Improved Cooperation by Balance Exploration and Exploitation in Intertemporal Social Dilemma Tasks

摘要：当个体的行为具有理性特征时，对群体而言可能导致非理性的总收益。人类与许多群居特性的动物往往会进化出合作这一社会属性来应对这一挑战。因此，个体间相互合作对于群居生物适应自然环境的变化具有重要的意义。基于多智能体强化学习，我们提出了一个由个体阶段累积收益和目标收益之间的差值来定义的学习率，该学习率能根据环境收益通过在 exploration 和 exploitation 之间的切换调整智能体的策略，从而形成对群体总收益相对较优的个体策略。结果表明，该策略在跨期社会困境问题中获得相对较优的总体策略，其策略不需要直接获取群体内其它个体的信息。尤其是，个体内在需求的异质性能够帮助个体更好地平衡 exploration 和 exploitation，从而促进群体的合作行为。

1 介绍

动物或者人类的群体行为结果不仅受到环境的影响，同时也会受到群体内个体行为策略的影响。比如，动物或者人类的迁徙会受到所处环境资源的影响，动物或者人类会倾向于向环境资源丰富的地方迁徙。同时，群体内个体从环境处获得的收益也受到群体内其他个体策略的影响，当群体内所有个体的策略都是往资源丰富的地方迁徙，那么处于该地的个体获得的收益会逐渐减少。这类个体理性导致群体非理性的问题，就是所谓的跨期社会困境问题 [1]。为了在跨期社会困境情形下获得最优总收益，群体内的个体需要能在个体短期收益和群体长期收益之间进行权衡。然而，对于个体如何在其短期收益与群体长期收益之间权衡的策略还不清楚。

多智能体强化学习能够模拟多个智能体在动态可变环境下的行为策略，目前已有许多研究发现多智能体强化学习可以模拟群体如何形成合作，从而在各类社会困境问题中获得最优解。这些模型往往通过内部奖励，来让智能体形成对群体理性的策略。内部奖励包括对不平等的厌恶 [2]、亲社会性 [3]和名声 [4]等，这些内部奖励将抑制智能体采取对群体总体收益不利的策略。比如，智能体厌恶不平等，当它发现自己的收益远远大于其它智能体的收益时，就会抑制其继续采用最大化其个体收益的策略。为了形成内部奖励，这些模型都假定群体内个体能直接获取其它同伴的信息，从而形成个体的亲社会属性。然而，对于那些只能间接获取群体内极少同伴信息的群体，如鱼群和蚁群等如何形成合作的机制还难以给出合理的解释。

本研究基于 Eric Charnov 的边际价值理论，提出智能体可以通过简单地调整学习率来权衡 exploration 和 exploitation，就可以在跨期社会困境任务中形成合作，从而获得较高的群体总收益。强化学习模型中，Exploration 表现为智能体为了避免局部最优解选择当前不能获得最优奖励的动作，而 Exploitation 则是智能体选择当前能获得最优奖励的动作。为了在 exploration 和 exploitation 之间权衡，我们基于深度 Q 学习，提出了一个由个体阶段累积收益和目标收益之间的差值来定义的学习率，该学习率能根据环境收益通过在 exploration 和

exploitation 之间的切换调整智能体的策略，从而形成对群体总收益相对较优的个体策略。

2 相关研究

面对社会困境问题，智能体如何逐渐形成合作行为一直是社会科学、经济学和心理学等学科的重要研究问题。通过构建两个参与者交互的策略游戏，Komorita 和 Parks [5]等（1995）发现，通过设定“担心”和“贪婪”两种策略的收益，可以形成促使两个参与者形成合作行为。而进化动力学模型发现，针锋相对的策略（Axelrod 1984 [6]）、与直接给予自己帮助的人合作（Nowak 2006 [7]）、或者惩罚他人从而获得足够的回报（Fehr 和 Gachter 2002 [8]）都能促进合作的形成。以上研究尽管给出了形成群体合作的可能存在的因素，但没有给出个体具体的策略。

