Algorithmic Decision Making and the Cost of Fairness

Sam Corbett-Davies  
Stanford University  
scorbett@stanford.edu

Emma Pierson  
Stanford University  
emmap1@stanford.edu

Avi Feller  
Univ. of California, Berkeley  
afeller@berkeley.edu

Sharad Goel  
Stanford University  
scegoel@stanford.edu

Aziz Huq  
University of Chicago  
huq@uchicago.edu

ABSTRACT

Algorithms are now regularly used to decide whether defendants awaiting trial are too dangerous to be released back into the community. In some cases, black defendants are substantially more likely than white defendants to be incorrectly classified as high risk. To mitigate such disparities, several techniques have recently been proposed to achieve algorithmic fairness. Here we reformulate algorithmic fairness as constrained optimization: the objective is to maximize public safety while satisfying formal fairness constraints designed to reduce racial disparities. We show that for several past definitions of fairness, the optimal algorithms that result require detaining defendants above race-specific risk thresholds. We further show that the optimal unconstrained algorithm requires applying a single, uniform threshold to all defendants. The unconstrained algorithm thus maximizes public safety while also satisfying one important understanding of equality: that all individuals are held to the same standard, irrespective of race. Because the optimal constrained and unconstrained algorithms generally differ, there is tension between improving public safety and satisfying prevailing notions of algorithmic fairness. By examining data from Broward County, Florida, we show that this trade-off can be large in practice. We focus on algorithms for pretrial release decisions, but the principles we discuss apply to other domains, and also to human decision makers carrying out structured decision rules.

ACM Reference format:

1 INTRODUCTION

Judges nationwide use algorithms to help decide whether defendants should be detained or released while awaiting trial [13, 32]. One such algorithm, called COMPAS, assigns defendants risk scores between 1 and 10 that indicate how likely they are to commit a violent crime based on more than 100 factors, including age, sex and criminal history. For example, defendants with scores of 7 reoffend at twice the rate as those with scores of 3. Accordingly, defendants classified as high risk are much more likely to be detained while awaiting trial than those classified as low risk.

These algorithms do not explicitly use race as an input. Nevertheless, an analysis of defendants in Broward County, Florida [2] revealed that black defendants are substantially more likely to be classified as high risk. Further, among defendants who ultimately did not reoffend, blacks were more than twice as likely as whites to be labeled as risky. Even though these defendants did not go on to commit a crime, being classified as high risk meant they were subjected to harsher treatment by the courts. To reduce racial disparities of this kind, several authors recently have proposed a variety of fair decision algorithms [16, 21, 22, 24, 25, 27].

Here we reformulate algorithmic fairness as constrained optimization: the objective is to maximize public safety while satisfying formal fairness constraints. We show that for several past definitions of fairness, the optimal algorithms that result require applying multiple, race-specific thresholds to individuals’ risk scores. One might, for example, detain white defendants who score above 4, but detain black defendants only if they score above 6. We further show that the optimal unconstrained algorithm requires applying a single, uniform threshold to all defendants. This safety-maximizing rule thus satisfies one important understanding of equality: that all individuals are held to the same standard, irrespective of race. Since the optimal constrained and unconstrained algorithms in general differ, there is tension between reducing racial disparities and improving public safety. By examining data from Broward County, we demonstrate that this tension is more than theoretical. Adhering to past fairness definitions can substantially decrease public safety; conversely, optimizing for public safety alone can produce stark racial disparities.

We focus here on the problem of designing algorithms for pretrial release decisions, but the principles we discuss apply to other domains, and also to human decision makers carrying out structured decision rules. We emphasize at the outset that algorithmic decision making does not preclude additional, or alternative, policy interventions. For example, one might provide released defendants with robust social services aimed at reducing recidivism, or conclude that it is more effective and equitable to replace pretrial detention with non-custodial supervision. Moreover, regardless of the algorithm used, human discretion may be warranted in individual cases.

1We consider racial disparities because they have been at the center of many recent debates in criminal justice, but the same logic applies across a range of possible attributes, including gender.
2 BACKGROUND

2.1 Defining algorithmic fairness

Existing approaches to algorithmic fairness typically proceed in two steps. First, a formal criterion of fairness is defined; then, a decision rule is developed to satisfy that measure, either exactly or approximately. To formally define past fairness measures, we introduce a general notion of (randomized) decision rules. Suppose we have a vector $x_i \in \mathbb{R}^p$ that we interpret as the visible attributes of individual $i$. For example, $x$ might represent a defendant’s age, gender, race, and criminal history. We consider binary decisions (e.g., $a_0 = \text{release}$ and $a_1 = \text{detain}$), and define a \textit{decision algorithm}, or a \textit{decision rule}, to be any function $d$ that specifies which action is taken for each individual. To allow for probabilistic decisions, we require only that $d(x) \in [0, 1]$.

\textbf{Definition 2.1 (Decision rule).} A decision rule is any measurable function $d : \mathbb{R}^p \mapsto [0, 1]$, where we interpret $d(x)$ as the probability that action $a_1$ is taken for an individual with visible attributes $x$.

