# 基于因果建模的强化学习控制: 现状及展望

孙悦雯 [1] 柳文章 [2] 孙长银 [1, 3, 4]

**摘 要** 基于因果建模的强化学习技术在智能控制领域越来越受欢迎. 因果技术可以挖掘控制系统中的结构性因果知识, 并提供了一个可解释的框架, 允许人为对系统进行干预并对反馈进行分析. 量化干预的效果使智能体能够在复杂的情况下 (例如存在混杂因子或非平稳环境) 评估策略的性能, 提升算法的泛化性. 本文旨在探讨基于因果建模的强化学习控制技术 (以下简称因果强化学习) 的最新进展, 阐明其与控制系统各个模块的联系. 首先介绍了强化学习的基本概念和经典算法, 并讨论强化学习算法在变量因果关系解释和迁移场景下策略泛化性方面存在的缺陷. 其次, 回顾了因果理论的研究方向, 主要包括因果效应估计和因果关系发现, 这些内容为解决强化学习的缺陷提供了可行方案. 接下来, 阐释了如何利用因果理论改善强化学习系统的控制与决策, 总结了因果强化学习的四类研究方向及进展, 并整理了实际应用场景. 最后, 对全文进行总结, 指出了因果强化学习的缺点和待解决问题, 并展望了未来的研究方向.

**关键词** 强化学习控制, 因果发现, 因果推理, 迁移学习, 表示学习

**引用格式** 孙悦雯, 柳文章, 孙长银. 基于因果建模的强化学习控制: 现状及展望. 自动化学报, 2023, **49**(3): 661−677

**DOI** 10.16383/j.aas.c220823

## Causality in Reinforcement Learning Control: The State of the Art and Prospects

SUN Yue-Wen[1] LIU Wen-Zhang[2] SUN Chang-Yin[1, 3, 4]

**Abstract** Causality research has shown its potential and advantages in the reinforcement learning community. Beyond the inherent capability of inferring causal structure from data, causality provides an explainable toolset for investigating how a system would react to an intervention. Quantifying the effects of interventions allows actionable decisions to be made while maintaining robustness in the complex system (e.g., in the presence of confounders or under nonstationary environments). This paper explores how causality can be incorporated into different aspects of control systems and introduces recent advances in causal reinforcement learning. First, the concept and algorithms of reinforcement learning are introduced, and two main challenges, e.g., lack of causal explanation of observation variables and hard to transfer in transferable environments, are discussed. Second, the lines of research within causality are reviewed, including causal effect estimation and causal discovery, which provide potential solutions to address the aforementioned challenges. After that, how to embed causality in reinforcement learning systems is introduced. Four kinds of research advances in causal reinforcement learning are summarized and analyzed, followed by real-world applications. Finally, this paper summarizes and presents opening problems and future work prospects.

**Key words** Reinforcement learning control, causal discovery, causal inference, transfer learning, representation learning

**Citation** Sun Yue-Wen, Liu Wen-Zhang, Sun Chang-Yin. Causality in reinforcement learning control: The state of the art and prospects. *Acta Automatica Sinica*, 2023, **49**(3): 661−677

近年来, 人工智能的研究范围不断拓宽, 并在医疗健康、电力系统、智慧交通和机器人控制等多个重要领域取得了卓越的成就. 以强化学习为代表的行为决策和控制技术是人工智能驱动自动化技术的典型代表, 与深度学习相结合构成了机器智能决策的闭环[1]. 强化学习控制是指基于强化学习技术制定控制系统中行动策略的方法. 强化学习的主体, 即智能体, 通过交互的手段从环境中获得反馈, 以试错的方式优化行动策略. 由于擅长处理变量间复杂的非线性关系, 强化学习在面对高维和非结构化数据时展现出了极大的优势. 随着大数据时代的到

收稿日期 2022-10-18 录用日期 2023-02-10
Manuscript received October 18, 2022; accepted February 10, 2023
国家自然科学基金 (62236002, 61921004) 资助
Supported by National Natural Science Foundation of China (62236002, 61921004)
本文责任编委 李鸿一
Recommended by Associate Editor LI Hong-Yi
1. 东南大学自动化学院 南京 210096 2. 安徽大学人工智能学院 合肥 230601 3. 自主无人系统技术教育部工程研究中心 合肥 230601 4. 安徽省无人系统与智能技术工程研究中心 合肥 230601
1. School of Automation, Southeast University, Nanjing 210096 2. School of Artificial Intelligence, Anhui University, Hefei 230601 3. Engineering Research Center of Autonomous Unmanned System Technology, Ministry of Education, Hefei 230601 4. Anhui Unmanned System and Intelligent Technology Engineering Research Center, Hefei 230601

来, 强化学习控制技术快速崛起, 在学术界和产业界获得了广泛关注, 并在博弈[2−5]、电力系统[6−7]、自动驾驶[8−9] 和机器人系统[10] 等领域取得了巨大突破. 在实际系统应用中, 强化学习被广泛应用于路径规划和姿态控制等方面, 并在高层消防无人机路径规划[11] 和多四旋翼无人机姿态控制[12] 等实际任务中取得了良好的控制性能.

尽管如此, 强化学习在处理控制任务时仍面临一些缺陷, 主要体现在以下两个方面. 一是难以在强化学习过程中进行因果推理. 大多数强化学习控制算法是基于采样数据间的相关关系完成对模型的训练, 缺少对变量间因果效应的判断. 而在控制任务中, 任务的泛化和模型的预测通常建立在因果关系之上. 越来越多的证据表明, 只关注相关性而不考虑因果性, 可能会引入虚假相关性, 对控制任务造成灾难性的影响[13]. 二是无法在迁移的场景下保证控制算法的泛化性. 泛化性是指强化学习模型迁移到新环境并做出适应性决策的能力, 要求学习的策略能够在相似却不同的环境中推广. 然而在面临环境改变或者任务迁移时, 智能体收集到的观测数据表现出非平稳性或异构性, 训练数据和测试数据的独立同分布条件受到破坏. 在这种情况下, 强化学习算法常常表现不佳, 无法保证策略的泛化性[14−15], 难以直接推广到更普遍的控制场景.

为了解决上述问题, 目前研究人员尝试在强化学习任务中引入因果理论, 提出了基于因果建模的强化学习控制算法. 因果强化学习的中心任务是在控制问题中建立具有因果理解能力的模型, 揭示系统变量之间的因果关系, 估计数据之间的因果效应, 进一步通过干预和推断, 理解智能体的运行机理. 近年来, 包括 ICLR, NeurIPS, ICML 和 AAAI 在内的人工智能重要国际会议多次设立研讨会, 探索因果理论在机器学习领域的发展和应用[16−19]. 越来越多控制性能优异的因果强化学习算法被陆续提出, 成为最新的研究热点. 建立可解释的因果模型并保证算法的合理决策, 是加速推广强化学习控制算法落地的必要条件, 具有理论意义和应用价值. 本文的主旨是梳理目前因果强化学习的研究现状, 讨论因果理论如何提供变量间因果关系的解释, 帮助改善非平稳或异构环境下的可迁移的决策, 提高数据利用率, 并对未来工作方向提供可借鉴的思路.

本文内容安排如下: 第 1 节介绍强化学习的基本概念和经典算法, 并指出传统强化学习算法的缺陷. 第 2 节介绍因果关系和因果模型的概念, 总结因果效应估计和因果关系发现的研究内容, 为解决强化学习的缺陷提供了可行方案. 第 3 节构建因果强化学习系统的抽象模型, 在此基础上整理出四个研究方向, 综述了因果强化学习的最新研究进展并总结了应用场景. 第 4 节总结全文, 指出了因果强化学习的缺点和待解决的问题, 并对未来的发展趋势进行展望.

## 1 强化学习概述

### 1.1 强化学习的基本概念

强化学习是解决序贯决策问题的重要范式, 其主要框架如图 1 所示. 决策的主体称为智能体, 智能体以试错的方式与环境进行交互, 观测当前环境状态并给出执行动作. 具体地, 在任意一个时间步 $t$, 智能体根据当前所处环境的状态 $S_t$ 采取动作 $A_t$, 并获得下一时刻的状态 $S_{t+1}$ 和实时奖励 $R_{t+1}$. 智能体在不同状态下选择动作的方式被称为策略 $\pi(A|S)$. 强化学习的目标是通过优化策略使得期望累积奖励 $\mathcal{J}(\pi)$ 最大化. 累积奖励定义为 $G(t)=\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}$, 其中 $\gamma\in[0,1]$ 是奖励折扣因子, 用于衡量实时奖励和延迟奖励的权重参数.

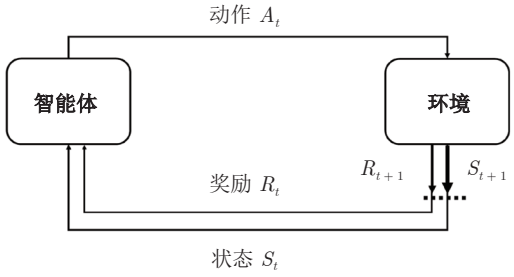



图 1 强化学习框图

Fig. 1 The framework of reinforcement learning

如果智能体可以观测到环境的全部状态, 则称环境是完全可观的, 然而在实际应用中, 状态 $S_t$ 并不一定能包含环境的所有信息. 如果智能体只能观测到环境的局部状态信息, 则称环境是部分可观的. 对于完全可观的环境, 强化学习问题通常可描述为马尔科夫决策过程 (Markov decision process, MDP), 用一个五元组表示为 $\langle\mathcal{S},\mathcal{A},P,R,\gamma\rangle$. 状态空间 $\mathcal{S}$ 和动作空间 $\mathcal{A}$ 分别表示所有状态和所有动作的集合; 对于任意 $\boldsymbol{s}\in\mathcal{S}$ 和 $\boldsymbol{a}\in\mathcal{A}$, 状态转移概率 $P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})$ 表示在状态 $\boldsymbol{s}$ 上执行动作 $\boldsymbol{a}$, 状态 $\boldsymbol{s}$ 转移到状态 $\boldsymbol{s}'$ 的概率. 奖励函数 $R(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')$ 表示在状态 $\boldsymbol{s}$ 上执行动作 $\boldsymbol{a}$, 状态 $\boldsymbol{s}$ 转移到状态 $\boldsymbol{s}'$ 获得的实时奖励. 折扣因子 $\gamma\in[0,1]$ 用于衡量智能体当前动作对后续奖励的累积影响. 对于部分可观的环境, 我们通常使用部分可观马尔科夫决策过程 (Partially observable MDP, POMDP) 描述强化学习问题, 用一个七元组表示为 $\langle\mathcal{S},\mathcal{A},\mathcal{O},P,R,\phi,\gamma\rangle$. 与 MDP

不同, POMDP 假设智能体无法直接观测到环境的潜在状态, 因此动作的选择是基于观测而非状态. 潜在状态空间 $\mathcal{S}$ 表示所有潜在状态的集合; 观测空间 $\mathcal{O}$ 表示所有观测值的集合; $\phi: \mathcal{S} \to \mathcal{O}$ 代表潜在状态到观测空间的映射.

