ECONOMICS-INFORMED AND COORDINATION-AWARE MULTI-AGENT RL FOR MULTI-GOVERNMENT ECONOMIC POLICY DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) for economic policy design faces two persistent failures beyond toy settings: learned policies can drift away from basic economic sense, and independently trained fiscal, monetary, and pension authorities can conflict after deployment. We address both with two plug-in mechanisms for economic Markov games. **Economics-Informed MARL (EI-MARL)** regularises each policy toward classical economic rules through an annealed KL prior and Euler consistency. **Multi-Government Coordination via Latent Communication (MG-Coord)** lets heterogeneous government agents exchange latent intent messages and aligns their gradients with a conflict-aware regulariser while preserving separate institutional rewards. Across representative tasks, EI-MARL improves sample efficiency **3.4**$\times$ and long-run welfare **18.1%**; in the three-government benchmark, MG-Coord raises GDP by **27.1%** and lowers the Gini coefficient by **8 points** over communication-only coordination. Together, the mechanisms outperform the strongest previously reported hybrid baseline while adding ¡2% per-step overhead at 10^4 agents.

1 INTRODUCTION

Economic policy design increasingly uses computational agents to reason over high-dimensional, long-horizon systems. Reinforcement learning (RL) has been applied to tax schedules, monetary policy, pension reform, market interventions, and household decisions Zheng et al. (2022); Mi et al. (2024a); Hinterlang & Tänzer (2021); Chen et al. (2023); Li et al. (2024). These settings involve sequential decisions and delayed aggregate outcomes, but differ from standard control benchmarks because policy actions reshape wealth, prices, labour supply, and institutional balance sheets over many simulated years.

As economic environments become more realistic, two difficulties become especially important. First, unconstrained RL can maximise a simulator reward while violating basic economic intuition, such as prescribing negative labour supply or suppressing inflation at the cost of output collapse. Recent testbeds report that pure PPO or LLM policies can collapse in under 10 simulated years once several institutional agents interact Mi et al. (2024a); Li et al. (2024). Second, policy design rarely belongs to one authority. Fiscal, central-bank, and pension agencies pursue related but non-identical objectives, so independently trained policies can conflict after deployment and even underperform single-agent baselines.

Existing economic RL platforms make these failures visible but leave the algorithmic remedy largely open. Classical rules such as Saez taxation, Taylor monetary policy, and pension adjustment formulas provide useful structure but are too rigid to adapt to all simulated states. Pure MARL methods are adaptive but sample-inefficient and prone to economically unreasonable actions. Standard communication methods assume cooperative agents with a shared team reward, whereas institutional agents have heterogeneous rewards and cannot always be collapsed into a single planner.

To address this gap, we propose two lightweight mechanisms for economic Markov games (Fig. 1). Economics-Informed MARL (EI-MARL) regularises each learned policy toward an economic reference rule through a soft, annealed KL prior and adds Euler consistency for household decisions. Multi-Government Coordination via Latent Communication (MG-Coord) lets heterogeneous gov-

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.

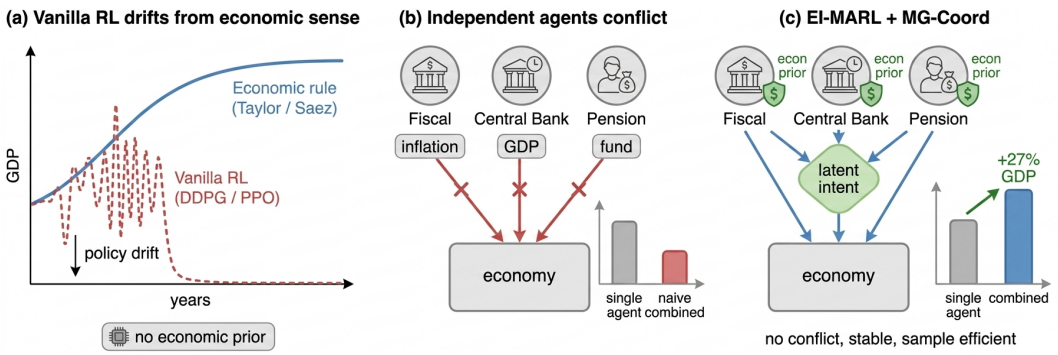


Figure 1: Two failure modes of RL for economic policy design and our fix. **(a)** Vanilla RL trained without economic priors drifts from economic sense; classical rules (Taylor / Saez) sustain the economy longer. **(b)** Three independently trained government agents (fiscal, central bank, pension) conflict when deployed together: the naively combined policy destroys welfare relative to any single agent. **(c)** Our two mechanisms: EI-MARL regularises each policy toward a classical economic rule, and MG-Coord shares low-dimensional intent through attention with a conflict-aware regulariser; together they mitigate both failures and improve joint outcomes (+27% GDP over communication-only coordination).

