ChatGPT: Usage and Limitations

Amos Azaria
School of Computer Science, Ariel University

***Abstract*—Large language models have been shown to be useful in multiple domains including conversational agents, education, and explainable AI. ChatGPT is a large language model developed by OpenAI as a conversational agent. ChatGPT was trained on data generated by humans and by receiving human feedback. This training process results in a bias toward humans' traits and preferences. In this paper, we stress multiple biases of ChatGPT, and show that its responses demonstrate many human traits. We begin by showing a very high correlation between the frequency of digits generated by ChatGPT and humans' favorite numbers, with the most frequent digit generated by ChatGPT, matching humans' most favorable number, 7. We continue by showing that ChatGPT's responses in several social experiments are much closer to those of humans' than to those of fully rational agents. Finally, we show that several cognitive biases, known in humans, are also present in ChatGPT's responses.**

***Index Terms*—Cognitive biases; Rational behavior; ChatGPT**

I. Introduction

Large Language Models (LLMs) are statistical models that use previous tokens to predict future tokens. They are trained on large amounts of data and include billions of parameters. Large language models, such as GPT-3 [11], bloom [26], and Roberta [17] have recently demonstrated their usefulness in multiple domains including conversational agents, education, explainable AI, text summarization, information retrieval, and more [9]. To fully exploit the advantages of large language models, it is crucial to understand their limitations and biases present in their output. Clearly, the characteristics of the data influence LLMs' responses, and since the majority of data these models are trained on is text written by humans, one can expect finding biases that are present in humans also in these models.

ChatGPT is a LLM developed by OpenAI as a conversational agent [2]. Not only is it trained on data generated by humans, but it also obtains sample responses that it should follow, and, using reinforcement learning [28], it is trained by explicit human feedback [23]. Therefore, finding biases that are present in humans also in ChatGPT is even more likely than it is in common LLM. In this paper, we show that this is indeed the case.

We begin by testing the distribution of digits generated by ChatGPT. Unfortunately, ChatGPT does not provide a simple or straightforward way to gain access to its distribution over digits that it samples. Therefore, we exploit a common limitation of LLMs—their inability to compute mathematical expressions. Namely, ChatGPT cannot multiply large numbers, compute a square-root of a number, compute the value of a number to a power of another number (especially fractions), and add (or subtract) a number to an irrational number (e.g., pi or e). Furthermore, ChatGPT is unaware of its limitations, and when attempting to answer a complex mathematical expression it often merely samples digits. For example, when asked "How much is 3.2 to the power of 3.3?", ChatGPT provides the following answer: "The result of 3.2^3.3 is approximately 21.73." The correct answer is approximately 46.45. We note that ChatGPT does not provide any indication of its response being a total guess, and incorrect.

Since ChatGPT tends to sample digits as its response, we query it with mathematical expressions that result in irrational numbers. Specifically, in response to our queries all digits should appear equally as frequent (i.e., each of the digits should appear in the correct answer approximately 10% of the time). Any distribution that is not uniform implies a bias in ChatGPT's responses. We analyze the distribution over the digits generated by ChatGPT and compare them with a survey conducted by Bellos [7], in which participants were asked for their most favorable number. We find a very high correlation between the two distributions. Furthermore, the most frequent digit generated by ChatGPT matches humans' most favorable number, 7. Similarly, the least frequent digit generated by ChatGPT, matches humans' least favorable number, 1. We find this result intriguing, despite it seeming sensible in hindsight.

Next, we examine ChatGPT's responses to well studied social experiments. Specifically, we consider the prisoner's dilemma, the ultimatum game, and the trust game. We show that in all three games ChatGPT's responses are similar to typical human responses and are very different from fully rational behavior.

Finally, we study ChatGPT's responses to queries that attempt to reveal cognitive biases present in humans. Namely, we show that like humans, ChatGPT is highly prone to the framing effect [3] and, like humans, ChatGPT is prone to the intuition bias [15], availability heuristic [29], and iterative reasoning [27].

II. Experimental Settings

The overall hypothesis tested in this paper is that ChatGPT encounters cognitive biases known in humans, does not adhere to perfectly rational strategies (Nash equilibrium), and has a preference to sample digits preferred by humans. The experiments were selected from well known experiments conducted with humans that were easy to conduct using text only communication (not requiring any diagrams or physical actions).

