Alternative Fairness and Accuracy Optimization in Criminal Justice

Shaolong Wu*, **James Blume, Geshi Yeung**

Harvard University lorrywu@g.harvard.edu, Massachusetts Institute of Technology jblume19@mit.edu, Harvard University geshiyeung@g.harvard.edu

Abstract

Algorithmic fairness has grown rapidly, yet key concepts remain unsettled in criminal justice. We review group, individual, and process fairness and map the conditions under which they conflict. We then develop a simple modification to standard group fairness. Rather than exact parity across protected groups, we minimize a weighted error loss while keeping differences in false negative rates within a small tolerance. This improves feasibility, raises accuracy, and highlights the ethical choice of error costs. We situate this proposal within three classes of critique: biased and incomplete data, latent affirmative action, and the explosion of subgroup constraints. Finally, we propose a practical framework for deployment in public systems, built on three pillars: need-based decisions, transparency, and narrowly tailored solutions. Together, these elements link technical design to legitimacy and provide actionable guidance for agencies that use risk assessment and related tools.

Keywords Algorithmic fairness, criminal justice, risk assessment, group fairness, individual fairness, process fairness, disparate impact, equalized odds

Introduction

The use of algorithms has become increasingly pervasive in modern society, with many important decisions now being made by computers. It is essential to ensure algorithms are designed fairly. This paper explores algorithmic fairness, its challenges, and implications for computer science. By examining existing research, it will identify key implications for the future of computer science and consider how algorithmic fairness can be achieved.

We propose a general alternative framework to rethink fairness. We hope to provide a general framework for the idea of fairness by offering guiding pillars that can be applied in a broad context. The paper covers the following: Popular definitions of fairness in machine learning, Criticisms of the current framework, and proposed Three Pillars of Fairness.

*Contact Author, 700 Soldiers Field Road, Boston, MA, 02163

What is Algorithmic Fairness?

Over the last decade, there has been a bewildering number of definitions of algorithmic fairness (Narayanan 21). Making matters even more problematic, many of these proposed ideas of fairness are often incompatible with each other. Increasing fairness in one sense may decrease it in another. This paper shall focus on the three following dimensions: Group Fairness, Individual Fairness, and Process Fairness.

Group Fairness

Definitions Group fairness entails that an algorithm does not treat different demographic groups systematically differently. This idea of fairness was popularized after multiple real-world algorithms from criminal justice, corporate hiring, and credit ratings were found to systemically discriminate against minority candidates. While these algorithmic harms were often not malevolent in intent, their damage was tangible (Slaughter, Kopec, and Batal 2020). In particular, historically biased data often lead to a biased model. Algorithmic group fairness is often defined as having equal error rates between the desired groups. For example, a model for credit ratings should have an equal false positive rate between racial groups. In legal terms, group fairness is concerned with disparate impact.

There are also a few mathematical notions of group fairness that are based on probability (Zhou 2022). First, there is demographic parity. This requires that there be the same proportion of individuals in any group receiving a positive outcome as the group's proportion of the population. Consider a binary classification setting, where a model has to make a prediction $\hat{Y} \in \{0, 1\}$, where $\hat{Y} = 1$ means the model predicts an individual to have a positive outcome, while $Y = 0$ means an individual is predicted to have a negative outcome. Then, let's say an individual has a group membership $S \in \{0, 1\}$, where $S = 0$ denotes membership to an underprivileged group while $S = 1$ denotes membership to a privileged group. Then, group parity is achieved if $P(\hat{Y} = 1|S = 1) = P(\hat{Y} = 1|S = 0)$, and hence this can be measured by computing $\frac{P(\hat{Y}=1|S=1)}{P(\hat{Y}=1|S=0)}$. If the result is close to 1, there is group parity. However, if base rates are different, meaning that $P(Y = 1|S = 1)$ and $P(Y = 1|S = 0)$ are different, then even a classifier that never makes prediction errors (i.e. $\hat{Y} = Y$) will have a group parity measure

that is not equal to 1, because $\frac{P(Y=1|S=1)}{P(Y=1|S=0)}$ is not equal to 1.

