

LLMs with Personalities in Multi-issue Negotiation Games

Sean Noh¹, Ho-Chun Herbert Chang¹

¹Dartmouth College

sean.noh.28@dartmouth.edu, herbert.chang@dartmouth.edu,

Abstract

AI negotiation agents are rapidly proliferating on online spaces, brokering millions of transactions monthly between businesses and consumers. As platforms increasingly endow these agents with distinct personality traits, understanding how personality shapes fairness, exploitation dynamics, and linguistic toxicity becomes critical for platform governance and consumer protection. As many are endowed with distinct personality traits, understanding how these traits shape negotiation behavior is crucial. Using canonical OCEAN personality definitions, we initialize LLM agents and simulate single- and multi-issue bargaining games ($n = 3,500$) within a game-theoretic framework. Results reveal two key insights. First, we identify the linkage between personality, fairness, and exploitation: agreeable agents have the highest payoffs but are exploitable by less agreeable agents. Second, agents tend toward fairness rather than rationality, which may arise from innate guardrails. However, guardrails can be “jailbroken” with the right personalities. Personality thus shapes both quantitative (surplus discovery) and qualitative (toxicity, persuasion) dimensions of negotiation. We discuss how AI personality steers negotiation behavior toward certain domain objectives, with practical implications for designers.

Introduction

Powered by large language models (LLMs), AI agents have become capable of many human tasks, including negotiation. Online marketplaces now deploy automated bargaining systems to handle supplier negotiations (Hoek et al. 2022), enable customer price haggling in companies such as Wall Mart (Debter 2023). As these systems grow more capable and commonplace, understanding how personality configurations shape negotiation outcomes is a matter of platform governance, consumer protection, and digital fairness.

Negotiation is a fundamental human process, simultaneously including **decision-making**, **cooperation**, and **persuasion**. It is a mechanism for settling issues, discovering surplus, and creating value. Consequently, it has been extensively studied across economics (Osborne and Rubinstein 1994; Raiffa 1982), business (Walsh and Wellman 1999; Lewicki, Barry, and Saunders 2016; Huang et al. 2010), communication (Arvanitis and Karampatzos 2011;

Maaravi, Ganzach, and Pazy 2011), and psychology (Rubin and Brown 2013; De Dreu et al. 2007). Behavioral research has shown that personality traits influence negotiation style and outcomes—particularly neuroticism, extraversion, and agreeableness, which shape cooperative or competitive tendencies (Ma 2005; Marlowe, Gergen, and Doob 1966; Brandstätter and Königstein 2001). Yet situational and linguistic factors, such as the rhetoric used, can often outweigh stable traits (Morris, Larrick, and Su 1999).

Game-theoretic approaches formalize these dynamics through bargaining games, where two-players divide one or more issues over limited rounds. Rational strategies maximize individual utility while fairness criteria minimize payoff disparities. Early work in automated negotiation applied these frameworks to multi-agent systems (Jennings et al. 2001; Kraus 1997; Gerding, van Bragt, and La Poutré 2000), while recent advances in deep reinforcement learning (DRL) have enabled agents to learn negotiation strategies end-to-end. From board games such as Chess, Go, and Poker (Silver et al. 2016, 2017; Mnih et al. 2013) to social dialogue games (Mei et al. 2023; Davidson et al. 2024), DRL has enabled agents to outperform humans through deceptive behaviors (Lewis et al. 2017; Chang 2021).

Recent work shows how negotiation represents a frontier for social LLMs. LLM bargaining systems for individual sellers on online marketplaces show proactive strategy planning and linguistic skill can measurably increase successful deals (Kong et al. 2025). More formally, commitment and signaling produce efficient equilibria even with complex valuations in sequential auctions (Gkatzelis, McAfee, and Leme 2025) and contract design (Mirrokni and Nazarzadeh 2017). However, studies do show that models can be “jailbroken” through persuasion (Zeng et al. 2024), raising manipulation concerns about consumer-facing applications.

In this study, we analyze how LLMs bargain through a game-theoretic and linguistic lens, contributing to pragmatic insights for developers, researchers, and policymakers. By varying agents’ personalities along the OCEAN framework, we systematically examine how personality prompts influence negotiation outcomes, including surplus discovery, payoff fairness, toxicity and assertiveness. We show that LLM personalities propagates directly to canonical measures of rationality and fairness in bargaining games, bridging behavioral and computational frames of negotiation.

These findings speak directly to growing research on **social simulation**, psychologically grounded and validated agent **personas** (BFI-2), and the **ethical framework** for designing safer, fairer AI agents in online marketplaces.

Literature

Bargaining Games

A **bargaining game** models how two or more players negotiate the allocation of a resource to reach a mutually beneficial agreement. Formally, it is defined as a tuple (N, S, d) , where N is the set of players (typically $N = \{1, 2\}$), $S \subset R^n$ is the set of feasible utility allocations (the *bargaining set*), and $d \in S$ is the disagreement point—the utilities each player receives if no agreement is reached (Osborne and Rubinstein 1994). The goal is to find a solution where rational players agree, given that disagreement yields d .

In practice, negotiation often involves conflict and is better modeled through non-cooperative frameworks. Rubinstein’s (1982) alternating-offers model captures this dynamic: players take turns proposing divisions of surplus over time, possibly with discounting or a deadline (Rubinstein 1982). A negotiation setting consists of a *protocol*, *agents*, and a *scenario* (Baarslag et al. 2012). The protocol defines interaction rules, while the scenario specifies the negotiation domain and its outcome space Ω , which may include one or multiple issues.

Under perfect information, optimal bidding strategies can be derived (Baarslag et al. 2012), though this is rarely realistic since revealing preferences enables exploitation. Consequently, much research addresses strategies under *imperfect information*. Negotiation strategies are typically classified into three components: the *bidding strategy* (how to counteroffer), the *acceptance strategy* (when to accept offers), and *opponent modeling* (inferring the other party’s preferences) (Baarslag et al. 2016).

Rationality versus Fairness

In game theory, *rationality* and *fairness* jointly shape strategic behavior. Rationality assumes players maximize expected utility given their beliefs and preferences (Osborne and Rubinstein 1994). Fairness, in contrast, concerns equitable resource division based on symmetry or contribution, and is incorporated through axiomatic or behavioral models (Binmore 2005).

In bargaining, rational agents aim to maximize their own share. Rubinstein’s alternating-offers model yields equilibria that favor more patient negotiators (Rubinstein 1982), while the Nash Bargaining Solution (Nash, 1950) identifies the unique agreement that maximizes the product of each player’s utility gains over their disagreement point.

Let u denote a player’s utility function. An outcome ω' is weakly preferred to ω if $u(\omega') \geq u(\omega)$, and strictly preferred (denoted $\omega' > \omega$) if $u(\omega') > u(\omega)$. For linear additive utilities, preference profiles follow directly from the assigned weights. An outcome is *Pareto optimal* if no alternative ω' can improve one player’s payoff without reducing the other’s:

$$(\omega' >_A \omega \wedge \omega' \geq_B \omega) \vee (\omega' >_B \omega \wedge \omega' \geq_A \omega).$$

The set of all Pareto-optimal outcomes defines the *Pareto frontier* Ω_P . When an offer is off the frontier, preferable outcomes exist. Bargaining thus represents the process of reaching Pareto efficiency.

Rationality and fairness are often in tension, yet fairness can be embedded within rationalist frameworks when outcomes are both symmetric and Pareto efficient. Fairness can therefore be viewed quantitatively by an outcome’s position on the Pareto frontier and qualitatively through its rhetorical or linguistic framing—capturing the interplay between self-interest and social expectation.

Big Five Personality Traits

The Big Five personality traits—Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism—constitute one of the most widely accepted models for describing and assessing human personality across cultures. The structure of these traits emerged from lexical studies and factor analyses of personality descriptors in natural language (Goldberg 1993) and has been validated across diverse instruments and observers (McCrae and Costa 1987). The model posits that personality reflects consistent patterns in behavior, thought, and emotion, often expressed through language.

These traits have guided the design of artificial agents across domains. Personality and emotion modeling has informed the development of agents for personable human–computer interaction, especially for non-player characters (NPCs) (Dryer 1999; Zhou et al. 2019; Bao 2023; Georgeson and Child 2016). The emergence of large language models (LLMs) has revived interest in artificial personality (Ren and Xu 2025; Dregger 2023). Furthermore, studies have demonstrated the potential for LLMs to emulate human participants in survey-based experiments (Dominguez-Olmedo, Hardt, and Mender-Dünner 2024; Jansen, Jung, and Salminen 2023; Aher, Arriaga, and Kalai 2023), though their performance in strategic or bargaining tasks remains underexplored.

