

强化学习：迈向知行合一的智能机制与算法

王飞跃^{1,2,3}, 曹东璞⁴, 魏庆来^{1,2,3}

(1. 中国科学院自动化研究所复杂系统管理与控制国家重点实验室, 北京 100190;

2. 中国科学院大学, 北京, 100049;

3. 青岛智能产业技术研究院, 山东 青岛 266109;

4. 滑铁卢大学, 加拿大 滑铁卢 N2L 3G1)

摘要: 简要地评论了强化学习的历史、现状与未来的发展途径, 认为强化学习应从先行后知、先知后行向知行合一的平行强化学习迈进, 实现在虚拟世界“吃一堑”, 在物理世界“长一智”, 真正成为智慧机制和智能算法的基础学习理论。

关键词: 强化学习; 动态规划; 深度学习; 平行学习; 人工智能; 平行智能

中图分类号: TP301

文献表示码: A

doi: 10.11959/j.issn.2096-6652.202011

Reinforcement learning: toward action-knowledge merged intelligent mechanisms and algorithms

WANG Fei-Yue^{1,2,3}, CAO Dongpu⁴, WEI Qinglai^{1,2,3}

1. The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

2. University of Chinese Academy of Sciences, Beijing 100049, China

3. Qingdao Academy of Intelligent Industries, Qingdao 266109, China

4. University of Waterloo, Waterloo N2L 3G1, Canada

Abstract: This article discusses briefly the history, the state of the art and the future of reinforcement learning, and outlines a roadmap of evolution from learning by doing, doing with planning to parallel intelligence that combining learning virtually in artificial systems and acting accordingly in actual systems.

Key words: reinforcement learning, dynamic programming, deep learning, parallel learning, artificial intelligence, parallel intelligence

1 引言

人工智能的异军突起, 除计算能力和海量数据之外, 最大的贡献者当属机器学习, 其中最引人注目的核心技术与基础方法是深度学习和强化学习(reinforcement learning), 前者是前台的“明星”, 后者是背后的“英雄”^[1]。与新兴的深度学习相比, 强化学习相对“古老”, 其思想源自人类“趋利避害”和“吃一堑、长一智”的朴素意识, 其最初的“尝试法”或“试错法”, 远在人工智能技术出现

之前就在各行各业广为流行, 并成为人工智能起步时的核心技术之一。AlphaGo 在围棋人机大战中的胜利使社会大众普遍认识到有监督的深度学习和无监督的强化学习的威力^[2-4]。近年来, 人工智能算法在一些多角色游戏中大胜人类顶级专业选手, 使人们对强化学习的功力有了更加深刻的印象和理解。

例如, 以强化学习为核心技术之一的人工智能系统 Pluribus 在六人桌无限制的德州扑克比赛中, 在一万手回合里分别以单机对五人和五机对单人

收稿日期: 2020-05-08; 修回日期: 2020-05-20

通信作者: 王飞跃, feiyue.wang@ia.ac.cn

的方式，共击败 15 名全球最佳专业玩家，突破了过去人工智能仅能在国际象棋等二人游戏中战胜人类的局限，成为游戏中机器胜人的又一个里程碑，被《科学》杂志评选为 2019 年十大科学突破之一。Pluribus 这项工作之所以重要，主要原因如下。

- 人工智能算法必须处理不完备信息，需要在不知道对手策略和资源的情况下进行决策，并在不同博弈之间寻求平衡。

- 博弈最佳的理论结果是纳什平衡，但随着玩家数目的增加，求解纳什平衡的计算复杂度呈指数增长，算法要求的算力在物理上不可能实现，必须引入智力。

- 掌握“诈唬”等心理技巧是游戏胜利的关键之一，必须考虑并采用此类心理“算计”，在博弈中有效推理并隐藏意图，产生让对手难以预测和分析的策略。

解决这些问题正是人工智能进一步发展必须面对的核心任务，也是强化学习之所以关键的主要因素^[5]。这些问题的有效解决和其解决方案的广泛应用，不但可为多角色、多玩家场景下的博弈和电子竞技做出贡献，更将为人工智能在工业控制、商务决策、企业管理和军事控制等重大领域的大规模实际应用提供有效的方法和坚实的技术支撑。

