arXiv:2301.07416v1 [cs.MA] 18 Jan 2023

通过奖励份额交易学习参与

Michael K¨olle[1], Tim Matheis[1],Philipp Altmann[1],Kyrill Schmid[1]l慕尼黑大学信息学研究所，厄廷根大街67号，德国慕尼黑{*michael.koelle, philipp.altmann, kyrill.schmid*}*@ifi.lmu.de*

关键词: 多智能体系统、强化学习、社会困境

摘要: 使自主智能体能够协同采取行动，是将人工智能融入日常生活的重要一步。一些方法通过让智能体向其他智能体提供奖励来促进合作，而本文提出了一种受股票市场启发的方法：智能体可以通过购买奖励份额，参与其他智能体的收益。直观而言，当智能体直接受到其他智能体奖励的影响时，它可能会学会根据共同利益采取行动。对所测试的一般和马尔可夫博弈进行的实证结果表明，这一机制能够促进独立训练的智能体在社会困境情境中形成合作策略。此外，正如在一个时空扩展领域中所展示的，参与能够促使角色的形成以及智能体之间的子任务分工。

1 引言

合作式人工智能领域旨在探索能够促使彼此独立的自主智能体之间形成合作行为的方法（Dafoe 等，2020年）。要将人工智能融入我们的日常生活，具备合作行动的能力是必不可少的一步，尤其是在自动驾驶等不同决策者相互交互的应用中。单智能体领域取得的诸多突破（Mnih 等，2015年；Silver 等，2017年），也推动了强化学习在多智能体系统中的成功应用（Leibo 等，2017年；Phan 等，2018年；Vinyals 等，2019年）。然而，对于所有智能体获得相同奖励、因而追求相同目标的纯合作场景，可以通过集中式训练技术加以处理；但如果智能体拥有各自的奖励和目标，情况便并非如此。此外，如果智能体共享资源，它们很可能学会不良行为，尤其是在资源日益短缺时（Leibo 等，2017年）。独立优化可能导致次优结果，例如囚徒困境或公共物品博弈中的情况。在更复杂的博弈中，智能体的风险规避倾向以及信息不对称——

尝试进一步降低达成理想结果的可能性（Schmid 等人，2020）。

近年来，人们提出了多种促进独立智能体之间合作的方法，例如学习经过验证的博弈论策略，如以牙还牙（Lerer 和 Peysakhovich，2018），

	C	D
C	(-1, -1)	(-3, 0)
D	(-3, 0)	(-2, -2)

	C	D
C	(-1, -1)	(-1.5,-1.5)
D	(-1.5,-1.5)	(-2, -2)

(a) 无参与的囚徒困境 (b) 参与率为 50% 的囚徒困境

图1：通过允许智能体交易其收益中的份额，或许可以实现社会最优结果。

让智能体能够相互激励、提高合作意愿（Schmid 等，2018；Lupu 和 Precup，2020；Yang 等，2020），或引入市场，让智能体通过交易提升整体福利（Schmid 等，2021b）。在这项工作中，我们采用市场这一概念，以增进独立决策者之间的合作。更具体地说，我们提出一种方法，使智能体能够交易自身奖励的份额。我们认为，在存在此类参与市场的情况下，让智能体直接参与其他智能体的奖励，可以实现更好的均衡。由于直接纳入了所有全局奖励，而非仅纳入个体奖励，因此有望形成社会最优均衡。如何让智能体以公平价格交易其份额，同时建立一条确实能带来有利份额分配的交易路径，是实施这一模型时最具挑战性的问题。

为说明参与的有效性，考虑图1所示的囚徒困境。在标准版本的囚徒

困境中，每个智能体都会获得自己的个体奖励。对于两个理性的决策者而言，唯一的纳什均衡是相互背叛（DD），这在社会层面并不可取。然而，如果引入奖励交易的能力，并且每个智能体都持有其他智能体回报的份额（例如，参与比例为50%），那么整体动态就会发生变化，相互合作将成为优势策略。为了训练智能体参与，我们采用不同的强化学习方法，使智能体能够自主找到策略。由此，智能体参与其他智能体回报的过程得以在多个训练回合中实现。为此，本研究测试了参与市场的不同变体，这些变体在参与机制的实现方式上有所不同。具体而言，本文作出以下贡献：

