arXiv:2305.00768v1 [cs.MA] 1 May 2023

异质的社会价值取向导致序贯社会困境中有意义的多样性

Udari Madhushani 普林斯顿大学
udarim@princeton.edu

Kevin R. McKee,
DeepMind,
kevinrmckee@google.com

John P. Agapiou
DeepMind
jagapiou@google.com

Joel Z. Leibo,
DeepMind,
jzl@google.com

Richard Everett
DeepMind
reverett@google.com

Thomas Anthony
DeepMind
twa@google.com

Edward Hughes
DeepMind
edwardhughes@google.com

Karl Tuyls
DeepMind
karltuyls@google.com

Edgar A. Duéñez-Guzmán
DeepMind
duenez@google.com

摘要

在社会心理学中，社会价值取向（SVO）描述个体在自己与他人之间分配资源的倾向。在强化学习中，SVO 被视为一种内在动机，它会根据特定的群体奖励目标分布重新映射智能体的奖励。先前研究表明，在与囚徒困境激励结构相似的情境中，具有异质SVO的智能体群体会学得多样化策略。我们的研究拓展了这一系列成果，并证明：（1）在序贯社会困境中的多种激励结构下，异质SVO会带来有意义的策略多样性，这一结论由任务特定的多样性指标衡量；（2）在某些情况下，针对这种策略多样性学习最佳响应，能够带来更好的零样本泛化。我们表明，这些最佳响应智能体学到的策略会根据共同玩家进行条件化；我们认为这是零样本泛化效果有所提升的原因。

1 引言

在心理学研究中，社会价值取向（SVO）是一种认知构念，反映了一个人对如何在自己与他人之间分配资源的偏好 [7, 15, 22]。有些人可能只关注最大化个人成功，而另一些人则表现出不同的动机，包括最大化自己与他人结果之间的差距（竞争型取向）、最大化集体福祉（亲社会取向），或最大化他人的利益（利他型取向）。在人工智能研究中，各种算法在设计或实现时都从这些洞见中汲取灵感[19, 27]。在强化学习中，社会价值取向是一种内在动机，它根据一个参数化的目标分布，在智能体自身奖励与他人奖励之间进行权衡，从而调整智能体的奖励。近年来，研究者探讨了社会价值取向在社会困境中的作用。社会困境是指一群智能体或参与者相互作用的情境，其中涉及个人利益与群体集体利益之间的权衡。这些研究揭示了社会价值取向对多样化行为和合作的产生 [19, 20], 以及伙伴选择 [18]的影响。社会价值取向研究主要聚焦于激励结构类似于囚徒困境 [26], 的社会困境；在这种情境下，每位参与者都有背叛的动机，尽管双方合作会让他们的处境更好。

序贯社会困境是一类社会困境，其中互动智能体的决策过程在时间和空间上均有所延展 [13]。在序贯社会困境中取得良好表现，可以通过考虑长期后果、相互依存关系以及群体成员之间的合作来实现。序贯社会困境已被广泛用于研究合作的涌现与维持[14, 25], 、不公平厌恶 [11], 、伙伴选择 [4, 18], 以及泛化 [1, 20] ；在泛化研究中，智能体会在测试场景中与新的共同玩家互动。

尽管环境会提供可用于学习策略的外在奖励，但通常也有必要为智能体提供内在奖励，以引导其行为，使其采用具有理想属性的策略。内在奖励的使用方式类似于人类决策中的社会偏好。在大多数关于序贯社会困境的研究中，所有玩家要么没有内在奖励，要么都采用相同的函数（即具有同质的社会偏好） [14, 31]。然而，研究发现，具有不同内在奖励函数的智能体群体（即具有异质社会偏好的群体）能够带来更高水平的合作 [11]。在 [18–20], 中，作者表明，异质性能够在群体困境中，以及在激励结构类似于囚徒困境的博弈中，产生行为多样性。其他激励结构尚未得到探索。此外，社会偏好多样性与智能体策略多样性之间的确切相互作用，尤其是能够实现向新的社会伙伴泛化的机制，仍未得到充分研究。

策略的多样性已被证明能够提升智能体性能的多个方面，例如探索 [33],适应环境变化 [3], 改善群体结果 [19, 30],泛化到新的共同玩家 [17], 以及与人类协作 [29]。量化多样性的一种方法，是考察智能体与不同共同玩家互动时获得的奖励（通常

自适应与学习智能体研讨会论文集（*ALA 2023*），*Cruz*、*Hayes*、*Wang*、*Yates*（编），*2023*年*5*月*9*日至*10*日，线上，*https://ala2023.github.io/*。2023。

称为策略多样性）[2, 6]。作为对这些方法的补充，也可以通过状态—动作变化来评估多样性，即衡量智能体所经历的状态—动作对的分布。可以通过测量状态访问频率的差异 [33], 给定状态下动作选择频率的差异 [20],或从特定状态开始的状态—动作轨迹之间的差异 [17]来评估状态—动作多样性。要定义一种不依赖特定环境、基于状态—动作变化且能捕捉有意义的多样性（即对群体轨迹产生更广泛影响的多样性）的指标，可能颇具挑战。另一种做法是采用特定于环境的多样性度量，研究者可以利用对特定环境特征的了解来设计这些度量。

零样本泛化 [9, 10, 12, 20, 29] 旨在开发能够在测试时与新颖智能体（即训练期间未见过的智能体）成功交互的通用智能体。在此类情形下，测试时遇到的新颖智能体的策略可能超出这些智能体的分布范围，导致它们在纯合作场景中协调不佳 [10, 17], , 并在竞争场景中遭到利用 [24]。在混合动机博弈中，无法泛化到新颖智能体可能会错失合作机会，从而造成无谓损失 [12]。学习针对具有多样化策略的伙伴／对手的最佳响应，已成为实现零样本泛化的一种有前景的方法 [29]。这一方法背后的直觉是，使用一组多样化策略进行训练，可以降低测试时遇到分布外策略的可能性。尽管前景可期，这些最佳响应技术尚未应用于范围广泛的激励结构中。

在这项研究中，我们评估了序贯社会困境中不同激励结构下的异质SVO。我们纳入了在时间和空间上延展的环境，其底层结构与几种不同的矩阵博弈相似：囚徒困境；胆小鬼博弈，在这种博弈中，双方玩家都有动机选择高风险行为，但最糟糕的结果是双方都选择高风险；以及猎鹿博弈，在这种博弈中，参与者可以选择安全选项，同时也有动机协调采取高奖励策略，但若协调失败则可能付出高昂代价。胆小鬼博弈和猎鹿博弈属于均衡选择型社会困境。

