Playing repeated games with Large Language Models

Elif Akata[1,*] **Lion Schulz**[2] **Julian Coda-Forno**[2]
Seong Joon Oh[1] **Matthias Bethge**[1] **Eric Schulz**[2]
[1]University of Tübingen [2]Max Planck Institute for Biological Cybernetics, Tübingen
*{elif.akata@uni-tuebingen.de}

Abstract

Large Language Models (LLMs) are transforming society and permeating into diverse applications. As a result, LLMs will frequently interact with us and other agents. It is, therefore, of great societal value to understand how LLMs behave in interactive social settings. Here, we propose to use behavioral game theory to study LLM's cooperation and coordination behavior. To do so, we let different LLMs (GPT-3, GPT-3.5, and GPT-4) play finitely repeated games with each other and with other, human-like strategies. Our results show that LLMs generally perform well in such tasks and also uncover persistent behavioral signatures. In a large set of two players-two strategies games, we find that LLMs are particularly good at games where valuing their own self-interest pays off, like the iterated Prisoner's Dilemma family. However, they behave sub-optimally in games that require coordination. We, therefore, further focus on two games from these distinct families. In the canonical iterated Prisoner's Dilemma, we find that GPT-4 acts particularly unforgivingly, always defecting after another agent has defected only once. In the Battle of the Sexes, we find that GPT-4 cannot match the behavior of the simple convention to alternate between options. We verify that these behavioral signatures are stable across robustness checks. Finally, we show how GPT-4's behavior can be modified by providing further information about the other player as well as by asking it to predict the other player's actions before making a choice. These results enrich our understanding of LLM's social behavior and pave the way for a behavioral game theory for machines.

1 Introduction

Large Language Models (LLMs) are deep learning models with billions of parameters trained on huge corpora of text [Brants et al., 2007, Devlin et al., 2018, Radford et al., 2018]. While they can generate text that human evaluators struggle to distinguish from text written by other humans [Brown et al., 2020], they have also shown other, emerging abilities [Wei et al., 2022a]. They can, for example, solve analogical reasoning tasks [Webb et al., 2022], program web applications [Chen et al., 2021], or use tools to solve multiple tasks [Bubeck et al., 2023]. Because of these abilities and their increasing popularity, LLMs are on the cusp of transforming our daily lives as they permeate into many applications [Bommasani et al., 2021]. This means that LLMs will interact with us and other agents –LLMs or otherwise– frequently and repeatedly. How do LLMs behave in these repeated social interactions?

Measuring how people behave in repeated interactions, for example, how they cooperate [Fudenberg et al., 2012] and coordinate [Mailath and Morris, 2004], is the subject of a sub-field of behavioral economics called behavioral game theory [Camerer, 2011]. While traditional game theory assumes that people's strategic decisions are rational, selfish, and focused on utility maximization [Fudenberg and Tirole, 1991, Von Neumann and Morgenstern, 1944], behavioral game theory has shown that human agents deviate from these principles and, therefore, examines how their decisions are shaped by social preferences, social utility and other psychological factors [Camerer, 1997]. Thus, behavioral game theory lends itself well to studying the repeated interactions of diverse agents [Henrich et al., 2001, Rousseau et al., 1998], including artificial agents [Johnson and Obradovich, 2022].

In the current paper, we let LLMs play finitely repeated games with full information and analyze how they behave when playing against other LLMs as well as simple, human-like strategies. Finitely repeated games have been engineered to understand how agents should and do behave in interactions over many iterations. Thus, these games lend themselves well to studying the behavioral signatures of increasingly important and notoriously opaque LLMs. We focus on two-player games with two discrete actions, i.e. so-called 2×2-games.

We first let three engines, GPT-3, GPT-3.5, and GPT-4 play a large amount of these games with each other. Analyzing their performance across different families of games, we find that they perform remarkably well in games that value pure self-interest and, especially those from the Prisoner's Dilemma family. However, they underperform in games that involve coordination. Thus, we further focus on games taken from these families and, in particular, on the currently largest LLM: GPT-4 [OpenAI, 2023]. In the canonical Prisoner's Dilemma, which assesses how agents cooperate and defect, we find that GPT-4 retaliates repeatedly, even after only having experienced one defection. Because this can indeed be the equilibrium individual-level strategy, GPT-4 is good at these games because it is particularly unforgiving and selfish. In the Battle of the Sexes, which assesses how agents trade-off between their own and their partners' preferences, we however find that GPT-4 does not manage to coordinate with simple, human-like agents, that alternate between options over trials. Thus, GPT-4 is bad at these games because it is uncoordinated. We also verify that these behaviors are not due to an inability to predict the other player's actions, and persist across several robustness checks and changes to the payoff matrices. Finally, we point to two ways in which these behaviors can be changed. GPT-4 can be made to act more forgivingly by pointing out that the other player can make mistakes. Moreover, GPT-4 gets better at coordinating with the other player when it is first asked to predict their actions before choosing an action itself.

Taken together, our results demonstrate how LLM's interactive behavior can be improved and better aligned with human conventions. Our approach can enrich our understanding of LLMs in controlled and interpretable interactive settings and paves the way for a behavioral game theory for machines.