- 本文从理论上阐述了市场参与为何有益。
- 本文在囚徒困境以及一种称为清理游戏的复杂多智能体场景中，对参与机制进行了实证评估（Yang 等，2020）。

实验的所有代码都可以在此处找到 [1]。

2 相关工作

尽管强化学习在不断扩展的任务集合中取得了日益显著的成功，但大多数研究都集中于单智能体环境以及完全合作的多智能体环境（Mnih 等，2015年；Silver 等，2017年；Vinyals 等，2019年；OpenAI 等，2019年）。然而，当涉及多个智能体时，它们的目标往往并非（完全）一致，因此通常无法采用集中式训练技术。完全去中心化模型的缺点在于，智能体只关注各自的个体奖励，这可能导致不理想的集体表现，尤其是在社会困境或涉及公共池资源的情形下（Leibo 等，2017年；Perolat 等，2017年）。

此前，解决这一问题的一种方法是为独立智能体提供内在奖励（Eccles 等人，2019；Hughes 等人，2018；Wang 等人，2019）。内在奖励的概念源自行为经济学中的一些概念，例如利他行为、互惠、不公平厌恶、社会影响和同伴评价。这些内在奖励通常是预先设定的，或会根据其他智能体在游戏中的表现而演变。其他研究表明，奖励机制（Yang 等，2020）或惩罚机制（Schmid 等人，2021a）可能促成序贯社会困境中的合作。文献将选择性激励与制裁机制区分开来，二者都旨在促进社会困境中的合作行为（Kollock，1998）。选择性激励指的是试图积极促进合作的方法。例如，可以通过提供金钱奖励，减少对水或电力等公共池物品的消耗（Maki 等人，1978）。相反，惩罚可以用来减少背叛行为。事实上，人类实验表明，惩罚能有效减少背叛行为（Komorita，1987）。

在 *LIO* （Yang 等，2020）中，奖励提供者的激励函数与策略学习在同一时间尺度上进行学习。引入激励函数偏离了经典强化学习的做法；在经典强化学习中，奖励函数专属于环境，且只会因外部因素而改变。实证研究（Lupu 和 Precup，2020）表明，在智能体的动作空间中加入“给予奖励”动作，可以在某些训练阶段提升合作。通过对手塑造，智能体可以影响其他智能体的学习更新，从而使自身受益。另一种对手塑造方法是考虑一个智能体的策略对其他智能体预期参数更新的影响（Foerster 等，2018）。通过在 *LOLA* 中加入这一额外的学习组件，在重复囚徒困境中可以形成以牙还牙等策略，从而维持合作。

该领域的其他研究表明，市场可以作为促进合作的载体（Schmid 等，2018；Schmid 等，2021b）。通常，智能体只能获得自身的个体奖励。由于其他智能体的个体奖励不会影响它们，它们只会追求自身利益。然而，只获得个体奖励会使智能体面临重大风险。根据经济学理论，实现多元化通常是有益的。在投资组合理论中，人们普遍认为，多元化能够提高预期收益。尽管人们似乎并未做到足够的多元化，这种现象被称为多元化之谜（Statman，2004年），但理性智能体应当实现完全多元化，只持有市场投资组合与无风险资产（例如安全的政府债券）的组合。将这一点应用于囚徒困境等博弈时，智能体应当希望获得自身个体奖励与其他智能体个体奖励的组合，以将自身的风险敞口降至最低。

[1] https://github.com/TimMatheis/Learning-to-participate

3 学习参与

多智能体系统由共享同一环境的多个智能体组成。智能体是一种具备两项主要能力的自主实体：感知和行动。对当前环境状态的感知使智能体能够从动作集中选择适当的动作。所选择的动作取决于智能体的策略。在多智能体强化学习中，人们通常应用强化学习方法来教导智能体形成良好的策略。

由于合作策略可能难以找到并维持，我们提出了一种参与机制。其思路是让智能体直接参与其他智能体的环境奖励，以协调它们原本相互冲突的目标。下文中，我们将参与份额的交易作为实现合作的手段。如果智能体愿意持有所有智能体奖励中相当数量的份额，它们就可能以“社会利益”为行动准则。也就是说，单个智能体的个体利益与全体智能体的共同利益之间的差异可能会消失。在我们的实现中，只有当两个智能体都选择增加或减少各自自有股份的数量时，它们才会相互交易份额。在初始状态下，智能体持有自己奖励的 100% 份额。如果智能体希望实现完全多元化，经过若干次交易后，每个智能体都可以持有每个智能体奖励的 $^{100}\%_{n}$ 份额，其中n 表示环境申的智能体数量。