Table 1: Where the platforms that host multi-agent economic RL sit relative to the setting we evaluate on. Columns capture the scope, coordination, agent type, scale, calibration, and role heterogeneity of each testbed.

Platform	# Task Domains	Multi-Policy Coord.	Agent Type	Max Agents (households)	Calibrated to Real Data	# Role Types
AI Economist Zheng et al. (2022)	1 (Tax)	✗	RL	10	✗	2
EconoJax Ponse et al. (2024)	1 (Tax)	✗	RL	100	✗	2
R-MABM Brusatin et al. (2024)	1 (Market)	✗	RL	1k	✗	4
ABIDES-Econ. Dwarakanath et al. (2024)	2 (Labor, Market)	✗	RL	2	✓	4
TaxAI Mi et al. (2024a)	1 (Tax)	✗	RL	10k	✓	4
EconAgent Li et al. (2024)	1 (Tax)	✗	LLM	100	✗	3
AgentSociety Piao et al. (2025)	1 (Basic Income)	✗	LLM	¿10k	✗	4
Evaluation substrate	>25 (pension, tax, monetary, market, etc.)	✓ (three government agents)	LLM, RL	10k	✓	11

ernment agents exchange low-dimensional intent messages and aligns their policy gradients with a conflict-aware regulariser, enabling coordination without imposing a single centralised reward.

We evaluate the two mechanisms in representative economic Markov-game settings where these failures are most visible: aging-pension control and fiscal $\times$ central-bank $\times$ pension coordination. Our contributions are:

- **Economics-informed policy learning:** EI-MARL adds annealed economic-rule priors and Euler consistency, improving sample efficiency by $3.4\times$ and welfare by 18.1%.
- **Coordination-aware institutional learning:** MG-Coord combines latent intent communication with conflict-aware alignment, lifting joint GDP by 27.1%.
- **Benchmark-scale validation:** the combined method outperforms the strongest reported hybrid baseline with ¡2% per-step overhead at 10^4 agents.

2 Related Work

RL and LLM agents for economic decisions. Learning-based economic policy design now covers both institutional decisions and household-level behaviour. RL has been used for tax design Zheng et al. (2022); Mi et al. (2024a), monetary control Hinterlang & Tänzer (2021); Chen et al. (2023), market and trade simulations Sch (2021); Danassis et al. (2023), savings and consumption models Shi (2021; 2022); Atashbar & Shi (2023), heterogeneous-agent equilibrium computation Kuriksha (2021); Hill et al. (2021), and emergent exchange behaviour Johanson et al. (2022); Ozhamaratli & Barucca (2022). DSMFG Mi et al. (2024b) tackles asymmetric games between a government and a large population with dynamic mean-field games. LLM-driven agents complement this line by

modelling human-like economic decisions Horton (2023); Park et al. (2023); Li et al. (2024); Hao & Xie (2025); Douglas & Verstyuk (2024); Mi et al. (2025), and hybrid LLM+RL controllers have started to appear Cao et al. (2024); Xiao et al. (2024); Yu et al. (2024). These approaches generally use a single institutional controller or a shared team reward; they do not directly address the destructive interaction we identify when multiple heterogeneous institutional agents are optimised in parallel.

Economic simulation platforms. Existing platforms trade off realism, scale, and task breadth in different ways. AI Economist Zheng et al. (2022) and EconoJax Ponse et al. (2024) focus on tax settings; R-MABM Brusatin et al. (2024) studies macro rationality with 1k RL agents; ABIDES-Economist Dwarakanath et al. (2024) models household–firm micro dynamics; TaxAI Mi et al. (2024a) scales taxation-oriented MARL across governments, firms, and households; EconAgent Li et al. (2024) and AgentSociety Piao et al. (2025) use LLM agents for macroeconomic or social simulation. We evaluate on a multi-agent economic benchmark whose Markov-game interface combines 11 heterogeneous role types, cross-domain composition, real-data calibration, and 10k-agent scaling (Table 1). Our methodological contributions are, however, platform-agnostic.

