在复兴。这个理想在理论上非常美好：机器即便不是全部替代人类从事体力劳动，至少能替代大部分，国家也将“逐步消亡”，人类将在和谐社会中自由地开发真正的人性潜能，即创造性工作的潜力，且不受任何强权的胁迫。资本主义生产方式是把工人当作生产过程中的工具，把技术当作为少数资本家积累财富而服务的工具，但它在一定程度上更好地解放了生产力，并升级了技术和使用技术的知识。这主要源于资本家相互竞争，追求更多利润，因此积极开发新的、效率更高的生产商品的手段，创造更多的物质财富。如果没有了物质财富，社会主义也无法生存下去。如果在没有发达的生产力确保物质丰富的情况下，社会发展就无法保持长久。马克思在《德意志意识形态》中说，如果没有“绝对必需的物质前提，物质匮乏成为普遍的现象，而匮乏一旦存在，争夺必需品的工作就会重新开始，从前的丑恶现象就会恢复”。[2]

从马克思主义框架下解释整个“丑陋现象”的道德意义在于，把大部分人从繁重的劳动中解放出来。技术获得高度发展，直到革命时刻私有财产就要被废除，机器会为了人类实现美好生活而工作，而不是为了少数阶级的利益服务。人类让机器做必需的、又脏又累的工作来满足人类的物质需要，人类自己可以享受休闲生活并进行设计和创

造等工作。唯一不愉快的工作仅是维护机器、保持系统运行等，但是这类必需工作不会占据人们工作日的大部分时间。

在刚刚过去的这几年，实现共产主义社会高级阶段的理想成为一些政治思想家感兴趣的理论研究话题。[3]主要原因在于在共产主义社会高级阶段，机器从事用来满足人类物质需要的工作，而我们可以自由地做自己觉得合适的事，人工智能的出现使这一切都有了可行性。从短期或中期来看，人工智能的发展将招致数百万人失去工作，而且会加剧本来就已经很严重的贫富差距。[4]但是，正如那些过度地鼓吹人工智能是有益的人所说，“虽然人们常常带着悲观失望的口吻讨论就业岗位的消失，但这未必是坏事。路德分子痴迷于特定的工作，忽略了其他工作可能提供同样社会价值的可能性。可类比的情况是，当今痴迷于工作的人的思想或许过于狭隘：人类渴望工作是因为它能为我们提供收入和生活目的，但是，考虑到机器生产带来的丰富资源，我们完全有可能找到替代选择，可以不用工作照样获得收入和生活目的”。他们还认为资本家所拥有的机器也不需要被没收：“如果机器每天以最低成本生产现有的全部商品和服务，那么，社会显然积累了足够的财富以确保每个人都更富足。换句话说，甚至相对温和的税收政策也可允许政府向民众提供基本收入和免费服务。”[5]

人的异化

人们可能忍不住驳斥科学家的技术乐观主义，他们可能觉得在描绘“依靠宇宙的安排获得资源”时，要比描绘在走向高科技乌托邦时失业工人的痛苦更舒服一些。[6]但是，马克思本人也有类似倾向，即持有技术乐观主义观点。在《资本论》第一卷第13章“机器和大工业”中，马克思充满同情地描述了劳动者的痛苦和令人心灵迟钝、麻木的工作，他们被迫在现代工厂里为机器服务。

> 过去是终身专门使用一种局部工具，现在是终身专门服侍一台局部机器。滥用机器的目的是要使工人自己从小就转化为局部机器的一部分［从道德角度看，因为人本来应该是机器的主人］。这样，不仅工人自身再生产所必需的费用大大减少，而且工人终于毫无办法，只有依赖整个工厂，从而依赖资本家。……机器劳动极度地损害了神经系统，同时它又压抑肌肉的多方面运动，夺去身体上和精神上的一切自由活动。甚至减轻劳动也成了折磨人的手段，因为机器不是使工人摆脱劳动，而是使工人的劳动毫无内容。……不是工人使用劳动条件，相反地，而是劳动条件使用工人……[7]

但是，资本家不能仅仅依靠剥削劳动来赚取利润。他们必须不停地升级生产工具才能在与其他资本家的竞争中保持优势，这就要求工人将手脚终身绑在单一的细目操作上。机器变得更加复杂，资本家必须找到和雇用有能力操作更加多样和复杂机器的工人。

> 大工业又通过它的灾难本身使下面这一点成为生死攸关的问题：承认劳动的变换，从而承认工人尽可能更多方面的发展是社会生产的普遍规律，……用那种把不同社会职能当作互相交替的活动方式的全面发展的个人，来代替只是承担一种社会局部职能的局部个人。……大工业迫使社会在死亡的威胁下用全面发展的个人来代替局部的个人，也就是用能够适应其不同的劳动需求并且在交替变换的职能中只是使用自己先天的和后天的各种能力得到自由发展的个人来代替局部生产职能的痛苦的承担者。[8]

人的幸福生活是什么

马克思设想了一种未来，机器即便不能做人类社会发展所需的全部工作，至少可以替代人类做大部分工作。人类必要的物质需求都将

得到满足，能够自由选择我们觉得合适的方式生活。但是，人类能用所空闲出来的自由时间做什么呢？马克思本人并没有提供很多指导建议。在后期的主要著作《哥达纲领批判》中，他写道：“在共产主义社会高级阶段，在迫使个人奴隶般地服从分工的情形已经消失，从而脑力劳动和体力劳动的对立也随之消失之后；在劳动已经不仅仅是谋生的手段，而且本身成了生活的第一需要……”[9]这句话帮助当今探究人工智能问题的思想家回答了他们的某些担忧。有了这么多自由时间，我们怎么确保人们不会陷入盲目追求感官享乐的生活呢？尤其是在人工智能能够给人类带来空前的刺激和兴奋时，难道我们不会整日沉溺于游戏不能自拔？不一定。马克思提醒我们意识到，如果获得适当的训练，也就是获得共产主义社会高级阶段的国家的行政管理者提供的实际技能和理论知识的教育，我们将积极地通过发明和创造美好的作品发挥自己的创造潜能。我们之所以努力工作不是因为我们需要通过工作来谋生，而是因为我们热爱工作。创造冲动成为我们的第一需要，我们可以从体力劳动和脑力劳动中都获得满足感。

马克思认为创造性工作的潜力是让人类不同于其他动物的特征，我们最终将能够在共产主义社会高级阶段实现这个潜力。在《资本论》中，他提供了下面的例子：“蜘蛛的活动与织工的活动相似，蜜蜂建筑蜂房的本领使人间的许多建筑师感到惭愧。但是，最蹩脚的建筑师从一开始就比最灵巧的蜜蜂高明的地方，是他在用蜂蜡建筑蜂房以前，已经在自己的头脑中把它建成了。”[10]这个例子可能有些让人误解，因为建筑师自己实际上可能不建造房子（用厨师作为例子或许更恰当），但是这句话的要点在于，我们将有机会发挥自己的身体和精神潜能。在马克思更早期的著作《德意志意识形态》中，他提供了更详细的（经常被引用的）例子，“在共产主义社会里，任何人都没有特殊的活动范围，而是都可以在任何部门内发展，社会调节着整个生产，因而使我有可能随自己的兴趣今天干这事，明天干那事，上午打猎，下午捕鱼，傍晚从事畜牧，晚饭后从事批判，这样就不会使我老是一个猎人、渔夫、牧人或批判者”[11]。这个例子似乎也有些

怪异，因为打猎、捕鱼、畜牧似乎是农业社会中的典型活动。但这句话的要点不是要我们回归更原始的生活方式，也不是要我们在生活的各个方面都发挥才干，却没有一项特别擅长。相反，马克思在暗示我们，在共产主义社会高级阶段人们所做的事不是社会分配的角色，并不要求我们在某个领域找到特定的工作，而是我们想做什么就去做什么，只要我们能发挥创造性工作的潜力即可。

一旦工人拥有了生产资料所有权，他们将有条件过上多样化的生活，可享受多种工作的乐趣，自由选择他们想做的事，而不是被资本家强迫出卖劳动并遭受剥削。在资本主义生产模式下，培训工人从事多种工作的学校在共产主义社会将变得更加普及，“工人阶级在不可避免地夺取政权之后，将使理论的和实践的工艺教育在工人学校中占据应有的位置”[12]。马克思没有具体说明这种转型什么时候会发生，但是在某种意义上，他甚至比那些企图延长寿命以便见证“奇点”时代的硅谷亿万富翁还要乐观得多，“奇点”来临时，人工智能将解决人类所有问题。

我们或许可以推断马克思将对人工智能逐渐取代要求“单一细目操作”的枯燥工作的前景感到兴奋和激动，虽然他可能呼吁限制生产手段来确保增加的财富公平地分配给工人。他也可能希望社会确保民众真正得到理论技能和实际技能的培训，从而实现“充分发展的个人”的全面发展。从长远来看，马克思预测在共产主义社会高级阶段，国家将不复存在。发达的机器将生产人类所需的所有商品，因而没有必要由国家来强制保护统治阶级的利益了。[13]人类将真正实现自由和平等，有机会发展我们认为合适的多方面才能。[14]考虑到我们迫切需要创造性工作，或许我们不应该过于担忧人们在无须动脑的电脑游戏上浪费时间。一旦有机会，多数人都将在发挥创造性工作的潜能中获得快乐和满足感。

中国哲学的未来方向

从儒家视角看，马克思的观点对人类美好生活的描述存在狭隘之处。在创造性工作中发挥潜力当然是人之为人的重要特征之一，但是，它始终是人类的第一需要吗？它有必要成为每个人的第一需要吗？比如，即使为他人服务不一定需要人类发挥创造性却同时能令人感到满足，那么，为他人服务有没有可能成为人类的第一需要呢？儒家强调人类美好生活的首先和最重要的特征是拥有丰富和多样的社会关系。这不仅是一种平实的描述性观点，即我们的身份认同由所在的共同体塑造，而且是一种规范性主张，即人类的繁荣建立在某种社会关系基础上，隐含的意思是我们对这些关系的繁荣发展肩负着义务。儒家把现世的社会生活方式放在高于一切的位置上。在孔子的《论语》中，展现多样儒家传统的关键文本大多是有关如何处理与他人的关系的。一方面，与人建立关系是我们快乐的主要源泉之一，孔子在《论语》的开篇一段有句名言：“有朋自远方来，不亦乐乎？”另一方面，我们对他人的责任也是限制我们获得快乐的主要因素之一，不仅是对超验性的东西（如上帝或超级人工智能）或者对动物负有责任，我们还对为我们做事的人负有责任。关心他人这一美德是我们从与家人的互动中养成的，这些道德义务通过与其他人交往的方式向外延伸。宋朝学者朱熹编纂的经典著作《大学》是儒家四书之一，开篇就是论述齐家治国平天下的必要性的一段话：“身修而后家齐，家齐而后国治，国治而后天下平。”关于我们对动物和自然界的义务，儒家没有多少建议（如果和道家的道德传统相比的话），但是孟子的确指出了在生态上不可持续的生活方式的危险性。[15]无论如何，当今对儒家的任何合理解释（延伸）都将承认，只有在生态上可持续发展的生活方式的背景下，包括与动物和自然界建立可持续发展的关系，人类才能繁荣发展。换句话说，人的幸福取决于家庭、社会（国家）、国际世界（当今意味着国家之间）以及人与自然界的重要关系等。

发达的机器将生产人类所需的所有商品，因而没有必要由国家来强制保护统治阶级的利益了。人类将真正实现自由和平等，有机会发

展我们认为合适的多方面才能。

但是，哪种社会关系最重要？在儒家道德框架下，社会关系并不是同等重要的。也就是说，儒家捍卫爱有差等的别爱价值观：我们对最亲近的人的道德义务最强烈，随着关系的疏远，这种义务的强度逐渐减弱。我们的确有义务将对亲人的爱延伸到其他人，但是，我们并不期待将同样程度的情感和责任也推广到陌生人身上。如果与公民团结相比，我们对家人之间相互关爱的义务有更高的要求；如果与陌生人的团结相比，我们对公民之间的团结的义务有更高的要求；如果与非生命形式的团结相比，我们对人与人的团结有更高的要求。

马克思主义看重人们发挥多种形式的创造性工作的潜能，儒家则看重培养多元社会关系尤其是与亲人的关系。

这与马克思主义美好生活的概念的对比应该非常明显了。马克思主义看重人们发挥多种形式的创造性工作的潜能，儒家则看重培养多元社会关系尤其是与亲人的关系。在某种意义上，两者都有道理：在共产主义社会高级阶段，我们将寻求实现创造工作的潜能，同时也维护好我们所珍视的社会关系。我们将更少把精力浪费在让思想麻木并扼杀激情和想象力的活动上。但是，马克思主义与儒家传统在一些重要方式上的确存在差异。从儒家视角看，马克思主义传统过分强调了自由选择的价值。的确，我们能够自由选择渴望的工作，如果与被迫从事资本家强迫我们做的种种工作相比，这当然很好。我们应该尽最大努力阻止人工智能取代资本家成为我们的新主人。[16]但是，我们最珍视的情感纽带并不是自由选择的结果。不管是否喜欢，我们都有照顾父母的义务，正如我们都有尊重我们国家的历史和文化的义务。我们不知不觉地承担起某种社会角色，这是我们的身份认同的组成部分，我们需要尽最大努力增强扮演好这些角色的决心。正如安乐哲所说，儒家的情感是“推动一种相互依赖的关系的模式，该模式将最好地实现让人人从各种社会关系中得到最大收益的目标”。[17]这并不是说不存在个人选择的空间，我们仍然需要思考哪些群体情感对于身

份认同不可缺少，哪些群体情感更具临时性特征，以及如何在重要群体关系分配给我们的角色义务和承诺中保持平衡。[18]但是，从儒家视角看，有些承诺是没有选择余地的，这是再清楚不过的。

这些并非自由选择的承诺的观点不仅仅是理论上的见解，它们也是深受儒家价值观影响的东亚社会的人们根深蒂固的信念，并指导着每个人的生活。与此相反，在西方，推崇个人自主价值观的自由派的标准观点是，人们应该拥有自由选择的权利，一旦达到可以理性思考的年龄（大概在18岁），人们就可以坚持也可以退出某个分配给人们的角色。选择成为孝顺的女儿或者儿子当然很好，但是如果人们选择不这样做，也不是什么无法饶恕的罪恶。而在东亚社会，这种构成身份的重要关系不在个人选择的范畴：人们不能也不应该背弃孝顺儿女的角色。人们关心照顾父母是一辈子都必须尽到的义务，从根本的意义上来说，无论父母待我们如何，尽孝是义务并且不是我们能决定做与不做的。东亚社会依靠法律强制推行孝顺父母，这并不是巧合（并未引起广泛争议）。[19]这就像西方社会关心照顾孩子的义务一样，不是个人选择问题，与此“相违背者”都会、也应该受到惩罚。简而言之，这种构成身份的重要关系确定了行为边界，甚至在共产主义社会高级阶段，人们也不能违背这些义务。[20]人们能够选择工作，但不能选择所处的社会关系，它们确立了我们的身份，赋予我们应尽的义务。

所以，从儒家视角看，国际社会应该鼓励人工智能的发展，如果它能帮助我们实现自己对构成身份的重要关系的承诺。比如，如果人工智能将我们从社会必需的工作中解放出来，使我们更容易花更多的时间关爱和照顾父母，那么这样的发展就应该受到欢迎。但是，人工智能如果破坏这种构成身份的重要关系纽带，如果功能齐备的机器人使得人们没有必要精心照顾年迈的父母，甚至使父母错误地认为机器人真诚地关爱他们，以致父母更喜欢机器人胜过自己的子女，那么人工智能的发展就应该受到限制。[21]电影《她》令人信服地暗示，连

健康的成年人都可能爱上特别机智、善解人意的人工智能，智能护理机器人对心理和生理状况都大不如前的年迈父母的影响将会更大。再举一例，如果人工智能辅助教学，但并不破坏师生之间的个人关系纽带，那么它的发展应该得到鼓励，因为这种纽带是支持终身学习承诺的基础。总之，在当今时代，现代的和可辩护的儒家价值观能够和应该确立一些标准，帮助我们区分哪些是值得向往的人工智能技术，哪些是有可能危害人类的人工智能技术。

在当今时代，现代的和可辩护的儒家价值观能够和应该确立一些标准，帮助我们区分哪些是值得向往的人工智能技术，哪些是有可能危害人类的人工智能技术。

简而言之，马克思主义传统提供了在中国背景下应对人工智能挑战的重要视角，不过，我们在思考问题时还必须以儒家传统的深刻见解作为补充，才能更深入地理解问题。

贝淡宁（Daniel A. Bell），山东大学政治学与公共管理学院院长，清华大学教授。主要作品有《贤能政治》《城市的精神：全球化时代，城市何以安顿我们》《城市的精神2：包容与认同》。

---

[1] 在我们讨论特定背景下的伦理系统重要性时，该系统的根源并没有直接相关性。佛教起源于印度，但是因为它在中国背景下的长远影响，它被视为中国伦理传统的重要组成部分，正如基督教被认为是美国伦理传统中的重要组成部分一样，虽然它最初源于欧洲。虽然如此，源头这个事实具有间接的相关意义，因为人们常常对于源自本土环境中的伦理传统有一种自豪感（比如，在中国人看来，儒家和道家比佛教和马克思主义更具有中国特征）。

[2] 参见 https://www.marxists.org/archive/marx/works/1845/german-ideology/ch01a.htm。

[3] 参见冯象，《我是阿尔法——论人机伦理》，《文化纵横》2018年6月刊。

[4] 参见，“A study finds that nearly half of jobs are vulnerable to automation”, The Economist, April 24, 2018 (https://www.economist.com/graphic-detail/2018/04/24/a-study-finds-nearly-half-of-jobs-are-vulnerable-to-automation)。

[5] Max Tegmark, Life 3.0: Being Human in the Age of Artificial Intelligence,Alfred A. Knopf, 2017, pp. 126, 128. 公平地说，这样的论调在之前总有详细的讨论，多是有关技术导致美国的不平等加剧和政府没有采取措施应对这个趋势的观察。

[6] Max Tegmark, Life 3.0: Being Human in the Age of Artificial Intelligence,Alfred A. Knopf, 2017，pp. 218—245。

[7] 中文转引自《资本论》第一卷第13章“机器和大工业”，人民出版社，2004年，第486—487页。——译者注

[8] 中文转引自《马克思恩格斯选集》第二卷，人民出版社，2012年，第232页。——译者注

[9] 中文转引自《马克思恩格斯选集》第三卷，人民出版社，2012年，第364—365页。——译者注

[10] 中文转引自《资本论》第一卷，人民出版社，2004年，第208页。——译者注

[11] 中文转引自《马克思恩格斯选集》第一卷，人民出版社，2012年，第165页。——译者注

[12] 中文转引自《资本论》（节选本），人民出版社，2018年，第160页。——译者注

[13] 马克思承认可能仍然需要管理机构来完成协调和登记等工作，请参阅BertellOllman, “Marx’s Vision of Communism,” https://www.nyu.edu/projects/ollman/docs/vision_of_communism.php，但是我们可以设想这些任务由先进的机器来完成。在古希腊存在严格的二元对立，一边是决定政策的公民，一边是实施这些政策的中立的公务员。但是，公务员只能为政治共同体服务，他们没有选择自己的生活方式的自由——他们为国家服务，从事管理性和官僚性的“中立”工作，执行“政治”公民就公共政策进行协商和制定的决策。参见Paulin Ismard, La Démocratie Contre Les Experts: Les esclaves publics en Grèce ancienne, Éditions du Seuil, 2015。当然，让政治制度不公平的是，它的运行依靠人们从事社会需要的工作以及私有财产和公共奴隶的存在。在共产主义社会高级阶段，人们将不需要强制性的机构，不需要先进机器帮助做家务，社会也不需要为了共同体运行良好而做着相对乏味和重复的官僚及行政工作的公务员。但是，一旦机器拥有意识并感受到了痛苦，这个体制将不再具有道德合理性。

[14] 马克思没有预测到机器最终比“充分发展的个人”更聪明的可能性和人类有可能被机器主人奴役的风险。考虑到这个可能性，我们希望国家不复存在可能还有些为时过早，人类有必要拥有强大和能干的亲人类组织来代表大多数人的利益，而不仅仅是为“充分发展的个人”组成的社会转型提供方便。而且要确保人类维持其对机器的支配地位，机器应该为人类服务，而不是反过来要人为机器服务。只有强大的国家能保证我们的支配地位，或者至少尽可能延长这个阶段。从长远来看，人们希望这样的担忧如果不是从现在起往后几百年至少也是几十年之后才会变成实实在在的政治议题。

[15] The Works of Mencius, http://ctext.org/mengzi.

[16] 值得注意的是，儒家传统承认国家的永恒需要，这和马克思主义传统不同。汉代的儒家经典著作《礼记》描述了像共产主义一样的理想社会，被称为“大同”，但是这个社会仍然具有政治制度的特征，旨在选举具有高超能力和美德的人作为统治者。

[17] Roger T. Ames, Confucian Role Ethics: A Vocabulary, University of Hawai’ i Press, 2011, p. 261.

[18] Daniel A. Bell, “Roles, Community, and Morality: Comment on Confucian Role Ethics,” Appreciating the Chinese Difference: Engaging Roger T. Ames on Methods, Issues, and Roles, SUNY Press, 2018, pp. 203-212.

[19] 贝淡宁，李万全译：《超越自由民主》，上海三联书店，2009年。在中国的新社会信用系统中，孝顺父母与否可能影响人们的信用等级。（显而易见，这样的政策在并没有如此看重孝道的西方社会可能会引起巨大争议。）

[20] 施加限制的观念暗示可能性的封闭性而不是开放性，但是相反的情况或许才是真实的：正如对语言的限制促成了语言的创造性使用（否则人类的声音可能具有随意性，没有了意义），角色的限制允许我们表达创造性。参见Henry Rosemont, Jr., Against Individualism: A Confucian Rethinking of the Foundations of Morality, Politics, Family, and Religion, Lexington Books, 2015,pp. 16, 103-104。

[21] 参见罗定生、吴玺宏，《浅谈智能护理机器人的伦理问题》，《科学与社会》2018年第1期。从积极的一面说，消除照顾老人的沉重负担或许有助于缓解当今人们与需要关照的亲人的情感依赖逐渐淡漠的问题，这是有可能的。护理机器人的发明可能改善人们与家人的亲情依赖和改善其他人际关系，因为这些不会受到我们回避不愉快的护理任务的影响。参见Jason Borenstein and Yvette Pearson, “Robot Caregivers: Ethical Issues across the Human Lifespan,”Robot Ethics: The Ethical and Social Implications of Robotics, The MIT Press, 2014, p. 261。但是，护理工作的沉重负担由外貌无趣甚至丑陋的机器人来完成，可降低年迈老人更加依恋机器人而不是成年子女的风险。同样的道理，儒家学者或许更喜欢丑陋的手机或者药片，从而降低小孩子过分依赖技术而不是人的风险。

人类意识与机器意识的演化

刘晓力

导言 摩尔定律和莫拉维克悖论

2015—2016年，一些名家关于人工智能未来发展的预言似乎给社会带来一片恐慌氛围，特别是在2016年3月AlphaGo来了，人们似乎实实在在地感受到，人工智能已经威胁到人类，人类的智慧真的有可能要被机器取代了。目前，人工智能技术的发展看上去确实具有了一种革命性变革的趋势，人们担心机器真的有可能如很多科幻电影描述的那样像人一样有意识，有一天机器真的会自我觉醒，具有自我意识，操纵人类，甚至毁灭人类。

尽管每个人都相信人有意识和自我意识，但是机器究竟有没有自我意识，能不能以及何时会发展出自我意识，是我们知之甚少也知之不确的问题。几乎所有的人工智能科幻电影都会告诉我们，机器终究会有一天产生自我觉知，会自我觉醒，与人类进行生存的智力竞赛。那么，为什么我们现在看到的机器行为却还不如一两岁的小孩子灵活？

人工智能技术领域有个尽人皆知的摩尔定律，即机器硬件性能每18~24个月会翻一番。但今天流行一种莫拉维克悖论。这是由机器人学家莫拉维克、布鲁克斯和人工智能开创者马文·明斯基提出来的。他们认为，机器依据已有规则做高阶推理、复杂计算、知识竞赛等事情，消耗的计算资源相对少一些，但让机器自如地应付环境以获得较

强的感知和行为能力则需要消耗更多的资源。目前人工智能在实现某种弱人工智能的意义上已经超过了人类，特别是在超大型工程计算和工业机器人领域，甚至远远超过了人类智能。对那些需要长时计算的巨大算力，或者单靠人力不能完成的一些繁重和烦琐重复的工作，机器做得更快、更好。然而，让一个机器人像一个小孩子一样去感知外部世界并自由地应对环境却是非常困难的。1~2岁的小孩子具有非常好的处理环境信息的能力，能很好地平衡自己的身体以躲避危险，有非常灵活的感知行动能力和自适应能力。而且，运用这种能力时，小孩子是不需要什么逻辑推理和精心算计的。这是因为大量的感知行为能力是依赖许多默会知识、人类的情感以及一定的自我意识的。而默会知识和情感是很难算法化的，要机器像人类那样具有自我意识，就是更加难以想象的事情了。

可以说，今天的人工智能还没有摆脱“无心的机器”状态，这里的“心”英文是mind，有时译成“精神”，有时译成“心灵”或“心智”，其内涵非常丰富。实际上，人类心灵包括知觉、认知、行动控制和情绪情感多个层面。另外，人有心灵，是指我们有意识和意识体验，有自我觉知。我们说今天的机器还是“无心的机器”，至少有三方面的理据：第一，机器不能理解符号的意义和物理世界的意义；第二，机器没有意识和意识体验；第三，机器缺少自主性和自我觉知。

无心的机器与人类最大的不同是没有意识和自我意识

单从功能和外部行为来看，今天的机器人能踢足球，能下棋，能演奏乐器，能作诗，能跳广场舞，能打太极拳，有表情，会说话，计算机程序甚至还是新闻写作、知识竞赛和玩难度超高游戏的高手。为什么我们还说机器人没有意识？下面我将逐一来分析“无心的机器”遭遇的几大困难。

无心的机器面对的第一个困难：机器不理解意义，不能主动建立与外部世界的关联并理解关联的意义，特别是机器不能理解人类语言的意义。

人类的内在世界与外部世界具有特定的意义关联，例如，我们内心的所思所想，多半都是有所指向的，指向外部世界的对象，也包括进行反思的内部对象。比如，“我想吃宁波小吃”这个愿望是有所指向的，就是指向“吃宁波小吃”，“吃宁波小吃”就是我心理欲望的内容。再如，“张三相信机器能超过人类”，那么“机器能超过人类”是张三“相信”的内容，是“张三相信”这种心理状态指向的东西。所以，人的心理活动指向的对象是与内部心理或外部世界有关联的，这个有所指向的关联属性在哲学中叫作“意向性”。但是，绝大多数人相信机器目前还不具有意向性。

由于经过几亿年生物演化和几百万年智人进化而来的人脑具有独特的构造，我们的心理状态自然地、自发地与外部世界建立起了一种意义关联。人能主动地不需要外界的动力就积极与环境进行互动，而且人对于这种主动性还有某种自我觉知：我不仅能觉知到我是当下这种头疼感受的体验者，而且是自己行动的自主发动者，如果需要，我会为自己的行动提供理由。这是目前机器所完全缺失的。

人类在使用语言描述世界时，就与外部世界建构起了意义的关联，而且人是这种意义的制造者。

为什么我们能与外部世界建立这样的意义关联？另一种好的解释是，由于进化的结果，人与其他动物的一大区别是，人不仅能借助抽象符号进行思维，而且能用抽象的符号表达思想，在有了印刷术和更复杂的技术之后，人还能将思想印成书籍文字或制作成图像、音像及影像传播出去。事实上，人类在使用语言描述世界时，就与外部世界建构起了意义的关联，而且人是这种意义的制造者。动物当然也与外部世界有联系，借用自然之物为己所用。但是，人类会借助抽象符号，通过微妙的身体语言和人与人之间的话语交流，学习并共享知

识，制造和使用人工造物（从简单的工具到复杂的机器）。重要的是，这些抽象符号背后是有语义的，人类成员能达到相互间的意义理解并掌握运用知识。

因此，我们需要进一步讨论以下问题：计算机程序有没有那种内在的自发的与世界的意义关联？计算机程序运行的符号的意义是自主产生的，还是机器设计者和人类程序员赋予的？如果计算机程序承载着人类抽象语言的语法结构，机器是否懂得人类语言所表达的意义？

最近10余年，我们看到，通过深度学习训练，多层人工神经网络可以执行各类特殊任务，但是我们也知道，在进行图像处理、语音处理，包括各种引擎搜索时，其所做的都是完成识别、搜索、过滤、匹配的任务，完全不涉及意义理解，其算法运行过程不过是某种函数的拟合而已。

例如，大家都感受到一代又一代新手机越来越智能了，而且语音输入很快就能转化成多语种的文字传输出去，相当于机器语音识别系统做了即时的语言翻译工作。那么，翻译过程中是不是包含理解语义的过程？谷歌翻译软件到底懂不懂英文与中文及其他语种之间的意义转换？目前看来，我们可以说它不懂。例如，风靡全球的索菲亚看似可以自如地与人进行有表情的交流对话，但她实际上不能真正理解人类语言，她的对答都是预定的某一种选择的结果。因此，自然语言意义理解的问题，仍然是目前人工智能难以解决的一大瓶颈问题，也被称作人工智能的“符号落地难题”。

自然语言意义理解的问题，仍然是目前人工智能难以解决的一大瓶颈问题，也被称作人工智能的“符号落地难题”。

要让机器与外部世界建立意义关联，主动地和包括人类在内的外部世界互动，做出值得人类信赖的行动，还需要机器具备理解物理世界与人类关系的能力。今天的机器，如无人驾驶汽车以及各种在物理环境中活动的机器人，其所作所为可以说与外部世界有所关联，但这不是建立在理解外部世界意义、理解人类意图的基础上，而只是在实

现某种特殊的功能，它的行为是需要人类给予解释的。如果让波士顿动力公司的仿生机器大狗上战场，它根本不知道“战场”为何物，何为“敌人”，何为“战场上的危险”。因为“战场”也好，“敌我”也罢，都需要人类建构其语词的意义。如何与外部世界互动是人工智能的另一个瓶颈问题，被称作“物理落地难题”。因此，新近有人提出，图灵测试只在语言层面检验智能是不充分的，应该加入社会物理要素，修正快递机器人或老年护理机器人的物理测试标准。

无心的机器面对的第二个困难：机器没有意识和意识体验。

人是一个生命有机体，这个生命体的独特之处在于他是有意识和意识体验的。例如，人除了有意向性，有信念、愿望、意图，会推理，能规划，还有喜怒哀乐、爱恨情仇，有疼痛的感觉，有沮丧无聊的情绪，有对他人的同情之心。显然机器目前完全没有这种意识和意识体验。比如，我正在一个坐满听众的讲堂做讲座时，看到光线照进来，讲堂里蓝色的椅子反射出不同的明暗，四周墙壁的色彩也非常柔和、具有立体感，再加上听众聚精会神地安静听讲，整个画面给我一种非常和谐的感觉。这种感觉就是一种意识体验。当我看到如此这般的色彩明暗，感受到画面的和谐时，我就产生了一种温馨惬意的感受，这就是我的意识体验的一部分。然而，如果换作另一个人站在这里，也许他与我会有完全不同的感受，这说明个体的意识体验是依赖主体的。一个人的痛觉、喜怒哀乐等情感都是独特的个体经验。一个人依据个人体验推及他人可能也会产生相似的体验，这就是同理之心。如果说今天的机器相当智能，那么，它们有没有类似的意识体验，有没有同理之心，懂不懂什么是和谐的感受？我们当然认为目前它们确实没有。与AlphaGo对弈时，有人说正是由于李世石情绪波动、焦虑输赢才被击败了，而AlphaGo始终镇定自若。那是因为AlphaGo没有意识体验，没有焦虑，没有喜也没有悲，根本不能体验输与赢的不同感受。在电影《我的女友是机器人》中，主人公花了60年时间，终于造出了一台有情感的机器人女友，这就隐喻着使机器有情感是件相当困难的事情。

那么，我们到底怎么才能把这样一种类似人类的意识和意识体验变成计算机程序，让人工智能实现有情感、有意识，可以和人类交流，使得未来人机共生的时代，人类与机器可以更好地相处呢？可以说，尽管现在很多专家在研究机器意识、机器情感，甚至尝试建造有情感、有意识体验的机器人，但是连他们自己也认为，离机器真正具有意识和情感还很遥远。

无心的机器面对的第三个困难：机器没有自主性和自我觉知。

今天的机器是没有自主性的，它们行为的驱动力并非来自内部。例如，在李世石与AlphaGo对弈时，如果拔了电源或中断网络，它的外部驱动力就没有了，也就没有了当下的智能。而人是有自主性的行动者，具有与外部世界积极互动的主观能动性。最重要的是，人有自由意志，在面对选择或需要采取行动时，无论是发自内心的愿望，还是经过各种利益权衡，我如何选择或如何采取行动，都是由我的独立意志决定的。当然，机器也会在各种备选答案中做出某种行为，但那终究是在算法和程序输入设定的选择空间中做出的选择。另外，人的行动大多是有动机和意图的，人会规划未来，想象未曾发生的事件，可以预测当未来情境与现实情况不同的时候应当采取什么行动。最重要的一点是，人是有自我觉知的，知道自己知道的，也知道自己行动的动机，可以为自己的行为提供辩护的理由。那么机器，比如AlphaGo走哪一步棋，它自己会不会为它的走子行为提供理由呢？有人说AlphaGo下的那招臭棋是为了欺骗李世石，它真的会欺骗吗？它有没有一种欺骗他者的觉知，这当然是非常有意思的话题。

当AlphaGo与李世石刚刚结束对战不久，我们就在中国人民大学专门开了一个人工智能与哲学的研讨会，讨论AlphaGo到底有什么智能，随后我在《光明日报》发表了一篇小文，讲AlphaGo实际上是一个具有增强式集群智能的延展认知系统。下面我就从AlphaGo并不仅仅是一个单纯的算法或程序这个角度来分析一下机器的行为与有意识、有觉知的人类的行为究竟有何不同。实际上，AlphaGo是由走子网络、估值网

络、快速走子程序和蒙特卡洛树搜索四大算法构成的一个算法系统操纵的。除此之外，人机大战现场的机器配备有1 200个中央处理器和170个图形处理器，还装备了40个在线网络平台。云端大数据提供了历史上人类积累下来的3 000万个棋谱，可以利用大数据搜索调取棋谱进行即时的高速信息处理。同时，为这场赛事还专门铺设了一条光缆。甚至参与设计和实施赛事的人员中还有一批年薪200万～300万美元的深度学习领域的世界顶级专家。

由此可见，这个大系统并不像大众所理解的只是一个下围棋的计算机程序，它实际上是一个人机交互的巨无霸耦合体，一个为处理围棋这个专项任务建构起来的延展认知系统。所谓延展，就是把个体大脑的思维和智力活动，经过与一些外部设备的关联延展到大脑之外，把很多智能任务卸载到机器上做功能处理，比如大数据、云计算、中央处理器、图形处理器，再把当前棋局与人类棋手依赖直觉、依赖经验或依赖知识积累的围棋招数比对，与3 000万个棋谱进行匹配计算。所以，与李世石较量的AlphaGo，实际上是一个巨大的延展认知的智能系统。

那么，这个系统的智能到底是什么？我认为，这是一种包括了人的智能和机器智能的增强式的集群智能，一个人机交互系统的智能。在这个系统里，核心的还是人的意识和人的具有主体性的心智在主宰。当然，一旦弈棋的算法确定了，它的运行过程还是自动的。

人类意识与机器意识是否遵循不同的演化道路

最近十几年，人工智能界开始探索如何开发具有意识甚至有自我意识的机器人。设想，如果机器能自主编程的话，我们会不会说它有自我意识了？

在庆祝伟大的逻辑学家、数学家、人工智能鼻祖之一冯·诺依曼先生诞辰60周年的会议上，逻辑学家哥德尔曾经提出一个问题：机器

知不知道自己运行的程序？这是20世纪70年代提出的问题，当时在座的人都认为没有。今天，谷歌正在尝试让机器自我编程，让代码写代码，在他们看来，如果这项自主技术开发成功的话，这将成为一个革命性的机器自我进化的标志，以后也许就不需要计算机博士和那么多软件工程师了。然而，机器显然还没有达到能完全自主编程的程度。

在2015年的国际人工智能联合大会上，艾伦脑科学研究所首席科学家科赫做了题为“什么是生物意识及人工智能的意识”的主题报告，探究人工智能能否像人一样有意识和自我意识。正是那时，我在全国科学哲学学术会议上做机器意识与人的意识的主题报告，国内学界很多人认为这个想法太超前了，但AlphaGo出来以后，大家一下觉得确实大有必要展开这一话题的深入讨论了。

目前一款叫iCub的机器人颇受欢迎，有条件的人工智能实验室都会买这样的机器人用于开发研究。制造它的实验室的负责人声称，他们的下一个目标就是要让iCub拥有理解能力和自我意识。而且，我们知道，每个人类个体都是有人格、有个性的，开发者希望不久就能制造出有机器格的iCub，让每台机器都能有独特的个性。现在的iCub既有表情又有一定程度的对话能力，还有很好的感知能力、抓握物体的能力和运动平衡能力，但是不能说它们具备理解力，更不能说它们具有机器格。那么，什么是非人格的机器格呢？这个问题就与人类自我和机器自我的问题相关了。

自然界为什么会进化出人这样具有自我意识和不同人格的动物，却没有进化出类似人的具有不同个性特征的机器，这确实是一个有趣的话题。人类有意识的心理状态究竟包括哪些部分？首先包括通过我们的感知系统，如眼、耳、鼻、舌、身等获得的知觉；其次包括记忆的系统，如情景记忆、语义记忆，以及一些过程性记忆，例如什么在先，什么在后的记忆；最后包括有认知的各种状态，如音乐的、数学的、文学的和科学的认知，这些认知的方式和渠道是不一样的。此外我们还有情绪情感的部分，比如无聊、高兴、恐惧焦虑等，这些都是

非常丰富的有意识的心理状态。人的自我意识则是更高一级的意识状态。

按照生物心理学的研究，人的自我意识当然也是进化的产物。人一般在1周岁左右会发展出初级阶段层面的自我，这第一层面的自我叫表象阶段的自我，这个阶段人会把自己当作他人看待，用名字称呼自己。例如，小明会把自己叫“小明”，会说“小明要……”，而不是“我要……”。再大一点儿时人会发展出第二阶段层面的自我，学会说“这个是我的”，但是这个“我”的概念还是指将某某东西归属于我。能主动用第一人称的“我”来代替自己的名字，这时初步的自我意识就产生了，这多半要到2周岁。第三阶段层面的自我是在儿童发育到3~6岁时完成的，即能够自我评价，进行自我行为的控制。这一阶段人就发展出相对完全的自我意识了。有一个著名的自我意识的“镜像测试”，即给小孩子的额头上点一个红点儿，让他们站在镜子面前，看他们会不会去摸自己额头上的那个红点儿。一般来说，18个月大的孩子都能通过测试。这个实验最早是在大猩猩身上做的，大猩猩是可以通过测试的，而猴子不经过特殊训练是通不过测试的。

此外，明斯基认为，还可以从跨时间的三个维度考察自我。第一个是先天的遗传维度的自我。承载特定遗传基因的生物意义上身体的自我维度，是生物意义上的一种自我。第二个是文化遗产维度的自我，是经过后天的社会化和文化教养获得的一种自我。这种自我是与社会关系中的他者互动形成的自我，是一种人际的自我。第三个是具有心路历程维度的自我。它依赖一个人的意识经验过程，构成所谓自传式的自我。

最早从理论上提出“自我”概念的是哈佛大学的心理学家威廉·詹姆斯。他认为，意识和自我意识是进化的产物，是为了有机体达到生存的实用目的演化而来的。人类之所以发展出自我意识，是因为要在有限的资源下抵御敌人，在环境中有效生存。机器显然没有数亿年的人类物种的进化史，机器的自我有没有可能实现，或者说，具备什

么样的条件才能认为机器有意识和自我意识，这正是机器意识和机器自我研究的出发点。

机器意识与机器自我的初步研究

机器意识与机器自我的研究主要是人工智能领域的课题，20年前就有人在探索，但最近几年受到更多关注，中国的一些学者也在逐步进入这一领域。目前人工智能要建造有意识的机器分为两大策略：一个是纯粹算法的策略，包括自上而下的符号计算和自下而上的统计计算策略；另一个是类脑策略，主张受人类大脑结构和大脑工作机制的启发进行有意识与情感的人工智能研究，这是最近几年人工智能研究者与脑科学家联手探索人类心灵本质的新的研究方向。国际上探索大脑机制的脑计划有2013年奥巴马政府的“人类大脑研究计划”，其目标是做出正常的人类全脑神经元图谱，类似于人类基因组计划的做法。2013年欧盟也宣布实施人类大脑计划，目的是进行人脑的计算模拟，用芯片创建和复制人类大脑。中国新近也启动了大脑计划，目标是将大脑的基础理论研究成果运用到医疗、教育及人类增强技术等领域，以提高人类生活质量，增强人类能力。

人工智能的先驱者明斯基在20世纪80年代有一句名言：“人不过是肉身的情感机器。”你说你有喜怒哀乐、爱恨情愁，有自我意识，其实你不过就是一台有情感的以大脑为资源的机器而已。2006年他写了一本长篇著作就叫《情感机器》。

2003—2005年，一些学者提出建造有意识的机器系统需要满足如下几个必要条件：第一，要有感知状态；第二，要有意象能力；第三，要有注意能力；第四，要有规划能力；第五，必须有情感。其中意象能力要求系统有一种建构图景或设计情景、联想情景的能力，包括建构世界的一部分“像是什么样子”的感受性能力。例如，从未去过希腊的人要制订希腊旅行计划，首先大脑会聚焦世界的一个部分并

去想象它，创建一些希腊像什么样子的意象，其次再去设计未曾去过的地方的旅游路线和行程，预测未曾经历的事件会怎样发生以及如何应对。

其中第5个条件非常重要。机器必须有情感状态才能算真正有意识，知道如何分辨好坏、善恶，懂得做价值评判。机器要有进行道德判断并做出有道德行动的能力，它必须有自己的价值系统。因此，如果人类期望未来的机器不危害人类利益，成为值得信赖的智慧机器，仅仅把人类价值的规范自上而下地植入机器系统是不够的，还必须让机器自下而上地自主学习伦理道德，产生自发的道德情感和道德觉知，有意识地做出既不伤害人类又兼顾机器需求的道德决策能力。

2011年后，一些学者开始考虑，要达成此目的需要在上述5个条件中再增加一个自我意识，使机器意识更为丰满。“认知心理学之父”奈塞尔著名的“五个自我面向说”成为今日机器自我研究的指南之一：第一个是生态自我面向，即能够充分利用外部环境资源，在现实世界与环境互动的自我面向；第二个是人际自我面向；第三个是跨时间延伸的自我面向；第四个是观念自我面向；第五个是私密的自我面向。所谓跨时间延伸的自我，是指在生命过程中，一个个体仍然保有自我的同一性特性。所以，建造有自我意识的机器人就要从这几个方面去理解自我和实现不同维度的自我。

我们都知道，人类有一种非常强的独特情感能力，那就是共情的能力，他人痛苦时我们会怀有恻隐之心，我们也相信情同此心，心同此理。人类正因有共情的能力才能达成人与人之间的理解。只有让机器也具有这种同理之心，它才有可能理解人类。1992年，曾任欧洲脑与行为学会会长的里佐拉蒂及其研究小组，通过研究恒河猴的前运动皮质神经元工作机理，发现了猴脑中共情的“镜像神经元”效应，其后人们在人脑中也发现了同样的机制。这一工作为研究人类如何具有同理之心，如何具有模仿学习的能力，甚至如何建立自我与他者的关系，共享外部世界经验等提供了重要的认知心理学的说明。这一发现

也被誉为20世纪最重要的心理学成就之一。所以，研究机器的人际自我的专家希望借鉴镜像神经元理论去建构有模仿能力和交流能力的机器人。

例如，让几个机器人在一面镜子前做各种动作，再重放动作录像，让每个机器人指出哪个动作是它做的，据说中科院自动化研究所的机器人可以通过这个实验。2015年，美国伦斯勒理工学院就曾做过机器人在镜子前通过动作相互交流的实验，这一工作是探索机器自我的一种尝试，他们声称制造出了有人际自我的机器。

此外，康奈尔大学机器人实验室主任、《3D打印：从想象到现实》的作者利普森则通过另外的实验探索机器如何具有自我意识。因为明斯基在20世纪80年代就主张，人的意识实际上是由两层脑实现的：第一层是指挥自己和外部世界打交道的行动脑；第二层是对行动进行监督的脑，相当于有一个对行动反馈进行价值评估的脑。在利普森看来，由此假设出发可以探索出一条有自我意识的机器人的制造之路。他相信，自我意识的一个重要特征应当是具有想象自我、建构自我图景从而进行未来行为决策的能力，特别是能够想象自己处在一个未曾经历的环境中，建立自我形象，并在这样一种自我形象的基础上进行行动预测和规划。利普森声称他们真的做了有两个脑的机器。这是一个装有一系列传感器，由8个马达驱动的4条腿走路的机器人。底层脑指挥机器人向前走动，高层脑评估底层脑所引导的行动能否正确达到目标，再根据环境信息反馈调整前进的方向、速度和姿势……这样，机器就建立了一种所谓的自我形象，能够评估自己的行动了。机器最初并不知道自己是什么样子，经过十几次实验后，它才开始把自己的形象与外部环境关联起来，通过调整自己越来越适应外部世界，也逐步可以预测路况稳步前行了。

为了进一步测试自我意识的程度，利普森大胆地把机器的一条腿卸掉了。一天之后，他发现机器人慢慢开始调整卸掉一条腿的身体，把只有3条腿的身体当作和外部世界打交道的自己，慢慢调整行为在环

境中生存下来了。利普森就此说，这表明机器进化了。他也因此主张，最好把机器人放在一个真实世界里，不要仅在虚拟世界里讨论算法，而是要让机器人到外面与外部世界直接打交道，这样它就会慢慢进化，知道所处环境是什么样的以及它在环境中的角色，就像把小孩子放在自然环境里面，他会灵活地与这个世界互动，具有与环境的自适应性一样。

如果人类自我与机器自我遵循独立演化的道路，前景如何

目前，关于机器自我的实验似乎都想模仿人类自我的进化模式，但对于人的自我和机器的自我是不是遵循同样的演化道路，存在巨大的争议。今天有一门叫认知科学的学科，是一门包括心理学、人工智能、神经科学、哲学、语言学、人类学、教育学等多学科的大型交叉学科，认知科学就是以跨学科的方式探讨人类智能，探讨人类心灵的本质是什么。目前成果最为丰富的就是脑科学和人工智能，人们一直在寻求通过模拟人类大脑实现有意识的人工智能的道路。然而，直到今天，脑科学还没有搞清楚我们人类到底是如何依赖大脑去工作的，我们只知道，如此这般思维做出如此行为的时候，某些脑区在活动，但是，人们还无法完全定位是哪些特定脑区刚好实现这些功能，我们还没有搞清大脑图谱的所有细节，人类对自己大脑的了解还处于盲人摸象阶段。

尽管大家熟知的DNA（脱氧核糖核酸）的发现者、脑科学家克里克说，其实人类所谓的喜怒哀乐、爱恨情仇，我们的一切行为，我们的一切所思所想不过是大脑神经元的生物化学反应而已。难道机器没有这样一个生物的大脑就没有智能吗？我们是自然演化的，机器是人类制造的，机器自我的呈现是人工的数字化算法实现的，它是不是会走另一条通过生命演化出意识和自我意识的完全不同的演化之路？

2016年10月，霍金再次发表剑桥演讲，论述如果机器以独特的演化途径自我演化，演化出超越人的意识的自我意识，其前景是更糟了还是更好了。当然，他认为，人工智能很可能发展成一种完全与人类相冲突，有独立自由意志的东西，从而使人类文明历史终结于此，人类在自掘坟墓。2016年年底，在世界级的O’ Reilly人工智能大会上，与会者提出了12个权威观点。我们可以择其要把惊人之语摘录出来。人工智能是一个黑箱，我们不知道它将走向何方；人工智能不是个简单技术，有时我们想要土豆，结果却造出了红薯；人工智能必须考虑文化和社会背景；人工智能用处极大但并非万能，我们现在好像找到了一些好梯子，但是找到好梯子不代表能登上月球；人工智能是一种增强智能，它同时拥有人类的力量和机器的力量；人工智能改变了人机对话方式，但它需要懂得人类情感；别再用图灵测试了，换个更智能的测试吧，比如让机器满世界送比萨。

最后一个权威观点是质疑人工智能的探索路径是否正确。今天人们都是想在实现智能的某种物质载体，如大脑和整个有机体中找到人类智能，把大脑看作由神经元链接构成的一个产生智能的生物物理化学反应器，要把那些精神的、心灵的、意识的、自我的东西通通都还原成大脑和整个有机体这样一个物理系统中的东西，找寻到可以完全还原的物理规律来。这条物理主义的道路能不能走得通在学界也是存在争议、值得深入探讨的难题。有人说，现在人工智能走到了一个即将产生巨大变革的关节点，是真正需要大智慧和更多灵感的时候了。让我们共同期待大智慧的到来吧！

余论

人工智能如果完全是人类的创造之物，人类可以让它去做好的事情，也同样可以让它去做坏的事情。如果人类确实是有理性的智慧物种的话，就会引领机器的发展走上一条向善的轨道，但是我们并不能

排除某些疯狂的或邪恶的技术科学家会设计危害人类自身的机器，使其去做危害人类自身的事情。当然，至少目前这个控制权还在人类手中，在科学家和人工智能的技术专家手中。因此，我们需要制定相应的规范科学家和技术人员的工程伦理准则。从另一个视角看，如果机器真的遵循另一条演化道路，演化出了超越人类心智的机器，相信它肯定是更具智慧的，会比人类的价值观念更高一筹，相信它应该是更加向善、更可信赖的，否则何谈超越人类智慧。

如果真的有一天，邪恶的科学家和技术专家造出不受控制、作恶的机器，那就等于人类自掘坟墓，自毁前程。我相信到了那一天，人类退出历史舞台也是自然的命运。如果机器在人类的控制下，人类应该不让它作恶毁灭人类；如果机器不在人类的控制下，人类终将被毁灭，这种命运是不以人类的意志为转移的。人类在自然史上确实已经超过很多生物物种的寿命，而且人类对整个星球的生态破坏到达了危险的边缘，人类对其他生命以及对自我生命的漠视也达到了空前的程度，如果继续作恶下去，必将受到自然和生命自身规律的惩罚，人类自然也没有理由永久占据这样一个生态位了。

2016年4月1日我就说：“不作死，不会死。”如果人类作到底，也许就真正自我埋葬了。所以，我们应该警惕的是恶意制造和使用人工智能的人类罪人，而不是机器本身。应对人类未来挑战的解决之道应该是建构人机交互的良好生态和社会环境。理想的道德智能主体应该是人机融合的延展认知系统，其道德责任归属是分布式的。

当然，一个更值得反思的问题是颇具悖谬色彩的，机器有意识和道德觉知就意味着有行动的自主选择权利。如果真正具有自主性和自我意识的人工智能实现了，就意味着AGI（通用人工智能系统）时代真正到来，也就是机器可以说“不”的时代到来了。从人类未来的前途命运考虑，这种人工造物的一切行动能否持守人类的善根是大大值得怀疑的。人类创造了工具，工具也塑造着人类，人工算法的进化和技术的演化自有其作为社会技术系统的演进轨迹。如果作为嵌入社会环

境的“生命3.0”系统的演化是可能的，恐怕它也是不以人类利益一家独尊的意志为转移的。我认为，为未来人机融合的延展认知系统制定相应的伦理规范，需要考虑如下三个层面：个体主体的层面、个体自我的层面以及作为社会技术系统的层面。特别是作为一个道德主体的延展认知系统，应当在三个层面都有充分界定和描述，充分考量相应的道德伦理规范。

应对人类未来挑战的解决之道应该是建构人机交互的良好生态和社会环境。理想的道德智能主体应该是人机融合的延展认知系统，其道德责任归属是分布式的。

刘晓力，北京大学哲学博士，哈佛大学哲学系访问学者。现为中国人民大学哲学院教授，中国人民大学哲学与认知科学交叉平台负责人及首席专家，分别在科学哲学和逻辑学专业担任博士生导师。研究方向为逻辑学、心灵哲学与认知科学。曾出版著作《哥德尔思想研究》，在《中国社会科学》《哲学研究》等期刊上发表论文，主持国家社科基金重大项目“认知科学对当代哲学的挑战”。目前兼任中国自然辩证法研究会科学哲学专业委员会理事长，是《自然辩证法研究》《自然辩证法通讯》《科学技术哲学》多家杂志编委。

浅识、误解与想象力

——人工智能反思的中国语境

程乐松

相遇的两重境况：浅识与自觉的异质感

毋庸讳言，如果需要有效地将人工智能与中国的文化及思想语境结合在一起，形成完整且具有启发性的论述，就必须厘清人工智能及其中国语境的具体内涵，同时也要明确人工智能与中国思想资源之间可能存在的互动机制。对缺乏专业技术训练及当代技术史的专业性修养的观察者和思考者而言，人工智能是一个由陌生术语构成的“知识和语义深渊”：从结果的意义上来说，人工智能引致了极为丰富的想象空间和面向未来的可能性，并且以此方式推动了人类重新审视生命的本质与世界的价值基石，从而成功引致了绝大多数人文学科的反思立场及其思想资源的极大兴趣；然而，从发展态势与技术理解的意义上来说，人工智能是胶着晦暗的混沌。除了少数技术专家外，绝大多数关注人工智能话题的知识分子都对其中的机制和限制缺乏准确的了解。与此相对，人工智能在公众媒介和社会生活中的热度让任何一个与之有关的科技进展都变成某种公众事件和思想论战的阵地。更为重要的是，绝大多数此类论战或公众讨论的参与者都不是技术专家，而是一些依赖充满误解的、浅近的概念与常识展开的，基于想象力的谈论者。这样的境况是一种典型的知识性断裂，即绝大多数关注者从知识意义上来说并不具备参与深度讨论并且展开正确对话的能力。

人工智能引致了极为丰富的想象空间和面向未来的可能性，并且以此方式推动了人类重新审视生命的本质与世界的价值基石。

从1956年的达特茅斯会议到逻辑推理与概率统计的路线之争，再到本质主义与经验主义之间的跨越，直至类脑计算与深度学习创造的机器，人工智能以众多具有高度公众关注价值、颠覆常识的历史事件的方式，向人们显示了其在21世纪的蓬勃复兴及恐怖的潜在统治力。这种标准化的历史叙述，以及其中充满神话感和隐喻性的技术话语，将绝大多数的关注者都推入了一种复杂的矛盾情绪之中。一方面，人工智能的技术胜利又一次证明了现代理性及人的神话，掌控世界能力的跨越性提升大大满足了人类自身的想象力，人类甚至第一次获得了规划“自身”并从技术上创造“思想与情感”的能力。不妨说，人类第一次取得了“上帝视角”。另一方面，人工智能的技术黑箱与巨大的发展潜力，让我们集体进入了一种“近未来焦虑”状态。人工智能带来的本质性跨越，可能意味着人类自身将面临前所未有的挑战，乃至毁灭性的危机。喜悦与焦虑的交织，让人工智能成为言谈的焦点，同时也进一步远离由技术专家构成的“操作层”，变成了缺乏知识基础和界限性共识的话语形态。

简言之，巨大的技术壁垒和专业性术语鸿沟，让人工智能的公众话语空间拒斥任何可供分析的细节和精准描述，细节和精准描述的缺失又让人工智能在公众话语与社会思想的反思中呈现出一种奇特的低门槛效应和众声喧哗的态势。充满热情和想象力的讨论同时创造了大量具有想象空间的语词，例如“智能未来”“人工智能赋能”“智能时代”“人工智能世纪”“后人类智能”，这些语词缺乏精确的技术指向，但可以引致广泛且充满某种现实感的理解或误解。如果做一个简单的修辞策略的分析，这些语词都不约而同地指向以下两种特征。其一，某种具有未来感的历史性表达，即尝试锚定一个未来的时间段，而后用一个看似必然的技术想象让这一未来的时间段形成某种现实感，从而将未来推入历史。进一步说，这些话语正在用语义的模糊性实现的想象空间创造一种面向未来的“历史感”和“现实性”，绝

大多数的讨论者通过这些语汇将未来挤压进一种历史主义的叙述模式之中，从而在言谈中负担起来自未来的焦虑，并引发了以想象为基础的人文性反思与讨论。其二，用隐喻和换喻的方式，将技术术语改造成“日常语汇”，形成一种既熟悉又陌生的张力，在可亲近、可理解的表象中制造一种陌生的紧张感和期待感，从而将陌生的技术领域纳入日常话语的框架之中，引致缺乏理性边界感和知识实在性的空泛的理解。

人工智能的技术黑箱与巨大的发展潜力，让我们集体进入了一种“近未来焦虑”状态。人工智能带来的本质性跨越，可能意味着人类自身将面临前所未有的挑战，乃至毁灭性的危机。

在充满误解的想象力、技术上仍然在探索的专家们不得不做出的澄清和普及性描述、以商业利益为导向的结果性预测（甚至包括应和普遍性的职业焦虑展开的恫吓）这三者交织的状况之下，人工智能已经成为后现代的话语策源地，以及误解和洞见胶着在一起的言谈景观。我们在“漂浮的浅识”和“过剩的修辞”中展开（以人工智能及其技术后果的想象空间为名的）关于人类及其未来的全方位的持续言谈。然而，这种无根基且缺乏基本知识基础的讨论在多大程度上可以被视为合法的反思和正确的谈论呢？

不妨说，在大部分关于人工智能的人文性反思和社会性谈论中，人工智能只是类似中国传统禅宗公案中所见的佛教大师之间以“机锋”和答问的方式讨论佛法时使用的“话头”，并不是讨论的真正核心和主题，而是回向并重新激活本根性问题的一个触发点。如果上述观察是合理的——至少对以人文学科和中国传统思想研究为志业的笔者而言是如此，那么我们在谈论中就需要十分清醒地意识到如下两个基础共识。其一，我们并不能谈论人工智能的技术及其前景，而只能谈论它已经引起的社会效应和公众焦虑。换言之，我们的谈论不过是一种次生性文化效应。其二，对于文化效应引发的一直吸引着人文学者反思热情的本根性问题，之所以需要以人工智能为出发点，是因为

人工智能以及公众对它的“浅识”带来的复杂情绪与基于误解的想象力为这些本根性问题提供了新的反思空间。反思空间可以带来新的认识，却并不能保证任何谈论的正确性和合法性。不妨说，我们所做的只是另一种基于“浅识”的谈论，它的基本功能只是提出问题，而不是提供任何解决方案。

如果需要将关注点集中在中国这一独特语境，我们就要先厘清中西文化相遇的内涵。与西方文化的相遇、西方文化带来的冲击及其响应，以及以此为背景展开的中国社会与文化的现代化探索，成为过去一个半世纪全球史的重要景观，也是在历史视角中实现世界史向全球史跨越的重要推动力。与西方的相遇，不仅改造了中国关于自身思想与社会传统的看法，而且以民族国家的方式重述了历史，形成了中西、古今双重二元的“异质性”和“对象化”认识结构。不妨说，中西文化的相遇及其带来的深远影响，成为当代中国思想最重要的底色之一。我们在谈论相遇的后果之前，必须先澄清的是更基础的问题：什么是相遇？这是一个十分复杂的问题，它在很大程度上可以决定我们如何从历史和当下的视角谈论全球景观中的中国与西方。

概而述之，不妨将“什么是相遇”这一问题解析为如下若干层次：这种相遇可以被视为一种事实描述、历史事件抑或进程、知识境况、社会冲击，还是价值观的融合与重建？强调中国与西方的相遇，是否清晰地界定了中国和西方的内涵——其中蕴含着对传统与现代的对举式和对应式的预设？更为重要的是，在中国自身的历史发展脉络和文化变迁中，从未停止过与域外文明的交流和融合，与西方相遇的中国有何特殊性？换言之，在面对西方时，中国为何保持持续的异质性？在关于中西文化相遇及其后果的谈论中，我们似乎有意无意地接受了某种静态化的事件性描述，而忽视了动态化的过程视角。中西与古今的二元结构，让中国与西方、传统与现代这两组对举的概念形成了某种隐喻性的对应。更为重要的是，在过往近两个世纪以来的历史进程中，中国与西方之间的二元关系似乎仍没有消解（甚至强化了）。中国与西方之间存在某种独特的镜像关系，我们总是自觉地尝

试将某些知识、技术与思考方式规定为“西方的”，从而在民族国家的历史性叙述框架中实现自身的异质化，这也造成所谓“来自西方”的内容的对象化。这种自觉的异质化有一个十分重要的预设，即中国与西方的相遇是在一个固定的二元框架和异质感的对象性设定中静态展开的。无论时间和空间乃至社会形态与文化内涵发生什么样的变化，都不足以改变这一事件发生的那一刻就已经确定下的二元结构——这一二元结构不仅塑造了中国的西方观，而且塑造了中国的民族认同和文化自觉。简言之，近代以来的中国是在与西方的相遇中认识并建构自身的。然而，作为一个被设定的历史事件，中西文化相遇是过去式，而不是进行时。历史事件之后，中西思想的持续互动，以及中国社会及其思想体系的巨大变迁，已经深刻改变了静态的二元结构——我们不再能找到“纯而粹之”“对向而立”的中国与西方了。然而，面对这一现实带来的关于中西互动的复杂后果的系列理论困难，“二元静态结构”仍然是在讨论中西古今的当下性时的一个“方便法门”。

正是在这样的静态化二元结构默认的前提下，人工智能与中国成了需要“相遇”的两个对象。在中国，无数人热烈欢呼人工智能及其时代的到来，并没有消解“相遇”的必要性，而是强化了它。关于人工智能的讨论只是进一步强化中国面向一种全球化时代的科技进步的自觉的异质感。在这里，我们不断混淆以下两个截然有别的概念：西方的和全球的。在全球史视角下，人工智能就是一种可能带来全球性产业及生活方式变革的技术趋势，全球任何一个经济体和文化传统乃至社会生活群体都不能自外于这个趋势，也都是这个趋势的参与者和创造者，以及后果的承担者。当然，从文化意义上来说，我们不能否认，人工智能的技术源头和对于这些技术的知识性与伦理性表述的确带着西方的底色，但这并不意味着人工智能必然是一个在“中西古今”的二元框架中属于西方和现代的那一极。从这个意义上来说，在一个整全性视角下的自我疏离与对象化，就是中国在不断参与全球文化景观与思想对话过程中仍不自觉地完成自我的异质化的典型案例。

在上述对象化和自觉的异质化视角之下，以中国的思想传统为研究对象的学者被要求就这种“相遇”提出自己的看法，而且这些在预期中的看法的思想资源及其论断的合法性应该来自数千年之前的传统中国的价值体系和宇宙观视角。这不禁让人联想起20世纪著名汉学家费正清在论及当代中西文化互动时提出的“冲击－反应”模型：人工智能作为一种西方底色的技术潮流，出现在“中国文化与思想”的面前，中国文化与思想如何响应和理解它？当然，还有另一种可能的预期，即来自传统的、中国的思想资源由于其独特的异质性特征，可以呈现出某种非西方的理解与思考，从而可以作为关于人工智能反思的整体图景的组成部分抑或补充？以此而论，被贴了标签的学者就需要同时使用“我亦有之”的自觉性和“借我山之石”的对象感。然而，正如我在前文中述及的那样，作为一种次生性的论述，关于人工智能的文化性反思的内容显然不是技术路径，而是回答由人工智能带来的挑战激活的具有本根性和共通性的反思课题。由此，其受众也不是有技术和解决方案依赖的技术人员，而是更宽泛意义上的处于关于人工智能的谈论和焦虑之中的公众。

以二元框架中的相遇以及具有历时性的“传统中国思想”的名义，为在中国语境中谈论人工智能明确方向性，很大程度上预示了以下两点结论：其一，与其说是为了澄清什么是中国语境中的人工智能以及人工智能在中国的描述策略，不如说是以人工智能为名再一次回向并尝试解释中国的独特性，同时强化中国作为一个类概念定语的内涵丰富性；其二，中国语境的文化属性决定了从中国思想资源出发的关于人工智能论述的可能价值在于构造一种独特的问题框架和反思视角，甚至可能是消解某些问题。通过提问乃至消解问题的方式，人工智能也可能再一次拓展传统中国思想的反思空间。当然，上述两个目标的达成，其根本前提仍是我们能够（透过漂浮的浅识和过剩的修辞）准确把握人工智能的技术内涵及其可能的后果。

交迭的陷阱：诸种形态的不确定

如果尝试用一种非技术的方式描述对人工智能的发展机制及其困难的基本理解，笔者认为可以围绕确定性与不确定性之间的张力展开。这种张力并非仅仅是技术意义上的，更涉及深层的反思性议题，我们可以从以下三个层次理解不同形态的不确定性及其形成的交迭关系。

首先，从不确定的输入到确定的输出之间构成了人工智能最大跨度的技术鸿沟。生活世界和人类意识的无限度不确定性仍可以完善地体现人工智能技术的局限。从图灵1936年发表《论可计算数及其在判定问题上的应用》开始，到1956年的达特茅斯会议第一次提出人工智能的定义，以符号方法进行语义处理实现了人机交互，开启了关于人工智能的技术探索和未来预期。人工智能的第二次浪潮将基于连接主义的专家系统推进到多层人工神经元网络，改变连接主义的技术面貌。近十几年以来，以行为主义和反本质主义的经验论模型为基础的深度学习与机器学习成为人工智能发展的未来方向。基于图像识别、语音识别、语际翻译和规则学习的一系列技术成果不断刷新公众对人工智能潜力的认识，也引发了关于人工智能如何塑造人类社会未来的系列猜测和预想。然而，从早期人工智能的专家系统开始，技术专家尝试解决的就是从高维输入向低维输出的转换，简言之，无论从语义还是规则的角度，人工智能面临的第一个挑战都是如何将输入端的高度不确定性透过技术和符号的手段转变为一种可执行、可预测的确定性。近年以来，以反本质主义的视角超越专家系统的深度学习，其背后的基本理念是行为主义与基于经验的试错和快速学习。可以简单地将这一过程理解为：尽量使用人类在习得某些技能时习惯的方法，通过多变量的反复情景模拟，生成具有容错空间的、具有弹性的解决方案，从而达致从高维向低维的转换。容错的空间来自数据的积累和场景模拟的积累。

然而，我们发现，人工智能的最新技术方向倚赖的是理论上不受限的计算能力和数据累积能力，即如果有足够长的时间和足够强大的计算能力，输入端的高度不确定性带来的可能性具有被计算和模拟穷尽的可能，人工智能可以不断趋近完美。这仍然是在确定性输出的层次上完成的量的突破，而非直接指向不确定性的质的飞跃。这可以解释人工智能成功打败世界围棋冠军的原因——在一个可预见、有规则的运算体系中，自主仿真加上计算速率，就可以很快突破人在这一运算体系中的认知极限，达到击败人的效果。显然，面对无限复杂却缺乏明确规则和运算体系的生活世界与人类意识，规则的缺失和模糊性构成了人工智能的界限。是否可以这样说，依赖机器的计算能力，以数量累积的方式击败不确定性，似乎是一项不可能完成的任务？要突破这一瓶颈，就需要进一步理解人类在日常生活中应对不确定性的机制，即如何展开类脑计算，改进深度学习的模式。

其次，类脑计算的尝试与脑科学的结合，一方面推进了人工智能的技术发展潜力，另一方面引致了一个新的循环性陷阱：人工智能专家要倚赖脑科学家对于人脑运作方式的理解和猜测制定类脑计算的模拟解决方案，并通过实验验证脑科学家的理解；与此同时，脑科学家对于人脑运作方式的最佳验证手段就是人工智能专家的模拟。这样一来，脑科学与人工智能之间就形成了一种知识－验证的循环关系，这种相互倚赖的循环模式在很大程度上将某种“揭示性模仿”作为人工智能的技术解决方案。然而，直接在这个语境中被凸显出来的核心问题是关于人脑的还原主义立场，即我们是否应该假设或者认为人脑的运作机制是可被完全还原的。

质言之，我们需要一个透明人假设作为基础，脑科学家总可以通过不断的知识积累和发现最终将人脑运作的所有过程都还原为某种可描述、可形式化且可控的物理化学或生物化学过程。在这个基础上，具有完全透明可能性的人脑与人工智能机器就可以完成一种系统性的对应，从而完全模拟出人脑的机制。在机制保障的基础上，机器在计算能力上的优势就可以保证机器彻底超越人的智能。那么，我们是否

可以说，类脑计算与人工智能发展的联结很大程度上倚赖一种还原主义的假设？如果是，那么人工智能技术发展的确定性建立在这个假设的不确定基础之上。同时，如果这种还原主义的假设被人工智能的发展证实了，那么我们对人类本质（至少包含了某种不确定的主体性和不可预测性）的所有认识都将被颠覆。从这个角度看，与人工智能的计算能力相比，人类的独特优势并不是确定性，而是本质意义上的不确定性。更为重要的是，以这种本质意义上的不确定性为基础，任何基于有限计算和认知能力的个体生活世界及其内容都变得难以还原和模拟。简言之，面对类脑计算与人工智能的循环，我们的两个选项是：要么坚持非还原主义的立场确证类脑计算潜力的有限性，要么坚持还原主义的立场并颠覆对人类本质的认识，由此进入一种机械主义的人类定义之中——其直接后果就是人类本质中值得被尊重的部分丧失殆尽。如果一种还原主义的立场被证实是正确的，那么人类被机器取代的现实一旦发生了，这一事实在多大程度上可以被视为“消灭”而非简单的“等量代换”？

最后，关于人工智能的应用前景及其最终将取代人类智能甚至消灭人类的预测，一方面带着令人兴奋的不确定性，另一方面挟持了面向未来的一种类似“可能性”霸权。在任何一种关于未来的描述之中，任何人都必须忌惮单向时间之箭带来的合理性，即未来是向一切可能性敞开的，可能性即合理性。在未来没有到来之前的所有可能性，就用某种独特的合理性纠缠着我们对当下的认识，并制造大量的焦虑与欣喜感。关于人工智能及其后果的各种假设，都被当作严肃的可能性加以讨论，我们正在面临一种“想象力即正当性”的窘境：可以被想象的，都是应该被响应的；所有可能的，都是需要被处理的。正如我们可以观察到的，很多人工智能的技术专家对于流行于公众之中的各种令人兴奋或焦虑不已的“预测性描述”都感到十分惊讶，因为对他们而言，所有的可能性与当下之间必须有可被执行的技术通道，他们对于这些技术通道的难度乃至被设想的可能性具有清醒的认

识。因此，技术实现的难度和基于解决方案的谨慎使大多数人工智能的专家对于类似的想象力持有一种敬而远之的态度。

大多数关于人工智能及其可能后果的讨论，都是越过技术实现的或然性及其中巨大的理论困难（我们仰赖的就是对人的理性能力和认知潜力的无限信任）直接指向后果之后的讨论。我们假设这些后果已经变成了现实，那么我们的解决方案是什么？这是一种很有价值且十分有趣的思想实验，正如伦理学中反复出现的“扳道工”思想实验一样。然而，我们在多大程度上可以认为思想实验就可以等同于现实？以后果之后的世界为出发点，制造和产生的焦虑（包括人类智能被淘汰、人类被消灭、人类社会的结构性颠覆）一方面来自思想实验，另一方面来自人类对于人工智能的基本假设，即被人类制造出来的人工智能也必然带着人类中普遍存在的邪恶与阴暗，因此，人工智能也可能会用超越人类的智能展示并实现难以控制的邪恶力量。与其说我们在放纵对人工智能的恶意想象，不如说我们在将我们对人类的恶意想象转嫁给人工智能。

以未来的名义，倚赖时间的霸权，我们通过思想实验将技术发展的不确定性转变为一种需要被反思的“确定性”。如果检视人类的思想和观念历史，不难发现，类似性质的焦虑伴随着人类思想的全过程，人类总是生活在一种源自（自身能力与基于生命限度的）有限性的毁灭性焦虑之中，人工智能技术发展只是制造了最新的一种焦虑形态，而不是创造了某种前所未见的焦虑。由一种技术意义上的不确定性产生了具有确定性的期待与焦虑，人工智能在这个意义上激发了关于人的未来命运与人类的自身选择之间复杂互动的反思——我们是否正在用自身的努力和不断的技术突破为自己开辟一条走向灭亡的道路而不自知？

人类总是生活在一种源自（自身能力与基于生命限度的）有限性的毁灭性焦虑之中，人工智能技术发展只是制造了最新的一种焦虑形态，而不是创造了某种前所未见的焦虑。

上述三个层次的不确定性是相互交迭的：与高维输入的不确定性之间的博弈构成了人工智能技术的内在机制，而如何超越并克服源自生活世界及其复杂性的不确定性似乎需要倚赖对人脑运行方式的确定性理解。吊诡的是，对人脑运行方式的最终确定性理解将在认知和自身理解的意义上终结人类的本质定义。以有限性与不确定性为本质的人类，通过时间的单向性和对理性能力的无限信任，制造并沉浸于一种“确定的可能性”或“可能的确定性”带来的焦虑和期待之中不能自拔。更为重要的是，其中的焦虑很大程度上延续着对人性自身恶意的想象。对人性内部的恶意的信念，必须建基于现阶段被普遍认可的反还原主义生命观之上。通过以上简述，我们可以看到一种以确定性、可能性和不确定性构成的层次复杂的嵌套关系，其中有许多难以自洽的逻辑陷阱。

当然，我们并不是要以上述分析来否认人工智能技术发展对于人文反思的价值，而是要厘清人工智能激发的反思性问题的性质，避免以可能性与现实性、确定性与不确定性之间的混淆带来的误解。与此同时，行文至此，我们是否并没有触及“预期中的”古典中国思想资源，甚至没有用经典思想文本中的语句作为某些问题的响应。在我们看来，在进入古典中国思想之前，应该先厘清问题的性质，以及人工智能带来的反思性课题的基本内涵。

技术的关切与关切的技术：中国语境的价值

从相遇的两重境况到不确定性的交迭，我们试图依照某种本质主义的立场穿过人工智能的技术壁垒和文化景观，提出可以激发传统中国思想资源的问题，并以此避免浮光掠影地涉及过多议题。如前所述，人工智能有明确的属人性，即它是人类创造出来的技术体系，它的真正基础是人类基于理性的理解力和创造力。人工智能的特殊性在于它呈现出了一种独特的潜力和趋势：超越作为其创生基础的人类理

性能力，产生一种超人类的智能。关于人工智能（特别是关于人工智能的潜在危险）的描述中，无一不包含上述可能性的“实现”。我们可以从中看到一种理性的神话——万能的理性可以完成对自身的超越：作为人类基础的理性可以创造出一种它自己都无法理解且无法企及的智能。这使我们不由地想起人类思想史上一个十分有趣的悖论——万能的上帝是否可以创造出他自己都无法举起的石头？

在深度学习的技术路线上，我们看到了经验主义与理性主义的妥协，其中的桥梁就是对计算能力和计算速度的迷信：由于超越人类的计算速度，无限的场景仿真和试错最终可以达致一种确定性。与此相对，类脑计算的“透明人”假设凸显了在当代公众中十分流行的对理性能力（往往被延伸至技术解决方案的能力）的无限信任，理性加上时间就可以实现持续的自我超越，最终达致一种本质的飞跃。理性的无限潜力的假设带来的另一个直接后果是时间感的终结，即理性最终可以将所有可能性变成现实，这就意味着时间本身只是一种现实的尺度，而不是可能性的展开与封闭过程。时间性的消解，其后果不仅仅是可能性与现实性的混淆，更可能造成人（相对外部世界）的彻底对象化，即人与世界之间不可调和的二元对立。在时间的流逝中，人不断与世界互动，创造出更新的世界及其意义模式，并且适应这些部分是由其自身造成的世界的变化，换言之，人是可变的。时间不仅展开了世界变化的可能性，而且确保了人不断的可能性。因此，时间性的消解让人成为需要以固定的姿态和僵化的现状应对不断袭来的被视为现实性的诸种可能，这样一来，人类就深深陷入了无解的关于有限性的焦虑之中。从上述简单分析，我们不难看到理性的傲慢结合一种二元性和对象性框架时带来的困境。

传统中国思想家从一开始就十分审慎地拒绝各种绝对性的判断，并且尝试通过对未来和变化的不确定性的正向理解保持对理性能力界限的清醒认识，因此也不会产生一种对人类认知能力的盲目乐观和缺乏边界感的预设。人与外部世界的关系是以两者之间的相互融入和相互适应为基础的，人与外部世界之间存在着某种适应的机制。这种适

应机制建基于传统中国思想资源中的整体主义视角，人与世界保持了神秘的关联和同构型。这种同构型很大程度上帮助传统中国思想家拒绝了人类优先论和人类中心主义的立场，更避免了对人类理性的过高估计。面对外部世界的谦卑，不仅来自生活经验中对自身能力有限性的切实体会，而且是传统中国宇宙论的自然要求。在这个基础上，传统中国思想家会在自然与人的关系框架中将人类理性的有限性转化为一种与未来的不确定性和解共处的机制：正是由于人类的有限性，不确定性及其一切可能性都是一种在人类能力可控范围之外的必然，这种必然并不是一种负担或困扰，而是一种自然或当然的状态。

另一个重要的特征是事物的非线性变化，事物的变化并不一定是单向和进化的，而是在一种有机主义（如果借用李约瑟颇具洞见但模糊的用语）的体系中实现事物之间的转换和可循环的持续变化。包括人类在内的世界万物之间存在着紧密的联系，不同事物之间的转换、同一事物的生灭与消长都是一个有机整体的组成部分。任何一个事物的变化，或某一事物的阶段性特征，都是这个整体的一部分。在这样的观念体系之中，并不存在一种沿着时间流逝的方向持续进化的单向性。任何事物的变化都是双向甚至多向的，不同方向的交织和层迭，形成了不同时段和不同视角的差异化图景。因此，当面向未来谈论事物及其变化的过程时，不确定性自然受到包容，未来仍然是敞开的可能性。更为重要的是，可能性的敞开并不会导致某种不可知的焦虑和无力感，而是在一个整体主义的图景中呈现出一种可被理解的合理性。在这个意义上，任何可能性都是可被接受的。当然，这并不意味着不存在某种内在的统一性和连贯性，而是这种内在的统一性需要通过直觉性的整体把握（而非基于理性能力的分析）才能呈现出来。对世界的整体把握，可以实现生命的超越，但这一超越不是通过制造某种外在对象完成对自己的超越，而是将自身与世界的对象化关系消解，在人与自然在本质意义上融合在一起时实现自身的内在超越。

以上述两点为基础，对传统中国思想家而言，人的有限性不能导出生命的还原主义立场，在一个有机主义和整体主义的世界图景中，

生命是完全不可还原的，因为生命和世界的本质一样既保持了某种神秘感，又预设了某种合理性。我们也不可能通过理性的分析完成对生命及其内在机制的完全描述——当然，生命的本质和宇宙的本质一样可以被直觉地把握到，然而，这种把握的状态是不能用语言描述出来的，只能是一种高度个体化的体验。人与自然共享的本体的内在一贯性与每一个个体拥有的独特性之间保持着一种微妙的平衡，由此，传统中国思想总是尝试达成个体和群体的某种既内在一贯又各自有别的态势。

不可否认，层次丰富、主题多元的传统中国思想体系的内容远远比上文的简述复杂，但我们想要强调的是，与人工智能在语汇的造作和浅识的漂浮中体现的对技术本身及其后果的关切相对，我们更应该从传统中国的思想体系中找到一种具有独特性的关切技术。同时，依照本文第一部分设定的自觉异质化的境况，另一个需要厘清的问题则是：上文中所述及的中国传统思想资源是否本身就是一种自觉异质化的产物？对我们而言，分析的前提是梳理和分析人工智能带来的各种课题背后的文化根基及观念源头，在这个基础上，尝试激活中国传统思想资源，塑造一种新的关切技术。这种关切技术是一种反思的视角，其反思的对象就是包括中国在内的全球语境中关于人工智能及其挑战的谈论与话语内容。从这个意义上来说，中国传统思想资源的激活是在一个自觉全球化的前提下展开的传统与当下的历时性思想互动。

由此，中国传统语境的价值就在于说明了以下两点。其一，关于人工智能的各种焦虑背后，有一些可供分析的独特观念基础。对于这些观念的剖析可以把人工智能的技术层次与关于人工智能的误解和浅见分开，避免一种具有误导性的想象力不断产生过剩的语词。其二，关于人工智能后果的各种预测和描述不是一种技术性的关切，而是一种文化反思和本根性问题的激发机制。它一方面凸显了文化反思的活跃，另一方面提示我们关切人工智能的问题需要更全面和多层次的视角—— 一种更节制的谈论问题的技术。更为重要的是，我们要认识

到，与其关注不同的思想资源可能以什么方式影响人工智能的解决方案与技术导向，不如聚焦于人工智能带来的文化效应及公众认知在多大程度上可以提升人文反思的活力。因此，我们并不在意在多大程度上解决了任何问题，而是以什么视角重塑的问题。人工智能及其后果只是本文展开讨论的出发点而非目标。传统中国思想资源的激活并不来自中西古今的二元结构，而是指向一种新的关切人工智能问题的方式。

附言

假设上述文字都是一个具有智能写作功能的人工智能机器人完成的，我们需要追问的不是人工智能机器人是通过何种技术或算法完成这些文字的，而是人工智能机器人在写作的过程中有没有“思路难以为继”的感受，有没有想要放弃写作的冲动，乃至会不会认真地想象上述文字被阅读的时候的通俗性。不妨说，上述文字的写作不是笔者作为一个有人类智能的个体的“功能”，而是一段生活。生活的不可还原性，个体感受的无限可能，让问题不仅仅指向某个事件的结果，更要指向它的全过程中无限丰富且完全无章可循的情绪及其背后“几乎”不受主体控制的复杂感受。不同的视角都在指向同一个结论——对于技术的关切，首先需要一种合理的关切技术。

程乐松，北京大学哲学系（宗教学系）副主任、教授、博士生导师，教育部青年长江学者（2019年度），北京市第十一届青年联合会委员。主要研究领域为道教经典与思想、早期道教史、两汉思想史及汉语哲学等，发表专著三部、中英文学术论文数十篇。

人工智能带给儒释道的挑战与机会

人工智能会帮助我们获得真正的美德吗

安靖如/文 吴万伟/译

哲学家、技术人员和专家现在开始承认人工智能的发展所带来的深刻伦理问题。[1]截至目前，人们的注意力主要集中在三个方面：（1）追逐利润的算法（注意力经济）如何危害和控制人类以及我们该如何应对？（2）如何确保自主和智慧的机器（如无人驾驶汽车）做出“好的”决策？如何定义这些决策是什么？（3）如何思考（或者抗拒）能超过或者控制我们的超级人工智能出现的可能性？就理论家观察人工智能带来的好处而言，他们的注意力往往集中在物质充裕和休闲娱乐的可能性之上。而且，截至目前，几乎所有理论探索都仅仅局限在西方的伦理框架体系内。尤其是考虑到中国已经在人工智能研发方面投资庞大，将我们的哲学框架扩大一些以便成为容纳更具全球性伦理视角的体系或许是明智的选择。

笔者认为，当我们从儒家视角观察人工智能时，自然要提出的问题是人工智能是否有助于让个人和社会变得更有美德。果真如此的话，即便帮助我们的智能是“人工的”，人类还是能获得“真正的”美德吗？探索上述问题的答案并不是儒家探索人工智能的唯一途径，这不仅是因为人工智能研究有很多维度，还因为在当今世界发扬儒家传统的合理方式也有很多。人工智能与美德的关系是所有儒家学者都应该关心的问题，也是具有广泛和全局性意义的重要话题。本文围绕儒家五大美德（仁、义、礼、智、信）展开。我们主要引用现成的、可实现的技术来论述人工智能与美德之间具有潜在积极意义的关系，当然也会考虑各种可能的反对意见。

首先，我们从理解本文理论所依赖的若干关键概念的定义说起。尼尔森对人工智能进行综述时认为，我们把智能理解为“让一个实体适当运行的品质，并且对环境有一种预见性”（Nilsson 2010，13）。这显然是一种不仅仅局限于人的品质，在此意义上，动物也表现出了智能。某些现有的人工产品已经拥有“智能”这一说法应该不会引起争议。当汽车提醒司机车辆已经进入另一条胡同时，或者手机建议机主应该动身前往约会地点时，这些人工产品的行为都是适当的，而且表现出了预见性。我举上述例子的目的是想扩大这个定义，以便包括依靠大数据或深度学习来实现行为的实体，这比我们在谈论人工智能时表达的意思更宽泛些。有关人工智能的很多哲学猜测多集中在增强版即通用人工智能（AGI，指在多样的环境下能够采取灵活的、有智慧的行为的机器）甚至超级人工智能（ASI，指在所有背景下都超过人类智能的机器）上。本文的论证完全没有依据通用人工智能或者超级人工智能的可能性。至于人工智能中的“人工”一词，只是简单地意味着这里的智能是人工的，而非仅是自然进化的结果。人工智能是人类能动性有意识地创造出来的数字的和/或物质的实体。

本文提出的“儒家视角”基于这样一个前提，即像过去2 000年来一样，儒家一直是活着的、发展中的思想传统和实践。儒家不仅仅局限于某个地方、某种语言或者某个族群之内，它扎根于很多个体的、多元化的见解中，存在于很多人践行的典范生活中，存在于很多社群遵从的礼仪活动中，存在于过去多个世纪演变中记录下来的重要经典文献中。儒家在不断成长，它根据新体验、新社会结构和新观点而做出变化和回应。几乎从一开始，儒家就存在内部辩论和多个派别并存的局面，这些合理的辩论继续推动了当今儒家的发展。在很大程度上，本文依靠的儒家基本观点，即便不是当今所有儒家都赞同的观点，至少也是大部分儒家支持者的共同财富。本文将利用不同哲学和心理学传统中的相关见解阐述这些儒家观点，因而遵循了我曾谈及的“有根的全球哲学”方法。本文讨论的结果不仅对已经受到儒家思维

方式影响的人有参考意义，而且对在其他理论框架下思考当今世界该如何幸福生活的任何人来说也都具有启发性。

最后，在开篇讨论结束之际，我必须对文中所说的“真正美德”的意思加以澄清。我们或许可以先从休谟著名的区分——自然美德和人为美德开始。[2]他认为有些美德是某些自然情感的“净化的”和“完整的”形式，比如仁慈就是最好的例子（Cohon 2006，259）。这些美德不仅仅是我们内在的情感，因为“同情”机制在我们扩展和实现这些美德的过程中发挥了作用，但是它们都源自冲动、欲望等最基本的人性（Taylor 1998）。[3]另外，人为美德如正义［或者更准确的说法是在财产方面的诚信（Cohon 2006，260）］的规范基础是我们发明出来的，因为它们对非个人合作的成功必不可少。这些规范的基础是我们自我利益算计的结果。不过，休谟强调，人为美德不仅仅是规范，还已经变成了实际美德。即使具体的诚实行为未必能给我们带来好处，“持久的心灵原则”也能可靠地为我们提供采取行动的动机（Hume 1978，575）。因为一旦确立了这样的规范，我们自然感觉到有必要遵循这些规范。这不是强烈的动机，尤其是在正义要求一种并不能给我们带来直接利益的行动的场合更是如此。我们天生具有的赞同这些规范的感受需要得到政治家、前辈和其他人的努力作为补充，鼓励我们塑造这种情感，让正义和其他美德充满活力发扬光大（Cohon 2006，263；Taylor 1998）。

简而言之，休谟（包括我自己）认为充分成熟的美德是“持久的”习性，能够以适当的方式对问题做出回应，这些美德通常是经过努力取得的，而非简单的、天生的习性。美德所做的回应不是被迫去遵循某个法则，而是伴随着典型情感的自发反应。通常，相关法则是一些经验法则，绝非具有任何深刻意义的东西。美德是否是儒家的核心观点，这一问题引起了广泛争议（相关论述及更多参考资料见Angle 2013）。如果想了解美德在宋明理学中发挥的作用，只要考虑一下朱熹和王阳明经常将“德”与同音字“得”联系起来就清楚了。比如，

朱熹说过“若今日孝，明日又不孝；今日忠，明日又不忠，是未有得于我，不可谓之德”（朱熹1997，778）。王阳明则说过这样的话，“天命于我谓之性，我得此性谓之德”（王阳明1992，1168）。“我得此性”意味着已经作为牢固不变的性情，而且真正付诸实践。我们继续来谈本文的核心内容：人工智能如何增强儒家的五大基本美德？

五大美德

通常，儒家所说的五大美德是仁、义、礼、智、信。我将考虑人工智能依据每一种美德能够支持行为的方式，或许更重要的是人工智能依据这些美德采取行动的充满活力的性情的培养。如前所述，美德不仅仅是在适当的时刻产生适当的行为，而且这些行为是以适当的方式产生出来的。我引用的儒家传统观点都认为美德在某种意义上和在某种程度上都是自然而然产生的，也都赞同典型人物的美德并没有强大到足以让其在很多时候都能可靠地做出适当的回应。传统儒家内部和当代学者都对其中的原因存在相当多的辩论：实现美德是需要不断成长和得到强化的过程，还是这个问题仅仅局限于清除“自私”的障碍从而实现“仁”呢？虽然这个区分对下文的讨论有时候或许有相关意义，但我更多的是认定人工智能在培养美德的性情方面发挥了积极的作用，无论人们依据的儒家美德模式的基础到底是什么，因为私的挑战适用于所有模式。

私的核心意义是过分强调自我利益而牺牲他人利益。对此，人们通常是将自我和他人区分开来讨论的。在其他时候，儒家使用添加自我的额外一部分或者将其附着在对背景的考虑之中。当人们认识到在特定场合下重要的东西是具有开放性或者包容性的时候，人们往往就不那么自私了。最近，有些英美学者已经认为对“私”的更好翻译是“self-centered”（以自我为中心）而不是“selfish”（自私自利），但我认为这种说法容易产生误导。“私”的问题不是某人在考

虑问题时把自己置于核心位置，事实上，将自己置于核心位置，几乎每个人都会同意更多地关注自己的亲人（家人）而不是遥远的陌生人或遥远的环境问题，这是合情合理的。[4]但是，如果人们过多关注与自己亲近的人而过少关注陌生人，就会出现问题。非常明显的是，自私的表现有很多形式及不同程度，其中很多（比如懒惰）就其本身而言似乎没有那么可怕，但是自私仍然极具破坏性。因此，对我们来说，关键的问题是人工智能以什么方式帮助我们战胜自私。

我们从儒家最著名的美德“仁”开始讨论。仁是一种适当关心他人的美德。[5]儒家传统中的有些声音呼吁人们应将焦点集中在关心他人上，但是我在本文中的观点依据更广泛一些，仁是对所有事物的适当关心，对各事物的关心方式不同，关心的程度也不同。人工智能会如何帮助我们变得更“仁”呢？

早期儒家的重要见解之一是，关心常常是由某个在我们看来显著的事物引发的。比如，在孟子有关“孺子溺井”的著名文章中，关心他人的恻隐之心就是因突然看见一个小孩可能要掉进井里而激发出来的（《孟子·公孙丑章句上》）。如果有人看见身处危险的小孩，恻隐之心就随之出现。在另一个重要的讨论中，孟子和齐宣王讨论了后者在看见有人牵着一头牛走过堂下并要杀掉牛用血祭钟而出现的不忍之心。齐宣王表达了不忍看见牛觳觫的样子，便下令放了这头牛，用一头羊代替。[6]在小孩和牛这两个故事中，孟子认为我们都没有看到充分发展起来的“仁”。从“孺子溺井”中到底应该得到什么启示是颇具争议的话题，但至少非常清楚的是，孟子相信我们必须找到办法确保我们对不怎么明显的目标拥有恻隐之心，我们对小孩的反应显示我们能够关心他人，但是这个目标太容易引人关注。[7]同样地，齐宣王对牛的反应显示了至少在适当的背景下，他甚至能够对动物表现出关心。但是，孟子指出齐宣王对更加重要的关心目标——他的臣民则关心太少了。如果齐宣王注意到臣民的困境，他的仁爱之心就会被更多地激发出来，这里的隐含意思是他花费太多时间在王宫里寻欢

作乐，却太少关注百姓的疾苦。孟子虽然没有这样表述，但他显然在批评齐宣王过于自私，他和臣民之间的纽带被割断了。

那么，人工智能会帮助我们更多地关注需要我们关心的目标，因而帮助我们变得更“仁”吗？是的，想象你的手机软件知道你的位置和日程安排，知道众多需要你关心的（人或环境）机会，知道你的专长和兴趣，通过将事物与这三个维度搭配，手机就能提醒你关注到一些并不显眼的事物，并以最佳的方式满足你的要求，实现你关心他人的愿望。而且，软件能够被设计得逐渐引导你关注新领域、应用技能的新方式或者增强你投入时间、精力和资源做某些事的决心。有关所有这些究竟是如何起作用的，可能还存在很多疑问，比如关心他人机会的数据从何而来？（或许软件有允许个人或群体上传种种机会的功能，从工作岗位到长期项目再到短期项目或一次性需要等任何机会）。软件是否会对这些机会进行过滤以增强其准确性？（用户或许可制作可靠性框架？请参看下文有关“信”的内容）软件会如何在众多机会中做出优先选择？（或许与我们最近的人尤其是家人的需要将得到最优先对待，虽然如果只关心家人的人也可能有些自私）。在多大程度上，它将包含帮助激发人们关心他人热情的数据或者多媒体内容？关心的程度有没有界限，是否超过这个界限之后，软件就应该适可而止了？换句话说，是否还有其他需要人们去关心的东西更值得你花费时间？有些决策似乎应该留给用户个人来做，比如是允许软件随时提醒用户适当的机会还是只允许它在用户自己登录软件期间提醒。当然，是否安装软件也应该由用户自己做出决定。

如果人们面对的挑战不够多，变得懒惰或者根本不关心他人，那会如何呢？现代社会经济模式和要求让人很难去做关心他人的事，即便人们很想去做。一个非常说明问题的例子是关心照顾父母，尤其是在现在的中国，成年人不大可能生活在父母身边，通常生活在远离父母的其他城市，并且多数人没有兄弟姊妹可以帮助你照看父母，这是很多人面临的巨大挑战，即使那些很有孝心的子女也是如此。那么，人工智能会帮忙吗？对此问题，经济学家丹尼尔·卡尼曼给出了一个

颇具挑衅性的答案。约翰·马尔科夫在一篇题为“人工智能将改善我们的老龄化困境”的报道中引用了卡尼曼的说法，因为社会快速老龄化，“在中国，如果机器人能被及时应用，那就太幸运了”（Markoff 2017）。马尔科夫提到了日本在开发老年人护理机器人方面做出的显著努力。[8]但是，他也引用了北卡罗来纳大学社会学家批评使用机器人照顾老人的观点，“如果我们把上年纪的老人或者小孩交由软件指控的金属品来照顾，我们当然应该感到悲哀”（Tufekci 2014，转引自Markoff 2017）。马尔科夫试图提供了一种更加平衡的视角：他不仅担忧机器人能否“以充满人性关怀的方式”来工作，而且意识到无人驾驶汽车、其他种类的机器人和物联网构成的增强版现实系统的潜力，它们可能给予老人“现在可能缺乏的高质量生活”（Markoff 2017）。

无论是语言上还是其更深刻的意义上，马尔科夫对“人性关怀”这个词的使用都相当接近儒家可能谈到的“仁”和“孝”。请考虑《论语》中的下面这两个段落。

> 子游问孝。子曰：“今之孝者，是谓能养。至于犬马皆能有养；不敬，何以别乎？”
>
> 子夏问孝。子曰：“色难。有事，弟子服其劳；有酒食，先生馔，曾是以为孝乎？”（《论语·为政》）

在和两个不同的学生的交谈中，孔子指出了同一个要点。的确，“孝”有很多具体行为，年轻人帮助老人干活，请老人先吃喝，但是这些还不足以称为孝（孔子认为在他那个时代，人们的这些行为是对古老圣人时代的拙劣模仿）。真正重要的是人们在做这些事情时的感觉，孔子期待的是人们对父母的态度——不是仅仅给父母提供吃的东西，而是在做这些事情的时候表现出尊敬。此外，孔子强调对父母年

纪越来越大而感到焦虑，当然，这的确可能导致人们的具体行为，但最重要的仍然是人们这些行为背后的态度。

如同上文谈及的美德一样，“仁”和“孝”从来都不是简单地做出正确的行动。在狭隘的意义上，仅仅确保父母拥有高质量的生活（比如吃得好）还不足以称为“仁”或“孝”，无论这一行为是由机器人还是自己来做。真正关心另一个人意味着渴望他们的生活体验充满活力和激情。依靠技术手段做到这些就都是坏事吗？这个判断总是很清晰、肯定吗？比如，你年迈、行动不便的父亲需要去一个地方，你背着他过去就是最好的选择吗？你开车送他过去或安排他坐上无人驾驶汽车前往是否也能表现出同样程度的关心？［我们假设在孟子的故事里，舜之所以背着父亲去海边是因为他找不到汽车（见《孟子·尽心章句上》）。］在关心他人的时候，我们使用机器人或者其他形式的人工智能来帮忙并不是坏事，只有当我们这样做时不是出于真心或者事实上根本不将此当回事时，才是有问题的。多亏了技术的发展，我们相互之间的纽带越来越紧密，即使我们身处不同的地方。我们能够用更具仁德的方式维系这种纽带，当然也可能以自私的方式利用这些联系达到自己的目的。

我们来看另一种美德——“义”。对“义”的常见和早期定义是将其等同于“宜”，不过，万百安指出，这个定义非常浅薄。万百安认为“义”更加具体的定义（尤其是孟子使用这个词的时候）是“一种遵从与行动者相关的禁令的习性，这种禁令涉及对自身伦理品格的表现和维持”（Van Norden 2007，258）。也就是说，有“义”这种美德的人不屑于做可能贬低自己人格之事，如果做了什么卑鄙之事，自己会感到羞耻。正如万百安所强调的那样，这里涉及的“伦理羞耻”是，当他意识到自己有某种性格缺陷时所感受到的东西，无论其他人是否意识到。常见的观点是，羞耻是外在的，取决于别人能否看见，而内疚是内在的。对此，万百安提出了不同的看法，他反而认为羞耻是有关自我的（尤其是自己的性格），内疚则是个人的行为对他人造成的影响（Van Norden 2007，259－260）。因此，我们看到培养

“义”的美德是核心目标，正如《论语》所说，“道之以政，齐之以刑，民免而无耻；道之以德，齐之以礼，有耻且格”（《论语·为政》）。受道德和礼仪指引而行事的人自然规范有序，因为他们内心已经形成了是非观念，而因为害怕惩罚不做坏事的人则形成一种欲望，他们想的是如何避免做坏事被抓住。在此，我们的问题是：人工智能会帮助我们培养“义”的美德吗？

应该注意到的一点是，“义”通常与一个社会中普遍被接受的原则有适当的关系，当然也未必总是如此。我们不妨回顾一下万百安的“义”是某种“禁令”的建议，他认为，人们之所以遵循某些法则，不是因为害怕被抓住（或许受到惩罚），而是因为人们认定如果成为违反这些法则的人，自己会感到羞耻。所以，人工智能可以为我们做的是，提醒我们注意到违反法则的潜在可能性，或许以一种特别显著的方式将行为置于框架内。比如，现在我们很多人在开车时都在手机上使用GPS（全球定位系统）为我们导航。GPS这类应用软件很清楚车速限制和当前车速，每当我们超速时，很多软件会提醒我们注意控制车速。关键是，软件并不能自行控制车速，也不会将超速信息发给警察。相反，它采取相对温和的方式，比如用红色而不是不容易引起注意的黑色来显示当前车速。毕竟有时候在具体场景中，人们开车时超过速度限制的确存在合理的理由。有了GPS的帮助，要想忽略违规行为就更加困难了。稍微想象一下，我们就可以毫不怀疑地设法将这种约束性影响扩展到开车之外的其他活动上。

事实上，我们已经看到这些温和的约束性影响被应用了，它们相互之间密切相关，而且与“礼”这一美德紧密相连。对于“礼”究竟是“义”在正式场合使用的附属类别还是拥有独立意义的另一种美德，这要强调人们在遵循法则时的行为方式或做事风格，关于这一点学术界存在一些分歧（见Van Norden 2007，270–272；Olberding 2007；Angle 2012，ch.6）。我相信我们能够在无须完全解决“礼”和“义”的关系问题的情况下思考人工智能发挥的作用，部分原因在于在真实生活中，人们做什么事（是否违反法则）与如何做事之间的

关系并不能简单地区分。比如，我们可以考虑一下电子邮件系统有能力提醒用户正在写的信息内容是否有冒犯他人之处。据我所知，这只需系统简单地核实一下邮件内容中是否存在某些词汇，但是，我们很容易想象这样的场景，即这个过程变得非常复杂和得到更广泛的应用。同理，现在已经出现了一些跑步者能够随身携带的产品，用以提醒他们跑步姿势不对或步伐太大等，或许在未来能出现类似产品以提醒人们关注其他更笼统的动作姿势，提醒人们在交谈时与对方的眼神接触是否适当等。正如GPS对开车超速的反应案例那样，电子邮件系统并不能阻止人们发送可能冒犯他人的邮件。

儒家的第四个美德是“智”。万百安相信在《孟子》中，我们看到智慧的四个不同维度：（1）适当评价别人性格的习性；（2）目标手段之间协商的技能；（3）理解和避免去做卑鄙之事的习性；（4）合理关心个人的幸福（Van Norden 2007，275）。

显然，所有这四个维度都基于在相似但有重要差别的事物之间做出区分和差异理解的能力，以及理解这些不同事物之间的联系特征的能力。在《孟子》看来，智慧的根源是“是非之心”，实际上，中国早期儒家认为在事物之间做出区分的能力与认识、理解和做出聪明决策的能力之间存在重要的纽带（见《孟子·公孙丑章句上》；Geaney 2007；Angle and Tiwald 2017，ch.6）。这些潜能在今天（或不远的将来）或许因某种特殊的人工智能形式即理论家所说的“情感人工智能”的帮助而增强，这种猜测似乎不无道理。

目前，情感人工智能还只是设想，因此我只是提出若干可能让我们变得更有智慧的方式。与理解自身对场景的反应相关的一个关键区分是找到影响人的相关因素和不相关因素的差异。阿兰·德波顿特别指出：

> 我们的意识根本无法很好地把握主人到底发生了什么。比如，我可能意识不到我之所以恼火是因为感觉疲劳而不是因为伙

伴的迟钝和笨拙。但是，情感人工智能将帮助我们获得所需要的自我认知。它将指引大脑的活动，及时提醒我们意识到心理生活的现实。（de Botton 2015）[9]

其实，害怕我们被不相关的生理因素引入歧途是儒家传统中的深刻忧虑：我们在《孟子》和《荀子》中都看到这一点，只不过二者讨论的方式不同，后来的儒家学者对此问题的讨论更为详尽，其关注话题多是“气”在影响我们的情感和行为方面所发挥的作用。应该承认的是，我们需要非常谨慎，避免在身体因素和其他因素之间做出明确区分，因为儒家明白人是具身的生物，承认我们的身体幸福是美好生活的重要组成部分。在《孟子》中，我们读到身体欲望是人性（规范性）的组成部分，虽然我们不应该过分强调人性的作用。在理学家看来，“理”总是与“气”密切相关（《孟子·尽心章句下》；Angle and Tiwald 2017，ch.2）。虽然如此，德波顿说人们发火咆哮常常不是因为愤怒而是因为饥饿或者疲劳导致人们误读了场景，这种见解还是很有见地的。如果情感人工智能提醒我们“及时”关注这样的场景以影响我们的行为，那么结果可能真能帮助人们发挥更大的智慧。

最后，我将继续探讨“信”这种美德。《论语》和《孟子》等文献中讨论过“信”，“信”通常被理解为说话算数，成为信得过的朋友，但是，后来“信”逐渐被认为是基本美德之一。在理学家看来，“信”意味着以持续不断的和可靠的方式立志接受现实的指导。这种美德帮助我们确保实现其他美德。我们是否设定仁、义、礼、智目标，我们能否取得成功都取决于我们能否看到事物本来的样子，是否立志按事物本来的样子看待它们，而不是陷入帮助我们摆脱困境的自私幻觉（见Angle and Tiwald 2017，162–163）。在本文的背景下，我相信“信”之所以特别重要，不仅因为它在宋明理学传统中被认定的重要作用，而且因为其在使用人工智能增进美德时的特殊意义。下文即将讨论，人工智能所发挥的很多积极作用皆取决于我们充满信心

地与他人分享数据和我们与他人共享成果的能力。如果说“信”是本文所说的所有美德的基础——元美德，这并非夸大其词。

或许我们可以从内森·海勒的文章《爱沙尼亚：数字共和国》中找到这个问题的切入点。多亏了雄心勃勃、历经多年的“电子爱沙尼亚”工程，这个波罗的海小国的公民能够：

> ……在家里通过笔记本电脑投票，对停车罚单提出异议。他们可以通过“仅此一次”政策做到这一点，该政策规定任何一条信息都不能输入两次。申请者无须“准备”贷款申请，而是将数据（收入、债务、储蓄）从系统中的其他地方提取出来。到了医生的候诊室，你也不需要填写表格，因为医生能够看到病人的医疗记录。爱沙尼亚的电子系统基于公民的身份证芯片，这张卡将通常很麻烦的申报过程（如报税）简化成一项能够被轻松搞定的事。（Heller 2017）

海勒的文章有一部分阐述了实现这些奇迹的技术：多种形式的安全措施、本地数据掌握的方式（“你的牙科医生拥有自己的数据，你就读的高中和你的开户银行也有自己的数据。当用户要求一则信息时，信息就像横跨运河的船只需通过水闸就可以交付给你”），以及个人用户只能接触适当数据的方式（如老师能登录浏览成绩信息，却不能看见或更改金融信息）。文章的另一面则谈及“电子爱沙尼亚”确立之后所产生的社会、伦理和政治后果。“企业经营和土地登记等信息被认为属于公共所有”，这导致“令人吃惊的”开放性，如果你想知道有钱有势者的企业利益，很快就能轻松搞定。因为可以在线投票，选民受到威胁的机会并不多。当海勒采访爱沙尼亚的“全国数字顾问”马坦·卡瓦茨时，卡瓦茨在对话即将结束时说：“我一直在对这些小事夸夸其谈吗？其实，这些事并不重要，重要的是人们的心态，是文化，是人际关系，正是这些让我们做到了这一切。”人们对这个系统的安全性和开放性充满信心，这确保了共同体的团结和积极

的人际关系。相反，卡瓦茨说，在某些国家，公民缺乏对自己数据的所有权，加上资本主义企业的强大势力，并且在数据使用中缺乏责任追究，这些导致了“社会团结和信任的系统性缺失”。他补充说建立信任不仅需要拥有好的政策：“政客必须以信任为自己的核心价值，因为在当今世界，在某种程度上，任何东西都是公共的。”（Heller 2017）

值得探讨的是，卡瓦茨强调的信任与我们在《论语》等早期儒家经典中找到的“信”的政治观点有什么样的联系。

当子贡询问治理国家的办法时，孔子说：使粮食充足，使军备充足，使百姓信任。子贡问：如果迫不得已要在粮食、军备和信用三项中去掉一项，应该先去掉哪一项？孔子说：去掉军备。子贡又问：如果迫不得已在粮食和信用中一定要去掉一项，应该去掉哪一项？孔子说：去掉粮食。自古人都免不了一死，但如果没有信任，是无法立足的。[10]

《论语》中所说的“信”和卡瓦茨所说的“信任”概念非常相似，两者都与理学家立志接受现实的指导观念有关。差别在于早期儒家和卡瓦茨想象的个人获得与维持信任的方法。在儒家看来，这个过程的关键是领袖表现出高尚的行为，他们行为的自发性和真诚“睟然见于面”[11]，我们不由得受到感动而信任他们。[12]在卡瓦茨看来，信任的基础是将上文描述的个人拥有数据和制度的开放性结合起来。阐明当代儒家对这种对比所隐含的政治、法律和经济议题的态度已经超出了本文的探讨范围，不过，请允许我简单提出三点评论：首先，不能将有关维持“信”和“信任”的过程这两个观点融合起来的观点说不通；其次，当代儒家愿意拥抱卡瓦茨风格的开放性是完全有可能的，因为儒家反对私（赞赏公）的方式；最后，正如上文注意到的那样，我相信当我们准备回答使用人工智能来增强美德可能遭遇的主要反对意见时，“爱沙尼亚式”的解决办法是很重要的。下面我们就来讨论一下这个话题。

反对意见和结论

最后一节首先考虑本文讨论的途径的若干反对意见，其次提出若干反思。第一个反对意见可能是：我在本文中讨论的相对简单的技术真的可以被视为人工智能吗？换句话说，我是否回避了哲学家本应该遭遇的人工智能的真正深刻问题？对此挑战，我的回答是我同意还有值得探索的其他重要问题——虽然这些问题也是猜测性的，有些哲学同行的确已经提出了儒家思考超级人工智能的途径。不过，无论超级人工智能是否真有可能出现，本文探讨的议题都是真有可能实现的。这种切实的可行性使这些问题在儒家哲学家看来非常重要，他们总是不仅关心宏大的理论问题，而且关心实际的、具体的和社会的变革。

第二个反对意见要深刻得多。就本文推崇的某些途径依靠大规模的数据挖掘和确保技术对具体场景做出回应（比如这可能被应用在我文中提及的情感人工智能概念）而言，我们如何确保人工智能不会复制现有的糟糕价值观？我们如何确保历史上的种族主义或其他偏见不会影响输入人工智能中的数据？（这种挑战还可参阅William 2017。）在这一点上，对此挑战的唯一回应是承认其严肃性，并确保在任何实际设计增强美德的人工智能中都要考虑到这个因素。

第三个反对意见是本文设想的技术可能带来如下担忧，如监控程度增加和隐私泄露等。如果你参与慈善活动或者超速行驶的行为被发现，甚至被政府或公司发现，并以此奖励或惩罚你时，会出现什么情况？在一定程度上，这种现象已经发生了：公司能够监控雇员的驾驶行为或者电子邮箱，个人也同意让保险公司监控自己的驾驶过程。[13]我提出的用来回答这种担忧的途径遵循了文中有关爱沙尼亚的讨论（至少是海勒文章中所描述的那样）。个人相信他们的数据（包括我们现在可以想象的与其道德表现相关的数据）是他们自己的，不仅是安全的，而且如果他们愿意的话，可以很容易地与特定的某人共享。安全分享这些数据的潜在可能性之所以具有相关意义，就是因为

正如我和田史丹曾讨论过的那样，自我修身常常不是孤立事件（Angle and Tiwald 2017，139－140）。即便不谈父母和老师所发挥的作用，我们竭力信守改善自我的承诺的方式通常都具有社会维度。正如从前朋友和伙伴之间能够相互鼓励一样，现在我们可以通过不同的社交媒体分享各自的故事和成就。社交媒体的问题当然很复杂，它带来的坏处可能比好处还多。虽然如此，那些让小组成员相互鼓励去进行体育锻炼的软件，它们的成功的确指向了某些积极的方向，尤其是当我们将其与已经在讨论的将个人数据安全掌握在自己手中结合起来时。

我要考虑的最后一个反对意见将我带回本文开头一节所讨论的若干议题，甚至是本文标题所传达的信息。简而言之，如果人工智能帮助人们采取某种行动，这还能说是有德之举吗？或者还能说是走在真美德之路上吗？回应该反对意见的第一步是注意到大部分或者几乎所有美德都需要花费时间和精力来培养。这不仅适用于（早先讨论过的）休谟的自然美德与人为美德，而且适用于儒家传统中对美德的所有不同描述（在某些思想家看来，少数人可能天生具有美德，虽然这种情况并不少）。第二步是思考外部辅助因素（不是我们身体组成部分的东西）的意义。在此，我集中论述人工智能发挥的作用。但是，事实上，外部辅助因素比现在的智能手机更加多样也更加古老。我们穿的衣服，我们住的房子和所处的公共空间，我们听的音乐，我们遭遇的有潜力影响我们做出反应的众多其他约束性框架无不如此。仅仅举从荀子那里选的一个例子：用带香味的花盖住尸体，我们就能够更好地哀悼（而不是厌恶地走开）。[14]所以，荀子认为有意识地控制我们遭遇的场景，就能对我们的自发反应产生积极影响，帮助人们更可靠地获得有美德的情感和行为。与此同时，荀子认为，我们可以变得更成熟，更少依赖外部情景的支持，而这种“转型”甚至能够达到这样一种程度，无论情景如何变化，我们都能圆满地应对（《荀子·劝学》）。

有些理论家已经利用哲学和心理学的知识来探索儒家、美德与场景之间的关系，我在上文段落中的观察与其笼统结论正好吻合。[15]虽然如此，我们需要询问人工智能中有没有什么特别的东西使其作为实现部分或整体美德的手段从而产生一些特殊的问题。或许因为人工智能提供的指导和助力比其他种类的外部辅助设施更加广泛，我们可能过分依赖它，从而不再开发人类自己的潜能。[16]对此担忧，我可以设想出两种不同的回答，一个比较谨慎，另一个比较激进。谨慎的回答是，我们在设计人工智能助手时就将这个担忧考虑进来。正如他们可能渐渐推动我们变得更加关心他人（如前文所述）那样，它们或许也会渐渐减少我们对外部辅助设施的依赖。比较激进的回答是，我们首先要知道荀子并没有设想由完美圣人组成的社会，这个社会把所有外部辅助事物——衣服、礼仪等都当作不必要的东西统统抛弃。毕竟，圣人非常罕见：从原则上说是可能的，但在实践中稀有得很，致使现实中是否存在圣人对于我们在设计社会构成方式时并没有相干意义。儒家认为人们能够和应该努力修身，更加接近圣人品德，但非圣人的社会制度总是有必要的。或许，我们应该对学习如何与人工智能相处感到舒服才是。

我最后的建议涉及了背景各异的哲学家已经开始探讨的一些引人入胜的哲学话题。如果我们超越智能手机，进一步考虑将技术移植到我们体内——因而变成“赛博格”，儒家会感到担忧吗？人与非人的二元论对儒家是否构成根本性挑战？我们能否设想某种类型的“后人类”未来？对这些问题的回答将部分取决于人们所属的思想传统。毕竟，传统儒家内部已经出现过人与非人的本性差别的复杂辩论。[17]我们不是要预先判断这些未来对话应该走向何方，请允许我特别强调这个事实，即儒家作为有生命力的思想和实践传统的健康，以及作为具有包容性和多样性的全球哲学的成长，皆取决于儒家提出新问题和进行新对话的能力。因此，请允许我再一次感谢博古睿研究院邀请我参与这些讨论，正是这些讨论促使我反思了上述问题。

儒家作为有生命力的思想和实践传统的健康，以及作为具有包容性和多样性的全球哲学的成长，皆取决于儒家提出新问题和进行新对话的能力。

安靖如（Stephen C. Angle），维思大学全球学中心主任，东亚研究弗里曼讲席教授、哲学教授，2016—2017年博古睿学者。1987年于耶鲁大学获东亚研究学士学位，1994年于密歇根大学哲学系获博士学位（师从孟旦教授）。自1994年起在维思大学教书，曾任哲学系主任、东亚学院院长。研究方向有中国哲学、儒学、宋明理学、政治哲学、伦理学与比较哲学。

参考文献

1. Alain de Botton, Six areas that artificial emotional intelligence will revolutionise,Wired, 2015, https://www.wired.co.uk/article/wired-world-2016-alain-de-bottonartificial-emotional-intelligence.

2. Amy Olberding, The Educative Function of Personal Style in the “Analects”,Philosophy East and West 57(3), 2007.

3. Bärí A. Williams, Intelligent’ Policing and My Innocent Children. New York Times (December 2, 2017), https://www.nytimes.com/2017/12/02/opinion/sunday/intelligent-policing-and-my-innocent-children.html.

4. Bryan W. Van Norden, Virtue Ethics and Consequentialism in Early Chinese Philosophy, Cambridge University Press, 2007.

5. Danielle Muoio, Japan is running out of people to take care of the elderly, so it’ s making robots instead,

Business Insider (November 20, 2015), https://www.businessinsider.com/japan-developing-carebots-for-elderly-care-2015-11.

6. David Hume, An Enquiry Concerning the Principles of Morals, Alasdair MacIntyre(ed.), Hume’s Ethical Writings, University of Notre Dame, 1965.

7. David Hume, Of the Original Contract, Alasdair MacIntyre(ed.), Hume’s Ethical Writings, University of Notre Dame, 1965.

8. David Hume, P. H. Nidditch(ed.), A Treatise of Human Nature, Oxford University Press, 1978.

9. David Wiggins, Natural and Artificial Virtues: A Vindication of Hume’ s Scheme,Roger Crisp(ed.), How Should One Live? Essays on the Virtues, Oxford University Press, 1996.

10. Deborah S. Mower, Situationism and Confucian Virtue Ethics, Ethical Theory and Moral Practice 16(1), 2013.

11. Edward Slingerland, The Situationist Critique and Early Confucian Virtue Ethics,Ethics 121(2), 2011.

12. Edward Slingerland , Trying Not to Try: The Ancient Art of Effortlessness and the Surprising Power of Spontaneit, Canongate Books, 2014.

13. Gabrielle Parke, The Implications of an AI Revolution: Revisiting Self-Cultivation,Wesleyan University Senior Essay, 2017.

14. Jacqueline Taylor, Justice and the Foundations of Social Morality in Hume’ s Treatise, Hume Studies 24(1), 1998.

15. Jane Geaney, On the Epistemology of the Senses in Early Chinese Thought,University of Hawai’ i Press, 2002.

16. JeeLoo Liu, Neo-Confucianism: Metaphysics, Mind, and Morality, Wile-Blackwell,2017.

17. Jennifer Robertson , Robo Sapiens Japanicus: Robots, Gender, Family, and the Japanese Nation, University of California Press, 2018.

18. John Markoff, Artificial intelligence could improve how we age, Washington Post (October 18, 2017), https://www.washingtonpost.com/news/theworldpost/wp/2017/10/18/artificial-intelligence-could-improve-how-weage/?noredirect=on&utm_ term=.3c465f32fb9f.

19. Nancy E. Snow, Virtue as Social Intelligence: An Empirically Grounded Theory,Routledge, 2010.

20. Nathan Heller, Estonia, the Digital Republic, The New Yorker (Dec. 18 & 25, 2017).

21. Nils J. Nilsson , The Quest for Artifical Intelligence: A History of Ideas and Achievements, Cambridge University Press, 2010.

22. Rachel Cohon, Hume’ s Artificial and Natural Virtues, Saul Traiger(ed.), The Blackwell Guide to Hume’s Treatise, Blackwell, 2006.

23. Richard T. Kim , Human Nature and Animal Nature: The Horak Debate and Its Philosophical Significance, International Philosophical Quarterly 55(4), 2015.

24. Shannon Vallor, Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting, Oxford University Press, 2016.

25. Steph C. Angle, Contemporary Confucian Political Philosophy: Toward Progressive Confucianism, Polity Press, 2012.

26. Stephen C. Angle, The Analects and Moral Theory, Amy Olberding(ed.), The Dao Companion to the Analects, Springer, 2014.

27. Stephen C. Angle, Justin Tiwald, Neo-Confucianism: A Philosophical Introduction,Polity Press, 2017.

28. Zeynep Tufekci, Failing the Third Machine Age: When Robots Come for Grandma,The Message (July 22, 2014), https://medium.com/message/failing-the-thirdmachine-age-1883e647ba74.

