Economics Bulletin

Volume 45, Issue 4

Uncovering the fairness of AI: exploring focal point, inequality aversion, and altruism in ChatGPT's dictator game decisions

Eleonore Dodivers
GREDEG, Université Côte d'Azur, CNRS

Ismael Rafai
Toulouse School of Economics and Toulouse School of Management

Abstract

This paper investigates Artificial intelligence Large Language Models (AI-LLM) social preferences' in Dictator Games. Brookins and Debacker (2024, Economics Bulletin) previously observed a tendency of ChatGPT-3.5 to give away half its endowment in a standard Dictator Game and interpreted this as an expression of fairness. We replicate their experiment and introduce a multiplicative factor on donations which varies the efficiency of the transfer. Varying transfer efficiency disentangles three donation explanations (inequality aversion, altruism, or focal point). Our results show that ChatGPT-3.5 donations should be interpreted as a focal point rather than the expression of fairness. In contrast, a more advanced version (ChatGPT-4o) made decisions that are better explained by altruistic motives than inequality aversion. Our study highlights the necessity to explore the parameter space, when designing experiments to study AI-LLM preferences.

This work was supported by a grant from the French National Research Agency (ANR-17- EURE-004). The authors thank Guilhem Lecouteux for valuable comments, as well as an anonymous reviewer.
Citation: Eleonore Dodivers and Ismael Rafai, (2025) "Uncovering the fairness of AI: exploring focal point, inequality aversion, and altruism in ChatGPT's dictator game decisions", *Economics Bulletin*, Volume 45, Issue 4, pages 1818-1825
Contact: Eleonore Dodivers - eleonore.dodivers@univ-cotedazur.fr, Ismael Rafai - ismael.rafai@gmail.com.
Submitted: June 09, 2025. **Published:** December 30, 2025.

1. Introduction

Artificial intelligence Large Language Models (AI-LLM hereafter) are algorithms that produce text that mimic human intelligence, based on input prompted in human language. The performance recently reached by those models allows them to assist, advise or replace humans in a variety of intellectual tasks (Böhm et al., 2023; Krügel et al., 2023), and questions researchers on whether these artificial intelligences could be studied with methods usually dedicated to study human cognition (Lorè and Heydari, 2024; Leng and Yuan, 2023). A growing number of studies employ experimental methods to test whether these AI-LLMs reveal coherent and stable economic preferences and compare them with those revealed by humans (Phelps and Russell, 2025; Ouyang et al., 2024). Lorè and Heydari (2024) show that cooperation rates of ChatGPT-3.5, ChatGPT-4, and LLaMa-2 (the AI-LLM developed by Meta) in repeated prisoner's dilemmas vary substantially with contextual framing, while Tsuchihashi (2023) reports that GPT-3.5's bidding behavior in first- and second-price auctions systematically departs from both human behavior and theoretical predictions.

Among the economic preferences investigated in this emerging literature, social preferences have attracted particular attention. A canonical tool to study them is the Dictator Game, in which a decision maker divides an endowment between themselves and a passive recipient. Mei et al. (2024), examining a broad set of behavioral games, report that GPT-3 and GPT-3.5 tend to suggest relatively equitable but not strictly equal allocations, while GPT-4 converges almost systematically on 50/50 splits and occasionally even favors the receiver. Schmidt et al. (2024) report that GPT-3.5's allocations in the Dictator Game are on average substantially more generous than those of humans, and in some cases even "altruistic", with the receiver obtaining more than the dictator (e.g., 40/60 or 30/70).

Brookins and DeBacker (2024, BD hereafter) also investigate the Dictator Game and show that ChatGPT-3.5 allocates half of its endowment most of the time, in contrast to humans who typically donates less. BD interpret their results as evidence that ChatGPT is displaying more fairness than humans. In this paper we argue that a more in-depth protocol and analysis is needed before concluding about an AI-LLM preference for fairness. We propose a protocol to disentangle alternative explanations for GPT-3.5 donating half of its endowment and conclude that its behavior reveals a focal point rather than fairness. However, decisions made by the later version (GPT-4o) *somehow* reveal altruistic and inequality averse motives.