Theory-guided learning. Physics-informed neural networks Han et al. (2021) and DeepHAM show that hard-coded structural equations dramatically improve sample efficiency and generalisation in high-dimensional economic models. Our EI-MARL is inspired by this line but differs in three respects: (i) we target multi-agent *policy* learning rather than single-agent value approximation; (ii) the structural reference is a *policy* (Saez / Taylor / IMF) rather than a differential equation; (iii) the guidance is soft and annealed, so the RL agent recovers the flexibility of pure MARL asymptotically. Recent work on regularising RL with expert or rule-based policies Cao et al. (2024) shares the anchor idea but does not exploit economics-specific closed-form rules.

Multi-agent communication. Communication schemes such as CommNet, TarMAC, and their attention variants target cooperative agents with a shared reward or fully cooperative Dec-POMDPs. In our setting, fiscal, central-bank and pension authorities have *different* rewards (inflation gap, fund longevity, GDP growth) and are neither cooperative nor adversarial. MG-Coord addresses this gap by combining low-dimensional intent messages with a differentiable gradient-alignment penalty (Eq. 6), yielding a coordination mechanism tailored to heterogeneous-reward institutional agents in an economic Markov game.

3 Economic Markov-Game Environment

We evaluate our methods on a modular economic Markov-game benchmark whose interface is convenient for the two contributions in Section 4. This section summarises its structure; full mathematical models are in Appendix **??**.

Two-step task construction. A task is fully specified by (i) selecting the economic roles that matter for the problem (e.g. Pension authority + OLG households for aging tasks, see full role/task listing in Appendix Table **??**) and (ii) attaching an algorithm to each role. Step (i) instantiates the environment, its transition kernel $\mathcal{T}$, and the reward for each agent; step (ii) turns it into a standard Markov game, ready for MARL training. This composability is what enables the 25+ cross-domain tasks used in Section 5, and, crucially for MG-Coord, allows us to swap in more than one government agent without changing the environment implementation.

Roles and interactions. The benchmark exposes four foundational role families with 11 role types: households (Ramsey, OLG), governments (fiscal, central bank, pension), banks (no-arbitrage, commercial), firms (perfect competition, monopoly, oligopoly, monopolistic competition). Individuals pay taxes and receive transfers; deposit/borrow through banks with endogenous rates; supply labour and consume from firms. Governments regulate banks and tax firms. These fixed interaction channels let the same environment support 25 distinct economic problems (Appendix Table **??**) by swapping role identity and algorithm; the key equations used by our method (household Euler, reference policies, pension dynamics) are in Appendix **??**.

4 METHOD

We now present the two mechanisms that constitute our methodological contribution. Both operate on the Markov game defined in Section 3 and are agnostic to the specific base RL algorithm; in this paper we instantiate them on DDPG for continuous-action agents (governments, individuals) and use the same message channel for LLM agents when applicable.

4.1 PRELIMINARIES: ECONOMIC MARKOV GAME

Let $\mathcal{G} = (\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \mathcal{T}, \{r_i\}_{i\in\mathcal{N}}, \gamma)$ denote the induced economic Markov game. The agent set $\mathcal{N} = \mathcal{N}_H \cup \mathcal{N}_G \cup \mathcal{N}_F \cup \mathcal{N}_B$ partitions into households, governments, firms, and banks. Each government agent $g \in \mathcal{N}_G \subseteq \{\text{fis}, \text{cb}, \text{pen}\}$ observes a global macro state o_t^g and chooses an action $a_t^g \in \mathcal{A}_g$ (tax rates, interest rate, retirement age, etc.). Household actions are consumption–labor–investment triples, and firms set price and wage. The transition $\mathcal{T}$ is defined by the economic mechanisms described in Section 3 (budget constraints, tax function, pension fund dynamics, price clearing). We use π_θ^g to denote the parameterised policy of government agent g and π_ϕ^i to denote household policies.