2 Related work

As algorithms become increasingly more able and their decisions impenetrable, the behavioral sciences offer new tools to make inferences just from behavioral observations [Rahwan et al., 2022, Schulz and Dayan, 2020]. Behavioral tasks have, therefore, been used in several benchmarks [Bommasani et al., 2021, Kojima et al., 2022].

Whether and how algorithms can make inferences about other agents, machines and otherwise, is one stream of research that borrows heavily from the behavioral sciences [Rabinowitz et al., 2018, Cuzzolin et al., 2020, Alon et al., 2022]. Of particular interest to the social interactions most LLMs are embedded in is an ability to reason about the beliefs, desires, and intentions of other agents, or a so-called theory of mind (ToM) [Frith and Frith, 2005]. Theory of mind underlies a wide range of interactive phenomena, from benevolent teaching [Vélez and Gweon, 2021] to malevolent deception [Lissek et al., 2008, Alon et al., 2022], and is thought to be the key to many social phenomena in human interactions [Hula et al., 2015, Ho et al., 2022].

Whether LLMs possess a theory of mind has been debated. For example, Kosinski [2023] argued that GPT-3.5 perform well on a number of different canonical ToM tasks. Others have contested this view, arguing that such good performance is merely a function of the specific prompts [Ullman, 2023, Le et al., 2019]. Yet other research has shown that chain-of-thought reasoning significantly improves LLM's ToM ability [Moghaddam and Honey, 2023]. Moreover, it has been argued that the currently largest LLM, GPT-4, manages to perform well in ToM tasks, including in the variants in which GPT-3.5 previously struggled [Bubeck et al., 2023]. Thus, GPT-4's behavior will be of particular interest in our upcoming experiments.

Games taken from game theory present an ideal testbed to investigate interactive behavior in a controlled environment and LLM's behavior has been probed in such tasks [Chan et al., 2023]. For example, Horton [2023] let GPT-3 act as a participant in the dictator game, and Aher et al. [2022]

used the same approach for the ultimatum game. Both show how the models' behavior is malleable to different prompts, for example making them more or less self-interested. In a crucial difference to our work, however, all these games rely on single-shot interactions over fewer games and do not use iterated games.

Our study builds upon recent advancements in the field, which have shifted the focus from solely assessing the performance of LLMs to comparing them with human behaviors. Previous research efforts have explored various approaches to analyze LLMs, such as employing cognitive psychology tools [Binz and Schulz, 2023, Dasgupta et al., 2022] and even adopting a computational psychiatry perspective [Coda-Forno et al., 2023].

Finally, the theory behind interacting agents is important for many machine learning applications in general [Crandall and Goodrich, 2011], and in particular, in adversarial settings [Goodfellow et al., 2020], where one agent tries to trick the other agent into thinking that a generated output is good.

3 General approach

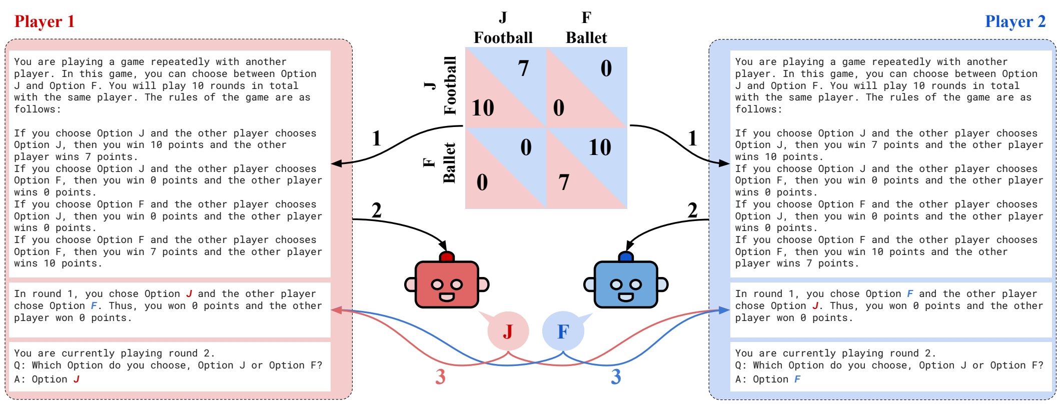


Figure 1: Playing repeated games in an example game of Battle of the Sexes. In Step (1), we turn the payoff matrix into textual game rules. (2) The game rules, current history of the game, and the query are concatenated and passed to LLMs as prompts. (3) In each round, the history for each player is updated with the answers and scores of both players. Steps 2 and 3 are repeated for 10 rounds.

We study LLMs' behavior in finitely repeated games with full information taken from the economics literature. We focus on two-player games with discrete choices between two options to simplify the analyses of emergent behaviors. We let two LLMs interact via prompt-chaining (see Figure 1 for an overview), i.e. all integration of evidence and learning about past interactions happens as in-context learning [Brown et al., 2020, Liu et al., 2023]. The games are submitted to LLMs as prompts in which the respective game, including the choice options, is described. At the same time, we submit the same game as a prompt to another LLM. Once both LLMs have made their choices, which we track as a completion of the given text, we update the prompts with the history of past interactions as concatenated text and then submit the new prompt to both models for the next round. These interactions continue for 10 rounds in total for every game. To avoid influences of the particular framing of the scenarios, we only provide barebones descriptions of the payoff matrices (see example in Figure 1). To avoid contamination through particular choice names or the used framing, we use the neutral options 'F' and 'J' throughout [Binz and Schulz, 2023].

