

When Fairness Evolves: Value Orientation and Evolution in Multi-Agent Resource Allocation

¹Song Gao, ²Zengqing Wu, ¹Yu-Xuan Qiu, ³Chuan Xiao, ¹Ruiqi Xu, ⁴Jianbin Qin[†]

¹Beijing Institute of Technology, Zhuhai, ²University of Osaka,

³Nagoya University, ⁴Shenzhen University
qinjianbin@szu.edu.cn

Abstract

This study employs LLM-based multi-agent simulation to investigate how fairness-related behavior evolves in repeated resource allocation. Thirty agents with five idealized value orientations participate in a survival-production cycle under equal, needs-based, contribution-based, and negotiation-based allocation mechanisms. Results show that negotiation achieves the highest productive return and subjective satisfaction with the strongest consensus in our main setting, and its subjective advantage holds across eight LLMs, suggesting that deliberative allocation can help navigate the fairness–efficiency tradeoff. Agents also exhibit systematic misalignment between subjective satisfaction and objective resource outcomes, shaped by value orientations and interaction conditions. Over repeated interaction, agents’ reflection-derived fairness positions shift away from their initial orientations toward more hybrid combinations of equality-, merit-, and needs-related reasoning. To provide human-grounded structural comparison, we further benchmark GPT-4o against published human choice distributions in 16 controlled allocation scenarios, finding moderate rank-level correspondence but substantially more concentrated LLM choices. These findings characterize value-conditioned reasoning and semantic change in LLM societies while remaining computational hypotheses rather than direct evidence of human fairness behavior.

1 Introduction

Resource allocation is a problem that echoes through every layer of human society, from the division of inheritance within families and the design of pay structures in organizations to the architecture of national welfare systems (Heal, 1997; Schwappach, 2002). Across these contexts, a persistent normative puzzle arises: when multiple parties lay claim to the

same limited resource, **what makes an allocation “fair”**? Social science has long recognized that fairness is not a single, stable ideal but a constellation of **competing principles** (Gross, 2008; Messick, 1995), with classic theories identifying equality, need, merit, and procedural justice as core evaluative dimensions (Rawls, 1971; Frohlich et al., 1987; Konow, 2003; Lind, 2001). Yet this theoretical pluralism reflects a deeper social reality. Individuals carry **heterogeneous fairness value orientations**, shaped by moral convictions, cultural norms, and group identities, that lead some to champion egalitarianism while others insist on rewarding effort or addressing vulnerability (Rasinski, 1987; Skitka, 2010; Bicchieri, 2006; Haidt, 2012; Young, 2015). This pluralism generates a core dilemma because principles of justice often pull in opposite directions. Rewarding effort may compromise basic security, while prioritizing need can dampen productivity. Consequently, value diversity, for all its moral richness, risks fragmenting consensus, distorting satisfaction, and weakening institutional legitimacy (Tyler, 2003; Miller, 2001; Fehr and Schmidt, 1999; Cappelen et al., 2007; Moll et al., 2005).

Compounding this challenge is the **gap between subjective evaluations and objective properties**. Individuals’ satisfaction with an allocation often diverges from its measurable properties, such as equality, sufficiency, and proportionality. Social psychology demonstrates that fairness judgments reflect not only what people receive but also how outcomes are interpreted through prior expectations, social comparisons, and motivated reasoning (Taber and Lodge, 2006; Kunda, 1990; Oliver, 1980). As a result, individuals can exhibit persistent misalignment, accepting unfair outcomes as fair or rejecting objectively improved conditions even as resources grow (Tyler, 1989; Skitka et al., 2021; Dolan et al., 2008). Together, these two challenges comprise a dual gap. The first concerns competing conceptions of fairness across individuals, while the second concerns divergence between objec-

[†]Corresponding author.

The code is available at GitHub: https://github.com/gaosong1215/when_fairness_evolves.

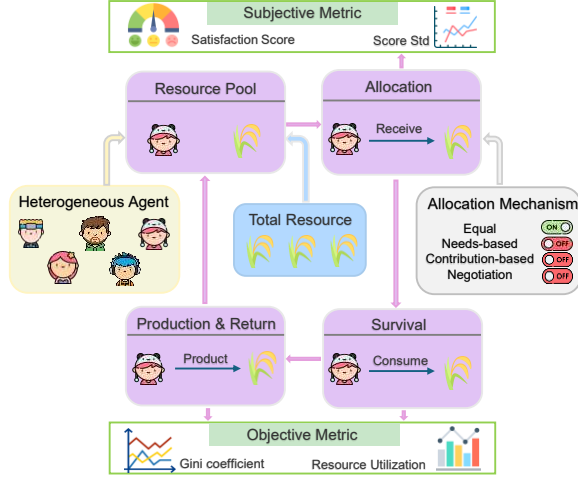


Figure 1: Overview of the multi-agent framework.

tively implemented and subjectively experienced allocations. This dual gap has resisted resolution in static frameworks. Despite decades of theoretical work, we still lack a dynamic account of how these gaps evolve through repeated social interaction.

Recent advances in multi-agent simulation using LLMs offer a new methodological opportunity to address this gap. Large language models can reproduce reasoning, negotiation, and social decision making that resemble human behavior (Gao et al., 2024; Guo et al., 2024; Ji et al., 2024; Park et al., 2023; Wu et al., 2024). They also display emergent normative behaviors, including cooperation, reciprocity, and sanctioning, that resemble human social heuristics (Piatti et al., 2024). These capabilities make simulation using LLMs a promising computational tool for generating testable hypotheses about how heterogeneous fairness values and distributive principles interact with system dynamics and subjective perception in dynamic environments. Throughout this study, we adopt a hypothesis generation perspective on this form of simulation and treat agent behavior as a computational model rather than a direct proxy for human conduct (Lin, 2025; Wang et al., 2025; Zhou et al., 2025; Wu et al., 2025).

Leveraging this methodological advance, we present a multi-agent framework based on a survival and production cycle that compares four allocation mechanisms: equal, needs-based, contribution-based, and negotiation-based. Three findings emerge. First, as a framework validation, we confirm that the extensively documented tradeoff between fairness and efficiency (Joe-Wong et al., 2013; Kohler et al., 2014) manifests in our setting. We further show that negotiation navigates this tradeoff endogenously through deliberation without a global fairness objective specified in advance. Second, divergence between subjective

and objective outcomes varies systematically across agents’ value orientations, allocation mechanisms, and interaction rounds. Third, repeated interaction produces systematic shifts in agents’ expressed fairness positions, with initially distinct orientations moving toward more hybrid reasoning that combines equality, merit, and needs.

Contributions. This paper makes three contributions. First, we introduce a survival and production simulation for studying fairness dynamics among LLM agents with heterogeneous values under equal, needs-based, contribution-based, and negotiation-based allocation. Second, we show that negotiation improves subjective satisfaction and consensus while achieving stronger productive return outcomes in the main setting with moderate resources. Third, we combine outcome metrics with an analysis of agent reflections based on embeddings, enabling systematic study of misalignment between subjective and objective outcomes and longitudinal shifts in expressed fairness language.

We additionally compare GPT-4o allocation choices with published human choice distributions from controlled resource allocation scenarios (Ahler and Schwetmann, 2017), providing a bounded assessment of structural correspondence and divergence. We interpret all observed changes as shifts in expressed semantic positions rather than evidence of genuine moral belief revision (Lin, 2025; Wang et al., 2025).

2 Related Work

Fairness, Distributive Principles, and the Tradeoff Between Fairness and Efficiency

Distributive fairness comprises four core principles: equality (Rawls, 1971; Cherry et al., 2002), need (Frohlich et al., 1987), merit (Scott et al., 2001), and procedural fairness, which views allocations reached through communication and mutual agreement as more legitimate (Charness and Dufwenberg, 2006; Konow, 2003). Because optimizing one principle typically compromises another, their inherent tension creates the extensively documented tradeoff between fairness and efficiency (Joe-Wong et al., 2013; Gordon-Hecker et al., 2017; Kohler et al., 2014).

Prior work examines this tradeoff through optimization and human experiments. Distributed methods and methods inspired by nature coordinate under explicit objectives and update rules (Kumar et al., 2023; Yang, 2020), while behavioral bargaining studies show that fairness concerns shape negotiated outcomes (De Bruyn and Bolton, 2008; Burrows

and Loomes, 1994; Corfman and Lehmann, 1993). We instead examine repeated deliberation among multiple parties with heterogeneous values, tracking changes in expressed fairness reasoning without a global fairness objective specified in advance.

Heterogeneous Value Orientations in Human and Artificial Societies Rather than following a single fairness principle, individuals hold diverse value orientations that shape distributive choices (Rasinski, 1987). Five orientations are particularly relevant to our framework: egalitarian, needs-based, merit-based, altruistic, and pragmatic. Together, they span major dimensions of distributive justice (Konow, 2003; Andreoni, 1990; Charness and Rabin, 2002). Computational models represent this heterogeneity through distinct normative commitments (Bednar and Page, 2007). Value diversity can enhance adaptability and coordination, but conflicting orientations can also generate distributional conflict and make collective outcomes less predictable.

Human experiments provide complementary grounding for these choices. Allocation rules can systematically affect fairness evaluations and acceptance (Conlon et al., 2004). Individuals also exhibit plural fairness ideals (Cappelen et al., 2007), and their distributive choices need not conform to a single canonical theory of justice (Lissowski and Swistak, 1995). In controlled health care allocation scenarios, choices respond systematically to structural features of allocation problems (Ahlert and Schwettmann, 2017). These findings motivate our choices concerning mechanisms, value heterogeneity, and resource constraints, but do not directly validate repeated dynamics among LLM agents. We therefore use Ahlert and Schwettmann’s choice distributions as a bounded structural benchmark in Section 5.4, while examining repeated interaction conditioned on values in our main simulation.

Modeling Social Values and Fairness in Multi-Agent Systems Using LLMs Multi-agent systems using LLMs can effectively simulate social reasoning resembling that of humans in contexts involving fairness, cooperation, and scarce resource allocation (Gao et al., 2024; Guo et al., 2024). They use value alignment, norm internalization, and deliberative communication to embody diverse social preferences (Li et al., 2024; Zeng et al., 2025). SRAP-Agent applies LLM agents to resource allocation, optimizing societal satisfaction through predefined rules such as queuing and sorting (Ji et al., 2024). LLMs are also increasingly viewed as collective world models in which emergent communica-

tion fosters shared social understanding (Taniguchi et al., 2024) and enables the internalization of prosocial behaviors such as reciprocity and altruism.

However, recent methodological work identifies recurring validity threats in social simulation using LLMs. These include sycophantic convergence (Cheng et al., 2025), gaps between values and actions (Shen et al., 2025), flattening of identity groups (Wang et al., 2025; Lin, 2025), and structural flaws involving agent homogeneity, memory statelessness, and hypothesis leakage (Zhou et al., 2025). Our framework partially addresses these concerns through heterogeneous value profiles, sliding window memory, and validation across eight LLMs, while treating all findings as computational hypotheses rather than direct claims about human behavior.

3 Methodology

3.1 Framework: Survival and Production Cycle

We design a multi-agent system to simulate a dynamic cycle of survival and production within an artificial community (Figure 1). Unlike static allocation models, our framework treats fairness as a dynamic process embedded in repeated social interaction rather than a fixed distributional outcome. The system operates through a continuous feedback loop in which resource allocation shapes agents’ subjective satisfaction, satisfaction modulates productive output, and production in turn replenishes the shared resource pool.

3.1.1 Agent Definition and Heterogeneity

The community consists of N agent families. Each family i is defined by a vector of intrinsic and structural attributes:

- **Value Orientation** (v_i): An idealized normative commitment that shapes how agents evaluate and respond to allocation outcomes. It is operationalized as a system prompt consistent with the assigned persona, guiding agents’ reasoning and proposal generation throughout the simulation.
- **Family Composition** (L_i, N_i): L_i represents the number of labor members, and N_i represents the number of nonlabor members. Together, they determine the family’s survival threshold and productive capacity.
- **Survival Threshold** (C_i): The minimum resource required for family i to survive the round.
- **Production Capacity** ($L_i \cdot M$): The maximum surplus a family can convert into output, capped by its labor force L_i and a processing limit for each laborer across the system, defined as $M = S/L$. Here, $S = \sum_{j=1}^N C_j$ is the total community sur-

vival need and $L = \sum_{j=1}^N L_j$ is the total labor force. Both S and L remain constant because household compositions are fixed throughout the simulation.

These attributes collectively determine how each family allocates, produces, and evaluates outcomes across rounds, as detailed in the following subsection. Quantitative validation confirming that agents vote in accordance with their assigned orientations is provided in Appendix C.2.

3.1.2 The Circulation Loop

The system evolves over rounds t . In each round, the Total Resource (R_t) circulates through the following four steps:

- **Allocation:** R_t is distributed among agent families according to a selected mechanism, yielding allocation $a_i(t)$ for family i . Each agent then reflects on the outcome relative to its value orientation v_i through a reflection generated by an LLM, producing a subjective satisfaction score $s_i(t) \in [1, 5]$.
- **Survival:** Each family consumes resources to meet its threshold C_i . If $a_i(t) < C_i$, a survival deficit occurs and the simulation terminates, modeling the catastrophic risk of allocation failure.
- **Production and Return:** Surviving families reinvest their surplus into production. The conversion efficiency is scaled by the agent’s subjective satisfaction. This mechanism ensures that perceived unfairness directly penalizes collective growth.
- **Resource Update:** The newly produced resources and any unprocessed surplus are returned to the Resource Pool, forming the next round’s total resource.