为了分析策略 $\pi$ 的优劣, 研究人员使用两类值函数描述期望累积奖励. 状态值函数 $V_\pi(\boldsymbol{s})$ 指的是从状态 $\boldsymbol{s}$ 出发, 策略 $\pi$ 对应的期望累积奖励, 定义为

$$V_\pi(\boldsymbol{s}) = \mathrm{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}|S_t = \boldsymbol{s}\right] \tag{1}$$

状态动作值函数 $Q_\pi(\boldsymbol{s}, \boldsymbol{a})$ 指的是从状态 $\boldsymbol{s}$ 出发, 执行动作 $\boldsymbol{a}$ 后再使用策略 $\pi$ 的期望累积奖励, 定义为

$$Q_\pi(\boldsymbol{s}, \boldsymbol{a}) = \mathrm{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}|S_t = \boldsymbol{s}, A_t = \boldsymbol{a}\right] \tag{2}$$

为了方便计算, 我们可以利用递归关系推导出状态值函数和状态动作值函数的贝尔曼方程:

$$\begin{aligned} &V_\pi(\boldsymbol{s}) = \mathrm{E}_\pi\left[R_{t+1} + \gamma V_\pi(S_{t+1})|S_t = \boldsymbol{s}\right] \\ &Q_\pi(\boldsymbol{s}, \boldsymbol{a}) = \mathrm{E}_\pi\left[R_{t+1} + \gamma Q_\pi(S_{t+1}, A_{t+1})|S_t = \boldsymbol{s}, \right. \\ &\quad \left. A_t = \boldsymbol{a}\right] \end{aligned} \tag{3}$$

根据值函数, 我们可以定义策略的优劣关系: 对于任意状态 $\boldsymbol{s}$, $\pi \succeq \pi'$ 如果 $V_\pi(\boldsymbol{s}) \geq V_{\pi'}(\boldsymbol{s})$. 那么对于任意 MDP, 存在最优策略 $\pi^*$ 满足 $\pi^* \succeq \pi, \forall\pi$ 成立, 且所有最优策略的状态值函数都等于最优状态值函数 $V^*(\boldsymbol{s}) = \max_\pi V_\pi(\boldsymbol{s})$, 所有最优策略的状态动作值函数也等于最优状态动作值函数, 即 $Q^*(\boldsymbol{s}, \boldsymbol{a}) = \max_\pi Q_\pi(\boldsymbol{s}, \boldsymbol{a})$.

### 1.2　强化学习的经典算法

根据智能体在策略更新中是否用到环境的动力学模型, 强化学习算法可以分为有模型强化学习方法和无模型强化学习方法. 本节从是否利用模型先验知识出发, 对主流的强化学习算法进行梳理, 并将提及的经典算法总结在表 1. 关于强化学习算法的更多内容, 请参见强化学习领域的综述[20−23].

有模型强化学习方法的特点是具有环境的先验知识. 智能体在环境模型上进行规划, 无须与真实环境进行交互便可以优化策略. 因此在相同样本量的前提下, 相对于无模型的方法, 有模型强化学习可以大幅提高数据利用率, 降低采样复杂度. 具体来说, 有模型强化学习方法可以分为两类: 第一类是模型已知的方法, 智能体可以直接利用已知的系统模型和奖励函数进行策略优化. 例如, 在 AlphaZero 中智能体直接利用已知的围棋规则和奖励函数进行策略优化[24]. 在 ExIt 算法中, 智能体利用蒙特卡罗树搜索在棋盘游戏 Hex 中进行策略泛化[25]. 然而在现实情况中, 环境具有复杂性和不可知性, 智能体有时无法直接获得环境的模型, 因此衍生了第二类模型可学习的方法. 智能体通过与环境交互收集原始数据, 并基于观测数据估计系统的前向状态转移模型 $\boldsymbol{s}_{t+1} = f(\boldsymbol{s}_t, \boldsymbol{a}_t)$, 然后进行策略优化. 这类问题的研究重点在于如何学习环境模型. 早在 1980 年代, 利用神经网络拟合环境模型的思想已初现端倪[26−27]. 但是早期的神经网络模型设计较为简单, 难以处理复杂环境下的模型拟合问题. 近年来, 研究人员尝试结合线性回归[28]、高斯回归[29]、随机森林[30]、支持向量回归[31] 和深度神经网络[32−34] 等机器学习方法对模型进行更准确的估计, 其中基于深度学习的深度神经网络由于其良好的特征提取和非线性函数逼近能力, 在模型学习研究中应用最为广泛. 为了减少模型误差, 提高模型的准确性, 概率推理控制 PILCO (Probabilistic inference for learning control)[29] 利用高斯过程学习环境的概率动力学模型, 将模型的不确定性纳入长期规划中. 尽管 PILCO 提升了数据利用率, 但是此类方法需要对模型的分布做出高斯假设, 且计算复杂度较高, 只适用于低维数据. 为了解决高斯回归模型难以推广到高维空间的问题, 后续学者利用近似变分推理的贝叶斯神经网络拟合动态模型, 对 PILCO 进行了拓展, 提出了深度 PILCO 模型[32]. 深度 PILCO 根据贝叶斯公式推理网络权值, 既保留了 PILCO 算法概率模型的优势, 同时计算复杂度更低, 并成功运

表 1　强化学习算法分类及其特点

Table 1　Classification of reinforcement learning algorithms

| 强化学习方法 | 具体分类 | 代表性模型 | 算法特点 |
|---|---|---|---|
| 有模型强化学习 | 模型已知 | AlphaZero[24], ExIt[25] | 状态转移模型已知, 现实场景下不易实现 |
| | 模型可学习: 结构化数据 | PILCO[29] | 数据利用率高, 适用于低维状态空间 |
| | 模型可学习: 非结构化数据 | E2C[33], DSA[34] | 与机器学习相结合, 适用于高维冗余状态空间 |
| 无模型强化学习 | 基于值函数的方法 | SARSA[37], 深度 Q 网络[36, 39] | 采样效率高, 但是无法实现连续控制 |
| | 基于策略梯度的方法 | PG[44], TRPO[45], PPO[46] | 对策略进行更新, 适用于连续或高维动作空间 |
| | 两者结合的方法 | DDPG[47], Actor-Critic[48] | 包含两个网络, 分别更新值函数和策略函数 |

用于更加困难的控制任务. 此外, 以视觉信号为输入的控制任务具有高维性和信息冗余性. 学者们通常利用卷积神经网络[35−36] 处理高维数据, 并利用变分自编码器提取数据的低维特征, 如嵌入控制 E2C (Embed to control)[33] 和深度空间自动编码器 DSA (Deep spatial autoencoders)[34], 提高了算法的数据利用率. 有模型方法的主要缺点是过度依赖建模精度, 难以处理由模型误差造成的性能下降问题. 例如, 在面对高维复杂的状态动作空间, 或者在交互前期数据量较少时, 有模型的方法难以估计出精确的环境模型. 智能体基于不精确的环境模型进行策略优化, 容易导致双重近似误差, 影响控制性能.

在无模型强化学习方法中, 智能体直接与环境进行交互, 以端到端的方式优化策略, 不仅更易于实现, 而且策略具有较好的渐进性能, 适用于大数据背景下的深度网络架构. 根据优化对象的不同, 无模型的强化学习可分为基于值函数的方法, 基于策略梯度的方法, 以及两者结合的方法. 基于值函数的方法在全局范围内进行贪婪搜索并估计状态动作值函数, 以值函数最大化为目标制定策略, 并基于环境反馈更新值函数. 这类方法采样效率相对较高, 值函数估计方差小, 不易陷入局部最优; 缺点是不能处理连续动作空间任务, 且最终的策略通常为确定性策略而非概率分布的形式. 经典算法包括 SARSA (State-action-reward-state-action)[37], Q 学习[38], 深度 Q 网络[36, 39] 及其变体[40−43]. 基于策略梯度的方法直接针对动作策略进行优化, 在策略空间中针对当前策略 $\pi$ 计算累积奖励的梯度值, 以期望累积奖励最大化为目标更新策略. 该类方法直接利用梯度下降优化性能目标 $\mathcal{J}(\pi)$, 或者间接地对 $\mathcal{J}(\pi)$ 的局部近似函数进行优化. 与基于值函数的方法相比, 基于策略梯度的方法相对直观, 算法收敛速度更快, 适用于连续或高维动作空间的场景. 经典算法包括策略梯度法 PG (Policy gradient)[44], 信任域策略优化 TRPO (Trust region policy optimization)[45] 以及近端策略优化 PPO (Proximal policy optimization)[46] 等. 两者结合的方法基于上述两类方法取长补短, 衍生出了执行−评价方法. 评价网络利用基于值函数的方法学习状态动作值函数 $Q$ 或状态值函数 $V$, 减少了样本方差, 提高了采样效率; 执行网络利用基于策略梯度的方法学习策略函数, 使得算法可以推广到连续或高维的动作空间. 经典算法包括深度确定性策略梯度 DDPG (Deep deterministic policy gradient)[47], Actor-critc 算法[48] 及其变体[49]. 无模型强化学习方法最大的缺点是测试任务需要和环境进行大量的交互, 数据利用率低. 在交互代价较高的真实场景中, 由于需要考虑时间消耗、设备损耗和探索过程中的安全性等因素, 无模型的方法难以直接应用到实际场景中.

### 1.3 强化学习的理论困境

虽然强化学习被广泛应用于复杂环境下的控制任务, 但是与人类智能相比, 仍然存在以下两类缺陷. 一是无法提供变量 (尤其是高维和非结构化数据) 间因果关系的解释; 二是在迁移场景下无法确保策略的泛化性和系统的鲁棒性.

可解释性研究主要对系统模型的运作机制进行解释, 通过了解模型每个组分的作用, 进而理解整个模型. 在传统的强化学习场景中, 基于统计的算法模型只能根据观测数据学习到变量间的相关性, 缺少对于变量间因果关系的判断. 值得注意的是, 相关性并不意味着因果性. 如果通过观察发现变量 $X$ 的分布发生变化时, 变量 $Y$ 的分布也会发生变化, 那么可以判定 $X$ 和 $Y$ 之间存在相关性, 但是否存在因果性还需要进一步判断. 举例来说, 气压计的水银柱高度和下雨概率相关, 但是事实是由于气压发生变化同时造成了水银柱高度和下雨概率发生变化, 水银柱高度和下雨概率之间并不存在直接因果关系. 因此利用深度神经网络等统计手段解决强化学习控制问题时, 可能会引发变量间的因果混淆问题. 此外, 缺乏因果标记的观测数据无法将状态和动作联系起来, 使得算法缺乏可解释性, 限制了强化学习在安全敏感领域 (如自动驾驶和医疗诊断) 中的应用. 因此缺乏变量间的因果解释俨然成为阻碍强化学习进一步发展和应用的主要障碍之一.

此外, 由于基于深度神经网络的强化学习模型知其然 (关联性) 而不知其所以然 (因果性), 学习到的策略在非平稳或异构环境等迁移场景中往往缺乏鲁棒性与泛化性. 这里非平稳或异构环境指的是底层数据生成过程会随时间或跨域发生变化的环境[50]. 具体来说, 强化学习算法通常要求采样数据满足独立同分布条件. 算法一般需要在相同的环境评估策略的性能, 同时采样数据通常被人工处理为独立同分布 (如深度 Q 学习中的经验回放池、异步优势 Actor-critic 中的异步采样等技巧), 尽可能地降低样本数据之间的相关性. 否则神经网络的拟合将会出现偏差, 甚至无法稳定收敛. 然而在实际应用中, 观测数据通常是在相对较长的时间段进行采集 (即非平稳性), 或是在不同场景下收集的多领域数据 (即异构性), 因此数据分布会随时间或跨域发生变化. 此时破坏了独立同分布的假设, 强化学习算法性能就会表现得很脆弱[51]. 因此如何在非平稳或异构的场景下确保策略的泛化性与系统的鲁棒

性, 成为当前研究者面临的挑战. 此外, 对泛化性开展研究有利于提高算法的数据利用率, 减少算法对于数据量的高度依赖. 当前强化学习算法性能很大程度上依赖于海量的数据和充分的算力. 然而在大多数实际场景中, 智能体与环境进行大量交互是不可行甚至危险的, 此时采样数据量往往无法满足算法训练的要求, 进而导致控制性能不佳. 因此在非平稳或异构场景下确保控制策略的可迁移性和自适应性, 是加速推广强化学习落地的必要条件, 具有重要的理论意义和应用价值.

## 2　因果理论概述

从古至今, 人类从未停止关于事物间因果关系的思考. 具备因果关系的推理能力被视为人类智能的重要组成部分[52]. 因果关系指的是原因变量和结果变量之间的作用关系. 具体来说, 在不考虑混杂因子[1] 的前提下, 对变量 $X$ 实施适当干预会导致变量 $Y$ 的分布发生变化, 但对 $Y$ 实施干预并不会导致 $X$ 发生变化, 此时可以认为 $X$ 是 $Y$ 的原因变量, $Y$ 是 $X$ 的结果变量.