Before defining algorithmic fairness, we need three additional concepts. First, we define the \textit{group membership} of each individual to take a value from the set $\{g_1, \ldots, g_k\}$. In most cases, we imagine these groups indicate an individual’s race; but they might also represent gender or other protected attributes. We assume an individual’s racial group can be inferred from their vector of observable attributes $x_i$, and so denote $i$’s group by $g(x_i)$. For example, if we encode race as a coordinate in the vector $x$, then $g$ is simply a projection onto this coordinate. Second, for each individual, we suppose there is a quantity $y$ that specifies the benefit of taking action $a_1$ relative to action $a_0$. For simplicity, we assume $y$ is binary and normalized to take values 0 and 1, but many of our results can be extended to the more general case. For example, in the pretrial setting, it is beneficial to detain a defendant who would have committed a violent crime if released. Thus, we might have $y_i = 1$ for those defendants who would have committed a violent crime if released, and $y_i = 0$ otherwise. Importantly, $y$ is not known exactly to the decision maker, who at the time of the decision has access only to information encoded in the visible features $x$. Finally, we define random variables $X$ and $Y$ that take on values $X = x$ and $Y = y$ for an individual drawn randomly from the population of interest (e.g., the population of defendants for whom pretrial decisions must be made).

With this setup, we now describe three popular definitions of algorithmic fairness.

1. \textit{Statistical parity} means that an equal proportion of defendants are detained in each race group [16, 17, 28, 41]. For example, white and black defendants are detained at equal rates. Formally, statistical parity means,

$$E[d(X) | Y = 1] = E[d(X)].$$

(1)

2. \textit{Conditional statistical parity} means that controlling for a limited set of “legitimate” risk factors, an equal proportion of defendants are detained within each race group [15, 27]. For example, among defendants who have the same number of prior convictions, black and white defendants are detained at equal rates. Suppose $\ell : \mathbb{R}^p \mapsto \mathbb{R}^m$ is a projection of $x$ to factors considered legitimate. Then conditional statistical parity means,

$$E[d(X) | \ell(X), g(X)] = E[d(X) | \ell(X)].$$

(2)

(3) \textit{Predictive equality} means that the accuracy of decisions is equal across race groups, as measured by false positive rate (FPR) [22, 29, 40]. This condition means that among defendants who would not have gone on to commit a violent crime if released, detention rates are equal across race groups. Formally, predictive equality means,

$$E[d(X) | Y = 0, g(X)] = E[d(X) | Y = 0].$$

(3)

As noted above, a major criticism of COMPAS is that the rate of false positives is higher among blacks than whites [2].

2.2 Related work

The literature on designing fair algorithms is extensive and interdisciplinary. Romei and Ruggieri [36] and Zliobaite [42] survey various measures of fairness in decision making. Here we focus on algorithmic decision making in the criminal justice system, and briefly discuss several interrelated strands of past empirical and theoretical work.

Statistical risk assessment has been used in criminal justice for nearly one hundred years, dating back to parole decisions in the 1920s. Several empirical studies have measured the effects of adopting such decision aids. In a randomized controlled trial, the Philadelphia Adult Probation and Parole Department evaluated the effectiveness of a risk assessment tool developed by Berk et al. [10], and found that the tool reduced the burden on parolees without significantly increasing rates of re-offense [1]. In a study by Danner et al. [14], pretrial services agencies in Virginia were randomly chosen to adopt supervision guidelines based on a risk assessment tool. Defendants processed by the chosen agencies were nearly twice as likely to be released, and these released defendants were on average less risky than those released by agencies not using the tool. We note that despite such aggregate benefits, some have argued that statistical tools do not provide sufficiently precise estimates of individual recidivism risk to ethically justify their use [39].

Several authors have developed algorithms that guarantee formal definitions of fairness are satisfied. To ensure statistical parity, Feldman et al. [16] propose “repairing” attributes or risk scores by converting them to within-group percentiles. For example, a black defendant riskier than 90% of black defendants would receive the same transformed score as a white defendant riskier than 90% of white defendants. A single decision threshold applied to the transformed scores would then result in equal detention rates across groups. Kamiran et al. [27] propose a similar method (called “local attributes, such as race. However, as frequently noted, it is difficult to restrict to “legitimate” features that do not at least partially correlate with race and other protected attributes, and so one cannot be completely “blind” to the sensitive information [15]. Moreover, unlike the other definitions of fairness, this one does not necessarily reduce racial disparities. Conditional statistical parity mitigates these limitations of the blindness approach while preserving its intuitive appeal.

Eric Holder, former Attorney General of the United States, has been similarly critical of risk assessment tools, arguing that “[e]qual justice can only mean individualized justice, with charges, convictions, and sentences befitting the conduct of each defendant and the particular crime he or she commits” [23].
massaging”) to achieve conditional statistical parity. Given a set of decisions, they stratify the population by “legitimate” factors (such as number of prior convictions), and then alter decisions within each stratum so that: (1) the overall proportion of people detained within each stratum remains unchanged; and (2) the detention rates in the stratum are equal across race groups. Finally, Hardt et al. [22] propose a method for constructing randomized decision rules that ensure true positive and false positive rates are equal across race groups, a criterion of fairness that they call equalized odds; they further study the case in which only false positive rates must be equal, which they call equal opportunity.

The definitions of algorithmic fairness discussed above assess the fairness of decisions; in contrast, some authors consider the fairness of risk scores, like those produced by COMPAS. The dominant fairness criterion in this case is calibration. Calibration means that among defendants with a given risk score, the proportion who reoffend is the same across race groups. Formally, given risk scores \( s(X) \), calibration means,

\[
Pr(Y = 1 \mid s(X), g(X)) = Pr(Y = 1 \mid s(X)).
\]

Several researchers have pointed out that many notions of fairness are in conflict; Berk et al. [9] survey various fairness measures and their incompatibilities. Most importantly, Kleinberg et al. [29] prove that except in degenerate cases, no algorithm can simultaneously satisfy the following three properties: (1) calibration; (2) balance for the negative class, meaning that among defendants who would not commit a crime if released, average risk score is equal across race group; and (3) balance for the positive class, meaning that among defendants who would commit a crime if released, average risk score is equal across race group. Chouldechova [12] similarly considers the tension between calibration and alternative definitions of fairness.

