Journal of Data Science 24 (1), 203–217
January 2026
DOI: 10.6339/26-JDS1219
Statistical Data Science

A Designed Look at Artificial Intelligence from the Lens of Fairness

Md Borhan Uddin[1,*], Mengqi Yin[1], and Nairanjana Dasgupta[1]
[1]*Department of Mathematics and Statistics, Washington State University, United States*

Abstract

As the use of Artificial Intelligence (AI), especially Generative AI, becomes ubiquitous, we take a look at the performance of these methods. We specifically focus on concept of fairness element of trustworthiness. We use Statistical Parity Difference and Equalized Odds Difference to mathematically measure fairness. To systematically study how various factors like bias, access to protected categories, types of intervention affect fairness and accuracy, we performed a simulation as a multi-factor experiment. Our results indicate that accuracy and fairness (in terms of statistical parity and equalized odds) tend to go in opposite directions. This opens up the question of whether we can look at methods that can consider both accuracy and fairness simultaneously.

Keywords *equalized odds; ethical principles of science; factorial design; statistical parity; unbiasedness*

1 Introduction

For the last few years with the advent of GPT-4, Artificial Intelligence (AI), especially Generative Artificial Intelligence (GenAI), is everywhere. From large to small, from government to industry to academia, this is something everyone is talking about and using. There are multiple GenAI platforms available easily and *for free*. Hence, there is no cost, no necessary training, or age cut-off for people using this technology. Children use it for homework, banks use it for loan decisions, courts use it to consider recidivism, universities use it for admissions, college students use it for writing assignments and large industries use it for crucial marketing decisions. In short, it has permeated to every section of society with not as much thought about how trustworthy some of the methods are.

In the past, programmers or data scientists who used data to predict decisions were taught the how, what, why and *when* as part of their training, but with the deployment of GenAI, this is not the case. This democratization of GenAI is not necessarily a good thing since people who are *using* GenAI do not often understand how the methods work and more importantly what its limits are and its appropriateness and *trustworthiness*. This lack of scrutiny may contribute to the production and spread of massive and dangerous misinformation (Capraro et al., 2024). Thus, we have a very powerful tool in the hands of the general population, and so, the immediate question is, how often is this used and applied in situations where it is *not appropriate* mathematically, sociologically or ethically. Further, there is little understanding of how often are these tools being used, when it is mathematically feasible for use, but there are ethical questions.

*Corresponding author. Email: mdborhan.uddin@wsu.edu.

Received November 30, 2025; Accepted January 24, 2026

Ethical challenges introduced by GenAI can be defined by five primary domains, such that, *bias and discrimination*, *misinformation and deep fakes*, *data privacy violation*, *intellectual property issues*, and *accountability and explainability* (Surbakti, 2025).

In this paper, we will address trustworthiness from the angle of *fairness* that falls under the domain of *bias and discrimination*. We would like to reiterate (Chouldechova, 2017) that "fairness" is not a statistical concept rather a social and ethical concept. Further, methods could be fair but for a particular criteria may exhibit "disparate impact" for another criteria. It is also important to note that both Kleinberg et al. (2017) and Chouldechova (2017) have pointed out the *impossibility result* that when the underlying groups have different prevalence, then the different fairness measures used like *calibration*, *error-rate balance*, *being better than a random guess*, will not be compatible with each other. To avoid these issues we focus on the criteria of statistical parity and equalized odds which fall under the umbrella of *error-rate balance*. Unlike (Chouldechova, 2017; Kleinberg et al., 2017) and others, we do not look at the *compatibility of different measures of fairness*.

If we really get down to it, GenAI is data based decision making. Thus, it is important to look at the history and trajectory of data based decision making and why ideas of fairness, informed consent and subject welfare are crucial. Looking at history is critical so that we do not make the mistakes that were made in the past.

The idea of using data and empirical evidence or data for decision making has been going throughout history. There is evidence of data collection in terms of census, crop yield and trade from prehistoric era. Chinese, Egyptians, Babylonians used empirical evidence to impose taxes and allocate resources. This topic was discussed by Graunt (1662) in the 17th century, discussing mortality in public health decisions. As probability theory developed – the idea of statistical inference started taking shape. Though there are multiple evidences of using data to make decisions (Florence Nightingale and the Crimean War etc.) (Nightingale, 1858), the idea of collecting data for making specific decisions emerged in the 1920. However, then the concepts around ethics and fairness were not front and center. As a result, there are various examples of unethical data collection throughout history where no effort was made to instill trust. Stark examples include Tuskegee Syphilis Study (1932–1972) (Reverby, 2009) and Nazi experiments (Lifton, 1986) where testing and data collection were done without permission and violating both trust and ethical principles.

Figure 1 traces the progression of data analysis from simple, early practices to more advanced analytical approaches, showing how modern data-driven methods developed from earlier foundations. The first major breakthrough in terms of ethics in data collection was the Nuremberg Code (1947) (Nuremberg Military Tribunals, 1949). This was followed by The Declaration of Helsinki (1964) (World Medical Association, 2013) and the Belmont Report (1979) (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). These were the guidelines to research principles of ethical collection and use of data. The Belmont report emphasized the concepts of respect for persons, beneficence and justice and implicitly trustworthiness. And these remain the cornerstone for data collection, statistical design, data analysis, and reporting. Ideas of informed consent, privacy, confidentiality and transparency in reporting are now the accepted norms of research practice. But these days the questions are creeping up again. As we face the age of large repositories of generated data we are again faced with same ethical dilemma. The data used to train some of the GenAI predictions are often not collected with explicit consent. The current trend of "publish or perish" in academia has led to issues with reproducibility and transparency in research (Ioannidis, 2005). Ethical challenges arising from GenAI have been studied in different capacities in different fields (Doshi-Velez et al.,

