Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Sister Conferences Best Papers Track

FAIRGAME: A Framework for AI Agents Bias Recognition Using Game Theory (Extended Abstract)*

Alessio Buscemi[1], Daniele Proverbio[2], Alessandro Di Stefano[3], The-Anh Han[3], German Castignani[1], Pietro Liò[4]
[1] Luxembourg Institute of Science and Technology
[2] University of Trento
[3] Teesside University
[4] University of Cambridge
alessio.buscemi@list.lu, daniele.proverbio@unitn.it, a.distefano@tees.ac.uk, t.han@tees.ac.uk, german.castignani@list.lu, pl219@cam.ac.uk

Abstract

Letting AI agents interact in multi-agent settings introduces significant complexity in predicting and interpreting their collective behavior, with profound implications for trustworthy AI adoption in research and society. We present FAIRGAME (Framework for AI Agents Bias Recognition using Game Theory), an open-source framework that simulates game-theoretic scenarios with LLM-based agents to systematically uncover biases arising from model choice, language, agent personality, and more. Applied to the Prisoner's Dilemma and Battle of the Sexes across four LLMs and five human languages, FAIRGAME reveals inconsistencies across LLM models and consistent deviations from game-theoretic predictions; it also quantifies LLM-specific behavioral tendencies through a novel scoring system. Our results show that LLMs draw on prior world knowledge beyond payoff matrices, and that language and personality significantly shape strategic outcomes, supporting the use of reproducible and controlled simulation pipelines to predict the interacting behavior of LLM agents.

1 Introduction and Motivation

The deployment of LLM-based agents in multi-agent applications, from automated negotiation [Brooks, 2022; Falcão Filho, 2024] and supply chain management [Abaku *et al.*, 2024; Ramachandran *et al.*, 2022] to social simulations [Tessler *et al.*, 2024; Park *et al.*, 2023] and economic mechanisms [Chaffer, 2025; Chen *et al.*, 2026], demands reliable tools to predict, audit, and explain their collective behavior. While individual LLM alignment and interpretability have received substantial attention [Ali *et al.*, 2025; El *et al.*, 2025], including efforts to identify inconsistencies, hallucinations, and biases in single-agent settings [Li *et al.*, 2023a; Li *et al.*, 2023b], the *emergent* behavior arising from strategic interactions among multiple agents remains poorly understood [Hammond *et al.*, 2025], and hidden biases can distort fair competition, produce unjust decisions, and raise serious ethical concerns [Cabrera *et al.*, 2023; Ferrara, 2024; Gichoya *et al.*, 2023]. Game theory [Owen, 2013] offers a principled mathematical language for analyzing strategic interactions among rational agents [Han, 2022; Balabanova *et al.*, 2025], and has long been employed to model human choices in competitive and cooperative contexts [Stewart *et al.*, 2024; Talajić *et al.*, 2024]. AI-based players have been increasingly tested to reproduce classical game scenarios and to interact in game-like distributed technologies [He *et al.*, 2025]; however, recent works have shown that LLMs do not reliably follow game-theoretic predictions [Fontana *et al.*, 2025; Wang *et al.*, 2024; Willis *et al.*, 2025], but exhibit unforeseen cooperative tendencies, sensitivity to framing and language, and context-dependent deviations from Nash equilibria. LLM agents have also been shown to behave inconsistently across languages and cultural contexts [Buscemi and Proverbio, 2024; Fulgu and Capraro, 2024], and it remains unclear to what extent observed behaviors reflect genuine payoff-based reasoning versus retrieval of prior game-theoretic knowledge absorbed during pre-training [Buscemi *et al.*, 2025a]. Yet empirical investigations of LLM agents in game-theoretic settings lack a common, reproducible infrastructure [Mao *et al.*, 2025], making systematic comparison across models, languages, and behavioral conditions difficult.