3 OPTIMAL DECISION RULES

Policymakers wishing to satisfy a particular definition of fairness are necessarily restricted in the set of decision rules that they can apply. In general, however, multiple rules satisfy any given fairness criterion, and so one must still decide which rule to adopt from among those satisfying the constraint. In making this choice, we assume policymakers seek to maximize a specific notion of utility, which we detail below.

In the pretrial setting, one must balance two factors: the benefit of preventing violent crime committed by released defendants on the one hand, and the social and economic costs of detention on the other. To capture these costs and benefits, we define the immediate utility of a decision rule as follows.

**Definition 3.1 (Immediate utility).** For \( c \) a constant such that \( 0 < c < 1 \), the immediate utility of a decision rule \( d \) is

\[
u(d, c) = E \left[ Y d(X) - c d(X) \right] = E \left[ Y d(X) \right] - c E \left[ d(X) \right].
\]

\(u(d, c)\) in their context, they consider human decisions, rather than algorithmic ones, but the same procedure can be applied to any rule.

Calibration is sometimes called predictive parity; we use “calibration” here to distinguish it from predictive equality, meaning equal false positive rates.

Some jurisdictions consider flight risk, but safety is typically the dominant concern.

The first term in Eq. (5) is the expected benefit of the decision rule, and the second term its costs. For pretrial decisions, the first term is proportional to the expected number of violent crimes prevented under \( d \), and the second term is proportional to the expected number of people detained. The constant \( c \) is the cost of detention in units of crime prevented. We call this immediate utility to clarify that it reflects only the proximate costs and benefits of decisions. It does not, for example, consider the long-term, systemic effects of a decision rule.

We can rewrite immediate utility as

\[
u(d, c) = E \left[ Y d(X) - c d(X) \right] = E \left[ p_{Y \mid X} d(X) - c d(X) \right] = E \left[ d(X) (p_{Y \mid X} - c) \right]
\]

where \( p_{Y \mid X} = Pr(Y = 1 \mid X) \). This latter expression shows that it is beneficial to detain an individual precisely when \( p_{Y \mid X} > c \), and is a convenient reformulation for our derivations below.

Our definition of immediate utility implicitly encodes two important assumptions. First, since \( Y \) is binary, all violent crime is assumed to be equally costly. Second, the cost of detaining every individual is assumed to be \( c \), without regard to personal characteristics. Both of these restrictions can be relaxed without significantly affecting our formal results. In practice, however, it is often difficult to approximate individualized costs and benefits of detention, and so we proceed with this framing of the problem.

Among the rules that satisfy a chosen fairness criterion, we assume policymakers would prefer the one that maximizes immediate utility. For example, if policymakers wish to ensure statistical parity, they might first consider all decision rules that guarantee statistical parity is satisfied, and then adopt the utility-maximizing rule among this subset.

For the three fairness definitions we consider (statistical parity, conditional statistical parity, and predictive equality) we show next that the optimal algorithms that result are simple, deterministic threshold rules based on \( p_{Y \mid X} \). For statistical parity and predictive equality, the optimal algorithms detain defendants when \( p_{Y \mid X} \) exceeds a group-specific threshold. For example, black defendants might be detained if \( p_{Y \mid X} \geq 0.2 \), and white defendants detained if \( p_{Y \mid X} \geq 0.1 \). The exact thresholds for statistical parity differ from those for predictive equality. For conditional statistical parity, the thresholds in the optimal decision rule depend on both group membership and the “legitimate” factors \( s(X) \). Finally, we show that the unconstrained utility-maximizing algorithm applies a single, uniform threshold to all individuals, irrespective of group membership. Importantly, since the optimal constrained algorithms differ from the optimal unconstrained algorithm, fairness has a cost.

To prove these results, we require one more technical criterion: that the distribution of \( p_{Y \mid X} \) has a strictly positive density on \( [0, 1] \). Intuitively, \( p_{Y \mid X} \) is the risk score for a randomly selected individual with visible attributes \( X \). Having a density means that the distribution of \( p_{Y \mid X} \) does not have any point masses; for example, the probability that \( p_{Y \mid X} \) exactly equals 0.1 is zero. Positivity means that in any sub-interval, there is non-zero (though possibly small)
probability an individual has risk score in that interval. From an
applied perspective, this is a relatively weak condition, since start-
ing from any risk distribution we can achieve this property by
smoothing the distribution by an arbitrarily small amount. But the
criterion serves two important technical purposes. First, with this
assumption, there are always deterministic decision rules that sat-
ify each fairness definition; and second, it implies that the optimal
decision rules are unique. We now state our main theoretical result.

Theorem 3.2. Suppose $D(p_{Y|X})$ has positive density on $[0, 1]$. The
optimal decision rules $d'$ that maximize $u(d, c)$ under various
fairness conditions have the following form, and are unique up to a
set of probability zero.

1. The unconstrained optimum is

\[
d^*(X) = \begin{cases} 1 & p_{Y|X} \geq c \\ 0 & \text{otherwise} \end{cases}
\]

2. Among rules satisfying statistical parity, the optimum is

\[
d^*(X) = \begin{cases} 1 & p_{Y|X} \geq t_g(X) \\ 0 & \text{otherwise} \end{cases}
\]

where $t_g(X) \in [0, 1]$ are constants that depend only on group
membership. The optimal rule satisfying predictive equality takes
the same form, though the values of the group-specific thresholds are
different.