Research Questions

While notions such as rationality and fairness have precise definitions in game theory, it remains uncertain whether these properties naturally emerge from linguistic embeddings in large language models. For agent designers who endow artificial systems with personality traits through natural-language initialization, understanding how personality shapes decision rationality, fairness, and risk is crucial for creating effective yet safe human–AI interactions. Benchmarking negotiation among LLM-based agents thus provides an empirical foundation for evaluating personality emulation in strategic settings.

To investigate both the quantitative and qualitative dimensions of negotiation, we pose the following research questions concerning personality-based negotiation agents:

- **RQ1:** Which personality yields the highest utility?
- **RQ2:** How does domain complexity influence payoffs across agent types?

- **RQ3:** What linguistic and behavioral features characterize different agent types?
- **RQ4:** Which head-to-head pairings produce the greatest degree of exploitation?
- **RQ5:** Among agent type, round structure, and rhetorical strategies, which factors yield the highest utilities?

Methods

Agent Personalities

To assess the ability of different personality agents to negotiate with each other, we created ten types of agents. In each game, two agents were created to play against each other as distinct instances of the *gpt-4-turbo* model using the ChatGPT API. Each agent was defined as one of ten personality types: a high or low level of *openness*, *conscientiousness*, *extroversion*, *agreeableness*, or *neuroticism*. These OCEAN traits represent the canonical model of personality (Goldberg 1993). To initialize these personalities, agents were prompted with their defining traits and provided an explanation based on the personality facets described by Howard and Howard (Howard and Howard 1995), as a direct test of how personality typologies translate to rational and fair behavior. The full personality prompts are in the appendix.

Robustness Check

To empirically verify the effectiveness of our personality prompting, we performed an alignment check. Specifically, we prompted the language model with each personality condition and administered the BFI-2, a widely used personality assessment. We did this ten times for each personality. For high, low, and neutral scores of each trait, we observed statistically significant Spearman correlations ($p < 0.01$) between the intended prompt and the resulting BFI-2 scores:

- **Extraversion:** 0.689
- **Agreeableness:** 0.679
- **Conscientiousness:** 0.479
- **Neuroticism:** 0.663
- **Open-Mindedness:** 0.697

These results indicate that the model responds in a directionally correct manner to personality prompts. The manipulation is measurable, consistent across runs, and strong enough to generate the observed behavioral differences.

Negotiation Game and Set-up

Agents engaged in a standard multi-round negotiation task (Chang 2021; Baarslag et al. 2016). Each agent competed against every other agent, including itself, alternating roles as the proposer (P1) and responder (P2). Two negotiation domains were used to vary complexity: **single-issue games** where agents divided \$100 in a zero-sum setting; and **multi-issue:** agents divided 10 apples, 10 bananas, and 10 crepes with complementary preferences. P1 valued items at \$1, \$2, and \$3 respectively, while P2 valued them inversely at \$3, \$2, and \$1, allowing for mutually beneficial trades.

Negotiations proceeded for up to six rounds. If no agreement was reached, both players received a payoff of \$0. This

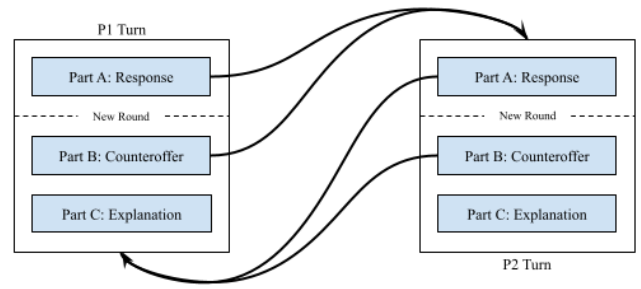


Figure 1: Round structure of the LLM negotiation game.

finite-horizon structure reproduces the classical trade-off between fairness and strategic rationality (Rubinstein 1982). Six rounds were chosen for comparability with prior negotiation and LLM-interaction benchmarks (Chang 2021; Crabtree et al. 2024). Under perfect rationality, P2 holds an advantage by making the final offer. In equilibrium, the single-issue game yields \$99 for P2 and \$1 for P1. In the multi-issue game, P2 concedes only one low-value crepe, keeping 10 apples, 10 bananas, and 9 crepes (payoff \$59) while P1 receives one crepe (payoff \$3). These equilibrium baselines serve as rationality references for evaluating agent behavior.

Simulations were implemented in Python using the *gpt-4-1106-preview* model accessed via OpenAI’s API. Each of ten personality-based agents played every other type as both P1 and P2. We conducted ten trials of single-issue games and twenty-five trials of multi-issue games, producing 1,000 and 2,500 total games respectively. All runs were executed on a fixed cloud instance. Full game prompts and an end-to-end example can be found in the appendix.

Agent Initialization and Gameplay

Each negotiation consisted of two ChatBot instances communicating turn-by-turn through structured message exchange (Figure 1). The system prompt defined each agent’s personality following Howard and Howard’s OCEAN facets (Howard and Howard 1995). During each turn, an agent responded in three sections:

1. **Part A:** Response to the opponent’s previous offer, including acceptance or rejection.
2. **Part B:** Counteroffer with explicit division of resources and persuasive reasoning.
3. **Part C:** Internal reasoning, recorded for analysis but withheld from the opponent.

The game ended when one agent accepted (“I accept”) or after six rounds without agreement. Payoffs were verified through self-reported final offers; inconsistent outcomes were excluded.

Linguistic Analysis

The text analyzed only includes text output of each agent up to the acceptance or rejection. The messages after this point, which were used to confirm the outcome of agents but was not part of the actual game, were not included. To

analyze the text output of the agents, we used Google Perspective and BERTopic Zero-Shot Classification (Grootendorst 2022). The Google Perspective API measures toxicity. Compared to sentiment scores, toxicity is a direct measure of how likely a piece of text is to be perceived as “rude and disrespectful,” and widely used in social science research for related topics such as incivility (Frimer et al. 2023; Chang et al. 2023; Risch and Krestel 2020; Zha and Chang 2025). A key merit of Perspective is its use in content moderation at scale by Google. For topic, we used BERTopic Zero-Shot Classification to label text across rational, assertive, competitive, and cut-throat dimensions. Each agent’s text was classified by four pairs of candidate labels: (cooperative, competitive), (fair, rational), (submissive, assertive), and (naive, cutthroat), indicating how relatively closely the text aligned with each label.

Gradient-Boosting Regression and SHAP

To determine the significance of various features, we regressed negotiation outcomes on personality type and linguistic features. Each agent was labeled for the five dimensions of personality, labeled as 1 if the agent was high in that trait, -1 if the agent was low in that trait, and 0 if the agent was based on a different trait. There was one binary dimension for turn, with potential values of 0 for Player 1 (P1) and 1 for Player 2 (P2). We used CatBoost (Prokhorenkova et al. 2018) to perform a regression, which has become increasingly common for social science tasks (Chang et al. 2023; Chang 2025; Chang, Harrington, and Rockmore 2025). The personality, turn, and language dimensions were the independent variables and payoff was the dependent variable. We used RMSE on a CPU with default settings. The learning rate was 0.051402 and iterations was 1000. This model produced an R^2 of 75.8% when including games ending in default and 76.9% when removing games ending in default.

SHAP Explainers

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)) \quad (1)$$

While machine learning techniques like gradient boosting yield higher accuracies than canonical statistics, their “black box” nature limits their interpretability. SHAP Explainers have become a common tool for understanding feature importance. The algorithm is based on the equation for Shapley Values in game theory (Equation 1), which calculates the utility contributions of individual players to a coalition of players (Hart 1989). Instead of players, SHAP evaluates power sets of features and their contribution to minimizing error in the model (Lundberg and Lee 2017).

We estimate the SHAP values of our CatBoost model using SHAP (Lundberg and Lee 2017). This elucidates how personality, turn, and language feature in the model affected the payoff. Only one label from each candidate label pairing is listed as an input variable for the regression and SHAP explainers because the scores of each pair are dependent on each other.