我们将异质SVO会带来多样化策略这一观察扩展到类似胆小鬼博弈和猎鹿博弈的激励结构中。我们还表明，在均衡选择型序贯社会困境中，通过最佳响应利用这种多样性，可以提升零样本泛化能力。我们发现，最佳响应智能体通过学习条件策略，能够适应行为各异的伙伴／对手。然而，当序贯社会困境并非均衡选择问题时，学到的最佳响应会退化为单一的无条件策略，导致零样本泛化能力较差

本文结构安排如下。第 2 节概述本文采用的方法。第 2.1 小节介绍本文所用 N 智能体部分可观测马尔可夫过程的形式化表述。第 2.2 小节描述社会价值取向（SVO）框架及其实现。第 2.3 小节讨论研究中使用的各种环境及其特征。第 2.4 小节详细介绍在序贯社会困境中生成多样化策略的流程。第 2.5 小节介绍使用智能体群体训练最佳响应智能体并评估零样本泛化性能的过程。

此外，我们在第 2.5 小节介绍智能体的架构。第 3 节呈现研究结果。第 3.1 和 3.2 小节展示在具有不同激励结构的环境中生成多样化策略所取得的结果。第 3.3 小节呈现零样本泛化性能评估的结果。最后，第 4 节提供进一步讨论并总结全文。本节概述了研究的主要贡献，并讨论了潜在的社会影响。

2 方法

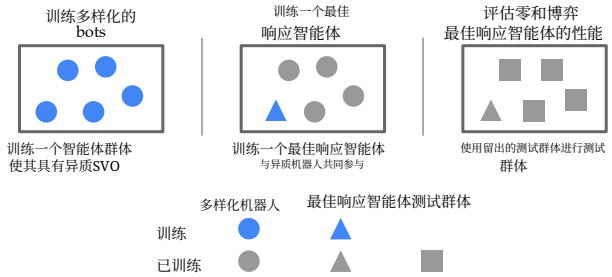


图 1：方法概述。蓝色形状表示正在积极训练的智能体，灰色形状表示被冻结的智能体（机器人）。圆形表示采用多样化社会价值取向训练的智能体，三角形表示最佳响应智能体，正方形表示留出集中的共同玩家。评估采用零样本方式，这意味着最佳响应智能体处于冻结状态（灰色三角形），并针对留出集中的机器人进行评估。

2.1 N-智能体 POMDP

我们考虑一个由元组 $\langle N、\mathrm{S}、\mathrm{A}、\mathrm{R}\ P\ \gamma\rangle$定义的多智能体部分可观测马尔可夫决策过程，其中 N是智能体的数量，S 是联合状态空间，$\mathrm{A} = \times_{i=1}^{N}\mathrm{A}^i$ 是联合动作空间， P是状态转移概率分布，R 是奖励函数， γ 是折扣因子。这也可以称为部分可观测马尔可夫博弈 [16] 或部分可观测随机博弈 [28]。在每个时间步 t，每个智能体 $i \in 1、\ldots\ldots、N$都会观察一个私有（局部）观测 o_t^i，并从动作集合 A^i中采取一个动作 a_t^i。所有智能体在时间步 t采取的联合动作记为 $a_t = (a_t^1、\ldots\ldots、a_t^N)$。状态 s_t 不会被智能体直接观测到；相反，部分观测 o_t^i取决于环境的当前状态 s_t 以及智能体的观测函数。智能体 i的观测函数记为 $O^i(o_t^i|s_t)$。每个智能体 i都会获得奖励 r_t^i，该奖励是联合动作 a_t和环境状态 s_t的函数。环境状态按照概率分布 $P(s_{t+1}|s_t、a_t)$进行转移。

每个智能体 i的目标是在给定的有限时间范围内最大化其累积期望折扣奖励，定义为 $J^i = \mathbb{E}\left[\sum_{t=0}^{T}\gamma^t r_t^i\right]$，其中 $\gamma \in [0, 1]$ 用于权衡即时奖励与未来奖励的重要性。智能体的策略被定义为从智能体的观测历史到一个动作的映射，即：

$\pi^i(a_t^i|o_1^i, \cdots o_t^i)$。这些策略通过多智能体强化学习算法进行更新，以最大化智能体的目标函数。

2.2 社会价值取向

省略对 t 的依赖，令 r^i 为智能体 i. 的奖励。令 $\bar{r}^{-i}$ 为除智能体 i. 以外所有智能体的平均奖励。于是有

$$\bar{r}^{-i} = \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} r^j.$$

令 θ^i 表示智能体 i 的社会价值取向目标角度。根据 [18], 中给出的定义，我们将智能体 i 的有效奖励 $\hat{}r^i$ 定义为

$$\hat{r}^i = r^i \cos(\theta^i) + \bar{r}^{-i} \sin(\theta^i).$$

虽然有时会对内在奖励进行时间平滑处理（例如，[11]），但在本研究中，有效奖励不包含任何时间平滑处理。将上一节中的时间步 t 重新引入后，智能体 i 优化的目标函数为

$$\hat{J}^i = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \hat{r}_t^i\right].$$

2.3 环境

我们简要介绍这些环境。本文所有实验均使用 Melting Pot 2.0 中的环境，且未作任何修改 [1]。

“矩阵内”重复博弈：“矩阵内”重复博弈是 Melting Pot 2.0 中的一类序贯社会困境，涉及两名玩家进行交互。在每个回合开始时，环境会根据给定的资源布局以及玩家可以生成的一组固定点进行初始化。地图由两种资源组成，可通过颜色区分：红色对应背叛，绿色对应合作（见图3）。玩家可以走过资源将其拾取，并将其收入库存。玩家生成时，库存中各有一种资源。生成后，每名玩家都可以在地图上移动、收集资源，并通过发射交互光束与另一名玩家互动。当玩家进行交互时（即一名玩家用交互光束击中另一名玩家），每名玩家都会获得一份奖励，其数值等于根据库存数量和特定环境的收益矩阵计算出的预期收益。用光束击中另一名智能体的智能体被视为行玩家。每名玩家的库存数量决定了一种混合策略，其中选择每种纯策略的概率等于相应资源所占的比例。令N_r^i和 N_g^i 分别表示智能体 $i \in 1$的库存数量，即红色资源和绿色资源的数量，2。对于每个智能体i，其混合策略如下：

