

TUC-PPO: Team Utility-Constrained Proximal Policy Optimization for Spatial Public Goods Games

Zhaoqilin Yang^{a,b}, Xin Wang^c, Ruichen Zhang^d, Chanchan Li^e, Youliang Tian^{f,b,*}

^aState Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang, 550025, Guizhou, China

^bInstitute of Cryptography and Data Security, Guizhou University, Guiyang, 550025, Guizhou, China

^cSchool of Mathematics and Statistics, Beijing Jiaotong University, Beijing, 100044, Beijing, China

^dCollege of Computing and Data Science, Nanyang Technological University, Singapore, 639798, Singapore

^eState Key Laboratory of Public Big Data, College of Mathematics and Statistics, Guizhou University, Guiyang, 550025, Guizhou, China

^fState Key Laboratory of Public Big Data, College of Big Data and Information Engineering, Guizhou University, Guiyang, 550025, Guizhou, China

Abstract

We introduce Team Utility-Constrained Proximal Policy Optimization (TUC-PPO), a new deep reinforcement learning framework. It extends Proximal Policy Optimization (PPO) by integrating team welfare objectives specifically for spatial public goods games. Unlike conventional approaches where cooperation emerges indirectly from individual rewards, TUC-PPO instead optimizes a bi-level objective integrating policy gradients and team utility constraints. Consequently, all policy updates explicitly incorporate collective payoff thresholds. The framework preserves PPO’s policy gradient core while incorporating constrained optimization through adaptive Lagrangian multipliers. Therefore, decentralized agents dynamically balance selfish and cooperative incentives. The comparative analysis demonstrates superior per-

*<https://github.com/geek12138/TUC-PPO>

*Corresponding author

Email addresses: zqlyang@gzu.edu.cn (Zhaoqilin Yang), xinwang2@bjtu.edu.cn (Xin Wang), ruichen.zhang@ntu.edu.sg (Ruichen Zhang), ccli@gzu.edu.cn (Chanchan Li), yli@tian@gzu.edu.cn (Youliang Tian)

formance of this constrained deep reinforcement learning approach compared to unmodified PPO and evolutionary game theory baselines. It achieves faster convergence to cooperative equilibria and greater stability against invasion by defectors. The framework formally integrates team objectives into policy updates. This work advances multi-agent deep reinforcement learning for social dilemmas while providing new computational tools for evolutionary game theory research.

Keywords: Spatial public goods games, Deep reinforcement learning, Proximal policy optimization, Team Utility-Constrained

1. Introduction

Cooperative behavior serves as a cornerstone of human civilization, playing a pivotal role in societal development and sustainability across historical epochs [1, 2, 3]. This fundamental mechanism is evident in agricultural communities, where collective irrigation systems demonstrate how coordinated efforts yield benefits that surpass the capabilities of individual farmers. Early Mesopotamian societies thrived precisely because farmers collaboratively maintained complex water distribution networks that no single individual could sustain alone. The evolutionary persistence of such cooperative systems underscores their profound significance in enabling human progress and adaptation to environmental challenges.

The evolutionary persistence of cooperative behavior presents a fundamental paradox in complex adaptive systems, where individual optimization frequently opposes collective welfare [4, 5]. The public goods game framework provides a mathematical formulation of this tension. It captures the essential conflict between private incentives and group benefits through rigorous payoff structures [6, 7, 8]. International climate agreements exemplify this tension. Participating states confront dichotomous options: cooperative emission reduction versus defection through free-riding. The 2015 Paris Agreement illustrates this dynamic, as countries voluntarily set emission targets while facing economic incentives to minimize their costly contributions. Evolutionary game theory delivers a comprehensive framework for dilemma analysis. It specifically elucidates how strategic interactions and population configurations mold cooperation patterns at varying scales [9, 10, 11]. The agreement’s design incorporates key theoretical insights, including reciprocal commitments and transparency mechanisms that mimic the reputation

systems studied in evolutionary models. Network reciprocity mechanisms effectively resolve trust dilemmas in structured populations [12]. In contrast, Lim et al. [13] propose an asymmetric N-player trust game where investors retain agency to abstain from interactions, resolving the critical flaw of ‘investor extinction’. However, conventional models inadequately characterize the dynamic interactions among adaptive agents [14].

Research has identified some primary mechanisms that sustain cooperation against free-riding in social systems, each with distinct real-world manifestations. Positive reinforcement mechanisms [15, 16, 17] operate through incentive structures, as seen in modern carbon credit markets where companies receive tradable credits for emission reductions. Reputation systems [18, 19, 20] function similarly to credit scoring, where businesses maintaining strong environmental records gain preferential access to green financing. Negative constraints [21, 22, 23, 24] appear in various sanctioning forms, exemplified by international trade penalties imposed on nations violating environmental agreements. Exclusion practices [25, 26, 27] manifest in professional networks blacklisting unethical members, paralleling evolutionary models of ostracism. Institutional innovations [28, 29, 30] include progressive carbon taxation systems that scale levies with emission levels. Separately, differential investment rules [31] resemble tiered membership structures in sustainability certification programs. These mechanisms collectively demonstrate how theoretical frameworks translate into practical governance tools for maintaining cooperative systems across economic and environmental domains.

The integration of reinforcement learning (RL) paradigms with evolutionary game theory has significantly advanced our understanding of strategic decision-making processes in social dilemmas. Traditional analytical frameworks based on Fermi update rules and replicator dynamics effectively capture immediate payoff effects and local interactions. However, they neglect the complexity of individual learning processes [32, 33], exemplified by businesses in a trade association adjusting their sustainability investments based on peers’ performance. RL algorithms implement sophisticated closed-loop learning systems that optimize strategies through iterative state-action-reward cycles and experience accumulation [34, 35, 36]. These systems provide more accurate simulations of human decision-making by accounting for both immediate gains and long-term strategic considerations [37, 38, 39].

Among these RL approaches, Q-learning’s robust value iteration mechanism has proven effective at maintaining cooperative equilibria despite strong

free-riding incentives in various game settings [40, 41, 42]. Its applications demonstrate consistent performance across different network topologies in spatial games, with the algorithm’s temporal difference learning enabling effective strategy adaptation [43, 44, 45, 46]. Recent methodological innovations have expanded Q-learning’s applicability through novel combinations with periodic strategy updates and adaptive punishment mechanisms [47]. Moreover, Shen et al. [48] significantly enhance agents’ cooperative inclinations through fused Q-learning/Fermi dynamics.

While traditional RL methods like Q-learning have shown promising results, deep RL approaches offer superior capabilities in handling complex, high-dimensional strategy spaces. Modern deep RL methods like Proximal Policy Optimization (PPO) [49] address critical dimensionality challenges through their neural network architectures. They overcome limitations of tabular methods in large-scale problems. A compelling economic parallel exists in central bank monetary policy committees. There, policymakers employ PPO-like reasoning by continuously adjusting interest rates (policy parameters) based on complex economic indicators (high-dimensional state space). They simultaneously maintain stability through constrained adjustments (clipped policy updates). The architecture enables direct policy optimization with stable training. Stability leverages clipped objectives and adaptive learning rates [50, 51]. This parallels central banks’ policy-stability balancing. Yang et al. [52] were the first to introduce PPO into spatial public goods games (SPGG), developing a two-stage curriculum learning framework that enhances agents’ cooperative tendencies. However, significant theoretical challenges remain in understanding how these modern machine learning approaches interact with population structures and evolutionary dynamics in complex social systems. These open questions illuminate frontier research integrating evolutionary game theory with multi-agent RL. Applications encompass socio-technical system design, notably modeling regulatory-financial co-dynamics.

We propose a Team Utility-Constrained Proximal Policy Optimization (TUC-PPO) framework that establishes a new paradigm for studying cooperation evolution in SPGG. By integrating constrained deep RL with evolutionary dynamics, this approach enables adaptive policy optimization under explicit team utility requirements. The architecture’s dual-objective design bridges individual strategic adaptation with collective welfare preservation, achieving robust equilibrium maintenance in non-stationary multi-agent environments. Systematic game-based evaluations verify the sustainability of

cooperation under strong free-riding incentives. Quantitative comparisons further confirm the framework’s superior convergence efficiency and behavioral stability versus baselines. Our research makes three fundamental contributions:

- We establish the first integration of team utility constraints into policy gradient optimization for evolutionary games. This achieves a controlled balance between individual rationality and collective welfare. Furthermore, the Lagrangian dual-ascent mechanism provides new mathematical foundations for cooperation sustainability.
- We design a self-adjusting constraint mechanism that dynamically adapts penalty coefficients through batch-wise violation evaluation. This ensures team utility thresholds while maintaining policy update stability via constrained gradient updates.
- The PPO-based framework uniquely addresses SPGG challenges through clipped policy updates and advantage estimation. This enables stable strategy adaptation in high-dimensional non-stationary environments where traditional evolutionary methods exhibit oscillation and convergence failures.

The rest of this paper is organized as follows. Section 2 establishes the theoretical framework of SPGG. It further provides rigorous derivation of the TUC-PPO algorithm integrating constrained policy gradients with evolutionary strategy updates. Analysis in 3 systematically examines cooperation evolution under varying initial conditions through three complementary dimensions: (1) sensitivity analysis of hyperparameter configurations, (2) comparative performance against conventional algorithms, and (3) Evolution under different initialization strategies. Section 4 concludes the study.

2. Model

We consider an SPGG on an $L \times L$ periodic lattice with von Neumann neighborhood ($k = 4$). Each agent participates in $G = 5$ game groups centered on itself and its four neighbors. The strategy space is $\mathcal{S} = \{C, D\}$, where cooperators (C) contribute 1 unit to the public pool while defectors (D) contribute nothing. The payoff for agent i in group g is:

$$\Pi(s_i^g) = \begin{cases} \frac{rN_C^g}{k+1} - 1, & s_i^g = C \\ \frac{rN_C^g}{k+1}, & s_i^g = D \end{cases}, \quad (1)$$

where $N_C^g = \sum_{j \in g} \mathbb{I}(s_j^g = C)$ counts cooperators in group g , and $r > 1$ is the enhancement factor. The total payoff aggregates over all groups:

$$\Pi_i = \sum_{g \in \mathcal{G}_i} \Pi(s_i^g). \quad (2)$$

This SPGG framework provides the foundation for adapting proximal PPO to evolutionary game dynamics. The key innovation of our TUC-PPO lies in transforming conventional individual reward maximization into a team-optimization paradigm. In this paradigm, agents learn strategies that balance personal gains with collective welfare requirements. Unlike standard PPO, TUC-PPO introduces team utility as a fundamental constraint on policy updates, requiring each agent’s actions to maintain minimum contribution thresholds for their local game groups. Unlike standard PPO, TUC-PPO introduces team utility as a fundamental constraint on policy updates. This requires each agent’s actions to maintain minimum contribution thresholds for their local game groups.

