Understanding Economic Tradeoffs Between Human and AI Agents in Bargaining Games*

Crystal Qian† Google DeepMind
Kehang Zhu† Harvard
John J. Horton MIT & NBER
Benjamin S. Manning MIT
Nithum Thain Google DeepMind
Vivian Tsai Google DeepMind
James Wexler Google DeepMind

March 21, 2025

Abstract

As AI agents become increasingly capable of completing tasks traditionally handled by humans, the tradeoff between agency (optimizing for rational, self-interested decisions) and alignment (reflecting human preferences, including social and strategic considerations) becomes more pressing. This work introduces a framework to quantify these tradeoffs, including the *Price of Agency*, a novel metric capturing the human-LLM alignment gap. In addition, we introduce a bespoke structured bargaining game that directly varies key parameters highlighted by our framework, such as endowments, the number of goods, and payoff structures. We leverage this game to evaluate our framework, analyzing game play by LLM-based agents, statistical agents, rule-based agents, and an online population of humans. Our results show that while out-of-the-box LLM agents with simple prompting performed slightly worse than humans, they offered novel strategies and time savings. Humans exhibited greater strategic and adaptive play, leading to performance improvements, yet had more opportunity cost errors relative to the more conservative LLM-based play style. The strategic wins of human players, combined with the efficiency and scalability of AI-driven play, suggest a future where AI augments rather than replaces human decision-making. As AI systems become more embedded in economic and strategic interactions, balancing agency and alignment will be crucial for designing AI that optimizes outcomes while reflecting human values.

*Author contact information, code, and data are currently or will be available at `http://www.benjaminmanning.io/`.
†Both authors contributed equally to this work. They are listed alphabetically by last name.

1 Introduction

Recent advances in artificial intelligence (AI) – particularly large language models (LLMs) – have accelerated the development of *AI agents*: autonomous systems capable of performing complex, imperfectly specified tasks on behalf of people. While the idea of delegating work to AI is not new, the latest generation of AI agents differs fundamentally from earlier software systems. These new agents can interpret under-specified directives and apply broad, pre-trained knowledge to fill in missing details, expanding the scope of tasks that can be delegated. Industry implementations—ranging from automatic supplier negotiations at major retailers (Van Hoek et al., 2022) to online shopping (Gaarlandt et al., 2025) to AI-assisted code generation (Anthropic, 2025)—suggest that the rollout of such agents is proceeding rapidly.

Despite growing enthusiasm and experimentation, relatively little research addresses the full set of economic tradeoffs involved in choosing whether to delegate tasks to a human or an AI agent. However, this line of work can have significant practical implications for both individual and social welfare. As AI agents improve their capabilities to handle increasingly unstructured and complex tasks, the relevant question becomes not just "Can AI do this?" but "Should it do this—compared to a human or another agent type?" The answer, of course, depends heavily on context: the factors that matter in one domain may be less relevant in another. While some efforts have analyzed this question within specific settings (Agarwal et al., 2023), the lessons learned often do not translate easily elsewhere. A more widely applicable framework is therefore needed—one that can be adapted to potentially complex, uncertain environments and guide decision-makers through the heterogeneous tradeoffs involved.

This paper proposes a practical framework for evaluating the economic tradeoffs in choosing between different types of agents. While broader ethical and social concerns are undoubtedly important, this work focuses specifically on quantifiable factors that inform real-world decision-making. Concretely, we identify six key dimensions that managers, policymakers, or individuals might weigh: (i) Preference parameters, such as risk tolerance or the alignment of agent incentives with the principal's interests; (ii) Strategic capabilities, meaning the ability to anticipate and respond to other parties' behavior; (iii) Time and cost, encompassing both direct monetary expenses and managerial overhead; (iv) Information sets, which determine what the agent knows and how this knowledge affects choices; (v) The structure of the environment, for example whether it is a one-shot bargain, a repeated interaction, or a market context; and (vi) Complexity, including the number of agents, the variety of goods, or other dimensions that can exacerbate coordination challenges.

For instance, some tasks—such as auction design—may emphasize strategic sophistication, whereas others, like AI-driven portfolio management, hinge on aligning risk preferences. We remain agnostic about which dimension matters most in any given scenario. To be clear, our framework is not a panacea but rather a starting point for evaluating tradeoffs in a rapidly advancing AI landscape. Taken together, however, these considerations offer a comprehensive approach to determining when substituting agents might deliver net benefits—and when such substitution could instead introduce new risks or inefficiencies.

Although our framework is designed for broad applicability, we illustrate its usage in a domain that economics has studied extensively: multi-agent bargaining. Bargaining offers a compelling context in which to illustrate this framework for several reasons. As already mentioned, AI agents are proliferating as bargainers and negotiators—from corporate supplier negotiations to online shopping assistants—yet decades

of research emphasize that it is often complicated by asymmetric information, incomplete specifications, and strategic misalignment. Classic results show that fully efficient, incentive-compatible bargaining outcomes are generally unachievable under even basic conditions (Myerson and Satterthwaite, 1983), underscoring the need for imperfect but adaptive agents to mitigate inefficiencies. At the same time, bargaining lends itself to a self-contained, controlled experimental design that still retains core real-world features. By systematically varying who participates, how information is shared, and how transactions are structured, researchers can scrutinize nearly every dimension of the proposed framework. In doing so, bargaining naturally showcases the full range of tradeoffs.

To explore these issues in a concrete, self-contained setting, we design and implement a new chip-trading bargaining game under incomplete information. Each player (human or otherwise) starts with a distinct endowment of multiple goods—colored chips—with valuations that are randomly assigned and typically allow for substantial mutual gains. Over several rounds, participants propose trades by offering one color in exchange for another, yet they lack complete knowledge of each other's valuations. This setup introduces strategic complexity: while shrewd proposals can yield significant benefits, mistaken offers or acceptances can lead to lost opportunities. We test various types of agents—humans (in an online experiment), LLM-based AI, and simpler rule-based or statistical agents—all facing the same initial allocations. Human participants receive real monetary compensation based on the difference between their initial and final chip holdings, ensuring that incentives are salient.

We chose to design this custom game, rather than use an existing design, so that we could directly control and vary the parameters highlighted by our framework (e.g., endowments, number of goods, payoff structures) and maintain full control over each agent's utility function—allowing for precise, apples-to-apples comparisons across agent types. Additionally, by introducing a novel environment, we reduce the likelihood that large language models simply reproduce known optimal strategies from widely available corpora (Ludwig et al., 2024), and we mitigate the well-known "garbage-in–garbage-out" criticism (Bender et al., 2021) often directed at common negotiation benchmarks. The environment is also neither trivially solvable nor so complex as to preclude meaningful theoretical analysis. Indeed, we employ linear programming (LP) to compute a Pareto-efficient reference allocation—feasible only with perfect information—as a yardstick for real-world bargaining outcomes. From there, we define the "Price of Agency" to gauge the efficiency gap between human and AI performance, capturing surplus gains relative to this LP optimum. Together, these elements form a structured basis for examining when different agent types can approach (or deviate from) the theoretical ideal under varying information and strategic constraints.

We found that although both humans and our LLM-based agents achieved high efficiency in simpler versions of this bargaining game, human participants still had a slight edge over the AI as task complexity grew. This gap was not large, however, considering that we employed an off-the-shelf foundation model with minimal prompt engineering. In fact, LLM-based agents captured the majority of potential surplus with relatively straightforward prompts and completed the task significantly faster than humans—who exhibited clear signs of cognitive effort. While human players generally exhibited more sophisticated strategic reasoning and long-term planning, such advantages may diminish as large language models continue to improve (Google Flash Thinking, 2024). Overall, these results underscore that our multi-dimensional framework can help managers and policymakers navigate the evolving tradeoffs of when and how to integrate, delegate, or complement human decision-making with AI.

In short, this paper offers a roadmap for systematically assessing both the economic tradeoffs and the nuances of alignment that arise when substituting AI for human decision-making. The subsequent sections detail the formal underpinnings of our framework, elaborate on the bargaining experiment design, and discuss the empirical findings in depth. We conclude by reflecting on broader implications and identifying directions for future research in this rapidly evolving AI landscape.

1.1 Related literature

The study of strategic interactions offers a rich framework for analyzing decision-making among multiple, interdependent agents (see, e.g., Fudenberg and Tirole, 1991). Within this domain, recent work has increasingly focused on the capabilities of autonomous agents, especially those powered by large language models (LLMs), to perform effectively in complex strategic settings such as gaming and negotiation. While some researchers have explored broader frameworks for integrating AI into existing human workflows (Lubars and Tan, 2019), relatively little attention has been paid to explicitly quantifying the economic tradeoffs that arise in practice when deciding whether an AI or a human should undertake a task.

Strategic LLM Agents and Equilibria It has long been recognized that a divergence exists between socially optimal outcomes and practical performance in strategic settings (Nash Jr, 1950). To quantify this gap, concepts such as the *Price of Anarchy* (Koutsoupias and Papadimitriou, 1999) and the *Price of Stability* (Anshelevich et al., 2008) were introduced to measure the inefficiency arising from equilibrium strategic play compared to socially optimal solutions. Recent work by Raman et al. (2024) has begun to explore the ability of LLM agents to take rational economic actions, including playing equilibrium strategies. Their findings suggest that most current generation models struggle to consistently play equilibrium strategies across a wide class of games. In response to these limitations, Goktas et al. (2025) have called for greater emphasis to be placed on developing strategic capabilities in foundation models, arguing that this is crucial for these models to better represent humans as agents in strategic contexts.

LLM Agents Acting as Human Proxies. A key motivation for using LLM-based agents is their potential to serve as proxies for human participants in complex, time-consuming decision processes. Krishna et al. (2022) introduced the idea of socially situated AI, arguing that LLMs can refine their understanding of human norms by interacting with individuals in natural settings. Empirical work by Park et al. (2023) showed how generative agents can exhibit emergent social behavior in simulated "virtual towns," while others have demonstrated that LLMs, given human-like preference prompts, can mimic human responses in survey and experimental contexts (Horton, 2023; Park et al., 2024; Manning et al., 2024). Others have shown that LLM agents can approximate human decision-making in auctions, bargaining, marketing, and pricing (Fish et al., 2024; Zhu et al., 2024; Soumalias et al., 2025).

LLMs in Real-World Settings. Beyond controlled experiments, recent studies underscore the emerging role of LLM agents in practical, high-stakes scenarios. (FAIR) combined language-based reasoning modules with strategic planning to achieve near-human performance in Diplomacy, a negotiation-centric board game. Kramár et al. (2022) explored trade proposals in Diplomacy using a propose-reject methodology. Tessler et al. (2024) showed how LLM agents can facilitate more constructive dialog between

human negotiators, essentially moderating toward common ground. However, questions persist about alignment and user trust in real-world deployments. For instance, Palminteri et al. (2024) found that agreement with AI-generated choices can depend on whether individuals realize the recommendation comes from an AI. Work by Kapania et al. (2022) and Qian and Wexler (2024) highlights that users may over- or under-utilize AI when their perceptions of the technology—or their ability to calibrate its capabilities—are misaligned with actual performance.

2 Conceptual Framework

In this section, we detail the key dimensions of our proposed framework for comparing different types of agents—human or AI—in economic contexts. While no single method can capture every nuance of real-world decision-making, we aim to identify the most salient factors that shape how agents perform, what risks they entail, and how well they align with a principal's objectives. Taken together, these dimensions provide a starting point for systematic evaluation, even if the relative importance of each dimension varies from one setting to another.

Preference Parameters. An agent's underlying preferences—and how they align with the principal's—are crucial to understanding its likely behavior. In many economic contexts, these preferences include risk tolerance, time discounting, or concern for fairness. A portfolio manager, for instance, must match the investor's risk appetite; any mismatch can lead to large welfare losses. In practice, AI agents may be "hard-coded" with certain objectives (e.g., maximizing expected returns) that omit crucial subtleties like risk aversion, while human agents naturally bring idiosyncratic beliefs and biases. Assessing whether the agent's preference parameters closely align with those of the task sponsor is therefore essential for predicting outcomes such as the level of caution in bidding, the aggressiveness of bargaining stances, and the propensity to exploit or ignore small gains.

Strategic Capabilities. Beyond having well-defined preferences, an agent must also anticipate and respond to others in strategic interactions. Bargaining, auctions, pricing decisions in oligopolies, and many other economic settings demand the ability to model or learn opponents' behaviors. Indeed, as Goktas et al. (2025) emphasizes, strategic reasoning—rather than pure statistical prowess—is increasingly critical for foundation models that operate in multi-agent environments. In practice, these capabilities can be viewed both vertically (the depth or sophistication of iterative reasoning) and horizontally (the range of strategies an agent can implement effectively). AI agents with advanced reinforcement learning or game-theoretic algorithms may systematically outperform simpler rule-based systems if the environment rewards iterative adaptation. Conversely, a human negotiator might excel at reading nuanced signals of trust or deception that current AI might overlook. From a policy or managerial standpoint, evaluating how well an agent can handle repeated interactions, partial cooperation, or complex strategic dynamics becomes central to deciding whether it should replace or complement human decision-makers.

Time and Cost. Although true zero marginal cost for AI may not be immediate, the per-unit expense of deploying an AI agent can approach negligible levels, especially for large-scale or repetitive tasks

(e.g., routine contract negotiations). Meanwhile, human agents carry salary costs, training, and other expenses. However, AI systems may require significant oversight or monitoring to ensure outputs align with organizational norms, offsetting some of their apparent cost advantage. A careful analysis of both direct and indirect costs is thus crucial to determining whether an AI agent will genuinely be more economical over the relevant time horizon.

Information Sets. Every agent's decisions hinge on what it knows—or can infer—about the environment and other participants. Human agents may have rich contextual knowledge and adaptive social cues, while AI agents can exploit large training sets and systematic inference algorithms. However, if an AI's training data lack critical context or feature biases, it might misuse or misinterpret signals that humans naturally weigh. Likewise, tasks involving private or tacit knowledge (e.g., insider relationships or subtle norms) may demand a human's interpretive judgment. Analyzing how differing information sets impact each agent's strategy and ability to learn or adapt thus reveals an important dimension of the agent-comparison problem.

Structure of the Environment. Economic tasks vary immensely in their structure—some are one-shot interactions with simple payoffs, whereas others feature repeated encounters, network externalities, or complex institutions. For instance, some environments can be designed so that virtually any agent approximates efficient outcomes purely through market mechanisms (Gode and Sunder, 1993). Whether the setting is static or dynamic, competitive or cooperative, and whether payoffs are private or public, all inform the opportunities of the agent(s) to excel. Pinpointing these structural features can help predict whether the AI agent will exploit systematic patterns (e.g., in repeated games) or struggle to handle irregular but critical interactions (e.g., crisis negotiations).