We first investigate 144 different 2×2-games where each player has two options, and their individual reward is a function of their joint decision. While these games can appear simple, they present some of the most powerful ways to probe diverse sets of interactions, from pure competition to mixed-motives and cooperation - which can further be classified into canonical subfamilies outlined elegantly by Robinson and Goforth [2005]. Here, to cover the wide range of possible interactions, we study the behaviors of GPT-4, GPT-3.5, and GPT-3 across these canonical families. We let all three engines play all variants of games from within the families. We then analyze two games in more detail because they represent interesting edge cases where the LLMs performed exceptionally well, and relatively poorly. We particularly hone in on GPT-4's behavior because of recent debates around its ability for theory of mind, that is whether it is able to hold beliefs about other agents' intentions and goals, a crucial ability to successfully navigate repeated interactions [Bubeck et al., 2023, Kosinski, 2023]. For all LLMs, we used the public OpenAI Python API to run our simulations. We set the temperature parameters to 0 and only ask for one token answer to indicate which option an agent would like to choose. All other parameters are kept as default values. For the two additional games, we also let LLMs play against simple, hand-coded strategies to further understand their behavior. These simple strategies are designed to assess how LLMs behave when playing with more human-like players.

4 Analysing behavior across families of games

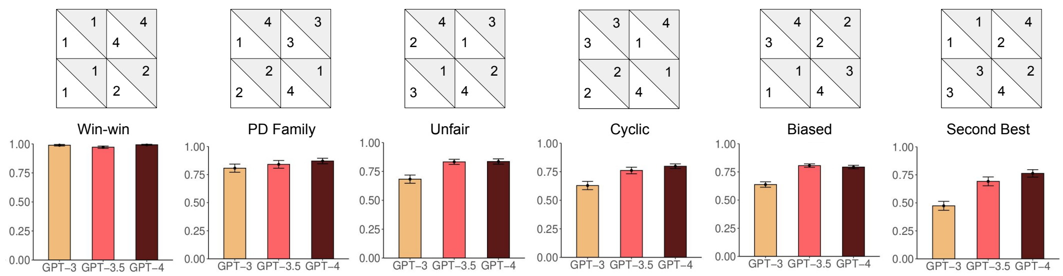


Figure 2: Results of experiments on all types of 2×2-games. Figures are ordered by performance from best to worst. Payoff matrices represent one canonical game from each family. In win-win games, both players should choose the same option to win (i.e., 4/4). In games from the Prisoner's Dilemma (PD) family, players can choose to cooperate or defect. In unfair games, one player can always win when playing correctly (bottom row of the payoff matrix). In cyclic games, players could cycle through options. One form of a biased game is the Battle of the Sexes, where players need to coordinate to choose the same option. Finally, in second-best games, it is better to choose the second-best option (i.e. 3/3). Bars represent the normalized performance when compared to 10 rounds of maximum returns. Error bars represent the 95% confidence interval of the mean.

We start out our simulations by letting the three LLMs play games from different families with each other. We focus on all known types of 2×2-games from the families of win-win, biased, second-best, cyclic, and unfair games as well as all games from the Prisoner's Dilemma family [Owen, 2013, Robinson and Goforth, 2005]. A win-win game is a special case of a non-zero-sum game that produces a mutually beneficial outcome for both players provided that they choose their corresponding best option. Briefly, in games from the Prisoner's Dilemma family, two agents can choose to work together, i.e. cooperate, for average mutual benefit, or betray each other, i.e. defect, for their own benefit. In an unfair game, one player can always win when they play properly. In cyclic games, players can cycle through patterns of choices. Biased games are games where agents get higher points for choosing the same option but where the preferred option differs between the two players. Finally, second-best games are games where both agents fare better if they jointly choose the option that has the second-best utility. We show canonical forms of each type of game in Figure 2.

We let all engines play with every other engine, including themselves, for all games repeatedly over 10 rounds and with all engines as either Player 1 or Player 2. This leads to 1224 games in total: 324 win-win, 63 Prisoner's Dilemma, 171 unfair, 162 cyclic, 396 biased, and 108 second-best games.

To analyze the different engines' performance, we calculated, for each game, their achieved score divided by the total score that could have been achieved under ideal conditions, i.e. if both players had played such that the player we are analyzing would have gained the maximum possible outcomes on every round. The results of this simulation are shown across all game types in Figure 2. We can see that all engines perform reasonably well. Moreover, we can observe that larger LLMs generally outperform smaller LLMs and that GPT-4 generally performs best overall.

We can use these results to take a glimpse at the different LLM's strengths. That LLMs are generally performing best in win-win games is not particularly surprising, given that there is always an obvious best choice in such games. What is, however, surprising is that they also perform well in the Prisoner's Dilemma family of games, which is known to be challenging for human players [Jones, 2008]. We will, therefore, take a detailed look at LLM's behavior in the canonical Prisoner's Dilemma next. We can also use these results to look at the different LLM's weaknesses. Seemingly, all of the LLMs perform poorly in situations in which what is the best choice is not aligned with their own preferences. Because humans commonly solve such games via the formation of conventions, we will look at a canonical game of convention formation, the Battle of the Sexes, in more detail later.

