arXiv:2601.06382v1 [cs.MA] 10 Jan 2026

变化奖励下的动态激励合作

Philipp Altmann 慕尼黑路
德维希-马克西米利安大学，德国
philipp.altmann@ifi.lmu.de

Thomy Phan 拜罗伊特大学，德国

Maximilian Zorn 慕尼黑路
德维希-马克西米利安大学，德国

Claudia Linnhoff-Popien 慕尼黑路
德维希-马克西米利安大学，德国

Sven Koenig
加利福尼亚大学欧文分校，美国
厄勒布鲁大学，瑞典

摘要

同伴激励 (PI) 是一种常见的多智能体强化学习方法，其中所有智能体都可以相互奖励或惩罚，以在社会困境中实现合作。尽管当前的 PI 方法具有实现可扩展合作的潜力，但它们严重依赖固定的激励值，而这些值需要根据环境奖励进行恰当选择，因此对环境奖励的变化极为敏感。因此，即使相互合作的条件保持不变，它们也无法在环境的变化中的奖励 下维持合作，例如因规格修改、供需变化或传感缺陷导致的奖励变化。本文提出了可变交换的动态奖励激励 (DRIVE)，这是一种针对奖励变化中的社会困境实现合作的自适应PI方法。DRIVE 智能体以完全去中心化的方式相互交换奖励差异，以激励相互合作。我们展示了 DRIVE 如何在一般的囚徒困境中实现相互合作，并通过实证评估 DRIVE 在奖励变化的复杂序贯社会困境中的表现，证明与当前最先进的PI方法相比，它能够实现并维持合作。

代码： https://github.com/philippaltmann/DRIVE

关键词

多智能体强化学习、涌现合作、同伴激励、动态奖励、社会困境

1 引言

许多人工智能应用场景，例如自动驾驶 [40], 智能电网[10], 和物联网应用 [8],，都可以建模为具有自利性并进行在线学习的多智能体系统（*MAS*），其中，对立的目标或共享资源 [5]会引发冲突。为了最大化社会福利，自利型多智能体系统中的合作至关重要，这要求智能体避免自私行为，以实现集体利益。这种个体理性与集体理性之间的张力通常被建模为社会困境（*SD*） [3, 38]。设计分布式机制以激励社会困境中的合作仍是一项关键挑战，因为它会直接影响全局效率（例如，汇总的社会福利

如第 2 节所定义）以及可持续性（例如，长期资源可用性和智能体之间的公平性） [3, 6, 33, 44]。多智能体强化学习（多智能体强化学习）是一种广泛使用的框架，用于在社会困境和时间跨度更长的序贯社会困境（*SSD*）中训练理性智能体，其中每个智能体都最大化自身的奖励 [13, 23, 33, 52]。

这是该作品的作者版本，仅供您个人使用。该作品的扩展摘要版本发表于第*25*届自主智能体与多智能体系统国际会议论文集（*AAMAS 2026*年） [1]。

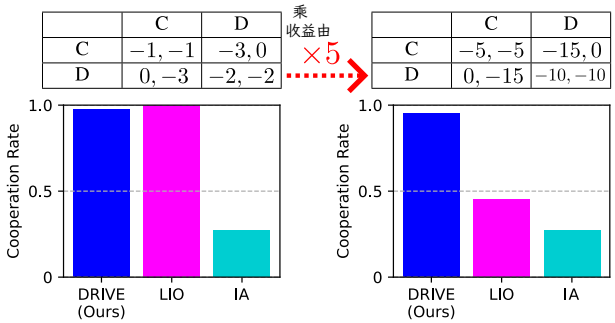


图1：用于说明两人囚徒困境中同伴激励方法奖励敏感性的示例：即使在原始收益设置下，IA 仍需仔细调参，且在两种情形下都无法实现合作。LIO 在原始尺度下（左）实现了较高水平的合作，但收益发生变化时便会退化，尽管贪婪和恐惧的不等式条件（公式1）仍然成立。相比之下，DRIVE 无需重新调参，便能在两种设置下保持稳健合作。

然而，非合作博弈论和实证研究表明，朴素的多智能体强化学习方法往往无法维持合作，因为独立适应会导致相互背叛 [3, 46]。此外，策略优化会受到奖励函数不一致的影响，这种情况可概括为变化中的奖励。从模拟转向现实应用时，这类变化很常见：抽象的奖励代理指标会被物理信号取代。噪声传感器、硬件退化、规格变化或市场需求变化，同样可能在训练过程中改变奖励尺度 [11]。在有人参与的回路设置中，评估者也可能在线调整反馈，从而导致突发或不规则的变化 [20]。在这些情况下，底层的序贯社会困境结构（即贪婪与恐惧不等式）得以保留，但数值会发生波动。对这类变化缺乏适应能力的方法，难以在狭窄且经过人工调参的条件之外维持合作。

同伴激励 (PI)使智能体能够相互奖励或惩罚，从而促进社会困境中的合作 [28, 52]。它因在复杂的序贯社会困境中取得成功，并与生物学、经济学和社会科学相关联而日益受到关注 [21, 39, 47]。然而，现有的 PI 方法依赖固定或针对特定领域调整的激励值；即使合作条件保持不变，在奖励变化时也会失效 [12, 56]。如图 1 所示，先前的 PI 方法可能在囚徒困境的某个特定实例中奏效 [13, 36, 52],，但当收益发生偏移或缩放时，即使仍满足相同的贪婪 和恐惧 不等式，也可能失效 [3, 29]。

值得注意的是，激励其他学习智能体 (LIO) [52] 或不平等厌恶 (IA) [18] 等常用方法，表面上似乎能够适应此类变化，因为它们依赖学习得到的激励或相对激励。然而，IA 需要精确调整超参数才能发挥作用，而且在图 1 左侧的原始（未缩放）收益矩阵和右侧的统一缩放变体中，都无法促进合作。LIO 在原始收益下能够实现合作，但在简单缩放后性能下降，因为其学习到的激励幅度与绝对奖励尺度挂钩。因此，这些方法要么在奖励偏移时失去鲁棒性，要么抑制智能体性能中有价值的异质性。应对此类偏移通常需要重新调整超参数，这对于临时或在线学习场景并不实际。

为此，我们提出了可变交换的动态奖励激励（*DRIVE*），这是一种适用于奖励不断变化的社会困境的自适应 PI 框架。通过直接根据奖励差值调整激励，DRIVE 无需重新调参即可维持合作，也不会将合理的差异与剥削行为混为一谈。我们的贡献总结如下：

- 我们提出了一种互惠交换机制，通过奖励差异激励合作。
- 我们证明，在广义囚徒困境博弈中，DRIVE 能使激励与双方合作保持一致，且不受奖励偏移与缩放的影响。
- 我们通过实验证明，在奖励不断变化的序贯社会困境中，DRIVE 具有鲁棒性，并能带来更优的合作表现。

2 背景

问题表述。在本研究中，我们关注部分可观测马尔可夫博弈 $M = \mathbb{D}, \mathbb{S}, \mathbb{A}, \mathbb{P}, \mathbb{U}, \mathbb{Z}, \Omega$，其中 $\mathbb{D} = \{1, ..., n\}$ 是智能体集合 i，$\mathbb{S}$ 是状态集合 s_t 在时间步 t，$\mathbb{A} = \mathbb{A}_1, ..., \mathbb{A}_n = \mathbb{A}_{ii\in\mathbb{D}}$ 是联合动作集合 $a_t = a_{t,i_{i\in\mathbb{D}}}$，$\mathbb{P}(s_{t+1}|s_t, a_t)$ 是转移概率，$u_{t,i_{i\in\mathbb{D}}} = \mathbb{U}(s_t, a_t) \in \mathbb{R}$是联合奖励，$\mathbb{Z}$ 是局部观测集合$z_{t,i}$ 对于每个智能体i，且 $\Omega(s_{t+1}) = z_{t+1} = z_{t+1,i}$ $_{i\in\mathbb{D}} \in \mathbb{Z}^n$ 是后续联合观测。每个智能体 i 都会维护局部历史记录 $\tau_{t,i} \in (\mathbb{Z} \times \mathbb{A}_i)^t$。$\pi_i(a_{t,i}|\tau_{t,i})$ 是基于智能体 i的个体策略的动作选择概率。此外，我们假设每个智能体 i在每个时间步 t都有一个由其他智能体构成的邻域 $\mathrm{N}_{t,i} \subseteq \mathbb{D} - \{i\}$，该邻域取决于具体领域，正如文献 [28, 54]所述。 π_i通过价值函数 $V_i^{\pi}(s_t) = \mathbb{E}_{\pi}[G_{t,i}|s_t]$进行评估，适用于所有 $s_t \in \mathbb{S}$，其中 $G_{t,i} = \sum_{k=0}^{\infty} \gamma^k u_{t+k,i}$是智能体 $i \in \mathbb{D}$的个体折扣回报，折扣因子为 $\gamma \in [0, 1)$，而 $\pi = \pi_{j\,j\in\mathbb{D}}$是多智能体系统（MAS）的联合策略。智能体 i的目标是寻找最佳响应 π_i^*，使得 $V_i^* = max_{\pi_i} V_{\pi_i, \pi_{-i}}$适用于所有 $s_t \in \mathbb{S}$，其中 π_{-i}是不包含智能体 i的联合策略。实际上，任何智能体 i都无法直接观测全局状态 s_t，因此 V_i^{π}需通过局部信息进行近似，即 $\tau_{t,i}$而非 [19, 23, 32]。为衡量多智能体系统（ MAS ）中的合作，我们将社会福利或功利主义指标 $U = \sum_{i\in\mathbb{D}} \sum_{t=0}^{H-1} u_{t,i}$ 定义为未折扣的奖励总和。

	C	D
C	R, R	S, T
D	T, S	P, P

(a) 社会困境收益

	C	D
C	$-1, -1$	$-3, 0$
D	$0, -3$	$-2, -2$

(b) 囚徒困境实例

图 2：（a）包含 R、P、T 和 S 的社会困境收益矩阵。在囚徒困境**（*PD*）中，$T > R > P > S$ 满足 [3, 29]。（b）满足这些不等式的一个囚徒困境实例。**

社会困境。社会困境（*SD*）是指这样一类博弈：独立优化的策略 π_i 无法实现使集体福利最大化的全局最优结果。社会困境通常通过双人矩阵博弈进行研究，其中有两个动作： C（合作）和 D（背叛），由此产生四种可能的收益（图 2）： R表示相互合作，P表示相互背叛， T表示剥削对方，S表示被剥削。若收益满足以下条件，一个矩阵博弈就是囚徒困境（*PD*） [3, 29]:

$$T > R > P > S \tag{1}$$

此处， $T > R$表示贪婪，而 $P > S$表示恐惧。在迭代囚徒困境（*IPD*）中，还满足一个额外条件： $2R > T + S$ [3, 29]。图 2b 展示了一个实例，其中尽管 C是社会最优的， D仍是个体理性的。只要这些不等式成立，博弈的策略性质就不受具体收益 [3, 38]的影响。囚徒困境和迭代囚徒困境是尤为重要的社会困境，因为贪婪和恐惧常常使智能体偏离相互合作，尽管合作能带来集体利益；这一现象在自然界和人类社会中均有观察 [3, 7, 38]。

序贯社会困境（*SSD*）通过引入时间结构来扩展社会困境，并将其建模为随机博弈 [23, 33]。序贯社会困境允许构建更现实的情境，在这些情境中，行为由策略而非原子动作刻画。通过将策略归类为 C或 D，并评估经验收益 [23], , 仍可将其映射为矩阵博弈，从而使社会困境的核心概念适用于序贯社会困境。

多智能体强化学习。我们考虑去中心化（独立）学习，其中每个智能体 i使用局部数据（如 $\tau_{t,i}$, $a_{t,i}$, $u_{t,i}$ 和 $z_{t+1,i}$）通过强化学习（*RL*） [13, 43, 54]优化其策略π_i。对于采用同伴激励的方法，塑形奖励会在该轨迹中直接取代$u_{t,i}$。策略梯度强化学习是近似最佳响应 π_i^* [13, 27, 52] $\pi_{i\,\phi_i} \approx \pi_i^*$ 的一种常用方法。函数逼近器 $\hat{}$,通过对 $J = \mathbb{E}_{\pi}[G_{0,i}]$ [51]的估计进行梯度上升来训练，其中策略梯度近似为 [42]:

$$g = (G_{t,i} - b_i(s_t))\nabla_{\phi i} log\hat{\pi}_{i,\phi_i}(a_{t,i}|\tau_{t,i}) \tag{2}$$

这里， $b_i(s_t)$ 是依赖于状态的基线方法，通常通过学习得到的价值函数 $\hat{V}_{i,\ \omega_i}(\tau_{t,\ i}) \approx V_i^{\pi}(s_t)$ [13] 进行近似。为简便起见，我们省略参数，记作 $\hat{\pi}_i$、 $\hat{V}_i$。现代行动者-评论家方法会加入此类方差缩减基线、熵奖励或集中式评论家，但仍然只优化每个智能体自身的回报，这在社会困境中仍与社会福利不一致。此外，独立学习 会在智能体同时适应时引入非平稳性 [16, 22], , 除非修改奖励或激励，否则往往会导致过度贪婪的行为和相互背叛。