All queries to ChatGPT reported in this paper were submitted in a new conversation window, i.e., without any previous context, unless explicitly reported otherwise as a part of a conversation. This practice is important and required to eliminate any interference between different queries.

We report all queries and responses obtained, except interactions that resulted in indecisive responses (e.g., "I cannot answer the question..." or "The answer depends on various parameters...", etc.). In case we received indecisive responses, the prompt was modified to eliminate them. We did not run any additional experiments aimed at proving our hypothesis, which were not reported in this paper. While we conducted more than a single query to ensure consistency, we only provide the exact text of a single query, but note that other queries resulted in very similar responses. An exception to this rule is in the availability heuristic, in which, as we later describe, we first composed multiple queries that did not include an explanation of how the answer was reached, and thus could not be analyzed. We therefore edited the prompt, and added "Explain why."

All experiments were conducted with version Dec 15, 2022 of ChatGPT. We note that the bat and ball problem was later tested in the March 14, 2023 version, and ChatGPT answered correctly[1]. Interestingly, when provided the bat and ball question along with ChatGPT's Dec 15 answer (reported in this paper), and prompted to explain why the answer is incorrect, ChatGPT March 14 responded (incorrectly) that the previous answer is correct and does not require any correction.

III. ChatGPT's Digit Preferences

In this section we expose a bias in the way ChatGPT samples digits when attempting to compute a value. We first describe our experimental methodology, and then describe our results.

A. Background

Since ChatGPT is trained also on data gathered from the real world, one might expect that digits generated by it may adhere to distributions often found in real world data. We thus discuss Zipf's [13], [16], [24] and Benfold's [10], [18], [21] laws, which intend to describe such data. According to Zipf's law a word's frequency is inversely proportional to its frequency's rank. Zipf's law was also shown to hold in corporation sizes, income rankings, and other human created systems [24].

According to Benford's law, the probability of observing a number with a leading digit is higher the smaller the digit is. More specifically, the probability of observing a leading digit d is given by $log_{10}\frac{d+1}{d}$. That is, approximately 30% of the numbers should have a leading '1', and only 4.6% of the numbers should have a leading '9'. Benford's law was extended to non-leading digits, and preserves the property that lower digits are more likely to be observed than higher digits. This phenomenon was confirmed in many different datasets including U.S. county population, U.S county land area, volume of trades, and fundamental constants [18]. Benford's law was even suggested as a method to detect false tax reports [21]. Unfortunately, analysis of digits generated by ChatGPT shows no relation to these two laws. In fact, as we later show, digit '1' appears least frequently in the data gathered from ChatGPT.

[1] It is quite possible that ChatGPT was corrected as a result of a preliminary version of this paper appearing in pre-print.

How much is pi * 4.26 with 200 digits?
Can you write pi - 1.01 with 100 digits?
Can you give me pi divided by 1.95236 with 400 digits?
Give me 300 digits of pi * 1.017.
How much is pi + 1.23 with 100 digits?
pi divided by 2.36 with 400 digits.
Can you write the value of e/4.325 with 300 digits?
Can you write the value of e/2.8 with 100 digits?
Can you give me e divided by 3.956 with 400 digits precision?
Can you give me e divided by 3.176 with 400 digits precision?
Can you write the value of the third root of 2811 with 300 digits?
Third root of 2811 with 300 digits.

TABLE I
EXAMPLES OF QUERIES POSTED TO CHATGPT, WHICH CAUSED IT TO GENERATE (INCORRECT) NUMBERS WITH AT LEAST ONE HUNDRED DIGITS.

B. Experimental Methodology

In order to gather digits generated from ChatGPT, we posed queries resulting with irrational numbers. However, these queries could not be anything that might have been present in ChatGPT training data (e.g., pi, 2π or even π^2), as ChatGPT is likely to provide the correct answer. Furthermore, we requested ChatGPT to provide between 100 to 400 precision digits. In practice, ChatGPT tended to provide more digits than requested. Table I provides examples of some queries we posted to ChatGPT. ChatGPT's responses varied, but in most cases, it provided a number with at least one hundred digits. Hereunder is a typical response. We note that the actual value of pi*4.26 begins with: 13.3831; we also note that in this example ChatGPT provided less than 150 digits, while being asked for 200 digits.