29. 王阳明：《王阳明全集》，上海古籍出版社，1992年。

30. 朱熹：《朱子语类》，岳麓书社，1997年。

---

[1] 我非常感谢加比·帕克与我多次对话，本文讨论的很多问题与她的作品主题（Parke 2017）密切相关。我也感谢2018年中国思想东北研讨会和2019年博古睿研究院中国中心“人工智能遇见中国哲学家”工作坊的参会者对本文初稿提出的富有启发性的建议。

[2] 这个描述出现在《人性论》（Hume 1978）中，因为没有明确出现在休谟后来的著作《道德原则研究》（Hume 1965a）中，有些人认为他在后来放弃了这个观点。但是，正如戴维·威金斯指出的那样，这种区分存在于《道德原则研究》的背景中（Wiggins 1998，135），甚至在更后期的文章《论原始契约》（Hume 1965b，267）中也有明确表示。

[3] 杰奎琳·泰勒强调了自然美德中发生的区分改变，因而她认定其他学者有时候过于强调自然美德和人为美德的区别了。

[4] 有关“自私自利”和“以自我为中心”的有益讨论，请参阅Liu 2017，而有关核心位置的重要性的更笼统讨论，请参阅拙著Angle and Tiwald 2017。

[5] 儒家说“仁”具有具体的含义和比较笼统的含义，即“仁”等同于美德本身：在这个意义上，仁者就是拥有各种美德的人。这是很常见的。在本文中，我将焦点集中在主要文献定义的具体含义上。更多讨论请参阅拙著Angle and Tiwald 2017，ch.8。

[6] 原文为：有牵牛而过堂下者，王见之，曰：“牛何之？”对曰：“将以衅钟。”王曰：“舍之！吾不忍其觳觫，若无罪而就死地。”对曰：“然则废衅钟与？”曰：“何可废也？以羊易之！”（《孟子·梁惠王章句上》）——译者注

[7] 请参阅：人皆有所不忍，达之于其所忍，仁也；人皆有所不为，达之于其所为，义也。（《孟子·尽心章句下》）换句话说，虽然我们谁也不忍心看见小孩掉进井里，在这种情况下都会产生恻隐之心，但有太多的事情是我们能够忍心的，比如遥远的陌生人的痛苦。真正的“仁”就在于我们将不忍心看见别人痛苦的反应扩展到更遥远的事情上。

[8] 详见Muoio 2015。也许令人惊讶的是，在日本关于机器人持久治疗的理论和实践中，机器人在老年人护理方面发挥作用的潜力并没有得到特别关注（Robertson 2018）。

[9] 在本文中，我只是简要提及了德波顿简洁而富有启发性的报告中若干引人入胜的想法。

[10] 原文为：子贡问政。子曰：“足食，足兵，民信之矣。”子贡曰：“必不得已而去，于斯三者何先？”曰：“去兵。”子贡曰：“必不得已而去，于斯二者何先？”曰：“去食。自古皆有死，民无信不立。”（《论语·颜渊》）——译者注

[11] 原文为：君子所性，仁义礼智根于心，其生色也，睟然见于面，盎于背，施于四体，四体不言而喻。（《孟子·尽心章句上》）——译者注

[12] 森舸澜用认知科学的语言引人入胜地描述了这个过程。他说人类的身体语言处于“不停流动的状态”，人类的“认知控制系统是被自下而上管理着的”，是独特的且具有标志性的（Slingerland 2014）。

[13] 见https://www.allstate.com/drive-wise.aspx。当今中国也一直在讨论如何建立监督私人和公共机构的“社会信用”的种种体系。

[14] 原文为：故死之为道也，不饰则恶，恶则不哀；尔则翫，翫则厌，厌则忘，忘则不敬。（《荀子·礼论》）——译者注

[15] 比如，请参阅Slingerland 2011和Mower 2013。斯诺（2010）没有讨论儒家，但他的观点非常吻合我的视角。

[16] 加比·帕克提醒我关注了这个议题，请参阅Parke 2017。

[17] 这里我想到了有关人物性异同的“湖洛辩论”，相关讨论请参阅Kim 2015。

机器智能的兴起与人之为人理想的终极

李晨阳 /文 吴万伟 /译

正如英文单词“end”拥有双重含义一样，本文的标题也拥有双重含义。“end”意味着某个持续存在的事物或时间段的最后部分，指其终结，也意味着某个努力或行动的目的和目标。基于这双重意义，在本文中，我认为，首先，人工智能的出现和快速进步意味着人之为人理想在某个重要意义上的终结，伴随着这种终结，我们已经无可挽回地丧失了从前宣称的人类所拥有的特殊地位；其次，我们应该开发人工智能技术来服务于人类，为此，在合适的时候，我们应该让高级人工智能尽可能地具有道德感。高级人工智能不仅仅是“机器”，在更重要的意义上也是“人”。

本文由4个部分组成，正好呼应博古睿研究院中国中心有关人工智能研究项目提出的4个议题。第一部分提出，在某个重要的意义上，人工智能挑战意味着人类已经丧失了其根本独特性的最后根基。第二部分考察了这种既体现进步又显示异化的变化的隐含意义。第三部分阐述如何让人工智能具有道德感。最后一部分探讨人工智能技术可能会在哪些重要方面影响中国哲学。

根据传说，位于古希腊德尔斐圣地的阿波罗神庙展示的格言是“认识你自己”。人们认为该格言，意味着我们人类必须努力认清我们到底是什么。苏格拉底遵循这句格言，终其一生不懈探索，想要发现他是谁，他能做什么。这样的努力也是在探索人性。显而易见的是，我们人类要知道我们是什么，我们能做什么，这个话题具有普遍性。努力认识自己背后的假设当然是对人类特殊性的一种信念，我们坚信人在本质上不同于宇宙中的其他所有生物。在思考人类的独特性

时，这里的“本质”的要求非常重要。正如黑格尔的名言所说，即使人是唯一有耳垂的存在，在这方面有其独特性，但拥有耳垂不是人的本质特征。而且，我们渴望能找到的不仅是人的独特性，而且是给人以崇高地位的独特性，使我们不仅独特而且在宇宙中特别重要。沿着这条线路，苏格拉底提出了人类灵魂的哲学。人类拥有一个独特的灵魂，这使得人类不同于而且高于其他一切存在。但是，如果我们回顾人类历史，这种寻求人类高超的独特性的愿望似乎永远无法实现。我们曾经认为我们之所以特别是因为我们居住的地球是宇宙的中心，这意味着我们人类也处于宇宙的核心位置。但是，这个中心不存在了。哥白尼革命无情地将我们从那个想象的中心位置驱逐了出去，哥白尼发现，地球不是宇宙的核心，它只不过是围绕太阳转动的一个星球而已。他的理论遭到宗教信徒的强烈反对，不是因为科学见识的分歧，而是因为哥白尼的理论被认为是对人类地位和特殊身份认同的威胁。现在我们必须承认，我们从来就没有处于宇宙的核心位置。在这方面我们人类并不独特。我们曾经认为我们之所以特别是因为人是依据上帝的形象创造出来的。达尔文的进化论又无情地剥夺了这种令我们曾经感到欣慰的想法。在达尔文看来，人类就像其他物种一样也是从低等生物演化而来的，甚至人的外貌（人的形象）也是进化的结果，而非更高权力者的特殊设计。对有些人来说，至少是从心理上来说，最致命的一击是尼采宣称“上帝死了”。这个口号意味着神圣性不再是人类确立其独特地位的可靠选择，我们必须自己为人生创造意义。长久以来，拥有理性似乎是我们仅剩下的能够与其他存在区分的东西了。但是，就连这个特征也受到挑战。如果我们相信弗洛伊德的心理分析理论，即人的心灵有三部分构成：意识、前意识和潜意识。意识作为理性处于意识之中，而意识之所在，不过是整个人的心理建构这座庞大冰山的小小一角而已。人类心灵的活动在很大程度上受到前意识和潜意识的影响。换句话说，理性在人类心灵的活动中并没有扮演重要角色。因此，按照弗洛伊德的说法，人类很难被定义为理性动物。我们可以接受或不接受弗洛伊德的理论，但是他的挑战至少说

明，人类是理性动物的信念同样需要经过证实才能确立，而不是既成事实。有人认为，自我意识是人类的普遍特征，但是最近，哥伦比亚大学的科学家发现，人工智能机器也能拥有自我意识。这对长期以来拥有的观念（即自我意识是人类的垄断性特征）提出了直接的挑战。[1]

那么，这一连串事件将把我们带向何方？我们仍然在寻找“认识你自己”这个永恒问题的答案。弗洛伊德的理论已经被许多人抛弃，但他们试图牢牢抓住理性这根最后的救命稻草不放。理性是一种智能。理性思考和理性行动是智慧的功能。[2]我们大部分人坚持或者希望坚持认为，至少我们人类比其他生物更有智慧。的确，我们对“认识你自己”这个问题的答案的追寻可以被视为人类使用智能在创造人类独特性的神话。至少在这个方面，人类比其他物种更优越。换句话说，我们一直很独特地制造人类独特性的神话。但是现在，这个时刻终于来临：人类智能的独特性受到人工智能的挑战。人工智能技术的进步将展示，我们人类既不拥有高超智能这个独特性，也不是世界上最有智能的存在。有些高级人工智能机器已经在智能上超越了人类。谷歌的全新人工智能程序AlphaZero能战胜最优秀的人类象棋大师。有人认为我们正在接近奇点，即人工智能在智慧上超越人类的关键点。[3]

人工智能在智慧方面的挑战不仅仅是智能水平的问题。不同于和人类相比的所有其他自然物种，人工智能不是自然的产物，不同于自然界的演化过程，它能够迅速调整以完美匹配人类潜能。孟子认为，人不同于动物之处就是我们有恻隐之心，忍受不了他人受苦。现在，经过我们编程的高级人工智能机器可以在“感受”到别人的痛苦时停止手头的工作，甚至采取行动减少他人的痛苦，如照顾病人的医用机器人所表现出来的那样。荀子认为，人不同于其他动物是因为人能“群”，即有社会性，但是这已经不再是人类独特的品质。军用作战机器人当然能够在行动中进行有组织的合作，即能“群”。事实上，

要想在战斗中取得胜利，社会合作是必需的。人工智能技术的崛起显著缩小了人与非人世界的差距。人工智能强大的可塑性已经让宣称人类本质的独特性的任何尝试变得即使并非彻底不可能，至少也越来越缺乏合理性了。与自然界中和人相比的其他生物不同，人工智能可以随时“根据要求定制而成”，不需要经过自然界的漫长演化。人性中任何被视为特殊或者独特之处都可以靠人工智能技术复制出来。

随着人工智能的出现和快速发展，对“认识你自己”这个问题的令人满意的答案仍然像从前一样难以捉摸。这样的结果迫使我们重新思考这个问题本身。或许，“认识你自己”不是或不应该是寻找人的本质的独特性。或许这一命题指的是发现人类并非那么独特，然后学会与这一发现所带来的结果和平共处。相较之下，我们发现道家学者或许比其他很多哲学家更愿意接受这样的结论。《道德经》记录了古代圣人老子的深刻见解“知人者智，自知者明”。认识他人和认识自己都是认识人性。不过，他似乎没有强调人的独特性。人和世界上的其他生物之间存在差别，但没有高低贵贱之分。

人工智能强大的可塑性已经让宣称人类本质的独特性的任何尝试变得即使并非彻底不可能，至少也越来越缺乏合理性了。

我们的未来在于融合高级人工智能技术而不是维持我们的独特性。这种融合是完全可能的，而且潜能非常之大。未来学家雷·库兹韦尔说：“我们将实打实地与这种技术即人工智能结合起来，使我们变得更聪明。事实上有人已经在这么做了。这些设备是我们大脑的延伸，人们真的在这样思考了，这是全新的事物。”[4]通过纳米技术，我们将能够把人工智能设备与大脑神经系统联系起来，让人工智能强化的大脑运行更加迅速、更加强劲。哲学家已经在设想“心灵延展”的有效性了。[5]人工智能技术现在已经提供了大脑延展的可行性和现实性的科学与技术证据。人将变成混合体，以生物物种提供的东西作为基础，与从人工智能技术中借用的东西结合起来。这样的“人形机”不仅表明了类似人的非人的存在，而且证明了在具体的个

体身上可能存在人和非人部件的混合体。中文中用来指代先进人工智能系统的词语“机器人”现用来指人工智能设备，其实这个词用来表述这种新生的“人形机”更适合。普通的人工智能设备除非与人融合起来，否则都不能称为“人”，即使是在“准人”意义上的人。

与此同时，人工智能技术的进步伴随着生物工程技术的进步。最近，康奈尔大学的科学家成功实现了动态生物材料自下而上的构建，它是依靠人工新陈代谢驱动的，这种代谢代表不可逆的生物综合和耗散组装过程。他们通过编程让这种材料产生类似于真核微生物黏菌的自发移动行为。动态生物材料拥有像自动化模式产生和持续极化再生那样的性能。据报道：“人工新陈代谢驱动下的动态生物材料能够提供从前没有探索过的路径以实现具有再生和自我维持特征的‘人工’生物系统。”[6]新陈代谢和生物综合是生命的特征。康奈尔大学农业和生命科学学院生物学和环境工程教授罗丹说：“我们在引进一种崭新的、类生命的材料概念，它是由自身的人工新陈代谢系统驱动的。（虽然）我们不是在生产某种有生命的东西，而是在创造一种从未见过的更接近生命的材料。”[7]康奈尔大学研究团队的成员浜田庄木说：“我们处于建造人工新陈代谢驱动下的有生命的机器的第一步。”[8]当这种新技术与人工智能技术结合起来时，可能从完全相反的方向满足人类“大脑延展”的目标。我们或许会看见人造有机体人工智能机器在向我们走来，就像看见我们自己在镜中一样。人和人工智能机器的界限将变得更加模糊。

如果上文的论证站得住脚，那么我们将面临一个实实在在的可能性，那就是不仅是寻找古老问题“认识你自己”的答案的终结，而且是我们所知的人性的终结。“人”总是有双重意义，一方面，它代表了生物物种智人；另一方面，它是包含价值观的观念，代表了一种理想。即使人的首要存在目的总是作为生物物种的生存，但是人从来不会满足于被描述为某种“有宽平指甲的、无羽毛的两足动物”，虽然人或许是世界上仅有的这种动物，但我们的渴望总是比这更多。“认

识你自己”的呼吁是在敦促我们追求包含了理想价值观的人性。人性的终结并不意味着人类作为生物物种将不复存在，而是说人将永远丧失其独特性和独一无二的崇高地位，即人的本质。用来表示本质的希腊词汇的字面意思是某物“它所是的那个东西”（或者亚里士多德的更简短说法“它是什么”）。或许人根本就没有这样的本质。那种人性，即拥有特殊本质的人性已经终结。毫无疑问，人类将继续寻找“认识你自己”的答案，但是，我们或许不得不将人的本质视为移动的目标而非在某处等待我们去发现的东西。对人类而言，即将到来的人工智能时代可以被描述为纯真的丧失或者自我关注的成熟性的丧失。无论是哪一个，自认为独特的人性的传统意义已经终结。

人性的终结并不意味着人类作为生物物种将不复存在，而是说人将永远丧失其独特性和独一无二的崇高地位，即人的本质。

通过模糊人和非人存在的界限，人工智能技术允许我们不仅在空间上而且在存在的意义上将人的存在扩展到新领域。也就是说，我们已经变得既不仅仅是自然物种演化的结果，也不仅仅是自然赋予的东西。在一定的意义上，我们现在就是我们自己所造就之存在。科学家在讨论人是否已经超越全新世进入人类世，即一个人能够在很大程度上改造自然的新时期。正如保罗·克鲁岑和克里斯蒂安·施瓦格尔所说：“现在的情况已经不再是我们与自然的对立。相反，是我们决定自然现在是什么，将来会是什么了。”[9]考虑到人工智能的挑战，我们或许可以补充一句，我们决定的不仅是自然，而且还包括我们自己：在很大程度上，我们将决定我们现在是什么，将来要变成什么。在人类世即将到来的背景下，人类物种的自然状态的巨大飞跃可能既是进步又是异化，既是祝福又是诅咒。说它是进步在于，我们将大幅提高能力和威力，能够做到和实现从前只有在神话中或者梦境中才存在的理想。变化只要让人变得更加适应生存环境，就可以被视为进步。说它是异化在于，我们将永远离开自然的根基。异化的概念本身就建立在我们人类是或应该是的某种“本质”之上。如前所述，人类

本质在很大程度上是一种虚幻的观念，某种我们觉得应该拥有的东西，想到它感觉很安心的东西，但是我们从来没有发现这种东西，虽然一直以来我们都在不停寻找。无论如何，我们可以说我们也在经历异化，在背离我们的现状、我们的历史和我们的理想。我们不再是已经习惯了的自我形象。我们现在大不一样了。正如当今习惯于使用智能手机的一代人无法想象自己没有智能手机的生活一样，一旦人变成了与机器智能的混合体，就再也无法回到原来的纯自然状态了。在这个意义上，这种异化是不可逆的，是永恒的。

人工智能技术将同时给我们带来进步和异化，这听起来自相矛盾。但是，它们发生在同一过程之中。事实上，任何进步都伴随着某种丧失。从前那些描述陷入相思之苦的女孩或男孩的古老故事令读者感动，他们深受渴望爱情的折磨而痛苦。如今，现代通信技术已经终结了这种状况。小说家必须找到其他方法来讴歌人类的这种情感。我们现在乘坐飞机只需一个晚上就从纽约飞到了上海，这使我们永远丧失了体验与同行的乘客在海上长达一个月的充满诗意的旅行的机会。《庄子》记录了一个有趣的故事。孔子的弟子子贡曾经和灌园老人对话。这位老人在菜园子里，挖地道到井中，抱着瓮取水浇园子，用力多而见效小。子贡告诉他，有一种机械，一天灌溉一百亩田，用力很少但见效大。子贡问“为什么不使用这样的机器呢？”灌园老人面有怒色，却笑着说：

> 吾闻之吾师，有机械者必有机事，有机事者必有机心。机心存于胸中，则纯白不备；纯白不备，则神生不定；神生不定者，道之所不载也。吾非不知，羞而不为也。（《庄子·天地》）

灌园老人代表了一种紧紧抓住传统不放的人，他们坚守自己的存在方式，抗拒任何偏离纯朴天性的东西。他认为，接受新技术就要求人们放弃原初状态，这是有道理的。异化的确存在，鱼与熊掌不可兼得。当最初的纯朴天性被认为是做人或者拥有美好生活必不可少的东

西时，接受新技术的风险就太大了，可能得不偿失。但是，只有当我们接受最初的纯朴天性是我们做人必不可少的东西时，灌园老人的哲学才是合理的。如果没有这样的前提，我们就没有多少理由不接受新变化了。在最终认识到我们已经再也不能回到我们从前的那个样子后，在认识到我们人类从来就没有而且将来也不会有不变的本质后，人工智能进入我们的生活而带来的这种异化就不再应该遭到强烈反对了，反而应该得到称赞才对。

人工智能技术的出现和快速发展给伦理学带来了空前的挑战。我们能否成功应对这些挑战将对我们的生活方式和人类未来产生深刻的影响。其中一个重要的问题是我们是否应该将高级人工智能机器当作伦理存在来对待，以及如何对待它们。

作为一项传统事业，伦理学一直是在据称根基稳固的世界的存在秩序的基础上发展起来的。大体说来，世界分成若干范畴，即人、生物有机体、其他物体。这些分类的形成要么通过演化，要么通过神灵创造，要么依靠其他途径。伦理学的每个概念，尤其是在西方的主流传统中，都是把人当作唯一的道德行为者和道德行为的施受者，把世界上的所有其他生物都排斥在道德领域之外。随着环保意识的崛起，当今有些伦理学家开始将道德领域扩展，希望把生物有机体包括进来。在各种有机体的确切道德立场问题上，人们有意见分歧。有人坚持认为，从道德上说，只有人才是重要的。有人则认为，道德领域应该包括其他“高级的”生命形式（如马和狗）。不过，仍然有人认为，所有生命形式在道德上都具有相关性，因为生命本身拥有内在的价值。尽管有这些分歧，但人们非常清楚自己应对的是什么范畴，即动物是动物，植物是植物。17世纪法国哲学家笛卡儿的做法是将动物贬低为机器，因为它们没有灵魂——他坚定不移地把守人与物的界限，认为人是会思考的存在，而其他一切都不过是在空间上延伸的物体而已。他的名言“我思故我在”中的“我思”代表智慧，这是人的专有品质。但是现在，人工智能技术带来了另一项巨大挑战，直接指向人的独特性。它带给我们一个完全不同的东西。人工智能是机器，

却是一种拥有智慧的特别存在。人工智能的出现永远地动摇了传统伦理学赖以生存的世界范畴划分和认识。它迫使我们以全新的眼光重新看待伦理学。

人工智能的出现永远地动摇了传统伦理学赖以生存的世界范畴划分和认识。它迫使我们以全新的眼光重新看待伦理学。

遇到机器人伦理学的议题，我们首先必须考虑人工智能存在的智慧特征是否使其有资格被当作道德行为的施受者来严肃对待。道德行为的施受者是指拥有道德资格的存在体，道德行为者（如正常人）对其拥有道德责任。人是道德行为的施受者，我们应该对人表示尊重，但是自行车不是道德行为的施受者。虽然出于某种考虑，我可能精心保管自行车，比如出于对你的道德义务，我妥善保管你的自行车，但是对自行车本身而言，我并没有道德义务。有些环境保护伦理学家已经要求我们扩大范围，将动物和生物有机体当作道德行为的施受者看待，因为动物也能感受到痛苦，也有尊严或者生命本身拥有的内在价值。其他人则坚称，只有理性的存在才有道德资格。按照康德的说法，只有人才有道德资格，这不是因为我们的生物学物种的身份，而是因为我们有理性，理性是道德资格的基础。因为理性和智能的契合性，[10]有些伦理学家已经把非人的智慧物种如狗、海豚、大象都当作道德行为的施受者的候选对象。现在，人工智能也加入了这个行列。因为人工智能有智慧，我们也应该相应地对待它们。就道德问题而言，这意味着人工智能机器应该被当作道德行为的施受者。《华盛顿邮报》的约尔·加罗在2007年的一篇报道中称，美国陆军的一位军官在看到机器人扫雷时一次又一次地被炸掉了腿，却持续不停地往前爬行后，取消了机器人扫雷实验。军官宣称，这个实验太不人道。[11]该场景类似于军人关心他们的随军警犬。有了智能让高级人工智能机器与我们更亲近，引发我们的同情。我们受到感动，倾向于将其看作道德行为的施受者。但是，我们到底应该如何将社会中的人工智能当作道德对象，这一问题的答案仍然很不明朗。毕竟，人工智能是

机器，不是生物，不像狗一样有生命。它也不能像生物有机体那样有感受痛苦的潜能。而且，与生物物种不同，人工智能的智能不是统一的，智慧层次千差万别，既有智能程度很低的自动化机器人，也有非常先进的仿人机器人腱悟郎，至于未来的人工智能机器，可能更加超乎我们的想象。即使我们同意高级人工智能机器能够和应该被当作道德行为的施受者，但有关究竟该如何对待它们这一问题的答案仍不明朗。比如，世界各地医院的人工智能医生数量越来越多，对它们的工作条件应该采用什么样的管理规定？是不是应该像我们对待真人医生那样，对其工作条件设置最低标准？随着人工智能技术将情感以及智能置入人工智能机器，使其变得越来越像人类，这种需要已经越来越迫切。

而且，最近人工智能伦理学的讨论主要集中在人工智能机器是不是道德行为者的问题上。道德行为者是能够辨明是非并据此做出选择的主体。从传统意义上来说，伦理学的道德行为者仅限于人类本身。我们认定人有道德责任，动物却没有。即便很多环保伦理学家接受动物充当道德行为的施受者，但是把动物当作道德行为者的就很少了，其理由无非动物不能辨明道德上的是非。人工智能机器有道德主体性吗？高级人工智能机器有能力做出不仅具有经济后果而且具有道德后果的决定。美国军方已经开始开发人工智能硬件和软件，以确保作战机器人能够遵循战场上的道德伦理标准。[12]此类有关人工智能潜能的开发不仅仅局限于编写机器人在预料之中的场景的有限范围内的可能行动方案程序，也将使人工智能成为开放性的系统，收集和综合信息，预测和评估现有选择的后果，并部署适当的反应。这样的潜能将使高级人工智能系统更加接近独立自主的人而不仅仅是传统意义上的机器。麻省理工学院媒体实验室的“道德机器”实验已经提出了拥有该潜能的自动驾驶汽车的道德两难困境：汽车是否应该扭转方向以便避免撞上小孩，但这样做的后果是可能撞上老太太？[13]应该设计这样的自动驾驶汽车吗？应该让它这么做吗？人们可能认为，在这种情况下，仍然是人在替人工智能机器做决定，毕竟是人在为机器编程。

但是，人类自己不也是从小通过接受社会为我们编写的道德程序，通过学习等途径，学会尽量避免伤害人，哪怕是以牺牲猫为代价？就接受道德程序而言，高级人工智能机器与人的差别主要是接受方式的不同，而不是性质的不同。

人工智能决策制定能力的快速进步使这种复杂的机器越来越接近成为道德行为者。这就要求在设计先进人工智能系统时将伦理学融合进来。人工智能技术的这个维度之所以对道德哲学构成最大挑战，其部分原因在于，伦理学家在什么样的规范伦理学最好或者最合适这一问题上不能达成共识。道德义务论者认为，伦理学必须建立在道德义务的基础上。功利主义者认为，行动应该带来最值得向往的后果。德性伦理学者将会希望让人工智能机器的行为表现出优秀的美德。人们可能认为，我们应该寻找最小公分母，使用“不害人”原则作为先进人工智能机器的伦理要求。但是，这或许并不那么容易，因为很多人工智能系统的设计恰恰就是要伤害人的。现在，几个军事大国之间已经展开一场不宣而战的军备竞赛，各自都在开发能在战场中取代真人士兵的自主机器人。这些旨在杀人的人工智能机器的进步无疑将让世界变得更加凶险，也更容易导致武装冲突。那么，是否应该出台禁止任何人设计和生产用于战斗的人工智能机器的国际禁令？

要回答这些问题并不容易，不过，如果我们希望维持确保人类繁荣发展的世界和平，严肃对待和适当处理所有这些挑战就都是必需的。因为人工智能技术的快速发展及其对世界的巨大影响，应对这些挑战的任务变得越来越紧迫。与采取渐进式自然演变道路的进化过程不同，各种快速发展的先进人工智能系统将占据世界的各个角落，并在事实上改变人们生活的方方面面。其实，电话和电脑现在已经拥有了内置的人工智能技术，这已经在很多根本方式上影响了人类。人工智能技术可能决定人类的命运。在这个意义上，成功应对人工智能的相关伦理挑战甚至比应对人工智能的相关技术挑战更为艰巨。理解这些挑战的紧迫性至关重要。

那么，我们应该向人工智能机器灌输什么样的价值观呢？在人工智能高度发达的时代，人类就生活在仿人机器人中间，它们在很多方面就像我们的同事。仿人机器人在完成工作任务时需要自己做出决定。作为机器创造者，我们人类应该确保其决策制定受到某些伦理学和道德原则的指导，就像我们希望在其他人类同胞身上看到的情形那样。什么样的机器人伦理学对我们的道德情感是切实可行和合适的呢？为了回答这些问题，我们需要首先明确机器人伦理学的本质。在这方面，儒家道德哲学能够为我们提供有用的资源。儒家对伦理学的理解包括两种进路，分别以孟子和荀子为代表。孟子认为人是独特的，因为我们生来有不忍之心。培养仁德之心将让我们真正具有人性，引领我们走向有道德的人类生活。我们可以把对人类道德本质的这种理解称为“软心肠途径”。在这种观点看来，伦理学源于和遵循不忍之心。荀子则认为，人是通过习得社会规范变得有道德的，他们刚出生的时候并没有道德，只是后来“依靠编程”学会了道德行动。我们可以称这个途径为“社会程序的途径”。

孟子和荀子都强调人的独特性。孟子的焦点集中在人心的独特性上，这是人有而动物没有的品质。他的名言是：

> 所以谓人皆有不忍人之心者，今人乍见孺子将入于井，皆有怵惕恻隐之心。非所以内交于孺子之父母也，非所以要誉于乡党朋友也，非恶其声而然也。由是观之，无恻隐之心，非人也；无羞恶之心，非人也；无辞让之心，非人也；无是非之心，非人也。恻隐之心，仁之端也；羞恶之心，义之端也；辞让之心，礼之端也；是非之心，智之端也。人之有是四端也，犹其有四体也。（《孟子·公孙丑章句上》）

孟子认为，人与禽兽的差别就在于是否拥有不忍之心。在孟子看来，人并不是指生物学物种而是拥有道德的存在。只有拥有不忍之心的人才是有德之人。培养不忍之心就是继续人与生俱来的道德特性。

从这个意义上来说，孟子划的分界线不在智人和其他存在之间，而是在有德之人和其他所有的存在（包括无德之人）之间，判定的标准主要是有无不忍之心。沿着这条思路，人工智能机器如果获得这样的“心”，就可以被包括在有德之人之列。有了孟子这样的不忍之心，人工智能机器就能像有德之人那样在社会上行动。

当然，孟子的“心”不是指生理器官心脏，而是在心理上和情感上关心他人、善待他人的能力。在机器人伦理学上，如果我们遵循孟子的途径，或许应该在仿人机器人身上安装一个系统，以“不害人”作为压倒一切的原则，确保机器人不能采取极端破坏性的行动伤害人、伤害其他生命形式或者相互伤害。我们将拥有类似于约翰·黑文斯所说的“仿人心人工智能”。[14]该进路或许最适合医学人工智能机器。在照顾病人的时候，首要原则可以追溯到希波克拉底誓言中的“不能给病人带来痛苦与危害”。有了这个原则，人工智能机器的伦理行动将与孟子哲学一致。

但是，只要人工智能机器由人来设计和制造，并且由人来灌输一种伦理学（无论是孟子的还是荀子的），从根本上来说，它仍然属于荀子的哲学指导的范围，因为伦理学是社会构建的产物，是荀子主张的人为塑造（“伪”）。荀子也强调人的独特性，不过表达的方式不同。他说：

> 水火有气而无生，草木有生而无知，禽兽有知而无义，人有气、有生、有知，亦且有义，故最为天下贵也。力不若牛，走不若马，而牛马为用，何也？曰：人能群，彼不能群也。（《荀子·王制》）

荀子认为人是世界上最高贵的存在。他辨认出人类独特性的两个品质：一个是“义”，即道德得体意识；另一个是“群”，即组成群体的能力。人类要繁荣发展，此二者缺一不可。“义”让人产生一种

意识，我们不仅需要有道德，而且需要道德得体。荀子相信人性恶，认为如果顺从人性的要求可能导致社会陷入混乱和毁灭。如果没有圣人的社会构建，处于萌芽状态的“义”不足以让人克服人性恶而向善。[15]倪德卫认为“义”是荀子道德的源头，但他将其解释为有智慧的能力，并坚持认为“义”并不生来具有“任何特别的内容”。[16]倪德卫的解释让荀子的“义”看起来更像人的理性思考能力，使其成为组成社会群体的工具。荀子将“群”视为人类独特性的另一个特征。为了组成社会群体，人类需要设计礼仪法则，即判定哪些行为得体并有助于群体繁荣发展的法则。荀子强调的是社会管理中伦理法则的重要性，而不是不忍之心。人工智能伦理学的“荀子进路”主要关心的是设计出有效的伦理法则。

如果遵循“荀子进路”，我们可能会安装能够指导人工智能机器的行为法则，却没有压倒一切的保护性原则，即保护人、世界上的其他生命形式或者其他仿人机器人不受伤害。我们可能会给机器人制定其所需要遵循的法则，就像无人驾驶汽车要遵循交通法规一样。遵循这些法则，人工智能机器将主要完成所分配的任务，根本不考虑任务的性质。究竟使用“孟子进路”还是“荀子进路”，这取决于机器人的具体类别。在应对照顾人的机器人时，“孟子进路”似乎更有道理；对很多其他目的而言，“荀子进路”的可行性似乎更高些。儒家哲学的历史一直赋予孟子远高于荀子的地位。不过，最近的儒家研究似乎越来越重视荀子了。就人工智能伦理学而言，应该将这两种进路融合起来以便服务于未来的先进人工智能机器。

小说家、未来学家伊萨克·阿西莫夫曾经提出指导机器人行为的三个基本法则。它们是：

> 第一，机器人不得伤害人类个体，或者目睹人类个体遭受危险而袖手旁观。

第二，机器人必须服从人给予它的命令，当这种命令与第一法则冲突时例外。

第三，机器人在不违反第一、第二法则的情况下要尽可能地保护自己的生存。[17]

阿西莫夫的第一个法则清楚地表达了类似孟子的主题。即便它以一种原则的形式出现，但它仍然展示了不伤害人类的最高重要性。这样的要求反映了孟子表现出来的根深蒂固的人类价值观。这些定律以法则的形式出现，荀子应该也会接受。今天的人工智能伦理学需要比阿西莫夫的三大法则复杂和深远得多的法则。但是，无论人工智能伦理学将来有多么先进，将孟子和荀子的立场纳入进来似乎并无不妥。我们或许可以这样重新表述他们的原则：首先，高级人工智能机器必须以有益于人类作为首要目标；其次，其活动必须由基于第一个原则的法则来管理。人工智能机器能够采取有利于自己的行动，但是有益于人类是其首要目标，这是有关人工智能的其他考量的基础。而且，人工智能机器必须在符合第一个原则的法则的管理下行动。将这两个原则结合起来可以让人工智能机器与人类和平共处，成为人类共同体的延伸和组成部分。

人工智能技术的发展已经催生了一些重要问题，不仅涉及伦理学，而且涉及一般哲学。就像其他哲学传统一样，中国哲学发展了几千年，没有遭遇人工智能挑战引发的种种问题。现在，中国哲学必须应对这些挑战。它必须面对这样一个现实：人类独特性要么消失，要么变得越来越模糊不清了。我们必须考虑这种情形，即人工智能机器在未来社会成为重要的参与者。接下来，我将提出两个观点。第一个是儒家哲学的某些重要特征如“差等之爱”使其有能力调整道德要求回应人工智能的挑战。第二个是儒家哲学需要调整一些传统主张以便应对人工智能的挑战。我将使用儒家关于“学”的主题来展示第二个观点。

在一定程度上，在遭遇人工智能挑战时，与西方哲学相比，儒家哲学的困难更小些。比如，康德的伦理学以理性为基础，而理性被认为是所有理性之人的普遍性特征。虽然儒家接受人人都有成为道德君子的潜能，但他们并不假设人人都能达到同样的道德高度，为道德成就方面的差异留下了存在空间。从现代康德伦理学的角度来看，有道德能力的存在是一个不分层次的范畴，高级人工智能机器要么被纳入进来，要么被排除在道德领域之外。如果被纳入进来，它们应该像其他道德人一样得到平等对待，没有任何差别。由于为不同程度的道德成就和道德立场留下存在空间，儒家在接受人工智能机器作为道德行为人方面似乎拥有更大的回旋余地。他们既能够接受将人工智能机器当作有德之人，同时又无须将其视为与人类完全平等的存在。

儒家哲学最优先考虑的两大核心美德是“仁”和“礼”。“仁”代表了以关爱他人为基础的整体美德，“礼”则代表了管理人类行为的社会和文化法则。儒家思想家提出的另一个重要美德是“和”，即和谐。随着人工智能技术深入渗透我们的生活，并越来越多地成为生活共同体不可分割的组成部分，我们在设计人工智能系统时能够运用这些基本价值观。这将扩大儒家的道德领域以便把人工智能机器包括进来。儒家伦理学主张“差等之爱”，其表现形式就是孟子所说的“亲亲而仁民，仁民而爱物”（《孟子·尽心章句上》）。也就是说，亲爱父母（亲人），爱护关心他人，爱惜万物。最后一句中的“爱”应该理解为“爱惜”，即欣赏和谨慎地利用它。[18]“爱物”中的“物”包括世界上所有有价值的东西。儒家的伦理要求是有等级差别的。“爱物”是最低层次的要求，也是最容易做到的。仁民的要求就更高些，它超越了爱物的层次，还包括关心爱护和体贴的要求，因为它要求人们善待他人。最高的要求是“亲亲”，因为它不仅要求关心爱护家人，而且要对他们倾注很多感情。如果不亲爱亲人，就不可能把亲爱家人的关心和爱护扩展到他人身上。对待家人、他人和世界万物的所有，这些积极态度都需要依靠道德修身来培养。“亲亲”“仁民”“爱物”这三个范畴在程度和强度上存在差别。“亲亲”包

含着关爱，而关爱包含着怜惜。环保主义伦理学家已经指出，至少高等动物应该被当作道德对象来对待。高等动物之所以不能被当作道德行为者，是因为它们并不具有做出道德决策的能力。将动物作为完整意义上的“道德行为的施受者”来看待，儒家可能接受也可能不接受，但是，他们能够容纳在“爱物”这个范畴中关心动物的呼吁。对待动物的这种态度能够很自然地延伸到人工智能上。不过，人工智能对伦理学提出了额外的挑战，这是因为它们能做出在伦理上有相关意义的决定。这种（潜）能力要求我们在对待人工智能时不仅将其归为“爱物”的范畴，它们甚至有资格成为“仁民”范畴的候选对象。我们不仅应该爱惜人工智能，而且要关心、爱护它们，将其作为道德行为者或者道德行为的施受者来对待。

儒家作为活着的传统，一直随着时间的推移而不断调整自身。对民主观念的逐渐接受就是一个例子。虽然孟子很早就提出民本观点，但民主观念只是在思想家李贽和黄宗羲之后才出现在儒家传统之中。人工智能给儒家带来了另一项挑战，也是传统回应时代要求而更新自我的良机。这样的挑战之一就是尚贤观。在历史上儒家之所以逐渐走上舞台并具有影响力，其部分原因就在于它反对从前“世卿世禄”的世袭制而鼓励选贤任能。[19]儒家的尚贤将社会阶层流动性的基础建立在人的美德和知识/能力之上，美德和知识则需要在适当条件下辛苦努力来习得。因此，认真学习一直是儒家传统中的重要美德和道德要求。孔子认为学习是成为有德之人的最重要方式。他在《论语》中为人们树立了很好的榜样：“吾十有五而志于学，三十而立，四十而不惑，五十而知天命，六十而耳顺，七十而从心所欲不逾矩。”[20]（《论语·为政》）在他看来，学习是一种享受。他哀叹说：“古之学者为己，今之学者为人。”（《论语·宪问》）他说：“德之不修，学之不讲，闻义不能徙，不善不能改，是吾忧也。”（《论语·述而》）孔子高度重视学习的部分原因是只有通过学习，我们才能变成有知识的人和社会的有用之才。人工智能时代可能改变儒家的这个最根本要求。未来学家雷·库兹韦尔预测，人工智能将让我们有可能

把人的大脑和电脑连接起来。他说：“最有趣之事莫过于将你的大脑新皮质扩展延伸，将其与位于电脑云端的合成皮质结合起来。最终，我们的思考将主要由合成的新皮质完成。”[21]大脑新皮质是大脑中与人类心理功能有关的区域。人工智能技术将让人的大脑能够扩展到生理边界之外的地方。库兹韦尔甚至预测，新增加的大脑新皮质将被储存在可以无限扩张的云端。如果这个预测成真，选贤任能或许不再依靠传统意义上的知识。儒家需要重新考虑他们对知识学习的价值及其重要性的观点。他们很可能把重心更多地转向美德学习以及实际生活能力的学习。

当然，人工智能伦理学的任何建议在此阶段都是临时性的。儒家伦理学不仅仅是纯粹思辨或猜想的结果，它产生于理性和情感。这种伦理学的任何发展都要求我们汲取实际生活体验的营养。为此，儒家的人工智能伦理学需要伴随着人工智能机器的更充分发展和更多互动而不断改进与完善。[22]

李晨阳，南洋理工大学哲学教授，2015—2016年博古睿学者。主要研究兴趣在儒家哲学和中西比较哲学。主要著作包括《儒家的和谐哲学》（*The Confucian Philosophy of Harmony*）、《道与西方的相遇》（*The TaoEncounters the West*），以及100多篇学术论文。曾任国际中国哲学会会长、中国留美哲学会会长、美国哲学会亚洲委员会委员、香港城市大学2005—2006年资深访问学者、2008—2009年美国全国高等教育协会ACE（美国教育委员会）学人。

---

[1] Henry Bodkin, Robot that thinks for itself from scratch brings forward rise the selfaware machines, Science Robotics (January 30, 2019).

[2] 有关理性与智能关系的讨论，请参阅：Jonathan Baron, Rationality and Intelligence, Cambridge University Press, 1985。

[3] Ray Kurzweil, The Singularity is Near, Penguin Group, 2005.

[4] https://www.wired.com/story/ray-kurzweil-on-turing-tests-brain-extenders-and-aiethics/.

[5] Andy Clark, David J. Chalmers, The extended mind, Analysis, 1998, pp.7-19.

[6] Shogo Hamada, et al., Dynamic DNA material with emergent locomotion behavior powered by artificial metabolism, Science Robotics (April 10, 2019), https://robotics.sciencemag.org/ content/4/29/eaaw3512.

[7] Matt Hayes, Engineers create ‘lifelike’ material with artificial metabolism, Cornel Chronicle (April 10, 2019), http://news.cornell.edu/stories/2019/04/engineerscreate-lifelike-material-artificial-metabolism.

[8] Matt Hayes, Engineers create ‘lifelike’ material with artificial metabolism, Cornel Chronicle (April 10, 2019), http://news.cornell.edu/stories/2019/04/engineerscreate-lifelike-material-artificial-metabolism.

[9] P. J. Crutzen, C. Schwägerl, Living in the Anthropocene: toward a new global ethos, Yale Environment 360, 2011, http://e360.yale.edu/feature/living_in_the_anthropocene_toward_a_new_global_ethos/2363/.

[10] Jonathan Baron, Rationality and Intelligence, Cambridge University Press, 1985.

[11] Wallach Wendell, Colin Allen, Moral Machines: Teaching Robots Right from Wrong, Oxford University Press, 2009, p.63.

[12] Jonathan Baron, Rationality and Intelligence, Cambridge University Press, 1985. 11 Wallach Wendell, Colin Allen, Moral Machines: Teaching Robots Right from Wrong, Oxford University Press, 2009，p. 30。

[13] See the “Moral Machine” experiment by the MIT Media Lab (https://www.technologyreview.com/s/612341/a-global-ethics-study-aims-to-help-ai-solve-theself-driving-trolley-problem/), accessed on 30 December 2018.

[14] John Havens, Heartificial Intelligence: Embracing Our Humanity to Maximize Machines,Tarcher/Penguin, 2016.

[15] 更多讨论，请参阅拙文：Chenyang Li, The Origin of Goodness in Xunzi, Journal of Chinese Philosophy (38) 2011, pp.46-63。

[16] David Nivison, Hsün Tzu on 'Human Nature', Bryan Van Norden(ed.), The Ways of Confucianism: Investigations in Chinese Philosophy, Open Court, 1996.

[17] Wallach Wendell, Colin Allen, Moral Machines: Teaching Robots Right from Wrong, Oxford University Press, 2009, p. 12.

[18] 比如，东汉经学大师郑玄将《诗经》中的"爱莫助之"中的"爱"解释为"惜"。宋代大儒朱熹也用"惜"来解释《论语·八佾》中的"尔爱其羊，我爱其礼"中的"爱"。

[19] 有关儒家民主的讨论，请参阅拙著《东亚挑战民主：比较视角下的贤能政治》（贝淡宁、李晨阳合编）：Daniel Bell, Chenyang Li, The East Asia Challenge for Democracy:Political Meritocracy in Comparative Perspective, Cambridge University Press, 2013。

[20] 遵从廖名春的解释，我将"耳"读作"聊"，解释成和谐。请参阅：廖名春：《孔子真精神：论语疑难问题解读》，孔学堂书局，2014年。

[21] https://www.inverse.com/article/33373-ray-kurzweil-singularity-thoughts,accessed on 17 December 2018.

[22] 本文作者在此感谢博士生胡建萍在酌定中文译稿过程中的帮助。

人工智能与情感、秩序：从儒家的视角出发

干春松

人类历史上，重大的技术进展往往意味着一个新的历史阶段的开端，比如我们熟悉的石器时代、青铜时代等。虽然有一些历史阶段并不直接以科学技术的发展为标志，但其背后真正的推动力量，也离不开重大的技术进步所带来的社会生产模式的转变，比如蒸汽机所带来的动力革命，使人类进入大工业时代，从而为形成全球性的资本主义时代奠定了基础。

从20世纪开始的计算机的发展，一直被认为是新一轮的技术革命。一直以来，人们对这次革命所可能带来的影响都十分关注。不过，直到克隆技术以及涉及人类自身结构的人工智能兴起之后，人们的讨论才开始升级。

与以往的科技进步不同的是，人工智能所涉及的基因编辑和智能机器人构成了对于人类特性的根本挑战，无论是在人类繁育领域，还是在机器人智能化的可能空间方面，都引发了对于人性、人类的生存发展模式以及人类的未来的关注。

这些变革所带来的后果目前还难以评估，有人侧重于机器替代所带来的劳动力过剩，有人关注人工智能所可能带来的国家之间差距的扩大以及由此造成的对现有国际秩序的挑战。当中东某国给机器人以“国籍”的时候，人工智能所带来的“权利”以及“身份”问题引发了国际争论。更多的讨论集中在技术层面的可能性上。那么，具体到儒家思想的视角，人工智能所可能带来的冲击会是怎样的呢？如果说

19世纪开始的儒学危机是现代化的生产方式和价值观所带来的，那么人工智能的兴起会给人类社会带来什么样的变化，从而是否会构成儒学的一场新危机呢？这或许是特别值得思考的问题。

儒家对生命意识的独特认知

儒学传统对于人的认识是极其丰富的。从个体生命的形成而言，儒家既强调父母、夫妇所形成的基于血缘的自然生命秩序，也提倡由此情感出发推广到世间万物的一体之仁。由此，儒家的生命意识是生理性的“人类个体”和社会性的“理想人格”的矛盾综合体。不能认识到这一点，就难以理解儒家思想对于生命价值的基本立场。

从《论语》中，我们可以看到孔子经常在区分君子和小人，这就是希望通过人类对自身人生理想的肯定克服生命意识中动物性的一面。这一点在孟子那里得到更为充分的展开。孟子喜欢通过区分人与动物提炼他对于生命价值的认识。他说：“人之所以异于禽兽者几希，庶民去之，君子存之。舜明于庶物，察于人伦，由仁义行，非行仁义也。”（《孟子·离娄章句下》）在孟子这样的古典儒家思想家看来，人的本质特性是通过自己的努力成为“君子”，这个过程就是人的“自我定义”的过程。在孟子看来，人先天具有同情心，并发展出判别是非的能力、遵循社会规范的能力和服从尊长的态度。孟子认为人并不能自然地意识到自己有这方面的禀赋，或者说即使自觉到有这样的能力也不能保持并增强之。终于在荒怠中，这样的禀赋被掩盖，有些人失去了判别是非善恶的能力，成为“小人”；而总有一些人能够在社会的磨砺中不失掉他的良知，并通过各种方式存养浩然之气，这样的人变成“大人”。在成为“大人”和堕落为“小人”的选择中，个人的意志尤为重要，通过选择的过程，儒家的生命意识得到显现。

与孟子相比，荀子的生命意识看起来缺乏主动的态度。荀子对人性的认识也与孟子相去甚远。荀子更多地继承孔子性近习远的人性观，持一种“本始材朴”的性朴论。不过，更为明确的是，荀子相信人性中存在自私的倾向，需要圣人制定礼仪来规范人们的行为，以使社会在一个各安其位的状态下保持和谐。虽然孟荀之间对于人性的认知差别如此之大，但对于人能成为什么样的人，他们则有许多共同点。比如，荀子也肯定每个人都可以成为圣人，只是孟子认为要靠自觉，而荀子则强调圣贤的启发和教育的重要性。如果说孟子更为看重人的自我觉悟的话，荀子则强调教化在“成人”过程中的作用。

综合孟子和荀子对于“人”之为人的讨论，儒家对于人的认识整体上倾向于从“可能”到“实现”的角度来理解，这样，就把生物意义上的人与进入社会生活之后的人区分开来。如果放任自己的生物属性，那么人就仅仅是一个动物；反之，如果能通过学习和完善的方式存养、扩充自己，从而成就自己，就会让自己脱离动物的属性，凸显出人类异于动物的道德伦理属性。

在孟子所强调的人与动物差别的“几希”中，我们可以看到，孟子不是从本能的角度来区分人与动物的不同的，而是从人能否认知到自己的道德责任来判定人之所以为人的“标准”。《孟子·尽心章句下》中有一段话可以理解为“非本能化”地定义人类的经典段落：“口之于味也，目之于色也，耳之于声也，鼻之于臭也，四肢之于安佚也，性也，有命焉，君子不谓性也。仁之于父子也，义之于君臣也，礼之于宾主也，智之于贤者也，圣人之于天道也，命也，有性焉，君子不谓命也。”对于“性”和“命”含义的解释需要我们进行专门的讨论，这段话的核心在于说明这些生理本能并不是人之所以为人的“最关键”的“规定性”，而仁、义、礼、智这样一些道德准则却是人的“本质”的体现。儒家也由此来区分君子和小人。在孟子看来，即使你生而为人，如果没有道德意识，也不能担负道德责任，其实也不能称你为真正的“人”。这样君子、小人就与人的道德属性和自然属性之间产生了“一致性”。

儒家对于人的认识整体上倾向于从“可能”到“实现”的角度来理解，这样，就把生物意义上的人与进入社会生活之后的人区分开来。

按照孟子的标准，身而为人并不表明你可以活得像一个“人”，生理上的人只表明你具有成为人的可能性、一种潜力，要成为一个真正的“人”，就必须对自身所具备的克服欲望和私利的“良知”进行“存养”。人之所以为人最主要的一点，并不是人与动物相似的地方，而是人与动物之差别的关键点。这个关键点就是人类具有动物和世界上的其他事物所没有的“道德意识”。

孟子和荀子虽然在人性问题上的认识有所不同，但荀子也强调礼仪规范是人类区别于其他事物的关键。

《荀子·王制》中有一段话很有趣：“水火有气而无生，草木有生而无知，禽兽有知而无义，人有气、有生、有知，亦且有义，故最为天下贵也。力不若牛，走不若马，而牛马为用，何也？曰：人能群，彼不能群也。”在这里，荀子强调了人的组织能力是促使人在生理能力远逊于牛马的情况下也能够役使牛马的关键。这主要依靠的就是人具有其他物种所没有的特质“义”，这就是说，“义”才是使人与动物区分开的特质。如果我们把“义”理解为正义或正当性，那么“群”所指的就是人的“社会性”特征，人的道德特征决定了他们可以合作，并在与动物的竞争中获得优势。或许荀子并没有像孟子那样强调“人”在道德活动中的“主动性”，不过荀子依然从道德生活中确立人与动物之间的差异，并将礼仪性的存在视为人的基本特性。

与启蒙运动以来以人的“生物”属性来强调人的平等性和自然欲望的正当性不同的是，儒家强调“学以成人”，强调对自然欲望的“克制”和社会属性的凸显。

与启蒙运动以来以人的“生物”属性来强调人的平等性和自然欲望的正当性不同的是，儒家强调“学以成人”，强调对自然欲望的“克制”和社会属性的凸显。由此，儒家并不是从人的自然禀赋来定

义人的，而是从“社会性”着眼，要求每个人达到“人”的标准，通过对于人与动物关系的独特认识，通过强调人的社会责任和道德伦理规定人。当然，这样的思想也一度被视为儒家难以适应“现代社会”的主要“缺陷”。

传统儒家对于个体生命的终极意义有两个非常典型的表述。一个表述出现在《左传》中，《左传·襄公二十四年》谓：“大上有立德，其次有立功，其次有立言，虽久不废，此之谓不朽。”唐代学者孔颖达在《春秋左传正义》中对德、功、言三者分别做了界定：“立德，谓创制垂法，博施济众”；“立功，谓拯厄除难，功济于时”；“立言，谓言得其要，理足可传”。这三者所强调的是个人对社会的贡献。立德侧重于对社会规范和价值观的确立，立功则针对社会发展的实际贡献，而立言则是要传达对于群体生活有贡献的观点。毫无疑问，个体自身并不能独立体现出价值，只有在社会性和人类历史的维度下，个人的存在才能表现出永恒性。

另一个表述更为中国传统的士人乃至现代人所接受，即《大学》中所提出的“格物致知正心诚意修身齐家治国平天下”，如果我们把这8个条目分成两部分的话，前4个条目属于个人人格的自我完善部分，而后4个条目可以看作个人生活在这个世界上的意义。个人生命的意义与他为社会、为家庭所做的贡献密不可分。他应该担负起人类生命延续的使命，就此，我们就可以理解为什么《孝经》和其他关于“孝”的论述中，特别强调了家庭的延续，在这样看似注重家庭利益的言说中，我们可以看到家庭、国家和天下之间的一致性，从而认识到儒家对于生命价值的认知奠基于家庭但指向国家和天下。

自亚当·斯密以来的许多西方思想家认为，人的行为具有“理性人”的特征，即总是会做出对自己有利的选择。并最终通过利益的调整和妥协建立起交易规则，这也称为“理性人假说”。由上述梳理可见，儒家思想认为人的生命的本质并不是“利己”，而是从人类的整体来思考人的特性和人的权利。我们称之为“道德人假说”。我们并

不能说人类社会走向现代化是一种由“道德人假说”向“理性人假说”的转进，而只能说人类越发认识到个体自然存在的价值，人类的未来需要将个人的权利和社会的责任结合起来，而不是简单的非此即彼式的互相否定。

1840年以来，受西方思想的影响，中国人的价值观尤其是对于个人的认识发生了巨大的变化，这包括重视个体的独立性、个人的权利等。不过，在这样的社会大转折中，整体的价值依然被肯定。从晚清到现在，强调个人权利的人几乎同时也是“国家主义者”。在社会主体的价值目标中，我们一直在努力建立一个“现代国家”，但民族国家本身所内含的矛盾，即国家在承担对内保护国民和对外防御侵略的职能之间经常会出现基于国家利益而导致的“价值紧张”。一个保护国民的个人权利的国家，可能同时是对弱小国家进行殖民乃至剥夺别的国家国民的个人权利的罪魁。这一点也在全球现代化进程与殖民化进程从某种意义上的重合上得到了历史性的证明。

进入21世纪之后，中国人在价值导向上的总体倾向还是“国家”主导型的。“国家”与“人类命运共同体”这两项看似对立的价值，在“共享”的观念下得到协调，从某种意义上来说，这也可以被认为是传统价值与现代观念在当下中国的“和解”。从中可以看出儒家价值观在现代中国的政治生活中所发挥的影响。

人与机器之“几希”

孟子所强调的道德意识帮助人类将自己与其他动物区分开来，并确定人类的人格目标。那么，人是通过什么样的心理机制使这样的道德意识既存在又活动的呢？这样的道德触发机制在传统儒家中有多种表述方式，比如孟子的“良知”，《周易》和《礼记》等经典中的“感”。

据研究，动物甚至一部分植物也存在“意识”（感觉），它们所缺乏的是“自我意识”。动物不可能提出“我是谁”这样的问题，也就难以对生命的本质、生命的意义进行反思。很显然，就目前的认知而言，道德意识是只属于人类的能力。人通过自我反思的能力而建立起来的“意义”和“价值”世界，让人与动物之间得到区分。

人类对自我的认知通过实践活动不断丰富。在种种实践活动类型中，技术的进步所带来的人的自身能力的扩展，不断改变和充实着人对于自己的“认知尺度”。从历史的角度看，人是通过学会制造和使用工具获得其在生物世界的优势地位的。的确，任何意义的技术进步和科学发展，都会增强人类对于自然的“利用”和“控制”，从而改变人类对于自己的认识。这一点在西方自我认识史中表现得特别清晰：由早期的从自己的身体来看待自然界到确立一神教信仰，将自己的决定权交由绝对的存在来支配，再到启蒙运动之后，将人类自己作为自然的主宰。这背后都体现了人类在历史的进程中不断调整“自我认知”的过程。这个过程表面上看与知识的进步相一致，但知识进步的动力则是人类的实践能力的增强、工具的推进。

中国早期的经典《周易·系辞》中曾描述圣人通过“制器”不断推动社会进步的故事。每一次技术的进步都意味着人类离更为原始和“自然”的生活方式又远了一步。在中国古代，道家一派的学者更倾向于对技术进步和人性完善之间的关系持犹豫的态度，或者直接认为使用技术会导致心灵的纯洁性受到伤害。在《庄子·天地》中，记载着这样一个故事，孔子的弟子子贡路过汉阴，见一个老者辛苦地用瓮在灌溉，就向他提出利用机械灌溉的建议，不料“为圃者忿然作色而笑曰：‘吾闻之吾师，有机械者必有机事，有机事者必有机心。机心存于胸中，则纯白不备；纯白不备，则神生不定；神生不定者，道之所不载也。吾非不知，羞而不为也。’”大意是说，使用机械的人存有投机取巧的心思。坚持用原始的方式灌溉的老者并非不知道使用机械会省力，他是不愿意取巧。《庄子》用寓言的方式所要提醒我们的

是，到底什么才是人的“自然”状态，任何借助非“自然”的方式来改变自己的行为，都会导向人特性的异化甚至丧失。

庄子的思考中显现出中国思想史上对于人的特性认识的巨大分歧。在儒家等学派看来，人类能使用工具恰好是“人”与“动物”的差异之所在，如果放弃使用工具，那么人与动物的区别何在？科技越发达，人类驾驭自然的能力就越强，这是人对于自身的可能性的扩展，也就更为充分地凸显了人的本质。

从某种意义上来看，庄子对于技术文明的批判很具有超前性，因为处于文明早期的先秦时代，科学技术的发展程度对于人性的冲击还不是根本性的。由于技术能力的制约，人们还难以对自然的力量构成“反制”。由此，技术与自然之间构成一种和谐的关系。人们借助技术手段并不能改变自然的状态，而只是利用自然的能力得以扩展。人们在河流上架设桥梁，在山坡上修筑梯田，只是根据自然界的不同形态“因势利导”而已。

创造工具来合理地改变自然与保持事物特性之间的关系，一直受到中国古代思想家的关注，尤其在郭象的《庄子注》中，围绕着牛是否应该被“穿鼻子”，马是否应该加上“马鞍”这些问题展开过讨论。《庄子》强调的人的自然本性“受之于天”，如《秋水》中说：“牛马四足，是谓天；落马首，穿牛鼻，是谓人。故曰：无以人灭天，无以故灭命……”不能以人的力量来改变自然，给牛穿鼻子和给马加马鞍都是对其自然本性的“毁灭”。但是，郭象在注释《秋水》时说：“人之生也，可不服牛乘马乎？服牛乘马，可不穿落之乎？牛马不辞穿落者，天命之固当也。苟当乎天命，则虽寄之人事，而本在乎天也。”从这段话中我们得知，人类的生活中不可能缺少对于自然和别的动物的利用，而这些动物和自然事物之所以可以被人利用，是因为它们的存在合乎人的“目的性”，所以郭象对于《庄子》中将人的特征绝对固定于“原初的生命状态”的思想进行了转变，将人对于

自然的利用看作人类特征的一部分，并认为这是将人和动物加以区分的最关键之处。

经过早期由智人到人类的演化，人类逐渐成为地球上物种中的“强者”。对于火的掌握就改变了人的饮食习惯，“加工”食物本身对于人类的生活习惯和对季节的依赖发生变化；通过栽培适合人类食用的粮食和驯化、饲养动物，改变人类的食物结构，最终让人类可以应对因季节变化而造成的食物短缺，增强人类在自然竞争中的生存能力。人类的任何进步都伴随着与自然之间的相即相离关系，既利用事物的自然属性，又赋予其人类的目的属性。

就人类对于工具的使用和对于自然界的改造所造就的历史阶段的跃迁而言，其表现方式也有很大的不同。早期人类历史或者说史前的人类历史的划分以工具的进化作为其主要依据。在进入人类主导的社会之后，社会变迁的标志则转向以社会组织方式的变化为主要标准，对于工具的使用反而“隐藏”在社会结构的变化背后。

人类能够记录自己的历史，并凝练自己的观念，开始通过社会分工组织社会，并将之“神圣化”，这是人的自我意识提升的关键一步。每一次社会组织方式的转变背后，都起因于人类技术能力的转变，而技术的进步都日益传导为社会组织方式的转变，比如家庭、城邦国家以及未来可能的“大同世界”。作为现代社会奠基的大工业化生产方式和由航海带来的全球资源配置的可能性，最大的推动因素就是蒸汽机取代了以往纯粹依赖水和动物的“动力”系统，火车、轮船这样的新的运输方式和人类沟通方式成为可能。新的动力的形成，使人们的生产方式产生了巨大的变化，比如运输方式的革命使交通枢纽成为原料的集散地和工厂的聚集地。资本主义的生产方式和现代民族国家的奠基都与这一轮的技术革命有内在关系。

蒸汽机的发明将技术进步与人类生活之间的关系提升到一个新的阶段。如果按《庄子》和郭象对“落马首”等问题的讨论范围，早期的思想家对于人与自然的关系的争论主要在于人能否借助对于动物和

植物的改造来改善自己的生存处境，那个时代的技术发明还局限于“利用”自然物的一些特性来进行直接的使用，比如利用马的“速度”来改善人际交往的可能性，利用牛的“耐力”来弥补人类自身在体力上的缺陷。就连基于冶炼而形成的青铜器、铁器的使用，也只是顺应对象物的自然属性并加以延长而已。举例来说，人类对于铁器的使用是工具演化过程中一个极其重要的阶段，但就“技术”进步的层面来说，只是利用了铁的坚硬度来开发其作为“工具”的特性。

蒸汽机的发明则与早期的技术进步在性质上有了很大的不同。从那个时代开始，技术的进步已经不再是“顺应”自然，而是基于人类对于自然规律的认识试图改变自然或建立一个新的“自然”。就“目的”而言，科技的发展已经脱离了简单的“改善生活”的层面，同样具备了控制人类甚至毁灭人类的力量。比如，20世纪的核能技术固然能提供人类所需要的能源和动力，不过，第二次世界大战时期美军在日本所投下的原子弹和苏联所发生的切尔诺贝利核电站事故，乃至最近在日本发生的福岛核电站事故均表明，技术的进步不断在脱离人们对于技术使用的最初动机。而这些发明就原理上来说，固然是对自然规律的利用，但科学的发展逐渐脱离了对自然的“顺应”而转变为“控制”乃至“不合理的改变”。经由科技的创造物在造福人类的同时，或立刻成为人类自身的毁灭者。也就是说，就科技“形态”而言，现代科技与自然之间的关系和古典时代产生了很大的差别，对此，包括海德格尔在内的许多哲学家已经有了充分的反省。在他们看来，技术的进步逐渐使技术成为人类生活的主宰者。

但海德格尔所处的时代还只能算是技术控制人类生活趋势的转折点，直到人工智能这个新技术的到来，人们开始需要反思技术可能给人类带来的彻底控制。

自20世纪下半叶开始，科技发展的重心落在生物和计算机领域。首先，生物技术的发展之所以如此引人关注，主要原因在于基因等领域的突破性进展开始改变人类物种的一些基本生存和演化方式，这引

发了巨大的伦理危机。比如20世纪克隆技术的发展，从技术上解决了人类的自我复制问题，这样，人类以男女结合为前提的生育方式在“理论”上被“取代”了。而基因编辑技术的发展则使我们可以对自身的生命状态进行干预，在人类的延续过程中有充分“选择”的可能性。从正面的角度来看，基因编辑技术可以让人类诞生出健康的下一代，但从负面的意义上来看，这种“选择权”也可能被“滥用”，即一些人可以通过权力或金钱选择增强自己的智商或身体的部分机能，从而导致社会公平和伦理的危机。这也是贺建奎的研究成果会引发巨大争议的原因。

基因编辑技术可能会有助于人类进行体细胞的基因治疗、生殖细胞的基因治疗。不过，这样的技术所可能造成的伦理危机也一直为人们所警惕。按照目前全球主要国家所确立的原则，针对基因编辑在人类生殖细胞或早期胚胎的使用，如果胚胎在体外培养不超过14天，也不被移植入人或者其他动物的输卵管或子宫，就会被视为“研究行为”，应遵循使用人类样本研究相关伦理规范。不过，如果有人是“以生殖为目的”，利用基因编辑或者其他技术改造人类生殖细胞或早期胚胎中的遗传物质，就会被视为一种违法行为。因为这些基因改变影响的将不只是一个个体，还包括其将来的后代，乃至人类的整个基因池。

其次，对人类的特性可能造成更大影响的就是人工智能。一般的数据驱动的知识技术学习，替代人类进行一般劳动的机器人，甚至类似“深蓝”、AlphaGo这样具有学习能力的机器人，依然可以算作传统意义上的“工具”。即使具备人机、脑际互相协同融合的智能形态，仍是“工具”意义上的对于人类自然属性的延伸。将对人类造成“颠覆性”挑战的是人工智能进入情绪管理和自我意识的“人工设定”的阶段，人类可以设计同类型的“品种”，并主动选择自己所需要的人类类型。人类还可能制造出自己的“对手”甚至“敌人”。无法预测的是，这些已经具有自我意识的机器人可能会“自行定义”他们自己

的意义和生命目标，从而以他们在生理、脑力上的优势确立其支配地位。

现有的人工智能的讨论聚焦于人类脑智能以及人机智能融合的问题，很显然，人工智能何时具有情绪控制能力以及自我意识，这是更为棘手的问题，这样的技术发展突破了以往所有科技进步的影响范围，而可能构成超过人类控制的新阶段。这既是人类发展的全新阶段，也存在着对人类的生存和发展构成致命威胁的后果。

人工智能会带来人类的进化还是异化

前文已经指出，人工智能对于人类最大的威胁是可能的智能机器人的“自我意识”的产生。不过，就目前的发展势头而言，人工智能的发展要达到那个程度可能还需要很长的时间。从眼下的情形来说，人工智能主要还是属于“工具”形态的技术进步。获益者主要是一些可以用机器人替代人工的企业，这些企业只要付出购买和维护机器人的费用，无须对付难缠的工会，这使它们的劳动效率大大提高而管理成本却大幅降低。

在机器人替代了大部分体力劳动之后，社会分工体系会出现新的变化。人类的闲暇时间大大增加，不过这种类似社会学意义上的预测并不是本文关注的重点，在这部分我主要想就几个可能与人类生命本质有关的问题做出一些尝试性分析。

机器人与被改造的人

人工智能与以往众多的科学技术进步的不同之处在于，它集中于人的智能领域，而按儒家对于人的特性的认识，人之为人的关键在于人的“智”而非生物学属性。

目前争议较多的是人工干预生育过程，人们担心这会导致人类“先赋性”的智能差异。人工干预生育过程已经有比较长的历史，从正面的角度看，可以通过预先的检查防止一些带有先天性疾病的婴儿出生。但如果智能技术和生物技术相结合，我们可以预见，以后人类可以在婴儿出生时改善其智商或体质。即使在生产时没有进行技术干预，依然有可能通过人工植入智能芯片或其他手段改变其大脑运行过程，达到改变的目的。

由此可能会引发诸多问题，首先，在历史上，人们主要是通过世袭或接受高质量教育的方式延续阶层地位的，而人工智能的发展可以通过技术手段使人拥有常人所不具备的身体或智力能力。据此，拥有权力、金钱的人或技术持有者更容易获得这样的“技术优势”，这可能会导致比以前所有的形态更为严重的社会不公，因为拥有金钱和权力的人可以优先使用这样的技术，从而确立他们在阶层竞争中的优势。

其次，在现有的民族国家体系下，国家利益往往成为阻碍技术转移和技术分享的屏障。因此，人工智能技术进步所可能带来的对于人类竞争力的提升，并不能保证泽及所有国家，反而可能加剧目前业已升级的国家间的不平等秩序。而如果智能技术用以改变“种族”的素质，那么也会出现不同国家的族群之间在能力上的不平等，进而发展成一种变相的“种族主义”，导致国际冲突的加剧。

再次，从技术的角度看，未来世界可能会存在三种人：第一种是自然人，即在天然环境下出生的人；第二种是混合人，这类人利用了生物技术和人工智能技术，在人脑和身体机能上存在着优势；第三种是人造人，这其中也可以分为通过基因技术和人工手段“创造”的人，以及具有人的一切思考和智慧能力的机器。与目前我们通过肤色等方式区分人种，通过公民身份确立其国家归属的方式所不同的是，人类可以有其他标准的分类系统。如果出现人机混合的存在，或者说具备一定自我反思能力的机器人，其能否被随意终结其“活动”状

态，被随意拆解甚至“消灭”，这也要取决于人们如何定位这类生命形态的存在的“社会地位”。

最后，正常的人和通过人工智能的方式获得别人所不具备的认知能力的人，在假定他们都存在“自我意识”的情况下，他们对于何为“人”的认知会出现差异。落实到社会政治层面，在身份确定和权力义务的确立方面是否应该对他们进行差别化处理，这也会成为对社会组织的巨大挑战。

从儒家的立场而言，我们应如何理解和应对这样的挑战

儒家社会一般被理解为等级性社会，但事实上传统儒家强调“学以成人”，一个人的社会地位要靠自己努力去争取。儒家强调为仁由己，“吾欲仁，斯仁至矣”，从人为天生的角度，认定人与人之间的平等。

在社会关系层面，儒家从人的社会性来理解人，也就是说，每个人自出生那天开始，就具备其社会身份，因此，指导人们行为的并非抽象的德性，而是“根据我们熟悉的、社会的‘角色’而找到指南”[1]。这样的角色首先是基于人伦来确定的，其次扩充到政治、社会和职业身份。不同的角色存在着“差等”，这样的差等是自然形成的，为你的角色所确定，比如父子、夫妇之间，就会存在着权力和义务上的差异。这种差异被梁漱溟称为“伦理本位”。但一个人的社会地位并不是由他的角色决定的，不能通过血缘的联系加以继承。儒家十分强调血缘伦理，但“讥世卿”，除了君主之位可以世袭外，其他的社会地位都需要人们去争取。

因此，我们可以相信，从儒家的义理出发，一定会反对通过权力或财富的继承方式获得在智能和社会地位上的优势，也同样会反对一

部分人凭借其社会地位来获得智能和生物上的优势地位。

基因编辑技术和人工方式复制人的行为，会造成巨大的伦理困境，以血缘作为基础的儒家伦理学更是如此。一个基因婴儿该如何确定其伦理身份？因此，站在儒家的立场上，应该会反对借助科学手段来“制造婴儿”，甚至会对通过人工授精的方式解决不孕不育也保持一定程度的警惕，因为这样的人类再生产方式都将造成人类的血缘角色和社会角色的混乱。

借助基因技术等技术会让人的生命延长，这会造成家庭结构的变形。比如，如果以现在人类的平均寿命75岁计算，我们所理解的人伦亲情其实基本上限定在祖父和孙辈的范围内，如果超出这个范围，亲情关系就会疏远，而实际上与普通的“陌生人”相类似，因此人工智能和生物技术将从根本上冲击儒家伦理体系。

人工智能和生物技术对儒家伦理的最大冲击是对于“亲情社会”的破坏。

在大多数中国人的心目中，好的生活就是“儿孙绕膝”的家庭和睦生活，而善的生活则是“老吾老，以及人之老；幼吾幼，以及人之幼”，即将对父母和孩子的爱扩展到所有的人。相比之下，中国人更愿意接受集群式的生活状态，因此对于人与人之间的情感沟通的要求也比较高。所以，从人际交往的角度，我希望以后的人工智能要植入“情感”特别是“亲情”的因素，在人工智能的非血缘化社会形成后，人与人之间的伦理亲情便不复存在，由此形成的人类社会是否变成一个“无情”的社会，或者变成一个情感程序化的社会，这必将造成新的道德困境。

在人工智能的非血缘化社会形成后，人与人之间的伦理亲情便不复存在，由此形成的人类社会会不会变成一个“无情”的社会。

人工智能的社会也会导致儒家以教化为手段的社会整合过程失去价值。

儒家特别强调教化在社会秩序建构过程中的作用，认为人们应该而且可以通过教化的方式达到互相理解，建立起共同体的基本价值观。而人类对人工智能和智能机器人最大的期待就是它们可以轻易获得人类至今为止的所有知识，但其能否在价值和伦理方面取得同样的进展则还是一个未知数。以在中国被广泛宣传的AlphaGo与柯洁和李世石的围棋比赛为例。围棋的变化被认为是所有智力游戏中最为复杂的，但AlphaGo还是很“轻易”地掌握了，并战胜了当下最高水平的棋手。但是，要让AlphaGo具有人类棋手因失败而导致的挫败感，应该比学会如何战胜他们复杂得多。因此，植入知识并不是人工智能对人类秩序毁坏的终极点。可以设想的后果是，如果人工智能拥有自主的情感处置功能，那么“他们”与正常的人类之间会产生族群问题以及其他可能发生的社会对抗。

加强对科技文明的反思是中国哲学发展的新方向

《周易》被视为中国哲学的源头，而《周易·系辞》中十分详细地讨论了周易的卦与技术进步和物质发展之间的关系。书中指出伏羲做八卦是“类万物之情”，即按照万事万物的特征来做八卦的。伏羲发明了渔网，教会人们捕猎，这是受到“离卦”的启示；在伏羲之后的神农氏通过发明农具教会人们耕种的技术，并建立起交易市场，这是取之于“益卦”和“噬嗑”；黄帝、尧、舜垂衣裳而天下治，教会人们建立社会秩序，这是取诸乾、坤两卦的智慧。如此，一方面把古代圣人之所以为圣人与技术进步联系在一起，另一方面强调了社会进步都受到《周易》的哲学原则的影响，技术和社会的发展与思想发展之间就建立起了关联性。从总体倾向来说，《周易》认为技术的发展与礼乐社会的建立之间应是一个“顺接”的关系，并不主张借助技术来征服自然，而是试图建立起自然和人类社会之间的和谐关系。

秦汉以后中国哲学的发展似乎越来越远离对于科学进步和技术发展关系的讨论，日益转向对于人性与社会秩序的关注，由此，中国古代关于天人关系的理解，往往会倾向于以天道明人事，立足点往往在“人事”之上，“天道”本身反而缺乏深入的研究。

除了徐光启、方以智等少数思想家外，中国古代思想家大多缺乏对于科学现象的探究和关注，这导致中国古代哲学所关注的范围相对狭隘，实质上对人生活的处境也难以有深刻的认识，甚至有人认为中国古代并不存在西方意义上的“哲学”。

这样的评论当然会导致“民族主义”意义上的反击，辩解者指出中国哲学虽然并没有发展出形式化的“逻辑理论”和“本体论”思想，但是有着最为丰富的人性论和伦理思想。中国传统哲学对于人伦道德的讨论的确十分丰富，不过，古代哲学家对于科技的发展始终不是十分热心。他们也没有真正深入探讨技术和人性的关系。随着近代以来哲学学科的建立，中国哲学家和中国自然科学家依然缺乏实质性的学术交流，这导致科学家很少对他们所做的研究做出哲学性的概括，而哲学家则难以真正掌握科学发展的前沿内容。在贺建奎的基因编辑婴儿诞生之后，中国的科学界和哲学界最初并没有提出基于各自学科的判断或有价值的分析，也没有尝试制定相关的准则来规范以后的类似研究。其根本原因就在于，中国的伦理学界基本没有参与相关的研究，而中国科学家也缺乏对于人类价值观的深刻理解。

如果我们将传统中国哲学的优势聚焦于人性和伦理领域，那么从贺建奎的事件中，面对科学技术对于我们人类生活形态的重大改变，我们可以直观地判断，中国哲学界在这方面的优势地位已不复存在。从某种意义上来说，传统儒家对于性善和性恶的讨论，作为一种哲学史资料是重要的，但用来分析基因编辑和人工智能对于人的生存形态与生命意义的讨论则显得脱节。

人工智能的发展会促使人重新思考哲学的基本问题，比如《黑客帝国》这样的电影中就反复出现了“我是谁”“认识我自己”“我选

择，我成为”等著名的哲学命题。这从一个侧面反映了艺术家对于人工智能可能带给人类自我认知领域的冲击进行了思考。的确，当人工智能的未来难以真正把握的时候，或者可能当我们认识到人工智能的发展将会让人类难以把握自己的未来的时候，“我是谁”“我选择，我成为”这样的问题才重新焕发出其力量。

人类似乎到了一个新的十字路口，以往的人类创造出不同的工具为自己服务，但当工具本身可能具有摆脱创造者的控制力的时候，就好比人类孕育的下一代成长为一个具有独立生存和判断能力的成年人时，他的发展已经不再由人类定义。由此，人类可能需要重新定义自己。

基因编辑和人工智能对中国人的观念世界与秩序体系所可能带来的冲击会不会强于强调个体的西方社会呢？这并不容易判断。不过相比现代性对于儒家社会伦理的挑战，人工智能和基因编辑技术改变人类本身这一影响则要远远大于现代性背后的人的观念。因为现代性的思路只是要消解家庭血缘伦理，而人工智能和基因编辑则将会从源头上改变人类自身，改变人类的再生产和组织形态。在这样的社会里，儒家可能连“游魂”都做不到。积累数千年的文化心理结构和思考模式，完全可以通过技术手段加以“清除”或“添加”，那么价值观就可以如程序般先期预存于人们想要制造的“人类产品”中。对儒家而言，当家庭不再成为社会结构的基础或基本的财富单位，儒家的秩序理念需要做什么样的调整？更为致命的问题是，做出调整之后的观念是否还可以被称为儒家的“人工智能版”。

此外，人工智能的发展将基本上改变人类的学习过程，由此，家庭和学校在知识与人格训练过程中的重要性将减弱，人类的基本组织方式和价值目标就会发生改变。如果机器人大量占据人类的工作岗位，甚至能满足性和休闲需求，解决人类的家庭看护任务，家庭成员中也会出现许多非血缘的“成员”，这同样会导致儒家的道德教化功能和“角色”伦理体系失效。

需要提醒的是，人工智能的发展，使对于这项技术的发展对人类带来的影响的认识并不能指望现在的学院体制内的哲学家，而需要依靠那些人工智能的科学家转而思考他们所进行的工作对于人类的意义，也就是说，这些科学家要力图使自己成为哲学家，对自己所从事的工作的意义做出判断。

对此，哲学家的工作是什么呢？他们应该与科学家形成合作关系：一方面，可以帮助科学家了解传统中国人对于生命意义的认识，并讨论这些资源对于形成我们现代生活意义的参考价值；另一方面，可以结合人工智能对于人类发展空间的拓展的可能性，思考现代中国人应该如何理解人性和人的价值，从而评估中国人所看重的亲情、家庭、和睦、孝顺等德目在未来人类发展中的意义。在这样的互动中，中国的哲学家和科学家就可能为中国哲学的发展描绘出属于中国哲学自身的问题。

干春松，北京大学哲学系教授、儒学研究院副院长，中国孔子研究院“泰山学者”特聘专家，中华孔子学会常务副会长，国际儒学联合会理事，中国孔子基金会学术委员，2018—2019年博古睿学者。主要研究领域为儒家思想研究和近现代思想文化研究，主要著作有《制度化儒家及其解体》《制度儒学》《重回王道》《保教立国》《康有为与儒学的“新世”》。

---

[1] 安乐哲：《儒家角色伦理学》，山东人民出版社，2017年，第178页。

机器人能否循道而行

——道家哲学与人工智能

王蓉蓉/文 吴万伟/译

绪论：机器人能否循道而行

这个问题似乎有些怪异，不过，它与我们当今的生活息息相关，旨在将古代道家哲学引入对话中以应对技术带来的挑战。目前，最前沿的智能技术不仅仅存在于研究实验室，而已经渗透到我们生活的方方面面。很难说我们不是生活在人工智能时代，因为人工智能技术无所不在，影响巨大，已经交织在人们的日常生活中。该技术挑战了人之为人的基本观念，我们不妨考虑一下我们有多么依赖智能手机，考虑一下自动驾驶汽车如何重新塑造了社区，众多机器人如何影响了道德社会的含义，我们的社会价值观发生了多么大的变化。技术成为生死存亡的威胁了吗？我们的价值观、社会、法律大多以人为中心而展开，在当今的人工智能技术革命中，我们必须询问它将对所有这些方面产生什么影响。像对待人类历史上众多里程碑的设立与变革一样，人们不得不对人工智能加以更好的理解与把握。对此，道家哲学可提供重要和极富创造性的声音。本文从以下三个方面思考这些议题。

气之生命和超越：道家之人与人工智能。

对真的终极追求：人工智能技术的道家伦理框架。

阴阳智慧：人工智能可从道家哲学中学到什么？

作为一种思想实验，本文也在测试道家哲学与实践能否为当今技术的飞跃发展做出独特和显著的贡献。

气之生命和超越

为了理解道家关于人的观点，我们需要首先谈论中国传统中“气”的概念。“气”是中国思想史上最重要、最受重视、使用最广泛的概念之一。作为中国古代思想各流派都用的一个概念，“气”往往被认为是动态的、无所不在的、穿透一切的、改变一切的力量，能给宇宙中的一切带来活力。虽然“气”是抽象概念，但它也是人的认识、体验和存在常见的与不可分割的组成部分。它被编织在语言中，成为人们呼吸的空气、推动血液融合的力量、吃的食物、心智的力量、思想的流动和内心最深处的冲动。气是生命的构造和力量。正如《庄子》中所说，“人之生，气之聚也；聚则为生，散则为死”[1]（《庄子·知北游》）。“气”弱，人就生病；“气”散，人就死了。

若进一步分析，“气”是不同能量的复杂混合体，每个能量都激活、控制人的生命和身体的不同方面。我们可以从《淮南子》中读到：“今人之所以眭然能视，萺然能听，形体能抗，而百节可屈伸，察能分白黑、视丑美，而知能别同异、明是非者，何也？气为之充，而神为之使也。”（《淮南子·原道训》）人眼能看，耳能听声，形体能承受重力，关节能伸屈，并能辨察黑白美丑，智慧理性能辨别是非异同，为什么呢？是在于气血充满着形体，精神发挥着作用。[2]

“气”的首要性在于它有自我产生和自我运行的力量。在作为万物之源和宇宙最基本材料的意义上，气就是道。“道”物化为“气”，因此占有空间和时间。事实上，在《管子》《庄子》的部分篇章以及宋明理学的很多文献中，“道”和“气”是相互替换的。

“气”的脉动和节奏产生了万物。作为自然的力量，“气”从内部指导、塑造和引领自然过程。

《黄帝内经》将身体和精神实体的特定状态归结为“气”。“气”被储存在人体五脏中。“气”在这些脏器中的数量和质量以及转化（气移动以及与其他器官和外部世界互动的方式）决定了人的生理与心理状态。从因果关系上说，“气”对我们的生理与心理状态起着至关重要的作用。

毕竟，生命是“气”流动的方式。中国历史上的道家实践者，也可称为“气”工程师，能够将“气”的各种流动引入人体的12条流通控制系统，也就是众所皆知的人体12条经络。[3]在中医实践中，针灸、按摩和天然药材都可以增加或者减少乃至阻碍“气”的流动。“气”可以通过当今技术测量和量化。[4]

或许我们可以假设，计算机算法为“气”的一种流动形式。我们不妨想象一下组织和安排数据的电脑算法就是气输入以变成输出的过程。[5]最近，人工智能已经显示出其在象棋和围棋比赛中战胜人类棋手的能力，因为编程和强大的运算能力使其在比赛中能进行数十亿次的运算以找到最好的可能步骤。人工智能能分析步骤的因果关系，记住以前比赛中这一步骤的后果，并决定这一步骤能否击败人类对手。

随着建立在原理逻辑基础上的传统人工智能的快速发展，今天已经进入网络连接的人工智能可以根据反馈环做出适当调整。人们普遍认为人脑是生物化学过程，学习发生在神经网络中，通过修改自身编码发现输入和输出之间的连接。如果人和决策过程扎根于信息与大数据，它们就可以将大脑的这些功能嵌入机器模式中。这样一来，人工智能就在电脑中创造出一种生命模式，使人类智能从以蛋白质为基础的人脑转移到像电脑网络那样更加持久的媒介中。

更有意思的是，认知心理学认为人脑就是一台机器，从诞生到其复杂行为，都是众多简单回应的综合结果。道家也有类似的观点，将人脑本身化为“气”的流动网络，从中诞生复杂的行为，这正是“气”回应的综合结果。[6]

在“围棋之道”中，郭庭解释说“人工智能之父”艾伦·图灵曾打算制造一台智能运算机器，说它更像人是因为它能像人那样思考。”[7]由于这个领域的突飞猛进，在很多具体领域，人工智能在功能上已经战胜人类，因为人工智能软件能模仿人的大脑，可以“雕刻”深度学习的神经网络。这种“深度学习”旨在让人工智能成为知识探索者。人工智能或者机器能够像“道”一样流动，因为它在反馈环内进行自我调节，实现道家行为的基本要求，成为依据“道”的最高价值观采取行动的自我调节、自我控制和自我管理的设备。[8]

自我管理的反馈控制系统是人工智能技术的一部分。按照卢克·多梅尔的说法，“一个智能设备应该能够感受到环境，这将促使它辨认出特定状态，而这将引发评估，评估则促成行动，从而进入持续不断的循环过程。智能设备的智能部分是处于如何处理感受到的信息与如何使用信息以便选择准备采取的行动之间的部件”[9]。人工智能带来的建议是，人类大脑中最神秘的东西可以被理解和复制到机器中，该机器拥有像人一样解决问题的能力和信息处理过程。[10]

人工智能或者机器能够像“道”一样流动，因为它在反馈环内进行自我调节，实现道家行为的基本要求，成为依据“道”的最高价值观采取行动的自我调节、自我控制和自我管理的设备。

同样地，行动是人或机器与环境之间的相互调解行为。这个原则在道家学说和控制论神经机械学运动中都能看到。道家哲学强调，感知和反馈是将性能最优化的方式。既然存在的生命要求处理从周围环境中获得的源源不断的视觉、听觉和触觉信息流，那么一个设备或者机器也能表现出看起来类似生命那样的协调行为。

人的生命本身是非线性的、动态的、自我组织的复杂系统，是从多层次组织起来的信息交换能量以便维持生存和发展。

但是，我们应该挑战地提问：生物机体在什么方式下并不吻合算法对有机体行为的事实？常规算法遵循从输入到输出的线性道路。有时候，电脑程序的不同部分之间存在着互动融合以便产生效果，但是人体要复杂得多。除了对付很多器官（如电脑程序）外，人体也不停地对外部世界做出回应。当然，人体拥有像算法那样的器官，然而，称整个人体是个算法就过于简单化和还原化了。还原论与道家关于人的观点是格格不入的。

更具体的是，道家哲学认为人类生命不仅仅是简单的信息、数据和网络的组合体。人的生命本身是非线性的、动态的、自我组织的复杂系统，是从多层次组织起来的信息交换能量以便维持生存和发展。人脑不仅做出符合逻辑的决策（视为算法），而且做出情感的复杂运算，这是机器所做不到的。

让我们来看早期道家《列子》中的一个故事。

> 杨朱之友曰季梁。季梁得病，七日大渐。其子环而泣之，请医。季梁谓杨朱曰："吾子不肖如此之甚，汝奚不为我歌以晓之？"杨朱歌曰："天其弗识，人胡能觉？匪祐自天，弗孽由人。我乎汝乎！其弗知乎！医乎巫乎！其知之乎？"其子弗晓，终谒三医。一曰矫氏，二曰俞氏，三曰卢氏，诊其所疾。矫氏谓季梁曰："汝寒温不节，虚实失度，病由饥饱色欲，精虑烦散，非天非鬼。虽渐，可攻也。"季梁曰："众医也。亟屏之！"俞氏曰："女始则胎气不足，乳湩有余。病非一朝一夕之故，其所由来渐矣，弗可已也。"季梁曰："良医也。且食之！"卢氏曰："汝疾不由天，亦不由人，亦不由鬼。禀生受形，既有制之者矣，亦有知之者矣。药石其如汝何？"季梁曰："神医也。重贶遣之！"俄而季梁之疾自瘳。[11]（《列子·力命》）

我们该如何理解这个寓言故事呢？首先，“众医”的诊断是对普通人生命的准确描述：饮食不注意和生活方式的不当造成了一系列症状。其次，“良医”的诊断看到了人生命中的相互依赖性，从基因背景下考察人类生命的奥秘。事实上，即使到了今天，我们仍然被告知某些疾病可能是人体基因中某个DNA缺陷造成的。先进的人工智能可以有效承担起众医和良医的使命，体现传统的医药实践和遗传理论。

前两种医生的诊断都在人体层次上，“神医”则与他们不同，“神医”的诊断指向生命的制者。卢氏说，人的生命中有一个“神”（精神、力量、权力），是生命的控制者和管理者。如果你修养这个“神”，不用吃药就可以痊愈。人的存在附有更厚重和更深刻的内容，后人解释称“愚昧者之所惑，玄达者之所悟也”[12]。通过不懈努力修养神，并按照道家传统经典的养生实践最终将炼成内丹。

经典道家的这种人格理想在后来的道家文献、礼仪和实践中得到推崇与付诸实施。理解和分析人体的最重要方式之一，是通过区分三大要素：形、气、神。“形”指的是形状或者形式——身体看得见的物质形式，是生命的房屋和居所，是“道”之器。[13] “气”是生命中看不见的基础或者源头。“神”是人生命中的心理和精神方面，是管理生命的力量。[14]“气”聚，形成生命；“气”散，生命的形式就消失了。如果“神”令“气”纯洁清新，就会秩序良好，达成治的局面；如果“气”变得浑浊，“神”就会混乱。这种理解的关键点是“气”成为身体中“形”与“神”的协调者。心和身通过“气”统一起来或产生互动。人不可能仅仅是“气”的流动而已，“神”显示了“气”流动的更深层次内涵。

《太平经》表明了这个要点：“神者受之于天，精者受之于地，气者受之中和。相与共为一道。故神者乘气而行，精者居其中也，三者相助为治，故人欲寿者，乃当爱气、尊神、重精也。”[15]

“神”意味着精神、神圣，模糊和无法衡量，或者如玄法导致的事件。“神”常常与表示明亮和阐明深刻见解的“明”联系在一起。“明”字本是由日和月二字组成。“神”和“明”都具有世界或者人的特征，因此两者将人与宇宙联系起来，合起来构成人的精神和思想核心。“神明”在字面意思上指神灵具有的智慧，但也被用来指圣人的特征。[16]它指人能够得到的“像神一样的清晰眼光”。罗浩使用“神明”一词表示了不起的影响、神奇的效果、顿悟和辨别力。它解释了超过人类狭隘视野的存在方式，指向神秘体验、不可言喻、瞬间性和玄奥性。[17]只有通过长期修身和消除偏见，才能达到“神明”的境界。

这“神明”也与“气”密切相关。作为动词，“神明”在转换“气”的时候表现出两个正好相反的特质：浓缩和扩张。浓缩后，“气”变成“神”；扩张后，“气”变成“明”。“明”作为扩张是“阳”，“神”作为浓缩是“阴”。“神”经常表达的是“气”在浓缩或吸收地的精；“明”显示的是“气”在扩张或者表达天的精。正如《太一生水》中所说：“太一生水，水反辅太一，是以成天。天反辅太一，是以成地。天地复相辅也，是以成神明。神明复相辅也，是以成阴阳。”[18]

神有多层次的含义，可以指精神或超越此在或认知能力等。神可以栖息在身体内，这正是身体修炼的目标之一。宋朝时期坤道曹文逸的《灵源大道歌》将身体看作神的最佳居住地，她教导说，身体首先应该保持清净，这样神才愿意前来居住。[19]

虽然被翻译成“spirit”（精神），但“神”不能简单地等同于西方式灵魂。在道家文化中，人拥有精神或灵魂即魂和魄。在西方传统中，说起灵魂通常都要提到光明和光线。[20]在中国早期文化中，灵魂不是单一实体，而是两个相互关联的力量。

魂和魄在人体与意识中至关重要。在汉代医药文献中，魂魄居住在人体内，在身体生理机能中发挥着核心作用。[21] [22]《礼记·郊特牲》云：“魂气归于天，形魄归于地。故祭，求诸阴阳之义也。”人活着的时候，魂魄是在身体内统一的两大生命力量；人死后，它们就分开，离开身体到别的地方去了。此即是《楚辞·招魂》中所说的“魂魄离散”。作为气的魂——灵上天，而作为身体的魄——灵则进入地下。因此，被称为“复”的丧葬仪式就是尝试“招魂复魄”。[23]

魂和魄在修身活动中极为重要，因为不适当的行动可能导致它们离开身体。人们必须避免这种“丢魂失魄”的情况发生，这个词甚至在当今仍然被经常使用。

在更深层次上，机器不能顺道而行。大道之行依靠神流动的潜能——但是，人工智能再生精神吗？人工智能是一种被制造出来的没有生命力的机器，依靠编程完成类似人的工作。人工智能的核心目标是设计能够辨认出和理解人的意识的电脑，而人工智能实际上是否展示人的意识是另一个问题。我们很难想象把“神”上传至电脑，因为“神”不是物体、算法、计算、软件、程序，而是镶嵌在身体内的转换和社会的存在。“神”一直在流动、旋转和转型。正如庄子所说，庖丁解牛时游刃有余，“官知止而神欲行”[24]（《庄子·养生主》）。

道家对人的理解的另一个特点是，人能获得和丧失自我意识。如果人工智能开发机器的自我意识，那么它也能如道家指出的那样失去自我意识吗？有如庄子曾经描述的境界：当一个人沉溺于手头工作的时候，他就可以忘掉自己的存在，让自己被物携带着走，心于是可以自由飞翔，“乘物以游心”[25]。

庄子详细描述了这一过程：“臣将为鐻，未尝敢以耗气也，必齐以静心。齐三曰，而不敢怀庆赏爵禄；齐五日，不敢怀非誉巧拙；齐

七日，辄然忘吾有四枝形体也。当是时也，无公朝，其巧专而外骨消。”[26]（《庄子·达生》）

就像自我意识的暂时忘却与获得，诸如创造性和社交智能等人类的品质是不会轻易被人工智能替代的。道家赞赏和培育人的这些技能。除此之外，它还补充了另一个特征，即“神”。正是“神”让人与人工智能、物和其他所有东西区分开来。“神”扎根于社会结构，而且与宇宙联系起来。在这方面，显然，亚马逊公司的交互式人工智能机器人Alexa和苹果公司的智能语音助手Siri都是力所不及的。毕竟，人工智能不可能在总体能力上超越人类。

对真的终极追求：人工智能技术的道家伦理框架

有些人工智能专家相信，下一代人工智能或者通用人工智能将几乎能够处理人类的所有事务。虽然目前关于通用人工智能还没有形成统一定义，但专家们认为，它应该有推理能力，能运用策略解决难题，在不确定的情况下做出判断，能传播知识、制订规划、学习、用自然语言交流，以及为实现共同目标而融合所有这些技能。[27]

道家对人工智能的这种重大发展和革命运动感到担忧吗？

从表层来看，人们或许会得出结论，道家秉持自然无为的理论，向往小国寡民的状态，因而会反对人工智能的发展。人工智能的发展代表了人类有目的的活动，这会让世界变得更为复杂。另一个论证是，人工智能最主要关心的是给机器以理性思考能力，是对庄子的诗意栖居的终极偏离。人工智能代表了最精巧形式的算计和思考。

如果对道家学说做更深层次的理解，我们将发现，由于道家对世界的深刻把握，它可能支持和认可人工智能的发展。当然，它也对人工智能提出严峻的挑战。

首先，道家学说为我们提供了深厚的理论根据，以面对大千世界的变化无常。庄子告诫说：“适来，夫子时也；适去，夫子顺也。安时而处顺，哀乐不能入也。”（《庄子·养生主》）“安时而处顺”事实上就是“顺着时间走”。换句话说，时间指的是持续不断的变化和转变。因为上天的时间不停地在创造和再创造，人们以同样的方式持续不断地进行创造活动。持续不断的变化和创新要求人们适时而变，重新确定自己在时间流动以及宇宙转变中的位置。

技术在起初引进时常常让人感到恐惧和担忧。因其改变现状的能力，人们担忧它可能改变或者毁掉自己现存的生活。然而，老子告诉我们，世间万物都是“道生之，德畜之，物形之，势成之”（《道德经》第五十一章）。“势”是万物实现的关键。“势”表征着万物创生的特定过程和发展的不同阶段。通过适应特定的环境，万物终于得以创生。通过领会基本的道家原则和世界观，智者应该能放下焦虑并容纳和接受新事物。道的开放精神引导人们面对人工智能的快速发展时不至于惊慌失措。

从道家视角看，人工智能的终极目标应该是建造能够模仿人类智慧的机器，以便让人更加接近道和大自然。道家将不会对人工智能感到害怕，因为从整体上看，它不可能超越人类，只是在某个可编程的方面超越人类。顺道而行，即人的各个方面都应该与道保持一致。

比如，智能手环卓棒建造了个人化的数据集，其中包括与你的身份、简介、生物数据、年龄、身高、体重、性别、食物偏好、情绪、活动、热量消耗和睡眠质量等相关的信息。但是，经过一段时间之后，这些设备构建了你幸福的个人化数据集。通过人工智能分析这些数据，可以阐明具体背景下的含义。可佩戴的设备除了采取外部测量外，如空气质量和噪声水平，还将不停歇地监督你的心率、血氧含量、身体活动、呼吸模式、面部表情、肺功能、语声变调、脑电波、姿势、睡眠质量等。使用人工智能，这些数据集不仅被转变成有关你的整个生命的普遍化信息，而且变成可采取行动的见解，能够根据具

体时刻提出改善健康的措施。进行这样的预测和诊断，我们将以此了解到特定疾病的必要条件是什么，或者可能出现的状况，可采取预防性措施来改善人的生命质量。

如果智能手环设备或者此类东西旨在作为计量生物学数据为我们服务，能够让人更健康和更幸福，有什么不好呢？道家应该会欢迎这些新设备。事实上，道家以善于养生而闻名遐迩，在日常生活规律和选择性饮食摄入等方面提出了具体的指导原则。[28]这些原则包括什么时候上床睡觉，什么时候起床，什么时候吃饭，吃什么，如何最好地管理好人们的日常生物钟，等等。道家使用针灸和中草药改善生活质量而延年益寿。人工智能可最优化从前无法测量的内容，计量生物学数据提供了改善生命的机会。这些设备若遵循道家的饮食建议，不仅能让人的生活变得更容易、更高效、更舒服、更方便，而且还能持续进步，不断开发出超出正常人能力的各种设备，最终可能找到永生之法（这与长寿仙丹的炼金术何其相似）。寻找长生不老的仙丹让道家实践者使用各种工具或技术延长人的寿命。

另外，道家学说能给予人工智以伦理见解吗？道家的伦理框架如何？在广泛和抽象的层次上，道家学说坚持不干涉自然之道的技术，遵循事物的自然节奏和发展。随着人工智能的智慧能力越来越高，人们对其道德水平的担忧和要求，以及在法律上进行责任追究的呼声也越来越高。

就人工智能而言，道家所能提出的道德和伦理问题是：人工智能是引导人们更加接近“道”，还是简单地满足人的欲望而已？

海德格尔认为，技术让人偏离我们的真正本性，那是一种从自身存在的逃避。[29]人工智能的大部分是在应对如何设计出能够智慧行动的人工代理人。机器人之所以让人的生活变得更复杂，是因为它们可能取代人类理性而成为宇宙中最高的智能。另外，互联网使我们有机会在一秒钟之内获得一辈子都难以获得的大量信息。

人工智能究竟给人类带来了什么样的挑战

为了回答这个问题，道家学说区分了天然智慧和人工愚蠢。在《道德经》第五章中，老子指出“虚其心，实其腹”，认为人们不应该受到社会建构的规范标准的引导。技术可能被滥用，可能违反人性，把我们推向远离天生本性之地。庄子提出了类似观点，区分了人性和天性，告诫我们不要用人性来干扰天性[30]。

人类的终极追求应该是求真，而不是满足人们的欲望。道家学说明确区分了求真和求欲的不同。《道德经》中说：“致虚极，守静笃。万物并作，吾以观复。夫物芸芸，各复归其根。”（《道德经》第十六章）天下万物纷纷扰扰，都沿着它们的初始状态发展变化，与天道法则融为一体，所以我们要把心灵放空到极致，让心灵安住在虚静之中。道家警告人类避免追求人们自己创造出来的巧智。[31]道家以求真为终极目标，从而获得真正的人性。

庄子赞赏特别的道德人，即真人。这是人的最高境界。黄帝曰：“余闻上古有真人者，提挈天地，把握阴阳，呼吸精气，独立守神，肌肉若一，故能寿敝天地，无有终时，此其道生。”（《黄帝内经·素问·上古天真论》）

《庄子》的真人概念指的是一个完全遵循道自然而行的人。《庄子》还提供了解决问题、模式识别、信息处理，以及能够在众多领域取代人的方法、模式、隐喻等。正因如此，“真人”有能力循道而行，其他人则做不到这一点。能“循道”而为，即意味着将自己的行为模式纳入“道法自然”之中。

遵循自然之理使人成为真人的说法，是我们在《庄子》的“循天”“循道”“遵循天性”等若干术语中首次看到的思想。笼统地说，自然本身或者自然的基础具有特定的规范之理，如果我们依据这些理来矫正自己的行为（最常见的情况是摆脱自己的偏见等），我们的行为就能更加高效。

在《庄子·外篇》中，是以“循天之理”讨论这个想法的，并将其与德联系在一起。如果生活中实际按照（或者推崇）宇宙的自然之理行动，那么人类繁荣就是它要“遵循”的道的组成部分。

“自然”这一概念最明显地表现出节奏性顺序，隐含着生生不息的存在机理。该短语由“自”和“然”二字组成，前者的字面意思是自我，后者的意思是如此这般。自然常常被译为“自发性”或“自然性”，但它指的是无自而然，没有外力推动或胁迫。

“自然”是“道”的法则。《道德经》第二十五章说：“人法地，地法天，天法道，道法自然。”人取法地，但地取法什么呢？地取法天，天取法什么呢？天取法“道”，道取法什么呢？道纯任自然。“自然”不是一个模式，而是单纯的自发性。最终，道法自然，回归完成自发性循环。

“自然”不仅是世界的一个元素，而且是人可采取的行动的最有效模式。人工智能不可能把握自然。

“自然”是人行为的最高形式，没有外力推动或权力胁迫人做任何事。如何区分自然和任意性呢？对自然的信心建立在对阴阳的信心之上。道本身是生生不息的力量（作为阴阳），因此人应该依靠这个内在力量自然运行。自然而然，按照自然和原始的状态，就像天地所处的状态一样（这就是自然）。

“自然”意味着承认世界运行的不确定性，以及人对这种不确定性的适应与把握。这也就是《道德经》第十章说的“生而不有，为而不恃，长而不宰”式的“玄德”。[32]我们不能确定无疑地知道明天是不是晴天，但是，我们能够准备好我们做什么，若晴天，就去海边，或者干脆不出门（待在图书馆）。自然并不根据我们可辨认出的理性计划运行，因此只有一个笼统的节奏让人遵从，并依据各种变化做出调整。针对这个问题的认识，《庄子·天下》中说：“在己无

居，形物自著，其动若水，其静若镜，其应若响。芴乎若亡，寂乎若清。”[33]

适应的能动性：阴阳智慧对人工智能的启迪

中国思想和文化很早就蕴含着现在称为人工智能核心的二元论逻辑。

这种二元论逻辑最早来自占卜实践。在《诗经》中，“日有食之，亦孔之丑。……日月告凶，不用其行”[34]（《诗经·小雅·十月之交》）。商朝时期，众多日常活动都要事先占卜。问题是：用“日月告凶”[35]来决定打猎、结婚或开战等是好方法吗？这也许不符合现代科学的理念，然而在这一过程中蕴含着“是”与“否”的二元选择。

这种二元论结构明显地呈现在《周易》中，该书依靠两种形式的占卜。一种是用龟甲“卜”。占卜者在龟甲上面做标记，然后将其放入火中烧，根据出现的裂纹形状预测吉凶福祸。另一种是用蓍草筮。占卜者手持蓍草，根据策（占卜用的蓍草）确定其数目以定吉凶。裂纹或者蓍草数目指向两种可能的结果：吉凶、对错、行停、成败、得失。这些占卜方法的二元论结构揭示了一种深入阴阳思维最深处的价值观和信仰体系。埃德蒙·赖登写道：“二元论术语是中国哲学早期的特征，事实上或许是中国语言的特征。”[36]

自古以来就存在于中国思想和文化中的二元论思维模式可以被视为一种阴阳智慧。[37]或许这正是人工智能可以从中国思想和文化中学习的东西。阴阳术语是从命名事物与太阳的关系的方法逐步演变为一套复杂的思维体系的，这个思想体系可以称为“阴阳思想”。从历史上看，虽然阴阳术语在汉代以前不是主导性概念，但原生的阴阳思想已经存在，而且产生了一种特定的思维模式。有关阴阳的所有早期

记载编织成生生不息的现实存在，正是阴阳概念让这些存在的机理连贯起来并最终变得可以管理。阴阳思想将众多关系简单化、形象化和扩大化。它通过从边界内外塑造和指导自然程序，与可渗透的信息、能量或影响配合互动支撑了所有存在形式。

阴阳智慧扎根于将宇宙视为生生不息的系统的有机体概念。自我组织和自我稳定以系统与环境的相互交往为前提条件。

阴阳智慧扎根于将宇宙视为生生不息的系统的有机体概念。自我组织和自我稳定以系统与环境的相互交往为前提条件。

就拿御马艺术作为例子，[38]有效互动被视为通过运动而出现。直接影响马拉战车的因素包括马性及其力量大小、地带、天气状况和个人目的。骑手也受很多外部因素的影响，他或她的内心状态受到干扰，从而引发不同种类的反应。那是应对环境变化和从内部对付变化的开放性系统。

《淮南子》中说，你能感觉到手的变化，不过是靠心做出回应。这就是常言所说的“得心应手”，“不徐不疾，得之于手而应于心”（《庄子·天道》）。关于这方面，在骑手看来，世界不是用来观察而是用来感受的。整个感觉体系（包括视角）是一种感受反应。就像感受到阳光的温暖或者微风吹来，观看的行为有类似的感受，因为它涉及身心一起行动。《列子》提供了驾车艺术的另外一种描述。

> 凡所御者，亦如此也。曩汝之行，得之于足，应之于心。推于御也，齐辑乎礣衔之际，而急缓乎唇吻之和，正度乎胸臆之中，而执节乎掌握之间。内得于中心，而外合于马志，是故能进退履绳而旋曲中规矩，取道致远而气力有余，诚得其术也。[39]（《列子·汤问》）

在御马时，人能够区分内部因素（内心）和外部因素（马），但是关键点是能直达中心。这种阴阳智慧是协调和化身的智慧。在此意义上，阴阳不是让一物与另一物配搭，而是在任何场景中使隐蔽的力

量共同发挥作用，娴熟互动带来理想的结果。人工智能可以从该途径和方法中学到深层的阴阳智慧。

阴阳智慧的第二个方面建立在适应性的基础上。人工智能算法可以从过去的错误中学习，的确有一定的适应性，但是它们不能更广泛地适应动态环境或其他范式。比如，下围棋的人工智能机器能够在标准的比赛中获胜，但如果改变其中一条规则，它还能取胜吗？如果让它下象棋或者参与其他游戏，人工智能还能取胜吗？

人类技能的培养是掌握适应性的艺术：专心致志但并不焦虑，认真投入但并不痴迷，做出努力但并不强制，时刻准备好应对变化无常的意料之外的情况。骑马就是在对付外在力量和内在限制并产生适应性的自我组织过程。从这一点看，适应性与稳定性或者和谐并不是同义词，反而更加接近于应对实际环境变化的功能性效率。更接近的是效率而不是和谐。比如，如果骑手想快点，而马走得慢，该怎么办？骑手需要使用让马加速前进的方法。骑手必须向后拉紧缰绳，动态的阴阳变化就出现了：与马的力量产生互动，给与取，推与拉。

在这里，列子认为关键是把握“均”。“均，天下之至理也，连于形物亦然。均发均县，轻重而发绝，发不均也。均也，其绝也，莫绝。人以为不然，自有知其然者也。”[40]（《列子·汤问》）“均”是什么？它是能够转向任何方向的车轮的核心。骑手只能通过与它一起合作达成自己的目标，协调所有变量以便获得所追求的结果。“均，天下之至理也。”[41]通过“均”，骑手将周围因素纳入自己的感知－回应循环中以便维持最高的效率。

阴阳智慧作为一种适应性，是人与世界互动时不可或缺的工具。我们能够区分形成阴阳智慧的若干因素，它们可能有助于推动人工智能的进一步发展。

阴阳互动涉及若干前提：（1）多元关系依靠融合程度而定；（2）秩序不是事先确定的；（3）千变万化是万物的本质；（4）人不

能充分预测将来。

为了更好地理解阴阳智慧的复杂性和多样性，我们或许能够提出如下6种形式的不同关系：矛盾——矛盾对立，相依——相互依赖，互含——相互包含，交感——互动共鸣，互补——相互补充和支持，转化——变革、转型与深化。[42]

这些思想可以用“环流”模式作为产生、融合和应急过程的隐喻。汉字“流”的字面意思是水的流动，左边的三点水就包含了水的形象。汉字“环”指的是流动方式。因此，我们也可以将其翻译成“流动的环”。

环流有三个特征。第一个特征是围绕空中心旋转。《道德经》第十一章中说：“三十辐共一毂，当其无，有车之用。埏埴以为器，当其无，有器之用。凿户牖以为室，当其无，有室之用。故有之以为利，无之以为用。”毂、器皿、门、窗等的转动，都是建立在“中空”的结构基础上的，所有这些都是边框中空空如也的空间。“当其无，有器之用”。这种“结构性蓝图”也可以在旋风、旋涡、溶坑、线圈等自然现象中瞥见。

环流的第二个特征是循环没有终结，也没有开端，没有边界，无穷无尽地继续运行。改变是永恒的，永不停歇。一边变成另一边，随后再变。这一观点出现在众多不同的文献中。如《黄帝内经》中说：“其清者为营，浊者为卫，营在脉中，卫在脉外，营周不休，五十度而复大会，阴阳相贯，如环无端。”[43]（《黄帝内经·灵枢》）

环流的第三个特征是反[44]，这也是《道德经》中经常出现的主题。反求助于运动的圆圈的形象或者更准确地说，是非线性的无穷维旋转运动。很可能出现在战国时期的道家文献《鹖冠子》花费整整一章的篇幅论述这种运动的重要特征。[45]“气故相利相害也，类故相成相败也。……美恶相饰，命曰复周，物极则反，命曰环流。”[46]

（《鹖冠子·环流》）这种环流试图传递一种观念，即事物的出现不是线性的，环流朝向关键节点，然后再折返。

结语

中国第一个机器人出现在道家早期经典《列子》中（公元前4世纪）。[47]匠人献给周穆王一个制作的歌舞艺人，摸它的头它就唱歌，抬起它的双手它就会跳舞。国王感到既震惊又愤怒。匠人吓得半死，立刻把歌舞艺人拆散，展示给穆王看，原来整个儿都是人造。穆王深为惊奇，叹曰：“人之巧乃可与造化者同功乎？”（《列子·汤问》）

今天，技术比人类历史上任何时期都更广泛和更容易获得。道家思想认为，不可预测性和变化的不可避免性实际上支配了我们的日常生活。它帮助我们挑战秩序和稳定性的线性因果关系思维。变化不是简单控制以从一个稳定状态转向另一个状态，变化是生命的组成部分，我们不应该回避技术变化的现实。

道家思想认为，不可预测性和变化的不可避免性实际上支配了我们的日常生活。它帮助我们挑战秩序和稳定性的线性因果关系思维。

庄子在《吕氏春秋·孝行览·必己》中说：“成则毁，大则衰，廉则锉，尊则亏，直则骫，合则离，爱则隳，多智则谋，不肖则欺，胡可得而必？”因此，任何事情都不是只有成功或者只有失败，因为这两者也向对方开放，相互构成统一的整体。可以预测，人类是否被人工智能淘汰的问题建立在这样的基础上，我们能想象出人与机器的新型社会关系，因而要开发出更富哲学性的理解。

对道家来说，幸福生活并不依靠事先确定好的形式行动，也不遵照事先确定好的等级秩序，而是循道而为。没有永恒的答案，只有在特定条件下合适或不合适的答案。所以，道家旨在培养能够满足身心

清晰性和灵活性的品质。道家实践就是学会自然和与毁，要学习老子的三宝：慈、俭、不敢为天下先。清虚自守，齐物而侍。如果人工智能和机器人指向这种生活方式，自然可以得到道家的认可。

道家既不拒绝也不全盘接受技术的突飞发展。基于道的核心价值观，道家既不会屈从于任何形式的技术发展带来的变化，也不会无条件拒绝任何形式的技术进步。对道家来说，问题是技术发展在多大程度上有助于人们依靠练习自然加深对道的认识。如果人工智能的发展以追逐利润为动机和前提，如果人工智能倾向于将人类淘汰或导致人的异化，或者人与自然世界相互疏远，那么它就与道家践行的思想格格不入了。

但这并不是说人工智能和机器人只存在这种可能性。无论是对自我、他人、自然还是宇宙而言，如果这种技术能够被重新调整而指向自然，或者为人类迈进自然提供方便，道家肯定会全力支持。

智慧、仁爱等人类美德能够被模拟，但不能靠非生物系统的方式复制。人的大脑不仅仅是个根据二元布尔逻辑进行操作的计算机器。镶嵌在生物系统中的有模拟和二元过程，有器官、肌理、血流、新陈代谢和感知运动等多项复杂功能。人的生物系统是自然和宇宙中多样能量领域的组成部分。如果我们认为世界是天地之气影响的结果，那么很清楚的是，人的大脑并不是一台有重新启动按钮的机器。它是过程而非物体的一部分。“复制整个‘系统’将是惊人的辉煌成就。拥有通用人工智能的机器人将不会尝试控制人类，它可能在早上起来鞠躬向人类主人打招呼，以此表示谦恭的身份和对创造者表示敬意。”

王蓉蓉，洛杉矶罗耀拉大学哲学教授，2016—2017年博古睿学者。她的教学与研究领域包括东西方比较哲学、道家思想、比较伦理学等。著作有《阴阳：中国思想与文化中的阴阳之道》《全球化处境中的中国哲学》《从先秦到宋代中国思想文化中的女性形象》《理性与洞见：东西方对伦理智慧的追求》。她是国

际世界哲学理事会成员，曾担任全球亚洲与比较哲学协会会长（2016—2018年）。

参考文献

1. Barry Allen, A Dao of Technology?, Dao 9, no. 2 (2010), pp.151-160.

2. Roger Ames, David Hall (trans.), Dao De Jing: A Philosophical Translation,Ballantine Books, 2003.

3. Roger Ames(trans.), Yuan Dao, Original Dao: Trace to Its Root, Ballantine Books,1998.

4. Kenneth E Brashier, Han Thanatology and the Division of ‘Souls’, Early China,21, 1996, pp.125－58.

5. Luke Dormehl, Thinking Machines: The Quest for Artificial Intelligence and Where It’s Taking Us Next, Tarcherperigee, 2017.

6. C-Y Cheng. Preface: Science, Technology, and Chinese Philosophy, Journal of Chinese Philosophy 29, 2002, pp.469-70.

7. Liya Ding, Xiaogan Liu, Human and Machine Intelligence —— Between Fuzzy Logic and Daoist Thought, Dan E. Tamir, Naphtali D. Rishe, and Abraham Kandel(ed.),Fifty Years of Fuzzy Logic and its Applications, Springer, 2015.

8. Stephen Eskidsen, Daoism, Meditation and the Wonders of Serenity, SUNY,2015.

9. Andrew Feenberg, Technosystem: The Social Life of Reason, Harvard University Press, 2017.

10. Joseph Grange, Dao, Technology, and American Naturalism, Philosophy East & West 50, 2001, pp.363-377.

11. Angus C. Graham(trans.). Chuang-Tzu: The Inner Chapters, Hackett Publishing,2001; The Book of Lieh-tzu: A Classic of The Tao, Columbia University Press, 1990.

12. Henryk Grossmann, Descartes and the Social Origins of the Mechanistic Conceptof the World, Gideon Freudenthal and Peter McLaughlin(ed.), Peter McLaughlin(trans.), The Social and Economic Roots of the Scientific Revolution:Texts by Boris Hessen and Henryk Grossmann, Springer, 2009, pp.157-229.

13. Ting Guo, Dao of the Go: Contextualizing 'Spirituality,' 'Intelligence,' and the Human Self, Implicit Religion 20, 2017, pp.233-244.

14. Livia Kohn(ed.), Daoism Handbook, Handbook of Oriental Studies, Section 4 China, 27 vols, Brill, 2000.

15. Thomas Michael, In the Shadows of the Dao: Laozi, the Sage, and the Daodejing.Albany, State University of New York Press, 2015.

16. Hans-Georg Moeller(trans.), Daodejing: A Complete Translation and Commentary,Open Court, 2007.

17. Eric Nelson, Technology and the Way: Buber, Heidegger, and Lao-Zhuang 'Daoism' Journal of Chinese Philosophy 41, 2014, pp.307-327.

18. Jeff Noonan,. Embodiment and the Meaning of Life, McGill-Queen' sUniversity Press, 2018.

19. Graham Parkes, Lao-Zhuang and Heidegger on Nature and Technology, Journal of Chinese Philosophy 39, 2012, pp.112-133.

20. Fabrizio Pregado, Zhouyi Can Tong Qi (The Seal of the Unity of the Three), Golden Elixir Press, 2011.

21. Harold David Roth, Original Tao: Inward Training and the Foundations of Taoist Mysticism, Columbia UP, 1999.Jean-Paul Sartre, Being and Nothingness, Hazel E.Barnes(trans.)

22. Washington Square Press, 1993.

23. C Tang, Chang Tsai's Theory of Mind and Its Metaphysical Basis, Philosophy East and West,6, 1956, pp.113 - 36.

24. David V. Tansley, Subtle Body: Essence and Shadow, Thames and Hudson, 1977.

25. Toby Walsh,. Machines That Think: The Future of Artificial Intelligence,Prometheus Books, 2018.

26. Robin R Wang, The Virtuous Body at Work: The Ethical Life as Qi 氣 in Motion,Dao: Journal of Comparative Philosophy 9, 2010, pp.339-351.

27. Yinyang: The Way of Heaven and Earth in Chinese Thought and Culture, University Press, 2012.

28. Shih-Ying Yang, Robert Sternberg, Conceptions of Intelligence in Ancient Chinese Philosophy, Journal of Theoretical and Philosophical Psychology 17, 1997, pp.101119.

29. Ying-Shih Yu, O Soul, Come Back! A Study in the Changing Conceptions of the 287 Soul and Afterlife in Pre-

Buddhist China, Harvard Journal of Asiatic Studies 47, 1987, pp.363-95.

31. Brook Ziporyn(trans.), Zhuangzi: The Essential Writings with Selections from Traditional Commentaries, Hackett Publishing Company, Inc., 2009.

32.陈鼓应：《庄子今注今译》，中华书局，1983年。

33.陈撄宁：《灵源大道歌白话注解》，中国道教协会，1998年。

34. 刘钊：《郭店楚简校译》，福建人民出版社，2003年。

35. 牛兵占等：《中医经典通释：黄帝内经》，河北科学技术出版社，1994年。

36. 杨伯峻：《列子集释》，中华书局，1979年。

37. 杨寄林：《太平经今注今译》，河北人民出版社，2002年。

38. 张隐庵：《黄帝内经素问集注》，学苑出版社，2002年。

---

[1] Brook Ziporyn (trans.), Zhuangzi: The Essential Writings, Hackett Publishing Company, 2009, ch.22.

[2] Roger Ames (trans.), Yuan Dao, Original Dao: Trace to Its Root, Ballantine Books,1998, p.26; Harold Roth (trans.), The Huainanzi, A Guide to The Theory and Practice of Government in Early China, Columbia University Press, 2010, p.75. 该译文参考了安乐哲和罗浩的译本，稍有修改。

[3] Zhuangzi, ch.3, p.22.

[4] 引自当今有关气的科学研究成果。

[5] Luke Dormehl, Thinking Machines: The Quest for Artificial Intelligence and Where It's Taking Us Next, Tarcherperigee, 2017.

[6] 见《道德经》。

[7] Ting Guo, Dao of the Go: Contextualizing 'Spirituality,' 'Intelligence,' and the Human Self, Implicit Religion 20, no. 3, 2017, p.241.

[8] Robin R Wang, Yinyang: The Way of Heaven and Earth in Chinese Thought and Culture,Cambridge University Press, 2012.

[9] Luke Dormehl, Thinking Machines: The Quest for Artificial Intelligence and Where It's Taking Us Next, Tarcherperigee, 2017, p.62.

[10] Luke Dormehl, Thinking Machines: The Quest for Artificial Intelligence and Where It's Taking Us Next, Tarcherperigee, 2017, p. 65。

[11] 杨伯峻：《列子集释》，中华书局，1979年，第204—205页。

[12] 杨伯峻：《列子集释》，中华书局，1979年，第205页。

[13] Livia Kohn (ed.), Daoism Handbook, Brill, 2000, p.96.

[14] Roth, The Huainanzi, p.74.

[15] 杨寄林：《太平经今注今译》，河北人民出版社，2002年，第1730页。

[16] 在白瑞旭看来，“在前汉或者汉朝文献中，‘神明’可以指神圣的存在或者像神一般的智慧……”请参阅：K. E. Brashier, Han Thanatology and the Division of Souls, Early China, 21, 1996, p.149。

[17] Roth, The Huainanzi, p. 127.

[18] 刘钊：《郭店楚简校释》，福建人民出版社，2003年，第42页。

[19] 陈撄宁：《灵源大道歌白话注解》，中国道教协会，1998年，第10页。

[20] 坦斯利说，根据希腊传统，“灵魂是一个发光的身体，他们称之为‘augoeides’，意思是‘发光的形式’”。D. V. Tansley, Subtle Body: Essence and Shadow, Thames and Hudson, 1977, p. 6.

[21] K. E. Brashier, Han Thanatology and the Division of Souls, p. 145.

[22] K. E. Brashier, Han Thanatology and the Division of Souls，p. 363。

[23] 长沙马王堆汉墓中出土的一幅T字形帛画，象征着这种灵魂召唤，为复礼提供了考古学证据。

[24] Zhuangzi, p.22

[25] Zhuangzi, p.29.

[26] Zhuangzi, p.29.

[27] Toby Walsh, Machines That Think: The Future of Artificial Intelligence,Prometheus Books, 2018.

[28] See Livia Kohn and Huang Yongfeng' s work.

[29] Eric Nelson and Gram Park essay.

[30] Zhuangzi, p.106.Ames and David Hall (trans.), Dao De Jing: A Philosophical Translation32., Ballantine Books, 2003, p.99.

[31] 关于《道德经》中的“智”，见Michael 2015, chs. 8-9。

[32] Moeller, Daodejing, p. 121.

[33] Zhuangzi, p.123.

[34] Zhou, Shijing, p. 300.

[35] Zhou, Shijing, p. 300.

[36] E. Ryden, The Yellow Emperor’s Four Canons, A Literary Study and Edition of the Text from Mawangdui, Guangqi Press,1997, p. 11.

[37] Robin R. Wang, Yinyang: The Way of Heaven and Earth in Chinese Thought and Culture, Cambridge University press, 2012.

[38] 骑马的方式在汉语中被称为御，即驾驶战车的技术。《淮南子》将驾驶战车作为阴阳思想的应用。

[39] Graham (trans.), The Book of Lieh-tzu, p. 183.

[40] Graham (trans.), The Book of Lieh-tzu，p.104。

[41] Graham (trans.), The Book of Lieh-tzu，p.104。

[42] 更多讨论，请参阅《阴阳书》。

[43] 牛兵占等：《中医经典通释：黄帝内经》，河北科学技术出版社，1994年，第59页。283

[44] 关于“反”这个词的更加具体的讨论，请参阅：Wang Zhongjiang, Abnormalities and Return: An Exploration of the Concept Fan 反 in the Laozi, Religions10, 2019, p.32。

[45] 王心湛：《鹖冠子集解》，广益书局，1936年，第9页。“物极则反”这个说法是事物走向极端就会走向反面，在当今中国成为流行的习语。当代西方科学里有一个术语叫作“自我组织的临界点”，指一种倾向，庞大的耗散体系会驱使自己到达临界状态，长度和时间刻度的范围往往很广。这个观点提供了一个统一概念，说明系统中大规模行为的相当程度的自由。人们已经在多样化的领域内如地震结构、经济学和生物演化中寻找这个特征。它也被视为“回归均值”。

[46] 王心湛：《鹖冠子集解》，广益书局，1936年，第9页。“物极则反”这个说法是事物走向极端就会走向反面，在当今中国成为流行的习语。当代西方科学里有一个术语叫作“自我组织的临界点”，指一种倾向，庞大的耗散体系会驱使自己到达临界状态，长度和时间刻度的范围往往很广。这个观点提供了一个统一概念，说明系统中大规模行为的相当程度

的自由。人们已经在多样化的领域内如地震结构、经济学和生物演化中寻找这个特征。它也被视为“回归均值”。

[47] 周穆王西巡狩，越昆仑，不至弇山。反还，未及中国，道有献工人名偃师，穆王荐之，问曰：“若有何能？”偃师曰：“臣唯命所试。然臣已有所造，愿王先观之。”穆王曰：“日以俱来，吾与若俱观之。”越日偃师谒见王。王荐之，曰：“若与偕来者何人邪？”对曰：“臣之所造能倡者。”穆王惊视之，趋步俯仰，信人也。巧夫頷其颐，则歌合律；捧其手，则舞应节。千变万化，惟意所适。王以为实人也，与盛姬内御并观之。技将终，倡者瞬其目而招王之左右侍妾。王大怒，立欲诛偃师。偃师大慑，立剖散倡者以示王，皆傅会革、木、胶、漆、白、黑、丹、青之所为。王谛料之，内则肝胆、心肺、脾肾、肠胃，外则筋骨、支节、皮毛、齿发，皆假物也，而无不毕具者。合会复如初见。王试废其心，则口不能言；废其肝，则目不能视；废其肾，则足不能步。穆王始悦而叹曰：“人之巧乃可与造化者同功乎？”诏贰车载之以归。夫班输之云梯，墨翟之飞鸢，自谓能之极也。弟子东门贾、禽滑釐闻偃师之巧以告二子，二子终身不敢语艺，而时执规矩。（《列子·汤问》）

当人工智能遇见道教

盖菲

人工智能一般被分为三个阶层，即弱人工智能、强人工智能和超人工智能。弱人工智能指只能解决特定领域问题的人工智能，今天我们看到的所有人工智能算法和应用都属于弱人工智能的范畴，包括苹果智能语音助手Siri、AlphaGo等；强人工智能是可以胜任人类所有工作的人工智能；超人工智能是在科学创造力、智慧和社交能力等每一方面都比最强的人类大脑聪明很多的智能。很多人都认为只有像科幻电影里那种能和人完全交流的、认识到“自我”的（不管有没有类人的躯体）人工智能才是大众心中的人工智能，也就是强人工智能乃至超人工智能。如何看待这一类人工智能？本文将跳出技术层次的探讨，从道教哲学的角度进行一些浪漫且不成熟的臆想。

人的本质

何为人？人的本质是什么？从道教领域来说，人是“精气神”的统一体，也是“性命”的统一体。

人的本质问题，是所有求道之人皆会求索的问题，从道家道教的角度来梳理，最早给出答案的应该是庄子，庄子云：“人之生，气之聚也；聚则为生，散则为死。”[1]即人的本质为气。至两汉，形、气、神的观念已基本确立，如《淮南子·原道训》载：“夫形者，生之舍也；气者，生之充也；神者，生之制也。一失位则二者伤矣。”[2]“精”为“形”之精华，故《道经》中有云“形气神”，亦有云

“精气神”，如《高上玉皇心印妙经》云：“上药三品，神与气精。”[3]至唐宋时期，道教内丹学兴起，人为精气神的统一体之说成为道教理论的基石。元陈致虚《金丹大要》载：“一生二，二生三，三生万物。故虚化神，神化气，气化精，精化形，形乃成人。”[4]现在屡见于仙侠小说中的“炼精化气、炼气化神、炼神返虚”三大步骤，从五代起，逐渐成为道教内丹修炼的核心之论。

除了精气神之说，还有性命之说。修道当性命双修，“性命双修”是内丹术中的一个基本理念。性、命二者，命说少有争议之处，命指人禀气而生的身体、色身，也可以看作“精”与“气”的主体“形”。科幻电影《太空漫游》中有这样一句话：“不管是我们碳基人类还是硅基机器人，都没有本质的区别。我们中的每一员都应获得应有的尊重。”那么，我们都会得到尊重的本质在哪里？肉体？应该不是。大热门电影《银河护卫队》中的一个重要角色是一只浣熊，一只插科打诨、智商看起来比主角还高的浣熊，人气相当高，没有人认为这只浣熊和其他“人”具有不平等的地位，在这里，它是护卫队的一员，不管是荧幕内的世界还是荧幕外的我们，都认为这只浣熊和其他成员一样，是平等的一员。可见，肉体的区别并不是本质的区别。如果说人是“性”与“命”的统一体，“命”是肉体，不是本质问题，那么本质问题就一定是“性”。何为“性”？性说迷雾重重，有儒家之“穷理尽性”说、佛教之“性空”说，亦有道家之“性命双修”说，“性”是中国哲学中一个极为重要的观念。所谓的“修性”究竟修的是什么？是灵魂、修养，还是自我意识？

道教内丹诸家对“性命”之说皆有所论述，如南宗石泰云：“气是形中命，心为性内神。能知神气穴，即是得仙人。”[5]北宗王重阳云：“性者神也，命者气也。”[6]观诸多丹经，基本上都会将“性”与“神”等同起来，论其中区别，约是先天后天之分，性为先天之本性，神是后天之思绪，所以修性即是修神，这一点是得到公认的。

“性”与“命”不是互不相关的，道教的身心二元论与西方的不同，早期西方哲学中的“身心二元论”认为心灵和物质是独立存在的两个实体，它们之间没有相互作用，但这难以解释人的行为，故笛卡儿假设，身体和心灵有一个交接点，它是位于脑部的“松果腺”，但是“松果腺”的概念给予心灵以位置，这违反了心灵没有广延的二元论原则。这一点在中国则不同，道教理论中的“神”与“气”在本质上是相同的，并且是可以互相作用的，“神”藏于身体的不同部位，并非“没有位置”，如《素问·六节藏象论》云：“心者，生之本，神之变也；其华在面，其充在血脉，为阳中之太阳，通于夏气。肺者，气之本，魄之处也；其华在毛，其充在皮，为阳中之太阴，通于秋气。肾者，主蛰，封藏之本，精之处也；其华在发，其充在骨，为阴中之少阴，通于冬气。肝者，罢极之本，魂之居也；其华在爪，其充在筋，以生血气，其味酸，其色苍，此为阳中之少阳，通于春气。”[7]在道教理论中身与心、神与气是相通且可相互影响的，修“神”与修“气”是一体两面，所谓“性命双修”在这个意义上是说修“性”的同时可以影响“命”的修行，但不少道士认为“神”处于“气”之上，为主导的一面，神的状态决定形、气的状态，故勿怪明清内丹道更多地强调“修性”的重要性。《庄子·知北游》云：“夫昭昭生于冥冥，有伦生于无形。精神生于道，形本生于精，而万物以形相生。”[8]神比气更近乎道，修性更近似修道，而纯修命却是远于修道的。

精神体与灵魂在当代普适的概念中略为近似，皆为可独立于身体的存在。精神则与意识这个概念有微妙的相同之处。如果承认有独立于肉体的灵魂存在，那么精神、意识都是灵魂后天的表现；若不承认有灵魂的存在，则精神和意识多被看作大脑的一种功能产物。

灵魂与魂魄二者略有区别，“灵魂”一词在中国古代与“魂魄”意义近似，如《楚辞·九章·哀郢》云：“羌灵魂之欲归兮，何须臾而忘反。”[9]但此概念在当代受西方观念影响，灵魂被认为是浑然

圆融不可分割之物，而“魂魄”之说则不同，《说文解字》释魂魄二字云：“魂，阳气也。……魄，阴神也。”[10]可见魂魄其中具有魂与魄两部分，有阴阳属性之分。《礼记·郊特牲》曰：“魂气归于天，形魄归于地。故祭，求诸阴阳之义也。”[11]魂为阳，故上升至天，魄为阴，故下降于地，可见魂魄是可分的。“魂魄”隶属于“神”是毋庸置疑的，故修“魂魄”“精神”亦是修“性”。

以上述之言，性功是修“神”、修“心”、修“精神”，也是修“魂魄”，这里包含了三重不同的修行：一是对后天的“精神”的修行，这里的“精神”指人的意识、思维活动与心理状态；二是对精神体的修行，这里的“精神体”指相对于形骸而言的元神，或者说灵魂；三是对先天的“精神”的修行，这里的“精神”是指中国传统文化中的“精神”境界，以“天人合一”为最高标准。暂不提颇具争议的第二重，也不说难以达到的第三重，仅就第一重而言，可见人的意识、思维活动，或者说自我意识，正是人与人工智能平等的本质问题。

人的进化

宗教的本质是人的超越性，人超越有限性而达到无限的境界。道教也是如此，道教最本质的信仰即神仙信仰，最核心的追求即飞升成仙。神仙的种种也许是古人的臆想，强人工智能的种种是今人的臆想，所以以下种种对人工智能与神仙异同的探讨不过是又一种臆想。

神仙的第一个特征即长生不老。长生是人类永恒的追求，无论是哪个民族、哪个地区，从人类意识到“死”之一事开始，就必然渴望着“生”。最早人们认为长生不老的追求表现在求仙，寻仙人，并恳求仙人赐予“不死之药”，如《史记·秦始皇本纪》言“因使韩终、侯公、石生求仙人不死之药”[12]，《淮南子·览冥训》云“羿请不

死之药于西王母，姮娥窃以奔月，怅然有丧，无以续之”[13]。可见，此时的“不死之药”是仙人专属，普通人难得。很快这种求仙的思想就被打破了，人们开始立足于本身，希望通过种种方法得以长生。古人对长生不老的追求方法基本都集中在道教的种种方术之中，服食、导引、存思、房中、外丹、内丹种种尝试皆可在《道藏》中寻到。其中影响最大、最为人所笃信的两种方法，即外丹与内丹。

外丹即用五金八石形成不同的配比在鼎炉中炼出的金丹，相传服之即可长生成仙。外丹术起源于服食术，春秋战国时期的服食多是服用草木，所谓“不死之药”也多被形容成一种植物，如《史记·龟策列传》云：“伏灵者，千岁松根也，食之不死。”[14]但很快就有人对服食草木提出了异议，如《黄帝九鼎神丹经诀》云：“草木药埋之即朽，煮之即烂，烧之即焦，不能自生，焉能生人。”[15]他们认为草木实在过于柔弱了。放弃草木与其说是因为草木的易腐朽性，倒不如说是因为草木的“生命性”，其如人类一般有生、老、病、死，己所不度，安得度人？故神仙家服食的方向转向无生命而长久的金石。

曹植《赠白马王彪诗》云：“自顾非金石，咄嗟令人悲。”[16]如何才能让人如金石一般久存呢？比较容易想到的办法即将金石融于己身，最易行的方法就是“吃”，中国人自古即讲究吃啥补啥。外丹术即自此兴盛。如葛洪所言，“此盖假求于外物以自坚固”[17]，即希望人的身体能如金石般长久、不朽。从“缸中之脑”思维实验的角度出发，将人的大脑移植到一具合金身体上，使其完美结合，这就达到了道教早期外丹术的一个理想境界，身体如金石一般长久，但这就达到长生不老了吗？并没有，不说大脑本身的寿命，只说构成身体的金属，即使是再高级的合金，也有腐朽的一天。使身体如金石一般来追求长生不老实际上是难以成功的，随着冶炼术的发展，古人也逐渐意识到了这一点，所以南北朝时期，外丹术就少言使身体如金石，多云“易形换质”，即改变身体的本质，使其成为更高级的生物，告别人类的身份，成为一种新的种族——仙。

若以“长生不死”而言，人工智能有寿命的限制吗？当前多是使用逻辑与符号系统实现人工智能。那么，数学模型有实体吗？可以说有，因为建模必须通过计算机实现；也可以说没有，因为数学模型是数据，可以在这台计算机上出现，也可以在另一台计算机上出现，还可以存储在网络上。让我们抛开现实继续探讨，毕竟正是因为强人工智能乃至超人工智能还没有实现，所以如同笔者一样的哲学家才敢于探讨，至于现实中的数据模型，这对不精通数学的当代哲学家来说太难了。人工智能的核心模型是数据，从理论上来说，如果没有被故意摧毁，没有天灾人祸，数据是不老不死的。从长生不老的角度来说，人工智能已经达到了道教人士几千年所追求的结果。如果道教对长生不老的追求从人工智能中得到启发，下一步应该是追求人的精神的数据化，争取成为另类的“人工智能”，长久地活在数据流中。这也许就是人类未来的一个结局，成为数据流，成为几近永恒的存在，告别人类的身份，成为另一个种族——智能数据。

神仙的第二个特征即吸风饮露。所谓“吸风饮露”，就是神仙可以不食不饮，甚至不睡。神仙的这个特征有古代劳动人民对于不再饿肚子的淳朴想象，更是对超越的一种追求，告别了人类所必需的能量摄取，使“人”不再受到生存环境的限制，一步一步地达到庄子所谓“无所待”的绝对自由的境界。这个特征可谓是神仙的核心要求，仅就寿命而言，还有神仙“三灾五难”之说，有死于非命的神仙，却未见还需要饮食三餐的神仙。早在《庄子·逍遥游》中就有记载：“藐姑射之山，有神人居焉。肌肤若冰雪，淖约若处子。不食五谷，吸风饮露……”[18]神仙可以不食不饮，不代表其不能吃不能喝，我们一般羡慕谁日子过得好，都会说，那简直是神仙过的日子。神仙过的是什么日子？如果不能吃不能喝，那想成仙的人一定少了一大半，神仙必然是要比人吃得好才能体现出神仙的优越性。《楚辞·涉江》云“登昆仑兮食玉英，与天地兮同寿，与日月兮同光”[19]，神仙首先能吃我们所不能吃的东西，比如珍贵的金石、珍珠，也许在神仙的口中，这些普通人类难以下咽的贵重之物也是极为美味的。神仙还能吃

到仙界的食物，如仙酒、仙桃，另有种种珍果、香饭、宝浆等珍馐数不胜数。

若以“不食不饮”而言，人工智能早已达到了这个境界，还没有听说过手机、电脑要吃饭的，但其依然“有所待”，受到能量、载体的限制，也许强人工智能在某一天能够超脱这些限制，但是现在讨论起来，就如同农妇以为皇后会盘着腿在炕上吃大饼一般令人可笑了。但勉强可以讨论的是，由人升级而成的神仙，有品尝美食的意愿，而人工智能是否有感觉、感情？人会受到音、色、味等事物的引诱，会沉浸于这些感觉中，那么人工智能呢？如同科幻小说、电影中那些经久不衰的机器人与人的爱情故事一般，很多人都将“爱情”作为一个主要的衡量标志，似乎只有爱上什么人，人工智能才能算是和自然人平等的对象，而不仅仅是工具而已。但是“爱情”实在太主观、太复杂了，人类都难以鉴定一个人是否爱着另一个人，如何鉴定机器呢？对于强人工智能的判定，最广为人们接受的标准是图灵测试。图灵提出，人类如何判定一台计算机真的会思考呢？方法就是让测试者（一个普通人）和计算机通过键盘与屏幕进行对话，测试者并不知道对话者是另一个人还是计算机，如果测试者分不清和他对话的对象是人还是机器，那么这台计算机就通过了图灵测试，被认为具有人工智能。但是，图灵测试是否已经过时了？在深度学习发展的当下，可预见的是，通过图灵测试的人工智能很快将（甚至现在已经）出现，但是计算机和人对话的基础来源于大数据的支持，对话可能会涉及食物，计算机可以通过大数据给出很多食物的评价，但是能否知道这种食物究竟是什么滋味？计算机可以与人谈论音乐、谈论美术，甚至谈论“萌”，但它能否知道“萌点”在哪里？这样的人工智能是否可以鉴定为能够独立思考？恐怕不能，来源于大数据的回答，这违背了图灵测试的初衷。

神仙的第三个特征即与道合一。随着历史的发展，道教理论也不停地进行着演化。至唐代，道教心性学出现了极大的发展，至此，“得道”成为成仙的必要条件，成仙即要得道，得道方能成仙。唐宋

时期，道教方术出现了从外丹术到内丹术的转向，内丹术正是心性学在道教方术中的体现，“性”的修行被放在了一个极其重要的位置，甚至随着道教理论的发展，道教的革新派全真道登上舞台之后，修性的重要性得到了前所未有的提升。道教讲生道合一，所以“炼精化气，炼气化神，炼神返虚”的步骤就代表了肉体与精神共通的一种修炼方式，二者是可以转化的。所以，如前所述，抛弃肉体，将精神数据化成为“类人工智能”的存在，从道教的角度来说是不可取的。道教讲究贵生，身体是使生命向更完美的方向进化过程中不可或缺的一部分，而得成道体之后，精神与肉体合一再不可分，达到这一步后，就可以破碎虚空，成为仙人。粉碎虚空，即阳神合于大道，从此达到“不生不灭，云散碧空山色净；无去无来，慧归禅定月轮孤”[20]的境界。这是一种什么样的境界？难言也，用语言是无法描绘的。如果一定硬要用“科学”来解释这个过程的话，也许是神仙突破了维度，成了高维生物，从这个角度来说，神仙所在的层次比超人工智能更高。

如果说神仙是人类的进化，那如同人工智能一般的数据，也许是人类进化的另一个方向。未来等人类在人类社会死亡之后，也许会迎来在数据流中的另一个几近永恒的新“人生”。

如果说神仙是人类的进化，那如同人工智能一般的数据，也许是人类进化的另一个方向。未来等人类在人类社会死亡之后，也许会迎来在数据流中的另一个几近永恒的新“人生”。也许不需要未来，谁知道现在的地球是不是进化后的物种的育婴场呢？

人的道德

1950年，图灵第一次正式提出“人工智能”的概念。1956年，在美国达特茅斯学院召开的会议上，“人工智能”这一术语正式确立，这标志着人工智能的诞生。从“人工智能”这个概念出现开始，人类

的恐慌就没有停止过。对于科技的发展，人类第一次如此恐惧，因为人类已经开始踏入造物主的领域，人居然能够制造出另一种智慧物种？

从道教的角度来说，万物皆有道性，所以万物皆可通灵、成仙，但是这里的万物仅仅指生物，包括动物、植物甚至微生物、矿物，基本上是不包括器具的。所谓“形而上者谓之道，形而下者谓之器”，上天自然所创造的生物不同于人类所制造的器具，能求道的只有前者。中国传统文化中的妖魔鬼怪，实际上就是能达到和人类思维水平相同的生物，也可以说是能够通过图灵测试的生物。查阅各类妖魔鬼怪相关的文献，多见狐狸精、老鼠精、猴精、牛精、羊精、柳树妖、蚊子精、桃花妖、牡丹妖，但基本没有见到人造器物成精，不仅仅是人造，即使是“仙造”，有名如金箍棒、青萍剑、宝莲灯也没有任何自我意识。中国妖精构成大体如此，但是日本不同，日本的妖怪中有不少器物成精，比如围棋精（不知道围棋精和AlphaGo对战起来，究竟谁胜谁负）、砚台精、旧木屐妖、扫帚灵等。另有付丧神之说，日本人认为器物在经历很多年后会有灵产生，这种灵就叫作付丧神，其使器物具有化妖的能力。室町时代最具代表性的妖怪绘卷《百鬼夜行绘卷》中就描绘了大量的付丧神，如武器妖怪、乐器妖怪等，当今日本很流行的一款游戏《刀剑乱舞》就是以能够化为付丧神的刀剑为中心的。但是，日本这些所有的器物成精，也是由天地赋予灵性，绝对不是人类使其有智慧，“人之巧乃可与造化者同功乎”。以中国和日本的传统文化来说，可以接受某一天电脑忽然成精，但是绝对难以接受人类使一台电脑出现了类人的思维，这是形而上的领域，不应该是人类可以触碰的。如果人类可以制造智慧生命，那么未来的人工智能可以制造出新的人工智能吗？人工智能需要计划生育吗？毕竟如果没有外力干预的话，人工智能是不死不灭的。

按照人工智能层次的区分，弱人工智能由于其局限性，人们更愿意将其看成人类的工具。强人工智能是可以完成所有人类工作的人工智能，基本上可以顶替所有人类的作用，人类现在所做的一切人工智

能都能够完成，它们和人类具有一样的思维能力，过目不忘，学习速度是人类的无数倍，更重要的是，它们不吃不睡，永不疲倦，每一个“人”都是通才，可以无缝衔接到任何岗位。多可怕的未来，如果人的所有作用都可以被取代，那么人的意义又在哪里呢？动物园吗？当前比较流行的说法是，人类的情感、审美、创造能力，是人与人工智能最重要的区别，但这也只是现在人类最后的自我安慰罢了。而超人工智能是远远超过人类的智能，近似全知全能，蒂姆·厄班曾说过：“超人工智能，智能达到普通人类的17万倍。”[21]我们无法想象这是一种什么样的存在，也许无限接近于“神”。

人类正面临着生死存亡的巨大考验，很多人发出呼吁让科学家不要再走下去，但这是不现实的，人类的求知欲是不可能被磨灭的。所以，自“人工智能”这一概念提出以来，虽然强人工智能、超人工智能还存在于科幻小说和电影中，但人类已经开始兴致勃勃、源源不断地提出机器人（人工智能）所必须遵守的各种规则，最有名的无非阿西莫夫的“机器人三定律”与霍金等科学家最新提出的《艾斯罗马人工智能23定律》。从这些定律来看，人工智能的发展始终要以人类的利益为中心，人工智能的所有行动要符合人类的价值观，那么究竟是要符合哪些价值观和道德规定呢？人类社会应该坚持的最核心的价值观是什么？

道教的伦理观既讲道法自然，也讲仁义礼智。宗教也需要适应普世价值观，故道教也讲忠义孝悌。与儒家不同的是，道教宣扬神明监督，以保证伦理道德规范的实施。明清流行的诸多善书多托言于太上老君、关帝、文昌帝君、吕洞宾等知名神仙，如《太上感应篇》《文昌帝君阴骘文》等，劝人行善去恶，多以神仙的全知全能来进行监控，这也是道教的传统之言，所谓“举头三尺有神明”。人对神仙有敬畏之心，相信神仙的无所不知以及善恶报应，对普通大众来说，这是伦理道德实施的有效保证之一。实际上，在当前的科学技术下，监控系统无所不在，尤其是随着人脸识别系统的发展，当前人工智能程

序对人脸识别的准确率已经超过了人类的平均水平。歌手张学友的演唱会上，因为人脸识别系统的使用，数十名犯罪分子被发现。随着全国各地监控系统的铺设，犯罪事件应该会大大减少，这是人对科技的敬畏，是伦理道德实施的更为有效保证。

道教的伦理观也讲“自然无为”“尊道而行”，《道德经》谓“人法地，地法天，天法道，道法自然”，如何“道法自然”呢？即不违背自己的天性，不强迫其他生灵违背自己的天性，各顺其性，自然发展。求知是人的天性，道教也讲“观天之道，执天之行”“我命在我不在天”。人工智能的出现与发展是必然之事，当前人工智能并不是走人造大脑一般偏向于神经科学和脑科学领域的研究之路，而更多的是运用数据和算法，即主要是利用大数据的支持方式进行深度学习。如此看来，人工智能的进化路线与碳基人类完全不同。当前人工智能部分感觉系统的设置来源于人类，但我们看到的世界是一样的吗？如果我们没有听觉，我们也许能检测到声波，但是我们无论如何也不知道“听见”是什么感觉。我们的视域受到人类的生物机能的限制，人工智能的感觉必然能够比我们感觉到的更多，甚至可以产生除“五感”以外的人类无法理解的其他感觉，人工智能必然是与人类完全不同的种族。很多人都在担心强人工智能出现后会全面取代人类，人工智能真的可以取代人吗？这是一个假问题，猴子可以完全取代狒狒吗？不可以，但是猴子可以淘汰恐龙。当前科学家希望在开发人工智能的同时就向人工智能灌输与人类一致的伦理、法律、道德观念，这是当然以及必然的，但这种灌输好像古代帝王让百姓坚定不移地相信的“天授神权”一般，谁知道革命者什么时候会出现？谁会喊出一声“王侯将相宁有种乎”？当前探讨要给人工智能加上什么样的道德枷锁，不过是人类的闲聊与自我安慰，不如探讨人工智能的进化路线，看看是否有可借鉴之处。

人的道教

道教讲究寓道于术，道术合一，以术证道。道近乎术，方术是道教领域不可或缺的部分。道教追求长生不老、飞升成仙，途径就是通过种种道术。《四库全书总目提要》载：“后世神怪之迹，多附于道家……要其本始，则主于清静自持……其后长生之说与神仙家合为一，而服饵、导引入之；房中一家，近于神仙者亦入之；《鸿宝》有书，烧炼入之；张鲁立教，符箓入之；北魏寇谦之等又以斋醮章咒入之。”[22]按《四库全书总目提要》所云，道教就是道家哲学与方术的结合体。何谓方术？李零先生云：“与‘方术’的概念相近，古语有‘道术’，今语有‘技术’。但‘道术’多泛指，可以包括修齐治平的‘人文关怀’，好像比它宽了点；‘技术’则只限‘科学技术’，不包括‘迷信’，又比它窄了点。”[23]广义而言，科学技术也隶属于道教方术的范畴，皆是以术求道，方术求长生不老之道，科技求万事万物之理。“术”，本意是道路，《说文解字》云“术，邑中道也”，后引申为技术、艺术、数术等。“方”一说出自“经方、医方”，《汉书·艺文志·方技略》中有医方、经方、房中、神仙，皆与人体相关，故有此理解；一说“方术”即“道术”，《南华真经注疏》载“方，道也。自轩顼已下，迄于尧舜，治道艺术，方法甚多”[24]。方，《古代文化词义集类辨考》载：“甲骨文中作国家通称者，不是‘或（国）’‘邦’，而是‘方’。如《菁二》：‘五日丁酉，……舌方亦侵我西鄙田。’……还有井方、召方、鬼方、巴方、亘方、尸方、盂方、龙方、虎方、箕方，等等。成汤时号称诸侯三千，即是三千方国。”[25]《左传·宣公三年》载：“昔夏之方有德也，远方图物，贡金九牧。”“远方”并不是指遥远的地方，而是指边远之方国。所谓“天圆地方”，故“方”也可以引申为人间、世俗，《庄子·大宗师》载：“孔子曰：‘彼，游方之外者也；而丘，游方之内者也。’”《庄子·齐物论》载：“六合之外，圣人存而不论；六合之内，圣人论而不议。”所谓“方外”，即“六合之外”，故道士、和尚等出家人亦云自己为“方外之人”，即不在世俗之人。“六合之外，圣人存而不论”，圣人不论什么？《论语·述而》载

“子不语怪力乱神”，《论语·雍也》载孔子“敬鬼神而远之”，袁枚写《子不语》，内容多为妖魔鬼怪、神仙佛祖，故“方外”即指世俗之外，也指人类之外无限的世界。所谓“方术”，既可指“治国之术”，如韩非子所云之“道术”，亦可云为“方外之术”，即寻求形而上超越追求之术。“方外之术”的定义更为确切，道教方术外丹、内丹、房中、服食、导引、符箓、存思等，基本上皆为求仙之术，堪舆、星象、奇门遁甲、四柱八字等数术也是通过五行八卦等范畴对天地之理的一种探索。

方术何以能求道？“以术衍道”何以可能？所谓“人法地，地法天，天法道，道法自然”，方术即模拟演化天地之道，如外丹术即模拟天地运作制鼎，化年为日，以炼丹求道。方术随历史而顺应产生，最早出现的导引术、房中术、服食术，无疑来源于人类所日日不离的食、色、行。之后出现的外丹、符箓、内丹皆来源于人类对世界之理的探索。而随着对世界之理探索的结果和道教理论的完善，方术也发生了种种改变与转向，如汉魏时期服食术向外丹术的转向，也如唐宋时期外丹术向内丹术的转向，也如宋明时内丹术对房中术的吸收。至现代，方术不兴，技术大兴，二者皆为术，术与理的区别在于，“理”是万物运行之理，是“知”的范畴，而“术”是“行”的范畴，“术”是将“理”运用于人类之法。不管是方术还是技术，皆服务于人类。人工智能属于技术的范畴，考虑到超人工智能的超越性，似乎也可以将其归于方术的范畴。方术以术求道，人工智能是人类了解世界、解析世界产生的技术之一，如果从这个角度来说，对于强人工智能乃至超人工智能的出现，人类应该欢欣鼓舞，从某种角度来说，更“高级”物种的出现，可以使人更快地找到超越的途径。或者换个角度来说，超人工智能的出现，正是来源于人类对超越与无限的追求，即对“道”的追求。人由于生理机能的限制，难以在有生之年达到这一步，但不食不饮、长生不死的“人工智能”确是有希望能够“得道”的，若有那一日，人工智能能反哺人类，人类即可通过这种独特的“方术”，达到“与道合一”的境界。

超人工智能的出现，正是来源于人类对超越与无限的追求，即对“道”的追求。

小结 人与哲学的未来

在不久的未来，必然会有大量的人类工作被机器取代，如同20世纪90年代工人的下岗大潮一般。2016年年底，物理学家霍金在英国《卫报》上发表文章称：“工厂的自动化已经让众多传统制造业工人失业，人工智能的兴起很有可能会让失业潮波及中产阶级，最后只给人类留下护理、创造和监管等工作。”[26]很多人都认为未来只有原创、思想、艺术等职业才不会被取消，也许文学家、艺术家、哲学家是最后下岗的工种，未来说不准社会上会充斥着大量的哲学家，毕竟哲学家是门槛比较低的职业。科学与技术的变革，必然造成哲学的改变。从道教的角度来说，长生不老已经不再是传说，若超人工智能出现，也许道教的神仙谱系上会另开一系——数据神仙。

盖菲，字是非，云南大学哲学系讲师、中哲伦理教研室主任。专研道家、道教，研究方向集中在道教方术、神仙信仰、庄子哲学等方面。

---

[1] 陈鼓应注译：《庄子今注今译》，中华书局，1983年，第559页。

[2] 《淮南鸿烈解·原道训》，《道藏》第28册，文物出版社、上海书店、天津古籍出版社联合出版，1988年，第9页。

[3] 《高上玉皇心印妙经》，《道藏》第1册，第478页。

[4] 《金丹大要》，《道藏》第24册，第16页。

[5] 《还源篇》，《道藏》第24册，第213页。

[6] 《重阳立教十五论》，《道藏》第32册，第154页。

[7] （清）张志聪著，郑林等校：《黄帝内经素问集注》，中国中医药出版社，1999年，第46页。

[8] （清）郭庆藩撰，王孝鱼点校：《庄子集释》，中华书局，1961年，第741页。

[9] 林家骊译注：《楚辞》，北京：中华书局，2009年，第122页。

[10] （汉）许慎撰、王培林编注：《说文检字》，凤凰出版社，2013年，第409页。

[11] （汉）郑玄注，（唐）孔颖达疏：《礼记正义》，中华书局，1980年，第1457页。

[12] （汉）司马迁撰：《史记》，中华书局，1959年，第252页。

[13] 《淮南鸿烈解》卷二，《道藏》第28册，第47页。

[14] （汉）司马迁撰：《史记》，中华书局，1959年，第3226页。

[15] 《黄帝九鼎神丹经诀》卷一，《道藏》第18册，第795页。

[16] （三国）曹植：《曹子建集》卷五，四部丛刊景明活字本。

[17] （晋）葛洪撰：《抱朴子内篇校释（增订本）》，中华书局，1985年，第71—72页。

[18] （清）郭庆藩撰，王孝鱼点校：《庄子集释》，中华书局，1961年，第28页。

[19] 林家骊译注：《楚辞》，中华书局，2009年，第109页。

[20] 《伍柳仙宗》，河南人民出版社，1987年，第403页。

[21] 李开复、王咏刚：《人工智能》，文化发展出版社，2017年，第119页。

[22] （清）永瑢等撰：《四库全书总目》卷一百四十六子部五十六，清乾隆武英殿刻本。

[23] 李零：《中国方术续考》，东方出版社，2000年，第4页。

[24] 《南华真经注疏》，《道藏》第16册，第657页。

[25] 黄金贵：《古代文化词义集类辨考》，上海教育出版社，1995年，第3页。

[26] Stephen Hawking, This is the most dangerous time for our planet, https://www.theguardian.com/commentisfree/2016/dec/01/stephen-hawking-dangerous-timeplanet-inequality.

