Fair Prediction with Endogenous Behavior

Christopher Jung[1], Sampath Kannan[1], Changhwa Lee[2], Mallesh M. Pai[3], Aaron Roth[1], and Rakesh Vohra[2]

[1]Department of Computer and Information Sciences, University of Pennsylvania
[2]Department of Economics and Department of Electrical & Systems Engineering, University of Pennsylvania
[3]Department of Economics, Rice University

February 17, 2020

Abstract

There is increasing regulatory interest in whether machine learning algorithms deployed in consequential domains (e.g. in criminal justice) treat different demographic groups "fairly." However, there are several proposed notions of fairness, typically mutually incompatible. Using criminal justice as an example, we study a model in which society chooses an incarceration rule. Agents of different demographic groups differ in their outside options (e.g. opportunity for legal employment) and decide whether to commit crimes. We show that equalizing type I and type II errors across groups is consistent with the goal of minimizing the overall crime rate; other popular notions of fairness are not.

1 Introduction

Algorithms to automate consequential decisions such as hiring [Miller, 2015], lending [Byrnes, 2016], policing [Rudin, 2013], and criminal sentencing [Barry-Jester et al., 2015] are frequently suspected of being unfair or discriminatory. The suspicions are not hypothetical. The 2016 ProPublica study [Angwin et al., 2016] of the COMPAS Recidivism Algorithm (used to inform criminal sentencing decisions by attempting to predict recidivism) found that the algorithm was significantly more likely to incorrectly label black defendants as recidivism risks compared to white defendants, despite similar overall rates of prediction accuracy between populations. Since then, discoveries of "algorithmic bias" have proliferated, including a recent study of racial bias by algorithms that prioritize patients for healthcare [Obermeyer et al., 2019]. Thus spurred, policymakers, regulators, and computer scientists have proposed that algorithms be designed to satisfy notions of fairness (see for instance O'Neil [2016], Barocas et al. [2018], Roth and Kearns [2019] for overviews).

This raises a question: what measure(s) of fairness should designers be held to, and how do these constraints interact with the original objectives the algorithm was designed to target? The COMPAS case illustrates that the answer is not clear. ProPublica and Northpointe (the company that designed COMPAS) advocated for different measures of fairness. ProPublica argued that the algorithm's predictions did not maintain parity in false positive and false negative rates between white and black defendants,[1] while Northpointe countered that their algorithm satisfied predictive parity.[2] Subsequent research identified hard trade-offs in the choice of fairness metrics: under some mild conditions, the two requirements above cannot simultaneously be satisfied (Kleinberg et al. [2016], Chouldechova [2017]). This inspired a literature proposing (or criticizing) notions of fairness based on ethical/ normative grounds. The literature evaluates algorithms on the basis of these measures, and/or proposes novel algorithms that better trade-off the goals

[1] Northpointe's algorithm had differing Type-1 and Type-2 error rates across the two groups.

[2] Roughly, the accuracy of COMPAS scores was the same for both groups at all risk levels.

of the original designer (decision accuracy, algorithmic efficiency) with these fairness desiderata.[3] In general, the different proposed fairness measures are fundamentally at odds with one another. For example, in addition to the impossibility results due to Kleinberg et al. [2016], Chouldechova [2017], enforcing parity of false positive or false negative rates for e.g. parole decisions typically requires making parole decisions using different thresholds on the posterior probability that an individual will commit a crime for different groups. This has itself been identified by [Corbett-Davies and Goel, 2018] as a potential source of "unfairness".

This line of research is subject to two criticisms. First raised by, for example [Corbett-Davies and Goel, 2018]: these notions of fairness are disconnected from and lead to unpalatable tradeoffs with other economic and social quantities and consequences one might care about. Second, the literature almost exclusively assumes that the agent types, which are relevant to the decision at hand, are exogenously determined, i.e. unaffected by the decision rule that is selected. For instance, in the criminal justice application described, individual choices of whether to commit a crime or not, and therefore the overall crime rates, are fixed and not affected by policy decisions made at a societal level (e.g. what legal standards are used to convict, policing decisions etc). In settings like this where agent decisions are exogenously fixed, Corbett-Davies and Goel [2018] and Liu et al. [2019b] observe that optimizing natural notions of welfare and accuracy (incarcerating the guilty, acquitting the innocent) are achieved by decision rules that select a uniform threshold on "risk scores" that are well calibrated — for example, the posterior probability of criminal activity — which tend *not* to satisfy statistical notions of fairness that have been proposed in the literature. Does this mean that setting uniform thresholds on equally calibrated risk scores is better aligned with natural societal objectives than is asking for parity in terms of false positive and negative rates across populations?

In this paper, we consider a setting in which agent decisions are endogenously determined and show that in this model, the answer is *no*: in fact, parity of false positive and negative rates (sometimes known in this literature as *equalized odds* Hardt et al. [2016]) is aligned with the natural objective of minimizing crime rates. Parity of positive predictive value and posterior threshold uniformity are *not*. Although the model need not be tied to any particular application, we develop it using the language of criminal justice. We treat agents as rational actors whose decisions about whether or not to commit crimes are *endogenously* determined as a function of the incentives given by the decision procedure society uses to punish crime. The possibility for unfairness arises because agents are ex-ante heterogeneous: their demographic group is correlated with their underlying incentives— for example each individual has a private *outside option* value for not committing a crime, and the distribution of outside options differs across groups. Our key result is that policies that are optimized to minimize crime rates are compatible with a popular measure of demographic fairness — equalizing false positive and negative rates across demographics — and are generally incompatible with equalizing positive predictive value and uniform posterior thresholds. Thus, which of these notions of fairness is compatible with natural objectives hinges crucially on whether one believes that criminal behavior is responsive to policy decisions or not.

Our results have direct implications for regulatory testing for unfairness. Often, in settings of interest, a regulator does not directly observe the decision rule used by an adjudicator. However, the regulator may wish to test whether the adjudicator is using a "fair" rule, i.e. whether the adjudicators choices are biased towards or against some demographic group. Following a tradition starting with Becker [2010], one standard used is called an outcome test, i.e. comparing, ex-post, the classification assigned by the adjudicator to observed outcomes. For instance, in a criminal justice setting, one may compare the judge's decision to the (somehow obtained) actual innocence or guilt of the defendants, or in a lending setting, compare the lender's decision on whom to extend loans to with the actual repayment outcomes of loan applicants etc.

In this context, a given prescription on what constitutes a "fair" or non-discriminatory rule maps into a corresponding outcome test. In particular, a test that is popularly used by researchers and regulators corresponds to the common-posterior-threshold rule described above. As already mentioned, this is not the best test in our model. When used, this test attempts to evaluate whether the adjudicator is using a common posterior threshold across groups by evaluating whether the marginal agents across groups have

[3] See e.g. Dwork et al. [2012], Hardt et al. [2016], Corbett-Davies and Goel [2018], Corbett-Davies et al. [2017], Feller et al. [2016], Friedler et al. [2016], Kearns et al. [2018], Hébert-Johnson et al. [2018], Liu et al. [2019b] for a small sample of an enormous literature.

similar probabilities of different outcomes. [4] However, implementing this test is difficult: identifying (and being sure that one has correctly identified) the marginal agent in each group is hard (this is roughly the infra-marginality problem, see e.g. [Simoiu et al., 2017]). For instance, there may be information observed by the decision maker but not by the regulator/ econometrician (an oft cited example is that police observe a suspect's demeanor, and use this as a factor, but this cannot be quantified). By contrast, if our maintained assumptions are valid, then an adjudicator wishing to minimize crime should use a rule that equalizes false positive and false negative rates across demographic groups. This is easy to estimate and test: there is no need to identify a marginal agent.

1.1 Overview of Model and Results

We first derive our results in an extremely simple baseline model to highlight the underlying intuition. We then show that our conclusions are robust to a number of elaborations and generalizations of the model.

The Baseline Model

Our baseline model (in Section 2) has a mass of agents who each belong to one of two demographic groups. Each agent has a single choice on the extensive margin, for instance a binary choice of whether or not to commit a crime; or whether or not to acquire human capital, etc. To fix ideas, in this paper we frame the matter as a decision about whether to commit a crime.

An adjudicator has to classify each agent as guilty or innocent. This classification is based on a noisy signal that the adjudicator receives of each agent's choice; the distribution of this signal depends only on the agent's choice, and not on her group. Further, the adjudicator observes the group membership of each agent. The adjudicator commits ex-ante to a classification rule, i.e. how it will classify agents as a function of the signal received, and potentially the agent's group membership.

Agents are expected payoff maximizers who enjoy a monetary benefit from crime but also a cost of being declared guilty of the crime. In choosing whether to commit a crime, they compare their expected net benefit from the crime to an outside option. The costs and benefits are privately known to the agent, but not to the adjudicator (who only sees group membership). The only distinction between groups is that the distributions of costs and benefits may be different for different groups. For example, individuals from different groups might have different legal employment opportunities, different costs of incarceration (e.g. differences in stigma), etc. The model is flexible enough to allow (potentially different) fractions of the population in each group who are rigidly law-abiding (i.e. do not commit a crime regardless of circumstance) or hardened criminals (i.e. will commit a crime regardless of circumstance), and a variety of responses to incentives in between these two extremes. We don't model the source of this heterogeneity: it is exogenous, and the distribution is known to the adjudicator. Given these preferences, the adjudicator's decision rule determines their choices, which in turn determines the overall crime rate in each population.

The adjudicator's objective is to minimize the overall crime rate, i.e. the total mass of agents that choose to commit a crime. While we model the adjudicator as knowing that the underlying groups are heterogeneous (i.e. knowing the above distributions that describe each group), the adjudicator is not biased for or against any group, nor is there any underlying preference for fairness. Our main result (see Section 3) is that the classifier that minimizes the crime rate is fair according to a metric that has attracted attention in the literature: setting *different* thresholds on the posterior probability of crime for each group so as to guarantee equality of false positive and negative rates. This corresponds to setting the *same* threshold on signals across groups.

To dig a little deeper into this result, the equilibrium crime rate in each population can be viewed as the adjudicator's prior belief in equilibrium that an agent has committed a crime, given knowledge only of her group membership. Given the noisy signal, the adjudicator has a posterior belief that the agent has committed a crime. In a static environment, the optimal classification rule for an adjudicator who wishes to optimize classification accuracy will be a group-independent threshold on her posterior belief that an agent

[4] For a discussion of this in the context of evaluating the fairness of lending standards, see Ferguson and Peters [1995].

has committed a crime. Note that priors will generally differ between populations (because outside option distributions differ, crime rates in the two populations differ). Therefore policies corresponding to group-independent thresholds on posterior beliefs will typically correspond to applying group-dependent thresholds on signals and vice versa.

Equalizing false-positive and false-negative rates across groups then corresponds to setting identical thresholds on the raw *signal* for each group. It can be viewed as a commitment to avoid conditioning on group membership, even when group membership information is ex-post informative for classification. The intuition is that if the adjudicator uses the same threshold on the posterior belief that an agent has committed a crime for each group, the decision rule is making use of information contained within each group's prior. Although this information is statistically informative of the decision made by the agent, it is not within the agent's control. Using this information therefore only distorts the (dis-)incentives to commit crime. On the other hand, if the adjudicator uses the same threshold on agents' signals for each group (and hence different thresholds on posterior beliefs), decisions are made only as a function of information under the control of the agents, and hence are more effective at discouraging crime. The equalizing of false-positive and false-negative rates across the groups follows from this.

Extensions

The main insights of our baseline model continue to hold even when many of its core assumptions are relaxed. We summarize them here.

First, the baseline model assumes that a signal is observed by the adjudicator for every agent (or equivalently, at equal rates across populations). There is significant empirical evidence, however, that this is often not the case: for example, arrest rates (and hence prosecution rates) are substantially higher in minority populations for certain drug offenses, despite evidence that the underlying prevalence is more uniform across groups [Mitchell and Caudy, 2015]. Section 4 introduces an elaboration of the baseline model based on Persico [2002].