强化学习为何有如此强大的功能和作用？其实强化学习的发展经历了漫长而曲折的过程，与有监督的学习方法不同，强化学习面对的是更加复杂艰巨而且“不知对错、无论好坏”的学习任务：决策或行动实施之前，没有关于正确与错误的理性推断依据；决策实施之后，没有关于好与坏的客观评价依据。然而，一百年来，科学家们坚持不懈地尝试了许多方法，包括经典条件反射（classical conditioning）、试错法（trial and error method）等“先行后知”的动物学习方法，系统模型、价值函数、动态规划、学习控制等“先知后行”的最优控制方法以及集估计、预测、自适应等于一体的时序差分（temporal difference）学习方法^[6-7]。目前，强化学习正在整合算力、数据、知识图谱、逻辑推理、智能控制和知识自动化等技术，统一关于现状、回顾、展望等因素的分析，迈向“知行合一”的复杂自适应智能机制与算法体系。图 1 给出了由 F. Woergetter 和 B. Porr 总结的强化学习前因后世^[8]，比较完整地反映了这一方法的核心内容与相关问题。

2 先行后知的起步

作为一项科学研究，强化学习始于“摸着石子过河”的启发式思维，在学术文献上可追溯到英国著名学者 Alexander Bain 的“摸索与实验（groping and experiment）”学习原理。Alexander Bain 是著名杂志 *Mind* 的创办人，正是这份杂志于 1950 年发表了 Alan Turing 的文章“计算机与智能（computing machinery and intelligence）”，提出使用“图灵测试”判断机器智能水平，开启了人工智能研究领域。作为一种方法，强化学习源自“试错学习（trial-and-error learning）”，由英国生物和心理学家 Conway Morgan 正式提出，并以“摩根法则（Morgan's canon）”为指导原则，即尽可能用低级心理功能解释生物行为的节约原则，后被美国心理学家、学习理论专家、联结主义创始人之一的 Edward Thorndike 进一步简化为“效果定律（law of effect）”，这成为后来的学习规则——Hebb 定律和神经网络误差反向传播（back propagation, BP）算法的鼻祖。强化学习的正式出现要归功于生理学家巴甫洛夫及其经典条件反射理论和激励响应（stimulus-response）理论，特别是他通过狗进行的一系列刺激反应试验总结出来的条件反射定律。在美国，心理学家 Burrhus Frederic Skinner 提出的工具条件反射（operant or instrumental conditioning）和工具学习（instrumental learning）及其利用老鼠进行试验的 Skinner-Box 技术也推动了强化学习的行为分析试错法研究。

自此之后，强化学习在动物行为研究、生理和心理学以及认知科学等领域发挥了重要作用，成为相应的核心方法与技术。在人工智能之初，从图灵基于效果定律的“快乐-痛苦系统（pleasure-pain system）”、Marvin Minsky 基于加强学习的随机神经模拟强化计算器（stochastic neural analog reinforcement calculator, SNARC）及其“迈向人工智能（steps toward artificial intelligence）”一文中提出的复杂强化学习的“基本信用（功劳）分配问题”，再到 Donald Michie 基于强化学习的 MENACE 和 GLEE 学习引擎、Nils Nilsson 学习自动机（learning automata）和学习机器（learning machines），还有 John Holland 的“分类系统（classifier systems）”及其遗传算法，强化学习的思想和方法对许多人工智能机制和算法的设计产生了深刻的影响。然而，