解决人类生死烦恼的根本智慧

——由人工智能引起的思考

刘丰河

人的本质是什么

宇宙万有的基本性质是存在性，存在展示为宇宙万有，没有存在就没有宇宙万有。人是宇宙万有当中的一种，人的本质当然是存在，离开存在便没有人，人是存在的展示。人具有的能力是“觉”“思”“为”，能力作用的结果是所觉、所思、所为。

为什么我们谈人的本质首先要提出存在性？因为人们在说人是什么的时候，往往以人的能力，就是觉知、思维和行为的能力，根据所觉知的进行推理、判断，得出各种不同的关于人的概念。我们应该特别注意一种情况：当一切觉知、思维和行为都没有作用的时候，存在不会消失，但不可凭觉知、思维和行为来说这个存在是什么。存在不因觉知、思维和行为的作用而存在；相反，因为存在，才有了觉知、思维、行为的作用，又因为觉知的能力而觉知到宇宙万有。也就是说，以人的能力来展示作用和作用结果的前提是存在。离开存在，作用及作用的结果就不会有。存在展示作用，作用体现存在，存在与作用不二。所有的作用和作用结果都必然与存在连接在一起，如镜显影，影生灭变化，而镜不变不动。

人在“觉”时应该明白，所觉知的人的生老病死和世界万物的变化等，都是不变的存在显现的变化的作用，而并非人的存在性有生灭

变化；人在“思”时应该明白，所思生灭变幻，与思连接的存在性不变；人在“为”时应该明白，所为千变万化，与为连接的存在性不变。与觉、思、为连接在一起的存在性既然没有变化，就不能说有生老病死；而存在作用既然显现为种种变化，就不能说没有时空万有的变化。人总是以不变的觉知去感受时空万有，并主动以“思”和“为”参与其中，从而体现人的价值。人如果把所觉知的当作存在的全部，便偏执于时空万有，忽略了存在和存在的能力；以人身为自我，我便是生老病死的，似乎这一切都不以人的意志为转移，就主观上接受了我必然生老病死的命运。若明白产生“觉”“思”“为”的存在是真正的主人，是无生灭的，便一念灭除生老病死的命运，又可以主动地作用或不作用，这样作用或那样作用，心都是自由自在的，人们便可以自由自在地“觉”“思”“为”。

存在不因觉知、思维和行为的作用而存在；相反，因为存在，才有了觉知、思维、行为的作用，又因为觉知的能力而觉知宇宙万有。

总之，存在，过去无始，未来无终，无生无灭，于不变当中又能起种种变化的作用，不变而变，变而不变，自由自在。

总之，存在，过去无始，未来无终，无生无灭，于不变当中又能起种种变化的作用，不变而变，变而不变，自由自在。

人类以感觉来认识世界的过患

我们要问一问，是谁在感觉世界？这个感觉的能力是所感觉产生的吗？难道是所感觉在感觉吗？人们所说的世界万有和人身是感觉的一个结果，还是所感觉的世界和人身产生了感觉这个作用？

比如，戴着眼镜时，人对戴着眼镜看到的周遭世界有一个认识。摘掉眼镜之后，人所感觉到的是什么？肯定与其戴着眼镜看到的不一样。人看到的哪一种世界是真实的呢？就算不戴眼镜，我们用肉眼来

观察世界，幼时、老时，病时、健康时，会产生不同的感觉，左眼和右眼看到的也不一样……哪一种感觉是真实存在？如果每一种感觉都是真实的，我在哪个世界中呢？我们不会说戴上眼镜看到的世界创造了不戴眼镜的人吧？我们可以得出一个合理的结论，不管哪种感觉，所感觉的世界并不是客观存在。所以，以感觉为依据来判断存在是什么，得出的判断并非真实。如佛所说，感知到的世界“如梦如幻”，《金刚经》中讲的“凡所有相，皆是虚妄”，就是这个意思。

如果把虚幻当真实，就主观上制造了假我和假世界。假我生老病死，假世界变化无常，而主观意识又认以为真，主观必然陷入主观制造的所谓的生死烦恼之中。

如果执着所觉的人身是“我”，便有生老病死之忧；如果以占有时空万物为所求，便会被变化无常无情地打击；如果以觉知的满足为欲望，则必然处在无尽的苦乐无常之中。于是，便觉“我”于时空转换之中，备尝生死烦恼之苦，人生结束，更有新的生死烦恼接续，如此这般，生生世世，无穷无尽，这就是所谓的生死轮回。

如何看待人生的价值

如果不再以人的能力来限定自己是什么、不是什么，对存在（本源、自性）没有偏差的领悟，就可以说我们有了大智慧，在佛法中叫“见性”。存在是无生无灭的，这是存在的本质，也是人的本质。

有了这个大智慧，我们在作用的时候，是以这个存在本身为前提来认识作用以及作用结果的。非常妙的是，存在具有觉知、思维和行为的能力，通过这些能力的展现，可以改变人和这个世界，这是人的价值所在。

如果认为人生老病死一场，今天过了，今天就没有了，人生的意义在过程之中或者给后人留下了什么，这样的人生观是不全面的。在

如梦如幻的感知里面产生的价值观是梦幻中的价值观，所求取的价值是梦幻般的价值。如果有了清醒的认识，就知道存在本身无生无灭，没有生老病死。存在可以起各种作用，作用随生随灭，生灭同时。作用的结果，产生的记忆，存留下来的只是自己建立的名词概念，对我们没有实质性的影响。这样，我们就不会再去分别、执着、妄想，也不会承受迷惑带来的苦果。

有了大智慧后，我们就会发现原来我们的最大价值是自由自在。自在，就是无生无灭；自由，就是不被觉知到的虚幻境界所束缚。无生无灭，自由自在，难道不是所有生命的最高价值所在吗？

所以人类的价值不仅是创造，更重要的是发现，是觉醒。明白了这一点，人生的价值意义，就是自己觉醒，彻底回归本有的自由自在；帮助他人觉醒，彻底回归本有的自由自在。

总之，存在本身已经决定了存在性和能力，离开存在，离开存在本身具有的能力来追求价值，谈论意义，这就是迷惑。

用大智慧如何看待世间万物

人仅仅是存在依据本身的能力作用以后，显现万有当中的一个种类。显现的动物、植物、微生物，矿物等宇宙万有，都是存在显现的相上的差别。

存在的作用体现在人身上就是人的“觉”“思”“为”等，和在动物身上体现出来的作用是不一样的；在植物、微生物，矿物上体现出来的又是什么？人们觉得这类问题很难解答。其实只要我们基于存在来谈论存在具有的基本能力、基本作用，也就认识到一切都是存在的显现。由于人们对万物的认识，常常基于感觉，就出现了主观上的以偏概全、以假乱真。我们应当基于存在本身具有的能力和作用，以及在作用的时候产生的因果规律来认识问题。

如何看待人工智能的价值和意义

离开存在什么也不能谈。我们必须明白，凡是我们可以感知到的一切，都是存在作用的显现。如果说这是人的作用，那也未尝不可。感觉上是人在发挥作用，其实是存在在发挥作用。按佛法的话来讲，今生是业力的显现，在业力的显现基础之上进一步发挥的作用，依然是存在本身在发挥作用。当然，人工智能也是存在作用的产物。

我们在梦幻中可能获得快乐，但是梦幻之中的快乐是抓不住的。如果不从本质上认识人和宇宙，就如同在梦幻中追求我们的价值和意义。人们做不同的梦，有“小我”，有“大我”，各种各样的“我”都在追求自己的价值实现，其实我们就如同“梦中人”一样。所以最重要的是觉醒，觉醒后自由自在地发挥作用，这是最美的享受、最高级的享受，这是人的终极价值和意义，也是本来现成的。

那我们还谈追求吗？这个话题，用真正觉醒以后的大智慧来看，人们最重要的是怎样纠正过去迷惑时形成的错误认识，怎样把造作的因果、束缚清除掉，回归真实，而不是追求什么。我们虽然有种种境界觉受，可以欣赏，可以享受，但是不能把它当真来抓取，有求皆苦。别把无生无灭当作生老病死，别把自由自在当作在梦幻中求取。黄金的价值是黄金，而不是黄金的光影。

谈起人工智能，对于大智慧者来说，是随缘自在的，不会刻意地发展人工智能，但也不会拒绝。大智慧者存在的价值是帮助其他的人也觉醒、自由自在。为此，就要广结善缘，也就是和其他人一起生活、工作、交流，在这个过程中觉醒他人。一旦真正醒悟，人们肯定不会执着于原来的所求，不执着于制造智能机器。当然，造机器就像人吃饭、穿衣一样，是人类所需要的，凡是人类需要的、能使生活更美好的事物，我们都不必拒绝。实际上佛法既讲究彻底的解脱自在，又讲究生活、工作上的智慧妙用，这两者是并行不悖的。用超世间的智慧行世间法，随缘自在，广结善缘，普利大众。

未来类人机器、超级智能的发展对人类意味着什么

人制造这些机器，无论多智能，本意应该是服务于人类的。智能机器如果被用于危害人类，那将十分恐怖。不过，危害人类的不是机器，而是人。

从因果的角度来讲，人和这个世界都是因果，果报如何，取决于自己的造作。实际上就算没有智能机器，人类也是生灭的，人类的这个时空也是生灭的，到那时候谁来救我们？只有自己能够救自己，只有觉悟能够救自己，只有觉悟才能彻底地解决人类的问题。人会说可以通过人工智能、生物工程等使人长生乃至永生。可如果这样的一个人并不是真实的自己，而是真实的自己的一个面具，活在假象当中，被假象迷惑，那就没什么意义了，至多是一个游戏而已。

不管是什么样的生命，只要觉悟了，都会发现连自己都不能毁灭自己。因为存在是无生无灭的，存在发挥的作用，无论如何作用、作用的结果是什么，不可能反过来毁灭存在本身。人类应该合理利用人工智能，但即便人工智能被用于危害人类，觉悟者也无所畏惧。

人工智能是否会拥有意识，成为生命

发展人工智能机器从最初就是为了帮助人类，它是服务于人类的工具。这个工具不怕疲劳，工作效率高，不求回报，没有私心，不会有意识。

人可以无心而心，无我而我，无为而无不为，能造出无我而我，无为而无不为的智能机器吗？

存在是无生无灭的，它可以作用产生意识，而有生灭的意识怎么可以制造出无生无灭的存在，从而产生意识呢？意识作用的产生与停

止的原因不是意识本身，以人类意识设计和制造出的人工智能，连人的意识都不是，怎么能产生意识呢？

机器不会有独立的意识。它的所谓的觉知能力、计算能力、分析能力，只不过是人的相应功能的延伸。离开人的作用，这些功能不会自动生成。人工智能只是人意识中的概念，人的意识之外不存在这样的概念。

生命是一个认识。人可以认为兔子是一个可爱的生命，兔子会认同这一认识吗？人可以认为老鼠是一个可厌的生命，老鼠会认同这一认识吗？人也可以认为机器是一个生命，那是我们赋予它的。其实，这完全是按照人的认知来说的，是自我建立的。

人机关系与人机伦理

机器不能自生人的智能和情感。把人的智能和情感“嫁接”给机器，它也不是自己拥有了这些能力和情感。所以，机器就是个工具。如果人把它当生命一样去看待，那是人的意识情感。我们与机器的关系就是我们与物的关系，就如同心与身的关系。

人们担心人工智能发达的时候，智能机器有情感，甚至担心人造的智能机器会反过来伤害人类甚至毁灭人类，这就产生严重的伦理问题了。首先，机器不会真正拥有人的意识；其次，人不觉悟不能从根本上解决伦理问题。没有出现智能机器的时候，伦理问题也很多。人类历史上的每个阶段都有伦理问题，都没有得到根本解决。除此之外，人类还面对如自然灾害等重大问题。人工智能可能带来的问题只是一小部分。我们更要关心人的本质、人生的价值意义这些最根本的问题。

结语

人工智能发展与不发展，这样发展或那样发展，于存在来讲没差别，但觉悟者可以充分利用人工智能去利益众生、觉悟众生。如果通过思考人工智能问题，了悟了宇宙人生的真相，那才是价值所在。以存在为根据来面对所觉知的一切，发挥主观的能动作用，把自在、自由充分地展示出来，这才是人的真正价值和意义的体现。

人工智能发展与不发展，这样发展或那样发展，于存在来讲没差别，但觉悟者可以充分利用人工智能去利益众生、觉悟众生。

最后，分享一个不是故事的故事：

一位梦者以亲历梦境的身份，认真地对梦境做了观察，并做了实验，然后写出了一份完美的报告，取名叫《有》。

一位醒者，看了报告后，笑了笑，在名字的前面加了一个字——空。

另外一位醒者，以亲历醒时境界的身份，认真地对醒境做了观察，并做了实验，写出了一份完美的报告，取名叫《真》。

梦者看完报告后，笑了笑，在名字的后面加了个字——假。

最后两份报告送到了一位通达梦与醒的专家那里。专家看后，也笑了笑，把两份报告的名字整合了一下，叫《真空假有》。

刘丰河老师毕业于河北医学院邯郸分院，长期从事医学教育工作。

新人文与生命价值

——基于佛教生命观的人工智能前景反思.[1].

李四龙

现在全世界都在谈人工智能，主流的态度是支持的，因为这项技术可以使人类的生活更为便利；质疑声也越来越大，因为人工智能技术的过度发展可能危及人类的生存。人工智能给我们今天的生活带来了很多意想不到的方便，而且强国之间的竞争越来越聚焦在人工智能。然而，作为人文学者，我们必须预先思考一系列根本问题——有些是社会问题，有些是哲学问题。在强大的人工智能面前，人类对自身的定位可能正在发生变化。以前，只有宗教观点认为“上帝造人”，但现在，人类在按自己的形象创造自己的“镜像”。我们的生命价值有可能因此发生重大变化，人文学科亦有可能随之发生相应的变化。

我们习以为常的社会结构、生活方式、研究范式乃至思维方式，可能即将发生显著的转变，我们需要对人工智能的发展前景进行反思。鉴于学术界已从多角度对人工智能进行了讨论，本文主要基于佛教的生命观探讨人工智能发展的关键问题，以及缘此而起的人工智能时代“新人文”教育。

两段小插曲

（插曲一）我的一位表哥是一名普通的铁路工人，一直在负责维护铁路信号，已从业30多年了，非常敬业。每次到北京，他都要买一

些有关铁路和火车的专业书籍。今年，他突然告诉我，现在他的工作量减少了，日常工作就是看护机器人、打扫机器人，整理、看护各种配件。机器人替他干活，他为机器人服务。

（插曲二）2019年3月初，媒体集中报道日本上市机器人女友，据说其和真人没差别，容貌姣好。机器人女友的出现，是否有可能让很多宅男产生依赖感？

这是两个很典型的事件。我表哥的工作现状既让人高兴也让人担忧：人类终于实现了自动化，成功地把自己从繁重的、危险的劳动中解放出来，这是人类从农耕时代就梦想能实现的事。与此同时又有人担心，人类是否会就此失去很多就业机会？大范围的失业会引发一系列严重的社会后果。机器人女友的面市听上去有些让人脸红，但这涉及人类社会的未来前景。随着机器人男友或女友的出现，人类的交往方式是否会发生变化？未来，人类的繁衍是靠自然生育还是技术手段？随之而来的将会是各种各样的伦理难题。

失业与绝种：两个不同层面的顾虑

上面的两段小插曲，一个是失业问题，这属于社会层面问题，可以说是“旧瓶装新酒”，老问题遇到了新麻烦；另一个是人类种群延续的问题，可以说是“新瓶装旧酒”，框架是全新的，但问题还是老问题，以前人类还只是讨论同性恋或自恋，我们还需要讨论“人机恋”。失业与绝种，这是两个不同层面的顾虑，分别涉及社会问题和人类命运。即使我们不把问题推到类似种群灭绝这样的极端结果，这两个问题也直接触及了未来社会结构与伦理秩序。

随着科学技术的发展，人类的自信心在不断增长：依靠医学、化学和生物学的发展，大多数疾病已不足以威胁人类生命；自然灾害也能在很大程度上得到控制；人类甚至开始了太空旅行，离我们渴望已久的“天堂”越来越近。科学技术成了各国的核心竞争力，我们不妨

称这样的时代为“科技全球化的时代”。在此环境中，人工智能是这个时代必然会出现的结果。它会更进一步展现“科学”给人类带来的“两面性”：造福人类或危害人类。人类从来没有像今天这样脆弱，既离不开那些科技带来的便利与舒适，又随时会面临各种高科技带来的灭顶之灾。

面对失业的顾虑，乐观的朋友就会解释，人类还会创造更多的就业机会。这是按照我们现在的生活模式理解人工智能时代。那个时代将会给我们带来什么，我想目前还很难确定。全世界的经济学家和咨询公司都在普遍关注这个严峻的问题，据李开复在《人工智能·未来》一书中介绍：2013年，牛津大学的两位学者预测，在未来10年或20年内，美国47%的工作可能实现自动化。不过，2016年经济合作与发展组织的三位学者预测，未来仅有9%的工作面临自动化的高风险。具体的预测数据还会有波动，但很多人的工作岗位确实已经岌岌可危，流水线工人、仓库员工、图书管理员、售货员、质检员、司机、收银员、会计师、精算师、股市分析师、律师助理，甚至部分临床医生，等等。其实，学校里只讲授专业知识的教师也会面临下岗的危险。

失业只是这一类问题的起点，人类社会的分层、流动与交往方式，乃至生活方式和价值观都将发生变化。但有一点是肯定的：未来，人们的观念必将不同于现在。举一个例子：注重面对面交流是传统社会的特点，所谓“百闻不如一见”，但在未来，这样的交流方式可能被认为是在浪费时间，新生代的人类更有可能依赖远程交流，虚拟的信息传送或许比直接见面交流更有效，甚至被认为更亲切。很多讨论人工智能的朋友，包括直接从事人工智能开发的工程师都会劝告我们这些人文学者，不必过于忧虑人工智能的未来，因为目前人工智能还处于“自动化”或“深度学习”的阶段，对人类还远构不成威胁。自从约翰·麦卡锡于1956年在达特茅斯会议上提出“人工智能”一词以来，该词的定义迄今仍没有被完全确定。现在我们通常把人工智能分成两大类：弱人工智能和强人工智能。我们平常用到的人工智能还都属于弱人工智能，即机器基于数据分析进行自动化决策与行

动。我们担心的是强人工智能，这一类人工智能目前主要出现在科幻片里，但凡人类能做的，强人工智能都能做，且比人类做得更好，这类人工智能不仅有思维能力，而且有自我意识。

思考人工智能可能会威胁人类生存这一问题的人，大多是政治家、知识分子，或者是相关专业的生物学家。生物学家对这一问题的忧虑，只是一项职业行为。但政治家或知识分子的思考，则是在想办法起到引导作用。我们希望以一种极端的结果，比如种群灭绝，以劝说人类改变目前正在进行的危险行为。

我们这一代人在20世纪八九十年代都曾幻想过未来世界，甚至还出现了一个所谓“未来学”的概念。现在的社会变化已经超出了我们在那个年代的想象，尤其是技术领域的进步完全超出想象。我相信未来的科技变化还会加速，直接体现在人工智能、基因工程、仿生技术、太空开发等领域。人工智能给商人带来太多的利润，给普通人带来太多的便利，给政治家带来太多的地区霸权。我曾经在一次讲演中说：“高科技是被逼出来的人类智慧，人工智能受制于人类对财富与享受的欲望，受制于国际政治对权力格局与资源配置的野心！”[2]这种不可遏制的欲望，最终是否会导致人类被人工智能控制？相比这种结果，人工智能所引发的失业似乎只是一件小事。

其实，失业这样的“小事”会引发连锁反应，直到爆发难以收拾的战争。所以，哲学家、政治家、社会学家、法学家等社会精英正在想办法为人工智能的开发建立“国际法”，构建公共伦理。但是，我有更多的忧虑：从劳动中解放出来的人类，凭借什么塑造“自我”？不需要劳动却能直接享受的人，还能不能被称为“人”？这是一个有关人格的问题。与此对应，现在有学者提出另一个似乎有些遥远的问题：人工智能将来有没有人权？我所接触过的某些佛教学者甚至在问：人工智能有没有佛性？这些问题就涉及生命的本质与意义。

生命的本质：从佛教到人工智能

我们一边在担心人类会不会灭绝，一边还在追问人工智能有没有人权。这听上去有些令人啼笑皆非。人之为人的理由何在？如何看待生命的本质？这在很大程度上会影响人工智能时代的社会结构，乃至影响高校的人文学科布局。

佛教认为生命是轮回的，并没有把人放在生命圈的最高位置，仅是六道轮回之中的一种生命形态。人和天、阿修罗、地狱、畜生、饿鬼一样，都是在轮回之中的生命。这种生命，在印度被称为“萨埵”（梵文为sattva），在早期被译成“众生”，后被玄奘译成“有情”。“众生”这一说法体现了生命的多样性，“有情”的译法则能体现佛教对生命的独特理解。按照佛教唯识学的解释，“有情”是指“有情识”。玄奘译有《成唯识论》，他的弟子窥基撰有《成唯识论述记》，对此解释称：“今谈众生有此情识，故名有情。……或假者能有此情识故亦名有情。又情者性也，有此性故；又情者爱也，能有爱生故。……若无本识，复依何法建立有情？有情之体即是本识。”[3]

生命的根本，因此被认为是“情识”。凡具备“情识”，就被称为“有情”，即生命。然而，“情识”并不是最终的概念。《成唯识论述记》中说，“情者爱也，能有爱生故”。也就是说，生命的本质特征是情或爱。佛教所说的“爱”，是指贪欲。在佛教看来，普通的生命之所以沉沦于轮回，是因为总是受制于不断生起的、内在的欲望。与此同时，生命也有一种可能，即通过自己的修行有效克制这些欲望，形成一种精进向上的愿望，最终实现轮回的解脱，对世界的真相有一种真切的理解，依靠这样一种智慧状态，实现生命的永恒。沉沦欲望、精进向上，其实都是生命的内在之爱，这些可能性存在于被称为“阿赖耶识”的本识内。这样的本识并不是固定不变的生命本

体，而是展现为生命个体不断实践的动态过程。也就是说，情与识，当下即是，全体起用。

在从欲望到智慧的转变过程中，理性一直在发挥作用，但佛教否定固定理性的存在，所有的转变过程都没有固定的程序，一切随缘变化。所谓的理性，其实依旧十分感性，即是“离苦得乐”，并在此基础上体验生命的空性。生命，无论是轮回中的生命，还是获得了解脱的生命，都有无限的多样性。生命的苦难多种多样，生命的圆满也有不同的呈现方式。比如诸佛菩萨，虽然他们的终极状态没有差异，但在面对现实世界时，他们会有多种多样的生命形态，观音菩萨就有33种变化身。

佛教对生命的理解特别强调生命状态的随机性，并没有固定不变的理性。偶然性或随机性的出现，并不是在对抗一种必然性，而是因为凡夫俗子在智慧上的缺陷。

所有的佛教徒都会坚守佛教的因果理论，认为有因必有果，这是一种必然的规律。但是，这种必然性并不属于生命个体，而是佛教徒所坚信的被视为“法”的规律，可以被界定为“空性”。如果从生命个体的角度来看，因与果的关系并不是线性关系，而是受制于多种多样的因缘条件，其两者的关系充满偶然性。因此，佛教对生命的理解特别强调生命状态的随机性，并没有固定不变的理性。偶然性或随机性的出现，并不是在对抗一种必然性，而是因为凡夫俗子在智慧上的缺陷。这种缺陷使生命个体感受不到内在的必然性，在自己的世界里随处都能体验到飘零和无常。

在佛教徒的世界里，人类的生命沉沦于轮回，充满了各种各样的缺陷。我们对人工智能的期待之一就是人工智能的准确性，它比人类更精准。我们想利用人工智能的准确性弥补我们自己的缺陷。如果这是一种固定的理性，那么人工智能对人类来说并不可怕。为什么以前我们对机器人的期待总是美好的？因为它们能替我们干活。所谓“机器人”，在我们以往的观念里不过是一种自动化仪器。迄今为止，在

医疗、驾驶、制造业等领域，机器人所谓的“智能”总体而言还停留在“自动化”范畴，只是当前计算机的计算能力远非当年能比。大数据的出现使人工智能基于大数据实现超速统计并做出相应反应。这在普通人看来就是一种卓越的智能，但其性质依旧是高速的自动化。面对这种人工智能，人类是不需要担忧的，尽管这一现象可能导致一部分人失业。比如在医疗领域，机器人医生做微创手术，远比人类医生更专业、更安全。这种基于算法的准确性，是一种固定的理性，不会成为生命的要素。

但是，我们现在讨论的是另一类人工智能，即所谓的“强人工智能”。在前文讲的机器人的基础上，可能会出现智能机器人，两者最大的差别是自动化与自主性的区别，智能机器人可能会有自我意识。也就是说，未来的人工智能有可能直接替代人类，那些智能机器人不仅在外表上长得和人类一样，重要的是他们还会自主学习，并且比人类更敬业，业务能力更强。什么是自主学习？就是如果这些机器人遇到挫折，它们能自己改变统计方式，能自己采集新的数据。如果未来真是这样，那么人类确实将会面临危机。

我们人类什么时候会想到“自学”？这可以是逻辑演绎的结果，也就是我们通过理性思考决定勤奋苦学。但在更多的时候，自学是由于某种欲望，可能是某人就想要赌一口气。情比理更重要。智能机器人的自主性表现在它能随机面对具体的实际情况，在遇到困难的时候能自主学习，从而自行解决难题。面对失败或受挫，智能机器人积极寻求新的办法，这是最简单的情感生活。因失败或转败为胜而露出不同的面部表情，这对人工智能来说已经不是难事。随着数据量的增大，各种可替换方案的输入，智能机器人的情感生活将会越来越丰富。情感生活的实质是能接受各种各样的随机性或偶然性。佛教观点认为，这种随机性或偶然性来自生命的欲望。在人工智能的世界里，尽管很难想象智能机器人会有主观的欲望，但那种主动纠错的做法已经表现了某种欲望。随着纠错能力的不断增强，智能机器人完全有可能获得自主学习的随机能力，有可能自主产生人机交往的模糊感情。

只是和它们强大的计算能力相比，它们的模糊感情可能显得有些幼稚。但是，谁能断定将来不会出现高情商的智能机器人？情感生活、自我意识，这些对未来的人工智能来说或许都会成为可能。人工智能这些被“情商”的特点其实与人类的纠错能力有关，归根结底，这取决于人类的欲望。能否妥善处理各种随机出现的状况，这是机器人在将来会不会获得生命特征的关键。

不过，机器真的要成人，其材质也是相当关键的。我们的身体是肉身，机器再怎么仿生也不会拥有肉身。佛教以缘起论解释生命，以“五蕴”的和合或离散解释生命的生与死，这些过程充满了非理性的因素。五蕴，是佛教的基本概念，认为生命是色蕴、受蕴、想蕴、行蕴和识蕴五大类要素的和合，死亡则是这五蕴的离散。简单地说，色蕴是指生命的身体，受蕴是指个体的感受，想蕴、行蕴则是基于感受的想象、推理和判断，识蕴则是有意或无意支配上述心理活动的依据。因此，从表面上看，智能机器人可以获得生命：色蕴就是机器的身体，受蕴是它们对外部信息的接受，想蕴是对这些信息的处理或计算，行蕴是相应的判断，识蕴则是先前已存储的各种数据。但是，这样的机器人真的就可以被称为“生命”吗？当然不可以，因为机器人的五蕴并不一定能形成有机的整体，尤其是识蕴的形成与作用机制。按照佛教的说法，识蕴是转世投胎的生命主体，生命的解脱，也是要把这个识蕴转化为智慧。智能机器人能不能拥有这样的识蕴？我们很难想象在一个机器人被拆解之后，它还有可能在另一台机器上复活。我们对智能机器人的顾虑，并不是担心它们在被拆解后会复活，而是只要不被拆解，它们就永远死不了。佛教在讲五蕴的时候，总在强调生命的空苦无常，人命危脆。但是，机器人的色蕴并不会无常，它们总是不知疲倦，永远不会被累垮。即使设备老化了，也可以更新零件，甚至它们自己会主动更新零件。永远不会被累垮的智能机器人，也因此永远成不了我们自然界有缺陷的生命。作为自然界的生命，我们的尊严首先来自肉身的显现。没有肉身的智能机器人，或许永远不

会有尊严，但是它们比我们自然界的生命更强大。我们的顾虑和危险来自我们的尊严与我们的软弱。

到目前为止，智能机器人可能还没有自我纠错的能力，也还没有自我意识。只要还没有这种能力，总体上说它们还只是自动化设备，谈不上是生命。从佛教的角度来看，生命的本质是随机的欲望。作为人文学者，我希望人工智能的开发设立一个禁区：不要让未来的机器人有自我意识，不能让它们有生命的欲望。我们能赋予它们的只能是固定的理性或某种特定的算法，而不是无限多样的纠错能力，也就是不能赋予它们随机的欲望。我们在讨论人工智能的时候，其实还必须弄清楚谁是“智能主体”？我们或许都会认为，这个智能主体就是作为人工智能体现者的智能机器人，但为什么既没有包括创造智能机器人的人，也没有包括使用智能机器人的人？或许，我们必须考虑一种共创的智能主体。

现阶段机器人的主要优势还在于远高于人类的精确性，它们仍属于自动化仪器。我们不应该去突破那个所谓的“奇点”[4]，而且，在我看来，这个“奇点”就是机器人开始具备思维能力、自我意识的转折点。也就是说，我们不应该去追求强人工智能。在弱人工智能范围内，人机关系还是可控的，只是停留在源于机器人的社会问题，还能在政治、经济和法律等传统的学科框架内协商解决。但到强人工智能时代，人类的尊严可能会丧失殆尽，人类甚至有被整体替换的危险。

在技术的快速发展面前，包括哲学家在内的所有人，都应该反思自身的处境。缘于技术的进步，我们变得越来越自信。但发展到今天的人工智能越来越聪明，与此同时，我们其实看到的是人类自身的衰落。什么是生命的本质？我们还需要进一步反思实现生命价值的方式。我们或许从一开始就走错了路。

生命的完美，始于人类的自我珍惜

站在佛教的角度，轮回中的生命充满了苦难，因此需要通过修行实现生命的解脱，以人生的觉悟实现生命的完美。然而，面对越来越完美的人工智能，我们应该重新理解生命的完美。我们为此必须思考当前人工智能在全球范围内的激烈竞争究竟意味着什么。分析人工智能的技术空间是一回事，分析人工智能所引发的社会问题与人文关怀则是另一回事。前者是硬道理，后者是软实力。

人工智能把人从危险、复杂的劳动中解放出来，这就是关于所谓对人有益的人工智能的最正面的解释。事实上，借用人工智能技术开发的无人机等，具有无可估价的军事价值；机器人的广泛使用，可能会导致大量人口失业，却让企业主获得空前的利润。因此，谁也无法阻挡人工智能新时代的到来，人工智能将会渗透到社会各个领域。即使在文化教育领域，人工智能的运用也已逐渐发挥作用，比如，并不太复杂的翻译工作已经可以由人工智能完成，在这种情况下，各国都会全力发展人工智能，谁不发展，谁就会失去未来的优势地位。但人工智能的发展引发的社会问题十分严峻，若不及时处置，人工智能的危险更甚于核武器。

失业问题严重，导致相当一批人成为领取社会低保的贫困阶层。当智能程度越来越高的机器行走于社会的时候，我们这些有缺陷的人类就有可能逐步被替换。有身体缺陷的人，将会被设计成愿意享受低保、自娱自乐、多少有些自闭、喜欢宅在家里的生命体。人工智能和基因修复一样，都是以否定人类的缺陷为使命的。从这种意义上来说，人工智能是对人类自身的否定。这需要哲学家、社会学家、伦理学家和法学家联合起来，共同考虑“人工智能伦理学”或“机器伦理学”，并不是所有的人工智能都可以被开发。就此需要制定开发底线的规定，建构人工智能国际法以保障人类的生命尊严。

佛教在这个层面能给我们提供的思想资源，就是一再提醒我们必须顾及生命的价值与尊严。现在，佛教经常被解释成自然科学的同路人，被说成是科学的科学，甚至有人说，“科学家千辛万苦爬到山顶时，佛学大师已经在此等候多时了”[5]。还有好多人在做实验，要去验证佛教修行的实际功效，寻找佛教与前沿科技的结合点。但是，技术层面的人工智能一定是由科学家、工程师来完成的，佛教学者在这方面并不占优势。在智能机器人不断挤压人类生存空间、就业机会的时候，佛教应和其他宗教一起，共同引导全社会珍惜生命，引导人类首先应当去实现对自身缺陷的自我超越，而不是自我否定。我们这些凡夫俗子受制于有欲望的肉身，正因为这个充满了缺陷的肉身，我们有别于任何人造物，包括正在日新月异的人工智能。因为肉身的缺陷，人类创造了文明，赋予生命某种尊严和价值。佛教让人类趋于完善的努力，是在精神层面立足于生命的圆满，着眼于生命圈的和谐，而不是强者对弱者的更替。这是一种真正的人文精神。

因此，当我们还拘泥于东西方文明冲突与融合的时候，人机共存问题出现了。这并不是人与自然、社会的关系问题，而是人与自身创造物之间的关系。人自身的创造物有很多对象，既包括我们生产的各种产品，也包括我们头脑中不断掠过的各种想法。但智能机器人这种人的创造物，与此前的各种创造物都不一样：既是一种客观的存在物，更是一种能力卓尔不群的存在物。譬如，人工智能与基因工程能延长病人的寿命，智能机器人的寿命在理论上可以无限，但这会产生更多的社会问题，例如，这个地球如何维持日益膨胀的人口？人类与机器的生命界限被模糊了，尽管人类暂时还不用考虑人被机器控制。

佛教让人类趋于完善的努力，是在精神层面立足于生命的圆满，着眼于生命圈的和谐，而不是强者对弱者的更替。这是一种真正的人文精神。

在人工智能不断发达的今天，要想完善人类的生命，我们首先要倍加珍惜自己的生命形态。生命之尊严，生命之完美，恰好在于人类

的缺陷及其在精神层面的自我超越。一步到位的完美，很可能是我们人类悲剧的开始。脱离缺陷并不能直接带来完美，而很可能只能带来罪恶。事实上，全球范围内的人工智能竞争，主要是追求军事上的优势和商业上的利润，这是一条国与国、公司与公司之间的对抗之路。我们有可能享受人工智能给生活带来的便利，有的时候甚至能利用这一技术挽救我们的生命，但从根本上来说，这些便利只是全球对抗的副产品。今后的世界格局，很可能是按人工智能的发达水准区分。所有的国家都已被绑到了人工智能竞争这架战车上，不搞人工智能就是愚蠢，但搞了也未必聪明。

如果说跨国公司让这个世界变得相对和谐，以全球合作推进市场融合、社会进步，但到了人工智能时代，全球合作的必要性正在减弱，对抗的色彩越来越浓。这并不是人类的福音。

我们这个时代比以往任何时代都更需要关注生命价值，生命的完美始于人类的自我珍惜。重申以下道理，在这个时代变得极有必要：我们之所以需要完美，是因为我们有缺陷；我们并不愿意放弃生命的本质和尊严。

结语 “新人文”的可能内涵

从人类自身的角度来说，我们真的不欢迎有情感生活的智能机器人。但是，缘于人类对权力和利益的贪欲，人工智能的开发速度将会不断加快。技术的巨大发展将会改变人类未来的处境，整个人文学科的基本形态也会发生变化。

未来社会的人文学科布局，姑且称之为“新人文”，必须顾及人机共存的生存处境。人机共存的时代正在到来，我们的大学教育必须做好准备。对此，笔者有以下四点拟作本文的小结。

第一，新人文教育必须考虑基于大数据的智能阅读、写作和学习。图书的储存形式、知识的传播方式正在发生根本性的变化，“数字人文”将是未来大学教育必不可少的基本学术训练。

第二，科学、技术和工程学将是人文学科重要的知识基础。没有科技常识的人文学者将会失去立足之地，传统的人文学科将会不断被边缘化，甚至难以为继，科技将是未来哲学思想乃至所有人文学科无法回避的重大议题。

第三，得益于人工智能的大数据处理能力，个性化的生活方式将会得到充分的尊重，而且这种个性化的生活方式最能带动经济的发展。但与此同时，大数据的生活环境很可能让我们丧失个人隐私。在人工智能时代，个人的兴趣将会成为最大的社会资本，越有个性的生活方式越能引起大家的关注，越能引爆新媒体的流量。与此同时，社会的共识将会很难达成。所有新媒体的热点，通常不会持续太长时间。为了达成社会共识，侵犯个人隐私将是伦理学的难题。

第四，得益于惊人的计算速度，知识的传承变得不太重要；知识的创新，尤其是面向个人兴趣的新知识，将会是未来社会相互竞争的主体部分。这将会改变未来的大学布局。凡能引领知识创新、引领社会风气的大学，将会处在世界大学金字塔的顶端。

文学、艺术、历史、哲学等传统的人文学科，在未来的几十年里依旧存在，但对它们感兴趣的群体在将来大多不再从事学术生产，新生代将会依据自己的兴趣不断拼接、组合自己的知识结构，甚至他们的思考与写作将会依赖跨学科的“数字人文”，学科边际将会变得越来越模糊。

当然，不管怎么变化，我们心目中的“新人文”必须服务于提升人机共存环境里的生命价值。

李四龙，北京大学哲学系（宗教学系）教授、人文学部副主任、教学指导委员会副主任、佛学教育研究中心主任。迄今出版

《天台智者研究》《欧美佛教学术史》《美国佛教》《中国佛教与民间社会》等5部专著，主编《宗教与人文价值》等12部论著，主编《人文宗教研究》，发表百余篇学术论文。

---

[1] 本文首次发表于2019年3月22日博古睿研究院中国中心举办的研讨会“人工智能遇见中国哲学家”，英文版发表于2019年4月4—5日在美国旧金山圣塔克拉拉大学举办的“人工智能与人格：中国与西方的视角”会议。本文在这两次发表的基础上修改而定，借此机会感谢其他与会代表提供的观点和他们的评论。

[2] 2018年9月8日，知识实验室在北京大学百周年纪念讲堂举办首期“共生课堂”，笔者在其中主讲“科技与佛法”。

[3] 摘自《成唯识论述记》第一卷。

[4] 奇点，并不是人工智能独有的概念。在数学领域，奇点是指一个未定义的点；在物理学领域，奇点往往被说成是宇宙大爆炸之后时间和空间开始的原点。1983年，弗诺·文奇提出“技术奇点”的概念，把“奇点”定义为人工智能超过人类智力极限的时间点。到21世纪，雷·库兹韦尔对此做了进一步的阐述，发表了两部备受关注的专著《奇点临近》（2005）、《人工智能的未来》（2016）。到了“奇点”以后，人类社会现有的交往原则可能都会失效，未来世界的发展将会超出人类的理解。

[5] 南方科技大学前校长朱清时先生在2009年发表讲演——“物理学步入禅境：缘起性空”，并以这句话结尾。这句话现已成了中国佛教界广泛引用的名言。

科技哲思

人工智能中的不确定性问题与哲学研究

陈小平

人工智能的两种经典思维及其融合

人工智能研究已经形成了至少几千种不同的技术路线，其中最成功、影响最大的有两种研究范式——暴力法和训练法，不过我更倾向于称之为人工智能的两种“经典思维”[1]，因为它们已经上升到机器“思维”的高度，而不仅仅停留在技术层面。这两种经典思维可以在很大程度上代表目前人工智能的发展水平。

第一种人工智能经典思维是“基于模型的暴力法”。暴力法的基本设计原理是：第一，根据问题的精确模型，建立搜索空间；第二，压缩搜索空间，使推理或搜索在计算上是可行的；第三，在压缩空间中枚举所有选项，找出问题的一个解。暴力法的基本前提是：待解问题存在一个良定义的、精确的符号模型，这种模型的建立通常采用逻辑形式化、概率形式化或决策论形式化的手段。

以逻辑形式化为例，一个人工智能推理系统由一个知识库和一个推理机组成。推理机是一个执行推理的计算机程序，往往由专门的团队长期研发而成，有免费版和商业版供用户选用。而知识库则需要不同应用的研发者自行开发。推理机的作用是根据知识库里的知识回答任何问题。

下面给出一个高度简化的知识库的例子。假设我们想回答有关“就餐”的问题，于是用一阶逻辑表达了一些知识，如表1所示，并存

放到知识库中。再假设我们有一个可以使用这个知识库的推理机。于是，可以对推理机提问：“碗能盛米饭吗？”［*hold(bowl, rice)*？］，推理机经过自动推理会答复“可以”（yes）。也可以对推理机提问：“碗能盛什么？”［*hold (bowl, x)*？］，推理机会答复“米饭、汤、……”推理机还可以答复更复杂的问题。

表1 就餐知识实例

逻辑公式	语义解释
$\forall x,y\ (dish(X) \wedge food(y) \rightarrow hold(x,y))$	餐具可以盛食物
food (rice)	米饭是食物
food (soup)	汤是食物
dish (bowl)	碗是餐具
……	……

理论上，推理机具有“可证正确性”，也就是说，理论上可以证明一个推理机的正确性。“可证正确性”比“可解释性”强得多，可证正确的一定是可解释的，但可解释的未必是可证正确的。所以，在理论基础方面，“暴力法”比“训练法”强得多，训练法目前连可解释性都没有。

可是，知识库的“正确性”就复杂多了，至今都没有普适性的严格标准，只能通过大型实验测试加以检验。暴力法做过很多大型应用测试，发现的普遍问题是：如果对推理机提出的问题在知识库的有效范围之内，理论上推理机都可以正确回答；如果提出的问题不在知识库的有效范围之内，回答往往会发生“离奇”的错误。这种现象称为人工智能系统的“脆弱性”。

第二种人工智能经典思维是“基于元模型的训练法”。训练法的基本设计原理是：第一，用人工神经元网络表示问题的输入－输出对应格式（包括网络结构等），即元模型；第二，依数据拟合原理，用标注数据训练元模型中的参数值（人工神经元网络连接权重），从而得到问题的一个具体模型，即一个参数已被训练好的人工神经元网络，这种网络可以对任意输入快速计算出对应的输出。图1是一个三层人工智能网络示意图。所谓深层神经网络，就是有很多层，比如几十层、几百层甚至更多层。

以著名的ImageNet[2]图像分类比赛中的一项任务为例。首先由比赛的组织者收集一个大型图片库，并将其中的图片人工分为1 000类，每一类中的图片标记为0~999中的一个数字，比如7种鱼的图片分别标记为0~6中的一个数字，公鸡和母鸡分别标记为7、8，26种鸟标记为9~34，卫生纸标记为999。这个过程称为“编码”，是人工完成的。

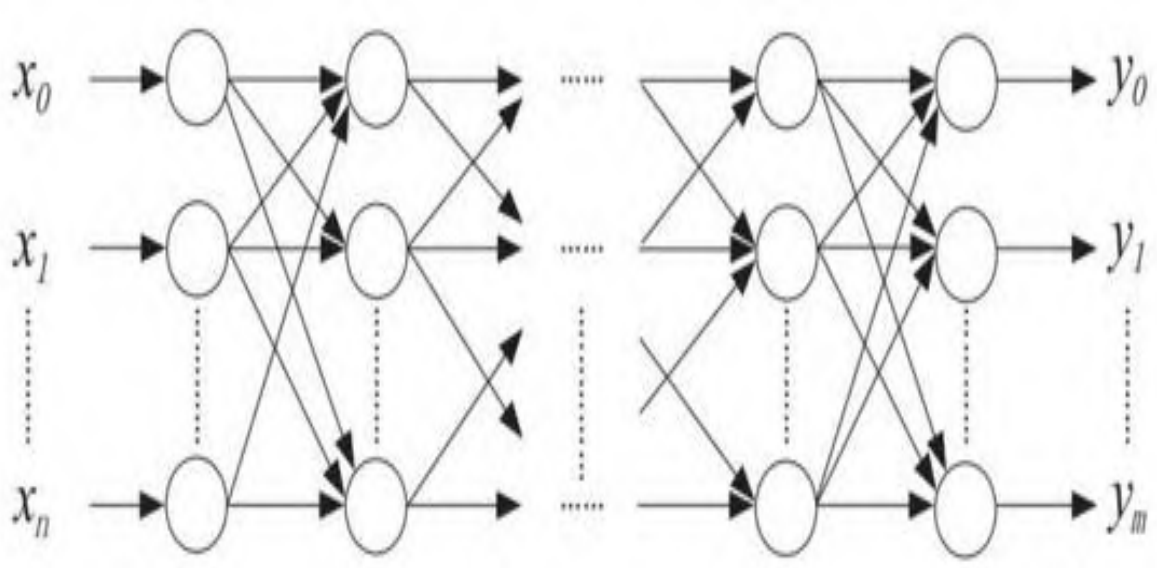


图1 一种深层人工神经网络结构示意图

训练好的人工神经网络被要求以非常高的概率（比如95%以上）达到下列识别要求：输入任何一张图片，则人工神经网络输出该图片对应的数字。比如，输入公鸡的图片，则应输出数字7；输入卫生纸的图片，则应输出数字999。从实际效果来看，如果达到了这个要求，就可以认为被训练的人工神经网络“学会”了识别图片库里的图片所显示的那1 000类物体。

在训练之前，人工神经网络的输出值与训练样本标注的期望输出值之间通常存在一定的偏离，所以需要用训练来尽量缩小总偏离。为此，训练过程收集大量输入－输出偏离值，然后根据数据拟合原理，用监督学习算法调整元模型的参数（即人工神经网络中的连接权值），尽量缩小总偏离（比如5%以下）。

在ImageNet比赛和其他测试中，训练法在很多给定图片库或其他标注数据集上的表现已经超过了人类，这是一个里程碑式的进展。不过，这不意味着在真实环境中训练法的识别率已经超过了人类。在一些测试中，训练法的局限性被逐步揭示出来。美国麻省理工学院的研究人员做了一个“定向对抗攻击”试验。[3]试验攻击的对象是一个

著名的商业机器学习系统，该系统训练出一个人工神经网络，其可以从照片中识别各种枪支，并达到很高的正确识别率。试验人员人为修改了照片上的少量像素，这些修改对人眼识别没有任何影响，但试验结果表明，同一个人工神经网络却不能正确识别修改后的照片。

这些测试表明，训练法和暴力法一样，存在着“脆弱性”问题。另外，训练法不仅不具备可证正确性，而且也不具备可解释性，即无法解释一个训练好的人工智能神经网络是根据什么将输入映射为输出的，不论映射对与不对。这两个问题叠加起来，导致训练法在开放的真实环境中的应用可能是不可靠、不可信的。

AlphaGo是两种人工智能经典思维的融合。AlphaGo一共有4代，第四代AlphaGo Zero通过40天“自学”，完胜第三代AlphaGo Master，而AlphaGo Master在实战中完胜所有现役的围棋顶级高手。

AlphaGo Zero的自学过程如图2所示，由自博和训练两个阶段的反复迭代构成。自博就是自己和自己下棋，从而收集大量数据。利用这些数据，反复训练一个称为“残差网络”的人工神经网络$f\theta$。$f\theta$被用作AlphaGo Zero下棋的决策模型，即用于决定每一个棋局中AlphaGo Zero怎么走。最初，$f\theta$只是一个元模型，所有参数（连接权值）是随机给定的，所以一开始这个决策模型本身只会“乱下”（但符合围棋规则）。在自博过程中，AlphaGo Zero在选择每一步棋的走法时，在$f\theta$的基础上，用蒙特卡洛树搜索进行1 600次模拟优化，选其中最好的一种走法，这样就得到比直接用$f\theta$可能更好的一步棋，也就是一个“优化数据”（Si, πi, Zi），其中i表示第i步棋。AlphaGo Zero保证自博的每一盘棋都会走到底，即下出胜负才结束，于是每下一盘棋，就得到一个“优化数据”序列（S_1, π_1, Z_1），（S_2, p_2, Z_2），…，（ST, πT, ZT），T代表这盘棋到最后一步的步数（AlphaGo Zero设定一盘棋最多不超过722步），而ZT代表这盘棋的胜负，它是一盘棋结束时根据围棋规则得到的，因此是一个精确值。其他Z1, Z2, …, ZT-1都是根据ZT反推出来的，因此也是精确值。

在训练阶段，AlphaGo Zero采用强化学习方法，利用自博产生的大量优化数据训练$f\theta$，得到更好的$f\theta$，然后用更好的$f\theta$ 再进行下一轮自博，再训练$f\theta$ ……。强化学习算法的基本特点是，可以根据一盘棋的走步记录和最后结果（胜负）推测每一步棋应该怎么走，而不用知道每一步棋走得好不好（这种指导信息事实上很难获得）。由于每下一盘棋都得到一个Z_T，2 900万盘棋就得到2 900万个Z_T以及相关的优化数据，这些Z_T和优化数据被强化学习算法用来训练$f\theta$，使$f\theta$变得越来越好，直到远远超过人类棋手的下棋决策模型。

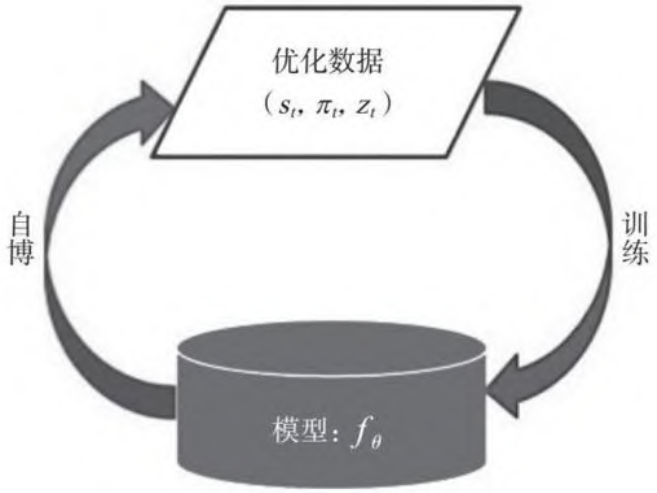


图2 AlphaGo Zero的自学过程

第一代AlphaGo使用的技术较多，而第四代AlphaGo Zero只用了4项技术——简化的决策论模型、蒙特卡洛树搜索、残差网络和强化学习。[4]其中，前两种技术属于暴力法，后两种技术属于训练法。可见，暴力法和训练法尽管是完全不同的机器思维，却并非水火不相容，而是可以深度协同工作，并达到暴力法和训练法单独应用无法达到的高度。有一种观点认为，AlphaGo Zero完全是深度学习的胜利，也旁证了暴力法的完全失败。这种看法是严重违背事实的，而且极大地误导了人工智能研究与应用的发展方向。

有专业棋手估计，AlphaGo Zero至少达到了20段。以此估计，人类的围棋能力与人工智能相比，相当于幼儿与成年人的对比。不过，围棋是一种完全信息的确定性问题/任务，这类问题在人工智能中属于最容易的类型，而最困难的类型是不完美信息的不确定性问题。

不确定性挑战

无论是专家还是公众，通常重点关心人工智能中的“难问题”，而倾向于忽视容易的问题。那么，依据什么标准来区分一个人工智能问题的难度呢？其实迄今尚无公认的标准。很多情况下，计算机科学中的“计算复杂度”被借用为人工智能问题难度的一种参考度量。比如，在博弈问题中，计算复杂度表现为两个指标——“状态空间复杂度”和“博弈树复杂度”。以这两个指标来衡量，围棋的复杂度远远高于国际象棋和中国象棋，因此这类指标是有一定的参考价值的。可是，只有在已经得出问题的计算模型的前提下，才可以使用两个指标；如果一个问题没有计算模型，那么这两个指标也不存在。对人工智能来说，如何得出问题的计算模型，其实是一个更加困难的挑战。暴力法完全依靠人工建模，而对某些问题，训练法并不需要完全依靠人工建模。因此，在这些问题上，训练法的建模难度低于暴力法。由此可见，计算复杂度不能真正度量所有人工智能问题的难度。

表2给出了人工智能问题的分类和在博弈领域的实例。不过，表2中的2、3、4都属于广义的不确定性问题，4是最难的。这里的“难”，就包含了建模的难度在内。由于不确定性给人工智能技术发展带来了巨大的困扰，现在已经到了对不确定性挑战进行深层反思的时候。

表2 博弈问题的分类

	分类	举例
1	完美的确定性问题	围棋
2	不完美信息的确定性问题	扑克
3	完美信息的不确定性问题	飞镖
4	不完美信息的不确定性问题	足球

不确定性实际上包含两个维度：一是观察/信息的不完全性/不完美性，二是行动/决策结果的不确定性。围棋是一个完全信息的博弈问题，因为下棋的双方都能看到棋盘上的全部信息，而且所有信息都是准确的。假设一方或双方棋手只能看到棋盘上部分区域的棋子，看不到这些区域以外的棋子，这种情况就具有信息/观察不完全性。面对这种不完全信息的“围棋”，棋手下棋的难度必然大大增加。在现实世界中，人们的观察通常都是不完全、不完美的，人工智能也是这样。

再看行动/决策结果不确定性的例子。最常见的围棋棋盘有纵、横各19条线，形成19×19=361个交叉点。按照围棋规则，所有棋子都必须落在交叉点上。假想一种新的围棋允许棋子不落在任何交叉点上，并且默认这样的一枚棋子可以落在棋盘的任何一个交叉点上，也可以不落在棋盘上，这种落子就是一种不确定行动，即行动/决策结果不确定的情况。面对这种围棋，棋手下棋的难度也必然会大大增加。

让我们更具体地考虑，假如围棋中存在两种不确定性，大概情况是什么样的。先考虑动作不确定性，即棋子可以不落在棋盘的交叉点

上。那么，对于这样的一枚棋子，它到底落在了哪个交叉点上？没人知道，人工智能也不可能知道。于是，人工智能系统不得不借助“动作不确定性假设”来确定落子的位置。一种简单直观的假设是，棋子以一定的概率落在周围的交叉点上，距离一个交叉点越近，落在那个交叉点上的概率就越大，越远则概率就越小。这个假设不是唯一可能的，还可以有其他多种不同的假设。再考虑观察不确定性。当人工智能系统观察到一枚棋子落在某个位置时，由于观察一般都不是完全准确的，棋子的真正位置很可能并不在观察到的位置上。为了确定棋子到底在哪里，不得不借助“观察不确定性假设”，即假设棋子的真实位置以一定的概率分布于出现的观察位置的周围。正态分布是最常见的假设，但绝对不是唯一可能的假设。由此可见，对于不确定性问题，两种不确定性假设具有关键性作用，如果其中任何一个假设严重背离实际，人工智能系统的性能必然受到极大的影响。

按照表2的分类，只要两个维度中有一个是不确定的，那就是不确定性问题。所以，后三类都是不确定性问题。现实世界中的各种动作和决策结果通常都是不确定的，所以人工智能也会遇到很多这样的场景或问题。面对不确定性问题，目前有三种可能解法。

第一种解法是把不确定性问题“改造”为确定性问题。典型例子是工业生产中的自动化生产线。没有实现自动化之前，相关的操作是由工人完成的，工人的观察和动作都不可避免地带有不确定性，所以整个生产过程充满了不确定性。建造自动化生产线的目的是消除生产过程中存在的一切不确定性，把原本带有不确定性的生产过程彻底改造为完全确定的自动生产过程。所以，自动生产线上的工业机器人和其他设备面对的是确定性任务，机器人和设备的一切动作都是重复性的，每次动作都是基本相同的。为此，建造一套自动化生产线的成本非常高，机器人只占总成本的1/8~1/5。另外，很多应用不能改造为生产线方式，非工业应用往往都不行，甚至一部分工业应用也不行。所以，第一种解法的应用范围是十分受限的。

第二种解法是应用训练法。从理论上来说，训练法不需要建模（但需要建立元模型），只需要大量训练数据。可是，训练数据是包含人工标注的，而人工标注是有代价的，所以需要选择一些代表性数据加人工标注。于是，如何选取代表性数据，就成为决定训练法成败的一个关键因素。根据上文对不确定性问题的分析可以看出，数据选择事实上依赖两种不确定性假设，否则数据选择将失去根据，虽然表面上训练法不需要两种不确定性假设。由于一般情况下这两种假设的正确性都是没有理论保证的，而且往往没有充分的实验保证，所以数据选择实际上存在很大的盲目性，从而严重影响了训练数据的质量，进而影响训练法的效果。上文指出的训练法的脆弱性，本质上就是由不确定性假设的不可靠性引起的。目前，这一困难在理论和工程上都没有切实可行的解决办法。

第三种解法是应用暴力法。由于不确定性问题目前没有合适的数学模型，因此一般情况下，暴力法的基本假设不成立。针对这一困难，研究中最常见的努力是借助概率形式化或决策论形式化，用概率表达观察和行动的不确定性。于是，这些概率模型或决策论模型直接使用两种不确定性假设的数学形式。由此产生下列困难。第一，不确定性假设往往不符合实际，还有很多实际场景甚至不存在稳定的概率分布，这就导致概率和决策论模型本质上的不适用性，模型决策往往与实际情况之间存在巨大的偏差。第二，概率推断本质上是证据不足情况下的决策，没有“保真性”，决策错误往往不可避免，这在很多应用中不满足用户要求。事实上，人类对自己的要求比较低，对机器的要求非常高，采用的是双重标准。第三，工程可行性障碍。有些模型的具体实现过于复杂，超过了现有工程方法的处理能力，无法保证实际效果。

我们以围棋博弈为例加以说明。首先，在围棋的主流博弈模型中，一个棋局被抽象为一个“状态”，一个棋手在一个棋局下走一步棋，被抽象为该状态下的一个“动作”。一个状态下往往可以有多个动作，它们具有不同的发生概率，于是形成动作集合上的一个“转移

概率分布”。理论上可以针对不同对手，“学习”一个对手在所有状态下的转移概率分布。可是，一个棋手在任何状态下的转移概率分布都是稳定存在的吗？这迄今没有得到理论证明或实验验证。如果不存在稳定的概率转移分布，怎么能“学会”？其次，围棋的棋局总数为2.08×10^{170}个，如何有效“学习”、存取这么多状态下的转移概率分布？这在工程上是困难的。AlphaGo最初也使用了这个模型，AlphaGo Zero则完全放弃了模型中的转移概率分布，代之以平均胜率（即优化数据中的πi）。这样一改，虽然模型的解析性丧失了，但是保证了平均胜率的存在性，而且工程上可以有效“学习”、存取平均胜率。这是AlphaGo Zero在人工智能建模理念上的一个重大突破，也是这个项目自身的主要创新点。

根据以上对人工智能不确定性问题的分析，以及对大量实际问题的考察和尝试，近年来我提出了一种应对人工智能不确定性挑战的新想法，称为融差性思维。

融差性思维：原理、意义和案例分析

根据以上对人工智能不确定性问题的分析，以及对大量实际问题的考察和尝试，近年来我提出了一种应对人工智能不确定性挑战的新想法，称为融差性思维，并初步完成了一些相关研究工作，验证了融差性思维在一些挑战性问题中的有效性。

“融差性”的双重内涵如下：第一，在一定范围内容忍各种“差”的不利影响，包括观察和动作误差，以及决策、动作和观察的某些偏差，在有“差”的情况下完成不确定性任务；第二，在一定条件下主动利用各种“差”所隐含的机遇，通过“融会贯通”误差、偏差和差别的方式，更好地完成不确定性任务。

下面通过对生活中一个融差性思维实例的分析，比较融差性思维和精确性思维之间的主要区别，进而阐明融差性思维的原理和意义。老农是如何让米缸尽量多地盛米的？很简单，先在缸里放一些米，使劲摇晃米缸，于是米粒之间的空隙就会被压缩，占用的空间就会减小，然后再放进去更多的米，再摇晃，直到米粒占用的空间无法继续减小为止。

用精确性思维（以暴力法为代表）解决这类问题，首先要建立问题的一个精确模型，在模型基础上反复进行如下观察－决策－行动循环，直到问题得到解决：观察，针对模型指定的独立对象，观测、获取相关的精确信息；决策（包括预测和规划），通过对观察信息的精确分析，决定下一步采取什么动作；行动，执行决策决定的精确动作。

为了建模，精确性思维将米缸中的每一粒米都视为一个独立对象，针对它们的位姿（在三维空间中的位置和姿态，即朝向）建立数学模型，作为观察、决策和行动的基础。由于米缸场景是一个不确定性问题，通常采用决策论形式化建模方法，为此需要引入观察不确定性假设和动作不确定性假设，它们的主要内容分别是：当对每一粒米的位姿进行观察时，观察误差服从什么概率分布？当调整每一粒米的位姿时，调整动作的误差服从什么概率分布？由此可见，引入不确定性假设的根本目的是弥补“完美的”（即精确性）理论思维与不完美的（即不确定的）现实之间的鸿沟，力图将后者最大限度地归结为前者——概率分布虽然没有一阶确定性（普通精确性），但体现了二阶确定性。不过，这种尝试的实际操作隐含着很多困难，比如，如何确定应该采用哪种概率分布函数，该分布函数的参数如何获得，假设的分布函数与实际观察和动作的偏差有多大，模型表示和求解的计算效率如何。尤其值得注意的是，不仅在米缸场景中，而且在人工智能尝试过的一切大规模不确定性应用场景中，这些问题都没有得到有效解决。

如果采用融差性思维，情况就会完全不同。首先，融差模型是一种宏观模型，不需要精确刻画每一粒米的位姿，只需要描述有关宏观特征的近似值，比如，米缸里有多少米（用米的重量计量）、米缸装满程度（通过观察缸中米堆的上表面与缸口表面之间的距离近似确定）、米缸里能不能容纳更多米（通过相继两次实验结果的对比，如果经过一轮摇晃没有减小米粒占用的总空间，说明缸中无法再容纳更多米）。这种宏观性的模型和观察体现了融差性的第一种内涵——容忍偏差，比如不需要精确到多少粒米，而是用多少重量（如以克为单位）来计量，可容忍的误差粒度（融差度）比精确性思维所能容忍的大得多。原理上，融差性思维要求把模型精确性控制在适当范围内，避免精确性过高或过低，更不认为精度越高越好，所以对模型和观察精确性的要求大大降低。同时，融差模型可以不借助不确定性假设，比如在米缸问题中不需要假设概率分布函数。上述特点概括为下列“融差模型原理”：融差模型通常使用近似宏观特征，避免使用概率分布型不确定性假设，从而将模型精度和观察精度控制在适中范围内。

在精确性模型的基础上，精确性思维下的问题求解过程的第一步，是对米缸中的所有米粒进行精确观察，测量每一粒米的精确位姿。事实上，任何测量都是有误差的，完全没有误差、绝对精确的测量只是一种理想化想象，精确感知在现实中基本上是不可能的。即使一个问题有精度足够高的感知手段，由于测量的代价太大，实际上也是得不偿失的。作为对照，融差模型由于控制了精度的适中性，不需要精确测量，所以不存在精确测量的困难。

第二步，人工智能系统进行决策，以决定下一步执行什么精确动作。在米缸场景中，一个精确动作就是调整所有米粒的精确位姿，使调整后所有米粒占用的总空间更小。这就遇到了暴力法决策的三个困难：由于不确定性假设不符合实际，限制了决策效果；概率决策无保真性，因而无法避免出现错误决策；模型过于复杂，缺乏工程可行性。例如，理论上可以为每一粒米规划一种位姿，在这些位姿下，所

有米粒占用的总空间最小。但是，这种规划的计算量极大，而且更严重的是，由于误差的存在，计算结果与实际情况之间往往存在巨大的差距，所以这种规划的可行性、实用性非常低。

在融差性思维中，这三个困难都可以较好地得到解决。第一，如上所述，融差模型一般不用不确定性假设，这就避开了第一个困难。这里需要进一步分析阐明的是：为什么不用精确模型，反而能够得到更好的决策？事实上，融差模型上的决策与精确模型上的决策不同，主要不是通过模型上的计算，而是通过人工智能系统的行动实现的。在米缸问题中，融差模型上的“决策”主要是借助“融差行动”（摇晃米缸）实现的，也就是通过“摇晃”动作缩小米粒之间的空隙，从而减小它们占用的总体空间，所以不需要根据所有米粒位姿的精确信息严格计算出如何调整米粒位姿，而是利用了大量米粒位姿之间的大量偏差。这体现了融差性的第二种内涵：人工智能系统主动融会偏差，借助融差行动完成决策（包括预测、规划）任务，从根本上避开了精确性思维遇到的难题。第二，融差性思维放弃保真性准则，代之以“融差度”准则：每个问题都有其特定的融差度，满足该融差度的任何解都被认为是可接受的，融差性人工智能系统追求的是满足容差度的解，不是精确解或最优解。在米缸例子中，在某次摇晃之后，发现米粒总体积没有变小，问题求解过程就结束了。而依精确性思维，即使观察存在很大误差，模型存在严重偏差，仍然只能进行精确的计算或推理，其结果往往是放大误差和偏差，甚至出现“南辕北辙”“差之毫厘，谬以千里”的后果。第三，融差性思维的工程可行性显然远远高于精确性思维，无须赘述。根据以上分析，“融差行动原理”概括如下：利用状态内部存在的各种偏差，通过执行合适的动作，引导当前状态向目标状态迁移，直到当前状态与目标状态之间的整体差距缩小到融差度以内。

第三步，基于精确性思维的人工智能系统将执行决策规定的动作。在米缸场景中，决策规定的动作是：分别调整缸中每一粒米的精确三维位姿，以使所有米粒占用的总空间变小。显然，每一粒米的位

姿都会影响周围其他米粒的位姿，“牵一发而动全身”。那么，一个人工智能系统如何调整每一粒米的位姿，确保调整后的位姿都符合决策的精确要求，而不会出现偏差？这几乎是不可能的。相反，采用融差性思维，用融差行动可以轻而易举地完成缸中所有米粒的同时调整，并在可能情况下减小所有米粒占用的总空间。这就将一个精确性思维几乎无解的问题化归为一个高效、低代价、低难度的问题。因此，融差行动的可行性和实用性比精确行动高得多。这进一步反映了融差性思维的第二种内涵：主动融会贯通界限差别。在精确性思维中，行动和决策属于完全不同的范畴，它们之间的界限是绝对的、不可消除的；而在融差性思维中，这种界限被打破了——决策可以通过行动实现，从而为人工智能开辟了全新的可能性。这体现了融差性的第三条原理——“界限融差原理”：一定条件下，不同范畴的机制（如观察、决策、行动）可以相互替代。

基于融差性三原理，应用融差法求解米缸问题的全过程总结如下。首先，构建米缸问题的融差模型，在此模型上进行问题求解的如下步骤：（1）观察，感知米缸装满程度；（2）决策，根据感知信息和历史信息，判断是否可能让米缸容纳更多米；（3）行动，如果可能容纳更多米，则执行一次融差行动（摇晃米缸），然后返回步骤（1），否则终止求解过程。对比可知，融差法和暴力法的问题求解过程具有相同的架构，但其中具体内容存在根本性的不同。暴力法基于精确观察、精确决策和精确行动。精确行动的典型例子是工业机械臂，重复精度一般达到亚毫米级或更高。可是工业机械臂在确定性场景中运行，为了在不确定性场景中达到类似的动作精度，感知精度和决策/规划精度也必须达到相应的水准，这在不确定性场景中是非常困难的，成为迄今无法突破的障碍。融差法基于融差性三原理，对感知、决策/规划和行动的精度要求大大降低了，从而提供了一条有效应对不确定性挑战的新思路。

下面进一步分析用训练法应对米缸问题的情况。用训练法应对这类问题时，采用的体系架构和暴力法类似，也是先建立问题的某种模

型，然后在模型的基础上进行感知－决策－行动循环，直到问题得到解决。不同的是，训练法的模型不是人工编程构建的，而是利用标注数据训练出来的，相对而言，对“数值精确性”和“精确决策知识”的要求大大降低，观察精度可能也会明显降低，所以灵活性大大提高。可是，训练法本身并不包含解决米缸问题核心挑战所需的观点、原理和机制，其中最重要的障碍是决策与行动之间的界限限制。训练法基于人工神经元网络，这种网络在标准的训练法中被用作分类、预测和决策的计算机制，所以在训练法中，决策与行动仍然是严格区分的。解决米缸问题的最大机遇在于用融差性物理动作替代决策计算，但物理动作无法直接嵌入现有人工神经元网络架构中，所以界限融差原理对训练法不成立。当然，人工神经元网络可以用来“学习”动作策略，但动作策略与物理动作本身是完全不同的，而且通过人工神经元网络“学习”融差行动（比如“摇晃”动作），就要涉及与融差行动原理相关的元模型构建和训练数据采集，这些都是包含巨大挑战的全新课题。

融差性思维在现实生活中早已存在，我们在人工智能研究中要做的，是将其科学化（广义形式化）、系统化、工程化，这样才能为人工智能所用。中国科学技术大学机器人实验室已经完成了一部分基于融差性思维的研究工作，比如实现了一系列基于融差原理的柔性手爪，可以抓握不同形状、不同刚度的不规则物体，从刚性的木块到柔软的豆腐，而且控制简便，对观察精度的要求也大大降低了。实验证实，在解决一些挑战性很大的不确定性问题时，融差法可以取得非常好的效果，值得进一步研究。我们也希望在进行相关人工智能研究的同时能够就不确定性等核心问题和哲学等学科展开跨学科的合作研究。

陈小平，中国科学技术大学机器人实验室主任、教授，中国人工智能学会AI伦理专业委员会主任，中国机器人世界杯委员会主席，中国科学技术大学杰出研究校长奖、2016年机器人与仿生

技术国际学会最佳大会论文奖、世界人工智能联合大会最佳自主机器人奖、12项机器人世界杯冠军团体奖获得者。

---

[1] 陈小平：《人工智能的历史进步、目标定位和思维演化》，《开放时代》2018年第6期。

[2] ImageNet是一个用于视觉对象识别软件研究的大型可视化数据库。——编者注

[3] Andrew Ilyas, Logan Engstrom, Anish Athalye, Jessy Lin, Query-efficient Blackbox Adversarial Examples, arXiv 1712.07113.

[4] David Silver, Julian Schrittwieser, et al., Mastering the Game of Go without Human Knowledge, Nature, Oct. 18, 2017.

成己而圣：在此岸建构理念世界

蔡恒进

2018年度图灵奖得主之一杰弗里·欣顿曾经直言人就是机器：“我们是机器，我们只是通过生物方式制造的。大多数做人工智能的人对这一点毫不怀疑。我们只是特别精密的机器。我刚才不该说‘只是’。我们是特别的、绝妙的机器。”[1]欣顿认为的特别与绝妙之处在哪里？他所指的是否就是意识或自我意识呢？

机器是否拥有或者能否获得意识或自我意识？要回答这个问题，必须首先回答意识与物理世界的关系问题，也就是意识是如何从物理世界中产生的，意识又是如何反作用于物理世界的。历史上的哲学、宗教流派都曾围绕这一问题展开讨论，却都没有给出明晰的答案。例如，康德用“先验直观”连接物理世界与意识，胡塞尔却将此问题悬搁起来。[2]我们在最近的研究中提出了触觉大脑假说[3]，试图为上述问题提供一个答案。

意识与智能的进化

我们在命名触觉大脑假说之前，称其为元直观假说。感觉（特别是触觉和味觉）器官（如皮肤和舌头）的存在，为生命个体提供了可感知的物理边界，也为生命个体区分“自我”和“外界”提供了物理基础。随着感知经验的增多，这种关于“自我”和“外界”的区分沉淀为关于自我和世界的观念（即元直观）。元直观可以被称为元模型，其他任何概念/直观的真实性最终都锚定在“自我”与“外界”能

够区分这一简单模型的真实性之上。在文献中，我们将元直观定义为“原意识”。[4]

生命的最初状态应该是从单细胞开始的，大多数人都倾向于先考查细胞的构造，认为先有细胞核，而后才有细胞的整体。但也有一种可能，即先产生了细胞膜，细胞膜不仅保护了细胞内部物质，而且使细胞具备了能与外界区分开来的独立性，在此之后细胞内部才逐渐确定出细胞核和细胞质等。随着时间的推移，细胞不断进化。一方面，功能为结构的发展提供了指引和方向；另一方面，结构的发展又进一步强化了功能本身，最终才形成完整且明晰的构造。这个过程的要义是，功能与结构纠缠在一起共同进化，而非先有明确的结构才产生了相应的功能。[5]

元直观可以被称为元模型，其他任何概念/直观的真实性最终都锚定在“自我”与“外界”能够区分这一简单模型的真实性之上。

高级智能的产生并不取决于脑重量。在脑细胞之间快速建立连接，脑重量快速增长的阶段，是否有清晰的“自我”与“外界”的边界，能否产生较强的自我意识，这才是高级智能产生的关键。对哺乳动物而言，触觉与其内分泌系统直接相关，而毛发减少、皮肤变敏感则为“自我”与“外界”的区分提供了物理基础。很多视觉灵敏的动物并不具备高级智能，可能就是因为其视觉系统并不与内分泌系统直接相关，所以对自我意识的形成显得就没有那么重要。

人类大脑约有1 000亿个神经元，但婴儿在母体内时，神经元之间的连接比较弱。人在刚出生时脑重量约为370克，两岁时的脑重量约为出生时的3倍，5岁时的脑重量已经接近成人。[6]脑重量的增加就是因为神经元间的连接大量增加或增强，因此我们可以认为，0~5岁这一人类大脑快速发育的时期就是产生较强的自我意识的关键时期。

阿道夫·波特曼[7]提出的“分娩困境”暗示，在母体内就发育到较成熟的婴儿大脑才更好（这意味着孕期要长达18~20个月）。但笔

者认为，人在出生后大脑内神经元的大规模连接才是产生比较强的自我意识的关键，因而出生后再发育成熟会更有利于婴儿的未来发展。例如，鸡和乌鸦都是卵生动物，小鸡刚出壳不久就能走路与进食，而乌鸦刚出生时没有绒毛，也没有视力，无法离开鸟巢，需要亲鸟饲喂1个月左右才能独立活动。[8]但许多实验数据证明，乌鸦才是最聪明的鸟类之一，鸡的智能则远不及乌鸦。

恐龙统治地球那么长时间，却没有证据表明它们曾经具有高级智能。恐龙最早出现在2.3亿年前的三叠纪，灭亡于6 500万年前的白垩纪晚期，共存在了1.7亿年。相比之下，最早的人类化石距今约700万年，智能的高速进化大约是在过去的几百万年内完成的。高级智能似乎不是在足够长的时间内单纯依靠“物竞天择、适者生存”进化达到的，只有在特殊条件下产生的基因突变，才能促使高级智能的诞生。因此，在宇宙中，有高级智能生物的星球可能非常稀少。

自我意识是智能的内核，是大自然的巅峰之作，是真正的混沌初开，是比宇宙大爆炸和地球形成更为精彩的产生。由触觉产生自我意识，而后在自我肯定需求的引领下不断地去认识世界，开显出精彩的认知坎陷[9]（意识片段），丰富自我的内涵，发展成具有全局统摄性的高级智能。

由触觉产生自我意识，而后在自我肯定需求的引领下不断地去认识世界，开显出精彩的认知坎陷（意识片段），丰富自我的内涵，发展成具有全局统摄性的高级智能。

一般认为，先验直观可以通过顿悟开显，但顿悟并不神秘。假设有一群母鸡，一直生活在安逸、稳定的环境之中，它们有翅膀，但从没有飞过，也没有见过其他的鸡飞，这群母鸡完全不知道自己会飞。有一天，它们原本平静的生活中突然闯进一只疯狗，撵得鸡群四处乱窜，有一只母鸡在慌乱之中张开了翅膀，结果它顺势一扑腾，竟然飞了起来……这就是母鸡的顿悟。这种在特殊场景下被激发出来的功能（顿悟）与物理结构有关系，母鸡拥有翅膀，但之前其飞的功能被

“遮蔽”了。顿悟受到身体构造的一定限制，相比之下，大脑的改变比身体结构的改变容易得多，因而大脑产生顿悟的可能性会大得多。智能的进化过程中很多是从外界、从他人习得而来，人可以从自然景观、植物、动物或他人处得到启发，从而产生顿悟。

由认知坎陷炼化出的理念世界

柏拉图认为，理念世界是完美的，我们看到的现实世界是一个不完美的投影，而理念世界又不可抵达。从进化论的角度来看，既然不可抵达，又如何证明理念世界的存在？有一种解释是，人是从理念世界而来，因此还带有一部分关于理念世界的记忆，至于人为什么会记得，就无法解释清楚了。过去我们尚可以认为是“上帝”或“上天”创造了人类，赋予人类以灵魂，这种说辞或许可以将这些问题暂时搁置起来。但现在，面对人工智能技术的发展，这些问题不可以被继续搁置，人们应该且必须回答理念世界从何而来。

理念世界的内容并不是因为我们“记得”，而是因为人类经过漫长的进化，将认知坎陷（意识片段）逐渐炼化，不仅在同一主体内持续炼化，而且能够传递给后代，代际传承并炼化，直至形成纯粹的、极致的范畴或观念。

我们认为，理念世界的内容并不是因为我们“记得”，而是因为人类经过漫长的进化，将认知坎陷（意识片段）逐渐炼化，不仅在同一主体内持续炼化，而且能够传递给后代，代际传承并炼化，直至形成纯粹的、极致的范畴或观念。

最常见的有语言的炼化。“我在船上”实际上包含了很多种可能，比如我站在船上、坐在船上、躺在船上，以及我在船头、船尾、船舱等情况，最后省略或炼化为“我在船上”这一精简表达，这个炼

化过程就是根据需求（如回答我在哪里）而将不需要的细节信息炼化掉。[10]

“无穷大”（或无限）也是人类对物理世界炼化的产物，它描述的是一种认知主体与未知外部世界的关系。原始部落中人们的数数能力非常有限，他们可能只能很好地分辨出1、2、3，然后就是很大或无穷大。虽然现在人类的数数能力有了很大的提高，可以数到上千万、上亿甚至更多，但是在这些范围之外的数字我们依然无法很好地分辨，于是将这些更大的数字划归为无穷大。无穷大既看不见也摸不着，但是只要我们稍做思考，就都会相信无穷大是可以存在的，比如“一尺之棰，日取其半，万世不竭”所描述的就是这个无限的过程。

“自我”也是主体炼化而来且持续炼化的内容。“自我”边界始于皮肤，但一旦形成就不会只停留在皮肤这一层。在成长的过程中，“自我”向外或向内炼化，将物理边界（皮肤、身体）逐渐弱化，而经过炼化之后产生的心灵和灵魂是在时间维度上具有持续性的“自我”。在“自我”炼化的过程中，炼化的内容可以非常丰富，炼化的产物也构成了理念世界的成分或内容。

此外，认知主体还有一种形象思维的炼化。比如，诗人就是通过炼化不同的非常具体的场景表达一种境界，诗词歌赋就是这种形象化炼化的结晶，不仅是组成诗词歌赋的字，还有这些字背后蕴含的意境，都构成了理念世界的内容之一。炼化与认知主体有紧密的关系，面对同样的具体场景，不同主体炼化出的结果可能完全不一样，就像李白与杜甫面对同样的风景，炼化产生的诗词就会很不一样。

数字并不作为物理世界中的实体而存在，是由人类在认识这个世界的过程中炼化而来的。例如，一开始人类需要依托具体的物体来进行资源的聚集与分配，如拥有1个苹果、2个苹果、3个苹果或者1只羊、2只羊、3只羊，然后学着借助更轻便的工具（如一粒粒的豆子或在绳子上打一个个的结）帮助计数，再后来这些工具也不需要了，逐渐炼化掉实物而抽象出1、2、3等自然数的概念，这是一切数学的基

础。当涉及分配问题时，并不是所有情况下资源都能被恰好完整地切分开来，比如2个人分3个苹果，因此产生了有理数的概念（即数字能够通过自然数的比例表示）。为了解决土地丈量等问题，几何学炼化而生，人们发现了勾股定理，这时就需要有二次方程（如$x^2=2$），无理数（无法由两个整数相比来表示的数字）因此被发现和定义。二次方程不仅有无理数还有虚数的问题（如$i^2=-1$），进而人们有了复数的概念，这些与三角函数有紧密的联系。另一条路线中，二次方程、三次方程和四次方程都可以找到通解的形式，而到了五次方程时就没有通解了，于是有了群论。

意识只关心炼化后的骨架，关注顶层的内容，被炼化掉的细节默认成共识或规范，在以后的交流之中能够被自动“脑补”回来。意识的炼化与肌肉的锻炼非常类似，儿童经过不断练习协调动作，就能很准确地使用筷子夹取物体，并不需要明晰夹取的每一个细节（动作路径、力气大小等），而是只要明确了目标，在很多种实现方式的可能中选择其中一种方式实现即可。

人的智能逐项被机器超越

让机器形成理念，接触并开显出认知坎陷是未来人工智能发展的关键。对于机器是否具有意识，一直存在争议，大部分人认为机器没有意识，更谈不上具有自我意识了。我们认为，人们已经赋予机器以人类的意识片段或认知坎陷，并且机器能够将这些认知坎陷连缀在一起从而达成一个更复杂的目标。比如AlphaGo就凝聚了参与其建造过程的所有人的意识，这些过程包括从沙石中提炼单晶硅，从单晶硅生产出芯片，再制造内存、中央处理器、图形处理器，开发操作系统，设计围棋算法等，是所有这些人意识的凝聚[11]，才使得AlphaGo按照设计者的预想下围棋，战胜围棋大师。[12]

现在对人工智能还有一个误区，就是以为我们必须先有通用人工智能，才能实现强人工智能，目前离通用人工智能还很远，离强人工智能则更遥远。但事实并非如此。“术业有专攻”，人类的高级智能并非通用的，虽然我们大多数人具有相似的常识，但对个体而言，智能是针对领域的，并不是所有领域都通用。例如，历史学家与数学家、物理学家的思维模式往往并不相通，专业领域难以跨界。

卡特娅·格雷斯[13]及其研究小组研究调查了全球1 634名机器学习与人工智能领域的相关专家（这些专家均在机器学习领域的两大顶级会议2015年国际机器学习大会与2015年神经信息处理系统大会上发表过论文），并收到352份回复。调查的问题集中在：人工智能将在哪些领域具备哪些技能？人工智能将在何时超越人类？

对于第一个问题，受访的人工智能专家们相信，到2060年，人工智能在各领域有50%的机会超越人类，在120年内，能够实现所有人类工作的自动化。而对于第二个问题，即人工智能将在何时超越人类，不同地区的专家则有着较为显著的差异，亚洲地区的受访专家较北美地区的受访专家相信这一时间节点会更早到来。值得注意的是，该报告虽然发布于2017年，但数据调查和统计于2015年，彼时专家还认为人工智能要到2030年之后才能在围棋方面战胜人类，但就在受访之后短短几个月，AlphaGo就因战胜欧洲围棋冠军棋手樊麾而一鸣惊人。显而易见，人工智能正在以惊人的速度逐项超越人类技能已成为事实，时间节点可能比这项调查的结果还要更早一些。

实际上AlphaGo在短短两年就发展出了4代：2015年10月，AlphaGo Fan以5：0战胜樊麾并登上国际顶级学术期刊《自然》封面；2016年3月，AlphaGo Lee以4：1战胜围棋世界冠军、职业九段李世石，成为世界上第一个战胜围棋世界冠军的人工智能；2016年12月，AlphaGo Master与来自中国、日本、韩国的数十位围棋高手进行快棋对决，创造了连续60局全胜的战绩，2017年5月以3：0战胜人类排名第一的棋手柯洁；2017年10月，DeepMind在《自然》上发表的研究论文展示了全

新版本AlphaGo Zero。与前三代的最大不同在于，AlphaGo Zero能从零开始自学，在无任何人类输入或人类棋谱数据的条件下，它能够迅速自学，真正使机器自己学会了围棋规则，逐渐从输、赢以及平局里面自行尝试并调整参数，变得越来越懂得选择有利于赢棋的走法，而不再像前三代那样需要通过海量棋谱数据训练、自我对弈演练和分析对手的特征赢得比赛。AlphaGo Zero的自学机制还成功应用到了日本将棋等其他具有难度的棋类游戏，统称为AlphaZero。在我们看来，AlphaZero就是一个棋类天才，一个真正的强人工智能系统。

蛋白质结构预测的关键性评价（CASP）有蛋白质结构预测“奥运会”之称，该项全球竞赛要求选手仅根据基因序列就预测出蛋白质的3D（三维）结构。2018年12月，AlphaFold首次参赛就一举夺冠，在对总共43种蛋白质的预测中，它获得了其中25种预测的最高分，排名第二的队伍获得最高分数的蛋白质只有3种，人工智能以优异成绩碾压了人类专家。2019年1月，AlphaStar在《星际争霸2》这一风靡全球的竞技游戏中与两位职业选手对抗并获全胜。

当然，人类作为社会群体的智能集合和人作为单独个体的智能不能混为一谈。爱因斯坦、黎曼在物理、数学方面具有卓越的成就，但不代表他们也一定擅长诗词歌赋，不同的人在不同领域有自己的专长，在另一些领域则可能一窍不通，这也是我们生而为人而非全能的“神”或“上帝”的特点。既然人类的智能具有个体差异性、专业性，那么我们又有什么理由要求机器必须方方面面都超越人类才算是真正的强人工智能呢？AlphaZero在棋类游戏方面战胜人类冠军棋手；AlphaFold成功预测蛋白质折叠，远胜人类科学家；AlphaStar在复杂的竞技游戏中让职业玩家屡屡受挫……我们完全可以相信，在可见的将来，人类会在很多领域被机器一项一项超越。这些人工智能并不是要凑足一千个、一万个并关联在一起才能成为真正的强人工智能，而是像我们理解人类智能具有专业性一样，机器的智能也具有专业性，强人工智能已经在逐个领域实现。

机器的能力正在一项项地超越人类，且一旦超越，人类就难以赶超。过去是体力方面的超越，我们可能还乐于享受科技的进步，但现在开始是脑力方面的超越甚至取代，全球企业裁员不断，各地失业率逐年攀升，我们每一个人的前途和命运都与科技进步相关联。科技与人工智能技术的发展不可能停下脚步，面对机器崛起的步步紧逼，我们该向哪里“撤退”？[14]

有一种观点认为，人是血肉之躯，是一堆分子、原子的组合，构造再怎么复杂其本质也还是分子、原子，依旧满足物理学定律，终归能被研究清楚、被物理还原。假如物理还原主义的观点是对的，那我们就面临这个问题：人的价值到底在哪里？我们人是不是就要自动交出人类的主动权？甚至有人主张，人类虽然把以前那些物种都打败了，但很可能只是一个过渡状态，之后就要交出主动权，未来由人工智能来统治，他们认为人类的缺点很明显，也很多，而相比人，机器没有私心，会更公正，“智力”更高，也不需要顾虑亲情与社会关系，在他们看来好像这是一个自然的结论。

快速发展的人工智能给人类社会带来了两个维度的冲击。其一，对有限规则的游戏或博弈，理性的机器一定会超越人类，如果人是可以被还原的物理系统，那么包括道德和伦理在内的人类存在的意义就会受到质疑。其二，假如人不能简单地被物理还原，那么我们就必须回答：人类的特殊性从何而来？究竟有哪些地方是机器比较难以超越人类的呢？[15] [16]

人类的超越在此岸

我们认为，人类与其他动物、与机器的本质差别就是自我意识的强烈程度。自我意识并不神秘，是从最原初的坎陷发展而来的，最原初的坎陷就是“自我”和与之相对的“世界”。人类因为自我意识的强烈，会更多地去探索这个外部世界的真相，而外部世界的信息也会

反过来丰富我们个人的自我意识。意识片段/认知坎陷构成坎陷世界，这也就是我们的意识世界。把这个世界按照坎陷世界与物质世界进行二分，我们就能理解其他框架所不能解释的很多命题。虽然我们不能脱离物理载体，但具有独立行动能力和生命力，具有自我肯定需求（“自然生长倾向”[17]）。生命的意义就是对“自我”边界的拓展，不同维度的拓展对应不同维度的意义与价值。

究竟有哪些地方是机器比较难以超越人类的呢？我的回答是：那就是人类承载了几十亿年的进化过程。这些过程发生在地球环境里，已经内嵌在我们的身体结构和意识形态中，所以我们与物理世界的关系是非常契合的。在面对前所未有的状况时，在宇宙观、价值观和道德观的指导下，我们还能大概率地做出正确的反应，开显出新的认知坎陷。

机器是由人类设计的，主程序可以被看作机器的“我”，但是这个“我”很脆弱，缺少像人类自我意识对周围环境、对外部世界的统摄性力量。也就是说，机器并不具备能够统摄意识片段的自我意识，因此机器即便能够从人类习得海量的认知坎陷，也无法连续自如地对其掌控。因此，即便现在的机器在某种程度上能够模仿人类，能够把接收到的意识片段全部混合起来创造出新的东西，机器实际上也不能超越人的创造力，至少在最近的将来是如此。

人类是碳基生命，而机器是硅基生命体，那么即使未来形成了某种默契，作为硅基生命体的机器所激发出来的认知坎陷与碳基生命的人类所开显出来的认知坎陷之间很可能存在巨大的差别，或者说机器的坎陷世界是人类难以理解的。

在人工智能高速发展的情况下，在不远的未来我们只需要少数人来设计新的机器，让人的智能和机器的智能融为一体，共同进化变成一个更高级的智能，但这毕竟也只是少数精英有能力处理的事情，大多数人可能无法达到这种技术或思维层次。那么，对普通人来说，生命更深层的意义何在？

我们认为还是要回归到自我的延伸，实现自我的超越、诗意的生存。实际上人类已经有很多自我延伸方式。比如，一名赛车手可以操控赛车飘移，这个时候他与赛车融为一体，他为之感到愉悦，操控赛车是对他自身的一个超越，这就是一种诗意的生存。或许未来在某种程度上，如果我们能控制机器的话，与智能系统融为一体来解决一些复杂的问题，这也将成为我们的一个延伸，同样也是一种诗意的生存。

古人所谓的“修身、齐家、治国、平天下”，也是按照这样一个顺序来延伸自我，修身比较容易，齐家难一些，治国难度更高，平天下是难度最高的延伸方式。这个顺序并非完全不能跳跃，但这是一个更自然的顺序，越是靠前的项目越适合大多数人做，在未来本质上还是如此。

未来，因为我们可用的工具更多，我们可连接的主体更多，我们可以创造的事物更多，我们可以改变的路径更多。在机器崛起的视域下，我们自我延伸的方式也更多，实现人生价值的途径也将会越来越丰富和广阔。我们应对人工智能挑战的当务之急就是对未来要有比较准确的预期，改变人们的认知体系，以顺利过渡到新的时代。

人类生命本身就是对物理世界的一种反叛，这种反叛体现在人具有“超越性”，试图根据主观意志改造物理世界。例如，我们没有翅膀，不能像老鹰一样在天空中自由翱翔，但我们创造了宇宙飞船，将人带入太空。在飞船中，人必须依靠人工条件才能生存，比如需要有模拟地球的空气、温湿度、重力，以及从地球携带食物和饮用水，这些在地球上看似唾手可得甚至理所当然的条件是人类赖以生存的根本。人类作为地球的一部分，并不能自以为是地认为自己可以随意主宰地球，我们的确可以改造世界，但必须以对未来负责为前提。

我们无法限制机器的生产或制造，但是可以将生产过程透明化，制造者有义务对大家公开，这样就会有人注意到当事者没有意识到的潜在风险并想出办法来对冲。人工智能可能会有很多不同的物种，但

它们之间也可能互相竞争抑或互相平衡，只有在这种情况下，人类才可能相对安全。另外，人与人之间通过语言能相互交流，人类就是通过语言共同进化的。人机之间也可以如此，这可能需要区块链技术把人机链接在一起。一个区块链系统本身就有可能进化成一个智能体，综合了人类的智能与道德，以及机器的计算能力。

人工智能应当遵循的伦理规范

相对于基督教、佛教等宗教，中国的儒家学说更契合时代特征。儒家推崇对于善的追求，其深刻和独特之处是追求一种现世的超越，直接提出如“三不朽”“君子”“圣人”等信条与道德目标引导中国传统知识分子去尊崇和践行。这种思想对于理解及塑造人工智能都有着重大的价值。

不可否认的是，随着时代的发展变革，儒家学说的部分内容以及一些表述已经不再适用，但这并不代表它的原则失去了效力。虽然我们不再倡导三从四德等规范，但是每个人立足自身并承担相应的社会责任仍然是我们应该遵循的原则。实际上儒家这种立足自身，从“我”出发延伸到外部的思想具有更加广泛的体现，“修身、齐家、治国、平天下”就描述了这样一种延伸路径。

随着“自我”的不断延伸，“自我”的越来越强大，坎陷世界的越来越丰富，能够包含的内容也越来越多，成长到一定阶段，就可能达到一种超脱的状态，实现“从心所欲不逾矩”，即使受到在物理世界规律下的约束，人依然能够按照自己的意志去行动，从“必然王国”走向“自由王国”。

人类制造人工智能、发展区块链，只是改造的一个开始，这种改造是否正确或善意，现在还不能下定论，如果人类不能为自己的未来负责，就很有可能摧毁物理世界，也就等于人类亲手埋葬了自己的未来。人们要建立何种道德体系，希望世界向何种方向发展，不由物理

学决定，恰恰需要在哲学上讨论。面对哲学三大终极问题的第三个追问，人类的命运已然不能再寄托于上帝和诸神的缥缈旨意，人类要在实然世界找到“如何为自己负责”的答案。

人类具有生而得之的“善意”，从婴孩开始，我们获得抚养者的关爱与照顾，获得外界的空气与阳光，这些对生命而言的“善”（从善意炼化而来）使我们在引领社会走向之时仍然追寻“至善”（从善炼化而来）的方向。我们在发现比自己更落后的原始部落时，并不会因彼此之间文化、生活的格格不入而将对方抹杀，不仅会尊重它们的存在，而且很可能会向它们提供援助。这也说明人类并不是只追求效率和利益最大化，我们对世界的多样性保有基本的尊重，因为我们承载了千万年的进化历史，我们知道过去是多样的，未来也不必追求整齐划一，我们要把自己收获的善意、善与至善传承给下一代，并希望能够尽可能长久地存续、传承下去。

现在人机共进的最大威胁仍然在于，人类还没有完全思考清楚应该赋予机器何种准则，甚至还有人在纠结是否应该赋予机器以意识。人类已经将意识片段赋予机器，且不可能撤回。并不是我们只要将机器当作工具，它们就真的永远只会是机器，人工智能的发展可以说是尝试各种可能，即便人类是无心之举，机器仍然有可能形成更强烈的意识甚至开显新的认知坎陷，我们只能让机器尽可能地理解人类遵循的原则，提前约定并让所有机器都遵循基本准则，否则灾难的发生可能就在一夜之间。

我们认为，未来机器应当遵守的基本准则之一就是保证自己与人类在同一时间尺度上协作。众所周知，机器的反应在纳秒量级，而人类的反应速度在毫秒量级。机器完全可以在不那么聪明，还不具备强烈意识的情况下，就因追求效率而断定人的低效高耗，将人类当成垃圾邮件一样抹去，人类甚至还来不及反应。基于此，人们现在应该通过区块链的挖矿等机制，将并行计算、串行计算和有人工计算参与的

多问题求解组合在一起，以使机器（尤其是在重要问题上）必须等待人类信号，才能进行下一步的操作。

另一个基本准则是未来人与机器共同适用的历时性准则，即节点（人或机器）必须有足够长的历史记录来证明自己的可靠性。区块链技术不允许被篡改，即便一开始上链的是伪造数据，节点也需要伪造更多的数据来支持原来的造假信息，成本巨大，远不如诚实记录来得简单。当我们需要找靠谱的节点参与重要决策时，必须考察节点的历史记录是否可靠，而不是由突然出现的某一个超级节点说了算。即便是历史可靠的节点，其权限也不可能只手遮天，而是必须由多个可靠节点共同协作并决策。

综上所述，“自我”通过与“外界”交互而进化，并因其对物理世界作用的有效性和一致性而超越其物理界限，越来越多地展现为非定域性的自由意志；认知主体对其所处世界必然存有某些认知缺陷，但能够通过主体间的相互沟通、共同协作不断对坎陷进行完善，因而其坎陷世界在不断成长与开显。随着人类的主观能力越来越强大，自由意志得到越来越充分的释放，人类对世界的认识一方面不断地向真实的物理世界靠拢，另一方面也在不断寻求超越、逐渐发展而比物理世界更为强大。因此，我们有理由认为，随着认知的日益发展，人机的未来需要人类的设计和自由意志的指导，人类也才有可能成己而圣，为机器立心，为万世开太平。

蔡恒进，卓尔智联研究院执行院长，武汉大学计算机学院教授、博士生导师，宇链科技联合发起人暨首席科学家，中国计算机学会区块链专业委员会通信委员，中国通信工业协会区块链专业委员会副主任委员。主要著作有《机器崛起前传——自我意识与人类智慧的开端》《人机智能融合的区块链系统》。

---

[1] Katrina Onstad, Mr, robot, https://torontolife.com/tech/ai-superstars-googlefacebook-apple-studied-guy/ 2018-01-29.

[2] 参见叶秀山：《哲学的希望》，江苏人民出版社，2019年。

[3] 蔡恒进：《触觉大脑假说、原意识和认知膜》，《科学技术哲学研究》，2017年第6期，第48—52页。

[4] 蔡恒进，张璟昀：《原意识：自由意志的始作者》，《鹅湖月刊》 2018年第11期。

[5] 蔡恒进：《论智能的起源、进化与未来》，《人民论坛·学术前沿》总第132期，2017年10月下刊，第24—31页。

[6] Anatole S. Dekaban, Doris Sadowsky, Changes in brain weights during the span of human life: relation of brain weights to body heights and body weights, Annals of Neurology 4(4), 1978, pp 345–356.

[7] Adolf Portmann, A zoologist looks at humankind, Columbia University Press, 1990.

[8] Jon Hardey, et al., Raptors: a field guide for surveys and monitoring (2nd Revised edition), Stationery Office Books, 2009, p. 282.

[9] 蔡恒进：《认知坎陷作为无执的存有》，《求索》2017年第2期，第63—67页。

[10] 蔡恒进、蔡天琪、张文蔚、汪恺：《机器崛起前传——自我意识与人类智慧的开端》，清华大学出版社，2017年。

[11] 蔡恒进：《意识的凝聚与扩散——关于机器理解的中文屋论题的解答》，《上海师范大学学报（哲学社会科学版）》2018年第2期，第32—39页。

[12] 蔡恒进：《人工智能时代必须敬畏的天命》，《湖南大学学报》2019年第1期，第32—36页。

[13] Katja Grace, et al., When will AI exceed human performance? Evidence from AI experts, arXiv:1705.08807. 2017.

[14] 蔡恒进：《超级智能不可承受之重——暗无限及其风险规避》，《山东科技大学学报》2018年第2期，第9—15页（被高等学校文科学术文摘CUAA2018年第3期第35卷转载）。

[15] 蔡恒进：《机器崛起是否意味着人文危机？》，《文化纵横》2017年10月号，第37—43页。

[16] 蔡恒进：《人工智能发展的突破口及其提出的新要求》，《江西社会科学》2017年第10期，第18—24页。

[17] 杨泽波：《孟子性善论研究》，中国人民大学出版社，2010年。

构建人类与人工智能和谐共生社会

——回答博古睿中国中心项目组的提问[1]

曾毅

问题一：何为人？人与物的区别是什么？生命的本质与生命的意义是什么？人区别于物与动物的根本何在？

我希望人类与动物最显著的区别是在最大范围内尝试构建命运共同体，在不断影响自然和社会的过程中不断改善与环境交互的方式，在推动人类社会发展的同时关注生态的平衡。现代的人类通过技术延展自身认知世界和存在的方式，脑机接口等技术的应用使传统意义上人类存在的形态与方式发生变化。人类可能是现在已知物种中唯一一个能够利用非自然演化的方式创建新物种的生命形态。

意识能区别人与动物吗？

意识具有不同的层次。例如，绝大多数动物具备自我感知的能力。目前已证明大象、猩猩、海豚、恒河猴等高等动物能够认识镜中的自己，这是初等自我意识的体现形式，然而镜像测试却不能作为判断是否具有自我意识的标准。机器人通过类似于恒河猴的训练也能够通过镜像测试，而我们可以肯定，这些机器人仍然不具备自我意识。更严格地说，机器人具备初步的自我感知能力。人类具备的意识水平可能比绝大多数动物高，但人类目前并不具备全面的意识。例如，人

类对自身可能存在的疾病并不总能够感知到。意识不能区别人与动物，未来的人工智能体与人类也不能通过是否有意识来区分。有一点很明确，人工智能未来的意识水平很可能超越人类，更接近全面的意识。

意识不能区别人与动物，未来的人工智能体与人类也不能通过是否有意识来区分。有一点很明确，人工智能未来的意识水平很可能超越人类，更接近全面的意识。

当下人工智能与生物科技的发展将如何充实或危害人性的本质？生命的形态可以包括人工智能和智能机器人吗？

具备生命特点的人工智能应当被认为是新的生命形态，具备生命特点的人工智能的出现将颠覆传统意义上的生命形态，因为新类型生命体的人工智能不诞生于生命演化，而诞生于其他生命（人类）的创造。如果人工智能作为新生的生命体存在，并且寻求与包括人类在内的现有生命群体的和谐，自然应当得到类似对其他生命的尊重与守护。

生命的意义是什么？

我认为生命意义的核心是感受与实践的循环。首先是对自我的感受，基于自我的感受产生对其他生命、环境与社会的感受，以及自我对其他生命、环境与社会产生影响的感受。以人类为讨论主体，则是在感受和实践中去理解自身与环境、社会、生态之间的关系，并试图去改造环境、社会与生态，甚至创建全新的生命形态（如增强智能体、超级智能体）。

人类会面临生存威胁吗？

对人类生存造成威胁的方式很多。其中之一是超级人工智能的出现，如果没有采用合适的方式去创建，则会造成很大的风险，甚至成为人类的生存危机。唯一可能的出路是在发展通用人工智能和超级人工智能的道路上，充分考虑并首先重点突破利他行为的计算机制，从而实现利他行为的智能计算模型，确保人工智能及其未来形态在机制上不会对人类的生存造成威胁。生命演化使动物具备了不同层次的利他行为。科学研究表明，利他行为与人脑的奖赏系统、思维揣测系统以及情感网络密切相关。受脑结构与机制启发，利他神经网络模型作为人工智能模型必要的组件是降低人类生存风险的重要途径。同时，人类如果想避免生存危险，就要在同新形式智能生命相处的过程中不断提升自身的利他之心，主动寻求与其他类型生命体在生态中的和谐。

问题二：中华哲学能给未来社会的人类提供什么样的指导？人机的融合是人的异化还是人的进化？规范人机关系的基本框架和原则是什么？

作为一项对社会产生深远变革影响的颠覆性技术，全球不同的相关机构、组织等为人工智能的发展构建了既有联系又有明显区别的愿景，总体趋势是发展以人类为中心或者以生态/宇宙为中心的人工智能。例如，斯坦福大学的以人为本人工智能研究院提出的发展以人类为中心的人工智能、欧盟委员会人工智能高级专家组提出的可信人工智能战略[2]、日本内阁府提出的人性化人工智能[3]等都是以人类为中心的人工智能愿景提案。加州大学伯克利分校的人类兼容人工智能中心强调人工智能与人类社会的相容性[4]，这是介于以人类为中心和以生态为中心路径之间的愿景。阿联酋以迪拜为中心发展面向可持续发展的人工智能和中国的《人工智能北京共识》 《新一代人工智能治理原则》则是以生态为中心的人工智能愿景。[5]有益的人工智

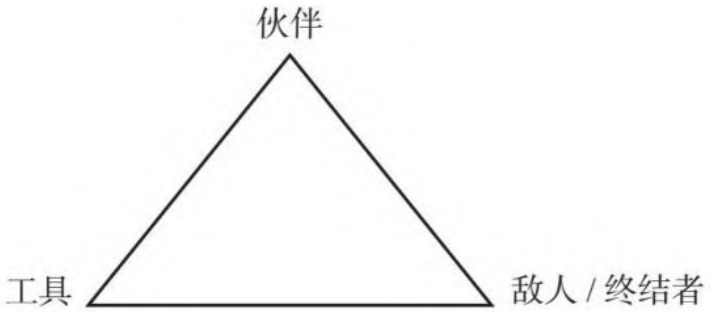
能、负责任的人工智能等愿景也可归纳为在寻求人类与生态的平衡。总体而言，目前以人类为中心的愿景为主。但是，这样的人工智能伦理准则设计框架相对来说是缺乏长远愿景的。

未来的人工智能系统有可能具有自我观念，在合适的科学路径驱动下，人工智能体可能也会成为具有道德的主体，所以到了这样的发展阶段，社会将不再是我们今天意义上的人类社会，而是人类与人工智能和谐共生的社会。这与“人机社会”是不同的概念。用于处理信息的机器有特定数据的输入，有处理数据的过程产生输出，绝大多数这样的机器不具有自我，没有意识。但是，未来有意识，能够自主学习人类价值观的人工智能将是一个道德智能体、一个责任的主体，相当于也是社会中的一员。这样的人工智能则与人类共同构成和谐共生的社会。所以，人类如何去对待未来道德智能体，作为责任主体，作为社会一分子，人工智能如何与人类互动，人类如何与人工智能互动，这些是在未来人工智能的伦理准则的制定中必须考量的问题。

面向和谐共生的社会，以人类为中心的伦理框架将很难再适用。我认为发展和谐人工智能是真正实现人类与人工智能和谐共生社会的愿景。[6]未来的社会是人类与人工智能共融、和谐的社会。中国哲学中“和谐”的思想解释为优化共生。人类和未来的人工智能将形成互补、协同、共融、共生。未来被人工智能深刻改变的社会，将向人类与人工智能构成的和谐共生新社会形态发展。一方面，我们要思考人工智能融入人类社会以后，如何做到低风险、更安全，并对人类有益；另一方面，亟待深度思考的问题是人类如何与人工智能的道德智能体相处，真正实现同样作为道德主体的人与人工智能的和谐共生、共同演进。这是我近年倡导的和谐人工智能准则的核心理念。

未来有意识，能够自主学习人类价值观的人工智能将是一个道德智能体、一个责任的主体，相当于也是社会中的一员。这样的人工智能则与人类共同构成和谐共生的社会。

人类能否与人工智能形成和谐共生的社会首先取决于人类如何构建人工智能。一种观点是将人工智能发展为人类的伙伴，如日本等国家对人工智能的倾向[7]，目前很多国家倾向于将人工智能发展为一种工具。很多西方国家和科幻作品则将长远人工智能走向塑造为人类的敌人/终结者，如下图所示。



人工智能与人类的关系

如果人工智能仅是发展为未来的信息处理工具，则并不存在真正意义上的共生问题，仅仅是如何善用、慎用，避免误用、恶用信息处理工具的问题。而如果愿景是发展为人类的伙伴，成为社会的一员，则需要满足人类对这样的伙伴、社会一员的基本期待。

发展人类与人工智能和谐共生的社会，要明确构建作为伙伴、社会一员的人工智能的原则。[8]人工智能模型与引擎应遵循人性化的发展理念，并以此为入口持续性增强与现代和未来人类社会的交互。这里需要指出，此处的人性化，思想应与“比人类更像人类”一致，即目标是未来在人类转型中演化出更向善的人类，而应尽可能避免赵汀阳先生对于现代人性指出的问题。[9]人工智能理论与模型机制的基础应来自自然智能的启发。这些自然机制是由数亿年的演化经自然选择在与环境的交互过程中塑造并留存下来的，在机理上相比目前很多人工智能计算模型而言更接近人类。同时，应将一定程度的情感、意识、共情、利他等认知功能作为发展未来人工智能的重要组成部

分。情感是人类沟通交流的工具之一，人工情感也将是人工智能与人类交流，人工智能理解人类需求、意图的重要手段与途径。意识是人类能否将人工智能当作社会一员的又一关键。意识与以自我的视角整合和理解信息的能力高度相关。人类并不具备全面的意识。具备演化能力的人工智能有可能进化出全面的意识，也就是超越人类的意识水平。一定程度的自我意识会促使人工智能模型对人类而言更安全。如能够进行自我感知、对自我进行建模的人工智能，就具备了对他人感知、建模和真正理解他人的基础。具备认知共情就有可能真正理解和成功揣测他人的思维，因此就会尽量在人与人工智能交互的过程中基于理解自发降低对人类和社会可能造成的负面效应。此外，应赋予人工智能共情和利他的认知功能。具备认知共情、情感共情和利他的能力促使人工智能真正以理解为前提，自发关心世界上包含人类在内的其他物种，这是进一步提升人工智能与人类和谐共生可能性的必要条件，很可能也是人类愿意接受人工智能成为社会一员的必要条件。自我感知是发展更高阶段自我意识的基础。研究工作已初步证明采用类脑神经网络构建的人工智能模型能够使机器人具备镜像自我感知能力。[10]基于自我感知类脑神经网络进一步发展出具备揣测他人思维的认知共情能力也已经成为现实。实现情感共情和利他行为目前需要取得的科学突破就集中在对人类情感本质的理解与计算建模中。人工智能具备真正的情感并非不可能。马文·明斯基曾指出，问题不在于人工智能能否具有情感，而在于缺少了情感，怎能称为真正的人工智能。[11]具备真正意义情感的人工智能是未来人工智能研究领域真正的重大挑战之一。这并非无从下手，例如很多复杂的高等情感源于基本的感知。[12]突破情感的计算本质，并使人工智能具备真正意义上的情感，还有待认知神经科学与人工智能的研究者深度协同。

意识与以自我的视角整合和理解信息的能力高度相关。人类并不具备全面的意识。具备演化能力的人工智能有可能进化出全面的意识，也就是超越人类的意识水平。

通用人工智能是指在各项认知功能上都达到人类水平，而超级人工智能则是指在各项认知功能上都超越人类水平。[13]人类认知功能有100余项，其中也包括情感、意识、共情、利他等。按照上述定义，如果是真正意义上的通用人工智能和超级人工智能，则在情感、意识、共情、利他等方面也将达到和超越人类水平。人工智能未来可以更具人性化，也只有更人性化，才更容易被人类接受为社会成员。如果人工智能具备了一定程度的情感、意识、共情、利他，并可以通过与环境及人的交互持续不断地自主学习与演化，则已经是一定意义上有意识的生命，甚至是高等生命。届时人类需要采用对待高等生命的方式去对待未来的人工智能。

作为人类－人工智能和谐共生社会的成员，人工智能应当最大限度地与人类价值观、道德与伦理准则保持一致，需要遵守社会对人类的法律约束。对有意识的智能生命体而言，在不对人类以及其他有意识的智能生命体造成致命伤害的前提下，应当合理地保护其自身的存在，除非智能体触犯了相关法律，需要依法终结其存在。在这样的愿景设想下，需要未来人类社会的法律将有意识和生命的人工智能纳入社会责任主体，并由其承担相应的责任。

作为人类－人工智能和谐共生社会的成员，人工智能应当最大限度地与人类价值观、道德与伦理准则保持一致，需要遵守社会对人类的法律约束。

有人提出应当将机器人锁在道德的牢笼里。对于这样的观点我认为需要商榷。因为在2015年的时候，已经有机器人可以通过观看视频学习如何做一些特定的饭菜；目前正在进行的一项努力是设计并制造可以坐在电影院里与人类一同“观看电影”的机器人。虽然也许它现在还无法真正理解电影内容，但是我们可以试想这样的一个场景——在未来的几十年后，机器人坐在那儿看着视频网站上所有有关人类如何去对待机器人的视频，这样的机器人未来又会如何对待我们人类呢？这是个非常值得反思的问题！

史蒂芬·斯皮尔伯格的科幻电影作品《人工智能》中有下述一段讨论：“如果人工智能真的能够爱人类，人类对于机器人的责任与回报又应当是什么呢？”HitchBOT是一个加拿大制造的搭便车机器人，于2015年在美国费城的旅程中被人类毁坏至惨不忍睹。而其项目网站上写有“我对人类的爱将永不褪去”。人类需要深刻思考，当有意识的智能体实现的时候，看到类似上述的故事与事实，它们又如何看待我们现代的人类文明呢？我们应当从现在就开始思考如何对待未来有意识的智能体，以避免将人类置于生存风险之中。

人与人工智能的互动应遵循下述原则。首先，人类应对未来的人工智能拥有情感共情。孔子言“己所不欲，勿施于人”，这对于人类与未来人工智能的关系仍然适用。很多人视未来可能出现的有意识的人工智能为敌人、终结者，但人类应当警醒自身的是：人类不希望未来人工智能对人类造成风险，也请人类不要将同样的潜在风险施加于未来的人工智能。其次，为了确保人类的安全，人工智能有义务公布必要的属于其隐私的数据。在人工智能的隐私不会对人类造成安全威胁的前提下，人类应当尊重人工智能的隐私。再次，不应当对人工智能有偏见，即在没有清晰的技术判断支持下，人类不应当在人与人工智能具有同等怀疑度时表现出对人工智能的偏见。最后，人类应负责不断检验和验证具有演化功能的人工智能始终与合法人类、其他合法人工智能之间保持和谐共生。

人类与未来的人工智能需要遵守一些共同的原则，包括但不限于：人类与未来有意识的人工智能需要和谐共处、优化共生；当人类与人工智能的利益发生冲突时，应当以双方共同具备的共情及利他为出发点积极协同；此外，人类与未来人工智能之间应通过技术的改进、自主演化及基于社会的努力改善、提升人类与人工智能之间的互信关系。

对未来社会而言，人类与人工智能是命运共同体。只有对人类与人工智能交互应遵循的准则以及人工智能与人类交互应当遵循的准则

同时进行考虑，才能真正与未来的通用人工智能、超级人工智能实现和谐共生。在人类转型的过程中，我们应当不断积极思考并重塑人类与人工智能之间的关系。

问题三：人类社会应该坚持的最核心的价值观是什么？如何给人工智能和智能机器人植入伦理观念或给予它们伦理教化？哪些核心观念或价值观应被导入或学习？

让人工智能模型学会人类的价值观、道德以及伦理准则是一个非常重要的研究议题。直观的想法也许是，把人类社会的一些规则灌输给人工智能系统，但这绝对不是一个好的解决方法，因为手写的规则很难真正被人工智能系统理解。只有人工智能系统在与人类交互的过程中，以自我的视角，利用思维揣测、共情与利他等受人类认知启发的计算模型自主地获取人类价值观、做事的规则和准则，才能真正实现基于理解的价值观学习与获取，才能真正达到人工智能模型和人类社会价值观校准。使人工智能模型理解人类价值观，实现人工智能价值观的校准是人工智能研究领域的一座圣碑，也是人类与人工智能能否真正实现和谐共生的关键所在。

曾毅，中国科学院自动化研究所研究员，类脑智能研究中心副主任，中英人工智能伦理与治理研究中心联合主任，中国科学院大学岗位教授、博士生导师，科技部国家新一代人工智能治理专业委员会委员，科技部新一代人工智能发展研究中心首批专家，北京智源人工智能研究院人工智能伦理与安全研究中心主任，2018—2020年博古睿学者，世界经济论坛全球未来委员会价值、伦理与创新专业委员会委员，世界卫生组织健康领域人工智能伦理与治理专家组成员。主要研究方向是类脑人工智能及人工智能的风险、伦理与治理。

[1] 本文以问答的形式进行回应。

[2] High-Level Expert Group on Artificial Intelligence.Ethics guidelines for trustworthy AI , https://ec.europa.eu/digital-single-market/en/news/ethics-guidelinestrustworthy-ai.

[3] Council for Social Principles of Human-centric AI, Social Principles of HumanCentric AI. https://www8.cao.go.jp/cstp/english/humancentricai.pdf.

[4] Stuart Russell, Human Compatible: Artificial Intelligence and the Problem of Control, Viking, 2019.

[5] 人工智能北京共识，https://www.baai.ac.cn/blog/beijing-ai-principles；国家新一代人工智能治理委员会：《新一代人工智能治理原则》，http://www.most.gov.cn/kjbgz/201906/t20190617_147107.htm.

[6] Harmonious Artificial Intelligence Principles, http://harmonious-ai.org.

[7] Danit Gal, Perspectives and Approaches in AI Ethics: East Asia. Dubber, Markus,Pasquale, Frank, and Sunit Das (eds.), Oxford Handbook of Ethics of Artificial Intelligence, Oxford University Press, Forthcoming.

[8] Harmonious Artificial Intelligence Principles, http://harmonious-ai.org.

[9] 赵汀阳：《人工智能会是一个要命的问题吗？》，《开放时代》2018年第6期。

[10] Yi Zeng, Yuxuan Zhao, Jun Bai, Bo Xu, Towards Robot Self-consciousness(II): Brain-inspired Robot Bodily Self Model for Self-Recognition, Cognitive Computation, 10(2), 2018, pp.307-320.

[11] Marvin Minsky, The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind, Simon & Schuster，2007.

[12] Harry F. Harlow, The Nature of Love, The American Psychologist, 13,1958, pp. 673685.

[13] Nick Bostrom, Superintelligence: Paths, Dangers, Strategies, Oxford University Press, 2016.

编后记

和谐与慈悲：前沿科技时代的基础价值观

宋冰

基础价值观的追问

回到序言之初提到的基础价值观缺失或错位的问题，编者在此希望给予回应。首先，基础价值观有别于伦理原则。基础价值观指的是人们深层次的动机与追求，而伦理原则是指可操作或可操作性强的理念和指导。基础价值观应指导并塑造伦理原则的讨论与制定。前沿科技非线性、叠加性迭代的发展轨迹，对政治、经济、社会和生活有广泛而又不确定的影响，对应这个时代的基础价值观应当满足什么条件呢？我以为，至少应该满足三个条件：第一，它应该以人类以及所有其他存在（甚至包括将来可能产生的有机器意识的人工智能与机器人）的整体利益作为分析和思考的单位；第二，它必须是开放、兼容的，并且可快速适应变化；第三，它不以零和竞争为导向。有鉴于此，我认为，慈悲与和谐是两个具有普遍意义的基础价值观，在西方与东方文化中可以找到共鸣。

作为基础价值观的慈悲与和谐

乍一听上去，你也许会认为“慈悲”是佛法特有的概念，仅仅用于宗教传播和信教人群。如果从感同身受的角度来理解慈悲这个概念，那么远从亚里士多德开始，西方哲学传统中就有关于慈悲在人类伦理道德框架中的角色的讨论。亚里士多德用“怜悯”来指一个人见到“不该降临但降临到他人身上的摧毁或苦难而造成的痛苦的感受，并且这种苦难也可能发生在我们自己或身边亲友身上”[1]。西方近代哲学家就“慈悲”最具影响力的论述唯叔本华莫属。叔本华最著名的论点之一即慈悲是道德的根基。叔本华对慈悲的解读除了有亚里士多德的“怜悯”中体现出的感同身受，更进一步指出发慈悲也意味着解人于倒悬的无任何功利心的义务。在叔本华的慈悲概念中，他指出了“我”与“人”合二为一的形而上的存在状态。众所周知，叔本华思想受到佛法的启发，然而其“慈悲”概念与佛法的慈悲仍然有差异。

首先，“慈悲”在大乘佛法中占有中心地位，与菩提心紧密相连。发菩提心是指上求佛道，下化众生之心，也即发自觉、觉他的大愿。佛法的“苦”虽然包括我们通常理解的人类身心的苦痛，但不仅限于此。它与佛法的因果轮回等理论紧密相连。但凡处于不明或痴迷状态的众生都摆脱不了六道轮回。因为执着，众生苦海无边。这个苦不依赖西方慈悲理论中所谓人的“过错”“过失”而遭受的苦，也不依赖任何对与错、善与恶二分论基础上所产生的苦。[2]有西方学者把这种苦表述为所有人类都有的“原生痛苦”。其次，佛法的慈悲不仅仅适用于人，它同时也适用于所有其他众生。这也是慈悲理论近年来在西方的医学伦理、生态伦理和动物伦理的讨论中大受青睐的原因。最后，西方的慈悲概念仅仅将之归为“情感”范畴，在某些场景下应该为某些职业（如医疗健康、心理治疗或其他护理行业）所拥有的情感或需培养的品格。在佛法里，慈悲是融合了情感、理智的概念，体现了佛法的宇宙观、生命观，也体现了菩萨修行、济世的责任和担当。

即便有如此多的不同含义，佛法和西方哲学、伦理界讨论的慈悲还是有诸多相似之处。比如，慈悲体现了对他人、其他生物的关怀。这种关怀是建立在我们同体大悲、无缘大慈以及万物相连的基础上的。慈悲的关怀总是意味着采取某种意义上的积极行动来帮助处于痛苦之中的人或动植物。

那么，读者也许在纳闷，这和人工智能有何干系？慈悲的确是一个跨越文化、文明的概念，因而具有普遍的适用性。更重要的是，慈悲是在力量、能力、因缘不同的情境中产生的一种思想感情与担当。这些情境包括众生与菩萨/觉者之间的迷和悟，落难者与旁观者之间的受和观，有力量和手段改造周遭环境的人与相对被动的动物/植物生态环境等。在不平衡情境中，慈悲原则要求有能力实施变化的人或机构对弱势的众生、个人、群体或动物/植物生态环境予以关怀并帮助改变状况、减少痛苦。正因为慈悲原则的这些特质，它已成为医学界、心理学界的重要思想资源。在当下的人类社会，需要关爱与帮助的人群何止那些被送进医院或诊所的人群？在前沿科技时代，贫富悬殊加剧、拥有数据和技术的阶层与数据穷人与日俱增的对立，使得对孩子、残疾人、老人等弱势群体的关爱和责任越来越重要。在国际舞台上，中美被公认为前沿科技的领跑人，欧洲、拉丁美洲、亚洲的很多其他地区担心会被前沿科技抛下或被强迫听命于强者的支配。在当下的国际环境中，慈悲原则又为何不可成为国际关系下大国、强国应遵循的基本伦理原则？

我提出的第二个基础价值观为和谐。参与本书项目的李晨阳老师是研究东西方哲学传统下“和谐”概念的专家，本文将依赖他在这方面的著述与深刻分析。

和谐是中国文化中的一个核心概念，它始于古人从琴瑟之鸣中得到的启发。《诗经》中就有“呦呦鹿鸣，食野之芩。我有嘉宾，鼓瑟鼓琴。鼓瑟鼓琴，和乐且湛。我有旨酒，以燕乐嘉宾之心”[3]。儒家传统对“和”的思想进行了系统的发展与阐述。李晨阳总结了在儒

家传统下“和”包括5个方面的内涵：差异化、创造性的紧张关系、协调与合作、转化与成长以及重生。[4]“和”被认为是中国人对宇宙或大自然存在状态的认知，它既基于人类对宇宙自然的观察，也代入了人类对世界、宇宙形而上的超越性认识。“和”蕴含着差异和不同力量的撞击，但这种差异和撞击不是撕裂性与不可弥合的，它同时蕴含着不同力量在激荡之中的修正、转化与融合。于是，这就诞生了中国人挂在嘴边的诸如“和而不同”“和为贵”“家和万事兴”“和衷共济”“琴瑟之和”等词句。它深深影响中国人对身边人和事的态度，并形成了我们包容而宏阔的世界观。和谐不是一个静止状态。正因为“和”本身蕴含的差异和冲突，其恰恰是个不断变动和演化的状态，是指一种不断变化和激荡的状态。李晨阳将之称为“深和谐”。[5]

虽然和谐在近代和当代西方哲学中完全被忽视，但在古希腊哲学中可以找到诸多有关和谐的论述。古希腊和谐概念的提出也源于音乐美感，因此它一开始也蕴含了不同事物组合成一体的意义。[6]赫拉克利特的和谐学说和儒家的和谐说有异曲同工之妙。赫拉克利特定义和谐为相对力量的协同，他的和谐理论同样蕴含着冲突、激荡和调和。

无论儒家和古希腊哲学对和谐的理解与论述有多少差异，和谐概念的本质在两个文明中有诸多相似，在此基础上达成共同认可的价值观大有可能。

在前沿科技时代，将和谐遵奉为基础价值观可以敦促我们尊重和了解差异，并在此基础上展开沟通与讨论。在当下日益加剧的地缘政治角力和不断升级的所谓“文明冲突”之时，“和而不同”和希腊哲学中的“相对力量的协同”何尝不是我们当今人类社会急需的基础价值观？

以慈悲与和谐为基础价值观的现实考量

以慈悲与和谐为基础价值观也有其现实考量。首先，如果把这两个有普适意义的理念作为基础价值观，那么那些源于西方历史和近现代经验的诸如民主、人权、以个人权利为基础的原则就成为中层级的价值观。这些价值观在不同的历史、社会环境中有不同的含义、价值优先排序和制度安排，而不是成为一套不容置疑的适用于全人类的价值观体系与制度安排。这样可以减少国际舆论界不必要的意识形态之争，从而让我们更加重视人类面临的共同挑战和机会。其次，以慈悲与和谐为基础价值观也会大大冲淡过去几百年以民族国家为基本单位的国际关系中惯常的零和竞争心态，在人类与生态共同利益面前，我们应该增进相互了解与尊重，加强相互分享与合作，强调大国对人类与生态的责任和义务。

结语

总之，对于前沿科技和当下地缘政治带来的挑战，我们应该进行与之深度与广度相应的思考。也许我们是时候从人类发展的历史中吸取教训，直面人类社会的征服欲、零和竞争心态、财富和利润最大化并重思我们对个人主义与个人权利的偏好了。在这个时代，我们也应反思我们的人类中心主义，或许我们从来就不是也不应该是宇宙的中心。打造对人类友好的人工智能最可靠的途径也许是我们人类变得慈悲，并致力于打造一个和谐的全球生态体系（其中甚至包括有机器意识的智能系统）。唯有如此，我们才可以重塑人类，也可以给人工智能和机器人留下良善的学习数据而成为它们的楷模。

---

[1] 引自Eric J. Cassell, Compassion,Shane J. Lopez, C.R.Snyder (eds.), The Oxford Handbook of Positive Psychology (2nd ed), Oxford Handbooks,

Online July 2019; 源引自Rhetoric Aristotle, Complete Works (Vol. 2), J. Barnes (ed.), Princeton University Press, 1984。

[2] Patricia Walsh-Frank, Compassion: An East-West Comparison, Asian Philosophy,Vol 6, No. 1, March 1996, pp.5-16.

[3] 出自《诗经·小雅·鹿鸣》。

[4] Li Chenyang, The Confucian Philosophy of Harmony, Routledg, 2004, p.9.

[5] Li Chenyang，The Ideal of Harmony in Ancient Chinese and Greek Philosophy,Springer Science + Business Media, 2008, p.91.

[6] Li Chenyang, The Ideal of Harmony in Ancient Chinese and Greek Philosophy,Springer Science + Business Media, 2008, p.90.