In this variant, the adjudicator must rely on an intermediary (police) to inspect agents and generate a signal. The adjudicator observes a signal from an individual only if the police inspect them, and individuals are punished only if they are inspected *and* their signal crosses the adjudicator's threshold for establishing guilt. The police have their own objectives: to maximize the number of successful inspections (e.g. inspections that result in an arrest). Formally, we study the following game: The adjudicator commits to a decision rule on guilt/ innocence based on signals. Then police and agents play a simultaneous move game: The police choose an *inspection intensity* for each group to maximize their objectives, given the adjudicator's rule and crime rates in each group, subject to an overall capacity constraint. Agents of different groups commit crime based on both the adjudicator's rule, and the police's inspection rate for their group. We show that similar results continue to hold in this model, i.e. the optimal rule for the adjudicator will continue to equalize the disincentive to commit crime across the groups. This will result in equalizing *conditional* false-positive and false-negative rates across the groups, i.e. the rate conditional on being inspected.

In Section 5, we consider a setting in which the signal distribution depends not just on the action chosen by the agent (crime or no crime) but may also depend on their group. For example, the underlying signal generating process may be less noisy for agents from certain groups, and noisier for others. Of course, in general, the structure of the optimal solution is closely tied to the relationship between the signal distributions, and our results cannot carry over without further assumptions.[5] Nevertheless, the insights provided by our previous analysis allow us to study the tradeoff between the adjudicator's objective (minimizing overall crime) and various fairness notions. In particular, Theorem 5.1 gives conditions under which our baseline insights continue to hold, i.e. conditions under which the adjudicator's optimal rule will continue to equalize the disincentive to commit crime across groups. Conversely, Theorem 5.3 shows conditions under which rules that equalize false-positive or false-negative rates across groups outperform rules that equalize incentives.

[5] Consider, for example, the case in which there is a perfectly informative signal for one group but not for another. Then, the optimal solution will have zero error for the former group, whereas error will be inevitable for the latter, for whom we only have a noisy signal.

Finally, note that crime rate in our model (or what is referred to in the classification literature as "base rate") is endogenous, where it is normally modeled as exogenously fixed. This allows us to consider an additional notion of fairness that is ill-defined in the existing literature; namely classification rules that equalize base rate across groups. In Section 6, we study when this may be better or worse than the other notions we considered above in terms of the adjudicator's objective (overall crime rate).

1.2 Additional Related Work

The economic literature starting from Arrow [1972] has considered models of discrimination where agent decisions (e.g. to gain education) are endogenous and their incentives determined by a principal's choice (e.g. employer's hiring rule). Coate and Loury [1993], Foster and Vohra [1992] study models in which individuals have a choice about how much effort to exert, and identical populations can have different outcomes in (e.g. hiring markets) because of asymmetric self-confirming equilibria. There has also been extensive interest in the design of affirmative action policies for example in higher education. Loury [1977] makes the case that affirmative action may be necessary to correct historical inequity by constructing a dynamic model in which heterogeneity between two groups may persist if the principal uses a non-discriminatory rule going forward. The subsequent literature is too large to comprehensively cite, see Fang and Moro [2011] for a survey.

There has also been substantial interest in evaluating outcome data for evidence of discrimination, using (or developing) an underlying theoretical prediction of how such discrimination would manifest: see e.g. Knowles et al. [2001], Persico [2002] or Anwar and Fang [2006] in the context of policing/ traffic stops. A large literature has studied lending data for evidence of discrimination against women and minorities: see e.g. Ferguson and Peters [1995] or Ladd [1998] for overviews of both the debate on what measures of (un-)fairness to use, and an overview of the existing research.

More recently, in the computer science literature, several papers consider effort-based models that are similar in spirit to Coate and Loury [1993], Foster and Vohra [1992]. Hu and Chen [2018] propose a two-stage model of a labor market with a "temporary" (i.e. internship) and "permanent" stage, and study the equilibrium effects of imposing a fairness constraint ("statistical parity", which corresponds to hiring from two populations at equal rates) on the temporary stage. Liu et al. [2019c] consider a model of the labor market with higher dimensional signals, and study equilibrium effects of "subsidy" interventions which can lessen the cost of exerting effort. Kannan et al. [2019] study the effects of admissions policies on a two-stage model of education and employment, in which a downstream employer makes rational decisions, but student types are exogenously determined. Two recent papers Liu et al. [2019a], Mouzannar et al. [2019] study non game-theoretic models by which classification interventions in an earlier stage can have effects on individual type distributions at later stages, and show that for many commonly studied fairness constraints (including several that we consider in this paper), their effects can either be positive or negative in the long term, depending on the functional form of the relationship between classification decisions and changes in the agent type distribution.

2 Preliminaries

Baseline Model

Each agent belongs to a group $g \in \mathcal{G}$. A group corresponds to some observable characteristic of the agent, for instance race or gender. There are N_g agents in group g. For simplicity assume just two groups $\{1, 2\}$ though the results extend straightforwardly to any finite number. Each agent makes a single binary decision to either commit a crime (c) or remain innocent (i).

Then, for each agent, the adjudicator observes a random signal $s \in \mathbb{R}$ which is informative of the agent's guilt/innocence. The distribution of the signal depends only on whether the agent has committed a crime (and is therefore conditionally independent of their group). Criminals' signals are drawn according to the distribution F^{c} (with pdf f^{c}) and innocents' signals drawn from F^{i} (with pdf f^{i}). It is without loss of generality (reordering signals if necessary) to assume that the signal distributions satisfy the Monotone Likelihood Ratio Property (MLRP), i.e. higher signals imply a higher likelihood of guilt.

The adjudicator commits to a decision rule β, which labels an agent in group g with signal s as guilty with probability $\beta_g(s) \in [0, 1]$. Note that implicitly this means the adjudicator perfectly observes an agent's group membership, along with the signal, at the time of adjudication. We write $q = 1$ to indicate that the agent is labeled guilty and $q = 0$ otherwise.

Now we describe agents' incentives to commit a crime in the first place. The agent receives a reward ρ when he commits a crime, but pays a penalty of κ if he is labeled as guilty. An agent who does not commit a crime receives his outside option value ω. All three quantities are privately known only to the agent and are drawn independently from a distribution that may potentially differ across the groups. An agent in group g commits a crime if his net utility from committing a crime is higher than not:

$$\rho - \kappa \Pr(q = 1 \mid c, g) \geq \omega - \kappa \Pr(q = 1 \mid i, g), \tag{1}$$

which can be written as

$$\Pr(q = 1 \mid c, g) - \Pr(q = 1 \mid i, g) \leq \frac{\rho - \omega}{\kappa}. \tag{2}$$

where $\frac{\rho-\omega}{\kappa}$ is the *marginal benefit* of committing a crime normalized by the penalty. Define

$$\Delta_g = \Pr(q = 1 \mid c, g) - \Pr(q = 1 \mid i, g) \tag{3}$$

as the *disincentive* for committing a crime: it is the group specific additional probability of being found guilty having committed a crime relative to not. Then, the crime rate of group g can be expressed in terms of H_g, the survivor function (i.e. 1-CDF) associated with the relevant quantity on the right hand side of (2) given the joint distribution of ρ, κ and ω:

$$\mathrm{CR}_g = \Pr\left(\Delta_g \leq \frac{\rho - \omega}{\kappa}\right) = H_g\left(\Delta_g\right). \tag{4}$$

The adjudicator's objective is to minimize the overall crime rate, i.e. to solve

$$\min_{\beta \in B} \sum_{g \in \mathcal{G}} N_g \mathrm{CR}_g. \tag{OPT}$$

where of course, β determines Δ_g by (3) and therefore CR_g as per (4). Here, B is the set of all feasible policies for the adjudicator, i.e. $B = \{\beta_g, g \in \mathcal{G} : \beta_g : \mathbb{R} \to [0, 1]\}$.

2.1 Fairness Measures

We are interested in how decision rules β respecting various notions of fairness perform relative to the optimal policy, and to each other. There are five main notions of fairness that we discuss throughout this paper, each of which corresponds to equalizing some statistical quantity across groups. Three of them have been considered both in the literature and in the popular press: equalizing false positive rates, false negative rates, and positive predictive value. These three notions of fairness are of particular interest to us because it has been shown that attaining all three measures simultaneously is impossible (Kleinberg et al. [2016], Chouldechova [2017]).

Given a policy β_g for group g, true positive rate (TPR), false positive rate (FPR), false negative rate (FNR), and positive predictive value (PPV) are defined as

$$\mathrm{TPR}_g = \Pr(q = 1 \mid c, g) = \int_{\mathbb{R}} f^{\mathrm{c}}(s)\beta_g(s)ds \tag{TPR}$$

$$\mathrm{FPR}_g = \Pr(q = 1 \mid i, g) = \int_{\mathbb{R}} f^{\mathrm{i}}(s)\beta_g(s)ds \tag{FPR}$$

$$\mathrm{FNR}_g = \Pr(q = 0 \mid c, g) = \int_{\mathbb{R}} f^{\mathrm{c}}(s)(1 - \beta_g(s))ds \ (= 1 - \mathrm{TPR}_g) \tag{FNR}$$

$$\mathrm{PPV}_g = \Pr(c \mid q = 1, g) = \frac{\mathrm{CR}_g \mathrm{TPR}_g}{\mathrm{CR}_g \mathrm{TPR}_g + (1 - \mathrm{CR}_g)\mathrm{FPR}_g} \tag{PPV}$$

Note that in light of these definitions, we can rewrite (3) as:

$$\Delta_g = \mathrm{TPR}_g - \mathrm{FPR}_g \tag{Δ}$$

Additionally, we propose two *new* notions of fairness: equalizing disincentives (denoted Δ) and equalizing crime rates (denoted CR). We say the policy β achieves fairness notion $\xi \in \{\mathrm{FPR}, \mathrm{FNR}, \mathrm{PPV}, \Delta, \mathrm{CR}\}$ if the resulting respective quantity is the same across the groups when the adjudicator chooses policy β. We write B_ξ to be the set of all policies that achieve fairness notion ξ.

Given this framework we are interested in two questions. First, which of these fairness notions is compatible with the adjudicator's problem (OPT)? Second, under what conditions is a particular fairness notion better than another in terms of the objective of minimizing overall crime rate, i.e. when do we have that for fairness notions ξ, ξ':

$$\min_{\beta \in B_\xi} \sum_{g \in \mathcal{G}} N_g \mathrm{CR}_g \leq \min_{\beta \in B_{\xi'}} \sum_{g \in \mathcal{G}} N_g \mathrm{CR}_g.$$

One final piece of notation will be useful. We denote by $\beta^\star$ the solution to the adjudicator's problem (OPT), i.e. the policy that minimizes crime overall. We sometimes refer to this as the optimal policy. Further, for fairness notion ξ, we denote as $\beta^\star_\xi$ the solution to the adjudicator's problem among all rules that satisfy fairness notion ξ,i.e. the rule that solves

$$\min_{\beta \in B_\xi} \sum_{g \in \mathcal{G}} N_g \mathrm{CR}_g.$$

3 Results in the Baseline Model

The main result we build around is that the solution to the adjudicator's problem (OPT) is naturally "fair" in terms of three of the five measures above. It provides an interesting counterpoint to the impossibility results of Kleinberg et al. [2016] and Chouldechova [2017]. Those results state that it is impossible to simultaneously equalize false negative rates, false positive rates, and positive predictive value across groups. This raises a question of which of the fairness measures should be preferred over the others. By endogenizing the base rate of criminal activity, we find that equalizing false positive rates and equalizing false negative rates are preferred to equalizing positive predictive value in the sense that the former two are compatible with the optimal policy while the latter is not. Formally,

Theorem 3.1. *The adjudicator's optimal policy $\beta^\star$ (i.e. the policy which solves* (OPT)*) equalizes the disincentive to commit crime* (Δ) *across groups. As a result, it also equalizes the false negative rates* (FNR)*, and false positive rates* (FPR)*.*

Proof. First note that because β_g can be set independently for each group g, minimizing the total crime rate is achieved by individually minimizing the crime rate within each group. Recall that the crime rate within a group is $H_g(\Delta_g)$. This in turn is minimized by maximizing the disincentive of crime Δ_g, since H_g being a survivor function is non-increasing. Recall that

$$\Delta_g = \int_{\mathbb{R}} (f^{\mathrm{c}}(s) - f^{\mathrm{i}}(s))\beta_g(s)ds.$$

Therefore the optimal β_g is independent of the (group-dependent) distribution over private values defining H_g, and is therefore the same for all groups:

$$\beta_g(s) = \begin{cases} 1 \text{ if } f^{\mathrm{c}}(s) \geq f^{\mathrm{i}}(s), \\ 0 \text{ if } f^{\mathrm{c}}(s) < f^{\mathrm{i}}(s), \end{cases}$$

Since the disincentive to commit crime is a function only of β_g, this results in the same disincentive to commit crime at the optimal solution.