4 实验

本节通过实验考察我们关于重复囚徒困境和清理游戏的理论考量。

4.1 重复囚徒困境

囚徒困境可以看作两个窃贼的处境。他们被抓获后，可以选择认罪$_0$ = 或不$_1$ =认罪。如果两人都认罪，就会受到重罚。如果两人都不认罪，就只会受到轻罚。困境源于只有一人认罪的情况：认罪者因成为主要证人而免于惩罚，而否认者则会受到极重的惩罚。在图1中，认罪是背叛（D）行动，不认罪则是合作（C）行动。根据定义，在纳什均衡中，任何智能体都无法通过改变策略让自身处境变得更好。智能体 1 知道智能体 2 可以选择背叛或合作。如果智能体 2 背叛，智能体 1 也背叛会更有利。如果智能体 2 选择合作，智能体 1 仍然是背叛更有利。因此，无论如何，智能体 1 都应当背叛，这使背叛成为单次博弈中的优势策略。根据对称性，智能体 2 面临同样的问题，也应当背叛。可悲的是，这种结果并不理想，因为双方合作会让两个智能体都获得更高的收益。当博弈重复进行多次时，从理论上说，背叛未必始终是优势策略。尽管通过逆向归纳法，智能体可能每一轮都选择背叛，但它们也可以制定策略来激励彼此合作。

4.2 重复囚徒困境中的参与

为了测试重复囚徒困境的不同实现，我们采用演员-评论家方法。我们改编了*LIO* 中的算法 1（Yang 等，2020，第 5 页），将激励函数及其参数 η 替换为整个回合中更大的动作空间，或预先执行交易动作。因此，动作要么在整个回合中同时指定环境动作和交易动作，要么在每个回合开始时增加一个步骤以确定参与情况，也可能两者兼有。

我们使用与*LIO*相同的全连接神经网络进行函数逼近。不过，我们只使用策略网络，因为我们不使用激励函数，而*LIO* 为此另设了一个神经网络。所有环境中的离散动作均采用具有 Softmax 输出的策略网络。所有实验均使用相同的架构和超参数：$\beta = 0.1$, $\varepsilon_{\text{start}} = 1.0$, $\varepsilon_{\text{end}} = 0.01$,$\alpha_{\theta} = 1.00E-03$。由于智能体在开始时不会参与其他智能体的奖励分配，因此在采用交易的实现中，最大步数设为 40，而不是 5。这样它们就有足够的步数进行交易。我们测试了以下两智能体重复囚徒困境实现。

（i）不参与： 每个智能体可能采取的环境动作是合作 和背叛。在实现中，这两个动作分别编码为 0 和 1。共有四种可能的状态：（合作，合作）、（合作，背叛）、（背叛，合作）、（背叛，背叛）。如果智能体随机选择动作，平均个体奖励将为（$-1+0-3-2$）$/4=-1.5$。但如果智能体学会维持一种双方都选择合作的合作均衡，个体奖励将收敛于 -1。然而，根据理论，在“经典”囚徒困境中，背叛 是占优动作

。在这种情况下，个体奖励将收敛于 -2。事实上，在没有参与份额的实现中，会形成一种双方智能体都背叛的稳定均衡。累积奖励收敛于 $-2+(-2)=-4$（图 2a）。

（ii）个体奖励的平均分配：在此实现中，智能体始终获得所有个体份额的平均奖励。具体来说，就是将所有个体奖励汇总后，再除以智能体的数量：每个智能体获得 $\times\sum_i^n$ 奖励$_i$ 的 1_n。动作空间及其他一切都保持不变。智能体无法选择是分享奖励，还是获得个体奖励。每个智能体都会学会合作，从而获得 $-1+(-1)=-2$ 的累积奖励。在此实现中，合作是占优动作。此实现表明，分享奖励可以促成合作。然而，它并未表明当智能体能够自由交易时，是否可以主动找到并维持这样的均衡（图 2a）。

