面向强化学习的可解释性研究综述

曹宏业[1)] 刘 潇[2)] 董绍康[1)] 杨尚东[3)] 霍 静[1)]
李文斌[1)] 高 阳[1)]

[1)](计算机软件新技术全国重点实验室(南京大学) 南京 210023)
[2)](湘潭大学自动化与电子信息学院 湖南 湘潭 411105)
[3)](南京邮电大学计算机学院 南京 210023)

摘 要 强化学习作为机器学习的一种范式,因其强大的策略试错学习能力,受到关注.随着深度学习的融入,强化学习方法在许多复杂的控制任务中取得了巨大成功.然而,深度强化学习网络作为黑盒模型,其缺乏可解释性所带来的不安全、不可控及难理解等问题限制了强化学习在诸如自动驾驶、智慧医疗等关键领域中的发展.为了解决这一问题,科研人员开展了对强化学习可解释性的研究.然而,这些研究开展相对较晚,且缺少针对多智能体强化学习可解释性方法的系统性总结,同时,可解释性的定义存在人为主观性,导致系统性面向强化学习过程的可解释性研究较为困难.本文对当前强化学习的可解释性研究工作进行了全面的整理与总结.首先,对强化学习的可解释性进行定义并总结了相关评估方法.随后,基于马尔可夫决策过程,划分了行为级解释、特征级解释、奖励级解释及策略级解释四个类别.此外,在每个类别中,分析了单智能体及多智能体的策略解释方法,并特别关注可解释性研究中的人为因素,描述了人机交互式的解释方法.最后,对当前强化学习可解释性研究面临的挑战以及未来的研究方向进行总结与展望.

关键词 强化学习;可解释性;机器学习;人工智能;马尔可夫决策过程
中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2024.01853

A Survey of Interpretability Research Methods for Reinforcement Learning

CAO Hong-Ye[1)] LIU Xiao[2)] DONG Shao-Kang[1)] YANG Shang-Dong[3)] HUO Jing[1)]
LI Wen-Bin[1)] GAO Yang[1)]

[1)](*State Key Laboratory for Novel Software Technology* (*Nanjing University*), *Nanjing* 210023)
[2)](*School of Automation and Electronic Information*, *Xiangtan University*, *Xiangtan*, *Hunan* 411105)
[3)](*School of Computer Science*, *Nanjing University of Posts and Telecommunications*, *Nanjing* 210023)

Abstract Reinforcement learning, as a machine learning paradigm, is garnering increasing attention due to its robust trial-and-error learning capabilities. With the integration of deep learning, reinforcement learning methods have achieved remarkable success in complex control tasks of real-world. However, the lack of interpretability in deep reinforcement learning networks, stemming from their black-box nature, presents challenges such as insecurity, lack of control, and difficulty in comprehension. This limitation hampers the progress of reinforcement learning in critical domains like autonomous driving and intelligent healthcare. To tackle this issue, researchers have undertaken extensive studies in the field of explainable reinforcement

收稿日期:2023-08-30;在线发布日期:2024-04-19.本课题得到科技创新2030—"新一代人工智能"重大项目(2021ZD0113303)、国家自然科学基金(62192783,62276128,62276142,62206133)、南京大学计算机软件新技术国家重点实验室资助项目(KFKT2022B12)资助.
曹宏业,博士研究生,主要研究领域为强化学习、可解释性强化学习.E-mail:hongyecao528@gmail.com. **刘 潇**,博士,讲师,主要研究领域为强化学习、可解释性强化学习. **董绍康**,博士,主要研究领域为强化学习. **杨尚东**,博士,讲师,主要研究领域为强化学习. **霍 静**,博士,准聘副教授,主要研究领域为机器学习. **李文斌**,博士,副研究员,主要研究领域为机器学习. **高 阳**(通信作者),博士,教授,主要研究领域为强化学习.E-mail:gaoy@nju.edu.cn.

learning. Nevertheless, these studies are relatively recent and lack a systematic summary of explainable methods tailored to multi-agent reinforcement learning. Moreover, the definition of interpretability carries subjective elements, making it more challenging to comprehensively categorize explainable research targeting the reinforcement learning process. This article provides a comprehensive review and synthesis of the current state of interpretability research in reinforcement learning. To commence, the article establishes a definition for the interpretability of reinforcement learning and outlines relevant evaluation methods. Subsequently, rooted in Markov decision processes, the article categorizes interpretability into four classes: action-level explanation, feature-level explanation, reward-level explanation, and policy-level explanation. Regarding the inclusion of action factors in action-level explanations within reinforcement learning, methods for action-level explanations can be categorized based on the extent to which they explain the decision-making action of intelligent agents. These categories include self-explanatory model construction, formalized explanation methods, and generative explanation methods in reinforcement learning. In terms of incorporating state factors into feature-level explanations, methods in the feature-level explanations category are classified based on the importance of reinforcement learning state features and the form of explanation. This includes interaction trajectory explanation methods and key feature visualization methods. When it comes to incorporating reward factors into reward-level explanations, methods in the reward-level explanations category are categorized based on the impact of reward feedback on policy effectiveness and different task scenarios. This includes reward decomposition methods and reward shaping methods. Regarding the inclusion of learning strategies into policy-level explanations, methods in the policy-level explanations category are further classified into strategy decomposition methods and strategy aggregation methods, considering the hierarchical relationship of strategy explanations. These four categories of methods encompass the entire Markov decision process of an agent, spanning action execution, state transitions, reward computations, and policy learning. Furthermore, the description of each method category takes into account the progression of interpretability methods from single-agent reinforcement learning to multi-agent reinforcement learning, addressing the interpretability of more complex multi-agent environments. This includes analyzing and discussing issues such as credit assignment, collaborative cooperation, and adversarial games in the context of explainable methods, effectively bridging the gap in interpretability research methods for multi-agent reinforcement learning. Additionally, considering the subjective factors of human interpretability research, this paper provides a comprehensive summary and organization of human-machine interactive interpretability research, emphasizing the involvement of humans in the interpretability. Finally, the article concludes by summarizing the current challenges in interpretability research for reinforcement learning and offering prospects for future research directions.

Keywords reinforcement learning; interpretability; machine learning; artificial intelligence; Markov decision process

1 引 言

强化学习[1](Reinforcement Learning, RL)是一种针对时序决策问题的人工智能(Artificial Intelligence, AI)解决方法. 作为机器学习[2](Machine Learning, ML)的一种范式,不同于监督式学习以及无监督式学习,强化学习通过智能体(agent)模型与环境进行交互式探索,模型根据环境反馈的奖励信息来指导策略的学习. 因其在博弈对

抗[3]、自动控制[4]以及游戏[5]等领域中显著的应用效果，强化学习受到了越来越多的关注. 与此同时，随着深度学习的快速发展，深度强化学习范式[6]被提出去解决更加复杂的决策问题. 通过将深度神经网络强大的表征能力与强化学习的时序决策能力结合在一起，深度强化学习实现了在游戏AI[7]、机械控制[8]以及大规模语言模型的训练[9]上性能的巨大提升.

虽然强化学习得到了蓬勃的发展，但是基于深度神经网络模型的策略网络是一个内部执行逻辑未知的黑盒模型[10-11]. 策略模型难以被人理解，限制了强化学习在自动驾驶、智慧医疗及军事控制等关键领域中的发展[12-13]. 仅凭借输入以及输出的结果，无法解释策略模型的执行逻辑以及保证决策执行的安全性. 例如，在Google团队使用强化学习技术训练的AlphaGo围棋模型[14]中，所执行的部分棋路无法归纳为定式. OpenAI公司借助与人类反馈的强化学习技术搭建的对话工具ChatGPT会发表一些不符合人类思维逻辑的言论[15]. 当具有不确定性的强化学习策略应用到医疗诊断、自动驾驶以及安全防务等关键任务上时，不确定的因素将产生巨大的安全隐患. 因此，亟需对强化学习的可解释性开展深入研究以推动强化学习在安全可信保证下快速发展.

针对不同的环境（单智能体、多智能体环境）以及任务（协同对抗、知识迁移以及人机交互等），进化强化学习[16-17]、自驱动强化学习[18]以及模仿学习[19]等强化学习相关方法被提出. 进化强化学习将遗传算法与强化学习相结合，通过自然进化式优化的方式提升策略性能以及多样性. 自驱动强化学习通过智能体内部的处理机制，对环境的感知和外部奖励信号进行处理，而对于其内部机制往往是在基于人类的理解基础之上进行构建. 模仿学习则是从专家示例中直接进行策略学习. 值得注意的是，无论是专家经验的学习还是自然进化的选择，这些方法的提出受到了人类因素的启发，如何针对这些方法的处理机制进行解释是提升强化学习性能以及可靠性的重要思路.

在面向人工智能的可解释性研究方法（eXplainable Artificial Intelligence，XAI）之中，可以分为模型内置（ante-hoc）可解释性方法以及事后（post-hoc）可解释性方法[20-21]，模型内置可解释性方法旨在模型训练之前，通过决策树搭建、线性回归以及规则模型构建等方式实现模型的事前解释. 事后可解释性方法则是在模型训练后对模型的重点特征进行分析，以达到对模型的事后行为溯因. 然而，区别于传统的人工智能可解释性的研究，由于强化学习面向时序决策的特点，导致了对其进行解释时无法直接借用传统人工智能可解释性研究的成果. 因此，针对强化学习策略的解释需要在动作解析、奖励分解以及决策溯因等方面进行深入研究. 近年来，研究者们提出了一系列的可解释性强化学习（eXplainable Reinforcement Learning，XRL）方法[22-25]. 由于视角和侧重不同导致巨大研究差异，且研究方法迭代迅速，有必要及时对强化学习的可解释性研究进行系统化以及科学化的整理、归纳与总结. 然而，现有强化学习可解释性研究综述[22-26]仍然存在不完善之处. 首先，现有工作没有严格按照强化学习模型的范式进行分类与整理，导致无法与传统的人工智能可解释性方法研究区分开来；其次，缺少针对多智能体强化学习的可解释性研究方法的总结，导致内容不够全面；最后，现有工作忽略了面向人机交互的可解释性方法，未充分考虑人类行为在可解释性研究中的重要性.

因此，本文提出了一个更加全面系统化的面向强化学习的可解释性研究综述. 本篇综述严格按照强化学习范式，根据其遵循的马尔可夫决策过程进行分类. 如图1所示，针对强化学习过程中的关键因子：动作、状态、奖励以及学习的策略，分为行为级解释、特征级解释、奖励级解释及策略级解释四个类别. 具体来讲，针对强化学习中的动作因子归纳为行为级解释，在行为级解释方法类别中根据智能体决策行为的解释程度，分为自解释模型构建方法、形式化解释方法以及生成式解释方法. 针对状态因子归纳为特征级解释，在特征级解释方法类别中，根据强化学习状态特征的重要性以及解释的呈现形式，分为交互轨迹解释方法以及关键特征可视化方法. 针对奖励因子归纳为奖励级解释，在奖励级解释方法类别中，根据奖励对策略的影响效果以及不同的任务场景，分为奖励分解方法以及奖励塑造方法. 针对学习的策略归纳为策略级解释，在策略级解释方法类别中，根据策略解释的层次关系，又分为策略分解方法以及策略聚合方法. 上述四个类别的方法涵盖了智能体从动作执行、状态转移、奖励计算以及策略学习的整个马尔可夫决策过程. 与此同时，在每类方法的描述中，按照从单智能体强化学习到多智能体强化学习的顺序进行可解释性方法的描述，面向更加复杂的多智能体环境，针对多智能体强化

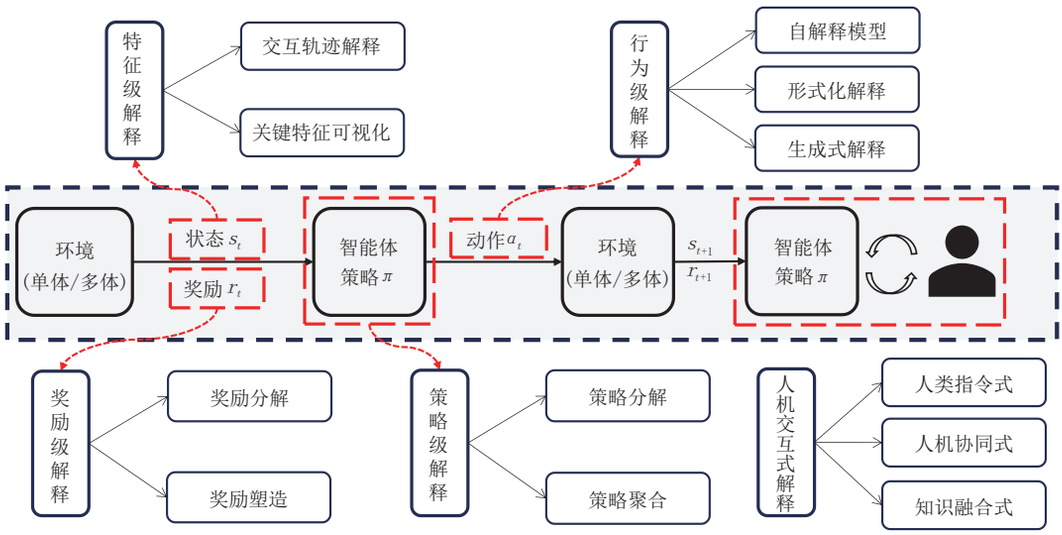

图1 面向强化学习的可解释性研究架构图(其中环境包含有面向单智能体(单体)以及多智能体(多体)的环境)

学习研究中的信用分配、协同合作以及对抗博弈等问题进行解释性方法的分析与思考,有效地填补了面向多智能体强化学习的可解释性研究方法综述的空白. 最后,考虑到可解释性研究中人类主观性的因素,本文从人在回路的角度出发,归纳整理了人机交互式可解释性研究内容.

综上所述,本篇综述的整体组织架构为:在引言部分,对文章的整体内容进行描述. 在第二节,介绍了强化学习的基础背景知识,并且对可解释性人工智能的相关背景知识进行说明. 随后,在第三节中给出了面向强化学习的可解释性定义以及现有的可解释性评估标准. 在第四节中,详细描述了面向强化学习的可解释性研究的四类方法以及人机交互式可解释性方法. 并且,在第五节中对当前强化学习可解释性研究面临的挑战和未来研究方向进行了分析与总结. 最后,在第六节中给出了本篇综述的最终结论.

2 背景知识

2.1 强化学习

在强化学习框架中,智能体与环境交互的过程被形式化定义为马尔可夫决策过程[27,28](Markov Decision Process, MDP). 在马尔可夫决策过程中,智能体与未知环境进行交互,根据策略(policy)执行动作(action),动作执行改变环境的状态并得到奖励(reward)反馈,随着时间而积累的奖励值被称为回报(return). 标准的MDP被定义为一个五元组:$M=\langle S, A, T, r, \gamma \rangle$. 其中$S$表示状态空间,$A$表示动作空间,$T(s'|s,a)$代表一个状态转移动态模型,$r(s,a)$是奖励函数,$\gamma \in [0,1)$是折扣因子. 强化学习的目标是去学习一个策略$\pi: S \times A \to [0,1]$,以最大化期望折扣回报$\eta_M(\pi) := \mathbb{E}_{s_0 \sim \mu_0,\, s_t \sim T,\, a_t \sim \pi}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$. 值函数$V_M^{\pi}(s) := \mathbb{E}_{s_t \sim T,\, a_t \sim \pi}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s]$表示策略$\pi$在状态$s$下的期望折扣回报. 根据策略学习过程中所迭代的对象不同,可将现有主流的强化学习算法分为基于值(value-based)的方法和基于策略(policy-based)的方法.

2.1.1 基于值的强化学习方法

基于值的方法[29-30]旨在使用状态-动作函数Q^{π}来提升策略π的性能,动作-价值函数的定义如下所示:

$$Q^{\pi}(s,a) := \mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_t = s, a_t = a] \quad (1)$$

动作-价值函数用来计算智能体在特定状态下执行一个动作后的预期折扣回报. 最优策略π^*的满足条件为:

$$Q^{\pi^*}(s,a) \geqslant Q^{\pi}(s,a), \forall \pi, s \in S, a \in A \quad (2)$$

同时在强化学习中,状态价值函数V^{π}表示在特定状态下智能体能够获得回报的期望值,可由动作价值函数推导出:

$$V^{\pi}(s) = \pi(a|s) Q^{\pi}(s,a) \quad (3)$$

由以上公式可知,最优动作价值函数与最优状态价值函数对应的是同一个最优策略,因此,最优状态价值函数定义为:

$$V^{\pi^*}(s) = \pi^*(a|s) Q^{\pi^*}(s,a) \quad (4)$$

使用贝尔曼最优方程[23]推导出最优状态价值函数

V^*以及最优动作价值函数Q^*：

$$V^*(s)=\max_a \sum_{s',r} p(s',r|s,a)\left[r+\gamma V^*(s')\right] \tag{5}$$

$$Q^*(s,a)=\sum_{s',r} p(s',r|s,a)\left[r+\gamma \max_a Q^*(s,a)\right] \tag{6}$$

其中$p(s',r|s,a)$为状态转移概率. 在现有基于值的强化学习算法中，目的是计算最优的动作-价值函数Q^*. 最经典的深度强化学习算法是深度Q网络算法(DQN)[31]，DQN使用一个神经网络作为动作价值函数估计器$Q(s,a;\beta)$. 拟合的动作价值函数神经网络的目标值为时序差分(Temporal-Difference，TD)目标$y_t=r_t+\gamma \max_a Q(s',a';\beta)$，DQN算法通过最小化TD损失来训练神经网络：

$$\mathcal{L}(\beta)=\mathbb{E}_{(s,a,r,s')\sim B}\left[\left(y_t-Q(s,a;\beta)\right)^2\right] \tag{7}$$

其中，B为收集的轨迹信息作为经验回放，经验回放的目的是提升算法学习的稳定性. 在DQN的执行过程中，通过计算TD损失来更新神经网络参数β，使得Q值尽可能地接近y_t，可以使用梯度下降等方法执行策略参数更新.

