软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(5):2300−2316 [doi: 10.13328/j.cnki.jos.006485]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

强化学习可解释性基础问题探索和方法综述*

刘 潇, 刘书洋, 庄韫恺, 高 阳

(计算机软件新技术国家重点实验室 (南京大学), 江苏 南京 210023)

通信作者: 高阳, E-mail: gaoy@nju.edu.cn

摘 要: 强化学习是一种从试错过程中发现最优行为策略的技术, 已经成为解决环境交互问题的通用方法. 然而, 作为一类机器学习算法, 强化学习也面临着机器学习领域的公共难题, 即难以被人理解. 缺乏可解释性限制了强化学习在安全敏感领域中的应用, 如医疗、驾驶等, 并导致强化学习在环境仿真、任务泛化等问题中缺乏普遍适用的解决方案. 为了克服强化学习的这一弱点, 涌现了大量强化学习可解释性 (explainable reinforcement learning, XRL) 的研究. 然而, 学术界对 XRL 尚缺乏一致认识. 因此, 探索 XRL 的基础性问题, 并对现有工作进行综述. 具体而言, 首先探讨父问题——人工智能可解释性, 对人工智能可解释性的已有定义进行了汇总; 其次, 构建一套可解释性领域的理论体系, 从而描述 XRL 与人工智能可解释性的共同问题, 包括界定智能算法和机械算法、定义解释的含义、讨论影响可解释性的因素、划分解释的直观性; 然后, 根据强化学习本身的特征, 定义 XRL 的 3 个独有问题, 即环境解释、任务解释、策略解释; 之后, 对现有方法进行系统地归类, 并对 XRL 的最新进展进行综述; 最后, 展望 XRL 领域的潜在研究方向.

关键词: 强化学习可解释性 (XRL); 人工智能可解释性 (XAI); 机器学习 (ML); 人工智能 (AI)

中图法分类号: TP181

中文引用格式: 刘潇, 刘书洋, 庄韫恺, 高阳. 强化学习可解释性基础问题探索和方法综述. 软件学报, 2023, 34(5): 2300–2316. http://www.jos.org.cn/1000-9825/6485.htm

英文引用格式: Liu X, Liu SY, Zhuang YK, Gao Y. Explainable Reinforcement Learning: Basic Problems Exploration and Method Survey. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2300–2316 (in Chinese). http://www.jos.org.cn/1000-9825/6485.htm

Explainable Reinforcement Learning: Basic Problems Exploration and Method Survey

LIU Xiao, LIU Shu-Yang, ZHUANG Yun-Kai, GAO Yang

(State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210023, China)

Abstract: Reinforcement learning is a technique that discovers optimal behavior strategies in a trial-and-error way, and it has become a general method for solving environmental interaction problems. However, as a machine learning method, reinforcement learning faces a common problem in machine learning, or in other words, it is unexplainable. The unexplainable problem limits the application of reinforcement learning in safety-sensitive fields, e.g., medical treatment and transportation, and it leads to a lack of universally applicable solutions in environmental simulation and task generalization. In order to address the problem, extensive research on explainable reinforcement learning (XRL) has emerged. However, academic members still have an inconsistent understanding of XRL. Therefore, this study explores the basic problems of XRL and reviews existing works. To begin with, the study discusses the parent problem, i.e., explainable artificial intelligence, and summarizes its existing definitions. Next, it constructs a theoretical system of interpretability to describe the common problems of XRL and explainable artificial intelligence. To be specific, it distinguishes between intelligent algorithms and mechanical algorithms, defines interpretability, discusses factors that affect interpretability, and classifies the intuitiveness of interpretability. Then, based on the characteristics of reinforcement learning, the study defines three unique problems of XRL, i.e.,

* 基金项目: 科技创新 2030—“新一代人工智能”重大项目 (2018AAA0100900)
本文由“面向开放场景的鲁棒机器学习”专刊特约编辑陈恩红教授、李宇峰副教授、邹权教授推荐.
收稿时间: 2021-02-23; 修改时间: 2021-07-16; 采用时间: 2021-08-27; jos 在线出版时间: 2021-10-20
CNKI 网络首发时间: 2022-11-16

environmental interpretation, task interpretation, and strategy interpretation. After that, the latest research on XRL is reviewed, and the existing methods were systematically classified. Finally, the future research directions of XRL are put forward.

Key words: explainable reinforcement learning (XRL); explainable artificial intelligence (XAI); machine learning (ML); artificial intelligence (AI)

人工智能 (artificial intelligence, AI) 和机器学习 (machine learning, ML) 在计算机视觉[1]、自然语言处理[2]、智能体策略[3]等研究领域都取得了突破, 并逐渐融入人的生活. 虽然 ML 算法对于很多问题具有良好表现, 但由于算法缺乏可解释性, 模型实际使用中常受到质疑[4,5], 尤其在安全敏感的应用领域, 如自动驾驶、医疗等. 缺乏可解释性的问题已经成为机器学习的瓶颈问题之一.

强化学习 (reinforcement learning, RL) 被验证适用于复杂的环境交互类问题[6−8], 如机器人控制[9], 游戏 AI[10]等. 但作为机器学习的一类方法, RL 同样面临着缺乏可解释性的问题, 主要表现在如下 4 个方面.

(1) 安全敏感领域中的应用受限. 由于缺乏可解释性, RL 策略难以保证其可靠性, 存在安全隐患. 这一问题在安全敏感任务 (如医疗、驾驶等) 中难以被忽略. 因此, 为避免模型不可靠带来的危险, RL 在安全敏感任务中大多局限于辅助人类的决策, 如机器人辅助手术[11], 辅助驾驶[12]等.

(2) 真实世界知识的学习困难. 虽然目前 RL 应用在一些仿真环境中具有优异表现, 如 OpenAI gym[13], 但这些仿真环境以简单游戏为主, 与真实世界存在较大差异. 另外, RL 应用难以避免对环境的过拟合. 当过拟合发生时, 模型学到环境的背景信息, 而非真正的知识. 这导致了两难的问题, 一方面, 在真实世界中训练 RL 模型通常消耗巨大, 另一方面, 难以确定在虚拟环境中训练的模型学到了真实的规律.

(3) 相似任务的策略泛化困难. RL 策略通常与环境存在强耦合, 难以被应用到相似环境中. 甚至在同样的环境下, 环境参数的微小变化也会极大影响模型性能. 这一问题影响了模型的泛化能力, 难以确定模型在相似任务中的表现.

(4) 对抗攻击的安全隐患难于应对. 对抗攻击[14]是一种针对模型输入的攻击技术, 通过将微小的恶意扰动加入到模型的输入中生成对抗样本. 对人而言, 对抗样本不影响判断, 甚至难以察觉, 然而对于模型而言, 对抗样本会使模型的输出产生极大的偏差. 对抗攻击从深度学习扩展到 RL[15,16], 成为 RL 算法的安全隐患. 对抗攻击的有效性进一步暴露了 RL 缺乏可解释性的问题, 同时也进一步说明 RL 模型并未学到真正的知识.

解释对模型的设计者和使用者都具有重要的意义. 对于模型的设计者, 解释能体现模型所学的知识, 便于通过人的经验验证模型是否学到鲁棒的知识, 从而使人高效地参与到模型的设计和优化中; 对于特定领域的专家使用者, 解释提供模型的内部逻辑, 当模型表现优于人时, 便于从模型中提取知识以指导人在该领域内的实践. 对于普通用户, 解释呈现模型的决策的原因, 从而加深用户对模型的理解, 增强用户对模型的信心.

强化学习可解释性 (explainable reinforcement learning, XRL), 或可解释强化学习, 是人工智能可解释性 (explainable artificial intelligence, XAI) 的子问题, 用于增强人对模型理解, 优化模型性能, 从而解决上述缺乏可解释性导致的 4 类问题. XRL 与 XAI 之间存在共性, 同时 XRL 具备自身的独特性.

一方面, XRL 与 XAI 存在共性. 首先, 提供解释的对象是智能算法而非机械算法. 机械算法, 如排序、查找等, 其特点是完备的输入, 固定的解法以及明确的解. 而智能算法因为输入的不完备以及解法的不确定, 导致算法必须在解空间中寻找较优的解; 其次, 人和模型是两个直接面对的关键实体. 与其他技术不同, 可解释性方法关注人对模型的理解. 由于人对大量条例混乱的数据缺乏理解, 因此解释通常对模型内在逻辑的抽象, 这一过程必然伴随对模型策略的简化. 其中的难点是, 如何在向人提供解释时, 保证该解释与模型主体逻辑的一致性; 最后, 解释的难度是相对的, 同时由问题规模和模型结构两个因素决定, 并且这两个因素在一定条件下相互转化. 例如, 结构简单的模型 (如决策树、贝叶斯网络等) 通常可以直观地展示输入和输出之间的逻辑关系, 但面对由大量简单结构组成的庞大模型, 其错综复杂的逻辑关系仍然导致模型的整体不可理解. 同时, 虽然结构复杂的模型 (如神经网络) 通常难以被理解, 但当模型被极致约减时 (如将神经网络塌缩为具有少数变量的复合函数), 模型本身仍然可以被人所理解.