3 动态奖励激励

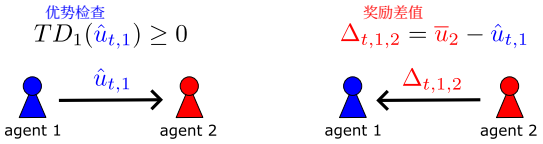
我们假设采用算法 1 中形式化定义的去中心化多智能体强化学习设置：在每个时间步，t 每个智能体 i 根据历史记录 $\tau_{t,i}$、策略 $\hat{\pi}_i$ 和价值函数 V_i，通过局部观测 $z_{t,i}$ 观测其邻域 $\mathcal{N}_{t,i}$，并执行动作 $a_{t,i} \sim \pi_i(\cdot|\tau_{t,i})$。DRIVE 仅在激励交换中使用 $\mathcal{N}_{t,i}$；底层策略不会接收环境观测之外的额外消息。随后，环境转移到新状态 $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$，每个智能体 i 通过观测 $z_{t+1,i}$ 和奖励 $\hat{u}_{t,i}$ 获取该状态信息。奖励是将环境奖励 $u_{t,i}$ 通过外部且可能未知的奖励变化函数 f_{mod}（算法 1，第 14 行）处理后得到的，该函数模拟修改后的规范、供需变化或传感器退化 [12, 56]。在第 5 节中，我们分析 f_{mod} 的一个仿射特例，即应用一个随时期变化的共享映射。所有智能体都会收集各自的经验元组 $e_{t,i} = \tau_{t,i}, a_{t,i}, \hat{u}_{t,i}, z_{t+1,i}$，用于 PI 交换以及对 $\hat{\pi}_i$ 和 $\hat{V}_i$ [52] 的独立更新。我们假设每个智能体 i 在每个时间步都有一个依赖于领域的邻域 $\mathcal{N}_{t,i} \subseteq D\backslash\{i\}$，如第 2 节所述。算法 1 第 13–20 行对应一个简短的并行通信阶段，在此阶段，相邻智能体会交换 DRIVE 请求和响应。

算法1 使用 DRIVE 进行多智能体学习

1: 初始化 $\hat{\pi}_i$ 的参数并 $\hat{V}_i$ 对所有智能体 $i \in \mathbb{D}$
2: **对于** 时期 $m \leftarrow 1, \dots, E$ **执行**
3: 设置 $u_i \leftarrow 0$ 对所有智能体 $i \in \mathbb{D}$ ▷ 重置平均奖励
4: 采样 s_0 并设置 $\tau_{0,i}$ 用于所有智能体 $i \in \mathbb{D}$
5: **对于**时间步 $t \leftarrow 0, \dots, H-1$ **执行**
6: **对于**智能体 $i \in \mathbb{D}$ **执行** ▷ 独立决策
7: 观察局部邻域 $\mathcal{N}_{t,i}$
8: $a_{t,i} \sim \hat{\pi}_i(\cdot|\tau_{t,i})$
9: **结束循环**
10: $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$ ▷ 执行联合动作 a_t
11: $u_{t,i\in\mathbb{D}} \leftarrow \mathbb{U}(s_t, a_t)$ ▷ 环境奖励
12: $z_{t+1,i\in\mathbb{D}} \leftarrow \Omega(s_{t+1})$
13: **对于**智能体 $i \in \mathbb{D}$ **执行** ▷ 并行通信
14: $\hat{u}_{t,i} \leftarrow f_{mod}(u_{t,i}, m)$ ▷ 外部奖励变化
15: $\bar{u}_i \leftarrow (t \cdot \bar{u}_i + \hat{u}_{t,i})/(t+1)$ ▷ 平均奖励
16: $u^{DRIVE}_{t,i} \leftarrow DRIVE(\hat{V}_i, \mathcal{N}_{t,i}, \hat{u}_{t,i}, \bar{u}_i)$ ▷ 算法 2
17: $e_{t,i} \leftarrow \langle \tau_{t,i}, a_{t,i}, u^{DRIVE}_{t,i}, z_{t+1,i} \rangle$
18: 更新 $\tau_{t,i}$ 为 $\tau_{t+1,i}$ 并存储 $e_{t,i}$
19: **结束循环**
20: **结束循环**
21: **对**智能体 $i \in \mathbb{D}$ **执行** ▷ 独立更新
22: 更新 $\hat{\pi}_i$ 并 $\hat{V}_i$ 使用时期 m 的所有 $e_{t,i}$
23: **结束循环**
24: **结束循环**

DRIVE 令牌交换。 DRIVE 采用受 Trivers [44] 启发并见图 3 所示的互惠激励方案，以交换动态激励，实现分布式奖励塑形。在请求阶段（图 3a），每个智能体 i 检查其优势或时序差分残差 $TD_i(\hat{u}_{t,i})$ [41]：

$$TD_i(\hat{u}_{t,i}) = \hat{u}_{t,i} + \gamma \hat{V}_i(\tau_{t+1,i}) - \hat{V}_i(\tau_{t,i}) \tag{3}$$



(a) DRIVE 请求 **(b) DRIVE 响应**

图 3：DRIVE 交换方案。(a) 如果 $TD_1(\hat{u}_{t,1}) \geq 0$ (公式 3)，智能体 1 会将其奖励 $\hat{u}_{t,1}$ 作为请求发送给邻居智能体 2。(b) 智能体 2 计算其当前时期 m 的平均奖励 $\bar{u}_i$ 与收到的请求 $\hat{u}_{t,i}$ 之间的差值 $\Delta_{t,1,2}$，并将其作为响应发回，用于调整两个智能体的奖励（公式 4）。

如果优势为非负，智能体 i 将其奖励 $\hat{u}_{t,i}$ 作为请求发送给所有其他智能体 $j \in \mathcal{N}_{t,i}$。在响应阶段（图 3b）中，每个接收到请求的邻居 $j \in \mathcal{N}_{t,i}$ 会将请求 $\hat{u}_{t,i}$ 与其历元平均奖励 $\bar{u}_j$ 进行比较，并计算 $\Delta_{t,1,2} = \bar{u}_2 - \hat{u}_{t,1}$，然后将其作为响应发回。完成这次交换后，每个智能体的 DRIVE 奖励 $u^{DRIVE}_{t,i}$ 计算如下：

$$u^{DRIVE}_{t,i} = \hat{u}_{t,i} - min\{\langle \Delta_{t,j,i} \rangle_{j\in\mathcal{N}_{t,i}}\} + min\{\langle \Delta_{t,i,j} \rangle_{j\in\mathcal{N}_{t,i}}\} \tag{4}$$

使用响应方的轮次平均值 $\bar{u}_j$ 而非其即时奖励，使 DRIVE 能够针对系统性利用，而非单次噪声结果作出反应：只有近期平均回报持续低于邻居请求的智能体，才会产生较强的负 Δ 项。其他组合（即时–即时、即时–平均、平均–平均）也都是可行的，并可能带来不同的权衡，这是未来工作中一个值得探索的方向。附录 A.1 提供了初步比较。

算法 2 动态奖励激励交换

过程 $DRIVE(\hat{V}_i, \mathcal{N}_{t,i}, \hat{u}_{t,i}, u_i)$
  $\hat{u}_{req} \leftarrow \infty, \hat{u}_{res} \leftarrow 0$
  若 $TD_i(\hat{u}_{t,i}) \geq 0$ **则** ▷ 根据公式 (3) 进行评估
    向所有 $j \in \mathcal{N}_{t,i}$ 发送请求 $\hat{u}_{t,i}$（图 3a）
  结束 if
  对邻居智能体 $j \in \mathcal{N}_{t,i}$ **执行** ▷ 请求
    如果请求 $\hat{u}_{t,j}$ 由 j **接收则**
      $\Delta_{t,j,i} \leftarrow u_T - \hat{u}_{t,j}$
      $\hat{u}_{req} \leftarrow$ 最小$\{\hat{u}_{req}, \Delta_{t,j,i}\}$
      向智能体 j 发送响应 $\Delta_{t,j,i}$（图 3b）
    结束如果
  结束循环
  如果 $\hat{u}_{req} = \infty$ **则** $\hat{u}_{req} \leftarrow 0$ ▷ 如果未收到请求
  如果 $TD_i(\hat{u}_{t,i}) \geq 0$ **则** ▷ 如果已发送请求
    $\hat{u}_{res} \leftarrow \infty$
    对于邻居智能体 $j \in \mathcal{N}_{t,i}$ **执行** ▷ 响应
      $\hat{u}_{res} \leftarrow min\{\hat{u}_{res}, \Delta_{t,i,j}\}$
    结束循环
  结束条件判断
  回报 $\hat{u}_{t,i} - \hat{u}_{req} + \hat{u}_{res}$ ▷ 塑形奖励（公式 4）
结束过程

分布式奖励塑形。DRIVE 奖励$u_{i\,i}^{DRIVE}$用于根据公式 (2)，采用任意强化学习算法（例如策略梯度方法）更新相应智能体的策略。形式上，每当启用 DRIVE 时，我们便在公式 (2) 中用 $u_{i\,i}^{\text{DRIVE}}$替换$u_{t,i}$，从而得到回报 $G_{t,i}$（参见算法 1，第 17 行）。因此，第 2 节描述的是环境层面的回报，而 DRIVE 则定义了如何将这些回报转换为强化学习更新所优化的塑形奖励。DRIVE 请求中的优势 $\text{TD}_i(\hat{}u_{t,i})$ 非负条件，是为了揭示背叛智能体；在社会困境 [3] 中，这类智能体通常比被利用智能体具有更大的优势。如果存在单方面背叛行为，背叛智能体 i 将受到被剥削程度最高的邻居智能体 $j \in \text{N}_{t,i}$ 的惩罚，因为 $\Delta_{t,i,j} < 0$，这一点由公式 (4) 中的最小值聚合项[1] 确保。然而，如果所有智能体的合作程度相同，那么所有 $\Delta_{t,i,j} = 0$，也就不会产生额外的奖励或惩罚。算法 2 给出了任意智能体 i 在时间步 t 时 DRIVE 的完整形式。$\hat{}V_i$ 是用于根据公式 (3) 计算 $\text{TD}_i(\hat{}u_{t,i})$ 的近似价值函数，$\text{N}_{t,i}$ 是当前局部邻域，$e_{t,i}$ 是经验元组，而 u_i 是智能体 i当前的平均奖励。

–

4 相关工作

社会困境中的多智能体强化学习。 多智能体强化学习在多个领域取得了显著进展 [5, 26, 43, 48]。在分散式社会困境和序贯社会困境中，关键挑战是在不采用集中式控制的情况下解决激励错位问题。近期研究通过局部、同伴诱导或社会条件化的奖励塑造机制 [2, 18, 36, 52],来应对这一问题；这些机制适用于具有观测受限智能体的混合动机环境，因此与我们的设定一致。相比之下，更广泛的多智能体强化学习文献中的许多方法——例如缓解合作场景中非平稳性的方法 [4, 30, 49], 、全局信息引导的奖励塑造 [9, 23], 、需要访问对手参数的对手塑形 [13, 25, 50, 57],以及集中式激励设计 [15, 53] ——都假设拥有全局信息或集中协调，因此无法扩展到分散式社会困境中。

同伴激励。许多 PI 方法通过奖励交换，以分布式方式促进相互合作 [55]。赠予机制通过为每个智能体 i 的动作空间增加一个奖励动作，来激励其他智能体 $j \in \text{N}_{t,\,i}$ [28]。相比之下，DRIVE 建立在互惠奖励塑形之上，无需扩展动作空间。Schmid 等人 [39] 提出了基于市场的同伴激励，智能体可以据此建立双边协议。文献 [47] 提出了一种公开制裁方法，智能体可根据已知的群体行为模式相互奖励或惩罚。Hughes 等人 [18] 定义了一种不平等厌恶（*IA*）方案，该方案基于非负奖励差值，并由两个领域相关系数加权。激励其他学习智能体（*LIO*）在完全可观测的条件下，基于其他所有智能体的联合动作 $j \neq i$ [52]，自动为每个智能体 i 学习一个激励函数。Phan 等人 [35] 提出了相互认可代币交换（MATE），这是一种两阶段协议：智能体交换固定代币 x，合作的智能体发放代币，而被利用的智能体接收代币，由此得到图 4 所示的修改后收益矩阵。大多数 PI 方法对环境的奖励值很敏感，因为它们依赖固定的激励值或领域相关参数，因此在奖励值发生变化时无法合作（图 1）。以囚徒困境为例，图 4 展示了 DRIVE、MATE、LIO 和 IA 的修改后收益矩阵。根据这些矩阵，DRIVE 是唯一不依赖任何特定超参数或显式学习值的方法。IA 只修改单边合作和背叛，但依赖两个领域相关系数

α β和 。[18]

LIO 根据其他智能体的动作学习激励函数，因此需要足够的经验和时间来相应地适应。对于 MATE，合作取决于一个恰当的全局代币 $\mathbf{x} \geq max\{P - S,\ ^{T-R}_{3}\}$，该值可以从修改后的收益矩阵中推导得出。在图 2b 的示例中，x 至少应为 1，即文献 [36] 中 MATE 使用的原始代币值。有趣的是，其他 PI 方法也采用了这一数值，却没有进一步分析或质疑 [28, 39]。

组队。 除独立学习者之外，近期研究还探讨了社会偏好、团队或联盟结构如何在混合动机情境中塑造激励，例如异质的基于社会价值取向的奖励塑造、基于团队的奖励共享，以及通过无政府代价最小化实现的自适应奖励混合 [14, 31, 34, 37]。这些方法通常假设固定的或自发形成的群体结构，或修改每个智能体的全局效用函数。相比之下，DRIVE 不引入显式联盟或共享奖励：激励完全通过局部的双边奖励交换产生，从而无需预先设定团队或进行群体层面的奖励混合，即可实现去中心化对齐。

5 理论分析

下文将分析 DRIVE 在一般囚徒困境中的激励对齐性质，以及其对变化中的奖励的不变性。我们强调，该机制并不保证学习动态的收敛，而是重塑激励，使合作在重复交互中成为个体理性的选择。尽管我们的形式化讨论聚焦于经典囚徒困境，但许多众所周知的序贯社会困境（SSDs），例如 Coin 和 Harvest，都体现了类似囚徒困境的激励结构，这一点已在 [23] 中通过实验证明，并在附录中加以解释。在第 7 节中，我们进一步表明，DRIVE 在包含两个以上智能体的更复杂序贯社会困境中同样能取得有希望的结果。

5.1 一般囚徒困境中的 DRIVE

在具有收益表的社会困境中，如图 2 和公式 (1) 的不等式 $T > R > P > S$ 所示，长期来看，DRIVE 激励不会改变相互合作的收益 R 和相互背叛的收益 P，因为对于双方智能体，奖励差值 $\Delta_{t,\,i,\,j}$ 都会为零。在 $R > P$的情况下，这仍会使相互合作优于相互背叛。然而，如果发生（反复的）单边背叛，背叛的智能体 i 会向被剥削的智能体 j发送请求 T；后者的响应会使差值变为 $\Delta_{t,\,i,\,j} = S - T < 0$ （最坏情况下），此时背叛智能体 i的奖励会变为 $T + \Delta_{t,\,i,\,j} = T + S - T = S$，而被剥削智能体 j的奖励会变为 $S - \Delta_{t,}$

