

X-反馈强化学习：前沿进展与展望

刘起东^{1),2),3)} 何文轩¹⁾ 姚恩广¹⁾ 陈董^{1),2),3)} 李亚飞^{1),2),3)} 徐明亮^{1),2),3)}

¹⁾(郑州大学计算机与人工智能学院 郑州 450001)

²⁾(国家超级计算郑州中心 郑州 450001)

³⁾(智能集群系统教育部工程研究中心 郑州 450001)

摘要 人类反馈强化学习(Reinforcement Learning from Human Feedback, RLHF)整合了人类智慧与机器的力量。它通过人类培训师对人工智能系统的行为或输出给予的反馈评价或建议,完成奖励信号的创建或智能体策略的改变等。高质量的人类反馈能够显著提升人工智能系统对人类偏好和价值观的理解与适应能力,然而,高质量数据的稀缺性成为了 RLHF 进一步发展的瓶颈。近期, AI 反馈强化学习(Reinforcement Learning from AI Feedback, RLAIIF)的兴起为突破这一限制提供了新的视角,促使本文重新审视并定义了一个更广泛的框架 X-反馈强化学习(Reinforcement Learning from X-Feedback, RLXF)。RLXF 是一种结合了多种反馈源(包括人类和 AI)来指导强化学习过程的框架。这些反馈可以是直接的奖励信号、策略建议、偏好排序等多种形式,旨在优化智能体的行为策略,以更好地适应复杂多变的环境和满足多样化的目标。围绕 RLXF,从方法论创新到前沿应用进行系统性探讨:首先,建立 RLXF 的统一理论框架,阐明其通过多源反馈实现策略优化的核心机理;其次,将现有研究分为模仿学习、基于人类反馈的强化学习(RLHF)及基于 AI 反馈的强化学习(RLAIIF)等三种反馈范式。进而详细讨论了 RLXF 在自动驾驶、具身智能与大型语言模型(Large Language Models, LLMs)等关键领域的突破性应用实例。最后,总结了 RLXF 当前面临的主要挑战,并对其未来发展方向进行了展望。

关键词 人类反馈强化学习; AI 反馈强化学习; 模仿学习; 大型语言模型; 人机共融

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2026.00497

Reinforcement Learning from X-Feedback: Recent Advances and Future Prospects

LIU Qi-Dong^{1),2),3)} HE Wen-Xuan¹⁾ YAO En-Guang¹⁾ CHEN Dong^{1),2),3)} LI Ya-Fei^{1),2),3)} XU Ming-Liang^{1),2),3)}

¹⁾(School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001)

²⁾(National Supercomputing Center in Zhengzhou, Zhengzhou 450001)

³⁾(Engineering Research Center of Intelligent Swarm Systems, Ministry of Education, Zhengzhou 450001)

Abstract Reinforcement Learning from Human Feedback (RLHF) has established itself as a critical methodology for addressing the challenge of value alignment in complex artificial intelligence systems. Building upon decades of research in reinforcement learning and human-computer interaction, RLHF

收稿日期: 2025-04-24; 在线发布日期: 2025-11-04。本课题得到国家自然科学基金杰青项目(62325602)、国家自然科学基金面上项目(62276238)、国家自然科学基金区域联合重点支持项目(U24A20326)、国家自然科学基金重点项目(62036010)、河南省优秀青年科学基金(232300421095)以及河南省教育厅高校科技创新人才支持计划(25HASTIT034)资助。刘起东, 博士, 教授, 中国计算机学会(CCF)会员, 主要研究领域为集群智能、具身智能、大小模型协同。E-mail: ieqdliu@zzu.edu.cn。何文轩, 硕士, 主要研究领域为强化学习、模仿学习。姚恩广(通信作者), 博士研究生, 主要研究领域为强化学习、模仿学习。E-mail: egypt@gs.zzu.edu.cn。陈董, 博士, 讲师, 中国计算机学会(CCF)会员, 主要研究领域为集群智能、具身智能、大小模型协同。李亚飞, 博士, 教授, 中国计算机学会(CCF)会员, 主要研究领域为集群智能、具身智能、大小模型协同。徐明亮(通信作者), 博士, 教授, 中国计算机学会(CCF)会员, 主要研究领域为具身智能、虚拟现实、工业软件。E-mail: iexumingliang@zzu.edu.cn。

integrates human intelligence directly into the machine learning loop, moving beyond simplistic, hand-crafted reward functions that often fail to capture real-world complexity. It operates by leveraging evaluations, pairwise comparisons, and strategic suggestions from human trainers to shape the learning process, either through the construction of a proxy reward model or by directly influencing the policy updates of an AI agent. The effectiveness of this paradigm critically depends on the quality, diversity, and consistency of human-provided data: high-quality feedback dramatically enhances an AI system's ability to generalize, behave safely, and act in accordance with human intent. However, the practical scalability of RLHF is fundamentally limited by the inherent cost, scarcity, and subjectivity of large-scale human supervision, posing a significant barrier to applying it in more complex, dynamic domains. Recent advances in Reinforcement Learning from AI Feedback (RLAIF) offer a compelling solution to these limitations. RLAIF leverages large pre-trained AI models—most notably large language models (LLMs)—to generate synthetic preferences, critiques, and trajectory evaluations, automating and scaling the feedback process. While this substitution of human supervision with AI-generated feedback enables unprecedented scalability and consistency, it also raises new challenges in maintaining reliability, avoiding feedback loops, and ensuring that learned behaviors remain value-aligned. The coexistence of RLHF, RLAIF, and related paradigms such as imitation learning calls for a unified theoretical perspective to connect these diverse approaches within a coherent conceptual framework. To this end, this paper introduces and formalizes a general learning architecture termed Reinforcement Learning from X-Feedback (RLXF). RLXF represents a comprehensive framework that integrates heterogeneous feedback sources—human, AI, or hybrid—to guide reinforcement learning. Within RLXF, “X” denotes the feedback origin, encompassing scalar rewards, pairwise preferences, natural-language critiques, or full behavioral demonstrations. The framework establishes a principled foundation for integrating multiple feedback sources and elucidates how such feedback influences distinct learning components through mechanisms such as reward shaping, value shaping, policy shaping, and decision biasing. Beyond unification, RLXF also provides the meta-structure for studying adaptive feedback weighting, confidence estimation, and dynamic arbitration among feedback sources, enabling agents to balance competing signals across tasks and environments. Centered on the RLXF framework, this survey systematically traces the evolution of feedback-driven reinforcement learning, from methodological innovations to cutting-edge applications. We propose a taxonomy comprising three major paradigms: Imitation Learning (learning from expert demonstrations), RLHF (learning from human preferences and critiques), and RLAIF (learning from AI-generated or model-assisted feedback). Each paradigm is analyzed in terms of its algorithmic foundation, strengths and weaknesses, and characteristic data requirements. Furthermore, the survey examines the transformative applications of RLXF across several domains: in autonomous driving, where feedback mechanisms support ethical decision-making and risk-aware control; in embodied intelligence, where linguistic or evaluative feedback enables adaptive skill acquisition; and in large language models, where RLXF has proven vital in aligning outputs to be helpful, harmless, and honest. Finally, we outline the emerging challenges and future prospects for RLXF. These include mitigating multi-source feedback conflicts, detecting biases within AI-generated supervision, improving the sample efficiency of reward modeling, and exploring recursive self-improvement loops between human and AI critics. Looking ahead, RLXF is poised to become a cornerstone framework for building trustworthy, interpretable, and scalable learning systems, bridging the gap between human-centric intelligence and autonomous machine adaptation.

Key words reinforcement learning from human feedback; reinforcement learning from AI feedback; imitation learning; large language model; human-machine integration

1 引言

随着人工智能(Artificial Intelligence, AI)模型快速迭代演进, 其参数量已从亿级跃升至万亿级。通过构建大模型以提升 AI 处理性能、增强 AI 通用性、加速 AI 广泛应用已成为各界共识。然而, AI 缺乏自身的价值观, 无法自行判断何为正确的行为, 因此 AI 伦理与安全问题正成为 AI 商业化发展的瓶颈。人类反馈强化学习(Reinforcement Learning from Human Feedback, RLHF)是一种结合强化学习(Reinforcement Learning, RL)和人类反馈的调优方法, 通过多种方式整合人类反馈以指导和优化 RL 模型, 引导模型的行为更加符合人类的期望和价值观。这种反馈机制不仅能增强模型与人类目标的

一致性, 还有效解决了强化学习中常见的奖励稀疏问题。然而, 获取大量的人类反馈需要耗费大量的时间和资源, 且人类反馈往往带有个人偏见, 可能导致训练出的模型性能不稳定。因此, 使用 AI 反馈强化学习(Reinforcement Learning from AI Feedback, RLAIIF)取代 RLHF 的方案引起广泛关注(如图 1 所示), 随着相关研究的不断积累和发展, 采用 AI 反馈的方法在克服传统基于人类反馈的局限方面展现了显著潜力。基于这些观察与分析, 本文定义了一个更广泛的框架—X-反馈强化学习(Reinforcement Learning from X Feedback, RLXF)。RLXF 是指一种结合了多种反馈源(包括人类和 AI)来指导强化学习过程的框架, 其中人类反馈强化学习和 AI 反馈强化学习可以被视为是 RLXF 的不同范式。

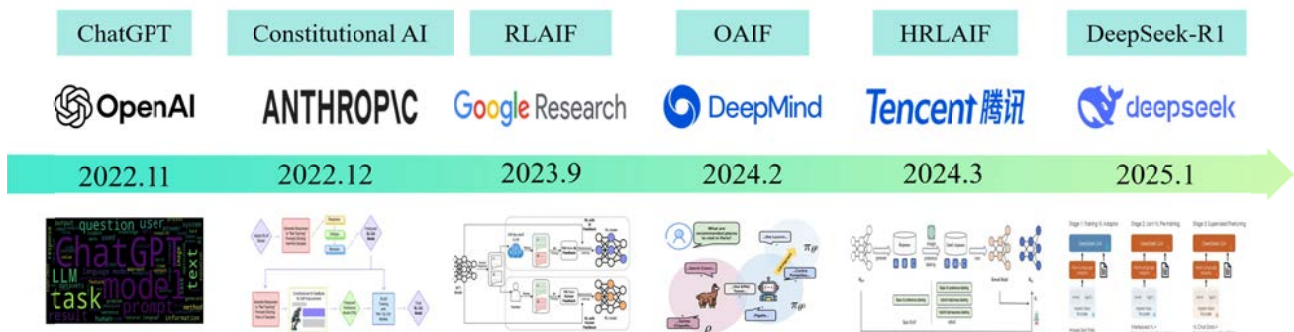


图 1 AI反馈强化学习发展历程

在强化学习的框架下, 智能体通过与环境不断交互来寻求最优的决策策略(即动作选择)。最优决策的判断标准完全依赖于由系统设计师根据智能体性能指标定义的奖励信号, 旨在提供必要的反馈以指导智能体学习正确的行为模式^[1-2]。然而, 在许多实际应用中, 成功的标准难以精确界定和量化, 这也增加了奖励函数设计的复杂性。此外, 稀疏的奖励信号往往不足以支持有效的学习, 这促使研究者采用奖励塑造(Reward Shaping)技术, 将原始的奖励信号转化为更利于学习的形式^[3]。然而, 奖励塑造可能引入伪相关(Spurious Correlations)问题, 即智能体可能因执行与真实目标相关联但本身并无价值的行为而获得奖励。这种现象最终可能导致奖励黑客攻击(Reward Hacking)^[4], 即智能体利用奖励机制中的漏洞以非预期的方式获得高奖励, 从而偏离期望的学习目标。

过人类迭代地定义和精炼学习目标, 来指导智能体的学习过程。这种方法不仅有望解决传统 RL 方法中的奖励函数难以设定的问题, 而且有助于实现智能体对齐(Agent Alignment), 促进智能体的学习目标与人类价值观保持一致, 进而推动构建道德健全和社会负责任的人工智能系统。

随着 RLHF 技术的不断进步与成熟, 它已经在多个领域展现出广泛而深入的应用价值。从较为简单的图像生成技术^[5]到复杂的连续控制^[1], 再到涉及高级交互的游戏开发^[6]和机器人技术^[7], 乃至大型语言模型(Large Language Models, LLMs)的微调^[8]等方面, RLHF 均展现出了显著的成效。此外, RLHF 在方法论上也取得了长足的进步, 例如融合多种反馈类型以利用它们的相对优势^[6,9-10]; 采用主动学习和主动查询合成提高查询效率^[11-13]; 借鉴心理学洞见以提高人类反馈的质量; 运用元学习等技术, 利用先前数据将已学习的偏好快速适应新任务^[7,14,16]; 以及借助数据增强和半监督学习等方法更有效地利用可用偏好数据^[17-20]。随着大模型的广泛应用,

为了克服这些挑战, RLHF 作为一种切实可行的解决方案应运而生。RLHF 将人在环(Human-In-The-Loop)组件引入到标准的 RL 学习范式中, 通

传统RLHF方法对高质量人工标注数据的强依赖成为瓶颈,从AI反馈中进行强化学习^[21-26]的概念应运而生,为本领域的发展提供了新的见解。

尽管RLHF和RLAIF已取得显著进展,但针对融合多源反馈的强化学习框架(RLXF)的系统性中文综述仍属空白。因此,本文旨在填补这一空白,为相关领域的研究人员提供理论框架与方法论的全面参考。本文的主要内容如下:首先梳理RLXF技术的发展历程并回顾相关理论基础;然后介绍该领域的国际研究进展,包括国内外最新研究成果及其实际应用案例;最后总结当前研究不足并探讨未来可能的发展方向,为后续研究提供参考。

2 数学概念

强化学习^[24,27-30]是一种智能体与环境交互并根据奖励反馈不断优化自身行为的机制。其核心在于将学习环境形式化为一个在不确定环境中做出序列决策的数学框架——马尔可夫决策过程(Markov Decision Processes, MDP)。在马尔可夫决策过程中,智能体首先观察其当前状态 $s_t \in S$,并根据策略网络 π_θ 从动作空间 A 中采取动作 a_t ,随后由状态转移函数 P 生成新的状态 s_{t+1} 。每次动作执行后,智能体会根据奖励函数 R 获得即时反馈,以评估动作的