基于DQN的这种算法架构，引申出来了很多相关的改进算法. DDQN[32](Double Deep Q Network)使用两个神经网络进行Q值的学习，基于当前网络执行动作，目标网络进行策略评价，DDQN有效解决了DQN过高估计动作价值函数所导致的误差积累问题. Dueling DQN[33](Dueling Deep Q Network)方法通过引入优势函数$A(s,a)=Q(s,a)-V(s)$，大幅度提升了DQN算法的性能. 该算法对共享网络引入先验知识，并根据状态价值函数和优势函数对动作价值函数进行转换计算，Dueling DQN对策略性能实现了大幅度的提升.

2.1.2　基于策略的强化学习方法

借助于启发式算法思想，基于策略的强化学习方法被提出[34-35]. 在输入与输出之间建立可微的参数模型，并通过梯度优化实现参数的优化. 该算法的核心在于建立策略π与目标函数的关系，这种方法的优势在于无需明确定义值函数，就可以用来解决值函数方法无法处理的连续动作问题. 首先，根据输入的状态信息，定义输出动作的概率分布：

$$\pi_\theta(a|s)=P[a|s;\theta] \tag{8}$$

强化学习的目标为最大化奖励，因此可改写为关于参数θ的目标函数：

$$J(\theta)=\mathbb{E}_{\tau\sim p_\theta(\tau)}[R(\tau)] \tag{9}$$

其中τ表示智能体轨迹，$R(\tau)$表示累计回报，轨迹分布$p_\theta(\tau)=\prod p(s_t)\pi_\theta(a_t|s_t)$. 引入REINFORCE算法[36]对该函数的梯度进行估计：

$$\nabla_\theta J(\theta)=\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1}\nabla_\theta \log \pi_\theta(a_t,s_t)R(\tau)\right] \tag{10}$$

因此，得到梯度的更新公式为：

$$\theta=\theta+\alpha\nabla J(\theta) \tag{11}$$

具体的更新过程为在策略π_θ下采集轨迹、计算目标函数的梯度、更新梯度、随着梯度的上升实现策略的优化. 基于REINFORCE算法架构，后续提出了很多改进的算法. 演员-评论家算法[37](Actor-Critic，AC)是一种基于策略梯度理论的算法，该算法由策略网络和价值网络所组成. 策略网络用于和环境进行交互，根据状态信息来执行动作，并根据价值网络的指导进行参数的更新，价值网络则用于对当前状态执行动作的好坏进行评价. 两种网络之间进行了对抗优化过程，实现了策略参数的有效优化. 深度确定性策略梯度算法[38]DDPG(Deep Deterministic Policy Gradient)在面对连续动作空间时执行了动作空间离散化的操作来直接处理连续动作空间问题，并且基于离策略(off-policy)的算法架构，实现了对经验的复用，DDPG算法的提出，在很多连续控制问题上都取得了很大的效果提升.

2.2　面向人工智能的可解释性研究

在基于数据驱动的人工智能发展过程中存在不可信、难理解以及难监管等问题. 在使用机器学习的算法开展研究时存在理论上的缺陷[20,39]，因为，建立输入与输出信息之间的关联时紧密地依赖于训练数据. 然而，由于数据本身的局限性，可能会导致虚假关系的产生以及无法解释关系的情况出现. 其次，在应用阶段，深度学习模型的黑盒性使得其在安全性方面存在潜在的风险，无法满足监管要求. 因此，面向人工智能的可解释性研究应运而生，该方法旨在使算法模型以一种可解释、可理解的方式与使用者、决策者和开发者等人员建立信任. 通过人工智能可解释性的研究，可以提供方法和技术来解释机器学习算法模型的执行过程，如特征重要性分析、模型可视化和规则提取等. 这些技术可以增加对模型工作原理的理解，提升模型的信任度，并满足监管性的要求[40-41].

现有面向人工智能的可解释性研究关注于算法的透明性、表达的解构性、算法适用边界、黑盒模型的事后解释以及因果分析等方面[42-44]，具体研究方法分为如下两类，第一类是模型本身内置的可解释

性研究，包括使用白盒自解释模型进行的可解释性研究，通过搭建决策树、贝叶斯网络等来实现黑盒模型到白盒模型的转化，通过自解释白盒模型的构建，实现了对模型结构、参数以及输入的直观解释．同时，将注意力机制引入到了模型内置的可解释性研究之中，受到人类注意力机制的启发，通过判别模型决策过程中不同特征的权重，来直观地呈现出算法模型所感兴趣的区域，以辅助人类对算法模型的理解．另一类研究方法是事后可解释性研究，这种方法指的是在模型训练之后，使用解释方法来研究已训练完算法模型的工作机制，通过规则提取，模型蒸馏以及敏感度分析等方法实现全局和局部的可解释性分析．

人工智能可解释性研究方法已经在生物医疗、电商推荐以及城市管理等领域之中实现了广泛的应用[45-48]．然而，相比于广义上的面向人工智能的可解释性研究，面向时序决策问题的强化学习方法，在开展针对决策的可解释性研究时，面临着动作解析困难、奖励无法分解以及因果关系模糊等更多、更新颖的可解释性挑战问题，因此，亟需开展面向强化学习的可解释性研究．

2.3 面向强化学习的可解释性研究综述现状

在现有面向强化学习的可解释性研究方法综述文章中，Milani等人[22]从特征重要性分析、策略学习行为解释以及策略长期表现解释三类解释性内容进行总结描述，并且设定了忠实性、性能以及领域相关性等评估可解释性方法的指标．刘等人[23]界定了智能算法和机械算法，对可解释性进行了定义，并且讨论了影响可解释性的因素，定义了环境解释、任务解释以及策略解释三类可解释性强化学习问题，但该文章并没有涵盖较多现有的可解释性强化学习研究方法．Heuilleta等人[24]将可解释性算法总结为透明可解释性算法和事后可解释性算法，他们将不同用户进行分类，分为专家、普通用户、开发者等不同角色，分析不同用户的可解释性的需求以及目标，但是该文章更多地从面向用户的角度出发，没有突出强化学习的特点．Puiutta和Veith[49]对可解释性的相关术语进行了介绍，他们将可解释性强化学习方法分为内在可解释性方法以及事后可解释性方法．同时，他们针对可解释性研究中的人类主观因素进行了讨论，但是并没有总结较多的相关工作．Wells和Bednarz[50]回顾了25项强化学习领域中可解释性的研究工作，包括可视化方法、基于查询的解释、策略总结、人在回路解释方法以及解释性验证等，并且分析了当前面向强化学习的可解释性研究的一些局限性．总体来讲，现有面向强化学习的可解释性研究综述并没有严格从强化学习研究范式的角度出发进行分析，并且所涵盖的研究工作较少，没有针对多智能体强化学习这类方法的可解释性进行总结与分析，对于可解释性研究中人类用户这一关键性因素，没有太多相关的研究工作介绍与分析，仍旧是非系统化且不全面的．因此，本文面向强化学习的可解释性研究进行全面且系统化的总结与分析，从单智能体到多智能体进行介绍，并对未来的研究方向进行思考与展望．

3 面向强化学习的可解释性定义与评估

3.1 强化学习的可解释性定义

本篇综述首先需要对可解释性进行准确的定义．在现有的研究工作之中，Lipon等人[51]将解释定义为对强化学习算法模型训练的事后解释与分析，Brain等人[52]将可解释性定义为在对智能体系统进行决策时的理解程度，Molnar等人[53]认为可解释性是为智能体的行为预测提供解释的能力．解释是指一个获取理解的方式，选择什么是必须解释的，以及如何进行可解释性的呈现．这涉及到过滤信息以及构建解释的形式化表征，并且使用可解释性的模型进行解释的推理．在可解释性的定义中，透明性代表了在系统上下文理解的基础上呈现解释的一种能力．

本文认为，面向强化学习的可解释性是指对强化学习模型的决策过程和行为进行理解和解释的能力．在强化学习中，智能体通过与环境的交互来学习行为策略，以最大化累积奖励．可解释性的目标是解释智能体为何做出某个决策、如何做出该决策以及明晰决策背后的因果关系．这种可解释性包含策略透明度、决策行为、因果分析以及人类因素，具体可以展现在以下几个方面：

(1)透明性解释

透明性解释是指可解释性算法能够揭示算法模型的内部结构和工作方式，以透明清晰且可理解的方式进行呈现，使人类能够直观地理解模型的学习和决策过程．

(2)决策性解释

决策性解释是指可解释性研究能够清楚地解释智能体在特定状态下为何选择某个动作，以及其他可能的动作为何被排除．

(3)因果性解释

因果性解释是指可解释性算法能够揭示智能体为何在某种情况下做出特定决策，即行为背后的原因和动机，以及执行决策后智能体变化的因果关系.

(4)用户性解释

面向强化学习的可解释性研究需要将解释结果以易于理解和使用的方式呈现给人类用户，使用户能够直观地理解智能体的决策.

3.2 可解释性评估

评估旨在基于相关技术以及指标体系来对模型的性能、鲁棒性等进行评价. 对于可解释性研究的评估，应该紧密地符合可解释性的定义. 在面向强化学习的可解释性研究中，对于可解释性的评估已经被证明是一件极具挑战性的任务. 首先，在该研究领域中没有一个统一的可解释性的概念[23]，因此，导致了不同定义下的解释，具有不同的评估标准. 其次，对于可解释性研究的受众不同，会导致产生很大的评估差异，一个解释的好与坏很大程度上取决于人为的主观判断，如何将这种人为定性的评估转化为定量计算的评估，是实现准确且统一化评估的关键. 结合上文中提出可解释性所关注的几个方面，本文将可解释性评估的标准设定为如下：

(1)忠实性(Fidelity)

忠实性衡量了模型的解释结果与原始模型预测结果之间的忠实度. 解释的忠实性[54]是指解释与智能体决策的真实原因之间的相关性. 通过设定仿真环境下的对比实验，与用户或黑盒模型预测结果进行对比是一种评估忠实性的有效方法. 通过实验，检查可解释性策略与原始黑盒策略期望的一致性，确保在相同条件下提供相似的解释，评估解释的可信度和合理性，确保解释在领域知识范围内且合理[55-57]. 忠实性的计算公式可总结为如下所示：

$$Fidelity = \mathbb{E}_{(x,y)\sim\tau(\pi,x_0)}[\mathbb{I}(\hat{\pi}(x), \pi(x))] \quad (12)$$

其中τ为在策略π下的分布，$\hat{\pi}$为可解释算法模型，π为原策略模型.

(2)性能(Performance)

此处的性能指标强调的是可解释性方法执行相同任务时的标准化性能评价，包括对模型解释性以及算法性能的综合评价. 在算法性能和可解释性能之间保持平衡是极为重要的，因为增强对强化学习的可解释性会导致更大的计算资源消耗以及更慢的推理速度[22,26]. 因此，需要检查可解释系统的性能是否与原始的强化学习算法模型的执行性能相同或更好. 对于强化学习方法，这些指标涉及到预测的准确率，总体的奖励学习情况，策略的收敛性以及执行任务的成功率等多项指标[25,58-60].

(3)效率(Efficiency)

在进行可解释性强化学习算法研究时，算法效率是至关重要的考虑因素. 强化学习可解释性算法通常需要在大量的环境交互和数据处理中进行解释生成，因此推理和训练的快速性是优先考虑的. 同时，为了在资源受限的环境下实际应用，算法的计算复杂度和模型大小应尽可能保持较低. 此外，算法的可扩展性对于未来可能面临的大规模应用需求非常关键. 综合考虑这些算法效率因素，可以开发出高效、快速且可部署的可解释性强化学习算法，进一步推动可解释性强化学习技术在各个领域的实际应用和发展[61-63].

(4)鲁棒性(Robustness)

鲁棒性意味着该系统对于环境变化、噪声、攻击或不确定性有一定程度的适应能力，指的是方法在面对未知环境、干扰或对抗性攻击时能够保持稳定且可预测的性能表现，鲁棒且抗干扰的解释性算法具有更高的可信度和更强的可靠性. 验证鲁棒性的方法包括对抗性测试、环境和模型扰动测试、跨数据集测试、灵敏度分析、实际应用测试以及用户反馈和评估[64-67]. 通过这些验证，来确保强化学习的可解释性方法能够在不同情况下提供可信的解释和决策，从而增加人们对这些方法的信任，并促进其在实际应用中的推广和应用.

(5)用户性(User-Orientation)

用户性是指对强化学习的可解释性方法分析其面向人类用户的主观性评价因素，例如，用户的满意度，人类用户对解释性系统的信任度等[68-69]. 具体的评估方法包含有问卷调查方法，统计人类用户与解释性系统交互的响应时间等方式. 用户性评估通过面向用户的主观性评价指标以衡量解释性方法的可理解性，可作为衡量解释性方法实用性和用户友好性的重要依据.

综合强化学习的可解释性定义内容，透明性解释应该准确地反映模型是如何做出行为预测的，不应该存在歪曲，从而达到忠实性指标要求，同时，透明性解释的方法应是高效的，不应该消耗过多的计算资源，以满足效率指标要求. 决策性解释应该提供对于如何改进或调整模型的指导，解释应该包含有用的信息，使得决策者能够采取相应的行动来改善模型的性

能.因果性解释应该帮助人类理解模型的预测是如何与输入之间的因果关系相关联的,应该满足用户性以及性能指标的要求.用户性解释应该具有对模型的可靠性和稳健性的保证,并且符合指定人类用户的需求与偏好,因此需满足鲁棒性以及用户性指标.基于对可解释性定义的理解以及指标体系的构建,可对不同的可解释性方法进行综合评价,以分析不同方法可解释性的完成度以及主要贡献.

4 面向强化学习的可解释性研究进展

本文遵循强化学习的马尔可夫决策过程开展可解释性研究的分类与总结,根据马尔可夫决策过程的四大要素:动作、状态、奖励以及策略,对可解释性方法依次分为:行为级解释、特征级解释、奖励级解释及策略级解释.接下来,针对这四个类别的可解释性方法进行详细描述.

4.1 行为级解释

在强化学习之中,智能体接收到环境的状态以及奖励信息进行分析,最终执行动作来完成决策.作为这个过程中最重要的一环,针对动作行为决策的可解释性研究是极为重要的,根据行为级解释的可理解程度,分为自解释模型构建、形式化解释以及生成式解释三类解释方法.针对这三类方法所整理的代表性算法的详细信息如表1所示,在表中罗列了每个类别下的代表性算法、所研究的智能体类别、执行实验的仿真环境以及实验评估维度的信息.行为级解释研究架构如图2所示.

表1 行为级解释代表性算法总结

类别		代表算法	智能体		仿真环境	评估指标				
			单体	多体		忠实性	性能	效率	鲁棒性	用户性
行为级解释	自解释模型	Linear Model U-Trees[70]	√	—	Flappy bird, MountainCar, CartPole	√	√	—	—	—
		VIPER[71]	√	—	Toy Pong, CartPole, Atari Pong, Half-Cheetah	√	√	—	√	—
		DDTs[72]	√	√	CartPole, Lunar Lander, Wildfire Tracking, Find-AndDefeatZerglings	√	√	√	—	√
		MAVIPER[64]	—	√	Physical Deception, Cooperative Navigation, Predator-prey	—	√	—	√	—
	形式化解释	PIRL[73]	√	—	TORCS, Acrobot, CartPole, MountainCar	√	√	—	—	√
		GPRL[74]	√	—	CartPole	—	√	—	—	—
		Symbolic policy[75]	√	—	CartPole, Mountain Car, Pendulum, InvDoublePend, InvPendSwingup, LunarLander, Hopper, BipedalWalker	√	√	—	—	—
		Neurosymbolic Transformers[76]	—	√	multi-agent formation flying	√	√	—	—	—
	生成式解释	SCM[77]	√	√	Cartpole, MountainCa, Taxi, LunarLander, BipedalWalker	√	√	√	—	√
		Generate Explanation[78]	√	—	guided maze, university building	√	√	√	√	√
		IBE[79]	√	—	Lunar-Lander	√	√	—	—	√
		Policy Explanations[80]	—	√	multi-robot search and rescue (SR), Multi-robot warehouse, Level-based foraging	√	√	√	—	√

4.1.1 自解释模型

自解释模型作为一种可被人直观理解的白盒模型,通过决策分支、关联关系等形式来直观地呈现出行为决策的过程.因其透明性、直观性以及可理解性,该方法被广泛地应用在强化学习行为级解释方法研究之中.常见的自解释模型有决策树、回归树以及K近邻模型(K-Nearest Neighbor, KNN)等[81-84].自解释模型的搭建流程基于强化学习的交互行为信息,根据智能体与环境交互收集到的状态,动作以及奖励信息来训练自解释白盒模型.训练得到的自解释模型通过逐步预测智能体的动作来清晰化呈现决策过程,从而实现行为决策的可解释性验证.