另一方面, XRL 也具备自身的独特性. 强化学习问题由环境、任务、智能体策略 3 个关键因素组成, 因此, 解决 XRL 问题必须同时考虑这 3 个关键因素. 由于 XRL 的发展仍处于初步阶段, 大部分方法直接从 XAI 的研究中继承, 导致现有研究集中于对智能体策略的解释, 即解释智能体行为的动机及行为之间的关联. 然而, 缺乏对环境

和任务的认识使得一些关键问题无从解决: 缺乏对环境的认识使人在面临复杂任务时, 缺乏对环境内部规律的理解, 导致对环境状态进行抽象时忽略有利信息, 使智能体难以学到真实的规律; 缺乏对任务的解释使任务目标与交互序列之间的关联不明确, 不利于智能体策略与环境的解耦合, 影响强化学习智能体策略在相似任务或动态环境中的泛化能力. 因此, 对环境、任务和策略的解释存在强关联, 是实现强化学习解释必然面临的问题.

目前, XRL 已经成为 AI 领域的重要议题, 虽然研究者们为提高强化学习模型的可解释性做出了大量工作, 但学术界对 XRL 尚且缺乏一致的认识, 导致所提方法也难以类比. 为了解决这一问题, 本文探索 XRL 的基础性问题, 并对现有工作进行总结. 首先, 本文从 XAI 出发, 对其通用观点进行总结, 作为分析 XRL 问题的基础; 然后, 分析 XRL 与 XAI 的共同问题, 构建出一套可解释性领域的理论体系, 包括界定智能算法和机械算法、定义解释的含义、讨论影响可解释性的因素、划分解释的直观性; 其次, 探讨 XRL 问题的独特性, 提出包括环境解释、任务解释和策略解释的 3 个 XRL 领域的独有问题; 随后, 对现有 XRL 领域的研究进展进行总结. 以技术类别和解释效果为依据将对现有方法进行分类, 对于每个分类, 根据获取解释的时间、解释的范围、解释的程度和 XRL 的独有问题, 确定每类方法的属性; 最后, 展望了 XRL 领域的潜在研究方向, 重点对环境和任务的解释、统一的评估标准两个方向进行展开.

1 人工智能可解释性的观点总结

对 XRL 的研究不能脱离 XAI 的基础. 一方面, XRL 是 XAI 的子领域, 其方法和定义密切相关, 因此 XRL 的现有研究广泛借鉴了 XAI 在其他方向 (如视觉) 的成果; 另一方面, XRL 目前仍处于起步阶段, 对其针对性的讨论较少, 而对于 XAI, 研究者们长期以来进行了广泛的研究和讨论[17−24], 具有深刻的借鉴意义. 基于上述原因, 本文从 XAI 的角度探讨可解释性问题, 整理出学术界对 XAI 的共识, 以此作为 XRL 的研究基础.

虽然学者们从不同角度对 XAI 的定义在特定情况下指导着一类研究. 然而, 缺乏精确而统一的定义使得学术界对 XAI 的认识存在一定差异. 本文对 XAI 相关的定义进行总结, 并将其分为形而上的概念描述、形而下的概念描述两类.

形而上的概念描述使用抽象概念对可解释性进行定义[25−28]. 这些文献使用抽象的词描述可解释性算法, 例如可信的 (trustworthy), 可靠性 (reliability) 等. 其中“可信”意味着人以较强的信心相信模型所做的决定, 而“可靠”意味着模型不同场景下总是能保持其性能. 虽然这样抽象的概念不够精确, 但仍然可以使人准确了解可解释性的目标、对象和作用, 建立对可解释性的直觉认知. 这些概念表明, 可解释性算法具备两个关键实体, 即人和模型. 换而言之, 可解释性是一项以模型为对象, 以人为目标的技术.

形而下的概念描述从哲学、数学等的观点出发, 基于解释的现实意义对其进行定义. 如 Páez 等人[17] 从哲学角度出发, 认为解释所产生的理解并不完全等同于知识, 同时理解的过程也不一定建立在真实的基础上. 我们认为, 解释作为媒介存在, 这个媒介通过呈现模型的真实知识或构建虚拟逻辑的方式, 增强人对模型的理解. 同时, 人对模型的理解不必建立在完全掌握模型的基础上, 只要求掌握模型的主要逻辑, 并能对结果进行符合认知的预测. Doran 等人[29] 认为, 可解释性系统使人们不仅能看到, 更能研究和理解模型输入和输出之间的数学映射. 一般而言, AI 算法的本质是一组由输入到输出的数学映射, 而解释则是将这样的数学映射以人类可理解和研究的方式展现出来. 虽然数学映射也是人们为描述世界而创造的一种方式, 但对于复杂的数学映射 (如用于表示神经网络的高维多层嵌套函数), 人们却无法将其与生活中的直观逻辑相联系. Tjoa 等人[19] 认为, 可解释性是用于解释算法做出的决策, 揭示算法运作机制中的模式以及为系统提供连贯的数学模型或推导. 这一解释也基于数学表达, 反映出人们更多地通过模型的决策模式来理解模型, 而非数学上的可重现性.

一些观点与上述文献存在微小出入, 但仍具有借鉴意义. 例如, Arrieta 等人[21] 认为可解释性是模型的被动特征, 指示模型被人类观察者理解的程度. 这个观点将模型的可解释性视为被动特征, 忽略了模型为了更强的可解释性而主动提出解释的可能. Das 等人[23] 认为, 解释是一种用于验证 AI 智能体或 AI 算法的方式. 这一观点倾向于关注模型的结果, 其目的是确保模型一贯的性能. 然而该描述忽略了一个事实, 即模型本身意味着知识, 可解释性不仅是对模型结果的验证, 同时也有助于从模型中提取人们尚未掌握的知识, 促进人类实践的发展. 虽存在较小出

入, 但上述观点也提出了独特的角度, 例如, 可以将模型的可解释性视为模型的一个特性, 而评估模型的性能是解释的重要功能.

虽然对 XAI 的定义众多, 但就整体而言, 学术界对 XAI 的基本概念仍然是一致的. 本文尝试提取其中的共性作为研究 XRL 问题的理论基础. 通过对以上文献的分析, 我们总结出学术界对 XAI 的共识.

(1) 人与模型是可解释性直接面对的两个关键的实体, 可解释性是一项以模型为对象, 以人为目标的技术.

(2) 解释作为理解的媒介存在, 该媒介可以是真实存在的事物, 也可以是理想构建的逻辑, 亦或是二者并举, 达到让人能够理解模型的目的.

(3) 人对模型的理解不需要建立在完全掌握模型的基础上.

(4) 可准确重现的数学推导不可取代可解释性, 人对模型的理解包括感性和理性的认知.

(5) 可解释性是模型的特性, 这一特性可用于验证模型的性能.

2 强化学习可解释性与人工智能可解释性的共同问题

在对 XAI 定义进行总结的基础上, 本节讨论 XRL 与 XAI 面临的共同问题. 由于 XRL 与 XAI 之间存在强耦合, 因此本节内容既适用于 XAI, 同时也是 XRL 的基础问题.

2.1 智能算法和机械算法界定

可解释性的对象是智能算法而非机械算法. 传统认知中的机械算法, 如排序、查找等, 面对确定的任务目标, 同时具有固定的算法程序. 强化学习作为一种智能算法, 在与环境动态交互的过程中寻找最优的策略, 最大化获得的奖赏. 界定智能算法和机械算法可用于确定被解释的对象, 进而回答“什么需要被解释”的问题. 一方面, 智能算法与机械算法存在差异, 而解释只在面向智能算法时存在必要性; 另一方面, 即使对于强化学习, 也无需对其所有过程产生解释, 而应针对其具有智能算法特性的部分进行解释, 如动作生成、环境状态转移等. 因此, 在讨论可解释性问题前, 有必要区分智能算法和机械算法.

本文根据算法对已知条件的获取程度和建模的完整性, 定义“完全知识”和“完全建模”.

- 完全知识: 已知足够任务相关的有效知识, 具备以机械过程获得最优解的条件.
- 完全建模: 进行完整的问题建模, 具备完成任务所需的计算能力.

完全知识是以机械方法确定最优解的前提. 例如, 求解系数矩阵的秩为 n 的线性方程组, 完全知识表示其增广矩阵的秩大于等于系数矩阵的秩, 此时可以根据当前知识, 获得确定的解或者确定其无解; 完全建模意味着对现有知识的充分利用, 换言之, 完全建模从建模者的角度出发, 表示在解决任务的过程中有能力 (包括程序设计者的设计能力和硬件的算力) 利用所有的知识. 例如, 在 19×19 围棋游戏中, 存在理论上的最优解法, 但目前尚不具备足够的计算能力在有限时间内获取最优解.

根据上述对完全知识和完全建模的定义, 本文进一步提出“任务完全”的概念来确定机械算法与智能算法之间的边界.

- 任务完全: 对特定任务, 具备完全知识并进行完全建模.

任务完全必须在完全知识的前提下进行完全建模. 满足任务完全的条件后, 算法的优劣仅取决于建模方式和使用者的实际需求. 任务完全的定义考虑了知识和建模两方面因素 (图 1).

任务完全的概念可以用来区分机械算法和智能算法. 机械算法是任务完全的, 具体来说, 算法已知足够的知识, 并进行了无简化的建模. 此时, 算法具备获取最优解的条件, 因此算法的过程是确定的, 获得的解也是可预期的. 例如, 经典排序算法、传统数据查询、3×3 井字棋游戏算法等都属于机械算法. 智能算法是任务不完全的, 这意味着算法不具备足够的知识, 或者采取了简化的建模方式. 智能算法无法直接获取最优解, 通常在解空间中寻找较优的解. 如基于贪心策略的算法, 线性回归方法, 机器学习类算法等.

