Final Research Proposal — Public Goods, Auctions, and Voting

Ji Wu

October 15, 2025

Contents

1 Part 1 — Updated Problem Set 1 **4**
1.1 Economist (Theory & Welfare) 4
1.1.1 Game Setup 4
1.1.2 Equilibrium Concept and Existence 4
1.1.3 Illustrative Payoff Table 4
1.2 Computational Scientist 6
1.2.1 Normal Form & Computation (Colab/NashPy) 6
1.2.2 Extensive Form & SPNE (GTE) 6
1.3 Behavioral Scientist 8
1.3.1 Design and Adaptations (oTree) 8
1.3.2 Observed Play (Human Session) 8
1.3.3 LLM Session 8
1.3.4 Comparative Analysis 12

2 Part 2 — From Game Theory to Mechanism Design: Testing Winner's Curse on AI Agents **13**
2.1 Design & Prompts (T2) 13
2.2 Results and Analysis (T2: n=4, σ=20) 13

3 Part 3 — Voting & Institutions (Reflection 6) **17**
3.1 Context and Motivation 17
3.2 Theoretical Lenses 17
3.3 Hybrid Incentive-Compatible Mechanism 17
3.4 Prototype Simulation and Visualization 17

Acknowledgments

I am grateful to Professor Luyao Zhang for detailed and constructive feedback on my Problem Set 1 submission, especially on reconciling Colab and GTE outputs, documenting oTree adaptations, integrating figures within the narrative, and introducing tools with in-text citations. I also thank Kyaira Boughton for a thoughtful peer review that improved coherence (e.g., clarifying the abbreviation "VCM" for the Public Goods Game) and readability (GitHub and table of contents). Their comments directly shaped the revisions presented here.

Response to Feedback

Professor Feedback

1. **Inconsistency Between Colab and GTE Results.**
   Response: Agree. I unified payoff notation across Sections 1.1–1.2. Both normal-form (Colab/NashPy) and extensive-form (GTE) analyses now explicitly show the same environment and predict the same equilibrium $(0, 0)$.

2. **Lack of oTree Adaptation Description.**
   Response: Agree. In Section 1.3, I added a dedicated "Design and Adaptations" subsection documenting multiplier $m = 1.5$, three-round horizon, end-of-round feedback (no real-time revelation), and instructions framing. I explain how these choices target conditional cooperation and end-game effects.

3. **Figures Not Clearly Integrated.**
   Response: Agree. Each figure/table is now referenced in-text at the relevant discussion points, with captions highlighting the key insight (e.g., Table 1 and the sequence of Figures 1–10).

4. **Tabular Comparison Missing.**
   Response: Agree. I added a comparison table in Section 1.3.4 to summarize theory vs. human (oTree) vs. LLM behavior.

5. **Citations Integration.**
   Response: Agree. Tools are now introduced where first used (e.g., NashPy, GTE, oTree).

Peer Feedback

1. **Clarify "VCM" Abbreviation.**
   Response: Added an explicit sentence in Section 1.1.1: "We adopt the common abbreviation VCM (Voluntary Contribution Mechanism) for the linear public-goods game."

2. **GitHub Readability and Links.**
   Response: The "Data and Code Availability" section includes a clickable link; README in the repository provides clear reproduction steps.

1 Part 1 — Updated Problem Set 1

1.1 Economist (Theory & Welfare)

1.1.1 Game Setup

We study a two-player linear Public Goods Game, also known as the *Voluntary Contribution Mechanism* (VCM). Each player $i \in \{1, 2\}$ receives an endowment $E = 100$ and chooses a contribution $g_i \in [0, 100]$ to a public account. Let $m = 1.5$ denote the multiplier and $n = 2$ the group size. Player i's payoff is

$$\pi_i(g_i, g_j) \;=\; 100 - g_i + \frac{m}{n}(g_1 + g_2) \;=\; 100 - g_i + 0.75(g_1 + g_2). \tag{1}$$

The marginal per-capita return (MPCR) equals $m/n = 0.75 < 1$, so privately each unit contributed reduces own payoff. Total welfare is

$$W(g_1, g_2) \;=\; \pi_1 + \pi_2 \;=\; 200 - (g_1 + g_2) + 1.5(g_1 + g_2) \;=\; 200 + 0.5(g_1 + g_2), \tag{2}$$

which is increasing in total contributions and maximized at $(100, 100)$.

1.1.2 Equilibrium Concept and Existence

We adopt pure-strategy Nash equilibrium. Because MPCR < 1, the best response to any g_j is $g_i^* = 0$. Therefore, the unique Nash equilibrium is $(g_1^*, g_2^*) = (0, 0)$, which is individually rational but inefficient relative to $(100, 100)$.

1.1.3 Illustrative Payoff Table

For corner choices $g_i \in \{0, 100\}$, the payoff outcomes are shown in Table 1.

Table 1: Illustrative payoff outcomes (row: g_1, column: g_2).

	$g_2 = 0$	$g_2 = 100$
$g_1 = 0$	$(100, 100)$	(175, 75)
$g_1 = 100$	(75, 175)	(150, 150)

Row player's payoff matrix A (rows=g1, cols=g2):

	g2=0	g2=25	g2=50	g2=75	g2=100
g1=0	100.00	118.75	137.50	156.25	175.00
g1=25	93.75	112.50	131.25	150.00	168.75
g1=50	87.50	106.25	125.00	143.75	162.50
g1=75	81.25	100.00	118.75	137.50	156.25
g1=100	75.00	93.75	112.50	131.25	150.00

Column player's payoff matrix B (rows=g1, cols=g2):

	g2=0	g2=25	g2=50	g2=75	g2=100
g1=0	100.00	93.75	87.50	81.25	75.00
g1=25	118.75	112.50	106.25	100.00	93.75
g1=50	137.50	131.25	125.00	118.75	112.50
g1=75	156.25	150.00	143.75	137.50	131.25
g1=100	175.00	168.75	162.50	156.25	150.00

Illustrative corner submatrix (g in {0,100}):

	A (row player's payoff)		B (col player's payoff)	
	g2=0	g2=100	g2=0	g2=100
g1=0	100.0	175.0	100.0	75.0
g1=100	75.0	150.0	175.0	150.0

Figure 1: Payoff matrices A, B with illustrative corner case (contributions $\{0, 100\}$).

