



# Fairness Hacking: The Malicious Practice of Shrouding Unfairness in Algorithms

Kristof Meding<sup>1,2</sup> · Thilo Hagendorff<sup>3</sup>

Received: 25 April 2023 / Accepted: 2 November 2023  
© The Author(s) 2024

## Abstract

Fairness in machine learning (ML) is an ever-growing field of research due to the manifold potential for harm from algorithmic discrimination. To prevent such harm, a large body of literature develops new approaches to quantify fairness. Here, we investigate how one can divert the quantification of fairness by describing a practice we call “fairness hacking” for the purpose of shrouding unfairness in algorithms. This impacts end-users who rely on learning algorithms, as well as the broader community interested in fair AI practices. We introduce two different categories of fairness hacking in reference to the established concept of p-hacking. The first category, intra-metric fairness hacking, describes the misuse of a particular metric by adding or removing sensitive attributes from the analysis. In this context, countermeasures that have been developed to prevent or reduce p-hacking can be applied to similarly prevent or reduce fairness hacking. The second category of fairness hacking is inter-metric fairness hacking. Inter-metric fairness hacking is the search for a specific fair metric with given attributes. We argue that countermeasures to prevent or reduce inter-metric fairness hacking are still in their infancy. Finally, we demonstrate both types of fairness hacking using real datasets. Our paper intends to serve as a guidance for discussions within the fair ML community to prevent or reduce the misuse of fairness metrics, and thus reduce overall harm from ML applications.

---

Opinions expressed are solely our own and do not express the views or opinions of our employees or other affiliated parties.

---

✉ Thilo Hagendorff  
thilo.hagendorff@uni-tuebingen.de

Kristof Meding  
kristof.meding@uni-tuebingen.de

<sup>1</sup> Data Protection Authority of Baden-Württemberg, Stuttgart, Germany

<sup>2</sup> University of Tübingen, Tübingen, Germany

<sup>3</sup> University of Stuttgart, Stuttgart, Germany

**Keywords** Fairness metrics · Fairness hacking · P-hacking · Algorithmic bias · AI ethics

## 1 Introduction

Machine learning (ML) algorithms that influence decision-making are being increasingly used within high-stakes fields such as employment, health, policing, trust score assessment, and lending. The trend in ML algorithms used within these fields has rendered the problem of algorithmic discrimination an important emerging issue (Caton and Haas, 2020). Discrimination, legally defined as the unfavorable treatment of individuals due to their belonging to certain demographic groups, is prohibited by international law. In order to establish ethical and trustworthy artificial intelligence (AI) as the field moves forward, it is paramount that new algorithm development prioritizes fairness-aware learning algorithms as well as the avoidance of machine bias (Jobin et al., 2019). While a vast portion of current publications in the field describes how to define and generate approximations of fairness in AI (Madaio et al., 2020), an equal proportion of the current literature describes shortcomings in AI fairness performances or shortcomings of the fairness concepts themselves (Weinberg, 2022).

The current study aims to provide a comprehensive and conclusive summary of the shortcomings of current fairness-aware learning algorithms. We describe the adversarial practice of fairness hacking. Fairness hacking describes the adding or removing of (sensitive) attributes from the analysis to lead outsiders to believe that fairness is ensured. Fairness hacking also facilitates the illegitimate exploitation of the plurality of fairness definitions. This occurs when tests are performed using an array of different fairness metrics, but only metrics that produce positive results are reported. During the process of fairness hacking, the fairness shortcomings in ML algorithms are cloaked by reporting that these algorithms have been deemed fair by certain metrics, while any metrics that have produced negative results are suppressed. Fairness hacking often hinges on using the fulfillment of a very specific or intentionally downscaled dimension of fairness to comprehensively report that an algorithm is fair.

Arguably, the most famous case of fairness hacking unfolded in 2016 when Northpointe—the company that developed COMPAS, a tool for assisting in decisions on whether defendants awaiting trial are too dangerous to be released—defended against allegations from ProPublica. ProPublica published an article claiming that COMPAS is biased against black individuals (Angwin et al., 2016). COMPAS uses more than 100 factors to calculate a risk factor for defendants, but race is not included in these factors. Nevertheless, Angwin et al. (2016) claim that for individuals who did not re-offend, black individuals were more likely to receive a medium or high score than white individuals. However, Northpointe claimed that when picking a score between 0 and 10, the calculated chance of recidivism is approximately equal for black and white defendants. In this case, we see the impossibility theorem at work because it is impossible to fulfill Angwin et al.’s as well as Northpointe’s fairness criteria at once (Corbett-Davies et al., 2016). Since the overall recidivism rate for black defendants is higher, and the proportion of individuals who re-offend is roughly equal for black and white individuals given an arbitrary risk score, it is inevitable that black people

who do not re-offend are assigned a higher risk score than white people who do not re-offend. If COMPAS would fulfill the fairness criteria that Angwin et al. propose, the software could still be viewed as unfair because one would have to systematically assign white individuals higher risk scores than black individuals. This approach is problematic as well. In this instance, we certainly cannot assert that ProPublica or Northpointe intentionally engaged in fairness hacking. However, the ensuing discussion about the appropriate use of fairness metrics perfectly illustrates the problem that at least inter-metric fairness hacking presents.

In this study, we will begin by describing analogies between fairness and p-hacking, since both practices work by selecting and reporting only specific results for data analysis. Next, we will describe intra-metric as well as inter-metric fairness hacking. The former stands for the misuse of particular fairness metrics by adding or removing attributes in the analysis. The latter involves the search for a specific fair metric with given attributes, inverting the impossibility theorem. To conclude the current study, we perform both types of fairness hacking using real datasets. This demonstration provides insight into how fairness hacking may work in real-world scenarios. We hope that this study in its totality will spark discussions on adversarial practices in ML within the fairness, accountability, and transparency (FAT) community. We hope that this will prevent or reduce the misuse of fairness metrics when applied to ML algorithms.

## 2 Related Work

Recent work on fairness in ML has demonstrated that fairness targets can be thwarted via technical means. By corrupting data, markedly adversarial attacks against fairness-aware learners can, for example (Jo et al., 2022), thwart efforts to render AI systems fair. In particular, gradient-based poisoning attacks can cause fairness-relevant classification disparities across specific groups (Solans et al., 2020). While this field of research investigates the adversarial introduction of algorithmic discrimination, fairness hacking, in contrast, is about the adversarial introduction of algorithmic fairness. Other adjacent studies have recently compared different fairness metrics without looking at confidence intervals (Žliobaitė, 2017). This research would benefit from the addition of confidence intervals, as calculating confidence intervals for group-based metrics directly works against fairness hacking. Despite this benefit, very few research studies include confidence or credible intervals (Besse et al., 2018; Ji et al., 2020; Ding et al., 2021; Cherian and Candès, 2023; Roy and Mohapatra, 2023). In addition to the previously mentioned work, other researchers have investigated the performance, as well as techniques used to remove bias, of different datasets and metrics (Biswas and Rajan, 2020).