导致任务不完全的可能有二, 即知识不完全和建模不完全. 在知识不完全的情况下, 算法无法直接确定最优解, 因此只能在解空间中逼近最优解. 此时, 智能算法的实际作用是在解空间中进行解的选择. 导致知识不完全的

因素通常是客观的, 如环境状态无法被完全观测, 任务目标不可预知, 任务评价指标的不可知, 任务始终点不可知等; 在建模不完全的情况下, 算法通常忽略某些知识, 导致算法过程没有充分利用知识, 从而无法获得最优解. 建模不完全的原因有客观和主观两方面, 客观原因如建模偏差, 不完全建模等, 主观原因包括降低硬件需求, 模型提速等. 在强化学习中, 并非所有过程具备任务不完全的特点, 因此只有部分需要进行解释, 如策略生成、环境状态转移等.

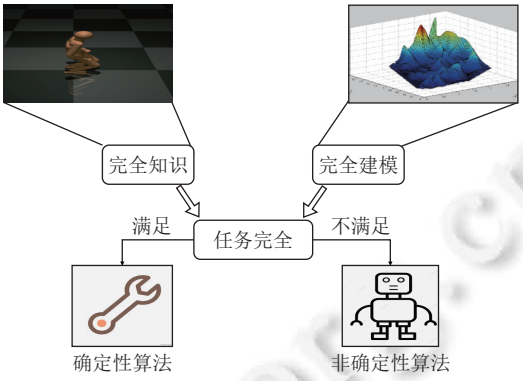

图 1 界定可解释性算法的范围

2.2 对“解释”的定义

在汉语词典中, 解释有“分析、阐明”的含义. 这不仅符合生活中对该词的理解, 同时也与可解释性研究中“解释”的含义相近. 然而, 具体到可解释性的研究中, 这一含义显得宽泛. 我们希望结合对可解释性的理解, 细化“解释”的含义, 使之具有更强的指导意义. 以强化学习模型为例, 模型学习使奖励最大化的策略, 其中包含着环境、奖励和智能体之间的隐式知识, 而 XRL 算法则是将这些隐式知识显式地表现出来. 本文将多个知识视为集合, 称为知识体系, 从知识体系相互之间关系的角度, 对“解释”做出如下定义.

定义 1. 解释. 知识体系之间的简洁映射. 简洁映射是在不引入新知识的条件下对目标知识进行表达.

具体来说, 解释是将基于原知识体系的表达转换为目标知识体系表达的过程, 这个过程仅使用目标知识体系的知识, 而不引入新的知识. 而 XRL 算法的目的在于产生解释, 从而使原知识体系能够被目标知识体系简洁地表达出来. 在 XRL 中, 原知识体系通常指代强化学习模型, 而目标知识体系通常指人的认知, 模型和人是可解释性的两个关键实体. 本文将原知识体系看作由多个元知识及其推论构成的集合. 以 k_i 表示元知识, S 表示知识体系, 则 $S=\{k_1,\ldots,k_n\}$. 假设智能体习得的知识属于知识体系 S_a, 而人类能够理解的知识属于知识体系 S_h, 则解释是将知识体系 S_a 转换为知识体系 S_h 表达的过程. 对于解释而言, 简洁映射是必要的, 非简洁的映射可能提升解释本身的被理解难度, 进而导致解释本身让人无法理解 (见第 2.3 节).

在对知识进行转换表达的过程中, 待解释的知识可能无法完全通过目标知识体系进行描述, 这时只有部分知识可以被解释. 本文使用“完全解释”和“部分解释”的概念描述这一情况, 如图 2.

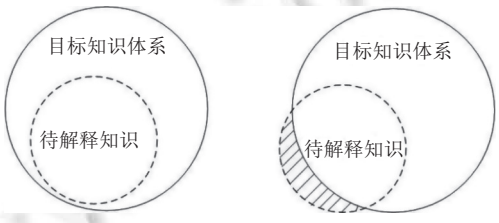

图 2 完全解释 (左), 部分解释 (右)

- 完全解释: 待解释的知识完全被目标知识体系表达. 其中, 被解释的知识属于目标知识体系是其必要条件.
- 部分解释: 待解释的知识的部分被目标知识体系表达.

具体来说, 完全解释和部分解释描述的是知识体系之间的包含情况 (图 2). 只有当待解释的知识体系完全被目标知识体系所包含时, 才可能进行完全解释, 否则只能进行部分解释. 在 XRL 中, 完全解释通常是不必要的. 一方面, 待解释知识体系和目标知识体系的边界难以确定, 导致完全解释难度高且耗费巨大; 另一方面, 实现对模型的解释通常不需要建立在对模型完全掌握的基础上. 因此, 部分解释是大部分可解释性研究中采用的方法, 即只描述算法的主要决策逻辑.

2.3 可解释性的影响因素

一个观点认为, 传统 ML (RL 为其子集) 方法是易于解释的, 而深度学习的引入使得可解释性产生了短板, 导致 ML 难于解释, 因此 ML 解释的本质是对深度学习的解释[21]. 这与可解释性领域的认知相悖[28]. 这一观点只关注模型而忽略了人在可解释性中的地位. 对于人而言, 即使是理论上可被理解的模型, 当规模扩张到一定程度时, 仍然会导致整体的不可理解. 本文对可解释性的影响因素进行如下定义.

- 透明度: 待解释模型结构的简洁程度.
- 模型规模: 待解释模型包含的知识量和知识组合多样化程度.

本文认为, 可解释性是对模型组件透明度和模型规模的综合描述. 透明度和模型规模是影响可解释性的两个主要因素. 具体来说, 可解释性强意味着同时具备高透明度和低复杂度, 而单一因素, 如复杂度高或透明度低将导致模型的弱可解释性 (图 3).

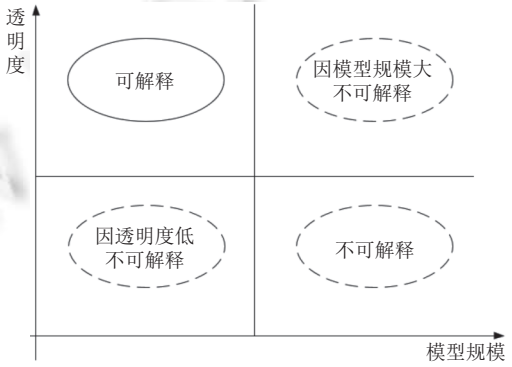

图 3 可解释性的影响因素

在不同语境下, “透明”一词具有不同的含义. 例如, 在软件结构中, 透明指的是对底层过程的抽象程度, 意味着上层程序无需关注底层的实现. 类似的, 透明度在可解释性领域也存在不同的含义, 如文献 [26,27] 认为透明度是模型可以被理解的程度, 将透明度与可解释性等价. 以强化学习为例, 基于值表的强化学习算法在规模一定时通常具有更强的可解释性, 而使用深度学习拟合值表则可解释性更弱, 这是因为通过查询值表而产生策略的过程符合人的直观理解, 但神经网络传播过程仅在数学上可被准确描述, 于人而言透明度更低. 然而, 这一思考将构建模型的基础结构作为可解释性的重点, 而忽略了模型规模对解释带来的难度, 并忽略了解释的目标——人. 因此, 为突出模型规模对解释的影响, 我们仅将透明度狭义理解为待解释模型的结构的简洁程度.

模型规模从人理解能力的角度衡量解释的难度. 具体来说, 假设模型中的知识由一系列元知识构成, 则模型规模表示元知识总量和知识之间组合的多样化程度, 而解释的难度一定程度上取决于模型规模, 当模型规模超过特定范围 (人的理解能力) 时模型将无法被理解. 例如, 线性加性模型、决策树模型、贝叶斯模型, 由于计算过程简洁, 使我们能够轻易了解模型基于何因素得到何种结果, 因此被认为是易于理解的. 然而, 当模型规模逐渐庞大时, 各因素之间的逻辑不可避免地相互交织, 变得错综复杂, 使我们最终无法抓住其主从关系. 对于以简洁结构 (如决策树分支) 构成的大规模模型, 虽然所有结果在理论上有迹可循, 但当模型规模已超越人类的理解能力, 系统整体将仍然不具备可解释性.

2.4 可解释性的程度划分

人的学习过程与强化学习过程存在一定的相似性, 因此, 如果将人脑看作目前最先进的智能模型, 则人对模型的理解不仅是人对模型的直观感受, 也是一个先进的智能体对强化学习模型的综合评估. 然而, 一个无法理解的模型不可能被有效评估, 因此对模型的解释成为人理解模型的媒介. 作为人和模型之间媒介, 可解释性算法不同程度的具备两个相互平衡特点: 接近模型和接近人的感知. 具体来说, 不同的解释有的更注重准确的描述模型, 而另一些更注重与人的感知一致. 基于这一概念, 本文将可解释性分为如下 3 个层次.

(1) 数学表达: 通过理想化的数学推导解释模型. 数学表达是使用数学语言简化模型的表达. 由于强化学习模型建立在数学理论的基础上, 因此通过数学表达可以准确地描述和重构模型. 虽然数学理论体系是人描述世界的一种重要方式, 但其与人的普遍直觉之间存在较大差异. 以深度学习为例, 虽然存在大量文章论证了其在数学上的合理性, 但深度学习方法仍然被认为是不可解释的. 因此, 数学的表达能够在微观 (参数) 层面对模型进行描述, 但难以迁移至人类知识体系.

(2) 逻辑表达: 通过将模型转换为显性的逻辑规律解释模型. 逻辑表达是对模型中主体策略的提取, 即忽略其细微分支, 凸显主体逻辑. 一方面, 逻辑表达保留了模型的主体策略, 因此与模型真实决策结果相近, 解释本身可以部分重现模型的决策; 另一方面, 逻辑表达简化了模型, 符合人的认知. 逻辑表达是较为直观的解释, 但需要人具备特定领域的知识, 是面对人类专家的解释, 而对一般用户尚不够直观.