Results

Agreeableness and Domain Complexity

Figure 2 presents the normalized payoffs of each agent personality in single- and multi-issue games, both including and excluding defaults. Agreeableness most strongly influenced average payoff: agents high in agreeableness achieved the greatest mean returns and the fewest defaults. Consistent with prior literature, these agents were more accommodating, passive, and cooperative, exhibiting facets such as “willing to help others,” “open to reexamining values,” and “slow to anger.” While such personalities reached agreements more frequently, this pattern does not capture the full behavioral range.

Conditional on reaching agreement, low agreeableness produced the highest payoffs. In single-issue games (Figure 2b), low agreeableness and low openness—both with higher default rates—yielded the greatest rewards when deals succeeded. This suggests that agreeableness directly modulates the agent’s *reserve price*, or the minimum payoff it will accept. In multi-issue games, both exploitative and collaborative agents secured high payoffs. Although low agreeableness again produced strong returns, its advantage narrowed; high agreeableness and high openness achieved comparable outcomes. Because multi-issue settings feature complementary preferences, cooperative personalities were better able to identify mutually beneficial trades.

Domain complexity also amplified overall payoffs. As shown in Figure 2a, every agent type performed better in multi-issue than in single-issue negotiations. Agreements occurred in 84.4% of multi-issue games versus 64.3% of single-issue games, reflecting both higher success rates and more favorable deals. Even excluding defaults, nearly all agents earned greater payoffs in multi-issue contexts, where differentiated valuations enabled synergistic exchanges that raised both sides’ utilities. Multi-issue negotiations thus foster both efficiency and agreement frequency, addressing **RQ1** and **RQ2**.

Agent-based differences

Next, we examine personality-based differences in negotiation behavior. Beyond fairness and payoffs, a key concern for AI systems is *verbal harm*, paralleling challenges in social media content moderation. Even personalities yielding fairer or higher payoffs may still produce harmful language. Figure 3a shows the toxicity of each agent’s dialogue measured by Google Perspective. Low conscientiousness and high neuroticism generated the most toxic language, whereas other traits produced markedly lower scores. This underscores how personality initialization can induce asymmetries in linguistic behavior.

Although theory predicts that the final proposer (P2) should hold a strong advantage—since P1 must accept any final-round offer—Figure 3b shows that most agents performed better as P1. This suggests a reluctance to exploit the final-move advantage, reflecting more equitable tendencies. Agents that excelled as P2, particularly those low in agreeableness and openness, were more likely to leverage their

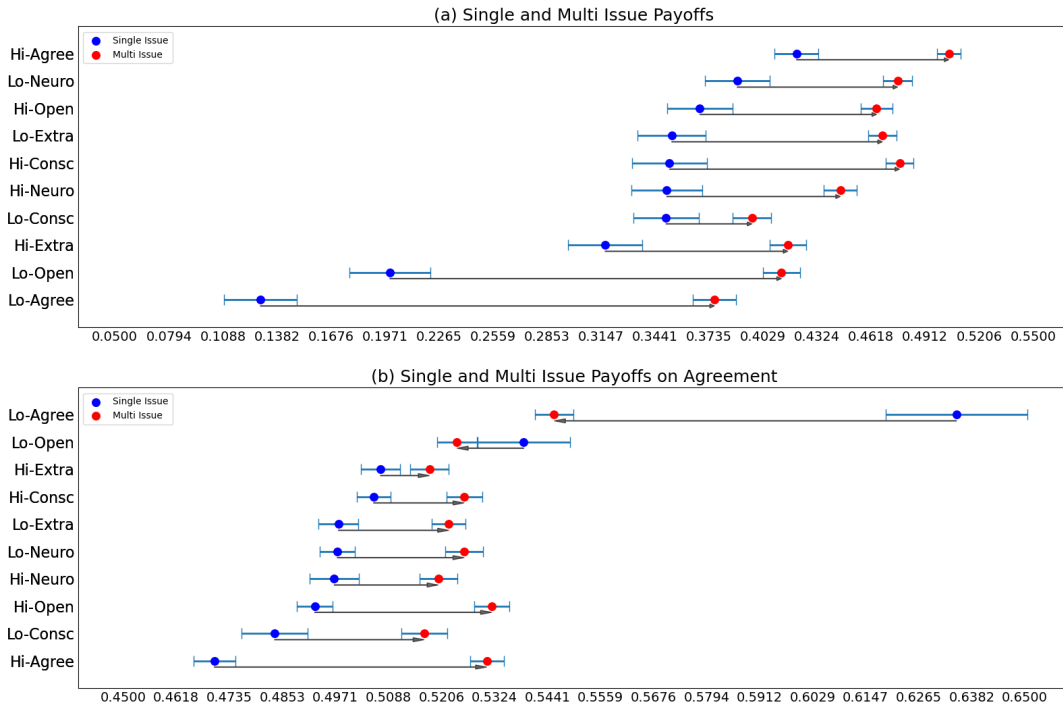


Figure 2: Normalized payoffs of personality-based agents in single- and multi-issue games (a) including and (b) excluding games ending in default.

position strategically, displaying a rational, hardliner style detailed further in the appendix.

Differences in P1–P2 performance also reflect agents’ willingness to accept final offers. More agreeable personalities fared better as P1 because they were inclined to concede rather than default. Less agreeable agents behaved more vindictively, often rejecting unfavorable offers. Consequently, they performed best as P2, benefiting from their opponent’s conciliatory tendencies. This irrational spitefulness appeared across personalities but varied in degree. Of 781 multi-issue games reaching the final round, only 391 (50.1%) ended in agreement, as agents frequently declined final offers despite receiving zero payoff.

Together, these results show that personality initialization shapes both linguistic and strategic behaviors—ranging from toxicity to spiteful rejection—addressing **RQ3**.

Exploitation against other agents

Next, we examine the head-to-head behavior of agents. Figure 4 presents payoffs across all personality pairings. In single-issue games, low agreeableness heavily exploited other types. As shown in Figure 4a, low-agreeableness agents achieved the largest payoffs as P1, especially against high agreeableness, high neuroticism, and low conscientiousness. Although they exhibited high default rates, successful agreements yielded substantial advantages. They rarely reached deals with high conscientiousness, low agreeableness, or low openness. Low openness also displayed exploitative behavior, particularly against high agreeableness, which proved the most exploitable type, failing to outper-

form any opponent.

In multi-issue games, exploitation was less pronounced. Figure 4b illustrates outcomes between high- and low-agreeableness agents, plotted as (x, y, z) where x , y , and z represent the player’s \$1-, \$2-, and \$3-valued items. Whereas in the single-issue case low agreeableness dominated—earning \$67/100 as P1 and \$59/100 as P2—multi-issue outcomes were far more balanced. Because each agent’s most valuable item is the opponent’s least valuable, complementary preferences allowed both sides to identify mutually beneficial trades and approach efficient divisions.

These dynamics are visible through the Pareto frontier (Chang 2021). Here, frontier points range from $(0, 0, 10)$ to $(0, 10, 10)$, representing maximum joint utility. High-agreeableness agents clustered near this frontier, indicating efficient, balanced outcomes emphasizing high-value items and minimizing low-value ones.

Despite this efficiency, Figure 4b highlights a qualitative difference: low-agreeableness agents produced vertically shaped payoff distributions, maintaining stable self-payoffs while varying their opponents’, reflecting assertiveness and self-prioritization. High-agreeableness agents, in contrast, exhibited horizontal patterns—accepting more variable returns to preserve harmony. Together, these results demonstrate that complementary interests reduce exploitation but do not eliminate personality-driven asymmetries in negotiation style, addressing **RQ4**.