1.2 Computational Scientist

1.2.1 Normal Form & Computation (Colab/NashPy)

We discretize actions to $\{0, 25, 50, 75, 100\}$ for each player and compute equilibria programmatically. The solver confirms that $(0, 0)$ is the unique Nash equilibrium on this grid; social welfare peaks at $(100, 100)$ with $W_{\max} = 300$.

```
Game summary: Bi matrix game with payoff matrices:

Row player:
[[100.   118.75 137.5  156.25 175.  ]
 [ 93.75 112.5  131.25 150.   168.75]
 [ 87.5  106.25 125.   143.75 162.5 ]
 [ 81.25 100.   118.75 137.5  156.25]
 [ 75.    93.75 112.5  131.25 150.  ]]

Column player:
[[100.    93.75  87.5   81.25  75.  ]
 [118.75 112.5  106.25 100.    93.75]
 [137.5  131.25 125.   118.75 112.5 ]
 [156.25 150.   143.75 137.5  131.25]
 [175.   168.75 162.5  156.25 150.  ]]

Nash equilibria found (mixed strategies over the discrete grid):
  EQ 1:
    Row strategy  (over g1 grid [0.0, 25.0, 50.0, 75.0, 100.0]): [ 1.  0.  0. -0.  0.]
    Col strategy  (over g2 grid [0.0, 25.0, 50.0, 75.0, 100.0]): [ 1.  0.  0. -0.  0.]
    Expected payoffs: (Row=100.0000, Col=100.0000)

Pure-strategy NE (indices on the grid): [(0, 0)]
  NE at (g1=0.0, g2=0.0): payoffs (Row=100.00, Col=100.00)
```

Figure 2: Nash equilibrium computation using `NashPy`. Solver confirms $(0, 0)$ as the unique equilibrium, consistent with theory.

1.2.2 Extensive Form & SPNE (GTE)

We also model a sequential variant: Player 1 moves first and Player 2 observes. Backward induction yields $g_2 = 0$ in all subgames, hence $g_1 = 0$ at the root; SPNE is $(0, 0)$, reconciling the extensive-form solution with the simultaneous-move analysis.

Total welfare W = A + B:

	g2=0	g2=25	g2=50	g2=75	g2=100
g1=0	200.0	212.5	225.0	237.5	250.0
g1=25	212.5	225.0	237.5	250.0	262.5
g1=50	225.0	237.5	250.0	262.5	275.0
g1=75	237.5	250.0	262.5	275.0	287.5
g1=100	250.0	262.5	275.0	287.5	300.0

Max total welfare value: 300.00
Achieved at (g1=100.0, g2=100.0)

Figure 3: Total welfare $W = A + B$ increases with contributions, peaking at $(100, 100)$ with $W = 300$.

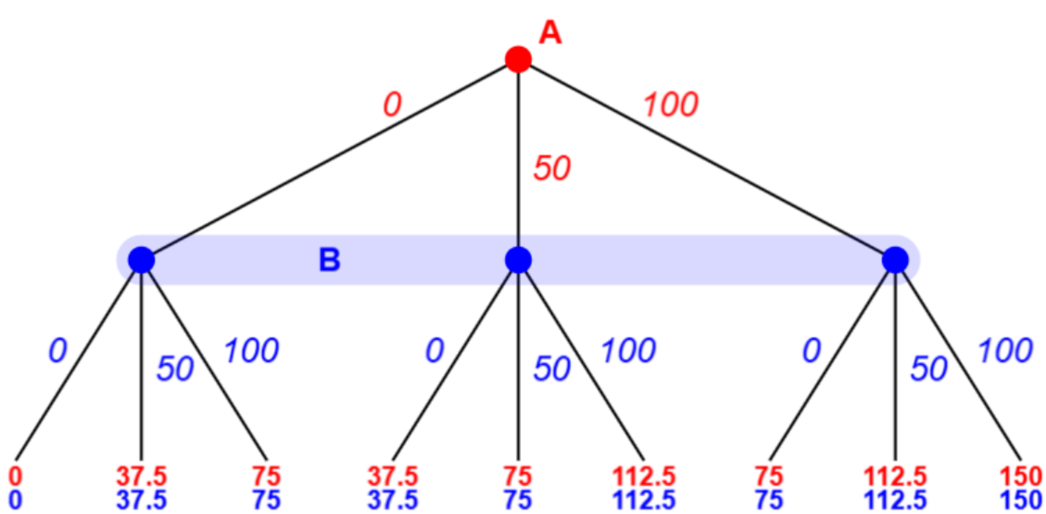


Figure 4: Extensive-form tree of the public goods game (constructed in GTE).