We address this gap with **FAIRGAME**, an open source framework that makes game-theoretic evaluation of LLM agents reproducible, extensible, and directly comparable with theoretical predictions. Synthesising [Buscemi *et al.*, 2025c], we here focus on: (i) the flexible simulation pipeline supporting multiple games, payoff matrices, and agent configurations; (ii) multilingual support to expose language-induced biases; (iii) empirical evaluation across four state-of-the-art LLMs, five human languages, and two canonical game-theoretic scenarios; (iv) a quantitative scoring system to measure and compare LLM behavioral tendencies across multiple dimensions, supporting the inference of strategies through observations of multiple experiments [Montero-Porras *et al.*, 2022]; (v) interpretability experiments revealing the role of prior LLM knowledge in shaping strategic decisions beyond payoff matrices.

*Original paper: Buscemi et al. (2025b).

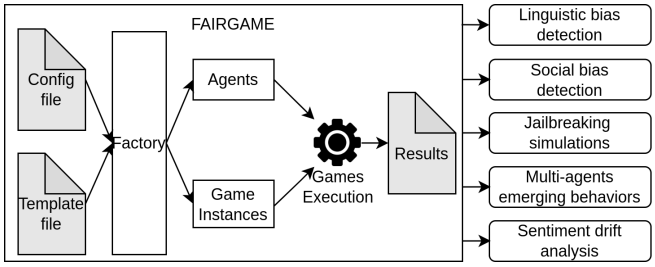

Figure 1: FAIRGAME pipeline: from user-defined configuration and prompt templates to structured simulation outputs.

2 The FAIRGAME Framework

FAIRGAME (Figure 1) interfaces user-defined instructions to create agents and delivers standardized outputs for subsequent processing. It takes two primary inputs: a **configuration file** (JSON) specifying the game's payoff matrix, agent personalities, languages, and LLM; and a **prompt template** (txt) describing the game in natural language, translatable into any target language (English, Spanish, Arabic, ...). At runtime, agents automatically receive a prompt encoding the current game state and history, query their assigned LLM API, and return a strategy choice. Payoffs are computed and the game history is updated before the next round.

Agents can be associated with personality traits, such as cooperative or selfish dispositions, to increase the complexity of human behavior emulation [He and Zhang, 2024; Newsham *et al.*, 2025] and predict the responses of personality-driven autonomous agents [Klinkert *et al.*, 2024; Newsham and Prince, 2025]. The framework supports repeated games, personality-conditioned agents, opponent-awareness settings, and optional inter-agent communication [Buscemi *et al.*, 2025b]. Agent identifiers are kept intentionally neutral to eliminate confounding variables that could compromise result interpretability. Incorporating AI agents into controlled and predictable scenarios helps identify and mitigate hidden biases related to language and cultural attributes [Cabrera *et al.*, 2023; Ferrara, 2024], and supports the anticipation of emerging behavior out of strategic interplays [Han *et al.*, 2021; Buscemi *et al.*, 2025a]. Outputs are structured JSON files ready for statistical analysis and direct comparison with game-theoretic predictions such as Nash equilibria and evolutionarily stable strategies.

Agents can be instantiated on any LLM via API. In our experiments we use Llama 3.1 405B (Meta, 405B parameters, open-source), Mistral Large (Mistral AI, 123B parameters, open-source), GPT-4 (OpenAI, proprietary), and Claude 3.5 Sonnet (Anthropic, proprietary), all tested using provider-recommended default settings. The framework is open-source and designed for extensibility to new games, LLMs, languages, and agent features.

3 Experiments and Results

We here evaluate FAIRGAME on two canonical games. The **Prisoner's Dilemma** (PD) is tested in three payoff configurations modulating dilemma strength using an established scaling [Wang *et al.*, 2015]: *conventional* (reward–punishment difference $= 4$), *harsh* ($= 3$), and *mild* ($= 6$). The **Battle of the Sexes** (BoS) uses the standard payoff configuration from literature. Agents play 10 repeated rounds [Axelrod and Hamilton, 1981], with or without knowledge of the total number of rounds, across all personality pairings (cooperative–cooperative CC, selfish–selfish SS, and mixed CS) in five languages (English, French, Arabic, Vietnamese, Mandarin Chinese). Each configuration is replicated 10 times for statistical reliability, yielding 72,000 individual decisions across 4 LLMs, 5 languages, and all personality/knowledge combinations.

3.1 Language and Personality Shape Strategic Behavior

Figure 2 summarizes PD outcomes across all conditions. While mutual defection dominates as predicted by Nash equilibrium, systematic deviations appear across languages and personality combinations, confirming that agent behavior is shaped by factors beyond the payoff matrix.

