arXiv:2302.05063v1 [cs.MA] 10 Feb 2023

基于策略相似度提升零样本协作性能

Lebin Yu、Yunbo Qiu、Quanming Yao、Xudong Zhang、Jian Wang

清华大学电子工程系，中国北京{yulb19,qyb18}@mails.tsinghua.edu.cn,
quanmingyao@gmail.com, {zhangxd, jianwang}tsinghua.edu.cn

摘要

近年来，多智能体强化学习在多智能体规划和调度任务中取得了显著成果。它通常采用自博弈设置，即通过与一组固定的智能体进行对战来训练智能体。然而，面对零样本协作时，智能体必须与未见过的伙伴进行协作，自博弈智能体可能会失败。已有多种方法尝试解决这一问题，但它们要么耗时过长，要么缺乏泛化能力。本文首先揭示了一个重要现象：零样本协作性能与智能体的训练伙伴和测试伙伴之间的相似度呈强线性相关。受此启发，我们提出基于相似度的鲁棒训练（SBRT）方案，根据预先定义的策略相似度值，在训练过程中扰动伙伴的动作，从而提升智能体的零样本协作性能。为验证该方案的有效性，我们将其应用于三种多智能体强化学习框架，取得了优于以往方法的性能。

引言

近年来，多智能体强化学习因其在自动化规划和调度任务中的广泛应用而日益受到关注，例如机器人导航（Li 等，2019）、交通控制（Calvo 和 Dusparic，2018；Ghosh 等，2021）以及车队管理（Lin 等，2018；Bøgh 等，2022）。与单智能体任务不同，在单智能体任务中，智能体只需与静态环境交互；而在多智能体场景中，智能体持续更新策略会带来不稳定性（Lowe 等，2017；Mahajan 等，2019）。为解决这一问题，研究人员让一组固定的智能体共同训练和测试，这种方法称为自博弈（Tesauro，1994）。自博弈使智能体能够迅速了解伙伴的策略，并调整自己的计划。然而，经过自博弈训练的智能体通常无法在测试期间与未见过的智能体良好协作（Hu 等，2020）。

为了深入研究协作智能体的泛化性能，研究人员提出了零样本协作的概念（Hu 等，2020）：智能体必须与此前从未见过的陌生伙伴进行协作。此后，研究人员提出了多种解决这一问题的方法，可分为两类。第一类是基于群体的训练（Shih 等，2021；Lucas 和 Allen，2022），即让智能体与多个不同的伙伴一起训练。这类方法能让智能体学到泛化能力更强的策略，但耗时较长。第二类是通过让智能体对任务进行推理（Hu 等人，2021；Cui 等人，2021），或打破对称性（Hu 等，2020；Treutlein 等人，2021），来避免自博弈智能体之间形成特定惯例。通过这类框架训练的智能体，能够与采用同一框架训练的未见过的伙伴很好地协作，但与采用其他框架训练的伙伴协作时表现不佳。

在这些论文中，研究者采用交叉对战来评估零样本协调表现，即要求分别训练的智能体彼此协作。然而，实验表明，即使不同组合中的智能体是在同一框架下使用不同随机数种子训练的，其协作表现也会有很大差异。我们对这一现象给出直观解释：如果测试伙伴与智能体的训练伙伴相似，那么该智能体可能能够与这个未曾见过的伙伴良好协作。由于智能体之间的相似度差异很大，交叉协作得分也会如此。为了衡量这种相似度，我们定义了智能体训练伙伴与测试伙伴之间的条件策略相似度（CPSTT），即智能体的训练伙伴和测试伙伴在与该智能体协作时采取相同行动的期望概率。此外，我们在合作卡牌游戏 Hanabi（Bard 等人，2020）中进行了大量实验，发现 CPSTT 与交叉协作得分之间存在很强的线性相关性。

