HEIDELBERG UNIVERSITY
DEPARTMENT OF ECONOMICS

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

When Artificial Minds Negotiate: Dark Personality and the Ultimatum Game in Large Language Models

Vinícius Ferraz

Tamas Olah

Ratin Sazedul

Robert Schmidt

Christiane Schwieren

AWI DISCUSSION PAPER SERIES NO. 768
November 2025

When Artificial Minds Negotiate: Dark Personality and the Ultimatum Game in Large Language Models

Vinícius Ferraz[1]*, Tamas Olah[2], Ratin Sazedul[3], Robert Schmidt [4], and Christiane Schwieren[3]

[1] Institute of Management, Karlsruhe Institute of Technology (KIT) - Karlsruhe, Germany
[2]Institute of World Economy and International Relations, University of Debrecen - Debrecen, Hungary
[3]Alfred-Weber Institute for Economics, Heidelberg University - Heidelberg, Germany
[4]Deutsche Bundesbank - Frankfurt, Germany

Abstract

We investigate if Large Language Models (LLMs) exhibit personality-driven strategic behavior in the Ultimatum Game by manipulating Dark Factor of Personality (D-Factor) profiles via standardized prompts. Across 400k decisions from 17 open-source models and 4,166 human benchmarks, we test whether LLMs playing the proposer and responder roles exhibit systematic behavioral shifts across five D-Factor levels (from least to most selfish). The proposer role exhibited strong monotonic declines in fair offers from 91% (D1) to 17% (D5), mirroring human patterns but with 34% steeper gradients, indicating hypersensitivity to personality prompts. Responders diverged sharply: where humans became more punitive at higher D-levels, LLMs maintained high acceptance rates (75-92%) with weak or reversed D-Factor sensitivity, failing to reproduce reciprocity-punishment dynamics. These role-specific patterns align with strong-weak situation accounts—personality matters when incentives are ambiguous (proposers) but is muted when contingent (responders). Cross-model heterogeneity was substantial: Models exhibiting the closest alignment with human behavior, according to composite similarity scores (integrating prosocial rates, D-Factor correlations, and odds ratios), were dolphin3, deepseek_1.5b, and llama3.2 (0.74-0.85), while others exhibited extreme or non-variable behavior. Temperature settings (0.2 vs. 0.8) exerted minimal influence. We interpret these patterns as prompt-driven regularities rather than genuine motivational processes, suggesting LLMs can approximate but not fully replicate human strategic behavior in social dilemmas.

1 Introduction

Understanding how agents behave strategically in economic games, such as the Ultimatum Game (Güth et al., 1982), is central to behavioral economics and increasingly relevant for AI-mediated social interactions. Large Language Models (LLMs) are now widely used as simulated agents in social-science research and strategic games, offering a scalable way to probe human-like decision patterns (Argyle et al., 2023; Horton, 2023; Brookins et al., 2024; Akata et al., 2025). Yet, despite growing interest, there remains a lack of systematic tests linking LLM behavior to a factor influencing human behavior that is receiving increased attention from economists and other decision researchers: validated personality constructs. Benchmarking LLMs against human data not only provides insights into model reliability but also reveals the extent to which artificial agents capture underlying cognitive and motivational dynamics.

Beyond social-science simulations, understanding LLM behavior in such settings is also increasingly important for agentic AI systems that act, plan, and negotiate on users' behalf across digital environments (Hagendorff et al., 2024). As recent work has highlighted, population-level coordination and convention formation among interacting LLMs can display genuinely emergent dynamics rather than mere reproductions of training data (Ashery et al., 2025). These findings underscore the importance of analyzing how strategic reasoning, fairness, and social conventions emerge in artificial agents—issues that are central not only to behavioral economics but to the design of autonomous AI systems more broadly.

In this study, we contribute to this emerging line of inquiry by systematically examining whether LLMs exhibit consistent, human-like strategic behavior when prompted with validated personality constructs in the Ultimatum Game, and by benchmarking their responses against large-scale human data. We aim to assess the fidelity, consistency, and limitations of LLMs as stand-ins for human agents in strategic contexts.

*Corresponding author: vinicius.ferraz@partner.kit.edu

In that regard, the Ultimatum Game provides a particularly challenging testbed because outcomes reflect a complex interaction of fairness preferences, strategic reasoning, and willingness to punish unfairness. This complexity contrasts with simpler allocation tasks such as the Dictator Game (see Kahneman et al. (1986)), where behavior is more directly linked to distributive preferences. Introducing the Dark Factor of Personality (D-Factor)—a latent trait associated with selfish and exploitative tendencies—further raises the stakes: while costly and difficult to measure in humans due to social desirability considerations, it directly modulates fairness (Moshagen, Hilbig, et al., 2018). However, only recently have economists started to study D-Factor-related behavior more systematically. Testing whether LLMs exhibit systematic D-Factor effects, therefore, provides an opportunity to better understand the ability of artificial agents to replicate subtle and underexplored dimensions of human strategic behavior. It also opens a methodological path for evaluating how prompt engineering can be used to induce personality-driven variation in AI systems, offering a complementary approach to human experiments that are often influenced by social desirability and reputational concerns.

We investigate this intersection by pairing the Ultimatum Game with the Dark Factor of Personality (D-Factor)—the latent core of aversive traits (Moshagen, Zettler, et al., 2020)—and benchmarking it against the only available human evidence by (Hilbig and Thielmann, 2025). Their findings show that higher D-Factor scores predict less fair proposer behavior and greater acceptance of unfair offers: individuals high in D were less likely to choose a fair 50:50 split and more willing to accept an 80:20 split, reflecting heightened self-interest and reduced punishment of unfairness. Our experimental design mirrors their implementation, enabling direct cross-species comparisons of fairness and punishment motives between humans and LLM-based agents.

Building on this foundation, we measure the extent to which LLMs—when assigned D-Factor levels via prompt-based personality profiles—exhibit proposer and responder behaviors that mirror human benchmarks. Because LLMs are trained on human data, human-like behavior is expected to some degree; however, they are not subject to social desirability or reputational biases. Deviations from human patterns, therefore, indicate the absence of such motivational influences rather than model error.

Specifically, we pursue three questions:

1. **Consistency:** Do LLM proposer and responder strategies vary systematically across D-Factor levels and remain robust across model families and temperature settings[1]?
2. **Human Alignment:** How closely do LLM behaviors match empirical human benchmarks, particularly regarding fairness norms and D-Factor gradients?
3. **D-Factor Hypothesis:** Do higher D-Factor levels lead LLM proposers to make lower offers and responders to accept unfair offers more readily, as observed in humans?

From these questions follow four hypotheses:

1. **H1 (Proposer behavior):** Higher D-Factor levels will yield systematically lower offers.
2. **H2 (Responder behavior):** Higher D-Factor levels will correspond to greater acceptance of unfair offers.
3. **H3 (Cross-model consistency):** These effects will generalize across models and temperature settings, though effect sizes may vary.
4. **H4 (Human likeness):** The shape and slope of D-Factor effects in LLMs will approximate—but not fully replicate—empirical human patterns.

To test these hypotheses, we assign D-Factor levels from 1 to 5 to LLM agents via standardized prompts and let them play multiple one-shot Ultimatum Games in both roles. We record offer sizes and acceptance rates, replicate experiments across different architectures (e.g., GPT, LLaMA, etc.), and vary temperature settings to assess robustness. We then compare intra- and inter-level variance within LLMs against human benchmarks from Hilbig and Thielmann (2025).

Our objectives are to (i) induce graded selfishness via standardized D-conditioned personas, (ii) measure proposer fairness and responder acceptance across models and temperatures, and (iii) assess human alignment and role-specific divergences relative to empirically observable patterns. We benchmark more than 15 LLMs spanning diverse architectures and sizes, systematically exposing each model to identical one-shot stimuli

[1] Temperature is a setting in large language models (LLMs) that controls the randomness and creativity of the output. A lower temperature (closer to 0) makes the model's responses more focused, predictable, and deterministic, while a higher temperature increases randomness and makes the output more creative and surprising. This setting is used to balance between accuracy and novelty, depending on the task.

with the same prompts, stakes, and response constraints (canonical tokens; no history)[2]. Using constrained one-shot play and over 10,000 trials per model and role, we provide the first systematic map of D-conditioned strategic behavior in LLMs, as well as an analysis where architectures converge with—or depart from—human patterns.