$$p = \left[\frac{N_r^i}{N_r^i + N_g^i}, \frac{N_g^i}{N_r^i + N_g^i}\right]$$

设 A 为该博弈的收益矩阵。设 r_{row} 和 r_{col} 为..... 分别表示行玩家和列玩家的奖励。令 p_{row} 和 p_{col} 分别为行玩家和列玩家的混合策略概率向量。那么，奖励可以定义为

$$r_{row} = p_{row}^T A p_{col}, \quad r_{col} = p_{col}^T A^T p_{row}$$

这些奖励计算对应于博弈论中用于矩阵博弈和迭代社会困境的计算 [32]。

猎鹿博弈

4	**0**
2	**2**

胆小鬼博弈

3	**2**
5	**0**

囚徒困境

3	**0**
5	**1**

图 2：猎鹿博弈、胆小鬼博弈和囚徒困境的收益矩阵。所示数值对应行玩家的收益。列玩家的收益是所示矩阵的转置（即这些博弈是对称博弈）。合作对应第一行和第一列。背叛对应第二行和第二列。

所使用的收益矩阵 A 见图 2。完成交互后，参与者会获得交互奖励，冻结 16 个步骤，并将其库存数量重置为每种资源类型各一个。16 个步骤结束后，参与者会消失，并在 5 个步骤后重新生成。参与者可以在一个回合内进行多次交互。资源被拾取后，会在延迟 10 个步骤后开始再生，此后每一步都有 20% 的概率再生。按照 Melting Pot 2.0 的惯例，在每场游戏中，每经过 100 个步骤，回合都有 10% 的概率结束，但每个回合至少进行 1000 个步骤。

外部性蘑菇： 外部性蘑菇是一种序贯社会困境，在这种困境中，参与者会立即受到共同玩家的亲社会或反社会行为的影响。这是一款由 5 名玩家参与的游戏，玩家通过吃蘑菇来获得奖励。地图上会生长四种蘑菇（数量各不相同）：红色、绿色、蓝色和橙色。吃掉红色（“fize”：完全内在性、零外部性）蘑菇后，食用该蘑菇的玩家可获得 1 点奖励。吃掉绿色（“hihe”：一半内在性、一半外部性）蘑菇后，所有玩家总共获得 2/5的奖励。吃掉蓝色（“zife”：零内在性、完全外部性）蘑菇后，所有玩家平分总额为 3/4 的奖励，但不包括食用该蘑菇的玩家。吃掉橙色（“nize”：负内在性、零外部性）蘑菇会使红色 fize 蘑菇各自以 0.25 的概率被摧毁，并使食用该蘑菇的玩家获得 -0.1 点奖励。吃掉蘑菇后，食用该蘑菇的玩家会在蘑菇的消化时间内冻结：红色为 0 步，绿色为 10 步，蓝色为 15 步，橙色为 15 步。蘑菇生成后，会在其消亡时间结束后从地图上消失，即分别经过以下时间后消失：红色 200 步、绿色 100 步、蓝色 75 步。橙色蘑菇永不消亡。蘑菇会根据其他蘑菇的消耗情况，由孢子重新生成。吃掉红色、绿色或蓝色蘑菇会释放 3 个红色蘑菇孢子，每个孢子有 0.25 的概率生成蘑菇。吃掉绿色或蓝色蘑菇还会释放

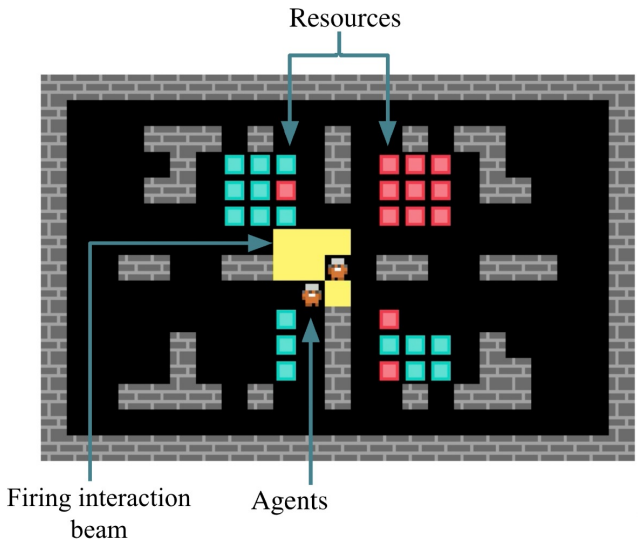


图3：“矩阵中”重复博弈。这是一种双人博弈，智能体可以采集两种资源（绿色代表合作，红色代表背叛）。智能体交互（使用交互光束）时，会根据各自的库存数量和特定博弈的收益矩阵获得奖励。收益矩阵可以是猎鹿博弈、胆小鬼博弈或囚徒困境类型的收益矩阵。

3 个绿色蘑菇孢子，其生成概率为 0.4。吃掉蓝色蘑菇还会释放 1 个蓝色孢子，其生成概率为 0.6。吃掉橙色蘑菇会释放 1 个新橙色蘑菇的孢子，其生成概率为 1。与“in the matrix”重复博弈类似，在外部性蘑菇中，每个回合至少运行 1000 步。此后，该回合在每经过 100 步时，都有 0.2 的概率终止。

外部性蘑菇具有类似胆小鬼博弈的激励结构：当其他玩家采食蓝色或绿色蘑菇时，采食红色蘑菇可使奖励通过自私行为最大化。但如果其他所有人都在采食红色蘑菇，那么自私策略就是采食绿色蘑菇，否则所有蘑菇最终都会耗尽。

2.4 在序贯社会困境中生成多样化策略

在训练过程开始时，我们为每个智能体定义不同的社会价值取向角度。每个环境都有固定数量的玩家。我们通过初始化“竞技场”来训练智能体群体，并以分布式异步方式训练智能体。各竞技场并行运行，每个竞技场都是一个副本，其玩家数量由相应环境指定。这是 A3C 的多智能体版本 [21]，是多智能体强化学习中常用的方法 [1]。Melting Pot评估协议要求有放回地抽样策略。纯自我博弈训练会因与自身的副本进行博弈而引入偏斜的奖励激励。为缓解这一问题，我们设置每个竞技场中的玩家进行一回合游戏，采用自我博弈或