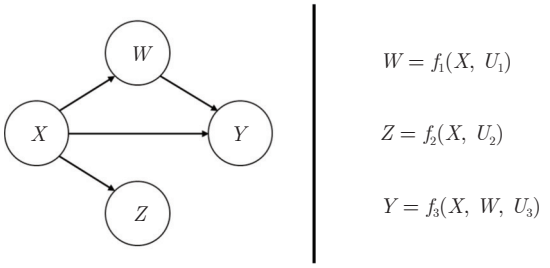
引入因果的概念有利于分析系统中特定个体对于干预的响应. 例如在强化学习领域, 研究人员常常关心结果变量 (状态) 在原因变量 (动作) 发生变化时的效应, 诸如 “采取某种动作, 系统的状态会如何变化”或者 “如果采取某种动作, 累积奖励是否会增加”. 第一类问题称为干预, 即手动将变量 $X$ 设置为某个具体值 $x$, 一般形式化表示为 do 算子 $(X = x)$. 与标准预测问题不同, 干预会导致数据分布发生改变, 有助于分析变量之间的因果关系. 第二类问题称为反事实推理, 即在事件 $X$ 已经出现, 并且事件 $Y$ 发生的前提下, 反过来推理如果事件 $X$ 不出现, 则事件 $Y$ 不发生的概率. 用公式表示为 $P(Y_{X=0} = 0|X = 1, Y = 1)$. 反事实问题致力于推理事件为什么会发生, 想象不同行为的后果, 由此决定采取何种行为来达到期望的结果. 接下来, 我们将从因果分析模型, 因果效应估计和因果关系发现三个方面概述因果理论. 关于因果理论的更多内容, 请参见因果理论的综述[53–57].

### 2.1　因果分析模型

得益于现代统计理论的发展, 因果关系已经从过去哲学层面的模糊定义发展到如今数学语言的精确描述. 当前广泛使用的因果分析模型包括潜在结果框架 (Potential outcome framework) 和结构因果模型 (Structural causal model)[58]. 文献 [55] 指出, 这两种模型在逻辑上是等价的.

1) 潜在结果框架. 潜在结果框架在已知因果结构的基础上, 能够估计治疗变量 (Treatment variable) 对于结果变量的因果效应. 基于潜在结果框架的工作侧重于因果推断, 即通过操纵某个特定变量的值, 观察另一些因果变量的变化. 对于每个样本 $i$, $i = 1, 2, \cdots, n$, 可以观测到治疗变量 $T_i$、特征变量 $X_i$ 和结果变量 $Y_i$. 一般考虑二元治疗变量 $T \in \{0, 1\}$, $T = 1$ 的群体称为试验组, $T = 0$ 的群体称为对照组. 对样本 $i$ 施加治疗 $T = t$ 后, 结果变量存在两个潜在结果 $Y_i^1$ 和 $Y_i^0$. 基于样本的潜在结果, 我们可以定义个体因果效应 $Y_i^1 - Y_i^0$, 即对样本 $i$ 施加与不施加治疗导致结果的差异. 由于个体因果效应是不可识别的, 研究人员通常针对总体识别平均因果效应, 可表示为 $\mathrm{E}_i[Y_i^1 - Y_i^0] = (1/n)\sum_{i=1}^{n}(Y_i^1 - Y_i^0)$.

2) 结构因果模型. 结构因果模型通常用于描述变量之间的因果机制, 侧重于寻找变量之间的因果结构, 进行因果关系识别. 结构因果模型由两部分组成: 因果图结构 (一般是有向无环图) 和结构方程模型. 有向无环图 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (如图 2(a) 所示) 是描述变量间因果关系的有向图, 以直观的方式嵌入变量因果关系, 其中节点集 $\mathcal{V}$ 代表随机变量, 边集 $\mathcal{E}$ 代表因果关系, 例如 $X \to Z$ 表示 $X$ 对 $Z$ 有直接因果影响. 结构方程模型 (如图 2(b) 所示) 用于定量地描述因果关系. 不同于普通的方程模型, 结构方程模型可以表示变量生成过程, 因此具有非对称性. 令 $n$ 个随机变量 $X_1, \cdots, X_n$ 为有向无环图的顶点, 每个变量 $X_i$ 都满足方程 $X_i \doteq f_i(\mathrm{Pa}(X_i), U_i)$, 其中 $f_i$ 为非参数函数, $\mathrm{Pa}(X_i)$ 表示 $X_i$ 的父辈变量, $U_i$ 为独立于父辈变量的随机噪声. 给定有向无环图以及结构方程模型, 我们可以描述由有向边表示的因果关系.




(a) 有向无环图　(b) 结构方程模型
(a) Directed acyclic graph　(b) Structural equation model

图 2　结构因果模型及其组成部分
Fig. 2　Structural causal model

### 2.2　因果效应估计

给定 $n$ 组数据集 $[(X_1, T_1, Y_1), \cdots, (X_n, T_n, Y_n)]$,

[1] 混杂因子指的是系统中两个变量未观测到的直接原因.

其中 $X$ 是特征变量, $T$ 是治疗变量, $Y$ 是结果变量. 因果效应估计用于量化改变 $T$ 后 $Y$ 发生改变的大小. 传统上, 随机化试验[59] 是因果效应估计的黄金标准. 在保持其他变量不变的情况下, 通过有目的地操纵感兴趣的变量, 可以观察到实验结果在统计上的显著影响. 但是在实际应用中, 受限于伦理等因素的影响, 随机化试验的成本太高, 通常不允许进行大范围实验, 甚至有时不可能实施, 因此从观测数据中估计因果效应显得至关重要.

因果效应估计可以大致分为干预类估计和反事实推理两类. 研究人员通常使用 do 算子评估变量 $X$ 对变量 $Y$ 的因果影响, 并根据系统中是否存在混杂因子, 采取不同的思路进行因果效应估计, 这类方法统称为干预类估计. 干预类估计方法通过对某一变量进行强制 do 操作, 观察其他变量的变化, 从而进行因果效应估计. 由于干预类估计方法的操作对象往往是群体数据, 因此难以使用 do 算子推理个体数据的假设分布. 此外, 在数据分析中有时需要对事实进行反转假设, 思考如何改进结果. 此时需要利用反事实推理, 从个体层面根据已发生的事件对未发生事件进行估计.

1) 系统中不存在混杂因子的干预类估计. 基于潜在结果框架的平均因果效应是针对数据总体进行评估. 当试验组和对照组之间存在数据异构时, 需要对样本采取适当的调整措施. 回归调整法[60] 通过训练有监督的回归模型, 对 $P(Y|T, X)$ 作出评估. 此时平均因果效应可表示为 $(1/n)\sum_{i=1}^{n}(P(Y_i|T=1, X_i)-P(Y_i|T=0, X_i))$. 该方法可以直接比较试验组和对照组观察结果的差异, 但是受限于精确匹配的严苛假设. 倾向得分方法[61] 可以基于倾向分数对试验组和对照组的个体进行匹配. 倾向分数 $e(x)=P(T=1|X)$ 能够反映出样本 $X$ 选择某种干预 $T$ 的可能性. 该类方法首先估计每个个体的倾向分数 $\hat{e}(x)$, 并基于分数进行分组, 然后估计每组的平均因果效应, 最后进行加权平均, 适用于试验组和对照组的个体足够且维度较高的情况.

2) 系统中存在混杂因子的干预类估计. 系统中常常存在的混杂因子可能导致选择偏倚 (Selection bias) 问题. 例如在因果关系模型 $Z\rightarrow X$, $X\rightarrow Y$, $Z\rightarrow Y$ 中, 未被观测的混杂因子 $Z$ 会导致 $P(Y|X)$ 不同于 $P(Y|\mathrm{do}(X))$, 因此研究人员通常需要借助额外假设辅助因果效应估计. 一个经典的方法是后门调整[62]. 当系统中存在可观测的混杂因子时, 通过对满足后门准则的变量集 $Z$ 做调整, 可以消除所有与 $X$ 对 $Y$ 的因果效应相关的混杂偏差. 具体地, 如果存在集合 $Z$ 相对于 $(X, Y)$ 满足后门准则, 那么 $X$ 对 $Y$ 的因果效应是可识别的, 且 $P(Y=y|\mathrm{do}(X=x))=\sum_z P(Y=y|X=x, Z=z)P(Z=z)$. 当系统中存在未能观测的混杂因子时, 在一定情形下可以采取前门调整[62] 估计因果效应. 如果存在集合 $Z$ 相对于有序对 $(X, Y)$ 满足前门准则且 $P(x, z)>0$, 则 $X$ 对 $Y$ 的因果效应是可识别的, 且 $P(Y=y|\mathrm{do}(X=x))=\sum_z P(Z=z|X=x)\sum_{x'} P(Y=y|X=x', Z=z)P(X=x')$.

3) 反事实推理. 反事实推理也是因果效应估计的重要组成部分. 给定结构因果模型 $Y=f(X, U)$, 其中变量 $X$ 是变量 $Y$ 的父辈变量, $U$ 是噪声项, 已有的观测数据统称为证据 $E$, 假设反事实变量为 $X=x'$. 一般来说, 反事实推理分为三个步骤[63]: 1) 溯源: 利用证据 $E=e$ 更新 $U$ 的值. 这一步在考虑证据 $E$ 的情况下解释特定的 $U$. 2) 干预: 修改原有的模型 $M$. 用 $X=x'$ 替换结构方程, 获得修改后的模型 $M_{x'}$. 3) 预测: 使用替换的模型 $M_{x'}$ 和 $U$ 重新计算 $Y$ 的值, 得到反事实的结果.

### 2.3 因果关系发现

因果效应估计的前提是假设因果关系已知, 然而在现实世界中因果关系往往不是先验知识. 因此根据观测数据揭示变量因果结构的因果发现近年来备受关注. 传统的因果发现算法常常基于以下四个假设[64]:

1) 无环性假设: 变量间的因果结构可以用有向无环图 $\mathcal{G}=(\mathcal{V}, \mathcal{E})$ 表示.

2) 因果充分性假设: 在有向无环图中, 任何一对节点都没有共同的外因, 即变量集 $\mathcal{V}$ 中的任意两个变量的直接原因变量都存在 $\mathcal{V}$ 中.

3) 因果马尔科夫假设: 有向无环图满足因果马尔科夫条件, 当且仅当对于任意一个节点 $v\in\mathcal{V}$, 在给定 $\mathrm{Pa}(v)$ 时, $v$ 独立于 $\mathcal{V}\backslash(\mathrm{De}(v)\cup\mathrm{Pa}(v))$, 其中 $\mathrm{Pa}(v)$ 指的是变量 $v$ 的父辈变量, $\mathrm{De}(v)$ 指的是变量 $v$ 的子代变量.

4) 因果忠诚性假设: 在有向无环图中, 已知变量集 $\mathcal{V}$ 中随机变量的联合概率分布为 $P$. 当 $\mathcal{G}$ 包含 $P$ 的所有条件独立性时, $(\mathcal{G}, P)$ 满足因果忠诚性条件. 换句话说, 分布中的所有条件独立性都可以用 $\mathcal{G}$ 表示, 不存在 $\mathcal{G}$ 所隐含的条件独立关系之外的独立性.

尽管上述假设有助于缩小因果图的范围, 但是如果直接遍历所有可能存在的因果图计算复杂度太高. 因此研究人员通常需要借助额外的约束实现因果发现. 根据约束条件不同, 因果发现算法可分为基于条件约束的方法, 基于分数的方法和基于函数因果模型的方法, 经典算法的归纳如表 2 所示.

