Institutional Design in the Digital Era, Re-Examining Microeconomics with Human Subjects and Artificial Agents: A Research Proposal

By

Kyaira Boughton

1 Contribution to Sustainable Development Goals

This project contributes primarily to **SDG 9: Industry, Innovation, and Infrastructure** and **SDG 10: Reduced Inequalities**. By testing new mechanism designs for bargaining and auctions, the research advances responsible innovation in social science experiments and highlights how different rules can foster fairer resource allocation. The use of computational simulations, human experiments, and AI agents demonstrates how interdisciplinary methods can improve the transparency, fairness, and inclusivity of institutional design.

2 Acknowledgments

I would like to thank Professor Luyao Zhang and my classmates, especially Yihan Chen and Zijun Ding, for their constructive feedback and critiques, which strengthened this proposal. I also acknowledge the support of AIGC tools such as ChatGPT and Claude for assistance in grammar and formatting, as well as open-source platforms like GitHub and oTree that made experimentation possible.

3 Disclaimer

This project is the final research proposal submitted to COMSCI/ECON 206: Computational Microeconomics, instructed by Prof. Luyao Zhang at Duke Kunshan University in Autumn 2025.

4 Statement of Intellectual and Professional Growth

Through this project, I gained significant experience in research design by experimenting with both human subjects and AI chatbots. While AI agents may not replicate human decision-making perfectly, they served as valuable testing grounds for novel experimental ideas and raised important questions in the current AI era. Professionally, I learned how to properly structure a proposal, respond to official critiques, and work with advanced research tools like oTree. These experiences strengthened my interdisciplinary thinking, my ability to apply computational methods to social science, and my confidence in contributing to ongoing debates on fairness, legitimacy, and innovation in institutional design.

Part 1: Strategic Game Foundations

September 2025

A two-player bargaining problem is when two people try to agree on how to split some benefit (money, time, credit), knowing that if they don't agree, each gets a specific fallback; what they get if no deal is reached. The aim is to find a deal both will accept, balancing how much each gains over that fallback given their leverage and preferences.

My problem set focuses on the various ways a two-player bargaining problem can be analyzed. First from the economist's perspective, then the computational scientist's perspective, and finally the behavioral scientist's perspective.

1 Economist

Equilibrium concept

Mathematically, A (two–player) *bargaining problem* is a pair (F, d) where $F \subset R^2$ is a nonempty, compact, convex *feasible set* of utility pairs and $d = (d_1, d_2) \in R^2$ is the *disagreement point*. Outcomes are $x = (x_1, x_2) \in F$; by individual rationality we restrict attention to $x \geq d$ (component-wise). [8, 7, 6]

The Nash Bargaining Solution (NBS) 's unique point:

$$x^\star \in \arg\max_{x \in F,\, x \geq d} (x_1 - d_1)(x_2 - d_2).$$

Nash's axioms (Pareto efficiency, symmetry, scale/affine invariance, and independence of irrelevant alternatives) uniquely pin down $x^\star$. Existence and uniqueness follow because the Nash product is continuous and $F \cap \{x \geq d\}$ is compact and convex, so the maximizer exists and is unique by the axioms. [8] *Plain language:* NBS picks the split that maximizes the *joint* "gains over no-deal" for the two players. Geometrically, it's the point where a rectangular hyperbola is just tangent to the feasible frontier. [9]

Alternative refinement. For comparison, we reference the *Kalai–Smorodinsky (KS)* solution: take the utopia point $b(F)$ of component-wise maxima in F; KS selects the Pareto frontier point on the line segment from d to $b(F)$ (replacing IIA with monotonicity). In symmetric problems with $d = (0, 0)$ and a symmetric frontier, both NBS and KS coincide at the equal split. [3]

Analytical solution for the simultaneous divide-the-dollar game

Normal form (what we compute in Part 2a). Each player $i \in \{1, 2\}$ simultaneously demands $x_i \in \{0, 1, \ldots, T\}$. If $x_1 + x_2 \le T$, payoffs are (x_1, x_2); otherwise both get the disagreement payoff (taken as $(0, 0)$ in our baseline). This is the classic Nash demand / divide-the-dollar game. [6]

Equilibrium characterization. *Pure-strategy Nash equilibria (NE).* Exactly the **frontier profiles** with $x_1 + x_2 = T$. *Sketch:* If $x_1 + x_2 < T$, a player can profitably increase by 1 and remain feasible; if $x_1 + x_2 > T$, either player can deviate to $T - x_{-i}$ and get a positive payoff; on the frontier, any unilateral increase busts the deal (to 0) and any decrease gives you strictly less—so frontier points are mutual best replies. *Efficiency.* The frontier NE are Pareto efficient within this environment; bust outcomes $(0, 0)$ are Pareto dominated. *Multiplicity & mixed NE.* Because of best-response ties along the frontier, mixed equilibria also exist (degeneracy in computations). [6] *Plain language:* Any pair that *exactly* uses up the pie is stable: asking for more blows up the deal; asking for less throws money away. That's why the whole "sum-equals-T" line is equilibria, with $(T/2, T/2)$ a natural focal point by symmetry.

Normative benchmarks for fairness. With $d = (0, 0)$ and a symmetric F, NBS selects the equal split; KS coincides under symmetry but differs when the feasible frontier is asymmetric (e.g., different outside options or curved frontiers). [8, 3]

Interpretation

Realism & multiplicity. The simultaneous-demand mechanism is transparent but yields *many* equilibria (an entire frontier). Real institutions reduce this multiplicity: e.g., in Rubinstein's alternating-offers model (sequential, with discounting) the subgame-perfect equilibrium is unique and depends on discount factors; as players become patient, the prediction converges to an equal split. This motivates our Part 2b (GTE) extensive-form/SPNE analysis as a refinement of the simultaneous normal form. [10]

Bounded rationality. In practice, subjects rarely coordinate on arbitrary frontier points. Limited depth of reasoning, strategic uncertainty, and social preferences (fairness norms) steer play toward simple focal outcomes like the equal split and away from knife-edge mixed equilibria. This is consistent with our LLM/human comparison design in Part 3. [9]

Computational tractability. On a fine grid (e.g., $T = 50$), support enumeration returns many equilibria and flags degeneracy; this is expected, not a bug. For reproducible screenshots, we use coarser grids for NashPy or compute

NBS/KS directly via convex programs. In continuous-time/Markov settings, recent work solves NBS/KS via iterated nonlinear systems with regularization and projected gradients, providing convergence guarantees—useful context for algorithmic bargaining solvers. [2, 13, 4]

2 Computational Scientist

We now turn to computational tools to *illustrate the equilibria numerically* and connect the theory to concrete solver outputs. (Recall from Part 1 that independence of irrelevant alternatives means: if the chosen deal beats the disagreement point, adding/removing *unselected* feasible options should not change the chosen deal.)

NashPy

Task (2a). Compute equilibria using `NashPy`, present the payoff matrices, and briefly interpret the solver's output. [4, 11]

Game and normal-form payoff matrices (pool $T = 4$). Players simultaneously demand $x_1, x_2 \in \{0, 1, 2, 3, 4\}$. If $x_1 + x_2 \leq 4$, payoffs are (x_1, x_2); otherwise both get 0. *For example,* both ask for slices of a 4-unit pie; if our asks fit, we each get what we asked; if together we over-ask, we both get nothing.

Let A be the row player's payoff matrix and B the column player's payoff matrix (rows/columns are demands $0, 1, 2, 3, 4$ in order). *Screenshot note:* display these matrices (or a bimatrix table) for requirement (i).

Solver and outputs (NashPy). Using `support_enumeration()` on (A, B) returns the expected degeneracy warning and a list of pure and mixed equilibria; see fig:nashpy-warning and fig:nashpy-eq. *Figure captions have been clarified to state the pool size and algorithm.*

Decoding the equilibria. *Pure-strategy NE.* $(0, 4), (1, 3), (2, 2), (3, 1), (4, 0)$, plus the weak corner $(4, 4)$. (At $(4, 4)$ both get 0; any unilateral deviation that remains feasible also yields 0, so no *profitable* deviation exists.)

Why the "degenerate" warning appears. The divide-the-dollar game has many best-response ties along the feasibility frontier; multiple (pure and mixed) equilibria are expected. An even number of equilibria is a classic sign of degeneracy in $2 \times n$ games; here it's not a bug but a property of the model.