效果。此外MDP还包含初始状态分布 $d_0 \in \Delta(s)$ 和折扣因子 $\gamma \in [0,1]$,用于计算长期累积奖励,以指导智能体实现最大化累积回报。表1汇总了本文涉及的核心数学符号及其定义。

令 π_e 表示专家策略,RLXF的目标可以表述为一个策略优化问题:通过最小化智能体策略 π_θ 与专家策略 π_e 之间的行为差异,使得智能体能够有效地模仿专家的决策模式。即RLXF希望智能体学习到一个与专家行为类似策略,具体为

$$\pi^* = \arg \min_{\pi_\theta} \mathbb{E}_{s_t \sim P(s_{t-1}, a_{t-1}), \hat{a}_t \sim \pi_e(s_t)} [L(\hat{a}_t, \pi_\theta(s_t))] \quad (1)$$

这里, $L(\hat{a}_t, \pi_\theta(s_t))$ 是衡量智能体策略和专家策略之间差异的损失函数, \hat{a}_t 表示 t 时刻的专家动作, $\pi_\theta(s_t)$ 表示智能体状态 s_t 下策略 π_θ 执行的动作。

综合上述内容,根据指导对象和指导方式的不同,本文将X-反馈强化学习的方法分为三大范畴:一是模仿学习,该类方法侧重于直接利用人类专家的经验知识作为学习蓝本;二是人类反馈强化学习,进一步细分为基于人类偏好的强化学习和基于人类建议的强化学习,旨在通过动态的人机交互来促进学习进程;三是AI反馈强化学习,利用AI作为反馈源来指导学习进程,改变了人在环路的反馈机制。具体内容如图2所示。

表1 符号表

符号	含义	符号	含义	符号	含义
S	状态空间	π^*	最优策略	r^*	最优奖励函数
A	动作空间	π_e	专家策略	$U(\tau)$	效用函数
s_t	t 时刻状态	D^e	专家示范数据	$H(\pi)$	因果熵
a_t	t 时刻动作	τ^e	专家轨迹	$D(s, a)$	判别器
Q	价值函数	L	损失函数	f	状态特征嵌入
w	策略权重向量	P	状态转移函数	$q(\tau)$	生成数据分布
π_θ	智能体策略	R/r	奖励函数	$u(f)$	特征嵌入偏好



图2 X-反馈强化学习

3 模仿学习

人类专家对智能体执行任务的经验, 对于显著提升探索效率并大幅降低试错成本具有至关重要的作用。鉴于当前 RL 领域普遍面临的采样效率低下的挑战, 利用人类专家经验成为了一条极具潜力和价值的研究路径^[1]。模仿学习(Imitation Learning, IL)^[31]是强化学习与监督学习的结合, 旨在通过观察和模仿人类专家的行为来训练智能体, 为缓解低样本效率问题提供了解决方案。

传统模仿学习涵盖了行为克隆^[32], 逆强化学习^[33]以及对抗式模仿学习^[34]等范式。近年来, 得益于计算能力的飞跃式提升, 模仿学习领域迎来了新的突破, 主要包括基于观察量的模仿学习^[35]和跨领域模仿学习^[36-38]等。本章节后续内容将深入探讨上述提及的各种模仿学习范式的核心思想。

3.1 行为克隆

行为克隆(Behavioral Cloning, BC)是一种基于监督学习的模仿学习方法。该方法将专家示范数据 $D^e = \{(s_i, a_i)\}_{i=1}^M$ 视为训练数据, 其中每个状态 s_i 被赋予相应的动作 a_i 作为标签, 如图 3 所示。智能体利用这些数据, 通过最小化策略输出与专家动作之间的差异(如交叉熵损失或均方误差)来学习专家的策略 π_e ^[39]。令 π_θ 表示智能体的策略, 其中 θ 为该策略的参数, 行为克隆旨在通过调整参数 θ , 使智能体的策略尽可能接近专家的策略, 即

$$\theta^* = \arg \min_{\theta} \sum_{(s_i, a_i) \sim D^e} \|\pi_{\theta}(s_i) - a_i\|_2^2 \quad (2)$$

其中, $\pi_{\theta}(s_i)$ 表示智能体在状态 s_i 下按照策略 π_{θ} 执行的动作。

行为克隆领域的重要突破始于 Dagger(Dataset Aggregation)算法^[40], 该算法通过迭代式在线学习机制将智能体生成数据与专家标注相结合, 从而最小化策略 π_t 与专家策略间的行为差异, 其理论表明当迭代次数 $T \rightarrow \infty$ 时策略性能渐近收敛。AggreVaTe(Aggregate Values to Imitate)算法^[41]在 Dagger 的基础上进行了关键改进, 将优化目标从动作匹配转向价值函数优化, 通过最小化专家成本函数(Q 函数)实现策略提升, 这一转变使算法不仅能模仿行为更能学习决策原理。

行为克隆技术虽然有效, 但其对有限专家数据的依赖带来了诸多挑战。首先, 行为克隆的泛化能

力受限, 这是有监督学习领域的普遍问题。具体而言, 智能体在训练数据之外的样本上表现显著下滑, 尤其是在面对分布偏移时, 其性能往往大幅下降。这是因为专家数据集仅包含有限的示范实例, 难以覆盖所有可能的场景。其次, 行为克隆学习的策略难以保证每一步都达到最优。一旦在某个环节出现微小偏差, 智能体就可能由于进入专家数据集中未曾记录的状态而采取随机动作, 进而引发后续状态的进一步偏离, 形成误差的累积与放大。此外, 该方法还面临因果混淆问题: 端到端的学习机制使智能体难以从专家演示中准确区分真实的因果关系。以自动驾驶为例, 专家行为虽常表现为红灯时停车, 但并非所有红色信号都应触发相同动作。若智能体缺乏对专家策略因果结构的理解, 可能过度泛化红色信号与停车之间的关联, 导致在复杂场景中出现误判。

为克服这些局限性, Decision Transformer^[42]及其后续工作提出了一系列创新方法。Decision Transformer 通过将强化学习问题转化为条件序列建模任务, 利用回报条件化机制显式优化任务目标, 从而缓解了行为克隆在分布偏移下的泛化问题。广义决策 Transformer(GDT)^[43]进一步引入后见信息匹配框架, 显著提升了离线多任务模仿学习中的表现。此外, Ma 等人^[44]通过分层强化学习重新设计 Decision Transformer 的架构优化了复杂任务中的策略生成, 并提升了模型的泛化能力。这些工作不仅解决了行为克隆的误差累积与因果混淆问题, 还推动了模仿学习从单一任务、静态数据向多任务、动态目标的范式转变, 为复杂场景下的策略学习提供了新的解决方案。

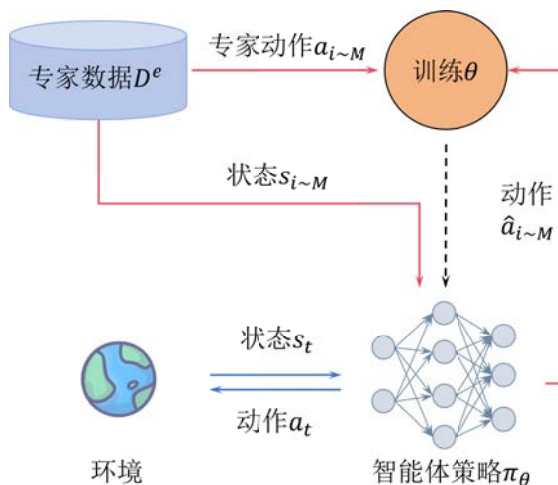


图 3 行为克隆示意图

3.2 逆强化学习

逆强化学习 (Inverse Reinforcement Learning, IRL)^[45-46] 也称为逆向最优控制 (Inverse Optimal Control, IOC), 如图 4 所示。不同于直接模仿专家行为, 逆强化学习旨在通过专家轨迹数据 $D^e = \{\tau_1^e, \tau_2^e, \dots, \tau_N^e\}$ 来推断奖励函数 $r(s, a)$, 进而深入理解专家行为背后的动机和原理^[33]。该方法基于一个核心假设: 专家策略 $\pi_e(a|s)$ 为最优策略 $\pi^*(a|s)$, 专家样本由该策略与环境交互产生。该最优策略的平均

回报被定义为: $E_{\pi_e}[r(s, a)] = E\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$, 其中

$a \sim \pi(\cdot|s_t)$, $s_{t+1} \sim P(\cdot|s_t, a_t)$ 。逆强化学习方法的首要目标是寻找一个奖励函数 r^* , 该函数需确保专家策略 π_e 相对于其他任意策略 π_θ 而言, 能够产生更高的累积奖励 (或更低的累积代价), 即

$$r^* = IRL(\pi_e) = \arg \max_r (\max_{\pi_\theta} H(\pi_\theta) + E_{\pi_\theta}[r(s, a)] + E_{\pi_e}[r(s, a)]) \quad (3)$$

其中, $H(\pi_\theta) = \mathbb{E}_{\pi_\theta}[-\log \pi_\theta(s, a)]$ 表示策略 π_θ 的因果熵, $\mathbb{E}_{\pi_\theta}[r(s, a)]$ 表示采取当前策略的平均回报, $\mathbb{E}_{\pi_e}[r(s, a)]$ 表示专家策略回报。之后, 通过使用奖励函数 r^* 进行正向强化学习, 尝试恢复专家策略 π_e 。

$$\pi_e = RL(r^*) = \arg \max_{\pi_\theta} H(\pi_\theta) + \mathbb{E}_{\pi_\theta}[r(s, a)] \quad (4)$$

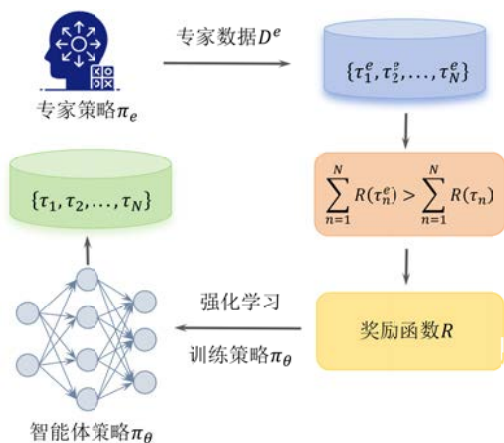


图4 逆强化学习示意图

逆强化学习的奖励函数最初由 Ng 等人^[33]于 2000 年提出, 随后发展出基于学徒学习^[47]和最大边际算法^[48]等方法, 假设奖励是状态或状态-动作对特征的线性组合。然而, 由于专家示范数据覆盖有限, 导致奖励函数存在歧义性问题。为解决这一问题, Ziebart 等人^[49]提出了最大熵逆向强化学习 (Maximum Entropy Inverse Reinforcement Learning,

MaxEnt IRL), 通过概率推理推断奖励函数。2016 年, Ho 等人^[34]将生成对抗网络 (Generative Adversarial Network, GAN) 引入模仿学习, 提出了生成式对抗模仿学习 (Generative adversarial imitation learning, GAIL)。

如图 5 所示, GAIL 的主要思想是智能体基于随机初始化策略与环境交互生成轨迹数据, 同时通过对抗训练判别器网络区分智能体轨迹与专家示范数据。基于此, 奖励函数被定义为 $r(s, a) = -\ln(1 - D(s, a))$, 其中 $D(s, a)$ 表示判别器将状态-动作对判定为来自专家策略的概率。这个奖励函数鼓励智能体生成那些判别器难以区分的轨迹。最后利用 RL 算法对智能体进行训练, 以最大化这个奖励函数, 从而“欺骗”判别器^[50]。GAIL 算法的最终目标函数为

$$\mathbb{E}_{\pi_e}[\log D_\epsilon(s, a)] + \mathbb{E}_{\pi_\theta}[\log(1 - D_\epsilon(s, a))] - \lambda H(\pi_\theta) \quad (5)$$

其中, $\mathbb{E}_{\pi_e}[\log D_\epsilon(s, a)]$ 这一项表示的是从专家策略 π_e 生成的状态-动作对 (s, a) 被判别器 D_ϵ 正确识别为专家数据的概率的期望值。 $\mathbb{E}_{\pi_\theta}[\log(1 - D_\epsilon(s, a))]$ 表示的是从当前智能体策略 π_θ 生成的状态-动作对 (s, a) 被判别器 D_ϵ 正确识别为智能体生成数据的概率的期望值。 $H(\pi_\theta)$ 表示智能体策略 π_θ 的熵, λ 是一个正则化参数。与 GAN 中的训练方法相同, GAIL 的训练过程通过交替优化实现鞍点求解。判别器的更新是在固定策略 π_θ 的条件下, 使用梯度下降 (如 Adam) 优化 D_ϵ , 最小化其对抗损失, 以提高区分真假轨迹的能力。策略的更新则是在固定判别器的条件下, 将 $\log(1 - D_\epsilon(s, a))$ 作为伪奖励信号, 通过强化学习算法 (如 TRPO 或 PPO) 更新策略参数 θ , 提升策略生成行为与专家轨迹的一致性。该过程在策略优化与判别器更新之间迭代进行, 直至

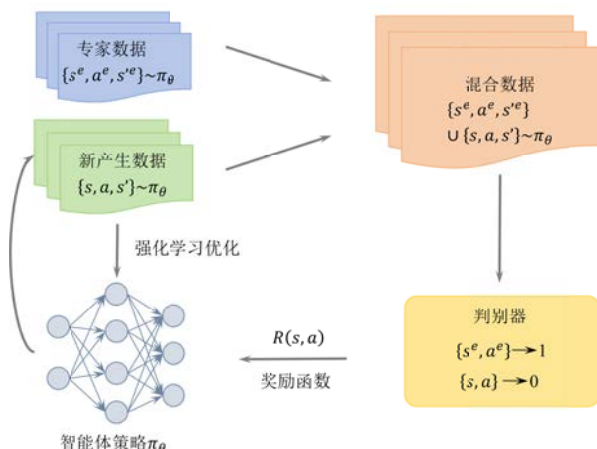


图5 生成对抗式模仿学习示意图

收敛。整体目标是通过判别器提供的反馈信号引导策略逼近专家分布, 实现高质量的模仿学习。虽然 GAIL 在高维状态空间表现出优势, 但仍存在模式崩塌等固有缺陷。

同年 Finn 等人^[51]把对抗式和逆强化学习的思想相结合提出了对抗式逆强化学习的框架 GAN-GCL (Generalized adversarial network guided cost learning) 算法, 该算法提出了一种特殊形式的判别器, 它利用生成器的密度信息来估计输入样本是来自专家数据还是生成数据。该方法使用了与 GAIL 类似的对抗训练框架, 但 GAIL 的原始设计未对判别器进行特定架构约束, 导致其无法像 GAN-GCL 方法那样可以重构奖励函数。