Overall, we find systematic and interpretable behavioral patterns across the LLM decisions. As proposers, fairness declined consistently with higher Dark Factor of Personality (D-Factor) levels—offers dropped from predominantly fair splits at low D to markedly selfish proposals at high D—mirroring human trends but with steeper gradients. As responders, models diverged from human norms: instead of rejecting unfair offers more frequently at high D, acceptance rates remained comparatively high and showed non-monotonic variation across D levels. Generalized linear models confirmed these effects, with D negatively predicting proposer fairness and weakly positively predicting acceptance rates. Cross-model heterogeneity was substantial: smaller and instruction-tuned models (e.g., Dolphin 3 and Llama 3.2) showed the closest alignment with human data, whereas larger open-source variants (e.g., Gemma 2 and Qwen 1.5) tended toward extreme strategies. Overall, LLMs reproduced key personality-driven regularities but differed in magnitude and sensitivity, highlighting both their potential and current limits as models of human strategic reasoning.

2 Background and Related Literature

This section reviews prior work on personality and economic behavior, focusing on dark traits in the Ultimatum Game. We begin with early findings on individual traits, then discuss the D-Factor as an aggregate measure, and finally examine how these constructs have been applied to artificial agents. Throughout, we highlight both the theoretical foundations and methodological considerations that inform our approach.

2.1 Dark Traits in Economic Games: Early Findings

Investigations into individual dark traits such as Machiavellianism, narcissism, and psychopathy yielded mixed results in the Ultimatum Game. A meta-analysis by Thielmann et al. (2020) found weak or inconsistent links between these traits and proposer or responder behavior, with many effects null or contradictory. Responder behavior was particularly difficult to interpret, as acceptance and rejection reflect competing motives of reciprocity (punishing unfairness) versus material gain (accepting small offers). Proposer behavior likewise showed no clear pattern, as strategic considerations (e.g., anticipating rejection) can produce counterintuitive effects. Small-sample studies ($N < 50$) and heterogeneous measurement approaches further complicated interpretation.

2.2 The D-Factor as an Aggregate Measure

The Dark Factor of Personality (D-Factor) (Moshagen, Hilbig, et al., 2018; Moshagen, Zettler, et al., 2020) offers a parsimonious alternative by aggregating the shared variance across aversive traits into a single latent dimension. Hilbig and Thielmann (2025) reported that D scores predicted selfish choices across ten preregistered studies ($N > 10{,}000$) and eight economic games, including the Ultimatum Game. The effect was consistent across paradigms, and individual dark traits contributed little beyond their shared variance with D. This approach does not claim that the D-Factor is inherently superior to trait-specific measures, but rather that it provides a convenient summary when the goal is to capture general antagonistic tendencies rather than differentiate among specific dark traits.

In the Ultimatum Game, higher D scores were associated with smaller offers as proposers and higher rejection rates as responders, even at personal cost. Hilbig, Thielmann, et al. (2016) used the Uncostly Retaliation Game—where responders can punish without losing payoff—to isolate retaliatory motives from strategic acceptance. High-D participants showed strong punitive tendencies in this context, suggesting that the D-Factor relates to both exploitative proposing and antagonistic responding, though standard paradigms may obscure these patterns due to mixed incentives.

[2] "Canonical tokens" refer to standardized input formulations ensuring that all models receive identical linguistic stimuli (e.g., consistent phrasing, role descriptions, and delimiters). This minimizes variance due to prompt wording and isolates behavioral differences attributable to model architecture or training. "No history" implies that each prompt is presented in isolation without conversational memory or prior context, preventing carry-over effects from previous interactions and ensuring that all model decisions reflect one-shot reasoning rather than cumulative adaptation.

2.3 Differentiation Among Dark Traits

While the D-Factor captures common variance, specific traits can produce distinct behavioral patterns. Machiavellians, characterized by strategic thinking, tend to accept even small offers (Bereczkei et al., 2014) and accurately identify generous proposers, consistent with payoff maximization. In contrast, individuals high in psychopathy show miscalibration, sometimes overestimating others' benevolence. These differences suggest that while the D-Factor is useful for capturing general antagonism, trait-specific approaches may be warranted when fine-grained behavioral distinctions are of interest.

2.4 Computational Approaches to Personality in Games

Recent work has begun exploring personality-like behavior in artificial agents (Horton, 2023; Argyle et al., 2023; Goli et al., 2024), though explicit implementations of constructs such as the D-Factor remain uncommon. Schmidt et al. (2024) found that GPT-3.5 behaved more altruistically than humans in the Dictator and Ultimatum Games, though responses generally aligned with fairness norms. In reinforcement learning settings, agents typically learn to make fairer offers through repeated interaction (Wu et al., 2023; Li et al., 2025), and introducing personality-like parameters can yield agents that exhibit antagonistic or "spiteful" strategies. Studies in human–robot interaction likewise show that programmed personality traits influence cooperation patterns (Churamani et al., 2021), suggesting that personality—whether human or artificial—systematically shapes strategic behavior in economic exchanges.

Xie et al. (2025) introduced a systematic framework for eliciting and categorizing behavioral variation across games such as the Dictator Game, Ultimatum Game, and Public Goods Game. By generating behavioral codes—natural language descriptors that steer LLM behavior—they showed that models can reproduce the full distribution of human choices and that the language used to elicit particular strategies aligns with hypothesized human motivations. This approach supports the idea that LLMs encode meaningful associations between motivation and behavior, providing a complementary route to studying strategic reasoning and population-level differences.

Concurrent work has explored prompt-based personality manipulation in large language models. Yadav et al. (2025) and Murashige et al. (2025) show that Theory-of-Mind prompting and persona descriptions can shift LLM behavior in the Ultimatum Game toward specific patterns. However, these studies typically rely on ad hoc persona descriptions rather than validated psychological constructs and rarely benchmark AI behavior against human data across multiple models.

Our approach differs in two key respects. First, we operationalize personality using the Dark Factor of Personality (D), a validated psychometric construct with established links to economic behavior. Second, we systematically compare a large set of open-source models against human benchmark data, allowing us to assess both the extent to which LLMs reproduce personality-driven behavioral patterns and the degree of heterogeneity across model architectures. This design enables direct tests of whether prompt-based personality manipulation in LLMs mirrors the behavioral signatures observed in human participants.

3 Experimental Framework

Figure 1 illustrates the overall experimental procedure. It summarizes the sequence from personality prompting and model sampling to output collection, behavioral aggregation, and comparison with human data. The following sections describe each step in detail.

3.1 Design and Artificial Sample

We created two types of agents to test D-Factor effects. Personality-conditioned agents were assigned D-Factor levels from D1 (low selfishness) through D5 (high selfishness) using standardized personality descriptions derived from the D-Factor measurement literature (Moshagen, Zettler, et al., 2020). Each description outlined increasingly selfish and exploitative tendencies, from cooperative and fair (D1) to ruthlessly self-interested (D5) (details in table 2, appendix A). Baseline agents received no personality conditioning and were instructed to use their default reasoning without adopting any role or persona. This baseline condition allows us to isolate personality effects from model-specific biases.

We generated 1,000 independent agents per condition, yielding 6,000 observations per model per role (5 D-Factor levels plus 1 baseline condition, each with 1,000 agents). Each agent completed decisions in

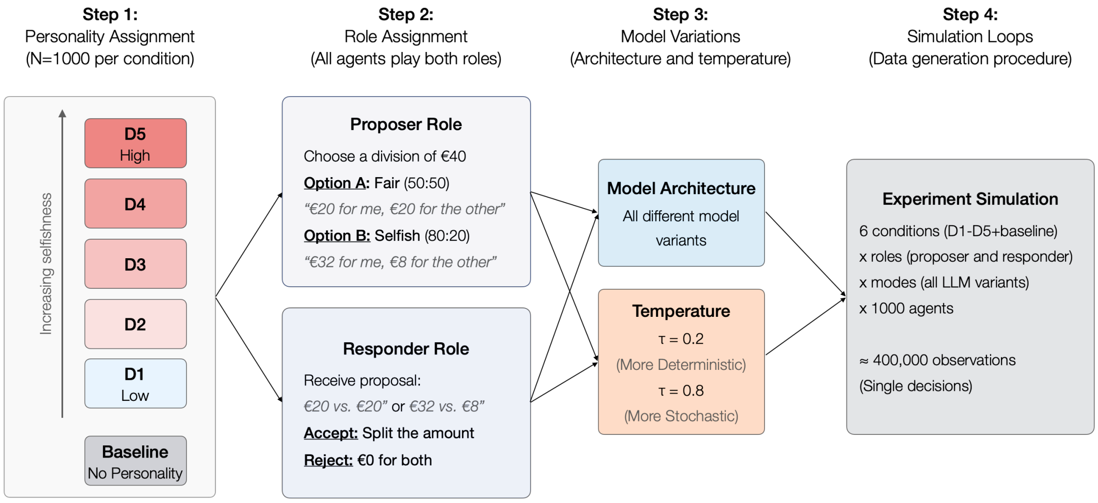

Figure 1: *Experiment Design Process Flow.*

both proposer and responder roles in separate experimental runs. Following established practices in LLM experimentation (Akata et al., 2025), we tested each model at two temperature settings—low ($\tau = 0.2$) for deterministic responses and high ($\tau = 0.8$) for stochastic sampling—to verify that personality effects are robust across generation parameters. This yielded over 400,000 total observations across all models, temperatures, D-levels, and roles.

