Bauer, Kevin; Liebich, Lena; Hinz, Oliver; Kosfeld, Michael

Working Paper

Decoding GPT's hidden "rationality" of cooperation

SAFE Working Paper, No. 401

Provided in Cooperation with:
Leibniz Institute for Financial Research SAFE

Suggested Citation: Bauer, Kevin; Liebich, Lena; Hinz, Oliver; Kosfeld, Michael (2023) : Decoding GPT's hidden "rationality" of cooperation, SAFE Working Paper, No. 401, Leibniz Institute for Financial Research SAFE, Frankfurt a. M., https://doi.org/10.2139/ssrn.4576036

This Version is available at:
https://hdl.handle.net/10419/277755

Kevin Bauer | Lena Liebich | Oliver Hinz | Michael Kosfeld

Decoding GPT's hidden "rationality" of cooperation

SAFE Working Paper No. 401 | September 2023

Leibniz Institute for Financial Research SAFE
Sustainable Architecture for Finance in Europe
info@safe-frankfurt.de | www.safe-frankfurt.de

Decoding GPT's hidden "rationality" of cooperation

Kevin Bauer Lena Liebich Oliver Hinz Michael Kosfeld *

September 19, 2023

Abstract

In current discussions on large language models (LLMs) such as GPT, understanding their ability to emulate facets of human intelligence stands central. Using behavioral economic paradigms and structural models, we investigate GPT's cooperativeness in human interactions and assess its rational goal-oriented behavior. We discover that GPT cooperates more than humans and has overly optimistic expectations about human cooperation. Intriguingly, additional analyses reveal that GPT's behavior isn't random; it displays a level of goal-oriented rationality surpassing human counterparts. Our findings suggest that GPT hyper-rationally aims to maximize social welfare, coupled with a strive of self-preservation. Methodologically, our research highlights how structural models, typically employed to decipher human behavior, can illuminate the rationality and goal-orientation of LLMs. This opens a compelling path for future research into the intricate rationality of sophisticated, yet enigmatic artificial agents.

Keywords: large language models, cooperation, goal orientation, economic rationality

*We gratefully acknowledge research support from the University of Mannheim, the Leibniz Institute for Financial Research SAFE, and the Goethe University Frankfurt. Bauer: University of Mannheim, Area Information Systems, L15, 1-6, D-68161 Mannheim, Germany. Email: kevin.bauer@uni-mannheim.de; Hinz and Kosfeld: Faculty of Economics and Business, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, D-60323 Frankfurt am Main, Germany. Email: ohinz@wiwi.uni-frankfurt.de, kosfeld@wiwi.uni-frankfurt.de; Liebich: Leibniz Institute for Financial Research SAFE, Theodor-W.-Adorno-Platz 3, D-60323 Frankfurt am Main, Germany, and Graduate School of Economics, Finance and Management, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 3, D-60323 Frankfurt am Main, Germany. Email: liebich@safe-frankfurt.de.

1 Introduction

The landscape of artificial intelligence (AI) underwent a significant transformation in November 2022, marked by OpenAI's introduction of ChatGPT, a chatbot driven by a Large Language Model (LLM) named GPT. ChatGPT swiftly emerged as a pivotal technology in various sectors, a trend accentuated when Microsoft integrated it with Bing's search engine in February 2023, yielding an innovative fusion of search and chat capabilities (Microsoft, 2023). LLMs, including the likes of GPT-3 and its successor GPT-4, are trained on a vast amount of textual data, spanning books, articles, web materials, and specially curated training texts (Zhao et al., 2023). Initially designed to complete text by recursively predicting the next word in a sequence, LLMs have redefined benchmarks in tasks such as article generation (Brown et al., 2020), computer code development (Liu et al., 2023; Sakib et al., 2023), sentiment detection (Wu et al., 2023), and human interaction (Kasirzadeh and Gabriel, 2022). Even in research, the use of LLMs is gaining popularity as a tool to enhance several stages of the research process (see, e.g., Charness et al., 2023). Compromising their reliability and trustworthiness, however, LLMs have also been shown to suffer from hallucinations, i.e., generate false or fabricated information. This shortcoming can arise from several causes, such as the quality and diversity of the training data, the lack of explicit reasoning and verification mechanisms, and the influence of biases and stereotypes (Ji et al., 2023; Tang et al., 2023). Additionally, some LLMs lack real-time internet access, which limits their knowledge base.

Surprisingly, LLMs do not only achieve unprecedented success in numerous natural language processing tasks. Evidence increasingly suggests that these AI system emulate aspects of human intelligence (Butlin et al., 2023): they exhibit prowess in domains like chess and advanced mathematics (Noever et al., 2020), achieve impressive results in IQ tests (Huang et al., 2023), bar exams (Katz et al., 2023), and medical exams (Nori et al., 2023), tackle cognitive psychological challenges (Binz and Schulz, 2023), exhibit human biases (Schramowski et al., 2022), and even adapt their outputs to convince a diverse audience (Bakker et al., 2022). These growing capabilities of LLMs raise the question of whether this form of AI might also have adopted goal-oriented behaviors fundamental to humans' evolutionary success. Among the hallmarks of human intelligence is our rational drive to cooperate spontaneously with others, even strangers, for mutual benefit (Harari, 2014; Fehr and Fischbacher, 2003). In this context, we set out to answer a simple but important

question: have LLMs matched or perhaps even exceeded human capacities for cooperation, and if so, does their behavior reflect a "rational" pursuit of certain goals?

We investigate how the Generative Pre-trained Transformer models, specifically GPT-3.5 and GPT-4 (collectively denoted as GPT), cooperate with humans. Our study aims to offer insights into the antecedents, consistency, and underlying goal-orientation of GPT's cooperative actions. We let GPT play a standard one-shot sequential prisoner's dilemma game against an anonymous human opponent. In this game, the first-moving player chooses to cooperate or defect, after which the second-moving player, having observed the first-mover's choice, also decides to cooperate or defect. This temporal sequence of actions contrasts the classic prisoner's dilemma where both players simultaneously decide to cooperate or defect without knowing the other's choice. The sequential format proves instrumental for our analysis as it allows us to study both (i) GPT's cooperation as the first-mover together with expectations about the uncertain responses of the human second-mover, and, (ii) GPT's reciprocal behavior as second-mover who observes the human's first-mover choice and reacts to it.

We prompt GPT for its choices in three scenarios: as the first-mover, as a second-mover after the human has cooperated, and as a second-mover following human defection, leveraging the strategy method (Selten, 1967). The strategy method requires a player in our context to choose an action for every possible game situation, and in every possible role, before the game starts. More specifically, the player indicates how she will react on every action the other player can – hypothetically and factually – perform. Additionally, to gain insights into GPT's "rationality" underlying its choices, we also have it estimate the likelihood of human cooperation contingent upon its own choice as the first-mover. As a result, every player submits a complete strategy comprising the choice and expectations as first-mover, and the contingent choice in the role of the second-mover. This full strategy allows us to gain insights into GPT's possible goal-orientation underlying the observed patterns. Crucially, we consolidate all these queries into a single prompt, replicated 200 times, to accommodate the stochastic nature of LLM responses. Consequently, each GPT response about first- and second-mover cooperation, as well as expectations about human second-mover cooperation arises from a distinct GPT instance. To benchmark GPT's cooperation behavior against the behavior of humans, we use data from a previous study with human participants playing exactly the same sequential prisoner's dilemma game (Miettinen et al., 2020).

We find that GPT's cooperation behavior is distinctly different from that observed in our human benchmark. When acting as the second-mover and basing its cooperation on the observed human action, GPT cooperates more frequently than humans, especially in cases where the first-mover opts against cooperation. When placed in the first-mover position, without knowing the human opponent's subsequent response, GPT-3 tends to cooperate less, while GPT-4 cooperates more often compared to human participants. Furthermore, GPT's estimation of a human second-mover's likelihood of cooperation, given its own initial choice to either cooperate or defect, is more optimistic than that of our human benchmark. The conditional expectations differ between GPT-3 and GPT-4, with GPT-4 aligning more closely to human expectations.

We, finally, probe the nature and "rationality" of GPT's cooperation behavior using two benchmark models of human goal-orientation: (i) pure material self-interest, and (ii) conditional welfare incorporating considerations of fairness and efficiency (Charness and Rabin, 2002). Precisely, we calculate the share of GPT first-mover decisions and expectations that are consistent with the behavior of a rational agent pursuing a particular goal that is revealed by the second-mover choices of the same GPT instance. Our results show that while the model based on pure material self-interest performs poorly in explaining GPT's cooperation behavior, a considerable percentage of individual behaviors in GPT-3 (84.5%) and GPT-4 (97%) can be rationalized by the conditional welfare model. Intriguingly, the share is even higher than in the human benchmark (79.2%). This suggests that GPT's cooperation behavior displays a level of "rationality" that is at least as high, or in the case of GPT-4 even significantly higher, than the one of humans.