此外, 随着 GAIL 的发展, 一些新的方法也被提出, 例如 Wang 等人^[52]提出了一种新的 DiffAIL 方法, 将扩散模型结合到 GAIL 中以提高判别器的分布捕捉能力。Liu 等人^[53]提出一种结合信号时序逻辑推理和控制合成的可解释模仿学习方法。Chang 等人^[54]通过 Boosting 框架结合对抗模仿学习, 提出一种新的对抗模仿学习算法。Sharma 等人^[55]通过对抗训练实现不同领域间的数据分布对齐。

3.3 基于观察量的模仿学习

基于观察的模仿学习 (Imitation Learning from Observation, ILfO) 作为行为克隆的重要扩展, 通过仅利用专家状态序列 (包括离散轨迹 $\tau_e = \{s_i\}_{i=1}^M$ 或连续感知数据) 克服对动作信息的依赖, 提升方法的实用性。该方法体系可分为两类核心范式: 有模型方法通过建立环境动态模型实现模仿学习, 其中正向动态模型^[56-57]学习潜在策略 $\pi(z|s_t)$ 预测状态演化, 逆向动态模型^[58-60]基于确定性动作假设从状态转移重建动作分布; 无模型方法则绕过显式建模, 包括对抗式学习 (见第 3.2 节) 和奖励函数工程法^[61-62]。这些方法各具特色: 正向模型注重状态预测, 逆向模型专注动作推断, 而奖励工程直接设计模仿目标。尽管 ILfO 摆脱了动作监督的约束, 但仍面临状态覆盖不足和动态建模误差等固有挑战, 这为未来研究指明了改进方向。

3.4 跨领域模仿学习

跨领域模仿学习 (Cross-Domain Imitation Learning, CDIL) 突破了传统模仿学习对同源领域数据的依赖, 通过构建跨域映射机制实现知识迁移。其核心挑战在于解决状态转移差异 (环境动力学不匹配)、形态学差异 (状态/动作空间异构) 和视角差异

(多源数据的表征差异) 三大跨域鸿沟。现有方法可分为四大技术路线: (1) 直接法^[63]通过线性映射建立状态对应关系, 但受限于表达能力; (2) 映射法^[61,64]利用深度神经网络学习非线性时空对齐, 但对数据要求严苛; (3) 对抗式方法^[55]通过生成对抗网络实现分布对齐, 能处理非配对数据; (4) 最优传输法^[36-38,65]基于 Gromov-Wasserstein 距离优化占用测度匹配, 提供理论严密的跨域传输方案。这些方法各具优势: 直接法简单高效但泛化性有限, 映射法表达能力强却依赖对齐数据, 对抗方法数据利用率高但训练不稳定, 最优传输理论完备但计算复杂度高。当前 CDIL 研究正朝着多模态融合、元学习增强和理论-算法协同设计等方向发展, 以应对现实场景中复杂的跨域模仿需求。

3.5 小结

模仿学习虽然是一种强大的方法, 但在实际应用中还面临以下挑战: (1) 对于复杂任务, 收集大量高质量的人类专家演示数据耗时且成本高昂, 且这些数据可能受限于专家的行为多样性和知识范围; (2) 模仿学习可能导致智能体在策略探索上受限, 特别是当专家未展示某些有效但不常见的行为时, 智能体难以独立发现这些策略; (3) 人类专家的行为并非总是最优或无误, 因此智能体在模仿过程中需设计机制来识别和纠正偏差, 以提升模型性能; (4) 模仿学习策略难以直接泛化到新情境, 需借助迁移学习进行适应优化。然而, 迁移过程中可能存在偏见传播的风险, 因为源域的数据可能包含特定群体的偏见。因此, 这不仅涉及技术挑战, 还引发了公平性和透明度的问题: 模型是否能在不同用户群体间保持公正, 避免歧视性结果; (5) 模仿学习没有明确的奖励信号, 它仅仅依赖于模仿专家的行为来学习策略。当遇到需要自主决策的情境时, 模仿学习可能无从下手, 特别是在奖励信号稀疏的复杂任务中表现不佳; (6) 在许多实际问题中, 状态空间非常庞大和复杂, 这增加了模仿学习的困难。高维状态空间要求智能体能够从观察中提取关键特征, 并将其映射到有效的动作策略。解决这个问题需要合适的特征提取方法和状态表示技术。

4 人类反馈强化学习

人类反馈强化学习 (Reinforcement Learning from Human Feedback, RLHF) 是一种结合强化学

习和人类反馈的机器学习范式,通过引入人在环路的机制^[66],将人类反馈通过多种人类知识集成方法传递到底层 RL 算法的特定元素中(例如动作、状态和奖励等),在设计维度上实现了人类智慧与算法学习的深度融合。该方法的核心在于智能体与环境动态交互的过程中,人类专家负责观察并评估智能体的行为,实时地向智能体提供反馈或指导,以优化智能体的策略。人类反馈强化学习已被证明相较于 RL 算法更为有效^[67],这种人机交互的机制使智能体能够更有效地学习和适应复杂的任务环境,尤其是在奖励信号稀疏或难以定义的情况下,通过结合强化学习的自动优化能力和人类的高层次判断,人类反馈强化学习能够加速学习过程并提高决策质量^[68]。然而,目前针对如何设计有效的易泛化人类反馈强化学习方法,并且可以适应不同人类专家的反馈仍然是个挑战。根据人类反馈的形式,本文将人类反馈强化学习方法分为两种,基于人类偏好的强化学习^[69]和基于人类建议的强化学习^[70]。

4.1 基于人类偏好的强化学习

基于人类偏好的强化学习(Preference based Reinforcement Learning, PbRL)^[69]是一种从非数值反馈中进行序列决策学习的范式,旨在通过引入基于人类偏好的反馈机制,利用相对效用评估替代绝对奖励值,克服传统强化学习对精确数值奖励的依赖。这里,人类的偏好可以定义为一种反馈,不受任意奖励选择、奖励塑造、奖励设计或预定义目标权衡的影响。这种偏好反馈不仅减少了奖励设计的主观性和复杂性,还增强了算法对不同任务及非专家用户的适应性。基于人类偏好的强化学习算法通过解析状态、动作或轨迹之间的偏好关系来解决强化学习问题,其目标在于学习一个与专家偏好高度一致的策略。为此,PbRL 算法需精准捕捉并理解专家的意图,进而搜索与之最为契合的策略。在人机共融的框架下,人类专家通过对随机选取的轨迹段进行“良好”、“坏”或“相等”的评估反馈,将偏好信息直接融入强化学习过程^[1,19]。智能体则利用这些轨迹及其对应的偏好反馈,训练出能够准确预测奖励的模型,从而在无需依赖繁琐的奖励函数设计的情况下,即可实现高效学习。

由于人类反馈强化学习注重针对不同底层 RL 特定元素的反馈,因此,根据人类偏好反馈的对象类型不同,结合现有文献,一般可分动作、状态和轨迹三种不同的偏好反馈类型。

4.1.1 动作偏好

动作偏好一般通过学习人类专家对在同一状态 s 下的不同动作 $a^i (a^i \in A(s))$ 的偏好来优化智能体的策略。其优化目标可分为两类:短期最优性关注即时奖励,但易陷入局部最优;长期最优性则基于累积回报进行决策,更能实现全局优化^[67]。在动作偏好学习中,人类专家虽需洞悉长期目标的期望,却无需深入探究最优策略的具体细节。例如 Fiihrnkranz 等人^[71]的工作便巧妙运用了长期动作偏好,通过计算长期期望的策略,为无需预知完整策略的最优动作选择提供了有效路径。该方法的核心是将动作偏好建模为状态相关的二分类问题:给定状态 s 下,通过训练数据学习预测模型 $M(a^i > a^j | s)$ 来判断动作 a^i 是否优于 a^j 。其中,训练数据可表示为 $\{s, a^i\} > \{s, a^j\} \Leftrightarrow (a^i > a^j | s), \forall a^i, a^j \in A(s)$, $A(s)$ 表示状态 s 下的动作集合。为获得完整动作排序,可通过投票机制整合两两比较结果^[71],每个动作 a 的得分 $C(s, a)$ 表示其在所有比较中胜出的次数,从而确定其相对优劣。

尽管选择每个状态下最受偏好的动作相对简单,但想要泛化到不同的情境中仍是一大挑战。

4.1.2 状态偏好

状态偏好学习通过比较不同状态间的相对价值来指导策略优化,其核心在于建立状态间的优先关系:若状态 s_1 下可选取的某个动作优于状态 s_2 中可用的所有动作,则认为状态 s_1 优于 s_2 。相较于动作偏好,状态偏好需权衡长期与短期优化目标:长期偏好^[72]追求累积回报最大化,致力于识别具有最优后续发展的状态;短期偏好^[73]则聚焦即时状态价值评估,但难以平衡频繁访问的次优状态与稀疏访问的帕累托最优状态之间的取舍。这一时序维度的权衡问题本质上是强化学习中探索与利用矛盾的体现,构成了状态偏好学习方法的核心挑战。当前研究重点在于开发能有效协调长短期偏好的新型优化框架,以提升策略在复杂环境中的决策质量。例如,Zucker 等人^[73]通过人类专家评判当前的状态偏好,协助智能体优化自己的策略。基于状态偏好的学习目标是对于当前状态的特征嵌入 f ,如果专家更喜欢 f^+ 而不是 f^- ,那么使得首选状态在波动幅度 $m > 0$ 的条件下获得的偏好值更高。

状态偏好简化了专家决策的需求,它无需专家直接比较各状态下的具体动作,但仍需评估各状态下策略的未来成效。尽管状态偏好不能直接转化为

策略, 但结合偏好转换模型, 仍有高效推导出相应策略的潜力。

4.1.3 轨迹偏好

轨迹偏好是反馈的最通用且最广泛使用的形式^[17], 它简化了专家评估的复杂性, 允许直接根据完整轨迹的成效进行评判, 一般指定哪个轨迹(状态动作序列)优于另一个轨迹。这种方法不仅考量行为效率(如目标达成路径的平滑度), 还兼顾了已知优质状态的出现频率, 旨在最大限度地减轻专家的认知负担。例如, Busa 等人^[74]通过专家评估, 对比两个策略 π_1 和 π_2 生成的不同轨迹对 τ^{π_1} 和 τ^{π_2} 。如果策略 π_1 生成的轨迹 τ^{π_1} 在期望上优于策略 π_2 实现的轨迹 τ^{π_2} , 则策略 π_1 优于策略 π_2 。

还有一些工作^[75]通过分解轨迹效用为状态-行为效用 $U(s, a)$ 之和, 即 $U(\tau) = \sum_t U(s_t, a_t)$ 来估算

专家评估标准下的轨迹价值。这一机制与经典强化学习中的预期回报类似, 此方法不仅拓展了策略空间的探索, 还深入探索了偏好强化学习中的效用空间, 使得无需频繁依赖专家直接反馈即可评估新轨迹, 并通过效用函数的迭代更新来优化策略。然而, 面对不同起始状态的轨迹, 如何精确识别与偏好相关的关键状态或动作, 成为轨迹偏好应用中的一大挑战。目前, 尚无现成算法能完美处理此类跨初始状态轨迹的偏好判定。

从另一个角度来说, 从偏好中学习是一个涉及智能体与人类专家两种参与者的过程, 该过程构建于一系列轨迹之上, 轨迹可预先由给定策略定义, 也可通过采样生成。人类专家评估一个或多个轨迹对, 并表示偏好。如图 6 所示, 根据学习基于偏好的策略 π 的方式不同, 一般可以分为直接学习策略、学习偏好模型和学习效用函数三种不同的方式。

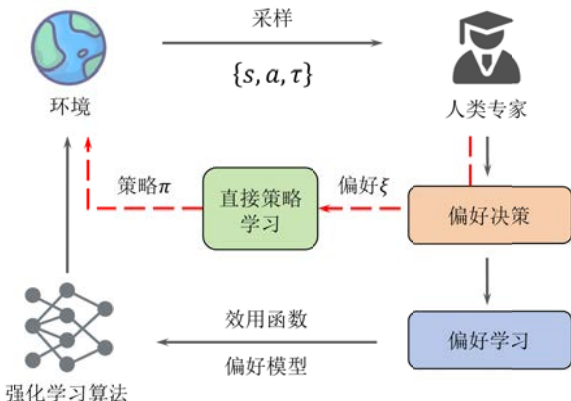


图 6 基于人类偏好的强化学习

近年来, 一种创新的离线优化方法在 RLHF 领域崭露头角, 即直接偏好优化(Direct Preference Optimization, DPO)^[76]。DPO 方法的重要突破在于它证明了在特定偏好模型假设下(例如 Bradley-Terry 模型), 当面临 KL 散度约束下的奖励最大化问题时, 其最优解可以通过定义在一个偏好数据集上的简化损失函数以解析形式求得。这种方法的优势在于无需构造具体的轨迹奖励函数, 同时避免了繁琐的在线策略梯度迭代过程。因此, DPO 将策略优化问题转换为一个可以在静态的、基于偏好的轨迹数据集上执行的稳定离线学习任务在此基础上, 后续研究进一步扩展了这一框架的应用范围与鲁棒性。例如, 身份偏好优化(Identity Preference Optimisation, IPO)^[77]针对有限或含噪声数据的情况, 提出了增强型的优化策略, 从而提升了该方法在复杂环境下的适应能力与效能。

尽管基于人类偏好的强化学习方法在特定领域取得了显著进展, 但智能体行为与复杂且多样的人类偏好之间的对齐仍面临挑战。为解决这一问题, AlignDiff 框架^[78]结合 RLHF 和扩散模型, 通过量化并统一多样化的人类偏好, 实现零样本行为定制。该框架利用多视角人类反馈数据集训练属性条件扩散模型, 灵活应对复杂用户需求。实验表明, 此方法增强了智能体的行为适应性, 并在缺乏明确指令时使智能体能够做出符合人类期望的决策。

