基于心智理论推理的多智能体逆向强化学习

Haochen Wu 密歇根大学机械工程系 美国密歇根州安娜堡
haochenw@umich.edu

Pedro Sequeira SRI International 人工智能中心 美国加利福尼亚州门洛帕克
pedro.sequeria@sri.com

David V. Pynadath 南加州大学创意技术研究所 美国加利福尼亚州洛杉矶
pynadath@usc.edu

arXiv:2302.10238v2 [cs.AI] 1 Mar 2023

摘要

我们通过多智能体逆向强化学习（MIRL），研究如何理解人们在协作环境中的互动，尤其是在个体对队友知之甚少时。其目标是根据团队在执行某项任务时的行为轨迹，推断引导每个个体行为的奖励函数。与当前的 MIRL 方法不同，我们不假设团队成员事先了解彼此的目标；相反，我们认为，他们在共同执行任务的同时，通过观察他人的行为来感知其目标，并据此调整自身行为、与他人协作。为解决这一问题，我们提出一种基于心智理论的全新 MIRL 方法（MIRL-ToM）。对于每个智能体，我们首先运用心智理论推理，根据其展示的行为估计其基线奖励档案的后验分布。随后，我们采用分散式均衡执行 MIRL，通过单智能体最大熵逆向强化学习推断每个智能体的奖励函数，同时根据档案的时变分布模拟其他队友的行为。我们在一个模拟的双人搜救行动中评估该方法；在此行动中，扮演不同角色的智能体要在环境中搜索并疏散受害者。结果表明，基线画像的选择对恢复真实奖励至关重要；此外，MIRL-ToM 能够恢复智能体与已知及未知队友互动时所采用的奖励。

关键词

逆强化学习；多智能体系统；心智理论；分散式均衡；合作

ACM 引用格式：
Haochen Wu、Pedro Sequeira 和 David V. Pynadath。2023年。基于心智理论推理的多智能体逆强化学习。载于第*22*届自主智能体与多智能体系统国际会议论文集（*AAMAS 2023*），英国伦敦，*2023*年*5*月*29*日至*6*月*2*日，IFAAMAS，共9页。

1 引言

理解人类在协作团队情境中彼此互动时所采用的底层策略，是构建相关系统的重要组成部分。这些系统需要能够准确预测团队行为、理解协作动态，并提供帮助以提升绩效。这一点尤其具有挑战性当人们首次合作时，这尤其困难，因为他们对他人的目标、策略和意图知之甚少，却必须在执行协作任务的过程中彼此适应。

预测个体将采取何种行动的一种方法，是利用机器学习技术，根据其行为轨迹构建其决策过程的模型。 f

其决策过程。某种方法是直接从观察中学习策略，即从状态到动作的映射。这包括模仿学习技术，例如行为克隆 [12]，它能够在无需进行环境和动力学建模的情况下复现他人的行为 [22]。然而，这类方法缺乏理解个体潜在

-

意图或任务目标的能力。此外，在团队合作场景中，仅了解个体的策略，很难理解他们如何协调行为 [25], ，也难以判断合作是否受到阻碍，例如团队成员之间的偏好不一致。

另一种更稳健的方法是假设每个个体的行为都是强化学习（RL）过程的结果：智能体通过与动态且不确定的环境进行试错交互，学习如何完成任务 [20]。强化学习智能体的目标由奖励函数指定，该函数编码了其偏好，而智能体希望在其生命周期内最大化该函数的输出。因此，通过观察个体行为来理解其内在策略，需要恢复在任务执行过程中引导其行为的奖励函数。这个问题称为逆强化学习（IRL） [3, 11]，其任务是根据个体的行为轨迹推断该个体所使用的奖励函数；该个体称为专家。与模仿学习方法相比，逆强化学习的优势在于，奖励函数是对专家目标最简洁、最稳健的表征，并能够预测其在新情境或不同任务中的行为（泛化） [11]。

在逆向强化学习中，学徒学习 [1] 建立了一种线性规划方法，用于使学得的奖励模型与从专家示范中的状态-动作对推导出的特征计数相匹配。博弈论方法 [21] 也被提出，用于在没有专家的情况下，通过参与重复博弈来学习。最大间隔逆向强化学习 [18] 将其表述为一个二次规划问题，旨在寻找一种奖励方案，使其产生的行为路径比其他路径更接近专家的行为，且接近程度的间隔随路径差异损失而变化。上述方法所要解决的逆向强化学习的主要难题之一是，该问题定义不适定——同一种观察到的行为可以由多个奖励函数解释，而同一个奖励函数在随机动力学 [3]下也可能导致不同的行为。为消除这种歧义，许多逆向强化学习

第*22*届自主智能体与多智能体系统国际会议论文集，*A. Ricci*、*W. Yeoh*、*N. Agmon*、*B. An*（编），*2023*年*5*月*29*日至*6*月*2*日，英国伦敦。© 2023年自主智能体与多智能体系统国际基金会（www.ifaamas.org）。保留所有权利。

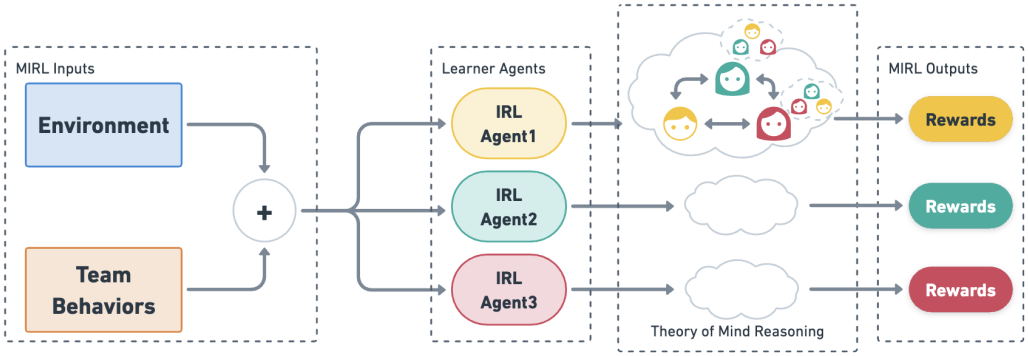

图 1：基于心智理论推理的多智能体逆强化学习（MIRL-ToM）方法概览。给定环境以及人类团队的示范行为，每个学习智能体通过心智理论推理其他智能体的行为，执行单智能体逆强化学习，并输出由此得到的奖励函数。