随着强化学习在解决诸如围棋 [9]和多人合作游戏 [10, 11]取得的显著成果，已有许多研究者开始利用多智能体强化学习模型来研究群体如何形成合作的机制 [12, 13]。通过设定智能体的决策任务和智能体形成合作所需要的策略参数，利用模型解释动物或者人类在形成合作行为时可能的策略参数。Sequeira [14]等提出智能体通过探索内在动机来形成社会属性。Foerster[15]等（2017）通过智能体对其它个体学习结果的建模，来让智能体在多轮囚徒困境游戏问题中形成合作。Peysakhovich[16]等（2018）发现当智能体更多关注其它个体收益时，能让智能体在 Stag Hunt games 中形成亲社会的策略。Hughes [12]等（2018）将对不平等的厌恶融合进智能体内部奖励，从而在自己收益远大于群体内其它个体时或自己收益远小于群体内其它个体收益时调整策略，从而形成合作。Jaques [17]等（2018）通过将个体动作对群体的影响转换为内部奖励，来让处于社会困境的智能体形成合作。Wang [18]等（2019）提出进化深度强化学习，将其它个体过去和将来的奖励定义为智能体的内部奖励来进化出合作的策略。Khadka [19]等人（2019）设计了一种方法来学习具有共享重放缓冲区的多种策略，并动态地选择最佳学习者从而逐渐进化出多智能体间的合作。Badjatiya [13]等（2020）提出设计一个现状（Status-Quo）损失函数来让智能体尽量跟随现状，从而在社会困境环境中进化出合作行为。McKee [20]等（2020）智能体在从具有异质性特征的群体内采样它们的奖励，可以让智能体获得亲社会的属性。Danassis [21]等（2021）发现智能体通过在学习过程融合公共信号（如时间、日期等周期性数字）可以提升智能体的合作行为。以上模型一个共同的特征是，智能体为了形成合作行为，都需要直接获取所有其他智能体的相关信息，这些模型对于那些只能间接获取群体内少部分同伴信息的群体，如鱼群和蚁群等如何形成合作的机制还难以给出合理的解释。

由于环境状态的动态变化和不确定性，智能体要么只利用现有经验进行 Exploitation，要么冒着当前不能获得较好收益的风险去 Exploration，以期望得到可能更好的策略。因此，Exploration 和 Exploitation 一直是强化学习中重要的研究主题。早期在求解多臂赌博机问题时，Epsilon-greedy [22]、Upper confidence bounds [23]和 Boltzmann exploration [24]等能在 Exploration 和 Exploitation 之间进行权衡从而获得最佳的总体收益。然而，现实的环境中由于奖励信号的稀疏特性以及环境状态存在的异常噪声，以上简单的探索策略并不能获得较好的总体收益。

一个较为通用的方法是设计一个内部奖励函数以便形成智能体的内部动机 [25]，从而通过诸如好奇心来引导智能体进行探索。好奇心包括发现了新的状态，或者提升智能体对环境变化估计的准确性等[22]。这类基于内部奖励的探索策略有可能存在收敛速度慢，以及探索回报非平稳导致难以形成固定探索策略的问题。由此，发展出基于记忆的探索策略 [26]，以及重采样 Q 值探索策略 [26]等从而避免基于内部奖励探索策略存在的不足。然而，以上单智能体的探索策略并不一定适合多智能体的协同探索。

对于多智能体探索的情形，不仅需要鼓励智能体探索新的状态和应对奖励信号稀疏的问题，也需要协同智能体之间的行动从而形成合作来对环境进行探索。Agogino 和 Tumer [27]定义了在较小规模的状态空间，用于评估多智能体所获奖励功能有效性的方法。Jaques [17]等人为多智能体强化学习定义了一个内在奖励函数，鼓励智能体采取对其他智能体行为影响最大的行动，从而获得协同的探索策略。Mahajan [28]等人引入了一种实现“承诺”探索的机制，允许智能体探索临时扩展的共同策略。Wang [18]等人定义了基于影响力的奖励，鼓励智能体访问其行为影响其他智能体转变和奖励的区域。最近，Iqbal 和 Sha（2021）[29]提出了一类基于内部奖励的探索方法，该方法主要的特点是可以协同智能体之间的探索策略，并且能让智能体更好的获得总体收益。以上多智能体下的探索策略仍然需要智能体脂解获取其他智能体的信息，对于只能获取少部分其他智能体，甚至不需要其他智能体信息下的探索策略还需要做进一步研究。

3 多智能体强化学习与决策任务

3.1 多智能体强化学习

我们将多智能体强化学习模型定义为一个四元组，它包括状态集合 S、状态转移函数 T、动作集合 A 和奖励 r，即$< S, T, A, r >$。在环境中有 N 个智能体，每个智能体能感知到的状态为$O_n: S \to \mathbb{R}^d$，表示智能体 n 能观察到状态的 d 个维度，也就是智能体只能部分观察到其所处状态。环境中的每个智能体通过其动作 A_n 与环境交互，智能体的动作会引起环境状态的变化，变化由状态转移函数来刻画：$T: s \times A_1 \times A_2 \times \ldots \times A_n \longrightarrow s'$，也就是环境中所有智能体的动作共同作用将环境状态从s改变为另一状态s'。

