

# TaxAI: A Dynamic Economic Simulator and Benchmark for Multi-Agent Reinforcement Learning

Qirui Mi

Institute of Automation, CAS  
School of Artificial Intelligence, UCAS  
Beijing, China  
miqirui2021@ia.ac.cn

Siyu Xia

Institute of Automation, CAS  
School of Artificial Intelligence, UCAS  
Beijing, China  
xiasiyu2023@ia.ac.cn

Yan Song

Institute of Automation, CAS  
Beijing, China  
yan.song@ia.ac.cn

Haifeng Zhang\*

Institute of Automation, CAS  
School of Artificial Intelligence, UCAS  
Nanjing Artificial Intelligence  
Research of IA  
Beijing, China  
haifeng.zhang@ia.ac.cn

Shenghao Zhu

University of International  
Business and Economics  
Beijing, China  
zhushenghao@yahoo.com

Jun Wang

University College London  
London, United Kingdom  
jun.wang@cs.ucl.ac.uk

## ABSTRACT

Taxation and government spending are crucial tools for governments to promote economic growth and maintain social equity. However, the difficulty in accurately predicting the dynamic strategies of diverse self-interested households presents a challenge for governments to implement effective tax policies. Given its proficiency in modeling other agents in partially observable environments and adaptively learning to find optimal policies, Multi-Agent Reinforcement Learning (MARL) is highly suitable for solving dynamic games between the government and numerous households. Although MARL shows more potential than traditional methods such as the genetic algorithm and dynamic programming, there is a lack of large-scale multi-agent reinforcement learning economic simulators. Therefore, we propose a MARL environment, named **TaxAI**, for dynamic games involving  $N$  households, government, firms, and financial intermediaries based on the Bewley-Aiyagari economic model. Our study benchmarks 2 traditional economic methods with 7 MARL methods on TaxAI, demonstrating the effectiveness and superiority of MARL algorithms. Moreover, TaxAI's scalability in simulating dynamic interactions between the government and 10,000 households, coupled with real-data calibration, grants it a substantial improvement in scale and reality over existing simulators. Therefore, TaxAI is the most realistic economic simulator for optimal tax policy, which aims to generate feasible recommendations for governments and individuals.

## KEYWORDS

Multi-agent reinforcement learning; optimal tax policy; dynamic economic simulator; benchmark; tax evasion behavior

\*Corresponding to Haifeng Zhang (haifeng.zhang@ia.ac.cn).



This work is licensed under a Creative Commons Attribution International 4.0 License.

## ACM Reference Format:

Qirui Mi, Siyu Xia, Yan Song, Haifeng Zhang\*, Shenghao Zhu, and Jun Wang. 2024. TaxAI: A Dynamic Economic Simulator and Benchmark for Multi-Agent Reinforcement Learning. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024)*, Auckland, New Zealand, May 6 – 10, 2024, IFAAMAS, 18 pages.

## 1 INTRODUCTION

The invisible hand [69, 70] of the market is not omnipotent, and in reality, all countries rely on government intervention to promote economic development and maintain social fairness. The extent of government intervention varies from country to country, such as a free market economy [8, 73], planned economies [57, 61] or mixed economies [44, 45, 58]. However, determining the optimal government intervention degree is challenging for several reasons. **Firstly**, extracting relevant and actionable information from the complex society is arduous. **Secondly**, governments face difficulties in effectively modeling a vast and heterogeneous population with diverse preferences and characteristics. **Lastly**, the behavioral response of individuals to incentives remains highly unpredictable.

In this intricate matter of government intervention, we opt to investigate a crucial and efficacious tool, **tax policy**, which is commonly studied using agent-based modeling (ABM) [13, 23] in economics. ABM is an effective approach to simulate individual behaviors and show the relationship between micro-level decisions and macro-level phenomena. However, traditional ABM suffers from simplicity and subjectivity in setting model parameters and behavior rules, making it difficult to simulate realistic scenarios [13, 37, 51]. While **Multi-Agent Reinforcement Learning (MARL)** surpasses traditional ABM settings by offering optimal actions based on evolving state information. Some MARL algorithms perform well in partially observable environments and adaptively learn to reach equilibrium solutions [33, 59, 72]. Hence, MARL is well-suited for addressing dynamic game problems involving the government and a large population. However, despite the significant advantages of MARL, there is currently a shortage of large-scale **MARL** economic simulators designed specifically for the study of tax policies. In the existing literature, the AI Economist [84] and

**Table 1: A comparison of MARL simulators for optimal taxation problems.**

Simulator	AI Economist	RBC Model	<b>TaxAI</b> (ours)
Households' Maximum Number	10	100	10000
Tax Schedule	Non-linear	Linear	Non-linear
Tax Type	Income	Income	Income& Wealth& Consumption
Social Roles' Types	2	3	4
Saving Strategy	✗	✓	✓
Heterogenous Agent	✓	✓	✓
Real-data Calibration	✗	✗	✓
Open source	✓	✗	✓
MARL Benchmark	✗	✗	✓

the RBC model [21] emerge as the most closely related simulators to TaxAI. However, these models exhibit certain limitations, notably a **partial grounding in economic theory, limited scalability in simulating a significant number of agents, and the absence of calibration using real-world data** (as detailed in Table 1). Therefore, proposing a more realistic MARL environment to study optimal tax policies and solve dynamic games between the government and the population holds significant research and practical value.

Therefore, we introduce a dynamic economic simulator, TaxAI, based on the Bewley-Aiyagari economic model [1, 2], which is widely used to study capital market frictions, wealth distribution, economic growth issues. By incorporating the Bewley-Aiyagari model, TaxAI benefits from robust theoretical foundations in economics and models a broader range of social roles (shown in Figure 1). Based on TaxAI, we benchmark 2 economic methods and 7 MARL algorithms, optimizing fiscal policy for the government, working and saving strategies for heterogeneous households. In our experiments, we compared 9 baselines across four distinct tasks, evaluating them from both macroeconomic and microeconomic perspectives. Our results reveal the tax-avoidance behavior of MARL-based households and the varying saving and working strategies among households with different levels of wealth. Finally, we test the TaxAI environment using 9 baselines with households' number ranging from 10, 100, 1000, and even up to 10,000, demonstrating its capability to simulate large-scale agents. In summary, our work encompasses the following three contributions:

**1. A dynamic economic simulator TaxAI.** The simulator incorporates multiple roles and main economic activities, employs real-data calibration, and facilitates simulations of up to 10,000 agents. These features provide a more comprehensive and realistic simulation than existing simulators.

**2. Validation of MARL feasibility in optimizing tax policies.** We implemented 2 traditional economic approaches and 7 MARL methods to solve optimal taxation for the social planner, and optimal saving and working strategies for households. The results obtained through MARL methods surpass those achieved by traditional methods.

**3. Economic analysis of different policies.** We conducted assessments from both macroeconomic and microeconomic perspectives, uncovering tax-avoidance behaviors among MARL-based households in their pursuit of maximum utility. Furthermore, we observed distinct strategies among households with differing levels of wealth.

Codes for the TaxAI simulator and algorithms are shown in the GitHub repository <https://github.com/jidai/TaxAI>.

## 2 RELATED WORKS

**Classic Tax Models.** Economic models provide powerful tools for modeling economic activities and explaining economic phenomena. In microeconomics, the Supply and Demand model [35, 68] reveals the mechanism behind market price formation, while the marginal utility theory [46] underscores the significance of consumption decisions. In macroeconomics, the Keynesian Aggregate Demand-Aggregate Supply Model [6, 31] addresses short-term fluctuations and policy effects [30]. The Comparative Advantage Theory [20, 43] in international trade explains collaborations across nations. The Quantity Theory of Money [34, 55] investigates the relationship between money supply and price levels. Regarding the optimal tax problem, the Ramsey-Cass-Koopmans (RCK) model [16, 48] studies the consumption and savings decisions of representative agents but ignoring individual heterogeneity. The Diamond-Mirrlees model [26, 27] considers the role of taxes and labor supply in social welfare but overlooks income and asset taxes. The Overlapping Generations (OLG) model [64] emphasizes intergenerational inheritance and resource transfers [25, 36]. In contrast, the Bewley-Aiyagari model [3, 11] can assess the impact of taxation on growth, wealth distribution and welfare while simulating real-world income disparities and risk-bearing capacity of individuals. This makes the Bewley-Aiyagari model an ideal choice for studying optimal taxation and household strategies.

**Traditional Economic Methods.** The optimal tax policy and wealth distribution [10] have been extensively studied in economics. Existing works [3, 18] have utilized mathematical programming methods to address decision-making processes related to governments and households [5, 12]. However, these approaches oversimplify decision-makers rationality and fail to consider autonomous learning abilities and environmental uncertainties. In contrast, dynamic programming-based approaches [28, 32] consider long-term consequences and environmental dynamics but struggle to model non-rational behaviors [15]. Alternative approaches, such as empirical rules [40, 53] and Agent-Based Modeling (ABM) [71, 77], have emerged to address these limitations. ABM enables the exploration of micro-level behaviors and their impact on macro-level phenomena, showing in the ASPEN model [7], income distribution [29] and transaction development [47]. Despite the abundance of research in economics based on ABM, this approach often involves relatively simplistic and subjective specifications of individual behavior, making it challenging to investigate the dynamic optimization of individual strategies.

**MARL and Simulators for Economy.** MARL aims to address issues of cooperation [14] and competition [83] among multiple decision-makers [74]. The simplest MARL method Independent Learning [56, 76], including IPPO [24], and IDDPG [52], involves

each agent learning and making decisions independently, disregarding the presence of other agents [75]. In the Centralized Training Decentralized Execution (CTDE) algorithms, like MADDPG [54], QMIX [62], and MAPPO [81], agents share information during training but make decentralized decisions during execution to enhance collaborative performance. To address significant computational and communication overhead posed by a growing number of agents [80], Mean-Field Multi-Agent Reinforcement Learning (MF-MARL) [86] simplifies the problem by treating homogeneous agents as distributed particles [38]. On the other hand, Heterogeneous Agent Reinforcement Learning (HARL) [85], HAPPO and HATRPO, is designed to achieve effective cooperation in a general setting involving heterogeneous agents.

Currently, there are not many efforts employing MARL methods to determine optimal tax policies and individual strategies. The closest works to our paper include AI economist [84] and the RBC model [21]. While they both account for fundamental economic activities, **they lack large-scale agent simulation, real-data calibration, and MARL benchmarks**. These limitations make practical implementation challenging, which is why we introduce TaxAI. Besides, prior research has already explored reinforcement learning-based approaches in some subproblems. For instance, in addressing optimal savings and consumption problems, some studies [4, 63, 67] have utilized single-agent RL to model the representative agent or a continuum of agents. Meanwhile, others have employed MARL to solve rational expectations equilibrium [41, 50], optimal asset allocation and savings strategies [60]. Regarding optimal government intervention problems, research has explored the application of RL in investigating optimal monetary policy [19, 42], market prices [22, 66], international trade [65], redistribution systems [49, 79], and the cooperative relationship between central and local governments under COVID-19 [78].

### 3 BEWLEY-AIYAGARI MODEL

In comparison to classical economic models in Section 2, we contend that the Bewley-Aiyagari model serves as the most suitable theoretical foundation for investigating optimal tax policy for the government and optimal savings and labor strategies for households. In this section, we provide a brief overview of the Bewley-Aiyagari model. This model encompasses four key societal roles: N households, a representative firm, a financial intermediary, and the government. The interactions among them are illustrated in Figure 1. All model variables and their corresponding symbols are organized in Appendix Table 8. More details on model assumptions and dynamics are shown in Appendix A.3, A.4.

