Bias in AI Essay

I. Introduction

This essay demonstrates our point of view on the overall subject of algorithmic bias. First, we discuss the justifications on the reasons that bias in AI-based solutions should be addressed. Second, we demonstrate various ways to measure the fairness of a dataset and algorithm. Third, we discuss different ways to mitigate algorithmic bias. Last of all, we discuss what we expect to see in the fair machine learning solutions in the future.

II. Justifications on the reasons that bias in AI–based solutions should be addressed

One reason to address bias in AI-based solutions is to ensure that the decisions made by the solutions do not reflect discriminatory behaviour toward certain groups or populations, as stated by [13]. To justify this, they highlight the canonical example of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software, which was found to more likely assign a higher risk score (of recommitting another crime) to African-American offenders than to Caucasians with the same profile[1].

Another reason that bias in AI-based solutions should be addressed is to avoid perpetuating any systemic discrimination, under a misleading veil of data-driven objectivity, as said by [9]. To justify this, they point to the broader debate concerning disparate impact, which is discussed extensively by [3]. They highlight 'redlining', which is the refusal of opportunities to people base solely on their zip code, for loans as a classic example of disparate impact.

III. Ways to measure the fairness of a dataset and algorithm

A. Ways to measure the fairness of a dataset

One way to measure fairness in a dataset is to address Simpson's Paradox [6]. This can be done by comparing the regression for the entire population, regressions for each subgroup, and the unbiased regression. An example arose in [5], where it seemed like there was bias toward women in graduate school admissions, but, at the same time, women also had an advantage over men in some cases, such as when the data was separated and analysed over the departments.

Another way to measure fairness in a dataset, which is traditional and statistical, is selection bias [16]. Selection bias can be measured by calculating the divergence of the probability distribution over the space of inputs in the training data against the true data distribution.

[1] propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Other measures of fairness in a dataset include historical bias and representation bias. As described by [16], historical bias can be measured by evaluating the representational harm — such as reinforcing a stereotype — to a particular identity group, while representation bias can be measured by calculating the percentage a minority group makes up of the true distribution.

B. Ways to measure the fairness of an algorithm

We can use definitions of fairness, many of which have been compiled by [13], as ways to measure the fairness of an algorithm.

One measure is *Fairness Through Unawareness*, which says that '*An algorithm is fair as long as any protected attributes are not explicitly used in the decision-making process*' [10], [12]. Protected attributes are specified in the Fair Housing and Equal Credit Opportunity Acts (FHA and ECOA) [7], and include race, colour, national origin, religion, sex, and more.

Another measure is *Demographic Parity*. This states that the likelihood of a positive outcome should be the same regardless of whether the person is in the protected group [17]. In mathematical terms, '*A predictor $\hat{Y}$ satisfies demographic parity if $P(\hat{Y}|A=0) = P(\hat{Y}|A=1)$*', where A is a protected attribute and $\hat{Y}$ is a binary predictor.

Closely related are *Equalised Odds* and *Equal Opportunity*. The equalised odds definition states that the protected and unprotected groups should have equal rates for true positives and false positives (i.e., $P(\hat{Y}|A=0, Y=y) = P(\hat{Y}|A=1, Y=y), y \in 0, 1$); while, the equal opportunity definition states that the protected and unprotected groups should have equal true positive rates (i.e., $P(\hat{Y}|A=0, Y=1) = P(\hat{Y}|A=1, Y=1)$) [11].

Additional metrics measure disparate impact specifically, which are described in [20] and summarised by [9]. The measure recommended is the *Mean Difference* divided by a normalisation constant, which measures the difference between the means of the targets of the protected group and the general group, where no difference indicates no discrimination. The mean difference can be modified to obtain a *Conditional Mean Difference*, which accounts for distributional differences between the protected populations and the overall population.

To tackle the problem of discrimination in data mining in a rule-based setting, [14] introduces the notion of discriminatory classification rules as a criterion to identify the potential risks of discrimination. (Strong) α-protection is one such criterion that measures the discriminatory power of rules which occur in approaches such as decision trees and rule-based classifiers. An example of a potentially discriminatory (PD) rule they give, in the context of the German credit dataset [8], is as follows:

```
personal_status=female div/sep/mar
    savings_status=no known savings
    ==> class=bad,
```

which contains the PD *attribute personal_status*.

IV. Ways to mitigate algorithmic bias

There are a number of ways to mitigate algorithmic bias, which have also been compiled by [13].

Methods to discover Simpson's paradoxes in data automatically have been proposed by [1] and [2], while [11] proposes a criterion to satisfy equalised odds and equal opportunity in supervised learning.

Another way to mitigate algorithmic bias is mentioned in [9], which is to introduce augmented cost functions during the model training phase. In [18] and [19], both augment a standard log-likelihood loss function with a fairness regulariser which takes into account differences in how the learning algorithm classifies protected and non-protected classes.

The final way of mitigating algorithmic bias we mention is to enforce fairness criteria, such as *Independence*, *Separation*, and *Sufficiency*.

V. What we expect to see in the fair machine learning solutions in the future

Recently, researchers have begun introducing tools that can assess the amount of fairness in a tool or system [13]. One example is Aequitas [15], which lets users test models for different population subgroups. Others include IBM's AI Fairness 360 [4] and Google's What-If Tool[2]. We expect to see more of these tools in the fair machine learning solutions in the future.

References

[1] Nazanin Alipourfard, Peter G. Fennell, and Kristina Lerman. Can you trust the trend: Discovering simpson's paradoxes in social data. *CoRR*, abs/1801.04385, 2018.

[2] Nazanin Alipourfard, Peter G. Fennell, and Kristina Lerman. Using simpson's paradox to discover interesting patterns in behavioral data. *CoRR*, abs/1805.03094, 2018.

[3] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104(3):671–732, 2016.

[4] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018.

[5] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175):398–404, 1975.

[6] Colin R. Blyth. On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366, 1972.

[7] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Jan 2019.

[8] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[9] Brian d' Alessandro, Cathy O'Neil, and Tom LaGatta. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data*, 5(2):120–134, Jun 2017.

[10] Nina Grgic-Hlaca, M. Zafar, K. Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. 2016.

[11] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016.

[12] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076. Curran Associates, Inc., 2017.

[13] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635, 2019.

[14] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 560–568, New York, NY, USA, 2008. Association for Computing Machinery.

[15] Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *CoRR*, abs/1811.05577, 2018.

[16] Harini Suresh and John V. Guttag. A framework for understanding unintended consequences of machine learning. *CoRR*, abs/1901.10002, 2019.

[17] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7, New York, NY, USA, 2018. Association for Computing Machinery.

[18] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. *30th International Conference on Machine Learning, ICML 2013*, pages 1362–1370, 01 2013.

[19] Indre Zliobaite. On the relation between accuracy and fairness in binary classification. *CoRR*, abs/1505.05723, 2015.

[20] Indre Zliobaite. A survey on measuring indirect discrimination in machine learning. *CoRR*, abs/1511.00148, 2015.

[2] github.com/PAIR-code/what-if-tool