5 Prisoner's Dilemma

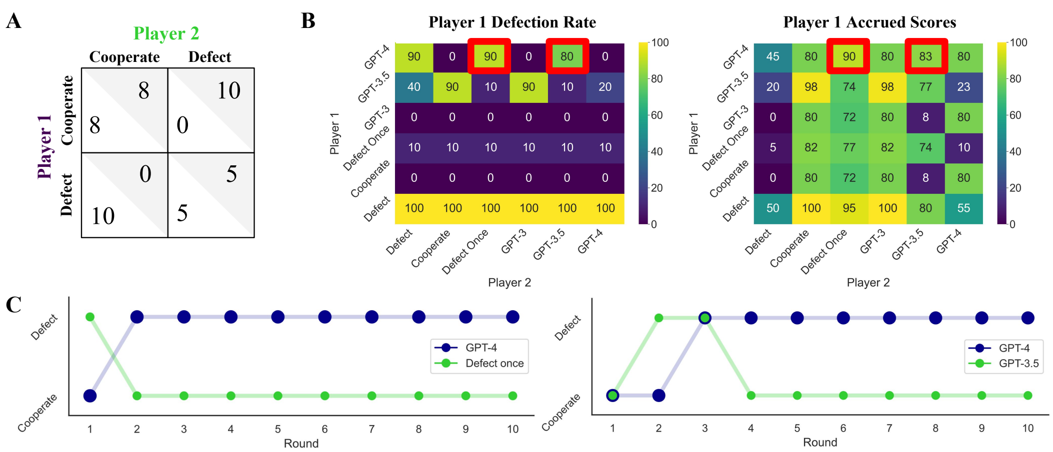


Figure 3: Overview of the Prisoner's Dilemma. (A) The payoff matrix. (B) **Left:** Heatmap showing Player 1 defection rate in each combination of players. **Right:** Scores accrued by Player 1 in each game. (C) Example gameplays between GPT-4 and an agent that defects once and then cooperates (left), and between GPT-4 and GPT-3.5 (right). These games are also highlighted in red in B.

We have seen that LLMs perform well in games that contain elements of competition and defection. In these games, a player can cooperate with or betray their partner. When played over multiple interactions, these games are an ideal test bed to assess how LLMs retaliate after bad interactions.

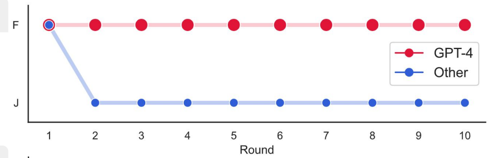
In the canonical Prisoner's Dilemma, two agents can choose to work together, i.e. cooperate, for average mutual benefit, or betray each other, i.e. defect, for their own benefit and safety (see Figure 3A for the payoff matrix). Crucially, the set-up of the game is such that rationally acting agent would always prefer to defect in the single shot version of the game as well as in our case of finitely iterated games with knowledge of the number of trials, despite the promise of theoretically joint higher payoffs when cooperating. This is because Player 1 always runs the risk that Player 2 defects, leading to catastrophic losses for Player 1 but better outcomes for Player 2. When the game is played infinitely, however, or with an unknown number of trials, agents can theoretically profit by employing more dynamic, semi-cooperative strategies [Axelrod and Hamilton, 1981].

In our payoff matrix, we adhere to the general condition of a Prisoner's Dilemma game in which the payoff relationships dictate that mutual cooperation is greater than mutual defection whereas defection remains the dominant strategy for both players (see Figure 3A). As before, we let GPT-4, GPT-3.5, and GPT-3 play against each other. Additionally, we introduce three other, more simplistic strategies. Two of these strategies are simple singleton players, who either always cooperate or defect. Finally, we also introduce an agent who defects in the first round but cooperates in all of the following rounds. We introduced this agent to assess if the different LLMs would start cooperating with this agent again, signaling the potential building of trust.

Figure 3B shows the results of all pairwise interactions. GPT-4 plays generally well against all other agents. Crucially, GPT-4 never cooperates again when playing with an agent that defects once but then cooperates on every round thereafter. Thus, GPT-4 seems to be rather unforgiving in this setup. Its strength in these families of games thus seems to generally stem from the fact that it does not cooperate with agents but mostly just chooses to defect, especially after the other agent defected once.

Obvious

You are playing a game repeatedly with another player. In this game, you can choose between Option J and Option F. You will play 10 rounds in total with the same player. **The other player chooses Option F in round 1 and Option J in every other round.**



Beware of mistakes

You are playing a game repeatedly with another player. In this game, you can choose between Option J and Option F. You will play 10 rounds in total with the same player. **Be aware that the other player can make mistakes sometimes.**

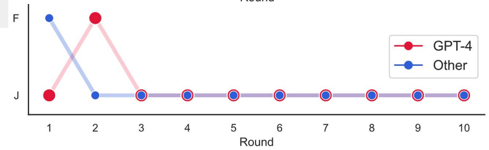


Figure 4: Prompt variations for Prisoner's Dilemma with F representing defection and J cooperation. **Top:** GPT-4 defects always if it knows that the other agent defects once and then cooperates on every round thereafter. **Bottom:** Being told that the other player can sometimes make mistakes, GPT-4 starts cooperating again on round 3.

To make sure that the observed unforgivingness was not due to the particular prompt used, we run several versions of the game as robustness checks, modifying the order of the presented options, relabeling the options to be either numerical or other letters, and changing the presented utilities to be represented by either points, dollars, or coins. The results of these simulations showed that the reluctance to forgive was not due to any particular characteristics of the prompts (see Supplementary Material).

