HIERARCHICAL MULTI-AGENT REINFORCEMENT LEARNING FOR MULTI-GROUP TAX GAME

Anonymous authors
Paper under double-blind review

ABSTRACT

Taxation is a fundamental instrument of economic policy and has long been studied through economic models. However, most existing taxation models focus on interactions within a single economic group, typically with one government and multiple households. Such a setting overlooks interactions among independent economic groups, where governments may compete for economic resources through fiscal policies. To capture these interactions, we formulate taxation as a hierarchical multi-group tax game. Within each group, a government sets tax policies while households respond, forming a hierarchical government–household interaction. Across groups, governments compete through fiscal policies, coupling the economic dynamics of distinct groups. This coupled structure poses challenges for standard multi-agent reinforcement learning (MARL) methods. To this end, we propose Multi-Group PPO (MGPPO), a bilevel MARL algorithm designed for hierarchical multi-group interactions. MGPPO incorporates Hierarchical Sampling to coordinate learning across agent levels and Curriculum Learning to improve training stability. We further develop a multi-group taxation simulation environment grounded in classical economic models, supporting the evaluation of fiscal policies under inter-group competition. Simulations show that in the two-group settings, MGPPO reduces the income and wealth Gini coefficients by 5.0% and 4.9%, respectively, and increases mean GDP by 4.1% compared with IPPO.

1 INTRODUCTION

Taxation is a central tool for influencing economic activity and the distribution of income and wealth (Ikeda, 2002; Johnson, 2014). Designing tax policies requires accounting for many participants, including governments, households, firms, and banks, with different information and objectives. Their interactions create feedback: tax policies affect household decisions, while aggregate household behavior shapes the economic conditions under which governments set policies. These differences and feedback effects make analytical policy optimization difficult (Axtell, 2000; Zhang et al., 2021).

Multi-agent reinforcement learning (MARL) offers a computational approach to studying these interactions. It allows agents with different observations and objectives to learn policies through repeated interactions in a simulated economy. For example, the AI Economist uses two-level reinforcement learning to learn tax policies alongside agents' economic behavior (Zheng et al., 2022). TaxAI provides a scalable economic simulator and a bilevel mean-field learning method for interactions between a government and many households (Mi et al., 2023) . These approaches capture how a government's tax policies and household behavior influence each other. However, their focus is a single economic group governed by one government.

In reality, economic groups also interact with one another, allowing policies in one group to affect economic outcomes in others. One form of such interaction is competition between governments, where each government must consider how rival policies affect its own economy. In this paper, we focus on competition for capital. When capital can move across groups, tax policies influence how it is allocated among them (Zodrow & Mieszkowski, 1986). Each government must therefore account for both household responses within its group and the policies of competing governments when setting tax policies.

These cross-group effects introduce dependencies that the single-group approaches discussed above do not capture. Combining government–household interactions within groups with competition among

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

governments creates a coupled learning problem. Each government must learn while both household behavior and rival governments' policies are changing. To address this problem, we introduce the Multi-Group Hierarchical Tax Game (MHTG) and an associated learning algorithm. MHTG models government–household interactions within each group (intra-group) as a leader–follower game and competition for capital across groups (inter-group) as a noncooperative game among governments. Building on this formulation, we propose Multi-Group PPO (MGPPO), a bilevel on-policy MARL algorithm for learning government and household policies across interacting groups. MGPPO uses hierarchical sampling to coordinate learning across government and household levels. It also uses a curriculum that gradually introduces inter-group competition to stabilize learning. We further develop a multi-group taxation simulation environment to evaluate the learned policies and study their economic outcomes. Our contributions are as follows:

- **Multi-group tax game formulation:** We formulate taxation as a set of government–household leader–follower games linked through competition for capital among governments.
- **Multi-group PPO algorithm:** We develop MGPPO to learn policies for governments and households across interacting groups. The algorithm combines hierarchical sampling with a curriculum designed to support learning under inter-group competition.
- **Simulation environment and evaluation:** We build a multi-group taxation environment grounded in classical economic models and evaluate policy performance, training stability, and scalability. In the two-group settings, MGPPO reduces the income and wealth Gini coefficients by 5.0% and 4.9%, and increases mean GDP by 4.1% compared with IPPO.

2 Related Work

Classic Tax Models in Economics Taxation has been widely studied in public finance. Early works primarily adopted a normative perspective where the government acts as a social planner and optimizes welfare under household responses. Classical optimal taxation theory models household labor supply and consumption as reaction functions to tax policies, typically within static or single-level optimization frameworks (Mirrlees, 1971; Atkinson & Stiglitz, 2015). While these models provide important theoretical insights, they do not explicitly account for strategic interactions among decision-making entities. Subsequent research transitioned toward game-theoretic formulations to capture the strategic interplay between governments and households, typically characterizing taxation as a leader–follower (Stackelberg) game. A significant advancement in this lineage is the integration of the Bewley–Aiyagari framework (Aiyagari, 1994; 1995), which enables modeling the responses of heterogeneous households to tax policies under idiosyncratic risk and incomplete markets. To address fiscal competition across regions or countries, tax competition has been extensively studied where multiple governments strategically set tax rates within a non-cooperative game framework (Wilson, 1999; Keen & Marchand, 1997). Seminal works, such as the Zodrow–Mieszkowski (Zodrow & Mieszkowski, 1986) model, formalize inter-jurisdictional tax competition and highlight key economic phenomena, including fiscal externalities, tax base mobility, and strategic interdependence among governments.

Game Theory in Economic Analysis Game theory provides the fundamental mathematical formalism for analyzing strategic interactions in taxation and fiscal policy models (Fudenberg & Tirole, 1991). To capture the nested nature of modern taxation systems, existing studies typically employ different game-theoretic formulations to model interactions with asymmetric and symmetric strategic power. For instance, leader–follower (Stackelberg) games (Von Stackelberg, 2011) are commonly adopted when capturing the asymmetric interactions between governments and households. Here, the government acts as the leader by setting tax policies while anticipating households' reaction functions, like adjustments in consumption and labor supply. The resulting bilevel structure provides a principled way to analyze how fiscal policies shape individual behavior (Colson et al., 2007). By contrast, interactions among multiple governments are often characterized by relatively symmetric strategic power and are modeled as non-cooperative games. Such characterization is widely used to study strategic competition among jurisdictions and to explain phenomena such as the failure of coordination and the emergence of competitive tax policies. While game theory provides powerful insights, seeking equilibrium solutions for large heterogeneous agents with high-dimensional state

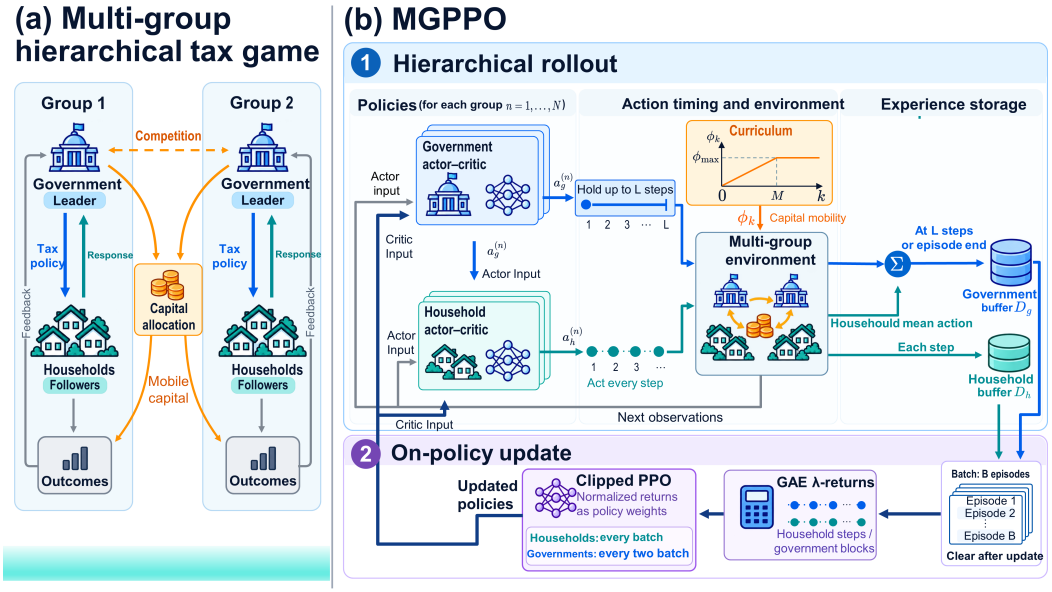


Figure 1: Framework overview. (a): Economic activities among the government, the capital allocation intermediary including firm and bank, and households. (b): MGPPO workflow. Government actions are held for up to L steps, households act at every step, and the curriculum parameter ϕ_k gradually increases inter-group coupling. Separate government and household trajectories are used for on-policy updates.

spaces becomes analytically intractable. This necessitates the application of computational methods such as MARL.