每个智能体 i 根据其观察$o_i = O(s, i)$学习到策略$\pi(a_i|o_i)$，智能体执行动作a_i后将获得奖励r_i，通过奖励对动作结果进行评估。智能体的目标是习得一个优化策略，以便获得最大的长期收益。智能体的长期收益定义为：

$$Q_{\vec{\pi}}(s_0, \overrightarrow{a_0}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \, \overrightarrow{r_t} | \overrightarrow{a_t} \sim \overrightarrow{\pi_t}, s_{t+1} \sim T(s_t, \overrightarrow{a_t})\right]$$

公式 1

其中，γ是取值 0 到 1 之间的折扣因子。简化起见，记$\vec{a} = (a_1, \ldots, a_n)$。对于智能体 i，为了获得最大的期望收益，可以根据以下函数更新 Q 函数 [30]，

$$Q_{t+1}^{i}(s_t, a_t) = Q_t^i(s_t, a_t) + \eta^i[r_{t+1}^i + \gamma^i \max_{a'} Q_t^i\left(s_{t+1}^i, a'\right) - Q_t^i(s_t^i, a_t^i)]$$

公式 2

3.2 基于目标收益的学习率

我们考虑通过智能体的阶段累积收益和目标收益来定义学习率，学习率反应的是环境的变化对智能体策略的影响。为了达到这个目的，我们将阶段累积收益$\tilde{R}$与目标收益$\hat{R}$的差值定义为学习率，即

$$\eta^i = \frac{max\{\hat{R}^i - \tilde{R}^i, 0\}}{\hat{R}^i} * \beta$$

公式 3

β是常数，大小设置为 0.001。阶段累积收益$\tilde{R}$是智能体在时间τ内的累积奖励值，它反映的是其他智能体间接地对个体在某一时间段内收益的影响。目标收益$\hat{R}$是一个固定值，每个智能体都有一个目标收益，它反映的是智能体的满足度。当目标收益较大时，表示智能体需要获得较多的累加收益才能满足。如果智能体的阶段累积收益小于目标收益，表明这个智能体的目标没有达到，其表现出更多的 Exploration。当智能体的阶段累积收益接近目标收益，表明智能体的策略达到了它的预期，其表现出更多的 Exploitation。

根据以上学习率的定义，当环境处于稳定状态时，智能体的策略逐渐收敛，其学习率处于较低水平。当环境状态存在突变，智能体的策略要能很快适应这个变化，此时期学习率处于较高水平。

3.3 决策任务

根据 Hughes [12]等（2018）的资源采集任务，我们设计了一个类似的跨期社会困境任务。任务环境中包含两个价值不同的资源区域，即苹果区和垃圾区。环境地图大小 S 为 12*20 个单位，垃圾分布在环境的上半区域，苹果分布在环境的下半区域。垃圾在其所在区域内以概率δ_g出现，环境中垃圾的数量记为N_g。苹果在其所在区域内以概率δ_a出现，环境中苹果的数量记为N_a。苹果的增长率与垃圾的数量呈负相关，其关系为:

$$\delta_a = -\frac{\sigma}{\Delta S_g} * N_g + \sigma$$

公式 4

其中σ是苹果的最大增长率，ΔS_g取值为地图中垃圾区域的一半。

环境中分布若干数量的智能体，智能体通过在环境中的移动，要么收获所在位置的奖励，要么清理所在位置的垃圾。智能体收获苹果的收益记为r_a，而清理垃圾的收益记为r_g。智能体的目的是获取最多的总收益。在该任务中，各个智能体仅能感知其周围有限视野内的信息，智能体的视野范围大小记为v。

该决策任务的困境如下，苹果和垃圾的增长相互影响。由于苹果的收益大于垃圾的收益，智

能体的个体策略会倾向采苹果而非清理垃圾。然而，苹果数量的减少会导致垃圾数量的增加，从而抑制苹果出现的概率。因此，对智能体群体而言，需要一部分智能体清理垃圾，一部分智能体采集苹果，才能获得对整体而言较多的总收益。