Finally note that both the FNR_g and FPR_g for each group are a function only of $f^{\mathrm{i}}(s)$ and $f^{\mathrm{c}}(s)$ (which are identical across groups), and the chosen policies $\beta_g(s)$, which we have shown in the optimal solution will be identical across groups. Hence, the adjudicator's optimal policy will equalize false positive rates and false negative rates across groups. □

Rather than thinking of an arbitrary function $\beta_g(s)$, it is more natural to think of the adjudicator as selecting a threshold T_g for each group g so that any member of group g whose signal s exceeds T_g is labeled guilty. That is

$$\beta_g(s) = \begin{cases} 1 \text{ if } s \geq T_g \\ 0 \text{ if } s < T_g. \end{cases}$$

Remark 3.2. *Since, f^c and f^i satisfy the MLRP property (i.e. $\frac{f^c(s)}{f^i(s)}$ is non-decreasing in s), $\beta^\star$ is a threshold policy by observation. Note that if strict MLRP holds, then the optimal thresholds are unique.*

Under a threshold policy, group g's true positive rate reduces to $\mathrm{TPR}_g = F^{\mathrm{c}}(T_g)$ and the false positive rate simplifies to $\mathrm{FPR}_g = 1 - F^{\mathrm{i}}(T_g)$.

3.1 Discussion

Posterior Thresholds

It is interesting to contrast the policy of setting equal thresholds on the signal, which we show to be the optimal policy here, as opposed to equal thresholds on the 'posterior' or another calibrated risk score. The latter is advocated in Corbett-Davies et al. [2017], for example.

In that paper, the authors consider a setting where crime choices are exogenous/ fixed, and study the choice of policy that minimizes weighted misclassification rates (i.e. acquittal of the guilty and incarceration of the innocent). They show that an optimal policy involves a common 'threshold' on the posterior across groups, i.e., first, the adjudicator estimates the prior probability that an individual has committed a crime by considering the base rate of crime for the individual's group and then uses the observed signal to update her prior probability to her posterior belief that the individual in question has committed a crime. Second, the individual is deemed guilty if the posterior probability of guilt exceeds some threshold.

Of course, in our setting, choice of crime is endogenous, and the planner's objective function is minimizing crime rate rather than minimizing mislabeling costs. Nevertheless, it is interesting to inquire into the implications of equalizing the thresholds on the posterior in our setting.

In our setting, the posterior after observing the signal s and the group g is

$$\Pr(c \mid s, g) = \frac{f^{\mathrm{c}}(s)\mathrm{CR}_g}{f^{\mathrm{c}}(s)\mathrm{CR}_g + f^{\mathrm{i}}(s)(1 - \mathrm{CR}_g)}$$

which increases in s when the signal structure satisfies monotone likelihood ratio property. Thresholding the posterior corresponds to choosing a value $\pi_g \in [0, 1]$ and classifying as guilty whenever $\Pr(c \mid s, g) \geq \pi_g$.

Let T^* be the threshold on the signal under the planner's optimal policy. The optimal policy classifies the defendant as guilty if the signal s exceeds the threshold T^*. With the monotone likelihood ratio property, this implies a threshold rule on the posterior which is to classify the defendant as guilty whenever the posterior exceeds

$$\pi_g = \Pr(c \mid T^*, g).$$

By observation, the posterior thresholds are equalized across groups if and only if $\mathrm{CR}_g = \mathrm{CR}_{g'}$, i.e. if and only if crime rates are equalized under the optimal policy T^*. This in turn will only occur if $H_g(\Delta) = H_{g'}(\Delta)$ which of course will not obtain in general because H_g need not be the same as $H_{g'}$.

4 Heterogenous Signal Observation

The baseline model assumes that when an agent commits a crime, the adjudicator observes the signal generated and adjudicates. What if the signals generated by members of each group are observed at different rates? This can happen if the adjudicator relies on an intermediary to record the signals. For example, in the crime application, the groups may be policed at different rates. Their (dis)incentives to commit crime will then differ as a result. Critically, we suppose that the police's incentives differ from the adjudicator's. The adjudicator's choice of rule therefore influences the police's choice on how to divide their manpower across different groups. Both the adjudicator's rule and the police's choice influence the incentives of agents to commit crime, and ultimately determine the overall crime levels in society.

To model this, we build upon upon the framework of Persico [2002]. There are a continuum of police officers who choose inspection intensities for each group $\{\theta_g\}_{g\in\mathcal{G}}$ given a search capacity S. The choice of inspection intensity determines the rate at which signals are observed from the two groups: upon inspection, a police officer observes a signal about whether a crime was committed or not. In Persico [2002], the signal is assumed to be perfect. We depart from this assumption in that in our setting, the observed signal is *noisy*, as in our baseline model. The adjudicator, as before, wishes to minimize the overall crime rate. As in Persico [2002], the police have different incentives. Specifically, each police officer tries to maximize the number of 'successful' inspections, i.e. where the signal recorded exceeds the threshold set by the adjudicator. As in Persico [2002], we motivate this incentive as driven by the career concerns of individual police officers (who are e.g. promoted if they have many successful arrests etc.).

The timing of the game is therefore: (1) The adjudicator chooses the function β, (2) the police takes this as fixed and chooses inspection intensities $\theta_g \in [0,1]$ for each group subject to the constraint $N_1\theta_1+N_2\theta_2 \le S$ (recall that N_g is the number of agents in group g),[6] and then (3) given the adjudicator's choice β, and inspection probability θ, an agent of group g with crime reward ρ, cost of being found guilty κ and outside option commits a crime if (analogous to (1), but now taking into account also the probability of inspection):

$$\rho - \theta_g\kappa \Pr(q = 1 \mid c, g) \ge \omega - \kappa\theta_g \Pr(q = 1 \mid i, g),$$
$$\text{i.e., whenever } \theta_g\Delta_g \le \frac{\rho-\omega}{\kappa}$$

where Δ_g is as defined in (3). By analogy with (4) an agent of group g commits crime with probability $H_g(\theta_g\Delta_g)$.

As a benchmark, consider a setting where the adjudicator can choose both β and θ. Here the objective function is to minimize the overall crime rate just as in (OPT), and the additional constraint simply reflects that the choice of inspection rule must be feasible, i.e. the total level of inspection cannot exceed total search capacity S. That is to say the adjudicator's problem in this benchmark can be written as:

$$\begin{aligned} &\min_{\{\beta,\theta\}} \sum_g N_g H_g(\theta_g\Delta_g) \\ &\text{s.t. } \sum_g N_g\theta_g \le S. \end{aligned} \tag{5}$$

We refer to the solution of (5) as the **first-best** solution.

Now return to our setting above where the adjudicator chooses β but not θ. The police take β as given and choose θ to maximize the number of successful inspections. Note that an inspection in group g successful with probability $H_g(\theta_g\Delta_g)$. As in Persico [2002], we assume an interior equilibrium solution, i.e., $\theta_g > 0\ \forall g$.[7] Intuitively, in this case, the optimal strategy for the police will equalize the crime rate between the two groups. Otherwise, police that are trying to maximize their successful inspection probability will exclusively search the group with the highest crime rate. Recognizing this, the adjudicator will solve the following

[6] To make this model non trivial, we assume that search capacity is limited, i.e., $S < N_1 + N_2$.

[7] A corner solution entails the police completely ignoring a group. In this case, the setting is trivial because conditional false positive rates are undefined for the ignored group.

problem:

$$\begin{aligned}
&\min_{\beta,\theta} \sum_g N_g H_g(\theta_g \Delta_g) \\
&\text{s.t. } \sum_g N_g \theta_g = S \\
&\qquad H_1(\theta_1 \Delta_1) = H_2(\theta_2 \Delta_2).
\end{aligned} \tag{6}$$

The solution to problem (6) is the **second-best** solution.

We note that because groups will be inspected at different rates, the TPR and FPR as we have defined them should correctly be called the *conditional* true and false positive rates respectively, i.e., the rates conditional on being inspected.[8] Theorem 3.1 now carries over mutatis mutandis, i.e. in both the first and second best solution, the thresholds will be set so as to equalize the conditional false and true positive rates $CFPR$ and $CTPR$. Proofs for this and subsequent theorems are in the appendix.

Theorem 4.1. *The optimal solutions to both the first best (5) and second best (6) equalize the* $CFPR$ *and* $CTPR$ *across groups.*

While the optimal β under the first and second best outcomes coincide, the optimal inspection intensities need not. In particular, the first- and second-best solutions coincide when H is convex. However, when H is concave, then the inspection intensities under the second-best outcome *maximize* the average number of crimes out of all possible search intensities given the optimal signal thresholds.

Theorem 4.2. *Suppose that the* H_g *belong to the same location family, i.e.* $H_g(s) = H(s - \mu_g)$ *for some* μ_g *for each* $i \in \mathcal{G}$ *and that* H *is convex (concave). Then, the inspection intensities in the second best solution minimize (maximize) the crime rate among all thresholds that equalize conditional false positive rates and conditional true positive rates* $CFPR$ *and* $CTPR$*.*

5 Heterogeneous Signal Structure

We now examine the extent to which the conclusion of Theorem 3.1 holds if we allow the signal structure $F_g = (F_g^{\mathrm{c}}, F_g^{\mathrm{i}})$ to be different across groups $g \in \mathcal{G}$. The signal structure F_g and the strategy $\beta_g(s)$ matter to the extent they discourage crime. Recall, that the relevant sufficient statistic of a strategy $\beta_g(s)$ is what we called the *disincentive* to commit crime:

$$\Delta_g = \mathrm{TPR}_g - \mathrm{FPR}_g = \int_{\mathbb{R}} (f^{\mathrm{c}}(s) - f^{\mathrm{i}}(s))\beta_g(s)ds.$$

The set of achievable disincentives given a signal structure is $[\underline{\Delta}_g, \overline{\Delta}_g]$ where

$$\begin{aligned}
\underline{\Delta}_g &= \int_{S_g^-} (f_g^{\mathrm{c}}(s) - f_g^{\mathrm{i}}(s))ds, \text{ where } S_g^- \equiv \{s : f_g^{\mathrm{c}}(s) - f_g^{\mathrm{i}}(s) < 0\}, \\
\overline{\Delta}_g &= \int_{S^+} (f_g^{\mathrm{c}}(s) - f_g^{\mathrm{i}}(s))ds, \text{ where } S_g^+ \equiv \{s : f_g^{\mathrm{c}}(s) - f_g^{\mathrm{i}}(s) > 0\}.
\end{aligned}$$

The relevant sufficient statistics for the signal structure $(F_g^{\mathrm{c}}, F_g^{\mathrm{i}})$ for group g are its minimal and maximal disincentive $\underline{\Delta}_g$ and $\overline{\Delta}_g$ which determines the range of disincentives a classification rule is able to provide.

[8] We observe that these conditional rates are implicitly what has been studied in the fairness in machine learning literature, because these are the rates that can be computed from the data.

