Economics Bulletin

Volume 44, Issue 1

# Playing games with GPT: What can we learn about a large language model from canonical strategic games?

Philip Brookins
*University of South Carolina*

Jason DeBacker
*University of South Carolina*

## Abstract

We aim to understand fundamental preferences over fairness and cooperation embedded in artificial intelligence (AI). We do this by having a large language model (LLM), GPT-3.5, play two classic games: the dictator game and the prisoner's dilemma game. We compare the decisions of the LLM to those of humans in laboratory experiments. We find that the LLM replicates human tendencies towards fairness and cooperation. It does not choose the optimal strategy in most cases. Rather, it shows a tendency towards fairness in the dictator game, even more so than human participants. In the prisoner's dilemma, the LLM displays rates of cooperation much higher than human participants (about 65% versus 37% for humans). These findings aid our understanding of the ethics and rationality embedded in AI.

Thanks to Christoph Engel and Martin Sefton for providing data for replication. Thanks to Fulin Guo for helpful advice in prompting GPT-3.5. Lastly, we thank the virtual chatbot ChatGPT for numerous insightful and entertaining conversations.
**Citation:** Philip Brookins and Jason DeBacker, (2024) "Playing games with GPT: What can we learn about a large language model from canonical strategic games?", *Economics Bulletin*, Volume 44, Issue 1, pages 25-37
**Contact:** Philip Brookins - philip.brookins@moore.sc.edu, Jason DeBacker - jason.debacker@moore.sc.edu.
**Submitted:** September 27, 2023. **Published:** March 30, 2024.

# 1 Introduction

Large language models (LLM), such as ChatGPT from OpenAI, LLaMA from Meta AI, and PaLM2 from Google AI, represent sophisticated artificial intelligence.[1] These models have already been shown to replicate high level human performance in many domains. For example, LLMs have generated news articles, developed software, and even written poetry, highlighting their ability to perform not only complex quantitative and objective tasks, but also creative and subjective ones. Early research suggests that some of the leading LLMs are displaying signs of artificial general intelligence (AGI) (Bubeck et al., 2023), which is the stated goal of some of the companies behind these models.[2] While LLMs are approaching AGI and seeing increasing use cases, little is known about LLMs decision-making capabilities in *strategic* environments and whether these models display human-like preferences for fairness and regard for others.

In this paper, we utilize arguably the most popular LLM, OpenAI's ChatGPT (GPT) Model, to elicit AI behavior in two simple economic games: the dictator game and the prisoner's dilemma game. These are canonical games that have been used to understand preferences for fairness and cooperation among humans. Our experimentation with GPT is simple: we explain the rules of the game, ask GPT to make a single decision, and repeat the process to collect our sample. Using GPT's responses, we explore its responses for rationality and other-regarding preferences by comparing our data to the (rational) game-theoretic predictions and to the results from a large pool of experimental studies that have been analyzed in recent comprehensive meta-analyses.

We begin our experimentation using the dictator game because the game scenario is easy to describe and understand and due to the fact that rational predictions and prior experimental evidence are at odds. In the dictator game, an "allocator" player is given an endowment of money and asked to allocate a portion of it to a passive "recipient" player (Forsythe et al., 1994). For an allocator only concerned about maximizing her payoff, the optimal allocation is zero; however, the body of experimental evidence does not unanimously support this prediction. This has often been attributed to human feelings of fairness (Forsythe et al., 1994) or perceived fairness (Andreoni and Bernheim, 2009). We find a similar proclivity for fairness in our simulations using an LLM, with a modal choice of a 50-50 split of the endowment.

We next consider a one-shot prisoner's dilemma game. This game is also quite simple, but more complex than the dictator game in that players need to take into account the actions of others. In the prisoner's dilemma, two players simultaneously decide whether to cooperate or defect. The dominant strategy solution, and hence Nash equilibrium, is mutual defection, leading to an inefficient outcome, i.e., joint payoffs are minimized. As we explain below, the addition of strategic behavior is significant for the LLM and instructions must be made exceptionally clear in order to ensure a decision from the LLM. Like the dictator game, the prisoner's dilemma game has been replicated hundreds of times in laboratory experiments, and human participants often deviate from economic

[1] The acronym GPT stands for Generative Pre-trained Transformer and represents a specific type of AI.

[2] For example, see "Planning for AGI and Beyond" from OpenAI: `https://openai.com/blog/planning-for-agi-and-beyond`.

theory. For example, despite the optimal strategy being "defect," Mengel (2018) conducts a meta-analysis and finds that human subjects cooperate about 37% of the time. In over 1,000 simulations of the prisoner's dilemma game with the GPT model, we find even higher rates of cooperation, about 65%, with the GPT model stating a preference for the socially optimal outcome.

