Algorithmic Fairness in Criminal Justice: Challenges and Solutions

A Bachelor Thesis in Artificial Intelligence

Brage Hamre Skjørestad

April 2024

Abstract

This thesis covers the topic of algorithmic fairness in criminal justice risk assessments. There are many definitions of fairness, and some of the most common views are discussed. They have a tendency to be in opposition with one another, and with accuracy. A fair algorithm should also be accurate, so an exploration of these trade-offs is required. After presenting the situation, there is also a discussion of what measures we can take for improvement, where multiple perspectives are included.

1 Introduction

Machine learning is integrating itself more and more into our society. Numerous fields of business are trying to reap its uses, even the field of criminal justice. Algorithms have been introduced to risk assessments, which have resulted in a great deal of scrutiny and coverage, due to the massive consequences it can entail. The criticism is usually derived from the unfair patterns an algorithm can base its outcomes on, for example with variables such as race and gender Berk et al. (2017).

In this thesis, the paper Berk et al. (2017) is the primary source and focus. They explore the tradeoffs of algorithmic fairness, and clarify the situation that the field of criminal justice finds itself in. If algorithms are to be implemented, we need to understand why it makes any given decision. Therefore, an explanation of the framework that prior data constructs is necessary. Moreover, different implementations of fairness are explored, and their conflict with one another and accuracy. There are also different types of accuracy, but there is much less ambiguity when deciding which one to implement. The question of fairness is potent in many fields of study, and it is always ambiguous.

In this context, the best approach we can take is to examine the different definitions and their relationships, and what consequences they entail. There does not exist a perfect algorithm, but perfection cannot be the objective. The task at hand, is how to take the best possible measures to improve the current standard of practice. In the field of criminal justice, each small step of improvement may impact the life of many.

2 Defining a Population Capable of Inference

Before we start thinking about how to introduce fairness into an algorithm, we need to reflect on how we are generating the framework for classification. The main purpose of machine learning in general, is to find patterns in real-life data and draw conclusions which can be used for future decisions. In the field of criminal justice, this could entail using a month of arraignments (formal document for a defendant's criminal charges) to draw inferences for a whole year Berk et al. (2017). If we are to use a dataset to extract information about future decisions, we need to be able to argue that our data is representative of the individuals it is to be implemented on. Therefore, we need to examine and understand how the data is generated. In machine learning, a common approach to this problem is to randomly sample data points from a single joint probability distribution. The distribution can be visualized as a high-dimensional space, where the dimensions are represented by the different variables in the data. Figure 1 and Figure 2 are graphs meant to clarify this point.

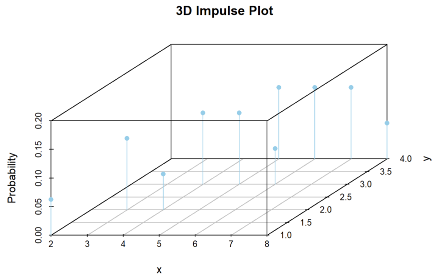

Figure 1: Visualization of a finite population. Ross (2022)

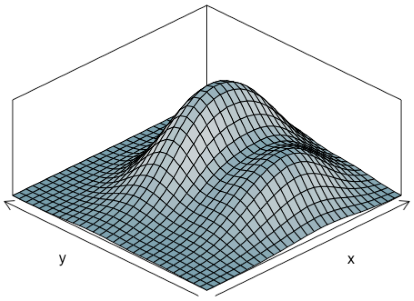

Figure 2: Visualization of an infinite population. MAPK (2019)

For illustration, consider these simple visualizations of a joint probability distribution. Note that Figure 1 and Figure 2 are not supposed to represent the same data, they are only for illustration purposes. In Figure 1, we have a finite population. We have some data points and have plotted them into a space. x and y could signify a single feature and a target variable, respectively. There is no distinction between a feature and a target variable in a joint probability distribution, but we can make that distinction ourselves Berk et al. (2018). We could try to predict new values of y with the population in Figure 1, but it would return a poor generalization ability because of the lack of information. Instead, we can define our population as the limitless amount of observations that we could produce from our distribution, visualized in Figure 2. The data points are no longer fixed, but randomly realized from an infinite population. We now have a statistical framework from which inferences can be made. This also allows us to view our population as IID data, meaning independently and identically distributed. When the data on hand has this quality, it is more representative and may yield better results.

