SIT 799

Task 3.1: Mini literature review on bias and discrimination in AI

1. Introduction

AI algorithms have impact on every sphere of our lives, from banking [1] to employee onboarding [2], AI algorithms have started a new revolution in automating mundane tasks. While adapting AI algorithms to do such routine tasks is turning out to highly efficient practice for businesses, there is a downside to it. AI algorithms inherit the bias from the data, this implies that AI algorithms are prone to the same biases which a human would have. This poses a threat of partial and prejudiced decision making by these algorithms. In other words, these algorithms tend to favour a particular group of people or are skewed towards a specific subgroup hence paving way for *bias* and *discrimination* in AI.

This survey majorly focuses on highlighting and understanding the concept of bias and its origins, while also exploring some approaches to handle bias. Thus, this survey aims to create a general understanding of practices that might lead to a biased AI solution and offers approaches being used across AI community as a safeguard against these situations.

2. Literature Survey

To explore how bias can affect user experience and promote discriminatory behaviour, [3] present a case study of employee shortlisting where job postings were automated, and candidates would receive job postings advertised by an AI solution. The aim of this solution was to perform this automation in a gender agnostic manner, although the solution was evidently biased in jobs that were advertised. The jobs were grouped according to genders and specific jobs that strengthen the stereotypes against women like receptionist would be shown to younger women whereas high paying jobs that are stereotyped for males would be preferred for men, this not only rejected valuable candidates based on their gender but also encouraged discriminatory hiring practices at a workplace.

Another glaring and more threatening example of implications of a biased AI was shown in [4] where an AI solution for predicting future criminals was used and it was biased against black individuals. Moreover, this solution was used by several judges to award pretrial sentences, which led to multiple biased decisions. Face recognition systems [5], recommendation systems [6] and even Automatic Speech Recognition systems [7] have displayed evidence of bias. To examine this, the following section explores the concept of bias and highlight possible sources which induce this bias into AI solutions:

2.1 Understanding bias

Generally, AI algorithms are fuelled by data driven approaches, these approaches require hand labelled data which is responsible for the functioning of these algorithms. These handcrafted data sources lead to several types of biases:

1. **Representation Bias**: This bias exists due to lack of access to a data resource which represents every subgroup of the society. This leads to the AI model being trained on selective groups which lead to *Representation Bias.* In other words, the data sampling is the source of bias and lack of diversity in the dataset accounts for biased AI solutions. Such data sources lead to biased AI solutions like [5] and [7].
2. **Generalization bias:** This type of bias creeps in the AI solution when the model starts picking up traits which should not be associated with an individual because they belong to a specific subgroup of society. In [8], authors show that how machine learning algorithms can sometimes pick up hidden traits from data which yield in generalisations about an individual

based on the ethnic group or gender they belong. This is sometimes also referred to as *Historical Bias.* [9] also show an example scenario where HIV is linked to homosexual or bisexual males.

3. **Linking Bias:** This bias arises when machine learning model begins to interpret user's behaviour based on the interactions/connections from their social media accounts. This bias can be seen in recommendation engines deployed at big tech ecommerce platforms and social media platforms when unsuitable content is served to a user based on their internet interactions [10].

Apart from these biases, some biases can be algorithm induced, this can result from organic user interactions with the AI solution. This might lead to scenarios where AI solution can become a victim to fake users/bot revies and starts preferring popular content as opposed to organic content [11].

3. Conclusion

For future developments and sustainable use of AI it is paramount that safeguards against biased AI solution is employed. [12] categorise algorithms which are used to target biases. Apart from algorithmic measures, legislative measures such as AI-WATCH, an organisation ideated by European Union to monitor impact of AI and robotics on society. Also, companies like IBM are building tools like AI Fairness tools which help in creating benchmark for evaluating algorithms on terms of fairness. Moreover, tools such as Aequitas [14] are also democratising benchmarking standards for researchers and developers to employ best practises to handle biases in their data and AI solutions.

This survey shows why ethical standards should be maintained via businesses during designing an AI solution. Both legislative and technological frameworks should be employed to create a robust infrastructure to tackle bias in AI solutions.

4. References

[1] A. Mukerjee, R. Biswas, K. Deb, and A. P. Mathur, "Multi-objective evolutionary algorithms for the risk-return trade-off in bank loan management," *Int. Trans. Oper. Res.*, vol. 9, no. 5, pp. 583–597, 2002.

[2] M. Bogen and A. Rieke, "Help wanted: an examination of hiring algorithms, equity, and bias," *Upturn*, 2018.

[3] A. Lambrecht and C. Tucker, "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads," *Manage. Sci.*, vol. 65, no. 7, pp. 2966–2981, 2019.

[4] J. Angwin, J. Larson, L. Kirchner, and S. Mattu, "Machine bias," *ProPublica*, 23-May-2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. [Accessed: 10-Aug-2022].

[5] I. D. Raji and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society - AIES '19*, 2019.

[6] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims, "Recommendations as treatments: Debiasing learning and evaluation," *arXiv [cs.LG]*, 2016.

[7] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," *arXiv [eess.AS]*, 2021.

[8] H. Suresh and J. V. Guttag, *A framework for understanding unintended consequences of machine learning*. 2019.

[9] "HIV surveillance," *Cdc.gov*, 11-Jul-2022. [Online]. Available: https://www.cdc.gov/hiv/library/reports/hiv-surveillance.html. [Accessed: 10-Aug-2022].

[10] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, "Social data: Biases, methodological pitfalls, and ethical boundaries," *Front. Big Data*, vol. 2, p. 13, 2019.

[11] L. Introna and H. Nissenbaum, "Defining the Web: the politics of search engines," *Computer (Long Beach Calif.)*, vol. 33, no. 1, pp. 54–62, 2000.

[12] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, 2021.

[13] J. S. Fulda, "AI watch: Data mining and the Web," *ACM SIGCAS Comput. Soc.*, vol. 28, no. 1, pp. 42–43, 1998.

[14] P. Saleiro *et al.*, "Aequitas: A bias and fairness audit toolkit," *arXiv [cs.LG]*, 2018.