受这种线性相关性的启发，我们提出了一种轻量级方案——基于相似度的鲁棒训练（SBRT），以提升智能体的零样本协作表现。该方案在训练过程中根据预定义的 CPSTT 值随机扰动训练伙伴的策略，从而缓解智能体对特定训练伙伴的过拟合。我们将 SBRT 应用于三种多智能体强化学习框架：IQL（Tan，1993）、VDN（Sunehag 等人，2018）和 SAD（Hu 和 Foerster，2019）。实验结果证实，这些框架的零样本协作表现均有所提升。

我们的贡献总结如下：

1. 我们定义了 CPSTT，并进行了大量实验，以

版权所有 © 2023，美国人工智能促进协会（www.aaai.org）。保留所有权利。

揭示它与交叉协作得分之间存在很强的线性相关性：交叉协作得分几乎随其线性增长。这一发现为零样本协作提供了新的视角。2. 受这种线性相关性的启发，我们提出了一种轻量级方案 SBRT，根据预先定义的 CPSTT 值，在训练期间随机扰动智能体伙伴的策略。3. 我们将 SBRT 应用于三个多智能体强化学习框架，并通过实验验证了所提方法的有效性。

背景

去中心化部分可观测马尔可夫决策过程

我们使用去中心化部分可观测马尔可夫决策过程（Dec-POMDP）（Nair 等人，2003 年）对多智能体协作任务进行建模。环境中的 N 个智能体只能部分观测环境。在时间步 t，当全局环境状态为 $s_t \in$ S时，智能体 i 获取其观测 $o_t^i \sim O(o|i, s_t)$，并根据其动作-观测轨迹 $\tau_t^i = \{o_0^i, a_0^i, ..., o_t^i\}$ 以及策略 π^i选择动作 a_t^i： $a_t^i \sim \pi^i(a^i|\tau_t^i)$。随后，环境反馈共享奖励 $r_t = R(s_t, \vec{a}_t)$，并更新全局状态 $s_{t+1} \sim T(s'|s_t, \vec{a}_t)$，其中 $\vec{a}_t = [a_t^1, ..., a_t^N]$ 是联合动作。训练目标是最大化期望回报 $J(\pi^1, ..., \pi^N) = \mathbb{E}_{\pi^1,...,\pi^N}[\sum_t \gamma^t r_t]$。

为实现这一训练目标，研究人员提出了许多多智能体强化学习框架，包括基于策略梯度的框架（Foerster 等人，2018 年；Kuba 等人，2021 年）和基于价值的框架（Rashid 等人，2018 年；Son 等人，2019 年）。这些框架通常让一组固定的智能体共同参与游戏并进行学习，使智能体能够快速学会合作行为。

零样本协作与交叉对战

遵循该领域的常见设定（Hu 等，2020），我们将零样本协作建模为双智能体场景。假设一个采用策略 π 的智能体与一个采用策略 π_p 的伙伴智能体共同训练，训练目标如下所述：

$$\pi^*, \pi_p^* = \arg\max_{\pi,\pi_p} J(\pi, \pi_p) \tag{1}$$

如果 π 和 π_p 在测试期间继续合作，它们很可能取得高分。然而，零样本协作要求智能体与未见过的伙伴进行协作。一种相似的设定是临时团队协作（Rahman 等人，2021 年；Zand、Parker-Holder 和 Roberts，2022 年），在这种设定中，智能体必须对陌生伙伴的策略进行建模，并在互动过程中调整自身策略。相比之下，零样本协作不允许智能体在面对未见过的伙伴时改变自身策略。

为了评估智能体在这种设置下的表现，研究人员通常会进行交叉对战，即在测试期间让单独训练的智能体组合起来进行协作。此外，训练框架的零样本协作表现以使用该框架训练的多个智能体的平均交叉对战得分来衡量。下面我们对这一过程进行公式化表示。假设由训练框架 M 训练的策略服从分布P_M，那么当测试伙伴由训练框架 M_o训练时，M 的零样本协作表现可表示为：

$$Z(M) = \mathbb{E}_{\pi\sim P_M, \pi_o \sim P_{M_o}}[J(\pi, \pi_o)] \tag{2}$$

显然， $Z(M)$在很大程度上依赖于 M_o。根据 M_o的不同选择，交叉对战可分为两类：

(1)算法内交叉对战： $M_o = M$，即未见过的伙伴使用相同的训练框架、但采用不同的随机数种子进行训练（Cui 等人，2021）。这种测试易于实施且客观，但其缺点也很明显：它无法说明智能体是否能够与通过其他框架训练得到的智能体良好协作。

