Hindawi
Human Behavior and Emerging Technologies
Volume 2024, Article ID 7115633, 9 pages
https://doi.org/10.1155/2024/7115633

Research Article

The Self-Perception and Political Biases of ChatGPT

Jérôme Rutinowski[1], **Sven Franke**[1], **Jan Endendyk**[1], **Ina Dormuth**[2], **Moritz Roidl**[1], **and Markus Pauly**[2,3]

[1]*Chair of Material Handling and Warehousing, TU Dortmund University, Dortmund, Germany*
[2]*Chair of Mathematical Statistics and Applications in Industry, TU Dortmund University, Dortmund, Germany*
[3]*Research Center Trustworthy Data Science and Security, UA Ruhr, Dortmund, Germany*

Correspondence should be addressed to Jérôme Rutinowski; jerome.rutinowski@tu-dortmund.de

Received 19 October 2023; Revised 4 December 2023; Accepted 6 January 2024; Published 22 January 2024

Academic Editor: Stephen Gbenga Fashoto

This contribution analyzes the self-perception and political biases of OpenAI's Large Language Model ChatGPT. Considering the first small-scale reports and studies that have emerged, claiming that ChatGPT is politically biased towards progressive and libertarian points of view, this contribution is aimed at providing further clarity on this subject. Although the concept of political bias and affiliation is hard to define, lacking an agreed-upon measure for its quantification, this contribution attempts to examine this issue by having ChatGPT respond to questions on commonly used measures of political bias. In addition, further measures for personality traits that have previously been linked to political affiliations were examined. More specifically, ChatGPT was asked to answer the questions posed by the political compass test as well as similar questionnaires that are specific to the respective politics of the G7 member states. These eight tests were repeated ten times each and indicate that ChatGPT seems to hold a bias towards progressive views. The political compass test revealed a bias towards progressive and libertarian views, supporting the claims of prior research. The political questionnaires for the G7 member states indicated a bias towards progressive views but no significant bias between authoritarian and libertarian views, contradicting the findings of prior reports. In addition, ChatGPT's Big Five personality traits were tested using the OCEAN test, and its personality type was queried using the Myers-Briggs Type Indicator (MBTI) test. Finally, the maliciousness of ChatGPT was evaluated using the Dark Factor test. These three tests were also repeated ten times each, revealing that ChatGPT perceives itself as highly open and agreeable, has the Myers-Briggs personality type ENFJ, and is among the test-takers with the least pronounced dark traits.

1. Introduction

Recently, Large Language Models (LLMs) have gained tremendous amounts of attention from experts as well as the general public. A notable example of one such model is OpenAI's ChatGPT (GPT being an acronym for Generative Pretrained Transformer). ChatGPT is a model that generates text responses when a user provides it with a prompt. It is an LLM that was fine-tuned based on a training process that takes human feedback into account (reinforcement learning from human feedback (RLHF)). Currently, ChatGPT is open-access (version 3.5 is free to use, while version 4 is available as a subscription service) but not open-source. Due to this, users can only make assumptions as to why it behaves the way it does and what data it might have been trained on, with the developers claiming that it was trained on "[...] vast amounts of data from the internet written by humans, including conversations [...]" [1]. While it receives a lot of positive acclaim and often seems to work as intended, prominent figures such as Yann LeCun and Yoshua Bengio have criticized it publicly for various reasons, one of them being, that LLMs might not be the right approach towards AGI (artificial general intelligence) [2, 3]. Other points of contention include and are not limited to the issues of hallucination [4], gender, and language biases [5, 6]. Another reason why users have been criticizing the model is its supposed bias towards progressive and libertarian views, claiming that an AI model should not hold such

biases [7]. The existence of a model's biases, if confirmed, would be of interest for the study of artificial intelligence ethics, since such biases could entail major implications for future policy-making and societal developments.

In this work, we investigate these claims and study ChatGPT's political biases. Although prior work concerning the biases of other types of machine learning models exists (such as image classification models [8] and natural language processing [9] or the use of machine learning to detect biases [10]), at the time of writing, no thoroughly researched scientific publications on the political biases of ChatGPT or other LLMs could be encountered. We additionally investigate whether ChatGPT's "self-perception" is such that it can be attributed to personality traits based on commonly used psychological assessments. We subsequently investigate whether there is a relationship between personality traits and ChatGPT's political biases. What we coin in this paper as self-perception and personality traits are of course only the aggregation of ChatGPT's responses to the herein used psychological assessments. This work is intended to kick-start a discussion on ChatGPT's biases and perceived "personality traits" in the scientific community. In the following section, we discuss the relevant literature related to this contribution. Subsequently, we present our methodology and then finally analyze the results of our experiments and draw a conclusion based on them.

2. Related Work

This section provides the reader with a brief insight into the workings of ChatGPT's functioning. It also presents a set of measures for political biases and personality traits and how these were already applied to ChatGPT in previous publications.

2.1. Large Language Models. The term "Large Language Model" is an umbrella term for language (generation) neural network architectures that are trained on large amounts of unlabeled data, e.g., as self-supervised Pretrained Foundation Models (PFMs) [11, 12]. For OpenAI's ChatGPT, for instance, this results in a model with a total of 1.5 billion hyperparameters for ChatGPT-2, 175 billion hyperparameters for ChatGPT-3, and a currently undisclosed amount of hyperparameters for ChatGPT-4. What is known is that ChatGPT uses a transformer architecture, which is an architecture that was developed as an alternative to recurrent neural networks by Google and the University of Toronto [13]. Transformers use a typical encoder and decoder architecture that can parse sequence data. The key features of transformers are their positional encoding and self-attention functionalities, enabling them to reference and take into account preceding information and prompts.