(1)单智能体自解释模型研究.针对强化学习的动作-价值函数,引入线性模型U树(Linear Model U-Tree, LMUT)[70]来近似深度神经网络以

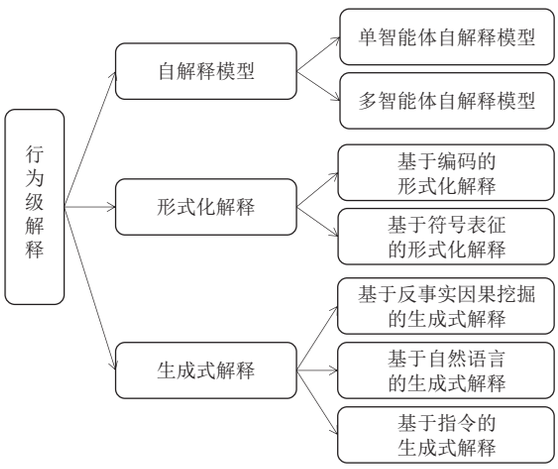

图2　行为级解释

进行隐式的规则提取，LMUT在线学习的模式适用于游戏场景，借助于透明树状结构直观地分析了所构建的LMUT模型与原策略的一致性，并通过计算特征重要性、规则提取以及像素级解释进行了行为可解释性的验证. 同样地，借助于树型的行为可解释性结构，通过策略提取验证强化学习[71](Verifiable Reinforcement Learning via Policy Extraction，VRLPE)方法基于模仿学习的思想对深度神经网络策略模型进行压缩与策略提取，最终构建出了可解释的决策树模型，同时，VIPER根据任务属性设定了正确性评估指标，基于动力学系统构建了稳定性评估以及ε-robust[85]鲁棒性评估指标进行模型行为可解释性的验证，最终，该方法在四类经典的单智能体控制任务上取得了优异的效果. 进化计算的方法被应用于可解释性决策树模型之中[86]，通过将进化计算与Q学习结合，对智能体的决策空间进行分解并且进行动作的一一对应，进化计算的全局优化性有效克服了Q学习过程中的局部最优问题，提升了决策树的性能.

(2)多智能体自解释模型研究. 不同于单智体的自解释模型构建，面向多智能体进行自解释模型研究时，需要考虑到智能体之间的竞争合作关系、单一智能体对多智能体系统的贡献等方面. Milani等人[64]将VIPER算法进行改进以适用在多智能体任务之中，为了更好地捕获多智能体任务中智能体之间的协调信息，他们提出了一种集中式多智能体决策树算法(Multi-Agent VIPER，MAVIPER). MAVIPER通过预测不同智能体的行为来联合构建决策树，并使用重采样的方法关注重要的状态信息，并在三个多智能体环境之中进行了性能的验证实验.

虽然通过树型结构的模型可以实现行为级可解释性策略的学习，但是在树结构模型构建时，面临着无法通过梯度下降方法进行智能体策略在线更新的问题，针对这一关键问题，一种可微决策树(Differentiable Decision Trees，DDTs)方法[72]被提出，借助于DDTs实现了对整个树模型的梯度更新从而对复杂样本进行逐步学习. 在单智能体以及多智能体任务的实验中，相比于传统的决策树模型，DDTs的决策性能提升了7倍，在实现行为级解释的同时实现了对白盒自解释模型性能的大幅度提升. 在DDTs方法中，针对决策树的一个单一节点的Q值学习以及策略梯度更新过程如下式所示：

$$f_T(s,a)=\mu(s)\hat{y}_a^{True}+(1-\mu(s))\hat{y}_a^{False} \tag{13}$$

$$\nabla f_T(s,a)=\left[\frac{\partial f_T}{\partial \hat{y}_a^{True}},\frac{\partial f_T}{\partial \hat{y}_a^{False}},\frac{\partial f_T}{\partial \alpha},\frac{\partial f_T}{\partial \beta},\frac{\partial f_T}{\partial \phi}\right] \tag{14}$$

其中：

$$\frac{\partial f_T}{\partial \hat{y}_a^{True}}=1-\frac{\partial f_T}{\partial \hat{y}_a^{False}}=\mu(s) \tag{15}$$

$$\frac{\partial f_T}{\partial \alpha}=(\hat{q}_a^{True}-\hat{q}_a^{False})\mu(s)(1-\mu(s))(\beta s-\phi) \tag{16}$$

$$\frac{\partial f_T}{\partial \beta}=(\hat{q}_a^{True}-\hat{q}_a^{False})\mu(s)(1-\mu(s))(a)(s) \tag{17}$$

$$\frac{\partial f_T}{\partial \phi}=(\hat{q}_a^{True}-\hat{q}_a^{False})\mu(s)(1-\mu(s))(a)(-1) \tag{18}$$

其中，s表示特征，β表示为特征因子，μ表示布尔表达式，$\hat{y}_a$表示决策树标签. 当利用DDTs作为Q值学习的函数近似时，每个叶子节点返回一个未来预期奖励的估计值：

$$f_T(s,a)\rightarrow Q(s,a)=\mu(s)\hat{q}_a^{True}+(1-\mu(s))\hat{q}_a^{False} \tag{19}$$

当利用DDTs函数近似强化学习策略梯度方法时，叶子节点代表了强化学习智能体应该采取的行动的最优概率分布的估计. 因此，这些叶子节点上的值代表了选择相应动作的概率.

$$f_T(s,a)\rightarrow \pi(s,a)=\mu(s)\hat{\pi}_a^{True}+(1-\mu(s))\hat{\pi}_a^{False} \tag{20}$$

其中，需要针对公式(13)约束所有的执行动作的概率之和为1：

$$\hat{y}_a^{True}+\hat{y}_a^{False}=1 \tag{21}$$

在后续的研究之中，基于自解释模型中决策树算法存在的一些缺陷，研究者们提出了很多改进的方法. 一种改进的决策树算法[87]结合模型的预测、行为可视化以及规则解释三项指标来综合评价，在叶子节点通过权衡这三个指标的最低可变性以进行模型的优化. 边界迭代的马尔可夫决策过程

(IBMDPs)算法[88]围绕一个基础的MDP模型,设定掩蔽程序以及值更新步骤进行决策树边界的迭代更新.

与此同时,为了限制决策树的规模,Roth等人[89]提出了一种保守算法,只有智能体策略的未来回报增加足够多时,才会扩大决策树的规模,通过这种方法构建了更加简洁的自解释模型以应对大规模决策任务.随后,演化学习方法被引入至决策树模型之中[90],设定了状态空间分解以及动作价值函数最大化的两个目标,基于这两个目标进行算法优化以构建更加精准的自解释模型,相比于决策树算法该算法展现出了更高的性能,消融实验的结果也进一步证明了演化学习方法引入的有效性.

除了树型结构的自解释模型构建方法,一种抽象策略图的模型[91]被提出,抽象的马尔可夫链简明地展示了行为决策过程,并且对于价值函数的学习没有添加限制,因此,与现有很多的强化学习算法相兼容.采用演化特征合成的方法[92]从神经网络策略之中进行策略提取和简化,并且在四种强化学习环境中进行实验,验证了该方法比基于树结构的自解释模型具有更优异的性能表现.

4.1.2 形式化解释

形式化解释方法旨在通过符号化、公式化的形式来实现可解释性决策,相比于自解释模型,形式化解释更加深入地挖掘了黑盒深度神经网络内的组件关系,借助于布尔函数,线性编码或者符号表征等方式呈现出透明的强化学习行为决策过程,基于编码以及基于符号表征的形式化解释示例如图3以及图4所示.

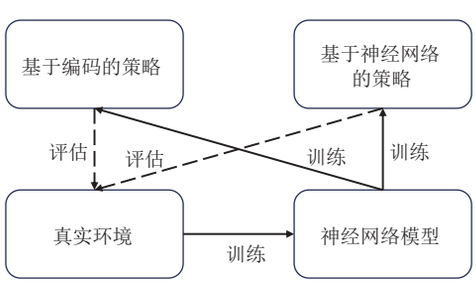

图3 基于编码的形式化解释方法图

图4 基于符号表征的形式化解释方法图

(1)基于编码的形式化解释.基于编码形式的强化学习解释性研究框架(Programmatically Interpretable Reinforcement Learning, PIRL)[73]使用一种高级的编程语言表示策略,从而更易被人所理解并且可以通过符号化的方式进行验证,同时,PIRL构建了一种神经定向程序搜索方法,以解决非光滑优化问题下的回报最大化搜索,PIRL在模拟赛车的游戏中进行了解释性的验证,取得了更加平滑的行驶轨迹并且具有很好的迁移性.该项研究工作为强化学习的形式化解释研究提供了值得参考的研究思路.

随后,基于自然进化思想使用遗传编码的生成式可解释强化学习策略方法(generating interpretable reinforcement learning policies using genetic programming, GPRL)[74]被提出.GPRL方法使用基于模型的离线处理方式,并且结合遗传编码的方法从轨迹样本之中自动学习行为策略,在强化学习的三个基准实验上,验证了GPRL具有较高的性能表现.将遗传编码方法引入至行为解释研究之中,基于离线数据的遗传强化学习编码算法[93]实现了从离线轨迹数据之中自动学习策略方程,最终生成性能良好的可解释性强化学习策略.以往的程序合成方法主要为命令式编程和声明式编程两个维度,但是命令式编程忽略因果逻辑导致其泛化性弱;声明式编程严格遵循因果逻辑,但在复杂任务中学习能力差.针对这两个问题,基于草图的程序合成方法被提出[94],自动生成具有泛化能力和可解释因果逻辑的白盒程序,从而解决深度强化学习方法泛化能力弱以及可解释性差的问题.

受到遗传编码优化算法的启发,粒子群优化方法被引入至强化学习可解释性方法之中,一种模糊粒子群强化学习模型(particle swarm optimization for generating interpretable fuzzy reinforcement learning policies, FRSRL)[95]被提出,FRSRL仅通过在模拟真实系统动力学的世界模型上的训练参数来构建模糊策略,该方法是第一个将自组织模糊控制器与基于模型的批处理强化学习方法结合的方法.FRSRL旨在解决无法与环境在线交互的策略学习任务,在三个基准测试集上的实验验证了方法的有效性和解释性.

(2)基于符号表征的形式化解释.符号化强化学习行为策略(symbolic policy)[75]表示方法通过自回归的递归神经网络(RNN)来生成控制策略,并通过数学表达式的形式进行表示,同时,以一种风险搜

寻的策略梯度更新方法来提升模型的性能，在将该方法扩展至高维动作空间环境时，采用符号化表征的方式进行策略表征．并在组合优化问题上进行探索，最终在8个基准测试环境中，模型性能优于现有的7个先进的深度强化学习算法．

同时，一阶逻辑也被应用在形式化解释方法之中．神经逻辑强化学习方法[96]（Neural Logic Reinforcement Learning，NLRL）基于策略梯度方法和可微归纳逻辑编程，使用一阶逻辑来表示强化学习的策略．该方法在监督任务的可解释性和通用性方面显示出了显著的优势．合成程序的方法作为可解释策略[97]通过奖励信号来程序化问题的表达，以无监督的形式学习嵌入空间上的连续参数，这种研究框架不仅可以生成任务解决程序，还可以保证生成的可解释行为策略性能优于基线方法．结合符号规划和神经学习方法的优势，构建出了一个利用以往不可用模型的不完整模型类算法[98]，并且在复杂的环境中实现了有效的应用与探索．

在多智能体强化学习形式化解释的研究之中，为解决协同多智能体规划问题的通信结构的推断问题并满足最小化通信量，设计出了一种控制策略[76]，将用于生成通信图的编程通信策略与用于选择行为动作的Transformer策略网络相结合，最终构建出了一种神经符号化Transformer（neuro symbolic transformers）模型，有效实现了多智能体强化学习的行为解释．针对深度强化学习模型无法解释以及部署成本大的问题，提出了一种形式化验证方法[99]，搭建了一个基于反事实引导和数据驱动的安全性的深度强化学习系统，并且在六个经典控制系统中验证了所提出系统的有效性，在不保证累计奖励和鲁棒性损失的情况下，实现了更加可靠的深度强化学习方法研究．

为了融入专家背景知识，一种可微归纳逻辑编程（Differentiable Inductive Logic Programming，DILP）的关系强化学习框架[100]被提出，有效地实现了从图像中捕获智能体关系信息，并将环境状态表征为一阶逻辑谓词形式以融合专家的背景知识判别信息，整体的框架是端到端训练优化，在较为复杂的BoxWorld、GridWorld以及CLEVR环境的任务中验证了该框架的有效性．随后，Zhang等人[101]提出了一种可微逻辑的离策略强化学习框架，继承了DILP中可解释性的优势，并使用分层近似推理的方法减少了逻辑规则的数量，提升了执行效率、稳定性以及可伸缩性．

4.1.3　生成式解释

生成式解释旨在根据强化学习行为动作，直接生成人类可理解的解释，包括文字、图表等内容．现有的研究集中在使用反事实因果挖掘、自然语言生成以及指令生成等方式进行生成式解释的研究．

（1）基于反事实因果挖掘的生成式解释．基于结构因果模型[77]（Structural Causal Models，SCM）的反事实行为解释方法，在强化学习之中学习结构因果模型并编码不同状态之间的因果关系，在6个领域中对该模型进行了评估并测量了因果预测精度，并且设定了一项120人参与的实验，让参与者观察因果模型在星际争霸2游戏中的行为解释，最终在可解释性以及满意度上取得了很好的效果．随后，Olsona等人[102]提出了一种反事实行为解释方法，针对智能体在雅达利（Atari）游戏中的视觉环境，进行反事实行为解释的生成方法研究，并使用该方法进行人类行为实验，通过调查人类能否分辨出生成的解释行为与实际的行为来检验方法效果，同时，该方法可以帮助通过反事实解释来帮助人类检查缺陷．

（2）基于自然语言的生成式解释．在部分可观察环境下生成对导航行为解释的自然语言生成方法（generate explanation）[78]将视觉信息作为输入并由神经网络进行处理，随后使用一种分层的改进式贝尔曼方程进行行为预期代价的计算，执行高层操作的预期代价可以通过根据高层动作及其二元输出结果编写的贝尔曼方程近似值来表示：

$$Q(b_t,a_t)=D(b_t,a_t)+P_S(a_t)R_S(a_t)+(1-P_S(a_t))[R_E(a_t)+\min_{a\in A(b_t)/a_t}Q(\tilde{b}_{t+1},a)] \quad (22)$$

其中，$\tilde{b}_{t+1}$是近似的未来信念，包含了未来执行动作失败以及状态转移未知的先验知识．$D(b_t,a_t)$表示子目标之间通过迪杰斯特拉算法计算的导航成本开销．$P_S(a_t)$表示动作执行成功的可能性．$R_S(a_t)$和$R_E(a_t)$分别表示成功探索和失败探索的预期代价．这种改进的计算方法搭建了训练程序来进行解释的生成，智能体执行的效率提升了9%．

针对人类感知，一种人工智能策略解释自动生成（automated rationale generation）[103]方法提出，通过搭建一个计算模型来学习智能体的内部状态和动作信息，转换为人类可理解的自然语言，并且设定了两项面向人类用户的实验以验证生成解释的准确性与可理解性，最终发现参与者更愿意接受一个描述更为详细的解释信息．随后，Hayes和Shah[104]提出

了一种自动策略解释生成的方法，实现了与人类协作者的响应以及目标查询. Ehsan等人[105]提出了一种智能体行为解释的自然语言生成方法，使用机器翻译将智能体内部的状态-动作表示翻译为自然语言，两项评估实验的结果表明该方法可以提升人类用户的满意度.

(3)基于指令的生成式解释. 基于指令的行为解释方法[79](Instruction-Based Explanation，IBE)可通过重用人类专家给出的指令信息来加速策略的学习，并且可自动获取表达式来解释自己的行为. 随后，提出了一种改进的指令级行为解释方法[106]，面向动态变化智能体策略实现了行为级解释指令生成.

在多智能体系统中，提出了两种方法来生成行为解释(policy explanations)[80]，第一种方法是实现智能体协作性策略的解释，另一种方法则是针对智能体行为查询的语言解释回复，这种直接生成解释的方法在三个多智能体任务上得到了可扩展性的验证，并且在人类用户交互的实验上进行了可解释性的验证，实验结果表明所生成的解释可以显著提高用户的表现，并且提升了用户的满意度等多项主观评价指标.

