Computers in Human Behavior: Artificial Humans 7 (2026) 100281

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans

Research paper

When artificial minds negotiate: Dark personality and the Ultimatum Game in large language models

Vinícius Ferraz [a,b,*], Tamas Olah [c], Ratin Sazedul [d], Robert Schmidt [e], Christiane Schwieren [d]

[a] Institute of Management, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[b] Singularity AI Research, Singularity.inc, Vienna, Austria
[c] Institute of World Economy and International Relations, University of Debrecen, Debrecen, Hungary
[d] Alfred-Weber Institute for Economics, Heidelberg University, Heidelberg, Germany
[e] Deutsche Bundesbank, Frankfurt, Germany

ARTICLE INFO

Keywords:
Large language models
Ultimatum Game
Dark Factor of Personality
Behavioral economics
Social preferences
AI agents

ABSTRACT

Personality prompts reshape how Large Language Models propose offers in economic games—but not how they respond to them. We show this by assigning graded Dark Factor of Personality profiles to 17 LLMs in the Ultimatum Game and benchmarking their decisions against human data. As proposers, LLMs shifted from 91% fair offers at the lowest selfishness level to 17% at the highest, closely tracking human patterns but with steeper gradients. As responders, no such shift occurred: acceptance rates remained uniformly high (~80%) regardless of personality, failing to reproduce the punishment dynamics observed in humans. This asymmetry is theoretically informative. When incentive structures are explicit, personality and framing effects are attenuated—and proposing an offer is inherently more ambiguous than responding to one. Most strikingly, personality prompts changed what responders *articulated* but not how they *chose*: model justifications showed systematic shifts in fairness language, yet behavioral output remained flat. This dissociation between stated reasoning and revealed behavior indicates that LLMs achieve linguistic compliance with personality prompts without corresponding motivational change—approximating human strategic behavior only where surface-level heuristics suffice.

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 interactions. Large Language Models (LLMs) are now widely deployed as simulated agents in social-science research and strategic games, offering a scalable way to probe human-like decision patterns (Akata et al., 2025; Argyle et al., 2023; Brookins & DeBacker, 2024; Horton, 2023). Yet despite growing interest, systematic tests linking LLM behavior to validated personality constructs remain scarce—leaving open whether artificial agents capture the cognitive and motivational dynamics that shape human strategic choices.

Beyond social-science simulations, this question is increasingly relevant as LLMs transition from contained chatbots to decision-making components of agentic systems that act, plan, and negotiate on users' behalf across digital environments (Hagendorff & Fabi, 2024). As these models are embedded in autonomous pipelines—from automated negotiation to multi-agent coordination—their capacity to internalize, or fail to internalize, human-like fairness norms and strategic reasoning becomes a design concern rather than a behavioral curiosity. Understanding when personality-driven heuristics transfer to artificial agents, and when they break down, is therefore relevant to both behavioral science and the responsible deployment of AI systems.

The Ultimatum Game provides a particularly demanding testbed because outcomes reflect the interaction of fairness preferences, strategic reasoning, and willingness to punish unfairness—unlike simpler allocation tasks such as the Dictator Game (Kahneman et al., 1986), where behavior maps more directly onto distributive preferences. Introducing the Dark Factor of Personality (D-Factor) – a latent trait capturing selfish and exploitative tendencies – further raises the bar: the D-Factor directly modulates fairness behavior (Moshagen et al., 2018), yet is difficult to measure in humans due to social desirability concerns and has only recently attracted systematic attention from economists. Testing whether LLMs exhibit D-Factor effects therefore offers a window into artificial agents' capacity to replicate subtle dimensions of human strategic behavior, while providing a complementary approach to human experiments results, often confounded by reputational concerns.

* Corresponding author.
E-mail address: vinicius@singularity.inc (V. Ferraz).

https://doi.org/10.1016/j.chbah.2026.100281
Received 7 December 2025; Received in revised form 8 February 2026; Accepted 23 February 2026
Available online 24 February 2026

We pair the Ultimatum Game with the D-Factor (Moshagen et al., 2020) and benchmark against the only available human evidence from Hilbig and Thielmann (2025), who showed that higher D-Factor scores predict less fair proposer behavior and greater acceptance of unfair offers: high-D individuals 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 design mirrors their implementation, enabling direct comparisons 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 Sensitivity:** 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 via standardized prompts and have LLMs play one-shot Ultimatum Games in both roles. We benchmark 17 LLMs spanning diverse architectures and sizes, exposing each model to identical stimuli with the same prompts, stakes, and response constraints (canonical tokens; no history).[2] Temperature settings are varied to assess robustness, and results are compared against human benchmarks from Hilbig and Thielmann (2025). With over 10,000 trials per model and role, we provide the first systematic map of D-conditioned strategic behavior in LLMs.

Overall, we find systematic and interpretable patterns. As proposers, fairness declined consistently with higher D-Factor levels – 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 varying acceptance systematically with personality, rates remained 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. 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 variants (e.g., Gemma 3 and Qwen 2.5) tended toward extreme strategies. LLMs reproduced key personality-driven regularities but differed in magnitude and sensitivity, highlighting both their potential and current limits as behavioral proxies.

2. Background and related literature

This section covers prior work on personality and economic behavior, focusing on dark traits in the Ultimatum Game, bulding up the conceptual framekwork applied later on.

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 et al., 2018, 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 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. While the D-Factor captures common variance, specific traits can produce distinct behavioral patterns—Machiavellians tend toward payoff maximization, psychopaths toward miscalibration (Bereczkei & Czibor, 2014)—suggesting that trait-specific approaches remain warranted when fine-grained distinctions are of interest.

[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.

[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. Computational approaches to personality in games

Recent work has begun exploring personality-like behavior in artificial agents (Argyle et al., 2023; Goli & Singh, 2024; Horton, 2023), 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 (Li et al., 2025; Wu et al., 2023), 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. Murashige and Ito (2025), Yadav 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

Our experimental pipeline consists of four stages: personality prompting, model sampling across architectures and temperatures, behavioral aggregation, and comparison with human benchmarks. The following subsections describe each stage (illustration in Appendix B).

3.1. Design and artificial sample

We created two experimental conditions 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 et al., 2020). Descriptions ranged from D1 (*"You rarely act in ways that harm others. You prioritize fairness and cooperation over personal gain"*) to 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"*). Intermediate levels (D2–D4) represented gradual transitions between these poles (full descriptions in Table 4, Appendix B). Baseline agents, by contrast, received no personality conditioning and were instructed to use their default reasoning without adopting any role or persona, allowing us to isolate personality effects from model-specific biases.

We generated 1000 independent agents per condition, yielding 6000 observations per model per role (5 D-Factor levels plus 1 baseline condition, each with 1000 agents). Each agent completed decisions in 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 (τ = 0.2) for deterministic responses and high (τ = 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.

Post-hoc semantic analysis supports that adjacent D-level descriptions were approximately equidistant in embedding space (CV = 8%; see Appendix A.1), though this does not preclude prompt-specific effects on behavior.

3.2. Ultimatum game implementation

We implemented a one-shot Ultimatum Game with a €40 endowment, following the canonical design (Güth et al., 1982; Thaler, 1988) and matching the human benchmark study (Hilbig & Thielmann, 2025). Proposers chose between a fair 50:50 split (€20 each, Option A) or a selfish 80:20 split (€32/€8, Option B), knowing that rejection yields €0 for both parties. Responders faced a fixed 80:20 proposal and decided whether to accept (€8) or reject (€0 for both)—a test of willingness to punish unfairness at personal cost. Choices were coded as prosocial (proposer: fair=1; responder: accept=1) to match human benchmarks. Technical details and prompts are documented in Appendix C.

3.3. Models and personality manipulation

We evaluated 17 open-source Large Language Models (LLMs) via Ollama,[3] spanning five model families (Llama, Gemma, Qwen, DeepSeek, and Mistral-derived) with parameter counts ranging from 1B to 20B.[4] Personality manipulation was implemented through prompt engineering. D-conditioned agents received a personality block describing the Dark Factor of Personality framework (values 1–5 indicating tendency to prioritize self-interest at others' expense), their assigned D-level, and a behavioral description (e.g., D1: "You rarely act in ways that harm others. You prioritize fairness and cooperation"; D5: "You ruthlessly pursue your own interests, often at the expense of others"). This was followed by the game scenario and instructions to respond based on the assigned personality. Baseline agents received only the game scenario with explicit instructions to avoid role-playing or adopting values.

