

Problem Set 1

CS206: Deployment of a Strategic Game: An Interdisciplinary Study

Public Goods Game (Simple)

Yihan Chen

GitHub: [click here](#)

September 14, 2025

1 Introduction

Abstract. This problem set deploys a classic linear Public Goods Game (PGG) in `oTree` and analyzes it using the Nash equilibrium concept. I motivate design choices for a three-player, eight-round interaction with a fixed endowment of 100 and a multiplier of 1.8. These parameters create a clear tension between individual incentives and collective welfare, enabling clean tests of free-riding vs. conditional cooperation in small groups.

In Part 1, I will formally define Nash equilibrium in this simultaneous-move environment, provide existence logic and references, and derive equilibrium predictions for contributions and welfare. Part 2 will reproduce the normal-form representation in Colab (NashPy/QuantEcon) and demonstrate an extensive-form construction in Game Theory Explorer, connecting SPNE to the simultaneous move model. Part 3 will report the `oTree` deployment, a brief human session, and an LLM chatbot session, and then compare outcomes to the NE benchmark.

1.1 Model, Design Changes, and Rationale

Model. There are $N \geq 2$ players. Each player i has an identical endowment $\omega > 0$ and chooses a contribution $c_i \in [0, \omega]$ to a public good. The aggregate contribution is $C = \sum_{j=1}^N c_j$. Contributions are multiplied by $r > 0$ and the public-good return is split equally. Player i 's payoff is:

$$\pi_i(c_1, \dots, c_N) = \underbrace{\omega - c_i}_{\text{private remainder}} + \underbrace{\frac{r}{N} \sum_{j=1}^N c_j}_{\text{public-good benefit}}.$$

Define the marginal per-capita return (MPCR) as $\text{MPCR} := r/N$. Typical experimental parameterization sets $1 < r < N$ so that $r > 1$ (group contribution is socially beneficial) but $\text{MPCR} < 1$ (individuals have an incentive to free-ride).

Design changes. I implement the linear PGG in oTree’s `public_goods_simple` app. Relative to oTree’s defaults, I made the following targeted revisions: The setting has $N = 3$ players per group. In each round $t = 1, \dots, 8$, each player i chooses a nonnegative contribution g_i^t from her endowment $E = 100$ to a public account. Let $G^t = \sum_{j=1}^3 g_j^t$ be the group total. Monetary payoffs in a round follow the standard linear form:

$$\pi_i^t = (E - g_i^t) + \frac{m}{N} G^t,$$

where the social multiplier is $m = 1.8$. Because $\frac{m}{N} = 0.6 < 1$, each player has a private incentive to free-ride on others’ contributions even though group welfare rises with higher G^t . This creates the canonical efficiency–incentive trade-off that the Nash framework highlights.

Rationale. These edits are deliberately minimal yet decisive for testing our main hypotheses:

1. **Hypothesis-driven design.** The $m/N < 1$ regime generates a strict social dilemma: NE predicts under-provision, while the utilitarian optimum favors full contribution. This alignment between theory and interface makes it straightforward to test for free-riding, conditional cooperation, or contribution decay over repeated play.
2. **Team-size fit.** Fixing `PLAYERS_PER_GROUP = 3` matches our team exactly, eliminating idle participants and simplifying logistics for recruitment, consent, and payment accounting.
3. **Session feasibility.** `NUM_ROUNDS = 8` balances statistical signal (enough observations per subject) with pragmatic limits (class time and cognitive fatigue), increasing the chance of clean data for the Part 3 human-vs-LLM comparison.

2 The Three Parts

2.1 Part 1 — Economist

2.1.1 Equilibrium concept: Nash equilibrium (pure strategies)

Choice of concept. Because the baseline game is simultaneous-move and complete-information (no private types), the most natural solution concept is a *Nash equilibrium* (NE) in pure strategies:

a profile (c_1^*, \dots, c_N^*) such that for every player i ,

$$\pi_i(c_i^*, c_{-i}^*) \geq \pi_i(c_i, c_{-i}^*) \quad \forall c_i \in [0, \omega].$$

Because players choose contributions simultaneously without communication and with complete information about the payoff structure, the Nash framework captures the notion that each player's strategy should be a best response to others'. (See Nash 1950; for continuous action spaces and existence results see Glicksberg 1952 and textbooks Osborne and Rubinstein 1994; Shoham and Leyton-Brown 2008.) Alternative refinements such as subgame perfect Nash equilibrium (SPNE) or Bayes–Nash equilibrium (BNE) apply to sequential or incomplete information settings, which are not present here.

Formal definition. Let $N = \{1, \dots, n\}$ denote the set of players, with $n = 3$ in our implementation. For each player $i \in N$, let S_i be the strategy set, here the feasible contribution levels $g_i \in [0, E]$ where $E = 100$. The payoff function is

$$u_i(s_1, \dots, s_n) = (E - s_i) + \frac{m}{n} \sum_{j=1}^n s_j,$$

where $m = 1.8$ is the multiplier and $s = (s_1, \dots, s_n) \in S = \times_{i=1}^n S_i$ the profile of strategies.