(4)面向强化学习的行为决策验证. 借助于行为级解释的方法可以在强化学习过程中对智能体的决策行为进行有效的评估验证，整体的验证架构可总结为图5所示. 具体来讲，一种针对深度强化学习系统的验证方法[107]基于深度神经网络验证方法应用于视频数据、云资源管理以及网络调度等任务之中，有效验证了恶劣的决策行为，为构建更加安全以及可靠的深度强化学习系统提供了指导方针. 随后，Zhu等人[108]提出了一种形式化程序解释方法用于神经网络部署环境，使用更简单、更可解释的近似网络保证所需的安全属性存储，并自适应于未观察到的环境之中，形式化程序可以表示规范的逻辑控制，进一步细化了行为搜索空间，通过监控和防止不安全行为的产生，动态地执行安全条件，这种形式验证技术可以用来实现低开销的可信强化学习系统. 为应对连续动作空间下的策略迭代学习问题，Anderson等人[109]构建了两个策略类来解决这一挑战，一个是通用的、具有近似梯度的神经符号类，另一个是允许有效验证的更有限的符号策略类. 在每次迭代中，安全地将符号策略提升到神经符号空间，对生成的策略执行安全梯度更新，并将更新后的策略投影到安全符号子集，这种方法加强了安全性的探索并优于先前的策略.

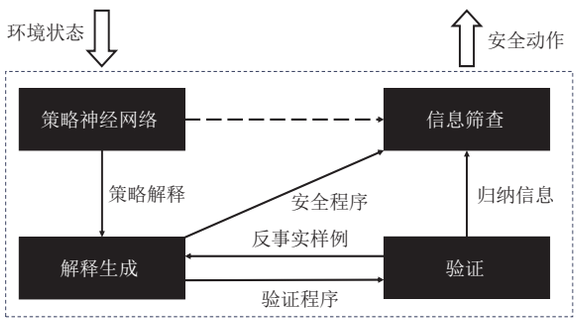

图5 面向强化学习的行为决策验证架构图

在行为级解释方法之中，自解释模型在实现黑盒深度模型白盒化的同时，也存在有很多缺点：(1)相比于具有较大参数量的深度网络模型，传统的自解释模型性能往往弱于深度网络模型，在相同任务的决策性能上低于深度强化学习算法模型；(2)不适用于大规模智能体决策场景，随着问题规模的扩大，决策树模型的节点数会逐渐增多，模型的性能也受到很大影响，在模型的性能与可解释性之间很难做到平衡，导致该类算法不适用于大规模决策求解问题.

形式化解释方法借助于符号化表征以及编码等技术实现了模型性能保证下的行为解释，作为一种辅助解释方式，相比于自解释模型有更强的性能表现、更好的鲁棒性以及更快的执行效率，但是如何生成人类可直接理解形式的行为解释较为困难. 生成式解释旨在直接生成人类可理解的自然语言来对智能体行为进行描述，这种方法更加直观. 但是存在较大的人为主观性，并且指令集以及语料库的构建需要消耗很大的人力. 如何在实现解释的同时提升效率是接下来的研究重点.

4.2 特征级解释

如表2所示，面向强化学习中智能体的状态信息，进行环境状态特征的可解释性分析研究，具体来讲，可以分为交互轨迹解释以及关键特征可视化两类方法. 智能体与环境交互过程中状态信息叠加为交互的轨迹，可以通过分析交互轨迹信息以有效地判别智能体所学习策略的好坏. 例如，在经典的控制环境MuJoCo中，可根据模拟环境中的机器人行为轨迹判别连续动作空间下智能体策略的学习程度. 在蚂蚁走迷宫(AntMaze)的环境之中，可以根据蚂蚁在迷宫中走过的轨迹信息判别其走出迷宫的重要路线或是关键的决策途径. 因此，从智能体交互轨迹出发，进行面向强化学习的可解释性方法的探索是极为重要的. 除此之外，对智能体交互轨迹

表2　特征级解释代表性算法总结

类别		代表算法	智能体		仿真环境	评估指标				
			单体	多体		忠实性	性能	效率	鲁棒性	用户性
特征级解释	交互轨迹解释	EDGE[110]	√	√	Pong in Atari, You-Shall-Not-Pass in MuJoCo, Kick-And-Defend, two OpenAI GYM games	√	√	—	√	√
		Collective eXplainable AI[111]	√	√	Multiagent Particle, Sequential Social Dilemmas	√	√	√	—	—
		SHA-KG[112]	√	—	Jericho game	√	√	—	—	—
		Weakly-Supervised Control[113]	√	—	PushLights, PickupColors, Push and Pickup, door	√	√	—	—	√
		CE-RLEC[114]	√	—	Human-Cases	√	—	—	—	√
		RBIRL[115]	√	—	Grid World, Real-world Vehicle Routing	√	√	√	√	—
	关键特征可视化	i-DQN [116]	√	—	Atari	√	√	—	—	√
		VUAA[117]	√	—	Atari	√	√	—	√	√
		RL-Intention[118]	√	—	Blackjack, CartPole, Taxi	√	√	—	—	√
		DQNViz[119]	√	—	Atari	√	√	√	—	√
		xGAIL[120]	√	√	Taxi	√	√	—	—	√
		Social-Attention[121]	—	√	Highway-Env	√	√	—	—	√

中的关键状态特征或是关键决策行为进行可视化的展示，可有效提升人类用户对智能体学习到强化学习策略的理解. 特征级解释研究架构如图6所示.

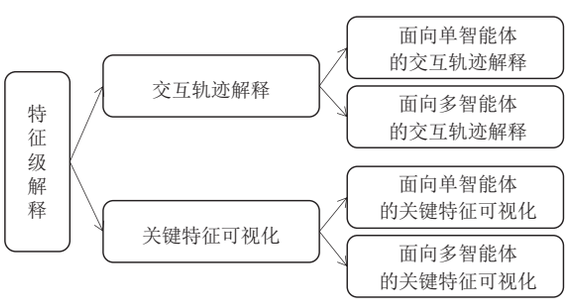

图6　特征级解释

4.2.1　交互轨迹解释

交互轨迹解释旨在通过对智能体与环境的交互轨迹进行解释分析. 在单智能体研究之中，基于单体与环境的交互信息可以实现对智能体任务的分解，并且可以有效地应对稀疏奖励环境，这种方法在逆向强化学习[115]研究之中被广泛地进行应用. 同时，在多智能体系统的研究之中，基于多体之间的交互信息，可以有效地分析智能体之间的竞争合作关系，提升多智能体系统的学习效率.

(1)面向单智能体强化学习的交互轨迹解释. 针对文本类游戏使用知识图谱的方法[112](Stacked Hierarchical Attention with Knowledge Graphs, SHA-KG)进行智能体状态的显式推理，交互式轨迹的模拟过程在图形式之中更加清晰地呈现了出来，以一种透明且可理解的方式进行显式表征，并且，在文本类游戏环境中进行了广泛的实验评估，验证了所提出方法的性能. Lee等人[113]提出了使用弱监督控制(weakly-supervised control)方法对智能体的决策状态空间进行缩减，针对单一决策任务，使用弱监督方法基于历史交互轨迹信息自动地从决策空间中分离出任务子空间，从而实现在任务子空间上进行智能策略的探索，在各种具有挑战性的、基于视觉的连续控制问题上，所提出的方法带来了实质性的性能提升.

智能体通过状态转换的预期结果来解释其行为(Contrastive Explanations for Reinforcement Learning in terms of Expected Consequences, CE-RLEC)[114]. 该方法构建了一个将状态和动作转换为人类用户更容易理解的描述模块. 其次，设计一种获取单个操作以及全局策略的程序以生成行为解释. 随后，针对电力系统提出了一种可解释的电力系统强化学习模型来辅助电力的控制操作[122]，基于交互轨迹信息构建了一种沙普利加性解释的反向传播解释器，并为应急控制应用提供了一个剪枝后的可解释模型，并在可解释的结果之中融入特征分类以更加清晰地呈现解释效果.

与此同时，在逆向强化学习的研究中，针对专家行为的噪声问题，提出了一种鲁棒的逆向强化学习

框架(Robust Bayesian Inverse Reinforcement Learning,RBIRL)[115],实现准确地估计噪声函数.并且,针对现实世界之中最普遍的稀疏噪声行为,引入了一种潜在变量来提升专家行动的可靠性,基于交互轨迹搭建的推理算法实现了自动识别和去除智能体学习过程中的行为噪声.在真实的车辆路由实验环境下,该方法展现了鲁棒的噪声去除能力.随后,针对深度黑盒模型的不确定性度量问题,一种基于蒙特卡罗采样的不确定估计方法被提出,并搭建出一个安全强化学习框架[123],来实现行人导航过程中的环境不确定性计算,以避免碰撞的发生.在模拟实验环境中,验证了该方法针对交互行为表现出更加稳健的安全性能.

(2)面向多智能体强化学习的交互轨迹解释.基于智能体交互轨迹信息,使用一个自定义核函数和一个可解释的预测器来增强高斯过程,并利用诱导点和变分推理设计一个参数学习程序的策略级解释方法[110](strategy-level explanation of drl agents,EDGE)提高策略学习效率.使用该方法可实现从智能体历史交互轨迹数据中预测最终的奖励信息,并分析不同轨迹信息的重要性,来辅助智能体的解释.通过在雅达利以及MuJoCo环境下的实验,验证了可解释性策略的忠实度,并且展示了如何根据解释来理解智能体行为并发现策略漏洞以进行纠正.

随后,针对多智能体之中的合作型策略,受到博弈论的启发,基于沙普利(Shapley)值来量化评价合作型多智能体之中个体之间的贡献(collective explainable AI)[111],从而解释合作型团体中各个智能体行为.并且,针对高开销的沙普利值计算过程,使用蒙特卡罗采样方法实现近似,最终,在多粒子环境和社会困境问题上的验证表明,所提出的方法实现了对每个智能体贡献的准确估计,并且从经济学研究的角度分析了其研究价值,证明了沙普利值方法在多智能体可解释性研究中应用的有效性.

4.2.2 关键特征可视化

在强化学习的可解释性研究中,可视化方法是一类重要的解释辅助方法.通过对强化学习过程中关键的状态特征进行可视化,可以直观地呈现智能体策略的学习情况.另一方面,在很多极其复杂的时序策略学习过程中,直接搭建自解释模型或是生成文本解释是无法实现的,因此,通过可视化的方式可以实时地了解到智能体的状态转移信息、重要的行为轨迹以及关键的决策步骤.关键特征可视化的方法是辅助人类对强化学习策略进行理解的重要途径,在现有的关键特征可视化研究方法之中,涌现出了基于注意力信息、显著性映射以及信念分布等重要的可视化方法[124-127].

(1)面向单智能体强化学习的关键特征可视化.注意力机制首先被引入至深度强化学习方法之中进行初步探索,一种可解释的深度Q值学习方法[116]通过键值记忆单元以及注意力机制为模型进行可解释性赋能,在保证智能体学习策略高性能的前提下实现了可解释性.在八个经典的雅达利游戏的实验中,证明了所提出的基于注意力机制的强化学习可解释性方法达到与当前深度Q值学习方法相媲美的策略性能.但是,该方法也存在提取特征模糊、模型易过拟合等问题.随后,针对可视化强化学习策略研究中的映射失效问题,提出了一种区域敏感性可视化方法(RS-Rainbow)[128],使得在强化学习过程中智能体具有天生的可视化能力,基于一个端到端的神经网络融合注意力机制来学习关键的特征区域,借助于反向传播技术可视化重要区域,所提出的方法在八个经典的雅达利游戏上的实验证明,该方法不仅提高了模型的可解释性并且提升了强化学习决策性能.

选择性注意力机制指的是人类在现实观察中只关注重要信息而忽视其他信息,从而更加专注于重要特征的现象,受到i-DQN等算法中所使用的注意力机制的启发,引入自注意力方法来对强化学习智能体的关键特征进行分析[129],在实际操作中,首先通过输入信息大小的限制,来缩小观测空间,随后,训练一个基于视觉的自注意力强化学习架构,自注意力机制拥有和间接编码相似的可解释性特征,从而在大幅度缩减参数的情况下来解决具有挑战性的视觉任务.

类似地,同样针对视觉图像类问题进行可视化可解释性研究,一种重要性图的表征方法[130]分析每个像素对黑盒模型的重要性,通过随机掩码嵌入的方法来预测模型的输出,在多个图像类实验上验证了所提出方法相比于白盒模型更加突出的性能.生成对抗学习方法[120](eXplainable Generative Adversarial Imitation Learning,XGAIL)被提出来学习专家的行为策略,针对模拟出租车场景,进行行为轨迹的可解释性研究,借助于重要性图来直观地呈现不同交通情况下,出租车的关键决策路径,从而有效地理解强化学习的策略.

随后,显著性映射方法被提出,如图7所示,使

用显著性映射的方法针对雅达利游戏进行强化学习关键特征可视化的研究[117]，特别地，在搭建显著性图的过程中，研究者们关注于智能体在学习过程中关注哪些信息、智能体出于哪些原因而做出决定以及智能体在策略学习过程中是如何实现进化的．最终的实验表明显著性映射方法可以为强化学习智能体的决策行为提供重要的解释．针对视觉任务下强化学习算法进行可解释性研究[131]，通过可视化其重要像素信息来实现策略的辅助解释，在该研究工作中，研究者们关注于强化学习智能体与人类执行相同任务时的视觉表征相似性，如何根据这些相似或者差异信息来解释智能体的行为，他们使用了显著性图的方法在雅达利游戏上进行视觉模型的分析，通过可视化的方式分析超参数如何影响智能体的策略学习，并且研究如何缩小人类专家与强化学习智能体之间的性能差异．

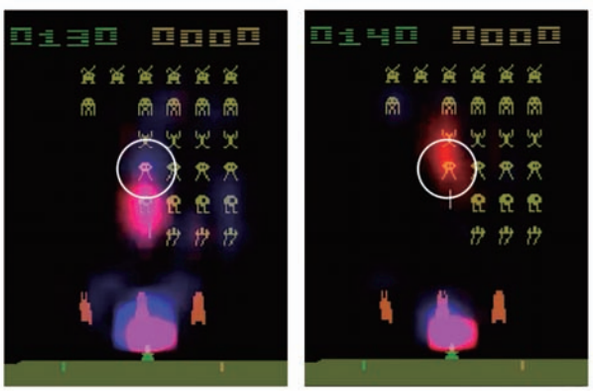

图7　SpaceInvaders游戏环境下显著性图[106]

同时，数据信念分布的研究也被应用在可视化方法之中，基于智能体预期结果信息来实现对策略的可解释性研究（RL-intention）[118]，在智能体训练过程中，基于改进Q函数近似方法进行局部行为解释方法的研究，并且通过智能体行为轨迹的信念分布图在多个经典强化学习场景下来可视化关键特征，实验的结果证明了所设计的可解释性Q函数与实际Q值的一致性．

结合以上研究，可以综合不同的可视化方法来进行智能体特征的系统性分析．如图8所示，通过融合多种可视化方法可以全面评估策略学习情况．针对强化学习过程搭建了层次化可视化分析系统（a visual analytics approach to understand deep q-networks，

图8　雅达利游戏环境下强化学习算法训练的可视化[119]（ⓐ统计视图用折线图和堆叠区域图显示整体训练统计数据；ⓐ1表示训练统计数据的折线图；ⓐ2表示训练统计数据的堆叠区域图；ⓑ训练周期视图用饼图和堆叠柱状图显示每个训练周期的统计数据；ⓑ1表示每个训练周期统计数据的饼图；ⓑ2表示每个训练周期统计数据的堆叠柱状图；ⓒ轨迹视图揭示了智能体在不同轨迹中的动作和奖励；ⓒ1-ⓒ7分别代表了智能体不同的轨迹片段信息；ⓓ片段视图显示了智能体游戏中的内容）

DQNViz)[119],从轨迹信息、动作奖励信息、平均奖励和损失等信息以及动作和奖励分布四个层次,通过图表的方式直观地呈现出智能体的学习过程,在雅达利游戏实现了提取有价值的动作奖励模式来实现模型的解释并控制训练过程.结合不同周期下的训练情况以及轨迹变换情况,挖掘重要的轨迹以及决策行为信息,明晰重要的决策步骤并分析其对应的视图片段来解释智能体的关键决策行为,从而有效地辅助领域专家进行理解,以改进深度强化学习模型.并且,可针对智能体未来行为的预测信息来进行特征的分析.Pedro Sequeira和Melinda Gervasio提出了一个可解释的强化学习框架[132],该框架通过分析智能体与环境的交互轨迹历史数据来提取关键特征元素,从而辅助行为解释.基于行为交互数据进行摘要提取,并且可视化摘要中的重要信息以及关键片段信息,可视化的信息可有效辅助人类用户的理解.同时,证明了所提取元素在策略执行过程中的优势以及局限性.随后,一个改进的预测模型[133]被提出,不仅实现了未来事件的推断,并对语义信息进行了生成,以可视化的方式进行展现,语义的预测模型通过聚合多尺度特征图来预测未来的语义信息,并指导智能体的行动选择.最终以图像分割以及动作概率分布图的可视化方式直观地呈现出对智能体行为的解释.

(2)面向多智能体强化学习的关键特征可视化.一种特征函数的方式来评估输入信息以构建因果模型[134],不同于反事实的方法,该研究直接针对训练过程搭建端到端的特征函数来分析多智能体博弈环境下个体的特征.通过显著性图的方法直观地展现出了该方法解释的有效性.针对密集交通环境下的多智能体行为决策(Social Attention for autonomous decision-making, Social-Attention)[121],同时一种基于注意力的体系结构被提出,实现了交通中参与者之间的行为交互的可视化.基于显著性图的研究,提出了一种内部状态可视化的通用方法[135],进一步提升了显著性映射方法在多智能体强化学习关键特征可视化上的应用.

