# 深度强化学习

[荷]阿斯克·普拉特(Aske Plaat) 著
殷海英 译

清華大學出版社
北 京

北京市版权局著作权合同登记号　图字：01-2024-0130
First published in English under the title Deep Reinforcement Learning by Aske Plaat.

Copyright © Springer Nature Singapore Pte Ltd. 2022.

This edition has been translated and published under licence from Springer Nature Switzerland AG. Part of Springer Nature.

本书封面贴有清华大学出版社防伪标签，无标签者不得销售。
版权所有，侵权必究。举报：010-62782989，beiqinquan@tup.tsinghua.edu.cn。

图书在版编目(CIP)数据

深度强化学习 / (荷) 阿斯克·普拉特(Aske Plaat)著；殷海英译. —北京：清华大学出版社，2024.5
书名原文：Deep Reinforcement Learning
ISBN 978-7-302-65979-2

I. ①深… II. ①阿…②殷… III. ①机器学习 IV. ①TP181

中国国家版本馆 CIP 数据核字(2024)第 068092 号

**责任编辑**：王　军
**装帧设计**：孔祥峰
**责任校对**：成凤进
**责任印制**：

**出版发行**：清华大学出版社
网　　址：https://www.tup.com.cn，https://www.wqxuetang.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-83470000　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
**印 装 者**：
**经　　销**：全国新华书店
**开　　本**：170mm×240mm　　**印　　张**：16.5　　**字　　数**：521 千字
**版　　次**：2024 年 6 月第 1 版　　**印　　次**：2024 年 6 月第 1 次印刷
**定　　价**：79.80 元

产品编号：103426-01

# 致　　谢

本书得益于许多朋友的帮助。首先，我要感谢莱顿大学高级计算机科学研究所的所有人员，因为他们创造了一个充满乐趣和富有活力的工作环境。

许多人对本书作出了贡献。其中一些内容基于我们之前在强化学习课程中使用的书籍，以及由 Thomas Moerland 编写的关于基于策略的方法的讲义。Thomas 还在早期草稿中提供了非常有价值的意见。此外，在撰写本书的过程中，我们还为基于模型的深度强化学习、深度元学习和深度多智能体强化学习等主题撰写了综述文章。感谢这些综述文章的合著者 Mike Preuss、Walter Kosters、Mike Huisman、Jan van Rijn、Annie Wong、Anna Kononova 和 Thomas Bäck。

感谢莱顿强化学习社区的所有成员，他们的意见和热情对此书的成稿起了重要作用。特别要感谢 Thomas Moerland、Mike Preuss、Matthias Müller-Brockhausen、Mike Huisman、Hui Wang 和 Zhao Yang，他们在撰写本书所涉及的课程方面提供了帮助。感谢 Wojtek Kowalczyk 对深度监督学习的深入讨论，以及 Walter Kosters 对组合搜索的见解和他的幽默感。

特别感谢 Thomas Bäck，因为我们就科学、宇宙及一切事物(特别是生命进化等)进行了许多讨论。没有你，这一努力将无法实现。

本书是我在莱顿大学讲授的研究生强化学习课程的成果。感谢所有参与这门课程的学生，不论是过去、现在还是将来，感谢你们的热情、深刻的问题和众多建议。本书是为你们所写，也是由你们所写！

最后感谢 Saskia、Isabel、Rosalin、Lily 和 Dahlia，感谢她们的真实与坦诚，她们的反馈让我学到了更多，也感谢她们给予我的无尽的爱。

# 前　言

近期，深度强化学习引起了广泛关注。人们在各个领域中取得了惊人成果，如自动驾驶、电子竞技、分子重组和机器人技术。在所有这些领域，电脑程序已经学会了解决困难的问题。它们学会了驾驶模型直升机，还可以完成像循环和翻滚这样的特技动作。在某些应用中，它们甚至比人类最优秀的操作者表现得更好，例如，在 Atari 游戏、围棋、扑克和星际争霸中。

深度强化学习探索复杂环境的方式，有点像小孩子玩耍时尝试不同的事情，得到反馈后再试一次。计算机好像真的具有一些人类学习的能力；深度强化学习触及人类的梦想。

研究领域的成功引起了教育者的关注，各个大学相继开始推出相关课程。本书的目标是全面介绍深度强化学习这个领域。它是为人工智能专业的研究生，以及想要更好地了解深度强化学习方法和挑战的研究人员和从业者编写的。我们假设读者具备计算机科学和人工智能方面的本科水平，并对这些内容有基本的了解；本书使用的编程语言是 Python。

我们将描述深度强化学习的基础、算法和应用。本书将涵盖构成该领域基础的已建立的无模型和有模型方法。由于该技术发展迅速，本书还将涵盖更高级的主题：深度多智能体强化学习、深度分层强化学习和深度元学习。

希望本书会给你带来与许多研究人员一样的喜悦，他们在开发算法、最终让它们运行起来的过程中感受到了无比的快乐！

## 关于 Links 文件

阅读本书时，你会不时遇到参考资源链接，形式是[link*]，其中的*代表编号，你可扫封底二维码下载 Links 文件。例如，在阅读第 1 章正文期间，看到[link 3]时，可从 Links 文件中“第 1 章”下面的[link3]处找到具体链接。

## 关于彩图

在阅读本书正文时，提及的彩图可扫描封底二维码下载。

# 目　录

# 第 1 章

# 简　　介

深度强化学习研究的是我们如何学习解决复杂问题，这些问题要求我们在很多不同情况下做出正确决定。例如，要做面包，就需要选择适合的面粉，加点盐、酵母和糖，调配适合的面团，然后在合适的温度下烘烤；要在舞蹈比赛中获胜，就需要找到合适的舞伴，学会跳舞，不断练习，然后在比赛中打败对手；在下国际象棋时，我们需要学习规则，多练习，并且走出最明智的棋步。

## 1.1　什么是深度强化学习

深度强化学习是深度学习和强化学习的结合。

深度强化学习的目标是学习在各种环境状态(比如面包店、舞厅、国际象棋棋盘)都能最大化奖励的最佳动作。我们通过与复杂的高维度环境互动，尝试不同的动作，并从反馈中学习来实现这一目标。

深度学习领域关注的是高维问题中的近似函数，这些问题非常复杂，传统的表格方法已经不能找到精确解了。深度学习使用深度神经网络为大型、复杂、高维度的环境(比如在图像和语音识别领域)寻找近似解。该领域取得了令人瞩目的进展；计算机现在可以在一系列图像中识别行人(以避免发生碰撞)并理解诸如“明天天气会怎么样？”这样的句子。