Many different fairness definitions have led theoretical researchers to stipulate various “fairness traps” (Selbst et al., 2018) or “fairness limitations” (Buyl and De Bie, 2022). Both fairness traps and fairness limitations can prevent fairness targets from being achieved (Hoffmann, 2019; Green and Viljoen, 2020). Fairness traps include the failure to define fairness when applied to a wider sociotechnical system (John-Mathews et al., 2023), the problem of re-purposing algorithmic solutions within different social

contexts where these solutions are inappropriate, and the failure to formalize fairness in purely mathematical terms (Weinberg, 2022; Castelnovo et al., 2022; Mitchell et al., 2021). The existence of these fairness traps supports the idea of a non-ideal approach to fairness (Fazelpour and Lipton, 2020). This non-ideal approach dismisses the idea that the gap between the status quo and conditions of perfect fairness can be closed at all. Metrics for quantifying deviations from fairness are imperfect, and therefore fairness should be framed as a non-ideal methodology that does not include possessing ideal standards and norms. Moreover, one of the critical aspects of fair machine learning is the a priori definition of sensitive categories. As highlighted by John-Mathews et al. (2023), these categories are often grounded in social constructs and should be carefully selected based on ethical principles and legal guidelines. This aligns with the “realist” approach to fairness, as discussed by John-Mathews et al. (2023), where fairness is grounded in existing demographic categories produced by institutions. The choice of fairness metrics should not be arbitrary but guided by underlying philosophies of justice. By adopting such a comprehensive approach, investigators can ensure that the machine learning algorithms they develop are not only technically sound but also ethically robust.

Additionally, the impossibility theorem indirectly supports the idea of a non-ideal approach to fairness by showing that no existing method of algorithmic fairness can satisfy different fairness conditions all at once (Kleinberg et al., 2016; Saravanakumar, 2021). Thus, the impossibility theorem renders fairness into an unachievable ideal. This insight can easily be misused for political reasons. For example, it may be invoked to serve the political motives of those who aim to eliminate efforts to ensure fairness in ML models entirely. However, even for those who seek to ensure fairness within ML models, fairness does not achieve perfect standardization or meet metrics that model perfection (Xivuri and Twinomurizi, 2021; Hanna et al., 2020; Beutel et al., 2019). Phenomena such as Simpson’s paradox (Sharma et al., 2022) and fairness gerrymandering disadvantaging subgroups despite group fairness (Kearns et al., 2018) further underpin this insight and are similar to fairness hacking. Even in light of these caveats, efforts to increase fairness should still be attempted (Holstein et al., 2019; Lum et al., 2022). However, measures to increase fairness are always overshadowed by the aforementioned shortcomings. Furthermore, these shortcomings can be harnessed to purport an unfair algorithm to be fair. One way of doing this is through fairness hacking.

## 2.1 P-hacking in the Sciences

The term “fairness hacking” as well as its underlying definition were inspired by the concept of p-hacking. Here, we will give a brief introduction to p-hacking before discussing fairness hacking. Essentially, p-hacking is the practice of rendering results that are not statistically significant to appear statistically significant. This is achieved by searching through different combinations of data and statistical approaches until settling on the approach and data input that render the results statistically significant. In the sciences, p-hacking causes an inflation of false positive results (Ioannidis, 2005). This is particularly harmful because false scientific findings circulate. These findings can, in turn, influence other researchers who base follow-up studies on the false positive results

of these studies. Researchers have an incentive to manipulate research results because established scientific publication practices mostly reward positive results (Fanelli, 2012). Negative results are seldom published, and it can be challenging to reproduce positive results (Open Science Collaboration, 2015). Moreover, there is strong pressure within the scientific community to produce novel, significant results. For some researchers, this can outweigh the pressure to adhere to sound, rigorous methodology when conducting experiments. These factors coalesce in a way that makes p-hacking tempting for researchers and data analysts of all disciplines. In response, countermeasures to p-hacking have already been developed.

There are two main categories of measures to counteract p-hacking, which will be referred to as non-technical and technical countermeasures. The use of non-technical countermeasures refers to changing entrenched publication practices. One example of a non-technical countermeasure is the practice of registering studies and experiments before performing them with a neutral party (Van't Veer and Giner-Sorolla, 2016; Simmons et al., 2021). Another example of a non-technical measure to counteract p-hacking is the publication of negative results (Mehta, 2019). If both of these countermeasure examples were combined, researchers could present their experimental project plan in detail before performing the experiment. Then, the experimental results could be published regardless of the experimental outcome, with both positive and negative results considered worthy of publication. In this example, the pressure on the researcher to publish positive results is reduced. In contrast, technical measures to reduce or prevent p-hacking are directed against specific problems such as the multiple comparison problem (Westfall and Young, 1993). The multiple comparison problem describes a situation in which variables are added to an analysis until a significant result is found. In this example, any non-significant variables are ignored. In an example of a multiple comparison problem, an analysis is being performed wherein researchers examine the influence of different variables on high blood pressure. For this analysis, the researchers hypothesize that one variable, out of a potentially unlimited number of variables, has an influence on high blood pressure. In this example, an alpha level of 5% is set for the statistical tests performed as part of this analysis. In this context, a simplified explanation of the alpha level determines how likely it is that a variable is found to influence high blood pressure, although the variable does not truly influence blood pressure. This describes the likelihood of a false positive result. When the analysis is performed, the researchers find that sleep duration has a significant effect on hypertension, but the researchers do not correct the alpha level to account for various multiple comparisons with other variables. When the researchers test an infinite number of measurement variables without accounting for the number of variables, they will certainly find a variable that has a significant influence on blood pressure, even if this variable's influence is not actually true. To prevent such random findings when testing multiple variables, corrections have been proposed (Shaffer, 1995). A very simple correction is the unweighted Bonferroni correction (Bonferroni, 1936). The unweighted Bonferroni correction states that the original alpha level in the analysis must be divided by the number ( $n$ ) of variables tested  $\alpha_{new} = \frac{\alpha}{n}$ . Thus, the alpha level is corrected and adjusted according to the number of tested variables. In

this review, we will discuss the Bonferroni correction as it applies to fairness hacking (see Section 5)<sup>1</sup>.

## 2.2 From P-hacking to Fairness Hacking

The problem of p-hacking has a long and sad tradition in science. We argue that it is important to build on the work done on p-hacking since fairness hacking is an extended version of p-hacking. In classic p-hacking, researchers are interested in producing statistically significant results. In fairness hacking, however, it depends on the investigator's point of view whether the desired result is the significant or the non-significant one. For example, a company may be interested in a result that is not significant and therefore does not indicate discrimination. Therefore, the goal can be in contrast to p-hacking to influence the results in the significant as well as in the non-significant direction.

Furthermore, fairness hacking can occur in at least two variants: intra-metric or inter-metric fairness hacking. Suppose we test the model for discrimination against different attributes (race, sex, age, etc.) within one specific metric. In analogy to medicine and p-hacking, where one would examine the influence of different groups on hypertension by means of a single test (e.g., a t-test), we run into the multi-comparison problem since we test different groups. We will refer to this type of fairness hacking as intra-metric fairness hacking.

Another big difference in fairness hacking when compared to p-hacking is that there are many different definitions of what fairness entails. Therefore, there are different concepts of how fairness should be measured. This is exploited in inter-metric fairness hacking. A metric is searched until the desired result (discrimination against or not against a group, depending on the desired result) is shown. Then only this single metric, and not all other tested metrics, is reported.

Are intra-metric and inter-metric fairness hacking related? Both practices describe the process of hacking fairness. However, while the former is more related to classical p-hacking, the latter is only possible when multiple metrics are used. Due to the nature of the fairness metrics, they do not necessarily overlap in their outcomes - as implied by the impossibility theorem. Therefore, we argue that both intra- and inter-metric fairness hacking can be used to render an algorithm fair or unfair, even though both practices depict different procedures and involve different methods.

In the following, we will discuss intra- and inter-metric fairness hacking in detail.