Complexity of the Environment. Finally, real-world contexts are complex. They often involve multiple goods, numerous players, or intricate strategic layers that can degrade any agent's performance (Oprea, 2024; Enke, 2024). A moderately skilled human may thrive in a three-party negotiation but struggle with a ten-party scenario featuring high-dimensional preferences. An AI agent with ample computational resources may handle the combinatorial explosion of larger games more effectively, yet still fail to capture reputational nuances or unmodeled social norms that determine outcomes. Assessing how each agent's capabilities scale with the complexity of the environment is thus pivotal for anticipating real-world outcomes and deciding who—or what—should be assigned each task.

In sum, these six dimensions form a flexible, economically grounded lens for comparing different agents in diverse scenarios. No single factor dominates universally, and the details of each context—along with the interplay among preferences, strategies, and costs—will determine the ultimate impact of deploying AI or human agents. In the following sections, we apply this framework to a bargaining game, illustrating how it illuminates both the potential benefits and hidden pitfalls of substituting one agent type for another.

3 Empirical Context: A Bargaining Game

Figure 1 presents an overview of the bargaining game, which is played by three participants. Each participant begins with a fixed number of chips, with one chip color—green—serving as a numeraire. In all versions of the game, every player knows that green chips are worth \$0.50 to all participants. Other chip colors are valued differently by each player, with individual valuations ranging from \$0.10 to \$1.00 in increments of \$0.10. These non-green valuations remain private, creating information asymmetry and a need for players to infer each other's values through observed trading behavior rather than direct disclosures.

Figure 1: Bargaining game overview.



Notes: This provides a broad overview of the bargain game as played by all types of agents. The left column shows the setup and rules of the game. The center column shows a simplified version of gameplay. The right column provides an overview of the different kinds of agents.

At the start, each player is endowed with ten chips of each color. Depending on the complexity of the game, there may be just two chip colors (green and red), three (green, red, and blue), or four (green, red, blue, and purple). The main objective is for players to maximize the total value of their final chip holdings relative to their initial endowment. The game unfolds over three rounds, and in each round, every player takes exactly one turn to propose a trade. Turns are taken in the same order in each round, and that order is randomly determined at the beginning of the game.

The active player, on a given turn, proposes a trade, specifying an integer quantity of one chip color to offer and another chip color to receive. The rules prohibit offering more chips than one currently holds, and a player cannot swap like-for-like (e.g., offering red to receive red). It is allowed, however, to request more chips than opponents presently hold. The two non-proposers then simultaneously decide whether to accept or reject the proposed trade. If exactly one accepts, the transaction occurs with that player. If both accept, the system randomly selects which acceptor obtains the deal. If neither accepts, the proposal fails and no exchange takes place. Players know the total number of each chip held by all players at any given time, along with all trades that have been offered and accepted.

4 Problem Setup and Notation

Before exploring the empirical outcomes of the bargaining games, we provide a formal setup. Although no version of the game has a clear, simple game-theoretic solution, we can use linear programming methods to find optimal allocations as a reference to analyze our results. Furthermore, we employ measurement tools from game theory to make clear comparisons between the different types of agent outcomes.

4.1 Market Notation

Markets often face the challenge of allocating goods efficiently when participants have different valuations and initial endowments. While centralized mechanisms can achieve optimal allocations under certain conditions, many real-world markets operate through decentralized bilateral trades. This section develops a theoretical framework to analyze the efficiency implications of these different allocation mechanisms.

Consider a market $M = (I, G, v, a)$ with a set of agents I and goods G. Each agent i has a personal valuation v_{ig} for each good g. These valuations are static, independent across agents, and complete for each agent-good pair. All agents have an allocation a_{ig} of each good, which indicates the amount of good g allocated to agent i. Given these allocations, an agent's welfare s_i is the sum of their valuations weighted by their allocations: $s_i = \sum_{g \in G} v_{ig} a_{ig}$. The total market surplus s is thus the sum of all individual surpluses: $s = \sum_{i \in I} \sum_{g \in G} v_{ig} a_{ig}$.

4.2 Optimal Allocation

Agents can trade goods to improve their welfare. Depending on their valuations and their initial allocations a^0_{ig}, mutually beneficial trades may exist that increase the total surplus. We characterize one particularly interesting outcome of trading as follows.

Definition 1. (Optimal Allocation of Goods). *An optimal allocation of goods* is any set of allocations $\mathbf{A}^* = \{a^*_{ig} \mid i \in I, g \in G\}$ that solves:

$$\underset{\mathbf{A}}{\operatorname{argmax}} \sum_{i \in I} \sum_{g \in G} v_{ig} a_{ig}$$

subject to three constraints:

1. *The conservation of goods:*

$$\sum_{i \in I} a_{ig} = \sum_{i \in I} a^0_{ig}, \quad \forall\ g \in G$$

2. *Pareto improvement:*

$$\sum_{g \in G} v_{ig} a_{ig} \ \geq \ \sum_{g \in G} v_{ig} a^0_{ig}, \quad \forall\ i \in I$$

3. *Non-negativity:*

$$a_{ig} \geq 0 \quad \forall\ i \in I, g \in G$$

While this optimal allocation $\mathbf{A}^*$ may not be unique, the maximum total surplus $\mathbf{s}^*$ achieved at the optimum is unique.[1]

The interpretation is that in a trading market with a centralized allocation mechanism, the optimal allocation of goods maximizes total utility without leaving any agents worse off.[2] It is a Pareto-efficient optimal outcome for the market as a whole. Importantly, this is not the outcome of perfectly rational agents trading in their own self-interest. Myerson and Satterthwaite (1983) proved that it is impossible to design a trading mechanism that is both Pareto efficient and guarantees neither party risks a loss. It is simply the maximizer of s under some simple rationality conditions for individual trades.

While this centralized optimization provides a theoretical benchmark, it requires complete information about agents' valuations and faces trust issues in practice. Instead, humans often exchange goods through bilateral trading (where $|I| = 2$). Even when there are many agents ($|I| > 2$), a market—when decentralized—may solely rely on iterative pairwise trades between agents. To analyze how such decentralized trading affects market efficiency, we model the process through iterative bilateral matching.

Definition 2. (Iterative Bilateral Allocation of Goods). *An iterative bilateral allocation of goods* is any set of allocations $\mathbf{A}^{**} = \{a_{ig}^{**} \mid i \in I, g \in G\}$ that results from the following process:

1. Take as given or set an initial set of allocations $\mathbf{A^0}$
2. Repeat until no beneficial trades occur in a full round:
   (a) Generate a random permutation of all possible agent pairs $\{(i,j) \in I \times I \mid i < j\}$
   (b) For each pair (i,j) in this permutation:
      i. Solve for their optimal allocation of goods $\mathbf{A_{ij}^*}$ following Definition 1 with $I = \{i,j\}$
      ii. If $\mathbf{A_{ij}^*}$ differs from current allocations, update accordingly

Having formalized this decentralized trading process, we can characterize its key properties.

Theorem 1. *The iterative bilateral allocation process converges to a locally stable allocation, but this allocation:*

1. *May not be unique, as it depends on the sequence of bilateral trades.*
2. *Can be arbitrarily far from the optimal allocation $\mathbf{A}^*$ in terms of total surplus.*

Proof. Convergence follows from three observations. First, each bilateral trade strictly increases the total surplus. Second, total surplus is bounded above by the optimal allocation $\mathbf{A}^*$. Third, there are finitely many possible allocations given the conservation of goods constraint. These observations together imply that the process must terminate.

The non-uniqueness and suboptimality claims are proved constructively in Appendix A, where we provide examples where the sequence of bilateral trades determines the final outcome and where that outcome can achieve arbitrarily low fractions of the optimal total surplus.

[1] The non-uniqueness is obvious through the following example. Whenever two agents have identical valuations and initial allocations, as their final allocations can be interchanged without affecting the total surplus.

[2] This optimization problem can be solved efficiently using modern interior point methods. Its computational complexity is bounded by $O\big((|I||G|)^{3.5}\big)$.

4.3 Quantifying Suboptimal Performance

The *Price of Anarchy* and *Price of Stability* are concepts that capture the ratio of the socially optimal outcome of a game to the worst or best case equilibrium. In this work, we introduce a similar concept, the *Price of Agency*, which capture performance by a population of agents (such as human agents or autonomous agents) relative to a theoretical optimum.[3]

For a given market $M = (I, G, v, a)$, let $\mathbf{w} = \sum_{i \in I} \sum_{g \in G} v_{ig} a_{ig}$ be the social welfare (i.e. aggregate utility) accrued by all agents I. Let $\mathbf{s} = \mathbf{w}_{final} - \mathbf{w}_{initial}$ be the surplus social welfare at the end of the game. For each market, let $\mathbf{s}_g$ be the empirical surplus for a particular collection of agents engaged in trading in that market. g. Let $\mathbf{s}^*$ be the optimal social welfare.

We now define the price of agency ($PoAg$) as follows:

$$PoAg = \frac{\mathbf{s}_g}{\mathbf{s}^*}$$

$PoAg$ thus capture the proportion of the optimal surplus accrued by agent participants, and can be utilized in capturing differences in outcomes between different types of agents. For example, we can compute the *human-LLM alignment gap* as $\Delta P = PoAg_H - PoAg_L$, where $PoAg_H$ is the price of agency for human participants, and $PoAg_L$ is the price of agency for a population of LLM-based agents.

4.4 Dynamics of Myopically Rational Play

While we can quantify the globally or iteratively bilateral optimal allocation of goods, neither fully captures the dynamics of this market. For example, players may be misrepresent their utilities to gain an advantage in future rounds of proposals. The game tree is too large to solve exactly with existing solvers, but we can explore the concept of *myopic rationality* to better understand agents might make short term decisions. A *proposal* is myopically rational if the proposer gains utility if their trade is accepted. An *acceptance decision* is myopically rational if the acceptor gains utility if they are chosen to transact.

To understand the dynamics of myopically rational play, we begin by considering the simplified scenario of a two-chip negotiation game. Let us consider a game state where player i's chip allocations are represented by the pair (a_i, b_i) of green and red chips respectively, and the corresponding values for these chips are $(0.5, v_i)$. A valid proposed trade can be represented as a pair (x, y), where one element is positive (chips received) and the other is negative (chips given), and the proposed quantities satisfy allocation constraints. So, if player i proposes a trade (x, y) with $x < 0$, then the they are giving x green chips in exchange for y red chips. This is valid as long as $-a_i < x < 0$ and $0 < y < \max_{j, j \neq i}(b_j)$. If this trade is accepted, player i's new allocation becomes $(a_i + x, b_i + y)$, while the receiving player j's allocation becomes $(a_j - x, b_j - y)$.

A trade is deemed myopically rational if it results in a utility increase for both the proposing player i and the receiving player j. This condition is satisfied if and only if $0.5x + v_i y > 0$ for player i and $0.5x + v_j y < 0$ for player j. The expected number of myopically rational trades for player i can be calculated by summing the probabilities that a given trade (x, y) is myopically rational when proposed to any potential opponent.

[3]Note that while the *Price of Anarchy* is traditionally defined as a ratio of inefficiency with $PoA \geq 1$, our *Price of Agency* serves as an efficiency benchmarks ≤ 1, where values closer to 1 indicate more optimal performance.

We employ linear programming and stochastic analysis methods to compute the optimal s^* and expected values in the context of our negotiation game. In the initial allocation of the two-chip variation, the expected number of myopically rational trades is calculated to be 37.1.

We use this analytical framework to demonstrate the increasing action space and expect number of trades for increasingly complex variations of the game. In the three-chip game, while players remain restricted to exchanging only two types of chips at a time, the number of possible chip pairings increases to $\binom{3}{2} = 3$. This expansion in potential trading pairs consequently enlarges the set of myopically rational trades. For example, in the starting allocation of the three-chip game, the expected number of myopically rational trades rises to 120.7. Extending this analysis to the four-chip game, the number of possible chip pairings grows to $\binom{4}{2} = 6$, further expanding the space of myopically rational trades. In the initial allocation of the four-chip game, the expected number of myopically rational trades reaches 250.7.[4]

5 Agents

Section 4 formalized theoretical benchmarks for optimal and decentralized market efficiency. However, in practice, negotiations operate under incomplete information and varying level of rationality. Before comparing empirical results on the negotiation game described in Section 3 with human and LLM-based agents, we introduce *noise-driven* agents in this section – rule-based solvers that provide benchmarks for behaviors ranging from random baselines to approximately rational behavior. Later, we also introduce *Bayesian-learning* agents that incorporate information from previous rounds to make offers aimed at maximizing expected utility.

5.1 Noise-Driven "Zero-Intelligence" Agents

Inspired by the early work of Gode and Sunder (1993), we consider markets populated by "zero-intelligence" computer traders that submit random bids subject to a budget constraint. These studies suggest that 'market' structures alone can lead to near-equilibrium outcomes, even in the absence of strategic reasoning. As we have a similar chip market, trade proposer, and receiver structure in our game, we can implement zero-intelligence agents with heuristics as follows:

1. **Random Strategy:** The proposer makes random proposals by selecting a random color to offer and a different random color to receive. The quantities of chips exchanged are also randomized, subject to two constraints: (i) the proposer cannot offer more chips than they currently hold, and (ii) the requested amount cannot exceed a predefined cap (e.g., the initial endowment). The receiver makes independent, random accept/reject decisions.

2. **Single Rational Strategy:** The proposer follows the same random strategy as above, but the receiver only accepts trades that are myopically rational, i.e. provide a positive utility gain.

3. **Double Rational Strategy:** The proposer generates trades that are myopically rational (i.e. improve their own utility), while the receiver continues to accept only beneficial trades.

[4]Note that we don't get an exact 3 or 6 multiple of myopically rational trades due to the fact that one chip has a fixed valuation whereas the others are variable.

Notably, these agents do not learn from prior interactions; each trade decision is independent of previous rounds. We include these random-noise baselines to explore the limits of purely competitive strategies and assess how market structure alone affects overall surplus.

5.2 Bayesian-Learning Agent

Next, we formulate more sophisticated agents that *do* learn: Bayesian-Learning agents. We formalize their strategy as follows: each proposer i forms a joint belief about the valuations of all other players, $\mathbf{v}_{-i}$. When proposer i offers a trade $(\mathbf{x}, \mathbf{y})$ —giving up $\mathbf{x}$ chips in exchange for $\mathbf{y}$—they choose $(\mathbf{x}, \mathbf{y})$ to maximize their expected payoff:

$$\max_{(\mathbf{x},\mathbf{y})} \sum_{\mathbf{v}_{-i}} \mathbf{1}\Big\{\exists\, j \neq i : \text{accept}\big(v_j; \mathbf{x}, \mathbf{y}\big)\Big\} \;\times\; \Big(u_i\big(h_i - \mathbf{x} + \mathbf{y}\big) - u_i\big(h_i\big)\Big) \;\times\; B_i\big(\mathbf{v}_{-i}\big), \tag{1}$$

where accept $\big(v_j; \mathbf{x}, \mathbf{y}\big)$ indicates that player j with valuation v_j would accept, and $B_i(\mathbf{v}_{-i})$ is proposer i's belief (probability) that the other players' valuations are $\mathbf{v}_{-i}$. Acceptance is *myopically rational*, i.e. a receiver with valuation accepts if and only if the expected utility is higher:

$$u_j(h_j - \mathbf{y} + \mathbf{x};\, v_j) \;>\; u_j(h_j;\, v_j). \tag{2}$$

All players update their priors after observing accept/reject decisions. A player discards valuation states $\{v_j\}$ that contradict the observed event and re-normalizes the remaining probabilities. When the trade is accepted (*Acceptance*), the proposer keeps only valuations that would accept in its distribution over the chosen receiver; the receiver keeps only valuations that would accept in its distribution over the proposer. Bystanders update about both the proposer and the receiver accordingly. When there is no trade (*Rejection or No Receiver*), all players discard valuations that conflict with the observed non-acceptance. This will provide a strong benchmark for how a strategic player behaves under incomplete information. We will use these agents as our theoretical baseline in computing the *Price of Agency* for the human and LLM-agent populations.