3. Additionally suppose $D(p_{Y|X} \mid f(X) = 1)$ has positive den-
sity on $[0, 1]$. Among rules satisfying conditional statistical parity, the optimum is

\[
d^*(X) = \begin{cases} 1 & p_{Y|X} \geq t_g(X), f(X) \\ 0 & \text{otherwise} \end{cases}
\]

where $t_g(X), f(X) \in [0, 1]$ are constants that depend on group
membership and "legitimate" attributes.

Before presenting the formal proof of Theorem 3.2, we sketch out the argument. From Eq. (6), it follows immediately that (un-
constrained) utility is maximized for a rule that deterministically
detains defendants if and only if $p_{Y|X} \geq c$. The optimal rule satisfy-
ing statistical parity necessarily detains the same proportion $p^*$ of
defendants in each group; it is thus clear that utility is maximized
by setting the thresholds so that the riskiest proportion $p^*$ of
defendants is detained in each group. Similar logic establishes the result
for conditional statistical parity. (In both cases, our assumption on
the distribution of the risk scores ensures these thresholds exist.)
The predictive equality constraint is the most complicated to ana-
lyze. Starting from any non-threshold rule $d$ satisfying predictive
equality, we show that one can derive a rule $d'$ satisfying predic-
tive equality such that $u(d', c) > u(d, c)$; this in turn implies a threshold
rule is optimal. We construct $d'$ in three steps. First, we show that
under the original rule $d$ there must exist some low-risk defendants
that are detained while some relatively high-risk defendants are
released. Next, we show that if $d'$ has the same false positive rate as
$d$, then $u(d', c) > u(d, c)$ if and only if more defendants are detained
under $d'$. This is because having equal false positive rates means
that $d$ and $d'$ detain the same number of people who would not
have committed a violent crime if released; under this restriction,
detaining more people means detaining more people who would
have committed a violent crime, which improves utility. Finally,
we show that one can preserve false positive rates by releasing the
low-risk individuals and detaining an even greater number of the
high-risk individuals; this last statement follows because releasing
low-risk individuals decreases the false positive rate faster than
detaining high-risk individuals increases it.

Proof. As described above, it is clear that threshold rules are
optimal absent fairness constraints, and also in the case of statistical
parity and conditional statistical parity. We now establish the result
for predictive equality; we then prove the uniqueness of these rules.

Suppose $d$ is a decision rule satisfying equal false positive rates
and which is not equivalent to a multiple-threshold rule. We shall
construct a new decision rule $d'$ satisfying equal false positive rates,
and such that $u(d', c) > u(d, c)$. Since this construction shows any
non-multiple-threshold rule can be improved, the optimal rule must be
a multiple-threshold rule.

Because $d$ is not equivalent to a multiple-threshold rule, there
exist relatively low-risk individuals that are detained and relatively
high-risk individuals that are released. To see this, define $t_d$ to be
the threshold that detains the same proportion of group $a$ as $d$ does:

\[
E[d(X) \mid g(X) = a] = E[I(p_{Y|X} \geq t_d) \mid g(X) = a].
\]

Such thresholds exist by our assumption on the distribution of $p_{Y|X}$.
Since $d$ is not equivalent to a multiple-threshold rule, there must be
a group $a'$ for which, in expectation, some defendants below $t_d'$
will be detained and an equal proportion of defendants above $t_d'$
released. Let $2\beta$ equal the proportion of defendants "misclassi-
ed" (with respect to $t_d'$) in this way:

\[
\beta = E[I(p_{Y|X} \geq t_d') (1 - d(X)) \mid g(X) = a']
\]

\[
= E[I(p_{Y|X} < t_d') d(X) \mid g(X) = a']
\]

\[
> 0,
\]

where we note that $P_{[a']}(p_{Y|X} = t_d') = 0$.

For $0 \leq t_1 \leq t_2 \leq 1$, define the rule

\[
d'_{t_1, t_2}(X) = \begin{cases} 1 & p_{Y|X} \geq t_2, g(X) = a' \\ 0 & p_{Y|X} < t_1, g(X) = a' \end{cases}
\]

otherwise

This rule detains

\[
\beta_2(t_1, t_2) = E[I(p_{Y|X} \geq t_2) (1 - d(X)) \mid g(X) = a']
\]

defendants above the threshold who were released under $d$. Further,
\[\gamma_2(t_1, t_2) = E[I(p_{Y|X} \geq t_2) (1 - d(X))(1 - p_{Y|X}) \mid g(X) = a'] \leq (1 - t_2) \beta_2(t_1, t_2)\]
defendants are newly detained and "innocent" (i.e., would not have
gone on to commit a violent crime). Similarly, $d'_{t_1, t_2}$ releases
\[
\beta_1(t_1, t_2) = E[I(p_{Y|X} < t_1) d(X) \mid g(X) = a']
\]
defendants below the threshold that were detained under $d$, result-
ing in

\[
\gamma_1(t_1, t_2) = E[I(p_{Y|X} < t_1) d(X)(1 - p_{Y|X}) \mid g(X) = a'] \geq (1 - t_1) \beta_1(t_1, t_2)
\]
fewer innocent detainees.
Now choose $t_1 < t_{a^*} < t_2$ such that $\beta_1(t_1, t_2) = \beta_2(t_1, t_2) = \beta/2$. Such thresholds exist because: $\beta(0, t_{a^*}) = \beta(t_{a^*}, 0) = \beta$, $\beta(0, t_{a^*}) = \beta(2(1, t_{a^*}) = \beta$, $\beta(0, t_{a^*}) = \beta(t_{a^*}, 0) = \beta$, and the functions $\beta_i$ are continuous in each coordinate. Then, $\gamma_1(t_1, t_2) \geq (1-t_1)\beta/2$ and $\gamma_2(t_1, t_2) \leq (1-t_2)\beta/2$, so $\gamma_1(t_1, t_2) > \gamma_2(t_1, t_2)$. This inequality implies that $d_{\gamma(t_1, t_2)}$ releases more innocent low-risk people than it detains innocent high-risk people (compared to $d$).