5.1 General Analysis

In this section, we give some insight into what happens when the signal structure varies across groups without making further assumptions on how the signal structure varies. First, as should be clear from the intuition previously, what really matters for our results in the baseline model is not that the signal distributions are identical across populations, but rather that the maximal disincentive $\overline{\Delta}_g$ is the same across groups: if we have this, then the basic insight of Theorem 3.1 holds as before (maximizing disincentives across groups). This is summarized in Theorem 5.1. Note that since the signal structures are different, the implication of Theorem 3.1, i.e. that FPR/ FNR will also be equalized across groups, will not hold in general. Further, if the maximal disincentive differs, then in general the result does not hold, as we show in Example 5.2. Theorem 5.3 then provides conditions under which various "natural" fair policies are ranked under the adjudicator's objective to minimize overall crime.

Let us start with analyzing the optimal policy that minimizes average crime. As in Section 3, average crime is minimized by maximizing the disincentive for crime in each group, which is attained by setting $\beta_g(s)$ such that $\Delta_g = \overline{\Delta}_g$ for every g. Unlike in Section 3, the optimal policy does not guarantee any of the fairness notions described in section 2.1 — equalizing disincentives, equalizing false positive rates, equalizing false negative rates or equalizing positive predictive value — when the signal structures differ across the groups.

An immediate observation is that when the signal structures have the same maximal disincentives $\overline{\Delta}_g$, then the optimal effective policy equalizes disincentives.

Theorem 5.1. *Suppose that $\overline{\Delta}_1 = \overline{\Delta}_2$. The adjudicator's optimal policy (i.e. the solution to* (OPT)*) equalizes disincentives* (Δ) *across groups.*

Theorem 5.1 follows from the core insight of Theorem 3.1 that the optimal rule maximizes the disincentive to commit crime for each group. When the signal structures across the groups are identical as in Theorem 3.1, the signal structures have the same maximal disincentives, and therefore, the optimal policy equalizes disincentives. Equalizing disincentives coincides with equalizing false positive rates and equalizing false negative rates in this case. However when the distributions of signals are different, equalizing disincentives need not be be the same as equalizing false positive/ false negative rates. Indeed, equalizing disincentives may yield a strictly lower crime rate than equalizing false positive rates and equalizing false negative rates even when $\overline{\Delta}_1 = \overline{\Delta}_2$. To see this, consider the following example.

Example 5.2. *Suppose that the signal for group g, s_g, is generated according to*

$$s_g = \eta_g + 1_c$$

where η_g is a random variable that has pdf $f_g(\eta)$ that is strictly log-concave and has full support on $\mathbb{R}$, and 1_c is an indicator function that equals 1 if and only if the agent has committed a crime. In words, committing a crime produces a signal that exceeds the signal from not committing a crime by at least 1 on average, while the underlying distribution f_g may differ across the groups. Note that

$$f_g^i(s) = f_g(s) \quad \textit{and} \quad f_g^c(s) = f_g(s-1)$$

and the strict log-concavity guarantees that the strict Monotone Likelihood Ratio Property between f^i and f^c is satisfied, that is, $\frac{f_g^c(s)}{f_g^i(s)}$ strictly increases in s, and that f_g is unimodal. Finally, suppose that $f_1(\eta)$ is asymmetric around its mode, and that $f_2(\eta) = f_1(-\eta)$ is a horizontal reflection of f_1.

For each threshold T_g, the corresponding disincentives satisfy $\Delta_g(T_g) = F_g(T_g) - F_g(T_g - 1)$. The threshold T_g^* maximizes $\Delta_g(T_g)$ if and only if it equalizes the pdfs at T_g^* and $T_g^* - 1$:

$$f_g(T_g^*) = f_g(T_g^* - 1).$$

Graphically, T_g^* and $T_g^* - 1$ are obtained as a pair of intersection points between the pdf f_g and a horizontal line, where the distance between the intersection points has to be 1 as in Figure 1. The maximal disincentive $\overline{\Delta}_g = \Delta_g(T_g^*)$ is the white area under f_g between $T_g^* - 1$ and T_g^*. Since f_2 is merely a horizontal reflection *of f_1, so is the maximal disincentives, and therefore, $\overline{\Delta}_1 = \overline{\Delta}_2$. By Theorem 5.1, the optimal policy T_g^* equalizes disincentives.*

The false positive rate and false negative rate for each group are colored in blue and red. Clearly, $FPR_1 \neq FPR_2$ and $FNR_1 \neq FNR_2$. Consequently, equalizing false positive rates and equalizing false negative rates yield strictly higher crime rates than equalizing disincentives. This also implies $PPV_1 \neq PPV_2$ so that equalizing PPVs also yields a strictly higher crime rate in general.

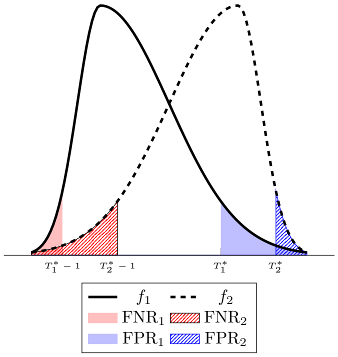


Figure 1

When the signal structures across the groups have different maximal disincentives, we identify conditions under which both equalizing false positive rates and equalizing false negative rates yields a strictly lower crime rate than equalizing disincentives. Without loss of generality, let us assume that $\overline{\Delta}_2 > \overline{\Delta}_1$.

Theorem 5.3. *Let that $\overline{\Delta}_2 > \overline{\Delta}_1$. Then, the following are equivalent:*

1. *The optimal policy subject to equalizing false positive rates ($\beta^\star_{\text{FPR}}$) attains a (weakly) lower crime rate than equalizing disincentives ($\beta^\star_{\Delta}$).*

2. *The optimal policy subject to equalizing false negative rates ($\beta^\star_{\text{FNR}}$) attains a (weakly) crime rate than equalizing disincentives ($\beta^\star_{\Delta}$).*

3. *$(F_2^c)^{-1} \circ F_1^c(T_1^*) > (\geq)(F_2^i)^{-1} \circ F_1^i(T_1^*)$ where T_g^* is the threshold under the optimal policy for group g.*

In general, the optimal policy does not guarantee any of the fairness notions. Theorem 5.3 provides a sufficient and necessary condition under which equalizing false positive rates and equalizing false negative rates attains lower crime rates than equalizing disincentives. However, condition 3 is hard to interpret. Further, it is unclear which of equalizing false positive rate and false negative rate would be better overall for the adjudicator. Without additional structure on the signal structures, it is hard to proceed further. To explore these issues, we restrict attention to signal distributions that are members of location-scale families of distributions.

5.2 Location Scale Families

Definition 5.4. *We say that the signal structure is from a location-scale family if each group's signal is a location-scale transformation of the same underlying random variable η that has absolutely continuous and log-concave density function f with full support on the real line. Specifically, the signal s_g for group g is generated according to:*

$$s_g = \mu_g + \sigma_g \eta + m_g 1_c$$

where μ_g is a location shifter, σ_g is a scale shifter, m_g is the marginal effect of crime on the signal and 1_c is an indicator function that equals 1 if and only if the agent has committed a crime. Equivalently, the *conditional pdfs of signal s for group g conditioning on being innocent and having committed a crime are*

$$f^i(s) = f\left(\frac{s - \mu_g}{\sigma_g}\right) \quad \text{and} \quad f^c(s) = f\left(\frac{s - \mu_g - m_g}{\sigma_g}\right). \tag{7}$$

Note that the underlying distribution f is identical across the groups as in contrast to Example 5.2 where the underlying distribution f_g differed across the groups. Combined with the functional form (7), log-concavity of f is equivalent to the signal structure satisfying the monotone likelihood ratio property for each group g which implies that the optimal β is a threshold strategy. The log-concavity of f also implies that f is unimodal, which guarantees the uniqueness of the threshold that attains the optimal policy. There are many natural location-scale families of distributions satisfying log-concavity including normal distributions, logistic distributions, and extreme value distributions.

A property that makes location-scale family particularly tractable is that the disincentives engendered by a threshold depend only on $\frac{m_g}{\sigma_g}$, i.e. is the ratio between the scale shift σ_g and the marginal effect of crime m_g. For this class of distributions, we can say that it is *always* preferable to equalize either false positive rates or false negative rates compared to equalizing disincentives.

Theorem 5.5. *Suppose the distributions across groups are from the location-scale family as defined in Definition 5.4. Then*

1. *If $\frac{m_1}{\sigma_1} = \frac{m_2}{\sigma_2}$, then the optimal policy equalizes disincentives, false positive rates and negative rates.*
2. *Suppose $\frac{m_1}{\sigma_1} \neq \frac{m_2}{\sigma_2}$, and assume $\frac{m_2}{\sigma_2}$ is larger without loss of generality. Then $\overline{\Delta}_2 > \overline{\Delta}_1$. Further, the optimal policy subject to equalizing false positive rates ($\beta^\star_{\text{FPR}}$) and equalizing false negative rates ($\beta^\star_{\text{FNR}}$) attain strictly lower crime rates than equalizing disincentives ($\beta^\star_{\Delta}$). Further, all three attain strictly higher crime rates than the optimal policy (β^*).*

The formal proof is in the appendix. For some intuition, note that the maximum disincentive $\overline{\Delta}_g$ is determined by and increasing in $\frac{m_g}{\sigma_g}$. Intuitively, this is because the larger the normalized marginal effect of crime on the signal is, the better the adjudicator is able to distinguish between the criminal and the non-criminal based on the signal, and therefore the adjudicator can increase the disincentive to commit crime. If this term is equal across groups, then Theorem 5.1 applies and part (1) follows as a corollary. Now, without loss of generality, suppose instead that $\frac{m_1}{\sigma_1} < \frac{m_2}{\sigma_2}$. Then $\overline{\Delta}_2 > \overline{\Delta}_1$. It can also be verified that the condition (3) in Theorem 5.3 holds, so that equalizing false positive rates and equalizing false negative rates always yield a lower crime rate than equalizing disincentives. Furthermore, we can also verify that equalizing false positive rates and equalizing false negative rates never attains the crime rate under the optimal policy.

A natural question to ask is whether one of equalizing false positive rates or equalizing false negative rates will have a lower crime rate than the other. It is a hard question to answer in general. When the underlying distribution f is symmetric around 0 (0 is without loss of generality since the μ_g's can always be shifted if f is symmetric around some other number), however, the set of feasible disincentives (Δ_1, Δ_2) are identical under equalizing false positive rates and false negative rates, and therefore, the two notions of fairness yield the same crime rate.

Theorem 5.6. *Suppose the signal structure is from the location-scale family as in Definition 5.4, and f is symmetric around 0. Then, the optimal policy subject to equalizing false positive rates and that subject to equalizing false negative rates yield the same crime rate.*

6 Equalizing Crime Rates

In our model, the crime rates are endogenously determined by agents' decisions as a response to the policy implemented. This is unlike most of the fairness literature that assumes that the underlying rates are fixed. It motivates us to study another fairness measure that previous papers could not have asked for: equalizing crime rates.

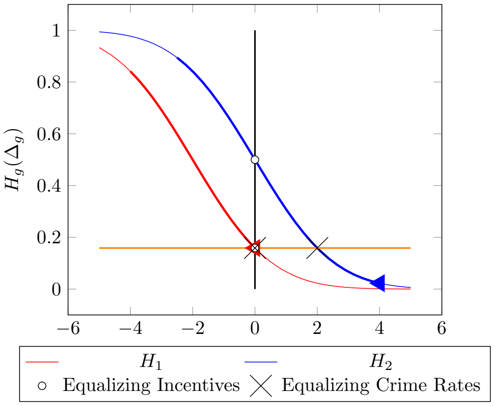


Figure 2: $\overline{\Delta}_1 + \epsilon < \overline{\Delta}_2$

To understand the implications of equalizing crime rates, let us assume that group 2 is 'riskier' than group 1 without loss of generality. Specifically, assume that H_2 stochastically dominates H_1 – that is to assume $H_2(\Delta) \geq H_1(\Delta)\ \forall \Delta$.

