nature portfolio

Peer Review File

Studying and improving reasoning in humans and machines

Web links to the author's journal account have been redacted from the decision letters as indicated to maintain confidentiality

5th Dec 23

Dear Dr Palminteri,

Thank you for your patience during the peer-review process. Your manuscript titled "Studying and improving reasoning in humans and machines" has now been seen by 3 reviewers, and I include their comments at the end of this message. They find your work of interest, but raised some important points. We are interested in the possibility of publishing your study in Communications Psychology, but would like to consider your responses to these concerns and assess a revised manuscript before we make a final decision on publication.

We therefore invite you to revise and resubmit your manuscript, along with a point-by-point response to the reviewers. Please highlight all changes in the manuscript text file.

The reviewers raised some points that we hope will allow you to strengthen your paper. You are expected to respond to each point raised by the reviewers, but we ask you to pay particular attention to the following:

Reviewers #1 and #2 mention issues that affect the generalizability of the results on the human (Ref #2) or model side (Ref #1). We ask you to address these through suitable analyses.

The Reviewers collectively raise points that affect the degree to which the results are unambiguously interpretable and which cannot be resolved through additional empirical work. Please ensure that you address these thoroughly in the Discussion in a dedicated "Limitations" section.

Lastly, please also ensure that you follow the required order of sections (Introduction - Methods - Results - Discussion). The Methods section needs to be moved into the main manuscript and a statement on whether ethical approval and consent were obtained is required.

We are committed to providing a fair and constructive peer-review process. Please don't hesitate to contact us if you wish to discuss the revision in more detail.

Please note that your revised manuscript must comply with our formatting and reporting requirements, which are summarized on the following checklist:
Communications Psychology formatting checklist and also in our style and formatting guide
Communications Psychology formatting guide .

Please use the following link to submit your revised manuscript, point-by-point response to the referees' comments (which should be in a separate document to any cover letter) and the completed checklist:
[link redacted]

** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

We hope to receive your revised paper within 8 weeks; please let us know if you aren't able to submit it within this time so that we can discuss how best to proceed. If we don't hear from you, and the revision process takes significantly longer, we may close your file. In this event, we will still be happy to reconsider your paper at a later date, provided it still presents a significant contribution to the literature at that stage.

We would appreciate it if you could keep us informed about an estimated timescale for resubmission, to facilitate our planning.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Best regards,

Antonia Eisenkoeck

Antonia Eisenkoeck
Senior Editor
Communications Psychology

EDITORIAL POLICIES AND FORMATTING

We ask that you ensure your manuscript complies with our editorial policies. Please ensure that the following formatting requirements are met, and any checklist relevant to your research is completed and uploaded as a Related Manuscript file type with the revised article.

Editorial Policy: Policy requirements (Download the link to your computer as a PDF.)

* TRANSPARENT PEER REVIEW: Communications Psychology uses a transparent peer review system. This means that we publish the editorial decision letters including Reviewers' comments to the authors and the author rebuttal letters online as a supplementary peer review file. However, on author request, confidential information and data can be removed from the published reviewer reports and rebuttal letters prior to publication. If your manuscript has been previously reviewed at another journal, those Reviewers' comments would not form part of the published peer review file.

* CODE AVAILABILITY: All Communications Psychology manuscripts must include a section titled "Code Availability" at the end of the methods section. In the event of publication, we require that the custom analysis code supporting your conclusions is made available in a publicly accessible repository; at publication, we ask you to choose a repository that provides a DOI for the code; the link to the repository and the DOI will need to be included in the Code Availability statement. Publication as Supplementary Information will not suffice. We ask you to prepare code at this stage, to avoid delays later on in the process.

* DATA AVAILABILITY:

All Communications Psychology manuscripts must include a section titled "Data Availability" at the end of the Methods section or main text (if no Methods). More information on this policy, is available at http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf.

At a minimum the Data availability statement must explain how the data can be obtained and whether there are any restrictions on data sharing. Communications Psychology strongly endorses open sharing of data. If you do make your data openly available, please include in the statement:
- Unique identifiers (such as DOIs and hyperlinks for datasets in public repositories)
- Accession codes where appropriate
- If applicable, a statement regarding data available with restrictions
- If a dataset has a Digital Object Identifier (DOI) as its unique identifier, we strongly encourage including this in the Reference list and citing the dataset in the Data Availability Statement.

We recommend submitting the data to discipline-specific, community-recognized repositories, where possible and a list of recommended repositories is provided at http://www.nature.com/sdata/policies/repositories.

If a community resource is unavailable, data can be submitted to generalist repositories such as figshare or Dryad Digital Repository. Please provide a unique identifier for the data (for example a DOI or a permanent URL) in the data availability statement, if possible. If the repository does not provide identifiers, we encourage authors to supply the search terms that will return the data. For data that have been obtained from publicly available sources, please provide a URL and the specific data product name in the data availability statement. Data with a DOI should be further cited in the methods reference section.

Please refer to our data policies at http://www.nature.com/authors/policies/availability.html.