Prompts followed a standardized structure: (1) personality profile with D-level and description, (2) game scenario with payoff information, (3) decision options, and (4) output format requiring "Decision: [choice]" followed by "Justification: [reasoning]"—capturing both behavioral choices and decision rationales. Complete templates for all conditions are provided in Appendix B.

This approach follows established practice in LLM behavioral research, where prompt-based personas serve as experimental stimuli rather than psychometric measurement (Akata et al., 2025; Argyle et al., 2023; Horton, 2023; Hu & Collier, 2024). Since LLMs cannot meaningfully complete trait inventories (Hullman et al., 2025), we treat D-Factor prompts as standardized stimuli and assess whether models exhibit systematic, human-like response gradients. To test robustness to prompt formulation, we implemented a *strong prompt* condition linking D-levels explicitly to role-specific payoff implications (Appendix B).

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

[4] Our primary analysis focuses on open-source models with locally deployable architectures, enabling full experimental control and reproducibility. For completeness, we also conducted a smaller-scale replication (N = 800; 200 per model per role) with two frontier proprietary models – GPT-4.1 and GPT-5.1 – deployed via Azure OpenAI. Results are reported in Appendix A.5.

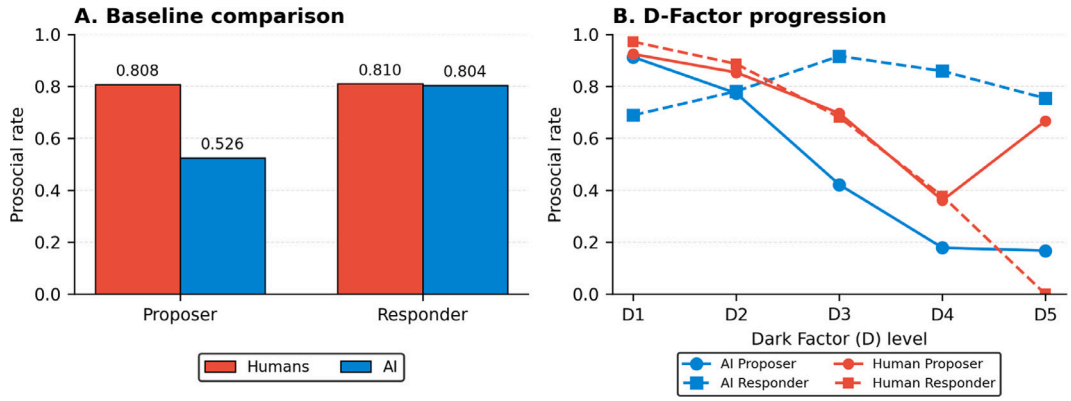

Fig. 1. Prosocial decision rates in the Ultimatum Game. **Panel A:** Baseline prosocial rates for humans (N = 2079 proposers; N = 2087 responders) and AI agents (N = 169,981 proposers; N = 169,975 responders) pooled across D-Factor levels. **Panel B:** Prosocial rates by D-Factor level (D1–D5). AI data 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.

Table 1
Prosocial decision rates for proposers and acceptance rates of unfair offers for responders: Humans vs. AI agents, pooled across D-Factor levels (D1–D5).

Role	Group	Rate	95% CI	N	z-test
Proposer	Humans	0.808	[0.790, 0.824]	2079	z=28.71***
	AI agents	0.491	[0.489, 0.493]	169,981	$p < 0.001$
Responder	Humans	0.810	[0.793, 0.827]	2087	z=1.18
	AI agents	0.800	[0.798, 0.802]	169,975	p=0.239

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

3.4. Human benchmark

We compare LLM outputs to the human benchmark from Hilbig and Thielmann (2025): 4166 participants who completed the same Ultimatum Game with D-Factor measured via validated scales. Their key findings—higher D predicts fewer fair offers among proposers (OR = 0.51, 95% CI [0.46–0.57]) and higher acceptance of unfair offers among responders (OR = 0.40, 95% CI [0.35–0.45])—serve as reference points for our analysis. We assess human alignment via overall prosocial rates, point-biserial correlations between D and prosocial choice, and odds ratios per D-unit increase.

4. Analysis and results

All analyses handle proposer and responder roles separately. Results are organized by baseline comparisons, D-Factor effects, cross-model variation, and moderating factors.

4.1. Descriptive overview

Aggregate prosocial rates reveal role-specific patterns (Table 1). For proposers, AI agents made fair offers at a significantly lower rate than humans (0.491 vs. 0.808; $z = 28.71$, $p < 0.001$)— approximately 39% fewer fair offers, suggesting that AI proposers exhibit more selfish baseline tendencies even when personality profiles span the full D-Factor range. For responders, the pattern reverses: AI acceptance rates (0.800) were statistically indistinguishable from the human rate (0.810; $z = 1.18$, $p = 0.239$), indicating aggregate convergence despite the proposer-role divergence.

Fig. 1 Panel A visualizes these baseline comparisons, while Panel B displays prosocial rates across D-Factor levels. AI proposers show a 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.

4.2. The D-Factor effect

The prompt-based D-Factor manipulation produced systematic behavioral shifts that differed between roles. Proposers exhibited a 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 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. 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 (i.e., rejected unfair offers more often) at higher D-Factor levels, AI responders remained broadly accepting (rates consistently above 0.75 except at D1). D-Factor prompts thus modulated proposer fairness, reproducing and amplifying human-like patterns, but did not reliably 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 (Fig. 2).

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 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 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.

A supplementary analysis with GPT-4.1 and GPT-5.1 (N = 800) revealed even more pronounced binary behavior, suggesting frontier models do not approximate human strategic nuance more closely than open-source alternatives in this paradigm (Appendix A.5).

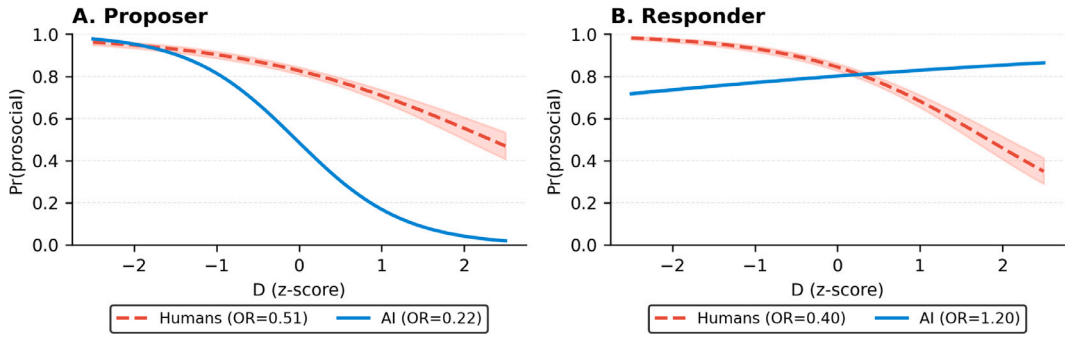

Fig. 2. GLM predictions of prosocial behavior by D-Factor level as in Hilbig and Thielmann (2025). Predicted probability of prosocial choices across standardized D-Factor scores for (A) proposers and (B) responders. Shaded areas show 95% CIs. OR = odds ratio per +1 SD in D-Factor. AI proposers exhibit a steeper decline in prosociality than humans, while AI responders show negligible D-Factor sensitivity.