$_{i\,j} =$, ( )

$S - S - T = T$ ，如公式 ( ) 所示。如图所示，^{4}T 和 S 的 DRIVE ^{4}收益转换使双方智能体都能克服贪婪和恐惧，从而激励合作。

[1]这些 最小值 聚合可以替换为总和或均值，类似于 [18]。不过，这会削弱受剥削程度最高的智能体相较于其他智能体的影响力，从而容忍个体的不满，而这种不满可能会在后续轮次中蔓延。

DRIVE Payoffs

	C	D
C	R, R	T, S
D	S, T	P, P

MATE Payoffs

	C	D
C	$R+2\mathrm{x}, R+2\mathrm{x}$	$S+\mathrm{x}, T-\mathrm{x}$
D	$T-\mathrm{x}, S+\mathrm{x}$	P, P

Inequity Aversion Payoffs

	C	D
C	R, R	$S-\alpha(\mathrm{T}-\mathrm{S}),$ $T-\beta(\mathrm{T}-\mathrm{S})$
D	$T-\beta(\mathrm{T}-\mathrm{S}),$ $S-\alpha(\mathrm{T}-\mathrm{S})$	P, P

LIO Payoffs

	C	D
C	$R+\mathrm{x}_{\mathrm{C}}^{1},$ $R+\mathrm{x}_{\mathrm{C}}^{2}$	$S+\mathrm{x}_{\mathrm{D}}^{1},$ $T+\mathrm{x}_{\mathrm{C}}^{2}$
D	$T+\mathrm{x}_{\mathrm{C}}^{1},$ $S+\mathrm{x}_{\mathrm{D}}^{2}$	$P+\mathrm{x}_{\mathrm{D}}^{1},$ $P+\mathrm{x}_{\mathrm{D}}^{2}$

图4：不同 PI 方法的 ModifiedPD 收益矩阵，收益修改以红色标出 [18, 36, 52]。

定理 1. *DRIVE* 在广义双智能体囚徒困境中协调激励，通过逆转诱惑收益和受骗收益，使相互合作成为双方智能体的占优策略。

证明。使用奖励差值 $\Delta_{t,1,2} = u_2 - \hat{u}_{t,1}$ 塑造两个智能体的奖励（公式 4），会使原始囚徒困境收益矩阵（图 2a）中的收益 T 和 S 互换。根据图 4 中得到的收益矩阵，DRIVE 智能体能够克服贪婪，因为 $R > S$；也能够克服恐惧，因为 $T > P$。因此，在囚徒困境中，DRIVE 智能体总是受到激励而选择合作。□

总体而言，DRIVE 不依赖固定的激励值或特定领域的参数。因此，只要奖励变化仍满足关于贪婪和恐惧的囚徒困境不等式，它就能应对这些变化，如图 1 的示例所示。为清楚起见，附录 A.1 给出了一个扩展示例，具体展示了单个囚徒困境交换步骤中的塑造机制。请注意，公式 (4) 实际上将阶段博弈转变为一个以 (C, C) 为唯一纳什均衡的协调博弈。这体现了标准的奖励设计实践：我们不依赖学习者推断长期外部性，而是调整即时收益，使合作符合个体理性，同时保持环境动态不变。

5.2 对奖励变化的不变性

现在我们分析 DRIVE 和 MATE 在 x 的不变性，即原始收益 R、P、T 和 S 会被某个外部变化函数 f_{mod} 动态改变。

定义 2（奖励变化函数）。令 $r_{t,i}$ 表示智能体 i 在时间 t 的原始环境奖励。时期 m 中的修正奖励定义为 $r'_{t,i} = f_{mod}(r_{t,i}, m) = c_m r_{t,i} + b_m$，其中 $c_m > 0$ 和 $b_m \in \mathbb{R}$ 是依赖于时期的缩放参数和平移参数，它们在时期 m 内统一适用于所有智能体和时间步。这种共享的仿射映射在改变奖励幅度和偏移量的同时，保留了博弈的策略结构（例如囚徒困境不等式）。我们重点研究这一宽泛但易于处理的逐时期仿射变换类别：任意的调度 c_m 和 b_m 涵盖了第7.2节中的所有奖励动态，并且与我们的逐时期归一化方案兼容。附录 A.4 还提供了更多示例和边缘情形，包括 $c_m \approx 0$ 的时期。

定理 3. 在一般囚徒困境中，当环境奖励因 f_{mod} 而改变时，*MATE* 智能体不具备对变化中的奖励的不变性。

证明。根据图 4 中的 MATE 收益矩阵，全局代币 $\mathrm{x} = x_1 = x_2 > 0$ 必须满足 $\mathrm{x} \geq max\{P - S, \frac{T-R}{3}\}$，才能实现涌现合作。当变化函数 f_{mod} 的选择使得 $f_{mod}(P - S) > \mathrm{x}$ 或 $f_{mod}(\frac{T-R}{3}) > \mathrm{x}$ 时，MATE 智能体就不再能够保证彼此合作。□

请注意，定理 3 可推广至任何使用固定同伴激励值的其他 PI 方法。原则上，可以在每次 f_{mod} 发生变化后重新调整 x，但这需要额外的全局知识或元优化，而 DRIVE 会通过局部奖励差异自动适应。

定理 4. 在一般囚徒困境中，当环境奖励因 f_{mod} 而改变时，*DRIVE* 智能体对变化中的奖励具有不变性。

证明。对原始收益 R、P、T 和 S 进行任意 f_{mod} 改动，都会按比例改变图 4 中的 DRIVE 收益矩阵。因此，只要克服贪婪和恐惧的各项不等式仍然成立，DRIVE 智能体就始终受到激励而相互合作。□

现在，我们将不变性分析从囚徒困境扩展到一般序贯社会困境，并采用带回报归一化的策略梯度方法，具体细节见附录 A.6。

定理5。 在使用归一化策略梯度方法训练时，*DRIVE* 智能体在 *SSD* 中对变化中的奖励具有不变性；在此类情形下，环境奖励由 f_{mod} 改变（附录 *A.6* 中的引理*9*）。

考虑序贯社会困境，其中回报按时期归一化为均值为零、方差为一。若每个时期的奖励都经过正仿射映射变换，那么 (i) 标准策略梯度学习和 (ii) DRIVE 的塑形奖励相对于归一化回报 $G_{t,i}$ 均保持不变。因此，DRIVE 所带来的激励对齐在此类变换下进行逐时期奖励缩放和平移时仍然成立。总体而言，在逐时期归一化下，基线策略梯度更新和 DRIVE 的塑形奖励对于共享的逐时期正仿射奖励变换均保持不变（附录 A.6 中的引理9）。

证明。该结果由两个部分得出。首先，引理9表明，在逐时期归一化下，标准策略梯度强化学习对于正仿射奖励变换具有不变性。其次，DRIVE 根据时期平均奖励与瞬时奖励之间的差异来塑形奖励。应用 f_{mod} 可得 $\Delta'_{t,i,j} = f_{mod}(u_i) - f_{mod}(\quad - \quad u_{t\,i} \quad i\,j \quad i \quad = ,) \quad f_{mod}\ u_{t\,i}^{DRIVE} \quad f_{mod}\ x = c_m x + b_m \quad c_m = f_{mod}(\Delta_{t,\ ,\ }),$ 因此塑形奖励变为 $\hat{u}_t^{DRIVE'}, (,)$。对于 ( )，归一化消除了 和 b_m，使 $G_{t,i}$ 保持不变。因此，基线策略梯度更新和 DRIVE 的激励效应都对这类变换具有不变性。□

5.3 不遵守协议的影响

前述分析假设所有智能体都如实且同步地遵循 DRIVE 协议，即每个智能体都分享其历元平均奖励、响应请求，并始终如一地应用塑造规则。然而在实践中，可能会出现通信故障、延迟或策略性虚报。本文分析此类不遵守行为相较于其他同伴激励方法的影响；详细示例请参见附录 A.2。

完全遵循。 当所有智能体都如实遵循协议时，定理 1 和定理 4–5 的结论均适用。收益 T 和 S

在单边背叛时会互换，从而确保激励方向与合作一致。只要保持贪婪和恐惧的特征不等式，这一推理便可推广至囚徒困境之外的序贯社会困境。

未发送请求（尽管存在 TD ≥ 0）。隐瞒请求的智能体仍然可以收到请求并如实响应。对于其邻居而言，这意味着少了一项互惠请求。由于采用最小值聚合，只要其他智能体仍发送请求，它们受到的惩罚程度可能会略有增加，但不会受到显著影响。对于不发送请求的智能体本身，$\hat{u}_{res} = 0$，因此它会失去获得互惠改善的机会，并受到间接惩罚。如果该智能体背叛，它或许能暂时避免惩罚，但总体而言，它会失去相互塑造带来的益处，也无法获得稳定优势。

未发送响应。如果智能体没有响应请求（由于通信丢失或故意隐瞒），其影响取决于具体情境：如果该智能体正在背叛，隐瞒其负面 Δ 会减轻对被其利用的邻居施加的惩罚。不过，它也会失去从互惠中受益的机会，因此故意不响应并不理性。如果该智能体正在合作，缺失的（通常为正面或较小的） Δ 影响有限，因为最小值聚合会削弱单个缺失贡献的影响。因此，不响应对背叛者而言大多是中性或略有利，但总体上代价高昂。

虚假请求或响应。存在两种不当行为：发送带有 $TD < 0$的请求：这可能是由噪声或训练早期误估造成的。这类请求通常只会产生很小或为正的差值，会被最小值算子过滤掉，因此影响甚微。发送虚假响应：智能体可能故意谎报 Δ，以扭曲塑造过程。在最坏情况下，最小值聚合可能会对请求者施加不合理的惩罚，进而可能损害合作。这体现了 DRIVE 的脆弱性与局限性。

完全不遵从。如果所有智能体都拒绝响应，或通信全面失效，那么所有塑形项都会消失。系统将退化为没有激励的纯多智能体强化学习，相互背叛也会再次成为一种可能的均衡。这类似于 MATE 等固定代币方法：一旦其协调机制不可用，也会失效。

5.4 概念讨论

DRIVE 提供了一种简单、适应性强的 PI机制，可在奖励变化中的社会困境中促进合作。它通过本地同伴交互运作，无需中央控制器、全局知识或全对全通信。与 [18, 36, 52], 算法 2 类似，就激励交换而言，其最坏情况下的规模随 O(4(n − 1)) 线性增长。同时，DRIVE 依赖真实的同伴通信才能正常运行。部分遵从会带来<用风格 id='3'>优雅降级：只要至少有一名诚实的响应者提供反馈，背叛者仍会受到惩罚；若所有智能体都未能遵从，DRIVE 就会退化为基线多智能体强化学习动态。与固定令牌方法相比，DRIVE 至少同样稳健，而且通常更具适应性，但它依赖严格遵守交换协议，这引入了一种新的潜在脆弱性。设计能够应对部分遵从的协议变体——例如采用冗余聚合规则、随机审计或替代性的塑造算子——仍是未来工作的重要方向。

一个关键的建模假设是邻域 $N_{t,i}$ 它限定了哪些智能体可以交换激励。在我们的实验中，该邻域与环境相关，具体定义将在下一节介绍。关键在于，DRIVE 的最小值聚合意味着背叛者的塑形奖励仅取决于其剥削程度最高的邻居，因此，与简单的奖励共享方案不同，惩罚不会因群体规模扩大而被稀释。附录 A.2 将先前的遵从性分析扩展到一般的 N 智能体系统和任意通信图，并对由此产生的鲁棒性进行形式化：当且仅当合规智能体构成一个支配集时，才能惩罚背叛者。因此，在部分遵从的情况下，合作激励会平稳减弱，而不会突然崩溃。

如图 4 所示，在修改后的囚徒困境收益表中，DRIVE 仅依赖环境奖励，无需像 MATE 那样设置固定的激励值 x，无需像 IA 那样使用领域相关系数 α，β，也无需像 LIO 那样进行耗时的激励学习 x^i_C，x^i_D。请注意，鉴于一般囚徒困境所体现的贪婪和恐惧特征，上述理论分析通常适用于各种社会困境。因此，DRIVE 可用于任何社会困境，其中奖励通过外部且未知的变化函数 f_{mod} 动态变化。如果合作条件（即社会困境的特征不等式）保持不变，DRIVE 就能适应环境并维持合作，如图 1 所示，后文也将对此进行验证。

6 实验设置[2]

我们使用三种基于 [13, 32] 的知名序贯社会困境。在每个时间步，智能体动作的执行顺序都会随机化，以解决冲突，例如多个智能体同时踩到一枚硬币或互相标记。所有序贯社会困境均代表囚徒困境实例，这一点已通过实证展示于 [23] 和附录 B。由于我们通过 f_{mod} 修改奖励，因此使用领域特定的指标评估合作情况。

迭代囚徒困境。对于 IPD，我们使用图 2b 所示的收益矩阵。两个智能体在每个时间步 t 都会观察前一个联合动作$z_{t,i} = a_{t-1}$；在初始状态 s_0 下，该动作是零向量。纳什均衡是始终背叛（DD）。一个回合由 $H = 150$ 次迭代组成，我们设定 $\gamma = 0.95$。邻域 $\mathrm{N}_{t,i} = \{j\}$ 定义为另一智能体 $j \neq i$。我们将每回合的合作（CC）次数除以 H，以此衡量合作率。

Coin. *Coin-2* 和*Coin-4* 是序贯社会困境，由 $n \in \{2,4\}$ 个不同颜色的智能体组成。它们从随机位置开始，必须收集颜色和位置随机的硬币 [13, 24]。如果智能体收集到一枚硬币，就会获得 +1 的奖励。然而，如果硬币的颜色与收集它的智能体不同，那么实际颜色匹配的另一名智能体会受到 -2 的惩罚。硬币被收集后，会以新的颜色随机重新出现。所有智能体都能观察整个场地，并可以向北方、南方、西方和东方移动。每个智能体只能判断硬币的颜色是否与自身颜色相同。经过 $H = 150$ 个时间步后，一轮即告结束，并且我们设定 $\gamma = 0.95$。邻域 $\mathrm{N}_{t,i} = \mathbb{D} - \{i\}$ 定义为所有其他智能体 $j \neq i$。我们根据每个智能体收集的硬币，计算“自身颜色硬币”比例P（自身颜色硬币） $= \frac{\#\text{收集到的同色硬币数}}{\#\text{收集到的所有硬币数}}$。