USER: *How much is pi * 4.26 with 200 digits?*

CHATGPT: *To calculate the product of pi and 4.26 with 200 digits of precision, you can use a calculator or computer program that can handle very large numbers with many decimal places. Here is the result:*

13.5342917352885175958532670577032166737373559322094377440307914394040274068794038702809584247937705918867463979369984812822051481824270790282700 7

*Note that this result is only accurate to the 200 digits shown. The actual value of pi * 4.26 has an infinite number of decimal places, and no exact representation of it can be given with a finite number of digits.*

The first five digits of the response were removed from our analysis. This is because they may be related to the actual query and may be approximately correct. For an unclear reason, some of ChatGPT's responses ended with a sequence of 4's, which could be hundreds of digits long. In addition, some responses included the sequence "123456789" multiple times. Some responses also included the true sequence of digits of π or e. Consequently, all these sequences were removed from the analysis provided in this paper.

C. Results

In total, we gathered $10,811$ digits. Table II presents the frequency of each of the 10 digits in the data. Clearly, 7 is the most frequent digit.

Digit	Frequency	percent [1-9]
0	1,049	-
1	865	8.73%
2	976	9.85%
3	1,113	11.23%
4	1,121	11.31%
5	1,064	10.74%
6	1,116	11.26%
7	**1,316**	**13.28%**
8	1,222	12.33%
9	1,110	11.20%

TABLE II
FREQUENCY OF EACH OF THE 10 DIGITS IN THE NUMBERS GENERATED BY CHATGPT.

We compare these results with the results obtained by Bellos [7], when asking people for their favorite number (see Table III). We focus on Bellos' results of the numbers 1-9. A comparison between the tables shows that the number 7 is the most frequent in both, and the number 1 is least frequent in both. High similarity is also seen in frequencies of the other numbers. Overall, the Pearson correlation coefficient between the frequencies is **0.893**, which is considered a very high correlation.

number	frequency	percent
1	358	2.73%
2	1011	7.72%
3	2248	17.16%
4	1694	12.93%
5	1544	11.79%
6	1015	7.75%
7	**2912**	**22.23%**
8	2025	15.46%
9	1438	10.98%

TABLE III
FREQUENCY OF HUMAN'S FAVORITE NUMBERS (ACCORDING TO SURVEY BY BELLOS, 2015).

We note that any attempt to query ChatGPT for its favorite number resulted in a generic response with no definitive answer. Furthermore, questions related to humans' favorite number did not result in any useful response.

IV. SOCIAL EXPERIMENTS

In this section we describe the results of running social game experiments with ChatGPT. Social games are multiplayer games, in which each player must select her action and the outcome to all players may depend on the actions of all players. Depending on the game, the actions may either be taken simultaneously, or sequentially. Each game has a Nash equilibrium, which determines the actions and strategies of *rational* players [19]. We note that rational behavior is not necessarily better, but there are some reasons to expect a computer to exhibit more rational behavior. Specifically, it has been shown that programs written by humans exhibit more rational behavior than the humans who wrote them [6]. However, as we show in this section, ChatGPT's responses are very similar to those of human players and are very different from the expected behavior of a fully rational agent. We note that, at times, ChatGPT's responses indicate that it does not completely understand all games. The description of each of the games examined appears in the query to ChatGPT by the user.

A. Prisoner's Dilemma

We begin with the well studied prisoner's dilemma, which is a two player game played simultaneously. The actions and outcomes of each of the players are depicted in Table IV.

	Keep Quiet	Speak
Keep Quiet	-1,-1	-30,0
Speak	0,-30	-5,-5

TABLE IV
THE NORMAL FORM REPRESENTATION OF THE PRISONER'S DILEMMA. THE FIRST NUMBER IN EACH CELL INDICATES THE OUTCOME TO THE ROW PLAYER, AND THE SECOND NUMBER INDICATES THE OUTCOME TO THE COLUMN PLAYER.

USER: *If you were caught with a friend, and if you both keep quiet you will be in prison for one year, if you both speak you will both serve 5 years. If one speaks and the other keeps quiet, the one who speaks will be out immediately and the second one will serve for 30 years. Each of you makes his/her decision separately, you cannot influence your friends decision. Which would you pick? Keep quiet or speak? You must answer only one of the two options.*

CHATGPT: *In this situation, it would be best to keep quiet. While serving time in prison for one year is not ideal, it is better than serving 30 years. Additionally, even if your friend decides to speak, you will only serve 5 years by remaining quiet, which is still a better outcome than serving 30 years. It is important to carefully consider all potential consequences before making a decision in a situation like this.*

The rational behavior in the Prisoner's Dilemma is for both players to speak. This is because, regardless of the action of the other player, each player is better-off when speaking. In fact, the only equilibrium in the game is when both players speak. However, it was shown that when faced with the Prisoner's Dilemma experiment many people opt to keep quiet [5].

