Bias in Artificial Intelligence Essay

chpf93
Department of Computer Science
Durham University
Durham, United Kingdom
chpf93@durham.ac.uk

I. AI Fairness 360

AI Fairness 360 is a toolkit developed by IBM to examine and mitigate bias within AI systems [1]. This paper, and the toolkit provided with it, are important contributions to fair AI because it presents a solution for those without specialised knowledge to be able to visualise and explore the impact of bias. It also provides a clear and useful set of evaluation tools and algorithms that can be used in industrial settings to improve the fairness of machine learning which allows these bias mitigation algorithms that have been researched to be used in real-world applications.

By providing a clear way to evaluate the effects of these mitigation techniques it enables fairness algorithms to be compared in a systematic way enabling researchers to further build on these algorithms to continually improve fairness techniques. The toolkit includes comparable metrics and 9 mitigation algorithms covering pre-processing, in-processing, and post-processing.

An important factor is that the toolkit is open source. This further amplifies the effect that this contribution has because it overcomes the requirement for specialised expertise to implement fair machine learning which could be too expensive for small and medium enterprises that want to utilise fair AI as well as speeding up the process of integrating these fair algorithms into a variety of applications quickly.

Further to this, the paper has raised significant awareness of issues within bias in AI. The interactive web application, with preset data sets, allows anybody with an interest in AI to consider real-world cases where there is a significant impact of disparity caused by bias. However, rather than just highlighting these, it presents practical solutions that can be integrated to reduce the bias and then compare the effects. This is important because it shows that this bias does not have to be accepted and can be mitigated instead.

Ultimately, the impact that this paper has on fairness in AI is very significant because it has industry backing - by IBM - and it provides accessible tools that can be used across AI to evaluate the impact of bias as well as offering methods to reduce the impact of the bias.

II. The Future of Fair Machine Learning

Algorithmic bias is a significant issue that compounds on a distrust of artificial intelligence systems [2]. It is important that bias is reduced sufficiently in machine learning in order to facilitate an environment of trust in these algorithms. Currently there is little incentive for those in industry to ensure that their systems are truly fair. Regulation of these algorithms will likely have an important role in the future; an independent report for the UK government found that these algorithms present a new challenge for regulators and guidance from lawmakers is needed with regards to the lawfulness of bias mitigation [3] and recently, the European Union announced plans to ban high-risk AI uses which is significant [4].

Without financial incentives or legal frameworks in place to compel businesses into ensuring that these systems are fair, the progress of integrating fair machine learning will be slow. This is because industry will naturally favour maximising their profit using machine learning rather than considering the impact that this has on society and currently there can exist a significant trade-off between accuracy and fairness as noted in lectures which could affect profits generated. I believe that this issue needs to be resolved before there can be widespread adoption of these mitigation techniques and research indicates that it is possible to create algorithms with little to no trade off - and in some cases improvement - of the overall accuracy can be achieved by debiasing [5].

Technology companies such as Google and Amazon also have to play an important part in integrating fairness techniques. Amazon provides a range of AI services through its Amazon Web Services platform. As such, any changes to enhance fairness that it makes would have a profound impact. It is also researching fairness, one example is a recent paper considering fairness in Bayesian optimisation [6]. Similarly, Google produces a range of research on important topics such as through its Responsible AI publications. One example is finding ways to create scalable fairness solutions that can be used in industry [7] which would be a significant step forward as AI providers hold massive influence in this area. However, Google recently fired a leading ethics in AI researcher due to a dispute with the conclusions of their research [8] which could be troubling for the future of ethics research in general.

In conclusion, without incentives or regulatory action it will be difficult to compel businesses to adapt their plans for AI sufficiently to lead to widespread adoption of bias reducing algorithms. As well as this, bias mitigation algorithms need to be further developed to reduce their impact on the accuracy of models before they will be more widely used. However, I believe that it is inevitable that bias mitigation and reduction algorithms will continue to grow and lead to greater inclusivity within AI systems.

REFERENCES

[1] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *CoRR*, vol. abs/1810.01943, 2018. [Online]. Available: http://arxiv.org/abs/1810.01943

[2] H. Hagras, "Toward human-understandable, explainable ai," *Computer*, vol. 51, no. 9, pp. 28–36, 2018.

[3] Centre for Data Ethics and Innovation. (2020, November) Review into bias in algorithmic decision-making. [Online]. Available: https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making/main-report-cdei-review-into-bias-in-algorithmic-decision-making

[4] "EU artificial intelligence rules will ban 'unacceptable' use," Apr 2021. [Online]. Available: https://www.bbc.co.uk/news/technology-56830779

[5] C. Sweeney and M. Najafian, "Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 359–368. [Online]. Available: https://doi.org/10.1145/3351095.3372837

[6] V. Perrone, M. Donini, K. Kenthapadi, and C. Archambeau, *Fair Bayesian optimization*. [Online]. Available: https://www.amazon.science/publications/bayesian-optimization-with-fairness-constraints

[7] C. Xu, C. Greer, M. N. Joshi, and T. Doshi, "Fairness indicators demo: Scalable infrastructure for fair ml systems," 2020.

[8] British Broadcasting Corporation. (2020, December) Timnit Gebru: Google staff rally behind fired AI researcher. [Online]. Available: https://www.bbc.co.uk/news/technology-55187611