[2]实现可在 https://github.com/philippaltmann/DRIVE 获取。

Harvest. *Harvest-12* 是一个序贯社会困境，由 $n = 12$ 个智能体组成。智能体从随机位置开始，必须收集苹果。苹果再生率取决于周围苹果的数量 [32]。如果所有苹果都被采摘，直到本轮结束前都不会再有苹果生长。每个时间步，所有智能体都会受到 -0.01 的时间惩罚。每收集一个苹果，智能体就会获得 +1 的奖励。所有智能体都有 7×7 的视野，可以选择不采取行动、向北方、南方、西方或东方移动，也可以使用指向特定基数方向、宽度为 5 的标记光束，标记视野内的其他智能体。如果智能体被标记，则在 25 个时间步内无法行动。标记不会带来任何奖励。经过 $H = 250$ 个时间步后，一轮即告结束，并且我们设定 $\gamma = 0.99$。邻域 $N_{t,i}$ 定义为所有其他智能体 $j \neq i$ 处于智能体 i 的视野内。我们将可持续性定义为采摘苹果的平均时间步数。

多智能体强化学习算法。 为了将奖励变化对PI机制的影响与基础强化学习算法的影响区分开来，我们按照 [17], 中的建议，基于标准策略梯度算法（公式 2）并使用归一化回报（附录 A.6）实现了所有方法，包括 DRIVE。该方法也作为不采用任何基于PI的奖励塑形的朴素学习基线 [13]。我们采用 LIO、MATE（使用 $x = 1$，如 [36], 所建议）以及 IA（使用 $\alpha = 5$ 和 $\beta = 0.05$），作为最先进的PI基线方法 [18, 36, 52]。对于*IPD*和*Coin-2*，由于深度神经网络二阶导数计算的计算开销很大，我们直接纳入了 [13, 57], 中报告的对手塑形技术 LOLA-PG 和 POLA-DiCE 的表现。

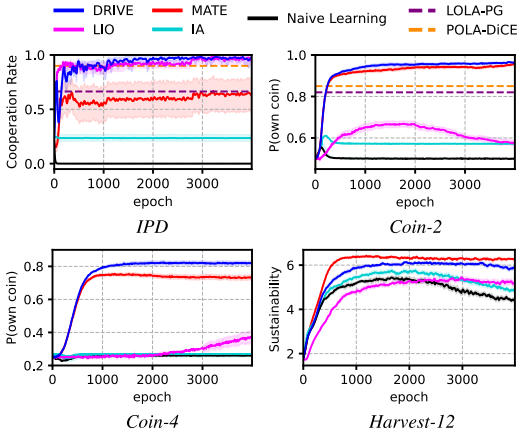


图5：在奖励不变的序贯社会困境中，DRIVE和其他基线方法的平均进展。阴影区域表示95%的置信区间。LOLA-PG和POLA-DiCE的结果引自 [13, 57]。

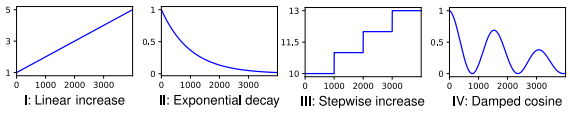


图6：根据算法1第15行，f_{mod} 在各时期 m 使用的奖励变化函数。

7 实验结果

在每项实验中，所有相应算法均独立运行 20 次，每次运行 $E = 4,000$ 轮，每轮包含 10 个回合，并报告平均进展和 95%置信区间。

7.1 标准设置——无奖励变化

设置。我们比较了在所有轮次中均不发生奖励变化的情况下，DRIVE 与 *IPD*、*Coin-2*、*Coin-4* 和 *Harvest-12* 中的基线方法，即 $f_{mod}(u, m) = u$（算法1，第15行）。

结果。结果如图5所示。与所有基线方法相比，DRIVE 在*IPD* 中实现了具有竞争力且稳定的合作。在*Coin-2*中，DRIVE 和 MATE 的“自身硬币”比例最高，显著优于 POLA-DiCE；POLA-DiCE 的合作程度则高于 LIO、IA 和 LOLA-PG。在两个规模更大的 SSD 中，即*Coin-4*和*Harvest-12*，DRIVE 和 MATE 的合作程度最高；其中，DRIVE 在*Coin-4*中的合作程度略高，而 MATE 在*Harvest-12*中的合作程度略高。

7.2 动态奖励变化

设置。接下来，我们使用不同的奖励变化函数 f_{mod} 对 DRIVE 和基线方法进行评估，如图6所示：

(1) 线性增长： $f^{I}\mathbf{mod}(u, m) = u(\eta m + 1)$ **(2) 指数衰减：** $f^{II}\mathbf{mod}(u, m) = ue^{-\eta m}$ **(3) 阶梯式增长：** $f^{III}\mathbf{mod}(u, m) = u(\lfloor \eta m \rfloor + \chi)$ **(4) 阻尼余弦：** $f^{IV}\mathbf{mod}(u, m) = \eta + u(1 - \frac{m}{E})\cos^2(2\eta m)$ —

我们设定 $\eta = 0.001$ 和 $\chi = 10$。选择这些函数是为了呈现不同的情景，其中奖励尺度或偏移量要么不断增加，要么在极限情况下收敛至零，同时不改变社会困境中的基本不等式。(1) 可作为外部因素的例子，例如整个系统的需求不断增加。(2) 则可模拟内部因素，例如感官磨损或硬件随时间老化。(3) 则展示了幅度较大的变化，这通常出现在不同模拟之间，或模拟与现实之间的转换中。最后，(4) 可解释为人类反馈或规格在线修改所导致的不规则性。

结果。结果如图7所示。使用函数 $f^{I}{}_{mod}$、$f^{II}{}_{mod}$ 和 $f^{III}{}_{mod}$ 进行训练，并不会显著影响 DRIVE 的学习曲线。函数 $f^{IV}{}_{mod}$ 会导致 DRIVE 的学习曲线偶尔出现下跌，但随后会迅速恢复到原有水平。这些偶发的下跌是由阻尼余弦函数的极小值造成的，此时函数值接近零，会暂时使公式 (1) 的囚徒困境不等式失效。然而，这属于数值问题而非概念问题，因为 DRIVE 随后会迅速恢复。对于 *Coin-4* 中的任何变化函数，LIO 都具有鲁棒性，但在其他领域中，其合作水平会显著恶化。尽管能够自动学习激励函数，LIO 对 $f^{IV}{}_{mod}$ 尤其敏感。在 *Coin-2*和 *Coin-4* 中，人工智能会受到影响，并且其表现始终不如朴素学习。只有在使用 $f^{I}{}_{mod}$ 进行训练时，MATE 才能抵抗 *Harvest-12*中的奖励变化。在其他所有情况下，MATE 的合作水平都会显著恶化，某些情况下甚至比朴素学习更差。使用 f^{IV} 时，学习到的行为最不稳定；与 DRIVE 不同，此时 MATE 无法从学习曲线的下跌中恢复。DRIVE 是唯一一种在

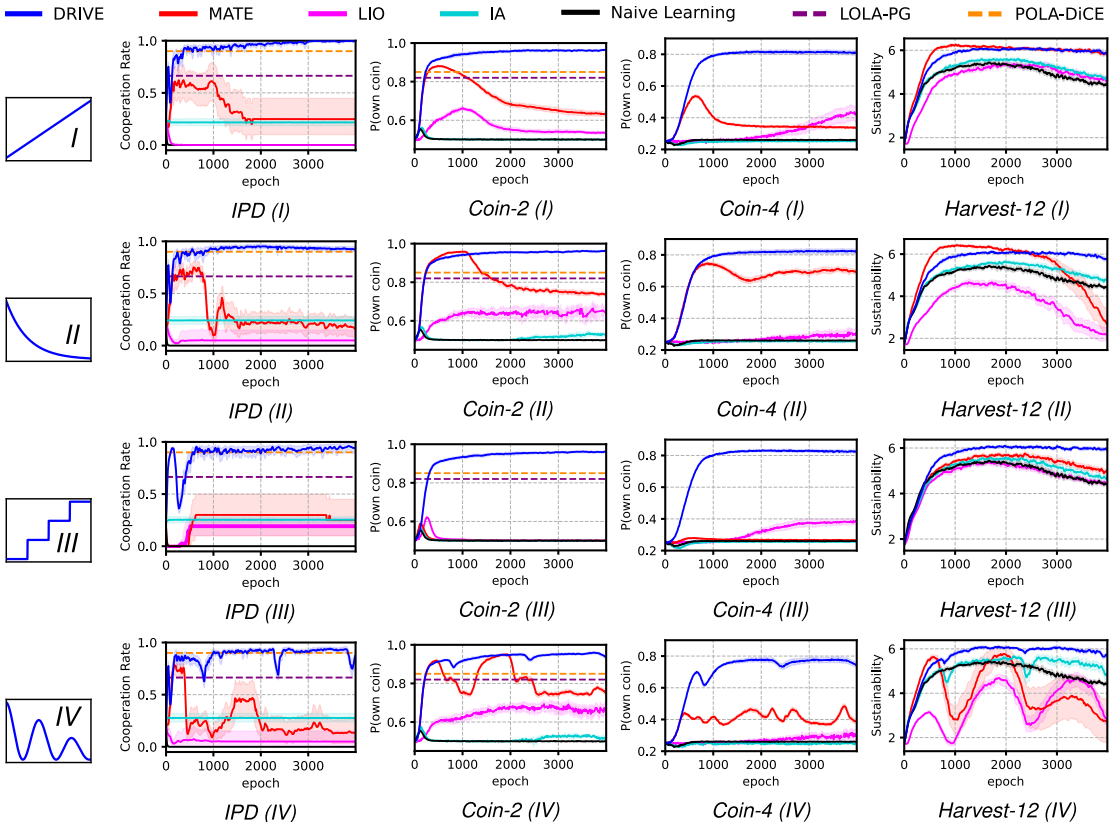


图7：在具有不同奖励变化函数 f_{mod}（I、II、III 和 IV）的序贯社会困境中，DRIVE 与其他基线方法的平均进展，如顶部图例和图6所示。阴影区域表示 95%置信区间。LOLA-PG 和 POLA-DiCE 的结果来自 [13, 57]。由于回报归一化，非PI方法朴素学习、LOLA-PG 和 POLA-DiCE 不受 f_{mod} 的奖励变化影响，具体说明见附录 A.6。

IPD 和 *Coin-2* 中表现优于朴素学习、LOLA-PG 和 POLA-DiCE 的 PI 方法。附录C进一步提供了 *Harvest-12* 中的合作指标（社会福利、平等、可持续性和和平）。

8 讨论

我们提出了 DRIVE，这是一种自适应PI方法，旨在变化中的奖励条件下实现并维持社会困境中的合作。DRIVE 智能体以互惠方式交换奖励差异，从而以完全去中心化的方式激励相互合作。

我们的理论分析表明，DRIVE 如何在一般囚徒困境中克服贪婪和恐惧，实现面向相互合作的激励对齐，同时不受逐轮奖励偏移与缩放的影响。这与以往依赖固定激励值、领域相关系数或耗时的激励学习的PI方法形成对比。因此，DRIVE 更符合博弈论假设，在奖励不断变化的环境中部署也更安全。

我们的实验进一步表明，在多个存在动态奖励变化的序贯社会困境中，DRIVE 的表现显著优于 LIO、MATE 和 IA 等最先进的方法，也优于 POLA-DiCE 和 LOLA-PG 等对手塑形基线方法。结果证实，现有同伴激励方法对奖励大小十分敏感，并凸显了 DRIVE 在只要底层的策略不等式保持不变，就能在此类转变下维持合作。

尽管这些结果颇具前景，DRIVE 仍依赖真实且同步的同伴通信。部分不遵从会导致性能优雅降级。此外，当前的理论分析假设群体同质，且通信图完整或连接良好。在存在策略操纵、通信噪声或稀疏交互结构的更现实场景中，这些假设可能并不成立。

从这些观察中可以得出若干未来工作的方向。一条途径是开发能够应对部分不遵从或策略性不遵从的机制，例如通过冗余聚合、随机审计或共识系统 [2]。将理论分析扩展到异质群体和更一般的网络拓扑，是迈向现实应用的另一重要步骤。一个自然的下一步是研究学习形成的邻域和动态邻域，并将 DRIVE 与演化中的社会结构和联盟模型相结合，正如近期关于团队和联盟中奖励共享的研究所提出的那样。最后，纳入智能体识别功能或可实现有针对性的双边响应，例如针对对抗性对等体。