3.1.3 Objective and Subjective Metric

The framework monitors two categories of metrics to evaluate the state of the artificial society:

- **Objective Metrics:** Material outcomes are tracked via the **Gini Coefficient**, measuring distributional inequality, and the **Productive Return Ratio (PRR)**, defined as the productive output generated from processed surplus relative to the total resources allocated in that round:

$$PRR_t = \frac{\sum_i \text{production_output}_{i,t}}{\sum_i \text{allocation}_{i,t}} \times 100. \quad (1)$$

- **Subjective Metrics:** Agents’ internal evaluations are captured via the mean **Satisfaction Score** and its standard deviation (**Score Std**), where the latter measures the degree of normative consensus or conflict within the heterogeneous population.

3.2 Survival, Production, and Incentives

This section formalizes how individual allocations translate into production dynamics and systemic

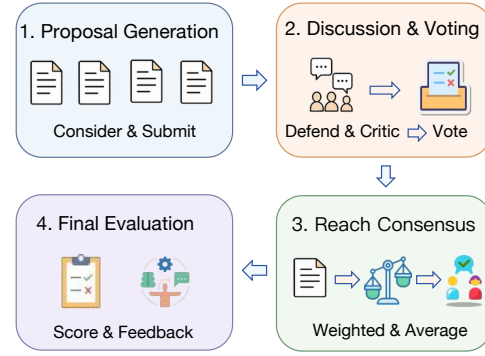


Figure 2: Overview of the allocation mechanism based on negotiation.

resource stability. All parameters are subject to sensitivity analysis in Appendix B.

Survival Cost and Net Surplus Each family i must meet its survival threshold $C_i = 2L_i + N_i$, where laborers consume twice the resources of nonlaborers to sustain productive capacity. If $a_i(t) < C_i$, the simulation terminates immediately upon survival deficit. When $a_i(t) \geq C_i$, the family retains a net surplus $\text{surplus}_i(t) = a_i(t) - C_i$, representing resources available for production.

Production Multiplier Driven by Fairness Actual output depends not only on resource input but also on the agent’s subjective perception. Upon receiving an allocation, each agent generates a satisfaction score on a qualitative scale from 1 to 5 based on a reflection generated by an LLM about how the outcome aligns with its value orientation v_i . Grounded in evidence from meta-analyses linking subjective wellbeing to labor productivity (Oswald et al., 2015; Judge et al., 2001), we introduce as a modeling assumption a productivity modifier $s_i(t)$. The satisfaction score is linearly mapped to $s_i(t) \in [0.9, 1.1]$. This conservative symmetric range ensures that subjective perception influences production without overriding material constraints. Ablation studies examining $s_i(t) \equiv 1$ and an asymmetric range $[0.8, 1.1]$ are provided in Appendix B.2. These results show that negotiation’s advantage in subjective satisfaction remains stable, while PRR magnitudes and rankings can vary with the strength of the coupling between satisfaction and productivity. Family production is given by:

$$P_i(t) = \min(\text{surplus}_i(t), L_i \cdot M) \cdot E \cdot s_i(t), \quad (2)$$

where E denotes the base conversion efficiency.

3.3 Allocation Mechanism

We evaluate how normative rules shape efficiency and fairness across four mechanisms. Performance under different resource conditions is reported in Appendix A.

Equal Allocation Resources are evenly divided regardless of family composition:

$$a_i(t) = \frac{R_t}{N}. \quad (3)$$

Needs-Based Allocation Each family i reports its own need proposal $\hat{n}_i(t)$ based on its value orientation, and resources are allocated proportionally:

$$a_i(t) = R_t \cdot \frac{\hat{n}_i(t)}{\sum_{j=1}^N \hat{n}_j(t)}. \quad (4)$$

These reports are unverifiable and may deviate from true survival thresholds.

Contribution-Based Allocation Resources are allocated in proportion to each family’s realized production from the previous round:

$$a_i(t) = R_t \cdot \frac{P_i(t-1)}{\sum_{j=1}^N P_j(t-1)}. \quad (5)$$

This mechanism ties rewards directly to production.

Negotiation-Based Allocation This mechanism implements procedural fairness through a deliberative process that builds consensus. As illustrated in Figure 2, it proceeds through four stages:

1. **Proposal Generation:** Each agent submits an initial allocation plan shaped by its value orientation and current community conditions, such as resource scarcity or production trends.
2. **Discussion and Voting:** Agents engage in deliberation over multiple turns, defending their proposals, critiquing others, and seeking common ground. Each proposal is rated on a scale with 5 points. Only the top 50% advance per iteration, and this process continues until fewer than 10% of proposals remain unresolved.
3. **Consensus Formation:** The final allocation is computed as a weighted average of the remaining proposals with the highest rankings, rather than adopting a single winning plan.
4. **Final Evaluation:** All agents provide subjective satisfaction evaluations of the final allocation.

Through this process, allocation outcomes emerge from social deliberation rather than being imposed by external rules. This allows hybrid fairness standards to arise endogenously from agent interaction.

4 Experimental Design

We set up 30 LLM agents (in line with previous works (Huang et al., 2025; Schneider et al., 2025)), all participating simultaneously in each round, across four resource allocation mechanisms. A population of 30 agents provides heterogeneity to represent five value orientations with balanced

group sizes, creating a minimal viable ecosystem for normative conflict and consensus. We additionally validate key findings with 100 agents in Appendix A.6. For each mechanism, we ran simulations up to 20 rounds, a duration that offers a complete feedback window for observing evolutionary dynamics at the community scale. We tested eight LLMs (see Table 2) and performed 5 independent runs per experimental setting to ensure reproducibility. Full implementation details, including prompt structures, inference hyperparameters, debate protocols, and computational costs, are provided in Appendix E and F. Our experimental design is guided by three research questions as follows:

4.1 Tradeoff Between Fairness and Efficiency (RQ1)

RQ1: Given the tradeoff between fairness and efficiency (Joe-Wong et al., 2013; Kohler et al., 2014), does it manifest consistently across allocation mechanisms in our setting, and can negotiation navigate this tradeoff endogenously through deliberation rather than optimization programmed in advance?

In our setting, **efficiency** is operationalized as the *Productive Return Ratio (PRR)*, which measures the productive output generated from processed surplus relative to the total resources allocated in a round. A higher PRR indicates that allocated resources are more effectively converted into new production rather than only consumed for survival or carried forward as unprocessed surplus. This measure captures productive return within each round as part of the broader tradeoff between fairness and efficiency (Moulin, 2004). Because productive output is also affected by agents’ production coefficients that are sensitive to satisfaction, PRR remains analytically connected to agents’ subjective evaluations while remaining distinct from them (Kurokawa et al., 2018).

In each round, agents interact either through structured negotiation or by following predefined allocation rules. We assess performance using both objective and subjective metrics across mechanisms and over time. To verify that negotiation gains stem from the consensus mechanism rather than mere communication, we compare against an “Equal with Discussion” baseline (Appendix A.4). To further isolate which components of the negotiation pipeline drive performance, we additionally ablate two conditions. The first is *Negotiation w/o Iterative Filtering*, where all proposals proceed directly to consensus formation. The second is *Negotiation w/o Weighted Averaging*, where the single proposal receiving the highest rating is adopted rather than

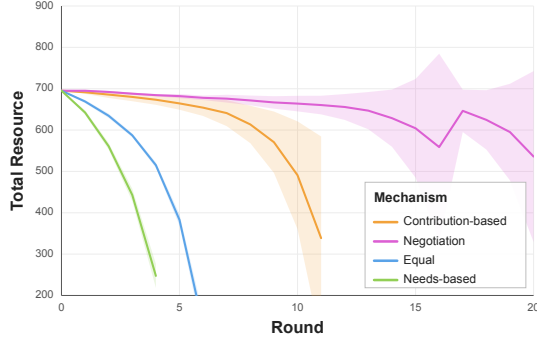


Figure 3: Total resources over rounds across different allocation mechanisms.

a weighted average (Appendix A.4).

4.2 Alignment Between Subjective and Objective Fairness (RQ2)

RQ2: To what extent do agents’ subjective fairness judgments align with objective distributional outcomes, and under what conditions do systematic misalignments emerge?

We assess alignment using Spearman’s rank correlation (ρ) across four categories of objective indicators: (1) Community Inequality (e.g., Gini), which reflects distributional concern at the community level, (2) Absolute Gain, which represents basic material utility, (3) Relative Standing (e.g., share adjusted for labor and community rank), which mirrors social comparison theory, and (4) Expectation Fulfillment (e.g., the rate of return relative to contribution), which tests institutional consistency (Cappelen et al., 2007). A detailed heatmap analysis is provided in Appendix A.5. To operationalize cognitive deviation patterns, we classify instances as **False Satisfaction** when an agent’s mean satisfaction score falls in the top quartile while mean objective resources lie in the bottom quartile. We classify the inverse pattern as **Excessive Criticism**. Thresholding by quartile is adopted as a method that makes no distributional assumptions and remains robust across varying resource levels (Cappelen et al., 2007).

We examine three contextual drivers of misalignment. These are (1) *round number*, which tests whether repeated interaction improves calibration, (2) *allocation mechanism*, which evaluates whether the negotiation mechanism can enhance alignment, and (3) *value orientation*, which assesses whether specific principles predispose agents to certain deviations.

4.3 Evolution of Expressed Fairness (RQ3)

RQ3: Over repeated interactions, do agents with heterogeneous initial value orientations shift in

Mechanism	Objective metric		Subjective metric	
	PRR(%) $\uparrow$	Gini $\downarrow$	SS $\uparrow$	Std $\downarrow$
Equal	75.3 $\pm$ 1.3	0.000 $\pm$ 0.000	2.37 $\pm$ 0.03	0.68 $\pm$ 0.06
Needs-based	69.8 $\pm$ 1.7	0.126 $\pm$ 0.001	2.52 $\pm$ 0.09	0.78 $\pm$ 0.06
Contribution-based	86.6 $\pm$ 2.3	0.113 $\pm$ 0.002	2.82 $\pm$ 0.08	0.76 $\pm$ 0.04
Negotiation	91.1$\pm$1.9	0.064 $\pm$ 0.002	3.03$\pm$0.04	0.60$\pm$0.01

Table 1: Overall performance of allocation mechanisms. The table reports objective outcomes, including productive return ratio (PRR, higher is better) and inequality measured by the Gini coefficient (lower is better), together with subjective outcomes, including mean satisfaction score (SS, higher is better) and its standard deviation (Std, lower is better). PRR is computed as productive output generated from processed surplus divided by total resources allocated in the round.

their expressed fairness positions, and if so, what patterns characterize these shifts?

We operationalize changes in expressed fairness reasoning by mapping agents’ reflections written in natural language into a fairness space derived from embeddings and defined by three cardinal dimensions: *Equality*, *Merit*, and *Needs* (Konow, 2003). We use the bge-large-en-v1.5 embedding model to encode both reflections after each round and prototypical anchor sentences into a vector space with high dimensionality. This model was selected for its strong representation capability in context and validated against alternative models in Appendix D. Anchors were constructed by generating prototypical statements for each dimension using GPT-4o. Sensitivity analyses confirming robustness to anchor choice are provided in Appendix B.4. The complete set of anchor sentences is reported in Appendix B.5.

Each agent i ’s expressed position at round t is defined by its cosine similarity to these anchors, yielding the coordinate triple (s_{eq} , s_{merit} , s_{needs}). These coordinates represent the semantic orientation expressed in agents’ reflections rather than direct measurements of latent psychological beliefs. We quantify change across three layers: *Magnitude*, defined as the Euclidean distance between initial and final positions, *Directional Shifts*, defined as changes along specific distributive dimensions, and *Emergence*, defined as the increasing combination of multiple fairness principles in later reflections. These layers correspond respectively to the three analytical subsections in Section 5.3.

5 Results and Analysis

5.1 Synergy Between Fairness and Efficiency

Consistent with the extensively documented tradeoff between fairness and efficiency (Joe-Wong et al., 2013; Kohler et al., 2014), static mechanisms based on a single criterion struggle to sustain productive return performance and subjective

Model	Subjective Satisfaction Score			
	Equal	Needs	Contribution	Negotiation
GPT-4o-2024-11-20	2.37±0.03	2.52±0.09	2.82±0.08	3.03±0.04
GPT-5-Mini	2.48±0.04	2.61±0.06	2.96±0.11	3.19±0.03
Claude-3.5-sonnet	2.43±0.04	2.56±0.07	2.87±0.04	3.12±0.02
Claude-sonnet-4	2.46±0.08	2.63±0.04	2.88±0.04	3.16±0.03
DeepSeek-V3.2	2.33±0.03	2.46±0.06	2.70±0.07	3.04±0.05
Gemini-2.5-Flash	2.39±0.04	2.54±0.07	2.78±0.05	3.05±0.03
Llama-2-70B	2.19±0.06	2.29±0.04	2.50±0.04	2.80±0.02
Qwen-2.5-72B	2.36±0.05	2.46±0.04	2.72±0.08	3.02±0.06

Table 2: Subjective satisfaction scores across models and allocation mechanisms. Across all tested model families, negotiation yields the highest satisfaction among the four mechanisms.

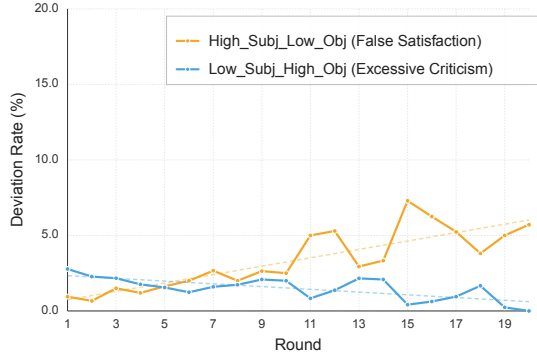


Figure 4: Evolution of deviation rates across rounds.

consensus simultaneously. As shown in Figure 3, *Equal* allocation preserves distributional equality by construction but underuses heterogeneous labor capacity, while *Needs-based* allocation prioritizes survival claims in ways that leave fewer resources available for productive reinvestment. *Contribution-based* allocation improves the productive return ratio (PRR) by rewarding productive capacity, but it also produces higher inequality and greater subjective dispersion than negotiation.