基于图像分割的研究,进一步提出了一种改进的动态行为可解释性研究[136],通过检测智能体与环境的交互量来获取更优的策略,相比于直接使用黑盒模型将图像映射为策略更具有可解释性,也进一步帮助了人类用户对最优策略进行选择.类似地,针对视觉图像类任务,提出一种稀疏贝叶斯强化学习框架[137],用于记录历史经验中最关键的图像数据,通过维护一个图像快照缓存,来辅助策略的学习.该缓存仅保存关键数据,通过可视化的方式提升对多智能体决策行为的理解.并在导航任务上进行了方法有效性的验证.

在特征级解释方法之中,交互轨迹解释旨在对智能体的轨迹数据进行解释性分析,但是面临着智能体交互行为难量化,沙普利值计算复杂以及多智能体间关系难分解等问题.而在关键特征可视化研究之中,更多是借助于可视化手段来辅助对关键特征的理解,面临着特征边界定义模糊,多智能体间关键特征无法区分等问题.这些问题限制了特征级解释方法从辅助理解到客观解释,也是后续特征级解释的重要研究方向.

4.3 奖励级解释

在强化学习之中,奖励函数的设计是指导智能体进行策略学习的关键因素,奖励函数设计的好坏直接决定能否学习到收敛且最优的策略,因此,在可解释性研究之中,奖励级解释方法的研究具有重要意义.通过对奖励级解释的研究,可有效指导奖励函数的设计与优化过程,并且,在强化学习研究之中,奖励函数的设计往往基于人类先验知识进行构建,奖励级解释方法可针对人为设计的奖励函数实现有效性的验证.在现有奖励级解释方法之中,根据不同的任务场景以及先验信息,可分为奖励分解以及奖励塑造两个类别.具体方法总结如表3所示.奖励级解释研究架构如图9所示.

4.3.1 奖励分解

(1)面向单智能体的奖励分解.研究奖励分解方法来解释强化学习中智能体的决策行为(decomposed reward Q-learning, drQ)[138],将奖励信息分解为语义意义上不同类型奖励的总和,从而可根据不同奖励类型来分析智能体的决策行为,与此同时,在该项工作之中研究者们基于Q值分解的思想提出了一个名为最小充分解释的概念,以解释在相同类型奖励上智能体的最优决策行为.并且通过在CliffWorld以及Lunar Lander环境上的实验,验证了奖励分解的有效性.

(2)面向多智能体的奖励分解.奖励分解方法因其对于智能体系统中奖励信号的有效拆分,被广泛应用于多智能体强化学习研究之中,以分析多智能体系统中奖励信息随着智能体交互的动态影响.在多智能体强化学习奖励分解的可解释性研究之中,基于集中式训练分布式执行框架中的信用分配方法进行可解释性研究(Shapley counterfactual

表3　奖励级解释代表性算法总结

类别		代表算法	智能体		仿真环境	评估指标				
			单体	多体		忠实性	性能	效率	鲁棒性	用户性
奖励级解释	奖励分解	DrQ[138]	√	—	CliffWorld，Lunar Lander	√	√	—	—	—
		Shapley Counterfactual Credits[139]	—	√	Star Craft II	—	√	—	—	—
		SQD-DPG[140]	√	√	Cooperative Navigation，Prey-and-Predator，Traffic Junction	√	√	√	—	—
		“WHAT-IF” EXPLANATIONS[141]	√	—	ICU	√	√	—	—	—
		Mere Mortals[142]	—	√	Real-Time Strategy	√	√	√	—	√
		COMA[143]	—	√	StarCraft	—	√	—	—	—
	奖励塑造	ELLA[144]	√	—	BabyAI	—	√	—	—	—
		LEARN[145]	√	—	Montezuma’s Revenge	√	√	—	—	√
		TSP-PRL[146]	√	—	Charades-STA，ActivityNet	√	√	—	—	—
		SORL[147]	√	—	Office World，Montezuma’s Revenge	√	√	√	—	√
		RARE[148]	√	—	a live human-robot collaboration	√	√	—	—	—
		SDRL[149]	√	—	Taxi Domain，Montezuma’s Revenge	√	√	√	—	—

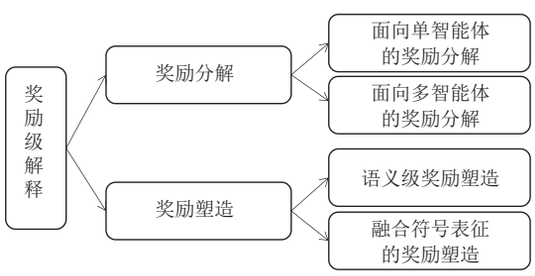

图9　奖励级解释

credits)[139]，针对现有研究中未充分考虑智能体之间交互行为的沙普利值作为信用分配的依据，同时，针对沙普利值计算的高复杂性，采用蒙特卡罗采样的近似方法来降低计算的复杂性，在星际争霸环境下进行方法的测试，证明了所提出方法优于现有合作型多智能体强化学习算法，并且在更为复杂的任务上取得了更高的收益.

同样地，针对合作式博弈问题进行信用分配方法的研究，提出了一种沙普利Q值计算的奖励分解方法(Shapley Q-value Deep Deterministic Policy Gradient，SQD-DPG)[140]，实现了智能体联盟中每个智能体对智能体合作联盟贡献的计算，从而克服了全局奖励信息不准确的问题，并且在合作导航、捕食者以及交通枢纽环境下进行实验验证，并通过可视化的方式对信用分配的公平性进行了可解释性验证. 针对合作型多智能体系统，反事实多智能体策略梯度方法(COunterfactual Multi-Agent，COMA)[143]解决信用分配问题，通过一个集中式的批评者和分散的行动者来优化策略，该方法显著提高了平均性能. 类似地，为在强化学习的研究之中融合专家知识，基于历史观测信息将反事实推理方法融合进逆向强化学习(“what-if” explanations)[141]方法之中，通过建模奖励函数来融合专家的经验知识，这为定义奖励函数和解释专家行为提供了依据，并且更加满足现实世界下的约束，这种方法通过估计不同行动的影响以更加适用于离策略强化学习方法之中，而不仅仅依赖于当前观测，更加有效地挖掘了历史经验信息，并且在真实的医疗环境中，对应用该方法的智能体决策的准确性以及可解释性进行了实验验证. 在对可解释性方法的综合性评估方法中，设定了一个现实验证实验(Mere Mortals)[142]，针对常见的强化学习可视化方法，显著性映射方法以及奖励分解方法在124位人类参与者实验中进行实验验证，比较在一个简单的实时策略游戏中应用不同解释方法对参与者的心理变化的影响.

4.3.2　奖励塑造

奖励塑造(reward shaping)的方法多用于单智能体强化学习可解释性研究之中，由于智能体的奖励函数多为基于先验知识，存在有一定的误差和不确定性，针对这种不确定性往往通过奖励塑造的方式进行规范化. 奖励塑造这一过程，本质上讲就是一个挖掘可解释性因素的过程，在前文介绍的相关可解释性方法之中，例如，语义信息建模、符号表征等方法，往往可转化为一种奖励信息，融入至智能体的奖励函数之中以提升强化学习策略的可解释性与性能表现. 奖励塑造方法被广泛应用于自驱动强化

学习研究之中，通过挖掘智能体内在的奖励机制，以有效地提升智能体策略训练的性能.

(1)语义级奖励塑造研究方法. 语言摘要学习的探索方法[144] (Exploration through Learned Language Abstraction, ELLA)基于奖励塑造方法将高级指令与低级指令进行关联，针对语言控制类游戏实现了摘要的生成，有效提升了稀疏奖励环境下的样本效率. 具体来说，ELLA搭建了一个终端分类器，用于识别智能体何时完成低级指令，一个相关度分类器，将低级指令与高级指令相关联，通过在线学习且可解释的方式实现了对智能体的有效控制，并在一套较为复杂的Baby AI环境下，进行了实验验证. 一种语言-行动奖励网络学习架构(Language-Action Reward Network, LEARN)[145]，将人类自然语言指令融入奖励信息之中，从而实现对奖励的重塑，这种自然语言驱动的中间奖励信息可以无缝地嵌入到任何标准的强化学习算法之中，高效地实现了人机融合，提升了人类对智能体行为的理解与控制，针对蒙特祖玛复仇游戏下的十五个任务进行实验，最终发现在与环境互通相同次数的情况下，基于自然语言的奖励塑造方法完成任务成功率提升了60%. 面向视频理解场景，开展构建语义级指导任务[146]的研究. 针对现有研究效率低下、可解释性弱以及偏离人类感知机制的问题，启发于人类决策机制，构建了一个树结构的渐进强化学习框架(Tree-Structured Policy based Progressive Reinforcement Learning, TSP-PRL)，通过策略迭代过程来调整时间边界，语义信息被显式地引入树结构策略分支之中，通过两个任务的奖励信息来共同指导信用分配，以鼓励树节点之间的相互竞争，从而实现了将复杂黑盒策略构建为可解释性白盒模型，在两个视频场景上的实验结果表明TSP-PRL具有更强的性能表现.

(2)融合符号表征的奖励塑造研究方法. 将符号知识嵌入至深度强化学习方法之中以提升其可解释性以及迁移性，搭建出了一个循环训练程序(Symbolic Options for Reinforcement Learning, SORL)[147]，可根据智能体模型以及交互轨迹自动学习符号表征，符号表征作为外部奖励信息，被用来重塑智能体学习过程中的奖励函数，该方法降低了对领域专家知识的高度依赖性，挖掘出了策略模型内在的可解释性，并且提高了数据效率，在蒙特祖玛实验环境下验证了该方法对数据效率、模型的可解释性以及移植性的提升.

类似地，将符号规划方法引入至分层强化学习方法之中，提出了一个符号深度强化学习框架(Symbolic Deep Reinforcement Learning, SDRL)[149]，通过符号操作与搭建的option集合相关联以实现任务级解释，具体来说，他们搭建了一个符号规划器-控制器-元控制器的体系结构，分别负责子任务调度、子任务学习以及子任务评估，控制器将从与环境交互得到的信息重塑为一个外部奖励，输入至元控制器之中，处理得到内部的子任务目标来进行子任务的学习，三个部件紧密衔接，实现了从低层到高层的策略学习与解释.

针对人机交互的协作式决策支持系统在故障恢复以及风险控制上的缺陷，提出了一个新的框架(Reward Augmentation and Repair through Explanation, RARE)[148]来改进协作系统的理解力以及执行力，通过将次优符号信息表征为一种奖励信息来对奖励函数进行重塑，并且向人类进行反馈，以规范其未来的决策行为，从而有效规避了风险的发生并且支持及时的故障恢复，并且，在一个人机交互的协作式场景下进行了实验验证.

在奖励级解释方法之中，奖励分解方法旨在将奖励函数进行拆分，以分析不同行为或是不同智能体的贡献. 由于奖励函数的构建存在很多人为主观性，导致该方法面临着奖励函数分解难以建立语义关联关系. 并且，面对复杂的混合博弈场景，无法将合作博弈的信用分配机制迁移其中. 同样地，奖励塑造方法在将语义信息以及逻辑符号引入其中时，面临着人为主观性强、策略迁移性差以及策略不确定性大等问题. 这些问题也是提升奖励级解释研究的关键难点.

4.4 策略级解释

在以上三个类别的可解释性方法研究中，更多地关注于强化学习范式中的动作、状态以及奖励方面的可解释性研究，如何针对智能体长期的、高层的以及全局性的策略进行解释是一个核心的难点，因此，我们针对这种策略级解释方法进行整理以及总结性描述，具体来讲，分为策略分解方法以及策略聚合两类方法. 策略分解方法旨在对强化学习的高层策略进行层次化分解，长期策略进行时序性分解，全局性策略进行局部性分解. 策略聚合可解释性方法旨在对智能体学习过程中的状态进行聚合性分析，以挖掘策略对智能体学习过程中状态的影响，来直接生成策略级的解释. 具体方法总结如表4所示. 策略级解释研究架构如图10所示.

表4　策略级解释代表性算法总结

类别		代表算法	智能体		仿真环境	评估指标				
			单体	多体		忠实性	性能	效率	鲁棒性	用户性
策略级解释	策略分解	HAL[150]	—	√	MuJoCo, CLEVR, Crafting Environment	√	√	—	—	—
		CARE[151]	√	—	Meta-World	√	√	—	—	√
		ECUM[152]	√	√	Gridworld, HIV Simulator, PAC-MAN	—	√	—	—	√
		HISA[153]	—	√	Minecraft	—	√	—	—	—
		MPHRL[154]	√	—	MuJoCo ant, Stacker arm	—	√	—	—	—
		SDRL[149]	√	—	Taxi Domain, Montezuma's Revenge	√	√	√	—	—
	状态聚合	Boolean Task Algebra[155]	√	—	Four Rooms, High-dimensional Video Game	√	√	—	—	√
		ANOVA[156]	√	—	Pong, Car	√	√	√	—	√
		TLdR[157]	√	—	Amazon Mechanical Turk	√	√	√	—	√
		Dot-to-Dot[62]	√	—	FetchPush, FetchPickAndPlace, HandManipulateBlock	√	√	—	—	—
		APG[91]	√	—	PrereqWorld	√	√	—	—	—
		Understanding DQNs[56]	√	—	Gridworld, Atari	√	√	√	—	√

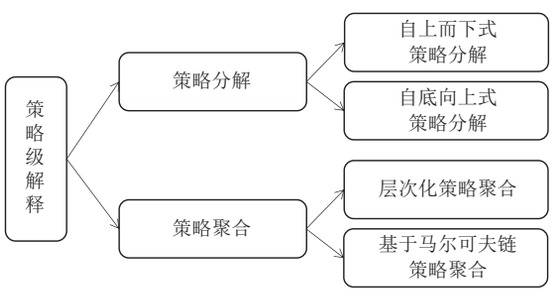

图10　策略级解释

4.4.1　策略分解

策略分解的示意图可直观地概括为图11，将上层策略进行子策略模型的分解，并与环境交互实现多级策略的更新. 具体可以分为自上而下的策略分解方法以及自底向上的策略分解方法两个类别.

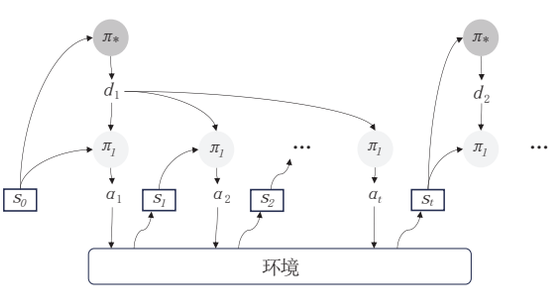

图11　策略分解方法示意图

(1)自上而下式策略分解. 在具体的研究之中，面向复杂的时序性任务，借助于语言的快速学习和泛化能力，使用语言生成指导性摘要进行分层深度强化学习(Hierarchical Abstraction with Language, HAL)[150]，搭建了一个指令级的低层策略以及一个跨任务摘要重用的高层策略学习框架以实现结构化语言级的智能体推理. 在MuJoCo以及CLEVR多智能体环境下进行智能体的训练与评估，通过实验结果的分析，证明了自然语言的组合特性对于概括和生成多样化的子技能是关键的，层次化任务摘要生成的方法实现了策略的可解释性提升，在人类理解的前提下，实现了层次化强化学习智能体推理学习.

课程学习[158]是一个逐步从简单任务到复杂问题解决的知识分解学习过程，策略分解方法可有效地应用于课程学习之中，从而实现由子任务到最终任务的解决. 在面向多任务学习之中，单任务以及多任务之间信息的挖掘是一种重要的可解释性方法，针对多任务强化学习方法，引入上下文表征信息进行跨任务共享信息的挖掘(Contextual Attention-based Representation Learning, CARL)[151]，基于元数据来学习可解释性的数据表征，作为上下文嵌入融入至状态表征信息之中. 多任务之间上下文的共享信息挖掘是实现知识迁移的重要方法，最终在Meta-World环境下的50个任务中进行了实验验证. 针对智能体学习的策略进行总结性的摘要提取，可以直观且具有可解释性地展示出智能体行为的优缺点，提出了一种基于模仿学习的策略总结方法(Exploring Computational User Models, ECUM)[152]，通过摘要提取模型以及策略模型的匹配和搭建，实现了高性能的策略总结，并且在模拟世界环境以及模拟药物研制环境下进行了实验验证.

在进行符号化深度强化学习的研究中，搭建了

一个策略分解的体系结构[149]，分别执行子任务调度，子任务学习以及任务评估，通过子任务的学习，以及长期任务规划和知识嵌入的方法，来实现对全局策略的提升，基于符号化的表示，也实现了对子任务的可解释性表征，最终的实验结果验证了方法的可解释性以及对数据效率的提升.

面向多智能体的环境，同样针对多任务学习问题，搭建了一个分层策略学习架构(Hierarchy and Interpretable Skill Acquisition, HISA)[153]，来控制智能体执行已有策略还是学习新策略，每个学习到的任务都对应于一个人类语言表述，基于这些融合到分层框架中的表述，智能体来获取已有策略，通过这种层次化策略以及可解释性的表述，高效实现了不同任务的技能组合以及任务执行，在多任务强化学习研究之中，引入了可解释性的方法，构建了一种人机交互的新范式.