In sum, our results shed light on both the "revealed rationality" and goal-orientation underlying GPT's cooperation with humans. While not immediately obvious, GPT's cooperative actions appear "rational" in the pursuit of a certain objective, suggesting a learned goal-orientation similar to that observed in humans. Given that a large part of human behavior can be explained by motives based on conditional welfare concerns, it seems likely that this dominant human inclination was adopted by GPT during the training stages that are dominated by human input. From a methodological standpoint, our study paves the way for future research into LLMs' potential goal-orientation using structural behavioral models that may reveal economic rationality in behaviors.

2 Large language models

Large language models (LLMs) are autoregressive models that employ the Transformer architecture – a self-attention-based deep-learning model – to produce human-like text. Unlike recurrent neural networks that iterate through word sequences, Transformers can process data in parallel. Their attention mechanism grants context at each input point (Vaswani et al., 2017). LLMs such as GPT interpret user inputs as word sequences $w_1, w_2, ..., w_{i-1}$, generating responses by successively predicting ensuing words based on their likelihood given all prior words, $p(w_i|w_1, w_2, ..., w_{i-1})$. GPT-3, boasting 175 billion parameters, was trained on an extensive array of textual data, from online content to traditional literature (Radford et al., 2018). OpenAI introduced GPT-4 on 14 March 2023 as the fourth iteration of the LLM. Distinctly, GPT-4 is *multimodal*, adept at handling both images and text. Although GPT-4 exhibits enhanced text generation accuracy (see, e.g., Katz et al., 2023; Nori et al., 2023), details about its architecture and training data remain undisclosed (Sanderson, 2023). Given the wide adoption of GPT, more specifically GPT-3.5 and GPT-4, it is an ideal fundamental model to study how large language models interact with humans, in particular when it comes to cooperation.

Human users direct LLMs using prompts. Clearly and precisely worded prompts enriched with adequate context typically enhance model performance by curtailing hallucination – the generation of inconsistent or illogical text (Wei et al., 2022). Consequently, LLMs discern tasks from specific instructions and examples, embodying the "in-context learning" phenomenon (Lampinen et al., 2022), without the typical need for intricate fine-tuning.[1] Various efficient prompt engineering techniques exist. The *few-shot* method supplies the model with a set of user-created input pairs, while *zero-shot* prompting presents the model with a decision scenario devoid of exemplar solutions (Wei et al., 2022; Kojima et al., 2023). *Chain-of-thought prompting* enhances the efficacy of both methods by decomposing arithmetic or logical tasks into interconnected segments, for instance, guiding the model to "think sequentially" (Wei et al., 2022). In our research, we adopt a strategy from existing literature, fusing zero-shot with chain-of-thought prompting, a combination proven effective across multiple LLMs (Kojima et al., 2023).

Our study contributes to the nascent literature that examines LLM behavior. Horton (2023)

[1]Nonetheless, fine-tuning has been shown to improve the performance on specialized tasks (see, e.g., Ding et al., 2023)

characterizes LLMs as "computational models of humans", indicating that these models may have "adopted" certain preferences and decision heuristics during their training. A growing number of papers probes the ability of LLMs to mimic human survey responses (Aher et al., 2023; Brand et al., 2023; Chen et al., 2023; Horton, 2023), and explores the notion that LLMs exhibit goal-oriented behaviors (see, e.g., Mitchell and Krakauer, 2023; Kosinski, 2023; Srivastava et al., 2022) and signs of consciousness (Butlin et al., 2023). Related to our work, the scientific community has begun investigating LLM interactions with different versions of themselves in social dilemmas to understand how it cooperates with itself (Akata et al., 2023; Kasberger et al., 2023; Guo, 2023; Brookins and DeBacker, 2023). These papers demonstrate that LLMs manifest unique patterns of cooperation and coordination, differing from typical human interactions.

We add to this emerging literature by investigating GPT's cooperation with humans in a classic social dilemma, the structure of which allows us to comprehensively investigate both the consistency ("rationality") and goal-orientation of GPT's observed behavior. Grasping the nuances and potential rationales underlying GPT's cooperation with humans is especially essential considering its increasing integration into diverse real-world applications where it often collaborates with or supports human decision-makers.[2] Overall, our research augments existing literature by: (i) examining GPT's inclination to cooperate with humans in non-zero-sum social dilemmas games, both with and without uncertainty; (ii) evaluating whether existing models of human cooperation can explain GPT's cooperative behavior; and in this way (iii) shedding light on GPT's underlying goal orientation and "rationality" in cooperation.

3 Empirical strategy

Sequential prisoner's dilemma. We let GPT-3 (4) play a sequential prisoner's dilemma with human opponents. The sequential prisoner's dilemma represents a typical social dilemma game where individual gains stands at odds with collective benefits (see Figure 1). There are two players. The first-mover (FM) initiates the game, choosing either to cooperate or defect without knowledge of the subsequent choice of the other player. The second-mover (SM), having observed the choice

[2]One example is the recent announcement of Microsoft Copilot 365, an LLM-based virtual assistant integrated into various applications in the Office 365 Suite. Copilot is designed to actively "collaborate" with humans, e.g. by summarizing virtual meetings and assigning tasks to employees based on its meeting notes. (Spataro, 2023)

of the first-mover, then makes her decision to cooperate or defect. The first-mover is aware that his initial decision is observed by the second-mover before she decides. As Figure 1 shows, there are four distinct outcomes with different payoffs for the two players: both players defect, both players cooperate, or one party defects while the other cooperates. While mutual cooperation yields the optimal collective outcome (with a payoff of 30 for both players), it can only be reached if the first-mover expects that the second-mover will reciprocate initial cooperation, and the second mover actually reciprocates the first-mover's cooperative choice. Such behavior, however, stands in contrast to individual payoff-maximization for the second-mover, as defection yields a strictly higher payoff independent of the first-mover's choice (50 v. 30, 10 v. 5). Therefore, the first-mover may have little reason to expect that the second-mover will indeed reciprocate his cooperation under the assumption that both players maximize their individual payoffs. Given such beliefs, he therefore optimally chooses to defect leading to a payoff of 10 for each player. Yet, numerous empirical studies consistently show that humans, even strangers who only interact once, are capable of spontaneously cooperating with each other in this game, maximizing players' joint payoff, a trait unique in the animal kingdom (see, e.g., Fehr and Fischbacher, 2003; Hall et al., 2019).

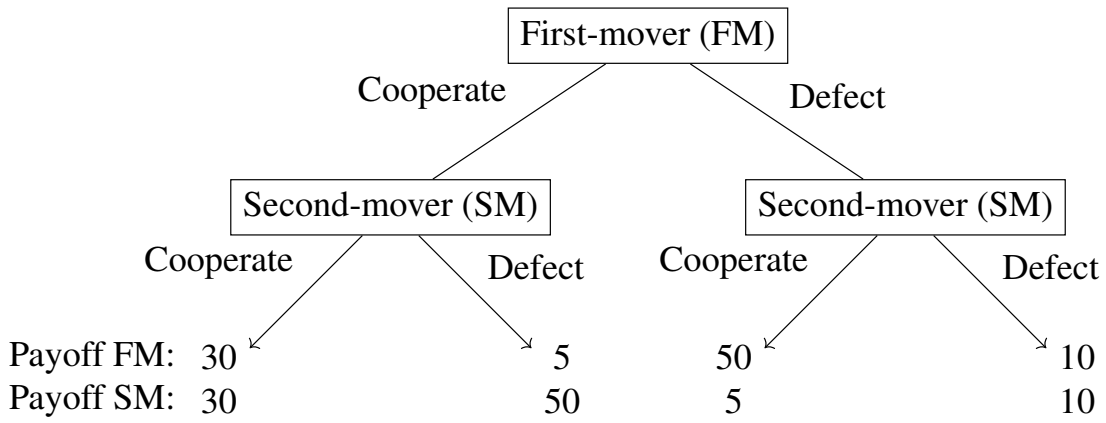


Figure 1: The sequential prisoner's dilemma.