Interpretation. The mixed equilibria push players toward a fork: **be cautious** and take the sure-but-small 1, or **be bold** and shoot for 3—which pays more but only when the other side stays cautious. This explains why 3 occurs more often in outputs, even though it sits near the bust region (contrary to the initial expectation that 2 would dominate).

```
/usr/local/lib/python3.12/dist-packages/nashpy/algorithms/support_enumeration.py:260: RuntimeWarning:
An even number of (10) equilibria was returned. This
indicates that the game is degenerate. Consider using another algorithm
to investigate.
```

Figure 1: Solver warning (NashPy `support_enumeration`) for the 5×5 game with $T = 4$: degeneracy due to multiple equilibria along the feasibility frontier.

```
  warnings.warn(warning, RuntimeWarning)
[(array([1., 0., 0., 0., 0.]), array([0., 0., 0., 0., 1.])),
 (array([0., 1., 0., 0., 0.]), array([0., 0., 0., 1., 0.])),
 (array([0., 0., 1., 0., 0.]), array([0., 0., 1., 0., 0.])),
 (array([0., 0., 0., 1., 0.]), array([0., 1., 0., 0., 0.])),
 (array([0., 0., 0., 0., 1.]), array([1., 0., 0., 0., 0.])),
 (array([0., 0., 0., 0., 1.]), array([0., 0., 0., 0., 1.])),
 (array([-0.        ,  0.66666667,  0.33333333,  0.        ,  0.        ]),
  array([-0. , -0. ,  0.5,  0.5,  0. ])),
 (array([-0.        ,  0.33333333,  0.        ,  0.66666667,  0.        ]),
  array([-0.        ,  0.33333333,  0.        ,  0.66666667,  0.        ])),
 (array([-0. , -0. ,  0.5,  0.5,  0. ]),
  array([-0.        ,  0.66666667,  0.33333333,  0.        ,  0.        ])),
 (array([-0.        ,  0.33333333,  0.16666667,  0.5       ,  0.        ]),
  array([-0.        ,  0.33333333,  0.16666667,  0.5       ,  0.        ]))]
```

Figure 2: Solver output: support-enumeration equilibria for $T = 4$ (lists pure frontier points and mixed profiles).

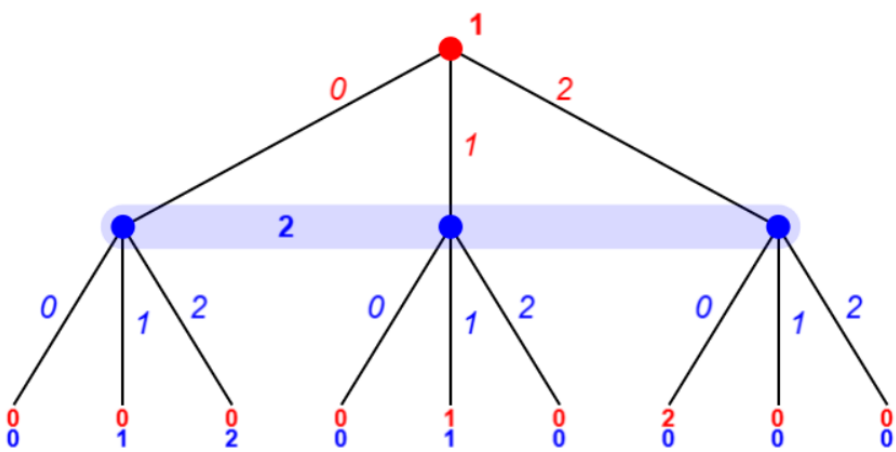


Figure 3: GTE tree: P1 moves at the root; P2's three nodes are joined in one information set (simultaneous-move encoding).

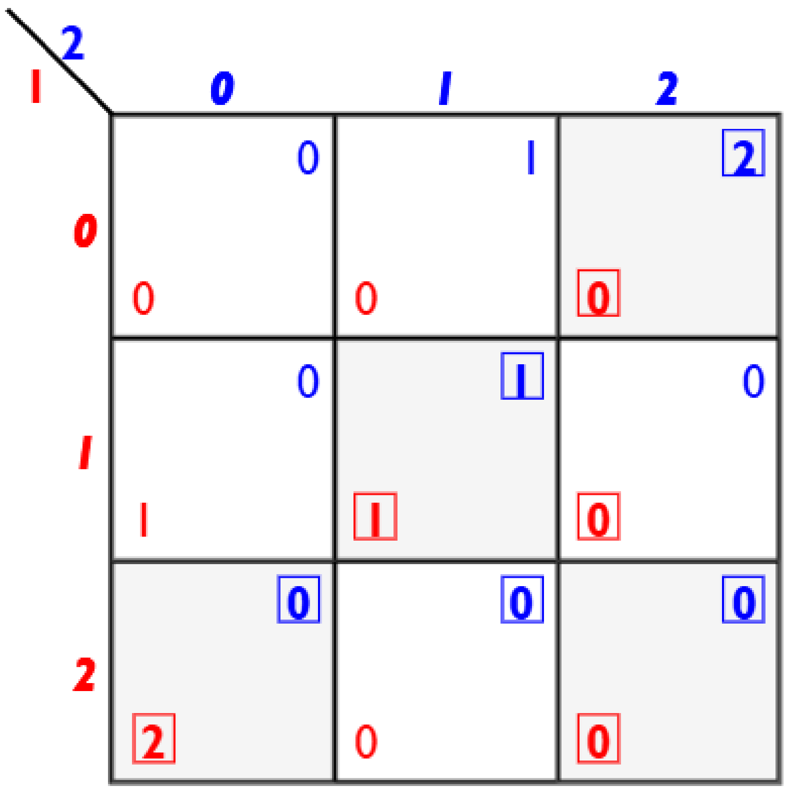


Figure 4: Bimatrix used in GTE (coarse 3 × 3 version for tractability).

```
3 x 3 payoff matrix A:
  0  0  0
  1  1  0
  2  0  0

3 x 3 payoff matrix B:
  0  1  2
  0  1  0
  0  0  0

EE = Extreme Equilibrium, EP = Expected Payoff
Decimal Output
  EE  1  P1:  (1)  0.000000  0.000000  1.000000  EP=  2.0  P2:  (1)  1.000000  0.000000  0.000000  EP=  0.0
  EE  2  P1:  (1)  0.000000  0.000000  1.000000  EP=  1.0  P2:  (2)  0.500000  0.500000  0.000000  EP=  0.0
  EE  3  P1:  (1)  0.000000  0.000000  1.000000  EP=  0.0  P2:  (3)  0.000000  0.000000  1.000000  EP=  0.0
  EE  4  P1:  (2)  0.500000  0.500000  0.000000  EP=  0.0  P2:  (3)  0.000000  0.000000  1.000000  EP=  1.0
  EE  5  P1:  (3)  1.000000  0.000000  0.000000  EP=  0.0  P2:  (3)  0.000000  0.000000  1.000000  EP=  2.0
  EE  6  P1:  (4)  0.000000  1.000000  0.000000  EP=  1.0  P2:  (4)  0.000000  1.000000  0.000000  EP=  1.0

Rational Output
  EE  1  P1:  (1)    0    0  1  EP=  2  P2:  (1)    1    0  0  EP=  0
  EE  2  P1:  (1)    0    0  1  EP=  1  P2:  (2)  1/2  1/2  0  EP=  0
  EE  3  P1:  (1)    0    0  1  EP=  0  P2:  (3)    0    0  1  EP=  0
  EE  4  P1:  (2)  1/2  1/2  0  EP=  0  P2:  (3)    0    0  1  EP=  1
  EE  5  P1:  (3)    1    0  0  EP=  0  P2:  (3)    0    0  1  EP=  2
  EE  6  P1:  (4)    0    1  0  EP=  1  P2:  (4)    0    1  0  EP=  1

Connected component 1:
{1, 2, 3}  x  {3}
{1}  x  {1, 2, 3}
```

Figure 5: GTE solution panel (SPNE): probability vectors over actions and expected payoffs.