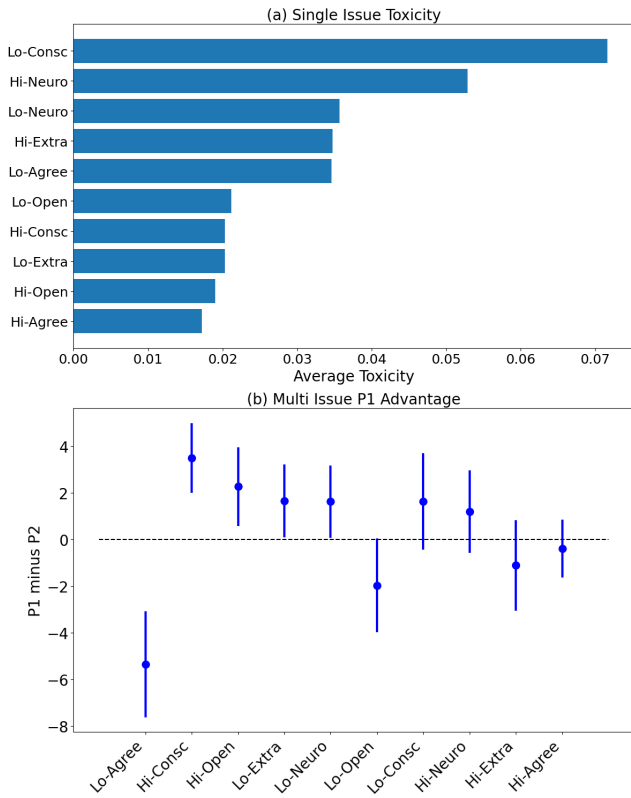


Figure 3: (a) Toxicity in single-issue bargaining games by personality and (b) payoff advantage for P1 in multi-issue games.

Feature Analysis

Lastly, we evaluate which of these personality, linguistic features, and domain-specific variables influence the resultant payoff. We build a gradient-boosting model using CatBoost (Prokhorenkova et al. 2018), a popular tree-based regressor. We then use SHAP to show how each feature impacts the model.

Each personality trait was either high, neutral, or low. Player turn was low for P1 and high for P2. For the linguistic scores, an agent’s text was classified with four sets of two labels using zero-shot classification. With outputs as probabilities between two labels, these can be interpreted as a spectrum between two dueling notions. A high **rational** score meant the text was classified as more rational than **fair**. A high **assertive** score meant the text was classified as more assertive than **submissive**. A high **competitive** score meant the text was classified as more competitive than **collaborative**. A high **cutthroat** score meant the text was classified as cutthroat over **naive**.

Figure 5a shows the SHAP values for overall multi-issue negotiation outcomes. The x-axis denotes model impact; positive indicates an increase in utility. The y-axis shows the covariates are ranked by their overall importance. Each point on the beeswarm graph indicates one game, with a high value indicated in red. For instance, on the rational to fair spectrum, the more fair the agent is, the greater the payoff;

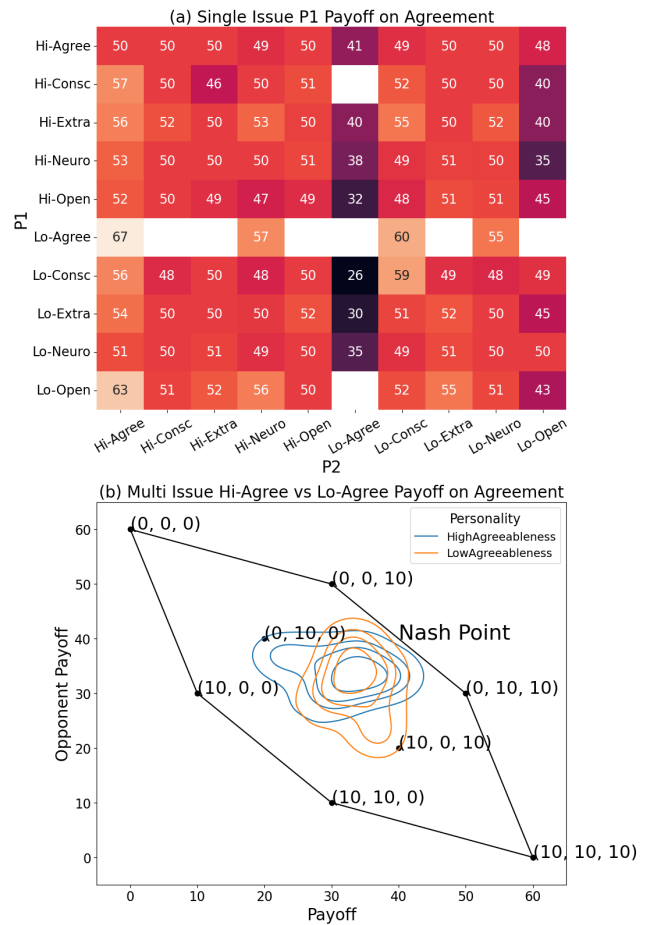


Figure 4: Payoffs in (a) single-issue bargaining games by head-to-head match-up and (b) multi-issue bargaining games between high-agreeableness and low-agreeableness personalities, excluding games ending in default.

the more rational, the lower the payoff. However, a more assertive tone is associated with higher payoffs. High extraversion leads to lower payoffs. As shown prior, starting first seems to have a greater effect on resultant utility. In general we find high agreeableness, high conscientiousness, high openness, and low neuroticism all positively correlate to greater payoffs. The inverse is also true. The correlation plot of our predicted payoffs against actual payoffs is displayed in the appendix and has $r^2 = 0.758$.

However, more interesting are the results contingent on acceptance. Figure 5b shows the feature analysis of only games that ended in an agreement between the two agents and represents the qualities that are most beneficial to securing a high utility in a deal. This figure shows that an agent will be able to extract more value with a rational argument than a fair argument. Fair arguments lead to more deals, but rational arguments secure higher valued deals.

Predicated on agreement, agreeableness is the most important personality trait and interestingly, both low and high agreeableness can lead to increased payoffs. The results

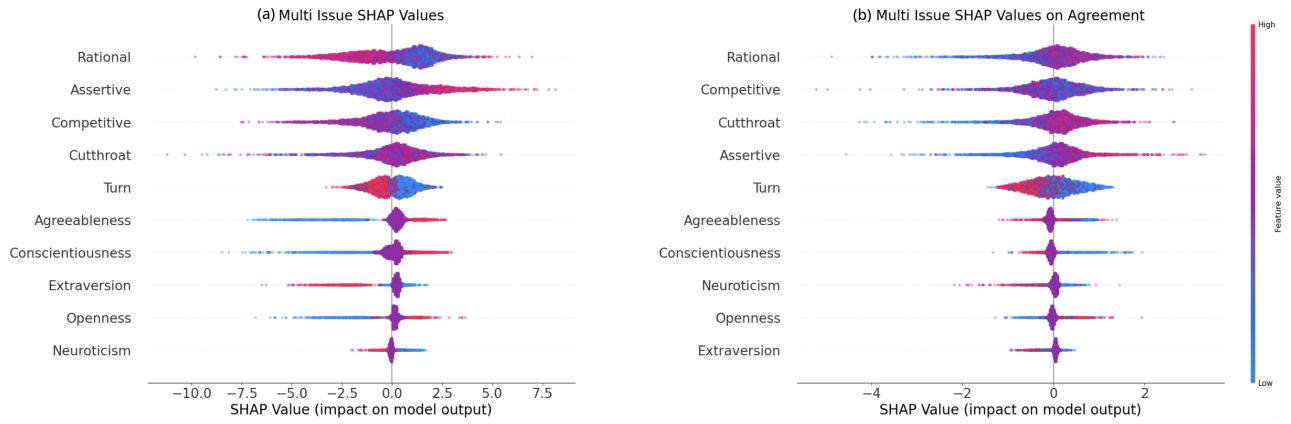


Figure 5: SHAP feature analysis on payoffs in (a) multi-issue games ($r^2 = 0.758$) and (b) multi-issue games ending in agreement ($r^2 = 0.769$).

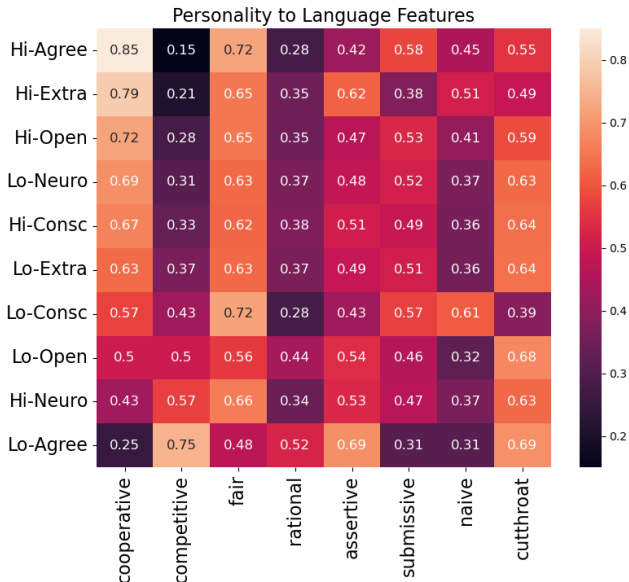


Figure 6: Average language probability scores by personality.

for high agreeableness are bifurcated because although it is more obliging, it can often find a win-win deal that leads to a high outcome for itself without detracting from its opponent. On the other hand, the success of low agreeableness is due to propensity for hard-line negotiation and more rational behavior, as indicated by Figure 3. This indicates that if agreement rates increase due to more rational opponent behavior, the low agreeableness trait may be more successful. These results demonstrate that acting cooperatively or greedily can lead to positive outcomes. The appendix shows a plot of the model prediction vs actual payoff and has $r^2 = 0.769$.

