Bias in AI Essay

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

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

II. Justifications

One reason to address bias in AI-based solutions is to ensure that the decisions do not reflect discriminatory behaviour toward certain groups or populations, as stated by Mehrabi et al. 2019 [12]. They justify this by highlighting the canonical example of the software COMPAS, 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 [8]. To justify this, they point to the broader debate concerning disparate impact, which is discussed extensively by [3]. They highlight "Redlining" (refusing opportunities to people base solely on their zip code) for loans as a classic example of disparate impact.

[13] echoes the sentiment of the authors in the previous paragraph, and tackles the problem of discrimination in data mining in a rule-based setting, by introducing the notion of discriminatory classification rules, as a criterion to identify the potential risks of discrimination.

III. Dataset fairness

One way to measure fairness in a dataset is to address Simpson's Paradox [5]. This can be done by comparing the regression for the entire population, regressions for each subgroup, and the unbiased regression. An example arose in [4], where it seemed like there was bias toward women in graduate school admissions, but at the same women also had an advantage over men, in some cases, when the data was separated and analysed over the departments.

A traditional statistical way to measure fairness in a dataset is selection bias. 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.

Other measures of fairness in a dataset include historical bias and representation bias, as introduced by [15]. Historical bias can be measured by evaluating the representational harm (such as reinforcing a stereotype) to a particular identity group. Representation bias can be measured by calculating the percentage a minority group makes up of the true distribution.

IV. Algorithm fairness

Mehrabi et al. 2019 [12] compiles some of the most widely used definitions of fairness which we can apply to measure an algorithm.

One measure is Fairness Through Unawareness. That is, "An algorithm is fair as long as any protected attributes A are not explicitly used in the decision-making process" [9], [11]. Protected attributes are specified in the Fair Housing and Equal Credit Opportunity Acts (FHA and ECOA) [6], and include race, colour, national origin, religion, sex, and more.

Another measure is Demographic Parity. This states that the likelihood of a positive outcome [16] should be the same regardless of whether the person is in the protected (e.g., female) group. 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., "$P(\hat{Y}|A = 0, Y = y) = P(\hat{Y}|A = 1, Y = y), y \in 0, 1$"). The equal opportunity definition states that the protected and unprotected groups should have equal true positive rates (i.e., $P(\hat{Y}|A = 0, Y = 1) = P(\hat{Y}|A = 1, Y = 1)$) [10].

[8] summarises the metrics for measuring disparate impact from [19]. The recommended measure is the "Mean Difference" divided by a normalisation constant. The mean difference can be modified to a "Conditional Mean Difference",

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

to account for distributional differences between the protected populations and the overall population.

[13] introduces (strong) α-protection to measure the discriminatory power of a rule, 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 [7], is as follows:

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

This contains the potentially discriminatory attribute personal_status.

V. Mitigations

Mehrabi et al. 2019 [12] compiles some ways to mitigate algorithmic bias.

[1], [2] proposed methods to discover Simpson's paradoxes in data automatically

Authors try to satisfy equality of opportunity and equalized odds in [10]

Another way to mitigate algorithmic bias mentioned in [8] is to introduce augmented cost functions during the model training phase. [17] and [18] are said to both augment a standard log-likelihood loss function with a 'fairness' regularizer, which takes into account differences in how the learning algorithm classifies protected vs. non-protected classes.

VI. Fairness in future

As mentioned by Mehrabi et al. 2019 [12], researchers have begun introducing tools that can assess the amount of fairness in a tool or system. The example given is Aequitas [14], which lets users test models for different population subgroups. I expect to see more of this 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] 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.

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

[6] 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.

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

[8] 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.

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

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

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

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