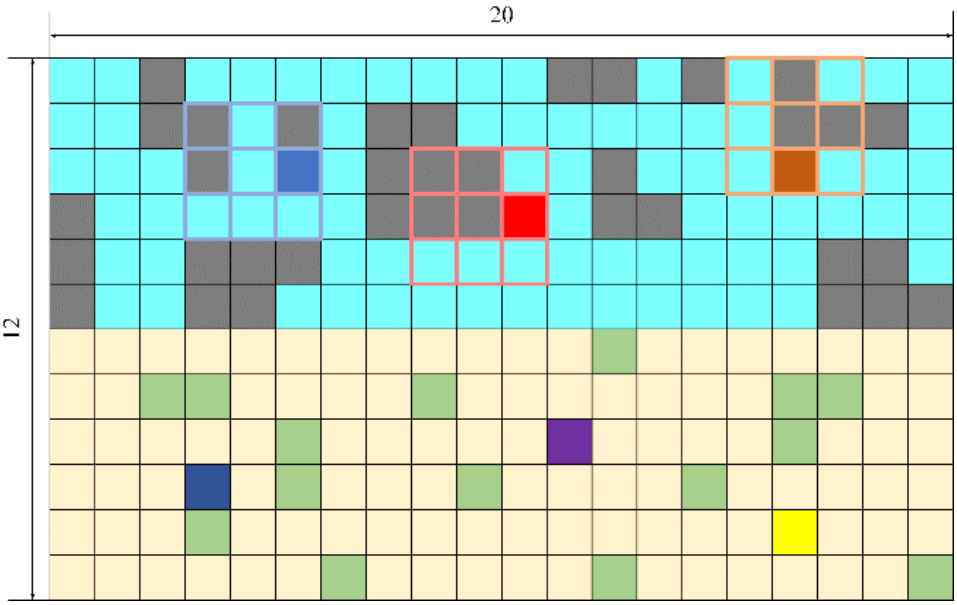


图 1. 游戏地图

由于垃圾与苹果的生长并不平衡，多智能体在清理垃圾时清理视野范围内所有的垃圾v_g，而采集苹果时只采集当前位置的苹果。因此，智能体清理垃圾获得的实际奖励为$reward_g = v_g * r_g$，采集苹果获得的实际奖励为$reward_a = r_a$。决策任务中设定$\tilde{R}^i = \sum_{t=1}^{\tau} reward_t$，表示智能体在时间跨度$\tau$内的总收益。

仿真中，环境中随机放置了 6 个智能体$Agent_{1\sim6}$，它们各自的学习函数见公式 1。我们将地图中每个单元格的位置映射到$[-120, 120]$区间内，其中$[0, 120]$表示垃圾区域内的单元格，$[-1, -120]$表示苹果区域内的单元格。M_i表示每个 Agent 在环境中的初始位置，且$M_i \in [-120, 120]$。对每个智能体而言，每个 episode 包括 100 trials，每组实验包括 300 轮 episode。

3.4 同质与异质性群体属性

根据$\hat{R}^i$的取值方法，将智能体群体分为异质性和同质性。异质性表示在给定范围内$\hat{R}^i$随机取值，这种异质性反映了智能体个体的多样性。当智能体目标收益满足$\hat{R}^i \geq reward_a * \tau$时，称其为高目标收益者，而当智能体目标收益满足$\hat{R}^i \leq reward_g * \tau$时，称其为低目标收益者，这两种情况下的智能体具有同质性群体属性。仿真中的群体属性相关的参数设置见表 1。

表 1 群体属性参数设置

$group$	$\hat{R}^{i=1,2,3}$	$\hat{R}^{i=4,5,6}$	M^i	τ	r_a	r_g	η^i
Heterogeneous	$10 < \hat{R}^i$	$\hat{R}^i > 50$	$M^{i=1,2,3} \in [\mathbf{0}, \mathbf{120}]$ $M^{i=4,5,6} \in [-\mathbf{1}, -\mathbf{120}]$	5	10	5	$\frac{max\{\tilde{R}^i - \hat{R}^i, 0\}}{\hat{R}^i} * r$
Homogeneous	$10 < \hat{R}^i$	$10 < \hat{R}^i$					
	$\hat{R}^i > 50$	$\hat{R}^i > 50$					

4 结果

我们使用阶段累积收益$\tilde{R}^i$和目标收益$\hat{R}^i$定义动态学习率η^i，验证该学习率在跨期社会困境任务中通过平衡 exploration 和 exploitation 以形成群体合作，从而获得相对较优的总体收益。如图 2 所示，我们比较了群体在固定学习率和动态学习率下执行跨期社会困境任务的收益。动态学习率下的群体总收益可以收敛达到 2200~2500 之间，而固定学习率(=0.001)的群体的总收益仅可以收敛到 1300~1600 之间。智能体仅采用随机策略，总收益收敛在 700~1000 之间。