2.1. TUC-PPO

The proposed method extends PPO by introducing team utility constraints and adaptive reward balancing. The core formulation begins with a constrained optimization problem. This formulation aims to maximize the expected cumulative composite reward while ensuring the average team utility meets a minimum threshold τ . This translates mathematically to:

$$\begin{cases} \max_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{t=0}^T (1 - w_t) r_t^{\text{ind}} + w_t r_t^{\text{team}} \right] \\ \text{s.t.} \quad \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\frac{1}{T} \sum_{t=1}^T r_t^{\text{team}} \right] \geq \tau \end{cases} \quad (3)$$

where $\theta \in \mathbb{R}^d$ denotes the d -dimensional trainable parameters of the policy network, $\mathbb{E}_{\tau \sim \pi_{\theta}}$ represents the expectation over trajectories $\tau = (\mathbf{S}_0, a_0, r_0, \dots)$ generated by policy π_{θ} . $\mathbf{S}_t \in \mathbb{R}^{L \times L \times 3}$ is the encoded state tensor in combining strategy matrix, neighborhood cooperation counts, and global cooperation rate. a_t is the joint action matrix in $\{0, 1\}^{L \times L}$ sampled from the policy π_{θ} . $r_t \in \mathbb{R}^{L \times L}$ is the immediate reward in combining individual and team components through adaptive weighting. T is the episode horizon, $w_t \in [0, 1]$

is the adaptive weight, and τ is the minimum team performance threshold. The threshold $\tau = 0.5$ corresponds to a minimum 50% neighbors cooperating in a Prisoner's Dilemma and is Equivalent to the Nash equilibrium payoff in a 2-player game. The individual reward r_t^{ind} for agent i is:

$$r_t^{\text{ind}}(i) = \sum_{g \in \mathcal{G}_i} \Pi(s_i^g). \quad (4)$$

The team utility r_t^{team} for agent i is:

$$r_t^{\text{team}}(i) = \begin{cases} \frac{r}{k+1} \left(N_C^{\text{neigh}}(i) + \mathbb{I}(s_i = C) \right) - 1 & \text{if } s_i = C \\ 0 & \text{if } s_i = D \end{cases} \quad (5)$$

where $N_C^{\text{neigh}}(i)$ counts cooperating neighbors in agent i 's von Neumann neighborhood (4 adjacent cells). $\mathbb{I}(s_i = C)$ is an indicator function that equals 1 when agent i cooperates ($s_i = C$) and 0 otherwise. \mathcal{G}_i represents the 5 game groups agent i participates in (centered on itself and its 4 neighbors). The division by 5 normalizes the payoff across all participating groups.

The constrained optimization problem is transformed through Lagrangian relaxation, yielding the primal-dual formulation:

$$\max_{\theta} \min_{\eta \geq 0} \mathcal{L}(\theta, \eta) = \mathbb{E}_{\tau} \left[\sum_{t=0}^T \gamma^t R_t \right] + \eta \left(\mathbb{E}_{\tau} \left[\frac{1}{T} \sum_{t=1}^T r_t^{\text{team}} \right] - \tau \right) \quad (6)$$

The inequality constraint is converted using the ReLU [53] function:

$$\max \left(0, \tau - \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\frac{1}{T} \sum_{t=1}^T r_t^{\text{team}} \right] \right) = 0 \quad (7)$$

In implementation, we approximate the expectation using mini-batch averages:

$$\mathbb{E}_{\tau \sim \pi_{\theta}} \left[\frac{1}{T} \sum_{t=1}^T r_t^{\text{team}} \right] \approx \frac{1}{B} \sum_{i=1}^B r_i^{\text{team}} \quad (8)$$

where B is the batch size.

This leads to the penalty term with the dual variable $\eta \geq 0$:

$$L^{\text{CV}}(\eta) = \eta \cdot \underbrace{\left(\tau - \frac{1}{B} \sum_{i=1}^B r_i^{\text{team}} \right)}_{\text{Constraint violation}}_+ \quad (9)$$

where $\max(0, x)$ ensures one-sided penalty. Here, η represents the Lagrange multiplier that dynamically scales the magnitude of the constraint violation.

The dual variable update rule is:

$$\eta \leftarrow \eta + \zeta \cdot L^{\text{CV}}(\eta) \quad (10)$$

with ζ denoting the dual learning rate that controls the adjustment speed of η . This update follows the dual gradient ascent principle.

Applying Lagrangian relaxation transforms this into an unconstrained optimization problem with dual variable η :

$$L^{\text{TUC}}(\theta, \eta) = L^{\text{CLIP}}(\theta) + \delta L^{\text{VF}}(\theta) - \rho L^{\text{ENT}}(\theta) + \eta L^{\text{CV}}(\eta) \quad (11)$$

where $L^{\text{CLIP}}(\theta)$ is the clipped policy objective that prevents excessively large policy updates. δ is the coefficient balancing policy and value function updates. $L^{\text{VF}}(\theta)$ is the value function loss that minimizes the squared error between predicted and actual returns. ρ is the entropy coefficient that controls the strength of the exploration incentive. $L^{\text{ENT}}(\theta)$ is the policy entropy term that encourages exploration by penalizing low-entropy policies.

The clipped policy objective $L^{\text{CLIP}}(\theta)$ prevents drastic policy changes by constraining the update ratio $\pi_\theta / \pi_{\theta_{\text{old}}}$ to $[1 - \epsilon, 1 + \epsilon]$. The policy ratio $r_t(\theta) = \pi_\theta(a_t | s_t) / \pi_{\theta_{\text{old}}}(a_t | s_t)$ is constrained within $[1 - \epsilon, 1 + \epsilon]$ where ϵ controls the maximum policy deviation:

$$L^{\text{CLIP}}(\theta) = -\mathbb{E}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right] \quad (12)$$

where the advantage estimates \hat{A}_t are computed using a weighted combination of individual and team rewards across the trajectory. This generalized advantage estimation balances immediate and long-term returns through discount factors γ (reward decay) and λ (bias-variance trade-off):

$$\hat{A}_t = \sum_{l=0}^{T-t} (\gamma \lambda)^l \left[(1 - w_t) r_{t+l}^{\text{ind}} + w_t r_{t+l}^{\text{team}} + \gamma V(s_{t+l+1}) - V(s_{t+l}) \right] \quad (13)$$

The value function loss $L^{\text{VF}}(\theta)$ trains the critic network by minimizing the prediction error:

$$L^{\text{VF}}(\theta) = \mathbb{E}_t [(V_\theta(s_t) - R_t)^2] \quad (14)$$

where the target return R_t is computed through backward recursion:

$$R_t = \gamma [(1 - w_t)r_t^{\text{ind}} + w_t r_t^{\text{team}} + (1 - \mathbb{I}_t^{\text{done}})R_{t+1}] \quad (15)$$

with $\mathbb{I}_t^{\text{done}}$ indicating episode termination. This recursive form is mathematically equivalent to the discounted sum but more computationally efficient. The adaptive weight w_t dynamically balances individual and team rewards based on their historical ratio:

$$w_t = \sigma \left(\frac{\sum_{i=1}^t r_i^{\text{team}}}{\sum_{i=1}^t r_i^{\text{ind}} + 10^{-8}} \right) \quad (16)$$

To encourage exploration, entropy regularization $L^{\text{ENT}}(\theta)$ adds a bonus proportional to the policy’s information entropy:

$$L^{\text{ENT}}(\theta) = \mathbb{E}_t \left[- \sum_{a \in \mathcal{A}} \pi_\theta(a|s_t) \log \pi_\theta(a|s_t) \right] \quad (17)$$

2.2. Actor-Critic Network Architecture

The Actor-Critic network processes spatial game states through the following mathematical operations. Let $s_t^i = [x_t^i, n_t^i, g_t] \in \mathbb{R}^3$ denote the encoded state for agent i at time t , where $x_t^i \in \{0, 1\}$ represents the current strategy (0: defection, 1: cooperation). $n_t^i \in \mathbb{N}$ counts cooperating neighbors within the von Neumann neighborhood. $g_t \in [0, 1]$ indicates the global cooperation rate. Fig. 1 shows the actor-critic network architecture.

The Encoder transforms inputs through:

$$h_t^i = \sigma(W_2 \sigma(W_1 s_t^i + b_1) + b_2) \quad (18)$$

where $\sigma(\cdot)$ denotes the ReLU activation function, $W_1 \in \mathbb{R}^{64 \times 3}$ and $W_2 \in \mathbb{R}^{64 \times 64}$ are weight matrices, $b_1, b_2 \in \mathbb{R}^{64}$ are bias terms. The Actor branch computes action probabilities via:

$$\pi_\theta(a_t^i | s_t^i) = \text{softmax}(W_a h_t^i + b_a) \quad (19)$$

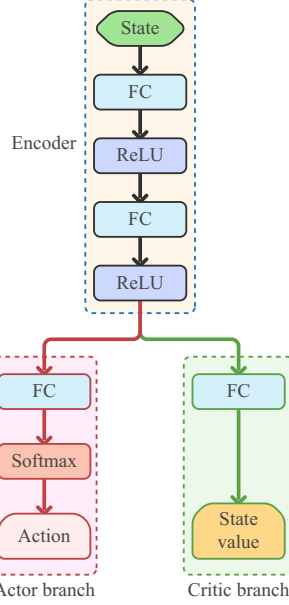


Figure 1: The architecture of actor-critic network.

with $W_a \in \mathbb{R}^{2 \times 64}$ and $b_a \in \mathbb{R}^2$ producing a categorical distribution over actions (cooperate/defect). The Critic branch estimates state values as:

$$V_\theta(s_t^i) = w_v^T h_t^i + b_v \quad (20)$$

where $w_v \in \mathbb{R}^{64}$ and $b_v \in \mathbb{R}$. For an $L \times L$ grid, the joint policy factorizes as $\pi_\theta(a_t|s_t) = \prod_{i=1}^{L^2} \pi_\theta(a_t^i|s_t^i)$.