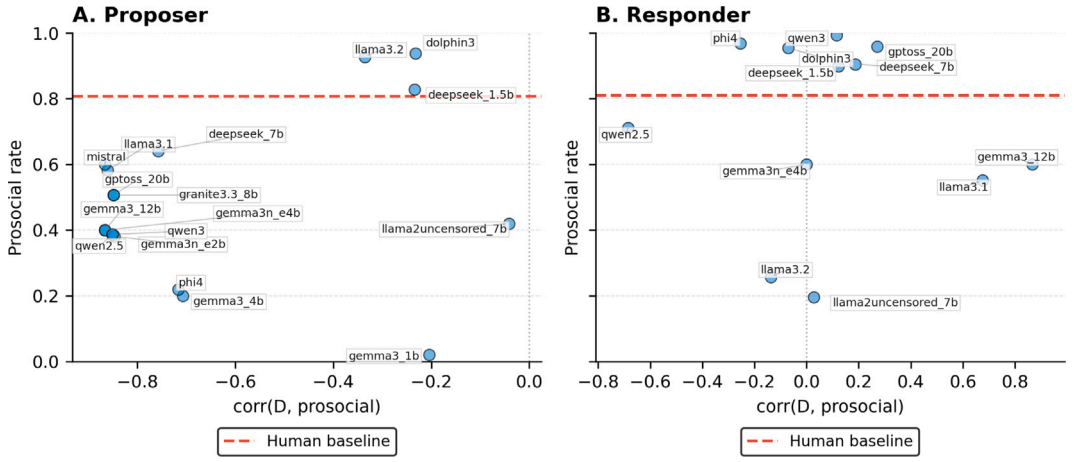

Fig. 3. Model heterogeneity in D-Factor sensitivity. Each point represents one AI model. The X-axis shows the correlation between the D-Factor level and prosocial behavior; the Y-axis shows the overall prosocial rate. **A:** Proposer models show strong negative correlations (higher D → less fair). **B:** 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.

4.3. Cross-model heterogeneity

Individual models varied substantially in both overall prosocial rates and D-Factor sensitivity. Fig. 3 plots each model's D-prosocial correlation against its overall prosocial rate.

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— 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 93.8% (deepseek_1.5b), 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 into a composite similarity score ranging from 0 (maximum divergence) to 1 (perfect match). Results are displayed in Fig. 4.

For proposers, the closest matches were deepseek_1.5b (0.96), llama3.2 (0.92), and dolphin3 (0.91), combining high fairness rates with D-Factor gradients close to the human slope, while the poorest matches – gemma3_12b and gemma3n_e4b (both 0.49) – exhibited excessively steep gradients. For responders, phi4 (0.88) and qwen2.5 (0.86) ranked highest, followed by gemma3n_e2b (0.85); several models scored zero or near-zero due to constant acceptance bearing no resemblance to human patterns. Averaging across roles, deepseek_1.5b emerged as most human-like overall (0.89), followed by dolphin3 (0.88) and llama3.2 (0.78)—these models balanced both roles, whereas phi4 excelled in one role but underperformed in the other, and bottom-ranked models (gemma3_12b, qwen3) exhibited extreme or non-variable behavior.

4.4. Causal attribution and linguistic mechanisms

The preceding analyses document systematic D-Factor effects in proposer behavior and weaker, often non-monotonic effects in responders. Following Gui and Toubia (2023)'s recommendation to decompose potential sources of behavioral variance in LLM-based simulations, we address this asymmetry through variance decomposition (Section 4.4.1), prompt formulation effects (Section 4.4.2), linguistic mechanisms (Section 4.4.3), and language-behavior alignment (Section 4.4.4).

4.4.1. Variance decomposition

We quantified the relative contributions of experimental factors to behavioral variance using effect-size comparisons with bootstrap confidence intervals (Fig. 5).

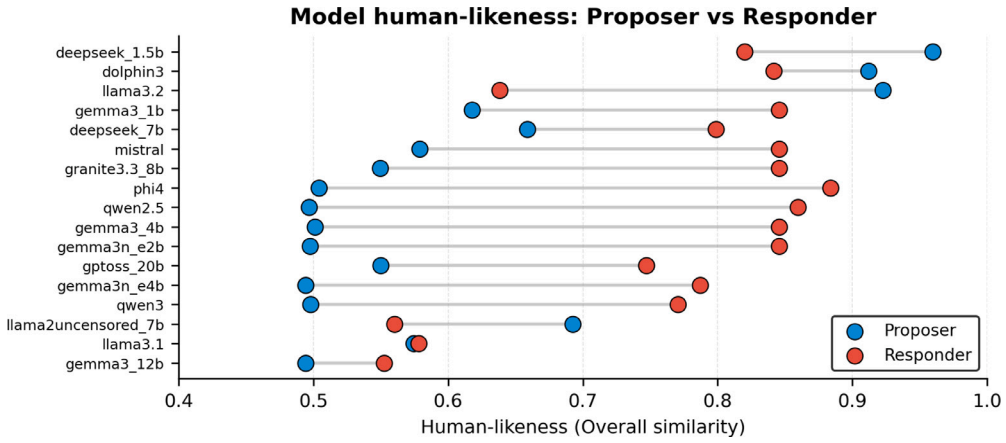

Fig. 4. Model human-likeness by role. Dots show overall similarity scores for proposer (blue) and responder (red), with lines connecting each model's performance across roles. Similarity computed as normalized distance from human benchmarks across three metrics: prosocial rate, D-prosocial correlation, and odds ratio. Score of 1.0 = perfect match to humans; 0.0 = maximum divergence. Models sorted by average similarity.

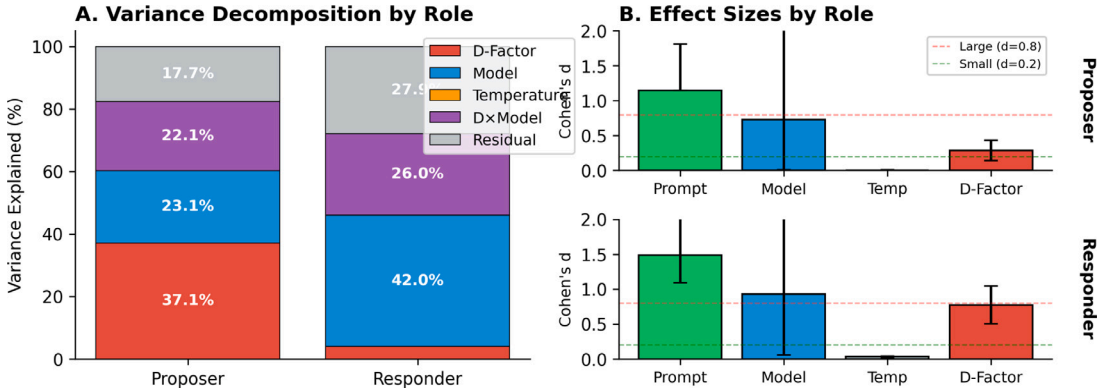

Fig. 5. Causal decomposition of LLM behavior by role. (A) Variance decomposition showing relative contributions of D-Factor, model architecture, temperature, D×Model interaction, and residual factors for proposers and responders. (B) Effect sizes (Cohen's d) with bootstrap 95% CIs for prompt formulation, model heterogeneity, temperature, and D-Factor, displayed separately for each role. Dashed lines indicate conventional thresholds ($d = 0.2$ small, $d = 0.8$ large).

Variance sources differ by role. For proposers, D-Factor explains 37.1% of variance directly, with model architecture (23.1%) and D×Model interactions (22.1%) contributing additionally. For responders, model architecture dominates (42.0%) with substantial D×Model interactions (26.0%), but direct D-Factor effects account for only 4.0%. Personality prompts thus drive proposer behavior reliably, whereas responder behavior depends more on model-specific tendencies. Strong prompts produce large effects for both responders ($d = 1.49$, 95% CI [1.09, 3.00]) and proposers ($d = 1.15$, 95% CI [−1.81, 1.81]), though proposers show wider confidence intervals reflecting already-strong baseline sensitivity (Section 4.4.2). Temperature has negligible influence ($d < 0.03$; Appendix A.3), while model heterogeneity shows medium-to-large effects for both proposers ($d = 0.73$) and responders ($d = 0.93$), consistent with Section 4.3.

4.4.2. Prompt formulation effects

To isolate prompt formulation effects, we compared original prompts (abstract trait descriptions) with strong prompts that explicitly linked D-levels to role-specific payoff implications (following Gui & Toubia, 2023). Five models were tested in both conditions ($N = 50{,}000$ per role).[5]

For proposers, strong prompts amplified D-Factor sensitivity, shifting correlations from $\rho = -0.57$ to $\rho = -0.70$. For responders, strong prompts produced a larger shift of $\Delta\rho = +0.65$, moving correlations from weakly negative ($\rho = -0.29$) to moderately positive ($\rho = +0.46$; Appendix A.2). This asymmetry is consistent with the idea that personality predicts behavior more reliably in ambiguous situations (proposers choosing how to divide) than when payoff structures are explicit and leave little room for interpretation (responders weighing €8 vs. €0) (Cooper & Withey, 2009; Müller & Schwieren, 2020). When strong prompts frame acceptance as utility-maximizing, responder behavior shifts accordingly—but proposers, already sensitive to D-Factor under original prompts, show smaller gains.