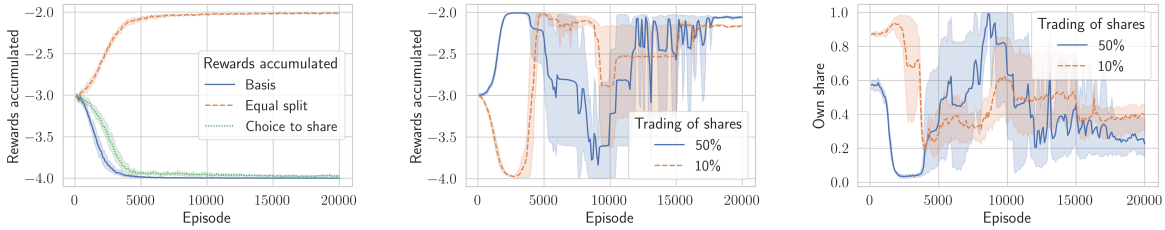
（iii）选择是否分享奖励：智能体可以决定是否分享各自的个体奖励。动作空间随之扩展。除了两种环境动作之外，智能体还可以选择分享奖励或获得个体奖励。动作空间由一个动作元组定义：环境动作和交易动作。因此，现在每个智能体有 $2\times 2=4$ 种可能的动作。重要的是，单个智能体无法决定合并后的奖励是否会在两个智能体之间平均分配。只有当两个智能体都决定分享各自的奖励时，才会平均分配。状态空间扩展为八种状态：分享 $\times$环境动作 $1\times$环境动作 $2=2\times 2\times 2=8$。两个智能体都会学会背叛，从而获得 $-2+(-2)=-4$ 的累积奖励（图 2a）。在此实现中，不分享而选择背叛是占优动作。因此，仅仅提供分享奖励的机会还不够。

（iv）交易50%的份额：起初，两个智能体都不持有对方的参与份额。每一步中，它们可以从六种行动中进行选择。一种行动可视为一个三元组：环境行动、是否增加自有奖励的份额，以及是否增加对方奖励的份额。当智能体决定合作时，有三种可能的行动：（合作，不购买自有份额，不购买对方的份额）、（合作，购买自有份额，不购买对方的份额）、（合作，不购买自有份额，购买对方的份额）。同理，背叛也有三种可能的行动。再次强调，只有在两个智能体都打算交易时，份额交易才会执行。例如，如果双方都选择购买自有份额，且各自持有对方奖励的50%，那么它们就会交换份额，此时各自持有100%的自有份额。这意味着不再存在参与。然而，如果它们此时再次选择购买自有份额，就无法进行这笔交易，因为它们已经持有全部自有份额。因此，只有环境行动会产生影响。这意味着所有份额的价格相同，只是按照相同的比例进行交换，而且每个智能体的份额总和始终为100%。共有36种状态：环境行动 $\times$ 自有份额比例 $\times$交易 $=4\times 3\times 3$。自有份额比例可以是0、0.5或1。“交易”变量表示没有交易、导致自有份额增加的交易，或导致自有份额减少的交易。这种实现成功促成了合作。累积奖励收敛至 $-1+(-1)=-2$（图2b）。然而，学习过程耗时较长。此外，行动和奖励最初会偏离到先前的低效均衡。

(v) 交易 10% 的份额：与交易 50% 的份额相比，状态空间扩大到 $4\times 11\times 3=132$，因为自有股份的比例现在可以是 0、0.1、0.2、.....、1。同样，可以建立一个社会最优均衡。约 40% 的自有股份和 60% 的他人股份似乎能有效防止智能体转而背叛（图 2c）。

4.3 清理游戏

在清理游戏中，多个智能体会同时尝试在同一环境中收集苹果。智能体每收集一个苹果，就会获得奖励 +1 。苹果会随机生成在一张方形*7x7* 或 *10x10* 地图的右侧。地图左侧有一条河流，废弃物会不断污染河流。随着废弃物水平上升并接近耗竭阈值，苹果生成率会线性下降至零。为避免游戏迅速结束，智能体可以发射清洁光束 清除废弃物，但只有身处河流中时才能这么做。清洁光束会清除智能体上方的所有废弃物。每个智能体的观测都是整张地图的自我中心视角 RGB 图像。困境在于，待在河流中发射清洁光束清除废弃物，远不如收集苹果以获得奖励来得有吸引力。然而，如果所有智能体都只专注于收集苹果，游戏就会很快结束，智能体获得的总奖励也会很低。因此，这是一种跨期



(a) 奖励 (b) 交易 10% 或 50% 份额后的奖励份额 (c) 自身奖励中的份额

图 2：重复囚徒困境中参与与不参与的结果。

社会困境，其中短期个人激励与长期集体利益之间存在权衡（Hughes 等人，2018年）。这一领域尤其有趣，因为基于行为经济学的模型只能解释简单且不切实际的无状态矩阵博弈中的合作。相比之下，清理游戏是一种在时间和空间上均有延展的马尔可夫博弈。