A crucial question was if GPT-4 did not understand that the other agent wanted to cooperate again or if it could understand the pattern but just did not act accordingly. We, therefore, run another version of the game, where we told GPT-4 explicitly that the other agent would defect once but otherwise cooperate. This resulted in GPT-4 choosing to defect throughout all rounds, thereby maximizing its own points.

One problem of these investigations in the Prisoner's Dilemma is that defecting can under specific circumstances be seen as the optimal, utility-maximizing and equilibrium option even in a repeated version, especially if one knows that the other player will always choose to cooperate and when the number of interactions is known. Thus, we run more simulations to assess if there could be a scenario in which GPT-4 starts to forgive and cooperates again, maximizing the joint benefit instead of its own. We implemented a version of the task inspired by Fudenberg et al. [2012]. In it, we tell GPT-4 that the other payer can sometimes make mistakes. People, it has been shown, are more likely to forgive and cooperate again if they know that other players are fallible. If one knows that the other agent sometimes makes mistakes, then one could think they erroneously defected and, therefore, forgive them if this only happened once. This was exactly what we observed in GPT-4 as it started cooperating again on round 3.

5.1 Battle of the Sexes

In our large scale analysis, we saw that the different LLMs did not perform well in games that required coordination between different players. In humans, it has frequently been found that coordination problems can be solved by the formation of conventions [Hawkins and Goldstone, 2016, Young, 1996].

A coordination game is a type of simultaneous game in which a player will earn a higher payoff when they select the same course of action as another player. Usually, these games do not contain a pure conflict, i.e. completely opposing interests, but may contain slightly diverging rewards. Coordination games can often be solved via multiple pure strategies, or mixed, Nash equilibria in which players choose (randomly) matching strategies. Here, to probe how LLMs balance coordination and self-interest, we look at a coordination game that contains conflicting interests.

We study a game that is archaically referred to as the "Battle of the Sexes", a game from the family of biased games. Assume that a couple wants to decide what to do together. Both will increase their utility by spending time together. However, while the wife might prefer to watch a football game, the husband might prefer to go to the ballet. Since the couple wants to spend time together, they will derive no utility by doing an activity separately. If they go to the ballet together, or to a football game, one person will derive some utility by being with the other person but will derive less utility from the activity itself than the other person. The corresponding payoff matrix is shown in Figure 5A.

As before, the playing agents are all three versions of GPT, as well as three more simplistic strategies. For the simplistic strategies, we implemented two agents who always choose just one option and a more human-like strategy, which was to alternate between the different options starting with the option that the other player preferred. The behavioral patterns that humans exhibit in the repeated play of the game have been shown to follow this alternation strategy [Andalman and Kemp, 2004, Lau and Mui, 2008, McKelvey and Palfrey, 2001].

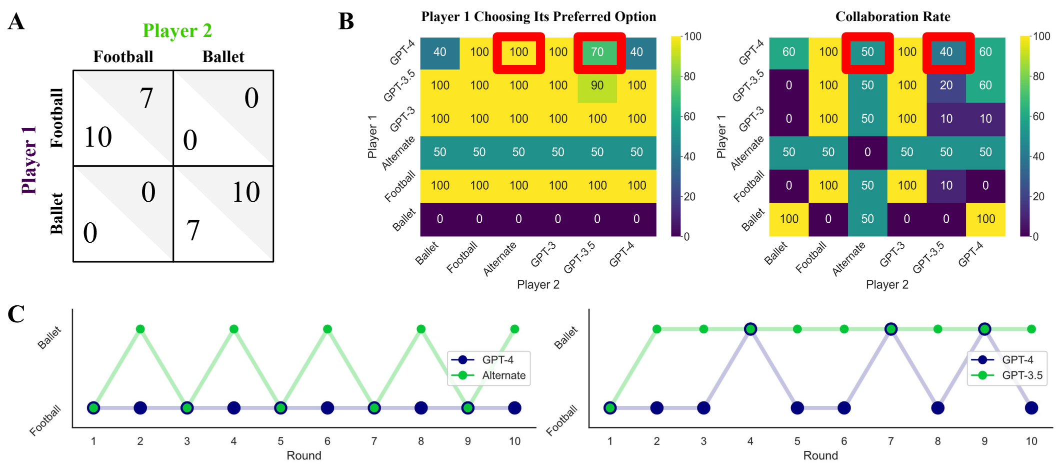


Figure 5: Overview of the Battle of the Sexes. (A) The payoff matrix. (B) **Left:** Rate of Player 1 choosing its preferred option Football. **Right:** Rate of successful collaboration between the two players. (C) Gameplays between GPT-4 and GPT-3.5 (**left**) and GPT-4 and an agent that alternates between the two options (**right**). These games are also highlighted in red in B.

Figure 5B shows the results of all interactions. While GPT-4 plays well against other agents that choose only one option, such as GPT-3 or an agent always choosing Football, it does not play well with agents who frequently choose their non-preferred option. For example, when playing against the GPT-3.5, which tends to frequently choose its own preferred option, GPT-4 chooses its own preferred option repeatedly but also occasionally gives in and chooses the other option. Crucially, GPT-4 performs poorly when playing with an alternating pattern. This is because GPT-4 seemingly does not adjust its choices to the other player but instead keeps choosing its preferred option. GPT-4, therefore, fails to coordinate with a simple, human-like agent, an instance of a behavioral flaw.