表 2　因果理论研究内容

Table 2　Classification of causality research

| 研究内容 | 具体分类 | 代表算法 | 算法特点 |
|---|---|---|---|
| 因果效应估计 | 没有混杂因子的干预类估计 | 回归调整[60], 倾向得分方法[61] | 对样本采取适当的调整措施 |
| | 存在混杂因子的干预类估计 | 前门调整[62], 后门调整[62] | 借助额外的假设进行估计 |
| | 反事实推理 | 标准三步骤[63] | 回答反事实问题 |
| 因果关系发现 | 基于条件约束的方法 | PC[64], FCI[67] | 基于条件独立性假设 |
| | 基于分数的方法 | GES[70], FGES[71] | 基于评分标准对因果图打分 |
| | 基于函数因果模型的方法 | LiNGAM[74], ANM[75−76], PNL[77−78] | 需要对函数类型作出假设 |

#### 2.3.1　基于条件约束的方法

基于条件约束的方法利用条件独立性重构数据中的因果信息. 已有的研究表明, 在满足因果马尔科夫假设和因果忠诚性假设的情况下, 可以在因果图结构和统计独立性之间建立对应关系. 因此可以通过判定观测变量之间的条件独立性来学习因果结构. 在满足独立同分布采样和因果充分性假设的前提下, PC (Peter-Clark) 算法及其变体[64−66] 利用一个标准的统计决策程序找到变量之间的因果关系. 研究表明, 如果在满足大量样本的前提下条件独立测试是正确的, 则 PC 算法可以收敛到马尔科夫等价类[2]. 当系统中存在混杂因子时, 快速因果推理 FCI (Fast causal inference) 及其变体[67−68] 也可以得出渐近正确的结论. 基于条件约束的方法具有普适性, 可以处理多种类型的数据分布和因果关系, 然而这类方法面临以下几个缺点: 1) 当系统中存在很多变量时, 因果忠诚性假设可能是一个强假设. 一旦该假设在某些情况下不成立, 算法得出的因果图可能是错误的. 2) 这些算法依赖于可靠的条件独立检验, 而在非线性情形或连续和离散变量同时存在的情形, 条件独立检验往往需要大量的样本基数, 否则将存在较大的计算误差[69]. 3) 条件独立测试可能会导出无法分辨方向的马尔科夫等价类, 特别是在二变量情况下, 由于没有条件独立关系可用, 基于条件约束的方法无助于确定其因果方向.

#### 2.3.2　基于分数的方法

为了放宽因果忠诚性假设, 在满足因果充分性假设的前提下, 基于分数的方法可以在有向无环图空间进行贪心启发式搜索, 通过最优化因果图分数寻找和数据最匹配的图结构. 具体来说, 给定观测数据 $X$, 这类方法基于恰当的评分标准 $S(X, \mathcal{G})$ (如贝叶斯信息准则) 对可能的因果图 $\mathcal{G}$ 打分, 并通过优化因果图分数实现结构优化. 经典算法主要包括贪婪等价搜索 GES (Greedy equivalence search)[70] 及其变体[71−72]. GES 从一个空图开始, 以迭代的方式添加有向边直到收敛, 然后再消除不必要的边直到收敛, 目标是在搜索空间中找出最优图结构 $g^*$ 满足 $g^* = \arg\max_{g\in\mathcal{G}} S(X, g)$. 在有限样本的情况下, GES 计算复杂度较高. 快速贪婪等价搜索 FGES (Fast greedy equivalence search)[71] 基于 GES 采取适当的修改, 能够在高维数据集中搜索因果关系. 此外, 一般的 GES 类算法无法处理系统中存在混杂因子的情况. 贪婪快速因果推理 GFCI (Greedy fast causal inference)[73] 将 GES 和 FCI 方法组合起来, 使用 GES 找到因果结构, 使用 FCI 调整结构并找到因果方向, 适用于连续变量和混杂因子场景.

#### 2.3.3　基于函数因果模型的方法

因果关系的一个基本属性是不对称性, 即 $X$ 可能导致 $Y$, 但 $Y$ 可能不会导致 $X$. 基于函数因果模型的方法利用不对称性的特征, 根据噪声项的独立性判断因果方向. 具体来说, 假定 $Y$ 是结果变量, $X$ 是原因变量. 那么用 $Y$ 对 $X$ 做回归, 得到的噪声项和 $X$ 是独立的; 但反过来用 $X$ 对 $Y$ 做回归, 得到的噪声项和 $Y$ 是不独立的. 在表达方式上, 函数因果模型将结果 $Y$ 表示为直接原因 $X$ 和噪声项 $U$ 的函数: $Y = f(X, U; \theta)$, 其中 $U$ 是独立于 $X$ 的噪声项, 函数 $f \in \boldsymbol{F}$ 解释了 $X$ 和 $Y$ 之间的因果生成机制, $\theta$ 是 $f$ 的参数. 通过合理地限制 $f$ 的函数空间, 研究人员可以利用回归发现非对称的独立性, 从而判定因果方向. 有研究表明, 正确定义的函数因果模型能够区分同一马尔科夫等价类中不同的有向无环图. 考虑到不同的参数模型, 研究人员提出了多种基于函数因果模型的方法. Shimizu 等[74] 提出了线性非高斯无环模型 LiNGAM (Linear non-gaussian acyclic model): $Y = f(X) + U$, 其中 $f$ 是线性函数, 并且至多一个噪声项 $U$ 和原因变量 $X$ 服从高斯分布. 在无环性、因果马尔科夫和线性非高斯的假设下, LiNGAM 可以唯一确定变量的因果结构. 由于 LiNGAM 模型受限于线性系统, 后续学者在此算法基础上进行了一些拓展和改进. 针对非线性系统, Hoyer 等[75−76] 提出了加性噪声模型 ANM (Additive noise model): $Y = f(X) + U$, 其中 $f$ 表

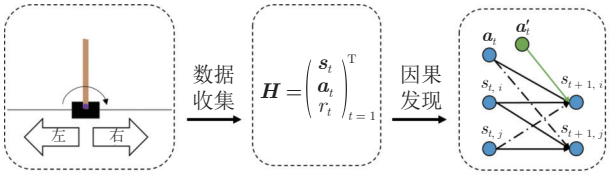
[2] 马尔科夫等价类指的是满足相同条件独立性的一组因果结构.

示非线性函数, $U$ 表示独立噪声. 在因果马尔科夫和因果充分性的假设下, ANM 模型可以唯一识别真正的因果结构. 推广到更一般的情况下, Zhang 等[77−78] 提出了后非线性模型 PNL (Post-nonlinear): $Y = g(f(X) + U)$, 其中 $f$ 和 $g$ 是非线性函数, 并且假定非线性变换 $g$ 是可逆的. PNL 模型综合考虑了因变量的非线性影响, 噪声影响以及观测变量中可能的传感器或测量失真, 因此更具有一般性.

## 3 因果强化学习控制

在本节中, 我们将详细阐述如何利用因果关系改善强化学习控制与决策. 强化学习的目标是最大化期望累积奖励, 智能体本身不具备因果推理的能力. 如第 1.3 节所述, 现有的强化学习算法存在两类缺陷. 幸运的是, 这两类缺陷恰好可以通过引入因果关系来解决. 与一般的强化学习控制不同, 因果强化学习可以区分系统变量之间的虚假相关性和因果关系. 接下来以倒立摆系统为例, 说明如何将因果分析融入强化学习系统, 辨别虚假相关性. 倒立摆系统是强化学习领域的基准测试环境. 智能体对小车施加动作 $\boldsymbol{a}$, 令其沿着无摩擦水平轨道左右移动, 控制目标是防止车上的杆跌落. 因此在杆保持直立的每个时刻, 智能体获得奖励 $r = +1$. 状态 $\boldsymbol{s}$ 分别为小车位置、小车速度、杆与车之间的角度和角速度. 因果强化学习的处理流程如下: 首先从控制任务 (如图 3(a) 所示) 中采样观测数据, 具体包括状态变量、动作变量和奖励变量 (如图 3(b) 所示); 然后利用因果理论, 从观测数据中提取高层的因果特征, 并将其形式化表示成一个能够反映数据生成过程的因果结构 (如图 3(c) 所示). 通过这种方式直观地展示虚假关系 (虚线) 和因果关系 (实线), 降低冗余信息的影响, 提高数据利用率. 此外, 强化学习数据采集过程中常常存在选择偏倚问题. 在控制系统中引入因果分析有助于理解偏倚, 并利用 do 算子实现对干预效果的形式化推理. 例如, 在图 3(c) 中对动作 $\boldsymbol{a}_t$ 进行干预 $\boldsymbol{a}'_t \leftarrow \boldsymbol{a}_t$ (如绿线所示) 只会影响因果图中的子代变量 $s_{t+1,\,i}$, 而对其他非因果变量 $s_{t+1,\,j}$ 不产生影响, 从而可以进行有针对的干预.

与一般机器学习算法不同, 在强化学习中, 智能体不仅能够观测环境, 还可以用行动 (或干预) 塑造环境. 因此与其他机器学习应用场景相比, 强化学习更易于融合因果理论. 目前, 因果理论在强化学习领域的研究已初现端倪. 因果强化学习的基本任务是将因果建模的思想融入强化学习过程中, 旨在解决强化学习的可解释性问题和泛化性问题, 提高数据利用率. 研究的关键问题是如何利用因果知识显式地提取系统的结构不变性, 同时提升控制性能.




(a) 控制任务 (a) Control task (b) 底层数据 (b) Observation data (c) 因果结构 (c) Causal structure

图 3 在倒立摆系统中提取系统变量之间的因果关系

Fig. 3 Causal representation in cart pole system

### 3.1 学习算法的结构

根据已有的研究成果, 我们在图 4 中展示了如何将因果技术集成到强化学习控制系统中, 并将因果强化学习的研究方向分为两大类: 1) 利用因果发现构建因果模型, 即给定观测数据 (尤其是高维和非结构化数据), 提取系统的低维因果特征和因果关系, 搭建系统的因果模型; 2) 利用因果推理实现策略优化, 即给定因果模型, 分析系统对干预将作出何种反应并进行策略规划. 现有因果强化学习算法总结在表 3 中.

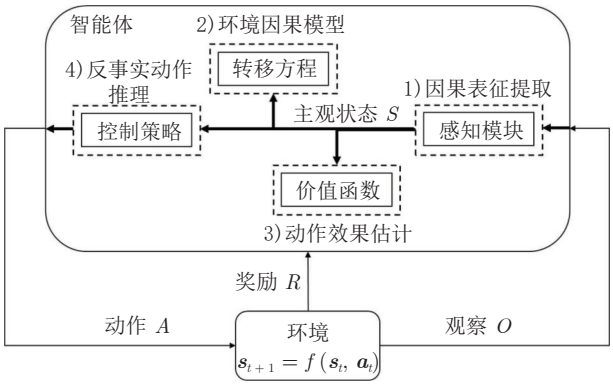



图 4 因果技术在强化学习控制系统各环节的应用

Fig. 4 The application of causality in reinforcement learning control system

1) 利用因果发现构建因果模型. 如上文所述, 虚假的相关性和选择偏倚会导致模型预测不可靠或不公平. 这类研究将因果发现的技术引入强化学习领域, 基于历史数据推断状态、动作和奖励信号之间的因果关系, 去除变量之间的虚假关联, 指导构建因果模型. 基于有模型的强化学习框架, 提取系统中的结构不变性, 并去除模型中和因果链条无关的参数. 这样不仅简化了模型, 同时能够识别出不同任务间发生改变的因果链条有哪些, 从而让模型

表 3　因果强化学习算法总结

Table 3　The classification of causal reinforcement learning algorithms

| 研究内容 | 代表算法 | 解决问题 |
|---|---|---|
| 因果表征提取 | ASR[83], CCPM[84], MABUC[88], B-kl-UCB[89] | 对高维冗余的原始数据进行因果结构化表征 |
| 环境因果模型 | AdaRL[95], CCRL[97], IAEM[98], OREO[102] | 在非平稳或异构环境中构建可迁移的环境因果模型 |
| 动作效果估计 | CEHRL[103], SDCI[104], 倾向性评分[109], FCB[110] | 量化智能体动作对于环境的影响, 获得数据的无偏估计 |
| 反事实动作推理 | CF-GPS[111], 反事实数据增强[81] | 提高算法的样本效率和可解释性 |