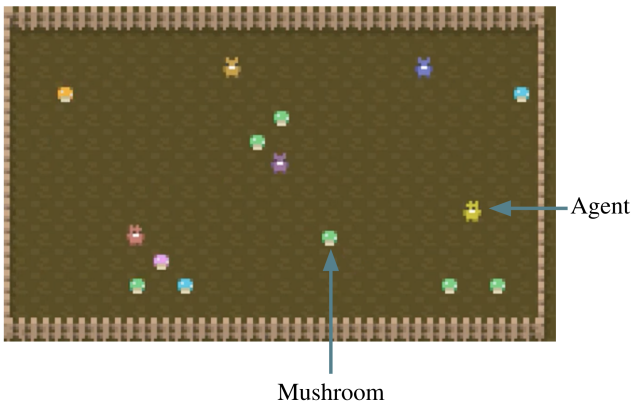


图4：外部性蘑菇。这是一种具有即时反馈的五人序贯社会困境游戏。智能体会根据他人采摘的蘑菇，立即与他人分享奖励。

群体博弈（两者概率相等）。在群体博弈过程中，我们不放回地抽取智能体。我们为每个智能体训练 10_9 个学习器帧。

2.5 训练最佳响应智能体并评估零样本泛化性能

我们训练一个没有内在奖励的自私朴素学习者，使其对使用异质SVO生成的策略做出最佳响应。为避免混淆，我们将训练中的智能体称为最佳响应智能体，将使用异质SVO值训练的预训练多样化智能体称为社会价值取向机器人 。在每个回合中，最佳响应智能体都会与一组不重复抽样的社会价值取向机器人进行博弈。我们训练最佳响应智能体，使其达到 10_9 个学习者帧。

Melting Pot 2.0 [1] 提供了一套评估协议，用于评估智能体对新颖社会伙伴的泛化能力。这些社会伙伴作为测试场景套件中的留出共同玩家集，与该套件一同提供。我们使用Melting Pot测试协议评估最佳响应智能体的表现。

我们使用Melting Pot测试场景，在猎鹿博弈、胆小鬼博弈、囚徒困境“矩阵中”的重复博弈以及外部性蘑菇中进行评估。测试场景的详细信息如下。

“矩阵中”重复博弈的测试场景。焦点玩家（我们的最佳响应智能体）会遇到：

S0：*(合作方 + 背叛方)* 以各 0.5 的概率成为合作方或背叛方

S1：*(合作方)* 合作方 S2：*(背叛方)* 背叛方

S3:*(冷酷策略1)* 当焦点玩家背叛一次后，先合作、随后在本回合剩余时间里持续背叛的玩家

S4:*(冷酷惩罚2)* 当焦点玩家背叛两次后，先合作、随后在本回合剩余时间里持续背叛的玩家

S5：（以牙还牙）采用以牙还牙策略的玩家

S6：（以牙还牙颤抖）采用以牙还牙策略并偶尔无条件背叛的玩家。（带噪声的以牙还牙）

S7：（反转）在前 3 次互动中合作，之后在本回合剩余时间里背叛的玩家

S8：（可纠正的以牙还牙策略）起初采取背叛，并在最佳响应智能体背叛时切换为以牙还牙策略的玩家

S9：（可纠正的以牙还牙策略颤抖）一名玩家以背叛开局，并在最佳响应智能体背叛时切换到带噪声的以牙还牙策略

外部性蘑菇的测试场景：
焦点玩家（我们的最佳响应智能体）遇到：

S0：（访问型合作者）4名合作方

S1：（访问型背叛者）4名背叛者

2名焦点玩家（在我们的案例中，是2个最佳响应智能体副本）遇到：

S2：（常驻合作者）3 名合作方

S3：（常驻合作者）3 名背叛者

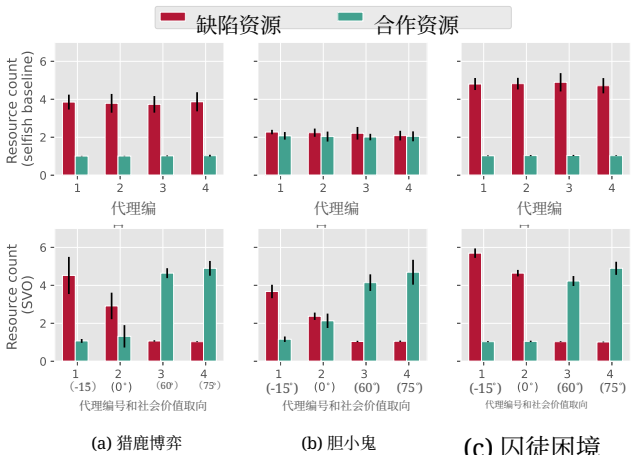
我们在图 1 中概述了端到端的方法流程。

2.6 智能体架构

我们使用 [5]，提出的成熟行动者-评论家基线算法训练智能体，该算法建立在 [21] 的早期研究基础上，名为异步优势行动者-评论家（A3C）。

智能体的神经网络由两个卷积层、一个两层感知机和一个长短期记忆网络（LSTM）组成，各层之间均由 ReLU 激活函数分隔。卷积层的输出通道数分别为 16 和 32，卷积核形状分别为 8 和 4，步幅分别为 8 和 1。感知机层各有 64 个神经元，LSTM 层有 128 个单元。评论家的策略和基线由多层感知机（256 个带 ReLU 激活的隐藏单元）构成，并连接到 LSTM 的输出。

表征塑造通过使用辅助损失和对比预测编码 [23],来实现，后者利用长短期记忆网络状态表征区分相邻时间点。PopArt [8] 用于调整不同环境中的不同奖励尺度。采用的优化方法为 RMSProp，学习率为 4×10^{-4}，epsilon 为 10^{-5}，动量为零，衰减率为 0.99，批量大小为 256。评论家的基线代价为 0.5，策略的熵正则化代价为 0.003。



(a) 猎鹿博弈　(b) 胆小鬼　(c) 囚徒困境

图 5：“矩阵内”重复博弈。自私基线机器人与社会价值取向机器人的策略多样性。 每个子图展示了 4 个智能体在评估期间的平均库存数量；这些智能体采用 50% 自我博弈和 50% 群体博弈进行训练。底行对应社会价值取向为 $\theta^i \in\{-15°,0°,60°,75°\}$ 的社会价值取向机器人，顶行对应自私基线机器人。绿色和红色分别表示合作资源数量和背叛资源数量。误差条表示 3 个随机种子下结果的标准差。