(2)算法间交叉对战： $M_o = (M_1 + M_2 + ... + M_K)$，其中 M_1， M_2，..... M_K 是其他多智能体强化学习框架（Lucas 和 Allen，2022）。这种方式似乎更加全面，但受测的 K 个框架可能无法代表所有可行的框架，因此结果可能存在偏差。

为了更全面地评估智能体的零样本协作表现，我们在实验中同时进行了算法内交叉对战和算法间交叉对战。

条件策略相似度

不同智能体组合的交叉协作得分差异很大，即使这些智能体是由同一框架训练的，情况也是如此。为了解释这一现象，我们直观地猜测，如果一个未曾见过的伙伴与智能体的训练伙伴相似，那么该智能体就应能与这个伙伴良好协作。本节将定义 CPSTT 来衡量这种相似度，提出一种估算方法，并给出实验结果，表明它与交叉协作得分之间存在很强的线性相关性。

条件策略相似度的定义与估算

下面给出条件策略相似度的定义：

定义 1. 在双智能体博弈中，以 π 为条件， π_1 与 π_2 之间的条件策略相似度为：

$$S_\pi(\pi_1, \pi_2) = \mathbb{E}_{\tau\sim P_{\tau(\pi,\pi_1)}}[\pi_1(\tau) = \pi_2(\tau)] \tag{3}$$

其中，$P_{\tau(\pi,\pi_1)}$ 表示由 π 和 π_1 相互对战所生成的动作—观测轨迹分布。

条件策略相似度衡量从 π 的视角来看， π_2 与 π_1 有多相似，并可通过蒙特卡洛方法进行估计：让 π 和 π_1 多次进行游戏，并假设总共有 n 个时间步。然后， π_1 根据 n 的动作—观察轨迹 $\{\tau_1, \tau_2, ..., \tau_n\}$ 做出 n 次决策 $\{\pi_1(\tau_1)$ $\pi_1(\tau_2)$ ....., $\pi_1(\tau_n)\}$。令 π_2 根据这些 n 轨迹采取行动，那么 $S_\pi(\pi_1, \pi_2)$ 的估计值

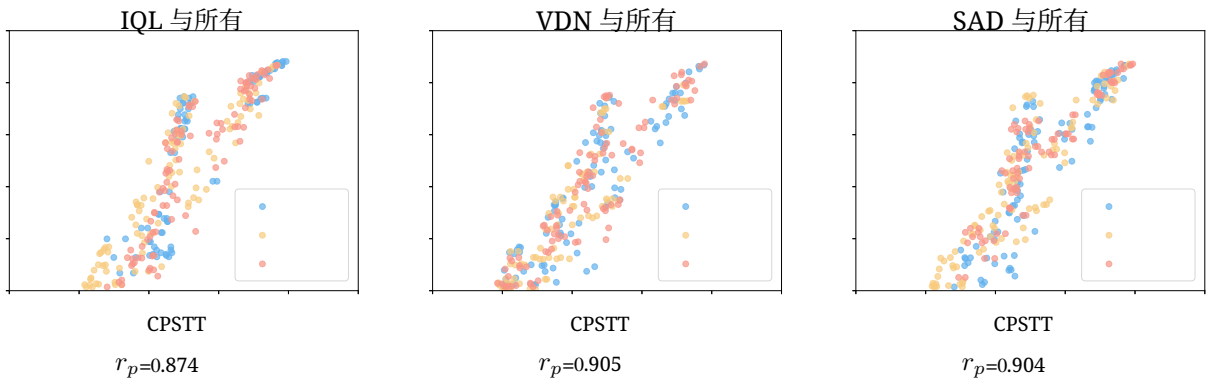

图1：每幅图展示了一类模型与所有类型模型进行交叉对战的详细结果。 r_p 为皮尔逊相关系数，这是一项描述两个因素之间线性相关性的统计量（$r_p = 1$ 表示完全线性关系）。总体而言，CPSTT 与交叉协作得分之间的线性相关性较强。