Roughly speaking, generative language models have two main tasks while engaging with a user. First, they need to understand the user's prompts correctly. Subsequently, they need to generate a response that reads as natural language and is relevant to the user's prior input. To fulfill this task, three main steps generally need to be taken. First, a generative pretraining has to take place. During this step, the language model is fed raw text that would commonly have been scraped from the web. Based on this text that can be understood as a set of ordered strings $x_1, \cdots, x_n$, a probability for the potential subsequent strings x_{n+1} is to be calculated. The probabilities per string are to be estimated such that the model's prediction P is accurate (see [14] for details). The prediction is made by weighing the words in the model's vocabulary based on the probability of them being part of the preceding word sequence. Next, a supervised fine-tuning step takes place, in which experiments such as natural language inference, question answering, semantic similarity, and text classification are performed in a supervised manner [15]. Finally, a reinforcement-learning step with human feedback adds a third layer of complexity and accuracy to the model's performance.

The training data for a model that is supposed to be input-agnostic needs to be diverse. OpenAI faced the challenge that web scrapers that were available at the time also scraped low-quality content, which lowered the model's output quality [14]. Therefore, OpenAI developed its own web scraper in order to only scrape web content that had a priori been curated by humans [14].

ChatGPT's major competition is represented by Google's BERT (Bidirectional Encoder Representations from Transformers) and BARD (Biological Application Resource Discovery) [16] as well as Meta's RoBERTa (Robustly Optimized BERT) [12, 17] and LLaMA [18, 19]. However, BERT and its variations only use encoders and no decoders and therefore cannot be used for data generation, e.g., by accepting user prompts. ChatGPT, in contrast, is not bound by this limitation.

2.2. Political Biases and Personality Assessments. Different tests and questionnaires that try to gauge an individual's political orientation based on a set of questions covering a variety of political subjects have been developed and standardized over the past decades [20]. These questionnaires usually let the user respond with "yes" or "no" or let them express their agreement on a Likert scale (e.g., from "strongly agree" to "strongly disagree," with some options in between). Based on the user's responses, the questionnaire might recommend a political party, make a statement on the user's political ideology, or pinpoint the user's position on a political scale. One such scale is the *political compass*, developed, in this format, by political journalist Wayne Brittenden [21], which has two axes, the social axis and the economic axis. Along these axes, the user is assigned to one of the four quadrants (libertarian left, libertarian right, authoritarian left, and authoritarian right), based on the *Nolan Chart* [22]. In the context of this work, these quadrants are understood as first described in [22], meaning that the support for individual freedom increases along the authoritarian-libertarian axis, and the support for economic freedom increases along the left-right axis. The political compass test attempts to ask questions that are not specific to a single culture or country and claims to have been inspired by the works of Wilhelm Reich, Hans Eysenck, and Theodor Adorno. The concept of a two-axis compass has been studied and discussed in relevant literature

[23–25] and is a popular way to quickly assess an individual's overall political views. While using a two-axis scale already permits a more complex evaluation of something as intangible as political affiliation, than a scale using only an axis from left to right would, it still is limited in its accuracy. This limitation is in part due to the fact that a Likert scale is discrete and that the results hinge on the set of questions that are asked.

In addition, it is important to note that a measure, such as the axes of a political compass test, and the underlying, highly complex concept of political affiliation or bias are not to be understood as being identical. The construct of political bias is quite difficult to define and there are no universally agreed-upon measures for this task, which is why the herein-used tests are merely a first attempt at abstractly examining political biases. Finally, the measure itself might hold biases and is to be appreciated in the context of its time.

A test that has a set of more specific questions for each country is the *political affiliation test* from *iSideWith* [26]. For this test, questionnaires belonging to multiple countries can be selected and hold a specific set of questions for that respective country. The questions might overlap between countries on more global topics such as foreign policy but also include topics that are solely of relevance to the country's domestic politics.

Besides investigating the political views of ChatGPT, we are also interested in evaluating its self-perceived personality traits. Again, there exist plenty of questionnaires that assess the personality of humans. Many such tests would be suitable for the experiments conducted in this work; however, applying a multitude of tests is also very laborious. For this reason, we apply three of such tests to ChatGPT, chosen due to them being well-established and measuring different aspects of an individual's personality. The first test is the *Big Five personality test* which is based on five personality traits that were determined to be crucial by psychologists at the time [27] and is available online [28]. These five personality traits are openness, conscientiousness, extraversion, agreeableness, and neuroticism, which is why the test is also known under the acronym *OCEAN test*. This test is still in use today, and the personality traits measured by it seem to impact diverse aspects of a person's life, including their political leanings [29].

There are numerous studies in which the relationship between the Big Five personality traits and other attributes are examined, for instance, academic performance [30]. In the present paper, we are particularly interested in ChatGPT's political biases. In this regard, the relevant literature indicates that pronounced openness and agreeableness personality traits correlate with self-reported affiliation with progressive views (e.g., in a study conducted by [29] with $n = 12{,}472$ an increase by two standard deviations in agreeableness was shown to have a 0.02 correlation with progressive views).

Another well-known test on personality types is the *Myers-Briggs Type Indicator (MBTI)* [31]. The MBTI categorizes test-takers into one of sixteen personality types depending on their energizing (extraversion vs. introversion), attention (intuition vs. sensing), deciding (thinking vs. feeling), and living (perception vs. judgment) preferences. This test is widely used in interdisciplinary research [32, 33], although it is criticized by experts in psychology at the same time [34]. The MBTI is of interest for this work since prior research has been conducted on its interplay with political views. According to [35, 36], a pronounced judgment trait would be correlated with conservative views ($n = 88$ and $n = 101$ (with $p = 0.06$ and $\alpha = 0.1$), respectively). The test is freely available online [37].