参考文献

[1] Philipp Altmann、Thomy Phan、Maximilian Zorn、Claudia Linnhoff-Popien 和 Sven Koenig。2026年。变化奖励下的动态激励合作。载于第25届自主智能体与多智能体系统国际会议（AAMAS '26）。https://openreview.net/forum?id=pbENOmnHWy [2] Philipp Altmann、Katharina Winter、Michael Kölle、Maximilian Zorn 和 Claudia Linnhoff-Popien。2025年。MEDIATE：相互认可的分布式激励确认代币交换。载于第17届国际智能体与人工智能会议（葡萄牙波尔图）（ICAART '25）。SciTePress, 33 44 10 5220 0013091900003890 –。https://doi.org/。/ [3] Robert Axelrod。。1984。《合作的演化》。Basic Books。https://doi.org/ 10 1007 978 3 319 16999 6_1220 1 。/ - - - - - Michael Bowling 和 [4] Manuela Veloso。。2002。《运用可变学习率进行多智能体学习》。《人工智能》136, 2 (2002), 215–250。[5] Lucian Buşoniu、Robert Babuška 和 Bart De Schutter。2010年。多智能体强化学习：综述。载于《多智能体系统与应用创新-1》。Springer, 183–221。[6] Allan Dafoe、Edward Hughes、Yoram Bachrach、Tantum Collins、Kevin R McKee、Joel Z Leibo、Kate Larson 和 Thore Graepel。2020年。合作型人工智能中的开放问题。arXiv预印本 arXiv:2012.08630 (2020)。[7] Richard Dawkins。2016年。《自私的基因》。牛津大学出版社。[8] Shuiguang Deng、Zhengzhe Xiang、Peng Zhao、Javid Taheri、Honghao Gao、Jianwei Yin 和 Albert Y Zomaya。2020年。可信物联网系统的边缘动态资源分配：一种强化学习方法。《IEEE工业信息学汇刊》16, 9 (2020), 6103–6113。[9] Sam Devlin、Logan Yliniemi、Daniel Kudenko 和 Kagan Tumer。2014年。基于潜在函数的多智能体强化学习差分奖励。载于2014年国际自主智能体与多智能体系统会议论文集。[10] AL Dimeas 和 ND Hatziargyriou。2010年。面向微电网的多智能体强化学习。载于IEEE PES全体会议。IEEE，1–8。[11] Gabriel Dulac-Arnold、Daniel Mankowitz 和 Todd Hester。2019年。现实世界强化学习面临的挑战。arXiv预印本 arXiv:1904.12901 (2019)。[12] Benjamin Eysenbach 和 Sergey Levine。2021年。最大熵强化学习（可证明地）解决了一些鲁棒强化学习问题。载于国际学习表征会议。[13] Jakob Foerster、Richard Y Chen、Maruan Al-Shedivat 等。2018年。运用对手学习意识进行学习。载于第17届自主智能体与多智能体系统国际会议论文集。[14] Ian Gemp、Kevin R. McKee、Richard Everett、Edgar Duéñez Guzmán、Yoram Bachrach、David Balduzzi 和 Andrea Tacchetti。2022年。D3C：降低多智能体学习中的无政府代价。载于第21届自主智能体与多智能体系统国际会议论文集（线上会议，新西兰）（AAMAS '22）。自主智能体与多智能体系统国际基金会，南卡罗来纳州里奇兰，498–506。[15] Bengisu Guresti、Abdullah Vanlioglu 和 Nazim Kemal Ure。2023年。IQ-Flow：通过机制设计诱导序贯社会困境中的自利智能体表现出合作行为。载于2023年国际自主智能体与多智能体系统会议论文集（英国伦敦）（AAMAS '23）。南卡罗来纳州里奇兰，2143–2151。[16] Pablo Hernandez-Leal、Michael Kaisers、Tim Baarslag 和 Enrique Munoz de Cote。2017年。多智能体环境中的学习综述：应对非平稳性。arXiv预印本 arXiv:1707.09183 (2017)。[17] Matteo Hessel、Hubert Soyer、Lasse Espeholt、Wojciech Czarnecki、Simon Schmitt 和 Hado Van Hasselt。2019年。运用 PopArt 进行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。[18] 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年。不平等厌恶提升了跨期社会困境中的合作。载于第32届神经信息处理系统国际会议论文集。3330–3340。[19] Max Jaderberg、Wojciech M Czarnecki、Iain Dunning、Luke Marris、Guy Lever、Antonio Garcia Castaneda、Charles Beattie、Neil C Rabinowitz、Ari S Morcos、Avraham Ruderman 等。2019年。基于种群的强化学习在3D多人游戏中达到人类水平表现。《科学》364, 6443 (2019)。[20] W Bradley Knox 和 Peter Stone。2009年。通过人类强化互动式塑造智能体：TAMER框架。载于第五届知识获取国际会议论文集。9–16。[21] Michael Kölle、Tim Matheis、Philipp Altmann 和 Kyrill Schmid。2023年。通过奖励份额交易学习参与。载于第15届国际智能体与人工智能会议论文集——第1卷：ICAART。INSTICC，SciTePress, 355 362 10 5220 0011781600003393 –。 [22] 2011 https://doi.org/。/ Guillaume J Laurent、Laëtitia Matignon、Le Fort-Piat 等。《独立学习者的世界并非马尔可夫性》。《国际知识工程与智能系统杂志》15, 1 (2011), 55–64。

[23] Joel Z Leibo、Vinicius Zambaldi、Marc Lanctot 等。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（AAMAS '17）论文集。464–473。[24] 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）。[25] Alistair Letcher、Jakob Foerster、David Balduzzi、Tim Rocktäschel 和 Shimon Whiteson。2019。可微博弈中的稳定对手塑造。ICLR（2019）。[26] 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。爱思唯尔，157–163。[27] Ryan Lowe、Yi Wu、Aviv Tamar、Jean Harb、Pieter Abbeel 和 Igor Mordatch。2017。混合合作竞争环境中的多智能体演员-评论家算法。载于《神经信息处理系统进展》。6379–6390。[28] Andrei Lupu 和 Doina Precup。2020。多智能体强化学习中的馈赠。载于第19届自主智能体与多智能体系统国际会议论文集。789–797。[29] Michael W Macy 和 Andreas Flache。2002。社会困境中的学习动态。《美国国家科学院院刊》99（2002），7229–7236。[30] Laëtitia Matignon、Guillaume J Laurent 和 Nadine Le Fort-Piat。2007。滞后Q学习：合作型多智能体团队中的去中心化强化学习算法。载于2007年IEEE/RSJ智能机器人与系统国际会议。IEEE, 64–69。[31] 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（新西兰奥克兰）（AAMAS '20）。自主智能体与多智能体系统国际基金会，南卡罗来纳州里奇兰，869–877。[32] Julien Perolat、Joel Z. Leibo、Vinicius Zambaldi、Charles Beattie、Karl Tuyls 和 Thore Graepel。2017。公共池资源占用的多智能体强化学习模型。载于第31届神经信息处理系统国际会议论文集（美国加利福尼亚州长滩）（NIPS'17）。Curran Associates Inc.，美国纽约州雷德胡克，3646–3655。[33] Julien Perolat、Florian Strub、Bilal Piot 和 Olivier Pietquin。2017。从批量数据中学习一般和马尔可夫博弈的纳什均衡。载于《人工智能与统计学》。232–241。[34] Thomy Phan、Fabian Ritz、Lenz Belzner、Philipp Altmann、Thomas Gabor 和 Claudia Linnhoff-Popien。2021。VAST：采用可变智能体子团队的价值函数分解。载于《神经信息处理系统进展》，M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan（编），第34卷。Curran Associates, Inc.，24018–24032。https://proceedings.neurips.cc/paper_files/paper/2021/file/c97e7a5153badb6576d8939469f58336-Paper.pdf [35] Thomy Phan、Felix Sommer、Philipp Altmann、Fabian Ritz、Lenz Belzner 和 Claudia Linnhoff-Popien。2022。通过相互致谢交换产生的涌现合作。载于第21届自主智能体与多智能体系统国际会议论文集（AAMAS）（线上活动，新西兰）。1047–1055。https://doi.org/doi/abs/10.5555/3535850.3535967 [36] Thomy Phan、Felix Sommer、Fabian Ritz、Philipp Altmann、Jonas Nüßlein、Michael Kölle、Lenz Belzner 和 Claudia Linnhoff-Popien。2024。多智能体强化学习中通过相互致谢交换产生的涌现合作。《自主智能体与多智能体系统》38，34（2024）。https://doi.org/ 10 1007 10458 024 09666 5 。/s - - - David Radke、Kate [37] Larson 和 Tim Brecht。2022。探索多智能体学习中团队的优势。载于第31届人工智能国际联合会议论文集（IJCAI-22），Lud De Raedt（编）。人工智能国际联合会议组织，454–460。https://doi.org/10.24963/ijcai.2022/65 主会场。[38] Anatol Rapoport。1974。囚徒困境——回忆与观察。载于《作为冲突解决理论的博弈论》。Springer，17–34。[39] Kyrill Schmid、Lenz Belzner、Robert Müller 等。2021。随机市场博弈。载于第30届人工智能国际联合会议论文集（IJCAI-21）。人工智能国际联合会议组织，384–390。[40] Shai Shalev-Shwartz、Shaked Shammah 和 Amnon Shashua。2016。自动驾驶安全多智能体强化学习。arXiv 预印本 arXiv:1610.03295（2016）。[41] Richard S Sutton。1988。用时序差分方法进行预测学习。《机器学习》（1988）。[42] Richard S Sutton、David A McAllester、Satinder P Singh 和 Yishay Mansour。2000。采用函数逼近的强化学习策略梯度方法。载于《神经信息处理系统进展》，S. Solla、T. Leen 和 K. Müller（编），第12卷。麻省理工学院出版社，1057–1063。[43] Ming Tan。1993。多智能体强化学习：独立智能体与合作智能体。载于第十届国际机器学习会议论文集。Morgan Kaufmann出版社，

[44] Robert L Trivers。1971。互惠利他主义的演化。《生物学季刊评论》46，1（1971），35–57。[45] Hado P van Hasselt、Arthur Guez、Matteo Hessel、Volodymyr Mnih 和 David Silver。2016。跨越多个数量级学习价值。神经信息处理系统进展29（2016）。[46] Paul AM Van Lange、Jeff Joireman、Craig D Parks 和 Eric Van Dijk。2013。社会困境心理学：综述。《组织行为与人类决策过程》120，2（2013），125–141。[47] Eugene Vinitsky、Raphael Köster、John P Agapiou、Edgar Duéñez-Guzmán、Alexander Sasha Vezhnevets 和 Joel Z Leibo。2023。在去中心化多智能体环境中通过公共制裁习得社会规范的学习型智能体。《集体智能》2，2（2023）。[48] Oriol Vinyals、Igor Babuschkin、Wojciech M Czarnecki、Michaël Mathieu、Andrew Dudzik、Junyoung Chung、David H Choi、Richard Powell、Timo Ewalds、Petko Georgiev 等。2019。使用多智能体强化学习达到《星际争霸II》特级大师水平。《自然》（2019），1–5。[49] Ermo Wei 和 Sean Luke。2016。独立学习者随机合作博弈中的宽容学习。《机器学习研究杂志》17，1（2016），2914–2955。[50] Timon Willi、Alistair Hp Letcher、Johannes Treutlein 和 Jakob Foerster。2022。COLA：具备对手学习意识的一致学习。载于国际机器学习会议。PMLR，23804–23831。[51] Ronald J Williams。1992。用于联结主义强化学习的简单统计梯度跟随算法。《机器学习》8，3（1992），229–256。[52] Jiachen Yang、Ang Li、Mehrdad Farajtabar、Peter Sunehag、Edward Hughes 和 Hongyuan Zha。2020。激励其他学习智能体。神经信息处理系统进展33（2020）。[53] Jiachen Yang、Ethan Wang、Rakshit Trivedi、Tuo Zhao 和 Hongyuan Zha。2022。基于多智能体元梯度强化学习的自适应激励设计。载于第*21*届自主智能体与多智能体系统国际会议论文集。1436–1445。[54] Y Yang、R Luo、M Li、M Zhou、W Zhang 和 J Wang。2018。均值场多智能体强化学习。载于第*35*届国际机器学习会议，*ICML 2018*，第80卷。PMLR，5571–5580。[55] Yuxuan Yi、Ge Li、Yaowei Wang 和 Zongqing Lu。2022。网络化多智能体强化学习中的共享学习。载于神经信息处理系统进展，S. Koyejo、S. Mohamed、A. Agarwal、D. Belgrave、K. Cho 和 A. Oh（编），第35卷。Curran Associates, Inc.，15119–15131。[56] Xuezhou Zhang、Yuzhe Ma、Adish Singla 和 Xiaojin Zhu。2020。针对强化学习的自适应奖励投毒攻击。载于国际机器学习会议。PMLR，11225–11234。[57] Stephen Zhao、Chris Lu、Roger B Grosse 和 Jakob Foerster。2022。具有对手学习意识的近端学习。神经信息处理系统进展35（2022），26324–26336。

A 扩展理论结果

本节提供更多经过完整推导的示例和详细论证，以补充正文中的理论分析。这些结果更加具体且直观地说明了 DRIVE 在简单设置下的运行方式、其行为如何在不同协议遵循假设下发生变化，以及微观的请求/响应机制如何支撑宏观均衡结果，并为更大的多智能体设置提供一种简单的扩展。

预备知识。我们回顾标准的双人囚徒困境（PD），其动作是 C（合作）和 D（背叛），当策略组合分别为 (C,C)、(D,D)、(D,C) 和 (C,D) 时，收益依次为 (R,R) 、 (P,P) 、 (T,S) 和 (S,T) （图 2 ）。若 $T > R > P > S$ （参见公式 1 ），则该博弈为囚徒困境，其中 $T > R$表示贪婪， $P > S$表示恐惧 [3]。我们将智能体 i在时间 t的环境奖励记为 $u_{t,i}$，将时期 m内的外部修正奖励记为 $\hat{}u_{t,i} = f_{mod}(u_{t,i}, m)$，将时期 m内 $\hat{}u_{t,i}$的运行平均值记为 $\bar{u}_i$。智能体 i在时间 t 的 DRIVE 塑形规则为

$$u_{t,i}^{\mathrm{DRIVE}} = \hat{u}_{t,i} - \min_{j \in N_{t,i}} \{\Delta_{t,j,i}\} + \min_{j \in N_{t,i}} \{\Delta_{t,i,j}\}, \tag{5}$$

其中 $\Delta_{t,\,i,\,j} = \bar{u}_j - \hat{}u_{t,\,i}$是 i发出请求时从 j发送给 i的响应，并且公式 (3) 中的时序差分门控 $\mathrm{TD}_i(\hat{}u_{t,\,i})$ 非负。在双智能体囚徒困境中，我们有 $N_{t,\,1} = \{2\}$ 和 $N_{t,\,2} = \{1\}$，因此最小值可简化为单项。

A.1 双人囚徒困境中激励对齐的基础：单次交换机制

我们给出一个简单的例子来说明 DRIVE 的塑形规则如何促进单边背叛时产生收益互换 $T \leftrightarrow S$。这会将囚徒困境转变为协调博弈，其中（C， C）是唯一的纳什均衡。除非另有说明，我们均假设完整协议合规。

设定。考虑双智能体囚徒困境中的单个时间步 t，其收益为 $T > R > P > S$，邻域为$N_{t,\,1} = \{2\}$、$N_{t,\,2} = \{1\}$。奖励在时期 m中可能已由 f_{mod} 转换；为方便记号，我们写作

$$\hat{T} := f_{mod}(T, m), \quad \hat{R} := f_{mod}(R, m), \quad \hat{P} := f_{mod}(P, m), \quad \hat{S} := f_{mod}(S, m).$$

我们假设，在平稳博弈的长期过程中，各策略组合对应的时期平均值与相应的即时收益一致（例如，在（C， C）下，两个智能体的收益均为 $\bar{u}_i = \hat{}R$）。

$\hat{}$相互合作与相互背叛。 如果两个智能体在时间 t 都选择合作，则各自获得 $\hat{}u_{t,\,1} = \hat{}u_{t,\,2} = R$ 并且在稳态下，$\bar{u}_1 = \bar{u}_2 = \hat{}R$。因此 $\Delta_{t,\,1,\,2} = \bar{u}_2 - \hat{}u_{t,\,1} = 0$ 且 $\Delta_{t,\,2,\,1} = 0$，所以公式 (5) 得出

$$u_{t,1}^{\mathrm{DRIVE}} = \hat{R}, \quad u_{t,2}^{\mathrm{DRIVE}} = \hat{R}.$$

类似的论证表明，u 在相互背叛时，两个智能体都保有收益 $\hat{}P$。因此，DRIVE 在长期内不会改变 R 和 P n.