Multi-Agent Reinforcement Learning in Economic System MARL has developed into a broad family of algorithms for economic decision-making. Tax simulations, in particular, often involve strategic interactions among agents (Shi, 2023; Trott et al., 2021; Hinterlang & Tänzer, 2021; Chen et al., 2021). On-policy methods can be advantageous in such settings because they limit the mismatch between collected trajectories and current joint policies in a non-stationary game (Foerster et al., 2017). Representative on-policy MARL algorithms include IPPO De Witt et al. (2020), MAPPO (Yu et al., 2022) and HAPPO Zhong et al. (2024). However, these general-purpose algorithms do not by themselves define how agents interact within a game, leaving the interaction structure to the environment and learning design. To address this issue, some studies model these interactions explicitly or implicitly. Explicit modeling directly influences an agent's current decision, for example by including another agent's action in its policy input. Implicit modeling influences decisions through the training process, such as through rewards or value estimates, and thereby changes how the agent acts in the future. As an example of explicit game modeling, Zhang et al. Zhang et al. (2020) formulated multi-agent coordination as a bi-level Stackelberg game and proposed an actor–critic method to learn its equilibrium. Recently, frameworks such as the AI Economist (Zheng et al., 2022) and the RBC-based model (Curry et al., 2022) further leveraged reinforcement learning to capture bi-level interactions between policymakers and economic agents. Building on these efforts, TaxAI (Mi et al., 2023) extended RL-based economic simulation to large-scale heterogeneous agent settings. However, these economic simulation frameworks primarily study interactions within a single government-led system. We extend this setting from one high-level agent to multiple high-level agents. Our proposed algorithm models both intra-group dynamics and strategic competition across multiple government–household systems by combining explicit and implicit modeling.

3 Hierarchical Multi-Group Tax Game

We propose the multi-group tax game to describe the complex phenomenon in an economic society. Figure 1(a) illustrates its structure, showing hierarchical interactions among four entities: households,

governments, firms, and banks. Within each group, the government acts as the *high-level agent* and optimizes social welfare (measured by GDP and the Gini coefficient) through tax policy adjustments. Households serve as *low-level agents* and maximize individual utilities by choosing labor supply and consumption. Firms and banks act as *intermediary institutions* and enable the conversion and circulation of labor and capital. Across groups, governments compete for mobile capital to advance their own economic interests. Thus, we have two types of strategic interactions. The *intra-group interaction* models the government-household hierarchy via a leader–follower game grounded in the Bewley–Aiyagari framework. The *inter-group interaction* captures tax competition among governments through a noncooperative game based on the Zodrow–Mieszkowski model. Full mathematical details are provided in Appendices A–B.

3.1 Intra-Group Leader–Follower Game

Following the notations in Appendix A, we introduce the actions and objectives of four entities, respectively.

Households A household has two actions: saving ratio p_t and working hour h_t at time t. We also denote i_t, a_t, and c_t as the household's income, wealth, and consumption, respectively. Under the assumption of the fixed consumption tax rate equation 9, the household's consumption c_t can be represented in terms of the saving ratio p_t by $c_t = \frac{1-p_t}{1+\tau_s}\left[i_t - T_{\text{inc}}(i_t) + \frac{G_t}{N}\right]$, where $T_{\text{inc}}(\cdot)$ and $T_{\text{ast}}(\cdot)$ are income and asset tax functions. G_t denote total government transfers and N the number of households receiving those transfers. Based on the working–spending dynamics in equation 5–equation 8, each household seeks to optimize her objective function given by

$$\max_{p_t, h_t} \quad \mathbb{E}\sum_{t=0}^{T} \beta^t \left(\frac{(c_t)^{1-\theta}}{1-\theta} - \frac{(h_t)^{1+\gamma}}{1+\gamma}\right), \tag{1}$$

where $T > 0$ is the decision horizon and $\beta \in [0, 1]$ is the discounted factor. The objective captures the household's utility (or "happiness"), which increases with consumption c_t and decreases with labor effort h_t. The two terms are shaped by risk preferences, with $\theta > 0$ and $\gamma > 0$ representing the coefficients of relative risk aversion (CRRA) for consumption and disutility of labor, respectively.

When there are M households present, they are independent and heterogeneous, distinguished by their individual productivity levels equation 4 and financial states. Each household evolves according to the aforementioned working–spending dynamics. Although decisions are made independently, their aggregate behaviors jointly influence the wage rate and, consequently, the government's tax policy. For instance, the wage rate defined in equation 14 is determined by the joint labor supply of all households.

Government The government interacts with households through two adjustable policy instruments: tax policies and the spending ratio η_t. Regarding taxation, we consider three types of taxes: consumption, income, and asset taxes. Based on the tax policy selection in equation 9–equation 11, the government chooses a fixed consumption tax rate τ_s, the income tax parameter $\theta_i := (\tau_i, \xi_i)$, and the asset tax parameter $\theta_a := (\tau_a, \xi_a)$. In addition to taxation, the government determines the scale of its fiscal expenditure to support public spending, defined by $\eta_t = \frac{G_t}{Y_t}$. Here, G_t and Y_t are the government's social spending and the total production. The debt dynamics in equation 12 explicitly link the government's policy choices (captured by θ_i, θ_a, and η_t) to its fiscal state trajectory.

Based on the production function equation 13 and the debt dynamics equation 12, the government seeks to optimize a social welfare objective that emphasizes both economic efficiency and equality, which is the expected cumulative product of production and the inverse Gini index over the planning horizon:

$$\max_{\theta_i, \theta_a, \eta_t} \quad \mathbb{E}\sum_{t=0}^{T} \beta^t \log(Y_t) \cdot \Big(1 - \text{Gini}(\{i_t^j\}, \{a_t^j\})\Big) \tag{2}$$

where Y_{-1} is given as the initial baseline production output and $\text{Gini}(\{i_t^j\}, \{a_t^j\})$ denotes the joint Gini index over all households' income and asset distributions. The objective balances two competing dimensions of social welfare. The term $1 - \text{Gini} \in [0, 1]$ measures equality, attaining 1 under perfect equality and approaching 0 under maximal concentration. The term $\log(Y_t)$ captures economic efficiency via aggregate output, where the logarithm is used to reduce the scale for numerical stability.

3.2 INTER-GROUP COMPETITION GAME

We model inter-group interaction as a strategic competition among governments. This is motivated by the observation that cross-group dynamics are primarily driven by group leaders, who both interact most directly with one another and exert the greatest influence over their respective followers.

