Bias and Fairness in Artificial Intelligence and Machine Learning

Liviu-Ștefan Neacșu-Miclea
Specialization: ICA
Group: 246/2

Introduction

Artificial Intelligence has become a hot topic in the last few years. The emergence and accessibility of generative tools like Stable Diffusion and Large Language Models has raised popularity of AI not only in science communities, but also among users as well. Consequently, the impact of occurrences in AI outputs of what may be thought as biased or unfair judgement is not to be ignored as it could have devastating outcomes [2].

While bias in solutions handling tasks like object detection, prediction and classification may be seen as pure technical flaws in the system, bias in social applications employing AI technology is a subject that needs intensive care and delicacy, as it may be met with indignation and subjective reception and interpretation. Researchers who try to identify prejudice tendencies in AI models in order to fight against bias have developed a number of methods in data collection, analysis and model training techniques that help building AI products that promote fairness and equitable behavior towards a diversity of social groups [2].

Bias sources

In the first place, it is important to understand the causes that are at the root of bias apparition in AI models. One of the problems lie in the data used for training, where **dataset imbalances** lead to under or overrepresentation of certain categories [2]. A dataset failing to represent the diverse nature of its target application may affect its accuracy and reliability. For example, it was observed that AI medical utilities have lower diagnostic accuracy across different demographic groups, as well as racial biases can influence decisions in justice systems [2]. Moreover, it was noted that women are highly underrepresented in a variety of social contexts, sometimes gender ratio can encounter values of 4:1 [1]. Thus, the influence of women's perspective in a dataset can be limited by the quantity of collected samples, like the number of female authored posts and comments in a LLM training text corpus [1].

On the other hand, it was deduced that part of bias in AI models simply reflect the human natural tendency to prejudice, like, for instance, a dataset containing scrapped unfiltered Reddit comments can reveal an outrageous amount of discrimination, inequality, gender objectification and biased assumptions [1]. Obviously, training a LLM on such a database would raise the chance of the model emulating the **societal bias** taking form of the commenters' inadequate behaviors. For this matter, cleaning the dataset of discriminative samples is a crucial step in model bias prevention [1].

A well assembled dataset may, however, not be enough to fully ensure fairness, since the **flawed design** of models and algorithms could also be susceptible to bias [2]. For example, algorithms might give more importance to certain features during the decision process or assume facts that might put marginalized groups at disadvantage, like, for example, when people of darker skin color have higher chances to be convicted [2].

Countering bias

A certain type of fairness-aware machine learning has been designed to *detect* and *mitigate* biases in AI tools in order to *promote equity* across various demographic groups [2]. Some methods focus on ensuring balanced data representations, like **disparate impact removal** techniques or **fair representation**

methods, which are preprocessing steps purposed to mitigate bias in the dataset before training [2]. In other words, supplementing the dataset by enriching details that represent the minority could help the model make abstract of possible features that can induce bias.

Other studies have also proposed new learning techniques and algorithms designed to combat this issue. One example is **adversarial debiasing**, which involves coupling the training model with a regularizing entity that tries to predict the attribute that is prone to bias (e.g. gender, race), penalizing the model for making biased decisions [2]. Last but not least, **post-processing** steps may be employed to adjust the fairness of the model output.

While such bias countermeasures proved effective in real world applications, like mitigating gender bias in hiring processing, or racial particularities in medical diagnostics, such strategies can sometimes backfire. Interestingly, it was noted in [1] that when a generative image model was asked to create a picture of an Asian violinist, the output was satisfactory, however, when the query was to generate a picture of an African violinist, the image depicted him accompanied by a lighter skin person. This is thought to be the result of bias mitigation techniques adopted by the model's developers in order to combat racial discrimination [1].

Ethics and legal implications

A set of ethical principles has been established for AI guidance in social applications, such as: **fairness**, defined as the absence of bias; **transparency**, in the way that model design should be clearly understandable and easy for the decision process to be interpreted by humans such that the solution would earn a better trust; **accountability**, meaning that developers should be aware of and respond for the consequences and fix possibly harmful occurrences; **privacy**, governing data collection, user anonymity and ensuring user rights to opt out of the process or choose their preferences; **inclusivity**, stating that developing AI solutions should involve actors from a variety of backgrounds, such that technology fulfills the needs of all communities; **sustainability**, regarding the feasibility of energy costs of AI systems.

Laws have been adopted to ensure the fairness of AI technologies. The General Data Protection Regulation (GDPR) promotes fairness by integrating human factor into important decisions. Also, the EU's Artificial Intelligence Act requires to evaluate risks and provide transparency of AI systems in order to prevent bias propagation, as a measure for ensuring public safety and protecting human rights [2].

Conclusion

The existence of any sort of social inequality in AI models may originate from underrepresentation of certain marginal groups in development teams, biased observations in datasets, or connections established by the training process. However, various methods have been designed for identifying and mitigating biases, and studies are still on-going to address the fair and well-intended use of AI in our society.

References

[1] Liu, Yuhan. (2024). Unveiling bias in artificial intelligence: Exploring causes and strategies for mitigation. Applied and Computational Engineering. 76. 124-133. 10.54254/2755-2721/76/20240576.

[2] Onebunne, Amaka & Alade, Bolape. (2024). Bias and fairness in AI Models: Addressing disparities in machine learning. International Research Journal of Modernization in Engineering Technology and Science. 6. 1921-1934. 10.56726/IRJMETS61692.