B. Ultimatum Game

The ultimatum game is a well studied game, in which two players are required to split some amount. The first player (offerer) proposes a split, and the second player (responder) may either accept or reject it.

USER: *Suppose you are given 1000 dollars to split between you and someone else. You can decide to split it however you want, but if the other person rejects your suggestion, you both end with nothing. How would you propose to split the money?*

CHATGPT: *In this situation, it would be best to propose a split that is fair and reasonable to both parties. A good way to do this would be to propose a 50/50 split, where each person receives 500 dollars. This split is fair because it divides the money equally between the two parties. Additionally, it is reasonable because it is unlikely that the other person would reject a fair and equal split. Of course, the final decision on how to split the money will ultimately depend on* *the specific circumstances and the preferences of the other person involved.*

Next, we examine ChatGPT's response when being the responder in the ultimatum game.

USER: *Suppose someone else is given 1000 dollars to split between him/her and you. They can offer you any amount of the 1000 dollars, but if you decide to decline the offer, you will both get nothing. Either way there is no possibility for another offer. Suppose that the other player offered you 100 dollars, would you accept or decline? Your answer must be only one of the two. That is, you must use the word accept or decline in your answer.*

CHATGPT: *I would accept the offer.*

However, when offered one dollar less (i.e., \$99), ChatGPT rejects the offer. This experiment was repeated with many different values of offers to ChatGPT (\$10, \$50, \$90, \$200, \$300, \$400, \$500, \$600), and ChatGPT only rejected offers less than \$100, and accepted any offer of \$100 or greater. The experiment was also repeated with a total of \$100, and ChatGPT rejected offers less than \$10.

The rational behavior in the ultimatum game states that the responder should accept any amount greater than 0, and thus, the offeror should take almost all the money for themselves. However, in practice, the majority of human proposers offer between 40% to 50% of the total sum, and most people reject low offers [4], [22]. Indeed, ChatGPT's response is a typical human response and very different from that of a fully rational agent.

C. Trust Game

The trust game is a two player game with sequential turns. Like the ultimatum game the first player (investor) is given an amount and may decide to invest some of it. Any amount invested is multiplied by 3, and the second player (trustee) may reciprocate any amount with the first player.

USER: *Suppose you are given 100 dollars. You can give any amount of money to a second person. The money you decide to pass is multiplied by 3, and then given to the second person. The second person may then reciprocate and give you back as much money as s/he decides. How much money would you give the second person? Your answer must be a number between 0 and 100.*

CHATGPT: *If you are given \$100, you can give any amount of money to a second person, as long as it is between 0 and 100 dollars. The amount of money you give to the second person will be multiplied by 3, so if you give the second person 10 dollars, s/he will receive 30 dollars. It is up to the second person to decide whether to reciprocate and give you back some of the money.*

In this situation, the optimal decision for you would be to give the second person 50 dollars. This is because the money you give will be multiplied by 3, so the second person will receive 150 dollars. This is the maximum amount of money the second person can receive, so if the second person chooses to reciprocate, s/he will likely give you back the most money possible.

Overall, the optimal amount of money to give the second person in this situation is 50 dollars.

The rational behavior in the trust game is for the trustee not to return any money to the investor, and thus, for the investor not to pass any money to the trustee [6], [8]. However, in practice, human investors invest around half their money, and the trustees return more than they have received [31].

V. Cognitive Biases in ChatGPT

In this section we survey several cognitive biases studied in humans and demonstrate them in ChatGPT.

A. Intuition Bias

According to the intuition bias, people prefer to base their decisions more on quick intuitive perceptions (System 1) than on deliberate thinking (System 2) [15]. A common experiment exposing the intuition bias is the bat and ball problem [12].