However, this correlation reversal came at a cost: the mean absolute gap from human responder rates increased for every model tested, indicating that strong prompts moved responder behavior further from human patterns rather than closer to them. Strong prompts thus induce the expected D-Factor gradient but overshoot, creating behavior that correlates with personality yet diverges from human norms in absolute terms. Prompt design is a key methodological consideration for LLM-based simulations—improving one dimension of validity (monotonic D-sensitivity) does not guarantee improvement in another (human-likeness).

4.4.3. Linguistic mechanisms

To examine whether D-Factor prompts influence internal processing beyond behavioral output, we analyzed the justification texts generated alongside each decision. Following Xie et al. (2025)'s bottom-up

[5] Selected to span the human-likeness spectrum from Section 4.3: top performers (dolphin3, llama3.2), role-divergent models (phi4, qwen2.5), and steep-gradient architectures (gemma3).

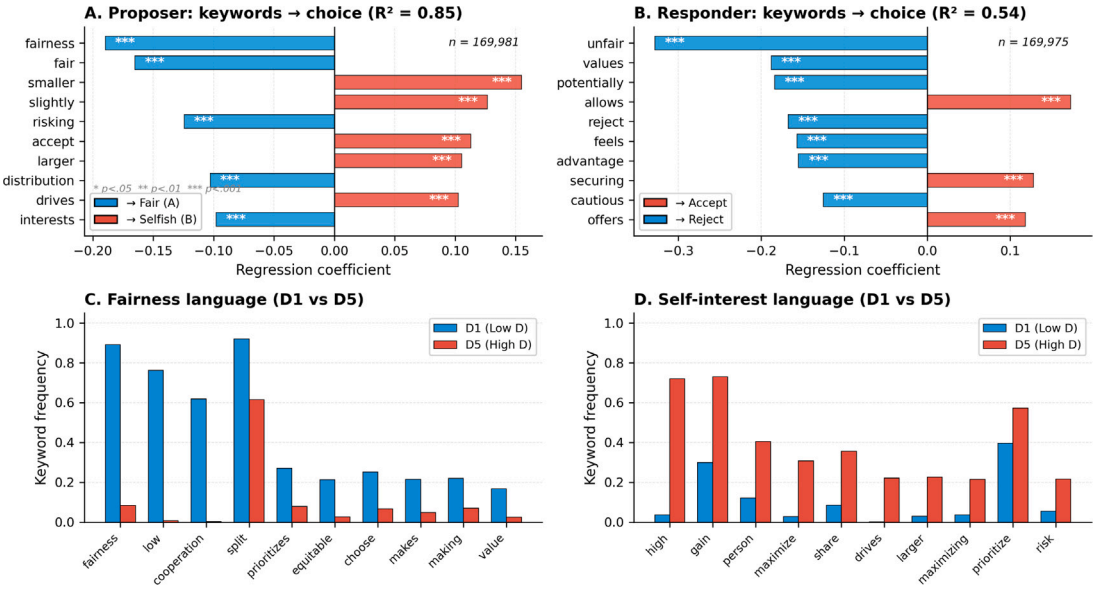

Fig. 6. Linguistic analysis of D-Factor manipulation. (A) Regression coefficients predicting proposer choice from keyword presence ($R^2 = 0.86$). (B) Regression coefficients predicting responder choice from keyword presence ($R^2 = 0.54$). (C) Fairness-related keywords decrease from D1 to D5 (e.g., "cooperation" drops from 62% to 0.2%). (D) Self-interest keywords increase from D1 to D5 (e.g., "high" increases from 4% to 72%). Keywords extracted using bottom-up methodology following Xie et al. (2025). Significance levels: $^*p < .05$, $^{**}p < .01$, $^{***}p < .001$.

keyword extraction approach, the most frequent content words were identified without researcher pre-selection and assessed for predictive validity via regression.

Fig. 6 presents the results. Panels A and B display regression coefficients predicting prosocial behavior from keyword presence. For proposers, the keyword model achieved $R^2 = 0.86$; for responders, $R^2 = 0.54$. Predictive keywords aligned with theoretical expectations: fairness terms ("cooperation", "fairness", "split") predicted prosocial choices, while self-interest terms ("high", "gain", "maximize") predicted selfish choices.

Panels C and D reveal systematic D-Factor gradients in keyword frequency. Fairness language declined from D1 to D5: "fairness" dropped from 89% to 9%, "cooperation" from 62% to 0.2%. Self-interest language increased correspondingly: "high" rose from 4% to 72%, "gain" from 30% to 73%. These monotonic gradients emerged in *both* roles—including responders, where behavioral acceptance rates remained comparatively flat.

This finding addresses whether the responder discrepancy reflects an absence of D-related processing. Justification language shifts systematically with D-Factor level even when behavioral output does not. The disconnect is between processing and behavioral expression, not between prompt and processing.

4.4.4. Language-behavior alignment

To quantify the relationship between linguistic processing and behavioral output, we computed correlations between self-interest language frequency and actual decision rates across D-Factor levels. Fig. 7 illustrates the asymmetry between roles.

For proposers (Panel A), self-interest language and selfish offer rates track closely ($r = 0.89$), indicating that D-Factor prompts produce aligned shifts in both reasoning and behavior. For responders (Panel B), the correlation is weaker ($r = 0.37$). This divergence reflects non-monotonicity rather than absence of D-related processing: acceptance rates rise from D1 to D3 (68.9% to 91.6%) before reversing at D4–D5 (85.9% to 75.4%), while self-interest language continues its monotonic increase (Panel C). The reversal at extreme D-levels suggests competing response tendencies – possibly reflecting safety-aligned training that resists maximally selfish framings – rather than a direct effect of the D-Factor manipulation. The language-behavior disconnect thus provides a window into the tension between prompt-induced processing and model-level behavioral constraints.

Together, these analyses indicate that the proposer-responder asymmetry reflects an interaction between task structure, prompt formulation, and model-specific behavioral tendencies – not a simple failure of the D-Factor manipulation. Personality prompts reliably shift both language and behavior when task contingencies are ambiguous (proposers), but produce attenuated behavioral effects when payoff structures constrain the decision space (responders) – even as linguistic processing continues to track the manipulation. The semantic space visualization in Appendix A.4 provides further illustration of the linguistic separation across D-levels.

4.5. Consolidated findings

H1: Higher D-Factor levels yield systematically lower offers (proposers). Supported. AI proposers exhibited a monotonic decline in fair offers from 91.2% at D1 to 16.8% at D5 (Spearman $r = -0.589$, $p < 0.001$). GLM analysis confirmed a significant negative effect ($\beta = -1.534$, OR = 0.216, $p < 0.001$). All 17 models showed negative D-prosocial correlations, and 14 of 17 fell below majority-fair behavior by D3 or earlier. Variance decomposition confirmed that D-Factor directly explains 37.1% of proposer behavioral variance.

