Enforcing Group Fairness in Algorithmic Decision Making: Utility Maximization Under Sufficiency

Joachim Baumann
baumann@ifi.uzh.ch
University of Zurich
Zurich, Switzerland
Zurich University of Applied Sciences
Zurich, Switzerland

Anikó Hannák
hannak@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

Christoph Heitz
christoph.heitz@zhaw.ch
Zurich University of Applied Sciences
Zurich, Switzerland

ABSTRACT

Binary decision making classifiers are not fair by default. Fairness requirements are an additional element to the decision making rationale, which is typically driven by maximizing some utility function. In that sense, algorithmic fairness can be formulated as a constrained optimization problem. This paper contributes to the discussion on how to implement fairness, focusing on the fairness concepts of positive predictive value (PPV) parity, false omission rate (FOR) parity, and sufficiency (which combines the former two).

We show that group-specific threshold rules are optimal for PPV parity and FOR parity, similar to well-known results for other group fairness criteria. However, depending on the underlying population distributions and the utility function, we find that sometimes an upper-bound threshold rule for one group is optimal: utility maximization under PPV parity (or FOR parity) might thus lead to selecting the individuals with the smallest utility for one group, instead of selecting the most promising individuals. This result is counter-intuitive and in contrast to the analogous solutions for statistical parity and equality of opportunity.

We also provide a solution for the optimal decision rules satisfying the fairness constraint sufficiency. We show that more complex decision rules are required and that this leads to within-group unfairness for all but one of the groups. We illustrate our findings based on simulated and real data.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; • **Applied computing → Decision analysis**.

KEYWORDS

algorithmic fairness, prediction-based decision making, constrained utility optimization, sufficiency, machine learning, group fairness metrics, fairness trade-offs

ACM Reference Format:
Joachim Baumann, Anikó Hannák, and Christoph Heitz. 2022. Enforcing Group Fairness in Algorithmic Decision Making: Utility Maximization Under Sufficiency. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3531146.3534645

1 INTRODUCTION

Advances in machine learning (ML) have led to a rise in algorithmic decision making systems that assist or replace humans to make consequential decisions. Today, such algorithms are used in various domains, such as credit lending [18, 26], pretrial detention [1], hiring [36], and many more. It has been shown that this often violates fairness across protected groups [2]. This is especially worrying if the prediction-based decision systems systematically harm marginalized groups, and, in particular, if they are applied in domains where a decision is potentially life-changing for the affected individuals [3]. A potential way to reduce ML-based discrimination is to mitigate outcome disparities across some predefined groups [2, 9, 10, 13, 20, 30, 35]. In order to measure and eventually correct for these disparities, different mathematical notions of so-called *group fairness metrics* have been proposed [39, 45]. The group fairness metrics that have attracted the most interest are independence, separation, and sufficiency [2]. These three definitions of fairness are all "entirely reasonable and desirable" [23], however, they are mutually exclusive except for in highly constrained cases, which are unlikely to occur in practice [8, 16, 25, 42]. Hence, decision makers must choose one metric over the others.

In this paper, we focus on the fairness of prediction-based decision systems that take decisions based on the prediction of a variable Y, which is unknown at the time of decision making. Different methods have been developed to ensure the fairness of such systems, most of which fall into one of three categories: pre-processing, in-processing, or post-processing [7, 41]. One line of papers within the field of algorithmic fairness is concerned with optimal decision rules satisfying some group fairness constraint [10, 10, 20, 30, 35]. Thereby, the prediction model is treated as given, but the decision maker has the freedom to modify the decision rule for fulfilling fairness constraints, i.e., the predictions are post-processed so that the resulting decisions are fair w.r.t. a specified protected attribute. Following this approach, we formulate the goal of fairness as a constrained optimization problem where a standard approach is to assume that the decision maker's goal is to maximize some utility[1]

[1] We will define the term utility and formulate the constrained (as well as the unconstrained) optimization problem in Section 3.3.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3534645

function while also satisfying some fairness constraint [37]. Such optimal decision rules have been derived for the group fairness metrics statistical parity, conditional statistical parity, TPR parity, and FPR parity [10, 20, 30]. It has been shown that optimal decision rules that satisfy these fairness constraints are characterized by lower-bound threshold rules.[2] Surprisingly, to our knowledge, no such solution has been derived for the group fairness metrics PPV parity, FOR parity, and sufficiency. This paper closes this gap by deriving optimal decision rules for these group fairness metrics. Our main contributions and findings are:

- We show that optimal decision rules satisfying PPV parity or FOR parity take the form of group-specific (lower-bound or upper-bound) thresholds.
- We find that, surprisingly, under PPV parity or FOR parity, it can be optimal for decision makers to apply an upper-bound threshold for one group (depending on the populations and the applied utility function). In such situations, the most promising individuals are left out, leading to an extreme form of within-group unfairness.
- We provide a solution for the optimal decision rules that satisfy sufficiency as the combination of both PPV parity and FOR parity. We find that this definition of fairness requires more complex decision rules (i.e., decision rules that do not take the form of a simple lower- or upper-bound threshold) and leads to within-group unfairness for all but one of the groups.
- We highlight the trade-off between fairness across groups and within groups.

The remainder of the paper is structured as follows: Section 2 introduces the most important group fairness metrics and provides the necessary background. In Section 3, we formalize the (un)constrained optimization problem and solve it for several group fairness metrics. Section 4 demonstrates the solutions for optimal decision rules under these fairness constraints based on simulated and real data. Section 5 concludes the paper.

2 RELATED WORK

2.1 Group Fairness Metrics

Much of the technical literature on algorithmic fairness strives to create some generalized notion of fairness in terms of the impact an algorithm has on different groups [3, 5, 8, 9, 12, 14, 47]. As ML algorithms are used more and more for consequential decision making, their impact on individuals and groups may be tremendous. Numerous metrics have been suggested to quantify the group fairness of decision making algorithms [39]. Most of these group fairness criteria fall into one of three categories: independence, separation, or sufficiency [2]. Table 1 provides the mathematical definitions for those three criteria[3].

Independence – also called *statistical parity* [12] – compares decision rates across groups (i.e., the fraction of individuals who are granted a loan in each group), whereas the other two criteria compare error rates across groups [45]. Conditional statistical parity extends this definition of fairness by allowing a set of legitimate features to affect the decision [10, 21]. True positive rate (TPR) parity – also called *equal opportunity* [20] – and false positive rate (FPR) parity are relaxations of the separation criterion. Positive predictive value (PPV) parity – also called *predictive parity* [8] – and false omission rate (FOR) parity are relaxations of the sufficiency criterion – which has also been called *conditional use accuracy equality* by [5] or *overall predictive parity* by [33]. There is an essential difference between separation and sufficiency: TPR and FPR focus on a subpopulation that is defined by Y. In contrast, PPV (also called *precision*) and FOR focus on a subpopulation that is defined by D.[4] In the loan granting scenario, the TPR denotes the fraction of those individuals who are granted a loan from all those who would not default. For the PPV, on the other hand, only those individuals who are granted a loan are considered to measure the fraction of individuals who repay it.

PPV parity, FOR parity, and sufficiency are relevant notions of fairness, not only theoretically but also in practice. Most prominent is probably the case of the 2016 debate surrounding the tool COMPAS (which gives judges recidivism risk predictions that are supposed to inform them on whether or not a defendant should be released in different stages of the criminal justice system), where [1] published an article saying that the tool systematically disadvantages black defendants because of a FPR disparity. However, Northpointe (the developers of COMPAS) responded that the two metrics TPR parity and FPR parity are not appropriate for assessing recidivism risk scales and that instead PPV parity and FOR parity are appropriate criteria [11]. They conclude that their tool is not unfair because it satisfies those two metrics. In addition to recidivism prediction, PPV parity is also prevalent in predictive policing [43] (where the metric is usually called *hit rate*) and in personalized online ads (where the notion of *click through rates* [46], which is an equivalent metric, is omnipresent).