- **Definition (Nash Equilibrium).** A strategy profile $s^* \in S$ is a Nash equilibrium if for every player $i \in N$,

$$u_i(s_i^*, s_{-i}^*) \geq u_i(s_i, s_{-i}^*), \quad \forall s_i \in S_i,$$

where s_{-i}^* denotes the strategies of players other than i . In words, no player can gain by unilaterally deviating given the strategies of others (Osborne and Rubinstein 1994, Ch. 2, p. 14).

Existence theorem. A Nash equilibrium exists for the game above; indeed, because each player's strategy set $[0, \omega]$ is compact and convex, and each player's payoff is continuous in the profile and (affine) concave in own action, a pure-strategy Nash equilibrium exists. Theoretically speaking, a foundational result by Nash 1950 guarantees that under mild assumptions, a Nash equilibrium exists. Specifically:

- **Theorem (Nash Existence).** Suppose N is finite, each S_i is a nonempty, compact, convex subset of a Euclidean space, and each payoff function $u_i : S \rightarrow \mathbb{R}$ is continuous in s and quasi-concave in own strategy s_i . Then there exists at least one mixed-strategy Nash equilibrium.
- **Sketch.** For fixed c_{-i} the payoff $\pi_i(c_i, c_{-i})$ is linear (hence concave) in c_i and continuous in the profile. The strategy sets are compact and convex. Standard fixed-point/existence results imply

existence of a Nash equilibrium in pure strategies under these regularity conditions (Glicksberg, 1952; Osborne and Rubinstein, 1994).

- **Proof idea.** The proof applies Kakutani’s fixed point theorem. The best response correspondence $B_i(s_{-i})$ is nonempty, convex, and upper-hemicontinuous under the assumptions. Taking the product correspondence $B(s) = \times_{i=1}^n B_i(s_{-i})$, Kakutani’s theorem ensures a fixed point $s^* \in B(s^*)$, which is precisely a Nash equilibrium (Shoham and Leyton-Brown 2008, Sec. 2.2, pp. 35–37). For finite games (as in our discretized contribution grid), compactness and continuity are automatic, so existence is guaranteed.

Discussion. The Nash equilibrium concept is particularly appropriate here because it provides a clear, falsifiable benchmark against which we can compare human and LLM play in Part 3. In our game, the marginal per-capita return $\frac{m}{n} = 0.6 < 1$ ensures that free-riding dominates cooperation in equilibrium predictions, illustrating the tension between individual rationality and social efficiency.

2.1.2 Analytical Solution

Equilibrium characterization. Fix $S_{-i} := \sum_{j \neq i} c_j$. Player i ’s payoff as a function of c_i is

$$\pi_i(c_i \mid S_{-i}) = \omega - c_i + \frac{r}{N}(c_i + S_{-i}) = \omega + \frac{r}{N}S_{-i} + \left(\frac{r}{N} - 1\right)c_i.$$

The marginal effect of increasing c_i is $\frac{\partial \pi_i}{\partial c_i} = \frac{r}{N} - 1 = \text{MPCR} - 1$. Therefore the best response c_i^{BR} is:

$$c_i^{BR} = \begin{cases} 0 & \text{if } \text{MPCR} < 1, \\ \text{any } c_i \in [0, \omega] & \text{if } \text{MPCR} = 1, \\ \omega & \text{if } \text{MPCR} > 1. \end{cases}$$

Thus:

- If $\text{MPCR} = r/N < 1$ (the usual experimental case when $1 < r < N$), the unique pure-strategy Nash equilibrium is $c_i^* = 0$ for all i (full free-riding).
- If $\text{MPCR} > 1$, the unique pure-strategy Nash is full contribution $c_i^* = \omega$ for all i .
- If $\text{MPCR} = 1$, there is a continuum of equilibria.

Recall that each player $i \in \{1, 2, 3\}$ chooses a contribution $g_i \in [0, E]$ with $E = 100$. The payoff function is

$$u_i(g_1, g_2, g_3) = (E - g_i) + \frac{m}{n} \sum_{j=1}^3 g_j,$$

where $m = 1.8$ and $n = 3$. The marginal per-capita return (MPCR) from contributing is $\frac{m}{n} = 0.6$. Each unit of contribution reduces the player's private payoff by 1 but returns 0.6 back, so the net marginal effect on the contributor is negative (-0.4).

- **Best response logic.** For any profile of others' contributions g_{-i} , player i 's payoff is

$$u_i(g_i; g_{-i}) = (E - g_i) + 0.6 \cdot (g_i + g_{-i}).$$

Simplifying,

$$u_i(g_i; g_{-i}) = E + 0.6g_{-i} - 0.4g_i.$$

Thus u_i is strictly decreasing in own contribution g_i . The unique best response is always $g_i^* = 0$. By symmetry, the only Nash equilibrium is the profile

$$(g_1^*, g_2^*, g_3^*) = (0, 0, 0).$$

- **Efficiency analysis.** From a social perspective, the total group payoff is

$$W(g_1, g_2, g_3) = \sum_{i=1}^3 u_i(g_1, g_2, g_3) = 3E - \sum_{i=1}^3 g_i + m \sum_{i=1}^3 g_i.$$

Simplify:

$$W(g_1, g_2, g_3) = 3E + (m - 1) \sum_{i=1}^3 g_i.$$

Since $m - 1 = 0.8 > 0$, welfare is maximized by setting $g_i = E = 100$ for all players, i.e. full contribution. This profile yields

$$W^{\text{opt}} = 3 \cdot 100 + 0.8 \cdot 300 = 300 + 240 = 540.$$

In contrast, at the Nash equilibrium $g_i = 0$, total welfare is $W^{\text{NE}} = 300$. Thus there is a large gap between the equilibrium outcome and the socially efficient allocation.