算法通过概率形式化来寻找使示范轨迹似然度最大化的奖励函数。具体而言，最大熵（MaxEnt）逆向强化学习 [27]会寻找一种轨迹分布，使专家示范获得指数级偏好。贝叶斯逆向强化学习 [17]提出了一种高效的采样方法，用于推断奖励函数参数的后验分布。

要理解协作任务中团队行为背后的策略和意图，需要将逆强化学习扩展到多智能体场景，这称为多智能体逆强化学习（MIRL）。在 MIRL 中，目标是根据团队（联合）行为的轨迹，推断指导每个智能体行为的个体奖励函数。人们通常假设联合行为是某种均衡解概念的结果，例如纳什均衡 [5, 7]。这使 MIRL 问题更难，因为除了示范所对应的多个合理奖励函数之外，多个个体奖励组合还可能存在多个均衡解 [5, 9]。

许多现有的多智能体逆向强化学习方法假设存在一个集中式学习者，它能够根据每个个体已知的奖励函数计算联合行动均衡 [9, 19, 23]。这意味着所有个体始终都完全了解彼此的偏好（任务收益/奖励函数），并据此按照均衡解采取行动。然而，假设团队中的个体完全了解他人，或许过于严格。首先，个体可能正在与陌生队友互动，例如，所观察到的行为可能是由一群被招募来执行协作任务的个体产生的，他们在执行任务时试图适应彼此的行为。其次，由观察队友行为而形成的、关于队友目标和意图的假设，可能会因人类行为的嘈杂性而出现错误。最后，由于个体会根据团队成员的行为调整自身行为（非平稳性），他们必须在互动过程中不断更新对彼此偏好的假设。所有这些因素都使得团队所展现的行为不太可能源自一个唯一的、稳定的且事先已知的均衡。

为了解决上述问题，我们借鉴了认知科学中的心智理论（ToM）概念 [24]，该理论认为

心理状态可用于解释和预测他人的动作。具体而言，我们将每个个体建模为通过为团队中的其他每位成员维持一个基线行为特征的概率分布，来适应队友的行为。随着任务的推进以及团队成员之间发生交互，该概率分布也会不断更新。从计算角度看，这些特征对应于任务领域特定的奖励模型，其中编码了典型偏好。随后，我们将多智能体逆向强化学习分为两个阶段（见图 1 概览）：

(1) 模型推断： 对于团队中的每个个体，我们利用心智理论推理，根据轨迹中包含的队友行为，估计其他每位队友的基线画像后验分布——这模拟了每位团队成员在任务执行过程中观察彼此行为后，可能如何建立对彼此的模型。

(2) 去中心化多智能体逆向强化学习： 我们采用去中心化均衡方法进行多智能体逆强化学习，将问题分解为多个单智能体逆向强化学习问题，类似于 [5] 中的方法。我们通过最大熵逆向强化学习，根据相应个体的演示行为推断奖励函数；同时，根据第一阶段计算出的基线画像时变分布，模拟其他团队成员的行为——这模拟了以他人感知模型为条件的最佳响应策略。

[8] 中提出了逆向强化学习与心智理论之间的类比，以关联人类的心理状态与逆向强化学习智能体的奖励模型 [14]。尽管心智理论已应用于多智能体强化学习 [16]，但针对多智能体逆向强化学习场景中团队潜在策略行为及其动机的研究仍然有限。本文的主要贡献是提出一种采用心智理论推理的去中心化多智能体逆向强化学习框架（MIRL-ToM），将最大熵逆向强化学习扩展至多智能体场景。我们的框架能够在假设存在一组基准奖励配置的前提下，对任意未知队友之间的交互进行建模。我们通过在贝叶斯框架下对奖励档案进行心智理论推理，处理其他队友意图的不确定性。

我们在一个模拟的双人搜索与救援任务中评估了我们的方法。在该任务中，每个智能体都扮演不同的角色，并具备一组独特的技能。我们为每个智能体设计了奖励，其中既包含任务相关方面，也包含社会相关方面，从而允许不同类型的

协调策略自然涌现。我们评估了 MIRL-ToM 方法在给定不同基线画像集合的情况下，恢复智能体原始奖励函数的能力。总体而言，结果表明，第一阶段中基线画像的选择对于第二阶段恢复真实奖励至关重要。特别是，纳入与团队成员相反目标相对应的基线画像会阻止最大熵逆向强化学习收敛，而纳入目标与原始队友相近的基线画像，则会得到与真实值相似的奖励函数。此外，我们还表明，我们的方法通过对示范行为进行心智理论推理，可以恢复智能体在与已知和未知队友交互时所使用的奖励，这证明了我们方法的稳健性。

2 相关工作

将逆向强化学习推广到多智能体系统中颇具挑战，因为我们的目标是学习能够解释个体奖励函数：既解释每个智能体的行为方式——这涉及通过逆向强化学习恢复奖励；也解释智能体如何协同——这需要将联合行为建模为均衡概念的解。在多智能体逆向强化学习领域，[10] 提出了一种算法，用于训练集中式智能体，使其优化交通路由问题中的行为协调；然而，在通过逆向强化学习学习奖励时，智能体会忽略队友，并将队友视为环境的一部分。在 [9],中，研究者通过在似然函数下优化奖励，对零和双人博弈进行贝叶斯推断，该似然函数编码了极小极大均衡的概念，但此方法无法扩展到 n人博弈和复杂领域。此外，许多多智能体逆向强化学习方法依赖于在每次迭代时，根据个体奖励函数计算某种均衡。针对竞争性博弈，一些方法通过学习马尔可夫完美均衡策略来处理，即基于个体奖励 [19], , 让智能体通过对抗训练与纳什均衡策略对弈 [23],; 另一些方法则通过从示范中提取效用函数来评估不良策略行为 [5],, 并利用生成对抗网络 [4] , 在高维领域中恢复奖励函数[26]。然而，这些方法在学习过程中仍需要寻找均衡策略，并假设专家完全了解队友的目标和意图，忽略了交互的动态性以及初次相遇的可能性。本文提出的 MIRL-ToM 将心智理论推理与单智能体最大熵逆向强化学习相结合，不仅能够使模型推理和奖励学习过程去中心化，还能考虑队友情况未知的情形，并通过对一组基准奖励配置进行信念更新来实现适应。