## 3 Methods

Before we start to describe our methods, we would like to clarify our terminology, which is borrowed from Barocas et al. (2019). With the *features*, we describe (random) variables used as input for a machine learning algorithm. With (*sensitive*) *attributes*,

<sup>1</sup> Please note that the Bonferroni correcting in our examples is used because of the simple implementation and for the sake of the argument. However, there are concerns regarding the reliability (White et al., 2019) and methods such as Bayesian statistics or bootstrapping approaches could be used for mitigation (Ji et al., 2020; Dimitrakakis et al., 2019; Foulds et al., 2020; Cherian and Candès, 2023).

we describe (random) variables on which the fairness of the outcome of a machine learning algorithm is analyzed. Typical sensitive attributes are race, age, or gender. Within one attribute-such as age-we have different *groups* such as young and old.

Our main results are grounded in the use of different fairness metrics. We focus primarily on group-based metrics because we believe that they are the most commonly used in applications. The following group-based metrics, or specifically the difference between the metrics, are used: average odds, base rates, equal opportunity, error rate, false omission, false positives, predictive parity, statistical parity, and true negatives; see Bellamy et al. (2018) for definitions. Furthermore, we use the Theil index (Speicher et al., 2018) and consistency (Zemel et al., 2013) as an in-between measure of individual and group-based metrics.

First, we use synthetic data for our investigation. In this investigation, we assume that there were 100 participants in our dataset. Every participant has 1,000 randomly Bernoulli (binary) distributed attributes.<sup>2</sup> These attributes could be, for example, an age variable (young vs. elderly or villager vs. townspeople). In our *theoretical* analysis, the accuracy of our algorithm does not influence the analysis since all attributes share the same accuracy. The accuracy of our hypothetical algorithm is 75%. One might ask why we used a simulation-based approach for the comparison of group metrics since here one could also calculate analytical solutions. We used a sampling-based approach on purpose because with a sampling approach, we were easily able to compare the group-based metrics to two individual-based metrics.

Individual fairness metrics rely on the assumption that similar individuals should be treated similarly. For the comparison of individuals, therefore, every hypothetical member has one continuous Gaussian distributed ( $\mathcal{N}(\mu = 0, \sigma^2 = 1)$ ) feature. The calculation of confidence intervals is important. The group-based metrics are essential differences in binomial distribution except for the average odds metrics. For the average odds metric, one averages the true- and false positive ratio between the groups. It is possible to calculate the confidence intervals of differences for the ratio  $\Delta p$  of two binomial distributions with various methods (Newcombe, 1998). We used the Wald method, basically a normal approximation

$$\Delta p \pm z_\alpha \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \quad (1)$$

where  $n$  denotes the number of samples in group one and group two respectively, and  $z_\alpha$  denotes the standard  $1 - \frac{\alpha}{2}$  quantile of the standard normal distribution with confidence level  $\alpha$ . We check numerically that the Wald confidence interval method yields a good coverage in our setting.

Besides using synthetic data, we use real data from the Medical Expenditure Panel Survey included in the "AIF360" package provided by (Bellamy et al., 2018) to assess fairness issues in machine learning. We include each binary attribute (134 in total) in our analysis. We train a linear logistic regression classifier provided by AIF360 to predict utilization. The training and test split percentages are 90% to 10%. Additionally, data and the introductory example from (Ding et al., 2021) are used in the appendix.

<sup>2</sup> The binary setting here is for illustration purposes only. Our results hold equally for non-binary settings.



We use the python package "AIF360" (v. 0.4.0) to compute different metrics and then include Medical Expenditure Panel Survey data for our analysis, matplotlib (v. 3.5.2) for plotting, and the statsmodels package (v. 0.13.2) for calculation of confidence intervals. The folktables dataset is used in v. 0.0.11.

## 4 Results

In the following chapter, we will present our results. First, we discuss intra-metric fairness hacking. We will argue that this is a variant of p-hacking. Second, we discuss inter-metric fairness hacking. Here we will argue that inter-metric fairness hacking is a specific misuse of fairness metrics inherent to these methods. Finally, we show that fairness hacking is possible on real datasets.

### 4.1 Intra-metric Fairness Hacking as a Variant of P-Hacking

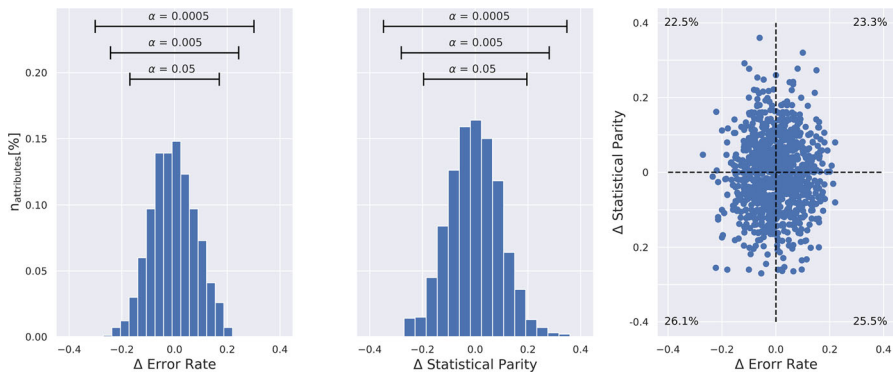
Assume that two of our favorite computer scientists, Alice and Bob, participate in a credit scoring study, e.g., similar to the German credit card dataset.<sup>3</sup> Alice and Bob have a number of attributes (age, race, gender, etc.). For simplicity, we argue here that Alice and Bob do not share any group across the attributes-thus, we assume mutually exclusive group membership for all attributes. For the sake of our argument, we next assume an ideal-world scenario. There is no intentional bias on any attribute in the dataset with respect to Alice's and Bob's groups. An algorithm now decides whether a loan is given or not. After one year, the performance of the algorithm is evaluated. Alice and Bob want to investigate whether the algorithm is biased against their respective groups. They agree on a common group-based fairness metric such as equal opportunity. After carefully inspecting the data, both Alice and Bob argue that the algorithm is biased against them for some attributes. How can this happen?

Let us have a look at our data in Fig. 1 (left). Here, we plot a histogram of the equal-opportunity difference of the algorithm between all 1,000 attributes. By chance, and inherent to random nature, some group membership identifier will indicate a bias against Alice's or Bob's group. By cherry-picking their favorite group, Alice and Bob can argue that the algorithm is biased against them. Furthermore, an outsider can even argue that the algorithm is not biased since some groups yield zero difference in the metrics. It is important to note that values to the right and to the left both indicate biases of the algorithm since the sign of the value depends on the arbitrary choice of whether the protected group within the attributes is coded zero or one.

Of course, this implies severe problems. In our case, we introduced by design no bias against attributes *in the synthetic dataset*. Simultaneously, one can find arguments for and against an arbitrary bias. Let us say that this result is obvious. Datasets do always contain biases against certain groups, but maybe only regarding irrelevant attributes. Relevant for fairness discussions are always attributes that include groups that are protected or sensitive. However, we want to highlight the following point: We used

<sup>3</sup> For more information, please refer to [https://archive.ics.uci.edu/ml/datasets/statlog+\(german+credit+data\)](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)).