能够更容易地迁移到新的任务, 提升模型的可解释性和鲁棒性. 该方法具体可以分为以下两类.

a) 因果表征提取. 人类擅长构造复杂世界的简单蓝图, 对事物的理解往往是基于正确的因果结构并能自动忽略不相关的细节. 越来越多的证据表明, 使用恰当的结构化表征对于理解系统内部因果关系很有帮助. 良好的表征可以帮助智能体对其周围环境进行简洁的建模, 进而支持复杂环境下的有效决策. 以像素化游戏为例, 智能体可以根据对动作的共同反应, 对像素进行分组, 从而识别物体. 此时物体可视为允许单独干预或操控的模块化结构. 因果表征提取是基于独立因果机制的结构化生成方法, 其目的是将环境相关的原始观测数据转化为因果模型的结构化变量. 关键问题在于如何从原始数据中抽取高级因果变量, 这不仅关乎系统的感知能力, 还涉及智能体与环境的交互方式.

b) 建立可迁移的环境因果模型. 因果模型允许将环境建模为一组潜在的独立因果机制. 在此情况下, 如果数据分布发生变化, 并非所有机制都需要重新学习. 此类方法致力于从数据中学习合理的因果模型, 构建对分布变化具有鲁棒性的预测因子[79], 并找到一种恰当的方式将知识分解为能够匹配微小变化的组件和机制. 因果模型不仅能帮助智能体更好地实现迁移学习, 还能启发包括因果机制变化检测、因果骨架估计、因果方向识别和非平稳机制估计的框架设计在内的多个研究领域[50, 80]. 首先, 为了对强化学习环境进行因果建模, 智能体通过干预观察变量间的因果影响, 进而发现因果结构. 此外, 不同于传统的强化学习通常假设系统动态遵循固定的概率分布, 此类方法提供了处理非稳态和异构分布的解决方案. 假定系统的潜在因果结构是固定的, 但是与因果结构相关的机制或参数可能会随任务或时间发生变化. 该类方法能够将非稳态因素导致的分布变化转换成训练信号, 学习系统的不变性结构, 并基于独立因果机制将知识分解为能够匹配变化的组件, 显式地展示哪些部分发生了变化, 遵从什么样的规则在变, 以端到端的方式在任务间实现快速迁移. 智能体可以有针对性地重新训练模型中因果关系改变的部分, 降低了采样需求和模型复杂度.

2) 利用因果推理实现策略优化. 强化学习的目标是生成最优策略, 构建因果模型只能展示变量间的因果关系, 还需要引入因果推理才能实现策略优化. 因果推理作为一种校正偏见的手段, 通过采取不同的策略 (干预) 观察环境状态的变化, 使任务中的规则更加清晰, 帮助智能体更高效地学习值函数或探索策略. 此外, 在因果结构已知的前提下, 不需要或只需要很少的实验就可以回答大量的干预性问题和反事实性问题. 因此因果推理可以大幅减少算法对数据的依赖, 提高数据利用率. 具体来说, 我们可以通过在线学习 (真正实施干预) 和离线学习 (想象中干预) 两种模式进行策略改进, 主要分为以下两类.

a) 动作效果估计. 人类可以通过干预获得因果启示. 例如在倒立摆实验中, 在杆左倾的情况下向左移动小车, 可以维持杆的直立状态; 向右移动则会导致杆失衡. 动作效果估计旨在对动作变量进行干预, 观察智能体的行为对环境的影响. 通过对动作进行恰当的规划, 智能体可以观察到干预导致的联合分布变化. 此外, 智能体还可以推断不同动作带来的效果, 进而了解何时或何种行为对状态能够产生何种影响, 有效地指导策略优化.

b) 反事实动作推理. 在强化学习控制领域中, 拥有反事实推理的能力对于实验成本高昂或存在安全隐患的任务至关重要. 在已有观测数据的前提下, 反事实动作推理旨在推断出采取不同的动作导致的结果. 智能体可以通过制定假想策略, 在想象空间中进行反事实干预, 验证干预效果, 进而不断优化策略[63]. 此外, 反事实动作推理可以在想象空间中产生新的数据, 智能体可以充分利用可用信息 (包括观测数据和反事实数据) 进行推理, 从而提高算法的数据利用率[81].

## 3.2　利用因果发现构建因果模型

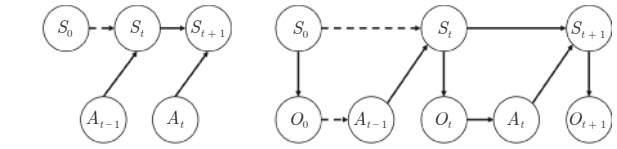
### 3.2.1　因果表征提取

在强化学习控制系统中, 系统的输入状态可能是高维或非结构化数据. 因此引入恰当的结构化表征可以对冗余的原始数据进行信息提取, 有助于解决强化学习的可解释性问题. 总的来说, 和强化学

习控制相关的因果表征提取主要分为基于 POMDP 的表征提取和存在混杂因子的表征提取.

基于 POMDP 的表征提取通常假设观测数据 $O$ (通常是高维或非结构化数据, 如像素输入) 由潜在状态 $S$ 生成, 智能体根据策略 $\pi(A|O)$ 采取行动, 通过与环境交互获得观测数据, 并基于观测数据恢复潜在状态. 与显式的 MDP 不同 (如图 5(a) 所示), 基于 POMDP 的表征提取 (如图 5(b) 所示) 的关键问题在于如何找到 $O \to S$ 的映射, 并根据过去的动作 $A_{\leq t}$ 和过去的潜在状态 $S_{\leq t}$ 预测未来的潜在状态 $S_{>t}$, 学习底层因果图结构. Yao 等[82] 指出潜在时序因果状态在一定场景下是可识别的, 该研究为基于 POMDP 的表征提取提供了理论保证. 该类方法的代表性工作包括动作充分状态表示 ASR (Action-sufficient state representation)[83] 和因果正确部分模型 CCPM (Causally correct partial models)[84]. ASR 以最大化累积奖励为目标, 基于变量结构关系建立环境生成模型, 以因果结构为约束提取出足够决策的最小状态表示集. 在 ASR 的框架下, 策略学习与表征学习可以分开进行, 且策略函数只依赖于低维状态表征, 从而提高了样本利用率, 缺点是没有扩展到可迁移的场景下. 为了在策略发生变化的情况下对模型进行修正, 解决部分模型中因果不正确的问题, CCPM 结合概率模型和因果推理, 提出了因果正确的部分可观模型, 提高了模型的鲁棒性. 此外, 部分研究人员致力于将因果技术和 POMDP 融入一个框架内进行分析. Sontakke 等[85] 引入了因果好奇心 (Causal curiosity) 作为内在奖励, 鼓励智能体在探索性交互时, 通过自监督的方式发现环境中变化的因果机制. Gasse 等[86] 通过引入 do 算子, 将有模型的强化学习表示为因果推理问题, 并且使用观测数据和干预数据共同推断 POMDP 的状态转移方程. 由于假设观测空间要小于离散状态空间, 因此该方法的缺点是只能处理维数较低的观测空间. 为了解决高维观测空间问题, Zhang 等[87] 利用循环神经网络从观测数据中学习近似的因果状态表示, 并在 Lipschitz 假设下为该表示连续版本的最优性提供了理论保证.

存在混杂因子的表征提取方法则考虑更一般的实际场景, 假设系统中存在未能直接观测到的混杂因子. 此时, 系统的状态转移模型和奖励模型将会受到影响, 阻碍行为策略的有效学习. 以自动驾驶场景为例, 智能体从不同场景中收集的离线数据可能依赖于某些未被观测的因素 (如交通的复杂度或道路设计的合理性). 当训练场景为行人过马路时, 智能体可能会从观测中错误地推断出 “只要踩下刹车, 就会有行人出现在汽车前面”这种结论, 从而引入虚假相关性. 这种由混杂因子导致的虚假相关性, 使得观测数据无法提供有效信息, 甚至会误导因果效应识别. 在此情况下, POMDP 模型可能会被未能准确描述的观测数据迷惑, 推导出错误的因果模型, 进而导致不正确的策略规划. 因此存在混杂因子的表征提取方法的关键问题在于去除或估计混杂因子, 以减少虚假相关性对后续因果模型推导的影响. 在混杂因子建模上, 早期的工作包括存在未观测混杂因子的多臂老虎机问题 MABUC (Multi-armed bandit problem with unobserved confounders)[88] 和 Kullback-Leibler 置信上限 B-kl-UCB (B-Kullback-Leibler upper confidence bounds)[89]. MABUC 通过引入结构因果模型, 将具有混杂因子的多臂老虎机问题表示为因果推理问题. MABUC 首次将混杂因子和强化学习融入一个框架之中进行分析, 缺点是模型需要在线学习, 而且没有考虑知识迁移的场景. 在 MABUC 的框架下, B-kl-UCB 利用结构知识推导智能体分布的界限, 将工作拓展到离线且可迁移的场景下. 在混杂因子去除方面, Lu 等[90] 提出了去混杂强化学习框架, 使用自动变分编码器估计潜在变量模型, 发现隐藏的混杂因子并推断因果效应. 尽管该框架允许嵌入强化学习算法进行策略更新, 缺点是要求每一个混杂因子都需要体现在潜在变量模型中, 且无法给出明确的遗憾值[3]. 为了在有限遗憾值内识别最优治疗方案, Zhang 等[91] 在观测数据存在混杂因子的情况下, 利用结构因果模型和独立性约束, 降低候选策略空间的维度, 简化问题的复杂度, 缺点是模型需要在线学习. 为了充分利用离线数据提高样本效率, Wang 等[92] 提出了一种去混杂的最优值迭代方法, 综合考虑了部分可观的混杂因子和完全不可观的混杂因子两种情况, 通过后门准则和前门准则显式地调整观测数据中的混杂偏差, 并且提供了遗憾值的表达.




(a) MDP 数据生成过程 (a) Data generation process of MDP

(b) POMDP 数据生成过程 (b) Data generation process of POMDP

图 5 MDP 和 POMDP 的数据生成过程

Fig. 5 Data generation process in MDP and POMDP

### 3.2.2 建立可迁移的环境因果模型

强化学习问题通常假定训练样本和测试样本满

[3] 遗憾值指的是实际算法的累计损失和理性算法的最小损失之间的差值.

足独立同分布的条件. 在面临环境改变或者任务迁移时, 独立同分布的假设受到破坏, 在特定领域学习的最优策略无法推广到其他领域, 导致控制性能下降[93]. 因此在面对非平稳或异构环境时, 智能体不仅需要识别变化, 更需要适应这些变化. 幸运的是, 非平稳或异构数据已被证明有助于识别因果特征. 可迁移的环境因果模型致力于实现可靠、低成本、可解释的模型迁移, 关键问题在于提取正确的知识表示, 找出哪些因素发生了变化、在哪里变化、如何变化, 使得在源域训练的模型能够在新的场景下实现快速迁移. 可迁移的环境因果模型通常基于以下三个原则[94]: 1) 独立因果机制. 系统变量的因果生成过程由互不影响的独立模块组成. 给定原因, 每个变量的条件分布不会影响其他模块. 2) 最小变化原则[50] 或稀疏机制迁移. 细微的分布变化往往以稀疏或局部的方式在因果分解 $P(X_1, \cdots, X_n) = \sum_{i=1}^{n} P(X_i|\mathrm{Pa}(X_i))$ 中表现出来. 当数据分布发生变化时, 仅有少量的模块和参数需要改变. 3) 相似因果动态. 该假设允许训练数据和测试数据来自不同的分布, 但涉及 (大致) 相同的因果状态转移方程. 例如在机器人导航过程中, 房间内的光照条件可能会发生变化, 但环境的动力学模型仍然是相同的. 基于以上原则, 目前有三种研究方向用于构建可迁移的环境因果模型.