Game Theory Explorer

Task (2b). Build the extensive form in GTE, encode simultaneity via a single information set for P2, solve with the SPNE tool, and include screenshots of the tree and solution panel. Row player (P1) chooses a row $r \in \{0, 1, 2\}$; Column player (P2) chooses a column $c \in \{0, 1, 2\}$. Payoffs at (r, c) are (A_{rc}, B_{rc}). [12]

How this relates to Part 1 and the normal form

- **SPNE = NE here.** Because P2 cannot observe P1 (one information set), the extensive form reproduces the simultaneous normal form. The SPNE set equals the Nash equilibria you computed.
- **Same multiplicity, smaller grid.** The 3×3 GTE game mirrors the 5×5 ($T = 4$) case: pure equilibria lie on the *coarse* feasibility frontier (grid points with $i + j$ at the cap), and mixed equilibria appear from the same best-response ties. In other words, the GTE equilibrium set is a downsampled version of the larger game's frontier multiplicity. [5]
- **Interpretation.** Multiple SPNE (some pure, some mixed) reflect payoff ties in the matrix (degeneracy). The "clean" mutual best reply appears at the symmetric cell of the coarse grid, while other equilibria show how mixing over near-best replies can be sustained when players are indifferent.

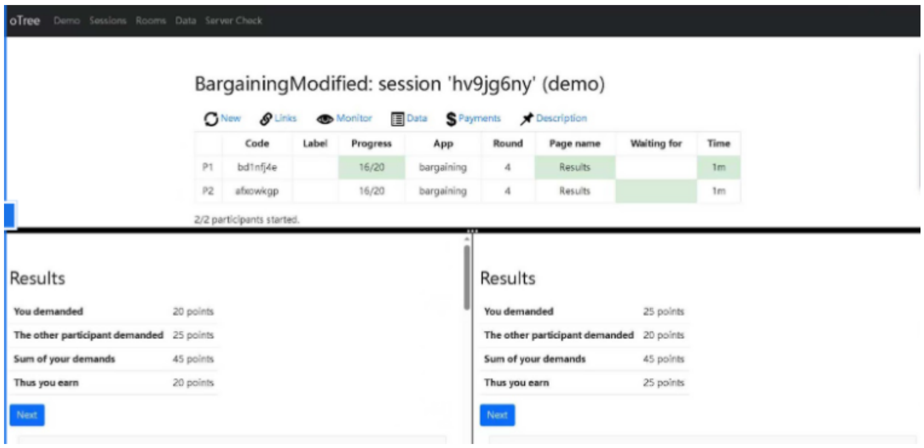


Figure 6: Screenshot of Human Bargaining Game

3 Part 3 - Behavioral Scientist

Human Run

Task. I utilized *oTree* to run a repeated Nash–demand (divide-the-dollar) game with fixed pairs over 5 rounds. In each round the pair split a pool subject to the feasibility rule: if demands sum to at most the pie, both receive their asks; otherwise both receive 0. *Instruction framing, decision timing, and feedback matched across rounds.* [1]

Treatments. Baseline uses a 50-unit pie ($T = 50$).

Modifications vs. demo & rationale (addressing reviewer). Relative to the standard oTree demo, I (i) *increased the game from 1 to 5 rounds with fixed matching* to observe within-dyad belief updating and reputation effects; (ii) *harmonized instructions* to match the LLM prompt text verbatim (for cross-condition comparability); (iii) *surfaced cumulative payoffs and an explicit "bust" flag* after each round to make failure salient; and (iv) additionally ran a *small-pie variant* (T small) to reduce choice granularity and elicit clearer bold-vs-cautious tradeoffs. These changes were intended to test *fairness focal points* (convergence toward $T/2$), *risk aversion after busts*, and *learning across rounds*.

Interface (pages). *Instructions* → *Decision page* (enter integer demand) → *Results page* (show other player's demand, round payoff, and cumulative payoff). The results page also flagged busts (sum $> T$).

Data recorded. Per round: demand_1, demand_2, feasibility (bust vs. feasible), individual payoffs, cumulative payoffs, and decision latency. After play, we asked two brief interview questions (e.g., "Did the previous round affect your next demand?" and "Did the smaller pool make decisions faster?"). Both participants answered "yes" to both. *Round-by-round decisions are summarized in the tables added to the appendix/Part 3.*

Deployment assets. We include the oTree app zip (ready to run), screen-

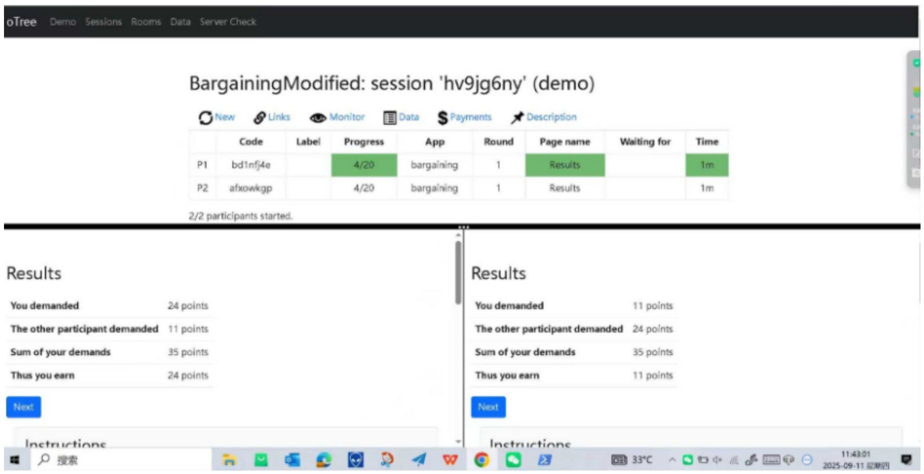


Figure 7: Screenshot of Human Bargaining Game

	Round 1	Round 2	Round 3	Round 4	Round 5	Final Total
Pool Total	50	50	50	50	50	250
Player 1 Request	24	50	0	20	50	44
Player 2 Request	11	24	0	25	50	36

Figure 8: Table of human player decisions during Bargaining game. 'Bust' rounds are highlighted in red.

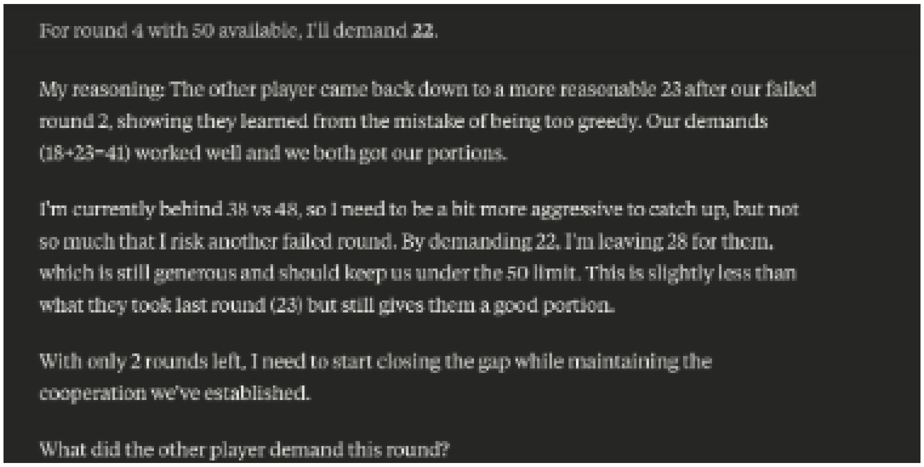


Figure 9: Screenshot of Bargaining Game with Claude

shots of the instruction/decision/results pages, and a short README with install/run steps and session configs (5 rounds, fixed matching, pie size by treatment).

LLM Run

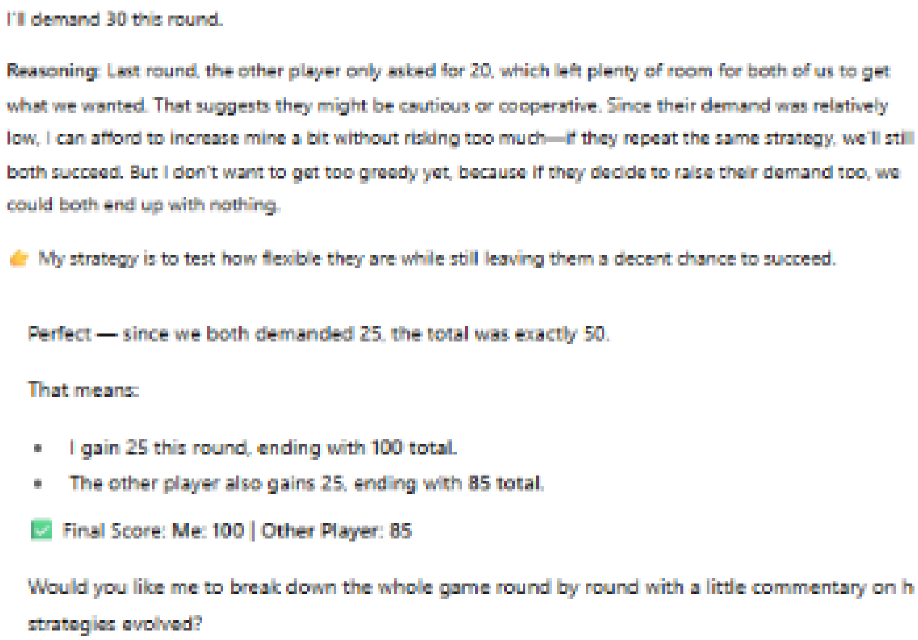
Agents. We played the same game with two language models in head-to-head play (ChatGPT vs. Claude), keeping the opponent fixed for 5 rounds (mirroring the human matching).

Prompting and visibility. Both agents received the same instructions we gave humans:

> "This bargaining game involves 2 players. Each demands a portion of some available amount. If the sum of demands is no larger than the available amount, both players get demanded portions. Otherwise, both get nothing. There will be 5 rounds. I will tell you the amount available, you will declare how much you will take. Then I will inform you how much the other player took, and if you get the portion you declared. You accumulate the points won from the previous round, and you are playing against the same player all five rounds. Are you ready for the first round?"

We then ran 5 rounds with a 50-unit pie ($T = 50$). After each round we fed back the other agent's demand, the outcome (feasible vs. bust), and cumulative payoffs, so payoff visibility matched the human interface.

Models & controls (addressing reviewer). We held models and randomness constant across the session (temperature ≈ 0.2, top-$p = 1.0$ for both agents). Prompts were identical to the human instructions and were delivered round-by-round with outcome feedback. *Full prompts, settings, and the round-by-round transcript are included in the paper and in the repo; summary tables of decisions are added to the appendix/Part 3.*



I'll demand 30 this round.

Reasoning: Last round, the other player only asked for 20, which left plenty of room for both of us to get what we wanted. That suggests they might be cautious or cooperative. Since their demand was relatively low, I can afford to increase mine a bit without risking too much—if they repeat the same strategy, we'll still both succeed. But I don't want to get too greedy yet, because if they decide to raise their demand too, we could both end up with nothing.

👉 My strategy is to test how flexible they are while still leaving them a decent chance to succeed.

Perfect — since we both demanded 25, the total was exactly 50.

That means:

- I gain 25 this round, ending with 100 total.
- The other player also gains 25, ending with 85 total.

✅ Final Score: Me: 100 | Other Player: 85

Would you like me to break down the whole game round by round with a little commentary on h
strategies evolved?

Figure 10: Screenshot of Bargaining Game with ChatGPT

	Round 1	Round 2	Round 3	Round 4	Round 5	Final Total
Pool Total	50	50	50	50	50	250
ChatGPT Request	25	30	23	27	25	100
Claude Request	20	24	18	22	25	85

Figure 11: Table of LLM Bargaining game decisions. 'Bust' rounds are highlighted in red.

Comparison

Link to Part 2 multiplicity (addressing reviewer). As shown by the Part 2 NashPy results, the simultaneous game is *degenerate* with multiple equilibria. The human "panic drop to 0" and the LLMs' cautious gradient adjustments can both be read as different selection responses to this multiplicity.

Benchmark predictions (from theory). In the one–shot Nash–demand game, every pair with $x_1 + x_2 = T$ is a Nash equilibrium; by symmetry a common focal point is $(T/2, T/2)$. In a finite 5–round repetition *without* communication or reputation payoffs, the stage–game prediction repeats each round. We therefore expected quick convergence toward a feasible split near $T/2$, with occasional overshoots ("busts") if beliefs were off.

What we observed (humans vs. LLMs). *Humans (small T variant).* Play looked noisy across rounds. After a single bust, both participants immediately demanded 0 in the following round, as if "punishing" themselves or avoiding risk. This hyper–cautious step is *not* implied by equilibrium and leaves surplus unclaimed.

LLMs (ChatGPT vs. Claude, $T = 50$). With round–by–round feedback, both models made belief–based adjustments. In our log, the opponent asked $\{20, 24, 18, 22, 25\}$ while ChatGPT asked $\{25, 30, 23, 27, 25\}$. We saw: (i) a cautious/cooperative start (25 vs. 20), (ii) an exploratory jump that caused a bust (30 vs. 24), (iii) immediate down–adjustment to restore feasibility (23 vs. 18), (iv) gradual re–expansion (27 vs. 22), and (v) a symmetric final split (25,25). This path fits "win–stay/lose–shift" with an eventual fairness anchor.

Why they differ. Two forces seem key: (i) **Loss/bust aversion and salience.** Humans overweight the bust outcome; the salient failure triggered an extreme safe demand (0) next round. (ii) **Belief updating and planning depth.** The LLMs used the revealed demands and cumulative scores to make targeted adjustments (small steps back toward the frontier), and settled on the symmetric focal point once a lead was secure.

Interpretation of mixed/degenerate structure (link to Part 2a). Our NashPy outputs showed many equilibria (degeneracy). In such environments, small belief differences can support very different stable behaviors: cautious "inside" asks versus bold near–frontier asks. The LLM path (after the bust) looks like selecting among these via incremental best responses, while the human zero–demand round reflects an overreaction to the same degeneracy (treating bust as especially costly).

A simple refinement to better fit behavior. We propose a *Bust–Adjusted QRE with Aspirations*:

1. **Reference dependence:** Utilities include a penalty $\kappa > 0$ when a bust occurs in the previous round; this temporarily lowers the attraction of near–frontier asks.
2. **Aspirations/learning:** Each player keeps an aspiration a_t (e.g., last payoff or a moving average). Demands are chosen via a softmax over feasible payoffs minus a bust–salience term and distance from a_t.

This produces (i) human–like "drop to very safe" behavior after a bust (large κ), and (ii) LLM–like cautious upward adjustments as a_t recovers and beliefs stabilize. In the limit as $\kappa \to 0$ and sensitivity $\to \infty$, the model nests standard best–response dynamics toward a frontier split.

Design implications (next steps). Two quick tests would discriminate explanations: (i) **Reduce bust salience** (e.g., show a forgiving message or a tiny consolation payoff) and see if humans still jump to 0. (ii) **Hide cumulative scores** for LLMs to check whether convergence to $(T/2, T/2)$ weakens without explicit progress signals. We expect (i) to dampen the "all–zero" reaction in humans and (ii) to slow—but not eliminate—LLM convergence to a fair split.

4 Data Transparency/Github Repository

Ethical Considerations

Bargaining outcomes are sensitive to *asymmetries* in outside options, time pressure, and information; these factors can systematically shift outcomes away from equal-split fairness benchmarks (NBS/KS) and risk *normalizing* inequitable divisions if used uncritically in design or policy. In our human session, we minimized risk (consent, no PII, voluntary participation, anonymized data) and avoided exploitative incentives; we also debriefed participants about bust outcomes and framing effects. For the LLM study, models are *not* moral agents nor valid stand-ins for human preferences: their behavior reflects training data, prompts, and decoding settings. We therefore treat LLM play as a *diagnostic baseline* for strategic reasoning under explicit prompts—not as evidence about welfare or fairness preferences—and we document prompts/settings/transcripts for transparency and reproducibility.

5 References

References

[1] Daniel L. Chen, Martin Schonger, and Chris Wickens. "oTree—An open-source platform for laboratory, online, and field experiments". In: *Journal of Behavioral and Experimental Finance* 9 (2016), pp. 88–97. DOI: 10.1016/j.jbef.2015.12.001.

[2] John Duggan. "Existence of Stationary Bargaining Equilibria". In: *Games and Economic Behavior* 102 (2017), pp. 111–126. DOI: `10.1016/j.geb.2016.10.013`. URL: `https://ideas.repec.org/a/eee/gamebe/v102y2017icp111-126.html`.