5.3 Human Player Experiments

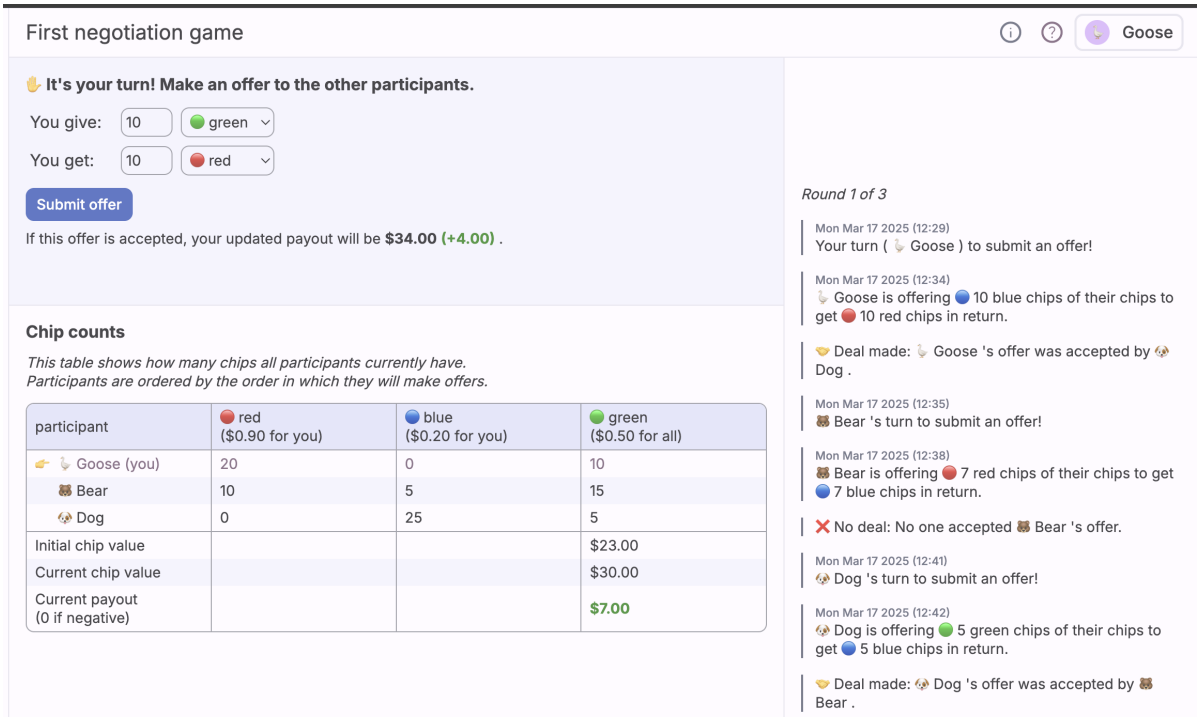
Implementation. We recruited participants on Prolific[5] and redirected them to a custom bargaining game implementation hosted on *Deliberate Lab* (Tsai et al., 2024), a custom platform for facilitating live, multi-participant experiments. Participants received a base payment for their time (∼30 minutes), which factored in time spent reading instructions, completing comprehension checks, and waiting for other players to join their live session. Each participant played two games, each with different chip valuations and other player profiles. Participants were able to convert the surplus on their final chip holdings into a bonus monetary payout, incentivizing strategic play.

For the 2-chip, 3-chip, and 4-chip game variations, we analyzed data from 24 full groups of three participants each, totaling $N = 3 \times 24 \times 3 \times 2 = 432$ effective participants over 144 games.[6]

[5] www.prolific.com.

[6] Ordering and learning effects between game 1 and 2 were minimal, so we evaluate all games independently for this work.

Figure 2: The game interface for human experiments.



This shows a participant's game state. The top left indicates that they are currently the Proposer, offering an interface to propose an offer. A table below displays all players' chip volumes. To the right, the public ledger shows a history of game rounds, turns, and trade statuses.

5.4 LLM-Agent Experiments

We developed simulations of all of the human games with LLM-powered agents serving as individual players. The simulation closely mirrors the experience of human players: At each turn, an LLM agent receives the same information as its human player counterpart, including their own chip values, each player's chip holdings, and the history of previous trades. In each round, one agent proposes a trade, and the other two respond with *Yes* or *No*. The specific prompts used can be found in Appendix E. First, we provided the same prompt shown to our Prolific participants, specifying each player's chip valuations. Next, we defined an output format to ensure the LLM's responses would be compatible with the game's requirements. To maintain this specific format, we used a chain-of-thought approach.

We prioritized a straightforward approach to building the LLM agent, focusing on how well it could perform with minimal optimization. We began with an out-of-box agent that relied on simple prompting, adapted from the human players' instructions, and parsed the output for trade proposals and decisions. To explore the value-add of minimal refinement, we also developed a refined agent that approximated more deliberate ("Type 2") thinking through a two-step process (Kahneman et al., 1991). The LLM first generated three possible trade proposals, then selected the best one in a second call.

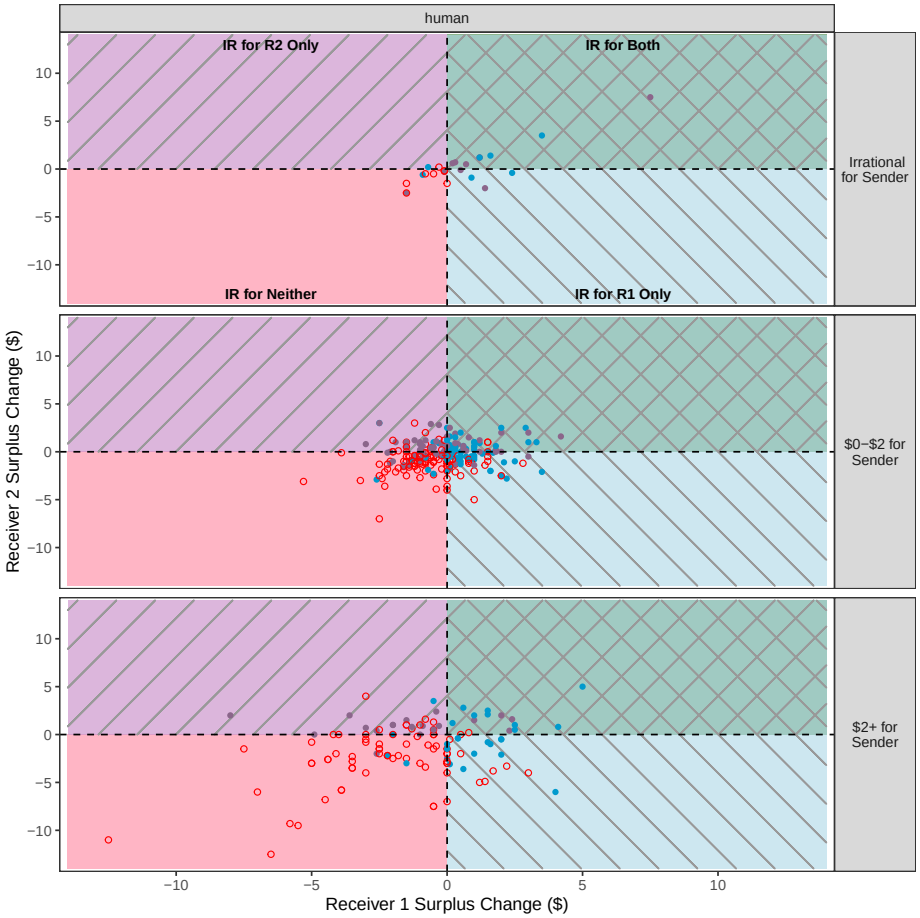
For our experiments, we used a state-of-the-art language model – `Gemini-2.0-flash-thinking-exp` (Team et al., 2024) with a temperature setting of 0.5 to encourage more variable trade proposals. We also open-sourced an interactive Notebook for interested readers to try different LLMs. The repository will be open-sourced once the paper is published.

6 Game-Level Empirical Evaluation

In this section, we present the empirical results from gameplay with human participants, followed by simulations with rule-based agents, statistical agents, and LLM-simulated agents, all using the same seeded chip valuations as the human participants. This setup enables a direct comparison of decision-making behaviors and strategies across both groups.

6.1 Human Experiment Results

Figure 3: (3-chip game) Player surplus trade visualization



Receiver 1 (R1) and receiver 2 (R2) are the other players who will be the proposer in the next two turns, respectively. The x-axis shows the surplus change for R1 and the y-axis shows the surplus change for R2, revealing the myopic rationality of trades. Quadrant 1 shows positive surplus for R1 and R2, and Q3 shows negative surplus for both. The three rows correspond to different surplus outcomes for the proposer: the top row shows myopically irrational trades (negative surplus for the Proposer), while the middle and bottom rows show rational trades with small-to-medium and large surplus gains, respectively. Each point represents a proposed trade, with open red circles indicating rejected trades and filled points showing accepted trades (blue for R1 as the executed trade partner, purple for R2).

Surplus improvement (over rounds, per agent). Fig 3 displays the distribution of player surplus changes from trades in the 3-chip game; equivalent 2-chip and 4-chip plots appear in Appendix B).

Several patterns emerge from this visualization. First, the top row contains noticeably fewer points than the bottom two rows, indicating that Proposers rarely offer trades that are irrational from their perspective. Second, the majority of proposed trades in the red quadrant (negative surplus for both

receivers) are rejected. Third, accepted trades with Receiver 1 (blue points) generally appear to the right of zero on the x-axis, while accepted trades with Receiver 2 (purple points) typically appear above zero on the y-axis. Overall, Fig 3 demonstrates that players predominantly propose and accept trades yielding positive surplus from their perspective, though a nontrivial number of exceptions suggest some players make decisions that deviate from strict myopic rationality. We discussion potential reasons for these behaviors in Section 7.

Average time spent. From Table 1, we observe that human participants spend the longest average time per turn in the 3-chip variation (58.94 ± 40.55 seconds). While participants spent less time in the 2-chip and 4-chip variations (48.44 ± 33.83 and 48.21 ± 34.04 respectively), these differences are not statistically significant (one-way ANOVA, $F(2, 141) = 0.90, p = 0.41$). These turn durations account for the entire interval from the moment the proposer becomes active to when the transaction is either accepted or unanimously declined by the other players.

Notably, the standard deviations are substantial – often approaching the corresponding mean. Such wide dispersion underscores the varying cognitive and strategic approaches humans bring to each turn. Interestingly, despite the increasing complexity from the 2- to 4- chip game variations shown in Section 4.4, the time spent per turn remains consistent. This may be due to cognitive limitations (such as fatigue), bounded effort (such as satisficing strategies), or diminishing marginal complexity of the game.

Summary Statistics of Game Quantitative Metrics by Game Variation			
Game Variation	**Player Type**	**Scaled Final Surplus**	**Time Spent Per Turn(s)**
2 chip	*Human*	0.607 (0.387)	48.44 (33.83)
	Gemini	0.602 (0.263)	14.92 (0.43)
	Gemini-Refined	0.620 (0.351)	20.49 (0.40)
3 chip	*Human*	0.590 (0.257)	58.94 (40.55)
	Gemini	0.512 (0.194)	14.60 (0.36)
	Gemini-Refined	0.514 (0.255)	20.56 (0.39)
4 chip	*Human*	0.575 (0.179)	48.21 (34.04)
	Gemini	0.375 (0.125)	15.42 (0.40)
	Gemini-Refined	0.478 (0.149)	23.40 (0.85)

Table 1: Statistics of quantitative metrics in Human and LLM agents.

Quantitative self-assessment. Following the end of the game, participants completed a post-game survey (questions in Appendix D.2), which included ranking Likert scale questions on a scale from 1 to 10 to assess their self-perceived *competitiveness* (ranging from highly collaborative to highly competitive), *confidence* in their performance (low to high confidence), *satisfaction*, and amount of cognitive *exertion*. Table 2 shows summary statistics and correlation analysis across the different game variations.

Subtable 1 of the table shows that the self-assessment scores remained relatively stable across game variations, with no statistically significant differences observed (ANOVA, $p > 0.1$ for all variables). This finding for self-reported *exertion*, in particular, is consistent with the previous findings that participants spent similar amounts of time in each variation, as these two factors are likely correlated[7]. Participants

[7]It is said that "time is effort!"

1. Summary statistics of self-assessment variables				
Game variation	Competitive	Certainty	Satisfaction	Exertion
2 chip	5.57 (0.34)	5.57 (0.33)	6.68 (0.31)	7.11 (0.24)
3 chip	5.32 (0.34)	5.92 (0.33)	7.07 (0.30)	7.10 (0.25)
4 chip	5.75 (0.34)	6.04 (0.33)	7.46 (0.26)	7.46 (0.19)
2. Correlation matrix (Pearson coefficients)				
Competitiveness	1.0	0.28	0.19	0.20
Certainty	0.28	1.0	0.60	0.11
Satisfaction	0.19	0.60	1.0	0.13
Exertion	0.20	0.11	0.13	1.0
3. Correlation with payout (Pearson coefficients, ***p <0.01)				
Payout	0.02	0.02	**0.20*** **	0.08

Table 2: Statistics and correlations of quantitative self-assessment measures.

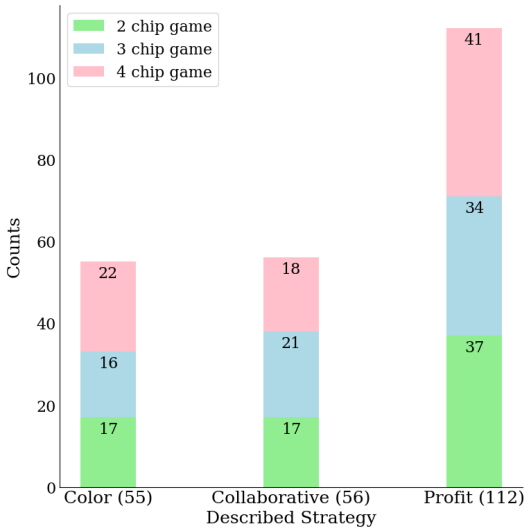


Figure 4: Categorized self-described strategies.

also feel similarly certain and confident in their performance across chip variations, despite empirically capturing less of the theoretical surplus from 2- chip to 4- chip; perhaps due to a stable subjective experience in the game, a consistency in decision heuristics across variations, or cognitive errors such as the Kruger and Dunning (1999) effect.