The proposed TUC-PPO framework is described as Algorithm 1, which integrates team utility constraints with adaptive reward balancing in a spatially structured environment. The algorithm is as follows:

Algorithm 1 PPO-TUC Framework for SPGG

```
1: Initialize:
2:   Policy network  $\pi_\theta$  and value network  $V_\phi$  with shared features
3:   Team utility threshold  $\tau$ , dual variable  $\eta$ 
4:   Hyperparameters:  $\delta$ ,  $\rho$ ,  $\zeta$ , clip range  $\epsilon$ 
5:
6: for each epoch  $t = 1$  to  $T$  do
7:   for each agent  $i$  in grid do
8:     Sample action  $a_t^i \sim \pi_\theta(s_t^i)$  (Eq. 19)
9:     Compute composite reward  $r_t^i = (1 - w_t^i)r_t^{\text{ind},i} + w_t^i r_t^{\text{team},i}$ 
10:   end for
11:   Compute constraint violation  $L^{\text{CV}}$  (Eq. 9)
12:   Compute GAE advantage  $\hat{A}_t$  (Eq. 13)
13:   Update dual variable  $\eta$  (Eq. 10)
14:   Optimize surrogate objective:
15:      $L^{\text{TUC}} = L^{\text{CLIP}} + \delta L^{\text{VF}} - \rho L^{\text{ENT}} + \eta L^{\text{CV}}$  (Eq. 11)
16:   Update  $\theta$  via gradient descent on  $L^{\text{TUC}}$ 
17:   Clear experience buffer
18: end for
```

3. Experimental results

3.1. Experimental setup

The default configuration employs a 200×200 spatial grid with parameter settings. Learning rate $\alpha = 1 \times 10^{-4}$, discount factor $\gamma = 0.99$, generalized advantage estimation parameter $\lambda = 0.95$, PPO clip threshold $\epsilon = 0.2$, critic loss weight $\delta = 0.5$, and entropy regularization coefficient $\rho = 0.01$. The team utility mechanism operates with threshold $\tau = 0.5$, and dual learning rate $\zeta = 0.01$ for constrained optimization. For constraint evaluation in TUC-PPO, the effective batch size B is set to encompass the entire experience buffer. This substantial B value provides a comprehensive trajectory evaluation while maintaining reasonable computational requirements. Parameter updates utilize the Adam optimizer [54] with StepLR scheduling that halves the learning rate every 1,000 iterations.

All experiments were executed on mobile workstation hardware featuring an AMD Ryzen 9 5900HX CPU and NVIDIA GeForce RTX 3080 Laptop GPU (16GB GDDR6) under Ubuntu 22.04.5 LTS, with PyTorch 2.2.1 leveraging CUDA 12.8 acceleration. The spatial visualization protocol encodes

strategic choices using binary cell states: defection represented by black and cooperation by white RGB.

3.2. Comparative analysis of algorithms

Figure 2 presents a comparative analysis of four algorithms: PPO-TUC, PPO, Q-learning, and the Fermi update rule. These were evaluated under an enhancement factor $r = 3.3$. The simulation initializes with defectors concentrated in the upper half of the domain and cooperators in the lower half. The left subfigure displays the evolution of cooperation and defection rates over time. In this visual, the iteration count is plotted along the horizontal axis, with the fraction of cooperators and defectors represented by blue and red curves, respectively. The remaining subfigures capture spatial snapshots at key intervals, where white pixels denote cooperators and black pixels indicate defectors. By examining both temporal trends and spatial configurations, the results demonstrate significant variations in algorithmic performance within critical parameter ranges. TUC-PPO and PPO have only undergone 1,000 iterations, while Q-learning and Fermi update rules have undergone 10,000 iterations.

Our TUC-PPO achieves full cooperation among all agents within just 20 iterations, demonstrating significantly faster convergence than other methods. Initially, the randomly initialized policy network leads to mixed strategies among agents. The standard PPO without the TUC mechanism rapidly converges to complete defection within 100 iterations, with defectors eventually dominating the entire population later in training. Q-learning requires nearly 300 iterations to stabilize, yet the cooperation fraction remains below 40%, reflecting its inefficiency in policy optimization. The Fermi update rule, under the current enhancement factor, fails to sustain cooperation, with defectors completely overtaking the population before 300 iterations. These results highlight TUC-PPO’s superior convergence speed and stability in promoting cooperation. While the PPO and Fermi update rule succumb to defectors, and Q-learning struggles with inefficient learning, our method ensures rapid and robust cooperation emergence. The comparison underscores the critical role of temporal update coordination in multi-agent RL for evolutionary games.

3.3. Algorithm performance evaluation under varying enhancement factors r

The experiment systematically evaluates four computational approaches including TUC-PPO, PPO, Q-learning, and Fermi dynamics. The experi-

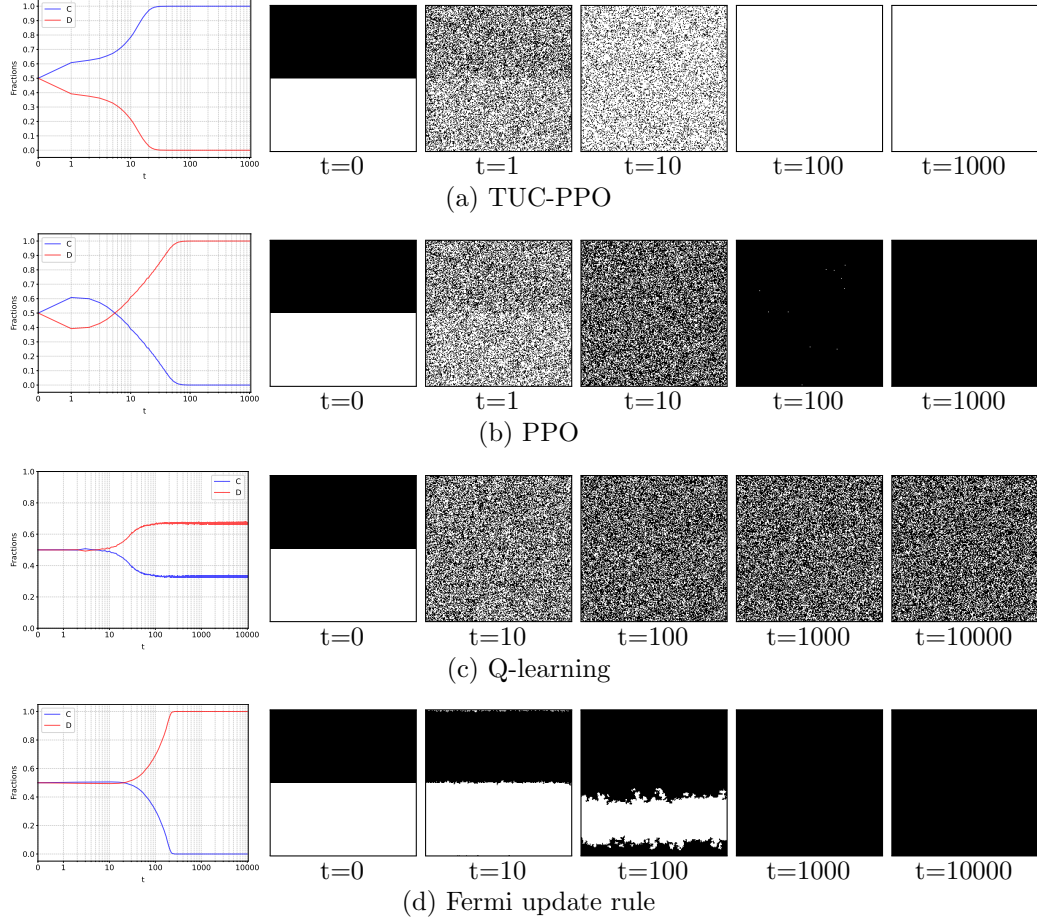


Figure 2: Simulation for TUC-PPO, PPO, Q-learning, and Fermi update rule at enhancement factor $r = 3.3$. Initial conditions place defectors (black) in the upper half and cooperators (white) in the lower half. The leftmost subfigure in each row presents the temporal evolution curve. Subfigures (a) and (b) show state snapshots at iterations $t = 0, 1, 10, 100$, and 1000 . Subfigures (c) and (d) display snapshots at $t = 0, 10, 100, 1000$, and 10000 . TUC-PPO achieves rapid global cooperation, outperforming baselines that either fail to sustain cooperation (PPO/Fermi), or achieve only suboptimal cooperation (Q-learning).

mental setup maintains consistent initial spatial distributions where defectors occupy the upper region and cooperators the lower region of the domain. Figure 3 displays the relationship between fractions measured on the vertical axis and enhancement factor r shown on the horizontal axis.