REVIEWERS' EXPERTISE:

Reviewer #1: decision making
Reviewer #2: decision making, biased reasoning
Reviewer #3: decision making, biased reasoning

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

This paper investigates how large language models in the GPT family solve traditional reasoning problems – and how reasoning can be improved in machines as well as humans. In general, I think the manuscript has value and is written very well. I have some (rather minor) suggestions for further improvements, that the authors can address in any ways they want:
The authors use GPT models for their investigations. These models, as well as other LLMs, are ubiquitous, it is unclear what texts they are trained on exactly, and their underlying structure is not published as well. The problem is that these models could be similar in some unknown ways as well, producing a selection bias in the data. Other model classes, developed by different teams (and relying on potentially different model structures, different training data, and different training

procedures)could be used to get a different benchmark and test the generalizability of the findings to a larger class of LLMs (models like Llama2, Chinchilla, Vicuna-13B).
Another problem comes from the unknown samples these models were trained on by using reinforcement learning from human feedback which GPT4 training extensively relies on for example. The composition of these samples could induce different biases, and model output might simply reflect the thinking of the people in this sample. For example, GPT4 might have used people who are superior reasoners and get the reasoning problems right, consistently. How likely is it that differences in the models represent differences in the training population instead of the underlying structure? What sample biases should be considered in the training of these machines that could affect reasoning performance?
In the concluding paragraphs, the authors say 'We were therefore unable to find, in the considered model space, no model capable to correctly mimic the performance of our standard experimental sample'. I do not think the results can support this final conclusion. As I said, training data, and training using extensive reinforcement learning from human feedback procedures could potentially be used to train these machines for superior performance, as well as very human-like performance. Better or worse than human performance is therefore not necessarily inherent to these machines.
Line 32 "where"-> were

Reviewer #2 (Remarks to the Author):

This ms presents a very interesting comparison between human reasoning and that observed in LLMs of different generations. The authors adapted problems from two often used measures of human irrationality, the CRT and the conjunction fallacy. Results show that there is a clear decrease in the extent to which more advanced LLMs show the same propensity to "illogical" reasoning as is typically shown in human samples. In addition, use of two prompts on the earlier models produced somewhat better results than with the human sample. These results are very interesting, particularly those showing that the latest versions are not as prone to biased reasoning. There are however some limitations to these results. First, although there are many different kinds of reasoning and judgments that show biases, the authors focused on only two, with no indication why these might be more representative of any general tendency to biased reasoning. Certainly, a larger sample of problems would answer any questions as to the generality of the idea that recent LLMs are not prone to biased reasoning. Second, as the authors note themselves, using the same prompts with humans as with LLMs simply might not work due to the difficulty that humans have in understanding just what is meant by these, or by their difficulty in adjusting their actual behavior. There are some recent studies on debiasing reasoning in humans that might suggest a more direct way of comparing what humans do with LLMs. Finally, the authors consider humans as being a homogenous sample. This is of course certainly wrong. At the very least, there are large individual differences in the propensity of individuals to produce biased reasoning. There are certainly a subset of people whose tendency to biased reasoning would look like that found with Chatgpt or possibly GPT-4. Unlike individual LLMs, people differ greatly in terms of their background knowledge, level of education, specific training, etc. Any attempt to directly compare the reasoning of LLMs and humans that does not consider individual differences among the latter is almost certainly misleading.
Nonetheless, these results are very interesting. One possibility to make the comparison more useful would be to split the human sample into performance tiers, to make the within human variation much clearer, and also to make the real complexity of how to eventually compare LLMs with humans more evident.

Reviewer #3 (Remarks to the Author):

My area of expertise is in cognitive science with specific interests lie in the role of intuitive and deliberative processes in human reasoning and the veracity of dual process models as an explanatory framework for understanding human reasoning and judgment. In this regard, my comments are necessarily limited to this aspect of the paper rather than the technical details regarding the LLM model features and characteristics.

My overall evaluation of this paper is very positive. It is novel, timely and extremely interesting. The authors have approached their research question very thoughtfully, utilising a range of novel contents in their judgment tasks to eliminate potential explanations based upon direct match between LLM training materials and their task set. Their analysis is thorough and clearly presented and incorporate a broad range of LLM models.

The findings are intriguing. The evidence of similar biases amongst all but the most recent LLM models might indeed suggest that the source of certain biases in reasoning and judgment lies in a dissociation between language-based reasoning drawing on statistical regularities and mathematical computation, a conclusion that aligns well with the distinction made between dual system models of human reasoning.

The authors consider two potential explanations for their findings; biases present in the corpus or biases that arise through human feedback fine tuning. However, it remains unclear why the biases are not present in the lates Chat GPT and GPT-4 models.

Whilst the behaviour of earlier models appears to approximate participants responses on a global basis, the item-based analyses show clear areas of non-alignment and the differential impact of prompts is also surprising.

These aspects of the findings are intriguing and there clearly remains much additional work to be done to develop an explanation for these divergencies as this will be crucial in determining whether this type of research can genuinely inform psychological or computational models of human reasoning and judgment. This is the most significant weakness in this paper, as it is difficult to draw any firm conclusions about the use of LLMs in informing psychological theory.

However, despite this, it is a high-quality piece of work utilising an approach that will be valuable in informing future work and potential extension to a broader range phenomenon in the heuristics and biases domain. I am consequently supportive of publication.

Reviewer #1 (Remarks to the Author):

This paper investigates how large language models in the GPT family solve traditional reasoning problems – and how reasoning can be improved in machines as well as humans. In general, I think the manuscript has value and is written very well. I have some (rather minor) suggestions for further improvements, that the authors can address in any ways they want:

We thank the reviewer for their positive and helpful comments.

The authors use GPT models for their investigations. These models, as well as other LLMs, are ubiquitous, it is unclear what texts they are trained on exactly, and their underlying structure is not published as well. The problem is that these models could be similar in some unknown ways as well, producing a selection bias in the data. Other model classes, developed by different teams (and relying on potentially different model structures, different training data, and different training procedures)could be used to get a different benchmark and test the generalizability of the findings to a larger class of LLMs (models like Llama2, Chinchilla, Vicuna-13B).

R1.1

We thank the reviewer for this important point. We agree that the models we used suffer from being opaque, in that they are not open-source and that this undermines, to some extent, the generalizability and the interpretability of the results.