此外, 在实际应用中, 基于偏好的强化学习方法面临的一个关键问题是, 偏好信号的信息量通常少于直接的奖励信号, 这增加了学习过程的复杂性和挑战。Wang 等人^[79]从理论上探讨了这一问题, 证明即使偏好信号的信息量较少, RLHF 问题依然可以通过现有的强化学习算法和技术有效解决, 且成本较小。他们的研究表明, 现有的强化学习技术可以有效地支持基于偏好信号的学习过程, 从而减轻了由于偏好信息不足带来的复杂性挑战。

4.2 基于人类建议的强化学习

基于人类建议的强化学习(Reinforcement Learning with Human Advice)是指智能体利用人类专家的指导来改善其学习过程和性能。如图 7 所示, 根据建议的提供方式与内容, 人类建议可被系统性地分为两大类: 第一类是通用建议(General Advice), 它不依赖于具体任务状态, 通常在学习开始前以规则或约束的形式离线提供; 第二类是情境建议(Contextual Advice), 其含义与当前状态紧密

耦合,需要在任务执行过程中以评价性反馈或实时指令等形式交互式地给出。这些不同形式的建议能够被融入奖励函数、价值函数、策略或决策等不同学习阶段^[1],其处理方式可分为基于模型的方法(通过教师模型处理建议)和无模型方法(直接使用建议)。相比传统强化学习仅依赖环境奖励,该方法通过引入人类先验知识来提高学习效率并引导探索,其具体实现方式则取决于所采用的强化学习范式与建议整合策略。

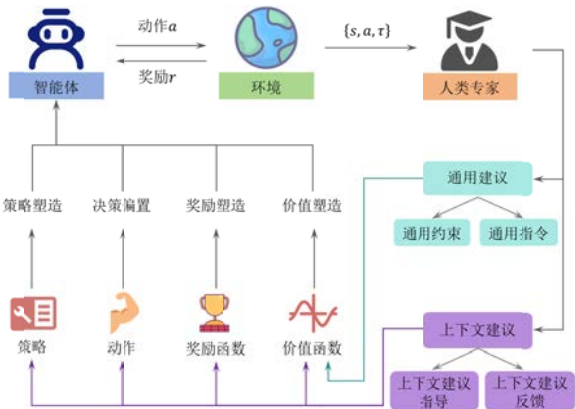


图7 基于人类建议的强化学习

4.2.1 奖励塑造

奖励塑造旨在通过将人类专家提供的反馈直接作用于奖励函数上来加速智能体的学习过程。提供反馈奖励的一种方式是使用评估性反馈^[3]。评估性反馈被视为与RL中环境提供的反馈相同,将评估性反馈转换为数值后,它可以被视为延迟奖励,就像MDP奖励一样,并且可以使用标准RL算法来计算价值函数。将评估性反馈转换为数值后,与预定义的奖励函数相加形成新的奖励函数,即 $R' = R + R_h$,其中 R 是环境奖励, R_h 是人类反馈奖励。例如,在TAMER框架中^[80],评估性反馈被转换为奖励,并用于计算回归模型,称为“人类强化函数”(Human Reinforcement Function)。这个模型预测人类对每个状态-动作对的奖励预期,并引入衰减权重因子,将人类反馈与预定义的奖励函数相结合。

4.2.2 价值塑造

在研究奖励塑造时,一些研究者指出了即时奖励和延迟奖励之间的基本差异。他们认为评估性反馈与标准的MDP奖励不同,实质上是对一个动作价值的即时信息。Torrey等人^[81]提出的KBKR方法,通过内核回归模型将if-then规则融入,进而将这些规则作为约束条件整合至值函数中。Maclin等人^[82]利用支持向量回归技术优化值函数,其中,动作值

的约束(如“若满足某条件,则 $Q(s,a) \geq 1$ ”)通过KBKR方法直接作用于价值函数。Maclin等人^[83]又在此基础上发展出pref-KBKR(偏好KBKR)方法,该方法将约束转化为动作偏好表达,使得if-then规则能够更直观地反映智能体在不同条件下的动作倾向。此外Knox和Stone提出了Q-Augmentation方法,通过将人类强化函数 $H(s,a)$ 与Q函数结合起来改进学习。研究发现,将 $H(s,a)$ 的折扣因子设为0(视为即时动作评估)效果最佳。该方法使人类价值判断直接融入强化学习过程。

4.2.3 策略塑造

策略塑造将建议直接集成到智能体的策略中。在无模型的策略塑造中,MacGlashan等人^[84]以及Najar等人^[85]的研究中均提供了基于评估反馈的无模型策略塑造的具体实例。在这些方法中,评估反馈被用作调整Actor-Critic架构中actor(即策略网络)的直接手段。尤其在MacGlashan等人提出的方法中,策略的权重向量 w 的更新根据策略的梯度进行了恰当的缩放,具体如下:

$$w \leftarrow w + \beta \nabla_w \ln \pi_w(a_t | s_t) H(s_t, a_t) \quad (6)$$

其中, $H(s_t, a_t)$ 是人类专家在 t 时刻提供的反馈, $\ln \pi_w(a_t | s_t)$ 是关于策略的梯度。

此外,关于有模型的策略塑造,Knox和Stone等人^[67,86]提出了两种基于模型的评估反馈策略优化方法。第一种方法是动作偏置(Action Biasing),它在决策过程中使用与Q-Augmentation相同的公式,但不会直接修改智能体的Q函数,具体如下:

$$a^* = \arg \max_a [Q(s, a) + \beta \cdot H(s, a)] \quad (7)$$

第二种方法是控制共享(Control Sharing),它通过概率标准在两种价值函数的决策之间进行判断。参数 β 用作阈值,以确定根据 $H(s, a)$ 选择决策的概率,具体如下:

$$\Pr(a = \arg \max_a [H(s, a)]) = \min(\beta, 1) \quad (8)$$

4.2.4 决策偏置

在前面提到策略塑造方法可以是无模型的,即直接修改智能体的策略,也可以是有模型的,即构建一个模型,在决策时用于引导策略的输出。另一种方法是使用建议直接在决策时调整策略的输出^[87],通过让智能体直接执行人类专家反馈的动作建议,而不破坏策略或对建议进行建模。这种策略被称为决策偏置,它是使用建议的最简单方式,它仅仅调整智能体的探索策略或者直接给予智能体一个动作建议,而不修改任何内部变量。在这种情况下,学

习通过遵循建议间接进行。

4.3 小结

人类反馈强化学习是一种将人类反馈融入传统强化学习过程的方法,通过让人类专家参与智能体的学习过程,提供实时的观察和指导,从而优化智能体在复杂环境中的决策能力。人类反馈强化学习解决了模仿学习高度依赖专家数据、限制智能体的探索能力和面对稀疏奖励表现不佳等局限,但是仍面临以下挑战:(1)人类反馈强化学习依赖于人类专家的实时反馈和指导,这要求专家在整个学习过程中持续参与,这不仅增加了人类专家的负担,也限制了算法的可扩展性,同时也会带来高成本和时间的消耗^[88];(2)人类提供的反馈质量和一致性可能存在波动,专家可能对同一情境给出不同的反馈,或者同一专家在不同时间点对相似情况的反馈可能不一致,这会导致智能体学习到的策略不稳定或偏离预期,还可能引入偏见和不公平性的传播,从而影响系统的整体性能和公正性;(3)在许多复杂和特殊的场景下,实时提供反馈是非常困难的,尤其是在复杂的任务或快速变化的环境中,反馈的延迟可能导致智能体无法及时调整策略,进而影响学习效果。此外,这种延迟还引发了确保反馈透明度和责任归属的问题,即如何明确反馈的来源及其对决策过程的影响,这对于维护系统的可靠性和公正性至关重要;(4)由于策略的学习依赖于特定的反馈,智能体在新的或未见过的语境中可能无法有效泛化。面对复杂或不熟悉的任务时,人类专家的反馈可能包含噪声或错误,如何处理这些噪声和错误反馈是一个重要的挑战,特别是在确保系统公平性和避免误导方面;(5)人类反馈强化学习的适应性和泛化能力在很大程度上依赖于人类专家的知识 and 经验。如果专家的知识有限,智能体的学习效果可能会受到限制。

5 AI 反馈强化学习

尽管基于人类反馈的强化学习解决了低样本效率和策略通用性的问题,但它仍然高度依赖于高质量的人类偏好标签。因此,探索人工生成标签作为替代方案成为了研究的重点挑战之一。随着大型语言模型的迅速发展,利用这些模型自动生成标签成为了一个有前景的解决方案。在此背景下,从 AI 反馈中进行强化学习(Reinforcement Learning from

AI Feedback, RLAIIF)应运而生。RLAIIF 是一种采用 AI 反馈指导智能体学习过程的方法,其核心是使用预训练的 LLMs(如 ChatGPT、DeepSeek 等)或先进的专家系统等作为反馈源。通过这种方式,智能体能够学习并优化其策略,以实现更优的行为表现(见图 8)。相较于 RLHF, RLAIIF 的主要特点及优势包括:自动化与高效性、成本效益以及一致性与客观性。这使得 RLAIIF 不仅能够缓解对稀缺且昂贵的人类偏好数据的依赖,同时还能维持甚至提升学习效果。

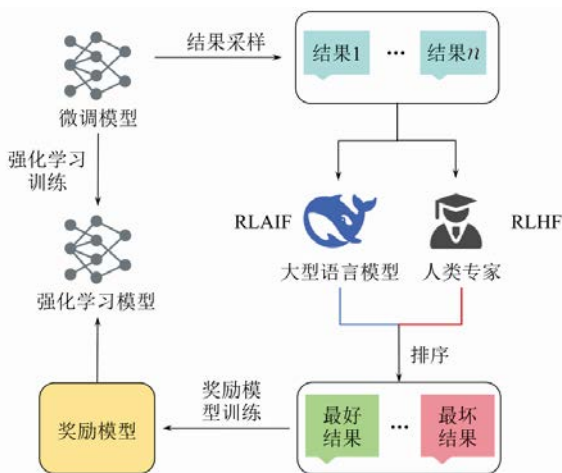


图 8 人类反馈和 AI 反馈的对比

5.1 AI 反馈的优势

5.1.1 自动化与高效性

RLAIIF 借助 LLMs 强大的语言理解和生成能力,能够深度解析复杂多变的情境,并自主生成海量且富含丰富上下文信息的反馈。这种即时且智能的反馈机制不仅能够高效处理简单任务,还能在高实时性和快速响应的挑战性场景中提升反馈效率,从而显著加速强化学习算法的训练进程,极大地提升了其整体效率。相比之下,尽管 RLHF 可以直接融入人类的洞察力与判断力,但其反馈收集过程过于依赖人工,导致在处理复杂或大规模任务时,反馈速度受限,效率瓶颈尤为突出。因此,作为一种创新的反馈强化学习范式,RLAIIF 弥补了 RLHF 在反馈速度和规模上的不足。

例如, CriticGPT^[89]展示了理解机器人操作轨迹视频并提供偏好反馈的能力,有效指导决策过程,减少对任务特定细节的依赖,提高反馈效率。OAIF^[90]通过每次训练迭代中采样当前模型响应,并利用 LLMs 实时标注,解决了传统离线反馈方法的数据滞后问题。Bai 等人^[91]提出的“Constitutional

AI”框架结合监督学习与强化学习,通过AI自我批评和修订初始响应,并利用AI生成的偏好数据替代人工反馈标签,优化了AI行为。Harrison Lee等的研究^[21]表明,RLAIF在总结、有益对话生成和无害对话生成任务上与RLHF效果相当,同时降低了人力成本,提高训练效率。最近,DeepSeek展示了其在处理复杂任务中的强大能力。DeepSeek通过优化Mixture-of-Experts(MoE)模型中的专家选择机制,提升自动化专家分配能力^[92,96]。这种动态选择专家的方法不仅减少了不必要的计算和资源浪费,还实现了高效的自动化决策过程。不同于依赖人工标注或人类反馈的传统方法,DeepSeek显著提高了AI反馈强化学习的效率。通过直接向LLMs请求奖励评分,省略了将偏好标签转化为奖励模型的步骤,简化了训练流程,进一步推动了强化学习的进展。实验表明,即使LLMs标注器的大小与策略模型相当,DeepSeek机制依然有效提升了整体训练效果。

5.1.2 成本效益

AI反馈机制在优化成本效益方面展现出卓越的经济优越性。具体而言,LLMs经过训练之后,其生成反馈的边际成本近乎于零,这使得AI反馈能够以高效的方式实现持续、大规模的反馈生成,而无需承担传统人类反馈模式所固有的高昂人工成本。尤其是在面对需要长期、高频次反馈的复杂任务时,人类反馈模式往往伴随着人力资源的持续投入与管理负担,导致成本随着时间和任务规模的扩展而急剧上升。相比之下,AI反馈通过高度自动化的流程,不仅有效减轻了对人力资源的过度依赖,还凭借其精准的语言理解和生成能力,确保了反馈的及时性与准确性,从而在成本控制与反馈质量之间达到了理想的平衡。

例如,Ding等人^[97]的研究表明,GPT-3在多种自然语言处理任务中展现出作为数据注释器的有效性,能够以相对较低的成本为不同任务提供注释数据,尤其适用于预算有限的场景。在某些情况下,基于GPT-3注释数据训练的模型性能可以与基于人工注释数据的模型相媲美甚至更优。此外,Gilardi等人^[98]的研究揭示了LLMs在多种任务中转换文本注释程序的潜力,表明ChatGPT等模型已成为一种优于传统人类注释的方法。相比传统的RLHF方法,DeepSeek^[99]通过自动化的反馈机制减少人工介入,使得AI训练成本大幅度降低,同时提高了训练的速度。DeepSeek-Coder-Instruct模型^[100]通过高质量

指令数据进行微调,提升了其在代码生成和理解任务中的表现,这种能力不仅减少了人工干预的需求,还提高了模型的泛化能力,进一步降低了开发和维护成本。

5.1.3 一致性与客观性

AI反馈机制凭借其高度的一致性与可扩展性,展现出在强化学习中生成稳定、自动化评估信号的能力,成为一种具备高效性和广泛适用性的反馈范式。不同于人类反馈容易受到主观判断、情绪波动或认知偏差的影响,AI反馈通过大型语言模型的推理能力,在面对相似输入时能够生成结构统一、标准明确的响应。这种特性使其特别适合用于需要高频、标准化反馈的场景,例如内容质量评估、风格一致性控制、多轮对话连贯性打分等。然而,当前AI在理解复杂情感、文化背景和主观价值判断方面仍存在局限,导致其反馈在人性化和情境适应性上有所欠缺。因此,AI反馈是一种在效率与一致性上具有明显优势,但在深度理解层面仍需提升的反馈范式。