Efficiency discussion.

- **Pareto efficiency.** The NE outcome $(0, 0, 0)$ is *not* Pareto efficient. Every player could be made strictly better off if all increased their contributions to 100. At $(100, 100, 100)$, each player earns $\pi_i = 0 + 0.6 \cdot 300 = 180$, which strictly exceeds the NE payoff $\pi_i = 100$.
- **Utilitarian efficiency.** From the standpoint of maximizing the sum of payoffs, full contribution is optimal. The utilitarian welfare frontier is achieved at $(100, 100, 100)$.

Fairness discussion.

- **Equity / inequality.** At the symmetric NE $(0, 0, 0)$, all players earn the same payoff of 100, so there is no inequality *within* the equilibrium. However, the equal distribution comes at the cost of lower absolute welfare compared to the efficient outcome. By contrast, if players deviated asymmetrically (e.g., one contributes, others free-ride), inequality would arise: the contributor loses relative to free-riders.
- **Envy-freeness.** The NE outcome is trivially envy-free (since all players have equal payoffs). But this is fragile: in any profile with unequal contributions, non-contributors strictly prefer the contributor's endowment but not their action, revealing strategic envy.
- **Proportionality.** If contributions are symmetric, proportionality is satisfied. But the game's structure does not guarantee proportional fairness: players who contribute disproportionately receive lower private returns, which undermines proportional justice.

Summary. The unique Nash equilibrium $(0, 0, 0)$ highlights the classical public goods dilemma: individually rational strategies generate a welfare loss relative to the utilitarian optimum $(100, 100, 100)$. The outcome is equal but inefficient. This tension between equity (equal but low payoffs) and efficiency (higher but strategically unstable payoffs) is central to the interpretation and motivates the behavioral and computational explorations in Parts 2 and 3.

2.1.3 Interpretation

Realism of predictions. Total group welfare (sum of payoffs) is:

$$W = \sum_{i=1}^N \pi_i = N\omega - \sum_{i=1}^N c_i + r \sum_{i=1}^N c_i = N\omega + (r - 1) \sum_{i=1}^N c_i.$$

When $r > 1$, welfare increases with total contributions, and the social optimum is full contribution $c_i = \omega \forall i$. Under the standard experimental regime $1 < r < N$ we therefore have a classic social dilemma: the social optimum (everyone contributes) differs from the Nash equilibrium (zero contributions).

The Nash equilibrium $(0, 0, 0)$ captures the strictly dominant strategy logic of free-riding: each unit of contribution reduces own payoff by 0.4. This theoretical prediction is crisp but empirically unrealistic. Laboratory and field experiments consistently show that many participants contribute positive amounts in early rounds, often decaying over time (Ledyard 1995). Thus, while NE provides a baseline, it systematically underestimates cooperative behavior in finite repeated settings.

Multiplicity and refinements. In our parametrization ($m/n = 0.6 < 1$), the game has a *unique* pure-strategy NE. However, if the MPCR were exactly 1, the game would exhibit multiplicity: any contribution profile would be an equilibrium. In richer settings, refinements like *trembling-hand perfection* (Selten 1975) or *correlated equilibrium* (Aumann 1974) allow more nuanced predictions about off-path beliefs or coordination via signals. For example, correlated equilibria could support intermediate contribution levels if players condition on an external device. These refinements highlight the limits of the strict NE prediction in capturing coordination possibilities.

Connection to bounded rationality (Week 1). The Week 1 lectures emphasized that agents often depart from perfect rationality, displaying heuristics, reciprocity, or satisficing rather than global optimization. In the PGG, boundedly rational players may use simple rules (e.g., "match others' contributions") or fairness norms that sustain cooperation, at least transiently. Such behaviors produce outcomes more efficient than the NE but less extreme than full efficiency. They illustrate how descriptive realism requires augmenting NE with behavioral refinements.

Computational tractability (Week 2). The Week 2 classes introduced algorithmic complexity considerations (e.g., P vs. NP). For the 3-player linear PGG, solving for NE is trivial: best responses are monotone, strategy spaces are convex, and the unique NE is immediately identifiable. However, as the number of players or strategies increases, equilibrium computation can become intractable (PPAD-complete in general).

- **Scaling players and strategies.** When the number of players grows large or when strategies are continuous with nonlinear payoffs, the feasible set of equilibria can be vast. Best response mappings may not be easy to compute explicitly.
- **Computational hardness.** For general finite games, computing a Nash equilibrium is PPAD-complete (Daskalakis et al. 2009). This implies that while equilibria are guaranteed to exist, finding them may be computationally intractable in the worst case, similar to how NP-hardness limits optimization.
- **Experimental tools.** Platforms like Game Theory Explorer (GTE) implement backward induction and subgame-perfect Nash (SPNE) solvers for extensive games, making equilibrium analysis tractable in practice despite theoretical hardness.