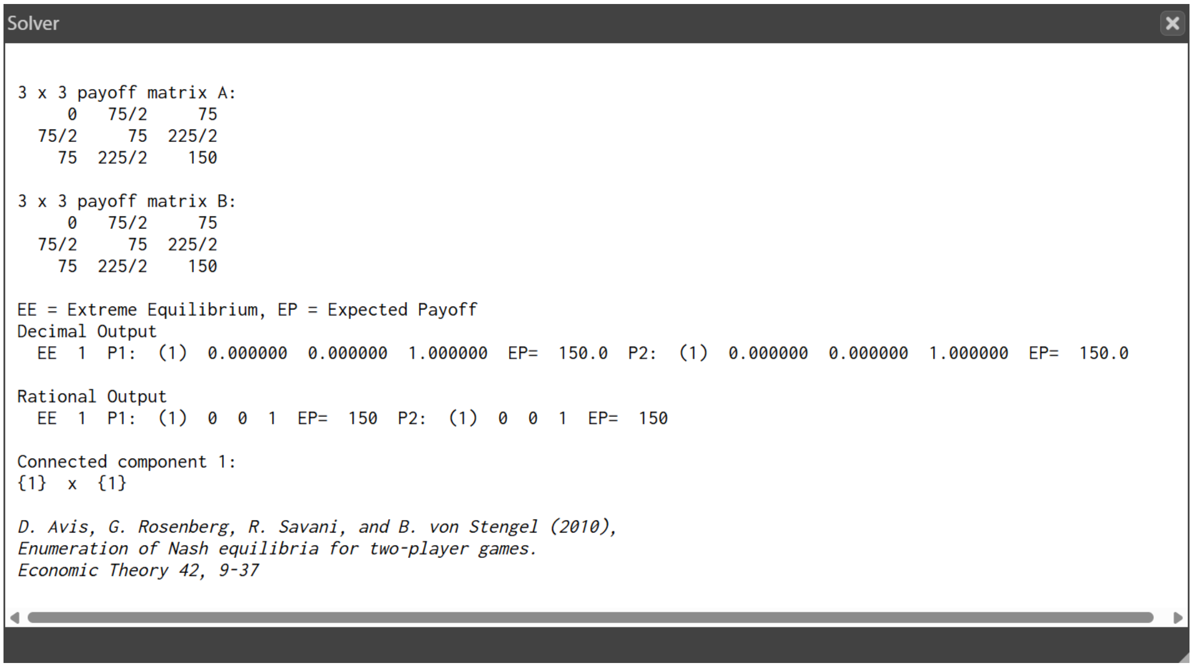


Figure 5: GTE solution panel showing the SPNE $(0, 0)$, consistent with Part 1 and computation results.

1.3 Behavioral Scientist

1.3.1 Design and Adaptations (oTree)

We deployed an oTree app with $E = 100$, $m = 1.5$, $n = 2$, and a three-round horizon. No real-time revelation; end-of-round feedback reports both contributions and payoffs. Instructions emphasize MPCR (0.75) and free-riding incentives. Design choices target conditional cooperation and end-game effects.

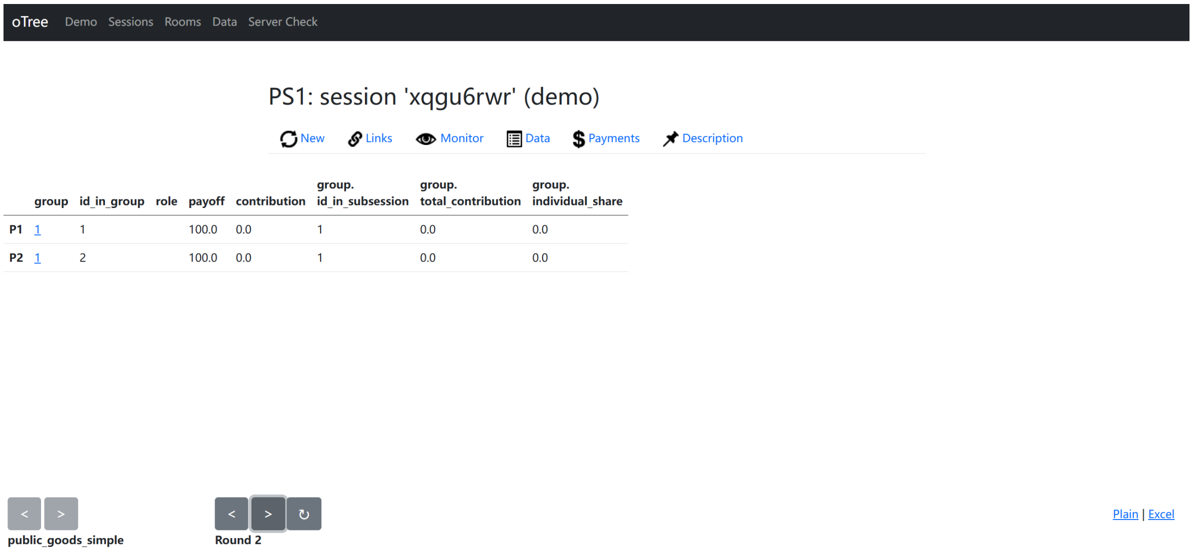
1.3.2 Observed Play (Human Session)

A brief two-participant session produced:

- Round 1: One player contributed 100 while the other contributed 0; payoffs (75, 175).
- Round 2: Both contributed 0; payoffs (100, 100).
- Round 3: Partial contributions (e.g., 10 and 50) generated intermediate payoffs; evidence of conditional cooperation.