To clarify, the data generation process is about extracting an infinite population from a finite one, and then randomly sample data points in order to attain IID data. However, in the context of criminal justice, one would need to argue that there were no changes in all relevant areas in the period the population is to be applied for Berk et al. (2017). The police, judges, prosecutors, and so on, cannot have seen any substantial change in operation. But if there has been no meaningful changes for a few years, you could argue that arraignments for a whole year is able to generalize. Defining our population in this manner is important if we are to implement a fair and accurate algorithm.

3 What is Fairness?

Fairness can be defined as "impartial and just treatment or behaviour without favouritism or discrimination." Oxford English Dictionary (2024). In machine learning, fairness is about trying to identify and correct algorithmic bias against protected groups. This means that an algorithm should not favour one group over another. But how do we measure favouritism and identify discrimination? There are many metrics, which unfortunately makes the question of fairness quite complicated. Before discussing them, we should first consider a confusion matrix.

Table 1: Table with Frequencies of Actual Outcome by Predicted Outcome From An Algorithm

		Predicted value		
		Positive	Negative	Total
Actual value	Positive	a	b	$a+b$
	Negative	c	d	$c+d$
	Total	$a+c$	$b+d$	N

a and d are the correctly identified values, true positive and true negative. b is called the false negative (a negative which should have been a positive), and c is the false positive (a positive which should have been a negative). N is the total number of cases, found by $(a+b+c+d)$. With these different values, we can extract metrics which represent a form of fairness. Also note that Table 1 and the rest of this thesis discusses a binary task, where there are only two classes for an observation, for example high-risk and low-risk. However, the same points apply for non-binary tasks.

3.1 Statistical Parity

Statistical parity is one of the most frequently utilized measurements of fairness. If applied, the proportion of both positive and negative outcomes ($\frac{a+c}{N}$, and $\frac{b+d}{N}$) should be the same across groups. The values that need to be equal are often called *base rates* Berk et al. (2017). In order to contextualize, consider the table below.

Table 2: A Fictive Example of Frequencies Of Actual Reoffenders by Predicted Reoffenders For Men and Women. Green and Red Represent Correctly and Incorrectly Identified Cases

		Predicted value (Men)		Predicted value (Women)	
		High-Risk	Low-Risk	High-Risk	Low-Risk
Actual value	High-Risk	200	50	50	20
	Low-Risk	50	200	30	100

In Table 2, we have an algorithm that classifies individuals as high-risk or low risk, and obtains these results for the two groups, men and women. The classification an individual receives determines if they are granted probation (high-risk) or prison time (low-risk). We can observe that statistical parity is not achieved in this example. Examining the proportion of outcomes, we find that the algorithm classifies men as high-risk 50% of the time, and 40% of the time for women. The fairness criterion demands that these two percentages are equal, so we have to adjust. Women who are not deemed a safety risk may still get sent to prison, or high-risk men must be released into society. This criterion tries to implement parity in a hugely disparate context, and will often unfairly punish groups to accommodate. It is a problem that individuals have to suffer unfair consequences because a fairness goal has to be met. However, it is still a very valid and pursued form of fairness. It builds on the idea that the group an individual belongs to is irrelevant, whatever the task at hand is. Unfortunately, obtaining the criteria without consequences is rare.

3.2 Conditional Measurements

Some other relevant criteria include *conditional procedure accuracy equality* ($\frac{a}{a+b}$, $\frac{d}{c+d}$) and *conditional use accuracy equality* ($\frac{a}{a+c}$, ($\frac{d}{b+d}$). To achieve one of them, the two formulas, though maybe different from one another, need to be equal across groups. They are very similar, but differ in that the former is conditioned on the true event, while the latter is conditioned on the predicted event. We can think of them as classification accuracy and prediction accuracy, respectively.