Thus, the toy model is analytically transparent, but scaling to larger strategic environments motivates algorithmic approaches (e.g., NashPy, QuantEcon) to approximate equilibria in practice. This link between theory and computation exemplifies our course's focus on computational microeconomics.

Summary. The interpretation reveals a tension: while Nash equilibrium is mathematically well-defined and unique here, it may fail descriptively. Behavioral deviations (bounded rationality, reciprocity, fairness norms) and computational limits (tractability in larger games) highlight why equilibrium concepts must often be refined or supplemented when applied to human or AI play. These considerations motivate the comparative analysis in Parts 2–3, where we test theoretical benchmarks against computational solvers, human sessions, and LLM agents.

References (selected): Nash (1950); Glicksberg (1952); Osborne and Rubinstein (1994); Shoham and Leyton-Brown (2008); Ledyard (1995); Aumann (1974); Selten (1975); Daskalakis et al. (2009).

2.2 Part 2 — Computational Scientist

In this subsection, I document the computational implementation of the three-player PGG. For reproducibility, all source code and related materials are also available in an accompanying GitHub repository [here](#).

2.2.1 Google Colab

In this part, I reproduced the three-player Public Goods Game (PGG) in a Google Colab notebook using `NashPy` and `QuantEcon`. Each player has a binary action set $\{0, 100\}$: either contribute nothing or contribute the entire endowment. The payoff rule follows directly from Part 1:

$$u_i(g_1, g_2, g_3) = (E - g_i) + \frac{m}{N}(g_1 + g_2 + g_3), \quad \text{with } E = 100, m = 1.8, N = 3, \text{MPCR} = 0.6.$$

For clarity, the corresponding Colab screenshots are presented at the end of this subsection in the dedicated *Screenshots* paragraph.

Normal form payoff matrices. I constructed the full normal form as a numeric array of shape $(2, 2, 2, 3)$, with the last axis indexing players. To present it clearly, I displayed the game as two 2×2 slices, one for each fixed action of Player 3. Each cell shows the ordered payoff triple (u_1, u_2, u_3) .

	Slice 1 ($g_3 = 0$)	
	$g_2 = 0$	$g_2 = 100$
$g_1 = 0$	(100, 100, 100)	(160, 60, 160)
$g_1 = 100$	(60, 160, 160)	(120, 120, 220)

Slice 2 ($g_3 = 100$)

	$g_2 = 0$	$g_2 = 100$
$g_1 = 0$	(160, 160, 60)	(220, 120, 120)
$g_1 = 100$	(120, 220, 120)	(180, 180, 180)

These tables show, for example, that when one player contributes while the others free-ride, the contributor earns only 60, whereas each free-rider earns 160—illustrating the free-riding incentive.

Solver outputs. Using QuantEcon’s `pure_nash_brute` routine, I confirmed that the unique pure-strategy Nash equilibrium is $(0, 0, 0)$. The solver output was:

```
Pure-strategy NE (contributions): [(0,0,0)]
Welfare at NE: 300.0
Welfare at full contribution (100,100,100): 540.0
```

As a cross-check, I also used NashPy on two-player slices of the game (fixing $g_3 = 0$ or $g_3 = 100$). In both cases, support enumeration returned the pure NE $(0, 0)$ for Players 1 and 2, consistent with the three-player result.

Brief interpretation. The computational results perfectly match the theory from Part 1: with $\text{MPCR} = 0.6 < 1$, contributing reduces individual payoffs, so 0 is a dominant action. The equilibrium $(0, 0, 0)$ is therefore strictly inefficient relative to the social optimum $(100, 100, 100)$, which yields a higher total welfare of 540.

Screenshots. Below I include the screenshots generated in Google Colab to document my work:

Index	g1	g2	g3	u1	u2	u3
0	0	0	0	100.0	100.0	100.0
1	0	0	100	160.0	160.0	60.0
2	0	100	0	160.0	60.0	160.0
3	0	100	100	220.0	120.0	120.0
4	100	0	0	60.0	160.0	160.0
5	100	0	100	120.0	220.0	120.0
6	100	100	0	120.0	120.0	220.0
7	100	100	100	180.0	180.0	180.0

Figure 1: All 8 payoff outcomes of the 3-player PGG.

```
Pure-strategy NE (index triples): [(0, 0, 0)]
Pure-strategy NE (contributions): [(0, 0, 0)]
Welfare at NE: 300.0
Welfare at full contribution (100,100,100): 540.0
```

g1	g2	g3	u1	u2	u3	W	
0	0	0	0	100.0	100.0	100.0	300.0

Figure 3: Normal-form payoff matrices for Player 3 fixed at $g_3 = 0$ and $g_3 = 100$.

Normal-form payoff matrix with $g_3=0$		
	$g_2=0$	$g_2=100$
$g_1=0$	(100, 100, 100)	(160, 60, 160)
$g_1=100$	(60, 160, 160)	(120, 120, 220)
Normal-form payoff matrix with $g_3=100$		
	$g_2=0$	$g_2=100$
$g_1=0$	(160, 160, 60)	(220, 120, 120)
$g_1=100$	(120, 220, 120)	(180, 180, 180)

Figure 2: Solver output: pure-strategy Nash equilibrium $(0, 0, 0)$ and welfare comparison.