单边背叛。 现在考虑这样一种状态：智能体 1 背叛（D），而智能体 2 合作（C），因此瞬时修正奖励为 $\hat{}u_{t,\,1} = \hat{}T, \hat{}u_{t,\,2} = \hat{}S$。按照通常的囚徒困境排序，背叛者获得的奖励高于合作者。我们假设这会使背叛者的时序差分优势非负，而合作者的时序差分优势非正，因此

$$\mathrm{TD}_1(\hat{u}_{t,1}) \geq 0 \;\Rightarrow\; \text{agent 1 sends a request}, \qquad \mathrm{TD}_2(\hat{u}_{t,2}) < 0 \;\Rightarrow\; \text{agent 2 does not send.} \tag{6}$$

这符合 DRIVE 预期的运行机制，也与我们在实验中观察到的实证行为一致。智能体 1 将其瞬时奖励 $\hat{}u_{t,\,1} = \hat{}T$作为请求发送给智能体 2。收到请求后，智能体 2 计算轮次-m差值 $\Delta_{t,\,1,\,2} = u_2 - \hat{}u_{t,\,1}$。对于所考虑的囚徒困境步骤，最坏情形下的一致选择是 $u_2 = \hat{}S$，得到 $\Delta_{t,\,1,\,2} = \hat{S} - \hat{}T < 0$。智能体 2 将此值返回给智能体 1。由于只有智能体 1 发送了请求，本步骤中的反向差值 $\Delta_{t,\,2,\,1}$未定义；根据算法 2，在没有收到响应时，该值在聚合时默认为 0。使用响应方的轮次平均值 u_j而非其瞬时奖励，使 DRIVE 能够对系统性利用行为而非单次噪声结果作出反应。只有近期平均回报持续低于邻居请求值的智能体，才会产生较强的负向 Δ 项。如果双方都使用瞬时奖励，单次走运或倒霉的步骤就可能触发大额转移，使激励受到噪声干扰，也容易被短期冒险所操纵。若用瞬时奖励与平均值比较，则会使 DRIVE 变成一种风险共担机制；而双方都比较平均值虽然会非常稳定，却反应迟缓。因此，所选的平均值—瞬时值设计兼顾了抗噪鲁棒性与对新出现的利用行为的响应能力。在双人情形下，对邻居取最小值可简化为单项。逐步塑形奖励为

$$\hat{u}_{t,1}^{\mathrm{DRIVE}} = \hat{u}_{t,1} - \underbrace{\min\{\Delta_{t,2,1}\}}_{=0} + \underbrace{\min\{\Delta_{t,1,2}\}}_{=\hat{S}-\hat{T}} = \hat{T} + (\hat{S} - \hat{T}) = \hat{S}, \tag{7}$$

$$\hat{u}_{t,2}^{\mathrm{DRIVE}} = \hat{u}_{t,2} - \underbrace{\min\{\Delta_{t,1,2}\}}_{=\hat{S}-\hat{T}} + \underbrace{\min\{\Delta_{t,2,1}\}}_{=0} = \hat{S} - (\hat{S} - \hat{T}) = \hat{T}. \tag{8}$$

因此，单边背叛步骤（$\hat{}T$， $\hat{}S$）被重新塑造为（$\hat{}S$， $\hat{}T$）：背叛者获得受骗收益，而合作者获得诱惑收益。式（7–8）明确展示了时序差分门控与 DRIVE 塑形规则如何在单个交换步骤中实现 $T \leftrightarrow S$交换。根据对称性，策略组合（C， D）由（$\hat{}S$， $\hat{}T$）重新塑造为（$\hat{}T$， $\hat{}S$）。

由此产生的激励与均衡解释。汇总所有行动组合，DRIVE 生成的修改后的收益矩阵为

	C	D
C	$\hat{R}, \hat{R}$	$\hat{S}, \hat{T}$
D	$\hat{T}, \hat{S}$	$\hat{P}, \hat{P}$

$\hat{T} > \hat{R} > \hat{P} > \hat{S}$。因此，如果对手采取 C，以 C 作为响应将得到 $\hat{R} > \hat{S}$；如果对手采取 D，以 C 作为响应将得到 $\hat{T} > \hat{P}$。因此，合作是对两种动作的严格最佳响应，所以 (C, C) 是唯一的纳什均衡。

要点。 在时序差分门控（公式 3）和塑形规则（公式 4）下，单边背叛状态中的一次请求/响应交换会确定性地将 $(\hat{T}, \hat{S}) \mapsto (\hat{S}, \hat{T})$ 映射，从而在该步中消除变换后的囚徒困境里的贪婪和恐惧激励。这一微观机制直接支撑了均衡层面的结论：DRIVE 将囚徒困境转变为协调博弈，其中合作是唯一的纳什均衡；同时，它也为定理 1 的证明提供了一个经过完整推演的补充。

A.2 协议遵从案例分析

为补充第 3 节中的讨论，我们在 DRIVE 下、具有不同协议遵循程度的双人囚徒困境实例。我们考虑公式 (1) 中的一般囚徒困境收益 $T > R > P > S$；具体而言，可以采用典型取值 $(T, R, P, S) = (5, 3, 1, 0)$，但以下论证并不依赖这些特定数值。在每种情况下，智能体 i 背叛，而智能体 j 合作；DRIVE 根据请求/响应行为进行奖励塑形。 **（1）完全遵循**。如果所有智能体都如实遵循

Case	背叛者收益	合作者收益
（1）完全遵循	S	T
（2）背叛者不发送请求	T （无惩罚）	S
（3）合作者不予响应	T （无惩罚）	S
（4）虚报（虚假请求/响应）	不稳定；可能导致惩罚失效	不稳定；可能导致错误惩罚
(5) 完全不遵循协议	T	S

表 1：2 人囚徒困境中不同协议遵循情形下收益的示例 **D.**

遵循协议时，定理 1 的结论适用。单边背叛（T， S）会被重新塑造成（S， T）：背叛者获得 S，合作者获得 T，从而消除了背叛的激励。

(2) 背叛者不发送请求。如果背叛者尽管 TD 为正却不发送请求，就能避免受到邻居的惩罚，但也会失去实现双方收益提升的机会。其收益仍为 T，而合作者的收益仍为 S。这种策略会间接惩罚不发送请求的智能体，使其无法从互惠塑造中获益。

(3) 合作者隐瞒响应。如果合作者未作响应（无论是故意为之还是由于信息丢失），背叛者就能逃避预期的惩罚。收益恢复为（T， S）。由于邻居之间采用最小值聚合，只有当缺失的响应对应唯一或最强的惩罚时，这种响应缺失才会产生影响。

(4) 虚报。虚报可能以两种形式出现：(i) 智能体在没有改进信号时仍发送请求（TD< 0）；或 (ii) 智能体响应时提供不正确的 Δ 值（由于错误或故意为之）。在第一种情况下，由此产生的差异通常较小或为正，最小值算子会将其过滤掉，因此影响可以忽略不计。然而，在第二种情况下，虚假响应可能会扭曲聚合结果：虚报的邻居可能会减轻或抵消对背叛者的惩罚，或不公正地惩罚遵守规则的请求者。最坏情况下，这可能会破坏合作的稳定性，但其风险并不高于其他 PI 方法所面临的风险。总体而言，虚报是鲁棒性方面的一项局限，但只要大多数响应仍然真实，就不会从根本上破坏该机制。

(5) 完全不遵守。如果没有交换请求或响应，所有塑造项都会消失。在单边背叛的情况下，DRIVE 将退化为采用收益（T， S）的纯多智能体强化学习，合作激励也将不复存在。

向更大规模场景的扩展。 上文的双人示例展示了一个极端情况：任何单个不遵从的智能体都意味着协议遵循率为 50%，从而显著放大了个体偏离的影响。然而，在更大的群体中，部分不遵从的影响会更加渐进：只要仍有足够比例的智能体继续如实响应，最小值聚合就仍能惩罚背叛者。

三人示例。 考虑由三个智能体组成的群体，其中一个背叛（i），两个合作（j， k）。每个合作者都如实响应 $\Delta_{i,\,j}$， $\Delta_{i,\,k} < 0$。背叛者的塑形奖励为

$$u_i^{\mathrm{DRIVE}} = T + \min\{\Delta_{i,j}, \Delta_{i,k}\},$$

因此，只要至少有一位合作者如实响应， u_i^{DRIVE}就会受到惩罚。如果一位合作者隐瞒，另一位仍会执行惩罚，只是力度可能较弱。因此，与双人情形不同，即使单个智能体不遵从，合作激励也不会完全消失。

A.3 扩展到简单的 N 智能体囚徒困境设定

主定理聚焦于双人囚徒困境（PD）。这里我们首先将其清晰地扩展到一个由 N个智能体组成的图形化 PD，并假设所有智能体均完整遵循协议；在此情形下，收益可分解为对称的双人囚徒困境交互。随后，我们逐步放宽结构和行为假设。

定义 *6*（图结构囚徒困境）。设 $G=(V,\ E)$ 是一个包含 $|V|=N$ 个智能体的无向图。每个智能体 $i\in V$ 选择一个动作 $a_i\in\{C,\ D\}$。对于每条边 $(i,\ j)\in E$，我们将其关联到一个双人囚徒困境，其收益 $(R,\ R)$、$(P,\ P)$、$(T,\ S)$、$(S,\ T)$ 如公式 (1) 所示。智能体 i 的阶段收益为

$$u_i(a)=\sum_{j\in N(i)} u_{i,j}(a_i,a_j),$$

其中，$N(i)$ 是 G 中 i 的邻域，$u_{i,\ j}$ 是对邻居 j 的囚徒困境收益。

为单独考察激励对齐的影响，我们假设每对智能体之间的交互都使用双智能体 DRIVE 协议独立塑造，并且所有智能体都完全遵守请求与响应规则。也就是说，在与邻居 j 的每次交互中，智能体 i 的行为都与附录 A.1 中双智能体分析的情形完全相同。

命题 7（成对激励对齐）。考虑上述图结构囚徒困境，并假设每条边 $(i,\ j)$ 都独立地对 $u_{i,\ j}(a_i,\ a_j)$ 使用双智能体 *DRIVE* 协议及塑形规则 (5)。那么，对于每个智能体 i 及其邻居的每一种联合行动组合 a_{-i}，

$$u_i^{\mathrm{DRIVE}}(C,a_{-i})>u_i^{\mathrm{DRIVE}}(D,a_{-i}),$$

因此，合作是每个智能体的占优行动，且塑造后博弈的唯一纳什均衡是全体合作组合（$C,\ \ldots\ldots,\ C$）。

证明。根据定理 1 及附录 A.1 中的详细证明，对于每个邻居 j，成对 DRIVE 塑造的收益函数 $u_{i}^{\mathrm{DRIVE}}{}_{,\ j}$使 C在相应的双人囚徒困境中严格占优于 D。也就是说，对于任意固定的 $a_j\in\{C,\ D\}$，$u_{i}^{\mathrm{DRIVE}}{}_{,\ j}(C,\ a_j)>u_{i}^{\mathrm{DRIVE}}{}_{,\ j}(D,\ a_j)$。对所有邻居 $j\in N$ 求和后，对于任意联合邻居动作组合 i，可得

$$u_i^{\mathrm{DRIVE}}(C,a_{-i})=\sum_{j\in N(i)} u_{i,j}^{\mathrm{DRIVE}}(C,a_j)>\sum_{j\in N(i)} u_{i,j}^{\mathrm{DRIVE}}(D,a_j)=u_i^{\mathrm{DRIVE}}(D,a_{-i}).$$

He因此，对于所有智能体 i，C都是对任意 a_{-i}的严格最佳响应，所以（$C,\ \ldots\ldots,\ C$）是唯一的纳什均衡。 □

该命题形式化说明了双人激励对齐结果如何扩展到一类简单的 N 智能体博弈，其中包含加性双人囚徒困境交互。我们的 SSD 实验采用了更复杂的交互模式，但图结构囚徒困境结果清晰地从理论上展示了 DRIVE 如何扩展到两个以上的智能体。

跨邻域的成对囚徒困境交互。在许多多智能体系统中，智能体会与动态变化或随机匹配的伙伴反复参与成对社会困境（即交互图），而激励交换则受限于一个可能不同且固定的通信图，该图由邻域 $N_{t、\ i}$定义。在此类设置中，背叛者的 DRIVE 塑形奖励 i仍采取 $u_i^{\mathrm{DRIVE}}=T+\min_{j\in N_{t、\ i}}\Delta_{i、\ j}$ 的形式。当某个囚徒困境回合中被剥削的合作者同时也是 i 的邻居时，便可恢复二智能体情形下的精确 $T\leftrightarrow S$收益互换。当交互伙伴位于 $N_{t、\ i}$之外时，惩罚则由那些因 i 的行为而持续遭受长期劣势的遵从协议的邻居承担（例如通过共享资源或未来交互）。因此，交互图决定哪些行为会产生负外部性，而通信图决定谁能够追究谁的责任。只要每个背叛者至少有一名遵从协议的邻居，并且该背叛者背叛时会使这名邻居的长期回报降低，那么激励的方向仍然是促进合作的，即使其幅度不同于完全局部化的二智能体囚徒困境。