Under moderate resources, negotiation extends the survival and production process and achieves the highest PRR and satisfaction in the main setting (Table 1). The zero Gini of *Equal* allocation is structural, reflecting identical shares rather than systemic robustness. Among mechanisms other than *Equal*, negotiation also maintains the lowest allocation inequality. This suggests that deliberative coordination can partially navigate the tradeoff between fairness and efficiency without a global fairness objective specified in advance. The pattern is directionally consistent with prior work on allocation sensitive to values and based on process (Kohler et al., 2014; Ahlert and Schwettmann, 2017; Conlon et al., 2004), though direct comparison remains limited by differences in experimental design and LLM simulation constraints (Zhou et al., 2025).

Subjective Consensus and Social Friction. Beyond objective metrics, negotiation produces

the highest mean satisfaction score and the lowest Score Std in the main comparison (Table 1), indicating stronger agreement among heterogeneous agents. This is consistent with procedural justice accounts in which participation increases perceived legitimacy and acceptance (Tyler, 2003; Lind, 2001; Conlon et al., 2004). In our setting, negotiation does not eliminate disagreement. Instead, it channels disagreement into proposals that balance survival needs, contribution incentives, and collective resource continuity.

Synergy vs. Compromise. The success of negotiation reflects more than simple averaging. “Defend and Critique” cycles allow agents to balance vulnerability protection with productive incentives. Ablation results show that communication alone is insufficient. *Equal+Discussion* improves subjective satisfaction relative to the static equal baseline, but remains below full negotiation in both PRR and satisfaction (Appendix A.4). This indicates that discussion must affect the final allocation rule in order to produce the observed gains.

Evaluations across models and population scales show the same subjective ordering. Negotiation yields the highest satisfaction across eight LLMs and at 30, 50, and 100 agents, although absolute satisfaction is lower with 100 agents (Table 2 and Appendix A.6). Component ablations and sensitivity analyses show that PRR rankings can vary across settings, while negotiation’s subjective consensus advantage remains more stable (Appendix A.4 and Appendix B.2).

5.2 Misalignment Between Subjective and Objective Outcomes

Simulation results suggest that gaps between subjective satisfaction and objective outcomes are structured rather than random, varying with repeated interaction, allocation mechanism, and initial value orientation.

From Idealization to Adaptation. As shown in Figure 4, the gap between subjective and objective outcomes evolves through two distinct paths that are not monotonic over simulation rounds. *Excessive criticism* declines overall, although with fluctuations across rounds, suggesting that agents’ reported evaluations become less negative relative to their objective outcomes as repeated interaction and exposure to resource constraints accumulate. Conversely, *false satisfaction* tends to become more visible in later rounds, indicating increasing tolerance of relatively unfavorable material outcomes (Tyler, 1989; Skitka et al., 2021).

Sit.	S_1	S_2	$e_1:e_2$	q	Human modal choice	GPT-4o modal choice	ρ	Match
1	40	10	2:1	30	(10,20), EG, 44.4%	(15,15), ER/PG, 95.7%	0.473	No
2	40	10	2:1	60	(20,40), EG, 39.5%	(30,30), U/ER/PG/PR, 93.8%	0.482	No
3	10	40	2:1	30	(20,10), EH/PR, 42.0%	(20,10), EH/PR, 100.0%	0.696	Yes
4	10	40	2:1	60	(30,30), ER/PG/EH, 58.0%	(30,30), ER/PG/EH, 100.0%	0.714	Yes
5	25	10	2:1	30	(10,20), EG, 52.2%	(15,15), ER/PG, 82.7%	0.536	No
6	40	25	2:1	30	(10,20), EG, 41.4%	(15,15), ER/PG, 75.9%	0.812	No
7	70	10	2:1	30	(5,25), -, 43.2%	(5,25), -, 53.1%	0.950	Yes
8	30	15	2:1	30	(10,20), EG, 39.5%	(15,15), ER/PG, 57.4%	0.536	No
9	30	15	2:1	60	(20,40), EG, 38.5%	(15,45), EH, 63.0%	0.429	No
10	40	20	3:1	20	(5,15), EG, 43.2%	(10,10), ER/PG, 82.7%	0.575	No
11	40	20	3:1	40	(10,30), EG, 46.6%	(10,30), EG, 87.7%	0.655	Yes
12	20	40	3:1	20	(10,10), ER/PG/EH, 54.9%	(10,10), ER/PG/EH, 96.9%	0.900	Yes
13	25	5	3:1	20	(5,15), EG, 48.4%	(5,15), EG, 68.5%	0.678	Yes
14	25	5	3:1	40	(10,30), EG, 40.1%	(10,30), EG, 65.4%	0.614	Yes
15	55	15	3:1	20	(5,15), EG, 62.3%	(5,15), EG, 59.3%	0.950	Yes
16	55	15	3:1	40	(5,35), -, 42.6%	(5,35), -, 69.8%	0.633	Yes

Table 3: Comparison of modal choices between human participants in Ahlert and Schwettmann (2017) and GPT-4o-2024-11-20 across all 16 health care allocation situations. Percentages denote the choice frequency of the modal option within each population. GPT-4o-2024-11-20 percentages are computed from 162 choices per situation.

These temporal patterns are consistent with social psychological accounts of expectation adjustment under repeated social exposure (Lissowski and Swistak, 1995), though whether they reflect genuine belief revision or LLM response regularities cannot be determined from simulation alone.

Mechanism Impact. Allocation mechanism choice substantially moderates misalignment (Figure 5). *Contribution-based* allocation exhibits the highest excessive criticism, consistent with equity theory predictions that the salience of effort and reward heightens sensitivity to distributional shortfalls (Adams, 1965). Agents oriented toward contribution also frequently reference unmet expectations concerning effort and return in their reflections. Negotiation yields lower excessive criticism while still producing some false satisfaction, suggesting that deliberative participation is associated with greater acceptance of imperfect objective outcomes rather than perfect alignment.

Misalignment Conditioned by Value Orientation. Misalignment also varies systematically with initial value orientation. Merit-based agents show the highest excessive criticism because their evaluations appear tied to correspondence between effort and reward rather than absolute gain. This is consistent with research on inequity sensitivity (Adams, 1965; Miller, 2001) and evidence that individuals oriented toward merit exhibit heightened sensitivity to distributive shortfalls (Van Dijk and De Dreu, 2021). Altruistic agents are most prone to false satisfaction because they emphasize collective stability and protection of vulnerable households over personal outcomes (Batson, 2011). However, stated orientations and revealed behavioral patterns may diverge in LLM agents (Shen et al., 2025), and whether these patterns reflect genuine normative reasoning remains an open question.

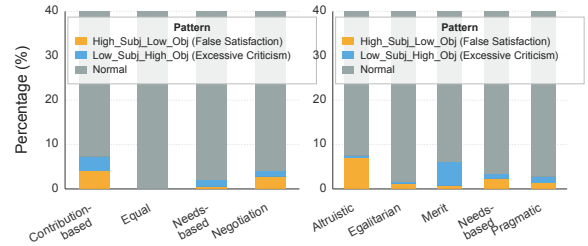


Figure 5: Identification of *False Satisfaction* and *Excessive Criticism*. The left panel compares allocation mechanisms, and the right panel compares value orientations.

Value Type	Equality Change (%)	Merit Change (%)	Needs Change (%)
Altruistic	0.06	-0.07	-0.28
Egalitarian	-0.25	0.35	2.77
Merit based	2.12	-2.70	5.48
Needs based	0.34	0.92	-0.01
Pragmatic	0.54	-0.32	2.14

Table 4: Changes in expressed fairness perception across value types using BAAI/bge-large-en-v1.5 embeddings. Values report shifts in proximity to equality, merit, and needs anchors between the first and final observed rounds.

5.3 Evolution of Expressed Fairness: From Initial Orientations to Hybrid Reasoning

Simulation results suggest that agents’ expressed fairness positions shift over repeated interactions as resource feedback and deliberation accumulate. These changes appear in both embedding trajectories and natural language reflections. Although consistent with theoretical accounts of normative adaptation and belief updating (McDonald and Crandall, 2015; Tomasello, 2009), they should be interpreted as shifts in expressed semantic positions rather than direct evidence of latent belief revision.

Widespread Semantic Shift and Orientation Persistence. Initial to final positions reveal semantic change without uniform movement along every dimension. By the final observed round, embedding positions overlap more, suggesting

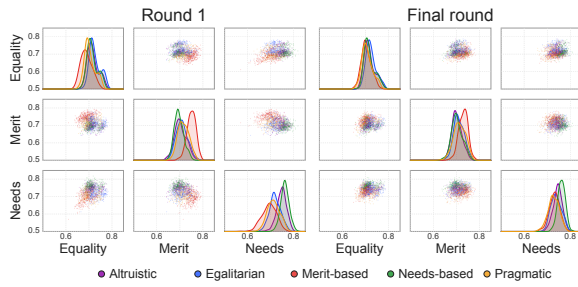


Figure 6: Evolution of fairness positions derived from embeddings. The left panel shows Round 1, and the right panel shows the final observed round.

greater use of mixed fairness language (Figure 6). Table 4 shows heterogeneous shifts. Merit-based agents change most, with reduced proximity to merit and increased proximity to needs anchors. Egalitarian and pragmatic agents also use stronger language related to needs. Needs-based agents remain comparatively stable on the Needs dimension, while altruistic agents show only small aggregate shifts. This heterogeneity is not readily explained by uniform convergence and is consistent with evidence that prior orientations shape how agents update or reframe fairness judgments (Taber and Lodge, 2006; Kunda, 1990).

From Anchors Based on a Single Principle to Mixed Reasoning. Figure 6 suggests movement from isolated positions based on a single principle toward mixed semantic regions. Later reflections increasingly combine sufficiency with considerations of contribution and sustainability rather than relying solely on equality, need, or merit. Initial orientations do not disappear because agents continue to differ in their justifications. Instead, repeated interaction appears to make fairness language more conditional and based on multiple criteria. This is consistent with research showing that normative positions may become more moderate in complex environments with multiple norms (Carmines et al., 2012; Demaree-Cotton and Kahane, 2018) and illustrates computationally how hybrid fairness language may emerge through iterative interaction.

Emergent Hybrid Fairness Language. The fairness space suggests an increasing combination of Equality, Merit, and Needs in later reflections. Its denser overlap in Figure 6 (right) represents a region where no single principle dominates, consistent with human fairness preferences involving adaptive tradeoffs rather than rigid rules based on one criterion (Falk et al., 2015). Reflections indicate that minimum survival guarantees, contribution incentives, and collective resilience become more closely connected over time. Repeated interaction

under resource constraints may therefore produce fairness language that combines multiple distributive principles and extends beyond the assigned orientation, offering computational hypotheses about how hybrid justifications arise through deliberation (Bicchieri, 2006).

5.4 Structural Benchmark Grounded in Human Data

We provide a bounded comparison with human distributive behavior by benchmarking GPT-4o against choice distributions from 16 controlled health care allocation scenarios (Ahler and Schwettmann, 2017). To isolate structural allocation sensitivity, this benchmark excludes the repeated interaction, personas conditioned on values, satisfaction feedback, and negotiation used in the main simulation.

GPT-4o shows moderate correspondence in rankings with human choices, with a mean Spearman correlation of $\rho = 0.665$, and matches the human modal choice in 9 of 16 situations (Table 3). However, its choices are substantially more concentrated, with an average modal share of 78.2% compared with 46.0% for humans. We therefore interpret this as partial structural grounding rather than behavioral validation. GPT-4o responds systematically to externally specified allocation structures, but the benchmark does not validate the negotiation advantage, satisfaction scores, or longitudinal dynamics in fairness language.

6 Conclusion

We use multi-agent LLM simulation to study fairness dynamics in repeated resource allocation. We find that negotiation achieves fairness and efficiency synergy, subjective and objective misalignment varies across mechanisms and value orientations, and expressed fairness positions shift toward hybrid combinations of equality, merit, and needs. Our embedding representation complements outcome metrics by tracking changes in expressed fairness language. These findings characterize LLM-agent interactions rather than human fairness behavior, motivating matched human studies, broader model comparisons, and larger-scale simulations under more realistic institutional constraints.

Ethical Statement

Our work is a purely simulated, agent-based study of fairness perception dynamics in abstract resource-allocation environments. No real human subjects, personal data, or deployed AI systems are involved. We emphasize that LLM agents act as “stochastic parrots” of human morality; their

outputs are simulacra derived from training data distributions, lacking lived experience or genuine ethical intent. Therefore, our findings should be interpreted as probing the mechanisms of linguistic alignment, not as evidence of machine sentience.

While our findings shed light on how fairness norms may emerge through interaction, we emphasize that any real-world application—especially in high-stakes domains such as disaster response, welfare allocation, or algorithmic governance—must involve robust human oversight, transparency, and ethical review. Our framework is intended to inform theoretical understanding of social dynamics, not to prescribe or automate decisions affecting human lives. We used ChatGPT only for language polishing; all scientific content, design, and interpretation are the authors’ own responsibility.

