Summary Post: Balancing Ethics and Innovation in Technology

by Natali Nikolic - Wednesday, 15 January 2025, 8:17 PM

https://www.my-course.co.uk/mod/forum/discuss.php?d=271299#p517932

If society wants to leverage AI's (full) potential, e.g. in form of Large Language Models (LLMs) and other AI technologies, while at the same time being able to manage uncertainties and risks, then ethical innovation and societal adaptability is critical in achieving this.

Rodrigo offered an additional perspective, supported by concrete examples, that added value to my initial post. One notable example he gave was that of a lack of quality control and the tool in question was Air Canada’s virtual assistant that led to reputational damage (Olavsrud, 2024). Furthermore, in his opinion fears of catastrophic AI outcomes, such as job loss or those illustrated by "ChaosGPT,", are often overstated (Zhan et al. 2024). Rodrigo therefore concludes that regulating AI systems should aim to reduce the consequences of failures. At the same time, in order to avoid a fear-driven overreach or unnecessary centralization, the capabilities of AI need to be understood as well.

Stuart contributed to the discussion by highlighting that LLMs can be exploited to produce harmful content by using harmless keywords and then text substitutions, effectively bypassing existing safety measures (Bianchi and Zou, 2024). Fine-tuning has the potential to easily enable harmful content generation, even with few examples, while ethical fine-tuning may unintentionally compromise safety (Qi et al., 2023). Despite these challenges, LLMs are great tools that offer the possibility to detect inaccuracies and explain communicative manipulation tactics comprehensively (Chen and Shu, 2024; Jones, 2024).

The conclusion is one that seems to be the current consensus in many literary sources that I have read since the beginning of this AI course in early 2024 and it doesn't just apply to AI writing, but AI applications in general: Leveraging AI's potential while managing its risks requires a balanced approach that prioritizes ethical innovation, effective regulation, and societal adaptability.

References

- Bianchi, F. and Zou, J. (2024) 'Large Language Models are Vulnerable to Bait-and-Switch Attacks for Generating Harmful Content'. arXiv. Available at: https://doi.org/10.48550/arXiv.2402.13926.
- Chen, C. and Shu, K. (2024) 'Combating misinformation in the age of LLMs: Opportunities and challenges', AI Magazine, 45(3), pp. 354–368. Available at: https://doi.org/10.1002/aaai.12188.
- Jones, D. G. (2024) 'Detecting Propaganda in News Articles Using Large Language Models', Engineering: Open Access, 2(1), pp. 01–12. Available at: https://doi.org/10.33140/EOA.01.02.10.
- Olavsrud, T. (2024) 12 famous AI disasters. Available from: https://www.cio.com/article/190888/5-famous-analytics-and-ai-disasters.html [Accessed 16 December 2024].
- PermalinkShow parentReply
- Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P. and Henderson, P. (2023) 'Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!' arXiv. Available at: https://doi.org/10.48550/arXiv.2310.03693.
- Zhan, E.S., Molina, M.D., Rheu, M. & Peng, W. (2024) What is There to Fear? Understanding Multi-Dimensional Fear of AI from a Technological Affordance Perspective. International journal of human-computer interaction 40(22): 7127-7144. DOI: https://doi.org/10.1080/10447318.2023.2261731