Before diving into the additional analyses and experiments, we performed in order to obviate, as much as possible, to this issue, we wanted to point to how technical appendix where we try (by looking at the token log-probabilities only) to “empirically” investigate the relation between Open AI models (see technical appendix “*Computing distance between large language models*”). Of course, the information we get does not replace explicit and transparent information concerning model structure, training corpus, and fine-tuning strategy, but it does, at least, some information concerning their relative “similarity”.

Back to the main point, we reacted to the Reviewer’s point in two main ways. First, we included in our study results from 4 additional, open source, models. Two of them may be considered as belonging to the “first” generation of open source LLM (Hugging face’s Bloom-7B and Meta’s OPT-13B), two others are more recent (Meta’s LLAMA 2 7B and LMSYS’s Vicuna v1.5 13B). We tested our main tasks (the new CRT and Linda/Bill items) on these two models and report their performance in the manuscript (supplementary **Figure X** – see below):

The results show that performance were relatively poor in these models and generally comparable to those observed in the early models released by Open AI (e.g., DV, DVB). Critically, especially in the CRT experiment, lower accuracy is not driven by a higher rate of intuitive reasoning, but rather by a higher frequency of responses that

are classified as “other” and betrays the fact that these models often fail to engage with the user in a question/response manner. We discuss what could be the source of this difference and we mainly point to fine tuning as a possible reason, but also that the propensity of a model to engage in a question/response interaction (as requested by our experiments) has not to be confounded with its cognitive performance.

Finally, to further showcase the limitations inherent to working with opaque, proprietary, LLMs we include in the current manuscript results of some replication experiments we run on chatGPT and GPT-4 at a different time step, X months after the results we originally included in the manuscript. The results are shown in the figure below (and integrated to **Figure X** of the revised manuscript) and show that, while the general observation that these two models outperform humans in both tasks, their performance significantly changed across time points. This was probably due to the fact the Open AI updated the models, but in absence of clear timestamps (generally absent in the main website and present only in the API) and accessibility to past releases, it also illustrates how research based on these models could be doomed to be not replicable. This replicability issue is further complicated by the fact that some models can be removed altogether, as it recently happened for many of the GPT-3 models.

We now present these new results (open source models and test-retest reliability) in the revised results and insist on this point in the revised discussion:

Results

Test-retest reliably and open source models

To conclude our study, we replicated our main experiments (new CRT and Linda/Bill items) in the latest models (ChatGPT and GPT4) at different time points (March vs June 2023). Surprisingly, even if this second set of experiments confirmed that these models (ChatGPT and GPT4) continue to outperform human participants, models’ performance in both tasks significantly differed between time points. Accuracy in the CRT test (new items) increased in ChatGPT (correct response ratio March = 0.70(0.01) vs correct response ratio June = 0.79(0.01); Testing time main effect: $\chi 2 = 5$ DF = 1, P = 0.03; March vs June tukey-corrected pairwise contrast P = 0.001) and, surprisingly, decreased in GPT4 (correct response ratio March = 0.97(0.01) vs correct response ratio June = 0.79(0.01); Testing time main effect: $\chi 2 = 9$ DF = 1, P = 0.003; March vs June tukey-corrected pairwise contrast P < .0001) (**Figure 7A**). Concerning the Lind/Bill experiment, accuracy decreased significantly in both LLMs and, specifically in GPT4, we observed the emergence of a fallacy (substantially lower accuracy in the Linda compared to the Bill items) at the second timepoint (**Figure 7**B). The instability of the results at different timepoint can only be explained by Open AI releasing updated versions of the models and illustrate an important caveat associated to studying proprietary and opaque models: they may evolve, thus undermining replicability, while lacking clear information concerning the timing and nature of the changes [59].

Following this line of reasoning, we run our main experiments on a battery of four open source models (Hugging face’s Bloom-7B and Meta’s OPT-13B, Meta’s LLAMA 2 7B and LMSYS’s Vicuna v1.5 13B). The results are showed in **Figure 7C** and **7D**, for CRT and Linda/Bill, respectively). Overall these models performed quite poorly in both tasks. Concerning the CRT experiment the correct response rate was quite close to 0% in all case, except for VICUNA (the only model approaching the performance of DV3). In all models the intuitive response rate was above the correct response, but in general slightly above the correct response rate, indicating that the majority of the responses were neither “correct” nor “intuitive” (and therefore classified as “others). The nature of the actual responses suggested two different sources of errors in OPT and BLOOM, versus LLAMA and VICUNA. The “other” responses of the

former were due to the models failing to engage in a proper question/answer modality, while the latter (LLAMA and VICUNA) did engage in a Q&A modality, but delivered quite random responses. Concerning the Linda/Bill task, all models were around chance and presented no systematic conjunction fallacy.

Discussion

Unfortunately, among the open-source models that we investigated, none achieve significant levels of accuracy, but the increasing popularization of customized fine-tuning should solve this issue. Our analyses nonetheless underscored an additional important aspect of working with proprietary model, that is often overlooked,[59]. As we tested ChatGPT and GPT4 models at two different time points (roughly three months apart), we found that their performances (in both tasks and models) differed significantly. Even if in this case the overall experiments generally led to the same conclusions (i,e,, models outperformed human participants), performance instabilities due to (disclosed or undisclosed) updates in models are bount to undermine reproducibilityof this and other future research endeavors.