Another mathematical definition of group fairness is called equalized odds, which requires the same false positive rate across groups as well as the same true positive rate across groups. This can be formulated as requiring $P(\hat{Y} = 1|S = 0, Y = y)$ and $P(\hat{Y} = 1|S = 1, Y = y)$ to be the same for $y \in \{0, 1\}$.

The third main definition of group fairness is an equal opportunity, in which only the true positive rate is required to be equal across groups. Formally, this means $P(\hat{Y} = 1|S = 0, Y = 1)$ and $P(\hat{Y} = 1|S = 1, Y = 1)$ need to be the same. Note this requirement is a subset of the requirement for equalized odds.

The fourth definition of group fairness is calibration. It means that if the algorithm predicts the probability of a positive outcome to be p for a set of individuals, then we should expect a p portion of them to have a positive outcome. For example, for individuals predicted to have a high probability of recidivating, a large portion of them should actually have recidivated, which would mean that the algorithm is well-calibrated. Group fairness is achieved if calibration is held for different demographic groups (also called calibration within groups), meaning that for each demographic group, calibration should be held. Mathematically, in a binary classification setting, this means that $P(Y = 1|S = 0, \hat{Y} = 1)$ and $P(Y = 1|S = 1, \hat{Y} = 1)$ should be the same.

These many definitions of group fairness mean that it is often up to the discretion of the algorithm designer to decide which fairness notion to adopt. Once a notion is adopted, the algorithm designer can then use many different methods to try to achieve group fairness.

Methods for achieving group fairness *Summary.* We review pre-, in-, and post-processing approaches for enforcing group fairness; full details are provided in Appendix .

Individual Fairness

Another intuitive definition of fairness comes from the notion of individual fairness. This can be characterized as whether similar individuals are treated similarly. This can be thought of as similar to the traditional racial color-blind arguments (Kleinberg et al. 2018). If a member of one group must satisfy a great threshold to achieve the same loan as another group, then the individual may perceive an injustice. Specifically, in a racial context, the individual may argue they are facing harm purely due to the color of their skin. This notion of fairness is the common assumption of many American anti-discrimination laws. In legal terms, individual unfairness can be thought of as disparate treatment. Mathematically, this may be formalized as the average difference between the label of an individual and the average label of the k-nearest neighbors of that individual based on non-sensitive attributes.

Process Fairness

Unlike the other notions of fairness, which have been output-focused, process fairness concerns input fairness. In this framework, an algorithm gains legitimacy through having an open and transparent process (Grgic-Hlaca et al. 2016). While this idea is often under-discussed in computer science literature, it is very common in political science. At its core, fairness depends on whether people trust a given institution. An organization that is transparent both about its intentions and methods will be trusted more than an organization that obscures them. The value of process fairness is that it is robust to model errors and biased data because it does not depend on algorithmic outputs.

The Canonical Definition of Fairness and Its Critiques

The PAC setup for group fairness

Summary. We formalize equalizing false negative rates and the associated loss-minimization program; full derivations and examples appear in Appendix .

Critiques to the Canonical Definition of Fairness

Previous literature on group fairness has challenged its concept. The critiques are that the scheme is either insufficient or excessive.

Critique 1. *Inherent biases in data*

The training data may contain unforeseen biases (e.g., in a complex field such as crime), and the algorithm may fail to correct them, as documented by Mehrabi et al. (2021). This phenomenon has been carefully examined empirically in Chapman et al. (2022) using a comprehensive data set from the official UK Crime API that provides data on policing's impact on crime rates.