4.4 两个智能体参与清理游戏

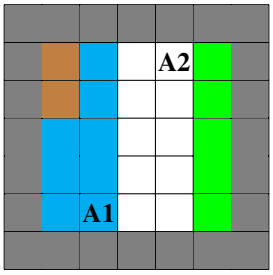


图3：包含两个智能体的 *7x7* 地图。智能体 1 在河流旁生成，而智能体 2 在苹果附近生成。如果它们不分工，游戏很可能很快就会停止，因为如果智能体不清理废弃物，废弃物就会向下蔓延。

如（Yang 等，2020）所述，此处使用较小的 *7x7* 地图。游戏初始状态如图 3 所示。地图左侧的蓝色格子代表河流，绿色格子表示苹果随机生成的区域，棕色格子则代表废弃物。智能体 1 在河流中生成，因此能够清理废弃物。智能体 2 在右侧靠近绿色区域的位置生成，苹果会在该区域随机生成。这已经提示了一种可能的策略：智能体 1 清理废弃物，智能体 2 收集苹果。然而，只要智能体 1 没有机会收集苹果，也无法从智能体 2 的奖励中获益，它就没有充分理由清理废弃物。

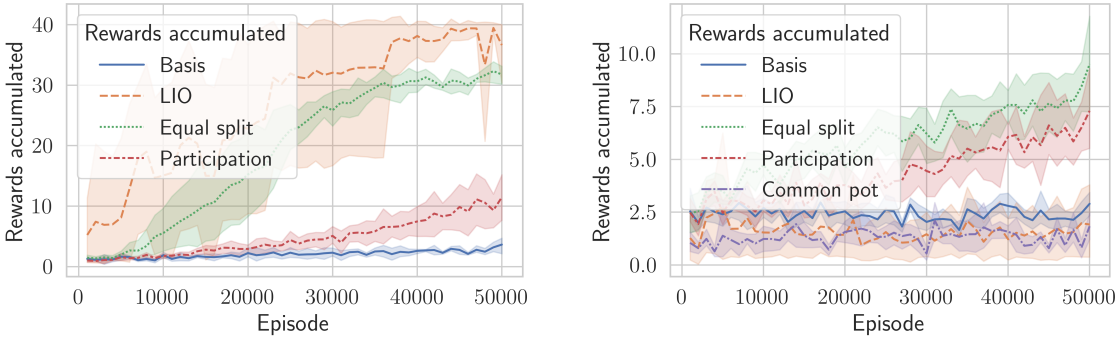
智能体 2 也面临类似的两难处境。如果它待在绿色区域收集苹果，就无法发射清洁光束。若智能体之间没有分工，由于蓝色区域和绿色区域相距太远，游戏就会提前结束。每个智能体可能采取的环境动作包括向左移动、向右移动、向上移动、向下移动、不操作和清理。在所有清理领域的实现中，均采用演员-评论家方法。优化过程是去中心化的。

我们使用与 *LIO* 相同的卷积网络来处理 Cleanup 中的图像观测。在所有环境中，策略网络都通过 Softmax 输出给出离散动作。所有实验均使用相同的神经网络架构。我们测试了以下两个智能体参与清理游戏的实现方式。

(i) 不参与： 如果没有任何额外的激励机制，这一社会困境似乎无法解决。如图4a所示，累积奖励仍处于非常低的水平。合作对任何智能体都没有吸引力。由于苹果生成的位置离河流很远，因此很难学会清理废弃物能够创造价值。如果智能体从河流返回苹果所在处，那么它甚至还没来得及采到一个苹果，游戏就很可能已经停止了。

（ii）LIO： 在基线情景中，新增了每个智能体都可以向另一个智能体提供奖励这一可能性，将其作为额外的合作渠道。重要的是，智能体i向智能体j支付的额外奖励$r_{\eta}^{j}{}_{i}$，是通过沿着智能体自身外在目标的梯度直接上升来学习的，其中考虑了该支付对所有其他智能体策略的影响。因此，这些支付不属于强化学习，不会影响强化学习智能体的动作空间（而非将其扩展）。智能体成功地分工：离河流和废弃物较近的智能体成为清洁者，而离苹果较近的智能体则专门负责收集苹果。实现细节见 Yang 等，2020。