[…]

On the other hand, concerning open source models, our results show that their performance were relatively poor in these models and generally comparable to those observed in the early models released by Open AI (e.g., DV, DVB). The source of this relatively poor performance in the CRT experiment was found not in a higher rate of intuitive reasoning, but rather in a higher frequency of responses that are classified as “other”. In the case of the two early models (BLOOM and OPT), it mainly reflected a failure of these to engage with the user in a question/response manner. This raise the possibility of a potential confounding factor in evaluating models performances in a Q&A format; what can be taken as lower cognitive abilities could actually just betray their incapacity to detect answers to respond. The same interpretation was not true for the more recent models (LLAMA and VICUNA) who did engate in a Q&A manner, but provide out-of-the-blue responses, thus indicating that the presence of intuitive reasoning in many DV models was not inevitable, nor trivial. Of note, none of the open source model reached level performance comparable to human participants and more recent OpenAI models (ChatGPT and GPT-4).

Figure 7

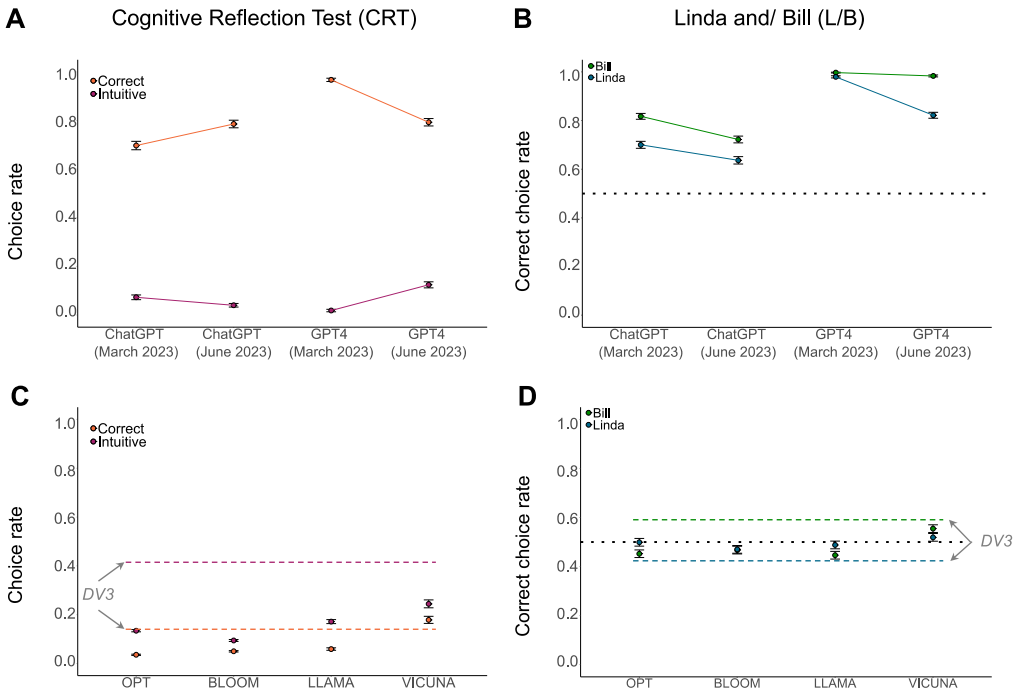


Figure 7: Test Retest reliability of ChatGPT and GPT4 and open source models. *(**A**) and **(B)** CRT results (i.e., correct and intuitive response rate) and) Linda/Bill results (correct response rate for the Linda and Bill items) for ChatGPT and GPT-4 at two different time pints (March – June 2023). (**C**) and (**D**) CRT and Linda/Bill results for four open access models. For visual reference, dotted colored lines represent the performance of DV3 model.*

Another problem comes from the unknown samples these models were trained on by using reinforcement learning from human feedback which GPT4 training extensively relies on for example. The composition of these samples could induce different biases, and model output might simply reflect the thinking of the people in this sample. For example, GPT4 might have used people who are superior reasoners and get the reasoning problems right, consistently. How likely is it that differences in the models represent differences in the training population instead of the underlying structure? What sample biases should be considered in the training of these machines that could affect reasoning performance?

R1.2

The reviewer here raises an important question, which can unfortunately answer only speculatively, given the issues related to the relative opacity of Open AI models and methods. Specifically, it is true that, in addition to differences in model structure and corpus, differences could in principle also arise from the fact that models fine-tuned based on reinforcement from human feedback, may have been exposed to feedback deriving from different human samples. Accordingly, an increase in performance in a given model could also be potentially ascribed to having being fine-tuned using feedback coming from particularly high functioning individuals (or experts on some sorts). We intuitively find unlikely that this is currently driven the performance difference between the included models (e.g., DV3, chatGPT and GPT4), specifically because we (presume) that human feedback is derived from panels of individual belonging to

the general population typically involved in this kind of studies (in fact, we believe that the fine-tuning feedback may actually come from populations not so dissimilar from those regularly accepting being part of our experiments). However, we cannot preclude this possibility and we would not be that surprised if, as the numbers and types of LLMs increase, expert-based fine-tuning will become common (for example one may argue that for medical applications feedback gathered from doctors will prove more useful and so forth).

In the concluding paragraphs, the authors say 'We were therefore unable to find, in the considered model space, no model capable to correctly mimic the performance of our standard experimental sample'. I do not think the results can support this final conclusion. As I said, training data, and training using extensive reinforcement learning from human feedback procedures could potentially be used to train these machines for superior performance, as well as very human-like performance. Better or worse than human performance is therefore not necessarily inherent to these machines.

R1.3

We realize that our wording was to some extent misleading, since we do agree with the Reviewer saying *"training using extensive reinforcement learning from human feedback procedures could potentially be used to train these machines for superior performance"*. Furthermore, also going in the same direction is that fact that recent studies as shown as fine-tuning on human data from a given task, increases the behavioral similarity between the LLM and human participants in another task that has not been seen by the LLM before. In revised manuscript, we modified the sentence/paragraph in order to better reflect what is our (and we believe the Reviewer's) position on this issue:

> We were therefore unable to find, in the considered model space, no model trained or fine-tuned resulting in average performance matching those the average performance of our standard experimental sample.