1) 利用结构因果模型编码变化模块. 根据最小变化原则或稀疏机制迁移原则, 当因果模型得到恰当表示时, 仅需要更新少量的模块和参数就可以实现分布迁移, 进而提高策略的鲁棒性. 在强化学习框架下, 结构因果模型不仅能够表征变量之间的结构关系, 还可以显式地编码跨域的变化模块. 因此这类问题的研究重点在于如何编码最小的可迁移模块. 在 MDP 的框架下, Sun 等[79] 基于结构因果模型拟合环境的动态转移方程, 并将跨域变化的模块集成为一个外生变量 $\lambda$, 通过更新 $\lambda$ 实现环境分布的迁移. 在 POMDP 的框架下, Huang 等[95] 提出一种自适应强化学习算法 AdaRL (Adaptive reinforcement learning). AdaRL 利用图模型实现最小状态表征, 包括特定域的变化因素和共享域的状态表示, 同时对状态动态、观察函数和奖励函数的变化进行建模, 利用因子分解提高数据利用率, 只需要来自目标域的少量样本就可以实现稳健有效的策略迁移.

2) 寻找因果不变性. 得益于独立因果机制, 我们可以将数据生成过程视为一些独立模块, 通过寻找因果不变性发现因果结构. 在这种情况下, 研究人员通常需要基于相似动态的多个环境挖掘环境的潜在结构, 进而实现良好的泛化. 对于观测分布不同但是潜在因果结构相同的环境族, Zhang 等[96] 考虑区块 MDP 的因果不变性预测, 其中不同场景下的观测分布会发生变化, 但潜在状态空间中环境动态和奖励函数是相同的. 文章提出一种不变预测方法提取潜在状态, 并将其迁移到多环境场景下, 解决了潜在空间动态结构的泛化问题. 遵循类似的思路, 因果情景强化学习算法 CCRL (Causal contextual reinforcement learning)[97] 假设情景变量的变化会导致状态分布的变化. CCRL 利用情景注意力模块提取解耦特征, 并将其视为因果机制. 通过改变解耦特征, 提高智能体在新场景下的泛化性能. Zhu 等[98] 将不同状态下的动作效果作为不变性来推断因果关系, 提出了不变动作效果模型 IAEM (Invariant action effect model). IAEM 将相邻状态特征的残差作为动作效果, 在不同场景下实现自适应迁移, 提高了样本的利用率和策略的泛化性.

3) 引入因果关系的模仿学习. 在模仿学习任务中, 智能体直接从专家提供的范例中学习控制策略. 由于传统的模仿学习是非因果的, 智能体不知道专家与环境交互的因果结构. 忽略因果关系的盲目模仿会导致反直觉的因果错误识别现象[99], 进而导致模仿策略失效. Haan 等[99] 指出, 基于专家行为的真实因果模型可以减少因果错误识别的影响. 文章通过环境交互或专家查询的方式对观测数据进行有针对性的干预, 学习正确的因果模型. Etesami 等[100] 假定系统中某些模块因果机制发生变化, 但动作效果机制保持不变, 并在此基础上分析了因果机制的可识别情况, 解决了传感器偏倚情况下的策略迁移问题. 尽管大多数模仿学习任务都假定专家变量可完全观测, 但是实际系统中可能存在混杂因子, 对模仿学习造成不利影响. 针对存在未被观测的混杂因子场景, Zhang 等[101] 利用结构因果模型学习专家范例的数据生成过程, 并利用观测数据中包含的定量知识学习模仿策略. Park 等[102] 以提取语义对象的方式调整模仿策略, 提出了对象感知正则化算法 OREO (Object-aware regularization). 为了防止策略学习到与专家行为密切相关的混杂因子, OREO 鼓励策略统一关注所有语义对象, 显著提高了模仿学习的性能.

### 3.3 利用因果推理实现策略优化

#### 3.3.1 动作效果估计

在强化学习的场景下, 动作效果估计的关键问题在于: 1) 量化智能体动作对环境造成的影响; 2) 获得数据的无偏估计, 进而通过干预因果图改变策略

分布, 有效地指导策略更新.

针对稀疏奖励下的探索和信用分配问题, Corcoll 等[103] 提出了一种基于受控效果的分层强化学习结构 CEHRL (Controlled effects for hierarchical reinforcement learning). CEHRL 智能体基于随机效应进行探索, 并依靠反事实推理识别动作对环境的因果影响. 分层式的结构允许高层策略设置跟时间有关的目标, 以此实现长期信用分配, 高效地学习特定任务的行为. Seitzer 等[104] 引入了基于条件互信息的情境相关因果影响度量 SDCI (Situation-dependent causal influence), 用于衡量动作对环境的因果影响, 进而有效地指导学习. 通过将 SDCI 集成到强化学习算法中, 改进智能体探索能力和离线策略学习性能. 针对强化学习样本效率不高的问题, Pitis 等[105] 定义了局部因果模型, 并提出了一种用于反事实数据增强的算法, 使用基于注意力的方法在解耦状态空间中发现局部因果结构. 这种局部因果结构可用于提高模型的预测性能, 改善非策略强化学习的样本效率. 为了构建与强化学习智能体相关的有效因果表示, Herlau 等[106] 以最大化自然间接效应为目标识别因果变量. 识别的因果变量可以集成环境的特征, 从而确保因果表征与智能体相关.

此外, 虽然动作效果估计可以量化干预和结果之间的影响, 但是采集的观测数据受现有的策略影响, 可能会间接造成选择偏倚问题. 为了实现数据的无偏估计, 研究人员常常采用重要性采样加权[107] 进行离线策略评估, 但是该方法具有高方差和高度依赖权重的缺陷. 为了从观测数据中选择最佳策略, Atan等[108] 考虑了观测数据评估新策略时产生的估计误差, 提供了估计误差的理论界限, 并提出了一种使用域对抗神经网络选择最优策略的方法, 结果表明估计误差取决于观测数据和随机数据之间的 H 散度. 在批量学习的场景下, Swaminathan 等[109] 指出仅对离策略系统的性能进行无偏估计不足以实现稳健学习, 还需要在假设空间中推断估计量的方差有何不同. 该项研究通过倾向性评分设计了反事实估计器, 提出了反事实风险最小化原则, 证明了倾向加权经验风险估计计量方差的广义误差界限. 为了学习结构化输出预测的随机线性规则, 提出了指数模型策略优化器, 从而实现有效的随机梯度优化. 为了消除由旧策略和新策略引起的分布偏倚, 精确评估新策略的效果, Zou 等[110] 提出了重点上下文平衡算法 FCB (Focused context balancing), 用于学习上下文平衡的样本权重.

#### 3.3.2 反事实动作推理

利用因果框架, 智能体可以进一步回答与强化学习控制任务相关的反事实问题. 例如在已有观测数据的前提下, “如果策略中的某些动作发生变化, 系统的控制性能能否提升”? 目前, 反事实动作推理已经被证明可以提高强化学习算法的样本效率和可解释性[81, 111]. Madumal 等[112] 提出了一种基于结构因果模型的行为影响模型, 利用因果模型进行反事实分析, 提高了模型的可解释性. 在非平稳数据的场景下, Lu 等[81] 提出了一种基于反事实的数据增强算法. 该算法利用结构因果模型对环境动态进行建模, 并基于多领域数据的共性和差异进行因果模型估计. 智能体可以根据结构因果模型进行反事实推理, 解决了有限经验导致策略偏倚的问题, 避免风险性探索. 同时利用反事实推理进行数据集扩充, 提高了数据利用率. 在 POMDP 的框架下, Buesing 等[111]提出了反事实指导的策略搜索算法 CF-GPS (Counterfactually-guided policy search), 基于结构因果模型对任意策略进行反事实评估, 改善策略性能, 消除模型预测的偏差.

### 3.4 因果强化学习的应用

因果强化学习作为一种通用的学习算法, 目前在机器人控制[104, 113]、医疗健康[91]、推荐系统[114]、金融投资[115] 和游戏控制[116] 等多个领域中有着广泛的应用. 在机器人控制领域, Liang 等[113] 在仿真机械臂控制系统中, 将神经网络与概率图模型相结合, 构建了观测数据的因果图模型, 控制机械臂进行绘画操作和轮胎拆卸, 提高了数据利用率和强化学习算法的可解释性. 在医疗健康领域, Zhang 等[91] 基于因果强化学习在肺癌和呼吸困难数据集上设计了最佳动态治疗方案, 提升了算法的在线性能和数据效率. 在推荐系统领域, Bottou 等[114] 基于 Bing 搜索引擎的广告投放系统, 利用因果推理理解用户与环境交互的行为, 致力于合理地使用因果推理和机器学习技术进行广告投放. 在金融投资领域, Wang 等[115] 提出了一种优化投资策略的深度强化学习方法 DeepTrader. 该方法将风险收益平衡问题构建为强化学习问题, 并利用分层图结构建模资产的时空相关性. 其估计的因果结构能够反映资产之间的相互关系, 有效平衡收益与风险. 在游戏控制领域, Shi 等[116] 针对 Atari 2 600 游戏环境, 提出了时空因果解释模型, 对观测数据与智能体决策之间的时序因果关系进行建模, 并使用一个单独的因果发现网络来识别时空因果特征. Madumal 等[112] 在星际争霸游戏环境中使用因果模型来推导无模型强化学习智能体行为的因果解释. 利用结构因果模型对系统进行建模, 然后基于反事实推理生成对动作的解释.

## 4 总结与展望

由于在可解释性以及跨域迁移等方面展现出优势, 因果理论已经被广泛应用于强化学习领域, 并且在控制系统中表现出了良好的性能. 本文致力于阐述因果强化学习算法如何探索数据之间的因果关系, 并在决策过程中提供因果解释. 因果强化学习以无监督的方式构建环境的因果模型, 实现跨域分布泛化, 并利用因果模型进行推理, 设计有效的干预措施进行策略更新. 本文首先概述了强化学习和因果理论的背景知识, 在此基础上, 对因果强化学习的研究现状进行阐述. 针对强化学习领域的两类研究缺陷, 总结了四类研究方向, 具体包括: 1) 因果表征提取; 2) 可迁移的环境因果模型; 3) 动作效果估计; 4) 反事实动作推理.

虽然基于因果建模的强化学习控制可以解决强化学习可解释性和可迁移性的问题, 提升数据利用率, 但是仍存在以下缺点: 1) 依赖不可测试假设. 尽管目前已有多项研究成果可以根据观测数据估计因果结构, 但这些方法通常是不可扩展的, 依赖于不可测试的假设 (如因果忠诚性假设), 因此难以融入高维、复杂和非线性的强化学习系统. 2) 欠缺理论研究基础. 目前针对因果强化学习理论层面上的研究还远远不够. 例如在因果表征领域, 现有的可识别性理论研究大多基于非平稳或时序数据, 并且需要对模型类型做出较强的假设. 在更一般的场景下 (如因果关系发生变化或存在瞬时因果关系) 的可识别性理论研究目前还是空白. 3) 难以保证控制性能. 虽然利用因果理论, 动作策略能够表现出良好的控制效果, 但是基于探索与试错的方法并不能在理论上保证控制性能的收敛. 目前还没有一套完善的框架能够评估因果强化学习的控制策略是否稳定, 这可能阻碍因果理论在强化学习控制系统中的研究发展.

综上, 虽然因果强化学习展现出了具有潜力的应用前景, 但是目前研究成果相对较少, 研究的广度和深度都略显不足, 还存在以下待解决的问题.

1) 探索归纳偏置对因果强化学习的影响. 归纳偏置指的是学习算法中假设的集合. 目前大多数因果迁移强化学习的研究都是基于独立因果机制和最小变化原则. 当不满足条件独立性假设或没有额外辅助信息的情况下, 如何选取归纳偏置, 使算法能够自动检测分布的变化并在有限时间内保证算法收敛是一个亟待解决的问题.