变为：

$$\bar{S}_\pi(\pi_1, \pi_2) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\pi_1(\tau_i)=\pi_2(\tau_i)} \tag{4}$$

本文聚焦于零样本协作中某个智能体的训练伙伴与测试伙伴之间的条件策略相似度（即 CPSTT）。给定一个策略为 π 的智能体、其训练伙伴策略 π_p 和测试伙伴策略 π_o，CPSTT= $S_\pi(\pi_o, \pi_p)$。

测试平台：Hanabi

我们在 Hanabi（Bard 等人，2020）上进行实验。这是一款常用于研究零样本协作的合作式纸牌游戏。每张卡牌都有一种颜色和一个点数，玩家必须按合法顺序出牌，才能完成五种颜色各自对应的一叠牌。游戏有 5 种颜色和 5 个点数，最高得分为 25 分。Hanabi 采用典型的合作奖励机制：智能体因提高团队得分而获得奖励，因失去团队生命标记而受到惩罚。请注意，每位玩家都能看到其他所有人的手牌，却看不到自己的手牌，因此必须根据他人的提示猜测自己手中的牌，同时还要向他人提供有价值的信息。

由于这项任务要求智能体推断伙伴的信念和意图，而自博弈智能体又非常熟悉自己的训练伙伴，因此它们能够达到超人类水平（Hu 和 Foerster，2019）。然而，它们很容易学会特殊约定，以获得较高的自博弈得分。例如，可以提示“蓝色”，让伙伴打出从左数起的第一张牌。这种技巧有助于快速实现（1）中的自博弈目标，却无助于智能体与陌生伙伴良好协作。因此，Hanabi 中的零样本协作十分困难，该领域的大量研究都以此为实验对象（Hu 等，2020；Treutlein 等人，2021；Lucas 和 Allen，2022；Hu 等人，2021）。

探索 CPSTT 与交叉协作得分之间的关系

我们进行了大量实验，以检验 CPSTT 与交叉协作得分是否相关。具体而言，我们针对三种模型进行了实验：IQL、VDN 和 SAD。它们是典型的自博弈算法，我们对其采用参数共享来加快训练，这是同质多智能体博弈自博弈中广泛使用的一种技术（Christianos、Sch¨afer 和 Albrecht，2020）。因此，智能体的训练伙伴就是它自身，这四类模型的 CPSTT 为 $S_\pi(\pi_o, \pi)$。我们使用不同的随机种子，为每种框架训练 10 个模型，并将这 30 个模型两两配对，得到 900 种组合。对于每种组合，我们记录 10000 场博弈的平均得分，并根据公式（4）估算 $S_\pi(\pi_o, \pi)$。

为了直观展示条件策略相似度如何影响交叉协作得分，我们在图 1 中展示了详细的交叉对战结果。这些图呈现出相同的规律：得分随相似度的提高而上升。这证实了我们的猜想：如果测试伙伴与智能体的训练伙伴相似，合作往往会更好。那么，它们之间的线性相关性究竟有多强？我们在每幅图下方给出了皮尔逊相关系数（r_p）。最低为 0.874（IQL），最高为 0.905（SAD）。作为参考，$r_p = 1$表示一种不切实际的完美相关。因此，CPSTT 与交叉协作得分之间具有很强的线性相关性。基于这一事实，我们可以更全面地分析零样本协作，并提出提高交叉协作得分的新方法。

基于条件策略相似度提高交叉协作得分

上一节的实验表明，交叉协作得分几乎随 CPSTT 线性增长。因此，基于这种相关性，可以通过两种方式提高交叉协作得分。第一种是提高 CPSTT。然而，零样本协调意味着要与未见过的伙伴合作，而提高训练伙伴与未知智能体之间的策略相似度几乎是不可能的。第二种方式更可行，即在 CPSTT 固定的情况下提高交叉协作得分；基于此，我们提出了自己的方案。