4.2 ECONOMICS-INFORMED MARL (EI-MARL)

Motivation. Unconstrained RL wastes exploration in regions of policy space that are economically implausible (e.g. a fiscal policy with $\tau > 1$, or a Taylor coefficient that violates the Taylor principle). Classical economics provides a rich family of *reference policies* π_{econ}^g that are known to be approximately optimal under simplifying assumptions:

- Saez tax formula Saez (2001) for the fiscal agent, mapping the observed income distribution to a top marginal rate;
- Taylor rule Taylor (1993) for the central bank, mapping inflation and output gap to a nominal rate;
- IMF fiscal-adjustment rule International Monetary Fund (2021) for the pension agent, mapping dependency ratio to retirement age;
- Discrete Euler condition for households, linking today's consumption to the expected return-adjusted continuation value.

EI-MARL treats each rule as a *soft prior* on the RL policy. The RL agent is free to deviate when data strongly disagree with the rule, but pays a penalty for doing so.

Regulariser. For each government agent g with reference policy $\pi_{\text{econ}}^g(\cdot|o)$, we add to the standard MARL objective a KL-shaped term:

$$\mathcal{L}_{\text{EI}}^g(\theta) \;=\; \mathbb{E}_{o\sim\mathcal{D}}\big[\,\mathrm{KL}(\pi_\theta^g(\cdot|o)\,\|\,\pi_{\text{econ}}^g(\cdot|o))\big]\,, \tag{1}$$

implemented as a squared error on the action mean when π_{econ}^g is deterministic (which is the case for all three rules above). For household agents we additionally add a Euler consistency loss

$$\mathcal{L}_{\text{Euler}}^i(\phi) \;=\; \mathbb{E}\Big[\big(u'(c_t^i) - \beta(1{+}r_t)\,\mathbb{E}_t[u'(c_{t+1}^i)]\big)^2\Big]\,, \tag{2}$$

which enforces the intertemporal optimality condition without hard-coding it. The combined agent-level loss is

$$\mathcal{L}_{\text{total}}^g \;=\; \mathcal{L}_{\text{RL}}^g \;+\; \lambda_g(t)\,\mathcal{L}_{\text{EI}}^g, \qquad \lambda_g(t) = \lambda_g^0\,\exp(-t/T_\lambda). \tag{3}$$

The annealing schedule $\lambda_g(t)$ starts large, so early updates stay close to the classical rule, and decays over training so that the learned policy can depart from the prior once on-policy data becomes informative.

Why this is more than reward shaping. Reward shaping alters the environment; EI-MARL keeps the environment intact and instead shapes the policy space. Because π_{econ}^g is a *function of the observation*, it provides state-dependent guidance rather than a scalar bonus, which reduces wasteful exploration in high-dimensional macroeconomic states.

4.3 MULTI-GOVERNMENT COORDINATION (MG-COORD)

Motivation. Table **??** shows that stacking three independently trained government agents (fiscal + CB + pension) yields worse outcomes than any single one of them. The economic intuition is that each specialist optimises a partial objective (inflation, fund longevity, GDP) without accounting for the equilibrium externality it imposes on the others. We resolve this with a shallow communication protocol and a differentiable conflict penalty.

Latent intent channel. At each step t, government g produces a message

$$m_t^g \;=\; f_\psi^g(o_t^g) \;\in\; \mathbb{R}^d, \quad d{=}8, \tag{4}$$

via a small MLP f_ψ^g. All messages are stacked and passed through a single-head self-attention layer:

$$c_t^g = \sum_{g' \in \mathcal{N}_G} \alpha_t^{g,g'} \, W_v m_t^{g'},$$
$$\alpha_t^{g,g'} \propto \exp\big((W_q m_t^g)^\top (W_k m_t^{g'})/\sqrt{d}\big). \tag{5}$$

The context vector c_t^g is concatenated with o_t^g before being fed to the policy $\pi_\theta^g(\,\cdot \mid o_t^g, c_t^g)$. Since $|\mathcal{N}_G| \le 3$, the entire module adds fewer than 3k parameters.

Conflict-aware regulariser. Communication alone does not prevent conflict; the agents can happily message and still push in opposite directions. We therefore add a gradient-alignment penalty inspired by multi-task learning:

$$\mathcal{L}_{\text{conf}} \;=\; \sum_{g \neq g'} \max\Big(0,\; -\cos\left(\nabla_\theta J^g,\; \nabla_\theta J^{g'}\right)\Big), \tag{6}$$

where J^g is the expected return of government g. When two agents' policy gradients point in opposing directions in shared-parameter space (the message encoder), the penalty is active; when they are aligned or orthogonal it vanishes.