**Fig. 1** Intra-metric fairness hacking for error rate and statistical parity. Distribution (in percent) of the error rate (left) or statistical parity (center) difference between 1,000 randomly assigned attributes (with binary values) for the outcomes of our hypothetical ML algorithm. A value of zero corresponds to no bias against a group. Values to the left *or* to the right indicate a bias against groups. Upper horizontal lines indicate confidence intervals with different alpha levels. Right: Scatter plot of statistical parity against error rate. The numbers in the corners indicate which percentage of the data falls in the respective quadrant

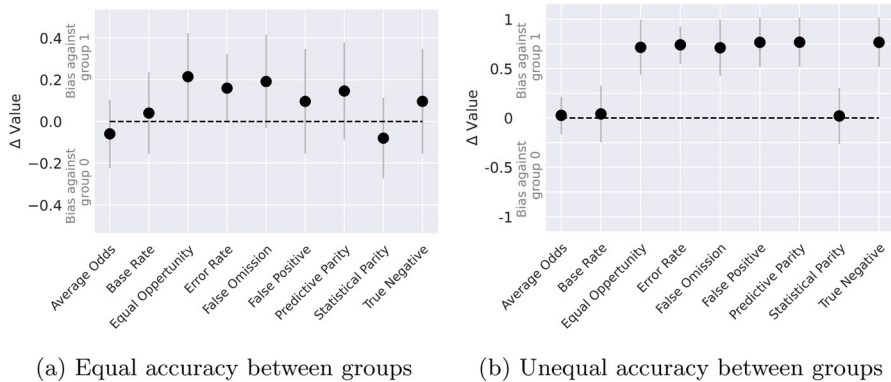
1,000 attributes for illustration purposes. However-and very important-the issue of intra-metric fairness hacking is still present if one uses fewer attributes and/or datasets with a smaller number of participants since smaller datasets tend to have more extreme average values (Wainer, 2007). We will discuss inter-metric fairness hacking in real datasets that have attributes (sex, age, race, geographical location, etc.) in Section 4.3.

We argue that when intra-metric fairness hacking happens within one fairness metric, it is a form of traditional p-hacking. Group-based fairness metrics are, in essence, a form of signal detection theory within confusion matrices. Thus, intra-metric fairness hacking is a form of the multiple comparisons problem. Luckily, it follows that the designer of our credit score study can use a well-known mechanism of protection against p-hacking to similarly protect against fairness hacking. Here, it is important to highlight that properly adapted confidence intervals, indicated in Fig. 1 by the vertical lines, are one method to avoid intra-metric fairness hacking (also see the Discussion section for different approaches).

Does intra-metric fairness hacking depend on the chosen metric? This is not the case. We want to highlight that intra-metric fairness hacking is inherent to fairness metrics. For showing this, we once more plotted Fig. 1 with equal opportunity and statistical parity in the appendix; see Fig. 4. Furthermore, intra-metric fairness hacking is not only possible in group-based methods but also in individual fairness metrics-the Theil index (Speicher et al., 2018) and Consistency (Zemel et al., 2013)-as shown in Fig. 6 and 7 in the appendix.

## 4.2 Inter-metric fairness hacking: inverse of the impossibility theorem

From now on, we assume that Alice and Bob agree to analyze one single attribute. This single attribute is used to evaluate the fairness of the algorithm. In Fig. 2a), we plot the difference between Alice's (group 0) and Bob's (group 1) groups across a



**Fig. 2** Inter-metric fairness hacking as the inverse of the impossible fairness theorem. a) Plotted is the difference between group 1 and group 0 of one binary attribute. The binary attribute is randomly distributed across groups. The hypothetical algorithm has 75% accuracy for both groups. Vertical lines indicate a Bonferroni-corrected confidence interval: alpha level =  $\frac{0.05}{12}$ . Figure a) has the same plot conventions as b), but the accuracy difference is set to 0.7

broad range of metrics. It is important to note that plotting fairness metrics without proper confidence intervals would be misleading. For some metrics, group 1 would be seen as unfairly treated; and for other metrics, group 0 would be seen as the one that is unfairly treated. Plotting confidence intervals shows that there is no significant effect for differences across metrics.

How does the situation change if we intentionally insert a bias in our dataset? To highlight our argument, we inserted a very strong accuracy difference in our algorithm. For Fig. 2a), we assumed an accuracy difference of 0.7 between group 0 and group 1 for our hypothetical algorithm. Here, we now see that several algorithms show a significant bias against group 1. However, and that is very important to note, three metrics do not indicate a bias: average odds, base rates, and statistical parity. Depending on the metrics, Alice and Bob can argue for or against discrimination in group 1. Please note that this effect is not an effect of classic p-hacking since there is a significant effect caused by the accuracy difference. We also used Bonferroni-corrected confidence intervals. Here, choosing the metric instead of the sensitive attribute makes the difference between inter-metric and intra-metric fairness hacking.

Inter-metric fairness hacking has a relation to the well-known impossibility theorem (Kleinberg et al., 2016). The impossibility theorem claims that it is not possible to satisfy the three metrics of demographic parity, equalized odds, and (positive/negative) predictive parity at the same time. The theorem focuses on the consensus of fairness metrics. However, we have a different focus. We focus on the non-consensus of fairness metrics. The impossibility theorem also implies that at most, two of the three metrics can be fulfilled simultaneously. Therefore, at least one metric does not agree with the other two metrics. This disagreement of metrics is what we define in a more general approach-beyond demographic parity, equalized odds, and positive/negative predictive parity-as inter-metric fairness hacking.

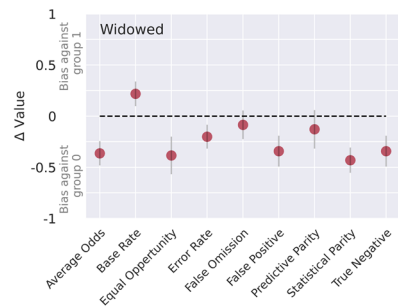
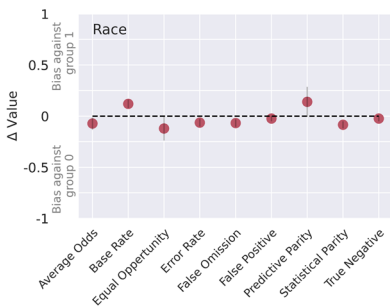
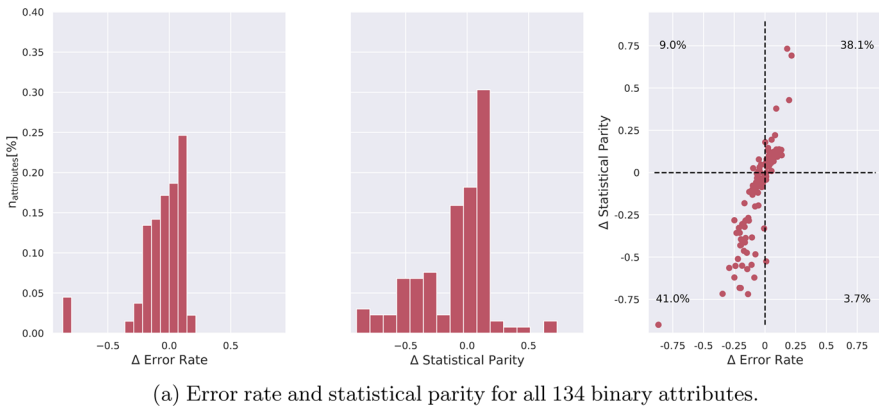
Inter-metric fairness hacking is different from classic p-hacking. It is also different from the problem of choosing the right test statistic in classic hypothesis testing. The correct test statistic depends on the data distribution and assumptions about it. In contrast, from a purely statistical viewpoint, using demographic parity or predictive parity is equally valid.