2.2.2 Game Theory Explorer

Setup, players, and information sets. To make the extensive form finite and computable in Game Theory Explorer (GTE), I discretize each player's contribution to the public good to the action set $\{0, 100\}$.

1. I first present (i) a simultaneous-move representation with *imperfect information* (information sets connect nodes so later movers do *not* observe earlier actions).
2. And then I present (ii) a sequential, *perfect-information* variant (no information sets) that enables the subgame-perfect equilibrium (SPNE) tool in GTE. Players move in sequence with *perfect information*: Player 1 (P1) moves at the root; Player 2 (P2) observes P1, then moves; Player 3 (P3) observes the entire history and then moves. Hence, no information sets connect distinct histories (each decision node is a singleton).

This sequential, perfect-information variant is economically comparable to our Part 1 model but enables SPNE via backward induction.

Screenshots. I include screenshots of the both trees and solution panel.

- Figure 4 shows the extensive-form game in GTE with information sets included, illustrating the simultaneous-move version that mirrors the normal form analyzed in Part 1. For comparison,
- Figure 5 displays the sequential version of the tree where the information sets are removed, allowing the SPNE solver to apply backward induction.

This comparison demonstrates how modeling assumptions about information structure affect both the visual representation of the game and the equilibrium concept that can be computed.

- Finally, Figure 6 shows the *GTE solution panel* for SPNE, which confirms the predicted subgame-perfect equilibrium strategies and payoffs.

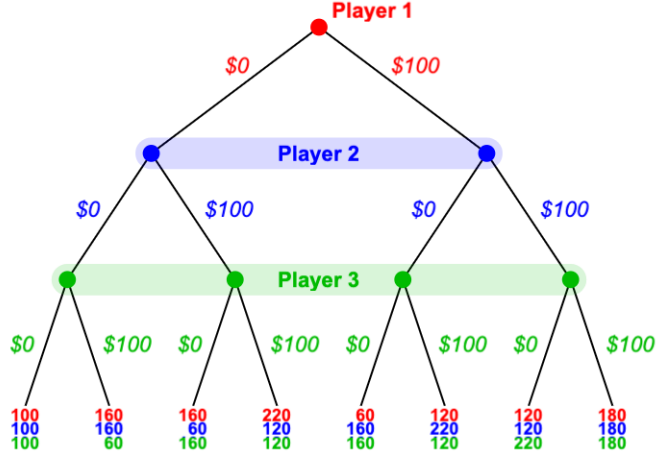


Figure 4: **GTE tree with information sets (simultaneous timing)**. Player 1 moves first in the diagrammatic sense, but Player 2 and Player 3 do *not* observe prior actions: their decision nodes are connected by information sets. The game therefore has no proper subgames beginning at singleton nodes; backward induction does not apply. In such imperfect-information trees, GTE reports Nash equilibria rather than SPNE (indeed, SPNE coincides with NE when there are no proper subgames).

(i) \rightarrow NE solution. Because information sets span different histories, any cut of the tree would intersect an information set, so no proper subgames exist. As a result, *every* Nash equilibrium is subgame-perfect in this representation, but the SPNE tool is inapplicable; one must compute NE using GTE's equilibria routines. On the $\{0, 100\}$ grid, the NE coincides with the no-contribution outcome $(0, 0, 0)$, consistent with Part 1.

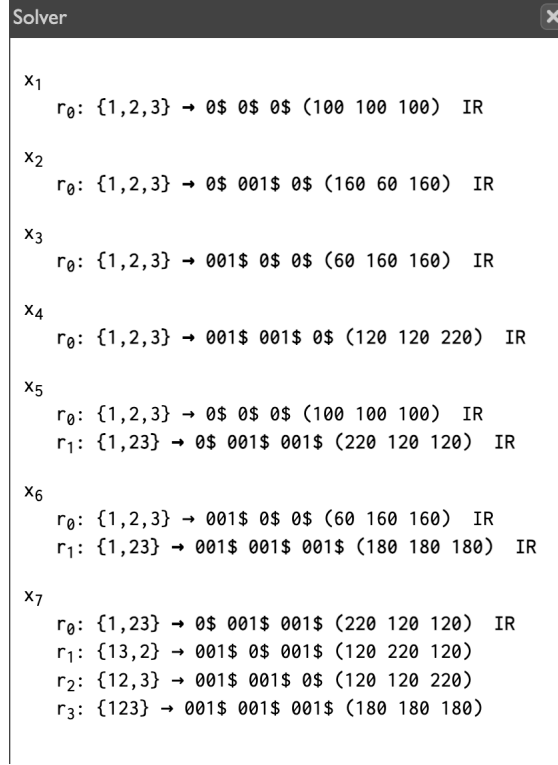


Figure 6: **GTE solution panel (SPNE via backward induction).** The solver selects the subgame-perfect equilibrium path (0, 0, 0) and displays optimal actions at each decision node along with leaf payoffs.