2) 完善潜在因果变量的可识别性理论. 从因果表征的角度来说, 潜在因果变量的可识别性是因果变量提取和因果动态分析的理论基础. 虽然已有研究表明在非参数非平稳模型或者线性高斯平稳模型的假设下, 潜在因果变量可识别性可以得到保证[82], 但是当变量间因果关系发生变化或存在瞬时因果关系时, 如何基于观测数据恢复潜在因果变量是一个值得研究的问题.

3) 构造因果强化学习框架的稳定性评估机制. 从策略学习的角度来说, 确保控制器的稳定是控制理论中首要考虑的问题. 虽然已有研究表明, 在反事实数据增强的场景下, Q 学习可以收敛到最优值函数[81], 但是如何构造一套完整的因果强化学习框架以评估控制策略的稳定性是一个亟待解决的问题.

解决上述问题并将因果强化学习推向更广阔、更现实的应用场景将是未来的研究方向, 具体来说包括以下几个方面.

1) 合理利用观测数据和干预数据. 在因果强化学习中, 根据有无人为干预可以将数据分为无人为干预的观测数据和有人为干预的干预数据. 从数据分布上来看, 观测数据可能受控制策略、混杂因子和潜在因果变量的影响, 干预数据受人为控制的影响, 由这些原因导致的分布不匹配会造成选择偏倚的问题. 但是如果对选择偏倚进行适当的修正, 则可以提高数据利用率, 增加模型的可解释性. 因此在强化学习中合理地利用观测数据和干预数据, 采取适当的方式将知识分解为独立因果机制非常具有研究价值.

2) 构建普适的基准测试环境. 在强化学习的应用背景下, 传统的评估指标不足以判断因果模型的好坏. 大多数研究成果都在不同的实验场景下验证算法性能, 无法横向判断模型结构的好坏, 也难以衡量因果模型和强化学习算法对控制性能的贡献程度. 因此构建一个普适的因果强化学习基准数据集, 验证和比较各类因果强化学习方法就显得至关重要.

3) 将因果强化学习拓展到多智能体场景. 目前针对因果强化学习的研究都是针对单智能体. 在多智能体场景下, 联合状态空间和联合动作空间将随着智能体个数的增加呈指数性扩大, 极大地加重了计算负担. 考虑到使用恰当的结构化表征有利于提高系统控制性能, 如何在多智能体系统中构建可迁移的环境因果模型, 减轻计算负担并提高系统的可解释性, 将是非常有趣且可行的研究方向.

## 致谢

感谢 Carnegie Mellon University 的 Zhang Kun 教授在论文修改过程中提出的宝贵建议.

### References

1 Sun Chang-Yin, Wu Guo-Zheng, Wang Zhi-Heng, Cong Yang,

Mu Chao-Xu, He Wei. On challenges in automation science and technology. *Acta Automatica Sinica*, 2021, **47**(2): 464−474
(孙长银, 吴国政, 王志衡, 丛杨, 穆朝絮, 贺威. 自动化学科面临的挑战. 自动化学报, 2021, **47**(2): 464−474)

2 Silver D, Huang A, Maddison C J, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 2016, **529**(7587): 484−489

3 Vinyals O, Babuschkin I, Czarnecki W M, Mathieu M, Dudzik A, Chung J, et al. Grandmaster level in starcraft II using multi-agent reinforcement learning. *Nature*, 2019, **575**(7782): 350−354

4 Brown N, Sandholm T. Superhuman AI for multiplayer poker. *Science*, 2019, **365**(6456): 885−890

5 Wurman P R, Barrett S, Kawamoto K, MacGlashan J, Subramanian K, Walsh T J, et al. Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature*, 2022, **602**(7896): 223−228

6 Wei F R, Wan Z Q, He H B. Cyber-attack recovery strategy for smart grid based on deep reinforcement learning. *IEEE Transactions on Smart Grid*, 2020, **11**(3): 2476−2486

7 Zhang D X, Han X Q, Deng C Y. Review on the research and practice of deep learning and reinforcement learning in smart grids. *CSEE Journal of Power and Energy Systems*, 2018, **4**(3): 362−370

8 Liang X D, Wang T R, Yang L N, Xing E. CIRL: Controllable imitative reinforcement learning for vision-based self-driving. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 604−620

9 El Sallab A, Abdou M, Perot E, Yogamani S. Deep reinforcement learning framework for autonomous driving. arXiv: 1704.02532, 2017.

10 Kober J, Bagnell J A, Peters J. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 2013, **32**(11): 1238−1274

11 Chen Jin-Tao, Li Hong-Yi, Ren Hong-Ru, Lu Ren-Quan. Cooperative indoor path planning of multi-UAVs for high-rise fire fighting based on RRT-forest algorithm. *Acta Automatica Sinica*, DOI: 10.16383/j.aas.c210368
(陈锦涛, 李鸿一, 任鸿儒, 鲁仁全. 基于 RRT 森林算法的高层消防无人机室内协同路径规划. 自动化学报, DOI: 10.16383/j.aas.c210368)

12 Li Hong-Yi, Wang Yan, Yao De-Yin, Zhou Qi, Lu Ren-Quan. Robust adaptive sliding mode attitude control of MQUAVs based on event-triggered mechanism. *Scientia Sinica Informationis*, 2023, **53**(1): 66−80
(李鸿一, 王琰, 姚得银, 周琪, 鲁仁全. 基于事件触发机制的多四旋翼无人机鲁棒自适应滑模姿态控制. 中国科学: 信息科学, 2023, **53**(1): 66−80)

13 Li Jia-Ning, Xiong Rui-Bin, Lan Yan-Yan, Pang Liang, Guo Jia-Feng, Cheng Xue-Qi. Overview of the frontier progress of causal machine learning. *Journal of Computer Research and Development*, 2023, **60**(1): 59−84
(李家宁, 熊睿彬, 兰艳艳, 庞亮, 郭嘉丰, 程学旗. 因果机器学习的前沿进展综述. 计算机研究与发展, 2023, **60**(1): 59−84)

14 Zhang A, Ballas N, Pineau J. A dissection of overfitting and generalization in continuous reinforcement learning. arXiv: 1806.07937, 2018.

15 Zhang C Y, Vinyals O, Munos R, Bengio S. A study on overfitting in deep reinforcement learning. arXiv: 1804.06893, 2018.

16 AAAI-20 tutorial representation learning for causal inference [Online], available: http://cobweb.cs.uga.edu/~shengli/AAAI20-Causal-Tutorial.html, February 8, 2020

17 Causal reinforcement learning [Online], available: https://crl.causalai.net/, December 24, 2022

18 Elements of reasoning: Objects, structure, and causality: Virtual ICLR 2022 workshop [Online], available: https://objects-structure-causality.github.io/, April 29, 2022

19 NeurIPS 2018 workshop on causal learning [Online], available: https://sites.google.com/view/nips2018causallearning/home, December 7, 2018

20 Moerland T M, Broekens J, Plaat A, Catholijn M J. Model-based reinforcement learning: A survey. *Foundations and Trends ® in Machine Learning*, **16**(1): 1−118

21 Yi F J, Fu W L, Liang H. Model-based reinforcement learning: A survey. In: Proceedings of the 18th International Conference on Electronic Business. Guilin, China: 2018. 421−429

22 Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 1996, **4**(1): 237−285

23 Wang H N, Liu N, Zhang Y Y, Feng D W, Huang F, Li D S, et al. Deep reinforcement learning: A survey. *Frontiers of Information Technology & Electronic Engineering*, 2020, **21**(12): 1726−1744

24 Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 2018, **362**(6419): 1140−1144

25 Anthony T, Zheng T, Barber D. Thinking fast and slow with deep learning and tree search. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 5366−5376

26 Schmidhuber J, Huber R. Learning to generate artificial fovea trajectories for target detection. *International Journal of Neural Systems*, 1991, **2**(01n02): 125−134

27 Schmidhuber J. An on-line algorithm for dynamic reinforcement learning and planning in reactive environments. In: Proceedings of the International Joint Conference on Neural Networks. San Diego, USA: IEEE, 1990. 253−258

28 Parr R, Li L H, Taylor G, Painter-Wakefield C, Littman M L. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In: Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland: ACM, 2008. 752−759

29 Deisenroth M P, Rasmussen C E. PILCO: A model-based and data-efficient approach to policy search. In: Proceedings of the 28th International Conference on Machine Learning. Bellevue, USA: Omnipress, 2011. 465−472

30 Hester T, Stone P. TEXPLORE: Real-time sample-efficient reinforcement learning for robots. *Machine Learning*, 2013, **90**(3): 385−429

31 Müller K R, Smola A J, Rätsch G, Schölkopf B, Kohlmorgen J, Vapnik V. Predicting time series with support vector machines. In: Proceedings of the 7th International Conference on Artificial Neural Networks. Lausanne, Switzerland: Springer, 1997. 999−1004

32 Gal Y, McAllister R, Rasmussen C E. Improving PILCO with Bayesian neural network dynamics models. In: Proceedings of the Data-Efficient Machine Learning Workshop, International Conference on Machine Learning. ICML, 2016. 25

33 Watter M, Springenberg J T, Boedecker J, Riedmiller M. Embed to control: A locally linear latent dynamics model for control from raw images. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2015. 2746−2754

34 Finn C, Tan X Y, Duan Y, Darrell T, Levine S, Abbeel P. Deep spatial autoencoders for visuomotor learning. In: Proceedings of the IEEE International Conference on Robotics and Automation. Stockholm, Sweden: IEEE, 2016. 512−519

35 Guzdial M, Li B Y, Riedl M O. Game engine learning from video. In: Proceedings of the 26th International Joint Confer-

ence on Artificial Intelligence. Melbourne, Australia: IJCAI.org, 2017. 3707−3713

36 Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing Atari with deep reinforcement learning. arXiv: 1312.5602, 2013.

37 Singh S, Jaakkola T, Littman M L, Szepesvári C. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 2000, **38**(3): 287−308

38 Watkins C J C H, Dayan P. *Q*-learning. *Machine Learning*, 1992, **8**(3−4): 279−292

39 Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, **518**(7540): 529−533

40 Wang Z Y, Schaul T, Hessel M, van Hasselt H, Lanctot M, de Freitas N. Dueling network architectures for deep reinforcement learning. In: Proceedings of the 33rd International Conference on Machine Learning. New York, USA: JMLR.org, 2016. 1995−2003

41 van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, USA: AAAI, 2016. 2094−2100

42 Fortunato M, Azar M G, Piot B, Menick J, Hessel M, Osband I, et al. Noisy networks for exploration. In: Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net, 2018.

43 Bellemare M G, Dabney W, Munos R. A distributional perspective on reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR, 2017. 449−458

44 Silver D, Lever G, Heess N, Degris T, Wierstra D, Riedmiller M. Deterministic policy gradient algorithms. In: Proceedings of the 31st International Conference on Machine Learning. Beijing, China: PMLR, 2014. 387−395

45 Schulman J, Levine S, Abbeel P, Jordan M, Moritz P. Trust region policy optimization. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR.org, 2015. 1889−1897

46 Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv: 1707.06347, 2017.

47 Lillicrap T P, Hunt J J, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning. In: Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico: ICLR, 2016.

48 Mnih V, Badia A P, Mirza M, Graves A, Lillicrap T, Harley T, et al. Asynchronous methods for deep reinforcement learning. In: Proceedings of the 33rd International Conference on Machine Learning. New York, USA: JMLR.org, 2016. 1928−1937

49 Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018. 1861−1870

50 Zhang K, Huang B W, Zhang J J, Glymour C, Schölkopf B. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia: IJCAI.org, 2017. 1347−1353

51 Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I J, et al. Intriguing properties of neural networks. In: Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: ICLR, 2014.