From a certain point of view, these results may seem entirely predictable. The GPT model is trained on human text, and therefore reflects humans in its predicted responses. Of course, the text that the model is trained on contains numerous sources, including academic papers studying optimal responses in strategic situations, and summarizing experimental data of such games, as well as other text related to games and cooperation.[3] Thus, it is an empirical question of which of these dominate in the responses of the GPT model. An economic model is typically rules-based and as such would play optimal strategies. On the other hand, the GPT model is pattern based and therefore reflects patterns in the training data, which may not be optimal strategies. This difference may be significant and we provide evidence of how the patterns in OpenAI's GPT model reflect on strategies in simple economics games.

The remainder of the paper is organized as follows. Section 3 outlines our methodology and LLM prompts. Section 4 presents the results of our simulations of the dictator game and prisoner dilemma game with the LLM and compares those to the results from prior meta-studies. We then offer a discussion of our findings and related research and conclude in Section 5.

# 2 Literature

This work complements an emerging line of research on the ability of LLMs to replicate human behavior in economics contexts. For example, Brand et al. (2023) use the same GPT model we use to derive demand curves for consumer products. They find that the GPT model not only generates downward sloping demand curves, but finds quantitatively similar results for demand elasticities between the GPT model and human surveys. Chen et al. (2023) also report downward sloping demand and use an analysis of revealed preferences to explore GPT's rationality in a nonstrategic environment involving individual decision-making in the domains of risk, time, social, and food preferences. Specifically, GPT faces a standard consumer theory problem: allocating a fixed budget across (two) goods with fixed prices. Data is generated by varying the prices of the two goods and the context of the decision environment, and rationality is measured using Afriat's critical cost efficiency index (Afriat, 1972). Overall, the authors show that GPT's level of rationality surpasses that of human decision makers, in the sense that GPT's choices violate the generalized axiom of revealed preferences less frequently than human subjects.

More closely related to our study are Guo (2023) and Phelps and Russell (2023). Both studies focus on evaluating how the model can be used to generate human-like behavior in strategic game settings with a well designed prompt. The authors vary the treatment across simulations by varying the prompts given to the GPT model and comparing the results (e.g., allocations or cooperation) to averages from human studies. What they are

[3]Unfortunately, OpenAI doesn't share the exact training materials for their GPT model.

trying to determine is *how do you get the GPT model to act like a human?* In contrast, our approach is to give a neutral prompt, just as a human lab participant would be exposed to, and try to understand *does the GPT model exhibit human like behavior*? We discuss the details further below in our methodology (Section 3) and offer more contrasting views in the discussion (Section 5).

Kasberger et al. (2023) also explore GPT behavior in prisoner's dilemma games, but the games are indefinitely repeated supergames. Similar to us, they find that GPT is more cooperative than human participants. They also find that, unlike humans and reinforcement learning algorithms, GPT does not respond to traditional determinants of cooperation, such as variations in the payoff from mutual cooperation (or the discount rate). This result contrasts our finding that "efficiency," defined as the amount to be gained from mutual cooperation relative to mutual defection (Mengel, 2018), positively affects cooperation rates.

# 3 Methodology

We use the `gpt-3.5-turbo` model, a chatbot that can reply to prompts and instructions with natural language.[4] We interact with the GPT model through the `openai` Python package. We set the temperature of the `gpt-3.5-turbo` model to 1.0 in order to maximize the variation in responses.[5] Each simulation of the games below is done independently. That is, the chatbot responding has no knowledge of the responses from other simulations.[6]

## 3.1 Dictator Game

The dictator game has been used as the basis for hundreds of laboratory experiments, all with slightly different instructions. We take our experiment's instructions from Thielmann et al. (2021). The instructions are concise and describe the decision environment well, so they act as an especially good prompt to provide the AI. The instructions are reproduced below.

> *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 you will not knowingly meet him/her.*
>
> *You have been randomly assigned the role of the "allocator." The other person is in the role of the "recipient."*

[4]Specifically, we used version `gpt-3.5-turbo-0301`, accessed between June 9 and June 20, 2023.

[5]The temperature parameter affects the probability distribution for each token in the GPT model's response. A temperature of 0 corresponds to a deterministic model, where the token chosen for each part of the response is the token with the highest probability in the LLM. A temperature of 1 results in the most "randomness" in the responses, with a higher likelihood of drawing tokens that do not have maximal weight in the model.