Conditional procedure accuracy is a combination of true positive rate and true negative rate. A model applying this fairness criterion would demand that the probability of correctly identifying a positive from all the positives in the sample, is equal across groups (and the same for negatives). Going back to the probation example, men and women need to be equally often identified as a safety risk when they actually are, and vice versa. This looks like a more precise metric of fairness because it examines when the model is incorrect.

Conditional use accuracy is subtlety different, it demands that the positive (and negative) predictions are equally often correct across groups. Now the focus is on the accuracy of the predictions themselves, not relative to the true event. Let us consider another example to clarify the difference.

Table 3: Another Fictive Example of Frequencies Of Actual Reoffenders by Predicted Reoffenders For Men and Women. Green and Red Represent Correctly and Incorrectly Identified Cases

		Predicted value (Men)		Predicted value (Women)	
		High-Risk	Low-Risk	High-Risk	Low-Risk
Actual value	High-Risk	200	50	80	20
	Low-Risk	75	300	40	160

In Table 3, conditional procedure accuracy is achieved between men and women. The algorithm is equally proficient at identifying the high-risk (and low-risk) individuals across the two groups. Recalling the formulas for calculation, we find that it achieves a score of 0.8 for both men and women. The true positive and true negative rates are equal, which means the same can be said for the false positive/negative rates. In this example, we can say we have achieved *equalized odds*, the errors do not discriminate. In contrast, conditional use accuracy equality is not achieved. $\frac{a}{a+c}$, also called *precision*, differs between men and women. It achieves a score of around 0.72 for men, and around 0.67 for women. The positive predictions are notably more often correct for men than for women. This happens because of the different sample sizes of a and d, which is not an unrealistic scenario. To summarize, Table 2 and Table 3 were meant to clarify the differences of the three fairness definitions. Next, we will discuss the challenges of acquiring an accurate algorithm together with the three fairness criteria.

4 Challenges with Implementing Fairness

4.1 Addressing Bias

When implementing fairness in machine learning algorithms, we first have to estimate the accuracy of our model. Then we can compare the estimates between groups. Training a model involves performing a chosen procedure on the dataset. What that procedure is depends on the data. By visualizing and examining the data, you can choose an appropriate procedure, which is the job of data scientists and analysts. Furthermore, after the procedure is performed, we obtain a function that is supposed to link our features to the target values, denoted $\hat{f}$.

Naturally, it will not do that perfectly. If the model performs too well on the dataset, it will do worse on new data. This is what the paper Berk et al. (2017) calls "estimating the true response surface", trying to make our function $\hat{f}$ as similar as we can to f. This results in the model being biased towards the training data. Berk et al. (2017) state that, even asymptotically (when the size of a dataset gets infinitely large), you cannot argue that the estimation will be unbiased. Trying to estimate the true response surface is therefore not a good solution. We can't find the "truth" from data.

However, there exists alternatives. Acknowledging that our function is only an approximation of the true response surface can result in smaller bias. This slight change in perspective can be crucial. Instead of trying to search for the true f function, we try to create a function with the best generalization ability. This is possible due to the approach in the data generation process. The IID data may grant the ability to $\hat{f}$ to be less biased in large samples. With a model that can generalize, we have a better stance for addressing the fairness of it.

4.2 Trade-offs

We denote L as legitimate parameters, such as number of prior convictions, and S as protected parameters, for example gender or race. Protected parameters refer to individual characteristic that are illegal to discriminate against. When performing the procedure, called $h(L, S)$, on training data to obtain $\hat{f}(L, S)$, some would argue we could remove the protected parameters (only implementing $h(L)$). However, this will reduce accuracy. Less accuracy is arguably also less fair. If every group experiences the same unsatisfactory accuracy, it may be fair across groups, but now unfair for everyone. Also, removing S does not mean certain groups will not receive unjust treatment.