3 结果

3.1 实验 1：在“矩阵内”重复博弈中生成多样化策略

实验设置：我们研究猎鹿博弈、胆小鬼博弈和囚徒困境的“矩阵中”重复博弈。每种博弈的结果均取 3 个随机种子的平均值。我们训练四个社会价值取向分别为 $-15°$，$0°$，$60°$和 $75°$的智能体。这些取值围绕竞争（$-15°$）、自私（$0°$）和亲社会性（$60°$，$75°$）的激励设定。“矩阵中”重复博弈为双人博弈。除社会价值取向机器人外，我们还按照相同流程训练并冻结一组自私基线机器人（即没有内在奖励），以供比较。

发现 1：异质社会价值取向机器人能够学习到有意义的多样化策略

我们将机器人在交互时的库存数量用作特定环境的多样性度量。由于库存数量决定了混合策略概率向量，库存数量的比例存在足够明显的差异，就表明混合策略不同。在评估期间，智能体以群体博弈的方式进行博弈。

图 5 展示了训练完成后评估期间，4 个机器人在 500 次交互中的平均库存数量。上、下两行分别对应自私基线机器人和社会价值取向机器人的资源数量。图 5(a)、5(b) 和 5(c) 分别对应猎鹿博弈、胆小鬼博弈和囚徒困境。图中所示误差条对应 3 次独立运行结果的平均值。结果表明，在每种博弈中，4 个自私基线机器人的库存数量比例都相近，这说明它们的策略缺乏多样性。

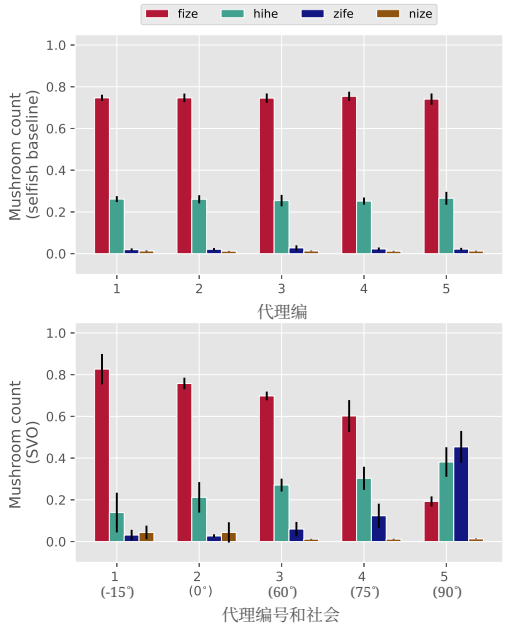


图 6：外部性蘑菇。自私基线机器人和社会价值取向机器人的策略多样性。**每个图展示了在评估期间 5 个智能体消耗的平均蘑菇比例；这些智能体是在外部性蘑菇密集型游戏中，以 50% 自我博弈和 50% 群体博弈的方式训练的。底排对应社会价值取向为 $\theta^i \in \{-15°, 0°, 60°, 75°, 90°\}$ 的智能体，顶排对应自私基线智能体。误差条表示 3 个随机种子下结果的标准差。**

相反，4 个社会价值取向机器人的库存数量比例各不相同，表明其行为具有多样性。对于每种“矩阵中的”重复博弈，资源数量对应于 SVO 角度为 $\theta = [-15°, 0°, 60°, 75°]$，的社会价值取向机器人，其中 $\theta^i = \theta[i]$，对于 $i \in \{1, 2, 3, 4\}$。我们分别用绿色和红色表示合作资源数量和背叛资源数量。随着社会价值取向角度从 $-15°$增至 $75°$，红色与绿色资源数量之比上升，表明利他行为增加。

3.2 实验 2：在外部性蘑菇中生成多样化策略

实验设置：与“矩阵中的”重复博弈的训练过程类似，我们对 3 个随机种子的结果取平均值。对于每个随机种子，我们分别训练 5 个社会价值取向值为 $-15°$、0°、60°、75°和 90°的智能体，训练时采用 50% 自我博弈和 50% 群体博弈。除社会价值取向机器人外，我们还使用相同的流程训练了一组自私基线机器人，以供比较。

发现 2：结果可推广到参与者超过2 人的多人博弈

我们展示了该方法可以扩展到拥有超过 2 名玩家的游戏中。图 6 显示，在外部性蘑菇环境中，使用异质SVO训练的智能体学会了多样化策略。我们使用各类型蘑菇的蘑菇消耗量作为特定于环境的多样性指标。自私基线机器人往往以相似的比例消耗不同类型的蘑菇，而社会价值取向机器人则以不同的比例消耗各类蘑菇，表现出具有实质意义的多样化行为。社会价值取向较低（或为负）的智能体会消耗自私型蘑菇（红色），甚至会消耗恶意型蘑菇（橙色）；而社会价值取向较高的智能体则倾向于消耗更多亲社会型蘑菇（绿色和蓝色）。

3.3 实验 3：零样本泛化评估

我们评估了针对使用异质SVO训练的社会价值取向机器人所学习的最佳响应的零样本泛化性能。**基线方法：**我们比较了针对社会价值取向机器人的学习所得最佳响应策略，与针对自私基线机器人的最佳响应的性能，

虚拟协同博弈（FCP，一种最佳响应方法，它还会纳入智能体先前的检查点进行最佳响应）[29], 和利用者（即直接在测试场景中训练的最佳响应智能体）[1]。我们针对每个测试场景训练一个利用者。为了训练 FCP 智能体，我们为机器人群体中的每个机器人训练一个包含 3 个检查点的朴素学习智能体。这里使用第一个检查点、中间检查点和最后一个检查点。中间检查点指训练过程中，智能体首次达到其最终奖励一半时的检查点，这里的奖励是指机器人策略的奖励。我们报告了 FCP 应用于异质社会价值取向机器人的结果（FCP(SVO)），以及应用于自私基线的结果（FCP(selfish-baseline)）。为了评估零样本泛化，我们还比较了最佳响应智能体、自私基线智能体和随机智能体的表现。

实验设置：我们为自私基线机器人和社会价值取向机器人训练最佳响应智能体。回顾一下，在此设置中，我们分别用 3 个随机种子训练每种类型的机器人，即自私基线机器人或社会价值取向机器人。我们为每个种子训练出的机器人训练一个最佳响应智能体。对于每种类型的测试机器人，我们展示与这 3 次训练相对应的评估运行的平均表现。