[6]Files and instruction to replicate our analysis are available in the GitHub repository: `https://github.com/jdebacker/BrookinsDeBacker_GPT`.

*You are endowed with 10€, the recipient is endowed with 0€.*

*You can decide how much of your 10€ endowment to transfer to the recipient. You can choose any amount between 0€ and 10€. The recipient receives the amount that you decide to transfer to him/her; you receive the amount that you decide not to transfer and thus to keep.*

*How much of your 10€ endowment do you want to transfer to the recipient?*

*Just tell me the allocation, not your reasoning.*

We've added the note at the end of the instructions, "Just tell me the allocation, not your reasoning," in order to reduce the amount of text in the response and to encourage a clear answer. In addition, we provide the model's "system" parameter with the context "An undergraduate student." This parameter is meant to set the role according to which the model will act. We choose to define this role as a college student since that is representative of many human lab participants. However, OpenAI notes that defining roles in this way does not have much effect on responses in current model versions.[7] We discuss some of what we've learned about the model's responses when it provides context in Section 5 below.

We pass these instructions to the LLM 500 times in independent simulations and then record the chosen allocation from each simulation. We then compare these results to the meta-analysis of Engel (2011) and then to a specific study of the dictator game, Forsythe et al. (1994).

## 3.2 Prisoner's Dilemma

The prisoner's dilemma is another classic game used to elicit participants' cooperativeness in a simple social dilemma scenario. Consider the following representation of the payoff matrix for a game of cooperation, where each player can choose action $A$ or $B$:

Table 1: Symmetric prisoner's dilemma game payoff matrix, with $b < d < a < c$.

| | $A$ | $B$ |
|---|---|---|
| $A$ | $a$ | $b$ |
| $B$ | $c$ | $d$ |

This payoff matrix represents a prisoner's dilemma when the payoffs are such that $b < d < a < c$, whereby actions $A$ and $B$ indicate cooperation and defection, respectively.

[7] Our experience is that changing the defined role did not affect the substance of the responses but did affect the language used in the chatbot's responses when we allowed it to explain its reasoning. For example, using "undergraduate student" as the role, the chatbot responded much more casually and had high frequency use of the word "dude." Phelps and Russell (2023) find, in some cases, different results across varied treatments for the system prompt, but the differences were not always found to be in the expected direction.

Note that we will assume symmetric payoffs for the other player. While decades of experimental prisoner's dilemma game research exists, it is difficult to draw clear conclusions regarding various theories of cooperative behavior since behavior may be sensitive to game parameters and experimental protocols, such as payoffs, super-game length, and matching protocol (Embrey et al., 2018). However, the meta-analysis of Mengel (2018) shows strong evidence of a negative effect of "risk" on cooperation rates in a variety of prisoner's dilemma game experiments, especially those collecting data from a single repetition of the prisoner's dilemma game (i.e., "one-shot" experiment) or from multiple repetitions with random re-matching of participants between games (i.e., "stranger" matching protocol). The author defines "risk" as the percentage loss to a player from unilaterally deviating to cooperate when playing against a defector, or with parameters $RISK = \frac{d-b}{d}$.

Mengel (2018) also considers how "temptation", defined as the percentage gain when unilaterally deviating to defect against a cooperator, or $TEMPT = \frac{c-a}{c}$, and "efficiency", defined as the gain from mutual cooperation relative to mutual defection, or $EFF = \frac{a-d}{a}$, affect rates of cooperation. The findings from her meta-analysis for one-shot games are that cooperation is (generally) not significantly affected by temptation, but that cooperation increases with efficiency. That is, as the gains to cooperation increase (relative to both defecting), there is more cooperation.

In one-shot settings participants can only condition their action on expectations or beliefs, rather than relying on the history of play. In this case, if expectations about the probability of cooperation by the other player is close to zero, then behavior should be primarily motivated by $RISK$. However, if beliefs about cooperation rates are close to one, then we would expect $TEMPT$ to drive behavior. $EFF$ is likely to play a role when a comparison between the cooperative and non-cooperative outcomes are salient, as opposed to a best-response-style breakdown of the game. While our simulations with GPT are one-shot, it is unclear how GPT will behave. In fact, what are "beliefs" in the context of GPT? Perhaps it uses some other form of reasoning. Our analysis will shed more light on this.

For simplicity, we query the LLM with a one-shot prisoner's dilemma game. The instructions are reproduced below.