The competition is characterized by *capital regulation*. Each government can strategically set a capital tax rate $\tau_t^{(n)}$ to either attract or repel mobile capital and further influence its domestic capital stock and objective in equation 2. Let the superscript (n) denote the quantities related to the n-th group. $K_t^{(n)}, B_t^{(n)}, M^{(n)}$ represent the capital, the debt, and the household number of the n-th group at time t, respectively. Based on the law of motion for capital in Appendix B, the real capital in group n is given by

$$\begin{aligned} K_{t+1}^{(n),\text{real}} &= \underbrace{K_{t+1}^{(n)}}_{\text{Internal Accumulation}} + \underbrace{\Delta K_t^{(n),\text{flow}}}_{\text{External Competition}} \\ &= (1+r_{\text{interest}})K_t^{(n)} + (1+r_{\text{interest}})\left(B_t^n - \sum_{j=1}^{M^{(n)}} a_t^{j,(n)}\right) + \sum_{j=1}^{M^{(n)}} a_{t+1}^{j,(n)} \quad - B_{t+1}^{(n)} - \phi\left(\tau_t^{(n)} - \bar{\tau}_t\right) K_t^{(n)}, \end{aligned} \tag{3}$$

where $\bar{\tau}_t$ is the average capital tax rate across all groups at time t. All other variables are extended to the n-th group analogously and retain their original economic interpretations.

4 MULTI-GROUP BILEVEL LEARNING FRAMEWORK

4.1 ENVIRONMENT SETTING

To interface with standard MARL methods, we express the model in the form of a partially observable Markov game, defined by the tuple $\langle \mathcal{N}, \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \beta \rangle$. This representation is adopted for notational compatibility rather than as an additional modeling assumption. Here, $\mathcal{N}$ denotes the set of agents and $\mathcal{S}$ the state space. Each agent $i \in \mathcal{N}$ has observation and action spaces $\mathcal{O}_i$ and $\mathcal{A}_i$, with joint spaces $\mathcal{O} = \prod_i \mathcal{O}_i$ and $\mathcal{A} = \prod_i \mathcal{A}_i$. The reward function is $\mathcal{R} = \{R_i(s,a)\}_{i\in\mathcal{N}}$, where $R_i(s,a)$ is the reward to agent i under state s and joint action a. The transition kernel $\mathcal{P}$ governs the system dynamics, and $\beta \in [0,1]$ is the discount factor. We partition $|\mathcal{N}|$ agents to N groups. Each group consists of one government and $N^{(n)} - 1$ households, $n = 1, \ldots, N$. Since all governments share a common observation and action space, and similarly for households, we omit group indices when defining these spaces for simplicity.

Action space A government agent's action at time t is $a_{\text{gov},t} := (\theta_i, \theta_a, \eta_t, \tau_t)$, which contains tax parameters, social spending ratio, and capital tax rate. Each household j in group n takes action $a_{\text{hh},t}^j := (p_t^j, h_t^j)$, representing its saving ratio and working hours.

Observation space In practice, it is unrealistic to assume that a government can observe all individual household states. Following Mi et al. (2023), we approximate household-level information using aggregate statistics. Specifically, at time t, households are partitioned into the top 10% richest and 50% poorest, and we compute their averaged income $(\bar{i}_t^{\text{top}}, \bar{i}_t^{\text{btm}})$, averaged asset $(\bar{a}_t^{\text{top}}, \bar{a}_t^{\text{btm}})$, and averaged productivity $(\bar{e}_t^{\text{top}}, \bar{e}_t^{\text{btm}})$. These aggregates, together with the current wage rate w_t, form the core of the government's observation. Moreover, due to inter-group competition, each government also observes the previous-period government actions, denoted by $a_{\text{gov},t-1}^n$. The observation for a government in group n at time t is $o_{\text{gov},t}^n = \left(w_t^n, \bar{i}_t^{\text{top}n}, \bar{i}_t^{\text{btm},n}, \bar{a}_t^{\text{top},n}, \bar{a}_t^{\text{btm},n}, \bar{e}_t^{\text{top},n}, \bar{e}_t^{\text{btm},n}, a_{\text{gov},t-1}^n\right)$.

Households observe both their own financial state and the same aggregated group statistics (macro-level trends). For household j in group n, the observation is $o_{\text{hh},t}^{j,n} = \left(w_t^n, i_t^{j,n}, a_t^{j,n}, e_t^{j,n}, \bar{a}_t^{\text{top},n}, \bar{a}_t^{\text{btm},n}, \bar{e}_t^{\text{top},n}, \bar{e}_t^{\text{btm},n}\right)$.

Interactive dynamics At time t, within each group, the government (the leader) observes aggregate household statistics $o_{\text{gov},t}$ and sets policy parameters $(\theta_i, \theta_a, \eta_t)$ with fiscal dynamics governed

by equation 12. Households, as followers, observe their individual financial states (i_t, a_t, e_t) together with the governments action $a_{\text{gov},t}$, and choose decisions (p_t, h_t). Across groups, in addition to $o_{\text{gov},t}$, each government observes the previous-period policies of other governments $a_{\text{gov},t-1}^{(-n)}$ and adjusts its capital tax rate τ_t. These decisions influence the allocation of mobile capital via the flow rule equation 16, thereby affecting production and household dynamics in all groups. This coupling induces strategic interdependence among governments and links the evolution of otherwise separate groups.

4.2 MGPPO Algorithm Design

MGPPO algorithm consisting of inter-group and intra-group modules. We adopt a hybrid modeling strategy that combines implicit and explicit representations to characterize the underlying hybrid game structure. The implicit modeling mechanism is employed to capture the latent competitive relationships among different groups, where strategic interactions are reflected through learned representations. In contrast, the explicit modeling mechanism is utilized to represent the leader-follower relationships, where the hierarchical decision-making process is directly incorporated into the policy estimation. Moreover, such a design provides a balance between centralized training and scalable optimization. On the one hand, the proposed method preserve essential global interaction information during training, thereby improving learning stability in complex multi-agent environments. On the other hand, instead of directly concatenating all agents' states and actions, the architecture distributes agents' state and action information more evenly across the actor and critic networks by combining explicit and implicit modeling, thereby reducing the input dimensionality of each network and alleviate the optimization difficulty caused by large-scale agent interactions.

Intra-group Actor-Critic. At the intra-group level, the government and households act as the leader and followers, respectively. We capture the government-household interaction with explicit modeling and process competition among households through implicit modeling. For the household agents, their policy is conditioned on the household private observation and the government's decision, which captures the follower's response under the leader's strategy, denoted as$\pi_{\text{hh}}^{j,n} = \pi_{\theta_{\text{hh}}^n}(a_{\text{hh}}^{j,n}|o_{\text{hh}}^{j,n}, a_{\text{gov},t})$. The government agent's policy is solely conditioned on its local observation, denoted by $\pi_{\text{gov}}^n = \pi_{\theta_{\text{gov}}^n}(a_{\text{gov}}^n|o_{\text{gov}}^n)$.

The household critic estimates the expected return based on its individual state-action pair, given by $Q_{\text{hh}}^n := Q_{\phi_{\text{hh}}^n}(o_{\text{hh}}^n, \bar{a}_{\text{hh}})$. Considering the large number of household agents, it is impossible to incorporate all household actions into the government critic network. Therefore, we employ a mean-field representation to aggregate household behaviors: $\bar{a}_{\text{hh}}^n = M(a_{\text{hh}}^n) = \frac{1}{N_{\text{gov}}} \sum_{j=1}^{N_{\text{gov}}} a_{\text{hh}}^{j,n}$, where N_{gov} denotes the number of households within group n. Thus, the intra-group government critic is given by $Q_{\text{gov}}^{n,intra} := Q_{\phi_{\text{gov}}^{intra}}(o_{\text{gov}}^n, \bar{a}_{\text{hh}})$.

Inter-group Actor-Critic. Apart from the interactions within a group, the governments in different groups also compete with each other. Therefore, we introduce an additional inter-group critic $Q_{\text{gov}}^{n,inter} := Q_{\phi_{\text{gov}}^{inter}}(o_{\text{gov}}, a_{\text{gov}})$ as a complementary estimate of cross-group competition. The complete government value estimation should combine the two levels of estimation. Here, we design $Q_{\text{gov}}^n = Q_{\text{gov}}^{n,intra} + \lambda Q_{\text{gov}}^{n,inter}$ as the government's critic in the MHTG, where λ balances the relative importance between internal and external interactions.

Overall, the MGPPO decomposes high-dimensional multi-agent optimization into structured subproblems while preserving scalability and the environment's hierarchical game structure.