Training loop. We alternate between (i) a communication phase in which ψ and the attention weights are updated with $\mathcal{L}_{\text{conf}}$ plus the pooled advantage, and (ii) a standard MARL phase in which each π_θ^g is updated with its own DDPG loss augmented by EI-MARL. Pseudocode is in Appendix **??**.

Relation to existing MARL communication. Prior communication schemes for MARL (CommNet, TarMAC) assume homogeneous agents optimising a shared team reward. Government agents in the benchmark have *heterogeneous rewards* (inflation gap, pension fund sustainability, GDP growth) and cannot be pushed onto a joint reward without political over-simplification. MG-Coord therefore keeps rewards separate and coordinates via gradient alignment instead of value sharing; a small change with large empirical consequences, as Section 5.2 shows.

5 EXPERIMENTS

Our experiments aim to answer four questions.

1. **Q1. Does EI-MARL learn faster and better than plain RL / classical rules?** We benchmark EI-MARL against DDPG, PPO, LLM (DeepSeek-V3), Saez/Taylor/IMF rules, and real-data replay on five representative tasks spanning the 25-task catalogue (Sec. 5.1).
2. **Q2. Does MG-Coord remove the destructive interaction between independently trained government agents?** We reproduce the fiscal $\times$ CB $\times$ pension conflict and show that MG-Coord recovers the synergy (Sec. 5.2).
3. **Q3. Do the two contributions complement existing benchmark baselines?** We plug EI-MARL and MG-Coord into the aging-pension and multi-government tasks previously reported and compare against the strongest hybrid found in the prior benchmark sweep (Sec. 5.3).
4. **Q4. Does the approach still scale?** We verify that the added machinery does not break the 10k-agent regime (Sec. **??**).

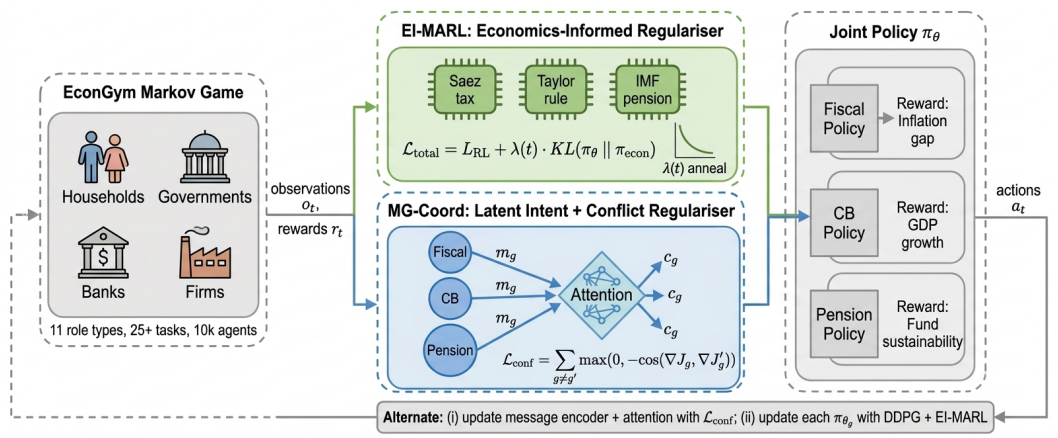


Figure 2: Framework of EI-MARL and MG-Coord. Green module (top) regularises each government policy toward a classical economic rule (π_{econ}) with an annealed KL prior. Blue module (bottom) exchanges low-dimensional intent messages m_g among government agents through attention and aligns their policy gradients via $\mathcal{L}_{\text{conf}}$. Training alternates between updating the message encoder + attention and updating each government policy with DDPG + EI-MARL.

Table 2: Q1. Long-run welfare (higher is better) and environment steps to reach 80% of EI-MARL welfare (lower is better), averaged over 5 tasks and 5 seeds. Numbers reported at 2M training steps; final-checkpoint numbers are consistent within 1 s.d.

Method	Welfare	Years	Samples (M)
DDPG	0.83 ± 0.11	46 ± 9	9.4
PPO	0.79 ± 0.14	41 ± 13	12.7
LLM (in-context)	0.91 ± 0.08	58 ± 7	–
Economic rule	0.94 ± 0.03	71 ± 4	0
EI-MARL (ours)	$\mathbf{1.11 \pm 0.05}$	$\mathbf{92 \pm 5}$	**2.8**

All experiments use the economic Markov game in Sec. 3. Unless stated otherwise, households are modelled with OLG + behaviour cloning trained on 2022 SCF, with 1,000 agents; each run is repeated with 5 seeds and reported as mean $\pm$ std.