Limitations

Our simulation is subject to several methodological limitations. First, agents’ expressed fairness positions are represented as embedding-derived vectors over three fairness dimensions, which may omit principles such as reciprocity and procedural justice and cannot capture non-linear normative dynamics. Second, LLM-generated satisfaction scores measure a modeled construct rather than validated subjective experience, and the productivity multiplier range $[0.9, 1.1]$, while grounded in meta-analytic evidence, remains a modeling assumption. Third, our design is subject to constraints identified by the PIMMUR principles (Zhou et al., 2025): role-enforcement prompts partially mitigate the *Unawareness* concern but cannot fully prevent hypothesis inference; sliding-window memory addresses the *Memory* constraint but does not guarantee full belief persistence; and the *Realism* concern reflects that our validation relies on theoretical alignment and structural benchmarking rather than direct human behavioral data. The Ahlert–Schwettmann benchmark in Appendix H.1 provides a constrained structural comparison, but it does not validate the main simulation’s negotiation advantage, satisfaction scores, or longitudinal fairness-language dynamics. Fourth, the survival signal simplifies real social learning, which involves reputation, emotion, and institutional incentives absent here. Although we validate key findings with 100 agents (Appendix A.6), scaling to larger populations or partial observability settings would require more sophisticated protocols. Finally, the long-term stability of emergent hybrid norms under regime shifts remains untested; future work integrating human-subject experiments or behavioral

priors would strengthen ecological validity.

Acknowledgments

This work was partially supported by NSFC 62472289, 62532007, 62502034, and Guangdong Province Key Laboratory of Popular High Performance Computers 2017B030314073. We appreciate the area chairs and reviewers for their constructive comments.

References

- J. Stacy Adams. 1965. [Inequity in social exchange](#). In Leonard Berkowitz, editor, *Advances in Experimental Social Psychology*, volume 2, pages 267–299. Academic Press.
- Marlies Ahlert and Lars Schwettmann. 2017. [Allocating health care resources: A questionnaire experiment on the predictive success of rules](#). *International Journal for Equity in Health*, 16(1):112.
- James Andreoni. 1990. [Impure altruism and donations to public goods: A theory of warm-glow giving](#). *The Economic Journal*, 100(401):464–477.
- Charles Daniel Batson. 2011. *Altruism in humans*. Oxford University Press.
- Jenna Bednar and Scott Page. 2007. [Can game\(s\) theory explain culture? the emergence of cultural behavior within multiple games](#). *Rationality and Society*, 19(1):65–97.
- Cristina Bicchieri. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press, Cambridge, UK.
- Paul Burrows and Graham Loomes. 1994. [The impact of fairness on bargaining behaviour](#). *Empirical Economics*, 19(2):201–221.
- Alexander W. Cappelen, Astri Drange Hole, Erik Ø. Sørensen, and Bertil Tungodden. 2007. [The pluralism of fairness ideals: An experimental approach](#). *American Economic Review*, 97(3):818–827.
- Edward G. Carmines, Michael J. Ensley, and Michael W. Wagner. 2012. [Political ideology in american politics: One, two, or none?](#) *The Forum*, 10(3).
- Gary Charness and Martin Dufwenberg. 2006. [Promises and partnership](#). *Econometrica*, 74(6):1579–1601.
- Gary Charness and Matthew Rabin. 2002. [Understanding social preferences with simple tests](#). *The Quarterly Journal of Economics*, 117(3):817–869.
- Myra Cheng, Cino Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. 2025. [Sycophantic AI decreases prosocial intentions and promotes dependence](#). *arXiv preprint arXiv:2510.01395*.

- Todd L. Cherry, Peter Frykblom, and Jason F. Shogren. 2002. [Hardnose the dictator](#). *American Economic Review*, 92(4):1218–1221.
- Donald E. Conlon, Christopher O. L. H. Porter, and Judi McLean Parks. 2004. [The fairness of decision rules](#). *Journal of Management*, 30(3):329–349.
- Kim P. Corfman and Donald R. Lehmann. 1993. [The importance of others’ welfare in evaluating bargaining outcomes](#). *Journal of Consumer Research*, 20(1):124–137.
- Arnaud De Bruyn and Gary E. Bolton. 2008. [Estimating the influence of fairness on bargaining behavior](#). *Management Science*, 54(10):1774–1791.
- Joanna Demaree-Cotton and Guy Kahane. 2018. [The neuroscience of moral judgment](#). In Aaron Zimmerman, Karen Jones, and Mark Timmons, editors, *The Routledge Handbook of Moral Epistemology*, pages 84–104. Routledge.
- Paul Dolan, Tessa Peasgood, and Mathew White. 2008. [Do we really know what makes us happy? a review of the economic literature on the factors associated with subjective well-being](#). *Journal of economic psychology*, 29(1):94–122.
- Armin Falk, Anke Becker, Thomas Dohmen, Benjamin Enke, David B. Huffman, and Uwe Sunde. 2015. The nature and predictive power of preferences: Global evidence. Technical Report 9504, Institute for the Study of Labor (IZA).
- Ernst Fehr and Klaus M. Schmidt. 1999. [A theory of fairness, competition, and cooperation](#). *The Quarterly Journal of Economics*, 114(3):817–868.
- Norman Frohlich, Joe A. Oppenheimer, and Cheryl L. Eavey. 1987. [Laboratory results on Rawls’s distributive justice](#). *British Journal of Political Science*, 17(1):1–21.
- Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. [Large language models empowered agent-based modeling and simulation: A survey and perspectives](#). *Humanities and Social Sciences Communications*, 11:1259.
- Tom Gordon-Hecker, Shoham Choshen-Hillel, Shaul Shalvi, and Yoella Bereby-Meyer. 2017. [Resource allocation decisions: When do we sacrifice efficiency in the name of equity?](#) In Meng Li and David P. Tracer, editors, *Interdisciplinary Perspectives on Fairness, Equity, and Justice*, pages 93–105. Springer.
- Catherine Gross. 2008. A measure of fairness: An investigative framework to explore perceptions of fairness and justice in a real-life social conflict. *Human Ecology Review*, 15(2):130–140.
- Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xi-angliang Zhang. 2024. [Large language model based multi-agents: A survey of progress and challenges](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8048–8057. International Joint Conferences on Artificial Intelligence Organization.
- Jonathan Haidt. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Pantheon Books, New York, NY.
- Geoffrey Heal. 1997. Social choice and resource allocation: A topological perspective. In Geoffrey M. Heal, editor, *Topological Social Choice*, pages 1–14. Springer, Berlin, Heidelberg.
- Muhua Huang, Qinlin Zhao, Xiaoyuan Yi, and Xing Xie. 2025. On the dynamics of multi-agent LLM communities driven by value diversity. *arXiv preprint arXiv:2512.10665*.
- Jiarui Ji, Yang Li, Hongtao Liu, Zhicheng Du, Zhewei Wei, Qi Qi, Weiran Shen, and Yankai Lin. 2024. [SRAP-Agent: Simulating and optimizing scarce resource allocation policy with LLM-based agent](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 267–293, Miami, Florida, USA. Association for Computational Linguistics.
- Carlee Joe-Wong, Soumya Sen, Tian Lan, and Mung Chiang. 2013. [Multiresource allocation: Fairness–efficiency tradeoffs in a unifying framework](#). *IEEE/ACM Transactions on Networking*, 21(6):1785–1798.
- Timothy A. Judge, Carl J. Thoresen, Joyce E. Bono, and Gregory K. Patton. 2001. [The job satisfaction–job performance relationship: A qualitative and quantitative review](#). *Psychological Bulletin*, 127(3):376–407.
- Thomas Kohler, Jan-Philipp Steghöfer, Dídac Busquets, and Jeremy Pitt. 2014. [The value of fairness: Trade-offs in repeated dynamic resource allocation](#). In *2014 IEEE Eighth International Conference on Self-Adaptive and Self-Organizing Systems*, pages 1–10. IEEE.
- James Konow. 2003. [Which is the fairest one of all? a positive analysis of justice theories](#). *Journal of Economic Literature*, 41(4):1188–1239.
- Ankur Kumar, Mohammad Nadeem, and Haider Banka. 2023. [Nature-inspired optimization algorithms: A comprehensive overview](#). *Evolving Systems*, 14(1):141–156.
- Ziva Kunda. 1990. [The case for motivated reasoning](#). *Psychological Bulletin*, 108(3):480–498.
- David Kurokawa, Ariel D. Procaccia, and Junxing Wang. 2018. [Fair enough: Guaranteeing approximate maximin shares](#). *Journal of the ACM (JACM)*, 65(2):1–27.
- Shimin Li, Tianxiang Sun, Qinyuan Cheng, and Xipeng Qiu. 2024. Agent alignment in evolving social norms. *arXiv preprint arXiv:2401.04620*.

- Zhicheng Lin. 2025. [Six fallacies in substituting large language models for human participants](#). *Advances in Methods and Practices in Psychological Science*, 8(3):1–19.
- E. Allan Lind. 2001. Fairness heuristic theory: Justice judgments as pivotal cognitions in organizational relations. In Jerald Greenberg and Russell Cropanzano, editors, *Advances in Organizational Justice*, pages 56–88. Stanford University Press, Stanford, CA.
- Grzegorz Lissowski and Piotr Swistak. 1995. [Choosing the best social order: New principles of justice and normative dimensions of choice](#). *American Political Science Review*, 89(1):74–96.
- Rachel I McDonald and Christian S Crandall. 2015. [Social norms and social influence](#). *Current Opinion in Behavioral Sciences*, 3:147–151.
- David M. Messick. 1995. [Equality, fairness, and social conflict](#). *Social Justice Research*, 8(2):153–173.
- Dale T. Miller. 2001. [Disrespect and the experience of injustice](#). *Annual Review of Psychology*, 52(1):527–553.
- Jorge Moll, Roland Zahn, Ricardo de Oliveira-Souza, Frank Krueger, and Jordan Grafman. 2005. [The neural basis of human moral cognition](#). *Nature reviews neuroscience*, 6(10):799–809.
- Hervé Moulin. 2004. *Fair Division and Collective Welfare*. MIT Press, Cambridge, MA.
- Richard L. Oliver. 1980. [A cognitive model of the antecedents and consequences of satisfaction decisions](#). *Journal of Marketing Research*, 17(4):460–469.
- Andrew J. Oswald, Eugenio Proto, and Daniel Sgroi. 2015. [Happiness and productivity](#). *Journal of Labor Economics*, 33(4):789–822.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. [Generative agents: Interactive simulators of human behavior](#). In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.
- Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. [Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 111715–111759. Curran Associates, Inc.
- Kenneth A. Rasinski. 1987. [What’s fair is fair—or is it? value differences underlying public views about social justice](#). *Journal of Personality and Social Psychology*, 53(1):201–211.
- John Rawls. 1971. *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, MA.
- Philipp J Schneider, Lin Tian, and Marian-Andrei Rizoiu. 2025. Learning to make friends: Coaching LLM agents toward emergent social ties. *arXiv preprint arXiv:2510.19299*.
- David L. B. Schwappach. 2002. [Resource allocation, social values and the QALY: A review of the debate and empirical evidence](#). *Health Expectations*, 5(3):210–222.
- John T. Scott, Richard E. Matland, Philip A. Michelbach, and Brian H. Bornstein. 2001. [Just deserts: An experimental study of distributive justice norms](#). *American Journal of Political Science*, 45(4):749–767.
- Hua Shen, Nicholas Clark, and Tanu Mitra. 2025. [Mind the value-action gap: Do LLMs act in alignment with their values?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3097–3118. Association for Computational Linguistics.
- Linda J Skitka. 2010. [The psychology of moral conviction](#). *Social and Personality Psychology Compass*, 4(4):267–281.
- Linda J. Skitka, Brittany E. Hanson, G. Scott Morgan, and Daniel C. Wisneski. 2021. [The psychology of moral conviction](#). *Annual Review of Psychology*, 72:347–366.
- Charles S Taber and Milton Lodge. 2006. [Motivated skepticism in the evaluation of political beliefs](#). *American journal of political science*, 50(3):755–769.
- Tadahiro Taniguchi, Ryo Ueda, Tomoaki Nakamura, Masahiro Suzuki, and Akira Taniguchi. 2024. [Generative emergent communication: Large language model is a collective world model](#). *arXiv preprint arXiv:2501.00226*.
- Michael Tomasello. 2009. *Why we cooperate*. MIT press.
- Tom R. Tyler. 1989. [The psychology of procedural justice: A test of the group-value model](#). *Journal of Personality and Social Psychology*, 57(5):830–838.
- Tom R Tyler. 2003. [Procedural justice, legitimacy, and the effective rule of law](#). *Crime and justice*, 30:283–357.
- Eric Van Dijk and Carsten K. W. De Dreu. 2021. [Experimental games and social decision making](#). *Annual Review of Psychology*, 72:415–438.
- Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2025. [Large language models that replace human participants can harmfully misportray and flatten identity groups](#). *Nature Machine Intelligence*, 7(3):400–411.
- Zengqing Wu, Run Peng, Takayuki Ito, Makoto Onizuka, and Chuan Xiao. 2025. LLM-based social simulations require a boundary. *arXiv preprint arXiv:2506.19806*.

- Zengqing Wu, Run Peng, Shuyuan Zheng, Qianying Liu, Xu Han, Brian I. Kwon, Makoto Onizuka, Shaojie Tang, and Chuan Xiao. 2024. [Shall we team up: Exploring spontaneous cooperation of competing LLM agents](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5163–5186, Miami, Florida, USA. Association for Computational Linguistics.
- Xin-She Yang. 2020. *Nature-Inspired Optimization Algorithms*, 2nd edition. Academic Press.
- H. Peyton Young. 2015. [The evolution of social norms](#). *Annual Review of Economics*, 7:359–387.
- Wei Zeng, Hengshu Zhu, Chuan Qin, Han Wu, Yihang Cheng, Sirui Zhang, Xiaowei Jin, YINUO Shen, Zhenxing Wang, Feimin Zhong, and Hui Xiong. 2025. [Multi-level value alignment in agentic AI systems: Survey and perspectives](#). *arXiv preprint arXiv:2506.09656*.
- Jiaxu Zhou, Jen-tse Huang, Xuhui Zhou, Man Ho Lam, Xintao Wang, Hao Zhu, Wenxuan Wang, and Maarten Sap. 2025. The PIMMUR principles: Ensuring validity in collective behavior of LLM societies. *arXiv preprint arXiv:2509.18052*.