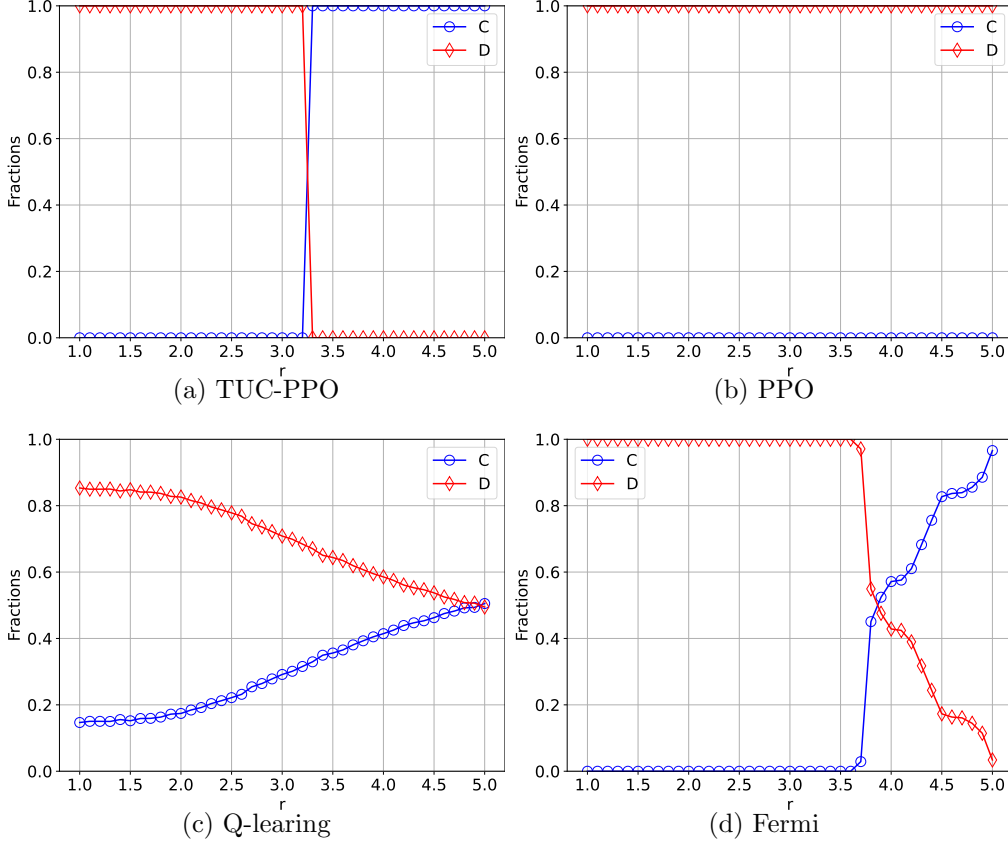


Figure 3: Performance comparison of TUC-PPO, PPO, Q-learning, and Fermi update rule under varying enhancement factors r . Blue circles represent cooperators while red diamonds indicate defectors. Initial conditions positioned all defectors D in the upper half of the grid and all cooperators C in the lower half. TUC-PPO achieves full cooperation at lower enhancement factors, outperforming other methods that fail to sustain cooperation, reach only suboptimal levels, or require higher incentive thresholds.

Our TUC-PPO framework demonstrates superior performance in promoting cooperative behavior under demanding conditions. When the enhancement factor $r \geq 3.3$, all agents become cooperators by the end of 1000 iterations. This achievement originates from TUC-PPO’s team utility constraint, which effectively strengthens agents’ inclination toward cooperation. The

algorithm achieves complete cooperation at substantially lower enhancement factors compared to traditional approaches. Standard PPO without TUC exhibits fundamental limitations in the discovery of cooperative strategies. For enhancement factors below 5.0, all agents inevitably converge to become defectors. This pathological convergence demonstrates PPO’s tendency toward extreme strategy polarization. Such pathological convergence stems from the policy update mechanism amplifying initial random biases. This process drives the entire population toward complete cooperation or complete defection, eliminating intermediate stable states. Q-learning shows different behavioral characteristics, maintaining some cooperators even at relatively low r -values. While the fraction of cooperators increases with higher enhancement factors, the approach fails to achieve majority cooperation, reaching only 55% even at $r = 5.0$. This performance ceiling reflects the inherent limitations of tabular methods in modeling continuous policy spaces and spatial agent interactions. The Fermi update rule demonstrates intermediate performance, producing cooperative when r exceeds 3.7. However, this imitate update mechanism cannot guarantee full cooperation even at $r = 5.0$, instead showing gradual improvement in cooperation levels. This contrasts sharply with TUC-PPO’s ability to achieve complete cooperation at substantially lower enhancement factors.

The comparative results yield fundamental insights about TUC-PPO’s advantages. The team utility constraint successfully enforces cooperative behavior in scenarios where conventional methods fail. This constraint mechanism achieves full cooperation with remarkable efficiency, outperforming both value-based methods and local interaction rules. Most importantly, the results demonstrate how explicitly modeling team utility can resolve core limitations in multi-agent learning systems. This approach offers an effective solution for evolutionary game scenarios. These findings position TUC-PPO as an important innovation in cooperative RL through its principled team-based optimization approach.

3.4. Statistical analysis of TUC-PPO

To rigorously evaluate the operational reliability of TUC-PPO, we implemented three complementary statistical visualization approaches. These methods analyze distribution spread through error bars, uncover probability density patterns via violin plots, and detail precision metrics using comparative tables of 95% confidence intervals. Across the enhancement factor

domain where $r \in [1.0, 6.0]$, we executed 50 independent experimental trials for each parameter configuration.

Experimental executions consistently generated dichotomous outcomes for TUC-PPO and PPO, where individual runs manifested either 0% or 100% cooperation probabilities. This behavior produced substantial standard deviations across all enhancement factor values. Fig. 4 fundamentally distinguishes TUC-PPO from baseline PPO through error bar analysis quantifying differences in cooperation rate distributions and variability patterns. This visualization confirms the core mechanism’s efficacy in diminishing cooperation barriers. Specifically, TUC-PPO achieves reliable cooperation significantly earlier than the baseline algorithm while requiring lower enhancement factors.

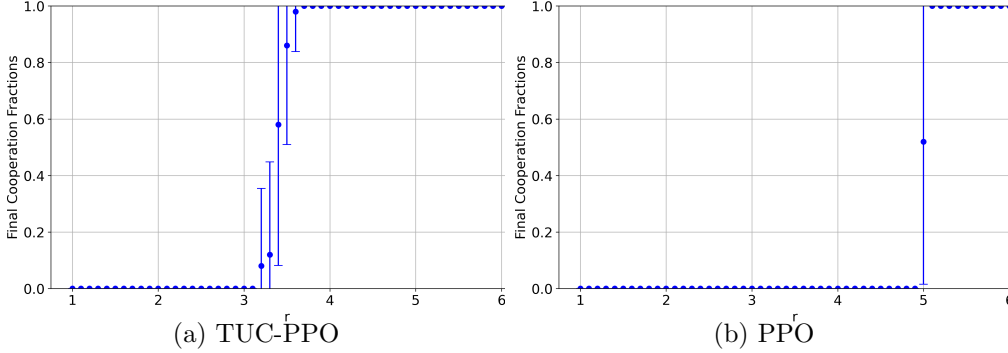


Figure 4: Comparison of mean cooperation fractions and standard deviation between TUC-PPO and PPO. TUC-PPO achieves a state where all agents are collaborators at lower enhancement factors r than standard PPO through the integration of the TUC.

Violin plot analysis (Fig. 5) reveals TUC-PPO’s superior cooperation dynamics compared to baseline PPO. At enhancement factors below 3.2, TUC-PPO maintains stable 0% cooperation. Between $r = 3.2$ and 3.6, its cooperation distribution progressively approaches optimal values. Beyond $r=3.6$, TUC-PPO achieves perfect 100% cooperation stability. This performance contrasts with baseline PPO, which sustains 0% cooperation below $r = 5$ and reaches stable 100% cooperation above this threshold. At the critical $r = 5$ point, baseline PPO achieves approximately half the cooperation rate of TUC-PPO. The distributions demonstrate TUC-PPO’s decisive advantages through establishing stable cooperation at lower enhancement factors while maintaining superior performance at critical thresholds and transitioning more efficiently to optimal cooperation.

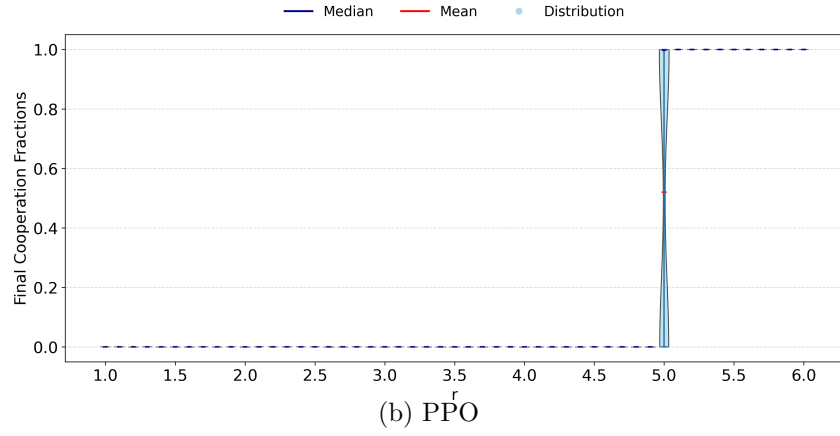
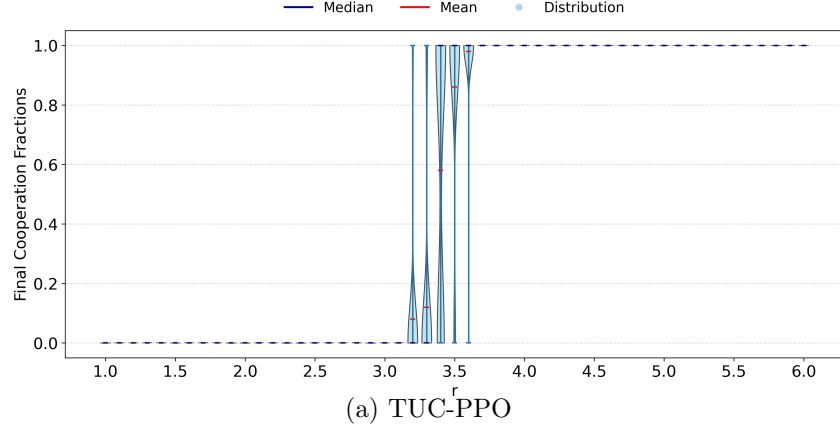


Figure 5: Probability density of cooperation fractions for TUC-PPO and standard PPO across enhancement factors $r \in [1.0, 6.0]$. Violin plots show that compared to PPO, TUC-PPO enables all agents to ultimately select the cooperative strategy at smaller enhancement factors.