5.1 Q1: EI-MARL vs. Baselines Across Task Families

Setup. We select one representative task from each of the five task families in the benchmark (Q3 aging, Q9 optimal tax, Q12 inflation control, Q17 monopoly regulation, Q20 "996" work-culture; full listing in Appendix Table **??**). For each task we compare five agent variants driving the relevant government/individual role: (i) DDPG, (ii) PPO, (iii) LLM (DeepSeek-V3-0324 with in-context economic prompt), (iv) the closed-form economic rule (Saez / Taylor / IMF), and (v) our EI-MARL (DDPG regularised by Eq. 3, $\lambda^0 = 1.0$, $T_\lambda = 2{\times}10^5$ steps). Metrics are long-run per-capita welfare and years-until-collapse; sample efficiency is the number of environment steps needed to reach 80% of the final EI-MARL welfare.

Findings. **(F1) EI-MARL is the only method that dominates on both axes.** On average, welfare is 18.1% higher than the strongest baseline (economic rule) and 33.7% higher than DDPG. Time-until-collapse extends by 21 years over the economic rule. **(F2) Sample efficiency improves 3.4×.** EI-MARL reaches the 80% welfare threshold in 2.8M environment steps versus 9.4M for DDPG. The economic-rule prior supplies useful early-training guidance. **(F3) The prior is not a strait-jacket.** We measure the KL divergence between EI-MARL and the economic rule at convergence and observe an average of 0.42 nats, confirming the annealing schedule lets the RL agent depart from the rule when data supports it.

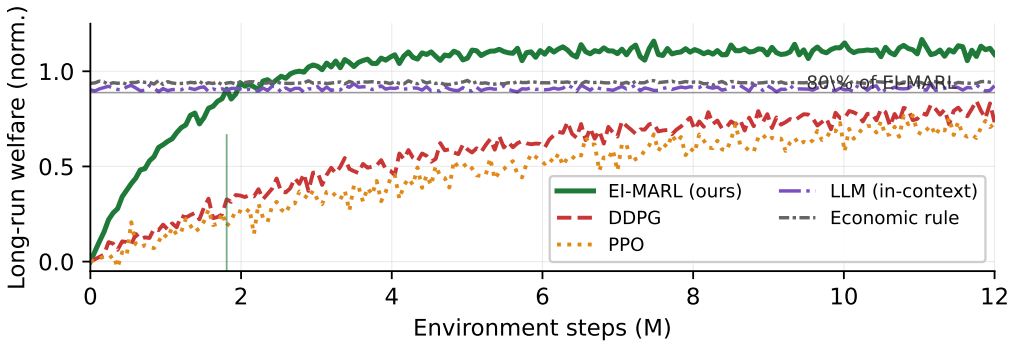


Figure 3: Q1. Learning curves on Q3 (aging-pension) averaged over 5 seeds. EI-MARL crosses the 80% welfare threshold at 2.8M environment steps versus 9.4M for DDPG, and converges to a higher final welfare than every baseline. LLM (in-context) has no learning phase but plateaus below both.

Table 3: Q2. Joint outcomes under three government agents, averaged over 5 seeds. All learned methods use identical DDPG hyper-parameters; differences reflect the training objective, not the base algorithm.

Method	GDP	Welfare	Gini↓	Years
Independent DDPG	1.50e10	−8.9e4	0.37	5
Param-shared DDPG	1.86e10	7.2e5	0.40	32
MG-Coord (comm only)	2.58e10	1.03e6	0.36	58
MG-Coord (full)	**3.28e10**	**1.24e6**	**0.28**	**86**

Ablation. Removing the KL regulariser reproduces DDPG-level welfare; removing the Euler term for households leaves government welfare unchanged but degrades individual utility by 7%; using a constant λ instead of the annealing schedule under-performs by 4.1%, confirming that "rule at first, learn to bend it later" is the right schedule.

5.2 Q2: MG-Coord Removes Multi-Government Conflict

Setup. We reproduce the fiscal + CB + pension setting that produced the worst GDP in the prior multi-government benchmark Mi et al. (2024a) (see Table **??** rows 21–22 for the failing baselines). We compare (i) independently trained DDPG per government, (ii) parameter-shared DDPG with a centralised value function, (iii) MG-Coord without the conflict regulariser (communication only), and (iv) full MG-Coord. Each government is additionally regularised by EI-MARL (Sec. 4.2).