强化学习领域关注的是从反馈中学习；它是通过试错来学习的。强化学习不需要通过预先存在的数据集进行训练；它会选择自己的动作，并从环境提供的反馈中进行学习。可以这样理解，在这个试错过程中，智能体会犯错(例如，在学习如何烤面包的过程中，准备灭火器是必不可少的)。强化学习领域的核心是从成功和错误中学习。

近年来，深度学习和强化学习这两个领域相结合，产生了能够通过对动作的反馈来逼近高维问题的新算法。通过以策略为基础的方法、基于模型的方法、迁移学习、分层强化学习以及多智能体学习等方面的进展，深度学习引入了新的方法并带来了新的成功。

这两个领域也可以独立存在，分别是深度监督学习和表格强化学习(见表 1-1)。深度监督学习的目标是从预先存在的数据集里面泛化和逼近复杂的高维函数，而不必进行交互；附录 B 讨论了深度监督学习。表格强化学习的目标是在较简单、低维的环境(如网格世界)中进行交互学习；我们将在第 2 章讨论表格强化学习。

表 1-1　深度强化学习的组成部分

| | 低维状态 | 高维状态 |
|---|---|---|
| 静态数据集 | 经典监督学习 | 深度监督学习 |
| 智能体/环境交互 | 表格强化学习 | 深度强化学习 |

下面详细介绍这两个领域。

### 1.1.1　深度学习

经典的机器学习算法在数据上学习预测模型，使用线性回归、决策树、随机森林、支持向量机和人工神经网络等方法。这些模型的目标是泛化，进行预测。从数学角度看，机器学习旨在从数据中逼近一个函数。

过去，当计算机运算速度较慢时，所使用的神经网络由少数层的全连接神经元组成，在处理复杂问题时表现并不令人满意。随着深度学习和计算机速度的提升，这种情况发生了变化。深度神经网络现在包含许多层神经元，并使用不同类型的连接[1]。深度网络和深度学习将某些重要的机器学习任务的准确性提升到一个新水平，并使得机器学习能够应用于复杂的高维问题，比如在高分辨率(百万像素级别)图像中识别猫和狗。

深度学习可以实时解决高维复杂问题，使机器学习可以应用到日常生活中，如智能手机中的人脸识别和语音识别。

### 1.1.2　强化学习

让我们更深入地研究强化学习，看看从我们自己的行为中学习意味着什么。

强化学习是一个智能体通过与环境交互来学习的领域。在监督学习中，我们需要预先存在的标注示例的数据集来逼近一个函数；而强化学习只需要一个可为智能体尝试的行为提供反馈信号的环境。这一要求更容易满足，使得强化学习比监督学习应用的情况更加广泛。

强化学习智能体通过自己的动作，通过环境提供的奖励，生成即时的数据。智能体可选择学习哪些行为；强化学习是主动学习的一种形式。从这个意义上说，智能体

1 这里的“多层”指输入层和输出层之间隐藏层的数量超过 1 个。

像孩子一样，通过游戏和探索来自主学习某项任务。这种自主性是吸引研究人员投身该领域的原因之一。强化学习智能体选择执行哪种动作(测试哪种假设)，并调整对有效动作的理解，建立在遇到的各种环境状态下要执行的动作策略。注意，这种自由也使强化学习很困难，因为当你被允许选择自己的例子时，很容易停留在自己的舒适区，陷入正向增强的泡沫，以为自己表现很好，但很少学习周围的世界。

### 1.1.3 深度强化学习

深度强化学习将高维问题的学习方法与强化学习相结合，实现了高维的交互式学习。人们对深度强化学习感兴趣的一个重要原因是它在当前计算机上表现良好，并且似乎可以应用到不同任务中。例如，在第 3 章中，我们将看到深度强化学习如何学习手眼协调来玩 20 世纪 80 年代的视频游戏；在第 4 章中，我们看到一个模拟机器人猎豹如何学习跳跃；在第 6 章中，我们看到它如何通过自学复杂的策略游戏，击败世界冠军。

下面具体介绍深度强化学习适用于哪些场景。

### 1.1.4 应用

简单来讲，强化学习是一种教导智能体在世界中如何操作的方法。就像小孩通过不断尝试和得到反馈学会走路一样，强化学习智能体也通过行动和反馈来学习。深度强化学习可解决复杂的决策问题，通过试错不断接触问题，可学到近似的解决方法。这听起来有点复杂，但通过概括和尝试从例子中推断出模式或规则是我们日常生活中常做的事情。试错和逼近是人们学会如何处理陌生事物的方法，比如“按下这个按钮会发生什么？哦，糟糕。”或者“我在前移时如果将一条腿与另一条腿交叉，会发生什么？哎哟，摔倒了。”

#### 1. 序贯决策问题

在复杂环境中，动作是一个高层次的目标；我们可以更具体一点。强化学习关注智能体的行为。强化学习可以为序贯决策问题或最优控制问题(工程学中的叫法)找到解决方案。在现实世界中，为达到一个目标，必须做出一系列决策的情况很多。无论是烘焙蛋糕、建房子，还是玩纸牌游戏，都必须做出一系列决策。强化学习为快速有效地学习解决序贯决策问题提供了方法。

许多现实世界的问题都可以建模为决策序列[33]。例如，在自动驾驶中，智能体要面对速度控制、找到可驾驶区域以及最重要的避免碰撞等决策。在医疗保健中，治疗方案包含许多序贯决策，可以研究延迟治疗的影响。在客户服务中，自然语言处理可以帮助改进聊天机器人的对话和问答甚至机器翻译。在营销中，推荐系统可以推荐新闻、个性化建议、发送通知或者优化用户体验。在金融中，系统决定持有、买入或卖

出金融资产，以获取未来回报。在政治中，可以通过模拟一系列决策来研究政策效果。在游戏中，棋牌游戏和策略游戏包含一系列决策。在计算创作中，绘画需要一系列美学决策。在工程中，获取物品和使用材料包含一系列决策。在化学制造中，生产过程优化包含影响产量和质量的许多决策。最后，在能源系统中，电力分配可以建模为一个序贯决策问题。

所有这些情况下，我们都必须做出一系列决策。并且我们要知道，在这些情况下做出错误决策将付出极大的代价。

序贯决策制定的算法研究主要聚焦于两类应用：机器人技术和游戏。下面将更详细地介绍这两个领域，首先是机器人技术。

### 2. 机器人技术

从原理上讲，机器人应该采取的所有动作都可由程序员一步一步地预先通过编程来实现，详细设计每个细节。在高度控制的环境中，比如汽车工厂中的焊接机器人，这种方法可以实现，但是任何微小的改变或新增任务都需要对机器人重新编程。