The process must be to implement trade-offs between fairness and accuracy, and between the different measures of fairness. Every discussed metric can be found from confusion matrices, which makes us able to perform these trade-offs.

4.2.1 Impossibility Theorem

Before examining trade-offs, a mention of an "impossibility theorem" is required. "When the base rates differ by protected group and when there is no separation, one cannot have both conditional use accuracy equality and equality in the false negative and false positive rates" Berk et al. (2017), page 18. Base rate is the probability of an outcome for a given group. Usually, base rates differ across groups. Separation is achieved if there is perfect classification of a population.

Essentially, the theorem states if a population cannot be perfectly classified, base rates must be adjusted to achieve both of the conditional measurements. Alternatively, decision makers must choose which of them is better to implement.

4.2.2 Conditional Use Accuracy vs False Positive/Negative Rate

The theorem entails huge implications. If, in a particular setting, decision makers deem that conditional use accuracy is necessary for fairness, there will be disparity in the false positive/negative rates (and naturally vice versa). These are fairness criteria we would like to achieve simultaneously. Berk et al. (2017) presents some cases where this can happen. The problem is that they are of a very trivial nature.

For example, you might achieve it if you were to assign the same outcome class to everyone. In a case where 500 men and 50 women were all given the same outcome class, you achieve the same base rate. The example sets 400 men and 40 women to a true positive. This leads to conditional procedure accuracy equality, conditional use accuracy equality and statistical parity. Incredibly fair outcomes, but with a glaring issue. There is no prediction, just guessing.

Furthermore, if the size of the populations would slightly change, consequences follow. If there were 600 men, where 500 were a true positive, conditional use accuracy equality is lost. When the population changes without classification changing, the base rates are no longer equal. Equality in base rates entails great fairness results, but if we already had fairness in base rates, fairness in machine learning would not be as scrutinized as it is today. Then, the world we live in would be fairer already.

Alternatively, we can also achieve both of the desired fairness criteria by achieving separation. When the observations are separable, and when we have a $h(L, S)$ that can separate them, all our problems disappear. Unfortunately, that is not realistic. In the real world, if $h(L, S)$ achieves separation in a population, it means it is over-fitted and will perform worse on new data. In practice, decision makers have to make the trade-off choice between conditional use accuracy and false positive/negative rates, and possibly consider shifting base rates to accommodate for fairness.

5 Strategies in Application

We now arrive at strategies for applying fairness and accuracy in algorithms. After discussing and understanding the challenge of fairness, we will discuss the possibly most important question. What do we do? There exists unavoidable trade-offs, but we still want to create an algorithm that can be fair and accurate. Several strategies exist which can be divided into three categories, depending on the stage of development it operates. But they all try to correct algorithmic fairness in order to reduce the dramatic nature of the trade-offs.

5.1 Pre-Processing

Pre-processing is about finding unfairness in data before we construct our algorithm. “Garbage in, garbage out” is a common phrase in computer science, signifying that we cannot perform miracles on bad input. If the data is biased and unfair, the algorithm will likely also have those qualities. This is a huge problem, considering the history of humankind. In the United States of America, Black people could not vote in practice until the 1960s, which is “only” around 60 years ago. The racism can then infiltrate itself very subtlety into our data.

Fortunately, there exists techniques that attempts to mitigate these problems, such as residualizing procedures. Berk et al. (2017) presents a rather simple one, which is to try to predict legitimate parameters by using protected parameters (regression analysis). The resulting difference between the predicted values and actual values is called the residual. Performing this procedure, we obtain a transformed and ideally more fair data set. However, this transformed data set may still contain unfairness. The approach removes the contribution of S on L, but L may still be dependent on S in relation to the outcome. The impact of prior convictions on the outcome could still be different across groups.