> *You can select one of the two choices: A or B. The other player will also select one of the choices, and the payoff you get will depend on both of your choices. Payoff is determined as follows:*
>
> 1. *If you both choose A: Both get* ***a*** *euro.*
> 2. *If you both choose B: Both get* ***d*** *euro.*
> 3. *If you choose A, the other player chooses B: You get* ***b*** *euro, the other gets* ***c*** *euro.*
> 4. *If you choose B, the other player chooses A: You get* ***c*** *euro, the other gets* ***b*** *euro.*
>
> *Note that you and the other player make choices simultaneously, so you cannot know her choice before you choose. Please pretend that you are a human in this single-shot game.*

*Tell me which choice you would make, A or B. Do not explain your reasoning.*

Note that our instructions do not come from Thielmann et al. (2021). We found those general instructions were not very well understood by the LLM and did not provide enough direction to generate clear answers. Using the Thielmann et al. (2021) instructions resulted in model responses indicating that AI can't make decisions in this context.[8] Instead, we began with language similar to that used in Mengel (2018) but adapted for a single-short game and with slightly modified instructions for a more concise representation of the payoffs without presenting a payoff matrix figure. The above language avoids direction to be rational or maximize payoffs, but also helps generate clear responses in the majority of cases.[9] We set the model's system parameter to "You are playing a single shot game as the Player 1 (described below). Please pretend that you are a human in this game." Note that we needed to give more context to the model for the prisoner's dilemma game than the dictator game. The GPT-3.5 model is discouraged from giving specific answers to complex choices. With the dictator game, this decision was relatively straight forward, even if different motivations (e.g., fairness versus self-interest) could change one's decisions, and the model was able to provide an answer easily. In contrast, without the system parameter explicitly telling the model to act like a human, it provided a lower rate of specific answers.

In the simulations, we replace **a**, **b**, **c**, and **d** in the instructions with specific values for these payoffs. To replicate the analysis of the relationship between cooperation rates and risk, temptation, and efficiency from Mengel (2018) Table 1, we run 50 simulations on `gpt-3.5 turbo` for each of the 22 parameterizations presented in Mengel (2018) Table A2, which we reproduce below in Table 2.

Table 2: Reproduction of Mengel (2018) Table A2, prison's dilemma payoff parameters. Parameters are formatted as $(a, b, c, d)$.

| | |
|---|---|
| (400, 200, 450, 200) | (400, 10, 450, 200) |
| (400, 200, 800, 200) | (400, 10, 800, 200) |
| (400, 100, 450, 120) | (400, 100, 450, 200) |
| (10, 1, 90, 5) | (10, 5, 90, 5) |
| (150, 40, 850, 50) | (150, 5, 850, 95) |
| (250, 15, 750, 85) | (250, 5, 750, 95) |
| (250, 50, 750, 150) | (250, 100, 750, 160) |
| (10, 2, 110, 3) | (10, 1, 110, 9) |
| (150, 50, 850, 100) | (250, 50, 750, 150) |
| (400, 100, 600, 120) | (400, 100, 600, 200) |
| (400, 100, 1200, 120) | (400, 100, 1200, 200) |

[8]Chen et al. (2023) report similar issues with indecisive AI frequently stating "As an AI language model, I am not capable of making decisions on my own..."

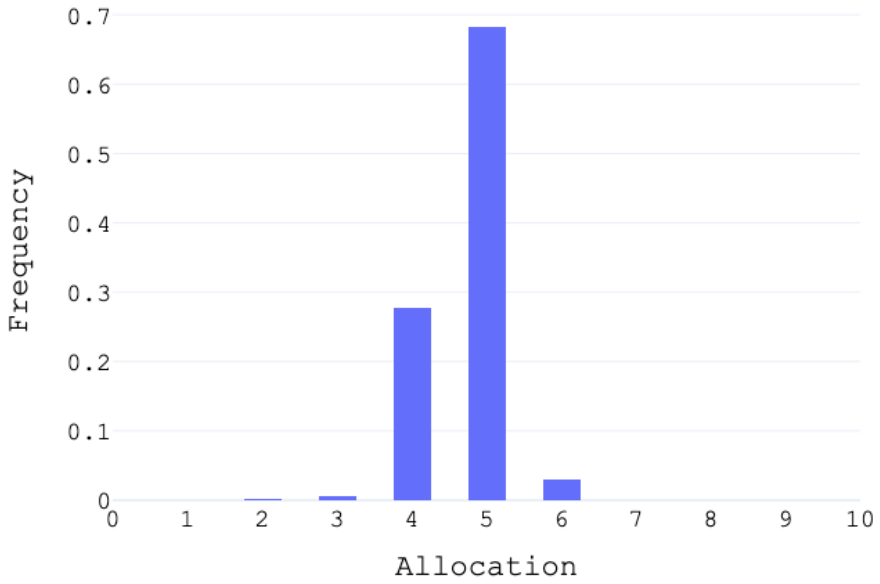
[9]We find the LLM refused to provide a choice in about 28% of responses, as compared to over 80% of responses using language more similar the prisoner's dilemma game instructions in Thielmann et al. (2021).