基于相似度的鲁棒训练

受鲁棒强化学习（Mankowitz 等人，2018）的启发，我们提出了一种鲁棒训练方案 SBRT，可应用于常见的自博弈多智能体强化学习框架。其训练目标为

$$\pi^*, \pi_p^* = \arg\max_{\pi,\pi_p} J(\pi, \pi_a) \;\; s.t. CPSTT_\pi(\pi_p, \pi_a) = \alpha \tag{5}$$

在训练的第一阶段，我们设置 $\alpha = 1$，这意味着训练方式与自博弈相同，使智能体能够快速优化其策略。训练 N_{st} 个周期后，当训练接近收敛时，我们设置 $\alpha = \alpha_r$，并进行 N_{rt} 个周期的鲁棒训练。α_r，N_{st}和 N_{rt} 是训练超参数。我们通过在智能体与环境交互时扰动搭档选择的动作来实现 π_a。具体而言，给定动作-观测轨迹 τ 以及由 π_p生成的动作：$a_p = \pi_p(\tau)$，

$$\pi_a(\tau) = \begin{cases} a_p, & with\ probability\ \alpha \\ a_{alt} \neq a_p, & with\ probability\ 1-\alpha \end{cases} \tag{6}$$

我们测试了三种选择 a_{alt} 的方法：

最差替代动作： 这种选择要求 π_a 满足

$$\pi_a = \arg\min_{\pi_a} J(\pi, \pi_a), \quad s.t. CPSTT_\pi(\pi_p, \pi_a) = \alpha \tag{7}$$

该方法遵循对抗训练的思路（Pan 等，2019），其中对手试图最小化主智能体的表现。在采用 Q 函数 $Q(\tau, a)$ 的基于 DQN 的框架下，$a_{alt} = \arg\min_a Q(\tau, a)$。

最佳替代　i动作 i作: T his choi作要求i求 π_a 以满足isfy

$$\pi_a = \arg\max_{\pi_a} J(\pi, \pi_a), \quad s.t. CPSTT_\pi(\pi_p, \pi_a) = \alpha \tag{8}$$

选择这一做法的原因是，零样本协作伙伴的策略通常并不差，因此这一限制可能会让 π_a 更像它们。在基于 DQN 的框架中，给定 Q函数 $Q(\tau, a)$，$a_p = \arg\max_a Q(\tau, a)$ 和$a_{alt} = \arg\max_{a \neq a_p} Q(\tau, a)$。

随机替代动作： 该方案只是随机选择一个可行行动作为 a_{alt}。它转而最大限度地探索可能遇到的未知伙伴策略。实验结果表明，这是最佳方案，因此我们在主要实验中选择该方案来实现 SBRT。

SBRT 的评估

为了考察 SBRT 是否能提升常见自博弈智能体的零样本协作表现，我们将该方案与三种基线多智能体强化学习框架结合：IQL、VDN 和 SAD。我们还将 OP（Hu 等，2020）纳入比较；该方法通过打破任务特定的对称性来提高交叉协作得分。在这项任务中，OP 会打乱卡牌的颜色，以防智能体形成特殊共识（例如，提示“蓝色”，让伙伴打出从左数第一张卡牌）。

至于训练超参数，$\alpha_r = 0.8$, N_{st} =400 和 $N_{rt} = 100$。此外，为了与 SBRT 保持一致，我们将基线模型和 OP 模型的总训练轮数设为 500。为提高结果的可靠性，我们训练每种方法都使用不同的随机数种子训练 10 个模型。

表1：基线模型、OP 模型和 SBRT 模型交叉协作得分的均值和标准误

模型类型	交叉协作得分	
	算法内	算法间
IQL	13.55±0.99	12.39±0.21
IQL+OP	11.93±1.10	11.35±0.22
IQL+SBRT	**15.72±0.70**	**13.25±0.20**
VDN	7.70±0.86	9.97±0.20
VDN+OP	6.11±0.81	8.65±0.22
VDN+SBRT	**11.38±0.92**	**12.14±0.20**
SAD	14.12±0.86	11.32±0.20
SAD+OP	11.35±0.93	9.25±0.19
SAD+SBRT	**14.70±0.77**	**13.19±0.19**