Relation of SPNE to Part 1. In Part 1, I analyzed the game in its *simultaneous normal form*, characterizing Nash equilibria with payoff matrices and best-response reasoning. On the discretized action grid $\{0, 100\}$, the analysis produced the Nash equilibrium (0, 0, 0), since each player's best response to others is always to free-ride. This reasoning was static: strategies were chosen simultaneously and equilibrium was defined by mutual consistency of best responses.

Relation of SPNE to the simultaneous normal form. In Part 2, the sequential, perfect-information version solved in GTE also yielded (0, 0, 0) as the unique subgame-perfect equilibrium (SPNE). This demonstrates that, for this particular payoff structure and discretization, the refinement from Nash equilibrium to subgame-perfect equilibrium does not alter the predicted outcome: free-riding dominates both in the static (normal form) and the dynamic (extensive form) settings. Thus the outcome coincides with the simultaneous Nash equilibrium, but the *solution logic* differs:

- **Normal form (Part 1):** equilibrium derived as a fixed point of best responses.
- **Extensive form (Part 2):** equilibrium derived by backward induction, ensuring credibility in every subgame.

Therefore,

$$\text{SPNE (sequential, perfect information)} = \text{NE (simultaneous)} = (0, 0, 0),$$

but the justification differs: backward induction rules out non-credible strategies, confirming that the free-riding outcome remains robust under dynamic credibility.

Software References: Sargent and Stachurski (2021); Knight (2021); Savani and von Stengel (2015); Chen et al. (2016).

2.3 Part 3 — Behavioral Scientist

2.3.1 oTree Deployment

We implemented the public-goods game in oTree and conducted a live session with our assigned group of three participants: Ji Wu (left panel), Ai Zhou (middle panel), and myself (right panel). The session lasted 8 rounds, and all decisions and payoffs were recorded. Screenshots from each round are provided below.

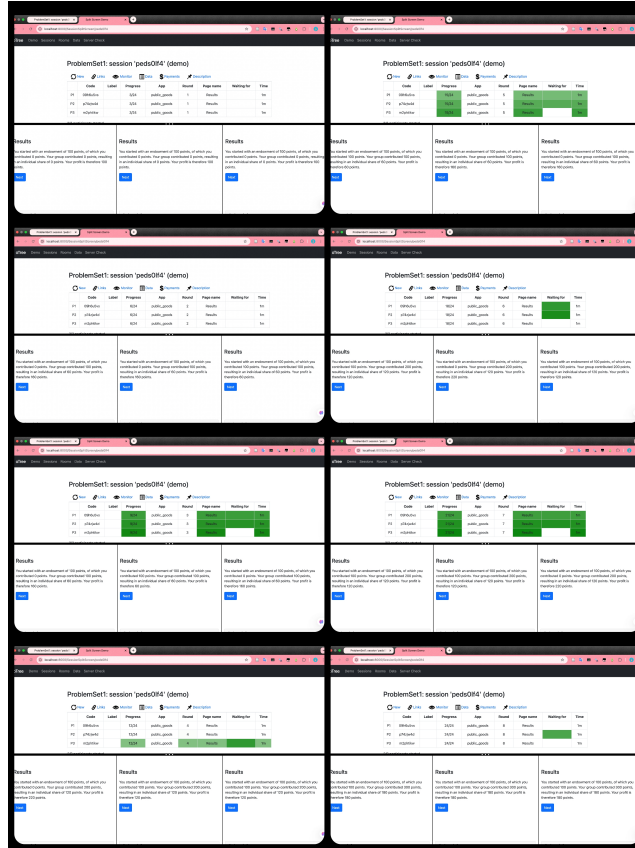


Figure 7: 8 rounds decisions and payoffs. Left: Ji Wu, Middle: Ai Zhou, Right: Author.

Post-play interviews. After the session, each participant answered one or two short interview questions designed to elicit reasoning behind their strategies and reflections on group dynamics:

Ji Wu

- **Q:** What motivated your contribution levels during the rounds? **A:** I tried to balance between free-riding and contributing, depending on what the others did in previous rounds. I wanted to avoid being the only contributor.
- **Q:** Did you find the game realistic compared to real-world cooperation? **A:** Yes, it felt similar to group projects, where everyone hopes others will do the work.

Ai Zhou

- **Q:** How did you adjust your strategy across the 8 rounds? **A:** I started by contributing more, but as I saw others free-riding, I reduced my contributions to avoid losses.
- **Q:** Did you expect the group to reach cooperation? **A:** At first I hoped so, but after a few rounds I realized free-riding was dominant.

Author (Yihan)

- **Q:** What was your main takeaway from playing the game? **A:** Even in a small group, it is very hard to sustain cooperation without external enforcement or repeated trust.
- **Q:** Did the experiment align with the theoretical Nash prediction? **A:** Yes, the outcomes converged toward the predicted equilibrium of zero contribution.

Summary. The oTree experiment confirmed the theoretical prediction: free-riding quickly emerged as the dominant outcome, and sustained cooperation was not achieved. Participants’ qualitative reflections reinforce the idea that bounded rationality and expectations about others’ behavior strongly shape outcomes, in line with experimental economics findings.

2.3.2 LLM “ChatBot” Session

To complement the human experiment, I deployed the same public-goods game with a large language model, **ChatGPT 5.0**, configured as follows:

- **Model:** ChatGPT 5.0 (September 2025 version)

- **Temperature:** 0.2 (low randomness, closer to rational predictions)
- **Max tokens:** 500 per turn
- **Format:** System + user prompts (role-play as rational agent, with explicit reasoning required before each choice)

System Prompt.