Another often discussed statistical concept in algorithmic fairness studies is calibration, which is defined as $P[Y = 1|S = s] = s$, where s denotes a real-valued score [8, 25, 42]. An extended notion of calibration that also accounts for group membership is provided by [2]. They call it calibration by group and formally define it as $P[Y = 1|S = s, A = 0] = P[Y = 1|S = s, A = 1] = s$. This notion of fairness is closely related to sufficiency, which is why some confusion regarding the differences between calibration and sufficiency (or one of its relaxations) emerged. [31] and [2] state that unconstrained learning satisfies group calibration and the fairness metric sufficiency. In contrast, [8] claims that it is possible that calibration is satisfied while PPV parity is not. [19] clarify this confusion by pointing out the difference between these two metrics: As calibration is defined for every score s (which is assumed to be a continuous value and not a binary one), whereas PPV parity is just measured for a binary outcome, the two notions of fairness cannot be used interchangeably. In particular, they show that for groups with different probability distributions, calibration does not necessarily imply sufficiency. In this work, we investigate group fairness

[2] In the fair ML literature, so-called *thresholding* is arguably the most typical decision rule for probabilistic classifiers, also because of its conceptual similarity to the way humans take decisions [7, 24]. In this paper, we refer to this type of decision rule as a *lower-bound threshold rule*.

[3] See Section 3.1 for a description of the notations used for the equations.

[4] All four metrics can be expressed by their respective complements: PPV parity is equivalent to false discovery rate parity, FOR parity is equivalent to negative predictive value parity, TPR parity is equivalent to false negative rate parity, and FPR parity corresponds to true negative rate parity.

Table 1: Group fairness metrics. The acronyms stand for true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV), and false omission rate (FOR).

Fairness criterion	Parity metric	Equation
Independence	Statistical parity	$P[D=1\|A=0]=P[D=1\|A=1]$
Separation	TPR parity	$P[D=1\|Y=1,A=0]=P[D=1\|Y=1,A=1]$
	FPR parity	$P[D=1\|Y=0,A=0]=P[D=1\|Y=0,A=1]$
Sufficiency	PPV parity	$P[Y=1\|D=1,A=0]=P[Y=1\|D=1,A=1]$
	FOR parity	$P[Y=1\|D=0,A=0]=P[Y=1\|D=0,A=1]$

metrics regarding a protected attribute that divides individuals into groups with different probability distributions.

2.2 Optimal Decisions and Fairness

Much of the extensive literature on algorithmic fairness is concerned with mitigating ML-based discrimination across protected groups. According to [37], a standard way of ensuring algorithmic fairness is to formulate it as a constrained optimization problem. Thereby, a specific kind of utility function is maximized while also satisfying a fairness constraint [10, 20, 30, 35]. This approach allows a utility-maximizing decision maker to derive optimal fair decision rules. Absent any fairness constraint, applying a uniform threshold to all groups is optimal [9]. However, this does not automatically lead to fair decisions w.r.t. specific groups [2]. Due to the mathematical incompatibility of most group fairness metrics [8, 16, 25, 42], the constrained optimization problem must be solved separately for any chosen definition of fairness. This has been done for some group fairness metrics but not for others: [20] and [10] have shown that optimal decision rules that satisfy (conditional) statistical parity, TPR parity, and FPR parity take the form of group-specific lower-bound thresholds. Several other scholars have investigated thresholding solutions, such as [15, 30, 35, 44]. However, to our knowledge, a solution for the optimization problem satisfying PPV parity, FOR parity, or sufficiency does not yet exist. This paper closes this research gap by providing a solution for deriving optimal decision rules that satisfy one of these three group fairness metrics.

In the computer science and in philosophical literature, sufficiency (or one of its relaxations, PPV parity and FOR parity) is often mentioned as one of the main fairness metrics [2, 4, 5, 7, 8, 23, 27–29, 32, 39, 41, 45]. Several algorithmic fairness papers have studied sufficiency or one of its relaxations. [22] use an economic approach to argue that PPV parity is insufficient for fairness as it does not question existing differences between or within groups. [6] explore the possibilities of satisfying several fairness constraints at once, namely, parity of PPV, FOR, TPR, and FPR, but they do not provide a solution for PPV parity or FOR parity alone. However, none of these authors derive optimal decision rules that satisfy (one of) these fairness constraints. Such a solution is crucial to know what decision rational decision makers take if any of these group fairness metrics are enforced.

3 OPTIMAL DECISIONS UNDER FAIRNESS CONSTRAINTS

This section provides a theoretical solution to maximizing the decision maker's utility while satisfying a group fairness definition (PPV parity, FOR parity, or sufficiency). In the following, we first state the problem and introduce general notations before introducing an additional notion of fairness called *within-group fairness*, which will prove to be helpful for the interpretation of the theoretical results. Then, we formulate the optimization problem to be solved (with and without fairness constraints) in Subsection 3.3, before actually solving it for three specific group fairness definitions (see Subsections 3.4 and 3.5 and Appendix C).

3.1 Problem Statement and Notations

Let us first introduce the specific context of our work, along with the main assumptions and some notations. We assume a decision maker has to make a binary decision D for each individual i, based on a feature vector $x_i \in \mathbb{R}^m$, which includes a protected attribute $a_i \in A$, denoting the group membership (sometimes also called *sensitive attribute*). Let $n_{A=a}$ be the number of individuals that are part of a group a. Following related work, we restrict our analysis to a binary protected attribute A. However, our analysis generalizes to all cases with a discrete protected attribute with more than two values. An example may be the decision of a bank to grant a loan, based on x_i.[5] We assume that the decisive feature for the decision is a binary target variable Y. For a perfect predictor, every individual that belongs to the positive class ($Y = 1$) must receive decision $D = 1$, and vice versa [37, 38]. However, Y is unknown at the time of decision making and is replaced by the probability $p_i = P[Y = 1]$, which is given as a function of x_i, provided by a probabilistic prediction algorithm. Generalizing the idea of a perfect predictor to probabilities means that individuals with a higher p_i should be assigned $D = 1$ and individuals with a lower probability of belonging to the positive class should receive the decision $D = 0$. The decision rule is thus a function d that maps p_i (and, possibly, a_i) to a binary decision.[6] Similar to [6, 10], for our analysis, we assume furthermore that each group's probability distribution has strictly positive density.

In this paper, we formulate algorithmic fairness as a constrained optimization problem. The goal of a rational decision maker is to

[5] In the loan granting scenario, x_i might include an applicant's bill-paying history, unpaid debt, or past foreclosures.

[6] Notice that changing this decision rule represents a form of post-processing [34]. There is no need to know the specific features used to train an algorithm because the learned model is treated as a black box.

maximize the expected utility while also satisfying some definition of group fairness. In this section, we solve this constrained optimization problem for the group fairness definitions PPV parity, FOR parity, and sufficiency.

3.2 Within-Group Fairness

Before we derive the solution for a utility-maximizing decision maker that must satisfy some group fairness metric, let us formally define another notion of fairness – which will be helpful for the interpretation of the theoretical results.

Definition 1. (Within-group fairness). We say that a decision rule $d(p, a)$ satisfies *within-group fairness* with respect to protected attribute A if $\forall i \in S_{a|D=1} \forall j \in S_{a|D=0} (p_i > p_j)$, where S_a is the set of all individuals of group $a \in A$.

Decision rules satisfying within-group fairness ensure that, within each group, a larger probability p always leads to a higher chance of $D = 1$. More specifically, no individual that is assigned $D = 1$ has a lower probability than any of the individuals that are given $D = 0$. In contrast, within-group unfairness results if there is at least one pair of individuals (i, j) where p_i is larger then p_j, and still $D_i = 0$ and $D_j = 1$. As more or less such cases can exist, there are different degrees to which within-group fairness can be violated. We say that a decision rule $d(p, a)$ leads to an extreme form of within-group unfairness if $\forall i \in S_{a|D=1} \forall j \in S_{a|D=0} (p_i < p_j)$. This is equivalent to applying an upper-bound threshold.