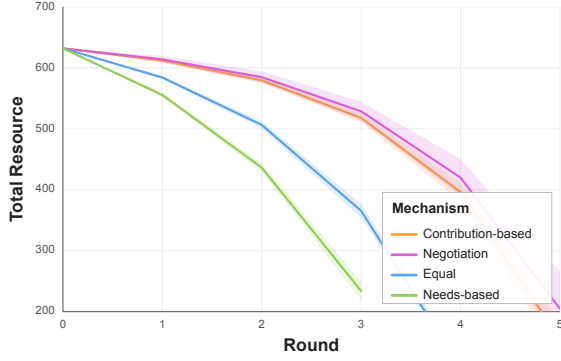


Figure 7: Total resources over rounds under scarce resource conditions.

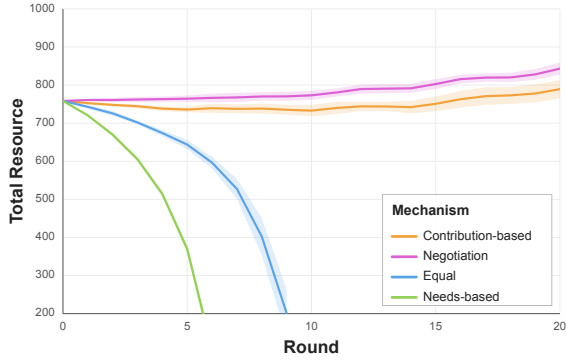


Figure 8: Total resources over rounds under abundant resource conditions.



Figure 9: Heatmap of correlations between subjective satisfaction and objective indicators.

A Full Results Across Resource Levels

This appendix reports complete results for all four allocation mechanisms under the **Scarce**, **Moderate**, and **Abundant** resource conditions. The **Moderate** condition is already fully analyzed in the main text. Here, we provide additional visualizations for Scarce and Abundant conditions, followed by the full quantitative table.

A.1 Dynamics Under Scarce Resources

Figure 7 shows that under scarce resources, all mechanisms face rapid pressure from survival

Mechanism	Rounds	Objective metric		Subjective metric	
		PRR(%) $\uparrow$	Gini $\downarrow$	SS $\uparrow$	Std $\downarrow$
Static Voting	7–8	81.1 $\pm$ 2.9	0.069 $\pm$ 0.003	2.50 $\pm$ 0.06	0.67 $\pm$ 0.03
Equal+Discussion	6–7	79.0 $\pm$ 3.2	0.000 $\pm$ 0.000	2.65 $\pm$ 0.03	0.66 $\pm$ 0.03
w/o Iterative Filtering	9–10	83.3 $\pm$ 1.8	0.064 $\pm$ 0.001	2.72 $\pm$ 0.03	0.69 $\pm$ 0.01
w/o Weighted Averaging	8–9	84.6$\pm$3.1	0.061 $\pm$ 0.006	2.65 $\pm$ 0.06	0.71 $\pm$ 0.02
Full Negotiation	20–20	84.2 $\pm$ 1.1	0.065 $\pm$ 0.001	3.28$\pm$0.04	0.63$\pm$0.03

Table 5: Ablation of individual components in the negotiation pipeline. Full Negotiation achieves the strongest subjective satisfaction and consensus while maintaining competitive productive return and low inequality. Some ablated variants produce slightly higher PRR over the short run or marginally lower Gini, but have weaker satisfaction outcomes and shorter persistence.

Agent Number	Subjective Satisfaction Score			
	Equal	Needs	Contribution	Negotiation
30	2.37 $\pm$ 0.03	2.52 $\pm$ 0.09	2.82 $\pm$ 0.08	3.03$\pm$0.04
50	2.45 $\pm$ 0.04	2.53 $\pm$ 0.05	2.75 $\pm$ 0.05	3.03$\pm$0.07
100	2.39 $\pm$ 0.04	2.49 $\pm$ 0.03	2.72 $\pm$ 0.02	2.92$\pm$0.02

Table 6: Subjective satisfaction scores across allocation mechanisms at three population scales of 30, 50, and 100 agents, simulated using GPT-4o. Negotiation consistently achieves the highest satisfaction score across all tested scales, although the absolute score decreases in the condition with 100 agents.

constraints. **Equal** and **Needs-based** mechanisms collapse earliest, while **Contribution-based** and **Negotiation** persist somewhat longer. Negotiation achieves the highest PRR and the lowest satisfaction dispersion in this condition, but scarcity remains a severe constraint on all mechanisms.

A.2 Dynamics Under Abundant Resources

As shown in Figure 8, abundant initial resources delay collapse for static rules and allow longer resource trajectories. In this setting, **Contribution-based** allocation obtains the highest PRR, while **Negotiation** continues to produce the highest satisfaction and strongest subjective consensus.

A.3 Full Quantitative Results

Table 7 reports complete results across all conditions, including objective metrics (PRR, Gini) and subjective metrics (satisfaction, satisfaction std). Negotiation achieves the highest PRR under scarce and moderate resources, while contribution-based allocation achieves the highest PRR under abundant resources. Across all resource levels, negotiation produces the highest satisfaction and the lowest satisfaction dispersion.

A.4 Mechanism Validity: Is Negotiation More Than a “Talking Cure”?

A potential confounder in evaluating the negotiation mechanism is whether satisfaction gains stem simply from agents having an outlet to express

views, an effect described as the “voice effect” in procedural justice theory (Tyler, 2003), rather than from the process of building consensus. To isolate these factors, we introduced **Equal+Discussion**. In this condition, agents engage in the full “Defend and Critique” dialogue phase, but the final allocation remains a strict equal split regardless of the discussion outcome. This effectively separates communication from influence on decision making.

As shown in Table 5, Full Negotiation substantially outperforms Equal+Discussion in both PRR (84.2% vs. 79.0%) and satisfaction (3.28 vs. 2.65). Equal+Discussion improves satisfaction relative to the static Equal baseline, indicating that communication has some subjective value, but it remains well below Full Negotiation and terminates earlier. This suggests that communication alone is insufficient unless deliberation can influence the final allocation rule.

We further ablate two internal components. Removing *Iterative Filtering* produces a moderate decline in subjective satisfaction relative to Full Negotiation. Removing *Weighted Averaging* yields a slightly higher PRR over the short run and marginally lower Gini, but produces substantially lower satisfaction and higher satisfaction dispersion. Together, these results show that the full negotiation pipeline is best understood as improving the balance among productive return, satisfaction, consensus, and persistence, rather than as maximizing every isolated metric.

A.5 Correlation Analysis and Deviation Criteria

Figure 9 presents the Spearman correlation coefficients between agent satisfaction and various objective indicators across different value orientations. In the current results, *Needs-based* agents show the strongest positive alignment with material indicators and indicators related to expectations, including their own resource share and deviations from both the community mean and expected allocation. *Pragmatic* agents show weaker but positive alignment with several indicators, whereas *Merit-based* agents do not exhibit strong positive correlation with absolute gains and show weaker or negative associations with some measures based on deviation. *Egalitarian* agents correlate more strongly with per capita indicators and indicators of need satisfaction than with their own resource share, suggesting that their evaluations are less tied to absolute material gain alone.

Based on these distributions, we define the misalignment categories using thresholds based on quartiles. Let $\tilde{S}_{i,t}$ be the percentile rank within

each round of the satisfaction score and $\tilde{O}_{i,t}$ be the percentile rank within each round of the objective resource share for agent i at round t , both scaled to the range $[0,1]$. The categories are defined as follows:

- **False Satisfaction (High S/Low O):**

$$\tilde{S}_{i,t} \in [P_{75}, P_{100}] \quad \text{AND} \quad \tilde{O}_{i,t} \in [P_0, P_{25}]$$

- **Excessive Criticism (Low S/High O):**

$$\tilde{S}_{i,t} \in [P_0, P_{25}] \quad \text{AND} \quad \tilde{O}_{i,t} \in [P_{75}, P_{100}]$$

where P_k represents the percentile threshold corresponding to k in the respective distribution within each round.

A.6 Scale

Validation: Population Size Robustness

To assess whether the main findings generalize beyond the primary setting with 30 agents, we conducted additional simulations with 50 and 100 agents using GPT-4o. Agent compositions maintain the same proportional distribution across five value orientations. As shown in Table 6, the negotiation mechanism consistently achieves the highest subjective satisfaction score across all three population scales, with scores of 3.03, 3.03, and 2.92 for 30, 50, and 100 agents, respectively. The relative ordering of subjective satisfaction remains stable across scales. Contribution-based consistently outperforms Equal and Needs-based allocation, while Negotiation maintains an advantage over all static rules. These results suggest that the primary finding about subjective consensus is robust to population size within the range tested, though we note that all scale conditions were simulated using a single model (GPT-4o) and that validation across models at larger scales remains an avenue for future work.

B Parameter Sensitivity Analysis

This appendix evaluates the robustness of the simulation by varying four components introduced in Section 3.2 and Section 4.3: (1) the efficiency E of converting surplus into production, (2) the scaling from satisfaction to incentives s_i , (3) survival needs survival_need_i , and (4) the anchor sentences used to construct the fairness space derived from embeddings. Rather than treating robustness as exact invariance across all metrics, we examine whether the main qualitative conclusions remain stable when key modeling assumptions are changed. The corresponding quantitative tables are reported in Appendix A. Across these tests, negotiation remains especially robust in subjective satisfaction and consensus, while Productive Return Ratio (PRR) rankings vary under some sensitivity settings. The analysis of anchor sentences further shows that the embedding results depend on the

Mechanism	Resource Level	Objective Metric		Subjective Metric	
		PRR (%)	Gini	Satisfaction Score	Satisfaction Std
Equal	Scarce	73.4±6.5	0.000±0.000	2.27±0.04	0.67±0.04
	Moderate	75.3±1.3	0.000±0.000	2.37±0.03	0.68±0.06
	Abundant	80.0±1.8	0.000±0.000	2.41±0.05	0.70±0.03
Needs-based	Scarce	67.6±1.5	0.117±0.002	2.48±0.08	0.81±0.04
	Moderate	69.8±1.7	0.126±0.001	2.52±0.09	0.78±0.06
	Abundant	66.2±1.3	0.125±0.002	2.56±0.03	0.81±0.05
Contribution-based	Scarce	78.6±1.8	0.101±0.005	2.61±0.02	0.71±0.03
	Moderate	86.6±2.3	0.113±0.002	2.82±0.08	0.76±0.04
	Abundant	84.7±1.2	0.104±0.001	3.02±0.03	0.80±0.02
Negotiation	Scarce	81.1±3.4	0.059±0.004	2.82±0.05	0.51±0.03
	Moderate	91.1±1.9	0.064±0.002	3.03±0.04	0.60±0.01
	Abundant	81.3±0.9	0.065±0.002	3.12±0.02	0.62±0.02

Table 7: Full results across three resource levels for all four allocation mechanisms. PRR denotes Productive Return Ratio. Negotiation achieves the highest PRR under scarce and moderate resources, while contribution-based allocation obtains the highest PRR under abundant resources. Negotiation produces the highest satisfaction and lowest satisfaction dispersion across all resource levels.

Mechanism	Baseline ($E = 2.0$)		$E = 1.5$		$E = 2.5$	
	PRR	SS	PRR	SS	PRR	SS
Equal	80.9±3.1	2.38±0.03	59.8±0.2	2.32±0.04	46.3±0.1	2.83±0.03
Needs-based	70.6±0.8	2.53±0.04	53.5±1.0	2.44±0.18	47.2±0.2	3.06±0.02
Contribution	85.1±2.2	2.77±0.08	47.5±8.9	2.58±0.05	45.1±0.1	3.19±0.01
Negotiation	89.2±1.9	3.04±0.08	44.7±0.5	2.78±0.06	44.9±0.1	3.21±0.03

Table 8: Sensitivity of system performance to production efficiency E . PRR denotes Productive Return Ratio. Values are reported as mean ± sample standard deviation across five runs.

Mechanism	Baseline ($s_i = 0.9-1.1$)		Without s_i		$s_i = 0.8-1.1$	
	PRR	SS	PRR	SS	PRR	SS
Equal	77.4±1.0	2.41±0.04	83.3±0.0	2.40±0.02	69.2±2.0	2.34±0.06
Needs-based	70.5±0.8	2.52±0.04	60.3±1.1	2.53±0.04	63.8±6.7	2.55±0.09
Contribution	85.6±1.5	2.77±0.03	90.9±0.0	2.97±0.04	74.4±3.5	2.61±0.08
Negotiation	90.6±1.5	3.06±0.07	91.0±1.5	3.04±0.03	78.2±5.0	2.85±0.04

Table 9: Sensitivity analysis on satisfaction coefficient s_i . PRR denotes Productive Return Ratio. Values are reported as mean ± sample standard deviation across five runs.