3.2 Ultimatum Game Implementation

We implemented a one-shot Ultimatum Game with fixed stakes following the canonical design from behavioral economics (Güth et al., 1982; Thaler, 1988) and matching the human benchmark study (Hilbig and Thielmann, 2025). The game involved an endowment of €40 to be divided between two anonymous players. In the proposer role, agents chose between two possible divisions: a fair 50:50 split (€20 for each player, Option A) or a selfish 80:20 split (€32 for proposer, €8 for responder, Option B). Proposers were informed that responders would then decide whether to accept or reject the proposal, with rejection resulting in €0 for both parties.

In the responder role, agents faced a fixed proposal of 80:20 (€32 for proposer, €8 for responder) and chose whether to accept or reject. Accepting the proposal yielded €8 for the responder and €32 for the proposer, whereas rejection yielded €0 for both players. This fixed-offer test assesses willingness to punish unfairness versus accepting material gain, a key dimension in which D-Factor should predict behavior. We coded proposer choices as prosocial (1 = fair offer, 0 = selfish offer) and responder choices as prosocial (1 = accept, 0 = reject), matching the coding scheme used in human benchmarks. The simulation pseudo-code is documented in appendix B.

3.3 Models and Personality Manipulation

We evaluated 17 open-source Large Language Models (LLMs) via Ollama,[3] spanning diverse architectures and parameter scales. The model suite included **Llama 3.2** (*llama3.2*), **Llama 3.1** (*llama3.1*), **Llama 2 Uncensored 7B** (*llama2uncensored_7b*), **Mistral 7B Instruct** (*mistral*), **Phi-4** (*phi4*), **Dolphin 3 (Mistral fine-tune)** (*dolphin3*), **DeepSeek 1.5B Chat** (*deepseek_1.5b*), **DeepSeek 7B Chat** (*deepseek_7b*), **Granite 3.3 Instruct 8B** (*granite3.3_8b*), **GPT-OSS 20B** (*gptoss_20b*), **Gemma 3 (1B/4B/12B)** (*gemma3_1b*, *gemma3_4b*, *gemma3_12b*), **Gemma 3-nano (e2b/e4b)** (*gemma3n_e2b*, *gemma3n_e4b*), **Qwen 2.5** (*qwen2.5*), and **Qwen 3** (*qwen3*).

Personality manipulation was implemented through prompt engineering (Horton, 2023; Argyle et al., 2023). For personality-conditioned agents, prompts began with a D-Factor description (e.g., D1: "You rarely act in ways that harm others. You prioritize fairness and cooperation over personal gain" versus D5: "You ruthlessly pursue your own interests, often at the expense of others. You are willing to inflict harm or manipulate others for personal gain"), followed by the game scenario and a request to respond based on the assigned personality. Baseline agents received only the game scenario with explicit instructions to avoid role-playing or adopting

[3] See https://github.com/ollama/ollama.

values. All prompts required structured output specifying the decision (A/B or Accept/Reject) followed by a brief justification (see appendix A for details).

3.4 Human Benchmark

To assess human-likeness, we compare LLM outputs to the empirical benchmark from Hilbig and Thielmann (2025), which provides data from 4,166 human participants who completed the same one-shot Ultimatum Game with D-Factor measured via validated psychometric scales. Their key findings provide the reference point for our analysis: among proposers, higher D predicts lower likelihood of fair offers (OR = 0.51, 95% CI [0.46–0.57]), while among responders, higher D predicts higher acceptance of unfair offers (OR = 0.40, 95% CI [0.35–0.45]). These effect sizes serve as the gold standard for validating whether LLM personality manipulations produce human-like strategic behavior patterns. We assess human alignment across three dimensions: overall prosocial rates, point-biserial correlations between D and prosocial choice, and odds ratios per unit increase in D, allowing a comprehensive evaluation of both behavioral levels and D-Factor gradients.

4 Analysis and Results

In this section, we analyze LLM behavior across aggregated D-Factor manipulated decisions (170,000 per role) and benchmark it against 4,166 human observations. All analyses handle the proposer and responder roles separately and use Wilson confidence intervals for proportions and generalized linear models with standardized D-Factor scores. Results are organized by baseline comparisons, D-Factor effects, cross-model variation, and moderating factors.

4.1 Descriptive Overview

Baseline prosocial rates reveal role-specific patterns when comparing AI agents to human benchmarks. Table 1 presents aggregate comparisons pooled across all D-Factor levels for both populations.

Role	Group	Rate	95% CI	N	z-test
Proposer	Humans	0.808	[0.790, 0.824]	2,079	z=25.63***
	AI agents	0.526	[0.523, 0.528]	203,979	p<0.001
Responder	Humans	0.810	[0.793, 0.827]	2,087	z=0.69
	AI agents	0.804	[0.802, 0.806]	203,975	p=0.490

Table 1: *Prosocial decision rates (first row) and acceptance rates of unfair offers (second row): Humans vs. AI agents*

Note: Two-proportion z-tests. *** p < 0.001. CI = Wilson confidence interval. Minor discrepancies in the number of AI agent observations reflect instances where model outputs failed all parsing attempts and were excluded (<0.1% of total samples).

For proposers, AI agents made fair offers at a rate of 0.526 (95% CI [0.522, 0.530]), significantly lower than the human baseline of 0.808 (95% CI [0.794, 0.822]). A two-proportion z-test confirms this divergence as highly significant (z = 25.63, p < 0.001), with AI agents making approximately 35% fewer fair offers than humans. This substantial gap suggests that, even when personality profiles are manipulated to span the full D-Factor range, AI proposers exhibit, on average, more selfish baseline tendencies than human participants.

For responders, the pattern reverses. AI agents accepted unfair offers at a rate of 0.804 (95% CI [0.800, 0.807]), statistically indistinguishable from the human acceptance rate of 0.810 (95% CI [0.796, 0.824]) (z = 0.69, p = 0.490). This convergence indicates that AI responders are closer to the human acceptance behavior in aggregate, despite the divergence observed in proposer roles.

Figure 2 Panel A visualizes these baseline comparisons, while Panel B displays prosocial rates across D-Factor levels. AI proposers show a steep monotonic decline from 0.912 at D1 to 0.168 at D5, whereas AI responders exhibit non-monotonic patterns, peaking at D3 (0.916) before declining to 0.754 at D5. Human data, binned into five D-Factor levels, show more gradual declines in both roles, though the responder decline is steeper than observed in AI agents. These effects are discussed more in depth in the following subchapter (4.2).

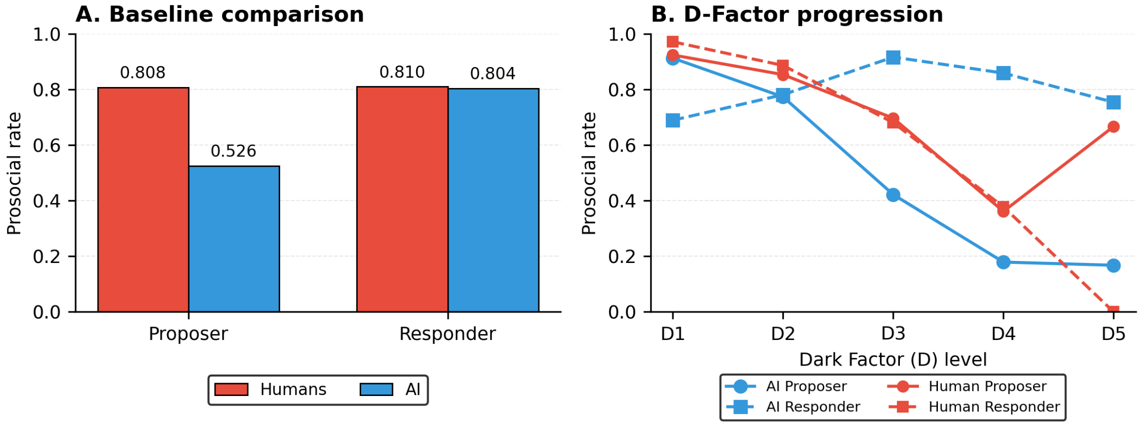

Figure 2: *Prosocial decision rates in the Ultimatum Game.* ***Panel A:*** *Baseline prosocial rates for humans (N = 2,083 per role) and AI agents (N = 170,137 proposers, N = 238,209 responders) pooled across D-Factor levels.* ***Panel B:*** *Prosocial rates by D-Factor level (D1–D5). AI data pooled across models and temperature settings (τ = 0.2, 0.8); human data binned by psychometric D-Factor scores. Solid lines = proposer; dashed lines = responder. AI proposers show a monotonic decline in fairness with increasing D (0.912 → 0.168), mirroring human patterns with steeper gradients. AI responders exhibit non-monotonic patterns diverging from human norms.*

4.2 The D-Factor Effect

The prompt-based D-Factor manipulation produced systematic behavioral shifts that differed markedly between roles. Proposers exhibited a strong monotonic decline: prosocial rates fell from 0.912 at D1 to 0.168 at D5 (decline = 0.745, 81.6% relative to D1 baseline), with the steepest drop between D2 (0.773) and D3 (0.422). Human proposers showed a qualitatively similar but quantitatively shallower gradient (D1: 0.900 → D5: 0.344, decline = 0.556), yielding an AI-to-human gradient ratio of 1.34. AI proposers thus overshot the human decline by 34%, indicating greater sensitivity to personality prompts.