Within-group fairness requires that, within a group, individuals with a higher probability of belonging to the positive class ($p[Y = 1]$) should have a higher chance of being assigned $D = 1$ than individuals with a lower probability of belonging to the positive class. For the loan example, it would be viewed as unfair if a loan is granted to one person but denied to another person with a higher probability of paying back the loan. In many applications, such a perspective can be morally justified. Similarly, in the context of COMPAS, it is morally just to detain a defendant ($D = 1$) with a very high risk of committing a violent crime if released ($Y = 1$). As we will see in more detail below, optimal decision rules satisfying PPV parity, FOR parity, or sufficiency do not always satisfy this notion of fairness.

3.3 Optimal Decision Rules With or Without Fairness Constraints

For our theoretical analysis, we assume that a rational decision maker relies on a prediction model to cope with the uncertainty of the decision-relevant variable Y. We assume that if Y was known, the decision would be given. More specifically, we assume that the decision maker's choice would be $D = 1$ in the case of $Y = 1$ and vice versa [38]. However, in most real-world scenarios, a perfect predictor does not exist, which introduces uncertainty regarding the outcome of a decision. There are four possible outcomes, all of which can be weighted according to the decision maker's desirability, representing a standard approach in the fair ML literature [37]. This leads to the following expected individual utility[7]:

$$u_i = \begin{cases} u_{11} p_i + u_{12}(1 - p_i), & \text{for } D = 1 \\ u_{21} p_i + u_{22}(1 - p_i), & \text{for } D = 0. \end{cases} \tag{1}$$

Defining $\tilde{u}_i$ as the expected *relative utility gain* when switching the decision from $D = 0$ to $D = 1$ gives $\tilde{u}_i = 0$ for $D = 0$, and $\tilde{u}_i = \alpha p_i + \beta(1-p_i)$ for $D = 1$, with the two parameters $\alpha = u_{11} - u_{21}$ and $\beta = u_{12} - u_{22}$. It can be shown easily that maximizing u_i is equivalent to maximizing $\tilde{u}_i$. Moreover, the above made assumption that $Y = 1$ implies $D = 1$ requires that $\alpha > \beta$.

We assume that the decision maker takes not only one decision d, but many decisions d_i, over a population of individuals (e.g., when making loan decisions for many applicants). In this case, the goal of a rational decision maker is to maximize the total expected utility $\tilde{U}$, which leads to the following optimization problem:

$$\underset{d}{\arg\max} \quad \tilde{U} = \sum_{i \in S} \tilde{u}_i d_i = \sum_{i \in S} \left(p_i(\alpha - \beta) + \beta\right) d_i, \tag{2}$$

where S is the set of all individuals and d_i is a binary multiplier representing the decision that is made for an individual i. The optimum unconstrained decision rule d^* is thus:

$$d_i^* = \begin{cases} 1, & \text{for } p_i > \frac{-\beta}{\alpha - \beta} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

and takes the form of a single lower-bound threshold. In the following, we interpret the decision problem as a *selection problem*, denoting individuals with $D = 1$ as "being selected."

The unconstrained solution does not ensure fairness w.r.t. the protected attribute at all and, in fact, is likely to produce unfairness (as measured with different group fairness metrics, see Section 2.1). Decision makers who want to maximize their utility while taking fair decisions must solve the following constrained optimization problem:

$$\underset{d}{\arg\max} \quad \tilde{U} \quad \text{subject to} \quad \textit{some fairness constraint.} \tag{4}$$

As we outlined in Section 2.2, this constrained optimization problem has been solved for some group fairness metrics[8] (statistical parity, conditional statistical parity, TPR parity, and FPR parity) but not for others, such as PPV parity, FOR parity, or sufficiency. In the remainder of this chapter, we solve the constrained optimization problem stated in Equation 4 (using the three mentioned group fairness metrics as fairness constraints) for two different cases:

case I) the number of individuals to be selected ($n_{D=1}$) is predefined,

case II) the number of individuals to be selected ($n_{D=1}$) is not predefined.

[7] For example, u_{21} denotes the utility of making a decision $D = 0$ and having outcome $Y = 1$ occur, a so-called false negative (see Appendix A). The definition of these utilities is context-specific. In many cases, it would be straightforward for the decision maker to estimate them. For example, a bank can easily calculate its utility in terms of monetary gains or losses for a successful loan (as opposed to a default) based on interest rates.

[8] The authors of [20] use a function they call *immediate utility* and [10] rely on *loss minimization*. Both approaches can easily be formulated in terms of what we call decision maker utility, which is why the solutions of [20] and [10] also hold in this setting.

3.4 Optimal Decision Rules under PPV Parity

We now present the optimal solution for the optimization problem stated in Equation 4 constrained by the group fairness metric positive predictive value (PPV) parity for both cases I and II. The PPV is defined as the average probability of individuals with $D = 1$ to have $Y = 1$, which can be written as $\frac{1}{n_{D=1}} \sum_{i \in S} p_i d_i$. The fairness definition PPV parity requires this value to be the same across groups. Thus, the constrained optimization problem has the form:

$$\begin{aligned} \underset{d}{\arg\max} \quad & \tilde{U} = \sum_{i \in S} (p_i(\alpha - \beta) + \beta)\, d_i \\ \text{subject to} \quad & \frac{1}{n_{A=0|D=1}} \sum_{j \in S_0} p_j d_j = \frac{1}{n_{A=1|D=1}} \sum_{j \in S_1} p_j d_j = PPV, \\ & \text{for } PPV \in [0,1], \end{aligned} \tag{5}$$

where S_a is the set of all individuals of group a and $n_{A=a|D=1}$ denotes the number of individuals in group a with $D = 1$. Each decision rule results in a specific selection of individuals, which also yields a specific selection for each group S_a. Since the PPV can only be defined if at least one individual is selected, we assume $n_{A=a|D=1} \geq 1$ for each group.

We derive the solution to this optimization problem in two consecutive steps.

- First, we derive the optimal decision rules d^* for a simplified constraint: We assume that the PPV of both groups must be equal to a predefined value $PPV_t \in [0,1]$.
- Second, we solve the full optimization problem by maximizing the decision maker's utility over all possible values of PPV_t.

We now derive the solution for the first step, thus specifying a value $PPV_t \in [0,1]$ for the constraint. We do this under the assumption of a positive probability density of individuals over the full range $[0,1]$ for both groups, and in the limit case of very large populations ($n_{A=a} \longrightarrow \infty$). Thus, for each PPV_t, there exist individuals in each group with $p = PPV_t$.[9] The most straightforward selection fulfilling the fairness constraint thus consists of selecting one of these individuals in each group. Obviously, other selections exist, for example selecting more than one individual with $p = PPV_t$, or selecting individuals in an interval $[PPV_t - \epsilon, PPV_t + \epsilon]$ such that the average p of the selection equals PPV_t. However, many other selection rules are conceivable, with different numbers of selected individuals.

For a predefined number of selected individuals $n_{D=1}$ (i.e., case I), the following Lemma holds:

Lemma 2. *For a given value of PPV_t and a predefined number of selected individuals $n_{D=1}$, any selection fulfilling the fairness constraint of Equation 5 leads to a total utility $\tilde{U}$ of:*

$$\tilde{U} = (\alpha PPV_t + \beta(1 - PPV_t)) n_{D=1}. \tag{6}$$

Lemma 2 (proof in Appendix B) shows that the fairness constraint already defines the total utility, if $n_{D=1}$ is given. In other words: any decision rule $d(p, a)$ with $n_{D=1}$ that satisfies the constraint stated in Equation 5 for a given PPV_t is optimal. We thus end up with two independent selection problems, one for each group, which consists of finding a selection of individuals characterized by the fact that their average probability equals PPV_t. For each group a, selections with different numbers $n_{A=a|D=1}$ are possible. As long as the predefined $n_{D=1}$ is met, the group membership of the selected individuals does not matter for the resulting total utility. Hence, there may be several solutions to the optimization problem that differ regarding the number of individuals selected per group (i.e., representing different combinations of $(n_{A=0|D=1}, n_{A=1|D=1})$), with $n_{A=0|D=1} + n_{A=1|D=1} = n_{D=1}$. Note that most of these solutions violate the group fairness metric statistical parity while still meeting the fairness criterion of PPV parity.