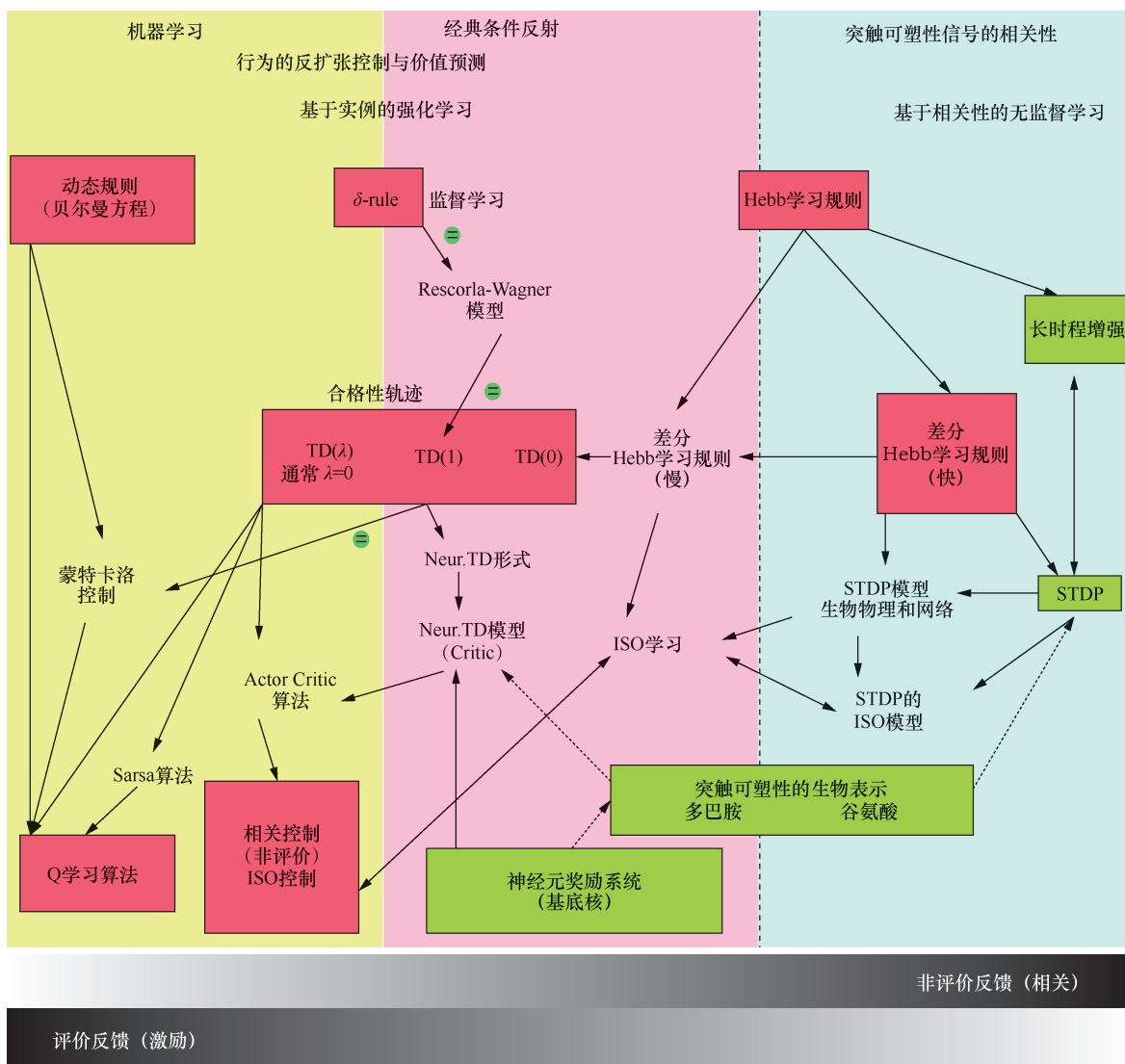


图1 强化学习的核心内容与相关问题

相对于许多机器学习方法而言，人们对强化学习的期望远大于其成果，在相当长的时间里，强化学习实际上并不是人工智能及其相关领域的主流方法和技术。

3 先知后行的重铸

基于生物和心理学并以试错法为主的强化学习没有用到太多的数学概念和工具，而且也很少在工程上应用，直到20世纪50年代，随着工程数学化的深入和现代控制理论的兴起，特别是基于系统动力学模型的最优控制的出现，加上Richard Bellman的杰出工作，强化学习走上了一条数学化和工程应用的崭新道路，局面大为改观：朴素的奖励惩罚变成了“价值函数 (value function)”，简

单的行为选择升华为“动态规划 (dynamic programming)”，非线性随机微分方程来了，伊藤积分 (Ito integral) 用上了，马尔可夫随机过程成了离散情况下的标配，有时还必须引入博弈论。强化学习从极其具体实在的动物行为学习突然变为十分复杂抽象的马尔可夫决策过程 (Markov decision process, MDP) 和Bellman方程，甚至是更难认知和求解的哈密顿-雅可比-贝尔曼 (Hamilton-Jacobi-Bellman, BHJ) 偏微分方程。一时间，原来“先知后知”的试错行为不见了，取而代之的是“先知后行”式的方程求解，尽管看起来有些“风马牛不相及”，但其一下子成为控制理论与工程的一部分，让许多研究者惊奇之余看到希望和曙光。