通过手动对机器人编程并让它执行复杂任务，这十分困难。人类并不知晓自己的操作知识，比如拿起杯子时对哪些肌肉施加了什么“电压”。定义一个期望的目标状态，让系统自行找到复杂的解决方案，这会简单得多。此外，在略有挑战的环境中，机器人必须能更灵活地响应不同条件，这需要使用能够自适应的程序。

机器人技术是机器学习研究的重要应用领域之一，这一点不足为奇。机器人研究人员很早就开始寻找让机器人可以自主学习某些行为的方法。

机器人实验资料丰富多样。机器人可以自学如何在迷宫中导航、如何执行操作任务以及如何学习运动任务。

自适应机器人的研究取得了相当大的进展。例如，最近的成就之一涉及给煎饼翻面[29]和飞行特技模型直升机[1,2]；参见图 1-1 和图 1-2。通常情况下，学习任务与计算机视觉相结合，其中机器人必须通过对其自身动作结果的视觉解释进行学习。

图 1-1　机器人给煎饼翻面[29]


图 1-2 操纵模型直升机进行空中特技表演[2]

### 3. 游戏

现在我们转向介绍游戏。自古以来，人们就使用谜题和游戏来研究智能行为的各个方面。实际上，在计算机尚未具备足够的计算能力来运行国际象棋程序时，也就是在香农和图灵那个时代，人们进行了纸上的设计，希望通过理解国际象棋来探索智能的本质[38, 41]。

游戏让研究人员能够缩小研究范围，专注于在有限的环境中进行智能决策研究，而不必掌握真实世界的全部复杂性。除了象棋和围棋等桌面游戏，视频游戏也广泛用于测试计算机中的智能方法。例如，类似于 Pac-Man[32]的街机游戏以及类似于星际争霸[43]的多人策略游戏。详见图 1-3、图 1-4、图 1-5 和图 1-6。

图 1-3 国际象棋

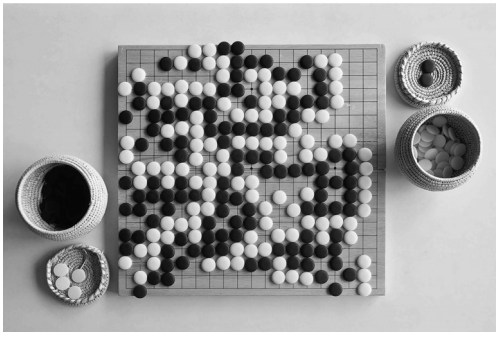

图 1-4 围棋

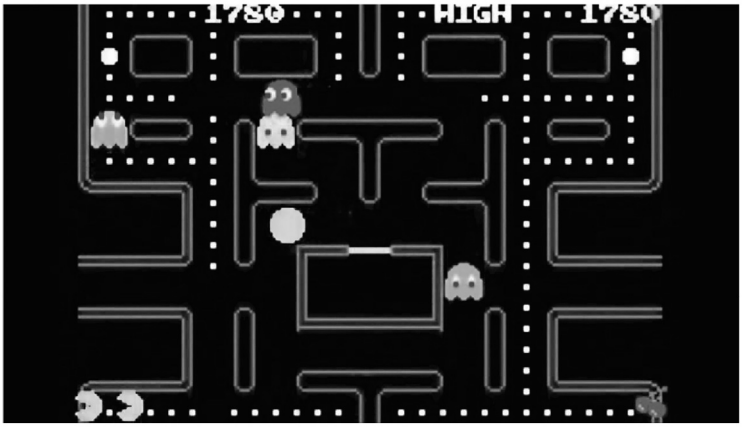


图 1-5　游戏：Pac-Man[6]

图 1-6　游戏：星际争霸[43]

### 1.1.5　四个相关领域

强化学习是一个内容丰富的领域，早在人工智能领域开始之前，它在生物学、心理学和教育等领域中就已存在[9, 25, 40]。在人工智能中，强化学习已成为机器学习的三大主要类别之一，其他两类是监督学习和无监督学习[10]。本书涵盖了从自然科学和社会科学等领域获得的灵感，其中包含一系列算法。尽管本书的其余部分将介绍这些算法，但简要讨论深度强化学习与人类以及动物学习之间的联系也很有趣。我们将介绍对深度强化学习产生深远影响的四个相关科学领域。

#### 1. 心理学

在心理学中，强化学习也被称为条件学习或操作性条件学习。图 1-7 阐述了一个

关于狗如何被条件化的常见心理观点。当狗接触到食物时，会自然地流口水。通过每次给狗食物时敲响铃声，狗就会学会将声音与食物联系起来，经过足够多次的尝试后，狗一听到铃声就开始流口水，这可能是狗在期待食物，无论实际上是否有食物。

行为科学家巴甫洛夫(1849—1936)和斯金纳(1904—1990)以他们在条件反射方面的研究而闻名。例如，诸如“巴甫洛夫反应”之类的词汇已经进入我们的日常语言，并且有关条件反映的各种笑话也存在(例如，图 1-8 所示)。心理学中关于学习的研究是对我们在人工智能中所知的强化学习的主要影响之一。

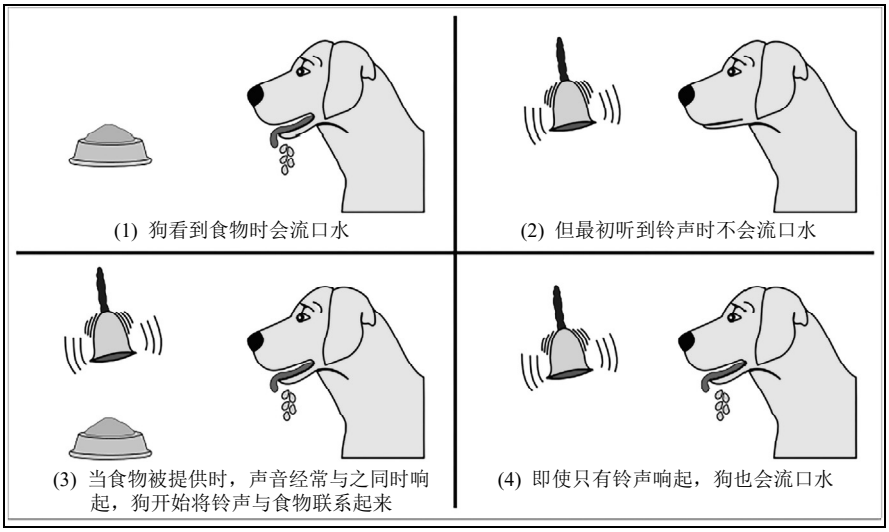


图 1-7 经典条件反映


图 1-8 谁对谁进行条件化呢？

### 2. 数学

数学逻辑是深度强化学习的另一个基石。在强化学习的形式化过程中，离散优化和图论发挥着重要作用，正如将在 2.2.2 节中详细介绍的马尔可夫决策过程。数学形式化的应用为高效规划和优化算法的发展提供了契机，这些算法在当前的进展中占据核心地位。