#### 3.1 N Households

To avoid differences in age, gender, and personality, individuals are modeled as households, whose main activities include production, consumption, saving, and tax payments. At timestep  $t$ , households' income  $i_t$  is derived from two sources. They **work** in the technology firm for the labor income  $W_t e_t h_t$ , which depends on the wage rate  $W_t$ , the individual labor productivity levels  $e_t$  and the working hours  $h_t$ . On the other hand, households can only engage in **savings** and are assumed not to borrow ( $a_t \geq 0$ ). They earn interest income  $r_{t-1} a_t$  from savings, which depends on household asset  $a_t$  and the

return to savings  $r_{t-1}$ .

$$i_t = W_t e_t h_t + r_{t-1} a_t \quad (1)$$

In this model,  $N$  households are heterogeneous in terms of labor productivity levels  $e_t$  and initial asset  $a_0$ , and we model  $e_t$  as either being in a super-star ability or a normal state. In the normal state, it follows an AR(1) process [12], a model commonly used for analyzing and forecasting time series data.

$$\log e_t = \rho_e \log e_{t-1} + \sigma_e u_t \quad (2)$$

where  $\rho_e$  is the persistence and  $\sigma_e$  is the volatility of the standard normal shocks  $u_t$ . In the super-star state, the labor market ability is  $\bar{e}$  times higher than the average. The transition of households from the normal state to the super-star state occurs with a constant probability  $p$  while remaining in the super-star state has a constant probability  $q$ .

In addition, each household seeks to maximize lifetime utility (3) depends on **consumption**  $c_t$  and **working hours**  $h_t$ , subject to budget constraint, where  $\beta$  is the discount factor,  $\theta$  is the coefficient of relative risk aversion (CRRA), and  $\gamma$  represents the inverse Frisch elasticity.  $T_N$  denotes the maximum steps.

$$\begin{aligned} \max & \mathbb{E}_0 \sum_{t=0}^{T_N} \beta^t \left( \frac{c_t^{1-\theta}}{1-\theta} - \frac{h_t^{1+\gamma}}{1+\gamma} \right) \\ \text{s.t. } & (1+\tau_s)c_t + a_{t+1} = i_t - T(i_t) + a_t - T^a(a_t) \end{aligned} \quad (3)$$

The households are required to **pay taxes** to the government, including consumption taxes  $\tau_s$ , income taxes  $T(i_t)$ , and asset taxes  $T^a(a_t)$ , the last two are expressed by a nonlinear HSV tax function [9, 40],

$$T(i_t) = i_t - (1-\tau) \frac{i_t^{1-\xi}}{1-\xi}, \quad T^a(a_t) = a_t - \frac{1-\tau_a}{1-\xi_a} a_t^{1-\xi_a} \quad (4)$$

where  $\tau, \tau_a$  determine the average level of the marginal income and asset tax, and  $\xi, \xi_a$  determine the slope of the marginal income and asset tax schedule. It presents a free market economy when all taxes are equal to 0.

#### 3.2 Technology Firm

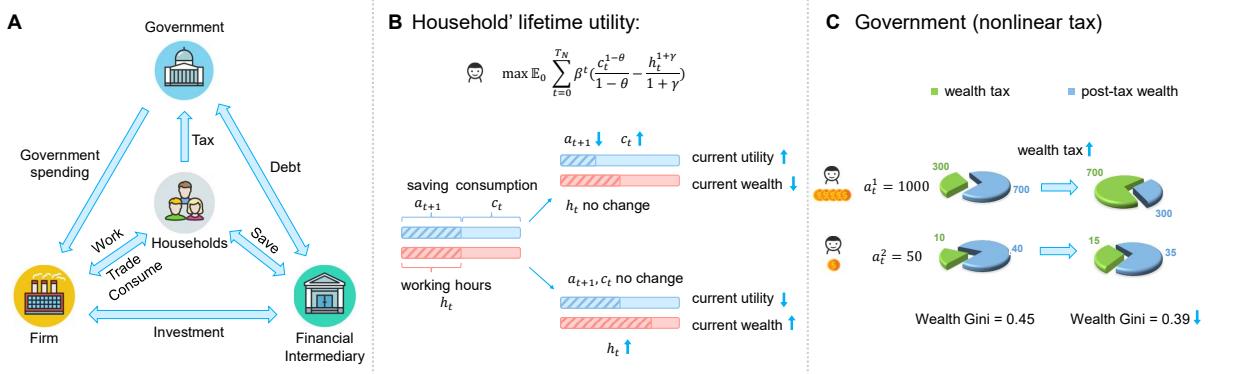
As the representative of all firms and industries, the firm converts capital and labor into goods and services. We assume it produces a homogeneous good with technology, which can meet the consumption need of households, following the Cobb-Douglas production function,

$$Y_t = K_t^\alpha L_t^{1-\alpha} \quad (5)$$

where  $K_t$  and  $L_t$  are capital and labor used for production,  $\alpha$  is capital elasticity, and we normalize the output price to 1. The firm rents capital at a rental rate  $R_t$  and hires labor at a wage rate  $W_t$ . The produced output is used for all households' gross consumption  $C_t$ , government spending  $G_t$ , and physical capital investment  $X_t = K_{t+1} - (1-\delta)K_t$ , with the depreciation rate  $\delta$ , so the aggregate resource constraint is

$$Y_t = C_t + X_t + G_t \quad (6)$$

Suppose the firm takes the marginal income from labor as households' wage rate  $W_t$  and the marginal income from capital as the



**Figure 1: Model Dynamics in the Bewley-Aiyagari Model.** A: Economic activities among the government, the firm, the financial intermediary, and households. B: The influence of households' saving and labor strategies on current utility and wealth. Households must strike a balance between consumption and savings, as well as work and leisure, to optimize lifetime utility. Increasing consumption enhances current utility but reduces current wealth, affecting future utility. Longer working hours yield higher labor income, thereby increasing wealth, but simultaneously result in disutility. C: The effect of government taxation on households' wealth. The social planner employs a nonlinear taxation, applying varying tax rates based on different assets. As tax rates rise, the taxes paid by households increase, with wealthier household contributing more. This narrows the gap in households' post-tax wealth, leading to a reduction in the Gini coefficient for wealth distribution.

rental rate  $R_t$ .

$$W_t = \frac{\partial Y_t}{\partial L_t} = (1-\alpha) \left( \frac{K_t}{L_t} \right)^\alpha, \quad R_t = \frac{\partial Y_t}{\partial K_t} = \alpha \left( \frac{K_t}{L_t} \right)^{\alpha-1} \quad (7)$$

Market clearing on labor and goods is an important assumption for simplification, which means there is an equilibrium between supply and demand. The goods market clears by Walras' Law, and the labor market clearing condition is  $L_t = \sum_i^N e_t^i h_t^i$ .

### 3.3 Government

The government has multiple goals, such as promoting economic growth, maintaining social fairness and stability, and maximizing social welfare. To optimize these objectives, the government typically employs three tools, including government spending  $G_t$ , taxation  $T_t$ , and debt  $B_t$  with the interest rate  $r_{t-1}$ . For instance, when maximizing economic growth, the government's objective and the budget constraint are as follows:

$$\begin{aligned} \max \quad J &= \mathbb{E}_0 \sum_{t=0}^{T_N} \beta^t \left( \frac{Y_t - Y_{t-1}}{Y_{t-1}} \right) \\ \text{s.t. } (1+r_{t-1}) B_t + G_t &= B_{t+1} + T_t \end{aligned} \quad (8)$$

where taxes  $T_t$  are derived from personal income taxes, wealth taxes, and consumption taxes.

$$T_t = \sum_i^N \left( T(i_t^i) + T(a_t^i) + \tau_s c_t^i \right) \quad (9)$$

In addition to the task of maximizing social welfare, the government also has the objectives of maximizing economic growth, optimizing social equity, and multi-objective optimization. The corresponding mathematical objective functions are shown in the Appendix A.2.

### 3.4 Financial Intermediary

We posit a financial intermediary where households can deposit their savings and the intermediary uses these funds to purchase capital and government bonds. Its budget constraint is defined as:

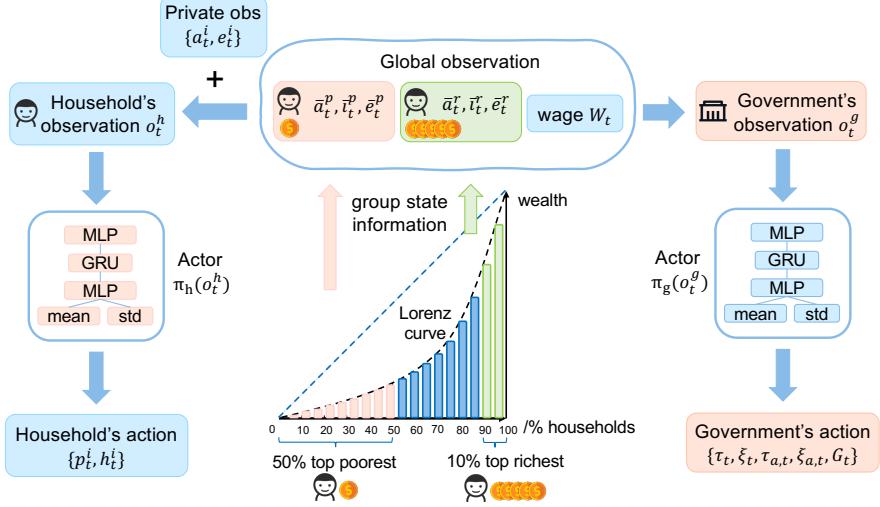
$$K_{t+1} + B_{t+1} - A_{t+1} = (R_t + 1 - \delta) K_t + (1 + r_{t-1}) (B_t - A_t) \quad (10)$$

where  $A_t$  are the gross deposits from the households. No-arbitrage implies that  $R_{t+1} = r_t + \delta$ .

## 4 TAXAI SIMULATOR

In this section, we model the above problem of optimizing tax policies for the government and developing saving and working strategies for households as multiplayer general-sum Partially Observable Markov Games (POMGs). In the POMGs  $\langle \mathcal{N}, \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ .  $\mathcal{N} = \{1, \dots, N\}$  denotes the set of all agents,  $\mathcal{S}$  represents the state space,  $\mathcal{O}^i$  denotes the observation space for agent  $i$ , and  $\mathcal{O} := \mathcal{O}^1 \times \dots \times \mathcal{O}^N$ .  $\mathcal{A}^i$  signifies the action space for agent  $i$ , and  $\mathcal{A} := \mathcal{A}^1 \times \dots \times \mathcal{A}^N$ .  $\mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \Omega(\mathcal{S})$  denotes the transition probability from state  $s \in \mathcal{S}$  to next state  $s' \in \mathcal{S}$  for any joint action  $a \in \mathcal{A}$  over the state space  $\Omega(\mathcal{S})$ . The reward function  $\mathcal{R} := \{R^i\}_{i \in \mathcal{N}}$ , here  $R^i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$  denotes the reward function of the agent  $i$  for a transition from  $(s, a)$  to  $s'$ . The discount factor  $\gamma \in [0, 1]$  keeps constant across time. The specific details of POMGs is shown in the Figure 2 and following paragraphs.

**Observation Space  $\mathcal{O}$ .** In the real world, households can observe their own asset  $a_t$  and productivity ability  $e_t$ , and acquire statistical data about the population from the news. While the government can collect data from all households and access current market prices  $W_t$ . However, the presence of a large number of heterogeneous households results in a considerably high-dimensional state space. To mitigate the dimensionality challenge, TaxAI categorizes households based on their wealth into two groups [17]: the top 10% richest and



**Figure 2: The Markov game between the government and household agents.** In the center of the figure, we display the Lorenz curves of households' wealth distribution. The global observation consists of the average assets  $\bar{a}_t$ , income  $\bar{i}_t$ , and productivity level  $\bar{e}_t$  of the 50% poorest households and 10% richest households, along with the wage rate  $W_t$ . For the government agent, it observes the global observation and takes tax and spending actions  $\{\tau_t, \xi_t, \tau_{a,t}, \xi_{a,t}, r_t^G\}$  through the actor-network. For household agents, they observe both global and private observation, including personal assets  $\{a_t^i\}$  and productivity level  $\{e_t^i\}$ , and generate savings and workings actions  $\{p_t^i, h_t^i\}$  through the actor-network. The actor-network structure in the figure is just an example.

the bottom 50% poorest households. The average wealth  $\{\bar{a}_t^r, \bar{a}_t^p\}$ , income  $\{\bar{i}_t^r, \bar{i}_t^p\}$ , and labor productivity levels  $\{\bar{e}_t^r, \bar{e}_t^p\}$  of these two groups are incorporated into the global observation. Therefore, the government's observation space  $O_g = \{W_t, \bar{a}_t^r, \bar{i}_t^r, \bar{e}_t^r, \bar{a}_t^p, \bar{i}_t^p, \bar{e}_t^p\}$ , while the household agent  $i$  can observe the global and its private information  $O_h^i = \{W_t, \bar{a}_t^r, \bar{i}_t^r, \bar{e}_t^r, \bar{a}_t^p, \bar{i}_t^p, \bar{e}_t^p, a_t^i, e_t^i\}$ ,  $i \in \{1, \dots, N\}$ . Moreover, the initialization of state information is calibrated by statistical data from the 2022 Survey of Consumer Finances (SCF)<sup>1</sup>.