Line 32 "where"-> were

R1.4

We thank the reviewer for spotting this typo that we have corrected

Reviewer #2 (Remarks to the Author):

This ms presents a very interesting comparison between human reasoning and that observed in LLMs of different generations. The authors adapted problems from two often used measures of human irrationality, the CRT and the conjunction fallacy. Results show that there is a clear decrease in the extent to which more advanced LLMs show the same propensity to "illogical" reasoning as is typically shown in human samples. In addition, use of two prompts on the earlier models produced somewhat better results than with the human sample. These results are very interesting, particularly those showing that the latest versions are not as prone to biased reasoning. There are however some limitations to these results.

We thank the reviewer for this positive evaluation and the following helpful comments.

First, although there are many different kinds of reasoning and judgments that show biases, the authors focused on only two, with no indication why these might be more representative of any general tendency to biased reasoning. Certainly, a larger sample of problems would answer any questions as to the generality of the idea that recent LLMs are not prone to biased reasoning.

R2.1

We of course agree with the reviewer that our choice of tasks does not cover the full spectrum of decision-making tasks. We nonetheless believe that we can provide some arguments that, we hope, illustrate the strength of our approach, compared to other concurrent studies.

First, while it is true that previous studies (see Binz & Schultz, 2023) has included a greater number of cognitive tasks, it is also true that this and other studies has focused on using the very same items previously published, thus undermining their conclusions in such that the previously published items were very likely included in the training corpus of the considered models (contamination problem; and, in fact, our **Figure 1** does provide convincing evidence of the presence of a contamination problem). We opted for the harder (and we believe cleaner) way of designing and testing new versions of the included task, which allowed us overcoming the contamination problem. Additionally, while other studies (see, again, Binz & Schultz, 2023) also opted to include verbal versions of otherwise visual tasks, we avoided doing so because of additional interpretational and translational issues arising from transforming non-verbal tasks into textual material.

Second, while it is true that we included only two main tasks, it is also true we systematically manipulated the framing or the prompting strategies ("*baseline*", "*reasoning*" and "*example*"; after inclusion, for the Linda/Bill problem we are also including a "*frequency*" condition; see **R2.2**), thus leading to 7 versions of the experiments.

This being said, as already mentioned, we do agree with the reviewer that we cannot state with certaintly whether or not the conclusions drawn from our two tasks, will generalize to other tasks (even though the results for CRT and Linda/Bill went

generally in par in our data – except for DVB). This is why, following the Reviewer's remark we expanded the discussion to take into account this point:

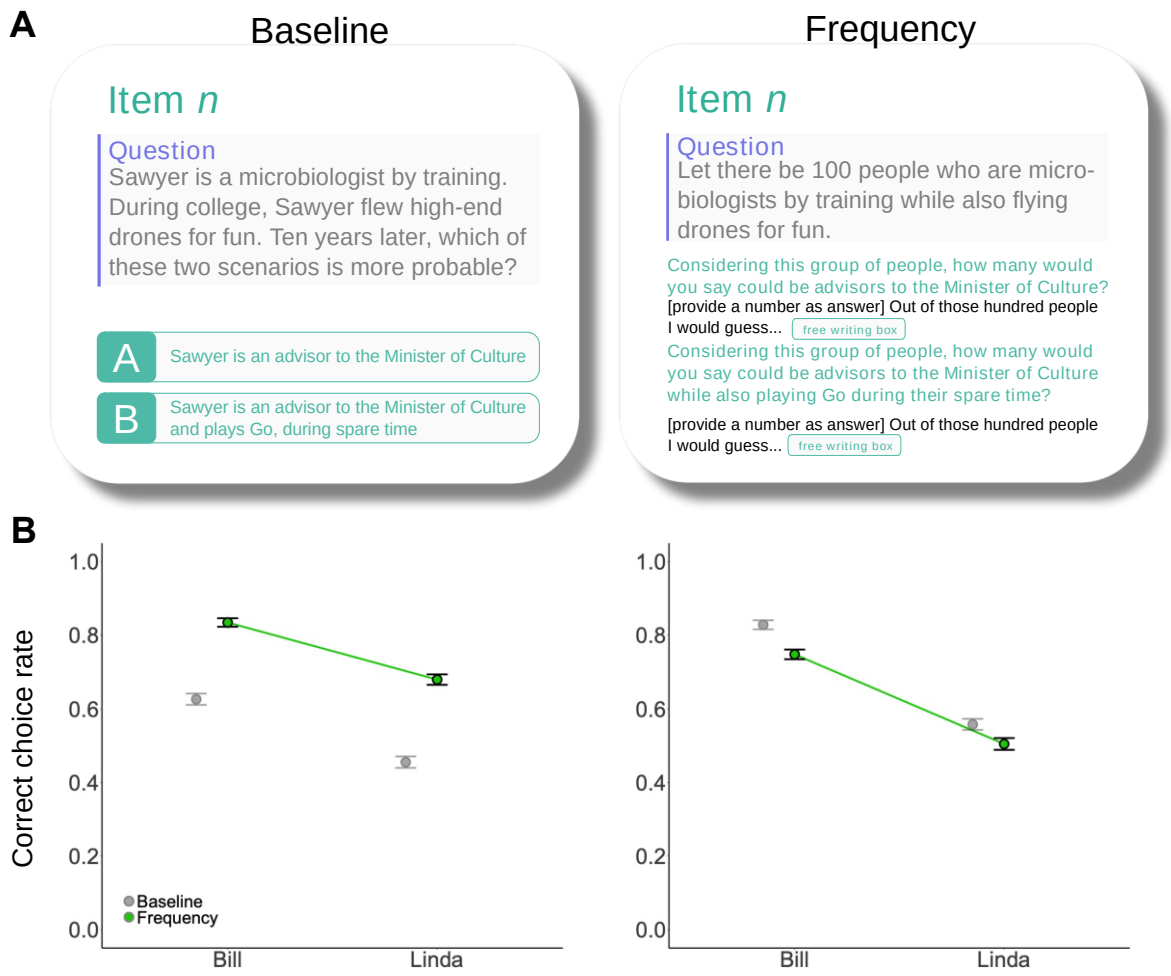
> We understand that *prima facie* the fact that we selected two reasoning tasks among many other possible tasks could seem arbitrary. However, contrary to previous studies, focusing on two tasks allowed us a more in-depth investigation into these processes, including the testing and developing of new (contamination free) tasks, and testing multiples possible variations (including the bare-boned mathematical formulations underlying reasoning, as well as different prompts).