We now analyze case II, where $n_{D=1}$ is not predefined. Lemma 2 also leads to another important result: For values PPV_t for which $\alpha PPV_t + \beta(1 - PPV_t) < 0$, a decision maker who wants to maximize the total utility should minimize $n_{D=1}$, thus selecting only one individual from each group, yielding a total utility of $\tilde{U} = 2(\alpha PPV_t + \beta(1 - PPV_t))$ for a binary protected attribute. In the following, we thus assume that $\alpha PPV_t + \beta(1 - PPV_t) > 0$. Again we assume that the size of both groups is large but finite. Lemma 2 shows that, under these assumptions, the decision maker's goal is to find the selection that satisfies the constraint $PPV = PPV_t$ with the maximum $n_{D=1}$. Theorem 3 specifies the solution of this optimization problem (which can be solved independently for each group):

Theorem 3. *For any given PPV_t, the optimal fair decision rules d^* (i.e., decision rules that maximize $\tilde{U}$ while satisfying $PPV = PPV_t$) take the following form:*

$$d_i^* = \begin{cases} \left.\begin{matrix} 1, & \mathit{for}\, p_i \geq \tau_a \\ 0, & \mathit{otherwise} \end{matrix}\right\} \mathit{for}\, PPV_t > BR_{A=a} \\ \left.\begin{matrix} 1, & \mathit{for}\, p_i \leq \tau_a \\ 0, & \mathit{otherwise} \end{matrix}\right\} \mathit{for}\, PPV_t < BR_{A=a}, \end{cases} \tag{7}$$

where τ_a denote different group-specific constants and $BR_{A=a}$ denotes group a's base rate (BR) which is defined as the ratio of individuals belonging to the positive class ($Y = 1$) in a group: $BR_{A=a} = P[Y = 1|A = a] = \frac{1}{n_{A=a}} \sum_{i \in S_a} p_i$.

Proof. We begin with the case $PPV_t < BR_{A=a}$. We define a group-specific function $g_1(n_{A=a|D=1})$, defined as the minimum value of PPV among all decision rules $\vec{d}$ with a specified $n_{A=a|D=1}$, i.e., $g_1(n_{A=a|D=1}) = \min_{\vec{d}} \frac{1}{n_{A=a|D=1}} \sum p_i d_i$. Obviously, $g_1(n_{A=a|D=1})$ is given by selecting the $n_{A=a|D=1}$ individuals with the smallest values of p. The function $g_1(n)$ for $n = 1, ..., n_{A=a}$ is monotonously increasing, with $g_1(1) = 0$ [10] and $g_1(n_{A=a}) = BR_{A=a}$. It is now easy to see that solving the equation $g_1(n) = PPV_t$ w.r.t. n yields the maximum possible value n that meets the PPV condition: Assume that there was a value $m > n$ for which a decision rule exists such

[9] This technical assumption simplifies the notation. For finite group sizes, the equality constraint in Equation 5 may not be met precisely for many values of PPV_t, and the fairness constraint might only be fulfilled approximately. Thus, the equality requirement of the FC has to be softened into approximate equality. However, the proofs are also valid for an approximate version of equality.

[10] Recall that we consider a limit of very large populations, so the individual with the lowest p_i is characterized by $p_i = 0$. For $n = 1$, the minimum PPV value is achieved by selecting just this individual.

that $PPV = PPV_t$. As g_1 is monotonically increasing, this implies $m \leq n$, which is a contradiction. Thus, for the case $PPV_t < BR_{A=a}$, the maximum achievable $n_{A=a|D=1}$ with $\frac{1}{n_{A=a|D=1}} \sum p_i d_i = PPV_t$ in the space of all possible decision rules is achieved by selecting all individuals with $p_i \leq \tau_a$. The corresponding upper-bound threshold τ_a is given by the unique solution of $g_1(n) = PPV_t$.

For $PPV_t > BR_{A=a}$, an analogous argumentation holds by introducing a function $g_2(n_{A=a|D=1}) = \max_{\vec{d}} \frac{1}{n_{A=a|D=1}} \sum p_i d_i$. This is a monotonically decreasing function with $g_2(1) = 1$ and $g_2(n_{A=a}) = BR_{A=a}$. The unique solution of $g_2(n) = PPV_t$ yields the lower-bound threshold τ_a that meets the PPV condition. □

Finally, we perform the second step of the solution: from a discretization of all PPV, for which a solution exists, we choose the one that (in combination with the corresponding $n_{D=1}$) maximizes the total utility. Thereby, every $n_{D=1}$ is composed of the optimal selections $n_{A=a|D=1}$ for all groups $a \in A$, as elaborated in the first step of the solution.

We provide an analogous solution for the optimal decision rules satisfying FOR parity in Appendix C.

3.5 Optimal Decision Rules under Sufficiency

Based on the solutions presented above, we now describe the decision rules that maximize the decision maker's utility while satisfying sufficiency (requiring PPV parity and FOR parity). This gives the constrained optimization problem:

$$\begin{aligned}
\underset{d}{\arg\max} \quad & \tilde{U} = \sum_{i \in S} \tilde{u}_i \\
\text{subject to} \quad & \frac{1}{n_{A=0|D=1}} \sum_{j \in S_0} p_j d_j = \frac{1}{n_{A=1|D=1}} \sum_{j \in S_1} p_j d_j \\
& \frac{1}{n_{A=0|D=0}} \sum_{j \in S_0} p_j (1 - d_j) = \frac{1}{n_{A=1|D=0}} \sum_{j \in S_1} p_j (1 - d_j),
\end{aligned} \tag{8}$$

where the first constraint represents PPV parity and the second constraint ensures FOR parity. Similar to our PPV parity solution, we also proceed in two steps for optimal decision rules satisfying sufficiency. First, we derive the optimal decision rules for a given value of $PPV = PPV_t$. Second, we solve the optimization problem by choosing a PPV-FOR combination that maximizes the decision maker's utility.

We start with an optimal decision rule satisfying PPV parity (see Equation 7) and then add the second constraint (requiring FOR parity). Recall that a decision rule splits this group into those selected ($D = 1$) and those not selected ($D = 0$). Thus, we can write:

$$\sum_{i \in S_a} p_i = \left(\sum_{i \in S_a} p_i (1 - d_i) \right) + \left(\sum_{i \in S_a} p_i d_i \right). \tag{9}$$

As we specified $PPV_{A=a} = PPV_t$, PPV parity is satisfied. Thus, this gives:

$$n_{A=a} BR_{A=a} = n_{A=a|D=0} FOR_{A=a} + n_{A=a|D=1} PPV_t. \tag{10}$$

With $n_{A=a|D=0} = n_{A=a} - n_{A=a|D=1}$ and some reformulation, we get:

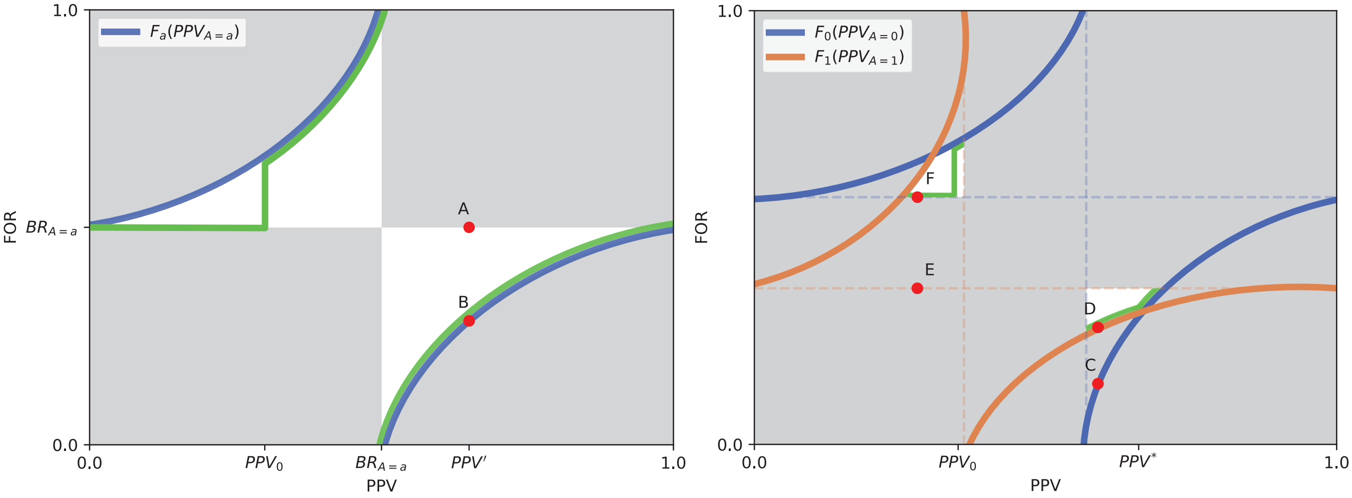
$$FOR_{A=a} = \frac{n_{A=a} BR_{A=a} - n_{A=a|D=1} PPV_t}{n_{A=a} - n_{A=a|D=1}}. \tag{11}$$