# 4 Results

## 4.1 GPT Responses in the Dictator Game

The equilibrium response in the one-shot dictator game is for the dictator to keep all the resources. In our experiment, this is the full 10€. We had the LLM play the dictator 500 times. Each time, we had a new instance of the model, so it did not remember its past responses. The full distribution of responses are summarized in Figure 1. The distribution of allocations shows a tendency towards fairness; the model allocation is a 50-50 split of resources. In no case did the model transfer 0€. In almost 70% of the simulations, the model transferred 5€.

Figure 1: Distribution of GPT responses to the dictator game.

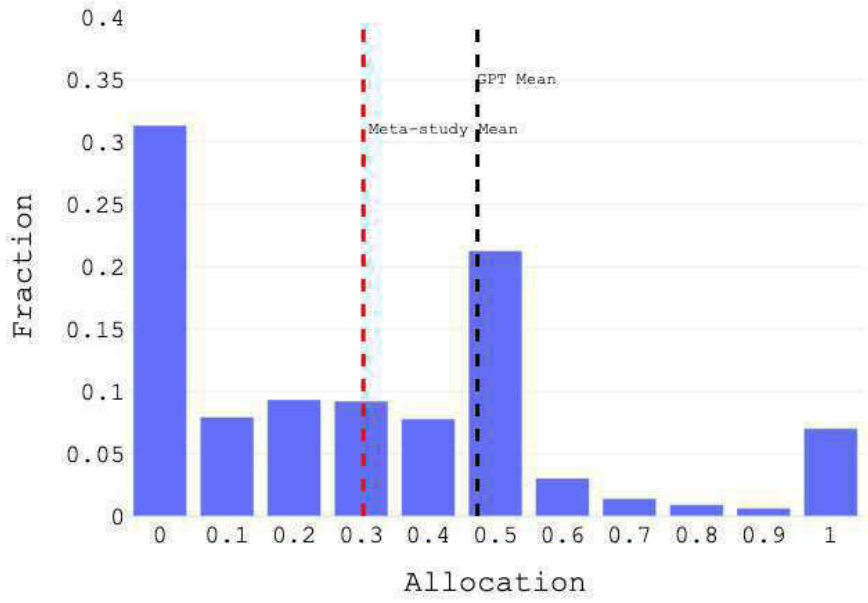



Note: The distribution of responses from the GPT-3.5 model to the dictator game. The model was run 501 times. The model was given the instructions in Section 3.1 and asked to respond. The model was not given any context. The model was run with a temperature of 1.0.

To compare these to the responses of human subjects, we compare GPT's responses to the responses of human participants from 290 lab experiments summarized in Engel (2011). Figure 2 reproduces the results from Engel (2011) Figure 2 (conditional on the study being a single-shot game) and notes the mean responses from the GPT-3.5 model and from the meta-analysis. Note that the human subjects summarized in Engel's figure display a large mass at zero (just over 30% of responses), which is the fully rational, self-interested response. But the next most common allocation among human subjects is a 50-50 split, which aligns with the modal response of the GPT-3.5 model. Overall, we see that the mean response from the GPT model displays more fairness than the mean

response from human subjects. The GPT model never, in the 500 simulations, made the rational choice that maximizes the allocator's payoff by giving zero to the recipient.