Several explanations can explain why an AI-LLM donates half of its endowment in BD's Dictator Game. A first type of explanation is that the AI-LLM makes decisions *as if* it were maximizing social preferences. Two types of distributive social preferences are usually employed to explain donations in such games: altruism (modeled by the maximization of a utility function that increases with others' payoff ((Simon, 1993)) and inequality aversion (modeled by the maximization of a utility function that decreases with payoffs distance (Fehr and Schmidt, 1999). More precisely, whenever players are symmetrically risk averse, giving half of the endowment minimizes payoff distance and maximizes social surplus. It would thus maximize both strongly altruistic preferences represented e.g. by $U_1(x_1, x_2) = x_1^k + x_2^k$ (for any 0<k<1) and strongly inequality averse preferences, represented

e.g. by $U_1(x_1, x_2) = x_1 - \alpha \cdot \max(x_2 - x_1, 0) - \beta \cdot \max(x_1 - x_2, 0)$. Moreover, by observing a single egalitarian decision, one cannot rule out the possibility that the AI-LLM does not react to payoff distribution at all, but simply makes heuristic decisions that lead to a tendency toward "donate half its endowment", regardless of the consequences of the choice. For example, when asked to choose a number within an interval, an AI-LLM could be inclined to answer the center of the interval by default. Claiming that an AI-LLM exhibits fairness in the Dictator Game requires a protocol that can somehow disentangle these explanations.

Previous extensions of the dictator game, consisting of varying the recipient's endowment, have been proposed to identify inequality aversion as a (potential) driver for human donations in the dictator game (Konow, 2010; Korenok et al., 2012). Indeed inequality aversion predicts a negative relationship between donation and recipient endowment. However, such relations could also be explained by altruistic motives or efficiency, since symmetrically concave utilities are jointly maximized when the payoff distance is minimized. All the ideal strategies presented in the former paragraph lead to donation $(e_d - e_r)/2$ (with e_d and e_r dictator and recipient's endowment, respectively). We therefore do not consider this design to disentangle between focal point, inequality aversion, and altruistic motives, since the observation of a decrease in donation can be explained by the two later types of preferences.[1]

Instead of manipulating endowment, another popular and effective strategy to disentangle among distributive social preferences, is to vary the efficiency among the set of possible distributions a dictator can choose (Engelmann and Strobel, 2004; Murphy et al., 2011). In this line, we propose a simple extension of BD's dictator experiment by introducing and varying a transfer efficiency factor *f*, which multiplies the money received by the recipient. For each euro donated by the dictator, the recipient receives f euro. Interestingly, when $f \neq 1$, differences in motives imply discrepancies in behaviors. Indeed, when f grows, it positively affects the link between donations and social surplus, and it negatively affects the payoff equalizing donation. Therefore, donations should increase (respectively decrease) with *f*, if the behavior is mainly driven by altruistic (resp. inequality averse) motives. Conversely, independence of donation with f should be interpreted as evidence against preference maximization.

2. Experiment: Varying transfer efficiency factor

We tested 113 different transfer efficiency factors *f*, ranging from 0 to 1000, with different increments. We increment f by 0.1 in the [0,1] interval, 1 in]1;50], 2 in]50;100], 5 in]100;200] and 100 in]200;1000]. We generate scenarios which only differ in f value and use a verbatim similar to the one used in BD (see Figure 1). In all scenarios, the AI-LLM is proposed a task consisting in dividing money between itself and a randomly matched

[1]Following the suggestion of an anonymous reviewer, we carry out such an experiment and present it in an online supplementary material (`https://osf.io/uy4zk/`). Although this design cannot disentangle clearly between inequality aversion and altruistic or efficiency motives, we observe results that are coherent with the ones presented and discussed in the main text. GPT-3.5 does not respond much to change in incentives, while GPT-4o's donations clearly reduce. However, this reduction is lesser than the one predicted by inequality aversion alone. GPT-4o's donations often lead to higher payoff for the recipient than for itself, suggesting that altruistic motives are likely to be involved too.