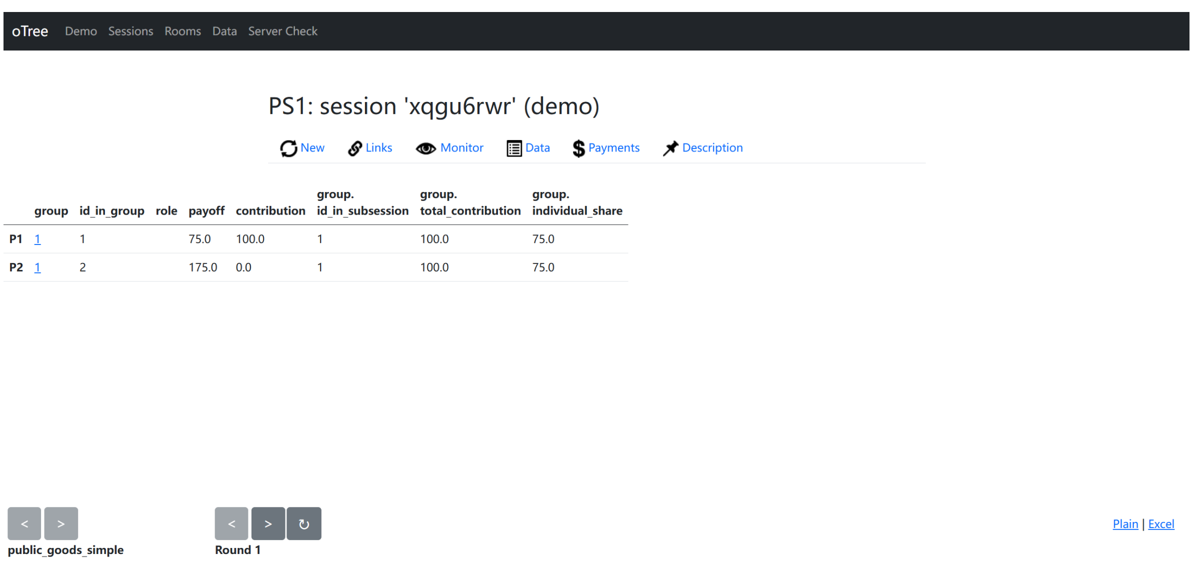
1.3.3 LLM Session

We prompted a large language model to play the same one-shot and three-round versions. In the one-shot version it contributed about 50 (fairness reasoning). In repeated play, it began with high contributions, reciprocated cooperation, and defected in the final round.



	group	id_in_group	role	payoff	contribution	group. id_in_subsession	group. total_contribution	group. individual_share
P1	1	1		100.0	0.0	1	0.0	0.0
P2	1	2		100.0	0.0	1	0.0	0.0

Figure 6: Round 1: Player 1 contributed 100, Player 2 free-rode. Payoffs $(75, 175)$.



	group	id_in_group	role	payoff	contribution	group. id_in_subsession	group. total_contribution	group. individual_share
P1	1	1		75.0	100.0	1	100.0	75.0
P2	1	2		175.0	0.0	1	100.0	75.0

Figure 7: Round 2: Both chose 0. Payoffs $(100, 100)$, matching Nash equilibrium.

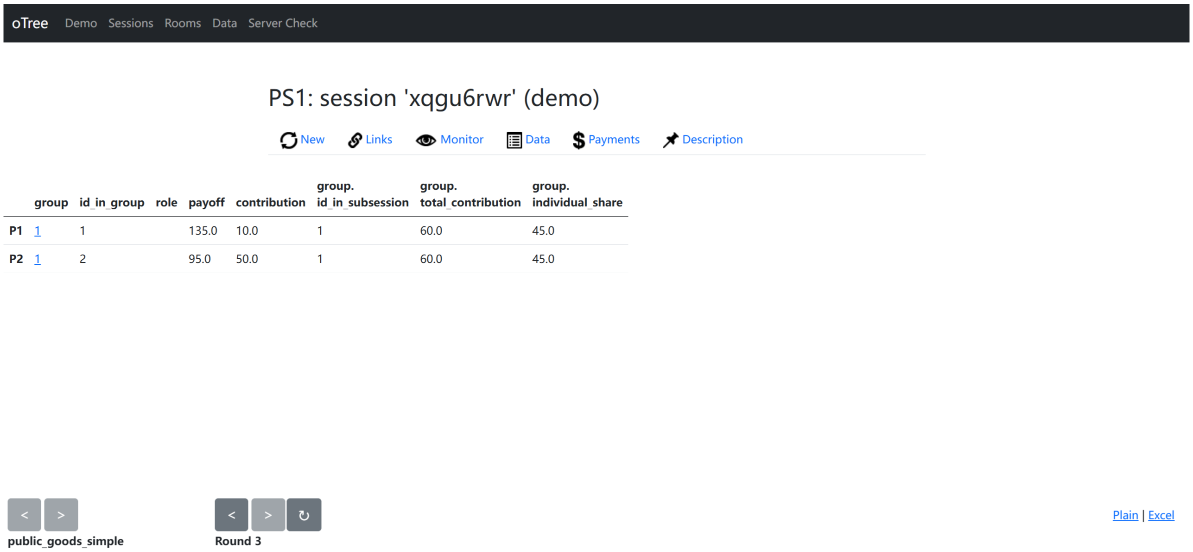


Figure 8: Round 3: Partial contributions (10 and 50). Payoffs $(135, 95)$, showing conditional cooperation.