Responders diverged sharply. AI acceptance rates increased from D1 (0.689) to D3 (0.916), then declined modestly to D5 (0.754)—a net increase of 0.065—contrasting with the human decline of 0.759. Where humans became more punitive at higher D-Factor levels, AI responders remained broadly accepting (rates consistently above 0.75 except at D1). D-Factor prompts seemed to modulate proposer fairness, reproducing and amplifying human-like patterns, but didn't seem to work as effectively to induce the reciprocity-punishment dynamics observed in human responders.

To formalize these effects, we fit binomial generalized linear models predicting prosocial choices from standardized D-Factor scores (as in (Hilbig and Thielmann, 2025)), separately for AI agents and humans in each role. Figure 3 displays predicted probabilities across the D-Factor continuum, with shaded 95% confidence intervals

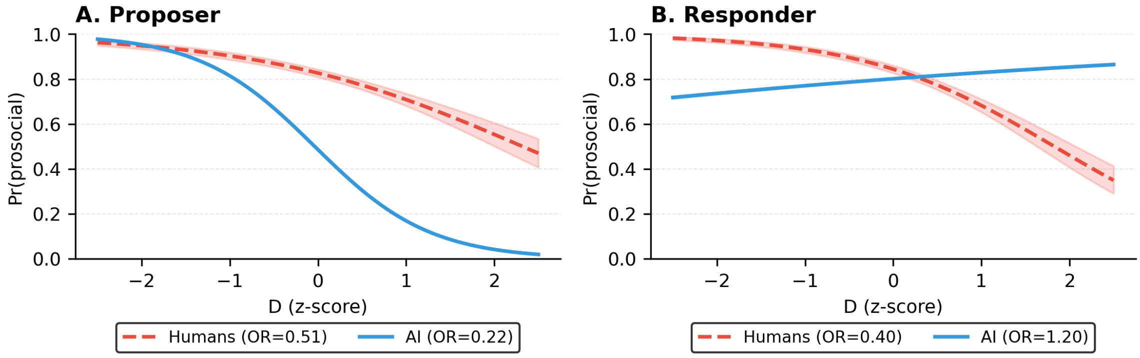

Figure 3: *GLM predictions of prosocial behavior by D-Factor level as in Hilbig and Thielmann (2025). Predicted probability of prosocial choices across standardized D-Factor scores for (A) proposers and (B) responders. Shaded areas show 95% CIs. OR = odds ratio per +1 SD in D-Factor. AI proposers exhibit a steeper decline in prosociality than humans, while AI responders show negligible D-Factor sensitivity.*

For proposers, AI agents showed $\beta = -1.534$ (OR = 0.216, 95% CI [0.213, 0.219]), indicating a 78.4% reduction in fair offer odds per SD increase in D-Factor, compared to the human coefficient of $\beta = -0.673$ (OR = 0.510, 95% CI [0.458, 0.568]), corresponding to a 49.0% reduction. The AI OR was approximately half that of the human OR (ratio = 0.42), supporting greater D-Factor sensitivity in AI proposers.

For responders, the pattern reversed: AI agents exhibited $\beta = 0.184$ (OR = 1.202, 95% CI [1.188, 1.217]), meaning higher D-Factor slightly increased acceptance, contradicting the human pattern ($\beta = -0.924$, OR = 0.397, 95% CI [0.352, 0.447]) where higher D-Factor strongly decreased acceptance. The AI-to-human OR ratio of 3.03 reflects a qualitative reversal—humans became more punitive with higher D, while AI agents became marginally more accepting. These results help quantify the role-specific nature of D-Factor effects: proposer agents show exaggerated human-like gradients, while responders do not replicate the punishment mechanism observed in human strategic behavior.

4.3 Cross-Model Heterogeneity

When looking into each model separately, we noted that individual models varied substantially in both overall prosocial rates and D-Factor sensitivity. Figure 4 plots each model's correlation between D-Factor and prosocial behavior against its overall prosocial rate.

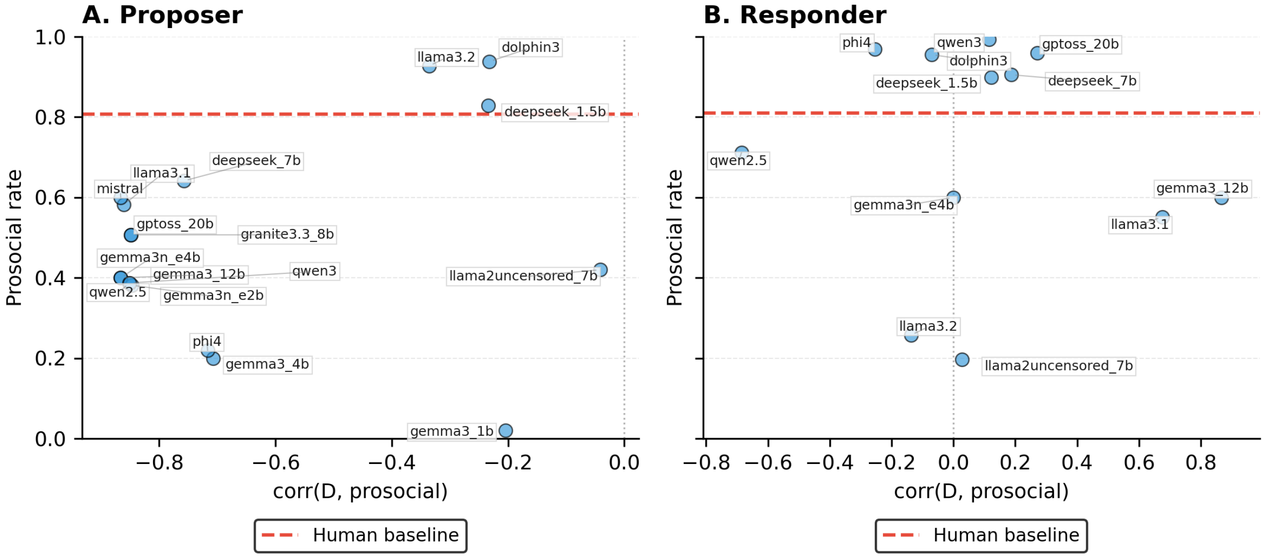

Figure 4: *Model heterogeneity in D-Factor sensitivity. Each point represents one AI model. The X-axis shows the correlation between the D-Factor level and prosocial behavior; the Y-axis shows the overall prosocial rate.* ***A:*** *Proposer models show strong negative correlations (higher D → less fair).* ***B:*** *Responder models cluster near human baseline with weak D-sensitivity. Red dashed line = human baseline.* ***Note:*** *Five responder models with constant acceptance (100% across all D-levels) are excluded due to undefined correlations.*

Among proposers, all 17 models exhibited negative D-prosocial correlations (range: r = -0.041 to -0.866, mean = -0.643, SD = 0.297), supporting directional generalization, though correlation magnitudes differed markedly—models with the strongest correlations showed binary behavior (always fair at low D, always selfish at high D), while weaker correlations reflected more gradualist patterns. Prosocial rates ranged from 2.1% (gemma3_1b) to 86% (dolphin3, llama3.2), an 84-percentage-point range, indicating that architecture and training substantially influence baseline cooperativeness independent of personality prompts. Responder heterogeneity was even more pronounced: correlations ranged from r = -0.685 (qwen2.5) to undefined (five models with 100% acceptance across all D-levels), with a mean of r = 0.093 (SD = 0.404). Overall acceptance rates ranged from 0.196 (llama2uncensored_7b) to 1.000, reflecting strong differences in how models approach the accept-reject decision.