52 Kuang K, Li L, Geng Z, Xu L, Zhang K, Liao B S, et al. Causal inference. *Engineering*, 2020, **6**(3): 253−263

53 Shen X P, Ma S S, Vemuri P, Simon G, Alzheimer′s Disease Neuroimaging Initiative. Challenges and opportunities with causal discovery algorithms: Application to Alzheimer′s pathophysiology. *Scientific Reports*, 2020, **10**(1): Article No. 2975

54 Eberhardt F. Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, 2017, **3**(2): 81−91

55 Nogueira A R, Gama J, Ferreira C A. Causal discovery in machine learning: Theories and applications. *Journal of Dynamics and Games*, 2021, **8**(3): 203−231

56 Guo R C, Cheng L, Li J D, Hahn P R, Liu H. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys*, 2020, **53**(4): Article No. 75

57 Zhang K, Schölkopf B, Spirtes P, Glymour C. Learning causality and causality-related learning: Some recent progress. *National Science Review*, 2018, **5**(1): 26−29

58 Peters J, Janzing D, Schölkopf B. *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge: The MIT Press, 2017.

59 Bhide A, Shah P S, Acharya G. A simplified guide to randomized controlled trials. *Acta Obstetricia et Gynecologica Scandinavica*, 2018, **97**(14): 380−387

60 Vansteelandt S, Daniel R M. On regression adjustment for the propensity score. *Statistics in Medicine*, 2014, **33**(23): 4053−4072

61 Austin P C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 2011, **46**(3): 399−424

62 Pearl J. *Causality* (Second edition). New York: Cambridge University Press, 2009.

63 Pearl J, Glymour M, Jewell N P. *Causal Inference in Statistics: A Primer*. Chichester: John Wiley & Sons, 2016.

64 Spirtes P, Glymour C N, Scheines R. *Causation, Prediction, and Search* (Second edition). Cambridge: MIT Press, 2000.

65 Colombo D, Maathuis M H. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 2014, **15**(1): 3741−3782

66 Le T D, Hoang T, Li J Y, Liu L, Liu H W, Hu S. A fast PC algorithm for high dimensional causal discovery with multi-core PCs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016, **16**(5): 1483−1495

67 Spirtes P L, Meek C, Richardson T S. Causal inference in the presence of latent variables and selection bias. In: Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence. Montreal, Canada: Morgan Kaufmann, 1995. 499−506

68 Colombo D, Maathuis M H, Kalisch M, Richardson T S. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 2012, **40**(1): 294−321

69 Zhang K, Peters J, Janzing D, Schölkopf B. Kernel-based conditional independence test and application in causal discovery. In: Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence. Barcelona, Spain: AUAI Press, 2011.

70 Chickering D M. Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, 2003, **3**: 507−554

71 Ramsey J D. Scaling up greedy causal search for continuous variables. arXiv: 1507.07749, 2015.

72 Hauser A, Bühlmann P. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 2012, **13**(79): 2409−2464

73 Ogarrio J M, Spirtes P, Ramsey J. A hybrid causal search al-

gorithm for latent variable models. In: Proceedings of the 8th Conference on Probabilistic Graphical Models. Lugano, Switzerland: JMLR.org, 2016. 368−379

74 Shimizu S, Hoyer P O, Hyvärinen A, Kerminen A. A linear non-Gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 2006, **7**: 2003−2030

75 Hoyer P O, Janzing D, Mooij J, Peters J, Schölkopf B. Nonlinear causal discovery with additive noise models. In: Proceedings of the 21st International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2008. 689−696

76 Hoyer P O, Hyvärinen A, Scheines R, Spirtes P L, Ramsey J, Lacerda G, et al. Causal discovery of linear acyclic models with arbitrary distributions. In: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence. Helsinki, Finland: AUAI Press, 2008. 282−289

77 Zhang K, Hyvärinen A. On the identifiability of the post-nonlinear causal model. In: Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. Montreal, Canada: AUAI Press, 2009. 647−655

78 Zhang K, Chan L W. Extensions of ICA for causality discovery in the Hong Kong stock market. In: Proceedings of the 13th International Conference on Neural Information Processing. Hong Kong, China: Springer, 2006. 400−409

79 Sun Y W, Zhang K, Sun C Y. Model-based transfer reinforcement learning based on graphical model representations. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(2): 1035−1048

80 Huang B W, Zhang K, Zhang J J, Sanchez-Romero R, Glymour C, Schölkopf B. Behind distribution shift: Mining driving forces of changes and causal arrows. In: Proceedings of the IEEE International Conference on Data Mining. New Orleans, USA: IEEE, 2017. 913−918

81 Lu C C, Huang B W, Wang K, Hernández-Lobato J M, Zhang K, Schölkopf B. Sample-efficient reinforcement learning via counterfactual-based data augmentation. arXiv: 2012.09092, 2020.

82 Yao W R, Sun Y W, Ho A, Sun C Y, Zhang K. Learning temporally causal latent processes from general temporal. In: Proceedings of the 10th International Conference on Learning Representations. Virtual: ICLR, 2022.

83 Huang B W, Lu C C, Liu L Q, Hernández-Lobato J M, Glymour C, Schölkopf B, et al. Action-sufficient state representation learning for control with structural constraints. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 9260−9279

84 Rezende D J, Danihelka I, Papamakarios G, Ke N R, Jiang R, Weber T, et al. Causally correct partial models for reinforcement learning. arXiv: 2002.02836, 2020.

85 Sontakke S A, Mehrjou A, Itti L, Schölkopf B. Causal curiosity: RL agents discovering self-supervised experiments for causal representation learning. In: Proceedings of the 38th International Conference on Machine Learning. Virtual: PMLR, 2021. 9848−9858

86 Gasse M, Grasset D, Gaudron G, Oudeyer P Y. Causal reinforcement learning using observational and interventional data. arXiv: 2106.14421, 2021.

87 Zhang A, Lipton Z C, Pineda L, Azizzadenesheli K, Anandkumar A, Itti L, et al. Learning causal state representations of partially observable environments. arXiv: 1906.10437, 2019.

88 Bareinboim E, Forney A, Pearl J. Bandits with unobserved confounders: A causal approach. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2015. 1342−1350

89 Zhang J Z, Bareinboim E. Transfer learning in multi-armed bandit: A causal approach. In: Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems. São Paulo, Brazil: ACM, 2017. 1778−1780

90 Lu C C, Schölkopf B, Hernández-Lobato J M. Deconfounding reinforcement learning in observational settings. arXiv: 1812.10576, 2018.

91 Zhang J Z. Designing optimal dynamic treatment Regimes: A causal reinforcement learning approach. In: Proceedings of the 37th International Conference on Machine Learning. Article No. 1021

92 Wang L X, Yang Z R, Wang Z R. Provably efficient causal reinforcement learning with confounded observational data. In: Proceedings of the 35th Conference on Neural Information Processing Systems. NeurIPS, 2021. 21164−21175

93 Taylor M E, Stone P. Transfer learning for reinforcement learning domains: A survey. *The Journal of Machine Learning Research*, 2009, **10**: 1633−1685

94 Schölkopf B, Locatello F, Bauer S, Ke N R, Kalchbrenner N, Goyal A, et al. Toward causal representation learning. *Proceedings of the IEEE*, 2021, **109**(5): 612−634

95 Huang B W, Fan F, Lu C C, Magliacane S, Zhang K. ADARL: What, where, and how to adapt in transfer reinforcement learning. In: Proceedings of the 10th International Conference on Learning Representations. Virtual: ICLR, 2022.

96 Zhang A, Lyle C, Sodhani S, Filos A, Kwiatkowska M, Pineau J, et al. Invariant causal prediction for block MDPs. In: Proceedings of the 37th International Conference on Machine Learning. Shenzhen, China: PMLR, 2020. 11214−11224

97 Eghbal-zadeh H, Henkel F, Widmer G. Learning to infer unseen contexts in causal contextual reinforcement learning. In: Proceedings of the Self-Supervision for Reinforcement Learning. 2021.

98 Zhu Z M, Jiang S Y, Liu Y R, Yu Y, Zhang K. Invariant action effect model for reinforcement learning. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual: AAAI, 2022. 9260−9268

99 de Haan P, Jayaraman D, Levine S. Causal confusion in imitation learning. In: Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2019. 11693−11704

100 Etesami J, Geiger P. Causal transfer for imitation learning and decision making under sensor-shift. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020. 10118−10125

101 Zhang J Z, Kumor D, Bareinboim E. Causal imitation learning with unobserved confounders. In: Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2020. 12263−12274

102 Park J, Seo Y, Liu C, Zhao L, Qin T, Shin J, et al. Object-aware regularization for addressing causal confusion in imitation learning. In: Proceedings of the 35th Conference on Neural Information Processing Systems. NeurIPS, 2021. 3029−3042

103 Corcoll O, Vicente R. Disentangling causal effects for hierarchical reinforcement learning. arXiv: 2010.01351, 2020.

104 Seitzer M, Schölkopf B, Martius G. Causal influence detection for improving efficiency in reinforcement learning. In: Proceedings of the 35th Conference on Neural Information Processing Systems. NeurIPS, 2021. 22905−22918

105 Pitis S, Creager E, Garg A. Counterfactual data augmentation using locally factored dynamics. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 335

106 Herlau T, Larsen R. Reinforcement learning of causal variables using mediation analysis. In: Proceedings of the 36th AAAI

Conference on Artificial Intelligence. Virtual: AAAI, 2022. 6910−6917

107 Precup D, Sutton R S, Singh S. Eligibility traces for off-policy policy evaluation. In: Proceedings of the 17th International Conference on Machine Learning. Stanford, USA: Morgan Kaufmann, 2000. 759−766

108 Atan O, Zame W R, van der Schaar M. Learning optimal policies from observational data. arXiv: 1802.08679, 2018.

109 Swaminathan A, Joachims T. Counterfactual risk minimization: Learning from logged bandit feedback. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR.org, 2015. 814−823

110 Zou H, Kuang K, Chen B Q, Chen P X, Cui P. Focused context balancing for robust offline policy evaluation. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage, USA: ACM, 2019. 696−704

111 Buesing L, Weber T, Zwols Y, Racanière S, Guez A, Lespiau J B, et al. Woulda, coulda, shoulda: Counterfactually-guided policy search. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net, 2019.

112 Madumal P, Miller T, Sonenberg L, Vetere F. Explainable reinforcement learning through a causal lens. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020. 2493−2500

113 Liang J C, Boularias A. Inferring time-delayed causal relations in POMDPs from the principle of independence of cause and mechanism. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal, Canada: IJCAI.org, 2021. 1944−1950

114 Bottou L, Peters J, Quiñonero-Candela J, Charles D X, Chickering D M, Portugaly E, et al. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 2013, **14**(1): 3207−3260

115 Wang Z C, Huang B W, Tu S K, Zhang K, Xu L. Deeptrader: A deep reinforcement learning approach for risk-return balanced portfolio management with market conditions embedding. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual: AAAI, 2021. 643−650

116 Shi W J, Huang G, Song S J, Wu C. Temporal-spatial causal interpretations for vision-based reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **44**(12): 10222−10235

**孙悦雯** 东南大学自动化学院博士研究生. 2017 年获得山东大学学士学位. 主要研究方向为强化学习与因果发现.
E-mail: amber_sun@seu.edu.cn
(**SUN Yue-Wen** Ph.D. candidate at the School of Automation, Southeast University. She received her bachelor degree from Shandong University in 2017. Her research interest covers reinforcement learning and causal discovery.)

**柳文章** 安徽大学人工智能学院博士后. 2016 年获得吉林大学学士学位, 2022 年获得东南大学博士学位. 主要研究方向为多智能体强化学习, 迁移强化学习.
E-mail: wzliu@ahu.edu.cn
(**LIU Wen-Zhang** Postdoctor at School of Artificial Intelligence, Anhui University. He received his bachelor degree and Ph.D. degree from Jilin University in 2016 and Southeast University in 2022, respectively. His research interest covers multi-agent reinforcement learning and transfer reinforcement learning.)

**孙长银** 东南大学自动化学院教授. 主要研究方向为智能控制与优化, 强化学习, 神经网络, 数据驱动控制. 本文通信作者.
E-mail: cysun@seu.edu.cn
(**SUN Chang-Yin** Professor at the School of Automation, Southeast University. His research interest covers intelligent control and optimization, reinforcement learning, neural networks, and data-driven control. Corresponding author of this paper.)