然而, 这道曙光仅带来了短暂的黎明, 很快又沉入“黑暗”, 强化学习在新的道路上刚起步就遭遇“维数灾难(curse of dimensionality)”, 当问题变得复杂(维数增加)时, 动态规划求解方程的计算量呈指数增加, 没有计算机可以应对, 强化学习的“先知后行”变得无法实施。

为了克服“维数灾难”, 智能这面旗帜被再次举起, 最优控制从以数学推理为主演化为以智能技术为主的学习控制和智能控制。智能控制最初的代表人物是美国普渡大学的 King-Sun Fu 和 George N. Saridis, 后来两人分别成为模式识别和机器人与自动化领域的创始人和早期开拓者之一。受当时人工智能逻辑化、解析化思潮的影响, 智能控制在 30 多年的初创时期主要围绕形式语言、语法分析、决策自动机、图式学习、随机逼近、蒙特卡洛法、最小二乘法、参数识别、自适应算法、自组织系统、迭代学习、强化学习等技术展开, 并被应用于模式识别、机器人与自动化、无人系统、计算机集成制造、金融科技等领域, 但无论是在规模还是在效益方面都没有完全摆脱“维数灾难”的阴影, 其发展到 20 世纪 90 年代中期就陷入瓶颈, 相关工作几乎停滞不前。

“山重水复疑无路, 柳暗花明又一村”。Paul Werbos 在 20 世纪 70 年代中期推出神经网络误差反向传播算法的同时, 就开始研究新优化方法在策略分析中的应用, 并于 20 世纪 80 年代末正式提出近似动态规划(approximate dynamic programming)的思想。同一时期, Saridis 和王飞跃也针对非线性确定系统和随机系统提出了类似的次优控制迭代策略。经过 Wurren B. Powell、Dimitri Panteli Bertsekas、John N. Tsitsiklis 等人的研究, 这一方法进一步与神经网络技术结合, 从近似动态规划发展到神经动态规划(neuro-dynamic programming), 最后发展到目前的自适应动态规划(adaptive dynamic programming, ADP)。自 20 世纪 80 年代中期以来, 王飞跃、刘德荣和魏庆来从不同角度开展 ADP 相关研究工作, 经过十余年的努力, 形成中国科学院自动化研究所复杂系统管理与控制国家重点实验室自适应动态规划团体, 致力于 ADP 方法的进一步发展和应用, 从智能控制的角度推动了强化学习的理论与工程实践。

4 时序差分的再生

尽管 Werbos 在 20 世纪 70 年代末就试图整合

统一试错学习和最优控制的学习方法, 但在相当长的时间里, 基于这两种方法的强化学习几乎各自独立、没有交叉, 直到 20 世纪 80 年代以 Andrew G. Barto、Richard S. Sutton 和 Charles W. Anderson 为核心的学者重新推出时序差分(temporal difference, TD) (TD 也有时间差分的译法, 但本意是暂时差别)的概念和方法, 局势才开始改变, 强化学习渐渐走上“先行后知”与“先知后行”为一体的“知行合一”之途。简言之, 时序差分集现状、回顾、展望的不同需求和分析于一体, 在试错和规划上充分考虑并利用不同时段系统预估与环境反馈之间的差别, 显著地提高了学习和决策的系统性和效率。由此, 强化学习进入了时序差分学习阶段, 理论研究和工程应用的水平都得到了很大的提升。

时序差分学习的理念源自动物学习心理学中与主要强化因子匹配的“次要强化因子(secondary reinforcers)”概念。Minsky 在人工智能之初就认定这一心理学方法对人工学习系统具有重要的意义, 计算机游戏博弈技术的开创者 Arthur Lee Samuel 在其著名的跳棋程序中也采用了时序差分的理念, 使“机器学习”一词成为广为人知的术语。20 世纪 70 年代初, A. Harry Klopff 认识到强化学习与监督学习的本质不同, 强调强化学习内在的趋利(hedonistic)特性, 试图将试错学习与时序差分学习结合, 提出了“局部强化(local reinforcement)”和“广义强化(generalized reinforcement)”等概念, 但与现代的时序差分并不完全相同; 加上 A. Harry Klopff 英年早逝, 其工作不算十分成功。新西兰学者 Ian H. Witten 在其 1976 年的博士论文中第一次明确指出了时序差分学习规则。A. Harry Klopff 的工作对 Andrew G. Barto、Richard S. Sutton 和 Charles W. Anderson 的启发很大, 促使他们在 20 世纪 80 年代初将时序差分学习与试错学习结合, 提出著名的“行动者-评论者框架(actor-critic architecture)”, 时序差分的强化学习由此正式登场。然而, 将时序差分与动态规划和试错方法全部整合在一起是在 20 世纪 80 年代末, 这归功于英国学者 Chris J. Watkins 在其 1989 年的博士论文中提出的 Q 学习(Q-learning)算法。1992 年, IBM 公司的 Gerald Tesauro 利用时序差分构造了多层神经网络 TD-Gammon, 并在古老的西洋双陆棋中战胜了人类世界冠军, 引起广泛关注, 这也使时序差分的强化学习方法广为人知。同年, Chris J. Watkins