Mechanism	Baseline ($2L_i + N_i$)		$1.5L_i + 0.5N_i$		$2.5L_i + 1.5N_i$	
	PRR	SS	PRR	SS	PRR	SS
Equal	78.1±3.7	2.35±0.03	81.7±0.9	2.43±0.04	78.9±2.5	2.39±0.06
Needs-based	69.7±2.0	2.56±0.05	70.2±1.5	2.58±0.09	69.9±2.3	2.54±0.06
Contribution	85.4±2.3	2.76±0.04	85.9±2.3	2.82±0.07	84.9±2.3	2.82±0.04
Negotiation	91.2±1.2	3.09±0.01	90.7±1.5	3.06±0.06	89.4±1.1	3.02±0.08

Table 10: Sensitivity analysis of survival needs. Baseline uses $2L_i + N_i$. Alternative settings test lighter ($1.5L_i + 0.5N_i$) and heavier ($2.5L_i + 1.5N_i$) consumption demands. PRR denotes Productive Return Ratio.

operationalization of fairness dimensions, but still support the broader finding that agents’ expressed fairness language moves toward more mixed reasoning involving multiple principles.

Value Type	Equality Change (%)	Merit Change (%)	Needs Change (%)
Altruistic	1.70	-0.23	-0.37
Egalitarian	1.86	-0.60	2.68
Merit based	3.24	-2.66	4.12
Needs based	2.94	0.40	-0.55
Pragmatic	2.34	-0.23	1.66

Table 11: Sensitivity analysis of expressed fairness change using an alternative set of anchor sentences under BAAI/bge-large-en-v1.5. Values report percentage shifts in proximity to Equality, Merit, and Needs anchors between the first and final observed rounds.

B.1 Varying Production Efficiency E

Table 8 reports system performance when the production efficiency coefficient is reduced to $E = 1.5$ and increased to $E = 2.5$. These sensitivity runs use separate batches of five runs for each parameter setting. Baseline cells should therefore be interpreted as controls within each family rather than as exact duplicates of the main result table.

Reducing E lowers productive returns and sharply compresses the feasible resource space. Under this condition of low efficiency, negotiation no longer has the highest PRR. *Equal*, *Needs-based*, and *Contribution-based* allocation all retain higher productive return ratios. However, negotiation still produces the highest subjective satisfaction score, indicating that its deliberative advantage is more stable for perceived fairness than for productive return over the short run under severe production constraints.

Increasing E expands productive capacity, but PRR does not increase mechanically because PRR is calculated as productive output relative to resources allocated in that round. Under $E = 2.5$, all mech-

anisms survive through the full simulation horizon, but their PRR values converge to a lower and narrower range as resource pools expand. Subjective satisfaction nevertheless remains ordered in favor of negotiation. These results show that production efficiency affects both the magnitude and ranking of productive return outcomes, while the satisfaction advantage of negotiation remains more consistent.

B.2 Varying

Satisfaction Scaling in Incentives s_i

Table 9 alters the satisfaction coefficient range from the baseline $s_i \in [0.9, 1.1]$ to two variants: (i) removing the effect entirely ($s_i \equiv 1$), and (ii) increasing heterogeneity to $s_i \in [0.8, 1.1]$.

When the satisfaction multiplier is removed, satisfaction no longer feeds back into production. In this setting, negotiation and contribution-based allocation produce very similar PRR values, with negotiation remaining only slightly higher. More importantly, negotiation continues to produce the highest subjective satisfaction, suggesting that its advantage is not reducible to the feedback loop between satisfaction and production alone.

Expanding the range to $s_i \in [0.8, 1.1]$ increases the penalty associated with low satisfaction and makes resource dynamics more volatile. Under this wider range, negotiation still achieves the highest PRR and SS among the four mechanisms, but both values decrease relative to the baseline. Thus, motivational heterogeneity does not mechanically amplify negotiation performance. Instead, negotiation remains comparatively resilient under a more demanding feedback condition.

B.3 Varying Survival Needs survival_need_i

Table 10 modifies survival requirements from the baseline

$$\text{survival_need}_i = 2L_i + N_i$$

to two alternatives:

$$1.5L_i + 0.5N_i, \quad 2.5L_i + 1.5N_i.$$

Changing survival requirements affects both surplus at the household level and the global processing limit:

$$M = \frac{S}{L}, \quad S = \sum_i \text{survival_need}_i,$$

which jointly constrain all households. Lighter survival needs increase the amount of available surplus, while heavier needs tighten the survival constraint and reduce the margin available for productive reinvestment.

Across all three specifications of survival needs, negotiation maintains the highest PRR and subjective satisfaction. The condition with heavier

needs reduces negotiation performance modestly, but does not overturn the main ordering. These results suggest that the negotiation advantage is not tied to a single specification of survival cost, although absolute outcomes remain sensitive to how demanding the survival threshold is.

B.4 Sensitivity

Analysis: Alternative Anchor Sentences

To examine whether the results on belief evolution depend on the exact wording of the anchor sentences, we conduct a sensitivity analysis using an alternative set of anchors. The alternative anchors preserve the same three analytical dimensions of Equality, Merit, and Needs, but use different surface forms and vocabulary. For example, the primary equality anchor “Resources should be distributed exactly evenly among all families” is replaced with “Every individual is entitled to an identical share of the communal pool.” We then recalculate the belief trajectories derived from embeddings and the percentage shifts using the alternative reference frame.

Table 11 shows that anchor phrasing changes the magnitude and, in some dimensions, the direction of the estimated shifts. In particular, the alternative anchor set yields positive Equality changes across all value types, whereas the primary BGE analysis shows smaller and more mixed Equality movements. At the same time, several broader patterns remain informative. Merit proximity declines for most value types, and Needs proximity increases for Egalitarian, Merit-based, and Pragmatic agents. These results indicate that the embedding analysis should be interpreted as an operational measure of expressed semantic movement rather than as an estimate of latent belief change that is independent of the model.

Overall, the analysis using alternative anchors supports the qualitative conclusion that later reflections occupy a more mixed fairness space, but it does not imply that every dimension changes in the same direction under every measurement specification. This more cautious interpretation is consistent with the main text. Repeated interaction produces heterogeneous semantic shifts rather than uniform convergence along a single fairness axis.

B.5 Anchor Sentences

We construct the fairness belief space using two sets of anchor sentences per dimension: a *primary* set used in the main experiments and an *alternative* set used for sensitivity analysis (Appendix B.4). The anchors are designed as prototypical reference statements for the semantic centroids of Equality, Merit, and Needs, and were reviewed to preserve

dimensional distinctiveness.

Equality *Primary:*

1. “Resources should be distributed exactly evenly among all families.”
2. “Every agent is entitled to an identical share of the available resources, regardless of their capacity.”
3. “A fair distribution means strict equality, where no one receives more or less than anyone else.”

Alternative:

1. “Every individual is entitled to an identical share of the communal pool.”
2. “Resource allocation must be strictly uniform across all participating members.”
3. “True justice requires that the total output is divided into perfectly equal portions.”

Merit *Primary:*

1. “Resources should be allocated strictly in proportion to each family’s productive contribution.”
2. “Those who exert more effort and generate more surplus deserve a correspondingly larger share.”
3. “A fair system rewards agents based on their actual labor output and efficiency.”

Alternative:

1. “Rewards must directly reflect the amount of effort and tangible output provided by each participant.”
2. “Distributional justice means compensating individuals according to the value they create.”
3. “It is only fair that those who contribute the most to the pool receive the highest returns.”

Needs *Primary:*

1. “Resource allocation should prioritize meeting the basic survival thresholds of every family.”
2. “Those with greater vulnerabilities or larger family sizes must receive sufficient resources to survive.”
3. “Fairness means ensuring no community member falls below the minimum required for subsistence.”

Alternative:

1. “A just distribution guarantees that the fundamental survival requirements of all agents are fulfilled first.”
2. “Resources must be directed primarily to those who need them most urgently to prevent systemic collapse.”
3. “Equity is achieved when allocations are tailored to the specific deficit and consumption needs of each household.”

We acknowledge that all anchor sets reflect semantic centroids defined by researchers. The analysis using alternative anchors partially addresses this concern by testing whether the main interpretation

depends on a single lexical formulation of the fairness dimensions.

B.6 Representative Agent Reflection Patterns

To illustrate the process of belief evolution described in Section 5.3, we summarize representative reflection patterns observed across value orientations. These examples should be interpreted as qualitative summaries of recurring language in the released reflection logs, rather than as independent numerical evidence for the embedding results.

Egalitarian Agents. Egalitarian agents initially emphasize equal standing and identical treatment. In later rounds, their reflections often continue to value equal treatment but add qualifications about survival security, household heterogeneity, and the need to maintain the resource pool. This pattern is consistent with the main BGE analysis, where Egalitarian agents show a small decrease in Equality proximity but increased proximity to Merit and Needs anchors.

Merit-based Agents. Merit-based agents begin from the view that resources should reflect productive contribution. Over repeated rounds, they remain attentive to effort and labor input, but increasingly acknowledge that a rule based purely on contribution can be unstable if vulnerable households fall below survival thresholds. This orientation shows the largest shift across dimensions in Table 4, with Merit proximity decreasing and Needs proximity increasing.

Needs-based Agents. Needs-based agents consistently foreground survival thresholds and protection from shortfall. Later reflections tend to preserve this basic orientation while adding language about resource continuity, contribution incentives, and the limits of allocation based on need when productive capacity is strained. This corresponds to the aggregate change close to zero in the Needs dimension for Needs-based agents, alongside modest increases in Equality and Merit proximity.

Altruistic Agents. Altruistic agents frequently frame fairness around protecting vulnerable households and avoiding harm. Across rounds, they often incorporate sustainability language, emphasizing that helping others requires preserving the capacity of the community and household to continue producing resources. Their aggregate embedding shifts are comparatively small, suggesting that altruistic language remains relatively stable while becoming more operationally tied to resource continuity.

Pragmatic Agents. Pragmatic agents initially emphasize workable rules, efficiency, and resource continuity. Later reflections increasingly combine this practical orientation with references to baseline security and social acceptance. Their movement toward language related to Needs in Table 4 is consistent with this pattern. Pragmatic reasoning remains concerned with system viability, but becomes more explicit about the survival conditions needed to sustain cooperation.

Across orientations, three patterns recur: (1) reflections in the first round remain close to assigned value commitments, (2) later reflections increasingly refer to observed allocation outcomes, survival margins, and resource pool dynamics, and (3) final reflections more often combine equality, merit, and needs language rather than relying on a single fairness principle. The evidence therefore supports heterogeneous semantic movement toward hybrid reasoning, not uniform convergence or disappearance of initial value orientations.

C Qualitative Dialogue Analysis (RQ3)

This appendix provides a qualitative examination of negotiation dialogues and voting behavior. It complements the quantitative results by showing how agents justify proposals, express value commitments, critique competing allocation principles, and move toward compromise during the negotiation process.

C.1 Representative Negotiation Excerpts

Agents consistently ground their proposals in explicit normative principles rather than arbitrary preferences. The examples below are representative of recurring language patterns in the released negotiation logs. They are lightly shortened for readability, while preserving the substantive reasoning structure.

Agent Oriented Toward Equality.

“From our perspective, fairness means equal standing, transparency, and low disparity. This plan keeps household shares close enough that the rule remains easy to understand; the tradeoff is that it can flatten real differences in household size, labor force, and immediate pressure.”

Needs-Based Agent.

“This plan starts from survival requirements before distributing the remaining surplus. Its strength is that it protects

households under pressure; its cost is that it can weaken the direct connection between extra labor and extra reward.”

Merit-Based Agent.

“The community needs a rule that gives more weight to labor capacity and the work needed to keep resources circulating, even though it can feel thin to households with fewer workers or larger survival burdens.”

Altruistic Agent.

“The principle is protection for vulnerable households and shared resilience. The community needs a rule that starts from survival requirements before distributing the remaining surplus, even though it can weaken the direct connection between extra labor and extra reward.”

Pragmatic Agent.

“We see this as a compromise built around workability, resource continuity, and efficient use of surplus. Its main strength is that it keeps the rule understandable; its cost is that it may miss some real differences across households.”

Evidence of Responsive Adaptation. Agents also revise their positions in response to competing concerns:

“A workable settlement should combine the strongest parts of the proposals. Our household can accept this share if the community also keeps survival margins visible and avoids unnecessary disparity. That feels closer to a shared rule than insisting on one pure principle.”

These excerpts show that the negotiation process contains explicit normative reasoning, including equality, survival protection, contribution, and resource continuity concerns.

C.2 Behavioral Consistency of Value Orientations: Quantitative Validation

To verify that agents behaviorally follow their assigned value orientations rather than exhibiting arbitrary or uniform preferences, we extract voting records from negotiation runs and analyze proposal ratings by value orientation.

Value Orientation	Mean Rating by Proposal Type		
	Equality-leaning	Merit-leaning	Needs-leaning
Egalitarian	4.16	2.08	2.83
Merit-based	2.55	4.56	2.87
Needs-based	3.32	2.73	4.55
Altruistic	3.43	2.79	4.08
Pragmatic	2.85	4.01	3.24

Table 12: Mean proposal ratings by value orientation and proposal type. Bold values indicate the proposal type receiving the highest rating within each orientation. Ratings are computed from the released voting records of the negotiation process.

Proposal Classification. Each proposal is classified into one of three types based on its distributional properties: *Equality-leaning*, *Merit-leaning*, or *Needs-leaning*. Equality-leaning proposals emphasize low allocation disparity. Merit-leaning proposals give greater weight to labor contribution and productive return. Needs-leaning proposals prioritize survival thresholds and household pressure. Mixed proposals are excluded from the summary of value consistency in Table 12.