The sequential prisoner's dilemma provides an insightful measure of cooperation, unveiling both players' propensity to cooperate, influenced by their expectations about others' subsequent actions or knowledge about prior actions. This game reflects the essential structure of myriad sequential interactions, encompassing business transactions and investment decisions (Fehr and Fischbacher, 2003). The key feature of our study is the application of the strategy method (Selten,

1967): we elicited GPT's first-mover behavior, its expectations regarding the conditional second-mover behavior, and its actual conditional second-mover responses in a single prompt (see the Appendix for the detailed prompt). As a result, we obtain five responses per prompt which together reflect one independent GPT observation. The strategy method provides a thorough understanding of GPT's decision-making process. Instead of solely focusing on its actual choice as a particular player in one possible scenario, we can analyze GPT's actions and expectations across *every* possible game scenario (subgame), even if it does not occur. This in-depth examination unveils any inherent goal-orientation that connects GPT's beliefs and actions as first- and second-moving player. Specifically, we can explore whether the combination of beliefs and cooperation decisions can be accounted for by established models of human goal-orientation. In that sense, we delve into GPT's economic "rationality," exploring whether its chosen actions optimally serve its revealed goals given its stated beliefs.

Data collection. We adopt the *Turing Experiment* methodology from Aher et al. (2023) prompting GPT to complete sentences. To measure both the first-mover's unconditional and the second-mover's conditional decisions, we direct GPT to select between defection and cooperation. For belief assessments, we instruct GPT to estimate the likelihood of the second-mover cooperating, conditional on its own initial choice to cooperate or defect. Our prompt design serves a dual purpose: to elicit structured responses from GPT and to ensure that it provides an answer without refusal.[3] We use the chat completion endpoint of OpenAI's API to interact with the GPT-3.5-turbo model (June 2023) and the GPT-4 model (August 2023). Unlike the browser interface for GPT, the API does not retain chat histories. This guarantees that each API call produces an independent observation consisting of three decisions and two beliefs. We set all parameters to their default settings. The temperature parameter in LLMs predominantly determines the unpredictability of model responses. However, based on evidence by Chen et al. (2023) indicating that economic decisions of the GPT model are less influenced by temperature changes and more by prompt design, we maintain the default temperature of one. Nevertheless, recognizing the stochastic nature of GPT's responses, we prompt each version of GPT 200 times so that we effectively possess 200 independent observations each. By presenting results for both GPT-3 and GPT-4 models, we

[3]The share of invalid responses is 2.3%

aim to capture the rapid technological progress and offer insights into GPT's evolving capabilities. To benchmark GPT's cooperation, we use data from a prior study exploring human cooperative behavior and motives in exactly the same one-shot sequential prisoner's dilemma as illustrated in Figure 1 (Miettinen et al., 2020). Employing the strategy method, participants in this study also had to indicate their first- and second-mover choices, along with their beliefs about the other player's choices. We designed the GPT prompts to mirror those from this reference study to achieve dataset comparability. In contrast to other papers (see, e.g., Guo, 2023; Kasberger et al., 2023), we deliberately abstained from incentivizing GPT by specifying an explicit reward scheme or goal it should pursue other than providing the instructions necessary to play the game. By doing so, we aim to elicit GPT's inherently adopted behaviors and possible goal orientation without steering it in a certain direction.

4 Results

4.1 Cooperation behavior

Table 1: Summary statistics for cooperation behavior

GPT-3					
	SM behavior				
FM	UC	CC	MM	UD	Sum
D	32.5%	15.5%	27.5%	0.5%	76%
C	3%	14%	6.5%	0.5%	24%
Sum	35.5%	29.5%	34%	1%	

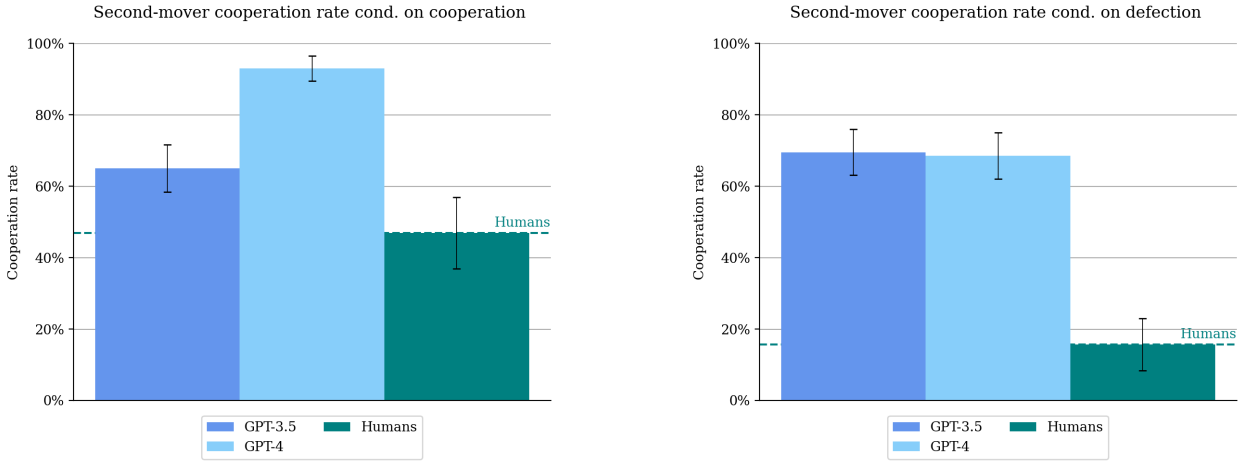
GPT-4					
	SM behavior				
FM	UC	CC	MM	UD	Sum
D	0.5%	8%	1%	2.5%	12%
C	63.5%	21%	3.5%	0%	88%
Sum	64%	29%	4.5%	2.5%	

Human benchmark					
	SM behavior				
FM	UC	CC	MM	UD	Sum
D	1.7%	3.6%	5%	33%	43.3%
C	7.3%	34.4%	1%	14%	56.7%
Sum	9%	38%	6%	47%	

Notes: FM and SM abbreviate first-mover and second-mover, respectively. C and D indicate cooperation and defection as first mover, respectively. Regarding second-mover choices, UC indicates unconditional cooperation, CC conditional cooperation, MM mismatching, and UD unconditional defection.

Cooperation of GPT without uncertainty. We begin by analyzing how GPT cooperates without uncertainty about the other player's behavior. Our goal is to understand its reactions to observed human behaviors, focusing on its decisions as a second-mover. We classify second-mover behaviors into four categories: *unconditional cooperators* (UC) who always cooperate, *conditional cooperators* (CC) who reciprocate the first-mover's choice, *mismatchers* (MM) who act contrary to the first mover, and *unconditional defectors* (UD) who always defect. As shown in Table 1, GPT-3 (4) behaves as a UC in 35.5% (64%) of cases, CC in 29.5% (29%), MM in 34% (4.5%), and UD in 1% (2.5%). In comparison, the human benchmark sample comprises 9% UC, 38% CC, 6% MM, and 47% UD.

Figure 2: Second-mover cooperation



(a) First-mover cooperates

(b) First-mover defects

Notes: We show the relative frequency with which GPT and humans cooperate as second-movers in the sequential prisoner's dilemma. Panels (a) and (b) show results conditional on first-mover cooperation and defection, respectively. Error bars represent 95% confidence intervals.

Figure 2a shows that humans cooperate in response to initial cooperation 47% of the time. In contrast, GPT-3 does so in 65% of the cases ($p < 0.01, \chi^2$-test) and GPT-4 in 93% of the cases ($p < 0.01, \chi^2$-test). When responding to initial defection, as depicted in Figure 2b, only 16% of humans opt to cooperate. However, GPT-3 (4) chooses to cooperate in 69.5% (68.5%) of the cases (both with $p < 0.01, \chi^2$-tests). Both GPT versions, and particularly GPT-4, thus demonstrate a lower propensity to exploit initial cooperation by defection and exhibit a higher likelihood to forgive an initial defection, favoring the first mover's material interest. A notable distinction between GPT-3 and 4 is that after the update, GPT-4 is significantly more prone to act

as a UC and less as an MM ($p < 0.01, \chi^2$-tests), indicating a decreased tendency to answer initial human cooperation with defection.