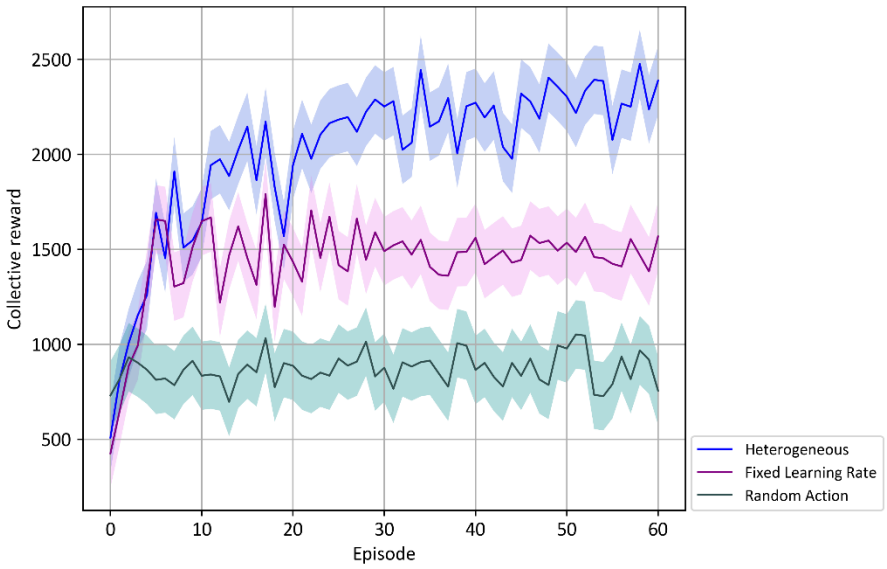


图 2. 动态学习率和固定学习率的收益比较

我们通过设定目标收益不同的分布，来验证目标收益在权衡 Exploration 和 Exploitation 时的作用。我们根据目标收益的分布，定义了 Heterogeneous、Homogeneous High 和 Homogeneous Low 三种群体属性。如图 3 所示，这三类群体中 Heterogeneous 群体的总收益比两种 Homogeneous 群体的总收益都高。Homogeneous Low 群体的总收益最低，甚至低于随机策略的总收益。

为了更清楚的表示每一个智能体在环境中 Exploration 和 Exploitation 的变化，我们在图 4 中画出每一个智能体在环境中的活动位置。Heterogeneous 群体根据自身的目标收益进行 Exploitation 形成分工（图 4 A）。这种分工会让部分智能体（低目标收益的智能体）在垃圾区采集垃圾，而部分智能体（高目标收益的智能体）在苹果生长区采集苹果，正是这种分布导致 Heterogeneous 群体的总体收益最高。当群体内每个智能体的目标收益都高，导致它们一直在环境中采用 Exploration 去获取高收益的苹果（图 4 B）。而当群体内每个智能体的目标收益都低时，他们在苹果区域获得奖励大于个体本身的期望，模型会抑制低目标收益者贪婪地采集苹果，当智能体 Exploration 到匹配自身目标收益的区域时(垃圾区域)才会更新策略（图 4 C）。完全随机选择动作的群体一直在进行没有目标的 Exploration（图 4 D）。