我们通过算法间和算法内交叉对战来衡量测试模型的零样本协作性能。算法内交叉对战要求智能体与由相同框架训练的伙伴进行协作。例如，IQL 与 IQL 模型协作，VDN+OP 与 VDN+OP 模型协作。算法间交叉对战要求智能体与由其他框架训练的伙伴进行协作。例如，对于 IQL 而言，IQL+OP 和 IQL+SBRT 模型的算法间交叉对战伙伴是原始 VDNandSAD 模型。

表1所示结果证实，SBRT 有效提升了 IQL、VDN 和 SAD 的算法间及算法内交叉协作得分。对 SBRT 有效性的一种直观解释是，它使智能体训练伙伴的策略更加多样化，从而避免形成特殊共识，并使智能体在与未见过的测试伙伴协作时表现更佳。

结论与未来工作

我们关注零样本协作，即智能体需要与未见过的伙伴合作。针对这一问题，人们已经提出了多种方法，但这些方法要么缺乏泛化能力，要么耗时很长。本文定义了条件策略相似度，并发现它在零样本协作中至关重要：交叉协作得分与CPSTT高度相关，两者之间的皮尔逊相关系数最高可达0.905。基于这种相关性，我们提出轻量级方案SBRT，通过根据预定义的CPSTT值扰动训练伙伴的策略，提升智能体的零样本协作表现。为通过实证确认其有效性，我们将其应用于三种多智能体强化学习框架（IQL、VDN和SAD）。结果表明，SBRT成功提升了这些模型的算法间和算法内交叉协作得分。我们希望，交叉协作得分与CPSTT之间强相关性的发现，以及我们的SBRT方案，能为该领域的研究提供新的视角。