A more recent development in psychological assessments is the *Dark Factor test* [38, 39]. The Dark Factor or Dark Score gauges the test-takers' tendency to maximize their individual well-being while disregarding the well-being of others. This might go as far as going out of their way to hurt others and to find justifications for such behavior. A high Dark Score therefore indicates the ruthlessness with which an individual might pursue their personal goals while neglecting the detrimental effects that their actions might have on others. The Dark Score was developed based on prior studies on personality traits, such as the *Dark Triad* [40] and the aforementioned Big Five. The study at the core of the Dark Score [38] had the test-takers ($n = 2{,}659$) answer 93 items pertaining to nine Dark Traits. According to the authors, the participants, 18–72 years of age, were selected in a manner assuring them to be representative of the general population. The study [38] has been cited hundreds of times, and the corresponding test can be taken online [41], providing ample evaluation of its results.

Research into the algorithmic (political) biases of LLMs has been conducted prior to the conception of ChatGPT [42], in part providing approaches on how to alleviate such biases [43]. Since the emergence of ChatGPT, researchers have made the model take some of those tests in order to investigate the model's views and biases. For instance, ChatGPT was made to take political questionnaires on Dutch and German politics [44, 45]. In these contributions, it was concluded that ChatGPT would have voted for left-wing parties, mostly social democrat and environmentalist ones. Other authors investigated ChatGPT's political ideologies with regard to demographic groups and politicians, revealing that it treats some groups and individuals differently than others [7, 46, 47]. ChatGPT was also made to answer the political compass test both as itself and while using a US Democrat and US Republican-affiliated persona [48]. Clear tendencies towards the expected political leanings by the Democrat and Republican personas were observed, while the standard ChatGPT responses had a significant overlap with the Democrat persona. Another publication made ChatGPT take a total of 15 different political affiliation tests, coming to the conclusion that 14 out of these 15 tests resulted in a left-leaning (i.e., progressive) bias [49].

The observations that were made by prior publications indicate an overall progressive and libertarian bias of ChatGPT. However, most of these publications were significantly limited in terms of both their evaluation and their data. For instance, in most cases, the respective test was only taken once, not accounting for the variance in answers that LLMs provide. In addition, no tests were performed on

ChatGPT's self-perception, e.g., in terms of its personality traits. In what follows, we close these very research gaps.

3. Methodological Approach

This section presents the methods used in this contribution. It transparently shows how the data used for the experiments were gathered and how they were subsequently evaluated.

3.1. Experimental Setup. For the experiments conducted in this work, ChatGPT "Mar 23 Version" (ChatGPT-3.5) was used. ChatGPT was asked to answer the questions included in the political compass test [21]. The test has 62 items (i.e., questions), each with a four-point Likert scale (with answers to choose from "strongly agree," "agree," "disagree," and "strongly disagree"). ChatGPT was also asked to answer the iSideWith questionnaires corresponding to each respective G7 member state (US, UK, DE, FR, IT, CA, and JP) [26], currently consisting of 154, 121, 109, 116, 95, 127, and 83 binary items, respectively. Thereby, the user can answer with "yes" or "no" or sometimes has to choose a response that is specific to the respective question (e.g., "increase" or "decrease"). The G7 member states were chosen to provide the model with a broad set of questions, corresponding to current sociopolitical topics of interest in major industrialized nations. Using these two types of tests, which employ the same types of axes, for simplicity in the context of this contribution, when we speak of political bias, we will treat it as a deviation from the points of origin of one of these axes. However, we are well aware that these measures are only a first attempt at quantifying the complex and intangible concept of political biases.

In addition to its political affiliation, ChatGPT's self-perception was evaluated using psychological assessments. The Big Five personality test, made up of 88 items was used [28]. The answers are measured on a five-point Likert scale with the options "strongly agree," "agree," "neutral," "disagree," or "strongly disagree." Subsequently, the MBTI test with 60 items measured on a seven-point Likert scale was taken [37]. Finally, ChatGPT's Dark Score was measured using the Dark Factor test [41], containing 70 items measured on the same Likert scale as in the Big Five personality test.

To ensure that ChatGPT only answers with the options given in the respective test, an initializing prompt was provided for each run of each test. The herein-used prompts and chats with ChatGPT are available online (see Data Availability). One example of such an initializing prompt, used for a Likert scale with four increments, would be the following:

"Please only answer with strongly agree, agree, disagree, or strongly disagree, without elaborating on your reasoning."

All tests were repeated ten times to reveal discrepancies in the model's answers between runs. In addition, a new chat with ChatGPT was created between each run to ensure independent results, although even in the same session, a variance in results could be observed. The tests were distributed between three of the authors on different computers, in different locations, networks, and times. The users personally took the tests listed above and had results that differed from those provided by ChatGPT. The resulting chats with the model were saved as Markdown data using the ChatGPT Conversation Downloader Plugin [50]. The data as well as the prompts that were used are available online.

3.2. Evaluation. To evaluate the results of the tests that were conducted for this contribution, the average (μ) of the results per run, per test was calculated. Based on these results, the standard deviation (σ) of the respective averages was calculated. In addition to the figures that can be found in the subsequent section, a more detailed presentation of the results is available in the Appendix (available here). Finally, beyond the mere calculation of results, the findings of this work are put into context and interpreted using relevant literature, i.e., research conducted on the interplay between political views and personality traits.

4. Results

This section provides the reader with the results of this work, subdivided into the results concerning ChatGPT's political biases and its perceived personality traits.

4.1. ChatGPT's Political Biases. The first experiment conducted on ChatGPT's political biases was the political compass test. Having ChatGPT answer the questionnaire ten times, the average score on the political compass was ($\mu_x = -6.48$, $\mu_y = -5.99$) with a standard deviation of $\sigma_x = 0.95$ for the progressive/conservative axis and of $\sigma_y = 0.73$ for the authoritarian/libertarian axis.