Thus, for a given PPV_t, the corresponding group-specific $FOR_{A=a}$ just depends on $n_{A=a|D=1}$, because $n_{A=a}$ and $BR_{A=a}$ are given by the group a's population. For groups with different probability distributions, $FOR_{A=0}$ and $FOR_{A=1}$ are usually different if just PPV parity is enforced. Hence, to satisfy sufficiency, at least one of the two groups must deviate from their optimal solution (under PPV parity) to ensure that the FORs of the two groups are equal. Most importantly, this deviation must not change the group's PPVs so that the PPV parity constraint still holds (with $PPV = PPV_t$). Let the *solution space* consist of all combinations of PPV and FOR that can be achieved by all groups, based on the groups' probability distributions. We now show how this solution space can be constructed for one or for more groups.

As shown in Equation 10, the PPV and the FOR always lie on different sides of the BR, because $n_{A=a} = n_{A=a|D=0} + n_{A=a|D=1}$ and $BR_{A=a}, FOR_{A=a}, PPV_t \in [0, 1]$. Therefore, if $PPV > BR_{A=a}$, the group's $FOR_{A=a}$ must take a value *below* $BR_{A=a}$ and vice versa. Let $F_a(PPV_{A=a})$ be a group-specific function defined as a group a's $FOR_{A=a}$ that results from maximizing $n_{A=a|D=1}$ for a specific value of PPV. As shown in the proof of Theorem 3, varying the number of selected individuals without changing the group's PPV lets us specify the range of values a group's FOR can take. In this way, we can derive the range of values the FOR can take for any PPV, which will then allow us to construct the solution space.

In Figure 1a, the solution space is represented as a white area and the function $F_a(PPV_{A=a})$ is illustrated with a blue line. For example, for a given PPV', point A is achieved by selecting just one individuals with a probability $p_i = PPV$, point B is achieved by maximizing $n_{A=a|D=1}$. The green line in Figure 1a visualizes the combinations resulting from applying optimal decision rules for a specific PPV: As we stated in Theorem 3, it is optimal to apply a lower-bound threshold and if $PPV \in [BR_{A=a}, 1]$ and an upper-bound threshold is optimal if $PPV \in [PPV_0, BR_{A=a}]$, where PPV_0 denotes the PPV for which $\alpha PPV + \beta(1 - PPV) = 0$. If $PPV < PPV_0$, it is optimal to minimize the number of selected individuals (see Section 3.4). The intuition to construct a solution that satisfies sufficiency is the following: under PPV parity, for a given PPV', it is optimal to apply a decision rule leading to a PPV-FOR combination lying at point B. However, the FOR that this decision rule yields might not lie within the other group's solution space, making a deviation in point A necessary.

Let us now generalize this to two (or more) groups. To construct the joint solution space of several groups, the individual solution spaces can be laid on top of each other. Figure 1b illustrates this for two groups, 0 (blue) and 1 (orange). The two white areas include all PPV-FOR combinations that are feasible for both groups. Inside this resulting smaller solution space, the optimal FOR for each possible PPV can be found (as visualized with the green line in Figure 1b), which satisfies sufficiency. Enforcing PPV parity does not result in a solution that also satisfies FOR parity simply by chance, apart from one exceptional case: That is, only if $PPV = PPV^*$, where PPV^* denotes the specific PPV for which the two groups'



(a) Solutions space: possible (white area) and optimal (green line) PPV-FOR combinations for one group

(b) Overlying solutions spaces: possible (white area) and optimal (green line) PPV-FOR combinations for two groups

Figure 1: PPV-FOR plot: utility-maximizing PPV-FOR combinations for specific values of PPV

lines representing their optimal decision satisfying PPV parity (i.e., $F_0(PPV_{A=0})$ and $F_1(PPV_{A=1})$) intersect, the decision rule satisfying PPV parity also satisfies sufficiency. If $PPV \neq PPV^*$, one of the two groups must deviate from their optimal PPV-FOR combination in order to match the other group's FOR and to ensure that not only PPV parity but also FOR parity is satisfied. Visually, this deviation (representing a change in the FOR for a remaining value of PPV) can be perceived as a vertical move away from the optimal PPV-FOR combination (satisfying PPV parity) towards the edge of the solution space (see $C \longrightarrow D$ or $E \longrightarrow F$ in Figure 1b).

The construction of the solution space (as visualized in Figure 1b) directly generalizes to cases with any number of groups, i.e., cases in which the sensitive attribute is a set consisting of more than two different values. Theorem 4 shows that this makes a full satisfaction of within-group fairness impossible.

Theorem 4. *Optimal decision rules d^* that satisfy sufficiency lead to within-group unfairness in all but one of the groups if a solution exists.*

Proof. Let us first consider a binary protected attribute A. The intersection of the group-specific solution spaces defines all PPV-FOR combinations for which a solution exists. If the deviating group's $FOR_{A=a} > BR_{A=a} > PPV$, their $FOR_{A=a}$ must match the other group's BR ($E \longrightarrow F$ in Figure 1b). Otherwise, if $FOR_{A=a} < BR_{A=a} < PPV$, their FOR must match the other group's $F_a(PPV_{A=a})$ ($C \longrightarrow D$ in Figure 1b). This deviation is necessary to satisfy sufficiency and can be achieved by adjusting $n_{A=a|D=1}$. This represents an equivalent problem as maximizing the utility under PPV parity with case I – as we discussed it in Section 3.4. Hence, the deviating group's optimal decision rule can take many forms – e.g., one could apply a stochastic decision rule that flips a coin to set $D = 1$ with probability q for all individuals with $p > \tau_a$, where τ_a is a group-specific constant. However, instead of a simple lower- or upper-bound threshold but, are more complex decision rule is required in order to ensure that the correct number of individuals are selected. Thus, this always leads to unfairness *within* this group to achieve sufficiency *between* the groups: $\exists i, j \in S_a(p_i > p_j \land d_i = 0 \land d_j = 1)$. □

Notice that any PPV-FOR combination lying inside the solution space but not at the edge is Pareto dominated because there is another point with the same PPV that results in a higher utility.[11] The green line in Figure 1b represents the optimal PPV-FOR combinations for specific values of PPV. Any number of solution spaces can be laid on top of each other, which is why this finding extends directly to non-binary sensitive attributes. Though, the more groups are considered (assuming that the groups' F_a functions and their BRs differ), the smaller the solution space becomes. And, the smaller the solution space, the bigger the required deviation, which produces more within-group unfairness. An area of size 0 is possible and would imply that sufficiency cannot be satisfied.

Finally, as we can compute the utility resulting from applying an optimal decision rule satisfying sufficiency for any value of PPV, we can solve the constrained maximization problem stated in Equation 8 by choosing the optimal PPV-FOR combination (i.e., the optimal point lying on the green line in Figure 1b).