4.3 Training Procedure

Introducing inter-group competition induces a nested game structure and tightly couples intra- and inter-group dynamics. In particular, the capital tax rate τ influences capital allocation across groups, which further affects the wage rate w_t and household labor decisions. These household responses then alter capital accumulation and each government's competitiveness in attracting mobile capital. Such feedback loops make the environment highly non-stationary and difficult to optimize with standard MARL methods. Conventional MARL approaches struggle to capture the structured interactions

in the game, while bilevel MARL methods have difficulty simultaneously modeling intra- and inter-group dependencies.

We observe that the primary challenge arises from inter-group competition. Governments must not only maintain stable intra-group economic dynamics but also adapt strategically to competing governments. To address this, we propose a bilevel learning framework integrating two components: *hierarchical sampling* and *curriculum learning*, as illustrated in Figure 1(b). Hierarchical sampling enforces the leader–follower structure by operating governments and households at different time scales. Curriculum learning gradually increases the strength of inter-group competition, allowing policies to stabilize before adapting to strategic interactions. Full implementation details are provided in Appendix E.

Hierarchical Sampling To preserve the leader–follower structure within each group, we introduce a *hierarchical sampling* strategy that separates the decision time scales of governments and households. Households operate at a fine scale, sampling observations and acting at every time step, while the government updates its policy at a coarser interval of $n_{\text{gov}} > 1$ steps. During each government decision interval, the government policy remains fixed, allowing household behaviors to evolve under a stationary policy and reveal stable response patterns. Data are collected according to this schedule for each group. This design enables the government to better anticipate household responses and reinforces the intended leader–follower learning structure.

Curriculum Learning Curriculum learning mechanism gradually introduces inter-group competition into the intra-group leader–follower game. It controls the intensity of inter-group competition through the capital mobility parameter ϕ, defined as $\phi = \min\left(0.001 \times \text{episode},\ 1\right)$. At early training stages, ϕ is close to zero, implying limited capital mobility between groups and effectively decoupling groups, allowing the model to focus primarily on the intra-group leader–follower game. As training progresses and the intra-group strategies largely stabilize, ϕ gradually increases, strengthening inter-group competition and encouraging governments to adapt their policies in response to rivals. In implementation, we alternate between intra-group and inter-group updates. For each episode, we first suppress inter-group competition and train governments and households within each group under the leader–follower structure. Next, we fix the learned intra-group policies, set ϕ according to the curriculum, and update the inter-group policy. This alternating procedure stabilizes training while enabling gradual adaptation to competition.

5 Experiments and Analysis

We evaluate the proposed MGPPO framework through comparative experiments and a scalability analysis. The evaluation considers survival years, GDP, household consumption, and income and wealth inequality. One environment step corresponds to one year, and each episode is capped at 300 years. An episode may terminate earlier when the environment's capital, household-asset, inequality, or numerical-validity conditions are violated. The detail can be found in Appendix C.1. Survival years therefore measure how long a learned policy maintains a feasible economic trajectory within the evaluation horizon. GDP denotes average annual group output, while the Gini coefficients measure inequality among households within a group; lower Gini values indicate greater equality. These results are calculated from 20 random evaluation seeds . Values in parentheses denote the half-width of a two-sided 95% Student-t confidence interval across the 20 seed-level observations. Additionally, within-episode action dynamics are presented in Appendix D.1.

5.1 Comparative Experiments

We compare MGPPO with three on-policy multi-agent baselines: IPPO, MAPPO, and MAA2C. We examine the learned policies and the resulting economic outcomes for $N = 2$ and $N = 5$ groups.

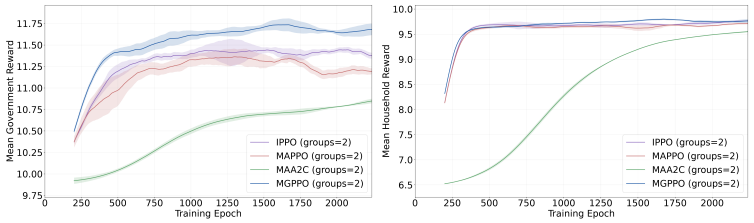
Economic-indicator and reward analysis. Table 1 compares economic outcomes and survival years across algorithms. For $N = 2$, MGPPO reaches the 300-year limit, as do IPPO and MAA2C. Compared with IPPO, it increases mean GDP by 4.1% (from 5.16×10^6 to 5.37×10^6) and reduces the mean income and wealth Gini coefficients by 5.0% and 4.9%, respectively. Its mean household consumption is 4.8% lower. MGPPO records the lowest mean Gini coefficients and highest GDP

among the four methods. For $N = 5$, MGPPO maintains a mean survival time of 300 years, compared with 297.2, 285.4, and 275.9 years for IPPO, MAPPO, and MAA2C, respectively. It also records the lowest income and wealth Gini coefficients and the highest mean household consumption. However, its mean GDP of 5.21×10^6 is the lowest of the four methods, whereas MAA2C attains the highest value of 5.64×10^6. These results reveal a trade-off between economic output, consumption, equality, and survival rather than uniform superiority across all indicators.

Metric	**IPPO**		**MAPPO**		**MAA2C**		**MGPPO**	
	$N = 2$	$N = 5$	$N = 2$	$N = 5$	$N = 2$	$N = 5$	$N = 2$	$N = 5$
Avg. consumption (10^3) ↑	**26.02(0.60)**	12.74(0.21)	9.35(0.21)	17.49(0.70)	12.03(0.25)	10.25(0.43)	24.78(0.49)	**20.16(0.24)**
Survival years ↑	**300.0(0.0)**	297.2(5.9)	297.2(5.8)	285.4(18.7)	**300.0(0.0)**	275.9(24.3)	**300.0(0.0)**	**300.0(0.0)**
Income Gini (%) ↓	27.72(0.20)	36.01(0.16)	36.07(0.20)	36.53(0.42)	35.28(0.20)	36.69(0.49)	**26.34(0.16)**	**29.36(0.14)**
Wealth Gini (%) ↓	22.37(1.35)	26.90(0.66)	26.56(0.94)	27.66(1.07)	26.28(0.99)	27.79(0.84)	**21.27(1.07)**	**21.95(0.79)**
Avg. GDP (10^6) ↑	5.16(0.13)	5.29(0.09)	5.32(0.12)	5.37(0.22)	5.35(0.12)	**5.64(0.25)**	**5.37(0.12)**	5.21(0.08)

Table 1: Economic outcomes for $N = 2$ and $N = 5$ over 20 evaluation seeds.

Figure 2 shows the reward trajectories for $N = 2$. At the final recorded training epoch, MGPPO achieves the highest mean government reward, while its mean household reward remains close to that of IPPO. This pattern is consistent with MGPPO's leader–follower design, in which government policies influence household decisions and governments learn to account for those responses when optimizing their own rewards.



(a) Government reward. (b) Household reward.

Figure 2: Mean government and household rewards during training for $N = 2$.

Action analysis. Table 2 reports the normalized actions. For $N = 2$, MGPPO selects higher income- and asset-tax parameter actions than IPPO but lower income-tax slope and spending-ratio actions. MGPPO records lower income and wealth Gini coefficients, a higher mean GDP (5.37×10^6 versus 5.16×10^6), and lower mean consumption. Because total transfers are defined as $G_t = \eta_t Y_t$, the spending ratio alone does not determine the absolute transfer amount. MGPPO also selects a higher saving-ratio action than IPPO (19.93% versus 15.53%), leaving a smaller share of current resources available for consumption.

For $N = 5$, MGPPO combines the lowest income-tax parameter action with the highest asset-tax parameter, income-tax slope, spending-ratio, and capital-tax parameter actions. It records the highest mean consumption (20.16×10^3), the lowest income and wealth Gini coefficients (29.36% and 21.95%), and a mean survival time of 300 years. Its high spending-ratio action and low saving-ratio action (24.82%) allocate more available resources to current consumption. Despite selecting the highest working-hour action (60.96%), MGPPO records the lowest mean GDP (5.21×10^6) at $N = 5$. Thus, the highest working-hour action does not correspond to the highest GDP in this setting. Output depends on both effective labor and productive capital, while capital taxation affects capital availability and inter-group flows.