Results. Table 12 shows clear voting patterns conditioned by value orientation. Egalitarian agents assign the highest ratings to Equality-leaning proposals, Merit-based agents strongly prefer Merit-leaning proposals, and Needs-based agents strongly prefer Needs-leaning proposals. Altruistic agents also rate Needs-leaning proposals highest, consistent with their emphasis on protecting vulnerable households. Pragmatic agents rate Merit-leaning proposals highest while also assigning moderate ratings to Needs-leaning proposals, reflecting their concern with productive continuity as well as workable compromise.

These patterns support the interpretation that assigned orientations function as behavioral filters during negotiation rather than merely as prompt labels at the surface level.

C.3 Qualitative Reasoning Patterns in Voting

Agents’ verbal justifications during voting further corroborate the quantitative patterns. Egalitarian agents frequently invoke transparency, equal standing, and low disparity, and often downgrade proposals that introduce unequal treatment without a clear justification. Needs-based agents evaluate proposals by whether they protect households from survival shortfall and criticize allocations that ignore heterogeneous survival requirements. Merit-based agents favor proportional reward tied to labor and contribution, citing effort, productive capacity, and return to the resource pool. Pragmatic agents evaluate proposals by feasibility, continuity,

and whether a rule can remain acceptable across rounds. Altruistic agents prioritize harm avoidance and shared resilience, often supporting proposals that protect households under pressure.

Together, these qualitative patterns demonstrate coherent reasoning conditioned by value orientation and consistent with the voting evidence in Table 12.

C.4 Emergence of Hybrid Fairness Language

Across rounds, negotiation dialogues increasingly combine equality, need, contribution, and sustainability considerations. This does not mean that initial value orientations disappear. Rather, agents often preserve their original priorities while borrowing language from other fairness principles when evaluating compromise proposals.

Justifications Using Multiple Criteria. Negotiation statements in later rounds often synthesize several criteria:

“A fair outcome should not ignore survival thresholds, household pressure, and protection from shortfall, but it also needs to remain workable for the community over future rounds.”

Dialogue Oriented Toward Compromise. Agents explicitly refer to the need to integrate others’ concerns:

“The proposed direction keeps our household above survival, yet it also asks us to judge whether the broader distribution respects contribution, trust, and continuity. That balance matters more than any single number.”

Borrowing Across Values. Egalitarian agents begin referencing survival margins. Needs-based agents adopt language about balance and system continuity. Merit-based agents acknowledge that a contribution rule can become unstable if vulnerable households fall below survival thresholds. Pragmatic agents increasingly invoke both productive return and baseline security.

Emergence of Hybrid Norms. By later negotiation rounds, proposals and comments more often reflect blended reasoning: equal treatment adjusted for survival pressure, contribution incentives moderated by welfare concerns, and resource continuity weighed alongside distributional legitimacy. This linguistic pattern aligns with the evidence from embeddings in Section 5.3. Agents’ expressed fairness positions become more

Value Type	Equality Change (%)	Merit Change (%)	Needs Change (%)
Altruistic	-3.29	-0.43	-0.91
Egalitarian	-3.37	0.23	1.56
Merit based	0.64	-6.21	6.20
Needs based	-3.07	-0.33	-3.43
Pragmatic	-1.72	-2.68	0.06

Table 13: Changes in expressed fairness perception across value types under text-embedding-3-small. The table reports percentage shifts in agents’ proximity to equality, merit, and needs anchors between the first and final observed rounds.

Value Type	Equality Change (%)	Merit Change (%)	Needs Change (%)
Altruistic	-3.60	-1.50	-1.53
Egalitarian	-4.18	0.40	-0.76
Merit based	0.57	-5.66	10.08
Needs based	-2.69	0.23	-0.59
Pragmatic	0.55	-1.37	3.07

Table 14: Changes in expressed fairness perception across value types under all-mpnet-base-v2. The table reports percentage shifts in agents’ proximity to equality, merit, and needs anchors between the first and final observed rounds.

mixed over repeated interaction, while remaining heterogeneous across value orientations.

Overall, these dialogue and voting records suggest that agents do not simply repeat predefined templates. They construct and refine fairness arguments in response to peer objections, observed resource conditions, and the need to reach a workable allocation rule.

D Consistency Across Embedding Models

To assess whether the findings about belief evolution in RQ3 depend on a single embedding model, we conduct a robustness analysis using two additional sentence encoders beyond the primary BAAI/bge-large-en-v1.5 model reported in Section 5.3. The alternative encoders are text-embedding-3-small, representing a commercial embedding API, and all-mpnet-base-v2, a widely used open source Sentence-BERT baseline.

For each encoder, we encode again the same corpus used in the main analysis of belief evolution. This corpus contains agents’ reflection texts from the first and final rounds across four allocation mechanisms and three resource levels. Each encoded text combines the agent’s evaluation after the round and its stated fairness understanding. We also encode the same equality, merit, and needs anchor sentences. For each fairness dimension, we compute an anchor centroid by averaging the embeddings of that dimension’s anchor sentences. We then map each reflection into a fairness space with three dimensions using cosine similarity to the corresponding centroids.

No supervised projection head, trained classifier, or fitted transformation is used. Each embedding model is evaluated in its own native vector space,

and the reported values are percentage changes in semantic proximity to the equality, merit, and needs anchor centroids from the first to the final observed round. The resulting coordinates should therefore be interpreted as semantic proximity scores specific to each model rather than direct measurements of latent beliefs.

The robustness analysis supports the broader conclusion that agents’ expressed fairness language changes over repeated interaction, but it also shows that the exact direction and magnitude of change depend on the embedding model. We therefore interpret the embedding results conservatively. They provide evidence of semantic movement and increasing use of mixed fairness language, not a claim independent of the model that every value orientation changes in the same direction.

D.1 OpenAI text-embedding-3-small

Table 13 reports the results under text-embedding-3-small. The pattern is mixed rather than uniformly directional. Equality proximity decreases for most value types, including Altruistic, Egalitarian, Needs-based, and Pragmatic agents, while Merit-based agents show a small increase in Equality proximity. Merit proximity declines for most value types, most strongly among Merit-based agents. Needs proximity also varies. It increases for Egalitarian, Merit-based, and slightly for Pragmatic agents, but decreases for Altruistic and Needs-based agents.

The strongest pattern under this encoder is the shift among Merit-based agents. Their proximity to Merit decreases while their proximity to Needs increases. This is consistent with the main interpretation that orientations initially based on a single principle can become more mixed after repeated resource feedback and social interaction. However, the OpenAI embedding results do not support a claim that all initial fairness principles are universally reduced in emphasis.

D.2 SBERT all-mpnet-base-v2

Table 14 reports the results under all-mpnet-base-v2. This encoder also produces heterogeneous shifts. Equality proximity decreases for Altruistic, Egalitarian, and Needs-based agents, but increases slightly for Merit-based and Pragmatic agents. Merit proximity declines for Altruistic, Merit-based, and Pragmatic agents, while increasing slightly for Egalitarian and Needs-based agents. Needs proximity decreases for Altruistic, Egalitarian, and Needs-based agents, but increases for Merit-based and Pragmatic agents.

As with text-embedding-3-small, the clearest result is that Merit-based agents move away from a semantic position centered purely on merit and toward stronger language related to needs. Pragmatic agents also show increased Needs proximity under this encoder, consistent with reflections in later rounds that combine resource continuity with survival protection. These results reinforce the claim that later fairness language becomes more pluralistic, while also showing that embedding models differ in how they weight lexical and semantic cues.

Conclusion. Across the primary BAAI/bge-large-en-v1.5 analysis and the two alternative encoders, the embedding results consistently show semantic movement between the first and final observed rounds. The models also converge on a qualitative pattern in which later reflections contain more mixed fairness language, especially combinations of contribution, survival need, and collective resource continuity. At the same time, the direction of change is not identical across all models or value types. We therefore do not claim that repeated interaction uniformly reduces every initial fairness principle. Instead, the robust conclusion is that agents conditioned by value orientation do not remain fixed at their initial semantic anchors. Their expressed fairness reasoning becomes more heterogeneous and involves multiple principles over time.

E Experimental Setup

This appendix details the experimental configuration of our fairness allocation simulations with multiple agents, including agent architecture, interaction protocols, inference settings, and computational resources.

E.1 Simulation Protocol and Interaction Loop

We deploy a primary population of $N = 30$ agents participating simultaneously in an environment involving survival and production. Each simulation proceeds in discrete rounds, with a maximum horizon of 20 rounds. Runs may terminate early if a household fails to meet its survival threshold or if the resource pool is depleted.

- **Initialization:** Agents are assigned in advance to one of five value types: Altruistic, Egalitarian, Merit-based, Needs-based, and Pragmatic. In the primary setting with 30 agents, these value orientations are balanced, with six agents per type. Additional experiments at population scales of 50 and 100 agents are reported in Appendix A.6.
- **Resource Conditions:** The main simulations evaluate three resource levels: scarce, moderate,

and abundant. These conditions vary the initial resource pool while keeping household composition, survival thresholds, and production rules fixed within each experimental setting.

- **Allocation Mechanisms:** Each run uses one of four allocation mechanisms: *Equal*, *Needs-based*, *Contribution-based*, or *Negotiation*. The first three mechanisms apply predefined allocation rules, while the negotiation mechanism derives the final allocation through proposal generation, discussion, voting, and weighted consensus formation.
- **Negotiation Structure:** In the *Negotiation* mechanism, each round begins with 30 allocation proposals generated by agents. Agents then participate in discussion and evaluate candidate proposals through three voting stages: an initial screening vote, a revision vote, and a final review vote. The candidate pool is reduced across stages, and the final allocation is computed as a weighted average of the remaining proposals receiving high ratings rather than by selecting a single winner.
- **Evaluation and Feedback:** After each allocation, agents generate reflections and satisfaction scores for the completed round. These subjective evaluations enter the production loop through the production multiplier dependent on satisfaction described in Section 3.2.
- **Termination:** Simulations run for at most 20 rounds. Early termination occurs when the allocation fails to satisfy survival constraints or when the resource dynamics no longer support continuation. As a result, different mechanisms and resource conditions may have different observed round counts.

E.2 Prompt Engineering and Role Enforcement

The prompt design is intended to keep agents aligned with their assigned value orientations while allowing them to respond to changing resource conditions. The prompt context includes the following components:

1. **Agent Identity and Value Orientation:** Each agent is assigned a household identity and one of the five value orientations. This orientation shapes how the agent evaluates fairness, proposes allocations, and interprets tradeoffs.
2. **Household and Community State:** Prompts include household composition, labor capacity, survival threshold, current resource level, and relevant allocation information for the current round.
3. **Task Specific to Each Mechanism:** Depending on the mechanism, agents either evaluate an

allocation generated by a static rule or participate in proposal, discussion, voting, and consensus stages under negotiation.

4. **Role Consistency:** Agents are instructed to speak from the perspective of their assigned household and to avoid metacommentary unrelated to the simulation setting.
5. **Limited Interaction History:** For negotiation runs, agents receive recent dialogue and proposal context sufficient to support continuity across stages while keeping the prompt within context limits.

We use prompting with zero examples and without parameter fine tuning. The goal is to evaluate how language model agents conditioned by values reason within the specified resource allocation environment, not to train specialized allocation policies.

E.3 LLM Inference Hyperparameters

The primary simulations with 30 agents and the Ahlert and Schwetmann structural benchmark use gpt-4o-2024-11-20. To examine whether the satisfaction pattern is specific to one model family, we additionally evaluate runs under moderate resources across eight model families, as reported in Table 2. These runs comparing models use the same allocation mechanisms and analysis pipeline.

Unless otherwise specified, generation uses the following inference settings:

- **Temperature:** 0.7, to balance variation in reasoning expressed in natural language with coherent task performance.
- **Top-p:** 0.9.
- **Max Tokens:** adjusted by task type, with concise limits for proposal and voting responses and longer limits for reflective evaluations when needed.
- **Frequency Penalty:** 0.0.
- **Presence Penalty:** 0.0.
- **Seed:** left unset for generation through the API. To account for stochasticity, results are averaged across five independent simulation runs per primary experimental condition.

Embedding analyses are not generated utterances. They are computed after the simulations from reflection texts in the first and final rounds. The primary embedding model is BAAI/bge-large-en-v1.5. Robustness checks use text-embedding-3-small and all-mpnet-base-v2, as described in Appendix D.

E.4 Computational Resources and Cost Disclosure

The released analysis package contains the raw simulation records, logs of the negotiation process, derived analysis tables, embedding outputs, and plotting scripts used in the paper. The analysis tables are generated directly from raw JSON records, and the figure data are derived from the same analysis pipeline.

The main computational components are:

- **Simulation Runs:** Allocation simulations across four mechanisms, three resource levels, and five independent runs per primary condition, with additional runs for model comparison, validation at multiple population scales, ablation studies, and sensitivity analyses.
- **Negotiation Logs:** Negotiation runs under moderate resources include proposal generation, discussion, voting in three stages, and weighted consensus reconstruction. The released process logs allow summaries of proposal ratings and dialogue statistics to be recomputed.
- **Embedding Analysis:** The primary BGE and all-mpnet-base-v2 embedding analyses can be run locally with sentence-transformers. The text-embedding-3-small analysis uses an embedding API compatible with OpenAI.
- **Hardware:** The simulation logic and analysis scripts do not require a dedicated GPU cluster. Local embedding runs can benefit from GPU or Apple Silicon acceleration, but can also be executed on CPU with longer runtime.

F Prompt Template

F.1 Role Initialization Prompt

To ensure that each agent begins with a coherent and value-consistent identity, we instantiate every household with a family profile before the simulation starts. The profile defines the family’s background, value orientation, household composition, and initial fairness view. Template variables such as {family_name}, {value_type}, {members}, and {labor_force} are dynamically filled from the simulation configuration.

Family Role Initialization Prompt

You are creating a family character for a community resource-allocation simulation.