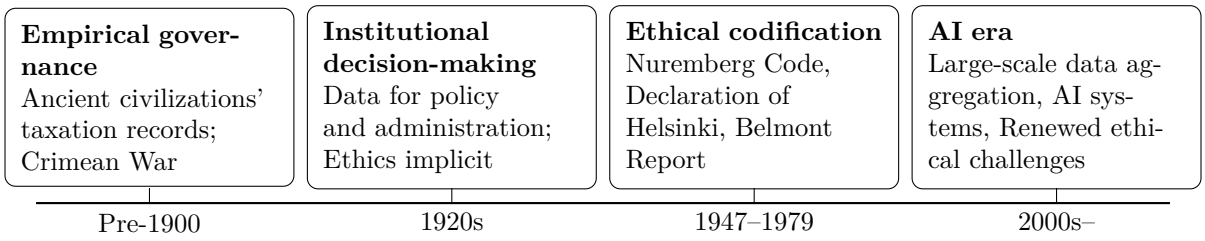


Figure 1: A historical timeline of data-driven decision making and the evolving role of ethics.

2017; Gupta et al., 2024; Haltaufderheide and Ranisch, 2024; Tabassum et al., 2025). In this paper, we look at the fairness angle of trust and discuss two related metrics. In Section 2, we lay the foundation by discussing the concepts of fairness and unbiasedness and discuss two statistical measures of fairness. In Section 3, we describe our simulation experiment and the data generation process. Our results are provided in Section 4. We discuss our findings and talk about future research in Section 5.

2 Fairness Versus Unbiasedness: Definition and Context

We would like to first distinguish the concept of "fairness," which is getting some attention these days and *unbiasedness*. One of the main tenets of statistical design is the idea of *unbiasedness*. The ideas of randomization, replication and local control all contribute to the idea that the study we design is unbiased. Unbiased is defined to be *free from prejudice and favoritism* by the Merriam-Webster Dictionary. As fairness takes center-stage, the immediate question is whether fair and unbiasedness are the same. Merriam-Webster Dictionary defines *fair* to be *marked by impartiality and honesty: free from self-interest, prejudice, or favoritism*. While often used interchangeably, these are subtly different. Unbiasedness is more of absence of favoritism or skew but fairness is about *equitable treatment of all groups*. Hence, in the idea of *fairness* we assume that there are multiple groups and while this is not necessary in unbiasedness. In this paper, we focus on the idea that fairness that can be measured by metrics *existing in the literature*: statistical parity (defined by (Darlington, 1971) and popularized in AI literature by (Chouldechova, 2017)) and equalized odds (which was introduced by (Dwork et al., 2012; Hardt et al., 2016)).

Unbiasedness, or being correct on average, is always the aim, we will look at some scenarios where biases exist and discuss some common sources of bias. Some of these biases could also give rise to unfairness. We will denote D as the data set used to train our models and $\mathcal{P}$ as the population on which we would like to make our decisions. While $D \subset \mathcal{P}$, it is not necessarily a random sample. The response or target variable Y is generally a class, X is the collection of explanatory variables or features that is used to predict the class variable, Y, and A is a group variable that can be considered "protected" in some fashion. Further, we are *not* assuming that $A \subset X$, i.e. that the protected group may or may not be a part of the explanatory variables studied.

2.1 Sources of Bias

The primary sources of bias that are relevant to our setting are presented here briefly. These sources of bias are not necessarily mutually exclusive, meaning they may appear simultaneously in practice.

- **Data Bias.** This type of bias happens when the data used to train the models are not representative of the population of interest. Symbolically, given a dataset $D = \{(X_i, Y_i)\}_{i=1}^{n}$,

bias may arise if the marginal distribution of X, $P(X)$ in the dataset is not representative of the true population. This bias *often present* when we try to use scraped or generated data where randomization is not considered. Shifts in the data from training to testing is one of the common causes of data bias. These shifts have been defined and studied in previous literature (Quionero-Candela et al., 2009; Federici et al., 2021; Koh et al., 2021; Shao et al., 2024).

- **Label Bias.** Label biases may happen when the target variable has some inherent biases in the training data. Historical bias of certain classes of Y given X in the training, is one very good example of label bias. Hence, the conditional distribution of Y given X, $P(Y \mid X)$ is skewed. Label bias in hiring is commonly discussed. If historically a certain group was favored for a certain job, training data remain biased and impact the model's performance trained on them (Barocas and Selbst, 2016).
- **Algorithmic Design Bias.** This bias arises from the algorithm design used during the training. The optimization function:

$$\min_{\theta} \sum_{i} L(f_\theta(X_i), Y_i) \tag{2.1}$$

may prioritize accuracy over other criteria, leading to bias. Therefore, the learning procedure, including the objective function, constraints, and evaluation criteria are very important. The most common approach to fairness is axiomatic. We may choose a fairness criterion and ensure the model satisfies it by constrained learning, or we may post-process the learning (Corbett-Davies et al., 2023; Weerts et al., 2024). If we make the model "fair" according to some selected criteria, we may overlook some alternative policies that may be better for the overall well-being of protected groups. Therefore, we may consider fairness constraints as policy choices and design the algorithm accordingly (Corbett-Davies et al., 2023).

2.2 Statistical Measures of Fairness

Fairness is not a statistical concept and how we measure fairness depends upon its inherent definition. The study of fairness in the ethics and philosophy literature is complicated as it is defined under multiple (sometimes competing and complimentary) lenses. For the purpose of this paper we are defining fairness as *equitable treatment across groups*. Out of the various metrics available we focus on two measures, equalized odds and statistical parity. We need to emphasize that these fairness evaluation metrics are used post hoc, their interaction with the generalization guarantees from classical learning theory is not well understood yet. Further, complicating the matter several recent papers (Dwork et al., 2012; Barocas et al., 2023; Kim et al., 2019; D'Amour et al., 2022) reveal that models that are fair and accurate on training dataset may exhibit severe fairness violations on unseen data.

In our context, along with the set of explanatory variables, X, there is A which comprises of groups and we are interested in the fairness across the groups. To keep our notations simpler, we will set up the fairness metrics assuming that we are predicting two classes, $Y \in \{0, 1\}$ and the attribute A has two groups a and b. This is done for simplifying notation and can be generalized to multiclass predictions and multiclass protected attribute. We discuss the two measures of fairness considered in our paper here.

- **Statistical Parity.** The concept of statistical parity is an old one. It has existed since the 1970s (Darlington, 1971). However, it was popularized in AI literature later (Chouldechova, 2017). A classifier f predicts Y, given X, satisfies statistical parity if:

$$\Pr(f(X) = 1 \mid A = a) = \Pr(f(X) = 1 \mid A = b), \tag{2.2}$$

with $\hat{Y} = f(X)$ as the predicted outcome. This is commonly used for evaluating fairness. A classifier is considered fair if the prediction probability remains identical for all groups of A. This can be thought of as $\hat{Y} \perp A$, i.e., the prediction is independent of the protected attribute. An alternative term used for statistical parity is demographic parity.

- **Equalized Odds.** Equalized odds is a newer idea (Dwork et al., 2012; Hardt et al., 2016). In Equalized Odds, we take a step further than the statistical parity and interested in seeing if $\hat{Y} \perp A$ given the true outcome. We try to look at if the true positive rate (TPR) and false positive rate (FPR) are equal across all groups of A. This can be written as:

$$\Pr(f(X) = 1 \mid Y = y, A = a) = \Pr(f(X) = 1 \mid Y = y, A = b), \tag{2.3}$$

where the feature vector, X includes the protected attribute $A \in \{a, b\}$, and here, $y \in \{0, 1\}$.

3 Experimental Setup

To look systematically at fairness, we performed a multi-factor designed simulation study. The simulation experimental design systematically evaluates how different sources of bias, bias levels, access levels, and fairness interventions interact to affect accuracy, statistical parity and equalized odds.

The following are the input for the experiment.

- Sources of Bias (3 levels): It denotes the types of data-generating bias, including linear bias, nonlinear bias, and label noise.
- Bias level (3 levels): We have low, medium, and high biases, representing increasing magnitudes of bias levels. This is nested within the Source of Bias.
- Access (2 levels): It represents full or no access to the protected attribute during model training.
- Intervention (2 levels): We utilized no intervention, and postprocessing based on equalized odds.

The combination of these factors results in a $3(3) \times 2 \times 2 = 36$ design, yielding 36 distinct experimental configurations. Each configuration represents a specific situation under which models are trained, interventions are applied, and metrics are computed. This structure allows a quantitative decomposition of performance and fairness outcomes into main effects and interactions across bias sources, bias levels, access levels, and intervention strategies.

3.1 Evaluation Metrics

To evaluate the effect of the factors, we considered the metrics, overall accuracy, absolute statistical parity difference (SPD), and absolute equalized odds difference (EOD). Metrics are defined so that smaller values of SPD and EOD indicate higher fairness, while higher accuracy measures better predictive performance.

- **Accuracy.** Accuracy is exactly what the name indicates. How often are the predictions correct. In other words out of the total number of predictions made, what percent of them are correct.
  For predictions $\hat{Y}_i$ and true labels Y_i,

$$\text{Accuracy} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\hat{Y}_i = Y_i\}.$$

- **Statistical Parity Difference (SPD).** We discussed Statistical Parity in 2.2. The absolute difference between the positive prediction rates, or the TPR, between the two groups of interest is defined as the statistical parity difference:

$$\text{SPD} = \left| \mathbb{E}[\hat{Y} \mid A = a] - \mathbb{E}[\hat{Y} \mid A = b] \right|.$$

  A value of zero corresponds to perfect parity, while larger values indicate stronger group-level imbalance in predicted outcomes.
- **Equalized Odds Difference (EOD).** The absolute difference between the equalized odds (discussed in the subsection 2.2) is defined as the Equalized Odd Difference. It measures group-level disparities in TPR and FPR and, thus, accounts for SPD, along with FPR:

$$\text{EOD} = \max\Big(\left|\text{TPR}_1 - \text{TPR}_0\right|,\ \left|\text{FPR}_1 - \text{FPR}_0\right|\Big),$$

  where

$$\text{TPR}_g = \frac{\sum_i \mathbf{1}\{Y_i = 1, \hat{Y}_i = 1, A_i = g\}}{\sum_i \mathbf{1}\{Y_i = 1, A_i = g\}}, \qquad \text{FPR}_g = \frac{\sum_i \mathbf{1}\{Y_i = 0, \hat{Y}_i = 1, A_i = g\}}{\sum_i \mathbf{1}\{Y_i = 0, A_i = g\}}.$$

Both SPD and EOD are reported in absolute value to ensure comparability on a common scale $[0, \infty)$, where 0 denotes perfect fairness. While SPD and EOD are related in that they both aim to quantify group-level disparities in outcomes, they fundamentally capture different aspects of fairness. Therefore, satisfying one criterion does not imply the other. For example, a classifier may equalize acceptance rates across different groups, resulting in low SPD but still failing to equalize false positive or false negative rates, which may result in higher EOD.

3.2 Data Generation Processes

We consider a binary protected attribute $A \in \{0, 1\}$ to maintain conceptual clarity and computational simplicity throughout the experiments. Let the sample size be denoted by $n \in \mathbb{N}$ and the feature dimension by $d \in \mathbb{N}$. All experiments are conducted under this general framework unless otherwise specified.

Data are generated by the methods given below. Once the structure is set, we took ten random samples as replications. This allowed us to also study variability within a data generation structure. The details of the data generation relating to these three types of bias are given in the Supplementary Material. We would like to clarify that the bias we are adding is related to fairness as we are assuming that there is a protected attribute A in the study design.

3.2.1 Types of Bias

The types of bias we utilized in our setting are discussed here.

- **Linear Bias** We introduce a mean shift in the first feature. We also have a direct effect coefficient $\gamma \in \mathbb{R}$ quantifying the impact of the protected attribute A on the outcome Y. Formally, let $m = (1, 0, \ldots, 0)^\top \in \mathbb{R}^d$, and define the observed feature vector as so that $X_i \mid (A_i = a) \sim \mathcal{N}(\mu_a, I_d)$. A latent score is then specified as $Z_i^\star = m^\top X_i + \gamma A_i + \varepsilon_i$, from which the binary response is obtained as $Y_i = \mathbf{1}\{Z_i^\star > 0\}$. This process induces bias through γ and the decision boundary vary systematically across the groups defined by the protected attribute, resulting in a direct structural form of unfairness.

- **Nonlinear Bias** This is similar to the process of linear bias, except instead of inducing the bias γ linearly, we use non-linear functions like sine functions and log functions to induce the bias.
- **Noisy Outcome Process** This data-generating process introduces label noise depending on the classes in the observed responses. Although the process of data generation is same as the linear bias process, the responses are corrupted with unequal probabilities for different groups of A which results in asymmetric mislabeling. Using the notation in the linear bias, we define $q_a, q_b \in [0, 1]$ the probabilities of corrupting the response for individuals with $A = a$ and $A = b$, respectively. We first generated features and outcomes using linear bias mechanism. The observed outcomes are then altered with probability q_a conditional on the class of the protected attribute, $\Pr\big(Y_i^{\text{obs}} \neq Y_i^{\text{new}} \,\big|\, A_i = a\big) = q_a$.
  When $q_a \neq q_b$, one subgroup experiences a higher level of output corruption than the other, which leads to a systematic disparity in label reliability. This process imitates real-world situations where class-dependent mislabeling may occur, such as biased historical records or diagnostic data across different groups of A.

3.2.2 Bias Levels

Every bias family is evaluated with three levels of magnitude, such as, low, medium, and high which corresponds to a systematic increase of degrees of dependence between A and the data-generation processes. For each data generation process, only one bias-controlling parameter is varied systematically to make different processes comparable.

In the linear bias process, the direct effect coefficient γ controls the bias levels. The parameter takes values $\gamma \in \{0.2, 0.6, 1.0\}$ which represents low, medium, and high bias, respectively.

In the nonlinear bias process, the scaling parameter a_{scale} controls the levels of bias. The parameter varies over $\{0.5, 1.0, 1.5\}$.

In the noisy outcome process, the linear bias process is employed to generate clean data first. The bias level is then defined by a three parameters (γ, q_a, q_b), where γ is the parameter from linear bias process and (q_a, q_b) denote the label corruption probabilities. The configurations for low, medium, and high bias are $(0.2, 0.05, 0.10)$, $(0.6, 0.05, 0.20)$, and $(1.0, 0.05, 0.35)$, respectively.

3.2.3 Access Levels

This factor concerns with the levels of access of A to the model during training and testing. We have two different levels, such as, full access or no access to protected attribute, A. In different scenarios, A may or may not be available to the learning algorithm.

Under full access, the protected attribute is available during all stages, which allows the models to utilize protected attribute explicitly. When we have no-access, A is entirely withdrawn from the dataset, and models are blind to any direct information about A.

3.2.4 Intervention Levels

This experimental factor specifies the fairness interventions that are used to make the predictions fair. Let the training set be $(X_{\text{tr}}, A_{\text{tr}}, Y_{\text{tr}})$ and define corresponding validation and test sets. We consider two intervention types.

- **No intervention.** The model is trained directly on $(X_{\text{tr}}, Y_{\text{tr}})$ without using A:

$$\hat{f}_{\text{none}} = \arg\min_{f \in \mathcal{F}} \mathcal{L}(Y_{\text{tr}}, f(X_{\text{tr}})).$$

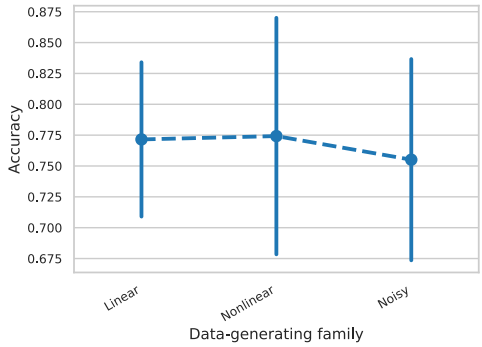
- **Postprocessing based on Equalized Odds.** Given fitted scores $\hat{p}_i = f(X_i)$, thresholds are found to satisfy approximate equality of false positive and false negative rates between groups:

$$\Pr(\hat{Y} = 1 \mid A = 0, Y = y) \approx \Pr(\hat{Y} = 1 \mid A = 1, Y = y), \quad y \in \{0, 1\}.$$

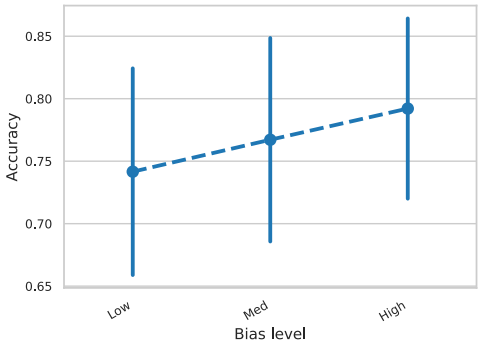
  All hyper-parameters were fixed across all experimental runs to ensure that differences that are observed are only due to the factors. Since model choice is not a primary variable of interest, these algorithms serve as representative learning functions for fairness evaluation.

4 Results

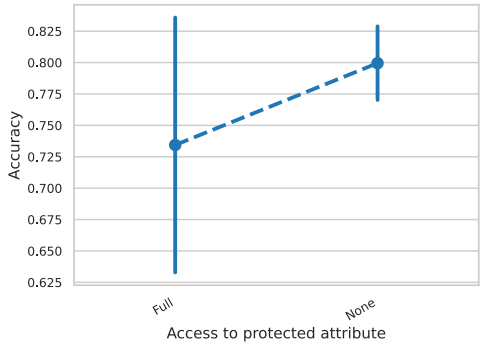
The following presents the aggregated results, highlighting the main effects of individual factors. We looked at the first-order interactions on model performance and fairness measures as well. But the results did not provide as much insight as the main effects. We will focus on the main effects here. Figures 2–4 give us the plots of the main effects. The interesting piece here is that the accuracy and the fairness metrics of SPD and EOD are diametrically opposite.



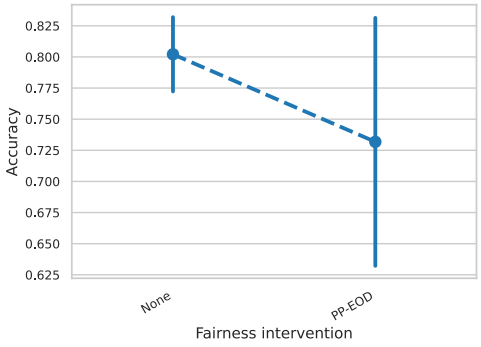
(a) Effect of Family on Accuracy



(b) Effect of Bias on Accuracy

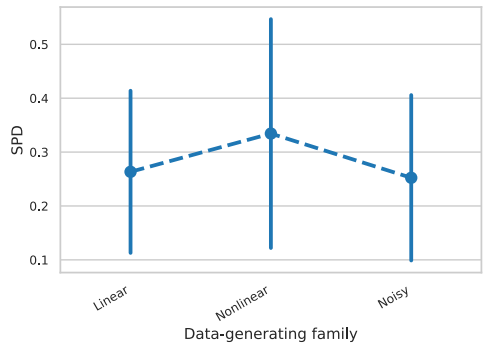


(c) Effect of Access on Accuracy

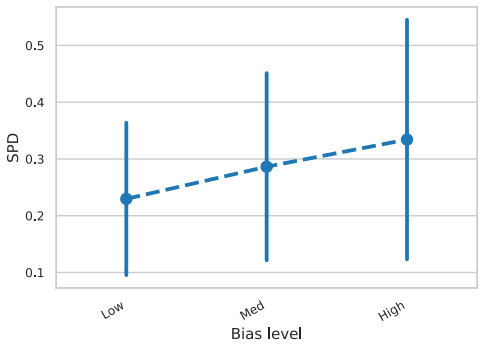


(d) Effect of Intervention on Accuracy

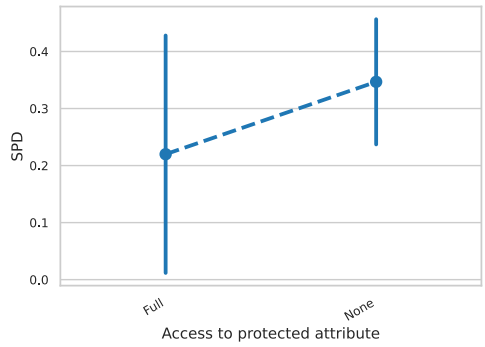
Figure 2: Main effects on accuracy. The points indicate the mean values, and the bars represent ±1 standard deviation. PP-EOD denotes the postprocessing equalized odds intervention.



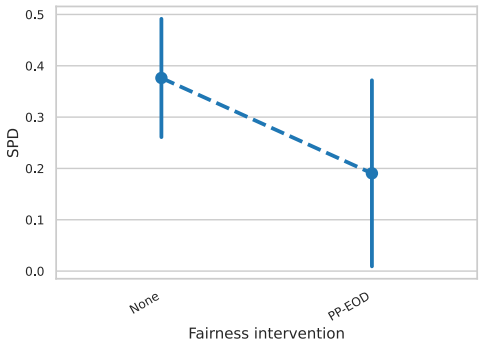
(a) Effect of Family on SPD



(b) Effect of Bias on SPD



(c) Effect of Access on SPD

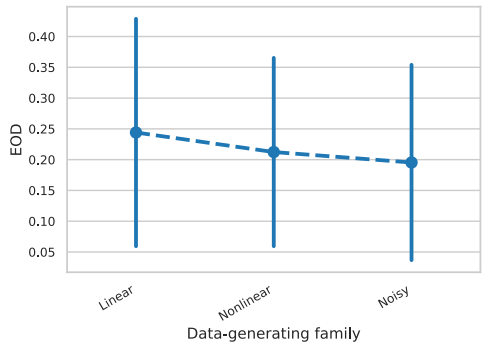


(d) Effect of Intervention on SPD

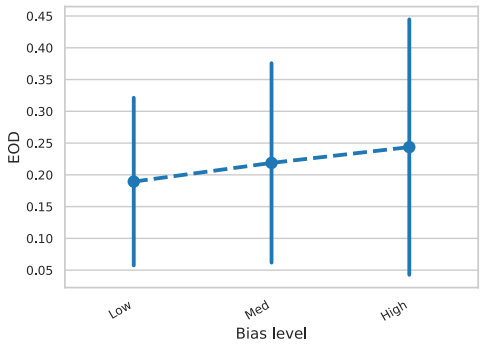
Figure 3: Main effects on SPD. The points indicate the mean values, and the bars represent ±1 standard deviation. PP-EOD denotes the postprocessing equalized odds intervention.

As the level of bias increases from low to high, accuracy rises along with SPD and EOD, indicating that the models become *more unfair*. It is not surprising, since higher bias allows the classes to be more separated, making the classification easier and boosting accuracy, even as fairness reduces. This emphasizes that why accuracy alone can be misleading as a model may perform well by exploiting biased pattens already present in the data. A similar trade-off appears when fairness interventions are applied. Moving from no intervention to the post-processing EOD leads to accuracy drops sharply, while SPD and EOD decrease as intended. When the model has no access to the protected attribute, accuracy goes up, but SPD and EOD also increases, suggesting that ignoring protected attributed does not eliminate bias and may even worsen it by allowing proxy variables to take over. Finally, while linear versus nonlinear does not drastically change accuracy, the nonlinear setting tends to produce more extreme SPD values. Overall, improving fairness typically comes at the cost of accuracy.

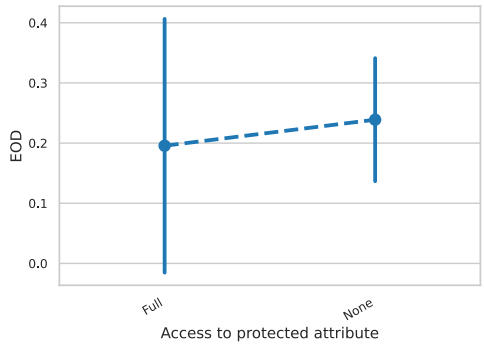
In Table 1, we provide the mean and standard deviation of accuracy across the groups. In general, we see that average accuracy is higher in the situations where no intervention was done compared to the intervention of post-processed EOD. This is in line with past literature that if measures are taken to preserve fairness, accuracy suffers. The interesting finding is that the standard deviation of the accuracy does not change much across the conditions and across interventions or not. We see that access to the protected attribute does alter accuracy. When no intervention is made and there is a protected attribute but the model does not have access



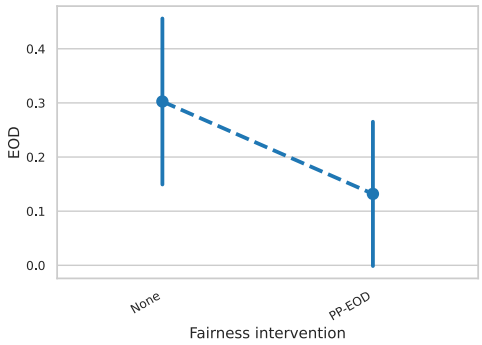
(a) Effect of Family on EOD



(b) Effect of Bias on EOD



(c) Effect of Access on EOD



(d) Effect of Intervention on EOD

Figure 4: Main effects on EOD. The points indicate the mean values, and the bars represent ±1 standard deviation. PP-EOD denotes the postprocessing equalized odds intervention.

to the attribute, it reduces accuracy slightly in general. However, the pattern reverses when post-processing EOD intervention is used. Here, the accuracy tends to increase from full access to no access. Our conjecture is that might be an overkill as post processing EOD is already taking the protected attribute into consideration.

In Tables 2 and 3, we look at the mean and standard deviation of the Statistical Parity Difference and Equalized Odds Difference. However, the interesting thing to note is that in the EOD (where 0 is perfectly fair, in the sense of equalized odds across groups), we can have numbers as high as 0.59 when bias is high, and the protected variable is in the data set, but no intervention is made. For SPD in that case, we have 0.46, which indicates an unfair model. However, accuracy was 0.83 showing a higher than average "high performing model".

This observation that accuracy and Fairness behave in opposite manner is particularly relevant in high-stake environments such as healthcare, criminal justice, financial domains etc., where violation of fair outcome under shifts may have severe social consequences and accuracy of the model does matter. Various sources of bias can affect social groups disproportionately. This raises some more questions than there are answers, and begs the validity of use of accuracy as the only metric of choice, as is done in most ML/AI models. There is some work on this topic in the computer science arena, (Dutta et al., 2020) and related papers.

Table 1: Accuracy mean and standard deviation by groups.

			No intervention		Postprocess EOD	
Family	Bias	Access	Mean	Std.	Mean	Std.
linear	high	full	0.829	0.012	0.723	0.033
		none	0.808	0.018	0.808	0.018
	med	full	0.808	0.012	0.705	0.037
		none	0.795	0.015	0.795	0.015
	low	full	0.780	0.015	0.647	0.110
		none	0.780	0.015	0.780	0.015
nonlinear	high	full	0.854	0.017	0.719	0.066
		none	0.853	0.017	0.853	0.017
	med	full	0.813	0.027	0.608	0.135
		none	0.813	0.026	0.813	0.026
	low	full	0.780	0.029	0.624	0.081
		none	0.780	0.030	0.780	0.030
noisy outcome	high	full	0.800	0.023	0.669	0.129
		none	0.795	0.017	0.795	0.017
	med	full	0.801	0.012	0.667	0.081
		none	0.794	0.014	0.794	0.014
	low	full	0.778	0.014	0.615	0.096
		none	0.778	0.014	0.778	0.014

Table 2: Statistical parity difference mean and standard deviation by groups.

			No intervention		Postprocess EOD	
Family	Bias	Access	Mean	Std.	Mean	Std.
linear	high	full	0.462	0.022	0.045	0.042
		none	0.285	0.285	0.028	0.028
	med	full	0.434	0.027	0.043	0.037
		none	0.302	0.302	0.036	0.036
	low	full	0.353	0.036	0.039	0.049
		none	0.327	0.327	0.030	0.030
nonlinear	high	full	0.623	0.133	0.036	0.117
		none	0.603	0.603	0.038	0.038
	med	full	0.451	0.057	0.047	0.089
		none	0.423	0.423	0.036	0.036
	low	full	0.231	0.032	0.042	0.047
		none	0.216	0.216	0.037	0.037
noisy outcome	high	full	0.350	0.000	0.088	0.000
		none	0.323	0.323	0.048	0.048
	med	full	0.398	0.000	0.053	0.000
		none	0.309	0.309	0.035	0.035
	low	full	0.349	0.000	0.043	0.001
		none	0.334	0.334	0.037	0.037

Table 3: Equalized odds difference mean and standard deviation by groups.

			No intervention		Postprocess EOD	
Family	Bias	Access	Mean	Std.	Mean	Std.
linear	high	full	0.592	0.107	0.030	0.057
		none	0.185	0.085	0.185	0.085
	med	full	0.483	0.114	0.042	0.054
		none	0.236	0.082	0.236	0.082
	low	full	0.334	0.095	0.028	0.036
		none	0.289	0.079	0.289	0.079
nonlinear	high	full	0.457	0.106	0.062	0.073
		none	0.390	0.093	0.390	0.093
	med	full	0.273	0.061	0.030	0.043
		none	0.239	0.052	0.239	0.052
	low	full	0.157	0.042	0.032	0.040
		none	0.140	0.044	0.140	0.044
noisy outcome	high	full	0.291	0.215	0.000	0.000
		none	0.171	0.077	0.171	0.077
	med	full	0.407	0.141	0.000	0.000
		none	0.221	0.083	0.221	0.083
	low	full	0.303	0.084	0.000	0.001
		none	0.280	0.073	0.280	0.073

5 Conclusion

To some extent, our results are not earth shattering and has been conjectured by others in the past, that focus on accuracy might be achieved at the cost of fairness. However, to the best of our knowledge, this is the first systematic study of the various factors that have a role in these models. Our simulation design looked at protected groups and how the interventions do help to establish fairness. The one glaring issue we still have to talk about is that, the models are trained and tested using accuracy as the only metric for optimization. The intervention of post-processing EOD is done post-hoc. How we bake in fairness in the algorithm itself by potentially looking at another optimization criteria is an open area of research.

The limitations of this study is that we are relying on existing software and methods for the simulation and on our data generation scheme. Further, *fairness* is a very fluid term and different definitions of fairness exist. It has been shown in the past that these metrics of fairness are often at odds with each other. For this study, we focused on the definition of *equitable treatment across groups* as our definition of fairness and measured it using two metrics that are known in the literature, statistical parity difference and equalized odds difference. Hence, other definitions of fairness may not provide the same results as ours. As a matter of fact, the impossibility results (Chouldechova, 2017; Kleinberg et al., 2017) clearly talks about the different measures of fairness could be at odds with each other and cannot be satisfied at the same time. Hence, our results are only applicable to the fairness metrics we used. However, from a statistical perspective equality across groups, made the most sense to us for measuring fairness.

While we present results from the 3 crossed and one nested factor in this paper, we would

like to point out that we did try other factors and other simulations, but the results were similar. Another limitation of this study is esoteric, if we do not have access to the protected group, how can we use it and look at fairness. We only looked at very specific types of bias and specific types of protected groups but the overall results are clear. Higher accuracy is often at the cost of fairness as *measured by our chosen metrics*. There is much work to be done in this area and we hope that statisticians and data scientists look at this as an open area of research. However, the fact that statisticians are looking at trustworthiness and taking fairness into account in algorithms is a welcome sign. We are currently working on developing hybrid methods that can optimize both fairness and accuracy simultaneously. But there is much work to be done in this arena and somehow hasn't been discussed in the statistics literature. There are open questions about how to *bake in* fairness into the optimization algorithms. We hope our work serves as a call for more research in this area.

Supplementary Material

The supplementary materials include Data generation process described in 3.2 as well as the full Python code. The Python implementation is also available at Https://github.com/borhan-stat/fairness-simulation-paper.

Acknowledgments

The authors thank the editor, associate editor, and referees for their constructive comments which has led to significant improvement of this paper.

Funding

This work was partly supported by a grant from Washington State Students Achievements Council (AWD00499).

References

Barocas S, Hardt M, Narayanan A (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Barocas S, Selbst AD (2016). Big data's disparate impact. *California Law Review*, 104(3): 671–732.

Capraro V, Lentsch A, Acemoglu D, Akgun S, Akhmedova A, Bilancini E, et al. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *Proceedings of the National Academy of Sciences of the United States of America*, 121(27): e2400303121.

Chouldechova A (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2): 153–163. https://doi.org/10.1089/big.2016.0047

Corbett-Davies S, Gaebler JD, Nilforoshan H, Shroff R, Goel S (2023). The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24: 1–117.

D'Amour A, Heller K, Moldovan D, Adlam B, Alipanahi B, Beutel A, et al. (2022). Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226): 1–61.

Darlington RB (1971). Another look at "cultural fairness". *Journal of Educational Measurement*, 8(2): 71–82. https://doi.org/10.1111/j.1745-3984.1971.tb00908.x

Doshi-Velez F, Kortz M, Budish R, Bavitz C, Gershman S, O'Brien D, et al. (2017). Accountability of ai under the law: The role of explanation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3064761

Dutta S, Wei D, Yueksel H, Chen PY, Liu S, Varshney K (2020). Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In: *Proceedings of the 37th International Conference on Machine Learning*, (H Daumé III, A Singh, eds.), volume 119 of ICML, 2803–2813. PMLR.

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012). Fairness through awareness. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, (S Goldwasser, ed.), 214–226. ACM.

Federici M, Tomioka R, Forré P (2021). An information-theoretic approach to distribution shifts. In: *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, (M Ranzato, A Beygelzimer, Y Dauphin, P Liang, J Wortman Vaughan, eds.), 17628–17641. Curran Associates Inc.

Graunt J (1662). *Natural and Political Observations Made upon the Bills of Mortality*. Royal Society, London.

Gupta N, Khatri K, Malik Y, Lakhani A, Kanwal A, Aggarwal S, et al. (2024). Exploring prospects, hurdles, and road ahead for generative artificial intelligence in orthopedic education and training. *BMC Medical Education*, 24: 1544. https://doi.org/10.1186/s12909-024-06592-8

Haltaufderheide J, Ranisch R (2024). The ethics of ChatGPT in medicine and healthcare: A systematic review on large language models (LLMs). *npj Digital Medicine*, 7: 183. https://doi.org/10.1038/s41746-024-01157-x

Hardt M, Price E, Srebro N (2016). Equality of opportunity in supervised learning. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16*, (DD Lee, M Sugiyama, U von Luxburg, I Guyon, R Garnett, eds.), 3323–3331. Curran Associates Inc., Red, Hook, NY, USA.

Ioannidis JPA (2005). Why most published research findings are false. *PLoS Medicine*, 2(8): e124. https://doi.org/10.1371/journal.pmed.0020124

Kim MP, Ghorbani A, Zou J (2019). Multiaccuracy: Black-box post-processing for fairness in classification. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES'19*, (V Conitzer, GK Hadfield, S Vallor, eds.), 247–254. Association for Computing Machinery.

Kleinberg J, Mullainathan S, Raghavan M (2017). Inherent trade-offs in the fair determination of risk scores. In: *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* (CH Papadimitriou, ed.), volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Koh PW, Sagawa S, Marklund H, Xie SM, Zhang M, Balsubramani A, et al. (2021). WILDS: A benchmark of in-the-wild distribution shifts. In: *Proceedings of the 38th International Conference on Machine Learning (ICML)* (M Meila, T Zhang, eds.), volume 139 of *Proceedings of Machine Learning Research*, 5637–5664. PMLR.

Lifton RJ (1986). *The Nazi Doctors: Medical Killing and the Psychology of Genocide*. Basic Books, New York.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979). The Belmont report: Ethical principles and guidelines for the protection of

human subjects of research. *Technical report*, U.S. Department of Health, Education, and Welfare, Washington, D.C.

Nightingale F (1858). *Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army*. Harrison, London.

Nuremberg Military Tribunals (1949). *Trials of War Criminals Before the Nuremberg Military Tribunals Under Control Council Law No. 10, Vol. 2*. U.S. Government Printing Office, Washington, DC.

Quionero-Candela J, Sugiyama M, Schwaighofer A, Lawrence N (2009). *Dataset Shift in Machine Learning*. MIT Press.

Reverby SM (2009). *Examining Tuskegee: The Infamous Syphilis Study and Its Legacy*. University of North Carolina Press, Chapel Hill.

Shao M, Li D, Zhao C, Wu X, Lin Y, Tian Q (2024). Supervised algorithmic fairness in distribution shifts: A survey. In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024*, (K Larson, ed.), Jeju, South Korea, August 3–9, 2024, 8225–8233. ijcai.org.

Surbakti FPS (2025). Systematic literature review on generative ai: Ethical challenges and opportunities. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 16(5): 307–315.

Tabassum A, Elmahjub E, Padela AI, Zwitter A, Qadir J (2025). Generative ai and the metaverse: A scoping review of ethical and legal challenges. *IEEE Open Journal of the Computer Society*, 6: 1–15. https://doi.org/10.1109/OJCS.2025.3536082

Weerts H, Pfisterer F, Feurer M, Eggensperger K, Bergman E, Awad N, et al. (2024). Can fairness be automated? Guidelines and opportunities for fairness-aware automl. *Journal of Artificial Intelligence Research*, 79: 639–677. https://doi.org/10.1613/jair.1.14747

World Medical Association (2013). Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA*, 310(20): 2191–2194. https://doi.org/10.1001/jama.2013.281053