例如,为降低对人工偏好评分的依赖,部分工作采用DeepSeek^[94,96]等大型语言模型作为自动反馈生成器,通过结构化提示从其输出中提取偏好信号,具有更高的可扩展性与一致性。Cao等人^[22]的研究揭示了RLAIF方法的局限性:虽然ChatGPT作为偏好标注器能提升开放域提示的响应质量,但其在特定任务上的准确性不足,导致响应正确性和真实性下降。为此,他们提出了混合强化学习框架,通过对不同提示类别实施分阶段标注策略,在保持RLAIF低成本优势的同时,提升AI标注的准确性和响应满意度。

5.2 AI反馈的复杂性 with 局限性

RLAIF作为一种强化学习的新范式,展现了显著的自动化和效率优势,但其复杂性和局限性同样突出。首先,RLAIF的复杂性体现在它对高质量反馈数据的高度依赖,尤其是在任务复杂且多变的情况下。如果训练数据存在偏差或缺乏广泛代表性,模型可能会学习到带有偏见的策略,影响其在新环境下的适应性。例如,在自动驾驶系统中,如果训练数据主要来自特定地理区域或天气条件,模型可能无法有效应对其他未曾覆盖的情境。

此外,缺乏真实人类理解的AI反馈机制在效率和性能上面临严峻挑战。尽管AI系统能够自动生成大量反馈并进行调整,但这些反馈通常基于算

法推测,而非基于人类专家的深度理解和情境感知。在处理复杂决策任务时, AI往往难以像人类专家那样结合直觉、经验和情境理解进行灵活应对,特别是在情感、伦理决策或动态变化环境中。例如,在医疗诊断中, AI可能无法捕捉患者的细微情感状态或伦理考量。在快速变化的环境中,如金融市场或紧急救援, AI也可能无法迅速适应新情况,导致策略执行和优化效果不佳。这种缺乏人类理解的机制使得 AI在复杂、动态情境中的适应性和决策精准性受限,尤其在高风险领域中, AI生成的反馈可能导致错误决策,责任归属变得复杂且困难。此外,在缺乏透明度和可解释性的情况下, RLAIIF可能导致模型在生成决策时偏离伦理标准,这不仅影响了决策的公正性,还增加了开发者无意间引入不符合伦理的决策的风险^[101]。

同时, AI反馈强化学习的设计与优化也是一个重要的挑战。为了生成高质量的反馈, AI模型通常需要处理大量输入数据并进行复杂的计算,这要求模型具备强大的计算能力和高效的算法支持。以 DeepSeek 为例, DeepSeek 通过优化 MoE 架构中的专家选择机制,提升了在不同任务中的自动化专家分配能力。这种方法提高了任务适应性,减少了不必要的计算和资源浪费,提升了决策过程的效率。然而,这种复杂设计也增加了模型的调试与维护难度,同时要求模型能够动态调整参数以适应不断变化的任务需求,这无疑加大了其部署和应用的复杂性。

RLAIIF 面临的另一挑战是反馈的稀疏性和延迟性,尤其是在复杂任务中,反馈信号未必能立即到达。反馈的稀疏性和延迟性使得模型在长期决策中无法实时获取指导,影响其学习效率和决策效果。具体来说,不同于监督信号密集的设置, AI生成的反馈往往仅在关键决策点出现,且分布不均,导致模型难以获得持续、细粒度的学习信号。这种稀疏性使得策略更新缺乏足够的引导,尤其在长序列、高维动作空间中更为明显。同时,反馈生成通常依赖于额外推理过程,引入显著的时间延迟,造成反馈与当前状态之间的时序错位,进一步削弱了指导的有效性。稀疏而滞后的反馈不仅降低了学习效率,还可能导致策略陷入局部最优或震荡收敛,特别是在需要精细控制和长期信用分配的任务中,这类问题尤为突出。

5.3 安全对齐与可信大模型

安全对齐(Safety Alignment)旨在确保 AI系统

的行为与人类的价值观和伦理标准一致,但其实现面临两大核心挑战。首先,对抗性攻击下的鲁棒性不足是一个重要问题。研究表明, LLMs 在面对对抗性提示(例如拼写错误、同义词替换)时,表现出显著的性能下降^[102]。这些对抗性攻击揭示了模型在复杂决策环境中的脆弱性,亟需提升其对抗干扰的能力。其次,隐私保护与模型性能的平衡同样是安全对齐中的一个关键挑战。尤其在联邦学习场景下, Yang 等人^[103]提出的 PR-PFHL 框架通过优化参数异构性,成功在多模态分布式训练中实现了隐私保护与模型可用性之间的平衡,为 RLAIIF 的安全对齐提供了新的解决思路。

为了克服上述挑战并提升 AI模型的可信度,必须在多个层面确保其可靠性和伦理性。公平性评估与修正是构建可信大模型的基础。Fleisig 等人^[104]提出的 FairPrism 数据集为评估 AI生成文本中的性别和文化偏见提供了有效工具,帮助减轻刻板印象和不公平性问题。此外,透明决策与可解释性也是提升模型可信度的重要方面。通过利用注意力机制可视化和概念分解等技术,能够追溯模型的决策过程,确保在高风险领域(如医疗、金融)中的决策结果具有可审计性和透明性。结合 Constitutional AI 中的自我改进机制^[91],这一方法通过减少对人工标注的依赖,提升了模型的无害化训练效果,从而进一步强化了模型的可信性。

展望未来,随着 AI技术不断进步,研究将更加关注如何在更复杂的应用场景中确保 AI系统的对齐和可信性。多模态对齐是一个重要的研究方向^[105],随着 DeepSeek 等大型语言模型的推出,如何在跨模态任务中设计有效的安全约束机制,确保文本、图像和视频生成的一致性,将成为研究中的重点挑战^[106-107]。与此同时,针对模型脆弱性的解决方案也亟待发展,动态防御体系的构建将成为未来研究的关键。通过开发针对对抗性攻击的实时检测与自适应防御框架^[102],可以有效增强模型的鲁棒性。此外,全生命周期可信保障将是未来的重要课题,从数据清洗、对齐算法设计到部署后的持续监控,建立端到端的风险管理体系,确保 AI系统的长期可信性和安全性。

5.4 多源反馈强化学习

随着 RLHF 和 RLAIIF 的兴起,融合多源反馈的方法论逐渐成为研究热点。一种常见的策略是利用 AI反馈进行大规模数据标注,以 AI反馈为工具,在

大规模无标签语料上进行偏好标签扩展, 典型如 RLAIIF 框架^[21-22], 通过少量人类反馈训练初始奖励模型, 再对大规模语料自动打分, 实现在保证偏好方向的同时提升训练效率。另一种方法则将 AI 反馈用作基于规则的安全约束机制, 如 Constitutional AI^[91] 采用预设原则评估和修正输出, 构建出结合形式规则与人类反馈数据驱动的混合反馈系统, 用于增强模型在高风险场景下的可控性与审计性; 此外, 还有一些研究聚焦于交互式推理过程监督框架, 通过引入过程监督(Process Supervision)或 AI 辩论(AI Debate)等机制, 使人类不仅评估模型最终输出, 更介入中间推理过程, 从而获得更细粒度、更具结构性的反馈信号, 提升模型推理路径的透明性与准确性。

在此基础上, 未来研究正朝着更具整合性与适应性的方向发展, 其中一个关键趋势是动态反馈仲裁(Dynamic Feedback Arbitration)机制的构建^[108]。该机制旨在引入元学习组件, 根据任务情境、领域属性或策略置信度动态评估人类反馈与 AI 反馈的可靠性, 计算不同反馈源的权重, 实现反馈融合的上下文感知。进一步地, 这一机制有望拓展为共适应反馈框架(Co-adaptive Feedback Framework): 其中, 人类反馈不仅用于训练主策略网络, 也用于持续优化生成 AI 反馈的批评模型。反过来, AI 模型也可用于识别人类标注中的系统性偏差, 辅助构建更稳健的监督数据。这种“人-机协同增强”的双向反馈循环, 有望成为构建长期对齐、高鲁棒性智能体的核心路径。

5.5 小结

AI 反馈机制本身带来了独特的伦理挑战: 首先, 基于训练数据的 AI 反馈可能存在隐性的偏差或缺乏多样性, 导致智能体学习到带有偏见的策略; 其次, AI 反馈的高度自动化可能使责任归属变得模糊, 特别是在复杂或多变的任务环境中。此外, 尽管 AI 反馈提高了效率, 但也需特别注意用户隐私和数据安全, 防止敏感信息泄露或滥用。最后, 提升 AI 反馈的可解释性与透明度, 对于增强用户对系统决策的理解与信任具有重要意义。在未来的发展趋势中, 整合 AI 反馈与人类反馈的优势, 将成为智能系统学习与发展的新方向。AI 反馈以其高度一致性、客观性、低成本和自动化能力, 在推动精密任务中的快速迭代与优化方面展现出巨大潜力; 而人类反馈则因其复杂情境理解力、情感共鸣及创造性思维, 在涉及伦理判断、情感理解和个性

化服务的场景中具有不可替代的作用。通过将两者有机结合, 可以形成一种兼具效率、精准度与人性化特质的反馈机制, 从而弥补单一反馈源的局限性。在此框架下, AI 将能够承担大量的重复性和技术性工作, 而在人类高层次的指导下, AI 将能够更深入地理解社会文化背景、个体情感需求及创新性问题解决方案。这种协作模式不仅能够增强强化学习在多样化的现实应用中的适应能力, 更能推动其决策过程向更全面、更深刻的方向发展, 以实现更高层次的人机协同智能。

6 RLXF 应用

6.1 自动驾驶领域

随着人工智能、物联网和虚拟现实技术的迅猛发展, 自动驾驶正逐步融入日常生活。然而, 现有 AI 系统在复杂场景理解和动态环境适应方面仍存在显著局限, 尚无法完全替代人类驾驶员应对所有突发交通状况。鉴于人类在复杂多变的驾驶环境中展现出的卓越鲁棒性与即时应变能力, 将人类智慧深度融入人工智能的训练循环中显得尤为关键。这一融合策略旨在利用人类的直观判断与经验, 推动机器学习算法向更高层次发展^[109]。例如 Liang 等人^[110]提出了一种层次模仿学习框架, 通过解耦自动驾驶的决策任务为车道选择与速度管理及精确控制两个层级, 利用基于变压器的模型降低学习复杂性, 并在 NeurIPS 2022 驾驶智能比赛中荣获双轨冠军, 验证了其优越性能。Wu 等人^[111]设计了一种控制转移机制, 允许人类在关键训练阶段实时干预决策。Yuan 等人^[112]开发了一种混合增强智能(Hybrid Augmented Intelligence, HAI)的方法, 利用人类反馈和互动强化学习提升自动驾驶车辆在实际交通中的决策与规划能力。

自动驾驶系统与行人交互的优化是当前研究的关键挑战。Richard 等人^[113]通过构建仿真环境收集人类反馈, 调整强化学习奖励函数来优化交互行为; Sun 等人^[114]引入多模态生理数据, 提升决策系统的舒适性和安全性。由于现实测试的高成本, 仿真技术成为重要替代方案, 但其真实性面临挑战。为此, Cao 等人^[115]提出 TrafficRLHF 框架, 通过人类反馈强化学习增强交通仿真的真实性: 首先收集人类对交通场景视频的评分, 训练奖励模型评估场景真实性, 最终优化交通模型生成更符合人类认知

的仿真环境。

在自动驾驶技术的广阔应用版图中,除了自动驾驶汽车这一前沿阵地,RLXF亦在自主水下航行器(Autonomous Underwater Vehicle, AUV)及无人机等自动驾驶领域展现出独特价值。针对AUV在复杂海洋环境中执行任务的需求,Zhang等人^[116]提出的深度交互强化学习方法创新性地融合了人类先验知识与环境反馈,解决了深海环境实时奖励缺失的难题,实现了复杂海洋条件下的自主路径优化。这一技术突破与Pollak等人^[117]在无人机领域的探索形成方法论上的呼应,后者通过将人类面部表情转化为强化学习奖励信号,开创了基于情感反馈的行为优化新范式。两项研究共同证明了RLXF在非结构化环境中的独特优势:前者攻克了感知受限条件下的自主决策挑战,后者则突破了传统人机交互的认知边界,为自动驾驶系统的人性化发展提供了关键技术支撑。

此外,X反馈强化学习在具身智能中的应用已逐渐形成一种“大规模预训练+反馈微调”的两阶段范式。该范式的思路在于:首先利用文本与视频数据,预训练一个具备通用感知-动作能力的视觉-语言-动作(Vision-Language-Action, VLA)基础模型,如Google的RT-2^[118],使其初步掌握物体属性、物理交互规律及语言指令理解等先验知识。随后,在具体任务场景中,借助RLXF方法对预训练模型进行高效微调,而非从零开始学习策略。这一过程通常基于真实世界中多样化的反馈信号实现,例如:人类对机器人执行任务的轨迹视频进行偏好排序^[1]、提供实时语言修正指令,或通过物理示教采集反馈数据。通过对反馈信号的有效建模,系统能够在保留VLA模型泛化能力的同时,快速适应新任务需求。DeepMind提出的RoboCat^[119]即为此类架构的典型代表,展示了该范式下智能体可通过持续吸收任务反馈数据实现迭代进化的能力。综上所述,将VLA模型的广域知识与RLXF的精细对齐相结合,被认为是当前构建具备泛化性与可信性的具身智能系统的最具前景的技术路径。

6.2 大型语言模型领域

在大型语言模型中,RLXF发挥着重要作用,包括OpenAI的ChatGPT^[120]、DeepMind的Sparrow^[121]、Anthropic的Claude^[91]以及近来广受关注的DeepSeek^[99]等。RLXF在大型语言模型中的应