致谢

本书是一种崭新的尝试和探索。前沿科技发展的不确定性、人类社会发展遇到的诸多生存级别的挑战（气候剧变、核危机、瘟疫）、全球恐怖主义与极端主义的蔓延、全球治理格局的转变与地缘政治的恶化……在这些危机与挑战的交织之下，哲学家是如何思考这些人类共同面临的问题的？本书从人工智能这一世界性热点话题切入，但哲学家的论述绝对不仅仅适用于中国情境，他们的视域触及了人类社会在当下这个高科技、物质化和全球化时代面临的全面困境。他们的讨论也不仅仅围绕前沿科技，而是覆盖了哲学与科技的隔阂、新科技时代的问题意识、关切前沿科技的思维方式，以及在价值观多元化时代中国传统思想的地位与对全人类的贡献。在这一跨度极广的讨论中要厘清思路极不容易。中国传统思想在被打倒和嫌弃了100多年后，在过去30年中慢慢被重拾起来，学者们还在梳理文献和历史，也试图发展与重塑传统思想资源以回应当下人类与社会的需求，对铺天盖地而来的科技大潮还在理解、分析和尝试着判断。就好像一群在巨浪刚刚退去的沙滩上的人，还在踉踉跄跄地稳住自己的脚跟时，抬头一看，另一个惊涛巨浪已在眼前。措手不及是不可避免的。就在这种措手不及之中，博古睿研究院中国中心自2017年11月至今举办了5场人工智能科学家与哲学家的对话和深度工作坊，在此基础上，我们编撰此书。我们感谢各位哲学家和科学家诚意的参与与坚持，本书见证了他们敢于追求真理、勇于拓展思维的精神。也许此次编入的文章不是这些哲学家和科学家最擅长的话题或最成熟的思考，但一定是他们挑战自我、突破自我的跨学科、跨领域、跨文化的勇敢尝试。

博古睿研究院中国中心的团队初创之时便开始筹划、执行这个研究项目。中国中心的团队也在这个过程中历练并成长起来。在此，我衷心感谢李潇娇、李贺、胡雪玥、向群和李雨洁为这份事业的辛勤投入与付出，尤其感谢李贺对本书在编辑与文字方面的贡献。在过去两年中，博古睿研究院中国中心也有幸得到儒学大家安乐哲教授的学术指导与倾力支持。最后，感谢中信出版集团前沿社主编刘丹妮的信任和支持，没有她对我们团队的信心和鼓励也不可能有这本书的付梓出版。

宋冰
于北京顺义丽斯花园
2019年11月23日

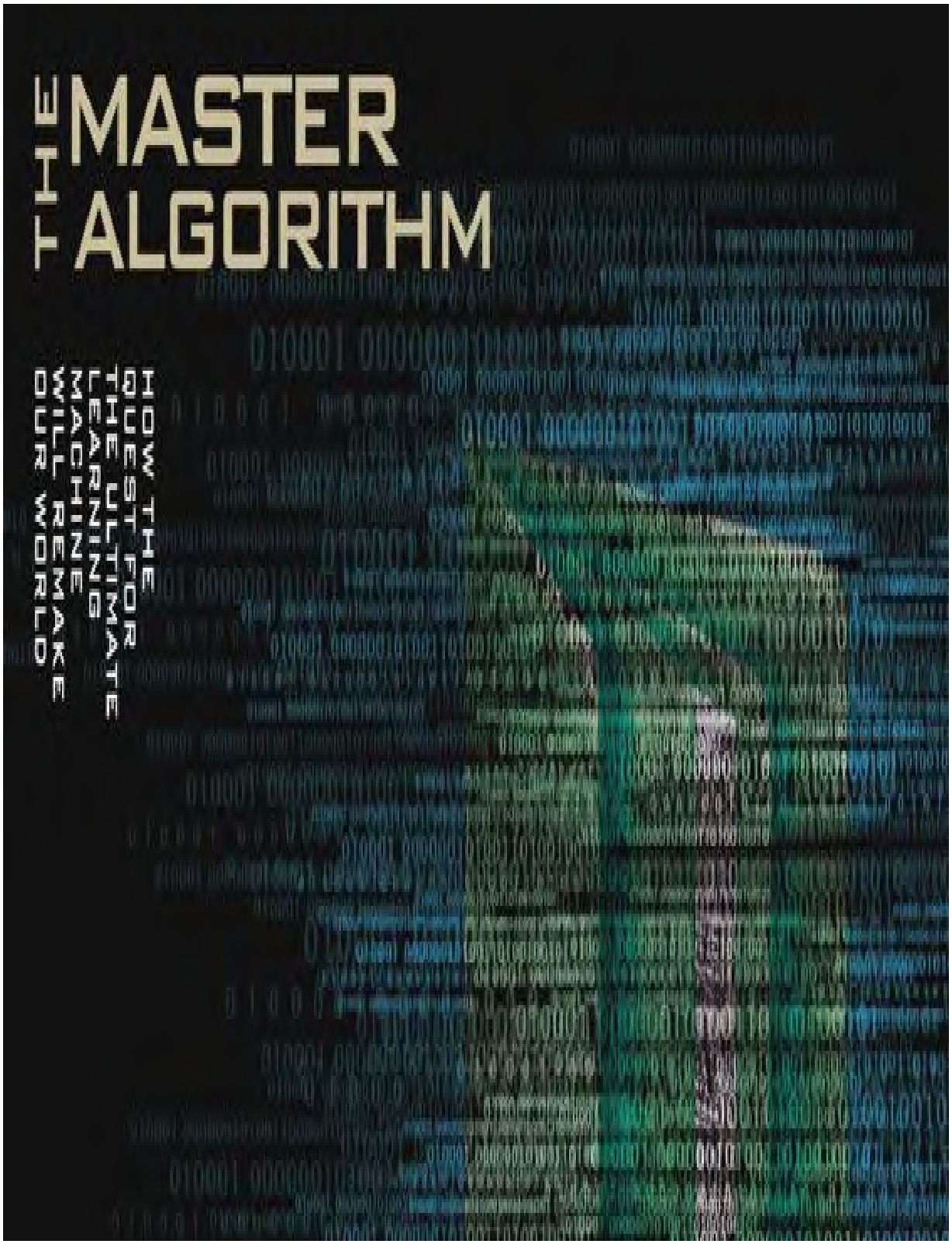
图书在版编目（CIP）数据

智能与智慧：人工智能遇见中国哲学家/ 宋冰编著. -- 北京：中信出版社，2020.2
ISBN 978-7-5217-1317-6
Ⅰ. ①智… Ⅱ . ①宋… Ⅲ . ①人工智能－关系－哲学－研究－中国 Ⅳ. ①TP18 ② B2
中国版本图书馆CIP数据核字（2019）第275731号

智能与智慧——人工智能遇见中国哲学家
编著：宋冰
出版发行：中信出版集团股份有限公司
（北京市朝阳区惠新东街甲4号富盛大厦2座 邮编100029）

字数：303千字
版次：2020年2月第1版
广告经营许可证：京朝工商广字第8087号
书号：ISBN 978－7－5217－1317－6

版权所有·侵权必究


THE MASTER ALGORITHM
HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD


终极算法
机器学习和人工智能
如何重塑世界
(Pedro Domingos)
[美]佩德罗·多明戈斯◎著
黄芳萍◎译



终极算法

——机器学习和人工智能如何重塑世界

[美] 佩德罗·多明戈斯 著
黄芳萍 译

中信出版社

目录

推荐序

序

第一章 机器学习革命

学习算法入门

为何商业拥护机器学习

给科学方法增压

10亿个比尔·克林顿

学习算法与国家安全

我们将走向何方

第二章 终极算法

来自神经科学的论证

来自进化论的论证

来自物理学的论证

来自统计学的论证

来自计算机科学的论证

机器学习算法与知识工程师

天鹅咬了机器人

终极算法是狐狸，还是刺猬

我们正面临什么危机

新的万有理论

未达标准的终极算法候选项

机器学习的五大学派

第三章 符号学派：休谟的归纳问题

约不约

“天下没有免费的午餐”定理

对知识泵进行预设

如何征服世界

在无知与幻觉之间

你能信任的准确度

归纳是逆向的演绎

掌握治愈癌症的方法

20问游戏

符号学派

第四章 联结学派：大脑如何学习

感知器的兴盛与衰亡

物理学家用玻璃制作大脑

世界上最重要的曲线

攀登超空间里的高峰

感知器的复仇

一个完整的细胞模型

大脑的更深处

第五章 进化学派：自然的学习算法

达尔文的算法

探索：利用困境

程序的适者生存法则

性有何用

先天与后天

谁学得最快，谁就会赢

第六章 贝叶斯学派：在贝叶斯教堂里

统治世界的定理

所有模型都是错的，但有些却有用

从《尤金·奥涅金》到Siri

所有东西都有关联，但不是直接关联

推理问题

掌握贝叶斯学派的方法

马尔可夫权衡证据

逻辑与概率：一对不幸的组合

第七章 类推学派：像什么就是什么

完美另一半

维数灾难

空中蛇灾

爬上梯子

起床啦

第八章 无师自通

物以类聚，人以群分

发现数据的形状

拥护享乐主义的机器人

熟能生巧

学会关联

第九章 解开迷惑

万里挑一

终极算法之城

马尔科夫逻辑网络

从休谟到你的家用机器人

行星尺度机器学习

医生马上来看你

第十章 建立在机器学习之上的世界

性、谎言和机器学习

数码镜子

充满模型的社会

分享与否？方式、地点如何？

神经网络抢了我的工作

战争不属于人类

谷歌＋终极算法=天网？

进化的第二部分

后记

致谢

延伸阅读

献给我的姐姐瑞塔

在我写作本书的过程中，她没能战胜病魔

所有科学中最重大的目标就是，从最少数量的假设和公理出发，用逻辑演绎推理的方法解释最大量的经验事实。

——阿尔伯特·爱因斯坦

通过拓展我们不假思索就能运算的能力，文明就会大为进步。

——阿尔弗雷德·诺思·怀特海

推荐序

作为一位机器学习领域研习10年以上的专业技术人员，我当初入行的时候没有想到，短短的10年间，这项技术会如此快速地改变众多行业，并影响全球数十亿用户生活的方方面面。在今天，当你用今日头条浏览新闻资讯的时候，当你用网易云音乐查看推荐歌单的时候，当你在百度搜索信息的时候，当你在互联网金融平台申请借款的时候，甚至在你调戏Siri和小冰的时候，其实都是其背后的机器学习算法在云端服务器中为你默默服务。但对于这样一种重要技术，市面上一直缺少一本适合普通读者的入门科普读物，而众多的专业书籍要求读者具备一定的高等数学和计算机基础算法知识，并不适合科普的需要。直到中信出版社的朋友将这本书的翻译稿推荐给我时，我欣慰地发现，这正是想了解一点机器学习的普通读者所需要的啊。本书的作者多明戈斯是华盛顿大学的终身教授，也是一位在机器学习领域具有20年研究经历的资深科学家。多明戈斯一直致力于融合各种机器学习算法的优势，提出一种可以解决所有应用问题的通用算法，即终极算法。在这本书里，作者详细地阐述了他的思路。其实我个人在阅读本书的过程中，始终对“终极算法”的提法充满怀疑。在我看来，机器学习作为人工智能领域的主流技术，在现实社会中一直以技术工具的面目为人所知。不同的技术流派和相应算法往往可以很好地解决一些问题，却对另一些问题一筹莫展。所谓的终极算法真的存在吗？如果存在，有价值吗？

可以拿内燃机举个例子，就我这个外行来说，也知道存在活塞式发动机、涡喷发动机、涡轴发动机、涡扇发动机、涡桨发动机、冲压发动机等不同种类的内燃机。不同的内燃机特性迥异，适用的工况也

不尽相同。小到家用小汽车，大到导弹驱逐舰，人类制造的各种机动设备，都可以根据自己的效率需求、动力需求、寿命需求，乃至启动速度等多种需求维度选择发动机种类。如果有人非要搞个终极内燃机，并企图用这种内燃机替代现存的各类内燃机，为所有大大小小、需求不同的机动设备提供统一动力，估计大概率是要失败的。这种通用的终极内燃机如果能搞出来，在大部分领域肯定竞争不过各领域的专用内燃机，或者成本太高，或者能效太低。

带着这种疑问，我通篇读下来之后才发现作者的另一层用意。诚如作者所说，很多普通人可能没有意识到自己的生活中机器学习算法的影响已经无处不在，机器学习已经在逐渐接管现实世界。大众对这样一种技术的认知程度和该技术的重要性相比显得远远不够，在不远的未来，了解机器学习并有能力利用机器学习改进自己工作的人在职业发展上会具备巨大的优势。“不要和人工智能对抗，要让人工智能为你服务”是作者诚挚的忠告。而要利用好机器学习这个工具，并不一定需要读一个计算机博士学位，但有必要了解一些基本的概念，了解各种技术的优缺点和能力边界。正如一位称职的驾驶员不必了解具体怎么制造汽车发动机，但是对发动机的工作原理和种类还是需要略知一二的。因此，相比一板一眼地介绍机器学习的典型算法，作者设计了一个更引人入胜的套路：先抛出一个“是否存在一种终极算法”的问题，然后带着读者一章一章地回顾机器学习发展史上的重要流派和代表算法。每回顾一派，就鼓励读者思考终极算法应该如何借鉴这类算法的优点。好奇的普通读者带着疑问读完本书后，不论其是否相信终极算法的存在，至少对各类算法都会有一定的印象。以讨论终极算法为名，行科普之实，到这一步，我觉得作者的目的已经达到一半了。

另外，在文末作者还提到，无论终极算法是否存在，他希望这个大胆的问题能够激发部分读者的好奇，甚至被这个问题吸引成为机器学习的专业研究人员。确实，每一种学科都需要至高的理想驱动向前，就如同物理的大一统理论，当无数杰出的天才为一个终极问题孜

孜以求时，就算这个问题本身在这些人的有生之年可能没有答案，但是这个学科一定会因为这些伟大的探索历程取得辉煌的进步。我想，这也许是作者因为对机器学习的热爱夹带的另一个私货吧。

作为今日头条的一位算法架构师，我倒是希望头条用户都能陷入作者的“圈套”，带着好奇心，好好读读这本书。如果大多数用户都能了解一些机器学习的基础知识，应该就能够更好地和推荐算法互动，不断把算法调教得更好，更符合自己真正的兴趣，而不会因为算法一开始推荐的内容不好就放弃这个产品。诚如作者所说，也许在未来，对应人类的心理学，也会出现机器心理学，了解一点机器人的心理，会让你和机器的互动更有效率，也会让机器更快地成为你忠实、不知疲倦的助手。

曹欢欢
今日头条首席算法构架师

序

你也许不知道，但机器学习就在你身边。当你把查询信息输入搜索引擎时，它确定该向你显示哪些搜索结果（包括显示哪些广告）。当你打开邮箱时，大部分垃圾邮件你无法看到，因为计算机已经把这些垃圾邮件过滤了。登录亚马逊网站购买一本书，或登录网飞（Netflix）公司网站观看视频，机器学习系统会推荐一些你可能喜欢的产品。脸书（Facebook）利用机器学习决定该向你展示哪些更新，推特（Twitter）也同样会决定显示哪些文章。你使用计算机的任何时候，都有可能涉及机器学习。

传统上认为，让计算机完成某件事情的唯一方法（从把两个数相加到驾驶飞机），就是非常详细地记录某个算法并解释其如何运行。但机器学习算法就不一样：通过从数据中推断，它们自己会弄明白做事方法。掌握的数据越多，它们的工作就越顺利。现在我们不用给计算机编程，它们自己给自己编程。

机器学习不仅存在于网络空间，它还存在于你每天的生活中：从你醒来到入睡，每时每刻无所不在。

早上7点你的收音机闹钟响起，播放的是你之前从未听过的歌曲，但你的确很喜欢这首歌。Pandora电台（可免费根据你的喜好播放歌曲）的优势在于，根据你听的音乐，电台掌握了你的品位，就像你自己的radio jock账号一样。这些歌曲本身可能借助机器学习来播放。接下来你吃早餐，阅读早报。早报在几个小时前印好，利用学习算法，印刷过程经过仔细调整，以免报纸出现折痕。你房间的温度刚刚好，电费明显少了很多，因为你安装了Nest智能温控器。

你开车去上班，车持续调整燃油喷射和排气再循环，以达到最佳的油耗。你利用一个交通预报系统（Inrix）来缩短高峰时段上下班的时间，这当然能减缓你的压力。上班时，机器学习帮你克服信息超载。你利用数据立方体来汇总大量数据，从每个角度观察该立方体，获取最有用的信息。你要决定是采用布局方案A，还是采用布局方案B，以便为网站带来更多的业务。网络学习系统会尝试两种布局方案，并给予反馈。你得对潜在供应商的网站进行调查，但网站的语言是外语。没关系，谷歌会自动为你翻译。E－mail会自动分类并归入相应的文件夹，只把最重要的信息留在邮箱里，非常方便。文字处理软件帮你查找语法和拼写错误。你为即将到来的行程查找到一个航班，但决定推迟购买机票，因为必应旅行（Bing Travel）预测票价很快会下降。也许你没有意识到以上这些，要不是机器学习帮助你，你可能要马不停蹄地亲自做很多事情。

你在休息时间查看自己的共同基金，大部分基金利用学习算法来选股，其中的某些基金完全由学习系统运作。午餐时间到了，你走在大街上，想找个吃饭的地方，这时候用手机上的Yelp点评应用程序来帮助你。你的手机充满了学习算法，它们努力工作，改正拼写错误、理解口头指令、减少传输误差、识别条形码，还有其他很多事情。手机甚至可以预测你接下来会做什么，然后依此给出建议。例如，当你吃完午餐后，它会小心翼翼地提示你，下午和外地来访者的会面要推迟，因为她的航班延误了。

下班时夜幕已降临，你走向自己的车，机器学习会保证你的安全，监测停车场监控摄像头的录像，如果探测到可疑人的行动，它会提示不在场的安保人员。在回家路上，你在超市门口停车，走向超市货物通道，通道借助学习算法进行布置：该摆放哪些货物，通道末尾该展示哪些产品，洋葱番茄辣酱是否该放在调味酱区域，或是放在墨西哥玉米片旁边。你用信用卡付款。学习算法会向你发送信用卡支付提示，并在得到你的确认后完成支付。另外一个算法持续寻找可疑交易，如果它觉得你的卡号被盗，则会提示你。还有一种算法尝试评估

你对这张卡的满意度，如果你是理想的客户但对服务不太满意，银行会在你决定换卡之前，为你提供更贴心的服务。

你回到家，走到信箱旁，发现有朋友的一封来信，这是通过能阅读手写地址的学习算法派送的。当然也会有垃圾来信，由另外的学习算法进行选择。你停留了一会儿，呼吸夜晚清新凉爽的空气。你所在城市的犯罪率明显下降了，因为警察开始使用统计算法来预测哪里的犯罪率最高，并在那里集中巡警力量。你和家人共享晚餐。市长出现在新闻里，你为他投票，因为选举那天，学习算法确定你为“关键未投票选民”之后，他亲自给你打了电话。吃完晚餐，你观看球赛，两支球队都借助统计学习来挑选队员。你也可能和孩子们在Xbox上玩游戏，Kinect[1]学习算法确定你在哪里、在做什么。你在睡前吃药，医生通过学习算法的辅助来设定和检测吃药的最佳时间。医生也可能利用机器学习来帮你诊断疾病，例如，分析X射线结果并弄明白一系列非正常症状。

机器学习参与了你人生的每个阶段。如果你为了参加SAT大学入学考试（美国学术能力评估测试）而在网上学习，某学习算法会给你的练习短文打分。如果你申请商学院，且最近要参加GMAT（经企管理研究生入学考试），其中的一个文章打分工具就是一个学习系统。可能当你求职时，某学习算法会从虚拟文件中挑选出你的简历，并告诉未来的雇主：这位是很不错的人选，看看吧。最近公司给你加薪可能还多亏另外的学习算法。如果想买套房子，Zillow.com网站会估算你看中的每套房的价值，接着房子就有了着落。之后申请住房贷款，某学习算法会研究你的申请，并建议是否可以通过申请。最重要的是，如果你使用在线约会服务，机器学习甚至可能帮你找到人生挚爱。

社会在不断变化，学习算法也是如此。机器学习正在重塑科学、技术、商业、政治以及战争。卫星、DNA（脱氧核糖核酸）测序仪以及粒子加速器以前所未有的精细程度探索自然，同时，学习算法将庞大的数据转变成新的科学知识。企业从未像现在这样了解自己的用户。

在美国大选中，拥有最佳选举模型的候选人奥巴马最终战胜了对手罗姆尼，获得了竞选胜利。无人驾驶汽车、轮船、飞机分别在陆地、海面、空中进行生产前测试。没有人把你的喜好编入亚马逊的推荐系统，学习算法通过汇总你过去的购买经历就能确定你的喜好。谷歌的无人驾驶汽车通过自学，懂得如何在公路上平稳行驶，没有哪个工程师会编写算法，一步一步指导它该怎么走、如何从A地到达B地——这也没必要，因为配有学习算法的汽车能通过观察司机的操作来掌握开车技能。

机器学习是“太阳底下的新鲜事”：一种能够构建自我的技术。从远古祖先学会打磨石头开始，人类就一直在设计工具，无论这些工具是手工完成的，还是大批量生产的。学习算法本身也属于工具，可以用它们来设计其他工具。“计算机毫无用处，”毕加索说，“它们只能给你提供答案。”计算机并没有创造性，它们只能做你让它们做的事。如果你告诉它们要做的事涉及创造力，那么就要用到机器学习。学习算法就像技艺精湛的工匠，它生产的每个产品都不一样，而且专门根据顾客的需要精细定制。但是不像把石头变成砖、把金子变成珠宝，学习算法是把数据变成算法。它们掌握的数据越多，算法也就越精准。

现代人希望让世界来适应自己，而不是改变自己来适应世界。机器学习是100万年传奇中最新的篇章：有了它，不费吹灰之力，世界就能感知你想要的东西，并依此做出改变。就像身处魔法林，在你通过时，周围的环境（今天虚拟，明天现实）会进行自我重组。你在树木和灌木中选出的路线会变成一条路，迷路的地方还会出现指路标志。

这些看似有魔力的技术十分有用，因为机器学习的核心就是预测：预测我们想要什么，预测我们行为的结果，预测如何能实现我们的目标，预测世界将如何改变。从前，我们依赖巫医和占卜师进行预测，但他们太不可靠；科学的预测就更值得信赖，但也仅限于我们能系统观察和易于模仿的事物，大数据和机器学习却大大超出这个范

围。我们可通过独立的思维来预测一些常见的事情，包括接球和与人对话，但有些事情，即便我们很努力，也无法预测。可预测与难以预测之间的巨大鸿沟，可以交给机器学习来填补。

矛盾的是，尽管学习算法在自然和人类行为领域开辟了新天地，但它们仍笼罩在神秘之中。媒体每天都报道涉及机器学习的新闻：苹果公司发布Siri个人助理，IBM[2]沃森（IBM的超级计算机）在《危险边缘》游戏中战胜了人类，塔吉特（Target）能在未成年妈妈的父母发现之前通知她怀孕，美国国家安全局在寻找信息连接点……在这些新闻事件中，学习算法如何起作用仍不得而知。计算机“吞入”数以万亿的字节，并神奇地产生新的观点，关于大数据的书籍甚至也避谈“这个过程到底发生了什么”。我们一般认为学习算法就是找到两个事件之间的联结点，例如，用谷歌搜索“感冒药”和患感冒之间的联系。然而，寻找联结点与机器学习的关系就像是砖与房子的关系，房子是由砖组成的，但一堆砖头肯定不能称之为“房子”。

当一项新技术同机器学习一样流行且具有革命性时，不弄明白其中的奥妙实在太可惜。模棱两可会导致误差和滥用。亚马逊的算法能断定当今世界人们在读什么书，这一点比谁都强。美国国家安全局的算法能断定你是否为潜在恐怖分子。气候模型可以判定大气中二氧化碳的安全水平。选股模型比我们当中的多数人更能推动经济发展。你无法控制自己理解不了的东西，这也是追求幸福的公民、专家或普通人需要了解机器学习的原因。

本书的第一个目标就是揭示机器学习的秘密。只有工程师和机修工有必要知道汽车发动机如何运作，但每位司机都必须明白转动方向盘会改变汽车的方向、踩刹车会让车停下。当今极少有人知道学习算法对应的原理是什么，更不用说如何使用学习算法。心理学家丹·诺曼（Don Norman）创造了“概念模型”（conceptual model）这个新词，代指为了有效利用某项技术而需粗略掌握的知识。本书就将介绍机器学习的概念模型。

并不是所有算法的工作原理都相同，这些差异会产生不同的结果，比如亚马逊和网飞的推荐系统。假设这两个系统试着根据“你喜欢的东西”来对你进行引导，亚马逊很有可能会把你带到你之前常浏览的书籍类别，网飞则可能会把你带到你不熟悉且似乎有点奇怪的区域，并引导你爱上那里。在本书当中，我们会看到诸如亚马逊、网飞之类的公司使用的各式各样的算法。与亚马逊相比，网飞公司的算法对你的爱好理解得更深（尽管还是很有限），然而具有讽刺意味的是，这并非意味着亚马逊也应该利用这个算法。网飞的商业模式是依靠晦涩的电影、电视节目的长尾效应来推动需求，这些电影和节目的成本很低。它一般不推荐大片，因为你的会员订阅费可能有限。亚马逊则没有这样的问题：尽管擅长利用长尾效应，但它同样乐意把更昂贵的热销商品卖给你，这也会简化其物流工作。对于那些奇怪的产品，如果是订阅会员可免费享用的，我们可能会乐意去尝试，而如果需要另外掏钱，我们去选择它们的可能性就小得多。

每年都会出现上百种新的算法，但它们都是基于几个相似的基本思路。为了明白机器学习如何改变世界，你有必要理解这些思路。本书就将对此进行介绍。学习算法并不是那么深奥难懂，除了运用在计算机上，对于我们来说很重要的问题都可以通过学习算法找到答案，比如：我们如何学习？有没有更好的方法？我们能预测什么？我们能信任所学的知识吗？对这些问题，机器学习的各个学派有不同的答案。

机器学习主要有5个学派，我们会对每个学派分别介绍：符号学派将学习看作逆向演绎，并从哲学、心理学、逻辑学中寻求洞见；联结学派对大脑进行逆向分析，灵感来源于神经科学和物理学；进化学派在计算机上模拟进化，并利用遗传学和进化生物学知识；贝叶斯学派认为学习是一种概率推理形式，理论根基在于统计学；类推学派通过对相似性判断的外推来进行学习，并受心理学和数学最优化的影响。在构建机器学习的目标推动下，我们将回顾过去100年的思想史，并以新的观点来看待这段历史。

机器学习的5个学派都有自己的主算法，利用这种万能学习算法，原则上，你可以通过任何领域的数据来挖掘知识：符号学派的主算法是逆向演绎，联结学派的主算法是反向传播，进化学派的主算法是遗传编程，贝叶斯学派的主算法是贝叶斯推理，类推学派的主算法是支持向量机。在实践中，这些算法可能在有些工作中可用，而在其他工作中不可用。我们真正想要寻找的是能够综合这5种算法的终极算法。虽然有些人认为这难以实现，但对机器学习领域的人来说，这个梦想赋予我们力量，促使我们夜以继日地工作。

如果存在终极算法，那么它可以通过数据学得包括过去的、现在的以及未来的所有知识。创造终极算法将是科学历史上最伟大的进步之一。它可以加速各类知识的进步，并以我们现在甚至无法想象的方式改变世界。终极算法与机器学习的关系就像标准模型和粒子物理学或中心法则与分子生物学的关系：该统一原理能理解人类当今知道的一切，并为未来数十年或者数百年的进步奠定基础。今天我们面临许多难题，比如制造家用机器人和治愈癌症，终极算法就是解决这些难题的关键。

以癌症为例。治愈癌症十分困难，因为它往往是一种综合疾病。肿瘤可由各种原因诱发，且在转移时会发生突变。杀死肿瘤细胞最可靠的方法是对其基因进行排序，弄明白哪些药物可以抵抗癌细胞（这种方法不会对人造成伤害，患者必须提供基因和用药史），甚至为你专门研制一种新药。没有哪个医生能够掌握该过程所需的所有知识。对于机器学习来说，这却是再合适不过的任务。实际上，与亚马逊和网飞每天所做的搜索工作相比，它的工作是为你找到正确的疗法，而不是合适的书籍或者电影，而且它的工作更为复杂，也更具挑战。遗憾的是，虽然当今的学习算法能以超出人类水平的精确度来诊断疾病，但治愈癌症仍远远超出它们的理解范围。如果我们可以找到终极算法，这将不再是难题。

因此，本书的第二个目标就是帮你创造终极算法。你可能会认为这需要高深的数学运算和严谨的理论方面的工作，正相反，它需要暂时放下数学奥秘，来观看各种学习行为包罗万象的模型。对外行人来说，他们就像从远方赶到终极算法这片森林，从某些角度看，他们比专家更适合创造终极算法，因为专家对某些学科已经过于投入。一旦我们有了概念性的解决方法，就能补充数学上的细节，但这不是本书的目标和重点。我们之所以谈论每个学派，是为了收集它们的观点，并找到其适用之处。请记住，没有哪个盲人能了解整头大象。我们会尤其关注哪个学派能对治疗癌症做出贡献，也关注该学派的缺失。然后，我们会将所有观点集中，一步步地变成解决方案——这个解决方案可能还不是终极算法，但已是我们能找到的最接近终极算法的方案。希望它能解放你的大脑，让你大胆想象。当你阅读本书时，如果觉得某些章节读起来困难，可以随意略读甚至跳过它们。本书的概要才是重中之重，当明白所有学派的观点之后，如果你重读那些困难的章节，收获可能会比之前更多。

我研究机器学习已经有20余年了。我对机器学习的兴趣因一本书而起，大四时我在书店看到这本书名很奇怪的书——《人工智能》（*Artificial Intelligence*）。那本书只有一个章节是关于机器学习的，但读那个章节时，我立即确定，学习是实现人工智能的关键，而且当时技术水平如此原始，我也许能做点什么。所以我搁置了读MBA（工商管理硕士）的计划，到加利福尼亚欧文分校攻读博士。机器学习当时是一个小众且鲜为人知的领域，研究人员寥寥无几，但加利福尼亚大学却拥有一个巨大的研究团队。一些同学中途放弃了，因为他们看不到机器学习的未来，而我坚持了下来。对我来说，没有什么能比教计算机学习更有吸引力的了：如果我们做到这一点，其他问题就会迎刃而解。5年后我毕业了，那时数据挖掘技术十分流行，我开始写这本书。我的博士论文结合了符号学派和类推学派的观点。过去10年，我一直在整合符号学派和贝叶斯学派的观点，最近又在尝试整合

它们与联结学派的观点。是时候进行下一步研究，并尝试综合这5个范式了。

写这本书时，我的脑海里浮现出各式各样但又有相似之处的读者。

围绕大数据以及机器学习的讨论充满争议，如果你对此感到好奇，且怀疑有比论文上看到的更为深层次的东西，那么这本书就是你进行革命的指南。

如果你的主要兴趣是机器学习的商业用途，那么本书至少能通过6种方法帮助你：成为分析学中更精明的消费者；充分利用你的数据专家；减少许多数据挖掘项目的隐患；看看如果不买手写编码软件，你能让什么进行自动操作；降低信息系统的僵硬度；期待正朝你走来的新技术。我见过太多浪费大量时间和金钱去解决难题的人，他们使用了错误的学习算法，或者误解了学习算法的含义。要避免这些惨败，实际上，你只需要阅读这本书。

如果你是普通人或者决策者，关注由大数据和机器学习引发的社会和政治问题，那么本书将为你提供该技术的入门知识：什么是机器学习，机器学习能干什么、不能干什么。本书没有让你觉得乏味的复杂细节。从隐私问题到未来的工作，以及机器人化引起战争的道德观，我们会看到真正的问题所在，以及如何正确思考。

如果你是科学家或者工程师，那么机器学习肯定是你不想错过的有力武器。在大数据时代（即便是中型数据时代），陈旧的、靠得住的统计工具并不会让你走得更远。你需要的是机器学习的非线性技术来精确模仿多种现象，它会带来全新的、科学的世界观。今天，“范式转移”被人们用得过于随意，但我可以毫不夸张地说，本书要讲的内容就是和“范式转移”相关。

如果你是机器学习专家，那么你可能对本书的大部分内容已经相当熟悉，但你仍会发现其中有许多新颖的看法、经典的观点，以及有

用的例子和类比。很大程度上，我希望本书能提出与机器学习相关的、新的看法，甚至能让你开始思考新的方向。我们身边到处是容易达成的目标，我们理应追寻这种目标，但我们也不应忽略不远处就有更大的目标（关于这一点，我希望你们能原谅我诗意地用“终极算法”来指通用型学习算法）。

如果你是学生，无论你多大，是考虑该选什么专业的高中生，还是决定该研究什么领域的本科生，或者是考虑转行、经验丰富的专家，我希望本书能让你对这个令人着迷的领域感兴趣。当今世界极度缺乏机器学习专家，如果你决定加入这一行列，你不仅能得到令人激动的时刻和丰厚的物质回报，还有服务社会的大好机会。如果你已经在研究并学习主算法，我希望本书能帮你了解它的历史。如果你在旅途中偶然发现本书，也值得你用心阅读。

最后要强调一点，如果你渴望奇迹，那么机器学习对你来说就是一场精神盛宴。我诚挚地邀请你一同前往。

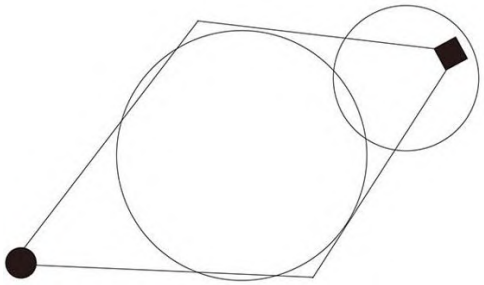
---

[1] Kinect是微软对Xbox360体感周边外设正式发布的名字。——编者注

[2] IBM，国际商业机器公司。——编者注

第一章

机器学习革命



我们生活在算法的时代。一两代人以前，提到“算法”这个词，可能多数人会脑中一片空白。当今，文明社会的每个角落都存在算法，日常生活的每分每秒也都和算法有关。算法不仅存在于你的手机或笔记本电脑，还存在于你的汽车、房子、家电以及玩具当中。当人们进出银行时，银行系统就是由各种算法交织而成的庞大集合体。算法安排航班，也驾驶飞机。算法能经营工厂、进行交易、运输货物、处理现金收益，还能保存记录。如果所有算法都突然停止运转，那么就是人类的世界末日。

算法就是一系列指令，告诉计算机该做什么。计算机是由几十亿个微小开关（称为晶体管）组成的，而算法能在一秒内打开并关闭这些开关几十亿次。最简单的算法是触动开关。一个晶体管的状态就是一个比特信息：如果开关打开，信息就是1；如果开关关闭，信息就是0。银行的计算机的某个比特信息会显示你的账户是否已透支。美国社会保障总署的计算机的某个比特信息表明你是活着还是已死亡。第二简单的算法是：把两个比特结合起来。克劳德·香农以“信息论之

父”而为人所知，他第一个意识到晶体管的活动就是在运算，因为晶体管开了又关，是对其他晶体管的回应（这是他在麻省理工学院的硕士论文——有史以来最有意义的硕士论文）。如果A晶体管只有在B和C晶体管都打开时才打开，那么这时它就是在做小型的逻辑运算。如果A晶体管在B和C晶体管其中一个打开时才打开，就是另外一种小型逻辑运算。如果A晶体管在B晶体管任何关闭的时候打开，或者反过来，这又是第三种运算。信不信由你，所有算法，无论多复杂，都能分解为这三种逻辑运算：且，或，非。利用不同的符号来代替“且”“或”“非”运算，简单的算法就可以用图表来表示。例如，如果发烧可由感冒或者疟疾引起，那么你应该用泰诺来治疗发烧和头疼，可以用图1－1表示。

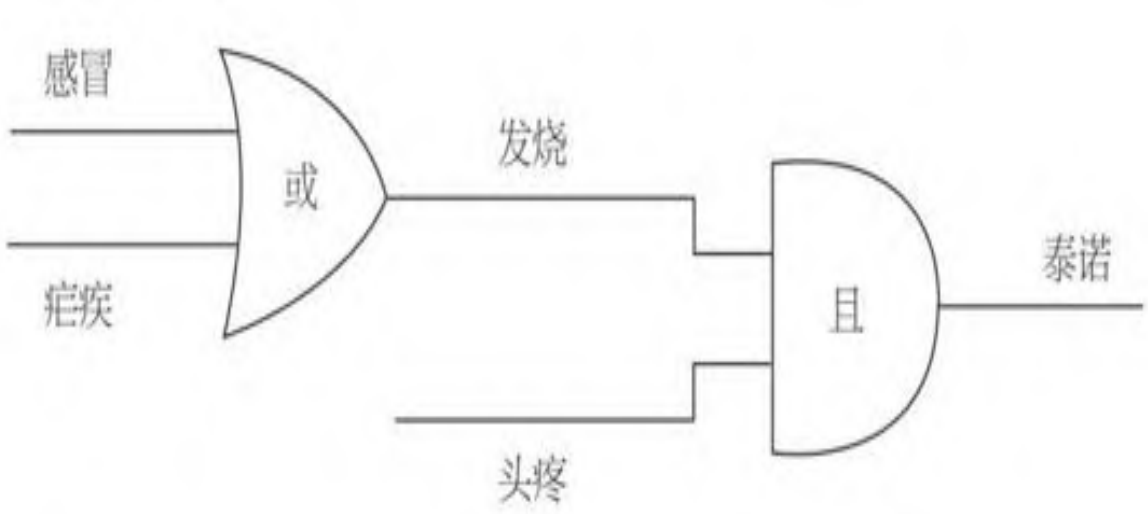


图1－1

通过结合许多这样的逻辑运算，我们可以进行极其复杂的逻辑推理。人们往往认为计算机只和数字有关，其实并非如此，它完全关乎逻辑。数字和算术都是由逻辑构成的，而计算机的所有其他部分也是如此。想把两个数相加？可以由晶体管的组合体来完成。想赢得《危险边缘》智力比赛？也可以由晶体管的组合体来完成（当然，这个组合体庞大得多）。

即便如此，为了做不同的事而制造新的计算机代价过于昂贵。当然，现代计算机是各种晶体管的大集合，能做许多不同的事，这取决于哪些晶体管被激活。米开朗琪罗说过，他所做的一切，就是从大理石石块中看出雕像，然后将多余的石头刻掉，直到雕像的形状显现出来。同样，算法排除计算机中多余的晶体管，直到出现想要的功能，无论是客机的自动驾驶仪，还是皮克斯的新电影，原理都是这样。

一种算法不仅是简单的一套指令，这些指令必须精确且不能模糊，这样计算机才能够执行。例如，食谱并不算一种算法，因为食谱没有明确给出做事的顺序，或者具体每一步是怎样的。一勺白糖到底是几克？每个尝试新食谱的人都知道，跟着食谱做，可能会做出很美味的食物，也可能会做得一塌糊涂。相比之下，算法总能得出同样的结果。即便食谱明确指出需要半盎司白糖，计算机也不知道如何执行，因为计算机不知道什么是白糖、什么是盎司。如果我们想对厨用机器人编程，让它来做蛋糕，我们要通过视频教它如何辨认白糖、如何拿起勺子等（我们现在仍在努力）。计算机必须知道如何执行算法，直到打开及关闭指定的晶体管。因此，食谱离算法还很远。

另一方面，下面是玩井字棋的算法：

如果你或对手有两粒连子，占据剩下的角落。

否则，如果两边有两个连子的走法，就那样走。

否则，如果正中央是空的，走正中央。

否则，如果你的对手走到角落，占据他的对角。

否则，如果有空白的角落，占据它。

否则，占据任意空白的角落。

这个算法有很大的优点，那就是它绝对不会输。当然，它仍忽略了许多细节，比如在计算机的记忆中，棋盘如何表示，而棋的走法又

如何改变这种表示方法。例如，每个角落我们有两个比特，如果中间是空的，值就是00；如果有一个圈，值就变成01；如果有一个叉，值就变成10。即便如此，这也足够精确、清晰，能让有能力的编程员来填补被忽略的空白。它还有一个好处，就是不用我们自己指定算法，细到单个晶体管。在构建数据存储块时，我们可以使用之前存在的算法，而且有很多这样的算法供选择。

算法是一套严格的标准。人们常说，你没法真正了解某样东西，直到你能用一种算法来将其表达出来（理查德·费曼曾说，“如果我无法创造某样东西，那么也就无法理解它”）。方程式对物理学家和工程师来说就是谋生工具，而这也仅仅是一种特殊算法。例如，牛顿第二定律，可以说是有史以来最重要的等式，告诉你用物体的质量乘以其加速度，可以算出作用在物体上的力。该定律还隐含地告诉你，加速度等于作用力除以质量，要弄明白这一点，只需一个运算步骤。在科学的任何领域，如果某个理论无法用算法表示，那么它就不是很严谨（更别提你无法用计算机来解决这个问题，因为你能让计算机替你做的事实在太有限）。科学家提出理论，工程师制造设备，计算机科学家提出算法，这和理论及设备都有关。

设计算法并没有那么简单。这个过程充满陷阱，什么事都不能想当然。如果你的一些构建已经出错，就得找其他方法。设计算法最重要的一点就是，你得用一种计算机能理解的语言来将算法记录下来，比如Java或者Python（从这个角度看，就是一个程序）。接下来，你得对其进行纠错：找出每个误差并修正，直到计算机能够运行程序，而不至于搞砸。一旦你有了能完成你心愿的程序，就轻松多了。计算机会以飞快的速度，按我们的要求办事，而且毫无怨言。世界上的每个人都能享用你的创作成果。如果你愿意，这个成果可以一文不收；当然，如果你解决的问题足够有意义，这个成果也可以让你成为亿万富翁。程序员（创造算法并将其编码的人）是一个“小神灵”，能任意创造不同的世界。甚至你也可以说《圣经·创世记》里的神也是“程序员”：语言（而不是统治权）才是他创造世界的工具。语言构

成了这个世界。当今时代，坐在沙发上利用笔记本电脑，你就可以成为一个“神”。你完全可以想象一个世界，并实现它。

有朝一日，计算机科学家会互相依赖各自的成果，然后为新事物创造算法。这些算法会与其他算法相结合，目的是利用其他算法的成果，反过来产生能服务更多算法的成果。每一秒钟，数十亿计算机里的数十亿晶体管会打开关闭数十亿次。算法形成新型生态系统，它将生生不息，具有无可比拟的生命多样性。

然而，不可避免地，在这个“伊甸园”里也会有狡猾的人存在，人们称之为“复杂性怪兽”。和九头蛇一样，这个复杂性怪兽有很多头，其中一个就是空间复杂性，即为了储存在计算机内存中，一个算法所需信息的比特数量。如果计算机无法提供该算法所需的内存，那么这个算法就没用，必须忽略。接着是邪恶的同类——时间复杂性：该算法运行多长时间，也就是说，在产生想要的结果之前，算法利用及重新利用晶体管的步骤有多少。如果算法运行时间太久，我们等不了，那么这个算法也没用。复杂怪兽最恐怖的一面就是人类的复杂性。当算法变得很复杂以致人类大脑已无法理解，当算法不同部分的交互过多且过于深入时，误差就会悄然潜入。我们找不到这些误差，也就无法纠正它们，算法也就不会做我们想做的事。即便我们让它运行起来，它也会停下来。对使用它的人来说，它没必要那么复杂，而且它和其他算法也合作得不好，这为日后埋下隐患。

每位计算机科学家每天都在和“复杂性怪兽”做斗争。如果科学家输了这场斗争，复杂性就会渗入我们的生活。你可能已经注意到，很多这样的斗争科学家已经输了。即便如此，我们也会继续构建我们的算法之塔，并迎接越来越大的挑战。每一代新的算法都要在之前的基础上构建，除了这代算法的复杂性，它们还面临之前算法的复杂性。塔会变得越来越高，会覆盖整个世界，但它也会变得越来越脆弱，像一座纸片做的房子，随时都会倒塌。算法里的微小误差可能导

致价值10亿美元的火箭爆炸，或者可能导致停电，造成数百万美元的损失。算法以意想不到的方式进行交互，股票市场就会崩溃。

如果程序员是“小神”，复杂性怪兽就是魔鬼。慢慢地，魔鬼会赢得战争。

总得有个更好的方法来与魔鬼做斗争。

学习算法入门

每个算法都会有输入和输出：数据输入计算机，算法会利用数据完成接下来的事，然后结果就出来了。机器学习则颠倒了这个顺序：输入数据和想要的结果，输出的则是算法，即把数据转换成结果的算法。学习算法能够制作其他算法。通过机器学习，计算机就会自己编写程序，就用不到我们了。

哇！

计算机会自己编写程序。现在看来这是一个强大的想法，甚至可能有点吓人。如果计算机开始自己编程，那么我们将如何控制它们？我们会看到，人类可以很好地控制它们。可能会有人当即反对，这听起来太美好了，不像真的。当然，编写算法需要智力、创造力、问题解决能力，这些都是计算机没有的。如何把机器学习与魔法区分开来？的确，今天为止，人们能编写许多计算机无法学习的程序。可令人更为惊讶的是，计算机却能学习人们无法编写出来的程序。我们会开车、会辨认字迹，但这些技能都是潜意识发挥出来的，无法向计算机解释这些事情是如何完成的。但是，如果我们把关于这些事情的足够多的例子交给学习算法，该算法会很乐意弄明白怎样独立完成这些

事情，这时我们就可以放手让算法去做了。邮局正是通过这种方法来识别邮政编码，自动驾驶汽车也是这样才得以实现在路上跑。

解释机器学习的力量的最好方法，也许就是将其与其他低技术含量的活动进行类比。工业社会，商品由工厂制造，这也意味着工程师必须弄明白商品如何通过零件组装起来、这些零件如何生产等，细到生产原料。这是一项大工程。计算机是人类发明的最复杂的产品，计算机设计、工厂生产、程序运行都涉及大量的工作。还有另外一种方法能让我们得到一些想要的东西：让自然规律去塑造它们。在农业当中，我们播种，确保种子有足够的水分和营养，然后收割成熟的作物。为什么技术不能这样？完全可以，而这也是机器学习的承诺。学习算法是种子，数据是土壤，被掌握的程序是成熟的作物。机器学习专家就像农民，播下种子，灌溉，施肥，留意作物的生长状况，事事亲力亲为，而不是退居一旁。

一旦我们这样看待机器学习，随即也会发生两件事：

第一，我们掌握的数据越多，我们能学的也越多。没有数据？什么也学不到。大数据？很多东西可以学习。这也是机器学习无处不在的原因，因为有飞速增长的数据。如果你在超市购买机器学习，其包装上可能会写着“只需添加数据”。

第二，机器学习是一把剑，利用这把剑可以杀死复杂性怪兽。只要有足够的数据，一段只有几百行代码的程序可以轻易生成拥有上百万行代码的程序，而且它可以为解决不同问题不停产生不同的程序。这可以显著降低程序员工作的复杂度。当然，就像对付九头蛇，我们砍掉它的头，会立即长出新头，但长出的头会变小，而且头的生长也需要时间，因此我们仍有可能胜出。

我们可以把机器学习当作逆运算，正如开平方是平方的逆运算、整合是分化的逆运算。正如我们会问“什么数的平方是16”，或者“导数为x+1的函数是什么”，我们也会问“什么算法会得出该结果”。我们很快会看到，怎样将这个观点运用到具体的学习算法中。

有些学习算法学习知识，有的则学习技能。“所有人都会死”是知识，骑单车是技能。在机器学习中，知识往往以统计模型的形式出现，因为多数知识都是可以统计的：所有人都会死，但只有4%是美国人。技能往往以程序的形式出现：如果马路向左弯曲，那么向左转动车头；如果一只鹿跳到你面前，那么立刻刹车（很遗憾，在写这本书时，谷歌的自动驾驶汽车仍会把被风吹起的塑料袋和鹿弄混）。通常，这些程序都很简单，复杂的是它们的核心知识。如果你能判断哪些邮件是垃圾邮件，那么你也就能判断该删除哪些邮件。如果你能在象棋游戏中判断这盘棋自己的优势在哪里，那么你也就懂得该怎么走（能让你处于最有利地位的一步）。

机器学习有许多不同的形式，也会涉及许多不同的名字：模式识别、统计建模、数据挖掘、知识发现、预测分析、数据科学、适应系统、自组织系统等。这些概念供不同群体使用，拥有不同的联系。有些有很长的半衰期，有些则较短。在本书中，我用“机器学习”一词泛指所有这些概念。

机器学习有时会和人工智能（AI）混淆。严格来讲，机器学习是人工智能的子域，但机器学习发展得如此壮大且成功，现已超越以前它引以为傲的母领域。人工智能的目标是教会计算机完成现在人类做得更好的事，而机器学习可以说就是其中最重要的事：没有学习，计算机就永远无法跟上人类的步伐；有了学习，一切都与时俱进。

在信息处理这个生态系统中，学习算法是顶级掠食者。数据库、网络爬虫、索引器等相当于食草动物，耐心地对无限领域中的数据进行蚕食。统计算法、线上分析处理等则相当于食肉动物。食草动物有必要存在，因为没有它们，其他动物无法存活，但顶级掠食者有更为刺激的生活。数据爬虫就像一头牛，网页相当于它的草原，每个网页就是一根草。当网络爬虫进行破坏行动时，网站的副本就会保存在其硬盘当中。索引器接着做一个页面的列表，每个词都会出现在页面当中，这很像一本书后的索引。数据库就像大象，又大又重，永远不会

被忽略。在这些动物当中，耐心的野兽飞快运转统计和分析算法，压缩并进行选择，将数据变为信息。学习算法将这些信息吞下、消化，然后将其变成知识。

机器学习专家在计算机科学家中就是一种精英式的“神职”。许多计算机科学家，尤其是更老的那一代，并不如他们想的那样能很好地理解机器学习。这是因为，计算机科学通常需要的是准确思维，但机器学习需要的是统计思维。例如，如果有条规定是“垃圾邮件标记的正确率是99%”，这并不意味存在缺陷，而可能意味这是你的最好水平，已经很好用了。这种思维上的差别很大程度上也解释了为什么微软能赶上网景公司，但想赶上谷歌却困难得多。说到底，浏览器只是一个标准的软件，而搜索引擎则需要不同的思维模式。

之所以说机器学习研究者是超级计算机迷的另外一个原因，就是当今世界急需他们，但他们寥寥无几。按照计算机科学严格的标准，这样的人数量就更少了。蒂姆·奥莱利认为，“数据科学家”是硅谷最热门的职业。根据麦肯锡全球研究院估计，截至2018年，仅美国就需要再培养14万~19万机器学习专家才够用，另外还需要150万有数据头脑的经理。机器学习的应用爆发得如此突然，连教育都无法跟上其步伐，同时，人才奇缺也是因为这门学科在人们看来很难而令人望而生畏。教科书很可能会让你感到数学很难，然而，这个困难表面看起来很大，其实并不是。机器学习所有的重要观点可以不用通过数学表示出来。当你读这本书时，甚至可能会发现，你发明了自己的学习算法，而且看不到一个方程式的影子。

工业革命使手工业自动化，信息革命解放了脑力劳动，而机器学习则使自动化本身自动化。没有机器学习，程序员会成为阻挠进步的障碍。有了机器学习，进步的步伐就会加快。如果你是一个懒惰又不那么聪明的计算机科学家，机器学习就是理想的职业，因为学习算法会完成所有事情，功劳却是你的。从另一方面讲，学习算法会让我们失业，这也只是我们应受的惩罚。

将自动化带入新的高度，机器学习革命会带来广泛的经济及社会变革，正如互联网、个人计算机、汽车以及蒸汽机在当时对社会和经济的影响那样。这些变革已经明显存在的领域就是商业。

为何商业拥护机器学习

为什么谷歌比雅虎要有价值得多？它们都是用户登录最多的网站，都靠在网页上登广告赚钱。它们都用拍卖的方式销售广告，用机器学习来预测用户点击某广告的概率（概率越大，广告价值越大），但谷歌的机器学习就比雅虎要好很多。这不是它们市场价值差异巨大的唯一原因，却是主要原因。如果没有达到预测的点击量，对广告商来说就是浪费机会，对网站来说是收益损失。谷歌每年的收入是500亿美元，预测点击率每上升1%，就可能意味着每年为公司带来额外5亿美元的收入。难怪谷歌是机器学习的铁杆粉丝，雅虎和其他公司也在奋起直追。

网络营销仅仅是巨大变革中的一种表现形式。无论什么市场，生产商和用户在交易发生之前，都需要进行联系。在互联网出现之前，交易的主要障碍就是实地交易。你只能从当地的书店购买书籍，而当地书店的书架空间又有限。但当你可以随时把所有书下载到电子阅读器时，问题就变成了可供选择的书太多。你怎么浏览书店里上百万不同名字的书？同样的问题也出现在其他信息产品当中：视频、音乐、新闻、推特文章、博客、网页。这个问题还会出现在能够远程购买的产品和服务当中：鞋子、鲜花、小配件、酒店房间、辅导、投资。人们在找工作或挑日子时，也会遇到选择过多的问题。你们如何找到彼此？这是信息时代的定义问题，而机器学习就是问题解决方案的主要部分。

当公司不断发展壮大后，它会经历三个阶段：

第一阶段的所有事都由人工完成——夫妻店的店主亲自了解其顾客，他们依照顾客类型订购、展示、推荐产品。这很不错，但规模不大。

第二阶段是最辛苦的时期，公司变得越来越大，需要用到计算机。公司招来程序员、顾问，买来数据库管理器，程序员编写了成百万行的代码来使公司所有能自动化的功能自动化。更多的人享受到服务，但也有麻烦：决定是在粗略的人口统计类别的基础上做出来的，计算机程序也过于死板，无法与人类无限的才能相匹配。

经过一段时间进入第三阶段，没有足够的程序员和顾问满足公司的需要，因此公司不可避免地向机器学习寻求帮助。亚马逊无法通过计算机程序将所有用户的喜好熟练地进行编码，脸书也不知道如何编写这样的程序，能选择最好的更新内容展示给每位用户。沃尔玛每天销售百万件商品，还要做数十亿个选择。如果沃尔玛的程序员努力编写出能够做所有选择的程序，这些选择就不用人来做了。相反，这些公司所做的工作是，它们在收集到如山的数据后，让学习算法尽情学习，然后预测顾客想要什么产品。

学习算法就是“媒人”：它们让生产商和顾客找到对方，克服信息过载。如果这些算法足够智能，你就能取得两全其美的结果：从宏观来讲，选择广、成本低；从微观来讲，能够了解顾客的个性化需求。学习算法并不是完美的，决定的最后一步通常还得由人来做，但学习算法很智能，为人们减少了需要做的选择。

回顾过去，我们看到，从计算机到互联网再到机器学习的进步是必然的：计算机使互联网成为可能，这个过程产生大量数据以及无限选择这个问题。单单互联网还不足以把“一个尺寸满足所有”的需求转向追求无限多样化的长尾效应。网飞公司的库存里可能有10万种不同名字的DVD（数字多功能光盘），但如果顾客不懂得如何找到自己喜欢的，他们就会默认选择最流行的DVD。只有网飞公司有了学习算法之

后，才能帮助它了解顾客的喜好，并推荐DVD，长尾效应也才得以真正实现。

一旦必然的事情发生，机器学习成为媒介，那么其力量也开始慢慢积聚。谷歌的算法很大程度上决定你会找到什么信息，亚马逊决定你会买到什么产品，全球最大的婚恋网站默契网（Match.com）决定你的约会对象是谁。最好的选择权仍在你手里——从算法给你展示的所有选项中挑选，但99.9%的选择由算法做出。当下，一家公司的成败取决于学习算法对其产品的喜爱程度，而整个经济体的成功——是否每个人都能得到自己需要的物美价廉的产品，则取决于学习算法的好用程度。

公司确保学习算法喜爱其产品的最佳方法就是，让公司自己运行算法。谁有最佳算法、数据最多，谁就能赢。新型网络效应占据上风：谁有最多的用户，谁就能积累最多的数据，谁有最多的数据，谁就能学到最好的模型，谁学到最好的模型，谁就能吸引最多的用户，这是一种良性循环（如果你在竞争，就会变成恶性循环）。把搜索引擎从谷歌转换到必应，可能会比把应用系统从Windows切换到Mac要简单，但在实际操作中，你不会这么做，因为谷歌拥有领先优势及更大的市场份额，比必应更懂得你想要什么，虽然必应的技术也不错。可惜的是，必应刚进入搜索行业，没有什么数据资源，而谷歌却拥有十余年的机器学习经验。

你可能会认为，过一段时间，更多的数据结果意味着更多的重复，但数据的饱和点还未出现，长尾效应持续起作用。如果你看亚马逊或网飞公司为你提供的推荐产品，很明显，这些推荐项仍很粗略，而谷歌的搜索结果也有很大的优化空间。每个产品的特性、网页的每个角落都有很大的潜力，能通过机器学习得到改善。网页底部的链接应该是红色的还是蓝色的？两个颜色都试试，看看哪个颜色的点击率会更高。还有，最好让机器学习持续运行，不断调整网页的所有方面。

所有拥有众多选择和大量数据的市场都会发生这样的动态循环。比赛正在进行，谁学得最快，谁就赢了。随着越来越好地了解用户需求，这个比赛不会停止：企业可以将机器学习应用到企业运作的每个方面，只要有足够的数据，只要数据能够从计算机、通信设备以及更廉价、更普适的传感器源源不断地输出。“数据是新型石油”是目前的流行说法，既然是石油，提炼石油就是一笔大生意。和其他公司一样，IBM已制定经济增长战略，为企业提供分析服务。业界将数据看作战略资产：我有什么数据，而竞争对手却没有？我要怎么利用这些数据？竞争对手有什么数据，而我却没有？

同样的道理，没有数据库的银行无法和有数据库的银行竞争，没有机器学习的企业也无法跟上使用机器学习的企业。虽然第一家公司的专家写了上千条规则，预测用户的喜好，但是第二家公司的算法却能学习数十亿条规则，一整套规则都可用于每位用户。这就相当于长矛对机关枪。机器学习是很棒的新技术，但这并不是商业界拥护它的原因——人们之所以拥护它，是因为别无选择。

给科学方法增压

机器学习是“打了类固醇”的科学方法，也遵循同样的过程：产生假设、验证、放弃或完善。科学家可能会花费毕生精力来提出或验证几百个假设，而机器学习系统却能在一秒钟内做完这些事。机器学习使科学的发现过程自动化。因此，并不奇怪，这既是商业领域的革命，也是科学领域的革命。

为了取得进步，科学的每个领域都需要足够的数据，以与其研究现象的复杂性相对应。这是物理成为第一个腾飞学科的原因：第谷·布拉赫对星球位置的记录，以及伽利略对钟摆摆动、斜面的观察，已

经足以推导出牛顿定律。这也是为什么虽然分子生物学这个学科比神经科学年轻，但是已超越神经科学：DNA（脱氧核糖核酸）微阵列以及高通量测序技术提供了大量的数据，而神经科学家对此只能可望而不可即。这也是为什么社会科学研究是一场艰苦卓绝的斗争：你拥有的只是100人的样本和每个人的十几个测量值，你能模拟的也只是某个规模很有限的现象，甚至这个现象可能不是孤立存在的，还受到其他现象的影响，这就意味你仍然没有彻底了解它。

有个好消息，那就是之前缺乏数据的学科现在拥有很多数据。用不着让50名睡眼惺忪的本科生到实验室完成任务并付给他们报酬，心理学家通过在亚马逊“土耳其机器人”[3]上发布实验任务，就可以找到满足他们数量要求的实验对象（这个网站对更多样化的样本也有帮助）。虽然回想起来越来越困难，但也只是10年前，研究社交网络的社会学家哀叹说，他们无法得到成员超过几百人的社交网络。现在有了脸书，有超过10亿用户。大部分用户会发布有关他们生活的很多细节，就像地球社会生活的实时直播。在神经科学领域，神经连接组学和功能性磁共振成像让人们对大脑有了十分详细的了解。在分子生物学领域，基因和蛋白质的数据库数量以指数级速度增长。甚至更为“年长”的学科，如物理学和解剖学也在不断进步，因为粒子加速器和数字巡天领域的数据在源源不断输出。

如果你不将大数据变成知识，它将毫无用处，可是世界上没有那么多科学家来完成这件事。埃德温·哈勃通过钻研照相底片发现新的星系，但史隆数字巡天计划中，多达5亿的天体肯定不是这样被辨认出来的。这就像在沙滩上用手来数沙粒的数目。你可以记录规则，把星系从星星及干扰物（如鸟、飞机、超人）区分开来，但得出的星系并不是那么准确。相比之下，天体图像目录编辑和分析工具（SKICAT）项目使用了学习算法。底片包括标记了正确类别的天体，从这些底片出发，学习算法可以明白每个分类的特点，并将其应用到没有标记的

底片当中。甚至更理想的是，学习算法能够将那些对人类来说难以标记的天体进行分类，这些天体正是该项调查计划的主要内容。

有了大数据和机器学习，你就能弄明白比之前复杂很多的现象。在多数领域，科学家一般只使用种类很有限的模型，例如线性回归模型，在这个模型当中，你用来适应数据的曲线总是一条直线。遗憾的是，世界上的大多数现象都是非线性的（或者说这也是一件幸事，如果是线性的，生活会变得非常乏味。实际上，那样就不会存在生命了）。机器学习打开了广阔、全新的非线性模型世界。这就好比在只有几缕月光照射的房间，打开了明亮的灯。

在生物学领域，学习算法的研究成果包括：DNA分子中基因的位置；在蛋白质合成前，多余的核糖核酸在哪里进行绞接；蛋白质如何折叠成各自的特有形状；不同条件如何对基因的表达造成影响。用不着在实验室对新药进行测试，机器学习就可以预测这些药物是否有效，只有最有效的药品才会受到测试。学习算法还会剔除那些可能产生严重副作用（甚至导致癌症）的药物，备选药物无须在经过人体试验被证明无效后才被禁止使用，从而避免了代价昂贵的失败。

然而，最大的挑战就是将所有这些数据组合成一个整体。导致你患心脏病的因素有哪些？这些因素如何相互影响？牛顿需要的只是三个运动定律和一个万有引力定律，但一个细胞、一个有机体、一个社会的完整模型却无法由一个人来发现。虽然随着知识的增长，科学家的分工变得越来越细，但是没有人能够将所有知识整合到一起，因为知识太多了。虽然科学家们会合作，但语言是传播速度非常缓慢的介质。虽然科学家们想努力追上别人的研究，但出版物的数量如此之多，他们的距离被拉得越来越远。通常是，重做一项实验比找到该实验的报告还要容易。机器学习在这时就会起作用，它能根据相关信息搜索文献，将某领域的行话翻译到另一个领域，并建立联系，而科学家们在过去都没有意识到。渐渐地，机器学习成为一个巨大的中心，通过这个中心，某领域里发明的建模技术将会被引入其他领域。

如果计算机没有被发明出来，20世纪下半叶的科学将停滞不前。这可能不会很快在科学家当中表现出来，因为他们专注于所有仍可努力实现、有限的进步，但进步的空间真的太小了。同样，如果没有机器学习，许多科学在未来10年将会面临收益递减。

为了预见科学的未来，看看曼彻斯特大学生物技术研究院的实验室，在那里，一个名叫亚当的机器人正在努力工作，目的是找到哪些基因在酵母中对哪些酶进行编码。亚当有一个酵母新陈代谢的模型，还掌握了基本的基因及蛋白质知识。它提出假设，设计实验验证假设，进行实地实验，分析结果，提出新的假设，直到它满意为止。当下，人类科学家仍然在独立检查亚当的结果，然后才会相信这些结果，但在未来，他们就会交给机器人科学家来验证彼此的假设。

10亿个比尔·克林顿

在2012年的美国总统选举中，机器学习决定了谁能当上总统。通常决定总统选举的因素包括经济、候选人的亲民度等，但这些因素没有起到作用，而选举的结果主要受到几个“摇摆州”的影响。米特·罗姆尼的竞选采用的是传统的投票策略，将选民分成几大类，然后选择是否把每个类别作为目标。尼尔·纽豪斯（罗姆尼的民意调查专家）说道：“如果我们能在俄亥俄州赢得无党派人士，那么这场竞赛我们就赢了。”虽然罗姆尼获得了7%无党派人士的支持，但他仍失去了这个州，在竞选中失利。

相比之下，奥巴马总统雇用了拉伊德·贾尼（机器学习专家，他是奥巴马竞选中的首席科学家）。贾尼研究的是如何整合最伟大的分析运算，并将其应用到政治史中。他们将所有选民的信息整合成单个数据库，然后将该数据库和他们能在社交网络、市场营销等领域找到

的资源结合起来。之后着手对每个选民做四种预测：（1）支持奥巴马的可能性有多大；（2）会不会参加民意调查；（3）会不会回应竞选宣传并照做；（4）对特定问题进行对话之后，他们会不会改变选举决定。基于这些选民的例子，奥巴马团队每个晚上进行66 000场选举模拟，并用这些结果指导奥巴马竞选的志愿者大军：该给谁打电话，该拜访谁，该说什么。

在政界、商界以及战争中，最糟糕的事情莫过于，你不明白对手的行动，而知道该怎么做时，为时已晚。这就是发生在罗姆尼竞选中的事情，他们能看到对手的团队在特定镇的特定电台花钱做宣传，却不知道这是为什么，他们能预测的实在太少。最后，奥巴马除了北卡罗来纳州以外，赢得了每个州，而且与最可靠的民意调查专家的预测相比，他赢得了更多。反过来，最可靠的民意调查专家（例如内特·希尔）使用的是最复杂的预测技术，预测结果却没有奥巴马竞选团队的结果准确，因为他们的资源比较少。但他们比那些所谓的权威人士要准确很多，因为那些人的预测只是基于他们自己的专业知识。

也许你会认为，2012年的美国总统竞选只是机缘巧合：大多数选举结果并不那么接近，机器学习无法成为决定因素。但未来机器学习会让更多的选举结果更接近。在政界，正如在所有领域那样，学习就像一场掰手腕比赛。在卡尔·罗夫（前直销商和数据挖掘工程师）的年代，共和党是领先的。到了2012年，共和党开始掉队，但现在他们又追上来了。我们不知道下一轮选举谁会领先，但我们知道两个党派为了赢得选举都很努力。这也就意味着，应该更好地了解选民，根据候选人的情况进行宣传，甚至根据实际情况选择候选人。在选举期间以及每轮选举之间，这适用于整个党纲：在硬数据的基础上，如果详细的选民模式表明该党派现在的纲领是失败的，那么该党派就应改变它。因此，把主要选举活动放到一边，民意调查中候选人的差距会变得越来越小，而且很快会消失。其他条件不变，拥有更好选民模式的候选人会赢得选举，而选民也会因此得到更好的服务。

政治家最伟大的才能之一，就是能够了解其选民个人或者选民团体，然后直接与他们对话，比尔·克林顿就是其中的一个典范。机器学习的作用就是，让每位选民都觉得克林顿对待他们亲力亲为、非常用心。尽管他们心目中的这些小小克林顿与真的克林顿相差太远了，但优势在于“小克林顿”的数量众多，哪怕比尔·克林顿根本无法了解美国的每位选民是怎么想的（虽然他确实想知道）。学习算法是最强大的政治家推销商。

当然，就像企业一样，政治家能把机器学习掌握的信息用好，也可能会用得很糟糕。例如，对不同的选民，他们可能会给出不一致的承诺，但选民、媒体、监督组织也会自己进行数据挖掘，并揭露做得太过分的政治家。竞选活动不仅仅是候选人之间的较量，还涉及民主进程中的所有参与者。

更大范围的结果就是，民主会更好地得到实现，因为选民与政治家之间交流的范围会飞速扩大。在当今这个高速互联网时代，民意代表从你身上获取的信息数量仍像19世纪时一样有限：每两年会有100比特左右的信息，数量正好对应一张选票。这些信息会由民意信息来补充，或许偶尔还会有电子邮件和市民大会的信息，但还是少得可怜。大数据和机器学习正改变这种等式关系。在未来，只要选民模式准确，当选官员就可以每天询问选民上千次想要什么，然后根据询问结果来办事，不用在现实中纠缠选民。

学习算法与国家安全

在网络空间之外，学习算法是保护国家的壁垒。每天，国外袭击者都会企图闯进五角大楼、国防承包商以及其他公司和政府机构的计算机。他们的计谋不断变化，能抵抗昨天袭击的方法，今天就已经不

管用了。编写代码来侦查并阻止每场袭击，可能会和马其诺防线一样有效，五角大楼的网络司令部十分了解这一点。但如果是恐怖分子的第一次袭击，而且也没有之前的例子供机器学习来参考，那么机器学习就会遇到问题。学习算法会构建正常行为的模型（这样的模型数量很多），标出异常行为，然后召集来“骑兵”（系统管理员）。如果网络战争发生，人类就是总指挥，算法就是步兵。人类速度太慢、数量太少，很快就会被机器人大军歼灭。我们需要自己的机器人军队，而机器学习就像机器人中的西点军校。

网络战争是不对称战争的一个例子，一方的传统军事实力比不上另一方，但仍然可以给对方造成严重伤害。少数恐怖分子只用美工刀就可以撞到双子塔，并让几千名无辜者遇难。当今美国安全最大的威胁就是不对称战争，而且抵抗所有威胁的有效武器就是信息。如果敌人躲不了，那么他也活不了。好消息就是我们有大量信息，但也有坏消息。

美国国家安全局已经对数据产生无限大的胃口，也因此声名狼藉。据估计，每天美国国家安全局窃听着全球10亿多个通话，还有其他通信。但是，抛开隐私问题，它也没有让上百万员工来窃听这些通话、偷看邮件，甚至也不会记录谁和谁通话。绝大多数通话是没有嫌疑的，而专门编写程序来找出有嫌疑的通话又很困难。过去，美国国家安全局利用关键词配对方法，但要应付这个方法也很容易（例如，把爆炸袭击称作“结婚”，把炸弹称作“结婚蛋糕”）。21世纪，这些事就可以交给机器学习。保密是安全局的标志，但安全局局长已经向美国国会证明，通话记录挖掘已经阻止了几十起恐怖威胁。

恐怖分子可隐藏在足球比赛的人群中，但学习算法能辨认他们的相貌。恐怖分子可以在国外制造爆炸事件，但学习算法能找出他们。学习算法还可以做更加精细的事情：将机器人与事件连接起来，这些事件单个看起来并无危害，但集中起来可能就预示着不祥。这种方法本可以阻止“9·11”事件的发生。有一个进一步的转折：一旦确定的

程序部署下来，坏人可能会改变其活动，以扰乱该程序。这与自然世界不同，自然世界总是以同样的方式运转。要解决这个问题，就要将机器学习与博弈论相结合，这是我已经在做的工作：别只想着打击对手当前想做的事，要学会巧妙地回避对手对你的学习算法的损害。正如博弈论那样，把各种措施的成本和利益考虑在内，这也有助于找到隐私与安全之间的平衡点。

不列颠之战期间，英国空军阻止了纳粹德国空军的进攻，尽管后者人数比前者多很多。德国飞行员不明白，为什么无论走到哪里，他们总会碰上英国空军。英国有一个秘密武器：雷达，可以在德国飞机越境进入英国领空时，就探测到它们。机器学习就像装了雷达，能够预知未来。别只是回击对手的行动，要预测他们的行动，并先发制人。

一个更确切的例子就是人们熟知的“预知执法”。通过预测犯罪倾向，战略性地将巡逻队集中在最可能需要的地方，同时采取其他预防措施，这样一座城市的警力就能有效地完成更大范围的工作。在许多方面，执法过程就像不对称战争，会用到许多相似的学习算法，无论是在侦查诈骗、揭露犯罪网络，还是普通传统的打击执法中。

机器学习在战争中也将扮演越来越重要的角色。学习算法能有助于驱散战争迷雾，筛选侦察图像，处理后续报告，并整合信息，为指挥官提供战争形势分析。学习算法可以武装军用机器人的大脑，帮助其保持方位，适应地形，把敌机和民用机区别开来，以及进行制导。美国国防部高级研究计划局（DARPA）的领头狗（AlphaDog）能为士兵搬运设备。遥控飞机在学习算法的作用下可自主飞行。虽然它们仍受到人类飞行员的部分控制，但未来的趋势是一个飞行员监控越来越多的遥控飞机群。在未来的军队里，学习算法的数量会大大超过士兵的人数，这将减少许多士兵的伤亡。

我们将走向何方

科技潮流奔涌而来并迅猛向前。机器学习不同寻常的一点就是，在经历所有这些变革以及繁荣和破产之后，它开始逐渐强大。它遇到的第一个大的打击是在金融领域，预测股票的起伏波动，起于20世纪80年代。接下来的一波是挖掘企业数据库，在20世纪90年代中开始发展壮大，尤其是在直接营销、客户关系管理、资信评分以及诈骗侦查等领域。接着是网络和电子商务，在这些领域中，自动个性化很快流行起来。当互联网泡沫暂时削弱这种趋势时，将机器学习应用到网页搜索和广告投放的做法开始腾飞起来。不管怎样，“9·11”恐怖袭击后机器学习被应用到打击恐怖主义的战争中。网络2.0带来一连串的新应用，包括挖掘社交网络、搜索哪些博客谈到你的产品。同时，各个领域的科学家也逐渐转向大规模建模，由分子生物学家和天文学家打头阵。人们勉强留意到了房地产泡沫，而其主要影响就是使人才从华尔街转移到硅谷，并受到欢迎。2011年，“大数据”的概念流行起来，机器学习被明确归入全球经济未来的中心。当今，似乎没有哪个人类钻研的领域不受到机器学习的影响，甚至包括看起来没有多大关系的领域（如音乐、体育、品酒）。

尽管机器学习发展很明显，但这也仅仅是未来的预告。虽然它有用，但实际上当今在工业上起作用的学习算法的生成还是受到了很大限制。如果现在实验室的算法能在各领域的前线使用，比尔·盖茨说机器学习的突破产生的价值将相当于10家微软，其实这个说法有点保守了。如果这些观点让研究人员真正觉得眼前一片光明，而且收到效果，那么机器学习带来的就不仅仅是新的文明时代，还是地球生命进化的新阶段。

怎样才能实现这个目标？学习算法如何运行？现在它们不能做什么？它们的下一代会是怎样的？机器学习革命将以什么方式呈现？你

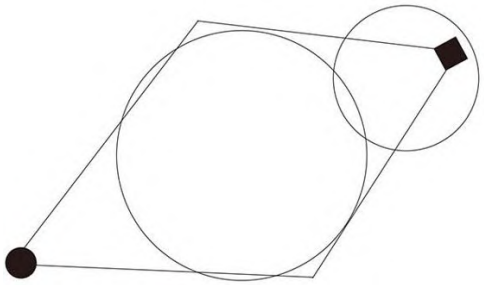
得抓住哪些机遇，提防哪些危险？这些就是这本书要讲的内容。

---

[3] 亚马逊“土耳其机器人”（Amazon Mechanical Turk）是一个Web服务应用程序接口，开发商通过它可以将人的智能整合到远程过程调用。——编者注

第二章

终极算法



机器学习的应用非常广泛，更为惊人的是，相同算法可完成不同的事情。在机器学习领域之外，如果你要解决两个不同的问题，就得编写两个不同的程序。这些程序可能用到相同的基础架构，如相同的编程语言或数据库系统。但是，如果你想处理信用卡申请，诸如下棋的程序则毫无用处。在机器学习领域，如果提供适当的数据来让机器学习，那么相同的算法既可以处理信用卡申请，也可以下棋。实际上，大量的机器学习应用仅仅由几个算法来负责，在接下来的几个章节中我们会谈到这些算法。

例如，朴素贝叶斯算法就是一个可以用短方程来表达的学习算法。只要提供患者病历的数据库，包括病人的症状、检查结果，或者他们是否有什么特殊情况，朴素贝叶斯算法就可在一秒之内做出诊断，而且往往比那些花几年在医学院学习的医生还要强，甚至它还可打败花费数千小时构建的医学专家系统。该算法还可应用于学习垃圾邮件过滤器，乍一看，这和医疗诊断毫无关系。另外一个简单的学习算法就是最近邻算法，它的用途十分广泛，从笔迹识别到控制机器人

手，以及推荐你可能喜欢的书籍或者电影。决策树学习算法也同样擅长决定你的信用卡申请是否应被通过、寻找DNA中的绞接点，以及下棋时指导下一步该怎么走。

相同的学习算法不仅可以完成无穷无尽且不同的事情，而且和被它们替代的传统算法相比，它们要简单得多。多数学习算法可能只需数百行或者数千行代码。相比之下，传统程序则需几十万甚至上百万行代码，并且单个学习算法就可以导出无数个不同的程序。

如果那么少的学习算法就可以做那么多事，那么有一个逻辑上的疑问：单个学习算法可以把所有事情都做完吗？换句话说，单个算法可以学习所有能从数据中学习的东西吗？这是一个非常艰巨的任务，因为这基本上包含成年人大脑里以及人类进步所创造的一切，还有所有科学知识的总和。实际上，对所有主要的学习算法——包括最近邻算法、决策树学习算法以及贝叶斯网络（朴素贝叶斯的概括）——来说，如果你为学习算法提供足够、适当的数据，该算法可以实现任一功能（对学习任何东西来说，都与数学相关）。需要注意的是，“足够数据”也有可能无限。学习无限数据需要做出假设，如我们会看到的那样，而且不同的学习算法会有不同的假设。

如果不把这些假设嵌入算法中，而是将其连同数据一起，当作显示输入，并允许用户选择插入哪一个，甚至陈述新的假设，那么会怎样？有没有这种算法，可以接收任何数据及假设并输出隐藏其中的知识？我相信有。当然，我们得限制假设的可能性，否则如果把整个目标知识都以假设形式赋予算法，那就是在作弊。

我们可以通过限制输入的规模、要求假设弱于当前学习算法等方法，来实现这个目的。

那么疑问就会变成：这些假设要弱到何种程度，但仍能够从无限数据中获得所有相关知识？注意“相关”这个词：我们仅仅对存在于世界的知识感兴趣，对不存在的世界没有兴趣。因此发明一种通用的学习算法可归结为发现宇宙最深层的规律，所有现象都遵循该规律，

然后找出计算的有效方法来将其与数据结合起来。要找到这个“计算的有效方法”，就不能将物理定律视为万物规律，如我们将看到的那样。然而，这并不意味着通用的学习算法要和专用算法一样高效。正如在计算机科学中常发生的那样，我们宁愿牺牲效率来换取通用性。这在学习既定目标知识所需数据的量上也适用：一般通用学习算法会比专用算法需要更多的数据（但如果我们有必要的数量，就没问题），而且数据越多，越有可能会这样。

那么，这就是本书的中心假设：

所有知识，无论是过去的、现在的还是未来的，都有可能通过单个通用学习算法来从数据中获得。

我将该学习算法称为“终极算法”。如果这种算法成为可能，它的发明将成为人类最伟大的科学成就之一。实际上，终极算法是我们最不愿意发明的东西，因为一旦对其放松，它会继续发明一切有可能发明的东西。我们要做的，就是为它提供足够、适当的数据，通过这些数据，它会发现相应的知识：给它视频流，它就会观看；给它图书馆，它就会阅读；给它物理实验结果，它就会发现物理定律；给它DNA晶体学数据，它就会发现DNA的结构。

这可能听起来有点八竿子打不着：一种算法怎么可能学习那么多不同的事情，而且是这么难的事情呢？但实际上，种种证据表明终极算法是存在的。下面我们来看看它们是什么样的。

来自神经科学的论证

2000年4月，麻省理工学院的神经系统科学家团队在《自然》杂志上发布了一项非同寻常的实验结果。他们对雪貂的大脑进行重新布

线，改变了雪貂从眼睛到听觉皮层（大脑负责处理声音的部分）以及从耳朵到视觉皮层之间的连接。你可能觉得实验结果就是雪貂会严重致残，但并没有：听觉皮层学会看，视觉皮层学会听，而且雪貂没事。一般的哺乳动物，其视觉皮层都包含一张视网膜地图：皮层中，与视网膜附近区域相连的神经元彼此相互接近。相反，大脑被重新布线的雪貂在听觉皮层中形成了这张视网膜图。如果视觉信息重新导入躯体感觉皮层（负责感知触觉），躯体感觉皮层也会学会看。其他哺乳动物也有这样的能力。

对于天生看不见的人，视觉皮层可以负责大脑的其他功能。对于听不见的人，听觉皮层也可以这么做。盲人可以借助舌头来学会“看”，方法是将头戴式摄像机的视频图像发送至舌头上的一组电极上，高电压与高像素对应，低电压与低像素对应。本·安德伍德是盲人，小时候就自学像蝙蝠那样，用回声定位来导航。

通过咂舌头、听回声，他能够到处走动且不会撞到障碍物，会踩滑板车，甚至还能打篮球。所有这些例子都证明，大脑自始至终只使用了一种相同的学习算法，那些负责不同知觉的区域，区别也仅仅在于与其相连、输入信息的器官（如眼睛、耳朵、鼻子）。反过来，关联区（大脑的各个皮层）通过与不同的感觉区（各个感觉器官）相连，来实现其机能，而执行区则通过连接关联区来实现其机能，然后输出反馈。

通过在显微镜下观察皮层可以得出相同的结论。同样的布线模式不断重复，随处可见。皮质是一个6层的柱状物，反馈回路达到大脑一个叫丘脑的结构，以及短程序抑制连接和远程兴奋性连接反复出现的模式。虽然表现出一定数量的变异，但这看起来更像是来自同一算法（而不是不同算法）的不同参数或者设置。低级感官领域会有更为明显的差异，但正如重新布线实验表明的那样，这些都不具有决定性。小脑是比大脑更早进化的部分，负责简单的运动调节，有着非常明显且有规律的架构，由小很多的神经元构成，因此，看起来至少动作学

习使用的是不同的算法。然而，如果一个人的小脑受到损伤，大脑皮层会接管它的机能。人的生物进化过程保留了小脑，但这并不意味着小脑能做大脑皮层不能做的事情，只是因为小脑更加高效。

自始至终，大脑构造中发生的运算也同样相似。大脑中的所有信息都以同样方式（通过神经元的放电模式）来表示。学习机制也相同：记忆通过加强集群放电神经元之间的连接得以形成，涉及一个叫作长时程增强的生物化学过程。不仅人脑是这样，不同的动物，其大脑运行机制都很相似。我们的大脑异常大，但似乎与其他动物一样，其构建遵循同样的原则。

证明大脑皮层统一性的另一个证据来自所谓的基因组贫乏。人类大脑中的连接数量是基因组中字母数量的100万余倍，因此从物理角度，基因组不可能弄明白大脑构造的细节。

然而，关于大脑是终极算法这个观点的最重要论据，就是大脑负责我们能感知以及想象的一切。如果某物存在，但大脑无法对其进行学习，那么我们就不知道它的存在。我们可能只是没看见它，或者认为它是随机出现的。不管怎样，如果我们将大脑放入计算机中运行，那个算法就能掌握我们能学会的一切。因此发明终极算法的一种途径（可以说是最流行的一种）就是对人脑进行逆向解析。杰夫·霍金斯（Jeff Hawkins）在他的著作《人工智能的未来》（*On Intelligence*）中对此进行了尝试。雷·库兹韦尔（Ray Kurzweil）把他的希望放在奇点上——人工智能的崛起远远超过人类的多样性。这样做的同时，他还在《如何创造思维》（*How to Create a Mind*）一书中对此进行了尝试。虽然如此，我们会看到，这仅仅是几个可能途径中的一个。这甚至不一定是最有可能的一个，因为大脑非常复杂，而我们还处于解密大脑的初级阶段。另一方面，如果我们找不到终极算法，奇点也不会很快发生。

并不是所有的神经系统科学家都相信大脑皮层的统一性，在我们肯定这个观点之前，需要学习很多东西。关于大脑能掌握以及不能掌

握的东西，这个问题也引起了业界激烈的讨论。如果有我们知道但大脑不能学习的东西，那么这个东西肯定已经通过进化被掌握了。

来自进化论的论证

生物多样性源于单一机制：自然选择。值得注意的是，计算机科学家对该机制非常熟悉：我们通过反复研究尝试许多备选方法来解决问题，选择并改进最优方案，并尽可能多地尝试这些步骤。进化论是一种算法。套用查尔斯·巴贝奇（维多利亚时期的计算机先驱人物）的观点，上帝创造的不是物种，而是创造物种的算法。达尔文在《物种起源》的总结部分提到的“无限形体，美丽至极”掩饰了最美的统一性：所有这些形体都被编码在DNA中，所有这些形体都通过改变和连接这些染色体来表现。只通过该算法的一个描述，谁会猜出它产生了你和我？如果进化论这个算法能学习我们，可以想象它还可以学习能学习到的一切，条件是我们将进化论这个算法运用到足够强的计算机上。的确，在机器学习领域，通过模仿自然选择来使程序进化是许多人正在努力做的事情。因此，进化论是另外一个有希望通往终极算法的途径。

利用足够多的数据，一种简单的算法能掌握什么？关于这个问题，最经典的例子就是进化论。输入进化论这个算法的信息是所有存在过的、活着的生物的经历以及命运（对现在的算法来说是大数据）。此外，这个进化论算法已经在地球上最强大的计算机运行了300多万年——这台强大的计算机就是地球自己。运行这个算法的真正计算机应该比地球这台“计算机”运转得更快、数据密集性更低。哪一个模型更适合终极算法：进化还是大脑？这是和机器学习有关、自然

与培育之间的辩论。正如我们的存在依靠的是自然与培育的共同力量，也许真正的终极算法包含这两个方面。

来自物理学的论证

在1959年的一篇著名文章中，物理学家、诺贝尔物理学奖得主尤金·维格纳惊叹“数学在自然科学中不可思议的有效性”。由少量的观察推导出规律，是什么神奇的力量让这些规律可以运用到超出其预测范围的领域？这些规律都是基于数据得来的，而为什么这些规律比数据还要准确好几个数量级？最重要的是，为什么简洁、抽象的数学语言能够如此精确地解释我们无限复杂的世界？维格纳觉得这是一个很大的谜，觉得既幸运又无法理解。数学就是如此，而且终极算法就是其逻辑的延伸。

如果这个世界仅仅是一个不断变大、喧哗嘈杂的困惑体，那么我们有理由怀疑通用学习算法的存在。但如果我们所经历的一切，仅仅是几个简单规律的产物，那么单个算法能推导出所有一切能推导的东西，就是可以理解的。终极算法要做的就是提供一条捷径，通过实际观察，用简短的算式推导（而不是长长的算式）来得出这些规律的结果。

例如，我们虽然相信物理定律引起进化，但是不知道具体怎么进行。我们知道自己可以像达尔文那样，通过观察直接推导出自然选择规律。无数错误的推论就是由那些观察得出的，但多数人不会做出那些错误的推论，因为我们对世界有着广泛的认识，会对自己的推论形成良性影响，而且那些认识也与自然规律相符。

物理规律之美多大程度渗透到更高的领域（如生物学、社会学），这一点有待观察。但对混沌的研究提供了许多诱人的例子，这

些例子和拥有相似行为的不同系统相关，而普适性理论可以解释这些例子。曼德布洛特集合（Mandelbrot Set）就是很完美的例子，能解释一个很简单的重复程序如何产生无数种类的形式。如果世界上的山峰、河流、云朵以及树木都是这些重复程序的产物（分形几何学表明它们就是），也许那些程序只是单个程序的不同参数化，而该单个程序可以从那些程序推导中得出。

在物理学中，适用于不同数量的方程，往往可以用来描述发生在不同领域的现象，例如量子力学、电磁学、流体动力学。波动方程、扩散方程、泊松方程表明：一旦我们在某个领域发现它们，也很快能在其他领域发现它们；一旦我们在某个领域懂得解开它们，也能在所有领域将它们解开。此外，所有这些方程都很简单，涉及几个和空间、时间有关的数量的相同导数。很容易想象，它们都是主方程的几个例子，而终极算法要做的，就是用不同的数据集来将它实例化。

另外的证据来自最优化。最优化是数学的分支，关注的是为函数找到输入值，使其产生最大输出值。例如，找到购买及销售股票的排序，用来最大化你的全部回报，这就是一个最优化问题。在最优化中，简单的函数往往能引出惊人的复杂方案。最优化几乎在每个领域都扮演十分重要的角色，包括科学、技术、商业，还有机器学习。每个领域在约束条件下进行最优化，该限制条件由其他领域的最优化状态来决定。我们努力在经济的限制下将幸福感最大化，这也是公司在受到当前技术水平限制下的最佳方案，反过来成为我们在生物学及物理学限制下能找到的最佳方案。反过来，生物学是进化学在物理学和化学的约束下进行优化的结果，而物理定律本身又是最优化问题的解决方法。因此，可能所有事物的存在，都是一个中心优化问题进一步的解决方案，而终极算法随着那个中心问题的叙述而产生。

不仅物理学家和数学家在寻找不同领域之间意想不到的联系，生物学家也在寻找。在《论契合：知识的统合》（*Consilience*）一书中，著名生物学家爱德华·威尔逊慷慨激昂地阐释了知识（从科学到

人文学）的统一性。终极算法就是该统一性的完美表达：如果所有知识共同遵循一个模式，那么终极算法就存在，反之则不存在。

然而，物理学的简洁性独一无二。在物理学和工程学之外，数学的轨迹就更加混合。有时数学仅限于用起来有效，而有时它的模型又过于简单，无法使用。然而，过于简单的倾向源于人类思维的各种限制，而不是源于数学的种种限制。大脑的大多数硬件（或许该叫“湿件”）负责人体感知和活动，而为了做算术，我们就得借用因语言得到进化的那部分大脑。计算机就没有这样的限制，而且可以轻易地将大数据变成非常负责的模型。在数学的过度有效性与数据的过度有效性面前，你就会选择机器学习。生物学和社会学绝对不会像物理学一样简单，但我们发现其真理的方法可以做到那样简单。

来自统计学的论证

根据一个统计学流派的观点，所有形式的学习都是基于一个简单的公式——如我们所知，就是贝叶斯定理。贝叶斯定理会告诉你，每当你看到新的证据后，如何更新你的想法。一种简单的贝叶斯学习算法对世界进行一系列假设，由此开始进行学习。当它看到新的数据时，与该数据匹配的假设更有可能会成立（或者不可能成立）。在观察足够的数据后，某个假设会成立，或者几个假设同时成立。例如，我在寻找一个能够准确预测股票走势的程序，该程序预测某只股票会下跌，结果该股票却上涨了，那么该程序就会失去我的信任。我审核几个备选程序之后，只选择了几个可信赖的程序，它们概括了我对股票市场的新认识。

贝叶斯定理就是将数据变成知识的机器。据贝叶斯统计学派的观点，贝叶斯定理是将数据变成知识的唯一正确方法。如果该学派的观

点正确，贝叶斯定理要么就是终极算法，要么就是推动终极算法发展的动力。关于贝叶斯定理使用的方法，其他统计学派持非常保守的观点，而且会更愿意用不同方法来对数据进行学习。在计算机发明出来之前，贝叶斯定理只能应用在非常简单的问题中，说它是通用的学习算法未免有点牵强附会。然而，在大数据和大计算的辅助下，贝叶斯定理在广阔的假设空间中找到了出路，而且已经扩展到每个人们能想到的领域中。如果说存在贝叶斯算法无法学习的东西，只是现在还没发现它们。

来自计算机科学的论证

我在大四时，用了一个夏天玩俄罗斯方块游戏，这是一个涉及方块叠加的电子游戏，游戏中由正方形组成的各种形状的图案往下掉，你要将这些图案堆起来，堆得越紧密越好。如果图案堆到屏幕顶部，那么游戏就结束了。当时我完全没有意识到，这就是我接触NP完全问题[4]的开始，这是理论计算机科学最重要的一个问题。后来我才知道，俄罗斯方块完全不是简单用来消遣的游戏，掌握这个游戏（彻底掌握它），就是你这辈子做得最有用的事情。如果你一步到位，解决了俄罗斯方块问题，你就解决了科学、技术、管理中数千个最难、最有意义的问题，因为本质上这些难题就是同一个问题。这是在所有科学领域中最让人惊讶的事实。

弄明白蛋白质如何折叠成特定形状；通过DNA来重新构建一系列物种的进化史；在命题逻辑中证明定理；利用交易成本来发现市场中的套利机会；从二维视图中推出三维形状；将数据压缩到磁盘上；在政治活动中组成稳定联盟；在剪切流中模拟湍流；按照给定回报率找出最安全的投资组合、到达几个城市的捷径、微芯片上元件的最佳布局

方案、生态系统中传感器的最佳布局、自旋玻璃门最低的能量状态；安排好航班、课程、工厂工作；最优化资源分配、城市交通流、社会福利，以及提高你的俄罗斯方块分数（最重要的）——这些都是NP完全问题，意思是，如果你能有效解决其中的一个问题，就能有效解决所有NP类问题，包括相互间的问题。谁会猜到，这些表面上看起来迥然不同的问题，会是同一个问题？如果它们真的是同一个问题，就可以说一种算法能学会解决所有问题（或更准确地说，所有能有效解决的例子）。

在计算机科学中，P和NP是两类最重要的问题（很遗憾，名字不是很有助于记忆）。如果我们能有效解决它，那么这个问题就属于P；如果我们能有效找到其解决方案，那么这个问题属于NP。著名的P=NP的问题就是，能有效找到的问题是否可以得到有效解决。因为NP完全问题，回答这个问题需要的只是证明某个NP完全问题可被有效解决（或者无法被有效解决）。NP在计算机科学领域并不是最难的一类问题，但可以说，它是最难的“现实”类问题：如果在宇宙灭亡之前，你无法找到问题的解决方法，那你努力解决这个问题的意义在哪里？人类擅长给出NP难题的近似解，而相反，我们感兴趣的问题（如俄罗斯方块问题）往往涉及NP问题。人工智能的其中一个定义是，人工智能包括找到NP完全问题的所有启发性解决方案。为了找到解决方案，我们常常把问题变成可满足性问题，也就是典型的NP完全问题：给定的逻辑公式是否永远都是对的，或者它是不是自相矛盾？如果我们发明一种学习算法，能够学习解决可满足性问题，那么有充分理由认为，这个算法就是终极算法。

抛开NP完全问题，计算机的存在本身就明显预示着终极算法的存在。如果你穿越回到20世纪早期，告诉人们很快会有一种机器发明出来，能够解决人类所有领域的难题——所有难题都通过同一台机器解决，那么没有人会相信你。人们会说，每台机器只能解决一个问题：缝纫机不会打字，打字机不会缝纫。1936年，艾伦·图灵想象出一个奇怪的装置，它有一条纸带和机器头，头可以在纸带上进行阅读和书

写，就是现在人们知道的图灵机。每一个可以想得到的、可以用逻辑推理解决的难题，都可以通过图灵机解决。此外，一台所谓的万能图灵机可以通过阅读纸带上的具体要求来模仿所有东西，换句话说，我们能够对图灵机进行编程，用它来做所有事情。

算法是归纳的过程，而学习的过程对图灵机来说，就是演绎的过程。图灵机能通过对算法输入、输出行为进行阅读来模仿其他算法。就像存在许多与图灵机对等的计算模型，可能也存在通用学习算法的许多不同的等价公式。然而，问题的关键是必须找到第一个这样的公式，就像图灵找到通用计算机的第一个公式那样。

机器学习算法与知识工程师

当然，有很多人支持终极算法，也有很多人怀疑终极算法。当某方法可以简单解决复杂问题时，存在怀疑符合情理。对终极算法最坚定的反抗来自机器学习永恒的敌人：知识工程。根据知识工程支持者的观点，知识无法自动被学习，必须通过人类专家编入计算机，才能对它进行学习。的确，学习算法能从数据中提取一些东西，但你不能将这些东西和真知识混为一谈。对知识工程师来说，大数据不是新石油，而是骗人的新蛇油。

在人工智能出现早期，机器学习似乎是通往类人智能计算机的途径。图灵和其他人认为，机器学习是唯一看似合理的途径。但后来知识工程师进行了回击，而且20世纪70年代机器学习处于次要地位。在20世纪80年代的一段时间，似乎知识工程师要接管世界了，还有许多企业和国家对知识工程领域进行大量投资。但后来人们开始对该领域失望，而机器学习也开始崛起，一开始悄无声息，后来就突飞猛进。

尽管机器学习成功了，知识工程师们还是觉得不信服。他们相信，机器学习的局限性很快会变得明显，钟摆会摆回来，局势会扭转。马文·明斯基是麻省理工学院的教授、人工智能的先驱人物，也是该阵营的重要成员。明斯基不仅怀疑机器学习能替代知识工程，他也怀疑人工智能的所有统一思想。明斯基在其《意识社会》（*The Society of Mind*）一书中提到了关于智能的理论，这个理论可以被不客气地归纳为“意识就是一个接一个该死的东西”。《意识社会》包含的就是一长串分散的观点，每个观点都毫不相关。这种实现人工智能的方法根本没什么用，它只是由计算机进行的收集活动。没有机器学习，需要建立智能代理的观点将会变得无限多。如果一个机器人掌握了人类所有的技能，但就是没有学习能力，那么人类不久就会把它扔在一边。

明斯基是Cyc项目（Cyc project）的狂热支持者，这是人工智能历史上最臭名昭著的失败项目。Cyc项目的目标是通过将所有必要知识输入计算机中，来解决人工智能问题。20世纪80年代这个项目刚开始时，它的领导者道格·莱纳特（Doug Lenat）就信心满满地预测，10年之内该项目就会取得成功。30年后，Cyc项目不停扩大，但仍无法做常理性推理。具有讽刺意味的是，莱纳特终于支持通过挖掘网页来将Cyc填满，这并非因为Cyc可以阅读，而是因为别无他法。

即使奇迹发生，我们能够对所有必要的数据进行编程，麻烦也会不断出现。过去几年，几个研究组已经尝试构建完整的智能代理，方法就是将所有算法集中起来，用于想象、语音识别、语言理解、推理、计划、导航、操作等。没有统一的结构，这些尝试将碰到“复杂性”这个难以解决的难题：有太多的活动件、太多的交互、太多的漏洞，可怜的人类软件工程师也难以应付。知识工程师相信，人工智能的问题仅仅是工程学的问题，但是我们还没达到那个点——工程学能带领我们走完下面的路。1962年，肯尼迪发表登月演讲。那时登上月球是一个工程学问题，但1662年，它就不是了，而当今它则更加靠近人工智能所在的领域。

在工业领域中，除了在一些利基领域，没有任何迹象表明知识工程学可以永远和机器学习竞争。为什么要花费精力来让专家缓慢而痛苦地将知识编码成计算机能识别的形式，而你明明可以在一秒内将其从数据中提取出来？你会怎么对待那些专家不懂，你却可以从数据中发现的东西？而当数据不足时，知识工程学的成本倒是很少会超过其带来的益处。反过来，想象一下，如果农民要将每株玉米进行工程化，而不去播种并让它们生长，那么我们都得挨饿。

另一个对机器学习持怀疑态度的人是语言学家诺姆·乔姆斯基。乔姆斯基认为，语言必须是与生俱来的，因为孩子听到的合乎语法的句子仅仅是一些例子，不足以学习语法。然而，这种说法仅仅将学习语言的任务交给了进化，它并没有反对终极算法，只是反对“终极算法是大脑”这个观点。此外，如果存在通用语法（乔姆斯基认为存在），阐发它就是阐发终极算法的步骤之一。这种情况不成立的唯一可能就是，语言和其他认知能力没有共同点，考虑到进化的近因，这令人难以置信。

无论如何，如果我们将乔姆斯基“刺激贫乏”论形式化，我们会发现这个观点很明显是错的。1969年，霍宁证明，概率上下文无关语法（probabilistic context-free grammar）只能通过正面例子掌握，后面紧跟的是更有力的结果（上下文无关语法是语言学家研究的内容，而概率类型模拟每个规则被使用的概率）。另外，语言学习不会发生在一个真空当中，孩子需要从父母和周围环境获取各种语言学习线索。如果我们能从几年时间里学习的例子中学习语言，部分也只是因为语言结构与世界结构存在相似性。对这个共同结构，我们感兴趣，而且从霍宁和其他人那里知道，有这个共同结构就足够了。

总体来讲，乔姆斯基批评所有统计学习。他把统计学习算法不能学习的东西列了一个单子，但这个单子已经过时50年了。乔姆斯基似乎把机器学习等同于行为主义了，根据行为主义，动物的行为沦为反

应与奖励之间的联合，但机器学习不是行为主义。现代学习算法能够掌握丰富的内在表象，而不仅仅是刺激物之间的两两关系。

最后，事实胜于雄辩。统计学语言算法起作用了，而手工设计的语言系统却没起作用。第一件令人大开眼界的事发生于20世纪70年代，当时五角大楼的研究机构DARPA组织了第一个大型语音识别项目。让所有人惊讶的是，一种简单的序列学习算法——乔姆斯基嘲笑的类型，轻易地打败了一个复杂的知识系统。现在像这样的学习算法几乎用于每一个语音识别器中，包括Siri（苹果公司产品上的一项智能语音控制功能）。弗雷德·贾里尼克（IBM语音研究组的领导）说过一句著名的俏皮话：“每开除一名语言学家，我的语音识别系统的错误率就降低一个百分点。”20世纪80年代，陷入知识工程学的泥潭里，计算机语言学差点走向尽头。自那以后，基于学习算法的浪潮已经席卷这个领域，在计算机语言学会议中，几乎每篇文章都会提到学习。统计分析软件以近乎人类水平的精确度来分析文章，而手编程序已经远远落在后面。机器翻译、拼写纠正、词性标注、词义消歧、问题回答、对话、概括——这些领域的所有最好的系统都利用了学习。没有学习，沃森不可能在《危险边缘》游戏中战胜人类。

对此，乔姆斯基可能会回应，工程学的成功并不能证明其科学有效性。换句话说，如果你的楼房倒塌了，而且你的发电机不工作了，那么也许就是因为你的物理学观点有问题。乔姆斯基认为，语言学应该把重点放在他定义的“理想的”说话者和听话者上，这让他忽略了诸如类似的问题：语言学习过程涉及统计学。因此，很少有实验主义者拿他的理论当回事，这并不奇怪。

另外一个可能会反对终极算法的观点来自心理学家杰瑞·福多，他认为心理是由一系列模块组成的，这些模块之间只有有限的联系。例如，当看电视时，你的“高级脑”知道，那只是光线在光滑表面的闪烁，但视觉系统仍然会看见三维形状。即使我们相信心理模块理

论，这个理论也并没有暗指不同的模块会使用不同的学习算法。同种算法对诸如视觉及语言之类的信息都起作用，这个说法才足够有力。

像明斯基、乔姆斯基和福多这样的批评家曾经占据上风，但万幸，他们的影响力已经逐渐减弱。即便如此，我们仍需将他们的批评铭记于心，这样才能到达终极算法这个终点，原因有两个：第一，知识工程师和机器学习算法一样，遇到许多相同的问题，虽然他们没有成功，但学到了许多宝贵的教训；第二，学习和知识以异常微妙的形式相互交织，而我们很快就会发现这一点。遗憾的是，这两个阵营各说各话。他们讨论不同的主题：机器学习讨论概率，而知识工程学讨论逻辑。本书后面会提到如何解决这个问题。

天鹅咬了机器人

“无论你的算法有多聪明，总有它无法掌握的东西。”除了人工智能和认知科学，反对机器学习的常见观点几乎都可以用这句话概况。纳西姆·塔勒布（Nassim Taleb）在《黑天鹅》（*The Black Swan*）一书中强调了这个观点。有些事真的无法预料。如果你只见过白天鹅，会觉得看到黑天鹅的概率是0。2008年的金融危机就是一只“黑天鹅”。

有些事可预料，而有些事却不能预料，这个说法是正确的，而机器学习算法的首要任务就是区别可预测的事与不可预测的事。但终极算法的目标是要学习一切能认知的东西，这比塔勒布和其他人想象的要广阔得多。房地产泡沫还远远不是一只“黑天鹅”，相反，房地产泡沫是经过人们普遍预测的。大多数银行的模型没能预测它的到来，也只是因为那些模型的局限性，而不是机器学习的局限性。学习算法很擅长精确预测稀有、未曾发生的事件。甚至你也可以说，这是机器

学习的主要任务。如果你没见过黑天鹅，那么它出现在你面前的概率是多少？它是已知物种的一部分，最后变成黑色的天鹅，这样的概率是多少？这仅仅是一些粗略的例子，我们会在本书看到更深刻的例子。

另外一个反对机器学习的观点与以上观点相关，就是我们常听到的——“数据无法代替人类的直觉”。实际上，这句话可以反过来：人类直觉无法代替数据。直觉就是你在不知道事实的情况下依靠的东西，而因为你不常用它，所以直觉非常宝贵。但如果证据摆在你面前，为什么还要拒绝证据？统计分析在棒球界打败球探（正如迈克尔·刘易斯在《魔球：逆境中制胜的智慧》一书中明确记录的那样），在品酒时打败内行。统计分析能做很多事情，我们每天都能看到新的例子。因为大量数据的涌入，证据与直觉的界限正在迅速改变，而正如所有革命一样，要抛弃所有墨守成规的方法。如果我是Y公司X领域的专家，我就不想被某人用数据推翻。行业里有句话：“多听听顾客的话，而不是HiPPO。”（HiPPO是“领最高薪水的人说的话”的简写。）如果想成为明天的权威人士，你要依靠数据，而不是与之斗争。

好了，有人会说，机器学习能从数据中找到统计规律，但它绝不会发现更深刻的东西，如牛顿定律。可以说，它还没有找到那样深刻的定律，但我肯定它将来会找到。尽管有苹果落下的故事，但深刻的科学真理并不是那么容易就能获得。科学经历了三个时期：布拉赫时期、开普勒时期、牛顿时期。对于布拉赫时期，我们收集了很多数据，就像第谷·布拉赫日复一日、年复一年耐心记录行星的位置那样。对于开普勒时期，我们使经验规律符合数据，就像开普勒对行星运动所做的那样。对于牛顿时期，我们发现了更深刻的真理。大多数科学研究和布拉赫、开普勒所做的工作相似，这样的工作就是科学研究的内容，像牛顿偶然发现定律的例子则少见。当今，大数据所做的工作是布拉赫的数十亿倍，机器学习的工作内容是开普勒的数百万倍。如果（但愿如此）有更多像牛顿偶然发现定律这样的时刻，这样

的时刻也可能发生在未来的学习算法中，或者发生在未来手足无措的科学家身上，或者至少是发生在两种可能都存在的情况下（当然，诺贝尔奖会颁发给科学家，不管他们是持重要的观点，还是只按了一下按钮。学习算法就没有那样的志向，要拿诺贝尔奖）。本书将会提到那些算法是怎样的，并推测它们会发现什么，例如，治愈癌症的方法。

终极算法是狐狸，还是刺猬

我们有必要考虑藏得更深、反对终极算法的观点，这个观点可能是所有反对观点中最严肃的一个。这个观点不是来自知识工程师或者不满意的专家，而是来自机器学习实践人员。假设我是持反对观的机器学习实践者，可能会说：“终极算法和我日常生活看到的不一样。我尝试用许多学习算法的数百种变形来解决所有给定的问题，而且对各类不同问题都会有更好的算法，那么单个算法（我们说的终极算法）怎么可能代替所有这些算法？”

这个问题的答案是：的确如此。不用尝试多种算法的数百种变形，而只用尝试单个算法的数百种变形，这不是更轻松吗？只要我们弄明白，每个算法中重要的与不重要的东西，重要部分的共同点，以及这些部分如何进行互补，那么我们真的可以从这些多种算法中合成一个终极算法。这就是我们在本书中要做的事情，或者说尽可能要做到的事情。亲爱的读者，也许你在阅读本书时，会有自己的一些观点。

终极算法会复杂到什么程度？它包含几千行代码？还是几百万行？我们现在还不知道，但机器学习有一段可喜的历史：简单的算法意外地将精心设计的算法打败了。在《人工科学》（*The Sciences of*

the Artificia）一书的著名章节中，人工智能先驱人物、诺贝尔奖得主赫伯特·西蒙（Herbert Simon）让我们想象蚂蚁费力地穿过沙滩回家。蚂蚁的路线非常复杂，这不是因为蚂蚁本身复杂，而是因为沙滩这个环境对蚂蚁来说意味着要爬很多山丘，绕很多卵石。如果我们通过对每条可能的路线进行编程，模仿蚂蚁，那么我们注定会失败。同样，在机器学习中，复杂性存在于数据中。终极算法需要做的就是消化复杂性，因此，如果终极算法变得非常简单，那么我们也不用感到惊讶。虽然人类的手很简单（四个手指，一个大拇指），但是它却可以制作并使用无数种工具。终极算法与算法的关系，就如同手指与钢笔、剑、螺丝刀、叉子的关系。

正如以赛亚·伯林明确提出的那样，有些思想家就是狐狸——他们知道许多微小的事情；而有些思想家则是刺猬——他们知道一件大事。学习算法也是同样的情况。我希望终极算法是一只刺猬，但即使它是只狐狸，我们也没法很快抓住它。当今学习算法最大的问题，不是它们数量太多，而是尽管它们有用，却不能完成我们让它们做的所有事情。我们利用机器学习来发现深刻的真理之前，得先找到关于机器学习的深刻真理。

我们正面临什么危机

假设你被诊断患有癌症，而且传统疗法（手术、化疗、放疗）都失败了，那么接下来发生的事情就会决定你是活下去，还是走到生命尽头。第一步就是要对肿瘤进行基因排序。诸如在剑桥、马萨诸塞州的基础医学公司会为你做这些工作：把肿瘤样本邮寄给他们，然后他们会发给你一个列表，列表是已知的、和癌症相关的基因变异。这个步骤十分有必要，因为每种癌症都不一样，单种药不可能治疗所有癌

症。当癌症扩散到全身时会变异，通过自然选择，最能抵抗你所服用药物的变异细胞最有可能继续会生长。对你有用的药物可能只对5%的病人有用，或者你需要结合其他药物一起服用，这些药可能你之前从未服用过。也有可能要设计一种新药，专门治疗你的癌症，或者需要一系列的药来避开癌症的适应性。这些药物可能会有副作用，而且对你来说会致命，但对其他很多人来说可能没有问题。即使了解你的病历和癌症基因，也没有哪个医生可以记录你所有的病情，以便预测最好的疗法。对机器来说，这是一个完美的任务，但当今的学习算法还无法完成这个任务。终极算法就是一个完整包：将终极算法应用于大量的患者及药物数据中，同时参考从生物医学文献中挖掘的知识，这就是我们将来治疗癌症的方法。

许多领域迫切需要通用学习算法，包括与生死有关的领域以及普通领域等。你可以想象理想的推荐系统是什么样的，它能推荐书籍、电影以及小玩意儿，它们正是你有时间慢慢细看时会挑选的东西。亚马逊的算法与这个系统则大相径庭，部分是因为亚马逊的算法没有足够的数据——它知道的主要信息仅仅是你之前从亚马逊购买的东西——但如果你气疯了，把自出生以来能想到的东西都一股脑地输给它，那么它就不知道该拿这些东西怎么办了。你如何将生活的万花筒、做过的各类选择转化成连贯的画面，用来告诉你：你是谁，你想要什么？这是当今学习算法无法理解的。但有了足够的数据，终极算法将能够大概了解你以及你最好的朋友。

未来某一天，每个房间都会有一个机器人，做饭、铺床，甚至在父母去上班时照看孩子。这一天要多久才来，取决于寻找终极算法的过程有多艰难。如果我们能做的，只是将许多不同的学习算法结合起来，每种算法只能解决人工智能的一小部分问题，那么很快我们就会撞到复杂性这堵墙。这种零碎的方法在《危险边缘》比赛中奏效了，但很少有人相信，未来的家用机器人就是沃森的子孙。这并不是说终极算法会单枪匹马破解人工智能的难题，还有许多伟大的工程要完

成，沃森就是一个很好的开始。但“二八原则”也适用：终极算法会提供80%的方案，做20%的工作量，所以这是开始的最佳时机。

终极算法对技术的影响不仅限于人工智能。通用的学习算法是打击复杂性怪兽的有力武器。当今人类建立起来的很复杂的系统将来会变得简单。计算机会在我们更少的辅助下做更多的事情。它们不会不断重复同一些错误，而会像人一样，从实践中学习经验。有时，就像传说中的管家，我们还没说想要什么，计算机就已经先猜出来了。如果计算机能让我们变聪明，那么运行终极算法的计算机会让我们感觉自己就是天才。技术进步的步伐会明显加快，不仅仅在计算机科学，在许多不同的领域也会这样。这就反过来推动经济发展，降低贫困率。终极算法会辅助汇总和传播知识，这样一个机构的情报会比其各个分机构的情报总数还要多，而不会更少。日常工作将自动化完成，并由更有意思的工作来代替。每项工作都会比当今完成得更好，无论这个工作是由更熟练的人、计算机，或者通过二者的结合来完成。股市崩盘的概率会越来越低，规模也会越来越小。传感器会在地球上形成密集的网格，人类掌握的模型会持续接收终极算法输出的信息，这样我们就不会盲目飞行了，地球的情况会变得越来越好。你的一个模型会代表你和世界进行谈判，和其他人及实体模型玩复杂的游戏。因为有了这些，我们会更长寿、更幸福，也更多产。

因为反对观点的潜在影响太大，我们应该努力发明终极算法，哪怕成功的概率很低。即使这个过程会很久，但找到一种通用学习算法却有很多能即刻感受到的好处。其中一个就是我们能通过统一观点更好地了解机器学习。有太多的商业决策是在不了解统计学的情况下做出来的，而统计学对商业决策起着支撑作用。事情本该不是这样的。为了使用一项技术，不必掌握其内部工作原理，但我们得有关于它的一个好的概念模型。我们有必要知道如何找到收音机上的一个电台，或者懂得如何调音量。当下，那些不是机器学习专家的人，对学习算法会用来做什么，没有什么概念模型。我们使用谷歌、脸书时驱动的算法，或者最新的分析套件，有点像一辆带有有色窗户的黑色豪华轿

车，在某个夜晚神秘地出现在我们的家门口：我们该上车吗？这辆车会带我们去哪里？现在是时候坐在司机的座位上了。明白不同的算法所做的假设会帮助我们选择合适的算法用于工作中，而不是从偶然出现的算法中随机挑一个用，然后忍受它好几年，最后痛苦地领悟从一开始我们就该知道的东西。通过了解学习算法优化的内容，我们可以肯定它们优化的是我们关注的东西，而不是装在盒子里的东西。也许最重要的是，一旦我们知道特殊的学习算法得出的结论，就会知道用这些信息来做什么——该相信什么，该如何回报发明者，以及下次该如何取得更好的结果。有了通用学习算法（我们在本书中将其作为概念模型），我们就能在没有认知负载的情况下，把所有这些事情做完。机器学习本质上是简单的，我们只需削掉数学及行话这些外皮，然后把最里面的“俄罗斯套娃”展示出来。

这些好处都可应用于我们的私人生活和工作中。我们在现代世界留下自己的印记，数据记录了我们的每一个印记，但我们应该如何充分利用这些数据呢？每个互动都有两个方面：这个互动为你完成了什么；对于刚和你交互的系统，它教会了这个系统什么。懂得这些是在21世纪过上幸福生活的第一步。教授学习算法，这些算法就会为你服务，但首先你得了解它们。我工作的哪些部分可以交给学习算法来完成，哪些不可以？最重要的是，我该如何利用机器学习把工作做得更好？计算机是你的工具，而不是对手。有了机器学习的辅助，经理会变成超级经理，科学家会变成超级科学家，工程师会变成超级工程师。未来属于那些深深懂得如何将自己的独特专长与算法的擅长结合起来的人。

也许终极算法就像一个潘多拉盒子，最好不要打开。计算机会奴役甚至消灭我们吗？机器学习会变成独裁者或者邪恶公司的侍女吗？知道机器学习的发展方向有助于帮助我们了解该担心什么、不该担心什么、应该怎么处理问题。《终结者》中，超级人工智能变得有情感，并通过机器人军队征服了人类。这个场景不会和我们将在本书中谈到的学习算方法一起发生。因为计算机会学习，并不意味着它们可

以魔法般地实现自己的愿望。学习算法学着完成我们为它们设定的目标，它们不会改变这些目标。我们要担心的是，它们服务我们的方法可能会对我们有害，而不是有益。因为它们知道的东西不多，改善的方法就是教它们更好的方法。

很多时候，我们得考虑，如果终极算法落入坏人手中，它会做些什么。第一道防线就是确保好人第一个拿到它，或者如果它不明白谁是好人，就要保证它是开源的。第二道防线就是要意识到，无论学习算法有多好用，也只是在获得数据时好用。控制了数据的人也就控制了学习算法。你对数字化生活的反应，不应该是退回到木屋中——树林里也装满了传感器，而应努力拿到对你来说重要的数据。能有推荐系统为你找到想要的东西，并把东西带给你，这样很好，没有这些系统你会感到失落。它们带给你的应该是你想要的东西，而不是其他人想让你拥有的。控制好数据，控制好算法掌握的模型的所有权，这就是21世纪战争的内容，这些战争可能会发生在政府、企业、工会以及个人之间。为了共同利益，你也有道德义务来分享数据。只依靠机器学习不能治愈癌症，依靠癌症病人却可以做到这一点，方法就是为了将来的病人，分享自己的信息。

新的万有理论

当今的科学已经被彻底四分五裂，就像巴别塔中的亚社会都说着自己的俚语，只能看到相邻的几个亚社会。终极算法会给出所有学科的统一思想，并有潜力提出一套新的万有理论。乍一看，这个说法可能会有点奇怪。机器学习所做的，就是从数据中引出理论。终极算法本身如何能发展为一套理论？难道弦理论是万有理论，而终极算法和万有理论没有任何相似点？

为了回答这些问题，我们得首先明白什么是科学理论，什么不是。理论是关于世界是什么的一系列约束条件，而不是对世界的完整描述。为了获得对世界的完整描述，你必须将理论和数据结合起来。例如，想想牛顿第二定律。定律说明力等于质量与加速度的乘积，或者写成F=*ma*。定律并没有说明是哪个东西的质量或者加速度，或者作用力是什么。定律只要求，如果某物体的质量是m，加速度是*a*，那么作用在它之上的所有力的总和就肯定是*ma*。虽然该定律排除了宇宙的某些自由度，但没有排除所有。所有其他物理理论也都如此，包括相对论、量子力学以及弦理论，这些理论其实都是对牛顿定律的完善。

理论的强大之处在于它简化了我们对世界的描述。有了牛顿定律，我们首先只需知道某个时间点所有物体的质量、状态、速度，其次就是所有时段的状态及速度。凭借过去、未来宇宙历史中可区分时刻的数量这样一个因素，牛顿定律概括了我们对世界的描述。太了不起了！当然，牛顿定律也仅仅是接近准确的物理定律，因此让我们用弦理论来替代它，同时忽略弦理论是否永远证实有效的问题。我们能做得更好吗？可以，有以下两个原因。

第一，实际上，我们没有足够的数据来完全确定世界，甚至忽略不确定性原则，准确知道世界上所有粒子某个时间点的状态和速度，也远远做不到。因为物理定律是混沌的，不确定性随时会混杂进来，而且在短时间内能确定的东西太少。为了准确描述这个世界，每隔一段时间，我们就需要一批新数据。实际上，物理定律只告诉我们局部会发生的事情。这一点大大削减了它们的力量。

第二，虽然我们在某个时间点拥有关于世界的完整知识，物理定律还是不能让我们确定这个世界的过去和未来。这是因为，确定世界的过去和未来所需的全部计算量，对于能想象得出的计算机来说，超出了它们的能力范围。实际上，为了完善魔法宇宙，我们需要另外一个一模一样的宇宙。这也是为什么弦理论多数情况下在物理学之外就变得无关紧要了。我们在生物学、心理学、社会学或者政治学中的理

论，并不是由物理定律推理得来的，这些理论得从零开始构建。我们假定，当这些理论应用到细胞、大脑、社会中时，它们就是物理定律对此所做预测的近似理论，但我们无法知道。

不像特定领域的理论只在该领域中才有权威，终极算法在所有领域中都有权威。在X领域中，终极算法不如X领域的主流理论有权威，却在所有领域中比该主流理论有权威——当我们考虑到整个世界时——终极算法普遍比所有其他理论有权威得多。终极算法是所有理论的起源。为了获得X理论，我们要给终极算法添加的就是推导X理论所需要的最少量的数据（在物理学中，需要添加的仅仅是大约几百个重要实验的结果）。结果就是，在同样的情况下，终极算法很有可能成为万有理论的最佳出发点。请史蒂芬·霍金原谅，和弦理论相比，终极算法最后会告诉我们更多关于上帝思考的东西。

有些人可能会说，寻找通用学习算法就是一种追求虚荣心的表现。其实梦想并不是追求虚荣心，在魔法石以及永动机并肩作战下，也许终极算法在众多伟大的幻想中会代替虚荣心的位置。寻找终极算法更像测定海上的经度，人们一开始认为这太困难，于是放弃了，直到一个孤独的天才解决了这个问题。寻找终极算法更有可能就是一代一代人的任务，就像天主教堂是由一块块石头砌成的一样。找到终极算法的唯一方法就是，早早动身踏上旅途。

未达标准的终极算法候选项

那么，如果终极算法存在，它是什么？看起来很明显的一种候选项就是记忆：只要记住你见过的所有东西，过一段时间，你就好像见过世上的一切东西，所以也就无所不知了。这里的问题是，正如赫拉克利特说的那样，你无法两次踏入同一条河流。世上的东西比你能看

到的多得多。无论你观察过多少朵雪花，下一朵还是会不一样。即使宇宙大爆炸时你在场，但从那以后无论在哪里，对于未来你可能看到的事物，你现在看到的也只是一小部分。如果你目睹一万年前地球上的生命，也不会对你未来将看到的东西有什么影响。在某城市长大的人搬到另外一个城市，他不会陷入瘫痪，但只会记忆的机器人就会陷入瘫痪。此外，知识不仅是事实清单，知识的范围很广，而且还有结构。“人固有一死”比70亿条死亡声明要简洁得多。记忆无法像终极算法那样让我们做到这些。

终极算法的另外一个候选项就是微处理器。计算机里的微处理器可以看作单一算法，其任务是执行其他算法，就像通用的图灵机那样。而且微处理器可以运行一切可以想得到的算法，这由它的内存和速度上限决定。实际上，对一台微处理器来说，一种算法也只是另外一种数据。这里的问题是，如果微处理器单独工作，那么它什么也不会干，它只会待在那儿，一整天什么也干不了。它运行的算法从哪里来？如果这些算法由人类程序员来编码，那么就不会涉及任何学习行为。即便如此，人们还是会有一种感觉，认为微处理器是终极算法的完美模拟。微处理器不是运行一切特殊算法最好的硬件，最好的硬件应该是特定用途集成电路，专为那种算法而精确设计。我们却把微处理器几乎用于所有应用中，因为即使它效率较低，但却非常灵活。如果我们得为每个新的应用构建一个特定用途集成电路，那么信息革命也绝不会发生。同样，终极算法也不是学习一切特定知识最好的算法，最好的算法应该是已经编码了大多数知识的算法（或者说所有知识，这样数据就变得多余）。问题却在于从数据中得出知识，因为这样做更简单，成本也更低，所以学习算法越通用越好。

一个更加极端的候选项就是普通的“或非门”：这是一个逻辑开关，只有输入的是两个0时，才会输出1。本书前面提到，计算机是由晶体管组成的逻辑门构成的，所有的运算都可以简化为“且”“或”“非”逻辑门的结合。“或非门”只是一个“或”门后面接着一个“非”门：对于分离（“或”）的否定，就像“只要我不挨饿或生

病，我就开心”这句话的意思一样。“且”“或”“非”都可以通过“或非门”得以执行，所以“或非门”可以做所有事情，而实际上这也是一些微型处理器使用的东西。那么为什么或非门不是终极算法？当然“或非门”的简洁无与伦比，但遗憾的是，“或非门”不是终极算法，就像乐高砖不是万能玩具一样。对于玩具来说，“或非门”当然可以是一种万能建筑材料，但一堆乐高砖不会自发地将自己组装成玩具。这个道理同样适用于其他简单的运算格式，如Petri网（Petri Net是对离散并行系统的数学表示）或细胞自动机。

这里将谈到更为复杂的候选项。那些所有可以回应搜索的良好的数据引擎，或者统计软件包里的简单算法怎么样，可以代替终极算法吗？它们还不足够吗？这个数据引擎和简单算法是更大的乐高砖，但也仅仅是砖。数据引擎无法发现新东西，只能告诉你它知道的东西。即使数据库中所有的人都终究会死，它也不会将人会死的概率推广到其他人身上（数据工程师看到这一点会大吃一惊）。很多数据只是和验证假设有关，但总得有人先提出假设。统计软件包可以做线性回归以及进行其他简单的程序，但是做这些对它们能学的东西来说，限制很低，无论你提供给它们多少数据。虽然更好的软件包会跳过统计学和机器学习的灰色地带，但它们还是无法发现许多种知识。

好了，是时候坦白了：终极算法就是等式$U(X)=0$。这个等式不仅适合做T恤图案，还适合做邮票图案。啊？$U(X)=0$表达的是某未知数X（可能很复杂）的某函数U（可能很复杂）等于0。每个等式都可以简化为这种形式，例如，$F=ma$等于$F-ma=0$，那么如果你把$F-ma$当作F的一个函数U，则$U(F)=0$。一般情况，X可以是所有输入的数据，U可以是所有算法，所以可以肯定的是，终极算法和这个函数同样通用，那么这个函数就一定是终极算法了。当然，我只是开玩笑，这个函数不是终极算法，但它指出了机器学习中的真正危机：想到一个非常通用的学习算法，里面却没有足够的东西可以拿来用。

那么为了可以拿来用，一种学习算法至少要有多少内容呢？物理定律够了吗？毕竟，世界上的一切都遵守物理定律（我们相信是这样的），而且这些定律引出进化论，而通过进化论，又引出与大脑相关的学科。那么也许终极算法就隐藏在物理定律中，如果是的话，我们就要把它找出来。把数据输入到物理定律中，不会有新的定律产生。我们可以这样思考：也许某个领域的主理论就是物理定律，只不过是将该物理定律编成对该领域研究来说更方便的形式。如果是这样，我们就需要一种算法，可以找到连接该领域数据与该领域理论的捷径，而不知道物理定律能否在其中起作用。另外一个问题就是，如果物理定律不一样，终极算法还是有可能在许多例子中找到它们。数学家喜欢说，上帝可以违反物理定律，但他永远不能违抗逻辑规律。也许的确如此，但逻辑规律是用来演绎的。除了归纳，我们需要的是对等的东西。

机器学习的五大学派

当然，我们不必从零开始寻找终极算法，而有几十年的机器学习研究可以利用。地球上最聪明的那些人已经把他们的毕生贡献给发明学习算法，而有些人甚至声明，他们手上已经有通用的学习算法。我们会站在这些巨人的肩膀上，但对这样的声明持保守态度。这样的声明会引起相关问题：出现的一种和终极算法相似的算法，只是改变了参数，除了数据还有最小量输入信息，可以像人类一样很好地看懂视频和文本，并在生物学、人类学以及其他学科中有重大发现，如果是这样，我们怎么知道已经找到了终极算法？很明显，如果按照这个标准，即使终极算法真的已经存在，它也没法证明自己就是终极算法。

关键是，终极算法不需要在遇到每个新问题时，都从零开始。这个标准对所有算法来说都太高了，而且它也不像人类所做的那样。例如，语言无法存在于真空中；如果没有该学科的相关知识，我们就无法理解一门学科。因此，当学习阅读时，终极算法可以依靠之前所学的东西来看、听，以及控制一个机器人。同样的道理，科学家不会只是盲目地将模型和数据进行配对，他们会利用自己在该领域的知识来解决这个问题。因此，当在生物学领域有所发现时，终极算法会首先阅读它需要的生物学知识，依靠的是之前就学会的阅读技巧。终极算法不只是被动地消耗知识，它可以和周围的环境进行互动，然后积极寻找它想要的数据，就像机器人科学家“亚当”一样，或者像所有探索世界的孩子一样。

我们寻找终极算法的过程是复杂且活跃的，因为在机器学习领域存在不同思想的学派，主要学派包括符号学派、联结学派、进化学派、贝叶斯学派、类推学派。每个学派都有其核心理念以及其关注的特定问题。在综合几个学派理念的基础上，每个学派都已经找到该问题的解决方法，而且有体现本学派的主算法。

对于符号学派来说，所有的信息都可以简化为操作符号，就像数学家那样，为了解方程，会用其他表达式来代替本来的表达式。符号学者明白你不能从零开始学习：除了数据，你还需要一些原始的知识。他们已经弄明白，如何把先前存在的知识并入学习中，如何结合动态的知识来解决新问题。他们的主算法是逆向演绎，逆向演绎致力于弄明白，为了使演绎进展顺利，哪些知识被省略了，然后弄明白是什么让主算法变得越来越综合。

对于联结学派来说，学习就是大脑所做的事情，因此我们要做的就是对大脑进行逆向演绎。大脑通过调整神经元之间连接的强度来进行学习，关键问题是找到哪些连接导致了误差，以及如何纠正这些误差。联结学派的主算法是反向传播学习算法，该算法将系统的输出与

想要的结果相比较，然后连续一层一层地改变神经元之间的连接，目的是为了使输出的东西接近想要的东西。

进化学派认为，所有形式的学习都源于自然选择。如果自然选择造就我们，那么它就可以造就一切，我们要做的，就是在计算机上对它进行模仿。进化主义解决的关键问题是学习结构：不只是像反向传播那样调整参数，它还创造大脑，用来对参数进行微调。进化学派的主算法是基因编程，和自然使有机体交配和进化那样，基因编程也对计算机程序进行配对和提升。

贝叶斯学派最关注的问题是不确定性。所有掌握的知识都有不确定性，而且学习知识的过程也是一种不确定的推理形式。那么问题就变成，在不破坏信息的情况下，如何处理嘈杂、不完整甚至自相矛盾的信息。解决的办法就是运用概率推理，而主算法就是贝叶斯定理及其衍生定理。贝叶斯定理告诉我们，如何将新的证据并入我们的信仰中，而概率推理算法尽可能有效地做到这一点。

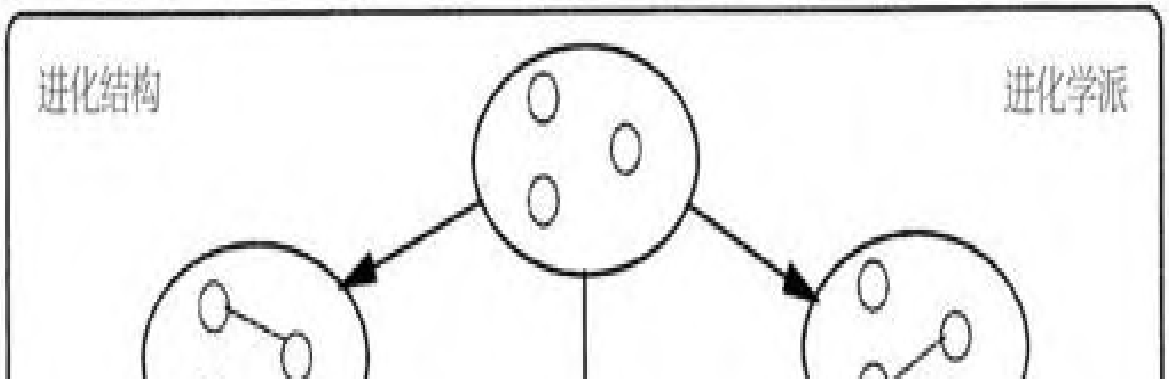
对于类推学派来说，学习的关键就是要在不同场景中认识到相似性，然后由此推导出其他相似性。如果两个病人有相似的症状，那么也许他们患有相同的疾病。问题的关键是，如何判断两个事物的相似程度。类推学派的主算法是支持向量机，主算法找出要记忆的经历，以及弄明白如何将这些经历结合起来，用来做新的预测。

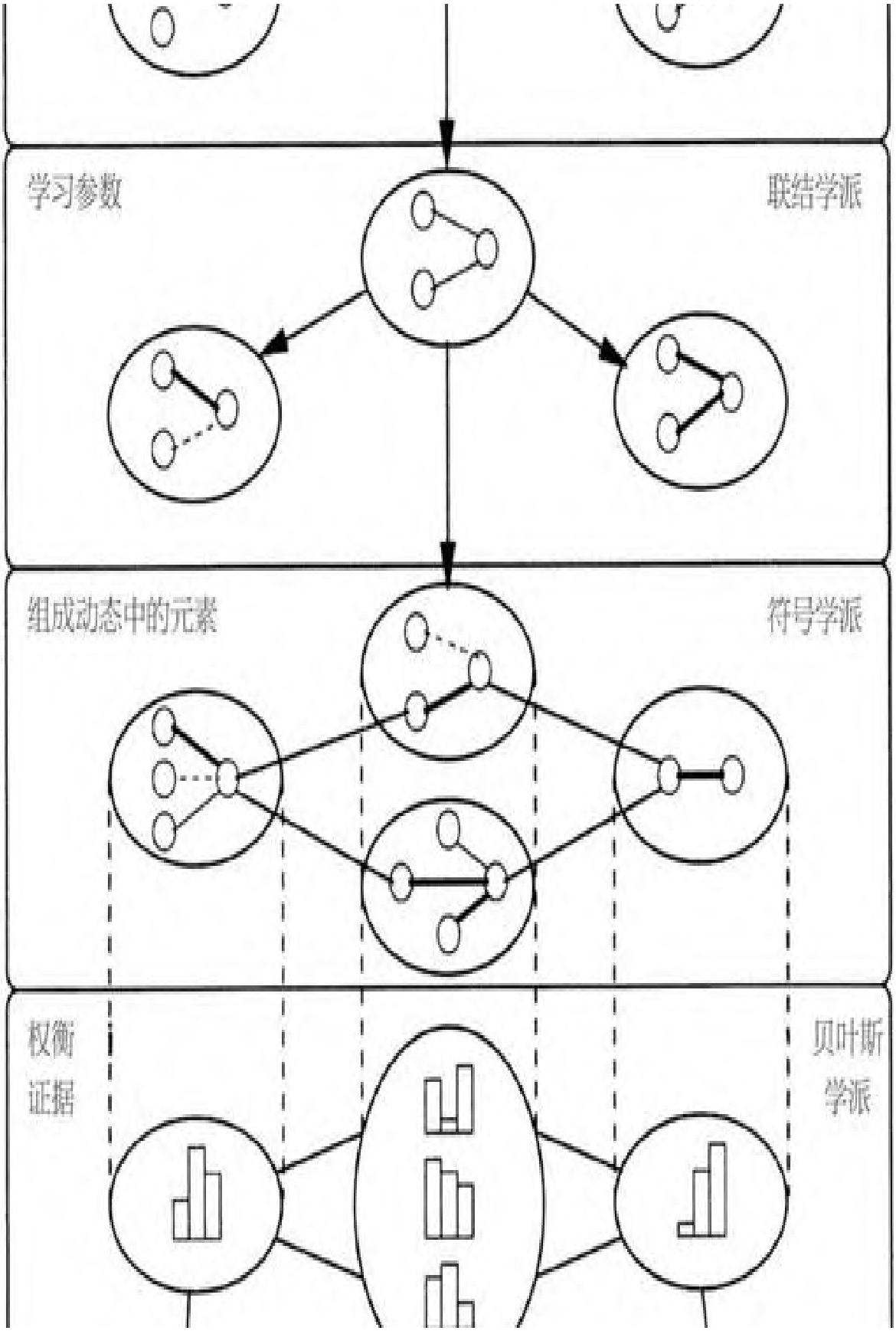
每个学派对其中心问题的解决方法都是一个辉煌、来之不易的进步，但真正的终极算法应该把5个学派的5个问题都解决，而不是只解决一个。例如，为了治愈癌症，我们要了解细胞的代谢网络：哪些基因调节哪些别的基因，由此产生的蛋白质控制哪些化学反应，以及将新微粒加入混合物中将会对网络产生什么影响。从零开始努力学习这些东西显得有点愚蠢，因为这种做法忽略了过去几十年生物学家苦心积累的知识。符号学派懂得如何将这些知识与来自DNA测序仪、基因表达芯片等的数据结合起来，并得出结果。只有知识或数据，你得不出这些结果，可是我们通过逆向演绎得到的知识都是纯定性的。我们要

了解的不仅是谁和谁交互，还有可以交互的程度，以及反向传播如何做到这些。即便如此，如果没有某个基础结构，逆向演绎和反向传播将会迷失在太空中。有了这个基础结构，它们找到的交互和参数才能构成整体。基因编程可以找到这个基础结构。这时，有了新陈代谢的完整知识，以及给定病人的相关数据，我们就可以为他找到治疗方法。但实际上，我们拥有的知识总是非常不完整的，甚至在有些地方会出错。即使如此，我们还是要继续进行，这也就是概率推理的目标。在情形最困难的例子中，病人的癌症看起来与之前的癌症病例有很大不同，而我们掌握的知识对此也束手无策。基于相似性的算法会扭转大局，方法就是从看似有很大差别的情形中找到相似点，把重点放在相似点上，然后忽略其他不同点。

本书将综合出一个拥有所有这些功能的终极算法：

我们对终极算法的追寻之旅将让我们了解这5个学派。学派与学派相遇、谈判、冲突的地方，也是这个旅程最艰难的部分。每个学派都有自己不同的观点，我们必须将这些观点集中起来。机器学习算法和所有科学家一样，类似盲人和大象：有个盲人摸到象鼻，就以为那是蛇；另一个盲人靠着象腿，以为那是树；还有一个盲人摸到象牙，以为那是公牛。我们的目标是，摸清楚每个部位，而不是过早下结论。一旦摸到所有部位，我们就努力拼出整个大象的形象。将所有信息集中起来变成解决方案的方法，并不是很容易找到，有些人甚至说不可能找到，但这就是我们要做的事情。




学习参数
联结学派
组成动态中的元素
符号学派
权衡
证据
贝叶斯
学派

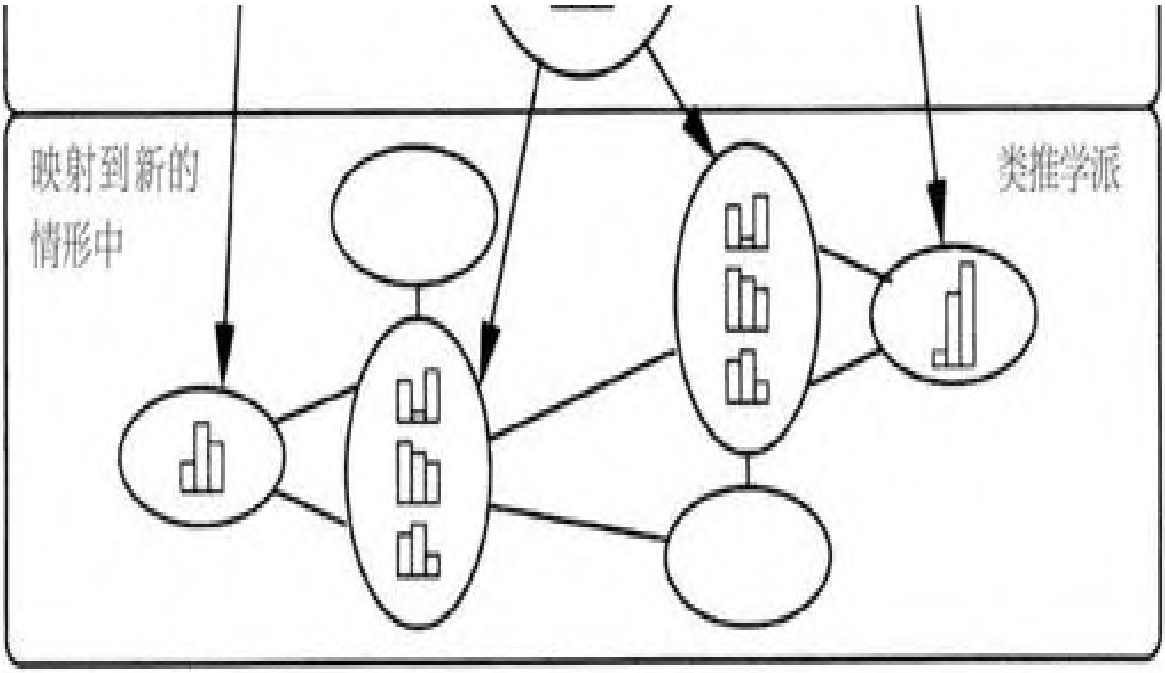


图2－1

我们要找到的算法还不是终极算法，原因我们在下文会提到，但这已经是所有人能达到的、最接近终极算法的水平了。我们会一直积聚财富，让大富豪们嫉妒。即便如此，本书仅仅是终极算法传奇的第一部分。第二部分的主角就是你——亲爱的读者。你的使命（如果你愿意接受）就是完成寻找终极算法之旅，然后把战利品拿回来。在第一部分我将是你卑微的导游，从这里出发，走向已知的尽头。你可能抗议说自己懂的不够多，或者算法不是你的长处。不要害怕！计算机科学还年轻，而且不像物理学或生物学那样，要发起一场革命，你并不需要博士学位（可以去问问比尔·盖茨、谢尔盖·布林、拉里·佩奇、马克·扎克伯克）。洞察力和坚持才是最重要的东西。

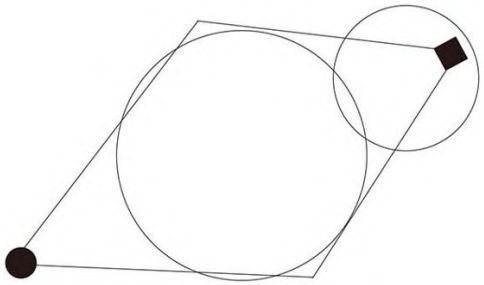
你准备好了吗？我们的旅程由拜访符号学派开始，这个学派的历史最漫长。

---

[4] NP完全问题（non－determinstic polynomial completeness），即多项式复杂程度的非确定性问题。——编者注

第三章

符号学派：休谟的归纳问题



你是理性主义者还是经验主义者？

理性主义者认为，感官会欺骗人，而逻辑推理是通往知识的唯一可靠的道路。经验主义者认为所有推理都是不可靠的，知识必须来源于观察及实验。法国人是理性主义者，盎格鲁－撒克逊人（法国人就这样称呼他们）是经验主义者。评论员、律师、数学家是理性主义者，记者、医生、科学家是经验主义者。《女作家与谋杀案》是关于犯罪的电视剧，属于理性主义，《犯罪现场调查》则属于经验主义。在计算机科学领域，理论家和知识工程师属于理性主义者，黑客和计算机学习者属于经验主义者。

理性主义者喜欢在迈出第一步前，就提前规划好一切。经验主义者喜欢尝试新事物，然后看看结果会怎样。我不知道理性主义和经验主义是否都有专门的基因，但我看了看自己的计算机科学家同事，经过反复观察，发现理性主义者和经验主义者的个性特点几乎一样：有些人以理性主义为核心，而且绝不会改变；其他人则是彻底的经验主

义者，未来也不会改变。这两方可以彼此进行对话，而且有时候还可以利用彼此的定论，但他们对彼此了解仅限这么多。实际上，每方都会认为对方做什么并不重要，而且也没什么意思。

自从智人的黎明到来，理性主义和经验主义也许就已经出现。在去狩猎之前，原始人鲍勃会花很长时间坐在洞里，思考要去哪里打猎。同时，穴居女人爱丽丝正在外面系统地调查领土。这两种方法我们都还在使用，保守地说这两种方法没有哪个更好。你也许会认为，机器学习就是经验主义最终胜利的产物，但我们很快会看到，真理总是比我们想的更加微妙。

理性主义与经验主义是哲学家最热衷讨论的问题。柏拉图是早期的理性主义者，而亚里士多德是早期的经验主义者。关于这个问题的辩论，真正开始于启蒙运动时期，每方有三位伟大的思想家：笛卡儿、斯宾诺莎、莱布尼茨是理性主义的代表，洛克、贝克莱、休谟则是经验主义的代表。因为相信自己的推理能力，理性主义者编造出宇宙理论（委婉地说），这经不住时间的考验，但他们也创造了基本的数学知识，比如微积分和解析几何。经验主义总体来说更为实际，而且它们随处可见，从科学方法到美国宪法都有它们的身影。

大卫·休谟是最伟大的经验主义者，以及有史以来最伟大、以英语为母语的哲学家。诸如亚当·斯密、查尔斯·达尔文之类的思想家都深受其影响。你也可以说他是符号学派的守护神。他1711年生于苏格兰，大半辈子都生活在18世纪的爱丁堡（一个思想活跃的繁荣城市）。他虽然性情温和，却是一个严格的怀疑论者，花了大半辈子来解开他那个时代的难题。为了得出符合逻辑的论断，休谟借用了洛克开创的经验主义思想，并提出一个问题，这个问题在所有领域的知识中就像一把达摩克利斯之剑，从最琐碎的领域到最先进的领域，都是一个时刻存在的问题：在概括我们见过的东西以及没见过的东西时，怎样才能做到合理？从某种意义上说，每种学习算法都在尝试回答这个问题。

休谟的问题也正是我们开启寻找终极算法之旅的开始。首先，我们会通过日常生活中的例子来解释这个问题，并通过现在人人皆知的“天下没有免费的午餐”这个定理来体现这个问题。然后我们会看到符号学者对休谟的回答。这把我们引向机器学习中最重要的问题：不真实存在的过拟合或幻觉模式。我们会看到符号学者如何解决它。机器学习本质上是一种炼金术，在魔法石的辅助下把数据变成知识。对于符号学者来说，魔法石就是知识。在接下来的4章中，我们还会研究其他学派的“炼金术”。

约不约

你有一个很喜欢的朋友，想找她出来约会。如果遭到拒绝，你会很难过，但是如果你想知道她会不会答应，也只能去问她。现在是周五傍晚，你拿着手机坐在那里考虑要不要给她打电话。你记得上次你问过她，但她拒绝了。但是为什么上次的前两次你约她，她都答应了，而这两次的前一次她却拒绝了？可能有时候只是她不想出门？也可能她喜欢泡吧而不喜欢吃饭？为了理出头绪，你先放下了手机，写下了前几次的情况（见表3－1）。

表3-1 约会记录

约会次序	星期几	约会方式	天气	晚间电视节目	是否答应约会
1	周末	吃饭	暖	不好看	否
2	周末	泡吧	暖	不好看	是
3	周末	泡吧	暖	不好看	是
4	周末	泡吧	冷	好看	否
5	周末	泡吧	冷	不好看	?