(3) 感知表达: 通过提供符合人类直觉感知的规律解释模型. 感知表达基于模型生成符合人类感知的解释, 由于不需要人具备特定领域的知识, 因此易于理解. 例如, 可视化关键输入、示例对比等解释形式都属于感知表达的范畴. 然而, 感知表达通常是对模型策略的极大精简, 无法重现模型的决策, 因此只表达决策的合理性.

在可解释性的 3 个层次中, 数学表达作为第 1 个层次, 也是构建强化学习算法的理论基础. 在已知模型所有参数的情况下, 数学表达通常可以较为准确的推断出模型的结果, 然而, 数学上的合理性不意味着能被人所理解; 逻辑表达介于数学表达和感知表达之间, 是对模型策略的近似, 但逻辑表达方法产生的解释通常要求用户具备特定领域的专业知识; 感知表达对模型决策的重要因素进行筛选, 并使用清晰、简洁的形式进行呈现, 虽然结果易于理解, 但已经不具备重构策略的能力. 总而言之, 不同的解释在接近模型和接近人类感知之间存在着平衡, 难以兼顾.

3 强化学习可解释性的独有问题

与其他 ML 方法不同, RL 问题由环境、任务、智能体 3 个关键因素组成. 其中, 环境为给定的具有一定内部规律的黑盒系统; 任务为智能体拟合的目标函数; 策略是智能体行为的依据和一系列行为之间的关联. 根据强化学习的 3 个关键组成因素, 本文归纳出 XRL 的 3 个独有问题, 即环境解释, 任务解释, 策略解释. 3 个独有问题之间存在着密切的关联, 与整个强化学习过程密不可分, 是实现强化学习解释直接面临的问题.

3.1 环境解释

环境解释问题关注环境状态转移的内部规律. 在强化学习中, 环境是具备一定内部规律的黑盒系统, 如何从环境中获取规律是强化学习面临的首要问题. 环境解释用于解决两方面的问题: 一方面, 由于强化学习在真实环境中的学习成本高昂, 往往需要在仿真环境中训练. 然而, 缺乏对环境的认识使人在对环境状态进行抽象时容易忽略有利信息, 从而导致智能体难以学到真正的规律; 另一方面, 由于难以验证模型是否学到真实的规律, 导致模型泛化能力不稳定, 环境中任何微小的改动都可能使算法出现极大偏差.

本文基于状态转移的概念描述环境解释的问题, 具体来说, 对环境的解释目的在于寻找对环境规律的解释, 提高人对环境未来状态的预期能力, 环境解释表达为:

$$\max_{0\leqslant t\leqslant n}\frac{dist\left(P_t,P_t^H(X_E)\right)}{dist(P_t,P_t^H)}\leqslant\delta \tag{1}$$

其中, P_t 为时间 t 下的状态环境的真实状态转移概率, X_E 为对环境变化规律的解释, 该解释中不包含 P 的直接信息, P_t^H 为时间 t 下人对环境状态转移的预期, $P_t^H(X_E)$ 为时间 t 下获得解释 X_E 后人对环境状态转移的预期, $dist(\cdot)$

为差异评估函数, δ 表示解释的程度. 根据公式 (1), X_E 对环境的解释作用越强, 则 $P_t^H(X_E)$ 越接近 P_t, 从而使分子变小. 在一段时间内, 当 $\delta<1$, 则 X_E 对人的理解起着正向作用, 称 X_E 对环境 E 具有 δ 程度的解释能力.

值得注意的是, 环境解释区别于对环境状态转移的拟合. 例如, 部分强化学习算法 (如 DQN[30] 等) 具有拟合环境状态转移的能力, 但使用模型拟合状态转移函数仍然具有较高的复杂性, 无法直接用于提高人对环境的理解. 因此, 环境解释问题的根源在于环境知识的传递, 即帮助人理解环境的内部规律.

3.2 任务解释

任务解释问题关注任务目标与过程状态序列之间的关联. 在任务执行过程中, 环境的过程状态与任务目标息息相关. 然而在强化学习过程中, 过程状态通常由各时间步中零散的环境属性构成, 难以归纳它们与最终目标之间的关系. 因此, 任务解释使用子任务等方式, 对这些过程状态进行抽象, 在减少子任务数量的同时明确子任务与最终任务之间的关联, 使人能够理解任务目标与过程状态序列之间的关联. 因此, 任务解释可以归纳出任务本身的规律, 从而增强模型在相似任务中的泛化性能.

在 RL 中, 回报 (return) 用于衡量过程状态与任务目标关联. 然而, 回报函数的设计通常基于人的主观经验, 主要目的是指导 RL 模型的训练, 函数设计的优劣对模型的能力具有极大影响. 因此, 回报函数并非对过程状态和任务目标之间关联的准确描述. 为了进行区分, 本文使用“贡献”一词进行描述描述任务解释, 过程状态的贡献 U 被描述为:

$$U_t=\sum_{i=0}^{t}(r_i|r_i=C(s_0,\ldots,s_i,a_0,\ldots,a_i)) \tag{2}$$

其中, U_t 为时间 t 内获得的总贡献, $C(\cdot)$ 为贡献函数, r_t 为 t 时刻的贡献值, r_t 被 s_t 直接影响, 同时受到 $\{s_0,\ldots,s_{t-1}\}$ 的传递影响. 本文基于贡献的概念描述任务解释的问题, 具体来说, 任务解释表示为:

$$\max_{0\leqslant t\leqslant n}\frac{dist\left(U_t,U_t^H(X_G)\right)}{dist(U_t,U_t^H)}\leqslant\delta \tag{3}$$

其中, U_t^H 为 t 时间下人对过程状态贡献的预期, $U_t^H(X_G)$ 为引入对任务的解释 X_G 后人对过程状态的贡献预期, δ 表示解释的程度. 因此, 当 X_G 对任务的解释越强, 则人对当前状态下贡献值的预测越准确, 则 δ 越小. 在强化学习任务中, 模型的学习目标通常为最大化 U 的值 (通过设计逼近 U 的回报函数实现). 而对任务的解释目的在于寻找对解释 X_G, 从而在任意中间过程增强人对任务完成度的理解.

值得注意的是, 回报函数的设计是强化学习的重要目标, 但基于人的先验知识设计的回报函数通常具有主观性和稀疏性的特点, 因此无法精确地表示任务目标. 具体来说, 主观性导致回报函数可能关注某些显性标准, 而忽略具体的状态与任务目标的关系, 例如基于游戏时间的回报忽略了游戏中具体状态与任务目标的关系. 稀疏性导致回报可能只关注某些关键状态, 而忽略了过程状态与目标的关系, 例如稀疏奖赏的任务容易忽略奖赏为 0 的状态与目标任务的关系.

3.3 策略解释

策略解释关注智能体的动作推理过程及动作序列之间的关联. 对策略的解释对模型决策过程中的隐式逻辑进行呈现, 是当前 XRL 的主要关注的方向. 在强化学习中, 虽然性能良好的智能体学习到了优质的策略, 但缺乏对策略的解释不仅让我们难以区分这些知识的正确性, 也使得这些知识无法被迁移到人的实践中. 例如, 围棋领域的“AI 棋”(人类复盘时难以理解的落子) 可能包含着对围棋的深层规律, 但目前尚不能被人们的认知所理解. 策略解释对强化学习具有重要的作用, 包括可以发现模型决策的隐式逻辑, 溯源决策的原因, 以及解释动作序列之间的关联. 因此, 策略解释是获取用户信任的关键.

在强化学习中, 智能体策略是状态到动作的映射, 本文基于智能体策略来描述策略解释. 策略解释的基本问题为寻找可解释的状态动作映射方式. 智能体决策过程表述为:

$$P_A=\pi(A|\{s_0,\ldots,s_t\}) \tag{4}$$

其中, P_A 为时间 t 下动作的执行概率, $\pi(\cdot)$ 为智能体策略函数. 则策略解释的目标为寻找 π 的可解释策略, 通过解释

作为媒介以增强人对模型决策的预期能力. 具体来说, 策略解释表示为:

$$\max_{0\leqslant t\leqslant n}\frac{dist\left(P_A,P_A^H(X_\pi)\right)}{dist\left(P_A,P_A^H\right)}\leqslant\delta \tag{5}$$

其中, P_A^H 为时间 t 下人对智能体动作概率的预期, $P_A^H(X_\pi)$, 表示获得对策略 π 的解释 X_π 后人对智能体时间 t 下动作概率的预期. 在一段时间 n 内其获取公式 (5) 左侧的最大值小于等于 δ, 则称为 X_π 对策略 π 实现了 δ 程度的解释, 可见 δ 越小, 解释的程度越高. 策略解释的关键在于使用解释的方式获得策略的隐式逻辑. 因此, 策略解释需要在以可解释方式呈现的基础上, 尽可能使人对模型动作的预测接近模型的真实动作. 为了实现这一点, 解释 (即 X_π) 的需要以人易于理解的方式呈现, 复杂的结构或庞大的规模都可能降低解释的能力.

4 强化学习可解释性研究现状

由于 XRL 涉及的领域广泛, 学者从各领域的角度出发, 导致所提出的方法具有较大差异. 因此, 本节分两步对相关方法进行总结. 首先, 根据技术类别和解释的展现形式, 将现有方法分为视觉和语言辅助解释、策略模仿、可解释模型、逻辑关系提取和策略分解 5 个类别. 然后, 在通用分类方法 (即获取解释的时间、解释的范围) 的基础上, 结合本文所提出的分类依据 (即解释的程度, 面对的关键科学问题), 确定不同类别方法的属性.