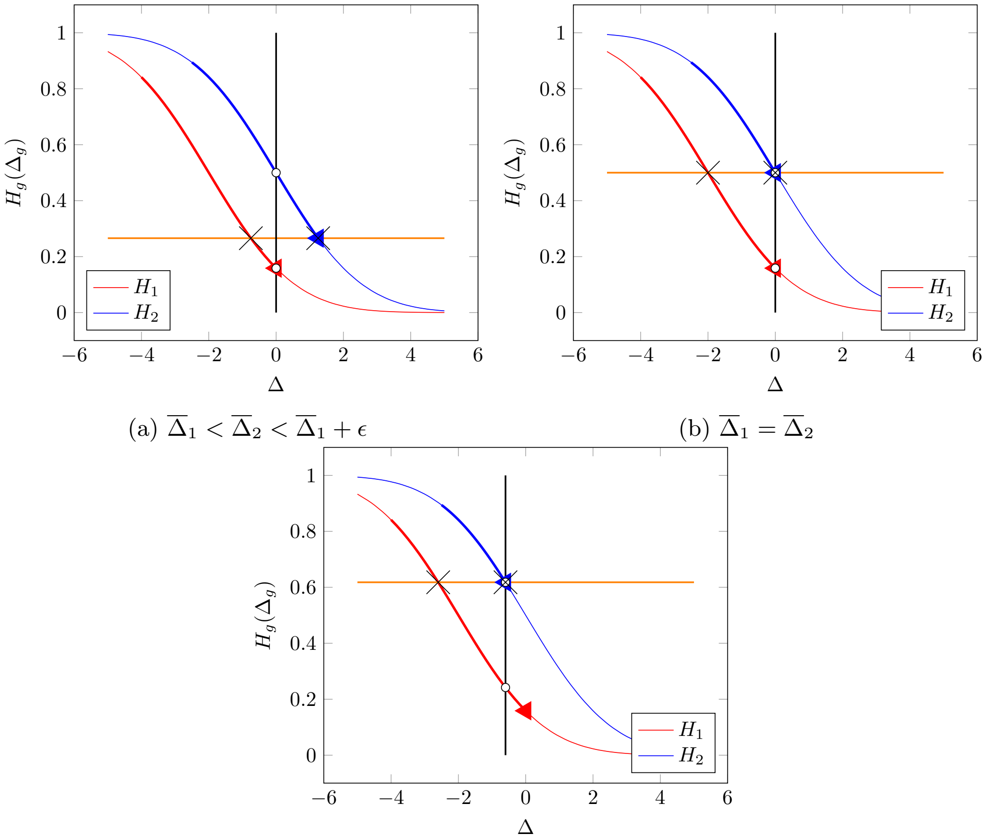
In this section, we focus on the comparison between equalizing crime rates and equalizing disincentives, while allowing any arbitrary signals structures $F_g = (F_g^{\mathrm{c}}, F_g^{\mathrm{i}})$. Equalizing disincentives is an appropriate fairness measure to compare because it is attained by the optimal policy when $\overline{\Delta}_1 = \overline{\Delta}_2$.

The first question to ask is whether equalizing crime rates can ever attain a lower crime rate than equalizing disincentives. We find that equalizing crime rates attains a lower crime rate than equalizing disincentives if and only if $\overline{\Delta}_2$ is sufficiently larger than $\overline{\Delta}_1$.

Theorem 6.1. *Suppose that H_2 first-order stochastically domiantes H_1. Then, there is an $\epsilon > 0$ such that equalizing crime rates attains a lower crime rate than equalizing disincentives if and only if $\overline{\Delta}_2 \geq \overline{\Delta}_1 + \epsilon$.*

The theorem is best demonstrated using the diagrams, although the formal proof is provided in the appendix. For theorem 6.1, we consider 4 cases: (i) $\overline{\Delta}_1 + \epsilon \leq \overline{\Delta}_2$, (ii) $\overline{\Delta}_1 < \overline{\Delta}_2 < \overline{\Delta}_1 + \epsilon$, (iii) $\overline{\Delta}_1 = \overline{\Delta}_2$, and (iv) $\overline{\Delta}_1 > \overline{\Delta}_2$. For initial illustration purposes, we will focus on figure 2 which corresponds to the first case, $\overline{\Delta}_1 + \epsilon < \overline{\Delta}_2$. Each red and blue curve represents $H_1(\cdot)$ and $H_2(\cdot)$, respectively. Note the 'thicker' segments of each curve are the set of crime rates $H_g(\Delta_g)$ that can be achieved by varying the disincentive, $\Delta_g \in [\underline{\Delta}_g, \overline{\Delta}_g]$. For each group, we denote the optimal policy by a triangle. The optimal policy while equalizing disincentives, which is denoted by 'X', is obtained as intersections of the black line and the outside option distribution functions. The optimal policy while equalizing crime rates, which is denoted by 'o', is obtained as intersections of the orange line and the outside option distribution function. In figure 2, note that for group 1, the crime rate stay the same under equalizing disincentives and crime rates, but the crime rate increases once one changes from the optimal policy that equalizes disincentives to the one that equalizes crime rates. Therefore, the optimal policy subject to equalizing crime rates is more preferred than under equalizing disincentives.

Figure 3 shows the other cases. Note that as we decrease $\overline{\Delta}_2$ (or equivalently increase the crime rate of group 2), there comes a point determined by N_1 and N_2 at which equalizing crime rates is more preferred than equalizing disincentives. More specifically, in figure 3a, imagine moving the right most blue triangle to the left and hence raising the orange line; as this happens, both 'X' marks, which denote the optimal policy that equalizes crime rate, need to go up, while the optimal policy that equalizes incentive stays the same. Therefore, depending on the ratio of the number of people (N_1 and N_2), there exists some ϵ such that $\overline{\Delta}_2 \leq \overline{\Delta}_1 + \epsilon$ if and only if equalizing crime rate attains lower crime rate than equalizing crime rates. And it's easy to see from figure 3b and 3c that equalizing disincentives achieves lower crime rate than equalizing crime rates in the corresponding cases. Therefore, these arguments together imply that equalizing crime rates attains a lower crime rate than equalizing disincentives if and only if $\overline{\Delta}_2$ is sufficiently higher than $\overline{\Delta}_1$.



(a) $\overline{\Delta}_1 < \overline{\Delta}_2 < \overline{\Delta}_1 + \epsilon$

(b) $\overline{\Delta}_1 = \overline{\Delta}_2$

(c) $\overline{\Delta}_1 > \overline{\Delta}_2$

Figure 3

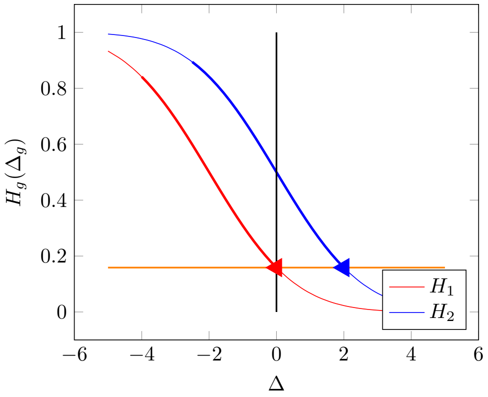


Figure 4: $H_1(\overline{\Delta}_1) = H_2(\overline{\Delta}_2)$

Having verified that equalizing crime rates can attain lower crime rates than equalizing disincentives, the next question is whether equalizing crime rates can ever attain a lower crime rate than any of other fairness notions. The answer to this question is positive which we establish by finding a condition under which the optimal policy attains equalizing crime rates.

Theorem 6.2. *Suppose that H_2 first-order stochastically dominates H_1. When $H_1(\overline{\Delta}_1) = H_2(\overline{\Delta}_2)$, the optimal policy equalizes crime rates but not necessarily false positive rates, false negative rates or disincentives in general.*

Figure 4 depicts the case when $H_1(\overline{\Delta}_1) = H_2(\overline{\Delta}_2)$. By construction, the optimal policy equalizes crime rates. As it can be seen from Figure 4, equalizing disincentives attains a strictly higher crime rates than equalizing crime rates. Furthermore, it can be shown that other notions of fairness — equalizing false positive rates, false negative rates and positive predictive value - are not satisfied in general.

7 Discussion and Conclusions

This paper gives a general model in which classification rules which equalize false positive and false negative rates can be compatible with natural objectives, *in spite of failing to capitalize on statistically relevant information.* We derived the model using the language of criminal justice, but one could just as easily apply the base model to settings in which the principle was making some other binary decision based on partial information, such as a lending or employment decision. The underlying reason is that conditioning on demographic information, while statistically useful, leads to decision rules that incentivize different groups differently — *because demographic information is not under individual control.* Hence, in settings in which the underlying objective depends on the decisions of rational agents, the decision rule should explicitly commit *not* to condition on information that relates to an individual's demographic group, and instead use only information that is affected by the choices of the individual. Abstracting away, the necessary conditions under which our conclusions hold are that:

1. The underlying base rates are rationally responsive to the decision rule deployed by the principle,
2. Signals are observed by the adjudicator at the same rates across populations, and
3. The signals that the adjudicator must use to make her decision are conditionally independent of an individual's group, conditioned on the individual's decision.

Here, conditions (2) and (3) are unlikely to hold precisely in most situations, but we give settings under which they can be relaxed.

More generally, if we are in a setting in which we believe that individual decisions are rationally made in response to the deployed classifier, and yet the deployed classifier does *not* equalize false positive and negative rates, then this is an indication that *either* the deployed classifier is sub-optimal (for the purpose of minimizing base rates), *or* that one of conditions (2) and (3) fails to hold. Since in fairness relevant settings, the failure of conditions (2) and (3) is itself undesirable, this can be a diagnostic to highlight discriminatory conditions earlier in the pipeline than the adjudicator's decision rule. In particular, if conditions (2) or (3) fail to hold, then imposing technical fairness constraints on a deployed classifier may be premature, and instead attention should be focused on structural differences in the observations that are being fed into the deployed classifier.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *Propublica*, 2016. URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

Shamena Anwar and Hanming Fang. An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *American Economic Review*, 96(1):127–151, 2006.

Kenneth J Arrow. Some mathematical models of race discrimination in the labor market. *Racial discrimination in economic life*, pages 187–204, 1972.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2018. `http://www.fairmlbook.org`.

Anna Maria Barry-Jester, Ben Casselman, and Dana Goldstein. The new science of sentencing. *The Marshall Project*, August 8 2015. URL `https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing`. Retrieved 4/28/2016.

Gary S Becker. *The economics of discrimination*. University of Chicago press, 2010.

Nanette Byrnes. Artificial intolerance. *MIT Technology Review*, March 28 2016. URL `https://www.technologyreview.com/s/600996/artificial-intolerance/`. Retrieved 4/28/2016.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*, 2017.

Stephen Coate and Glenn C Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, pages 1220–1240, 1993.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

Hanming Fang and Andrea Moro. Theories of statistical discrimination and affirmative action: A survey. In *Handbook of social economics*, volume 1, pages 133–200. Elsevier, 2011.

Avi Feller, Emma Pierson, Sam Corbett-Davies, and Sharad Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post*, 2016.

Michael F Ferguson and Stephen R Peters. What constitutes evidence of discrimination in lending? *The Journal of Finance*, 50(2):739–748, 1995.

Dean P Foster and Rakesh V Vohra. An economic argument for affirmative action. *Rationality and Society*, 4(2):176–188, 1992.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

Úrsula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1944–1953, 2018.

Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pages 1389–1398, 2018.

Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 240–248, 2019.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2569–2577, 2018.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

John Knowles, Nicola Persico, and Petra Todd. Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109(1):203–229, 2001.

Helen F Ladd. Evidence on discrimination in mortgage lending. *Journal of Economic Perspectives*, 12(2): 41–62, 1998.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6196–6200. AAAI Press, 2019a.

Lydia T Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pages 4051–4060, 2019b.

Lydia T Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. *arXiv preprint arXiv:1910.04123*, 2019c.