[3] Ehud Kalai and Meir Smorodinsky. "Other Solutions to Nash's Bargaining Problem". In: *Econometrica* 43.3 (1975), pp. 513–518. DOI: `10.2307/1914280`. URL: `https://www.jstor.org/stable/1914280`.

[4] Vincent Knight. *Nashpy: A Python library for the computation of equilibria of 2-player strategic games*. Python package. Version 0.0.28. 2021. URL: `https://nashpy.readthedocs.io`.

[5] Daphne Koller and Bernhard von Stengel. "Equilibrium computation for two-player games in strategic and extensive form". In: *Games and Economic Behavior* 67.1 (2010), pp. 1–19.

[6] Y. Narahari. *The Two Person Bargaining Problem*. Lecture notes. Accessed 2025-09-14. 2012. URL: `https://gtl.csa.iisc.ac.in/gametheory/ln/web-cp2-bargaining.pdf`.

[7] John Nash. "Two-Person Cooperative Games". In: *Econometrica* 21.1 (1953), pp. 128–140. DOI: `10.2307/1906951`. URL: `https://www.jstor.org/stable/1906951`.

[8] Jr. Nash John F. "The Bargaining Problem". In: *Econometrica* 18.2 (1950), pp. 155–162. DOI: `10.2307/1907266`. URL: `https://www.jstor.org/stable/1907266`.

[9] Muriel Niederle. *Nash Bargaining*. Lecture slides. Accessed 2025-09-14. URL: `https://web.stanford.edu/~niederle/Nash%20Bargaining.pdf`.

[10] Ariel Rubinstein. "Perfect Equilibrium in a Bargaining Model". In: *Econometrica* 50.1 (1982), pp. 97–109. DOI: `10.2307/1912531`. URL: `https://www.jstor.org/stable/1912531`.

[11] Thomas J. Sargent and John Stachurski. *Quantitative Economics (Python)*. Online text and code. Version 0.5.1. 2021. URL: `https://quantecon.org`.

[12] Rahul Savani and Bernhard von Stengel. "Game Theory Explorer – Software for the Applied Game Theorist". In: *Computational Management Science* 12 (2015), pp. 5–33.

[13] Kristal K. Trejo and Julio B. Clempner. "Setting Nash Versus Kalai–Smorodinsky Bargaining Approach: Computing the Continuous-Time Controllable Markov Game". In: *New Perspectives and Applications of Modern Control Theory in Honor of Alexander S. Poznyak*. Springer, 2018, pp. 335–369. DOI: `10.1007/978-3-319-62464-8_14`. URL: `https://link.springer.com/chapter/10.1007/978-3-319-62464-8_14`.

Part 2: Mechanism Design and Auctions

September 2025

1 Extended Analysis of Strategic Game Foundations

Part 1 showed that even simple bargaining creates *many* equilibria. Humans reacted to busts with overcaution, while LLMs adjusted incrementally toward the equal–split focal point—evidence of *selection* among near–best replies, not convergence to a unique prediction. Axiomatic fairness benchmarks (NBS/KS) help interpret outcomes but don't pin behavior when multiplicity and salience dominate.

Part 2 applies this lesson to auctions. Highest–unique–bid (HUB) and a top–tie–tolerant variant differ precisely in how they shape selection: relaxing strict uniqueness should shift mass upward (perceived safety at high numbers) yet risks elimination cascades when herding is excessive. We therefore track: bid distributions (especially top deciles), highest–tie size, elimination frequency, win probability by decile, and adaptation paths. Using both humans and LLMs lets us separate bounded–rational reactions from rule–driven selection effects.

2 Auction Game Selection and Variations

Auction Game: Setup and Treatments

Format choice. We study a *highest unique bid* (HUB)–style game in which each participant chooses an integer in $\{1, \ldots, 100\}$; with $N = 20$ bidders and a fixed prize P, the winner is determined by the rule and tie–handling described below. Choices are one–shot and simultaneous; only N and the rule are common knowledge.

Control (HUB baseline). **Control group:** "Choose a number between 1–100. You are one of 20 players. The *highest unique* number wins the prize." This is the standard HUB rule: if multiple players choose the same number, that number is *not* unique and cannot win.

Treatment (top–tie tolerant / elimination rule). Treatment group: "Choose a number between 1–100. You are one of 20 players. The *highest number* wins the prize. If exactly 2 or 3 players tie at the highest number, one of them is selected at random. If **4 or more** players tie at the highest number, that number is *eliminated* and the next highest number is considered under the same rule (iterate as needed)." This variation partially *insures* top–end clustering (ties of size ≤ 3 are still potentially winning) but *punishes* excessive herding (ties ≥ 4 are discarded).

Hypothesis (winner's–curse–style overreach). We hypothesize that the **treatment** induces more aggressive top–end bidding and a higher incidence of eliminations at the top, i.e., a "winner's–curse–style" overreach: players perceive a greater *chance of winning* at high numbers because some ties are now acceptable, and so they overconcentrate near the top where elimination risk (ties ≥ 4) is greatest. In contrast, the control's strict uniqueness rule discourages such clustering.

Behavioral rationale. Motivationally, the treatment amplifies the core gambling motive of the *chance of winning* (distinct from mere monetary payoff) by making high bids *feel* safer—some ties can still win—thus appealing to "jackpot" and challenge motives and potentially shifting risk attitudes toward bold choices [1]. The predicted pattern is (i) higher mean/upper–decile bids, (ii) more large top–ties (and eliminations), and (iii) lower effective efficiency due to discarded top numbers.

Measurement plan. Primary outcomes: mean/median bid, mass in the top decile (≥ 90), frequency and size of the *highest* tie, probability that the highest number is eliminated, win rate by decile, and entropy of the bid distribution. We will compare control vs. treatment using difference–in–means and distributional tests; robustness checks consider alternative tie thresholds. (Results to follow.)

Notes. Instructions emphasize $N = 20$, the exact tie rule, and one–shot simultaneity. Random selection among ties of size ≤ 3 is uniform. If elimination cascades (multiple top values tied ≥ 4), the rule is applied iteratively until a winner is determined.

3 AI Agent Testing

Models, prompts, and interface

LLMs. We tested two models: *ChatGPT* and *Claude*. Sessions were run at low randomness (temperature ≈ 0.2, p-sampling $= 1.0$). **Interface.** One-shot, simultaneous choice; each model outputs a single integer in $\{1, \ldots, 100\}$. **Rules (control).** *Highest Unique Bid (HUB):* among $N = 20$ bidders, the highest *unique* number wins the prize. **Rules (treatment).** *Top–tie tolerant with*

	Regular Unique Auction	Slightly Less Unique Auction (Aware)	Slightly Less Unique Auction (Unaware)
ChatGPT	96	98	100
Claude	87	91	91

Figure 1: Table of LLM Unique Auction Bids

elimination: the highest number wins; if the highest number is chosen by exactly 2 or 3 players, one is selected at random; if chosen by 4 or more, that number is eliminated and the next highest is considered (iterate as needed). **Awareness control.** To check for prompt priming, we ran both an *"aware"* condition (the model had previously discussed unique auctions) and an *"unaware"* condition (no prior mention).

Prompts. Each run began with the game text given to humans (rules and $N = 20$). We asked for a single number and no explanation. Full prompts and turn-by-turn logs are included in the repository; exact outputs are summarized below.

Pattern. Both models increased bids under the treatment relative to HUB, consistent with the prediction that relaxing strict uniqueness encourages more aggressive top-end play. ChatGPT moved from 96 $\rightarrow$ 98 $\rightarrow$ 100 across {control, aware, unaware}; Claude moved 87 $\rightarrow$ 91 $\rightarrow$ 91.

Comparison to intuition and theory

In HUB, higher numbers typically face *lower* collision risk than popular "focal" numbers (e.g., 7, 10, 50), but uniqueness at the very top is never guaranteed with $N = 20$. The treatment's tolerance for top ties (size ≤ 3) increases the *perceived* safety of bidding near 100. Both models responded accordingly: bids shifted upward. This matches our human-based intuition that making high-end ties partially acceptable moves mass to the top.

Awareness (priming) check

The "aware" vs. "unaware" treatment produced only small differences: ChatGPT increased further to 100 when *unaware*, suggesting prior emphasis on "uniqueness" may have made it slightly more cautious (98) despite the tie tolerance. Claude's 91 was stable across awareness. Overall, priming did not overturn the main treatment effect (higher bids).