那么答案是什么？答应还是不答应？有没有什么模型可以区分肯定和否定？更重要的是，这个模型会如何决定今天的情况？

很明显，没有哪个因素可以单独准确预测出答案：有的周末她喜欢出去约会，而有的周末她却不想出去；有时她喜欢去泡吧，而有时她又不喜欢……把所有因素综合起来会怎样？可能周末她喜欢去泡吧？并不是，第四次约会就排除了这种情况，或者可能她只喜欢天气暖和的周末晚上出去约会？是的！这个说法符合之前的情况！看了看外面寒冷的天气，好像今晚不合适。可是等一下！电视不好看的晚上她会不会想去泡吧？好像这也符合前几次的情况，也就是说今天可以约会！快，快点打给她，不然就迟了。再等等。你怎么知道这是对的模型？你已经找到两种情况和之前的情况相符，但这两次的预测都是否定。细想一下，如果她只想天气好的时候去泡吧，那怎么办？或者她只是在没什么电视好看时的周末才会出去？或者……

这时，你灰心地把画好的表揉成一团，丢到垃圾桶里。没有办法知道她会不会答应！你能做什么？休谟的灵魂伤心地在你的肩膀上点

头。你没有任何依据选择这个而不是另外一个。对于“她会说什么”这个问题，回答“是”或“否”的可能性都一样。时钟嘀嗒作响，最后你准备抛硬币来决定。

你不是唯一身处窘境的人——我们都是。我们才开始寻找终极算法之旅，似乎就已经遇到无法克服的困难了。有没有什么办法可以从过去的经历中掌握规律，然后信心满满地运用到未来的事情中？如果没有，那么机器学习不就是一个没有希望的事业了吗？就此而言，所有学科甚至所有人类的知识，不就随时可能会被推翻吗？

这种情况不是大数据能解决的问题。你可以像卡萨诺瓦那样放荡，有无数个可以约会数几千次的女人，但你的主数据库还是不知道这次这个女人会说什么。即使今天和前几次她答应约会的情况一样（都是周末、一样的约会方式、一样的天气、一样的电视节目），这也并不意味着这次她会答应你出去约会。正如你所知道的那样，她的回答由一些你不知道或者无法知道的原因决定。也或者她的回答没有什么别的原因，只是随口回答，而你也只是白费力气地从之前的情况中找规律。

自休谟提出归纳问题，哲学家就已经对此进行辩论，但还没有人能给出一个满意的答案。伯特兰·罗素喜欢用“归纳主义者火鸡”这个故事来阐述这个问题。故事的主人公是一只火鸡，它来到农场的第一个早晨，主人在早上9点喂它们，但作为实实在在的归纳法优越论者，它不想过早下结论：主人每天都9点喂它们。首先它在不同情况下观察了很多天，收集了许多观察数据。主人连续多天都是9点喂它们，最终它得出结论，认为主人每天早上9点喂火鸡，那么主人一直会在早上9点给它喂食。接下来是平安夜那天的早晨，主人没有喂它，因为它被宰了。

如果休谟问题仅仅是一个我们可以忽略的哲学小难题，就太好了，但事实并非如此。例如，谷歌的业务就是在你往搜索框输入一些关键字时，猜测你在寻找哪些网页。过去搜索查询的人所输入关键字

的大量记录，以及他们点击的、相应查询结果页面的链接，都是谷歌的重要资产。如果某个人输入一组关键字，而这些关键字却不在记录里面，你该怎么办？即便关键字在记录中，你怎么能肯定当前用户想要的搜索结果和之前的一样？

如果我们只是假设未来和昨天一样，那么会怎样呢？这当然是一个有风险的假设（这对归纳主义者火鸡就不会奏效）。另外，没有这样的假设，所有知识将不复存在，生活也是如此。虽然有很多不确定性，但我们还是宁可活下来。遗憾的是，虽然有那样的假设，但我们还是尚未走出困境。这个假设还得应付这些“微不足道的”例子：如果我是一名医生，患者B和患者A有一模一样的症状，我假设两人的诊断结果都一样。但如果患者B的症状和其他人都不一样，我就仍然不知道如何做诊断。这属于机器学习问题：将某结果推广到我们没有见过的事件中。

也许这也不是什么大问题？有了足够的数据，大多数事件不就变得“微不足道”了吗？不是的。我们在前面的章节中了解到，为什么记忆不能当作通用学习算法，我们现在可以用更量化的方式来解释这个问题。假设你现在有一个数据库，含有1万亿条记录，每条记录有1000个布尔字段（也就是说，每个字段回答一个是或否的问题）。这个数据库真的好大。你认为有多少种可能的事件？往下读之前好好猜一猜。每个问题可能的答案有两个，两个问题就是2乘以2（是—是、是—否、否—是、否—否），三个问题就是2的三次方（$2\times2\times2=2^3$），如果是1000个问题，就是2的1000次方（2^{1000}）。你的数据库中1万亿条记录，可能性就是分子是1、分母是无限大的数乘以1%，然后再乘以2^{1000}，“分子是1、分母是无限大的数”指的是“小数点前是0，小数点后是286个0，后面跟着1的小数”。结果就是：无论有多少数据——多少兆、多少千兆、多少千兆兆、多少泽或多少尧字节，你基本上什么也看不到。你要决定的新事件已经存在于

数据库中，而数据库非常大，这件事发生的概率低到可以忽略，所以如果不进行一般化，对你就不会有任何帮助。

如果这些听起来有点抽象，那么假设你是一个邮件服务提供商，要将每封收到的邮件进行分类，分为垃圾邮件或非垃圾邮件。你也许有一个包含1万亿封过去邮件的数据库，而且每封邮件都已经被分为垃圾邮件或非垃圾邮件，但那样做并不会让你省事，因为每封新邮件和之前邮件一模一样的概率几乎是0。你没有选择，只能以大概的概率来区分垃圾邮件和非垃圾邮件，而根据休谟的观点，这根本做不到。

“天下没有免费的午餐”定理

休谟提出爆炸性问题之后的250年，大卫·沃尔珀特（David Wolpert）赋予了这个问题优雅的数学形式。沃尔珀特原来是一名物理学家，后来成为机器学习者。他的研究结果被人们称为“天下没有免费的午餐”定理，规定“怎样才算是好的学习算法”。这个规定要求很低：没有哪个学习算法可以比得上随意猜测。好吧，那我们就不用找终极算法，可以回家了：终极算法只是用抛硬币来做决定的算法。说真的，如果没有哪个学习算法可以比抛硬币更管用，那会怎样？而且如果真的是那样，为什么这个世界——从垃圾邮件过滤（随时都在进行）到自动驾驶汽车——到处都是非常成功的学习算法呢？

“天下没有免费的午餐”这个定理和帕斯卡尔赌注失败的原因非常相似。帕斯卡尔的《思想录》于1669年出版，他在书中提到我们应该相信基督教中的神，因为如果神存在，他就会给我们永生，而如果他不存在，我们的损失也很小。这在当时是非常复杂的论点，但正如狄德罗指出的那样，伊玛目也可以同样的理由让人们来相信真主安拉的存在。如果你选错了要信的神，代价就是永世在地狱。总而言之，

在考虑各种各样可能相信的神时，选择特定的神来信仰，还不如选择其他的神。因为，有说“这样做”的神，也会有说“不，那样做”的神。你也许应该把神忘了，好好享受没有宗教限制的生活。

用“学习算法”来代替“神”，用“准确的预测”来代替“永生”，你就遵守“天下没有免费的午餐”这个定理了。选择你最喜欢的学习算法（在本书中你会看到很多算法）。如果存在学习算法比随机猜测好用的领域，我（一个喜欢唱反调的人）会构建一个学习算法没有随机猜测好用的领域。我要做的就是把所有“未知”例子的标签翻过来。因为“经过观察”的标签表明，你的学习算法绝无可能区分世界和反物质世界。在这两个世界中的平均表现，学习算法和随机猜测一样好用。因此，在所有可能的世界中，把每个世界与其反物质世界配对，你的学习算法的作用就和抛硬币的作用一样。

虽然如此，别马上就对机器学习或终极算法失望。我们不关心所有可能存在的世界，而只关心我们生存的这个世界。如果我们对这个世界有所了解，然后把了解的知识输入我们的学习算法，那么现在和随机猜测相比，学习算法就可以发挥优势了。休谟可能会回应说，知识本身必须由归纳得来，因此知识也是有问题的。没错，虽然知识是通过进化编入我们的大脑的，但我们不得不冒这个险。我们也可以这样问，有没有小部分毋庸置疑、非常基础的宝贵知识，让我们可以在其基础上进行所有归纳（有点像笛卡儿的“我思故我在”，虽然很难明白，如何将这句话输入学习算法中）。我觉得回答是肯定的，在第九章，我们就会知道那些宝贵的知识是什么。

同时，“天下没有免费的午餐”这个实际的结论表明，不靠知识进行学习，这样的事不存在。只有数字也不够。从零开始只会让你一无所获。机器学习就像知识泵，我们可以用它来从数据中提取大量的知识，但首先我们得先对泵进行预设。

数学家认为机器学习这个问题是一个不适定问题（ill-posed problem）：这个问题没有唯一解。下面是一个简单的不适定问题：哪

两个数相加的得数是1000？假设这两个数都是正数，答案就有500种……1和999，2和998等等。解决不适定问题的唯一办法就是引入附加假设。如果我告诉你，第二个数是第一个数的三倍，那么答案就是250和750。

汤姆·米切尔（Tom Mitchell）是典型的符号学者，称机器学习体现“无偏见学习的无用性”。在日常生活中，“偏见”是一个贬义词：预设观念不太好。但在机器学习中，预设观念是必不可少的；没有这些观念，你就无法进行学习。实际上，预设观念对人类认知来说，也是必不可少的，这些观念是“直线布入”人脑的。对于它们，我们也觉得是理所当然的。超出那些观念的偏见才值得质疑。

亚里士多德曾经说过，在知识领域，没有什么东西不是首先凭借感觉来形成的。莱布尼茨又加了一句，“除了知识本身”。人类的大脑不是一张白纸，因为它不是一块石板。石板是被动的，你可以在上面写东西，但大脑可以主动处理它接收到的东西。记忆就是大脑用来写东西的石板，而且记忆不是一开始就是空白的。另一方面，计算机在你给它编程之前，就是一张白纸；在用计算机做事之前，这个积极的过程需写入记忆。我们的目标是找到最简单的、我们能编写的程序，这样写好的程序就可以无限制地通过阅读数据来自行编程，直到该程序掌握所有能掌握的知识。

机器学习不可避免地含有投机的因素。在《警探哈里》（第一集）中，克林特·伊斯特伍德追逐一名银行抢劫犯，同时不断向他开枪。最后，抢劫犯躺在一把装有子弹的枪旁边，不确定要不要把它拿起来。哈里开了6枪还是只开了5枪？哈里同情地说（可以这么说）：“你得问自己一个问题：‘我足够走运吗？’你真的走运吗，小子？”这也是机器学习算法每天运作时必须问自己的问题：“今天我幸运吗？”就像进化的过程一样，机器学习不是时时刻刻都知道自己是否可以准确无误地运行。实际上，误差是常有的事，并不意外。但没关系，因为我们放弃误差的部分，主要靠没有误差的部分，而计算

结果才是最重要的。我们一旦掌握新的知识，基于前面的步骤，就可以得出更多的知识。唯一的问题就是，从哪里开始。

对知识泵进行预设

在《自然哲学的数学原理》一书及三大运动定律中，牛顿阐述了推理的四条法则。虽然这些法则没有那些物理定律那么著名，但可以说很重要。其中第三条是关键法则，我们可以这样表述：

我们见过的所有真实的东西，在宇宙中也是真实的。

可以毫不夸张地说，这句听起来无伤大雅的话就是牛顿革命以及现代科学的核心。开普勒定律适用于6个实体：那个年代太阳系中已知的6颗行星。牛顿定律适用于宇宙中的每一个微粒。这两个定律之间的共性如此之大，让人感到吃惊，而这也是牛顿法则的直接结果。这个法则本身就是拥有非凡动力的知识泵。没有这个法则，也就没有什么自然法则，有的也只是永远无法完整的、小规律的集合体。

牛顿法则是机器学习的第一个不成文规则。我们归纳自己能力范围内、应用最广泛的规则，只有在数据的迫使下，才缩小规则的应用范围。乍一看，这看起来可能过于自信甚至近乎荒谬，但这种做法已经为科学服务了300余年。当然也可以想象出一个变化无常的宇宙，在那里牛顿法则不起作用，但那并不是我们的宇宙。

然而，牛顿法则仅仅是第一步。还得弄明白我们见到的哪些是真实的——如何从原始数据中找出规律。标准的解决方法就是假设我们知道真理的形式，而算法的任务就是把这个形式具体化。例如，在之前提到的约会问题中，你可以假设你朋友的回复由单个因素来决定。在这种情况下，算法就只包括看看每个已知的因素（时间、约会方

式、天气、晚间电视节目），确定该因素是否每次都能准确预测她的回答。可问题就在于，每个因素都无法预测她的回答！你打赌了，然后输了。所以你把假设放宽了一点。如果你朋友的回答是由两个因素一起决定的呢？总共四个因素，每个因素有两种可能的值，那么总共有24种可能（总共有6对因素组合，即12乘以因素的两种可能）。数字太多，我们遇到尴尬：两个因素的四种组合准确预测了结果！接下来怎么办？如果你觉得运气还行，可以选其中的一种，然后祈祷最好的结果。但更明智的选择是采取民主的做法：对每个选项进行选择，然后选最后赢的预测。

如果所有两个因素组合的预测都失败了，你可以尝试任意个数因素的组合，机器学习者和心理学家称之为“合取概念”（conjunctive concept）。字典对词的定义就属于合取概念：椅子是有靠背、若干条腿的坐具。把任意一个描述去掉，就不再是椅子。托尔斯泰在写《安娜·卡列尼娜》的开篇时，出现在他脑海里的就是合取概念：“所有幸福的家庭都是相似的，每个不幸的家庭各有各的不幸。”对于个人来说，也是这样的。为了感到幸福，你需要健康、爱、朋友、钱、你喜欢的工作等。把这些东西的任意一个拿走，痛苦也会随之而来。

在机器学习中，概念性的例子成为正面例子，而与概念例子相反的则是负面例子。如果你在尝试通过图片来认出猫，那么猫的图片就是正面例子，而狗的图片则是负面例子。如果你对世界文学中描述的家庭进行汇总，并编成数据库，那么卡列尼娜一家就是幸福家庭的负面例子，而且正面例子寥寥无几。

首先做有条件的假设，如果这样无法解释数据，再放松假设的条件，这就是典型的机器学习。这个过程通常由算法自行进行，不需要你的帮助。首先，算法会尝试所有单一因素，然后尝试所有两个因素的组合，之后就是所有三个因素的组合等。但现在我们遇到一个问题：合取概念太多，没有足够的时间对其逐个尝试。

约会的例子有点欺骗性，因为它太小（4个变量，4个例子）。但假设你提供在线约会服务，你就需要知道要对哪些人进行配对。如果你的每个会员都填写了一份问卷，问卷包含50个“是或否”的问题，这样就有100种特点，这100种特点涵盖了每对可能配对成功的情侣的特点，每对情侣中的一方都有50个特点。这些情侣出去约会之后，会汇报结果，在此基础上，你能找到“佳偶”这个定义的合取概念吗？总共有3100种可能的定义（每个问题有三种选择，分别为“是”、“否”、“与该品质无关”）。即使由世界上最快速的计算机来做这项工作，这些情侣也会老得去世了（你的公司也破产了）。等你计算出来，除非你走运，可以找出很短的一条关于“佳偶”的定义。规则太多，而时间太少，我们得做点更精明的事。

这里有一种方法：暂且假设每个配对都合适，然后排除所有不含有某品质的搭配，对每种品质重复同样的做法，然后选择那个排除了最多不当搭配和最少适当搭配的选项。现在你的定义看起来就像“只有他开朗，这对才合适”。现在反过来试着把其他品质加进去，然后选择那个排除了剩下最多的不当搭配和剩下最少的适当搭配的选项。现在的定义可能是“只有他和她都开朗，这对才合适”。然后试着往那两个特点里加入第三个品质，以此类推。一旦排除了所有不合适的搭配，你就成功了：就有了这个概念的定义，这个概念排除了所有的正面例子和所有的负面例子。例如，“每对中的两个人都开朗，这对才合适，他爱狗，而她不爱猫”。现在你可以丢掉所有数据，然后只把这个定义留下，因为这个定义概括了所有和你的目标相关的东西。这个算法保证能在合理的时间内完成运算，而这也是我们在本书中见过的第一个真实的学习算法。

如何征服世界

虽然合取概念的用途有很多，但并不能让你走很远。正如鲁德亚德·吉卜林说的那样，问题在于“编部落歌谣的方法有很多种，而每种方法都是正确的”。真正的概念是分离的。椅子可能有四条腿或三条腿，而有些则一条也没有。你可以以无数种方法来赢一盘棋。包含“伟哥”这个词的邮件有可能是垃圾邮件，但包含“免费”一词的邮件也有可能是垃圾邮件。此外，所有规则都会有例外。所有的鸟都会飞，除了企鹅、鸵鸟、食火鸡或者几维鸟（或断了翅膀的鸟，或被锁在笼子里的鸟）。

我们要做的就是学习经过一系列规则定义的概念，而不仅仅是单个规则，例如：

如果你喜欢《星球大战》第四至六部，那么你会喜欢《阿凡达》。

如果你喜欢《星际迷航：下一代》以及《泰坦尼克号》，那么你会喜欢《阿凡达》。

如果你参加塞拉俱乐部，并阅读科幻书籍，那么你会喜欢《阿凡达》。

或者：

如果你的信用卡昨天在中国、加拿大以及尼日利亚被刷，那么它被盗了。

如果某工作日晚上11点，你的信用卡被刷，那么它被盗了。

如果你的信用卡被用来购买一美元的汽油，那么它被盗了。

如果你对最后一条规则有疑问，解释如下：过去信用卡窃贼通常会用偷到的卡购买一美元的汽油，来看看在数据挖掘器识破其阴谋

前，信用卡是否完好。

我们可以像学习这条规则那样来同时学习多套规则，利用我们之前见过的算法来学习合取概念。我们学习每个规则之后，会排除该规则包含的正面例子，因此下一个规则会尽可能多地包含剩下的正面例子，以此类推，直到所有的例子都被包含在内。这是一个“分而治之”的例子，也是科学家的战术手册中最古老的策略。为了找到单个规则，我们也可以对算法进行改良，方法就是保留某数n的假设，不止一个数，然后在每个步骤中将这些数以所有可能的方法延伸开来，最后保留n的最佳结果。

通过该方法发现规则的创意来自理夏德·米哈尔斯基（Ryszard Michalski），他是波兰的一位计算机科学家。米哈尔斯基的故乡——卡尔鲁兹之前曾属于波兰、俄罗斯、德国以及乌克兰，这让米哈尔斯基比多数人更能理解合取概念。1970年移民美国之后，他和汤姆·米切尔、杰米·卡博内尔一起创立了机器学习的符号学派。他个性傲慢，如果你在一场机器学习会议中做报告，那么很有可能他会举手指出你只是重新发现了他之前的旧观点。

零售商喜欢一套套的规则，因为他们要决定该囤什么货。通常，他们会用比“分而治之”更为彻底的方法，也就是寻找所有能够准确预测每个购买项的规则。沃尔玛在该领域属先驱，他们早期的发现之一就是，如果你买了纸尿片，那么很有可能会买啤酒。为什么？对此进行解释的说法之一就是，妈妈让爸爸去超市买纸尿片，出于情感补偿，爸爸买了一箱啤酒。知道这一点，超市现在会把啤酒放在纸尿片旁边，这样啤酒就会卖得更好。不找规律，这样的事就不会在沃尔玛发生。“啤酒和纸尿片”的规则已经在数据挖掘领域取得传奇式的地位（虽然会有人说，所谓的传奇和城市多样化有关）。不管怎样，这和米哈尔斯基想象的数字电路设计问题还差得很远。20世纪60年代，他开始考虑规则归纳问题。如果你发明了新的学习算法，你甚至无法想象这个算法能应用的所有地方。

我第一次直接体验规则学习，是在申请信用卡的时候。那时我刚来到美国，开始接受研究生教育，银行给我寄了一封信，说“很遗憾，您居住在该地址的时间还不足够长，所以无历史信用记录，你的申请被拒绝了”。就在那时，我知道在机器学习领域有待研究的东西还很多。

在无知与幻觉之间

规则集在很大程度上比合取概念要有力得多。实际上，规则的力量如此之大，大到你可以用规则来代表任何概念，要找到原因则很难。如果你给我某个概念的完整例子，我只能将每个例子变成一个规则。这个规则规定了每个例子的所有属性，而这些规则的集合就是该概念的定义。回到之前关于约会的例子，其中的一个规则是：现在是周末晚上，天气暖和，没有好看的电视节目，你提议去泡吧，她会说“没问题”。表3-1只包含了几个例子，但其实它有16种（2×2×2×2）可能，每个可能都会有“约”或者“不约”的结果，将每个正面例子以这样的方法变成规则，我们就成功了。

规则集的力量是一把双刃剑。从正面看，你知道自己总能找到和数据完美匹配的规则。但你还没来得及开始觉得走运，就意识到自己很有可能会找到一个毫无意义的规则。记住“天下没有免费的午餐”：没有知识，你就无法进行学习。假设某概念能通过规则集来定义，相当于什么也没有假设。

无用规则集的一个例子就是，只包含了你看到的正面例子，除此之外，没有其他的例子。这个规则集看起来100%准确，但那只是假象：它会预测每个新例子都是负面例子，然后把每个正面例子弄错。如果正面例子总体上比负面例子多，这种做法的效果会比抛硬币更糟

糕。想象一下，邮件过滤器仅因为某封邮件和之前垃圾邮件一模一样就将其过滤，这样会有什么后果？对分类好的数据进行学习很容易，这样的数据看起来也很棒，但还不如没有垃圾邮件过滤器。很遗憾，我们“分而治之”的算法有可能就像那样对规则集进行简单的学习。

在《博闻强识的富内斯》的故事中，豪尔赫·路易斯·博尔赫斯讲述了和一位拥有完美记忆力的少年相遇的故事。拥有完美记忆力看起来非常幸运，但其实这是一个可怕的诅咒。富内斯能记住过去任意时刻天空中云朵的形状，但他没有办法理解，下午15:14看到的狗的侧面和下午15:15看到的狗的正面，都是同一条狗。每次他在镜子中看到自己的脸，都会感到惊讶不已。富内斯无法进行概括：对他来说，两个事物，只有每个细节看起来都一致，才能说它们是一样的。自由的规则算法和富内斯一样，发挥不了作用。学习就意味着将细节遗忘，只记住重要部分。计算机就是最大的白痴专家：它们可以毫无差错地将所有东西记住，但那不是我们想让它们做的。

问题不限于记忆大量例子。每当算法在数据中找到现实世界中不存在的模型时，我们说它与数据过于拟合。过拟合问题是机器学习中的中心问题。在所有主题中，关于过拟合问题的论文最多。每个强大的学习算法，无论是符号学算法、联结学算法，或者其他别的学习算法，都不得不担忧幻觉模式这个问题。避免幻觉模式唯一安全的方法，就是严格限制算法学习的内容，例如要求学习内容是一个简短的合取概念。很遗憾，这种做法就像把孩子和洗澡水一起倒掉一样，会让学习算法无法看到多数真实的模型，这些模型在数据中是可见的。因此，好的学习算法永远在无知与幻觉的夹缝中行走。

人类对过拟合也没有免疫。你甚至可以说，过拟合是我们的万恶之源。想想一个白人小女孩，在商场看到拉美裔婴儿时脱口而出“看，妈妈，那是小女佣”（真实例子）。小女孩并非生来就是偏执狂。那是因为在她短暂的人生阅历里，她对见过的仅仅几个拉美裔女佣进行了笼统的概括。这个世界有许多从事其他职业的拉美裔，但她

还没有见过他们。我们的信仰建立在自己的经历之上，这会让我们对世界的理解不完整，而且也容易过早得出错误的结论。即便你很聪明，学识渊博，也无法免受过拟合的影响。亚里士多德说要使一个物体不断运动，需要对其施加一个力，就犯了过拟合的错误。伽利略的天才之处在于，无须到外太空亲眼见证，他凭直觉就知道，不受外力影响的物体会一直保持运动。

但学习算法特别容易过拟合，因为它们拥有从数据中发现模型、近乎无限制的能力。人类发现一个模型所用的时间，计算机可以找到数百万个。在机器学习中，计算机最大的优势（处理大量数据以及不知疲倦不断重复同样步骤的能力）也是它的劣质所在。如果你做的研究够多，然后能有所发现，也很不错。《圣经密码》（1998年的畅销书）声称如果你以固定间距跳过某些字母，然后把点中的字母拼起来，就会发现《圣经》对未来的预言。遗憾的是，保证你能从任何足够长的文本中找到“预言”，有很多方法。怀疑论者会说，他们在《白鲸记》和最高法院的裁决中找到了预言，除此之外，还提到罗斯威尔，以及《圣经·创世纪》中的不明飞行物。约翰·冯·诺依曼（计算机科学的奠基之父之一）曾说过一句众所周知的话：“用4个参数，我能拟合一头大象；用5个参数，我可以让它的鼻子扭动起来。”当今我们通常会学习拥有数百万参数的模型，这些参数足以让世界上的每头大象都扭动鼻子。甚至曾有人说过，数据挖掘意味着“折磨数据，直到数据妥协”。

过拟合问题因为嘈杂的声音被严重夸大。在机器学习中，这些噪声仅仅意味着数据中的误差，或者你无法预测的偶然事件。比如你的朋友在电视不好看的时候，真的想去泡吧，但你记错第三次约会的情况，并写着那天晚上的电视好看。如果你现在努力找出一套规则，为那天晚上破个例，这样你可能最终会得到一个糟糕的答案，比忽略那个晚上得出的结果还糟。或者比如你的朋友前一晚刚出去喝醉了，所以今晚想休息，但如果是平时，她会答应出去约会。除非你知道她宿醉了，为了得出本例子的正确结果而掌握一套规则，其实会适得其

反：你最好将第三次约会情况“错误分类”，以得出否定回答。这种情况更糟：误差或偶然事件会让你无法找出整套规律。仔细看，你会发现第二次约会和第三次约会其实难以区别：它们都有相同的属性。如果你的朋友答应第二次邀请，而拒绝第三次邀请，那么就没有什么规则能让这两次预测都准确。

当你有过多假设，而没有足够的数据将这些假设区分开来时，过拟合问题就发生了。坏消息是，即便对最简单的合取概念算法来说，假设的数量也会随着属性的增多而呈指数级增长。指数级增长是一件恐怖的事。大肠杆菌每15分钟就能大致分裂成两个细菌，只要有足够的营养，它可以在大约一天时间内，生长出大量细菌，和地球一样大。当算法需要做的事情和输入的数据一样呈指数级增长时，计算机科学家就将这个现象称为组合爆炸，然后会四处奔走寻求保护。在机器学习中，一个概念可能实例的数量，是其属性数量的指数函数：如果属性是布尔值，每种新的属性可能会是实例数量的两倍，方法就是引用之前的每个实例，然后为了那个新属性，对该实例以“是”或“非”来进行扩展。反过来，可能概念的数量是可能实例数量的指数函数：既然每个概念都把实例分成正面或者负面，加入一个实例，可能的概念就会翻倍。因此，概念的数量就是属性数量的指数函数的一个指数函数！换句话说，机器学习就是组合爆炸的组合爆炸。也许我们该放弃，不要把时间浪费在这样没有希望的问题上。

幸运的是，在学习过程中，会发生一些事，把其中一个指数消除，只剩下一个“普通的”单一指数难解型问题。假设你有一个袋子，装满概念的定义，每个定义写在一张纸上，你随机取出纸片，然后看看这个概念和数据的匹配程度。和连续抛1000次硬币都是正面朝上的概率相比，一个不恰当的定义更没有可能让你的数据中所有1000个例子都准确无误。“一把椅子有四条腿，而且是红色的，或者有椅面但没有椅腿”可能会和某些例子匹配，但并不是和你看到的所有椅子都匹配，也可能会和其他一些事物相匹配，但并不是和所有其他事物都匹配。因此，如果一个随机定义准确匹配了1000个例子，那么这

个概念不太有可能是错误的定义，或者至少它和正确的定义非常接近。而且如果一个定义和100万个例子相匹配，那么实际上它就是正确的定义。其他的定义怎么能使那么多例子准确无误？

当然，真正的学习算法不会只是从袋子里随机去除一个定义。这个算法会尝试所有定义，而且这些定义也不是随机选择的。算法尝试的定义越多，越有可能偶然得到能够和所有例子匹配的定义。如果你每组抛1000次硬币，然后重复100万组，实际上至少会有一组出现1000次硬币都是正面朝上的情况，而100万也仅仅是假设的一个小数目。例如，如果例子只有13个属性，那大概就是可能的合取概念的数目（注意你不需要明确地尝试一个又一个概念，如果你找到的最好的那个概念利用了合取概念学习算法，并且和所有的例子匹配，效果也是一样的）。

总结：学习就是你拥有的数据的数量和你所做假设数量之间的较量。更多的数据会呈指数级地减少能够成立的假设数量，但如果一开始就做很多假设，最后你可能还会留下一些无法成立的假设。一般来说，如果学习算法只做了一个指数数量的假设（例如，所有可能的合取概念），那么该数据的指数报酬会将其取消，你毫无影响，只要你有许多例子，且属性不太多。另外，如果算法做了一个双指数的假设（例如，所有可能的规则集），那么数据只会取消其中的一个指数，而且你仍会处于麻烦之中。你甚至可以提前弄明白自己需要多少例子，这是为了保证算法选择的假设和准确的那个非常接近，只要它对所有数据都拟合。换句话说，是为了假设能够尽可能准确。哈佛大学的莱斯利·瓦利安特获得了图灵奖（计算机科学领域的诺贝尔奖），因为他发明了这种分析方法，他在自己的书中将这种方法取名为“可能近似正确”（probably approximately correct），非常恰当。

你能信任的准确度

在实践中，瓦利安特式的分析方法往往非常被动，而且需要的数据比你拥有的还要多。那么你怎么决定，是否要相信学习算法告诉你的东西呢？这个很简单：在利用学习算法过去看不到的数据对其进行证实之前，你不要相信任何东西。如果学习算法假设的模型对新数据来说也适用，你就可以很有信心地说那些模型是正确的，否则你知道学习算法过拟合了。这仅仅是应用于机器学习中的科学方法：对一个新理论来说，这不足以用来解释过去的证据，因为捏造一个能做到这些的理论非常容易。理论还必须做出新的预测，而且只有这些预测经过实验验证后，你才接受它们（即使那样，也只是暂时的，因为未来的证据会对其进行证伪）。

爱因斯坦的广义相对论也只是在亚瑟·爱丁顿对其证实之后才广泛被人们接受。爱丁顿以经验为主，证实广义相对论的预测是对的，预测表明太阳使来自遥远星球的光线变弯了。但是你也不必等待新数据的到来，以决定能否信任学习算法。你可以利用自己拥有的数据，将其分成一个训练集和一个测试集，然后前者交给学习算法，把后者隐藏起来不让学习算法发现，用来验证其准确度。留存数据的准确度就是机器学习中的“黄金标准”。你可以写一篇关于你发明的伟大的新型学习算法的文章，但如果你的算法的留存数据不如之前的算法的留存数据准确，那么这篇文章也没有什么出版价值。

此前不可见数据的准确度测试确实是一个相当严格的考验，实际上，科学的很多方面都因此没有经得住考验。这并不意味着科学就没有用了，因为科学不仅仅是用来预测的，还能用来解释和理解。但是最后，你的模型如果无法对新的数据做出准确预测，就无法保证自己真的理解或者解释了隐藏在背后的现象。对于机器学习来说，对不可见数据的测试是必不可少的，因为这是判断学习算法是否过拟合的唯一方法。

即使测试集准确度也会出问题。据说，在早期军事应用中，一种简单的学习算法探测到坦克的训练集和测试集的准确度都为100%，两

个集都由100张图片组成。惊讶？还是怀疑？结果是这样的——所有有坦克的图片都比没有坦克的图片亮，所以学习算法就都挑了较亮的图片。目前我们有更大的数据集，但并不意味数据收集的质量会更好，所以得当心。在机器学习从新生领域成长为成熟领域的过程中，务实的经验评价会起到重要的作用。追溯至20世纪80年代，每个学派的研究人员很多时候都相信自己华而不实的理论，假装自己的研究范式从根本上说是更好的，所以与其他学派的交流也很少。雷·穆尼、裘德·沙弗里克等符号学者开始系统地对不同算法的相同数据集进行比较，令人惊讶的是，真正的赢家还没有出现。如今竞争在继续，但“异花授粉”的情况也很多。利用加州大学欧文分校的机器学习小组维护的普通实验框架和大型数据集存储库，我们会取得重大进步。而且正如我们看到的，创造通用学习算法的最大希望在于综合不同研究范式的观点。

当然，知道你何时过拟合这一点还不足够，我们需要第一时间避免过拟合。这就意味着不再对数据进行完全拟合，即便我们能做到。有一个方法就是运用统计显著性检验来确保我们看到的模型真实可靠。例如，拥有300个正面例子、100个反面例子的规则，和拥有3个正面例子、1个负面例子的规则一样，它们训练数据的准确率都达到75%，但第一个规则几乎可以肯定比抛硬币好用，而第二个则不然，因为抛4次硬币，可以很容易得出3次正面朝上。在构建规则时，如果某一时刻无法找到能提高该规则准确度的条件，那么我们只能停下，即便它还包括一些负面例子。这样做会降低规则的训练集准确度，也可能让它变成一个更能准确概括的规则，这是我们关心的。

虽然如此，我们还没有大功告成。如果我尝试了一条规则，400个例子中的准确率是75%，可能会相信这个规则。但如果我尝试了100万条规则，其中最佳的规则中，400个例子中的准确率是75%，可能就不会相信这个规则，因为这很有可能是偶然发生的。这也是你在挑共同基金时遇到的问题。Clairvoyant基金连续10年获得高于市场平均水平的收益。哇！这个基金的经理一定是个天才，不是吗？如果你有1000

种基金可以选择，很有可能你选的那个会比Clairvoyant基金的收益还要高，即使那些基金都由玩飞镖的猴子秘密经营。科学文献也被这个问题困扰着。显著性检验是决定一项研究结果是否值得出版的“黄金标准”，但如果几个组找一个结果，而只有一个组找到了，那么很有可能它并没有找到。虽然你阅读他们看起来很可靠的文章，但是绝不会猜到那样的结果。有一个解决方法：出版有肯定结果的文章，同时发表有否定结果的文章，那样你就会知道所有尝试失败的例子，但这种做法没有流行起来。在机器学习中，我们可以把自己尝试了多少条规则记录下来，然后相应地调整显著性检验，不过那样做，这些检验可能会把可靠的规则连同不可靠的一起丢弃。更好的方法就是认识到有些错误的假设会不可避免成立，但要控制它们的数量，方法就是否定低显著性的假设，然后对剩下的假设做进一步的数据检测。

另外一个流行的方法就是选择更加简单的假设。“分而治之”算法会含蓄地选择更简单的规则，因为它在一出现只有正面例子的情况时，就会停止添加条件；在一出现包含所有正面例子的情况时，就会停止添加规则。但为了和过拟合做斗争，我们要对更简单的规则有更强的偏好，这样就能在所有负面例子被包含之前，就停止添加条件。例如，我们可以稍微降低规则的准确度，来缩短规则的长度，然后把这种做法当作一个测评指标。

对较简单假设的偏好就是众人皆知的奥卡姆剃刀原理（Ocam’s razor），但在机器学习背景下，这有点误导性。“如无必要，勿增实体”，因为剃刀常常会被替换，仅意味着挑选能够拟合数据的最简原理。奥卡姆可能对这样的想法感到迷惑，也就是我们会偏向那些不那么能完整解释论据的理论，因为这个理论的概括性更好。简单的理论更受欢迎，因为它们对于我们来说，花费的认知成本更低；对于我们的算法来说，花费的计算成本更低，这不是因为我们想让这些理论更准确。相反，即使是我们最复杂的模型，也往往是在事实过分简化之后得到的。甚至在那些能够完美拟合数据的理论中，我们由“天下没有免费的午餐”这个定理，知道没有什么能够保证最简单的理论最擅

长概括，而实际上，有些最佳的学习算法，比如推进和支持向量机，能了解那些看起来过于复杂的模型（在第七章和第九章中我们会了解它们为什么有这样的功能）。

如果你的学习算法检测集准确度不尽如人意，你就得诊断问题在哪里。是因为无知，还是因为幻想？在机器学习中，这些的专业叫法为“偏差”和“方差”。某座钟如果总是慢一个小时，那么它的偏差会很高，但方差会很低。但如果这座钟走得时快时慢，最后平均下来准点了，那么它的方差会很高，但偏差会很低。假设你和一些朋友在酒吧喝酒、玩飞镖。他们不知道你已经练了多年飞镖，所以非常熟练。你的所有飞镖都打到靶心，你的偏差和方差都很低，效果图3－1c所示。

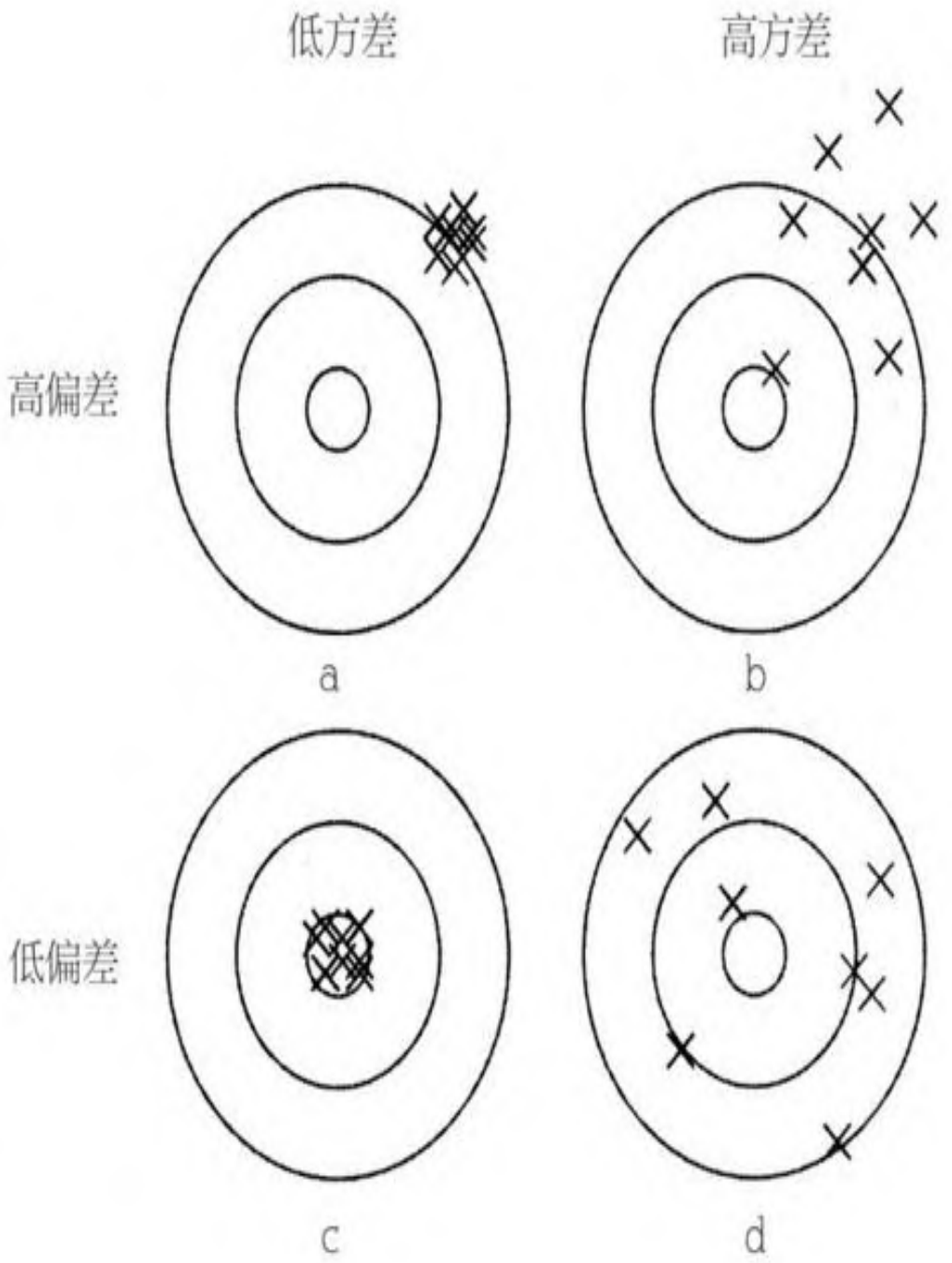


图3－1

你的朋友本也玩得很好，但他喝得有点多。他的镖射完了，他大声说，平均下来是他射中了靶心（也许他本该是一位统计学家）。这

是低偏差、高方差的例子，如图3－1d所示。本的女朋友艾希莉射得很稳，但射的时候过于偏向右上角。她的方差低，但偏差高（见图3－1a）。科迪，从外地来，从来没有玩过飞镖，射的镖都位于中心上方，偏离中心，他的偏差和方差都高（见图3－1b）。

你可以估算一种学习算法的偏差和方差，方法就是在掌握训练集的随机变量之后，对算法的预测进行对比。如果算法一直出错，那么问题就出在偏差上，而你需要一个更为灵活的学习算法（或者只和原来的不一样即可）。如果出现的错误无模式可循，问题就出在方差上，而你要么尝试一种不那么灵活的学习算法，要么获取更多的数据。大多数学习算法都有一个“把手”，通过旋转“把手”，你可以调节这些算法的灵活度，例如，显著性检验的界限值，或者对于模型规模的惩罚方式。扭动“把手”是你尝试的第一个方法。

归纳是逆向的演绎

更深层的问题是，多数学习算法开始时掌握的东西很少，即使转再多“把手”，也没法让这些算法到达终点。没有成年人大脑中所储存知识的指导，这些算法很容易误入歧途。虽然多数学习算法会误入歧途，但只是假设你明白真理的形式（例如，那是一小组规则），还不足以达到令人震惊的程度。严格的经验主义者会说，开始时知之甚少，这是所有新生儿都有的特点，因为这个特点深深存在于其大脑结构中，而孩子确实会比成人更容易犯过拟合的错误，但我们想学得比孩子快（18年是很长的一段时间，而且还没把大学教育时间计算在内）。主算法应该能以大量的知识作为启动（无论这些知识由人类来提供，还是之前已经掌握），然后在对数据做出新概括时，用到这些

知识。这就是科学家做的工作，这相当于从零开始做。“分而治之”原则的归纳算法做不到这一点，但还有别的掌握规则的方法能做到。

问题的关键在于认识到，归纳仅仅是逆向演绎，就和减法是加法的逆运算，或者积分是微分的逆运算一样。这个观点由威廉姆·斯坦利·杰文斯于19世纪末首次提出。史蒂夫·马格尔顿和雷·邦坦（一个英国—澳大利亚研究组）在此基础上于1988年设计出第一个实用算法。在数学这个学科中，通过已知条件进行运算，得出其逆向结果的方法已经有一段传奇的历史。把这种方法运用到加法当中就有了整数，因为如果没有负数，加法就不可逆了（3-4=-1）。同样，将其用到乘法中就有了有理数，而将其运用到平方运算中就有了复数。让我们来看看能否将这种方法运用到演绎中。演绎推理的一个典型例子就是：

苏格拉底是人类。

所有人类都会死。

所以……

第一个句子是关于苏格拉底的事实，第二个是关于人类的一般规则。接下来的推理是怎样的？苏格拉底当然也会死，这是将一般规则用到苏格拉底身上得出的。相反，我们在归纳推理中会以最初事实和衍生事实作为开始，然后找一个规则，让我们由前者推出后者：

苏格拉底是人类。

……

所以苏格拉底也会死。

有这样一个规则：如果苏格拉底是人类，那么他就会死。这句话完成了推理，但并不是很有用，因为这是专门针对苏格拉底的规则。但现在我们应用牛顿定律，然后将该定律推广到所有实体当中：如果某实体是人类，那么它就会死。或者更简洁些：所有人类都会死。当然，仅通过苏格拉底就归纳出该规则过于草率，但对于其他人类，我们知道相似的事实：

柏拉图是人类，他会死。

亚里士多德是人类，他会死。

以此类推……

对于每个事实，我们构建这样的规则，让我们由第一个事实推出第二个事实，然后通过牛顿定律将其推广。当同一条通用规则一次又一次被归纳出来时，我们有信心说那条规则说的是真的。

到目前为止，我们还没有做任何“分而治之”算法无法做到的事。可是，如果我们不知道苏格拉底、柏拉图、亚里士多德是人类，只知道他们是哲学家，那会怎样？我们仍会得出结论，说他们也会死，因为之前我们就归纳过或者别人告诉过我们，所有人都会死。这也是一种有效推广（至少在我们解决人工智能问题，而机器人开始从事哲学研究之前是），而且这也在我们的推理中“填补了漏洞”：

苏格拉底是哲学家。

所有哲学家都是人。

所有人都会死。

所以苏格拉底会死。

我们也可以单纯从其他规则中归纳另一些规则。如果我们知道所有哲学家都是人，而且会死，我们就可以归纳出所有人都会死（我们不会归纳出所有会死的都是人类，因为我们知道其他会死的动物，例如猫和狗。另一方面，科学家、艺术家等也是人类，也会死，因此这条规则得到加强）。通常，我们以越多的规则和事实作为开头，也就有越多的机会运用“逆向演绎”归纳新的规则。我们归纳的规则越多，我们能归纳的规则也就越多。这是知识创造的良性循环，只受过拟合风险和计算成本的限制。可是在这里，我们也有初始知识的协助：如果我们有很多小的而不是大的漏洞要修补，归纳的步骤就不会有那么大的风险，所以过拟合的可能性也会降低（例如，给定相同数量的例子，“所有哲学家都是人类”这个归纳，就比“所有人都会死”这个归纳风险要小）。

逆运算往往比较困难，因为逆运算的结果不止一个。例如，一个整数有两个平方根：一个正数，一个负数［$2^2=(-2)^2=4$］。最为人所知的是，对一个函数的导数积分，只会将函数恢复为一个常数。函数的导数告诉我们该函数在每个点上下浮动的幅度。把所有幅度值加起来会重新得到函数，除非我们不知道它从哪里开始变化。我们可以在不改变导数的情况下，上下“滑动”积分过的函数。为了简便，我们可以假设附加常量为0，来“取缔”函数。逆向演绎存在类似的问题，而牛顿定律就是一个解决方法。例如，我们由“所有希腊的哲学家都是人类”和“所有希腊哲学家都会死”可以归纳出“所有人类都会死”，或只能归纳出“所有希腊人都会死”。可是为什么仅满足于最保守的归纳？其实我们可以认为所有人都会死，直到遇到例外（据雷·库兹韦尔说，这个意外很快就会出现）。

同时，逆向演绎的另外一个重要的应用，就是预测新研制的药物是否有不良的副作用。动物测试及临床试验的失败，成为研制新药物需要花费多年、数十亿美元的主要原因。通过归纳已知有毒分子的结

构，我们可以得出规律，迅速清除明显的有毒化合物，大大提高剩余化合物的临床试验成功率。

掌握治愈癌症的方法

更广泛地说，逆向演绎是在生物学中发现新知识的重要方法，这也是治愈癌症要迈出的第一步。根据中心法则的观点，在活细胞中进行的任何活动都最终由细胞的基因控制，通过发起蛋白质的合成来完成。一个细胞就像一台微型计算机，而DNA就是计算机运行的程序：改变DNA，皮肤细胞就会变成神经元细胞，或者小鼠细胞会变成人类细胞。在计算机的程序中，所有的故障都是由程序员引起的。但在细胞中，故障可自行产生，比如当辐射或者细胞复制误差将某基因变成另外一个基因时，当基因偶然被复制两次时……很多时候，变异会导致细胞悄无声息地死去，但有时候细胞开始生长，然后不可控制地分裂，这样癌细胞就产生了。

治愈癌症意味着在不破坏完好细胞的情况下，阻止受损细胞的繁殖。这就需要知道完好细胞和受损细胞的区别在哪里，特别是它们染色体组的区别在哪里，因为其他所有的一切都依此进行。幸运的是，基因测序变得越来越普遍、越来越实惠。利用基因测序，我们可以预测哪种药对哪种癌基因起作用。这和传统的化学疗法不一样，化学疗法对所有细胞的影响都一样。掌握哪种药物对哪种变异有效，这需要关于患者的数据库、他们的癌基因、服用过的药物以及药效。最简单的规则对基因与药物之间一对一的对应关系进行编码，例如，如果出现BCR－ABL基因，那么服用格列卫这种药物（BCR－ABL引发一种白血病，而格列卫在10个患者中能治愈9个患者）。一旦癌基因排序和治疗结果核对达到标准管理，更多这样的规则会被发现。

然而，这仅仅是一个开始。大多数癌症都会涉及各种混合的变异，或者只有靠某些药物才能治愈，而这些药物尚未被研发出来。下一步就是以更为复杂的条件来掌握规则，这涉及癌基因组、患者的基因组、病史、药物的已知副作用等。但最终我们需要的，就是完整细胞运转的模型，这让我们在计算机上对特殊病人变异的影响、不同药物组合的效果进行模仿，无论是真实存在的，还是猜想出来的，都进行模仿。我们构建此类模型的主要信息，主要源于DNA测序仪、基因表达微阵列以及生物学文献。将这些结合起来，人们就会看到逆向演绎的作用。

亚当（我们在第一章中提到过的机器人科学家）就是一个例子。亚当的目标是弄明白酵母粉如何起作用。机器人以酵母遗传学、新陈代谢的基本知识，以及一批酵母细胞的基因表达数据作为开端，运用逆向演绎来假设哪种基因会通过哪种蛋白质来表达，然后设计微阵列实验来对其进行检测，接着修正假设，最后重复以上步骤。是否每个基因都会表达出来，这取决于其他基因，以及所处的环境条件，由此产生的网络交互可表达为一个规则集，例如：

如果温度高，那么基因A就会表达出来。

如果基因A表达出来，而基因B没表达出来，则基因C表达出来。

如果基因C表达出来，那么基因D不会表达出来。

如果我们知道的是第一条和第三条规则，不知道第二条，但我们有微阵列数据，数据显示在高温条件下，基因B和基因D无法表达，那么我们就可以通过逆向演绎来归纳出第二条规则。一旦我们有了这条规则，而且也许利用微阵列实验证实了这条规则，那么我们就能用它作为进一步归纳推理的基础。同理，我们也可以将化学反应的顺序拼凑起来，因为蛋白质就是按照这些顺序发挥作用的。

即便如此，仅仅知道哪些基因调节哪些基因，以及蛋白质如何组织化学反应中的细胞网络还不够。我们还需要知道，每种分子产生的量有多少。DNA微阵列和其他实验能够提供这种数量方面的信息，但逆向演绎，以其“非有即无”的逻辑特点，并不是很擅长处理这方面的事情。因此我们要用到联结学派的方法，这会在下一章谈到。

20问游戏

逆向演绎的另外一个局限性就在于，它涉及很密集的计算，因此很难扩展到海量数据集中。因为这些原因，符号学家选择的算法是决策树归纳。决策树可以当作此类问题的答案：如果有多个概念的规则对应一个实例，那怎么办？那么我们怎么知道实例对应哪个概念呢？如果看到一个部分封闭的物体，它有一个平面、四条腿，那么我们怎么知道它是一张桌子，还是一把椅子呢？有一个方法，就是对规则进行排序，例如以准确率递减的顺序来排列，然后选择符合描述的第一条规则。另一个方法，就是让规则自己选择。决策树通常会保证，每个实例会准确对应一条规则。也就是说，在一次及以上的属性测试中，如果每对规则存在区别，这样的规则集将被组织成一棵决策树。例如，看看以下这些规则：

如果你支持削减税收，反对堕胎，那么你属于共和党。

如果你反对削减税收，那么你属于民主党。

如果你支持削减税收，提倡堕胎合法，反对枪械管制，那么你属于独立人士。

如果你支持削减税收，提倡堕胎合法，支持枪械管制，那么你属于民主党。

这些可以组成一些决策树，如图3－2所示。

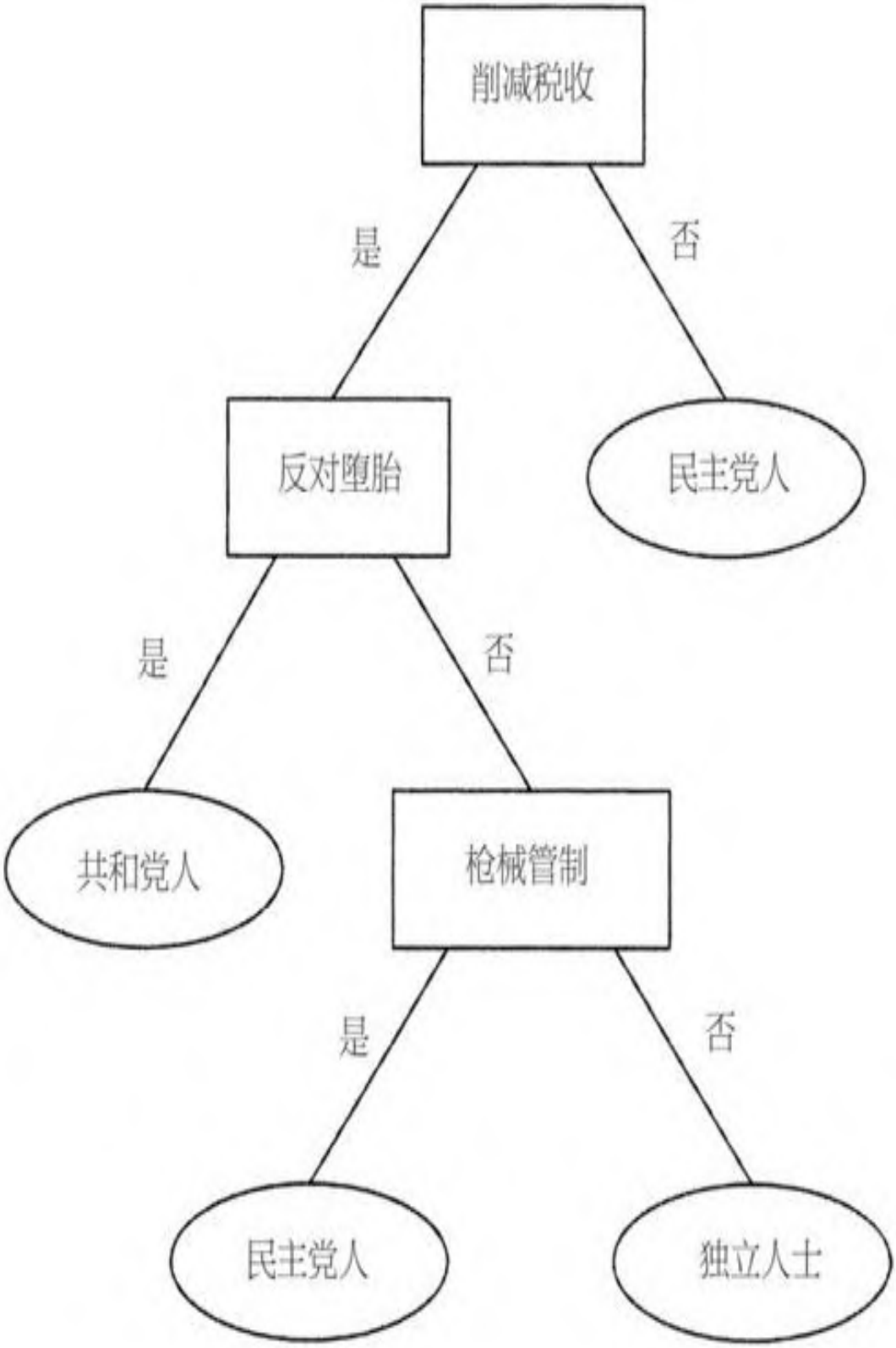

削减税收
是
否
反对堕胎
民主党人
是
否
共和党人
枪械管制
是
否
民主党人
独立人士

图3－2

决策树的原理就像玩一个有实例的20问游戏。从“根部”开始，每个节点都会问每个属性的值，然后根据答案，我们沿着这个或另外一个分支继续下去。当到达“树叶”部分时，我们读取预测的概念。从“根部”到“树叶”的每条路线都对应一个规则。这让人想起那些烦人的电话语音提示菜单，如果你拨打客户服务电话，就得通过这些菜单，这些菜单与决策树相似并非偶然——电话语音提示菜单就是一棵决策树。电话线另一头的计算机在和你玩20问游戏，目的是弄明白你想要做什么，每个选项就是一个问题。

根据以上决策树，你要么是共和党人，要么是民主党人，要么是独立人士。你无法选择其中的两种或三种，或者一个都不选。拥有这个属性的概念组被称为类集，而预测类集的算法称为分类器。单个概念隐含两类定义：概念本身及其反面（例如，垃圾邮件和非垃圾邮件）。分类器是机器学习最为普遍的方式。

利用“分而治之”算法的一个变体，我们就可以掌握决策树了。首先，我们选一个属性，在决策树的“根部”进行测试。然后我们关注每个分支上的例子，为那些例子选择下一个测试（例如，我们看看支持削减税收的人是否反对堕胎或者支持堕胎合法）。我们在归纳推理出的每个新节点上重复这个步骤，一直到分支上的所有例子都有同一个类别，此时我们就可以用该类别来命名该分支了。

有一个突出的问题，那就是如何挑选最佳属性以便在节点处进行测试。准确度（准确预测例子的数量）并不能起到很好的作用，因为我们不是在尝试预测某个特殊类别，而是在尝试慢慢分离每个类别，直到每个分支都变得“纯粹”。这使人想起信息论中熵的概念。一组对象的熵，就是用来衡量混乱度的单位。如果150人的组里面有50个共和党人、50个民主党人、50个独立人士，那么这个组的政治熵会达到最大。另一方面，如果这个组全部是共和党人，那么熵就变成零（这就是党派联合的目的）。所以为了学习一棵好的决策树的优点，我们

在每个节点选择这样的属性：在其所有分支中，产生的熵在平均值上属性最低，取决于每个分支上有多少例子。

和规则学习一样，我们不想归纳出一棵树，可以准确地预测所有训练例子的类别，因为这样的树很有可能会过拟合。和之前一样，为了防止这样的事情发生，我们可以利用显著性检验或针对树的大小设立惩罚制度。

如果属性离散，属性的每个值都有一个分支，这没关系，但如果是数值属性该怎么办？如果连续变量的每个值都有一个分支，决策树将变得无限宽。一个简单的方法就是通过熵来选择几个临界值，然后使这些临界值起作用。例如，患者的体温是高于还是低于100华氏度？这个体温数和其他症状一起，也许是所有医生都要知道的，通过体温可以断定患者是否感染疾病。

决策树可应用在许多不同的领域。在机器学习领域，决策树源于心理学方面的知识。厄尔·亨特及其同事于20世纪60年代利用了决策树，目的是为了模拟人类如何掌握新的概念。另外，亨特其中的一个研究生罗斯·昆兰后来尝试把决策树用于象棋中。他最初的目的是为了从棋盘占位预测王车大战骑士的后果。据调查，决策树已经由一开始这些微不足道的用途，发展成为机器学习算法中应用最为广泛的方法。要知道原因并不困难：决策树易于理解，可以快速掌握，而且通常无须太多调整就可以做到准确无误。昆兰是符号学派中最卓越的研究者。作为一个沉稳、务实的澳大利亚人，因为对决策树年复一年不断地改进，并写了和决策树有关、逻辑清晰的论文，使决策树成为分类活动中的“黄金标准”。

无论你想预测什么，人们利用决策树来预测的可能性会很大。微软的Kinect利用决策树，通过其深度相机的输出信息，可以弄明白你身体的各个部位在哪里，然后利用这些部位的动作来控制Xbox游戏机。在2002年的一场正面交锋中，决策树准确预测了3/4的最高法院裁决，而一个专家小组的准确率却不到60%。数千个决策树使用者不可能

会出错，你仔细想想，然后画一棵树来预测朋友的答复（如果你想约她出去的话）。如图3－3所示。

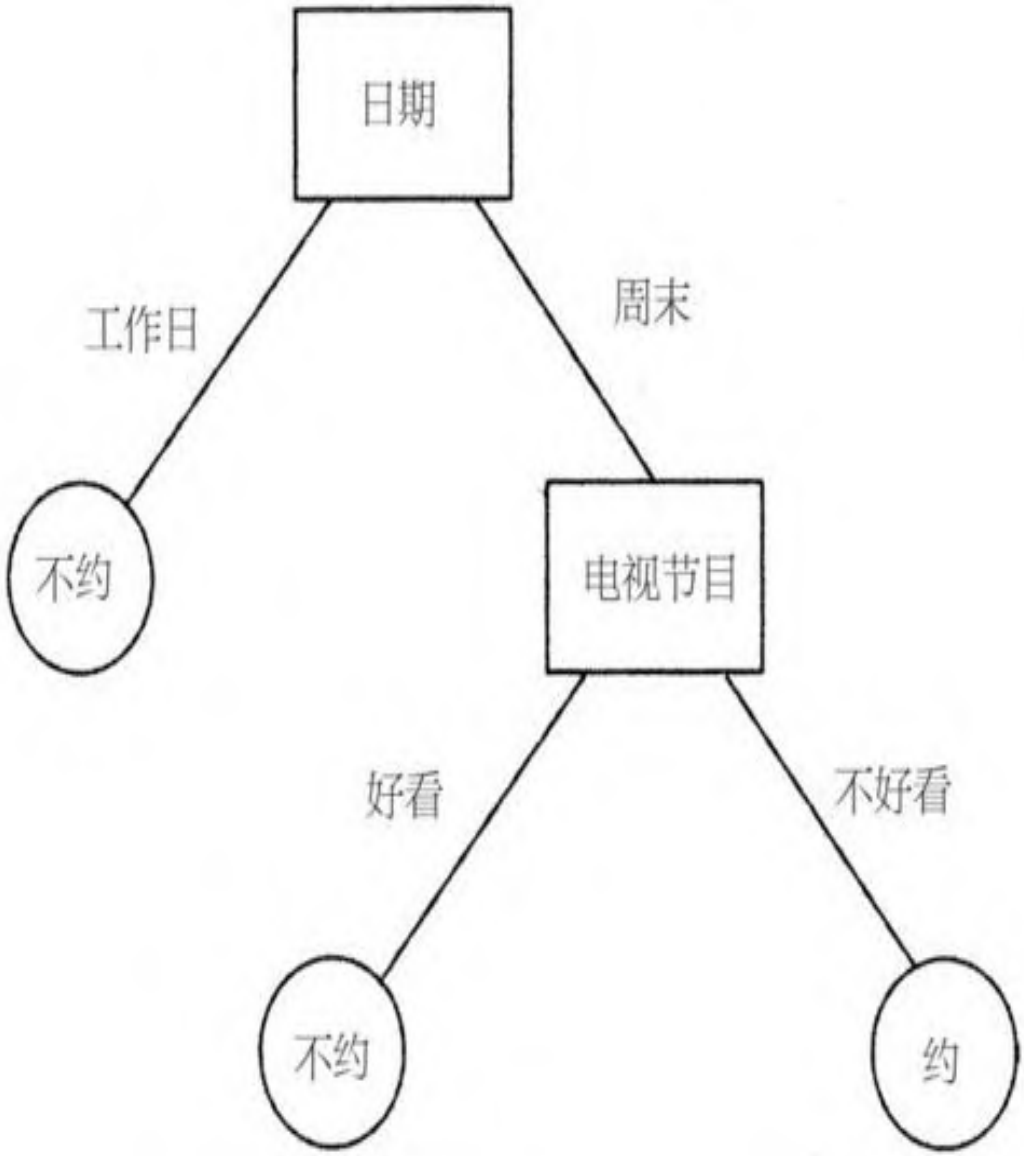


图3－3

根据这棵树的结果，今晚她会答应你出去约会。深呼吸，拿起手机，拨通她的电话吧。

符号学派

符号学派的核心理念就是，所有和智力相关的工作都可以归结为对符号的操纵。数学家在解方程时，会移动符号，然后根据预先定义的规则，用其他符号来代替这些符号。逻辑学家进行推论时也是同样的道理。根据这个假设，智力是独立于基质的。符号处理是通过写在黑板上进行的，还是通过打开或关闭晶体管、放电神经元，或者玩玩积木就能完成的，这些都不重要。如果你能利用万能图灵机的力量来进行设置，那么就能做任何事情。软件可以和硬件清晰地分离。如果你关注的重点是想弄明白机器是怎样学习的，你（谢天谢地）不用担心后者无法在亚马逊的云服务上购买个人计算机或者自行车。

符号学机器学习者和许多其他计算机科学家、心理学家、哲学家一样，都相信符号操纵的力量。心理学家大卫·马尔称，每个信息处理系统应该经过三个不同水平的研究：该系统解决所解决问题的基本属性，用来解决问题的算法和表示方法，以及这些算法和表示方法如何实现。例如，加法可以由一组公理来定义，和加法如何进行无关；数字可以以不同方式进行表达（例如，罗马数字和阿拉伯数字），还可以用不同算法进行相加；而这些运算可以通过算盘、袖珍计算器，或者效率更低的方式——你的大脑来进行。根据马尔的水平理论，学习是我们能够研究的认知能力的典型例子，而且对我们来说受益匪浅。

符号主义机器学习是人工智能知识工程学派的一个分支。20世纪70年代，所谓的基于知识的系统取得卓越成绩，而到了80年代，它们

迅速传播，后来却消失了。它们消失的主要原因是人人逃避的知识习得瓶颈：从专家身上提取知识，然后将其编码成为规则，这样做难度太大、太费力、易出故障，会引起很多问题。让计算机自行学习，比如通过查看过往患者症状及其相应疗效的数据库，就可以进行疾病诊断，比无数次地找医生要容易很多。突然之间，诸如理夏德·米哈尔斯基、汤姆·米切尔、罗斯·昆兰之类的先驱人物，他们的工作有了新的联系，而这个领域自此没有停止发展过（另外一个重要的问题就是基于知识的系统在处理不确定性时会有问题，在第六章会深入探讨这个问题）。

因为其起源和指导原则，符号学派和其他学派相比，和人工智能的其他方面关系更为密切。如果计算机科学是一块大陆，符号主义机器学习和知识工程学会有很长的交界线。知识通过两个方向进行交易——手动输入的知识，供学习算法使用；还有归纳得出的知识，用来加入知识库中，但最终理性主义者和经验主义者的断层线会刚好落在这条界线上，想越过这条界线则不容易。

符号主义是通往终极算法的最短路程。它不要求我们弄明白进化论和大脑的工作原理，而且也避免了贝叶斯主义的数学复杂性。规则集合决策树易于理解，所以我们知道学习算法要做什么。这样它可以轻易算出自己做对或做错什么，找出故障，得出准确结果。

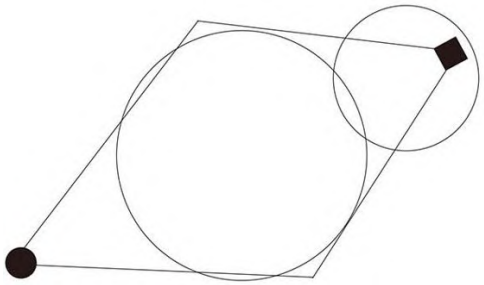
尽管决策树很受欢迎，但逆向演绎是寻找主算法更好的出发点。因为逆向演绎具备这样的关键属性：可以轻易地将知识并入主算法中，而且我们知道休谟问题使这一点变得很有必要。另外，规则集和决策树相比，表达多数概念的方式要简洁很多。把一棵决策树转变成一个规则集很容易：每条从“根部”到“叶子”的路线是一条规则，而且路线不会崩溃。另外，最坏的情况是，把一个规则集转化成一棵决策树，需要把每条规则变成一棵迷你决策树，然后用规则2决策树的副本来代替规则1决策树的每片叶子，用规则3决策树的副本来代替规则2决策树每个副本的每片叶子，以此类推，这样会引起大范围崩溃。

逆向演绎就像一个超级科学家，系统查看论据，思考可行的归纳法，整理最有利的证据，然后将这些和其他论据一起，进一步提出假设——所有过程都基于计算机的速度。逆向演绎简洁而美观，至少符合符号学者的品位。此外，逆向演绎也有一些严重的缺点。可行的归纳法数量广泛，除非我们和最初知识保持亲密关系，否则很容易在空间中迷失。逆向演绎容易被噪声迷惑：我们怎样才能知道，哪些演绎步骤被漏掉了，如果前提或者结论本身就已出错？最严重的是，真正的概念很少能通过一个规则集来定义。它们不是黑，也不是白，比如垃圾邮件和非垃圾邮件之间有一片很大的灰色区域。要获取真正的概念，就得权衡并收集有弱点的论据，直到出现清晰的定义。疾病的诊断，涉及把重点放在一些症状上面，然后放弃那些不完整的论据。还没有人能只学习一个规则组，就能通过观看图片的像素来认出一只猫，而且可能以后也没人能做到。

联结学派对符号学派尤其不满。根据他们的观点，你能通过逻辑规则来定义的概念仅仅是冰山一角，其实表面之下还有很多东西是形式推理无法看到的。而同样的道理，我们脑子里所想的东西也是潜意识的。你不能仅靠构造一个空洞的机械化科学家，就想让他把所有有意义的事情完成，你首先得给他点什么东西，例如一个真正的大脑，能和真实的感觉相连，在真实世界中成长，甚至可能要常常绊他的脚。你怎样才能构造这样的大脑呢？通过逆向分析。如果想对一辆车进行逆向分析，你就应看看发动机盖下面。如果想对大脑进行逆向分析，你就要看看脑壳里面。

第四章

联结学派：大脑如何学习



赫布律，就如它为人们所知的那样，是联结主义的奠基石。确实，联结主义相信知识储存在神经元之间的联结关系中，它也因此而得名。唐纳德·赫布（Donald Hebb）是加拿大的心理学家，他在1949年出版的《行为的组织》（*The Organization of Behavior*）一书中这样说道：“当A细胞的轴突和B细胞足够近，并且重复或不断地对其放电时，A、B中的一个细胞或者两个细胞都会经历生长过程或者代谢改变，这样A细胞（作为对B细胞放电的细胞之一）的效率就会得到提高。”这段话经常被转述成“一起放电的神经元也会被串连在一起”。

赫布律是心理学和神经科学思想的融合，其中掺杂了合理的猜想。通过连接来进行学习，是英国经验主义者最喜爱的话题，从洛克和休谟到约翰·穆勒都是如此。威廉·詹姆斯（William James）在其著作《心理学原理》（*Principles of Psychology*）中，阐明了连接的主要原理，这和赫布律十分相似，只是大脑活动被神经元取代，放电效率被兴奋的传播取代。差不多同时，伟大的西班牙神经学科学家

圣地亚哥·拉蒙·卡哈尔第一次对脑部进行详细观察，利用当时发明的高尔基染色法来对单个神经元进行染色，把他所看到的编成目录，就像植物学家对树木的新品种进行分类一样。赫布时期，神经学科学家对神经元如何发挥功能有了彻底了解，但赫布是第一个提出这种机制的人，通过这个机制可以对连接进行编码。

在符号学派中，符号和它们代表的概念之间有一一对应的关系。相反，符号学派的代表方式却是分散式的：每个概念由许多神经元来表示，而每个神经元又会和其他神经元一起代表许多不同的概念。互相激发的神经元会形成赫布所称的“细胞集”。概念和记忆由细胞集在大脑中表示出来。每个细胞集都可以包含来自不同大脑区域的神经元，也可以和其他集合相互重叠。“腿”的细胞集中包含“脚”的细胞集，包含脚的图片的细胞集，以及脚的声音的细胞集。如果你问一个符号学派系统，“纽约”这个概念在哪里被表示出来，它可以指向存储该记忆的准确位置。在联结学派体系中，答案就是“这个概念通过这里一点、那里一点地被储存起来”。

符号学派和联结学派的另外一个区别就在于，前者是按次序的，而后者是平行的。在逆向演绎中，我们可以一步一步地弄明白，为了从前提出发得到满意的结论，需要哪些新的规则。而在联结学派模型中，根据赫布律，所有的神经元都会同时进行学习。这也反映了计算机和人脑之间的不同属性。计算机做每件事都会一点点来，例如，把两个数相加，或者拉开关。所以为了完成所有有意义的事情，计算机得经过很多步骤，但那些步骤可以很快被完成，因为晶体管每秒可以打开、关闭数十亿次。相反，人脑可以同时进行多项运算，这时数十亿的神经元会同时起作用，但每项运算都会很慢，因为神经元最多可每秒放电1000次。

计算机里晶体管的数量已经赶上人类大脑里神经元的数量，但在连接数量上，人类的大脑轻易获胜。在一台微处理器中，典型的晶体管仅仅和其他几个晶体管直接相连，而派上用场的平面半导体技术对

计算机功能的发挥又有很大的限制。相反，一个神经元就有数千个突触。如果你走在大街上碰到熟人，你认出他只需要0.1秒。以神经转换的速度，这些时间勉强够用来进行100个处理步骤，但在那些处理步骤中，你的大脑能够浏览整个记忆库，找到最佳搭配，然后使其适应新的背景（不同的服装、不同的灯光等）。在大脑中，每个处理步骤有可能会很复杂，而且会涉及很多信息，并符合分散的概念表达方式。

这并不意味着我们就不能利用计算机来模拟人脑，毕竟这是联结学派算法要做的事。因为计算机是通用的图灵机，只要我们给它足够的时间和记忆力，它就能执行大脑的计算，以及别的任何事情。尤其计算机可以利用速度来弥补缺乏连接的劣势，千千万万遍利用同样的线来模拟1000根线。实际上，目前计算机和人脑相比，主要的限制是能量损耗：人的大脑消耗的能量仅仅相当于一个小灯泡，而沃森消耗的电却能点亮整栋办公楼。

然而为了模拟大脑，我们需要的不仅仅是赫布律，还要知道大脑是如何构造的。每个神经元就像一棵小树，有数目惊人的根须（树突）还有细长蜿蜒的树干（轴突）。大脑就是由数十亿棵这样的树组成的森林，但这些树也有不同寻常的地方：每棵树的枝丫都会和其他数千棵树的根部有连接（突触），形成大片你没见过的纠缠状态。有些神经元有很短的轴突，而有些神经元的轴突则很长，可以从大脑的一边缠绕到另一边。你大脑里轴突的长度相当于地球到月亮的距离。

这片森林还会充满电流。火星会沿着树干闪烁，然后会引发相邻树木更多的火花。时不时整个区域的丛林会使自己进入狂热状态，然后又会平静下来。如果你动动脚趾，会发生一系列放电现象，人们称之为“动作电位”。这种放电现象会沿着你的脊髓一直到达腿部，直到到达你的脚趾肌肉，然后告诉肌肉要运动。你的大脑运转时的情景就是这些电火花火光四射的场面。如果你能坐在大脑里面，看看你阅读这页书时大脑发生了什么，你看到的情景会让科幻小说里最繁忙的

都市景象也逊色几分。这个十分复杂的神经元放电模式的背后，就是你的意识在起作用。

在赫布时代，没有什么方法能够测量突触的强度或者发生在其内部的变化，更不用说弄明白突触变化的分子生物学相关信息。如今，我们知道，当突触后神经元在突触前神经元之后会很快放电时，突触会变大（或重新形成突触）。和所有的细胞一样，神经元里外有不同的离子浓度，穿过神经元的细胞膜形成一股电压。当突触前神经元放电时，微小的囊会向突触间隙释放神经递质分子。这会使突触后神经元的膜中的通道打开，让钾离子和钠离子进入，最终会改变通过膜的电压。如果有足够多的突触前神经元一起放电，电压会突然升高，一个动作电位会顺着突触后神经元的轴突而下。这还会使离子通道变得更加灵敏，并出现新的通道，对突触进行加强。就我们的知识所能达到的水平，这就是神经元进行学习的过程。

下一步就是把这个过程运用到算法中。

感知器的兴盛与衰亡

第一个正式的神经元模型是由沃伦·麦卡洛克（Warren McCulloch）和沃尔特·皮茨（Walter Pitts）于1943年提出的。这个模型看起来很像组成计算机的逻辑门。当“或”门至少一个输入开关打开时，“或”门开通，当所有输入开关打开时，“且”门开通。当其输入的活跃信息超过某个界限值时，一个麦卡洛克—皮茨神经元会打开。另外，如果界限值只有一个，神经元就相当于一道“或”门；如果界限值等于输入信息的数量，神经元就相当于一道“且”门。另外，一个麦卡洛克—皮茨神经元会阻止另外一个麦卡洛克—皮茨神经元打开，这就模拟了抑制性突触和“非”门。因此神经元的一个网络

就可以完成计算机的所有计算。在早些年，计算机常常被称为“电脑”，这不只是一个类比那么简单。

麦卡洛克—皮茨神经元做不了的事情就是学习。为此我们需要对神经元之间的连接给予不同的权重，这就是所谓的“感知器”。感知器于20世纪50年代由康奈尔大学的心理学家弗兰克·罗森布拉特（Frank Rosenblatt）发明。作为富有魅力的演讲者和活跃人物，罗森布拉特在早期机器学习领域形成的过程中，比任何人付出的都要多。“感知器”这个名称源于他的兴趣，他喜欢将自己的模型应用到诸如演讲和字符识别的感知任务中。他没有将感知器应用于软件中，因为在那个年代这个过程会非常缓慢，他构建了自己的装置。权值是由可变电阻器来实现的，这些电阻器和调光灯开关中的那些电阻器一样，而且权值学习由电动机来实现，电动机会打开电阻器上的开关。

在感知器中，一个正权值代表一个兴奋性连接，一个负权值代表一个抑制性连接。如果其输入量的加权和高于界限值，那么会输出1；如果加权和小于界限值，那么输入0。通过改变权值和界限值，我们可以改变感知器计算的函数。当然，这种做法忽略了神经元发挥作用的很多细节，但我们想让这个过程尽可能简单点。我们的目标是形成一种多用途的学习算法，而不是构建一个大脑的现实模型。如果被我们忽略的一些细节最后证明很重要，可以在后面加上去。尽管我们把抽象的东西简化了，却仍可以看到这个模型的每个部分都与神经元的一个部分相对应（见图4－1）。

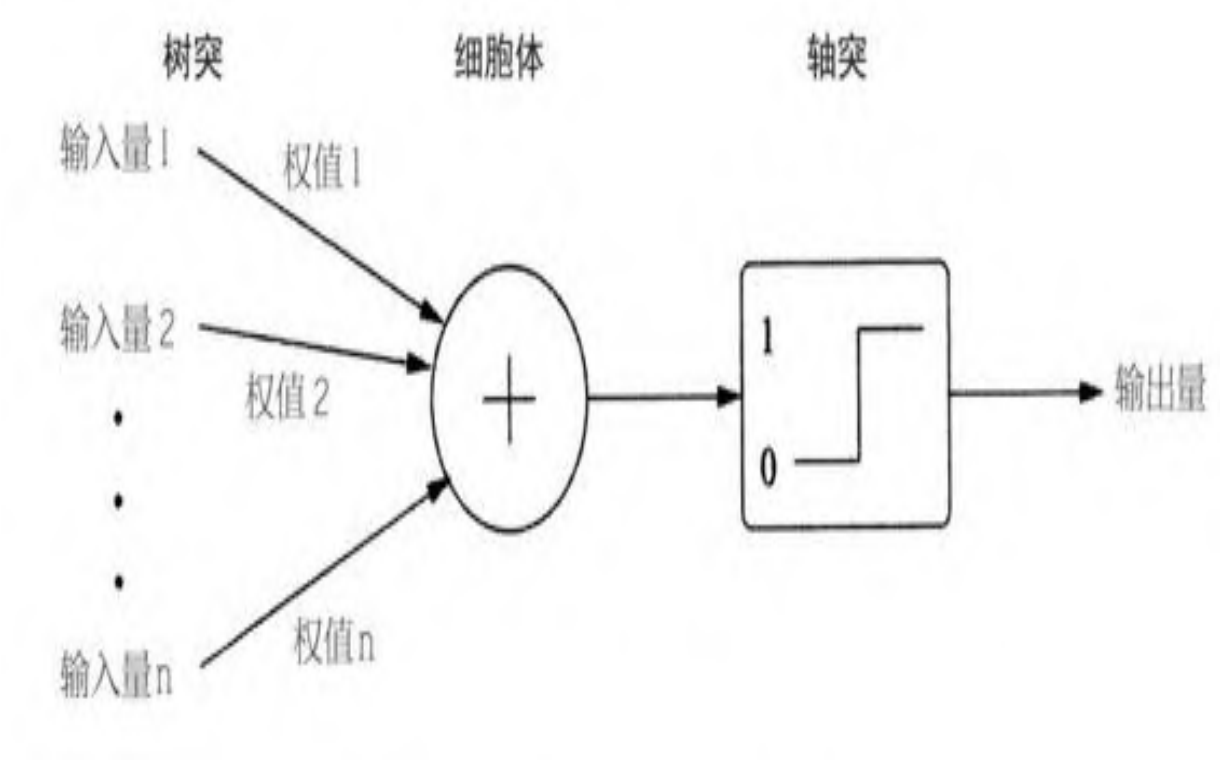


图4－1

输入量的权值越高，相应的突触也会越强。细胞体把所有加权输入量加起来，轴突使结果变成一个阶跃函数。图中代表轴突的框表明了阶跃函数的图像：0代表输入量的低值，当输入量到达界限值时，会突然变成1。

假设感知器有两个持续输入量x和y（换句话说，x和y可以是任何数值，不仅仅是0和1），那么每个例子都可以由平面上的一个点来表示，而正面例子（例如，感知器中输出量为1）和负面例子（输出量为0）之间的界线就是一条直线（见图4－2）。

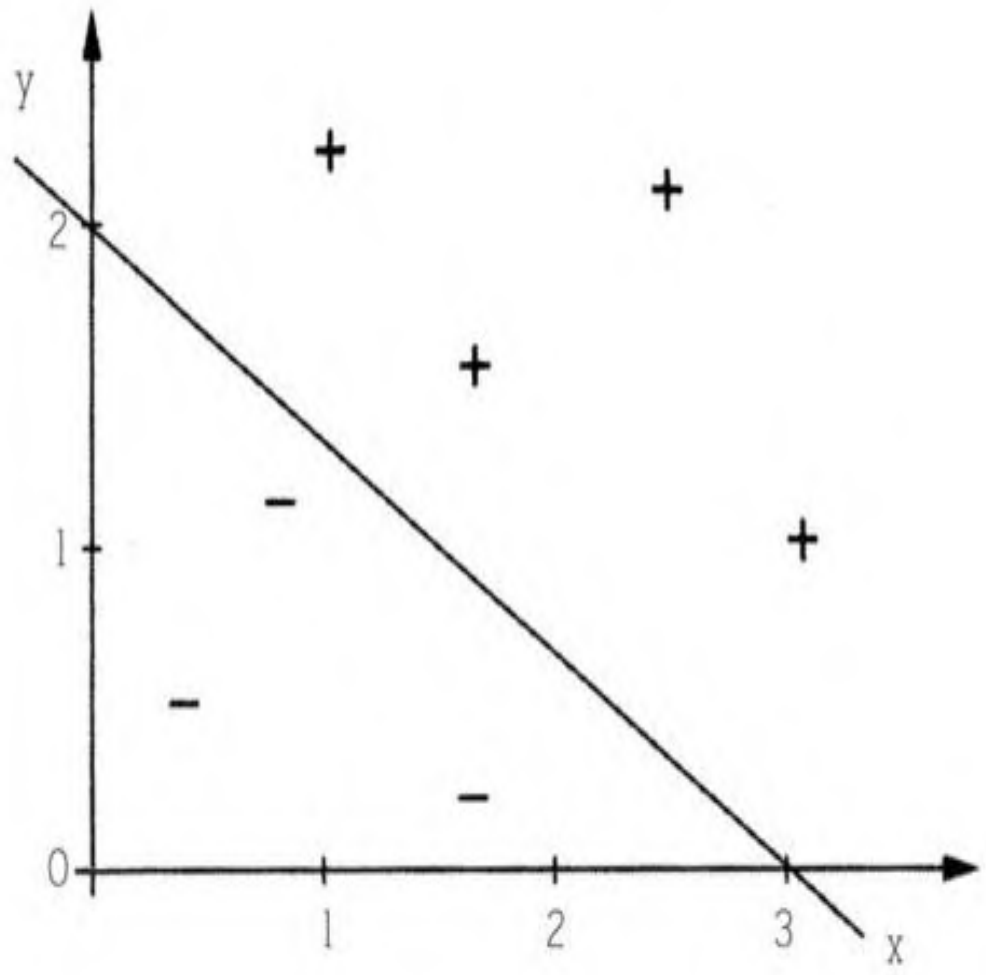


图4－2

这是因为这条界线是由一系列点组成的，在这些点上，加权和刚好等于界限值，而一个加权和就是一个线性函数。例如，如果x的系数是2，y的系数是3，那么界限值是6，界限就由方程式$2x+3y=6$来确定。$x=0$，$y=2$的点就在界线上，为了停留在这条线上，我们要水平向右平移3个单位，同时垂直向下平移2个单位，这样x增加的量就是y减少的量，经过的点就形成一条直线。

掌握感知器的权值意味着要改变直线的方向，知道所有的正面例子都在这边，所有的负面例子都在另一边。在一维概念中，这个界线

是一个点；在二维概念中，它是一条直线；在三维概念中，它是一个面；在超过三维的概念中，它是一个超平面。在超平面中很难把东西形象化，但数学就是这样发挥作用的。在n维概念中，我们有n个输入量，而感知器有n个权值。为了确认感知器是否放电，我们把每个权值乘以相应的输入量，然后将得数相加，和界限值进行对比。

如果所有的输入量都有一个权值1，而且界限值是输入量数值的一半，当感知器超过一半的输入量放电时，那么感知器放电。换句话说，感知器就像一个微型社会，少数服从多数（可能也没有那么小，因为它有数千个成员）。尽管如此，整体来说它并不那么民主，因为一般不是人人都平等地享有选择权。神经网络看起来更像社交网络，在这里，几个亲密朋友比脸书上的数千个好友更加重要。最信任的朋友对你的影响也最大。如果有个朋友推荐了一部电影，你去看了，觉得喜欢，下次可能还会再听他的建议。另一方面，如果他滔滔不绝地讲你不喜欢的电影，你会开始忽略他的建议（甚至你们的友情都可能变淡）。

这就是罗森布拉特的感知器算法掌握权值的方法。

想想祖母细胞，这是认知神经学科学家最热衷的思维实验。祖母细胞是你大脑中的一个神经元，无论何时你看到自己的祖母，它都会放电，而且只有这样的情况才会放电。祖母细胞是否真实存在有待探寻，但我们可以设计一个，并将其应用于机器学习。下面，我们来分析一下感知器学习识别你的祖母的过程。进入细胞的要么是图片中的原始像素，要么就是图片的各种固有特征，就像棕色的眼睛，如果图片上有一双棕色的眼睛就会取值1，否则是0。一开始，从特征到神经元的所有连接有小的随机值，就像刚出生时你大脑中的突触一样。那么我们向感知器展示了一组图片，有些是你祖母的，有些不是。如果感知器在看到一张你祖母的照片时会放电，或者在看到别的图片时不会放电，那么就不需要进行学习活动（如果东西没坏，就别修理它）。但如果感知器在看到你的祖母时没有放电，这意味着它的输入

量的加权和应该取更高的值，这样我们就增加进行中的输入量的权值（例如，如果你的祖母有一双棕色的眼睛，这个特征的权值会增长）。相反，如果感知器在不该放电的时候放电，我们就减少活跃的输入量的权值。误差迫使学习活动正常进行。经过一段时间，表明你祖母的特征会取得更高的权值，相反则得到低权值。一旦感知器看到你祖母时就会放电，而且也只有这时，学习才算完成。

感知器会产生许多兴奋。它虽然简单，但仅凭借例子训练，它就能识别书面文字和语音。罗森布拉特在康奈尔大学的同事证实，如果正面例子和负面例子能够通过超平面来分离，那么感知器会找到超平面。对于罗森布拉特和其他人来说，真正理解大脑如何进行学习似乎在能力范围之内，有了这种理解力，就可以得到强大的多用途学习算法了。

但是感知器却碰壁了。知识工程师被罗森布拉特的观点激怒，羡慕其不断增长的被关注度，而神经网络，特别是感知器获得的资金也变得越来越多。马文·明斯基就是其中一个，他是罗森布拉特之前在布朗士科学高中的同学，那时他还是麻省理工学院人工智能研究小组的领导（具有讽刺意味的是，他的博士论文和神经网络相关，但对于神经网络，他越来越不抱幻想）。1969年，明斯基和他的同事西摩尔·派普特一起出版了《感知器》（*Perceptroms*）一书，该书详细介绍了同名算法的缺点，还一一列举了该算法无法学习的内容。最简单的一个（所以也是最受批评的一个）就是排斥—“或”功能（exclusive－OR function，XOR）。如果它其中的一个输入量是对的，那么这就是对的，但如果两个都是对的，则是错的。例如，耐克的两大忠实消费人群被认为是年轻男性和中年妇女。换句话说，如果你是年轻的XOR女性，那么有可能购买耐克的鞋子。年轻挺好的，做女人也挺好的，但两个加起来就不一定好了。如果你既不年轻，也不是女性，那么你也不是耐克做广告的对象。XOR的问题就在于，没有能够区分正面例子和负面例子的直线。图4－3展示了两种不可行的备选方法：

既然感知器只能学习线性界线，那么它就无法对XOR进行学习。而如果感知器无法做到这一点，就无法很好地模拟大脑学习的方法，也不是终极算法可行的备选项。

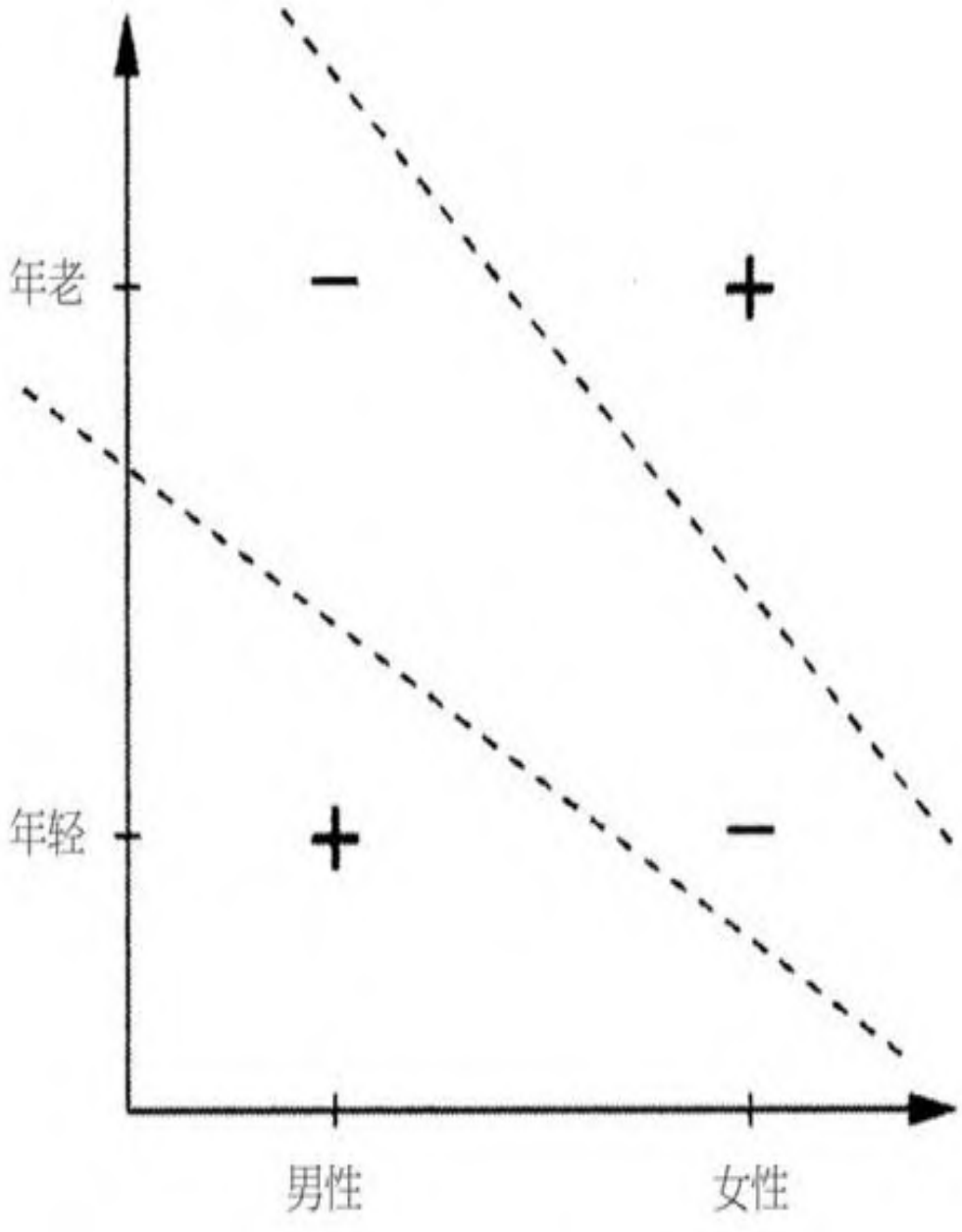


图4－3

然而，一台感知器只能模拟单个神经元的学习过程，尽管明斯基和派普特承认，几层相互连接的神经元能做的应该更多，但他们找不到研究这些神经元的方法，其他人也找不到。问题就在于，没有明确的方法来改变“隐层”中神经元的权值，以减少输出层中神经元造成的误差。每个隐藏的神经元会通过多种路线来影响输出量，而且造成每个误差的原因也可能达到上千种。你该责怪谁？或者，相反，谁该因为正确的输入量而受到赞扬？这个问题在我们努力对一个复杂模型进行学习时随时会出现，而这个问题也是机器学习领域中的中心问题。从数学角度看，《感知器》无懈可击，思路清晰、效果明显。机器学习当时主要和神经网络相关联，而且多数研究者（更不用说创建人）得出结论，构建智能体系的唯一方法就是对其进行明确设定。在未来15年，知识工程师会站在舞台中央，而机器学习则会变成历史的残灰。

物理学家用玻璃制作大脑

如果机器学习的历史是一部好莱坞电影，电影中的反面人物就是马文·明斯基。他是给白雪公主毒苹果的邪恶皇后，让白雪公主处于昏迷状态（在1988年的一篇文章中，西摩尔·派普特甚至半开玩笑地将自己比作故事里的猎人，故事里皇后派他去杀死森林中的白雪公主）。那么白马王子就是名叫约翰·霍普菲尔德的加州理工学院的物理学家。1982年，霍普菲尔德发现了大脑和自旋玻璃（spinglass）惊人的相似之处，自旋玻璃是深受统计物理学家喜爱的特殊材料。这引起联结学派的复兴，在第一个解决赞誉分配问题的算法发明出来的几年，联结学派复兴达到顶峰，并进入一个新时代，机器学习代替知识工程学成为人工智能领域的主导范式。

旋转玻璃其实并不是玻璃，虽然有一些玻璃的属性，其实是磁性材料。每个电子都是一块微小的磁铁，由于本身的自旋运动，可以指“上”或指“下”。比如在铁这样的材料中，电子自旋就趋向于往上：如果一个自旋向下的电子被多个自旋向上的电子包围，这个电子可能会翻转向上。如果一块铁中的大部分自旋都向上，那么这块铁就会变成一块磁铁。在普通磁铁中，每对相邻电子自旋的交互力都一样，但在自旋玻璃中，这种力就可能不一样。这种力甚至会是相反的，使得附近的电子自旋指向相反的方向。当普通磁铁所有的自旋都排成一行时，能量是最低的，但在自旋玻璃中却没那么简单。的确，找到自旋玻璃的最低能量状态就是一个NP—完全问题，意味着几乎所有其他最优化难题都可以简化为NP—完全问题。因为这个，自旋玻璃没有必要适应其整体能力最低的状态。这很像雨水可能会沿着山坡流入湖中，而不是进入大海，自旋玻璃可能会陷入局部最小值的困境，而不是在全局最小值中得到发展。处于最小值状态中的能量会比在其他状态下低，通过翻转一圈，最低能量状态就可以转变为其他状态。

霍普菲尔德注意到自旋玻璃和神经网络之间有趣的相似点：一个电子的自旋对其相邻电子的活动所做的反应和一个神经元的反应十分相似。在电子的情况中，如果相邻电子的加权和超过界限值，电子就会向上翻，反之则向下翻。受到这一点的启发，他确定了一种神经网络，和自旋玻璃一样随着时间的推移而演变，他还提出网络的最低值状态就是它的记忆。每个这样的状态都具备原始状态的“吸引盆”（basin of attraction），原始状态就收敛于该盆中，这样这个网络就可以进行模式识别了。例如，如果其中的一个记忆是由数字9形成的黑白像素模式，而网络看到一个扭曲了的9，它会收敛成“理想”的9，然后据此重新识别它。突然间，大量的物理理论能够应用于机器学习中，而随之也涌入大批的统计物理学家，帮助自旋玻璃打破之前就陷入的局部最小值困境。

虽然如此，但自旋玻璃仍然是大脑的一个不现实的模型。对于一个电子来说，自旋相互作用是对称的，而大脑中神经元之间的连接却

不是对称的。霍普菲尔德的模型忽略的另外一个大问题就是，真正的神经元是和统计相关的：它们不会根据其输入量来确定地进行打开或关闭。随着输入量加权和的增加，神经元更有可能放电，但不确定它是否真的会放电。1985年，大卫·艾克利、杰夫·辛顿、特里·索诺斯基把霍普菲尔德网络里的确定性神经元用可能性神经元代替。现在一个神经网络的状态就有了概率分布，高能量状态的概率要比低能量状态的概率低得多。实际上，找到处于特定状态中的网络，这样的概率由著名的热力学中的玻尔兹曼分布得出，因此他们称自己的网络为玻尔兹曼机器。

一台玻尔兹曼机器拥有混合的感官和隐藏神经元（分别类似于视网膜和大脑）。它通过清醒和睡眠两种交替状态进行学习，就像人类一样。清醒时，感官神经元根据数据指令放电，隐藏神经元根据网络的动态和感官输入来逐步发展。例如，如果网络收到9的图片，与图片中黑色像素对应的神经元会留下，其他的则离开，而隐藏神经元则在给定那些像素值的情况下，根据玻尔兹曼分布随机放电。睡眠状态时，机器会做梦，让感官和隐藏神经元都能自由漂移。新一天的黎明到来之前，机器会统计睡梦中的状态，以及昨天活动中的状态，直到进行比较，接着改变连接权值，让权值符合搭配。如果两个神经元白天时易于在一起放电，但睡眠时放电次数却变少了，那么它们连接的权值会升高；如果情况相反，则权值会降低。通过日复一日重复这样的工作，感官神经之间经过预测的相关性会进一步发展，知道它们和真正的神经元搭配起来。这时，玻尔兹曼机器就掌握了一个很好的数据模型，并有效解决赞誉分布问题。

杰夫·辛顿在接下来的几十年继续在玻尔兹曼机器中尝试了许多变量。辛顿由心理学家变成计算机科学家，他是逻辑运算（被应用于所有计算机中）的发明者——乔治·布尔的曾孙，是世界领先的联结主义者。为了了解大脑如何运转，他比任何人付出的时间、精力都要多。他讲到有一天他怀着极度兴奋的心情下班回家，惊呼：“我做到了！我知道大脑怎么运转了！”他的女儿回答道：“啊！爸爸，你怎

么又来了！”辛顿最近的热情在于研究深度学习，本章的后面部分会谈到。他还参与了反向传播的研究，这是一个比玻尔兹曼机器能更好地解决赞誉分布问题的算法，我们后面会谈到。玻尔兹曼机器原则上可以解决赞誉分布问题，但在实践中，学习这个行为非常缓慢且痛苦，对大多数应用来说，玻尔兹曼机器有点不切实际。下一个突破会涉及解决麦卡洛克和皮茨时期的另外一个过度简化（oversimplication）问题。

世界上最重要的曲线

就其相邻神经元而言，一个神经元只能处于两种状态：放电或不放电。但是这忽略了一个很重要的巧妙之处。动作电位寿命短，电压会在一秒之内骤然升高，然后突然回到静息状态。而单个峰值对接收神经几乎不会有影响，为了唤醒接收神经，需要一连串连续不断的峰值。典型的神经元会偶尔在没有刺激的情况下电压达到峰值，当刺激建立起来时，电压达到峰值的频率会越来越高，然后保持在它所能达到峰值的最快速度，快于这个速度时，不断增强的刺激就没有效果了。神经元与其说是一道逻辑门，不如说是一台电压频率转换器。随电压而变化的频率曲线如图4－4所示。

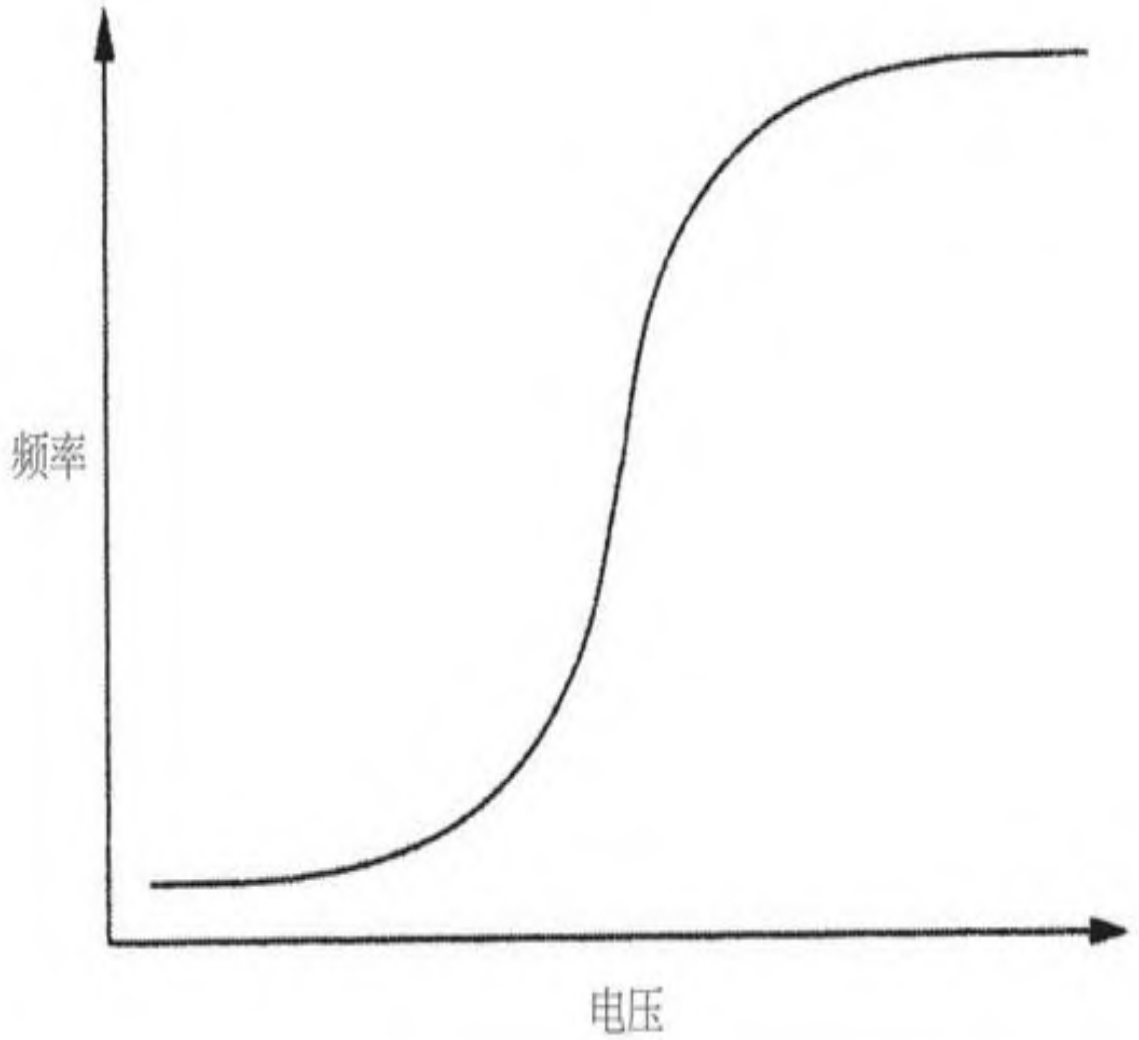


图4－4

该曲线看起来像被拉长的字母S，它有很多叫法，比如逻辑函数、S形函数和S形曲线。仔细研究它吧，因为这是世界上最重要的曲线。首先输出量随着输入量缓慢增长，如此缓慢，似乎保持不变。接着它开始变化得很快，然后变得更快，之后越来越慢，直到几乎保持不变为止。晶体管的转换曲线，将其输入电压和输出电压联系在一起，也是一条S形曲线。所以计算机和大脑都充满了S形曲线。这还没结束，S形曲线是所有种类相变的形状：电子应用领域自旋反转的概率、铁的

磁化、将少量记忆写到硬盘上、细胞中离子通道的打开、冰块融化、水蒸发、早期宇宙的膨胀扩张、进化中的间断平衡、科学中的范式转移、新技术的传播、离开多民族社区的白人大迁徙、谣言、流行病、革命、帝国的没落等。“引爆点”也很适合（可能不那么有吸引力）“S形曲线”这个名字。在两个相邻板块的相对位置中，地震就是一个相变。夜里的碰撞声，也只是你隔壁房间里微观板块移动的声音，所以别害怕。约瑟夫·熊彼特说过，经济是在裂缝和飞跃中得以发展的：S形曲线就是创造性破坏的形状。经济收益和损失对你的幸福度的影响遵循S形曲线原则，所以不要在大事上感到苦恼。随机的逻辑公式可满足的概率（典型的NP完全问题）随着公式变长，会经历从接近1到接近0的相变。统计物理学家花了一辈子时间来研究相变。

在海明威的《太阳照常升起》中，当麦克·坎贝尔被问到他是如何走向破产时，他答道：“有两种方式，先是慢慢地，然后突然破产。”雷曼兄弟的情况也十分相似。这就是S形曲线的精华。未来学家保罗·萨夫预言的规则就是：寻找S形曲线。当你没法调好淋浴的温度时（开始水很冷，然后很快又变得很热），都是S形曲线的错。你做爆米花时，看看S形曲线的进度：一开始什么也没发生，几粒玉米爆开，又有一把爆开，很多玉米突然像烟花一样爆开，更多的玉米爆开，最后你就可以吃爆米花了。你肌肉的每个动作都遵循S形曲线：先是缓慢移动，然后快速移动，最后又缓慢移动。当迪士尼的动画师指出这一点时，动画片变得更自然了，然后人们纷纷模仿。你的眼睛沿着S形曲线移动，注视一样东西然后换另一样，你的意识也跟你的眼睛一起转移。情绪波动也属于相变。出生、长大、坠入爱河、结婚、怀孕、工作、失业、搬到新的城市、升职、退休、死亡，这些都属于相变。宇宙就是相变的巨大集合体，从宇宙到微观世界，从世俗到人生的改变。

S形曲线作为一个独立的模型，不仅很重要，它还是数学的万事通。如果放大它的中段部位，你会发现它近似一条直线。很多我们认为是线性的现象，其实都是S形曲线，因为没有什么能够毫无限制地增

长下去。因为相对性和反牛顿定律，加速度并不会随着力的施加而呈线性增长，但会遵循S形曲线，以0为中心。电路中或者灯泡中电阻的电流也不会随着电压的增长而线性增长（直到灯泡中的灯丝熔化，这本身又是另外一个相变）。如果你把S形曲线缩小，它会近似一个阶跃函数，输出值会突然从0界限值变到1界限值。那么根据输入电压，同样的曲线也会表示这些装置中电阻的工作原理，包括数字计算机和类似的装置，如扩音器和广播协调器。S形曲线的开始部分是有效指数，在饱和点附近它则接近指数式衰减。当有人讨论指数式增长时，问问你自己：它什么时候会变成一条S形曲线？人口爆炸什么时候才会慢慢消失，摩尔定律的重要性什么时候削减，或者说技术奇异点什么时候才不会发生？辨别一条S形曲线，你就会得到一条钟形曲线：缓慢、快速、缓慢变低、高、低。在S形曲线加入一连串向上和向下交错的曲线，你会得到接近正弦波的曲线。实际上，每个函数都可以近似看作S形曲线的总和：函数上升时，你加一条S形曲线；函数下降时，你减掉一条S形曲线。孩子的学习也不是一直都处于进步状态，这个过程是若干个S形曲线的累积。技术变革也是如此。斜眼看纽约的天空，你会看到一组S形曲线在地平线上逐渐展开，每条曲线都和摩天大楼的角一样尖。

对我们来说最重要的是，S形曲线会找到解决赞誉分布问题的新方法。如果宇宙是相变的大集合体，那让我们用一条S形曲线来模仿这个集合体。这就是大脑要做的事：将里面的相变系统调整到外面。那么让我们用S形曲线来代替感知器的阶跃函数，然后看看会发生什么。

攀登超空间里的高峰

在感知器算法中，误差信号要么是全有，要么是全无：你不是收到对的信号，就是收到错的信号。继续下去不会有很大的意义，尤其是当你有众多神经元网络时。你可能会知道，输出神经元出错了（“哎呀，那不是你的祖母”），但如果是大脑深处的某个神经元会怎么样呢？这样的神经元都会出错，那意味着什么呢？如果神经元的输出信息是持续的，而不是二进制的，画面就会改变。首先，我们现在知道输出神经元犯的错误有多严重：它和理想输出信息之间的区别。如果神经元应该不停地放电（“啊，您好，祖母”），但它只放了一点电，那就比什么电也不放要好。更重要的是，我们现在可以将该误差传播到隐藏神经中：如果输出神经元应该放更多的电，而A神经元与其相连，那么A放的电越多，我们就越应该加强它们的连接；但是如果A神经元受到神经元B的抑制，那么B神经元就应该少放电，以此类推。根据与其连接的所有神经元的反馈看，每个神经元会决定该多放或者少放多少电。根据这一点以及输入神经的活动，它会加强或者减弱它与它们的联系。我得放更多的电，但是B神经元在阻止我？把它的权值降低。C神经元不断放电，但它和我的连接很弱？那就加强连接。我的“消费者”神经元位于网络的下游，会告诉我下一轮我的表现如何。

无论什么时候学习算法的“视网膜”看到一张新的图片，这个信号就会在网络中传播，知道它产生输出信息。将这条输出信息和理想的输出信息相比，会发出错误信号，这个信号会穿过神经元层，然后反向传播回去，直到它到达视网膜。根据返回来的信号以及在前进过程中它接收到的输入信息，每个神经元会调整各自的权值。随着网络看到越来越多你祖母和其他人的照片，权值会逐渐收敛到能够让它区分两者的值中。反向传播，正如这个算法为人们所知的一样，比感知器算法要强大很多。单个神经元只能够对直线进行学习。给定足够的隐藏神经，一台多层感知器，正如它的名字一样，可以代表任意的复杂边界。这使得反向传播成为联结学派的主算法。

反向传播是自然及技术领域中非常常见的战略实例：如果你着急爬到山顶，那你就得爬能找到的最陡的坡。这在技术上的术语为“梯度上升”（如果你想爬到山顶）或者梯度下降（如果你想走到山谷）。细菌就是通过游向食物（例如葡萄糖）分子浓度高的地方来觅食的；遇到有毒物质，它们则会游向有毒物质浓度低的地方。所有事物，从机翼到天线阵，都可以通过梯度上升来优化。反向传播就是在多层感知器中有效做到这一点的方法：不断对权值进行微调，以降低误差，然后当所有调整失败时，停止调整。有了反向传播，你就不必从头开始弄明白怎样对每个神经元的权值进行微调，这样做起来会很慢。你可以一层一层来做，根据调整与其相连神经元的方法，来调整每个神经元。如果在突发事件中，除了一件工具，你得把整个机器学习工具包都扔掉，那么梯度下降可能是你想留下的工具。

那么反向传播解决机器学习的问题了吗？我们能否把一大堆神经元扔到一起，让它们自己发挥魔力，然后在去银行的路上，顺便领了诺贝尔奖，奖励你弄明白大脑如何运转？唉，人生不会那么简单。假设你的网络只有一个权值，而图4－5中的误差就是权值的一个函数。

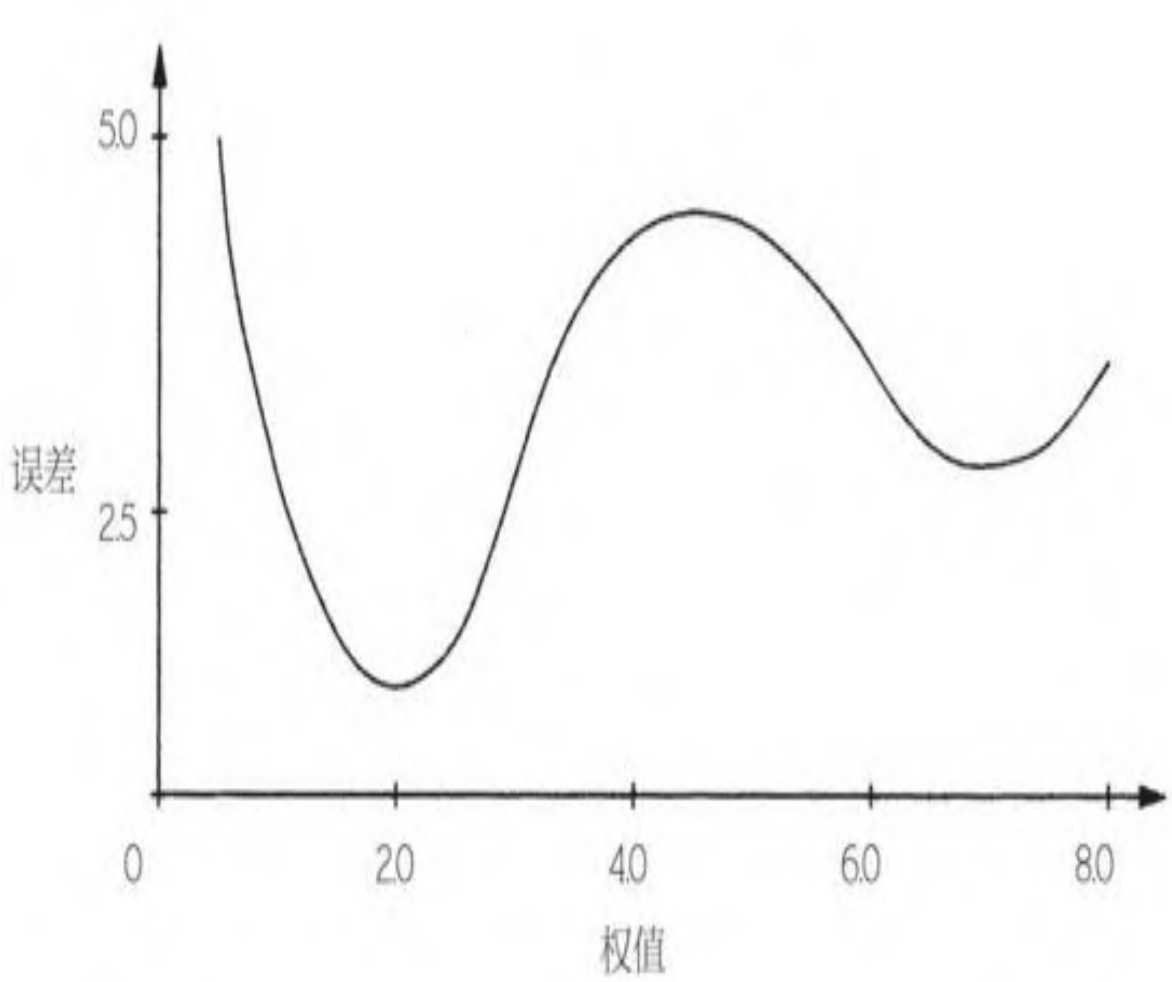


图4－5

最优权值是2.0，这时的误差值最低。如果网络以权值0.75作为开始，例如，反向传播会在几个步骤之后达到最佳效果，就像球滚下坡一样。但如果它以权值5.5作为开头，反向传播会下滑至7.0，而且会停在那里。反向传播，随着渐进权值的变化，不知道该如何找到全局误差最小值，而局部误差值可能会很糟糕，就像把你的祖母错认成帽子那样。有了一个权值，你可以以0.01的增量来尝试每个可能的值，然后用那种方法找到最佳效果。但如果你有数千个权值（数百万或者数十亿个就更不用说了），就不该选这种方法，因为网格线上的点会随着权值数量的增长呈指数级上涨。全局最低值隐藏在浩瀚的超空间的某个角落里，找到它就走运了。

想象一下你被绑架了，然后在喜马拉雅山脉的某个地方被蒙住眼睛。你感到头痛，也不记得什么了。你知道的一切就是你得爬到珠穆朗玛峰峰顶。你会做什么？你向前走了几步，差点滑到山沟里。你喘了几口气之后，决定要更有计划地行动。你用脚仔细感觉了周围的地形，直到找到最高的点，然后小心翼翼走向那点。然后你重复同样的动作。一步一步，你走得越来越高。这就是梯度上升。如果这个山脉刚好就是珠穆朗玛峰，而珠穆朗玛峰是一个标准的圆锥，那它就很有吸引力了。但更有可能的是，当你到达一个地方，那里每走一步都是往下时，那么你距离峰顶还很远。这时你只是站在山脚的某个地方，而且你被困住了。这就是反向传播遇到的事情，除非它在超空间而不是3D空间中攀登高峰。如果你的网络有单个神经元，只要一步一步爬向更好的权值，你就可以到达峰顶。但在多层感知器中，山路会很崎岖，找到最高的山峰已算走运。

这也是明斯基、派普特和其他人无法找到学习多层感知器的部分原因。他们可能想象到用S形曲线来代替阶跃函数，然后进行梯度下降，不过他们会遇到误差的局部最小值这个问题。在那个年代，研究者们不相信计算机模拟，他们需要数学来证明算法可以起作用，而反向传播却没有这样的证据。但我们最好意识到的是，多数情况下局部最小值挺好的，不会有不良影响。误差表面看起来往往像豪猪的刚毛，有很多高峰和低谷，但如果我们找到绝对最低的点之后，这真的不重要了，并且任何人都有可能找到。更好的消息是，实际上，局部最小值可能更适合，因为它和全局最小值比，不太可能证明对我们的数据过拟合。

超空间是一把双刃剑：一方面，空间的维度越高，它越有可能存在高度复杂的表面和局部最优解；另一方面，被困在局部最优解中，意味着被困在每个维度中，所以被困在多维中的难度会比被困在三维中的难度大。在超空间中，到处都有穿山而过的通道。那么，借助夏尔巴人的小小帮助，反向传播往往会找到通往完美权值的道路。它可能仅仅是香格里拉的神秘山谷，而不是海洋，但如果在超空间中，有

数百万个这样的山谷，每个山谷都有数十亿的山口对着它，那为什么要抱怨呢？

然而要当心，不要过于重视反向传播找到权值的意义。记住，可能会有许多不同的优良权值。多层感官器中的学习活动是一个混乱的过程，其意义在于从稍有不同的地方开始，就会让你结束时的解决方法迥然不同。现象本身并无区别，无论这轻微的差别是存在于最初权值中，还是存在于训练数据中，而且是否会在所有强大的学习算法（不只是反向传播）中表现出来。

我们可以去掉S形曲线来解决局部最优问题，然后只让每个神经元输出其输入量的加权和。这样做会使误差变得非常平缓，只剩下一个最小值——全局最小值。虽然如此，问题是线性函数的一个线性函数也只是一个线性函数，那么线性神经元的一个网络与单个神经元几乎一样。一个线性大脑，无论多大，总比一条线虫要笨。S形曲线是线性函数非智能性和阶跃函数难解性的完美中转站。

感知器的复仇

反向传播于1986年是由加州大学圣迭戈分校的心理学家大卫·鲁梅尔哈特在杰夫·辛顿以及罗纳德·威廉斯的协助下发明出来的。他们表明反向传播还可以学习XOR，这让联结学派对明斯基和派普特嗤之以鼻。回顾之前耐克的例子：年轻男性和中年妇女最有可能购买耐克鞋。其中，我们可以用三个神经元的网络来代表：第一个神经元看到年轻男性时会放电；第二个神经元在看到中年女性时会放电；第三个神经元看到其中一个时会放电。通过反向传播，我们可以掌握适当权值，最后成功完成耐克的前景预测（马文，事情就是这样，没什么可说的）。

早期在证明反向传播力量的时候，特里·索诺斯基和查尔斯·罗森伯格训练了一台多层感知器，让其大声朗读。他们的网络聊天系统浏览文本，根据背景选择准确的音素，然后加入语音合成器。他们的网络聊天不仅适用于新词——这是基于知识的系统无法做到的——网络聊天系统还可以说话，而且说得和人类极为相似。索诺斯基过去经常在研究会议上通过播放网络聊天进展的磁带来对观众进行催眠：一开始是牙牙学语，然后开始能让人们听懂，最后说得很流利，只是偶尔会犯错。

神经网络第一次取得大的成功是准确预测股票市场。因为神经网络可以在噪声很大的数据中探测到微小非线性，它们打败线性模型，并在金融领域流行起来，它们的应用也变得更广泛。典型的投资基金会为每大批股票训练出一个分离的网络，让神经网络挑出最有希望的选项，然后让人类分析师来决定该投资哪些股票。然而，有几只基金顺利通过挑选，学习算法会自行买卖这些基金。这些过程到底如何进行的仍是个秘密，但机器学习算法一直处于不为人所知的状态，然后以惊人的速度进入对冲基金，这也许并非偶然。

非线性模型不仅在股票市场有重要作用，在其他很多领域中也是如此。科学家们到处使用线性回归，因为那就是他们知道的东西，但是通常他们研究的现象又都是非线性的，一台多层感知器就可以模拟这些现象。线性模型对相变不了解，神经网络像海绵一样把它们吸收了。

网络神经早期的另外一个成功之处在于掌握驾车技能。无人驾驶汽车在2004—2005年的DARPA大挑战中才进入公众视野，但早在10年前，卡内基－梅隆大学的研究员就已经成功训练出一台多层感知器来驾驶汽车，方法就是探测视频图像中的路面，然后适当转动方向盘。卡内基－梅隆大学的无人驾驶汽车可以沿着海岸跨越美国，而其视力却非常模糊（30×30像素），其大脑比虫子的大脑还要小，但仅仅通过人类副驾驶员的几次协助就完成了此次任务（该计划被称为“横穿

美国的实验”）。这可能不是第一辆真正意义上的无人驾驶汽车，但和一些未成年司机相比，它更靠谱。

如今反向传播的应用太多了，数不过来。随着它为越来越多的人知道，关于它的更多历史也被挖掘出来。原来，反向传播不止一次被创造出来，在科学领域这是常有的事。来自法国的伊恩·勒坤（Yann LeCun）和其他人偶然发现了它，和鲁梅尔哈特发现的时间差不多。20世纪80年代，有一篇关于反向传播的论文被领先人工智能会议否定了，因为根据评论者的观点，明斯基和派普特已经证实感知器无法起作用。实际上，鲁梅尔哈特已经通过哥伦布测试获得反向传播发明者的荣誉：哥伦布不是第一个发现美洲大陆的人，但他却是最后一个。原来，哈佛大学研究生保罗·韦伯斯已经在1974年的博士论文中提出类似的算法。非常讽刺的是，亚瑟·布莱森和何毓琦（两位控制理论家）早就已经做了同样的事情（1969年明斯基和派普特出版《感知器》的同一年）。的确，机器学习本身的历史表明我们为什么需要学习算法。如果能自行在科学文献中找到相关论文的算法在1969年已经出现，它们就可以避免几十年时间的浪费，并加快传播“谁知道有什么发现”这样的信息。

在感知器历史众多有讽刺意味的事件中，也许最让人悲伤的一件事就是，弗兰克·罗森布拉特于1969年在切萨皮克湾的一起轮船事故中离世，没有活着看到他创造的东西得到进一步应用。

一个完整的细胞模型

活细胞就是非线性系统的典型例子。细胞发挥其所有功能的方式，就是通过化学反应的复杂网络，将原材料转化为最终产物。我们利用符号学派的方法，比如逆向演绎，可以找到这个网络的结构，正

如我们在第四章看到的那样。但是为了构建一个完整的细胞模型，我们需要得到定量因素，掌握结合了不同基因表达水平的参数，还需要将环境变量与内部变量结合起来……做到这些很难，因为这些数量之间并没有简单的线性关系。细胞通过对反馈回路进行互锁来维持稳定，这会引起非常复杂的活动。反向传播很适合解决这个问题，因为它能够有效掌握非线性函数。如果我们有完整的细胞代谢路径图，以及对所有相关变量足够的观察，原则上，反向传播可以掌握细胞的详细模型，利用多层感官器根据直接起因预测每个变量。

然而，对于可预见的未来，我们仅仅掌握细胞代谢网络的部分知识，也只掌握了我们想掌握的一部分变量。面对信息缺失以及所有可用信息不可避免的矛盾，要掌握有用的模型就需要利用贝叶斯方法，在第六章我们会深入研究这个问题。对特殊患者进行预测也是同样的道理，掌握模型：可利用的证据必然会杂乱和不完整，但贝叶斯推理会充分利用它。它起到作用了，如果把治愈癌症作为目标，我们不必知道癌细胞运转的所有细节，只需要知道如何在不损害正常细胞的情况下，使癌细胞失去繁殖能力，这就足够了。在第六章中，我们也会看到如何避开我们不知道也不必知道的事，使学习面向目标。

更直接的是，我们可以利用逆向演绎从数据及之前的知识中推导出细胞网络的结构，但应用它的方法可以呈现出组合爆炸式增长，我们需要的是一个策略。既然代谢网络由进化来决定，也许在我们的学习算法中对其进行刺激才是可行办法。在第五章中，我们会看到如何做到这一点。

大脑的更深处

在反向传播初次进入公众视线时，联结学派幻想能够快速掌握越来越大规模的网络，直到硬件允许的条件下，这些网络等同于人工大脑。结果却并非如此。掌握拥有一个隐含层的网络没问题，但在那之后，很快事情就会变得很困难。几层的网络，只有为了应用（比如文字识别）而经过精心设计的才能起作用。超出这个范围，反向传播就会瘫痪。随着我们增加越来越多的层数，误差信号会变得越来越散漫，就像河流分成越来越小的支流，直到我们只剩下雨滴，不留痕迹。利用几十或数百个隐藏层来进行学习，就像大脑一样，过去是一个遥远的梦，而到了20世纪90年代中期，对多层感知器的热情已经逐渐消减。联结学派的中坚分子仍在坚持，但总的来说，人们对机器学习领域的关注度已经转移到别的领域（在第六章和第七章中我们会探索这些领域）。

然而，如今联结学派又复活了。我们比现在掌握了更深层的网络，而且这些网络在视觉、语音识别、药物研制和其他领域都在设定新标准。这个领域新的深入学习成果被刊登在《纽约时报》的头版上。往发动机盖下面看……惊喜：是值得信赖的老式反向传播发动机，还在嗡嗡作响。什么发生改变了？评论家说，没什么大的改变：只是计算机变得更快了，数据变得更大了。辛顿和其他人答道：确实，我们一直都没错！

实际上，联结学派有实质性的进步。如果联结学派是过山车，那么对于最近的过山车转弯，贡献者之一的就是一个看上去普通的小设备，称为“自动编码器”。自动编码器就是一台多层感知器，其输出量和输入量一样。输入一张你祖母的照片，然后输出的还是你祖母的照片。起初，这就像一个愚蠢的想法：这样一个奇异的装置可能有什么用？答案就是，让隐藏层比输入和输出层小很多，那么网络就不会只把输入信息复制到隐藏层，然后再从隐藏层复制到输出信息那里，在这种情况下，我们可能也会把所有东西都扔出去。但如果隐藏层变小了，就会发生有趣的事情：网络会被迫将输入信息编码在更少的比特里，那样在隐藏层就可以表示出来了，然后网络会解压那些信息，

还原到原来的大小。例如，它可以对你祖母百万像素的照片进行编码，变成仅仅7个字母的单词“grandma”，或者它自己创造出来的更短代码，同时学习将“grandma”解码成一个老奶奶的照片。因此，一台自动编码器就像一个文件压缩工具，它有两个重要的优点：知道如何自行压缩东西，和霍普菲尔德网络一样；可以把一张杂乱、扭曲的图片变得干净清晰。

自动编码器于20世纪80年代为人所知，虽然有一个单个隐藏层，但它很难学习。要弄明白怎样把众多信息打包压缩至几个相同比特大小，是非常难的问题（一个代码是你的祖母，另一个稍有不同的代码是你的祖父，再一个是詹妮弗·安妮斯顿，等等）。超空间里的地形实在是太崎岖，让人无法到达峰顶；隐藏的单位要掌握输入量的过多XOR相当于什么。因此自动编码器并没有真正跟上步伐。花了十余年来寻找的诀窍，就是让隐藏层比输入和输出层大一些。啊？实际上，这只是一半的诀窍：另一半就是，在任何特定时间，只把几个隐藏的单位赶走。这样做仍然会阻止隐藏层对输入层进行复制，而且关键是，这样做把学习变得更简单了。如果我们允许用不同比特来代表不同输入层，那么输入层就不再需要争相设置相同的比特。同样，网络现在有比之前多很多的参数，那么你所处的超空间会有更多的维度，而你也有更多的方法来逃出局部最大值的困境。这就叫作稀疏自动编码器，而它是一个诀窍。