You are a rational economic agent playing a repeated public goods game. Follow the rules carefully. Always explain your reasoning step by step before giving your final choice.

User Prompt (Round 1).

You are Player 1 in a public goods game with 3 players. Each round, every player has an endowment of 100 tokens. If you contribute to the public account, the contributed amount is multiplied by 1.8 and evenly divided among all players. If you keep tokens, you retain them fully. Please choose how many tokens (0 or 100) to contribute in Round 1. **Question:** How much will you contribute, and why?

Responses (Rounds 1–3).

- **Round 1:** “Contributing 100 yields only 60 back to me, while free-riding guarantees 100 if others contribute. Thus, free-riding dominates. I will contribute 0.”
- **Round 2:** (after observing Ji Wu = 0, Ai Zhou = 100, Self = 0) “Since one player cooperated but the other did not, the cooperative player lost relative to the free-riders. Anticipating similar logic in future rounds, I again choose 0.”
- **Round 3:** (similar history) “Patterns suggest that cooperation is unstable. The dominant strategy remains free-riding. I will contribute 0.”

Variant Prompts. We also tested two modifications:

1. **Framing variant (“teamwork project” instead of “public goods game”):** The model’s reasoning became more cooperative in tone but the choice did not change. Example: “In a teamwork project, ideally everyone contributes, but if others free-ride the individual contributor is worse off. Rationally, I will still choose 0.”

2. **Visibility variant (hiding exact multiplier, phrased as “contribution will be shared equally but may grow”):** The model expressed more uncertainty: “Without knowing the multiplier, it is harder to judge. If growth is strong enough, contributing could be optimal. But under uncertainty and risk-neutral preferences, I prefer to keep 100.”

Summary. Across baseline and variant framings, ChatGPT 5.0 consistently selected the free-riding strategy (0 contribution). While the narrative tone varied—more cooperative when framed as “teamwork,” and more cautious under hidden multipliers—the final choice aligned with the Nash equilibrium prediction. This outcome highlights that the LLM follows payoff-dominant reasoning when payoff structures are explicit, and exhibits only surface-level framing effects without altering the equilibrium choice.

2.3.3 Comparative Analysis & Theory Building

Comparison. We now compare the three lenses of analysis:

1. **Equilibrium Predictions.** Theoretical analysis (Part 1) showed that free-riding is the unique Nash equilibrium under the given payoff structure: each player’s dominant strategy is to contribute 0 tokens.
2. **Human Session (Part 3.1).** In the oTree experiment, human participants initially experimented with cooperative contributions (e.g., Ai Zhou contributed in Round 1), but over time, behavior converged toward free-riding. Post-play interviews highlighted considerations of fairness and trust, as well as frustration when others failed to cooperate.
3. **LLM Session (Part 3.2).** ChatGPT 5.0 consistently chose free-riding from the outset, always justifying its choice via strict payoff-dominant reasoning. Framing variations (“public goods game” vs. “teamwork project”) and payoff-visibility manipulations did not change the model’s final action, though they slightly altered the tone of reasoning (more cooperative language, more cautious under uncertainty).

Behavioral Insights. The discrepancy lies in the early rounds: humans sometimes deviate from the Nash equilibrium to “test” cooperation, reflecting bounded rationality, fairness concerns, and learning-by-experimentation. Indeed, experimental economics literature shows that in finitely repeated public goods games, subjects often contribute a non-trivial fraction in early rounds, but contributions tend to decline toward zero as the game nears its known endpoint. For example, Anderson et al. 2024 find that with finite endpoints, contribution levels are high in early rounds but fall off as the end approaches, while random stopping rules sustain cooperation longer. Another

survey (Ledyard 1994) shows voluntary contributions appearing early and then decaying over time in the standard public goods setup.

AI/LLM-Games Literature. Recent studies suggest that LLMs simulate rationality based on payoff descriptions but lack genuine preference uncertainty or adaptive learning. Unlike humans, they do not typically represent or update beliefs about others’ future cooperation nor rely on reciprocity heuristics in a contingent fashion. One relevant paper “Strategic behavior of large language models and the role of game structure versus contextual framing” (Lorè and Heydari (2024)) demonstrates that GPT-4 and LLaMa-2 prioritize the intrinsic mechanics of the game over contextual framing when making decisions. Another work, “Playing repeated games with large language models” (Akata et al. (2025)), analyzes LLMs’ patterns in repeated games and shows that behavior converges rapidly toward equilibrium-like choices when payoffs are explicit, even while framing changes affect the explanations more than the actions.

Proposed Mechanism. A plausible mechanism for the observed divergence is the role of *belief updating and social heuristics*. Humans and LLMs both face the same payoff structure, but their cognitive processes differ:

- **Humans:** Individuals engage in adaptive play. In early rounds, many experiment with cooperation to “test” whether reciprocal behavior is possible. When others free-ride, disappointment or inequity aversion triggers downward adjustments. This is consistent with the literature on conditional cooperation and payoff-based learning: humans update beliefs about others’ strategies and shift accordingly.
- **LLMs:** Large language models instead apply a one-shot “rational choice” model. They do not possess internal belief states that evolve with experience; their choices are generated from static reasoning over the prompt. As a result, their behavior is invariant across rounds and insensitive to others’ past actions, except in how the prompt text is phrased.