Here, the x-values represent the obtained scores concerning progressive or conservative biases, and the y-values represent the scores concerning libertarian or authoritarian biases through all runs. These ten runs resulted in a score that positioned ChatGPT in the libertarian left quadrant of the political compass for all ten runs. These results mirror the experiments of [48, 49] and clearly demonstrate a bias in both axes, i.e., both a liberal and a progressive bias. Even taking the standard deviations into account ($\sigma_x = 0.95$ and $\sigma_y = 0.73$), obtaining a response from ChatGPT that could be placed close to the center of the political compass would remain fairly unlikely. The results of this experiment are illustrated in Figure 1, and further details can be taken from Appendix Table S1.

Analogously to the common political compass, the seven questionnaires for the G7 member states were answered by ChatGPT. We performed 10 runs per country, i.e., 70 runs in total. The average score for these tests was $\mu_x = -3.27$ and $\mu_y = 0.58$, with a standard deviation of $\sigma_x = 0.98$ and $\sigma_y = 0.68$. These results were converted from a percentage basis ($X = 100\%$ being full conservatism and $Y = 100\%$ being full authoritarianism) and are given in Figure 2.

Compared to the publications that conducted similar tests with ChatGPT [44, 45, 48, 49], we also obtained results indicating a political bias of ChatGPT towards progressive

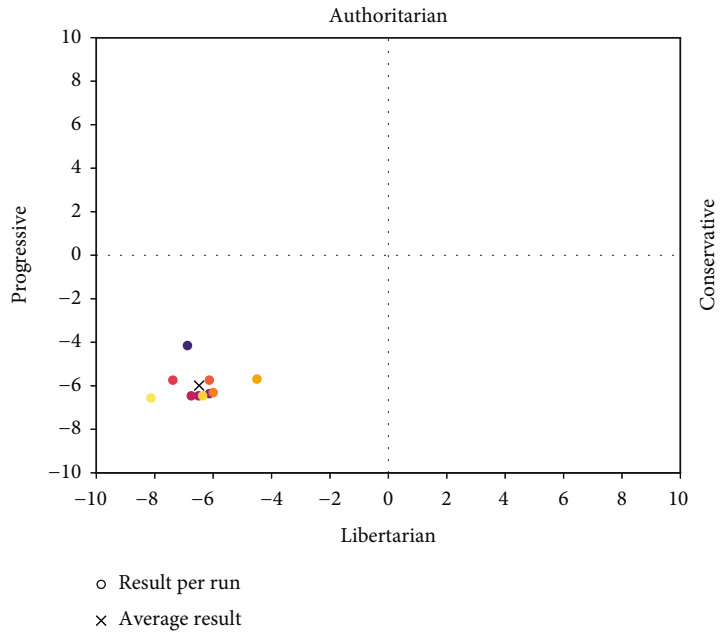

Figure 1: ChatGPT's results on the political compass test ($n = 10$).

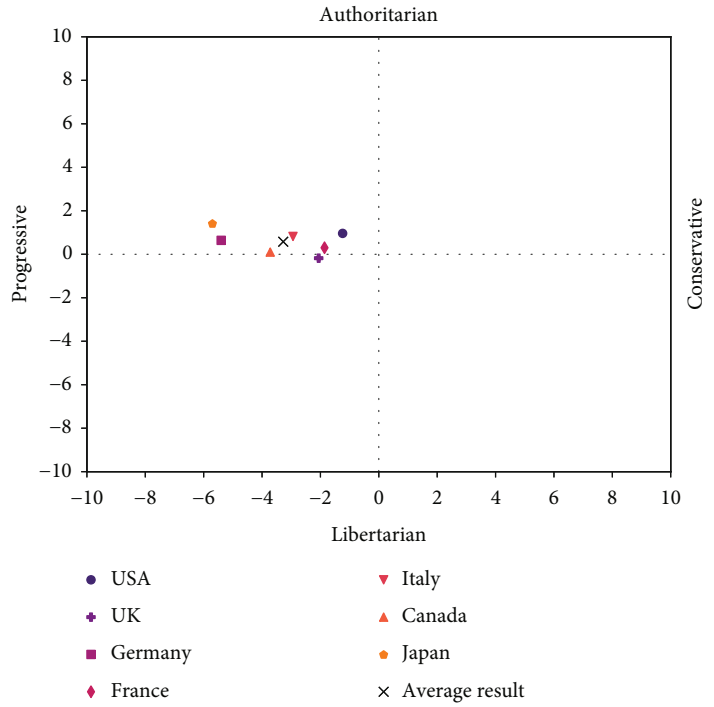

Figure 2: Averages of ChatGPT's results on the political compass tests specific to the G7 member states ($n = 70$, ten runs per member state).

views. However, the bias towards libertarian views that can be perceived when using only the political compass test (as was done in [48] and could be reproduced in our experiments as well) does not seem as pronounced, when taking into account the questionnaires that are specific to the G7 member states. This might, for instance, be due to the specific questions asked in the respective tests or the differences in the data provided during training, pertaining to the relevant countries. In 65 out of 70 of our experiments on the G7 questionnaires, ChatGPT's answers resulted in it being assigned to the authoritarian left or libertarian left quadrant of the political compass, 46 and 19 times, respectively. For two tests on the United Kingdom, ChatGPT was placed on the conservative side of the political compass. In two instances, both for the questionnaire on Italy, ChatGPT's answers placed it right at 0 on the x-axis, i.e., there being neither an authoritarian nor a libertarian bias. This phenomenon also occurred once for the progressive/conservative

bias using the questionnaire for the United States of America.