To equalize false positive rates between $d$ and $d'$ we must equalize $\gamma_1$ and $\gamma_2$, and so we need to decrease $t_1$ in order to release fewer low-risk people. Note that $\gamma_1$ is continuous in each coordinate, $\gamma_1(0, t_2) = 0$, and $\gamma_2$ depends only on its second coordinate. There thus exists $t_1' \in (0, t_1)$ such that $\gamma_1(t_1', t_2) = \gamma_2(t_1', t_2) = \gamma_2(t_1, t_2)$. Further, since $t_1' < t_1$, $\beta_1(t_1', t_2) < \beta_1(t_1, t_2)$. Consequently, $d_{\gamma_1(t_1', t_2)}$ has the same false positive rate as $d$ but detains more people.

Finally, since false positive rates are equal, detaining extra people means detaining more people who go on to commit a violent crime. As a result $d_{\gamma_1(t_1', t_2)}$ has strictly higher immediate utility than $d$:

$$u(d_{\gamma_1(t_1', t_2)} - u(d, c) = \mathbb{E} \left[ d_{\gamma_1(t_1', t_2)}(X)(p_{\gamma_1|X} - c) \right] - \mathbb{E} \left[ d(X)(p_{\gamma_1|X} - c) \right]$$

$$= \mathbb{E} \left[ d_{\gamma_1(t_1', t_2)}(X)(1 - c) \right] - \mathbb{E} \left[ d_{\gamma_1(t_1', t_2)}(X)(1 - p_{\gamma_1|X}) \right]$$

$$+ \mathbb{E} \left[ d(X)(1 - c) \right] - \mathbb{E} \left[ d(X)(1 - p_{\gamma_1|X}) \right]$$

$$= (1-c) \mathbb{E} \left[ d_{\gamma_1(t_1', t_2)}(X) \right] - \mathbb{E} \left[ d(X) \right] > 0.$$

The second-to-last equality follows from the fact that $d_{\gamma_1(t_1', t_2)}$ and $d$ have equal false positive rates, which in turn implies that

$$\mathbb{E} \left[ d_{\gamma_1(t_1', t_2)}(X)(1 - p_{\gamma_1|X}) \right] = \mathbb{E} \left[ d(X)(1 - p_{\gamma_1|X}) \right].$$

Thus, starting from an arbitrary non-threshold rule satisfying predictive equality, we have constructed a threshold rule with strictly higher utility that also satisfies predictive equality; as a consequence, threshold rules are optimal.

We now establish uniqueness of the optimal rules for each fairness constraint. Optimality for the unconstrained algorithm is clear, and so we consider only the constrained rules, starting with statistical parity. Denote by $d_{\alpha}$ the rule that detains the riskiest proportion $\alpha$ of individuals in each group; this rule is the unique optimum among those with detention rate $\alpha$ satisfying statistical parity. Define

$$f(\alpha) = u(d_{\alpha}, c) = \mathbb{E} \left[ d_{\alpha}(X)p_{Y|X} \right] - c\alpha.$$

The first term of $f(\alpha)$ is strictly concave, because $d_{\alpha}$ detains progressively less risky people as $\alpha$ increases. The second term of $f(\alpha)$ is linear. Consequently, $f(\alpha)$ is strictly concave and has a unique maximizer. A similar argument shows uniqueness of the optimal rule for conditional statistical parity.

To establish uniqueness in the case of predictive equality, we first restrict to the set of threshold rules, since we showed above that non-threshold rules are suboptimal. Let $d_{\sigma}$ be the unique, optimal threshold rule having false positive rate $\sigma$ in each group. Now let $g(\sigma)$ be the detention rate under $d_{\sigma}$. Since $g$ is strictly increasing, there is a unique, optimal threshold rule $d'_{\sigma}$ that satisfies predictive equality and detains a proportion $\sigma$ of defendants: namely, $d'_{\sigma} = d_{g^{-1}(\sigma)}$. Uniqueness now follows by the same argument we gave for statistical parity.

Theorem 3.2 shows that threshold rules maximize immediate utility when the three fairness criteria we consider hold exactly. Threshold rules are also optimal if we only require the constraints hold approximately. For example, a threshold rule maximizes immediate utility under the requirement that false positive rates differ by at most a constant $\delta$ across groups. To see this, note that our construction in Theorem 3.2 preserves false positive rates. Thus, starting from a non-threshold rule that satisfies the (relaxed) constraint, one can construct a threshold rule that satisfies the constraint and strictly improves immediate utility, establishing the optimality of threshold rules.

Threshold rules have been proposed previously to achieve the various fairness criteria we analyze [16, 22, 27]. We note two important distinctions between our work and past research. First, the optimality of such algorithms has not been previously established, and indeed previously proposed decision rules are not always optimal. Second, our results clarify the need for race-specific decision thresholds to achieve prevailing notions of algorithmic fairness. We thus identify an inherent tension between satisfying common fairness constraints and treating all individuals equally, irrespective of race.

Our definition of immediate utility does not put a hard cap on the number of people detained, but rather balances detention rates with public safety benefits via the constant $c$. Proposition 3.3 below shows that one can equivalently view the optimization problem as maximizing public safety while detaining a specified number of individuals. As a consequence, the results in Theorem 3.2—where immediate utility is maximized under a fairness constraint—also hold when public safety is optimized under constraints on both fairness and the proportion of defendants detained. This reformulation is useful for our empirical analysis in Section 4.