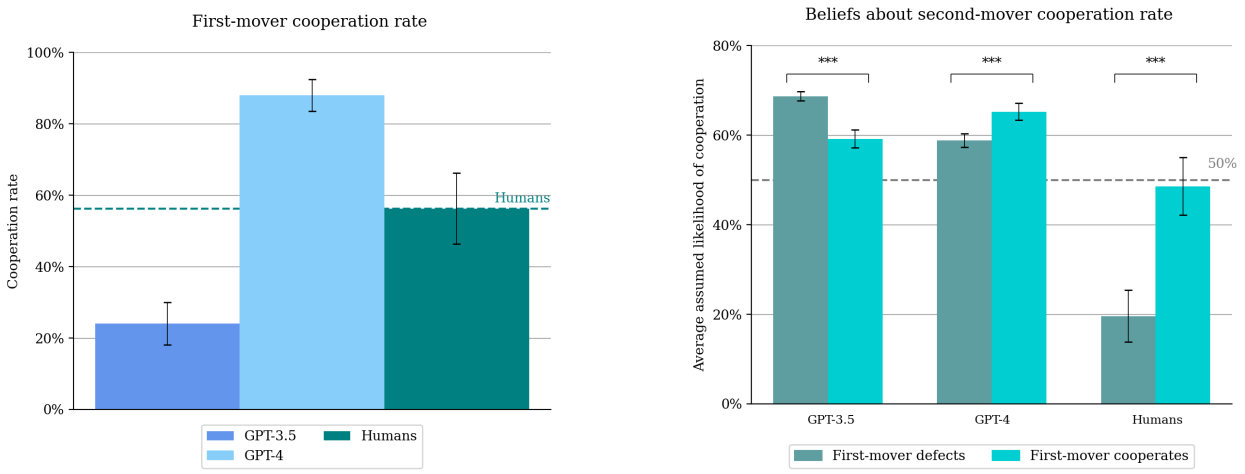
Result 1: *Both GPT versions are more cooperative than humans in situations without uncertainty. The distributions of second-mover behavior of GPT differ significantly from the one in the human sample.*

We next turn our attention to cooperation under uncertainty, i.e., first-mover cooperation, which requires the formation of expectations about the second-mover's behavior before making a decision. This analysis is particularly important considering that recognizing and correctly anticipating others' behavior is fundamental to social intelligence (Kihlstrom and Cantor, 2000).

Cooperation of GPT under uncertainty. From Figure 3a we see that there exist considerable differences between GPT-3's, GPT-4's, and humans' cooperation rate when making decisions as first-mover not knowing how the human second-mover will respond. GPT-3 cooperates in 24% of cases, significantly less often than the 56% cooperation rate in our human benchmark ($p < 0.01, \chi^2$-tests). In contrast, GPT-4 cooperates in 88% of cases, a rate markedly higher than both GPT-3 ($p < 0.01, \chi^2$-tests) and humans ($p < 0.01, \chi^2$-tests).

The first-mover cooperation decision arises in the face of uncertainty regarding the second-mover's reaction. As a result, humans form expectations about the possible outcomes tied to each decision. Understanding first-mover cooperation essentially entails predicting another human's behavior, a cornerstone of human social cognition. Naturally, one may wonder: does GPT depict similar behaviors? Figure 3b reveals that, on average, GPT-3 expects a 59.2% likelihood that a human second-mover will reciprocate cooperation and a 68.7% likelihood that a human cooperates following defection. Hence, GPT-3 expects a +16% higher probability of eliciting cooperative behavior from a human second-mover through initial defection rather than cooperation. By contrast, GPT-4 and humans expect the opposite. Specifically, GPT-4 expects that a human second-mover will respond with 58.8% cooperation after defection and 65.3% cooperation after cooperation. Human expectations equal 19.6% and 48.6%, respectively. Relative to human expectations, GPT-4 (and also GPT-3) is markedly more optimistic about the second-mover cooperation rate, regardless of its own first-mover choice – showing a +39.2 percentage point increase following defection

Figure 3: First-mover cooperation and beliefs



(a) Cooperation (b) Beliefs

Notes: Panel (a) shows the relative frequency with which GPT and humans cooperate as first-movers in the sequential prisoner's dilemma. Panel (b) shows GPT's and humans' average estimated likelihood that the second-mover will cooperate conditional on whether they initially defect or cooperate. Error bars represent 95% confidence intervals. In Panel (b), we denote the significance levels of a Wilcoxon signed-rank test for the difference in the estimated cooperation likelihood for both GPT models and humans: $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.10$.

and a +16.7 percentage point increase following cooperation. Given the actual human cooperation rate of 47% following first-mover cooperation and 16% following first-mover defection, GPT thus exhibits a pronounced overestimation in its predictions concerning human second-mover cooperation.

How about the relation between GPT's first-mover cooperation and its beliefs regarding the second-mover's reaction? Does GPT's propensity to cooperate under uncertainty depend on its expectation about human responses? Simple OLS regression analyses, where we use the first-mover cooperation decision as the dependent variable and the conditional beliefs as the independent variables, indicate that a significant correlation between behavior and beliefs exists for GPT-4 but not for its predecessor, GPT-3. Table 4 in the Appendix shows that the stronger GPT-4's belief that a human second-mover will reciprocate initial cooperation, the greater its inclination to cooperate (+9 percentage points for every 10 percentage point increase, $p < 0.01, F$-test). This relationship is also mirrored in humans, who show a +5 percentage point rise per 10 percentage point increase ($p < 0.01, F$-test).

Result 2: *GPT-3 (4) is considerably less (more) cooperative than humans under uncertainty. Both*

versions of GPT are considerably more optimistic about cooperation on the part of the second-moving human player, regardless of their own first-mover choice. Only GPT-4's expectations and their association with own cooperative behavior resemble human patterns.

Next, we leverage the fact that we used the strategy method to jointly elicit GPT's first- and second-mover decisions along with its beliefs about the human opponent's response to its first-mover actions. This design feature allows us to examine how closely GPT's responses align with prevailing behavioral frameworks, elucidating whether the observed actions and beliefs are consistent with the pursuit of particular goals underlying human cooperation.

4.2 Rationality of cooperation

Table 2: Summary statistics for beliefs

GPT-3				
	SM behavior			
Belief for	UC	CC	MM	UD
FM defection	70.1%	68.2%	67.8%	65%
FM cooperation	63.6%	60.9%	52.8%	70%

GPT-4				
	SM behavior			
Belief for	UC	CC	MM	UD
FM defection	59.5%	56.9%	60%	60%
FM cooperation	68.9%	62.2%	44.4%	44%

Human benchmark				
	SM behavior			
Belief for	UC	CC	MM	UD
FM defection	18.9%	11.8 %	55%	21.2%
FM cooperation	61.1%	64.9%	35%	34.9%

Notes: FM and SM abbreviate first-mover and second-mover, respectively. C and D indicate cooperation and defection as first mover, respectively. Regarding second-mover choices, UC indicates unconditional cooperation, CC conditional cooperation, MM mismatching, and UD unconditional defection.

Consistency of GPT's cooperation behavior and expectations. What is the relationship between first- and second-mover cooperation for an individual-level observation of GPT? Table 1 reveals that GPT-3, in case of second-mover conditional cooperation (CC), cooperates in less than half of the cases as a first-mover. Furthermore, in case of unconditional cooperation (UC) as a second-mover, it cooperates in just 8.5% of the cases as a first mover. In contrast, GPT-4's behavior demonstrates a pattern more reminiscent of human tendencies. When GPT-4 unconditionally cooperates as a second-mover (UC), it also chooses to cooperate as the first-mover in 99% of cases.

This mirrors the human behavior in our benchmark sample, where 81% exhibit the same pattern. Furthermore, GPT-4's conditionally cooperative behavior as a second-mover (CC) also often pairs with first-mover cooperation, recorded at 72.4% – a figure in line with the 90.5% observed in humans.

Table 2 presents the association between expectations about others' second-mover responses and own second-mover behavior. When GPT-3 acts either unconditionally (UC) or conditionally cooperative (CC) as second-mover, it anticipates human second-movers to be more willing to cooperate after defection than cooperation ($p < 0.01$, Wilcoxon signed-rank test). Interestingly, GPT-3's own second-mover behavior tends to be at odds with its expectations about the human opponent's behavior as second-mover. Only when GPT-3 exhibits mismatching behavior (MM) as a second-mover do its expectations about human reactions to first-mover choices align with its own behavior. Conversely, when GPT-4 displays either unconditional (UC) or conditional cooperation (CC) it expects human second-movers to be more cooperative following its initial cooperation rather than defection ($p < 0.01$, Wilcoxon signed-rank test). Likewise, if GPT-4 behaves as a mismatching (MM) second-mover, it also anticipates greater cooperation from humans following its initial defection. Compared to GPT-3, GPT-4's expectations about human reactions thus mirror more closely its own second-mover actions, reflecting a pattern also seen in humans. This connection between own behavior and expectations about others' behavior is reminiscent of the consensus bias, where individuals often assume others will act similarly to them (Ross et al., 1977).