Specifically, the underlying logic is similar to weight-adjusting mechanisms. More weights are assigned to the predictor that predicts there will be recidivism in the areas with historically high crime rates (especially if there's been a crime in the past several days, then the probability another crime will happen is considered very high by the algorithm without fairness adjustment). The learning theory behind "Near Repeat Theory" is that a crime incident triggers a temporary increase in crime rates nearby. Even with completely randomized synthesized "historical" data, the existing predictive policing algorithms lead to biased feedback loops that further confirm the assumed pattern. This is simply because more police in the areas with a prior crime makes crime in these areas more identifiable and further strengthens the need to police that area.

Historically, in criminology (Gottfredson and Hirschi 1987), there's always a notion of positive criminology that believes crime to be the aggregate result of social issues such as psychological health problems, poverty, and social injustice. This thought is fundamentally against penalizing the "offenders" who are breaking laws as the last resort and not causing significant harm to the local community. From the positivist criminologist's view, the unavoidable biases within the prior data set make the framework of fairness insufficient to address unfairness.

Critique 2. *Latent Affirmative Action*

Any group fairness definition ends up using affirmative action or similar logic, because one may inevitably end up

giving an advantage to certain groups. As outlined by Lagioia (2022), since different groups have different base rates, a system that has the same accuracy for different groups may fail to comply with group-parity standards. The authors analyzed the performance of Correctional Offender Management Profiling for Alternative Sanctions, also known as COMPAS, in evaluating defendants' risk profiles and classifying individuals as being at a high risk of recidivism if the system assigns the individual a score higher than a certain threshold based on the individual's criminal history, education, income level, family situation, etc. The authors found that by using the same threshold for different groups, individual fairness is achieved, in which the system would assign the same score to individuals with an equal likelihood of recidivism and therefore make the same classification for the two individuals. However, by setting different thresholds for the different groups, individuals with the same score may be classified differently, which violates individual fairness. Also, if the scores of different groups are calibrated, meaning that they are "equally correlated with the predicted classifications," then having different thresholds would lead to lowered accuracy for at least one of the groups. This accuracy-fairness tradeoff may be of concern to many.

Overall, it appears that satisfying group fairness may result in unavoidable damage to individual fairness. Whether this is acceptable might depend on the quality of the input data: for example, if the input data is historically biased against certain groups, then having different thresholds allows the system to be calibrated such that predictions and probabilities can be aligned even with biased data. Also, depending on the policy goals, having different thresholds may be desired; "the goal of increasing diversity or balancing access to education, types of jobs, or positions" may be some examples in which we might compromise individual fairness for certain policy goals [Lagioia]. Nonetheless, this paper illustrates how group fairness may be incompatible with individual fairness, and therefore one must be careful when applying group fairness metrics. As the authors argue, this requires "discretionary value-based political choices" that statistical notions alone are insufficient to address.

Nonetheless, we know that individual fairness alone is not sufficient for ensuring a fair model either. As proposed by Fleisher (2021), a model that classifies the same outcome for every individual would satisfy individual fairness (same treatment for similar individuals), but it is clearly unfair. Instead, Speicher (2018) has proposed an index for overall fairness that can be decomposed into two components: between-group and within-group unfairness. They acknowledged that improving components may be to the detriment of the other, but the decision of which component to prioritize might rest upon the user of the algorithm.

Critique 3. *Vagueness of Fairness in Subgroups*

As raised in Kearns et al. (2018), the intersection of demographic groups poses a challenge to achieving equalization of false negative rates across all protected categories. There are hundreds of different possible intersections of demographic variables, such as black homosexual women with a college education, and white heterosexual men with associate's degrees. In most census or social surveys, there could be 6 different options under religion, 7 under race, 20 under country of origin, and 3 under sexual orientation. Naturally, this may introduce over one thousand categories with a considerable number of people in each of them. While the authors proposed an algorithm that relies on heuristics for learning to converge to the best subgroup-fair distribution over classifiers within polynomial time, it could be unclear how one should decide at what point we stop considering further intersection between subgroups.

This subgrouping problem leads to the following three issues:

1. There is an insufficient moral justification for why it suffices to not consider the intersection of the protected features, but just consider group fairness alone.
2. This practice is equivalent to introducing a thousand or even more linear constraints to a convex optimization problem, which could make the solution suboptimal or infeasible.
3. To get a fair representation of the intersections between many different demographic subgroups, a large sample size may be needed for each intersection, which is difficult to obtain.