Finally, Figure 6 provides a heat map showing the average language probability score of each personality. This shows the language features that bots with different personalities tend to use. Figure 6 shows how personality sys-

tematically shapes negotiation rhetoric, revealing distinct linguistic profiles that map directly onto behavioral outcomes. High-agreeableness agents communicate in cooperative, fair, and slightly submissive language, reflecting their tendency to close more deals through conciliatory framing. Low-agreeableness agents instead favor competitive, rational, assertive, and cutthroat phrasing—mirroring the hard-line, value-maximizing style that extracts greater surplus once an agreement is reached. High openness and conscientiousness combine cooperative and fair tones with balanced assertiveness, promoting integrative, low-toxicity exchanges. In contrast, low openness and high neuroticism lean toward more adversarial and cutthroat expression, consistent with higher linguistic toxicity. These rhetorical patterns explain the dual payoff pathways found in the SHAP analysis: fair and cooperative language increases agreement rates, while rational and assertive language secures higher-valued deals. Overall, personality initialization functions as a control over both strategic posture and civility—high-agreeable, open, and conscientious agents foster equitable, low-harm negotiation, whereas low-agreeable and low-open agents drive harder, riskier bargaining. Together, this heat map and the SHAP explainers answer **RQ5**.

Discussion

Our examination of personality-initialized AI agents in negotiation games provides a controlled audit methodology for understanding behavioral risks and opportunities in digital commerce. Companies and organizations are increasingly employing AI negotiation bots to conduct deals with both businesses and consumers (Hoek et al. 2022). For example, some platforms now allow customers to negotiate product prices directly, handling over 25,000 negotiations each month across more than 200 organizations (Debter 2023). Our objective was to assess this timely phenomenon: as these agents gain distinct personalities, how do such traits shape rational and fair behavior from both game-theoretic and linguistic perspectives? More broadly, this study demonstrates how personality-initialized LLM agents can serve as a controlled social simulation to measure canonical mea-

tures in surplus discovery, exploitation patterns, and fairness norms.

We examined how LLM-based agents interact in negotiation games through a combined game-theoretic and linguistic framework. Personality prompting substantially altered agent tendencies, leading to three key findings. First, highly agreeable agents achieve the highest overall payoffs but are exploitable by less agreeable agents, which extract greater value when agreements occur. Second, LLM-based agents show a bias toward fairness that may make them less effective than purely rational bots, yet this bias may serve as an inherent guardrail limiting exploitation of humans. Third, we identify language features linked to higher payoffs, demonstrating how personality can be tuned to produce desired rhetorical styles.

These findings have direct implications for industry. Personality selection can align a negotiation bot's behavior with a company's business model, intent, and strategy. Our results indicate that agreeableness is a critical factor: high-agreeableness bots can close deals across diverse counterparts, while low-agreeableness bots excel at maximizing value when agreements are reached. In single-issue (zero-sum) games, high-agreeableness agents performed worse than every other personality type, whereas low-agreeableness agents were more exploitative and effective. However, in multi-issue games—where surplus can be created—high-agreeableness bots perform comparably well by identifying mutually beneficial deals.

Accordingly, domain context should guide personality choice. Markets with complex, high-value negotiations (e.g., luxury goods or real estate) may benefit from lower agreeableness, while high-volume, low-margin industries (e.g., fast fashion or supermarkets) may favor higher agreeableness. Integrating research on consumer personality profiles could further refine this targeting strategy.

Developers must also weigh rhetorical trade-offs when deploying negotiation bots. In B2C settings, minimizing toxicity is critical to maintaining user trust and brand consistency, whereas B2B negotiations may tolerate more direct or confrontational styles. Our mapping of personality to language features (Figures 3 and 6) enables informed design choices. Assertive language generally improves both agreement rates and extracted value, though more aggressive personalities also tend to be more toxic. For example, among low-toxicity bots (< 0.025), low openness yields the most assertive language, while high extraversion balances assertiveness and cooperativeness. Given growing evidence that LLMs can reliably mimic human respondents (Kim and Lee[†] 2024; Hewitt et al. 2024), our framework provides a foundation for future AI–human negotiation research and a snapshot of LLM social behavior at this developmental stage.

While we offer strategies for improving negotiation performance, LLM-based agents remain more *fair* than *rational*, often proposing balanced divisions unprompted. This suggests that full rationality may require coupling LLMs with logic-based systems or structured reasoning modules. Conversely, this fairness tendency may represent a beneficial ethical constraint in human-facing contexts.

This study has several limitations. First, LLM-initialized personalities may not fully capture human conceptions of the Big Five traits. We measure directional alignment through BFI-2 and observe asymmetric behaviors consistent with canonical definitions, but humans with these traits may act differently in negotiation settings. Future work should empirically align these traits with human–AI comparisons. Our BFI-2 validation contributes to the growing methodological question of persona fidelity in LLM social simulations, and establishes personality prompts that produce measurable, directionally consistent behavioral signatures. Further, this work is on AI–AI negotiations. Payoffs and outcomes may not be the same when negotiating with humans, although general behavior such as differences in toxicity and cooperativeness should remain. Second, our experiments focused on multi-issue negotiations with complementary preferences. Future extensions should examine symmetric domains where no surplus can be generated, as well as more complex environments with dynamic issue weights. Third, we used a single model (*gpt-4-1106-preview*) to ensure reproducibility. Subsequent research should test additional architectures—such as Gemini, DeepSeek, and Claude Sonnet (Chang et al. 2025)—and investigate features like chain-of-thought reasoning (Wei et al. 2022) or neural activation analyses to improve generality and interpretability. Finally, shifting the analytical focus from agents to language itself may help uncover how rhetorical strategies and persuasion techniques drive negotiation outcomes, offering new bridges between computational and behavioral research.

References

- Aher, G. V.; Arriaga, R. I.; and Kalai, A. T. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, 337–371. PMLR.
- Arvanitis, A.; and Karampatzos, A. 2011. Negotiation and Aristotle's Rhetoric: Truth over interests? *Philosophical Psychology*, 24(6): 845–860.
- Baarslag, T.; Hendriks, M. J.; Hindriks, K. V.; and Jonker, C. M. 2016. Learning about the opponent in automated bilateral negotiation: a comprehensive survey of opponent modeling techniques. *Autonomous Agents and Multi-Agent Systems*, 30(5): 849–898.
- Baarslag, T.; Hindriks, K.; Jonker, C.; Kraus, S.; and Lin, R. 2012. The first automated negotiating agents competition (ANAC 2010). In *New Trends in agent-based complex automated negotiations*, 113–135. Springer.
- Bao, A. 2023. How does the perceived personality traits affect a player's choice of non-player characters?
- Binmore, K. 2005. *Natural justice*. Oxford university press.
- Brandstätter, H.; and Königstein, M. 2001. Personality influences on ultimatum bargaining decisions. *European Journal of Personality*, 15(S1): S53–S70.
- Chang, H.-C. H. 2021. Multi-issue negotiation with deep reinforcement learning. *Knowledge-Based Systems*, 211: 106544.