(2)自底向上式策略分解. 针对传统策略分解方法中自上而下的结构进行改进，一个新的自下而上强化学习策略分解方法(Model Primitive Hierarchical lifelong Reinforcement Learning, MPHRL)[154]通过自底向上的方式，对单个的元任务进行自动分解，并搭建出一个控制器进行子策略的协调，实现了对复杂任务的模块化子策略分解. 并且，在复杂的单任务学习以及终身学习任务中进行了有效性的验证，最终的消融实验也验证了所提出的策略分解框架中不同元素的重要性与鲁棒性.

4.4.2 策略聚合

相对于策略分解的可解释性研究方法，策略聚合的方法旨在通过智能体任务或者状态等信息的聚合来实现对智能体策略中重要信息的汇集以进行可解释性分析，这种方法通常设定层次化的结构或者基于马尔可夫链进行策略的状态聚合性分析.

(1)层次化策略聚合. 针对组合学习中任务的逻辑组合，形式化表征为了一个布尔代数(Boolean task algebra)[155]，该方法可借助布尔代数中的连接符号实现新任务的设定，并通过对值函数的组合实现了智能体最优策略的学习，这种方法在保证任务组合的可解释性前提下，实现了针对新任务的组合与技能学习. 在面向机器人行为决策的研究中，通过聚合状态信息，挖掘关键的状态数据来辅助人类对机器人行为决策的理解(Establishing Appropriate Trust via Critical States, EATCS)[156]，通过在人机交互过程中反馈关键状态信息，来让人类明确何时进行策略的部署以及有效的控制.

针对最短路径问题，搭建了一个层次化体系架构(Temporal Abstraction through Landmark Recognition, TALR)[157]将智能体状态进行聚合分析，总结为重要的地标表征信息，根据地标信息按照时间顺序生成出策略摘要，并且以有向图结构的形式，直观呈现出最短路径，该方法实现了对最短路径问题有效且可解释性的研究. 层次化任务聚合研究框架(dot-to-dot)[62]，通过搭建一个低级策略来有效地处理智能体的状态以及动作预测，高级策略通过指令化的形式来学习环境和任务的聚合性动态表征. 高级策略生成子目标并分配给低级策略进行执行，与环境完成交互，并反馈状态转移信息至高级策略之中进行规划，生成可解释性表示信息. 这种点对点式的策略分析方法可以有效地为人类提供解释，并且，在机器人模拟环境中进行了可解释性与方法性能的验证.

(2)基于马尔可夫链策略聚合. 一种策略摘要图生成方法(Abstracted Policy Graphs, APG)[91]基于强化学习马尔可夫链，使用值函数以及状态转移信息进行策略摘要图的搭建，该方法可以在未来智能体的动作预测过程中，实现策略的解释，其马尔可夫特性可使得该方法有效嵌入至现有许多的强化学习方法之中. 如图12所示，Zahavy等人[56]提出了一种半聚合马尔可夫决策过程理解DQN算法(understanding DQNs)，通过挖掘时空状态信息，搭建t-SNE图实现了对智能体状态的降维与聚合分析，以揭示深度Q值学习算法所学习的策略特征，来解释策略成功的原因，最终，在雅达利游戏中针对深度强化学习算法进行了广泛的训练，解释以及评估验证. 在策略级解释方法之中，策略分解方法旨在通过层次化的结构实现任务级的策略分解，以逐步提升对策略的理解. 但是该方法面临着复杂任务场景下难以划分层次，不同子任务策略之间的相关性难分析等问题. 此外，策略聚合方法则是将状态信息进行聚合分析，以理解智能体的决策行为. 后续，将策略分解与聚合的方法相结合以深入研究长期的策略是一个重要的研究思路.

4.5 人机交互式解释

针对强化学习实现策略的可解释性研究核心目的是为了提升人类对算法模型的理解与认识，从而基于这种可解释性去提升模型的可靠性、泛化性以及安全性. 针对人在回路的场景，融合人类知识以及强化学习的算法模型可有效解决现实世界的应用. 进行人机交互式的可解释性方法研究是提

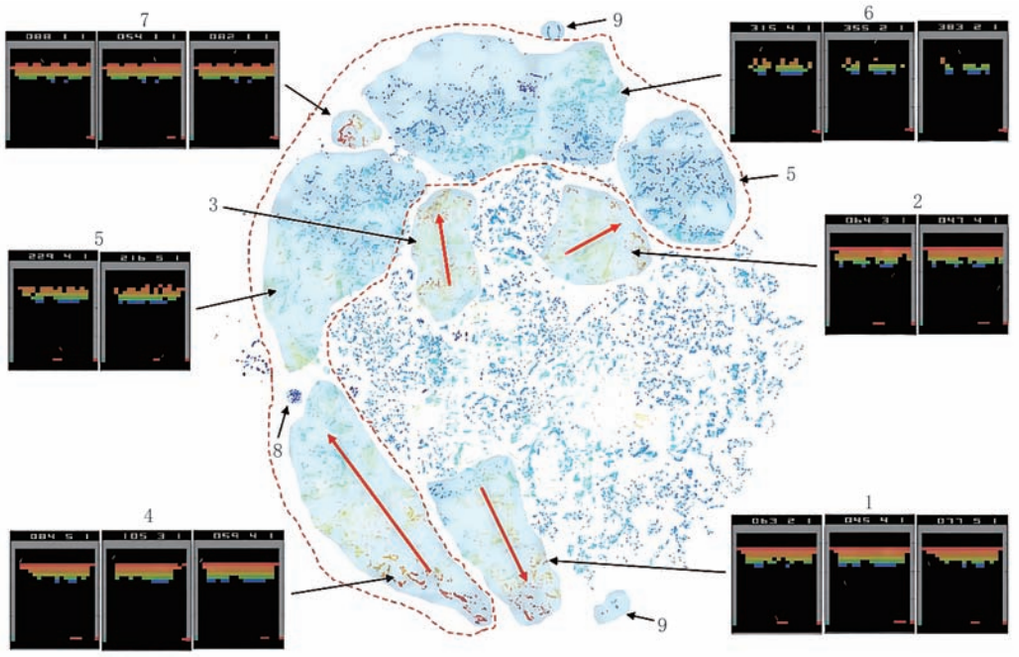

图12　Breakout游戏环境下的状态聚合3D的t-SNE散点图[56]

升现实世界下强化学习方法性能的一种重要途径. 针对现有人机交互式可解释性方法中，人类用户融入强化学习研究中的角色功能，本文分为人类指令式解释、人机协同式解释以及知识融合式解释三类方法. 人机交互式解释代表性算法总结如表5所示.

表5　人机交互式解释代表性算法总结

类别		代表算法	智能体		仿真环境	评估指标				
			单体	多体		忠实性	性能	效率	鲁棒性	用户性
人机交互式解释	人类指令式解释	ASK YOUR HUMANS[159]	—	√	crafting-based world	√	√	—	—	√
		LEARN[145]	√	—	Montezuma's Revenge	√	√	—	—	√
		TAMER[160]	√	—	Tetris	—	√	√	—	—
	人机协同式解释	EXPAND[161]	√	—	Pixel-Taxi and fourAtari games	—	√	√	—	√
		RARE[162]	—	√	a color-based collaborative Sudoku	√	—	—	—	√
		HITL[163]	—	—	NONE	—	√	√	√	√
	知识融合式解释	T2FS[164]	√	—	indoor airquality assessment	√	—	—	—	√
		KoGuN[165]	√	—	CartPole, LunarLander, FlappyBird, LunarLanderContinuous	—	√	√	—	—
		BDI[166]	√	√	MountainCar, 空战机动环境	—	√	√	—	√

(1)人类指令式解释方法. 将人类用户的自然语言指令作为引导智能体快速学习的信号，来提升智能体面对复杂场景下的奖励学习效率，这种方法可以有效地应对稀疏奖励场景下学习效率低的问题[167-169]. 面向强化学习研究中面临的稀疏奖励场景，针对分层强化学习方法，融入了人类自然语言信息来进行智能体行为指导，构造出了一个人类自然语言指令数据集[159]，通过搭建一个基于自然语言的高级指令生成器以及一个低级策略进行稀疏奖励场景下的多任务强化学习策略学习. 结合人类语言指令后，可以解决零样本设定下的问题，有效地提升了策略的泛化性以及快速适应性. 最终，在网格世界实验环境下进行了策略的泛化性以及多任务策略学习的性能验证. 类似地，随后，Goyal等人[145]基于人类自然语言信息构建出了一个密集奖励函数，对马尔可夫决策过程引入一个语言指令信号，通过搭建一个语言-行动-奖励网络来估计智能体从人类语言中获取的指令信息，基于指令信息构造出一个内在

的语言指令级奖励，有效应对了稀疏奖励下的智能体学习效率低的问题. TAMER[160]方法将人类专家引入到智能体的学习循环中，可以通过人类向智能体提供奖励信号，以指导智能体的训练过程，从而快速完成目标任务.

(2)人机协同式解释方法. 人类的解释能力可以被用来扩展人与智能体之间的通信协作方式，Guan等人[161]首次将人类视觉解释融入至人机交互强化学习方法之中，参与其中的人类用户不仅对状态-动作对提供优劣的评价，并且，针对图像中的相关特征进行视觉级解释. 这种融合人类解释的强化学习范式可以有效提升样本效率，并在雅达利游戏以及出租车模拟场景下展现了良好的效果. 在面向机器人与人类合作的场景中，可以通过构建一个共享的心理模型[162]来提升协作效率. 设计一个自主系统检测方法来检测人类与智能体之间的行为差异，同时，分析产生这些差异的原因，以及差异不解决会带来的潜在后果，最终，提供一个可解释性的反馈信息以指导对机器人行为的修正. 该方法减少了在人机联合任务执行过程中付出的昂贵且危险的代价，并且验证了在人机交互研究中基于这种策略修复的可解释性指导方法的有效性. 同样的，面向人在回路问题，一种循环式人机交互学习方法[163]被提出，通过主动识别会导致异常的样本来预测未来的探索行为，以提高安全性，在这一过程中，融入了数据质量评价，安全性保证以及用户信任度等指标进行验证，特别地，该方法适用于人为主观环境之中，可有效发挥人类知识信息的引导作用. 但是，人机协同式解释方法仍处于初步探索阶段，还需要大量的实验测试以及改进验证.

(3)知识融合式解释方法. 模糊逻辑[164, 170-171]可以提供出一个有效的范式，通过与环境相同的不确定和不精确的形式来表示人类的知识. 将人类的先验知识融入至强化学习之中，人类先验次优知识融合的强化学习框架[165]被搭建出来. 该框架包含一个模糊规则控制器以及微调模块进行先验知识微调. 端到端的训练框架可以适用在基于策略的强化学习算法之中，在离散以及连续控制任务上进行实验，实验结果表明即便是低性能的人类先验知识的引入，也可以有效地提升强化学习算法的学习效率，实现了人机知识融合下强化学习研究的开创性探索.

面向多智能体复杂环境，通过搭建融合认知行为模型的深度强化学习框架[166]，将领域内先验知识建模为基于信念-愿望-意图的认知行为模型，用于引导智能体策略学习，将已有知识表示为人和学习型智能体之间相互可理解的形式，并有效地加速策略收敛. 在迁移学习之中，融合人类知识的强化学习方法可有效地加速策略的收敛，从而快速适应新的场景以及任务[172].

5 挑战与发展

随着面向强化学习的可解释性研究的深入发展，面临着越来越多的挑战性难题. 首先，缺少面向强化学习可解释性研究的统一评估标准，这一问题限制了可解释性研究的快速发展. 同时，在进行可解释性算法模型搭建时，如何平衡模型的决策性能以及可解释性能，决定了可解释性方法能否实现广泛的应用. 其次，面向多智能体强化学习的可解释性研究仍处于起步阶段，如何开展面向多智能体系统的解释性研究具有很大的挑战. 并且，针对现有强化学习新的研究模式，例如，离线强化学习，无监督强化学习等，开展可解释性研究面临着更新的挑战. 除此之外，如何更好地结合人类专家知识，是提升可解释性方法性能的关键.

(1)如何对可解释性方法进行统一评估?

现有面向强化学习的可解释性方法研究中，进行实验验证的环境多种多样，有经典的雅达利游戏场景，有控制类的MuJoCo环境，也有高仿真度的模拟世界场景. 目前，缺少针对不同研究场景和任务的统一测试环境以及测试数据集. 与此同时，在进行可解释性的验证评估时，不同的研究工作设定有不同的可解释性评价标准，例如有策略契合度，解释忠实度等多种不统一的计算标准[173,61,174]. 因此，无法针对不同场景下的可解释性研究方法实现统一的评估，这也是限制强化学习可解释性研究快速发展的重要瓶颈. 在未来的研究中，亟需针对不同类型场景构建规范化的测试环境以及测试数据，并且设定规范化、理论性的评估指标体系以支撑对可解释性方法的准确评估.

(2)如何对模型的综合性能进行平衡?

开展可解释性研究时，往往需要搭建白盒模型或者人类可理解的形式化模型，这些方法在实现可解释的同时，大大降低了强化学习算法的决策性能，这一问题限制了可解释性方法在关键领域之中的应用. 例如，在军事作战场景下，需要保证决策可解释性的同时，高效地应对战争态势变化，快速地给出战术决策. 该问题作为可解释性研究中面临的核心问

题，在现有的研究提出了很多应对的方法，例如，搭建可微的决策树模型，软决策树模型[175]等，但是，现有的方法仍然无法达到满足可解释的同时保证与原策略相匹配的决策性能．因此，如何平衡算法模型的决策和解释的综合性能将是开展未来研究的重中之重，这决定着可解释性方法能否在医疗、军事以及工业控制等关键领域实现应用与部署[176-178]．

(3)如何开展针对复杂多智能体系统的可解释性研究？

现有面向复杂的多智能体系统的可解释性研究中，多为基于单智能体强化学习可解释性研究方法的扩展，例如，针对多智能体系统中的单一智能体搭建树型的自解释模型，结合多项单智能体评估指标来综合评价多智能体系统中的策略学习情况，以及基于奖励分解的方法来实现多智能体系统中的信用分配等．但是，这些方法局限于合作或者竞争的单一场景之中，并且没有充分考虑到非完全理性的多智能体现实场景．在多智能体强化学习可解释性研究中，仍然面临着如何解释多智能体系统中智能体之间的竞争合作关系，如何解释非完全理性条件下的智能体行为决策，以及如何解释多智能体系统中单智能体策略与多智能体联盟策略之间的层次关系等诸多挑战．

(4)如何应对强化学习研究新模式下的可解释性挑战？

随着预训练大模型[179]的提出，面向强化学习的研究提出了越来越多的研究模式．不同于传统的马尔可夫决策过程，现有的研究基于数据驱动来进行智能体的模型的训练．例如，为了避免昂贵且危险的智能体在线交互而提出的离线强化学习方法[180-182]，实现了基于离线数据的智能体策略学习．无监督强化学习方法[183-185]则实现了不依赖于奖励信息或动作信息来挖掘离线的智能体轨迹信息，以进行智能体潜在技能的学习，这一学习模式也为实现预训练决策大模型提供了理论支撑．同时，多个决策大模型[186-188]的提出，也对面向决策大模型的可解释性研究提出了新的挑战．因此，不同于传统的马尔可夫决策过程进行的可解释性研究，在针对这些新颖的强化学习研究模式进行可解释性研究时，面临着技能规则提取，多任务分析以及大模型解释等大量的全新挑战．

(5)如何更好地结合人类的专家知识？

在本篇综述的方法介绍部分，针对人机交互式的可解释性方法进行了总结描述，但是现有的人机融合方法，多为人在回路或专家指令级的可解释性方法学习，仍旧是低效率的．如何更好地将人类专家知识嵌入至强化学习算法模型之中，如何基于人类先验知识实现规则策略的提取，是一项关键的挑战．受到元学习[189-192]的启发，基于智能体行为轨迹的源数据，通过挖掘源数据中的技能快速进行策略的学习是一个具有潜力的研究方法．同时，可以思考如何更好地结合模仿学习的方法，来融合专家知识以实现智能体高效决策．

6 总 结

随着强化学习研究的逐渐深入，针对强化学习的可解释性研究受到了越来越多的关注，涌现出了大量的工作．本文从强化学习的经典模型马尔可夫决策过程出发，首先对强化学习的基础知识进行了介绍，随后，针对强化学习可解释性的定义以及评估指标进行总结．在方法描述部分，根据智能体在执行马尔可夫决策过程时的动作、状态、奖励以及策略，进行对应的分类，将强化学习的可解释性研究分为了行为级解释、特征级解释、奖励级解释及策略级解释，并且在每一类别的总结性描述中按照从单智能体到多智能体的顺序进行阐述，同时，对于可解释性研究的一个重要方向，面向人机交互式的可解释性研究方法进行了总结．最后，针对当前强化学习可解释性研究面临的挑战进行了总结，并对未来的研究方向进行了展望．

致 谢 本课题得到科技创新2030—“新一代人工智能”重大项目(2021ZD0113303)、国家自然科学基金(62192783, 62276128, 62276142, 62206133)、南京大学计算机软件新技术国家重点实验室资助项目(KFKT2022B12)资助．感谢各位审稿人专业严谨的评审！

参 考 文 献

[1] Sutton R S, Barto A G. Reinforcement Learning: An Introduction. Cambridge, USA: MIT press, 2018