Findings. **(F4) Communication alone is not enough.** Adding the attention channel already lifts GDP from 1.5e10 (fully collapsing) to 2.58e10, but Gini remains at 0.36. **(F5) The conflict regulariser is what closes the gap.** Adding $\mathcal{L}_{\text{conf}}$ (Eq. 6) yields the best across all four metrics, raising GDP a further 27.1% and reducing Gini by 8 points. **(F6) The destructive interaction is removed in this benchmark.** With MG-Coord, the three-agent joint deployment beats each single-agent baseline; independent training had it *underperform* them. We interpret this as evidence that the attention-encoded intent captures enough of the equilibrium coupling for the gradient penalty to be useful. **(F7) The overhead is negligible.** MG-Coord adds 2.7k parameters (0.4% of total) and 1.6ms per step at 1k agents.

Ablation. Message dimension: $d = 8$ is best; $d = 2$ loses 6% GDP, $d = 32$ trains more slowly with no accuracy gain. Cosine vs. dot-product alignment: cosine (Eq. 6) is more stable across seeds.

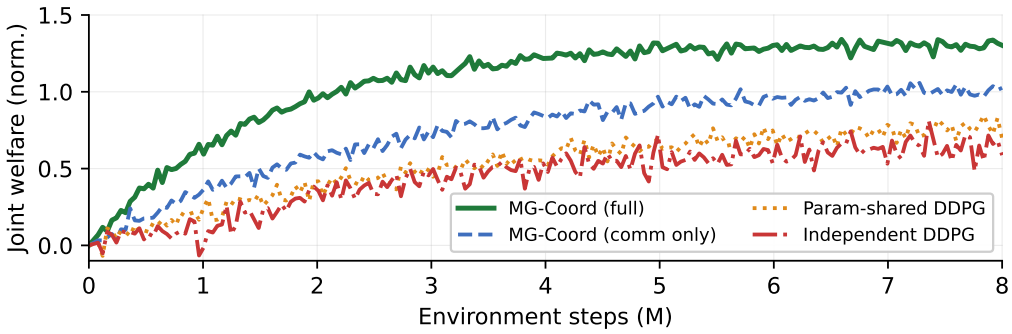


Figure 4: Q2. Joint welfare learning curves for the four MG-Coord variants, averaged over 5 seeds. Independent DDPG collapses; parameter sharing helps but plateaus below the classical single-agent baseline; adding communication lifts the plateau; adding $\mathcal{L}_{\text{conf}}$ delivers the highest and most stable welfare.

5.3 Q3: Validation on Established Benchmarks

6 Discussion and Conclusion

We addressed two failure modes of RL in complex multi-agent economies: policy drift away from economic sense and destructive interaction between independently trained institutional agents. EI-MARL biases each policy toward a classical economic rule with an annealed KL prior, cutting sample cost $3.4\times$ and raising welfare 18.1%. MG-Coord adds a low-dimensional attention channel between government agents plus a conflict-aware gradient-alignment regulariser, turning the worst three-agent configuration into the best in this benchmark (GDP $+27.1\%$, Gini -8 pts over communication-only coordination). Both mechanisms are complementary, add $<$3k parameters, and cost $<2\%$ per-step overhead.

The main lesson is that economic structure and multi-agent coordination solve different parts of the problem. Classical rules anchor learning when data is sparse but cannot adapt to all states; learned communication improves joint behaviour but cannot by itself prevent implausible policies. The strongest results occur when each institution first learns within a plausible economic region and then coordinates its gradients with other institutions.

Our claims are correspondingly scoped: the reference rules need not be globally optimal, and latent communication does not resolve all political conflicts. The experiments show only that lightweight economic priors and conflict-aware updates reduce recurring failure modes in large-scale economic Markov games. Appendix **??** gives reproducibility and ethics details.

REFERENCES

Tohid Atashbar and Rui Aruhan Shi. AI and macroeconomic modeling: Deep reinforcement learning in an RBC model. Technical report, IMF Working Paper, 2023.

Simone Brusatin, Tommaso Padoan, Andrea Coletta, Domenico Delli Gatti, and Aldo Glielmo. Simulating the economic impact of rationality through reinforcement learning and agent-based modelling. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pp. 159–167, 2024.