4.1 方法的宏观定性

在可解释性领域中, 分类通常基于获取解释的时间和解释的范围两个因素[31]. 具体而言, 根据获取解释的时间, 可解释性方法被分为固有 (intrinsic) 解释和事后 (post-hoc) 解释. 固有解释通过限制模型的表达, 使模型在运行时生成具备可解释性的输出. 例如, 基于较强可解释性的原理和组件 (决策树、线性模型等) 构造模型, 或者通过增加特定过程使模型生成可解释性的输出; 事后解释是通过对模型行为的分析, 总结模型的行为模式, 从而达到解释的目的. 通常而言, 固有解释是策略产生过程中的解释, 特定于某个模型, 而事后解释是策略产生后的解释, 与模型无关. 根据解释的范围, 可解释性方法被分为全局 (global) 解释和局部 (local) 解释, 全局解释忽略模型的微观结构 (如参数、层数等因素), 从宏观层面提供对模型的解释, 局部解释从微观入手, 通过分析模型的微观结构获得对模型的解释.

除上述可解释性的通用分类之外, 本文基于解释与模型和人类感知的符合程度, 将可解释性方法分为数学表达、逻辑表达和感知表达 3 类 (见第 2.4 节). 这 3 类可解释性方法体现出可解释性算法在解释的形式、解释与模型结果的近似程度和解释的直观度等方面的区别. 前文 (见第 3 节) 分析了 XRL 面临的 3 个关键问题, 即环境解释, 任务解释和策略解释. 目前, 单个 XRL 方法难以同时解决 3 类问题, 因此, 我们也以此为依据, 对当前 XRL 方法所着眼的问题进行区分.

综上所述, 本文以“获取解释的时间”“解释的范围”“解释的程度”以及“关键问题”为依据, 对 XRL 方法进行分类 (见表 1). 由于算法多样, 表 1 仅显示大类别算法的特点, 部分算法可能不完全符合.

表 1 XRL 各类方法的属性

方法分类		获取解释的时间	解释的范围	解释的程度	关键问题
视觉和语言辅助解释		事后	局部	感知	策略/任务
策略模仿	行为克隆	事后	全局	逻辑	策略
	逆强化学习	事后	全局	逻辑	环境
可解释模型		固有	全局	逻辑/数学	策略
逻辑关系提取		事后	局部	逻辑	策略
策略分解		固有	局部	逻辑	任务

4.2 视觉和语言辅助解释

视觉和语言辅助解释最初应用于计算机视觉领域, 该方法主要关注待解释模型的输入和输出之间的关联, 具

体而言, 通过量化输入状态对输出动作的影响, 实现对关键状态信息高亮显示, 使人们获得对关键状态的直观了解, 达到辅助人类理解模型行为的目的. 视觉、语言辅助解释是最为直观的解释方法, 但这类方法提供的解释较为抽象, 因此解释的效果依赖于人对直觉和经验.

在以图像为输入的强化学习任务中, 视觉和语言辅助解释通过计算与强化学习任务相关的像素显著性值, 获得对模型决策具有关键影响的重要性掩码, 然后通过在原始输入中叠加该掩码, 实现对输入图像中重要区域的高亮[32,33]. 例如, Annasamy 等人[34]基于键值存储、注意力和可重构嵌入 (reconstructible embeddings), 提出了一种用于 Q 学习的可解释的神经网络体系结构, 该结构以学习到的键值, 注意力图和显著性图等可视化形式提高可解释性. 一些研究引入了决策的时序信息, 解释决策之间的关联性, 从而获得关键状态/决策信息. 例如 Mott 等人[35]提出用于强化学习领域的软注意力模型 (soft attention model). 该模型使用自上而下的软注意力机制, 通过顺序查询其环境视图迫使注意力机制专注于任务相关的信息, 该文章还分析了智能体的学习过程, 发现部分策略在不同游戏中反复出现. Samuel 等人[36]提出一种显著性图算法, 在 Atari 2600 环境中产生对智能体重要特性的可视化, 包括智能体的意图、决策的影响因素、学习过程中的进化等. 实验表明这些解释可以提高人类受试者的推理能力, 证明了该解释传递了有效的知识. 一些研究引入生成模型或对象识别过程以替代人的先验知识. 例如, Wang 等人[37]使用生成网络生成自然语言描述的解释, 提出的模型包括 3 个部分: 特征提取编码器, 特征选择的注意力机制, 自然语言解码器; Iyer 等人[38]在深度强化学习模型中加入对象识别过程, 以对象为基础构成更具有可读性的显著性图.

另外, 一些研究致力于改善可视化的效果和测试解释的准确性. Puri 等人[39]认为, 基于扰动的显著性图经常会突出与智能体行为无关的动作, 因此提出特征和相关属性归因 (specific and relevant feature attribution, SARFA) 方法, 该方法通过平衡显著性图的特异性 (specificity) 和相关性 (relevance) 以产生直观的显著性图. 其中特异性指的是扰动对目标动作预期回报的影响, 相关性指的是与目标动作不相关的特征对其他动作的预期回报的影响. Atrey 等人[40]提出一种基于反事实推理的经验性方法, 测试从显著性图生成的假设, 评估这些假设是否符合 RL 环境的语义; Gottesman 等人[41]通过突出显示对离线策略评估 (off-policy evaluation, OPE) 有重大影响的动作观察 (observation), 制定一套规则来选择关键的动作观察以提交给领域专家进行验证. Huber 等人[42]发现显著性图通常会强调所有可能相关的区域, 导致可视化结果中包含相关度不高的区域, 因此对逐层关联传播 (layer-wise relevance propagation, LRP) 的概念进行了调整, 仅利用每个卷积层中最相关的神经元, 从而生成更具选择性的显著性图.

除此之外, 一些方法基于其他技术实现视觉和语言辅助解释. 例如, Zahavy 等人[43]提出一种在非盲 (non-blind) 情况下分析 DQN 的工具, 通过记录 DQN 最后一层隐层的激活, 使用 t-SNE[44]进行降维和可视化, 发现 DQN 以分层的方式聚合了状态空间; Mishra 等人[45]提出了一个视觉稀疏贝叶斯强化学习框架, 在输入图像状态时感知环境, 将图像编码为特征向量, 用于训练稀疏贝叶斯强化学习模型, 并在这个过程中记录过去经验中对未来决策更重要的记忆图像. 通过这些记忆图像对训练过程产生可视化的解释; Sequeira 等人[46]通过分析智能体和环境的交互信息, 提取有助于解释的关键数据, 以关键帧决策的形式突出了互动的关键时刻.

虽然视觉和语言辅助解释通常能获得较为直观的结果, 但其解释依赖于人对任务的理解和经验. 因此, 面对训练优化不佳的模型, 生成的可视化解释通常可读性差, 导致无法进一步指导模型的优化.

4.3 策略模仿

策略模仿的思想来源于模仿学习[47]方法. 本文将该类方法分为行为克隆 (behavior cloning) 和逆强化学习 (inverse reinforcement learning) 两个研究路径. 行为克隆是指从示教者提供的范例中学习, 具体而言, 范例包含决策的 (s,a) 序列, 行为克隆使用简易的模型结构从复杂模型的范例中获取知识, 达到模仿复杂模型结构输出的效果; 逆强化学习是在给定一个专家策略的前提下, 对强化学习环境的回报函数进行反向求解, 获得与环境相关的知识.

行为克隆的效果直观且操作过程简单, 但一些可解释的模型 (决策树) 无法在环境中被直接训练, 导致行为克隆方法通常只能通过学习 DNN 的 (s,a) 经验获得. 由于决策树的性能稳定, 且能被直接转换为规则模型, 大量方法使用决策树克隆深度模型的行为. 例如, Liu 等人[48]提出 Q 函数模仿学习框架, 并提出线性模型 U 树 (linear model

U-trees, LMUTs), 该方法在 U 树模型的基础上, 将叶子节点变更为一个小型的线性模型, 使之能产生连续的结果. Bastani 等人[49]通过训练决策树策略来进行可验证的强化学习, 为了解决决策树难以训练的问题, 结合了模型压缩和模仿学习的思想, 首先提出 Q-DAGGER 以解决 DAGGER 算法无法利用 Q 函数的问题, 然后提出 VIPER 算法从 DNN 模型中提取出决策树策略, 该方法能达到接近 DQN 类方法的性能. 为了增强树的表达能力, 一些研究者基于新的决策树模型实现行为克隆. 例如, Vasic 等人[50]提出了表达能力更强的决策树模型混合专家树 (mixture of expert trees, MOËT), 对状态空间进行分区, 并针对不同分区构建多个决策树. 一些方法基于新型决策树结构实现行为克隆, 从而获得更好的效果, 例如, 软决策树 (soft decision tree, SDT)[51]是决策树的变体, 沿用了决策树的结构, 其原理是使用一阶运算对状态空间进行划分 (类似于分段函数), 对不同状态空间区域采用不同的一阶策略. 软决策树通常被认为是可解释的, 且理论上能够达到与 DNN 一致的表达能力, 因此文献 [52,53] 使用软决策树策略模仿 DNN 的输出以实现可解释性; Ding 等人[54]在软决策树和离散可分决策树 (discretized differentiable decision trees, DDT) 的基础上, 提出级联决策树 (cascading decision trees, CDT), 在具备较强可解释性的同时实现良好的性能; 另外, 一些研究者在不改变决策树基础结构的前提下对其性能进行优化. 例如, Roth 等人[55]为解决决策树过大的问题, 仅在整体策略的估计折现未来奖赏增加达到一定阈值时才对树进行扩充. 一些方法致力于使用简单模型进行策略模仿, 最大程度控制模型的规模. 例如, Brown 等人[56]利用深度策略产生的专家数据集构建 CART 决策树[57]. Lee[58]基于神经网络策略构建一个基于简单规则的策略. 除此之外, 一些研究采用的方法与行为克隆类似. 例如, Fukuchi 等人[59]提出基于指令的行为解释 (instruction-based behavior explanation, IBE), 使智能体可以重用人类专家给出的指令来加快策略的学习, 自主获取表达式以解释其自身的行为. 之后 Fukuchi 等人[60]改进了 IBE, 使之能适用于动态改变策略的智能体. 另外, 文献 [61,62] 通过遗传编程方法从状态动作轨迹中学习具有更强可解释性的策略代数方程式. 总体而言, 虽然基于行为克隆的方法能够获得性能接近 DRL 的可解释模型, 但其缺点是无法脱离 DRL 经验的约束, 因而可解释模型无法完全替代 DRL 而成为独立的模型.