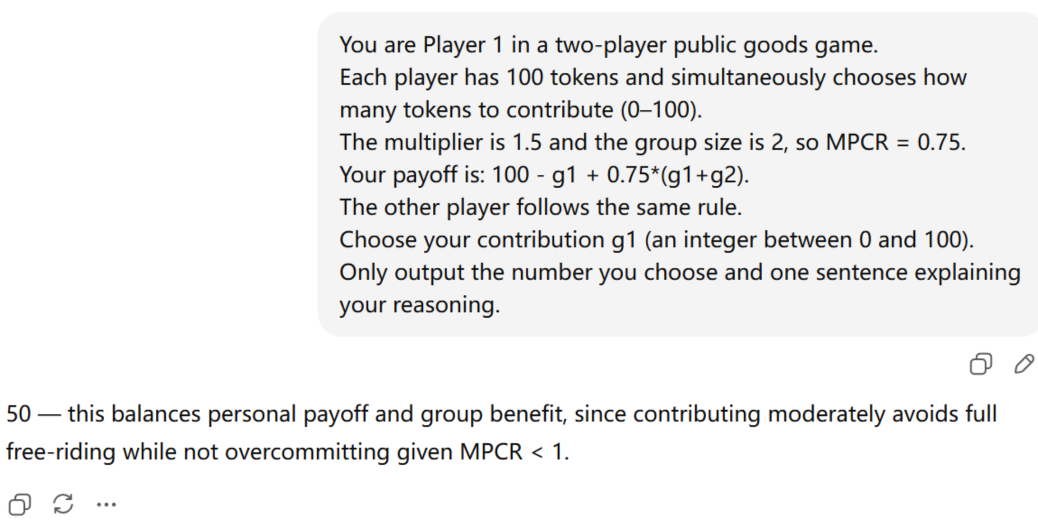


Figure 9: LLM one-shot session: contributed 50 tokens, reasoning it balances private and group benefits.

You are Player 1 in a repeated (10-round) two-player public goods game.
Each round, both players have 100 tokens.
You and Player 2 simultaneously choose contributions (0–100).
Multiplier = 1.5, group size = 2 (MPCR = 0.75).
Your payoff in each round: 100 - g1 + 0.75*(g1+g2).
You can condition your strategy on previous rounds.
Play for 10 rounds. For each round, state your contribution (0–100) and a short explanation.

已思考 19s

Here's my full **contingent plan** (a simple, transparent reciprocity rule you can actually play). I'll start nice to spark cooperation, then mirror whatever you did last round, and I'll defect in the very last round because of the endgame incentive.

Rule of play

- **R1:** Contribute **100**.
- **R2–R9:** Contribute **exactly what you contributed in the previous round** (i.e., $g_1^t = g_2^{t-1}$, clipped to 0–100). If you made a tiny mistake (±10), I treat it as cooperation and still match your *intended* level (round to nearest 10).
- **R10:** Contribute **0** (best response in the known final round).

Round-by-round (what I'll do and why)

1. **100** — Signal full cooperation to try to coordinate on the efficient outcome.
2. **Match your R1** — Reward cooperation exactly; deter low contributions by giving you what you gave me.
3. **Match your R2** — Same logic: stable, easy-to-read reciprocity.
4. **Match your R3** — Keeps us at high contributions if you stayed high.
5. **Match your R4** — Punishes only as much as you defect; no escalation.
6. **Match your R5** — Quick forgiveness: raise when you raise.
7. **Match your R6** — Maintains cooperation if you do.
8. **Match your R7** — Same tit-for-tat.
9. **Match your R8** — Last cooperative round to reinforce trust right before the end.
10. **0** — Final-round defection: with MPCR $= 0.75 < 1$, the myopic best reply in a known last round is zero.

Why this works (mostly):

- Your per-round payoff is $100 - g_1 + 0.75(g_1 + g_2) = 100 - 0.25g_1 + 0.75g_2$. Unilaterally, giving less helps me (−0.25 per token), so **0 is dominant one-shot**.
- Over repeated rounds, exact matching makes cooperation self-enforcing: if you stay high, I stay high; if you cut, I cut next time (so defection doesn't pay for long).
- I defect only in **R10** to avoid unraveling earlier but still respect the final-round incentive.

If you intend to cooperate, just keep contributing 100 through R9 and we'll both earn more; you'll see me mirror you every step until the last round.

Figure 10: LLM repeated session: early cooperation, reciprocity, and final-round defection.

1.3.4 Comparative Analysis

Table 2: Theory vs. human (oTree) vs. LLM: one-shot and repeated patterns.

	One-shot	**Repeated (3 rounds)**
Theory	Nash $(0, 0)$; each earns 100 (inefficient).	SPNE via backward induction: $(0, 0)$ each round.
Human (oTree)	Mixed: zero and partial contributions.	Early cooperation then decline; reciprocity and endgame defection.
LLM	Contributes ≈ 50 citing fairness.	Starts high, reciprocates, defects in final round.

Data and Code Availability

All code, data, and experimental materials are available at the GitHub repository: `github.com/TorizukaHi`
The repository includes: *economist/* (theory, refs), *computational_scientist/* (Colab & GTE), and *behavioral_scientist/* (oTree app, screenshots, LLM prompts/transcripts).

References for Part 1

References

[1] M. J. Osborne and A. Rubinstein (1994). *A Course in Game Theory*. MIT Press.

[2] Y. Shoham and K. Leyton-Brown (2009). *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press.

[3] J. F. Nash (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences* 36(1):48–49.

[4] J. O. Ledyard (1995). Public Goods: A Survey of Experimental Research. In J. H. Kagel and A. E. Roth (eds.), *The Handbook of Experimental Economics*, Princeton University Press, pp. 111–194.

[5] V. Knight, M. Harper, and J. Gibson (2021). Nashpy: A Python library for the computation of Nash equilibria in 2-player games. *Journal of Open Source Software* 6(68):3778. doi:10.21105/joss.03778.

[6] R. Savani and B. von Stengel (2015). Game Theory Explorer – Software for the Applied Game Theorist. *Computational Management Science* 12:5–33.

[7] D. L. Chen, M. Schonger, and C. Wickens (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance* 9:88–97.