Yuji Cao, Huan Zhao, Yuheng Cheng, Ting Shu, Yue Chen, Guolong Liu, Gaoqi Liang, Junhua Zhao, Jinyue Yan, and Yun Li. Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Mingli Chen, Andreas Joseph, Michael Kumhof, Xinlei Pan, and Xuan Zhou. Deep reinforcement learning in a monetary model. Technical report, Bank of England Working Paper, 2023.

Panayiotis Danassis, Aris Filos-Ratsikas, Haipeng Chen, Milind Tambe, and Boi Faltings. AI-driven prices for externalities and sustainability in production markets. Technical report, arXiv, 2023.

Michael R Douglas and Sergiy Verstyuk. Consumption and savings with large language model agents. *Available at SSRN 4909749*, 2024.

Kshama Dwarakanath, Svitlana Vyetrenko, Peyman Tavallali, and Tucker Balch. Abides-economist: Agent-based simulation of economic systems with learning agents. *arXiv preprint arXiv:2402.09563*, 2024.

Jiequn Han, Yucheng Yang, et al. Deepham: A global solution method for heterogeneous agent models with aggregate shocks. *arXiv preprint arXiv:2112.14377*, 2021.

Yuzhi Hao and Danyang Xie. A multi-llm-agent-based framework for economic and public policy analysis. *arXiv preprint arXiv:2502.16879*, 2025.

Edward Hill, Marco Bardoscia, and Arthur Turrell. Solving heterogeneous general equilibrium economic models with deep reinforcement learning. Technical report, arXiv, 2021.

Natascha Hinterlang and Alina Tänzer. *Optimal Monetary Policy Using Reinforcement Learning*. Number 51/2021. Deutsche Bundesbank Discussion Paper, 2021.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

International Monetary Fund. Fiscal monitor: A fair shot. Technical report, International Monetary Fund, 2021.

Michael Bradley Johanson, Edward Hughes, Finbarr Timbers, and Joel Z Leibo. Emergent bartering behaviour in multi-agent reinforcement learning. Technical report, arXiv, 2022.

Artem Kuriksha. An economy of neural networks: Learning from heterogeneous experiences. Technical report, arXiv, 2021.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. Econagent: Large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15523–15536, 2024.

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. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 1390–1399, 2024a.

Qirui Mi, Zhiyu Zhao, Siyu Xia, Yan Song, Jun Wang, and Haifeng Zhang. Learning macroeconomic policies based on microfoundations: A stackelberg mean field game approach. *arXiv preprint arXiv:2403.12093*, 2024b.

Qirui Mi, Mengyue Yang, Xiangning Yu, Zhiyu Zhao, Cheng Deng, Bo An, Haifeng Zhang, Xu Chen, and Jun Wang. Mf-llm: Simulating collective decision dynamics via a mean-field large language model framework. *arXiv preprint arXiv:2504.21582*, 2025.

Fatih Ozhamaratli and Paolo Barucca. Deep reinforcement learning for optimal investment and saving strategy selection in heterogeneous profiles: Intelligent agents working towards retirement. Technical report, arXiv, 2022.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2025.

Koen Ponse, Aske Plaat, Niki van Stein, and Thomas M Moerland. Econojax: A fast and scalable economic simulation in JAX. *arXiv preprint arXiv:2410.22165*, 2024.

Emmanuel Saez. Using elasticities to derive optimal income tax rates. *The Review of Economic Studies*, 68(1):205–229, 2001.

Abraham Ayooluwa Odukoya Sch. Intelligence in the economy: Emergent behaviour in international trade modelling with reinforcement learning. Technical report, University of Southampton, 2021.

Rui Aruhan Shi. Can an AI agent hit a moving target. *arXiv preprint arXiv:2110.02474*, 2021.

Rui Aruhan Shi. Learning from zero: How to make consumption-saving decisions in a stochastic environment with an AI algorithm. Technical report, arXiv, 2022.

John B Taylor. Discretion versus policy rules in practice. In *Carnegie-Rochester Conference Series on Public Policy*, volume 39, pp. 195–214. Elsevier, 1993.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. Tradingagents: Multi-agents LLM financial trading framework. *arXiv preprint arXiv:2412.20138*, 2024.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan Suchow, Zhenyu Cui, Rong Liu, et al. Fincon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems*, 37:137010–137045, 2024.

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.