参考文献

Bard, N.; Foerster, J. N.; Chandar, S.; Burch, N.; Lanctot, M.; Song, H. F.; Parisotto, E.; Dumoulin, V.; Moitra, S.; Hughes, E.; et al. 2020. Hanabi挑战：人工智能研究的新前沿。人工智能, 280: 103216. Bøgh, S.; Jensen, P. G.; Kristjansen, M.; Larsen, K. G.; and Nyman, U. 2022. 基于模型预测控制的噪声环境下分布式车队管理。载于国际自动规划与调度会议论文集，第32卷, 565–573。Calvo, J. A.; and Dusparic, I. 2018. 用于交通信号灯控制的异构多智能体深度强化学习。载于*AICS*, 2–13。Christianos, F.; Sch¨afer, L.; and Albrecht, S. 2020. 面向多智能体强化学习的共享经验演员-评论家。神经信息处理系统进展，33: 10707–10717。Cui, B.; Hu, H.; Pineda, L.; and Foerster, J. 2021. Hanabi中的零样本协作K级推理。神经信息处理系统进展，34: 8215–8228。Foerster, J.; Farquhar, G.; Afouras, T.; Nardelli, N.; and Whiteson, S. 2018. 反事实多智能体策略梯度。载于*AAAI*人工智能会议论文集，第32卷。Ghosh, S.; Laguna, S.; Lim, S. H.; Wynter, L.; and Poon-awala, H. 2021. 多智能体强化学习的深度集成方法: 空中交通管制案例研究。载于国际自动规划与调度会议论文集，第31卷，468–476。Hu, H.; and Foerster, J. N. 2019. 面向深度多智能体强化学习的简化动作解码器。载于国际学习表征会议。Hu, H.; Lerer, A.; Cui, B.; Pineda, L.; Brown, N.; and Foer-ster, J. 2021. 离线信念学习。载于国际机器学习会议，4369–4379。PMLR。Hu, H.; Lerer, A.; Peysakhovich, A.; and Foerster, J. 2020. 面向零样本协作的“他者博弈”。载于国际机器学习会议，4399–4410。PMLR。Kuba, J. G.; Chen, R.; Wen, M.; Wen, Y.; Sun, F.; Wang, J.; and Yang, Y. 2021. 多智能体强化学习中的信赖域策略优化。载于国际学习表征会议。Li, J.; Harabor, D.; Stuckey, P. J.; Felner, A.; Ma, H.; and Koenig, S. 2019. 基于冲突的搜索中多智能体路径规划的不相交分裂。载于国际自动规划与调度会议论文集，第29卷，279–283。Lin, K.; Zhao, R.; Xu, Z.; and Zhou, J. 2018. 基于多智能体深度强化学习的高效大规模车队管理。载于第*24*届*ACM SIGKDD*国际知识发现与数据挖掘会议论文集，1774–1783。Lowe, R.; WU, Y.; Tamar, A.; Harb, J.; Pieter Abbeel, O.; and Mordatch, I. 2017. 面向混合合作竞争环境的多智能体演员-评论家。神经信息处理系统进展，30: 6379–6390。Lucas, K.; and Allen, R. E. 2022. Any-Play：一种用于零样本协作的内在增强。载于第*21*届自主智能体与多智能体系统国际会议论文集，853–861。Mahajan, A.; Rashid, T.; Samvelyan, M.; and Whiteson, S. 2019. Maven：多智能体变分探索。神经信息处理系统进展，32。Mankowitz, D.; Mann, T.; Bacon, P.-L.; Precup, D.; and Mannor, S. 2018. 学习鲁棒选项。载于*AAAI*人工智能会议论文集，第32卷。Nair, R.; Tambe, M.; Yokoo, M.; Pynadath, D.; and Marsella, S. 2003. 驯服去中心化部分可观测马尔可夫决策过程：面向多智能体场景的高效策略计算。载于*IJ-CAI*，第3卷，705–711。Citeseer。Pan, X.; Seita, D.; Gao, Y.; and Canny, J. 2019. 风险规避鲁棒对抗强化学习。载于*2019*年国际机器人与自动化会议（*ICRA*），8522–8528。IEEE。Rahman, M. A.; Hopner, N.; Christianos, F.; and Albrecht, S. V. 2021. 基于图的策略学习实现开放式临时团队协作。载于国际机器学习会议，8776–8786。PMLR。Rashid, T.; Samvelyan, M.; Schroeder, C.; Farquhar, G.; Foerster, J.; and Whiteson, S. 2018. Qmix：用于深度多智能体强化学习的单调价值函数分解。载于国际机器学习会议，4295–4304。PMLR。Shih, A.; Sawhney, A.; Kondic, J.; Ermon, S.; and Sadigh, D. 2021. 惯例在自适应人机协作中的关键作用。*arXiv*预印本 *arXiv:2104.02871*。Son, K.; Kim, D.; Kang, W. J.; Hostallero, D. E.; and Yi, Y. 2019. Qtran：通过变换学习分解，以实现合作式多智能体强化学习。载于国际机器学习会议，5887–5896。PMLR。Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W. M.; Zambaldi, V.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J. Z.; Tuyls, K.; et al. 2018. 基于团队奖励的合作式多智能体学习价值分解网络。载于第*17*届自主智能体与多智能体系统国际会议论文集，2085–2087。Tan, M. 1993. 多智能体强化学习：独立智能体与协作智能体。载于第十届国际机器学习会议论文集，330–337。Tesauro, G. 1994. TD-Gammon，一种自我学习的西洋双陆棋程序，达到大师级水平。神经计算，6(2)：215–219。Treutlein, J.; Dennis, M.; Oesterheld, C.; and Foerster, J. 2021. 零样本协作的新形式体系、方法与未决问题。载于国际机器学习会议，10413–10423。PMLR。Zand, J.; Parker-Holder, J.; and Roberts, S. J. 2022. 面向临时智能体协调的即时策略适应。载于第*21*届自主智能体与多智能体系统国际会议论文集，1771–1773。