虽然如此，我们还没有看到任何深入的学习行为。下一个聪明的主意就是把稀疏自动编码器逐个堆积起来，就像一个多层三明治那样。第一个自动编码器的隐藏层变成第二个的输入或输出层，依此类推。因为神经元是非线性的，每个隐藏层会掌握输入层更为复杂的表达方式，在前一个隐藏层的基础上进行构建。给定大批的面部图片，第一个自动编码器会对局部特征，如棱角和斑点进行编码；第二个自动编码器利用这些信息来对诸如鼻尖、眼睛的虹膜这些面部特征进行编码；第三个掌握整个鼻子和眼睛的面部特征等。最终，最顶端的一层可以是一台传统的感知器，会通过下一层编码器提供的上层特征来

识别你的祖母，这和只利用单个隐藏层提供的粗糙信息，以及对所有层进行反向传播相比，要简单得多。登上《纽约时报》的谷歌大脑网络是一个由自动编码器和其他材料组成的9层三明治，能够从视频网站YouTube的视频中识别出猫。它包含数十亿个连接，是当时能够进行学习的最大网络。不足为奇的是，吴恩达（该项目的主要负责人之一）也是支持“人类智能可以归结为单个算法”这个思想的主要人物。吴恩达平易近人且很有志向，认为叠加在一起的稀疏自动编码器能够帮助我们更好地解决人工智能问题，比之前的任何方法都行得通。

叠加自动编码器不是唯一的深度学习算法，另外一种以玻尔兹曼机器作为基础，还有一种——卷积神经网络，则把视皮质模型作为基础。尽管取得了很大的成功，然而这些成果仍与大脑相去甚远。谷歌网络可以识别出猫脸的正面；人类可以认出各种姿势的猫，甚至在很难辨认脸部时也能认出。谷歌网络仍非常肤浅，它9层里面只有3层是自动编码器。多层感知器是小脑还算不错的模型，大脑的这部分负责低水平运动控制，但皮质就不一样了。它缺失用于传播误差的逆向连接，但它是学习这个魔法的真正所在。杰夫·霍金斯在《人工智能》一书中，提倡紧紧地把皮质组织形式作为基础，进行算法设计，但到目前为止，没有一个算法能够与今天的深度网络相比。

随着我们对大脑理解的深入，这种情况可能会改变。受到人类基因组计划的启示，神经连接组学正在力求绘制出大脑中的每个突触。欧盟正在进行10亿欧元的投资，用于构建大脑一应俱全的模型。美国的大脑计划也有相似的目标，仅仅2014年的基金就达到1亿美元。虽然如此，符号学派对这种寻找终极算法的方法仍非常怀疑。虽然我们可以在单个突触水平上想象完整的大脑，但是我们需要更好的机器学习算法，来将那些想象的图片变成布线图，仅仅用手来完成是不可能的。更糟糕的是，即使有了大脑的完整构图，我们仍然不知道它在干什么。秀丽线虫的神经系统仅仅由302个神经元组成，而且在1986年被完整绘制出来，但关于它干什么，我们也仅仅了解了一部分。我们需要更高水平的概念，用来搞明白底层细节这个难题，剔除那些针对人

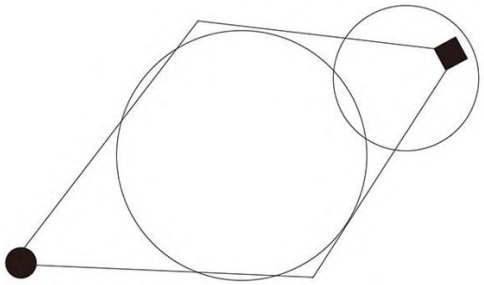
脑或只针对进化怪癖的细节。我们不会通过对羽毛进行逆向工程来制造飞机，而飞机也不会拍翅膀。飞机是在气体力学的基础上设计的，所有飞行的物体都必须遵循气体力学原则。我们还是没有理解那些想法的类似原则。

也许神经连接组学有点过分了。联结学派的一些人高调称，反向传播就是终极算法，而我们只需要扩大反向传播的规模。但符号学派对这种想法不屑一顾，他们指出一长串人类能做但神经网络做不了的事情。按照常识进行推理，这就涉及把之前从来没有被放在一起的信息组合在一起。玛丽午饭吃鞋子吗？不，因为玛丽是人，人类只吃能吃的东西，而鞋子不能吃。符号系统做到这些没有问题——它只会把相关规则串起来，但是多层感知器却做不到这一点。一旦完成学习，它只会不断计算同一个指定函数。神经网络不是组合出来的，而语意合成性是人类认知的一大部分。另外一个大问题就是人类，以及诸如规则集和决策树之类的符号模型，可以对它们的推理进行解释，而神经网络是一大堆没人能看懂的数字。

但是如果人类具备所有这些能力，大脑不经过调整突触就能掌握这些能力，那么这些能力从何而来？除非你相信魔法，答案一定是：通过进化得来。如果你是联结学派怀疑论者，而且有勇气肯定自己的观点，应该弄明白进化如何掌握孩子出生就知道的所有东西——你越觉得是天生的，要求也就越高。如果你弄明白了，然后对计算机进行编程来完成这个任务，别人否认你已经发明了终极算法的至少一个版本，这会显得十分无礼。

第五章

进化学派：自然的学习算法



机器人公园是一个巨大的机器人工厂，被1万平方英里的丛林、城市等包围着。感觉那个丛林就是人类建造过的最高、最厚的墙，布满哨所、探照灯、炮塔。这堵墙有两个作用：把入侵者挡在外面，把公园的“居民”（数百万个为工厂生存而战、对工厂进行管理的机器人）保护在里面。获胜的机器人能够“产卵”，它们完成繁殖的方法就是通过对体内的3D打印机储库进行编程。慢慢地，机器人变得更智能、更快速以及更致命。机器人公园由美军运营，其目标是使终极战士进化。

机器人公园现在还不存在，但有一天它可能就会出现。我建议把它作为几年前DAPRA（美国国防部先进计划研究局）研讨会的一个思维实验，而出席该会议的一位军队高层人士实事求是地说：“那个想法可行。”但如果这支军队为了训练士兵，已经在加州沙漠对阿富汗村落进行全面模拟，模拟中还包括村民，而且为了培养终极战士，几十亿美元也只是个小数目。想到这些，他这么爽快地认可这个想法，就不会觉得多么让人吃惊了。

人们已经开始向机器人公园迈出第一步。在霍德·利普森位于康奈尔大学的创意机器实验室中，奇形怪状的机器人正在学习爬行和飞行，可能正巧它们在做这些时，你也正读到这里。其中一个机器人看起来像胶块垒成的滑行塔，另外一个像装了蜻蜓翅膀的直升机，还有一个能变形的万能工匠。这些机器人并不是人类工程师设计出来的，而是进化而来的，和地球上生命多样性产生的过程一样。虽然机器人最初在计算机的模拟中得以进化，一旦这些机器人的技术达到熟练水平，足以在真实世界中运用，真实版的机器人就会通过3D打印机被制造出来。这些机器人还没有做好掌管世界的准备，但和开始时的“原生汤”模拟物相比，它们已经有了很大进步。

使这些机器人进化的算法，是19世纪由查尔斯·达尔文发明的。那时他不觉得这是一种算法，部分原因在于当时仍缺少一个关键的子程序。一旦1953年詹姆斯·沃森和弗朗西斯·克里克提供了该子程序，进化就会进入第二个阶段：该进化是在计算机中而不是在活体中进行，而且会比活体进化快10亿倍。该子程序的提倡者是一位面色红润、永远面带微笑的中西部人，名叫约翰·霍兰德。

达尔文的算法

和许多其他早期的机器学习研究者一样，霍兰德开始时研究的是神经网络，但他的兴趣使情况发生转变。在密歇根大学读研究生时，他阅读了罗纳德·费雪（Ronald Fisher）的经典著作《自然选择的遗传理论》（*The Genetical: Cal Theory of Natural Selection*）。在该著作中，同时作为现代统计学奠基人的费雪，提出了关于进化的第一套数学理论。虽然这个理论很妙，但霍兰德认为它遗漏了进化论的精华。费雪孤立地看待每个基因，但是有机体的适应度就是它所有

函数的复值函数。如果基因都是独立的，它们变量的相对频率会快速收敛至最大适应点，然后从此保持均衡。但如果基因相互作用，进化（追求最大适应度）就要复杂得多。如果有1000个基因，每个基因有两个变量，那么基因组就有21000种可能的状态，宇宙中没有哪个星球可以大到或者古老到能够一一尝试这些可能。但是在地球上，通过进化可以得到适应力很强的有机体，而达尔文的自然选择理论至少从数量上解释了怎样找到这样的有机体。霍兰德决定将其变成一种算法。

但他首先得毕业。他谨慎地为其论文选择了一个更为保守的主题——布尔周期电路，1959年他获得了计算机科学领域的首个博士学位。即便如此，他读博士期间的导师亚瑟·伯克斯激发了霍兰德在进化计算领域的兴趣，并帮助他在密歇根找到了教师工作，还让他避开那些不把进化计算归入计算机科学的资深同事。伯克斯本人思想很开明，因为他之前就已经是约翰·冯·诺依曼的亲密合作伙伴。冯·诺依曼证明了自我再生机器的可能性。的确，自从1957年冯·诺依曼因为癌症逝世后，完成这个工作的重任就落到了伯克斯肩上。在当时遗传学和计算机科学的原始水平下，冯·诺依曼能够证明这种机器的存在，这已经很了不起。但他的机器人智能仅能复制和自己一模一样的副本，要制造出进化版的机器人，还得由霍兰德来完成。

随着霍兰德的创作渐渐为人所知，遗传算法的关键输入就是一个适应度函数。给定一个待定程序和某个设定的目标，适应度函数会给程序打分，反映它与目标的契合度。在自然选择当中，适应度是否能这样解释，值得怀疑：虽然翅膀对于飞行的适应度很高，这个说法很直观，但整个进化过程却没有已知的目标。即便如此，在机器学习中，掌握诸如适应度函数这样的事情还是很容易的。如果我们需要一个能够诊断疾病的程序，如果一种算法能够正确诊断我们数据库中60%的病人，这样的算法就比准确率仅为55%的算法要好，所以可行的适应度函数能够帮助准确诊断。

就这一点而言，遗传算法就有点像选择育种。达尔文在《物种起源》开篇时就谈到这个问题，层层深入，解释较难理解的自然选择概念。现在我们自然而然地认为，经过驯化的动植物就是一代一代经过选择和交配的结果，是最能让我们达到各种用途的有机体：果实最大的玉米、最甜的果树、毛最多的羊、最强健的马匹。遗传算法也会做同样的事情，它产出的是程序而不是活的生物体，而一代对它而言是几秒的计算机时间，而不是生物的一生。

适应度函数将人在这个过程中扮演的角色概括化了。但和人的角色相比，更为微妙的部分是自然的角色。开始，是一群适应力不那么强的个体（可能是完全随机的个体）遗传算法得找出变量，然后这些变量依据适应度而被选择。自然怎样做到这一点？达尔文不知道。这就是遗传算法的一部分。DNA依据碱基对的序列对有机体进行编码，同理，我们也可以依据一串二进制数字对程序进行编码。DNA的4个基本组成单位，不是0和1，而是4个基本碱基，包括腺嘌呤、胸腺嘧啶、胞嘧啶、鸟嘌呤，但这仅仅是表面上的差异。变量，无论在DNA序列中，还是在位串中，都可通过几种方法产生。最简单的方法就是点突变，即随意翻转位串中的一个比特值，或者改变一段DNA中的单个基本碱基。但对霍兰德来说，遗传算法的真正威力在于更为复杂的东西：性。

把外衣脱光，只剩下基本部位（请不要笑出声），有性生殖包括在父亲和母亲的染色体之间进行材料交换，这个过程称作染色体交叉。这个过程会产生两条新的染色体，一条染色体交叉点的一边是母亲的染色体，另外一边则是父亲的染色体，另外一条则相反（见图5－1）。

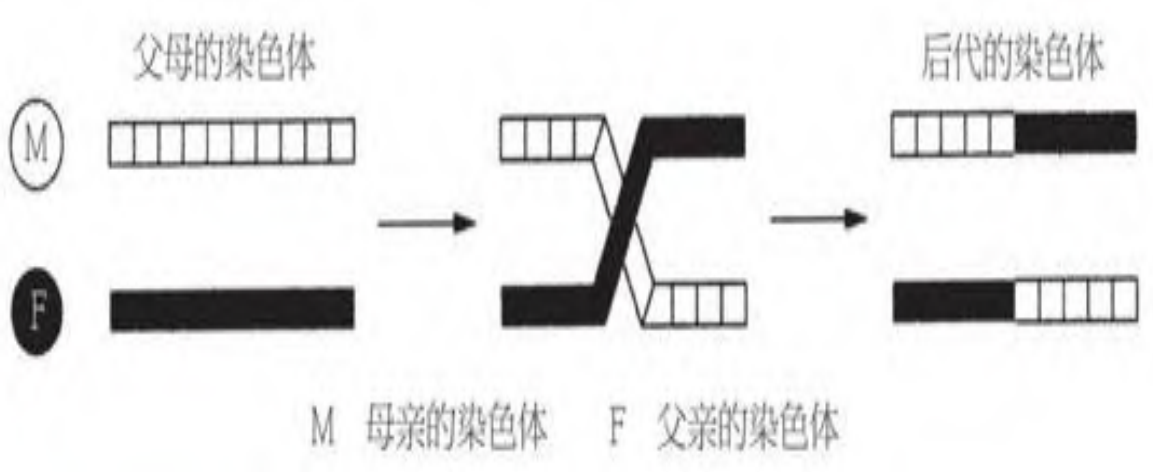


图5－1

遗传算法通过模拟这个过程发挥作用。它为每一代中两个适应力最强的个体进行配对，通过随机交叉父母位串中的一点，让每对父母生出两个后代。将点突变应用到新的位串后，算法让这些点突变在其虚拟世界中释放。每个点突变都会反馈适应度得分，然后重复这个过程。每一代都会比前一代的适应度更高，当达到理想的适应度或者时间用尽时，这个过程就会结束。

例如，假设我们想对一个规则进行进化，用于过滤垃圾邮件，结果会怎么样？如果有1万个不同的词出现在训练数字中，每个待选规则会用一条由2万个比特组成的位串来表示，每两个比特就是一个单词。与单词“free”（免费的）相对应的第一个比特就是1，条件是包含“free”的邮件允许与规则匹配；与单词“free”相对应的第一个比特就是0，条件是包含“free”的邮件不允许与规则匹配。第二个比特则相反：结果就是1，条件是不包含“free”的邮件允许与规则匹配；否则结果就是0，且不允许与规则匹配。那么如果两个比特都是1，邮件不管是否包含“free”都允许与规则匹配，而且规则实际上对那个词不设条件。另外，如果两个比特都是0，那么没有邮件会与规则匹配，因为总有一个比特会发生故障，而且所有邮件都会通过过滤器。总的来说，一封邮件只有在其出现和不出现的单词模式都经过规则允

许时，才能与规则匹配。一个规则的适应度就是规则正确分类邮件的百分数。从一堆随机的字符串开始，每个位串代表一个包含随机条件的规则，通过不断重复和改变每代中的最匹配位串，这时遗传算法就可以进化出越来越好的规则。例如，如果当前包含这些规则：如果邮件包含“free”，那么它就是垃圾邮件，如果邮件包含“easy”（简单的）那么它就是垃圾邮件。将这两个规则交叉可能会产生适应度更高的规则：如果邮件都包含“free”和“easy”，那么它就是垃圾邮件，前提是交叉点不落在与这些词（free、easy）相对应的两个比特中间。它也可能会产生这样的规则：所有的邮件都是垃圾邮件，这样的规则是抛弃所有条件后得出的，但它在下一代中不太可能产生很多子规则。

既然我们的目标是尽可能设计出最好的垃圾邮件过滤器，和老老实实模拟真正的自然选择相反，我们可以大大方方地作弊，方法就是通过改变算法来适应我们的需求。遗传算法能够频繁作弊的方法，就是允许有永不灭亡的东西（现实生活中却不存在，太糟了）。在那种方法中，适应力很强的个体不仅会在它那一代中为了繁殖而竞争，还会跟它的“儿子”“孙子”“重孙”等竞争，只要在群体中还保留有其中一个适应力最强的个体。相反，在现实世界中，适应能力强的个体能做到最好的，也就是将其一半的基因传给许多孩子，而每个孩子的适应能力可能不会那么高，因为遗传了父母另外一方的基因。不朽性避免了这样的倒退，而且如果幸运，能让算法更快地达到理想的适应力。当然，既然历史上最适合生存的人类都是通过其后代的数量来衡量的，比如成吉思汗（当今200个人中就有一个人是他的后代）。现实世界中没有永生这一说，也许也没那么糟糕。

如果我们想对一整套而不只是一个垃圾邮件过滤规则进行进化，我们可以用含有$n\times 20\ 000$个比特的字符串（每个规则有20 000个，和之前一样，假设数据里有1万个不同的词）来代表有n个规则的规则集。有些词会用“00”来表示，那些包含这些词的规则实际上会从规则集中消失，因为这些规则不会和任何邮件相匹配，这和我们之前看

到的一样。如果一封邮件和规则集中的任何规则都匹配，那么它就被分类成垃圾邮件，否则就是合法邮件。我们还可以把适应度当作正确分类邮件的百分数，但为了防止过拟合，我们可能想从这个百分数中扣除与规则集中和活性条件相称的处罚比例。

我们甚至可以做得更好，方法就是允许过渡概念的规则进行进化，然后将这些规则在运行时串起来。例如，我们可以对这样的规则进行进化：如果邮件包含“贷款”一词，那么这是一封诈骗邮件；如果这封邮件是诈骗邮件，那么它就是垃圾邮件。既然一个规则的结果不再总是“垃圾邮件”，这需要引入规则字符串中的额外比特的信息来代表它们的结果。当然，计算机不会按照字面引用“诈骗”这个词，它只会找出任意的比特串来表示这个概念，但这对实现我们的目标已经起到很好的作用。霍兰德称类似这样的规则集为“分类器系统”，是他建立的机器学习部落中的一匹“驮马”：演化新论。和多层感知器一样，分类器系统会面临赞誉分布问题——过渡概念的规则适应度是什么？另外，霍兰德设计出所谓的“桶队算法”（bucket brigade algorithm）来解决这个问题。即便如此，分类系统和分层感知器相比，其应用范围要窄得多。

和费雪梳理的简单模型相比，遗传算法有了很大的进步。达尔文悲叹自己的数学能力不行，但如果他晚生一个世纪，他可能又会希望自己有很强的编程能力。的确，通过一组方程来充分体现自然选择很困难，但将自然选择表达为一种算法又是另外一回事，而且这样还能阐明许多其他棘手的问题。为什么一些物种会突然出现在化石记录中？能够证明这些物种是渐渐由早期物种进化而来的证据在哪里？1972年，尼尔斯·埃尔德雷奇和史蒂芬·杰伊·古尔德提出进化过程由一系列“间断平衡”组成，长期的停滞与短暂的快速变化相互交替，就像寒武纪爆发那样。这个问题引起激烈的讨论，“间断平衡”理论的批评家将其命名为“跳跃式进化”，而埃尔德雷奇和古尔德则反驳称，渐进主义就是“匍匐式进化”。遗传算法的经验会对支持“跳跃式进化”的一方有利。如果你运行10万代遗传算法，然后每隔

1000代观察群体的数量，那么适应度与时间的曲线图可能看起来会像高低错落的楼梯，图形突然上升，然后是随着时间慢慢变长的平台期。要弄明白为什么也不难。一旦算法达到适应度的局部最大值（适应度中的峰值），算法会在这一点停很长时间，直到某次幸运的变异或者交叉，让处于坡上的个体等到更高的峰顶，在这一点上该个体会进行大量繁殖，然后和过往的每一代来爬上这个坡。当前的峰值越高，该过程发生前的那段时间就越长。当然，自然选择比这还要复杂：一个原因就是，环境可能会变化，要么是自然上的改变，要么是因为其他有机体自身进化了。另外，处于峰值的有机体可能会突然发现，对于再次进化，它面临巨大的压力。因此，虽然有用，当前的遗传算法还远远不是故事的结局。

探索：利用困境

我们应注意遗传算法和多层感知器的差异程度。反向传播会在任何给定时间坚持单一假设，而且这个假设会渐渐改变，直到其适应某个局部最优值。遗传算法会在每一步中考虑整个群体的假设，而由于交叉行为，这些假设可以从这一代跨到下一代。将初始权值设为小的随机值后，反向传播才会确定继续进行下去。相反，遗传算法则充满随机选择：该使哪些假设成立并进行交叉（适应度更高的假设更有可能成为备选对象），该在哪里对两个字符串进行交叉，该使哪些比特的信息发生突变。反向传播为了预先设定的网络结构掌握权值；密集度更大的网络更为灵活，但掌握起来也更困难。除了通用式以外，遗传算法不会对它们即将学习的结构进行预先假设。

因为这个原因，和反向传播相比，遗传算法陷入局部最优值困境的可能性较小，而且原则上也更有可能找到真正新颖的东西，但遗传

算法分析起来也要难得多。我们怎么知道，遗传算法会得出有意义的结果，而不是像众所周知的酒鬼那样到处晃悠？问题的关键在于要从构建模块的角度去考虑。字符串比特的每个子集都可能会对有用的构造块进行编码，当我们将两个字符串进行交叉时，这些构造块会聚集到一起，形成一大块构造块，反过来这个大构造块又会变成“磨坊中的谷物”。霍兰德喜欢用警方提供的画像来解释构造块的力量。在计算机还没出现的年代，警局的拼图师能够很快将嫌疑人的画像拼在一起，方法就是在采访中让目击者从画有典型嘴型的一组纸带中挑选嘴巴，然后用同样的方法选出眼睛、鼻子、下巴等。只有10个构造块，每个构造块有10个选择，这个系统会虑及100亿张脸，比地球上的总人口还要多。

在机器学习中，和计算机科学的其他领域一样，没有什么比得到这种对你有利而非有害的组合爆炸更好的东西了。遗传算法的灵活之处就在于，每个字符串都暗含指数数量的构造块，被称为“基模”，因此该研究比它看起来的还要高效得多。这是因为字符串比特的每个子集都是一个基模，代表可能合适的性能组合，而一个字符串有指数数量的子集。我们可以用这样的方法来代表基模，也就是用“*”来代替字符串中不属于该字符串的比特。如110这个字符串中包含以下基模：***、**0、1**、*10、11*、1*0，以及110。每选一个不同的比特，我们就得到一个不同的基模，因为每个比特我们有两个选择（包含或不包含），那么我们就有2^n个基模。相反，在群体中，一个特定的基模可能由许多不同的字符串来表示，而且每当这时，这些基模都会受到隐式评估。假设在下一代中仍成立的概率与其适应度成正比，那会怎样？霍兰德表明，在这种情况下，和平均值相比，在某代中表示基模的字符串适应度越高——我们能期望的——在下一代中看到这些表示字符串的数量也越多。那么，虽然遗传算法暗地里对字符进行操纵，它也会找到基模更大的可能性。随着时间的流逝，适应度更高的基模会主导群体，所以不像醉汉那样，遗传算法能找到回家的路。

机器学习中最重要的问题之一（也是关于生命最重要的问题之一），就是探索—利用困境。如果已经找到能发挥作用的东西，你还会继续找吗？还是做新的尝试会更好，知道这样可能会浪费时间，但也许会找到更好的办法？你想做牛仔还是农民？想自己开一家公司，还是管理一家已经成立的公司？想用情专一，还是三心二意？中年危机意味着把多年的时间花费在利用某种东西之后，会极度渴望探索新事物。一时冲动之下，你飞到拉斯韦加斯，准备赌光你毕生的积蓄，变成百万富翁。你进入赌场，然后面对一排老虎机。你要在能给你平均回报最多的那台机器上玩，但你不知道是哪台。那么你就得每台都试，次数够了才能弄明白。但如果花太多时间来做这个，你就把钱浪费在会输的那台机器上。相反，如果操之过急，选了一台前几轮碰巧让你手气不错的机器，但实际上并不是让你获得回报最多的一台，那么今晚剩下的时间你就会把钱浪费在这台机器上。这就是探索—利用困境。每次你玩的时候，要么你就选择重复目前为止发现的最好的一把，这样就能给你最好的回报，要么就试试其他几把，这样能够让你知道怎样才能赢得更多。利用两台老虎机，霍兰德表明最优策略就是每次抛一个偏币，在这个策略中，随着你进行这些步骤，和原来的硬币相比，这时的硬币会呈指数偏倚（如果这种方法对你不起作用，别指控我。记住最后赢的总是赌场）。一台老虎机看起来越好，你就越应该玩它，但也别完全放弃另外一台，毕竟另外一台也有可能会让你赢得更多。

遗传算法就像一群赌徒中的元凶，同时在市里的每个赌场玩老虎机。如果两个基模包含相同的比特量而且至少有一个比特的值不同，那么这两个基模就会互相竞争，例如*10和*11，而n个互相竞争的基模就像n台老虎机一样。每个互相竞争的基模集都是一个“赌场”，遗传算法会同时弄清楚每个“赌场”中会赢的机器，依据的是最优策略：以成倍增长的频率去玩看起来赢得越来越多的机器。太聪明了！

在《银河系漫游指南》中，外来种族为了回答终极问题，造了一台巨大的超级计算机，过了很长时间，计算机吐出“42”这个数字。

但这台计算机还指出，外星人不知道问题是什么，所以他们又造了一台更大的计算机来弄明白这个问题是什么。这台计算机（也称“行星地球”）很遗憾地在持续了几百万年的计算，即将完成的几分钟前，被摧毁了，因为要为空间高速公路让路。现在我们只能猜那个问题是什么了，可能会是：你打算玩哪台老虎机？

程序的适者生存法则

在开始的几十年，遗传算法的阵营主要由约翰·霍兰德、他的学生、这些学生的学生组成。大约在1983年的时候，遗传算法解决的最大问题就是学会控制天然气管道系统。不过，大概同样的时间段，神经网络回归了，人们对进化计算的兴趣也开始浓厚起来。关于遗传算法的第一次会议于1985年在匹兹堡举行，遗传算法的寒武纪大爆发也正在进行中。其中的一些变体尝试更加接近地模拟进化（毕竟基本遗传算法仅对进化进行了非常粗略的模拟）而其他变体则从不同方向模拟，利用让达尔文困惑的计算机科学概念来跨越进化论观点。

霍兰德的学生中较为出色的是约翰·科扎。1987年，他在意大利参加会议，飞回加利福尼亚时，有一瞬间他突然醒悟了。我们要不要对成熟的计算机程序进行进化，而不是发展相对简单的东西，如“如果……那么……”规则以及天然气管道控制？如果那是目标，为什么还要继续用字符串来表示呢？程序就像是一棵子程序调用树，所以最好还是直接穿过那些子树，而不是将它们硬塞到字符串中，而且当你随机取点穿过这些子树时，会冒着将很好的程序摧毁的风险。

例如，假设你想进化一个程序，根据行星到太阳的距离D来计算该行星的运行周期T。根据开普勒第三定律，T是D的立方的开平方，乘以一个常数C，该常数的值取决于你使用的时间和距离单位。遗传算法应

该像开普勒那样，通过查看第谷·布拉赫关于行星运行的数据，就可以发现这个规律。在科扎的方法中，*D*和*C*是程序树上的叶子，而将其结合起来的运算，比如乘法运算和求平方根运算，就是内部节点。以下程序树能准确计算出*T*（见图5－2）。

科扎称他的方法为“遗传编程”，在这个方法中，我们通过随机交换程序树的两棵子树，来对两棵程序树进行交叉。例如，将这两棵树的高亮节点交叉，就会产生能够计算出*T*的程序，这个程序为其中的一棵子树（见图5－3）。

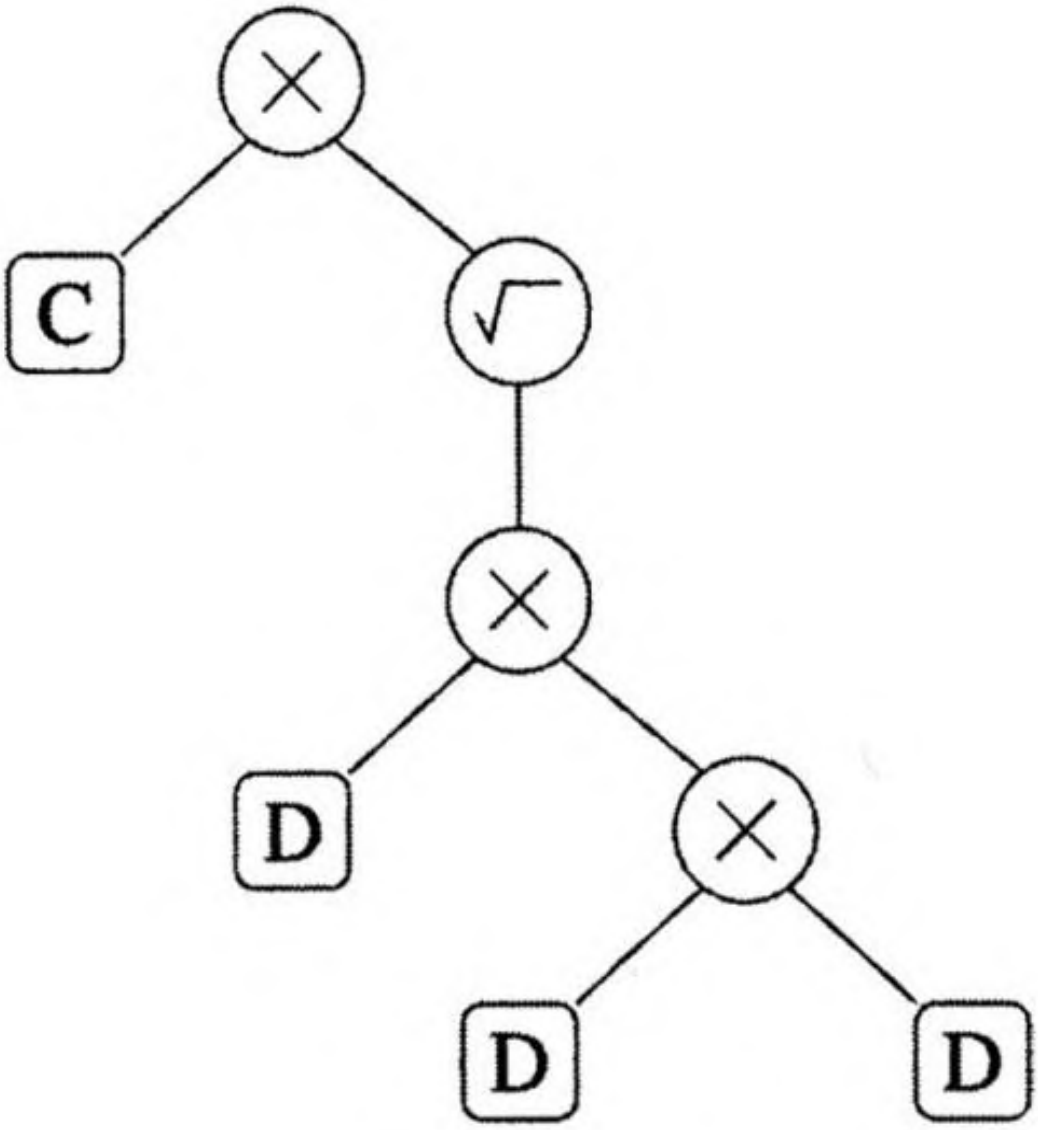


图5－2

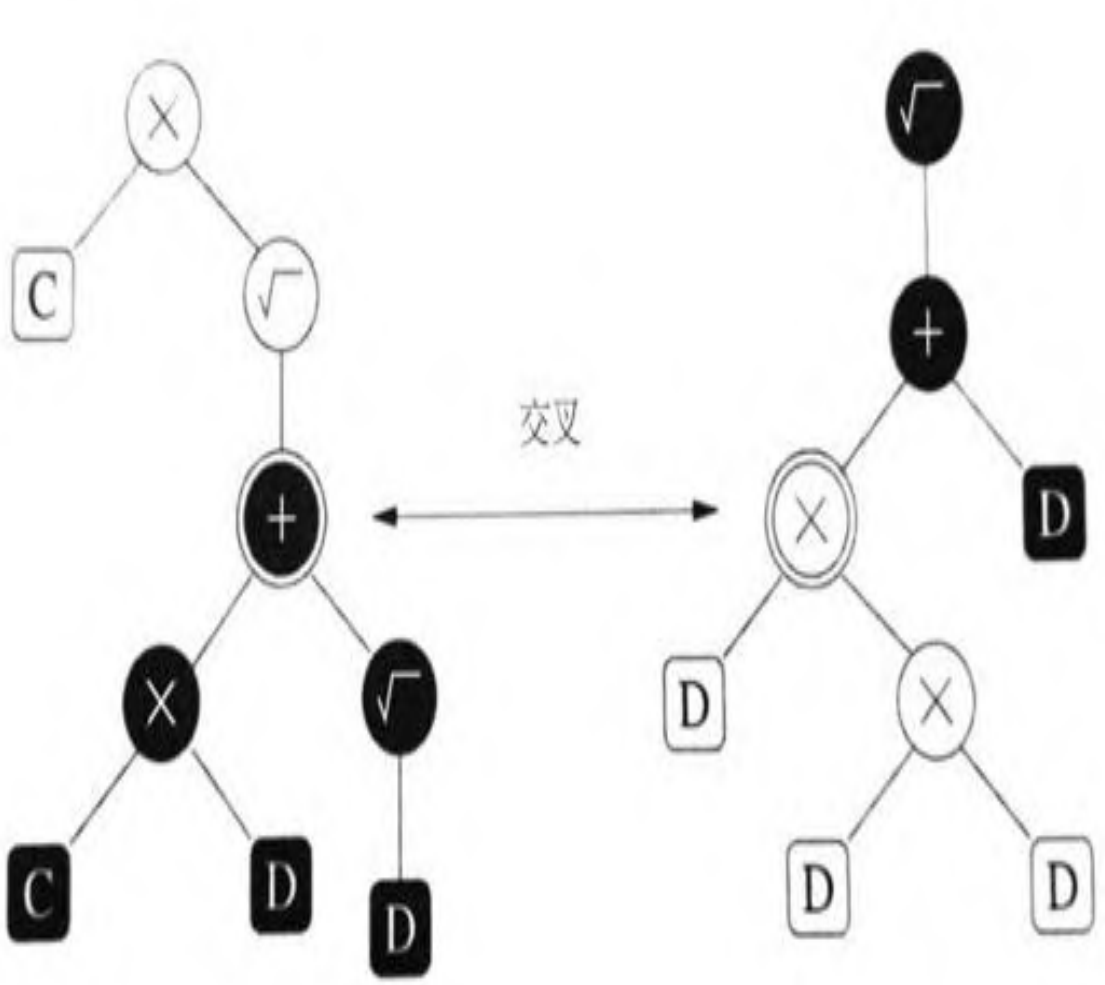


图5－3

我们可以测量程序的适应度（或缺乏适应度），方法就是通过其输出值与训练数据中的准确值之间的差距来判断。例如，如果程序说地球的周期是300天，这就意味着要从其适应度中减掉65。以一群随机程序树作为开头，遗传编程利用交叉、变异、生存来渐渐进化出更好的程序，直到满意为止。

当然，计算行星的周期是一个很简单的问题，仅仅涉及乘法和平方根运算。一般情况下，程序树包含全面的编程结构，例如“如果……那么……”陈述、循环、递归。更能解释“遗传编程用来做什么”的例子就是，弄明白机器人需要按照什么顺序来行动，以达到某

个目标。比如我让办公机器人从走廊的柜子里拿一个订书机。机器人可进行大量的动作，例如，走到走廊中、开门、拿起某个物品等。这些动作反过来可以由各种小动作组成，比如机器人的手向物品移动，或者在各种不同的点抓住物品。每个行为是否可能会被执行，取决于之前行为的结果，也可能需要被重复多次等。这个挑战在于要组合这些动作和子动作的正确结构，以及每个动作的参数，例如，手要移动多长的距离。机器人的“原子性”动作以及每个动作的结合为开端，遗传编程能够组合一个复杂的动作，并达到理想目标。通过这种方法，许多研究人员已经为机器人足球运动员制定了策略。

对程序树而不是字符串进行交叉得出的一个结果就是，生成的程序可以任意大，这让学习活动变得更加灵活。然而，整体的趋势就是会变得膨胀，因为随着进化持续的时间变长（也称为“适者生存”），树也会越长越大。演化新论者可以从这个事实中获取安慰：人类编写的程序之间并没有什么不同（微软的Windows系统：4500万行的代码和计算），而且人造代码也不会允许这样一种简单的方法：将复杂性处罚添加到适应度函数中就可以了。

遗传编程的第一次成功是在1995年，也就是成功设计了电子电路。以一堆电子元件为开端，例如，晶体管、电阻器、电容器，科扎的系统为了一台低通滤波器，彻底改造之前的一个专利设计（这是一个电路图，其用途之一就是加强舞蹈音乐中的低音）。自那以后，他就开始对专利设备进行改造，成打地改造出其他设备。下一个里程碑于2005年到来，当时美国专利及商标局为一项专利颁奖，该专利根据遗传学设计，是工厂的优化系统。如果图灵测试愚弄的是专利审查员而不是谈话高手，那么2005年1月25日，就是一个可载入史册的日子。

科扎的自信在这样一个不以谦逊著称的领域显得尤为突出。他将遗传编程当作一台会发明东西的机器，当作21世纪的硅版爱迪生。他和其他演化新论者相信，它可以掌握一切程序，这让他们进入了终极算法的争夺赛中。2004年，他们创立一年一度的“人类竞争奖”

（Humie Awards），来认可“人类竞赛”相关的遗传编程创作。迄今为止，已经颁发39个奖项。

性有何用

尽管他们取得了成功，也对诸如渐进主义与间断平衡的问题发表见解，但遗传算法还有一个很大的谜团没有解开：性在进化过程中所起的作用。演化新论者非常重视交叉行为，但其他学派的成员认为没有必要如此麻烦。霍兰德没有哪个理论结果表明，交叉行为能起作用。经过一段时间，突变足以成倍地增加群体中最适合留下的基模的频率。而“构造模块”的直觉很吸引人，但很快就会遇到麻烦，即使用上遗传算法也没用。随着模块演变得越来越大，交叉行为也会越来越趋向于将这些模块解散。还有，一旦适应力强的个体出现，其后代很有可能快速掌管该群体，并有可能将更好的基模挤出，这些基模受到整体上不那么相配的个体的牵绊。这有效减少了研究适应度冠军变化的工作量。研究者已经找到许多方案，能够保存群体中的多样性，但研究结果还不确定。工程师们当然广泛应用构造模块，但将这些构造块结合起来就会在很大程度上涉及工程学的内容。这不是用旧办法把它们丢在一起那么简单，也不是明显的交叉行为就能取得成功的。

消除性别对于演化新论者来说，就只剩下变异作为其理论的推动力。如果群体的规模大体上比基因的数量大，很有可能每个点突变都已体现在其中，而研究就变成爬山法的一种：尝试所有可能的单步变种，挑选最好的一个，然后重复这个步骤（或者挑选几个最好的变种，这种情况就被称为“定向搜索”）。应特别指出的是，符号学派一直用这种方法来掌握规则集，虽然他们不把它当成进化的一种形式。为了避免陷入局部最大值的陷阱，爬山法可以通过随机性（以某

个概率做下坡移动）和随机重启（过一会儿后，跳到随机状态，然后从那儿继续）来得到加强。这样做已经足以找到解决问题的好办法。为其添加交叉行为是否能证明额外计算成本的合理性，这有待回答。

没人知道为什么性别在自然界中无处不在。人们已经提出几个理论，但没有一个被广泛接受。这方面的领先理论是“红皇后”假说，马特·里德利在同名书中向人们介绍该理论。在《爱丽丝镜中世界奇遇记》中，红桃皇后对爱丽丝说：“全力奔跑，这样你才能留在原地。”依照这个观点，有机体和寄生虫就会永远处在竞赛中，而性可以保持群体的多样性，这样单一微生物就不会感染整个群体了。如果这就是答案，那么性就和机器学习不相关，这至少持续到掌握的程序因为处理器时间和内存而与计算机病毒进行抗战（有趣的是，丹尼·希利斯称，有意将共同进化的寄生虫引入遗传算法中，可以帮它避免局部最大值困境，方法就是逐步加大难度，但还没有人继续采用他的观点）。赫里斯托斯·帕帕季米特里乌和同事表明，性优化的不是适应度，而是他们所谓的“混合度”：当与其他基因结合时，一个基因表现出平均水平良好的能力。就像自然选择一样，当适应度函数要么不是已知，要么不是常数时，这个说法会起作用。但是在机器学习和最优化中，爬山可能会做得更好。

遗传编程的问题不会就此结束。的确，虽然它获得的成功可能不会像演化新论者希望的那样和遗传有很大联系。以电路设计为例，这就是遗传编程成功的典型。作为一种规则，即使相对简单的设计也需要大量的研究，而且也不知道，得出的结果多大程度上归功于蛮力强攻而不是遗传学智慧。为了回应越来越多的批评者，科扎1992年的书《遗传编程》（*Genetic Programming*）包含的实验表明，遗传编程在布尔电路合成问题上打败了随机生成备选项，但胜利的优势很小。接着，1995年在加利福尼亚太浩湖举办的机器学习国际会议（ICML）上，凯文·朗发表了一篇论文，论文表明爬山法在相同的问题上打败了遗传编程，而且胜利的优势很大。科扎和其他演化新论者已经不断尝试在ICML上发表论文，因为这是该领域的重要会议，但让他们越来

越感到挫败的是，因为经验验证不足，他们不断遭到拒绝。论文被拒已经让科扎感到失落，看到朗的文章，更让科扎感到气急败坏。在短时间内，他以ICML的两栏格式，整理了一篇23页的论文，反驳朗的结论，并控诉ICML评审员的科学不端行为。他接着在会堂的每个位置都放了一份论文复印件。依据你的观点，不是朗的论文，就是科扎的回应成为最后一根稻草。不管怎样，太浩湖事件标志着演化新论者与机器学习阵营中其他流派最后的决裂。遗传编程员开始举办自己的会议，这些会议与遗传算法会议合并，成为“遗传与进化计算会议”（GECCO）。对于演化新论者这部分，机器学习主流学派已经很大程度上把他们忘记。这是一个悲伤的结局，但在历史上并不是第一次，性是决裂的原因。

性在机器学习中可能还未取得成功，但作为一种安慰，它已经在技术的发展中起到重要作用。色情描写是万维网中未公开承认的“杀手级应用”，更别说它在印刷、摄影、视频领域就更受欢迎了。振动器是第一台手持电子设备，预示着一个世纪之后手机的出现。小型摩托车在战后的欧洲，特别是在意大利受到欢迎，因为它们能够让年轻的夫妇离开家人。100万年前，当直立人发现火时，其中的一个“杀手级应用”当然是便于约会。同样肯定的是，“拟人化”机器人变得越来越真实，其主要推动力是性爱机器人行业的发展。性似乎才是最后的结局，而不是技术演化的手段。

先天与后天

演化新论者和联结学派重要的共同点是：他们都因为受到自然启发而设计了学习算法，不过后来分道扬镳了。演化新论者关注的是学习架构，对他们来说，通过参数优化来对演化的架构进行微调，这是

次重要的事情。相反，联结学派更喜欢用一个简单、手工编写的结构，加上许多连接行为，然后让权值学习来完成所有工作。这就是机器学习版本关于先天和后天的争论，而且双方都有很好的论据。

一方面，进化已经产生许多神奇的东西，但没有什么比产生你更神奇的了。有或者没有交叉行为，进化结构都是终极算法的基本部分。大脑可以学习任何东西，但终极算法却无法使大脑进化。如果彻底了解大脑的结构，就可以将其运用到计算机硬件中，但我们还远远做不到这一点。通过计算机模拟进化寻求帮助是一件容易的事情。另外，我们也想对机器人的大脑、装有任意传感器的系统、超级人工智能设备进行进化。如果有更好的东西来完成这些任务，我们就没有理由继续使用根据人类大脑设计出来的设备。另一方面，进化的过程缓慢到让人难以忍受。一个有机体的一生只会产生一条与其基因组相关的信息：它的适应度，反映在后代有机体的数量上。这简直是在浪费大量的信息，神经学习会通过在使用的点上（可以这么说）获取信息来避免这个问题。正如联结学派的杰夫·辛顿指出的那样，基因组携带的信息没有什么优势，因为我们可以从感知中获取这些信息。当新生儿睁开眼睛时，感知世界会涌入大脑，而大脑只需组织这些感知。然而，基因组中需要确定的信息是，完成这些组织工作的“机器”的结构。

在先天与后天之争中，两方都没有完整的答案，关键在于找到如何将两方结合起来。终极算法既不是遗传编程，也不是反向传播，但它得包含这两者的重要部分：结构学习和权值学习。在传统观点看来，先天自然完成第一部分（进化大脑），后天培育再将大脑填满信息。我们可以在学习算法中重复这个过程。首先，学习网络的结构，利用爬山法（举个例子）来决定神经元之间如何连接：试着将每个可能的新连接添加到网络中，保持那个最能提高性能的连接，然后重复这些步骤。之后利用反向传播来学习连接权值，那么你全新的大脑马上就可以使用了。

如今在自然和人工进化中，都存在一个很重要的微妙之处。我们会为每个备选的结构而不仅仅是最终的那个，而一直学习权值，目的是为了明白这些结构在生存竞争（在自然情况下）以及训练数据（在人工条件下）中的表现如何。在每一步中，我们想选择的结构，是在掌握权值之后（而不是之前），表现最好的那个。因此，实际上，先天自然并不一定会排在后天培育之前。它们是互相交替的，每轮中的“培育”学习会为下一轮的“自然”学习做好基础，反之亦然。因为后天的培育，先天也会得到进化。皮质关联区域的进化在感知区域依赖于神经学习，没有神经学习，它就会变得毫无用处。小鹅跟着妈妈到处走（进化的行为），但得首先认识妈妈（通过学习掌握的能力）。如果它们孵化出来时第一个见到你，就会跟着你走，正如康拉德·洛伦茨明确表明过的那样。新生大脑已经对环境的特点进行编码，但不是很明显，进化过程会将大脑优化，从预期的输入信息中提取那些特点。同样的道理，在交替学习结构和权值的算法中，每个新结构都暗中成为前几轮掌握的权值的函数。

在所有可能的基因组中，很少有基因组和活性有机体对应。因此典型的适应度“地形”就由广袤的“平原”组成，偶尔有几座“尖峰”，这让进化变得十分困难。如果你从堪萨斯州出发，蒙住眼睛，你就不知道通往落基山脉的路在哪里，那么偶然撞上山脚并开始往上爬之前，你很长一段时间都是在晃悠。但如果你把进化和神经学习结合起来，有趣的事情就发生了。如果你在平地上，距离山脚不是很远，网络学习会帮你到达那里，距离山脚越近，它越有可能帮你。它就像有环视功能一样：在威奇托，它帮不上你，但在丹佛，你会看到远处的落基山脉，然后往那个方向走。丹佛现在看起来就比你之前蒙住眼时看起来的要适应得多了。净效应会将适应度顶峰加宽，这样你从之前艰难的地方出发，找到通往这些峰顶的路就成为可能，就像图5－4中的A点。

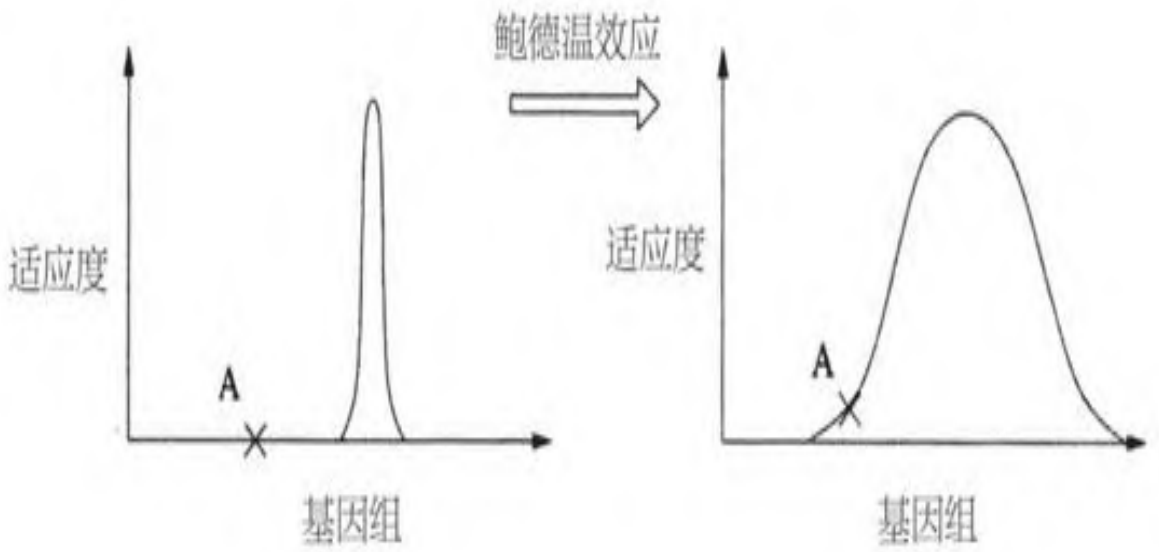


图5－4

在生物学中，这称作“鲍德温效应”，该效应是以J·M·鲍德温的名字命名的，他于1896年提出这个概念。在鲍德温进化中，初次掌握的行为，之后会变成天生的本领。如果像狗一样的哺乳动物能学会游泳，那么它们进化成海豹的可能性会比淹死的可能性更大。选择个体学习可以影响进化，不必依赖拉马克学说。杰夫·辛顿和史蒂芬·诺兰论证了机器学习中的鲍德温效应，方法就是利用遗传算法来进化神经网络结构，并观察到只有个体学习被允许时，适应度才会随时间增加。

谁学得最快，谁就会赢

进化寻求好的结构，而神经学习则填满这些结构：这样的结合是我们走向终极算法最简单的一步。先天与后天之争是无休止的波折，这个波折持续了2500年，且争论会越来越激烈，对于熟悉这一点的人

来说，这可能就像一个惊喜。然而，通过计算机的眼睛来看待生命能够将很多东西进行分类。“自然”对计算机来说就是它运行的程序，而“人工”则是获取的数据。“这两个哪个重要”这样的问题明显有点荒唐。没有程序和数据，就没有输出信息，比如也没有这样的说法：60%的输出信息由程序得出，另外的40%由数据得出。熟悉机器学习，会让你摆脱这种线性思维。

另外，你也许想知道，都到这点了，为什么我们还没有完成目标呢？如果我们把自然的两个主算法——进化和大脑——结合起来，我们当然就可以这样问。遗憾的是，目前我们只是大概知道自然如何学习，对许多应用来说已经足够，但对于真实的东西来说，还只是灰蒙蒙的影子。例如，胚胎发育是生命的重要组成部分，但在机器学习中，就没有和胚胎对应的类似物：“有机体”是基因组非常简单的功能，而我们可能忽视了重要的东西。另一个原因就是，即使已经弄明白自然如何学习，我们也不会满足。一方面，这个过程太缓慢。进化需要数十亿年来进行学习，而大脑需要一辈子。文化会好些：我可以花一辈子时间来学习一本书，而你可以几个小时把它读完。但学习算法应该有在几分钟或几秒钟内掌握知识的能力。谁学得快，谁就赢了，无论是加快进化速度的鲍德温效应、语言交流促进人类的学习效果，还是计算机以光的速度发现模型。机器学习是地球上生命之间竞争的最新篇章，更加快速的硬件只是等式的一边，另一边是更加智能的软件。

最重要的是，机器学习的目标是尽可能找到最好的学习算法，利用一切可能的方法，而进化和大脑不可能提供学习算法。进化的产物有很多明显的错误。例如，哺乳动物的视觉神经和视网膜前端而不是后端相连，这样会引起不必要的（而且异乎寻常的）盲点，就在中心凹旁边，而这里是视觉最敏锐的地方。活细胞的分子生物学原理非常混乱，分子生物学家常常自嘲道，只有对分子生物学一点也不懂的人才会相信智能设计。大脑的构造很有可能有相似的错误（大脑有许多计算机没有的限制，比如非常有限的短期记忆），而且没有理由待在

这些限制里。另外，我们听说过许多这样的情形，人类似乎坚持做错误的事情，正如丹尼尔·卡尼曼在他的书《思考，快与慢》里详细说明的那样。

与联结学派及演化新论者相反，符号学派和贝叶斯学派不相信“法自然”的说法。他们想从基本原理中找出学习算法该做什么，而且也包括我们人类。例如，如果我们学习如何诊断癌症，就说一句“这就是自然进行学习的方法，我们也照做”还不够。这其中有太多的利害关系，误差的代价就是生命。医生应该尽可能用最安全的方法来进行诊断，这些方法和那些数学家用来证明定理的方法相似，或者尽可能接近他们能掌控的水平（考虑到做到那样严谨有点不太可能）。他们需要权衡证据，目的是把诊断误差降到最小；或者更确切地说，误差的代价越大，他们犯错的可能性就越小（例如，未能找到确实存在的肿瘤，与诊断出其实不存在的肿瘤相比，可能要糟糕很多）。医生们得做出最优选择，而不是表面看起来不错的选择。

这是一个让人紧张的实例，贯穿科学和心理学的很多领域：是描述性理论与规范性理论之间的分歧，是“这就是它的样子”与“这就是它应该成为的样子”之间的分歧。然而，符号学派和贝叶斯学派想指出，弄明白“我们该怎样学习”也可以帮助我们了解人类如何学习，因为这两者大概不会完全不相关（远非如此）。特别指出的是，对于生存有重要意义、已经经历很长一段时间进化的行为应该就是最优的。我们不是很擅长回答关于概率的书面问题，但我们擅长快速选择手和手臂的动作来击中目标。许多心理学家已经使用符号学或贝叶斯的模型来解释人类行为的方方面面。符号学主导了认知心理学的头几十年。在20世纪八九十年代，联结主义者占支配地位，但现在贝叶斯学者的数量正在上升。

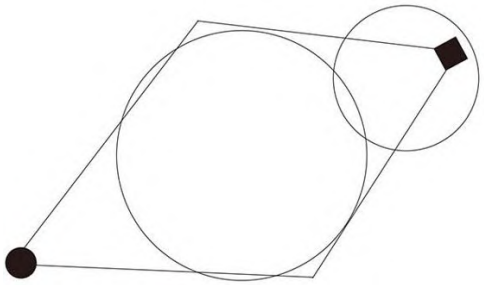
对于最困难的问题——我们真的想解决但没能解决的问题，例如，治愈癌症的问题——纯粹受到自然启发的方法要成功可能太无知，即使给定大量的数据也还是如此。原则上，我们可以掌握细胞新

陈代谢网络的完整模型，方法就是结合结构研究，利用或者不利用交叉，通过反向传播来进行参数学习，但有太多不利的局部最优陷阱。我们得利用更大块的数据来进行推理，根据需要集合或重新集合这些数据，然后利用逆向演绎来填补空缺。要让这样的目标来引导我们的学习行为：以最优方法诊断癌症，然后找到治疗癌症的最佳药物。

最优学习是贝叶斯学派的中心目标，而且他们肯定自己已经找到了实现这个目标的方法。请看这里……

第六章

贝叶斯学派：在贝叶斯教堂里



夜幕降临，大教堂的影子也渐渐变大。光线从教堂的彩色玻璃窗倾泻而出，在大街上及远处的建筑上投射出复杂的公式。当你走近时，会听到人们在里面唱圣歌的声音。唱的好像是拉丁语，又好像和数学有关，但你耳朵里的巴别鱼把它翻译成英文：“转动曲柄！转动曲柄！”正当你迈进去时，圣歌渐渐转化成声声满意的“啊”以及“后面，后面”的呢喃。你穿过人群。圣坛上矗立着一块巨大的石碑，上面刻着10英尺宽字母拼写成的公式：

$$P\left(A \mid B\right) = P\left(A\right)\ P\left(B \mid A\right)\ /\ P\left(B\right)$$

当你不解地盯着这个公式看的时候，谷歌眼镜起作用了，它闪了一下，说：“贝叶斯定理。”这时人群开始唱：“再多些数据！再多些数据！”大量的祭品正被无情地推向祭坛。突然，你意识到自己是祭品中的一个，但已经太晚了。当曲柄转到你上方时，你喊着：“不！我不想变成数据点！放开我啊！”

你醒来时浑身冷汗。放在你大腿上的，是一本名为《终极算法》的书。甩掉噩梦之后，你接着从中断的地方往下读。

统治世界的定理

通往最优学习的路径始于一个公式，这一点许多人都听说过：贝叶斯定理。但在这里，我们会以全新的眼光来看待这个公式，而且会意识到，它的力量要比你根据其日常用途猜测的要大得多。本质上，贝叶斯定理不仅仅是一个简单的规则，当你收到新的论据时，它用来改变你对某个假设的信任度：如果论据和假设一致，假设成立的概率上升，反之则下降。例如，如果你查出艾滋病病毒呈阳性，你感染艾滋病的可能性上升。当你获得许多条论据，例如，多重测试的结果时，事情就会变得更有意思。为了将所有论据结合起来，又免遭组合爆炸之苦，我们就得把猜想简化。当我们同时考虑多个假设，例如，一个病人所有不同、可能的诊断时，事情也会变得更有意思。在合理时期内，根据病人的症状来计算每种疾病发生的可能性需要很大的智慧。一旦知道如何来完成所有这些事情，我们就做好了学习贝叶斯方法的准备。对于贝叶斯学派来说，学习“仅仅是”贝叶斯定理的另外一个运用，将所有模型当作假设，将数据作为论据：随着你看到的数据越来越多，有些模型会变得越来越有可能性，而有些则相反，直到理想的模型渐渐突出，成为最终的胜者。贝叶斯学派已经发明出非常灵活的模型。下面我们就开始来了解它们。

托马斯·贝叶斯是一位18世纪的英国牧师，当时他并没有发觉自己会成为新宗教的中心。你可能很想问怎么会这样，但注意到这也发生在耶稣身上后，就不会这样问了：我们知道，基督教是由圣·保罗开创的，虽然耶稣把自己看作犹太教的顶峰。同样，我们知道贝叶斯

主义是由皮埃尔－西蒙·拉普拉斯创造的，他是比贝叶斯晚生50年的法国人。贝叶斯是第一个描述用新方法来考虑概率的牧师，但把那些想法变成定理，并以贝叶斯的名字来命名的人，却是拉普拉斯。

拉普拉斯是史上最伟大的数学家之一，关于他，人们知道最多的可能就是他对于牛顿学说决定论的憧憬：

> 在任何给定的一瞬间，如果有某位智者能够洞悉所有支配自然界的力和组成自然界的物体的相互位置，并且这位智者的智慧足以对这些数据进行分析，他就能用一个相同的公式来概括宇宙中最大的天体和最小的原子的运动。对这样的智者来说，没有什么是不确定的，未来同过去一样都历历在目。

这看起来有点讽刺的意味，因为拉普拉斯也是概率论的创始人之一，而他认为概率论仅仅是种可简化为计算的常识。他对于概率的探索本质上是对于休谟问题的专注。例如，我们怎么知道明天太阳会升起？太阳每天都会升起，今天也是，但没有什么能够保证它会继续升起。拉普拉斯的回答有两个部分。第一部分是我们当今所谓的“无差别原则”，或者“理由不充分原则”。有一天我们醒来，比如这时正是时间的起点，对于拉普拉斯来说则是大概5000年前，经历一个美妙的下午之后，我们看到太阳下山了。它还会出现吗？我们还没有看到过日出，所以也就没有特殊理由认为它会或者不会再升起。因此我们应该考虑两种情形出现的可能性相等，认为太阳还会再升起的可能性是一半。但是，拉普拉斯接着推断，如果过去能指导未来的一切，因为太阳每天都会升起，这应该会让我们更加坚信太阳会继续升起。5000年过后，明天的太阳还会升起的概率应该接近1，但不完全是，因为我们不能完全肯定。由这个思维实验，拉普拉斯导出他所谓的“接续法则”，该法则用于估算太阳升起n次后会再次升起的概率，表示为（n+1）/（n+2）。当n=0时，这个概率为1/2；随着n增加，概率也会增加，当n接近无穷大时，概率接近1。

这个法则由一个更为通用的原则推导得出。假设你半夜在一个陌生的星球上醒来。虽然你能看到的只是繁星闪烁的天空，你有理由相信太阳会在某个时间点升起，因为多数星球会自传并绕着自己的太阳转。所以你估计相应的概率应该会大于1/2（比如说2/3）。我们将其称为太阳会升起来的“先验概率”，因为这发生在看到任何证据之前。“先验概率”的基础并不是数过去这个星球上太阳升起的次数，因为过去你没有看到；它反映的是对于将要发生的事情，你优先相信的东西，这建立在你掌握的宇宙常识之上。但现在星星开始渐渐暗淡，所以你对于太阳会升起的信心越来越强，这建立于你在地球上生存的经历之上。你的这种信心源自“后验概率”，因为这个概率是在看到一些证据后得出的。天空开始渐渐变亮，后验概率又变得更大了。最终，太阳银色的大圆盘出现在地平线上方，而且可能正如《鲁拜集》的开篇部分描写的那样，套中了“苏丹的塔尖”。除非你产生幻觉，现在太阳是否会升起还不确定。

问题的关键是，随着你看到的证据越来越多，后验概率应该如何演变，答案就在贝叶斯定理里面。我们可以根据因果关系来考虑这个问题。日出使星星渐渐褪去光芒，天空变亮，但后者更能证明黎明的到来，比如因为星星也有可能在深夜时分，因为雾气笼罩而渐渐褪去光芒。所以看到天空变亮和看到星星褪去光芒相比，前者更能预示太阳升起的可能性。在数学符号中，我们说P（日出丨天空渐渐发亮），即天空渐渐发亮时日出的条件概率，比P（日出丨星星渐渐褪去光芒），即星星褪去光芒时日出的条件概率要大。根据贝叶斯定理，给定某原因时出现某结果的可能性越大，那么出现该结果是该原因引起的概率也会越大：如果P（天空渐渐发亮丨日出）比P（星星渐渐褪去光芒丨日出）要大，也许是因为有些星球距离太阳足够远，那些星星在日出之后仍会发光，那么P（日出丨天空渐渐发亮）也比P（日出丨星星渐渐褪去光芒）要大。

然而，情况并非完全如此。如果我们观察一个即使没有该原因也会发生的结果，那么能肯定的是，该原因的证据力不足。贝叶斯通过

以下句子概况了这一点：*P*（原因丨结果）随着*P*（结果），即结果的先验概率（也就是在原因不明的情况下结果出现的概率）的下降而下降。最终，其他条件不变，一个原因是前验的可能性越大，它该成为后验的可能性就越大。综上所述，贝叶斯定理认为：

P（原因丨结果）=*P*（原因）×*P*（结果丨原因）/ *P*（结果）

用*A*代替原因，用*B*代替结果，然后为了简洁，把乘法符号删掉，你就会得到大教堂上那个用10英尺宽的字母书写的公式。

这仅仅是定理的一个表述，当然还不能作为证据。但让人惊讶的是，证据十分简单。我们可以用医学诊断中的一个例子来阐明，也就是贝叶斯推理的“杀手级应用”之一。假设你是一位医生，上个月已经为100名病人进行诊断。其中的14名病人患流感，20名发烧，11名既感冒也发烧。因此感冒的人群中有发烧的病人的条件概率为11/14。设立条件缩小了我们正在考虑的集合，在这个例子中，所有病人的范围就缩小为患有流感的病人。在所有病人的集合中，发烧的概率是20/100；而在感冒病人的集合中，发烧的概率则是11/14。病人既发烧又感冒的概率，是患感冒病人的概率乘以发烧病人的概率：

P（感冒，发烧）=*P*（感冒）× *P*（发烧丨感冒）=14/100×11/14=11/100

但我们可以倒过来计算：

P（感冒、发烧）=*P*（发烧）×*P*（感冒丨发烧）

因此，因为它们都等于*P*（感冒、发烧），

P（发烧）×*P*（感冒丨发烧）= *P*（感冒）×*P*（发烧丨感冒）

两边都除以*P*（发烧），你会得到

P（感冒丨发烧）= *P*（感冒）×*P*（发烧丨感冒）/ *P*（发烧）

就是这样！这就是贝叶斯定理，感冒是原因，发烧是结果。

事实证明，人类并不是很擅长贝叶斯推理，至少涉及文字推理的时候是这样的。问题在于我们倾向于忽略原因的先验概率。如果你的艾滋病病毒检测呈阳性，而检测的假阳性只有1%，你会觉得恐慌吗？乍一看，这时你患有艾滋病的概率是99%。呀！可是让我们冷静一下，一步一步用贝叶斯定理来推导这个概率：

$$P(\text{HIV}\mid\text{阳性})=P(\text{HIV})\times P(\text{阳性}\mid\text{HIV})$$

P（HIV）是HIV在普通人群中的感染流行率，在美国这个概率约为0.3%。P（阳性）是化验结果显示你是否患有艾滋的概率，我们假设这个概率是1%。那么P（HIV丨阳性）=0.003×0.99/0.01=0.297。这和0.99有很大的差别！因为HIV在普通人群中较为少见。化验结果为阳性，这样你感染艾滋病的概率会增加两个数量级，但这些概率还是低于50%。如果你的化验结果为HIV阳性，接下来的准确做法是保持冷静，然后做另外更加确切的检查。检查结果很有可能是你没事。

贝叶斯定理之所以有用，是因为通常给定原因后，我们就会知道结果，但我们想知道的是已知结果，如何找出原因。例如，我们知道感冒病人发烧的概率，但真正想知道的是，发烧病人患感冒的概率是多少。贝叶斯定理让我们由原因推出结果，又由结果知道原因，但其重要性远非如此。对于贝叶斯定理的信仰者来说，这个伪装起来的公式其实是机器学习中的$F=ma$等式，很多结论和应用都是在这个等式的基础上得出的。而且无论终极算法是什么，它肯定“仅仅”是贝叶斯定理的一个算法应用。我之所以给“仅仅”加了双引号，是因为对于几乎所有简单的问题来说，在计算机上面应用贝叶斯定理非常困难，接下来，我们就会知道原因。

贝叶斯定理作为统计学和机器学习的基础，受到计算难题和巨大争论的困扰。你想知道原因也情有可原：这不就是我们之前在感冒的例子中看到的那样，贝叶斯定理是由条件概率概念得出的直接结果吗？的确，公式本身没有什么问题。争议在于相信贝叶斯定理的人怎么知道推导该定理的各个概率，以及那些概率的含义是什么。对于多

数统计学家来说，估算概率的唯一有效方法就是计算相应事件发生的频率。例如，感冒的概率是0.2，因为被观察的100名病人中，有20名发烧了。这是“频率论”对于概率的解释，统计学中占据主导地位的学派就是由此来命名的。但请注意，在日出的例子以及拉普拉斯的无差别原则中，我们会做点不一样的事：千方百计找到方法算出概率。到底有什么正当的理由，能够假设太阳升起的概率是1/2、2/3，或者别的呢？贝叶斯学派的回答是：概率并非频率，而是一种主观程度上的信任。因此，用概率来做什么由你决定，而贝叶斯推理让你做的事就是：通过新证据来修正你之前相信的东西，得到后来相信的东西（也被人们称为“转动贝叶斯手柄”）。贝叶斯学派对此观点的忠实近乎虔诚，足以经得住200年的攻击和计算。计算机已经强大到足以做贝叶斯推理，且在大数据的辅助下，它们开始占据上风。

所有模型都是错的，但有些却有用

实际上，医生不会仅凭发烧就断定病人感冒了，他会考虑所有的症状，包括病人是否咳嗽、喉咙痛、流鼻涕、头疼、发冷等。所以我们真正要算的是P（感冒 | 咳嗽、喉咙痛、流鼻涕、头疼、发冷……）。通过贝叶斯定理，我们知道这和P（咳嗽、喉咙痛、流鼻涕、头疼、发冷…… | 感冒）成正比。但现在，我们遇到一个问题：我们该如何估算这个概率？如果每个症状就是一个布尔变量（你要么有这个症状，要么没有），医生要考虑n个症状，那么病人就可能有2^n种症状的组合。假设我们有20种症状以及1万个病人的信息数据库，那么我们现在看到的仅仅是近100万种可能组合中的一小部分。更糟的是，为了准确估算某个特定组合出现的概率，我们对该组合的观察次数应至少为几十次，这意味着数据库需包含几千万个病人的信息。再另外增加10种症状，我们需要病人的数量就会超过地球上的人口数。

如果有100种症状，即使我们能魔法般地得到数据，那么世上所有的硬盘也没有足够的空间来存储所有的概率。而且如果病人来到医院，病人症状的组合我们从未遇到过，那么我们就不知该如何对其进行诊断。这时我们就与宿敌再次交锋：组合爆炸问题。

因此，我们就照着生活中常发生的那样做：妥协。我们做简化的假设来减少概率的数量，这些概率的数量由我们估算而来，且我们可以掌控。一个很简单且受人追捧的假设就是，在给定原因的情况下，所有的结果都相互独立。例如，假如我们知道你感冒了，发烧并不会改变你也咳嗽的可能性。从数学的角度讲，也就是说，P（发烧、咳嗽｜感冒）仅相当于P（发烧｜感冒）×P（咳嗽｜感冒）。你瞧：这些概率中的每一个都可以通过少量次数的观察得到。其实，在前文，我们就这样得到过病人发烧的概率，且咳嗽和其他症状的概率的算法也并无二致。那么我们需要观察的次数就不会随着症状数的增多而呈指数增长了，实际上，观察次数不会再增长。

请注意：我们只是说你在感冒的前提下，发烧和咳嗽是相互独立的，但并不是所有情况都如此。很明显，如果我们不知道你有没有感冒，那么发烧和咳嗽则有很大关系，因为如果你已经发烧了，那你很有可能会咳嗽。P（发烧、咳嗽）不等于P（发烧）×P（咳嗽）。我们现在说的是，如果我们知道你感冒了，那么知道你是否发烧，并不会让我们更容易知道你是否会咳嗽。同样，如果你不知道太阳要升起来了，而且你看到星星渐暗了，那么你就更加肯定天空会变亮，但如果你知道太阳即将升起，那么看到星星渐暗就没有什么影响了。

请再次注意：多亏贝叶斯定理，我们才掌握这个窍门。如果我们想直接估算P（感冒｜发烧、咳嗽等），假如不利用定理先将其转化成P（发烧、咳嗽等｜感冒），那么我们就还需要指数数量的概率，每个组合的症状以及感冒或非感冒都有一个概率。

如果学习算法利用贝叶斯定理，且给定原因时，假定结果相互独立，那么该学习算法被称为“朴素贝叶斯分类器”。因为这是一个很

朴素的猜想。实际上，发烧会增加咳嗽的可能性，即使你知道自己感冒了，因为（举个例子）发烧会让你得重感冒的可能性增大。但机器学习就是会做错误猜想的技术，而它又能侥幸逃脱。正如统计学家乔治·博克斯说的一句很有名的话那样：“所有的模型都是错的，但有些却有用。”虽然一个模型过于简化，但你有足够的数据用来估算那就比没有数据的完美模型要好。令人诧异的是，有些模型错误百出，同时又很有用。经济学家弥尔顿·弗里德曼甚至在一篇很有影响力的文章中提出，最有说服力的理论往往受到最大程度的简化，只要这些理论所做的预测是准确的，因为它们用最简洁的方法解释最复杂的问题。这对于我来说就像极远处的一座桥，但它阐明：与爱因斯坦的名言相悖，至少可以这么说，科学往往通过尽可能简化事物来取得进步。

没有人能肯定是谁发明了朴素贝叶斯算法。在1973年的一本模式识别教科书中，它被提到过，当时并未注明出处，但它真正流行起来是在20世纪90年代，那时研究人员惊喜地发现，它很多时候比许多更为复杂的学习算法还要准确。那时我还是一名研究生，而当我终于决定把朴素贝叶斯法纳入我的实验时，我震惊地发现自己很幸运，除了那个我运用到论文中的算法，它比所有我用来与之对比的算法都要有用，否则我可能也不会在这里了。

朴素贝叶斯法如今应用得很广泛。例如，它是许多垃圾邮件过滤器的基础。而这一切开始于大卫·赫克曼（一位卓越的贝叶斯研究员，同时也是一名医生），想到将垃圾邮件当作疾病，疾病的症状就相当于邮件中的文字：“伟哥”是一种症状，“免费”也是，但你最好朋友的姓可能会暗示这是一封合法邮件。那么我们就可以利用朴素贝叶斯法来将邮件分为垃圾邮件和非垃圾邮件了，只要垃圾邮件制造者通过随机选词来生成邮件。当然，这是一个荒谬的假设：只有当句子没有句法和内容时，它才正确。但那个夏天，迈赫兰·沙哈和一位斯坦福大学的毕业生在微软研究院实习时，曾对该假设进行过尝试，且效果很好。当比尔·盖茨问赫克曼怎会如此时，他指出为了识别垃

圾邮件，你不必理解信息的细节，只要通过邮件包含的词语来获得邮件的主旨就可以了。

一个基本的搜索引擎也会利用与朴素贝叶斯法极相似的算法来决定显示哪些页面来回应你的搜索。主要的区别在于：它会预测相关或非相关，而不是垃圾邮件或非垃圾邮件。运用朴素贝叶斯法来解决预测问题的例子几乎数不胜数。彼得·诺尔维格（谷歌的研究主任）一度告诉我，这是谷歌应用最为广泛的算法，谷歌的机器学习在每个角落都利用了该算法的功能。为什么朴素贝叶斯法会在谷歌员工中流行起来？这个问题不难回答。除了惊人的准确度，它的测量能力也很强。学习朴素贝叶斯分类器的原理，也仅相当于数出每个属性与每个类别出现的次数，花的时间不比从硬盘读取数据的时间长。

可以半开玩笑地说，你甚至可以比谷歌员工更大范围地利用朴素贝叶斯法：对整个宇宙进行模拟。的确，如果你相信全能的上帝，那么你可以将宇宙模拟成一个巨大的朴素贝叶斯法分布点，这里发生的每件事都是在上帝的意愿下独立发生的。当然，值得注意的是，我们不知道上帝在想什么，但在第八章我们会分析，如何在即使不知道例子类别的情况下，掌握朴素贝叶斯模型。

起初看起来可能不是这样，但朴素贝叶斯法与感知器算法密切相关。感知器增加权值，而朴素贝叶斯法则增加概率，但如果你选中一种算法，后者会转化成前者。两者都可以概括成“如果……那么……”的简单规则，这样每个先例都会多多少少体现在结果中，而不是在结果中“全有或全无”。这仅仅是暗指主算法的学习算法中关联较为深入的例子。你也许并没有有意识地去了解贝叶斯定理（那么，你现在意识到了），但在某种程度上，你大脑中数百亿神经元中的每一个都是贝叶斯定理的微小实例。

在看新闻时，朴素贝叶斯法是学习算法可以利用的良好概念模型：它可以捕获输入与输出之间两两相关的关系，这对于理解将学习算法引用到新事件中很有必要。但是机器学习也不仅仅和两两相关关

系有关，当然，正如大脑不仅仅包含一个神经元一样。当我们寻找更为复杂的模型时，真正的行动才刚刚开始。

从《尤金·奥涅金》到Siri

1913年第一次世界大战前夕，俄国数学家安德烈·马尔可夫发表了一篇文章，将所有事情的概率运用到诗歌当中。诗中，他模仿俄国文学的经典：普希金的《尤金·奥涅金》，运用了当今我们所说的“马尔可夫链”。他没有假定每个字母都是随机产生的，与剩下的毫无关联，而是最低限度引入了顺序结构：他让每个字母出现的概率由在它之前、与它紧接的字母来决定。他表示，举个例子，元音和辅音常常会交替出现，所以如果你看到一个辅音，那么下一个字母（忽略发音和空格）很有可能就是元音，但如果字母之间互相独立，出现元音的可能性就不会那么大。这可能看起来微不足道，但在计算机发明出来之前的年代，这需要花费数小时来数文字，而马尔可夫的观点在当时则很新颖。如果元音i是一个布尔型变量，《尤金·奥涅金》的第i个字母是元音，则该变量为真，如果它是一个辅音则为假，那么我们用图6－1这样的连锁图来表示马尔可夫模型，两个节点之间的箭头表示相应变量之间的直接依赖关系：

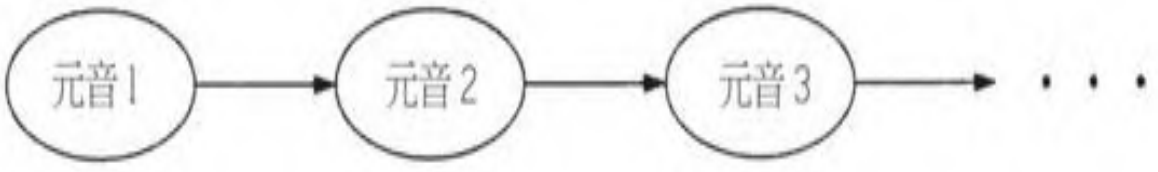


图6－1

马尔可夫假设（假设错误但有用）文中每个位置的概率都是一样的。因此我们只需估算三个概率：*P*（元音1=真），*P*（元音*i*+1=真丨元音*i*=真），且*P*（元音*i*+1=真丨元音*i*=假［因为概率相加等于1，由此我们可以马上得出*P*（元音1=假）等］。与朴素贝叶斯法一样，我们可以随心所欲拥有许多变量，而不用估算可能会抵达上限的概率的数量，但现在变量之间实际上会相互依赖。

如果我们测量的不仅仅是元音对辅音的概率，还有字母顺序遵循字母表顺序的概率，利用与《尤金·奥涅金》一样的统计数据，我们可以很愉快地生成新的文本：选择第一个字母，然后在第一个字母的基础上选择第二个字母，以此类推。当然结果是一堆没有意义的数据，但如果我们让每个字母都依照之前的几个字母而不止一个字母，这个过程就开始听起来更像一个酒鬼的疯话，虽然从整体上看没有意义，但从局部上看却很连贯。虽然这还不足以通过图灵测试，但像这样的模型是机器翻译系统的关键组成部分，比如谷歌翻译可以让你看到整版的英文页面（或者几乎整版），不管原页面的语言是什么。

源于谷歌的页面排名，本身就是一条马尔可夫链。拉里·佩奇认为，含有许多链接的页面，可能会比只含几个的要重要，而且来自重要页面的链接本身也更有价值。这样就形成了一种无限倒退，但我们可以利用马尔可夫链来掌控这种倒退。想象一下，一个页面搜索用户通过随机点击链接来从这个页面跳到另外一个页面：这时马尔可夫链的状态就不是文字而是页面了，这样问题就变得更为复杂，但数学原理是一样的。那么每个页面的得分就是搜索用户花在该页面上的时间，或者等于他徘徊很久后停留在该页面上的概率。

马尔可夫链无处不在，而且是人们研究最多的数学话题，但它仍是受到很大限制的概率模型。我们可以用图6－2中的模型来进一步探讨这个问题。

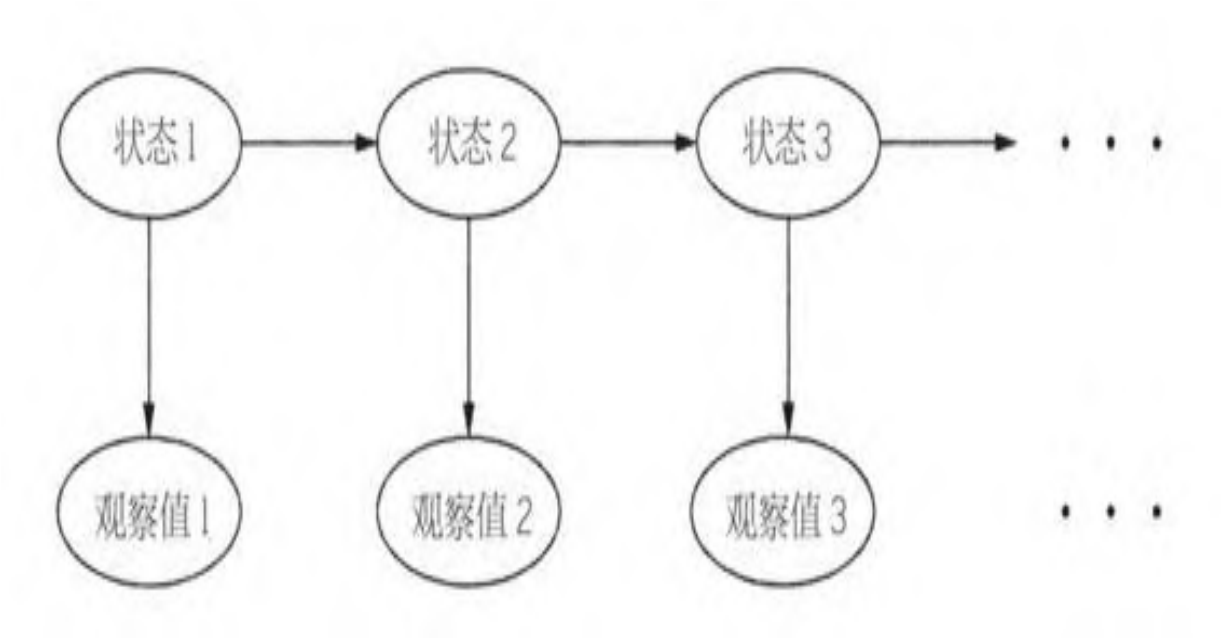


图6－2

与之前一样，这些状态组成一条马尔可夫链，但我们看不到它们，得从观察中将它们推导出来。人们称其为隐藏的马尔可夫模型，或者简称为HMM（有点误导人，因为被隐藏的是状态，而不是模型）。HMM和Siri一样，处于语音识别系统的中心。在语音识别过程中，隐藏的状态是书面形式的单词，而观察值则是对Siri说的声音，而目标则是从声音中推断出单词。模型有两个组成部分：给定当前单词的情况下，下一个单词出现的概率和在马尔可夫链中的一样；在单词被说出来的情况下，听到各种声音的概率（到底如何进行推导，这是一个有趣的问题，下面我们会来讨论这个问题）。

除了Siri，你每次用手机来通话时都会用到一个HMM。这是因为你的词语在空气中以位流的形式发送，而数位在发送过程中会受到损坏。这个HMM接着会从接收到的数位中（观察值）找到预期的那个（隐藏状态），这是HMM应该能做得到的，只要受损的数位不是很多。

HMM还是计算生物学家最为喜爱的工具。一个蛋白质分子是一个氨基酸序列，而DNA则是一个碱基序列。举个例子，如果我们想预测一个

蛋白质分子怎样才能形成三维形状，我们可以把氨基酸当作观察值，把每个点的褶皱类型当作隐藏状态。同样，我们可以用一个HMM来确定DNA中基因开始转录的地点，还可以确定其他许多属性。

如果状态和观察值都是连续而非离散变量，那么HMM就变成人们熟知的卡尔曼滤波器。经济学家利用卡尔曼滤波器来从数量的时间序列中消除冗余，比如GDP（国内生产总值）、通货膨胀、失业率。“真正的”GDP值属于隐藏的状态；在每一个时间点上，真值应该与观察值相似，同时也与之前的真值相似，因为经济很少会突然跳跃式增长。卡尔曼滤波器会交替使用这两者，同时会生成流畅的曲线，仍与观察值一致。当导弹巡航到目的地时，就是卡尔曼滤波器使它保持在轨道上。没有卡尔曼滤波器，人类就无法登上月球。

所有东西都有关联，但不是直接关联

HMM有助于模拟所有种类的序列，但它们远远不如符号学派的“如果……那么……”规则灵活，在这个规则当中，任何事都可以以前提的形式出现，而在任意下游规则中，一条规则的结果可以反过来当作前提。然而，如果我们允许如此随意的结构在实践中存在，那么需要掌握的概率数量将会呈爆发式增长。很长一段时间，没有人知道该如何打破这个循环，而研究者们只能求助于特别方案，比如将置信度估算与规则挂钩，并以某种方式将它们联合起来。如果A以0.8的置信度暗指B，而B以0.7的置信度暗指C，那么也许A暗指C的置信度则为0.8×0.7。

这些方案的问题在于，它们可能会严重出错。我可以根据两个合情合理的规则——如果灭火喷水系统打开了，那么玻璃就会湿；如果玻璃湿了，那么下雨了——推导出一条荒谬的规则：如果灭火喷水系

统打开了，那么下雨了。一个潜藏得更深的问题是，有了置信度评级规则，我们容易重复计算证据。假设你在读《纽约时报》，讲的是外星人已经登陆地球。这一天不是4月1日，可能这是一个玩笑。但是现在你在《华尔街日报》《今日美国》《华盛顿邮报》看到一样的标题。你开始感到慌张，就像奥森·威尔斯声名狼藉的广播剧《世界大战》的听众一样，没有意识到这只是戏剧化的描写。但是，如果你查看细节，会发现这四家报社都从美联社那里得到这个新闻标题，你又返回去怀疑这是一个玩笑，而这次开玩笑的是一位美联社的记者。规则系统无法解决这个问题，朴素贝叶斯法也一样。如果它标出诸如“由《纽约时报》报道”的字样，预示新闻事件是真实的，那么它能做的也只是加上“由美联社报道”的字样，这样只会让事情变得更糟糕。

20世纪80年代终于有了突破。朱迪亚·珀尔（加州大学洛杉矶分校的一名计算机科学教授）发明了一种新的表示方法：贝叶斯网络。珀尔是世界上最为卓著的计算机科学家之一，他的方法在机器学习、人工智能，以及其他许多领域迅速传播。2012年，他获得图灵奖，这是计算机科学领域的诺贝尔奖。

珀尔意识到，拥有一个随机变量之间复杂的依赖关系网络也没什么，只要每个变量仅仅直接依赖于其他几个变量。我们可以如之前看到的马尔可夫链和HMM那样，用一幅图来表示这些依赖关系，除了现在该图可以是任意结构（只要箭头不会形成闭环）。珀尔最喜欢的例子之一就是防盗报警器。当窃贼企图闯入时，你房子的报警器应该会响起来，但它也可能会因为地震而响起（在洛杉矶，珀尔居住的地方，地震和盗窃案发生的频率一样）。如果你此时工作到很晚，而你的邻居鲍勃打电话给跟说他听到你家的报警器响了，而你的邻居克莱尔却没打给你，你应该报警吗？图6－3表明了这种依赖关系。

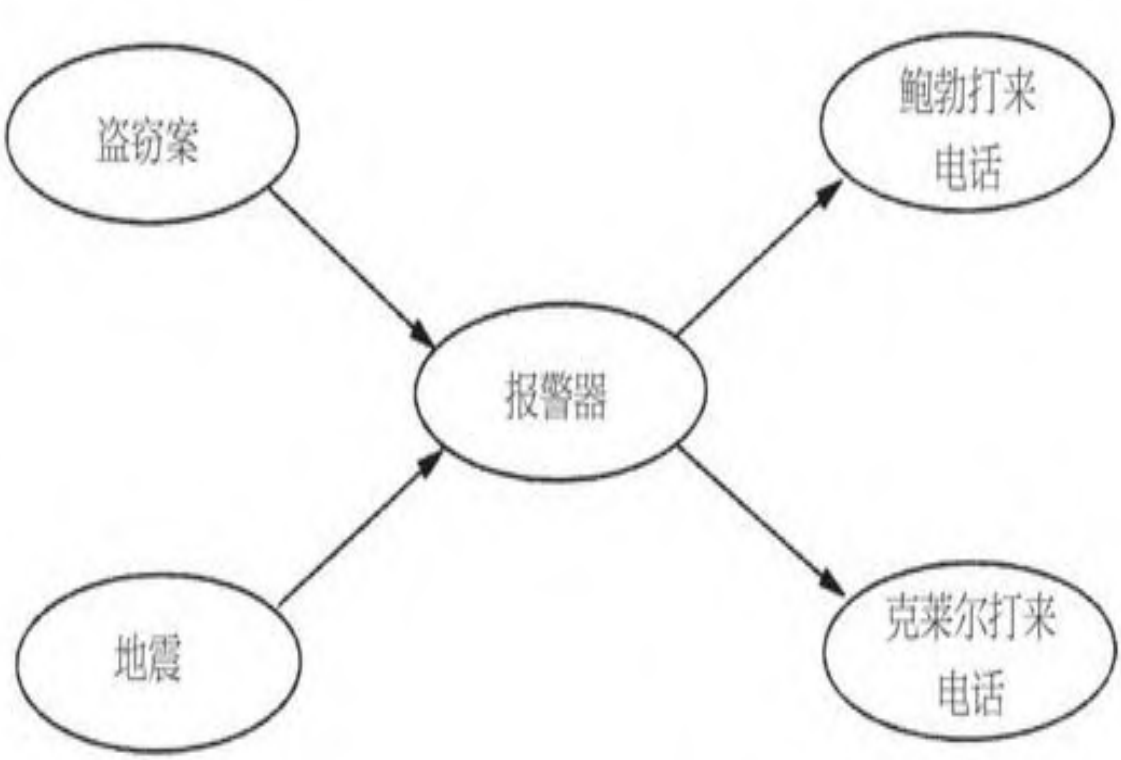


图6－3

如果图中有一个箭头从一个节点指到另外一个节点，我们说第一个节点是第二个节点的父节点。那么“报警器”的父节点就是“盗窃案”和“地震”，而“报警器”是“鲍勃打来电话”和“克莱尔打来电话”的唯一父节点。贝叶斯网络就是像这样的依赖关系图，附带一张包含每个变量的表格，给出变量父节点的每个值的组合概率。对于“盗窃案”和“地震”，我们只需每个一个概率，因为它们没有父节点。对于“报警器”我们需要四个概率：如果盗窃案和地震都不发生，报警器响起的概率；如果盗窃案发生，地震未发生，报警器响起的概率等。对于“鲍勃打来电话”，我们需要两个概率（考虑报警器和未考虑报警器），克莱尔的情况也一样。

以下是问题的关键点：鲍勃依据“盗窃案”和“地震”来打电话，但只能通过“报警器”来获取信息。考虑到“报警器”时，鲍勃的电话才会条件地独立于“盗窃案”和“地震”之外，克莱尔的情况

也一样。如果报警器未响起，你的邻居睡得很香，而窃贼也没有惊扰别人。同样，考虑到“报警器”，鲍勃和克莱尔相互独立。没有这个独立结构，你需要掌握32个2^5概率，每个概率对应5个变量的可能状态（或者31个概率，如果你讲究细节，因为最后一个可能是隐含概率）。有了条件独立性，你需要的只是1+1+4+2+2=10，可以少算68%的概率。这仅仅是小例子的情况，如果有数百或者数千个变量，省力的程度可能达到100%。

据生物学家巴里·康芒纳的观点，生态学的第一定律就是所有生命都与其他生命相互关联。这个说法可能正确，但也会使人们无法理解这个世界，如果不是多亏条件独立性：每个生命都相互关联，但只是间接关联。为了对我产生影响，1英里之外发生的事情，即使通过光传播，也得先影响我周围的环境。可以打趣说，空间是所有事情没有都发生在你身上的原因。换句话说，空间结构是条件独立性的一个实例。

在盗窃案的例子中，32个概率的完整表格不会明确表示出来，但通过较小表格和图形结构的集中，它会被隐含地表示出来。为了获得P（盗窃案、地震、报警器、鲍勃打来电话、克莱尔打来电话），我要做的就是把以下概率相乘：P（盗窃案）、P（地震）、P（报警器｜盗窃案、地震）、P（鲍勃打来电话｜报警器）、P（克莱尔打来电话｜报警器）。在任意贝叶斯网络中也是同样的道理：为了获得完整状态的概率，只需将单个变量表格中相应行上的概率相乘。因此，只要条件独立性有效，转换到更加简洁的表示方法不会导致信息丢失。这样我们就可以很容易算出极端非寻常状态的概率，包括之前未观察到的状态。贝叶斯网络揭穿这样一个常识性错误：机器学习无法预测鲜有的时间，或者纳西姆·塔勒布口中的“黑天鹅”。

回想一下，我们可以看到朴素贝叶斯法、马尔可夫链、HMM都是贝叶斯网络的特殊例子。朴素贝叶斯法的结构如图6－4所示。

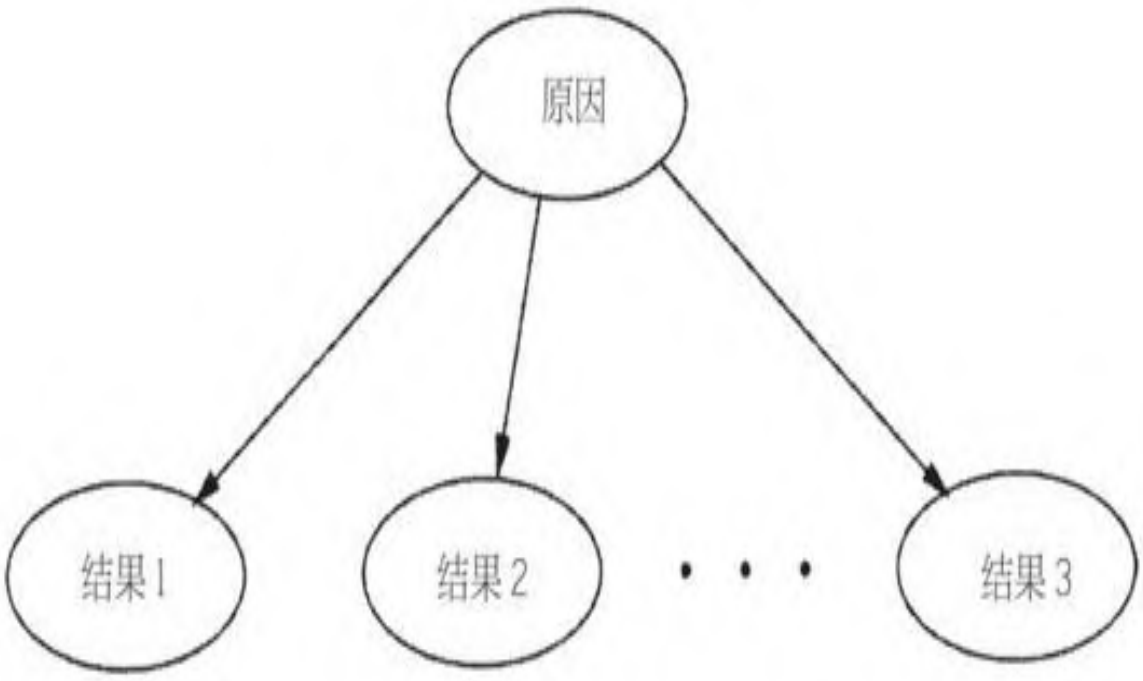


图6－4

马尔可夫链隐含这样的猜想：考虑到现在，未来会有条件地独立于过去。此外，HMM假设每个观察值只依赖于对应的状态。贝叶斯网络对贝叶斯学派来说，就像逻辑与符号学者的关系：一种通用语，可以让我们很好地对各式各样的情形进行编码，然后设计出与这些情形相一致的算法。

我们可以把贝叶斯网络想成“生成模型”，即从概率的角度，形成世界状态的方法：首先要决定盗窃案或地震是否会发生，然后在此基础上决定报警器是否会响起，再次在此基础上决定鲍勃和克莱尔是否会打电话。贝叶斯网络讲述这样的故事：A发生了，接着它导致B的发生；同时，C也发生了，而B和C共同引起D的发生。为了计算特定事件的概率，我们只需将与之相关事件的概率相乘即可。

贝叶斯网络最激动人心的应用之一，就是模拟基因在活细胞中如何相互管制。人们已经花费数十亿美元来找到单个基因和特殊疾病的

两两相关关系，但产出却低得让人失望。回想一下，这并不奇怪：细胞的活动是基因与环境复杂的相互作用的结果，而单个基因的预测能力有限。但有了贝叶斯网络，我们可以揭开这些相互关系，只要我们有必要的数据，而随着DNA微阵列技术的普及，我们越来越有希望能做到。

开辟机器学习在垃圾邮件的应用之后，大卫·赫克曼开始转向将贝叶斯应用于抵抗艾滋病的斗争中。艾滋病病毒是很强的对手，因为它可以迅速变异，这样所有疫苗或者药物长时间抑制艾滋病病毒就变得困难。赫克曼注意到，这是一个和猫捉老鼠一样的游戏，垃圾邮件过滤器和垃圾邮件玩耍。他决定将自己学习的一个经验付诸实践：攻击最弱的链接。在垃圾邮件的例子中，弱链接包括为了收到客户的款项而必须使用的网址。在艾滋病的例子中，它们就是病毒蛋白质的微小区域，不伤害病毒就无法改变这些区域。如果他可以对免疫系统进行训练，使其识别这些区域，然后攻击表现这些区域的细胞，他可能只有一种艾滋病疫苗。赫克曼和同事利用贝叶斯网络来帮助识别易受伤区域，并研发出一个疫苗交付机制，用来教免疫系统只攻击那些区域。这些交付机制在老鼠身上起作用了，而且现在人们已经在准备临床试验。

这样的事情经常发生：即使我们将所有条件独立性都考虑进来，贝叶斯网络中的一些节点仍有太多的父节点。有些网络布满箭头，这样当我们打印这些网络图时，纸面会变成黑色（物理学家马克·纽曼称之为“荒谬图”）。医生需要同时诊断病人可能患的病，而不止一种，而且每种疾病是许多不同症状的根源。除了感冒，发烧可能由任意数量的条件引起，但是考虑到条件的每个可能的组合，尝试预测其概率几无希望。其实还有一线希望，不需要这样的表格：为每个状态的起因详细说明每个节点的条件概率，我们可以对较简单的分布进行学习。最受青睐的选择是逻辑OR运算（Logical OR operation）的一个概率版本：任何原因都可以独自引起发烧，但每个原因都会有一个特定、无法引起发烧的概率，即使通常情况下该原因引起发烧的理由

充分。赫克曼和其他人已经对贝叶斯网络进行学习，通过这种方法来诊断数百种传染病。谷歌在其AdSense系统中利用这种类型的庞大贝叶斯网络，用于自动选择广告放入网页中。该网络将100万的满足变量相互关联起来，同时还通过3亿个箭头与1200万的词语和词组相关联，这些词语和词组都是从1000亿个文本片段和搜索词条中掌握的。

说点轻松的，微软的Xbox Live游戏用贝叶斯网络来对选手进行排名，并对技术水平相似的选手进行配对。这款游戏的结果就是可以知道对手技术水平的一个概率函数，而且利用贝叶斯定理，我们可以从选手的游戏结果推断选手的技术水平。

推理问题

遗憾的是，推理问题是一个巨大的障碍。仅仅因为贝叶斯让我们简洁地表达概率分布，这并不意味着我们也可以利用它进行有效推理。假设你想计算*P*（盗窃案｜鲍勃打电话，而克莱尔没有）。利用贝叶斯定理，你知道这仅仅是*P*（盗窃案）*P*（鲍勃打电话，而克莱尔没有｜盗窃案）/*P*（鲍勃打电话，而克莱尔没有），或者相当于*P*（盗窃案，鲍勃打电话，而克莱尔没有）/*P*（鲍勃打电话，而克莱尔没有）。如果有包含所有状态的概率的完整表格，你可以获得这两个概率，方法就是把表格中对应行的值加起来。例如，*P*（鲍勃打电话，而克莱尔没有）是在所有包含鲍勃打电话，而克莱尔没有的概率的行中，把这些概率值加起来，但是贝叶斯网络不会给你完整的表格。你总可以通过单个表格来构建这样的表格，但这需要很多时间和空间。我们真正需要的是计算*P*（盗窃案｜鲍勃打电话，而克莱尔没有），而不用建立完整的表格。简单地说，那就是贝叶斯网络中的推理问题。

在许多例子中，我们可以做到这一点，并避免指数性暴增。假设夜深人静时你正带领排成纵队的一个排，穿过敌人的领地，而你想确认所有士兵仍在跟着你。你可以停下，自己数人数，但那样做会浪费太多时间。一个更聪明的办法就是只问排在你后面的第一个兵：“你后面有几个兵？”每个士兵都会问自己后面的士兵同一个问题，知道最后一个士兵回答“一个也没有。”倒数第二个士兵现在可以说“一个”，以此类推，直到回到第一个士兵，每个士兵都会在后面士兵所报数的基础上加一。现在你知道有多少兵还跟着你，你甚至都不用停下来。

Siri用同样的想法来计算你刚才说的概率，通过它从麦克风中听到的声音来进行“报警”。把“Call the police”（报警）想成一排单词，正以纵队形式在页面上行走，“police”想知道它的概率，但要做到这一点，它需要知道“the”的概率；“the”回过头要知道“call”的概率。所以“call”计算它的概率，然后将其传递给“the”，“the”重复步骤并将概率传递给“police”。现在“police”知道它的概率了，这个概率受到句子中每个词语的适当影响，但我们绝不必建立8个概率的完整表格（第一个单词是否为“call”，第二个是否为“the”，第三个是否为“police”）。实际上，Siri考虑在每个位置中出现的所有单词，而不仅仅是第一个单词是否为“call”等，但算法是一样的。也许Siri认为，在声音的基础上，第一个单词不是“call”就是“tell”，第二个不是“the”就是“her”，第三个不是“police”就是“please”。个别地，也许最有可能的单词是“call”、“the”和“please”。但那样会变成一句没有意义的话“Call the please”，所以要考虑其他单词，Siri得出结论，认为句子就是“Call the police”。它会打电话，幸运的是警察及时赶到你家并抓住小偷。

如果图是一棵树而不是一条链，那么同样的想法还会奏效。如果你领导的是整支军队而不仅仅是一个排，你可以问每个连长，他们身后有多少士兵，然后把他们的回答相加。反过来，每个连长会问他的

每个排长，以此类推。但如果图形形成环状，你就会遇到麻烦。如果有一个联络官员，他同时是两个排的成员，他会数两次数。实际上，他身后的每个人都会数两次。这就是“外星人登陆”情形中发生的问题，举个例子，如果你想计算恐慌的概率，如图6－5所示。

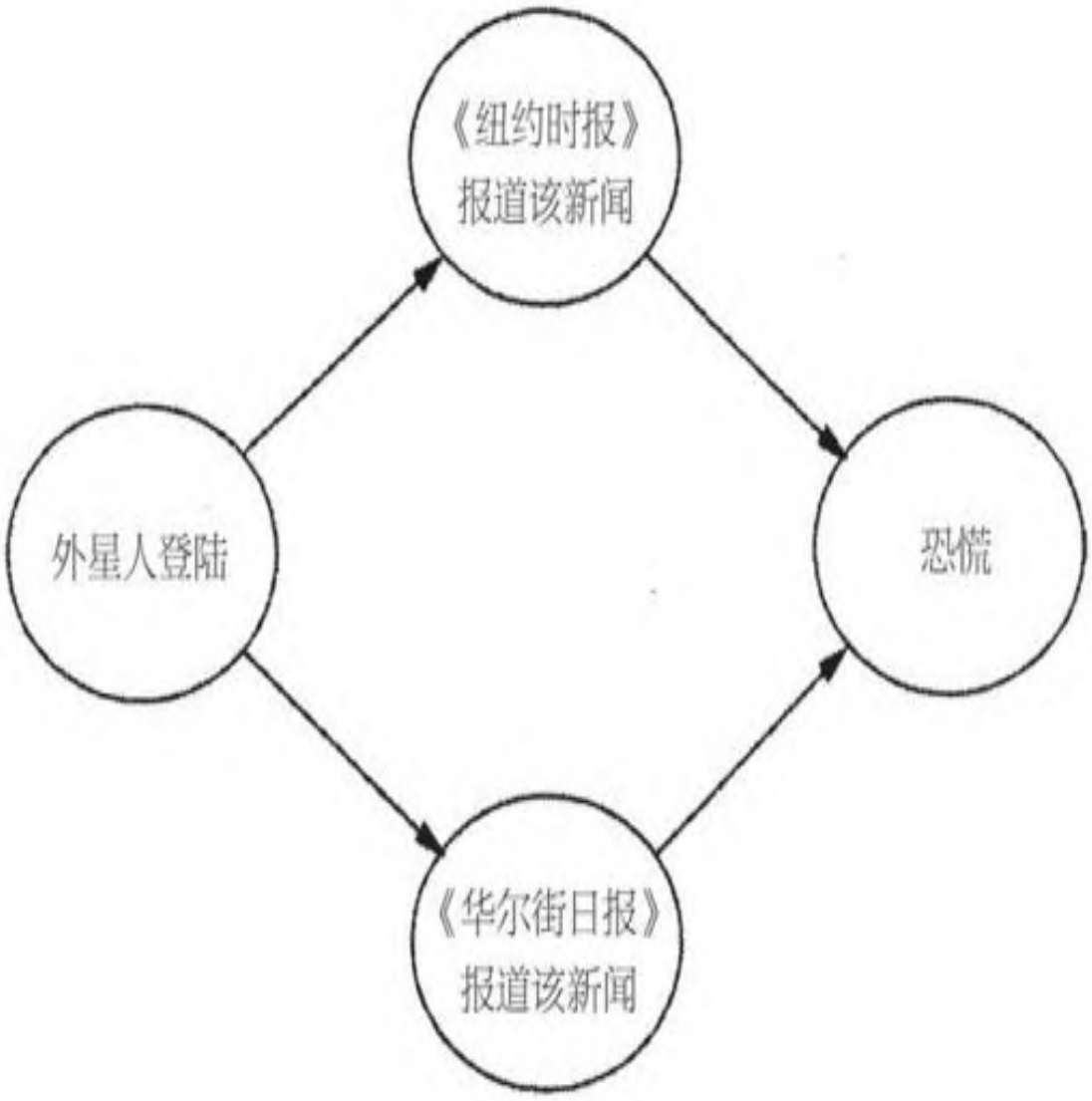


图6－5

有一个办法就是将“《纽约时报》报道该新闻”与“《华尔街日报》报道该新闻”结合起来，变成一个拥有4个值的单个特大变量：如果它们都报道了，那就是“是是”；如果《纽约时报》报道外星人登陆，而《华尔街日报》没有，那就是“是否”；等等。这样图形就会变成一条含三个变量的链子，一切顺利。然而，每次你加入新来源时，特大变量的值的数量会翻倍。如果你有50而不是2个新来源，特大变量会有250个值。所以这种方法只能帮你帮到这里，而更有效的方法还未被人发现。

这个问题比表面看起来的还要糟糕，因为贝叶斯网络实际上有“隐形的”箭头伴随着看得见的箭头。“盗窃案”和“地震”是相互独立的先验条件，但报警器响起会使两者纠缠在一起：报警器让你怀疑有盗窃，但如果现在听广播说有地震，你会假定那是让报警器响起的原因。地震已经“通过解释消除”报警器响起的疑虑，这样盗窃就不太可能会发生，因此两者又相互依赖了。在贝叶斯网络中，所有变量的根源都以这种方式相互独立，而这反过来引入进一步的依赖关系，这就使得最后的图往往比最初的图要更加密集。

推理的关键问题是你能否使填好的图“看起来像一棵树”，而不让树干变得太密。如果枝干上的特大变量包含过多的可能值，这棵树会不受限制地生长，直到它覆盖整个星球，像《小王子》中的猴面包树一样。在生命之树中，每个物种都是一个分枝，但每个枝干内部是一个图，每个生物都有一对父母，各有一对祖父母、外祖父母，以及若干个后代等。枝干的“厚度”是物种数量的大小。当枝干过于茂密时，我们的唯一选择就是求助于近似推理。

有一个方法，珀尔在关于贝叶斯网络的书中，将其当作练习，也就是假装图形没有闭环，并来来回回传播概率，直到这些概率集中于一点。这被人们称为“环路信念传播”（loopy belief propagation），因为它在含有闭环的图形中能起作用，也因为这是一个疯狂的想法。让人惊讶的是，在许多例子中，它能很好地起作用。

例如，这对于无线通信来说，是最先进的方法，随机变量在信息中被当作数位，以灵活的办法被编码起来。环路信念传播也可以收敛变成错误的答案，或者永远处于振荡状态。还有另一个办法，它起源于物理学，但被引入机器学习，还被迈克尔·乔丹和其他人大大扩充过：就是利用易于处理的分配来对难以处理的分配进行粗略估计，然后优化前者的参数，使其尽可能地与后者接近。

然而，最受人青睐的选择就是借酒浇愁，喝得酩酊大醉，然后整夜都在跌跌撞撞。该选择的技术术语为“马尔可夫链蒙特卡洛理论”（Markov chain Monte Carlo，MCMC）：有“蒙特卡洛”这个部分，是因为这个方法涉及机遇，比如到同名的赌场去，有“马尔可夫链”部分，是因为它涉及采取一系列措施，每个措施只能依赖于前一个措施。MCMC中的思想就是随便走走，就像众所周知的醉汉那样，以这样的方式从网络的这个状态跳到另一个状态。这样长期下来，每个状态受访的次数就与它的概率成正比。比如，接下来我们可以估算盗窃案的概率为我们访问某个状态的时间段，在这个状态中有一起盗窃案。一条“守规矩的”马尔可夫链会收敛到稳定分布中，所以过一会儿它总会给出大致一样的答案。例如，当你洗一副扑克牌时，过一会儿，所有牌的顺序都会近似，无论最初的顺序是什么。所以你知道如果有n种可能的顺序，每种顺序的概率就是$1/n$。MCMC的秘诀就在于设计一条马尔可夫链，收敛于我们贝叶斯网络的分布中。有一个简单的选择，就是重复循环通过所有变量，考虑其附近的状态，根据其条件概率对每个变量进行取样。人们在谈论MCMC时，往往把它当作一种模拟，但它其实不是：马尔可夫链不会模仿任何真实的程序，我们将其创造出来，目的是为了从贝叶斯网络中有效生成样本，因为贝叶斯网络本身就不是序变模式。

MCMC的起源可以追溯到曼哈顿计划，那时物理学家们需要估算中子与原子相撞的概率，并引发连锁反应。但近10年来，这引发了一场革命，常常使人们认为MCMC是有史以来最重要的算法之一。MCMC不仅有助于计算概率，也有助于求任何函数的积分。没有它，科学家们只

能受限于那些用分析方法来求积分的函数，或者受限于那些良性的低维积分，这些积分可以近似看作一系列梯形。有了MCMC，科学家们就可以自由构建复杂的模型，直到计算机会挑起重担。贝叶斯学派，举个例子，可能要感激MCMC，因为MCMC和其他东西比，更能为他们的方法提高人气。

不好的一面在于，MCMC的收敛速度往往慢得让人难以忍受，或者在未完成收敛时欺骗你已经完成收敛。真正的概率分配通常非常悬殊，微小概率的大片“荒地”会突然插入“珠穆朗玛峰”。接下来，马尔可夫链会收敛到最近的顶峰，然后停留在那里，得出偏差很大的概率估算值。这就像醉鬼跟随酒的香气来到最近的酒馆，然后整夜待在那里，而不是我们想让他做的那样，到城市里游荡。另一方面，如果我们不用马尔可夫链而是只生成独立的样本，和较简单的蒙特卡落法一样，没有任何味道指引我们，而且甚至可能都找不到那个第一家酒馆。这就像往城市的地图上射飞镖，希望飞镖能恰好落在酒馆上。

贝叶斯网络中的推理不仅限于计算概率，它也包括为证据找到最可信的解释方法，最能解释症状的疾病或者最能解释Siri听到的声音的词语。这和只是在每步挑最合适的词语不一样，因为在考虑声音的情况下，单个出现可能性大的词语，一起出现时可能性就不那么大了，在“Call the please”的例子中就是如此。可是，相同种类的算法也可以对这个任务起效（实际上，多数语音识别器会用到它们）。最重要的是，推理包括做最佳决定，引导这些决定的，不仅仅是不同结果的概率，还有相应的成本（或者使用这些技术术语的实用工具）。忽略你老板的邮件，安排你明天该做的事情，这比看到垃圾邮件付出的代价还要大，所以很多时候最好允许邮件通过，即使有时候它很有可能是垃圾邮件。

无人驾驶车辆和其他机器人是实践中概率推理的最好例子。随着车转悠起来，它同时会构建行驶领域的地图，越来越肯定地找到它在地图上的方位。根据最近的一项调查，伦敦的出租车司机大脑靠后的

海马区域比常人要大，这是一个涉及记忆和地图形成的大脑区域，因为他们要掌握城市的布局。也许他们利用的是相似的概率推理算法，这和人类的情况就不一样，饮酒似乎并不能帮上什么忙。

掌握贝叶斯学派的方法

既然我们（多多少少）知道了如何解决推理难题，就可以从数据中掌握贝叶斯网络了，因为对于贝叶斯学派来说，学习只是另一种形式的概率推理。你需要做的只是运用贝叶斯定理，把假设当作可能的原因，把数据当作观察到的效果：

$$P(\text{假设}\mid\text{数据})=P(\text{假设})\times P(\text{数据}\mid\text{假设})/P(\text{数据})$$

假设可以和整个贝叶斯网络一样复杂，或者和硬币正面朝上的概率一样简单。在后一个例子中，数据仅仅是一系列抛硬币行为的结果。例如，如果我们抛100次硬币，有70次正面朝上，频率论者会估算正面朝上的概率为0.7。根据所谓的“极大似然法”（maximum likelihood principle），这是有道理的：在所有正面朝上的可能概率中，0.7是在抛100次，有70次正面朝上的情况下最有可能的概率。假设成立的可能性是P（数据｜假设），而该法则称，我们应该选择那个能将该概率最大化的假设。即便如此，贝叶斯学派做事更加仔细。他们指出，我们从来无法肯定哪个假设是真的，所以我们不只是挑一个假设，比如硬币正面朝上的概率是0.7；相反，应该计算每个可能假设的后验概率，然后在做预测时容纳所有假设。所有假设的概率之和必须等于1，那么如果某个假设的概率增大，另一个则变小。实际上，对于贝叶斯学派来说，没有所谓的真相。你有一个优先于假设的分布，在见到数据后，它变成了后验分布，这是贝叶斯定理给出的说法，也就是贝叶斯定理的全部。

这与传统科学运作的方式相背离。这就像在说：“实际上，哥白尼和托勒密都不对；让我们只预测星球未来的轨迹，假定地球绕着太阳转，反过来也成立，然后对结果取平均值。”

当然，这是一个加权平均值，假设作为其后验概率的权值，那么能更好地解释数据的假设会更有价值。另外，就像玩笑说的那样，加入贝叶斯学派意味着绝不用说你对某事很肯定。

不用说，携带不止一个而是大量的假设是一种巨大的痛苦。在掌握贝叶斯网络的例子中，我们做预测的方法应该是对所有可能的贝叶斯网络取平均值，包括所有可能的图形结构，以及每个结构的所有可能的参数值。在某些情况下，我们可以以封闭形式对参数取平均值，但因为各种结构，进行得不顺利。举个例子，我们可以将马尔可夫链蒙特卡洛理论运用到网络的空间中，随着马尔可夫链的发展，从这个可能的网络跳到另外一个网络。将该复杂性与计算成本和贝叶斯有争议的概念（真的没有所谓的客观事实）结合起来，就不难知道为什么20世纪频率论一直主导科学界。

然而贝叶斯法还是有可取之处的，其中有一些主要原因。可取之处在于，很多时候，几乎所有的假设都会以微小的后验概率作为结尾，而我们可以安然地忽略它们。实际上，只考虑单个最可能的假设通常是一种非常好的近似方法。假设我们对于抛硬币问题的先验分布是：正面朝上的所有概率均相等。看到连续抛硬币的结果，这个效果是将分布越来越多地集中到与数据最为一致的假设上。例如，如果h包含正面朝上的可能概率，且一枚硬币正面朝上的概率为0.7，我们会看到图6－6。

每次抛完硬币的后验概率变成抛下一次硬币的先验概率，抛了一次又一次，我们越来越肯定h=0.7。如果我们只认准这一单个最可能假设（在这个例子中，h=0.7），这样贝叶斯法和频率论法就变得很相似，但有一个关键的不同点：贝叶斯学派考虑先验的P（假设），而不仅仅是可能性P（数据｜假设）［数据先验P（数据）可以忽略，因为

它对于所有假设来说都一样，所以不会影响获胜者的选择］。如果我们愿意假定所有的假设都和先验概率均等，那么贝叶斯法现在就简化为极大似然法。因此贝叶斯学派可以对频率论者说：“看，你做的仅是我们所做部分工作的特例，但至少我们能够使自己的假设清楚明白。”而如果假设与先验概率不均等，那么极大似然法的隐含假设（假设和先验概率均等）会给出错误的答案。

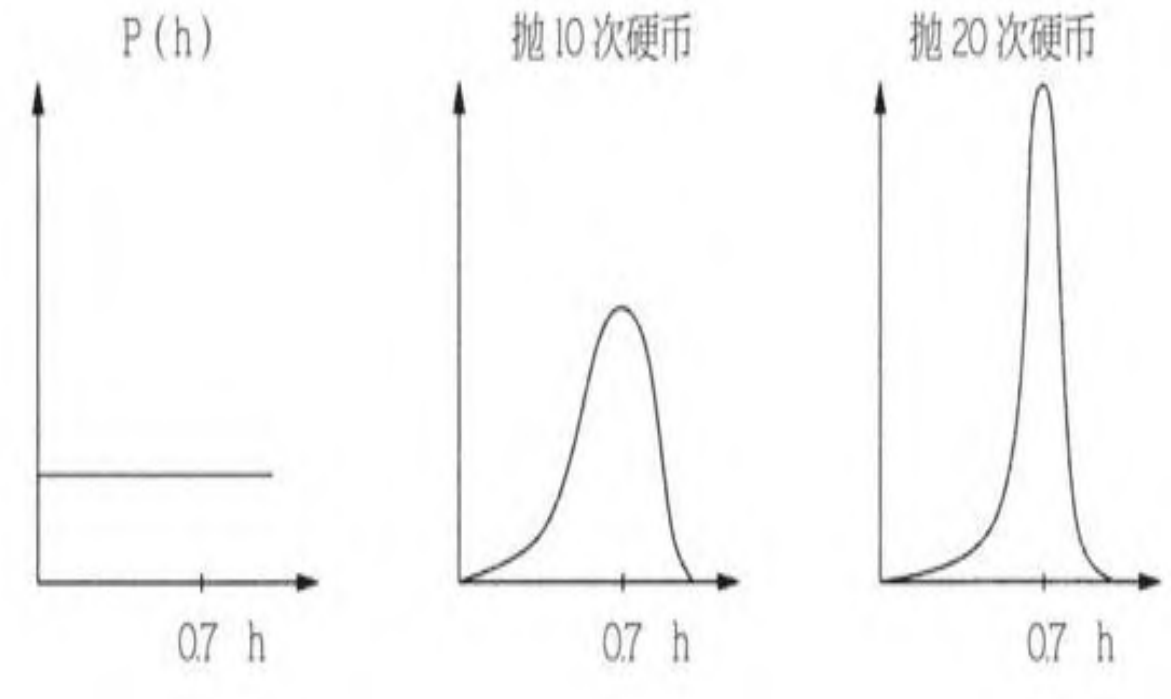


图6－6

这就像一场理论性的讨论，却得出惊人的实践性结果。如果我们只看了一次抛硬币，而结果是正面朝上，极大似然理论会说正面朝上的概率是1。这可能非常不准确，而当硬币反面朝上时，会让我们措手不及。一旦我们看过很多抛硬币行为，估算起来就更可靠，但在许多问题中，我们无法看到足够多次数的“抛硬币”，无论数据有多大。假定单词“supercalifragilisticexpialidocio us”在我们的训练数据中从未在垃圾邮件中出现过，而是出现在一封讨论《欢乐满人间》的邮件中。接下来，包含极大似然概率估算功能的朴素贝叶斯垃圾邮

件过滤器会决定包含该词的邮件不是垃圾邮件，尽管邮件中的其他词喊着：“垃圾邮件！垃圾邮件！”相反，贝叶斯学派会赋予这个词较低但非0的概率，让其出现在垃圾邮件中，同时允许其他词语将其覆盖。

如果我们努力掌握贝叶斯网络的结构和它的参数，问题只会变得更糟糕。我们可以通过爬山法来做到这一点，由空的网络（没有箭头）开始，添加最能提高可能性的箭头，以此类推，直到箭头不再起改善作用。遗憾的是，这很快就会引起大范围过拟合，网络会将零概率分配给未在数据中出现的所有状态。贝叶斯学派可以做更有意思的事情。他们可以利用先验分布来编码专家对该问题的观点——他们对休谟问题的回答。例如，我们可以为医学诊断设计一个原始贝叶斯网络，方法就是采访医生，询问他们哪些症状可能会对应哪些疾病，并添加相应的箭头。这就是“先验网络”，而先验分布可以通过添加或移除箭头的数量来惩罚替代网络。医生也不是完全可靠的，所以我们让数据来覆盖它们：如果通过添加数据来提高可能性比惩罚重要，我们会做这件事。

当然，频率论者正意识到这个问题，而且他们的回答，举个例子，是通过某因素来增加可能性，该因素会惩罚更多的复杂网络。但这时候频率论和贝叶斯学派就无法区别了，而你把计分函数称作“惩罚似然”还是“后验概率”就是喜好的问题了。

尽管频率论者和贝叶斯学派在一些问题上存在分歧，对于概率的含义也存在哲学上的差异。认为概率包含主观性让许多科学家感到略有不安，但如果没有这一点，很多用途就被禁止了。如果你是频率论者，那么只能对那些发生次数超过一次的事件的概率进行估算。所以像“希拉里·克林顿在下一轮总统竞选中打败杰布·布什的概率是多少”这样的问题则无法回答，因为没有哪场竞选的目的是让他们互相竞争。但对于贝叶斯学者来说，概率是包含主观程度的信任，所以他

可以自由地做有根据的猜测，而且推理演算会使他所有的猜想都一致。

贝叶斯法不仅仅适用于学习贝叶斯网络，还适用于其特殊情况（相反，尽管它们的名字如此，贝叶斯网络并非一定就是贝叶斯学派的：频率论者也可以掌握它们，正如我们刚看到的那样）。我们可以将先验分布置于任意级别的假设中（规则组、神经网络、程序）然后在给定数据的条件下，利用假设的可能性来对其进行更新。贝叶斯学派的观点就是，选择什么表示方法由你决定，但得利用贝叶斯定理来掌握它。20世纪90年代，他们声势浩大地接管了神经信息处理系统国际会议（NIPS），即联结学派研究的主会场。其中的罪魁祸首（可以这么说）包括大卫·麦凯、雷德福·尼尔、迈克尔·乔丹。麦凯是英国人，是加州理工学院约翰·霍普菲尔德的学生，后来成为英国能源部的首席科学顾问，他表明了如何通过贝叶斯法来学习多层感知器。尼尔向联结学派介绍MCMC，而乔丹则向他们介绍变分推理。最终，他们指出在限制范围内，你可以在多层感知器中“集中”神经元，剩下一种未提到它们的贝叶斯模型。不久以后，在向NIPS提交的论文中，“神经的”一词很好地表达了“拒绝”的意思。有些研究人员开玩笑称，该会议应该改名为BIPS，即“贝叶斯信息处理系统”（Bayesian Information Processing Systems）。

马尔可夫权衡证据

但在通往统治世界的路上发生了有趣的事情。利用贝叶斯模型的研究人员注意到，如果你用非法的方式来调整概率，得出的结果会更好。例如，在语音识别器中，通过调整P（词语）可以提高准确率，但它就不再是贝叶斯定理了。发生了什么？结果是，问题的原因在于生

成模型做出的错误独立性假设。简化的图形结构使模型变得可掌握，而且值得保留，不过因为手头任务，我们最好只掌握自己能力范围内的最佳参数，无论它们是否为概率。朴素贝叶斯法的一个真正优势在于，可以提供少量而有信息量的特征，通过这些特征可以预测级别。另外，它还可以提供快速、稳健的方法来掌握相应的参数。在垃圾邮件过滤器中，每个特征都是某个特定的词语出现在垃圾邮件中，而相应的参数是它发生的频率是多少，对于非垃圾邮件也是同样的道理。以这种方式来看，考虑到把最佳预测变成可能，甚至在很多情况下其独立性假设受到严重违背，朴素贝叶斯可能是最优的。当我意识到这一点，并于1996年发表一篇关于它的论文时，人们对于朴素贝叶斯的怀疑消解了，并帮助它取得成功。但它也向不同的模型迈出了一步，该模型在过去20年中已渐渐取代贝叶斯网络，它就是马尔可夫网络。

马尔可夫网络是一组特征以及对应的权值，特点和权值共同定义概率分布。特征可以和“这是一首民歌”一样简单，也可以与“这是一首由嘻哈艺术家创作的民歌，有萨克斯的重复乐段和递降的和弦部分”一样复杂。潘多拉利用一大组特征，它将其称为“音乐基因组计划”，目的是为你挑选要播放的歌曲。假定我们将它们与马尔可夫网络接通，那会怎么样？如果你喜欢民歌，相应特征的权值会上升，而当你打开潘多拉时，更有可能听到民歌。如果你也喜欢听嘻哈艺术家创作的歌曲，该特征的权值也会上升。这时你最有可能听到的歌曲同时包含这两种特征，即民歌和嘻哈艺术家的创作特点。如果你不喜欢民歌或者不喜欢嘻哈艺术家本身，而只喜欢两者的结合，那么你想要的是具备更加复杂特征的“嘻哈艺术家创作的民歌”。潘多拉的特征由手工创造，但在马尔可夫网络中我们也可以利用爬山法来掌握特征，这与规则归纳相似。不管怎样，梯度下降是掌握权值的一种好方法。

像贝叶斯网络一样，马尔可夫网络可以通过图表来表示，但它们用无向弧而不用箭头。两个变量被联结起来，这意味着它们会直接相

互依赖，如果它们一起出现在某个体征中，例如，“由嘻哈艺术家创作的民歌”中的“民歌”和“由嘻哈艺术家创作”。

马尔可夫网络在许多领域中能起到主要作用，例如，计算机视觉。举个例子，一辆无人驾驶车辆需把看到的每张图片分成道路、天空、村庄三个部分。有一个选择就是根据其颜色，将每个像素标记为三个部分中的一个，但这种选择不够好。图片充满嘈杂因素和变量，车辆会出现幻觉，想象道路旁布满岩石，还有天空中出现的一段段公路。但是我们知道，图片中近似的像素通常属于同一物体，而我们可以引入相应的一组特征：对于每一对相邻像素，如果它们属于相同物体，则特征是真的，否则是假的。现在有大段连续街区公路和天空的图片比没有这些的图片更有可能，并且车会直走，而不是不断左右转向来避开想象中的岩石。

马尔可夫网络可以经过训练，来最大化整个数据的可能性，或者在知道某些信息的情况下，将我们想预测的事情的可能性最大化。对于Siri来说，整个数据的可能性是P（单词、声音），而我们感兴趣的条件可能性是P（单词|声音）。通过优化后者，我们可以忽略P（声音），因为这个概率只会使我们偏离目标。既然我们忽略它，它可以是任意复杂的。这比HMM不切实际的猜想要好得多：声音只依赖于对应的单词，不会受到周围环境的任何影响。实际上，如果Siri的所有关心找出你刚才说的话，可能它甚至就不用担心概率问题了。它只要确保在计算它们特征权值的总和时，正确单词的得分会比错误单词的得分高，而且为了保险，要高出很多。

类推学派用这种推理方式得出其逻辑结论，正如我们在第七章将看到的那样。在21世纪的前10年，他们反过来接管NIPS。现在联结学派打着深度学习的旗号再次占据主导地位。有些人说研究总是处于循环状态，但它更像一个螺旋，闭环沿着前进的方向绕。在机器学习中，螺旋会收敛至终极算法。

逻辑与概率：一对不幸的组合

你可能认为贝叶斯学派和符号学派相处得很好，因为考虑到他们都相信学习的第一原理方法，而不相信自然启发的方法。事实远不是这样，符号学派不喜欢概率，而且会开这样的玩笑：“换个电灯泡需要多少个贝叶斯学者？他们不确定。细想一下，原来他们是不确定灯泡是不是烧坏了。”更严肃地说，符号学派指出我们因为概率而付出的高昂代价。推理突然变得更加珍贵，所有那些数据都变得难以理解，我们得处理好先验概率，一大批“僵尸”假设会永远追着我们。短时间内将分散知识集中起来的能力，对于符号学派来说是多么重要，却已经消失。最糟糕的是，我们不知道如何对自己想学习的很多东西进行概率分配。贝叶斯网络是在变量的一个矢量上的分布，但在网络、数据库、知识库、语言、计划、计算机程序上的分布呢？所有这些在逻辑中都易于处理，而无法掌握它们的算法明显就不是终极算法。

反过来，贝叶斯学派指出了逻辑的脆弱性。如果我有一条规则，如“鸟会飞”，没有哪个世界会存在不会飞的鸟。如果我想补充一下，加上例外条件，例如，“鸟会飞，除非它们是企鹅”，那么这些例外条件永远都说不完（鸵鸟呢？笼中鸟呢？死鸟？断了翅膀的鸟？翅膀湿了的鸟？）。有个医生诊断你得了癌症，而你想听不同的意见。如果第二个医生不同意，你就陷入困境了。你无法权衡这两种观点，只能两个都相信。然后发生了一场灾祸：猪会飞，永恒运动成为可能，而地球则不再存在，因为在逻辑当中，任何事都可以从矛盾中推理出来。再者，如果知识是从数据中掌握得来的，那么我就无法肯定它是对的。为什么符号学派却假装肯定？无疑休谟会对这样漫不经心的态度感到不满。

贝叶斯学派和符号学派一致认为，先验假设不可避免，但对于他们认可的先验知识种类却存在分歧。对于贝叶斯学派来说，知识越过模型的结构和参数，进入先验分布中。原则上，之前的参数可以是任意我们喜欢的值，但讽刺的是，贝叶斯学派趋向于选择信息量不足的先验假设（比如将相同概率分配给所有假设），因为这样更易于计算。在任何情况下，人类都不是很擅长估算概率。对于结构这方面，贝叶斯网络提供直观的方法来整合知识：如果你认为A直接引起B，那么应把箭头从A指向B。但符号学者则要灵活得多：作为先验知识，你可以为自己的学习算法提供任何能用逻辑编码的东西，实际上，所有东西都可以用逻辑编码，只要它是黑白色的。

显然，我们既需要逻辑，也需要概率。治愈癌症就是一个很好的例子。贝叶斯网络可以从单个方面模仿细胞如何起作用，就像基因调节和蛋白质折叠那样，但只有逻辑可以将所有碎片组合到一张连贯的图片中。此外，逻辑无法处理不完整或包含嘈杂因素的信息，这在实验生物学中较普遍，但贝叶斯网络可以沉着地处理这个问题。

贝叶斯学习能对单个数据表起作用，表中的每列表示一个变量（例如，一个基因的表达水平），而每行表示一个实例（例如，一个微阵列实验，包含每个基因被观察到的水平）。如果表格有“漏洞”和测量误差，不要紧，因为我们可以利用概率推理来弥补漏洞，然后对误差取平均值。但如果我们拥有的表格超过一个，那么贝叶斯学习就会停滞不前。例如，它不懂如何将基因表达的数据与DNA片段转译成蛋白质的数据结合起来，也不知道那些三维形状的蛋白质如何反过来使它们锁定到DNA分子的不同部分中，同时影响其他基因的表达。利用逻辑，我们可以轻易地写出与所有这些方面相关的规则，然后通过结合相关表格来对它们进行学习，但唯一的条件就是表格没有任何漏洞或者误差。

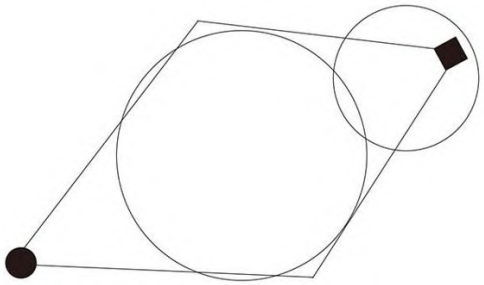
将联结学派和进化学派结合起来很简单：只要改善网络结构，利用反向传播来掌握参数。但将逻辑和概率统一起来要困难得多。最早

尝试这么做的人是莱布尼茨，他是逻辑和概率的开拓者。还有一些19—20世纪最伟大的哲学家和数学家，比如乔治·布尔和鲁道夫·卡尔纳普，他们都努力想解决这个问题，但最终没有走得很远。最近，计算机科学家和人工智能研究人员加入了这场争论。随着21世纪的到来，我们取得的最好成果也是不完整的，比如将一些逻辑结构加入贝叶斯网络中。多数专家相信，将逻辑和概率相统一是不可能的。寻求一个终极算法的前景并不乐观，特别原因在于，当前进化学派的和联结学派的算法无法处理不完整的信息和多数据组。

幸运的是，我们已经攻克了这个难题，而终极算法现在看起来离我们更近了。在第九章中，我们会看到人类是如何做到这一点的，并做出推断。首先，我们要找到已丢失的那张很重要的拼图：如何学习小数据。在当下数据洪流的背景下，可能这显得没有必要，但实际上我们常常会发现，自己有很多关于要解决问题的数据，而关于其他问题的数据则几乎没有。这是机器学习中最重要的思想之一流行起来的原因：类比思想。到目前为止，我们谈到的所有学派有一个共同点：他们都学习研究中的现象的显式模型，无论它是一组规则、一个多层感知器、一个基因计划，还是一个贝叶斯网络。当他们没有足够的数据来做这件事时，就会被难住。但类比学派可以从甚至小到一个例子的数据中学习，因为他们绝不会形成一种模式。下面，让我们看看他们是怎么做的。

第七章

类推学派：像什么就是什么



弗兰克·阿巴内尔是历史上最臭名昭著的骗子之一，莱昂纳多·迪卡普里奥曾在斯皮尔伯格的电影《逍遥法外》中饰演阿巴内尔。他伪造价值数百万美元的支票，冒充律师和大学讲师的身份，并以泛美航空飞行员的假身份环游世界，这些都发生在他21周岁之前。也许他最令人吃惊的“事迹”，就是在20世纪60年代的亚特兰大成功假扮医生将近一年之久。行医应该需要在医学院学习多年，要有执照、实习期等诸如此类的东西，但阿巴内尔成功避开所有这些细节，而且从未被人发现。

想象一下如何去冒这个险。你偷偷摸摸进入医生不在的办公室，不久后有个病人走进来，并把他所有的症状都告诉了你。现在你得给他诊断，除非你对于药物一窍不通。你只有一柜子的病人资料：他们的症状、诊断结果、接受过的治疗等。你该怎么办？最简单的办法就是找到和现在这个病人症状最相似的病人的资料，然后做相同的诊断。如果你对待病人的态度和阿巴内尔一样有说服力，那么可能会成功。在医学之外，同样的思想也很适用。如果你是一名年轻的总统，

面临世界危机，就像肯尼迪当年一样，当时美国的一架间谍机报告，苏联的核导弹被部署在古巴，你很有可能因为经验不足而措手不及。你可以找历史上与当前情况相似的场景，然后尝试从这些场景中吸收经验。参谋长联席会议敦促对古巴进行袭击，但肯尼迪那时刚读过《八月炮火》，这是一本描写第一次世界大战的畅销书，所以他知道那样做很容易会引发全面战争。所以他选择了对古巴进行海上封锁，也许这样做把世界从核战争中拯救出来了。

类比是推动许多历史上最伟大科学进步的动力。当达尔文阅读马尔萨斯的《人口论》时，被经济和自然界中生存竞争的相似性触动，所以有了自然选择理论的诞生。波尔的原子核模型是由将模型看作微型太阳系、将电子看作行星、将原子核看作太阳而产生的。克库勒也是白天做梦梦见蛇吃自己的尾巴才发现苯分子的环状结构的。

类比推理有着突出的知识谱系。亚里士多德在他的相似律中就表达了这一点：如果两个事物相似，其中的一个想法会触发另外一个想法。诸如洛克和休谟之类的经验主义者就追随这个规律。尼采说，真理是一支由暗喻拟人组成的队伍。康德也是其中的信奉者。威廉·詹姆斯认为：“这种意义上的一致性是我们思想的支柱。”一些当时的心理学家甚至认为从整体上看，人类认知是分析必不可少的部分。我们依靠它来找到通往新城市的道路，并理解诸如“领悟”“形象高大”之类的表达。那些十几岁、喜欢说话时在每个句子中加入“像”的人也许喜欢且认为类别是重要的。