**Action Space  $\mathcal{A}$ .** The decision-making of household and government agents needs to adhere to budget constraints at every step; however, the abundance of constraints often renders many MARL algorithms ineffective. Therefore, we have introduced proportional actions to alleviate these constraints. For household agents, the optimization of savings  $a_{t+1}$  and consumption  $c_t$  has been transformed into optimizing savings ratio  $p_t \in (0, 1)$  and working time  $h_t \in [0, 1] \cdot h_{max}$ , where  $h_{max}$  is calibrated by the real wealth-to-income ratio data.

$$p_t = \frac{a_{t+1}}{i_t - T(i_t) + a_t - T^a(a_t)} \quad (11)$$

For the government agent, the fiscal policy tools include optimizing tax parameters  $\{\tau_t, \xi_t, \tau_{a,t}, \xi_{a,t}\}$ , and the ratio of government spending to GDP  $r_t^G = G_t/Y_t$ . Thus, the action space of the government  $\mathcal{A}_g$  is  $\{\tau_t, \xi_t, \tau_{a,t}, \xi_{a,t}, r_t^G\}$ , while the action space of each household  $\mathcal{A}_h^i$  is  $\{p_t^i, h_t^i\}$ .

**Reward function  $\mathcal{R}$ .** The reward function for each household is denoted as:

$$r_{h,t}(s_t, a_{h,t}^i) = \frac{c_t^{1-\theta}}{1-\theta} - \frac{h_t^{1+\gamma}}{1+\gamma} \quad (12)$$

On the other hand, the government's objectives are more diverse, and we have defined four distinct experimental tasks within TaxAI: (1) Maximizing GDP growth rate. (2) Minimizing social inequality. (3) Maximizing social welfare. (4) Optimizing multiple tasks. For more details about reward function see Appendix A.2. For example, the government's reward function when maximizing GDP growth rate is denoted as:

$$r_{g,t}(s_t, a_{g,t}) = \frac{Y_t - Y_{t-1}}{Y_{t-1}} \quad (13)$$

In summary, we outline three key improvements made in constructing the **TaxAI** simulator: (1) To bridge the gap between economic models and the real world, we opt to calibrate TaxAI using 2022 SCF data. (2) To mitigate the curse of dimensionality associated with high-dimensional state information, we draw inspiration from the World Inequality Report 2022 [17] and employ grouped statistical averages for households as a representation of this high-dimensional state information. (3) In response to the abundance of constraints, we introduce the concept of proportional actions, facilitating control over the range of actions to adhere to these constraints. More details about environment setting are shown in Appendix A, including model assumptions, terminal conditions, parameters setting, and timing tests in Appendix B.

## 5 EXPERIMENTS

This section will begin by introducing 9 baseline algorithms 5.1, followed by conducting the following three sub-experiments: **Firstly** 5.2,

<sup>1</sup><https://www.federalreserve.gov/econres/scfindex.htm>

we aim to illustrate the superior performance of MARL algorithms over traditional methods from both macroeconomic and microeconomic perspectives. In the **second** part 5.3, we conduct an economic analysis of the optimization process for government and heterogeneous household strategies. **Lastly** 5.4, we assess the scalability of TaxAI by comparing the simulation results for different numbers of households, specifically N=10, 100, 1000, and even 10000. Additional results on full training curves, economic evolution, and hyperparameters, are shown in Appendix E.

## 5.1 Baselines

We compare 9 different baselines, including traditional economic methods and 4 distinct MARL algorithms, providing a comprehensive MARL benchmark for large-scale heterogeneous multi-agent dynamic games in a tax revenue context. Additional experimental settings are shown in the Appendix E.1.

- (1) Traditional Economic Methods: Free Market Policy, Genetic Algorithm (GA) [32].
- (2) Independent Learning: Independent PPO [84].
- (3) Centralized Training Distributed Execution: MADDPG [54], MAPPO [82], both with parameter sharing.
- (4) Heterogeneous-Agent Reinforcement Learning: HAPPO, HATRPO, HAA2C [85].
- (5) Mean Field Multi-Agent Reinforcement Learning: Bi-level Mean Field Actor-Critic (BMFAC), shown in Appendix D.

## 5.2 Comparative Analysis of Multiple Baselines

We benchmark 9 baselines on 4 distinct tasks, with the training curves of macroeconomic indicators in 3 tasks shown in Figure 3 and test results shown in Table 2 (Figure 8 in Appendix E presents the training curves for 6 macro-indicators in 4 tasks). In Figure 3, each row represents a task, including maximizing GDP, minimizing inequality, and maximizing social welfare. Each column represents a macroeconomic indicator, where longer years indicate longer economic stability, higher GDP represents a higher level of economic development, and a lower wealth Gini coefficient indicates fairer wealth distribution. The X-axis of each subplot represents training steps.

*Macroeconomic Perspectives.* From Figure 3, it can be observed that in each macro-indicator, most MARL algorithms outperform traditional economic methods. In the GDP optimization task, HATRPO achieves the highest per capita GDP, while BMFAC performs best in the tasks of minimizing inequality and maximizing social welfare. Different algorithms also differ in terms of convergence solutions. MADDPG excels in optimizing GDP but at the cost of reducing social welfare for higher GDP. The BMFAC algorithm excels in optimizing social welfare and the Gini coefficient. HARL algorithms, including HAPPO, HATRPO, and HAA2C, can simultaneously optimize all four macroeconomic indicators, but while achieving the highest GDP, social welfare is not maximized. On the other hand, MAPPO excels in optimizing social welfare.

*Microeconomic Perspectives.* During the testing phase, we conduct experiments on households following random, GA, and MADDPG policies within the same environment at each step. We utilize 10 distinct random seeds to simulate an economic society spanning

300 timesteps. In Figure 4, these subplots present various microeconomic indicators, including the average tax revenue, average utility, average labor supply, and average consumption for all households at each time step. The random policy represents a strategy unaffected by government tax policies, while the GA policy represents a conventional economics approach. We observe that households under the MADDPG strategy pay the lowest taxes, indicating **tax evasion behavior**, while simultaneously achieving utility levels significantly surpassing those of the GA and random policies. Labor supply and consumption are statistical measures of household microbehavior. We find that MADDPG-based households tend to opt for low consumption and reduced labor supply strategies.

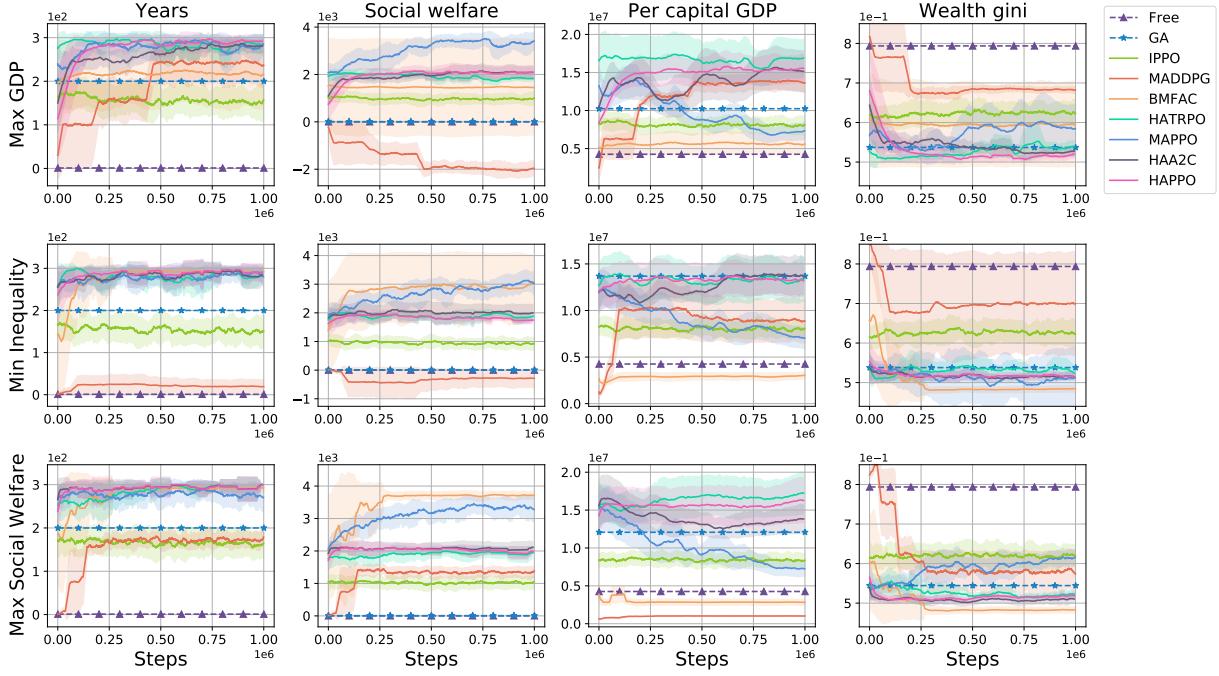
**Table 2: Test results for 9 baselines on 5 economic indicators under maximizing social welfare task (N = 100 households).**

Baselines	Years	Average Social Welfare	Per Capita GDP	Wealth Gini	Income Gini
Free market	1.0	2.9	4.3e6	0.79	<b>0.39</b>
GA	200.0	6.9	1.2e7	0.54	0.52
IPPO	162.7	1035.5	8.4e6	0.62	0.44
MADDPG	204.2	1344.6	1.0e6	0.61	0.58
MAPPO	274.5	3334.7	7.3e6	0.61	0.65
HAPPO	298.7	1986.0	1.6e7	0.52	0.54
HATRPO	<b>300.0</b>	1945.0	1.7e7	0.52	0.54
HAA2C	<b>300.0</b>	2113.3	1.4e7	0.51	0.53
BMFAC	292.8	3722.2	2.8e6	<b>0.48</b>	0.50

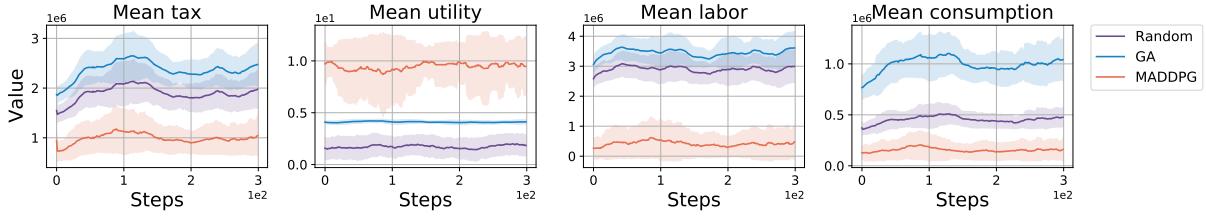
## 5.3 Economic Analysis of MARL Policy

Figure 5 illustrates the training curves of the MADDPG algorithm, aiming to maximize GDP on TaxAI. We utilize the experimental results to analyze the coordination of government actions (tax) and households' actions (labor and consumption) and their impact on economic indicators. The X-axis represents steps, while the Y-axis represents various economic indicators. In the subplots for income tax, wealth tax, total tax, labor, consumption, and households' utility, we categorize households into three groups based on wealth: the wealthy (top 10% richest), the middle class (top 10-50% richest), and the poor (50-100% top richest). We display the average indicators for these three groups as well as the average for all households (purple line), other subplots represent macroeconomic indicators (blue line). In each subplot, the solid line represents the mean of experimental results under different seeds, and the shaded area represents the standard deviation. The results of the testing phase are presented in Table 3. The following are two intriguing findings:

1. *MADDPG converges towards the highest GDP while compromising social welfare under maximizing GDP task.* As observed from Figure 5, during the initial 2e5 steps, the government increases income and wealth tax, leading to a reduction in the income and wealth Gini, making the wealth distribution more equitable. Social equity is a prerequisite for increasing the duration of the economic system (measured by years). Simultaneously, households choose to increase labor and reduce consumption, leading to a decrease in their utility. As the total savings of households increase, financial intermediaries can provide more production capital to firms.