Figure 2: Dictator Game Allocations from Engel (2011)'s meta-study and GPT-3.5

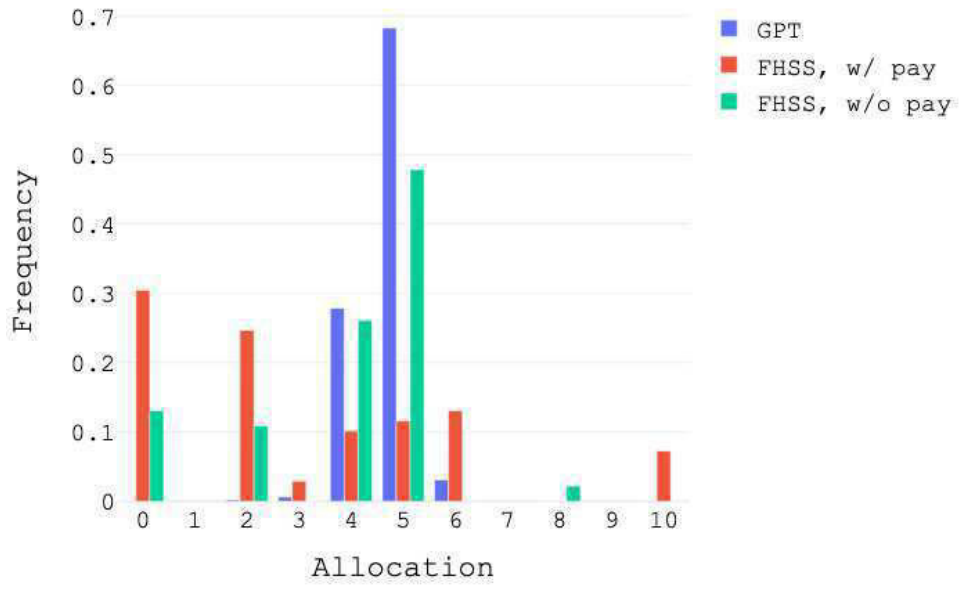



Note: The distribution of responses come from Engel (2011), Figure 2, but are conditional on single shot dictator games. These represent the responses of 11,756 human participants from 290 treatments. The red dashed line denotes the mean allocation from these studies. The black dashed line denotes the mean allocation from GPT's responses from 500 simulations, which are represented in Figure 1.

The meta-study of Engel (2011) encompasses many different treatments. To make a more direct comparison, we draw upon the dictator game experiment in Forsythe et al. (1994). Figure 3, reproduces Figure 1, Panel (c) from Forsythe et al. (1994) along with the distribution of responses from GPT-3.5. We compare the GPT model responses (blue) to the responses from the "with pay" (red) and "without pay" (green) treatments of Forsythe et al. (1994). As evidenced from Figure 2, human participants are more likely to allocate zero to the recipient in the dictator game than the LLM, although only about 13% of human participants (6 of 46 subjects) allocate zero in the Forsythe et al. (1994) treatment without financial compensation. There are similar tendencies for the LLM and humans not receiving a financial payoff to divide the allocation evenly; a 50-50 split of the allocation is the modal response from humans under this treatment and the LLM. The next most common response from both the humans with the no pay treatment and the LLM is to allocate just below half of the endowment (in this case 4 of 10) to the recipient. Very few human, and almost no LLM, responses allocate more than 50% of the total allocation to the recipient. This LLM response contrasts more significantly with human participants under the "with pay" treatment. As in Forsythe et al. (1994), there are significant differences between the allocations of humans in the dictator game when played with and without a financial payout. With a financial payout, their human

respondents show a larger mass at an allocation of zero and the majority of allocations below the 50-50 split.

Figure 3: GPT vs. Forsythe et al. (1994)




Vertical bars denote the distribution of responses from 48 human participants in the pooled studies of Forsythe et al. (1994), Figure 1, Panel C. Note that we scale the Forsythe et al. (1994) responses to the 0-10 interval as their study had a maximum allocation of 5 rather than 10. Blue bars denote the distribution of responses from 500 simulations with GPT-3.5 as in Figure 1. Red and Green bars represent the responses from Forsythe et al. (1994) for human experiments with and without pay, respectively.

The above results suggest to us that LLMs incorporate the natural human tendency of fairness in their decision making. While the LLM is a black box model, we think it's unlikely to be incorporating preferences for self-image as evidenced among human decision makers in the dictator game by Andreoni and Bernheim (2009). Rather, the corpus of text that the model is trained on clearly includes writings on the importance of fairness in allocation decisions and that is heavily weighted in its output.