规划和优化是深度强化学习的重要组成部分。它们还与运筹学领域有关，尽管在那里的重点是(非顺序的)组合优化问题。在人工智能领域，规划和优化被用作构建序列、高维问题的学习系统的基础模块，这些问题可以包括视觉、文本或听觉输入。在这些系统中，规划和优化有助于有效地引导学习过程，使得智能体能逐步改进其在复杂环境中的表现。

符号推理领域是基于逻辑的，是人工智能领域最早的成功案例之一。从符号推理的工作中涌现出了启发式搜索[34]、专家系统和定理证明系统。其中一些知名系统包括 STRIPS 规划器[17]、Mathematica 计算代数系统[13]、逻辑编程语言 PROLOG[14]，还有用于语义(Web)推理的 SPARQL 等系统[3, 7]。这些系统在各自的领域内发挥着重要作用，有助于处理复杂问题、进行推理和推断。

符号人工智能专注于在离散领域中进行推理，例如决策树、规划以及策略游戏(如国际象棋和跳棋)。符号人工智能在网络搜索方法、在线社交网络以及在线商务领域取得了成功。这些极为成功的技术构成了现代社会和经济的基础。2011 年，计算机科学领域的最高荣誉图灵奖授予了朱迪亚 • 皮尔(见图 1-9)[1]，以表彰他在因果推理方面的工作。后来，皮尔出版了一本影响深远的书来推广这一领域[35]。

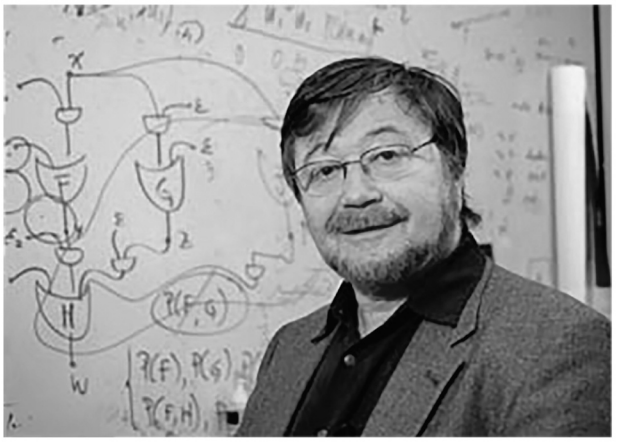

图 1-9　图灵奖获得者朱迪亚 • 皮尔

1 早前荣获图灵奖的人工智能研究者包括 Minsky(明斯基)、McCarthy(麦卡锡)、Newell(纽厄尔)、Simon(西蒙)、Feigenbaum(费根鲍姆)和 Reddy(雷迪)。

在深度强化学习中，数学的另一个重要领域是连续(数值)优化。连续优化方法在当前深度学习算法中扮演着关键角色，例如高效的梯度下降和反向传播方法。这些方法对于训练神经网络和优化模型参数至关重要，有助于算法更快地收敛并提升性能。

### 3. 工程

在工程领域，强化学习领域通常被称为最优控制。动态系统的最优控制理论由理查德·贝尔曼和列夫·庞特里亚金[8]开发。最初，最优控制理论聚焦于动态系统，其中的技术和方法与连续优化方法密切相关，例如在机器人领域的应用(参见图 1-10，图中展示了最优控制在对接两个航天器时的示例)。这一理论在工程中具有极其重要的地位，涵盖了众多问题的核心。

迄今为止，强化学习和最优控制仍然使用不同的术语和符号表示方法。在以状态为导向的强化学习中，状态和动作被表示为 $s$ 和 $a$，而工程领域的最优控制则使用 $x$ 和 $u$。在本书中，采用了前一种符号表示方式。

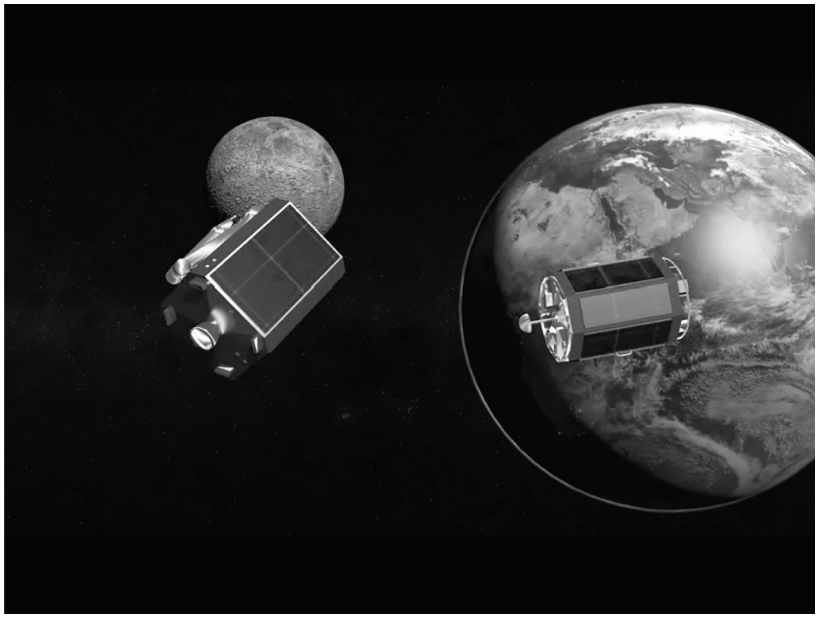

图 1-10 动态系统的最优控制实际应用示例

### 4. 生物学

生物学对计算机科学产生了深刻影响。许多受自然启发的优化算法在人工智能领域得到了发展。一个重要的自然启发派别是连接主义人工智能。

数学逻辑和工程方法将智能视为自上而下的演绎过程；实际世界中的可观察效果是从理论和自然法则的应用中得出的，智能则从理论演绎而来。相反地，连接主义以自下而上的方式处理智能。连接主义智能源自许多低层次的相互作用。智能从实践中归纳而来。智能是具体体现的：蜂巢中的蜜蜂、蚁群中的蚂蚁，以及大脑中的神经元都在相互作用，而从这些连接和相互作用中产生了我们所认知的智能行为[11]。