和 Peter Dayan 给出 Q 学习算法收敛性的第一个严格证明，更加加深了人们对 Q 学习和强化学习的兴趣。当前，时序差分已从专注预测的 TD (λ) 发展到预估决策控制一体的 SARSA (λ)，Barto 和 Sutton 合著的《强化学习导论 (reinforcement learning: an introduction)》已成为机器学习领域的经典之作。

5 平行强化的体系

基于大规模多层人工神经网络的深度学习的成功，特别是 AlphaGo 和 Pluribus 的巨大影响，使强化学习方法登上了一个更新、更高的层次。然而，随着深度强化学习 (deep reinforcement learning, DRL) 和深度 Q 学习 (deep Q learning, DQL) 等的不断涌现和广泛应用，数据再次成为重大问题，而且图 1 右边所示的强化学习大脑神经科学的部分内容，特别是 Hebb 学习规则的重新评估和计算复杂化与有效性问题，也更加引人注意。

在以试错法为主的先后知强化学习中，因实验周期长、成本高，数据来源受到“经济诅咒”的制约；而在动态规划类的先知后行强化方法中，算法实施又遇“维数灾难”，导致其无效、不可行，TD 强化学习，特别是 TD-Gammon 借助 Self Play 在一定程度上为解决数据生成和算法效率指明了一条道路，而 AlphaGo 和 Pluribus 进一步强化了这条道路的有效性。实际上，这是一条通过虚实平行运作，由“小数据”生成大数据，再与蒙特卡洛法或各类决策树等有效搜索技术结合，从大数据中锤炼出针对具体问题的“小智能”般的精确知识之道。人们应当通过知识图谱和知识范畴 (knowledge categories) 等工具，将这一数据生成和知识制造的过程形式化，并加以软件定义，为强化学习系统组态的生成和实际应用的自动化创造基础。

此外，抽象数字化的强化学习还必须与大脑生物化的功能强化实现平行互联。除了快慢过程的微分 Hebb 学习规则，人们更应关注强化学习与动物的无条件/工具性反射、典型惯性和目标导向行为以及认知图 (cognitive maps) 生成构造等问题的内在关联，并将其应用于针对不同病状的各种机器人辅助和智能康复系统的设计、操作、监控和运维以及与脑和神经相关的疾病的智能诊疗系统中。同时，强化学习机制应成为虚实互动的平行学习和平行

大脑的核心基础，扩展突触可塑、Hedonist 神经元、多巴胺神经元与响应、奖励预估误差机制、神经行动者评估者结构等大脑神经基础构成问题的计算和智能研究手段，使人类生物智能与人工智能的研发更加密切地结合到一起。图 2 给出了虚实互动、实践与理论融合的平行强化学习体系的基本框架，目前流行的数字双胞胎 (又称数字孪生) 是其中的一个重要组成部分。平行强化学习的目的是通过交换世界实现“吃一堑、长一智”：在虚拟的人工世界吃一堑、吃多堑，在现实的自然世界长一智、长多智，以此降低成本，提升效益，克服“经济诅咒”和“维数灾难”，走向智能知行合一的机器强化学习。

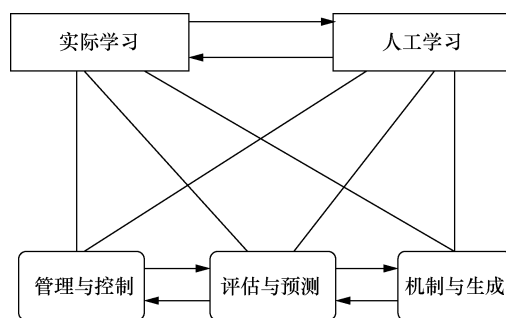


图2 平行强化学习体系的基本框架

6 知行合一的智能

学习是人类获取知识的通用且可靠的途径，这是人类文明有史以来的共识与实践。强化学习是机器具有机器智能的基础和关键手段，这是人工智能研究开创以来的认识与方向，在很大程度上，也是目前从事智能科学与技术研发工作者的共识。然而，要使强化学习真正成为机器学习的核心与智能机制和智能算法的基础与关键技术，从“不知对错、无论好坏”的先后知与先知后行，到知行合一、虚实互动的混合平行智能，仍有许多理论和实践的任务必须完成。