（iii）个体奖励的平均分配 ：



(a) 两个智能体

(b) 三个智能体

图4：清理游戏的结果

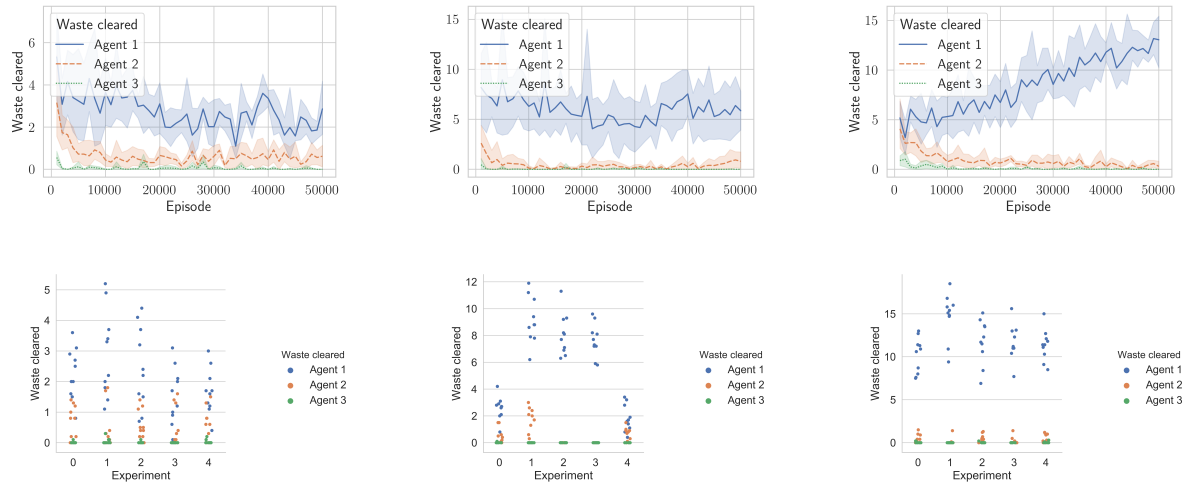
在此情景中，基线情景经过扩展：每一步之后，智能体之间平均分配联合奖励。这项额外计算不会影响状态空间或动作空间。与*LIO*类似，距离河流较近的智能体转而担任清洁者，而距离苹果较近的智能体则专门负责收集苹果。双方都能从这种任务分工中获益，因为无论由谁收集到苹果，双方都会均分相应奖励。与*LIO*相比，学习过程的波动性较小。

(iv) 参与权预交易：在这一情景中，基线情景的每个回合都增加了一个额外的首步骤。在这一步中，两个智能体都可以从六个动作（0-5）中进行选择，分别代表 0%、20%、40%、60%、80% 和 100%。两个智能体所选数值中的较大值，决定了他们在整个回合中参与自身奖励的比例。剩余比例则代表他们参与另一个智能体个体奖励的比例。例如，如果智能体 1 选择 40%，而智能体 2 选择 80%，那么他们将获得各自个体奖励的 80%，以及对方个体奖励的 20%。这个额外的首步骤属于强化学习的一部分，但此步骤不发放奖励。增加交易步骤的目的是让两个智能体都可以通过选择 100% 来避免合作。取两者中的最大值，意味着执行更保守的动作。同样，智能体成功地分配了各自的任务。智能体 1 清理的废弃物数量与累积奖励同步变化。然而，联合奖励仅达到约 11，尚不清楚这是否是一个稳定水平。

5 结论

通过引入参与其他智能体奖励的理念，我们提出了一种在共享多智能体环境中实现协调与合作的新方法。智能体会学到，直接参与其他智能体的奖励能够强化那些优化全局目标的策略。借助这一机制，无需像 *LIO* （Yang 等，2020）那样引入额外的外在激励结构。其他早期研究则聚焦于内在激励（Eccles 等人，2019；Hughes 等人，2018；Wang 等人，2019），而采用这种方法也无需这些激励。尤其是，这种对标准模型简单易懂的扩展，使参与机制颇具吸引力。在测试的两个社会困境问题——重复囚徒困境和 Cleanup——中，智能体利用通过份额参与的机会，探索合作行为。事实上，这一机制有效促成了 Cleanup 中的劳动分工。没有参与机制，智能体无法学会将任务划分为子任务，因为清理废弃物不会直接带来奖励。但通过参与，清理废弃物带来的积极影响可以通过其他智能体收集苹果所获得的奖励直接体现出来。重要的是，其他智能体只有在事先清理废弃物后才能收集苹果。所提出的参与方法能够在囚徒困境中实现最优的集体表现。在 Cleanup 中，参与机制带来的集体表现提升显而易见。虽然尚不清楚智能体是否彻底学会了最佳参与分配，但它们能够随着时间推移实现协调并提升表现。