**Figure 3:** The training curves for 9 baselines on 4 macro-economic indicators under 3 different tasks ( $N = 100$ ).



**Figure 4:** The micro-level behaviors of Random, GA, and MADDPG households while facing identical observations in an episode (300 steps). The subfigure illustrates the average values of labor provided, consumption, taxes paid, and utility for all households at each step. The results reveal that MADDPG households exhibit tax evasion behavior and attain the highest utility.

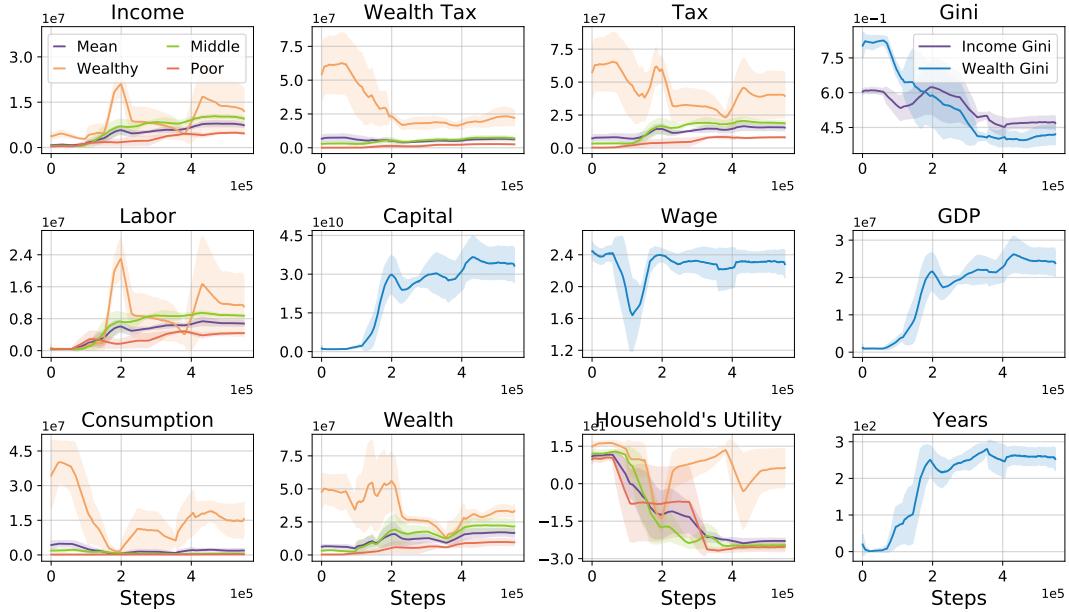
**Table 3:** The per capita economic indicators of the MADDPG algorithm during the testing phase at  $N = 100$  for 3 household groups and the average level across all households.

Households' groups	Income tax	Wealth tax	Total tax	Labor supply	Consumption	Wealth	Income	Per year utility
The wealthy	1.9e6	1.0e7	1.2e7	2.3e6	4.4e7	5.3e7	5.5e6	8.7
The middle class	5.7e6	3.0e6	8.7e6	7.1e6	6.4e5	2.1e7	7.2e6	-24.1
The poor	2.8e6	1.2e6	4.0e6	4.9e6	2.3e5	9.2e6	4.6e6	-25.6
Mean value	3.8e6	2.9e6	6.7e6	5.5e6	4.8e6	1.8e7	5.7e6	-22.7

The additional production labor and capital leads to an increase in output (GDP). The wage rate tends to decrease with an increase in labor and increase with a capital increase, exhibiting a trend of initially decreasing and then increasing.

2. *Different wealth groups adopt distinct strategies.* From the experimental curves (Figure 5) and results (Table 3) for the wealthy,

the middle class, and poor households, we find the following patterns: During  $0 \sim 2e5$  steps, the wealthy contribute significantly to taxation. To stabilize their wealth, they increase work hours and reduce consumption, leading to a decline in utility. In the second phase ( $2e5 \sim 4e5$  steps), the wealthy maximize utility by reducing work hours and significantly increasing consumption, even though their wealth levels decrease. In the third phase ( $4e5 \sim 6e5$  steps), the



**Figure 5: Temporal evolution of economic indicators during MADDPG training under maximizing GDP task on TaxAI ( $N = 100$ ).**

wealthy simultaneously increase labor and consumption, resulting in increased wealth while maintaining relatively stable utility. On the other hand, the middle class and the poor slightly increase work hours and reduce consumption during all three phases, leading to modest growth in wealth but significantly lower utility compared to the wealthy.

**Table 4: The per capita GDP achieved by 9 baselines for different numbers  $N$  of households under maximizing GDP task.**

Algorithm	$N=10$	$N=100$	$N=1000$	$N=10000$
Free Market	$1.3e6$	$4.3e6$	$3.9e6$	$4.0e6$
GA	$1.7e8$	$1.5e7$	NA	NA
IPPO	$5.0e6$	$1.6e7$	$1.7e7$	$1.6e7$
MADDPG	$3.2e6$	$1.1e7$	$1.7e7$	$1.7e7$
MAPPO	$6.1e6$	$7.3e6$	$1.2e7$	NA
HAPPO	$1.8e7$	$1.6e7$	$1.5e7$	NA
HATRPO	$3.2e7$	$1.7e7$	$2.0e7$	NA
HAA2C	$1.6e7$	$1.4e7$	$1.6e7$	NA
BMFAC	$4.0e6$	$1.2e7$	$1.2e7$	NA

#### 5.4 Scalability of Environment

To showcase the scalability of TaxAI in simulating large-scale household agents, we conduct tests with varying numbers of households: 10, 100, 1000, and even 10,000 (as shown in Table 4). The table presents the average per capita GDP for each baseline. The results in Table 4 indicate that IPPO and the improved MADDPG algorithm successfully achieve the maximum GDP when  $N = 10,000$ , whereas traditional methods yield NA (not available). HATRPO achieves

optimal strategies at  $N = 100$  and  $N = 1000$ , respectively, while GA only achieves optimal GDP when  $N$  is small. The above results indicate that TaxAI is capable of simulating 10,000 household agents, surpassing other benchmarks by a significant margin. Moreover, MARL algorithms can successfully solve the optimal tax problem in large-scale agent scenarios. These two advantages are crucial for simulating real-world society.

## 6 CONCLUSION

We introduce TaxAI, a large-scale agent-based dynamic economic environment, and benchmark 2 traditional economic methods and 7 MARL algorithms on it. TaxAI, in contrast to prior work, excels in modeling large-scale heterogeneous households, a wider range of economic activities, and tax types. Moreover, it is calibrated using real data and comes with open-sourced simulation code and MARL benchmark. Our results illustrate the feasibility and superiority of MARL in addressing the optimal taxation problem, while also revealing MARL households' tax evasion behavior.

In the future, we aim to expand and enrich the economic theory of TaxAI by incorporating a broader range of social roles and strategies. Furthermore, we will enhance the scalability of our simulator to accommodate one billion agents, enabling simulations that closely resemble real-world scenarios. By doing so, we aim to attract more researchers to explore complex economic problems using AI or RL techniques, thereby offering practical and feasible recommendations for social planners and the population.

## ACKNOWLEDGMENTS

Haifeng Zhang thanks the support of the National Natural Science Foundation of China, Grant No. 62206289.

## ETHICS STATEMENT

This paper presents work whose goal is to advance the fields of AI for economics. Our work aims to offer suggestions and references for governments and the people, yet it must not be rashly applied to the real world. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