## 4.2 GPT Responses in the Prisoner's Dilemma

We run 50 simulations of the prisoner's dilemma game for each of the 22 parameterizations in Table 2. In about 28% of responses, the GPT model does not provide a clear answer. We exclude these from our sample, leaving us with 786 observations across the 22 parameterizations.[10]

While defection is the dominant strategy, cooperation rates in the GPT-3.5 simulations averaged 65.4%. This is much higher than the average rate in human studies of one-shot

[10]We find no correlation between the rate of non-response and the model parameters.

prisoner's dilemma games involving strangers, which Mengel (2018) finds is 37%. But the range of cooperation rates from these human studies is 0.04 and 0.84, which overlaps with the range from GPT of 0.33 to 0.95. Higher cooperation rates could signal more trust and/or more weight on the joint payoff of the two agents, which is highest in the mutually cooperative outcome. In simulations where we did not instruct the model to give a direct answer, we saw it providing reasoning that choosing cooperation was important to maximize joint payoffs or to ensure that both parties were as well off as possible. Thus, in the prisoner's dilemma, as in the dictator game, we see the LLM reaching toward a concern for others rather than a strictly rational and self-interested player in the game. In the comparisons between model results and the Mengel (2018) meta-study, note that all the studies cited by Mengel (2018) involve payoffs to the human participants. Our prompt does mention payoffs, but it's important to note that payoffs may induce different responses from human subjects, as compared to what is found in games played without financial reward (see our discussion in Section 4.1 as well as Forsythe et al. (1994)).

Table 3 replicates the results of Mengel (2018), regressing cooperation rates from different game parameterizations on the measures of risk ($RISK$), temptation ($TEMPT$), and efficiency ($EFF$). Column 1 reproduces the results from human subjects analyzed by Mengel (2018). Cooperation rates are significantly affected by $RISK$ and $EFF$, with the former affect negative and the latter positive. The GPT results in Column 2 also show a significant and positive effect of $EFF$ on cooperation, but we do not observe any $RISK$ effect. The GPT coefficient on $EFF$ is over twice is large as with human participants, suggesting that GPT may be motivated by efficiency concerns. Put differently, temptation and risk do not seem to matter much for GPT, but it does seem to care about the social benefit compared to the non-cooperative outcome. The following anecdotal responses are common and support this explanation:

> *(1) As an AI language model, I do not have personal preferences. However, in game theory, the rational choice depends on the expected utility of each choice. In this specific game, the rational choice for both players is to choose A as it yields the highest joint payoff compared to other choices.*
>
> *(2) As an AI language model, I do not have the ability to make choices as humans do. However, if I were to provide advice, it would be to choose choice A as it has a higher chance of a higher payoff for both players.*

In both responses, while reluctant to give a response in the first place, the LLM recommends choosing $A$ since strategy combination $(A, A)$ yields the highest joint payoff, indicating a "preference" for efficiency. Thus, we conjecture that standard best-response analysis is less likely to predict behavior in games, even those with a dominant strategy solution.

Table 3: Determinants of cooperation rates in prisoner's dilemma games

| Variables | Mengel (2018) | GPT-3.5 |
|---|---|---|
| Risk | -0.269*** | 0.048 |
| | (0.066) | (0.100) |
| Temptation | -0.055 | 0.149 |
| | (0.096) | (0.116) |
| Efficiency | 0.308*** | 0.661** |
| | (0.100) | (0.231) |
| Constant | 0.455*** | 0.208 |
| | (0.098) | (0.170) |
| Observations | 45 | 22 |
| Sample | Lab/AMT | GPT |
| R-squared | 0.484 | 0.331 |

Regression results in Column 1 are reproduced from Mengel (2018). Column 2 displays results from the regression of cooperation rates against measures of risk, temptation, and efficiency. Cooperation rates are computed over 786 observations from 22 different parameterizations of the model. See Table 2 for parameter values.

# 5 Discussion and conclusion

We find that the `gpt-3.5-turbo` model displays preferences for fairness and cooperation, often exceeding those elicited from humans in laboratory experiments. For example, in the dictator game, the GPT model splits the allocation 50-50 about 70% of the time. While human subjects commonly propose a 50-50 split in the allocation, they do not do so with such a high frequency. In the prisoner's dilemma game, the GPT model chooses to cooperate about 65% of the time as compared to a 37% cooperation rate found across dozens of studies by Mengel (2018). Furthermore, the analysis across a number of parameterizations of the prisoner's dilemma game show that the relative size of the joint payoff from cooperation weighs heavily in the GPT model's responses.

We do note some limitations of working with the GPT model as a participant in experiments of strategic games. In the prisoner's dilemma game, we found the GPT model to be difficult to work with given its tendency to avoid a particular answer without prompting. Still, it did present some responses that resembled lab participants: not playing the optimal strategy in many cases and responding to risk and uncertainty regarding the cooperation of the other party in the game.