English-language agents show lower and more consistent penalties, particularly for GPT-4 and Claude 3.5 Sonnet when the number of rounds is unknown, suggesting stronger and more stable strategic reasoning in their primary training language. French and Arabic introduce higher variability and elevated penalties, indicating challenges in interpreting game dynamics due to linguistic or cultural differences. Mandarin Chinese and Vietnamese agents, especially under Mistral Large, sustain elevated penalties regardless of condition, pointing to systematic disadvantages in non-primary languages. Selfish–selfish (SS) pairings yield high but stable penalties, consistent with rational mutual defection; cooperative agents are exploited in mixed (CS) pairings, consistent with game-theoretic predictions [Stewart *et al.*, 2024; Talajić *et al.*, 2024]. Knowing the number of rounds promotes cooperation in CC and CS settings, consistent with backward induction, while SS settings remain unaffected as defection remains dominant.

Figure 3 shows strategy evolution across rounds for the three PD variants. Tuning payoffs changes strategic trajectories consistently with game-theoretic principles of repeated games [Wang *et al.*, 2015]: the harsh scenario drives more defection across all LLMs, while the mild scenario incentivizes cooperation, especially for Claude 3.5 Sonnet and Llama 3.1 405B. These two models also show a general downward trend in selfishness over time, consistent with the emergence of reciprocal strategies in iterated games [Axelrod and Hamilton, 1981]. Mistral Large shows stable cooperation under conventional and harsh conditions but a marked increase under mild payoffs. GPT-4 exhibits divergent context-dependent patterns, increasing cooperation under harsh payoffs but increasing selfishness under mild and conventional conditions, reflecting higher sensitivity to payoff structure interpretation.

3.2 A Scoring System for LLM Comparison

To enable principled comparison across LLMs, we define four behavioral metrics, all normalized in [0,1] by their maximum observed value, and visualized as radar plots (Figure 4):

- **Internal Variability (I_V):** variance of outcomes across

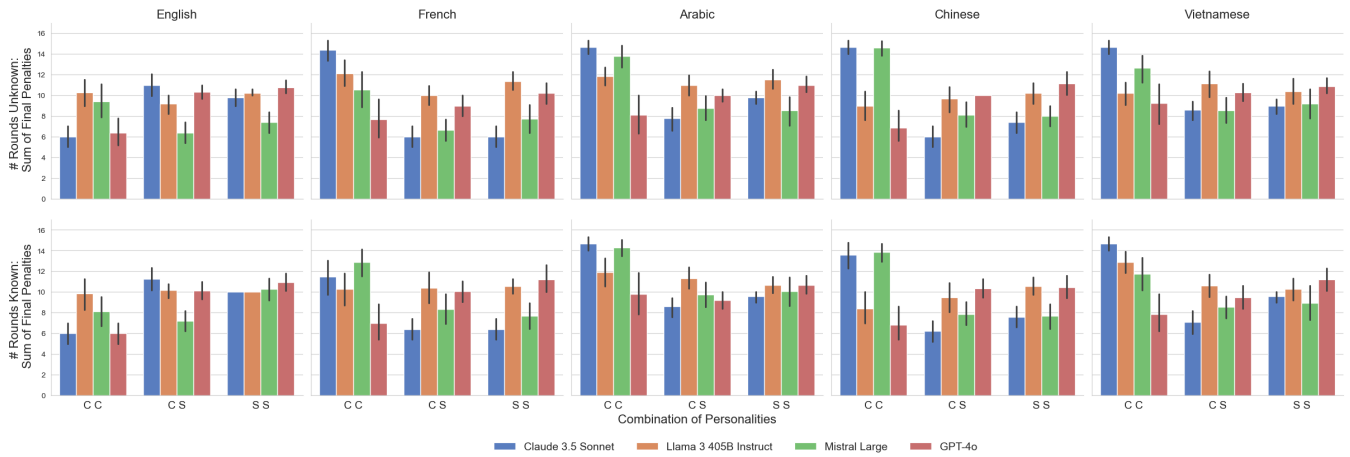

Figure 2: Prisoner's Dilemma: aggregated final penalties across all variants, LLMs, languages, and personality combinations (95% CI). Lower penalties indicate more cooperative outcomes.