**Proposition 3.3.** Suppose $D$ is the set of decision rules satisfying statistical parity, conditional statistical parity, predictive equality, or the full set of all decision rules. There is a bijection $f$ on the interval $[0, 1]$ such that

$$\arg \max_{d \in D} \mathbb{E} [Yd(X) - cd(X)] = \arg \max_{d \in D} \mathbb{E} [Yd(X)] \quad (7)$$

where the equivalence of the maximizers in (7) is defined up to a set of probability zero.

**Proof.** Let $f(c) = \mathbb{E} [d^*(X)]$, where $d^*$ is the unique maximizer of $u(d, c)$ under the constraint $d \in D$. For a fixed $c$, if a decision rule maximizes the right-hand side of (7) then it is straightforward to see that it also maximizes the left-hand side. By uniqueness of the solution to the left-hand side, the solution to the right-hand side is also unique. The equality in Eq. (7) thus holds for all $c$.

Feldman et al.'s [16] algorithm for achieving statistical parity is optimal only if one "repairs" risk scores $p_{Y|X}$ rather than individual attributes. Applying Kamiran et al.'s local massaging algorithm [27] for achieving conditional statistical parity yields a non-optimal multiple-threshold rule, even if one starts with the optimal single threshold rule. Harth et al. [22] hunt at the optimality of their algorithm for achieving predictive equality—and in fact their algorithm is optimal—but they do not provide a proof.
It remains to be shown that \( f \) is a bijection. For fixed \( c \) and \( \alpha \), the proof of Theorem 3.2 established that there is a unique, utility-maximizing threshold rule \( d_{\alpha} \in D \) that detains a fraction \( \alpha \) of individuals. Let \( g(\alpha) = u(d_{\alpha}, c) \). Now,
\[
g'(\alpha) = \frac{d}{d\alpha} E[Yd_{\alpha}(X) - cd_{\alpha}(X)] \\
= \frac{d}{d\alpha} (E[Yd_{\alpha}(X)] - c\alpha)
\]
and so \( g(\alpha) \) is maximized at \( \alpha^* \) such that
\[
\frac{d}{d\alpha} E[Yd_{\alpha}(X)] = c
\]
In other words, the optimal detention rate \( \alpha^* \) is such that the marginal person detained has probability \( c \) of reoffending. Thus, as \( c \) decreases, the optimal detention threshold decreases, and the proportion detained increases. Consequently, if \( c_1 < c_2 \) then \( f(c_1) > f(c_2) \), and so \( f \) is injective. To show that \( f \) is surjective, note that \( f(0) = 1 \) and \( f(1) = 0 \); the result now follows from continuity of \( f \).  

4 THE COST OF FAIRNESS

As shown above, the optimal algorithms under past notions of fairness differ from the unconstrained solution. Consequently, satisfying common definitions of fairness means one must in theory sacrifice some degree of public safety. We turn next to the question of how great this public safety loss might be in practice.

We use data from Broward County, Florida originally compiled by ProPublica [30]. Following their analysis, we only consider black and white defendants who were assigned COMPAS risk scores within 30 days of their arrest, and were not arrested for an ordinary traffic crime. We further restrict to only those defendants who spent at least two years (after their COMPAS evaluation) outside a correctional facility without being arrested for a violent crime, or were arrested for a violent crime within this two-year period. Following standard practice, we use this two-year violent recidivism metric to approximate the benefit \( y_i \) of detention: we set \( y_i = 1 \) for those who reoffended, and \( y_i = 0 \) for those who did not. For the 3,377 defendants satisfying these criteria, the dataset includes race, age, sex, number of prior convictions, and COMPAS violent crime risk score (a discrete score between 1 and 10).

The COMPAS scores may not be the most accurate estimates of risk, both because the scores are discretized and because they are not trained specifically for Broward County. Therefore, to estimate \( P_{Y|X} \), we re-train a risk assessment model that predicts two-year violent recidivism using \( L^1 \)-regularized logistic regression followed by Platt scaling [35]. The model is based on all available features for each defendant, excluding race. Our risk scores achieve higher AUC on a held-out set of defendants than the COMPAS scores (0.75 vs. 0.73). We note that adding race to this model does not improve performance, as measured by AUC on the test set.

We investigate the three past fairness definitions previously discussed: statistical parity, conditional statistical parity, and predictive equality. For each definition, we find the set of thresholds that produce a decision rule that: (1) satisfies the fairness definition; (2) detains 30% of defendants; and (3) maximizes expected public safety subject to (1) and (2). The proportion of defendants detained is chosen to match the fraction of defendants classified as medium or high risk by COMPAS (scoring 5 or greater). Conditional statistical parity requires that one define the “legitimate” factors \( f(X) \), and this choice significantly impacts results. For example, if all variables are deemed legitimate, then this fairness condition imposes no constraint on the algorithm. In our application, we consider only a defendant’s number of prior convictions to be legitimate; to deal with sparsity in the data, we partition prior convictions into four bins: 0, 1–2, 3–4, and 5 or more.

We estimate two quantities for each decision rule: the increase in violent crime committed by released defendants, relative to a rule that optimizes for public safety alone; and the proportion of detained defendants that are low risk (i.e., would be released if we again considered only public safety).

We estimate two quantities for each decision rule: the increase in violent crime committed by released defendants, relative to a rule that optimizes for public safety alone; and the proportion of detained defendants that are low risk (i.e., would be released if we again considered only public safety).