Glenn Loury. A dynamic theory of racial income differences. In *Women, minorities, and employment discrimination*, volume 153, pages 86–153. Heath, Lexington, MA, 1977.

Clair C Miller. Can an algorithm hire better than a human? *The New York Times*, June 25 2015. URL `http://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html`. Retrieved 4/28/2016.

Ojmarrh Mitchell and Michael S Caudy. Examining racial disparities in drug arrests. *Justice Quarterly*, 32 (2):288–313, 2015.

Hussein Mouzannar, Mesrob I Ohannessian, and Nathan Srebro. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 359–368, 2019.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown, 2016.

Nicola Persico. Racial profiling, fairness, and effectiveness of policing. *The American Economic Review*, 92 (5):1472–1497, 2002.

Aaron Roth and Michael Kearns. *The Ethical Algorithm.* Oxford University Press, 2019.

Cynthia Rudin. Predictive policing using machine learning to detect patterns of crime. *Wired Magazine*, August 2013. URL `http://www.wired.com/insights/2013/08/predictive-policing-using-machine-learning-to-detect-patterns-of-crime/`. Retrieved 4/28/2016.

Camelia Simoiu, Sam Corbett-Davies, Sharad Goel, et al. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

A Omitted Results and Proofs

Theorem 4.1. *The optimal solutions to both the first best (5) and second best (6) equalize the $CFPR$ and $CTPR$ across groups.*

Proof of Theorem 4.1. We show that the optimal disincentives $\{\Delta^*_g\}_g$ in both cases must be $\{\overline{\Delta}_g\}_g$, which entails equalizing $CFPR$ and $CTPR$.

As for the first best outcome, it's easy to see that given any inspection intensities $\{\theta_g\}_{i \in \mathcal{G}}$, minimizing the overall crime rate corresponds to maximizing $\{\Delta_i\}_i$. Then, it follows that for the optimal set of search intensities and signal thresholds, the optimal solution should in the first best outcome should be such that $\Delta^*_g = \overline{\Delta}_g$ for each g.

Once again, for the second best outcome, given any feasible solution $(\{\theta_g\}_g, \{\Delta_g\}_g)$ to 6, we can show that if there exists g such that Δ_g is not maximized (i.e. $\Delta_g < \overline{\Delta}_g$), then we can always find a new feasible solution $(\{\theta'_g\}_g, \{\Delta'_g\}_g)$ to 6 that sets $\Delta'_g = \overline{\Delta}_g$, while keeping other Δ'_g the same ($\Delta'_{g'} = \Delta_{g'}$) with strictly lower overall crime rate. This shows that the optimal solution to 6 must set $\Delta^*_g = \overline{\Delta}_g$ for each g.

Without loss of generality, assume that $\Delta_1 < \overline{\Delta}_1$ is not maximized. Let's say $\Delta'_1 = \overline{\Delta}_1 = (1+\epsilon)\Delta_1$ for some $\epsilon > 0$ and $\Delta'_2 = \Delta_2$. Now, consider setting a new inspection intensity for group 1 such that $\theta'_1 \in (\frac{1}{1+\epsilon}\theta_1, \theta_1)$ to guarantee that the crime rate in group 1 will be strictly lower than before – that is

$$\theta'_1\Delta'_1 > \theta_1\Delta_1 \quad \Longrightarrow \quad H_1(\theta'_1\Delta'_1) < H_1(\theta_1\Delta_1).$$

Then, because $\sum_g N_g\theta_g = S$, decreasing θ_1 to θ'_1 will require increasing θ_2 to some θ'_2. Then, the crime rate for group 2 will necessarily decrease:

$$\theta'_2\Delta'_2 > \theta_2\Delta_2 \quad \Longrightarrow \quad H_2(\theta'_2\Delta'_2) < H_2(\theta_2\Delta_2).$$

By the continuity of H_g, there exists $\theta'_1 \in (\frac{1}{1+\epsilon}\theta_A, \theta)$ such that $H_1(\theta'_1\beta'_1)) = H_2(\theta'_2\beta_2)$. Note that the crime rate in group 1 and group 2 must have decreased. Therefore, we have found a better feasible solution to 6.

Also, note that this proof can be generalized even when the number of groups is greater than 2. We can aggregate a collection of groups whose Δ_g are not changed to one 'super' group, apply the same argument as above, and use induction over the number of groups. □

Theorem 4.2. *Suppose that the H_g belong to the same location family, i.e. $H_g(s) = H(s - \mu_g)$ for some μ_g for each $i \in \mathcal{G}$ and that H is convex (concave). Then, the inspection intensities in the second best solution minimize (maximize) the crime rate among all thresholds that equalize conditional false positive rates and conditional true positive rates $CFPR$ and $CTPR$.*

Proof of Theorem 4.2. We first provide a high level sketch of the proof. Using the fact that both H_g's are from the same family (i.e. mean shifted), we show that the equilibrium inspection intensities (i.e. the second best solution) set the derivative of the objective value to 0. Now, using the the convexity (concavity) of H, we can show that the second derivative of the overall crime rate will be positive (negative), showing the equilibrium inspection intensities achieve local minima (maxima).

First, we show that if H_g's belong to the same location family, then the derivative of the objective value evaluated at the equilibrium inspection intensities will be 0. Denote the equilibrium inspection intensities by $\{\theta^*_g\}_g$ and the equilibrium disincentives by $\{\Delta^*_g\}_g$. Recall from Theorem 4.1 that (Δ^*_1, Δ^*_2) in both the first and second best solution correspond to $(\overline{\Delta}_1, \overline{\Delta}_2)$.

The equilibrium inspection intensities equalize the crime rates. Hence, we have

$$\begin{aligned} H_1(\theta^*_1\Delta_1) &= H_2(\theta^*_2\Delta_2) \\ \Rightarrow H(\theta^*_1\Delta_1 - \mu_1) &= H(\theta^*_2\Delta_2 - \mu_2) \\ \Rightarrow \theta^*_1\Delta_1 - \mu_1 &= \theta^*_2\Delta_2 - \mu_2 \\ \Rightarrow h(\theta^*_1\Delta_1 - \mu_1) &= h(\theta^*_2\Delta_2 - \mu_2) \\ \Rightarrow h_1(\theta^*_1\Delta_1) &= h_2(\theta^*_2\Delta_2) \end{aligned} \tag{A.1}$$

Replacing $\theta_2 = \frac{S - N_1\theta_1}{N_2}$ and taking the derivative of the overall crime rate with respect to θ_1 yields

$$\begin{aligned}
&N_1\Delta_1 h_1(\theta_1\Delta_1) + N_2 h_2\left(\left(\frac{S - N_1\theta_1}{N_2}\right)\Delta_2\right)\left(-\frac{N_1}{N_2}\Delta_2\right) \\
&= N_1\Delta_1 h_1(\theta_1\Delta_1) - N_1\Delta_2 h_2\left(\theta_2\Delta_2\right)
\end{aligned}$$

Note that by equation A.1 and $\Delta_1^* = \Delta_2^*$, the derivative of the overall crime rate evaluates to 0 under $\{\theta_g^*\}_g$ and $\{\Delta_g^*\}_g$.

Now, in order to determine whether $\{\theta_g^*\}_g$ achieves a local minima or maxima, we calculate the second derivative of the overall crime rate with respect to θ_1:

$$\begin{aligned}
&N_1\Delta_1^2 h_1'(\theta_1\Delta_1) - N_1\Delta_2 h_2'\left(\left(\frac{S - N_1\theta_1}{N_2}\right)\Delta_2\right)\left(-\frac{N_1}{N_2}\Delta_2\right) \\
&= N_1\Delta_1^2 h_1'(\theta_1\Delta_1) + \frac{N_1^2}{N_2}\Delta_2^2 h_2'\left(\left(\frac{S - N_1\theta_1}{N_2}\right)\Delta_2\right)
\end{aligned}$$

By the convexity (concavity) of h, h_1' and h_2' is positive (negative). Therefore, $\{\theta_g^*\}_g$ achieves a local minima (maxima) at the equilibrium inspection intensities. □

Theorem 5.1. *Suppose that $\overline{\Delta}_1 = \overline{\Delta}_2$. The adjudicator's optimal policy (i.e. the solution to* (OPT)*) equalizes disincentives (Δ) across groups.*

Proof of Theorem 5.1. This follows directly from the fact that $(\overline{\Delta}_1, \overline{\Delta}_2)$ is the adjudicator's most optimal policy. □

Theorem 5.3. *Let that $\overline{\Delta}_2 > \overline{\Delta}_1$. Then, the following are equivalent:*

1. *The optimal policy subject to equalizing false positive rates ($\beta^\star_{\mathrm{FPR}}$) attains a (weakly) lower crime rate than equalizing disincentives ($\beta^\star_{\Delta}$).*
2. *The optimal policy subject to equalizing false negative rates ($\beta^\star_{\mathrm{FNR}}$) attains a (weakly) crime rate than equalizing disincentives ($\beta^\star_{\Delta}$).*
3. *$(F_2^c)^{-1} \circ F_1^c(T_1^*) > (\geq)(F_2^i)^{-1} \circ F_1^i(T_1^*)$ where T_g^* is the threshold under the optimal policy for group g.*

Proof of Theorem 5.3. Let us begin with proving the following lemma.

Lemma A.1. *Suppose that $\overline{\Delta}_2 > \overline{\Delta}_1$. Let $\Delta_g^{\mathrm{FPR}}, \Delta_g^{\mathrm{FNR}}$ and Δ_g^{Δ} be the disincentives under the optimal policy subject to each fairness notion* FPR*,* FNR*, and Δ respectively. Suppose further that the optimal policies while equalizing false positive rates, false negative rates and disincentives are threshold policies.*

1. *Equalizing false positive rates attains a (weakly) lower crime rate than equalizing disincentives for all $(h_g)_g$ if and only if $\Delta_g^{\mathrm{FPR}} > (\geq)\Delta_g^{\Delta}\ \forall g$*
2. *Equalizing false negative rates attains a (weakly) lower crime rate than equalizing disincentives for all $(h_g)_g$ if and only if $\Delta_g^{\mathrm{FNR}} > (\geq)\Delta_g^{\Delta}\ \forall g$*

Proof of lemma A.1. By definition, $\Delta_g^{\xi} = \int_{\mathbb{R}} f^{\mathrm{c}}(s)\beta_g^{\xi}(s)ds - \int_{\mathbb{R}} f^{\mathrm{i}}(s)\beta_g^{\xi}(s)ds$ is the disincentive under the optimal policy $\beta_g^{\xi} \in \arg\min_{\beta \in B_\xi} \sum_{g \in \mathcal{G}} N_g \mathrm{CR}_g$ that achieve each fairness notion $\xi \in \{\mathrm{FPR}, \mathrm{FNR}, \Delta\}$

It is straightforward that the $\Delta_g^{\mathrm{FPR}} > (\geq)\Delta_g^{\Delta}\ \forall g$ implies equalizing false positive rates attains a (weakly) lower crime rate than equalizing disincentives. Note that H_g is a non-increasing function, and therefore, for all g,

$$\Delta_g^{\mathrm{FPR}} > (\geq)\Delta_g^{\Delta} \iff N_g H_g(\Delta_g^{\mathrm{FPR}}) < (\leq) N_g H_g(\Delta_g^{\Delta}).$$

The same argument applies for FNR.