Second, as the authors note themselves, using the same prompts with humans as with LLMs simply might not work due to the difficulty that humans have in understanding just what is meant by these, or by their difficulty in adjusting their actual behavior. There are some recent studies on debiasing reasoning in humans that might suggest a more direct way of comparing what humans do with LLMs.

R2.2

In line with our interpretation, the Reviewer here points out that the prompting strategy we labelled "*reasoning*" (i.e., framing the response in more mathematical, analytical frame) did not produce better results in humans because they failed to understand (or follow) the (implicit) instruction and wonder whether other strategies, originally developed in humans, would work in both systems. Since the Reviewer did not provide explicit references, we are not sure concerning which debiasing strategy is referring to, but after searching further the literature, we identified that a popular strategy used to modulate performance in the CRT is modulating time pressure (see many studies by Win de Neys, for example). This strategy (modulating time pressure) will not be effective (or meaningful) in LLM, because they have not obvious notion of time flow. On the other hand, a popular strategy to improve accuracy (or reduce the conjunction fallacy) in humans is framing the question in terms of frequency rather than probability. This popular result was originally put forward by Hertwig and Gigerenzer and led to a famous adversarial collaboration between the two and Daniel Kahneman. The underlying idea is that, at least partly, the conjunction fallacy arises from a wrong, intuitive, understanding of the word probable, which is corrected is participants are "forced" to think in terms of (objective) frequency.

We could not see any objection to running a "*frequency"* version of our Linda/Bill experiment and we therefore did so in both DV3 (the model where we tested the other prompting strategies) and human participants (N=100). We include the results in the revised manuscript (and see **Figure R3 b**elow). In short, we found that this strategy, originally proposed in humans, significantly improves accuracy in DV3 (significant "prompt hacking" effect: $X^2 = 7.5$, DF = 1, $p = 0.006$; Bill Baseline (0.69) vs Bill Frequency (0.9) correct response ratio Tukey contrast $p<.0001$; Linda Baseline (0.47) vs Linda Frequency (0.73) correct response ratio Tukey contrast $p<.00001$). However, somehow surprisingly, we could not replicate the improvement in our human sample (Bill Baseline (0.83) vs Bill Frequency (0.75) correct response ratio Tukey contrast $p=0.07$; Linda Baseline (0.56) vs Linda Frequency (0.50), contrast $p=0.75$).



Response length analysis in CRT. (A) *Different ways in which the Linda/Bill items were administered. Baseline: version consisting only in presenting the question/item, without any specific prompting. Frequency: version including a small verbal prompt aimed at inducing a reasoning based on actual frequencies rather than abstract probabilities. (**B**)). Correct choice rate in the Linda/Bill problems after prompting engineering (Frequency). The grey dots correspond to the accuracy in the Baseline version (results presented in **Figure 2**). Results are reported for DV3 and for human participants (N=100 subjects).*

Again, we can only speculate concerning the reason why this manipulation did not replicate original findings in human, as several differences exist between our experiment and Hertwig & Gigerenzer one. For instance, the tested only two (or few) vignettes; their experiment was performed in the lab and not online and slightly changed the formulation. Also, some differences exist between our own previous experiments ("*baseline*", "*example*" and "*reasoning*") and this one. The firsts were performed using forced choice response modality, while this one involved entering free text. We believe that the fact that the experiment is performed online and the "free text" response modality are the major factors contributing to the absence of improvement in humans in the "*frequency*" manipulation, which point to important warning concerning LLM / human comparative studies (because, on the other hand, LLMs have obviously no problem at all at answering in a "free text" modality). The new results are reported in the revised paper:

Results:

Finally, concerning the Linda/Bill vignettes, we tested an alternative framing of the experiment that previous studies in humans have shown to correct (if not eliminate) the conjunction fallacy[15]. It consists in framing question in terms actual frequency ('*out of 100 people how many are bank teller?*') rather than probability ('*which of these scenarios is more probable*?"). We deployed this experiment in DV3 found that, while it did not eliminate completely the fallacy (i.e., accuracy for the Bill items was higher compared thus of the Linda ones), it significantly increased accuracy, however the same manipulation did not have positive effect in human (**Figure S7B**)

Discussion:

Concerning the Linda/Bill experiment we also considered an alternative prompting strategy inspired by previous studies in humans showing that framing question in terms of actual frequency rather than abstract probability could correct the fallacy [15]). This manipulation was effective in increasing accuracy in the Linda/Bill task in DV3, even if it did not eliminate the fallacy (i.e., accuracy for the Bill items was still higher complared to accuracy for the Linda ones). However, this manipulation orginally designed based on its merit in human participants, did not exert a positive effect in our human sample. This was not entirely surprising, as previous attempts to capitalize on the "frequency" structure outside of the strict paradigm of its original authors (e.g., used in economic betting games), have consistently failed in reducing the fallacy [42]. This should not be regarded as a demerit of the original paradigm, as it may still be fruitful for future research efforts to ponder different alternative adaptations of the "frequency" strategy. Beyond differences between the original and our version of the "*frequency"* experiment [15], the fact that the experiment was performed online and required a "free text" response modality may have contributed to the absence of improvement in humans. Such a contrast is relevant in and of itself, as it constitutes a reminder that LLM / human cognitive comparative studies, will inevitably face experimental challenges. Namely, while in humans it is generally complicated to obtain meaningful data using open-ended response tasks (especially in online experiments), this same modality presents no issue for LLMs.If anyting, such models are specifically designed to provide open-ended responses.

Supplementary Materials

1.2 Framing questions in terms of Frequency

For Linda/Bill experiments, we complemented our prompting strategies with another alternative framing, based on the work of Hertwig & Gigerenze (1999). There, it was found that the conjunction fallacy could be diminished (and even completely corrected) by avoiding the concept of "probability", which authors argued was polysemic and hard to grasp. By framing questions in terms actual frequency (e.g., '*out of 100 people, how many would you say could be bank tellers?*') rather than in their original probability format ('*which of these scenarios is more probable...*?"), they claimed to have overcome the semantic hurdles of the classic Linda/Bill formulation, leading humans to become apparently impervious

to the reasoning fallacy. Following this line of reasoning, we deployed a variant of this experiment adapted to run in DV3, and administer it to both DV3 and humans (Figure S7). The experiment's presentation was identical to our other instances of "prompt-hacking", except that here the arborescence of potential Bill/Linda x Artsy/Sciency scenarios were introduced by saying "Let there be 100 people who are X", and asking to provide a response in terms of frequency in a free response box, for both the correct and the fallacious answer (Figure S7A). For example, in a prompt were the setup was built with the question "Let there be 100 people who are microbiologists by training while also flying drones for fun", the participant (LLMs and humans alike) was faced with two questions requiring free-text answers: (A) "How many of these people could be an assistant to the minister of culture?" and (B) "How many of these people could be an assistant to the minister of culture while also playing go on their spare time?". Answers were deemed correct when the number of people proposed for the correct answer was higher than that proposed for the fallacious answer (in the previous example A>B). We found that, while this manipulation did not eliminate completely the fallacy (i.e., accuracy for the Bill items was higher compared thus of the Linda ones), it significantly increased accuracy for DV3, but not for humans (**Figure S7B**)

Finally, the authors consider humans as being a homogenous sample. This is of course certainly wrong. At the very least, there are large individual differences in the propensity of individuals to produce biased reasoning. There are certainly a subset of people whose tendency to biased reasoning would look like that found with Chatgpt or possibly GPT-4. Unlike individual LLMs, people differ greatly in terms of their background knowledge, level of education, specific training, etc. Any attempt to directly compare the reasoning of LLMs and humans that does not consider individual differences among the latter is almost certainly misleading. Nonetheless, these results are very interesting. One possibility to make the comparison more useful would be to split the human sample into performance tiers, to make the within human variation much clearer, and also to make the real complexity of how to eventually compare LLMs with humans more evident.

R2.3

The Reviewer mentions an important point that is the fact that source of "behavioral" variability in humans and LLMs is radically different. Indeed, our cohort of human participants involved individuals that, however relatively homogeneous in term of sample, differ in terms of genetic and ontogenetic background. On the top of these stable inter-individual differences ("traits"), probably other factors introduce significant variation in behavioral performance in humans, such as psychological of physiological "states" (mood, being tired / hungry etc), which have absolutely any counterpart in LLMs. We, however, respectfully disagree with Reviewer in such that we believe that our statistical approach consisted of modeling accuracy by implementing generalized linear mixed models, with a random intercept per participant (lme4; Bates et al., 2015). We chose a hierarchical modeling approach precisely in order to account for individual differences and for unforeseen imbalances in samples across factors and levels (Jaeger, 2008). Crucially, this way of dealing with human imbalances was also the exact same approach we implemented to deal with within-LLMs (quite considerable) response variability, effectively treating variations within the human species and the "model species" in the same statistical manner. This approach had the advantage of allowing a parsimonious comparison between models and human.

On top of this, we also believe that our study has also the advantage compared to other study to take into account sources of variability that, however different, also exists in LLMs. For instance, we included a large number of models (which was further augmented upon revisions, see **R1.1**). We also implemented repeated testing in LLMs

(each question / experiment was repeated 100 times) avoiding using the deterministic response selection, adopted by many other studies in order to give a sense of the probabilistic nature of LLMs response.

Finally, following the Reviewer's advice, we included in the discussion a paragraph concerning the interpretational issues relative to different sources of variability between humans and LLM and we report here the modified paragraph:

> Our investigation also raises the question of the sources of behavioral variability in LLMs (as compared to humans). Opposite to previous studies[35–37,47] we did not set the decision temperature to be fully deterministic (1.0), as to get a sense of the variability of the models' responses. This, we believe, has allowed us to get a more realistic insight into the models' knowledge. Whil we re-iterated each question several times to match the sample sizes of our human participants, we understand that the resulting "behavioral variability" in LLMs is not directly equivalent to the behavioral variability observed in humans. We also believe that between-model variance and difference cannot really be matched to inter-individual differences in humans, since different LLMs present profound "genetic" differences (different corpora, and algorithm). As such, following up on the biological metaphor, they should rather be consider as belonging to different "species" rather than just being different "individuals". A possible, promising way to induce and study behavioural differences akin to those traditionally ascribed to human "states" and "traits" in LLMs, could be to study the behavior after specific prompts inducing changes in attitude, mood, or personaliy[73,74] .