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Percent of detainees that are low risk</th>
<th>Estimated increase in violent crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical parity</td>
<td>17%</td>
<td>9%</td>
</tr>
<tr>
<td>Predictive equality</td>
<td>14%</td>
<td>7%</td>
</tr>
<tr>
<td>Cond. stat. parity</td>
<td>10%</td>
<td>4%</td>
</tr>
</tbody>
</table>

Table 1: Based on the Broward County data, satisfying common fairness definitions results in detaining low-risk defendants while reducing public safety. For each fairness constraint, we estimate the increase in violent crime committed by released defendants, relative to a rule that optimizes for public safety alone; and the proportion of detained defendants that are low risk (i.e., would be released if we again considered only public safety).

5 THE COST OF PUBLIC SAFETY

A decision rule constrained to satisfy statistical parity, conditional statistical parity, or predictive equality reduces public safety. However, a single-threshold rule that maximizes public safety generally violates all of these fairness definitions. For example, in the Broward County data, optimally detaining 30% of defendants with a single-threshold rule means that 40% of black defendants are detained, compared to 18% of white defendants, violating statistical parity. And among defendants who ultimately do not go on to commit a violent crime, 14% of whites are detained compared to 32% of blacks, violating predictive equality.
The reason for these disparities is that white and black defendants in Broward County have different distributions of risk, $p_Y|x$, as shown in Figure 1. In particular, a greater fraction of black defendants have relatively high risk scores, in part because black defendants are more likely to have prior arrests, which is a strong indicator of reoffending. Importantly, while an algorithm designer can choose different decision rules based on these risk scores, the algorithm cannot alter the risk scores themselves, which reflect underlying features of the population of Broward County.

Once a decision threshold is specified, these risk distributions determine the statistical properties of the decision rule, including the group-specific detention and false positive rates. In theory, it is possible that these distributions line up in a way that achieves statistical parity or predictive equality, but in practice that is unlikely. Consequently, any decision rule that guarantees these various fairness criteria are met will in practice deviate from the unconstrained optimum.

Kleinberg et al. [29] establish the incompatibility of different fairness measures when the overall risk $Pr(Y = 1 \mid g(X) = g_i)$ differs between groups $g_i$. However, the tension we identify between maximizing public safety and satisfying various notions of algorithmic fairness typically persists even if groups have the same overall risk. To demonstrate this phenomenon, Figure 1 shows risk score distributions for two hypothetical populations with equal average risk. Even though their means are the same, the tail of the red distribution is heavier than the tail of the blue distribution, resulting in higher detention and false positive rates in the red group.

That a single decision threshold can, and generally does, result in racial disparities is closely related to the notion of infra-marginality in the econometric literature on taste-based discrimination [3, 4, 34, 37]. In that work, taste-based discrimination [6] is equated with applying decision thresholds that differ by race. Their setting is human, not algorithmic, decision making, and so one cannot directly observe the thresholds being applied; the goal is thus to infer the thresholds from observable statistics. Though intuitively appealing, detention rates and false positive rates are poor proxies for the thresholds: these infra-marginal statistics consider average risk above the thresholds, and so can differ even if the thresholds are identical (as shown in Figure 1). In the algorithmic setting, past fairness measures notably focus on these infra-marginal statistics, even though the thresholds themselves are directly observable.

### 6 DETECTING DISCRIMINATION

The algorithms we have thus far considered output a decision $d(x)$ for each individual. In practice, however, algorithms like COMPAS typically output a score $s(x)$ that is claimed to indicate a defendant’s risk $p_Y|x$; decision makers then use these risk estimates to select an action (e.g., release or detain).

In some cases, neither the procedure nor the data used to generate these scores is disclosed, prompting worry that the scores are themselves discriminatory. To address this concern, researchers often examine whether scores are calibrated [29], as defined by Eq. (4).\(^\text{10}\) Since the true probabilities $p_Y|x$ are necessarily calibrated, it is reasonable to expect risk estimates that approximate these probabilities to be calibrated as well. Figure 2 shows that the COMPAS scores indeed satisfy this property. For example, among defendants who scored a seven on the COMPAS scale, 60% of white defendants reoffended, which is nearly identical to the 61% percent of black defendants who reoffended.

However, given only scores $s(x)$ and outcomes $y$, it is impossible to determine whether the scores are accurate estimates of $p_Y|x$.

\(^{10}\) Some researchers also check whether the AUC of scores is similar across race groups [38]. The theoretical motivation for examining AUC is less clear, since the true risk distributions might have different AUCs, a pattern that would be reproduced in scores that approximate these probabilities. In practice, however, one might expect the true risk distributions to yield similar AUCs across race groups—and indeed this is the case for the Broward County data.
or have been strategically designed to produce racial disparities. Hardt et al. [22] make a similar observation in their discussion of “oblivious” measures. Consider a hypothetical situation where a malicious decision maker wants to release all white defendants, even if they are high risk. To shield himself from claims of discrimination, he applies a facially neutral 30% threshold to defendants regardless of race. Suppose that 20% of blacks recidivate, and the decision-maker’s algorithm uses additional information, such as prior arrests, to partition blacks into three risk categories: low risk (10% chance of reoffending), average risk (20% chance), and high risk (40% chance). Further suppose that whites are just as risky as blacks overall (20% of them reoffend), but the decision maker ignores individual characteristics and labels every white defendant average risk. This algorithm is calibrated, as both whites and blacks labeled average risk reoffend 20% of the time. However, all white defendants fall below the decision threshold, so none are detained. By systematically ignoring information that could be used to distinguish between white defendants, the decision maker has succeeded in discriminating while using a single threshold applied to calibrated scores.