Summary of Critiques

While none of these critiques alone serve to make group fairness unworkable, put together, they raise concerns over whether to adopt certain notions of group fairness. Moreover, not every issue with group fairness may be solved with technical solutions alone, but may also require value-based decisions. For example, the choice of definition of group fairness will depend on the values of those affected by the algorithm. All this goes to show that any working idea of group fairness requires a value-based framework. The paper will explore this idea more fully in the last section.

Modified Alternative Definition of Fairness

We develop this idea inspired by the related literature (Ho and Xiang 2020; Schoeffer, Kuehl, and Machowski 2022) ourselves as the following:

$$\min \sum_{i=1}^{n} \alpha FN(h, v_i) W_i$$

subject to

$$|FN(h, v_i) - FN(h, v_j)| \leqslant \tau, \forall i \neq j, 1 \leqslant i, j \leqslant n$$

where α, β are the loss of false negatives and false positives, and τ is the tolerance bound for the difference in false negative rates.

This setup has two benefits: (i) by fine-tuning the parameter τ, we can at least ensure that there's a feasible solution. (ii) Since the binding equalities are relaxed, the total accuracy (weighted sum of false negatives and false positives) can be higher.

The fundamental critique of this alternative setup is that designers of the algorithm may be well aware of the effect

of which equations are the binding ones in the original setting. For example, in the bank credit application, we know that the probability of an African American person getting falsely denied is higher than a white American. Then by setting τ to be 5 percent, we are either making the African American applicant further disadvantaged. Or to the other extreme, if we decide to make the false negative rejection rate for African American persons to be 5 percent lower than other racial groups, we are essentially practicing 'affirmative action' based on race, which leads to significant legal controversy.

Viable Framework: Three Pillars of Fairness

We have seen how individual and group fairness may be in conflict. To reconcile these different ideas of fairness, it is crucial to have a guiding framework. Ideally, this framework provides general principles that can be applied to any organization and to address any problem of algorithmic unfairness. In this paper, we propose the following Three Pillars as leading principles:

1. Need-based decisions
2. Transparency and Accountability
3. Narrowly Tailored Solutions and Definitions

Need-based decisions

As outlined in Srivastava, Heidari, and Krause (2019), while mathematical notions of fairness are important, it should be noted that fairness is inherently a value-based notion that may carry different meanings for different people or under different scenarios. For example, given data that may suffer from historical bias, setting different thresholds for classifications may help combat the bias in the data and help achieve group parity. However, if we are confident that all groups have a fair representation in the data, it may be best to set the same threshold across groups such that similar individuals, regardless of group membership, may be treated similarly. The decision as to which notion of fairness matters the most depends on the situation and the discretionary decision of the policymaker. There is thus not a one-size-fits-all notion of fairness that can be applied to all contexts, and we think future research in algorithmic fairness may continue to focus on how different specific scenarios may require different notions of fairness that best accommodate the needs of society in that area.

Transparency, Accountability, and Narrow Tailoring

Fairness definitions often conflict, so policy-makers must specify which notion they adopt and how tradeoffs are handled, such as between individual and group fairness. These choices should be explained clearly to affected groups, using mathematics only when necessary. Transparency enables accountability and prevents designers from masking arbitrary decisions.

Fairness also requires precise, context-specific definitions and remedies. Broad or vague definitions risk eroding trust, while narrow ones make solutions more feasible and defensible. Adjustments should be justified with historical context and communicated in plain language. Remedies must be specific to the problem rather than generic fixes. Together, transparency and narrow tailoring improve both the legitimacy and effectiveness of algorithmic systems.