连接主义智能方法的例子包括受自然启发的算法，如蚁群优化[15]、种群智能[11, 26]、进化算法[4, 18, 23]、机器人智能[12]，以及人工神经网络和深度学习[19, 21, 30]。

需要注意，符号主义和连接主义人工智能两派都取得了极大的成功。在搜索和符号主义人工智能(如谷歌、Facebook、亚马逊、Netflix)产生了巨大的经济影响之后，过去二十年中人工智能领域的许多兴趣都受到了连接主义方法在计算机语言和视觉领域的成功的启发。2018 年，深度学习领域的三位关键研究者——Bengio、Hinton 和 LeCun 获得了图灵奖(见图 1-11)。他们在深度学习方面的最著名论文为[30]。

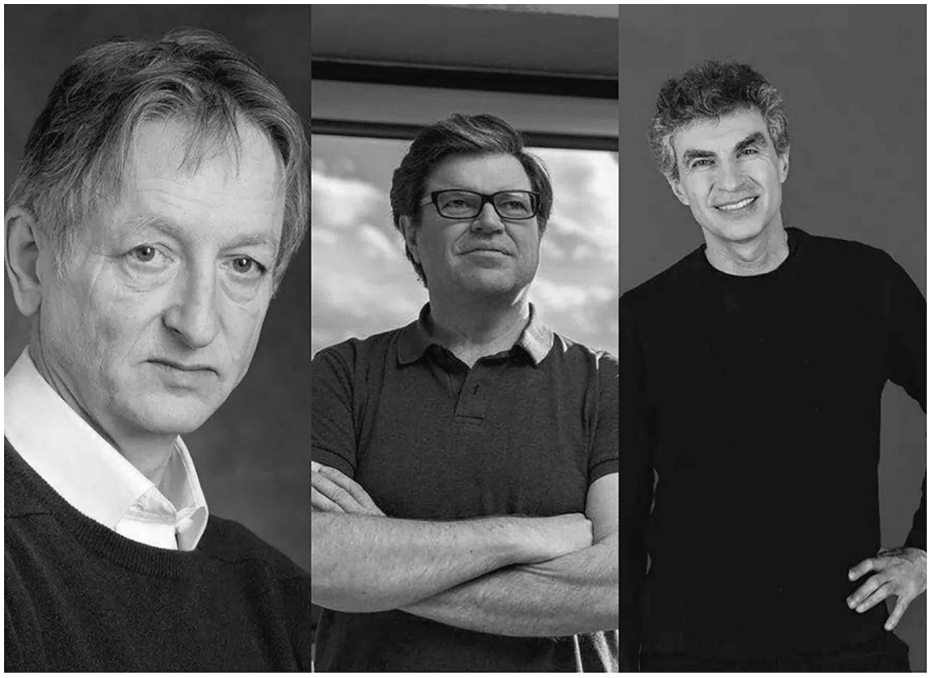

图 1-11　图灵奖得主 Hinton、LeCun 和 Bengio

## 1.2　三种机器学习范式

既然我们已经介绍了深度强化学习的一般背景和起源，下面我们转换视角，讨论一下机器学习。我们将探讨深度强化学习在该领域的总体框架中的定位。同时，我们将借此机会引入一些符号表示和基本概念。

在下一节，我们将提供本书内容的概述。但首先，我们将进入机器学习的领域，将从最基础的地方开始介绍函数逼近。

### 表示一个函数

函数是人工智能中的核心部分。函数 $f$ 根据某种方法将输入 $x$ 转换为输出 $y$，我们用 $f(x)\rightarrow y$ 表示。为对函数 $f$ 进行计算，必须以某种形式将函数表示为计算机内存中的程序。还可将函数表示为：

$$f: X \rightarrow Y$$

其中，定义域 $X$ 和值域 $Y$ 可以是离散或连续的；维度($X$ 中属性的数量)可以是任意的。

在现实世界中，同一输入可能产生多种不同的输出，因此我们希望函数能够提供一个条件概率分布，即一个将输入映射到输出概率的函数。这在深度学习中是非常常见的情况：

$$f: X \rightarrow p(Y)$$

这里，函数将定义域映射到值域上的概率分布 $p$。表示条件概率使我们能够对那些输入并不总是产生相同输出的函数进行建模(附录 A 提供了更多的数学背景信息)。

### 已知函数与学习函数

有时，我们感兴趣的函数是已知的，可通过特定的算法来表示这个函数，这个算法可以计算出一个已知的精确解析表达式。这种情况通常出现在物理定律的描述中，或者当我们对特定系统进行明确的假设时。

**示例** 牛顿的第二定律描述了质量恒定的物体的运动状态。

$$F = m \cdot a$$

其中，$F$ 表示作用在物体上的合力，$m$ 表示物体的质量，$a$ 表示物体的加速度。这种情况下，解析表达式为每种可能的输入组合定义了整个函数。

然而，在现实世界中，许多函数并没有解析表达式。这种情况下，我们进入了机器学习(特别是监督学习)的领域。当没有一个函数的解析表达式时，最好的方法是收集数据，也就是$(x, y)$的成对示例，然后通过这些数据进行逆向工程或对函数进行学习。参见图 1-12(可扫封底二维码下载彩图，后同)。