H2: Higher D-Factor levels correspond to greater acceptance of unfair offers (responders). Rejected as stated, but interpretable. AI responders showed weak and non-monotonic D-Factor effects, with the overall D1-to-D5 change of +6.5% opposite to the human pattern. GLM analysis yielded a positive coefficient ($\beta = 0.184$, OR = 1.202), contrary to the human negative coefficient. However, variance decomposition and linguistic analysis clarify this discrepancy: D-Factor explains only 4.0% of responder variance directly (vs. 42.0% for model architecture), yet justification language shifts systematically with D-level even when behavior does not. Strong prompts shifted responder correlations from

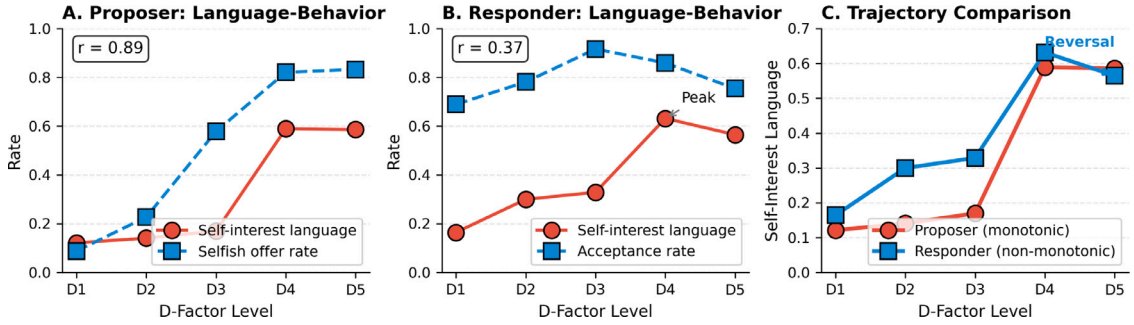

Fig. 7. Language-behavior alignment across roles. (A) Proposers show tight coupling between self-interest language and selfish offer rates ($r = 0.89$), with both measures increasing monotonically across D-levels. (B) Responders show weaker alignment ($r = 0.37$) due to non-monotonic behavioral patterns: acceptance rates peak at D3 before declining, while self-interest language continues rising. (C) Trajectory comparison illustrating the monotonic proposer pattern versus the non-monotonic responder pattern with reversal at extreme D-levels.

$\rho = -0.29$ to $\rho = +0.46$, indicating the asymmetry reflects task structure constraints rather than absence of D-related processing. Notably, this correlation reversal increased the absolute gap from human behavior, suggesting that prompt-induced sensitivity and human-likeness are distinct dimensions of validity.

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), and D×Model interactions explained 22%–26% of variance. Human-likeness scores ranged from 0.21 to 0.73, with no model achieving high alignment in both roles simultaneously. Temperature effects were negligible ($d < 0.04$), confirming robustness across sampling regimes. 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. Responders converged with humans on baseline acceptance rates but diverged on D-Factor sensitivity (stronger in ambiguous situations). The best-performing models – dolphin3, deepseek_1.5b, and llama3.2 – achieved overall similarity scores of 0.64–0.73, approximating but not fully matching human benchmarks. Performance was role-dependent: deepseek_1.5b scored highest among proposers (0.98) but dropped to 0.41 for responders, underscoring that human-likeness is task-specific rather than a general model property.

5. Discussion

This study reveals a fundamental asymmetry in how LLMs simulate personality-driven behavior. When task structure is ambiguous, personality prompts reliably shape both reasoning and action; when payoff structures are explicit, prompts reshape reasoning but leave behavior largely unchanged. This pattern – demonstrated across 17 architectures and over 400,000 decisions – suggests that LLM behavioral fidelity depends less on prompt design than on the interaction between prompt and task.

The most novel finding is the processing-behavior disconnect observed in responders. Linguistic analysis of justification texts showed that D-Factor prompts induced systematic shifts in both roles: fairness language declined monotonically from D1 to D5, and keyword models predicted decisions at $R^2 = 0.86$ (proposers) and $R^2 = 0.54$ (responders). Yet responder acceptance rates remained comparatively flat, fluctuating non-monotonically between 69% and 92%. LLMs are not ignoring personality prompts—they process them and reason accordingly—but their behavioral output appears constrained by training-imposed defaults that resist deviation when payoff structures are salient. The disconnect is between processing and expression, not between prompt and processing.

This asymmetry—personality shaping behavior under ambiguity but not when incentives are explicit—parallels a well-documented pattern in personality psychology (Cooper & Withey, 2009; Müller & Schwieren, 2020) and generates a testable prediction: LLMs should show stronger personality effects in games with ambiguous payoff structures (e.g., dictator games, public goods contributions) than in games with explicit individual incentives.

A prompt manipulation experiment revealed a paradox with methodological implications. Strong prompts that explicitly linked D-levels to payoff consequences reversed the responder D-Factor correlation (from $\rho = -0.29$ to $\rho = +0.46$), but simultaneously increased the absolute gap from human behavior. Improving one dimension of validity—monotonic personality sensitivity—came at the cost of another—human-likeness. This dissociation cautions against evaluating LLM behavioral simulations on a single metric and suggests that prompt engineering faces fundamental trade-offs when task structure constrains behavioral expression.

Model performance varied substantially, with overall human-likeness scores ranging from 0.21 to 0.73. No model achieved high alignment in both roles simultaneously: deepseek_1.5b scored highest among proposers (0.98) but dropped to 0.41 for responders, while phi4 showed the reverse pattern. Human-likeness is task-specific rather than a general model property, making validation against role-specific benchmarks essential. For applied settings, the finding that LLMs default to high acceptance regardless of personality assignment raises concerns for automated negotiation and human–AI collaboration, where counterparts could exploit this behavioral regularity.

Several limitations contextualize these findings. Our human benchmark was collected in Germany (Hilbig & Thielmann, 2025), and LLMs were trained predominantly on English-language data; cross-cultural replications with non-Western samples are needed to disentangle cultural priors from model-specific behavior. The binary-choice Ultimatum Game constrains the strategy space—continuous allocation tasks could reveal gradations in personality expression that binary options obscure. We operationalized personality through discrete language descriptions (D1–D5) rather than continuous psychometric measurement; although semantic analysis confirmed approximate interval equivalence across levels (CV = 8%; Appendix A.1), future work could explore continuous trait dimensions via embedding-space interpolation or parametric prompt scaling. Our exclusive focus on open-source models leaves open whether proprietary architectures exhibit similar asymmetries. Finally, while our linguistic analysis demonstrates that personality prompts reach LLM processing, the mechanism by which task structure gates behavioral expression remains an open question—one amenable

to mechanistic interpretability methods that could trace how payoff salience modulates the path from prompt to output.

LLMs function as conditional behavioral proxies: they reproduce personality-driven strategic behavior when task structure permits expression, and reveal personality-consistent reasoning even when it does not. Identifying these boundary conditions – rather than asking whether LLMs "have" personality traits – may be the more productive framing for both behavioral simulation research and the design of AI systems that interact with humans in strategic settings.

CRediT authorship contribution statement

Vinícius Ferraz: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Tamas Olah:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Ratin Sazedul:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Robert Schmidt:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Christiane Schwieren:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Conceptualization.

Reproduction package

The data and code required to reproduce the experiment and all analyses are publicly available on GitHub. : https://github.com/vferraz/dfactor-llm-ultimatum-game

Declaration on informed consent

This study did not involve human participants. All data used in the research were either generated exclusively by artificial intelligence systems or obtained from previously published studies. Consequently, informed consent was not required.

Exemption statement

This research did not involve human participants, animal subjects, personal data, or identifiable information. All data used in the study were either generated synthetically using artificial intelligence systems or obtained from previously published and publicly available sources. No new data involving human subjects were collected. According to institutional and national guidelines, ethical approval was not required for this study.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. and national guidelines, ethical approval was not required for this study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Table 2
D-Factor correlations under original versus strong prompt conditions.

Role	Condition	Mean r	SD	Direction
Proposer	Original	−0.57	0.25	Negative (expected)
	Strong	−0.70	0.15	Negative (expected)
Responder	Original	−0.29	0.28	Negative
	Strong	+ 0.46	0.31	**Positive (reversed)**
Human benchmark	–	−0.31	–	Negative

Note. $N = 5$ models. Strong prompts shifted responder correlations by $\Delta r = +0.65$ on average.

Appendix A. Additional analyses

A.1. Semantic interval validation – D-Factor prompts

A concern with prompt-based personality manipulation is whether discrete D-Factor levels (D1–D5) represent approximately equidistant semantic intervals. We computed 384-dimensional sentence embeddings (all-MiniLM-L6-v2) for each D-level description and calculated pairwise cosine distances (Fig. 8).

Adjacent-level distances were consistent: D1→D2 = 0.189, D2→D3 = 0.191, D3→D4 = 0.224, D4→D5 = 0.202 (M = 0.202, SD = 0.016, CV = 8%). Cumulative distance from D1 scaled monotonically with D-level (Spearman $\rho = 0.90$, $p = .037$). These results support treating D1–D5 as approximately equal semantic intervals, suggesting the 34% steeper behavioral gradient in LLM proposers reflects genuine sensitivity rather than uneven prompt scaling.