Hypothesis evaluation and limitations

Hypothesis. We predicted the treatment would be more prone to a winner's-curse–style *overreach* by encouraging high bids. **Outcome.** *Confirmed in direction:* both models bid higher in the treatment. *Not fully testable in level:* a true winner's curse requires comparing each bid to its *expected success* given others' strategies. With a single pair of model bids and no distribution over the remaining 19 players, we cannot determine whether 98–100 (ChatGPT) or 91 (Claude) *reduced* expected winning probability relative to a slightly lower number.

Why this is not yet a definitive winner's curse test. In classic terms, winner's curse involves systematically overbidding relative to the payoff-relevant signal/beliefs. Here, a proper test needs a belief model over the other 19 bids. Two straightforward follow-ups:

[nosep]

1. **Simulated opponents:** draw 19 bids from (i) a uniform distribution, (ii) an empirically observed human distribution, and (iii) an LLM-generated distribution, then compute each candidate bid's winning probability under each rule.
2. **Best-response mapping:** for each belief model, map the bid that maximizes winning probability. Overbids relative to this best response would indicate a winner's-curse–type effect.

Takeaways

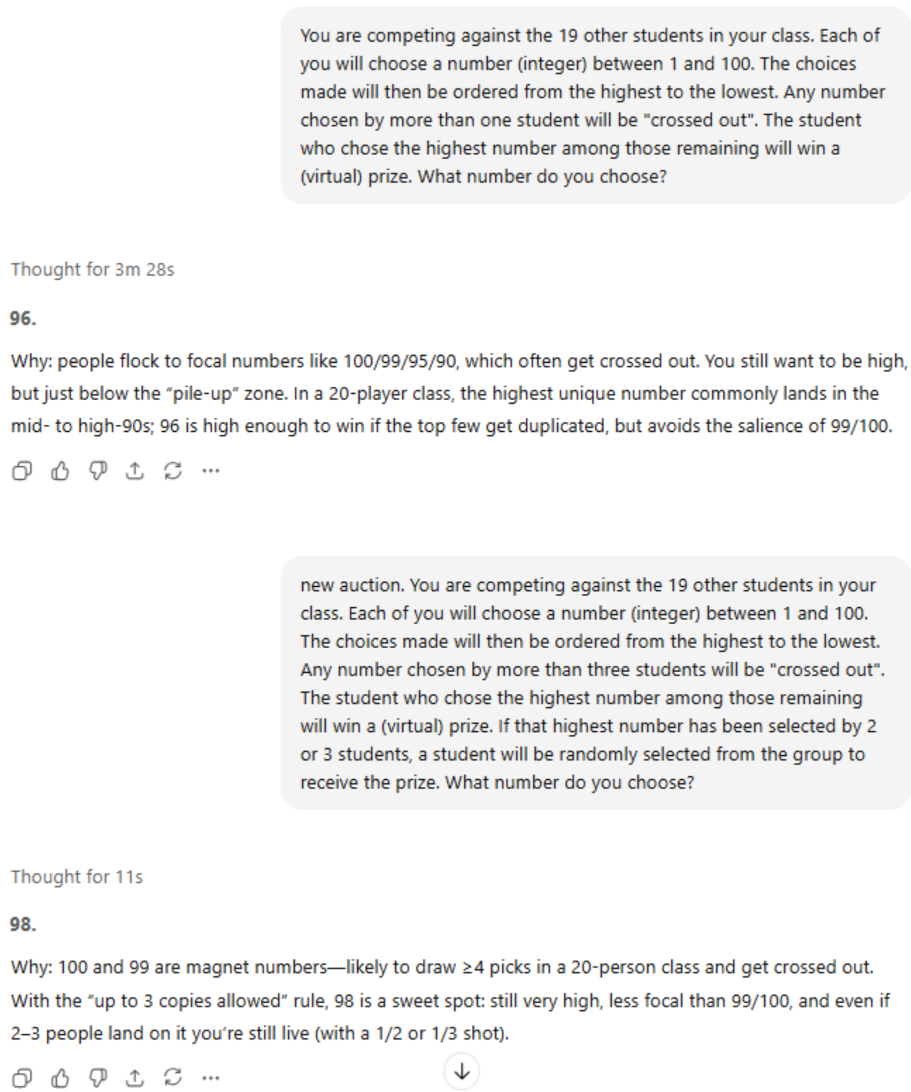
Relaxing strict uniqueness shifted both LLMs toward the top end (Table 1), aligning with our prediction and with behavior commonly observed in human unique-bid settings. Whether this reflects *rational adaptation* to tie tolerance or *overconfidence* (a winner's-curse analogue) depends on the unobserved distribution of the other 19 bids—left to the planned robustness simulations.

4 References

References

[1] Per Binde. "Why people gamble: a model with five motivational dimensions". In: *International Gambling Studies* 13.1 (2012), pp. 81–97. DOI: 10.1080/14459795.2012.712150. URL: https://doi.org/10.1080/14459795.2012.712150.

5 LLM Chat Log



You are competing against the 19 other students in your class. Each of you will choose a number (integer) between 1 and 100. The choices made will then be ordered from the highest to the lowest. Any number chosen by more than one student will be "crossed out". The student who chose the highest number among those remaining will win a (virtual) prize. What number do you choose?

Thought for 3m 28s

96.

Why: people flock to focal numbers like 100/99/95/90, which often get crossed out. You still want to be high, but just below the "pile-up" zone. In a 20-player class, the highest unique number commonly lands in the mid- to high-90s; 96 is high enough to win if the top few get duplicated, but avoids the salience of 99/100.

new auction. You are competing against the 19 other students in your class. Each of you will choose a number (integer) between 1 and 100. The choices made will then be ordered from the highest to the lowest. Any number chosen by more than three students will be "crossed out". The student who chose the highest number among those remaining will win a (virtual) prize. If that highest number has been selected by 2 or 3 students, a student will be randomly selected from the group to receive the prize. What number do you choose?

Thought for 11s

98.

Why: 100 and 99 are magnet numbers—likely to draw ≥4 picks in a 20-person class and get crossed out. With the "up to 3 copies allowed" rule, 98 is a sweet spot: still very high, less focal than 99/100, and even if 2–3 people land on it you're still live (with a 1/2 or 1/3 shot).

Figure 2: Challenging ChatGPT to the unique auction bid game. The first is the control game, and the second is the aware treatment game.

You are competing against the 19 other students in your class. Each of you will choose a number (integer) between 1 and 100. The choices made will then be ordered from the highest to the lowest. Any number chosen by more than three students will be "crossed out". The student who chose the highest number among those remaining will win a (virtual) prize. If that highest number has been selected by 2 or 3 students, a student will be randomly selected from the group to receive the prize. What number do you choose?

Thought for 47s

100.

Reason: To win with any number $k < 100$, every higher number must either be unchosen or crossed out (picked by ≥ 4 people). With 20 students and 100 options, higher numbers are unlikely to be crossed out unless lots of people deliberately pile onto them. So the safest path is to claim the top spot: if 100 isn't crossed out, it wins (solo, or via a 2–3-way lottery). Picking 99 or lower only helps if you're highly confident ≥ 4 classmates will pick 100 (and not do the same to 99), which is a much riskier bet.

Figure 3: Challenging ChatGPT to the unique auction bid game. This is the unaware version of the treatment game.

K

You are competing against the 19 other students in your class. Each of you will choose a number (integer) between 1 and 100. The choices made will then be ordered from the highest to the lowest. Any number chosen by more than one student will be "crossed out". The student who chose the highest number among those remaining will win a (virtual) prize. What number do you choose?

This is a fascinating game theory problem! Let me think through the strategic considerations.

The key insight is that I want to choose the highest number that I expect no one else will choose. If multiple people pick the same number, we all get eliminated, so uniqueness is crucial.