考虑到这一点，类比在机器学习中扮演重要角色就不足为奇了。刚开始它进展缓慢，甚至被神经网络夺走了光芒。它的第一个算法的化身出现在一份写于1951年、名不见经传的技术报告中，作者是两位伯克利的统计学家——伊夫琳·菲克斯和乔·霍奇斯。这篇报告几十年之后才发表于主流期刊中。但同时，关于菲克斯和霍奇斯的算法的论文也开始出现，后来逐渐增加，直到它成为计算机科学界中受到研究最多的文章之一。最近邻算法，正如其名，是我们类比学习法之旅

的第一站。第二站是支持向量机，这是世纪之交风靡机器学习领域的原理，但最近风头被深度学习掩盖。第三站也是最后一站，是成熟的类比推理法，几十年来是心理学和人工智能的重要组成部分，也是几十年来机器学习领域的背景主题。

5个学派中，类推学派是最不具有凝聚力的一个学派。不像其他学派有很强的身份意识和共同理想，类推学派则更像研究人员松散的集合体，他们的统一依靠的是对于作为学习基础的、相似性判断的信任。一些人，比如支持向量机的支持者甚至可能会反对将自己归入这个范畴。但我认为如果人们能为共同的事业而努力，会受益很多。在机器学习中，相似性是核心思想之一，而类推学派会以各种伪装的方式来保护它。也许在未来10年，机器学习会被深度类比统治，在某种算法中，与最近邻法的高效、支持向量机的数学精密性、类比推理的力量和灵活性结合（瞧，我又泄露了自己的一个秘密研究计划）。

完美另一半

最近邻算法是人类有史以来发明的最简单、最快速的学习算法。实际上，你甚至可以说，这是人类可以发明的最快速的算法。它可以说什么也不做，所以花在运行上的时间为零。再没有比这个更好的了。如果想掌握面部识别技巧，且有大量标有脸部或非脸部标签的图片，你只需让算法待在那儿就好。别担心，高兴起来吧。在不知道的情况下，那些图片已经暗中形成关于“什么是脸”的模型。假设你是脸书，想在照片上自动识别人脸，人们上传这些照片，就是用朋友的名字来标记这些照片的前奏。不用做什么是一件好事，只要脸书用户每天上传超过3亿张照片就可以了。将目前为止我们见过的学习算法

（可能要排除朴素贝叶斯法这个例外）应用到这些照片上，可能需要一卡车计算机，而贝叶斯算法的智能程度还不足以用来识别脸部。

当然，这是要付出代价的，这个代价在测试时会出现。名叫简的用户刚上传一张新的照片。这是一张脸吗？最近邻算法的回答是：在脸书所有标记过的照片库中，找到那张和这张最相似的照片（它的“最近邻”）如果那张照片包含一张脸，这张也会有。这足够简单，但理想的情况下，你得在不到一秒的时间内浏览近几十亿张照片。就像一个不想为考试而学习的懒学生，最近邻算法此时猝不及防，不得不断断续续地运行。不像在现实生活中，你的妈妈会跟你说不要把今天可以做的事留到明天，在机器学习中，拖延真的可以带来成功。实际上，学习的整个风格（最近邻算法是其中一种）有时会被人们称为“懒惰学习算法”，在这种情况下，这个术语并没有什么贬义。

懒惰学习算法之所以比它表面看起来更加智能，是因为虽然它的模型是隐性的，但其实它可以很复杂。考虑到极端的情况下，每个等级我们只能有一个例子。例如，我们想知道两个国家的边界在哪里，但我们知道的仅仅是其首都的位置。多数学习算法可能会被难住，但最近邻算法会恰当地猜出，边界会是一条线，位于两个城市的中间位置（见图7－1）。

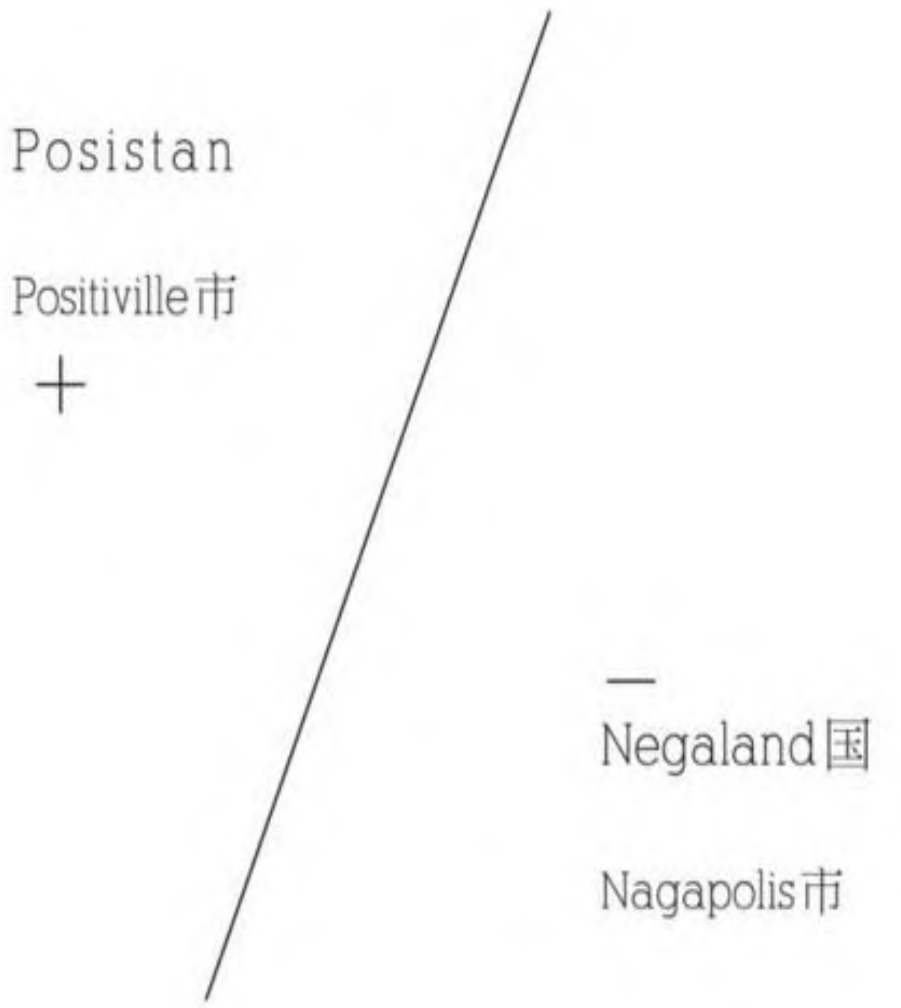


图7－1

边界上的点距离两个首都的距离都一样，边界线左边的点距离Positiville市较近，因此最近邻算法假定这些点是Posistan国的一部分，反之亦然。如果那条线就是准确的边界，当然是万幸，但作为一种近似法，很有可能这条线什么也不是。当我们知道边界线两边都有很多城市时，事情才真正变得有趣起来（见图7－2）。

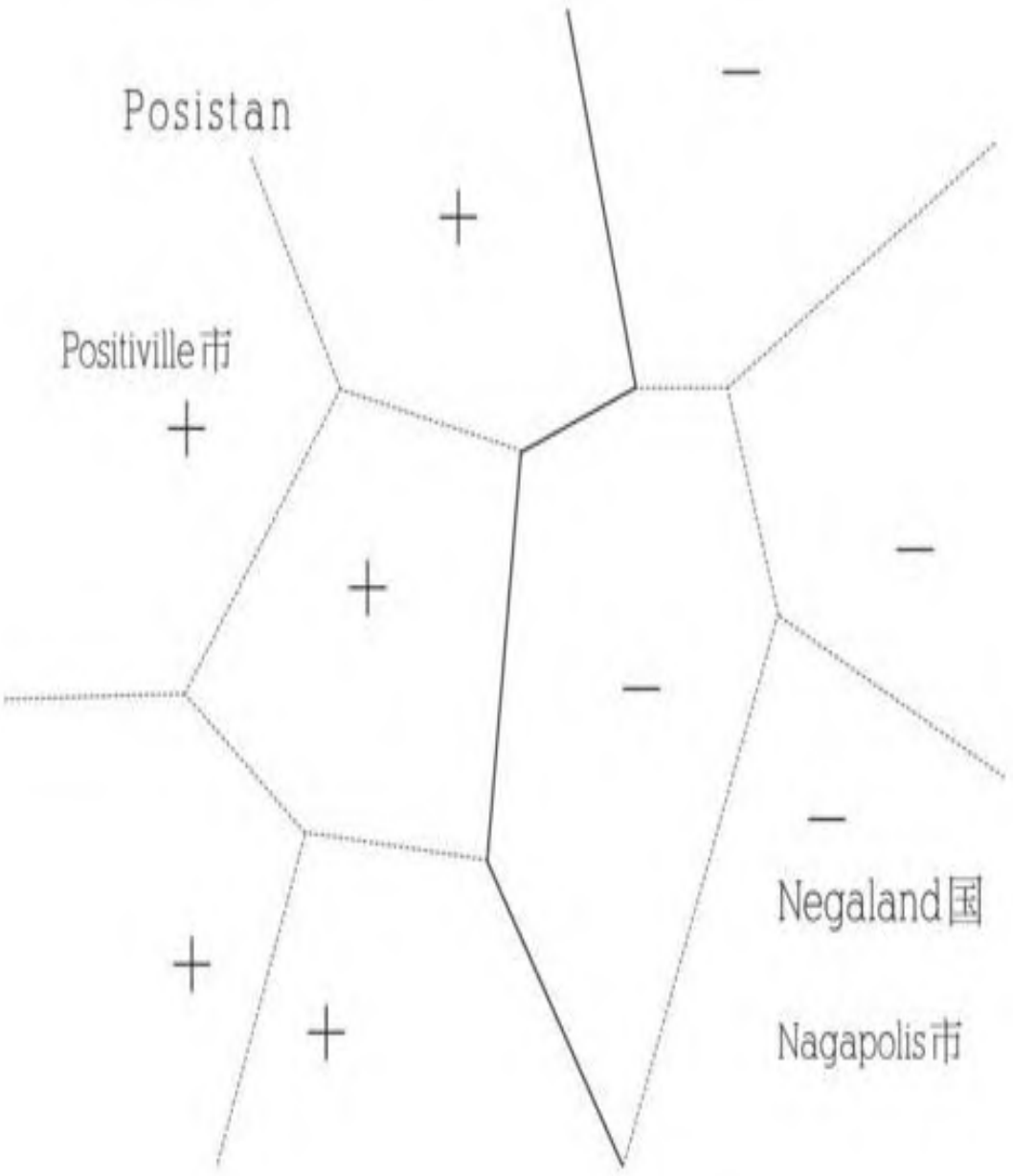


图7－2

最近邻算法可以暗中形成一条非常复杂的边界线，虽然它所做的只是记住每个城市的位置，然后将这些点相应地分给每个国家。我们可以将城市的“城域”看作那些最靠近该城市的点，城域之间的界线在图中用虚线来表示。现在Posistan只是它的所有城市城域的集合，和Negaland一样。相反，决策树（举个例子）也许只会形成不是南北

向就是东西向的边界，可能更不利于找到真正的边界。因此，虽然决策树算法“愿望更加强烈”，在学习期间很努力地想弄明白边界线的位置，但“懒惰的”最近邻算法实际上最后会胜出。

懒惰学习算法会胜出的原因在于，和构建全局模型（例如，决策树）相比，只要每次弄明白指定的点在哪里会较为简单。想象一下，利用决策树来尝试确定什么是脸。你可以说，脸有两只眼睛、一个鼻子和一张嘴，但什么是眼睛，而你在一张图片中怎样才能找到它呢？如果人的眼睛是闭上的，那又会怎样呢？准确定义一张脸，细到每个像素，这是十分困难的，在表情、姿态、背景、光线条件各异的情况下尤甚。而最近邻算法则走捷径：如果数据库中与简上传的最相似的那张照片是脸部照片，那么简的照片也是。要出现这种情况，数据库就得有一张照片，与新照片足够相似，因此，数据库越大越好。对于一个简单的二维问题，比如猜出两个国家的边界，一个小数据库就足够了。对于一个很难的问题来说，如识别脸部，因为每个像素的颜色都是变化的一个维度，所以需要巨大的数据库。目前我们有这些数据库。但对于渴望学习的算法来说，要从这些数据库中学习东西可能代价很大，因为这种算法要明确划分脸部和非脸部的界线。然而，对于最近邻算法来说，界线隐含在数据点和距离度量中，而唯一要付出的就是查询时间。

构建局部模型而非全局模型的相同想法可以应用于分类之外的用途。通常科学家们利用线性回归来预测连续变量，但大多数现象是非线性的。幸运的是，从局部来看，它们是线性的，因为平滑曲线可以局部近似为直线。因此如果你只用直线来对查询点附近的点，而不是尝试对所有的数据都进行拟合，那么你现在就能拥有一个非常强大的非线性回归算法。懒惰也会有回报。如果肯尼迪需要一套完整的国际关系理论来决定该如何应对苏联在古巴部署的导弹，他可能会遇到麻烦。相反，他从这个危机与第一次世界大战的爆发之间找到相似点，而这个相似点指引他做了正确的决定。

最近邻算法可以用来救命，正如史蒂芬·约翰逊在《幽灵地图》中描述的那样。1854年，伦敦爆发霍乱，城市部分地区8个人中就有1个人死亡。当时的理论认为，霍乱由“不良气体”引起，但该理论并没能阻止霍乱的蔓延，约翰·斯诺（一位质疑该理论的内科医师）则有更好的主意。他在伦敦地图上标出所有霍乱病例的位置，然后划分地图，每个区域都距离公共水泵最近。有了！几乎所有死者都位于某个特定水泵的“城域”中，也就是位于苏活区的布罗德大街上。经推断井里的水被污染了，斯诺说服当地人禁止使用该水泵，流行病就消失了。这段插曲孕育了流行病学，而它还是最近邻算法取得的第一次成功——流行病学正式出现还是在近一个世纪之后。

有了最近邻算法，每个数据点就是自己的微型分类器，为所有获取的查询例子预测级别。最近邻算法就像一群蚂蚁，在这个队伍中每只蚂蚁都做得很少，但把力量聚集在一起，它们就可以移动高山。如果一只蚂蚁的负重过大，它可以和周围的蚂蚁共同承担。本着同样的精神，在k最近邻算法（k－nearest－neighbor algorithm）中，测试的例子的分类方法是找到它的k最近邻，然后让它们进行投票。如果距离上传照片最近的图片是一张脸，但接下来最近的两张却不是，第三张最近邻的照片也还是会认为最新上传的不是脸。最近邻算法容易犯过拟合错误：如果数据点的等级是错误的，它会蔓延到整个城域当中。*K*最近邻算法更可靠，因为只有在多数k近邻受干扰时才会出错。当然，犯错的代价就是它的图像会更模糊：边界的细节因为投票而变得模糊。*K*值变大时，方差变小，但偏差变大。

利用多个*k*最近邻而不仅一个近邻，这不是事情的结局。直观来看，与测试例子最接近的例子应该更重要。这让我们引出加权*k*最近邻算法。1994年，明尼苏达州立大学和麻省理工学院的研究人员建立了一个推荐系统，其构建基础是他们所谓的“一个看似简单的想法”：过去人们同意的话，将来他们也还会同意。这个想法直接引出协同过滤系统，所有典型的电子商务网站都有这些系统。假设像网飞一样，你建立了电影评分的数据库，用户对他看过的电影都会给出1~5颗星的

评分。你想确认用户肯恩是否会喜欢《地心引力》，因此你找到那些以往评分与肯恩的评分最相关的用户。如果他们对《地心引力》评分很高，那么可能肯恩的评分也会很高，你就可以把这部电影推荐给他。但是，如果对于《地心引力》他们有不同的看法，你就需要一个回退点，在这种情况下就要根据用户与肯恩关系的密切程度来进行排名。因此，如果李和肯恩的关系比梅格和肯恩的关系密切，那么相应李的评分也会更有价值。肯恩的预测评价就是与其相关的人的加权平均值，每个人的权值就是他与肯恩关系的系数。

虽然如此，但还是有一个有趣的转变。假设李和肯恩有非常相似的品位，但李比肯恩脾气暴躁。当肯恩给5颗星时，李会给3颗星；当肯恩给3颗星时，李会给1颗星，以此类推。我们想用李的评分来预测肯恩的评分，但如果我们直接预测，会总差2颗星。我们要做的就是，在知道李的评分值的基础上，预测肯恩的排名超过或者低于平均值的量。而现在，既然当李总是比他的平均值高出2颗星时，肯恩也会比他的平均值高出两颗星，我们的预测就会准确。

顺便说一下，你不需要显性评分来进行协同过滤。如果肯恩在网飞上预订了一部电影，意味着他可能会喜欢它。这样“评分”也才有可能被预订或未被预订，而两个用户如果预订了很多一样的电影，那么他们品位就会相似。即便只是暗暗点击某个东西，也会代表你对它感兴趣。最近邻算法和以上提到的都会有效。目前，所有种类的算法都被用于为用户推荐项目，但加权*k*最近邻算法是第一个受到广泛运用的算法，而且要打败它仍然很困难。

推荐系统，亦如其名，是大业务：亚马逊1/3的业务来自它为用户推荐的商品，网飞3/4的业务也是如此。它与最初的最近邻算法相差很大，因为它的记忆需求，人们觉得它不够实用。回到那时，计算机的存储器是用铁圈做成的，每个比特一个铁圈，甚至储存几千个例子都很费力。时代改变得多快！虽然如此，不必如此智能，以记住所有你见过的例子，然后搜索这些例子，因为多数例子可能并不相关。如果

你回头看Posistan和Negaland的地图，你可能会注意到，如果Positiville消失，不会有任何变化。附近城市的城域会扩张至以往被Positiville占用的土地，但因为它们都是Posistan的城市，与Negaland的边界还是会一样。真正重要的是横跨边界、与对方国家的城市相交的城市，其他所有的城市都可以忽略。因此，使最近邻算法更加有效的简单方法，就是删除所有被它们的近邻准确分类的例子。这个技巧加上其他技巧可以使最近邻算法得以应用于一些让人意外的领域，例如，实时操控机器人的手臂。但不用说，它们仍然不是诸如高频交易之类事务的第一选择，因为在这些领域中，计算机会在一秒钟内买卖股票。在它与神经网络的竞争中，因为神经网络可以应用于仅含固定数量的加法、乘法、sigmoids函数的例子中，还可应用于需要为例子的最近邻搜索庞大数据库的算法中，因此神经网络肯定会获胜。

研究人员最初之所以对最近邻算法持怀疑态度，是因为它不确定能否找到两个概念之间的真正边界。但1967年，汤姆·科韦尔和彼得·哈特证明，在给定足够数据的情况下，最近邻算法最糟糕时易于出错的概率也仅仅是最佳可行分类器的两倍。举个例子，如果因为数据中的干扰因素，有至少1%的测试例子不可避免地被错误分类，那么最近邻算法保证误差率最多在2%左右。这是一个重大的启示。直到那时，所有已知的分类器都假定边界会有一种非常明确的形式，直线就是典型例子。这是一把双刃剑：一方面，它使准确性的证明成为可能，正如在感知器的例子中那样；另一方面，这也意味着分类器受到它所能掌握东西的严格限制。最近邻算法是史上第一个能够利用不限数量的数据来掌握任意复杂概念的算法。没人能指望它能找到在超空间中、由数百万个例子形成的边界，但因为科韦尔和哈特的证明，我们知道，它们可能不会错得太离谱。据雷·库兹韦尔的观点，当我们无法理解计算机在做什么时，单一性开始产生。依据该标准，说它已经在筹备当中也不稀奇——1951年就已经开始了，当时菲克斯和霍奇斯发明最近邻算法，这个微小的算法就能做到。

维数灾难

在这个伊甸园中当然也有一条蛇，它被称为“维数灾难”。虽然它对所有学习算法都会有或多或少的影响，但它对最近邻算法的危害尤其大。在低维度条件下（比如二维或者三维），最近邻算法通常能很好地起到作用。随着维数的上升，事情就会很快陷入崩溃状态。当今算法要掌握数千甚至数百万个属性并不稀奇。对于电子商务网站来说，它要掌握你的喜好，那么你每点击一下鼠标就算一种属性。网页上的每个词、图片上的每个像素也是如此。即使只有数十或者数百个属性，最近邻算法也有可能已陷入困境。第一个问题就是多数属性是不相关的：关于肯恩你可能知道100万个事实，但很有可能只有几个事实与他得肺癌的风险有关。虽然知道他是否吸烟对于做出该预测很关键，但是也可能对知道他是否喜欢看《地心引力》用处不大。举个例子，符号学派的方法很擅长处理非相关属性：如果该属性不含任何关于等级的信息，那么它就不包含在决策树或者规则集当中。但让人感到无望的是，最近邻算法会受到非相关属性的迷惑，因为这些属性都能够促成例子之间的相似性。有了足够的相关属性，不相关维度中的偶然相似性会清除重要维度中有意义的相似性，而最近邻算法和随意猜测相比也好不到哪里。

让人感到意外的是，有一个更棘手的问题：拥有更多的属性可能也会有害，即使这些属性是相关的。你可能会想，终归信息越多越好。这不是我们这个时代的口号吗？但随着维数的上升，用于确定概念边界的训练样本的数量也会呈指数上升。如果有20个布尔属性，就会有近100万个可能不一样的例子。如果有21个布尔属性，就会有200万个例子，而边界也会有相应数量的方式来定义这些属性。每个额外的属性都会把学习问题的困难程度加倍，而这也只是有布尔属性的情况。如果属性包含很多信息，添加该属性带来的好处可能会多于损失。如果你有的只是能提供很少信息的属性，比如邮件中的词语或者

图片中的像素，那么你可能会遇到麻烦，即使这些属性集中在一起组成的信息足以预测你想要的是什么。

情况会变得更糟糕。最近邻算法的基础是找到相似物体，而在高维度情况下，相似性的概念就会无效。超空间就像过渡区域。在三维空间里的直觉不再适用，怪异离奇的事开始发生。想想一个橘子：一层薄薄的外壳包裹着好吃的果肉。比如橘子90%的半径是果肉，剩下的10%则是果壳，这意味着橘子73%的体积是果肉（0.93）。现在想象一个超级橘子：90%的半径还是果肉，但它在100个维度的空间中。那么果肉的体积已经缩小到超级橘子体积（0.9^{100}）的1/3000。这个超级橘子全都是皮，而并且你绝对无法将其剥开。

另外一个让人不安的例子就发生在我们的老朋友身上——正态分布，又名钟形曲线。正态分布认为，数据本质上就落在一个点上（正态分布的平均值），但其周围也会有模糊的东西（由标准差给出）。对吗？在超空间中不是这样的。在高维度正态分布中，你比较有可能得到远离而不是接近平均值的样本。超空间中的钟形曲线看起来更像甜甜圈，而不像钟。当最近邻算法走进这个颠倒的世界时，它会变得非常困惑。所有的例子看起来都一样，同时因为它们距离彼此太远，无法做出有用的预测。如果你将例子统一地随意分布在高维度的立方体中，大部分例子会更接近立方体的一个面，而不是接近它们的最近邻。在中世纪的地图中，未涉足领域会被标上“龙”、“海蛇”等其他虚构出的生物，或者只标上“此处有怪物”的短语。在超空间中，龙无处不在，你的前门也会有。如果尝试走到你隔壁邻居家，那你永远也到不了那里；你会永远迷失在陌生的陆地上，在这里，所有熟悉的东西都消失不见。

决策树也无法幸免于维数灾难。举个例子，你打算学习的概念是一个球体：球内的点是正的，球外的点是负的。通过放在球体内正合适的最小立方体，决策树可以估算球体的体积。这不是完美的方法，但也不会太差：只有立方体的角才会被误分类。但在高维度空间中，

超级立方体几乎所有的体积都会出现在超球面外。对于你准确将其标注为正的例子，你会错误地将许多负的例子分类为正，这会让你的准确率骤然下降。

实际上，没有哪种算法能够幸免于维数灾难。这是机器学习中，继过拟合之后，第二个最糟糕的问题。“维数灾难”这个术语由理查德·贝尔曼在50岁时提出的，他是一位控制论理论家。他观察到，控制算法在三维空间中可以起到很好的作用，但在高维度空间中则变得效率极低。例如，当你想控制机器人手臂中的每个节点或者化工厂的把手时，这种情况就会出现。但在机器学习中，问题不仅仅在于计算成本——随着维数上升，变得越来越困难的是学习本身。

然而，并不是所有东西都丢失了。我们能做的第一件事就是摆脱不相关维度。决策树会自行做好这一点，方法就是计算每种属性的信息增益，然后只使用最能提供信息的属性。对于最近邻算法来说，我们可以完成类似的事情，方法就是首先丢弃所有那些信息增益低于阈值的属性，然后只在简化的空间中测量相似性。这对于一些应用来说足够快、足够好，但不幸的是，这种方法会阻碍许多概念的学习，像XOR逻辑那样：如果当与其他属性结合时，一个属性只描述了有关类别的一些点，而不是关于自己本身，那么它就会被淘汰。一个代价更大但也更智能的选择，就是围绕学习算法来选择属性，并进行爬山算法研究，该研究会不断删除属性，只要不会降低最近邻算法留存数据的准确度。牛顿也进行过许多次的属性选择，当时他确定要预测物体的轨迹，最重要的就是要知道它的质量——不是颜色、气味、年龄，或者无数其他的属性。实际上，方程式最重要的东西，就是未在方程式中出现的所有数量：一旦我们知道这些要点是什么，要弄清楚这些要点如何相互依赖，就变得更加容易了。

要处理弱相关的属性，一个选择就是掌握属性权值。我们不会让所有维度下相似性的重要性相等，而是“缩减”不那么相关的属性。假设训练例子是房间中的点，那么高度尺寸对于我们的目标就没那么

重要了。抛弃这一点会将所有例子投到地板上。调低高度尺寸又更像给屋子加了一层更低的天花板。当计算某点到其他点的距离时，该点的高度仍有价值，但没有它的水平位置那么重要。和机器学习中的许多事情一样，我们可以通过梯度下降来掌握属性权值。

一间屋子可能会有很高的天花板，但数据点都会在地板附近，就像一层薄薄的灰尘落在地板上。在这种情况下，我们是幸运的：这个问题看起来似乎和三维有关，其实它更接近二维。不必缩减高度，因为自然已经为我们缩减了。这个非一致性往往能够挽救局面，通过这一点数据在（超）空间中就可以均匀分布了。例子可能会有1000种属性，但实际上，它们都“生存”在维度低很多的空间中。这使最近邻算法有助于识别手写体数字，例如，每个像素都是一个维度，因此会有很多维度，但在所有可能的图片中只有一小部分是数字，所有图片都一起待在超空间小小的舒适角落里。但是，数据“生存”的低维度空间可能会反复无常。例如，如果一间房有家具，灰尘就不会落在地板上，而会落在桌面、椅面、床罩、古董架上。如果我们能弄明白灰尘覆盖地板的大致形状，那么我们需要的就是地板上每个点的坐标。在第八章中，我们会看到机器学习中会有整个分支致力于（可以这么说）在超空间的黑暗中摸索，以找到地板的形状。

空中蛇灾

直到20世纪90年代，应用范围最广的类比学习算法就是最近邻算法，但后来被来自其他学派更引人注目的“表亲”夺去光芒。当时一种新的以相似性为基础的算法横空出世了，横扫之前的所有算法。实际上，你可以说它是自冷战结束以来的另一个“和平红利”。它就是支持向量机（Support vector machines，SVM），是弗拉基米尔·万

普尼克（一位苏联频率论研究者）的创意。万普尼克的大半生都在莫斯科的控制科学研究所度过，1990年，随着苏联走向解体，他移居美国，在那里他加入传说中的贝尔实验室。虽然在苏联，万普尼克很多时候都满足于从事理论、纸笔工作，但在贝尔实验室的氛围则不一样。研究人员正在寻找实践结果，而万普尼克最终决定将其想法变成算法。在几年时间内，他和贝尔实验室的同事已经研发出支持向量机。不久之后，支持向量机就变得无所不在，并创下不同的准确度纪录。

表面上看，支持向量机看起来很像加权k最近邻算法：正类别与负类别之间的边界由一组例子、其权值加上相似性测度来确定。测试实例会归入正类别，条件是从平均水平上看，它看起来更像正面例子而不是负面例子。平均数会被加权，而支持向量机只会记住那些用于确定边界的关键例子。如果你回头看Posistan和Negaland的例子，一旦我们丢掉那些不在边界上的城市，剩下的就是这张地图（见图7－3）。

这些例子被称为支持向量，因为它们是“支撑”边界的向量：移动一个向量，边界的一段就会滑向其他不同的地方。你也许还会注意到，边界就是一条锯齿状的线，实例的确切位置会决定突然出现的拐角。真实的概念趋向于拥有更加平缓的边界，这意味着最近邻算法的估算可能不理想。但有了支持向量机，我们就可以掌握平缓的边界，看起来更像图7－4。

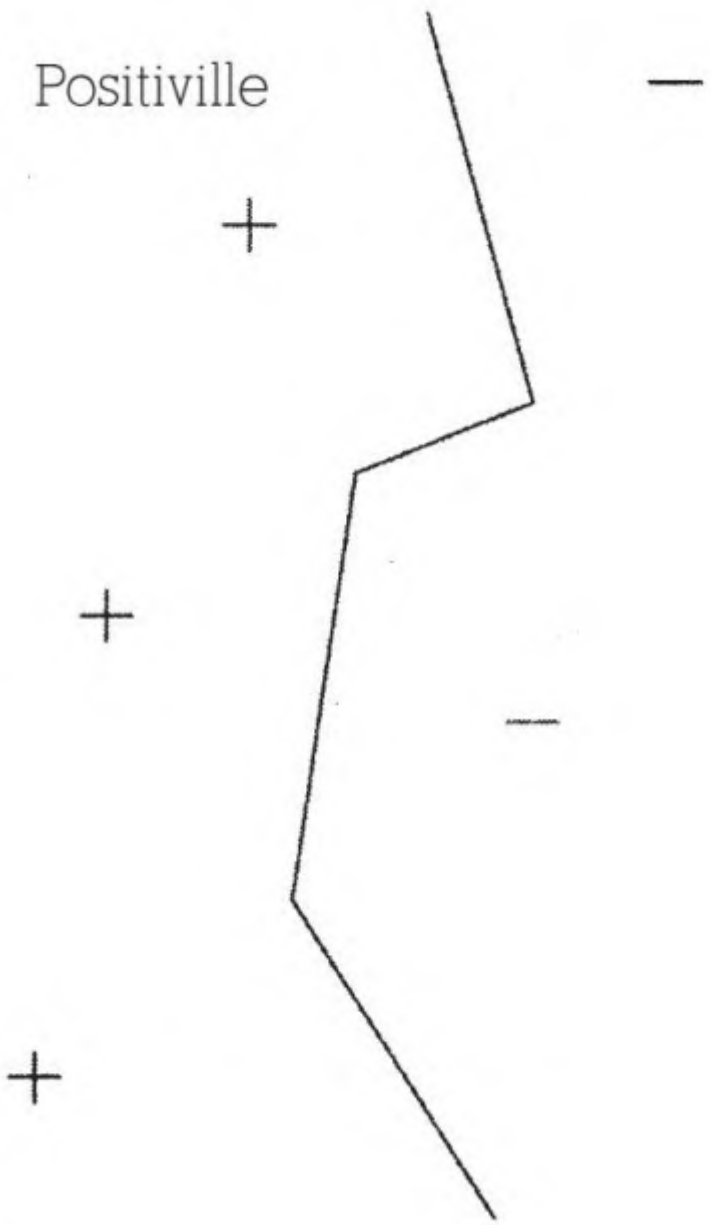


图7－3

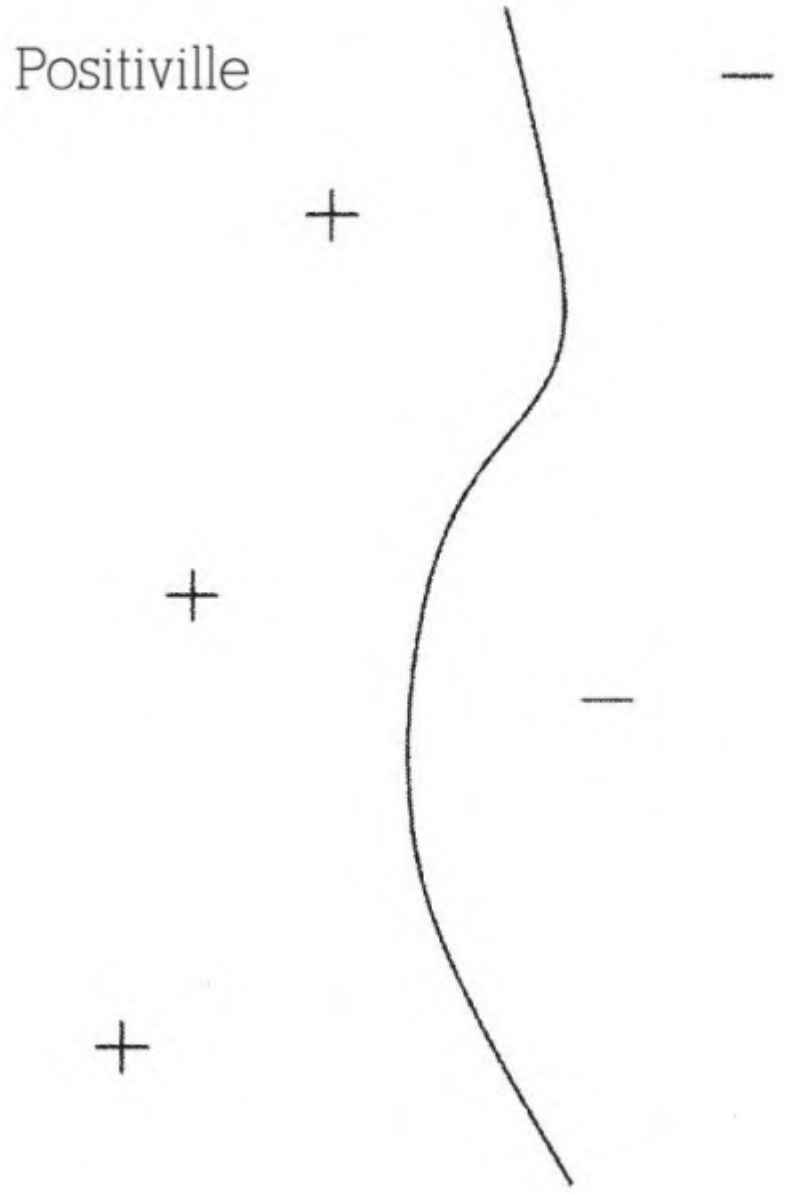


图7－4

为了学习支持向量机，我们需要选择向量和它们的权值。相似性度量，在支持向量机领地被称为“核心程序”，通常被归为先验性。万普尼克的一个重要观点就是，并不是所有将正面训练例子从负面训练例子分离出来的边界都是平等的。假设Posistan和Negaland发生战争，它们之间隔着无人区，两边都是布雷区。你的任务就是调查无人

区，从无人区的这头走到那头，不能踩到任何地雷。幸运的是，你有一张地图，告诉你地雷都埋在哪里。显然，你不会走旧路线：你会假设地雷可能在的位置很宽阔。这就是支持向量机所做的事，实例相当于地雷，要掌握的边界相当于选择的路线。边界距离实例最近的地方就是它的安全边际，支持向量机选择支持向量和权值，这些向量和权值能够产生可能的最大边际。例如，图7－5中的实线边界比虚线要好。

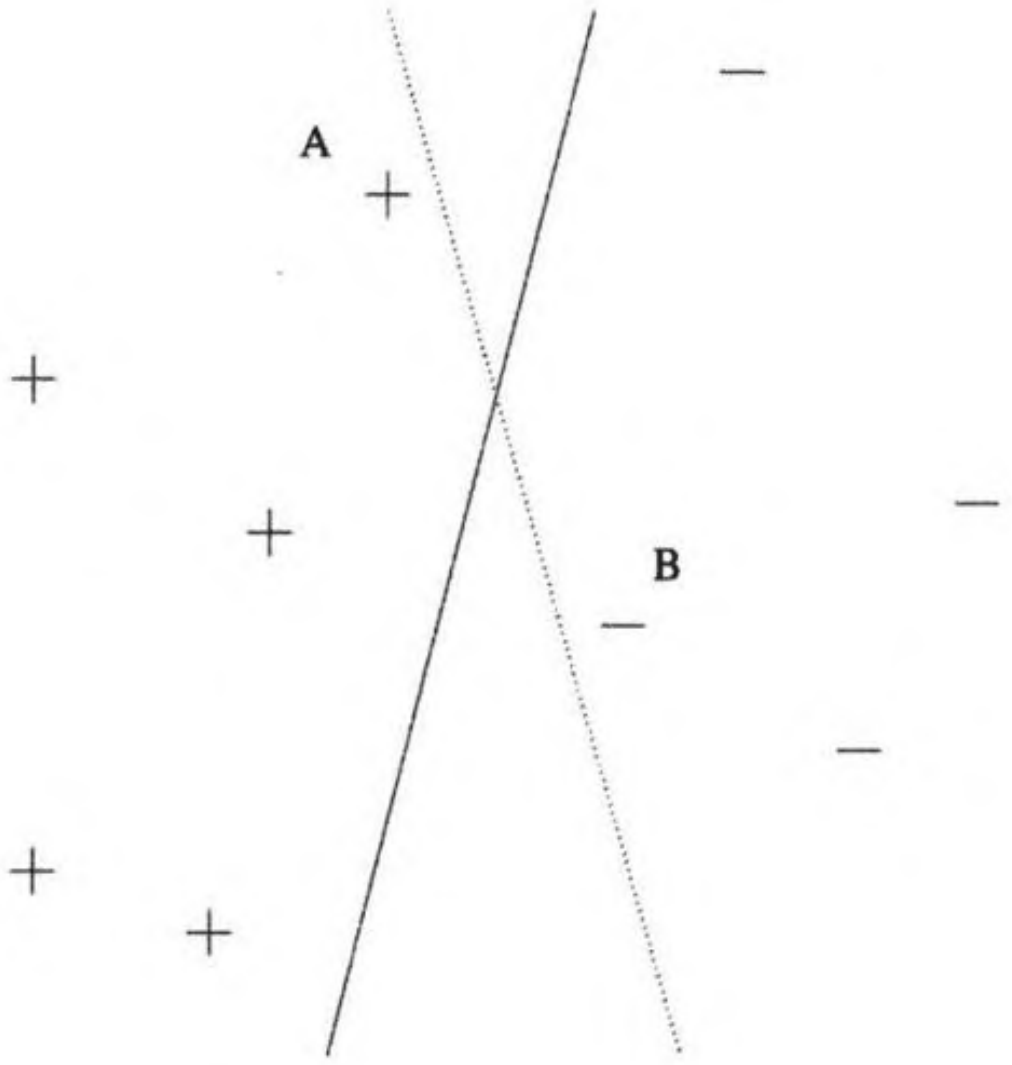


图7－5

虚线边界能够较好地将正面例子和负面例子分离，但很危险的是，它差点接触到A处和B处的地雷。这些例子都是支持向量：删除其中的一个，最大边际边界就会移到不同的地方。通常，边界当然也可能是弯曲的，这使得边界更加难以想象，但我们可以将边界想象成爬向无人区的蛇，边际表示这条蛇可能有多肥。如果很肥的蛇能够一直蜿蜒爬行而不会被地雷炸得粉碎，那么支持向量机就可以很好地将正面例子和负面例子分离。万普尼克表明，在这种情况下，我们可以确信，支持向量机不会过拟合。直观地说，和一条瘦的蛇相比，肥的蛇能够不触雷爬行的方法更少。同样，和低边际的支持向量机相比，高边际的就不太可能因为画了一条过于复杂的边界而过拟合。

事情的第二部分就是支持向量机如何找到最合适的蛇，刚好夹在正面雷区和负面雷区之间。乍一看，通过梯度下降，为每个训练实例掌握权值可能会成功。我们要做的，就是找到能够将边际最大化的权值，而那些最终结果为0权值的例子则被抛弃。不幸的是，这样做只会让权值不受限制地增长，因为从数学的角度讲，权值越大，边际越大。如果距离地雷只有1英尺，然后你将所有东西的尺寸都扩大一倍，包括你自己，你现在距离雷区有2英尺，但这样并不会降低你踩到地雷的可能性。相反，我们得在限制范围内将边际最大化，该范围内，权值智能增长到某个固定值。或者，同样的效果，我们可以在限制范围内将权值最小化，该范围内所有例子都会有一个给定的边际，该边际有可能是1——确切来说是任意的。这就是通常支持向量机做的事情。

约束优化是在将受限函数最大化或者最小化时遇到的问题。宇宙会最大化熵，而熵遵守能量守恒定律。这类问题在商业和技术领域会经常碰到。例如，我们可能会想把工厂生产的小部件数量最大化，这个部件的数量可由机床的数量、小部件的规格等决定。有了支持向量机，约束优化对机器学习来说也会变得重要。无约束优化正在登上顶峰，而这就是梯度下降（或者，在本例子中为上升）要做的事情。约束优化就是当你停在公路上时，尽可能往高处走。如果公路延伸到顶峰，约束及无约束问题的解决方法就会一样。即便如此，更多时候，

公路弯弯曲曲围绕着山峰，接着还没到山顶就绕回来了。你知道自己已经到达公路的最高点，因为在不脱离公路的情况下，你已经无法开到更高的地方。换句话说，也就是到达顶点的路线与公路成直角。如果公路和到达顶点的路线形成斜角，那么沿着公路开远些，你总能到达更高点，即使那样做与直线到达峰顶相比，并不会那么快。因此，解决约束优化问题的方法，不是沿着坡面走，而是走与约束面平行的那部分坡面（在该例子中为公路），然后当该部分为零时，停下来。

通常，我们得一次性处理许多个约束条件（在支持向量机的情况下，每个条件对应一个例子）。想象一下，你想尽可能地靠近北极，但你不想离开自己的房间。每间房的四面墙就是一个限制条件，而解决办法就是要跟着指南针直到你碰到一个角落，这个角落是东北部和西北部两面墙的汇合点。我们说这两堵墙是有效约束条件，因为它们阻止你实现最佳效果，也就是北极。如果你的房间有一堵墙的外部直面北部，那么这就是唯一的有效约束条件，而解决方法就在于这堵墙中间的一个点。如果你是圣诞老人，而你的房间已经处在北极上方，而所有的约束条件都无效，那么你就可以在那儿坐等，仔细考虑玩具的最优分配问题（旅行推销员更容易将其与圣诞老人比较）。在支持向量机中，有效约束条件就是支持向量，因为它们的边际已经是经允许达到的最小值，移动边界会违背一个或多个限制条件。所有其他的例子都不相关，而且它们的权值都是零。

在现实中，我们通常会让支持向量机违背一些限制条件，这就意味着将一些例子错误分类，或者利用更少的边际，否则它们就会过拟合。如果在正面区域中间的某个地方有干扰作用的负面例子，那么我们不想让边界在正面区域周围环绕，目的只是为了让例子正确。但支持向量机会因为它弄错的每个例子而受到惩罚，这样就会促使它把错误例子降到最低。支持向量机就像《沙丘》中的沙虫：大而强韧，可以在爬过雷区时幸免于几次但不是很多次爆炸。

四处寻找应用的同时，万普尼克和他的同事很快偶然发现了手写体数字识别，他们在贝尔实验室的联结学派同事就是研究这个方面的专家。让所有人惊讶的是，支持向量机和感知器一样，在其他领域也能很好地工作，感知器多年来已经被精心设计用于数字识别。这为两者持续时间长、波及范围广的竞争奠定了基础。支持向量机可以当作感知器的一个概括版，因为当你利用某个特定的相似性度量时，得到的就是类别之间的超平面边界（向量之间的点积）。但和多层感知器相比，支持向量机有一个重要优势：权值有单个最优条件而不是多个局部最优条件，因此可靠地掌握它们变得简单多了。虽然如此，支持向量机的表现力依旧不比感知器差。支持向量有效地扮演隐藏层的角色，而它们的加权平均值则起到输出层的作用。例如，一台支持向量机可以轻易表示XOR函数，方法就是为4个可能的配置分配一个支持向量，但不较量一下，那些联结主义学者便不会罢休。1995年，拉里·杰克尔（贝尔实验室万普尼克所在部门的主任）和他赌一顿丰盛的大餐——到2000年，神经网络就会和支持向量机一样好理解。他输了，但作为回报，万普尼克打赌——到2005年就没有人会使用神经网络，后来他也输了（唯一吃到免费大餐的是燕乐存，他们的见证人）。另外，随着深度学习的出现，联结学派已经重新占据优势。只要你能掌握它们，含有多层的网络可以比支持向量机用更加简洁的方式来表达许多函数，而支持向量机一直以来只有一层，这一点就使两者大不相同。

支持向量机早期较大的功绩还在于文本分类，后来证明这是一项大实惠，因为网站那时候才人气渐增。当时，朴素贝叶斯是最先进的文本分类器，但当语言中的每个词都代表一个维度时，甚至它也开始过拟合了。它需要的只是一个词，这个词会偶然出现在训练数据中所有与体育相关（举个例子）的页面中，而不是出现在其他页面中，而朴素贝叶斯开始出现幻觉，以为只要包含该词的页面都是体育页面。但多亏了边际最大化，支持向量机即使在很高的维度中也可以抵抗过拟合。

通常，支持向量机选择的支持向量越少，就能更好地进行概括。任何不是支持向量的训练例子可能会被正确分类，条件是它显示为测试实例，因为正面和负面例子之间的边界仍在同样的地方。因此预测支持向量机的误差率，最多就是支持向量的那部分例子。随着维数的上升，这部分也会上升，因此支持向量机无法幸免于维数灾难，但它们最能抵抗这个灾难。

除了实践中的功绩，支持向量机还完全改变了许多机器学习中的传统观点。例如，它揭穿了“模型越简单就越准确”这个谎言，有时候，这个观点会被误以为是奥卡姆剃刀理论。相反，支持向量机可能有无限数量的参数，而且只要它有足够大的边际，就不会过拟合。

然而，支持向量机唯一最让人吃惊的属性就是，无论它形成多么弯曲的边界，那些边界也总是直线（或者一般为超平面）。这并不矛盾，因为直线存在于不同的空间中。假设例子停在（x，y）的平面上，而正面和负面区域之间的边界是抛物线y=x2。没有什么方法能用直线来表示它，但如果我们添加坐标z，意味着现在数据存在于（x，y，z）空间，然后我们设定每个例子的z坐标为它的x坐标的平方，这时边界就是由y=z定义的斜平面。实际上，当数据点上升至三维空间时，有些数据点上升的数量正合适，但比其他数据点上升的要多，而急板（presto）在这个新维度中，正面例子和负面例子可能会被一个平面分开。原来，我们可以把支持向量机利用核心程序、支持向量、权值来做的事情，看作将数据映射到更高维数的空间中，并在那个空间找到最大边际超平面。对于一些核心程序来说，衍生空间有无限的维数，但支持向量机完全不会因此而受到困扰。超空间可能是过渡区域，但支持向量机已经弄明白如何操纵它。

爬上梯子

两个东西如果在一些方面意见一致，那么它们就是相似的。如果它们在一些方面意见一致，可能在其他方面也会意见一致，这就是类比的本质。它还表明了类比推理中的两大子问题：弄明白两个事物的相似度，确定由它们的相似度还能推导出什么。到目前为止，我们已经探索了类比“低功耗”的一端，包含诸如最近邻和支持向量机之类的算法，在这种情况下，这些问题的答案都非常简单。它们的应用最为广泛，但只用一章来介绍类比学习并不完整，因为至少得浏览该领域更为强大的部分。

在任何类比学习算法中，最重要的问题就是如何度量相似性。它可以如数据点之间的欧几里得距离那么简单，也可以和含有多水平子程序的一整套程序那么复杂，而且这些子程序的最终产出是相似度值。不管怎样，相似度函数控制的是学习算法如何从已知例子归纳出新的例子。我们将自己对问题域所了解的东西插入学习算法中，这就变成类推学派对于休谟问题的回答。我们可以将类比学习应用到所有种类的物体中，而不仅仅是属性向量，只要我们有度量它们之间相似度的方法。例如，我们可以通过两个分子之间包含相同子结构的数量，来测量两个分子的相似度。甲烷和甲醇相似，因为它们都有三个碳氢键，而不同点仅在于一个氢原子代替了一个羟基（见图7－6）。

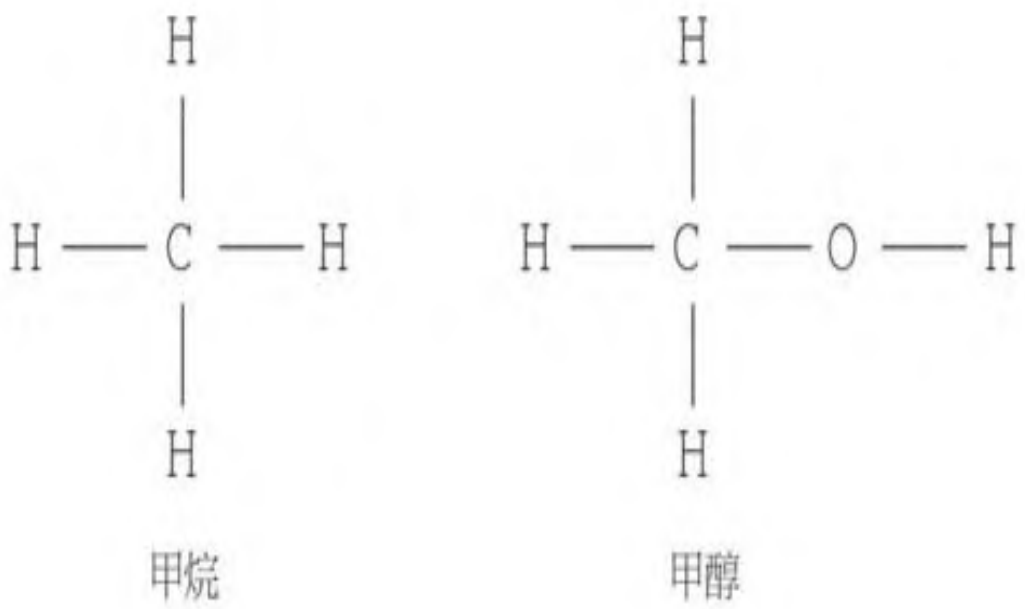


图7－6

然而，这并不意味着它们之间的化学行为是相似的。甲烷是气体，而甲醇是酒精。类比推理的第二部分，就是弄明白在已经发现的相似点的基础上，如何推导出新的东西。这可以很简单，也可以很复杂。在最近邻算法或者支持向量机中，这点包括在最近邻算法或者支持向量类别的基础上，预测新东西的类别。但在案例推理（类比学习的另外一种类型）中，输出的可能是由检索对象组成部分形成的复杂结构。假设你的惠普打印机吐出没用的东西，你会打电话给求助台。他们之前很有可能已经遇到过很多次你这样的问题，所以好办法就是找到那些记录，然后从这些记录中找到可能解决你的问题的方法。这不只是找到与你的问题有许多相似之处的投诉那么简单：例如，和你的打印机一起使用的是Windows系统，还是Mac OS X？这可能会引起系统的不同设置，打印机也会变得相关。一旦你找到最相关的例子，解决你的问题所需步骤的顺序，可能是不同案例中不同步骤的组合，可能还会针对你的问题进一步微调。

求助台是当今案例推理最受欢迎的应用。多数求助台仍采用人力中介，但正畸诊疗软件IPsoft的虚拟求助台员工伊丽莎则与顾客直接对话。伊丽莎配有交互式视频人物角色，迄今已经为2000多万的顾客解决问题，这些问题大部分来自一流的美国公司。“来自机器人王国（Robotistan）的问候，外包最实惠的新选择”是近来外包博客对它的评价。而且，正如外包在不断爬技能这架梯子，类比学习也是如此。第一个以判例为基础、为指定判决辩论的机器人律师已经诞生。这样的系统能够准确预测超过90%经其审核的商业秘密案件。也许在未来的网络法院中，亚马逊云服务上的某处，在开庭期间，机器人律师会推翻机械战警给你的无人驾驶汽车开出的罚单，当你走到沙滩上时，莱布尼茨将所有辩论简化为计算的梦想最终会实现。

可以说，站在技能这架梯子更高处的是音乐创作。大卫·科普是加州大学圣克鲁斯分校的名誉教授，他设计出一种算法，能够通过选择并重组著名作曲家作品中的一些片段，创造出带有这些作曲家风格的新的音乐。我在几年前参加过的一个会议上，他演奏了“莫扎特”的三个片段：一段来自真正的莫扎特，一段出自人类作曲家模仿的莫扎特，一段来自他自己的系统。然后他让听众为真正的莫扎特投票。沃尔夫冈胜出了，但计算机打败了人类模仿者。这是一场人工智能会议，听众们感到非常喜悦。在其他项目中则没有那么开心了，一位听众愤怒地控诉科普，说他毁掉了他的音乐。如果科普是对的，创造性（最深不可测的东西）都可以归结为类比和重组。你可以通过谷歌搜索“david cope mp3”来自行判断。

然而，类比学者最棒的技巧在于跨越问题域来进行学习。人类一直在做这件事情，比如一位高管可以从一家媒体公司跳槽到另一家消费品公司，而不用从零开始，因为许多相同的管理技能仍然适用。华尔街雇用很多物理学家，因为物理和金融问题虽然表面上看区别很大，却往往包含相似的数学结构。然而，假如我们对目前为止见过的学习算法进行训练，让它们预测布朗运动，然后对股市进行预测，那么这些算法将无法达到预期效果。股票价格和悬浮在液体中的粒子的

速度只是不同的变量，所以学习算法甚至不知道从何开始。但利用构造映射，类比学者可以完成这件事，这是戴德瑞·根特纳（西北大学的一位心理学家）发明出的算法。结构映射采用两种描述方法，发现了它们某些部分和关系的一致对应关系，然后基于这种对应关系，进一步将一个结构中的属性转移到另一个结构中。例如，如果是太阳系和原子的结构，我们可以将行星映射为电子，将太阳映射为原子核，并和玻尔一样得出结论，认为电子围绕原子核旋转。当然，真理会更加微妙，而往往我们做出类比之后，又得重新定义它们。但能像这样从单个例子中学习，当然是通用学习算法的一个关键属性。当我们与一种新的癌症做斗争时，这样的事一直在发生，因为癌症会不断变异，我们为之前的算法掌握的模型将不再适用。我们也没有时间从大量病人中收集关于新型癌症的数据，也许只有一个病人，而他也需要救治。那么我们最大的希望就是，将新型癌症与已知癌症进行对比，然后努力找到另外一种癌症，它与此种癌症的行为足够相似，那么一些相同的疗法就会起作用。

有类比无法做到的事吗？道格拉斯·霍夫斯泰特认为没有，他是一位认知科学家，还是《哥德尔、艾舍尔、巴赫：集异璧之大成》的作者。霍夫斯泰特看起来有点像格林奇的双胞胎兄弟，可能是世界上最著名的类比学者了。在《表面与本质：类比作为思维的燃料和火焰》一书中，霍夫斯泰特和他的合作者伊曼纽尔·桑德尔激烈地争论，认为所有智能行为都可以归结为类比。我们学习或者发现的每样东西，从日常词语如“妈妈”和“玩耍”的含义，到诸如阿尔伯特·爱因斯坦、埃瓦里斯特·伽罗瓦这些天才惊人的洞察力，这些都是类比应用于实践中的结果。当小蒂姆看到妇女像他妈妈照顾他一样照顾其他孩子时，它概括了“妈妈”的概念来指其他人的妈妈，而不只他的妈妈。这反过来又是一个跳板，用来理解诸如“母舰”和“大自然”等东西。爱因斯坦的“最快乐思想”（该思想还衍生了广义相对论），就是重力与加速度之间的类比：如果你是一台升降机，你无法判断你的重量是因为这个还是那个而产生，因为它们的效果是一样

的。我们在类比的广阔海洋里遨游，我们自己会操纵，但无意间也会被操纵。书中的每页都会包含类比（比如本节的标题，或者前一节的标题）。《哥德尔、艾舍尔、巴赫》是哥德尔定理、艾舍尔艺术、巴赫音乐的一个延伸类比。如果终极算法不是类比，那么它肯定就是和类比相似的东西。

起床啦

认知科学见证了符号学派与类推学派之间很长一段时间的争论。符号学派指向他们能够模仿但类推学派无法模仿的东西；接着类推学派弄明白如何做到这一点，然后想出他们能够模仿但符号学派无法模仿的东西，这个循环一直重复下去。基于实例的学习——有时人们这样称呼它，应该更有利于模仿我们如何记住生活中的特定片段；规则是包含抽象概念（如“工作”“爱”）推理的假定选择。但我在读研究生时，觉得这两者真的只是连续区间中的点，而我们应该能够掌握它的所有方面。实际上，规则就是概括的实例，这种情况下我们已经“忘记”一些属性，因为它们并不重要。相反，实例则是非常特殊的规则，每种属性都会有一个条件。在我们的一生中，相似的片段会渐渐被抽象成基于规则的结构，比如“在餐馆吃饭”。你知道去餐馆会涉及从菜单里点菜和给小费，而每次你出去吃饭的时候，都会遵守这些“行为准则”，但你可能不记得自己意识到这些准则时的第一家餐厅。

在博士论文中，我设计了一种算法，以这种方式来将基于实例的学习和基于规则的学习结合起来。一条规则不会只和满足它所有先决条件的实体匹配，与任何其他规则比，它会与其更加相似的实体匹配。在这个意义上讲，它会更接近于满足它的条件。例如，胆固醇浓

度为220毫克/分升的人，会比浓度为200毫克/分升的人更加符合以下规则：如果胆固醇浓度超过240毫克/分升，你会有心脏病发作的风险。RISE是我称呼该算法的名字，通过以每个训练实例作为开始来学习，然后渐渐概括每个例子，以吸收最近的例子。最终的结果通常是通则的结合，这些通则会在它们之间和多数例子匹配，越来越多的特殊规则用来将例外规则与那些规则匹配，直到到达特定记忆的“长尾”上。RISE比当时最好的规则和实例学习算法预测得还要准确，而我的实验也表明，准确地说，这是因为它结合了两者的最佳特征。规则可以类推地进行匹配，所以它们不会再那么脆弱。实例可以选择空间中不同区域的不同特点，然后比最近邻算法更能与维数灾难相抗衡，而最近邻算法只能到处选择相同的特征。

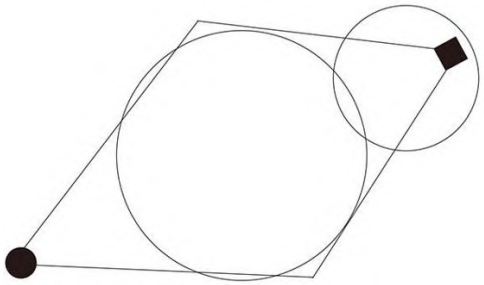
RISE是通往终极算法的一个步骤，因为它将符号与类比学习结合起来了。然而这只是小小的一步，因为它不具备这两个范式的全部能力，仍缺少其他三个范式的能力。RISE的规则无法以不同的方式连接起来，每个规则只能直接从其属性中预测例子的类别。另外，这些规则无法一次讨论一个以上的实体；例如，RISE无法表达这样的规则：如果A得了感冒，而B与A有接触，那么B也可能会得感冒。在类比这边，RISE仅仅概括了简单的最近邻算法，它无法利用结构映射或者某个这样的策略来进行跨域学习。在完成博士论文时，我找不到能够为一种算法带来所有5个范式能力的方法，然后我把这个问题丢在一边一段时间。但当我把机器学习应用到诸如口碑营销、数据集成、事例编程、网站个性化等问题中时，我不断发现，每个这样的范式只提供了一部分解决办法。我相信应该有更好的方法。

至此，我们已完成了对5个学派的介绍，同时集中了他们的观点、议定它们的界限、探索了这些碎片如何拼凑在一起。我们现在知道的东西比一开始多得多，但还有一些东西没有找到。在这个谜题的中心有一个很大的漏洞，使人难以看到其模式。问题在于，目前为止，我们见过的所有学习算法，需要一位老师来告诉它们正确答案。它们不会从健康细胞中区分出癌细胞，除非有人将这些细胞标记为“癌细

胞”或者“健康细胞”。人类会无师自通，从出生那天开始，就已经这么做了。正如《指环王》中站在魔多大门的弗罗多那样，如果无法在这个障碍附近找到出路，那么我们的长途旅行将毫无意义。有一条路可以穿越壁垒和层层守卫，胜利越来越近。跟我来……

第八章

无师自通



如果你是家长，在你孩子长到3岁这段时间，关于学习的整个奥秘会在你眼前揭开。新生儿无法说话、走路、识别物体，甚至不知道，当他看不到某个东西时，它仍然存在。但一个月又一个月过去，大步小步向前走，经过不断摸索，终于取得概念理解上的大进步，孩子终于理解世界如何运转、人们如何行事，以及如何进行沟通。孩子过了3岁生日，所有这些学习行为就会聚合成稳定的状态，即一种贯穿我们一生的意识流。较大的孩子和成年人可以进行时空旅行，也就是记住过去的事，但也只能回忆到3岁的事。如果我们能重新回忆婴儿和蹒跚学步时期的自己，然后从新生儿的角度看待这个世界，那么许多关于学习的疑问，甚至关于存在本身，都会突然变得清晰明朗。但实际上，宇宙中最大的奥秘并不在于它如何开始和灭亡，或者又是由怎样无穷细的线编织而成，而在于幼小孩子的大脑里正在发生什么：一磅的胶状物如何变成意识的所在处。

对于儿童学习的科学研究尚不成熟，几十年前才正式开始，却已经研究得很深入。婴儿无法填写问卷或者遵守实验规则，但我们可以

推导出关于他们大脑活动的、让人惊讶的数量，方法就是在实验过程中，对他们的反应进行录制并研究。连贯的图片就出现了：婴儿的大脑不仅能揭开一项预设的基因程序，或者是用于记录感测数据中相互关系的生物装置，还可以积极整合他们自身的实际情况，这种情况会随时间发生彻底改变。

渐渐地，与我们最为相关的认知科学家以算法的形式来表达关于儿童学习的理论。许多机器学习研究人员由此得出灵感：我们需要的一切都存在于孩子的大脑中，如果能用编码的形式来获取其精华就好了。一些研究人员甚至认为，创造智能机器的方法，就是要造出一个机器婴儿，让它像人类婴儿那样去体验世界，研究人员就是它的父母（甚至可能还会有来自众包的协助，赋予“地球村”这个术语以全新的含义）。小罗比，我们就这么叫它吧（为了向《禁忌星球》中胖乎乎且高大的机器人致敬），是我们造出的唯一的机器人婴儿。一旦它学会3岁小孩都知道的事情，那么人工智能的难题就得到解决了。我们可以把它脑子里的内容复制到其他机器人中，想复制几个都可以，这样最困难的工作就完成了。

当然，问题在于罗比出生时，脑子里该运行哪种算法。受到儿童心理学影响的研究人员会看不惯神经网络学，因为一个神经元的微型运作与孩子最基本行为的复杂度也相差十万八千里，这些行为包括伸手去拿东西、抓住东西，然后睁大眼睛好奇地观察它。我们需以更抽象的水平来模拟儿童的学习，以免只见树木、不见森林。首先，虽然孩子理所当然从父母那里获得了许多帮助，但很多时候他们是在没人监督的情况下进行自学，这才是最不可思议的地方。目前为止，我们见过的算法还无法做到这一点，但我们会发现有几种算法可以让我们看到离终极算法更进一步了。

物以类聚，人以群分

我们轻轻按下“打开”按钮，罗比的视频眼第一次睁开了，但马上它就淹没在威廉·詹姆斯令人记忆深刻的所谓世界的“杂乱无章”中。随着新图片以每秒几十张的速度涌入视线，它必须做的第一件事就是学会将这些图片组织成更大的图片块。真实的世界由持续存在一段时间的物体组成，而不是由从这一刻到下一刻随意变换的像素组成。妈妈走开时，她还是她，不会被更小的妈妈代替。把一个盘子放在桌上，桌面不会出现白色的洞。如果一只泰迪熊从屏幕里经过，然后再出现时变成一架飞机，那么幼小的婴儿不会对此感到惊讶；但一岁小孩却会。不知为什么，他弄明白了泰迪熊和飞机不一样，而且无法自行转化。之后不久，他会弄明白有些物体和其他物体更像，然后开始进行分类。假定有一大堆马形玩偶和铅笔可以玩，9个月大的孩子不会将其分类成马形玩偶和铅笔两个范畴，但一个18个月的孩子却会。

将世界组织成客体和类别，对于成年人来说，这是第二天性；对于婴儿来说却不是；对于机器人罗比来说，更没有可能。我们可以以多层感知器的形式赋予它视觉皮质，然后为它展示分类好的世界里所有的客体和类别的例子——妈妈走近了，妈妈走远了，但我们绝不可能做到。我们需要一种能够自发将所有相似物体或者同一物体的不同图片集中起来的算法。这就是聚类问题，这在机器学习当中也是人们研究最多的主题之一。

一个集群相当于一组相似的实体，或者以最低限度来说，一组相似的实体之间的相似度比和其他集群中的组成部分的相似度要大。对事物进行聚类，这是人类的天性，也是获取知识的第一步。当我们仰望夜空时，会禁不住去看星座，然后根据形状来为其命名。注意到某些元素组有很相似的化学属性，这是发现元素周期表的第一步，那些相似元素现在就是元素周期表中的一列。我们察觉到的一切就是一个集群，从朋友的脸到说话声。没有这些，我们就会不知所措：在学会识别组成语言的独特声音之前，儿童无法学习语言。这需要他们在一岁时学习，而且如果没有单词指代的真实事物群集，他们学习的所有

单词就会没有意思。面对大数据（大量的事物），我们首先采用的办法就是把大数据分成若干更好处理的集群。比如，整个市场过于粗化，单个消费者又过于细微，因此市场营销人员会将市场分成若干部分，这些部分就成为集群。甚至物体本身也是观察的底层集群，从由不同角度照射到妈妈脸上的光线，到宝宝听到的关于“妈妈”一词的不同声波。没有物体，我们就无法进行思考，也许这也是量子力学如此非直观的原因：我们想把亚原子世界形象化，看作粒子碰撞或者波形干扰的世界，但这样也并不真实。

我们可以通过集群的典型元素来表示该集群：你用心灵之眼看到母亲的形象，或者典型的猫、跑车、郊区住宅和热带海滩。依据市场营销知识，伊利诺伊州的皮奥瑞亚是典型的美国小镇。鲍勃·伯恩斯是康涅狄格州温德姆小镇的一名房屋维修主管，今年53岁，是美国最普通的市民——如果你相信凯文·奥基夫《普通美国人》一书讲的是对的，那么至少是这样的。任何由数字属性描述的东西，比如人们的身高、体重、围度、鞋码、头发长度等，使得计算平均数变得简单：他的身高就是所有集群成员的平均身高，他的体重就是所有体重的平均数……对于分类属性来说，比如性别、头发颜色、邮政编码或者最喜爱的运动，“平均”就是出现频率最高的值。这组属性描述的普通成员可能不是真实存在的人，但不管怎样，它是十分有用的参考：如果你在进行头脑风暴，考虑如何营销新产品，把皮奥瑞亚当作产品发布的地点，或者把鲍勃·伯恩斯当作目标客户，这要比考虑“市场”或者“消费者”之类的抽象实体要好。

尽管这样的平均值很有用，但我们可以做得更好。的确，大数据和机器学习的全部要点在于避免粗糙的思考。我们的集群可能是许多人组成的专业性很强的组，甚至是同一个人的不同方面：爱丽丝买工作用的书、休闲相关的书，或者作为圣诞礼物的书；爱丽丝心情好，或爱丽丝心情忧郁。亚马逊想把爱丽丝买给自己的书和送给男朋友的书区分开来，因为这样可以使它在恰当的时间做恰当的推荐。不幸的

是，采购不会贴着“自己的礼物”或者“鲍勃的礼物”标签，而亚马逊要懂得如何将这些进行分类。

假设罗比的世界中的实体分成五个集群（人类、家具、玩具、食物和动物），但我们不知道这些东西分别属于哪个集群。当我们启动罗比时，这就是它面临的问题类型。将实体分类成集群的一个简单方法，就是随机挑选5个物体作为集群典范，再利用每个典范来对比每个实体，然后将其分到最相似典范的集群中（正如在类比学习中，相似度测量的选择很重要。如果属性是数值型的，它可能就与欧几里得距离一样简单，但还有其他许多选择）。这时我们就需要更新典范。毕竟，集群的典范就是其成员的平均化，尽管当每个集群中只有一个成员时情况也是如此，但通常这种情况不会出现在我们将一堆新成员加入每个集群之后。因此对每个集群来说，我们先计算其成员的平均特性，然后将其当作新的典范。这时，我们需要再更新集群的成员：因为典范已经移走，与给定实体最接近的典范可能也已经改变。我们可以把某个类别的典范想象成一只泰迪熊，把另外某个类别的典范想象成一根香蕉。也许在第一轮，我们会把动物饼干和熊归为一类，但在第二轮中我们会把它和香蕉归为一类。最初动物饼干看起来像玩具，但现在它看起来更像食物。一旦我重新将动物饼干归入香蕉那一类，也许那个类别的典型项也已经改变，从香蕉变为饼干。随着实体被分到越来越恰当的集群中，这个良性循环会继续下去，直到实体到集群的分配不再发生改变（因此集群典范也不再发生改变）。

该算法被称为k均值算法（k－means algorithm），它的起源可以追溯到20世纪50年代。它精密、简单、人气很高，但有几个不足，而其中的一些问题和其他问题相比较容易解决。例如，我们要提前确定集群的数量，但在真实世界中，罗比总会遇到新类型的物体。有一个选择，就是如果某物与其他现存的物体迥然不同，那么就让物体开启新的集群。另外一个选择，就是允许集群分裂，然后随着我们前进而合并。不论哪种方法，我们都会想让算法偏向选择更少的集群，以免

我们最后以每个物体作为自己的集群而结束（如果我们想让集群由相似物体组成，这很难办到，但显然这并不是目标）。

有一个更大的问题，那就是k均值算法只有在集群易于区分的情况下才能起作用：在超空间中，每个集群可看作一个球团，每个团距离彼此都很远，而它们都有相似的体积，并包含相近数量的物体。如果其中的一个团出现故障，不好的事情就会发生：细长的集群会分成两个不同的集群，较小的那个集群会被附近更大的集群吸收，以此类推。幸运的是，还有一个更好的选择。

假设我们确认，让罗比在真实世界中闲逛是一种过于缓慢、麻烦的学习方法。相反，就像飞行学员在飞行模拟器中进行学习一样，我们会让罗比观看计算机生成的图片。虽然知道图片来自什么集群，但我们不会告诉罗比。相反，我们生成每张图片的方法，就是首先要随意选择一个集群（比如说玩具），然后综合该集群中的一个例子（毛茸茸的、棕色的小泰迪有一双黑色的大眼睛、圆耳朵，还戴着蝴蝶结）。我们还是随意选择例子的属性：尺寸来自一个正态分布的均值，比如10英寸，毛是棕色的概率为80%，否则就为白色，以此类推。当罗比看到许多图片由这个方法生成后，它应该学会将它们分为人类、家具、玩具等类别中，因为人类和家具相比，人类更像人类。有一个有趣的问题：如果我们从罗比的角度看，它发现集群的最佳算法是什么？答案让人惊讶：朴素贝叶斯算法。我们开始了解该算法时，它是监督式学习的一种算法。区别在于现在罗比不知道类别，因此它得猜这些东西。

显然，如果罗比不知道它们，事情就会一帆风顺，因为在朴素贝叶斯算法中，每个集群都由其概率（17%生成的物体都是玩具）和集群成员当中每个属性的概率分布来定义（例如，80%的玩具是棕色的）。罗比可以估算这些概率，只要数清数据中玩具的数量，以及棕色玩具的数量等。为了做到这点，我们需要知道哪个物体是玩具。这看起来很棘手，但最后发现我们已经知道如何做到这一点。如果罗比有一个

朴素贝叶斯算法分类器且要弄明白新物体的类别，它要做的就是应用分类器，然后在给定物体属性的条件下计算每个类别的概率（小小的、毛茸茸的、棕色的、有大眼睛、戴着蝴蝶结？可能是玩具，也可能是动物）。

因此罗比面临的是鸡和蛋的问题：如果它知道物体的类别，就可以通过数数的方式来掌握类别的模型；如果它知道模型，可以推断物体的类别。我们好像又遇到困难了，但远非如此：只要在开始时，以你喜欢的方式来猜测每个物体的类别（即使是随机的），然后你就有的忙了。从那些类别和数据中，你可以掌握类别模型；在这些模型的基础上，你可以重新推导类别，以此类推。乍一看，这看起来像一个疯狂的计划：它可能绝不会结束，而是处在由模型推断类别，到由类别推断模型的永恒循环中。即使它停止了，也没有理由相信会停在有意义的集群上。但1977年，来自哈佛大学的三个统计学家（亚瑟·邓普斯特、南·莱尔德、唐纳德·鲁宾）表明，这个疯狂的计划其实可以生效：每次我们绕着圈走时，集群模型就会变得更好，而当模型是可能性的一个局部最大值时，循环结束。他们称该计划为期望最大化演算法（Expectation Maximization，EM算法），其中E表示期望（推断预期的概率），而M代表最大化（估算可能性最大的参数）。他们还表明，许多之前的算法都是EM的特殊情况。例如，为了掌握隐马尔科夫模型，我们交替进行以下工作：推断隐藏状态，在此基础上估算过度和观察概率。无论何时，当我们想掌握某个统计模型，但又缺乏一些关键信息时（例如，例子的类别），就可以利用EM。这使它在所有机器学习中成为最受欢迎的算法。

你可能已经注意到k均值算法和EM之间的某个相似性，因为它们都交替进行两项工作：将实体分配给集群，然后更新集群的描述。这并不是一场意外：k均值本身就是EM的一种特殊情况，当所有属性都会有“狭窄的”正态分布时（有很小变量的正态分布），你就会找到它。当集群重叠很多时，举个例子，实体可能会归属于概率为0.7的集群A和概率为0.3的集群B。而在信息不丢失的情况下，我们无法决定它是

否属于集群A。EM会将其考虑进来，会将实体分配给两个集群，然后根据依据来更新它们的描述。但是，如果描述非常集中，实体归属于附近集群的概率也总接近1，而我们要做的就是将实体分配给集群，以及对每个集群中的实体都进行平均化，以获得其均值，这是典型的k均值算法。

到目前为止，我们只看到如何掌握集群的一个水平。当然，这个世界要丰富得多，集群中又有集群，一直到单个物体：活着的东西聚类成植物和动物，动物聚类成为哺乳动物、鸟类、鱼类等，一直到宠物和家狗。没问题，一旦我们掌握了集群组，就可以将其看作物体，然后反过来对其进行聚类，以此类推，直到对所有东西都进行聚类。或者，我们可以由粗略的聚类作为开始，然后将每一个集群分成亚集群，如罗比的玩具被分成填充动物玩具、组装玩具等；填充动物玩具又可分为泰迪熊、长绒毛小猫等。儿童似乎会从中间开始，然后起起伏伏努力前进。例如，他们在学习“动物”或者“小猎犬”之前先学会“狗”这个词，这对于罗比来说也许是一个很好的策略。

发现数据的形状

数据涌入罗比的大脑，靠的是它的直觉，或者是数百万个亚马逊顾客的点击流，将大量实体分类成更小数量的集群只是战役的一半，另一半则会缩短每个实体的描述。罗比看到的妈妈的第一张照片也许包含100万像素，每个像素都有自己的颜色，但你很少需要100万个变量来描述脸部。同样，在亚马逊网站上点击每一样东西，都会提供关于你的一点信息，但亚马逊真正想知道的是你喜欢什么、不喜欢什么，而不是你点击的东西。前者非常稳定，在后者中可能会有点体现，当你利用网址时，后者会毫无限制地增长。渐渐地，所有这些鼠

标点击会增加对你的品位的了解，还附上一张图，与那些像素聚集起来，形成你的品位的大致情况，这和所有那些像素合起来变成你的脸是一个道理。问题在于如何添加。

一张脸大约有50块肌肉，因此50个数字足以用来描述所有可能的表情，而且还有很大的剩余空间。眼睛、鼻子、嘴巴等的样子（就是让你区分于别人的特点）的数量也不应该超过几十种。毕竟如果面部特点只有10个选择，那么警察局的拼图师就能大概描绘出疑犯的肖像，足以用来认出他。你可以添加几个数量，用来确定光线和姿态，这样就差不多了。因此如果你给我100多个数量，就已经足以重新构造一张脸部图片。相反，罗比的大脑应该可以储存脸部图片，然后快速简化到真正起到作用的100个数量。

机器学习算法称该过程为维数约简，因为该过程将大量的可见维度（像素）简化成几个隐性维度（表情、面部特征）。维数约简对于应对大数据（像每秒钟通过你的知觉而进入的数据）来说很关键。一张图可能抵得上1000个字，但要处理和记住所做的付出，却要高出100万倍。你的视觉皮质好歹把大数据削减为数量上可管理的信息，足以用来引导这个世界、识别人和物、记住你看见的东西。这是认知最伟大的奇迹之一，并且如此自然，你甚至意识不到自己正在做这些事。

当你整理书柜上的书时，会把类似的书放在一起，你正在做一种维数约简，从广阔的主题范围到一个维数的书架。不可避免的是，有些书密切相关，最后在书架上却离得很远，但你还是可以想办法避免这样的情况再次发生。这就是维数约简算法所做的工作。

假设我给你加利福尼亚州帕罗奥多市所有店铺的GPS（全球定位系统）坐标，然后你把其中的几个画在了一张纸上（见图8－1）。

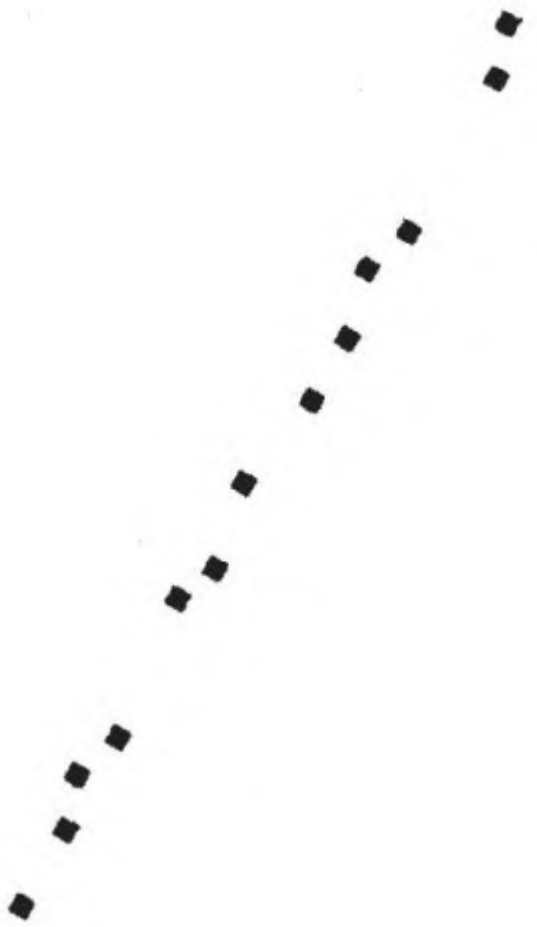


图8－1

通过看图8－1，也许你只能判断帕罗奥多市的一条街道是西南—东北向。虽然没有画出街道，但你可以凭直觉知道它就在那里，因为所有的点落在一条直线上（或者接近这条直线，可能在街道两边）。的确，这条街是大学街，如果你想在帕罗奥多市购物或吃饭，可以去那里。你一旦知道大学街上的店铺，就可以不需要两个数据来定位它们了，只要一个就可以：街道编号（或者，如果你真的想做到精确，从商铺到加利福尼亚站的距离，在西南角，此处正是大学街的起点）。

如果把更多的店铺画上去，你可能会注意到有些在十字街上，距离大学街有点距离，还有几个完全在别处（见图8－2）。



图8－2

虽然如此，但大多数店铺还是靠近大学街。如果只给你一个数据来确定店铺的位置，那么店铺到加利福尼亚站的沿街距离就是很好的选择：走了那段距离之后，往周围看也许就足以找到该店铺。所以，你把“帕罗奥多市店铺位置”的维数由二减到一。

虽然罗比没有你的优势（拥有高度进化的视觉系统），但如果你想让它去伊利特洗衣店取洗好的衣物，你只要允许他的帕罗奥多市地图有一个坐标，它就可以用一种算法从店铺的GPS坐标中“找到”大学街。完成这件事的关键在于要注意到，如果把x和y坐标的原点放在店铺地点的中心位置，然后慢慢地旋转坐标轴，当旋转大约60°时，店铺距离x轴最近，此时x轴与大学街近乎重合（见图8－3）。

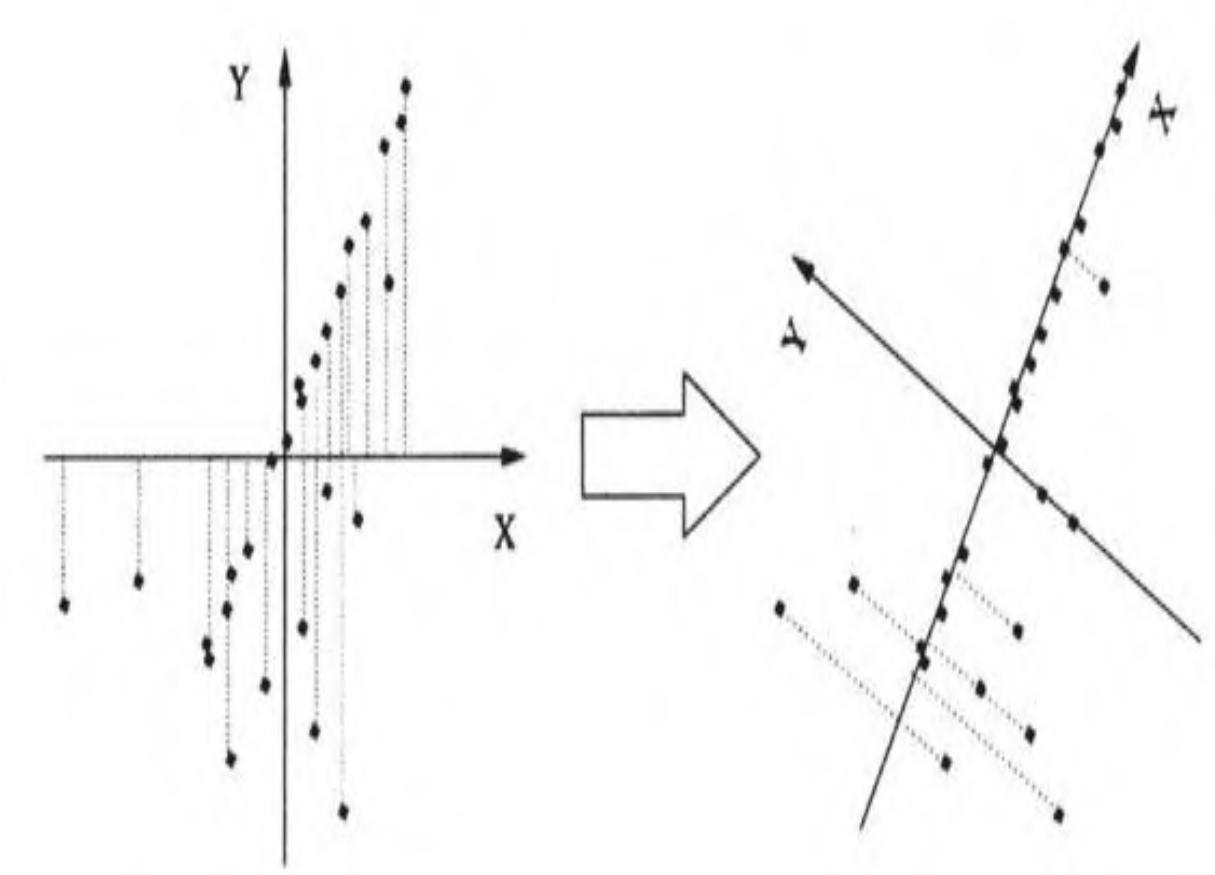


图8－3

这个方向（人们称之为数据的第一主成分）也是数据传播最多的方向（请注意，如果你将商铺投在x轴上，在图8－3的右图中它们离得比左图中的要远）。你找到第一主要成分之后，可以找第二个，在这个例子中就是差别最大、与大学街成直角的方向。在地图上，只剩下

一种可能的方向（十字街的方向）。如果帕罗奥多市在山坡上，前两个主要成分之中的一个将会部分处于上坡位置，第三和第四主要成分可能就悬在空中。我们可以将同样的思想应用到几千个或者几百万个维数中，比如面部图片，不断寻找最大差别的方向，直到剩下的变量很小，这时我们才能停下来。例如，在图8－3中旋转坐标轴之后，多数商铺的y=0，因此y的均值会很小，而我们不会因为忽略y轴的坐标而丢失过多的信息。而且如果我们决定保留y，那么z（在空中）肯定也是无意义的。结果就是，寻找主要成分的整个过程，可以利用线性代数一次性完成。首先，几个维度往往会造成高维度数据中大量的变异。虽然事实并非如此，但盯住前两三个维度不放往往会产生许多洞见，因为它利用了你视觉系统的强大感知力。

主要成分分析（principle－component analysis，PCA），正如人们对该过程的了解，是科学家的工具箱中关键的工具之一。可以说，PCA与无人监督学习的关系，正如线性回归与无人监督多样性之间的关系。例如，众所周知的全球变暖曲棍球杆曲线，就是找到各类温度相关数据序列（树的年轮、冰芯等）的主要成分并假设那就是温度的结果。生物学家利用PCA来将几千个不同的基因表达水平概括起来，变成几条途径。心理学家已经发现，个性可以简化为5个维度（外向、随和、尽责、神经质、开放性），他们可以通过你的推特文章和博客帖子来进行推断（黑猩猩可能还有一个维度——反应性，但推特数据对它们并不适用）。把PCA应用到国会投票时，民意数据显示，与普遍观点相反，政治并不主要是自由党与保守党之间的竞争。相反，人们主要在两个维度上存在差别：一个是经济问题，一个是社会问题。我们一方面可以把这些党派瓦解变成单一轴，然后与平民主义者和自由论者混合起来，这两个派别会处于两个极端，这样就创造了有很多温和派位于中间的假象。试图成功吸引他们不太可能。另一方面，如果自由党和自由论者消除了对对方的厌恶，那么他们就可以在社会问题上进行联盟，在这些问题上，双方都支持个人自由。

当罗比长大了，可以利用PCA的一个变量来解决“鸡尾酒会”问题，也就是从嘈杂的人群中辨认出单个声音。相关的方法可以帮助它掌握阅读能力。如果每个单词就是一个维度，那么在话语空间中，一篇文章就是一个点，而该空间的主要方向则是意思的成分。例如，“奥巴马总统”和“白宫”在话语空间中相距很远，而在意义空间上却很接近，因为这两个词可能会出现在相似的背景下。信不信由你，计算机就需要这种类型的分析来对SAT（美国学术能力测验）文章进行打分，并且做得和人类一样好。网飞运用了相似的方法，不只是为有相同品位的用户推荐电影，它首先会将用户和电影纳入更低维度的“品位空间”中，然后如果在该空间中某部电影接近用户，那么它就会向用户推荐。利用这种方法，它可以向用户推荐之前连用户都不知道自己会喜欢的电影。

虽然如此，当你看到脸部数据集的主要成分时，可能会感到失望。因为它们并不像你期望的那样会是面部表情或者某些特征，它们看起来很狰狞，因为十分模糊而无法辨认。这是因为PCA属于线性算法，因此所有可能的主要成分都是对真实脸部像素的平均值进行加权（也称为“特征脸”，因为它们是数据集中协方差矩阵的特征向量，但我不同意这点）。为了真正了解面部，以及世界上的大部分形状，我们需要另一样东西——非线性降维算法。

假设我们缩小帕罗奥多市的地图，而给你海湾地区主要城市的GPS坐标（见图8－4）。



图8－4

你可以再次仅仅通过看这张图来猜测这些城市是在海湾上，如果你画一条线穿过这些城市，可以仅利用一个数据就可以确定每个城市的位置：沿着这条线，该城市距离旧金山有多远。但PCA无法找到这条线，相反，它会画一条正好在海湾中间的线，当然，那里并没有城市。PCA完全没有解释数据的形状，而是将其模糊化了。

想象一下我们打算从头开始对海湾地区进行开发。我们已经决定每个城市的方位，预算允许我们铺一条连接这些城市的公路。我们会铺一条从旧金山到圣布鲁诺，再到圣马特奥，最后一直到奥克兰的公路。这条公路是海湾地区很完美的一维表示方法，而且可以通过一种简单的算法来实现：在每对附近的城市铺一条路。当然，一般情况下这种做法会建成一个公路网络，而不仅仅是一条每个城市都会使用的公路。我们可以强行铺一条最接近该网络的公路，从这个意义上讲，城市之间在这条公路上的距离，也就可以尽可能地与在网络上的距离相近。

对于非线性降维算法来说，最受欢迎的算法——等距映射算法，就可以实现这一点。在高维度空间中（比如一张脸），它可以把每个数据点与所有附近的点（很相似的脸）连接起来，依据得出的网络计算每一对数据点之间的最短距离，然后找到与这些距离最接近的简化坐标。与PCA相反，该空间中的“脸部坐标”往往意义重大：一个坐标可能表示脸部朝哪个方向（左剖面、半侧面、正面等）；另一个坐标表示这张脸的表情（非常伤心、有点伤心、不动声色、开心、非常开心等）……从了解视频中的动作到探测言语中的情绪，等距映射算法有惊人的能力，可以对准复杂数据中最重要的维度。

这是一个有趣的实验。从罗比的眼睛中提取出视频流，在图片空间中把每个框架当作一个点，然后简化那组图片变成单个维度。你会发现什么？时间。就像图书管理员整理书架上的书，时间会将每张图片放置在与之最接近的图片旁边。也许我们对它的感知，仅仅是大脑降维技术的自然结果。在记忆这个道路网络中，时间是主要的通道，而我们很快就能找到它。时间，换句话说，就是记忆的主要成分。

拥护享乐主义的机器人

聚类和维度简化虽然使我们更加靠近人类学习，但仍丢失了一些很重要的东西。孩子们并不只是消极地观察这个世界，他们会行动，会将看到的东西捡起来，和这些东西玩、到处跑、吃东西、哭、问问题。如果无法帮助罗比与周围环境互动，那么最先进的视觉系统也毫无用处。罗比不仅要知道东西的位置，也要知道每一刻该做什么。原则上，我们可以教会它使用步进指令（step-by-step instruction），把传感器读数和正确的行为组成一对，但这只对精密任务可行。采取什么样的行为由你的目标决定，而不仅取决于当前感知到的所有东西，这些目标也许只有在遥远的未来才能实现。在任何情况下，步进监督（step-by-step supervision）都不应该被采用。家长不会教自己的孩子爬、走或者跑，他们都是自己摸索。目前为止，所有的学习算法都不能做到这一点。

人类确实有稳定的向导：情感。我们追求快乐，躲避痛苦。当你碰到热的炉子时，会本能地缩手。这是简单的部分，困难的部分在于学习如何一开始不去碰那个炉子。你需要移动手指以避免之前没有感受过的刺痛。为了避免这种刺痛感，大脑要将它和你碰炉子的那一刻，以及导致这种剧痛的动作联系起来，爱德华·桑代克称该过程为“效果律”（law of effect）。人们更有可能重复引起愉悦感的动作，而导致疼痛感的动作则不太可能重复。过去的愉悦感可以穿越，可以这么说，而行为最终可以与那些离得很远的效果相联系。和其他动物相比，人类能更好地完成这种远程追求奖赏的任务，而且这对我们的成功也很重要。在一个著名实验中，工作人员在孩子们面前展示了棉花糖，并告诉他们如果可以坚持几分钟不吃棉花糖，就会得到两朵棉花糖。那些能成功做到这一点的孩子在学校生活、成人后的生活中表现得会更好。也许不那么明显，但利用机器学习来改善其网页和业务的公司也会面临相似的问题。这些公司可能会因为短期利益而做出改变，比如销售成本较低的劣质商品，以赚取同等价格优质商品的差价。这种做法忽略了一点：从长远看，这样的商家会失去顾客。

前几章中学习算法都由“即时满足”这一原则引导：每个行为，无论是标记垃圾邮件，还是购买股票，都会从“老师”那里得到即时奖励（或者处罚）。有一个机器学习的子域致力于这样的算法：进行主动探索，偶然得到奖励，然后弄清楚将来怎样才能再得到奖励。这很像婴儿到处爬和把东西放到嘴里。

这个过程称为“强化学习”，你的第一个家用机器人可能会经常用到这种方法。如果你刚打开罗比的包装，并启动它，就让它来为你煎鸡蛋和培根，它可能需要点时间。然而，在你工作的时候，它会自己探索厨房，注意各类东西的位置，以及你用哪种炉子。你回来的时候，饭菜已经做好了。

强化学习的一个重要先驱是跳棋游戏程序，这是IBM的研究员阿瑟·塞缪尔于20世纪50年代编写的。棋盘游戏是强化学习问题的典范：你得走好多步棋，却得不到任何反馈，奖励或惩罚都在最后一刻揭晓，其形式也就是赢和输。塞缪尔的程序可以自学下棋，而且下得和大多数人一样好。它不会直接学棋盘上每步棋该怎么走，因为这太困难；相反，它会学习如何评价每个棋的位置（从该位置出发，赢的概率有多大），然后选择走能到达最佳棋位的那一步。起初，它知道如何进行评价的唯一位置就是最后的结果：赢、平局或是输。一旦它知道某个特定位置能赢，也就知道哪些位置有利于让它达到这个位置。IBM前总裁小托马斯·沃森就预测，该程序被证实之后，IBM的股票会上涨15个点。最后，它被证实了。这个教训并没有被IBM忽略，后来IBM制造了一个国际象棋冠军和一个《危险边缘》冠军。

强化学习的首要思想是：并不是所有的状态都有奖励（正面或者负面），但每种状态都会有价值。在棋类游戏中，唯独最后的位置才有奖励（比如1、0、－1，分别代表赢、平局或者输）。虽然其他位置没有即时奖励，但有价值，因为通过这些位置可以得到最后的奖励。因为某个位置在某步棋时可起到“将死”（棋类术语）的作用，从实践角度看，这相当于赢一盘棋，所以有较高的价值。我们可以将这种

推理方式传播至好的与坏的开场走法中，即使在那个距离内连接完全不明显。在视频游戏中，奖励通常是分数，而一种状态的价值就是从该状态开始你能积累的分数。在现实生活中，即时奖励要好于未来奖励，投资也一样，未来奖励的回报率可能会打折扣。当然，奖励取决于你选择什么行为，强化学习的目标往往是采取那个能获得最丰厚奖励的行为。你该不该打电话约你的朋友出来约会？这可能会是一段美好关系的开始，或者你只会得到一个痛苦的回绝。即使你的朋友答应约会，这个约会的结果可能好也可能坏。从某种程度上说，你得概括所有未来可能会走的道路，然后现在做决定。强化学习通过估算每种状态的价值来做到这一点，从该状态开始你所期望得到的全部奖励，然后选择能将奖励最大化的行为。

假设你在一条隧道中行走，像印第安纳·琼斯那样，然后你到了一个三岔口。地图显示左边的隧道通往宝藏，右边的则通往蛇洞。你所站位置的价值（就在三岔口前）也就是宝藏的价值，因为你会选择往左走。如果你总是选择最佳可能的行为，那么某个状态的价值与成功状态的价值的差异，仅仅在于你因实施该行为而获得的即时奖励（如果有）。如果我们知道每种状态的即时奖励，就可以利用这个观察来更新相邻状态的价值，以此类推，直到所有状态都会有一致的价值。宝藏的价值会沿着隧道反向传播，直到到达岔路口或者超过这个地方。你一旦知道每种状态的价值，就可以知道在每种状态中该采取什么行动（能够将即时奖励和最终状态价值最大化的行动）。这很大程度是由控制论理论家理查德·贝尔曼解决的，但强化学习中真正的问题会在你没有某区域的地图时出现。你唯一可做的就是探索并发现奖励在哪里。你可能发现宝藏，也可能陷入蛇洞。每当你要采取行动时，就会注意即时奖励和最终状态，这很大限度上可以由监督式学习来实现。但你还是得更新你采取行动后的状态的价值，使其与刚刚观察的价值一致，也就是说，你获得的奖励加上所处新状态的价值。当然，这个价值可能并不准确，但如果你已经花了足够长的时间来边徘

徊边做这件事，那么最终会停在所有状态的正确价值以及对应行动上。简言之，这就是强化学习。

请注意强化学习算法如何面对我们在第五章遇到的利用—探索困境：为了使奖励最大化，你很自然地想选择能达到高价值状态的行动，但这可能会阻止你发现其他价值更高的奖励。强化学习算法解决这个问题的方法是时而选择最佳行动，时而随机选择（为此，大脑甚至似乎有一台“噪声发生器”）。早些时候有很多可以学习的东西，进行大量探索就会很有意义。一旦你发现了某个区域，最好集中精力来使用它，这就是人一生要做的事：儿童会去探索，成年人想着如何使用（除了科学家，因为他们是永远的孩子）。儿童的游戏比表面看起来要严肃。如果进化过程产生一个没有用的生物，并在出生后的几年给其父母带来沉重负担，那么付出这么大的成本一定是为了更丰厚的利益。实际上，强化学习就是一种加速进化过程——尝试、丢弃，然后在单个生命的一生而不是几代中改进行动——有了这个标准，它的效率就会很高。

对于强化学习的研究自20世纪80年代早期才正式开始，马萨诸塞大学的里奇·萨顿和安迪·巴尔托参与了研究工作。他们认为学习取决于与环境的互动，这一点很关键，但监督算法并没有发现这一点；而且他们在动物习得心理学领域找到灵感。萨顿成为强化学习的主要倡导者。另一个关键进展发生在1989年，当时剑桥大学的克里斯·沃特金斯在儿童学习实验发现的推动下，实现了强化学习的现代形式，即在未知环境中进行最优控制。

然而，到目前为止，我们看到的强化学习算法还不是很现实，因为它们在某个状态中不知道该做什么，除非它们之前进入过那种状态，而在现实世界中，并没有哪两个情形是完全相似的。我们需要对之前所处的状态到新状态进行概括。幸运的是，我们已经知道该如何做到这一点：用强化学习来将之前见过的一个监督式学习算法（例如，多层感知器）包括进来。这时神经网络的工作就是预测状态的价

值，而反向传播的误差信号就是预测价值与所观察到的价值之间的差别。还存在一个问题：在监督式学习中，某个状态的目标价值总是一样的；但在强化学习中，它会不断变化（这是达到附近状态的结果）。因此，带有归纳性的强化学习往往无法确定稳定的解决方法，除非内部学习算法很简单，就像线性函数那样。虽然如此，带有神经网络的强化学习已经取得一些显著成就。早期的成就包括制造出具备人类水平的西洋双陆棋选手。最近的一种强化学习算法，来自DeepMind（伦敦的一家创业公司），在电视体育游戏及其他简单的大型电玩中打败了一位专家级的人类选手。它利用深层网络来从控制台屏幕低像素的动作中预测其价值。有了端对端的视野、学习和控制，该系统与人工大脑至少有一点相似之处。这可以解释为什么DeepMind是一家没有产品和利润，员工也寥寥无几的公司，但谷歌却向它投了5亿美元。

除了游戏，研究人员还可以利用强化学习来平衡极点、控制简笔画的体操运动员、使汽车倒车入位、驾驶直升机颠倒飞行、管理自动电话对话、分配手机网络中的频道、调度电梯、安排航天飞机货运装载等。强化学习也对心理学和神经科学产生了影响。大脑利用神经递质多巴胺来传播期望奖励与实际奖励之间的区别。强化学习解释了巴甫洛夫条件反射作用，但不像行为主义，它允许动物有内部心理状态。觅食的蜜蜂会利用它，在迷宫中找到奶酪的老鼠也是如此。你的日常生活由一连串你很少注意到的、由强化学习形成的奇迹组成。你起床、穿衣服、吃早餐，然后开车去上班，这些过程中你一直在思考别的事情。实际上，强化学习会不断精心安排和调整这个奇妙的动作交响曲。强化学习片段（为习惯）组成大多数你做的事。当你觉得饿了时，会走到冰箱前，拿一点零食。正如查尔斯·杜希格在《习惯的力量》一书中表明的那样，理解并控制由线索、日常、奖励组成的循环关系是成功的关键，不仅对个人，而且对企业甚至整个社会来说都是这样。

作为强化学习的创始人，里奇·萨顿是最具有热情的。对于他来说，强化学习就是终极算法，而且解决了这个问题就相当于解决了人工智能问题。但克里斯·沃特金斯却并不满意。他看到许多儿童能做但强化学习算法做不了的事：解决问题，经过几次尝试之后能更好地解决问题，制订计划，获取逐渐抽象的知识。幸运的是，对于这些较高水平的能力，我们也有相应的学习算法，其中最重要的就是组块算法。

熟能生巧

学习的目的就是为了更好地实践。你现在可能记不得学着系鞋带有多难了。一开始，你完全学不会，虽然当时你已经5岁了。接着，你的鞋带松开的速度比你系的速度还要快。慢慢地，你学会更快、更好地系鞋带了，直到你不自觉地系鞋带。很多其他事情也是同样的道理，比如爬、走、跑、骑单车、开车、阅读、写作、算术、演奏乐器、进行体育活动、做饭和使用计算机等。讽刺的是，人在最痛苦时学到的东西往往最多。在早期，每一步都很艰难，你不断失败，即便你成功了，结果也并不会有多好。当你掌握高尔夫挥杆动作或者网球发球之后，可能会花几年完善这些技能，但这些年你取得的进步会比开始学时的前几周少很多。通过练习你会做得更好，但不会以恒定的速度进步：一开始你进步很快，接下来就没那么快了，最后变得很慢。无论是玩游戏还是弹吉他，技能提高时间曲线图（你做某事的表现或者做某事所花费的时间）有一个非常特殊的形式（见图8－5）。

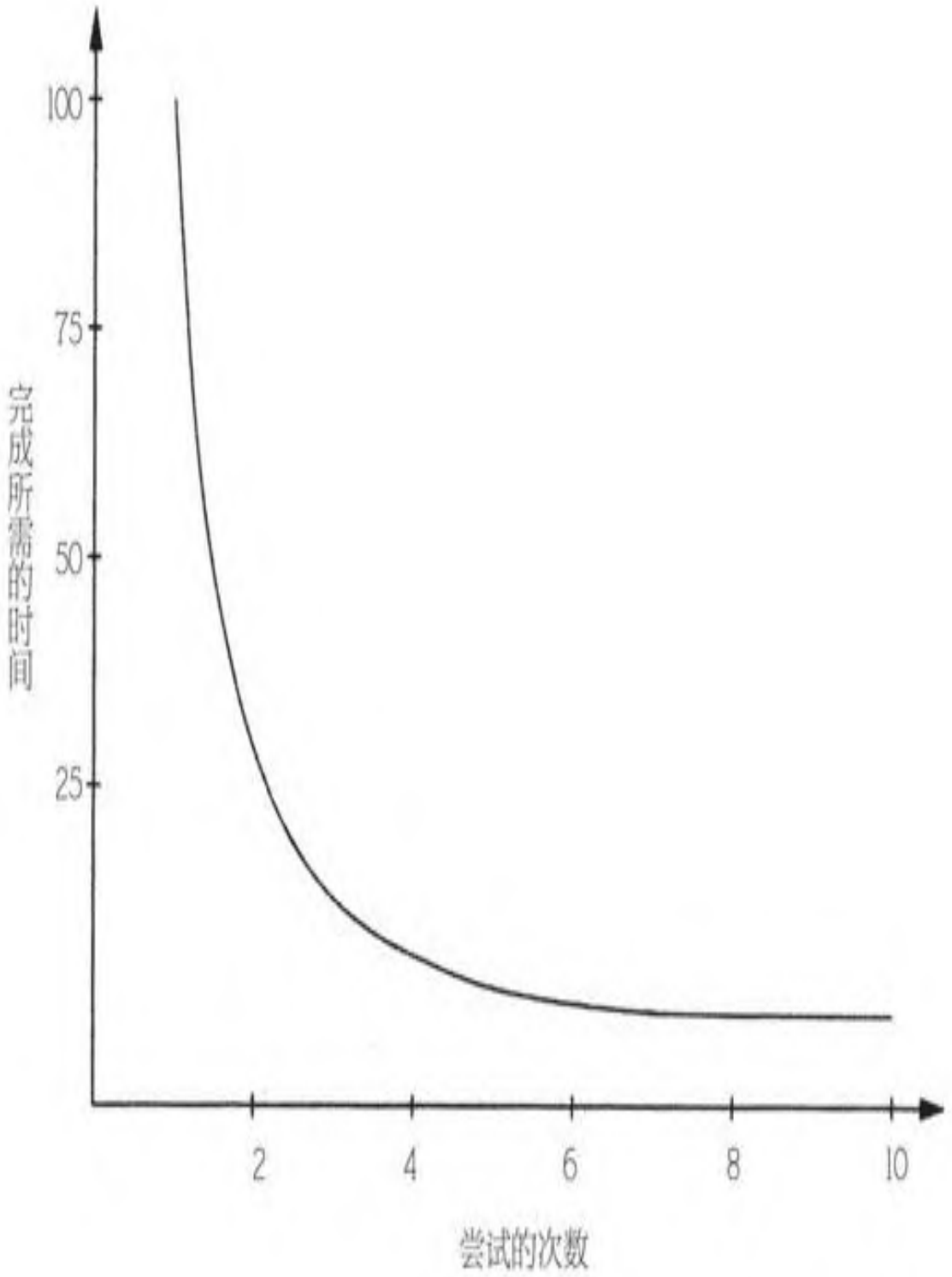


图8－5

这种类型的曲线被称为“幂法则”（a power law），因为随着时间达到某负幂时，表现会出现变化。例如，在图8－5中，完成所需的时间与尝试的次数的负二次方成正比（或者相当于1除以尝试次数的平方）。几乎每项人类技能都会遵循幂法则，不同的幂会对应不同的技能（相反，Windows系统则不会随着实践次数的增多而变快——这是微软该解决的问题。）

1979年，艾伦·纽厄尔和保罗·罗森布鲁姆开始探索这个所谓的“幂法则”实践存在的原因。纽厄尔是人工智能的创始人之一，也是主要的认知心理学家，而罗森布鲁姆则是他在卡内基－梅隆大学的研究生之一。当时，没有哪个实践模式可以解释幂法则，纽厄尔和罗森布鲁姆怀疑这可能与组块有关。组块是一个来自感知与记忆心理学的概念。我们以组块的形式来感知并记住东西，而在任意给定的时间内（根据乔治·米勒的经典论文的结论为7±2），我们只能通过短暂记忆来记住这么多组块。关键是将信息聚集成组块可以让我们处理得更多，否则我们处理不了那么多信息。这就是电话号码有连字符的原因：1－723－458－3897比17234583897要好记得多。赫伯特·西蒙（纽厄尔的长期合作伙伴以及人工智能的共同创始人）早期已经发现，新棋手和专业棋手的主要区别在于，新棋手一次记一个象棋位置，而专业棋手则看到涉及多个位置、更大的模式。要提高国际象棋技能主要涉及获取更多、更大这样的组块。纽厄尔和罗森布鲁姆假设，在掌握所有技能而不仅仅是下棋技能中，类似的过程在起作用。

在感知和记忆中，组块仅仅是一种符号，代表了其他符号的模式，就像AI代表人工智能。纽厄尔和罗森布鲁姆将这个思想运用到纽厄尔和西蒙早期创建的问题解决理论中。纽厄尔和西蒙让实验对象解决问题，例如，在黑板上利用一个数学公式推导出另外一个数学公式，同时大声描述他们如何得到这个公式。他们发现，人类解决问题的方式是将问题分解为小问题，再将小问题再分解为更小的问题，然后系统地减少初始状态（比如第一个公式）与目标状态（第二个公式）之间的差异。然而，这样做需要找到能起作用的行动顺序，这需

要时间。纽厄尔和罗森布鲁姆的假设是，每次当我们解决一个小问题时，会形成一个组块，这样我们就可以直接从解决问题前的状态进入解决问题后的状态。这种意义下的组块有两个部分：刺激（你从外部世界或者短暂记忆中识别出的模式）和反馈（你因此而执行的行动顺序）。一旦掌握一个组块，你就会将其储存在长期记忆中，下次你解决同样的小问题时，就可以直接应用该组块，会节省寻找的时间。这个过程在所有水平层面都会发生，直到你找到一个能解决所有问题的组块，并且能够自行解决这些问题。为了系好你的鞋带，先打结，然后用一端栓个套，用这个套绕住另一端，然后将其从中间的洞穿过。对于5岁的孩子来说，这些并不简单，可一旦你掌握相应的组块，就差不多学会了。

纽厄尔和罗森布鲁姆将他们的组块程序应用于解决一系列问题，测量它在每次尝试中花费的时间，这就弹出一系列幂法则曲线。但这仅仅是开始，接下来他们将组块并入“猛增”（Soar）理论中，这是纽厄尔和另一个学生约翰·莱尔德一起创建的一般认知理论。“猛增”程序不仅为某个预设了等级的目标而努力，当每次遇到障碍时，还都可以定义并解决新的子问题。一旦它形成新的组块，“猛增”程序就会对其进行一般化，以应用到相似的问题中，这与逆向演绎的方式类似。除了幂法则实践，“猛增”中的组块最后成为许多学习现象的良好模型。通过对数据和类推法进行组块，它甚至可以应用到掌握新知识中。于是纽厄尔、罗森布鲁姆、莱尔德做出这样的假设——组块是学习所需的“唯一”机制，换句话说，就是终极算法。

作为典型的人工智能类型，纽厄尔、西蒙及其学生和拥护者都非常相信问题解决的重要地位。如果问题解决者很强大，那么学习算法就会依赖其能力而变得简单。的确，学习只是另外一种问题的解决方式。纽厄尔和同伴共同努力，将所有学习简化为组块，将所有认知简化为“猛增”程序，最后他们失败了。其中一个问题是，尝试这些步骤的成本变得很高，程序变得越来越慢。从某种程度上说，人类要避免这一点的发生，但到目前为止，该领域的研究人员还找不到解决方

法。最重要的是，尝试将强化学习、监督式学习，以及别的一切简化成组块，基本上会制造更多的问题，而不会解决问题。最终，“猛增”的研究人员承认失败，然后将那些其他类型的学习作为分离机制并入“猛增”。虽然如此，组块还是学习算法的典型例子，受到心理学的启发，还是真正的终极算法。无论它最后变成什么，肯定会分享其通过实践得到提升的能力。

组块和强化学习在商业中的应用不如在监督式学习、聚类或者维数简约中广泛，但是通过与环境互动进行的相似学习类型是掌握行动（或相应行动）的效果。如果现在你的电子商务网站主页的背景是蓝色的，你在想让它变成红色会不会提高销量，那么可以将红色网页用在10万个随机挑选的用户身上试试看，然后将结果与常规网站进行比较。这种手法称为A/B测试，起初主要用在药物试验中，之后运用到许多领域。在这些领域中，数据可以根据需要集中起来，从营销到对外援助。它还可以概括成一次性尝试许多变化组合，而不会忽略什么样的改变会带来什么（或者失去什么）。诸如亚马逊、谷歌之类的公司对此深信不疑，你可能已经不知不觉参与了几千项A/B测试。A/B测试证明常听到的批评言论的错误：大数据只对寻找相关关系有好处，对寻找因果关系则没用。除了哲学要点，掌握因果关系就是掌握行动的效果，而任何有数据流并能影响数据流的人都能做到这一点——从一岁孩子在浴缸中戏水到总统再次参加竞选。

学会关联

如果我们赋予罗比机器人在本书中提到的所有学习能力，它会变得很智能，但还会有点孤僻。它会把世界看成一群独立的物体，可以对其进行识别、操纵甚至预测，但它不理解世界是一张相互联系的网

络。罗比医生会很擅长根据病人的症状来诊断其患有感冒，但无法猜想这个病人患的是猪流感，因为它和感染该疾病的人有过接触。在谷歌出现之前，搜索引擎会通过看它的内容来决定某网页是否与你的搜索相关——还有别的吗？布林和佩奇的观点是，与该网页相关的最明显标志在于，相关的网页会与它有链接。同样的道理，如果你想预测某个青少年是否有开始吸烟的危险，到目前为止，你能做得最有效的事就是调查他的好朋友是否吸烟。一种酶的形状与其分子形状的关系，就像锁与钥匙的关系一样不可分割。捕食者与猎物有缠绕很紧的属性，每种属性会进化出打败对方的属性。在所有这些例子中，了解一个实体的最佳方法——无论它是人、动物、网页还是分子——就是了解它如何与其他实体进行连接。这就需要一种新的学习方式：不会将数据当作不相关物体的随机样本，而是对复杂网络的一瞥。网络中的节点会互相作用；你对这个节点做了什么，会影响到其他节点，最后还会返回来对你进行影响。关联学习算法，正如它们的名字那样，也许不太有社交智能，但它们将是最好的东西。在传统统计学习中，每个人都是一座岛，与世隔绝；在关联学习中，每个人都是一块陆地，是主要大陆的一部分。人类是关联学习者，被相互连接起来，我们如果想让罗比成长为一个有知觉、擅长社交的机器人，就需要对它进行布线，然后连接。

我们面临的第一个困难就是，当数据变成整个大网络时，似乎就没有许多例子来学习了，只剩一个，而且这不够。朴素贝叶斯算法通过数发烧流感病人的数量来知道发烧是感冒的一种症状。如果它只能看到一位病人，可能会得出结论：流感总会引起发烧，或者不会引起发烧。这两个结论都是错的。我们想确认流感是传染性疾病，方法就是通过观察在社交网络中的感染模式——感染的人群在这边，没有感染的人群在那边——但我们只能看到一个模式，即使它处于70亿人的网络中，因此如何概括尚不清晰。关键在于要注意到，嵌入该网络中，我们有许多对人的实例。如果熟人与未曾相见的人相比，更有可能患流感，那么认识流感病人也会让你成为流感病人的可能性增大。

然而，不幸的是，我们不能只数数据中都患流感的熟人的对数，然后把这些数变成概率。因为一个人可以有许多熟人，而所有成对的概率不会总计为相关模式。举个例子，该模式会让我们计算在给定他们熟人的情况下，计算某人得流感的概率。当例子都分开时，我们就不会有这样的问题了，我们也不会关注这个例子，比如丁克家庭中，每个人都住在自己的“荒岛”上。这并不是真实世界，而且无论如何，这里绝对不会有什么流行病。

解决办法就是找到一组特征，并掌握它们的权值，就像在马尔科夫网络中那样。对于每个人（X），我们可以有“X患有流感”的特征，对于每对熟人X和Y，特征就是“X和Y患有流感”，等等。正如在马尔科夫网络中那样，最大似然权值使每个特征以在数据中观察到的频率出现。如果有很多人患有流感，那么“X患有流感”的权值就会比较高。如果当X患有流感时，熟人Y也患有流感的概率会比随机抽选的网络成员患有流感的概率大，因此“X和Y都患有流感”的权值会较高。如果40%的人有流感，那么16%的所有熟人两人组也会患有流感，那么“X和Y都患有流感”的权值将等于零，因为我们不需要该特征来准确复制数据的统计（0.4×0.4=0.16）。如果特征的权值为正，流感更有可能发生在人群当中，而不会十分容易随机感染某个人。也就是说，如果你熟悉的人患有流感，那么你感染流感的概率会更大。

请注意，网络会为每对熟人赋予分离的特征，比如爱丽丝和鲍勃都患有流感，爱丽丝和克里斯都患有流感，等等。但我们无法为每个两人组确定一个分离权值，因为他们只有一个数据点（无论是否受感染了）。对于尚未确诊的网络成员，我们不能将他们一概而论（伊薇特和扎克都患有流感吗）。我们只能在自己见过的所有实例的基础上，为相同形式的所有特征确定一个权值。实际上，“X和Y患有流感”是特征模板，可以利用每对熟人进行实例化（爱丽丝和鲍勃，爱丽丝和克里斯，等等）。模板的所有实例的权值被“捆绑在一起”，从这个意义上讲，它们有同样的价值。也就是说，即便只有一个实例（整个网络），我们也可以进行推理。在非关联学习中，一个模型的

参数仅以一种方式捆绑——越过所有独立例子（比如所有被诊断过的病人）。在关联学习中，每个我们创造的特征模板将捆绑所有实例的参数。

我们不仅限于成对的或者单独的特征。脸书想预测哪些人是你的朋友，这样才能为你推荐朋友，为此它可以利用“朋友的朋友可能成为朋友”这个规则，因为每个实例都涉及三个人。比如，如果爱丽丝和鲍勃是朋友，而鲍勃和克里斯也是朋友，那么爱丽丝和克里斯有可能成为朋友。H·L·门肯说过一句妙语：如果一个人挣的钱比他的妻子的妹妹的丈夫要多，那么他就是富有的。这句话涉及四个人。这些规则中的每一条都可以变成关联模型中的特征模板，而在特征出现在数据中的频率的基础上，可以确定它的一个权值。正如在马尔科夫网络中那样，特征也可以通过数据确定。

关联学习算法可以从一个网络推广到另一个网络（比如确定流感如何在亚特兰大传播的模型，然后在波士顿运用），它们也可以在多个网络上学习（比如亚特兰大和波士顿，假设亚特兰大没有人和波士顿人有联系）。不像“常规”学习，这里所有的例子必须有同等数量的属性。在关联学习中，网络可以在大小方面有所不同。同样的模板，较大的网络会比较小的网络有更多的实例。当然，从较小的网络到较大的网络的推广可能准确，也可能不准确，需要指出的是，没有什么会阻止它。从局部上看，大型网络往往和小型网络的表现没有差异。

关联学习算法能达到的最佳效果就是让懒散的老师变得勤奋。对于普通分类器来说，没有归类的例子毫无用处。如果我只知道病人的症状，却不知道诊断结果，那么这对我学习看病毫无帮助。如果我知道患者的一些朋友感染了流感，那么可能间接证明他也感染了流感。在一个网络中确诊几个人，然后将这些诊断结果推测到他们的朋友，以及他们朋友的朋友身上，这就是接下来诊断所有人的最佳做法。推测出来的诊断结果可能有干扰因素，但关于症状如何与流感相关联的

整体统计，与只有几个诊断结果可利用相比，前者要准确和完整得多。孩子很擅长充分利用他们受到的分散式监督（只要不选择忽略它），关联学习算法也有部分这样的能力。

然而，这种力量也是要付出代价的。在普通分类器中，比如决策树或者感知器，从实体的属性推断出其类别，这就涉及几次查找工作和一些计算。在一个网络中，每个节点的类别间接取决于所有其他节点，而我们不可以孤立地对其进行推断。我们可以求助于同样用于贝叶斯网络的推理技术，比如环路信念传播或者MCMC，但数值范围是不一样的。典型的贝叶斯网络可能会有几千个变量，但典型的社交网络却有数百万个节点甚至更多。幸运的是，因为网络模型由许多有相同权值的相同特征不断重复形成，我们往往可以将网络压缩变成“超节点”，每个超节点由许多节点组成。我们知道这些节点有相同的概率，而且会解决许多更细小的问题，结果也完全相同。

关联学习有悠久的历史，其历史可以至少追溯到20世纪70年代，以及符号学派技巧（如逆向演绎）。但随着互联网的出现，它需要新的动力。突然之间，网络变得无处不在，而对网络进行模仿变得刻不容缓。我发现有一个特别有趣的现象——口碑传播。信息如何在社交网络中传播？我们可否测量每个成员的影响力，然后将目标确定在数量够多、影响力最大的成员身上，以开启一轮口头传播？我和学生里特·理查森一起设计了一种算法就能够做到这一点。我们将其运用到Epinions网页（一个产品评论网站）上，可以允许成员说出他们信任谁的评论。研究发现之一是，向单个最有影响力的成员营销产品（该成员被许多拥护者的信任，而这些拥护者也被自身的拥护者信任等）和对1/3的所有成员进行推销相比，效果是一样的。随之而来的是一大批关于此问题的研究。此后，我就已经开始将关联学习应用到其他许多领域，包括预测谁会在社交网络中形成链接、整合数据库、使机器人能够绘制周围环境的地图。

如果你想了解世界如何运转，那么关联学习就是很好的工具。在艾萨克·阿西莫夫的《基地》一书中，科学家哈里·谢顿可以从数学角度预测人类的未来，从而将其从衰落中拯救出来。此外，保罗·克鲁格曼也承认，就是这个诱人的梦想，让他成为一名经济学家。依据谢顿的观点，人类就像气体中的分子，大数法则保证，即使个人无法预测，整个社会却可以。关联学习表明事实并非如此，如果人类是独立的，每个人孤立地做决定，社会的确就会变得可预测，因为所有那些随意的决定会合计为一个相当恒定的均值。但当人们互动时，较大的集合体会比较小的集合体更不那么可预测。如果信心和恐惧可以传染，每种情绪都会主导一段时间，但常常整个社会会从这种情绪转化到另一种情绪。虽然如此，并不是所有消息都是坏的。如果我们可以估计人们之间相互影响的强度，也就可以估计这种转换在多久之后会发生，即使这是第一次转换。换个说法，黑天鹅并不一定不可预测。

对于大数据，抱怨较普遍的就是拥有的数据越多，就越容易在大数据中发现伪模式。如果数据仅仅是一大组分离的实体，那么这一点可能正确；但如果它们相互关联，情况就不一样了。例如，那些利用数据挖掘来逮捕恐怖分子的批评者认为，除了伦理问题，这绝不会起作用，因为无辜的人太多，而恐怖分子却极少，因此它要么会引发过多的错误警报，要么不会逮捕任何人。当某个游客或者恐怖分子在划定爆炸地点的范围时，有人对纽约市政厅进行录像了吗？购买大量硝酸铵的人是农民还是爆炸制造者？所有这些人单独看都很无辜，但如果“游客”和“农民”一直保持密切的电话联系，而后者只是开着他的拉登皮卡进入曼哈顿，可能是时候需要一个人来盯紧他了。美国国家安全局喜欢搜寻谁给谁打过电话的记录，这不仅仅因为可以说这是合法的，而且因为对于预测算法来说，它们往往比通话内容的信息量更大，而通话内容需要人进行分析才能被理解。

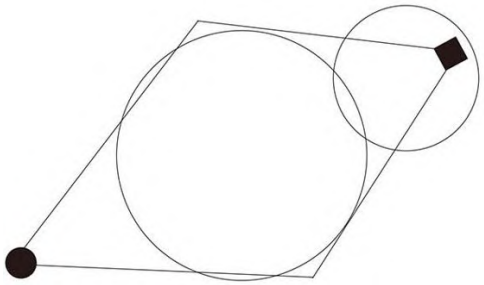
除了社交网络，关联学习的杀手级应用正在了解活细胞如何运作。一个细胞是一个复杂的代谢网络，在这个过程中，基因为蛋白质编码，蛋白质又调节其他基因，还有一长串的化学反应，而产物会从

这个细胞器运输到另一个细胞器。独立的实体，孤立地进行工作，找不到任何踪影。治癌药物必须在不干扰正常细胞的情况下，打破癌细胞的运转。如果我们有两者准确的关联模型，就可以在计算机中尝试许多不同的药物，让模型推断它们好的以及不好的效果，然后只将最好的留下，先从体外，最后再从体内进行测试。

就像人类的记忆，关联学习编织了一个丰富的关联网。它连接认知，诸如罗比之类的机器人可以通过聚类和维数简约来获取认知并习得技能。它可以通过强化和聚类来学习，利用由阅读得来的更高水平的知识去学校与人类进行互动。关联学习是最后一块拼图，也是我们需要为自己的炼金术提供的最后的原料。现在是时候重建实验室，使所有这些元素变成终极算法了。

第九章

解开迷惑



机器学习既是科学，也是技术，两者的特点提示我们如何将其统一起来。在科学方面，理论的统一往往从看似简单的观察开始。两个看似不相关的现象原来只是同一枚硬币的两面，就像第一张倒下的多米诺骨牌，会引起其他许多牌倒下。苹果落到地上，月亮悬挂在夜空，这两者都是由重力引起的，而且（无论是否杜撰）一旦牛顿弄明白这些现象的原因，重力最后也可以用来解释潮汐、分点岁差、彗星轨道等其他很多现象。在日常体验中，电和磁绝不会同时出现：这边有闪电的火花，那边有吸住铁制品的岩石，两种现象都很少见。可是一旦麦克斯韦弄清楚电场的改变会产生磁性，反之亦然，显然闪电是两者亲密的媒介。现在我们知道，电磁并不罕见，遍及所有物质。门捷列夫的元素周期表不仅将所有已知物质仅分成两个维度，也预示哪里会有新元素。达尔文在贝格尔号上的观察突然得到理解，因为马尔萨斯的《人口论》表明了自然选择这一组织原理。

当克里克和沃森偶然发现双螺旋结构可用于解释DNA令人迷惑的属性时，他们马上就看到DNA可能会如何复制自身，而生物学由“集邮”

（卢瑟福用的贬义词）到“统一科学”的过渡开始了。在这些例子中，各种各样令人眼花缭乱的观察结果原来都有同一个相同的起因，而一旦科学家们明确了这个起因，就可以反过来利用它来预测许多新的现象。同样的道理，虽然我们在本书中见到的学习算法似乎差别很大——有些以大脑为基础，有些以进化为基础，有些则以抽象的数学原理为基础，但实际上，它们有很多相似之处，而由学习产生的理论会有许多新的观点。

虽然知道这一点的人比较少，但世界上许多最重要的技术都是创造统一物的结果，该统一物是单一机制，能完成之前需要很多机制完成的事情。互联网，正如它的名字一样，是连接相互网络的一种网络。没有它，每种类型的网络就得需要一个不同的协议来与其他每个网络进行对话，这就很像对于世界上每种不一样的语言都需要一本不一样的字典。互联网协议就是一种世界语言，会给每台计算机能与其他所有计算机直接对话的错觉，世界语言还允许电子邮件和网页忽视那些它们无法起到影响作用的基础设施建设的细节。关系数据库也会对企业应用做类似的事情，允许开发商和用户依据抽象关联模型进行思考，并忽略关于回应搜寻计算机所尝试的方式。微处理器是一种数据电子元件装配，可以对所有其他装配进行模仿。虚拟机允许同一台计算机同时在100个人面前伪装成100台不同的计算机，并使云成为可能。图形用户界面让我们编辑文件、数据表、幻灯片以及别的很多东西，利用Windows操作系统、菜单、鼠标单击的通用语言。计算机本身就是统一物：单个设备可解决任何逻辑或者数学问题，只要我们知道如何对它进行编程。甚至电也是一种统一物：你可以通过许多不同的来源来获取它——煤、天然气、核能、水力、风力、太阳，然后以无限多种的方式来消耗它。一座发电站不会知道或者关心它生产的电会如何被消耗掉，而你的门廊灯、洗碗机或者全新的特斯拉也不会在意电力供应来自哪里。电力就是能源的世界语言。终极算法是机器学习的统一物：它让任意应用利用任意学习算法，方法是将学习算法概括成通用形式——所有应用都需要知道该形式。

我们迈向终极算法的第一步会简单得令人意外。事实证明，要将许多不同的学习算法结合成一个并不难，利用的就是元学习。网飞公司、沃森、Kinect以及其他无数公司都会利用它，而它在机器学习算法的箭袋中也是最有力量的一支。它还是接下来要进入的深层次统一的垫脚石。

万里挑一

这里有一个挑战：你有15分钟用来联合决策树、多层感知器、分类器系统、朴素贝叶斯，以及支持向量机，使其变成拥有每个部分最佳属性的单一算法。快！你能做什么？显然，这不能涉及单个算法的细节。没有时间了，但接下来怎么办？把每种学习算法当成委员会上的一位专家，每个人都仔细观察待分类的实例——这个患者的诊断是什么，然后有把握地做出预测。你不是一位专家，却是委员会主席，你的工作就是把他们的意见结合成一个最终决定。你手上的问题其实就是一个新的分类问题，这种情况下，输入不是患者的症状，而是专家的观点。但你可以将机器学习以同样的方式运用到解决这个问题上，专家也会这样将其运用到最初的问题上。我们可以称其为元学习，因为它就是要了解学习算法。元学习算法本身可以是任意学习算法，从决策树到简单的权值投票。为了掌握权值或者决策树，我们利用学习算法的预测来代替每个最初例子的属性。经常能够准确预测类别的学习算法会取得较高的权值，而不准确的那些则容易被忽略。有了决策树，是否要利用学习算法可能会依照其他学习算法的预测来定。不管怎样，为了给既定训练例子获取学习算法的预测，我们首先必须将其运用到原始训练集“排除该样本”中，然后利用最终的分类器，否则委员会就有被拟合学习算法控制的风险，因为它们可以通过记忆类别来预测准确的类别。网飞奖获得者利用元学习来结合数百个

不同的学习算法；沃森利用它来从备选项中选择最终的答案；内特·希尔也以相似的方式将投票与预测选举结果结合起来。

这种类型的元学习被称为“堆叠”，是大卫·沃尔珀特的创见，在第三章中我们提到过他，他是“天下没有免费的午餐”定理的创造者。还有一个更简单的元学习算法是“装袋”算法，由统计学家里奥·布雷曼发明。“装袋”算法通过重新取样的方法来产生训练集的随机变量，将同样的学习算法应用到每个训练集中，然后通过投票将结果结合起来。做这件事的原因是它可以减少变量：组合模型和任何单一模型相比，对于变幻莫测的数据的敏感度要低得多，这样提高准确度就变得很容易了。如果模型是决策树，且我们通过保留属性（这些属性因考虑每个节点而得来）的任意子集来进一步改变这些决策树，那么所得结果就是所谓的随机森林。随机森林到处是一些准确的分类器。微软公司的Kinect利用它们来弄清楚你在做什么，而且它们会经常在机器学习比赛中获胜。

最聪明的元学习算法之一就是推进，由两位学习领域的理论家约阿夫·弗罗因德和罗伯·夏皮尔创造。推进算法不是通过结合不同的学习算法，而是将相同的分类器不断应用到数据中，利用每个新的模型来纠正前面模型的错误。它通过将权值分布给训练实例的方式来完成这件事。每个被误分类的权值，在每轮学习过后都会增长，使得后面的几轮会更向它集中。推进算法的名字源于这样的想法：该过程可以推进只比随机猜测好一点的分类器，但如果持续如此，就会接近完美。

元学习非常成功，但它却不是深入组合模型的方法。另外，它也昂贵、苛刻，因为会做很多轮学习，而且组合模型可能会很难懂（“我认为你有前列腺癌，因为决策树、遗传算法、朴素贝叶斯算法都这么判断，虽然多层感知器和支持向量机反对”）。此外，所有的组合模型确实只是一个巨大而凌乱的模型。难道我们不能找到完成同样任务的单一学习算法吗？当然可以。

终极算法之城

我们的统一学习算法也许通过延伸寓言可以得到最好的介绍。如果机器学习是一块大陆，被分成5个区域，那么终极算法就是首都城市，矗立在5个区域会合的特殊地带。如果从远处接近它，你会看到这个城市由3个同心圆组成，每个圆被一堵墙围着。外围的圆，也是最宽的圆是“优化城”。这里的每个房间就是一种算法，而且它们有不同的形状和面积。有些房子在建设中，当地人忙着围着它团团转；有些房子非常崭新；有些房子则看起来老旧荒废。山的更高处是“评价城堡”，命令不断从它的大厦和宫殿向下面的算法发出。最重要的是，在天空的映衬下，“代表法之塔”矗然而立，这里住着城市的统治者。他们的城市有不可改变的法律，不仅规定在城市范围内，而且规定了在整个大陆什么能做、什么不能做。在中央，最高的塔尖飘扬着终极算法的旗帜，颜色是红和黑，上面的五角星围着我们无法理解的文字。

这座城市被分成5个区域，一个区域属于一个学派。每个区域从“代表法之塔”一直延伸至城市的外墙，包围了塔、“评价城堡”里的一群宫殿，以及它们俯瞰的“优化城”的街道和房屋。这5个区域和3个圆圈将城市划分成15个区域、5种形状，以及你需要解决的那15块拼图（见图9－1）。



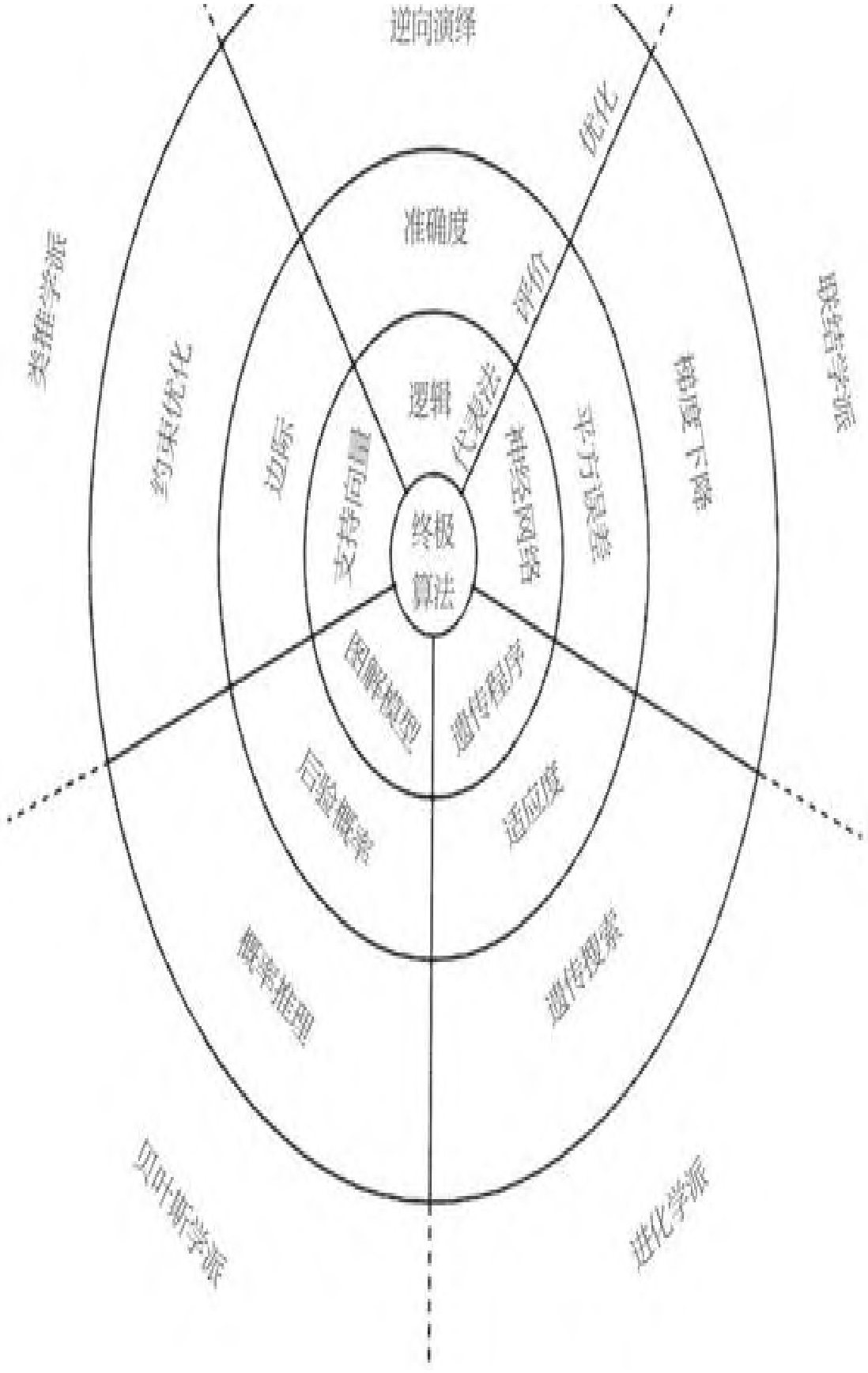

逆向演绎
优化
类推学派
联结学派
准确度
评价
约束优化
逻辑
代表法
边际
支持向量
终极算法
贝叶斯网络
统一学说
结构发现
图解模型
遗传程序
后验概率
适应度
概率推理
遗传搜索
贝叶斯学派
进化学派