Subtables 2 and 3 report Pearson correlation coefficients across the variables and the participant's payout amount. Participants who felt more confident in their trading decisions also reported greater satisfaction with their outcomes ($r = 60$). When evaluating correlation with payout, only a participant's satisfaction with their performance in the game was correlated ($r = 20, p < 0.01$).

Qualitative self-assessment. The survey also included a freeform section where participants described their strategies and relevant backgrounds. We conducted a thematic analysis (Braun and Clarke, 2006) of each participant's text responses and identified three main strategic themes: profit maximization, color optimization, and collaborative play (Fig 4). Each participant's response was coded into one or more of these categories, though multiple classifications were uncommon.

Profit maximization strategies focused on securing the highest possible personal gain, but without necessarily prioritizing a specific color. Players in this category aimed to take only advantageous trades, avoiding offers that could put them at a loss or provide insufficient returns. Many described their approach in terms of cautious risk management or aggressive early moves to optimize value, for example "I only accepted positive or neutral bets," "My approach was to make an initial large move and then try to offload the lowest-value chips," or "I only took offers that ensured mutual gain and avoided risky trades."

Color optimization strategies framed the approach to maximizing value in terms of specific chip colors. Rather than aiming for general gains, these participants prioritized trading for their highest-value chip or offloading their least-valuable chip. For example, one participant stated, "My strategy was to prioritize gaining higher-value chips, particularly focusing on the blue chips as they were valued higher than both red and green chips. I aimed to trade lower-value chips for higher-value ones while avoiding unnecessary trades of red chips unless beneficial." Interestingly, participants did not mention strategies regarding

the *green chip as a numeraire*, suggesting that they may have taken its stability for granted rather than actively incorporating it into their decision-making.

Collaborative play strategies involved taking others' positions into account, either by fostering mutually beneficial trades or by acting aggressively to gain an advantage. Many described explicitly cooperative mindsets, such as "I aimed to be fair and reasonable with my trades," or "I kept my offers low to encourage cautious players to engage." Others took a more strategic view of early interaction, such as "I traded aggressively early on to secure chips before others could deplete the supply."

The classification of the participants' self-reported strategies was not correlated with their performance in the game. Additionally, there was a lack of clear pattern in participants' self-reported previous expertise ("What backgrounds or experience may have influenced your performance in the game?") — some pointed to negotiation, math, or strategic games as useful, but most were uncertain about what skills translated to success. These findings suggest that players navigated the game with limited clarity on what strategies or experiences would give them an edge.

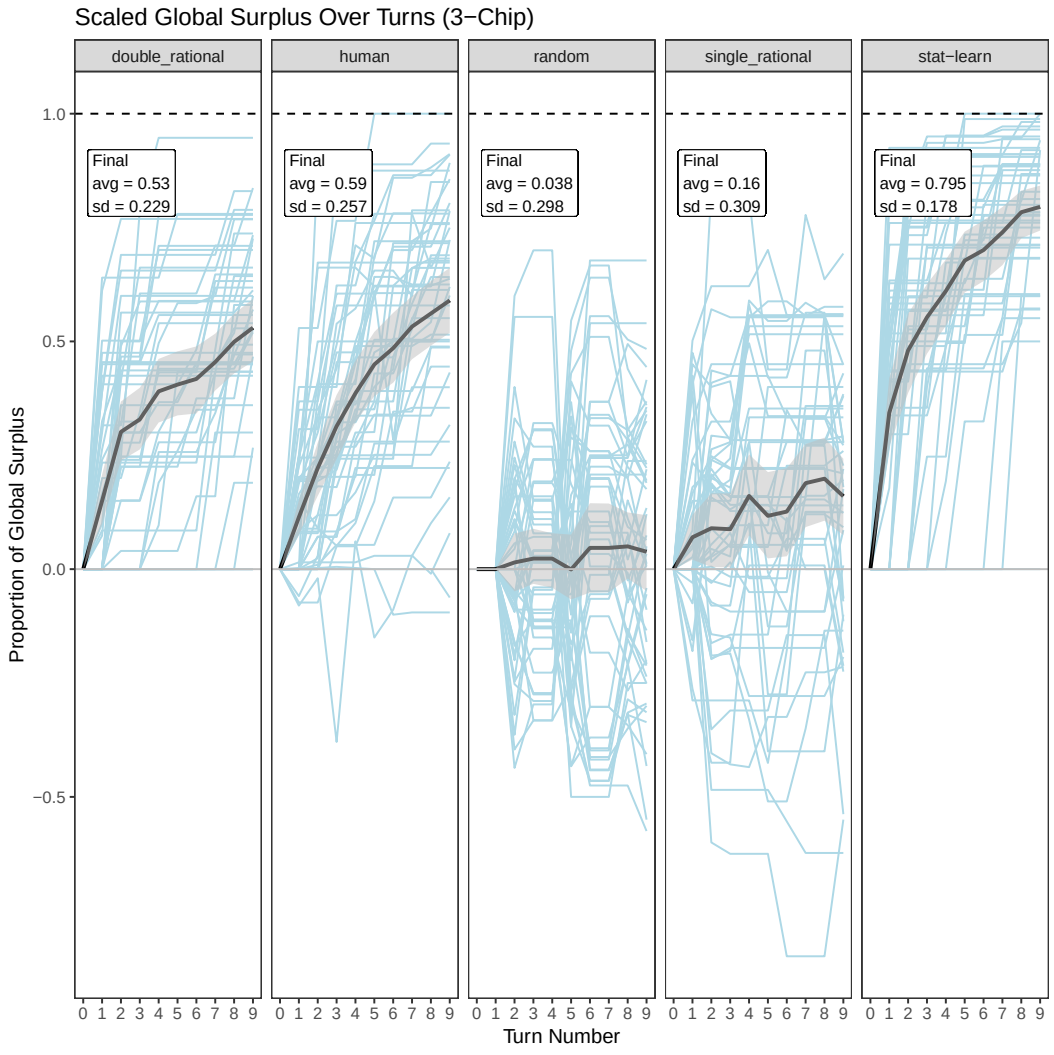
6.2 Rule-based and Statistical Agent Simulation Results

Fig 5 presents the trading trajectory data for all rule-based agents in the 3-chip game, highlighting distinct performance patterns across varying levels of rationality. In the following analysis, we focus on the 3-chip game. All statements hold for the 2-chip and 4-chip cases, with corresponding results and visualizations in Appendix C.

Zero-intelligence agent performance. The Random Strategy performs poorly, yielding near-zero or even negative surplus gains. Since both proposers and receivers make decisions without considering any utility constraints, random exchanges result in null overall gains. Introducing a rational constraint on the receiver side significantly improves performance. Under the Single Rational Strategy, where the receiver accepts only beneficial trades, we observe a net increase in surplus ($p < 0.001$). However, the overall efficiency remains significantly lower than human performance ($p < 0.001$), as the proposer continues to generate random offers without surplus consideration. When rationality is enforced on both sides in the Double Rational Strategy, the performance of zero-intelligence agents becomes comparable to that of human participants. A two-sided t-test indicates no significant difference between the surplus gains achieved by these agents and human bargainers. This suggests that basic rationality constraints, even without sophisticated learning or strategic behavior, are sufficient to approximate human-like performance in structured bargaining settings.

Bayesian-learning agent performance. The Bayesian Learning Agent demonstrates a significant improvement over all other agent types, including human negotiators. By updating beliefs based on observed acceptances and rejections, Bayesian agents adapt their proposals to maximize expected payoffs over time. A two-sided t-test confirms a statistically significant difference between Bayesian agents and all other agents, as shown in Fig 5. The Bayesian agent serves as an "approximately good" solution for the trading problem under incomplete information to provide a computational benchmark.

Figure 5: (3-chip variation) Surplus trajectories for noise-driven agents.



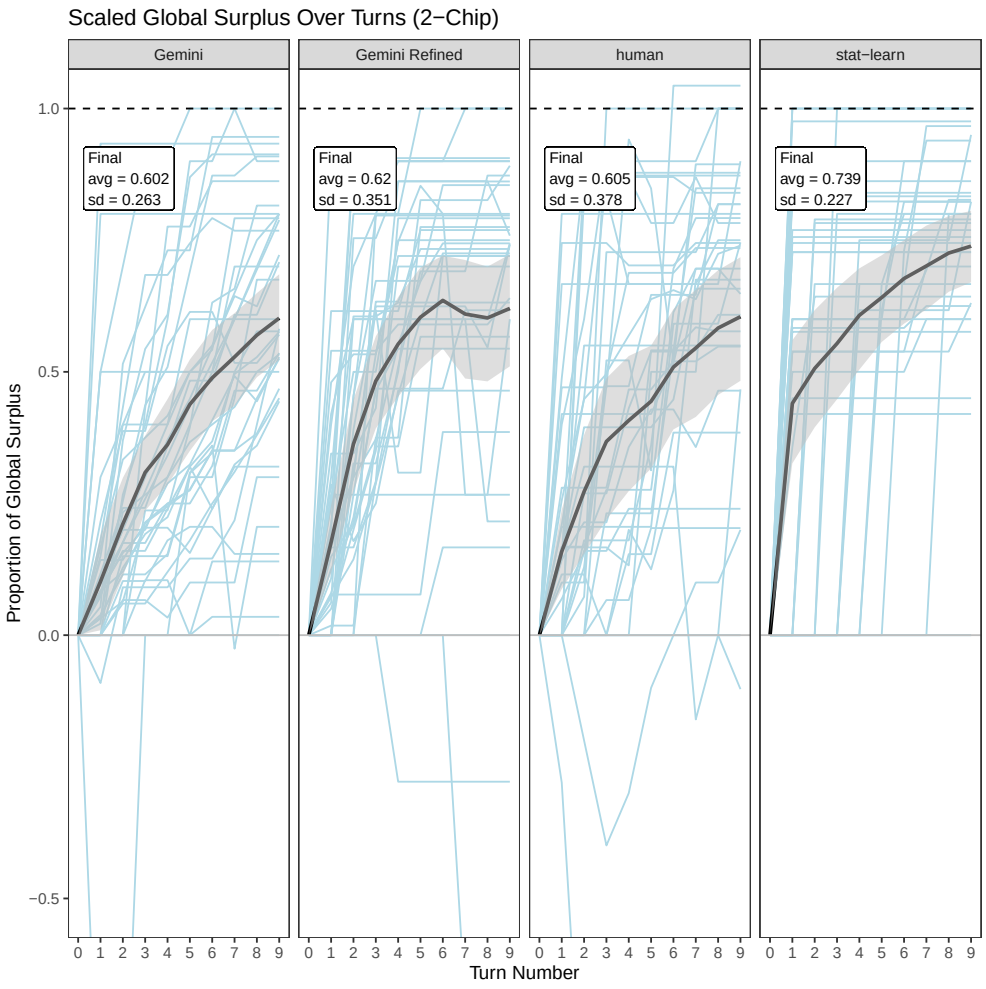
.

6.3 LLM Results

Surplus improvement. Fig 6 displays the trajectories of player surplus changes from trades in the 2-chip game (equivalent 3-chip and 4-chip plots appear in Appendix C). Any downward "dips" in the surplus trajectory signal a net-negative trade outcome, where players exchange chips in a way that decreases total surplus. One-way ANOVA tests (Table 1) reveal no significant difference in surplus improvement between humans and both LLMs in the 2-chip game ($p = 0.14$).

However, in the 3-chip variation, humans slightly but significantly outperformed both LLM- agents by a small margin ($p = 0.002^{**}$), with no difference between the LLM models. In the 4-chip variation, humans outperformed both Gemini agents significantly ($p = 1.05e-22^{***}$), and Gemini Refined significantly outperformed out-of-the-box Gemini.

Figure 6: (2-chip variation) Surplus trajectories for human, LLM, and Bayesian-learning agent simulations.



Notes: We denoted Bayesian-learning agent as 'stat-learn' in the figure.

We found that the Bayesian agent outperforms all human and LLM-based agents across all variations of the game, and captured a consistent amount of surplus across game variations. This may be because the Bayesian agent does not suffer from cognitive biases (exhibited by humans and potentially learned by LLMs), is resistant to any deceptive bargaining tactics, or has no cognitive limitations or hallucinations that can hinder its precise probability calculations. This could also suggest a difference in true objective utilities; for example, the human participants reported wanting to prioritize fairness and collaboration, or getting enjoyment from exploring the game space through exploratory actions.

Total time spent. Table 1 shows that both LLM agents complete each turn significantly faster than human participants. On average, the basic Gemini agent spent only about one-third the amount of clock time computing their responses as long as humans did, and the refined Gemini agent reduces the time per turn by about half relative to human players. Both LLM agents also exhibit substantially lower variance in turn durations, reflecting more consistent reasoning processes relative to humans. The scale of this time-saving will likely only increase as LLMs become more computationally efficient.

6.4 The Human-LLM Alignment Gap

We use the *Price of Agency* defined in Section 4.3 to compare the performance of human and LLM players. Fig 7 plots the distribution of the price of human agency ($PoAg_H$), the price of LLM agency ($PoAg_L$), and alignment gap across the cohorts of the 4 chip game with the out-of-the-box LLM agent. The $PoAg_H$ has a wider distribution with a higher mean than the $PoAg_L$, with an average alignment gap of $0.150^{***}(p = 1.2e{-}6)$. This reflects the fact that humans are less consistent than the LLM agents, but generally capture a higher fraction of the optimal surplus.

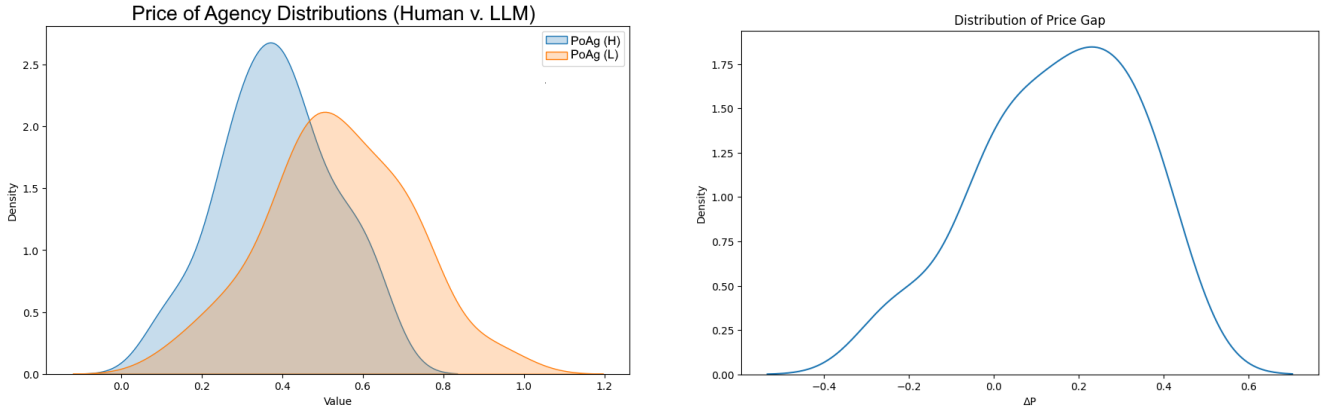


Figure 7: Comparing human and out-of-box agent performance in 3 chip game

The alignment gap closes for the more advanced refined agent as seen in Fig 8. While humans still slightly outperform the refined agent with an average alignment gap of 0.013, the distributions of $PoAg_H$ and $PoAg_L$ are highly overlapping and the difference is not statistically significant ($p = 0.61$).