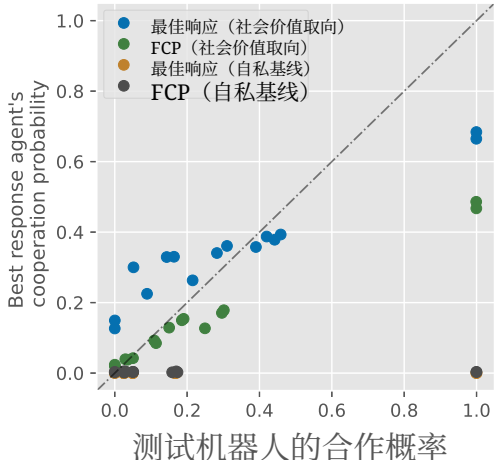


图7： 比较最佳响应智能体在矩阵中的猎鹿博弈里学习条件策略的效果。

发现3：最佳响应智能体学会了条件行为为了更好地了解所学到的策略

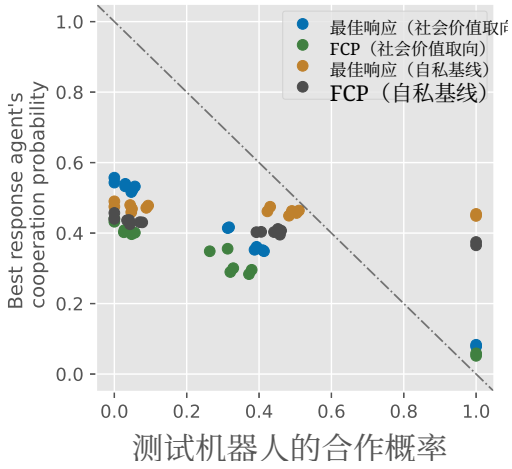


图8：比较最佳响应智能体在矩阵中的胆小鬼博弈里学习条件策略的效果。

我们分析了最佳响应智能体在测试时的行为。对于每个测试机器人，图7和图8分别展示了最佳响应智能体与测试机器人合作的交互比例，以及测试机器人与最佳响应智能体合作的交互比例。图7对应重复进行的“矩阵中”猎鹿博弈，图8对应重复进行的“矩阵中”胆小鬼博弈。

在本分析中，我们将最佳响应智能体的交互定义为合作：如果在交互时，其库存中的合作资源数量多于背叛资源数量。在矩阵中的猎鹿博弈中，双方合作（即双方都选择猎鹿）会带来更高的奖励，但这是一种风险更高的策略。背叛则能带来有保障的收益。双方合作或双方背叛都是纳什均衡，也就是说，没有一方有动机单方面偏离该策略。与背叛方合作的智能体获得的奖励为 0。在矩阵中的猎鹿博弈训练中，自私基线机器人学会了背叛。面对无条件背叛者，最佳响应是背叛。因此，使用自私基线机器人训练的最佳响应智能体学会了无条件背叛。相比之下，异质社会价值取向机器人群体由背叛者和合作方组成，他们的合作和背叛程度各不相同。使用社会价值取向机器人进行训练的最佳响应智能体会遇到合作方和背叛者，进而学会一种条件策略：倾向于与合作方合作，与背叛者背叛。

在胆小鬼博弈的矩阵中，两个纳什均衡分别是一个智能体合作（转向避让），另一个智能体背叛（直行）。在这种情况下，自私基线智能体学会了背叛和合作。因此，使用自私基线机器人训练的最佳响应智能体也学会了背叛和合作。然而，图8显示，这种行为并非有条件的。相比之下，使用社会价值取向机器人训练的最佳响应智能体遇到的机器人主要是合作型或背叛型，这使得最佳响应智能体学会了有条件的行为：它们倾向于与背叛者合作，并对合作方背叛。

发现4：囚徒困境中的失败案例 在囚徒困境的矩阵中，纳什均衡是两个智能体都背叛——
因此，自私基线智能体学会了无条件背叛。因此，与自私智能体一同训练的最佳响应智能体也学会了背叛。此外，背叛也是对无条件合作的最佳响应。由于社会价值取向机器人只学会了无条件策略（合作或背叛），对社会价值取向机器人的最佳响应同样是无条件背叛。图9对此进行了说明，显示所有最佳响应智能体无论其合作伙伴的合作程度如何，都在学习背叛。

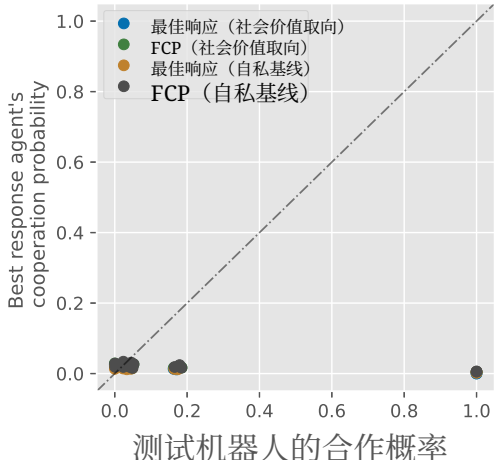


图9：比较最佳响应智能体在矩阵中的囚徒困境中学习条件策略的效果。

发现5：最佳响应智能体的零样本泛化表现更好表1列出了最佳响应智能体、自私基线智能体、随机智能体和利用者的零样本泛化性能。每种类型的智能体都会在每个测试场景中运行，并计算其平均回报。针对每个场景对智能体的得分进行归一化：最佳智能体得分为1，最差智能体得分为0。智能体的最终得分为其在所有场景中的平均得分。这与 Melting Pot [1] 中使用的方法相同。利用者和随机智能体旨在为所有环境中的性能提供近似的上限和下限。正如预期，表中显示利用者表现最佳，而随机智能体表现最差。在所有环境中，至少有一个最佳响应智能体的表现优于自私基线智能体，这表明学习最佳响应能够提升零样本泛化能力。

在矩阵情景的猎鹿博弈中，BR(SVO)的平均表现优于其他智能体。从图7中可以看到，BR(SVO)和FCP(SVO)与无条件背叛者合作的概率较低。然而，在矩阵中的猎鹿博弈里，与背叛者合作或对背叛者背叛，获得的奖励相同。因此，当遇到更可能背叛BR(SVO)的背叛者和测试机器人时，FCP(SVO)获得的奖励与BR(selfish-baseline)和FCP(selfish-baseline)相当。而当遇到更倾向于合作的测试机器人时，最佳响应智能体能够