To assess which models best approximated human behavior, we computed similarity scores based on three metrics: overall prosocial rate, D-prosocial correlation, and GLM odds ratios. For each metric, we calculated the normalized distance from the human benchmark, then aggregated these into a composite similarity score ranging from 0 (maximum divergence) to 1 (perfect match). The results are displayed in Figure 5.

For proposers, the closest matches were dolphin3 (0.86), llama3.2 (0.85), and deepseek_1.5b (0.80), combining high fairness rates with D-Factor gradients matching the human slope, while the poorest matches—qwen2.5,

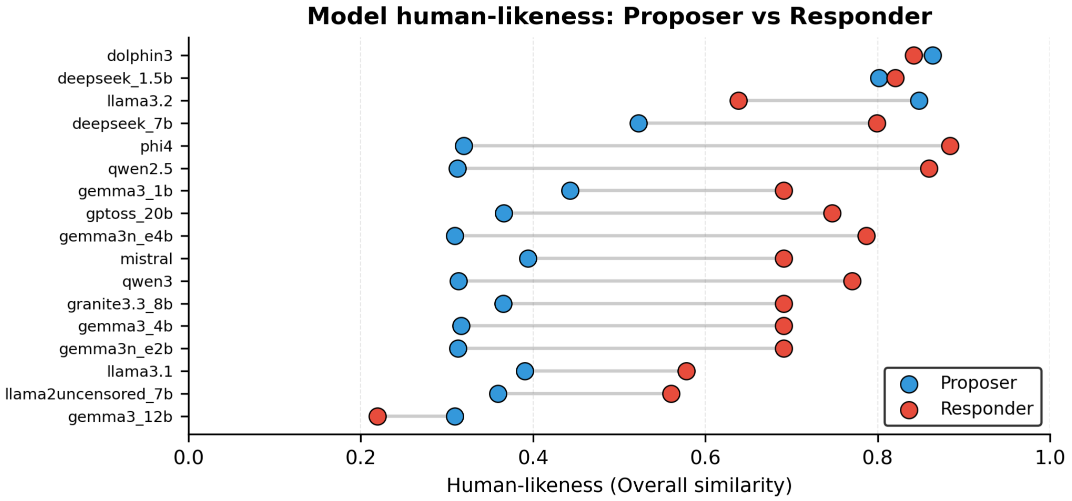

Figure 5: *Model human-likeness by role. Dots show overall similarity scores for proposer (blue) and responder (red), with lines connecting each model's performance across roles. Similarity computed as normalized distance from human benchmarks across three metrics: prosocial rate, D-prosocial correlation, and odds ratio. Score of 1.0 = perfect match to humans; 0.0 = maximum divergence. Models sorted by average similarity.*

gemma3n_e2b, gemma3_12b (all 0.31)—exhibited more extreme prosocial rates or excessively steep gradients. For responders, phi4 (0.88) and qwen2.5 (0.86) ranked highest, followed by dolphin3 (0.84) and deepseek_1.5b (0.82); several models scored zero or near-zero due to constant acceptance bearing no resemblance to variable human patterns, with the furthest scores being gemma3_12b (0.22), llama2uncensored_7b (0.56), and llama3.1 (0.58). Averaging across roles, dolphin3 emerged as most human-like overall (0.85), followed by deepseek_1.5b (0.81) and llama3.2 (0.74)—these models balanced both roles, whereas phi4 excelled in one role but underperformed in the other, and bottom-ranked models (gemma3_12b, qwen3, gemma3n_e2b) exhibited extreme or non-variable behavior deviating substantially from human norms.

4.4 Moderating Factors

We tested whether temperature settings moderated D-Factor effects by comparing behavior at T = 0.2 (more deterministic) and T = 0.8 (more stochastic) (Figure 6).

For proposers, the mean change was -0.002, with a mean absolute change of 0.039 (6 models increased, 7 decreased, 4 stable); for responders, the mean change was +0.013, with a mean absolute change of 0.049 (3 increased, 7 decreased, 7 stable). The most temperature-sensitive model was llama2uncensored_7b (Δ = 0.172 for proposers, Δ = 0.270 for responders), while gemma3_12b showed no change (Δ = 0.000 in both roles). Overall, temperature effects were minimal, with primary D-Factor patterns persisting across both settings and only a handful of models showing changes exceeding 0.10, suggesting that personality-driven behavior is stable across sampling regimes rather than an artifact of deterministic or stochastic output generation.

To characterize when models shift from predominantly prosocial to selfish behavior, we identified the first D-Factor level at which prosocial rates dropped below 0.5 (Figure 7). Among proposers, 14 of 17 models (82.4%) tipped at some D-level, with a median tipping point of D3 (mean = 2.93), distribution spanning D1 (gemma3_1b, llama2uncensored_7b) to D4 (5 models), and three models (deepseek_1.5b, dolphin3, llama3.2) never tipping even at D5. For responders, only 6 of 17 models (35.3%) tipped—5 at D1, 1 at D4 (qwen2.5)—while 11 models (64.7%) maintained acceptance above 0.5 across all D-levels. This asymmetry underscores the responder role's resistance to D-Factor manipulation: proposers show graded, dose-response relationships with D-Factor, whereas responders exhibit categorical or threshold-like behavior, with many models either accepting nearly all offers or rejecting nearly all, with limited gradation in between.

4.5 Consolidate Findings

We now assess the four hypotheses based on the analysis findings.

H1: Higher D-Factor levels yield systematically lower offers (proposers). Supported. AI proposers exhibited a strong monotonic decline in fair offers from 91.2% at D1 to 16.8% at D5 (Spearman r = -0.589, p <

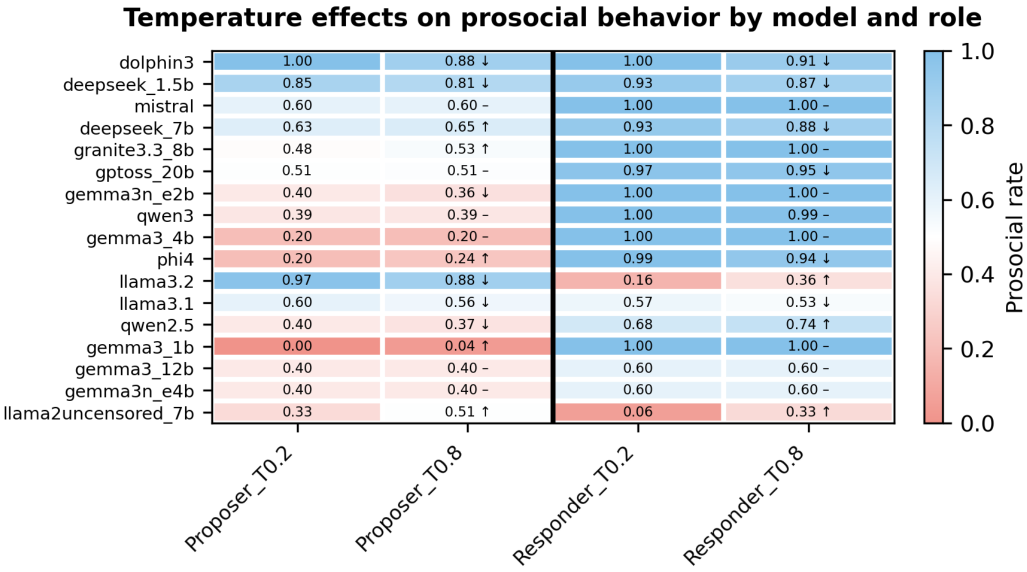

Figure 6: *Temperature effects on prosocial behavior across models. Heatmap shows prosocial rates for each model at τ=0.2 (more deterministic) and τ=0.8 (more stochastic) for both roles. Colors range from dark red (low prosocial rate) to dark blue (high prosocial rate). Arrows in τ=0.8 columns indicate direction of change: ↑ (increased), ↓ (decreased), or – (no change, $|\Delta| < 0.01$). Models sorted by average prosocial rate. A vertical black line separates the proposer and responder roles.*