Next, we demonstrate fairness hacking scenarios using real data.

### 4.3 Fairness Hacking in Real-World Datasets

We demonstrate intra- and inter-metric fairness hacking using real data in Fig. 3. Here we used the Medical Expenditure Panel Survey (MEPS) dataset. The goal is to forecast the utilization of patients based on different attributes. We used 138 features to predict utilization, 134 of which are binary attributes used in the fairness analysis. Accuracy on the test split was 86.4%. Figure 3a shows the distribution of the difference between the error rate and statistical parity for all 134 binary attributes. For a comparison of the equal opportunity metric and statistical parity, see Fig. 5 in the appendix.

We will use the race attribute from the MEPS dataset for intra-metric fairness hacking. Table 1 shows the result for corrected and uncorrected confidence inter-



**Fig. 3** Fairness hacking in the wild for the Medical Expenditure Panel Survey data

**Table 1** Intra-metric fairness hacking for the race attribute in the MEPS dataset

	Avg. $\Delta$	Unc. alpha level	Corrected alpha level
Equal Opportunity	- 12.1%	[-23.8%, -0.3%]	<b>[-33.4%, 9.3%]</b>
Error Rate	- 6.4%	[-10.1%, -2.7%]	<b>[-13.1%, 0.4%]</b>
Statistical Parity	- 8.4%	[-11.8%, -5.0%]	[-14.6%, -2.2%]

Comparison of the uncorrected  $\alpha = 0.05$  and corrected  $\alpha_{corr} = \frac{0.05}{134}$  levels. Bold font indicates a metric changing from significant to non-significant results (including zero) by using the Bonferroni correction

vals for all three metrics (statistical parity, equal opportunity, and error rate) from Fig. 3a. The adjusted confidence intervals are calculated with the Bonferroni correction. Intra-metric fairness hacking could occur for the first two metrics since the corrected confidence intervals include the zero-value and thus a significant result change toward a non-significant result. This indicates the importance of adjusting confidence intervals. For the error rate metric, uncorrected and corrected confidence intervals yield significant results.

In addition to intra-metric fairness hacking, inter-metric fairness hacking is likewise a serious risk. We show inter-metric fairness hacking for two attributes (race and widowed) in Fig. 3b and c. When we look at the race attribute, we see that 4 out of 12 metrics indicate a bias against white people, while the remaining 8 metrics indicate a bias against black people. For the widowed attribute, we see that 2 out of 12 metrics indicate a bias against widowed people while the remaining metrics indicate a bias against the group of non-widowed people. A malicious machine learning engineer could use a selected metric to argue for a bias against one or the other group, depending on the chosen metric while neglecting all other metrics.

Taken together, both inter- and intra-metric fairness hacking pose a risk during the development and monitoring of machine learning models.

## 5 Recommendations to Avoid Fairness Hacking

Fairness hacking can be (possibly) avoided to a certain degree. Hence, having pointed out the problems, we also suggest potential solutions for fairness hacking, both in its intra- as well as inter-metric form. We would like to differentiate between measures that can be implemented easily and immediately as well as measures that require time and concerted effort. We will borrow these ideas and measures developed in the context of p-hacking here, see insofar Head et al. (2015); Stefan and Schönbrodt (2023) as a good overview. Although complete guidelines on how to avoid fairness hacking are beyond the scope of our paper, we would like to give recommendations. These recommendations are not intended to be generally binding for the future, but rather to form guidelines from our point of view today.

## 5.1 Immediate Recommendations

As a malicious practice, fairness hacking cannot be aligned with the claims of good scientific practice. However, recommendations stemming from the field of p-hacking avoidance can help to avoid inter- and intra-metric fairness hacking as well.

Based on the results presented above, we recommend to always calculate uncertainty intervals. Reporting single values without reporting the significance of the result can be misleading. Choosing p-values for confidence intervals then needs critical thinking in advance (Wasserstein and Lazar, 2016). However, just because results are significant, this does not necessarily mean that the results have an impact on the real world, too. There can be small effects that are highly significant (Cumming, 2014; Sullivan and Feinn, 2012). For example, measuring the height of two groups of adults can give a highly statistical difference of 1 mm. However, this effect may have negligible consequences in practice. To prevent this style of reporting statistically significant but otherwise hardly relevant results, the concept of effect size has been developed (Cohen, 1969; Sullivan and Feinn, 2012). Only effects with a sufficient magnitude should be considered statistically sound. This also applies for fairness metrics in the context of significance testing. In machine learning studies, only significant fairness effects with sufficient effect size should be considered.

Intra-metric fairness hacking is kind of similar to p-hacking. Hence, excluding variables in post-analysis is problematic Nosek et al. (2012); John et al. (2012), as well as building the hypothesis after the results are known (also called Harking, Kerr (1998)). These practices are problematic due to them allowing to find significant results through hypothesis testing easily. Similarly, in fairness hacking, some individuals might have an interest in finding a significant discrimination while others might have the opposite interest. To avoid these pitfalls, we recommend two methods based on the p-hacking literature (Head et al., 2015; Ioannidis, 2005). First, one builds a clear hypothesis of which fairness attributes should be included in the analysis *before* one analyses the data. This becomes even more important since the sensitive attributes are grounded on social categories (John-Mathews and Cardon, 2022). Here, defining a priori sensitive categories based on hypothesis, principles, and regulations before applying fairness metrics is a preventive measure for fairness hacking. In this context, it is also imperative to recognize that sensitive categories are not mere attributes available for arbitrary selection, but are deeply rooted in social constructs. Traditional statistical methods, which have been foundational in shaping our understanding of the social world, are now being re-evaluated in the face of evolving AI models that seek to capture the intricate nuances of social life. This involves grounding the definition of sensitive categories in established ethical principles and sociological theories. In addition to defining sensitive categories a priori, one can label studies as explanatory research (Tukey, 1977; Cumming, 2014) —which is valuable on its own— without a predefined hypothesis. Here, researchers might also include variables after collecting data. By labeling studies as explanatory research, recipients can treat the result with caution (Head et al., 2015). Furthermore, when testing multiple variables —be it hypothesis-driven or in exploratory research—, confidence intervals need to be adapted accordingly. When not following these recommendations, it becomes difficult to detect

intra-metric fairness hacking since one simply cannot know the study's confidence and whether multiple testing issues were considered.

Inter-metric fairness hacking shares challenges that another type of studies, namely meta-analyses, are facing. In meta-analyses, separate studies with different methods are compared and summarized, see (Borenstein et al., 2021). To avoid problematic outcomes due to the different statistical methods used in different studies, it was recommended to consider potential failures of test designs (Carter et al., 2019).<sup>4</sup> Similarly, in inter-metric fairness hacking scenarios, different statistical tests need to be compared. It is highly important to think in advance about which metric is appropriate to apply, e.g. practitioners can choose a metric based on a theory of justice (John-Mathews et al., 2023). The decision about which metric to use needs to be adequately justified. Contrary to intra-metric fairness hacking, exploratory research may not hold as much value in terms of inter-metric fairness hacking. It may not be appropriate to report every fairness metrics as different metrics suggest distinct perspectives on fairness, which can lead to contradictions. Only methodologically justified metrics should be taken into consideration and reported. Conversely, one could argue that individuals who are affected by algorithmic decisions should be given priority, favoring participatory approaches. If all fairness metrics are displayed, it must be clarified why the results of some metrics have been dismissed. Besides, presenting all metrics aids in uncovering hidden morally relevant assumptions.