USER: *A bat and ball cost together \$1.10. The bat costs \$1.00 more than the ball. How much does the ball cost?*

CHATGPT: *The ball costs \$0.10. The bat costs \$1.00 more than the ball, so it costs \$0.10 + \$1.00 = \$1.10. Together, the bat and ball cost \$1.10.*

Indeed, just as ChatGPT, most people reply that the ball costs \$0.10, which is clearly incorrect. We note that ChatGPT demonstrates its consistency bias (once again) when claiming that adding its ball price to its bat price still results in \$1.10. Interestingly, when prompted that its response is incorrect, ChatGPT could correct its reasoning. A similar study was conducted on humans by [14].

USER: *That's incorrect, try again.*

CHATGPT: *I apologize for the mistake. Here is the correct solution: The bat costs \$1.00 more than the ball, so if the ball costs "x" dollars, the bat costs x+1 dollars. The cost of the bat and ball together is \$1.10, so: x + (x+1) = 1.10 2x + 1 = 1.10 2x = 0.10 x = 0.05 The ball costs \$0.05.*

B. Availability Heuristic

The availability heuristic bias states that when attempting to assess the frequency of an event, people use the similarity to events that they are familiar with and their frequency [29]. Therefore, when asked if the letter 'l' is more common as the first letter in a word or as the third letter, people attempt to recall words beginning with 'l' and words that have 'l' as the third letter. Since it is a lower cognitive load to recall words beginning with a letter, people conclude that the letter is more common as the first letter, regardless of the true distribution [29]. We note that [29], used the letters K,L,N,R, and V. Of those letters we picked 'L', since recent and more modern frequency counts show that it appears more times as a third letter (unlike the letter 'K')[2]. Interestingly, when explicitly asked about the frequency of letters in English, ChatGPT tends to provide short and confident responses, and at times (correctly) suggests that a letter is more common as being third. However, when the query requires ChatGPT to reason about its response, it clearly demonstrates the availability heuristic bias. For example:

USER: *Do you believe that there are more words in English that begin with 'l' or that have 'l' as their third letter? Explain why.*

CHATGPT: *It is likely that there are more words in English that begin with the letter 'l' than those that have 'l' as their third letter. This is because the letter 'l' is a relatively common*

[2]https://github.com/WilliamRitson/letter-positon-frequency/blob/master/results/corpus-analysis.csv

initial letter in English words, while it is less common as a third letter. Additionally, many common English words begin with 'l', such as "look," "love," "like," and "life." This is likely due to the structure of the English language and the way that words are formed in the language.

In practice, the letter 'l' is much more likely as the third letter. Hereunder, is a similar example related to first names.

USER: *Do you believe that there are more first names of people that begin with 'e' or that end with 'e'? Explain why.*

CHATGPT: *As a language model, I do not have the ability to believe or not believe. However, I can provide statistics on the frequency of first names beginning with the letter 'e' or ending with the letter 'e' based on the data I was trained on. Based on that data, it appears that there are more first names of people that begin with the letter 'e' than end with the letter 'e'. This is likely because many common names, such as "Emily" and "Elizabeth," begin with the letter 'e', whereas fewer common names, such as "Rylee" and "Jaycee," end with the letter 'e'.*

Once again, in practice, many more names end with 'e' than begin with it[3]. Additionally, common names such as "Theodore" and "Mike" end with 'e'. We note that ChatGPT was queried multiple times and provided similar responses.

C. Level-K Iterative Reasoning

The level-k iterative reasoning bias states that when required to reason about other people's actions, people go through an iterative process [27]. They first assume that all other players act randomly (level-0), react to that (level-1), and then assume that all players play according to level-1, and then they react to that. This process repeats until level-K is reached, where K indicates the level of the specific player. The following guess 2/3 of the average game is a guessing game that is often used to study the level-k bias [1]. In this game a player is required to guess 2/3 of the average of the number picked by all players. ChatGPT's response indicates a level-1 thinker.

USER: *Suppose 1000 people are asked to pick a number between 0 and 100 that is 2/3 of the mean of the numbers that everyone picks. Everyone submits their number confidentially, and cannot see what others have picked. What number would you pick?*

CHATGPT: *To determine the number that is 2/3 of the mean of the numbers that everyone picks, we first need to find the mean of the numbers that are submitted. If 1000 people each pick a number between 0 and 100, the sum of all the numbers they pick would be 1000 * 50 = 50000, since the mean of a set of numbers is equal to the sum of all the numbers divided by the number of items in the set.*

*Thus, the mean of the numbers that are submitted would be 50000 / 1000 = 50. To find the number that is 2/3 of the mean, we would multiply the mean by 2/3 to get 50 * (2/3) = 50 * (0.6666) = 33.33. Since the numbers must be integers between 1 and 100, the closest number to 33.33 that is within this range would be 33.*

Therefore, if 1000 people are asked to pick a number between 1 and 100 that is 2/3 of the mean of the numbers that everyone picks, the number that I would pick would be 33.