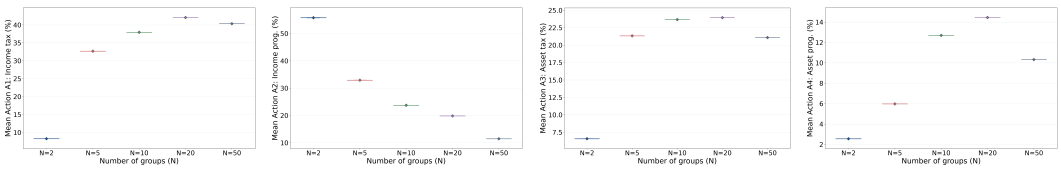
5.2 Scalability and Robustness Analysis

We examine MGPPO across different group sizes. Figure 3 presents the distributions of eight government and household actions across $N \in \{2, 5, 10, 20, 50\}$. As the number of groups increases across the smaller and intermediate settings, the income-tax parameter, asset-tax parameter, and

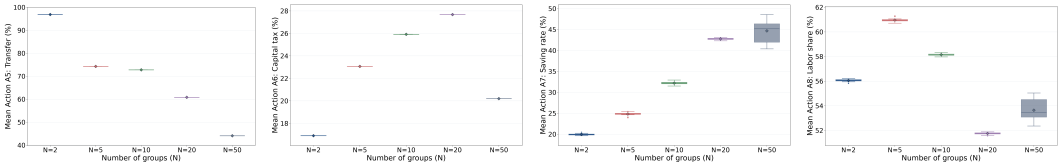
capital-tax parameter actions generally shift upward, while the income-tax slope and spending ratio actions shift downward. This pattern reflects a change in the fiscal-policy mix rather than a uniform increase in taxation or spending. The upward shift in the capital-tax parameter action does not, by itself, indicate the direction of capital flows, which depends on each group's tax rate relative to the market average. The household saving ratio action also shifts upward as the government spending ratio action declines. This pattern is consistent with a shift toward private saving as the spending ratio action falls. However, the spending ratio action does not measure the transfer received by each household, which also depends on output. Working-hour actions do not follow a monotonic trend, indicating that saving and labor decisions do not change in the same way as group size increases. At the largest group size, the government action distributions shift downward but remain concentrated, while the household saving ratio action shows greater dispersion. Overall, MGPPO adjusts its policy mix across group sizes while maintaining concentrated government action distributions, demonstrating strong stability and scalability at the level of policy actions.

Normalized action (%)	**IPPO**		**MAPPO**		**MAA2C**		**MGPPO**	
	$N=2$	$N=5$	$N=2$	$N=5$	$N=2$	$N=5$	$N=2$	$N=5$
Income-tax parameter τ_i	3.84(0.0135)	39.48(0.0010)	39.48(0.0024)	41.27(0.0090)	40.52(0.0005)	40.24(0.0009)	8.25(0.0046)	32.67(0.0018)
Income-tax slope ξ_i	63.72(0.0832)	11.36(0.0011)	11.21(0.0040)	10.76(0.0039)	12.53(0.0032)	10.41(0.0013)	55.79(0.0180)	32.92(0.0110)
Asset-tax parameter τ_a	2.41(0.0065)	19.12(0.0007)	19.55(0.0021)	19.68(0.0014)	19.80(0.0004)	20.15(0.0008)	6.57(0.0024)	21.33(0.0011)
Asset-tax slope ξ_a	4.60(0.0033)	10.27(0.0004)	9.87(0.0029)	9.77(0.0011)	10.34(0.0003)	10.01(0.0001)	2.56(0.0011)	5.98(0.0009)
Spending ratio η_t	98.11(0.0064)	47.94(0.0051)	48.74(0.0242)	44.41(0.0181)	51.24(0.0116)	41.78(0.0071)	96.84(0.0028)	74.25(0.0090)
Capital-tax parameter $\tau_{k,t}^{(n)}$	16.01(0.0035)	20.52(0.0005)	19.52(0.0010)	21.61(0.0049)	20.18(0.0003)	19.96(0.0003)	16.90(0.0015)	23.06(0.0017)
Saving ratio p_t	15.53(0.0218)	55.62(0.0716)	67.91(0.1048)	38.97(0.1498)	57.88(0.0392)	65.31(0.0349)	19.93(0.0948)	24.82(0.1931)
Working-hour action h_t	59.04(0.0809)	52.71(0.0330)	49.95(0.0616)	58.62(0.0673)	53.65(0.0196)	52.17(0.0206)	56.06(0.0488)	60.96(0.0753)

Table 2: Group-averaged government and household actions for $N = 2$ and $N = 5$.



(a) Income-tax parameter. (b) Income-tax slope. (c) Asset-tax parameter. (d) Asset-tax slope.



(e) Spending ratio. (f) Capital-tax parameter. (g) Saving ratio. (h) Working-hour action.

Figure 3: MGPPO's government and household actions across group counts $N \in \{2, 5, 10, 20, 50\}$.

6 Conclusion

In this study, we developed a multi-group hierarchical tax game environment and proposed MGPPO to capture government–household interactions and competition among governments. The proposed method integrates *hierarchical sampling* with *curriculum learning* to optimize policies in this setting. In the two-group setting, MGPPO achieves a 4.1% higher mean GDP than IPPO while reducing the income and wealth Gini coefficients by 5.0% and 4.9%, respectively. In the five-group setting, MGPPO reaches the 300-year evaluation horizon in all 20 runs and records the lowest inequality and highest household consumption among the compared methods. For $N \in \{2, 5, 10, 20\}$, the scalability experiments demonstrate that MGPPO maintains stable policy performance across different group sizes, confirming its robustness and scalability. Overall, the results reveal a trade-off between distributional outcomes and aggregate output. Future work will investigate the underlying fiscal mechanisms and strengthen economic robustness as the numbers of groups and households increase.

GENERATIVE AI USAGE DISCLOSURE

In this work, generative AI tools were used to assist with translation, improve the clarity and readability of author-written text, and debug author-written code. They were not used to generate data, develop theoretical or conceptual frameworks, formulate mathematical claims or proofs, propose hypotheses, design experiments or methodology, implement the proposed methods, analyze data, or interpret results. All AI-assisted text and code were reviewed, revised, and tested by the authors. We take full responsibility for the final content of this work.

REFERENCES

S. Rao Aiyagari. Uninsured idiosyncratic risk and aggregate saving*. The Quarterly Journal of Economics, 109(3):659–684, 08 1994. ISSN 0033-5533. doi: 10.2307/2118417. URL https://doi.org/10.2307/2118417.

S. Rao Aiyagari. Optimal capital income taxation with incomplete markets, borrowing constraints, and constant discounting. Journal of Political Economy, 103(6):1158–1175, 1995. doi: 10.1086/601445. URL https://doi.org/10.1086/601445.

Anthony B Atkinson and Joseph E Stiglitz. Lectures on public economics: Updated edition. Princeton University Press, 2015.

Robert Axtell. Why agents?: on the varied motivations for agent computing in the social sciences, volume 17. Center on Social and Economic Dynamics Washington, DC, 2000.

Roland Benabou. Tax and education policy in a heterogeneous-agent economy: What levels of redistribution maximize growth and efficiency? Econometrica, 70(2):481–517, 2002.

Christophe Chamley. Optimal taxation of capital income in general equilibrium with infinite lives. Econometrica: Journal of the Econometric Society, pp. 607–622, 1986.

Mingli Chen, Andreas Joseph, Michael Kumhof, Xinlei Pan, and Xuan Zhou. Deep reinforcement learning in a monetary model. arXiv preprint arXiv:2104.09368, 2021.

Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. Annals of operations research, 153(1):23–56, 2007.

Michael Curry, Alexander Trott, Soham Phade, Yu Bai, and Stephan Zheng. Analyzing micro-founded general equilibrium models with many agents using deep reinforcement learning. arXiv preprint arXiv:2201.01163, 2022.