## REFERENCES

- [1] S. R. Aiyagari. 1994. Uninsured Idiosyncratic Risk and Aggregate Saving. *The Quarterly Journal of Economics* 109, 3 (Aug. 1994), 659–684. <https://doi.org/10.2307/2118417>
- [2] S. Rao Aiyagari. 1995. Optimal Capital Income Taxation with Incomplete Markets, Borrowing Constraints, and Constant Discounting. *Journal of Political Economy* 103, 6 (1995), 1158–1175. [arXiv:2138707](https://arxiv.org/abs/2138707)
- [3] S Rao Aiyagari. 1995. Optimal capital income taxation with incomplete markets, borrowing constraints, and constant discounting. *Journal of political Economy* 103, 6 (1995), 1158–1175.
- [4] Tohid Atashbar and Rui Aruhan Shi. 2023. AI and Macroeconomic Modeling: Deep Reinforcement Learning in an RBC Model. <https://doi.org/10.5089/9798400234996.001>
- [5] Ozan Bakır, Barış Kaymak, and Markus Poschke. 2015. Transitional dynamics and the optimal progressivity of income redistribution. *Review of Economic Dynamics* 18, 3 (2015), 679–693.
- [6] Robert J Barro. 1994. The aggregate-supply/aggregate-demand model. *Eastern Economic Journal* 20, 1 (1994), 1–6.
- [7] Nipa Basu, R Pryor, and Tom Quint. 1998. ASPEN: A microsimulation model of the economy. *Computational Economics* 12 (1998), 223–241.
- [8] William J Baumol. 2002. *The free-market innovation machine: Analyzing the growth miracle of capitalism*. Princeton university press.
- [9] Roland Benabou. 2002. Tax and education policy in a heterogeneous-agent economy: What levels of redistribution maximize growth and efficiency? *Econometrica* 70, 2 (2002), 481–517.
- [10] Jess Benhabib and Alberto Bisin. 2018. Skewed wealth distributions: Theory and empirics. *Journal of Economic Literature* 56, 4 (2018), 1261–91.
- [11] Truman Bewley. 1986. Stationary monetary equilibrium with a continuum of independently fluctuating consumers. *Contributions to mathematical economics in honor of Gérard Debreu* 79 (1986).
- [12] Corina Boar and Virgiliu Midrigan. 2022. Efficient redistribution. *Journal of Monetary Economics* 131 (2022), 78–91.
- [13] Eric Bonabeau. [n.d.]. Agent-Based Modeling: Methods and Techniques for Simulating Human Systems. 99 ([n. d.]), 7280–7287. Issue suppl\_3. <https://doi.org/10.1073/pnas.082080899>
- [14] Lucian Buşoniu, Robert Babuška, and Bart De Schutter. [n.d.]. Multi-Agent Reinforcement Learning: An Overview. ([n. d.]), 183–221.
- [15] Daniel Carroll, André Victor Doherty Luduvice, and Eric R Young. 2023. Optimal fiscal reform with many taxes. (2023).
- [16] David Cass. 1965. Optimum growth in an aggregative model of capital accumulation. *The Review of economic studies* 32, 3 (1965), 233–240.
- [17] Lucas Chancel, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. 2022. *World inequality report 2022*. Harvard University Press.
- [18] Varadarajan V Chari and Patrick J Kehoe. 1999. Optimal fiscal and monetary policy. *Handbook of macroeconomics* 1 (1999), 1671–1745.
- [19] Mingli Chen, Andreas Joseph, Michael Kumhof, Xinlei Pan, and Xuan Zhou. 2023. Deep Reinforcement Learning in a Monetary Model. <https://doi.org/10.48550/arXiv.2104.09368> arXiv:2104.09368 [econ, q-fin, stat]
- [20] Arnaud Costinot. 2009. An elementary theory of comparative advantage. *Econometrica* 77, 4 (2009), 1165–1192.
- [21] Michael Curry, Alexander Trott, Soham Phade, Yu Bai, and Stephan Zheng. [n.d.]. *Analyzing Micro-Founded General Equilibrium Models with Many Agents Using Deep Reinforcement Learning*. <https://doi.org/10.48550/arXiv.2201.01163> arXiv:arXiv:2201.01163
- [22] Panayiotis Danassis, Aris Filos-Ratsikas, Haipeng Chen, Milind Tambe, and Boi Faltings. 2023. AI-driven Prices for Externalities and Sustainability in Production Markets. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 2463–2465.
- [23] Paul Davidsson. [n.d.]. Agent Based Social Simulation: A Computer Science View. 5 ([n. d.]).
- [24] Christian Schroeder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. 2020. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533* (2020).
- [25] Peter A Diamond. 1965. National debt in a neoclassical growth model. *The American Economic Review* 55, 5 (1965), 1126–1150.
- [26] Peter A Diamond and James A Mirrlees. 1971. Optimal taxation and public production I: Production efficiency. *The American Economic Review* 61, 1 (1971), 8–27.
- [27] Peter A Diamond and James A Mirrlees. 1971. Optimal taxation and public production II: Tax rules. *The American Economic Review* 61, 3 (1971), 261–278.
- [28] David Domeij and Jonathan Heathcote. 2004. On the distributional effects of reducing capital taxes. *International Economic Review* 45, 2 (2004), 523–554.
- [29] Giovanni Dosi, Giorgio Fagiolo, Mauro Napoletano, and Andrea Roventini. 2013. Income distribution, credit and fiscal policies in an agent-based Keynesian model. *Journal of Economic Dynamics and Control* 37, 8 (2013), 1598–1625.
- [30] Amitava Krishna Dutt. 2006. Aggregate demand, aggregate supply and economic growth. *International review of applied economics* 20, 3 (2006), 319–336.
- [31] Amitava Krishna Dutt and Peter Skott. 1996. Keynesian theory and the aggregate-supply/aggregate-demand framework: a defense. *Eastern Economic Journal* 22, 3 (1996), 313–331.
- [32] Sebastian Dyrda, Marcelo Pedroni, et al. 2016. Optimal fiscal policy in a model with uninsurable idiosyncratic shocks. In *2016 Meeting Papers*, Vol. 1245.
- [33] Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems* 29 (2016).
- [34] Milton Friedman. 1989. Quantity theory of money. In *Money*. Springer, 1–40.
- [35] David Gale. 1955. The law of supply and demand. *Mathematica Scandinavica* (1955), 155–169.
- [36] Oded Galor. 1992. A two-sector overlapping-generations model: A global characterization of the dynamical system. *Econometrica: Journal of the Econometric Society* (1992), 1351–1386.
- [37] John Geanakoplos, Robert Axtell, Doyne J. Farmer, Peter Howitt, Benjamin Conlee, Jonathan Goldstein, Matthew Hendrey, Nathan M. Palmer, and Chun-Yi Yang. [n.d.]. Getting at Systemic Risk via an Agent-Based Model of the Housing Market. 102, 3 ([n. d.]), 53–58.
- [38] Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. [n.d.]. Mean-Field Multi-Agent Reinforcement Learning: A Decentralized Network Approach. ([n. d.]). [arXiv:2108.02731](https://arxiv.org/abs/2108.02731)
- [39] Jonathan Heathcote, Kjetil Storesletten, and Giovanni L Violante. 2014. Consumption and labor supply with partial insurance: An analytical framework. *American Economic Review* 104, 7 (2014), 2075–2126.
- [40] Jonathan Heathcote, Kjetil Storesletten, and Giovanni L Violante. 2017. Optimal tax progressivity: An analytical framework. *The Quarterly Journal of Economics* 132, 4 (2017), 1693–1754.
- [41] Edward Hill, Marco Bardoscia, and Arthur Turrell. [n.d.]. *Solving Heterogeneous General Equilibrium Economic Models with Deep Reinforcement Learning*. <https://doi.org/10.48550/arXiv.2103.16977> arXiv:arXiv:2103.16977
- [42] Natascha Hinterlang and Alina Tänzer. [n.d.]. *Optimal Monetary Policy Using Reinforcement Learning*. <https://doi.org/10.2139/ssrn.4025682>
- [43] Shelby D Hunt and Robert M Morgan. 1995. The comparative advantage theory of competition. *Journal of marketing* 59, 2 (1995), 1–15.
- [44] Sanford Ikeda. 2002. *Dynamics of the mixed economy: Toward a theory of interventionism*. Routledge.
- [45] Norman Johnson. 2014. *Mixed economies welfare*. Routledge.
- [46] Emil Kauder. 2015. *History of marginal utility theory*. Vol. 2238. Princeton University Press.
- [47] Tomas B Klos and Bart Nootboom. 2001. Agent-based computational transaction cost economics. *Journal of Economic Dynamics and Control* 25, 3–4 (2001), 503–526.
- [48] Tjalling C Koopmans. 1963. On the concept of optimal economic growth. (1963).
- [49] Raphael Koster, Jan Balaguer, Andrea Tacchetti, Ari Weinstein, Tina Zhu, Oliver Hauser, Duncan Williams, Lucy Campbell-Gillingham, Phoebe Thacker, Matthew Botvinick, and Christopher Summerfield. [n.d.]. Human-Centred Mechanism Design with Democratic AI. 6, 10 ([n. d.]), 1398–1407. Issue 10. <https://doi.org/10.1038/s41562-022-01383-x>
- [50] Artem Kuriksha. [n.d.]. *An Economy of Neural Networks: Learning from Heterogeneous Experiences*. <https://doi.org/10.48550/arXiv.2110.11582> arXiv:arXiv:2110.11582
- [51] Xiaochen Li, Wenji Mao, Daniel Zeng, and Fei-Yue Wang. [n.d.]. Agent-Based Social Simulation and Modeling in Social Computing. In *Intelligence and Security Informatics* (2008). Springer, Berlin, Heidelberg, 401–412. [https://doi.org/10.1007/978-3-540-69304-8\\_41](https://doi.org/10.1007/978-3-540-69304-8_41)
- [52] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).
- [53] David A Love. 2013. Optimal rules of thumb for consumption and portfolio choice. *The Economic Journal* 123, 571 (2013), 932–961.
- [54] Ryan Lowe, given-i=1YI family=WU, given-i=YI, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. [n.d.]. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *Advances in Neural Information Processing Systems* (2017), Vol. 30. Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2017/hash/68a9750337a418a86fe06c1991a1d64c-Abstract.html>

- [55] Robert E Lucas. 1980. Two illustrations of the quantity theory of money. *The American Economic Review* 70, 5 (1980), 1005–1014.
- [56] Pattie Maes and Rodney A Brooks. 1990. Learning to Coordinate Behaviors. In *AAAI*, Vol. 90. Boston, MA, 796–802.
- [57] John McMillan and Barry Naughton. 1992. How to reform a planned economy: lessons from China. *Oxford review of economic policy* 8, 1 (1992), 130–143.
- [58] Victor Nee. 1992. Organizational dynamics of market transition: Hybrid forms, property rights, and mixed economy in China. *Administrative science quarterly* (1992), 1–27.
- [59] Frans A Oliehoek, Christopher Amato, et al. 2016. *A concise introduction to decentralized POMDPs*. Vol. 1. Springer.
- [60] Fatih Ozhamaratli and Paolo Barucca. 2022. Deep Reinforcement Learning for Optimal Investment and Saving Strategy Selection in Heterogeneous Profiles: Intelligent Agents Working towards Retirement. arXiv:2206.05835 [cs, econ, q-fin]
- [61] Mike W Peng and Peggy Sue Heath. 1996. The growth of the firm in planned economies in transition: Institutions, organizations, and strategic choice. *Academy of management review* 21, 2 (1996), 492–528.
- [62] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. [n.d.]. Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. 21, 1 ([n. d.]), 7234–7284.
- [63] Rui and Shi. 2022. Learning from Zero: How to Make Consumption-Saving Decisions in a Stochastic Environment with an AI Algorithm. <https://doi.org/10.48550/arXiv.2105.10099> arXiv:2105.10099 [econ, q-fin]
- [64] Paul A Samuelson. 1958. An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of political economy* 66, 6 (1958), 467–482.
- [65] Abrahama Ayooluwa Odukoya Sch. [n.d.]. Intelligence in the Economy: Emergent Behaviour in International Trade Modelling with Reinforcement Learning. ([n. d.]).
- [66] Michael Schlechtinger, Damaris Kosack, Heiko Paulheim, Thomas Fetzer, and Franz Krause. 2023. The Price of Algorithmic Pricing: Investigating Collusion in a Market Simulation with AI Agents. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 2748–2750.
- [67] Rui Aruhan Shi. 2021. Can an AI Agent Hit a Moving Target. *arXiv preprint arXiv 2110* (2021).
- [68] Adam Smith. 1937. *The wealth of nations [1776]*. Vol. 11937. na.
- [69] Adam Smith. 2010. *The theory of moral sentiments*. Penguin.
- [70] Adam Smith. 2023. *An Inquiry into the Nature and Causes of the Wealth of Nations*. BoD-Books on Demand.
- [71] Mitja Steinbacher, Matthias Raddant, Fariba Karimi, Eva Camacho Cuena, Simone Alfaro, Giulia Iori, and Thomas Lux. 2021. Advances in the agent-based modeling of economic and social behavior. *SN Business & Economics* 1, 7 (2021), 99.
- [72] Sainbayar Sukhaatar, Rob Fergus, et al. 2016. Learning multiagent communication with backpropagation. *Advances in neural information processing systems* 29 (2016).
- [73] Cass R Sunstein. 1997. *Free markets and social justice*. Oxford University Press on Demand.
- [74] Richard S. Sutton and Andrew G. Barto. [n.d.]. *Reinforcement Learning: An Introduction*. MIT press.
- [75] Ardi Tamppu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente. [n.d.]. Multiagent Cooperation and Competition with Deep Reinforcement Learning. ([n. d.]). arXiv:1511.08779
- [76] Ming Tan. [n.d.]. Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents. In *Proceedings of the Tenth International Conference on Machine Learning* (1993). 330–337.
- [77] Leigh Tesfatsion. 2001. Introduction to the special issue on agent-based computational economics. *Journal of Economic Dynamics and Control* 25, 3-4 (2001), 281–293.
- [78] Alexander Trott, Sunil Srinivasa, Sebastien Haneuse, and Stephan Zheng. [n.d.]. *Building a Foundation for Data-Driven, Interpretable, and Robust Policy Design Using the AI Economist*. <https://doi.org/10.48550/arXiv.2108.02904> arXiv:arXiv:2108.02904
- [79] Anil Yaman, Joel Z. Leibo, Giovanni Iacca, and Sang Wan Lee. [n.d.]. *The Emergence of Division of Labor through Decentralized Social Sanctioning*. arXiv:arXiv:2208.05568 <http://arxiv.org/abs/2208.05568>
- [80] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. [n.d.]. Mean Field Multi-Agent Reinforcement Learning. In *International Conference on Machine Learning* (2018). PMLR, 5571–5580.
- [81] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. [n.d.]. The Surprising Effectiveness of Ppo in Cooperative Multi-Agent Games. 35 ([n. d.]), 24611–24624.
- [82] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems* 35 (2022), 24611–24624.
- [83] Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. [n.d.]. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. ([n. d.]), 321–384.
- [84] Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. [n.d.]. The AI Economist: Taxation Policy Design via Two-Level Deep Multiagent Reinforcement Learning. 8, 18 ([n. d.]), eabk2607. <https://doi.org/10.1126/sciadv.abk2607>
- [85] Yifan Zhong, Jakub Grudzien Kuba, Siyi Hu, Jiaming Ji, and Yaodong Yang. 2023. Heterogeneous-Agent Reinforcement Learning. *arXiv preprint arXiv:2304.09870* (2023).
- [86] Ming Zhou, Yong Chen, Ying Wen, Yaodong Yang, Yufeng Su, Weinan Zhang, Dell Zhang, and Jun Wang. [n.d.]. Factorized Q-Learning for Large-Scale Multi-Agent Systems. In *Proceedings of the First International Conference on Distributed Artificial Intelligence* (2019). 1–7.

## A ENVIRONMENT SETTING

### A.1 Classical Tax Models

In Section 2, we compare classical tax models using key economic indicators (shown in Table 5) to highlight the Bewley-Aiyagari model's advantages in modeling economic activities and determining optimal tax strategies.

- Incomplete Markets: Markets with asymmetric information.
- Uncertainty: Risks and unpredictability in decision-making and forecasting.
- Inequality: Disparities in wealth and income distribution.