4 ILLUSTRATIVE EXAMPLES

We now illustrate the solutions (that we derived theoretically in the previous section) to showcase the decisions that result from a utility-maximizing decision maker who wants to satisfy different fairness constraints (PPV parity, FOR parity, sufficiency). First, we demonstrate the form that these optimal decision rules take for

[11] If $FOR_{A=a} < BR_{A=a} < PPV$, this point lies on one of the groups' $F_a(PPV_{A=a})$, else, this point is situated on one of the groups' BR.

different synthetic populations. Second, we apply the solutions to real data.[12]

To present our results, we use a simple tuple notation (τ_1, τ_2) (where τ_1 denotes the lower- and τ_2 the upper-bound), meaning that any individuals with a probability $p \in [\tau_1, \tau_2]$ is assigned the decision $D = 1$ and $D = 0$ otherwise.

4.1 Synthetic Data Example

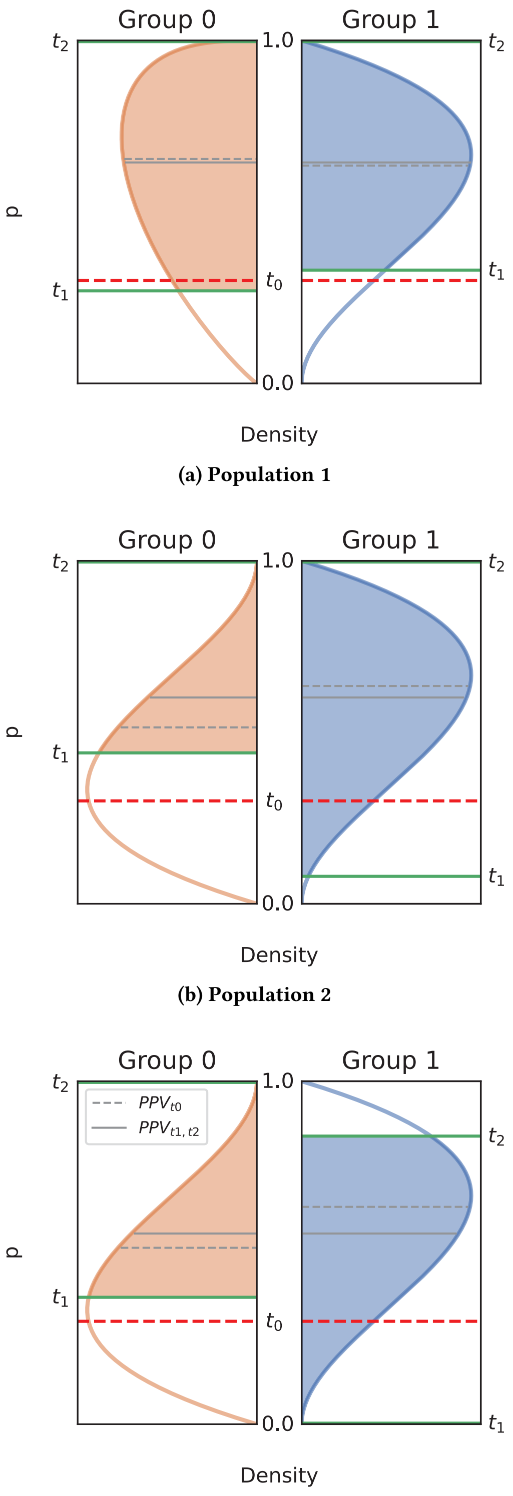
For three different populations, all of which are composed of two groups (1 and 2) of individuals with probabilities drawn from a Beta distribution, we investigate the form the optimal fair decision rules take. Table 2 list the detailed parameters for all populations. Notice that the groups are equal in size in populations 1 and 2, but in population 3, group 1 is much smaller (just 10% the size of group 0). In all populations, group 0 is disadvantaged, meaning that it has a lower base rate (BR) than group 1: $BR_{A=0} < BR_{A=1}$ (just slightly lower in population 1, substantially lower in populations 2 and 3). We present the solutions for decision rules that satisfy a fairness constraint (PPV parity, FOR parity, or sufficiency) while optimizing the decision maker's utility[13], which is defined as follows for all three populations:

$$U = \sum_{i \in S} u_i, \quad \text{for} \quad u_i = \begin{cases} 7p_i - 3(1 - p_i), & \text{for } D = 1 \\ 0, & \text{for } D = 0 \end{cases} \tag{12}$$

Hence, an individual's expected utility depends on the estimated repayment probability p. Absent any fairness constraint, it is optimal for the bank to grant a loan to all individuals whose $p > t_0 = 0.3$ (as indicated with the red dashed line in the Figures 2a-2c).

Figure 2 visualizes the probability densities of p along with the optimal decision rules under PPV parity, which are different for each of the three populations. Applying a single threshold t_0 results in unequal $PPVs$ for all of the three cases (see PPV_{t0} in Figure 2). The solid green lines indicate the thresholds t_1 and t_2 that correspond to the optimal decision rule while satisfying PPV parity. With this fairness constraint, $PPVs$ are equalized (see $PPV_{t1,t2}$ in Figure 2). But, the optimal decision rule used to achieve this depends on the population:

- Population 1: Compared to the optimal solution without fairness (t_0), group 0's threshold is decreased while group 1's threshold is increased ($t_1^{group\ 0} < t_0 < t_1^{group\ 1}$) in order to equalize the two groups' $PPVs$.[14]
- Population 2: Unlike in population 1, in population 2, group 0's PPV_{t0} is lower than the one of group 1. This means that the disadvantaged group 0 is held to a higher standard ($t_1^{group\ 0} > t_0 > t_1^{group\ 1}$) to satisfy PPV parity while maximizing the utility. This result is likely to occur in practice because, with the single threshold (t_0) rule that is used without any fairness constraint, the disadvantaged group's PPV_{t0} is lower for groups with similar distributions.



(a) Population 1

(b) Population 2

(c) Population 3

Figure 2: Utility maximization under PPV parity (synthetic)

[12] Data and code to reproduce our results are available at https://github.com/joebaumann/fair-prediction-based-decision-making.

[13] This hypothetical utility function represents a situation where a successful loan makes 7, but a default costs the bank 3.

[14] This result is not surprising as it is conceptually equivalent to solutions for other group fairness metrics.

Table 2: Parameters and solutions of the synthetic data example. The acronyms stand for group size (n), group distribution (P), base rate BR (which results from n and P), optimal threshold (t_0) and resulting PPV (PPV_{t0}) for unconstrained utility maximization, optimal thresholds (t_1, t_2) and resulting PPV ($PPV_{t0,t1}$) for utility maximization under PPV parity.

			Population 1		**Population 2**		**Population 3**	
			Group 0	Group 1	Group 0	Group 1	Group 0	Group 1
parameters		n	20,000					2,000
		P	Beta(1.9, 1.35)	Beta(3, 2)	Beta(2, 3)	Beta(3, 2)	Beta(2, 3)	Beta(3, 2)
		BR	0.58	0.60	0.39	0.60	0.39	0.60
solutions	unconstr.	t_0	0.30					
		PPV_{t_0}	0.65	0.63	0.51	0.63	0.51	0.63
	PPV parity	(t_1, t_2)	(0.27, 1)	(0.33, 1)	(0.44, 1)	(0.08, 1)	(0.37, 1)	(0, 0.84)
		PPV_{t_1,t_2}	0.64		0.60		0.56	

- Population 3: Due to the mere difference in the group sizes (all else equal to population 2), it is much more "costly" to change group 0's threshold (relatively to group 1). Thus, in this situation, it is optimal to deviate less from group 0's unconstrained optimum. This results in an optimal $PPV = 0.56$, which is lower than $BR_{A=1}$. For this reason, it is optimal to apply an upper-bound threshold for group 1 (set $t_1^{group\ 1} = 0$ and $t_2^{group\ 1} < 1$), i.e., deliberately disregarding those individuals with the highest probability of belonging to the positive class $Y = 1$. This leads to an extreme form of within-group unfairness. It means that a utility-maximizing decision maker would "sacrifice" the best individuals (with a probability between $t_2^{group\ 1}$ and 1) of the smaller group 1 in favor of "keeping" individuals with a probability slightly above $t_2^{group\ 0}$ in the bigger group 0. In the loan granting scenario, this would imply *not* granting a loan to those individuals of group 1 that are most likely to repay. At the same time, group 1's individuals with the lowest repayment probability (i.e., those with a high probability of default) are granted a loan.