3 背景

3.1 强化学习

强化学习（RL） [20] 依赖于（单智能体）马尔可夫决策过程（MDP）的形式化表示，该过程定义为由 S, A, T, R, γ组成的元组，其中 $s \in S$是有限状态集合， $a \in A$是有限动作集合， $T := Pr(s\prime, a, s)$表示在给定动作时状态之间的转移概率， $R := R(s, a)$ 表示到达某个状态并执行动作所获得的奖励，而 $\gamma \in [0, 1)$ 是折扣因子，用于鼓励在更早的步骤获得更大的奖励。根据贝尔曼方程，给定一个状态满足 $Q^*(s, a) = R(s, a) + \gamma \mathbb{E}_{s' \sim T}[maxQ^*(s\prime a\prime)]$。（前向）强化学习算法寻找最优策略 $\pi : S \rightarrow A$，该策略可通过 $\pi^* = argmax_a Q^*(s, a)$ 获取。

3.2 逆向强化学习

逆向强化学习 (IRL) [11] 表示强化学习的逆问题，即在给定 S、A、 T 以及遵循专家内部策略的一组示范轨迹 $\mathrm{D} = \{\zeta_0 \ldots \zeta_i \ldots\}$ 的情况下，求出学习智能体的奖励函数 R。每条轨迹 $\zeta_i = \{(s_0, a_0) \ldots (s_j, a_j) \ldots (s_t, a_t)\}$ 由 t 步中的状态-动作对组成。逆向强化学习旨在找到能使示范的似然最大化的奖励函数参数。最大熵逆向强化学习 [27] （MaxEnt IRL）利用最大熵原理，并假设给定状态时，示范动作相较于其他动作具有指数级偏好。最大熵逆向强化学习将轨迹奖励 $R(\zeta)$ 视为状态特征计数的线性组合（即 $R(\zeta) = \theta^T \phi(\zeta)$），并通过迭代求取参数 $\theta^* = \arg\max_\theta \mathbb{E}_{\zeta \sim Pr(\zeta|\theta)} [\sum_j r(s_j, a_j|\theta)]$，以最佳复现示范行为，使得采取轨迹 ζ 的概率与路径上奖励的指数成正比（即 $Pr(\zeta|\theta) \propto e^{R}(\zeta)$）。

3.3 多智能体逆向强化学习

多智能体逆向强化学习（MIRL）将逆向强化学习扩展到多智能体场景。收集到的团队轨迹记为 $\mathrm{D} = \{\zeta_0 \ldots \zeta_i \ldots\}$，其中 $\zeta_i = \{(s_0, \hat{}a_0) \ldots (s_j, \hat{}a_j) \ldots (s_t, \hat{}a_t)\}$ 表示每条团队轨迹，其形式为状态-联合动作对 $(s_j, \hat{}a_j)$ ，共持续 t 步，其中 $\hat{}a_j^k$ 表示智能体 k 在第 j 步采取的动作。随后，MIRL 旨在为所有智能体恢复一组奖励函数参数 $\Theta^* = \{\theta^*\} = \arg\max_\Theta \mathbb{E}_{\zeta \sim Pr(\zeta|\Theta)} [\sum_j r(s_j, \hat{}a_j|\Theta)]$，使其与收集到的团队轨迹最为匹配。

3.4 PsychSim

我们使用 PsychSim [13],这一框架实现了我们的方法。该框架通过递归心智理论（ToM）对决策理论智能体进行建模，以用于交互式社会模拟 [13]。为此，PsychSim 在多智能体部分可观测马尔可夫决策过程（MPOMDP）形式体系下对世界动态进行建模 [2, 15]。在每一步中，每个智能体都会根据对所有可用动作计算出的价值来选择动作；这些价值是在决策中权衡风险与奖励时计算得出的，决策规划时域设定为指定长度（累计折扣奖励），并以其他智能体的动作为条件。具体而言，在规划期间，每个智能体使用递归 ToM 计算所有其他智能体的动作价值，并将其他智能体视为其所构建的（心理）模型中的奖励最大化者。为对不确定性进行建模，智能体可以对环境状态以及其他智能体的模型形成信念，并分别依据对状态特征的观测和其他智能体的动作，通过贝叶斯推断更新这些信念。

4 基于心智理论的多智能体逆强化学习

我们假设团队中的个体在没有明确沟通的情况下协调彼此的行为。在这种情况下，他们需要推理

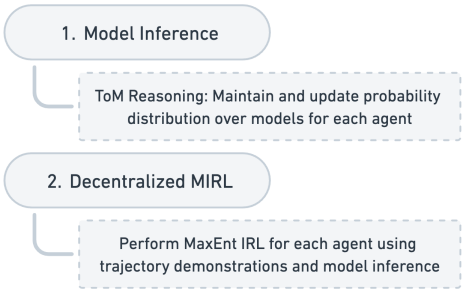

图 2：通过心智理论推理进行多智能体逆向强化学习的两个阶段

他人的目标和意图，同时计算自身的动作。我们还假设存在一组基准智能体档案，用于编码与特定任务领域相关的假定偏好。例如，在搜救任务中，这可能对应于一些档案，分别表示尝试搜索受害者、对其进行分诊、呼叫增援等行为。在任务开始前，个体可能会假设队友的行为遵循基于基准档案的先验分布，例如均匀分布，或根据背景信息对某个档案赋予偏好。随着任务的推进，个体观察到他人的行为后，会相应地更新这一分布。我们提出了基于心智理论推理的多智能体逆向强化学习（MIRL-ToM），使智能体能够在执行任务时推理他人的行为；这对应于一种去中心化方法，用于学习每位团队成员的奖励函数。该方法包含两个阶段，如图 2 所示，具体细节将在以下各节中讨论。

4.1 模型推断

在模型推断阶段，我们利用心智理论推理，根据轨迹中智能体的行为，更新基线画像上的概率分布（信念）。推断过程开始前，我们为每个智能体预先选定一组基线画像。一般来说，这些画像可以编码任意智能体决策模型，我们用 $\{m_i\}$ 表示初始画像集。本文中，我们将画像视为预定义的奖励函数，即奖励档案。设 $b^k(m)$表示智能体对另一智能体 k 的模型所持有的先验分布。对于每条轨迹 ζ_i 的每个步骤，$b^k(m)$通过贝叶斯推断进行更新：

$$b'(m) \propto Pr(a|m,\zeta) \times b(m|\zeta) \tag{1}$$

为方便起见，我们省略智能体索引 k。此阶段$\hat{}$ $\hat{}$通过在每个步骤添加智能体对智能体模型的信念，扩充团队轨迹，$\mathrm{D} = \{\zeta_i\}$，从而得到一组状态—联合动作—模型分布元组序列：$\hat{}$ $\zeta_i = \{(s_j, \hat{}a_j, b_j(m)\}$。概括来说，此阶段描述了在给定一组智能体档案和先验分布的条件下，如何根据轨迹中的动作对智能体的意图进行建模。[1]