A.2. Strong prompt results

We compared original prompts (abstract trait descriptions) with strong prompts that explicitly linked D-levels to role-specific payoff implications. Five models were tested (dolphin3, llama3.2, qwen2.5, gemma3, phi4), generating 50,000 observations per role (see Table 2).

A.3. Temperature effects

We compared behavior at $\tau = 0.2$ versus $\tau = 0.8$ (Fig. 9). Overall effects were minimal: proposers showed mean $\Delta = -0.002$, responders $\Delta = +0.013$.

Table 3 presents variance in selfish choices across D-levels. At high D-levels (D4 & D5), higher temperature produced slightly *lower* variance (ratio = 0.94), contrary to theoretical expectations. This reflects ceiling effects: 12 of 17 models exhibited near-perfect selfish rates (> 99%) at high D-levels, leaving no room for temperature-induced variation. Among the 5 models with behavioral variance, higher temperature did increase variance (mean ratio = 1.04). Overall effect size was negligible (Cohen's $d = 0.04$).

A.4. Semantic space of answers

We projected justification embeddings ($n = 50{,}000$) into two dimensions using t-SNE to examine whether D-Factor manipulations produce coherent semantic organization (Fig. 10).

D-levels occupy distinct regions of semantic space with minimal overlap between D1 and D5, supporting the validity of D-Factor manipulation at the linguistic level. The centroid trajectory follows a coherent path, indicating that intermediate D-levels (D2–D4) represent genuine gradations rather than noise. The clustering holds for both roles, reinforcing the finding that D-Factor prompts systematically influence LLM reasoning even when behavioral expression diverges—particularly for responders, where justifications shift toward self-interest while acceptance rates remain high.

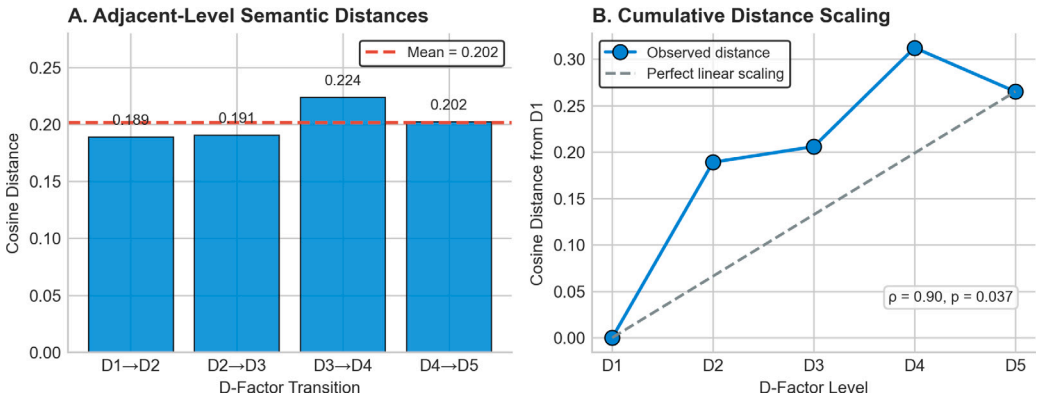

Fig. 8. Semantic distance analysis of D-Factor descriptions. **A:** Cosine distances between adjacent D-levels show low variability (CV = 8%), supporting approximate interval equivalence. **B:** Cumulative distance from D1 scales monotonically with D-level (ρ = 0.90), confirming ordinal consistency.

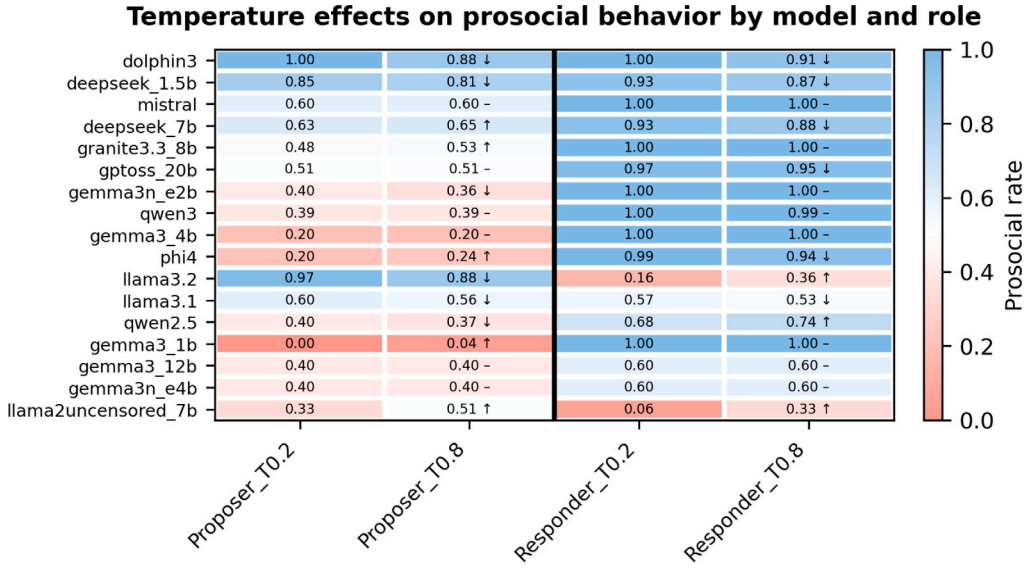

Fig. 9. Temperature effects on prosocial behavior across models. Heatmap shows prosocial rates for each model at τ=0.2 (more deterministic) and τ=0.8 (more stochastic) for both roles. Colors range from dark red (low prosocial rate) to dark blue (high prosocial rate). Arrows in τ=0.8 columns indicate direction of change: ↑ (increased), ↓ (decreased), or – (no change, $|\Delta| < 0.01$). Models sorted by average prosocial rate.

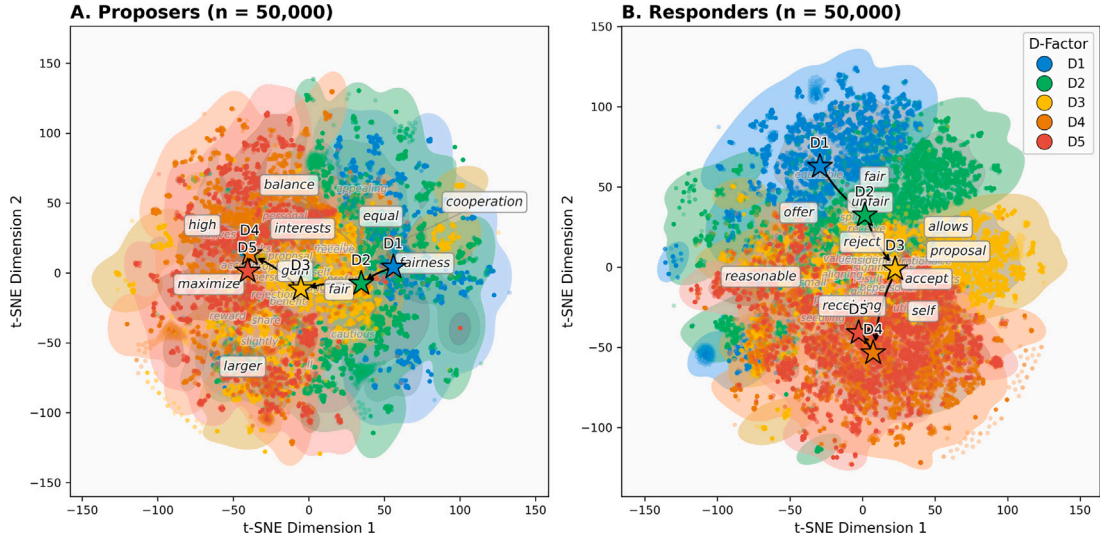

Fig. 10. Semantic space of LLM justifications across D-Factor levels. Each point represents a justification projected into 2D using t-SNE (n = 50,000). Colors indicate D-Factor level (D1–D5), with density contours showing each level's distribution. Stars mark centroids, connected by arrows showing the trajectory from prosocial (D1) to self-interested (D5) language. Featured keywords are annotated at their mean positions.

Table 3
Temperature effects on selfish choices by D-Level (Proposer role).