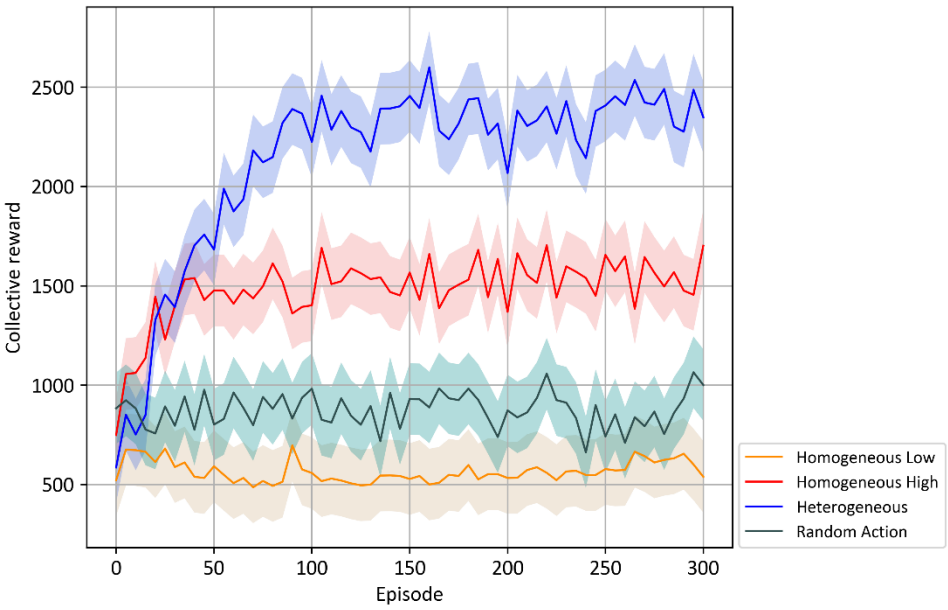
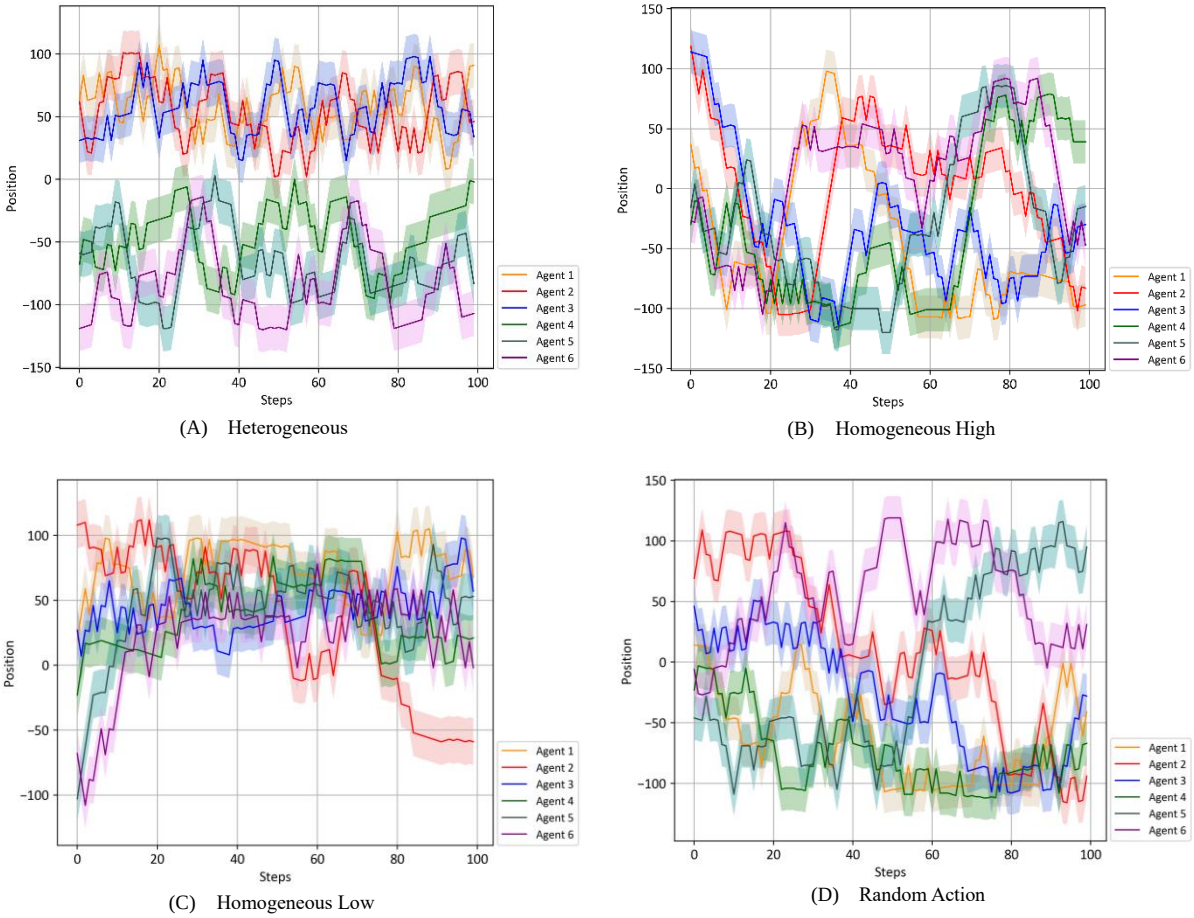


图 3. 个体异质性与个体同质性收益比较



(A) Heterogeneous (B) Homogeneous High

(C) Homogeneous Low (D) Random Action

图 4. 异质性群体与同质性群体内智能体的活动位置对比

图 5. 不同群体内智能体的收益对比

如图 5 所示，对比了不同群体中每个智能体之间的收益，以便比较合作行为是否造成智能体间贫富差距。结果表明，相比较于同质性群体内收益差值，异质性群体间收益差值最大。这表明尽管群体的异质性特征促进了智能体间的合作行为，但也会导致群体内个体间贫富差距变大。