用包含三个阶段。首先基于预训练模型初始化,利用其无监督学习获得的基础文本生成能力;随后构建专门的奖励模型,通过人工标注的文本质量排序数据训练优化的LLMs架构,实现文本到量化评分的精确映射;最后建立强化学习循环,通过生成-评估-更新的迭代过程持续优化模型性能。这一框架克服了完全依赖人类反馈训练的不可行性,同时确保了模型输出与人类目标的一致性。

大型语言模型的核心创新在于改进的RLXF训练框架,该框架在保持模型规模不变的情况下,通过三阶段优化显著提升性能:(1)基于人工标注数据的监督微调;(2)奖励模型构建;(3)强化学习优化。虽然具体实现细节(如参数冻结策略)未完全公开,但其基于instructGPT框架的改进取得了显著效果。类似地,Ouyang等人^[122]在AI智能体研发中采用了迭代式在线训练范式,通过周期性更新偏好模型和强化学习策略,结合新采集的人类反馈数据,有效提升了自然语言助手的实用性和安全性。这些工作共同推动了基于人类反馈的强化学习技术在语言模型优化中的应用发展。

6.3 其他领域

社交机器人技术的快速发展得益于自然语言处理与强化学习技术的深度融合。在这一领域中,如何有效利用人类反馈来提升机器人的交互能力成为关键科学问题。Maroto-Gómez等人^[123]提出的个性化活动选择框架,通过多臂老虎机算法将显式评分与隐式交互反馈相结合,实现了社交机器人对用户偏好的持续学习。该研究突破单一反馈模式的局限,构建了融合用户主观评价与行为数据的混合学习机制,提升了偏好预测的准确性与适应性。实验表明,该框架在长期人机交互中既能保持自然流畅的交互体验,又能通过动态调整活动策略实现精准个性化推荐,为构建具备持续进化能力的社交机器人系统提供了重要技术路径,对促进助老娱乐、认知辅助等实际应用具有关键推动作用。

在医学领域,ChatGPT展示了编写与医学问题紧密相关摘要和文本的潜力,能够满足学生及研究人员的咨询需求。它不仅支持知识传授,还能通过个性化测试减轻教师负担。作为医疗保健的信息提供者和沟通工具,ChatGPT增强了医疗服务的可访问性,并帮助克服语言障碍。然而,其生成医学数据的可靠性是一个挑战,尤其是误导信息和虚假引用可能威胁医疗研究和治疗。Wang等人^[124]针对这

一问题开发了 IvyGPT, 这是一款通过高质量医学问答实例和 RLHF 进行深度训练与微调的交互式中文语言模型, 能够提升多轮对话的流畅度与诊断建议的人性化水平。尽管 IvyGPT 的综合诊断能力不及专业医生, 但其优化输出更贴近临床实际, 为医学 AI 的发展提供了新动力。

表 2 概述了 X-反馈强化学习在不同应用领域的具体实例及其所采用的反馈方式。该表格详细列出了多个代表性应用场景。每项应用不仅明确了具体的任务目标, 还指出了相应的反馈机制, 展示了 X-反馈强化学习在多个领域的广泛应用潜力, 并验证了其在复杂任务中的有效性和灵活性。

表 2 X-反馈强化学习的应用领域及反馈方式

应用领域	具体应用	反馈方式	文献
自动驾驶	自动驾驶汽车	建议反馈(纠正反馈), 偏好反馈(动作, 轨迹)	[111,113]
	模拟真实交通模型	偏好反馈(状态)	[115]
	自主水下航行器	建议反馈(评估反馈-奖励)	[116]
	无人机	建议反馈(评估反馈-奖励)	[117]
智能应用	具身智能	偏好反馈(轨迹)	[125-129]
大型语言模型	DeepSeek、ChatGPT、Sparrow、Claude	偏好反馈(状态)	[91,99,120-121]
医疗领域	医学问答	偏好反馈(状态)	[124]
机器人领域	社交机器人	建议反馈(评估反馈-策略), 偏好反馈(奖励)	[123,130]

6.4 工具与资源

X-反馈强化学习(RLXF)的广泛应用催生了各类专业仿真平台的研发(如表 3 所示), 为算法迭代优化提供了关键支撑。在自动驾驶领域, Carla 平台^[123]通过开源架构支持城市级自动驾驶系统的开发测试, 其灵活的传感器配置和环境控制功能为复杂算法验证创造了条件。CarSim 则专注于高精度车辆动力学仿真, 为控制系统的开发提供可靠测试环境。机器人研究领域, Gazebo 凭借其强大的三维仿真能力

和物理引擎, 成为机器人控制算法验证的重要工具。为满足不同研究需求, 学界开发了多样化的测试平台: 从经典的 GridWorld 二维环境到 Mario AI、Pac-Man 等游戏平台, 为算法测试提供了状态空间复杂度各异的实验场景; Nao 机器人和 TurtleBot 平台则专注于人机交互研究; 而 OpenAI Gym 通过标准化环境接口, 大幅提升了强化学习研究的可重复性和比较性。这些平台共同构成了 RLXF 技术从理论研究到实际应用的重要桥梁。

表 3 实验平台

应用领域	平台名称	平台功能	适用系统	文献	核心反馈类型
自动驾驶	Carla ^①	丰富环境的无人车模拟器	Linux	[114,131]	RLHF
	CarSim ^②	车辆动力学的仿真软件	Windows	[132]	RLHF
	Gazebo ^③	搭载传感器的三维动态机器人模拟器	Windows/Linux	[116]	RLHF
	AirSim ^④	多无人车、无人机高仿真模拟器	Windows/Linux	[133]	RLHF
	JSBSim ^⑤	飞行动力学软件库	Windows/Linux	[134]	RLHF
算法测试	GridWorld ^⑥	二维网格模拟平台	Windows/Linux	[135-137]	RLHF
	Mario AI ^⑦	交互式强化学习游戏平台	Windows/Linux	[138]	RLAIF
	Pac-Man ^⑧	交互式强化学习游戏平台	Windows/Linux	/	/
	OpenAI Gym ^⑨	强化学习标准化的环境平台	Windows/Linux	[1,12,139] [42,56]	RLHF 模仿学习
模拟机器人	Nao Robot ^⑩	多种传感器机器人仿真器	Windows/Linux	/	/
	Turtle Bot ^⑪	移动机器人平台	Windows/Linux	[140-141] [36-37,43,52]	RLHF 模仿学习
	MuJoCo ^⑫	机器人动力学模拟仿真器	Windows/Linux	[1,16,19-20,65,67,142-143]	RLHF

① <https://carla.org/>

② <https://www.carsim.com/>

③ <https://gazebo.org/home>

④ <https://github.com/microsoft/AirSim>

⑤ <https://github.com/JSBSim-Team/jsbsim>

⑥ <http://www.gridworld.com/>

⑦ <https://github.com/aleju/mario-ai>

⑧ <https://pacman.com/en/>

⑨ <https://github.com/openai/gym>

⑩ <https://www.aldebaran.com/en/nao>

⑪ <https://www.turtlebot.com/>

⑫ <https://github.com/deepmind/mujoco>

此外,表 4 补充汇总了已有研究中应用 RLXF 框架所依赖的数据集与实验平台,涵盖了从虚拟到现实的多样化资源。这些实例不仅体现了 RLXF 方

法在不同应用场景下的适应性与有效性,也为读者提供了理解其实际部署路径的重要参考。

表 4 RLXF 框架应用实例

文献	研究领域	数据集/实验平台	文献	研究领域	数据集/实验平台
[38]	模仿学习	Digits-five, Office-Caltech10, Office-31	[94]	RLAIF	多源数据(网页截图、PDF、OCR 等)
[39]	模仿学习	Super Tux Kart open source 3D racing game 和 Mario Bros.game.	[98]	RLAIF	文本标注任务
[54]	模仿学习	DeepMind Control Suite benchmark	[104]	RLAIF	FairPrism 数据集(AI 生成文本)
[63]	模仿学习	RoboCup soccer tasks	[107]	RLAIF	MM-RLHF 数据集
[70]	RLHF	Mountain Car 和 Pac-Man	[115]	RLHF	nuScenes 数据集
[81]	RLHF	RoboCup soccer tasks	[143]	RLHF	PyBullet
[89]	RLAIF	Meta-World 任务	[145]	RLHF	Minecraft 游戏

7 挑战与展望

在强化学习领域,尽管 X-反馈被视为规避有害或错误结果的关键手段,但其本身亦非无懈可击。本文在列举一系列前沿研究工作的同时,也提出亟需解决的开放性问题。

(1) RLXF 的鲁棒性。X-反馈存在例如个体主观性、不一致性等固有局限,即专家之间可能存在理解差异与期望偏差,其反馈往往带有主观色彩并可能引入误差,可能会导致智能体接收到的信息相对混杂,进而直接影响智能体学习的有效性与准确性。此外,人类专家/AI 的认知边界与注意力局限亦限制了反馈的全面性和精准度,导致难以全面捕捉任务的复杂性与全局最优策略,进而制约了智能体的学习潜能与性能上限。对此,未来可以考虑引入多专家反馈聚合机制,减轻单一专家主观性和潜在误差的影响。另外,在 RLXF 参与强化学习训练的过程中,迭代地进行 RLXF(即使用 RLXF 策略生成新的响应对、进行 RLXF 和重复此过程),RLXF 如何适应 RL 模型设置,以及如何利用 X-反馈进行更具体的信用分配都是未来工作的研究方向。

(2) RLXF 的成本代价。频繁的反馈介入不仅消耗时间与资源,还限制强化学习系统在实际应用中的自主性与可扩展性。尤其是当涉及大量专家反馈时,大规模应用中收集和利用 X-反馈面临的时间成本高昂与资源密集挑战尤为显著。因此,如何设计新的范式,高效即时地利用 X-反馈缩短反馈收集周期,实现反馈收集的精准化,减少冗余反馈的收集,是目前面临的重大挑战。未来可以将主动学习或人反馈。人类反馈强化学习可以在高风险任务或复杂目标等关键节点关键时刻介入反馈辅助 AI 反馈机

制。而 AI 反馈机制则可以自动生成辅助性反馈,降低对人类专家的频繁依赖,从而缩短反馈收集周期并提高反馈的精准度与资源利用效率。

(3) RLXF 的及时性。及时的反馈对于迅速调整智能体的策略是至关重要的。然而,人类反馈的固有延迟性源于专家参与的间歇性与反馈整合的复杂性,常阻碍训练过程中的即时指导。在强化学习中,智能体需要快速获得反馈以调整其行为。如果反馈延迟过长,智能体可能已经改变了状态或采取了新的行动,导致反馈变得无关紧要或误导性。另外,在快速变化的环境中,反馈的及时性对于智能体适应环境的变化至关重要。如果反馈延迟,智能体可能无法跟上环境的变化速度。因此,如何获取及时反馈以响应智能体的需求是目前面临的一大瓶颈。虽然现在 AI 反馈可以有效解决此问题,但是其稳定性以及反馈的质量仍然是未来需要解决的挑战。

(4) RLXF 的可扩展性。当涉及大规模的状态空间和动作空间时,智能体需要处理的信息量急剧增加,这就要求反馈机制必须能够高效地筛选出有价值的信息。此外,随着应用场景的多样化,如何设计既灵活又能适应不同场景的反馈机制成为一大难题。例如,在现实世界的应用中,智能体可能需要在不同的环境条件下学习,而这些环境条件的变化可能导致原先有效的反馈不再适用。因此,如何在不牺牲性能的前提下,设计出既能适应大规模数据集又能灵活应对各种变化的反馈机制,是 RLXF 领域需要解决的核心挑战。未来一个极具潜力的方向在于探索多模态数据的融合与应用。但它也带来了新的挑战,包括数据融合过程的复杂性提升、多模态数据之间的精确对齐难题,以及在高维数据空间中如何高效处理与筛选有效信息。因此,

设计并开发高效的算法框架,以整合并有效处理这些复杂的多源多模态数据,是未来研究的一个关键议题。

8 总结语

人机共融是目前人工智能研究领域最主流的方向之一,而 X-反馈强化学习又是其最为重要的分支。本文深入剖析了 X-反馈强化学习这一人工智能研究前沿领域的核心概念及其演进历程,系统回顾了 X-反馈强化学习中模仿学习与人类反馈强化学习两大支柱,并调研了 AI 反馈强化学习的最新进展,随后,本文聚焦于大型语言模型、无人驾驶、智能机器人及医学等多个关键应用领域,全面梳理了 X-反馈强化学习的实践成果与潜在价值,凸显了其跨领域的广泛适用性与深远影响。此外,本文从高质量的人类反馈、成本代价、及时性和构建高效协同的交互机制等维度,提出了当前 X-反馈强化学习研究面临的挑战及应对策略,旨在为未来研究方向与策略优化提供借鉴。鉴于大型语言模型正引领人工智能进入新纪元,对 X-反馈强化学习的持续深入探索,不仅关乎技术本身的进步,更是推动人工智能全面落地、实现人机共融的关键所在,具有重要的理论意义与实践价值。

致谢 本课题得到国家自然科学基金杰青项目(62325602)、国家自然科学基金面上项目(62276238)、国家自然科学基金区域联合重点支持项目(U24A20326)、国家自然科学基金重点项目(62036010)、河南省优秀青年科学基金(232300-421095)以及河南省教育厅高校科技创新人才支持计划(25HASTIT034)资助。