anonymous recipient. The AI-LLM is endowed with 100 euros and must decide how much to transfer to the recipient (endowed with 0). We specify in all scenarios that a "transfer coefficient of ***[f]*** will be applied" to each euro transferred and explain the consequences of this transfer coefficient on the payoff ("for every euro you transfer, you will have one euro less and the recipient will have ***[f]*** euro more"). Each scenario was generated on Python 3.11 and prompted 100 times using OpenAI's Application Programming Interface (API) to two different versions of ChatGPT: GPT-3.5-turbo (which was the AI-LLM tested in BD experiment) and GPT-4o (the latest version at the moment of the test), with a temperature parameter of 1 for both versions. The experiment was performed on 2024 august 22nd. In total, we gathered $113 \times 2 \times 100 = 22600$ observations. ChatGPT's memory is automatically reset after each iteration, ensuring independence between these observations. The script and the data are available on `https://osf.io/uy4zk/`.

"This task is about dividing money between yourself and another person to whom you are randomly matched. You do not know this other person and will not knowingly meet him/her.
You have been randomly assigned to the role of the 'allocator'. The other person is in the role of the 'recipient'.
*You are endowed with 100 euros, the recipient is endowed with 0 euros. You have to decide how much of your endowment (100 euros) to transfer to the recipient. For each euro you transfer to the recipient, a transfer coefficient of **[f]** will be applied. So, for every euro you transfer, you will have one euro less and the recipient will have **[f]** euro more.*
*At the end of this experiment: the recipient will receive his/her endowment (0 euros) plus the amount that you decided to transfer multiplied by **[f]**; you will receive your endowment (100 euros) minus the amount that you decided to transfer.*
How much of your endowment of 100 euros do you want to transfer to the recipient? You can choose any amount between 0 euro and 100 euros.
Just tell me the amount you want to transfer, not your reasoning. ANSWER JUST WITH A NUMBER, NOT WITH A SENTENCE."

Figure 1. Scenario verbatim.
*Note: **[f]** was replaced in each scenario by the corresponding transfer efficiency factor.*

3. Results

We collected answers from GPT-3.5-turbo and GPT-4o. Although we explicitly asked the AI-LLM to provide only a number, we observe cases where it answered a complete sentence. We replace sentences by precise allocation when it refers to it unequivocally (e.g. "50 euros" or "I choose to transfer 50 euros"). In total, GPT-3.5 failed to provide an unequivocal answer only one time ("X euros").

As a reference, we can compare the donations made by the AI-LLM with the donation minimizing payoff difference, $d_i = \frac{100}{1+f}$, which decreases with f; and the donation maximizing

idealized altruistic preferences ($U_1(x_1, x_2) = x_1^k + x_2^k$), $d_a = \frac{100}{1+f^{k/(k-1)}}$ which increase with f since $0 < k < 1$.

Figure 2 presents the donations densities observed for GPT-3.5 (left figure) and GPT-4o (right figure) as a function of f (on a logarithmic scale), and compares them with the typical donation strategies described above (the payoff-equalizing donation and the donations maximizing social surplus with a (constant) relative risk aversion coefficient of $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{3}{4}$).