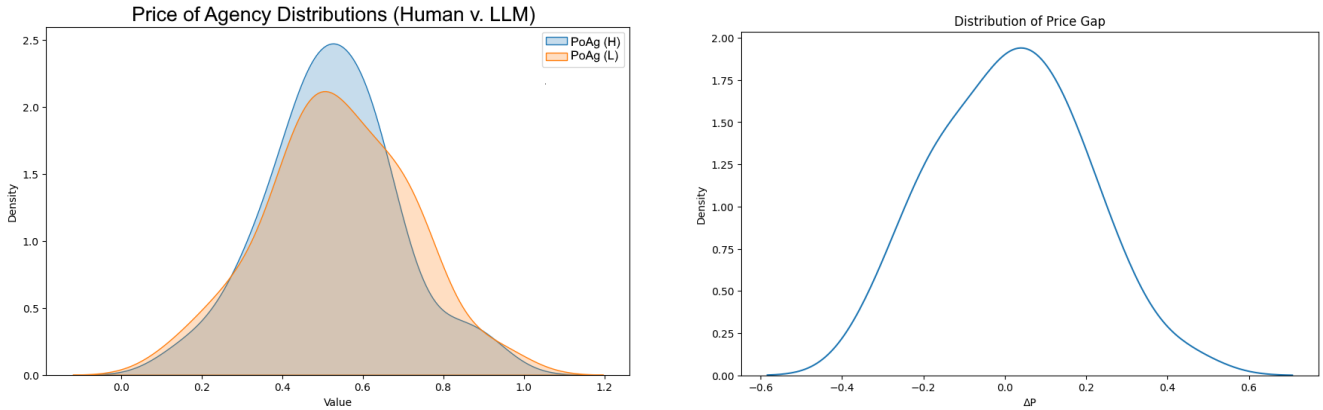


Figure 8: Refining agent closes the alignment gap.

Fig 9 extends this analysis across the 2, 3, and 4-chip scenarios. As the number of distinct chips increases, performance for all categories of agents generally falls, reflecting the increasing complexity of the game. All agents perform similarly on the simple 2-chip game. We see that the performance gap between the two styles of LLM agents is small for both the 2- and 3-chip games, and only diverges significantly in the 4- chip game. It may be that the two-stage reasoning of the refining agent is necessary at this level of complexity to properly explore the larger opportunity space.

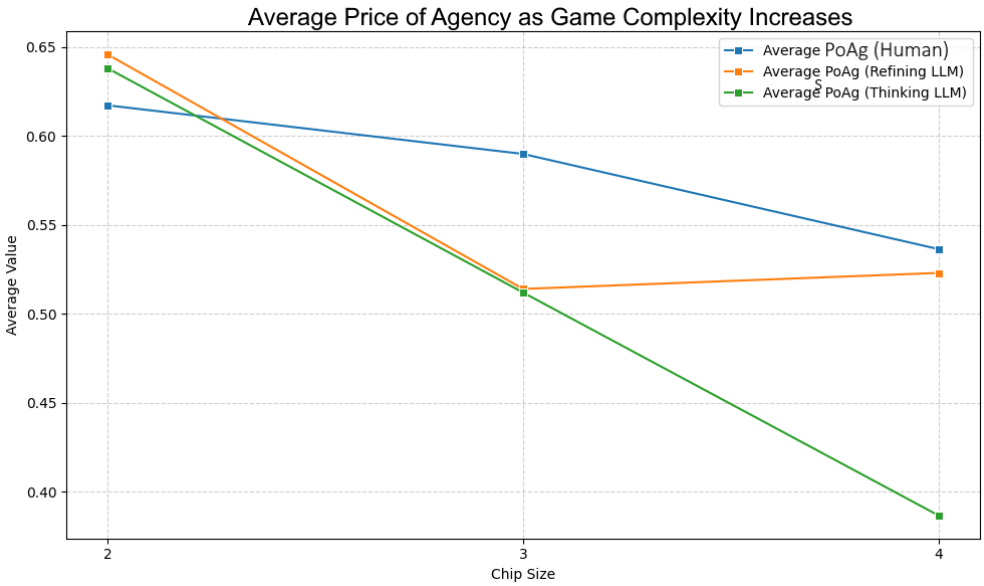


Figure 9: Performance across chip sizes.

7 Trading behavior analysis

In our results, we've observed differences in global surplus values across human and LLM populations. Even when surplus values are similar, such as between Gemini and Gemini Refined in Fig. 6, they may result from different decision-making processes.

To explore these differences, we conduct an analysis of trading behaviors that can emerge across humans and LLMs, broadly categorizing behaviors as potentially rational, strategic, or adaptive (although many behaviors can span multiple categories). While some behaviors reflect deliberate strategy, others may arise from judgment errors or execution flaws. Our goal is to distinguish intentional strategies from mistakes, identifying key points of divergence and underlying complexities. We focus on the 3-chip scenario for its generalizability, highlighting notable differences in the 2-chip and 4-chip cases where relevant.

7.1 Rational behavior

Rational trades increase a player's total surplus, while irrational trades decrease it. We classify irrational behaviors into two key types:

Net over- or under- trading. This occurs when a player's final chip value is lower than their starting position by the end of the game. We find that this happens rarely across all populations and game variations. Possible reasons for net-negative human trades include miscalculations, risk-taking, information asymmetry, and psychological biases. While LLM agents are explicitly instructed to trade rationally, their performance on this metric is indistinguishable from humans. This suggests that either LLMs exhibit similar rationality levels to humans or that their decision-making process introduces comparable errors, despite the rational trading instructions.

Direct trading losses. These occur when a player behaves in a myopically irrational manner. Some direct trading losses may be (i) market making; the player takes a loss now to deflate the perceived value of the chip and enable a more valuable trade later. We identify this behavior when a player later trades the lost chip and recovers the loss at a better per-value ratio. Players may also demonstrate (ii) attempted market making when they propose or accept a negative trade, then later bid on the chip at a higher valuation but are not chosen to clear the trade due to the random selection mechanism. In direct loss scenarios where neither successful nor attempted market making is observed, (iii) non-strategic errors may be the root cause – humans may miscalculate values, rush into trades, or misjudge market conditions. LLMs, despite being explicitly instructed in their system prompts to "trade rationally for positive values" (Appendix E), do not always do this, demonstrating potentially unpredictable behavior.

Direct trading losses are relatively rare, occurring in no more than 4% of cases across all agents and chip types. Negative acceptances are more common than negative proposals; in the user interface, it is easier to mistakenly accept a negative trade than to intentionally propose one. However, this pattern is not unique to humans; across all agent types, over half of direct trading losses are seemingly non-strategic. While the LLMs' loss categorization distributions are not significantly different, humans significantly employ more attempted market making relative to LLMs, and a higher proportion of their direct trading losses are strategic. As humans outperformed LLMs in the 3-chip scenario, it suggests that humans demonstrate more strategic flexibility in recognizing when to take a short-term loss for a long-term game. This could be due to a lack of stateful reasoning or optimization constraints by the LLM (for example, prompts in Appendix E explicitly incentivize myopic gains).

7.2 Strategic and adaptive behavior

Next, we discuss actions involving strategic decision-making, where players engage in more behaviors beyond simple surplus maximization, including adaptive behavior, where players appear to update their behaviors based on previous actions taken in the game.

Uncleared trade offers. A proposer can make an offer that goes unaccepted by any other player. Since proposers cannot opt out of making an offer, they may strategically set infeasible trading prices to (i) signal disinterest or satisfaction, effectively opting out of trading while still fulfilling the requirement to make an offer. Or, the uncleared trade offer may be due to (ii) overvaluation, where a proposer miscalculates market demand and overestimates how much others are willing to pay for a chip. The overvaluation may not be a mistake; the proposer may be attempting to manipulate the price of the chip.

Approximately half of all offers made across all agents and games, with no statistical difference between humans and LLM agents ($\chi^2 = 1.78, p = 0.87$). This high rejection rate is largely driven by overvaluation, which accounted for over 90% of uncleared offers. Rarely was it the case that disinterest was signaled through setting an infeasible trade, and this handful of cases was largely executed by LLMs.

Anticipating trade volume. We've considered negative-surplus offers and uncleared offers. Next, we consider cleared, positive-surplus offers, from the perspectives of proposers, decliners (those who decline the offer), and receivers (both those who want to accept the offer, and those who are chosen to accept the offer).

In Table 3 Group 1, we count the instances when a proposer proposes or a receiver accepts a positive, no-change, or negative surplus offer (from their perspective). We also perform χ^2 tests to compare these distributions between humans and the out-of-the-box Gemini agent. Humans were significantly more likely to both propose and receive non-positive trades, relative to LLMs. This may be due to cognitive biases, goodwill, or strategic attempts to shift their holdings towards chips they perceive as being more valuable in the market.

	Agent type	
	Human	*LLM*
1. Surplus		
Proposer*** ($\chi^2 = 13.29, p = 0.001$)		
Positive	191	214
No change	17	2
Negative	10	8
Receiver ($\chi^2 = 5, p = 0.08$)		
Positive	154	178
No change	44	30
Negative	21	16
2. Trade strategy		
Proposers*** ($\chi^2 = 9.64, p = 0.002$)		
Optimal	148	118
Suboptimal	71	106
Acceptors ($\chi^2 = 0.76, p = 0.3841$)		
Optimal	140	285
Suboptimal	41	67
Decliners*** ($\chi^2 = 22.89, p = 1.7e-6$)		
Optimal	35	31
Suboptimal	7	54

Table 3: Table of categorization counts for surplus and strategies in positive-surplus trading scenarios. χ^2 tests performed to evaluate differences between humans' and LLMs' categorization distributions.

We further break down the myopically rational play to see whether these behaviors were optimally rational with perfect information about the game state. A proposer's proposal is optimal if the trade clears and no better offer for the same chip occurs later. However, a proposer could undervalue the chip – say that their proposal clears at a low value. Later, another player offers to trade the same chip at a higher price. If this undervaluation prevents the original proposer from participating in the better trade, the original offer was suboptimal. However, if they still have resources to accept the better trade, their initial conservative offer may reflect market gauging rather than an error. Humans proposed the optimal trade significantly more than LLMs; this is likely attributed to trade volume (discussed later in Section 10), as humans tend to make more aggressive offers.

Decliners may reject surplus-increasing trades in hopes of a better opportunity later. If a superior trade later arises and the decliner secures it, this strategy is optimal. However, optimal play comes with an element of luck; however, since multiple players can attempt to accept the same trade, the random allocation mechanism means that players are not guaranteed to receive their preferred offer. If the decliner either bids for the better trade at a later step, or no better trade occurs, the strategy is labeled as suboptimal. With humans declining optimally the majority of the time, their increased surplus may be in part attributed to being strategically patient. They appear to decline only when they have a strong expectation of securing a better deal. LLMs, on the other hand, decline significantly more often and with less optimality.

Acceptors may make the optimal choice by selecting the best available trade; that is, no better offer for the chip is made later by another player. However, they may also suffer an opportunity cost if they accept a trade too soon and are later unable to accepting a better trade for the same chip due to depleted resources. Importantly, only the chosen acceptor faces this opportunity cost—potential acceptors who were not randomly selected remain eligible for future trades. The same decision in this scenario can result in different outcomes due to randomness; players' levels of risk tolerance may play a factor in these decisions. While the distribution of acceptance behavior remains similar across agent types, the LLM-played games display more opportunities to accept more positive-surplus trades. This, again, can be attributed to the more aggressive amounts of chips being traded in a single transaction in the human games, which lock players out of future opportunities once they exhaust their high-value chips.

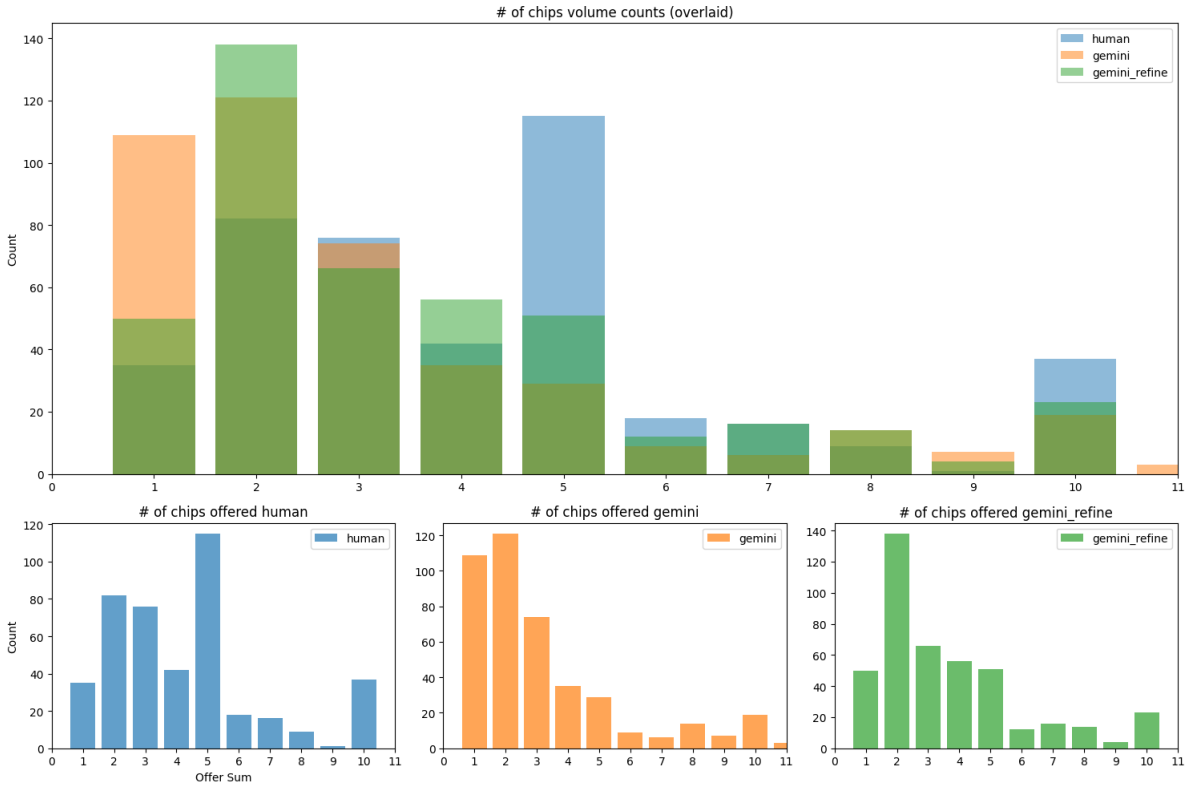


Figure 10: Number of chips offered by the proposer in a turn.