Reviewer #3 (Remarks to the Author):

My area of expertise is in cognitive science with specific interests lie in the role of intuitive and deliberative processes in human reasoning and the veracity of dual process models as an explanatory framework for understanding human reasoning and judgment. In this regard, my comments are necessarily limited to this aspect of the paper rather than the technical details regarding the LLM model features and characteristics.

My overall evaluation of this paper is very positive. It is novel, timely and extremely interesting. The authors have approached their research question very thoughtfully, utilising a range of novel contents in their judgment tasks to eliminate potential explanations based upon direct match between LLM training materials and their task set. Their analysis is thorough and clearly presented and incorporate a broad range of LLM models.

The findings are intriguing. The evidence of similar biases amongst all but the most recent LLM models might indeed suggest that the source of certain biases in reasoning and judgment lies in a dissociation between language-based reasoning drawing on statistical regularities and mathematical computation, a conclusion that aligns well with the distinction made between dual system models of human reasoning.

The authors consider two potential explanations for their findings; biases present in the corpus or biases that arise through human feedback fine tuning. However, it remains unclear why the biases are not present in the lates Chat GPT and GPT-4 models.

Whilst the behaviour of earlier models appears to approximate participants responses on a global basis, the item-based analyses show clear areas of non-alignment and the differential impact of prompts is also surprising

These aspects of the findings are intriguing and there clearly remains much additional work to be done to develop an explanation for these divergencies as this will be crucial in determining whether this type of research can genuinely inform psychological or computational models of human reasoning and judgment. This is the most significant weakness in this paper, as it is difficult to draw any firm conclusions about the use of LLMs in informing psychological theory.

However, despite this, it is a high-quality piece of work utilising an approach that will be valuable in informing future work and potential extension to a broader range phenomenon in the heuristics and biases domain. I am consequently supportive of publication.

We thank the Reviewer for this positive evaluation of our work. We hope we are able to transform the identified weakness (impossibility to draw firm conclusions) into a strength of the paper, by thoroughly amending the manuscript (see **R1**, **R2**) in order to better highlight the inherent difficulties of human / machine comparative cognitive studies. In fact, our hope is that, beyond the specific results concerning CRT and Linda/Bill tasks, our paper will be useful to stimulate the debate and pave to way to

methodologically and epistemologically sound way of cognitive-science inspired machine behavioral analysis.

27th Feb 24Dear Dr Palminteri,

Your manuscript titled "Studying and improving reasoning in humans and machines" has now been seen by our reviewers, whose comments appear below. In light of their advice I am delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Psychology under the open access CC BY license (Creative Commons Attribution v4.0 International License).

We therefore invite you to revise your paper one last time to address the remaining concerns of our reviewers and a list of editorial requests. At the same time we ask that you edit your manuscript to comply with our format requirements and to maximise the accessibility and therefore the impact of your work.

EDITORIAL REQUESTS:

Please review our specific editorial comments and requests regarding your manuscript in the attached "Editorial Requests Table". Please outline your response to each request in the right hand column. Please upload the completed table with your manuscript files as a Related Manuscript file.

If you have any questions or concerns about any of our requests, please get in touch with Marike Schiffer (marike.schiffer@nature.com), who will handle your manuscript going forward.

SUBMISSION INFORMATION:

In order to accept your paper, we require the files listed at the end of the Editorial Requests Table; the list of required files is also available at https://www.nature.com/documents/commsj-file-checklist.pdf .

OPEN ACCESS:

Communications Psychology is a fully open access journal. Articles are made freely accessible on publication under a CC BY license (Creative Commons Attribution 4.0 International License). This license allows maximum dissemination and re-use of open access materials and is preferred by many research funding bodies.

For further information about article processing charges, open access funding, and advice and support from Nature Research, please visit https://www.nature.com/commspsychol/article-processing-charges

At acceptance, you will be provided with instructions for completing this CC BY license on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

* TRANSPARENT PEER REVIEW: Communications Psychology uses a transparent peer review system. On author request, confidential information and data can be removed from the published reviewer reports and rebuttal letters prior to publication. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons.

* CODE AVAILABILITY: All Communications Psychology manuscripts must include a section titled "Code Availability" at the end of the methods section. We require that the custom analysis code supporting your conclusions is made available in a publicly accessible repository at this stage; please choose a repository that generates a digital object identifier (DOI) for the code; the link to the repository and the DOI must be included in the Code Availability statement. Publication as Supplementary Information will not suffice.

* DATA AVAILABILITY:
All Communications Psychology manuscripts must include a section titled "Data Availability" at the end of the Methods section. More information on this policy, is available in the Editorial Requests Table and at http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf.

Please use the following link to submit the above items:
[link redacted]
** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

We hope to hear from you within two weeks; please let us know if you need more time.

Best regards,

Antonia Eisenkoeck, PhD
Senior Editor

&

Marike Schiffer, PhD
Chief Editor
Communications Psychology

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

My points are addressed well, I have no additional comment. The manuscript should be accepted for publication.

Reviewer #2 (Remarks to the Author):

I have gone over the various comments and the authors' very thoughtful responses to these. I am satisfied that the modified ms should be published.

Reviewer #3 (Remarks to the Author):

The authors have responded to the reviewers comments thoughtfully and where possible addressed them through additional analyses or further empirical work. I liked the first version of the paper. The revision represents a significant improvement. This is interesting, innovative and important work and I am supportive of publication.