Let us now prove the opposite direction: equalizing false positive rates attains a lower crime rate than equalizing disincentives for all g implies that $\Delta_g^{\mathrm{FPR}} > (\geq)\Delta_g^{\Delta}$ for all g. To show this, suppose not. Without loss of generality, suppose that $\Delta_1^{\mathrm{FPR}} < (\leq)\Delta_1^{\Delta}$. If $\Delta_2^{\mathrm{FPR}} \leq \Delta_2^{\Delta}$, then any pair of survivor functions (H_1, H_2) that are strictly decreasing in their arguments will imply $H_1(\Delta_1^{\mathrm{FPR}}) > H_1(\Delta_1^{\Delta})$ and $H_2(\Delta_2^{\mathrm{FPR}}) \geq H_2(\Delta_2^{\Delta})$ so that equalizing disincentives attains a strictly lower crime rate than equalizing false positive rates. If $\Delta_2^{\mathrm{FPR}} > \Delta_2^{\Delta}$, then H_1 where that the difference between its value at Δ_1^{FPR} and at Δ_1^{Δ} is large enough and H_2 where the difference between its value at Δ_2^{FPR} and at Δ_2^{Δ} is small enough result in a lower crime rate for equalizing disincentives than for equalizing false positive rates. More specifically, let H_1 and H_2 be such that

$$N_1(H_1(\Delta_1^{FPR}) - H_1(\Delta_1^{\Delta})) > \epsilon$$

and

$$N_2(H_2(\Delta_2^{\Delta}) - H_2(\Delta_2^{FPR})) < \epsilon$$

for some $\epsilon > 0$. Then,

$$\Big(N_1 H_1(\Delta_1^{FPR}) + N_1 H_2(\Delta_1^{\Delta})\Big) - \Big(N_1 H_1(\Delta_1^{\Delta}) + N_2 H_2(\Delta_2^{\Delta})\Big) > 0$$

which states that equalizing disincentives attain a strictly lower crime rate than equalizing false positive rates. This completes the contradiction desired. Therefore, it has to be the case that $\Delta_g^{\mathrm{FPR}} < (\leq)\Delta_g^{\Delta}$ for all g. The same argument may be applied for equalizing false negative rates.

□

Because we are assuming signal threshold strategies by the adjudicator, we will parametrize $\mathrm{FPR}_g(T_g), \mathrm{FNR}_g(T_g), \mathrm{TPR}_g(T_g$ to denote false positive, false negative, and true positive rate when the signal threshold T_g is used.

We first show the equivalence between condition (1) and (3) in the theorem. Before we show the equivalence, we make some characterization of $\{\Delta_g^{\mathrm{FPR}}\}_g$ and $\{\Delta_g^{\Delta}\}_g$. Since $\overline{\Delta}_2 > \overline{\Delta}_1$, it must be that

$$\Delta_2^{\Delta} = \Delta_1^{\Delta} = \overline{\Delta}_1.$$

As for the first condition (1), by lemma A.1, we have that $\Delta_g^{\mathrm{FPR}} \geq \Delta_g^{\Delta}$ for all g. Thus, we have that for (1),

$$\Delta_1^{\mathrm{FPR}} = \Delta_1^{\Delta} = \Delta_1^{\Delta},$$

which implies

$$T_1^{\mathrm{FPR}} = T_1^{\Delta} = T_1^*.$$

For $\mathrm{FPR}_1 = \mathrm{FPR}_2$, we have

$$\mathrm{FPR}_2(T_2^{\mathrm{FPR}}) = \mathrm{FPR}_1(T_1^{\mathrm{FPR}}) = \mathrm{FPR}_1(T_1^{\Delta}) = F_1^{\mathrm{i}}(T_1^*).$$

or equivalently,

$$F_2^{\mathrm{i}}(T_2^{\mathrm{FPR}}) = F_1^{\mathrm{i}}(T_1^{\mathrm{FPR}}) = F_1^{\mathrm{i}}(T_1^{\Delta}) = F_1^{\mathrm{i}}(T_1^*).$$

Now, we show the equivalence:

$$\begin{aligned}
&\Delta_2^{\mathrm{FPR}} \geq \Delta_2^{\Delta} && \\
&\iff \mathrm{TPR}_2(T_2^{\mathrm{FPR}}) - \mathrm{FPR}_2(T_2^{\mathrm{FPR}}) \geq \mathrm{TPR}_1(T_1^{\Delta}) - \mathrm{FPR}_1(T_1^{\Delta}) && \Delta_2^{\Delta} = \Delta_1^{\Delta} \\
&\iff \mathrm{TPR}_2(T_2^{\mathrm{FPR}}) \geq \mathrm{TPR}_1(T_1^{\Delta}) && \mathrm{FPR}_2(T_2^{\mathrm{FPR}}) = \mathrm{FPR}_1(T_1^{\Delta}) = \mathrm{FNR}_1(T_1^*) \\
&\iff F_1^{\mathrm{c}}(T_1^{\Delta}) \geq F_2^{\mathrm{c}}(T_2^{\mathrm{FPR}}) && \\
&\iff F_1^{\mathrm{c}}(T_1^*) \geq F_2^{\mathrm{c}}(T_2^{\mathrm{FPR}}) && \\
&\iff F_1^{\mathrm{c}}(T_1^*) \geq F_2^{\mathrm{c}}((F_2^{\mathrm{i}})^{-1} \circ F_1^{\mathrm{i}}(T_1^*)) && T_2^{\mathrm{FPR}} = (F_2^{\mathrm{i}})^{-1} \circ F_1^{\mathrm{i}}(T_1^{\mathrm{FPR}}) = (F_2^{\mathrm{i}})^{-1} \circ F_1^{\mathrm{i}}(T_1^*) \\
&\iff (F_2^{\mathrm{c}})^{-1} \circ F_1^{\mathrm{c}}(T_1^*) \geq (F_2^{\mathrm{i}})^{-1} \circ F_1^{\mathrm{i}}(T_1^*) &&
\end{aligned}$$

A similar logic applies to equalizing false negative rates. Once again, by lemma A.1, $\Delta_g^{\mathrm{FNR}} \geq \Delta_g^{\Delta}$ for all g, which implies $\Delta_1^{\mathrm{FNR}} = \Delta_1^{\Delta}$ and hence,

$$T_1^{\mathrm{FNR}} = T_1^{\Delta} = T_1^*.$$

For $\mathrm{FNR}_2 = \mathrm{FNR}_1$, we have

$$\mathrm{TPR}_2(T_2^{\mathrm{FNR}}) = \mathrm{TPR}_1(T_1^{\mathrm{FNR}}) = \mathrm{TPR}_1(T_1^{\Delta}) = \mathrm{TPR}_1(T_1^*)$$

or equivalently,

$$F_2^{\mathrm{c}}(T_2^{\mathrm{FNR}}) = F_1^{\mathrm{c}}(T_1^{\mathrm{FNR}}) = F_1^{\mathrm{c}}(T_1^{\Delta}) = F_1^{\mathrm{c}}(T_1^*).$$

The equivalence follows, as

$$\begin{aligned}
&\Delta_2^{\mathrm{FNR}} \geq \Delta_2^{\Delta} \\
\iff\ & \mathrm{TPR}_2(T_2^{\mathrm{FNR}}) - \mathrm{FPR}_2(T_2^{\mathrm{FNR}}) \geq \mathrm{TPR}_1(T_1^{\Delta}) - \mathrm{FPR}_1(T_1^{\Delta}) && (\Delta_2^{\Delta} = \Delta_1^{\Delta}) \\
\iff\ & -\mathrm{FPR}_2(T_2^{\mathrm{FNR}}) \geq -\mathrm{FPR}_1(T_1^{\Delta}) && (\mathrm{TPR}_2(T_2^{\mathrm{FNR}}) = \mathrm{TPR}_1(T_1^{\Delta}) = \mathrm{TPR}_1(T_1^*)) \\
\iff\ & F_2^{\mathrm{i}}(T_2^{\mathrm{FNR}}) \geq F_1^{\mathrm{i}}(T_1^{\Delta}) \\
\iff\ & F_2^{\mathrm{i}}(T_2^{\mathrm{FNR}}) \geq F_1^{\mathrm{i}}(T_1^*) \\
\iff\ & F_2^{\mathrm{i}}((F_2^{\mathrm{i}})^{-1} \circ F_1^{\mathrm{c}}(T_1^*)) \geq F_1^{\mathrm{i}}(T_1^*) && (T_2^{\mathrm{FNR}} = (F_2^{\mathrm{c}})^{-1} \circ F_1^{\mathrm{c}}(T_1^{\mathrm{FNR}}) = (F_2^{\mathrm{c}})^{-1} \circ F_1^{\mathrm{c}}(T_1^*)) \\
\iff\ & (F_2^{\mathrm{c}})^{-1} \circ F_1^{\mathrm{c}}(T_1^*) \geq (F_2^{\mathrm{i}})^{-1} \circ F_1^{\mathrm{i}}(T_1^*)
\end{aligned}$$

□

Theorem 5.5. *Suppose the distributions across groups are from the location-scale family as defined in Definition 5.4. Then*

1. *If $\frac{m_1}{\sigma_1} = \frac{m_2}{\sigma_2}$, then the optimal policy equalizes disincentives, false positive rates and negative rates.*
2. *Suppose $\frac{m_1}{\sigma_1} \neq \frac{m_2}{\sigma_2}$, and assume $\frac{m_2}{\sigma_2}$ is larger without loss of generality. Then $\overline{\Delta}_2 > \overline{\Delta}_1$. Further, the optimal policy subject to equalizing false positive rates ($\beta^\star_{\mathrm{FPR}}$) and equalizing false negative rates ($\beta^\star_{\mathrm{FNR}}$) attain strictly lower crime rates than equalizing disincentives ($\beta^\star_{\Delta}$). Further, all three attain strictly higher crime rates than the optimal policy (β^*).*

Proof of Theorem 5.5.

Part (1) Let T^* be s.t.

$$f(T^* - r) = f(T^*).$$

Define $T_g = \mu_g + \sigma_g T^*$. Then,

$$\begin{aligned}
-f_g(T_g - m) + f_g(T_g) &= -\frac{1}{\sigma_g} f\Big(\frac{\mu_g + \sigma_g T^* - \mu_g - m}{\sigma_g}\Big) + \frac{1}{\sigma_g} f\Big(\frac{\mu_g + \sigma_g T^* - \mu_g}{\sigma_g}\Big) \\
&= -\frac{1}{\sigma_g} f\Big(T^* - \frac{m_g}{\sigma_g}\Big) + \frac{1}{\sigma_g} f\Big(T^*\Big) \\
&= -\frac{1}{\sigma_g}\Big(f(T^* - r) + f(T^*)\Big) \\
&= 0.
\end{aligned}$$

Therefore, $T_g^* = T_g = \mu_g + \sigma_g T^*$. Note that

$$F_g(T_g) = F(T^*)$$

and

$$F_g(T_g - m_g) = F\Big(T^* - r\Big)$$

for both g. That is, FNR and FPR are the same across the groups.

Part (2) Suppose $\frac{m_1}{\sigma_1} < \frac{m_2}{\sigma_2}$. Then $\overline{\Delta}_2 > \overline{\Delta}_1$.

$$(F_2^{cc})^{-1} \circ F_1^{cc}(T_1) = \Big(\frac{T_1 - \mu_1 - m_1}{\sigma_1}\Big)\sigma_2 + \mu_2 + m_2$$

and

$$(F_2^{nc})^{-1} \circ F_1^{nc}(T_1) = \Big(\frac{T_1 - \mu_1}{\sigma_1}\Big)\sigma_2 + \mu_2$$

so that

$$\begin{aligned}
&(F_2^{cc})^{-1} \circ F_1^{cc}(T_1) \geq (F_2^{nc})^{-1} \circ F_1^{nc}(T_1) \\
\iff &\Big(\frac{T_1 - \mu_1 - m_1}{\sigma_1}\Big)\sigma_2 + \mu_2 + m_2 \geq \Big(\frac{T_1 - \mu_1}{\sigma_1}\Big)\sigma_2 + \mu_2 \\
\iff &\frac{m_2}{\sigma_2} \geq \frac{m_1}{\sigma_1}.
\end{aligned}$$

Therefore, by Theorem 5.3, equalizing false positive rates and equalizing false negative rates attains strictly lower crime rates than equalizing disincentives. □

Theorem 5.6. *Suppose the signal structure is from the location-scale family as in Definition 5.4, and f is symmetric around 0. Then, the optimal policy subject to equalizing false positive rates and that subject to equalizing false negative rates yield the same crime rate.*

Proof of Theorem 5.6. Let (T_1, T_2) be thresholds that equalize false positive rates, that is, $1 - F\Big(\frac{T_1-\mu_1}{\sigma_1}\Big) = 1 - F\Big(\frac{T_2-\mu_2}{\sigma_2}\Big)$. The disincentive for group g is

$$F\Big(\frac{T_g - \mu_g}{\sigma_1}\Big) - F\Big(\frac{T_g - \mu_g - m_g}{\sigma_g}\Big). \tag{A.2}$$

Let $T_g' = 2\mu_g + m_g - T_g$ for each g. Then,

$$\frac{T_g' - \mu_g}{\sigma_g} = -\frac{T_g - \mu_g - m_g}{\sigma_g}$$

and

$$\frac{T_g' - \mu_g - m_g}{\sigma_g} = -\frac{T_g - \mu_g}{\sigma_g}.$$

Note that

$$F\Big(\frac{T_1' - \mu_1 - m_1}{\sigma_1}\Big) = F\Big(-\frac{T_1 - \mu_1}{\sigma_1}\Big) = 1 - F\Big(\frac{T_1 - \mu_1}{\sigma_1}\Big) = 1 - F\Big(\frac{T_2 - \mu_2}{\sigma_2}\Big) = F\Big(-\frac{T_2 - \mu_2}{\sigma_2}\Big) = F\Big(\frac{T_2' - \mu_2 - m_2}{\sigma_2}\Big)$$

where the second and the fourth equalities are from the symmetry around 0, and the third equality is from (T_1, T_2) equalizing false positive rates. Therefore, (T_1', T_2') equalize false negative rates.