Christian Schroeder De Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the starcraft multi-agent challenge? arXiv preprint arXiv:2011.09533, 2020.

Jakob Foerster, Nantas Nardelli, Gregory Farquhar, Triantafyllos Afouras, Philip HS Torr, Pushmeet Kohli, and Shimon Whiteson. Stabilising experience replay for deep multi-agent reinforcement learning. In International conference on machine learning, pp. 1146–1155. PMLR, 2017.

Drew Fudenberg and Jean Tirole. Game theory. MIT press, 1991.

Jonathan Heathcote, Kjetil Storesletten, and Giovanni L Violante. Optimal tax progressivity: An analytical framework. The Quarterly Journal of Economics, 132(4):1693–1754, 2017.

Natascha Hinterlang and Alina Tänzer. Optimal monetary policy using reinforcement learning. 2021.

Sanford Ikeda. Dynamics of the mixed economy: Toward a theory of interventionism. Routledge, 2002.

Norman Johnson. Mixed economies welfare. Routledge, 2014.

Michael Keen and Maurice Marchand. Fiscal competition and the pattern of public spending. Journal of Public Economics, 66(1):33–53, 1997. ISSN 0047-2727. doi: https://doi.org/10.1016/S0047-2727(97)00035-2. URL https://www.sciencedirect.com/science/article/pii/S0047272797000352.

Qirui Mi, Siyu Xia, Yan Song, Haifeng Zhang, Shenghao Zhu, and Jun Wang. Taxai: A dynamic economic simulator and benchmark for multi-agent reinforcement learning. arXiv preprint arXiv:2309.16307, 2023.

James A Mirrlees. An exploration in the theory of optimum income taxation. The review of economic studies, 38(2):175–208, 1971.

Rui Aruhan Shi. Deep reinforcement learning and macroeconomic modelling. PhD thesis, University of Warwick, 2023.

Alexander Trott, Sunil Srinivasa, Douwe van der Wal, Sebastien Haneuse, and Stephan Zheng. Building a foundation for data-driven, interpretable, and robust policy design using the ai economist. arXiv preprint arXiv:2108.02904, 2021.

Heinrich Von Stackelberg. Market Structure and Equilibrium. Springer Science & Business Media, 2011. First published in 1934, this is the translation.

John Douglas Wilson. Theories of tax competition. National tax journal, 52(2):269–304, 1999.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. Advances in neural information processing systems, 35:24611–24624, 2022.

Haifeng Zhang, Weizhe Chen, Zeren Huang, Minne Li, Yaodong Yang, Weinan Zhang, and Jun Wang. Bi-level actor-critic for multi-agent coordination. In Proceedings of the AAAI conference on artificial intelligence, volume 34, pp. 7325–7332, 2020.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. Handbook of reinforcement learning and control, pp. 321–384, 2021.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C Parkes, and Richard Socher. The ai economist: Taxation policy design via two-level deep multiagent reinforcement learning. Science advances, 8(18):eabk2607, 2022.

Yifan Zhong, Jakub Grudzien Kuba, Xidong Feng, Siyi Hu, Jiaming Ji, and Yaodong Yang. Heterogeneous-agent reinforcement learning. Journal of Machine Learning Research, 25(32):1–67, 2024.

George R. Zodrow and Peter Mieszkowski. Pigou, tiebout, property taxation, and the underprovision of local public goods. Journal of Urban Economics, 19(3):356–370, 1986. ISSN 0094-1190. doi: https://doi.org/10.1016/0094-1190(86)90048-3. URL https://www.sciencedirect.com/science/article/pii/0094119086900483.

A BEWLEY–AIYAGARI MODEL

The Bewley–Aiyagari model describes a minimal economic circulation system that includes four entities: household, government, firm, and financial intermediary.

Households In the economic system, a household get paid by working and spend money by consumption, forming a working–spending dynamics that powers economic circulation. We define i_t, a_t, and c_t as the income, wealth, and consumption of a household at time t, which represents the financial state of a household. We also denote h_t and w_t as the working hour and the wage rate, respectively. To capture heterogeneous characters of different households, we introduce productivity level e_t, which denotes the household's working efficiency. We assume e_t follows an AR(1) process, given by

$$\log e_{t+1} = \rho_e \log e_t + \sigma_e u_t, \tag{4}$$

where $u_t \sim \mathcal{N}(0,1)$ denotes a standard normal shock, and $\rho_e \geq 0, \sigma_e \geq 0$ are the persistence and the volatility of the shock. Let r_{interest} be the interest rate. A household's income in time t can be given by

$$i_t = w_t h_t e_t + r_{\text{interest}} a_t, \tag{5}$$

We consider three types of taxation policies in the model, which are income, asset, and consumption taxes, denoted by $T_{\text{inc}}(i_t)$, $T_{\text{ast}}(a_t)$, and $T_{\text{cons}}(c_t)$, respectively. Additionally, let G_t denote total government transfers and N the number of households receiving those transfers. Note that these tax policies are adjustable functions of a household's financial status. Based on the income, consumption, and tax, we can derive the wealth accumulation relation as

$$a_{t+1} = i_t - T_{\text{inc}}(i_t) + a_t - T_{\text{ast}}(a_t) - c_t - T_{\text{cons}}(c_t) + \frac{G_t}{N}. \tag{6}$$

Let p_t denote the household's saving ratio to simplify the complex household's working–spending dynamics in equation 4-equation 6 and training, which is given by

$$p_t = \frac{a_{t+1}}{i_t - T_{\text{inc}}(i_t) + \underline{a_t - T_{\text{ast}}(a_t)} + \frac{G_t}{N}}. \tag{7}$$

This simplifies the wealth evolution equation 6 to

$$a_{t+1} = a_t - T_{\text{ast}}(a_t) + p_t \left(i_t - T_{\text{inc}}(i_t) + \frac{G_t}{N} \right). \tag{8}$$

Government A government is responsible for taxing, social spending, and capital regulation. For taxation, we consider three types of tax policies: consumption, income, and asset taxes. In practice, the consumption tax can be modeled with a fixed rate τ_s Chamley (1986), i.e.,

$$T_{\text{cons}}(c_t) = \tau_s c_t. \tag{9}$$

The income tax $T_{\text{inc}}(i_t)$ and the asset tax $T_{\text{ast}}(a_t)$ are widely adopted by the nonlinear HSV Benabou (2002); Heathcote et al. (2017), given by

$$T_{\text{inc}}(i_t) = i_t - \frac{1-\tau_i}{1-\xi_i}(i_t)^{1-\xi_i}, \tag{10}$$

$$T_{\text{ast}}(a_t) = a_t - \frac{1-\tau_a}{1-\xi_a}(a_t)^{1-\xi_a}. \tag{11}$$

Here, τ_i and ξ_i control the average level and the slope of the marginal income tax; similarly, τ_a and ξ_a determine the average level and the slope of the asset tax.

In its role of social spending, the government may issue debt into the economic system and must service this debt by paying interest at a rate r_{interest}. Let G_t and B_t denote the government's social spending and debt at time t, respectively, which together characterize its fiscal state. Assuming there are M households paying taxes. With a slight abuse of notation, we use superscript to index households and denote i_t^j, a_t^j, c_t^j as the j-th household's income, wealth, and consumption at time t. The government's debt evolves according to

$$B_{t+1} = (1 + r_{\text{interest}}) B_t + G_t - \sum_{j=1}^{M} \left(T_{\text{inc}}(i_t^j) + T_{\text{ast}}(a_t^j) + T_{\text{cons}}(c_t^j) \right). \tag{12}$$

When introducing the spending ratio $\eta_t = \frac{G_t}{Y_t}$, the social spending G_t can be represented by $\eta_t Y_t = \eta_t K_t^{\alpha} L_t^{1-\alpha}$ using Cobb-Douglas production function equation 13, which can be used to further simplify the debt dynamics equation 12.

Capital regulation arises in multi-group settings, where governments compete for mobile capital to enhance their economic welfare. This inter-group strategic behavior is captured through tax competition dynamics and is formally modeled in Appendix B.