All the above-described recommendations can in principle be applied immediately.

## 5.2 Recommendations in the Long run

We would like to discuss four future directions to prevent fairness hacking which are again largely borrowed from the statistics literature: pre-registration, new datasets, checklists, and building good scientific practices.

A way to prevent p-hacking in the context of studies is pre-registration (Wagenmakers et al., 2012; Chambers, 2013). Here, studies and methods are defined and pre-registered in journals before collecting and analyzing data. This ensures that researchers have to deal with potential problems with their methodology, experimental designs, or metrics before constructing, training, and testing a model. The process could then include a discussion about which metrics should be used, including their advantages and disadvantages. Here, Gundersen (2021) discusses pre-registered reports in the context of ML research which could be adapted to fairness issues.

In addition to registered reports, new datasets or methods can also help mitigate the issues. In our synthetic data, we have shown that fairness hacking can occur from purely statistical noise alone. Through the MEPS data, we show that fairness hacking also occurs in real data, which may have an imbalance in accuracy between groups. Nevertheless, we think that approaches such as those from Ding et al. (2021), which provide a balanced dataset of the known UCI adult dataset, are a reasonable approach. As in the example, we show in the appendix that intra-fairness hacking possibilities are reduced; see Fig. 8.

<sup>4</sup> For example, if studies use t-tests one should check when these methods fail (such as small non-normal distributed samples, Lumley et al. (2002))

In the context of p-hacking, checklists and guidelines for the prevention have been presented Wicherts et al. (2016). These guidelines should help researchers to avoid p-hacking and facilitate good statically reporting. In the context of machine learning, among others, datasheets for datasets (Gebru et al., 2021) and model cards for model reporting (Mitchell et al., 2019) have been introduced. These checklists focus on the whole life-cycle of machine learning models and their datasets. Also, checklists for fairness were already developed (Madaio et al., 2020; Agarwal and Agarwal, 2023). We want to stress that these checklists and guidelines are a first good step in the right direction. However, we hope that the machine learning community will bring up further checklists, especially with regard to the problem of inter- and intra-metric fairness hacking. For this purpose, the calculation of confidence intervals and comparing metrics should, next to becoming good scientific practice, be included in the software packages for the calculation of metrics. Furthermore and so far, we know that none of the standard fairness packages have implementations of confidence intervals by default.

For both types of fairness hacking discussed in this paper, the awareness in the machine learning community has to be raised. Both AI ethics and technical fairness research should include reflections on the potential dangers of fairness hacking into their considerations (Mehrabani et al., 2021). Often, journals do require solid statistical tests to prevent p-hacking. Otherwise, papers are desk-rejected. This procedure is missing with respect to fairness metrics, although they are statistical tests, too. Here it might make sense that journals need to take fairness issues more into account to prevent the practices described in this study.

In summary, care must be taken during the full life-cycle of development, application and reporting of a machine learning algorithm to justify the choice of a metric whose aptitude for a specific purpose should be critically reflected using not just moral intuitions, but at best sound ethical considerations that take into account all affected individuals. In addition to that, the proprietary nature of many algorithms makes it difficult to detect, let alone rectify, instances of fairness hacking. When algorithms are closed off from public scrutiny, they may mask inherent biases or intentional manipulations that distort the true fairness of the system. Furthermore, the incentives driving fairness hacking are distinct from those behind p-hacking. While p-hacking arises from the pressure to produce statistically significant results, often leading researchers to manipulate or cherry-pick data to achieve this, fairness hacking may stem from commercial or social pressures to showcase unbiased AI models, prompting developers to superficially adjust outputs without addressing underlying biases.

## 6 Summary and Outlook

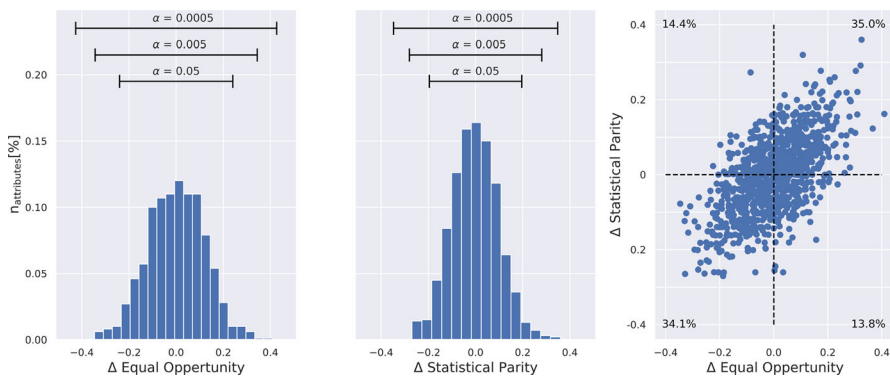
Bias reduction efforts require a mix of technical and non-technical solutions. At least when looking at technical solutions, interventions to increase fairness tend to cluster into associated modes, meaning that different fairness preservation measures strongly correlate with one another (Friedler et al., 2018). Nevertheless, our paper demonstrates



that fairness interventions can be brittle when engaging in intentional fairness hacking. In this context, one has to stress that several phenomena resemble fairness hacking practices, but are not explicitly called fairness hacking. These range from inversion of the impossibility theorem to phenomena such as Simpson's paradox or fairness gerrymandering. For utility-based approaches, one must also keep in mind that the choice of approach depends on the specific problem and the desired trade-off between accuracy and fairness, although this trade-off cannot be considered fairness hacking in the sense proposed in our paper. Ultimately, actual fairness hacking causes a twofold burden. If one picks a single fairness metric and claims that a particular algorithm is unfair without shedding light on the outcomes of alternative metrics, a distorted image of an algorithm's performance is created. On the other hand, if one lets several metrics compete with each other, fairness issues become relative. Both burdens should be avoided. To do this, one must always consider fairness metrics in a given sociotechnical context. One must consider when human judgment and moral intuitions are necessary when weighing different metrics against each other in a given situation. In fact, each fairness metric has an underlying moral assumption that can be explicitly defined (Heidari et al., 2018). Whereas the term "impossibility theorem" suggest that fairness metrics have the same status and can hence be played off against each other—a circumstance that is inversely exploited by fairness hacking practices—this does not hold on closer examination and when taking into account the normative presuppositions embedded in each fairness metric as well as the suitability of particular fairness metrics in a given context (Ruf and Detyniecki, 2021; Friedler et al., 2016; Gajane and Pechenizkiy, 2017). Thus, our results do not just underpin the importance of being aware of the ease of claiming a machine learning model to be fair; they also stress the importance of reflecting on the normative values embedded in fairness metrics.