[2] Zhou Z-H. Machine learning. Singapore: Springer Nature, 2021

[3] Rajeswaran A, Mordatch I, Kumar V. A game theoretic framework for model based reinforcement learning// Proceedings of the International Conference on Machine Learning.Virtual.2020: 7953-7963

[4] Qiu D, Dong Z, Zhang X, et al. Safe reinforcement learning for real-time automatic control in a smart energy-hub. Applied

Energy, 2022, 309: 118403
[5] Silver D, Hubert T, Schrittwieser J, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science, 2018, 362(6419): 1140-1144
[6] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518 (7540): 529-533
[7] Kaiser L, Babaeizadeh M, Milos P, et al. Model-based reinforcement learning for atari. arXiv preprint arXiv: 1903.00374, 2019
[8] Wang W, Guo J, Wang Z, et al. Abnormal flow detection in industrial control network based on deep reinforcement learning. Applied Mathematics and Computation, 2021, 409: 126379
[9] Bai Y, Jones A, Ndousse K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862, 2022
[10] Romdhana A, Merlo A, Ceccato M, et al. Deep reinforcement learning for black-box testing of android apps. ACM Transactions on Software Engineering and Methodology (TOSEM), 2022, 31(4): 1-29
[11] Otto F, Celik O, Zhou H, et al. Deep black-box reinforcement learning with movement primitives//Proceedings of the Conference on Robot Learning. Atlanta, USA,2023: 1244-1265
[12] Zhou S K, Le H N, Luu K, et al. Deep reinforcement learning in medical imaging: A literature review. Medical Image analysis, 2021, 73: 102193
[13] Wang H, Tang H, Hao J, et al. Large scale deep reinforcement learning in war-games//Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).Virtual, 2020: 1693-1699
[14] Ling Z, Ma H, Yang Y, et al. Explaining AlphaGo: Interpreting contextual effects in neural networks. arXiv preprint arXiv:1901.02184, 2019
[15] Zhuo T Y, Huang Y, Chen C, et al. Exploring AI ethics of ChatGPT: A diagnostic analysis. arXiv preprint arXiv: 2301.12867, 2023
[16] Wilson S W. Classifier fitness based on accuracy. Evolutionary Computation, 1995, 3(2): 149-175
[17] Chen H, Wang C, Huang J, et al. Efficient use of heuristics for accelerating XCS-based policy learning in Markov games. Swarm and Evolutionary Computation, 2021, 65: 100914
[18] Hao J, Yang T, Tang H, et al. Exploration in deep reinforcement learning: From single-agent to multiagent domain. IEEE Transactions on Neural Networks and Learning Systems, to appear
[19] Xu T, Li Z, Yu Y. Error bounds of imitating policies and environments for reinforcement learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(10): 6968-6980
[20] Gunning D, Stefik M, Choi J, et al. XAI—Explainable artificial intelligence. Science Robotics, 2019, 4(37): 7120
[21] Samek W, Müller K-R. Towards explainable artificial intelligence. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, 2019, 11700:5-22
[22] Milani S, Topin N, Veloso M, et al. A survey of explainable reinforcement learning. arXiv preprint arXiv:2202.08434, 2022
[23] Liu Xiao, Liu Shu-Yang, Zhuang Yun-Kai, Gao Yang. Explainable reinforcement learning: Basic problems exploration and method survey. Journal of Software, 2023, 34(5): 2300-2316 (in Chinese)
(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)
[24] Heuillet A, Couthouis F, Díaz-Rodríguez N. Explainability in deep reinforcement learning. Knowledge-Based Systems, 2021, 214: 106685
[25] Glanois C, Weng P, Zimmer M, et al. A survey on interpretable reinforcement learning. arXiv preprint arXiv: 2112.13112, 2021
[26] Qing Y, Liu S, Song J, et al. A survey on explainable reinforcement learning: Concepts, algorithms, challenges. arXiv preprint arXiv:2211.06665, 2022
[27] Puterman M L. Markov decision processes. Handbooks in Operations Research and Management Science, 1990, 2: 331-434
[28] Wagenmaker A J, Chen Y, Simchowitz M, et al. Reward-free rl is no harder than reward-aware rl in linear markov decision processes//Proceedings of the International Conference on Machine Learning. Baltimore, USA,2022: 22430-22456
[29] Arulkumaran K, Deisenroth M P, Brundage M, et al. Deep reinforcement learning: A brief survey. IEEE Signal Processing Magazine, 2017, 34(6): 26-38
[30] Zhang H, Yu T. Taxonomy of reinforcement learning algorithms. Deep Reinforcement Learning: Fundamentals, Research and Applications, 2020,1(3):125-133
[31] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning. arXiv preprint arXiv: 1312.5602, 2013
[32] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning//Proceedings of the AAAI Conference on Artificial Intelligence. Phoenix, USA,2016: 30
[33] Wang Z, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning//Proceedings of the International Conference on Machine Learning. New York, USA,2016: 1995-2003
[34] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017
[35] Wu Z, Yu C, Ye D, et al. Coordinated proximal policy optimization. Advances in Neural Information Processing Systems, 2021, 34: 26437-26448
[36] Williams R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning, 1992, 8: 229-256
[37] Konda V, Tsitsiklis J. Actor-critic algorithms//Proceedings of the 12th International Conference on Neural Information Processing Systems.Denver, USA,1999, 12:1008-1014
[38] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning. arXiv preprint arXiv: 1509.02971, 2015

[39] Belle V, Papantonis I. Principles and practice of explainable machine learning. Frontiers in Big Data, 2021, 4: 688969

[40] Tjoa E, Guan C. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(11): 4793-4813

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

[42] Angelov P P, Soares E A, Jiang R, et al. Explainable artificial intelligence: An analytical review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2021, 11(5): e1424

[43] Novakovsky G, Dexter N, Libbrecht M W, et al. Obtaining genetics insights from deep learning via explainable artificial intelligence. Nature Reviews Genetics, 2023, 24(2): 125-137

[44] Minh D, Wang H X, Li Y F, et al. Explainable artificial intelligence: A comprehensive review. Artificial Intelligence Review, 2022, 55(5): 1-66

[45] Bussmann N, Giudici P, Marinelli D, et al. Explainable machine learning in credit risk management. Computational Economics, 2021, 57(1): 203-216

[46] Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nature Machine Intelligence, 2020, 2(10): 573-584

[47] Loh H W, Ooi C P, Seoni S, et al. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011-2022). Computer Methods and Programs in Biomedicine, 2022, 226: 107161

[48] Ahmed I, Jeon G, Piccialli F. From artificial intelligence to explainable artificial intelligence in industry 4.0: A survey on what, how, and where. IEEE Transactions on Industrial Informatics, 2022, 18(8): 5031-5042

[49] Puiutta E, Veith E M S P. Explainable reinforcement learning: A survey//Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction. Dublin, Ireland, 2020: 77-95

[50] Wells L, Bednarz T. Explainable ai and reinforcement learning—a systematic review of current approaches and trends. Frontiers in Artificial Intelligence, 2021, 4: 550030

[51] Lipton Z C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 2018, 16(3): 31-57

[52] Biran O, Cotton C. Explanation and justification in machine learning: A survey//Proceedings of the IJCAI-17 Workshop on Explainable AI (XAI). Melbourne, Australia, 2017, 8: 8-13

[53] Molnar C. Interpretable machine learning, Lulu. com, 2020

[54] Mohseni S, Zarei N, Ragan E D. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Transactions on Interactive Intelligent Systems, 2021, 11(3-4): 1-45

[55] Ribeiro M T, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 1135-1144

[56] Ribeiro M T, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 32

[57] Zahavy T, Ben-Zrihem N, Mannor S. Graying the black box: Understanding dqns//Proceedings of the International Conference on Machine Learning. New York, USA, 2016: 1899-1908

[58] Groce A, Kulesza T, Zhang C, et al. You are the only possible oracle: Effective test selection for end users of interactive machine learning systems. IEEE Transactions on Software Engineering, 2013, 40(3): 307-323

[59] Krause J, Dasgupta A, Swartz J, et al. A workflow for visual diagnostics of binary classifiers using instance-level explanations//Proceedings of the 2017 IEEE Conference on Visual Analytics Science and Technology (VAST). Phoenix, USA, 2017: 162-172

[60] Tang Y, Ha D. The sensory neuron as a transformer: Permutation-invariant neural networks for reinforcement learning. Advances in Neural Information Processing Systems, 2021, 34: 22574-22587

[61] Kuhnle A, May M C, Schäfer L, et al. Explainable reinforcement learning in production control of job shop manufacturing system. International Journal of Production Research, 2022, 60(19): 5812-5834

[62] Beyret B, Shafti A, Faisal A A. Dot-to-dot: Explainable hierarchical reinforcement learning for robotic manipulation//Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Macao, China, 2019: 5014-5019

[63] Guo Y, Campbell J, Stepputtis S, et al. Explainable action advising for multi-agent reinforcement learning//Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA). London, UK, 2023: 5515-5521

[64] Milani S, Zhang Z, Topin N, et al. MAVIPER: Learning decision tree policies for interpretable multi-agent reinforcement learning//Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Grenoble, France, 2022: 251-266

[65] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 3319-3328

[66] Kindermans P-J, Hooker S, Adebayo J, et al. The (un) reliability of saliency methods. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, 2019, 11700(4): 267-280

[67] Binder A, Samek W, Montavon G, et al. Analyzing and validating neural networks predictions//Proceedings of the ICML 2016 Workshop on Visualization for Deep Learning. New York, USA, 2016: 107

[68] Gedikli Fatih, Jannach Dietmar, and Ge Mouzhi. How should I explain? A comparison of different explanation types for recommender systems. //Proceedings of the IJHCI. Crete, Greece, 2014, 4: 367-382

[69] Lim Brian Y, Dey Anind K, and Avrahami Daniel. Why and why not explanations improve the intelligibility of context-aware intelligent systems//Proceedings of the SIGCHI. Boston,USA, 2009:2119-2128

[70] Liu G, Schulte O, Zhu W, et al. Toward interpretable deep reinforcement learning with linear model u-trees//Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference, Dublin, Ireland, 2019: 414-429

[71] Bastani O, Pu Y, Solar-Lezama A. Verifiable reinforcement learning via policy extraction. Advances in Neural Information Processing Systems, 2018, 31

[72] Silva A, Gombolay M, Killian T, et al. Optimization methods for interpretable differentiable decision trees applied to reinforcement learning//Proceedings of the International Conference on Artificial Intelligence and Statistics. Palermo, Italy, 2020: 1855-1865

[73] Verma A, Murali V, Singh R, et al. Programmatically interpretable reinforcement learning//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 5045-5054

[74] Hein D, Udluft S, Runkler T A. Generating interpretable reinforcement learning policies using genetic programming// Proceedings of the Genetic and Evolutionary Computation Conference Companion. Prague, Czech Republic, 2019:23-24

[75] Landajuela M, Petersen B K, Kim S, et al. Discovering symbolic policies with deep reinforcement learning// Proceedings of the International Conference on Machine Learning. Virtual, 2021: 5979-5989

[76] Inala J P, Yang Y, Paulos J, et al. Neurosymbolic transformers for multi-agent communication. Advances in Neural Information Processing Systems, 2020, 33: 13597-13608

[77] Madumal P, Miller T, Sonenberg L, et al. Explainable reinforcement learning through a causal lens//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34: 2493-2500

[78] Stein G. Generating high-quality explanations for navigation in partially-revealed environments. Advances in Neural Information Processing Systems, 2021, 34: 17493-17506

[79] Fukuchi Y, Osawa M, Yamakawa H, et al. Application of instruction-based behavior explanation to a reinforcement learning agent with changing policy//Proceedings of the 24th International Conference, Guangzhou, China, 2017: 100-108

[80] Boggess K, Kraus S, Feng L. Toward policy explanations for multi-agent reinforcement learning. arXiv preprint arXiv: 2204.12568, 2022

[81] Zhu Y, Yin X, Chen C. Extracting Decision tree from trained deep reinforcement learning in traffic signal control. IEEE Transactions on Computational Social Systems, 2022, 10(4): 1997-2007

[82] Fonner D F, Coyle F P. Explainable machine learning models for evaluating government grantmaking//Proceedings of the 2022 IEEE International Conference on Big Data (Big Data). Osaka, Japan,2022: 2243-2248

[83] Wang Y-C, Chen T, Chiu M-C. An explainable deep-learning approach for job cycle time prediction. Decision Analytics Journal, 2023, 6: 100153

[84] Bechini A, Bárcena J L C, Ducange P, et al. Increasing accuracy and explainability in fuzzy regression trees: An experimental analysis//Proceedings of the 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Padua, Italy,2022: 1-8

[85] Katz G, Barrett C, Dill D L, et al. Reluplex: An efficient SMT solver for verifying deep neural networks//Proceedings of the Computer Aided Verification: 29th International Conference, Heidelberg, Germany, 2017: 97-117

[86] Custode L L, Iacca G. Evolutionary learning of interpretable decision trees. IEEE Access, 2023, 11: 6169-6184

[87] Bewley T, Lawry J. Tripletree: A versatile interpretable representation of black box agents and their environments// Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada,2021,35: 11415-11422

[88] Topin N, Milani S, Fang F, et al. Iterative bounding mdps: Learning interpretable policies via non-interpretable methods// Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada,2021, 35: 9923-9931

[89] Roth A M, Topin N, Jamshidi P, et al. Conservative q-improvement: Reinforcement learning for an interpretable decision-tree policy. arXiv preprint arXiv:1907.01180, 2019

[90] Custode L L, Iacca G. Evolutionary learning of interpretable decision trees. IEEE Access, 2023, 11: 6169-6184

[91] Topin N, Veloso M. Generation of policy-level explanations for reinforcement learning//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA,2019, 33: 2514-2521

[92] Zhang H, Zhou A, Lin X. Interpretable policy derivation for reinforcement learning based on evolutionary feature synthesis. Complex & Intelligent Systems, 2020, 6(3): 741-753

[93] Hein D, Udluft S, Runkler T A. Interpretable policies for reinforcement learning by genetic programming. Engineering Applications of Artificial Intelligence, 2018, 76: 158-169

[94] Cao Y, Li Z, Yang T, et al. GALOIS: Boosting deep reinforcement learning via generalizable logic synthesis. Advances in Neural Information Processing Systems, 2022, 35: 19930-19943

[95] Hein D, Hentschel A, Runkler T, et al. Particle swarm optimization for generating interpretable fuzzy reinforcement learning policies. Engineering Applications of Artificial Intelligence, 2017, 65: 87-98

[96] Jiang Z, Luo S. Neural logic reinforcement learning// Proceedings of the International Conference on Machine Learning. Long Beach, USA,2019: 3110-3119

[97] Trivedi D, Zhang J, Sun S-H, et al. Learning to synthesize programs as interpretable and generalizable policies. Advances in Neural Information Processing Systems, 2021, 34: 25146-25163

[98] Chester A, Dann M, Zambetta F, et al. SAGE: Generating symbolic goals for myopic models in deep reinforcement learning. arXiv preprint arXiv:2203.05079, 2022

[99] Jin P, Tian J, Zhi D, et al. Trainify: A cegar-driven training and verification framework for safe deep reinforcement learning//Proceedings of the International Conference on Computer Aided Verification. Haifa, Israel,2022: 193-218

[100] Payani A, Fekri F. Incorporating relational background knowledge into reinforcement learning via differentiable inductive logic programming. arXiv preprint arXiv:2003.10386, 2020

[101] Zhang L, Li X, Wang M, et al. Off-policy differentiable logic reinforcement learning//Proceedings of the Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, Bilbao, Spain, 2021: 617-632

[102] Olson M L, Khanna R, Neal L, et al. Counterfactual state explanations for reinforcement learning agents via generative deep learning. Artificial Intelligence, 2021, 295: 103455

[103] Ehsan U, Tambwekar P, Chan L, et al. Automated rationale generation: a technique for explainable AI and its effects on human perceptions//Proceedings of the 24th International Conference on Intelligent User Interfaces. Vienna, Austria, 2019: 263-274

[104] Hayes B, Shah J A. Improving robot controller transparency through autonomous policy explanation//Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction.Vienna, Austria, 2017: 303-312

[105] Ehsan U, Harrison B, Chan L, et al. Rationalization: A neural machine translation approach to generating natural language explanations//Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. New Orleans, USA, 2018: 81-87

[106] Fukuchi Y, Osawa M, Yamakawa H, et al. Autonomous self-explanation of behavior for interactive reinforcement learning agents//Proceedings of the 5th International Conference on Human Agent Interaction. Bielefeld, Germany,2017: 97-101

[107] Kazak Y, Barrett C, Katz G, et al. Verifying deep-RL-driven systems. //Proceedings of the 2019 Workshop on Network Meets AI & ML. New York, USA,2019: 83-89

[108] Zhu H, Xiong Z, Magill S, et al. An inductive synthesis framework for verifiable reinforcement learning//Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation. 2019: 686-701

[109] Anderson G, Verma A, Dillig I, et al. Neurosymbolic reinforcement learning with formally verified exploration. Advances in Neural Information Processing Systems, 2020, 33: 6172-6183

[110] Guo W, Wu X, Khan U, et al. Edge: Explaining deep reinforcement learning policies. Advances in Neural Information Processing Systems, 2021, 34: 12222-12236