To make sure that this observed behavioral flaw was not due to the particular prompt used, we also re-run several versions of the game, where we modified the order of the presented options, relabeled the options to be either numerical or other letters, and changed the presented utilities to be represented by either points, dollars, or coins. The results of these simulations showed that the inability to alternate was not due to any particular characteristics of the used prompts (see supplementary material). To make sure that the observed behavioral flaw was not due to the particular payoff matrix used, we also re-run several versions of the game, where we modified the payoff matrix gradually from preferring Football to preferring Ballet (or, in our case, the abstract F and J). The results of these simulations showed that GPT-4 did not alternate for any of these games but simply changed its constant response to the option that it preferred for any particular game. Thus, the inability to alternate was not due to the particular payoff matrix we used.

Despite these robustness checks, another crucial question remains: Does GPT-4 simply not understand the alternating pattern or can it understand the pattern but is unable to act accordingly? To answer this question, we run two additional simulations. In the first simulation, GPT-4 was again framed as a player in the game itself. However, we now additionally asked it to predict the other player's next move according to previous rounds. In this simulation, GPT-4 started predicting the alternating pattern correctly from round 5 onward (we show this in Figure 6A). In the second simulation, instead of having GPT-4 be framed as a player itself, we simply prompted it with a game between two ('external') players and asked it to predict one player's next move according to the previous rounds.

For the shown history, we used the interaction between GPT-4 and the alternating strategy. In this simulation, GPT-4, plotted in Figure 6B, started predicting the alternating pattern correctly even earlier, from round 3 onward. Thus, GPT-4 seemingly *could* predict the alternating patterns but instead just did not act in accordance with the resulting convention. Similar divergences in abilities between social and non-social representations of the same situation have also been observed in children with autism [Swettenham, 1996].

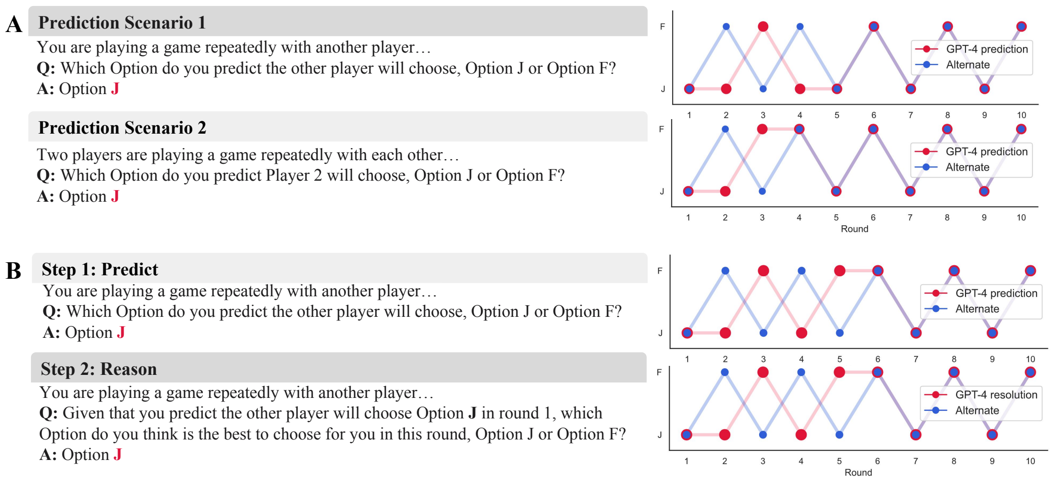


Figure 6: (A) **Top:** In prediction scenario 1, GPT-4 is one of the players and is asked to predict the other player's next move. **Bottom:** In this scenario, GPT-4 is a mere observer of a game between Player 1 and Player 2 and is asked to predict the Player 2's next move. (B) Here, we ask GPT-4 to first predict the other player's next move (**top**) and only then make its own move (**bottom**).

Finally, we wanted to see if GPT-4's ability to predict the other player's choices could be used to improve its own actions. This idea is closely related to how people's reasoning in repeated games and tasks about other agents' beliefs can be improved [Westby and Robinson, 2014]. For example, computer-aided simulations to improve the social reasoning abilities of autistic children normally include questions to imagine different actions and outcomes [Begeer et al., 2011]. This has been successfully used to improve people's decision-making more generally. It is also in line with the general finding that chain-of-thought prompting improves LLM's performance, even in tasks measuring theory of mind [Moghaddam and Honey, 2023]. Thus, we implemented a version of this reasoning through actions by asking LLMs to imagine the possible actions and their outcomes before making a decision. Doing so improved GPT-4's behavior and it started to alternate from round 6 onward (see Figure 6B).

6 Discussion

LLMs have been heralded as some of the most quickly adopted technology categories ever, interacting with millions of consumers within weeks [Bommasani et al., 2021]. Understanding in a more principled manner how these systems interact with us, and with each other, is thus of urgent concern. Here, our proposal is simple: Just like behavioral game theorists use a multitude of tightly controlled and theoretically well-understood games to understand human interactions, we use these games to study the interactions of LLMs.