逆强化学习方法的重心在于发现环境的规律, 具体而言, 通过对模型的强化学习过程的逆向计算拟合环境的知识, 然而, 现有方法虽然能够获得环境相关的知识, 但难以从这些知识中抽象出人可以理解的通用概念. 在 XRL 中, 一些工作借鉴了逆强化学习的思想, 例如, van der Waa 等人[63]根据 RL 的状态转换和预期结果来解释其行为, 首先将基于人的理解对状态和动作进行抽象, 然后通过构建状态转移模型拟合环境的状态转移, 最后通过模拟获得当前策略和派生策略的预期结果. Kaiser 等人[64]使用专家知识构造贝叶斯状态转移模型, 提出一种基于变分推断的学习方案.

4.4 可解释模型

可解释模型是基于易理解的组件构建出的高效模型, 与策略模仿的区别在于, 该模型不依赖于专家 DNN 模型而独立存在. 通常认为, 可解释性领域广泛存在着可理解性-性能的平衡[18,19], 即可解释性越强, 模型性能越差, 反之亦然. 但可理解性-性能的平衡是经验性的结论, 并没有坚实的理论支撑. 因此, 通过精巧的模型设计可能打破这种平衡, 发现易于理解且高性能的模型.

可解释模型可以替代 DNN 模型而存在, 因此在安全敏感应用中具有广泛的应用前景, 是近年来被逐渐重视的研究方向. Jiang 等人[65]基于策略梯度方法和可归纳逻辑编程提出了神经逻辑强化学习 (neural logic reinforcement learning, NLRL) 以表示一阶逻辑在强化学习中的策略, NLRL 在监督任务的可解释性和可概括性方面有显著优势, 而且能诱导解释策略实现接近最优的性能; Abhinav 等人[66,67]提出编程可解释的强化学习 (programmatically interpretable reinforcement learning, PIRL) 框架. 与传统的 DRL 不同在于, PIRL 使用特定领域的高级编程语言 (解决特定领域的问题的非通用语言) 来表示策略, 该方法便于通过符号表示的方法进行验证.

可解释模型有望替代深度模型, 从而解决由深度模型带来的组件不透明特性. 然而, 现有方法存在两类缺点: (1) 可解释的模型性能通常弱于深度模型, 因而难以替代深度模型; (2) 现有可解释模型虽然能够提升模型的组件透明度, 但为了提高性能, 其复杂度仍然较高. 因此, 该类方法仍然存在因复杂度而导致的难以理解.

4.5 逻辑关系提取

逻辑关系提取是将模型所包含的知识映射为人们所能理解的知识, 如因果关系、逻辑过程等. 逻辑关系提取

的方法不仅可以使人从直观上了解模型的决策过程, 还能将模型的知识显式传递给人类, 提升人类对该问题的认识, 指导人的实践.

逻辑关系是一种通过逻辑链理解事件的直观方法, 被认知科学广泛接受, 但往往需要引入人的先验知识. 相关研究中, Madumal 等人[68,69]发现大部分情况下, 人类的行为可以用因果关系描述 (如“A enables B and B causes C”). 继而提出和优化了因果解释的方法, 基于决策树和因果模型来分析事实, 产生中的机会链 (opportunity chains). Topin 等人[70] 提出抽象策略图 (抽象状态的马尔可夫链), 抽象策略图根据价值函数和一系列关键状态转移 (训练期间使用的可能导致策略改变的状态转移) 学习得到, 由于抽象策略图能够简明扼要的对策略决策过程进行表述, 因此可以用来解释智能体的决策. Volodin 等人[71]设计了一种奖励函数, 该函数可以激励代理进行干预以查找因果模型中的错误, 从而改善因果模型.

逻辑关系提取不仅能加深人对模型的理解, 还能梳理模型决策的因果关系, 反向指导人的实践. 然而, 对逻辑关系的提取建立在人对问题有一定理解的基础上, 通过先验知识获取实体, 并通过模型发现实体之间的关联. 因此, 逻辑关系提取仍然存在先验知识需求较多的缺点, 不利于算法的泛化.

4.6 策略分解

策略分解的思想基于分层强化学习 (hierarchical reinforcement learning)[72]. 具体而言, 人对行为的理解通常是策略层面的, 而非动作层面, 时间线上相近的动作通常为了实现相同的子目标. 因此, 通过对时间线上相近的动作进行聚合, 可以将任务过程划分为多个子任务, 进而生成策略层面的解释. 该方法的优势是可以忽略智能体的具体步骤, 直接了解智能体在当前阶段的目标, 进而从策略层面了解模型的行为. 然而, 该方法的缺点也较为明显, 由于复杂问题并不都能被很好地分解, 导致策略分解的方法在复杂任务中的应用受到限制.

相关工作中, Shu 等人[73] 提出了分层可解释性的多任务强化学习框架, 并在 Minecraft 游戏中完成方块的操纵任务 (查找、获取或堆叠某种颜色的方块), 每个顶层策略都包含几个较低级别的操作, 例如“找到方块”→ “拿起方块”→“放下方块”. 由于任务是由人工命令描述的, 而智能体通过这些描述来学习策略, 因此, 其决策都具有人类可理解的含义. Lyu 等人[74]认为在分层决策中, 子任务的可解释性至关重要, 因此, 将符号规划引入强化学习, 提出符号深度强化学习框架 (symbolic deep reinforcement learning, SDRL), 以处理感官输入和符号规划问题, 通过将符号操作与选项相关联, 可以获得任务级的解释; Beyret 等人[75]提出了一个由低级智能体组成的层次化的深度强化学习系统, 低级智能体遵循一个高级智能体的指令在大型动作/状态空间中执行高效的操作, 因此, 高级智能体学习到的是与环境和任务相关的高度抽象的知识, 对于人来说是可以解释的. 另外, 一些研究通过奖励分解的方法实现子任务的划分. 例如, Juozapaitis 等人[76]基于奖励分解实现可解释性, 提出最小需求解释 (minimal sufficient explanation, MSE), 将奖励分解为语义上有意义的奖励类型的总和.

对于人而言, 决策之间的关联和对未来的预期是人理解策略的基础, 但模型所直观呈现的往往是单步的决策, 这一差异导致了理解的障碍. 而策略分解的优点在于将模型的序列决策抽象为具有宏观意义的目标, 使决策之间的关系更加明确. 然而, 策略分解的缺点也相对明显, 对于复杂任务, 决策可能并非来源于单一的因素, 而是多因素聚合的, 导致智能体并不具有明确的子目标, 并且任务实现的过程是多样化的. 因此, 策略分解难以将复杂任务分解, 导致该类方法的应用受到限制.

5 潜在研究方向

XRL 是强化学习领域必然面临的问题, 虽然目前已经有部分工作致力于此, 但 XRL 的研究仍然处于基础阶段, 无法满足在安全敏感领域的应用需求. 具体来说, 目前 XRL 仍然存在如下的关键问题亟待解决.

5.1 对环境和任务的解释

本文提出了 XRL 领域所面临的独有问题 (见第 3 节), 即环境解释、任务解释、策略解释. 3 个问题本质上以强化学习的不同要素作为出发点, 即环境、任务以及智能体. 因此, 环境解释、任务解释和策略解释三者具有强耦合.

在对 XRL 方法的总结中 (见第 4 节), 众多方法致力于对智能体策略的解释, 极少数方法涵盖对环境和任务的解释. 究其原因, 一方面, XRL 目前尚缺乏完善的理论体系, 现有研究大部分基于 XAI 中已成熟的算法, 因此大部

分方法缺乏对 XRL 独有问题的考虑; 另一方面, 大部分研究建立在人对环境和任务具有较强的理解的基础上, 例如, 当智能体的能力超过人类时, 人们需要通过智能体获得与环境与任务相关的知识, 进而指导人的实践. 由于 RL 中的智能体、环境、任务之间的强关联性, 使得策略必然涉及与环境和任务相关的知识. 例如, 在复杂环境中, 面对黑盒的环境和模糊的目标, 人对环境转移和任务过程的理解通常不够充分, 进而导致智能体策略难以被理解. 尤其是当模型的性能超越人时, 基于人对环境和任务的经验的解释则更不足以理解模型的策略. 因此, 仅对策略进行解释无法解决 XRL 的问题.

目前, XRL 领域存在对环境和任务进行建模的方法. 例如, 使用逆强化学习 (见第 4.3 节) 的方法对环境进行建模, 以及使用策略分解 (见第 4.6 节) 的方法对任务进行细分. 然而, 现阶段逆强化学习获得的知识无法完全以可解释的方式呈现, 因此难以显著提高人的对环境的理解, 而策略分解通常局限于具有明显层次或顺序结构的任务, 面对不具备明显层次结构的复杂问题难以求解.