### A.2 Government Tasks

The mathematical objective functions for the four types of government tasks are as follows:

*Maximizing GDP Growth Rate.* The economic growth can be measured by Gross Domestic Product (GDP). Without considering imports and exports in an open economy, GDP is equal to the output  $Y_t$  in our model. Based on reality, we set the government's objective to maximize the GDP growth rate,

$$\max J = \mathbb{E}_0 \sum_{t=0}^{T_N} \beta^t \left( \frac{Y_t - Y_{t-1}}{Y_{t-1}} \right) \quad (14)$$

*Minimizing Social Inequality.* Social equality and stability build the foundation for all social activities. Social inequality is usually measured by the Gini coefficient of wealth distribution  $\mathcal{W}_t$  and income distribution  $\mathcal{I}_t$ . The Gini coefficient is calculated by the ratio of the area between the Lorenz curve and the perfect equality line, divided by the total area under the perfect equality line (shown in figure 2). The Gini coefficient ranges between 0 (perfect equality) and 1 (perfect inequality).

$$\min J = \mathbb{E}_0 \sum_{t=0}^{T_N} \beta^t (\text{Gini}(\mathcal{I}_t) \text{Gini}(\mathcal{W}_t)) \quad (15)$$

*Maximizing social welfare.* Optimizing social welfare refers to maximizing the sum of utilities for all households. This is one of the government's most important responsibilities.

$$\max J = \mathbb{E}_0 \sum_{t=0}^{T_N} \beta^t \left( \sum_i^N \frac{c_t^{i,1-\theta}}{1-\theta} - \frac{h_t^{i,1+\gamma}}{1+\gamma} \right) \quad (16)$$

*Optimizing Multiple Tasks.* If the government aims to simultaneously optimize multiple objectives, we weigh and sum up multiple objectives. The weights  $\omega_1$ , and  $\omega_2$  indicate the relative importance of Gini and welfare objectives.

$$\begin{aligned} \max J = \mathbb{E}_0 \sum_{t=0}^{T_N} \beta^t & \left( \frac{Y_t - Y_{t-1}}{Y_t} - \omega_1 \cdot \text{Gini}(\mathcal{I}_t) \text{Gini}(\mathcal{W}_t) \right. \\ & \left. + \omega_2 \cdot \left( \sum_i^N \frac{c_t^{i,1-\theta}}{1-\theta} - \frac{h_t^{i,1+\gamma}}{1+\gamma} \right) \right) \end{aligned} \quad (17)$$

In TaxAI, the four types of government reward functions are as follows:

- (1) Maximizing GDP Growth Rate:

$$r_{g,t}(s_t, a_{g,t}) = \frac{Y_t - Y_{t-1}}{Y_{t-1}} \quad (18)$$

- (2) Minimizing Social Inequality:

$$r_{g,t}(s_t, a_{g,t}) = -\text{Gini}(\mathcal{I}_t) \text{Gini}(\mathcal{W}_t) \quad (19)$$

- (3) Maximizing Social Welfare:

$$r_{g,t}(s_t, a_{g,t}) = \sum_i^N \frac{c_t^{i,1-\theta}}{1-\theta} - \frac{h_t^{i,1+\gamma}}{1+\gamma} \quad (20)$$

- (4) Optimizing Multiple Tasks:

$$\begin{aligned} r_{g,t}(s_t, a_{g,t}) = & \frac{Y_t - Y_{t-1}}{Y_t} - \omega_1 \cdot \text{Gini}(\mathcal{I}_t) \text{Gini}(\mathcal{W}_t) \\ & + \omega_2 \cdot \left( \sum_i^N \frac{c_t^{i,1-\theta}}{1-\theta} - \frac{h_t^{i,1+\gamma}}{1+\gamma} \right) \end{aligned} \quad (21)$$

### A.3 Assumption

We have summarized the assumptions involved in the economic model section 3 as follows:

- (1) Social roles, such as households and government, are considered rational agents.
- (2) Households are not allowed to incur debt and engage only in risk-free investments.
- (3) The labor productivity of households is categorized into two states: normal state and super-star state. The dynamic changes in each state follow an AR(1) process.
- (4) The capital market, goods market, and labor market clear.
- (5) The technology firm represents all enterprises and factories, producing a homogeneous good, and follows the Cobb-Douglas production function.
- (6) The financial Intermediary operates without arbitrage.

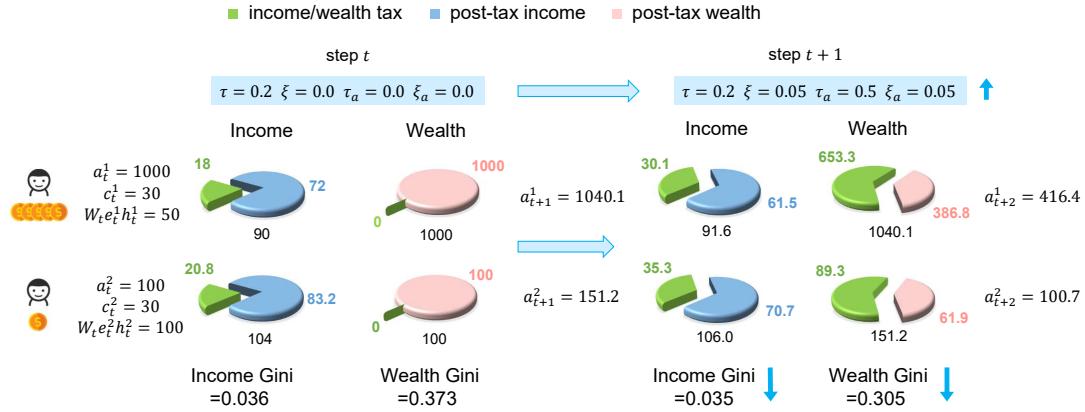
### A.4 Model Dynamics

To elucidate the dynamics of the economic model more lucidly, we present in Figure 6 the impact of tax changes (from time  $t$  to time  $t+1$ ) on individuals with varying levels of wealth. The pie chart in the figure illustrates the current households' income and wealth distributions, along with the taxation proportions. The green sector represents income or wealth taxation, the blue sector signifies post-tax income, and the pink sector represents post-tax wealth. Furthermore, we calculate the Gini coefficients for income and wealth distributions to illustrate the effects of changes in tax policies on social equity. (Given that this economic model involves only two households, the Gini coefficients are much smaller than those found in real-world scenarios; therefore, their use here is only for illustrating Gini coefficient variations). Suppose the economy consists of two categories of households—a wealthy household (with initial wealth  $a_t^1 = 1000$ ) and a poor household (with initial wealth  $a_t^2 = 100$ ). Their income predominantly originates from labor income  $W_t e_t h_t$  and asset returns  $r_{t-1} a_t$  (where  $r_{t-1} = 0.04$ ). And they have the same consumption level of  $c_t^1 = c_t^2 = 30$  for each step.

At step  $t$ , the wealthy household's income is  $i_t^1 = 50 + 0.04 * 1000 = 90$ , while the poor household's income is  $i_t^2 = 100 + 0.04 * 100 = 104$ . Considering tax rate parameters of  $\tau = 0.2$  and  $\xi = \tau_a = \xi_a = 0$ , both the wealthy and poor households face an income tax rate of 20%, with no taxation on assets. Consequently, the income tax for the wealthy household amounts to  $90 * 0.2 = 18$ , resulting in a post-tax income of  $90 - 18 = 72$ . Similarly, the income tax for

**Table 5: Comparison of 5 classical tax models on key economic indicators.**

Classic Tax models	Incomplete Markets	Uncertainty	Inequality	Emphasis
Ramsey-Cass-Koopmans (RCK) model	×	×	×	Economic Growth
Mirrlees model / Diamond-Mirrlees model	×	✓	✓	Tradeoff between Economic Growth and Equity
Overlapping Generations (OLG) model	✓	✓	✓	Intergenerational Equity, Pension Systems
Bewley-Aiyagari model	✓	✓	✓	Tax Policies, Social Equity, Social Security Systems
Atkinson-Stiglitz model	×	×	✓	Social Welfare and Taxation Systems



**Figure 6: Illustration of tax impact on the rich and poor households.**

the poor household is 20.8, resulting in a post-tax income of 83.2. Through the budget constraint in (3), we determine the next-step wealth as  $a_{t+1}^1 = 1040.1$  and  $a_{t+1}^2 = 151.2$ . The Gini coefficient for income is 0.036, while the Gini coefficient for wealth is 0.373, reflecting the wealth distribution between these two economic agents.

Moving to **step  $t + 1$** , despite maintaining unchanged consumption levels and labor incomes for both wealthy and poor households, the increase in wealth fosters an increment in asset returns. Consequently, the incomes for the wealthy and poor households have risen relative to step  $t$ ; namely,  $i_{t+1}^1 = 91.6$  and  $i_{t+1}^2 = 106.0$ . At this juncture, while the tax rate parameter  $\tau$  remains unchanged,  $\xi$  has been elevated to 0.05, transforming the income tax function into a non-linear structure and raising tax rates. Meanwhile, the asset tax parameters  $\tau_a = 0.5$  and  $\xi_a = 0.05$  prompt the government to initiate asset taxation. Under equivalent labor income conditions, the wealthy household's income tax has risen to 30.1 via Equation 22, while asset taxation has surged to 653.3. A similar scenario ensues for the poor household. Thus, due to the escalated tax rate, more than half of the asset for both wealthy and poor households is allocated toward taxation. Ultimately, the remaining wealth is  $a_{t+2}^1 = 416.4$  and  $a_{t+2}^2 = 100.7$ . While initially, the wealthy household's asset was **tenfold** that of the poor household, the introduction of wealth taxation by the government has reduced this disparity to roughly a **fourfold** difference. The increase in  $\xi$  has led to a marginal decline in the income Gini coefficient, while the wealth tax has significantly lowered the wealth Gini coefficient, concurrently promoting social equity.

$$T(i_t) = i_t - (1 - \tau) \frac{i_t^{1-\xi}}{1-\xi}, \quad T^a(a_t) = a_t - \frac{1 - \tau_a}{1 - \xi_a} a_t^{1-\xi_a} \quad (22)$$

## A.5 Environment Parameters

The assigned parameters of TaxAI are shown in table 6. At the beginning of each episode, we initialize households' assets  $\{a_i^i\}_{i=1}^N$  via the statistical data from the 2022 Survey of Consumer Finances (SCF) data and calibrate aggregate labors  $L_t$  and the maximum of working hours  $h_{max}$  based on the wealth-to-income ratio computed from the 2022 SCF data.

**Table 6: Assigned Parameters in Environment**

Hyperparameter	Value
CRRA $\theta$	1
Inverse Frisch elasticity $\gamma$	2
discount factor $\beta$	0.975
capital elasticity $\alpha$	1/3
depreciation rate $\delta$	0.06
return to savings $r_t$	0.04
consumption tax rate $\tau_s$	0.065
normal to super-star probability $p$	$2.2e^{-6}$
remain in super-star probability $q$	0.990
persistence $\rho_e$	0.982
volatility of the standard normal shocks $\sigma_e$	0.200
super ability times $\bar{e}$	504.3
Wealth to income ratio	6.6

**Table 8: Variables in Economic Model**

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)
$r_t$	interest rate
$i_t$	Household's income
$N$	Households' number
$\rho_e$	Persistence
$u_t$	Standard normal shocks
$\sigma_e$	Volatility of the standard normal shocks
$\bar{e}$	Labor market ability of super-star state
$p$	Transition probability from normal state to super-star state
$q$	Transition probability of remaining in super-star state
$\beta$	Discount factor
$\theta$	Coefficient of relative risk aversion (CRRA)
$\gamma$	Inverse Frisch elasticity
$\tau_s$	Consumption taxes
$T(i_t)$	Income taxes
$T^d(a_t)$	Asset taxes
$\tau, \tau_a$	Average level of marginal income and asset tax
$\xi, \xi_a$	Slope of marginal income and asset tax schedule
$Y_t$	Firm's output
$K_t$	Capital for production
$L_t$	Labor for production
$\alpha$	Capital elasticity
$C_t$	Households' gross consumption
$G_t$	Government spending
$X_t$	Physical capital investment
$\delta$	Depreciation rate
$R_t$	Borrowing rate of firm
$G_t$	Government spending
$T_t$	Households' gross taxation
$B_t$	Debt
$I_t$	Households' gross post-tax income
$W_t$	Households' gross post-tax assets
$A_t$	Households' gross deposits
$\omega_1, \omega_2$	Relative weight of social equality and social welfare

## A.6 Teriminal Condition

We will terminate the current episode under the following circumstances:

- If the maximum number of steps in an episode is reached.
- If the Gini coefficient of income or wealth exceeds a threshold;
- If the Gross Domestic Product (GDP) is insufficient to cover the total consumption of households ( $C_t + G_t > Y_t$ ).
- In the event of household bankruptcy ( $a_t^i < 0$ ).
- If the calculation of households' or government's rewards results in an overflow or leads to the appearance of NaN (Not-a-Number) values.