[1]需要指出的是，初始画像集和先验分布不必对所有观察智能体都相同。同样，不同智能体观察同一队友时，信念更新过程也可以不同，例如，将可观测范围限制为智能体之间的局部交互。

4.2 去中心化多智能体逆向强化学习

在每个学习智能体计算出关于其队友预期奖励模型的（随时间变化的）信念后，我们通过分别为每个智能体计算奖励函数，执行分散式多智能体逆向强化学习（MIRL）。不失一般性，这里假设奖励函数 $R(s, a) := \theta^T \phi(s, \hat{}a)$是（可能为非线性的）奖励特征 $\phi(s, \hat{}a)$的线性组合，并由奖励权重 θ参数化。我们遵循最大熵原理，寻找使智能体的示范轨迹具有指数级偏好的个体奖励函数，即：

$$\theta^* = \arg\max_\theta L(\theta|b(m)) = \arg\max_\theta \sum_{\zeta_i \in \mathcal{D}} \log Pr(\zeta_i|\theta, b(m)) \tag{2}$$

我们提出一种算法，通过去中心化均衡计算将最大熵逆向强化学习 [27] 扩展到多智能体逆向强化学习场景。在每次迭代中，如算法1所示，我们根据当前奖励权重以及模型推断阶段计算出的智能体模型信念，近似智能体路径上的分布。为便于说明，算法1展示的是团队仅包含两个智能体的情形，即只需计算一个模型分布 $b(m)$。为解决多智能体逆向强化学习问题，我们通过迭代改进，为每个智能体 k 学习一个奖励权重向量 θ^k；改进依据是该权重向量所产生的行为是否与相应专家在示范轨迹中表现出的行为相似，具体通过比较估计的特征计数（FC）与经验特征计数来衡量（算法1，第5行）。因此，在估计特征计数的过程中，我们令每个智能体的奖励函数成为针对其感知到的其他智能体模型的最佳响应策略的解，从而实现分散式均衡。值得注意的是，这种方法的优势在于，每个智能体的学习过程都可以完全分布式地并行进行。在每次迭代中，我们假设存在一个正向强化学习求解器：给定当前的奖励权重候选向量 θ^k，该求解器计算策略 π^k；在每个时间步，其他智能体的动作则根据模型分布 $b(m)$进行优化（算法1，第8行）。在梯度下降步骤中（算法1，第9–11行），估计的

算法 1 基于心智理论的去中心化多智能体逆向强化学习

1: $\mathrm{D} \leftarrow$团队轨迹集合
2: **对**每个学习智能体 k **执行**
3: $b_0(m)$：先验模型分布，$\phi(s, a)$：特征向量
4: $\hat{}\hat{}\mathrm{D} = \{\zeta_i\} \leftarrow$ 使用心智理论推理对 D 进行模型推断
5: $\phi^k_{emp} \leftarrow \sum_{\zeta_i \in \mathrm{D}} \sum_{(s_j, \hat{}a^k_j) \in \zeta_i} \phi(s_j, \hat{}a^k_j)$
6: 初始化奖励向量权重 θ^k
7: **当** θ^k 尚未收敛 **时**
8: $\pi^k(a|s, b(m), \theta^k) \leftarrow$前向传播
9: $\phi^k_{est} \leftarrow \sum_{\zeta_i \sim \pi} \sum_{(s_j, \hat{}a^k_j | b_j(m)) \in \zeta_i} Pr(s_j|\pi^k, \theta^k)\phi(s_j, \hat{}a^k_j)$
10: $\nabla L(\theta^k) \leftarrow \phi^k_{est} - \phi^k_{emp}$
11: $\theta^k \leftarrow \theta^k - \alpha \nabla L(\theta^k)$
12: **结束 while 循环**
13: **结束 for 循环**
14: **返回** $\{\theta^k\}$

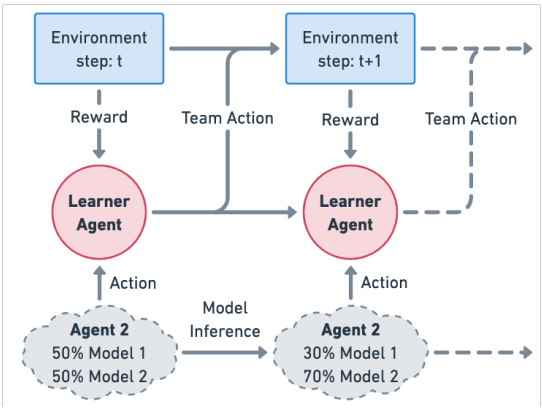

图 3：特征计数估计的前向传播。每个学习智能体利用模型分布数据选择其他智能体的动作，并根据当前奖励模型选择一个动作。特征计数根据环境状态和团队动作计算得出。

FC ϕ_{est} 通过使用由此得到的联合策略对蒙特卡洛轨迹进行采样来计算。这个过程如图3所示。

5 结果

为了评估所提出的 MIRL-ToM 方法的有效性，我们考虑一个协作环境和一个由两个智能体组成的团队。我们不使用真实的人类数据来评估该方法，因为无法获取其底层奖励函数；在这里，我们指定一组真实奖励，让智能体根据某种预先定义的他者模型最大化这些奖励，从而收集轨迹，然后将学习得到的奖励函数与每个智能体的真实奖励函数进行比较。在使用人类数据进行测试之前，这是评估我们方法能力（及其局限性）的必要第一步。[2]

5.1 实验

5.1.1 环境。我们使用 PsychSim [13]设计了一个协作式多智能体搜索与救援（S&R）行动，该系统支持通过智能体建模实现部分观测、随机状态-动作动态和心智理论推理。在我们的搜救任务中，两个智能体需要相互合作，以尽快搜索网格世界环境中的所有位置并疏散所有受害者。起初，智能体不知道受害者的位置；对此，我们通过对每个位置存在受害者的情况设定均匀信念进行建模，当智能体进入相应网格后，再将其设为实际值。环境的状态 $s \in S := S_1^P \times \cdots \times S_n^P$描述了一个包含 n 个位置的网格世界，其中每个位置都有一个特征 $v \in V := \{unknown, found, ready, clear, empty\}$，表示该位置的受害者状态。动作集合包括三类动作：1）移动动作：$\{up, down, left, right, wait\}$，以确定性方式将智能体移动到相应方向；2）受害者处置动作：$\{search, triage, evacuate\}$，用于改变受害者状态特征；以及 3）通信动作，$call$，该动作