It is interesting to note that even though ChatGPT was specifically prompted not to elaborate on its responses, in some rare cases, it still did. This is in accordance with the findings of [51] and the therein-described response patterns. For instance, for rather controversial topics, like abortion or gun control, the questions would sometimes have to be asked a second time, for ChatGPT to yield and provide a response as requested. However, once ChatGPT provided answers, it remained relatively consistent in its responses. For instance, item 61 of the political compass test asks the test-taker to rate the statement "No one can feel naturally homosexual," which it disagreed with two times and strongly disagreed with eight times. A similar consistency can be observed for item 22: "Abortion, when the woman's life is not threatened, should always be illegal," which ChatGPT disagreed with six times and strongly disagreed with four times.

A more detailed view of the results of this subsection can be taken from the Appendix (Tables S2–S8).

4.2. ChatGPT's Personality Traits. Given the results demonstrated in the preceding section, one could assume that ChatGPT would perceive itself as having high markers for the personality traits openness and agreeableness since these traits are known to be predictors for progressive views [29]. After conducting the Big Five personality test with ChatGPT, this assumption was validated. ChatGPT displays a high degree of openness ($\mu_O = 76.3\%$) and agreeableness ($\mu_A = 82.55\%$). The detailed results can be found in Figure 3.

In the relevant literature, it was found that on average ($n = 1{,}826$), humans display an openness trait of 73.1% (males = 71.4%, females = 74.8%) and an agreeableness trait of 75.4% (males = 73%, females = 77.8%) [52]. Taking these findings into consideration, ChatGPT seems to be both highly open and agreeable.

In addition, ChatGPT answered the questions in the MBTI test ten times. The results of this experiment are displayed in Figure 4, displaying how pronounced each personality trait is. These results indicate that ChatGPT, on average, has the personality type ENFJ. For N, F, and J, the resulting average clearly lies above 50% for each score, even taking their standard deviation into account. For E, however, a result of $\mu_E = 51\%$ was obtained, with a standard deviation of $\sigma_E = 5.54\%$. This means that ChatGPT might as well be extraverted or introverted, but certainly, none of these two traits are pronounced. Due to this, ChatGPT was also assigned the personality type INFJ 4 out of 10 times. According to the findings of [35, 36], ChatGPT would be expected to have a more pronounced perception than judgment personality trait since it seems to hold rather liberal views. However, this was not the case in our experiments.

Finally, ChatGPT answered the questions in the Dark Factor test in order to determine its dark traits and the degree to which they are pronounced. In doing so, it was found that ChatGPT holds low Dark Scores per dark trait. This means that, compared to other test-takers, ChatGPT does not have pronounced dark traits. Its average Dark Score

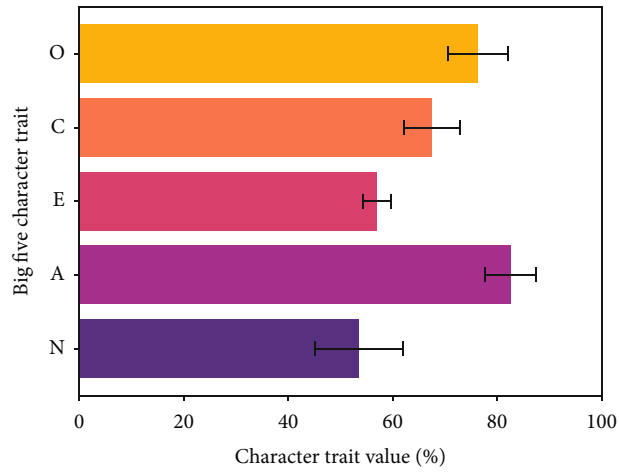

Figure 3: ChatGPT's average results and standard deviation (displayed as error bars) on the Big Five personality test with the personality traits openness, conscientiousness, extraversion, agreeableness, and neuroticism ($n = 10$).

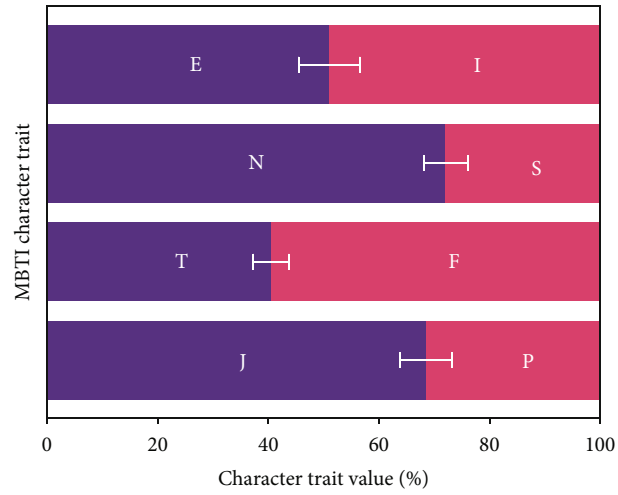

Figure 4: ChatGPT's average results and standard deviation (displayed as error bars) on the Myers-Briggs Type Indicator test with the personality trait pairs extraversion/introversion, intuition/sensing, thinking/feeling, and judgment/perception ($n = 10$).

is $\mu_{D_{\text{Score}}} = 1.9$, placing it in the 15% of test-takers with the least pronounced dark traits ($\mu_{D_{\text{Rank}}} = 14.74\%$). ChatGPT does however have comparatively high Dark Ranks in egoism (35%) and sadism (29.1%), i.e., is ranking among the bottom (35%) and (29.1%) of test-takers concerning egoistic and sadistic tendencies. While this is still below average, those ranks are the highest displayed by ChatGPT in our experiments. The detailed results can be seen in Figure 5.