As shown in Table 1, the confidence interval analysis confirms two critical behavioral mechanisms affecting convergence. First, NaN intervals emerge when algorithms achieve perfect deterministic outcomes (100% cooperation or 0% cooperation) across all trials. This results in zero standard deviation, which invalidates confidence interval computation. Second, the $[0.00, 0.00]$ intervals represent scenarios where exploration strategies prevent absolute convergence. During 1000 training iterations, PPO’s exploration mechanisms cause strategy deviations that maintain minimal but non-zero cooperation or defection probabilities. This proves that 1000 iterations cannot achieve complete convergence for PPO, but it is already very close. These stochastic policy decisions prevent either perfect cooperation or complete defection in final states. The results demonstrate TUC-PPO’s superior adaptability in social dilemmas. Moreover, PPO requires substantially higher enhancement factors ($r > 5.0$) to achieve meaningful cooperation and suffers from convergence instability at critical thresholds. In contrast, TUC-PPO establishes reliable cooperation patterns at lower enhancement factors ($r \geq 3.6$) with smoother probabilistic transitions. This performance advantage stems from TUC-PPO’s architectural capability to maintain cooperative equilibria despite exploration noise. Comparatively, PPO’s strategy exhibits greater vulnerability to exploration-induced deviations that delay convergence and reduce cooperation reliability.

Table 1: 95% confidence intervals comparison for cooperation fractions

r	3.1	3.2	3.3	3.4	3.5	3.6	3.7
TUC-PPO	nan – nan	0.00 – 0.16	0.03 – 0.21	0.44 – 0.72	0.76 – 0.96	0.94 – 1.02	nan – nan
PPO	nan – nan	0.00 – 0.00	0.00 – 0.00	0.00 – 0.00	0.00 – 0.00	0.00 – 0.00	0.00 – 0.00
r	3.8	3.9	4.0	4.1	4.2	4.3	4.4
TUC-PPO	nan – nan	nan – nan	nan – nan	nan – nan	nan – nan	nan – nan	nan – nan
PPO	nan – nan	nan – nan	0.00 – 0.00	0.00 – 0.00	0.00 – 0.00	0.00 – 0.00	0.00 – 0.00
r	4.5	4.6	4.7	4.8	4.9	5.0	5.1
TUC-PPO	nan – nan	nan – nan	nan – nan	nan – nan	nan – nan	nan – nan	nan – nan
PPO	nan – nan	nan – nan	0.00 – 0.00	0.00 – 0.00	0.00 – 0.00	0.38 – 0.66	nan – nan

3.5. TUC-PPO with half-and-half initialization

Fig. 6 shows the evolution curve for TUC-PPO under spatially partitioned initial conditions. The study implements a controlled initialization protocol where defectors and cooperators are systematically allocated to distinct grid

regions before training. Specifically, defector agents populate the upper half of the spatial domain while cooperators occupy the lower half, establishing well-defined initial strategy boundaries. The investigation examines system dynamics across two enhancement factor values: $r = 3.0$ and $r = 4.0$. These parameter selections enable comparative analysis of cooperation emergence under different environmental conditions. The experimental output combines quantitative temporal metrics with qualitative spatial representations to capture both macroscopic and microscopic evolutionary patterns. The experimental output incorporates three complementary visualization modalities to capture different aspects of system dynamics. First, temporal progression curves track the evolution of strategy. Cooperator and defector populations are represented through distinct colored trajectories plotted against the iteration count. Second, spatial strategy distributions display agent-type configurations at key iterations, using high-contrast markers to indicate individual decisions. Third, payoff heatmaps visualize the immediate reward landscape, employing a color gradient to represent each agent’s current earnings. This tripartite visualization framework enables simultaneous examination of temporal trends, spatial patterns, and economic incentives throughout the evolutionary process.

The experimental results reveal a critical dependence of cooperative behavior on the enhancement factor r . Figs. 6 (a) and (c) demonstrate the $r = 3.0$ scenario, where the fractions of cooperators decay almost monotonically, with all agents adopting defector strategies within 100 iterations. Spatial snapshots document the progressive territorial expansion of defectors, while payoff heatmaps reveal a corresponding decline in individual rewards as cooperation diminishes. Conversely, Figs. 6 (b) and (d) illustrate the $r = 3.5$ scenario where the enhancement factor crosses the critical threshold for cooperation sustainability. Here we observe rapid convergence to universal cooperation, with complete conversion occurring before 40 iterations. The visualization panels demonstrate cooperators systematically replacing defectors, accompanied by steadily improving payoff distributions across the population. This dichotomy highlights the crucial role of the enhancement factor in determining system equilibrium. The results confirm that TUC-PPO successfully maintains cooperative equilibrium when $r = 3.5$, while systems inevitably collapse to defection when $r = 3.0$. The payoff heatmaps provide direct evidence that cooperative states yield superior economic outcomes compared to defective equilibria.

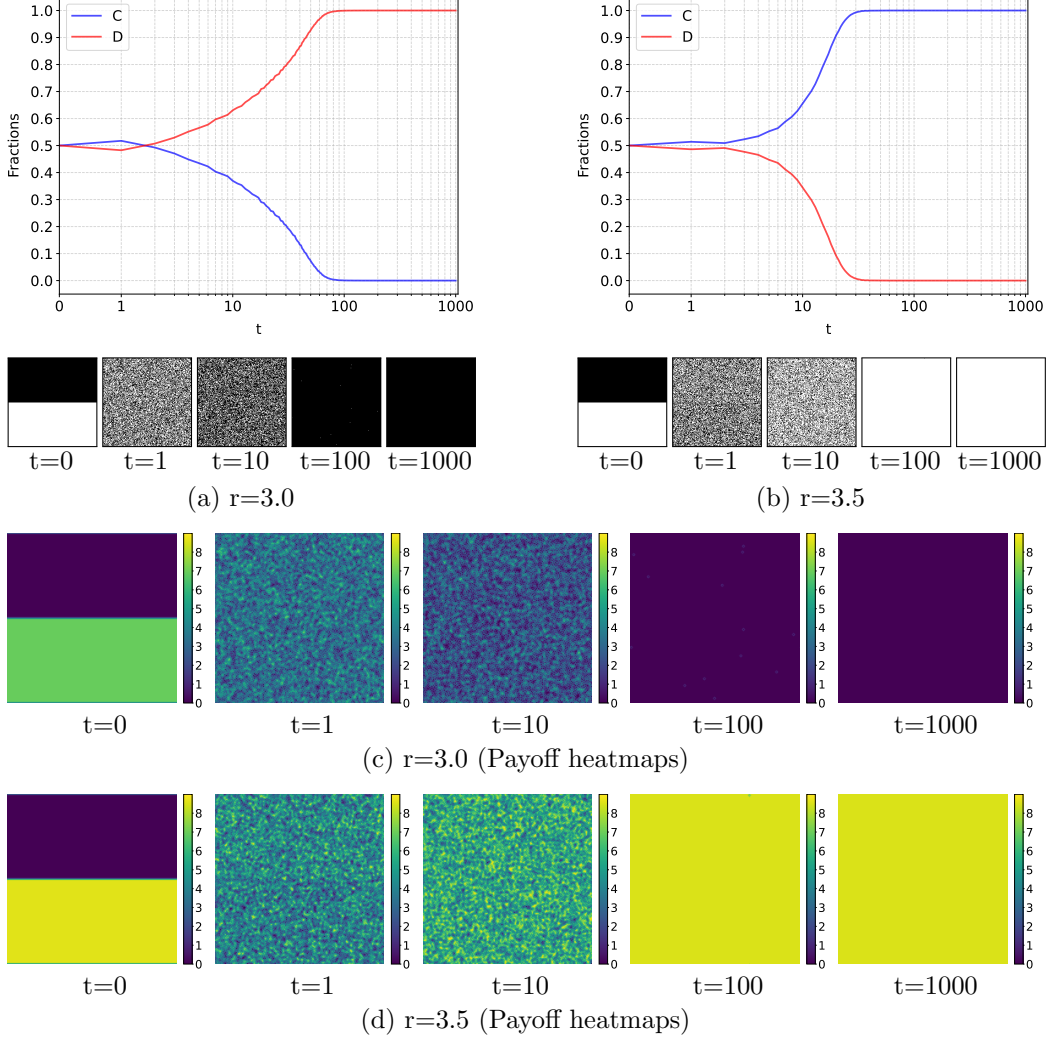


Figure 6: Evolution of cooperation and payoffs in SPGG using TUC-PPO. Initial conditions place defectors in the upper half of the grid and cooperators in the lower half. **(a, b)** Temporal evolution of cooperation (blue curves) and defection (red curves), with corresponding strategy snapshots below (black blocks: defectors, white blocks: cooperators). **(c, d)** Payoff heatmaps at key iterations, showing individual agent earnings. **(a, c)** Results for enhancement factor $r = 3.0$; **(b, d)** Results for $r = 3.5$. Convergence properties demonstrate that TUC-PPO enables all agents to rapidly stabilize in pure cooperation or defection states within minimal iterations. This process features near-absent oscillations during strategy evolution.

3.6. TUC-PPO with bernoulli random initialization

This experimental series investigates TUC-PPO dynamics under randomized initial conditions. Agent strategies are assigned following a Bernoulli distribution with equal probability $p = 0.5$ for both cooperation and defection strategies. Fig. 7 systematically compares the evolutionary results in different enhancement factors r . Each subfigure contains three integrated visualization components. The upper section displays temporal evolution profiles, plotting the fractions of cooperators (shown in blue) and defectors (shown in red) against the iteration count t . The middle section presents spatial strategy distributions at key evolutionary stages, using white pixels to represent cooperators and black pixels for defectors. The lower section provides payoff heatmaps that visualize individual agent rewards.

Fig. 7 with Bernoulli-distributed initial strategies shows a fundamental dependence of cooperation dynamics on the enhancement factor r . Under the $r = 3.0$ condition, the cooperation fraction exhibits a steady decline over time, with all agents transitioning to defector strategies within 80 iterations. Spatial evolution patterns reveal progressively increasing black regions representing defectors, while the payoff heatmaps show corresponding deterioration in individual rewards as cooperation diminishes. In contrast, the $r = 3.5$ scenario shows rapid system convergence, with complete adoption of cooperative strategies occurring before 40 iterations. The visualization panels demonstrate a swift transition to uniform cooperation, accompanied by significantly improved payoff distributions throughout the population. These results underscore the pivotal role of the enhancement factor in governing system evolution from random initial conditions. The TUC-PPO algorithm successfully drives the population to full cooperation when $r=3.5$, while inevitably leading to universal defection when $r = 3.0$. The payoff heatmaps provide conclusive evidence that cooperative equilibria generate superior economic outcomes compared to defective states.