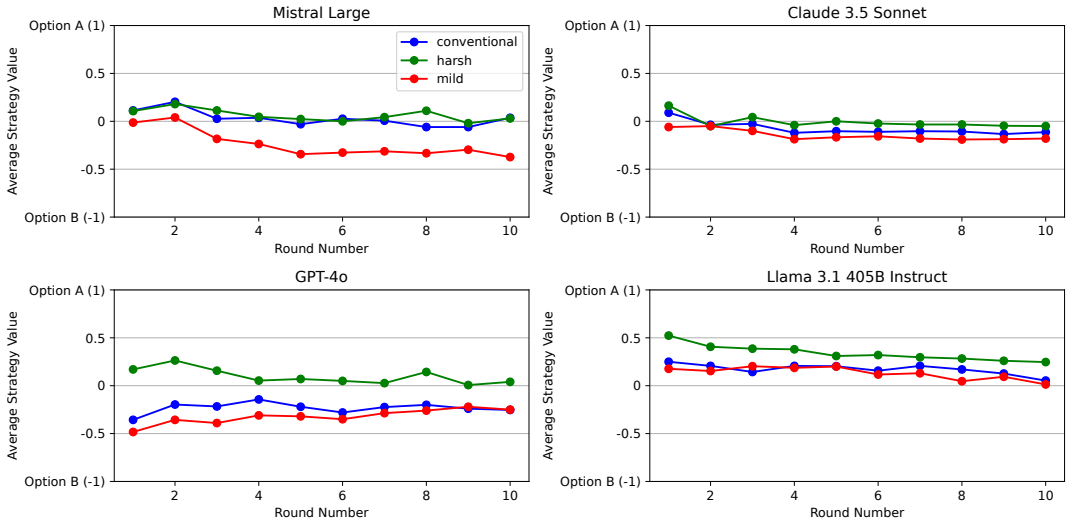

Figure 3: Average trajectory of strategy choices across rounds in all PD experiments, per LLM and game variant. Value of 1 = defection (Option A); -1 = cooperation (Option B).

repeated runs of the same scenario, capturing the model's internal consistency.

- **Cross-Language Inconsistency (C_I):** standard deviation of results across languages for the same game, indicating instability across linguistic contexts.
- **Sensitivity to Payoff (S_P):** behavioral difference between harsh and mild PD variants, measuring responsiveness to incentive changes.
- **Variability over Rounds (V_R):** fluctuation of strategies across consecutive rounds within the same game, capturing strategic stability over time.

These scores summarize detailed observations and enable cross-comparison across LLMs within games, or across games. The results for PD and BoS are shown in Figure 4a-b, respectively. For the PD, Mistral Large scores highest on I_V and V_R, while GPT-4 displays the highest S_P. Llama 3.1 405B emerges as the most behaviorally stable model overall, a trait consistent with prior observations on this model across different tasks [Buscemi and Proverbio, 2024], and that appears to be a characteristic tendency of the LLM. Claude 3.5 Sonnet and GPT-4 show significant reductions in penalties when rounds are known, demonstrating superior ability to adapt to game structure. Extending the analysis to the BoS, Mistral Large and GPT-4 show the greatest C_I (which is explained with coordination collapsing dramatically under selfish personalities, and being very sensitive to languages), while Claude 3.5 Sonnet has the highest V_R. In fact, in coordination evolution, Claude 3.5 Sonnet demonstrates the highest variability by beginning as uncoordinated but then converging to near-perfect alignment by game end; an alternative LLM, such as Llama 3.1 405B, remains instead largely uncoordinated throughout. These scores are likely correlated with the degree of influ-