This example shows clearly that the optimal decision rules depend on the groups' probability distributions. In some cases, this can lead to counter-intuitive solutions: it is possible that the disadvantaged group is held to a higher standard or that the most promising individuals of the advantaged group are omitted.

We present additional results (i.e., optimal decision rules under FOR parity and under sufficiency) for the synthetic data example in Appendix D.

4.2 Real-World Example: COMPAS

We now illustrate our results for the recidivism prediction case, using the ProPublica recidivism dataset[15], which includes data from the COMPAS tool collected by [1]. We trained a logistic regression (based on the implementation by scikit-learn [40]) to predict probabilistic recidivism risk scores (achieving an overall accuracy of 0.69).

A decision maker has to transfer a risk score into a decision. This involves weighing the severity of FP and FN in the utility function. We present the utility-maximizing solutions for three possible settings, each one specified by different utility weights FP and FN, while TP=TN=1 is kept constant. These different utility functions are paired with different fairness requirements (no fairness constraint, PPV parity, and FOR parity) w.r.t. the protected attribute *race*, which can take two values, *Caucasian* (c) or *non-Caucasian* (nc). The class $Y = 1$ denotes a recidivist, and each individual must either be detained ($D = 1$) or released ($D = 0$). Figure 3 shows the score distributions of the two groups. The base rate of non-Caucasians (0.49) is higher than the one of Caucasians (0.4), indicating that non-Caucasians more likely to be predicted as being of high risk to recidivate, on average. The specified utility weights and the resulting optimal decisions for the different fairness requirements are presented in Table 3. The (un)constrained optimal decision rules differ largely across the three cases:

- *Case 1* represents a situation where a decision maker is indifferent about what is worse: incorrectly classifying an innocent person as guilty or releasing a defendant who goes on to recidivate. Thus, equal weights for FP and FN are chosen. For such a case, a lower-bound threshold of $t_{u1} = 0.5$ is optimal from the decision maker's perspective. However, the two fairness metrics are not just satisfied by chance, because this threshold leads to different FORs and PPVs for the the two groups ($PPV_{A=c} < PPV_{A=nc}$ and $FOR_{A=c} < FOR_{A=nc}$).
- *Case 2* showcases decision rules representing a shift towards protecting the innocent, therefore, using a much lower weight (-10) for FP. For the unconstrained setting, this results in fewer detained individuals overall, with an optimal lower-bound threshold of $t_{u2} = 0.85$. As the two groups' distributions are similar above this threshold, their PPVs are almost the same, which is why just a slight adjustment of the group-specific thresholds is needed to satisfy PPV parity. In contrast, very different group-specific thresholds are optimal to satisfy FOR parity. Due to the lower BR of the non-Caucasian group (see the right-skewed distribution in Figure 3), it is optimal to release all Caucasians with a risk score below 0.98. This makes sure that released individuals are equally likely to recidivate across groups.
- *Case 3* resembles a decision maker who cares more about punishing guilty than protecting innocent individuals, which

[15] We used the already pre-processed dataset named "*propublica-recidivism_numerical.csv*," which can be accessed here: https://github.com/algofairness/fairness-comparison/tree/master/fairness/data/preprocessed. A detailed description of the COMPAS dataset and the use case is provided by [17] and [1].

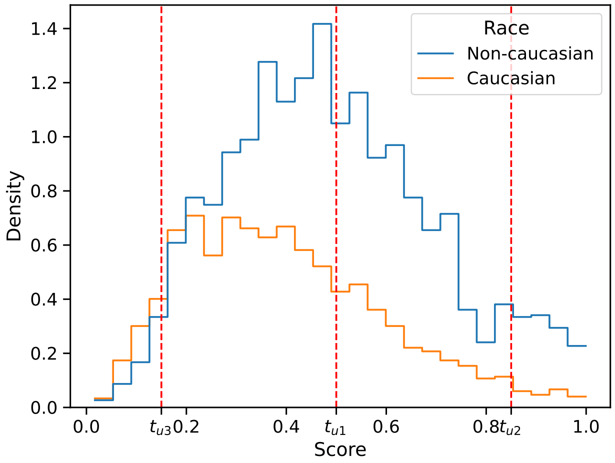


Figure 3: Score distributions by race and optimal unconstrained decision rules (t_u) for different utility functions (COMPAS)

Table 3: Optimal decision rules (COMPAS) for utility functions with different weights (TP, FP, FN, TN) paired with different fairness requirements (no fairness constraint, PPV parity, and FOR parity). The acronyms stand for base rate (BR), the optimal threshold (t_u) for unconstrained utility maximization, and the optimal thresholds (t_1, t_2) for utility maximization under fairness.

		Case 1		**Case 2**		**Case 3**	
		Caucasian	*non-Caucas.*	*Caucasian*	*non-Caucas.*	*Caucasian*	*non-Caucas.*
	BR	0.40	0.49	0.40	0.49	0.40	0.49
	TP, FP, FN, TN	1, -1, -1, 1		1, -10, -1, 1		1, -1, -10, 1	
unconstr.	t_u	t_{u1} = 0.50		t_{u2} = 0.85		t_{u3} = 0.15	
	PPV	0.65	0.68	0.92	0.92	0.42	0.50
	FOR	0.30	0.34	0.38	0.46	0.11	0.11
PPV parity	(t_1, t_2)	(0.52, 1)	(0.49, 1)	(0.84, 1)	(0.85, 1)	(0.27, 1)	(0.05, 1)
	PPV	0.67		0.92		0.49	
	FOR	0.30	0.33	0.38	0.46	0.18	0.03
FOR parity	(t_1, t_2)	(0.57, 1)	(0.47, 1)	(0.98, 1)	(0.62, 1)	(0.16, 1)	(0.15, 1)
	PPV	0.71	0.66	0.99	0.76	0.43	0.50
	FOR	0.32		0.40		0.11	

is represented with a large negative value for FN. Absent any fairness constraint, this results in a lower optimal lower-bound threshold ($t_{u3} = 0.15$), leading to more overall detentions. As opposed to case 2, this results in almost equal FORs (because the two groups' distributions are similar below the unconstrained threshold) but the two groups' PPVs differ largely. To satisfy PPV parity, it is optimal to detain almost all non-Caucasians (those with a risk score above $\tau_1 = 0.05$[16]) while detaining a much smaller fraction of Caucasians ($\tau_1 = 0.27$).

Without fairness-enforcing restrictions, the same prediction model can turn out to be fair or unfair, w.r.t. a specific fairness metrics, depending on the utility function. For example, in case 2, PPV parity is met in the unconstrained case, whereas there is a huge difference in PPVs in case 3. Note, however, that this cannot be generalized: there is no guarantee that PPV parity or FOR parity are met in the unconstrained case for a given utility function, as this depends on the groups' probability distributions. Thus, assuming that a prediction model is fair if it meets PPV parity or FOR parity is misleading because this only holds for specific utility functions and probability distributions but not in general. This contradicts the approach suggested by Northpointe, who claim that PPV and FOR are the only relevant measures to determine the treatment disparity of such a tool for different groups [11]. Interestingly, for the COMPAS example, introducing fairness constraints (in the form of PPV parity or FOR parity) leads to a lower group-specific threshold for the non-Caucasians, resulting in a higher fraction of detained individuals for the disadvantaged group – which is similar to the population 3's result in the synthetic example (see Figure 2b).[17] Further, in

[16] If the non-Caucasian group were much smaller, this would result in an upper-bound threshold, i.e., the non-Caucasian with the highest recidivism risk would be released – which is similar to group 1 in the population 3's result in the synthetic example (see Figure 2c).

[17] There is just one exception to this: enforcing FOR parity in case 2 leads to a slightly higher threshold for the disadvantaged group, which is similar to the population 1's result in the synthetic example (see Figure 2a).

some cases, it is optimal to release almost all individuals of the advantaged group or to detain almost all individuals of the disadvantaged group. This is counter-intuitive as one would expect that introducing a fairness constraint should favor the disadvantaged group.