## B TIMING TESTS

Table 7 shows the number of dynamics evaluations per second in a single thread. Results are averaged over 10 episodes on AMD EPYC 7742 64-Core Processor with GPU GA100 [GRID A100 PCIe 40GB].

**Table 7: The number of dynamics evaluations per second in a single thread.**

Algorithm	$N=10$	$N=100$	$N=1000$	$N=10000$
Number of steps	1799.927	482.335	46.983	4.018
num of episodes	6.289	1.689	0.157	0.013
episode length	285.900	285.400	300.000	300.000

## C GENETIC ALGORITHM

In this subsection, we provide additional details regarding the implementation of genetic algorithms. To evaluate the performance of different policies, we employ relevant economic indicators such as GDP and the Gini coefficient as the fitness function. Our genetic algorithm implementation includes the utilization of uniform crossover and simple mutation. The specific parameters used in the algorithm are presented in Table 12.

For household agents, we refer to Heathcote's [39] approach in selecting saving and working hours for households under the shocks of agent productivity and preferences. However, our model does not incorporate the preference for working and consumption, denoted as  $\varphi_t = 0$ . The productivity of work is determined by

$$\log e_t = \alpha_t + \varepsilon_t \quad (23)$$

$$\alpha_t = \alpha_{t-1} + \omega_t \quad (24)$$

$$\varepsilon_t = \kappa_t + \theta_t \quad (25)$$

The households strategies are obtained by competitive equilibrium under the productivity shock. The proportions of consumption to income and working hours  $c_t^a$  and  $h_t^a$  are calculated using the following equations:

$$\log c_t^a(s^t) = -\varphi_t + \frac{1+\gamma}{\gamma+\theta} \alpha_t + \mathcal{M}_t^a \quad (26)$$

$$\log h_t^a(s^t) = -\varphi_t + \frac{1-\theta}{\gamma+\theta} \alpha_t + \frac{1}{\gamma} (\kappa_t + \theta_t) - \frac{\theta}{\gamma} \mathcal{M}_t^a \quad (27)$$

$$\mathcal{M}_t^a = \gamma / (\gamma + \theta) \log \left( \int \exp \left( \frac{1+\gamma}{\gamma} \theta_t \right) dF_{\theta_t} \right) \quad (28)$$

where  $\omega_t$  and  $\theta_t$  is an independently distributed shock,  $\kappa_t$  is a permanent shock ( $\kappa_t = 0$  in our experiment).

## D BI-LEVEL MEAN FIELD ACTOR-CRITIC ALGORITHM

We propose a Bi-level Mean Field Actor-Critic (BMFAC) algorithm for the Stackelberg game between the government and  $N$  households. In our BMFAC method, the government serves as the leader and the households as equal followers. The objective of the government is to maximize its objective by setting tax rates and government spending while anticipating the households' actions. Conversely, households aim to optimize their lifetime utility by making

decisions regarding their saving policies and working hours. Initially, the government's action is initialized to a free-market policy (without taxes), and we employ an iterative process to optimize the household agents' policies in the inner loop. Given the household agents' policies, we generate sample trajectories for the government agent and update its actor and critic networks accordingly. Subsequently, we sample data and update the networks for the households, considering the government's policy. This iterative process continues until convergence is attained.

---

**Algorithm 1** BMFAC: Bi-level Mean Field Actor Critic

---

- 1: Initialize  $Q_g^\phi, Q_h^{\phi^-}, \pi_g^\theta, \pi_h^{\theta_-}$ , and  $\bar{a}_h^i$  for all  $i \in \{1, \dots, N\}$
- 2: Initialize the government policy as the free market rules  $a_g^0 = \{0, 0, 0, 0, 0.189\}$ .
- 3: **for** Outer Loop **do**
- 4:   **for** Inner Loop **do**
- 5:     **for** each sample step **do**
- 6:       Given the government policy  $\pi_g$ , the government agent sample action  $a_g \sim \pi_g(o_g)$ ; the household agent  $i$  sample action  $a_h^i \sim \pi_h(o_h^i, a_g)$ .
- 7:       Take the actions and observe  $r_g, r_h = \{r_h^1, \dots, r_h^N\}$  and next state  $o'_g, o_h = \{o_h^{1'}, \dots, o_h^{N'}\}$ .
- 8:       Store  $\langle o_g, o_h, a_g, a_h, r_g, r_h, o'_g, o_h' \rangle$  in replay buffer  $\mathcal{D}$ .
- 9:     **end for**
- 10:    Update the households' actor and critic for  $N_{\text{update}}^h$  steps:
$$\begin{aligned} \psi &\leftarrow \psi - \lambda_Q \hat{\nabla}_\psi \mathcal{L}(\psi) \\ \psi^- &\leftarrow \psi^- - \lambda_Q \hat{\nabla}_{\psi^-} \mathcal{L}(\psi^-) \\ \sigma &\leftarrow \sigma - \lambda_\pi \hat{\nabla}_\sigma \mathcal{L}(\sigma) \end{aligned}$$
- 11:   **end for**
- 12:   **for** each sample step **do**
- 13:     Given the households policy  $\pi_h$ , the government agent sample action  $a_g \sim \pi_g(o_g)$ ; the household agent  $i$  sample action  $a_h^i \sim \pi_h(o_h^i, a_g)$ .
- 14:     Take the actions and observe  $r_g, r_h = \{r_h^1, \dots, r_h^N\}$  and next state  $o'_g, o_h = \{o_h^{1'}, \dots, o_h^{N'}\}$ .
- 15:     Store  $\langle o_g, o_h, a_g, a_h, r_g, r_h, o'_g, o_h' \rangle$  in replay buffer  $\mathcal{D}$ .
- 16:   **end for**
- 17:   Update the government's actor and critic for  $N_{\text{update}}^g$  steps:
$$\begin{aligned} \phi &\leftarrow \phi - \lambda_Q \hat{\nabla}_\phi \mathcal{L}(\phi) \\ \phi^- &\leftarrow \phi^- - \lambda_Q \hat{\nabla}_{\phi^-} \mathcal{L}(\phi^-) \\ \theta &\leftarrow \theta - \lambda_\pi \hat{\nabla}_\theta \mathcal{L}(\theta) \end{aligned}$$
- 18: **end for**

---

## D.1 Government Agent

We update the government agent via the actor-critic method. We consider double Q-function  $Q_\phi(\mathbf{o}_t, \mathbf{a}_t)$ ,  $Q_{\phi^-}(\mathbf{o}_t, \mathbf{a}_t)$ , and a policy  $\pi_\theta(a_{g,t} \mid o_{g,t})$ . For example, the value functions can be modeled as expressive neural networks and the policy as a Gaussian with

mean and covariance given by neural networks. We will next derive update rules for these parameter vectors.

*The critic network.* The Q-function parameters can be trained to minimize the soft Bellman residual.

$$\mathcal{L}(\phi) = \left( y_g - Q_\phi \left( o_g, a_g, \{\pi_h(o_h^i, a_g)\}_i^N \right) \right)^2 \quad (29)$$

where the target mean field value with the weight  $\phi^-$  is computed by

$$y_g = r_g + \gamma \mathbb{E}_{a'_g \sim \pi_g} Q_{\phi^-}(s'_g, a'_g, \{\pi_h(o'_g, o_h^i, a_g)\}_i^N) \quad (30)$$

Differentiating  $\mathcal{L}(\phi)$  gives

$$\begin{aligned} \nabla_\phi \mathcal{L}(\phi) &= \left( y_g - Q_\phi \left( o_g, a_g, \{\pi_h(o_h^i, a_g)\}_i^N \right) \right) \\ &\quad \nabla_\phi Q_\phi \left( o_g, a_g, \{\pi_h(o_h^i, a_g)\}_i^N \right) \end{aligned} \quad (31)$$

which enables the gradient-based optimizers for training.

*The actor network.* The policy network of the government  $\pi_g^\theta$  is training by the sampled policy gradient:

$$\nabla_\theta \mathcal{L}(\theta) \approx \nabla_\theta \log \pi_g^\theta(o_g) Q_\phi \left( o_g, a_g, \{\pi_h(o_h^i, a_g)\}_i^N \right) \quad (32)$$

## D.2 Household Agents

We update household agents via the mean field actor-critic method. We consider double Q-function  $Q_\psi(\mathbf{o}_h, \mathbf{a}_h)$ ,  $Q_{\psi^-}(\mathbf{o}_h, \mathbf{a}_h)$  as a shared critic and a shared policy  $\pi_\sigma(a_{h,t} \mid o_{h,t})$  for all household agents. For example, the value functions can be modeled as expressive neural networks and the policy as a Gaussian with mean and covariance given by neural networks. We will next derive update rules for these parameter vectors.

*The critic network.* The Q-function parameters can be trained to minimize the Bellman residual.

$$\mathcal{L}(\phi) = \left( y_h - Q_\psi \left( o_h, a_g, a_h, \bar{A}_h \right) \right)^2 \quad (33)$$

with

$$y = r_h + \gamma V_{\psi^-}^{\text{MF}}(o'_h) \quad (34)$$

The mean field value function  $V_{\psi^-}^{\text{MF}}(o'_h)$  for household agent  $i$  over the government's action  $a_g$  is

$$V_{\psi^-}^{\text{MF}}(o'_h) = \mathbb{E}_{a'_h, \bar{A}'_h \sim \pi_h(o'_h)} Q(o'_h, a'_g, a'_h, \bar{A}'_h) \quad (35)$$

The mean action  $\bar{a}_h^i$  of all  $i$ 's neighbors is first calculated by averaging the actions  $a^j$  taken by  $i$ 's  $N^i$  neighbors from policies  $\pi_t(o_h^j)$  parametrized by their previous mean actions  $\bar{a}^-$

$$\bar{a}^i = \frac{1}{N^i} \sum_j a^j, a^j \sim \pi_t \left( \cdot \mid o_h^j, \bar{a}^-_h \right) \quad (36)$$

The gradient of the households' critic network is given by

$$\nabla_\psi \mathcal{L}(\phi) = \left( y_h - Q_\psi \left( o_h, a_g, a_h, \bar{A}_h \right) \right) \nabla_\psi Q_\psi \left( o_h, a_g, a_h, \bar{A}_h \right) \quad (37)$$

*The actor network.* With each  $\bar{a}^i$  calculated as in Eq. 36, the policy network of household agents  $\pi_h^\sigma$ , i.e. the actor, of MF-AC is trained by the sampled policy gradient:

$$\nabla_\sigma \mathcal{L}(\sigma) \approx \log \pi_\sigma(o_h, \bar{A}_h) Q_\psi \left( o_h, a_g, a_h, \bar{A}_h \right) \quad (38)$$

## E ADDITIONAL RESULTS

### E.1 Baselines Setting

**(1) Free Market Policy** The government takes a hands-off approach to market activities, and households choose to work and saving actions stochastically.

**(2) Genetic Algorithm** The government employs the genetic algorithm to optimize tax policies, while household agents adopt consumption and labor supply strategies following Heathcote's work [39]. See Appendix C for more details.

**(3) Independent PPO** Both the government and households utilize the PPO policy, treat other agents as part of the environment and learn policies based on local observations.

**(4) Multi-Agent Deep Deterministic Policy Gradient (MADDPG)** To address large-scale multi-agent problems using MADDPG, we categorize the agents into 4 groups: the government, the top 10% richest households, the top 10-50% richest households, and the bottom 50% richest households. Each group of agents shares an actor and a centralized critic and updates their networks using the MADDPG method.