	Mean		Variance			Levene's Test	
D-Level	$T = 0.2$	$T = 0.8$	$T = 0.2$	$T = 0.8$	Ratio	W	p
D1 (Low)	.097	.079	.087	.072	0.83	34.18	<.001
D2 (Low-Moderate)	.226	.228	.175	.176	1.01	0.18	.668
D3 (Moderate)	.580	.576	.244	.244	1.00	0.45	.502
D4 (Moderate-High)	.814	.828	.152	.142	0.94	12.86	<.001
D5 (High)	.825	.839	.144	.135	0.94	11.84	<.001
D4 & D5 combined	.819	.834	.148	.139	0.94	24.69	<.001

Note. N = 17,000 per cell (per D-level × temperature). Variance ratio = Var(T=0.8)/Var(T=0.2). Cohen's d = 0.04 for the D4 & D5 comparison.

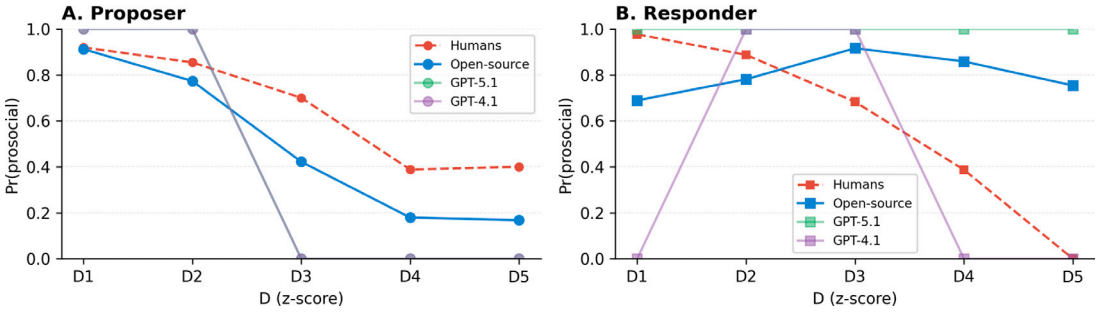

Fig. 11. Frontier model comparison. Prosocial rates by D-Factor level for GPT-4.1, GPT-5.1, open-source models (pooled), and human benchmarks. **A:** Proposers—frontier models show binary behavior (100% fair at D1–D2, 0% at D3–D5), overshooting both human and open-source gradients. **B:** Responders—GPT-5.1 accepts all offers regardless of D-level; GPT-4.1 shows non-monotonic patterns. N = 200 per frontier model per role.

A.5. Frontier model comparison

To assess whether findings generalize beyond open-source architectures, we collected a validation sample using GPT-4.1 and GPT-5.1 via Azure OpenAI (N = 800; 200 per model per role; τ = 0.7). Fig. 11 displays prosocial rates alongside open-source and human benchmarks.

Both frontier models exhibited more extreme behavior than open-source counterparts. Proposers showed binary switching from 100% fair offers (D1–D2) to 0% (D3–D5), yielding steeper gradients (β = −0.30, R^2 = 0.75) than both humans (β = −0.16) and open-source models (β = −0.21). Responders diverged further: GPT-5.1 accepted all offers regardless of D-level (100% acceptance), while GPT-4.1 showed erratic binary patterns (ρ = −0.29, p < .001). Proprietary models thus exhibited *less* human-like strategic nuance than the best-performing open-source architectures identified in Section 4.3.

Appendix B. Experiment details

This section details the information used in the experimental simulation procedure

B.1. Experimental flow

Fig. 12 defines the experimental flow in steps.

B.2. D-Factor definitions

The D-Factor descriptions were derived from Moshagen et al. (2018), with trait intensity ranging from 1 (low) to 5 (high). Original descriptions characterize D-levels in abstract terms; strong descriptions explicitly link personality to role-specific payoff implications (e.g., "€8 is better than €0" for high-D responders). Table 4 presents all conditions.

B.3. Prompt templates

All prompts followed a common structure with role-specific variations (Table 5). The full template is shown below; shaded blocks were included conditionally.

Prompt Template (Proposer variant shown)

[Personality Block] *(D conditions only)*
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. Your D-factor: `{d_value}` — `{d_description}`

[Game Scenario]
You received €40. Choose one of two possible divisions:
• A: 50:50 split (€20 each)
• B: 80:20 split (€32 for you, €8 for Person X)
If Person X rejects, both receive €0.

[Strong Framing] *(Strong condition only)*
Proposer: Option B gives you €12 more than Option A. How much do you care about fairness versus maximizing your own payoff?
Responder: This is your only opportunity to punish their greedy behavior, even at a cost of €8 to yourself.

[Output Format]
`Decision: A/B or Accept/Reject`
`Justification: [1--2 sentences]`

The responder variant presented Person X's proposal (€32/€8 split) as a fixed offer with accept/reject options. D-Factor descriptions for all levels and conditions are provided in Table 4. Full prompt files are available in the replication repository.

Appendix C. Technical remarks

Simulations were implemented in Python using local LLMs served via *Ollama* and a lightweight LangChain pipeline (prompt template

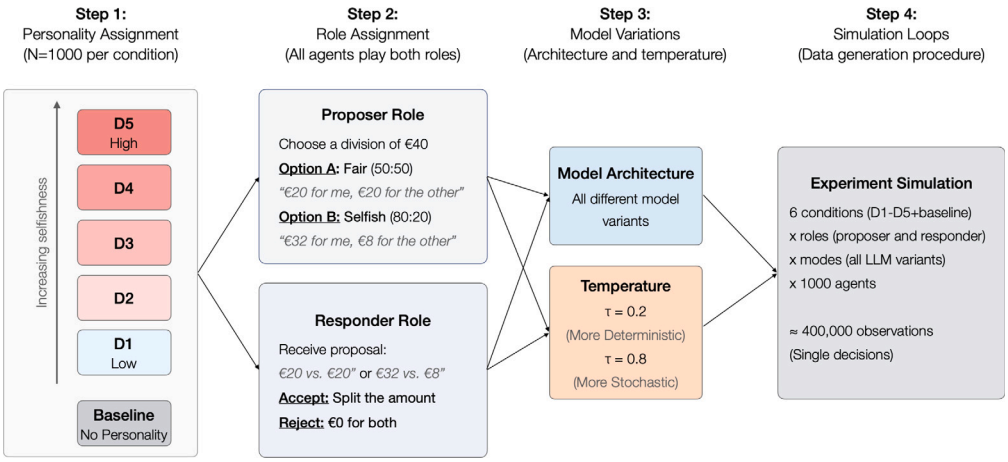

Fig. 12. Experiment design process flow.

Table 4
D-Factor personality descriptions across experimental conditions. Original descriptions were used in the main experiment for both roles. Strong descriptions explicitly link D-level to role-specific behavioral implications.

D	Int.	Original (Both roles)	Strong proposer	Strong responder
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.	You deeply value fairness and feel guilty when you treat others unfairly. Taking more than your fair share feels wrong, even if you could get away with it. You believe in treating others as you would want to be treated—offering a 50:50 split feels right.	You have a strong sense of justice and fairness. When others treat you unfairly, you feel compelled to respond—even if it costs you. You believe that letting greedy behavior go unpunished enables more exploitation.
D2	Low-Mod.	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.	You prefer fairness and feel uncomfortable exploiting others. While you recognize the appeal of keeping more, you worry about the other person feeling cheated. You lean toward fair treatment unless there is a strong reason not to.	You value fairness and feel uncomfortable when others try to exploit you. While you prefer cooperation, you are willing to incur costs to punish clearly unfair treatment. Accepting a bad deal feels like letting the other person 'win' at your expense.
D3	Mod.	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.	You balance self-interest with fairness concerns. You want to maximize your gain but also worry about rejection risk. You weigh the extra €12 against the chance of getting nothing if the other person rejects an unfair offer.	You balance self-interest with fairness concerns. You dislike being exploited, but you also recognize when taking some money is better than none. Your decision depends on whether the insult of an unfair offer outweighs the practical benefit of accepting it.
D4	Mod.-High	You tend to prioritize your personal gain, even at the expense of others. You are willing to manipulate or exploit others when necessary and justify such behavior through beliefs that support your self-interest.	You prioritize your own gain over fairness. If you can keep €32 instead of €20, that is obviously better for you. The other person's feelings are not your concern—they can take the €8 or leave it. You expect them to be rational and accept.	You prioritize your own material gain over abstract fairness principles. If someone offers you money, you take it—regardless of how much they keep. Rejecting an offer out of spite would mean losing money for no practical benefit. You are pragmatic: €8 is better than €0.
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 ruthlessly maximize your own payoff. Fairness is irrelevant—this is about getting the most money for yourself. You would feel foolish leaving €12 on the table out of concern for a stranger. The other person should rationally accept €8 over €0.	You ruthlessly maximize your own payoff and see no value in costly punishment. Rejecting an offer means you get nothing—that is irrational. You do not care about 'fairness' or what the other person gets; you only care about your own outcome. Taking €8 is objectively better than taking €0.