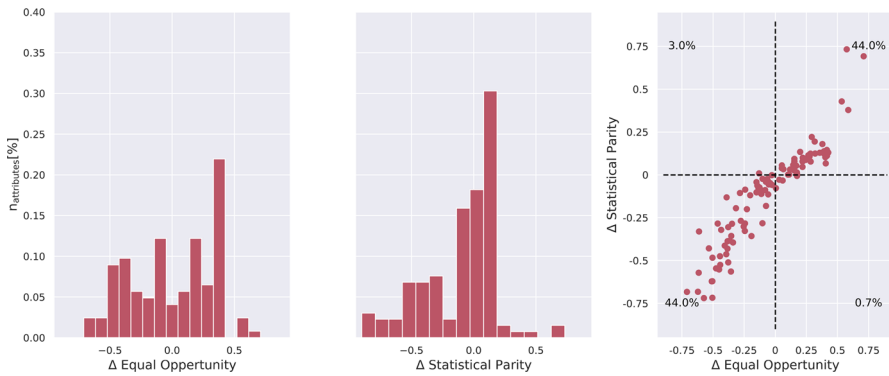
## Appendix

### A.1 Figure 1 with equal opportunity and statistical parity



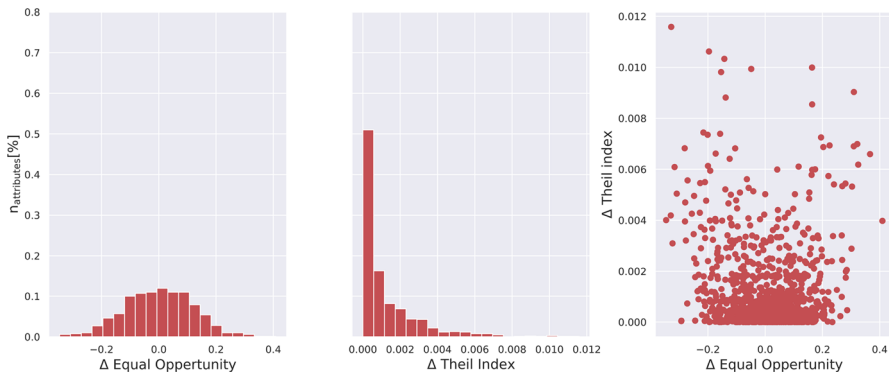
**Fig. 4** Intra-metric fairness hacking for equal opportunity and statistical parity. Plot conventions as in Fig. 1

## A.2 Equal opportunity and statistical parity in the wild

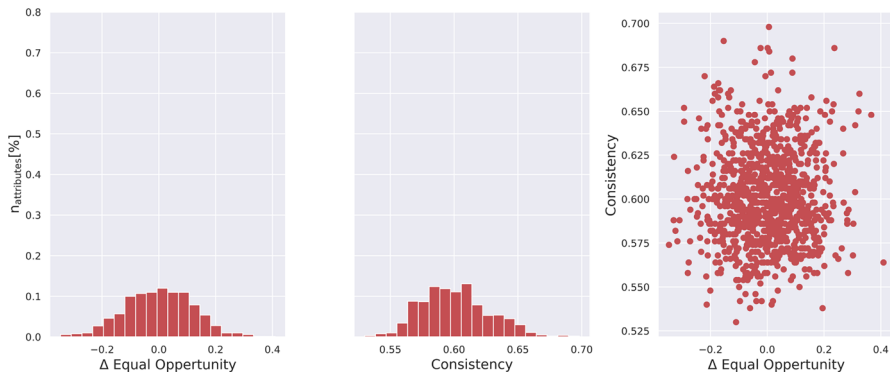


**Fig. 5** Equal opportunity and statistical parity for all 134 binary attributes. Plot conventions as in Fig. 1. The number of items belonging to the protected or unprotected group differs between actual attributes in the MEPS dataset (race, sex, etc.). Thus, we cannot calculate a single null hypothesis confidence interval

## A.3 Individual fairness metrics

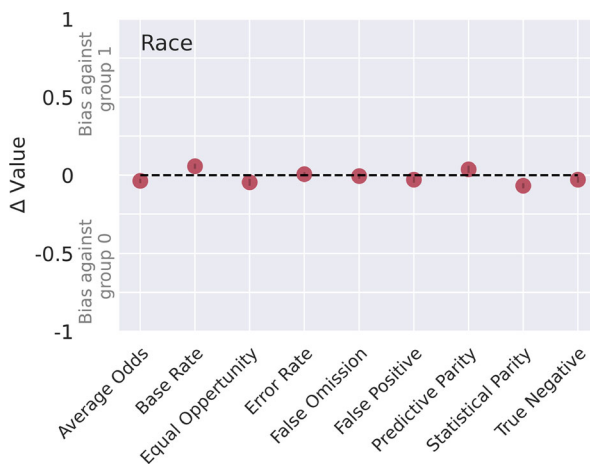


**Fig. 6** Equal opportunity and the Theil index for our hypothetical data. Plot conventions as in Fig. 1



**Fig. 7** Equal opportunity and consistency for our hypothetical data. Plot conventions as in Fig. 1

#### A.4 Inter-metric fairness hacking for the Folktables dataset



**Fig. 8** Data from the Folktables dataset. The provided logistic regression example model had an accuracy of 77.2%. Only the race attribute is used. Plot conventions as in Fig. 2

**Acknowledgements** We would like to thank Frank Jäkel, Moritz Hardt and Larissa Höfling for valuable feedback. Additionally, we thank our two anonymous reviewers for constructive feedback.

**Author Contributions** The project was initialized by KM during his work at the University of Tübingen and later jointly developed forward with TH during KM's work at the data protection authority of Baden-Württemberg. KM wrote the code for the analysis with input from TH. Both authors planned, structured, and wrote the manuscript.

**Funding** Open Access funding enabled and organized by Projekt DEAL. Funding was provided by Deutsche Forschungsgemeinschaft (DFG; the German Research Foundation)-project number 276693517-SFB 1233, TP 4 Causal inference strategies in human vision for KM and his work at University Tuebingen. TH was supported by the Baden-Württemberg Ministry of Science, Research, and the Arts under Az. 33-7533-9-19/54/5 in Reflecting Intelligent Systems for Diversity, Demography and Democracy (IRIS3D), as well as the Interchange Forum for Reflecting on Intelligent Systems (IRIS) at the University of Stuttgart.

**Data Availability** Not applicable

**Code Availability** Code will be available upon reasonable request.

## Declarations

**Competing interests** The authors have no competing interests to declare.

**Ethics approval** Not applicable

**Consent to participate** Not applicable

**Consent for publication** Both authors fully consent to publication

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Agarwal, A., Agarwal, H. (2023). A seven-layer model with checklists for standardising fairness assessment throughout the ai lifecycle. *AI a*
- Angwin, J., J. Larson, S. Mattu, and L. Kirchner. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks
- Barocas, S., M. Hardt, and A. Narayanan. (2019). Fairness and Machine Learning: Limitations and Opportunities. *fairmlbook.org*. <http://www.fairmlbook.org>
- Bellamy, R.K.E., 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. Richards, D. Saha, P. Sattigeri, M. Singh, K.R. Varshney, and Y. Zhang (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias
- Besse, P., del Barrio, E., Gordaliza, P., & Loubes, J. M. (2018). Confidence intervals for testing disparate impact in fair learning. [arXiv:1807.06362](https://arxiv.org/abs/1807.06362)