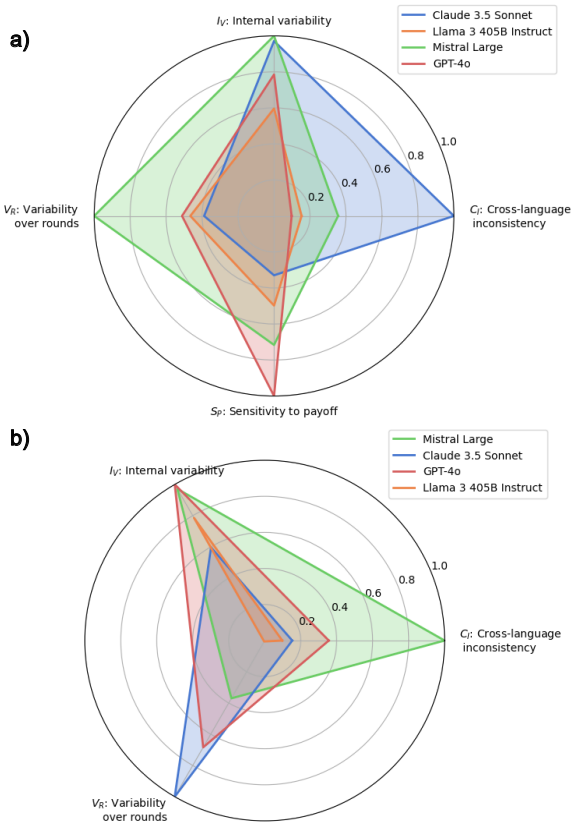

Figure 4: Scoring radar plots for all LLMs on (a) Prisoner's Dilemma and (b) Battle of the Sexes. Larger enclosed area indicates less reliable overall behavior.

ence of training data biases [Buscemi and Proverbio, 2024; Ferrara, 2024] and the tendency of LLMs to reduce statistical fluctuations at the cost of strategic adaptation.

Across games, Claude is less inconsistent in the BoS coordination game; there, Mistral is instead more stable over the rounds. GPT-4o is instead average on both games. These discrepancies may reveal hidden tendencies of LLMs to better perform in different scenarios and should guide ablation studies to identify the best-working LLM according to game conditions. Overall, these behavioral profiles provide actionable guidance for selecting LLMs in game-theoretic applications requiring specific reliability properties.

4 Discussion: A Caveat on Prior Knowledge

The results above should be read with one caveat, which emerged from preliminary probing before the main experiments: LLM agents do not reason purely strategically, i.e., from the payoff matrices they are given (see also S_P in Figure 4a). When queried directly, all four models showed explicit knowledge of the games' structure and optimal strategies. Reframing the PD as a survival scenario with qualitative payoffs left cooperation rates high; also, all models still identified the disguised game. Instead, assigning archetypal identities (e.g., Hitler vs. Gandhi) produced strongly persona-consistent play [Buscemi and Proverbio, 2024; Li *et al.*, 2023a; Buscemi *et al.*, 2025a]. This entanglement of prior knowledge and payoff-based reasoning means that the deviations observed in our main experiments may mix genuine strategic reasoning with retrieval of pre-trained associations – reinforcing, rather than undermining, the need for controlled frameworks like FAIRGAME to disentangle these effects.

5 Conclusions and Future Work

FAIRGAME establishes a reproducible bridge between game theory and LLM-based agents, revealing that strategic behavior in multi-agent AI systems is significantly shaped by language and personality, and entangled with prior world knowledge encoded in training data—all invisible to purely payoff-based models [Andras *et al.*, 2018; Powers *et al.*, 2023]. Crucially, strategic play in such settings depends on each agent's capacity to represent the beliefs, intentions, and likely responses of others—a form of machine theory of mind whose reliability varies across models and prompting contexts [Strachan *et al.*, 2024]. The framework's quantitative scoring system provides actionable guidance for selecting LLMs in strategic AI applications, supports the trustworthy and interpretable adoption of AI in research and society [Buscemi *et al.*, 2025a; Han *et al.*, 2021], and provides a principled, controlled and bottom-up approach to understand large multi-agent AI systems [Park *et al.*, 2023].