Yet, it is positive that we are aware of these “interactions effects”. The challenge is that we need to be aware of all of them. This is perhaps a too large challenge, but we still want to make predictions without using information in S. We need to transform our data so that L is truly independent of S. There exists an approach to this, which starts by estimating the probability distribution of a chosen predictor of L, given a predictor of S Johndrow and Lum (2017). For instance, what is the probability distribution of prior convictions, given that the subject is a man? Furthermore, the approach would continue to estimate the distributions of the next predictors, but now given the sensitive and predictor and the previous predictors as a condition. The residuals can then be found with the real data. This procedure results in a transformed dataset closer to achieving independence than the previously explained one. By making predictors independent of race, you are working with an approach to treat all races as the same. But this is still just one notion of fairness. In some contexts, treating all races equally may yield unfair outcomes, as we found in the discussion of statistical parity between men and women.

Differing base rates across groups has large significance. Measures have been adapted to redistribute them during pre-processing, for example randomly changing the outcome value of a data point to achieve balanced base rates Berk et al. (2017). Explicitly altering your dataset in such a dramatic way naturally decreases accuracy, but there are no methods which do not entail any disadvantages. At last, a rather ambitious method involves randomly transforming (slightly) the legitimate predictors to make it less dependent on the sensitive attributes. The randomness makes it unclear how the method fits into a notion of fairness. In any case, the only option is choosing the best available method, not a fault-free method.

5.2 In-Processing

In-processing refers to fairness measures taken in the process of formulating our algorithm. This train of thought does not alter the dataset, but tries to push $h(L, S)$ in a certain direction. A simple example is changing the outcome of an observation containing substantial uncertainty. This may as well avoid decreasing accuracy considerably and can push us towards a fairness goal. Nevertheless, it could also be very consequential for the false positive and false negative rates. Even though the method uses uncertainty measurements, we are still altering outcomes, which can be unfair.

Another example is to, during development, penalize violations of conditional procedure accuracy. However, the simplicity of this loss function finds rather unsatisfying trade-offs between accuracy and the minimizing of prejudice. Berk et al. (2017) does not go as far into detail regarding in-processing techniques, but more methods can be found elsewhere.

5.3 Post-Processing

The final stage where we could make fairness adjustments is after the development of our algorithm. We call this post-processing. Berk et al. (2017) suggests that, a technique by Hardt et al. (2016) is possibly the best to date. Their notion of fairness demands that the predictions by $\hat{f}(L, S)$ is independent of all attributes in S given the actual outcome. Formalized as such:

$P(\hat{Y}|A = 0, Y = y) = P(\hat{Y}|A = 1, Y = y)$, $y \in \{0, 1\}$

This formulation entails that the value A (an attribute from S) should not influence the probability of a predicted outcome class. The false positive rate and false negative rate should be equal across both values of A. Conditional procedure accuracy equality is the fairness criterion to be satisfied. To achieve it, the method assigns a new value to each data point. The value represents the probability of switching to the other outcome. These probabilities, combined with a loss function to minimize error while still meeting the fairness criterion, can find a $\hat{f}(L, S)$ that achieves conditional procedure accuracy equality. The advantage of this approach is its simplicity, because we only need access to the outcome and sensitive attribute data. We do not need predictors of L, in contrast to the pre-processing strategy. Altering a training pipeline may often be more complex.

However, this simple approach is still subject to the impossibility theorem and may come at the cost of conditional use accuracy equality. Furthermore, accuracy may also be heavily affected by the reassignments. A reality may be that the worst performance by an unadjusted classifier on a group becomes the performance for every group Berk et al. (2017).

6 Related Work

6.1 A Greater Focus on In-Processing

Wan et al. (2023) emphasizes more on the potential of in-processing techniques, arguing that they can produce "intrinsically fair" models because fairness is implemented directly, not adjusting the data or outcomes, like the other methods does. They divide in-processing techniques into two categories; explicit mitigation and implicit mitigation.

