# Heterogeneous of Intrinsic Requirements Improves Cooperation in Sequential Social Dilemmas

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。个体的内在需求往往会影响其决策的动机。基于多智能体强化学习，我们研究了具有内在需求异质性的群体，在完成跨期社会困境决策任务时的合作趋势。结果表明，个体内在需求的异质性促进了合作行为。此外，具有异质性特征的个体在环境中的分布同样对合作趋势的产生起到调制作用。

## 1 介绍

社会困境是指个人利益和群体利益发生了抉择上的冲突, 其主要表现是人们倾向于从自己的短期利益出发来决定自身的行动, 并未充分考虑集体的长期利益。

什么是社会困境。

当前社会困境的模型。

内在需求的异质性包括两个方面，一个是来源的异质，另外就是水平的异质。

## 2 多智能体强化学习与决策任务

### 2.1 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。在环境中有*N*个智能体，每个智能体能感知到的状态为，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*An*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：，也就是环境中所有智能体的动作共同作用将环境状态从改变为另一状态。

每个智能体*i*根据其观察学习到策略，智能体执行动作后将获得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：

其中，是取值0到1之间的折扣因子，表示收益函数，即智能体采取动作后的奖励。简化起见，记。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，

### 2.2 基于目标收益的学习率

我们考虑通过智能体的平均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，我们将平均收益与目标收益的差值定义为学习率，即

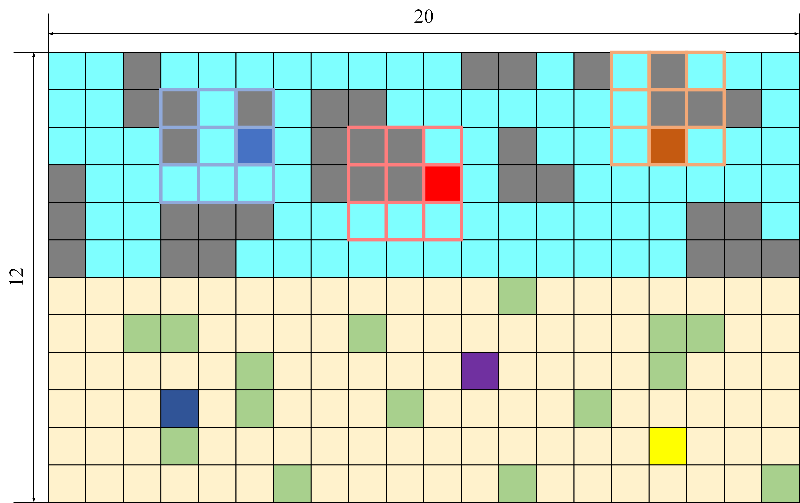
目标收益是一个固定值，每个智能体都有一个目标收益，它反映的是智能体的满足度。也就是说，当目标收益较大时，智能体需要获得较多的累加收益才能满足。平均收益是智能体在一段时间内的平均奖励值。如果智能体的平均收益小于目标收益，表明这个智能体的目标没有达到，其表现出更多的探索性。当智能体的平均收益接近目标收益，表明智能体的策略达到了它的预期，其表现出更多的Exploitation。

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。

### 2.3 决策任务

本文模拟了多智能体资源采集的社会困境，困境中包含了有利资源(苹果)、不利资源(垃圾)，且两种资源的生长区域均分模拟的任务环境。模拟的任务环境如图. 所示，环境地图大小为12\*20个单位，其中垃圾增长在环境的上半部区域，苹果增长在环境的下半部区域。垃圾增长区域的每个单位以概率生长垃圾，记环境中垃圾的数量为；苹果区域以概率生长苹果，记环境中苹果的数量为。其中苹果的增长率与垃圾的数量呈负相关，其关系为:

在该困境中多智能体只能观察到有限视野范围内的情况，他们可以根据自身的需求通过采集苹果获得较高的收益，或清理垃圾获得较低的收益。Agents将被随机或固定安置在环境中的垃圾区域和苹果区域，且每个episode任务中一个agent共执行100步动作，然后重新安置Agents的位置，本文每组实验进行了300轮episode任务。



由于垃圾与苹果的生长并不平衡，因此多智能体在清理垃圾时会清理视野范围内所有的垃圾，采集苹果时只采集当前位置的苹果。集体的行为会使环境中苹果和垃圾的数量发生变化，只有保证环境中资源的平衡增长才能使集体收益最大化，在该困境中多智能体需要形成既有采集苹果又有清理垃圾的合作行为。

### 2.4 参数设置

本节介绍了环境与多智能体的参数属性，在实验部分通过调整多智能体的参数探究了使多智能体形成合作并达到环境平衡的不同因素。

表1 参数

|  |  |
| --- | --- |
| **参数名称** | **描述** |
|  | 垃圾增长率 |
|  | 环境中垃圾的数量 |
|  | 采集垃圾的奖励 |
|  | 苹果增长率 |
|  | 环境中苹果的数量 |
|  | 采集苹果的奖励 |
|  | 智能体的目标收益 |
|  | 平均收益时间间隔 |
|  | 智能体的平均收益 |

## 3 结果

为研究多智能体之间的差异对合作的影响，在决策任务中调整多智能体参数分别从个体的目标收益、初始位置上的不同证明，具有异质性特征的个体在环境中的分布对合作趋势的产生起到调制作用。

### 3.1 目标收益的分布

### 3.2 智能体的初始位置

## 4 讨论

## 5 结论