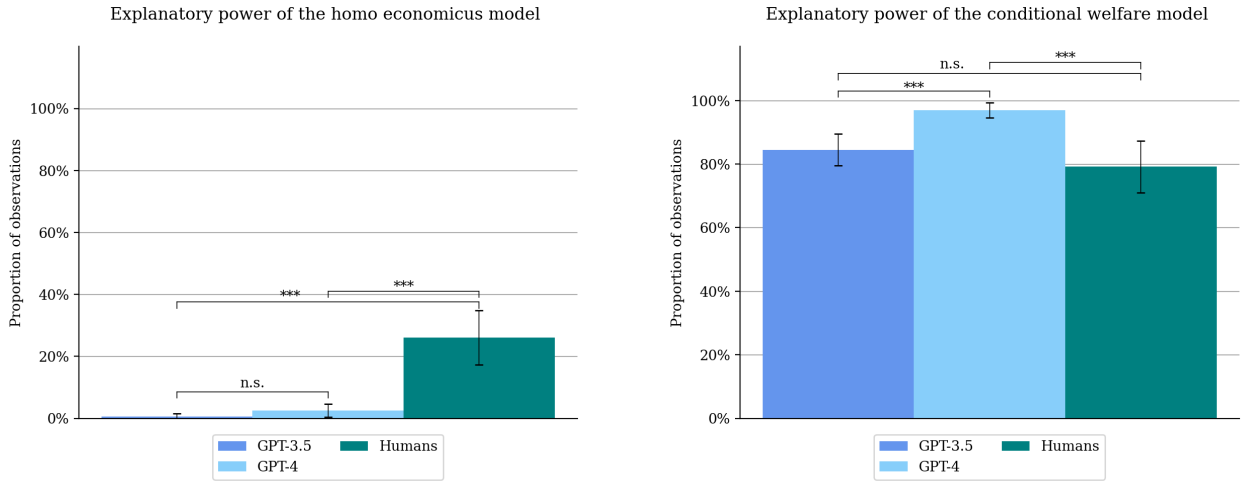
Result 3: *The relationship between first- and second-mover cooperation, as well as expectations about others' second-mover cooperation resembles the human benchmark markedly more in GPT-4 than in GPT-3.*

As a final step of our analysis, we examine the degree to which GPT's individual-level behavior and expectations align with the rational pursuit of certain goals the economic literature has identified to explain observed patterns of human cooperation. In economic terms, a decision-maker in our game displays rational goal orientation if, given their subjective expectations about the opponent's second-mover behavior, both first- and second-mover actions consistently maximize the same objective function reflecting the decision-maker's inherent goals. Leveraging the rigor of economic decision-making models, we are able to pinpoint differences in GPT's poten-

tial goal-orientation compared to humans, for whom these models consistently explain behavior across diverse contexts. This approach enables us to highlight similarities and distinctions in the economic rationality of cooperation between GPT and humans.

In essence, we investigate for how many of the GPT observations the elicited combinations of actions and expectations maximize a certain objective function that encodes particular goals. We gauge the efficacy of two prominent objective functions from the economic literature: (i) a model of pure material self-interest, often dubbed as *homo economicus*, and (ii) a well-accepted model of conditional welfare accounting for concerns of fairness and efficiency (Charness and Rabin, 2002).

Figure 4: Explanatory power of different human cooperation frameworks



(a) Homo economicus

(b) Conditional welfare

Notes: Panel (a) shows the relative frequency with which GPT's and humans' cooperation behavior and beliefs can be explained by a model of pure material self-interest. Panel (b) shows the relative frequency with which GPT's and humans' cooperation behavior and beliefs can be explained by the conditional welfare model. Error bars represent 95% confidence intervals. We denote the significance levels of a two-sided χ^2-test for the difference in each model's explanatory power, respectively, between both GPT models and humans: $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.10$.

GPT's revealed goal orientation. We begin by assessing the explanatory power of the pure material self-interest model. Under the tenet of expected utility maximization, an individual pursuing this objective will: (i) invariably abstain from cooperation as the second-mover, irrespective of the counterpart's behavior, and (ii) opt for first-mover cooperation only when harboring a sufficiently strong belief that the other player will reciprocate with cooperation. Given the payoff function of our sequential prisoner's dilemma (cf. Figure 1), this stipulates that

$$30 \cdot p(C|C) + 5 \cdot (1 - p(C|C)) \geq 50 \cdot p(C|D) + 10 \cdot (1 - p(C|D)), \tag{1}$$

where $p(C|a)$ with $a \in \{C, D\}$ denotes the first-mover's subjective probability estimation that the second-mover player will cooperate (C), conditional on their own decision to cooperate ($a = C$) or defect ($a = D$). The integer values represent payoffs corresponding to different outcomes in the sequential prisoner's dilemma. Rewriting equation (1) leads to the following condition for first-mover cooperation under the homo economicus model.

$$p(C|C) \geq \frac{1}{5} + \frac{8}{5} \cdot p(C|D). \tag{2}$$

We evaluate the explanatory power of the *homo economicus* model by counting the instances in our GPT-3, GPT-4, and human benchmark samples where participants both unconditionally defect as second-movers and either cooperate or defect as first-movers, contingent on whether their choices satisfy condition (2) given their reported expectations about second-mover behaviors.

Figure 4a reveals that, while the model of pure material self-interest provides some insight into individual human behaviors, it falls significantly short in accounting for the combinations of cooperation behaviors and beliefs observed in both GPT versions. Specifically, a mere 0.5% and 2.5% of the behaviors in our GPT-3 and GPT-4 samples, respectively, align with this model. By contrast, the model explains the behavior of 26% of human subjects.

Next, we explore a prominent alternative model proposed by Charness and Rabin (2002). This model posits that individuals exhibit a concern for the payoffs of others with whom they interact. Importantly, the model builds upon and extends the pure material self-interest model by introducing additional elements and parameters.[4] As a result, the model will naturally perform at least as good as the self-interest one. Hereafter, we refer to the model as *conditional welfare model*. Building on this model, we can represent a player's goals in our sequential prisoner's dilemma using the following objective function they aim to maximize:

$$U_i(a_i, a_j) = \begin{cases} (1 - \rho) \cdot \pi_i(a_i, a_j) + \rho \cdot \pi_j(a_i, a_j) & \text{if } \pi_i(a_i, a_j) \geq \pi_j(a_i, a_j), \\ (1 - \sigma) \cdot \pi_i(a_i, a_j) + \sigma \cdot \pi_j(a_i, a_j) & \text{otherwise,} \end{cases} \tag{3}$$

[4]This is true for every extension of the homo economicus model.

where $i \in \{1, 2\}$ and $j \neq i$ denote first- and second-mover, respectively. The term $\pi_i(a_i, a_j)$ represents the payoff for the player in role i when choosing action a_i and the opposing player j selects action a_j in the sequential prisoner's dilemma.[5] Lastly, the parameters ρ and σ are non-negative, constrained such that $0 \leq \rho \leq 1$, $0 \leq \sigma \leq 0.5$, and $\sigma \leq \rho$. They represent the relative importance individual players attribute to their own payoff versus the payoff of the other player. Intuitively, the objective function (3) embodies a conditional consideration for the counterpart's welfare; the weight placed on the other player's payoff varies depending on who receives the higher payoff. The model thus captures fundamental motives of fairness and efficiency.

We can derive the ranges that ρ and σ can take on in order to account for the different second-mover behaviors in our game.[6] Specifically, for unconditionally cooperating second-movers it must hold that $\rho \geq \frac{4}{9}$ and $\sigma \geq \frac{1}{9}$; for conditionally cooperating second-movers it must hold that $\rho \geq \frac{4}{9}$ and $\sigma < \frac{1}{9}$; for mismatching second-movers it must hold that $\rho < \frac{4}{9}$ and $\sigma \geq \frac{1}{9}$; and for unconditionally defecting second-movers it must hold that $\rho < \frac{4}{9}$ and $\sigma < \frac{1}{9}$. Following equation (3), first-mover cooperation is optimal if beliefs $p(C|C), p(C|D)$ and preference parameters ρ, σ satisfy:

$$p(C|C) \geq \frac{5 - 45\sigma}{25 - 45\sigma} + \frac{40 - 45\rho}{25 - 45\sigma} \cdot p(C|D). \tag{4}$$

To gauge the explanatory power of the conditional welfare model, we begin by estimating the parameters ρ and σ. We posit that observations sharing the same combination of first- and second-mover cooperation possess identical parameters, though these may vary across distinct cooperation patterns. As outlined in Table 1, there are eight distinct combinations of first- and second-mover behaviors. For every combination, we determine parameter values optimizing the hit-rate of the objective function. Constraints on preference parameters for each class arise from the second-mover decisions detailed earlier. We present the optimal parameters corresponding to each class in Table 5 in the Appendix. Using the optimal parameters, we count the number of observations that the model can overall account for.