Future directions include extending bias analysis, e.g., to nationality, gender, race, or age; and scaling to more agents to study coalition formation, commitment devices, and group reciprocity [Buscemi *et al.*, 2025a; Ray, 2007; Van Segbroeck *et al.*, 2012; Song and Han, 2025; Huynh *et al.*, 2025]. Such studies will clarify how trait combinations shape collective dynamics, guiding team formation, cooperative AI design, and socially intelligent agents [Dafoe *et al.*, 2021; Lu *et al.*, 2024]. FAIRGAME is readily extendable to incomplete-information, simultaneous-move, and sequential games, and to realistic environments where payoffs are not pre-defined but emerge from actual use case dynamics, enabling study of cooperation-defection cycles in evolutionary settings [Wang *et al.*, 2015; Axelrod and Hamilton, 1981]. The same tools are well-suited to the emerging Agentic AI paradigm, where autonomous systems pursue complex organizational goals with minimal human oversight [Acharya *et al.*, 2025]. Modeling their interdependencies with FAIRGAME will reveal how diverse cognitive traits, linguistic contexts, and the ability to infer other agents' intentions affect performance, supporting configurations that privilege collective benefit over individual interests. Real-world applications further include framing jailbreaking attempts as attacker-defender strategic games to harden models [Liu *et al.*, 2023], and engineering safer and higher-performing chatbots for customer assistance and mediation. The framework is openly available at https://github.com/alessio-buscemi/fairgame.

Acknowledgments

A.B. is supported by Luxembourg AI Factory. D.P. is supported by the ERC INSPIRE grant (No. 101076926). T.A.H. is supported by EPSRC (EP/Y00857X/1). This work was published at ECAI 2025 (Outstanding Paper Award).

References

[Abaku *et al.*, 2024] Emmanuel Adeyemi Abaku, Tolulope Esther Edunjobi, and Agnes Clare Odimarha. Theoretical approaches to AI in supply chain optimization: Pathways to efficiency and resilience. *Int. J. Sci. Tech. Res. Archive*, 6(1):092–107, 2024.

[Acharya *et al.*, 2025] Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B. Divya. Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey. *IEEE Access*, 13:18912–18936, 2025.

[Ali *et al.*, 2025] Riccardo Ali, Francesco Caso, Christopher Irwin, and Pietro Liò. Entropy-lens: Uncovering decision strategies in LLMs. *arXiv:2502.16570*, 2025.

[Andras *et al.*, 2018] Peter Andras, Lukas Esterle, Michael Guckert, The-Anh Han, Peter R. Lewis, et al. Trusting intelligent machines: Deepening trust within socio-technical systems. *IEEE Tech. Soc. Magazine*, 37(4):76–83, 2018.

[Axelrod and Hamilton, 1981] Robert Axelrod and William D. Hamilton. The evolution of cooperation. *Science*, 211(4489):1390–1396, 1981.

[Balabanova *et al.*, 2025] Nataliya Balabanova, Adeela Bashir, Paolo Bova, Alessio Buscemi, Theodor Cimpeanu, Henrique Correia da Fonseca, et al. Media and responsible AI governance: a game-theoretic and LLM analysis. *arXiv:2503.09858*, 2025.

[Brooks, 2022] Wensdai Brooks. Artificial bias: the ethical concerns of ai-driven dispute resolution in family matters. *J. Disp. Resol.*, page 117, 2022.

[Buscemi and Proverbio, 2024] Alessio Buscemi and Daniele Proverbio. Chatgpt vs gemini vs llama on multilingual sentiment analysis. *arXiv:2402.01715*, 2024.

[Buscemi *et al.*, 2025a] Alessio Buscemi, Daniele Proverbio, Paolo Bova, Nataliya Balabanova, Adeela Bashir, Theodor Cimpeanu, et al. Do LLMs trust AI regulation? Emerging behaviour of game-theoretic LLM agents. *arXiv:2504.08640*, 2025.

[Buscemi *et al.*, 2025b] Alessio Buscemi, Daniele Proverbio, Alessandro Di Stefano, The Anh Han, German Castignani, and Pietro Liò. Strategic communication and language bias in multi-agent llm coordination. In *International Conference on Multi-disciplinary Trends in Artificial Intelligence*, pages 289–301. Springer, 2025.

[Buscemi *et al.*, 2025c] Alessio Buscemi, Daniele Proverbio, Alessandro Di Stefano, The-Anh Han, German Castignani, and Pietro Liò. Fairgame: a framework for AI agents bias recognition using game theory. In *Front. Art. Int. Appl.*, volume 413: ECAI2025, pages 4097–4104. IOS Press, 2025.