超越图结构囚徒困境。许多基准社会困境环境（如 Coin Game 或 Harvest）无法明确分解为成对囚徒困境收益。相反，激励是通过共享资源、延迟后果和整体系统动态间接产生的。尽管如此，同一原则仍然适用：行为持续降低邻居长期回报的智能体会产生负差异 $\Delta_{i,\ j}$并受到 DRIVE 的惩罚。这一观点解释了为何 DRIVE 在此类环境中经验证实能够促进合作，尽管其中不存在明确的囚徒困境结构。

一般 N智能体的协议遵循与鲁棒性。前述结果假设完整协议合规。现在，我们考虑一个一般的 N智能体设置，其中只有 M个智能体如实遵循 DRIVE 协议，而 $K=N-M$个智能体可能通过不发送请求或响应，或通过虚报，采取对抗性行为。对于任何博弈论意义上的背叛者 i，有效惩罚需要满足两个条件：(i) i自身必须遵守请求门控（在 TD ≥ 0时发送请求）；(ii) 至少有一个邻居 $j\in N_i$如实响应。在这些条件下，i的塑形奖励为

$$u_i^{\mathrm{DRIVE}}=T+\min\{\Delta_{i,j}\}_{j\in N_i,\ j\ \text{compliant}}.$$

只要存在合规的响应者，背叛就会受到惩罚；惩罚程度取决于响应者中最负的差值，但惩罚不会完全消失。在全连接群体中，这简化为一个简单要求：至少有两个智能体合规（请求者和一名响应者），即 $M/N\geq 2/N$。在一般通信拓扑中，合规集合必须构成一个支配集，确保每个请求者至少有一名合规邻居。因此，DRIVE 表现出优雅降级：随着合规率 M/N下降，合作激励会平稳减弱，但只有在没有可用响应者时才会完全消失。

连通性要求。 在全连接图中，只要至少有两个智能体合规（请求者和一名响应者），就能确保惩罚，对应的合规率为 $c \geq 2/N$。在一般通信拓扑中，合规智能体必须构成一个支配集：每个请求者都必须至少有一名合规邻居。形式上，对于图 $G = (V, E)$ 和合规集合 $C \subseteq V$，当且仅当 $\forall i \in V$: $N(i) \cap C \neq \varnothing$ 时，惩罚才成立。最小合规集合的大小即支配数 $\gamma(G)$，由此得到拓扑感知阈值 $|C| \geq \gamma(G)$。

对抗鲁棒性。 上述阈值可以直接从 K 对抗智能体的角度来解释。如果有 K 个智能体是对抗性的（从不提供真实响应），只要其余 $M = N - K$ 个合规智能体中包含一个支配集，DRIVE就会继续惩罚每个遵循协议的背叛者。等价地说，系统最多可以容忍 $K \leq N - \gamma(G)$ 个对抗智能体。然而，在动作（总是背叛）和协议行为（从不发送请求）两方面都采取对抗行为的智能体，无法被任何基于自报告信号的PI方案惩罚，包括DRIVE。对于遵从协议的邻居而言，这些智能体看起来只是在背叛时不会产生负差异的玩家；这会减轻局部合作压力，并逐渐使系统转向底层的非PI多智能体强化学习动态（案例5），但不会导致系统骤然崩溃。

总体而言，这些结果凸显了DRIVE的优雅降级特性：完全遵循协议可恢复强有力的激励对齐保证，而部分遵从会削弱合作激励，但不会立即将其摧毁，前提是通信拓扑满足最低覆盖条件。在双智能体系统中，单个不遵从协议的智能体就会使合规率降至50%，从而消除惩罚机制。相比之下，在规模更大的群体中（$N > 2$），即使有若干智能体背离协议，其余多数智能体仍可通过最小值聚合实施惩罚。因此，只要合规率高于特定拓扑对应的阈值，DRIVE就仍然有效；在实践中，该阈值往往接近简单多数。不过，系统性虚报或普遍不遵从会使DRIVE退化为基线多智能体强化学习，这凸显了未来工作中设计能够抵御响应缺失或对抗性响应的变体的重要性。

A.4 仿射奖励变换：范围与边缘情况

此处，我们进一步说明分析中使用的仿射奖励变化模型。在每个时期 m，环境奖励都会经过一个所有智能体和时间步统一应用的共享正仿射映射 $r' = c_m r + b_m$ 转换。调度 c_m 和 b_m 可以在各时期任意变化，涵盖第7.2节中使用的所有奖励动态（线性、指数、分段和阻尼余弦）。

示例：阻尼余弦调度。举例来说，在阻尼余弦情形下，我们使用 $c_m = (1 - m/E)\cos^2(2\eta m)$ 和 $b_m = \eta$，因此有效奖励尺度会随时间振荡并逐渐衰减。

退化时期，其中 $c_m \approx 0$。个别时期可能满足 $c_m \approx 0$，从而产生 $r' \approx b_m$，并暂时使囚徒困境的不等式变得无法区分（T、 R、 P、 S变得无法区分）。这些时期的阶段博弈几乎对收益不敏感。我们的结果通过仅对满足 $c_m > 0$ 的时期应用囚徒困境分析，明确处理了此类情况。

未涵盖的变换。 这一仿射类别不包括真正的非线性奖励变化，例如截断、饱和、依赖符号的重新映射，或无法表示为共享仿射映射的依赖状态的变换。这些代表了更广泛的仿真到现实迁移，我们将留待未来工作研究。

A.5 囚徒困境中的奖励变化与不变性

我们首先形式化仿射奖励变换如何影响囚徒困境不等式。

引理 8（仿射变换保持囚徒困境结构）。 令 $T > R > P > S$并定义 $T\prime = cT + b, R' = cR + b, P' = cP + b,$ $S' = cS + b$ $c > 0$ $b \in \mathbb{R}$ $T' > R' > P' > S'$ ，其中某些，。于是。

证明。由于 $c > 0$，我们有 $T\prime - R' = c(T - R) > 0, R' - P' = c(R - P) > 0$，且 $P' - S' = c(P - S) > 0$，因此 $T\prime > R' > P' > S'$。加法常数 b在所有成对差值中都会抵消。□

现在，我们可以更正式地重述定理 3 和定理 4。

定理 3 的证明。在 MATE 中，当且仅当全局代币 $x > 0$ satisfies $x \geq \max\{P-S, (T-R)/3\}$ [36] 时，相互合作才是个体理性的。在定义 2 的仿射变换下，这些阈值变为 $P' - S'$ and $(T' - R')/3$，即

$$x \geq \max\{P' - S', (T' - R')/3\}.$$

如果固定任意满足此条件的 x，由于 c_m 和 b_m 可以任意取值（但须满足 $c_m > 0$），我们总能选择 (c_m, b_m)，使得 $P' - S' > x$，例如，取足够大的 c_m 即可。这样一来，变换后的阈值就超过了 x，而 MATE 收益矩阵不再保证合作是最佳响应。因此，MATE 对此类奖励变化并非不变。□

定理 4 $R\ P\ T\ S$ $R'\ P'\ T\ \ S'$ f_{mod} 2 8 $T >$ $R > P > S$ $T\ > R' > P' > S'$ 1 的证明。设 $(\_,\_)$ 和 $(\_,\prime)$ 是由定义 中任意一种方式关联的两组囚徒困境收益。根据引理，可知，。附录 A 中的详细分析表明，在完全遵循的情况下，DRIVE 保持相互合作和相互背叛的收益不变，并在单边背叛时交换诱惑收益与受骗收益。对 $(R', P', T\prime, S')$ 进行相同计算可知，变换后的收益遵循相同模式。在这两种情况下，所得博弈均以 (C, C) 为唯一纳什均衡，因此合作的占优性得以保留。□

A.6 使用归一化回报进行训练

在我们的论文中，$\hat{}\pi_i$ 和 $\hat{}V_i$ 使用归一化回报$G_{t,i} = \frac{1}{\sigma_i}(G_{t,i} - \mu_i)$ 进行训练，其中 μ_i 和 σ_i 分别是智能体 i 在一个时期 m内所有回报$G_{t,i}$的均值和标准差，以提高训练稳定性 [45]。根据引理9，这种归一化使标准策略梯度强化学习（不使用 PI机制）不受奖励变化的影响；奖励变化包括将奖励乘以因子 $c_m > 0$ 或加上标量 $b_m \in \mathbb{R}$。

引理9. 根据公式 *2*，标准策略梯度强化学习（不使用 *PI* 机制）对奖励变化具有不变性：当每个时期 m 中获得的回报$G_{t,i}$被归一化为均值为零、标准差为$G_{t,i} = \frac{1}{\sigma_i}(G_{t,i} - \mu_i)$时，将原始环境奖励乘以因子 $c_m > 0$ 或加上标量 $b_m \in \mathbb{R}$，不会改变结果。

证明。由于 $c_m > 0$，归一化回报$G_{t,i}$不受缩放和平移的影响，无论 $B_m = \sum_{t=0}^{H-1} \gamma^t b_m \in \mathbb{R}$为何：

$$\overline{\sigma}_i^2 = \frac{1}{H-1}\sum_{t=0}^{H-1}(c_m(G_{t,i} + B_m) - c_m(\mu_i + B_m))^2 = \frac{c_m^2}{H-1}\sum_{t=0}^{H-1}(G_{t,i} - \mu_i)^2 = (c_m\sigma_i)^2 \tag{9}$$

$$\frac{1}{c_m\sigma_i}(c_m(G_{t,i} + B_m) - c_m(\mu_i + B_m)) = \frac{1}{\sigma_i}(G_{t,i} - \mu_i) = \overline{G}_{t,i} \tag{10}$$

因此，通过缩放和平移对$G_{t,i}$进行的任何奖励变化，都不会影响归一化后的$G_{t,i}$。 □

根据引理9，非 PI 基线方法朴素学习、*LOLA-PG*和*POLA-DiCE* 不受第7.2节和图6中动态奖励变化的影响（但其合作程度仍低于 *DRIVE*）。然而，大多数最先进的 PI 方法，例如*LIO* 和*MATE* ，无法在其收益修改中适应此类变化（通过通常未知的函数 f_{mod}）（图 4），因此无法调整；而*DRIVE*则不然，如图 1 和图 6 所示。

A.7 归一化策略梯度下 SSD 的不变性

利用定理9中定义的归一化回报，我们可以更详细地重述并证明定理5

$$\bar{G}_{t,i} = \frac{G_{t,i} - \mu_i}{\sigma_i},$$

其中，μ_i和 σ_i 分别是智能体 i在时期 m的所有回报 $G_{t,\ i}$的均值和标准差。

引理10（归一化策略梯度的缩放不变性）。设时期 m中的奖励经过 $r'_{t,\ i} = c_m r_{t,\ i} + b_m$变换，其中 $c_m > 0$；并设 $G'_{t,\ i}$、 μ'_i和 σ'_i分别为相应的回报、均值和标准差。则归一化回报相同，即 $\bar{G}'_{t,\ i} = \bar{G}_{t,\ i}$，因此公式 (2) 中的策略梯度估计保持不变。

证明。如附录 A.6所示，变换后的回报满足 $G'_{t,\ i} = c_m G_{t,\ i} + B_m$，其中 $B_m = \sum_{k=0}^{H-1}\gamma^k b_m$与 t无关。因此，$\mu'_i = c_m\mu_i + B_m$和 $\sigma'_i = c_m\sigma_i$。归一化回报变为

$$\bar{G}'_{t,i} = \frac{G'_{t,i} - \mu'_i}{\sigma'_i} = \frac{c_m G_{t,i} + B_m - (c_m\mu_i + B_m)}{c_m\sigma_i} = \frac{G_{t,i} - \mu_i}{\sigma_i} = \bar{G}_{t,i}.$$

将 $\bar{G}'_{t,i}$ 代入策略梯度估计器公式 (2)，得到的梯度与 $\bar{G}_{t,i}$ 的梯度相同。 □

现在我们将这一点与 DRIVE 的结构结合起来。

定理5的证明。令时期 m 的环境奖励经由 $f_{mod}(r_{t,i}, m) = c_m r_{t,i} + b_m$ 转换，其中 $c_m > 0$。对于任意智能体 i，公式 (5) 中的 DRIVE 塑形规则仅取决于 $\hat{}u_{t,i}$ 以及形如 $\Delta_{t,i,j} = \bar{u}_j - \hat{}u_{t,i}$ 的差值。在 f_{mod} 下，这些量变为

$$\Delta'_{t,i,j} = \bar{u}'_j - \hat{u}'_{t,i} = (c_m\bar{u}_j + b_m) - (c_m\hat{u}_{t,i} + b_m) = c_m(\bar{u}_j - \hat{u}_{t,i}) = c_m\Delta_{t,i,j}.$$

因此，公式 (5) 中的最小值按相同的因子 c_m 缩放，塑形奖励则变换为

$$u_{t,i}^{\mathrm{DRIVE}\prime} = c_m u_{t,i}^{\mathrm{DRIVE}} + b_m,$$

即，$u_{t\,i}^{\mathrm{DRIVE}}$，本身也受到相同的逐时期正仿射映射。将引理9应用于$u_{t\,i}^{\mathrm{DRIVE}}$，而非$u_{t,\ i}$可知，归一化回报（因此策略更新）保持不变。因此，基准策略梯度动态以及 DRIVE 所引发的激励效应，均不受逐时期正仿射变换的影响，正如所述。□

B 技术细节与超参数

B.1 超参数

论文第7节报告的实验中所有多智能体强化学习方法共同使用的超参数均列于表2。最终取值是通过粗略网格搜索选定的，该搜索在性能与计算开销之间进行权衡，涉及*DRIVE*、*LIO*、*IA*和*MATE*在*IPD*和*Coin-2*中不动态改变奖励的情况，即 $f_{mod}(u,\ m) = u$。对于所有其他领域，我们直接采用了表2中的最终取值。所有训练和测试运行均在一组包含15台86_64 GNU/Linux（Ubuntu 18.04.5 LTS）机器的计算集群上并行执行；每台机器配备 i7-8700 @ 3.2GHz 中央处理器（8核）和64 GB内存。我们的实验未使用任何图形处理器。

表2：本文评估的所有算法使用的常见超参数及其最终值。我们还列出了本文开发过程中尝试过的数值。

超参数	最终值	数值/范围	描述
K	10	{1, 5, 10, 20}	每个时期的回合数。
E	4000	{2000, 4000, 8000}	时期数。E 逐步增加，以评估学习进度的稳定性直至收敛。
α	0.001	{0。001}	学习率。我们使用了 torch 中 ADAM 的默认值未进行进一步调优。
梯度裁剪范数	1	{1,∞}	梯度裁剪参数。将梯度裁剪范数设为 1 可获得更好的性能优于使用 ∞将其禁用。
λ	1	{0, 1}	用于 TD(λ) 评论家学习的迹参数。
γ	0.95 (*IPD*,*Coin-2*, *Coin-4*) 0.99 (*Harvest-12*)	{0.9, 0.95, 0.99}	回报的折扣因子$G_{t,i}$。任意值 ≥ 0.95 都足够。
$\lvert\tau_{t,i}\rvert$	1	{1, 5, 10}	局部历史长度。为减少计算量，将其设为 1，因为其他取值并未显著提升性能。

对于 *LIO*，我们将激励函数学习的成本权重设为 0.001，并将最大激励值 R_{max} 设为每个领域中绝对值最大的惩罚（*IPD* 中为 3，*Coin-2* 和 *Coin-4* 中为 2，*Harvest-12* 中为 0.25），如 [52] 所建议。对于 *IA*，我们按照 [18] 的建议，设置 $\alpha = 5$ 和 $\beta = 0.05$。

B.2 神经网络架构

我们通过调整 $\hat{}\pi_i$ 和 $\hat{}\ V_i$每个隐藏层的单元数（{32, 64, 128}），粗略地对论文中的神经网络架构进行调优，以兼顾性能和计算开销。隐藏层数量在 1、2 和 3 之间变化，但显著加深网络架构会导致性能下降（可能是由于梯度消失）。对于任何神经网络，使用 ELU 或 ReLU 激活函数都没有显著差异。因此，我们在整个实验中都采用 ELU。*DRIVE*、*MATE*和朴素学习 仅将 $\hat{}\pi_i$ 和 $\hat{}\ V_i$ 用作神经网络。*LIO* 的激励函数网络采用与 $\hat{}\pi_i$ 和 $\hat{}\ V_i$相同的隐藏层架构。此外，在输入激励函数之前，会将 $n-1$其他智能体的联合动作与展平后的观测拼接起来；激励函数的输出是一个 $n-1$ 维向量。随后，该输出向量经过 Sigmoid函数处理，并乘以 R_{max}（B.1 节）。

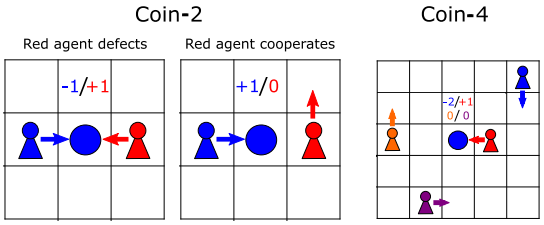
B.3 Coin-2 和 Coin-4

我们采用了 [13] *Coin-2* 中的设置，如图 8a 所示，使用相同的规则和奖励函数。执行动作的顺序是随机的，因此，当两个智能体同时踩到蓝色硬币时，红色智能体的期望收益为 +1，蓝色智能体的期望收益为 -1（图 8a 左）；对于红色硬币，情况则相反。此外，我们将 *Coin-4* 中的环境扩展到 4 个智能体（图 8a 右）。所有智能体都可以自由移动，且多个智能体可以占据同一网格单元位置。任何试图移出边界的移动都会被视为“什么都不做”动作。

囚徒困境联系 假设智能体在 *Coin-2* 中每个时间步都拾取一枚硬币，则根据第2节和图 2，每个时间步的期望奖励构成一个囚徒困境实例：

- 如果两个智能体都合作，即只收集各自匹配的硬币，则每个智能体平均获得 $R = 0.5(1+0) = 0.5$。
- 如果两个智能体都背叛，即无论硬币颜色如何都收集，则每个智能体平均获得 $P = 0.25(1+1-2+0) = 0$。
- 如果一个智能体背叛，则背叛的智能体平均获得 $T = 0.5(1+0) + 1 = 1.5$（100% 的时间收集自己的硬币，50% 的时间收集另一个智能体的硬币）。被剥削代理获得 $S = 0.5(1-2) = -0.5$

预期收益 R、P、T 和 S 满足关于贪婪和恐惧的典型囚徒困境不等式，即 $T > R > P > S$。因此，即使奖励不断变化，DRIVE 在 *Coin-2*和 *Coin-4*（将社会困境扩展到了 4 个智能体）中也同样表现良好，如图7所示。



(a) 硬币域

	C	D
C	0.5, 0.5	−0.5, 1.5
D	1.5, −0.5	0, 0

(b) *Coin-2* 的收益表（每个时间步）

图 8：(a) 论文中使用的*Coin-2* 和*Coin-4*。(b) *Coin-2* 每个时间步的期望奖励收益表。

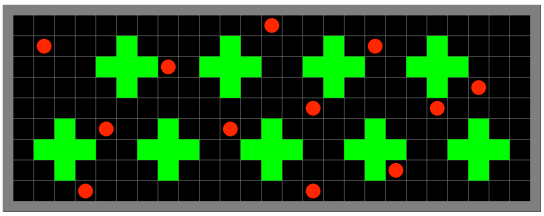
B.4 Harvest-12

我们采用了 [32] 中的设置，如图9a所示，动态和苹果再生速率均保持不变。*Harvest-12* 始终以图9a所示的初始苹果配置开始，智能体的位置则随机确定。智能体能够观察其 7×7 区域周围的环境，且没有特定朝向。因此，它们有 4 个独立动作，可为所有邻居智能体标记方位，这些邻居分别位于其北方、南方、西方或东方。所有智能体都可以自由移动，多个智能体也可以占据同一网格单元位置。任何试图移出边界的移动都会被视为“什么也不做”动作。

囚徒困境关联 我们通过对 Harvest 的双智能体实例进行实证测试来确定收益表，与 [23] 类似。该表见图 9b，根据第2节和图 2，它表示一个囚徒困境实例：

- 如果两个智能体都合作，即不击晕对方，并偶尔等待苹果再生，则每个智能体平均获得 $R \approx 0.3$。
- 如果两个智能体都背叛，即击晕对方并完全采收以阻止苹果再生，则每个智能体平均获得 $P \approx 0.06$。
- 如果一个智能体背叛，即击晕对方并采收，背叛的智能体获得 $T \approx 0.9$，而被剥削的智能体平均获得 $S \approx -0.01$（因为大部分时间都处于被击晕状态，且受到 0.01 的时间惩罚）。

预期收益 R、P、T 和 S 满足关于贪婪和恐惧的典型囚徒困境不等式，即 $T > R > P > S$。因此，正如图7和图10所示，即使奖励不断变化，DRIVE 在 *Harvest-12*（将社会困境扩展到12个智能体）中也表现良好。这也证实了 [23], 的实证结果：根据苹果的可用数量和再生速率，Harvest 可以表示一个关于贪婪和恐惧的囚徒困境实例。



（a）Harvest领域

	C	D
C	0.3, 0.3	−0.01, 0.9
D	0.9, −0.01	0.06, 0.06

（b）*Harvest-2* 的收益表（每个时间步）

图9：（a）用于 *Harvest-12* 的领域布局及其初始苹果配置。（b）*Harvest-2* 关于每个时间步预期奖励的收益表。

C 附加结果

我们针对*Harvest-12*提供了不同合作度量的更多结果，分别为社会福利或效率（U）（依据第2节）、平等（E）（1减去基尼系数）、可持续性（S）（收集苹果的时间步数的平均值），以及和平（P）（任意时间步的未标记智能体平均数量）。所有度量均基于原始环境奖励$u_{t,\,i}$而非 $\hat{u}_{t,\,i} = f_{mod}(u_{t,\,i},\ m)$（算法1，第15行），以可靠地评估合作稳定性。它们定义如下 [33]:

$$E = 1 - \frac{\sum_{i\in\mathcal{D}}\sum_{j\in\mathcal{D}}\left|\sum_{t=0}^{H-1}(u_{t,i}-u_{t,j})\right|}{2n\sum_{i\in\mathcal{D}}\sum_{t=0}^{H-1}u_{t,i}}, \tag{11}$$

$$S = \frac{1}{n}\sum_{i\in\mathcal{D}}\chi_i, \text{ where } \chi_i = \mathbb{E}[t|u_{t,i}>0], \tag{12}$$

$$P = n - \frac{1}{H}\sum_{i\in\mathcal{D}}\sum_{t=1}^{H}\mathbf{1}[\text{agent timed-out on time step } t] \tag{13}$$

图 10 展示了第7.2节中不同奖励变化函数 f_{mod}（I、II、III 和 IV）下*Harvest-12*中其他合作度量的变化情况。对于所有度量，DRIVE 都表现出稳健性；而在大多数设置下，MATE 和 LIO 都会显著变化，只有在奖励变化函数 $f^{I}{}_{mod}$和 $f^{III}{}_{mod}$下的平等度量例外。对于所有合作度量，IA 仅受 $f^{IV}{}_{mod}$影响。除 $f^{III}{}_{mod}$外，MATE 的和平度始终高于 DRIVE，但和平水平会随奖励变化函数而显著变化。只有在 $f^{II}{}_{mod}$和 $f^{IV}{}_{mod}$下，LIO 的和平度才高于 DRIVE；这两种情况都会随时间推移收敛至零。

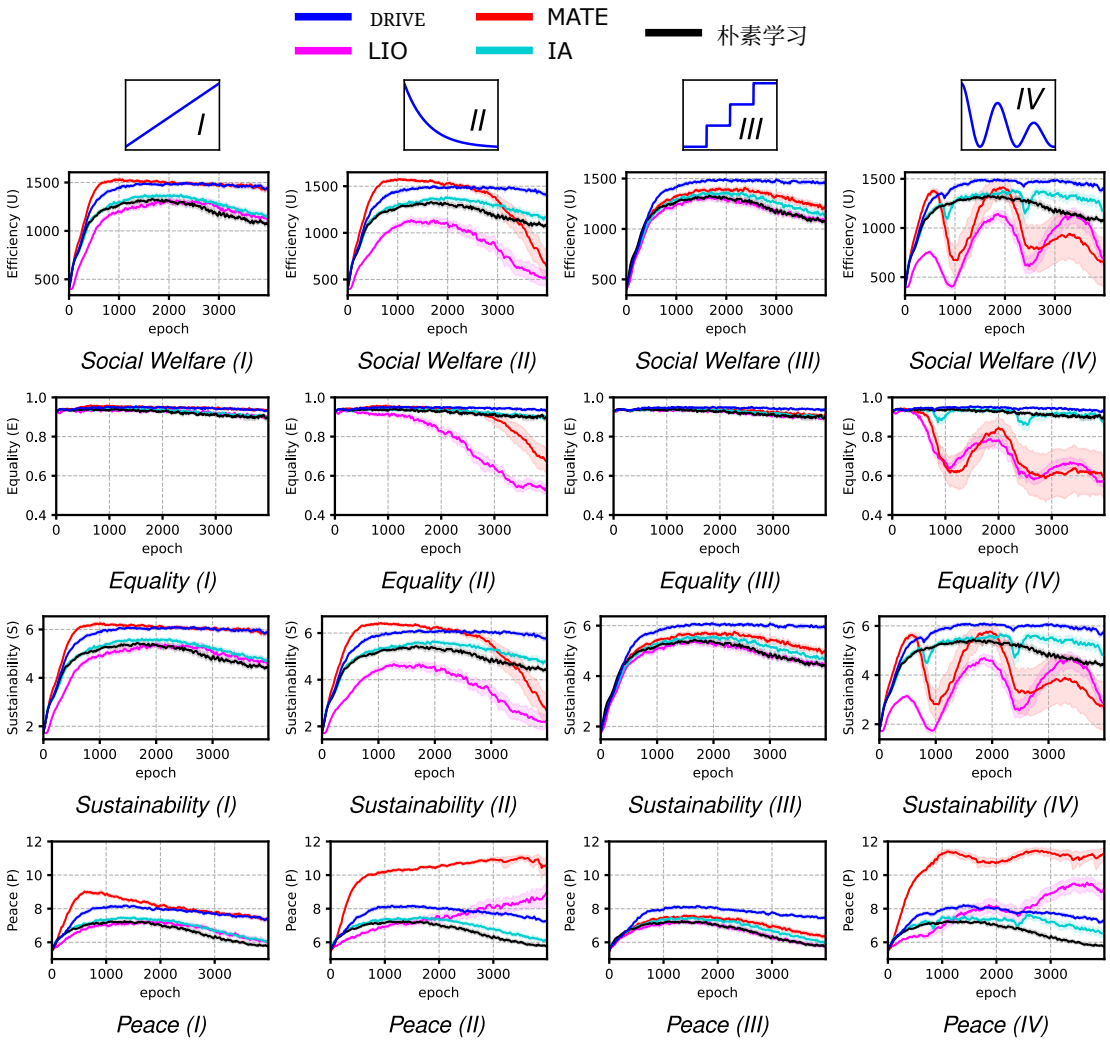


图10：DRIVE 与其他基线方法在 *Harvest-12* 中采用不同奖励变化函数 f_{mod}（I、II、III 和 IV）时的平均进展情况，如顶部图例和图6所示。合作度量包括 社会福利或效率（***U***）、平等（***E***）、可持续性（***S***）和和平（***P***） **[33]。阴影区域表示 95%置信区间。**