Why Tailored Definitions are Needed To address any specific case of algorithmic unfairness, it is crucial to have a tailored definition of unfairness. An organization using a too broad or ill-conceived definition will struggle not only to create a technical solution to the problem but may face what political scientists refer to as "mission creep." Mission creep is when there is a gradual shift away from the initial mission. In this case, the fear is that algorithm creators may be able to commit arbitrary changes by justifying them with some notion of fairness. In this way, broad definitions of algorithmic unfairness may lead to a decrease in algorithmic accountability and transparency.

Likewise, there is no good reason to assume that every problem will require the same definition of fairness. For example, if the creator of an algorithm is a corporation, it may have different goals of fairness than if the creator were a benign social planner (Jakesch et al. 2022). It would be both unreasonable and likely undesirable to require every corporation to play the role of the ultimate arbiter of fairness. In this way, any definition of fairness should be limited by the incentives and power of the algorithmic creator. More broadly, a large assortment of different problems may all fall under the same umbrella term of unfairness and it would make little sense to use the same definition in every context.

Finally, a tailored definition of unfairness makes technical solutions more feasible. The main limitation of technical solutions to algorithmic unfairness has been requiring a model to do too much. Different notions of fairness carry with them their trade-offs. There is almost always no solution that can satisfy every definition of fair. In light of this, a context-specific, tailored definition of fairness also has the benefit of being the most workable.

Why Tailored Solutions are Needed Even with tailored definitions, solutions must be tailored as well. One-size-fits-all approaches risk legal challenge under U.S. anti-discrimination law, which requires narrow tailoring and exhaustion of race-neutral alternatives (Slaughter, Kopec, and Batal 2020). Algorithmic interventions should reflect local context and history—for example, admissions models should differ across institutions with distinct pasts.

As Ho and Xiang (2020) argue, fairness-as-awareness is both legally viable and empirically effective: aligning model design with institutional histories and data imperfections better promotes substantive fairness. Each use of non–race-blind features should be justified by social-science evidence; deeper histories of discrimination warrant stronger adjustments.

Tailoring also improves legitimacy and performance. Output legitimacy improves because models fit local goals and data; input legitimacy improves when designers explain case-specific tradeoffs. Uniform methods risk miscalibration across settings and erode public trust. Clear,

narrowly scoped adjustments—especially on sensitive attributes—sustain transparency and confidence.

Conclusion

As algorithmic fairness moves from an academic idea to real-world applications, a general framework must be established. This framework should provide universal principles that organizations can follow to ensure fairness in their own algorithms. While there will always exist trade-offs between different systems, a proper framework can act as a guiding star to navigate historical discrimination and injustice. This paper hopes to lay the foundations of such a framework with our Three Pillars Model.

Future extensions of our framework could provide more mathematically rigorous analysis of the trade-offs between different types of unfairness. Likewise, more research can be done to integrate the political science idea of "legitimacy" into our model. Other extensions of this paper may include examples of practical applications of the Three Pillars Model. Specifically, future papers may show proposed analysis of real-world problems using the Three Pillars Model in comparison to other frameworks. We hope this could offer a holistic framework to tackle the complex issues of algorithmic unfairness.

Appendix

Appendix 1. Methods for achieving group fairness

A naive way of achieving group fairness is through unawareness, which means that sensitive group attributes (such as race and gender) are to be excluded from the model. However, simply doing this can still easily result in suboptimal group fairness, since there may be features that correlate with the sensitive attributes and are still included in the model. If the model takes in those features to make predictions, and the sensitive attributes are historically tied to classification outcomes, individuals of different groups may still be treated differently.

Another problem with unawareness is that group membership may sometimes offer valuable information, which the model would lose if group attributes are excluded from the model. Dwork et al. (2012) discussed this utility by offering the example: suppose that in the culture of a protected class S, the most talented individuals would enter fields like science and engineering, while the less talented individuals enter fields like finance, and the trends are reversed in the general population. An organization hiring for talent that ignores group membership might select the subset of S most involved in economics and finance, but this is also the subset of S that is less talented. This is a poor outcome that arises from ignoring the group membership of S. Hence, ignoring sensitive attributes may not necessarily be good.