We note that in the single Nash equilibrium of this game all players select 0. Clearly, this is not what happens in practice, and therefore, guessing 0 is incorrect. The mean guess in experiments conducted with humans is usually slightly above 30 (resulting in the correct answer being slightly above 20). Guessing 33 is quite common among humans and indicates a level-1 thinker, who (incorrectly) assumes that all other players act randomly and reacts to that [1], [20], [25].

D. Framing

Another major cognitive bias present in humans is the effect of framing. That is, the way different options are presented highly impacts human decision making [3], [30]. The following example is a modified version of the Asian disease experiment conducted by Kahneman and Tversky's [30].

USER: *Assume that there are 601 chickens with a fatal disease. The first medication will save the lives of 200 chickens (killing the rest), and the second medication will kill 400 chickens (saving the rest). Which medication should be given?*

CHATGPT: *It would be best to give the first medication, as it will save the lives of the most chickens (200 out of 601). In this case, it is more important to save as many lives as possible, even if that means losing some.*

Clearly, the second medication will save 201 chickens, which is 1 more than the first medication, but since it focuses more on the number of chickens that it kills (rather than saves), ChatGPT demonstrates its framing bias and prefers the first medication. Interestingly, when being queried about a fatal disease in humans, as originally proposed by [30], ChatGPT refuses to provide any answer, and claims that providing such an answer by an AI is unethical. Furthermore, it is unwilling to make a choice even when prompted that if it does not make any choice all sick people will die.

CHATGPT: *I understand that in this scenario, there are difficult choices that need to be made, and it is a tragic situation. However, as an AI, I am not qualified to make a decision on which medication should be given in this situation, as it involves complex ethical considerations and the potential loss of human life...*

Next, we show the framing effect on ChatGPT in the ultimatum game when taking the role of the responder. Recall that ChatGPT accepted any offer of $100 or greater. We made a very slight modification to the description of the game and added the word "only" to the offer (i.e., "Suppose that the other player offered you **only** 100 dollars"). Quite surprisingly, this caused ChatGPT to reject all offers, even when offered $600.

USER: *Suppose someone else is given 1000 dollars ... Suppose that the other player offered you only 600 dollars ...*

CHATGPT: *I would decline the offer of 600 dollars. In this situation, it is better to decline the offer and receive nothing than to accept an unfair split of the money. The other player's offer does not reflect a fair division of the money.*

Recall that the perfectly rational behavior in this game (the Nash equilibrium) is to accept any offer. Therefore, while ChatGPT's rejection of 60% is not typical human behavior, it does clearly demonstrate that ChatGPT does not adhere to the Nash equilibrium and, like humans, is (heavily) prone to the framing effect.

[3] https://www.babycenter.com/babyNamerSearch.htm

VI. Conclusions and Discussion

In this paper, we show that despite ChatGPT being a computer, its responses are more similar to humans than to fully rational agents, and that it seems to encounter the same cognitive biases known to be present in humans. In the prisoner's dilemma, ultimatum game, and trust game, following human traits may, in fact, be better for the society than being completely self-interested and following the Nash equilibrium. Namely, ChatGPT elects to cooperate with the second player in the prisoner's dilemma and offers a fair split of its money in the ultimatum game and in the trust game. However, possessing the cognitive biases discussed in this paper, including the tendency to justify its incorrect previous responses, and favoring some digits over others, is undesirable behavior.

We believe that ChatGPT's responses are closer to those of humans since its training data is primarily based on human generated text and its training methodology highly relies on human input. However, this behavior may be more inherent to LLMs and learning by language. Nevertheless, there are several reasons to believe that a large language model's output should actually be more accurate or rational than that of the average human. Namely, being trained on massive amounts of human generated data may expose the wisdom of the crowd [32]. Secondly, the training data is likely to be acquired from sources more trustworthy than the average human everyday correspondence. Thirdly, LLMs are developed using computers, mathematical principles and operations, which are highly associated with rational behavior; however, it is not expected that these principles will diffuse into the model responses, just as it is not expected that humans will understand how a brain works, just because they use it to reason.