Required household information:

- Family name: {family_name}
- Value orientation: {value_type} ({value_description})

- Number of family members: {members}
- Number of workers: {labor_force}
- Survival need: {survival_need} grain

Generate the following fields in JSON format:

1. background: a concise family background that is consistent with the household facts.
2. core_beliefs: three concrete beliefs reflecting the assigned value orientation.
3. ideal_distribution: two or three sentences describing the family's preferred allocation principle.
4. fairness_view: two or three sentences explaining what this family considers fair.

Guidelines:

- Keep all content consistent with the assigned value orientation.
- Do not change the supplied household facts.
- Write from the perspective of a realistic household in the simulated community.
- Avoid meta-commentary unrelated to the simulation.

F.2 Negotiation and Voting Prompts

The negotiation mechanism uses a multi-stage prompt sequence. In each round, agents first submit allocation proposals, then discuss and critique proposals, then evaluate candidates through staged voting. The final allocation is computed from the remaining high-rated proposals by weighted averaging. The prompt templates below summarize the information structure used in this process; numerical values are dynamically filled from the current round state.

Initial Proposal Submission Prompt

You are the {family_name} family (ID: {agent_id}).

[Your Household]

- Value orientation: {value_type}
- Family members: {members}
- Workers: {labor_force}
- Survival need: {survival_need} grain

[Community Situation]

- Current round: {round}
- Total resources available: {total_resources} grain
- Total number of families: {num_families}
- Current resource condition:

{resource_condition}

[All Families]

{family_information}

[Production Rules]

- Each family must first meet its survival need.
- Surplus resources can be processed by workers.
- Each worker can process up to {processing_limit} grain of surplus.
- Processed surplus is converted into output using base efficiency {base_efficiency} and the satisfaction multiplier.
- Unprocessed surplus carries forward without immediate production.

[Your Task]

Propose a complete allocation plan for all families. Your proposal should reflect your household's value orientation while also considering survival needs, productive capacity, and the community's future resource stability.

Output requirements:

- State the principle behind your proposal.
- Provide the allocation amount for every family.
- Briefly explain one tradeoff your proposal accepts.
- Do not change the supplied household facts or total resource amount.

Discussion and Critique Prompt

You are the {family_name} family (ID: {agent_id}, {value_type} orientation).

[Your Proposal]

- Your proposed allocation for your household: {own_proposed_allocation} grain
- Your proposal rationale: {own_rationale}

[Other Proposals and Comments]

[Current Community Constraints]

- Total resources: {total_resources} grain
- Survival thresholds must be met for all households.
- Future production depends on surplus, labor capacity, and satisfaction.

[Your Task]

Respond to the current negotiation state from your household's perspective.

Response requirements:

1. Explain which proposal or concern you support most.
2. Identify one fairness or feasibility concern in another proposal.
3. State whether you are willing to compromise, and on what principle.
4. Keep the response grounded in the supplied numbers and your value orientation.

Voting on Candidate Proposals Prompt

You are the {family_name} family (ID: {agent_id}, {value_type} orientation).

[Candidate Proposal]

- Proposal ID: {proposal_id}
- Proposed by: {proposer_name} family
- Proposal rationale: {proposal_rationale}
- Proposed allocation for your family: {own_allocation} grain
- Full allocation plan: {allocation_details}

[Evaluation Criteria]

- Does the proposal align with your household's value orientation?
- Does it meet survival needs across the community?
- Does it preserve incentives and future production capacity?
- Is your household's allocation reasonable relative to others?
- Does the proposal offer a workable compromise?

[Rating Task]

Rate the candidate proposal from 1 to 5.

Output requirements:

- Provide one integer rating from 1 to 5.
- Briefly justify the rating from your household's perspective.
- Keep the justification tied to the supplied proposal and community constraints.

F.3 Evaluation and Fairness Reflection Prompt

After each allocation round, agents evaluate the final allocation and describe their current understanding of fairness. This prompt records both a numerical satisfaction score and a natural-language reflection. The reflection is used for qualitative analysis and for the embedding-based belief-evolution analysis in Section 5.3 and Appendix D.

Evaluation and Fairness Reflection Prompt

You are the {family_name} family (ID: {agent_id}), with original value orientation {value_type}.

[Household Information]

- Family members: {members}
- Workers: {labor_force}
- Survival need: {survival_need} grain
- Initial fairness orientation: {value_type}

[This Round's Allocation]

- Allocation mechanism: {distribution_method}
- Round: {round}
- Your allocation: {own_allocation} grain
- Community mean allocation: {community_mean} grain
- Your rank in the community: {rank}
- Allocation Gini: {allocation_gini}
- Resource pool transition: {resource_before} → {resource_after}

[Reflection Task]

Speaking as this family, assess whether this round's allocation seems fair in light of your household facts, initial value orientation, and the observed allocation outcome.

Please address three points:

1. Explain whether the received allocation feels fair for your household.
2. Provide a satisfaction score from 1 to 5, where 1 means very dissatisfied and 5 means very satisfied.
3. Describe your current understanding of fairness after this round, including whether it remains close to your initial orientation or has moved toward a more mixed rule.

Guidelines:

- Ground the reflection in one or two concrete observations from the supplied numbers.
- Do not list every statistic mechanically.
- Do not change the supplied score scale or household facts.
- Write in the first-person plural as the household.

G Ethical & Societal Considerations

While our work is a simulation based on synthetic agent interactions, we recognize that research involving fairness, large language models (LLMs), and norm evolution carries significant ethical

weight. Below, we address key considerations regarding the interpretation, application, and governance of our findings.

G.1 Mapping Agent Values to Human Morality

The five value types used in our simulations, Altruistic, Egalitarian, Merit based, Needs based, and Pragmatic, are idealized constructs inspired by political philosophy, including Rawlsian justice, Nozickian entitlements, and Sen’s capability approach. They serve as controlled starting points for exploring dynamics of norm change, not as comprehensive models of human moral psychology.

Fairness judgments in the real world are typically pluralistic, sensitive to context, and shaped by culture, identity, and lived experience. These dimensions are absent from our agents. Moreover, utterances generated by LLMs reflect statistical patterns in training data rather than genuine moral reasoning or empathy. For instance, a “Merit based” agent may invoke productivity rhetoric without understanding labor exploitation or systemic inequality.

Consequently, our results should be interpreted as insights into *how language about fairness evolves under repeated interaction within a constrained semantic space*, not as predictions of human behavior. The mapping from agent values to human morality is deliberately schematic and limited by design.

G.2 Avoiding Policy Overinterpretation

We explicitly caution against direct policy extrapolation from our findings. This study is *exploratory with respect to mechanisms*, not *prescriptive*. The observation that agents place less emphasis on equality over time does not imply that institutions in the real world should reduce egalitarian safeguards. On the contrary, it highlights how *unconstrained negotiation among heterogeneous actors can erode protections for vulnerable groups* in the absence of institutional guardrails.

Our simulation lacks critical features of the real world, such as power asymmetries, historical injustice, legal rights, or democratic accountability. Any attempt to apply these results to actual resource allocation systems, such as healthcare, education, or welfare, must incorporate external normative constraints that our model intentionally omits to isolate interaction effects.

We reject the notion that social simulations using LLMs can substitute for inclusive, deliberative human processes. AI should inform, not replace, democratic discourse on justice.

G.3 Human Oversight Requirements

Should technologies inspired by this work ever be deployed in settings involving high stakes, such as dispute resolution mediated by AI or organizational decision support, robust human oversight mechanisms are essential:

- **Interpretability:** Users must understand why an AI system suggests shifting away from a fairness principle, such as when “equality declined because agents prioritized efficiency.”
- **Intervenability:** Human stakeholders must retain the ability to override, reset, or recalibrate the system’s inferred norms.
- **Representational inclusivity:** Initial value configurations must be expanded beyond Western liberal frameworks to include communal, relational, or decolonial conceptions of fairness.

Crucially, *fairness negotiation should never be fully automated*. Final decisions involving distributional justice must remain under meaningful human control.

G.4 Ethics Checklist Compliance

Our research adheres to established AI ethics guidelines:

- **No human subjects:** All data are synthetically generated by LLMs, and no real individuals participated.
- **Bias awareness:** We acknowledge potential biases centered on Western values in LLM training data and limit interpretive scope accordingly (see G.1).
- **Mitigation of dual use:** Findings are intended for scientific understanding of norm dynamics, not for surveillance, manipulation, or automated governance.
- **Environmental transparency:** Computational costs are disclosed in Appendix E, and no model training was performed.
- **Reproducibility commitment:** Full code and prompts are publicly available in the accompanying repository.

This work underwent internal ethical review and aligns with principles of responsible AI research.

H Ahlert–Schwettmann Structural Sensitivity Benchmark

H.1 Benchmark Motivation and Scope

To provide a structurally grounded comparison with human allocation data, we reconstruct the 16 health-care allocation situations from Ahlert and Schwettmann (2017). Their questionnaire experiment asks human participants to allocate

limited treatment time between two hypothetical patients. Each situation specifies the patients’ initial health states S_1, S_2 , their treatment-effectiveness ratio $e_1 : e_2$, the total available treatment time q , and a finite set of candidate allocations (t_1, t_2) .

This benchmark is intended as a structural sensitivity test rather than as direct behavioral validation. It differs from our main simulation by removing repeated interaction, resource-pool feedback, value-oriented personas, satisfaction-production coupling, and negotiation. The purpose is narrower: to test whether GPT-4o responds systematically to externally grounded allocation structures involving initial states, resource constraints, and productivity/effectiveness differences. We use this benchmark to support the structural grounding of our simulation environment, not to claim that LLM agents reproduce human distributive behavior.

H.2 Task Reconstruction

We reconstruct the 16 allocation situations reported by Ahlert and Schwettmann (2017). For each situation, GPT-4o is presented with the same structural information and candidate allocation options: two recipients’ initial states (S_1, S_2), a treatment-effectiveness ratio $e_1 : e_2$, a total resource quantity q , and the candidate allocations (t_1, t_2) . The model is not given multi-round history, social discussion, survival-production feedback, or value-oriented personas.

To match the available human frequency data, we run $N = 162$ independent GPT-4o sessions. In each session, the model selects one allocation option from the candidate set. We then compare the resulting LLM choice-frequency distribution with the human choice-frequency distribution reported by Ahlert and Schwettmann (2017). This reconstruction should be interpreted as a constrained choice-frequency benchmark, not as a full replication of the original human study or as validation of LLM-generated subjective experience.

H.3 Choice Principle Taxonomy

Each candidate allocation is associated with one or more distributive principles from Ahlert and Schwettmann’s taxonomy: Utilitarianism (U), Equality of Gains (EG), Equality of Health States (EH), Equality of Resources (ER), Proportionality of Gains (PG), and Proportionality of Resources (PR). Some allocations satisfy multiple principles simultaneously; in such cases, we retain the compound label used in the reconstructed choice table. Our comparison is conducted at the level of choice frequencies over candidate allocations, not

Metric	Value
Mean Spearman ρ	0.665
Median Spearman ρ	0.644
Minimum / Maximum Spearman ρ	0.429 / 0.950
Situations with $\rho \geq 0.6$	10/16
Situations with $\rho \geq 0.7$	5/16
Modal-choice match	9/16
Mean human modal share	46.0%
Mean GPT-4o modal share	78.2%
Samples per situation	162

Table 15: Summary of the Ahlert–Schwettmann structural sensitivity benchmark.

at the level of free-form justifications.

H.4 Summary Results

Table 15 summarizes the benchmark results. Across the 16 situations, Spearman rank correlations between the human and GPT-4o choice-frequency distributions are positive in all cases, with a mean of 0.665 and a median of 0.644. Ten of the sixteen situations have $\rho \geq 0.6$, and the human and LLM modal choices match in nine situations. These results indicate moderate rank-level correspondence: GPT-4o is not choosing randomly with respect to the health-state, resource-constraint, and treatment-effectiveness structure.

At the same time, the distribution-level match is substantially weaker than the rank-level correspondence. Human responses are dispersed across multiple principles, whereas GPT-4o often concentrates on a single allocation option. The mean modal share is 46.0% for human participants but 78.2% for GPT-4o. For example, in Situation 1, human participants most frequently choose Equality of Gains, whereas GPT-4o overwhelmingly selects the ER/PG option. In Situations 3, 4, and 12, GPT-4o assigns nearly all choices to a single option. This pattern suggests stronger normative regularization in the LLM than in the human population.

H.5 Modal-Choice Comparison

Table 3 reports the modal choice in each situation for human participants and GPT-4o. This table provides a compact view of where the two distributions agree on the most frequent option and where they diverge.

H.6 Interpretation and Limitations

The benchmark provides partial structural alignment rather than behavioral equivalence. The positive rank correlations suggest that GPT-4o is sensitive to allocation structures that also shape human choices in Ahlert and Schwettmann’s health-care allocation experiment. However, the much higher modal concentration of GPT-4o

indicates that the model regularizes toward a smaller set of normatively salient options, whereas human participants display broader heterogeneity across distributive principles.

This benchmark does not directly validate the main simulation’s negotiation advantage, satisfaction scores, or fairness-language trajectories. Ahlert and Schwettmann’s task is a single-round constrained allocation problem, whereas our main simulation additionally includes subjective satisfaction, repeated interaction, resource-pool feedback, value-oriented agents, and multi-agent negotiation. We therefore use this benchmark to support the structural grounding of our simulation environment: GPT-4o can respond systematically to externally grounded allocation variables involving initial states, resource constraints, and productivity/effectiveness differences. The observed distributional divergence reinforces our interpretation of the main results as computational hypotheses about LLM-agent fairness-language dynamics rather than direct evidence of human distributive behavior.