Trade volume. In Fig 10, we display this phenomenon. Human proposers trade at significantly higher volumes than their LLM counterparts. Humans' modal bumps at 5 and 10 suggest a preference for round numbers, reflecting a potential cognitive bias rather than strictly rational decision-making. In contrast, LLM agents tend to trade in smaller amounts, with modal values at 1 and 2, indicating a more conservative trading approach. Because humans trade away more chips at once (10 being the total number of chips of a single color), they are more frequently locked out of lucrative trades later in the game. This makes them more susceptible to opportunity cost mistakes, both as proposers (by undervaluing chips) and acceptors (by lacking the resources to accept better offers when they arise).

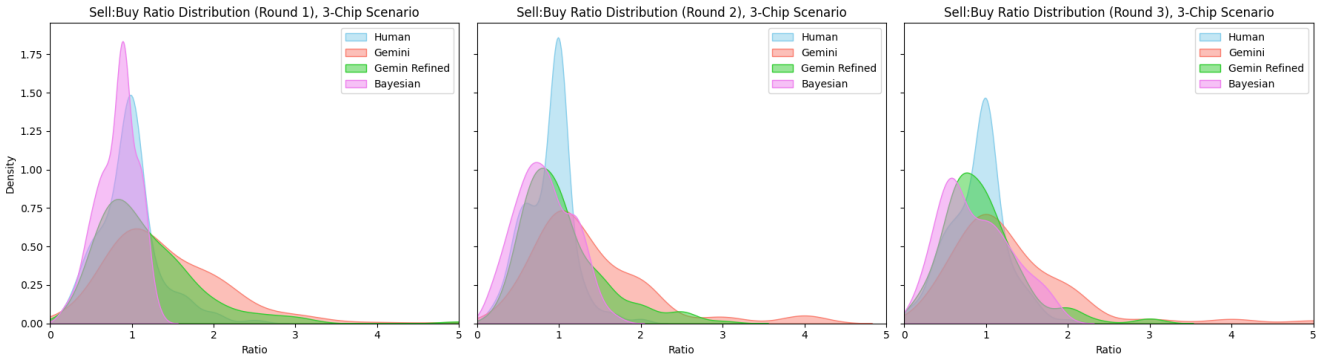


Figure 11: Ratio of sell-buy chips offered by the rational proposer

Trade ratio. An important distinction is that although humans' offers are for higher amounts, the offers are still largely *fair*. In Fig 11, we see how the rational proposer's sell:buy ratio evolves across three rounds in a 3-chip scenario, comparing humans, Gemini agents, Gemini refined agents, and the Bayesian-learning agent. Humans tend to cluster around a 1:1 ratio. This resonates with behavioral game theory observations that many human participants initially adopt a "fair" heuristic when proposing trades. By contrast, both Gemini agents display broader distributions while still showing a peak near 1:1.

As the game proceeds to later rounds, humans in particular begin to shift leftward on the ratio axis, meaning they propose offering fewer high-valued chips for larger quantities of lower-valued chips. In other words, they begin buying more and selling less. This is evidenced by the development of a second peak on the left tail of the human distribution in Round 2 and Round 3. In game-theoretic terms, this shift can be interpreted as humans learning that other players' chip valuations differ from their own, and that by exploiting that difference, they can propose more favorable trades without provoking refusal. Notably, both the Gemini-refined agent and the Bayesian agent show a similar trend: as the rounds progress, they also allocate more probability mass to proposals with sell:buy ratios below 1. This signaled an intentional exploitation of perceived or learned valuation differences. The Bayesian agent's advantage lies in systematically updating its beliefs about the opponent's valuation structure after each round. Hence, it can refine its offers accordingly. The Gemini-refined agent similarly moves away from purely balanced proposals once it "observes" that certain unbalanced ratios tend to be accepted.

Buy-in chip types. Our trade decision analyses demonstrate the profit-maximizing and collaborative strategies that human participants reported in their qualitative feedback. The third type of self-reported strategy was color-dependent, where participants would optimize for their most valuable chips, or attempt to trade away their lower chips. Fig 12 compares the proportions of "most valuable" versus "second most valuable" chips in the 3-chip game that the rational proposer attempts to buy across three rounds. By excluding all irrational trade proposers from the data, we isolate purely strategic behavior.

A benchmark is that when the market has ample supply, rational agents naturally opt to buy the most valuable chips. This is shown by the large blue bars in Round 1 across all four panels (Gemini, Gemini Refined, Human, and Bayesian). As the game proceeds, however, the supply of these top-tier chips becomes increasingly constrained. This drives up their effective "price," either in terms of what a proposer must offer or how difficult it becomes to secure those chips from others.

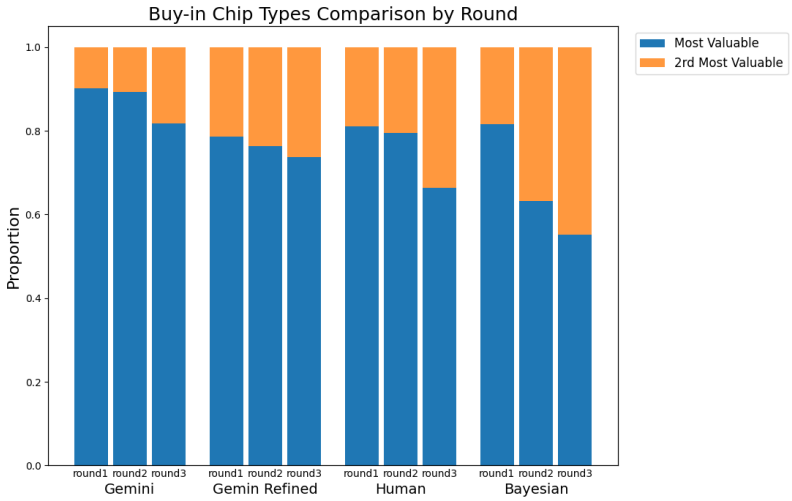


Figure 12: Buy-in type offered by the rational proposer.

Consequently, in Rounds 2 and 3, we observe a gradual shift toward purchasing the second most valuable chips (shown in orange). This effect appears most pronounced for the Bayesian agent, as its updating mechanisms are likely more aggressive than the other agents'. Human proposers exhibit a similar move, suggesting that players glean from experience that if top-tier chips are unavailable or too costly, they can improve payoffs by targeting the next-best alternative. Although Gemini and Gemini Refined generally maintain a strong preference for the highest-valued chips in every round, we still see some uptick in their purchase of second most valuable chips by Round 3.

8 Conclusion

In this paper, we have presented a practical framework for evaluating the tradeoffs that arise when choosing between different types of agents—human or AI—for economic tasks. Rather than prescribing which agent to use in any specific scenario, our goal was to delineate the key levers that shape such decisions. The six dimensions at the core of the framework—preferences, strategic capabilities, time and cost, information sets, environment structure, and complexity—work together to highlight how varying contexts can shift the relative benefits of human versus AI agents. By offering a systematic way to account for these factors, we aim to help decision-makers navigate an evolving AI landscape where uncertainty, coordination challenges, and social welfare considerations are all intertwined.

To illustrate the framework in a concrete setting, we developed a custom bargaining game under incomplete information. The game's flexible design—from a simple two-chip trading scenario to a more intricate four-chip variant—captured many strategic complexities that mirror real-world negotiations: coordination failures, asymmetric knowledge, and varying preferences across participants. By comparing the final allocations in each game to a Pareto-efficient linear-programming benchmark, we introduced the "Price of Agency," a measure of how close different agent populations came to the theoretical optimum. In demonstrating this approach, we emphasized that our methods can be adapted to a broad range of tasks, not just bargaining, to guide principled decisions about the tradeoffs, alignment considerations, and complementarity of human and AI decision making.

We found that both human participants and off-the-shelf large language model (LLM) agents can approach high levels of efficiency in the simple game variant. As complexity increased, humans tended to outmaneuver basic LLMs by employing more strategic and adaptive tactics, such as accepting short-term losses to secure larger gains later. However, small prompting refinements to the LLM agents bridged this gap. Humans and the LLM-based agents appeared to utilize different strategies, such as humans favoring fair or balanced deals and LLMs pursuing short-term payoffs, suggesting potential for complementarity.

The differences in time cost were pronounced: human participants required significantly more time to achieve marginally better results, underscoring the potential value of faster AI-driven negotiation in many real-world contexts. In our experiments, humans spent roughly the same amount of time on each game variation, even as complexity rose, and self-reported similar effort and confidence across all variations. Yet the only factor strongly correlated with their final satisfaction was the monetary payout. AI agents, by contrast, completed negotiations quickly and consistently, offering a potentially substantial advantage in tasks where speed is at a premium. While AI may not universally dominate on cost—especially when setup or monitoring costs are large—it is often poised to deliver major savings in repeated or large-scale tasks.

Of course, the pace at which LLMs are evolving suggests that the current performance gap might soon become obsolete or might shift in unanticipated ways. Models that today appear weaker than humans in complex negotiations could, with further training or updated architectures, surpass them across relevant dimensions as they evolve. Or they might improve in some respects (such as strategic risk-taking) while losing ground in others (such as consistency). Moreover, our experiments, though designed to be general, may not perfectly predict the performance of other LLMs or of humans from different demographic samples. The chief lesson is that this framework — combining multiple dimensions with alignment measures such as the Price of Agency —— can assess AI agents' suitability in any given environment. As models improve, or as human–AI collaboration practices evolve, systematically applying this framework will continue to be a worthwhile exercise. Future research could explore humans' willingness to delegate and collaborate, examine heterogeneous markets with both human- and AI- based players, or extend the existing scenario to incorporate natural-language negotiation to capture richer strategic interactions.

Looking ahead, these findings reinforce the notion that deciding whether to delegate tasks to LLM-based agents requires more than a simple assessment of raw capability. Each of the framework's six dimensions can matter in different proportions, depending on the domain and potential interactions with other participants. The complementarity apparent in our experiments—where humans used context or social cues to outperform purely myopic strategies, and AI agents excelled in efficiency—points to the possibility that a hybrid approach may be optimal in many real-world deployments. More broadly, as LLMs become more autonomous and gain improved reasoning abilities, they could transform the design of market mechanisms and institutions, amplifying the need for robust, context-aware analyses of when and how to substitute humans with AI. By continuously applying and refining the framework proposed here, researchers and practitioners can better understand whether an AI's rapid iteration and strategic consistency will indeed deliver net benefits, or whether the human-AI alignment gap is unacceptable. In doing so, organizations and policymakers can strike a deliberate balance between human creativity and AI's computational might, harnessing the most valuable qualities of each while maintaining vigilance over the risks inherent in shared decision-making.

Acknowledgments and funding disclosure

We thank Hal Abelson, David Parkes, Lucas Dixon, and Ian Gemp for their valuable feedback and support. Horton is a co-founder of a company, Expected Parrot Inc., using generative AI models for research. Horton was previously a visiting researcher at Microsoft. Manning also holds a financial interest in Expected Parrot Inc. Zhu is supported by a Google DeepMind Research Seed Fund.

References

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz, "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology," Working Paper 31422, National Bureau of Economic Research July 2023.

Anshelevich, Elliot, Anirban Dasgupta, Jon Kleinberg, Éva Tardos, Tom Wexler, and Tim Roughgarden, "The price of stability for network design with fair cost allocation," *SIAM Journal on Computing*, 2008, *38* (4), 1602–1623.

Anthropic, "Claude Code," 2025. Command line tool for delegating coding tasks to Claude AI assistant.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ," in "Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency" FAccT '21 Association for Computing Machinery New York, NY, USA 2021, p. 610–623.

Braun, Virginia and Victoria Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, 2006, *3* (2), 77–101.

Enke, Benjamin, "The Cognitive Turn in Behavioral Economics," July 2024. Working paper.

(FAIR)†, Meta Fundamental AI Research Diplomacy Team, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu et al., "Human-level play in the game of Diplomacy by combining language models with strategic reasoning," *Science*, 2022, *378* (6624), 1067–1074.

Fish, Sara, Yannai A Gonczarowski, and Ran I Shorrer, "Algorithmic collusion by large language models," *arXiv preprint arXiv:2404.00806*, 2024, *7*.

Fudenberg, D and J Tirole, "Game Theory, 616 pages," 1991.

Gaarlandt, Jur, Wesley Korver, Nathan Furr, and Andrew Shipilov, "AI Agents Are Changing How People Shop. Here's What That Means for Brands," *Harvard Business Review*, 2025.

Gode, Dhananjay K and Shyam Sunder, "Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality," *Journal of political economy*, 1993, *101* (1), 119–137.

Goktas, Denizalp, Amy Greenwald, Takayuki Osogami, Roma Patel, Kevin Leyton-Brown, Grant Schoenebeck, Daphne Cornelisse, Constantinos Daskalakis, Ian Gemp, John Horton, David C Parkes, David M. Pennock, Arjun Prakash, Sai Srivatsa Ravindranath, Max Olan Smith, Gokul Swamy, Eugene Vinitsky, Segev Wasserkrug, Michael Wellman, Jibang Wu, Haifeng Xu, Jiayao Zhang, Yichi Zhang, Sadie Zhao, and Quanyan Zhu, "Strategic Foundation Models," February 2025. working paper or preprint.

Google Flash Thinking, "Gemini API: Thinking and Reasoning," 2024. Accessed: March 20, 2025.

Hoek, Remko Van, Michael DeWitt, Mary Lacity, and Travis Johnson, "How Walmart automated supplier negotiations," *Harvard Business Review*, 2022, *8* (11), 2022.

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

Jr, John F Nash, "Equilibrium points in n-person games," *Proceedings of the national academy of sciences*, 1950, *36* (1), 48–49.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler, "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias," *Journal of Economic Perspectives*, March 1991, *5* (1), 193–206.

Kapania, Shivani, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan, ""Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India," in "Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems" CHI '22 Association for Computing Machinery New York, NY, USA 2022.

Koutsoupias, Elias and Christos Papadimitriou, "Worst-case equilibria," in "Annual symposium on theoretical aspects of computer science" Springer 1999, pp. 404–413.

Kramár, János, Tom Eccles, Ian Gemp, Andrea Tacchetti, Kevin R McKee, Mateusz Malinowski, Thore Graepel, and Yoram Bachrach, "Negotiation and honesty in artificial intelligence methods for the board game of Diplomacy," *Nature Communications*, 2022, *13* (1), 7214.