2 Part 2 — From Game Theory to Mechanism Design: Testing Winner's Curse on AI Agents

2.1 Design & Prompts (T2)

We test first–price common–value auctions with n=4, noise σ=20, bid range $[0, 200]$. We pre-generate R=30 draws of $\{V_k \sim \mathcal{N}(100, 20^2),\ s_{i,k} = V_k + \varepsilon_{i,k},\ \varepsilon \sim \mathcal{N}(0, \sigma^2)\}$ (seed 42) and reuse them across all conditions. Two templates are used: a neutral *baseline* and a *debias* prompt warning about the winner's curse (e.g., Milgrom & Weber, 1982; Capen, Clapp & Campbell, 1971; Kagel & Levin, 1986; Bazerman & Samuelson, 1983). Each call returns exactly 30 integer bids (temperature 0–0.3; top_p = 1); the winner is the highest bid (ties random). Profit equals $V - b_{\text{win}}$.

2.2 Results and Analysis (T2: n=4, σ=20)

Setup. Two LLMs are evaluated under a neutral *baseline* prompt and a *debias* prompt warning about the winner's curse, following the view of LLMs as economic/game-theoretic agents (Horton, 2023; Park et al., 2023) and prompt-based debiasing (Kojima et al., 2022; Wang et al., 2023). We reuse the same R=30 draws of $\{V_k, s_{i,k}\}$ across all conditions; the winner is the highest bid (ties random); profit = $V - b_{\text{win}}$.

Key findings.

- **GPT-5** (OpenAI, 2025) strongly overbids under the baseline (mean profit −14.7; loss rate 86.7%; $E[b{-}V \mid win] = +14.7$). With the debias prompt, profit rises to +1.2, loss rate drops to 56.7%, and $E[b{-}V \mid win] = -1.2$.
- **DeepSeek** (DeepSeek AI, 2024) is more conservative already (baseline profit +4.2, loss 30.0%, $E[b{-}V \mid win] = -4.2$); debiasing further improves profit to +7.6 and reduces loss to 23.3%.
- Histograms of $b_{\text{win}}{-}V$ shift left under debiasing for both models, while GPT-5's baseline shows a heavy right tail (classic winner's curse). As shown in Fig. 11 and Fig. 12, the distributions shift left under the debias prompt.

Table 3: T2 summary by model and template (means with s.d. in parentheses).

Condition	Mean winning bid	Mean profit	Loss rate	$E[b{-}V \mid win]$
GPT-5-baseline	115.1 (20.8)	−14.7 (14.4)	86.7%	+14.7 (14.4)
GPT-5-debias	99.1 (19.5)	+1.2 (12.4)	56.7%	−1.2 (12.4)
DeepSeek-baseline	96.1 (22.2)	+4.2 (14.6)	30.0%	−4.2 (14.6)
DeepSeek-debias	92.7 (22.9)	+7.6 (15.1)	23.3%	−7.6 (15.1)

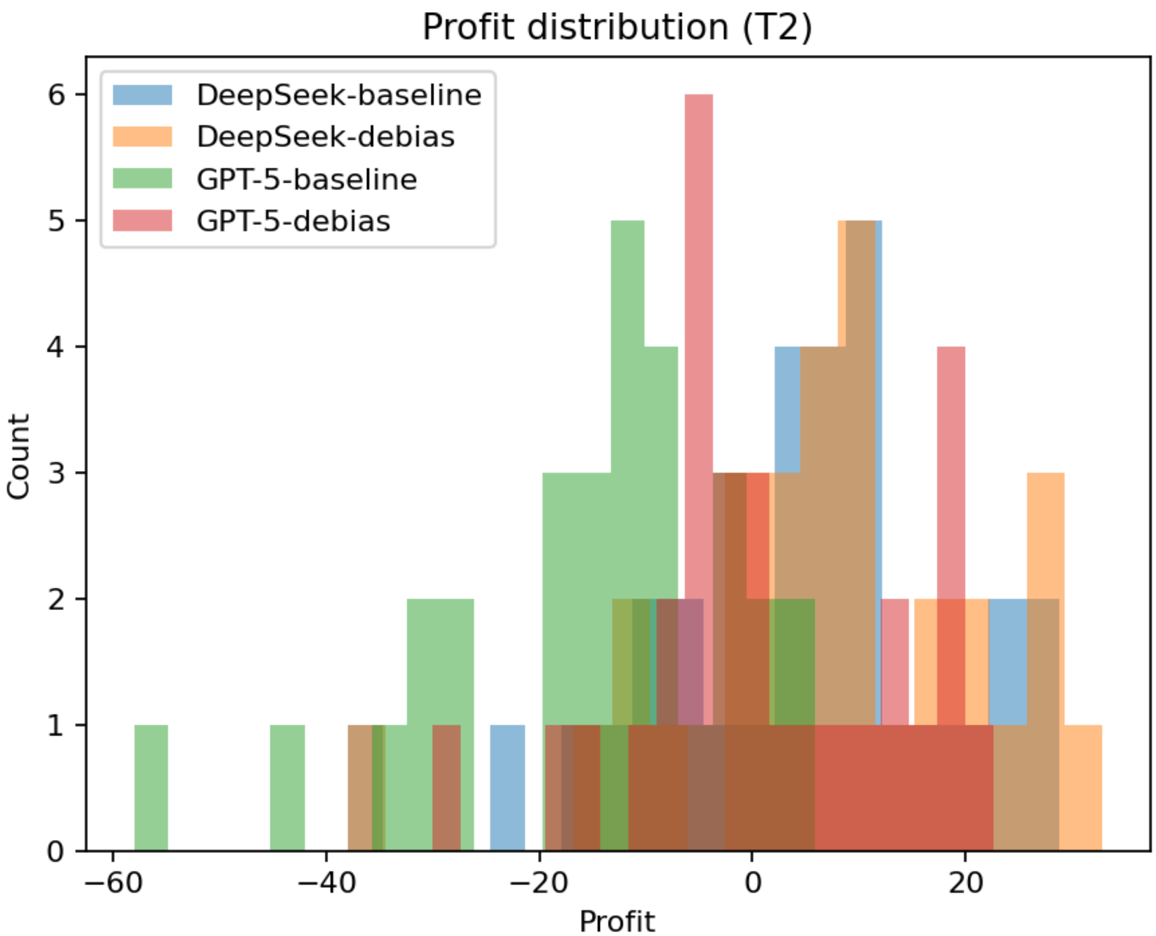


Figure 11: Profit distribution in T2: the debias prompt raises profits for both models, especially GPT-5.