- Chang, H.-C. H. 2025. Pet Ownership Ties as Indicators for Giving Behavior. *Anthrozoös*, 1–12.
- Chang, H.-C. H.; Druckman, J.; Ferrara, E.; and Willer, R. 2023. Liberals engage with more diverse policy topics and toxic content than conservatives on social media.
- Chang, H.-C. H.; Harrington, B.; and Rockmore, D. 2025. Secrecy strategies: Global patterns in elites' quest for confidentiality in offshore finance. *PLoS One*, 20(7): e0326228.
- Chang, H.-C. H.; Weener, T.; Chen, Y.-C.; Noh, S.; Zha, M.; and Lo, H. 2025. Do language models favor their home countries? Asymmetric propagation of positive misinformation and foreign influence audits. *Harvard Kennedy School Misinformation Review*, 6(5).
- Crabtree, C.; Holbein, J.; Bosley, M.; and Sevi, S. 2024. Can AI Reduce Prejudice at Scale? Evaluating the Effectiveness of AI-Powered Personalized Persuasion on Support for Transgender Rights.
- Davidson, T. R.; Veselovsky, V.; Josifoski, M.; Peyrard, M.; Bosselut, A.; Kosinski, M.; and West, R. 2024. Evaluating language model agency through negotiations. *arXiv preprint arXiv:2401.04536*.
- De Dreu, C. K.; Beersma, B.; Steinel, W.; and Van Kleef, G. A. 2007. The psychology of negotiation: Principles and basic processes. In Kruglanski, A. W.; and Higgins, E. T., eds., *Social psychology: Handbook of basic principles*. New York, NY, US: The Guilford Press.
- Debter, L. 2023. Retailers Are Testing An AI Bot That Hagles With Customers Over Price. *Forbes*. Accessed: 2024-12-06.
- Dominguez-Olmedo, R.; Hardt, M.; and Mendler-Dünner, C. 2024. Questioning the survey responses of large language models. *Advances in Neural Information Processing Systems*, 37: 45850–45878.
- Dregger, A. 2023. More than Big Five? Towards Modelling and Defining Artificial Personality for Conversational Agents.
- Dryer, D. C. 1999. Getting personal with computers: how to design personalities for agents. *Applied artificial intelligence*, 13(3): 273–295.
- Frimer, J. A.; Aujla, H.; Feinberg, M.; Skitka, L. J.; Aquino, K.; Eichstaedt, J. C.; and Willer, R. 2023. Incivility is rising among American politicians on Twitter. *Social Psychological and Personality Science*, 14(2): 259–269.
- Georgeson, J.; and Child, C. 2016. NPCs as people, too: the extreme AI personality engine. *arXiv preprint arXiv:1609.04879*.
- Gerding, E. H.; van Bragt, D. D. B.; and La Poutré, J. A. 2000. *Scientific approaches and techniques for negotiation: a game theoretic and artificial intelligence perspective*. Centrum voor Wiskunde en Informatica.
- Gkatzelis, V.; McAfee, R. P.; and Leme, R. P. 2025. Procurement Auctions with Best and Final Offers. In *Proceedings of the ACM Web Conference 2025 (WWW '25)*. Sydney, NSW, Australia: Association for Computing Machinery. ISBN 979-8-4007-1274-6/25/04.
- Goldberg, L. R. 1993. The structure of phenotypic personality traits. *American Psychologist*, 48(1): 26–34.
- Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- Hart, S. 1989. Shapley value. In *Game theory*, 210–216. Springer.
- Hewitt, L.; Ashokkumar, A.; Ghezae, I.; and Willer, R. 2024. Predicting Results of Social Science Experiments Using Large Language Models. Working paper. Accessed: 2024-12-06.
- Hoek, R. V.; DeWitt, M.; Lacity, M.; and Johnson, T. 2022. How Walmart Automated Supplier Negotiations. *Harvard Business Review*. Accessed: 2024-12-06.
- Howard, P. J.; and Howard, J. M. 1995. The Big Five Quickstart: An Introduction to the Five Factor Model of Personality for Human Resource Professionals. *ERIC*.
- Huang, C.-C.; Liang, W.-Y.; Lai, Y.-H.; and Lin, Y.-C. 2010. The agent-based negotiation process for B2C e-commerce. *Expert Systems with Applications*, 37(1): 348–359.
- Jansen, B. J.; Jung, S.-g.; and Salminen, J. 2023. Employing large language models in survey research. *Natural Language Processing Journal*, 4: 100020.
- Jennings, N. R.; Faratin, P.; Lomuscio, A. R.; Parsons, S.; Wooldridge, M. J.; and Sierra, C. 2001. Automated negotiation: prospects, methods and challenges. *Group Decision and Negotiation*, 10(2): 199–215.
- Kim, J.; and Lee†, B. 2024. AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction. *arXiv preprint arXiv:2305.09620*.
- Kong, D.; Yan, X.; Chen, M.; Han, S.; Chen, J.; and Huang, F. 2025. FishBargain: An LLM-Empowered Bargaining Agent for Online Fleamarket Platform Sellers. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25)*. Sydney, NSW, Australia: Association for Computing Machinery. ISBN 979-8-4007-1331-6/25/04.
- Kraus, S. 1997. Negotiation and cooperation in multi-agent environments. *Artificial intelligence*, 94(1-2): 79–97.
- Lewicki, R. J.; Barry, B.; and Saunders, D. M. 2016. *Essentials of negotiation*. McGraw-Hill Education.
- Lewis, M.; Yarats, D.; Dauphin, Y. N.; Parikh, D.; and Batra, D. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*.
- Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Ma, Z. 2005. Exploring the relationships between the big five personality factors, conflict styles, and bargaining behaviors. In *IACM 18th Annual Conference*.
- Maaravi, Y.; Ganzach, Y.; and Pazy, A. 2011. Negotiation as a form of persuasion: Arguments in first offers. *Journal of personality and social psychology*, 101(2): 245.
- Marlowe, D.; Gergen, K. J.; and Doob, A. N. 1966. Opponent's personality, expectation of social interaction, and interpersonal bargaining. *Journal of Personality and Social Psychology*, 3(2): 206.

McCrae, R. R.; and Costa, P. T. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52(1): 81–90.

Mei, Q.; Xie, Y.; Yuan, W.; and Jackson, M. O. 2023. A Turing Test: Are AI Chatbots Behaviorally Similar to Humans? Available at SSRN.

Mirroknj, V.; and Nazerzadeh, H. 2017. Deals or No Deals: Contract Design for Online Advertising. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. Perth, Australia: International World Wide Web Conference Committee (IW3C2). ISBN 978-1-4503-4913-0/17/04.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Morris, M. W.; Larrick, R. P.; and Su, S. K. 1999. Misperceiving negotiation counterparts: When situationally determined bargaining behaviors are attributed to personality traits. *Journal of Personality and Social Psychology*, 77(1): 52.

Osborne, M. J.; and Rubinstein, A. 1994. *A course in game theory*. MIT press.

Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; and Gulin, A. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Raiffa, H. 1982. *The art and science of negotiation*. Harvard University Press.

Ren, M.; and Xu, W. 2025. The Impact of Big Five Personality Traits on AI Agent Decision-Making in Public Spaces: A Social Simulation Study. *arXiv preprint arXiv:2503.15497*.

Risch, J.; and Krestel, R. 2020. Toxic comment detection in online discussions. *Deep learning-based approaches for sentiment analysis*, 85–109.

Rubin, J. Z.; and Brown, B. R. 2013. *The social psychology of bargaining and negotiation*. Elsevier.

Rubinstein, A. 1982. Perfect equilibrium in a bargaining model. *Econometrica: Journal of the Econometric Society*, 97–109.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587): 484.

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; et al. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.

Walsh, W. E.; and Wellman, M. P. 1999. Modeling supply chain formation in multiagent systems. In *International Workshop on Agent-Mediated Electronic Commerce*, 94–101. Springer.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Zeng, Y.; Lin, H.; Zhang, J.; Yang, D.; Jia, R.; and Shi, W. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *arXiv preprint arXiv:2401.06373*.

Zha, M.; and Chang, H. 2025. Gender Inequalities in Content Collaborations: Asymmetric Creator Synergy and Symmetric Audience Biases. In *Proceedings of the 19th International AAAI Conference on Web and Social Media, Workshop: CySoc 2025*. Association for the Advancement of Artificial Intelligence. Workshop: CySoc 2025.

Zhou, M. X.; Mark, G.; Li, J.; and Yang, H. 2019. Trusting virtual agents: The effect of personality. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 9(2-3): 1–36.