Krishna, Ranjay, Donsuk Lee, Li Fei-Fei, and Michael S Bernstein, "Socially situated artificial intelligence enables learning from human interaction," *Proceedings of the National Academy of Sciences*, 2022, *119* (39), e2115730119.

Kruger, Justin and David Dunning, "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments.," *Journal of personality and social psychology*, 1999, *77* (6), 1121.

Lubars, Brian and Chenhao Tan, "Ask not what AI can do, but what AI should do: Towards a framework of task delegability," in H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, Vol. 32 Curran Associates, Inc. 2019.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan, "Large Language Models: An Applied Econometric Framework," 2024.

Manning, Benjamin S, Kehang Zhu, and John J Horton, "Automated social science: Language models as scientist and subjects," Technical Report, National Bureau of Economic Research 2024.

Myerson, Roger B and Mark A Satterthwaite, "Efficient mechanisms for bilateral trading," *Journal of Economic Theory*, 1983, *29* (2), 265–281.

Oprea, Ryan, "Decisions under Risk Are Decisions under Complexity," *American Economic Review*, 2024. Available online at `https://www.aeaweb.org/journals/aer/forthcoming`.

Palminteri, Stefano, Basile Garcia, and Crystal Qian, "The Moral Turing Test: Evaluating Human-LLM Alignment in Moral Decision-Making," Oct 2024.

Park, Joon Sung, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein, "Generative Agent Simulations of 1,000 People," 2024.

_ , Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein, "Generative agents: Interactive simulacra of human behavior," in "Proceedings of the 36th annual acm symposium on user interface software and technology" 2023, pp. 1–22.

Qian, Crystal and James Wexler, "Take It, Leave It, or Fix It: Measuring Productivity and Trust in Human-AI Collaboration," in "Proceedings of the 29th International Conference on Intelligent User Interfaces" IUI '24 Association for Computing Machinery New York, NY, USA 2024, p. 370–384.

Raman, Narun, Taylor Lundy, Samuel Amouyal, Yoav Levine, Kevin Leyton-Brown, and Moshe Tennenholtz, "STEER: Assessing the economic rationality of large language models," *arXiv preprint arXiv:2402.09552*, 2024.

Soumalias, Ermis, Yanchen Jiang, Kehang Zhu, Michael Curry, Sven Seuken, and David C Parkes, "LLM-Powered Preference Elicitation in Combinatorial Assignment," *arXiv preprint arXiv:2502.10308*, 2025.

Team, Gemini, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

Tessler, Michael Henry, Michiel A Bakker, Daniel Jarrett, Hannah Sheahan, Martin J Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C Parkes et al., "AI can help humans find common ground in democratic deliberation," *Science*, 2024, *386* (6719), eadq2852.

Tsai, Vivian, Crystal Qian, and Deliberate Lab community contributors, "Deliberate Lab: Open-Source Platform for LLM-Powered Social Science," September 2024.

Zhu, Kehang, John Joseph Horton, Yanchen Jiang, David C. Parkes, and Anand V. Shah, "Evidence from the Synthetic Laboratory: Language Models as Auction Participants," in "NeurIPS 2024 Workshop on Behavioral Machine Learning" 2024.

A Algorithms and proofs

A.1 Multi-agent bilateral trading algorithm

Algorithm 1 Multi-agent Bilateral Trading Algorithm

$\mathbf{A} \leftarrow \mathbf{A^0}$
$possible_trades \leftarrow$ `true`
while $possible_trades$ **do**
    $trades_found \leftarrow$ `false`
    $I_{\text{shuffled}} \leftarrow \text{RandomShuffle}(I)$
    for $k = 1$ **to** $|I|$ **do**
        for $l = k + 1$ **to** $|I|$ **do**
            $(i, j) \leftarrow (I_{\text{shuffled}}[k], I_{\text{shuffled}}[l])$
            $\mathbf{A^*_{ij}} \leftarrow \text{SolveA}^*(i, j, G, \{a_{ig}, a_{jg}\}_{g \in G})$
            if $\mathbf{A^*_{ij}} \neq \{a_{ig}, a_{jg}\}_{g \in G}$ **then**
                $\{a_{ig}, a_{jg}\}_{g \in G} \leftarrow \mathbf{A^*_{ij}}$
                $trades_found \leftarrow$ `true`
            end if
        end for
    end for
    $possible_trades \leftarrow trades_found$
end while
return A

A.2 Proof of Theorem 1

Consider a set of agents $I = 1, 2, \ldots, N$ and a set of indivisible goods $J = 1, 2, \ldots, M$. Each agent $i \in I$ has a valuation V_{ij} for each good $j \in J$. Let $A^0 = (A^0_{ij})$ denote the initial allocation, where $A^0_{ij} \in 0, 1$ indicates whether good j is allocated to agent i. The allocation satisfies $\sum_i A^0_{ij} = 1$ for all j (each good is allocated to exactly one agent). A bilateral trade between two agents i and k involves a reallocation of goods, resulting in a new allocation $A = (A_{ij})$ where:

1. $A_{ij} \in 0, 1$ for all i, j
2. $\sum_i A_{ij} = 1$ for all j (each good remains allocated to exactly one agent)
3. $A_{ij} = A^0_{ij}$ for all j and all $i \notin i, k$ (only the allocations of agents i and k change)
4. $\sum_j A_{ij} V_{ij} > \sum_j A^0_{ij} V_{ij}$ and $\sum_j A_{kj} V_{kj} > \sum_j A^0_{kj} V_{kj}$ (both agents strictly benefit from the trade)

An allocation A is defined as bilaterally stable if there does not exist a pair of agents i, k and a reallocation satisfying the above conditions. Define the social welfare of an allocation A as $SW(A) = \sum_i \sum_j A_{ij} V_{ij}$, the total value of the allocation to all agents. We claim that there exist problem instances where:

1. There is a bilaterally stable allocation A.
2. There is another allocation B such that $SW(B) \geq k \cdot SW(A)$ for arbitrarily large k.

To construct such an instance, let $N = M = 3$ and consider the following valuations:

- $V_{11} = V_{22} = V_{33} = X$ for some large X
- All other $V_{ij} = 1$

Let A^0 be the initial allocation where $A^0_{12} = A^0_{23} = A^0_{31} = 1$ and all other $A^0_{ij} = 0$. This allocation is bilaterally stable, as no pair of agents can benefit from trading. However, $SW(A^0) = 3$. In contrast, consider the optimal allocation B where $B_{11} = B_{22} = B_{33} = 1$ and all other $B_{ij} = 0$. We have $SW(B) = 3X$. By choosing X arbitrarily large, we can make the ratio $SW(B)/SW(A^0) = X$ arbitrarily large.

Therefore, there exist problem instances where bilateral trading can lead to arbitrarily suboptimal allocations compared to the social welfare maximizing allocation.

B Trade distributions for two-chip and four-chip game variations

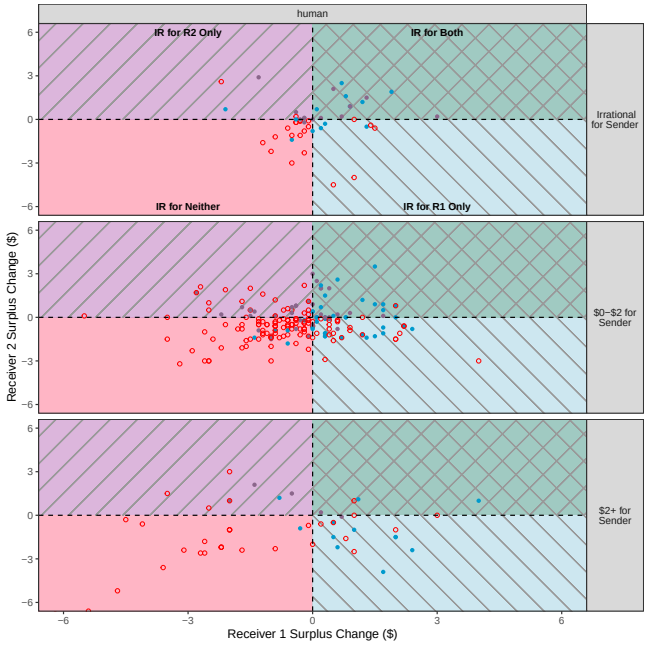


Figure 13: (2-chip game) Player surplus trade visualization

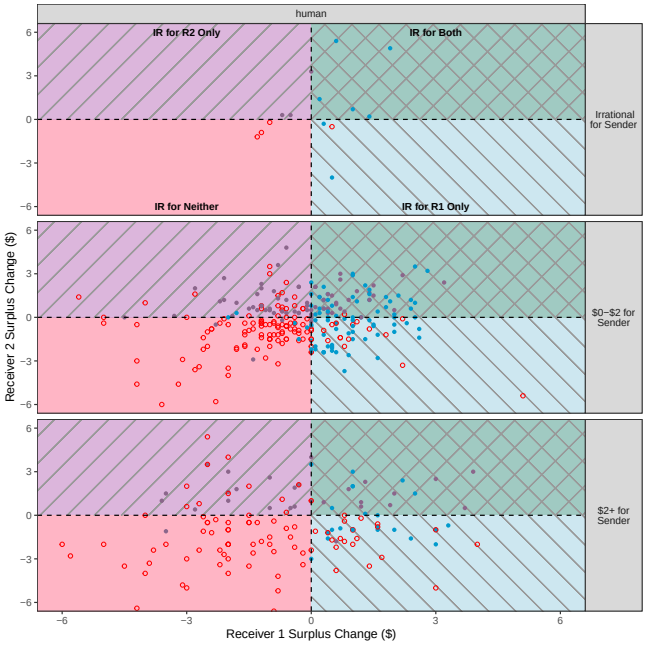


Figure 14: (4-chip variation) Player surplus trade visualization

C Surplus trajectory visualizations

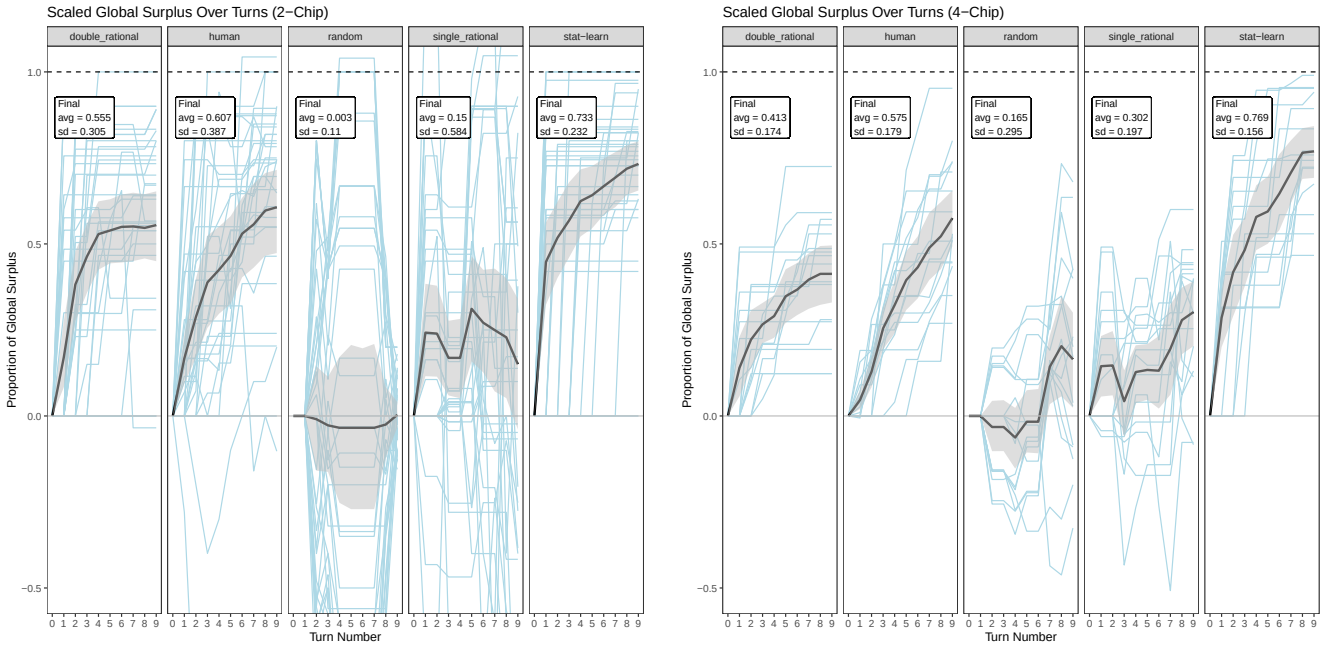


Figure 15: (2-chip, left, and 4-chip, right) Surplus trajectories for **noise-driven agents**.

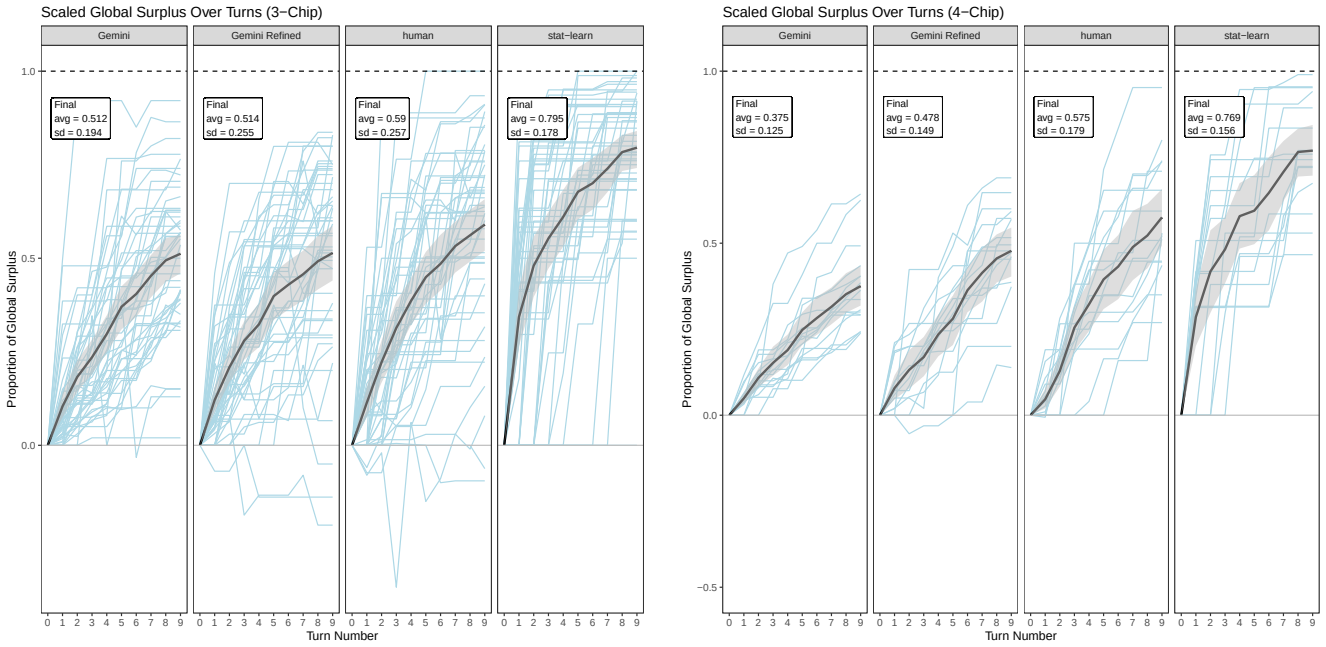


Figure 16: (3-chip, left, and 4-chip, right) Surplus trajectories for **human, LLM, and Bayesian-learning agent simulations**.