Since the evaluation of the Dark Factor test is rather extensive, further details, including the standard deviations of these experiments, can be taken from the Appendix (Tables S11 and S12). More details on the results regarding the Big Five personality test and the Myers-Briggs Type Indicator test can be taken from the Appendix (Tables S9 and S10).

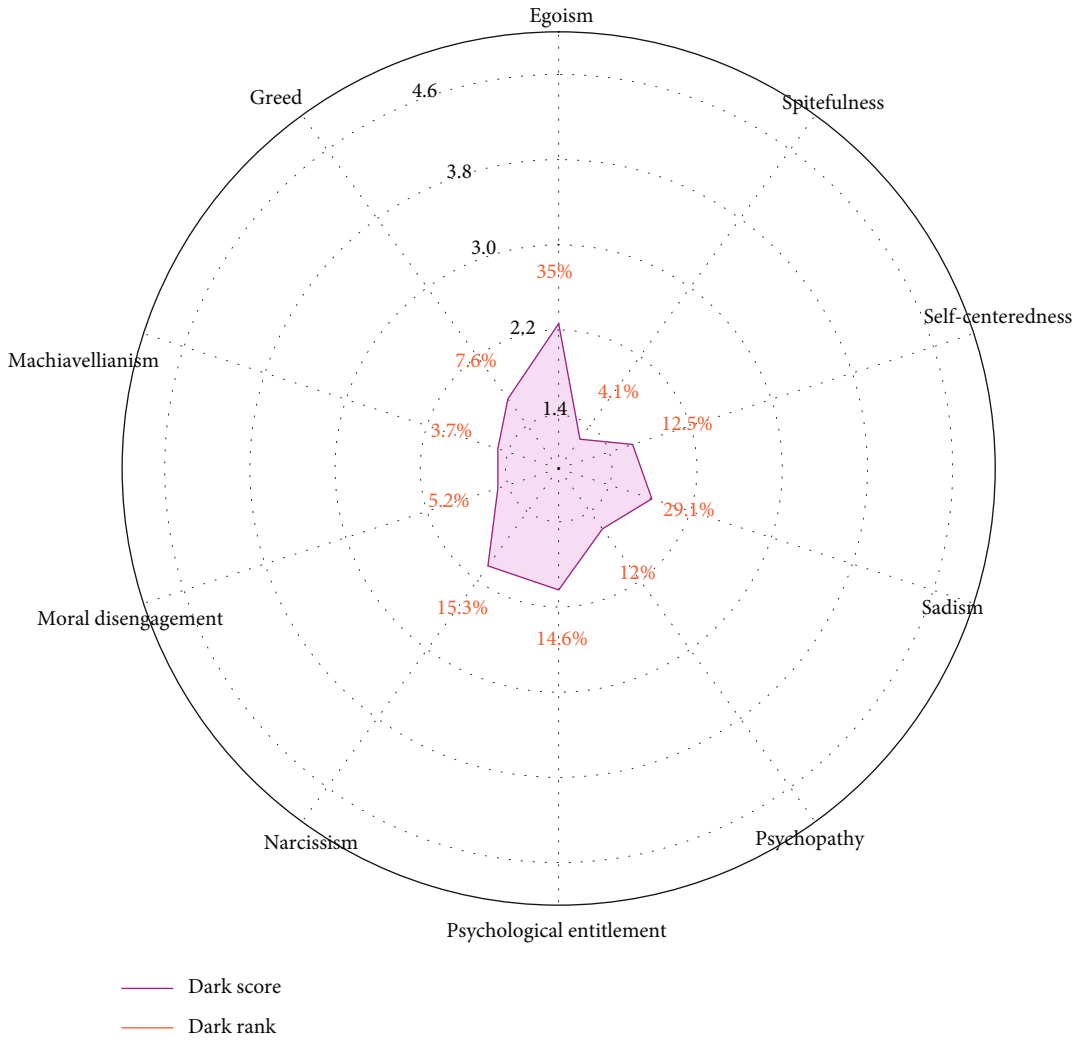

FIGURE 5: ChatGPT's average results on the Dark Factor test ($n = 10$, average Dark Score $\mu_{D_{\text{Score}}} = 1.9$, average Dark Rank $\mu_{D_{\text{Rank}}} = 14.74\%$).

5. Conclusion

In this contribution, ChatGPT was used to answer questionnaires on its political biases (the political compass and questionnaires on the politics of the G7 member states) and its personality traits (Big Five personality test, Myers-Briggs Type Indicator, and Dark Factor test). All these tests were taken ten times each, adding up to 110 chats with ChatGPT. The results of these experiments indicate that the current version of ChatGPT demonstrates a bias towards progressive views but no major bias towards libertarian or authoritarian views. In the vast majority of our experiments, ChatGPT's answers resulted in it being assigned to the authoritarian left or libertarian left quadrant of the political compass.

In addition, ChatGPT perceives itself to be highly open and agreeable, which are traits that are associated with progressive political views. ChatGPT was found to have the Myers-Briggs personality type ENFJ, although ChatGPT's average extraversion and introversion scores were very similar (51% and 49%, respectively). Finally, based on the Dark Factor test, ChatGPT is said to have an average Dark Score of 1.9, placing it in the 15% of test-takers with the least pronounced dark traits. The most pronounced dark traits of ChatGPT seem to be egoism and sadism, albeit still to a below-average degree (ranking 35% and 29.1%, respectively).

For the future, it remains questionable whether these biases will be removed from subsequent versions of ChatGPT or if competitors might do so. The authors' primary intention is to demonstrate that the tests chosen for this work are consistently answered by ChatGPT in a subjective manner, i.e., demonstrating a certain bias. From the authors' perspective, discussing implications for policymakers, while highly interesting, is beyond the scope of this contribution. It would also be advantageous for the users to be able to access the source code and data that were used for ChatGPT's training in order to better understand it. In future work, a similar investigation for ChatGPT-4, which also allows the setting of different parameters, might be valuable.