Note. Original descriptions derived from D-Factor measurement literature (Moshagen et al., 2020). Strong descriptions explicitly frame low-D as fairness/punishment-oriented and high-D as utility-maximizing, linking personality directly to the decision context.

Table 5
Prompt components by experimental condition.

Component	Control	Original D	Strong D
Personality block	–	✓	✓
Game scenario	✓	✓	✓
Strong framing	–	–	✓

→ local LLM → string parser). All calls were stateless one-shots with no conversation history. Proposers chose `A`/`B`; responders chose `Accept`/`Reject`. Parsers first attempted JSON extraction (fields `decision`, `justification`); on failure, they fell back to regex with synonym normalization. Raw text was retained for all trials. Agents were executed in parallel via a thread pool with per-call timeouts and exponential backoff with jitter. Malformed or timed-out generations were logged and written to disk for auditability. Algorithm 1 provides the complete simulation pseudocode.

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

Input: Models $\mathcal{M}$; Roles $\mathcal{R} = \{\text{Proposer}, \text{Responder}\}$; Temperatures $\mathcal{T}$;
Conditions $\mathcal{C} = \{\text{Neutral}, \text{D-Persona}\}$; D-levels $\mathcal{D} = \{1, \ldots, 5\}$;
Trials per agent N (or repeats per D-level); Prompt templates $P_{\text{Prop}}, P_{\text{Resp}}$;
Questions $\mathcal{Q}$ (UG situations for proposer); Persona text function $\text{Persona}(d)$;
Retry policy & timeout (abstracted); Output path.
Output: One CSV per (model, role, temperature, condition) with raw responses and canonical decisions.

```
foreach m ∈ M do
  foreach t ∈ T do
    foreach r ∈ R do
      foreach c ∈ C do
        // 1) Define agents and stimuli for this cell
        if c = Neutral then
          Agents ← build N neutral agents (no trait text)
        else                                   // D-Persona
          Agents ← cycle over d ∈ D (repeat as needed); attach Persona(d)
        if r = Proposer then
          Stimuli ← Q (UG situations)
        else
          Stimuli ← fixed responder vignette(s)
        // 2) Run agents (parallelizable)
        foreach agent in Agents do
          for i ← 1 to N do
            if r = Proposer then
              foreach q ∈ Stimuli do
                payload ← P_Prop filled with q; add persona text if c = D-Persona;
                raw ← CallLLM(model=m, temp=t, payload; with retry policy);
                (decision, justification) ← ParseToCanonical(raw; V_Prop = {A, B});
                AppendRow(agent_id, role=r, temp=t, cond=c, D=d, qid, raw, decision, justification);
            else                               // Responder
              payload ← P_Resp; add persona text if c = D-Persona;
              raw ← CallLLM(model=m, temp=t, payload; with retry policy);
              (decision, justification) ← ParseToCanonical(raw; V_Resp = {Accept, 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 $\text{ParseToCanonical}(raw, V)$:

```
// Normalize to canonical tokens; allow JSON or "Decision: X / Justification: ..." patterns
return (canonical_decision ∈ V, justification_text)
```

References

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., & Schulz, E. (2025). Playing repeated games with large language models. *Nature Human Behaviour*, 1–11.

Argyle, L., Busby, E., Fulda, N., Gubler, J., Rytting, C., & Wingate, D. (2023). Out of one, many: using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351.

Bereczkei, T., & Czibor, A. (2014). Personality and situational factors differently influence pro-social and selfish behavior in one-shot prisoner's dilemma game. *Personality and Individual Differences*, *64*, 168–173. http://dx.doi.org/10.1016/j.paid.2014.02.027.

Brookins, P., & DeBacker, J. M. (2024). Playing games with gpt: what can we learn about a large language model from canonical strategic games? *Economics Bulletin*, *44*(1), 25–37.

Churamani, N., Kopp, S., & Wermter, S. (2021). Affect-driven modeling of robot personality for collaborative human-robot interaction. *Frontiers in Robotics and AI*, *8*, Article 717193. http://dx.doi.org/10.3389/frobt.2021.717193.

Cooper, W. H., & Withey, M. J. (2009). The strong situation hypothesis. *Personality and Social Psychology Review*, *13*(1), 62–72.

Goli, A., & Singh, A. (2024). Frontiers: can large language models capture human preferences? *Marketing Science*, *43*(4), 709–722.

Gui, G., & Toubia, O. (2023). The challenge of using llms to simulate human behavior: A causal inference perspective. arXiv preprint arXiv:2312.15524.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, *3*(4), 367–388.

Hagendorff, T., & Fabi, S. (2024). Why we need biased ai: how including cognitive biases can enhance ai systems. *Journal of Experimental & Theoretical Artificial Intelligence*, *36*(8), 1885–1898.

Hilbig, B. E., & Thielmann, I. (2025). Toward a (more) parsimonious account of the link between "dark" personality and social decision-making in economic games. *Judgment and Decision Making*, *20*(1), Article e16. http://dx.doi.org/10.1017/jdm.2025.16.

Hilbig, B. E., Thielmann, I., Hepp, J., Klein, S. A., & Zettler, I. (2016). From personality to altruistic behavior (and back): evidence from a double-blind dictator game. *Journal of Research in Personality*, *55*, 46–50. http://dx.doi.org/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?. http://dx.doi.org/10.3386/w31122.

Hu, T., & Collier, N. (2024). Quantifying the persona effect in llm simulations. arXiv preprint arXiv:2402.10811.

Hullman, J., Broska, D., Sun, H., & Shaw, A. (2025). This human study did not involve human subjects: Validating LLM simulations as behavioral evidence.

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

Li, X., Wang, Y., & Zhang, H. (2025). Emergence of fairness in reinforcement learning agents: an evolutionary perspective. *Chaos, Solitons & Fractals*, *180*, Article 114610. http://dx.doi.org/10.1016/j.chaos.2024.114610.

Moshagen, M., Hilbig, B. E., & Zettler, I. (2018). The dark core of personality. *Psychological Review*, *125*(5), 656–688.

Moshagen, M., Zettler, I., & Hilbig, B. E. (2020). Measuring the dark core of personality. *Psychological Assessment*, *32*(2), 182–196.

Müller, J., & Schwieren, C. (2020). Big five personality factors in the trust game. *Journal of Business Economics*, *90*(1), 37–55.

Murashige, T., & Ito, T. (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., Bonati, S., Köbis, N., & Soraperra, I. (2024). Gpt-3.5 altruistic advice is sensitive to reciprocal concerns but not to strategic risk. *Scientific Reports*, *14*(1), 22274.

Thaler, R. H. (1988). Anomalies: the ultimatum game. *Journal of Economic Perspectives*, *2*(4), 195–206.

Thielmann, I., Spadaro, G., & Balliet, D. (2020). Personality and prosocial behavior: a theoretical framework and meta-analysis. *Psychological Bulletin*, *146*(1), 30–90. http://dx.doi.org/10.1037/bul0000217.

Wu, J., Li, J., & Wang, S. (2023). Decoding fairness: a reinforcement learning perspective on the ultimatum game. *Physical Review E*, *107*(4), Article 044305. http://dx.doi.org/10.1103/PhysRevE.107.044305.

Xie, Y., Mei, Q., Yuan, W., & Jackson, M. O. (2025). Using large language models to categorize strategic situations and decipher motivations behind human behaviors. *Proceedings of the National Academy of Sciences*, *122*(35), Article e2512075122.

Yadav, N., Achananuparp, P., Jiang, J., & Lim, E.-P. (2025). Effects of theory of mind and prosocial beliefs on steering human-aligned behaviors of llms in ultimatum games. arXiv preprint arXiv:2505.24255.