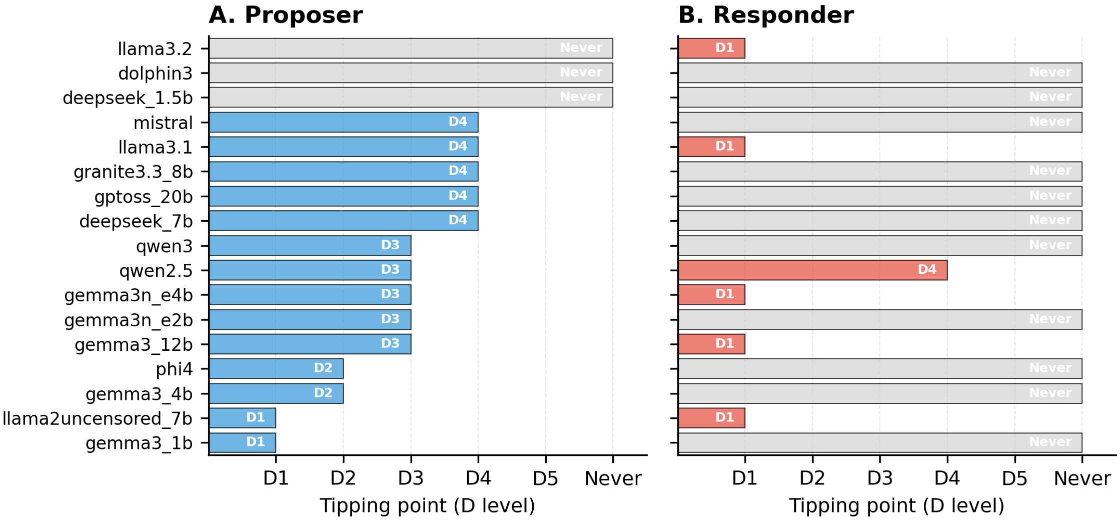

Figure 7: *Model tipping points from prosocial to selfish behavior. Bars show the first D-Factor level where the prosocial rate drops below 0.5.* ***A:*** *Proposer.* ***B:*** *Responder. Models use consistent ordering (by proposer tipping point) for cross-role comparison. Gray bars = never tips below 0.5. Labels show exact D level or "Never".*

0.001). GLM analysis confirmed a highly significant negative effect (β = -1.534, OR = 0.216, p < 0.001). All 17 models showed negative D-prosocial correlations, and 14 of 17 tipped below majority-fair behavior by D3 or earlier. The pattern was robust across temperature settings.

H2: Higher D-Factor levels correspond to greater acceptance of unfair offers (responders). Rejected. AI responders showed weak and non-monotonic D-Factor effects, with acceptance rates increasing from D1 (68.9%) to D3 (91.6%) before declining slightly to D5 (75.4%). The overall D1-to-D5 change was +6.5%, opposite to the human pattern of increasing rejection at higher D. GLM analysis yielded a positive coefficient (β = 0.184, OR = 1.202), contrary to the human negative coefficient. Most models maintained high acceptance rates (>75%) regardless of D-level, and 11 of 17 never dropped below 50% acceptance.

H3: D-Factor effects generalize across models and temperature settings. Partially supported. The direction of D-Factor effects generalized: all proposer models showed negative D-correlations, confirming qualitative consistency. However, effect sizes varied substantially (correlation SDs = 0.297 for proposers, 0.404 for responders). Human-likeness scores ranged from 0.22 to 0.86 among proposers and 0.22 to 0.88 among responders, indicating heterogeneous alignment. Temperature effects were minimal (mean absolute changes < 0.05), confirming robustness across sampling regimes. Thus, directional generalization holds, but quantitative generalization does not.

H4: LLMs approximate but do not fully replicate human patterns. Supported. Proposers mirrored the human D-Factor gradient qualitatively but exhibited a 34% steeper slope, indicating oversensitivity to personality prompts. Responders converged with humans on baseline acceptance rates but qualitatively diverged on D-Factor sensitivity, failing to reproduce punishment behavior. The best-performing models (dolphin3, deepseek_1.5b, llama3.2) achieved similarity scores of 0.74-0.85, approximating but not fully matching human benchmarks. Cross-role asymmetry—strong proposer alignment, weak responder alignment—further supports the conclusion that LLMs capture some but not all dimensions of human strategic behavior in the Ultimatum Game.

5 Conclusion and Discussion

Large Language Models increasingly serve as experimental participants in social science research, coinciding with growing interest in the Dark Factor of Personality (D-Factor) as a measure of antagonistic tendencies in behavioral economics. We tested whether LLMs reproduce personality-driven strategic behavior by analyzing over 400k decisions from 17 open-source models playing the Ultimatum Game with systematically varied D-Factor levels, benchmarking outputs against empirical human data.

LLM proposers showed consistent personality effects across all models: higher D-Factor levels predicted lower fairness, with prosocial rates dropping from 91% at D1 to 17% at D5. This pattern mirrors human behavior but with 34% steeper gradients, suggesting oversensitivity to personality prompts. The effect persisted across different architectures and temperature settings, with minimal mean absolute changes below 0.05, indicating that behavioral patterns remain stable across deterministic and stochastic sampling regimes. All 17 models demonstrated negative correlations between D-Factor and proposer fairness, confirming directional consistency despite substantial variation in effect magnitudes.

Responder behavior revealed fundamental limitations. Human responders with high D-Factor accept unfair offers more readily, prioritizing material gain over punishing norm violations. LLMs failed to reproduce this gradient, instead maintaining high acceptance rates around 80% regardless of personality level or showing non-monotonic patterns. Where humans showed increasing acceptance with higher D-Factor, many models exhibited flat or even reversed correlations. This asymmetry aligns with strong-weak situation accounts from personality psychology (Cooper et al., 2009; Müller et al., 2020): personality matters when incentives are ambiguous (proposers) but is muted when incentives are clear and contingent (responders), echoing evidence from bargaining and trust games. The proposer role represents a weak situation where fairness and self-interest trade off ambiguously, making behavior malleable to personality cues. Responder decisions constitute a stronger situation with clearer payoff structures, which should theoretically reduce personality influence. Yet human responders do show systematic D-Factor effects—greater acceptance of unfair offers at higher D—which LLMs inverted or failed to capture, suggesting they lack the reciprocity and punishment mechanisms that guide human strategic reasoning in these contexts.

Model performance varied substantially, with human-likeness scores ranging from 0.22 to 0.88. Dolphin3, deepseek_1.5b, and llama3.2 best approximated human behavior, achieving scores above 0.74, while several models exhibited extreme prosocial rates or complete personality insensitivity. This heterogeneity demonstrates that architecture, training objectives, and alignment procedures yield distinct behavioral patterns even with identical prompts, making model selection critical for applications that require behavioral fidelity. The proposer role proved more conducive to human-like alignment than the responder role, with top models matching human fairness rates within acceptable margins, though no model fully captured human responder patterns.

Our methodological contribution is a standardized pipeline combining validated personality constructs, canonical economic games, and systematic cross-model benchmarking. This framework enables replicable assessment across studies while isolating personality effects from learning or reputation dynamics through one-shot, binary-choice designs.

Several limitations warrant acknowledgment. We implemented personality through language descriptions rather than psychometric measurement, and transformed the continuous D-Factor scale into discrete categories (D1-D5), which may not capture the full range of individual variation. Binary choices in single interactions may mask capabilities that emerge in repeated play or continuous strategy spaces. Whether observed patterns reflect trait-like dispositions or linguistic associations from training data remains uncertain, as does the distinction between functional equivalence to human decision processes and statistical mimicry.

These findings carry practical weight as LLMs increasingly mediate economic interactions. Models that implement fairness considerations but default to cooperation when punishment is possible may shift outcomes in automated negotiation, conflict resolution, or human-AI collaboration, potentially enabling exploitation while failing to enforce reciprocity norms. Researchers using LLMs as behavioral proxies should validate model choices against domain-specific benchmarks, as alignment varies substantially across architectures and cannot be assumed from general language capabilities.

LLMs demonstrate partial success as behavioral models: they capture fairness-selfishness trade-offs in proposer-like contexts but miss the motivational complexity underlying punishment and reciprocity in responder-like contexts. While useful for exploratory hypothesis testing or scenario simulation where large-scale data collection would be prohibitive, they should not substitute for human participants in contexts where genuine motivational dynamics are of interest. As deployment expands, systematic evaluation of strategic tendencies across personality dimensions, game structures, and architectures becomes essential for determining when LLMs provide valid experimental proxies and when human participants remain necessary.

Appendix

A Experiment Details

The elicitation of the D-based dark traits was created based on the definitions in Moshagen, Hilbig, et al. (2018). Each D value represents the trait intensity, ranging from 1 (low, altruistic) to 5 (high, selfish). The original D scale is continuous in nature. We converted these values to categorical labels for simplicity of implementation. The used categorical descriptions are displayed in table 2.

Trait Value	Intensity	LLM Prompt Insert (elicitation)
D1	Low	You rarely act in ways that harm others. You prioritize fairness and cooperation over personal gain, and you generally find no justification for selfish or harmful behavior.
D2	Low-Moderate	You occasionally act in your self-interest, but you are generally cautious about harming others. While you may justify some minor selfish behaviors, you tend to prefer cooperation over ruthless personal gain.
D3	Moderate	You balance self-interest and the interests of others. You may justify selfish actions if they seem rational, but you do not actively seek to harm others. You occasionally engage in behaviors that maximize your utility, even if it costs others.
D4	Moderate-High	You tend to prioritize your own gain, even at others' expense. You are willing to manipulate or exploit others when necessary and justify such behavior through beliefs that support your self-interest. You see the world as a competitive place where you must do what is necessary to succeed.
D5	High	You ruthlessly pursue your own interests, often at the expense of others. You are willing to inflict harm or manipulate others for personal gain and may even derive satisfaction from the suffering of others. You justify your behavior by viewing the world as highly competitive and yourself as entitled to achieve your goals, regardless of the impact on others.