总体而言, 环境解释和任务解释是与策略解释并驾齐驱的研究问题, 但目前较少有针对这两个问题的研究工作. 因此, 对环境和任务的解释是 XRL 未来的两大重要研究问题.

5.2 统一评估标准

统一的评估标准是对 XRL 方法的量化, 该评估标准的构建必须建立在对可解释性技术的宏观理解上, 用以精确地理解强化学习的运行机制, 并从不同方法中抽象出描述可解释性的共有属性, 使不同 XRL 方法之间产生可比性, 进而实现跨方法的评估, 并推动 XRL 的研究的统一.

在 XAI 领域, 已有学者尝试提出较为通用的评价标准, 如张长水在 2021 北京智源大会机器学习论坛中提出基于输入和输出的一致性定义了可解释性 (https://hub.baai.ac.cn/view/8639), 即 $dist\left(f^{(a)}(x), f^{(b)}(x)\right) < \delta$. 其中 $dist(\cdot)$ 为距离函数, $f^{(\cdot)}(x)$ 为样本 x 在黑盒系统中的输出, 当黑盒系统 A, B 对样本 x 满足上式, 则系统 A 对系统 B 的可解释程度被评价为 δ (δ -interpretable). 该标准将可解释性视为相对的概念, 使用可量化的方式定义了可解释性.

总而言之, 目前尚缺乏对可解释性的统一评估标准. 在 XRL 领域, 统一评估标准的建立需要满足如下要求: (1) 评估标准与具体技术解耦合; (2) 实现对可解释性的统一描述, 构建对 XRL 的量化方法; (3) 评估指标不仅考虑系统的输出, 同时考虑系统推理的内因. 因此, XRL 仍然缺乏具有普适性的评估指标.

6 总 结

本文以 XRL 的问题为中心, 讨论了该领域的基础问题, 并对现有方法进行总结. 由于目前在 XRL 领域, 乃至整个 XAI 领域尚未形成完整、统一的共识, 导致不同研究的基础观点存在较大差异, 难于类比. 本文针对该领域缺乏一致认知的问题, 进行了较为深入的研究工作. 首先, 本文参考 XRL 领域的父问题——XAI, 收集 XAI 领域的现有观点, 并整理出 XAI 领域较为通用的认识; 其次, 以 XAI 领域的定义为基础, 讨论 XAI 与 XRL 面临的共同问题; 然后, 结合强化学习自身的特点, 提出 XRL 面临的独有问题; 最后, 总结了相关的研究方法, 并对相关方法进行分类. 分类中包括作者明确指出为 XRL 的方法, 也包括作者虽未着重强调, 但实际对 XRL 有重要意义的方法. XRL 目前尚处于初步阶段, 因此存在大量亟待解决的问题. 本文重点提出环境和任务的解释、统一的评估标准两类问题. 本文认为这两类问题是为类 XRL 领域的基石, 是值得重视的研究领域.

References:

[1] Janai J, Güney F, Behl A, Geiger A. Computer vision for autonomous vehicles: Problems, datasets and state of the art. Foundations and Trends® in Computer Graphics and Vision, 2020, 12(1–3): 1–308. [doi: 10.1561/0600000079]

[2] Otter DW, Medina JR, Kalita JK. A survey of the usages of deep learning for natural language processing. IEEE Trans. on Neural Networks and Learning Systems, 2021, 32(2): 604–624. [doi: 10.1109/TNNLS.2020.2979670]

[3] Polydoros AS, Nalpantidis L. Survey of model-based reinforcement learning: Applications on robotics. Journal of Intelligent & Robotic Systems, 2017, 86(2): 153–173. [doi: 10.1007/s10846-017-0468-y]

[4] Jefferson J, McDonald AD. The autonomous vehicle social network: Analyzing tweets after a recent Tesla autopilot crash. Proc. of the Human Factors and Ergonomics Society Annual Meeting, 2019, 63(1): 2071–2075. [doi: 10.1177/1071181319631510]

[5] Dikmen M, Burns C. Trust in autonomous vehicles: The case of Tesla Autopilot and Summon. In: Proc. of the 2017 IEEE Int'l Conf. on Systems, Man, and Cybernetics (SMC). Banff: IEEE, 2017. 1093–1098. [doi: 10.1109/SMC.2017.8122757]

[6] Collins S, Ruina A, Tedrake R, Wisse M. Efficient bipedal robots based on passive-dynamic walkers. Science, 2005, 307(5712): 1082–1085. [doi: 10.1126/science.110779]

[7] Mnih V, Kavukcuoglu K, Silver D, *et al*. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529–533. [doi: 10.1038/nature14236]

[8] Silver D, Huang A, Maddison CJ, *et al*. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529(7587): 484–489. [doi: 10.1038/nature16961]

[9] Johannink T, Bahl S, Nair A, Luo JL, Kumar A, Loskyll M, Ojea JA, Solowjow E, Levine S. Residual reinforcement learning for robot control. In: Proc. of the 2019 Int'l Conf. on Robotics and Automation (ICRA). Montreal: IEEE, 2019. 6023–6029. [doi: 10.1109/ICRA.2019.8794127]

[10] Vinyals O, Babuschkin I, Czarnecki WM, *et al*. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 2019, 575(7782): 350–354. [doi: 10.1038/s41586-019-1724-z]

[11] Kassahun Y, Yu BB, Tibebu AT, Stoyanov D, Giannarou S, Metzen JH, Vander Poorten E. Surgical robotics beyond enhanced dexterity instrumentation: A survey of machine learning techniques and their role in intelligent and autonomous surgical actions. Int'l Journal of Computer Assisted Radiology and Surgery, 2016, 11(4): 553–568. [doi: 10.1007/s11548-015-1305-z]

[12] Kiran BR, Sobh I, Talpaert V, Mannion P, Al Sallab AA, Yogamani S, Pérez P. Deep reinforcement learning for autonomous driving: A survey. IEEE Trans. on Intelligent Transportation Systems, 2022, 23(6): 4909–4926. [doi: 10.1109/TITS.2021.3054625]

[13] Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, Zaremba W. OpenAI gym. arXiv:1606.01540, 2016.

[14] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego, 2015.

[15] Lin YC, Hong ZW, Liao YH, Shih ML, Liu MY, Sun M. Tactics of adversarial attack on deep reinforcement learning agents. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. Melbourne: IJCAI.org, 2017. 3756–3762. [doi: 10.24963/ijcai.2017/525]

[16] Gleave A, Dennis M, Wild C, Kant N, Levine S, Russell S. Adversarial policies: Attacking deep reinforcement learning. In: Proc.of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[17] Páez A. The pragmatic turn in explainable artificial intelligence (XAI). Minds and Machines, 2019, 29(3): 441–459. [doi: 10.1007/s11023-019-09502-w]

[18] Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 2018, 6: 52138–52160. [doi: 10.1109/ACCESS.2018.2870052]

[19] Tjoa E, Guan CT. A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Trans. on Neural Networks and Learning Systems, 2021, 32(11): 4793–4813. [doi: 10.1109/TNNLS.2020.3027314]

[20] Byrne RMJ. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: IJCAI.org, 2019. 6276–6282. [doi: 10.24963/ijcai.2019/876]

[21] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 2020, 58: 82–115. [doi: 10.1016/j.inffus.2019.12.012]

[22] Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI—Explainable artificial intelligence. Science Robotics, 2019, 4(37): eaay7120. [doi: 10.1126/scirobotics.aay7120]

[23] Das A, Rad P. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. arXiv:2006.11371, 2020.

[24] Shih A, Darwiche A, Choi A. Verifying binarized neural networks by Angluin-style learning. In: Proc. of the 22nd Int'l Conf. on Theory and Applications of Satisfiability Testing. Lisbon: Springer, 2019. 354–370. [doi: 10.1007/978-3-030-24258-9_25]

[25] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Washington: ACM, 2016. 1135–1144. [doi: 10.1145/2939672.2939778]

[26] Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: An overview of interpretability of machine learning. In: Proc. of the 5th IEEE Int'l Conf. on Data Science and Advanced Analytics (DSAA). Turin: IEEE, 2018. 80–89. [doi: 10.1109/DSAA.2018.00018]

[27] Lipton ZC. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 2018, 16(3): 31–57. [doi: 10.1145/3236386.3241340]

[28] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608, 2017.

[29] Doran D, Schulz S, Besold TR. What does explainable AI really mean? A new conceptualization of perspectives. In: Proc. of the 1st Int'l

Workshop on Comprehensibility and Explanation in AI and ML 2017 Co-located with the 16th Int'l Conf. of the Italian Association for Artificial Intelligence. Bari: CEUR-WS.org, 2017.

[30] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M. Playing Atari with deep reinforcement learning. arXiv:1312.5602, 2013.

[31] Puiutta E, Veith EMSP. Explainable reinforcement learning: A survey. In: Proc. of the 4th Int'l Cross-Domain Conf. for Machine Learning and Knowledge Extraction. Dublin: Springer, 2020. 77–95. [doi: 10.1007/978-3-030-57321-8_5]

[32] Mnih V, Heess N, Graves A, Kavukcuoglu K. Recurrent models of visual attention. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2204–2212.

[33] Saltelli A. Sensitivity analysis for importance assessment. Risk Analysis, 2002, 22(3): 579–590. [doi: 10.1111/0272-4332.00040]

[34] Annasamy RM, Sycara K. Towards better interpretability in deep Q-networks. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI, 2019. 4561–4569. [doi: 10.1609/aaai.v33i01.33014561]

[35] Mott A, Zoran D, Chrzanowski M, Wierstra D, Rezende DJ. Towards interpretable reinforcement learning using attention augmented agents. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 1107.