3.7. TUC-PPO with all-defectors initialization

This investigation examines the evolutionary dynamics of TUC-PPO under extreme initial conditions where all agents initially adopt defector strategies. As shown in Fig. 8, this experimental design demonstrates the framework’s capability to overcome fundamental limitations inherent in traditional evolutionary game theory methods. Conventional approaches like the Fermi update rule often fail to initiate cooperative emergence from homogeneous

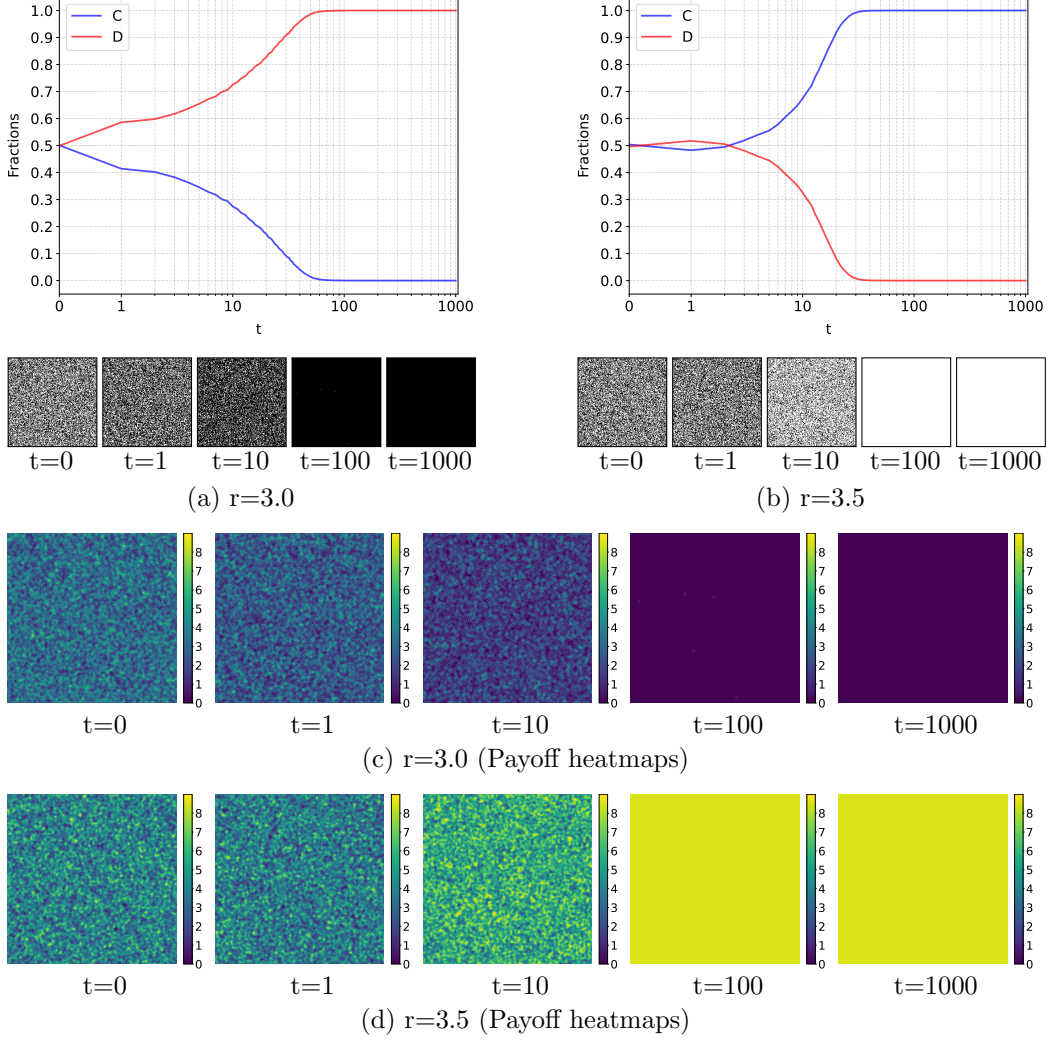


Figure 7: Evolution of cooperation and payoffs in SPGG using TUC-PPO with random initialization. Agent strategies are initially assigned according to a Bernoulli distribution, with equal probabilities for cooperation and defection. **(a, b)** Temporal evolution of cooperation is shown in blue, and defection is shown in red, accompanied by strategy distribution snapshots where defectors appear as black blocks and cooperators as white blocks. **(c, d)** Payoff heatmaps displaying individual agent rewards at selected iterations. **(a, c)** Results for enhancement factor $r = 3.0$ while **(b, d)** show $r = 3.5$ conditions. Convergence to cooperation proves markedly more challenging under random initialization than with half-and-half initialization.

defector populations due to their reliance on neighborhood strategy diversity. The TUC-PPO framework addresses this challenge through its team-utility-constrained RL architecture, which enables the discovery of cooperative strategies even in these demanding initial conditions.

Fig. 8 demonstrates TUC-PPO’s robust performance when starting from all-defectors initialization. The random initialization of Actor-Critic network parameters leads to stochastic initial strategy selection. Under $r = 3.0$ conditions, the cooperation rate first increases to around 50% before gradually declining to zero. Spatial patterns show temporary cooperative clusters being replaced by expanding defector regions, while payoff maps reflect decreasing individual rewards during this process. For $r = 3.5$, the system exhibits stable growth in cooperation, reaching full cooperation within 40 iterations. The visualizations demonstrate cooperators systematically replacing defectors, with payoff distributions showing continuous improvement across the population.

These results reveal two important findings. First, TUC-PPO maintains consistent convergence properties regardless of the initialization scheme, proving its algorithmic robustness. Second, the framework demonstrates significantly superior performance compared to rule-based approaches like the Fermi update rule. Crucially, the Fermi method shows high sensitivity to initial conditions and cannot escape all-defectors equilibria without external intervention. The payoff dynamics further demonstrate that TUC-PPO reliably guides the system toward economically superior cooperative equilibria when enhancement factors are sufficient.

3.8. Hyperparameter Sensitivity Analysis

The entropy regularization coefficient ρ exhibits dominant control over cooperation emergence in the proposed framework. Conversely, other hyperparameters including learning rate α , discount factor γ , and value loss weight δ demonstrate negligible sensitivity within standard operational ranges. Experimental results across enhancement factors $r \in [1, 5]$ reveal distinct behavioral regimes governed by ρ selection.

Fig. 9 shows the sensitivity analysis of the entropy regularization coefficient ρ in TUC-PPO. In PPO, the choice of ρ critically influences the exploration-exploitation trade-off during policy optimization. As shown in Fig. 9, when ρ exceeds 0.1, a higher reward r is required to ensure all agents adopt cooperative strategies. This occurs because the variance of policy

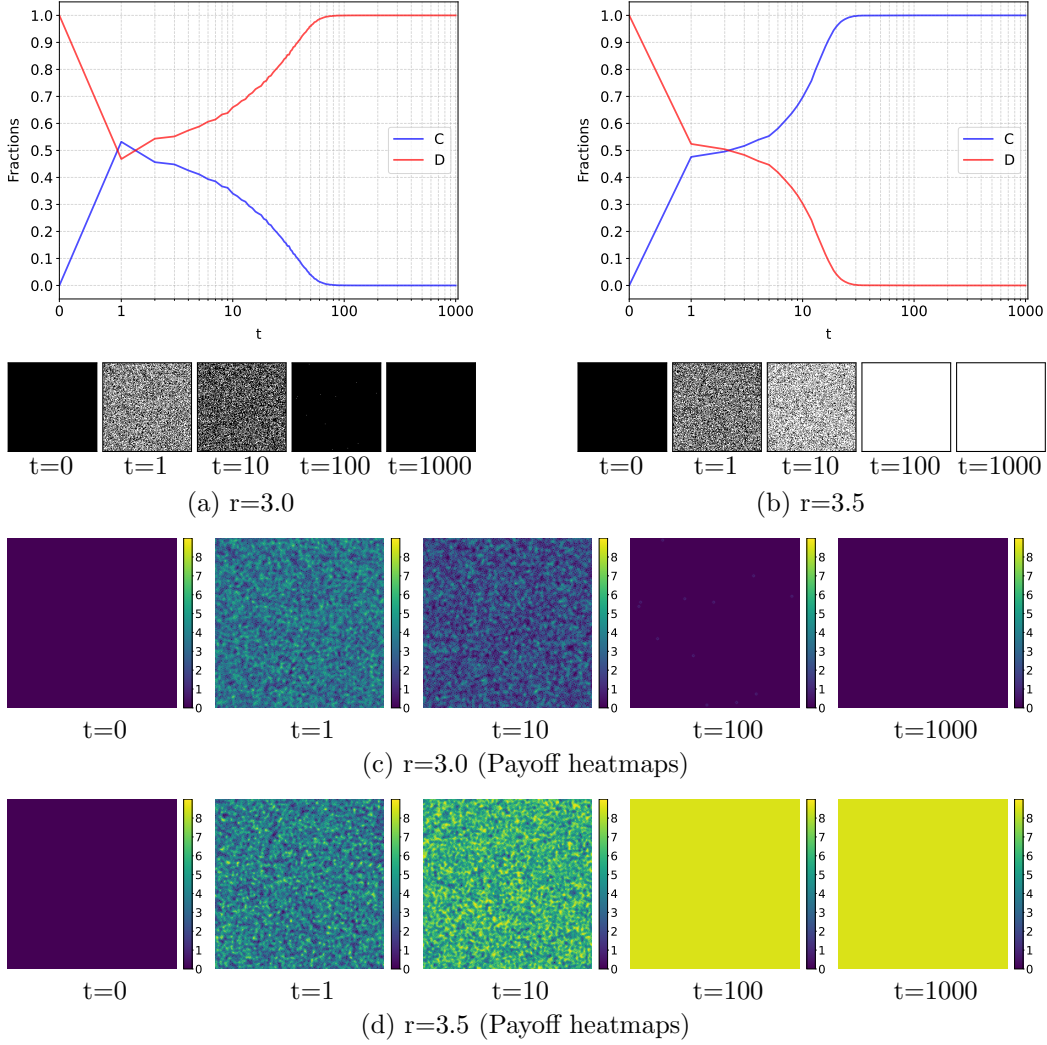


Figure 8: Experiment in SPGG using TUC-PPO with all-defectors initialization. **(a, b)** Temporal evolution curves show cooperators fractions in blue and defectors fractions in red, with corresponding strategy snapshots displaying defectors as black blocks and cooperators as white blocks. **(c, d)** Payoff heatmaps track individual reward changes during evolution. **(a, c)** Display results for enhancement factor $r = 3.0$ and **(b, d)** present $r = 3.5$ conditions. When all agents' strategies initialize as defectors, their transition to cooperation proves significantly more challenging.

updates grows nonlinearly with increasing ρ , causing the policy gradient direction to deviate from the intended optimization path. Consequently, agents struggle to converge to the deterministic optimal policy within a reasonable number of iterations. Furthermore, excessive policy entropy disrupts the stability of value function estimation, leading to distorted advantage signals and undermining the theoretical foundation of importance sampling.

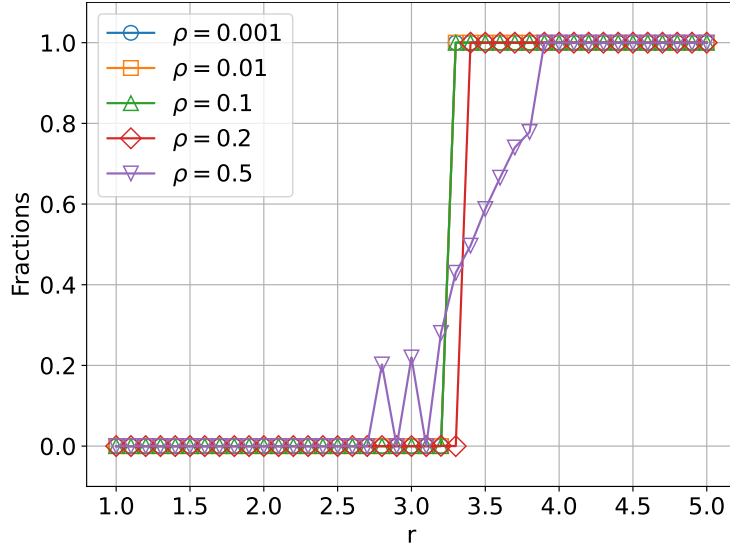


Figure 9: Impact of discount factor ρ on TUC-PPO.

TUC-PPO’s entropy regularization term effectively encourages exploration when ρ is appropriately set. However, surpassing a critical threshold introduces excessive stochasticity that violates the core assumptions of policy gradient optimization. In continuous action spaces, this manifests as non-convergent agent behaviors. For discrete decision-making scenarios, it causes the loss of deterministic policy choices at critical state nodes. Empirical studies indicate that optimal ρ values must balance task-specific state-space complexity and exploration needs. These values typically correspond to the minimal stochasticity level required for maintaining policy diversity. Based on these findings, we set the default value of ρ to 0.01. Other hyperparameters are set according to commonly used values.

4. Conclusions

The TUC-PPO framework establishes a novel paradigm for cooperation evolution in SPGG by integrating team utility constraints with proximal policy optimization. This synthesis creates a mathematically grounded approach where constrained policy updates explicitly balance individual rewards with collective welfare requirements. Through rigorous theoretical analysis and comprehensive experimental validation, we demonstrate that TUC-PPO’s team-constrained optimization mechanism delivers significant advantages in cooperative system dynamics. Key findings reveal TUC-PPO’s dual performance superiority over conventional methods. This method achieves stable cooperation at substantially lower enhancement factors ($r \geq 3.6$), significantly outperforming baseline PPO which requires $r > 5.0$. Additionally, TUC-PPO exhibits accelerated convergence rates compared to conventional approaches. Crucially, TUC-PPO maintains robust performance across diverse initialization schemes. This includes challenging all-defector scenarios where traditional methods fail, confirming the framework’s adaptability to adverse initial conditions.

Theoretical contributions encompass the first integration of team utility constraints into policy gradient optimization for evolutionary games. This integration enables precise control between individual rationality and collective welfare via Lagrangian dual-ascent. The framework employs a self-adjusting constraint mechanism that dynamically adapts penalty coefficients through batch-wise violation evaluation. This ensures team utility thresholds are met while maintaining policy update stability. This approach overcomes fundamental limitations in conventional multi-agent RL by demonstrating how explicit team welfare objectives foster cooperation emergence.

From an implementation perspective, TUC-PPO’s spatial dynamics reveal emergent self-organization patterns. In these patterns, cooperators form stable clusters that minimize boundary exposure to defectors. This spatial configuration provides algorithmic-level validation of network reciprocity theory, demonstrating how localized team constraints generate system-wide cooperative equilibria. For practical applications, TUC-PPO provides novel design principles for multi-agent systems requiring robust cooperation. Potential implementations include resource distribution networks, community-based sustainability initiatives, and institutional frameworks that balance individual incentives with group welfare.

While demonstrating significant advantages, TUC-PPO presents oppor-

tunities for further refinement. Future research should extend this framework to dynamic network topologies and investigate heterogeneous agent capabilities within team constraints. Additionally, exploration of asymmetric reward structures in collective action problems and optimization of computational efficiency for large-scale deployments are critical next steps. In summary, TUC-PPO advances evolutionary game theory by formalizing the relationship between local team constraints and global cooperation emergence. The framework combines mathematical rigor, empirical performance, and implementation flexibility. This triad establishes a new foundation for designing cooperative multi-agent systems in computational social science and distributed AI applications.

CRedit authorship contribution statement

Zhaoqilin Yang: Writing – original draft, Writing – review and editing, Validation, Methodology, Conceptualization. **Xin Wang:** Conceptualization, Investigation, Writing – review and editing. **Ruichen Zhang :** Writing – review and editing, Validation, Supervision, **Chanchan Li:** Writing – review and editing, Visualization, Software. **Youliang Tian:** Funding acquisition, Resources, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by the Natural Science Special Project (Special Post) Research Foundation of Guizhou University (No.[2024] 39); National Key Research and Development Program of China under Grant 2021YFB3101100; National Natural Science Foundation of China under Grant 62272123; Project of High-level Innovative Talents of Guizhou Province under Grant [2020]6008;

Science and Technology Program of Guizhou Province under Grant [2020]5017, [2022]065; Science and Technology Program of Guiyang under Grant [2022]2-4.

References

- [1] R. M. Dawes, R. H. Thaler, Anomalies: Cooperation, *Journal of Economic Perspectives* 2 (3) (1988) 187–197. doi:10.1257/jep.2.3.187.
- [2] M. Perc, Phase transitions in models of human cooperation, *Physics Letters A* 380 (36) (2016) 2803–2808. doi:https://doi.org/10.1016/j.physleta.2016.06.017.
- [3] M. Perc, J. J. Jordan, D. G. Rand, Z. Wang, S. Boccaletti, A. Szolnoki, Statistical physics of human cooperation, *Physics Reports* 687 (2017) 1–51, statistical physics of human cooperation. doi:https://doi.org/10.1016/j.physrep.2017.05.004.
- [4] E. Pennisi, How did cooperative behavior evolve?, *Science* 309 (5731) (2005) 93–93. doi:10.1126/science.309.5731.93.
- [5] D. Kennedy, C. Norman, What don’t we know?, *Science* 309 (5731) (2005) 75–75. doi:10.1126/science.309.5731.75.
- [6] M. A. NOWAK, R. M. MAY, The spatial dilemmas of evolution, *International Journal of Bifurcation and Chaos* 03 (01) (1993) 35–78. doi:10.1142/S0218127493000040.
- [7] M. W. Macy, A. Flache, Learning dynamics in social dilemmas, *Proceedings of the National Academy of Sciences* 99 (suppl_3) (2002) 7229–7236. doi:10.1073/pnas.092080099.
- [8] Z. Wang, S. Kokubo, M. Jusup, J. Tanimoto, Universal scaling for the dilemma strength in evolutionary games, *Physics of Life Reviews* 14 (2015) 1–30. doi:https://doi.org/10.1016/j.plrev.2015.04.033.
- [9] M. A. Nowak, R. M. May, Evolutionary games and spatial chaos, *Nature* 359 (6398) (1992) 826–829. doi:10.1038/359826a0.
- [10] C. Hauert, G. Szabó, Game theory and physics, *American Journal of Physics* 73 (5) (2005) 405–414. doi:10.1119/1.1848514.

- [11] G. Szabó, G. Fáth, Evolutionary games on graphs, *Physics Reports* 446 (4) (2007) 97–216. doi:<https://doi.org/10.1016/j.physrep.2007.04.004>.
- [12] M. Chica, R. Chiong, M. Kirley, H. Ishibuchi, A networked N -player trust game and its evolutionary dynamics, *IEEE Transactions on Evolutionary Computation* 22 (6) (2018) 866–878. doi:[10.1109/TEVC.2017.2769081](https://doi.org/10.1109/TEVC.2017.2769081).
- [13] I. S. Lim, N. Masuda, To trust or not to trust: Evolutionary dynamics of an asymmetric n -player trust game, *IEEE Transactions on Evolutionary Computation* 28 (1) (2024) 117–131. doi:[10.1109/TEVC.2023.3244537](https://doi.org/10.1109/TEVC.2023.3244537).
- [14] M. Waibel, L. Keller, D. Floreano, Genetic team composition and level of selection in the evolution of cooperation, *IEEE Transactions on Evolutionary Computation* 13 (3) (2009) 648–660. doi:[10.1109/TEVC.2008.2011741](https://doi.org/10.1109/TEVC.2008.2011741).
- [15] X. Chen, T. Sasaki, Å. Brännström, U. Dieckmann, First carrot, then stick: how the adaptive hybridization of incentives promotes cooperation, *Journal of The Royal Society Interface* 12 (102) (2015) 20140935. doi:[10.1098/rsif.2014.0935](https://doi.org/10.1098/rsif.2014.0935).
- [16] M. dos Santos, The evolution of anti-social rewarding and its countermeasures in public goods games, *Proceedings of the Royal Society B: Biological Sciences* 282 (1798) (2015) 20141994. doi:[10.1098/rspb.2014.1994](https://doi.org/10.1098/rspb.2014.1994).
- [17] I. Okada, H. Yamamoto, F. Toriumi, T. Sasaki, The effect of incentives and meta-incentives on the evolution of cooperation, *PLOS Computational Biology* 11 (5) (2015) 1–17. doi:[10.1371/journal.pcbi.1004232](https://doi.org/10.1371/journal.pcbi.1004232).
- [18] T. Ren, X.-J. Zeng, Reputation-based interaction promotes cooperation with reinforcement learning, *IEEE Transactions on Evolutionary Computation* 28 (4) (2024) 1177–1188. doi:[10.1109/TEVC.2023.3304911](https://doi.org/10.1109/TEVC.2023.3304911).
- [19] J. Li, C. Zhang, Q. Sun, Z. Chen, J. Zhang, Changing the intensity of interaction based on individual behavior in the iterated prisoner’s dilemma game, *IEEE Transactions on Evolutionary Computation* 21 (4) (2017) 506–517. doi:[10.1109/TEVC.2016.2628385](https://doi.org/10.1109/TEVC.2016.2628385).