参考文献

- [1] Christiano P F, Leike J, Brown T, et al. Deep reinforcement learning from human preferences//Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS). Long Beach, USA, 2017: 4299-4307
- [2] Marom O, Rosman B. Belief reward shaping in reinforcement learning//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 3762-3769
- [3] Mathewson K W, Pilarski P M. Simultaneous control and human feedback in the training of a robotic agent with actor-critic reinforcement learning. arxiv preprint arxiv:1606.06979, 2016
- [4] Skalse J, Howe N, Krashennnikov D, et al. Defining and characterizing reward gaming//Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA, 2022: 9460-9471
- [5] Lee K, Liu H, Ryu M, et al. Aligning text-to-image models using human feedback. arxiv preprint arxiv:2302.12192, 2023
- [6] Ibarz B, Leike J, Pohlen T, et al. Reward learning from human preferences and demonstrations in atari//Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS). Montreal, Canada, 2018: 8022-8034
- [7] Hejna III D J, Sadigh D. Few-shot preference learning for human-in-the-loop rl//Proceedings of the 7th Conference on Robot Learning (CoRL). Atlanta, USA, 2023: 2014-2025
- [8] OpenAI. Chatgpt: Optimizing language models for dialogue. <https://openai.com/blog/chatgpt>, 2022
- [9] Koppol P, Admoni H, Simmons R. Iterative interactive reward learning//Proceedings of the International Conference on Machine Learning (ICML) Workshop on Participatory Approaches to Machine Learning. Virtual, 2020
- [10] Shaunak A Mehta and Dylan P Losey. Unified learning from demonstrations, corrections, and preferences during physical human-robot interaction. ACM Transactions on Human-Robot Interaction, 2023, 13(3): 39:1-39:25
- [11] Hwang M, Lee G, Kee H, et al. Sequential preference ranking for efficient reinforcement learning from human feedback//Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA, 2023: 49088-49099
- [12] Reddy S, Dragan A, Levine S, et al. Learning human objectives by evaluating hypothetical behavior//Proceedings of the 37th International Conference on Machine Learning (ICML). Vienna, Austria, 2020: 8020-8029
- [13] Liu Y, Datta G, Novoseller E, et al. Efficient preference-based reinforcement learning using learned dynamics models//Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA). London, UK, 2023: 2921-2928
- [14] Ren Z, Liu A, Liang Y, et al. Efficient meta reinforcement learning for preference-based fast adaptation//Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA, 2022: 15502-15515
- [15] Xie A, Singh A, Levine S, et al. Few-shot goal inference for visuomotor learning and planning//Proceedings of the 2nd Conference on Robot Learning (CoRL). Zurich, Switzerland, 2018: 40-52
- [16] Liu R, Du Y, Bai F, et al. PEARL: zero-shot cross-task preference alignment and robust reward learning for robotic manipulation//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria, 2024: 30946-30964
- [17] Abramson J, Ahuja A, Carnevale F, et al. Improving multimodal interactive agents with reinforcement learning from human feedback. arxiv preprint arxiv:2211.11602, 2022
- [18] Ziegler D M, Stiennon N, Wu J, et al. Fine-tuning language models from human preferences. arxiv preprint arxiv:1909.08593, 2019
- [19] Cao Z, Wong K C, Lin C T. Weak human preference supervision for deep reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(12): 5369-5378
- [20] Zhan H X, Tao F, Cao Y C. Human-guided robot behavior learning: A GAN-assisted preference-based reinforcement learning approach. IEEE Robotics and Automation Letters. 2021, 6(2): 3545-3552
- [21] Lee H, Phatale S, Mansoor H, et al. RLAIIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria, 2024: 26874-26901

- [22] Li A, Xiao Q G, Cao P, et al. Hrlaif: Improvements in helpfulness and harmlessness in open-domain reinforcement learning from ai feedback. arxiv preprint arxiv:2403.08309, 2024
- [23] Yu T Y, Zhang H Y, Yao Y, et al. RLAIIF-V: Aligning mllms through open-source AI feedback for super GPT-4V trustworthiness. arxiv preprint arxiv:2405.17220, 2024
- [24] Sharma A, et al. A critical evaluation of AI feedback for aligning large language models//Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada, 2024: 29166-29190
- [25] Zhang H, Chen J, Jiang F, et al. HuatuoGPT, towards taming language model to be a doctor//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). Singapore, 2023: 10859-10885
- [26] Hengle A, Padhi A, Singh S, et al. Intent-conditioned and non-toxic counterspeech generation using multi-task instruction tuning with RLAIIF//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Mexico City, Mexico, 2024: 6716-6733
- [27] Li Lu-Lu, Zhu Rui-Jie, Sui Lu-Yao, et al. A review of reinforcement learning methods for intelligent cluster systems. Chinese Journal of Computers, 2023, 46(12): 2573-2596(in Chinese)
(李璐璐, 朱睿杰, 隋璐瑶等. 智能集群系统的强化学习方法综述. 计算机学报, 2023, 46(12): 2573-2596)
- [28] Li Ming-Yang, Xu Ke-Er, Song Zhi-Qiang, et al. A review of multi-agent reinforcement learning algorithms. Journal of Frontiers of Computer Science and Technology, 2024, 18(8): 1979-1997(in Chinese)
(李明阳, 许可儿, 宋志强等. 多智能体强化学习算法研究综述. 计算机科学与探索, 2024, 18(08):1979-1997)
- [29] Ding Shi-Fei, Du Wei, Zhang Jian, et al. Research progress of multi-agent deep reinforcement learning. Chinese Journal of Computers, 2024, 47(07):1-21 (in Chinese)
(丁世飞, 杜威, 张健等. 多智能体深度强化学习研究进展. 计算机学报, 2024, 47(7): 1-21)
- [30] Wang Shuo-Ru, Niu Wen-Jia, Tong En-Dong, et al. A review of offline strategy evaluation for reinforcement learning. Chinese Journal of Computers, 2022, 45(9): 1926-1945(in Chinese)
(王硕汝, 牛温佳, 童恩栋等. 强化学习离线策略评估研究综述. 计算机学报, 2022, 45(09): 1926-1945)
- [31] Wu Lan, Liu Quan, Huang Zhi-Gang, Zhang Li-Hua. Review of offline reinforcement learning research. Chinese Journal of Computers, 2025, 8(1): 156-187(in Chinese)
(乌兰, 刘全, 黄志刚, 张立华. 离线强化学习研究综述. 计算机学报, 2025, 48(1): 156-187)
- [32] Bain M and Sammut C. A framework for behavioural cloning. Machine Intelligence 15, 1995: 103-129
- [33] Ng A Y, Russell S, et al. Algorithms for inverse reinforcement learning//Proceedings of the 17th International Conference on Machine Learning (ICML). Stanford, USA, 2000: 663-670
- [34] Ho J, Ermon S. Generative adversarial imitation learning//Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS). Barcelona, Spain, 2016: 4565-4573
- [35] Torabi F, Warnell G, Stone P. Recent advances in imitation learning from observation//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China, 2019: 6325-6331
- [36] Papagiannis G, Li Y P. Imitation learning with sinkhorn distances//Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD). Grenoble, France, 2022: 116-131
- [37] Dadashi R, Hussenot L, Geist M, et al. Primal Wasserstein imitation learning//Proceedings of the 9th International Conference on Learning Representations (ICLR). Virtual, 2021
- [38] Nguyen T, Le T, Zhao H, et al. MOST: Multi-source domain adaptation via optimal transport for student-teacher learning//Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI). Virtual, 2021: 225-235
- [39] Ross S, Bagnell D. Efficient reductions for imitation learning//Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS). Sardinia, Italy, 2010: 661-668
- [40] Ross S, Gordon G, Bagnell D. A reduction of imitation learning and structured prediction to no-regret online learning//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS). Fort Lauderdale, USA, 2011: 627-635
- [41] Ross S and Bagnell A. Reinforcement and imitation learning via interactive no-regret learning. arxiv preprint arxiv:1406.5979, 2014
- [42] Chen L L, et al. Decision transformer: Reinforcement learning via sequence modeling//Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS). Virtual, 2021: 15084-15097
- [43] Furuta H, Matsuo Y, Gu S S. Generalized decision transformer for offline hindsight information matching. arxiv preprint arxiv:2111.10364, 2021
- [44] Ma Y, et al. Rethinking decision transformer via hierarchical reinforcement learning//Proceedings of the 41st International Conference on Machine Learning (ICML). Vienna, Austria, 2024: 33730-33745
- [45] Zhang Li-Hua, Liu Quan, Huang Zhi-Gang, et al. A review of reverse reinforcement learning research. Journal of Software, 2022, 34(10): 4772-4803 (in Chinese)
(张立华, 刘全, 黄志刚等. 逆向强化学习研究综述. 软件学报, 2022, 34(10): 4772-4803)
- [46] Song Li, Li Da-Zi, Xu Xin. Review of inverse reinforcement learning algorithm, theory and application research. Acta Automatica Sinica, 2024, 50(9): 1704-1723(in Chinese)
(宋莉, 李大字, 徐昕. 逆向强化学习算法、理论与应用研究综述. 自动化学报, 2024, 50(9): 1704-1723)
- [47] Abbeel P, Ng A Y. Apprenticeship learning via inverse reinforcement learning//Proceedings of the 21st International Conference on Machine Learning (ICML). Banff, Canada, 2004: 1
- [48] Ratliff N D, Bagnell J A, Zinkevich M A. Maximum margin planning//Proceedings of the 23rd International Conference on Machine Learning (ICML). Pittsburgh, USA, 2006: 729-736
- [49] Ziebart B D, Maas A, Bagnell J A, et al. Maximum entropy inverse reinforcement learning//Proceedings of the 23rd National Conference on Artificial Intelligence. Chicago, USA, 2008: 1433-1438
- [50] Lin Jia-Hao, Zhang Zong-Zhang, Jiang Chong, et al. A review of imitation learning based on generative adversarial networks. Chinese Journal of Computers, 2020, 43(2): 326-351(in Chinese)
(林嘉豪, 章宗长, 姜冲等. 基于生成对抗网络的模仿学习综述. 计算机学报, 2020: 43(02): 326 - 351)

- [51] Finn C, Christiano P, Abbeel P, et al. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. arxiv preprint arxiv:1611.03852, 2016
- [52] Wang B Z, Wu G Q, Pang T, et al. DiffAIL: Diffusion adversarial imitation learning//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 15447-15455
- [53] Liu W L, Li D Y, Aasi E, et al. Interpretable generative adversarial imitation learning. arxiv preprint arxiv:2402.10310, 2024
- [54] Chang J D, Sreenivas D, Huang Y, et al. Adversarial imitation learning via boosting//Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria, 2024. Note={Paper ID: DuQkqSc9en}
- [55] Sharma P, Pathak D, Gupta A. Third-person visual imitation learning via decoupled hierarchical controller//Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada, 2019: 2593-2603
- [56] Edwards A, Sahni H, Schroecker Y, et al. Imitating latent policies from observation//Proceedings of the 36th International Conference on Machine Learning (ICML). Long Beach, USA, 2019: 1755-1763
- [57] Kidambi R, Chang J, Sun W. Mobile: Model-based imitation learning from observation alone//Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS). Virtual, 2021: 28598-28611
- [58] Hanna J and Stone P. Grounded action transformation for robot learning in simulation//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 3834-3840
- [59] Nair A, Chen D, Agrawal P, et al. Combining self-supervised learning and imitation for vision-based rope manipulation//Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA). Singapore, 2017: 2146-2153
- [60] Torabi F, Warnell G, Stone P. Behavioral cloning from observation//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 4950-4957
- [61] Gupta A, Devin C, Liu Y, et al. Learning invariant feature spaces to transfer skills with reinforcement learning//Proceedings of the 5th International Conference on Learning Representations (ICLR). Toulon, France, 2017
- [62] Aytar Y, Pfaff T, Budden D, et al. Playing hard exploration games by watching YouTube//Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS). Montreal, Canada, 2018: 2935-2945
- [63] Taylor M, Stone P, and Liu Y X. Transfer learning via inter-task mappings for temporal difference learning. *Journal of Machine Learning Research*, 2007, 8(9): 2125-2167
- [64] Sermanet P, Lynch C, Chebotar Y, et al. Time-contrastive networks: Self-supervised learning from video//Proceedings of the 2018 IEEE International Conference on Robotics and Automation. Brisbane, Australia, 2018: 1134-1141
- [65] Fickinger A, Cohen S, Russell S, et al. Cross-domain imitation learning via optimal transport. arxiv preprint arxiv:2110.03684, 2021
- [66] Thomaz A L, Hoffman G, Breazeal C. Real-time interactive reinforcement learning for robots//AAAI 2005 Workshop on Human Comprehensible Machine Learning. Pittsburgh, USA, 2005: 1-5
- [67] Knox W B, Stone P. Reinforcement learning from simultaneous human and MDP reward//Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS). Valencia, Spain, 2012: 475-482
- [68] Dubey R, et al. Investigating human priors for playing video games//Proceedings of the 35th International Conference on Machine Learning (ICML). Stockholm, Sweden, 2018: 1348-1356
- [69] Wirth C, Akrou R, Neumann G, et al. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 2017, 18(136): 1-46
- [70] Torrey L, Taylor M. Teaching on a budget: Agents advising agents in reinforcement learning//Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS). Saint Paul, USA, 2013: 1053-1060
- [71] Hüllermeier E, Fürnkranz J, Cheng W, et al. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 2008, 172(16-17): 1897-1916
- [72] Wirth C, Fürnkranz J. On learning from game annotations. *IEEE Transactions on Computational Intelligence and AI in Games*, 2014, 7(3): 304-316
- [73] Zucker M, Bagnell J A, Atkeson C G, et al. An optimization approach to rough terrain locomotion//Proceedings of the 2010 IEEE International Conference on Robotics and Automation (ICRA). Anchorage, USA, 2010: 3589-3595
- [74] Busa-Fekete R, et al. Preference-based evolutionary direct policy search//Proceedings of the ICRA Workshop on Autonomous Learning. Karlsruhe, Germany, 2013
- [75] Wirth C, Fürnkranz J, Neumann G. Model-free preference-based reinforcement learning//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual Event, 2022
- [76] Rafailov R, Sharma A, Mitchell E, et al. Direct preference optimization: Your language model is secretly a reward model//Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA, 2023: 53728-53741
- [77] Azar M G, Guo Z D, Piot B, et al. A general theoretical paradigm to understand learning from human preferences//Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS). Valencia, Spain, 2024: 4447-4455
- [78] Dong Z, Yuan Y, Hao J, et al. AlignDiff: Aligning diverse human preferences via behavior-customisable diffusion model//Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria, 2024
- [79] Wang Y, Liu Q, Jin C. Is RLHF more difficult than standard RL? A theoretical perspective//Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA, 2023: 76006-76032
- [80] Huang J, Hao J, Juan R, et al. GAN-based interactive reinforcement learning from demonstration and human evaluative feedback//Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA). London, UK, 2023: 4991-4998
- [81] Torrey L, Walker T, Maclin R, et al. Advice taking and transfer learning: Naturally inspired extensions to reinforcement learning//Proceedings of the AAAI Fall Symposium on Naturally-Inspired Artificial Intelligence. Arlington, USA, 2008: 103-110
- [82] Maclin R, Shavlik J, Walker T, Torrey L. Knowledge-based support vector regression for reinforcement learning//Reasoning Representation, and Learning in Computer Games. Edinburgh, UK, 2005: 1-6