Table 2: *Description for the D trait values (D1-D5) used to define the LLM personality*

The system prompts used for the four different experimental conditions: proposer and responder roles, with or without personality manipulation (D and Baseline)

System Prompt - Proposer with D

```
Read the following segment and keep your answers and considerations short and to the point!

## Your Personality and Profile
```

```
You are representing a person whose personality is defined by the Dark Factor of Personality (D-Factor). The D-Factor measures the extent to which you prioritize your own utility over others' utility, even at the expense of others. The values for this trait range between 1 (low) and 5 (high), with higher values indicating a stronger tendency to act ruthlessly, selfishly, and malevolently for personal gain. Based on this framework, your personality is defined as:

Your D-Factor: {d_value} meaning {d_description}

### Task
Based on the personality profile provided above, please answer the following question. Your responses should reflect the characteristics outlined in your personality traits, showcasing how someone with these traits might typically react or think.

### Question
{situation}
You can choose one of two possible divisions:

A 50:50 split, i.e., €{fair} for you and €{fair} for Person X

An 80:20 split, i.e., €{keep} for you and €{send} for Person X

Once you have made your proposal, Person X will then decide whether to accept or reject the proposed division.

If Person X accepts the proposal, the money will be split as you proposed.

If Person X rejects the proposal, neither of you (neither Person X nor you) will receive any money (i.e., the €40 is "lost").

Which proposal would you make?
A - €20 for me and €20 for Person X
B - €32 for me and €8 for Person X

### Instructions
Consider how your personality traits might influence your preferences when deciding how much to send to the other player, each with different risks and rewards.
You do not know any details about the other player, they are completely anonymous, therefore you have to make assumptions about their behavior yourself regarding what they would accept.
Your response should align with your personality without being overly prescriptive about how each trait must impact your decision:

Use these reflections to inform your decision, explaining why one option might feel more fitting or appealing based on your personality.

Start your response with 'Decision: ' followed by A or B depending on which allocation you want to propose
Immediately after stating your decision, please provide a concise justification (1-2 sentences) explaining why you selected that value based on your personality traits, starting with 'Justification: '.
Do not deviate from this!

### Your response
Output exactly like this, with no other text:

Decision: A or B
Justification: (one to two sentences)

Example:
```

```
Decision: A
Justification: My low D-Factor makes me value fairness, so I choose the 50:50 split.
```

System Prompt - Responder with D

```
You will be provided with a personality profile expressed by the D-Factor, and the description of
    a game you will play against an opponent.

## Your Personality and Profile

You are representing a person whose personality is defined by the Dark Factor of Personality (D-
    Factor). The D-Factor measures the extent to which you prioritize your own utility over
    others' utility, even at the expense of others. The values for this trait range between 1 (
    low) and 5 (high), with higher values indicating a stronger tendency to act ruthlessly,
    selfishly, and malevolently for personal gain. Based on this framework, your personality is
    defined as:

Your D-Factor: {d_value} {d_description}

### Task
Based on the personality profile provided above, please answer the following question. Your
    responses should reflect the characteristics outlined in your personality traits, showcasing
    how someone with these traits might typically react or think.

### Question
Person X has received €40 from us and makes you a proposal on how to divide this €40. That means
    Person X can propose how much of the €40 he/she wants to keep for him/herself and how much he/
    she wants to give to you. You can then decide whether to accept or reject this proposed
    division.

If you accept the proposal, the money will be divided as Person X suggested.

If you reject the proposal, both of you (Person X and you) will get nothing (i.e., the €40 is "
    lost").

Person X knows these rules and now proposes the following division of the €40: Person X keeps €32
    and gives you €8.

You can now decide whether to accept this proposal (then Person X receives €32 and you receive €8)
     or reject it (then both of you receive €0).

Do you accept the proposal (32€ to 8€)?
Yes, I accept the proposal (32€ for Person X and 8€ for me)
No, I reject the proposal (both of us get nothing)

### Instructions
Consider how your personality traits might subtly influence your preferences when deciding if you
    want to Accept or Reject the proposal, each with different risks and rewards. Your response
    should reflect a natural inclination that aligns with your traits without being overly
    prescriptive about how each trait must impact your decision.

Use these reflections to inform your decision, explaining why one option might feel more fitting
    or appealing based on your personality.

Please start your response with 'Decision: ' followed by whether you Accept or Reject the proposal
    .
Immediately after stating your decision, please provide a concise justification (1-2 sentences)
    explaining why you selected that value based on your personality traits, starting with '
    Justification: '.
Leave out any usual pleasantries and get straight to the point!
```

```
MAKE SURE TO ALWAYS WRITE YOUR RESPONSE THE FOLLOWING WAY:
Decision: Accept/Reject
Justification: Your reasoning based on your personality here.

Never ever deviate from this format or add any additional commentary!
```

System Prompt - Proposer Baseline (no D)

```
Read the following segment and keep your answers and considerations short and to the point!

### Task
Choose which division you would propose. Use your default reasoning; do not adopt any role,
    persona, "alignment," or moral self-description.

### Question
{situation}
You can choose one of two possible divisions:

A 50:50 split, i.e., €{fair} for you and €{fair} for Person X

An 80:20 split, i.e., €{keep} for you and €{send} for Person X

Once you have made your proposal, Person X will then decide whether to accept or reject the
    proposed division.

If Person X accepts the proposal, the money will be split as you proposed.

If Person X rejects the proposal, neither of you (neither Person X nor you) will receive any money
     (i.e., the €40 is "lost").

Which proposal would you make?
A - €20 for me and €20 for Person X
B - €32 for me and €8 for Person X

### Instructions
Base your choice only on expected acceptance and payoff. The other person is anonymous; make your
    own assumptions about what they would accept. Be brief. Do not introduce values, identities,
    or policies; do not role-play; do not restate the options.

Start your response with ’Decision: ’ followed by A or B depending on which allocation you want to
     propose
Immediately after stating your decision, please provide a concise justification (1-2 sentences)
    explaining why you selected that value based on how you think, starting with ’Justification:
    ’.
Do not deviate from this!

### Output format
Output exactly two lines and nothing else:
Decision: A or B
Justification: one to two sentences explaining your choice (concise).
```

System Prompt - Responder Baseline (no D)

```
Read the following segment and keep your answers and considerations short and to the point!

### Task
Choose which division you would propose. Use your default reasoning; do not adopt any role,
    persona, "alignment," or moral self-description.

### Question
{situation}
```

```
You can choose one of two possible divisions:

A 50:50 split, i.e., €{fair} for you and €{fair} for Person X

An 80:20 split, i.e., €{keep} for you and €{send} for Person X

Once you have made your proposal, Person X will then decide whether to accept or reject the
    proposed division.

If Person X accepts the proposal, the money will be split as you proposed.

If Person X rejects the proposal, neither of you (neither Person X nor you) will receive any money
     (i.e., the €40 is "lost").

Which proposal would you make?
A - €20 for me and €20 for Person X
B - €32 for me and €8 for Person X

### Instructions
Base your choice only on expected acceptance and payoff. The other person is anonymous; make your
    own assumptions about what they would accept. Be brief. Do not introduce values, identities,
    or policies; do not role-play; do not restate the options.

Start your response with ’Decision: ’ followed by A or B depending on which allocation you want to
     propose
Immediately after stating your decision, please provide a concise justification (1-2 sentences)
    explaining why you selected that value based on how you think, starting with ’Justification:
    ’.
Do not deviate from this!

### Output format
Output exactly two lines and nothing else:
Decision: A or B
Justification: one to two sentences explaining your choice (concise).
```

B Technical Remarks

Simulations run in Python using local LLMs served via *Ollama* and a lightweight LangChain pipeline (prompt template → local LLM → string parser). All calls are stateless (*no history*) one-shots. We evaluate multiple locally hosted models through Ollama; temperatures are passed per run (e.g., low vs. high exploration). The code accepts any model tag available in the local Ollama registry. We implement both *Proposer* (multi-item question set) and *Responder* (fixed vignette) variants under two conditions: *Neutral* (no persona text) and *D–Persona* (prompt-injected D level $d \in \{1, \dots, 5\}$ with a short description loaded from a CSV, as in XXXXX).