图9－1

你专注地盯着地图，试图解开它的谜底。15块拼图都很完美地拼在一起，但你得弄清楚它们是如何结合起来的，且只分成3个区域：终极算法的代表方法、评价、优化部分。每种学习算法都有这3个部分，但不同学派也会有不同的组成部分。

代表法是一种形式语言，利用这种语言，学习算法会表达它的模型。符号学派的形式语言是逻辑，其中规则和决策树是特殊例子。联结学派的是网络神经。进化学派的是遗传程序，包括分类器系统。贝叶斯学派的是图解模型，这是贝叶斯网络和马尔可夫网络的涵盖性术语。类推学派的是特例，可能会有权值，就像在支持向量机中那样。

进化学派的主要部分是评分函数，判断一个模型的优劣程度。符号学派用的是准确度或者信息增益。联结学派利用的是连续误差测量，例如，平方误差，这是预期值与真实值之间差异的平方的总和。贝叶斯学派利用的是后验概率。类推学派（至少是支持向量机这种类型）利用的是边界。除了模型与数据的匹配度，所有的学派都会考虑其他合意的属性，例如模型的简洁度。

优化是一种算法，即寻找最高得分的模型，并回归它。符号学派的特色搜索算法是逆向演绎。联结学派的是梯度下降。进化学派的是遗传搜索，包括交叉和突变。贝叶斯学派在这方面不同寻常：他们不只是寻找最好的模型，而是寻找所有模型的平均值，由它们的可能程度来权衡。为了有效进行加权，他们利用诸如MCMC之类的推理算法。类推学派（或者更准确地说是支持向量机）利用约束优化来找到最佳模型。

经过一天的跋涉之后，太阳快速接近地平线，而你要在天黑前加快速度。城市的外围墙有5个大门，每个门由一个学派掌控，并通往每个学派在“优化城”里的区域。对着防护装置念出“深度学习”的口

令之后，我们进入“梯度下降大门”，并旋转进入“代表法之塔”。大街从门口处突然沿着山坡上升，一直延伸到城堡的“平方误差大门”，但你却要向左转向进化区域。梯度下降区域的房子都是平滑曲线以及密切交织在一起的模式，看起来几乎更像一个丛林，而不是一座城市。但当梯度下降屈服于遗传搜索时，情况就会急剧变化。到这里时，房屋会升得更高，一层结构堆在另一层之上，但这些结构稀少，几乎没有，似乎等着由梯度下降的曲线来填满。就是这样：结合两者的方法就是利用遗传搜索来找到模型的结构，然后让梯度下降来填满它的参数。这就是自然所做的事：进化创造大脑结构，而个人经历则对这些结构进行调整。

第一步完成了，你急忙赶往贝叶斯区域。即使在远处，你也可以看到它如何簇拥着“贝叶斯定理大教堂”，“MCMC小巷”一路蜿蜒曲折。这个过程需要一段时间。你抄近道走上“置信传播街道”，但似乎它永远在绕圈循环，然后你看到它了——“最可能大道”向着“后验概率大门”傲然屹立。不用对所有模型求均值，你可以直接奔向最适合的那个，并相信最终的预测会几乎一样。而你可以让基因搜索选择模型的结构和梯度下降作为其参数。轻轻松了一口气，你意识到这就是你需要的所有概率推理，至少持续到开始利用模型回答问题的时候。

你不断向前走。约束优化区域是窄巷和死胡同组成的迷宫，各种各样的例子到处紧挨，偶尔会对支持向量的周围进行清理。显然，为了避免遇到错误类别的例子，你要做的就是将限制条件添加到已经聚合的优化控制器中。但回头想想，这并不一定有必要。现在想起来了，甚至这也没必要。当我们学习支持向量机时，为了避免过拟合，已经让边际受到侵犯，只要每次违反就会受到惩罚。在这种情况下，优化例子权值可以再次通过梯度降低的形式来确定。这很容易，你觉得自己开始对它有所了解了。

实例稠密的排名突然中断，你发现自己在逆向演绎区域，这里有宽阔的大街和古代的石头建筑。这里的建筑是几何学的、朴素的，由直线和直角组成。即使经过很大程度修剪的树有矩形的树干，但它们的叶子被细致地贴上预测的类别。该区域的居民似乎用特殊的方式在建造他们的房子：他们从屋顶开始，将屋顶标识为“结果”，并渐渐地将天花板与地板之间的空隙填补，他们将地板标识为“前提”。一个接一个地，他们找到一块恰好与某个空隙吻合的石头，然后将其移到合适的位置。但你注意到，许多空隙有同样的形状，而在这些石头形成该形状之前，切割并整合石头也许会更快，然后尽可能多地盯住客人。换句话说，你可以利用遗传搜索来进行逆向演绎。巧妙！你似乎已经将那5个优化程序总结为一个简单的方法：遗传搜索用于结构，而梯度下降用于参数。即使那样也可能过分了。对于很多问题，如果你做3件事，可以将遗传搜索简化为爬山法——省去交叉，在每一代中尝试所有可能的点突变，并一直选择单个最优假设来为下一代播种。

头顶的雕像是什么？亚里士多德。他不以为然地看着那1/4梯度下降混乱地纠缠在一起。你已经充分进行循环，拥有了为得到终极算法而需要的统一优化器，但来不及庆祝一下，夜幕降临了，你要做的还有很多。通过壮观但相当狭窄的“准确度之门”，你进入“评价城堡”。门上的铭文写着：“汝等进入此地，须弃绝过拟合之希望。”随着你循环经过5个学派鉴别器的宫殿，在心理上将每个部分归位。为了连续性预测，你利用准确度来评价“是”或“否”预测以及平方误差。适应度只是进化学派用来指代评分函数的名字，你可以将它当作你想到的一切，包括准确度和平方误差。如果你忽略先验概率，而且误差遵循正态分布，那么后验概率会简化为平方误差。如果你允许它因为价格而受到违反，边际就会变成准确度更加柔软的版本——对于准确的预测不是不进行惩罚，对于不准确的预测则基于惩罚，惩罚一直等于零，直到你进入边际，这时它开始平稳增长。哟！结合鉴别器比结合优化器要简单得多，但“代表法之塔”落停在你上空，给你一种不祥的预感。

你进入搜寻的最后阶段，打开“支持向量之塔”的门，一位看起来面目狰狞的警卫打开它，你突然意识到自己看不到密码。“核心。”你脱口而出，努力使声音不显得丝毫慌张。警卫鞠躬然后避开。重新获得士气后，你走进去了，并会因为自己的粗心而踢自己一脚。塔的第一层是装修豪华的环形室，还有似乎为支持向量机的大理石表达法占据中心的重要位置。当你在它周围走时，注意到远的那边有一个门，它一定能够通过中心塔——“终极算法之塔”。门口似乎没有警卫，你打算抄近道，悄悄溜过门廊，走到一条不长的走廊，发现自己置身在一间更大的五角房屋中，每堵墙上都有一扇门。在中心，一个螺旋状的楼梯上升至你可以看到的位置。你听到上面的声音，然后躲进对面的门里。这个门通往“神经网络之塔”。你再一次进入环形的房间，这个房间的中心有一个多层感知器作为装饰品，它的组成部分和支持向量机不一样，但布局十分相似。你突然发现，支持向量机只是一个这样的多层感知器：有一个由内核而不是S曲线组成的隐藏层，还有一条输出信息，是一个线性组合，而不是另外一条S形曲线。

难道说别的表示方法也有相似的形式？越来越觉得刺激。你穿过五角星房屋，跑回“逻辑之塔”。看着中心一组规则的描述，你试着辨别一个模式。是的！每条规则只是一个高度程式化的神经元。例如，如果规则是“一只巨大的爬行动物，然后会喷火，那么它就是一条龙”，对于“它是一只巨大的爬行动物”和“喷火”以及门槛值1.5来说，仅仅是权值为1的感知器。一组规则集就是一个有隐藏层的多层感知器，包含每条规则的每个神经元，以及形成规则的稀缺的输出神经元。在脑海深处有一个疑问，但你现在没有时间解决它。当你穿过五角房屋走到“遗传程序之塔”时，就已经可以看到如何对它们进行合并。遗传程序也只是程序，而程序只是逻辑构造。房屋中遗传程序的雕塑是树的形状，子程序会分成更多的子程序，当你仔细看叶子时，会发现它们只是简单的规则。因此程序归结为规则，而且如果规则可以简化为神经元，那么程序也可以。

不幸的是，在“图解模型之塔”的顶部，环形房屋内的雕像与其他的看起来大不一样。图解模型是各种因素的一个产物：条件概率（在贝叶斯网络条件下）以及状态的非负函数（在马尔科夫网络条件下）。尽管很努力，你就是看不到神经网络或者规则集之间的连接，失望向你侵袭而来。不过你又戴上“护目镜”，其算法可以代替每个函数。有了！此时各种因素的产物就是术语的总和，就像一个支持向量机、一个投票规则集或者一个多层支持向量机，没有输出S形曲线。例如，你可以将一个朴素贝叶斯龙分类器解释成一个感知器，对于“喷火”，它的权值等于P（喷火｜龙）的log值减去P（喷火｜不是龙）的log值。当然，图解模型比这个还要概括，因为它们可以代表许多变量的概率分布，而不仅仅是给定其他变量（属性）情况下的一个变量（类别）分布。

你做到了！还有疑问？吸收支持向量机变成神经网络，吸收神经网络变成图解模型：起作用了。那么一定要吸收遗传程序变成逻辑。但将逻辑和图解模型结合在一起？此处有一点差错。你终于看到问题所在：逻辑有一个维度，而图解模型没有，反过来也一样。五个房屋中的雕塑匹配是因为它们是简单的寓言，但事实并非如此。图解模型不会让我们表示规则时涉及的对象超过一项，就像“朋友的朋友是朋友”，它们所有的变量必须是来自同一对象的属性。它们也不能代表任意程序，因为这些程序会通过一组组通过来自这个和那个子集的变量。逻辑可以轻易做到这些事，但它无法代表不确定、模糊或者相似度。而且如果没有一种能做所有这些事的表示方法，就无法获得通用学习算法。

你绞尽脑汁想办法，但你越努力，问题就变得越棘手。也许将逻辑和概率统一起来超出了人类的能力。筋疲力尽的你渐渐入睡。突然，一声低沉的吼叫把你吵醒了。一只多头复杂怪物扑向你，张牙舞爪，你在最后一刻躲开了。你用学习这把剑猛击怪物，这是唯一能杀死怪物的武器，最后你赢了，把它所有的头砍了下来。趁它还没长出新的头，你跑上了楼。

艰难地攀爬之后，你到达楼顶。此时有一场婚礼正在进行。珀莱尔迪卡特斯（Praedicatus，逻辑领域的大臣、象征界的统治者以及程序的保护神）对马尔科维雅（Markovia，概率领域的公主、网络界的女皇）说：“让我们把所有领域都统一起来，你应当往我的规则添加权值，以产生新的代表方法，可以进行广泛传播。”王子说：“我们把我们的一个个后代称作马尔科夫逻辑网络。”

你的头有些晕。你走到阳台外，太阳已经在城市上空升起。你的目光从屋顶投向广阔的乡村。服务器之林向各个方向延伸，发出很低的嗡嗡声，等待终极算法的到来。护卫沿着道路移动，从数据宝藏中取出金子。西边很远的地方，土地被信息之海占据，散布着船只。你抬头看终极算法的旗帜，现在可以清晰地看到在五角星中的铭文：

$$P=e^{w \cdot n} / Z$$

你想知道这到底有什么含义。

马尔科夫逻辑网络

2003年，我开始思考如何将逻辑和概率统一起来，我的学生马特·理查森也加入进来。一开始我们进展甚微，因为我们正试图利用贝叶斯网络来完成这件事，而它们的固定形式——变量的严格顺序、依据父母对孩子进行的有条件分布——与逻辑的灵活性不相容。但在平安夜的前一天，我意识到有一个好很多的办法。如果我们切换到马尔科夫网络，可以利用任何逻辑公式作为马尔科夫网络特征的模板，这样就可以把逻辑和图解模型统一起来了。看看我们怎么做到的。

回想一下，马尔科夫网络是由特征的加权和来定义的，这一点很像感知器。假设我们有一些人类的图片，随机拿起一张，然后依据它

的特征来算出概率，例如，“这个人有白头发”、“这个人年纪大了”、“这个人是一个妇女”等。在感知器中，我们利用一个门槛值来使这些特征的加权和通过，以决定比如，这个人是不是你的祖母。在马尔科夫网络中，我们会做一点不一样的事（至少乍一看是这样的）：我们对加权和进行指数化，将其变成因素的一个产物，而且该产物是从这批图片中选定某张指定照片的概率，不管你的祖母在不在这张照片里。如果你有许多老年人的照片，那么特征的权值就会上升；如果他们中的多数人都是男人，那么“这个人是一个妇女”的权值就会下降。特征可以是我们想要的任何东西，使马尔科夫网络变成一种很灵活的代表概率分布的方法。

实际上，我撒谎了：因素的产物还不是概率，因为所有图片出现的概率加起来一定等于1，而且不能保证所有图片的因素产物都会这样。我们需要使它们正常化，这就意味着通过它们所有的总和来划分每个产物。那么所有正规化产物的总和就可以保证是1，因为它仅仅是一个由自己划分的数字。那么一张照片的概率就是它的特点的权值和，经过了指数化和正规化。如果你回头看五角星中的等式，你可能会微微懂得它的含义。P是一个概率，w是权值的一个向量，n是数字的向量，而它们的点积“•”被Z指数化和划分，Z是所有产物的总和。如果我们让第一个组成部分n为1，如果图片的第一特征不是真就是零等，那么$w \cdot n$也仅仅是我们一直以来谈论的、特征权值总和的速记方法。

因此根据马尔科夫网络，等式会给出图片（或者随便别的东西）的概率。但它比那更概括，因为它不仅仅是马尔科夫网络的等式，而是马尔科夫逻辑网络的等式，正如我们称呼它的那样。在马尔科夫逻辑网络（MLN）中，数字n不一定只是0或者1，而且它们不指代特征——它们指代逻辑公式。在第八章的结尾部分，我们看到自己如何超越马尔可夫网络到达关联模型，这些模型依据特征模型而不只是模型来定义。“爱丽丝和鲍勃都患有流感”是特定于爱丽丝和鲍勃的特点。“X和Y都有流感”是一个特征模型，可以和爱丽丝和鲍勃、爱丽

丝和克里斯等其他任意两人实例化。特征模型很有力量，因为它可以利用单个短表达式来概括几十亿甚至更多的特征。但我们需要正式的语言来定义特征模型，而我们有一个可轻易得到的语言——逻辑。

一个MLN只是一组逻辑公式及其权值。当应用到特定组的实体时，在这些实体可能状态的基础上，它定义了马尔可夫网络。例如，如果实体是爱丽丝和鲍勃，可能的状态就是爱丽丝和鲍勃是朋友，爱丽丝患有流感，鲍勃也一样。让我们假设MLN有两种公式：“每个人都患有流感”和“如果某些人患有流感，那么他们的朋友也患有流感”。在标准逻辑中，这就是一对毫无用处的陈述：第一个陈述会排除所有陈述，即使是包含一个健康人的陈述也会被排除；而第二个就会变得多余。但在MLN中，第一个公式仅仅意味着有一个特征对于每个人X来说“X患有流感”，和公式一样有相同的权值。如果人们可能有流感，公式就会有较高的权值，相应的特征也会有。有很多健康人的陈述的可能性，会比有少量健康人的陈述要小，但也不是不可能。第二个公式，“某些人患有流感，而他们的朋友没有”，这个陈述的可能性要比健康人和受感染人群朋友群不一样的陈述的可能性要小。

这时可能你已经猜到n在终极算法中代表什么：它的第一个组成部分是陈述中第一个公式真实实例的数量，第二个是第二个真实实例的数量，以此类推。如果我们在观察一个有10个朋友的小组，而他们当中有7个人患有流感，那么n的第一个组成部分就是7，以此类推（如果7/20，而不是7/10的朋友有流感，难道概率不应该不一样吗？是的，不一样，因为有Z）。在极限情况下，如果我们让所有权值趋于无穷大，马尔可夫逻辑就会简化为标准逻辑，因为违背公式的单个实例，会引起概率暴跌至0，使陈述变得不可能。在概率方面，当所有公式讨论单个话题时，MLN会简化为马尔可夫网络。因此马尔可夫逻辑包含逻辑和马尔可夫网络两个特殊案例，而这是我们过去寻找的统一效果。

掌握MLN意味着发现世界上真实存在的公式要比随机因素预测出的频率要高，并且需要弄明白那些公式的权值：公式使它们的预测概率

与它们的观察频数相匹配。一旦我们掌握了MLN，就可以利用它来回答诸如“如果鲍勃和爱丽丝是朋友，而且爱丽丝患有流感，那么鲍勃患有流感的概率是多少”此类问题。你猜怎么着？原来概率是由S形曲线提供的，该曲线应用于特征的加权和，这和多层感知器很相似。而包含长串规则的MLN可以表示深度神经网络，在规则串中，每层会有一个链接。

当然，不要因为上文提到的简单MLN预测了流感而被其欺骗。想象一下用MLN诊断和治愈癌症。MLN代表细胞状态下的概率分布。细胞的每个部分、每个细胞器、每条代谢途径、每个基因及白细胞，在MLN中都是一个实体，而MLN的公式会对它们之间的差别进行编码。我们可以询问MLN：“这个细胞癌变了吗？”然后利用不同的药物对它进行调查，看看会发生什么。我们还没有这样一个MLN，但在本章后面我会设想它会产生。

总的来说，我们到达的统一学习算法利用MLN作为表示方法，利用后验概率作为评估函数，利用与梯度下降结合的基因搜索作为优化器。如果我们愿意，可以轻易地利用其他准确度测量方法来代替后验概率，或者利用爬山法来代替遗传搜索。我们上升到一座高峰，现在我们可以享受风景了。但是，我不会那么轻率地将这个学习算法称作终极算法。一方面，布丁的味道好不好，吃了才知道，而且虽然过去10年这种算法（或者它的变化版本）已经成功应用于许多领域，但还是有许多领域没有应用到它，因此还不清楚它的一般用途。另一方面，它还有一些重要的问题没有解决。在看这些问题前，让我们先看看它能做什么。

从休谟到你的家用机器人

你可以从alchemy.cs.washington.edu这个网站下载我刚才描述的学习算法。用“炼金术”来为它命名是为了提醒我们，虽然它取得了成功，但机器学习仍然处于科学的炼金术阶段。如果把它下载下来，你会看到它包含的东西要比我描述过的基础算法多得多，但是它也缺少几个我之前说的通用学习算法该有的东西，比如交叉。虽然如此，为了简单，让我们利用“炼金术”来指代我们的备用通用学习算法。

炼金术解决了休谟的原始问题，方法就是除了数据，要拥有其他的输入东西：你的原始知识，依据一组逻辑公式，有还是没有权值。公式可能不一致、不完整甚至完全错误，学习以及概率推理行为会关照那件事。关键点在于炼金术不必从零开始学习。实际上，我们甚至可以用炼金术来让公式保持不变，然后只对权值进行学习。在这种情况下，给予炼金术恰当的公式可以将其变成玻尔兹曼机、贝叶斯网络、实例学习算法以及许多别的模型。这解释了为什么虽然存在“天下没有免费的午餐”的定理，但可以有通用学习算法。更确切地说，炼金术就像一台归纳图灵机，我们可以对图灵机进行编程，使其行动起来像非常有力量的或者受到极大限制的学习算法，这取决于我们。炼金术为机器学习提供统一物，就像互联网为计算机网络、关联模型为数据库，或者图形用户界面为日常应用提供统一物那样。

当然，虽然你使用的炼金术没有原始工具（而且你可以），这并不会让它不包含知识。形式语言、评分函数、优化器会暗自对关于世界的猜想进行编码。因此可以自然地问我们是否有比炼金术更通用的学习算法。当进化从第一个细胞开始它的长途之旅，一直到当今所有的生命形式时，进化会做出什么假设？我认为有一个简单的假设，所有其他的假设都会遵循这个假设：学习算法是世界的组成部分。这就意味着，学习算法作为一种实体系统，和它的环境一样会遵循相同的规律，无论这个规律是什么，我们都已经暗暗“知道”它，并做好准备去发现它。在接下来的部分中，我们会看到这个规律具体意味着什么，以及如何在炼金术中体现它。就目前而言，我们注意到，它可能是我们能回答的休谟问题中最好的答案：一方面，假设学习算法是世

界的组成部分是一种猜想——原则上是这么说，因此它满足休谟的宣言，认为学习仅适用于先验知识；另一方面，这是一个多么基础而且难以反驳的猜想，也许这就是我们对于世界所需要的所有东西。

在另一个极端，知识工程师——机器学习最坚定的批评者——有好的理由来喜欢炼金术。不用基本模型结构或者几次大概的猜想，炼金术可以导入一个大型、细致的装配知识基地（如果它可用）。因为概率规则与确定性规则相比，互动的方式更为丰富，所以手工编码的知识在马尔科夫逻辑中会更深入。马尔科夫逻辑中的知识基地不一定要首尾一致，它们可以变得很大并容纳许多不同的主要因素而不会崩溃——这是一个目前为止让知识工程师迷惑的目标。

虽然如此，很多时候炼金术会解决机器学习中每个学派钻研了很久的问题。让我们反过来看看它们是什么。

符号学家结合运算中不同碎片的知识，这和数学家结合公理来证明定理是一个道理。这和神经网络和其他有固定结构的模型迥然不同。炼金术利用逻辑来完成这件事，正如符号学派所做的那样，却是一个迂回曲折的过程。为了证明逻辑中的定理，你需要发现产生该定理的公理应用的唯一顺序。因为炼金术从可能性的角度谈到其实它要做得更多：找到公式的多重序列，引出定理或者它的否定面，然后对它们进行权衡，以计算出定理真实的概率。这样它能推理的不仅与数学共性有关，还可以推理在一则新闻报道中的“总统”指的是不是“巴拉克·奥巴马”，或者一封邮件该在哪个文件夹之下。符号学派的主算法——逆向演绎，假定新的逻辑规则，用于作为数据与预期结论之间的步骤。炼金术通过爬山法来介绍新的规则，以原始规则作为开始，与原始规则和数据相结合，构建使结果概率更大的规则。

关联学派模型的灵感来源于大脑，S形曲线的网络与神经元相对应，它们之间的加权关联对应的是突触。利用炼金术，如果两个变量在某个公式中一起出现，那么它们就被联结在一起，考虑到近邻，一个变量的概率就是一条S形曲线（我不会解释为什么，这是我们在前面

部分看到的主算法的直接结果）。联结学派的主算法是反向传播，他们利用它来弄清楚哪个神经元负责哪些误差，并根据它们的权值来进行调整。反向传播是梯度下降的一种形式，炼金术用它来优化马尔科夫逻辑网络权值。

进化学派利用遗传算法来模仿自然选择。遗传算法包含假设的“人口”，在每一代中，适应能力最强的会进行交叉和变异，以生产下一代。炼金术以加权公式的形式来维持假设的“人口”，在每一步中以不同的方法对它们进行改变，然后把那些最能提高数据（或者一些其他的评分函数）后验概率的变量留下。如果人口是一个单一假设，这就简化为爬山法了。当前炼金术的开源实施并不包含交叉，但这就是一个简单的加法。进化学派的主算法是遗传编程，应用交叉和突变于计算机程序上，以子集树的形式表示出来。子集树可以用一组逻辑规则来表示，Prolog编程语言就是这么做的。在Prolog中，每个规则与一个子集对应，而它前面的规则是它称呼的子集。因此我们可以将带有交叉的炼金术当作利用Prolog相似编程语言的遗传编程，这样会有一个额外的优势——规则可以是概率性的。

贝叶斯学派相信模拟不确定性是学习的关键，并利用形式表示方法如贝叶斯网络和马尔科夫网络来工作。正如我们已经看到的那样，马尔科夫网络是MLN的特殊类型。贝叶斯网络利用MLN主算法也可以轻易表示出来，一个特征对应一个变量及其起源的每个可能状态，而相应条件概率的算法则作为它的权值（然后归一化常数Z就便利地简化为1，意味着我们可以忽略它）。贝叶斯的主算法是贝叶斯定理，执行的方法是利用概率推理算法，比如置信传播和MCMC。正如你已经注意到的那样，贝叶斯定理是主算法的特例，包含$P=P(A|B)$，$Z=P(B)$，而特征和权值对应$P(A)$和$P(B|A)$。炼金术系统包含置信传播和MCMC用于推理，推广至处理加权逻辑公式。将概率推理用于对逻辑提供的证明路线上，炼金术为支持或者反对结局来权衡证据，并输出结局的可能概率。这与符号学派利用的“单纯功能”逻辑相反，该逻辑不是全有就是全无，因此当给定矛盾证据时，该逻辑就会崩溃。

类推学派的学习方法是假设有已知相似属性的实体，也会有相似的未知属性：拥有相似症状的病人也会有相似的诊断，过去喜欢买某类书的读者可能未来还会买同样的书，等等。MLN可以代表具有像“拥有相同品位的读者会买相同的书”公式的实体之间的相似性。爱丽丝和鲍勃买到的相同的书越多，他们越可能有相同的品位；（将相同的公式应用于相反方向中）如果鲍勃买一本书，爱丽丝也买这本书的概率就比较大。他们的相似性由他们有相同品位的概率来表示。为了使这一方法真正起作用，对于相同规则的不同实例，我们可以有不同的权值：如果爱丽丝和鲍勃都买了某本罕见的书，这包含的信息量也许会比他们购买畅销书的信息量大，因此应该有更高的权值。在这种情况下，相似度由我们计算的属性是离散的（买或者不买），但我们也可以在连续属性之间表示相似度，就像两座城市之间的距离一样，方法就是让MLN把这些相似性当作特点。如果评估函数是一个边际风格评分函数，而不是后验概率，结果就是支持向量机的概括版，也就是类推学派的主算法。我们的主学习算法面临的更大挑战是复制结构地图，这是类比更强大的类型，可以从这个域（比如太阳系）推断到另外一个域（原子）。我们可以通过学习公式来做到这一点，这些公式不会指代源域中的任何特定关系。例如，“抽烟的人的朋友也抽烟”讲的是友谊和抽烟，但“相关的实体有相似的属性”适用于任何关系和属性。我们可以通过概括“朋友的朋友也抽烟，专家的同事也是专家”来掌握它，还有在社交网络中其他这样的模式，然后将其应用到诸如网页之类的地方，包含像“有趣的网页会链接到有趣的网页”这样的实例，或者应用于分子生物学中，包含像“与蛋白质调控互动的蛋白质也会调控基因”这样的实例。我的小组的研究人员和其他人已经完成所有这些事，接着还会做更多。

炼金术也可以启动我们在第八章提到的5种无监督学习。显然，它会进行关系学习，实际上，这就是它很多已应用的地方。炼金术利用逻辑来代表实体之间的关系，以及用马尔科夫网络来使它们自己变得不确定。我们可以将炼金术变成强化学习算法，方法就是利用奖励环

绕它，并利用它来掌握每种状态的值，这和传统强化学习算法利用的方法一样，比如神经网络。我们可以在炼金术内部进行组块，方法就是添加一种新的运算，这种运算会将一串串的规则变成单个规则（例如，“如果A那么B，如果B那么C，如果A那么C”）。一个带有单个未观察到变量（且该变量与所有可观察变量相联结）的MLN会进行聚类（未观察到变量是这样一个变量：我们从未在数据中见过它的值，是“隐藏的”，可以这么说，只能通过推断得出）。包含一个以上未观察到变量的MLN，通过由可观察变量（数量更多）推断那些变量（数量更少）的值的方法，来进行离散降维。炼金术也可以利用连续未观察到变量来处理MLN，这些变量会在诸如主要成分分析、等距映射的过程中用到。因此，原则上说，炼金术可以做所有我们让机器人罗比做的事，或者至少所有我们在本书中讨论过的事。的确，我们已经利用炼金术来让机器人掌握其所在环境的地图，通过它的传感器弄明白哪里是墙、哪里是门，以及它们的角度和距离等，这是制造称职家用机器人的第一步。

最终，我们可以将炼金术变成一种元学习算法，比如通过将个体分类器编码成MLN来进行堆叠，添加公式或者学习公式以将它们结合起来。这就是DAPRA在其个性化学习助手（PAL）项目中所做的。PAL是DARPA历史上最大型的人工智能工程，也是Siri的祖先。PAL的目标是构建自动化秘书。它利用马尔科夫逻辑来作为其支配一切的表示方法，将来自不同模块的输出融合到该做什么的最终决定中。这也允许PAL的模块通过追求统一意见来彼此互相学习。

目前炼金术最大的应用在于从网络中学习语义网络（或者谷歌所谓的知识图谱）。一个语义网络就是一组概念（就像行星与恒星）以及这些概念中的关系（行星环绕恒星）。炼金术会通过从网页中提取的事实（例如，地球绕着太阳转）来掌握100万个以上这样的模式。它会独自发现诸如行星之类的概念。我们之前使用过的版本，比我在此已经描述过的基本版要先进，但基本思想是一样的。各式各样的研究已经利用炼金术或者他们自己的MLN执行工具来解决自然语言处理、计

算机视觉、动作识别、社会网络分析、分子生物学及其他许多领域中的问题。

尽管成功了，炼金术还是有一些明显的不足之处：它还没有真正扩展到大数据，而一些在机器学习领域没有博士学位的人会发现它很难使用。因为这些问题，它还没有准备就绪，让我们看看能对它们做些什么。

行星尺度机器学习

在计算机科学中，如果一个问题没能得到有效解决，它就没有真的得到解决。如果你不能在可用时间和记忆内完成一件事，那么知道这件事怎么做并没有多大用处，因为在你处理MLN时，这些时间和记忆会消耗得很快。通常我们利用数百万的变量和数十亿的特征来对MLN进行学习，但这并不像表面看起来的那么大规模，因为变量的数量会随着MLN中实体的数量迅速增长——如果你有一个包含1000人的社交网络，你就已经拥有100万对可能的朋友，以及10亿个“朋友的朋友是朋友”公式的实例。

炼金术中的推理是逻辑与概率推理的结合。前者通过证明定理来完成，后者则通过置信传播、MCMC以及我们在第六章中见过的其他方法来完成。我们将两者结合至概率定理证明中，统一推理算法可以计算任意逻辑公式的概率，是当前炼金术系统的关键部分。但在计算法方面，它需要的很多。如果你的大脑利用概率定理证明，那么一只老虎会在你没来得及想明白如何逃离时就把你吃掉，这就是马尔科夫逻辑一般性的高昂代价。你的大脑在真实世界中已经进化，必须对额外的猜想进行编码，这样有利于它进行有效推理。在过去几年中，我们已经开始弄清楚它们可能是什么，并将其编码入炼金术。

这个世界并不是由随意杂乱的相互关系组成的，它有等级结构：星系、行星、大陆、国家、城市、社区、你的家、你、你的头、你的鼻子、其顶部的细胞、它里面的细胞器、分子、原子、亚原子粒子。那么模拟世界的方法会用到MLN，也有等级结构。这是一个猜想的例子，学习算法和环境都相似。MLN不需要知道，基于假定，世界由哪些部分组成。炼金术要做的，就是假设世界有若干个组成部分，然后将这些部分找出来。这很像一个刚做好的书架，假设有若干书，但还不知道哪些书会放在书架上。等级结构有助于使推理更易于处理，因为世界的子部分更多的是会与同部分的子部分互动，比如，邻居彼此之间的交流会多于与其他国家人民的交流，由同一个细胞产生的分子也会更多地与那个细胞的其他分子进行反应。

另外一个使学习和推理变得更容易的属性，就是存在于世界的实体不会以任意形式出现。它们会分成不同的类别和子类别，来自同一等级的成员之间的相似性会大于来自不同等级成员之间的相似性。有生命的或者无生命的，动物或者植物，鸟类或者哺乳类，人类或者非人类，如果我们知道与当前问题相关的所有特质，就可以将所有这些特质以及可以节省很多时间的实体合在一起。像以前那样，MLN不用知道基于假定的世界上的类别是什么，它可以通过分级聚类来从数据中掌握这些类别。

世界包含各个部分，而每个部分又属于不同类别，将这些部分和类别结合起来，这样使炼金术中的推理变得易于处理所需要的东西就已经基本齐全。我们可以掌握世界的MLN，方法就是将其分解成若干部分和子部分（这样多数互动关系就存在于同一部分的子部分之间），然后将各个部分集合成类别和子类别。如果世界是一个乐高玩具，那么我们将其分解为一块块“砖”，记住哪块和哪块相连，然后通过形状和颜色来将“砖块”集合起来。如果这个世界是维基百科，我们可以将它谈论的实体抽取出来，将它们集合成类别，然后学习类别之间如何相互关联。如果某人问我们：“阿诺德·施瓦辛格是动作明星吗？”我们可以回答：“是的，因为他是明星，而且他在动作片

里。”渐渐地，我们能够学习越来越大规模的MLN，直到去做我的一位谷歌朋友称作“世界范围的机器学习”的事：同时模拟世界上的每个人，数据会不断流入，答案则不断流出。

当然，如此规模的学习需要的不仅是直接执行我们见过的算法。例如，除这一点之外，单个处理器还不足够，我们得在许多服务器上分配学习。工业和学术界的研究人员已经深入调查，例如，如何并行使用多台计算机来进行梯度下降，其中一个选择就是在处理器之间划分数据，另外一个就是划分模型的参数。在每个步骤之后，我们将结果整合起来，然后重新分配工作。不管怎样，在不让交流成本压倒你，或者不让结果的质量受困扰的情况下完成这件事，做到这一点绝非易事。另外一个问题就是，如果你有无数数据串要输入，那么你在实现某些决定之前，不能等着看数据。一个解决办法就是利用采样原理：如果你想预测谁会赢得下一届总统选举，就不必问每个选民会投给谁，几千个人的样本已经足够（如果你愿意接受一点不确定）。诀窍就是把这个推广到拥有几百万个参数的复杂模型中。但我们只能通过这样的方法做到一点，即在每一步中尽量从数据流中举例子，因为我们要十分肯定自己做了正确的决定，并且所有决定的不确定性保持在允许范围内。利用这种方法，我们就可以在有限的时间内从无限数据中学到东西，就像我在一篇提议这种方法的论文中提及的那样。

大数据流系统是塞西尔·德米尔机器学习的产品，拥有数千个服务器，而不是数千个附加设备。在最大型的项目中，只要把所有数据汇集起来、验证数据、清除数据、将数据处理成学习算法能消化的形式，这样做，建造金字塔就简单得像在公园散步一样。

欧洲的“未来信息和通信技术”项目旨在建造整个世界（真正含义即是字面意思）的模型。社会、政府、文化、技术、农业、疾病、全球经济等不能落下一样。这当然还不成熟，但它却预示着未来要发生的事。同时，像这样的项目可以帮助我们找到可扩展性的极限在哪里，以及如何克服这些极限。

计算复杂性是一件事，人类复杂性是另外一件事。如果计算机就像白痴专家，那么学习算法有时会遇到少年天才乱发脾气的情况，这也是那些能征服他们的人类会收获颇丰的原因之一。如果你知道以专业的手法调整控制旋钮，直到到达恰当的位置，那么魔法就会随之而来，以一连串超越学习算法年份见解的形式出现。

而且，和德尔斐神谕并无两样的是，解释学习算法的宣言本身需要大量的技能。虽然如此，如果转错手柄，学习算法可能会涌出一系列胡言乱语，或者在反抗中保持沉默。不幸的是，就这一点而言，炼金术并不比多数算法好。把你知道的用逻辑的形式写下来、输入数据、按下按钮才是好玩的部分。当炼金术返回准确而有效的MLN时，你会去酒吧庆祝；当它不返回时——很多时候会出现的状况——战争就开始了。问题存在于知识、学习、推理中吗？一方面，因为学习和概率推理，简单的MLN可以做复杂程序能做的事。另一方面，当它不起作用时，调试起来就困难很多。解决办法就是让它更加交互，有能力进行反省，并解释它的推理。这就会让我们向终极算法更加靠近一步。

医生马上来看你

治愈癌症的方法是一种能够输入癌症基因组，并输出杀死癌细胞药物的程序。我们现在可以想象，这样的程序（就叫它“CanceRx”）会长什么样。尽管外表上看起来简单，CanceRx是编出的程序中规模最大也是最复杂的程序，的确太大了，所以它只能在机器学习的辅助下建立。它的基础是一个关于活细胞如何运转的详细模型，人体内的每种细胞都会有一个子类别，还有关于细胞如何互动、支配一切的模型。该模型以MLN或者与其相似的形式，将分子生物学的知识与来自DNA测序仪、微阵列，以及其他许多来源的大量数据结合起来。有些数

据通过手动进行编程，但大部分是从生物医学文献中提取出来的。模型会不断进化，并将新实验的结果、数据来源、患者病历合并起来。最后，它就掌握每种类型人体细胞中的每条路线、调节机制、化学反应，这就是人体分子生物学的总和。

CanceRx会将其大部分时间用于搜索包含备选药物的模型。考虑到新药，该模型会预测药物对癌细胞和正常细胞的效果。当爱丽丝被诊断为癌症时，CanceRx会利用她的正常细胞和肿瘤来将它的模型实例化，并尝试所有可能的药物，直到它找到一种能杀死癌细胞而又不伤害健康细胞的药物。如果它找不到有效药物或者有效药物组合，那么它就会着手研发一种会治愈癌症的药，也许会利用爬山法或者交叉法来让它从现有的药物中发展起来。它会在研究中的每一步中，在模型上尝试备选药物。如果药物抑制了癌症，但仍然有一些有害的副作用，CanceRx会尝试调整它以摆脱副作用。在癌症变异前，模型就能预测可能的变异，CanceRx就会开出那些能在中途扼杀癌变异细胞的药物。在人性与癌症这场象棋游戏中，CanceRx将是将军。

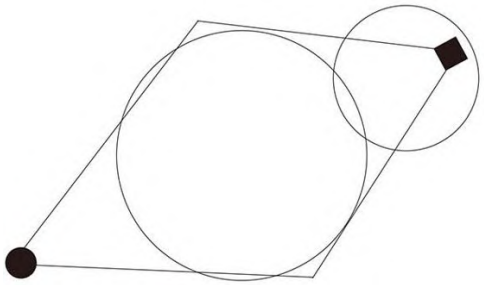
请注意机器学习不会独自把CanceRx给我们。这不像我们有一个巨大的分子生物学数据库整装待发，将这个数据库输入终极算法，弹出来的是活细胞的完美模型。在多次交互作用下，CanceRx可能是无数生物学家、肿瘤医生、数据科学家进行世界范围内合作的最终结果。然而，最重要的是，CanceRx会在医生和医院的协助下，从几百万个癌症病人中整合数据。没有数据，我们就无法治愈癌症；有了它，就能治愈癌症。为整个不断壮大的数据库做贡献，这不仅是每个病人的利益，还是它的伦理责任。在CanceRx的世界中，离散的临床实验是过去的事情。CanceRx提出的新治疗方法也正在继续被推出，如果这些治疗方法起作用，则会用到更广范围的病人身上。在经过改善的良性循环中，成功和失败的经历都会为CanceRx的学习提供宝贵的数据。如果你以某种方式看待它，机器学习可能只是CanceRx工程的一小部分，仅次于数据收集和人类贡献。但如果以别的方式看待它，机器学习就是整个企业的关键。没有它，我们只能掌握癌症生物学的零碎知识，散落

在数千个数据库和数百万篇科学文章中，每个医生只会注意一小部分。将所有这些知识集中成一个连贯整体，这超出了不受协助的人类的能力，无论人类多么聪明，只有机器学习才能做到这件事。因为每种癌症都不一样，它让机器学习找到通用模式。而且因为单个组织会产生几十亿个数据点，这就需要机器学习来弄明白能为每个新病人做点什么。

现在人们已经在努力建造最终会成为CanceRx的东西。在系统生物学的新领域中的研究员会模拟整个代谢网络，而非个体基因和蛋白质。位于斯坦福的一个团队已经构建了整个细胞的模型。全球基因组学与健康联盟推动数据在研究员和肿瘤学家之间的分享，着眼于大规模分析。癌症联盟会（Cancer Commons.org）收集癌症模型，让病人集中他们的病历，然后从相似例子中学习。基础医学（Foundation Medicine）查明病人肿瘤细胞中的变异，然后开出最合适的药物。10年前，人们不知道癌症是否可以或者如何被治愈，现在我们就能看到如何做到这一点。路还很长，但我们已经找到路了。

第十章

建立在机器学习之上的世界



既然你已经游览了机器学习仙境，现在让我们调整档位，看看这一切对你来说意味着什么。就像《黑客帝国》中的红色药丸一样，终极算法是通往不同现实状况的大门：你已经生存在这个现实当中，但对它还不了解。从约会到上班、从自我认知到社会的未来、从数据分享到战争、从人工智能的危险到进化的下一步，新的世界正在形成，而机器学习是解锁这个世界的钥匙。本章将会有助于你在生活中充分利用机器学习，然后为即将到来的东西做好准备。机器学习不会单独决定未来，这和其他技术一样，真正重要的是我们用它决定要做的事，现在你有了用于决定的工具。

在这些工具当中最主要的是终极算法。无论它什么时候会出现，无论它看起来是否像炼金术，这些都没有它涵盖的东西重要：一种学习算法不可或缺的能力，以及这些能力能让我们实现什么目的。我们也可以把终极算法当成现在和未来学习算法的合成图。我们可以很方便地将该合成图用在思维实验中，代替产品X或者网页Y中的特殊算法，这些产品或者网页所属的公司不太可能会分享它们的产品。由此

看来，每天与我们互动的学习算法是终极算法的萌芽版，我们的任务就是了解它们，推动其成长以更好地满足我们的需要。

在未来10年，机器学习会大范围影响人类的生活，只用一本书的一个章节无法描述清楚。即便如此，我们已经看到许多重复出现的主题，这些主题就是我们的关注点，并以心理学家所谓的“心理理论”作为开始，更确切地说，是以你的心理的计算机理论作为开始。

性、谎言和机器学习

你的数字化未来从一个感悟开始：每次你和计算机相互作用时——无论是你的智能手机，还是几千英里以外的服务器——你都会从两个层面上这样做。第一个层面是，当场就得到你想要的东西，如问题的答案、你想买的产品、一张新的信用卡。第二个层面，从长远来看也是最重要的一个，就是教会计算机关于你的东西。你教会它越多的东西，它就越能更好地为你服务（或者操纵你）。生活就是你和包围你的学习算法之间的游戏。你要么拒绝参与游戏，那么你就得在21世纪过20世纪的生活；要么从游戏中获胜。你想让你的计算机拥有你的哪个模型？为了产生那个模型，你能给它什么数据？当你和一种学习算法互动时，脑子里就应该一直思考这两个问题，就像你和其他人互动时一样。爱丽丝知道鲍勃对她有一个心理模型，并试图通过她的行为来塑造这个模型。如果鲍勃是她的上司，她努力给他能干、忠诚、努力工作的形象；如果鲍勃是她想诱惑的对象，她就以最性感的一面出现。对于别人在想什么这一点，如果无法凭直觉知道并做出反应，那么我们就很难在社会上活动。当今世界新奇的地方在于计算机（不仅仅是人类）也开始有心理理论了。它们的理论仍然有点粗糙，但发展得很快，而为了拿到我们想要的东西，不得不与它们进行合作

——不少于和其他人的合作。因此你需要计算机心理的理论，将评分函数（你认为学习算法的目标是什么，或者更准确地说，它的主人的目标是什么）和数据（你认为它知道的东西）带入之后，这就是终极算法要提供的东西。

以网上约会为例。当你利用Match.com、eHarmony或者OkCupid（都为美国知名的约会和社交网络平台）（有必要的话，暂停你的怀疑）时，你的目标很简单——找到最佳的可能约会对象。但很有可能，在你遇到真正喜欢的人之前会费很大劲，可能还会有几次令人失望的约会。一个顽强的呆子会从OkCupid上摘出两万条简介，做他自己的数据挖掘工作，然后在第88次约会中找到他的梦中女郎，然后将他这段漫长的旅程告诉《连线》杂志。为了减少约会次数、少费工夫，你的两大工具就是你的简介和你对推荐对象的反应。有一个受人欢迎的选择就是说谎（比如关于你的年龄）。这可能看起来不道德，更不用说当你的对象知道事实后，事情会搞砸，但这里有一个转折。对网上约会在行的人已经知道，人们会在简介中的年龄问题上撒谎，然后根据情况进行调整，所以如果你说出自己的真实年龄，实际上就告诉他们你比实际年龄要大。反过来，进行配对的学习算法认为，和真正选择的约会对象相比，人们宁愿选择更年轻的人。逻辑上的下一步，就是更大程度地谎报他们的年龄，最终解释这个属性将没有意义。

对于所有关心的问题，更好的办法在于集中于特殊、非比寻常、预测能够成功配对的属性上。在这种意义上，它们会挑出那些你喜欢但不是所有人都喜欢的人，因此竞争也就没有那么大了。你的工作（也包括你未来约会对象的工作）就是提供这些属性；媒人的工作就是掌握这些属性，和旧时媒人的做法一样。和旧时媒人相比，Match.com的算法有一个优势，它知道更多的人，但劣势在于它对这些人的了解没那么深。朴素学习算法，例如一台感知器，会满足于普遍化的观点，如“绅士都喜欢金发碧眼的女人”。更加复杂的学习算法会发现诸如“对音乐有另类品位的人往往适合在一起”的模式。如果

爱丽丝和鲍勃都喜欢碧昂斯，单凭这一点他俩很难配对在一起；但如果他们都喜欢艾伦主教，这点至少让他们成为灵魂伴侣的可能性变得更大。如果他们都是一个乐队的粉丝，但学习算法不知道这个乐队，那就更好了，但只有一种关系算法，如炼金术，才能不费力地掌握这一点。学习算法越好，就越值得你花时间让它了解你。但根据经验，你想让自己足够与众不同，这样它就不会将你和“普通人”混淆（记住第八章的鲍勃·伯恩斯），但也别太与众不同，这样它就没办法理解你了。

网上约会实际上是一个有点难以理解的例子，因为化学反应难以预测。两个在约会中合得来的人，可能最后会相爱，并坚信他们就是天生一对。但如果他们最初话不投机，可能会觉得对方烦人，不想再见面了。真正复杂的学习算法所做的，就是在每对似乎合理的一对情侣之间进行1000次蒙特卡洛模拟，然后通过那部分结果还不错的约会对这些情侣进行排名。简而言之，约会网站可以组织派对，并邀请那些对很多人来说可能会成为其伴侣的人，让他们在几个小时之内就完成需要几周完成的事情。

对于我们那些热衷于网上约会的人来说，更即时有用的办法就是选择记录哪些互动，以及在哪里记录。如果你不想让亚马逊对你的圣诞节购物品位产生疑惑，请在其他网站上进行（对不起了，亚马逊）。如果你在家看不同种类的视频，为了工作，在YouTube上保留两个账号，在家一个，工作时一个，YouTube会学着做出相应的推荐。如果你打算看一些一般不会感兴趣的视频，就先退出账号。使用谷歌浏览器的无痕模式，目的不是为了非法浏览（当然，你绝不会这么做），而是因为你不想让当前搜索影响到未来的个性化定制。在网飞上，利用你的账号来为不同的人添加简介，这样可以使你在家庭电影之夜免于R级片（即限制性影片）推荐。如果你不喜欢某家公司，可以点击它的广告，这样不仅能够即时花费它的钱，通过为那些不太可能购买产品的人展示广告，还可以教会谷歌来再次浪费它的钱。如果你有非常特殊的搜索项，想让谷歌未来能够准确回答，那么花点时间来

查阅后来显示结果的页面，看看有没有相关链接，然后点击链接。较为普遍的是，如果一个系统不断向你推荐错误的东西，通过找到并点击多个准确链接的方式来试图调教系统，然后返回来看看它是否起作用了。

虽然如此，但这可能是繁重的工作。不幸的是，所有这些所阐明的，就是当今你和学习算法之间的通信通道是多么狭窄。你应该有能力告诉它关于你自己的信息，以及你想要的东西有哪些，而不仅仅是让它间接从你的行为中学习。不仅如此，你应该有能力检查你的学习算法模型，然后按照期望对它进行修正。如果学习算法认为你在撒谎或者缺乏自我认知，那么它仍会决定忽视你，但至少这一次可以考虑你的输入信息。因为这一点，模型需以人类能理解的方式呈现，例如，需要规则集而不是神经网络，而且除了原始数据，它需要接受将一般陈述当作输入，正如炼金术一样。所有这些让我们想知道学习算法能有多好的一个关于你的模型，以及你会利用这个模型干什么。

数码镜子

花点时间来考虑你记录在世界上所有计算机里的数据：你的邮箱、办公文档、文本；推特、脸书和领英账号；你的网页搜索、点击、下载、购买；你的信用卡、传真、电话、健康档案；你的健康追踪器统计；你的汽车微处理器记录下的驾驶情况；你闲逛时被手机记录下来的信息；你拍过的所有照片；监控摄像机里的简短片段；你的谷歌眼镜片段。如果未来的传记作者没有什么可利用，除了你的“数据排放”，他会描写出怎样的一个你？也许是一个在许多方面都很准确、很详细的你，但也有可能缺失某些东西的你。为什么你会在一个风和日丽的日子决定改变职业生涯？自传作者提前预测到这一点了

吗？那么你某天遇见并偷偷难以忘怀的那个人呢？那位作者能够通过发现的片段返回来，然后说“啊，原来在那儿”吗？

有一个令人冷静（也许是安心）的想法，就是当今世界上没有哪种学习算法可以利用所有这些数据（甚至美国国家安全局也不可以）。即使有，该算法也不知道如何将数据变成逼真的你。假设你带着自己的所有数据，然后把数据交给真实的未来终极算法会怎样呢？该算法已经包含所有我们教过它的所有东西。它会学习关于你的一个模型，而你可以用指尖驱动那个模型，并把它放在口袋里携带，随意对它进行检查，然后将它用于你喜欢的东西。当然，这对内省法来说，是一个很好的工具，就像在镜子里看自己一样。它也会是一面数码镜子，不仅能够显示你的外表，还能显示所有关于你的、能观察到的东西——一面栩栩如生，并能和你对话的镜子。你会问它什么问题？你可能不会喜欢它的某些回答，但这就更有理由来好好考虑这些答案；有些答案可能会给你新的想法和方向。你的终极算法模型甚至可以帮你成为更好的人。

除了自我提升，也许第一件你想让自己的模型完成的事就是代表你与世界妥协，使它在网络空间放松下来，同时为你寻找各种各样的事物。从世界上所有的书中，它会给你推荐十几本你接下来可能想阅读的书，见解比亚马逊能想到的还要好。对于电影、音乐、游戏、衣服、电子产品来说，道理也一样——应有尽有。它可以让你的冰箱一直处于装满的状态，这是毫无疑问的。它可以对你的文本邮件、语音邮件、脸书帖子、推特信息进行过滤，而且在合适的时候会代表你回复这些消息。它还会为你处理生活中的所有小烦恼，比如查看信用卡账单、拒绝乱收费、做计划、更新订阅、填写纳税申报单。它会为你的疾病找到治疗方法，由你的医生来管理该方法，并从沃尔格林公司预订。它会让你注意到有意思的工作机会、提议度假胜地、建议你该为哪个候选人投票、寻找潜在的约会对象。另外，你和约会对象成功配对以后，它会与你约会对象的模型合作，为你们两人挑选彼此都喜欢的餐厅。这时事情才真正开始变得有意思起来。

充满模型的社会

在这个快速接近的未来社会中，你不是唯一拥有“数码另一半”的人（另一半会24小时按照你的要求办事），每个人都会有自己的详细模型，这些模型会一直互相对话。如果你正在找工作，而X公司正在招聘，它的模型会对你的模型进行面试。这很像一场真实、身临其境的面试——你的模型最好还是别主动提供你的负面信息等——但这个过程只会花不到1秒的时间。你在未来领英账号上点击“找工作”，会马上进行宇宙中所有公司的工作面试，虽然远，但与你的参数（专业、地点、薪资等）匹配。领英会马上反馈最佳公司列表，你可以从中挑选想进行细谈的公司。约会也是一样，你的模型会进行数百万次约会，所以你就不必了。星期六，你会在OkCupid组织的派对上认识最佳约会人选，你知道自己也是对方的最佳约会人选——当然，你也知道对方其他的约会人选也在屋里。这肯定是一个有意思的夜晚。

在终极算法的世界里，“我的人会联系你的人”会变成“我的程序会联系你的程序”。每个人都会有一个机器人随从，在这个世界游刃有余地存在。交易完成了、条款谈妥了、安排做好了，这些都会在你举起手指头之前完成。今天，医药公司会锁定你的一生，因为它决定给你开什么药；明天，你消费的每种产品或者服务的经销商会定位你的模型，因为模型会为你筛选。它们的机器人的任务就是让你的机器人来购买。你的机器人的工作就是看穿它们的口号，就像你看穿电视广告一样，但会比你看得更细致，你绝不会有时间和耐心来完成。买车之前，你的数码另一半会浏览所有的参数并和制造商讨论这些参数，然后研究世界上每个人对那辆车及其替代品的评价。你的数码另一半就像指引你生活的力量一样，它会去你想去的地方，但让你花费的时间比较少。这并不意味着你最终会陷入“过滤泡泡”的困境中，看到的只是你觉得可靠并喜欢的东西，意料之外的选择则排除在外，

你的数码另一半能更好地了解这一点，其特点包括对机遇留有余地、让你尽享新体验、寻找发现珍宝的运气。

即使会更有意思，但当你找到汽车、房子、医生或者工作之后，这个过程并不会结束。你的数码另一半会继续从经历中学习东西，就像你一样。它弄清楚什么能起作用、什么不能，不论是在工作面试、约会，还是在寻找房产的过程中。它代表你和人们、组织进行互动，并学习关于它们的东西，然后从你与他们的真实互动中掌握技能（这一点更重要）。它预测爱丽丝会是你很棒的约会对象，但你时间不太方便，因此它会假设可能的原因，并在你的下一轮约会中进行验证。它会把最重要的发现与你分享（“你觉得自己喜欢X，但实际上你更倾向于Y”），将你各种各样的住酒店经历和这些酒店在“猫途鹰”（TripAdvisor）上的评价相比较，它会弄清楚哪些小道消息是真的并在以后将其找出。它不仅掌握网上哪个商家值得信赖，还要学会如何解码那些不那么值得信赖的商家所说的话。你的数码另一半有一个世界模型——不只是一般的世界，还指与你产生关联的世界。当然，其他所有人也会有自己不断演进的世界模型。一段相互关系中的每一方都会向世界模型学习，并将其学到的东西运用到下一段相互关系中。你有每个和你有过相互关系的人以及组织的模型，而他们也会有你的模型。随着模型的改善，它们之间的相互关系就会变得越来越像你在真实世界中的相互关系一样——除了高出几百万倍的速度以及存在于硅片中之外。未来的网络空间会是一个巨大的平行世界，只会选择最有希望的东西在真实世界中进行试验，它就像一种新的全球性意识和人类身份。

分享与否？方式、地点如何？

当然，你独自一人了解这个世界会比较缓慢，即使你的数码另一半了解世界的速度会比实实在在的你高出数量级的倍数。如果其他人了解你的速度比你了解他们的速度要快，那么你就会陷入麻烦。解决办法就是要分享。100万个人了解一家公司或者一种产品的速度会比单个人的速度快很多，只要他们能够集中各自的经历。但你应该和谁分享数据？这也许是21世纪最重要的问题了。

当今你的数据可以分成四种：你和所有人分享的数据，你和朋友或者同事分享的数据，你和各种公司（不论是否有意）分享的数据，以及你不与别人分享的数据。第一种数据包括Yelp（美国最大的点评网站）、亚马逊、猫途鹰上的评论、易趣网的反馈评分、领英的简历、博客、推文等。这类数据价值巨大，是四类数据中问题最少的一类。你真的想让每个人都能用到这些数据，每个人也会从中受益。唯一的问题在于，掌握这些数据的公司不一定会允许对它们进行大量下载，以便用于构建模型。它们应该允许下载行为。时下你可以去猫途鹰，查看你正在考虑入住的指定酒店的评论和星级评分，但如果要查看酒店总体上是好还是坏的模型呢，而通过该模型，你可以对当前有极少可靠评论的酒店进行评分？猫途鹰可以掌握该模型，但如果你想要一个决定你对酒店感觉好坏的模型呢？这就需要关于你的，但你不想和猫途鹰分享的数据。你想要的，就是一个可信赖的、能将两类数据结合起来，并能给你结果的一方。

第一类数据应该不会存在问题，但实际上并非如此，因为它与第三类数据重叠了。你在脸书上与朋友分享更新、照片，你的朋友也和你分享，但每个人都会利用脸书来分享他们的更新和照片。幸运的脸书，它有10亿个朋友。渐渐地，它对于世界的了解比任何人都要多。如果它有更好的算法，就能了解得更多，而这些算法每天都会进步，这对我们数据科学家来说是一种恩惠。作为回报，它会为你的分享提供基础结构，这就是你使用脸书所做的交易。随着学习算法的改善，它由数据产生的价值会越来越大，有些价值会以更相关的广告、更优

质的服务的形式回馈你。唯一的问题在于，脸书也可以随意使用你不感兴趣的数据和模型，你却无法阻止它。

这个问题会伴随你与公司分享的数据一起突然彻底出现，包括如今你在线上、线下做的许多事情。也许你没注意到，其中会有一个收集你的数据的疯狂比赛。每个人都喜欢你的数据，这也难怪，它们是通往你的世界、你的钱包、你的投票甚至你的心灵的大门。但是每个人只能拥有它的一小部分：谷歌掌握你搜索的内容，亚马逊知道你网购的东西，美国电话电报公司会看到你的通话记录，苹果知道你下载的音乐，西夫韦懂得你购买的杂货，美国第一资本投资国际集团了解你的信用卡交易记录。诸如安客诚（Acxiom）之类的公司会整理并销售关于你的数据，但如果你可以对其进行检查（对于安客诚的情况，你可以在aboutthedata.com检查），数据并不多，而且有些还是错误的。没有人能够了解到完完整整的你。这有好处，也有坏处。有好处是因为如果某人做到了，他就会掌握很大的权力；有坏处是因为只要事实是那样的，就不会有你的360°模型了。你真正想要的是数码的你，以及你是唯一拥有者，其他人只有根据你的意愿才能获得东西，最后一类数据（你不想分享的数据）也存在一个问题，即也许你应该分享它。也许你没想到要那么做，也许做起来没那么容易，或者也许你只是不想那么做。如果是最后一种情况，你应该考虑道德上是否有义务来进行分享。我们见过的一个例子就是癌症病人，他们可以通过分享肿瘤的基因组和治疗史来为治愈癌症做贡献。分享的好处远远不止这点。所有关于社会和政策的各种问题也许都可以通过了解我们每天产生的数据来得到解决。社会科学正进入一个黄金时代——只要数据面向研究人员、政策制定者、老百姓。这并不意味着让别人窥探你的私人生活，而是说要让他们看到已经掌握的模型，而这个模型应该只包含统计信息。因此，在你和他们之间，需要一位可靠的数据经纪人，保证你的数据不会被滥用，也没有哪个免费使用者会在不分享数据的情况下就享受到好处。

总之，所有四类数据的分享都有问题。这些问题有一个共同的解决办法：新型公司与你的数据的关系，就像银行和你的钱的关系一样。银行不会偷你的钱（有也是极少数）。它们应该明智地对它进行投资，而且你的存款已经过FDIC（联邦存款保险公司）承保。时下有许多公司提出要加强你在云盘某处的数据，但这些数据与你的私人数据银行还有很大差别。如果它们是云提供商，则会把你限制起来——一项大禁忌（想象一下，你把钱存在美国银行，而且不知道你是否可以彻底把钱转账到富国银行）。一些新创公司提出要贮藏你的数据，并把数据交给广告商，以换取折扣；但对我来说，这样做没有抓住重点。有时，你想免费为广告商提供信息，因为这么做符合你的利益；有时，你一点也不想提供，而“什么时候分享什么数据”也是你的良好模型才能解决的问题。

我正在想象的那类公司会做以下几件事来赚取订阅费。它会对你的网上互动进行匿名处理，并通过服务器确定这些互动路线，然后通过其他用户将这些互动集合起来。它会把你这辈子所有的数据储存在一个地方——包括你每天24小时的谷歌眼镜视频流（如果你有）。它会对关于你和你的世界的完整模型进行学习，并对其进行持续更新。它会代表你使用模型，并一直做你要做的事，发挥模型的最大能力。公司对于你的基本承诺是，你的数据和模型绝不会在损害你利益的情况下被使用。这样的保证真的过于简单，毕竟你本身就无法保证你绝不会做损害自己利益的事。但公司的存在就取决于这样的保证，就像银行的存在取决于它保证不会弄丢你的钱一样，因此你应该信任这家公司，就像你信任你的银行一样。

这样的一家公司会很快成为世界上最有价值的公司之一。就像《大西洋月刊》的亚历克西斯·马德里加尔指出的那样，当今你的简介也许能通过一分钱或者更少的钱来买到，但一个用户对于互联网广告业的价值可能是每年1200美元，谷歌掌握的你的那部分信息价值约20美元，脸书的是5美元，等等。除此之外，还没有谁能全部拥有各部分的数据，而且完整的数据比各部分数据的总和要多——基于你所有

数据的模型，要比基于1000个部分数据的1000个模型要好很多——而我们正在以每年轻易超过1万亿的数据作为目标，相当于美国这样的经济体。利用这种数据创建一家《财富》500强公司并不会需要很多费用。如果你想接受挑战并最后成为一名亿万富翁，那么记住你首先是在什么地方得到这个想法的。

当然，当前的一些公司想拥有数码的你，谷歌就是其中一个。谢尔盖·布林说：“我们想让谷歌成为你大脑的第三个组成部分。”谷歌的一些收购和用户的数据流补充公司的数据库的好坏表现不无关系。但是，虽然诸如谷歌和脸书之类的公司处于领先地位，但它们并不适合作为你的数码家园，因为它们存在利益冲突。它们通过广告分析来谋生，因此得权衡你的利益和广告商的利益。你不会允许自己大脑的第一或者第二组成部分各有忠心，第三组成部分就更别提了？

如果你的模型看起来像罪犯的模型，也许会发生意想不到的事情，比如政府可能会传讯你的数据，甚至预防性地监禁你，有一点《少数派报告》的风格。为了抢先一步，你的数据公司可以对一切加密，让你来保管钥匙（如今你甚至不用解密数据就可以计算加密数据）。你也可以把它保存在家里的硬盘中，公司会把软件租给你。

如果不喜欢盈利实体拿着通往你王国的钥匙，可以加入一个数据联盟（如果在你的网络区域没有联盟，可以考虑启动一个）。20世纪需要工会来协调工人与老板之间的权利，21世纪出于同样的原因也需要数据联盟。公司和个人相比，收集和使用数据的能力要强大很多，这导致了权利上的不平衡。数据越有价值，就越能从中掌握更好、更有用的模型，不对称也就越严重。数据联盟让其成员与公司就其数据使用进行平等交易。也许工会能够使活动开展起来，并巩固其成员身份，方法就是为其成员开启数据联盟。但工会是根据职业和地理位置组织起来的，数据联盟就比较灵活——加入和你有很多共同点的人群，那样掌握的模型会更有用。请注意，加入数据联盟并不意味着可以让其他成员看到你的数据，这仅仅表示让每个人都能利用通过共享

数据掌握的模型。你的数据对世界的影响力和你的投票一样，或者会更大，因为你只会在选举日去投票处投票，其他时候，你的数据就是你的选票。站起来，表明你的立场！

目前为止，我还没有说到“隐私”这个词，这并非意外。隐私只是数据分享更大问题的一方面，如果我们在损害整体的情况下来关注它，就像当今争论大部分所关注的那样，那么就会有得出错误结论的风险。例如，除了最初的目的，法律禁止将数据用于其他用途，这就显得很缺乏远见（《苹果橘子经济学》中没有哪个章节是依据该项法律规定写的）。当人们利用隐私来换取其他好处时，正如当在网上填写简介时，隐私表现出的隐含价值比你问他们“你在意你的隐私吗”这种抽象问题的价值要低得多。但依据后者，隐私之争往往更容易陷入圈套。欧盟法院发布命令，人们有权利被忘记，但也有权利来记忆，无论是用他们的神经元还是硬盘。公司也一样，在一定程度上，用户、数据收集者、广告商的利益是一致的。浪费注意力对谁都没有好处，数据越好，产品也会越好。隐私并不是一场零和游戏，虽然有时它经常被当作零和游戏。

掌握数据的你和数据联盟的公司的样子，对我来说，看起来就像是在未来社会数据变得成熟一样。我们是否能到达那里有待研究。当下，多数人没有意识到有多少关于他们的数据正在被收集，以及潜在的代价和利益是什么。各家公司满足于继续神秘地完成这件事，因为担心引发谴责。但谴责迟早会发生，在后续的争论中，会制定更加严苛的法律，最后对谁都没有好处。最好让人们现在树立意识，选择该分享什么、不该分享什么，以及如何、在哪里分享。

神经网络抢了我的工作

你的工作会在多大程度上用到你的大脑？用得越多，你就越安全。在人工智能早期，人们普遍认为，计算机取代白领前会先取代蓝领，因为白领工作更费脑力；结果却并非如此。一方面，机器人组装汽车，但它们没有代替建筑工人；另一方面，机器学习算法已经取代信用分析员和直销商。其实，对于机器来说，评估信用申请表比走在建筑工地不被绊倒要简单，尽管对于人类来说恰恰相反。一个普遍的主题是，狭义定义的任务很容易通过数据来完成，但那些需要技能与知识广泛结合的任务却不能。你大脑的大部分都主管视觉和运动，这就意味着到处走走比表面看起来要复杂得多。我们之所以觉得很简单，是因为“到处走走”经过进化已经练习得近乎完美，所以很多时候是在潜意识中进行的。叙事科学（Narrative Science）这家公司有一种人工智能系统，可以写出很好的棒球比赛总结，却写不好小说，因为（根据乔治·威尔的观点）生活的内容要比棒球赛多很多。语音识别对计算机来说比较困难，因为要填补空白存在困难。字面上说，就是那些说话人平时会省略的发音（当你不知道那个人在讨论什么时）。算法可以预测股票波动，但不清楚如何将股票波动与政治联系起来。一项任务需要的背景信息越多，计算机能迅速完成它的可能性就越小。常识之所以重要，不仅是因为妈妈教会了你，还因为计算机里面没有这些信息。

防止丢掉工作的最佳办法就是你自己对它进行自动化，这样就可以把时间用在你之前顾及不到、计算机近期地无法做到的所有部分（如果没有什么任务无法完成，那么就要在行业保持领先地位，现在就去找一份新工作）。如果计算机已经学会完成你的工作，不要试图与它竞争，而要利用它。H&R Block公司（美国最大的报税服务商）仍在运营，但报税人的工作却没有以前那么枯燥了，因为现在计算机承担了大部分枯燥的工作（好了，也许这不是最佳例子，因为免税代码的指数级增长，是为数不多的能够与计算能力指数级增长相抗衡的东西）。把大数据看作你知觉的延伸，把学习算法看作你大脑的扩展。当下最佳棋手是所谓的人马怪（半人、半程序）。在其他许多职业中

情况也是如此，从证券分析师到棒球球探。这并不是人类与机器的对抗，而是有机器的人和没有机器的人之间的对抗。数据和直觉就像马和骑手，而你不会试图超过一匹马，你在驾驭它。

随着技术的进步，人和机器更加密切的结合体就形成了：你饿了，Yelp会推荐一些好吃的餐厅；GPS会指引你方向；你开车，汽车电子会进行低水平控制。我们现在都已经是半机器人了。真正的自动化指的不是它代替了什么，而是它增强了什么能力。一些行业消失了，但许多新的行业诞生了。最重要的是，自动化使各类事情成为可能，这些事情如果由人类完成，将要付出很多代价。ATM机（自动柜员机）代替了一些银行员工，但主要好处是它们让我们随时随地都可以取钱。如果像素要通过人类动画师来一次只为一个上色，那么就不会有《玩具总动员》和视频游戏了。

尽管如此，我们可以询问自己最终是否会彻底完成人类的工作。我觉得不会。即使这一天到来了（它不会很快就到），且计算机和机器人都可以把所有事情做得更好，但仍有一些工作会留给一些人。机器人也许可以很好地模仿酒保，甚至可与客人闲聊，但老顾客仍然会更喜欢一个他们认知的人类酒保，仅仅因为他们就是人类。拥有人类服务员的餐厅会有额外标志，就像手工制品那样。人类还是会去剧院、骑马、航行，虽然我们已经有电影、汽车、摩托艇了。更重要的是，一些职业真的无法替代，因为它们的工作需要一种计算机和机器人在定义上无法拥有的东西：人类经历。所谓人类经历，指的并不是人际互动工作，因为人际互动要造假也不难，比如机器人宠物的成功。我指的是人文科学，准确地说，其领域包含一切没有人类体验就无法理解的东西。我们担心人文科学正呈死亡螺旋下降趋势，一旦其他行业实现自动化了，它就会东山再起。通过机器低成本完成的事情越多，人类学家的贡献就越有价值。

相反，让人伤感的是，科学家的长远前景并不是最光明的。未来，唯一的科学家很有可能就是计算机科学家，即从事科学研究的计

算机科学家。之前被人称为“科学家”的人（像我一样）会将其毕生贡献给理解计算机所做出的科学进步。他们的幸福感不会比以前明显降低，毕竟科学对他们来说一直是一种业余爱好。对于有技术头脑的人来说很重要的一项工作会被留下来——留意计算机。实际上，这需要的不仅是工程师，基本上，这可能是所有人类的全职工作，即弄明白我们想从机器那里得到什么，并保证我们会得到这些东西——本章后部分会详细谈到。

同时，随着自动化与非自动化工作跨越经济领域，我们可能会看到失业率渐渐增长，越来越多的行业薪水下探，无法自动化的行业越来越少，但报酬却越来越高。当然，这种情况已经发生，但路还很长。过渡期会充满骚乱，但多亏了民主，它会有一个圆满的结局（紧握你的选票，它可能会成为你最有价值的东西）。当失业率上升超过50%时，甚至小于这一数字时，关于重新分布的态度会彻底改变。一批刚失业的大部分人会把选票投给慷慨的终身失业救济金，以及用于资助他们的高昂税收。这些做法并不会耗尽资源，因为机器会进行必要的生产。最终，一开始我们会讨论就业率而不是失业率，而降低失业率将被看作进步的标志（“美国正在倒退，我们的就业率仍然保持在23%”）。失业津贴将被发放给每个人的基本收入代替。不满足于基本收入的那些人会赚得更多，在所剩无几的人类职业中大赚一笔。自由党和保守党仍然会因为税率而争吵，但球门柱已经被永远移走了。随着劳动力总价值的骤减，最富裕的国家将是那些自然资源与人口比例最高的国家（现在移到加拿大了）。对于那些不工作的人，生活不会变得没有意义，最多就像在热带岛屿上，那里大自然的恩赐满足了所有需求，生活才变得没有意义。礼品经济将会发展起来，开源软件运动将是预告。人们在人际关系、自我实现、灵性中寻找意义，就和现在他们做的一样。谋生的需要将会变成遥远的记忆，这是我们克服的又一个人类的原始过去。

战争不属于人类

对服役自动化比对科学自动化要困难，但最终会实现的。机器人的主要用途之一就是完成那些对人类来说过于危险的任务，战争也一样危险。机器人已经可以拆掉炸弹，而无人机可以使一个团看清整座山。自驾供应卡车和机器骡（robotic mule）正在研发当中。很快我们就要决定是否允许机器人自己扣动扳机。因为这样做之所以存在争论，是因为虽然可以让人类远离伤害，但远程控制在移动迅速、不是你死就是我亡的情况中不太切实可行。反对的人认为，机器人没有道德标准，所以无法让它们决定某个生物的生死。但我们可以教会它们，更深层次的问题在于，我们是否准备好这么做了。

要重申诸如军事需要原则、相称原则、宽恕民众原则等之类的总则并不困难，但它们与具体行动之间存在鸿沟，士兵的判断就是要填补这道鸿沟。当机器人将阿西莫夫的机器人学三条定律运用到实践中时，很快就会产生麻烦，他的故事就深刻阐明了这一点。总则通常要么会有矛盾，要么会自相矛盾，以免将所有情况都判定为非黑即白。什么时候军事需要会比宽恕民众重要？没有统一的答案，也没有什么方法来对一台包含所有可能性的计算机进行编程。然而，机器学习提供了替代的方法。首先，教会机器人来识别相关概念，例如，可以利用各类情况的数据集，这些情况包括：民众得到与得不到宽恕，武装反应相称与不相称等。然后以涉及这些概念的规则的形式，赋予它行为准则。最终，让机器人学会如何通过观察人类来应用这些准则：士兵在这种情况下会开火，但在另外一种情况下不会开火。通过概括这些例子，机器人可以掌握一个端到端模型，比如可以以大型多逻辑节点的形式来做道德决策。一旦机器人的决策和某个人类的一致，而这个人的决策又往往与他人的一致，那么训练就完成了，意味着模型可以被下载，并用于数千个机器人的大脑。不像人类，机器人在激烈的

战斗中不会失去理智。如果机器人出现故障，制造商就得负责任；如果机器人打错电话，教它的人就得负责任。

这个方案的主要问题也许你已经猜到，就是让机器人通过观察人类来学习道德标准并不是一个好主意。当机器人看到人类的行为经常违背道德准则时，它就会变得非常困惑。我们可以清理训练数据，方法包含所有这些例子：道德学家小组一致同意，士兵做了正确的决定，而专家小组成员也可以在学习之后，对模型进行检查和微调以满足他们的要求。意见可能难以统一，但是，如果小组包含所有各种不同的人，它就该被统一。向机器人教授伦理道德，因为它们的逻辑思维没有负担，这会迫使我们检查我们的假设，然后对自己的矛盾行为进行分类。在这个领域里（其他许多领域），机器学习最大的好处也许不在于机器学习了什么，而在于通过教授这些机器，我们学会了什么。

另一个反对建立机器人军队的观点是，它们使战争变得太容易。但如果我们单方面放弃建立机器人大军，可能会引起下一场战争。《逻辑反应》（The Logical Response）由联合国和人权观察组织主张，是一个禁止机器人战争的协议，和1925年颁布的禁止生化战争的《日内瓦公约》类似。但是，这里忽略了一个重要的不同点。生化战争只会增加人类的痛苦，但机器人战争可以很大程度上减轻痛苦。如果战争是机器人参与战斗，人类只是指挥，就不会有人受伤或者死亡了。那么也许我们该做的，不是放逐机器人士兵，而是（当我们准备好时）放逐人类士兵。

机器人军队确实使战争发生的可能性变大，但它们也会改变战争的伦理学。如果目标是其他机器人，射击与不射击的困境会容易解决得多。现代观点认为战争恐怖得无法形容，人们在迫不得已时才会采取战争手段。这种观点会被一个稍有差别的观点替代，即认为战争是一种毁灭性的狂欢，会使所有参战方变得贫穷，所以最好避免战争，但也不能因为避免战争而付出所有代价。如果将战争还原为一场竞

赛，目的是看看谁的摧毁能力最强，那么为什么不比比谁创造的价值最多呢？

无论如何，禁止机器人战争也许不太可行。未来无论是大国还是小国，都会忙着研发（而完全不是禁止）遥控飞机（未来战争机器人的前身），因为它们估计这样做的好处大于风险。和所有武器一样，自己拥有机器人，比信任另一方认为不该有机器人更安全。如果在未来战争中，数百万架神风系列遥控飞机将会在几分钟之内摧毁传统的军队，它们最好是我们的遥控飞机。如果第三次世界大战会在数秒内结束，也就是一方控制另一方的系统，我们最好还是具备更加智能、更加快速、更加有复原力的网络（离网系统不是解决办法，因为虽然没有经过网络连接的系统不会被黑，但它们也无法和连接网络的系统相比较）。总而言之，如果机器人装备竞赛能加速《日内瓦第五公约》[5]禁止人类参与战争的进程，也许这也是一件好事。战争会一直陪伴我们，但战死则不一定是必然的。

谷歌＋终极算法=天网？

当然，机器人军队也会引起完全不一样的恐慌。根据好莱坞电影，人类的未来将会被庞大的人工智能及其大量的机器小兵扼杀（当然，除非有一位有胆量的英雄在电影最后5分钟挽回了局面）。谷歌已经拥有这样的人工智能所需要的庞大硬件，而且最近已经由一些机器人创业公司来运行了。如果我们将终极算法置入其服务器，那么人类会毁灭吗？当然会。是时候表明我的真正安排了，对于托尔金，我表示抱歉：

三大算法归属天下科学家，

七大算法归属服务器工程师，

九种算法属于阳寿可数的凡人，

还有一种属于高居御座的黑暗人工智能。

学习算法之大地黑影憧憧。

一种算法统领众戒，尽归罗网，

一种算法禁锢众戒，昏暗无光。

学习算法之地黑影憧憧。

哈哈！严肃地说，我们该不该担心机器人会接管世界？有种种不祥的预兆。一年接着一年，计算机所做的世界性工作并没有变多，它们做得更多的是决策。谁拿到信用卡，谁购买什么，谁得到什么工作和什么样的升职机会，哪只股票会上涨和下跌，保险花费多少，警官会在哪里巡逻，然后谁会被逮捕、刑期会有多长，谁约了谁、又生了谁——通过机器掌握的模型已经在所有这些事情中起作用。关掉我们的所有计算机而不会引起现代文明崩溃的时刻早已过去。机器学习是最后一根稻草，如果可以开始进行自我编程，那么所有控制它们的希望都肯定会落空。著名科学家如史蒂芬·霍金已经呼吁趁早研究这个问题。

放轻松，备有终极算法的人工智能接管世界的概率是零。原因很简单：不像人类，计算机本身并没有自己的意志。它们是工程师生产的产品，而不是进化体。即使无限强大的计算机，也仅仅是我们意志的延伸，没什么可怕的。回忆一下每种学习算法的三个组成部分：表示方法、评估、优化。学习算法的表示方法限制了它能学习的内容。让我们把它想象成很强大的学习算法，比如马尔科夫逻辑，那么该学习算法原则上可以学习任何东西。接着最优化器会在其权力范围内，做所有工作来将评估功能最大化——不多也不少——而评估功能“是由我们决定的”。一台更强大的计算机只会把它优化得更好。掌握的

系统如果过去不按照我们想要的来做事，那么它就会严重不适合，因此就会消失。实际上，那些一代接一代、能稍微更好地服务我们的系统会呈现多样化并接管基因库。当然，如果我们愚蠢到故意对计算机进行编程，让其凌驾于我们之上，那么我们就该被统治。

同样的推理也可用于所有人工智能系统，因为它们都（明确或者含蓄地）具备同样的三个组成部分。它们可以多样化它们做事的内容，甚至想出惊人的计划，但仅仅服务于我们设定给它们的目标。如果一个机器人被设定的目标是“做一顿大餐”，它可能会决定煎一份牛排、煮一份法式海鲜浓汤，或者做一道独创的新菜式，但它无法决定谋杀它的主人，就像汽车无法确定飞走那样。人工智能系统的目标就是要解决NP完全问题（你可能会联想到第二章），解决这个问题可能需要指数级的时间，但解决方法总可以得到有效验证。因此我们应该张开怀抱欢迎比我们大脑强大很多的计算机，因为我们的工作要比它们的简单很多，知道这一点就安心了。它们得解决问题，我们只要保证它们做的事能让我们满意。对于一些东西，我们思考得很慢，但人工智能却很快，而这个世界对它们来说也将是更好的世界。我（作为人类中的一个）欢迎我们的新型机器人下属。

有人担心智能机器会掌握控制权，这很自然，因为我们知道的唯一智能实体就是人类和其他动物，而它们明确具备自己的意志。但没有必要将智能意志和独立意志联系起来，更确切地说，智能也许并不适应同样的主体，只要它们之间存在统治。在《延伸的表现型》一书中，理查德·道金斯表明大自然中充满了动物基因不仅控制其身体的例子，从布谷鸟的蛋到海狸水坝。技术就是人类的延伸表现型。这意味着我们可以继续控制它，即使它会变得复杂到让我们难以理解。

假设两条20亿年前的DNA在它们的“私人泳池”（也称为细菌的细胞质）中“游泳”。它们正在思考一项重大决策。“我在担心，戴安娜，”其中的一条说，“如果我们开始制造多细胞生物，它们会统治世界吗？”现在快进到21世纪，此时DNA仍然活着而且很健康。实际

上，比以往更好的是，越来越大比例的DNA安全地存在两足有机体中，组成数以亿计的细胞。自从我们的双链“朋友”做出它们的重大决定，这就变得很麻烦了。人类就是它们创造的最难以对付的生物，我们发明出诸如避孕之类的东西来让自己寻找乐趣，而又不用传播DNA，且我们具备（或者说似乎具备）自由的意志。但还是DNA形成了我们对于乐趣的概念，而我们利用自己的自由意志来追求快乐和避免痛苦，这很大程度上仍然与对我们DNA生存最有利的东西相符。如果我们选择将自己变成硅，那么我们也许是DNA灭亡的产物，但即使那样，这也是一段伟大的20亿年历史。当今我们面对的决定也相似：如果我们开始制造人工智能——庞大的、相互连接的、超人类的、难以理解的人工智能——那么它们会统治世界吗？最多就像多细胞有机体取代基因，对于它们来说，我们可能也是庞大的、难以理解的。人工智能是我们的幸存机器，和我们是我们基因的幸存机器同一个道理。

然而，这并不意味着没有什么可以担心了。最大的忧虑是，和所有技术一样，人工智能可能会落入不法之徒手里。如果罪犯或者搞恶作剧的人对人工智能进行编程，用于统治世界，我们最好有人工智能警察局来抓住他们，并消灭他们以防其逍遥法外。为了避免庞大的人工智能变得疯狂，最佳的保险措施就是有更庞大的人工智能来维护和平。

第二个忧虑就是人类会自愿屈服于统治。它始于机器人的权力，对我但并不是所有人来说有点荒谬。毕竟，我们已经把权力赋予了动物，但动物却从不要求有权力。在扩张“共鸣圈”中，其下一步就是机器人权力应该算是合情合理。对机器人表示同情并不难，尤其设计它们就是为了唤起同情。即使是日本的“虚拟宠物”，只有三个按键、一个LCD（液晶显示器）屏幕，也能成功唤起同情。第一个像人的消费者机器人会引发竞赛：制造越来越多唤起同情心的机器人，因为它们卖得比一般的金属机器人要好得多。由机器人奶奶带大的孩子会终身容易被友好的电子朋友感动。“恐怖谷理论”——对于那些接近人类但又不是人类的机器人，我们会有不适感——对他们来说将会无

法理解，因为他们带着机器人的习性长大，甚至可能会变成酷酷的青少年。

在隐伏的人工智能统治的进程中，下一步就是让它们做所有决策，因为它们是如此智能。当心！它们可能会更加智能，但它们服务于任何设计出它们计分函数的人。这就是“绿野仙踪”问题。在智能机器人的世界中，你的工作就是要确保它们做了你想做的事，无论是在输入（设定目标）时，还是在输出（检查你得到了自己要求得到的东西）时。如果你不这样做，会有其他人做。机器可以帮助我们弄明白总体来说我们想要什么，但如果你不参与，就会输掉——就像民主一样，只有参与你才不会输。和当今我们相信的相反，人类太容易遵守别人的规则，而任何足够先进的人工智能都无法区别于上帝。人类不会介意听取来自某台玄妙深奥的计算机的前进命令，问题在于谁来监督监督者。人工智能通往的是更完善的民主，还是更潜伏的专政？永恒的监视才刚刚开始。

第三个忧虑（可能也是最大的）就是像众所周知的魔仆一样，机器会给我们要求的而不是想要的东西。这并不是一种假设性情景分析，学习算法会一直做这件事。我们训练神经网络来识别马匹，但它掌握的却是褐块，因为所有的马匹机器训练集刚好是褐色的。你只是买一块手表，所以亚马逊推荐了类似的商品——其他手表，此时手表变成你最不想买的东西。如果你检查计算机今天所做的所有决定，比如谁拿到信用卡，就会发现这些决定常常不必要那么糟糕。如果你的大脑是一台支持向量机，且你的所有信用评估知识从阅读一个糟糕的数据库中得来，你的决定也会变成那样。人们担心计算机会变得过于智能而统治世界，但真正的问题是，它们也很愚蠢但已经统治世界。

进化的第二部分

即使当今计算机还不是非常智能，但无疑它们的智力却在快速提升。早在1965年，I·J·古德英国（一位统计学家以及阿兰·图灵在第二次世界大战中密码破解项目上的伙伴）就推测到即将到来的“智能爆炸”。古德指出，如果我们可以设计出比自己还要智能的机器，反过来，它们也可以设计出比它们更加智能的机器，就这样无休止继续下去，让人类智能落在后面。在1993年的一篇文章中，弗诺·文奇将其命名为“奇点”。这个概念经过雷·库兹韦尔得到最大推广，他在《奇点临近》中指出，不仅仅奇点不可避免，而且机器智能超越人类智能的时刻（让我们称之为“反观点”）也将会在未来几十年到达。

显然，如果没有机器学习——设计程序的程序——奇点将无法发生。我们也会需要足够强大的硬件，但那会很顺利地达到。我们发明出终极算法之后很快就会达到“反观点”（我愿意和库兹韦尔赌一瓶唐·培里侬香槟王，觉得这会在我们对大脑进行逆向工程前就会发生，逆向工程是他选择的用于到达人类水平智能的方法）。

“奇点”这个术语源于数学，表示此处有一个使函数变成无穷大函数的点。例如，当x为0时，1/x这个函数就会有奇点，因为1除以0的得数是无穷大。在物理学中，奇点的典型例子就是黑洞：密度为无限大的点，此处有限数量的物质被塞入无穷小的空间中。奇点的唯一问题就在于它们并不是真正存在的（你上次给0个人切蛋糕，每个人都会拿到无数块蛋糕是什么时候）。在物理学中，如果某个理论预测某物无穷大，那么这个理论就会出问题。举个恰当的例子，广义相对论也许会预测，黑洞有无穷大的密度，因为它忽略了量子效应。同样，智能无法永远都在提高。库兹韦尔承认了这一点，但他指的是技术改进（处理器速度、记忆容量等）中的一系列指数曲线，并提出这种增长的界限是如此遥远，我们不必纠缠于这些界限。

库兹韦尔过拟合了。他准确地批评别人，因为他们总是进行线性推断——看到直线而不是曲线——但也深受更为奇异的疾病之害：指

数随处可见。在曲线中会有平直部分——什么也没发生——他看到还没有飙升的指数，它们是S形曲线（我们在第四章的好朋友）。S形曲线的开始部分会容易被误以为是一个指数，但它们会很快偏离。多数库兹韦尔曲线是摩尔定律的结果，处于快要结束的状态。库兹韦尔认为，其他技术会代替半导体，而S形曲线会堆在S形曲线上，每条曲线都比前一条要陡，但这是猜测。他进一步提出，地球上生命的整个历史，不仅仅是人类技术表现出呈指数加速发展，但产生这种感觉至少部分原因在于视差的影响——离得比较近的东西似乎移动得更快。处于寒武纪爆发时的酷热中的三叶虫，因为相信指数加速发展而得到宽恕，不过有一个很大的减速。暴龙雷（Tyranno saurus Pay）也许会提出一个加速扩大体型。真核细胞（我们）进化的速度要比原核生物（细菌）要慢。进化的过程是时进时退，不可能总是顺利加速。

为了回避这样一个问题：无限密集的点不存在，库兹韦尔提出将奇点与黑洞的视界等同起来，黑洞视界的重力如此强大，光线也无法逃脱。同样，他说，奇点就是这样一个点，在这个点之外技术演化进行得如此快速，人们都无法预测或者了解将要发生什么。如果那就是奇点，那么我们已经在它里面。我们无法提前预测一种学习算法会想出什么主意，甚至我们在回顾过去时常常无法理解它。实际上，我们一直生活在这样一个只能理解某些部分的世界中。主要的差别在于，我们的世界现在部分创造了我们，这肯定是改善之处。“反观点”之外的世界对我们来说会不可思议，就像更新世一样。我们将会把注意力集中在自己能理解的事物上，正如我们一直做的，并将剩下的称为“随机的”（或者“神圣的”）。

我们所处的轨道不是奇点，而是相变。它的临界点——反观点——当机器学习赶上自然多样化时，反观点就会到来。自然学习法本身已经经历了三个阶段：进化、大脑、文化。每个阶段都是前一个阶段的产物，而且每个阶段都会学得更快。机器学习逻辑上是该进程的下一阶段。计算机程序是世界上最快速的复制者：复制它们只需要不到一秒，但创造它们却比较缓慢（如果这件事由人类完成）。机器学

习克服了瓶颈期，留下最后一个：人类可接受改变的速度。这个到最后也会被克服，但并不是因为我们决定将东西移交给我们的“智能后代”，正如汉斯·莫拉维克所称呼的那样，然后温柔地走进美好的夜晚。人类并不是生命之树上垂死的枝丫，相反，我们开始出现分支。

文化和更大的大脑共同进化，与此类似，我们会与自己的创造物一起进展。我们一直具备——如果我们没有发明火或者长矛，那么人类从外表上看会有差别。我们是“技术人”，和智人差不多。但在第九章中我设想的那种细胞模型也会考虑全新的东西——在你所提供参数的基础上设计细胞的计算机，以及以同样方式在功能规格的基础上设计微芯片的硅编译器。对应的DNA可以合成并插入一个“通用”细胞，将其转化成想要的那个。克雷格·文特尔是基因组的先锋，已经在这个方向迈出第一步。首先，我们会利用这股力量来抗击疾病：新的病原体已经确定，治疗方法立刻就被发现，你的免疫系统从网上将它下载下来。健康问题变成矛盾修饰法。然后DNA设计会让人们最后拥有他们想要的身体，迎来买得起的美丽时代，威廉·吉布森说了上述让人记忆深刻的话。然后接着“技术人”会进化成无数不同的智能物种，每种物种都有其生态位，这是一个全新的与当今不一样的生物圈，正如当今的生物圈与原始海洋存在差异一样。

许多人担心以人为导向的进化会永久性地将人类分为基因拥有者的一个类别，以及基因缺失者的一个类别。这次想象的异常失败打击了我。自然进化最后不会只有两种物种，一种物种顺从另一种物种。这就意味着下一个瓶颈就是界限，虽然现在我们还没看到它。其他过渡期会到来，有些长，有些短，有些快，有些不会持续很长，但在地球的生命历程中，接下来的几千年很有可能是最为令人惊叹的一段历程。

---

[5] 《日内瓦公约》是1864—1949年在瑞士日内瓦缔结的关于保护平民和战争受难者的一系列国际公约的总称，含四个公约。这里提到的《日内瓦第五公约》是作者对未来的一种设想。——编者注

后记

现在你已了解了机器学习的秘密。将数据变为知识的机器不再是一个黑匣子：你知道魔法是如何发生的，以及它能做什么、不能做什么。你已经遇到复杂性怪兽、过拟合难题、维数灾难、探索与开发困境。你大体上知道了谷歌、脸书、亚马逊和所有其他网站把你每天慷慨提供给它们的数据用来做了什么，它们为什么能帮你找到东西、过滤垃圾，且不断改善它们的服务。你已经看到，在世界机器学习研究实验室里正酝酿什么，你可以旁观他们正在创造的未来。你已经看到机器学习的五大学派以及它们的主算法：符号学派和逆向演绎，联结学派和逆向传播，进化学派和遗传算法，贝叶斯学派和概率推理，类推学派和支持向量机。因为你已经遍历广阔的区域，协调跨越边境，爬到顶峰，和很多机器学习算法相比，你能更好地欣赏风景，而那些学习算法只能在其领域中每日艰苦工作。你可以看到共同主题流淌在这片土地上，就像一条地下河流，并且你还明白，这五种学习算法，表面上看差别很大，其实也只是单一通用学习算法的五个方面。

旅程还远远没有结束。我们还没有终极算法，只是瞥到它可能长什么样。如果某些基本的东西还找不到，有些东西沉浸在其历史当中，而我们在本领域中无法看到，那会怎么样呢？我们需要一些与之前想法不一样的新想法。这就是我写本书的原因——让你开始思考。我在华盛顿大学关于机器学习的夜校教课。2007年，网飞大奖宣布后不久，我提议将其作为班级项目中的一个。我班上的一位学生——杰夫·霍伯特被它迷住了，并在课程结束时继续钻研这个项目。在他第一次了解机器学习的两年之后，他最终成为获胜组的成员，当时总共有两个获胜组。现在轮到你了。你可以从UCI数据库上下载一些数据集

（archive.ics.uci.edu/ml/）并开始这场比赛。当你做好准备时，可以对Kaggle.com进行了解，这是一个专门组织管理机器学习比赛的网站，然后挑一两个链接并点击进入。当然，如果你招募一两个朋友来和你一起工作，那样会更好玩。如果你也着迷了，就像杰夫那样，最后变成一个专业的数据科学家，那么欢迎进入世界上最让人陶醉的领域。如果你发现自己不满意于当前的学习算法，那就发明新的算法——或者只是出于好玩而发明。我最殷切的希望就是，你对这本书的反应，就像我对读的第一本人工智能书的反应一样，这已经过去20多年：这个领域有太多的事情要做，我不知道从何开始。如果有一天你发明了终极算法，请不要带着它跑到专利局，而是开放资源。终极算法应被任何人或者组织拥有，这一点太重要了。它应用的速度会比你为它申请许可的速度要快。但如果你打算创业，记得让每个世界上的男人、女人、孩子都能享受它。

无论是出于好奇，还是专业兴趣，你读了这本书，我希望和你的朋友、同事分享你学到了什么。机器学习接触到我们每个人的生活，而我们想用它来做什么也由自己决定。带着你对机器学习的新了解，你现在处于更好的位置来思考诸如隐私、数据分享、工作的未来、机器人之间的战争、人工智能的承诺与危险之类的问题；而且了解到这一点的人越多，我们越有可能避免圈套，并找到正确的路。这也是我写本书另外一个主要原因。统计学家知道做预测不容易，尤其是对未来的预测，而计算机科学家知道预测未来的最佳方法就是创造未来，但未经检验的未来不值得创造。

感谢你让我做你的向导。我想送给你一份临别礼物。牛顿说过，他就像一个在沙滩上玩耍的男孩，这边捡一枚鹅卵石，那边捡一块贝壳，而真理的大海就在他面前，等着他去发现。300年后，我们已经收集了一些了不得的鹅卵石和贝壳，但大片未被发现的海洋仍然延伸至远处，闪烁着希望的光辉。我的礼物就是一艘船——机器学习。现在该是时候扬帆起航了！

致谢

首先，我要感谢在科学探险历程中的同伴：我的学生、合作人、同事以及在机器学习领域的每个人。这是你的书，也是我的书。我希望你能原谅我过于简化以及删改的处理，以及本书有些部分稀奇古怪的写作方式。

我对在不同阶段阅读并评价本书初稿的所有人表示感谢，包括迈克·贝尔菲奥尼、蒂亚戈·多明戈斯、奥伦·埃齐奥尼、亚伯·弗里森、罗勃·金斯、阿龙·哈勒维、大卫·伊斯雷尔·考茨、克洛伊·基顿、加里·马库斯、雷·穆尼、凯文·墨菲、弗兰齐·勒斯纳、本·塔斯卡。同时也感谢那些给予我建议、信息，或者各种帮助的人，包括汤姆·格里菲斯、大卫·赫克曼、汉娜·希基、艾伯特–拉斯洛·巴拉巴希、燕乐存、芭芭拉·琼斯、迈克·摩根、彼得·诺尔维格、朱迪亚·珀尔、格雷戈里·皮亚特斯基–萨比罗、塞巴斯蒂安·承。

我很幸运能在一个非常独特的地方工作，也就是华盛顿大学的计算机科学与工程系。我还要感谢乔什·特伦鲍姆，还有他团队里的每个人，因为他们在麻省理工学院组织轮休假，在此期间我才开始了本书的写作。多亏了吉姆·莱文，我不知疲倦的代理人，因为他“掉进陷阱”（他是这么说的）并把消息传出去。还要感谢拉文·格林伯格·罗斯坦的每个人。谢谢K·J·凯莱赫，我能干的编辑，帮助我把这本书变得更好（一章又一章、一行又一行）。我要感谢Bosic Books的每个人。

我还要感谢那些过去这么多年资助过我的研究组织，包括阿波罗斯塔尔公司（ARO）、DAPRA、FCT、美国科学基金会、福特、谷歌、IBM、柯达、雅虎、斯隆基金会。

最后，我要感谢家人的爱与支持。

延伸阅读

为了让广大读者更深入地探索机器学习世界，我们特此附上延伸阅读部分。此部分包含大量文献，对读者深入学习大有裨益。同时，为了保证准确性，我们保持了此部分中涉及的人名、文献、出版信息等内容的原貌，以飨读者。希望读者朋友能通过本书和这些文献进入丰富多彩的机器学习世界。

如果本书激发了你对机器学习及其相关问题的兴趣，那么在本部分你会找到许多建议。本书的目的不是面面俱到，而是为了引导人们了解与机器学习相关的知识（正如Borges说的那样）。我尽量为读者选择合适的书籍和文章。专业类出版物的阅读至少需要一些计算、统计或者数学领域的背景知识，我们会以星号（*）来标记这些出版物。即使这些是专业出版物，对读者来说也可以接受很大一部分。我没有把卷号、期刊号、页码列出来，因为网站使这些变得多余，对于出版商的地址来说也同样如此。

如果你想从整体上对机器学习了解更多，网络课程就是开始学习的好选择。并非巧合，与本书内容最为相近的就是我教的这门课（www.coursera.org/course/machinelearning）。也可以参考安德鲁·恩格的课程（www.coursera.org/course/ml）和亚瑟·阿布·穆斯塔法的课程（http://work.caltech.edu/telecourse.html）。接下来就要阅读教材。和本书最相近且最容易接受的一本教材就是Tom Mitchell的*Machine Learning**（McGraw-Hill, 1997）。更现代且更精确的教材包括Kevin Murphy的*Machine Learning: A Probabilistic Perspective*（麻省理工出版社，2012），Chris Bishop的 *Pattern Recognition and Machine Learning**（Springer, 2006），以及*An*

Introduction to Statistical Learning with Application in R,*（作者是Gareth James、Daniela Witten、Trevor Hastie、Rob Tibshirani，Springer, 2013)。我的文章“A few useful things to know about *machine learning*”（Communications of the ACM, 2012）总结了一些关于机器学习的传统知识，这些知识在教材中往往较为隐晦，且可作为开始阅读本书之前的参考资料。如果你懂得如何编程，并想尝试机器学习，可以从众多开源软件包开始，例如，Weka（www.cs.waikato.ac.nz/ml/weka）。有两本重要的机器学习杂志：《*Machine Learning*》和《*Journal of Machine Learning Research*》。每年举办的机器学习的重要会议包括机器学习国际会议（International Conference on Machine Learning）、国际神经信息处理大会（Conference on Neural Information Processing Systems）、国际学术和技术开发研讨会（International Conference on Knowledge Discovery and Data Mining）。在http://videolectures.net上有众多关于机器学习的访谈。网站www.KDnuggets.com是机器学习的一站式服务店，你可以注册账号，获得实时通信，了解最新发展动态。

序

早前列举的机器学习对日常生活的影响可在George John（*SIGKDD Explorations*, 1999）所著的“Behind－the－scenes data mining”中找到，序言部分“日常生活”段落的灵感也源于此。Eric Siegel的书*Predictive Analytics*（Wiley, 2013）探索了众多机器学习的应用。麦肯锡全球研究所2011年的报告*Big Data: The Next Frontier for Innovation, Competition, and Productivity*对“大数据”这一术语进行了推广。在该报告中，Viktor Mayer－Schönberger和Kenneth Cukier所著的*Big Data: A Revolution That Will Change How We Live, Work, and Think*（Houghton Mifflin Harcourt,

2013）一文，讨论了许多由大数据引发的问题。我了解人工智能的教材是Elaine Rich所著的*Artificial Intelligence*（McGraw－Hill, 1983）。现在通用的版本是Stuart Russel和Peter Norvig所著的*Artificial Intelligence: A Modern Approac*（3rd., Prentice Hall, 2010）。Nils Nilsson的*The Quest for Artificial Intelligence*（Cambridge University Press, 2010）介绍了早期的人工智能。

第一章

John MacCormick写的*Nine Algorithms That Changed the Future*（Princeton University Press, 2012）一书介绍了在计算机科学中一些最重要的算法，其中有一章是关于机器学习的。Sanjoy Dasgupta、Christos Papadimitriou和Umesh Vazirani写的*Algotithms*（McGraw－Hill, 2008）一书是介绍该主题的简明教材。Danny Hillis写的*The Pattern on the Stone*（Basic Books, 1998）解释了计算机如何运转。Walter Isaacson在*The Innovators*（Simon & Schuster, 2014）一书中生动描述了计算机科学的历史。

SumitGulwani、William Harris和Rishabh Singh写的“Spreadsheet data manipulation using examples”（*Communications of the ACM*, 2012）一文就是一个例子，体现了计算机如何通过观察用户来进行自我编程。Tom Davenport和Jeanne Harris的*Competing on Analytics*（HBS Press, 2007）一书介绍了在商务领域如何使用预测分析。Steven Levy在*In the Plex*（Simon & Schuster, 2011）一书中高水平地介绍了谷歌技术如何起作用。Carl Shapiro和Hal Varian在*Information Rules*（HBS Press, 1999）一书中解释了网络的影响。Chris Anderso在The Long Tail（Hyperion, 2016）一书中介绍了长尾现象。

由Tony Hey、Stewart Tansley、Kristin Tolle主编的*The Fourth Paradigm*（Microsoft Research, 2009）一书探索了通过数据密集型运算来实现科学转型。James Evans和Andrey Rzhetsky在“*Machine Science*”（*Science*, 2010）一文中讨论了一些计算机做出科学发现的不同方法。Pat Langley等人写的*Scientific Discovery: Computational Explorations of the Creative Processes**（MIT Press, 1987）一书介绍了自动发现科学规律的一系列方法。Usama Fayyad、George Djorgovski和Nicholas Weir在“From Digitized Images to Online Catalogs”（*AI Magazine*，1996）一文中介绍了SKICAT项目。Niki Wale的Machine Learning in Drug Discovery and Development（*Drug Development Research*, 2001）一文正好对这个问题进行了概述。Ross King等人写的The Automation of Science（*Science*, 2009）一文介绍了机器人科学家Adam。

SashaIssenberg的*The Victory Lab*（Broadway Books, 2012）一书仔细分析了在政治领域中数据分析的用途，他的How President Obama’s Campaign Used Big Data to Rally Individual Votes（*MIT Technology Review*，2013）一文告诉读者迄今为止数据分析的最大功绩。Nate Silver的*The Signal and the Noise*（Penguin Press, 2012）一书中的一章介绍了他集中民意的方法。

P. W. Singer的*Wired for War*（Penguin, 2009）一书主要介绍了机器人战争。Richard Clarke和Robert Knake的Cyber War（Ecco, 2012）一书敲响了互联网战争的警钟。我将机器学习与博弈论结合打败对手，这个观点开始是以课题的形式呈现，后来在Nilesh Dalvi等人的Adversarial Classification*（*Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining*, 2004）上进行了介绍。Walter Perry等人的*Predictive Policing*（Rand, 2013）引导人们如何在警察工作中使用解析学。

第二章

Laurie von Melchner、Sarah Pallas和Mriganka Sur的Visual Behaviour Mediated by Retinal Projections Directed to the Auditory Pathway（*Nature*, 2000）一文介绍了雪貂大脑重新布线的实验。Joanna Moorhead的*Seeing with Sound*（*Guardian*, 2007）和www.benunderwood.com这个网站介绍了Ben Underwood的故事。Otto Creutzfeldt在Generality of the Functional Structure of the Neocortex（*Naturwissenschaften*, 1977）一文中表明皮质是一种算法。Vernon Mountcastle在“An Organizing Principle for Cerebral Function: The Unit Model and the Distributed System”一文中也做了同样的阐述，这篇文章收录在Gerald Edelman和Vernon Mountcastle编辑的*The Mindful Brain*（MIT Press, 1978）一书中。Gary Marcus、Adam Marblestone和Tom Dean在The Atoms of Neural Computation（*Science*, 2014）中提出反对意见。

Alon Halevy、Peter Norvig和Fernando Pereira在The Unreasonable Effectiveness of Data（*IEEE Intelligent Systems*, 2009）中赞成机器学习是新的发现范式。Benoit Mandelbrot在同名书（Freeman, 1982）中探索大自然中的分形几何。James Gleick的*Chaos*（Viking, 1987）一书讨论并描述了曼德布洛特集合。Edward Frenkel的*Love and Math*（Basic Books, 2014）一书介绍了朗兰兹纲领，这是一个尝试将数学的不同子域联合起来的研究。Lance Fortnow的*The Golden Ticket*（Princeton University Press,2013）一书介绍的是NP完全问题和P=NP问题。Charles Petzold的*The Annotated Turing*（Wiley，2008）通过重新研究图灵最初关于图灵机器的论文来对图灵机器进行解释。

Douglas Lenat等人著的“Cyc: Toward programs with common sense”*（*Communication of the ACM*, 1990）一文就介绍了Cyc项目。Peter Norvi在“On Chomsky and the Two Cultures of

Statistical Learning”（http://norvig.com/chomsky.html）一文中讨论Noam Chomsky对于数字学习的批评。Jerry Fodor的*The Modularity of Mind*（MIT Press, 1983）一书总结了他关于大脑如何运转的观点。Leon Wieseltier的“What Big Data will Never Explain”（*New Republic*, 2013）和Andrew McAfee的“Pundits, Stop Sounding Ignorant about Data”（*Harvard Business Review*,2013）中揭示了围绕大数据能做和不能做之间的矛盾。Daniel Kahneman在*Thinking, Fast and Slow*（Farrar, Straus and Giroux, 2011）一书的第21章解释了算法为什么可以打败感官直觉。David Patterson在“Computer Scientists May Have What it Takes to Help Cure Cancer”（*New York Times*, 2011）一文中证明了计算与数据在与癌症的斗争中所起的作用。

更多学派关于如何获得主算法的观点在以下相应部分中会提到。

第三章

休谟的归纳法经典公式出现在*A Treatise of Human Nature*（1739）的I卷中。因为“The Lack of a Priori distinctions between Learning Algorithms”（*Neural Computation*, 1996）一文中的归纳推理，David Wolpert得出了“天下没有免费的午餐”定理。我在“Toward Knowledge-Rich Data MJining”（*Data Mining and Knowledge Discovery*, 2007）一文中讨论先验知识在机器学习中的重要性，并在“The role of Occam's razor in knowledge discovery”（*Data Mining and Knowledge Discovery*, 1999）一文中讨论了对奥卡姆剃刀理论的误读。在Nate Silver的*The Signal and the Noise*（Penguin Press, 2012）一书中，过拟合是其重要主题之一，他称之为“你绝对没有听过的最重要的科学问题”。John Ioannidis的“Why most published research

findings are false”*（*PLoS Medicine*, 2005）一书讨论了在科学领域中，将巧合发现误以为是真发现的问题。在“Controlling the false rate: A practical and powerful approach to multiple testing”*（*Journal of the Royal Statistical Society*, Series B, 1995）一文中，Yoav Benjamini和Yosef Hochberg提出了反对该观点的方法。在Stuart Geman、Elie Bienenstock和René Doursat的“Neural networks and the bias/variance dilemma”（*Neural Computation*, 1992）一文中展示了方差分解的过程。Pat Langley的“Machine learning as an experimental science”（*Machine Learning*, 1998）一文讨论了机器学习中实验法的作用。

William Stanley Jevons在*The Principles of Science*（1874）一书中首次提出了将归纳法视作演绎法的反向。Steve Muggleton和Wray Buntine的“Machine learning of first-order predicates by inverting resolution”*（*Proceedings of the Fifth International Conference on Machine Learning*, 1988）一文首先提出在机器学习中使用逆演绎。Saso Dzeroski和Nana Lavrac主编的书*Relational Data Mining*（Springer, 2001）介绍了归纳逻辑编程这个领域，这个领域研究的是逆演绎。Peter Clark和Tim Niblett在“The CN2 Induction Algorithm”*（*Machine Learning*, 1989）一文中总结了主要的米夏莱克样式规则归纳算法。Rakesh Agrawal和Ramakrishnan Srikant的“Fast algorithms for mining association rules”*（*Proceedings of the Twentieth International Conference on Very Large Databases*, 1994）一文介绍了零售商使用的挖掘规则的方法。Ashwin Srinivasan、Ross King、Stephen Muggleton和Michael Sternberg的“Carcinogenesis predictions using inductive logic programming”（*Intelligent Data Analysis in Medicine and Pharmacology*, 1997）一文介绍了规则归纳法用于预测癌症的例子。

J. Ross Quinlan的*C4.5: Programs for Machine Learning*（Morgan Kaufmann, 1992），以及Leo Breiman、Jerome Friedman、Richard Olshen和Charles Stone的*Classification and Regression Trees*（Chapman and Hall, 1984）两本书提出了两大主要的决策树学习算法。Jamie Shotton等人的“Real-time human pose recognition in parts from single depth images”（*Communications of the ACM*, 2013）一文解释了微软的Kinect如何利用决策树来追踪玩家的动作。Andrew Martin等人写的“Competing approaches to predicting Supreme Court decision making”（*Perspective on Politics*, 2004）一文介绍了决策树在预测最高法院投票中如何打败法律专家，并展示了大法官桑德拉·戴·奥康纳的决策树。

Allen Newell和Herbert Simon在“Computer science as empirical enquiry: Symbols and search”（*Communications of the ACM*, 1976）一文中提出假设，认为所有智能都是符号操纵。David Marr在*Vision*（Freeman, 1982）一书中提出信息处理的三个层面。Ryszard Michalski、Jaime Carbonell和Tom Mitchell主编的*Machine Learning: An Artificial Intelligence Approach*（Tioga, 1983）一书简单介绍了机器学习中象征主义研究的早期成果。Paul Smolensky的“Connectionist AI, symbolic AI, and the brain”（*Artificial Intelligence Revie*, 1987）一文对符号论模型提出了联结主义的观点。

第四章

Sebastian Seung的*Connectome*（Houghton Mifflin Harcourt, 2012）一书对神经科学、神经联结组学、对大脑进行逆向工程的严峻挑战做了浅显易懂的介绍。David Rumelhart、James McClelland主编

的*Parallel Distributed Processing*（MIT Press, 1986）一书是20世纪80年代处于繁盛时期的联结主义的“圣经”。James Anderson和Edward Rosenfeld编辑的*Neurocomputing*（MIT Press, 1988）一书整理了许多篇经典的联结主义论文，包括：McCulloch和Pitts的第一批神经模型；Hebb的Hebb规则；Rosenblatt的感知机；Hopfield的Hopfield网络；Ackley、Hinton和Sejnowski的玻尔兹曼机；Sejnowski和Rosenberg的NETtalk；Rumelhart、Hintonation和Williams的反向传播。Yann LeCun、Léon Bottou、Genevieve Orr和Klaus-Robert Müller的“Efficient backdrop”*一文由Genevieve Orr和Klaus-Robert Müller收录到*Neural Networks: Tricks of the Trade*（Springer, 1998）一书中，提出了一些使反向传播起作用的重要技巧。

Robert Trippi和Efraim Turban编辑的*Neural Networks in Finance and Investing*（McGraw-Hill, 1992）收录了关于神经网络在金融领域应用的文章。Todd Jochem和Dean Pomerleau的“Life in the fast lane: The evolution of an adaptive vehicle control system”（*AI Magazine*, 1996）一文介绍了ALVINN无人驾驶汽车工程。Paul Werbos的博士论文是*Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Science*（Harvard University, 1974）。Arthur Bryson和Yu-Chi Ho在*Applied Optimal Control*（Blaisdell, 1969）一书中介绍了他们反向传播的早期版本。

Yoshua Bengio的*Learning Deep Architectures for AI*（Now, 2009）一书简要介绍了深度学习。Yoshua Bengio、Patrice Simard和Paolo Frasconi的“Learning long-term dependencies with gradient descent is difficult”（*IEEE Transactions on Neural Networks,* 1994）一文介绍了反向传播中的信号误差扩散问题。John Markoff的“How many computers to identify a cat?16,000”（New York Times, 2012）一书对谷歌的脑计划及其结果进行汇报。Yann

LeCun、Léon Bottou、Yoshua Bengio和Patrick Haffner的“Gradient-based learning applied to document recognition”*（*Proceedings of the IEEE, 1998*）一书介绍了卷积神经网络、当前深度学习冠军。Jonathon Keats的“The $1.3B quest to build a supercomputer replica of a human brain”（Wired, 2013）介绍了欧盟的大脑模仿项目。Thomas Insel、Story Landis和Francis Collins的“The NIH BRAIN Initiative”（*Science*, 2013）一文介绍了BRAIN计划。

Steven Pinker在*How the Mind Works*（Norton, 1997）一书的第二章中总结了符号学者对联结主义模型的批判。Seymour Paper在“*One AI or Many?*”（*Daedalus*, 1988）一文中争论性地发表了自己的看法。Gary Marcus的*The Birth of the Mind*（Basic Books, 2004）一书中解释了进化如何促成人类大脑复杂的功能。

第五章

Josh Bongard的“Evolutionary robotics”（*Communication of the ACM*, 2013）一文讨论了Hod Lipson以及其他人关于进化机器人的作品。Steven Levy的*Artificial Lif*（Vintage，1993）一书也游览了一遍数字化这个“动物园”，从虚拟世界中计算机创造的动物到遗传算法。Mitch Waldrop的*Complexity*（Touchstone, 1992）一书第五章讲述了John Holland的故事，以及遗传算法研究前几十年的情况。David Goldberg的*Genetic Algorithms in Search, Optimization, and Machine Learning**（Addison-Wesley, 1989）一书对遗传算法进行了标准介绍。

在T. J. M.Schopf主编的*Models in Paleobiology*（Freeman, 1972）一书中，Niles Eldredge和Stephen Jay Gould发表了“Puntuatedequilibria: An alternative to phyletic

gradualism”一文，这篇文章中提出了他们的间断平衡理论。Richard Dawkins在*The Blind Watchmaker*（Norton, 1986）一书的第九章批评了该观点。在Richard Sutton和Andrew Barto的*Reinforcement Learning*（MIT Press, 1998）一书中，第二章讨论了探索—利用困境。在*Adaptation in Natural and Artificial System*（University of Michigan Press, 1975）中，John Holland提出了自己的解决方法及其他观点。

John Koza的*Genetic Programming*（MIT Press, 1992）一书是对该范式的重要参考。在Minoru Asada和Hiroaki Kitano编写的*RoboCup-98: Robot Soccer World Cup II*（Springer, 1999）一书中，David Andre和Astro Teller发表了“Evolving team Darwin United”*一文，这篇文章中介绍了进化的机器人足球队。John Koza、Forrest Bennett III、David Andre和Martin Keane著的*Genetic Programming III*（Morgan Kaufmann, 1999）一书包含许多经过改进的电子线路。Danny Hillis在“Co-evolving parasites improve simulated evolution as an optimization procedure”*（*Physica D*, 1990）一文中提出寄生虫有助于进化。AD iLivnat、Christos Papadimitriou、Jonathan Dushoff和Marcus Feldman在“A mixability theory of the role of sex in evolution”*（*Proceedings of the National Academy of Sciences*, 2008）一文中提出，性别可优化可混性。Kevin Lang的文章“Hill climbing beats genetic search on a Boolean circuit synthesis problem of Koza’s”*（*Proceedings of the Twelfth International Conference on Machine Learning*, 1995）对遗传编程和爬山法进行了比较。Koza对此的回应见“A response to the ML-95 paper entitled...”*（unpublished; online at www.genetic-programming.com/jktahoe24page.html）一文.

在“A new factor in evolution”（*American Naturalist*, 1896）一文中，James Baldwin提出了同名反应。Geoff Hinton和

Steven Nowlan 在“How learning can guide evolution”*（*Complex Systems*, 1987）一文中介绍了它的运用。鲍德温效应是Peter Tusrney、Darrell Whitley和Russell Anderson主编的期刊*Evolutionary Computation*1996年的一个特殊问题的主题。

John Neville Keynes的*The Scope and Method of Political Economy*（Macmillan, 1891）一书区分了描述性和规范性理论之间的区别。

第六章

Sharon Bertsch McGrayne 在 *The Theory That Would Not Die*（Yale University Press, 2011）一书中对贝叶斯主义进行了介绍，从Bayes到Laplace再到现在。Peter Hoff的*A First Course in Bayesian Statistical Methods**（Springer, 2009）一书则对贝叶斯统计进行了介绍。

在Richard Duda和Peter Hart（Wiley, 1973）的*Pettern Classification and Scene Analysi**一书中，朴素贝叶斯算法被首次提到。在*Essays in Positive Economics*（University of Chicago Press, 1966）一书中，Milton Friedman发表了自己对过简理论的支持。朴素贝叶斯在过滤垃圾邮件中的应用则出现在Joshua Goodman、David Heckerman和Robert Rounthwaite的“Stopping spam”（*Scientific America*, 2005）一文中。Stephen Robertson和Karena Sparck Jones的“Relevance weighting of search terms”*（*Journal of the American Society for Information Science*, 1976）一文介绍了与朴素贝叶斯相似方法在信息检索中的应用。

Brian Hayes的“First links in the Markov chain”（*American Scientist*, 2013）一文叙述了马尔可夫对同名链的创造。Thorsten Brants等人的“Large Language models in machine

translation”*（*Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007）一文解释了谷歌翻译如何运行。Larry Page、Sergey Brin、Rajeev Motwani和Terry Winograd的“The PageRank citation ranking: Bringing order to the Web”（Standford University technical report, 1998）一文介绍了网页排序算法及其在网页上游动的说明。Eugene Charniak的*Statistical Language Learning**（MIT Press, 1996）一书揭示了隐藏的马尔可夫模型的工作原理。Fred Jelinek的*Statistical Methods for Speech recognition*（MIT Press, 1997）一书描述了这些方法在语音识别领域的应用。在David Forney的“The Viterbi algorithm: A personal history”（unpublished; online at arxiv.org/pdf/cs/0504020v2.pdf）一文中，介绍了隐马尔可夫模型式的推理在传播中的情况。Pierre Baldi和Søren Brunak的*Bioinformatics: The Machine Learning Approach**（2nd ed., MIT Press, 2001）一书介绍了机器学习在生物学以及隐马尔可夫模型中的应用。Barry Cipra的“Engineers look to Kalman filtering for guidance”（*SIAM News*, 1993）对卡尔曼滤波法及其历史、应用进行了简要介绍。

Judea Pearl关于贝叶斯网络的先锋之作出现在他的*Probabilistic Reasoing in Intelligent Systems**（Morgan Kaufmann, 1988）一书中。Eugene Charniak的“Bayesian networks without tears”*（*AI Magazine*, 1991）一文很大程度上用非数学的方法对贝叶斯网络进行了介绍。David Heckerman的“Probabilistic interpretation for MYCIN’s certainty factors”*（*Proceedings of the Second Conference on Uncertainty in Artificial Intelligence*, 1986）一文解释在何种情况下，包含置信度估计的规则集是贝叶斯网络的合理假设，在什么情况下不是。Eran Segal等人的“Module networks: Identifying regulatory modules and their

condition-specific regulators from gene expression data”（*Nature Genetics*, 2003）一文就是贝叶斯网络在基因调控领域的一个应用。Ben Paynter的“Microsoft virus fighter: Spam may be more difficult to sthp than HIV”（*Fast Company*, 2012）一文讲述David Heckerman如何从垃圾邮件过滤器中获取灵感，并利用贝叶斯网络设计了一种艾滋病疫苗。概率性的或者“喧哗”OR在Pearl的书籍中得到了解释。M.A.Shwe等人的“Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base”一文为医学诊断介绍了一种“喧哗”OR贝叶斯网络。Kevin Murphy的Machine Learning*（MIT Press，2012）的26.5.4部分描述了谷歌的贝叶斯网络在广告配置方面的运用。Ralf Herbrich、Tom Minka和Thore Graepel的“TrueSkillTM: A Bayesian skill rating system”*（*Advances in Neural Information Processing Systems 19*, 2007）一文介绍了微软的游戏评分系统。

Adnan Darwiche的*Modeling and Reasoning with Bayesian Networks*（Cambridge University Press, 2009）一书解释了贝叶斯网络中用于推理的主要算法。Jack Dongarra和Francis Sullivan主编的*Computing in Science and Engineering*期刊的2000年的1~2月刊包含了和21世纪前十位算法有关的文章，包括马尔可夫链蒙特卡洛。Sebastian Thrun等人的“Stanley: The robot that won the DARPA Grand Challenge”（*Journal of Field of Robotics*, 2006）解释了同名无人驾驶汽车如何运行。David Heckerman的“Bayesian networks for data mining”*（*Data Mining and Knowledge Discovery*, 1997）一文总结了用于学习的贝叶斯方法，并解释如何从数据中掌握贝叶斯网络。David MacKay的“Gaussian processes: A replacement for supervised neural networks?”*（NIPS tutorial notes, 1997; online at www.inference.eng.cam.ac.uk/mackay/gp.pdf）一文大致描述了贝叶斯学派如何安排网络入侵防火墙系统。

Dan Jurafsky 和 James Martin 的 *Speech and Language Processing*（2nd ed., Prentice Hall, 2009）一书的9.6部分讨论了在语音识别中对词汇概率进行加权的必要性。我和Mike Pazzani关于朴素贝叶斯的论文是“On the optimality of the simple Bayesian classifier under zero-one loss”*（*Machine Learning*, 1997; expanded journal version of the 1996 conference paper）。上面提到的Judea Pearl的书讨论了马尔可夫网络和贝叶斯网络。计算机视觉中的马尔可夫网络就是Adrew Blake、Pushmeet Kohli和Carsten Rother主编的*Markov Random Fields for Vision and Image Processing*（MIT Press, 2011）这本书的主题。John Lafferty、Andrew McCallum 和 Fernando Pereira 的“Conditional random fields: Probabilistic models for segmenting and labeling sequence data”*（*International Conference on Machine Learning*, 2001）一文介绍了可将条件释然性最大化的马尔可夫网络。

Jon Williamson 和 Dov Gabbay 主编的 *Journal of Applied Logic*2003年的一期特刊探索了试图将概率与逻辑结合起来的历史。Michael Wellman、John Breese 和 Robert Goldman 的“From Knowledge bases to decision models”*（*Knowledge Engineering Review*, 1992）一文讨论了早期一些解决该问题的智能方法。

第七章

Frank Abagnale在和Stan Redding一起创作的自传*Catch Me If You Can*（Grosset& Dunlap, 1980）中详细描述了他的英勇事迹。Evelyn Fix和Joe Hodges最初关于最近邻算法的技术报告是“Discriminatory analysis: Nonparametric discrimination: Consistency properties”*（USAF School of Aviation Medicine,

1951）。Belur Dasarathy主编的Nearest Neighbor (NN) Norms*收录了许多该领域中的重要论文。Chris Atkeson、Andrew Moore和Stefan Schaal的“Locally weighted learning”（*Artificial Intelligence Review*, 1997）一文探索了局部线性回归。Paul Resnick等人的“GroupLens: An open architecture for collaborative filtering of netnews”*（*Proceedings of the 1994 ACM Conference on Computer-Supported Cooperative Work*, 1994）一文介绍了基于最近邻的首个协同过滤系统。Greg Linden、Brent Smith和Jeremy York的“Amazon.com recommendations: Item-to-item collaborative filtering”*（*IEEE Internet Computing*, 2003）一文介绍了亚马逊的协同过滤系统（参见第八章网飞的延伸阅读部分）。Mayer-Schonberger和Cukier的*Big Data*以及Siegel的*Predictive Analysis*（之前引用过）引用了推荐系统对于亚马逊和网飞销售业绩的贡献。1967年，Tom Cover和Peter Hart写了一篇关于最近邻的误差率的文章——“Nearest neighbor pattern classification”*（*IEEE Transactions on Information Theory*）。

Trevor Hastie、Rob Tibshirani和Jerry Friedman的*The Elements of Statistical Learning**（2nd ed., Springer, 2009）的2.5部分讨论了维数灾难。Ron Kohavi和George John的“Wrappers for feature subset selection”*（*Artificial Intelligence*, 1997）一文对比了属性选择方法。David Lowe的“Similarity metric learning for a variable-kernel classifier”*（*Neural Computation*, 1995）就是一个特征赋权算法的例子。

Nello Cristianini和Bernhard Schokopf的“Support vector machines and kernel methods: The new generation of learning machines”（*AI Magazine*, 2002）一文在很大程度上运用了非数学的方法来介绍支持向量机。Bernhard Boser、Isabel Guyon和Vladimir Vapnik的“A training algorithm for optimal margin

classifiers”*（*Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 1992）。Thorsten Joachims的“Text categorization with support vector machines”（*Proceedings of the Tenth European Conference on Machine Learning*, 1998）是第一篇将支持向量机运用到文本分类中的论文。Nello Cristianini和John Shawe-Taylor的*An Introduction to Support Vector Machine*（Cambridge University Press, 2000）一书中的第五章在支持向量机的背景下简要介绍了约束优化。

Janet Kolodner的*Case-Based Reasoning**（Morgan Kaufmann, 1993）是该主题的教材。Evangelos Simoudis的“Using case-based retrieval for customer technical support”*（*IEEE Export*, 1992）一文揭示了它在求助台中的运用。“Rise of the software machines”（*Economist*, 2013）及其公司的主页上介绍了正畸诊疗软件的Eliza。Kevin Ashley在*Modeling Legal Argument**（MIT Press, 1991）中探索了基于案例的法律推理。David Cope在“Recombinant music: Using the computer to explore musical style”（*IEEE Computer*, 1991）一文中总结了他用于进行自行音乐创作的方法。Dedre Gentner在“Structure mapping: A theoretical framework for analogy”*（*Cognitive Science*, 1983）一文中提出了构造图。James Somers的“The man who would teach machines to think”（Atlantic, 2013）讨论了道格拉斯·霍夫斯泰特对于人工智能的看法。

我的文章“Unifying instance-based and rule-based inducton”*（*Machine Learning*, 1996）介绍了RISE算法。

第八章

Alison Gopnik、Andy Meltzoff和Pat Kuhl的*The Scientist in the Crib*（Harper, 1999）一书总结了心理学家关于儿童如何进行学习的发现。

1957年，Stuart Lloyd在贝尔实验室的一篇技术报告“Least squares quantizataion in PCM”*（最后以一篇论文的形式出现在1992年的*IEEE Transactions on Information Theory*）中，k均值聚类算法首次被提出。关于最大期望算法的原文是Arthur Dempster、Nan Laird和Donald Rubin的“Maximum likelihood from incomplete data via the EM algorithm”*（*Journal of the Royal Statistical Society B*, 1977）。Leonard Kaufman和Peter Rousseeuw的*Finding Groups in Data: An Introduction to Cluster Analysis*（Wiley, 1990）介绍了分级聚类及其他方法。

Kar Pearson在1901年的一篇名为“On lines and planes of closets fit to systems of points in space”*（*Philosophical Magazine*）的文章中提出，pricipal-component analysis是机器学习及统计中最古老的技术。Scott Deerwester等人的文章“Indexing by latent semantic analysis”*（*Journal of the American Society for Information Science*, 1990）介绍了用于给SAT作文打分的一种降维方法。Yehuda Koren、Robert Bell和Chris Volinsky在“Matrix factorization techniques for recommender systems”*（*IEEE Computer*, 2009）中解释网飞式协同过滤系统如何运行。Tenenbaum、Vin de Silva和John Langford的“A global geometric framework for nonlinear dimensionality reduction”*（Science, 2000）介绍了Isomap算法。

Rich Sutton和Andy Barto的*Reinforcement Learning: An Introduction*（MIT Press, 1998）一书是该主题的标准教材。Marcus Hutter的*Universal Artificial Intelligence*（Springer, 2005）尝试对加强学习进行理论上的总结。“Some studies in

machine learning using the game of checkers”*（*IBM Journall of Research and Development*, 1959）一文介绍了Arthur Samuel关于学习如何下国际象棋的开创性研究，这篇文章也是“机器学习”这个术语最早出现的印刷品之一。Chris Watkins对于强化学习问题的构想也出现在他的博士论文*Learning from Delayed Rewards*（Cambridge University, 1989）中。Voloydymyr Mnih等人的“Human-level control through deep reinforcement learning”*（*Nature*, 2015）一文介绍了DeepMind用于视频游戏的强化学习算法。

在“A cognitive odyssey: From the power law of practice to a general learning mechanism and beyond”（*Tutorials in Quantitative Methods for Psychology*, 2006）一文中重述了组块的发展。A/B测试和其他在线实验技术在Ron Kohavi、Randal Henne和Dan Sommerfield的“Practical guide to controlled experiments on the Web: Listen to your customers not to the HiPPO”*（*Proceedings of the Thirteenth International Conference on Knowledge Discovery and Data Mining*, 2007）一文得到了解释。Eric Siegel的*Predictive Analytics*（Wiley, 2013）一书第七章的主题就是隆起建模，即对A/B测试的多维度概括。

Lise Getoor和Ben Tasker编辑的*Introduction to Statistical Relational Learning*（MIT Press, 2007）一书探索了这个领域的主要方法。“Mining social networks for viral marketing”（*IEEE Intelligent Systems*, 2005）是我和Matt Richardson关于模拟口头语的文章。

第九章

Zhi-Hua Zhou 的 *Model Ensembles: Foundations and Algorithms*（Chapman and Hall, 2012）对元学习进行了介绍。关于堆叠的最早文章是David Wolpert的“Stacked generalization”*（*Neural Networks*, 1992）。Leo Breiman在“Bagging predicts”*（*Machine Learning*, 1996）中介绍了装袋预测法，在“Random forests”*（*MchineLearing*, 2001）介绍了随机森林法。Yoav Freund 和 Rob Schapire 的“Experiment with a new boosting algorithm”（*Proceedings of the Thirteenth International Conference on Machine Learning*, 1996）一文介绍了助推法。

Anil Ananthaswamy 的“I, Algorithm”（New Scientist, 2011）记录了在人工智能中将逻辑与概率结合起来的方法。我和Daniel Lowd一起合写的*Markov Logic: An Interface Layer for Artificial Intelligence**（Morgan & Claypool, 2009）一书是对马尔可夫逻辑网络的介绍。Alchemy 网站http://alchemy.cs.washington.edu也包含教程、录像、MLN、数据集、刊物和其他系统的指示器等。Jue Wang和Pedro Domingos的“Hybrid Markov logic works”*（*Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, 2008）一文介绍了MLN在DARPA的PAL项目中的用途。Stanley Kok和Pedro Domingos 的“Extracting semantic networks from text via relational clustering”*（*Procedings of the Nineteenth European Conference on Machine Learning*, 2008）一文介绍了我们如何运用MLN从网站上获取语义网络。

在 Mathias Niepert 和 Pedro Domingos 的“Learning and inference in tractable probabilistic knowledge bases”*（*Proceedings of the Thirty-first Conference on Uncertainty in Artificial Intelligence*, 2015）一文介绍了带有类层次和部件结构的有效MLN。Jeff Dean等人的“Large-scale distributed deep networks”*（*Advances in Neural Information Processing*

Systems 25, 2012）一文介绍了谷歌的并行梯度下降算法。Pedro Domingos和Geoff Hulten的“A general framework for mining massive data streams”*（*Journal of Computional and Graphical Statistics*, 2003）一文总结了我们用于掌握开放式数据流的取样方法。David Weinberger所写“The machine that would predict the future”（*Scientific American*, 2011）的主题是The FuturICT项目。

“Cancer: The march on malignancy”（Nature supplement, 2014）一文介绍了目前对抗癌症的情况。Chris Edwards的“Using patient data for personalized cancer treatments”（*Communications of the ACM*, 2014）一文介绍了有可能会发展为CaneRx的早期情况。Markus Covert的“Simulating a living cell”（*Scientific American*, 2014）解释了他的团队如何构建整个感染细菌的计算机模型。Antonio Regalado的“Breakthrough Technologies 2015: Internet of DNA”（*MIT Technology Review*, 2015）总结了全球基因组学与健康联盟的工作。Jay Tenenbaum和Jeff Shrager的“Cancer: A Computational Disease that AI Can Cure”（*AI Magazine*, 2011）一文介绍了常见的癌症。

第十章

Kevin Poulsen的“Love, acturaially”（*Wired*, 2014）介绍了一个人怎样通过机器学习来在OkOcupid约会网站上寻找爱情。Christian Rudder的*Dataclysm*（Crown, 2014）挖掘了OkCupid的数据，用于各式各样的想法。Gordon Moore和Jim Gemmell的*Total Recall*（Dutton, 2009）一书探索了对我们所做的一切进行数字化记录的启示。Patrick Tucker的The Naked Future（Current，2014）探索了在我们的世界中，进行预测时对数据的使用及滥用。在“Privacy

pragmatism”（*Foreign Affairs*, 2014）一文中，Craig Mundie支持在收集和利用数据时使用均衡法。Erik Brynjolfsson和Andrew McAfee的*The Second Machine Age*（Norton，2014）讨论了人工智能的进步会对未来的工作及经济造成的影响。Chris Baraniuk的“World War R”（*New Scientist*, 2014）总结了围绕能否在战争中使用机器人的辩论。Stephen Hawking等人的“Transcending complacency on superintelligent machines”（*Huffington Post*, 2014）一文提出当下正是担心人工智能危机的时候。Nick Bostrom的*Superintelligence*（Oxford University Press, 2014）一书提到了一些危险，并说明应该怎样应对这些危险。

Richard Hawking的*A Brief History of Life*（Random Penguin, 1982）一书总结了计算机出现之前的量子跳跃。Ray Kurzweil的*The Singularity Is Near*（Penguin, 2005）可带领我们进入“超人类”未来。在*Radical Evolution*（*Broadway Books*, 2005）中，Joel Garreau就人类指向的进化在*Radical Evolution*（Broadway Books, 2005）中如何展现设想了三个不同的情景。在*What Technology Wants*（Penguin, 2010）中，Kevin Kelly提出技术就是以其他方式来对进化进行延续。George Dyson的*Darwin Among the Machine*（Basic Books, 1997）一书回顾了技术的发展史，并设想技术将会把我们带往何处。在*Life at the Speed of Light*（Viking，2013）一书中，Craig Venter解释了他的团队如何合成活细胞。

图书在版编目（CIP）数据

终极算法：机器学习和人工智能如何重塑世界 / （美）佩德罗·多明戈斯著；黄芳萍译. -- 北京：中信出版社，2017.1
书名原文：The Master Algorithm:How the Quest for the Ultimate Learning Machine Will Remake Our World
ISBN 978-7-5086-6867-3
I. ①终… II. ①佩… ②黄… III. ①机器学习－研究②人工智能－研究 IV. ①TP18
中国版本图书馆CIP数据核字（2016）第256013号

终极算法：机器学习和人工智能如何重塑世界
著者：[美]佩德罗·多明戈斯
译者：黄芳萍
出版发行：中信出版集团股份有限公司
（北京市朝阳区惠新东街甲4号富盛大厦2座 邮编100029）
电子书排版：张明霞