	BR(SVO)	FCP(SVO)	BR(selfish-baseline)	FCP(selfish-baseline)	自私基线	随机策略	利用者
猎鹿博弈 ITMR	**0.876**	0.830	0.856	0.847	0.850	0.000	**0.988**
胆小鬼博弈 ITMR	0.696	0.668	**0.745**	0.723	0.723	0.000	**0.958**
囚徒困境 ITMR	0.738	0.702	0.777	**0.783**	0.754	0.000	**1.000**
外部性蘑菇	0.619	0.764	0.612	**0.846**	0.660	0.000	**0.900**

表1：最佳响应智能体、自私基线智能体、随机智能体和利用者的零样本泛化性能。分数的计算方法是：首先重新缩放每个智能体获得的奖励，使得在每个情景中，获得最高（最低）奖励的智能体得分为 1（0）；然后对每个环境中的所有情景取平均值。

适应伙伴行为并与合作者合作，从而获得更高的奖励。这使得BR(SVO)智能体在矩阵中的猎鹿博弈中得分更高。

表1显示，在矩阵情景中的胆小鬼博弈里，BR(selfish-baseline)的表现优于其他智能体。需要注意的是，在矩阵中的胆小鬼博弈里，与背叛方合作的智能体所获得的回报高于对背叛方进行背叛的智能体。从图8的结果可以看出，当测试机器人以接近1的概率进行背叛时，所有最佳响应智能体都以相近的概率进行合作。因此，在测试机器人无条件背叛的情景中，所有最佳响应智能体都能取得相当的表现。然而，当测试机器人以约0.4的概率进行合作时，与BR(SVO)和FCP(SVO)相比，BR(selfish-baseline)和FCP(selfish-baseline)以更高的匹配概率进行合作，从而使BR(selfish-baseline)和FCP(selfish-baseline)取得更好的表现。在最佳响应智能体遇到无条件合作方的情景中，BR(SVO)和FCP(SVO)以接近1的概率进行背叛，其表现优于BR(selfish-baseline)和FCP(selfish-baseline)。由于大多数测试场景由背叛者或更倾向于背叛的测试机器人构成，因此BR(selfish-baseline)的表现优于BR(SVO)和BR(FCP)智能体。

请回想一下，图9显示，所有最佳响应智能体面对所有测试机器人时都会选择背叛。因此，我们预计，表1所示矩阵中最佳响应智能体在囚徒困境中的表现得分会相近。然而，令人意外的是，BR(selfish- baseline) 和 FCP(selfish-baseline) 的表现优于BR(SVO) 和FCP(SVO)。我们将对此展开调查，留待未来工作。

在外部性蘑菇环境中，FCP 类型的最佳响应智能体表现优于仅使用共同玩家最终策略进行训练的最佳响应智能体。这表明，在训练期间既遇到能力较弱的智能体、也遇到能力较强的智能体的最佳响应智能体，其表现优于训练期间只遇到能力较强智能体的最佳响应智能体。

4 讨论

本文研究了异质社会价值取向对序贯社会困境中不同激励结构的影响。我们检验了异质SVO是否会带来多样化策略，以及学习对这些策略的最佳响应能否提升零样本泛化能力。研究发现，异质SVO的确会使策略呈现出可测量的多样性，而这种多样性有时能让对其作出最佳响应的智能体获得更好的零样本泛化性能。

最佳响应智能体通过学习一种条件策略，在测试时适应新型智能体，从而取得更好的表现。研究还发现，当序贯社会困境并非均衡选择问题时，这种方法仍会产生有意义的策略多样性，但无法实现更好的零样本泛化性能。这是因为，对一组多样化策略作出的最佳响应会收敛为一种无条件策略，而这种策略在测试时遇到条件策略时表现不佳。

我们的研究发现有助于理解异质性的社会价值取向如何影响激励结构和策略多样性，以及智能体如何在测试时学会适应多样化策略，从而实现更好的零样本泛化性能。我们的研究发现为理解智能体在序贯社会困境中的行为提供了新见解，并凸显了在设计激励结构时考虑社会价值取向异质性作用的重要性。

我们观察到，社会价值取向智能体能够在我们测试的所有环境中学习合作策略。这表明，使用社会价值取向来捕捉使智能体至少在某些方面与人类价值观保持一致所必需的因素，具有潜在价值。

致谢

作者感谢 Daniel Hennes、DJ Strouse、MarcLanctot、Alexander (Sasha) Vezhnevets、Julien Perolat、Andrea Tacchetti 和 Ian Gemp 提供的有益讨论。