我们单独使用 Heterogeneous 群体作为实验对象，控制每个智能体的目标收益不变，来探究不同的阶段累计长度τ对群体合作行为的影响。如图 6 所示，随着τ的不断增大，群体的总收益不断下降；当τ的取值过小，智能体总体收益同样较小。当τ的取值较大时，智能体难以感知环境的变化，从而降低其采用 Exploration 的可能性。当τ的取值较小时，智能体仅关注当前收益，同样难以感知环境变化，从而使得智能体频繁采用 Exploration。

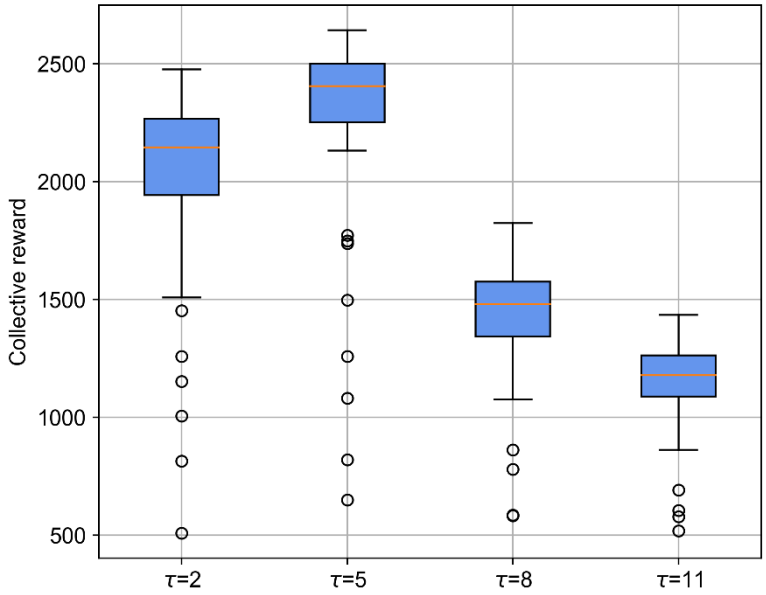


图 6. 不同阶段累计长度τ的总收益

5 结论

边际价值理论提出人对物品的欲望会随其不断被满足而递减。基于边际价值理论我们提出了一个通过平衡利用-探索，从而在跨期社会困境中多智能间形成合作的方法。面对跨期社会困境问题，智能体只需要计算在一段时期内的收益，而不需要从其他智能体获取额外的信息，就可以形成类似合作的行为。我们的结果表明，智能体间的异质性，如各个智能体的目标收益的异质性，是形成合作行为的关键。这个结论与 McAvoy 等（2020）和 McKee 等（2020）[31]最近的研究结果类似，他们发现智能体连接数量的异质性和社会偏好的异质性，能促进智能体社会行为的形成。我们的研究结果进一步揭示，其它类型的异质性，如本文设定的各个智能体目标需求的异质性，也能促进智能体间合作行为的形成。在此基础上，我们推断也许还存在其它可以促进智能体形成合作行为的异质性参数，这一点值得后续更为深入的理论与实验研究。

我们研究的一个主要不足之处在于，设计的智能体学习算法并没有考虑智能体间直接的互动和交互。智能体间通过一个时间段内总收益来产生相互的影响。对某个智能体而言，如果群体内的其它智能体行为都是非社会性，那么这个智能体一个时间段内总收益就会变少，这会引起该智能体根据其自身的目标收益来调整利用和探索策略。也就是说，其它智能体的非社会性行为间接引起了该智能体策略的变化。相似地，如果群体内的其它智能体行为都是亲社会性，他们的行为也会间接引起了该智能体利用和探索策略的变化。我们将在后续研究考虑，在算法中增加智能体间交互的参数。

需要特别指出，我们只是在跨期社会困境中验证了智能体通过调整利用与探索策略来形成合作，对于其它类型的社会困境问题能否得到同样的结论还需要进一步的验证。

[1] E. Hughes *et al.*, "Inequity aversion improves cooperation in intertemporal social dilemmas," *arXiv preprint arXiv:1803.08884,* 2018.

[2] D. Engelmann and M. Strobel, "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments," *American economic review,* vol. 94, no. 4, pp. 857-869, 2004.

[3] A. Peysakhovich and A. Lerer, "Prosocial learning agents solve generalized stag hunts better than selfish ones," *arXiv preprint arXiv:1709.02865,* 2017.

[4] J. A. Cuesta, C. Gracia-Lázaro, A. Ferrer, Y. Moreno, and A. Sánchez, "Reputation drives cooperative behaviour and network formation in human groups," *Scientific reports,* vol. 5, no. 1, pp. 1-6, 2015.

