Lara Dal Molin

Large Language Models as Gender-Shaping Infrastructures

Abstract

This paper engages with questions emerging from my proposed PhD project, due to start in October 2022. The project explores the design of Large Language Models (LLMs), Artificial Intelligence (AI) technologies that generate text. While LLMs are extremely sophisticated, to the point where their outputs are indistinguishable from human-generated text, they also reportedly display stereotypical language (Luitse and Dankena, 2021). An example of this may be completing the sentence "the woman worked as" with "a babysitter" (Sheng et al., 2019). This dynamic inspires questions regarding the infrastructural characteristics of LLMs, their implications with respect to the perpetuation of gender bias and exclusion, and their governance. What infrastructural processes determine stereotypical languages in LLMs? Who are the key stakeholders in such processes, and do they coincide with those who are most affected by such stereotypes?

LLMs, predominantly built by the company OpenAI, are still at research stages (Radford et al., 2018; Radford et al., 2018; Brown, 2020). This factor contributes to the significance of the outlined questions for two reasons. Firstly, the use of LLMs is currently limited to paying customers of OpenAI, resulting in the gatekeeping of design-related decisions. Secondly, however, it may be early enough to successfully adjust the design of LLMs before deployment. My proposed doctoral research offers Design Justice, a framework proposing participatory technology design processes led by marginalised communities, as a candidate solution for mitigating gender bias in LLMs (Costanza-Chock, 2020). Additionally, I propose algorithmic audits as methods to be applied within Design Justice, and as means to assess the performance of LLMs while considering the experiences of minority-gender communities (Brown, Davidovic and Hasan, 2021). As technology emerges as a major contemporary avenue for the reinforcement of gender-based oppression, addressing questions on infrastructure and governance may inspire human-centric approaches to technology design.

References

- Brown, S., Davidovic, J. & Hasan, A., 2021. The Algorithm Audit: Scoring the Algorithm that Scores Us. Big Data & Society, 8(1), pp. 1-8.
- Brown, T. et al., 2020. Language Models are Few-Short Learners. s.l., Advances in Neural Information Processing Systems 33 (NeurIPS 2020).
- Costanza-Chock, S., 2020. Design Justice: Community-Led Practices to Build the Worlds We Need. Cambridge, Massachussetts, United States: The MIT Press.
- Luitse, D. & Denkena, W., 2021. The Great Transformer: Examining the Role of Large Language Models in the Political Economy of AI. Big Data & Society, 8(2), pp. 1-14.
- Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I., 2018. Improving Language Understanding by Generative Pre-Training, San Francisco, California, United States: OpenAI.
- Radford, A. et al., 2018. Language Models are Unsupervised Multitask Learners, San Francisco, California, United States: OpenAI.
- Sheng, E., Chang, K.-W., Natarajan, P. & Peng, N., 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. Hong Kong, China, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing.