# Bias in Artifical Intelligence

fsdn71
*Department of Computer Science*
*University of Durham*
Durham, United Kingdom
fsdn71@durham.ac.uk

## I. Introduction

The definitions of bias are varied, but generally consistent: an inclination, leaning, tendency, bent; a preponderating disposition or propensity [1]. Bias is associated with, but subtly different from *discrimination*, the action of perceiving, noting, or making a distinction between things [2]. The difference is notable; where discrimination simply describes a difference in the perception of two objects, bias indicates a real positive or negative change in attitude or action made possible by the object's perceived difference. Bias, discrimination and their effects on individuals, populations and systems are real and measureable; this essay surveys the known methods of quantifying and reducing them, and, having established the current state of the art, discusses the moral imperatives at play.

### A. Algorithmic Bias

There is no difficulty in finding examples of bias in human thought [3] or actions [4]. With algorithms being nothing more than formalised extensions of human thought processes, bias - a natural component of human consciousness - extends to algorithms too. One example is the bias found in latent representations of words generated by NLP models [5] - associating technological fields such as science, physics and chemistry more closely with men than women, and the opposite for arts and humanities. Section II explores the existing biases in real-world algorithms and how these can be contextual and measured.

### B. Self-perpetuating Bias

Bias is difficult to measure and understand for a reason other than its subjective definition - its position as a phenomenon that both affects and is affected by its environment. For example, recommender systems are particularly prone to bias [6], due to the fact that they are initially trained on a data distribution with a particular bias, and subsequently retrained on new data collected by users informed by the model itself (users are recommended content, and often asked for feedback - but can only provide feedback on content which the model has recommended). Where a data-informed decision system is used in the real world, it has some real effect on the data distribution it learned from, and therefore has potential to perpetuate existing biases. These occurrences, and methods for mitigating bias are discussed in Section III.

### C. Moral Imperatives

Research into algorithmic bias and prevention methods is wide-ranging and ongoing. However, understanding the extent and effect of bias on society is only one part of the challenge; the available actions to mitigate it another; but the choice in which actions are morally justified is a highly subjective and debatable one. Equality is generally an important and respected quality in systems, but whether of outcome or treatment (the differences of which are discussed in [7]) is contested. The legal system has ramifications on which preventative actions are directly influenced or not by the state (for example, in the case of Lee v. Ashers, in which it was ruled that a baker could not legally refuse service by discimination against a customer's sexual orientation, but it could over the nature of a requested product [8]). However, as evidenced by the scale of controversy produced by such cases, the legal code is often detached from the personal beliefs of a substantial part of the population.

It could be argued that in a democracy, individuals writing algorithms with the potential for bias should follow the legislated guidelines set by elected representatives of the people, rather than whichever opinions are publicised and marketed most widely. On the other hand, the legislative process is slow and difficult to interpret, and even if it is fully representative of the moral preferences of the electorate, the electorate is not the entire population, and further, those moral preferences change over time [9], which suggests that if there is a 'correct' moral framework, humans do not consistently agree with it.

## II. Measuring Bias

Bias exists in a number of different forms, which may have implications on the appropriate way to measure and mitigate it. Both qualitative and quantitative explanations are useful for understanding and acting on bias.

### A. Types of Bias

Without decomposing bias into different types, no singular definition can accurately capture its current usage. In [10], the authors identify a number of different (but not mutually exclusive) types of bias, for example:

*a) Representation bias:* arises where a dataset contains observations from a sample of individuals substantially different than the population it claims to represent.

*b) Sampling bias:* occurs when a method of including individuals in a selection is non-random. For example, household surveys have been found to consistently under-sample top earners, due to lower (and therefore not entirely random) response rates [11].

*c) Aggregation bias:* is caused by identifying patterns in an aggregated dataset of multiple subgroups which did not exist in the disaggregated datasets. For example, consider the effect of a country-wide rise in top marginal income tax rates, with revenue distributed equally. The total change in revenue for the country is nil, but a county-level view would show high increases in revenue for less wealthy counties and decreases for more wealthy ones.

### B. Legislation

Relevant to the measurement of discrimination are the legal definitions set out in the Equality Act 2010: indirect discrimination is defined as the application of a policy to members of different groups, where one group is disadvantaged in comparison to another group. However, the strength of this clause is significantly reduced by a caveat - that this is not discrimination if the person applying the policy can show that it is a proportionate means of achieving a legitimate aim [12].

### C. Fairness metrics

Qualitative explanations of bias are useful, but difficult to reconcile with algorithms following entirely quantitative rules. A number of objective measures of fairness have been developed to provide directly comparable 'fairness' estimates.

*a) Group Unawareness:* The authors of [13] emphasised the importance of process unawareness of group memberships. Their model of fairness focuses analysis on whether each input feature is fair or unfair to make available to a machine learning model.

*b) Disparate Impact:* In [14], the authors propose a measure of the disparity between outcomes for different groups:

$$\frac{Pr(\hat{Y} = 1|X = 0)}{Pr(\hat{Y} = 1|X = 1)} <= D$$

where $\hat{Y}$ is a classified binary prediction and $X$ the membership of a majority protected class. They use a threshold of $D = 0.8$, informed by the US Equal Employment Opportunity Commission, to determine unfairness.

*c) Equalised Odds:* Proposed in [15] is a modification of the disparate impact measure which takes into account the true label of the predicted feature:

$$Pr(\hat{Y} = 1|X = 0, Y = y) = Pr(\hat{Y} = 1|X = 1, Y = y)$$

Ensuring that the error rates (for $y \in \{0, 1\}$) are consistent across groups allows models to account for existing disparities and ensure equality of treatment accordingly.

*d) Equalised Opportunity:* Similarly, equalised opportunity is calculated only on the cases with $Y = 1$:

$$Pr(\hat{Y} = 1|X = 0, Y = 1) = Pr(\hat{Y} = 1|X = 1, Y = 1)$$

All of the above metrics are useful for understanding biases of datasets; however, the nature of the biases they identify varies. Where disparate impact is used to unbias a model, this may result in a model does not provide equality of opportunity by actively rectifying historical bias. Equalised odds as a target metric instead preserves existing biases, seeking only to minimise the additional bias introduced by the model.

## III. Mitigating Bias

With a set of metrics for understanding the extent and directions of bias in datasets, we can by extension understand the effect of a dataset's bias on any model that is trained on it. Measures exist to mitigate the effects of bias before, during and after the model is informed by the dataset.

### A. Pre-processing

In [16], the authors describe the formal aims of pre-processing a dataset in order to reduce bias - to determine a randomised mapping from the original dataset (labelled with a protected class membership) to a new, debiased dataset:

$$p(\hat{X}, \hat{Y}|X, Y, D)$$

where $X$ is the feature matrix, $Y$ the labels and $D$ the protected class membership. Once the mapping function has been determined, a new dataset can be sampled from the original. While this approach makes use of re-sampling in order to change the characteristics of the overall dataset, other such as in [17] instead re-weight records. Both of these approaches have been shown to be effective when the labelled feature is a classification. However, this may not work for datasets where the label is a continuous feature, a circumstance that appears to be relatively neglected across the field of AI bias research compared to classification applications, since the distortion of the original dataset increases with the level of existing bias present.

*a) Generative adversarial networks:* The use of a GAN, entitled 'FairGAN', by [18] was found to reduce bias in datasets. This involved multiple adversarial models - firstly, a generator transforming random noise, given the sensitive attribute $s$, into a fake dataset sample $\hat{x}, \hat{y}$. Two discriminator models are trained - the first, to distinguish between real and synthetic dataset samples (ensuring the generated samples are accurate to the original dataset), and the second, to predict the sensitive attribute (incentivising the generator to produce samples without visible bias).

### B. In-processing

In-processing methods alter the model training process (or any point in which the model is informed by the data) to remove bias from the ouputs.

*a) Fairness regularizers:* In [19], a regularizer is proposed for usage during model training in order to constrain the degree of bias permitted. Referred to as the 'indirect prejudice index', this is the mutual information between the sensitive variable $S$, and $Y$, the target, in a dataset $D$:

$$PI = \sum_{(y,s)\in D} \hat{Pr}[y,s] \ln \frac{\hat{Pr}[y,s]}{\hat{Pr}[y]\hat{Pr}[s]}$$

If, for example, a pair $y$ and $s$ are independent, then their component will be nil; if not, it will be positive or negative depending on whether the $s$-classed individuals are over-represented or under-represented in their $y$-classed proportion.

*b) Adversarial learning:* In [20] a form of adversarial learning was used to prevent dataset bias from affecting model processing. From a sample $(x, y, s) \in D$, an encoder model $f$ transforms $x$ (optionally with $s$) into a representation $z$, such that a classifier model $g$ can predict $y$, (and optionally a decoder model $g$ can reconstruct the input) but an adversary $h$ cannot predict $s$. The latent representation can be thought of as a debiased version of the dataset that still retains predictive power.

### C. Post-processing

Post-processing methods do not alter the input dataset, or the model training process, but impose an additional processing step on models aiming to achieve fairness objectives.

*a) Optimal thresholding:* One such example is shown by [15], in which a threshold is applied to the result of the classifier in order to guarantee that equality of opportunity, as defined previously, is maximised (it may not be achievable for all groups, but can be optimized by selecting the point on the group-conditional ROC curves that minimises the difference between true-positive rates).

*b) Deferral of decisions:* However, there are other methods - including the ability for a model to defer prediction when confidence thresholds are not met [21]. This allows for both a model's decisions (tending to be fair but less accurate) and a decision maker's (more likely to be unfair but more accurate) to be improved by a partitioning decision that balances accuracy and bias.

## IV. Broader Context

Despite the increased research attention to the field of bias in artificial intelligence, the general public (ultimately a key arbiter of moral decisions) still believe that artificial intelligence should be more regulated than it is now - including on risks such as bias. However, the opinions of the general public are inconsistent and changeable, as demonstrated by the historical local popularity of policies of extreme bias and discrimination. Moral frameworks to debias models and data by researchers are potentially more reliable and desirable.

*a) Applications:* The methods outlined previously are shown to be effective at combating the biases that they are designed to reduce. Measuring the fairness of a dataset is unquestionable a useful process that can only increase the value of the models applied to it. Applying the relevant methods for debiasing data and models can in some applications be controversial, especially at the intersection of ethics, popularity and law.

*b) The future:* The future holds far greater space for artificially intelligent processes, especially machine learning. The problem that exists now of models affecting their own data distributions is so far only acute in specific domains, such as recommender systems, and less noticeable where user-bases are fast-changing: for example, the effects of bias causing a bank loan to be falsely denied are likely to be experienced by those people and processes interacting with the borrower, but are too small to be returned to the model in future data-gathering. However, as AI systems become more widely used, the effects of their self-reinforcement may become more problematic. This necessitates more agressive anti-bias mechanisms - a promising candidate for innovation is adversarial learning. As mentioned previously, models with continuous outputs are much less frequently used in anti-bias research, which may change as the demand for such models increases in future.

*c) Outcome and opportunity:* Featured in the introductions of many of the research papers cited in this essay are various opinionated assertions around equality of opportunity, equality of outcome and which of them is, in the mind of the author, irrelevant or nonsensical. A common argument is that conditioning models to produce equality of outcomes is a contradiction to the nature of equality and fairness, and therefore introduces more damaging unfairness into predictive models. This inconsistency across the research complicates the usefulness of various approaches, as they are only desirable if the person applying them to a model agrees with the underlying beliefs around ethical equality.

The outcome vs. opportunity debate focuses on a single question: should an algorithm seek to explicitly discriminate on a protected class in order to transform a biased 'default' distribution of data into a less biased distribution, or should the transformation itself be unbiased. However, I find that a single statement decides the entire issue: there is no default distribution. All data, models and humans (which are themselves data and models) are the outcome of an uncountable number of biased policies, environments and algorithms, with parameters set either by design or by ignorance. Algorithms which seek the pretense of 'unbiasedness' only seek to maintain existing bias and discrimination under the false notion that this is preferable to any change in the bias of a system. It seems remarkable that in seeking to make algorithms 'unbiased', researchers appear often to themselves reveal a fundamental bias: *status quo bias*. Equality of opportunity does not value all biases equally - instead showing a strong bias for *bias* which already exists. It is this fundamental contradiction which prevents me from seeing equality of oppotunity as a credible argument.

## REFERENCES


[1] Bias, *The Oxford English Dictionary*. Oxford University Press, 2021.

[2] Discrimination, *The Oxford English Dictionary*. Oxford University Press, 2021.

[3] C. FitzGerald, A. Martin, D. Berner, and S. Hurst, "Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: a systematic review," *BMC Psychology*, vol. 7, no. 1, p. 29, May 2019. [Online]. Available: https://doi.org/10.1186/s40359-019-0299-7

[4] M. Bertrand and E. Duflo, "Field experiments on discrimination," National Bureau of Economic Research, Working Paper 22014, February 2016. [Online]. Available: http://www.nber.org/papers/w22014

[5] M.-E. Brunet, C. Alkalay-Houlihan, A. Anderson, and R. Zemel, "Understanding the origins of bias in word embeddings," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 803–811. [Online]. Available: http://proceedings.mlr.press/v97/brunet19a.html

[6] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims, "Recommendations as treatments: Debiasing learning and evaluation," *CoRR*, vol. abs/1602.05352, 2016. [Online]. Available: http://arxiv.org/abs/1602.05352

[7] Z. Kodelja, "Equality of opportunity and equality of outcome," *Center for Educational Policy Studies Journal*, vol. 6, pp. 9–24, 02 2016.

[8] L. v Ashers Baking Company Ltd, October 1998. [Online]. Available: https://www.supremecourt.uk/cases/docs/uksc-2017-0020-judgment.pdf

[9] B. Garvey, "The evolution of morality and its rollback," *History and philosophy of the life sciences*, vol. 40, no. 2, pp. 26–26, Mar 2018, 29564652[pmid]. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29564652

[10] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *CoRR*, vol. abs/1908.09635, 2019. [Online]. Available: http://arxiv.org/abs/1908.09635

[11] S. Bach, A. Thiemann, and A. Zucco, "Looking for the missing rich: tracing the top tail of the wealth distribution," *International Tax and Public Finance*, vol. 26, no. 6, pp. 1234–1258, Dec 2019. [Online]. Available: https://doi.org/10.1007/s10797-019-09578-1

[12] Equality Act 2010, s. 19. [Online]. Available: https://www.legislation.gov.uk/ukpga/2010/15/section/19

[13] N. Grgic-Hlaca, M. Zafar, K. Gummadi, and A. Weller, "The case for process fairness in learning: Feature selection for fair decision making," 2016.

[14] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 259–268. [Online]. Available: https://doi.org/10.1145/2783258.2783311

[15] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *CoRR*, vol. abs/1610.02413, 2016. [Online]. Available: http://arxiv.org/abs/1610.02413

[16] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf

[17] F. Kamiran and T. Calders, "Data pre-processing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, 10 2011.

[18] D. Xu, S. Yuan, L. Zhang, and X. Wu, "Fairgan: Fairness-aware generative adversarial networks," *CoRR*, vol. abs/1805.11202, 2018. [Online]. Available: http://arxiv.org/abs/1805.11202

[19] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," 09 2012, pp. 35–50.

[20] D. Madras, E. Creager, T. Pitassi, and R. S. Zemel, "Learning adversarially fair and transferable representations," *CoRR*, vol. abs/1802.06309, 2018. [Online]. Available: http://arxiv.org/abs/1802.06309

[21] D. Madras, T. Pitassi, and R. Zemel, "Predict responsibly: Improving fairness and accuracy by learning to defer," in *NeurIPS*, 2018.