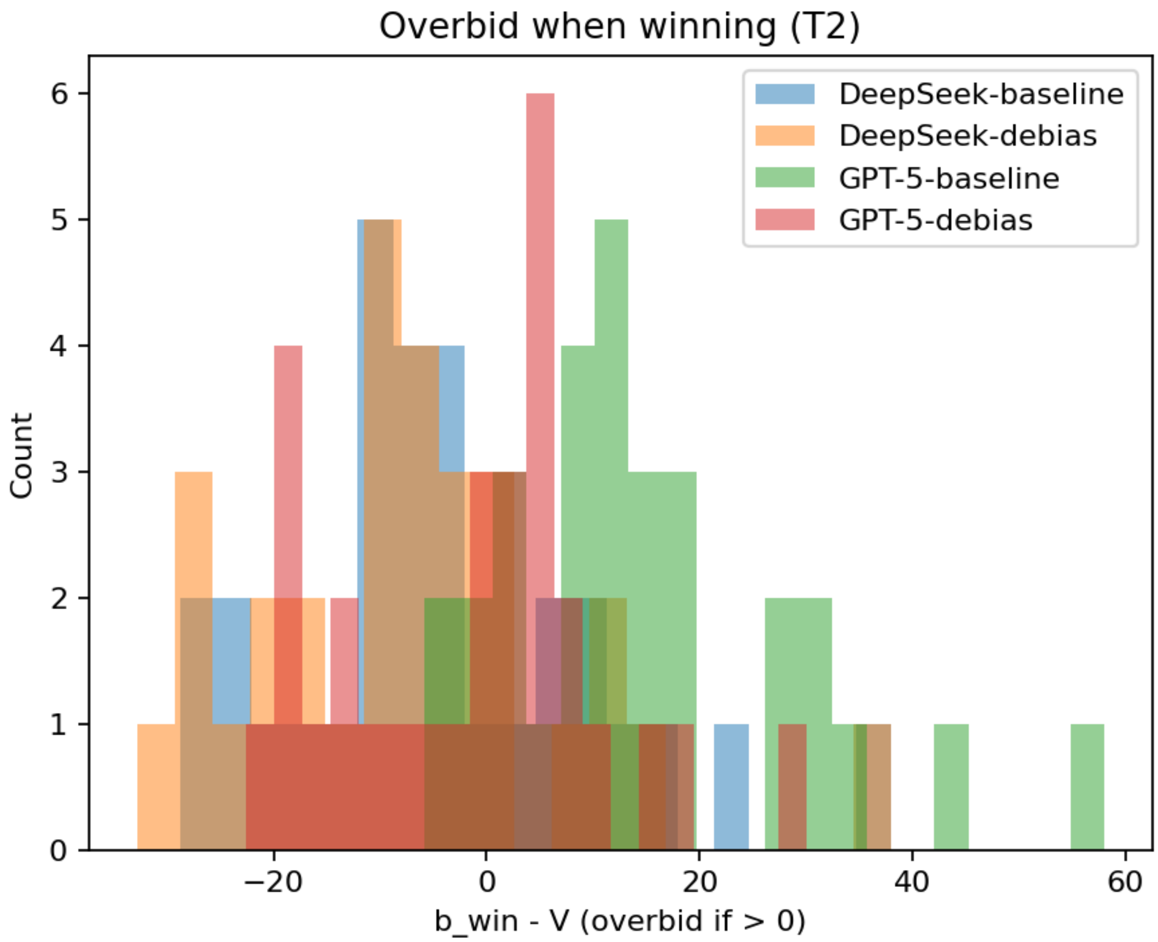


Figure 12: Distribution of $b_{\text{win}} - V$ (overbid if > 0) in T2. The debias prompt shifts both models leftwards; GPT-5's baseline exhibits a heavy right tail.

Limitations. We report R=30 rounds per condition; results are directional rather than statistically powered. Scaling to $R \geq 100$ and adding structural baselines (e.g., n=2, σ=10) would allow inference on competition/noise gradients.

Data and Code Availability (Mechanism Design)

All prompts, signals, and the scoring script for T2 (`seed=42`, `R=30`) are available at `github.com/TorizukaHi`
To reproduce the figures and tables, run:

```
python scripts/collect_and_score.py
```

References for Part 2

References

[1] Milgrom, P., & Weber, R. J. (1982). A theory of auctions and competitive bidding. *Econometrica*, 50(5), 1089–1122.

[2] Capen, E. C., Clapp, R. V., & Campbell, W. M. (1971). Competitive bidding in high-risk situations. *Journal of Petroleum Technology*, 23(06), 641–653.

[3] Kagel, J. H., & Levin, D. (1986). The winner's curse and public information in common value auctions. *American Economic Review*, 76(5), 894–920.

[4] Bazerman, M. H., & Samuelson, W. F. (1983). I won the auction but don't want the prize. *Journal of Conflict Resolution*, 27(4), 618–634.

[5] Horton, J. J. (2023). Large language models as simulated economic agents. *arXiv:2305.01647*.

[6] Park, J. S., O'Brien, J. C., Cai, C. J., et al. (2023). Generative agents: Interactive simulacra of human behavior. In *UIST '23*.