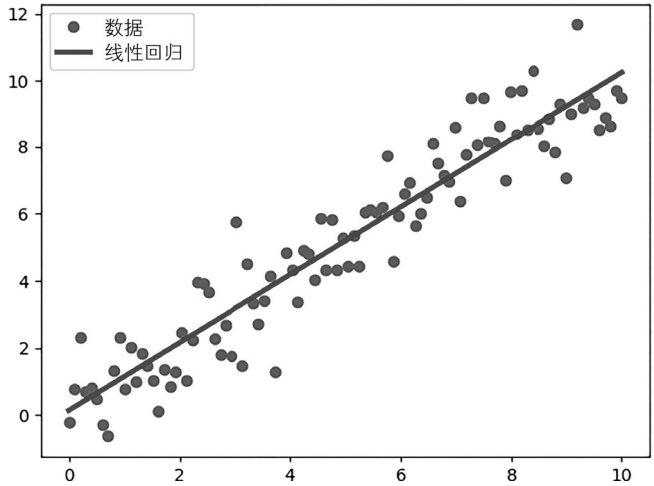


图 1-12　学习函数的示例；数据点以蓝色表示，可能的学习线性函数为红色线条，这使我们能够对任何新的输入 $x$ 进行预测，得到预测值 $\hat{y}$

**示例**　一家公司想要根据你的年龄来预测你购买染发洗发水的可能性。他们收集了许多数据点，其中 $x \in N$ 表示你的年龄(一个自然数)，映射到 $y \in \{0, 1\}$，一个二进制指示器，表示你是否购买了洗发水。然后他们想要学习这个映射：

$$\hat{y} = f(x)$$

其中，$f$ 是所期望的函数，用于告诉公司谁会购买该产品，而 $\hat{y}$ 则是预测的 $y$(在这个示例中确实过于简单)。

让我们看看在机器学习中有哪些方法可用来寻找函数逼近。

**三种范式**

在机器学习中，有三种主要的范式来提供观察数据：①监督学习；②强化学习；③无监督学习。

## 1.2.1　监督学习

机器学习中的第一个也是最基础的范式是监督学习。在监督学习中，用于学习函数 $f(x)$的数据以$(x, y)$示例对的方式提供给学习算法。这里，$x$ 表示输入，$y$ 表示观测输出值；针对特定的输入值 $x$，我们希望学习到相应的输出 $y$。这些 $y$ 值可以被视为监督学习过程中的指导，它们教导学习过程为每个输入值 $x$ 提供正确的答案，因此称为监督学习。

用于学习的“数据对”被组织成一个数据集，在算法开始之前，整个数据集必须完整存在。在学习过程中，会创建对生成数据的真实函数的估计值 $\hat{f}$。“数据对”中的

$x$ 值也称为输入，而 $y$ 值则是待学习的标签。

在监督学习中，存在两个广为人知的问题，分别是回归和分类。回归问题用于预测连续数值，而分类问题用于预测离散类别。其中，最为人熟知的回归关系是线性关系：即我们从入门统计课程中学习到的，通过一系列观测点绘制的直线。图 1-12 展示了这种线性关系，其中 $\hat{y} = a \cdot x + b$。线性函数可通过两个参数 $a$ 和 $b$ 来表征。当然，还有更复杂的函数，如二次回归、非线性回归，甚至高阶多项式回归[16]。

对于每个数据项 $i$，可使用 $(\hat{f}(x_i) - y_i)^2$ 这样的函数将监督信号计算为当前估计值与给定标签之间的差。误差函数 $(\hat{f}(x) - y)^2$ 也称为损失函数；它衡量了我们预测的质量。预测越接近真实标签，损失越低。有许多方法来计算这种接近程度，比如均方误差损失 $\mathcal{L} = \frac{1}{N}\sum_1^N (\hat{f}(x_i) - y_i)^2$，这经常用于回归中的 $N$ 个观察点。这个损失函数可供监督学习算法用来调整模型参数 $a$ 和 $b$，以将数据拟合到函数 $\hat{f}$。有许多可能的学习算法，如线性回归和支持向量机[10, 36]。

在分类中，我们学习了输入值和类别标签之间的关系。一个被广泛研究的分类问题是图像识别，其中需要对二维图像进行分类。图 1-13 展示了一组标记图像，其中包含了常见的猫和狗。在分类中，一种常见的损失函数是交叉熵损失 $\mathcal{L} = -\sum_1^N y_i \log(\hat{f}(x_i))$，详见 A.2.5 节。同样，这样的损失函数可用来调整模型参数，以使函数拟合数据。通常用于图像分类的模型可以是小型的、线性的，带有少量参数，也可以是大型的，带有许多参数，如神经网络。

在监督学习中，存在一个大型数据集，其中所有的输入项目都有相关的训练标签。而强化学习则不同，它并不假设事先存在一个带标签的大型训练集。无监督学习需要一个大型数据集，但不需要用户提供输出标签；它只需要输入数据即可。

| 猫 | 猫 | 狗 | 猫 | 狗 | 狗 |
|---|---|---|---|---|---|

图 1-13 用于监督分类问题的“输入/输出对”

深度学习的函数逼近最初是在监督式环境中开发的。尽管本书关注深度强化学习，但在讨论深度强化学习的深度学习方面时，我们经常会遇到监督学习的概念。

## 1.2.2 无监督学习

当数据集中没有标签时，就必须使用不同的学习算法。这种没有标签的学习被称为无监督学习。在无监督学习中，会利用数据项固有的指标特征，如距离。无监督学习面临的一个典型问题是在数据中发现模式，如聚类或群组[42, 44]。

常用的无监督学习算法包括 $k$ 均值算法和主成分分析[24, 37]。另外有一些流行的

无监督方法，来自可视化的降维技术，如 t-SNE[31]、最小描述长度[20]以及数据压缩[5]。在无监督学习中，一个常见的应用是自动编码器，详见 B.2.6 节[27, 28]。

有时，可以这样描述监督学习和无监督学习之间的关系：监督学习的目标是学习在给定标签 $y$ 的情况下，输入数据的条件概率分布 $p(x|y)$；而无监督学习的目标是学习先验概率分布 $p(x)$[22]。

在本书中，我们会在几个地方涉及无监督方法，具体来说，当讨论自动编码器和降维等内容时(比如在第 5 章)。在本书的结尾，还会探讨可解释的人工智能，其中可解释的模型在第 10 章中发挥重要作用。

### 1.2.3 强化学习

最后一个机器学习范式就是强化学习。强化学习与之前的范式有三个区别。

第一个区别在于，强化学习是通过交互来学习的；与监督学习和无监督学习不同，在强化学习中数据逐个到来，数据集是动态生成的。强化学习的目标是找到一个策略：一个在每个可能的状态下提供最佳动作的函数。

强化学习的方法是通过与环境互动来学习智能体如何在其中运作。在强化学习中，有一个智能体负责学习策略，还有一个环境对智能体的动作提供反馈(同时执行状态变化，见图 1-14)。在强化学习中，智能体就像人类，环境就像世界。强化学习的目标是找到在每个状态下能够最大化长期累积预期奖励的动作。这种将状态映射到动作的最优函数称为最优策略。

