《深度强化学习》序言：

格物致知，知行合一

强化学习研究序贯决策问题，它同监督学习、无监督学习一起构成机器学习的三大学习范式。强化学习像无监督学习一样不存在有标签的训练集，但它通过与环境交互并在奖惩制度的不断刺激下驱使系统学习如何最大化自己的利益或最小化自己的损失，这也与被动地获得有标签训练数据集的监督学习场景不同。强化学习植根于人工智能领域，但它与最优控制、运筹学、随机规划有着紧密的联系。它们都试图克服经典动态规划求解高维问题所面临的“维数诅咒”。

现代强化学习主要基于随机模拟思想，它的奠基性工作始于 1989 年 Chris Watkins 提出的 Q 学习方法。人工神经网络作为一种函数逼近技术自然被引入强化学习，由此，Dimitri Bertsekas 和 John Tsitsiklis (1996)提出了神经动态规划的概念。随着深度神经网络的突破性崛起，强化学习得以“深度强化学习”而复兴。深度学习和强化学习构成现代人工智能技术的两翼。深度学习提供了一种强大的数据表示或函数逼近途径，而强化学习则提供了一种求解问题的方法论或技术路线的通用途径。

我本人于 2017 年在北京大学开始讲授深度学习，次年又讲授强化学习。虽然这两门课都有非常经典的教材供参考，但是讲授难度还是比较大---既需要兼顾数学原理和动手实践，又需要兼顾经典方法和前沿成果。特别地，深度学习更多是各种方法、技术和应用场景的荟萃，缺乏一条清晰的脉络将知识点串联起来。相对而言，强化学习的数学脉络清晰且较为具体，因为它建立在马尔可夫决策过程基础上，而贝尔曼方程定义了问题求解的最优性准则。然而，强化学习在实践上又不如深度学习有这么丰富的开源平台。

本书是王树森博士根据自己讲授的深度强化学习课程材料整理而成（详见前言部分）。本书吸收了强化学习的经典方法和最新的前沿成果，同时兼顾了算法原理和实现，适合于强化学习初学者。由于我也有讲授强化学习课程的经验和体会，我欣然接受了王树森的邀请来一起修订完善书稿。为了帮助读者更好地理解和掌握相关内容，我们又邀请黎彧君博士加入来补充算法程序实现部分。算法开源包是我的博士研究生谢广增、陈昱和黎彧君在研究强化学习过程中陆续实现并整合成的，黎彧君基于 PyTorch 进行了改写和完善。

“运用之妙，存乎一心”。强化学习提供问题及其解两者的数学表示方法，集成了数学思维和工程思维。强化学习算法通常包含价值计算和策略调整两个要素，通过“trial-and-error”学习或基于“actor-critic”框架来求解问题，这诠释着知和行一体来寻找问题最优解的思路，暗合了中国儒家修心思想“格物致知，知行合一”。其实，这也是我们做学问的金科玉律。

张志华

北京大学燕园
2022 年 7 月 22 日