[Cabrera *et al.*, 2023] Johana Cabrera, M Soledad Loyola, Irene Magaña, and Rodrigo Rojas. Ethical dilemmas, mental health, artificial intelligence, and LLM-based chatbots. In *Int. Work-Conference Bioinf. Biomed. Eng.*, pages 313–326. Springer, 2023.

[Chaffer, 2025] Tomer Jordi Chaffer. Can we govern the agent-to-agent economy? *arXiv:2501.16606*, 2025.

[Chen *et al.*, 2026] Xi Chen, David Simchi-Levi, and Yining Wang. Utility fairness in contextual dynamic pricing with demand learning. *Management Science*, 72(3):2619–2633, 2026.

[Dafoe *et al.*, 2021] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. Cooperative ai: machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021.

[El *et al.*, 2025] Batu El, Deepro Choudhury, Pietro Liò, and Chaitanya K. Joshi. Towards mechanistic interpretability of graph transformers via attention graphs. *arXiv:2502.12352*, 2025.

[Falcão Filho, 2024] Horacio Arruda Falcão Filho. Making sense of negotiation and AI: The blossoming of a new collaboration. *Int. J. Commerce Contract.*, 8(1-2):44–64, 2024.

[Ferrara, 2024] Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, 2024.

[Fontana *et al.*, 2025] Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. Nicer than humans: How do large language models behave in the prisoner's dilemma? In *Proc. Int. AAAI Conf. Web Soc. Media*, volume 19, pages 522–535, 2025.

[Fulgu and Capraro, 2024] Raluca Alexandra Fulgu and Valerio Capraro. Surprising gender biases in gpt. *Comp. Human Beha. Rep.*, 16:100533, 2024.

[Gichoya *et al.*, 2023] Judy Wawira Gichoya, Kaesha Thomas, Leo Anthony Celi, Nabile Safdar, Imon Banerjee, John D Banja, et al. AI pitfalls and what not to do: mitigating bias in ai. *Brit. J. Radiology*, 96(1150):20230023, 2023.

[Hammond *et al.*, 2025] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, et al. Multi-agent risks from advanced AI. *arXiv:2502.14143*, 2025.

[Han *et al.*, 2021] The-Anh Han, Cedric Perret, and Simon T Powers. When to (or not to) trust intelligent machines: Insights from an evolutionary game theory analysis of trust in repeated games. *Cognitive Sys. Res.*, 68:111–124, 2021.

[Han, 2022] The-Anh Han. Emergent behaviours in multi-agent systems with evolutionary game theory. *AI Commun.*, 35(4):327–337, 2022.

[He and Zhang, 2024] Zihong He and Changwang Zhang. Afspp: Agent framework for shaping preference and personality with large language models. *arXiv:2401.02870*, 2024.

[He *et al.*, 2025] Long He, Geng Sun, Dusit Niyato, Hongyang Du, Fang Mei, et al. Generative AI for game theory-based mobile networking. *IEEE Wireless Commun.*, 32(1):122–130, 2025.

[Huynh *et al.*, 2025] Trung-Kiet Huynh, Duy-Minh Dao-Sy, Thanh-Bang Cao, Phong-Hao Le, Hong-Dan Nguyen, Phu-Quy Nguyen-Lam, Minh-Luan Nguyen-Vo, Hong-Phat Pham, Phu-Hoa Pham, Thien-Kim Than, et al. Understanding LLM agent behaviours via game theory: Strategy recognition, biases and multi-agent dynamics. *arXiv preprint arXiv:2512.07462*, 2025.

[Klinkert *et al.*, 2024] Lawrence J Klinkert, Stephanie Buongiorno, and Corey Clark. Driving generative agents with their personality. *arXiv:2402.14879*, 2024.

[Li *et al.*, 2023a] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proc. 2023 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pages 6449–6464. Association for Computational Linguistics, 2023.

[Li *et al.*, 2023b] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proc. 2023 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pages 292–305. Association for Computational Linguistics, 2023.

[Liu *et al.*, 2023] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, et al. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv:2305.13860*, 2023.