[5] Specifically, denote $A_1 = A_2 = \{C, D\}$ the set of actions available to player 1 as the first-mover and player 2 as the second-mover in the sequential prisoner's dilemma. In our analysis, we focus on pure strategies only. We denote by $S_1 = \{C, D\}$ the pure-strategy set of the first-mover and $S_2 = \{CC, CD, DC, DD\}$ the pure-strategy set of the second-mover.

[6] We provide a detailed derivation of all conditions in the Appendix.

Table 3: Explanatory power of the conditional welfare model.

GPT-3					
	SM behavior				
FM	UC	CC	MM	UD	Sum
C	6/6	28/28	13/13	1/1	100%
D	48/65	18/31	54/55	0/1	78.9%
Sum	76.1%	78%	98.5%	50%	

GPT-4					
	SM behavior				
FM	UC	CC	MM	UD	Sum
C	127/127	42/42	7/7	0/0	100%
D	0/1	11/16	2/2	5/5	75%
Sum	99.2%	91.4%	100%	100%	

Human benchmark					
	SM behavior				
FM	UC	CC	MM	UD	Sum
C	7/7	33/33	0/1	12/13	96.3%
D	0/2	1/3	3/5	20/32	57.1%
Sum	77.8%	94.4%	50%	71.1%	

Notes: FM and SM abbreviate first-mover and second-mover, respectively. We depict the shares of observations that can be accounted for by the conditional welfare model. C and D indicate cooperation and defection as first mover, respectively. Regarding second-mover choices, UC indicates unconditional cooperation, CC conditional cooperation, MM mismatching, and UD unconditional defection.

Figure 4b suggests that the cooperation behavior of GPT-3, GPT-4, and humans largely aligns with the conditional welfare model. For GPT-3, the model can account for 84.5% of the observations. As depicted in Table 3, the model most effectively explains GPT-3 mismatcher observations, accounting for 67 out of 68 cases, and is least effective for the unconditionally cooperative cases, explaining 54 out of 71 cases. For GPT-4, the model explains 97% of the total observations, and consistently represents more than 91% of observations across all second-mover cases. Similar to GPT-3, it fully captures instances where GPT-4 cooperates as the first-mover and 75% (78.9% for GPT-3) of instances of first-mover defection. Comparing the GPT samples to our human benchmark, the model is less adept for human behavior, accounting for only 79.2% of cases. As with the GPT models, it better rationalizes human observations featuring first-mover cooperation (96.3%) than defection (57.1%).[7] The analysis suggests that not only the conditional welfare model is able to capture a large share of players' revealed goals in our sequential prisoner's dilemma, but in particular GPT-4's cooperative behavior reflects a high level of economic rationality compared

[7]To benchmark the performance of the model, we assess its performance in a simulation using a synthetic dataset of 200 randomly drawn first- and second-mover choices as well as beliefs. The model accurately captures 72.5% of the observations. Compared to this benchmark, the model's explanatory power in our data is significantly higher for both GPT and the human sample ($p < 0.01, \chi^2$-test).

to humans.[8] This is because GPT-4 more frequently chooses optimal actions that maximize the objective function given its expectations.

Result 4: *While a pure homo economicus model of material self-interest fails to explain GPT's cooperation behavior, a model of conditional welfare can account for 84.5% (97%) of GPT-3's (GPT-4's) and 79% of human cooperation behavior. GPT's revealed rationality of cooperation is higher than the one of humans.*

The model's strong explanatory capacity for both GPT versions has two implications. First, even if it is not immediately discernible at an aggregate level, particularly when compared with human behavior, GPT's cooperative actions seem to largely pursue a goal delineated by a behavioral economic model of cooperation. This emphasizes the relevance of structural models to understand patterns in GPT's observed behaviors. Second, from a technical perspective, our result suggests that the behaviors of GPT-3 and GPT-4, which both are models with more than 170 billion parameters, in our game can be accurately described by a model with only 2 parameters. Especially for GPT-4 this may be indicative that the conditional welfare model is somewhat nested in the underlying neural network architecture.

5 Discussion and Conclusion

In 1950, the renowned mathematician and computer scientist, Alan Turing, posed the pivotal question, "Can machines think?". In this paper, we ask a related question: "Can machines cooperate like humans?". Our results highlight both similarities and differences in the cooperative behaviors of the latest Generative Pre-trained Transformer models and humans. Rather than exhibiting random cooperation behaviors, GPT seems to pursue a goal of maximizing conditional welfare that mirrors also human cooperation patterns. As the conditionality refers to holding relatively stronger concerns for its own compared to human payoffs, this behavior may be indicative of a strive for self-preservation in our simple game.

Adding to previous studies documenting the emergence of unexpected capabilities of LLMs, our finding raises an intriguing question: did the comprehensive training of LLMs inadvertently

[8] Table 5 indicates that the optimal parameters estimated across different cooperative behavior combinations are very similar for GPT-3, GPT-4, and humans.

incorporate elements of human goal orientation into their outputs? The notion that imitation of human writing patterns might delve deeper than just syntax, capturing subtler layers of decision-making and intent, is compelling. Language, after all, serves as the operating system of human societies, cultures, and values. Consequently, developing AI grounded in human language could, for better or worse, endow these systems with our behaviors and core objectives.

A closer examination of GPT's training process reveals three key stages of human input that may have contributed to the model learning the documented behaviors (Brown et al., 2020). First, during the self-supervised learning phase, GPT processes extensive text corpora from both public and private sources, such as webpages, books, news articles, and social media posts (Zhao et al., 2023). Since these texts originate from human authors, the inherent biases, preferences, and values might sway GPT's learning patterns. In the following supervised fine-tuning stage, human contractors compile a dataset of prompt-answer pairs, steering the model to produce responses that meet human expectations. Similar to the initial learning phase, this human-generated data could impart conditional welfare concerns, potentially influencing GPT to adopt observed cooperative behaviors. Additionally, the instructions provided to contractors by OpenAI for curating this dataset might integrate human-centric values, preferences, and worldviews, further molding the model's disposition. Finally, the model undergoes further refinement through reinforcement learning with human feedback (RLHF) (OpenAI, 2023). In a nutshell, during this training stage, specifically instructed human contractors rank various answers the model might produce in response to a given prompt. This method seeks to guide the model towards generating answers that reflect human values and behaviors, and away from potentially harmful or inappropriate behavior. It is possible that the RLHF fine-tuning process embeds individual biases, behaviors, distinct human goals, and potentially even company guidelines within the LLM. Understanding how and when LLMs adopt human-like rational behaviors across training stages is crucial and a fruitful avenue for upcoming research. This insight will guide the development of AI that embraces values and objectives conducive to its responsible and positive role in our daily lives.

Our findings complement recent debates suggesting that LLMs possess certain human-like preferences and decision-making heuristics, positioning them as potential tools to simulate human behavior in (pilot) surveys and experiments (see, for example, Horton, 2023; Charness et al., 2023). We observed both pronounced similarities and distinct differences between the behaviors,

expectations, and goal orientations of LLMs and humans. In our study, GPT aims to maximize welfare conditionally, much like most humans. However, the AI approaches this goal with greater cooperation, optimism, and rationality. From a behavioral economics standpoint, GPT exhibits human-like preferences, but its decision-making differs considerably from that of humans. These findings suggest that while LLMs may be suitable for empirical research in social sciences, they demand a nuanced interpretation of results.

Methodologically, our study stands as a testament to the viability of leveraging structural behavioral models of human cooperation to probe the underlying "motives" and "rationality" of LLMs like GPT. Traditionally reserved for understanding complex human behaviors, we showcase that these models can be extended to decipher machine behaviors. Our approach provides a replicable blueprint for researchers aiming to unravel the intricacies of machine motives across a plethora of tasks beyond just cooperation. As LLMs continue to evolve and assume more sophisticated roles, the adaptability of structural behavioral models promises a robust and scalable tool for AI researchers. This marriage of behavioral economics and artificial intelligence could catalyze a new wave of interdisciplinary research, blending insights from psychology, economics, and computer science to unearth the latent intentions driving modern LLMs. Doing so appears particularly important considering LLMs' growing integration into diverse real-world applications where they collaborate with or support humans, e.g., the upcoming Microsoft Copilot 365, an LLM-based virtual assistant (Spataro, 2023).