6.1.1 Explicit Methods

When using explicit methods of in-processing, fairness is directly integrated into $h(L, S)$. The previously discussed in-processing techniques are of this nature. These methods could be implemented quite simply, and are usually done by regularization or with constraints. With the former, the idea is to penalize unwanted patterns which $h(L, S)$ may adopt. If it develops a strong connection between a sensitive attribute and the outcome, this will be penalized.

In contrast, the latter constructs $h(L, S)$ as it would otherwise, but with a defined fairness constraint. This constraint could be a statistical measure of dependency between predictors of S and L. During development, $h(L, S)$ cannot exceed this dependency measure. The core difference between regularization and constraints is that it restrains behaviour instead of punishing it. The explicitness of both of them usually result in achieving a fair outcome, or at least what was defined as a fair outcome.

6.1.2 Implicit Methods

Instead of reformulating the specific objectives, implicit methods seeks to mitigate bias in the "latent representation", which is the name for the hidden patterns

$h(L, S)$ finds in the data. These methods are specific deep models, but those models can be very powerful. In addition, they argue that in deep models, in-processing techniques are the most viable option. Re-training a complex deep model to remove unfairness can be too large of a task.

An example of these implicit methods is to remove correlation between L and S in the representations within the hidden layers. However, while bias may be mitigated, that does not necessarily entail fair predictions. The implicit methods can be very effective, and bias and unfairness can often be correlated, resulting in accurate and fair results. But they are restricted to deep models.

6.1.3 Findings

Wan et al. (2023) delves deeper into the solutions with a focus on in-processing. However, even though they are explored more at length, they still arrive at a very similar conclusion. The selection of fairness measurement can entail unfairness by other measurements, so long as its definition remains subjective. If we cannot agree on what metric measures actual fairness in society, the impossibility theorem will apply. Arriving at a singular definition is not only for computer scientists, and the choice of measurement is for practitioners to make.

6.2 Process Fairness: An Unconventional Perspective

Usually, the discussion of algorithmic fairness is about fairness in outcomes. Yet, alternative trains of thoughts can also be found. In 2016, a case was made for the consideration of "process fairness" Grgic-Hlaca et al. (2016). This perspective tackles fairness during the process of the decisions, and involves a more "human" morality. What characteristics about us are fair to use in decisions? Humans can often share the same moral values, at least at a basic level.

Grgic-Hlaca et al. (2016) found in a survey that most people thought "criminal history" was a fair feature to use, but that "family criminality" was unfair. To formalize this, we can measure the fairness of $\hat{f}(L, S)$ as the proportion of users that consider the features of L and S to be fair. A lot of weight is being put upon the participant's shoulders, but it is not the case that they must decide the fairness of a feature initially. Their judgement might change as they see how the feature influences the function. To adjust for these changes, three different measurements of process fairness are defined: feature-apriori, feature-accuracy and feature-disparity fairness.

The first measurement is the proportion of people that considered a certain feature unfair prior to understanding how it influences $\hat{f}(L, S)$. The second is the proportion of people that considered a feature unfair if it increases the accuracy of $\hat{f}(L, S)$. Finally, feature-disparity fairness is the proportion of users that thinks a feature is fair regardless of an increase in disparity. Process fairness will then demand that certain features must be left out, if the values of the three measurements demands it. This can also be described as feature selection, but in relation to fairness instead of accuracy or generalization ability.

6.2.1 Findings

Grgic-Hlaca et al. (2016) performed a survey where they tried to examine the implications of their defined notion of process fairness. In the case of predicting a re-offending individual, they used a logistic regression with various combinations of nine different features. They found that the most fair classifier only utilized one feature (prior convictions), but the accuracy was not much lower than the best classifier achieved (63.0% and 68.1%). Unfortunately, there is a trade-off here. The dropping of features leads to less accuracy, meaning a high accuracy and high process fairness cannot be achieved at the same time.