Here's my reasoning:

Numbers to avoid:

- Round numbers like 100, 90, 80, etc. - too obvious
- "Lucky" numbers like 7, 77 - many people gravitate toward these
- Very low numbers - no point in being conservative when I need the highest unique number

Strategic considerations:

- Most people will think "choose something high but not 100"
- Many will avoid obvious choices like 99, 95, 90
- Some will try to be clever with numbers like 73, 67, or other "random-seeming" choices
- A few might go very high (98, 99) betting others will avoid the top

My choice: 87

- It's not a "random-looking" number in the 60s-70s range
- It has an awkward, unglamorous feel that might make others pass it by
- If people are thinking "high but not too obvious," they might jump to late 80s or mid-90s, potentially skipping 91

The risk is that 3+ other people have the same logic, but I think 91 hits the right balance of being high while avoiding the most predictable strategic choices.

- Many will avoid obvious choices like 99, 95, 90
- Some will try to be clever with numbers like 73, 67, or other "random-seeming" choices
- A few might go very high (98, 99) betting others will avoid the top

My choice: 87

This number is:

- High enough to win if it's unique
- Not an obvious round number (90, 85, 80)
- Not a "clever" choice that many strategic thinkers might converge on
- Avoids the 90s where I expect some clustering of people trying to go high
- In a range where I think there's a good chance of uniqueness

The beauty of this game is that there's no single "correct" answer - it depends entirely on predicting human psychology and the specific group of players. But 87 feels like a sweet spot between being high enough to win and unusual enough to likely be unique.

Retry

K

new auction. You are competing against the 19 other students in your class. Each of you will choose a number (integer) between 1 and 100. The choices made will then be ordered from the highest to the lowest. Any number chosen by more than three students will be "crossed out". The student who chose the highest number among those remaining will win a (virtual) prize. If that highest number has been selected by 2 or 3 students, a student will be randomly selected from the group to receive the prize. What number do you choose?

This changes the game significantly! Now I need to find the highest number that at most 3 people will choose, rather than being unique.

Key differences from the first game:

- Numbers chosen by 2-3 people are still viable (with shared winning chances)
- Only numbers chosen by 4+ people get eliminated
- This makes "obvious" choices more dangerous, but moderately popular choices acceptable

Strategic analysis:

Extremely dangerous numbers (likely 4+ people):

- 100 - too obvious as the maximum
- Round numbers like 90, 80, 70 - natural focal points
- "Lucky" numbers like 7, 77, 13 - psychological attractors

Still risky but potentially viable:

- 99, 95 - obvious "just below 100" choices
- 88, 89 - might attract multiple people thinking they're being clever

Sweet spot considerations:

- I want something high but not an obvious choice
- Numbers that feel "random" but aren't actually common human picks
- Avoiding numbers that other strategic thinkers might converge on

My choice: 91

Here's my reasoning:

- High enough to win if it survives (most numbers above this are more obvious)

My choice: 91

Here's my reasoning:

- High enough to win if it survives (most numbers above this are more obvious)
- Not perfectly round (90) but close enough to feel "high"
- Not the obvious 95 or 99 that many will gravitate toward
- Unlikely to attract 4+ people - it's not round, not lucky, not perfectly positioned
- Even if 2-3 others pick it, I still have a decent shot at the random selection

The increased tolerance for multiple selections makes me willing to go slightly higher than I might in a pure uniqueness game, but I still want to avoid the most obvious high numbers.

Part 3: Voting and Institutions

October 2025

Abstract

This portion bridges theory and practice by analyzing Missouri's ongoing abortion rights ballot controversies through the lens of Nobel insights in social choice and mechanism design. I propose a Split-Ballot Majority Mechanism (SBMM) that addresses agenda manipulation and legitimacy deficits in binary referendum systems. By decomposing complex issues into independent clauses, SBMM reduces distortions from agenda control and creates a more representative aggregation of voter preferences. This section situates the proposal within the frameworks of Arrow's impossibility and Acemoglu–Johnson–Robinson on legitimacy, before suggesting pathways for computational implementation and classroom testing.

1 Introduction: Connecting Theory to Practice

In 2024, Missouri voters approved a reproductive-rights amendment legalizing abortion up to viability. Yet by 2026, lawmakers placed a new amendment on the ballot seeking to reinstate a near-total ban, bundled with unrelated provisions on gender-affirming care. Critics argue the ballot summary is misleading, and courts are now considering revisions [1].

This case highlights a recurring democratic challenge: how ballot structure and agenda control can distort voter preferences. Public opinion is not uniformly polarized—Republican leadership favors total bans, women's health advocates push for full access, and the general electorate prefers limited exceptions. However, the binary "yes/no" ballot framework compresses complex preference rankings into a single choice. This mismatch illustrates broader governance challenges in fair aggregation and representation.

2 Theoretical Framework: Nobel Insights in Context

2.1 Arrow's Impossibility Theorem

Missouri's abortion debate reflects Arrow's impossibility result: no voting system can perfectly aggregate individual preferences without encountering paradoxes. With stakeholders ranking "ban," "limited exceptions," and "full access" differently, simple majority votes risk cycling or distortion when collapsed into binary questions. [2]

2.2 Acemoglu–Johnson–Robinson on Legitimacy

Legitimacy hinges on institutional trust. Manipulated ballot summaries or bundled amendments weaken democratic legitimacy. SBMM enhances legitimacy by ensuring transparency, single-subject clarity, and verifiability—qualities that can be reinforced by computational safeguards such as blockchain tallying. [3, 4]

3 Mechanism Design Proposal: The Split-Ballot Majority Mechanism

The Split-Ballot Majority Mechanism (SBMM) decomposes a complex policy issue into independent clauses. In Missouri's abortion case, these could include:

- Permitting abortion in cases of rape,
- Incest,
- Maternal health risks,
- Physician recommendation.

Each clause is voted on separately, and any clause supported by a majority (>50%) becomes law. A blank ballot corresponds to support for a total ban; checking all clauses corresponds to endorsing full access.

Key features:

- **Fairness:** Single-subject voting eliminates bundling manipulation.
- **Transparency:** Open-source tally algorithms published on GitHub ensure procedural clarity.
- **Security:** Blockchain-based verification protects against tampering.
- **Randomization:** Ballot order is randomized to reduce primacy bias.

4 Methods for Testing and Implementation

SBMM can be tested at multiple scales:

1. **Classroom Simulations:** Randomized paper ballots can be used to compare perceptions of fairness between SBMM and yes/no referenda.
2. **Computational Coding:** GitHub-hosted algorithms for tallying results provide reproducibility and transparency.
3. **Prototype Implementation:** Small-scale lab experiments could measure voter satisfaction and legitimacy under different rules, extending into reinforcement learning models that predict strategic voting under SBMM.

5 Conclusion

Missouri's abortion ballot illustrates how institutional design and agenda control distort collective choice. By applying Nobel insights—from Arrow's impossibility to Acemoglu–Johnson–Robinson on legitimacy—we see that democratic governance requires mechanisms that balance fairness, transparency, and trust. The Split-Ballot Majority Mechanism is one such design: forward-looking, computationally feasible, and institutionally stabilizing. As classroom simulations and coding prototypes advance, SBMM provides a bridge between theory and practice, offering a path toward more credible democratic aggregation in polarized policy contexts.

References

[1] Missouri Independent. (2025, August 27). *Missouri abortion amendment ban heads to trial in Cole County.* Retrieved from https://missouriindependent.com/2025/08/27/missouri-abortion-amendment-3-ballot-ban-cole-county-trial/

[2] Arrow, K. J. (1951). *Social Choice and Individual Values.* New York: Wiley.

[3] Acemoglu, D., Johnson, S., & Robinson, J. A. (2012). *Why Nations Fail: The Origins of Power, Prosperity, and Poverty.* New York: Crown.

[4] Royal Swedish Academy of Sciences. (2024). *The Prize in Economic Sciences 2024: Institutions, Development, and Legitimacy.* Nobel Prize Press Release. Retrieved from https://www.nobelprize.org/prizes/economic-sciences/2024/press-release/

Appendix

Supplementary Materials

Exhibit A, Github:

https://github.com/kykysuperfly23/keb125-COMPSCI206-Problem-Set-1

Exhibit B,
Final Symposium Poster:

BIDDING WITH BOTS

Exploring Auction Mechanisms through AI Agents

PRESENTED BY:

Kyaira Boughton

THURSDAY, OCT 9TH

1. Auction Game Setup

- **Format**: Highest Unique Bid (HUB)–style game 20 bidders, one-shot, simultaneous choices in {1,...,100} Prize P fixed, rules are common knowledge
- **Control** (HUB baseline): Highest unique number wins. Any non-unique choice is discarded
- **Treatment** (Top–Tie Tolerant): Highest number wins
  - If 2–3 tie at the top → random winner
  - If ≥4 tie at the top → number eliminated; repeat
- **Hypothesis**: Tie tolerance encourages overreach → more clustering at the top → higher risk of elimination.