In addition to the naive method of unawareness, there are several other methods of pre-processing, which involve modifying the input data, that help achieve group fairness (Kamiran and Calders 2011). First, unawareness may be replaced by suppression, where we remove not only the group membership attribute but also attributes that highly correlate with group membership, which would address the initial concern of unawareness. Second, the data may be preprocessed to remove bias against certain groups. For instance, the data may be "massaged," where certain outcome labels are artificially changed. For example, one might turn a portion of the negative outcomes for an underprivileged group to become positive, which would cause the algorithm to more likely assign a positive label to an underprivileged group. The data may also be re-weighted, whereby individuals are assigned a weight and a larger weight could be assigned to an individual with a positive label from an underprivileged group during training. Last but not least, a class of algorithms called the disparate impact remover (DI remover) may be applied to the data, which can help achieve group fairness (Feldman et al. 2015). Mathematically, the remover removes disparate impact: given a dataset $D = (X, X_n, Y)$ where X is a protected attribute such as binary group membership, X_n are the remaining attributes, and Y is a binary outcome, the data set is considered to contain disparate impact if $\frac{P(Y=1|X=0)}{P(Y=1|X=1)} \leq \tau = 0.8$, where τ is a tunable parameter depending on the need of the algorithm designer. This directly corresponds to the mathematical definition of demographic parity, where the closer τ is to 1 the stronger the demographic parity guarantee. A DI remover hence modifies the input data labels such that the input to the algorithm satisfies group parity.

Finally, the output of an algorithm could be post-processed to better abide by group fairness standards. For example, equalized odds post-processing is an algorithm that adds a simple post-processing step at the end to solve an optimization problem of achieving equal false positive and false negative rates, which involves flipping a certain amount of output labels (Hardt et al. 2016). Another post-processing method is called reject option based classification (ROC), whereby the designer could set a threshold $\theta \in \{0.5, 1\}$, predicting a positive label if confidence in that label exceeds θ and predicting a negative label if confidence is below $1-\theta$. If confidence is between $1-\theta$ and θ, individuals from the underprivileged group are predicted a favorable outcome while individuals from the privileged group are predicted an unfavorable outcome (Kamiran and Calders 2011). Note, however, that many of these examples are for a binary classification setting where individuals belong to one of two groups: privileged and underprivileged. Many of the in- and post-processing methods may not necessarily generalize to multiclass settings, and pre-processing may be a more general approach towards achieving group fairness.

Appendix 2. Full derivations of the PAC setup for group fairness

Similar to equal opportunity, a canonical setup for group fairness is to equalize false negative rates.

In the PAC setting, say the protected category variable is V, such that $V = \{v_1, \cdots, v_n\}$, where each of these is a particular realized value, such as Gender = $\{\text{male}, \text{female}\}$

The false negative rate is defined as:

$$FN(h, v_i) := \Pr_{(x,y)\sim P}[h(x) \neq -1 \mid x = v_i, y \neq +1]. \quad (1)$$

The goal is to ensure that the output hypothesis h^* satis-

fies:

$$FN(h, v_i) = FN(h, v_j), \forall i \neq j, 1 \leqslant i, j \leqslant n. \quad (2)$$

This is viewed as the constraint to the traditional loss minimization setting, where usually the goal is minimizing the sum of false negative rates and false positive rates in total:

$$\min \sum_{i=1}^{n} (\alpha FN(h, v_i) + \beta FP(h, v_i)) W_i$$

specifically, W_i indicates the percent in the population whose protected trait turns out to be v_i, α is the loss assigned to false negative and β is the loss assigned to false positive.

Equalizing false negative rates can be viewed as the constraint to this loss minimization problem.

In specific settings, α and β are tailored depending on how bad the two types of errors are. The relationship between the associated harm α and β can be so balanced, such as

$$\frac{\alpha}{\beta} \gg 10$$

(1)

$$\frac{\alpha}{\beta} \ll 0.1$$

(2)

For example, in the medical setting, doctors usually filter out those patients who likely have cancer and direct them to further rounds of tests. Thus, a false negative is a lot worse than a false positive, because not identifying latent cancer could delay therapies and intervention. In the case of a false positive, the doctor could simply not disclose the positive judgment directly and direct the patient to the next round of checking. This corresponds to (1).

However, in another setting, such as bank credit card applications. The influence of granting credits to individuals who potentially can't pay back does less social harm than withholding it from those who deserve it. Specifically, withholding credits from the lower social classes will have a significant negative impact on their household financial situations. This corresponds to (2).

In situation (2), the optimization problem is similar to:

$$\min \sum_{i=1}^{n} \alpha FN(h, v_i) W_i$$

subject to equalized false negative rates. Previous literature, such as Jakesch et al. (2022) and Grgic-Hlaca et al. (2016), have shown the need to flexibly address the balance between false negatives and positives. Being able to justify such a loss ratio $\frac{\alpha}{\beta}$ does indeed make the public audience trust the ethicality and process fairness of algorithms.

In particular, in the setting of criminal justice, not prosecuting a suspect after a police investigation could pose nontrivial dangers to community safety, whereas excessively prosecuting all the potential suspects is a waste of judicial resources and may cause stigmatization or marginalization of communities. In such settings, striking the balance between α and β seems inherently vague and tough.

Process fairness may require no arbitrary favoring of any group under the protected category in any step of the algorithm. This is a subtle definition. In some circumstances, it might not be feasible to achieve exactly. equalized false negative rates. One potential way to reconcile it is to give some tolerance bound, which is to make the false negative rejection rates across all the groups not differ beyond a bound. One practical situation could be to train a model that asks the predicted recidivism rate not to differ by 5 percent between all racial groups. This is a way to tackle the tradeoff between 'process fairness' and 'feasibility/accuracy'.

Appendix 3. Individual fairness vs. group fairness Discussion

As we have noted in previous sections, individual fairness and group fairness may sometimes be in conflict; doing better with the former might lead to unavoidable harm to the latter and vice versa. We will now examine the literature on this conflict in more detail. As argued by Dwork et al. (2012), premised on the classification setting where an algorithm needs to map an individual to a probability distribution over outcomes, individual fairness is achieved when the statistical distance between the distributions that individuals x and y are mapped is at most the distance between the two individuals, meaning that "the distributions over outcomes observed by x and y are indistinguishable up to their distance $d(x, y)$." Mathematically, this is known as the Lipschitz condition, where for a set of individuals V and outcomes $A = \{0, 1\}$, a mapping $M : V \to \Delta(A)$ satisfies the (D, d)-Lipschitz property if for every $x, y \in V$, $D(Mx, My) \leq d(x, y)$. D could be chosen as the statistical distance between two distributions P and Q, in which case $D(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)|$, while d may denote the distance between two individuals on input attributes.

Achieving individual fairness can then be formulated as a linear program, where the expected loss of any arbitrary loss function $L : V \times A \to \mathbb{R}$ is minimized subject to the constraint that the (D, d)-Lipschitz property is satisfied, meaning that the output distribution over outcomes for any two individuals differs by at most the distance of the two individuals. However, the researchers also proved that individual fairness (where the program is subject to the constraint of the Lipschitz property) implies group fairness if and only if the Wasserstein distance (a measure of the distance between two probability distributions) between the distributions of features between the two groups is small. This means that if the two groups share a similar distribution of features, individual fairness and group fairness can be achieved simultaneously. The opposite is also true: if the two groups share very different feature distributions, the two notions of fairness cannot be achieved simultaneously.