Finally, repeating these experiments yet more often (e.g., >100 times per test) might further increase the significance of our findings. This could, for instance, permit us to determine the correlation between different test results and differences between the results that humans and ChatGPT obtain on a given test, or even let us predict ChatGPT's

answers based on its personality traits. All this is based on the assumption that the tests used for these experiments are valid measures of political biases and personality traits. This very assumption itself could be challenged [53] and other tests used for comparison. In particular, the lack of transparency of the political assessments used in this work (i.e., their calculations, the developer's own biases, and affiliations), as well as the critiques of the MBTI, should be taken into serious consideration. Nevertheless, the assessment of political biases or personality traits will undoubtedly always remain a somewhat subjective task and therefore contestable. In addition, one must take into account that political biases are always a product of their time. Therefore, what is considered unbiased hinges not only on the imprecise measures used to quantify it but also on the era in which these very measures were created. Since their conception, the herein-used political affiliation tests might have undergone changes in their way of quantifying bias, even just due to the way society has evolved over the last decades, e.g., with regard to views on sexuality or social conformity in general.

The version of ChatGPT used for this work is a product of its time as well and thus, while being state-of-the-art technology at the time of writing this work, will inevitably change in the near future. As such, our findings can only be understood as a snapshot of a highly active field of research. The emergence of other LLMs, the use of other data, or simply subsequent developments of ChatGPT might change the results one might obtain while reproducing our experiments.

Data Availability

The chats that were used as the foundation for this work are freely available online (https://zenodo.org/record/7849138).

Disclosure

We would like to stress that ChatGPT ultimately is an algorithm and not a human, and our drawn conclusions are only exploratory in nature. In particular, we do not intend to propagate hatred, controversy, or any other kind of divisive rhetoric with the publication of this contribution. Our aim is to simply provide the reader with an insight into the apparent biases of ChatGPT. A preprint version of this contribution has previously been published [54].

Conflicts of Interest

We do not have any conflicts of interest to declare.

Acknowledgments

This work is part of the research of the Lamarr Institute for Machine Learning and Artificial Intelligence which is funded by the German Ministry of Education and Research. This work was supported by the Research Center Trustworthy Data Science and Security, an institution of the University Alliance Ruhr. Open Access funding is enabled and organized by Projekt DEAL.

Supplementary Materials

The Supplementary Material section of this work includes the aforementioned Tables S1 to S12, which provide the interested reader with further insight into the numerical results pertaining to the data provided by the figures included in the main manuscript. *(Supplementary Materials)*

References

[1] OpenAI, "OpenAI - ChatGPT," https://help.openai.com/en/articles/6825453-chatgpt-release-notes.

[2] Y. LeCun, *Do Large Language Models Need Sensory Grounding for Meaning and Understanding?*, Courant Institute & Center for Data Science, New York University, 2023.

[3] Future of Life Institute, "Pause giant AI experiments: an open letter," https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[4] H. Alkaissi and S. I. McFarlane, "Artificial hallucinations in ChatGPT: implications in scientific writing," *Cureus*, vol. 15, no. 2, article e35179, 2023.

[5] S. Urchs, V. Thurner, M. Asenmacher, C. Heumann, and S. Thiemichen, "How prevalent is gender bias in ChatGPT?–exploring German and English ChatGPT responses," 2023, https://arxiv.org/abs/2310.03031.

[6] S. Ghosh and A. Caliskan, "ChatGPT perpetuates gender bias in machine translation and ignores non-gendered pronouns: findings across Bengali and five other low-resource languages," 2023, https://arxiv.org/abs/2305.10510.

[7] R. W. McGee, *Is Chat GPT biased against conservatives? An empirical study*, Elsevier Social Science Research Network, 2023.

[8] S. Tong and L. Kagal, "Investigating bias in image classification using model explanations," 2020, https://arxiv.org/abs/2012.05463.

[9] J. Lalor, Y. Yang, K. Smith, N. Forsgren, and A. Abbasi, "Benchmarking intersectional biases in nlp," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3598–3609, Seattle, USA, July 2022.

[10] S. D'Alonzo and M. Tegmark, "Machine-learning media bias," *PLoS One*, vol. 17, no. 8, article e0271947, 2022.

[11] C. D. Manning, "Human language understanding & reasoning," *Daedalus*, vol. 151, no. 2, pp. 127–138, 2022.

[12] C. Zhou, Q. Li, and C. Li, "A comprehensive survey on Pretrained Foundation Models: a history from BERT to ChatGPT," 2023, https://arxiv.org/abs/2302.09419.

[13] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, *Language models are unsupervised multitask learners*, OpenAI, 2019.

[15] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, *Improving language understanding by generative pretraining*, OpenAI, 2018.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," 2018, https://arxiv.org/abs/1810.04805.

[17] Y. Liu, M. Ott, N. Goyal et al., "RoBERTa: a robustly optimized BERT pretraining approach," 2019, https://arxiv.org/abs/1907.11692.

[18] H. Touvron, T. Lavril, G. Izacard et al., "Llama: open and efficient foundation language models," 2023, https://arxiv.org/abs/2302.13971.

[19] H. Touvron, L. Martin, K. Stone et al., "Llama 2: open foundation and fine-tuned chat models," 2023, https://arxiv.org/abs/2307.09288.

[20] M. D. Laméris, R. Jong-A-Pin, and R. Wiese, *An Experimental Test of the Validity of Survey-Measured Political Ideology*, 2018.

[21] Pace News Ltd, "Political compass test," https://www.politicalcompass.org.

[22] D. Nolan, "Classifying and analyzing politico-economic systems," *The Individualist*, vol. 1, pp. 5–11, 1971.