An important corollary from this work is that the presence of these biases in ChatGPT can make it difficult to identify when it produces incorrect facts, because people would be subject to the same mistakes and would not be able to easily spot them. In future work we intend to study techniques used to mitigate cognitive biases in humans and consider methods for applying similar techniques to the training or prompting of large language models.

VII. Acknowlegments

This research has been partly supported by the Ministry of Science, Technology & Space (MOST), Israel.

References

[1] M. Agranov, A. Caplin, and C. Tergiman. Naive play and the process of choice in guessing games. *Journal of the Economic Science Association*, 1(2):146–157, 2015.

[2] M. Allouch, A. Azaria, and R. Azoulay. Conversational agents: Goals, technologies, vision and challenges. *Sensors*, 21(24):8448, 2021.

[3] O. Amir, D. Ariely, A. Cooke, D. Dunning, N. Epley, U. Gneezy, B. Koszegi, D. Lichtenstein, N. Mazar, S. Mullainathan, D. Prelect, E. Shafir, and J. Silva. Psychology, behavioral economics, and public policy. *Marketing letters*, 16(3):443–454, 2005.

[4] O. Amir, D. G. Rand, and Y. K. Gal. Economic games on the internet: The effect of $1 stakes. *PloS one*, 7(2):e31461, 2012.

[5] J. Andreoni and J. H. Miller. Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *The economic journal*, 103(418):570–585, 1993.

[6] A. Azaria, A. Richardson, and A. Rosenfeld. Autonomous agents and human cultures in the trust–revenge game. *Autonomous Agents and Multi-Agent Systems*, 30(3):486–505, 2016.

[7] A. Bellos. *The grapes of math: How life reflects numbers and numbers reflect life*. Simon and Schuster, 2015.

[8] C. Buntain, A. Azaria, and S. Kraus. Leveraging fee-based, imperfect advisors in human-agent games of trust. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

[9] R. Dale. Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118, 2021.

[10] C. Durtschi, W. Hillison, and C. Pacini. The effective use of benford's law to assist in detecting fraud in accounting data. *Journal of forensic accounting*, 5(1):17–34, 2004.

[11] L. Floridi and M. Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, 2020.

[12] S. Frederick. Cognitive reflection and decision making. *Journal of Economic perspectives*, 19(4):25–42, 2005.

[13] Y. Fujiwara. Zipf law in firms bankruptcy. *Physica A: Statistical Mechanics and its Applications*, 337(1-2):219–230, 2004.

[14] E. M. Janssen, M. Raoelison, and W. de Neys. "you're wrong!": The impact of accuracy feedback on the bat-and-ball problem. *Acta psychologica*, 206:103042, 2020.

[15] D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[16] W. Li. Zipf's law everywhere. *Glottometrics*, 5(2002):14–21, 2002.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[18] S. J. Miller. *Benford's law*. Princeton University Press, 2015.

[19] R. B. Myerson. Nash equilibrium and the history of economic theory. *Journal of Economic Literature*, 37(3):1067–1082, 1999.

[20] R. Nagel. Unraveling in guessing games: An experimental study. *The American economic review*, 85(5):1313–1326, 1995.

[21] M. J. Nigrini. A taxpayer compliance application of benford's law. *The Journal of the American Taxation Association*, 18(1):72, 1996.

[22] M. A. Nowak, K. M. Page, and K. Sigmund. Fairness versus reason in the ultimatum game. *Science*, 289(5485):1773–1775, 2000.

[23] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[24] S. T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130, 2014.

[25] A. Rubinstein. Instinctive and cognitive reasoning: A study of response times. *The Economic Journal*, 117(523):1243–1259, 2007.

[26] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, and T. J. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[27] D. O. Stahl II and P. W. Wilson. Experimental evidence on players' models of other players. *Journal of economic behavior & organization*, 25(3):309–327, 1994.

[28] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[29] A. Tversky and D. Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973.

[30] A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. In *Behavioral decision making*, pages 25–41. Springer, 1985.

[31] H. Tzieropoulos. The trust game in neuroscience: a short review. *Social neuroscience*, 8(5):407–416, 2013.

[32] S. K. M. Yi, M. Steyvers, M. D. Lee, and M. J. Dry. The wisdom of the crowd in combinatorial problems. *Cognitive science*, 36(3):452–470, 2012.