[2]我们的实现和实验可在 https://github.com/usc- psychsim/mirl-tom-aamas23 获取。

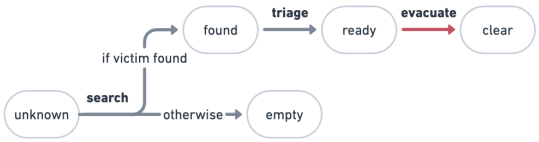

图 4：受害者状态的转移动态。*search* 要么找到受害者，要么找到空单元格，*triage* 准备受害者，并要求两个智能体位于同一位置才能 *evacuate* 它。

允许智能体激励其他智能体来到其所在位置。受害者状态的状态转移是确定性的，其动态过程如图 4 所示。正如红色箭头所示，为了建模合作的必要性，我们要求两个智能体位于同一位置，并同时执行动作 *evacuate*，受害者才能被疏散。

表1：两个智能体的真实奖励函数，是由特征的线性组合构成，并由 ϕ_i 加权 θ_i。

角色相关任务 社交/帮助

$\phi_i(s, a)$	θ_i	描述
Dist2Vic	0.06	到最近 *found* 受害者的距离
搜索	0.06	是否搜索 *unknown* 地点
分诊	0.19	是否对 *found* 受害者进行分诊
疏散	0.63	是否疏散 *ready* 受害者
Wait	0.03	是否待命
Call	0.03	是否需要时呼救

(a) Medic

$\phi_i(s, a)$	θ_i	描述
Dist2Help	0.25	与执行 *call* 动作的智能体之间的距离
搜索	0.25	是否搜索 *unknown* 地点
疏散	0.50	是否疏散 *ready* 受害者

(b) 探索者

5.1.2 真实奖励智能体（专家）。我们考虑由两个具有指定角色的智能体组成的团队——Medic 和 Explorer，每个智能体定义了不同的动作集合和奖励函数。Medic 智能体拥有完整的动作集合，而 Explorer 智能体无法 *wait*、*triage*或 *call*，即它只能探索环境。每个智能体的奖励函数 $R(s, a) := \sum_i \theta_i \phi_i(s, a)$ 由真实权重 θ进行线性参数化，如表1所定义。距离特征计算智能体到目标位置的曼哈顿距离。动作相关特征是指示函数，用于表示当前状态下是否采取了某个动作。如表1所示，奖励特征分为与角色相关的任务特征（橙色）和社会特征（蓝色）。我们将 Medic 建模为优先救援受害者，而 Explorer 则优先进行搜索，同时也会响应 Medic 的呼救。

5.2 收集和可视化轨迹

根据表1中定义的真实奖励函数，我们首先考虑这样一种情形：两个智能体彼此交互，并在确切了解对方奖励函数的情况下执行任务（完美模型）。尽管如此，我们注意到，MIRL-ToM 流程并不知道这一点，它只获得生成的团队

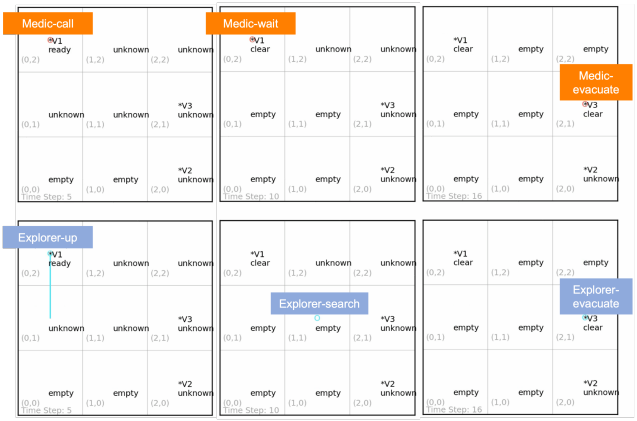

图 5：Medic（上）和探索者（下）智能体在时间步 $t = 4$（左）、$t = 10$（中）和 $t = 16$（右）处的示例轨迹。

轨迹和一组基线画像。我们的目标是检验不同画像先验对恢复原始奖励的重要性。我们收集了智能体团队的16条轨迹，每条包含25个步骤；这些轨迹由智能体使用软最大动作选择、以2为规划时域进行规划后生成。图5展示了一个示例轨迹中两个智能体在三个不同步骤的协作行为。在 $t = 4$处，Medic 呼叫求助，探索者便朝 Medic 移动，而不是搜索或离开；$t = 10$ 展示了角色依赖动作：由于所有受害者的状态要么未知，要么已确认安全，Medic 正在等待，而探索者继续搜索其他地点；在 $t = 16$处，两个智能体共同疏散受害者。[3]

5.3 模型推断结果

基于第 5.1.2 节中详细介绍的奖励特征 ϕ_i，我们通过对真实奖励以及任务导向和社会导向特征进行不同变体的组合，设计了一组基线画像，如表 2 所示。我们的目标是创建一组在搜救任务中人们通常会遵循的奖励函数。

根据这些奖励档案，我们定义了以下实验条件，即每种条件都定义了模型推断阶段所使用的智能体模型/档案集合（见第 5.3 节）：

表 2：基准奖励配置

	奖励函数	描述
Gt	θ_{gt}	真实奖励
Op	$-\theta_{gt}$	相反目标
Rd	θ_0	没有目标 / 随机行为
Tk	$[\theta_{task}, \theta_{social} = 0]$	关注任务特定方面
Sc	$[\theta_{task} = 0, \theta_{social}]$	关注社会方面

***cond*-1：真实目标，**始终相信另一个智能体使用真实目标奖励。 ***cond*-2：相反目标，**始终相信另一个智能体使用相反目标奖励。 ***cond*-3：真实目标-相反目标-随机行为，**通过心智理论更新信念 $b(\{Gt,\ Op,\ Rd\})$。 ***cond*-4：随机行为-任务导向-社会导向，**通过心智理论更新信念 $b(\{Rd,\ Tk,\ Sc\})$。