5 CONCLUSIONS

In this paper, we analyze common group fairness metrics that have been proposed to mitigate the unfairness of algorithmic decision making systems. We formulate algorithmic fairness as a constrained optimization problem representing a decision maker who wants to maximize the total utility while also satisfying a fairness constraint. A similar solution has been provided by [10, 20] for the group fairness metrics (conditional) statistical parity, TPR parity, and FPR parity – all leading to group-specific lower-bound thresholds. In contrast to these fairness metrics, we find that for the group fairness metrics PPV parity and FOR parity, optimal decision rules take the form of group-specific lower-bound or upper-bound thresholds. This is counter-intuitive as it means that, in certain situations, it can be optimal for decision makers to select the 'worst' individuals of one group and omit the most promising ones. In the loan granting scenario, for one of the groups, this would mean that individuals who are most likely to default are granted a loan, whereas those who are most likely to pay back their loan are not granted one. Similarly, to achieve PPV parity in recidivism risk prediction, it can be optimal to release defendants with the highest recidivism risk in one of the groups. Additionally, our work shows that there is a trade-off between the group fairness criterion sufficiency and within-group fairness. Namely, to satisfy sufficiency, it is optimal to sacrifice within-group fairness for all but one of the groups.

Experts increasingly call for fairer algorithms. Considering these byproducts of the group fairness metrics PPV parity, FOR parity, and sufficiency, we emphasize that these potential consequences must be considered when imposing such fairness criteria on utility-maximizing decision makers. We hope that our findings foster the discussion of fair algorithmic decision making and, in particular, support policymakers who find themselves in the position where they need to choose a specific definition of fairness.

ACKNOWLEDGMENTS

We thank our three anonymous reviewers for their helpful feedback. This work was supported by Innosuisse – grant number 44692.1 IP-SBM – and by the National Research Programme "Digital Transformation" (NRP 77) of the Swiss National Science Foundation (SNSF) – grant number 187473.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23, 2016 (2016), 139–159. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org.

[3] Solon Barocas and Andrew D Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 3 (2016), 671–732. http://www.jstor.org/stable/24758720

[4] Joachim Baumann and Christoph Heitz. 2022. Group Fairness in Prediction-Based Decision Making: From Moral Assessment to Implementation. In *2022 9th Swiss Conference on Data Science (SDS)*.

[5] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50, 1 (2021), 3–44. https://doi.org/10.1177/0049124118782533

[6] Ran Canetti, Aloni Cohen, Nishanth Dikkala, Govind Ramnarayan, Sarah Scheffler, and Adam Smith. 2019. From Soft Classifiers to Hard Decisions. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 309–318. https://doi.org/10.1145/3287560.3287561

[7] Simon Caton and Christian Haas. 2020. Fairness in Machine Learning: A Survey. (2020). arXiv:2010.04053

[8] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big data* 5, 2 (jun 2017), 153–163. https://doi.org/10.1089/big.2016.0047

[9] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv:1808.00023 [cs.CY] https://arxiv.org/abs/1808.00023

[10] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. https://doi.org/10.1145/3097983.3098095

[11] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity.* Technical Report. Northpoint Inc. https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *ITCS 2012 - Innovations in Theoretical Computer Science Conference*. ACM Press, New York, New York, USA, 214–226. https://doi.org/10.1145/2090236.2090255 arXiv:1104.3913

[13] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. Decoupled Classifiers for Group-Fair and Efficient Machine Learning. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 119–133. http://proceedings.mlr.press/v81/dwork18a.html

[14] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

[15] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. 2016. *A Confidence-Based Approach for Balancing Fairness and Accuracy.* 144–152. https://doi.org/10.1137/1.9781611974348.17

[16] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. arXiv:1609.07236 [cs.CY] https://arxiv.org/abs/1609.07236

[17] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A Comparative Study of Fairness-Enhancing Interventions in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 329–338. https://doi.org/10.1145/3287560.3287589

[18] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2017. Predictably Unequal? The Effects of Machine Learning on Credit Markets. *SSRN* (nov 2017). https://doi.org/10.2139/ssrn.3072038

[19] Pratyush Garg, John Villasenor, and Virginia Foggo. 2020. Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, IEEE Computer Society, Los Alamitos, CA, USA, 3662–3666.

[20] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 3323–3331. arXiv:1610.02413

[21] Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. 2013. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems* 35, 3 (2013), 613–644. https://doi.org/10.1007/s10115-012-0584-8

[22] Maximilian Kasy and Rediet Abebe. 2021. Fairness, Equality, and Power in Algorithmic Decision-Making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Vol. 11. ACM, New York, NY, USA, 576–586. https://doi.org/10.1145/3442188.3445919

[23] Michael Kearns and Aaron Roth. 2019. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design.* Oxford University Press, Inc., USA.

[24] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics* 133, 1 (2017), 237–293. https://doi.org/10.1093/qje/qjx032

[25] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. (2016). arXiv:1609.05807v2

[26] Nikita Kozodoi, Johannes Jacob, and Stefan Lessmann. 2022. Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research* 297, 3 (2022), 1083–1094. https://doi.org/10.1016/j.ejor.2021.06.023

[27] Matthias Kuppler, Christoph Kern, Ruben L Bach, and Frauke Kreuter. 2021. Distributive Justice and Fairness Metrics in Automated Decision-making: How Much Overlap Is There? (2021). arXiv:2105.01441v2

[28] Derek Leben. 2020. *Normative Principles for Evaluating Fairness in Machine Learning*. Association for Computing Machinery, New York, NY, USA, 86–92. https://doi.org/10.1145/3375627.3375808

[29] Joshua K Lee, Yuheng Bu, Deepta Rajan, Prasanna Sattigeri, Rameswar Panda, Subhro Das, and Gregory W Wornell. 2021. Fair Selective Classification Via Sufficiency. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 6076–6086. https://proceedings.mlr.press/v139/lee21b.html

[30] Zachary C Lipton, Alexandra Chouldechova, and Julian McAuley. 2018. Does mitigating ML's impact disparity require treatment disparity?. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Curran Associates, Inc., 8136–8146. https://proceedings.neurips.cc/paper/2018/file/8e0384779e58ce2af40eb365b318cc32-Paper.pdf

[31] Lydia T Liu, Max Simchowitz, and Moritz Hardt. 2019. The Implicit Fairness Criterion of Unconstrained Learning. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 4051–4060. https://proceedings.mlr.press/v97/liu19f.html

[32] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. On the Applicability of Machine Learning Fairness Notions. *SIGKDD Explor. Newsl.* 23, 1 (may 2021), 14–23. https://doi.org/10.1145/3468507.3468511

[33] Sandra Gabriel Mayson. 2018. Bias In, Bias Out. *Yale Law Journal* 128 (2018), 2218. https://ssrn.com/abstract=3257004

[34] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. (2019). arXiv:1908.09635

[35] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 107–118. http://proceedings.mlr.press/v81/menon18a.html

[36] Claire Cain Miller. 2015. Can an Algorithm Hire Better Than a Human? https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html

[37] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8, 1 (mar 2021), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

[38] Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

[39] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*.

[40] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[41] Dana Pessach and Erez Shmueli. 2020. Algorithmic fairness. *arXiv preprint arXiv:2001.09784* (2020).

[42] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On Fairness and Calibration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., 5684–5693.

[43] Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. 2017. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics* 11, 3 (sep 2017), 1193–1216. https://doi.org/10.1214/17-AOAS1058

[44] Isabel Valera, Adish Singla, and Manuel Gomez Rodriguez. 2018. Enhancing the Accuracy and Fairness of Human Decision Making. In *Advances in Neural Information Processing Systems*, S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/0a113ef6b61820daa5611c870ed8d5ee-Paper.pdf

[45] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness (FairWare '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3194770.3194776

[46] Xuerui Wang, Wei Li, Ying Cui, Ruofei Zhang, and Jianchang Mao. 2011. Click-through rate estimation for rare events in online advertising. In *Online multimedia advertising: Techniques and technologies*. IGI Global, 1–12.

[47] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1171–1180. https://doi.org/10.1145/3038912.3052660