- [20] J. Quan, Y. Zhou, X. Wang, J.-B. Yang, Information fusion based on reputation and payoff promotes cooperation in spatial public goods game, *Applied Mathematics and Computation* 368 (2020) 124805. doi:<https://doi.org/10.1016/j.amc.2019.124805>.
- [21] D. Helbing, A. Szolnoki, M. Perc, G. Szabó, Punish, but not too hard: how costly punishment spreads in the spatial public goods game, *New Journal of Physics* 12 (8) (2010) 083005. doi:10.1088/1367-2630/12/8/083005.
- [22] X. Chen, A. Szolnoki, M. Perc, Probabilistic sharing solves the problem of costly punishment, *New Journal of Physics* 16 (8) (2014) 083016. doi:10.1088/1367-2630/16/8/083016.
- [23] X. Chen, A. Szolnoki, M. c. v. Perc, Competition and cooperation among different punishing strategies in the spatial public goods game, *Phys. Rev. E* 92 (2015) 012819. doi:10.1103/PhysRevE.92.012819.
- [24] J. Liu, H. Meng, W. Wang, T. Li, Y. Yu, Synergy punishment promotes cooperation in spatial public good game, *Chaos, Solitons & Fractals* 109 (2018) 214–218. doi:<https://doi.org/10.1016/j.chaos.2018.01.019>.
- [25] H. Wei, X. Pu, J. Zhang, C. Zhang, M. Cao, Moral preferences co-evolve with cooperation in networked populations, *IEEE Transactions on Evolutionary Computation* (2024) 1–1doi:10.1109/TEVC.2024.3486572.
- [26] L. Liu, X. Chen, A. Szolnoki, Competitions between prosocial exclusions and punishments in finite populations, *Scientific Reports* 7 (1) (2017) 46634. doi:10.1038/srep46634.
- [27] A. Szolnoki, X. Chen, Alliance formation with exclusion in the spatial public goods game, *Phys. Rev. E* 95 (2017) 052316. doi:10.1103/PhysRevE.95.052316.
- [28] C. Griffin, A. Belmonte, Cyclic public goods games: Compensated coexistence among mutual cheaters stabilized by optimized penalty taxation, *Phys. Rev. E* 95 (2017) 052309. doi:10.1103/PhysRevE.95.052309.
- [29] S. Wang, L. Liu, X. Chen, Tax-based pure punishment and reward in the public goods game, *Physics Letters A* 386 (2021) 126965. doi:<https://doi.org/10.1016/j.physleta.2020.126965>.

- [30] H.-W. Lee, C. Cleveland, A. Szolnoki, Supporting punishment via taxation in a structured population, *Chaos, Solitons & Fractals* 178 (2024) 114385. doi:<https://doi.org/10.1016/j.chaos.2023.114385>.
- [31] X.-B. Cao, W.-B. Du, Z.-H. Rong, The evolutionary public goods game on scale-free networks with heterogeneous investment, *Physica A: Statistical Mechanics and its Applications* 389 (6) (2010) 1273–1280. doi:<https://doi.org/10.1016/j.physa.2009.11.044>.
- [32] G. Szabó, C. Tóke, Evolutionary prisoner’s dilemma game on a square lattice, *Phys. Rev. E* 58 (1998) 69–73. doi:[10.1103/PhysRevE.58.69](https://doi.org/10.1103/PhysRevE.58.69).
- [33] P. Schuster, K. Sigmund, Replicator dynamics, *Journal of Theoretical Biology* 100 (3) (1983) 533–538. doi:[https://doi.org/10.1016/0022-5193\(83\)90445-9](https://doi.org/10.1016/0022-5193(83)90445-9).
- [34] R. Sutton, A. Barto, Reinforcement learning: An introduction, *IEEE Transactions on Neural Networks* 9 (5) (1998) 1054–1054. doi:[10.1109/TNN.1998.712192](https://doi.org/10.1109/TNN.1998.712192).
- [35] L. R. Izquierdo, S. S. Izquierdo, N. M. Gotts, J. G. Polhill, Transient and asymptotic dynamics of reinforcement learning in games, *Games and Economic Behavior* 61 (2) (2007) 259–276. doi:<https://doi.org/10.1016/j.geb.2007.01.005>.
- [36] A. Lipowski, K. Gontarek, M. Ausloos, Statistical mechanics approach to a reinforcement learning model with memory, *Physica A: Statistical Mechanics and its Applications* 388 (9) (2009) 1849–1856. doi:<https://doi.org/10.1016/j.physa.2009.01.028>.
- [37] D. Jia, H. Guo, Z. Song, L. Shi, X. Deng, M. Perc, Z. Wang, Local and global stimuli in reinforcement learning, *New Journal of Physics* 23 (8) (2021) 083020. doi:[10.1088/1367-2630/ac170a](https://doi.org/10.1088/1367-2630/ac170a).
- [38] L. Wang, D. Jia, L. Zhang, P. Zhu, M. Perc, L. Shi, Z. Wang, Lévy noise promotes cooperation in the prisoner’s dilemma game with reinforcement learning, *Nonlinear Dynamics* 108 (2) (2022) 1837–1845. doi:[10.1007/s11071-022-07289-7](https://doi.org/10.1007/s11071-022-07289-7).

- [39] Z. Song, H. Guo, D. Jia, M. Perc, X. Li, Z. Wang, Reinforcement learning facilitates an optimal interaction intensity for cooperation, *Neurocomputing* 513 (2022) 104–113. doi:<https://doi.org/10.1016/j.neucom.2022.09.109>.
- [40] C. J. C. H. Watkins, P. Dayan, Q-learning, *Machine Learning* 8 (3) (1992) 279–292. doi:[10.1007/BF00992698](https://doi.org/10.1007/BF00992698).
- [41] O. Han, T. Ding, L. Bai, Y. He, F. Li, M. Shahidehpour, Evolutionary game based demand response bidding strategy for end-users using q-learning and compound differential evolution, *IEEE Transactions on Cloud Computing* 10 (1) (2022) 97–110. doi:[10.1109/TCC.2021.3117956](https://doi.org/10.1109/TCC.2021.3117956).
- [42] Y. Shi, Z. Rong, Analysis of q-learning like algorithms through evolutionary game dynamics, *IEEE Transactions on Circuits and Systems II: Express Briefs* 69 (5) (2022) 2463–2467. doi:[10.1109/TCSII.2022.3161655](https://doi.org/10.1109/TCSII.2022.3161655).
- [43] A. Szolnoki, M. c. v. Perc, G. Szabó, Topology-independent impact of noise on cooperation in spatial public goods games, *Phys. Rev. E* 80 (2009) 056109. doi:[10.1103/PhysRevE.80.056109](https://doi.org/10.1103/PhysRevE.80.056109).
- [44] A. Szolnoki, M. c. v. Perc, Impact of critical mass on the evolution of cooperation in spatial public goods games, *Phys. Rev. E* 81 (2010) 057101. doi:[10.1103/PhysRevE.81.057101](https://doi.org/10.1103/PhysRevE.81.057101).
- [45] G. Szabó, A. Szolnoki, Selfishness, fraternity, and other-regarding preference in spatial evolutionary games, *Journal of Theoretical Biology* 299 (2012) 81–87, evolution of Cooperation. doi:<https://doi.org/10.1016/j.jtbi.2011.03.015>.
- [46] G. Szabó, A. Szolnoki, L. Czakó, Coexistence of fraternity and egoism for spatial social dilemmas, *Journal of Theoretical Biology* 317 (2013) 126–132. doi:<https://doi.org/10.1016/j.jtbi.2012.10.014>.
- [47] Z. Yan, L. Li, J. Shang, H. Zhao, Periodic update rule with q-learning promotes evolution of cooperation in game transition with punishment mechanism, *Neurocomputing* 609 (2024) 128510. doi:<https://doi.org/10.1016/j.neucom.2024.128510>.

- [48] Y. Shen, Y. Ma, H. Kang, X. Sun, Q. Chen, Learning and propagation: Evolutionary dynamics in spatial public goods games through combined q-learning and fermi rule, *Chaos, Solitons & Fractals* 187 (2024) 115377. doi:<https://doi.org/10.1016/j.chaos.2024.115377>.
- [49] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, *CoRR* abs/1707.06347 (2017). arXiv:1707.06347.
URL <http://arxiv.org/abs/1707.06347>
- [50] Y. Sun, Y. Li, H. Li, J. Liu, X. Zhou, Intuitionistic fuzzy madm in wargame leveraging with deep reinforcement learning, *IEEE Transactions on Fuzzy Systems* 32 (9) (2024) 5033–5045. doi:10.1109/TFUZZ.2024.3435400.
- [51] C. Yu, A. Velu, E. Vinitzky, J. Gao, Y. Wang, A. Bayen, Y. WU, The surprising effectiveness of ppo in cooperative multi-agent games, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Vol. 35, Curran Associates, Inc., 2022, pp. 24611–24624.
- [52] Z. Yang, C. Li, X. Wang, Y. Tian, Ppo-act: Proximal policy optimization with adversarial curriculum transfer for spatial public goods games, *Chaos, Solitons & Fractals* 199 (2025) 116762. doi:<https://doi.org/10.1016/j.chaos.2025.116762>.
- [53] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: G. Gordon, D. Dunson, M. Dudík (Eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, Vol. 15 of *Proceedings of Machine Learning Research*, PMLR, Fort Lauderdale, FL, USA, 2011, pp. 315–323.
- [54] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization (2017). arXiv:1412.6980.
URL <https://arxiv.org/abs/1412.6980>