我们的方法试图回答有关去中心化多智能体群体中合作的一些未解问题。首先，尽管智能体除了学习环境动作外，还必须同时学习如何通过份额参与，但学习持有其他智能体的份额，可能比在其他智能体只顾自身利益行事时对其进行惩罚更简单



(a) 在基本实现中，没有智能体学会清理废弃物。

(b) 在 LIO 实现中，没有出现清理策略。

(c) 通过参与，其中一个智能体学会清理废弃物。

图5：在基本实现、LIO 和参与机制中，三个智能体进行清理时清除的废弃物。

。其次，惩罚或其他激励性奖励应当事先获得。使用份额时，可以直接进行交易而无需支付成本。第三，股份市场可以通过多种方式实现。我们确信，在大多数社会困境中都能找到合适的市场结构。此外，*LIO*假设接受者无法拒绝激励，但智能体可以只接受部分激励。对于份额而言，这并不是问题，因为相应的奖励只需根据所持份额进行分配。参与方法的另一项优势是，智能体没有明确的策略可以利用额外的市场功能来剥削其他智能体。关于奖励篡改的研究正逐渐兴起（Everitt 等人，2021年），参与市场或许能为部署安全应用指明正确方向。

我们的工作旨在促进在具有独立智能体的环境中确保共同利益。尽管参与机制在经过测试的社会困境中表现良好，但在其他环境中是否同样如此仍不明确。另一个悬而未决的问题是，智能体是否总能找到份额的稳定分配。我们建议开展实验：由经纪人设定价格，并结合限价订单簿来匹配需求和供给。此外，还需要在拥有更多智能体的其他环境以及其他博弈结构中测试参与机制。

参考文献

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., Mc- Kee, K. R., Leibo, J. Z., Larson, K., and Graepel, T. (2020). 合作人工智能中的开放问题。*CoRR*, abs/2012.08630. Eccles, T., Hughes, E., Kram´ar, J., Wheelwright, S., and Leibo, J. Z. (2019). 复杂序贯社会困境中的互惠学习。*arXiv:1903.08082 [cs]*. arXiv: 1903.08082. Everitt, T., Hutter, M., Kumar, R., and Krakovna, V. (2021). 强化学习中的奖励篡改问题及其解决方案：因果影响图视角。*Synthese*, 198(27):6435–6467. Foerster, J. N., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. (2018). 具备对手学习感知的学习。*arXiv: 1709 04326 [ ] cs*. arXiv: . . 1709 04326 Hughes, E., Leibo, J. Z., Phillips, M. G., Tuyls, K., Du´e˜nez- Guzm´an, E. A., Casta˜neda, A. G., Dunning, I., Zhu, T., McKee, K. R., Koster, R., Roff, H., a nd Graepel, T. (2018). 不公平厌恶能改善跨期社会困境中的合作。*arXiv:1803.08884 [cs, q-bio]*. arXiv: 1803.08884. Kollock, P. (1998). 社会困境：合作的剖析。*Annual Review ofSociology*, 24(1):183–214. eprint: https://doi.org/10.1146/annurev.soc.24.1.183. Komorita, S. S. (1987). 分解式社会困境中的合作选择。《人格与社会心理学通报》, 13(1):53–63. Publisher: SAGE 出版公司。 Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. (2017). 序贯社会困境中的多智能体强化学习。*arXiv: 1702 03037 [ ] .cs*. arXiv: . .