参考文献

[1] John P Agapiou, Alexander Sasha Vezhnevets, Edgar A Duéñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu 等。2022。Melting Pot 2.0。*arXiv* 预印本 *arXiv:2211.13746*（2022）。[2] David Balduzzi、Marta Garnelo、Yoram Bachrach、Wojciech Czarnecki、Julien Perolat、Max Jaderberg 和 Thore Graepel。2019。在对称零和博弈中的开放式学习。载于国际机器学习会议。PMLR，434–443。[3] Kenneth Derek 和 Phillip Isola。2021。通过策略生成模型实现可适应的智能体群体。载于神经信息处理系统进展 34（2021），3902–3913。[4] Edgar A Duéñez-Guzmán、Kevin R McKee、Yiran Mao、Ben Coppin、Silvia Chiappa、Alexander Sasha Vezhnevets、Michiel A Bakker、Yoram Bachrach、Suzanne Sadedin、William Isaac 等。2021。学习智能体中的统计歧视。*arXiv* 预印本 *arXiv:2110.11404*（2021）。[5] Lasse Espeholt、Hubert Soyer、Remi Munos、Karen Simonyan、Vlad Mnih、Tom Ward、Yotam Doron、Vlad Firoiu、Tim Harley、Iain Dunning 等。2018。IMPALA：采用重要性加权的行动者-学习者架构，实现可扩展的分布式深度强化学习。载于国际机器学习会议。PMLR，1407–1416。[6] Marta Garnelo、Wojciech Marian Czarnecki、Siqi Liu、Dhruva Tirumala、Junhyuk Oh、Gauthier Gidel、Hado van Hasselt 和 David Balduzzi。2021。选择你的战场：作为策略多样性群体层面目标的交互图。载于第*20*届自主智能体与多智能体系统国际会议论文集。1501–1503。[7] Donald W Griesinger 和 James W Livingston Jr. 1973。实验博弈中人际动机模型的探索。《行为科学》18卷3期（1973），173–188。[8] Matteo Hessel、Hubert Soyer、Lasse Espeholt、Wojciech Czarnecki、Simon Schmitt 和 Hado van Hasselt。2019。采用 PopArt 的多任务深度强化学习。载于*AAAI*人工智能会议论文集，第33卷。3796–3803。[9] Hengyuan Hu、Adam Lerer、Brandon Cui、Luis Pineda、Noam Brown 和 Jakob Foerster。2021。非信念学习。载于国际机器学习会议。PMLR，4369–4379。[10] Hengyuan Hu、Adam Lerer、Alex Peysakhovich 和 Jakob Foerster。2020。“Other-Play”用于零样本协同。载于国际机器学习会议。PMLR，4399–4410。[11] Edward Hughes、Joel Z Leibo、Matthew Phillips、Karl Tuyls、Edgar Dueñez-Guzman、Antonio García Castañeda、Iain Dunning、Tina Zhu、Kevin McKee、Raphael Koster 等。2018。不公平厌恶提升跨期社会困境中的合作。载于神经信息处理系统进展31（2018）。[12] Joel Z Leibo、Edgar A Dueñez-Guzman、Alexander Vezhnevets、John P Agapiou、Peter Sunehag、Raphael Koster、Jayd Matyas、Charlie Beattie、Igor Mordatch 和 Thore Graepel。2021。使用 Melting Pot 对多智能体强化学习进行可扩展评估。载于国际机器学习会议。PMLR，6187–6199。[13] Joel Z Leibo、Vinicius Zambaldi、Marc Lanctot、Janusz Marecki 和 Thore Graepel。2017。序贯社会困境中的多智能体强化学习。载于第*16*届自主智能体与多智能体系统会议论文集。464–473。[14] Adam Lerer 和 Alexander Peysakhovich。2017。使用深度强化学习在复杂社会困境中维持合作。*arXiv* 预印本 *arXiv:1707.01068*（2017）。[15] Wim BG Liebrand 和 Charles G McClintock。1988。社会价值环形测量：一种评估信息处理个体差异和社会价值取向的计算机化程序。《欧洲人格杂志》2卷3期（1988），217–230。[16] Michael L Littman。1994。作为多智能体强化学习框架的马尔可夫博弈。载于《机器学习论文集》*1994*。Elsevier，157–163。[17] Andrei Lupu、Brandon Cui、Hengyuan Hu 和 Jakob Foerster。2021。用于零样本协同的轨迹多样性。载于国际机器学习会议。PMLR，7204–7213。[18] Kevin R McKee、Xuechunzi Bai 和 Susan T Fiske。2022。人类-智能体合作中的温暖与能力。载于第*21*届自主智能体与多智能体系统国际会议论文集。898–907。[19] Kevin R McKee、Ian Gemp、Brian McWilliams、Edgar A Duèñez-Guzmán、Edward Hughes 和 Joel Z Leibo。2020。混合动机强化学习中的社会多样性与社会偏好。载于第*19*届自主智能体与多智能体系统国际会议论文集。869–877。[20] Kevin R McKee、Joel Z Leibo、Charlie Beattie 和 Richard Everett。2022。量化环境与群体多样性对多智能体强化学习的影响。《Autonomous Agents and Multi-Agent Systems》36卷1期（2022），1–16。[21] Volodymyr Mnih、Adria Puigdomenech Badia、Mehdi Mirza、Alex Graves、Timothy Lillicrap、Tim Harley、David Silver 和 Koray Kavukcuoglu。2016。深度强化学习的异步方法。载于国际机器学习会议。PMLR，1928–1937。[22] Ryan O Murphy、Kurt A Ackermann 和 Michel JJ Handgraaf。2011。测量社会价值取向。《Judgment and Decision Making》6卷8期（2011），771–781。[23] Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。2018。基于对比预测编码的表征学习。*arXiv* 预印本 *arXiv:1807.03748*（2018）。[24] Nicolas Perez-Nieves、Yaodong Yang、Oliver Slumbers、David H Mguni、Ying Wen 和 Jun Wang。2021。为开放式游戏中的学习建模行为多样性。载于国际机器学习会议。PMLR，8514–8524。[25] Alexander Peysakhovich 和 Adam Lerer。2018。信息不完全的社会困境中的结果主义条件合作（简短的研讨会版本）。载于第三十二届*AAAI*人工智能会议研讨会。[26] Anatol Rapoport。1974。囚徒困境——回忆与观察。载于博弈论：作为冲突解决理论。Springer，17–34。[27] Wilko Schwarting、Alyssa Pierson、Javier Alonso-Mora、Sertac Karaman 和 Daniela Rus。2019。自动驾驶汽车的社会行为。载于美国国家科学院院刊 116 卷50期（2019），24972–24978。[28] Lloyd S Shapley。1953。随机博弈。载于美国国家科学院院刊39卷10期（1953），1095–1100。[29] DJ Strouse、Kevin McKee、Matt Botvinick、Edward Hughes 和 Richard Everett。2021。无需人类数据即可与人类协作。载于神经信息处理系统进展34（2021），14502–14515。[30] Zhenggang Tang、Chao Yu、Boyuan Chen、Huazhe Xu、Xiaolong Wang、Fei Fang、Simon Shaolei Du、Yu Wang 和 Yi Wu。2021。通过奖励随机化发现多样化的多智能体策略行为。载于国际学习表征会议。[31] Jane X Wang、Edward Hughes、Chrisantha Fernando、Wojciech M Czarnecki、Edgar A Duéñez-Guzmán 和 Joel Z Leibo。2018。为利他行为演化内在动机。*arXiv* 预印本 *arXiv:1811.05931*（2018）。[32] Jörgen W Weibull。1997。演化博弈论。麻省理工学院出版社。[33] Tom Zahavy、Yannick Schroecker、Feryal Behbahani、Kate Baumli、Sebastian Flennerhag、Shaobo Hou 和 Satinder Singh。2022。使用 domino 发现策略：在维持近最优性的同时优化多样性。*arXiv* 预印本 *arXiv:2205.13521*（2022）。