Ethical Evaluation of the OpenAI API: The Importance of Understanding User's Linguistic Environment, Culture, and Context

Background

OpenAI has developed and is releasing a new API providing a 'general-purpose "text in, text out" interface' (https://openai.com/blog/openai-api/). In other words, the API would understand the context of a question, and find answers on the chosen website page without the presence of similar keywords. Here are some examples of its application: https://beta.openai.com/.

AI-powered tools are known to be biased. For example, gender and race biases might be seen in text sentiment analysis AI (Kiritchenko & Mohammad, 2018). The study proposed here aims to find some solutions while 'researching safety-relevant aspects of language technology (such as analysing, mitigating, and intervening on harmful bias)', as well as addressing the responsibility of the API's outcomes depending on context.

Goals

The outcome of the research should be concrete ethical recommendations and solutions in the context of the OpenAI API, and lessons learned should be applicable to other products, and different institutions.

Profile

Enthusiasm for biases in AI, ethics, fairness, and linguistics.

Contact

If interested please contact Auxane Boch at auxane.boch@tum.de or contact the IEAI Office ieai@mcts.tum.de.

Reference

Kiritchenko, S., & Mohammad, S. M. (2018). Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.