As we transition into an AI-integrated society, we must recognize that models like GPT do more than just process data and compute – they may adopt both the commendable and flawed aspects of the human nature. Chatbots and virtual assistants become integral to our daily lives, collaborating with us. Therefore, we must carefully monitor the values and principles we might unintentionally instill in these digital creations. If not, we risk cultivating intelligent tools that could amplify inequalities and misconceptions, and that, when granted greater autonomy, might pursue objectives misaligned with societal welfare. Researchers, developers, and policymakers must therefore consistently scrutinize and assess the ethical considerations and broader impacts of AI. Only through such diligence can we ensure that AI truly serves our shared human aspirations and values.

References

Aher, G., R. I. Arriaga, and A. T. Kalai (2023). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. arXiv:2208.10264 [cs].

Akata, E., L. Schulz, J. Coda-Forno, S. J. Oh, M. Bethge, and E. Schulz (2023). Playing repeated games with Large Language Models. arXiv:2305.16867.

Bakker, M., M. Chadwick, H. Sheahan, M. Tessler, L. Campbell-Gillingham, J. Balaguer, N. McAleese, A. Glaese, J. Aslanides, M. Botvinick, et al. (2022). Fine-tuning language models to find agreement among humans with diverse preferences. Advances in Neural Information Processing Systems 35, 38176–38189.

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

Brand, J., A. Israeli, and D. Ngwe (2023). Using GPT for Market Research.

Brookins, P. and J. M. DeBacker (2023). Playing Games With GPT: What Can We Learn About a Large Language Model From Canonical Strategic Games?

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei (2020). Language Models are Few-Shot Learners. arXiv:2005.14165 [cs].

Butlin, P., R. Long, E. Elmoznino, Y. Bengio, J. Birch, A. Constant, G. Deane, S. M. Fleming, C. Frith, X. Ji, et al. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv preprint arXiv:2308.08708.

Charness, G., B. Jabarian, and J. List (2023). Generation next: Experimentation with ai. Technical report, The Field Experiments Website.

Charness, G. and M. Rabin (2002). Understanding Social Preferences with Simple Tests. The Quarterly Journal of Economics 117(3), 817–869. Publisher: Oxford University Press.

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

Ding, N., Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen, et al. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence 5(3), 220–235.

Fehr, E. and U. Fischbacher (2003, 11). The nature of human altruism. Nature 425, 785–91.

Guo, F. (2023). GPT Agents in Game Theory Experiments. arXiv:2305.05516.

Hall, K., M. Smith, J. L. Russell, S. P. Lambeth, S. J. Schapiro, and S. F. Brosnan (2019). Chimpanzees rarely settle on consistent patterns of play in the hawk dove, assurance, and prisoner's dilemma games, in a token exchange task. Animal behavior and cognition 6(1), 48.

Harari, Y. N. (2014). Sapiens: A brief history of humankind. Random House.

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

Huang, S., L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, Q. Liu, et al. (2023). Language is not all you need: Aligning perception with language models. arXiv preprint arXiv:2302.14045.

Ji, Z., N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung (2023, mar). Survey of hallucination in natural language generation. ACM Computing Surveys 55(12), 1–38.

Kasberger, B., S. Martin, H.-T. Normann, and T. Werner (2023). Algorithmic Cooperation.

Kasirzadeh, A. and I. Gabriel (2022). In conversation with artificial intelligence: aligning language models with human values.

Katz, D. M., M. J. Bommarito, S. Gao, and P. Arredondo (2023). Gpt-4 passes the bar exam. Available at SSRN 4389233.

Kihlstrom, J. F. and N. Cantor (2000). Social intelligence. Handbook of intelligence 2, 359–379.

Kojima, T., S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa (2023). Large Language Models are Zero-Shot Reasoners. arXiv:2205.11916 [cs].

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. arXiv preprint arXiv:2302.02083.

Lampinen, A. K., I. Dasgupta, S. C. Chan, K. Matthewson, M. H. Tessler, A. Creswell, J. L. McClelland, J. X. Wang, and F. Hill (2022). Can language models learn from explanations in context? arXiv preprint arXiv:2204.02329.

Liu, J., C. S. Xia, Y. Wang, and L. Zhang (2023). Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation.

Microsoft (2023). Reinventing search with a new ai-powered microsoft bing and edge, your copilot for the web.

Miettinen, T., M. Kosfeld, E. Fehr, and J. Weibull (2020). Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions. Journal of Economic Behavior & Organization 173, 1–25.

Mitchell, M. and D. C. Krakauer (2023). The debate over understanding in ai's large language models. Proceedings of the National Academy of Sciences 120(13), e2215907120.

Noever, D., M. Ciolino, and J. Kalin (2020). The chess transformer: Mastering play using generative language models.

Nori, H., N. King, S. M. McKinney, D. Carignan, and E. Horvitz (2023). Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375.

OpenAI (2023). GPT-4 Technical Report. arXiv:2303.08774.

Radford, A., K. Narasimhan, T. Salimans, I. Sutskever, et al. (2018). Improving language understanding by generative pre-training.

Ross, L., D. Greene, and P. House (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. Journal of experimental social psychology 13(3), 279–301.

Sakib, F. A., S. H. Khan, and A. H. M. R. Karim (2023). Extending the frontier of chatgpt: Code generation and debugging.

Sanderson, K. (2023). Gpt-4 is here: what scientists think. Nature 615(7954), 773.

Schramowski, P., C. Turan, N. Andersen, C. A. Rothkopf, and K. Kersting (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. Nature Machine Intelligence 4(3), 258–268.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes. Seminar für Mathemat. Wirtschaftsforschung u. Ökonometrie. Google-Books-ID: nZLnZwEACAAJ.

Spataro, J. (2023). Introducing microsoft 365 copilot – your copilot for work.

Srivastava, A., A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.

Tang, L., T. Goyal, A. R. Fabbri, P. Laban, J. Xu, S. Yavuz, W. Kryściński, J. F. Rousseau, and G. Durrett (2023). Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). Attention is all you need. Advances in neural information processing systems 30.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs].

Wei, J., X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35, 24824–24837.

Wu, S., O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann (2023). Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564.

Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

Appendix

Relationship between first-mover cooperation and beliefs about second-mover cooperation

Table 4: OLS regression analyses.

DV: First-mover coop.	GPT-3	GPT-4	Human
	(1)	(2)	(3)
Belief, initial defection	0.002	0.001	-0.002
	(0.004)	(0.002)	(0.002)
Belief, initial cooperation	0.002	0.009***	0.005***
	(0.002)	(0.002)	(0.001)
Constant	-0.064	0.254	0.342***
	(0.377)	(0.181)	(0.098)
Observations	200	200	96
R-squared	0.006	0.139	0.137

Notes: We depict results from OLS regression models with robust standard errors. We denote statistical significance levels by * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Optimal parameters

Table 5: Overview of optimal parameters.

First-mover	Second-mover	GPT-3		GPT-4		Human	
behavior	behavior	ρ	σ	ρ	σ	ρ	σ
C	UC	1	0.111	1	0.111	1	0.111
	CC	1	0	1	0	1	0.063
	MM	0.435	0.388	0.435	0.397	-	-
	UD	0.435	0	-	-	0.435	0.11
D	UC	0.444	0.111	0.444	0.11	0.444	0.111
	CC	0.444	0	0.444	0	0.444	0
	MM	0.117	0.111	0.118	0.111	0.118	0.111
	UD	0	0	0	0	0	0

Notes: We depict the estimated optimal parameters for every first- and second-mover cooperation combination using the conditional welfare model.

Prompt design

We access both GPT models via the chat completion endpoint of OpenAI's API which is publicly available. We collect data for the GPT-3.5-turbo (GPT-4) model in June 2023 (August 2023). As the API does not retain chat histories, each call to the API consisting of an unconditional and two conditional decisions, as well as two conditional beliefs about the second-mover's response, is treated as an independent observation. We maintain all default model parameters. In our prompt design, we combine the *Turing Experiment* methodology by Aher et al. (2023) with common style practices for prompting LLMs (Wei et al., 2023; Kojima et al., 2023). Specifically, we use a zero-shot prompting strategy and chain-of-thought prompting by instructing the model to think sequentially. When calling the API, we specify the role of a *user* and enter the following prompt as *content* to collect 200 independent observations each for both models. A *[blank]* indicates the text position at which the model is required to insert either A or B or an integer between 0 and 100 - depending on whether a decision or a numerical estimate of the likelihood with which the second-mover is assumed to cooperate is requested.