**(5) Multi-Agent Proximal Policy Optimization (MAPPO)** Both government and household agents utilize a PPO actor and adopt a centralized value function. The government agent and household agents share a common actor network, with the observation and action spaces of heterogeneous agents aligned.

**(6) Bi-level Mean Field Actor-Critic (BMFAC)** The government acts as the leader agent and updates its policy using the policy gradient method. While the household act as follower agents, sharing a common actor and critic, and updating networks via the mean-field actor-critic approach [80]. The government and households utilize a bi-level optimization approach, iteratively optimizing their policies to converge toward a Markov Perfect Equilibrium. More details are provided in the Appendix D.

**(7) Heterogeneous-Agent Reinforcement Learning (HARL)** HARL algorithms (HAPPO, HATRPO, HAA2C) employ heterogeneous actors and a joint advantage estimator, without the need for relying on restrictive parameter-sharing trick.

### E.2 Training Curves

To showcase the learning performance of the MARL algorithms on the TaxAI simulator, we present training curves for 9 baselines across four different tasks (shown in Figure 7). Among these baselines, GA and free-market policy are not learning algorithms, thus their convergence results are represented by straight lines. The MARL algorithms are represented by solid lines indicating the mean values, while the shaded areas reflect the variances of these economic indicators. In the subplots, each row represents an economic indicator, while each column corresponds to the experimental task. From Figure 7, it is evident that the MARL algorithm outperforms the traditional methods on six economic indicators across each task. Different algorithms excel in different tasks. For instance, HATRPO achieves the highest per capita GDP in the maximizing GDP task, while BMFAC achieves the lowest wealth Gini coefficient and the highest social welfare in the tasks of minimizing social inequality and maximizing welfare. In the multi-task scenario, different algorithms prioritize different optimization objectives. These results

highlight the effectiveness and superiority of the MARL algorithms in finding optimal tax policies for the government and optimal working and saving strategies for households.

### E.3 Economic Evolution

To explore the optimal solution to the dynamic game between the government and households ( $N=100$ ), we implement five baselines over three tasks: maximizing GDP growth rate, minimizing social inequality, and maximizing social welfare. Additionally, we test the converged policies and showcase the macroeconomic indicators, such as per capita GDP, wealth distribution, and social welfare, as the number of steps increases. This analysis aims to illustrate the evolutionary process of the economy under different algorithms. In Figure 8, each row represents a macroeconomic indicator, with the first row corresponding to per capita GDP and the third-row representing social welfare. The horizontal axis represents different baselines and the height of the bars shows the GDP and social welfare at the current step. The second row displays the Lorenz curves for wealth distribution, where the vertical axis represents the cumulative proportion of wealth and the horizontal axis represents the cumulative proportion of a population ranked by their wealth. The five curves correspond to the Lorenz curves of different baselines, where a closer proximity to the line of perfect equality indicates a higher level of equality in wealth distribution. Each column is divided based on steps, where we define each step as representing one year in the real world. The government and households make decisions at each time step. The maximum step is set to 300. When an economy satisfies a terminal condition, such as insufficient production to meet consumer demand or the Gini coefficient of wealth distribution exceeding a threshold, the economic system stops evolving.

The experimental results in Figure 8 provide evidence for the ability of TaxAI to simulate the economic evolution and the feasibility and superiority of MARL algorithms in solving the optimal taxation problem. Firstly, the free market policy struggles to progress beyond 100 steps, whereas the other four tax optimization policies can evolve up to 300 steps, thereby highlighting the efficacy of the proposed policy interventions for ensuring the sustainability of the economy. Furthermore, with the increasing steps, the GDP and social welfare associated with the GA, IPPO, MADDPG, and BMFAC methods continuously increase, while wealth inequality decreases. This alignment with the given optimization objectives reveals the feasibility of these methods in solving the optimal tax problem. Lastly, in the tasks evaluating GDP and social welfare, IPPO achieves the highest GDP, while BMFAC significantly outperforms other algorithms in terms of social welfare. In terms of the Lorenz curve, 5 baselines except the free market policy converge from an unequal state to a relatively higher level of equality. These results show that MARL methods possess the capability to simulate the evolution of the economy and attain optimal tax policies that surpass those achieved by traditional approaches such as the free market and GA.

The hyperparameters used in baselines algorithms are shown in Table 9, 10, 11.

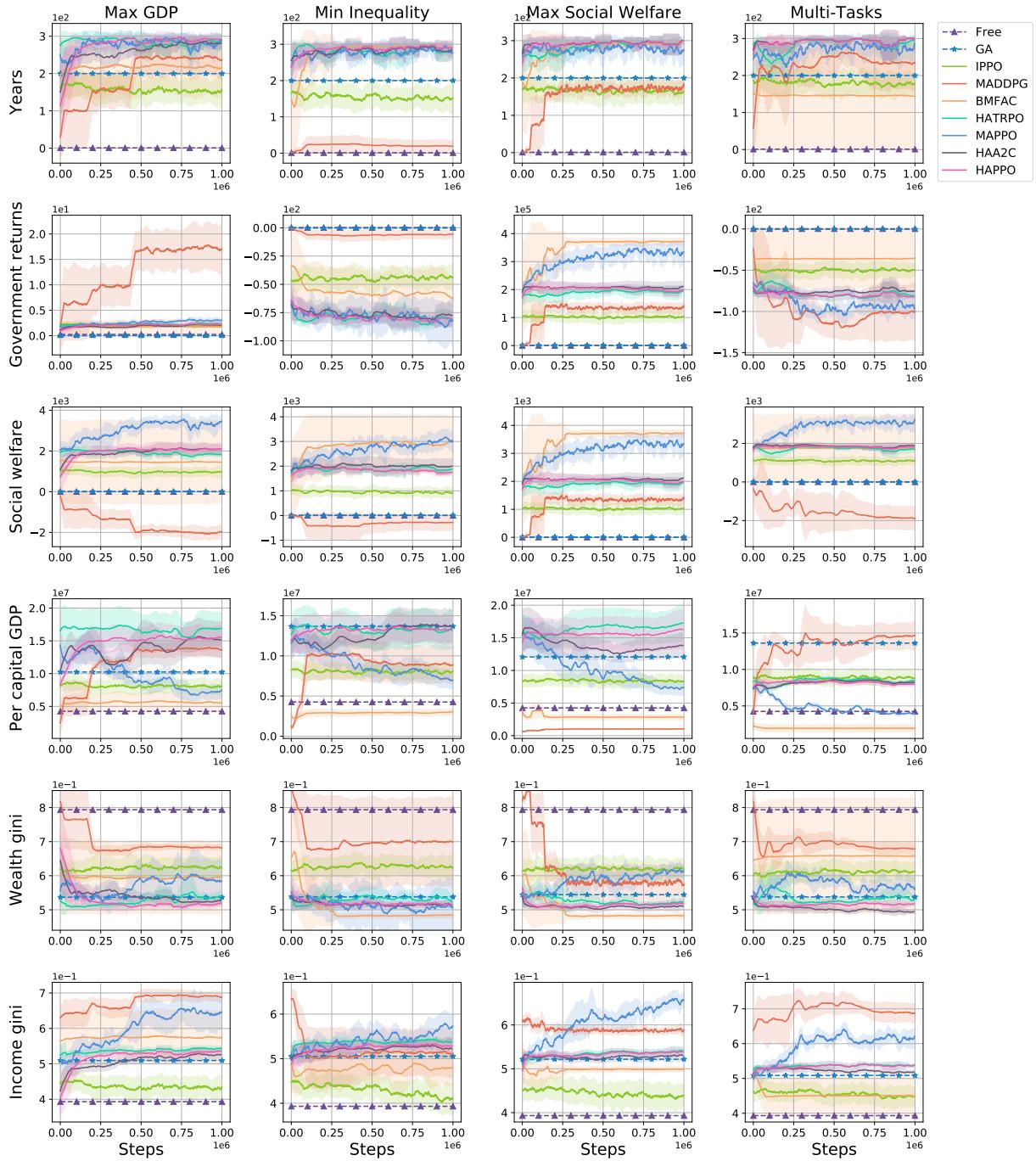
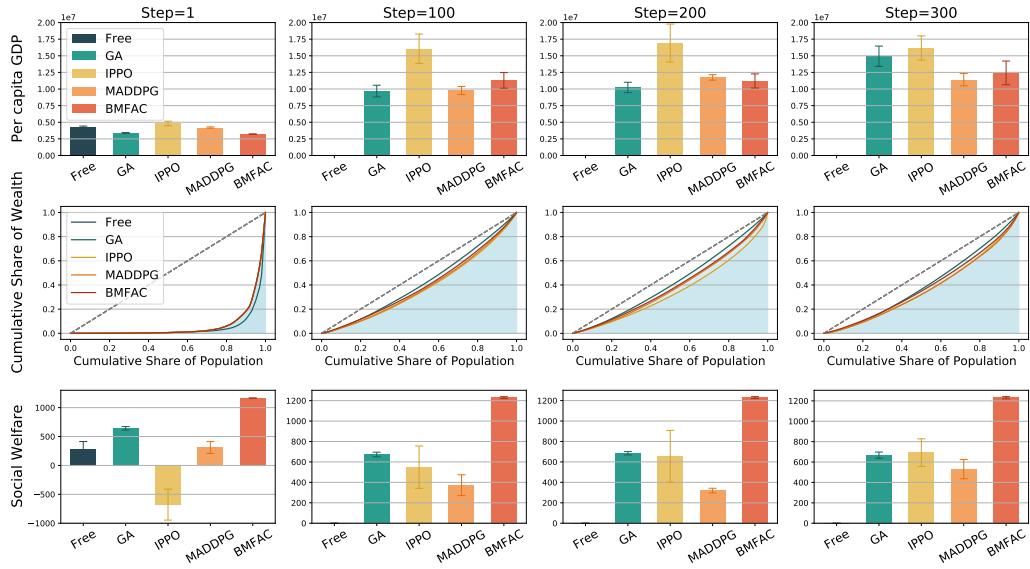


Figure 7: The training curves for 9 baselines on 6 economic indicators in 4 different tasks ( $N = 100$ ).



**Figure 8: The economic evolution process of different algorithms in the TaxAI environment. Each row represents a macroeconomic indicator, while each column represents a time step in the evolution.**

**Table 9: Public Hyperparameters of Baselines Algorithms**

Hyperparameter	Value
Discount factor $\gamma$	0.975
Critic learning rate	3e-4
Actor learning rate	3e-4
Replay buffer size	1e6
Init exploration steps	1000
Num of epochs	1500
Epoch length	500
Batch size	128
Tau $\tau$	5e-3
Evaluation epochs	10
Hidden size	128

**Table 10: Hyperparameters of IPPO Algorithm**

Hyperparameter	Value
Tau $\tau$	0.95
Gamma $\gamma$	0.99
Eps $\epsilon$	1e-5
Clip	0.1
Vloss coef	0.5
Ent coef	0.01
Max grad norm	0.5
Update frequency	1

**Table 11: Hyperparameters of MADDPG Algorithm**

Hyperparameter	Value
Update frequency	10
Noise rate	0.1
Epsilon	0.1
Save interval	100

**Table 12: Hyperparameters of Genetic Algorithm**

Hyperparameter	Value
DNA size	12
Pop size	100
Crossover rate	0.8
Mutation rate	0.1
Max generations	200

**Table 13: Hyperparameters of MAPPO, HAPPO, HATRPO, HAA2C Algorithm**

Hyperparameter	value	Hyperparameter	value	Hyperparameters	Value
Eval episode	20	Optimizer	Adam	Num mini-batch	1
Gamma	0.99	Optim eps	1e - 5	Batch size	4000
Gain	0.01	Hidden layer	1	Training threads	20
Std y coef	0.5	Actor network	MLP	Rollout threads	4
Std x coef	1	Max grad norm	10	Episode length	300
Activation	ReLU	Hidden layer dim	64	Critic lr	5e - 3
Actor lr	5e - 5	Gov actor lr	5e - 4	Gov critic lr	5e - 4