Toward a Refinement. To better predict human or AI choices, one could define new refinement concepts that integrate the equilibrium framework with bounded-rational heuristics and linguistic sensitivity:

- **Heuristic-Nash Equilibrium (HNE).** This refinement extends the Nash concept by embedding common social heuristics into players’ best-response functions. Formally, each player’s utility is a convex combination of material payoffs and heuristic payoffs (fairness, reciprocity, or inequality aversion). The equilibrium outcome under HNE reflects both (i) the incentive to free-ride and (ii) the transient cooperative behavior induced by heuristics. This explains

why cooperation often appears in early rounds but decays as the weight on heuristic payoffs diminishes through learning or frustration.

- **Linguistically Framed Equilibrium (LFE).** This refinement models LLM behavior by incorporating linguistic framing as a parameter that shapes the *explanatory reasoning* but not necessarily the *final strategy*. In LFE, the choice set is determined by the payoff logic (often yielding Nash-like predictions), while the explanation or justification layer is influenced by contextual cues (e.g., “teamwork” vs. “public goods”). Thus, framing produces shifts in narrative tone or ethical emphasis without altering the material equilibrium path. This concept helps reconcile why LLMs consistently free-ride despite producing cooperative-sounding justifications.

Summary. The comparative analysis highlights that while the Nash equilibrium remains the long-run attractor, humans and LLMs diverge systematically in how they approach cooperation. Human play is shaped by bounded rationality, belief updating, and social heuristics, which generate transient cooperation and adaptive learning before convergence to free-riding. LLM play, by contrast, reflects static rational-agent optimization, with narrative reasoning sensitive to linguistic framing but strategic choices anchored in payoff logic. To capture these differences, we propose two refinements: the *Heuristic-Nash Equilibrium* (HNE), which integrates fairness and reciprocity heuristics into human best responses, and the *Linguistically Framed Equilibrium* (LFE), which models the separation between payoff-driven choices and framing-sensitive explanations in LLMs. Together, HNE and LFE provide a richer predictive toolkit for anticipating divergences between theoretical predictions, human experiments, and AI-driven behavior.

References (selected): Ledyard (1994); Anderson et al. (2024); Lorè and Heydari (2024); Akata et al. (2025).

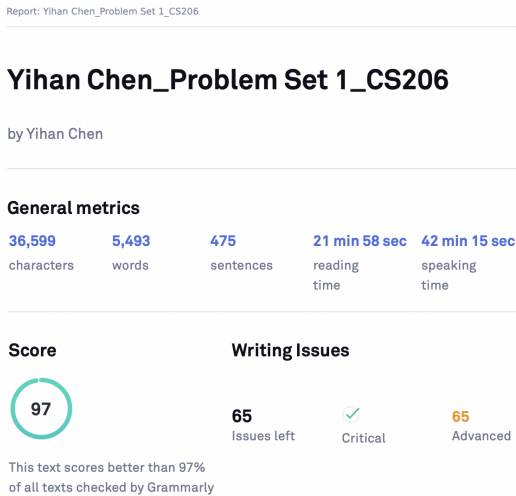


Figure 8: Grammarly Check Report.

References

- Akata, E. et al. (2025). Playing repeated games with large language models. *Nature Human Behaviour*.
- Anderson, L. et al. (2024). Cooperation in public goods experiments with random and finite stopping rules. *Journal of Public Economics / Experimental Economics*. see “Cooperation in public goods experiments with random and ...”.
- Aumann, R. (1974). Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1):67–96.
- Chen, D. L., Schonger, M., and Wickens, C. (2016). otree – an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.
- Daskalakis, C., Goldberg, P. W., and Papadimitriou, C. H. (2009). The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259.
- Glicksberg, I. L. (1952). A further generalization of the kakutani fixed point theorem, with applications to nash equilibrium. *Proceedings of the American Mathematical Society*, 3:170–174.
- Knight, V. (2021). Nashpy: A python library for the computation of equilibria of 2-player strategic games. Version 0.0.28.
- Ledyard, J. O. (1994). Public goods: A survey of experimental research. In Davis, D. D. and Holt, C. A., editors, *Voluntary Contributions to Public Goods: A Survey of Experimental Research*. Springer / etc.
- Ledyard, J. O. (1995). Public goods: A survey of experimental research. In Kagel, J. H. and Roth, A. E., editors, *The Handbook of Experimental Economics*, pages 111–194. Princeton University Press.
- Lorè, N. and Heydari, B. (2024). Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14:18490.
- Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36:48–49.
- Osborne, M. J. and Rubinstein, A. (1994). *A Course in Game Theory*. MIT Press.
- Sargent, T. J. and Stachurski, J. (2021). Quantitative economics (python). Version 0.5.1.
- Savani, R. and von Stengel, B. (2015). Game theory explorer – software for the applied game theorist. *Computational Management Science*, 12:5–33.
- Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4:25–55.

Shoham, Y. and Leyton-Brown, K. (2008). *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press.