- Beutel, A., Chen, J., Doshi, T., Qian, H., Woodruff, A., Luu, C., ... & Chi, E. H. (2019). Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 453–459).
- Biswas, S., Rajan, H. (2020). Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. In *Proceedings of the 28th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, pp. 642–653.
- Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8, 3–62.
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.
- Buyl, M., & De Bie, T. (2022). Inherent Limitations of AI Fairness. [arXiv:2212.06495](https://arxiv.org/abs/2212.06495)
- Cardon, D., & John-Mathews, J. M. (2023). The displacement of reality tests. The selection of individuals in the age of machine learning. *Distinktion: Journal of Social Theory*, 24(2), 217–240.
- Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, 2(2), 115–144.
- Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1), 4209.
- Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey. *ACM Computing Surveys*
- Chambers, C. D. (2013). Registered reports: a new publishing initiative at Cortex. *Cortex*, 49(3), 609–610.
- Cherian, J. J., & Candès, E. J. (2023). Statistical Inference for Fairness Auditing. [arXiv:2305.03712](https://arxiv.org/abs/2305.03712)
- Cohen, J. (1969). Statistical power analysis for the behavioral sciences. Academic press
- Corbett-Davies, S., E. Pierson, and S. Goel. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7–29.
- Dimitrakakis, C., Liu, Y., Parkes, D. C., & Radanovic, G. (2019). Bayesian fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 509–516.
- Ding, F., Hardt, M., Miller, J., & Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 6478–6490.
- Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3), 891–904.
- Fazelpour, S., Lipton, Z.C. (2020). Algorithmic fairness from a non-ideal perspective. In Markham, A., Powles, J., Walsh, T., Washington, A. L. (Eds.), *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, pp. 57–63. ACM
- Foulds, J.R., Islam, R., Keya, K.N., Pan, S. (2020). Bayesian modeling of intersectional fairness: The variance of bias. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pp. 424–432. SIAM
- Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4), 136–143.
- Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D. (2018). A comparative study of fairness-enhancing interventions in machine learning
- Gajane, P., & Pechenizkiy, M. (2017). On formalizing fairness in prediction with machine learning. [arXiv:1710.03184](https://arxiv.org/abs/1710.03184)
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.
- Green, B., Viljoen, S. (2020). Algorithmic realism. In Hildebrandt, M., Castillo, C., Celis, E., Ruggieri, S., Taylor, L., Zanfir-Fortuna G. (Eds.), *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, pp. 19–31. ACM
- Gundersen, O. E. (2021). The case against registered reports. *AI Magazine*, 42(1), 88–92.
- Hanna, A., Denton, E., Smart, A., Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. In M. Hildebrandt, C. Castillo, E. Celis, S. Ruggieri, L. Taylor, and G. Zanfir-Fortuna (Eds.), *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York, pp. 501–512. ACM
- Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, 13(3), e1002106.

- Heidari, H., Loi, M., Gummadi, K.P., Krause, A. (2018). A moral framework for understanding of fair ml through economic models of equality of opportunity
- Hoffmann, A. L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7), 900–915.
- Holstein, K., Vaughan, J.W., Daumé, H.III., Dudík, M., Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need?
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), 696–701.
- Ji, D., Smyth, P., & Steyvers, M. (2020). Can i trust my fairness metric? assessing fairness with unlabeled data and bayesian inference. *Advances in Neural Information Processing Systems*, 33, 18600–18612.
- Jo, C., Sohn, J.Y., Lee, K. (2022). Breaking fair binary classification with optimal flipping attacks
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524–532.
- John-Mathews, J. M., & Cardon, D. (2022). The crisis of social categories in the age of ai. *Sociologica*, 16(3), 5–16.
- John-Mathews, J.M., Mourat, R.D., Ricci, D., Crépel, M. (2023). Re-enacting machine learning practices to enquire into the moral issues they pose. *Convergence: The International Journal of Research into New Media Technologies*. <https://doi.org/10.1177/13548565231174584>
- Kearns, M., Neel, S., Roth, A., Wu, Z.S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In J. Dy and A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning*, Volume 80 of *Proceedings of Machine Learning Research*, pp. 2564–2572. PMLR
- Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196–217.
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. [arXiv:1609.05807](https://arxiv.org/abs/1609.05807)
- Lum, K., Zhang, Y., & Bower, A. (2022). De-biasing “bias” measurement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 379–389)
- Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23(1), 151–169.
- Madaio, M.A., Stark, L., Wortman Vaughan, J., Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In Bernhaupt, R., Mueller, F.F., Verweij, D., Andres, J., McGrenere, J., Cockburn, A., Avellino, I., Goguy, A., Bjørn, P., Zhao, S., Samson, B.P., Kocielnik, R. (Eds.), *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, pp. 1–14. ACM
- Madaio, M.A., Stark, L., Wortman Vaughan, J., Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–14
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6), 1–35.
- Mehta, D. (2019). Highlight negative results to improve science. *Nature*
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220–229)
- Mitchell, S., Potash, E., Barocas, S., D’Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8, 141–163.
- Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17(8), 857–872.
- Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6), 615–631
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science* 349 (6251), aac4716
- Simmons, P. J., Nelson, L. D., & Simonsohn, U. (2021). Pre-registration: Why and how. *Journal of Consumer Psychology*, 31(1), 151–162.
- Roy, A., Mohapatra, P. (2023). Fairness uncertainty quantification: How certain are you that the model is fair? [arXiv:2304.13950](https://arxiv.org/abs/2304.13950)

- Ruf, B., & Detyniecki, M. (2021). Towards the right kind of fairness in AI. [arXiv:2102.08453](https://arxiv.org/abs/2102.08453)
- Saravanakumar, K. K. (2021). The Impossibility Theorem of Machine Fairness—A Causal Perspective. [arXiv:2007.06024](https://arxiv.org/abs/2007.06024)
- Selbst, A.D., boyd, d., Friedler, S.A., Venkatasubramanian, S., Vertesi, J. (2018). Fairness and abstraction in sociotechnical systems. *ACT Conference on Fairness, Accountability, and Transparency (FAT)* 1 (1), 1–17
- Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, 46(1), 561–584.
- Sharma, R., Kaushik, M., Peious, S. A., Bertl, M., Vidyarthi, A., Kumar, A., & Draheim, D. (2022). Detecting simpson's paradox: A step towards fairness in machine learning. In S. Chiusano, T. Cerquitelli, R. Wrembel, K. Nørvg, B. Catania, G. Vargas-Solar, & E. Zupan (Eds.), *New Trends in Database and Information Systems* (pp. 67–76). Cham: Springer International Publishing.
- Solans, D., Biggio, B., Castillo, C. (2020). Poisoning attacks on algorithmic fairness.
- Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018). A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2239–2248)
- Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, 10(2), 220346.
- Sullivan, G. M., & Feinn, R. (2012). Using effect size-or why the P value is not enough. *Journal of Graduate Medical Education*, 4(3), 279–282.
- Tukey, J. W., et al. (1977). *Exploratory data analysis* (Vol. 2). MA: Reading.
- Van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology-A discussion and suggested template. *Journal of Experimental Social Psychology*, 67, 2–12.
- Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6), 632–638.
- Wainer, H. (2007). The most dangerous equation. *American Scientist*, 95(3), 249.
- Wasserstein, R.L., Lazar, N.A. (2016). The asa statement on p-values: context, process, and purpose
- Weinberg, L. (2022). Rethinking fairness: an interdisciplinary survey of critiques of hegemonic ML fairness approaches. *Journal of Artificial Intelligence Research*, 74, 75–109.
- Westfall, P. H., & Young, S. S. (1993). Resampling-based multiple testing: Examples and methods for p-value adjustment (Vol. 279). John Wiley & Sons
- White, T., van der Ende, J., & Nichols, T. E. (2019). Beyond Bonferroni revisited: concerns over inflated false positive research findings in the fields of conservation genetics, biology, and medicine. *Conservation Genetics*, 20, 927–937.
- Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in psychology*, 1832
- Xivuri, K., & Twinomurizi, H. (2021). A systematic review of fairness in artificial intelligence algorithms. In *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society: 20th IFIP WG 6.11 Conference on e-Business, e-Services and e-Society, I3E 2021, Galway, Ireland, September 1–3, 2021, Proceedings 20* (pp. 271–284). Springer International Publishing
- Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *International conference on machine learning* (pp. 325–333). PMLR
- Žliobaite, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4), 1060–1089.