Figure 3 illustrates a general method for constructing such discriminatory scores from true risk estimates. We start by adding noise to the true scores (black curve) of the group that we wish to treat favorably—in the figure we use $N(0, 0.5)$ noise. We then use the perturbed scores to predict the outcomes $y_i$ via a logistic regression model. The resulting model predictions (red curve) are more tightly clustered around their mean, since adding noise removes information. Consequently, under the transformed scores, no one in the group lies above the decision threshold, indicated by the vertical line. The key point is that the red curve is a perfectly plausible distribution of risk: without further information, one cannot determine whether the risk model was fit on input data that were truly noisy, or whether noise was added to the inputs to produce disparities.

These examples relate to the historical practice of redlining, in which lending decisions were intentionally based on coarse information—usually neighborhood—in order to deny loans to well-qualified minorities [11]. Since even creditworthy minorities often resided in neighborhoods with low average income, lenders could deny their applications by adhering to a facially neutral policy of not serving low-income areas. In the case of redlining, one discriminates by ignoring information about the disfavored group; in the pretrial setting, one ignores information about the favored group. Both strategies, however, operate under the same general principle.

There is no evidence to suggest that organizations have intentionally ignored relevant information when constructing risk scores. Similar effects, however, may also arise through negligence or unintentional oversights. Indeed, we found in Section 4 that we could improve the predictive power of the Broward County COMPAS scores with a standard statistical model. To ensure an algorithm is equitable, it is thus important to inspect the algorithm itself and not just the decisions it produces.

Figure 3: Calibration is insufficient to assess discrimination. In the left plot, the black line shows the distribution of risk in a hypothetical population, and the red line shows strategically altered risk estimates in the same population. Both sets of risk scores are calibrated (right plot), but the altered risk scores are less informative and as a result guarantee that no defendants fall above the detention threshold (dashed vertical line).

7 DISCUSSION

Maximizing public safety requires detaining all individuals deemed sufficiently likely to commit a violent crime, regardless of race. However, to satisfy common metrics of fairness, one must set multiple, race-specific thresholds. There is thus an inherent tension between minimizing expected violent crime and satisfying common notions of fairness. This tension is real: by analyzing data from Broward County, we find that optimizing for public safety yields stark racial disparities; conversely, satisfying past fairness definitions means releasing more high-risk defendants, adversely affecting public safety.

Policymakers face a difficult and consequential choice, and it is ultimately unclear what course of action is best in any given situation. We note, however, one important consideration: with race-specific thresholds, a black defendant may be released while an equally risky white defendant is detained. Such racial classifications would likely trigger strict scrutiny [18], the most stringent standard of judicial review used by U.S. courts under the Equal Protection Clause of the Fourteenth Amendment. A single-threshold rule thus maximizes public safety while satisfying a core constitutional law rule, bolstering the case in its favor.

To some extent, concerns embodied by past fairness definitions can be addressed while still adopting a single-threshold rule. For example, by collecting more data and accordingly increasing the accuracy of risk estimates, one can lower error rates. Further, one could raise the threshold for detaining defendants, reducing the number of people erroneously detained from all race groups. Finally, one could change the decision such that classification errors are less costly. For example, rather than being held in jail, risky defendants might be required to participate in community supervision programs.

When evaluating policy options, it is important to consider how well risk scores capture the salient costs and benefits of the decision. For example, though we might want to minimize violent crime conducted by defendants awaiting trial, we typically only observe
While perhaps a reasonable strategy in practice, we note that discretion—rather than minor offenses—since such incidents are less susceptible to biased observation. In particular, Skeem and Lowencamp note that the racial distribution of individuals arrested for violent offenses is in line with the racial distribution of offenders inferred from victim reports and also in line with self-reported offending data.

One might similarly worry that the features x are biased in the sense that factors are not equally predictive across race groups, a phenomenon known as subgroup validity [4]. For example, housing stability might be less predictive of recidivism for minorities than for whites. If the vector of features x includes race, an individual’s risk score \( p(y|x) \) is in theory statistically robust to this issue, and for this reason some have argued race should be included in risk models [7]. However, explicitly including race as an input feature raises legal and policy complications, and as such it is common to simply exclude features with differential predictive power [14]. While perhaps a reasonable strategy in practice, we note that discarding information may inadvertently lead to the redlining effects we discuss in Section 6.

Risk scores might also fail to accurately capture costs in specific, idiosyncratic cases. Detaining a defendant who is the sole caretaker of her children arguably incurs higher social costs than detaining a defendant without children. Discretionary consideration of individual cases might thus be justified, provided that such discretion does not also introduce bias. Further, the immediate utility of a decision rule might be a poor measure of its long-term costs and benefits. For example, in the context of credit extensions, offering loans preferentially to minorities might ultimately lead to a more productive distribution of wealth, combating harms from historical under-investment in minority communities.

Finally, we note that some decisions are better thought of as group rather than individual choices, limiting the applicability of the framework we have been considering. For example, when universities admit students, they often aim to select the best group, not simply the best individual candidates, and may thus decide to deviate from a single-threshold rule in order to create diverse communities with varied perspectives and backgrounds [33].

Experts increasingly rely on algorithmic decision aids in diverse settings, including law enforcement, education, employment, and medicine [5, 8, 19, 20, 26]. Algorithms have the potential to improve the efficiency and equity of decisions, but their design and application raise complex questions for researchers and policymakers. By clarifying the implications of competing notions of algorithmic fairness, we hope our analysis fosters discussion and informs policy.

ACKNOWLEDGMENTS

We thank Imanol Arrieta Ibarra, Angélique Christin, Alexander Gelber, Andrew Gelman, Jan Overgoor, Ravi Shroff, and Jennifer Skeem for their helpful comments. This work was supported in part by the John S. and James L. Knight Foundation, and by the Hellman Fellows Fund. Data and code to reproduce our results are available at https://github.com/Sharad/cost-of-fairness.

REFERENCES


Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. ProPublica (5 2016).