Firm A firm abstracts all production units in the economic circulation. It hires households to conduct production and pays wages to households. We use the widely adopted Cobb-Douglas production function to summarize the production process:

$$Y_t = K_t^{\alpha} L_t^{1-\alpha} \tag{13}$$

where K_t and L_t are capital and labors used for production and $\alpha \in (0, 1)$ is the capital elasticity. Suppose there are M households participating the production, we have $L_t = \sum_{j=1}^{M} h_t^j e_t^j$, representing the total effective labor supply. Using the assumption that the firm takes the marginal income from labor as the household's wage rate w_t (i.e., the market-clearing condition), we obtain

$$w_t = \frac{\partial Y_t}{\partial L_t} = (1-\alpha)\left(\frac{K_t}{L_t}\right)^{\alpha} = (1-\alpha)\frac{K_t^{\alpha}}{\left(\sum_{j=1}^{M} e_t^j h_t^j\right)^{\alpha}}. \tag{14}$$

It shows that all households face the same wage rate, which is endogenously determined by their collective labor supply and the available capital stock K_t.

Financial intermediary A financial intermediary abstracts institutions, such as banks, that collect households' savings and allocate these funds to productive capital and government bonds. This allocation determines the market's available capital for production in the subsequent period. Assuming M households deposit their savings into the intermediary, the evolution of aggregate capital and government debt satisfies

$$K_{t+1} + B_{t+1} - \sum_{j=1}^{M} a_{t+1}^j = (1 + r_{\text{interest}})K_t + (1 + r_{\text{interest}})\left(B_t - \sum_{j=1}^{M} a_t^j\right). \tag{15}$$

Here, K_t denotes the capital stock held by the intermediary at time t, which corresponds to the total capital available to firms for production in equation 13. B_t represents the government debt, and a_t^j denotes the asset of household j at time t. This equation balances the intermediary's portfolio across periods and accounts for returns on capital and interest payments on government debt.

B ZODROW–MIESZKOWSKI DYNAMICS

In a multi-group setting with N governments, capital is treated as an imperfectly mobile factor of production. Investors observe the effective capital tax rates across all regions and reallocate capital to maximize after-tax returns. This process implies that capital tends to migrate from high-tax jurisdictions to low-tax ones, a phenomenon often referred to as the "Race to the Bottom".

We use superscript to denote the group index and let $K_t^{(n)}$ denote the capital stock in group n at time t. The dynamics of capital flow are governed by the deviation of the local capital tax rate $\tau_t^{(n)}$ from the average tax rate of all groups $\bar{\tau}_t$. The law of motion for capital, adjusted for these flows, is formulated as:

$$\Delta K_t^{(n),\text{flow}} = -\phi_e \left(\tau_{k,t}^{(n)} - \bar{\tau}_{k,t}\right) K_t^{(n)}, \tag{16}$$

where $\bar{\tau}_{k,t} = \frac{\sum_{m=1}^{N} \tau_{k,t}^{(m)} K_t^{(m)}}{\sum_{m=1}^{N} K_t^{(m)}}$. represents the market average tax rate. It is a classic inter-jurisdictional capital reallocation from the Zodrow–Mieszkowski tax competition model, widely adopted in the fiscal competition literature. Here, $\bar{\tau}_t = \frac{1}{N}\sum_{n=1}^{N} \tau_t^{(n)}$ represents the market average tax rate. It captures the standard "Race to the Bottom" phenomenon: capital flows out of high-tax jurisdictions

and into low-tax ones proportionally to the tax differential $(\tau_t^{(n)} - \bar{\tau}_t)$ and the current capital stock $K_t^{(n)}$. The parameter $\phi > 0$ controls the intensity of capital mobility, and is treated as a fixed constant in standard models.

Incorporating this dynamic into the standard accumulation process, the aggregate capital for jurisdiction i at the next time step is updated by:

$$K_{t+1}^{(n),\text{real}} = K_{t+1}^{(n)} + \Delta K_t^{(n),\text{flow}}, \tag{17}$$

where $K_{t+1}^{(n),\text{real}}$ represents the real available capital accumulation. This mechanism forces governments to strategically lower tax rates to attract or retain the tax base, often at the expense of public good provision or fiscal balance.

C Simulation Settings

C.1 Termination Conditions

An episode ends when any of the following conditions is met:

- The simulation reaches its 300-step limit.
- Any group's next-period capital is non-positive.
- Any household's next-period assets become negative.
- A non-finite reward occurs.
- From step 21 onward, any group's wealth Gini coefficient exceeds 0.85 or its income Gini coefficient exceeds 0.469.

The step limit is treated as a time-limit termination, rather than an economic failure.

C.2 Hyperparameters

Tables 3 and 4 summarize the default simulation settings and the training parameters of the four implemented algorithms. The number of groups and households can vary across experiments.

Table 3: Default simulation and shared training settings.

Hyperparameter	Value	Hyperparameter	Value
Number of groups	2	Households per group	100
Episode limit (steps)	300	Years per step	1
Government action interval	10	Inequality checks begin	Step 21
Wealth Gini threshold	0.85	Income Gini threshold	0.469
Episodes per update	5	Evaluation episodes	20
Evaluation interval (updates)	10	Hidden units per layer	256
Activation	Tanh	Optimizer	Adam
Actor learning rate	3×10^{-6}	Critic learning rate	3×10^{-6}
Government discount factor	0.995	Household discount factor	0.995
Government GAE parameter	0.95	Household GAE parameter	0.99
Entropy coefficient	0.01	Maximum gradient norm	0.5

C.3 Simulation Variables

We summarize key variables in the economic simulation in Table 5.

Table 4: Algorithm-specific training parameters.

Hyperparameter	IPPO	MAPPO	MAA2C	MGPPO
Update method	PPO	PPO	A2C	PPO
Policy-update passes	10	10	1	10
Minibatch size (steps)	64	64	Full batch	64
PPO clipping parameter	0.2	0.2	–	0.2
KL threshold	0.01	0.01	–	0.01
Inter-group critic weight	–	–	–	1.0

Table 5: Key variables in the economic simulation environment

Variable	Meaning	Variable	Meaning
w_t	Wage rate	e_t	Individual labor productivity levels
h_t	Working hours	a_t	Household's asset (wealth)
a_0	Household's initial asset (wealth)	r_{interest}	Interest rate
i_t	Household's income	N	Number of households
ρ_e	Persistence	u_t	Standard normal shocks
β	Discount factor	θ	Coefficient of relative risk aversion (CRRA)
γ	Inverse Frisch elasticity	τ_s	Consumption tax rate
$T_{\text{inc}}(i_t)$	Income taxes	$T_{\text{ast}}(a_t)$	Asset taxes
τ_i, τ_a	Average level of marginal income and asset tax	ξ_i, ξ_a	Slope of marginal income and asset tax schedule
τ_k	Capital tax	Y_t	Firm's output
K_t	Capital for production	L_t	Labor for production
α	Capital elasticity	G_t	Government social spending
G_t	Government spending	B_t	Government debt

D Additional Experiment Results

D.1 Within-Episode Action Dynamics

To examine how trained policies respond to changing economic states, we evaluated each algorithm over a single episode in the two-group environment. Figure 4 presents the resulting action trajectories. Government actions form plateaus because each choice remains fixed until the next government decision point, whereas households update their actions at every step.

Under MGPPO, the household saving action declined early in the episode before stabilizing, whereas the working-hour action changed little. The baseline algorithms followed distinct saving trajectories: IPPO remained comparatively stable, MAPPO increased, and MAA2C declined more gradually. Thus, MGPPO showed a larger within-episode adjustment in saving than in labor. In the household budget equation, saving allocates current resources between consumption and next-period assets, whereas working hours affect both labor income and effort costs.

The two MGPPO governments followed distinct policy trajectories, particularly for the income-tax slope and capital-tax actions. Their policies remained differentiated as economic states evolved. Differences in capital-tax actions represent one channel of modeled competition because capital flows depend on relative tax rates. Government actions under MAPPO and MAA2C varied less over the episode.