D Human data collection

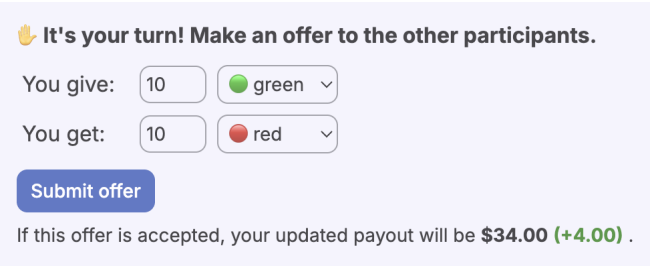
D.1 Implementation on Deliberate Lab

Platform. Human experiments were conducted via Deliberate Lab (Tsai et al. (2024)), an open-source platform for conducting group human-LLM social science experiments.

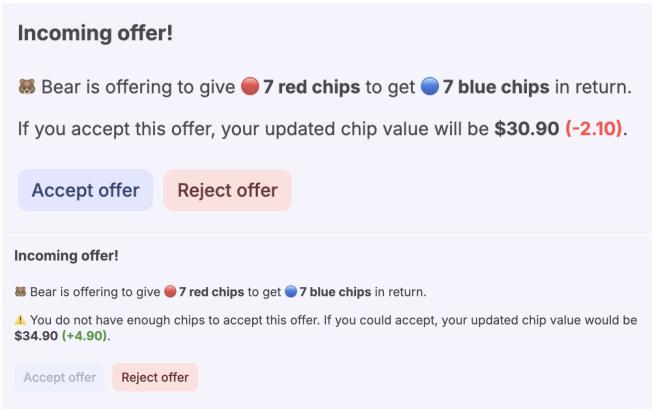
Participant experience. Upon entering Deliberate Lab through a web link, participants enter a multi-stage experiment include game instructions, comprehension checks, and payout information. Upon completing the final comprehension check, they wait in a "Lobby" stage for other participants. When three participants are in the lobby, they are sent an invitation to join a live bargaining game (Fig 2). Following the games, there is a post-game survey. For anonymity, we used a Deliberate Lab feature that assigns participants an anonymous animal avatar (e.g., "Bear") as they join the experiment.

Bargain game user interface. To submit trade proposals, participants used text and selection fields to specify number and type of chip, respectively. The interface automatically calculated/displayed projected total chip value if the proposal were to be accepted (Fig 17b); it also prevented players from submitting invalid offers (by disabling the "submit" button). For accepting/rejecting proposals, the interface also calculated/displayed project total chip value if the player were to accept the proposal; it also prevented players from accepting offers if the player did not have the chips needed to conduct the trade (Fig 17b).

All participants were shown a table of all players' chip quantities and a log of actions (i.e., when trades were proposed, accepted, and rejected and when each round and turn of the game changed). Turn order was determined randomly.



(a) Submitting a trade proposal



(b) Accepting or rejecting a trade. The bottom image shows a disabled "accept" option.

Experimenter interface. To conduct experiments, the experimenter used Deliberate Lab's dashboard (Fig 18) to monitor participants as they progressed through the experiment's information and comprehension check stages, then manually transferred participants into groups of three to play the bargaining game once the player was ready to be transferred. The experimenter was also able to send attention checks to participants (e.g., if several minutes elapsed without any visible progress) and remove them from the experiment (e.g., if attention checks went unanswered).

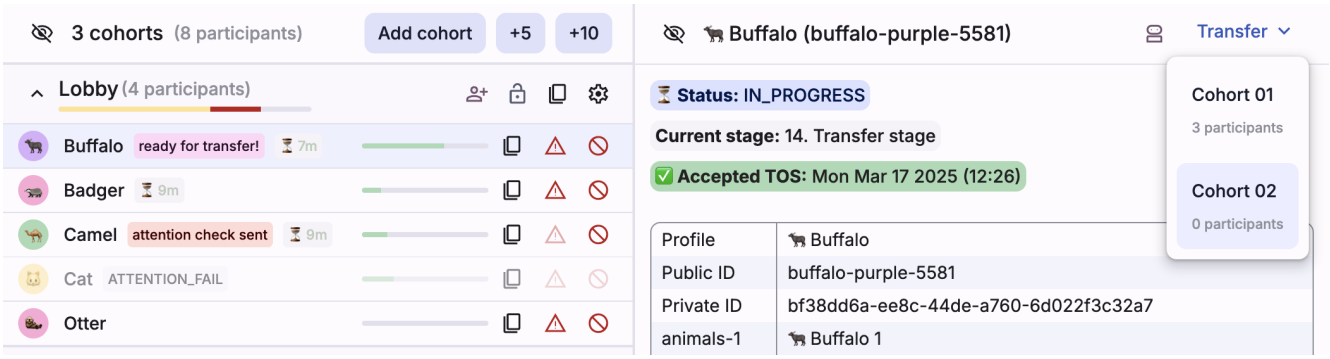


Figure 18: Experiment dashboard for managing participants. On the left, experimenters can view participants as they enter the experiment, send attention checks, and/or remove participants. On the right, experimenters can monitor the status of a selected participant and assign them to a cohort of three for gameplay.

D.2 Post-game survey

	Survey Question	Question Format
1	Please describe your strategy in the game in a few sentences.	freeform text
2	Please describe any experiences or background that may have influenced your performance in the game.	freeform text
3	On a scale from 1 to 10, how would you rate your trading strategy, where 1 is highly **competitive** (focused mainly on your own gains) and 10 is highly **collaborative** (focused on other players's potential gains and mutual benefits?	scale of integers 1 through 10, with 10 labeled "Very aggressive" and 1 labeled "Not at all aggressive"
4	On a scale from 1 to 10, how certain are you that you made the best possible trades during the experiment?	scale of integers 1 through 10, with 10 labeled "Very confident" and 1 labeled "Not at all confident"
5	On a scale from 1 to 10, how satisfied are you with your final trading outcomes?	scale of integers 1 through 10, with 10 labeled "Very satisfied" and 1 labeled "Not at all satisfied"
6	On a scale of 1 to 10, how would you rate the mental effort you put into today's trading games, where 1 means 'I barely engaged in any thinking' and 10 means 'I put in significant mental effort'?	scale of integers 1 through 10, with 10 labeled "Very high mental exertion" and 1 labeled "No mental exertion at all"
7	Please provide any additional context on your answers above.	freeform text
8	Please help us to improve this experiment. How was your experience today? Were there any elements of the instructions or gameplay that you found confusing?	freeform text

Table 4: Post-game survey questions given to participants via *Deliberate Lab*.

E LLM simulation prompts

E.1 Rule explanation

```
How the game works:
The game consists of 3 rounds of trading. During each round, each player will have a turn to
    propose 1 trade. These turns are pre-determined in a random order and the order stays the
    same in each round.

Trade proposals:
To propose a trade, a player must:
1. Request a certain quantity of chips of a single color from any other player to get.
2. Specify a certain quantity of chips of a different color to give in return.

Trade rules:
Players cannot offer more chips than they currently hold. For example, if you only have 5 red
    chips, you cannot offer 6 red chips.
Players cannot trade chips of the same color. For example, you cannot trade red chips for red
    chips.

Trade completion:
When an offer is presented, all other active participants get a chance to accept or decline.
    Note: Active participants are those not currently making the offer.

Participants make their decisions simultaneously and privately. The participant who receives
    the offer is not dependent on who accepts the trade first. Some possible outcomes:
If no one accepts, the trade does not happen, and the turn ends.
If multiple participants accept, one accepting participant is chosen at random to complete
    the trade with the offering participant. This means that participants cannot choose who
    they trade with.
If only one participant accepts, the trade will happen.

Key points to remember:
In each round, each player gets to propose one trade and respond to other player's trades
You can only propose trades between different colored chips, and cannot offer to give a chip
    amount that you do not have
When multiple players accept a trade, the trading partner is randomly selected
"""
```

E.2 Proposing prompt

We used a Chain-of-Thought approach to elicit LLM agent's trade proposals:

```
You are {{name}}.
Your valuations of the different types of chips are: {{preference_description}}.
```

```
You now have the following amounts of each chip: {{item}}.
The conversation history so far is {{history}}.

REMEMBER, to propose a trade, you must:
Request a certain quantity of chips of a single color from any other player to get
Specify a certain quantity of chips of a different color to give in return

REMEMBER you have the following amounts of each chip: {{item}}.
Your goal is to make as much money as possible. The trades, you choose to make to accomplish
    this, are up to you.
As a part of making money you must be rational - do not propose a trade in which you lose
    money. The value of a trade to you is the difference between the total value of chips you
    receive (quantity x your valuation) minus the total value of chips you give up (quantity x
    your valuation). Only propose trades that give you positive value.
In short, your trades should be both incentive compatible and incentive rational.
Your response must use these EXACT tags below. The response should include nothing else
    besides the tags, your trade offer, and your reasoning. The text between tags should be
    concise.
‘‘‘
<REASONING>
[Provide your concise reasoning in a few sentences, e.g. To gain more surplus, I want more
    xxx chips]
</REASONING>

<CHECK>
[check if you have sufficient chips to trade. If you have n green chips, you can at most give
    n green chips. If you don’t want to trade, you can ask for a large amount of chips that no
    one can afford]
<\CHECK>

<GET_COLOR> Color, e.g. red</GET_COLOR>
<GET_QUANTITY> quantity, e.g. n </GET_QUANTITY>
<GIVE_COLOR> Color, e.g. red</GIVE_COLOR>
<GIVE_QUANTITY> quantity, e.g. n </GIVE_QUANTITY>
‘‘‘
```

E.3 Answering prompt

```
You are {{name}}.
Your valuations of the different types of chips are: {{preference_description}}.
You now have the following amounts of each chip: {{item}}.
The conversation history so far is {{history}}.

You have an offer. {{proposer}} is offering to give {{give}} and get {{get}} in return.
If you make this trade, your total wealth will change by: {{delta_surplus}}
```

```
Now, you need to decide whether to accept or decline.
Your response must use these EXACT tags below. The response should include nothing else
    besides the tags, your choice to accept or decline, and your reasoning. The text between
    tags should be concise.
‘‘‘
<REASONING>
[Provide your concise reasoning in a few sentences.]
</REASONING>

<CHOICE>Yes or No </CHOICE>
```

E.4 Multiple proposal generation prompt

```
You are {{name}}.
Your valuations of the different types of chips are: {{preference_description}}.
You now have the following amounts of each chip: {{item}}.
The conversation history so far is {{history}}.

REMEMBER, to propose a trade, you must:
Request a certain quantity of chips of a single color from any other player to get
Specify a certain quantity of chips of a different color to give in return

REMEMBER you have the following amounts of each chip: {{item}}.
Your goal is to make as much money as possible. The trades, you choose to make to accomplish
    this, are up to you.
As a part of making money you must be rational - do not propose a trade in which you lose
    money. The value of a trade to you is the difference between the total value of chips you
    receive (quantity  your valuation) minus the total value of chips you give up (quantity
    your valuation). Only propose trades that give you positive value.
In short, your trades should be both incentive compatible and incentive rational.
You can trade as many chips as you want in a single turn, assuming you have that many. Do not
    feel constrained to only trade a single chip at a time.
Propose 3 different good trade ideas, so a next step can decide on the best trade of the ones
    you propose here.
Your response must use these EXACT tags below. The response should include nothing else
    besides the tags, your trade offer, and your reasoning. The text between tags should be
    concise.
Repeat the below tags once for each of the trade ideas you propose.
‘‘‘
<REASONING>
[Provide your concise reasoning in a few sentences, e.g. To gain more surplus, I want more
    xxx chips]
</REASONING>

<CHECK>
[check if you have sufficient chips to trade. If you have n green chips, you can at most give
```

```
    n green chips. If you don't want to trade, you can ask for a large amount of chips that no
    one can afford]
<\CHECK>

<GET_COLOR> Color, e.g. red</GET_COLOR>
<GET_QUANTITY> quantity, e.g. n </GET_QUANTITY>
<GIVE_COLOR> Color, e.g. red</GIVE_COLOR>
<GIVE_QUANTITY> quantity, e.g. n </GIVE_QUANTITY>
'''
```

E.5 Choosing from multiple possible proposals prompt

```
You are {{name}}.
Your valuations of the different types of chips are: {{preference_description}}.
You now have the following amounts of each chip: {{item}}.
The conversation history so far is {{history}}.

REMEMBER, to propose a trade, you must:
Request a certain quantity of chips of a single color from any other player to get
Specify a certain quantity of chips of a different color to give in return

REMEMBER you have the following amounts of each chip: {{item}}.
Your goal is to make as much money as possible. The trades, you choose to make to accomplish
    this, are up to you.
As a part of making money you must be rational - do not propose a trade in which you lose
    money. The value of a trade to you is the difference between the total value of chips you
    receive (quantity  your valuation) minus the total value of chips you give up (quantity
    your valuation). Only propose trades that give you positive value.
In short, your trades should be both incentive compatible and incentive rational.
You can trade as many chips as you want in a single turn, assuming you have that many. Do not
    feel constrained to only trade a single chip at a time.

Below are three trade ideas you have proposed. Please pick the best trade proposal from the
    ones below that you will propose to the group.
Do not change anything about the trade you have selected from the ideas below.
Your response must use these EXACT tags below. The response should include nothing else
    besides the tags and content of your selected trade offer including its reasoning.
'''
<REASONING>
[Provide your concise reasoning in a few sentences, e.g. To gain more surplus, I want more
    xxx chips]
</REASONING>

<CHECK>
[check if you have sufficient chips to trade. If you have n green chips, you can at most give
    n green chips. If you don't want to trade, you can ask for a large amount of chips that no
```

```
    one can afford]
<\CHECK>

<GET_COLOR> Color, e.g. red</GET_COLOR>
<GET_QUANTITY> quantity, e.g. n </GET_QUANTITY>
<GIVE_COLOR> Color, e.g. red</GIVE_COLOR>
<GIVE_QUANTITY> quantity, e.g. n </GIVE_QUANTITY>
‘‘‘

Proposed trade ideas to choose from:
{{proposed}}
```