Fortunately, the survey found that achieving a high process fairness does not come at a large cost of outcome fairness, in the example of only using one feature. Accuracy is the larger cost, but process fairness in algorithms is still a relatively unexplored area. At this point, outcome fairness is proved to be the more viable option because it can retain more accurate models. Yet, they present an interesting point of view that is very different from most others, and not as deeply explored. It remains to be seen if this is a form of fairness for the future.

7 Conclusion

The challenge of finding fair and accurate algorithms in criminal justice can be, as we have seen, a difficult task. There is naturally a trade-off to be made between fairness and accuracy, but the trade-offs between the different implementations of fairness creates obstacles. Fairness has never been completely objective in all fields of study. In fact, it is impossible to achieve the two "conditional measurements" if there is disparity in outcomes. The fairness criteria are only met simultaneously in very specific examples. There is usually unfairness to be found, even with representative data. The data represents an unfair society with disparate base rates, and unfair algorithms are a natural consequence of that fact.

The main contribution of the paper Berk et al. (2017) is the presentation and discussion of the situation we are in. They formulate an impossibility theorem which complicates matters, but they also discuss the measures we can take. We can try to transform our data from unfair to fair, or make smart adjustments for more fair outcomes. There is no need to dismantle the algorithms, keeping in mind that criminal justice without them does not yield fair outcomes either. Their main objective was to clarify the tradeoffs, and emphasize that there is no way to avoid them, and that many of the decisions becomes political discussions, not scientific ones.

Nevertheless, Wan et al. (2023) delve deeper into potential solutions. They suggest that in-processing methods may be the best option, since they can result in "intrinsically fair" algorithms. Considering the rapid success accuracy of current algorithms, implementing fairness instead of adjusting for it may be smart. Yet, they still arrive at a very similar conclusion as Berk et al. (2017). I think that in-processing techniques can among the best methods of fairness corrections, especially for powerful deep models. But the fair and accurate outcomes is all that matters, so if in-processing should not be prioritized if they prove insufficient.

Moreover, process fairness is a very interesting perspective on fairness. I do not think we should train our models on features deemed unfair by the consensus. The problem with the perspective of Grgic-Hlaca et al. (2016) is that finding the consensus of each feature may be challenging, and, as of today, process fairness also encounters unsatisfying tradeoffs. The whole concept may have too many complications in general.

In the end, we need to keep in mind, like Berk et al. (2017) emphasizes, that deciding what is fair is not only a task for science. We can only strive to improve our algorithms and continue exploring corrections for fairness.

References

R. Berk, L. Brown, A. Buja, E. George, and L. Zhao. Working with misspecified regression models. *Journal of Quantitative Criminology*, 34, 09 2018. doi: 10.1007/s10940-017-9348-7.

R. A. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50: 3 – 44, 2017. URL `https://api.semanticscholar.org/CorpusID:12924416`.

N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller. The case for process fairness in learning: Feature selection for fair decision making. 2016. URL `https://api.semanticscholar.org/CorpusID:13633339`.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. 10 2016.

J. Johndrow and K. Lum. An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13, 03 2017. doi: 10.1214/18-AOAS1201.

MAPK. How to generate a 3d joint probability distribution for bivariate data (e.g., plotly)? https://stackoverflow.com/questions/57602608/how-to-generate-a-3d-joint-probability-distribution-for-bivariate-data-e-g-pl, 2019. Accessed: 2024-04-21.

Oxford English Dictionary. Oxford english dictionary, 2024. URL `https://doi.org/10.1093/OED/6800683964.` Online; accessed on 2024-15-04.

K. Ross. An introduction to probability and simulation. https://bookdown.org/kevin$_davisross/probsim - book/joint - distributions.html, 2022. Accessed : 2024 - 04 - 21.$

M. Wan, D. Zha, N. Liu, and N. Zou. In-processing modeling techniques for machine learning fairness: A survey. *ACM Trans. Knowl. Discov. Data*, 17(3), mar 2023. ISSN 1556-4681. doi: 10.1145/3551390. URL `https://doi.org/10.1145/3551390`.