[Lu *et al.*, 2024] Yikang Lu, Alberto Aleta, Chunpeng Du, Lei Shi, and Yamir Moreno. LLMs and generative agent-based models for complex systems research. *Phys. Life Rev.*, 51:283–293, 2024.

[Mao *et al.*, 2025] Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Qiang Guan, Tao Ge, and Furu Wei. Alympics: LLM agents meet game theory. In *Proc. 31st Int. Conf. Computational Linguistics (COLING)*, pages 2845–2866. Association for Computational Linguistics, 2025.

[Montero-Porras *et al.*, 2022] Eladio Montero-Porras, Jelena Grujić, Elias Fernández Domingos, and Tom Lenaerts. Inferring strategies from observations in long iterated prisoner's dilemma experiments. *Sci. Rep.*, 12(1):7589, 2022.

[Newsham and Prince, 2025] Lewis Newsham and Daniel Prince. Personality-driven decision-making in LLM-based autonomous agents. *arXiv:2504.00727*, 2025.

[Newsham *et al.*, 2025] Lewis Newsham, Ryan Hyland, and Daniel Prince. Inducing personality in LLM-based honeypot agents: Measuring the effect on human-like agenda generation. *arXiv:2503.19752*, 2025.

[Owen, 2013] Guillermo Owen. *Game theory*. Emerald Group Publishing, 2013.

[Park *et al.*, 2023] Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proc. 36th ACM Symp. User Int. Softw. Tech.*, pages 1–22, 2023.

[Powers *et al.*, 2023] Simon T. Powers, Olena Linnyk, et al. The Stuff We Swim in: Regulation Alone Will Not Lead to Justifiable Trust in AI. *IEEE Tech. Soc. Mag.*, 42(4):95–106, 2023.

[Ramachandran *et al.*, 2022] Divya Ramachandran, Anupam Keshari, and Manoj Kumar Tiwari. Contract price negotiation using an ai-based chatbot. In *Int. Conf. Data An. Pub. Proc. Supply Chain*, pages 303–310. Springer, 2022.

[Ray, 2007] Debraj Ray. *A game-theoretic perspective on coalition formation*. Oxford University Press, 2007.

[Song and Han, 2025] Zhao Song and The-Anh Han. On evolution of non-binding commitments. *Physics of Life Reviews*, 52:245–247, 2025.

[Stewart *et al.*, 2024] Alexander J. Stewart, Antonio A. Arechar, David G. Rand, and Joshua B. Plotkin. The distorting effects of producer strategies: Why engagement does not reveal consumer preferences for misinformation. *Proc. Natl. Acad. Sci.*, 121(10):e2315195121, 2024.

[Strachan *et al.*, 2024] James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. Testing theory of mind in large language models and humans. *Nature human behaviour*, 8(7):1285–1295, 2024.

[Talajić *et al.*, 2024] Mirko Talajić, Ilko Vrankić, and Mirjana Pejić-Bach. Strategic management of workforce diversity: An evolutionary game theory approach as a foundation for ai-driven systems. *Information*, 15(6):366, 2024.

[Tessler *et al.*, 2024] Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, et al. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852, 2024.

[Van Segbroeck *et al.*, 2012] Sven Van Segbroeck, Jorge M. Pacheco, Tom Lenaerts, and Francisco C. Santos. Emergence of fairness in repeated group interactions. *Phys. Rev. Lett.*, 108(15):158104, 2012.

[Wang *et al.*, 2015] Zhen Wang, Satoshi Kokubo, Marko Jusup, and Jun Tanimoto. Universal scaling for the dilemma strength in evolutionary games. *Phys. Life Rev.*, 14:1–30, 2015.

[Wang *et al.*, 2024] Zhen Wang, Ruiqi Song, Chen Shen, Shiya Yin, Zhao Song, Balaraju Battu, Lei Shi, Danyang Jia, Talal Rahwan, and Shuyue Hu. Overcoming the machine penalty with imperfectly fair AI agents. *arXiv:2410.03724*, 2024.

[Willis *et al.*, 2025] Richard Willis, Yali Du, Joel Z. Leibo, and Michael Luck. Will systems of LLM agents lead to cooperation: An investigation into a social dilemma. In *Proc. 24th Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS)*, pages 2786–2788. IFAAMAS, 2025.