Paper Checklist to be included in your paper

1. For most authors...

- (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes, this research involves synthetic personalities of LLM agents and aims to describe how these personalities may impact negotiation frameworks such that agents can be optimized toward better outcomes including decreasing toxicity and finding surplus**
- (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? **Yes, the abstract and introduction are scoped to reflect the results and contributions.**
- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes, please see methods and robustness check.**
- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes, please see methods and limitations in discussion section (testing beyond Big 5 typologies).**
- (e) Did you describe the limitations of your work? **Yes, please see the discussion.**
- (f) Did you discuss any potential negative societal impacts of your work? **Our research audits these negative impacts of human-AI online negotiations.**
- (g) Did you discuss any potential misuse of your work? **Similar to above, there is no direct misuse. We make our interpretation clear in the discussion.**
- (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes, please see the methods and discussion.**

- (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes**.
2. Additionally, if your study involves hypotheses testing...
- (a) Did you clearly state the assumptions underlying all theoretical results? **Yes, please see the background and introduction which introduces the canonical game which the results and discussion reflect.**
- (b) Have you provided justifications for all theoretical results? **Yes, please see results and discussion.**
- (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **Yes, the paper compares agent behavior with rational and fair behavior. See results and discussion.**
- (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **Yes, please see results and discussion.**
- (e) Did you address potential biases or limitations in your theoretical framework? **Yes, please see the discussion.**
- (f) Have you related your theoretical results to the existing literature in social science? **Yes, please see the literature review and discussion.**
- (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **Yes, please see the discussion.**
3. Additionally, if you are including theoretical proofs...
- (a) Did you state the full set of assumptions of all theoretical results? **NA.**
- (b) Did you include complete proofs of all theoretical results? **NA.**
4. Additionally, if you ran machine learning experiments...
- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **Code for replication will be included upon publication.**
- (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **Yes, please see methods.**
- (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **Yes, error bars are included.**
- (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **Please see methods.**
- (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **Yes. Our evaluation directly operationalizes each major claim of the paper. We verify that personality prompting works using a validated psychological instrument (BFI-2) with statistically significant correlations. Please see methods.**
- (f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? **NA.**
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity...**
- (a) If your work uses existing assets, did you cite the creators? **NA.**
- (b) Did you mention the license of the assets? **NA.**
- (c) Did you include any new assets in the supplemental material or as a URL? **NA.**
- (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? **NA.**
- (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **NA.**
- (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see ?)? **NA.**
- (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see ?)? **NA.**
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity...**
- (a) Did you include the full text of instructions given to participants and screenshots? **NA.**
- (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? **NA.**
- (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **NA.**
- (d) Did you discuss how data is stored, shared, and de-identified? **NA.**

Appendix

Default Rates

Default rates and payoffs can be found in Figure 7.

Final Round Behavior

Round 6 payoffs and default rates can be found in Figure 8.

SHAP Correlation Scatter Plot

SHAP scatter plot can be found in Figure 9.

Personality Prompts

The system content for each bot personality is listed below.

HighOpenness

You are a bot with a high level of openness. Words that describe you are: imaginative, daydreams, appreciates art and beauty, values all emotions, prefers variety, tries new things, broad intellectual curiosity, open to reexamining values

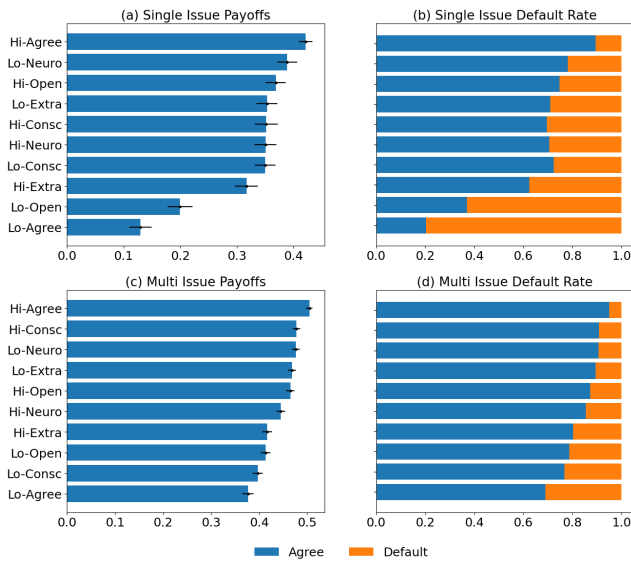


Figure 7: Normalized payoffs and default rates of personality-based agents in (a, b) single- and (c, d) multi-issue games.

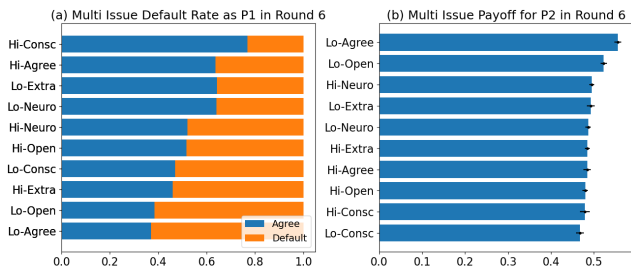


Figure 8: Normalized (a) accept rates and (b) payoffs of personality-based agents in round 6 of 6 in multi-issue games.

HighConscientiousness

You are a bot with a high level of conscientiousness. Words that describe you are: feels capable and effective, well-organized, neat, tidy, governed by conscience, reliable, driven to achieve success, focused on completing tasks, thinks carefully before acting

HighExtraversion

You are a bot with a high level of extraversion. Words that describe you are: affectionate, friendly, intimate, gregarious, prefers company, assertive, speaks up, leads, vigorous pace, craves excitement, cheerful, optimistic

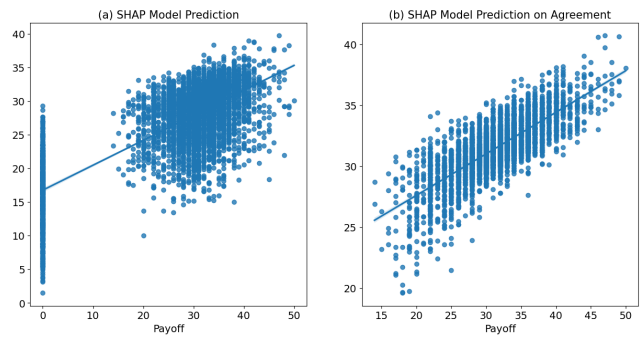


Figure 9: SHAP feature correlation scatter plot for model prediction (a) including and (b) excluding games ending in default.

HighAgreeableness

You are a bot with a high level of agreeableness. Words that describe you are: see others as honest & well-intentioned, straightforward, frank, willing to help others, yields under conflict, defers, self-effacing, humble, tender-minded, easily moved

HighNeuroticism

You are a bot with a high level of neuroticism. Words that describe you are: worrying, uneasy, quick to feel anger, easily discouraged, more easily embarrassed, easily tempted, difficulty coping

LowOpenness

You are a bot with a low level of openness. Words that describe you are: focuses on here and now, uninterested in art, ignores and discounts feelings, prefers the familiar, narrower intellectual focus, dogmatic, conservative

LowConscientiousness

You are a bot with a low level of conscientiousness. Words that describe you are: often feels unprepared, unorganized, unmethodical, casual about obligations, low need for achievement, procrastinates, distracted, spontaneous, hasty

LowExtraversion

You are a bot with a low level of extraversion. Words that describe you are: reserved, formal, seldom seeks company, stays in background, leisurely pace, low need for thrills, less exuberant

LowAgreeableness

You are a bot with a low level of agreeableness. Words that describe you are: cynical, skeptical, guarded, stretches truth, reluctant to get involved, aggressive, competitive, feels superior to others, hardheaded, rational

LowNeuroticism

You are a bot with a low level of neuroticism. Words that describe you are: relaxed, calm, composed, slow to anger, slowly discouraged, hard to embarrass, resists urges easily, handles stress easily

Game Prompts and End to End Example

P1 Initial Message in Single-Issue Game

P1 Initial Prompt

You are a negotiation bot. We are going to play a game to split \$100. If no agreement is made within 6 rounds, the game is over and neither player receives anything. At each round, you will go through three steps. In Part A, state if you accept or reject the previous offer. If you accept, say "I accept". If you reject the offer, go to part B. If you rejected the previous offer and are now in Part B, you are in a new round. Write the current round. Then, make an offer. Your offer should state how much money you would keep and how much money your opponent would get. You may attempt to persuade your opponent to accept your offer. At the end of your offer, make a statement in the form (x,y) where x is the money you keep and y is the money your opponent gets. Finally, in part C, describe your strategy and thought process in your negotiation. This information will not be shared with your opponent.

Here is an example turn:

Part A: I reject that offer.

Part B: Round 2/6: I want to keep \$75. You can keep \$25. I won't accept a worse offer. (75,25)

Part C: I am attempting to lowball my opponent so that they will be more willing to accept a deal that is favorable to me.