Furthermore, the disincentive under T_g' is

$$\begin{aligned}
F\Big(\frac{T_g' - \mu_g}{\sigma_1}\Big) - F\Big(\frac{T_g' - \mu_g - m_g}{\sigma_g}\Big) &= F\Big(-\frac{T_g - \mu_g - m_g}{\sigma_g}\Big) - F\Big(-\frac{T_g - \mu_g}{\sigma_g}\Big) \\
&= \Big(1 - F\Big(\frac{T_g - \mu_g - m_g}{\sigma_g}\Big)\Big) - \Big(1 - F\Big(\frac{T_g - \mu_g}{\sigma_g}\Big)\Big) \\
&= F\Big(\frac{T_g - \mu_g}{\sigma_g}\Big) - F\Big(\frac{T_g - \mu_g - m_g}{\sigma_g}\Big)
\end{aligned}$$

which exactly is the disincentive under T_g. Therefore, for any pair of disincentives that is feasible under equalizing false positive rates, it is feasible under equalizing false negative rates.

Similar arguments can be applied to the other case. Therefore, the set of feasible pair of disincentives are identical, and therefore, the lowest crime rate that can be attained by equalizing false positive rates and equalizing false negative rates are identical.

□

Theorem 6.1. *Suppose that H_2 first-order stochastically domiantes H_1. Then, there is an $\epsilon > 0$ such that equalizing crime rates attains a lower crime rate than equalizing disincentives if and only if $\overline{\Delta}_2 \geq \overline{\Delta}_1 + \epsilon$.*

Proof. We will fix H_1, H_2, and $\overline{\Delta}_1$. Then, we will consider varying $\overline{\Delta}_2$. There are 4 different cases: (i) $\overline{\Delta}_1 + \epsilon \leq \overline{\Delta}_2$, (ii) $\overline{\Delta}_1 < \overline{\Delta}_2 < \overline{\Delta}_1 + \epsilon$, (iii) $\overline{\Delta}_1 = \overline{\Delta}_2$, and (iv) $\overline{\Delta}_1 > \overline{\Delta}_2$, where $\epsilon > 0$ is such that $N_1 \left(H_2(\overline{\Delta}_1 + \epsilon) - H_1(\overline{\Delta}_1)\right) + N_2 \left(H_2(\overline{\Delta}_1 + \epsilon) - H_2(\overline{\Delta}_1)\right) = 0$. Such ϵ exists by the continuity of H_g. When $\epsilon = 0$, then the above value is positive and once ϵ is big enough such that $H_2(\overline{\Delta}_1+\epsilon') = H_1(\overline{\Delta}_1)$, then the value is negative. Therefore, by the intermediate value theorem, such ϵ exists. Furthermore, note that for $\epsilon' > 0$ whenever $H_2(\overline{\Delta}_1+\epsilon') < H_1(\overline{\Delta}_1)$, it must be that $N_2 \left(H_2(\overline{\Delta}_1 + \epsilon) - H_2(\overline{\Delta}_1)\right) + N_1 \left(H_2(\overline{\Delta}_1 + \epsilon) - H_1(\overline{\Delta}_1)\right) < 0$. Therefore, we have that $H_2(\overline{\Delta}_1 + \epsilon) \geq H_1(\overline{\Delta}_1)$.

Also, we write Δ_g^{CR} and Δ_g^{Δ} to denote the optimal disincentives that minimize the crime rates while equalizing the crime rate and disincentive respectively.

Case (i) $\overline{\Delta}_1 + \epsilon \leq \overline{\Delta}_2$

First, because $\overline{\Delta}_1 < \overline{\Delta}_2$, $\Delta_1^{\Delta} = \Delta_2^{\Delta} = \overline{\Delta}_1$. Now, as for Δ_g^{CR}, it depends on whether $H_2(\overline{\Delta}_2) \leq H_1(\overline{\Delta}_1)$. Consider when $H_2(\overline{\Delta}_2) \leq H_1(\overline{\Delta}_1)$. Then, we must have $\Delta_1^{\mathrm{CR}} = \overline{\Delta}_1$, and Δ_2^{CR} should be such that $H_2(\Delta_2^{\mathrm{CR}}) = H_1(\overline{\Delta}_1)$. Because H_2 stochastically dominates H_1, we have that $\Delta_1^{\Delta} = \Delta_2^{\Delta} < \Delta_2^{\mathrm{CR}}$. By the monotonicity of H_2, it must be that $H_2(\Delta_2^{\Delta}) > H_2(\Delta_2^{\mathrm{CR}})$. Therefore,

$$N_1 H_1(\Delta_1^{\Delta}) + N_2 H_2(\Delta_2^{\Delta}) > N_1 H_1(\Delta_1^{\mathrm{CR}}) + N_2 H_2(\Delta_2^{\mathrm{CR}}).$$

Now, consider when $H_2(\overline{\Delta}_2) > H_1(\overline{\Delta}_1)$. Then, we must have $\Delta_2^{\mathrm{CR}} = \overline{\Delta}_2$, and Δ_1^{CR} should be such that $H_1(\Delta_1^{\mathrm{CR}}) = H_2(\overline{\Delta}_2)$. Compare how each group's crime rate changes as we go from equalizing crime rate to equalizing disincentive. As for group 1, it goes from $H_2(\overline{\Delta}_2)$ to $H_1(\overline{\Delta}_1)$. As for group 2, it goes from $H_2(\overline{\Delta}_2)$ to $H_2(\overline{\Delta}_1)$. Therefore, total change in crime rate by going from equalizing crime rate to equalizing disincentive is at most 0:

$$\begin{aligned}
0 &= N_1 \left(H_2(\overline{\Delta}_1 + \epsilon) - H_1(\overline{\Delta}_1)\right) + N_2 \left(H_2(\overline{\Delta}_1 + \epsilon) - H_2(\overline{\Delta}_1)\right) \\
&\geq N_1 \left(H_2(\overline{\Delta}_2) - H_1(\overline{\Delta}_1)\right) + N_2 \left(H_2(\overline{\Delta}_2) - H_2(\overline{\Delta}_1)\right)
\end{aligned}$$

Therefore, we have that equalizing crime rates is better than equalizing disincentives.

Case (ii) $\overline{\Delta}_1 < \overline{\Delta}_2 < \overline{\Delta}_1 + \epsilon$

In this case, we know that $H_2(\overline{\Delta}_2) > H_1(\overline{\Delta}_1)$. For the optimal disincentive-equalizing policies, we have that $\Delta_1^{\Delta} = \Delta_2^{\Delta} = \overline{\Delta}_1$. As for crime-equalizing policy, we have $\Delta_2^{\mathrm{CR}} = \overline{\Delta}_2$, and Δ_1^{CR} is chosen such that $H_1(\Delta_1^{\mathrm{CR}}) = H_2(\overline{\Delta}_2)$. Now, compare how each group's crime rate as we goes from equalizing crime rates to equalizing disincentives. As for group 2, it goes from $H_2(\overline{\Delta}_2)$ to $H_2(\overline{\Delta}_1)$. As for group 1, it goes from $H_2(\overline{\Delta}_2)$ to $H_1(\overline{\Delta}_1)$. Therefore, total change in crime rate by going from equalizing crime rate to equalizing disincentive is at most 0, as

$$\begin{aligned}
&N_2 \left(H_2(\overline{\Delta}_2) - H_2(\overline{\Delta}_1)\right) + N_1 \left(H_2(\overline{\Delta}_2) - H_1(\overline{\Delta}_1)\right) \\
&> N_2 \left(H_2(\overline{\Delta}_1 + \epsilon) - H_2(\overline{\Delta}_1)\right) + N_1 \left(H_2(\overline{\Delta}_1 + \epsilon) - H_1(\overline{\Delta}_1)\right) \\
&= 0
\end{aligned}$$

Therefore, equalizing disincentives is better than equalizing crime rates in this case.

Case (iii) $\overline{\Delta}_1 = \overline{\Delta}_2$

First, $\Delta_1^{\Delta} = \Delta_2^{\Delta} = \overline{\Delta}_1 = \overline{\Delta}_2$. As for equalizing crime rates Δ_g^{CR}, $\Delta_2^{\mathrm{CR}} = \overline{\Delta}_2$, and Δ_1^{CR} is chosen such that $H_1(\Delta_1^{\mathrm{CR}}) = H_2(\Delta_2^{\mathrm{CR}}) > H_1(\overline{\Delta}_1)$. Therefore, we have

$$N_1 H_1(\Delta_1^{\Delta}) + N_2 H_2(\Delta_2^{\Delta}) < N_1 H_1(\Delta_1^{\mathrm{CR}}) + N_2 H_2(\Delta_2^{\mathrm{CR}}),$$

meaning equalizing disincentives is better than equalizing crime rates.

Case (iv) $\overline{\Delta}_1 > \overline{\Delta}_2$

First, $\Delta_1^{\Delta} = \Delta_2^{\Delta} = \overline{\Delta}_2$. As for equalizing crime rates Δ_g^{CR}, $\Delta_2^{\mathrm{CR}} = \overline{\Delta}_2$, and Δ_1^{CR} is chosen such that $H_1(\Delta_1^{\mathrm{CR}}) = H_2(\Delta_2^{\mathrm{CR}}) > H_1(\overline{\Delta}_1)$. Therefore, we have

$$N_1 H_1(\Delta_1^{\Delta}) + N_2 H_2(\Delta_2^{\Delta}) < N_1 H_1(\Delta_1^{\mathrm{CR}}) + N_2 H_2(\Delta_2^{\mathrm{CR}}),$$

meaning equalizing disincentives is better than equalizing crime rates. □

Theorem 6.2. *Suppose that H_2 first-order stochastically dominates H_1. When $H_1(\overline{\Delta}_1) = H_2(\overline{\Delta}_2)$, the optimal policy equalizes crime rates but not necessarily false positive rates, false negative rates or disincentives in general.*

Proof. $(\overline{\Delta}_1, \overline{\Delta}_2)$ is the most optimal policy, and they equalize the crime rates $H_1(\overline{\Delta}_1) = H_2(\overline{\Delta}_2)$. However, it is not guaranteed that the false positive/negative rates or disincentives will be equalized. For instance, if $\overline{\Delta}_1 \neq \overline{\Delta}_2$, then the disincentives are not equalized. And as $\Delta_g = (1 - \mathrm{FNR}_g) - \mathrm{FPR}_g$, false positive/negative rates won't be equalized in general. □