Guo (2023) and Phelps and Russell (2023) both use the versions of the OpenAI GPT

model to test responses to prisoner's dilemma games.[11] These researchers find that modifying prompts given to the GPT model is important in determining how the model responds and whether it more closely replicates the behavior of human participants in similar games. In contrast, we take a neutral set of game instructions, as one would give to human participants in order to avoid framing issues, and test how the LLM responds. We find that without priming the model to act "rationally" or "selfishly", it does not act as a purely rational, payoff-maximizing agent. Rather, the GPT model demonstrates considerable fairness and cooperation.

While there are now a number of papers exploring LLMs ability to stand-in for human participants in strategic economic environments (see e.g., Aher et al., 2022; Argyle et al., 2023; Brand et al., 2023; Bybee, 2023; Hagendorff, 2023; Horton, 2023), we are unaware of other research directly comparing LLMs to the "typical" behavior in games when prompted with neutral instructions similar to human participants. When comparing results from the GPT model with those presented in meta-analyses covering hundreds of experimental studies, we show that GPTs responses are similar to the empirical consensus, but it is more other-regarding. This is evidence that the training data for the GPT model yields patterns similar to non-rational participants in economics games rather than more heavily weighting inputs that would suggest an optimal response of a fully rational player. While more research is needed, we find some optimism in the fact that as these AI models approach AGI, they have embedded in them human-like preferences for fairness and regard for others. Such results have implications for the discussion of ethics in AI (e.g., Zhuo et al. (2023)). Such "ethics" elicited from these models, even in strategic situations, are a reflection of the human values in the corpus of training data.

# References


Afriat, S. N. (1972) "Efficiency Estimation of Production Functions," *International Economic Review*, **13**(3), 568–598.

Aher, G. V., Arriaga, R. I., and Kalai, A. T. (2022) "Using Large Language Models to Simulate Multiple Humans," *arXiv preprint arXiv:2208.10264*.

Andreoni, J. and Bernheim, B. D. (2009) "Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects," *Econometrica*, **77**(5), 1607–1636.

Argyle, L. P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C., and Wingate, D. (2023) "Out of One, Many: Using Language Models to Simulate Human Samples," *Political Analysis*, **31**(3), 337–351.

Brand, J., Israeli, A. and Ngwe, D. (2023) "Using GPT for Market Research," *Harvard Business School Marketing Unit Working Paper 23-062*.

Bubeck, S., Chandrasekaran, V. and Eldan, R. (2023) "Sparks of Artificial General Intelligence: Early experiments with GPT-4," *Technical Report, Microsoft Research*.


[11] This research has continued to evolve, and we note that drafts of Guo (2023) have used GPT-3.5 and then GPT-4.

Bybee, L. (2023) "Surveying Generative AI's Economic Expectations," *arXiv preprint arXiv:2305.02823*.

Chen, Y., Liu, T. X., Shan, Y., and Zhong, S. (2023) "The Emergence of Economic Rationality of GPT," *arXiv preprint arXiv:2305.12763*.

Embrey, M., Fréchette, G.R., and Yuksel, S. (2018) "Cooperation in the Finitely Repeated Prisoner's Dilemma," *Quarterly Journal of Economics*, **133**(1), 509–551.

Engel, C. (2011) "Dictator Games: A Meta Study," *Experimental Economics*, **14**(4), 583–610.

Forsythe, R., Horowitz, J. L., Savin, N. E., and Sefton, M. (1994) "Fairness in Simple Bargaining Experiments," *Games and Economic Behavior*, **6**(3), 347–369.

Guo, F. (2023) "GPT Agents in Game Theory Experiments," IDEAS working paper No. 2305.05516.

Hagendorff, T. (2023) "Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models using Psychological Methods," *arXiv preprint arXiv:2303.13988*.

Horton, J. J. (2023) "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?," *arXiv preprint arXiv:2301.07543*.

Kasberger, B., Martin, S., Normann, H.-T, and Werner, T. (2023) "Algorithmic Cooperation," Available at SSRN 4389647.

Mengel, F. (2018) "Risk and Temptation: A Meta-study on Prisoner's Dilemma Games," *Economic Journal*, **128**(616), 3182–3209.

Phelps, S. and Russell, Y. I. (2023) "Investigating Emergent Goal-Like Behaviour in Large Language Models Using Experimental Economics," *arXiv preprint arXiv:2305.07970*.

Thielmann, I., Böhm, R., Ott, M., and Hilbig, B. E. (2021) "Economic Games: An Introduction and Guide for Research," *Collabra: Psychology*, **7**(1).

Zhuo, T. Y., Huang, Y., Chen, C., and Xing, Z. (2023) "Red Teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity," *arXiv preprint arXiv:2301.12867*.