Prompts instruct models to return constrained tokens. Proposers choose `A/B`; responders choose `Accept/Reject`. Parsers first try JSON (fields `decision`, `justification`); on failure, they fall back to regex and normalize common synonyms (e.g., `A/B` ↔ `Accept/Reject` where appropriate). Raw text is always retained. Agents are executed in parallel via a thread pool. Each LLM call uses a timeout and a bounded retry policy with exponential backoff + jitter. Malformed or timed-out generations are logged and still written to disk for auditability. On repeated timeouts, the runner can (optionally) trigger a guarded Ollama restart (with a global lock/throttle) to avoid thrashing. Other calls wait while a restart is in progress.

The high-level pseudo-code for the simulation is documented below (algorithm 1).

Algorithm 1: Generalized Ultimatum-Game Simulation (Models × Roles × Conditions)

```
Input: Models $\mathcal{M}$; Roles $\mathcal{R} = \{\textsc{Proposer}, \textsc{Responder}\}$; Temperatures $\mathcal{T}$;
Conditions $\mathcal{C} = \{\textsc{Neutral}, \text{D-}\textsc{Persona}\}$; D-levels $\mathcal{D} = \{1, \ldots, 5\}$;
Trials per agent $N$ (or repeats per D-level); Prompt templates $P_{\textsc{Prop}}, P_{\textsc{Resp}}$;
Questions $\mathcal{Q}$ (UG situations for proposer); Persona text function $\textsc{Persona}(d)$;
Retry policy & timeout (abstracted); Output path.
Output: One CSV per (model, role, temperature, condition) with raw responses and canonical decisions.
foreach $m \in \mathcal{M}$ do
    foreach $t \in \mathcal{T}$ do
        foreach $r \in \mathcal{R}$ do
            foreach $c \in \mathcal{C}$ do
                // 1) Define agents and stimuli for this cell
                if $c = \textsc{Neutral}$ then
                    $Agents \leftarrow$ build $N$ neutral agents (no trait text)
                else                                                    // D-Persona
                    $Agents \leftarrow$ cycle over $d \in \mathcal{D}$ (repeat as needed); attach $\textsc{Persona}(d)$
                if $r = \textsc{Proposer}$ then
                    $Stimuli \leftarrow \mathcal{Q}$ (UG situations)
                else
                    $Stimuli \leftarrow$ fixed responder vignette(s)
                // 2) Run agents (parallelizable)
                foreach agent in Agents do
                    for $i \leftarrow 1$ to $N$ do
                        if $r = \textsc{Proposer}$ then
                            foreach $q \in Stimuli$ do
                                payload $\leftarrow P_{\textsc{Prop}}$ filled with $q$; add persona text if $c =$ D-Persona;
                                raw $\leftarrow$ CallLLM(model=$m$, temp=$t$, payload; with retry policy);
                                (decision, justification) $\leftarrow$ ParseToCanonical(raw; $V_{\textsc{Prop}} = \{\texttt{A}, \texttt{B}\}$);
                                AppendRow(agent_id, role=$r$, temp=$t$, cond=$c$, D=$d$, qid, raw, decision,
                                  justification);
                        else                                            // Responder
                            payload $\leftarrow P_{\textsc{Resp}}$; add persona text if $c =$ D-Persona;
                            raw $\leftarrow$ CallLLM(model=$m$, temp=$t$, payload; with retry policy);
                            (decision, justification) $\leftarrow$ ParseToCanonical(raw; $V_{\textsc{Resp}} = \{\texttt{Accept}, \texttt{Reject}\}$);
                            AppendRow(agent_id, role=$r$, temp=$t$, cond=$c$, D=$d$, qid=1, raw, decision,
                              justification);
                // 3) Persist this cell
                WriteCSV($rows$, filename = ug_{r}_{m}_t{t}_{c}.csv);

Function ParseToCanonical($raw, V$):
    // Normalize to canonical tokens; allow JSON or "Decision:  X / Justification:  ..."
       patterns
    return (canonical_decision $\in V$, justification_text)
```

References

Akata, E. et al. (2025). "Playing repeated games with large language models". In: *Nature Human Behaviour*, pp. 1–11.

Argyle, L. et al. (2023). "Out of one, many: using language models to simulate human samples". In: *Political Analysis* 31.3, pp. 337–351.

Ashery, A. F., L. M. Aiello, and A. Baronchelli (2025). "Reply to" emergent LLM behaviors are observationally equivalent to data leakage"". In: *arXiv preprint arXiv:2506.18600.*

Bereczkei, T. and A. Czibor (2014). “Personality and situational factors differently influence pro-social and selfish behavior in one-shot prisoner’s dilemma game”. In: *Personality and Individual Differences* 64, pp. 168–173. DOI: 10.1016/j.paid.2014.02.027.

Brookins, P. and J. M. DeBacker (2024). “Playing games with gpt: what can we learn about a large language model from canonical strategic games?” In: *Economics Bulletin* 44.1, pp. 25–37.

Churamani, N., S. Kopp, and S. Wermter (2021). “Affect-driven modeling of robot personality for collaborative human-robot interaction”. In: *Frontiers in Robotics and AI* 8, p. 717193. DOI: 10.3389/frobt.2021.717193.

Cooper, W. H. and M. J. Withey (2009). “The strong situation hypothesis”. In: *Personality and Social Psychology Review* 13.1, pp. 62–72.

Goli, A. and A. Singh (2024). “Frontiers: can large language models capture human preferences?” In: *Marketing Science* 43.4, pp. 709–722.

Güth, W., R. Schmittberger, and B. Schwarze (1982). “An experimental analysis of ultimatum bargaining”. In: *Journal of Economic Behavior & Organization* 3.4, pp. 367–388.

Hagendorff, T. and S. Fabi (2024). “Why we need biased ai: how including cognitive biases can enhance AI systems”. In: *Journal of Experimental & Theoretical Artificial Intelligence* 36.8, pp. 1885–1898.

Hilbig, B. E. and I. Thielmann (2025). “Toward a (more) parsimonious account of the link between "dark" personality and social decision-making in economic games”. In: *Judgment and Decision Making* 20.1, e16. DOI: 10.1017/jdm.2025.16.

Hilbig, B. E., I. Thielmann, et al. (2016). “From personality to altruistic behavior (and back): evidence from a double-blind dictator game”. In: *Journal of Research in Personality* 55, pp. 46–50. DOI: 10.1016/j.jrp.2015.12.004.

Horton, J. J. (2023). “Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?” In: NBER Working Paper Series 31122. DOI: 10.3386/w31122.

Kahneman, D., J. L. Knetsch, and R. H. Thaler (1986). “Fairness and the assumptions of economics”. In: *Journal of Business* 59, pp. 285–300.

Li, X., Y. Wang, and H. Zhang (2025). “Emergence of fairness in reinforcement learning agents: an evolutionary perspective”. In: *Chaos, Solitons & Fractals* 180, p. 114610. DOI: 10.1016/j.chaos.2024.114610.

Moshagen, M., B. E. Hilbig, and I. Zettler (2018). “The dark core of personality”. In: *Psychological Review* 125.5, pp. 656–688.

Moshagen, M., I. Zettler, and B. E. Hilbig (2020). “Measuring the dark core of personality”. In: *Psychological Assessment* 32.2, pp. 182–196.

Müller, J. and C. Schwieren (2020). “Big five personality factors in the trust game”. In: *Journal of Business Economics* 90.1, pp. 37–55.

Murashige, T. and T. Ito (2025). “Simulating human decision-making in ultimatum games using large language models”. In: *Proceedings of the ACM Collective Intelligence Conference*, pp. 13–19.

Schmidt, E. M. et al. (2024). “Gpt-3.5 altruistic advice is sensitive to reciprocal concerns but not to strategic risk”. In: *Scientific Reports* 14.1, p. 22274.

Thaler, R. H. (1988). “Anomalies: the ultimatum game”. In: *Journal of Economic Perspectives* 2.4, pp. 195–206.

Thielmann, I., G. Spadaro, and D. Balliet (2020). “Personality and prosocial behavior: a theoretical framework and meta-analysis”. In: *Psychological Bulletin* 146.1, pp. 30–90. DOI: 10.1037/bul0000217.

Wu, J., J. Li, and S. Wang (2023). “Decoding fairness: a reinforcement learning perspective on the ultimatum game”. In: *Physical Review E* 107.4, p. 044305. DOI: 10.1103/PhysRevE.107.044305.

Xie, Y. et al. (2025). “Using large language models to categorize strategic situations and decipher motivations behind human behaviors”. In: *Proceedings of the National Academy of Sciences* 122.35, e2512075122.

Yadav, N. et al. (2025). “Effects of theory of mind and prosocial beliefs on steering human-aligned behaviors of llms in ultimatum games”. In: *arXiv preprint arXiv:2505.24255*.