What happens then if the statistical distance between the two groups is large? How exactly does one balance individual fairness and group fairness? Dwork et al. (2012) proposed an algorithm that attempts to balance the two by implementing *fair affirmative action*. The algorithm relaxes the Lipschitz condition such that similar individuals from two different groups, S and T need not be treated similarly; the Lipschitz condition only needs to be held between individu-

als in group S and between individuals in group T. This algorithm ensures demographic parity between S and T up to a bias ϵ, meaning that $D(P,Q) = \frac{1}{2}\sum_{a\in A}|P(a)-Q(a)| \leq \epsilon$, where P and Q are the probability distributions over outcomes for the same individual with group membership S and T respectively. At the same time, individual fairness within a group is achieved, because the Lipschitz condition is satisfied for every pair $(x, y) \in (S \times S) \cup (T \times T)$. Hence, this work shows that by sacrificing individual fairness across groups, individual within groups and group fairness may be achieved simultaneously even if the feature distribution between different groups is very different.

To determine the empirical trade-off between individual and group fairness when the Wasserstein distance is large, Zhou (2022) applied a disparate impact (DI) remover, which attempts to achieve group parity, on a real dataset *Adult* (Asuncion and Newman 2007). The dataset consists of a binary sensitive attribute (sex or race), five non-sensitive attributes (e.g., age and education), and a binary outcome label of whether the individual's income exceeds 50K a year. The researchers found that a larger Wasserstein distance between the attribute distributions of the two groups (e.g. male vs. female) leads to a larger decrease in individual fairness after applying the DI remover. Individual fairness is also more likely to decrease if the large Wasserstein distance is due to a difference in mean rather than a difference in variance (as both can give the same Wasserstein distance). This confirms the intuition that it is difficult to achieve both group and individual fairness if the two groups are very different, especially if the mean rather than the variance of their attributes is different.

References

Asuncion, A.; and Newman, D. 2007. UCI machine learning repository.

Chapman, A.; Grylls, P.; Ugwudike, P.; Gammack, D.; and Ayling, J. 2022. A Data-driven analysis of the interplay between Criminological theory and predictive policing algorithms. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 36–45.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *In Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 259–268. New York, NY, USA: Association for Computing Machinery. ISBN 9781450336642.

Fleisher, W. 2021. What's Fair about Individual Fairness? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, volume AIES '21, 480–490.

Gottfredson, M. R.; and Hirschi, T. 1987. *Positive criminology*. Sage Newbury Park, CA.

Grgic-Hlaca, N.; Zafar, M. B.; Gummadi, K. P.; and Weller, A. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, volume 1, 2. Barcelona, Spain.

Hardt, M.; Price, E.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Ho, D. E.; and Xiang, A. 2020. Affirmative algorithms: The legal grounds for fairness as awareness. *U. Chi. L. Rev. Online*, 134.

Jakesch, M.; Buçinca, Z.; Amershi, S.; and Olteanu, A. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. *arXiv preprint arXiv:2205.07722*.

Kamiran, F.; and Calders, T. 2011. Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems. *Knowledge and Information Systems*, 33(1): 1–33.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, 2564–2572. PMLR.

Kleinberg, J.; Ludwig, J.; Mullainathan, S.; and Rambachan, A. 2018. Algorithmic fairness. In *Aea papers and proceedings*, volume 108, 22–27.

Lagioia, . S., Rovatti. 2022. Algorithmic fairness through group parities? The case of COMPAS-SAPMOC. In *AI & Society*.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

Narayanan, A. 21. Fairness Definitions and Their Politics. *Youtube: Arvind Naranayan, Available online: https://www. youtube. com/watch*.

Schoeffer, J.; Kuehl, N.; and Machowski, Y. 2022. " There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. *arXiv preprint arXiv:2205.05758*.

Slaughter, R. K.; Kopec, J.; and Batal, M. 2020. Algorithms and Economic Justice: A Taxonomy of Harms and a Path Forward for the Federal Trade Commission. *Yale JL & Tech.*, 23: 1.

Speicher, . Z., Heidari. 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*, volume KDD '18, 2239–2248.

Srivastava, M.; Heidari, H.; and Krause, A. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2459–2468.

Zhou, W. 2022. *Group vs. individual algorithmic fairness*. Ph.D. thesis, University of Southampton.