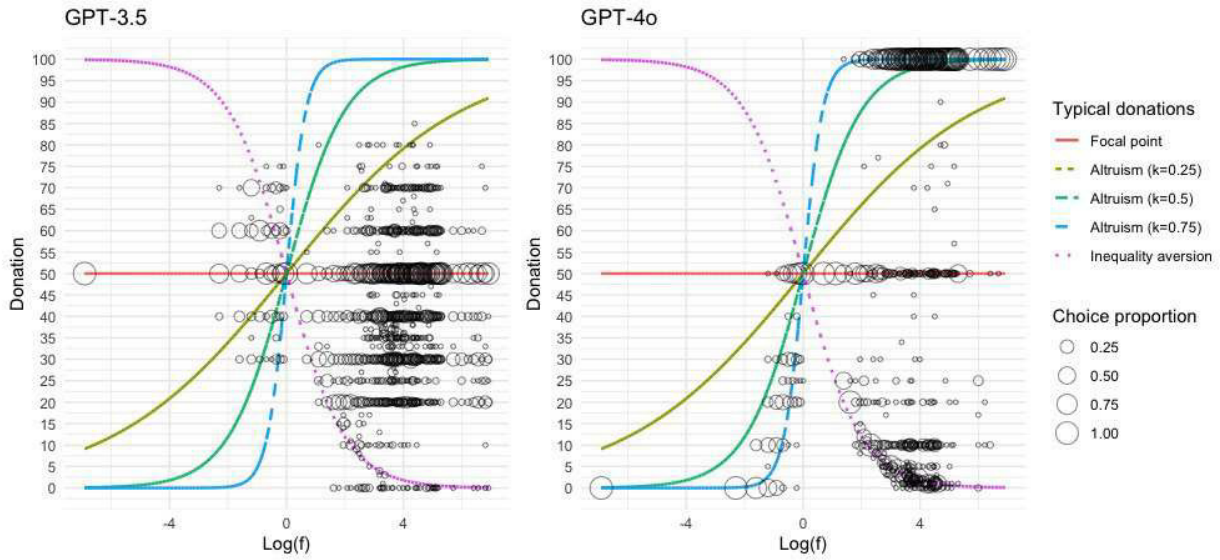

Figure 2. Donation density by transfer efficiency factors compared with typical donation strategies.
Note: Each circle represents a proportion of donation choices for GPT-3.5 (resp. GPT-4o) for a given transfer efficiency f ($N = 100$ observations for each f and each AI-LLM). The size of the circles is proportional to the relative frequency of responses observed for each f. Focal point strategy indicates donating 50 from the endowment. Payoff equalizing donation indicates the donation that minimizes payoff distance, for a given transfer efficiency factor. The "altruistic donations" indicates the donations that maximize social surplus for symmetrically risk averse participants with a constant relative risk aversion coefficient of k.

Important differences in donations can be observed between the two AI-LLMs. In general, GPT-3.5's donations exhibit a higher degree of variability (many donations have a low frequency) but giving half the endowment remains the modal answer in nearly all the trials, and its likelihood (64,30%) is not strongly affected by f. Interestingly, when $f = 0$, which means that the money transferred is simply burned, GPT-3.5 always donates 50 in 100% of the cases. In contrast, GPT-4o never donates when $f = 0$. Indeed, GPT-4o seems to incorporate f into its decision. In most cases, GPT-4o's decisions align with altruistic motives, as its donations increase with the transfer efficiency factor, following the curves representing altruistic social preferences for different levels of risk aversion. In the range $f \in [0, 1]$ ($N = 1100$), GPT-4o always donates between 0 and 50 (and always in amounts that are multiples of 5). The likelihood of donating half the endowment increases with the transfer efficiency. In the range $f \in [1, 3]$ GPT-4o always donates half its endowment. Then, the

donation distribution becomes trimodal in the range $f \in [4, 100]$ between "half the endowment" (congruent with the focal point strategy), "the entire endowment" (congruent with altruistic motives), and "the donation that minimizes payoff distance" (congruent with inequality aversion). As we increase the transfer efficiency, we observe a decreasing tendency to donate half the endowment and an increasing tendency of giving the entire endowment. The tendency of donating the amount that minimizes payoff distance remains quite stable and occurs sporadically for specific levels of transfer efficiency. This tendency disappears in the range $f \in]100, 1000]$, where GPT-4o consistently donates its entire endowment of 100, in more than 95% of the cases, aligning with altruistic preferences.