[5] Komorita, Samuel, S., Parks, Craig, and D., "Interpersonal relations: Mixed-motive interaction," *Annual Review of Psychology,* 1995.

[6] R. Axelrod and W. D. Hamilton, "The evolution of cooperation," *Quarterly Review of Biology,* vol. 79, no. 2, pp. 135-160, 1981.

[7] M. Nowak, "Five Rules for the Evolution of Cooperation," *Science,* vol. 314, no. 5805, pp. 1560-1563, 2006.

[8] S. Gaechter and E. Fehr, "Altruistic Punishment in Humans," *Nature,* 1997.

[9] D. Silver *et al.*, "Mastering the game of go without human knowledge," *nature,* vol. 550, no. 7676, pp. 354-359, 2017.

[10] N. Jaques *et al.*, "Social influence as intrinsic motivation for multi-agent deep reinforcement learning," in *International Conference on Machine Learning*, 2019: PMLR, pp. 3040-3049.

[11] Y. Du, L. Han, M. Fang, J. Liu, T. Dai, and D. Tao, "Liir: Learning individual intrinsic reward in multi-agent reinforcement learning," *Advances in Neural Information Processing Systems,* vol. 32, pp. 4403-4414, 2019.

[12] E. Hughes, J. Z. Leibo, M. Phillips, K. Tuyls, and T. G Ra Epel, "Inequity aversion improves cooperation in intertemporal social dilemmas," in *The 32nd Conference on Neural Information Processing Systems (NIPS)*, 2018.

[13] P. Badjatiya, M. Sarkar, A. Sinha, S. Singh, and B. Krishnamurthy, "Inducing Cooperation in Multi-Agent Games Through Status-Quo Loss," 2020.

[14] P. Sequeira, F. S. Melo, P. Rui, and A. Paiva, "Emerging social awareness: Exploring intrinsic motivation in multiagent learning," in *Development and Learning (ICDL), 2011 IEEE International Conference on*, 2011.

[15] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual Multi-Agent Policy Gradients," 2017.

[16] A. Peysakhovich and A. Lerer, "Prosocial learning agents solve generalized Stag Hunts better than selfish ones," 2017.

[17] N. Jaques *et al.*, "Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning," 2018.

[18] J. X. Wang, E. Hughes, C. Fernando, W. M. Czarnecki, and J. Z. Leibo, "Evolving intrinsic motivations for altruistic behavior," 2018.

[19] S. Khadka, S. Majumdar, T. Nassar, Z. Dwiel, and K. Tumer, "Collaborative Evolutionary Reinforcement Learning," 2019.

[20] K. R. McKee, I. Gemp, B. McWilliams, E. A. Duéñez-Guzmán, E. Hughes, and J. Z. Leibo, "Social diversity and social preferences in mixed-motive reinforcement learning," *arXiv preprint arXiv:2002.02325,* 2020.

[21] P. Danassis, Z. D. Erden, and B. Faltings, "Improved Cooperation by Exploiting a Common Signal," *arXiv preprint arXiv:2102.02304,* 2021.

[22] J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel, "Multi-agent reinforcement learning in sequential social dilemmas," *arXiv preprint arXiv:1702.03037,* 2017.

[23] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *Journal of Machine Learning Research,* vol. 3, no. Nov, pp. 397-422, 2002.

[24] M. Edman and N. Dhir, "Boltzmann Exploration Expectation-Maximisation," *arXiv preprint arXiv:1912.08869,* 2019.

[25] A. Zgonnikov and I. Lubashevsky, "Intrinsic motivation and learning dynamics," Citeseer, 2013.

[26] L. Weng, "Exploration strategies in deep reinforcement learning," *Online: https://lilianweng. github. io/lil-log/2020/06/07/exploration-strategies-in-deep-reinforcement-learning. html,* 2020.

[27] A. K. Agogino and K. Tumer, "Analyzing and visualizing multiagent rewards in dynamic and stochastic domains," *Autonomous Agents and Multi-Agent Systems,* vol. 17, no. 2, pp. 320-338, 2008.

[28] A. Mahajan, T. Rashid, M. Samvelyan, and S. Whiteson, "MAVEN: Multi-Agent Variational Exploration," 2019.

[29] S. Iqbal and F. Sha, "Actor-Attention-Critic for Multi-Agent Reinforcement Learning," 2018.

[30] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning,* vol. 8, no. 3-4, pp. 279-292, 1992.

[31] A. Mcavoy, B. Allen, and M. A. Nowak, "Social goods dilemmas in heterogeneous societies," *Nature Human Behaviour,* vol. 4, no. 8, pp. 1-13, 2020.