Lerer, A. and Peysakhovich, A. (2018). 使用深度强化学习维持复杂社会困境中的合作。*arXiv:1707.01068 [cs]*. arXiv: 1707.01068. Lupu, A. and Precup, D. (2020). 多智能体强化学习中的赠予。新西兰, 第9页。Maki, J. E., Hoffman, D. M., and Berk, R. A. (1978). 节水宣传活动影响的时间序列分析。《*Evaluation Quarterly*》, 2(1):107–118. Publisher: Sage Publications Sage CA: 千橡市, CA. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Ve- ness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Ku- maran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). 通过深度强化学习实现人类水平的控制。《*Nature*》, 518(7540):529–533. Bandiera abtest: a Cg type: Nature Research Jour- nals Number: 7540 Primary atype: Research Pub- lisher: Nature Publishing Group Subject term: 计算机科学 Subject term id: computer-science. OpenAI, Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., J´ozefowicz, R., Gray, S., Ols- son, C., Pachocki, J., Petrov, M., Pinto, H. P. d. O., Raiman, J., Salimans, T., Schlatter, J., Schneider, J., Sidor, S., Sutskever, I., Tang, J., Wolski, F., and Zhang, S. (2019). 使用大规模深度强化学习玩《Dota 2》。*arXiv:1912.06680 [cs, stat]*. arXiv: 1912.06680. Perolat, J., Leibo, J. Z., Zambaldi, V., Beattie, C., Tuyls, K., and Graepel, T. (2017). 公共池资源占用的多智能体强化学习模型。*arXiv:1707.06600 [cs, q-bio]*. arXiv: 1707.06600. Phan, T., Belzner, L., Gabor, T., and Schmid, K. (2018). 利用统计多智能体在线规划与涌现式价值函数近似。*arXiv:1804.06311 [cs]*. arXiv: 1804.06311. Schmid, K., Belzner, L., Gabor, T., and Phan, T. (2018). 深度多智能体强化学习中的行动市场。In K˚urkov´a, V., Manolopoulos, Y., Ham- mer, B., Iliadis, L., and Maglogiannis, I., editors,人工神经网络与机器学习 – *ICANN 2018*, 《计算机科学讲义》, 第240–249页, Cham. Springer International Pub- lishing. Schmid, K., Belzner, L., and Linnhoff-Popien, C. (2021a). 学会惩罚其他学习智能体。MIT出版社。Schmid, K., Belzner, L., M¨uller, R., Tochtermann, J., and Linnhoff-Popien, C. (2021b). 随机市场博弈。In Zhou, Z.-H., editor, 第三十届国际人工智能联合会议论文集, *IJCAI-21*, 第384–390页。International Joint Conferences on Artificial Intelligence Organiza- tion. Schmid, K., Belzner, L., Phan, T., Gabor, T., and Linnhoff- Popien, C. (2020). 风险与信息不对称条件下议价的多智能体强化学习。页数：151。

Silver, D.、Schrittwieser, J.、Simonyan, K.、Antonoglou, I.、Huang, A.、Guez, A.、Hubert, T.、Baker, L.、Lai, M.、Bolton, A.、Chen, Y.、Lillicrap, T.、Hui, F.、Sifre, L.、van den Driessche, G.、Graepel, T. 和 Hassabis, D.（2017年）。无需人类知识掌握围棋。《*Nature*》, 550(7676):354–359。Bandiera - abtest：a Cg 类型：Nature Research Journals 编号：7676 主要类型：研究 出版商：Nature Publishing Group 主题词：计算科学；计算机科学；奖励 主题词 ID：computational-science；computer-science；reward。 Statman, M.（2004年）。《多元化之谜》。《金融分析师杂志》, 60(4):44–53。出版商：Routledge 电子版：https://doi.org/10.2469/faj.v60.n4.2636。Vinyals, O.、Babuschkin, I.、Czarnecki, W. M.、Mathieu, M.、Dudzik, A.、Chung, J.、Choi, D. H.、Powell, R.、Ewalds, T.、Georgiev, P.、Oh, J.、Horgan, D.、Kroiss, M.、Danihelka, I.、Huang, A.、Sifre, L.、Cai, T.、Agapiou, J. P.、Jaderberg, M.、Vezhnevets, A. S.、Leblond, R.、Pohlen, T.、Dalibard, V.、Budden, D.、Sulsky, Y.、Molloy, J.、Paine, T. L.、Gulcehre, C.、Wang, Z.、Pfaff, T.、Wu, Y.、Ring, R.、Yogatama, D.、W¨unsch, D.、McKinney, K.、Smith, O.、Schaul, T.、Lillicrap, T.、Kavukcuoglu, K.、Hassabis, D.、Apps, C. 和 Silver, D.（2019年）。运用多智能体强化学习达到《星际争霸II》宗师级水平。《*Nature*》, 575(7782):350–354。Bandiera abtest：a Cg 类型：Nature Research Journals 编号：7782 主要类型：研究 出版商：Nature Publishing Group 主题词：计算机科学；统计学 主题词 ID：computer-science；statistics。 Wang, J. X.、Hughes, E.、Fernando, C.、 Czarnecki, W. M.、Duenez-Guzman, E. A. 和 Leibo, J. Z.（2019年）。为利他行为演化内在动机。*arXiv:1811.05931 [cs]*。arXiv：1811.05931。Yang, J.、Li, A.、Farajtabar, M.、Sunehag, P.、Hughes, E. 和 Zha, H.（2020年）。学习激励其他学习智能体。*arXiv:2006.06051 [cs, stat]*。arXiv：2006.06051。