Dr Joseph Stubbersfield and Dr Alberto Acerbi—written evidence (LLM0024)

House of Lords Communications and Digital Select Committee inquiry: Large language models

Human-like content biases in Large Language Models

Dr Joseph M. Stubbersfield is a lecturer in Psychology at the University of Winchester. In his research, he uses experimental methods to examine how psychological biases influence the dissemination and communication of information including misinformation and conspiracy theories.

Dr Alberto Acerbi is an Assistant Professor in the Department of Sociology and Social Research at the University of Trento, and member of the Centre for Computational Social Science and Human Dynamics. In his research he uses computational models and quantitative analysis of large-scale cultural data to examine contemporary cultural phenomena. He is author of the book *Cultural evolution in the digital age*.

This evidence is submitted in response to the government's call, so that the Communications and Digital Committee is aware of the implications of human-like content biases in LLM produced texts. Specifically, it responds to:

Question 3: 'How adequately does the AI White Paper (alongside other Government policy) deal with large language models? Is a tailored regulatory approach needed?'

Particularly question a) 'What are the implications of open-source models proliferating?'

1. LLMs are currently used, or their use has been proposed, **in journalism,**[1] **copywriting,**[2] **academia**[3] **and other writing tasks**[4] and the proliferation of open-source models will only increase the dissemination of LLM produced text in wider culture.
2. The risks of LLM texts reproducing human biases or stereotyping, particularly gender- or race-based prejudices have been broadly acknowledged[5,6] and are mentioned within the AI White Paper.

[1] Petridis, S., Diakopoulos, N., Crowston, K., Hansen, M., Henderson, K., Jastrzebski, S., ... & Chilton, L. B. (2023, April). Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-16). https://doi.org/10.1145/3544548.3580907

[2] Chen, G., Xie, P., Dong, J., & Wang, T. (2019). Understanding programmatic creative: The role of AI. *Journal of Advertising*, *48*(4), 347-355. https://doi.org/10.1080/00913367.2019.1654421

[3] Buruk, O. (2023). Academic Writing with GPT-3.5: Reflections on Practices, Efficacy and Transparency. *arXiv preprint*. https://doi.org/10.48550/arXiv.2304.11079

[4] Dale, R. (2021). GPT-3: What's it good for?. *Natural Language Engineering*, *27*(1), 113-118. https://doi.org/10.1017/S1351324920000601

[5] Lucy, L., & Bamman, D. (2021, June). Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding* (pp. 48-55).

3. However, research in social learning and cultural evolution demonstrates several psychological biases which influence the dissemination of information in culture**. In addition to stereotype-consistent information,**[7] **human memory and social transmission demonstrates biases towards content which is negative,**[8] **threat-related,**[9] **about social relationships,**[10] **and counterintuitive** (i.e., counter to intuitive expectations about physics, biology and psychology).[11] These 'content biases' have been demonstrated using both experimental methods and through analyses of cultural data, showing that **these biases are present in humans, and influence the content of wider culture**.[12]

4. Further, recent research shows that **LLMs demonstrate biases for content analogous to humans**. Across five experiments, replicating social transmission experiments in humans, text produced by **the LLM ChatGPT demonstrated content-based biases for gender stereotype consistency, negative information, threat-related information, social information, and information which is biologically counterintuitive**.[13]

5. **The issue of LLM biases for these contents is not currently addressed in the white paper. As these biases in LLM output align with biases in human psychology, they may be more difficult to recognise objectively and may have consequential downstream effects**. In particular, through processes of emotional contagion in digital media,[14] **negativity and threat bias in LLM generated material could contribute to broader negativity and overestimation of threats in culture**. Given that negativity and threat biases likely play a role in the

http://dx.doi.org/10.18653/v1/2021.nuse-1.5

[6] Cade, M. (2020, November 24). Meet GPT-3. It Has Learned to Code (and Blog and Argue). *The New York Times*. https://www.nytimes.com/2020/11/24/science/artificial-intelligence-ai-gpt3.html [archived]

[7] Kashima, Y. (2000). Maintaining cultural stereotypes in the serial reproduction of narratives. *Personality and Social Psychology Bulletin*, *26*(5), 594-604. https://doi.org/10.1177/0146167200267007

[8] Bebbington, K., MacLeod, C., Ellison, T. M., & Fay, N. (2017). The sky is falling: evidence of a negativity bias in the social transmission of information. *Evolution and Human Behavior*, *38*(1), 92-101. https://doi.org/10.1016/j.evolhumbehav.2016.07.004

[9] Blaine, T., & Boyer, P. (2018). Origins of sinister rumors: A preference for threat-related material in the supply and demand of information. *Evolution and Human Behavior*, *39*(1), 67-75. https://doi.org/10.1016/j.evolhumbehav.2017.10.001

[10] Mesoudi, A., Whiten, A., & Dunbar, R. (2006). A bias for social information in human cultural transmission. *British journal of psychology*, *97*(3), 405-423. https://doi.org/10.1348/000712605X85871

[11] Berl, R. E., Samarasinghe, A. N., Roberts, S. G., Jordan, F. M., & Gavin, M. C. (2021). Prestige and content biases together shape the cultural transmission of narratives. *Evolutionary Human Sciences*, *3*, e42. https://doi.org/10.1017/ehs.2021.37

[12] Stubbersfield, J. M. (2022). Content biases in three phases of cultural transmission: A review. *Culture and Evolution*, *19*(1), 41-60. https://doi.org/10.1556/2055.2022.00024

[13] Acerbi, A., & Stubbersfield, J. M. (2023, July 13). Large language models show human-like content biases in transmission chain experiments. *OSF Preprints*. https://doi.org/10.31219/osf.io/8zg4d

[14] Goldenberg, A., & Gross, J. J. (2020). Digital emotion contagion. *Trends in cognitive sciences*, *24*(4), 316-328. https://doi.org/10.1016/j.tics.2020.01.009

dissemination of misinformation,[15] and conspiracy theories online,[16] **they could also contribute to the role of LLMs in the spread of misinformation**.

6. When considering proportionate measures for bias detection, mitigation and monitoring, **regulators should have an understanding of how LLMs reflect human biases, including subtle biases such as towards negativity and threat-related content**, and the implications of such biases on the dissemination of information (and misinformation) in wider culture.

7. **Summary**

   i. Large Language Models (LLMs) are proposed as having potential to be used or are already being used in a wide range of writing tasks, including journalism, copywriting, academia.

   ii. Research in Cultural Evolution and Social learning demonstrates that humans are biased to attend to, remember, and transmit some types of information content over others.

   iii. In our evidence, we articulate that **LLMs show biases analogous to humans for content that is gender-stereotype consistent, negative, threat-related, about social relationships, and biologically counterintuitive.**

   iv. The presence of these **biases in LLM output may be difficult to detect, and has potential to magnify human tendencies for appealing, but not necessarily informative, or valuable, information content and could contribute to broader negativity and overestimation of threats in culture, and the appeal of online misinformation.**

September 2023

[15] Acerbi, A. (2019). Cognitive attraction and online misinformation. *Palgrave Communications*, *5*(1). https://doi.org/10.1057/s41599-019-0224-y

[16] Youngblood, M., Stubbersfield, J. M., Morin, O., Glassman, R., & Acerbi, A. (2021, October 26). Negativity bias in the spread of voter fraud conspiracy theory tweets during the 2020 US election. *PsyArXiv*. https://doi.org/10.31234/osf.io/2jksg