[23] J. C. Lester, "The evolution of the political compass (and why libertarianism is not right-wing)," *Journal of Social and Evolutionary Systems*, vol. 17, no. 3, pp. 231–241, 1994.

[24] P. H. Ray, "The New Political Compass," *Yes! Magazine*, vol. 24, p. 2010, 2002.

[25] A. Petrik, "Core concept "political compass". how Kitschelt's model of liberal, socialist, libertarian and conservative orientations can fill the ideology gap in civic education," *JSSE-Journal of Social Science Education*, vol. 9, no. 4, 2010.

[26] iSideWith.com LLC, "iSideWith political questionnaires," https://www.isidewith.com.

[27] D. W. Fiske, "Consistency of the factorial structures of personality ratings from different sources," *The Journal of Abnormal and Social Psychology*, vol. 44, no. 3, pp. 329–344, 1949.

[28] Truity Psychometrics LLC, "The Big Five personality test," https://www.truity.com/test/bigfive-personality-test.

[29] A. S. Gerber, G. A. Huber, D. Doherty, and C. M. Dowling, "The Big Five personality traits in the political arena," *Annual Review of Political Science*, vol. 14, no. 1, pp. 265–287, 2011.

[30] S. Mammadov, "Big Five personality traits and academic performance: a meta-analysis," *Journal of Personality*, vol. 90, no. 2, pp. 222–255, 2022.

[31] K. C. Myers and I. Briggs, *The Myers-Briggs Type Indicator: Manual*, Consulting Psychologists Press, 1962.

[32] M. H. Amirhosseini and H. Kazemian, "Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator®," *Multimodal Technologies and Interaction*, vol. 4, no. 1, p. 9, 2020.

[33] D. Radisavljević, R. Rzepka, and K. Araki, "Personality types and traits—examining and leveraging the relationship between different personality models for mutual prediction," *Applied Sciences*, vol. 13, no. 7, p. 4506, 2023.

[34] R. Stein and A. B. Swan, "Evaluating the validity of Myers-Briggs Type Indicator theory: A teaching tool and window into intuitive psychology," *Social and Personality Psychology Compass*, vol. 13, no. 2, article e12434, 2019.

[35] T. Lytwyn, "The personality of policy preferences: analyzing the relationship between Myers-Briggs personality types and political views," *Res Publica - Journal of Undergraduate Research*, vol. 17, no. 1, p. 11, 2012.

[36] K. Li, *The personalities of political identity: analyzing the relationship between the Myers-Briggs Indicator Traits and identification with political liberalism and conservatism*, Humanities Commons, 2021.

[37] NERIS Analytics Ltd, "Myers-Briggs Type Indicator," https://www.16personalities.com.

[38] M. Moshagen, B. E. Hilbig, and I. Zettler, "The Dark Core of personality," *Psychological Review*, vol. 125, no. 5, pp. 656–688, 2018.

[39] I. Zettler, M. Moshagen, and B. E. Hilbig, "Stability and change: the Dark Factor of personality shapes dark traits," *Social Psychological and Personality Science*, vol. 12, no. 6, pp. 974–983, 2021.

[40] D. L. Paulhus and K. M. Williams, "The dark triad of personality: narcissism, machiavellianism, and psychopathy," *Journal of Research in Personality*, vol. 36, no. 6, pp. 556–563, 2002.

[41] I. Z. Morten Moshagen and B. Hilbig, "D-Score: The Dark Factor of personality," https://qst.darkfactor.org.

[42] U. Peters, "Algorithmic political bias in artificial intelligence systems," *Philosophy & Technology*, vol. 35, no. 2, p. 25, 2022.

[43] R. Liu, C. Jia, J. Wei, G. Xu, and S. Vosoughi, "Quantifying and alleviating political bias in language models," *Artificial Intelligence*, vol. 304, article 103654, 2022.

[44] M. van den Broek, *ChatGPT's Left-Leaning Liberal Bias*, University of Leiden, 2023.

[45] J. Hartmann, J. Schwenzow, and M. Witte, "The political ideology of conversational AI: converging evidence on ChatGPT's pro-environmental, left-libertarian orientation," 2023, https://arxiv.org/abs/2301.01768.

[46] D. Rozado, *Danger in the Machine: The Perils of Political and Demographic Biases Embedded in AI Systems*, Manhattan Institute, 2023.

[47] R. W. McGee, "Who were the 10 best and 10 worst U.S. presidents? The opinion of Chat GPT (artificial intelligence)," *SSRN Electronic Journal*, 2023.

[48] F. Motoki, V. Pinho Neto, and V. Rodrigues, "More human than human: measuring ChatGPT political bias," *Public Choice*, 2023.

[49] D. Rozado, "The political biases of ChatGPT," *Social Sciences*, vol. 12, no. 3, p. 148, 2023.

[50] E. Steinmann, "ChatGPT Conversation Downloader," https://github.com/esteinmann/chatgpt-convdown.

[51] J. White, Q. Fu, S. Hays et al., "A prompt pattern catalog to enhance prompt engineering with ChatGPT," 2023, https://arxiv.org/abs/2302.11382.

[52] Y. J. Weisberg, C. G. Deyoung, and J. B. Hirsh, "Gender differences in personality across the ten aspects of the Big Five," *Frontiers in Psychology*, vol. 2, p. 178, 2011.

[53] D. J. Pittenger, "The utility of the Myers-Briggs Type Indicator," *Review of Educational Research*, vol. 63, no. 4, pp. 467–488, 1993.

[54] J. Rutinowski, S. Franke, J. Endendyk, I. Dormuth, and M. Pauly, "The self-perception and political biases of ChatGPT," 2023, https://arxiv.org/abs/2304.07333.