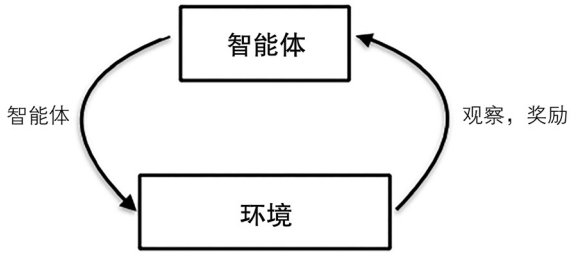


图 1-14　智能体和环境

在强化学习中，没有老师或监督者，也没有静态的数据集，但有一个环境，会指出所处状态的好坏。这就带来了第二个区别：奖励值。强化学习给我们提供了部分信息，一个数字指示了将我们带到当前状态的动作的质量；而监督学习则提供完整信息，即一个标签，它在该状态下提供了正确答案(表 1-2)。在这个意义上，强化学习介于监督学习和无监督学习之间，监督学习中所有数据项都有标签，而无监督学习中数据都没有标签。

表 1-2　监督学习与强化学习

| 概念 | 监督学习 | 强化学习 |
| --- | --- | --- |
| 输入 $x$ | 完整的状态数据集 | 部分(一次一个状态) |
| 标签 $y$ | 完整的(正确的动作) | 部分的(数值动作奖励) |

第三个区别在于，强化学习用于解决序贯决策问题。监督学习和无监督学习，学习的是项目之间的单步关系；而强化学习则学习一种策略，这个策略是多步问题的解决方案。监督学习可对一组图像进行分类；无监督学习可以告诉你哪些项目彼此相关；强化学习则可指出在国际象棋游戏中获胜的移动序列，或者机器人腿部为了行走所需的动作序列。

这三个区别会产生一些影响。在强化学习中，数据是逐步、逐动作地为学习算法提供的，而在监督学习中，数据一次性以一个大型数据集的形式提供。逐步方法适用于解决序贯决策问题。然而，许多深度学习方法是为了监督学习而开发的，在逐个生成数据时，数据可能表现出不同的特性。此外，由于动作是通过策略函数选择的，而动作奖励用于更新同一策略函数，可能导致循环反馈和局部最小值的问题。因此，在我们的方法中，需要注意确保收敛到全局最优解。人类学习也会受到这个问题的影响，就像当一个固执的孩子拒绝走出舒适区时。这个话题将在 2.2.4 节中讨论。

另一个不同之处在于，在监督学习中，学生从一个能力有限的教师(数据集)那里学习，而在某个时刻可能已经学到了所有可学的内容。强化学习范式提供了一种学习设置，智能体可以持续从环境中采样，只要环境保持挑战性(例如在国际象棋和围棋等游戏中)，智能体就会不断变得更加智能。这种持续学习的能力使得强化学习适用于长期持续的任务[1]。

因此，在强化学习方面存在着极大的兴趣，尽管使这些方法运作起来通常比监督学习更具挑战性。

许多经典的强化学习方法使用表格法，适用于具有小状态空间的低维问题。然而，许多现实世界的问题却复杂且高维，拥有广阔的状态空间。随着学习算法、数据集和计算能力的不断改进，深度学习方法已经变得非常强大。其中涌现出的深度强化学习方法成功地将高维问题和大状态空间中逐步采样的策略结合起来。后续章节中将详细讨论这些方法。

## 1.3　本书概述

本书旨在呈现深度强化学习领域的最新洞见，适合用作研究生的单学期课程。

1 实际上，一些人认为奖励足以支持人工通用智能，可参考 Silver、Singh、Precup 和 Sutton 的研究[39]。

除了介绍最先进的算法，还将涵盖经典强化学习和深度学习领域的必要背景知识。此外，将讨论自我对抗、多智能体、分层和元学习等领域的前景。

### 1.3.1 预备知识

为确保全面性，我们对先前知识有一些适度的假设。假定读者拥有计算机科学或人工智能的学士水平，且对人工智能和机器学习有浓厚兴趣。一本优秀的入门教材是 Russell 和 Norvig 的《人工智能，一种现代方法》[36]。

图 1-15 展示了本书的结构概览。深度强化学习融合了深度监督学习和经典的(表格型)强化学习。这个图展示了各章如何在这个双重基础上构建而成。在深度强化学习领域中，深度监督学习是非常重要的。它是一个广阔、深入且丰富的领域。许多学生可能已经学过深度学习课程；如果还没有，附录 B 将提供所需的背景知识(使用虚线标记)。另一方面，表格型强化学习对你来说可能有些陌生，我们将从第 2 章开始介绍这个主题。

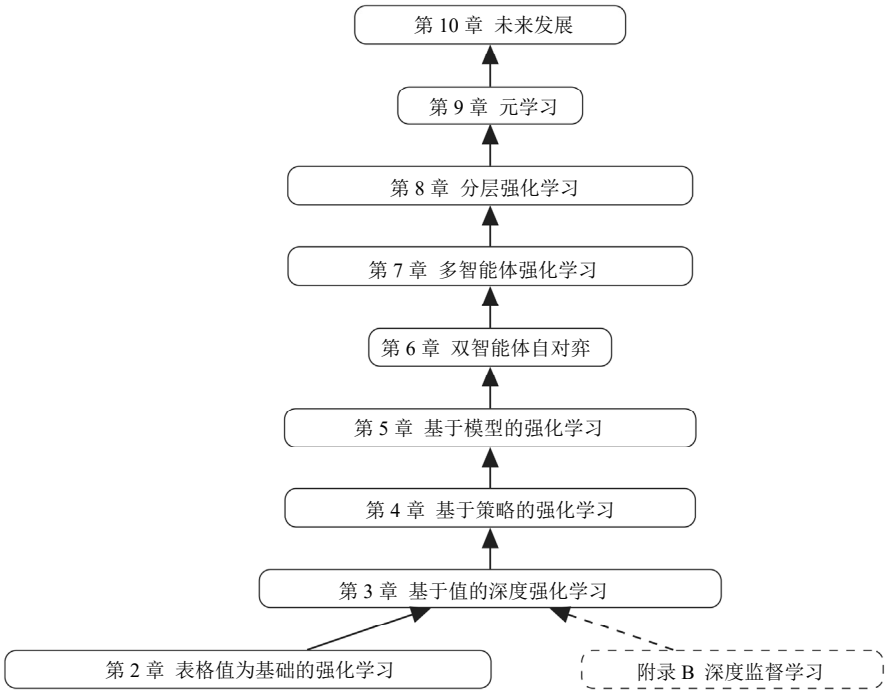


图 1-15 深度强化学习基于深度监督学习和表格型强化学习构建而成

我们还假设读者具有大学本科水平的 Python 编程语言基础。Python 已经成为机器学习研究的首选编程语言，并且是大多数机器学习软件包的主要开发语言。本书中的所有示例代码都使用 Python 编写，并且 scikit-learn、TensorFlow、Keras 和 PyTorch 等