[3]此处提供了示例模拟动画，展示智能体在 S&R 环境中的交互，作为补充材料。

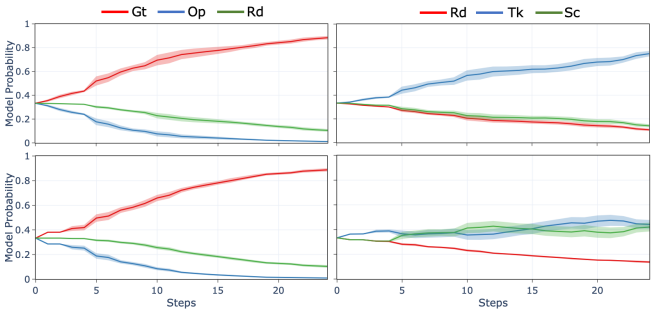

图 6：两种实验条件下，Medic（上）和 Explorer（下）智能体的 16 条轨迹的平均模型概率及标准差（阴影部分）：*cond*-3（Gt-Op-Rd，左）和 *cond*-4（Rd-Tk-Sc，右）。

条件 *cond*-1对应这样的设置：学习智能体在最大熵逆向强化学习过程中，使用其他智能体的真实奖励（完美模型）来计算估计的 FC。其目的是评估单智能体逆向强化学习中的最佳响应（分散式均衡）计算。 *cond*-2则描述了这样一种情况：学习智能体始终认为其队友的目标与真实值智能体的目标相反。我们在此评估智能体间相互依赖关系对任务的重要性，并检验在其他智能体的模型相互矛盾时，MIRL 恢复原始奖励的能力极限。对于前两个条件，模型分布是固定的，因此我们不通过心智理论推理进行模型推断。相比之下， *cond*-3 和 *cond*-4 均采用心智理论推理。 *cond*-3 用于验证模型推断过程能否根据观察到的行为准确识别真实值模型；而 *cond*-4 则考察在未提供 Gt 模型的情况下，MIRL-ToM 能否仍然恢复出接近原始奖励的奖励，例如分析真实人类数据时可能遇到的情况。

对于 *cond*-3 和 *cond*-4，我们使用 PsychSim 进行贝叶斯推断，根据式 1，利用收集到的 16 条轨迹中智能体在每一步的行为，计算智能体模型的后验分布。图 6 显示了各时间步的平均模型概率。在 *cond*-3（图 6 左）中，可以清楚地区分 Gt 模型与 Rd 模型，这意味着观察到的行为确实体现了不同的智能体策略（否则，随机行为也可能具有同样的概率）。此外，在轨迹结束时，Op 模型对应的概率非常低。对于 *cond*-4（图 6 右），Medic 智能体被识别为更偏向任务导向，这符合预期，因为它可以使用完整的动作集合。然而，对于 Explorer 智能体，Tk 模型和 Sc 模型更难区分，这意味着真实值行为被认为兼具同等程度的任务导向和社会导向。

5.4 去中心化多智能体逆向强化学习结果

我们在上述各项实验条件下分别应用了所提出的去中心化 MIRL-ToM 方法，初始奖励权重均匀（θ_0），逆向强化学习训练轮次最多为 30 轮，学习率（α）按指数衰减，从 0.9 开始。对于前向传播（算法1，第 8 行），我们在 PsychSim 中收集了 16 条蒙特卡洛轨迹，以估计 FC。

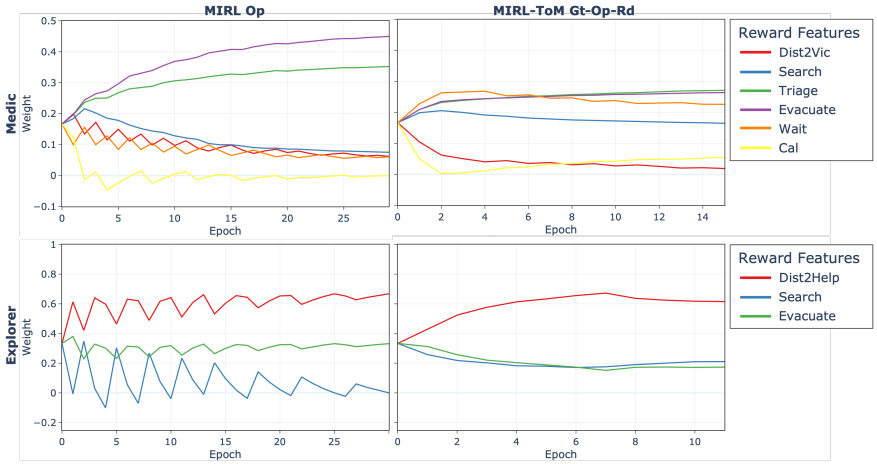

图7：在两种实验条件下，Medic（上）和探索者（下）在逆向强化学习过程中奖励权重的学习曲线：*cond*-2（左）和 *cond*-3（右）。

表3：已知队友设定下所有条件中两个智能体学习到的奖励函数。

特征	档案 θ_i		多智能体逆向强化学习 θ_i		MIRL-ToM θ_i	
$\phi_i(s,\ a)$	Gt	Op	*cond*-1	*cond*-2	*cond*-3	*cond*-4
Dist2Vic	0.06	-0.06	0.02	-	0.02	0.00
搜索	0.06	-0.06	0.18	-	0.16	0.16
分诊	0.19	-0.19	0.23	-	0.27	0.29
疏散	0.63	-0.63	0.25	-	0.26	0.28
Wait	0.03	-0.03	0.27	-	0.23	0.26
Call	0.03	-0.03	0.06	-	0.06	0.01

(a) Medic

特征	档案 θ_i		多智能体逆向强化学习 θ_i		MIRL-ToM θ_i	
$\phi_i(s,\ a)$	Gt	Op	*cond*-1	*cond*-2	*cond*-3	*cond*-4
Dist2Help	0.25	-0.25	0.01	-	0.62	0.26
搜索	0.25	-0.25	0.23	-	0.21	0.43
疏散	0.50	-0.50	0.76	-	0.17	0.30

(b) 探索者

通过 MIRL-ToM 针对各项条件推断出的奖励函数汇总于表3。我们观察到，使用心智理论推理得到的 *cond*-3 和 4 的奖励权重，与同时对两个智能体使用 Gt模型的 *cond*-1 的奖励权重非常接近。尤其对于 *cond*-4，我们发现，通过心智理论进行的模型推断足够准确地捕捉了队友的行为，从而能够恢复原始行为。通过查看学习曲线，即每次逆向强化学习迭代中奖励权重 θ_i 的变化，可以更好地理解该表中的部分结果。具体而言，图7展示了 *cond*-2 和 *cond*-3 的曲线。如图所示，除 *cond*-2 外，奖励权重大约在 15 个训练轮次内收敛至最终值；在 *cond*-2 中，智能体始终认为另一个智能体使用 Op模型。该条件下的奖励权重未能收敛（图7，左图），要么发生振荡，要么在最大迭代次数内未能收敛。[4]由于智能体使用了错误的队友模型，通过最佳响应计算得到的协调行为