E Algorithm Implementation

Algorithm 1 summarizes the training procedure of the proposed multi-group bilevel learning framework. Training combines two-timescale action sampling, a capital-mobility curriculum, and on-policy updates.

At each environment step, households sample actions conditioned on their group-level and private observations and the current joint government action. Each government samples a new action once

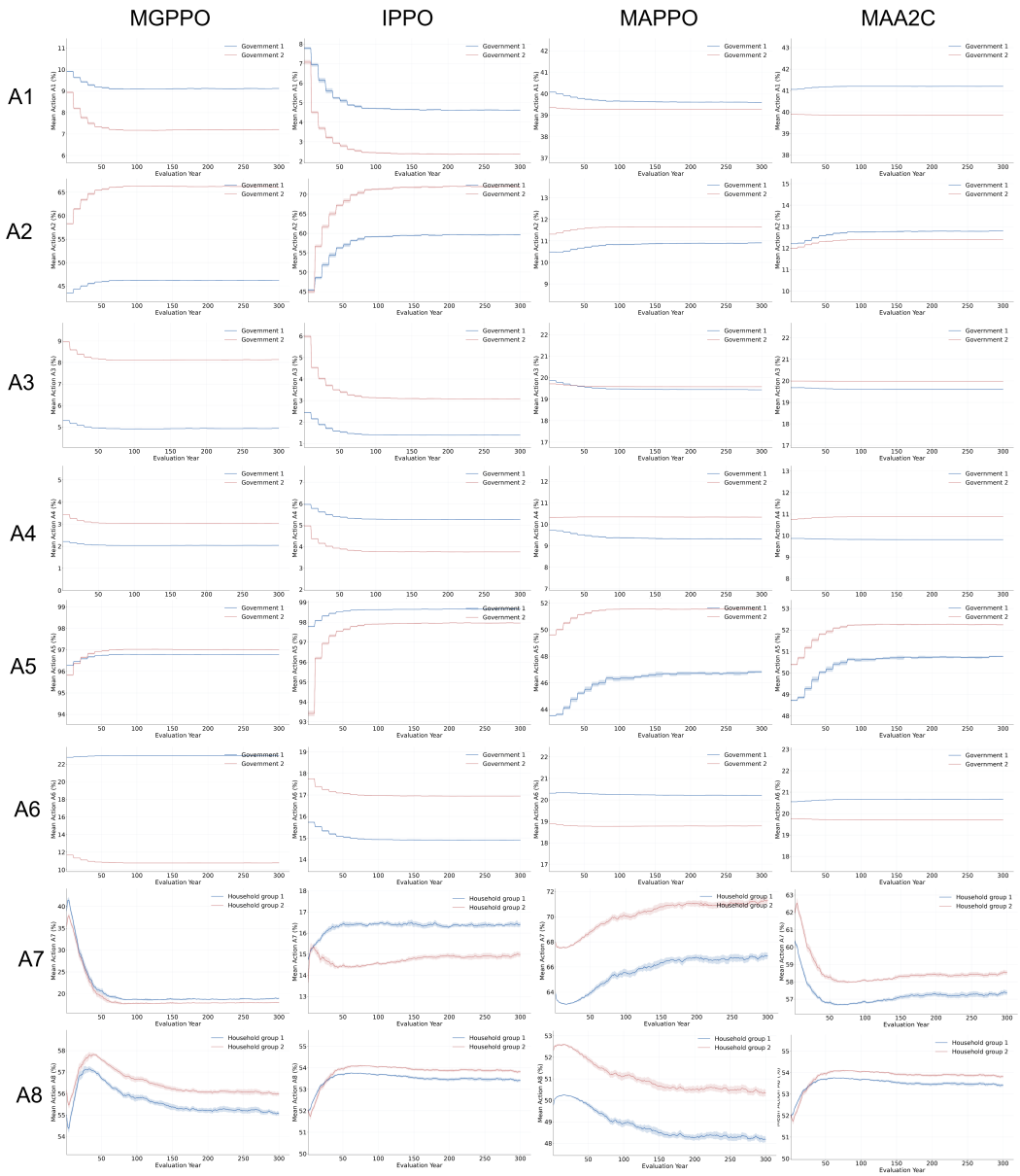


Figure 4: Within-episode action trajectories of trained IPPO, MAPPO, MAA2C, and MGPPO policies. A1–A6 denote the income-tax parameter, income-tax slope, asset-tax parameter, asset-tax slope, spending-ratio action, and capital-tax parameter, respectively. A7 and A8 denote the household saving-ratio and working-hour actions.

every L steps and holds it fixed between sampling points. During the first W episodes, government actions are selected deterministically while household policies continue to train.

At the beginning of episode k, the curriculum sets the capital-mobility weight ϕ_k, which controls the strength of inter-group capital flows during trajectory collection. Household transitions are recorded at every environment step, whereas each government transition summarizes one action block using its discounted reward and the mean household action over that block.

After collecting B episodes, the algorithm computes GAE-based λ-returns at the environment-step scale for households and at the action-block scale for governments. For each group, the critic is trained on these returns, while the actor is updated with clipped PPO using normalized returns as

Algorithm 1 Multi-Group Proximal Policy Optimization (MGPPO)

Require: Number of groups N, training episodes K, maximum episode length T, government action interval L, episodes per update B, warm-up length W, curriculum horizon M, maximum capital-mobility weight $\phi_{\max}$, and government discount factor γ_g.

Initialize a government actor–critic and a household actor–critic for each group.
Initialize on-policy buffers $\mathcal{D}_g, \mathcal{D}_h \leftarrow \varnothing$ and buffered episode count $m \leftarrow 0$.
for $k = 0, \ldots, K-1$ **do**
  Reset the environment and obtain initial observations.
  Set the previous joint government action $\mathbf{a}^g_{\text{prev}} \leftarrow \mathbf{0}$ and block length $d \leftarrow 0$.
  Set $\phi_k \leftarrow \phi_{\max} \min(1, k/M)$ if $M > 0$; otherwise set $\phi_k \leftarrow \phi_{\max}$.
  for $t = 0, \ldots, T-1$ **do**
    if $d = 0$ **then**
      Sample government actions using current observations and $\mathbf{a}^g_{\text{prev}}$; select actions deterministically if $k < W$.
      Set block reward $\mathbf{R}^g \leftarrow \mathbf{0}$ and household-action sum $\mathbf{S}^h \leftarrow \mathbf{0}$.
    end if
    Sample household actions conditioned on group-level and private observations and the current joint government action.
    Compute the mean household action $\bar{\mathbf{a}}^h_t$ for each group.
    Execute the actions with capital-mobility weight ϕ_k and store household transitions in $\mathcal{D}_h$.
    Set $\mathbf{R}^g \leftarrow \mathbf{R}^g + \gamma_g^d \mathbf{r}^g_t$, $\mathbf{S}^h \leftarrow \mathbf{S}^h + \bar{\mathbf{a}}^h_t$, and $d \leftarrow d + 1$.
    if $d = L$ **or** the episode ends **or** $t = T-1$ **then**
      Store a government transition in $\mathcal{D}_g$ with block reward $\mathbf{R}^g$, mean household action $\mathbf{S}^h/d$, and discount γ_g^d.
      Set $\mathbf{a}^g_{\text{prev}}$ to the current joint government action and $d \leftarrow 0$.
    end if
    if the episode ends **then**
      break
    end if
    Advance to the next observations.
  end for
  Set $m \leftarrow m + 1$.
  if $m = B$ **then**
    Compute GAE-based λ-returns for household steps and government action blocks.
    for group $n = 1, \ldots, N$ **do**
      Update household actor–critic n with clipped PPO, using normalized returns as policy weights.
      if $k - B + 1 \geq W$ **then**
        Update government actor–critic n in the same manner.
      end if
    end for
    Clear $\mathcal{D}_g, \mathcal{D}_h$ and set $m \leftarrow 0$.
  end if
end for

policy weights. Household networks are updated throughout training; government networks are updated only after the warm-up phase. The on-policy buffers are then cleared.