首先，强化学习面临的许多经典问题依然存在，并没有被彻底有效地解决，如“维数灾难”、信用分配、信息不完备、非稳环境、状态行动 Space Tiling、探索与利用的矛盾等，需要更加深入和系统地研究。其次，对于目前被广泛应用的各类深度强化学习方法来说，其引人注目的“超人”表现源自其解决特定问题的特点，但这也正是其难以被推广和普及的问题所在。必须考虑这些深度强化学习

方法构建与应用过程中的形式化和软件定义问题,从而使过程的迁移及其自动化成为可能,完成从特别应用到相对通用的转化。最后,必须引入针对强化学习的软硬件平台,边缘与云端的支撑环境和相应的开源基础设施使强化学习真正应用到生产、商业、交通、健康、服务等领域,使强化学习、强化控制、强化管理、强化医疗、强化经济、强化法律、强化安保等成为一个有效且普适的智能工程项目。

为此,研究者需要从更高更广的角度重新审视强化学习的方法和技术,使其真正成为“人机结合、知行合一、虚实一体”的“合一体”的核心与关键,化智能代理(agents)为知识机器人,深入推动和完善智能社会的知识自动化进程。

参考文献:

- [1] 郑南宁. 人工智能新时代[J]. 智能科学与技术学报, 2019, 1(1): 1-3.
ZHENG N N. The new era of artificial intelligence[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(1): 1-3
- [2] 张钹. 人工智能进入后深度学习时代[J]. 智能科学与技术学报, 2019, 1(1): 4-6.
ZHANG B. Artificial intelligence is entering the post deep-learning era[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(1): 4-6.
- [3] GARIBALDI J M, 陈虹宇, 李小双. 差异与学习: 模糊系统与模糊推理[J]. 智能科学与技术学报, 2019, 1(4): 319-326.
GARIBALDI J M, CHEN H Y, LI X S. Variation and learning: fuzzy system and fuzzy inference[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(4): 319-326.
- [4] LUCAS S, 沈甜雨, 王晓, 等. 基于统计前向规划算法的游戏通用人工智能[J]. 智能科学与技术学报, 2019, 1(3): 219-227.
LUCAS S, SHEN T Y, WANG X, et al. General game AI with statistical forward planning algorithms[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(3): 219-227.
- [5] 王飞跃. 人工智能在多角色游戏中获胜[J]. 中国科学基金, 2020, 34(2): 205-206.

WANG F Y. AI wins in multi-role games[J]. Bulletin of National Natural Science Foundation of China, 2020, 34(2): 205-206.

- [6] LUCIAN B, ROBERT B, BART D S, et al. Reinforcement learning and dynamic programming using function approximators[M]. Boca Raton: CRC Press, 2010.
- [7] 魏庆来, 王飞跃. 强化学习[M]. 北京: 清华大学出版社, 2020.
WEI Q L, WANG F Y. Reinforcement learning[M]. Beijing: Tsinghua University Press, 2020.
- [8] FLORENTIN W, BERND P. Reinforcement learning[M]. Cambridge: MIT Press, 2008.

[作者简介]



王飞跃(1961—), 男, 博士, 中国科学院自动化研究所复杂系统管理与控制国家重点实验室主任, 中国科学院大学中国经济与社会安全研究中心主任, 青岛智能产业技术研究院院长, 主要研究方向为平行系统的方法与应用、社会计算、平行智能以及知识自动化。



曹东璞(1978—), 男, 博士, 滑铁卢大学副教授, 认知自动驾驶实验室主任, 主要研究方向为自动驾驶、人车协同与平行驾驶。



魏庆来(1979—), 男, 博士, 中国科学院自动化研究所研究员, 复杂系统管理与控制国家重点实验室副主任, 中国科学院大学岗位教授, 青岛智能产业技术研究院智能技术创新中心主任, 主要研究方向为自学习控制、平行控制、自适应动态规划、智能控制、最优控制及其工业应用。