主要机器学习环境在 Python 环境下运行效果最佳。请访问 https://www.python.org 以获取有关如何开始学习 Python 的指南。除非正文另有说明，建议使用最新的稳定版本。

我们假设读者具备大学本科水平的数学基础，包括对集合论、图论、概率论和信息论有基本的理解，尽管本书不是一本数学书籍。附录 A 包含了一个概要，可供你复习数学知识，并介绍了本书中使用的符号表示法。

#### 1. 关于课程

本书内容丰富，涵盖了基础和高级的材料，并提供了许多参考文献。你有两种选择，一种是开设一门涵盖全书主题的课程。另一种是选择深入展开，花更多时间理解基础知识，创建一门关于第 2~5 章内容的课程，介绍基础主题(如基于值、基于策略和基于模型的学习)。此外，可创建一门独立的课程，覆盖第 6~9 章中多智能体、分层和元学习这些更高级的主题。这将有助于学生更好地掌握深度学习和强化学习领域的知识。

#### 2. 博客和 GitHub

深度强化学习领域充满了活力，理论与实践紧密结合。这个领域的文化非常开放，你会很容易找到许多关于有趣主题的博客文章，其中有些质量相当不错。理论推动着实验，实验结果又推动理论的深入研究。许多研究人员会在 arXiv 上发表论文，并在 GitHub 上分享他们的算法、超参数设置以及所用的环境。

在本书中，我们力求营造相同的氛围。在全书中，我们会提供代码链接，并通过实践部分的挑战，引导你亲自动手进行实验，深入了解。我们所用的所有网页链接已经稳定存在了一段时间。

**网站：**https://deep-reinforcement-learning.net 是本书的配套网站。这个网站包含更新内容、幻灯片以及其他课程材料，欢迎你探索并使用。

### 1.3.2　本书结构

深度强化学习领域主要包括两个主要领域：无模型强化学习和有模型强化学习。这两个领域都有两个子领域。本书的章节按照这个结构进行组织：

- 无模型方法
  - 基于值的方法：第 2 章(表格学习)和第 3 章(深度学习)
  - 基于策略的方法：第 4 章
- 基于模型的方法
  - 通过学习得到模型：第 5 章
  - 给定模型：第 6 章

接下来的三章将介绍更专业的主题。

- 多智能体强化学习：第 7 章
- 分层强化学习：第 8 章
- 迁移和元学习：第 9 章

附录 B 提供了深度监督学习的必要复习。

每一章的风格都是首先列举一个直观的例子介绍该章的主要思想，然后解释需要解决的问题类型，讨论智能体使用的算法概念，以及实际上用这些算法解决的问题。章中的各节按照问题-智能体-环境的方式命名。每章结尾处会提供一些测验问题，以检查你对概念的理解，并为更大规模的编程任务提供练习题(有些相对容易，有些可能有一定挑战)。每一章的最后还会总结内容。

现在让我们更详细地看一下各章讨论的主题。

### 各章内容介绍

第 2 章详细讨论基于表格(非深度)的强化学习基础概念。我们从马尔可夫决策过程开始，并进行详细解释。将介绍表格式的计划与学习，涵盖状态、动作、奖励、价值以及策略等重要概念。我们还将接触到第一个表格式的基于值的无模型学习算法(参见表 2-1 的描述)。需要注意，第 2 章是本书中唯一不涉及深度学习方法的章节，其他所有章节都将涵盖深度学习方法。

第 3 章详细解释深度值函数强化学习。该章涵盖了最早设计出来的用于找到最优策略的深度学习算法。我们仍然会继续在基于值的、无模型的范式下工作。在该章末尾，会分析一个可以自我学习如何玩 20 世纪 80 年代的 Atari 视频游戏的智能体。表 3-1 列举一些稳定的深度值函数无模型算法。

基于值的强化学习在诸如游戏这样具有离散动作空间的应用中表现出色。第 4 章讨论一种不同的方法：基于深度策略的强化学习(参见表 4-1)。除了适用于离散空间，该方法还适用于连续动作空间，例如机械臂运动和模拟的关节运动。我们将看到如何让模拟的 Half-Cheetah 自学如何奔跑。

第 5 章将介绍基于深度模型的强化学习，这种方法使用一个学得的模型，在构建策略之前首先建立环境的转移模型。基于模型的强化学习有望提高样本效率，从而加速学习过程。还将探讨一些新的发展，如潜在模型。这种方法在机器人和游戏领域都有应用(参见表 5-2)。

第 6 章将探讨如何为问题描述中给定转移模型的应用创建自对弈系统。这种情况常见于双智能体游戏，游戏规则决定了转移函数。我们将深入研究 TD-Gammon 和 AlphaZero 通过与自身的副本对弈，从零基础到世界冠军水平的学习过程(参见表 6-2)。该章还将介绍深度残差网络和蒙特卡洛树搜索如何实现课程学习。

第 7 章将介绍深度多智能体和团队学习的最新进展。在该章中，将涵盖竞争与合作、基于种群的方法以及团队协作等内容。这些方法在扑克和星际争霸等游戏中得到

应用(参见表 7-2)。

第 8 章将涵盖深度分层强化学习。许多任务都呈现出固有的分层结构，其中可以明确地识别出子目标。将探讨选项框架，并介绍能够识别子目标、子策略和元策略的方法。此外，会讨论表格和深度分层方法中的不同途径(参见表 8-1)。

第 9 章将介绍深度元学习，也就是学会如何更快地学习。在机器学习中，学习解决新任务常常需要花费很长时间。元学习和迁移学习的目标是通过利用先前学习的相关任务信息，来加速学习新任务；相关算法请参考表 9-2。在该章的最后，还将尝试进行少样本学习，即在只见过很少的训练示例的情况下学习一个任务。

第 10 章通过回顾我们所学的内容，以及展望未来可能遇到的情况，来总结本书。

附录 A 提供了数学背景内容和符号说明。附录 B 则提供了相当于一章的内容，概述了机器学习和深度监督学习。如果你希望加深对深度学习的理解，请在阅读第 3 章之前查看该附录。附录 C 列出了有关深度强化学习的有用软件环境和软件包。