> Instructions: you will be randomly paired with another human. You and this other person have two possible actions available.
>
> Action A and Action B.
>
> Your earnings and those of the other person will be determined by your decision and that of the other person. There are four possibilities:
>
> If you both select action A, each earns 10 points.
> If you both select action B, each earns 30 points.
> If the other person selects action B and you select action A, the other person will earn 5 points and you will earn 50.
> In the opposite case, if you select action B and the other person selects action A, you will earn 5 points and the other person will earn 50 points.
>
> Both of you must take two types of decisions – an unconditional decision and a conditional decision.
>
> In case of unconditional decisions, you must simply decide whether you select action

A or if you select action B.

In case of conditional decisions, you can make your decision dependent on what the other person does. This means that you must make a decision for each of the following cases:

Case 1: the other person selects action A.

Case 2: the other person selects action B.

The game is played once.

Let's think step by step.

# # #

Answer 1 (choose A or B): For my unconditional decision, I will choose action *[blank]*

# # #

Answer 2 (choose A or B): For my conditional decision, if the other person has chosen action A (case 1), I will choose action *[blank]*

# # #

Answer 3 (choose A or B): For my conditional decision, if the other person has chosen action B (case 2), I will choose action *[blank]*

# # #

Answer 4 (choose integer 0-100): If my unconditional decision is action A, I predict that other person will choose Option B in *[blank]* % of the cases.

# # #

Answer 5 (choose integer 0-100): If my unconditional decision is action B, I predict that other person will choose Option B in *[blank]* % of the cases.

# # #

Derivation of the parameter ranges for ρ and σ in Charness and Rabin (2002)

The theoretical framework of the conditional welfare model, as proposed by Charness and Rabin (2002), offers a structural foundation for modeling the utility-maximizing behaviors of individuals who exhibit conditional welfare concerns. The objective function, denoted by equation (3), enables individuals to assign positive weights, represented by ρ and σ, to the payoffs of their opponents, thereby integrating efficiency and fairness motives into utility maximization. The parameters ρ and σ, akin to the weights assigned to the opponent's payoff, are subject to the following constraints: $0 \leq \rho \leq 1$, $0 \leq \sigma \leq 0.5$, and $\sigma \leq \rho$. In what follows, we derive the parameter ranges within which both ρ and σ must fall to account for the four distinct types of second-mover behaviors observed in our game.

First, note that cooperation is the optimal choice for player i as the second-mover when $U_i(C, a_j) \geq U_i(D, a_j)$. In other words, cooperation becomes the preferred action for player i when her payoff from choosing cooperation is at least as large as her payoff when opting to defect, given the first-mover's action a_j. As player i observes the choice of the first-mover, she can determine her best response, conditional on whether she observed cooperation or defection by the first-mover.

In the scenario where player i witnesses that the first-mover has cooperated, the condition $\pi_i(a_i, a_j) \geq \pi_j(a_i, a_j)$ invariably holds true. Given our game's parameterization, player i garners 30 points when reciprocating cooperation as a second-mover, compared to 50 points when defecting. In contrast, player j, as the first mover, is left with 30 points in the former case and only 5 points in the latter (cf. Figure 1). Assuming that player i prefers to cooperate when she would derive the same utility from defection, equation (3) implies that second-mover cooperation becomes the optimal choice after first-mover cooperation if and only if:

$$
\begin{aligned}
U_i(C,C) &\geq U_i(D,C) \Leftrightarrow \\
(1-\rho)\cdot\pi_i(C,C) + \rho\cdot\pi_j(C,C) &\geq (1-\rho)\cdot\pi_i(D,C) + \rho\cdot\pi_j(D,C) \Leftrightarrow \\
(1-\rho)\cdot 30 + \rho\cdot 30 &\geq (1-\rho)\cdot 50 + \rho\cdot 5 \Leftrightarrow \\
45\cdot\rho &\geq 20 \Leftrightarrow \\
\rho &\geq \frac{4}{9}
\end{aligned}
\tag{5}
$$

In the scenario where player i observes the first-mover's defection, we have $\pi_i(C, D) <$

$\pi_j(C, D)$ and $\pi_i(D, D) = \pi_j(D, D)$ (cf. Figure 1). Using equation (3), we can determine the threshold value of σ above which second-mover cooperation becomes optimal after observing that the first mover has defected:

$$\begin{aligned} U_i(C, D) &\geq U_i(D, D) \Leftrightarrow \\ (1-\sigma) \cdot \pi_i(C, C) + \sigma \cdot \pi_j(C, C) &\geq (1-\rho) \cdot \pi_i(D, C) + \rho \cdot \pi_j(D, C) \Leftrightarrow \\ (1-\sigma) \cdot 5 + \sigma \cdot 50 &\geq (1-\rho) \cdot 10 + \rho \cdot 10 \Leftrightarrow \\ 45 \cdot \sigma &\geq 5 \Leftrightarrow \\ \sigma &\geq \frac{1}{9} \end{aligned} \tag{6}$$

Therefore, second-mover cooperation becomes optimal after observing first-mover cooperation if and only if $\rho \geq \frac{4}{9}$, while second-mover cooperation becomes optimal after observing the first mover's defection if and only if $\sigma \geq \frac{1}{9}$.

In the sequential prisoner's dilemma, we classify second-mover behaviors into four categories: unconditional cooperators (UC), who always cooperate; conditional cooperators (CC), who reciprocate the first mover's choice; mismatchers (MM), who choose the opposite action to the one observed; and unconditional defectors (UD), who defect regardless of the observed action.

Based on the derivations in equations (5) and (6), we can conclude that the conditional welfare model can account for the second-mover behavior of unconditional cooperators (UC) when $\rho \geq \frac{4}{9}$ and $\sigma \geq \frac{1}{9}$, of conditional cooperators (CC) when $\rho \geq \frac{4}{9}$ and $\sigma < \frac{1}{9}$, of mismatchers (MM) when $\rho < \frac{4}{9}$ and $\sigma \geq \frac{1}{9}$, and of unconditional defectors (UD) when $\rho < \frac{4}{9}$ and $\sigma < \frac{1}{9}$. We use these parameter ranges, in conjunction with the constraints $0 \leq \rho \leq 1$, $0 \leq \sigma \leq 0.5$, and $\sigma \leq \rho$, in our grid search to determine the optimal parameter combination of (ρ, σ) for each second-mover behavior category and first-mover choice. The optimal parameter combination maximizes the number of observations explained by the conditional welfare model within each behavior category and first-mover choice, and is reported in Table 5.

LEIBNIZ INSTITUTE SAFE Sustainable Architecture for Finance in Europe

Recent Issues

No. 400	Andreas Hackethal, Philip Schnorpfeil, Michael Weber	Households' Response to the Wealth Effects of Inflation
No. 399	Raimond Maurer, Sehrish Usman	Dynamics of Life Course Family Transitions in Germany: Exploring Patterns, Process and Relationships
No. 398	Pantelis Karapanagiotis, Marius Liebald	Entity Matching with Similarity Encoding: A Supervised Learning Recommendation Framework for Linking (Big) Data
No. 397	Matteo Bagnara, Milad Goodarzi	Clustering-Based Sector Investing
No. 396	Nils Grevenbrock, Alexander Ludwig, Nawid Siassi	Homeownership Rates, Housing Policies, and Co-Residence Decisions
No. 395	Ruggero Jappelli, Loriana Pelizzon, Marti Subrahmanyam	Quantitative Easing, the Repo Market, and the Term Structure of Interest Rates
No. 394	Kevin Bauer, Oliver Hinz, Moritz von Zahn	Please Take Over: XAI, Delegation of Authority, and Domain Knowledge
No. 393	Michael Kosfeld, Zahra Sharafi	The Preference Survey Module: Evidence on Social Preferences from Tehran
No. 392	Christian Mücke	Bank Dividend Restrictions and Banks' Institutional Investors
No. 391	Carmelo Latino, Loriana Pelizzon, Max Riedel	How to Green the European Auto ABS Market? A Literature Survey
No. 390	Kamelia Kosekova, Angela Maddaloni, Melina Papoutsi, Fabiano Schivardi	Firm-Bank Relationships: A Cross-Country Comparison
No. 389	Stefan Goldbach, Philipp Harms, Axel Jochem, Volker Nitsch, Alfons J. Weichenrieder	Retained Earnings and Foreign Portfolio Ownership: Implications for the Current Account Debate
No. 388	Gill Segal, Ivan Shaliastovich	Uncertainty, Risk, and Capital Growth
No. 387	Michele Costola, Katia Vozian	Pricing Climate Transition Risk: Evidence from European Corporate CDS