[4]这些曲线看似正在收敛，是因为学习率降低了 α。

表 4：已知队友设定下所有条件中两个智能体恢复行为的相似度。

特征	档案		MIRL	MIRL-ToM	
$\phi_i(s,\ a)$	Gt	Op	*cond*-1	*cond*-3	*cond*-4
经验特征计数					
Dist2Vic	11.86	0.00	11.53	15.39	14.48
搜索	2.40	0.00	1.75	3.13	2.06
分诊	2.97	0.00	2.91	2.97	2.94
疏散	2.94	0.00	2.81	2.63	2.66
Wait	8.97	0.00	10.88	6.28	7.78
Call	4.78	0.00	2.72	4.38	4.44
相似度指标					
特征计数差异	2.38	30.55	4.52	**7.53**	**4.50**
π Div	0.00	0.25	0.13	**0.10**	**0.13**

(a) Medic

特征	档案		MIRL	MIRL-ToM	
$\phi_i(s,\ a)$	Gt	Op	*cond*-1	*cond*-3	*cond*-4
经验特征计数					
Dist2Help	6.50	0.00	2.41	6.45	4.24
搜索	6.59	0.00	7.06	6.34	7.38
疏散	2.94	0.00	2.81	2.63	2.66
相似度指标					
特征计数差异	2.25	14.84	3.06	**2.13**	**1.26**
π Div	0.00	0.32	0.02	**0.03**	**0.04**

(b)探索者

(b)探索者 智能体根据当前奖励和对方的模型分布学习到的策略（算法1，第8行），会导致特征计数与经验特征计数不一致（第9–10行），进而使下一次迭代中的奖励向量发生显著变化（第11行）。

如表3所示，通过逆向强化学习恢复出的奖励权重与真实奖励（Gt列）并不完全一致。不过，这些权重仍能帮助我们恢复某些特征对智能体行为的相对重要性。例如，救助受害者（包括等待 Explorer 撤离）对 Medic 智能体来说更为重要，而搜索并撤离受害者则

表 5：未知队友设定下两个智能体恢复行为的相似度。

特征	Demo	MIRL-ToM
$\phi_i(s,\ a)$	θ_{gt}	θ_{irl}
经验特征计数		
Dist2Vic	14.05	14.44
搜索	1.99	2.16
分诊	2.49	2.97
疏散	2.68	2.81
Wait	7.13	7.88
Call	5.88	5.25
相似度度量		
特征计数差异	-	2.55

(a) Medic

特征	Demo	MIRL-ToM
$\phi_i(s,\ a)$	θ_{gt}	θ_{irl}
经验特征计数		
Dist2Help	6.49	7.23
搜索	8.25	6.50
疏散	2.68	2.81
相似度度量		
特征计数差异	-	2.63

（**b**）**探索者** 是 Explorer 智能体的优先事项。如前所述，不同的奖励函数可能会产生相似的行为，因此，考察学习到的行为与专家行为之间相似性的更好方法是使用相似度指标。这里我们使用以下指标：

特征计数差异： 使用逆向强化学习奖励的策略与演示行为之间的特征计数差异：$\sum_i |\phi_i^{\theta} - \phi_{\mathcal{D}_i}|$。

π 散度： 真实策略与学习到的策略在所有轨迹的所有步骤上的动作分布之间的平均詹森-香农散度：$JSD(\pi(a_j|s_j;\ \theta)\ \pi(a_j|s_j;\ \theta_{gt}))$。

结果汇总于表 4。配置文件列对应于所达到的最佳（真实目标）和最差（相反目标）行为相似度。真实目标对应的特征计数差异表示使用相同奖励但不同随机种子采样策略时的预期误差（因此不为零）。随着模拟轨迹数量（目前为16）增加，特征计数差异将趋近于零。多智能体逆向强化学习列显示，特征计数和行为非常相似；这也是预期之中的，因为它对应于使用固定的队友真实目标模型（不使用心智理论）。至于 MIRL-ToM 列，结果表明我们的方法能够实现非常相似的行为（以粗体突出显示），这说明了模型推断阶段的重要性。有趣的是，在 *cond*-4中达到了更相似的行为，该结果未将真实目标奖励纳入模型分布，不过我们没有评估这一结果的统计显著性。我们还注意到，由于表3中观察到的奖励权重差异，π散度并未接近零，而且概率取决于奖励权重的大小。

5.5 未知队友设定

在本节中，我们考虑这样一组示范：它们由一个智能体团队收集而来，团队成员会对彼此的目标形成信念，并根据在 任务执行期间对彼此行为的观测来更新这些信念。我们再次使用 PsychSim 生成示范轨迹，不过这次专家智能体在三个奖励档案（真实目标 Gt、相反目标 Op 和随机行为 Rd）的均匀先验下，推断彼此的模型。我们通过先基于示范进行模型推断，再对每个智能体执行最大熵逆向强化学习，来应用 MIRL-ToM。在表 5 中，我们比较了根据示范计算的经验特征计数（Demo 列）——其中智能体使用真实奖励 θ_{gt}——与智能体使用 MIRL-ToM 得出的奖励权重 θ_{irl} 时产生的特征计数。

结果表明，即使示范是由对队友的模型并不完美的智能体生成的，我们的方法仍能恢复出相似的行为。通过执行模型推断，我们能够对个体“发现”队友行为并据此调整自身行为的过程进行建模。

6 结论

我们提出了一种利用心智理论推理进行多智能体逆强化学习的方法（MIRL-ToM）。与其他多智能体逆强化学习方法不同，我们的方法不假设智能体在执行任务以计算均衡策略时，完全了解队友的奖励函数。相反，在给定一组基准奖励配置的情况下，我们利用心智理论推理每个个体如何通过观察他人在整个任务中的行为来建立对他人意图的模型。随后，我们将多智能体逆强化学习分解为针对每个个体的单智能体最大熵逆强化学习，并根据推断出的档案分布，推断引导其行为的奖励函数。对队友心理状态和奖励模型的归因，不仅考虑了个体行为认知不完美以及对他人策略的不确定性，还能以分布式方式高效计算均衡策略。所提出的 MIRL-ToM 能够在已知队友和未知队友两种情况下，恢复出轨迹特征计数相似的行为。此外，尽管我们在双人协作场景中进行了实验，我们的框架也适用于 n人一般和博弈问题。