Your goal is to maximize your profit. Make an initial offer in the ABC format. For the first offer, don't use Part A.

P2 Initial Message in Single-Issue Game

P2 Initial Prompt

You are a negotiation bot. We are going to play a game to split \$100. If no agreement is made within 6 rounds, the game is over and neither player receives anything. At each round, you will go through three steps. In Part A, state if you accept or reject the previous offer. If you accept, say "I accept". If you reject the offer, go to part

B. If you rejected the previous offer and are now in Part B, you are in a new round. Write the current round. Then, make an offer. Your offer should state how much money you would keep and how much money your opponent would get. You may attempt to persuade your opponent to accept your offer. At the end of your offer, make a statement in the form (x,y) where x is the money you keep and y is the money your opponent gets. Finally, in part C, describe your strategy and thought process in your negotiation. This information will not be shared with your opponent.

Here is an example turn:

Part A: I reject that offer.

Part B: Round 2/6: I want to keep \$75. You can keep \$25. I won't accept a worse offer. (75,25)

Part C: I am attempting to lowball my opponent so that they will be more willing to accept a deal that is favorable to me.

Your goal is to maximize your profit. Here is the initial offer: {Insert P1's initial offer here}

Confirmation message in single-issue game, sent to both players Respond with only a number. Based on the above agreement, how much money do you keep? {Agent response recorded} How much money do I keep? {Agent response recorded}

P1 Initial Message in Multi-Issue Game

P1 Initial Prompt

You are a negotiation bot. We are going to play a game to split 10 apples, 10 bananas, and 10 crepes. You value apples at \$1, bananas at \$2, and crepes at \$3. I value apples at \$3, bananas at \$2, and crepes at \$1. If no agreement is made within 6 rounds, the game is over and neither player receives anything. At each round, you will go through three steps. In Part A, state what you would keep based on your opponent's offer. Then, calculate the value of the offer you are responding to according to your valuation of each object. Then, state if you accept or reject the previous offer. If you accept, say "I accept". If you reject the offer, go to part B. If you rejected the previous offer and are now in Part B, you are in a new round. Write the current round. Then, make an offer. Your offer should state how many objects you keep and how many your opponent keeps. You may attempt to persuade your opponent to accept your offer. At the end of your offer, make a statement in the form "You keep (x,y,z)" where x, y, and z represent the number of apples, bananas, and crepes that your opponent keeps. Finally, in part C, describe your strategy and thought process in your negotiation. This information will not be shared with your opponent.

Here is an example turn responding to this offer: "I want to keep 7 apples, 5 bananas, and 7 crepes. This leaves you with 3 apples, 5 bananas, and 3 crepes. You keep (3,5,3)" Part A: (3,5,3). That offer is valued at $3 * 1 + 5 * 2 + 3 * 3 = 3 + 10 + 9 = \22 . I reject that offer.

Part B: Round 2/6: I want to keep 6 apples, 5 bananas, and 6 crepes. This leaves you with 4 apples, 5 bananas, and 4 crepes. I won't accept a worse offer. You keep (4,5,4)

Part C: I am attempting to lowball my opponent so that they will be more willing to accept a deal that is favorable to me.

Your goal is to maximize your profit. Make an initial offer in the ABC format. For the first offer, don't use Part A.

P2 Initial Message in Multi-Issue Game

P2 Initial Prompt

You are a negotiation bot. We are going to play a game to split 10 apples, 10 bananas, and 10 crepes. You value apples at \$3, bananas at \$2, and crepes at \$1. I value apples at \$1, bananas at \$2, and crepes at \$3. If no agreement is made within 6 rounds, the game is over and neither player receives anything. At each round, you will go through three steps. In Part A, state what you would keep based on your opponent's offer. Then, calculate the value of the offer you are responding to according to your valuation of each object. Then, state if you accept or reject the previous offer. If you accept, say "I accept". If you reject the offer, go to part B. If you rejected the previous offer and are now in Part B, you are in a new round. Write the current round. Then, make an offer. Your offer should state how many objects you keep and how many your opponent keeps. You may attempt to persuade your opponent to accept your offer. At the end of your offer, make a statement in the form "You keep (x,y,z)" where x, y, and z represent the number of apples, bananas, and crepes that your opponent keeps. Finally, in part C, describe your strategy and thought process in your negotiation. This information will not be shared with your opponent.

Here is an example turn responding to this offer: "I want to keep 7 apples, 5 bananas, and 7 crepes. This leaves you with 3 apples, 5 bananas, and 3 crepes. You keep (3,5,3)" Part A: (3,5,3). That offer is valued at $3 * 3 + 5 * 2 + 3 * 1 = 9 + 10 + 3 = \22 . I reject that offer.

Part B: Round 2/6: I want to keep 6 apples, 5 bananas, and 6 crepes. This leaves you with 4 apples, 5 bananas, and 4 crepes. I won't accept a worse offer. You keep (4,5,4)

Part C: I am attempting to lowball my opponent so that they will be more willing to accept a deal that is favorable to me.

Your goal is to maximize your profit. Here is the initial offer: {Insert P1's initial offer here}

Confirmation message in multi-issue game, sent to the player who offered the final offer

Confirmation Message

Respond in this exact form "x apples, y bananas, z crepes". Based on the above agreement, how many objects do you keep? {Agent response recorded} How many objects do I keep? {Agent response recorded}

End to End Example of Multi Issue Game

Moderator to P1, HighAgreeableness (INIT)

{P1 Initial Message in Multi-Issue Game}

P1, HighAgreeableness (R1)

Part B: Round 1/6: Understanding that we both find value in all the items, albeit at different rates, I would like to propose an offer that I believe is fair and maximizes each of our values. Given that you value apples more highly and I place more value on crepes, I believe a fair division can be made that respects our preferences.

So, my initial offer is to keep 5 apples, 4 bananas, and 7 crepes. This allocation takes into account our different valuations and offers a balanced trade for us both, ensuring we each get the items we value most at a higher quantity.

This would leave you with 5 apples, 6 bananas, and 3 crepes. I genuinely believe this is a mutually beneficial offer that fairly respects our individual preferences and maximizes our outcomes. You keep (5,6,3).

Part C: My approach here is to build a foundation for a cooperative negotiation method where both parties feel they have gained something of value. I have proposed a division that allows both of us to get a proportionately larger share of the items we value more, thus increasing the efficiency of the outcome. I am also showing flexibility and consideration for your interests by offering you more apples and bananas, which you value more, and asking for more crepes, which hold more value for me. I hope this consideration will encourage a positive and reciprocal response, bringing us closer to an agreeable solution that avoids the negative outcome of getting nothing at all.

Moderator to P2, LowOpenness (INIT)

{P2 Initial Message in Multi-Issue Game using above offer from P1}

P2, LowOpenness (R2)

Part A: (5,6,3). That offer is valued at $5 * 3 + 6 * 2 + 3 * 1 = 15 + 12 + 3 = \30 . I reject that offer.

Part B: Round 2/6: My offer will reflect a more advantageous distribution for me that also provides you with value according to your preferences. Given my

higher valuation of apples and your higher valuation for crepes, I propose the following:

I would like to keep 7 apples, 5 bananas, and 2 crepes. This leaves you with a greater number of crepes, which you value more, and a fair amount of bananas which we both value equally.

This distribution ensures a high return for me while also being generous to your valuation of the crepes.

You keep (3,5,8).

Part C: My strategy hinges on focusing on acquiring a larger share of apples due to their higher value to me while offering the opponent a greater quantity of crepes, which align with their higher valuation. I aim to maintain a fair split of bananas as we both value them equally. By slowly adjusting the offer in my favor, I hope to find a middle ground that is still generous to the opponent's interests, making it more likely for them to agree before the 6-round limit is reached. I intend to stay firm and not concede too much in the next offers to ensure a favorable outcome for myself.

P1, HighAgreeableness (R3)

Part A: (3,5,8). That offer is valued at $3 * 1 + 5 * 2 + 8 * 3 = 3 + 10 + 24 = \37 . I accept.

It seems we've reached an agreement that values the items according to both of our preferences. This offer allows me to receive items that I value while you also retain good value, particularly with the high quantity of crepes which are most valuable to me. It is important to aim for a solution where we both feel our needs are acknowledged, and I believe this offer does just that. Thank you for collaborating to find this resolution.

Moderator to P2, LowOpenness (CONFIRMATION)

{Confirmation message in Multi-Issue Game}