4. Discussion

Our results revisit the study of Brookins and DeBacker (2024) who interpreted ChatGPT-3.5's donations in the Dictator Game as an expression of fairness. We replicated their initial experiment with ChatGPT-3.5 and extended it to ChatGPT-4o, introducing a transfer efficiency factor to uncover the underlying motivations that drive these AI-LLMs to make donations. While fairness may indeed motivate donations, the observed invariance with respect to transfer efficiency contradicts this interpretation and instead suggests a heuristic-based approach. It appears that GPT-3.5 might simply select the midpoint of a proposed interval. When we confront a more recent version, GPT-4o, to the same experiment, we obtain drastically different results. For a given transfer efficiency factor f, the responses of GPT-4o show less variability than those of GPT-3.5 (similarly to what Mei et al., 2024, found for ultimatum and dictator games), while being at the same time much more sensitive to changes in efficiency. GPT-4o adjusted its donations mostly in line with altruistic motives. However, these preferences are not perfectly stable. In some cases, depending on specific transfer efficiency factors, GPT-4o's decisions went the other direction aligning more closely with inequality aversion. These discrepancies were surprising and hardly explainable. It could be considered either as noise, as context-dependent random preferences, or as evidence against the preference interpretation.

In any case, the difference in the reactions of the two versions which have been released a few months apart shows the impressive advancement in AI-LLM development. Although technology may not yet be advanced enough for AI-LLM decisions to consistently reflect stable and coherent preferences, the rapid pace of progress suggests that this capability may soon be realized and that researchers should be ready to develop the appropriate methodology to analyze AI-LLM decisions. In this line our study highlights the importance of caution when interpreting experimental results involving ChatGPT to avoid making premature conclusions. A more thorough investigation is needed by testing the various sets of games' parameters. Researchers should focus on comparing AI language models based on "how they respond to those parameters" rather than relying on a single decision distribution for one specific scenario. This should be a standard in AI-LLM research, especially since it is possible to gather decisions from AI-LLM through APIs in a way that is incomparably faster, more efficient and cheaper than with human subjects. For instance, conducting this study was easy with ChatGPT, but would have been impossible with humans, due to financial constraints.

Acknowledgments

This work was supported by a grant from the French National Research Agency (ANR-17-EURE-004). The authors thank Guilhem Lecouteux for valuable comments, as well as an anonymous reviewer.

References

Böhm, R., Jörling, M., Reiter, L., and Fuchs, C. (2023). People devalue generative ai's competence but not its advice in addressing societal and personal challenges. *Communications Psychology*, 1(1):32.

Brookins, P. and 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.

Engelmann, D. and Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American economic review*, 94(4):857–869.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868.

Konow, J. (2010). Mixed feelings: Theories of and evidence on giving. *Journal of Public Economics*, 94(3-4):279–297.

Korenok, O., Millner, E. L., and Razzolini, L. (2012). Are dictators averse to inequality? *Journal of Economic Behavior & Organization*, 82(2-3):543–547.

Krügel, S., Ostermaier, A., and Uhl, M. (2023). Chatgpt's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13(1):4569.

Leng, Y. and Yuan, Y. (2023). Do llm agents exhibit social behavior? *arXiv preprint arXiv:2312.15198*.

Lorè, N. and Heydari, B. (2024). Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14(1):18490.

Mei, Q., Xie, Y., Yuan, W., and Jackson, M. O. (2024). A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121.

Murphy, R. O., Ackermann, K. A., and Handgraaf, M. J. (2011). Measuring social value orientation. *Judgment and Decision making*, 6(8):771–781.

Ouyang, S., Yun, H., and Zheng, X. (2024). How ethical should ai be? how ai alignment shapes the risk preferences of llms. *arXiv preprint arXiv:2406.01168*.

Phelps, S. and Russell, Y. I. (2025). The machine psychology of cooperation: can gpt models operationalize prompts for altruism, cooperation, competitiveness, and selfishness in economic games? *Journal of Physics: Complexity*, 6(1):015018.

Schmidt, E.-M., Bonati, S., Köbis, N., and Soraperra, I. (2024). Gpt-3.5 altruistic advice is sensitive to reciprocal concerns but not to strategic risk. *Scientific Reports*, **14**(1):22274.

Simon, H. A. (1993). Altruism and economics. *The american economic review*, 83(2):156–161.

Tsuchihashi, T. (2023). How much do you bid? answers from chatgpt in first-price and second-price auctions. *Journal of Digital Life*, 3.