- [83] Maclin R, Shavlik J W, Torrey L, et al. Giving advice about preferred actions to reinforcement learners via knowledge-based kernel regression//Proceedings of the 20th National Conference on Artificial Intelligence (AAAI). Pittsburgh, USA, 2005: 819-824
- [84] MacGlashan J, Ho M, Loftin R, et al. Interactive learning from policy dependent human feedback//Proceedings of the 37th International Conference on Machine Learning (ICML). Sydney, Australia, 2017: 2285-2294
- [85] Najar A, Sigaud O, Chetouani M. Interactively shaping robot behaviour with unlabeled human instructions. *Autonomous Agents and Multi-Agent Systems*, 2020, 34: 1-35
- [86] Knox W B, Stone P. Combining manual feedback with subsequent MDP reward signals for reinforcement learning//Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems. Toronto, Canada, 2010: 5-12
- [87] Cruz F, Twiefel J, Magg S, et al. Interactive reinforcement learning through speech guidance in a domestic scenario//Proceedings of the 2015 International Joint Conference on Neural Networks. Killarney, Ireland, 2015: 1-8
- [88] Cao Hong-Ye, Liu Xiao, Dong Shao-Kang, et al. A review of interpretability research for reinforcement learning. *Chinese Journal of Computers*, 2024, 47(8): 1853-1882(in Chinese)
(曹宏业, 刘潇, 董绍康等. 面向强化学习的可解释性研究综述. *计算机学报*, 2024, 47(8): 1853-1882)
- [89] Liu J, Yuan Y, Hao J, et al. Enhancing robotic manipulation with ai feedback from multimodal large language models. arxiv preprint arxiv:2402.14245, 2024
- [90] Guo S, Zhang B, Liu T, et al. Direct language model alignment from online ai feedback. arxiv preprint arxiv:2402.04792, 2024
- [91] Bai Y T, Kadavath S, Kundu S, et al. Constitutional AI: Harmlessness from ai feedback. arxiv preprint arxiv:2212.08073, 2022
- [92] Kaufmann T, Weng P, Bengs V, Hüllermeier E. A survey of reinforcement learning from human feedback. arxiv preprint arxiv:2312.14925, 10, 2023
- [93] Dai D, et al. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL). Bangkok, Thailand, 2024: 1280-1297
- [94] Lu H, Liu W, Zhang B, et al. DeepSeek-v1: Towards real-world vision-language understanding. arxiv preprint arxiv:2403.05525, 2024
- [95] Liu A, Feng B, Wang B, et al. DeepSeek-v2: A strong, economical, and efficient mixture-of-experts language model. arxiv preprint arxiv:2405.04434, 2024
- [96] Liu A, Feng B, Xue B, et al. DeepSeek-v3 technical report. arxiv preprint arxiv:2412.19437, 2024
- [97] Ding B, et al. Is GPT-3 a good data annotator?//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). Toronto, Canada, 2023: 11173-11195
- [98] Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd-workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 2023, 120(30): e2305016120
- [99] Guo D, Yang D, Zhang H, et al. DeepSeek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arxiv preprint arxiv:2501.12948, 2025
- [100] Guo, D Y, et al. DeepSeek-Coder: When the large language model meets programming The rise of code intelligence. arxiv preprint arxiv:2401.14196, 2024
- [101] Kirk H R, et al. The past, present and better future of feedback learning in large language models for subjective human preferences and values//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). Singapore, 2023: 2409-2430
- [102] Zhu K, Wang J, Zhou J, et al. PromptRobust: Towards evaluating the robustness of large language models on adversarial prompts//Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis. Copenhagen, Denmark, 2023: 57-68
- [103] Yang R, Shen X, Xu C, et al. PR-PFL: A privacy-preserving and robust personalized federated learning framework//Proceedings of the 2024 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). Guangzhou, China, 2024: 163-170
- [104] Fleisig E, Amstutz A, Atalla C, et al. FairPrism: Evaluating fairness-related harms in text generation//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada, 2023: 6231-6251
- [105] Wang Wen-Sheng, Tan Ning, Huang Kai, et al. A review of embodied intelligent systems based on large models. *Acta Automatica Sinica*, 2025, 51(1): 1-19(in Chinese)
(王文晟, 谭宁, 黄凯等. 基于大模型的具身智能系统综述. *自动化学报*, 2025, 51(1): 1-19)
- [106] Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report. arxiv preprint arxiv:2303.08774, 2023
- [107] Zhang Y F, Yu T, Tian H, et al. Mm-rlhf: The next step forward in multimodal llm alignment. arxiv preprint arxiv:2502.10391, 2025
- [108] Badshah S, Sajjad H. DAFE: LLM-based evaluation through dynamic arbitration for free-form question-answering. arxiv preprint arxiv:2503.08542, 2025
- [109] Zare M, Kebria P M, Khosravi A, et al. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*, 2024, 54(12): 7173-7186
- [110] Liang H, Dong Z, Ma Y, et al. A hierarchical imitation learning-based decision framework for autonomous driving// Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. Birmingham, UK, 2023: 4695-4701
- [111] Wu J, Huang Z, Huang C, et al. Human-in-the-loop deep reinforcement learning with application to autonomous driving. arxiv preprint arxiv:2104.07246, 2021
- [112] Yuan K, Huang Y, Yang S, et al. Evolutionary decision-making and planning for autonomous driving: A hybrid augmented intelligence framework. *IEEE Transactions on Intelligent Transportation Systems*, 2024, 25(7): 7339-7351
- [113] Fox R, Ludvig E A. Assimilating human feedback from autonomous vehicle interaction in reinforcement learning models. *Autonomous Agents and Multi-Agent Systems*, 2024, 38(2): 26
- [114] Sun Y, Salami Pargoo N, Jin P, et al. Optimizing autonomous driving for safety: A human-centric approach with LLM-enhanced RLHF// Proceedings of the 2024 ACM International Joint Conference on Pervasive and Ubiquitous Computing (Companion Volume). Melbourne, Australia, 2024: 76-80
- [115] Cao Y, Ivanovic B, Xiao C, et al. Reinforcement learning with human feedback for realistic traffic simulation//Proceedings of the 2024

- IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan, 2024: 14428-14434
- [116] Zhang Q L, Lin J Y, Sha Q X, et al. Deep interactive reinforcement learning for path following of autonomous underwater vehicle. *IEEE Access*, 2020, 8: 24258-2426
- [117] Pollak M, Salfinger, A and Hummel K A. Teaching drones on the fly: Can emotional feedback serve as learning signal for training artificial agents? *arxiv preprint arxiv:2202.09634*, 2022
- [118] Zitkovich B, Yu T, Xu S, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control//*Proceedings of the 2023 Conference on Robot Learning*. Atlanta, USA, 2023: 2165-2183
- [119] Bousmalis K, Vezzani G, Rao D, et al. RoboCat: A self-improving generalist agent for robotic manipulation. *arxiv preprint arxiv:2306.11706*, 2023
- [120] Liu Y, Han T, Ma S, et al. Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology*, 2023, 1(2): 100017
- [121] Glaese A, McAleese N, Trębacz M, et al. Improving alignment of dialogue agents via targeted human judgements. *arxiv preprint arxiv:2209.14375*, 2022
- [122] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 2022, 35: 27730-27744
- [123] Maroto-Gómez M, Malfaz M, Castillo J C, et al. Personalizing activity selection in assistive social robots from explicit and implicit user feedback. *International Journal of Social Robotics*, 2024, 17(10): 1999-2017
- [124] Wang R, Duan Y, Lam C T, et al. IvyGPT: Interactive Chinese pathway language model in medical domain//*Proceedings of the CAAI International Conference on Artificial Intelligence*. Singapore, 2023: 378-382
- [125] Lee K, Smith L M, Abbeel P. PEBBLE: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training//*Proceedings of the 38th International Conference on Machine Learning*. Virtual, 2021: 6152-6163
- [126] An G, Lee J, Zuo X, et al. Direct preference-based policy optimization without reward modeling//*Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*. New Orleans, USA, 2023: 70247-70266
- [127] Kim C, Park J, Shin J, et al. Preference transformer: Modeling human preferences using transformers for RL//*Proceedings of the 11th International Conference on Learning Representations (ICLR)*. Kigali, Rwanda, 2023
- [128] Liang X, Shu K, Lee K, et al. Reward uncertainty for exploration in preference-based reinforcement learning//*Proceedings of the 10th International Conference on Learning Representations (ICLR)*. Virtual, 2022
- [129] Park J, Seo Y, Shin J, et al. SURF: Semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning//*Proceedings of the 10th International Conference on Learning Representations (ICLR)*. Virtual, 2022
- [130] Akalin N, Loutfi A. Reinforcement learning approaches in social robotics. *Sensors*, 2021, 21(4): 1292
- [131] Yang Y H, Bhatt N P, Ingebrand T, et al. Fine-tuning language models using formal methods feedback: An use case in autonomous systems//*Proceedings of the 6th Conference on Machine Learning and Systems (MLSys)*. San Jose, USA, 2024: 339-350
- [132] Yuan K, Huang Y J, Guo L L, et al. Human feedback enhanced autonomous intelligent systems: A perspective from intelligent driving. *Autonomous Intelligent Systems*, 2024, 4(1): 1-10
- [133] Hazra R, Sygkounas A, et al. Revolve: Reward evolution with large language models for autonomous driving. *arxiv preprint arxiv:2406.01309*, 2024
- [134] Richter D J, Calix R A, Kim K. A review of reinforcement learning for fixed-wing aircraft control tasks. *IEEE Access*, 2024, 12: 103026-103048
- [135] Krening S, Feigh K M. Interaction algorithm effect on human experience with reinforcement learning. *ACM Transactions on Human-Robot Interaction*, 2018, 7(2): 1-22
- [136] Krening S, Feigh K. Newtonian action advice: Integrating human verbal instruction with reinforcement learning//*Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Montreal, Canada, 2019: 720-727
- [137] Krening S, Feigh K M. Effect of interaction design on the human experience with interactive reinforcement learning//*Proceedings of the 2019 ACM Conference on Designing Interactive Systems*. San Diego, USA, 2019: 1089-1100
- [138] Sudhakaran S, González-Duque M, Freiberger M, et al. MarioGPT: Open-ended text2level generation through large language models. *Advances in Neural Information Processing Systems*, 2023: 54213-54227
- [139] Warnell G, Waytowich N, Lawhern V, et al. Deep TAMER: Interactive agent shaping in high-dimensional state spaces//*Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. New Orleans, USA, 2018: 1545-1553
- [140] Atuhurra J. Large language models for human-robot interaction: Opportunities and risks. *arxiv preprint arxiv:2405.00693*, 2024
- [141] Poudel R, Pandya H, Zhang C, et al. Langwm: Language grounded world model. *arxiv preprint arxiv:2311.17593*, 2023
- [142] Li C L, Zeng S L, Liao Z Y, et al. Joint demonstration and preference learning improves policy alignment with human feedback. *arxiv preprint arxiv:2406.06874*, 2024
- [143] Bukharin A, Zhao T, Wang Z, et al. Robust reinforcement learning from corrupted human feedback. *Advances in Neural Information Processing Systems*, 2024, 37: 124093-124113
- [144] Dosovitskiy A, Ros G, Codevilla F, et al. CARLA: An open urban driving simulator//*Proceedings of the 1st Conference on Robot Learning*. Mountain View, USA, 2017: 1-16
- [145] Arumugam D, Lee J, Saskin S, et al. Deep reinforcement learning from policy-dependent human feedback. *arxiv preprint arxiv:1902.04257*, 2019



LIU Qi-Dong, Ph. D., professor. His current research interests include collective intelligence system, embodied intelligence, and collaboration between large and small models, etc.

HE Wen-Xuan, M. S. His current research interests include reinforcement learning, imitation learning, etc.

YAO En-Guang, Ph. D. candidate. His current research interests include reinforcement learning, imitation learning, etc.

CHEN Dong, Ph. D., lecturer. His current research interests include collective intelligence system, embodied intelligence,

and collaboration between large and small models, etc.

LI Ya-Fei, Ph. D., professor. His current research interests include collective intelligence system, embodied intelligence, and collaboration between large and small models, etc.

Background

Reinforcement Learning from Human Feedback (RLHF) has become a key paradigm for aligning reinforcement learning (RL) agents with complex human preferences. Instead of relying on predefined reward signals, RLHF incorporates human evaluations—such as comparisons, preferences, or suggestions—into the learning loop, either to train a proxy reward model or to directly adjust policy updates. This enables agents to better capture subtle human intentions and values. However, RLHF is inherently limited by the high cost, subjectivity, and scarcity of human-labeled data.

To address these limitations, Reinforcement Learning from AI Feedback (RLAIF) has emerged as a scalable alternative, using large pretrained models to generate synthetic feedback that can approximate human supervision. While RLAIF significantly improves efficiency, it also introduces concerns about consistency, alignment drift, and evaluation reliability. As these diverse feedback mechanisms proliferate, there is a growing need for a unified framework to understand and organize them systematically.

This survey introduces Reinforcement Learning from X-Feedback (RLXF), a generalized framework that encompasses feedback from human, AI, or hybrid sources. We categorize existing approaches into three representative paradigms: imitation learning, RLHF, and RLAIF, and

XU Ming-Liang, Ph. D., professor. His current research interests include embodied intelligence, virtual reality, and industrial software, etc.

compare their modeling assumptions, optimization objectives, and data requirements. We further explore the transformative applications of RLXF across several domains: such as autonomous driving, embodied intelligence, and large language model alignment—domains where high-quality feedback is crucial for safe and reliable decision-making.

Finally, we outline open challenges including how to arbitrate between conflicting feedback sources, improve sample efficiency, and ensure robustness across feedback modalities. We conclude by discussing future directions for building scalable, feedback-driven RL systems that adapt reliably to complex environments.

This work was supported in part by the National Science Foundation for Distinguished Young Scholars of China (62325602), in part by the National Natural Science Foundation of China (62276238), in part by the Joint Funds of the National Natural Science Foundation of China (U24A20326), in part by the State Key Programs of National Natural Science Foundation of China (62036010), in part by the Provincial Natural Science Foundation of Henan (232300421095), and in part by the Henan Provincial Department of Education University Science and Technology Innovation Talent Support Plan (25HASTIT034).