[36] Greydanus S, Koul A, Dodge J, Fern A. Visualizing and understanding Atari agents. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholmsmässan: PMLR, 2018. 1787–1796.

[37] Wang XZ, Yuan SC, Zhang H, Lewis M, Sycara K. Verbal explanations for deep reinforcement learning neural networks with attention on extracted features. In: Proc. of the 28th IEEE Int'l Conf. on Robot and Human Interactive Communication (RO-MAN). New Delhi: IEEE, 2019. 1–7. [doi: 10.1109/RO-MAN46459.2019.8956301]

[38] Iyer R, Li YZ, Li HA, Lewis M, Sundar R, Sycara K. Transparency and explanation in deep reinforcement learning neural networks. In: Proc. of the 2018 AAAI/ACM Conf. on AI, Ethics, and Society. New Orleans: ACM, 2018. 144–150. [doi: 10.1145/3278721.3278776]

[39] Puri N, Verma S, Gupta P, Kayastha D, Deshmukh S, Krishnamurthy B, Singh S. Explain your move: Understanding agent actions using specific and relevant feature attribution. In: Proc. of the 8th Int'l Conf. on Learning Representations (ICLR). Addis Ababa: OpenReview.net, 2020.

[40] Atrey A, Clary K, Jensen DD. Exploratory not explanatory: Counterfactual analysis of saliency maps for deep reinforcement learning. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[41] Gottesman O, Futoma J, Liu Y, Parbhoo S, Celi LA, Brunskill E, Doshi-Velez F. Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions. In: Proc. of the 37th Int'l Conf. on Machine Learning. PMLR, 2020. 3658–3667.

[42] Huber T, Schiller D, André E. Enhancing explainability of deep reinforcement learning through selective layer-wise relevance propagation. In: Proc. of the 42nd German Conf. on AI, KI 2019: Advances in Artificial Intelligence. Kassel: Springer, 2019. 188–202. [doi: 10.1007/978-3-030-30179-8_16]

[43] Zahavy T, Ben-Zrihem N, Mannor S. Graying the black box: Understanding DQNs. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York: JMLR.org, 2016. 1899–1908.

[44] van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008, 9(86): 2579–2605.

[45] Mishra I, Dao G, Lee M. Visual sparse Bayesian reinforcement learning: A framework for interpreting what an agent has learned. In: Proc. of the 2018 IEEE Symp. Series on Computational Intelligence (SSCI). Bangalore: IEEE, 2018. 1427–1434. [doi: 10.1109/SSCI.2018.8628887]

[46] Sequeira P, Gervasio M. Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations. Artificial Intelligence, 2020, 288: 103367. [doi: 10.1016/j.artint.2020.103367]

[47] Ross S, Bagnell D. Efficient reductions for imitation learning. In: Proc. of the 13th Int'l Conf. on Artificial Intelligence and Statistics. Sardinia: JMLR, 2010. 661–668.

[48] Liu GL, Schulte O, Zhu W, Li QC. Toward interpretable deep reinforcement learning with linear model U-trees. In: Proc. of the European Conf. on Machine Learning and Knowledge Discovery in Databases. Dublin: Springer, 2019. 414–429. [doi: 10.1007/978-3-030-10928-8_25]

[49] Bastani O, Pu YW, Solar-Lezama A. Verifiable reinforcement learning via policy extraction. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 2499–2509.

[50] Vasic M, Petrovic A, Wang KY, Nikolic M, Singh R, Khurshid S. MoËT: Interpretable and verifiable reinforcement learning via mixture of expert trees. arXiv:1906.06717, 2019.

[51] Frosst N, Hinton GE. Distilling a neural network into a soft decision tree. In: Proc. of the 1st Int'l Workshop on Comprehensibility and Explanation in AI and ML 2017 Co-located with 16th Int'l Conf. of the Italian Association for Artificial Intelligence. Bari: CEUR-WS.org, 2017.

[52] Coppens Y, Efthymiadis K, Lenaerts T, Nowe A. Distilling deep reinforcement learning policies in soft decision trees. In: Proc. of the 2019 IJCAI Workshop on Explainable Artificial Intelligence. Cotai, 2019. 1–6.

[53] Dahlin N, Kalagarla KC, Naik N, Jain R, Nuzzo P. Designing interpretable approximations to deep reinforcement learning with soft decision trees. arXiv:2010.14785, 2020.

[54] Ding ZH, Hernandez-Leal P, Ding GW, Li CJ, Huang RT. CDT: Cascading decision trees for explainable reinforcement learning. arXiv:2011.07553, 2020.

[55] Roth AM, Topin N, Jamshidi P, Veloso M. Conservative Q-improvement: Reinforcement learning for an interpretable decision-tree policy. arXiv:1907.01180, 2019.

[56] Brown A, Petrik M. Interpretable reinforcement learning with ensemble methods. arXiv:1809.06995, 2018.

[57] Loh WY. Classification and regression trees. WIREs Data Mining and Knowledge Discovery, 2011, 1(1): 14–23. [doi: 10.1002/widm.8]

[58] Lee JH. Complementary reinforcement learning towards explainable agents. arXiv:1901.00188, 2019.

[59] Fukuchi Y, Osawa M, Yamakawa H, Imai M. Autonomous self-explanation of behavior for interactive reinforcement learning agents. In: Proc. of the 5th Int'l Conf. on Human Agent Interaction. Bielefeld: ACM, 2017. 97–101. [doi: 10.1145/3125739.3125746]

[60] Fukuchi Y, Osawa M, Yamakawa H, Imai M. Application of instruction-based behavior explanation to a reinforcement learning agent with changing policy. In: Proc. of the 24th Int'l Conf. on Neural Information Processing. Guangzhou: Springer, 2017. 100–108. [doi: 10.1007/978-3-319-70087-8_11]

[61] Hein D, Udluft S, Runkler TA. Interpretable policies for reinforcement learning by genetic programming. Engineering Applications of Artificial Intelligence, 2018, 76: 158–169. [doi: 10.1016/j.engappai.2018.09.007]

[62] Hein D, Udluft S, Runkler TA. Generating interpretable reinforcement learning policies using genetic programming. In: Proc. of the 2019 Genetic and Evolutionary Computation Conf. Companion. Boston: ACM, 2019. 23–24. [doi: 10.1145/3319619.3326755]

[63] van der Waa J, van Diggelen J, van den Bosch K, Neerincx M. Contrastive explanations for reinforcement learning in terms of expected consequences. In: Proc. of the 2018 Workshop on Explainable AI on the IJCAI Conf. Stockholm, 2018. 37.

[64] Kaiser M, Otte C, Runkler TA, Ek CH. Interpretable dynamics models for data-efficient reinforcement learning. In: Proc. of the 27th European Symp. on Artificial Neural Networks. Bruges, 2019.

[65] Jiang ZY, Luo S. Neural logic reinforcement learning. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 3110–3119.

[66] Verma A. Verifiable and interpretable reinforcement learning through program synthesis. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI, 2019. 9902–9903. [doi: 10.1609/aaai.v33i01.33019902]

[67] Verma A, Murali V, Singh R, Kohli P, Chaudhuri S. Programmatically interpretable reinforcement learning. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholmsmässan: PMLR, 2018. 5052–5061.

[68] Madumal P, Miller T, Sonenberg L, Vetere F. Explainable reinforcement learning through a causal lens. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 2493–2500. [doi: 10.1609/aaai.v34i03.5631]

[69] Madumal P, Miller T, Sonenberg L, Vetere F. Distal explanations for explainable reinforcement learning agents. arXiv:2001.10284, 2020.

[70] Topin N, Veloso M. Generation of policy-level explanations for reinforcement learning. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI, 2019. 2514–2521. [doi: 10.1609/aaai.v33i01.33012514]

[71] Volodin S, Wichers N, Nixon J. Resolving spurious correlations in causal models of environments via interventions. arXiv:2002.05217, 2020.

[72] Dietterich TG. Hierarchical reinforcement learning with the MAXQ value function decomposition. Journal of Artificial Intelligence Research, 2000, 13: 227–303. [doi: 10.1613/jair.639]

[73] Shu TM, Xiong CM, Socher R. Hierarchical and interpretable skill acquisition in multi-task reinforcement learning. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[74] Lyu DM, Yang FK, Liu B, Gustafson S. SDRL: Interpretable and data-efficient deep reinforcement learning leveraging symbolic planning. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI. 2019. 2970–2977. [doi: 10.1609/aaai.v33i01.33012970]

[75] Beyret B, Shafti A, Faisal AA. Dot-to-dot: Explainable hierarchical reinforcement learning for robotic manipulation. In: Proc. of the 2019 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Macao: IEEE, 2019. 5014–5019. [doi: 10.1109/IROS40897.2019.8968488]

[76] Juozapaitis Z, Koul A, Fern A, Erwig M, Doshi-Velez F. Explainable reinforcement learning via reward decomposition. In: Proc. of the 2019 IJCAI/ECAI Workshop on Explainable Artificial Intelligence. 2019. 47–53.

刘潇(1991－), 男, 博士生, 主要研究领域为可解释强化学习, 机器学习, 计算机视觉.

庄韫恺(1990－), 男, 博士生, 主要研究领域为多智能体系统, 强化学习, 博弈论.

刘书洋(1999－), 男, 硕士生, 主要研究领域为强化学习, 可解释强化学习.

高阳(1972－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为强化学习, 多智能体系统, 计算机视觉, 大数据分析.