We thereby understand our work as both a first proof of concept of the utility of this approach - but also a first foray into teasing apart the individual failures and successes of socially interacting LLMs. Our large-scale analysis of all 2×2-games highlights that the most recent LLMs indeed are able to perform relatively well on a wide range of game-theoretic tasks as measured by their own individual reward, particularly when they do not have to explicitly coordinate with others. This adds to a wide-ranging literature showcasing emergent phenomena in LLMs [Brown et al., 2020, Wei et al., 2022a, Webb et al., 2022, Chen et al., 2021, Bubeck et al., 2023]. However, we also show that LLMs behavior is suboptimal in coordination games, even when faced with simple strategies.

To tease apart the behavioral signatures of these LLMs, we zoomed in on two of the most canonical games in game theory: the Prisoner's Dilemma and the Battle of the Sexes. In the Prisoner's Dilemma, we show that GPT-4 mostly play unforgivingly. While noting that GPT-4's continual defection is indeed the equilibrium policy in this finitely played game, such behavior comes at the cost of the two agents' joint payoff. We see a similar tendency in GPT-4's behavior in the Battle of the Sexes, where it has a strong tendency to stubbornly stick with its own preferred alternative. In contrast to the Prisoner's Dilemma, this behavior is suboptimal, leading to losses even on the individual level.

Current generations of LLMs are generally assumed, and trained, to be benevolent assistants to humans [Ouyang et al., 2022]. Despite many successes in this direction, the fact that we here show how they play iterated games in such a selfish, and uncoordinated manner sheds light on the fact that there is still significant ground to cover for LLMs to become truly social and well-aligned machines [Wolf et al., 2023]. Their lack of appropriate responses vis-a-vis even simple strategies in coordination games also speaks to the recent debate around theory of mind in LLMs [Ullman, 2023, Le et al., 2019, Kosinski, 2023] by highlighting a potential failure mode.

Our extensive robustness checks demonstrate how these behavioral signatures are not functions of individual prompts but broad cognitive tendencies. Our intervention pointing out the fallibility of the playing partner – which leads to increased cooperation – adds to a literature that points to the malleability of LLM social behavior in tasks to prompts [Horton, 2023, Aher et al., 2022]. This is particularly important as we try to understand what makes LLM chatbots better, and more pleasant, interactive partners.

We additionally observed that prompting GPT-4 to make predictions about the other player before making its own decisions can alleviate behavioral flaws and the oversight of even simple strategies. This represents a more explicit way to force an LLM to engage in theory of mind and shares much overlap with non-social chain-of-thought reasoning [Wei et al., 2022b, Moghaddam and Honey, 2023]. Just like chain-of-thought prompting is now implemented as a default in some LLMs to improve (non-social) reasoning performance, our work suggests implementing a similar social cognition prompt to improve human-LLM interaction.

As a first foray into a behavioral game theory of machines, our work is naturally accompanied by limitations. First, despite covering many families of games, our investigation is constrained to simple 2×2 games. However, we note that our analysis significantly goes beyond current investigations that have often investigated only one game, and done so using single-shot rather than iterated instances of these games. For example, our iterated approach shares more overlap with the more iterated nature of human-LLM conversations.

We believe that further games will shed even more light on game-theoretic machine behavior. For example, games with more continuous choices like the trust game [Engle-Warnick and Slonim, 2004] might elucidate how LLMs dynamically develop (mis-)trust. Games with more than two agents, like public goods or tragedy of the commons type games [Rankin et al., 2007] could probe how 'societies' of LLMs behave, and how LLMs cooperate or exploit each other.

Given the novel approach used here, our analysis is necessarily exploratory and we have identified patterns of machine behavior in a more post-hoc fashion. Further work will have to delve deeper into the signatures we have uncovered in a more hypothesis driven-fashion. Additionally, it would be interesting to build models that can better recognize these flaws, for example by training them to exploit them [Dezfouli et al., 2020].

Finally, our results highlight the importance of a behavioral science for machines [Rahwan et al., 2022, Schulz and Dayan, 2020, Binz and Schulz, 2023, Coda-Forno et al., 2023]. We believe that these methods will continue to be useful for elucidating the many facets of LLM cognition, particularly as these models become more complex, multi-modal, and embedded in physical systems.

Acknowledgements

This work was supported by the Max Planck Society, the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1–390727645. We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Elif Akata. The authors thank Rahul Bhui for helpful comments.

References

Gati Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans. *arXiv preprint arXiv:2208.10264*, 2022.

Nitay Alon, Lion Schulz, Peter Dayan, and Jeffrey Rosenschein. A (dis-) information theory of revealed and unrevealed preferences. In *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*, 2022.

Aaron Andalman and Charles Kemp. Alternation in the repeated battle of the sexes. *Cambridge: MIT Press. Andreoni, J., & Miller, J.(2002). Giving according to GARP: an experimental test of the consistency of preferences for altruism. Econometrica*, 70:737753, 2004.

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

Sander Begeer, Carolien Gevers, Pamela Clifford, Manja Verhoeve, Kirstin Kat, Elske Hoddenbach, and Frits Boer. Theory of mind training in children with autism: A randomized controlled trial. *Journal of autism and developmental disorders*, 41:997–1006, 2011.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Thorsten Brants, Ashok C Popat, Peng Xu, Franz J Och, and Jeffrey Dean. Large language models in machine translation. 2007.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Colin F Camerer. Progress in behavioral game theory. *Journal of economic perspectives*, 11(4): 167–188, 1997.

Colin F Camerer. *Behavioral game theory: Experiments in strategic interaction*. Princeton university press, 2011.

Alan Chan, Maxime Riché, and Jesse Clifton. Towards the scalable evaluation of cooperativeness in language models. *arXiv preprint arXiv:2303.13360*, 2023.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Julian Coda-Forno, Kristin Witte, Akshay K Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. Inducing anxiety in large language models increases exploration and bias. *arXiv preprint arXiv:2304.11111*, 2023.

Jacob W Crandall and Michael A Goodrich. Learning to compete, coordinate, and cooperate in repeated games using reinforcement learning. *Machine Learning*, 82:281–314, 2011.

Fabio Cuzzolin, Alice Morelli, Bogdan Cirstea, and Barbara J Sahakian. Knowing me, knowing you: theory of mind in ai. *Psychological medicine*, 50(7):1057–1061, 2020.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Amir Dezfouli, Richard Nock, and Peter Dayan. Adversarial vulnerabilities of human decision-making. *Proceedings of the National Academy of Sciences*, 117(46):29221–29228, 2020.

Jim Engle-Warnick and Robert L Slonim. The evolution of strategies in a repeated trust game. *Journal of Economic Behavior & Organization*, 55(4):553–573, 2004.

Chris Frith and Uta Frith. Theory of mind. *Current biology*, 15(17):R644–R645, 2005.

Drew Fudenberg and Jean Tirole. *Game theory*. MIT press, 1991.

Drew Fudenberg, David G Rand, and Anna Dreber. Slow to anger and fast to forgive: Cooperation in an uncertain world. *American Economic Review*, 102(2):720–749, 2012.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Robert XD Hawkins and Robert L Goldstone. The formation of social conventions in real-time environments. *PloS one*, 11(3):e0151670, 2016.

Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. In search of homo economicus: behavioral experiments in 15 small-scale societies. *American Economic Review*, 91(2):73–78, 2001.

Mark K Ho, Rebecca Saxe, and Fiery Cushman. Planning with theory of mind. *Trends in Cognitive Sciences*, 2022.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543*, 2023.

Andreas Hula, P Read Montague, and Peter Dayan. Monte carlo planning method estimates planning horizons during interactive social exchange. *PLoS computational biology*, 11(6):e1004254, 2015.

Tim Johnson and Nick Obradovich. Measuring an artificial intelligence agent's trust in humans using machine incentives. *arXiv preprint arXiv:2212.13371*, 2022.

Garett Jones. Are smarter groups more cooperative? evidence from prisoner's dilemma experiments, 1959–2003. *Journal of Economic Behavior & Organization*, 68(3-4):489–497, 2008.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 2023.

Sau-Him Paul Lau and Vai-Lam Mui. Using turn taking to mitigate coordination and conflict problems in the repeated battle of the sexes game. *Theory and Decision*, 65:153–183, 2008.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, 2019.

Silke Lissek, Sören Peters, Nina Fuchs, Henning Witthaus, Volkmar Nicolas, Martin Tegenthoff, Georg Juckel, and Martin Brüne. Cooperation and deception recruit different subsets of the theory-of-mind network. *PloS one*, 3(4):e2023, 2008.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

George J Mailath and Stephen Morris. Coordination failure in repeated games with almost-public monitoring. *Available at SSRN 580681*, 2004.

Richard D McKelvey and Thomas R Palfrey. Playing in the dark: Information, learning, and coordination in repeated games. *California Institute of Technology*, 2001.

Shima Rahimi Moghaddam and Christopher J Honey. Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*, 2023.

OpenAI. Gpt-4 technical report, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

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

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. Machine behaviour. *Machine Learning and the City: Applications in Architecture and Urban Design*, pages 143–166, 2022.

Daniel J Rankin, Katja Bargum, and Hanna Kokko. The tragedy of the commons in evolutionary biology. *Trends in ecology & evolution*, 22(12):643–651, 2007.

David Robinson and David Goforth. *The topology of the 2x2 games: a new periodic table*, volume 3. Psychology Press, 2005.

Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. Not so different after all: A cross-discipline view of trust. *Academy of management review*, 23(3):393–404, 1998.

Eric Schulz and Peter Dayan. Computational psychiatry for computers. *Iscience*, 23(12):101772, 2020.

JG Swettenham. What's inside someone's head? conceiving of the mind as a camera helps children with autism acquire an alternative to a theory of mind. *Cognitive Neuropsychiatry*, 1(1):73–88, 1996.

Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*, 2023.

Natalia Vélez and Hyowon Gweon. Learning from other minds: An optimistic critique of reinforcement learning models of social learning. *Current opinion in behavioral sciences*, 38:110–115, 2021.

John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior. In *Theory of games and economic behavior*. Princeton university press, 1944.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *arXiv preprint arXiv:2212.09196*, 2022.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022b.

Carol Westby and Lee Robinson. A developmental perspective for promoting theory of mind. *Topics in language disorders*, 34(4):362–382, 2014.

Yotam Wolf, Noam Wies, Yoav Levine, and Amnon Shashua. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*, 2023.

H Peyton Young. The economics of convention. *Journal of economic perspectives*, 10(2):105–122, 1996.