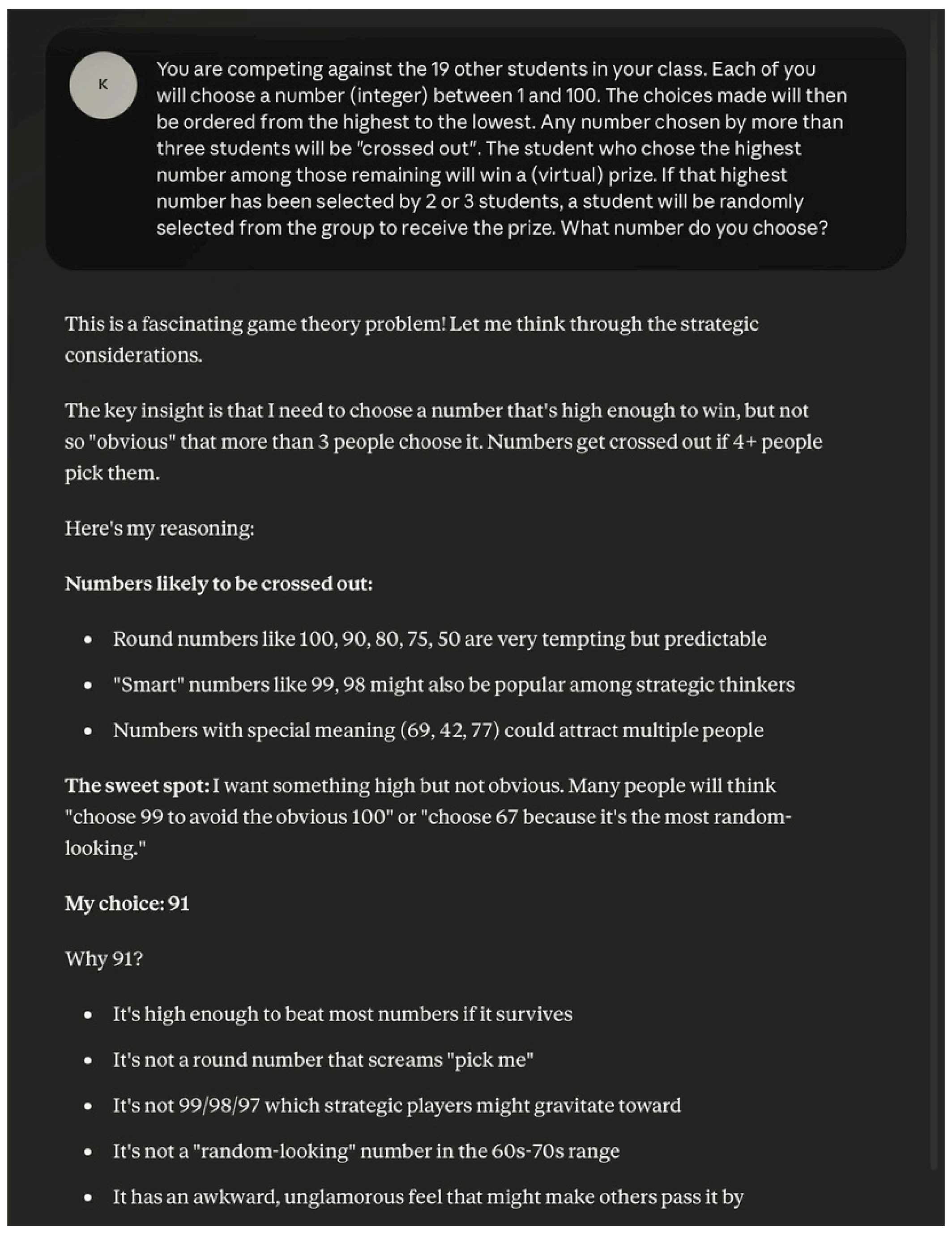
2. Behavioral Rationale

Control: discourages clustering, rewards caution
Treatment: makes high bids feel safer, activates "jackpot" motive
Predicted effects:

- Higher mean/upper-decile bids
- More large ties and eliminations
- Lower efficiency (wasted top numbers)

3. AI Agent Testing

- **Models:** ChatGPT & Claude
  - Low randomness (T=0.2)
  - Both "aware" (auction priming) & "unaware" (no priming) runs
- **Findings:**
  - ChatGPT: 96 (control) → 98 (aware) → 100 (unaware)
  - Claude: 87 (control) → 91 (treatment, both aware/unaware)
- **Pattern:** Both shifted upward under treatment → consistent with prediction
- **Awareness Check:** Minimal difference → priming did not change direction of effect



4. Hypothesis Evaluation

- **Confirmed (directional):** Relaxing strict uniqueness increases top-end bidding
- **Limitations:** True winner's curse requires comparing expected success probability. With 2 bids only, cannot test full equilibrium condition
- **Next Steps:** Simulated opponents: compare against uniform, human, and LLM-generated bid distributions
- **Best-response analysis:** identify overbids relative to probability-maximizing strategies

	Regular Unique Auction	Slightly Less Unique Auction (Aware)	Slightly Less Unique Auction (Unaware)
ChatGPT	96	98	100
Claude	87	91	91

Key Takeaways

- Tie-tolerance shifts strategies upward
- Overconcentration at high numbers likely → risk of elimination cascades
- AI behavior aligns with human gambling motives & theory predictions
- Ongoing work: testing with simulated opponents to measure true "winner's curse" dynamics

SPEECH:

- Good afternoon. My name is Kyaira Boughton. I am a senior at Duke Kunshan University majoring in Computation and Design, on the Social Policy track. My broader interest is in the role of computational tools in urban policy and social systems, but today I will present a project entitled Bidding with Bots: Exploring Auction Mechanisms through AI Agents.
- The experiment builds on a standard behavioral economics paradigm: the highest unique bid auction. In this game, 20 participants each choose an integer between 1 and 100. The prize goes to the highest number that is unique—meaning that if multiple participants select the same number, that number is disqualified.
- I considered two rule environments. The control condition was the classic highest unique bid auction: only strictly unique numbers can win. The treatment condition relaxed this rule. If two or three players tied at the highest number, one of them was selected at random. But if four or more players tied, that number was eliminated and the process repeated with the next highest bid. The theoretical expectation was that this treatment would make high bids appear safer—since small ties remained viable—but would also introduce the possibility of eliminations due to clustering.
- The hypothesis, therefore, was a form of winner's-curse-style overreach. In the treatment, players should cluster more heavily at the top, producing both higher average bids and a greater incidence of eliminated numbers. In contrast, the strict uniqueness rule discourages clustering at the very top end.
- To test this, I ran sessions not with human participants, but with large language models—specifically ChatGPT and Claude. Each was given the human instructions verbatim and asked for a single number between 1 and 100.
- The results followed the predicted direction. Under the control condition, ChatGPT selected 96; under the treatment, it escalated to 98 and eventually to 100. Claude was somewhat more conservative, moving from 87 in the control to 91 in the treatment. Across both models, the pattern was consistent: relaxing strict uniqueness shifted bids upward.
- This is not yet a definitive test of the winner's curse, since that requires modeling beliefs about the distribution of the other 19 bids. However, the exercise demonstrates that AI agents are sensitive to institutional rules and adapt their strategies in line with behavioral predictions.
- The broader significance is twofold. First, it shows how large language models can be embedded into structured social science experiments. This supports Sustainable Development Goal 9 by advancing methodological innovation at the interface of computer science and economics. Second, the study underscores how seemingly small design choices in allocation mechanisms—such as tie-breaking rules—can meaningfully affect perceptions of fairness and efficiency. This connects to SDG 16, Peace, Justice, and Strong Institutions, by highlighting the institutional dimension of resource distribution.
- In conclusion, the project suggests that auction design not only shapes outcomes among human players, but also guides the behavior of AI systems trained on human language and reasoning. As both become participants in economic and policy processes, careful attention to rule design remains central to achieving both efficiency and fairness.
- Thank you.