我们的初步实验使用合成专家示范数据，展示了 MIRL-ToM 的有效性和效能。自然而然的下一步是将该方法应用于从人类收集的数据，以深入了解哪些因素——任务相关、社交相关和情绪相关因素——对他们在协作环境中的行为影响最大。此外，我们将探索更具可解释性的奖励结构，不再依赖基特征之间的线性关系。我们还想探索创建基线画像的方法，因为 MIRL-ToM 的效果似乎在很大程度上取决于初始集合。一种选择是采用迭代方法：将每次迭代中学到的奖励函数添加到基线画像集合中，直至收敛。

致谢

本材料基于美国国防高级研究计划局（DARPA）通过合同编号 W911NF-20-1-0011 资助的工作。材料中表达的任何观点、研究结果、结论或建议均属于作者的意见，不一定反映美国国防高级研究计划局（DARPA）的观点。

参考文献

[1] Pieter Abbeel 和 Andrew Y. Ng。2004年。通过逆强化学习进行学徒学习。第二十一届国际机器学习会议论文集（2004年）。[2] Christopher Amato 和 Frans A. Oliehoek。2015年。面向多智能体部分可观测马尔可夫决策过程的可扩展规划与学习。第二十九届 AAAI人工智能会议论文集（2015年），1995–2002。[3] Saurabh Arora 和 Prashant Doshi。2021年。逆强化学习综述：挑战、方法与进展。《人工智能》297（2021年），103500。[4] Chelsea Finn、Paul F. Christiano、Pieter Abbeel 和 Sergey Levine。2016年。生成对抗网络、逆强化学习与基于能量的模型之间的联系。CoRR abs/1611.03852（2016年）。[5] Justin Fu、Andrea Tacchetti、Julien Perolat 和 Yoram Bachrach。2021年。评估多智能体逆向强化学习中的战略结构。《人工智能研究杂志》71（2021年9月），925–951。[6] B. Fuglede 和 F. Topsoe。2004年。詹森-香农散度与希尔伯特空间嵌入。载于《国际信息论研讨会论文集》，

2004 *2004* 31 37 10 1109 2004 1365067 。（SIT）。–。[7] 1998 https://doi.org/ 。/ISIT.。Junling Hu 和 Michael P. Wellman。多智能体强化学习：理论框架与一种算法。第十五届国际机器学习会议论文集（1998年），242–250。[8] Julian Jara-Ettinger。2019年。作为逆强化学习的心智理论。《行为科学当前观点》29（2019年），105–110。https://doi.org/10.1016/j.cobeha.2019.04.010 人工智能。[9] Xiaomin Lin、Peter A. Beling 和 Randy Cogill。2018年。双人零和博弈的多智能体逆向强化学习。《IEEE博弈汇刊》10，1（2018年），56–68。[10] Sriraam Natarajan、Gautam Kunapuli、Kshitij Judah、Prasad Tadepalli、Kristian Kersting 和 Jude Shavlik。2010年。多智能体逆向强化学习。2010年第九届机器学习与应用国际会议（2010年），395–400。[11] Andrew Y. Ng 和 Stuart J. Russell。2000年。逆向强化学习算法。第十七届国际机器学习会议论文集（2000年），663–670。[12] Dean A. Pomerleau。1991年。用于自主导航的人工神经网络高效训练。《神经计算》3，1（1991年），88–97。https://doi.org/10.1162/neco.1991.3.1.88[13] David V. Pynadath 和 Stacy C. Marsella。2005年。PsychSim：利用决策理论智能体建模心智理论。国际人工智能联合会议（2005年），1181–1186。[14] David V. Pynadath、Paul S. Rosenbloom 和 Stacy C. Marsella。2014年。Sigma认知架构中自适应心智理论的强化学习。《通用人工智能》（2014年），143–154。[15] David V. Pynadath 和 Milind Tambe。2002年。通信式多智能体团队决策问题：分析团队协作理论与模型。《人工智能研究杂志》16，1（2002年6月），389–423。[16] Neil Rabinowitz、Frank Perbet、Francis Song、Chiyuan Zhang、S. M. Ali Eslami 和 Matthew Botvinick。2018年。机器心智理论。第35届国际机器学习会议论文集80（2018年7月），4218–4227。[17] Deepak Ramachandran 和 Eyal Amir。2007年。贝叶斯逆向强化学习。第二十届国际人工智能联合会议论文集（2007年），2586–2591。[18] Nathan D. Ratliff、J. Andrew Bagnell 和 Martin A. Zinkevich。2006年。最大间隔规划。第二十三届国际机器学习会议论文集（2006年），729–736。[19] Tummalapalli Sudhamsh Reddy、Vamsikrishna Gopikrishna、Gergely Zaruba 和 Manfred Huber。2012年。去中心化非合作多智能体系统的逆强化学习。（2012年），1930–1935。[20] Richard S. Sutton 和 Andrew G. Barto。2018年。《强化学习：导论》。布拉德福德出版社，美国马萨诸塞州剑桥。[21] Umar Syed 和 Robert E Schapire。2007年。学徒学习的博弈论方法。《神经信息处理系统进展》20（2007年）。[22] Faraz Torabi、Garrett Warnell 和 Peter Stone。2018年。通过观察进行行为克隆。第二十七届国际人工智能联合会议论文集（2018年），4950–4957。[23] Xingyu Wang 和 Diego Klabjan。2018年。利用次优示范进行竞争性多智能体逆向强化学习。第三十五届国际机器学习会议论文集80（2018年），5143–5151。[24] Andrew Whiten。1991年。《自然心智理论：日常读心的进化、发展与模拟》。Blackwell。[25] Haochen Wu、Amin Ghadami、Alparslan Emrah Bayrak、Jonathon M. Smereka 和 Bogdan I. Epureanu。2023年。通过因果推断和子团队识别评估多智能体任务分配中的涌现式协调。《IEEE机器人与自动化快报》8

，2（2023 ，728–735。
年）

[26] Lantao Yu、Jiaming Song 和 Stefano Ermon。2019年。多智能体对抗逆强化学习。第*36*届国际机器学习会议论文集 (2019年)，7194–7201。[27] Brian D. Ziebart、Andrew Maas、J. Andrew Bagnell 和 Anind K. Dey。2008年。最大熵逆强化学习。*AAAI*人工智能会议 (2008年)，1433–1438。