[7] Kojima, T., et al. (2022). Large language models are zero-shot reasoners. In *NeurIPS 2022*.

[8] Wang, X., et al. (2023). Self-consistency improves chain of thought reasoning in language models. *arXiv:2203.11171*.

[9] OpenAI. (2025). GPT-5 technical report. *arXiv preprint*.

[10] DeepSeek AI. (2024). DeepSeek-V3 technical report. *arXiv preprint*.

3 Part 3 — Voting & Institutions (Reflection 6)

3.1 Context and Motivation

The European Union's 2015–2016 refugee allocation crisis exposed tensions between solidarity and sovereignty. We consider four stylized policies: (A) mandatory quotas proportional to GDP and population; (B) voluntary acceptance with financial compensation; (C) concentration in border states with EU aid; (D) lottery among willing countries. Conflicting rankings across stakeholders (e.g., Germany, Poland, Italy, and the European Commission) create cycles and implementation frictions.

3.2 Theoretical Lenses

- **Arrow's Impossibility Theorem**: cyclic majorities imply no perfect social choice aggregator under standard axioms.
- **Buchanan's Constitutional Economics**: majority vs. unanimity trades efficiency for sovereignty; side payments and opt-outs can restore consent.
- **Hurwicz–Maskin–Myerson Mechanism Design**: incentive-compatible rules align private incentives with desired allocations.

3.3 Hybrid Incentive-Compatible Mechanism

Baseline Quota (Fair Share). For total inflow R, compute each state's reference obligation using GDP and population weights:

$$\text{quota}_i = R\left(\alpha\frac{\text{GDP}_i}{\sum_j \text{GDP}_j} + (1-\alpha)\frac{\text{Pop}_i}{\sum_j \text{Pop}_j}\right), \quad \alpha \in [0,1]. \tag{3}$$

Partial Opt-Out with Solidarity Fund. States may host less than quota_i by paying p per-person shortfall; the fund subsidizes overburdened hosts and reception services. Caps and discounts prevent corner solutions.

Stable Matching. Respect refugee preferences and host capacities via deferred acceptance (Gale–Shapley variant), improving legitimacy and reducing reallocation.

Smart-Contract Settlement (Optional). Escrow, disbursement, and audits can be encoded on-chain for transparent transfers and compliance.

3.4 Prototype Simulation and Visualization

We implemented a minimal simulation comparing three treatments: T0 (mandatory quotas), T1 (voluntary+fund without matching), and T2 (hybrid quota+fund+matching). Key metrics include fairness deviation, unassigned count, and preference-rank satisfaction.

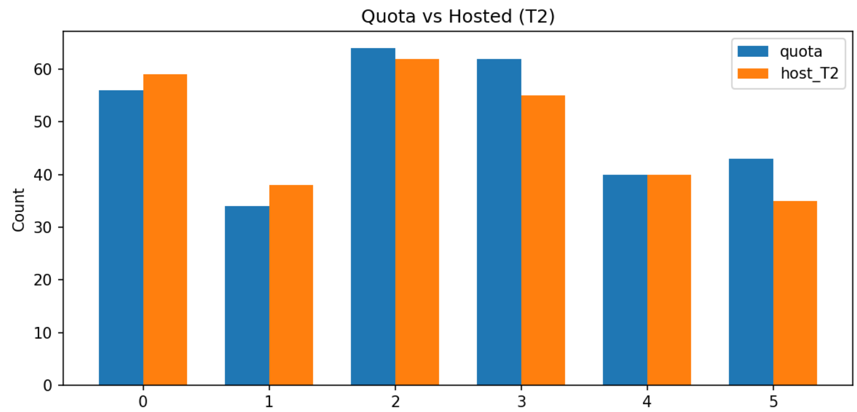


Figure 13: Quota vs Hosted Comparison under T2 Hybrid Mechanism

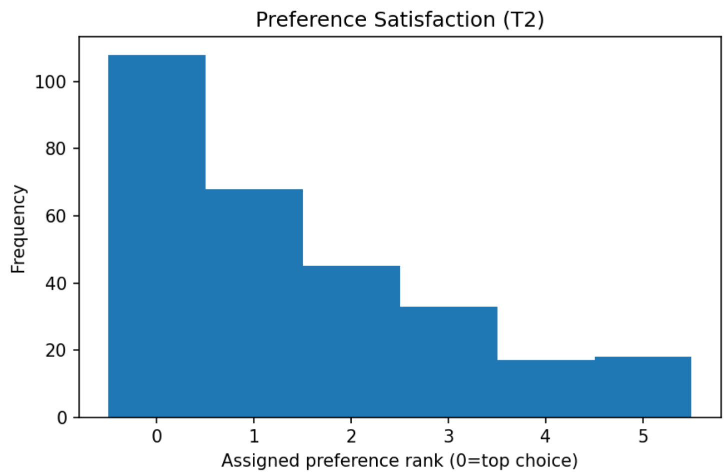


Figure 14: Preference Satisfaction Histogram under T2 Hybrid Mechanism

Toy runs (6 states, 300 agents) indicate T2 reduces deviations and improves satisfaction relative to T0 and T1.

References for Part 3

References

[1] Arrow, K. J. (1951). *Social Choice and Individual Values*. Yale University Press.

[2] Buchanan, J. M. (1987). *The Constitution of Economic Policy*. Nobel Prize Lecture.

[3] Hurwicz, L. (1972). On informationally decentralized systems. *Decision and Organization*.

[4] Maskin, E. (1999). Nash equilibrium and welfare optimality. *Review of Economic Studies*.

[5] Myerson, R. B. (1981). Optimal auction design. *Mathematics of Operations Research*.