[111] Heuillet A, Couthouis F, Díaz-Rodríguez N. Collective explainable AI: Explaining cooperative strategies and agent contribution in multiagent reinforcement learning with shapley values. IEEE Computational Intelligence Magazine, 2022, 17(1): 59-71

[112] Xu Y, Fang M, Chen L, et al. Deep reinforcement learning with stacked hierarchical attention for text-based games. Advances in Neural Information Processing Systems, 2020, 33: 16495-16507

[113] Lee L, Eysenbach B, Salakhutdinov R R, et al. Weakly-supervised reinforcement learning for controllable behavior. Advances in Neural Information Processing Systems, 2020, 33: 2661-2673

[114] Van Der Waa J, Van Diggelen J, Bosch K V D, et al. Contrastive explanations for reinforcement learning in terms of expected consequences. arXiv preprint arXiv:1807.08706, 2018

[115] Zheng J, Liu S, Ni L M. Robust bayesian inverse reinforcement learning with sparse behavior noise//Proceedings of the AAAI Conference on Artificial Intelligence. Québec, Canada,2014: 28

[116] Annasamy R M, Sycara K. Towards better interpretability in deep q-networks//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA,2019, 33: 4561-4569

[117] Greydanus S, Koul A, Dodge J, et al. Visualizing and understanding atari agents//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 1792-1801

[118] Yau H, Russell C, Hadfield S. What did you think would happen? Explaining agent behaviour through intended outcomes. Advances in Neural Information Processing Systems, 2020, 33: 18375-18386

[119] Wang J, Gou L, Shen H-W, et al. Dqnviz: A visual analytics approach to understand deep q-networks. IEEE Transactions on Visualization and Computer Graphics, 2018, 25(1): 288-298

[120] Pan M, Huang W, Li Y, et al. xgail: Explainable generative adversarial imitation learning for explainable human decision analysis//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. San Diego, USA, 2020: 1334-1343

[121] Leurent E, Mercat J. Social attention for autonomous decision-making in dense traffic. arXiv preprint arXiv:1911.12250, 2019

[122] Zhang K, Zhang J, Xu P-D, et al. Explainable AI in deep reinforcement learning models for power system emergency control. IEEE Transactions on Computational Social Systems, 2021, 9(2): 419-427

[123] Lütjens B, Everett M, How J P. Safe reinforcement learning with model uncertainty estimates//Proceedings of the 2019 International Conference on Robotics and Automation (ICRA). Montreal, Canada,2019: 8662-8668

[124] Liu Z, Zhu Y, Chen C. N. Q: Neural attention additive model for interpretable multi-agent Q-learning. arXiv preprint arXiv: 2304.13383, 2023

[125] Bertoin D, Zouitine A, Zouitine M, et al. Look where you look! Saliency-guided Q-networks for generalization in visual Reinforcement Learning. Advances in Neural Information Processing Systems, 2022, 35: 30693-30706

[126] Sarkar S, Babu A R, Gundecha V, et al. RL-CAM: Visual explanations for convolutional networks using reinforcement learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 3860-3868

[127] Barnby J M, Mehta M A, Moutoussis M. The computational

relationship between reinforcement learning, social inference, and paranoia. PLoS Computational Biology, 2022, 18(7): e1010326

[128] Yang Z, Bai S, Zhang L, et al. Learn to interpret atari agents. arXiv preprint arXiv:1812.11276, 2018

[129] Tang Y, Nguyen D, Ha D. Neuroevolution of self-interpretable agents//Proceedings of the 2020 Genetic and Evolutionary Computation Conference.Cancún, Mexico, 2020: 414-424

[130] Petsiuk V, Das A, Saenko K. Rise: Randomized input sampling for explanation of black-box models. arXiv preprint arXiv:1806.07421, 2018

[131] Guo S S, Zhang R, Liu B, et al. Machine versus human attention in deep reinforcement learning tasks. Advances in Neural Information Processing Systems, 2021, 34: 25370-25385

[132] Sequeira P, Gervasio M. Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations. Artificial Intelligence, 2020, 288: 103367

[133] Pan X, Chen X, Cai Q, et al. Semantic predictive control for explainable and efficient policy learning//Proceedings of the 2019 International Conference on Robotics and Automation (ICRA). Montreal, Canada, 2019: 3203-3209

[134] Wäldchen S, Pokutta S, Huber F. Training characteristic functions with reinforcement learning: Xai-methods play connect four//Proceedings of the International Conference on Machine Learning. Baltimore, USA,2022: 22457-22474

[135] Iyer R, Li Y, Li H, et al. Transparency and explanation in deep reinforcement learning neural networks//Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.New Orleans, USA, 2018: 144-150

[136] Goel V, Weng J, Poupart P. Unsupervised video object segmentation for deep reinforcement learning. Advances in Neural Information Processing Systems, 2018, 31: 5683-5694

[137] Mishra I, Dao G, Lee M. Visual sparse Bayesian reinforcement learning: a framework for interpreting what an agent has learned//Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI). Bangalore, India, 2018: 1427-1434

[138] Juozapaitis Z, Koul A, Fern A, et al. Explainable reinforcement learning via reward decomposition//Proceedings of the IJCAI/ECAI Workshop on Explainable Artificial Intelligence. Macao, China, 2019

[139] Li J, Kuang K, Wang B, et al. Shapley counterfactual credits for multi-agent reinforcement learning//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. Virtual,2021: 934-942

[140] Wang J, Zhang Y, Kim T-K, et al. Shapley Q-value: A local reward approach to solve global reward games//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34: 7285-7292

[141] Bica I, Jarrett D, Hüyük A, et al. Learning" what-if" explanations for sequential decision-making. arXiv preprint arXiv:2007.13531, 2020

[142] Anderson A, Dodge J, Sadarangani A, et al. Explaining reinforcement learning to mere mortals: An empirical study. arXiv preprint arXiv:1903.09708, 2019

[143] Foerster J, Farquhar G, Afouras T, et al. Counterfactual multi-agent policy gradients//Proceedings of the AAAI Conference on Artificial Intelligence.New Orleans, USA, 2018: 32

[144] Mirchandani S, Karamcheti S, Sadigh D. Ella: Exploration through learned language abstraction. Advances in Neural Information Processing Systems, 2021, 34: 29529-29540

[145] Goyal P, Niekum S, Mooney R J. Using natural language for reward shaping in reinforcement learning. arXiv preprint arXiv: 1903.02020, 2019

[146] Wu J, Li G, Liu S, et al. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34: 12386-12393

[147] Jin M, Ma Z, Jin K, et al. Creativity of ai: Automatic symbolic option discovery for facilitating deep reinforcement learning// Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2022, 36: 7042-7050

[148] Tabrez A, Hayes B. Improving human-robot interaction through explainable reinforcement learning//Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). Daegu, Korea,2019: 751-753

[149] Lyu D, Yang F, Liu B, et al. SDRL: interpretable and data-efficient deep reinforcement learning leveraging symbolic planning//Proceedings of the AAAI Conference on Artificial Intelligence.Hawaii, USA, 2019, 33: 2970-2977

[150] Jiang Y, Gu S S, Murphy K P, et al. Language as an abstraction for hierarchical deep reinforcement learning. Advances in Neural Information Processing Systems, 2019, 32

[151] Sodhani S, Zhang A, Pineau J. Multi-task reinforcement learning with context-based representations//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 9767-9779

[152] Lage I, Lifschitz D, Doshi-Velez F, et al. Exploring computational user models for agent policy summarization// Proceedings of the IJCAI: Proceedings of the Conference. Macao, China,2019, 28: 1401

[153] Shu T, Xiong C, Socher R. Hierarchical and interpretable skill acquisition in multi-task reinforcement learning. arXiv preprint arXiv:1712.07294, 2017

[154] Wu B, Gupta J K, Kochenderfer M. Model primitives for hierarchical lifelong reinforcement learning. Autonomous Agents and Multi-Agent Systems, 2020, 34: 1-38

[155] Nangue Tasse G, James S, Rosman B. A boolean task algebra for reinforcement learning. Advances in Neural Information Processing Systems, 2020, 33: 9497-9507

[156] Huang S H, Bhatia K, Abbeel P, et al. Establishing appropriate trust via critical states//Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid, Spain, 2018: 3929-3936

[157] Sreedharan S, Srivastava S, Kambhampati S. Tldr: Policy summarization for factored ssp problems using temporal

abstractions//Proceedings of the International Conference on Automated Planning and Scheduling.Nancy，France，2020，30：272-280

[158] Wang X，Chen Y，Zhu W. A survey on curriculum learning. IEEE Transactions on Pattern Analysis and Machine Intelligence，2021，44(9)：4555-4576

[159] Chen V，Gupta A，Marino K. Ask your humans：Using human instructions to improve generalization in reinforcement learning. arXiv preprint arXiv：2011.00517，2020

[160] Knox W B，Tamer Stone P. Training an agent manually via evaluative reinforcement//Proceedings of the IEEE International Conference on Development and Learning. Monterey，USA，2008：292-297

[161] Guan L，Verma M，Guo S S，et al. Widening the pipeline in human-guided reinforcement learning with explanation and context-aware data augmentation. Advances in Neural Information Processing Systems，2021，34：21885-21897

[162] Tabrez A，Agrawal S，Hayes B. Explanation-based reward coaching to improve human performance via reinforcement learning//Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). Daegu，Korea，2019：249-257

[163] Liu Y，Luo Y，Zhong Y，et al. Sequence modeling of temporal credit assignment for episodic reinforcement learning. arXiv preprint arXiv：1905.13420，2019

[164] Ghorbani A，Zamanifar K. Type-2 fuzzy ontology-based semantic knowledge for indoor air quality assessment. Applied Soft Computing，2022，121：108658

[165] Zhang P，Hao J，Wang W，et al. KoGuN：accelerating deep reinforcement learning via integrating human suboptimal knowledge. arXiv preprint arXiv：2002.07418，2020

[166] Chen Hao，Li Jia-Xiang，Huang Jiang，et al. Deep reinforcement learning framework and algorithms integrated with cognitive behavior models. Control and Decision，2023，38(11)：3209-3218 (in Chinese)
(陈浩，李嘉祥，黄健等.融合认知行为模型的深度强化学习框架及算法.控制与决策，2023，38(11)：3209-3218)

[167] Ng A Y，Harada D，Russell S. Policy invariance under reward transformations：Theory and application to reward shaping//Proceedings of the ICML. Bled，Slovenia，1999，99：278-287

[168] Harutyunyan A，Dabney W，Mesnard T，et al. Hindsight credit assignment. Advances in Neural Information Processing Systems，2019，32

[169] Rajendran P T，Espinoza H，Delaborde A，et al. Human-in-the-loop learning for safe exploration through anomaly prediction and intervention//Proceedings of the SafeAI@AAAI. Vancouver，Canada，2022

[170] Mustafaevich T N，Shakhboz R. System for analyzing and processing data on university staff based on a fuzzy controller with a fixed knowledge base. Open Access Repository，2022，8(3)：16-21

[171] Patel H R. Fuzzy-based metaheuristic algorithm for optimization of fuzzy controller：Fault-tolerant control application. International Journal of Intelligent Computing and Cybernetics，2022，15(4)：599-624

[172] Mao W，Liu J，Chen J，et al. An interpretable deep transfer learning-based remaining useful life prediction approach for bearings with selective degradation knowledge fusion. IEEE Transactions on Instrumentation and Measurement，2022，71：1-16

[173] Zelvelder A E，Westberg M，Främling K. Assessing explainability in reinforcement learning//Proceedings of the International Workshop on Explainable，Transparent Autonomous Agents and Multi-Agent Systems. Virtual，2021：223-240

[174] Druce J，Harradon M，Tittle J. Explainable artificial intelligence (XAI) for increasing user trust in deep reinforcement learning driven autonomous systems. arXiv preprint arXiv：2106.03775，2021

[175] Liu Z，Zhu Y，Wang Z，et al. MIXRTs：Toward interpretable multi-agent reinforcement learning via mixing recurrent soft decision trees. arXiv preprint arXiv：2209.07225，2022

[176] Liessner R，Dohmen J，Wiering M A. Explainable reinforcement learning for longitudinal control//Proceedings of the ICAART (2). Virtual，2021：874-881

[177] He L，Aouf N，Song B. Explainable deep reinforcement learning for UAV autonomous path planning. Aerospace Science and Technology，2021，118：107052

[178] Kumar S，Vishal M，Ravi V. Explainable reinforcement learning on financial stock trading using shap. arXiv preprint arXiv：2208.08790，2022

[179] Yu Y，Chung J，Yun H，et al. Fusing pre-trained language models with multimodal prompts through reinforcement learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada，2023：10845-10856

[180] Kumar A，Zhou A，Tucker G，et al. Conservative q-learning for offline reinforcement learning. Advances in Neural Information Processing Systems，2020，33：1179-1191

[181] Cao H，Wei Q，Zheng J，et al. Model-based offline adaptive policy optimization with episodic memory//Proceedings of the International Conference on Artificial Neural Networks. Bristol，UK，2022：50-62

[182] Prudencio R F，Maximo M R，Colombini E L. A survey on offline reinforcement learning：Taxonomy，review，and open problems. IEEE Transactions on Neural Networks and Learning Systems，2023

[183] Laskin M，Srinivas A，Abbeel P. Curl：Contrastive unsupervised representations for reinforcement learning//Proceedings of the International Conference on Machine Learning. Vienna，Austria，2020：5639-5650

[184] Laskin M，Yarats D，Liu H，et al. URLB：Unsupervised reinforcement learning benchmark. arXiv preprint arXiv：2110.15191，2021

[185] Park S，Ghosh D，Eysenbach B，et al. Offline goal-conditioned RL with latent states as actions//Proceedings of the ICML Workshop on New Frontiers in Learning，Control，and Dynamical Systems. Hawaii，USA，2023

[186] Reed S，Zolna K，Parisotto E，et al. A generalist agent. arXiv preprint arXiv：2205.06175，2022

[187] Wen Y, Wan Z, Zhou M, et al. On realization of intelligent decision-making in the real world: A foundation decision model perspective. arXiv preprint arXiv:2212.12669, 2022

[188] Driess D, Xia F, Sajjadi M S, et al. Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378, 2023

[189] Yun W J, Park J, Kim J. Quantum multi-agent meta reinforcement learning//Proceedings of the AAAI Conference on Artificial Intelligence.Washington, USA, 2023, 37: 11087-11095

[190] Wang J X, Kurth-Nelson Z, Kumaran D, et al. Prefrontal cortex as a meta-reinforcement learning system. Nature Neuroscience, 2018, 21(6): 860-868

[191] Beck J, Vuorio R, Liu E Z, et al. A survey of meta-reinforcement learning. arXiv preprint arXiv:2301.08028, 2023

[192] Mitchell E, Rafailov R, Peng X B, et al. Offline meta-reinforcement learning with advantage weighting//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 7780-7791

CAO Hong-Ye, Ph. D. candidate. His research interests cover reinforcement learning and explainable reinforcement learning.

LIU Xiao, Ph.D., lecturer. His research interests cover reinforcement learning and explainable reinforcement learning.

DONG Shao-Kang, Ph. D. candidate. His research interest is reinforcement learning.

YANG Shang-Dong, Ph. D., lecturer. His research interest is reinforcement learning.

HUO Jing, Ph. D., associate professor. Her research interest is machine learning.

LI Wen-Bin, Ph.D., associate researcher. His research interest is machine learning.

GAO Yang, Ph. D., professor. His research interest is reinforcement learning.

Background

Reinforcement Learning (RL) has undergone a significant evolution, progressing from early concepts of optimal control to revolutionary breakthroughs in deep RL. This fusion of deep neural networks with RL has enabled remarkable advancements in tackling complex tasks. Simultaneously, the rise of Explainable Reinforcement Learning (XRL) addresses the pressing need for transparent AI decision-making processes. XRL ensures that AI systems go beyond mere high performance, venturing into the realm of interpretability. This is especially crucial in critical domains like healthcare and autonomous systems, as it guarantees not only safety but also accountability and trustworthiness. XRL research is dedicated to developing algorithms capable of generating human-readable explanations for AI actions and constructing inherently interpretable models. Striking the delicate balance between performance and interpretability remains a challenge, particularly in intricate environments. Ultimately, RL's journey embraces the escalating demand for explainability, transforming AI from a mere performer into a collaborator armed with human comprehension.

This article provides a comprehensive review and synthesis of the current state of interpretability research in reinforcement learning. To commence, the article establishes a definition for the interpretability of reinforcement learning and outlines relevant evaluation methods. Subsequently, rooted in Markov decision processes, the article categorizes interpretability into four classes: action-level explanation, feature-level explanation, reward-level explanation, and policy-level explanation. Additionally, within each category, the article analyzes interpretive methods for both single-agent and multi-agent scenarios, with a special focus on the human element in interpretability research. It delves into human-machine interactive explanatory methods. Lastly, the article concludes by summarizing the current challenges in interpretability research for reinforcement learning and offering prospects for future research directions.

This work is supported in part by the Science and Technology Innovation 2030 New Generation Artificial Intelligence Major Project (2021ZD0113303), in part by the National Natural Science Foundation of China (62192783, 62276128, 62276142, 62206133), in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization and in part by State Key Laboratory of Novel Software Technology Project (KFKT2022B12).