Bias and discrimination in artificial intelligence

Joshua Fitch
Department of Computer Science
University of Durham
Durham, United Kingdom
mccb22@durham.ac.uk

I. Section 1 – "A Survey on Bias and Fairness in Machine Learning"

The authors of [1] reviewed multiple studies researching the effects of algorithmic bias, identifying two possible sources of unfairness in machine learning outcomes – those caused by biases in the data, and those from the algorithms. Alongside outlining between sources of bias, the review summarises different definitions of bias, fairness, and discrimination, and how all three concepts are connected within artificial intelligence (AI). Using real-world examples of harmful biases in AI systems, like in judicial systems and facial recognition, the authors reiterate the need to reduce inadvertent discrimination. They state that in order to address the algorithmic bias issue, knowledge of where these biases come from is very important to enable prevention.

Throughout the review, the difficulty of addressing bias and discrimination in a uniform way is shown. Studies such as [2] show incompatibility between the conditions of calibration and balance of positive and negative classes, which shows the need to apply fairness definitions with care and in the right context. The review separates fairness definitions into three types – individual, group, and subgroup – and each type needs different ways of addressing bias within them. By taking definitions used by other studies, the review assembles a full list of general types of bias, discrimination, and fairness, allowing researchers to identify the type of discrimination and bias they need to address in their studies, and what methods can be used to mitigate these biases. In this way, [1] encourages researchers to think deeper, and identify and apply methods to prevent bias in their future studies, as opposed to just providing solutions to AI fairness issues that cannot be widely applied.

As previously mentioned, [1] does not suggest many original methods to mitigate discrimination in AI, instead collecting bias mitigation methods proposed in other studies. General methods for fair practices are suggested, such as attaching datasheets including information on dataset creation methods and skews, and labels to categorise data. The authors of [1] then describe practices in more focused areas, describing specialist responses aiming to achieve fair machine learning, fair representation learning, and fair NLP, which includes word embedding and machine translation. One example is the use of sentence tags to identify the gender of the speaker in machine translation, so words commonly associated with a specific gender are translated correctly for both genders[4]. It also draws attention to new methods to detect bias, such as CLAN [3] that helps detect bias in online social networks. Furthermore, throughout the paper, areas requiring further research have been highlighted, supporting its role in educating the wider scientific community on how to reduce discrimination in AI.

Overall, it is my opinion that this paper is effective in helping mitigate discrimination in AI through its use of real-world applications and examples of successful fairness methods used by other researchers. The authors compilation of effective methods along with identifying areas requiring further research is useful and easily applied to new projects, encouraging readers to actively apply bias-limiting methods to their own research.

II. Section 2 – A Discussion on the Future of Discrimination-Aware AI

The growing public interest in AI bias and awareness of data protection has encouraged developments in discrimination-aware AI. Widespread investment in AI systems, in sectors such as the judicial system and job recruitment, highlight the need to maintain trust in AI, as loss of trust may lead to reluctance to adopt AI-driven services and products [5].

As a result, wide-ranging definitions of fairness and discrimination have been produced, which is significant as it enables researchers to develop their own solutions based on definitions that fit their own research focuses and skews. Yet, current fairness definitions are not always helpful, and may harm analysis over time for sensitive groups [6]. Current variation in definitions make general application of one fairness solution for all systems unrealistic. As a result, the authors of [1] stress that creating a common definition of fairness within discrimination-aware AI is very important in the future. Furthermore, [1] identifies several areas, like community bias detection at the subgroup level, and language modelling at an individual level, that have had limited research, and could benefit from further development. Like the authors of [6], we would expect future work to address specific types and levels of bias mitigation, and make it clear what problems each bias-mitigation method is addressing. Also, complexity requires attention. A 2015 survey commented on the recent concentration of research on binary classification, a relatively simple topic that doesn't involve more complex learning scenarios [7]. As algorithms correcting indirect discrimination in AI are developed, current methods will need to adapt to address bias problems within multi-class classification, together with other complicating characteristics like noisy input data.

Developments in discrimination-aware AI are very positive – increasing real-world applications of fair AI will help to eliminate discrimination and bias in machine learning. Although effective, current solutions are not widely applicable, because most are specific and non-transferable. However, public trust in the honesty and reliability of AI may have already been lost. Authors of [5] emphasise that the Chain of Trust is significantly harder to rebuild once broken, and as a result, developments in discrimination-aware AI must be effective and transparent to maintain a solid level of trust in the future.

References

[1] N. Hehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning." arXiv preprint arXiv:1908.09635v2 (2019).

[2] J. Kleinberg, S. Mullainathan, and M Raghavan, “Inherent trade-offs in the fair determination of risk scores.” arXiv preprint arXiv:1609.05807 (2016).

[3] N. Mehrabi, F. Morstatter, N. Peng, and A. Galstyan, “Debiasing Community Detection: The Importance of Lowly-Connected Nodes.” arXiv preprint arXiv:1903.08136 (2019).

[4] E. Vanmassenhove, C. Hardmeier, and A. Way, “Getting gender right in neural machine translation.” *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. (2018) 3003–3008.

[5] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. Gonzalez Zelaya, and A. van Moorselk, “The relationship between trust in AI and trustworthy machine learning technologies.” arXiv preprint arXiv:1912.00782v2 (2019).

[6] H. Suresh, and J. V. Guttag, “A Framework for Understanding Unintended Consequences of Machine Learning.” arXiv preprint arXiv:1901.10002v3 (2020).

[7] I. Žliobaité, “A survey on measuring indirect discrimination in machine leanring.” arXiv preprint arXiv:1511.00148v1 (2015).