# Rethinking Algorithmic Fairness for Human-AI Collaboration

HAOSEN GE, *WHARTON SCHOOL, UNIVERSITY OF PENNSYLVANIA*
HAMSA BASTANI, *WHARTON SCHOOL, UNIVERSITY OF PENNSYLVANIA*
OSBERT BASTANI, *UNIVERSITY OF PENNSYLVANIA*

**Abstract:** Existing approaches to algorithmic fairness aim to ensure equitable outcomes *if* human decision-makers comply perfectly with algorithmic decisions. However, perfect compliance with the algorithm is rarely a reality or even a desirable outcome in human-AI collaboration. Yet, recent studies have shown that selective compliance with fair algorithms can *amplify* discrimination relative to the prior human policy. As a consequence, ensuring equitable outcomes requires fundamentally different algorithmic design principles that ensure robustness to the decision-maker's (a priori unknown) compliance pattern. We define the notion of *compliance-robustly* fair algorithmic recommendations that are guaranteed to (weakly) improve fairness in decisions, regardless of the human's compliance pattern. We propose a simple optimization strategy to identify the best performance-improving compliance-robustly fair policy. However, we show that it may be infeasible to design algorithmic recommendations that are simultaneously fair in isolation, compliance-robustly fair, and more accurate than the human policy; thus, if our goal is to improve the equity and accuracy of human-AI collaboration, it may not be desirable to enforce traditional algorithmic fairness constraints. We illustrate the value of our approach on criminal sentencing data before and after the introduction of an algorithmic risk assessment tool in Virginia.

**Significance Statement:** Decision-makers attempt to enforce fair outcomes from human-AI collaboration by imposing constraints on the algorithm's decisions alone – this implicitly assumes that people follow AI recommendations perfectly, but in reality, people selectively choose when to comply. This selective compliance can unintentionally worsen inequities, even when the AI itself is designed to be fair. Our research introduces a new *user-aware* approach to fairness, focusing on algorithms that improve equity regardless of how much people follow their advice. We show how to design such "compliance-robust" algorithms while still improving performance. However, we find that achieving traditional algorithmic fairness, compliance-robustness, and improved performance simultaneously may not always be possible. This work highlights the need to rethink algorithmic fairness standards for AI systems when their goal is to support human decision-making.

**Classification:** Computer Sciences

**Keywords:** Algorithmic Fairness, Human-AI Collaboration, Selective Compliance, Compliance-Robust Fairness, Fairness Trade-Offs

## 1 INTRODUCTION

As machine learning algorithms are increasingly deployed in high-stakes settings (e.g., healthcare, finance, justice), it has become imperative to understand the fairness implications of algorithmic decision-making on protected groups. As a consequence, a wealth of work has sought to define algorithmic fairness [Chen et al., 2023, Corbett-Davies and Goel, 2018, Dwork et al., 2012, Hardt et al., 2016, Kleinberg et al., 2016] and learn machine learning-based policies that satisfy fairness constraints to ensure equitable outcomes across protected groups [Bastani et al., 2022, Basu, 2023, Joseph et al., 2016, Kearns et al., 2019, Kim et al., 2019].

Most of this work focuses on whether the algorithm makes fair decisions *in isolation*. Yet, these algorithms are rarely used in high-stakes settings without human oversight, since there are still considerable legal and regulatory challenges to full automation. Moreover, many believe that human-AI collaboration is superior to full automation because human experts may have auxiliary information that can help correct the mistakes of algorithms, producing better decisions than the human or algorithm alone. For example, while many powerful AI systems have been developed for diagnosing medical images [Esteva et al., 2017], the Center for Medicare and Medicaid Services only allows AI systems to *assist* medical experts with diagnosis [Rajpurkar et al., 2022].

However, human-AI collaboration introduces new complexities — the overall outcomes now depend not only on the algorithmic recommendations, but also on the subset of individuals for whom the human decision-maker complies with the algorithmic recommendation. Recent case studies have shown mixed results on whether human-AI collaboration actually improves decision accuracy [Ahn et al., 2024, Campero et al., 2022] or fairness [Van Dam, 2019]. For instance, a recent experiment examines diagnostic quality when radiologists are assisted by AI models [Agarwal et al., 2023]. The authors find that, although the AI models are substantially more accurate than radiologists, access to AI assistance does not improve diagnostic quality on average; the authors show that this is due to *selective compliance* of the algorithmic recommendations by humans, which they hypothesize is driven by improper Bayesian updating. Similarly, a recent study evaluates the impact of algorithmic risk assessment on judges' sentencing decisions in Virginia courts [Stevenson and Doleac, 2024, Van Dam, 2019]. Although risk assessment promised fairer outcomes [Kleinberg et al., 2018], the authors find that it brought no detectable benefits in terms of public safety or reduced incarceration; in fact, racial disparities *increased* in the subset of courts where risk assessment appears most influential. Once again, the mismatch is driven by selective compliance to algorithmic recommendations, which appears to be at least partly driven by conflicting objectives between judges and the algorithm (e.g., judges are more lenient towards younger defendants). Selective compliance has significant fairness implications, e.g., in this case, the authors note that "judges were more likely to sentence leniently for white defendants with high-risk scores than for black defendants with the same score." These case studies make it clear that ensuring equitable outcomes in human-AI collaboration requires accounting for humans' complex and unexpected compliance patterns. To this end, Gillis et al. [2021], Morgan and Pass [2019] show that the outcomes of human-AI collaboration can be arbitrarily less fair than either those of the human alone or the algorithm alone.

To resolve this state of affairs, we introduce the notion of *compliance-robust* algorithms — i.e., algorithmic decision policies that are guaranteed to (weakly) improve fairness in final outcomes, regardless of the human's (unknown) compliance pattern. In particular, given a human decision-maker and her policy (without access to AI assistance), we characterize the class of algorithmic recommendations that never result in collaborative final outcomes that are less fair than the pre-existing human policy, even if the decision-maker's compliance pattern is adversarial. Next, we prove that there exists considerable tension between traditional algorithmic fairness and compliance-robust fairness. Unless the true data-generating process is itself perfectly fair, it can be infeasible to design an algorithmic policy that is fair in isolation, compliance-robustly fair, and more accurate than the human-only policy, implying that compliance-robust fairness imposes fundamentally different constraints compared to traditional fairness. This raises the question of whether traditional fairness is even a desirable constraint to enforce for human-AI collaboration—if the goal is to improve fairness and accuracy in human-AI collaboration outcomes, it may be preferable to design an algorithmic policy that is accurate and compliance-robustly fair, but not necessarily fair in isolation.

Lastly, we use Virginia court sentencing data—leveraging variation from the introduction of an algorithmic risk assessment tool in 2002—to simulate the performance and fairness of compliance-robust policies versus natural baseline policies. We find that a compliance-robust policy performs favorably, both in terms of performance and fairness, across all 193 judges who exhibit very different compliance behaviors.

## 2 PROBLEM FORMULATION

We now introduce some notation to formalize the human-AI decision-making problem, and definitions of traditional fairness, compliance-robust fairness, and performance.

Consider a decision-making problem where each individual is associated with a type $x \in \mathcal{X} = [k] = \{1, ..., k\}$ (e.g., education, prior defaults), a protected attribute $a \in \mathcal{A} = \{0, 1\}$ (e.g., gender), and a true outcome $y \in \mathcal{Y} = \{0, 1\}$ (e.g., whether they can repay a loan).[1] Let $\mathbb{P}(x, a, y)$ over $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ denote the joint distribution of the types, protected attributes, and outcomes of individuals for whom decisions must be made (we will not require knowledge of $\mathbb{P}(x, a, y)$ to construct policies). We make the following assumption that our population has good "coverage" across variables/outcomes:

Assumption 1. *We have $\mathbb{P}(x, a, y) > 0$ for all $x \in \mathcal{X}$, $a \in \mathcal{A}$, and $y \in \mathcal{Y}$.*

[1] Note $\mathcal{X}$ can encompass multiple categorical features that are "flattened" into a single dimension. For mathematical simplicity, we restrict to categorical features (i.e., $\mathcal{X}$ has finite possible values), a binary protected attribute, and a binary outcome. Our results straightforwardly extend to protected attributes with multiple classes, but allowing for continuous features and outcomes requires modifying the primary fairness definition we use [Hardt et al., 2016] by introducing slack variables.

*Definition 2.1.* A *decision-making policy* is a mapping $\pi : \mathcal{X} \times \mathcal{A} \to [0, 1]$ that maps each feature-attribute pair to a probability $\pi(x, a)$; then, the decision is $\hat{y} \sim \text{Bernoulli}(\pi(x, a))$.[2]

The algorithm designer's goal is for the decision to equal the true outcome — i.e., $\hat{y} = y$ (e.g., we would ideally give each individual a loan if and only if they will repay the loan).[3] We consider a human decision-maker represented by a policy $\pi_H$ (without access to algorithmic assistance). When given access to recommendations from an algorithmic policy $\pi_A$, the human instead makes decisions according to a *compliance function* $c : \mathcal{X} \times \mathcal{A} \mapsto \{0, 1\}$, where $c(x, a) = 1$ indicates that the human adopts the algorithmic decision for individuals $(x, a)$. Then, the joint human-AI policy will be

$$\pi_C(x, a) = \begin{cases} \pi_A(x, a) & \text{if } c(x, a) = 1 \\ \pi_H(x, a) & \text{otherwise.} \end{cases}$$

Note that $\pi_H$ can be estimated via supervised learning on historical decision-making data in the absence of algorithmic recommendations; in contrast, the compliance function $c$ cannot be learned until an algorithm $\pi_A$ has already been deployed, potentially with poor consequences. Thus, we assume knowledge of $\pi_H$, but study compliance-robust policies that do not require any knowledge of $c$. We will show that our results extend starightforwardly to more complex compliance functions that are random or depend on the algorithmic recommendation itself (e.g., a human is more likely to comply when the algorithmic recommendation $\pi_A(x, a)$ is similar to their own judgment $\pi_H(x, a)$).

*Fairness.* We primarily analyze the well-studied notion of "equality of opportunity" [Hardt et al., 2016, Kleinberg et al., 2016], which requires that, for any chosen decision policy $\pi$, the true positive rates for each protected group should be equal:

$$\mathbb{P}[\hat{y} = 1 \mid y = 1, a = 0] = \mathbb{P}[\hat{y} = 1 \mid y = 1, a = 1].$$

In other words, on average, deserving individuals ($y = 1$) should have the same likelihood of access to the intervention ($\hat{y} = 1$) regardless of their protected group status ($a \in \{0, 1\}$). In Appendix A.4, we show that the qualitative challenges that we illustrate in this paper arise for a very general class of fairness definitions, subsuming demographic parity [Calders et al., 2009, Zliobaite, 2015] and equalized odds [Chen et al., 2023, Hardt et al., 2016].

For a policy $\pi$, we then marginalize out the types $x$ to obtain the average score for subgroup $a$ as

$$\overline{\pi}(a) = \sum_{x \in \mathcal{X}} \pi(x, a) \mathbb{P}(x \mid a, y = 1).$$

Traditional algorithmic fairness would require the algorithmic policy $\pi_A$ to satisfy $\overline{\pi}_A(0) = \overline{\pi}_A(1)$, without accounting for the human policy $\pi_H$ or the compliance function $c$.

Next, without loss of generality, we assume Group 1 is better off than Group 0 in terms of "opportunity" under the human-alone policy:

[2]We use the commonly employed Bernoulli distribution, but it suffices for $\mathbb{P}[\hat{y} = 1]$ to simply be increasing in $\pi(x, a)$.

[3]Note that the human's objective may vary, e.g., judges are more lenient towards younger defendants [Van Dam, 2019], but algorithm designers are typically restricted to predicting outcomes observed in the training data.

Assumption 2. *We have* $\overline{\pi}_H(1) \geq \overline{\pi}_H(0)$.

We now introduce some definitions. Let $\alpha$ be the slack in group fairness for a policy $\pi$:

$$\alpha(\pi) = |\overline{\pi}(1) - \overline{\pi}(0)|.$$

*Definition 2.2.* We say an algorithmic policy $\pi_A$ *reduces fairness* under compliance function $c$ if the resulting human-AI policy $\pi_C$ satisfies $\alpha(\pi_C) > \alpha(\pi_H)$.

Note that a human decision-maker can always choose to ignore all algorithmic advice (i.e., $c(x, a) = 0$ for all $x$ and $a$) resulting in the human's policy ($\pi_C = \pi_H$)—then, if $\pi_H$ is unfair, no choice of $\pi_A$ can guarantee a fair $\pi_C$. Thus, when designing $\pi_A$, we can at most demand that we do not reduce unfairness relative to the existing human policy $\pi_H$.

*Definition 2.3.* Given $\pi_H$, an algorithmic policy $\pi_A$ is *compliance-robustly fair* if there does not exist *any* compliance function $c$ that reduces fairness for $\pi_A$.

Let $\Pi_{\text{fair}}$ be the set of compliance-robustly fair policies; note that these policies need not be fair in the traditional algorithmic fairness sense. We will characterize $\Pi_{\text{fair}}$ in the next section.

*Performance.* Algorithmic assistance often aims to not only improve fairness but also the *accuracy* of decisions. Ideally, we would produce compliance-robustly fair recommendations that improve performance relative to the human policy. To define performance, we consider a loss function $\ell : [0, 1] \times \mathcal{Y} \to \mathbb{R}$, and define the expected loss

$$L(\pi) = \mathbb{E}[\ell(\pi(x, a), y)].$$

Let the performance-maximizing (but possibly unfair) optimal policy be

$$\pi_* = \operatorname*{argmin}_{\pi} L(\pi),$$

and the highest performing compliance-robustly fair policy be

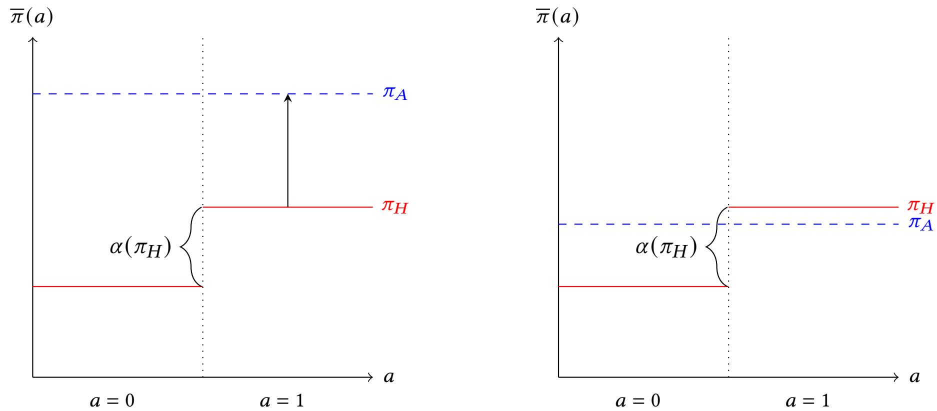
$$\pi_0 = \operatorname*{arg\,min}_{\pi \in \Pi_{\text{fair}}} L(\pi).$$

For analysis, we impose the following mild assumption on our loss:

*Definition 2.4.* We say a policy $\pi'$ has *higher deviation* than a second policy $\pi$ if for all $x \in \mathcal{X}, a \in \mathcal{A}$, if $\pi(x, a) \geq \pi_*(x, a)$, then $\pi'(x, a) \geq \pi(x, a)$, and if $\pi(x, a) \leq \pi_*(x, a)$, then $\pi'(x, a) \leq \pi(x, a)$. We say the deviation is *strictly higher* if the inequality is strict for any $x \in \mathcal{X}, a \in \mathcal{A}$.

Assumption 3. *For any policies* $\pi, \pi'$*, if* $\pi'$ *has higher deviation than* $\pi$*, then* $L(\pi') \geq L(\pi)$*; furthermore, if the deviation is strictly higher, then* $L(\pi') > L(\pi)$.

In other words, if $\pi'$ always deviates farther from $\pi_*$ than $\pi$ (i.e., for every $x$ and $a$), then $\pi'$ has higher expected loss. It can be easily checked that common loss functions (e.g., mean squared error, mean absolute error, cross entropy) satisfy the above definition. It is worth noting that we do not assume the loss is symmetric — i.e., if $\pi$ and $\pi'$ are on different sides of $\pi_*$ for any $x, a$ pair, this assumption does not say anything about which one attains a lower loss.




(a) Fairness reduced if $c(x, 0) = 0$ and $c(x, 1) = 1$

(b) Compliance-robust fairness

Fig. 1. Examples with trivial individual types (i.e., $\mathcal{X} = \{1\}$) with the same human policy $\pi_H$ and two different algorithmic policies $\pi_A$. The human policy is unfair ($\overline{\pi}_H(0) \neq \overline{\pi}_H(1)$), but the algorithmic policy in both cases is fair in isolation ($\overline{\pi}_A(0) = \overline{\pi}_A(1)$). Left: If the human selectively complies when $a = 1$, fairness is reduced (relative to $\pi_H$). Right: Fairness is never reduced for any compliance $c$, i.e., $\pi_A$ is compliance-robustly fair.

## 3 CHARACTERIZATION OF COMPLIANCE-ROBUST FAIRNESS

Our first main result characterizes the class of compliance-robust policies $\Pi_{\text{fair}}$. For intuition, consider the simple example depicted in Figure 1, where there are no types (i.e., $\mathcal{X} = \{1\}$). The left and right panels consider the same unfair human policy $\pi_H$ (i.e., $\overline{\pi}_H(0) \neq \overline{\pi}_H(1)$) but two different traditionally fair algorithmic policies $\pi_A$ (i.e., $\overline{\pi}_A(0) = \overline{\pi}_A(1)$). On the left, if the human selectively complies when $a = 1$, fairness reduces relative to $\pi_H$, i.e., $\pi_A$ is not compliance-robustly fair although it is fair in isolation. On the right, it is easy to check that no compliance function reduces fairness, i.e., $\pi_A$ is compliance-robustly fair. In general, as we formalize next, compliance-robustness holds exactly when $\pi_A$ is "sandwiched" between $\pi_H(x, 0)$ and $\pi_H(x, 1)$.

THEOREM 3.1. *Given $\pi_H$, an algorithmic policy $\pi_A$ is compliance-robustly fair if and only if*

$$\alpha(\pi_A) \leq \alpha(\pi_H) \tag{1}$$

$$\pi_H(x, 0) \leq \pi_A(x, 0) \qquad (\forall x \in \mathcal{X}) \tag{2}$$

$$\pi_A(x, 1) \leq \pi_H(x, 1) \qquad (\forall x \in \mathcal{X}). \tag{3}$$

We give a proof in Appendix A.1. In general, it is easier to find compliance-robustly fair policies when the human policy is rather unfair. In fact, the following corollary (proof in Appendix A.1) shows that when the human policy is perfectly fair, there are no nontrivial compliance-robust policies. This is because any deviation by the human that unequally affects the two classes $a \in \mathcal{A}$ provably results in unfairness.

Corollary 3.2. *If $\alpha(\pi_H) = 0$, then $\pi_A$ is compliance-robustly fair if and only if $\pi_A(x, a) = \pi_H(x, a)$ for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$.*

Theorem 3.1 allows us to write down a simple optimization problem (see Algorithm 1) to compute a compliance-robustly fair policy $\pi_0$ that performs the best (i.e., minimizes the loss $L$). Note that we have assumed nothing on the class of compliance functions $c(x, a)$ thus far. If we have a priori knowledge that $c$ only depends on some subset of the type variables, then we can trivially remove the constraints in Algorithm 1 corresponding to those types, enlarging the class of compliance-robust policies $\Pi_{\text{fair}}$.

**Algorithm 1:** Compliance-robustly Fair Algorithm.

1: **Input:** the human policy $\pi_H$.

2: Solve the following optimization problem:

$$\begin{aligned} \pi_0 = \quad & \arg\min_{\pi} L(\pi) \\ \text{subj. to} \quad & \alpha(\pi) \leq \alpha(\pi_H), \\ & \pi_H(x, 0) \leq \pi(x, 0), \quad \forall x \in \mathcal{X}, \\ & \pi(x, 1) \leq \pi_H(x, 1), \quad \forall x \in \mathcal{X}. \end{aligned}$$

3: **Return:** policy $\pi_0$.

*Extensions.* As noted earlier, it can be easily shown that Theorem 3.1 holds for a more general class of compliance functions. First, in practice, compliance may depend on the output of $\pi_A$—e.g., the human decision-maker may comply only when the recommendation $\pi_A(x, a)$ is sufficiently close to their own judgment $\pi_H(x, a)$. In particular, consider a policy-dependent compliance function which depends not only on the type and protected attribute, but also on the output of $\pi_A$, i.e., $\tilde{c} : \mathcal{X} \times \mathcal{A} \times [0, 1] \mapsto \{0, 1\}$. Then, since $\pi_A$ is itself a function of $x$ and $a$, there exists some compliance function from our original class such that

$$c(x, a) = \tilde{c}(x, a, \pi_A(x, a)).$$

Thus, Theorem 3.1 automatically subsumes this case.

Second, our results hold when the compliance function is random rather than deterministic. Specifically, consider a random compliance function of the form:

$$c_p(x, a) = \begin{cases} 1 & \text{with probability } p(x, a) \\ 0 & \text{otherwise,} \end{cases}$$

yielding the joint human-AI policy

$$\pi_C(x, a) = \pi_A(x, a) \cdot p(x, a) + \pi_H(x, a) \cdot (1 - p(x, a)).$$

The proof of Theorem 3.1 works without modification for this class.

Another issue that arises in practice is that we often do not directly observe $\pi_H$; rather, one must estimate $\hat{\pi}_H \approx \pi_H$ using supervised learning on historical data prior to the algorithmic intervention (we illustrate this on court sentencing data in Section 6.1). When using $\hat{\pi}_H$ instead of $\pi_H$, our compliance-robustness guarantee gracefully degrades in the estimation error of $\hat{\pi}_H$ as follows. In particular, suppose that we have an estimate $\hat{\pi}_H$ of $\pi_H$ satisfying $|\hat{\pi}_H(x,a) - \pi_H(x,a)| \leq \epsilon$ for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$. If we run our algorithm using $\hat{\pi}_H$, then we obtain an algorithmic policy $\pi_A$ that is compliance-robustly fair for $\hat{\pi}_H$. Now, let $\hat{\pi}_C$ be the joint policy combining $\pi_A$ and $\hat{\pi}_H$, and let $\pi_C$ be the joint policy combining $\pi_A$ and $\pi_H$. Note that for any compliance function $c$, we have $|\hat{\pi}_C(x,a) - \pi_C(x,a)| \leq \epsilon$, from which it follows that $\alpha(\pi_C) \leq \alpha(\hat{\pi}_C) + 2\epsilon$. As a consequence, we have

$$\alpha(\pi_C) \leq \alpha(\hat{\pi}_C) + 2\epsilon \leq \alpha(\hat{\pi}_H) + 2\epsilon \leq \alpha(\pi_H) + 4\epsilon,$$

where the second inequality follows from compliance-robust fairness of $\pi_A$ to $\hat{\pi}_H$, and the third follows since $\alpha(\hat{\pi}_H) \leq \alpha(\pi_H) + 2\epsilon$ by our assumption on the estimation error of $\hat{\pi}_H$. Thus, $\hat{\pi}_A$ satisfies a compliance-robust fairness guarantee within a slack of $4\epsilon$.

## 4 PERFORMANCE OF COMPLIANCE-ROBUSTLY FAIR POLICIES

We have so far established the best-performing compliance-robust policy $\pi_0$ (defined in Algorithm 1) as a strong candidate for algorithmic advice. However, in most cases, we would only provide algorithmic advice if we think it may perform better than the current human policy, i.e., $L(\pi_0) < L(\pi_H)$. In this section, we provide simple conditions (that can be easily verified with knowledge of $\pi_H$ and the performance-maximizing optimal $\pi_*$) to see if algorithmic advice is desirable.

Given the human policy $\pi_H$ and the optimal policy $\pi_*$, let

$$u(a) = \{x \in \mathcal{X} \mid \pi_H(x,a) \geq \pi_*(x,a)\}$$
$$\ell(a) = \{x \in \mathcal{X} \mid \pi_H(x,a) < \pi_*(x,a)\}.$$

Intuitively, $u(a)$ denotes regions within group $a$ where the human policy $\pi_H$ assigns weakly higher scores than the optimal policy $\pi_*$. Conversely, $\ell(a)$ corresponds to the regions where $\pi_H$ assigns scores that are strictly lower than those assigned by $\pi_*$.

We now construct the following policy $\pi_B$, which attempts to bridge between achieving high performance and ensuring compliance-robustness with respect to $\pi_H$:

$$\pi_B(x,a) = \begin{cases} \pi_H(x,a) & \text{if } x \in u(0) \cup \ell(1) \\ \pi_*(x,a) & \text{otherwise.} \end{cases}$$

This policy attempts to maximize performance by matching $\pi_*$ while satisfying constraints (2)-(3) in Theorem 3.1 to ensure compliance-robustness. To provide some intuition, $\ell(1)$ represents the regions where $\pi_H$ assigns lower scores than $\pi_*$ for the advantageous group. As stated in Assumption 3, we can improve the performance of $\pi_H$ by increasing $\pi_H$'s scores in $\ell(1)$. However, Theorem 3.1 shows that compliance-robustly fair policies cannot return higher scores for any type within the advantageous group. Therefore, the best-performing compliance-robustly fair policies must be the same as $\pi_H$ in $\ell(1)$. The same argument holds for $u(0)$.

The policy $\pi_B$ will be pivotal for us to understand the performance of compliance-robustly fair policies, as well as their relationship to traditional fairness (in the next section). First, $\pi_B$ is compliance-robustly fair if it (in isolation) does not reduce fairness relative to $\pi_H$ (Lemma A.2 in Appendix A.2). Consequently, $\pi_B$ provides a constructive upper bound on the performance of any compliance-robustly fair policy (Lemma A.3), which will be useful for examining when the performance of $\pi_0$ exceeds that of $\pi_H$. Intuitively, if there exists a compliance-robustly fair policy that can improve human performance, then $\pi_B$ must be more accurate than $\pi_H$.

Our next result, Theorem 4.1, shows simple conditions under which the optimal compliance-robustly fair policy $\pi_0$ is worth sharing with the human (i.e., when $L(\pi_0) < L(\pi_H)$). Namely, we require that human policy $\pi_H$ is not perfectly fair (in which case, there is no nontrivial compliance-robustly fair policy by Corollary 3.2), and $\pi_H$ deviates from the performance-optimal policy $\pi_*$ in a direction that we can plausibly correct with algorithmic advice.

Theorem 4.1. *Assume that $\alpha(\pi_H) \neq 0$, and that either $\pi_H(x,1) \neq \pi_*(x,1)$ for some $x \in u(1)$ or $\pi_H(x,0) \neq \pi_*(x,0)$ for some $x \in \ell(0)$. Then, we have $L(\pi_0) < L(\pi_H)$.*

We give a proof in Appendix A.2. As discussed earlier, $\pi_B$ must equal $\pi_H$ in $u(0)$ and $\ell(1)$. Consequently, compliance-robustly fair policies can only perform better than $\pi_H$ in $\ell(0)$ and $u(1)$. As long as the human policy $\pi_H$ doesn't perfectly match the optimal policy $\pi_*$ in at least one of these regions, we can construct a compliance-robustly fair policy that achieves strictly better performance than $\pi_H$.

## 5 COMPLIANCE-ROBUST FAIRNESS VS. TRADITIONAL FAIRNESS

As shown in the last section, compliance-robust fairness and performance improvement are often compatible; the same holds for traditional fairness and performance improvement [Hardt et al., 2016]. However, we will show that there is considerable tension between maintaining *both* types of fairness (compliance-robust fairness and traditional algorithmic fairness) while improving performance.

Building on the mild conditions required for a performance-improving compliance-robust policy in Theorem 4.1, the next lemma establishes additional conditions that are necessary and sufficient to find a policy $\pi_A$ that is also traditionally fair (i.e., $\alpha(\pi_A) = 0$).

Lemma 5.1. *Assume that $\alpha(\pi_H) \neq 0$, and that either $\pi_H(x,1) \neq \pi_*(x,1)$ for some $x \in u(1)$ or $\pi_H(x,0) \neq \pi_*(x,0)$ for some $x \in \ell(0)$. Then, there exists a compliance-robustly fair policy $\pi_A$ that is also traditionally fair ($\alpha(\pi_A) = 0$) and performance-improving ($L(\pi_A) < L(\pi_H)$) if and only if there exists a policy $\pi$ satisfying*

$$\overline{\pi}(1) \leq \overline{\pi}(0) \tag{4}$$

$$\pi(x,1) \leq \pi_B(x,1) \qquad (\forall x \in \mathcal{X}) \tag{5}$$

$$\pi(x,0) \geq \pi_B(x,0) \qquad (\forall x \in \mathcal{X}) \tag{6}$$

$$L(\pi) < L(\pi_H). \tag{7}$$

We give a proof in Appendix A.3. Next, we show a natural setting where we meet the above conditions — namely, when the data-generating process is such that the optimal performance-maximizing policy $\pi_*$ is already perfectly fair without any added constraints (i.e., $\alpha(\pi_*) = 0$).

Theorem 5.2. *Assume that $\alpha(\pi_H) \neq 0$, and that either $\pi_H(x,1) \neq \pi_*(x,1)$ for some $x \in u(1)$ or $\pi_H(x,0) \neq \pi_*(x,0)$ for some $x \in \ell(0)$. Then, if $\pi_*$ is fair, there is always a compliance-robustly fair $\pi_A \in \Pi_{fair}$ that is also traditionally fair and performance-improving.*

Proof. Consider $\pi_B$. Since $\pi_*$ is fair and $\pi_B(x,1) \leq \pi_*(x,1)$ and $\pi_B(x,0) \geq \pi_*(x,0)$, it immediately follows that $\overline{\pi}_B(1) \leq \overline{\pi}_B(0)$. Thus, the claim follows from Lemma 5.1 (with $\pi = \pi_B$). □

Unfortunately, it unlikely that an unconstrained performance-maximizing policy will be inherently fair; this insight has been the driving force of the algorithmic fairness literature. Rather, we may have to choose between the properties of compliance-robust fairness (to avoid disparate harm relative to the human policy), performance improvement (to ensure that algorithmic recommendations actually drive improved decisions), and traditional fairness (to ensure the algorithm is fair in isolation). To this end, we now construct a simple setting where we can only satisfy one criterion — performance improvement *or* traditional fairness — for all compliance-robustly fair policies.

Intuitively, this tension can arise when the human policy is not far from the performance-maximizing policy ($\pi_H \approx \pi_*$) and this policy is quite unfair ($\alpha(\pi_*) \gg 0$). Consider the extreme case where $\pi_H = \pi_*$ and $\alpha(\pi_H) > 0$. By Theorem 3.1, $\pi_*$ is compliance-robustly fair, and yet it is not traditionally fair. Thus, any traditionally fair policy must necessarily perform worse than the existing $\pi_H$ or not be compliance-robustly fair. The following proposition crystallizes this intuition in a nontrivial setting.

Proposition 5.3. *There exists $\mathcal{X}$, $\mathbb{P}$, $L$, and $\pi_H$ satisfying $\alpha(\pi_H) \neq 0$ and $\pi_H \neq \pi^*$ such that for any policy $\pi$, $\pi$ cannot simultaneously satisfy all of the following: (i) $\pi \in \Pi_{fair}$, (ii) $\alpha(\pi) = 0$, and (iii) $L(\pi) \leq L(\pi_H)$.*

We give a proof in Appendix A.3. Given these results, if the goal is to improve fairness and accuracy in human-AI collaboration outcomes, it may be preferable to design an algorithmic policy that is accurate and compliance-robustly fair, but not fair in isolation.

One may question whether the challenges arising from selective compliance and the resulting trade-offs are only relevant to our fairness definition — equality of opportunity [Hardt et al., 2016]. Therefore, we show in Appendix A.4 that selective compliance can lead to undesirable outcomes for a large class of fairness definitions that satisfies a mild assumption.

## 6 EMPIRICAL EVALUATION

We empirically simulate the performance of our compliance-robustly fair algorithm using criminal sentencing data from Virginia from 2000 to 2004. In July 2002, the Virginia Criminal Sentencing Commission (VCSC) introduced an algorithmic risk assessment tool to help judges identify *low-risk* individuals with a felony conviction, with the goal of diverting them from prison. Before making final sentencing decisions, judges were presented with the model's predicted risk score to facilitate risk assessment. We leverage data pre- and post- introduction of the risk assessment tool to assess the fairness and performance of different algorithmic advice policies.

### 6.1 Experimental Setup

Following Stevenson and Doleac [2024], we obtained criminal sentencing records through a Freedom of Information Act request, and defendant demographics from https://virginiacourtdata.org/. This data spans 22, 871 sentencing events made by 229 different judges (see Appendix A.5). We use data prior to the launch of the tool ($N = 15, 108$) to estimate each judge's policy $\pi_H$, and data from 2003 and 2004 ($N = 7, 763$) for evaluation.

The true outcome $y$ denotes whether a defendant recidivates within three years following release.[4] All decision-making policies, including $\pi_H$ and $\pi_A$, generate risk scores representing the estimated probability of recidivism for each defendant, based on the observed defendant features. Defendants with lower predicted risk scores are more likely to receive reduced sentences. The protected attribute $a$ is race, restricted to White or Black.

*Estimating the judges' policies.* To construct $\pi_H$, we need the judges' *independent* assessment of whether a defendant should be offered a reduced sentence. We estimate this by examining when judges overrode pre-existing VCSC sentencing guidelines to reduce a defendant's sentence (see Appendix A.5 for details), prior to the introduction of the algorithmic risk assessment tool. We train a gradient boosted decision tree [Ke et al., 2017] to predict reduced sentences based on observed defendant covariates as well as the the Judge ID (to obtain judge-specific policies).

*Estimating the judges' compliance functions.* After the introduction of the algorithmic risk assessment tool, compliance with the tool's recommendations is an observed variable in the data. We estimate a judge's compliance function $c$ by training a gradient boosted decision tree to predict compliance using the same set of defendant covariates as above.

*Estimating the original risk assessment model.* We do not have access to the original VCSC risk assessment tool, but we observe the tool's recommendations (i.e., low-risk or not). Thus, we train a gradient boosted decision tree to predict the tool's policy $\pi_A^{\text{actual}}$, using the same set of defendant covariates as above (except for Judge ID).

*Policies.* We then construct different human-AI collaborative policies for each judge using our estimates of judge-specific $\pi_H$, $c$ and $\pi_A$—(1) the actual observed policy $\pi_C^{\text{actual}}(x, a)$, (2) our compliance-robustly fair policy $\pi_C^{\text{robust}}(x, a)$, (3) the performance-maximizing policy $\pi_C^*(x, a)$, and (4) the traditionally fair policy $\pi_C^{\text{trad fair}}(x, a)$. Note that these are all human-AI policies and may not satisfy the properties guaranteed by their respective algorithmic policies $\pi_A$ alone. When simulating the performance and fairness of $\pi_A^{\text{robust}}$, $\pi_A^*$ and $\pi_A^{\text{trad fair}}$, we make a key assumption that judges' compliance functions would remain the same for these alternative algorithmic risk assessment tools as in the original VCSC algorithmic risk assessment tool. This may not be the case in practice, but our compliance-robust approach guarantees hold under *any* new compliance function that judges may adopt.

[4]In practice, we must also address the issue of *selective labels*—we only observe the true outcome when a defendant is released [Lakkaraju et al., 2017]. We set aside this issue for the purpose of the simulation exercise.

*Metrics.* For a human-AI policy $\pi_C$, we examine both performance improvement, $L(\pi_H) - L(\pi_C)$, and fairness improvement, $\alpha(\pi_H) - \alpha(\pi_C)$; details in Appendix A.5.

### 6.2 Results

Figure 2 shows a judge-level comparison of each of the four human-AI policies (relative to the $\pi_H$) in terms of performance and fairness. First, as discussed in the findings of Stevenson and Doleac [2024], we observe that the actual VCSC reduced performance (Fig 2a) and fairness (Fig 2b), relative to the prior human-alone policy, for nearly every judge. In contrast, our compliance-robust policy $\pi_C^{\text{robust}}$ benefits almost every judge in terms of both performance (Fig 2c) and fairness (Fig 2d). Only 2 of 193 judges have a negligible deterioration in fairness, likely due to finite sample estimation error. Then, as expected, the policy $\pi_C^*$ that relies on a performance maximizing algorithm significantly improves performance (Fig 2e), but comes at the cost of 34% of judges see deterioration in fairness in their sentencing outcomes (Fig 2f). Finally, we consider the policy $\pi_C^{\text{trad fair}}$ that relies on the highest-performing traditionally fair algorithm—we find that while it improves accuracy (Fig 2g) and fairness (Fig 2h) *on average*, 20% of judges see reduced performance and 13% of judges see deterioration in fairness due to selective compliance. In contrast, our compliance-robust approach guarantees weakly improved performance and fairness for *every* judge, regardless of their compliance pattern.

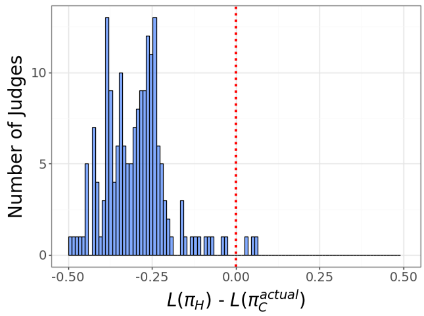
*Mechanism.* As illustrated in Figure 1, algorithmic recommendations can reduce fairness when decision-makers disproportionately comply with the algorithmic recommendations for an advantaged group whenever the algorithm offers a more favorable decision. To shed more light, we examine the compliance pattern $c_{\text{problem}}$ and human-alone policies $\pi_H^{\text{Ave}}$ for the subset of judges that worsen fairness the most (see Appendix A.5 for details). We define the variable "AI Low Risk" for a defendant $i$ with features $(x_i, a_i)$ as the indicator function of whether the algorithmic policy is more lenient than the human alone policy, $\pi_A(x_i, a_i) > \pi_H(x_i, a_i)$. Then, we test if judges comply more frequently for White defendants when the algorithmic policy is more lenient:

$$\begin{aligned} P(\text{Comply}_i = 1) = \text{Logistic}(&\beta_0 + \beta_1 \cdot \text{White}_i + \beta_2 \cdot \text{AI Low Risk}_i + \\ &\beta_3 \cdot (\text{AI Low Risk}_i \times \text{White}_i) + \epsilon_i). \end{aligned}$$

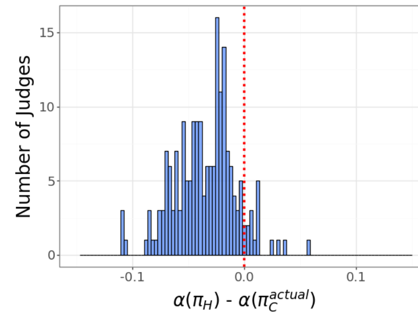
Indeed, we find that $\beta_3$ is positive and statistically significant for all human-AI collaborative policies except our compliance-robust policy, indicating that judges' compliance behaviors exacerbate existing racial biases under these policies. In contrast, our compliance-robustly fair policy ($\pi_A^{\text{robust}}$) effectively guards against such problematic compliance behaviors.

## 7 CONCLUSION

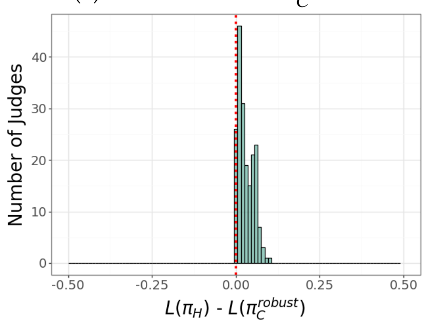
This paper illustrates the perils of selective compliance for equitable outcomes in human-AI collaboration. In particular, even algorithms that satisfy traditional algorithmic fairness criteria can amplify unfairness in decisions (relative to the human making decisions in isolation). Unfortunately, a human decision-maker's compliance pattern is a priori unknown, and may even change over time, affecting fairness in outcomes. Therefore, we introduce the concept of compliance-robust fairness and demonstrate how to derive algorithmic policies that weakly improve fairness regardless of the human's compliance pattern. Naturally, it is also important that the algorithmic advice achieves better performance than the human alone. We show that, as long as the human policy is slightly sub-optimal and not perfectly fair, the best performance-improving compliance-robust policy still generates improvements over the human in isolation. However, it is not always the case that we can also achieve the third property of traditional fairness — we may need to rely on algorithmic policies that are unfair in isolation to achieve compliance-robustly fair human-AI collaboration. We illustrate our approach on criminal sentencing data from Virginia. We demonstrate significant gains in fairness compared to a traditionally fair policy that does not account for judges' selective compliance patterns. Our findings contribute to the design of human-AI collaboration systems that are "user-aware," enhancing rather than diminishing fairness in collaborative decisions.



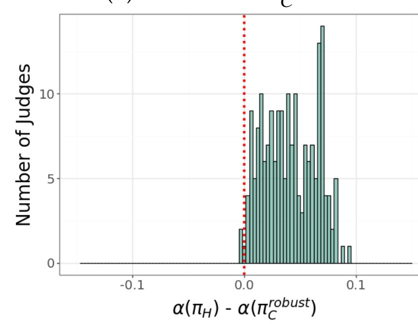

(a) Performance of $\pi_C^{actual}$



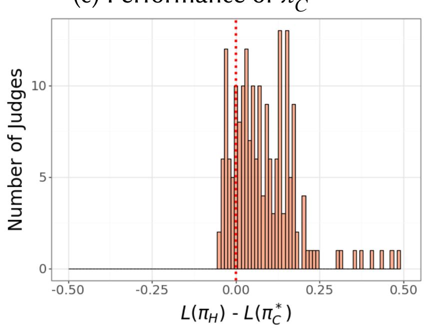

(b) Fairness of $\pi_C^{actual}$



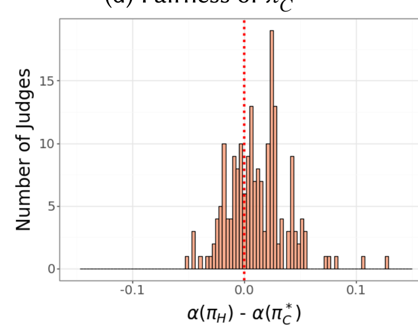

(c) Performance of $\pi_C^{robust}$



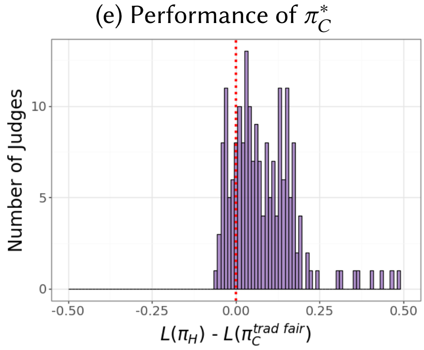

(d) Fairness of $\pi_C^{robust}$



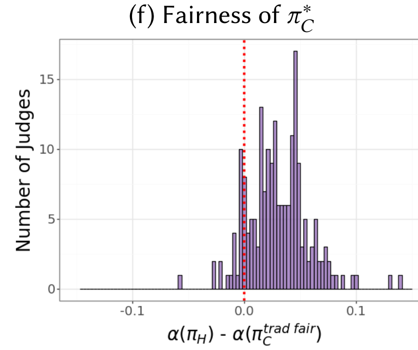

(e) Performance of $\pi_C^*$




(f) Fairness of $\pi_C^*$




(g) Performance of $\pi_C^{\text{trad fair}}$




(h) Fairness of $\pi_C^{\text{trad fair}}$

Fig. 2. We show the performance and fairness comparisons for $\pi_C^{\text{actual}}$, $\pi_C^{\text{robust}}$, $\pi_C^*$ and $\pi_C^{\text{trad fair}}$ across the 193 judges in our evaluation sample. Bars to the right of the red dotted line correspond to judges whose accuracy or fairness improve with the algorithmic recommendation.

## REFERENCES


Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. *Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology*. Technical Report. National Bureau of Economic Research.

Daehwan Ahn, Abdullah Almaatouq, Monisha Gulabani, and Kartik Hosanagar. 2024. Impact of Model Interpretability and Outcome Feedback on Trust in AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–25.

Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. 2022. Practical adversarial multivalid conformal prediction. *Advances in Neural Information Processing Systems* 35 (2022), 29362–29373.

Anirban Basu. 2023. Use of race in clinical algorithms. *Science Advances* 9, 21 (2023), eadd2704.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*. IEEE, 13–18.

Andres Campero, Michelle Vaccaro, Jaeyoon Song, Haoran Wen, Abdullah Almaatouq, and Thomas W Malone. 2022. A test for evaluating performance in human-computer systems. *arXiv preprint arXiv:2206.12390* (2022).

Richard J Chen, Judy J Wang, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. 2023. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering* 7, 6 (2023), 719–742.

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

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118.

Talia Gillis, Bryce McLaughlin, and Jann Spiess. 2021. On the fairness of machine-assisted human decisions. *arXiv preprint arXiv:2110.15310* (2021).

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems* 29 (2016).

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 100–109.

Michael P Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 247–254.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

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

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 275–284.

Andrew Morgan and Rafael Pass. 2019. Paradoxes in fair computer-aided decision making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 85–90.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. 2022. AI in health and medicine. *Nature medicine* 28, 1 (2022), 31–38.

Megan T Stevenson and Jennifer L Doleac. 2024. Algorithmic risk assessment in the hands of humans. *American Economic Journal, Economic Policy* (2024).
A Van Dam. 2019. Algorithms were supposed to make Virginia judges fairer. What happened was far more complicated. *Available via The Washington Post. Retrieved February* 18 (2019), 2020.
Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. 2023. Fairlearn: Assessing and Improving Fairness of AI Systems. , 8 pages. http://jmlr.org/papers/v24/23-0389.html
Indre Zliobaite. 2015. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723* (2015).

# A APPENDIX: THEORETICAL RESULTS

## A.1 Proof of Results in Section 3

THEOREM 3.1. *Given $\pi_H$, an algorithmic policy $\pi_A$ is compliance-robustly fair if and only if*

$$\alpha(\pi_A) \leq \alpha(\pi_H) \tag{1}$$

$$\pi_H(x,0) \leq \pi_A(x,0) \qquad (\forall x \in \mathcal{X}) \tag{2}$$

$$\pi_A(x,1) \leq \pi_H(x,1) \qquad (\forall x \in \mathcal{X}). \tag{3}$$

PROOF. First, we show that (1), (2), and (3) are sufficient. Note that (2) and (3) imply

$$\overline{\pi}_H(0) \leq \overline{\pi}_C(0) \leq \overline{\pi}_A(0) \tag{8}$$

$$\overline{\pi}_A(1) \leq \overline{\pi}_C(1) \leq \overline{\pi}_H(1), \tag{9}$$

respectively, for any compliance function $c$. Now, we have

$$\overline{\pi}_C(1) - \overline{\pi}_C(0) \leq \overline{\pi}_H(1) - \overline{\pi}_H(0) \leq \alpha(\pi_H),$$

where the first inequality follows from (8) and (9). Additionally, we have

$$\overline{\pi}_C(0) - \overline{\pi}_C(1) \leq \overline{\pi}_A(0) - \overline{\pi}_A(1) \leq \alpha(\pi_A) \leq \alpha(\pi_H).$$

where the first inequality follows from (8) and (9), and the third from (1). The claim follows.

Next, we show that (1), (2), and (3) are necessary. Note that (1) is clearly necessary, or the compliance function $c(x,a) = 1$ for all $x, a$ (i.e., the human always complies with the algorithmic decision) reduces fairness. To see that (2) is necessary, suppose to the contrary that $\pi_H(0, x_0) > \pi_A(0, x_0)$ for some $x \in \mathcal{X}$. Then, consider the compliance function

$$c(x,a) = \begin{cases} 1 & \text{if } x = x_0, a = 0 \\ 0 & \text{otherwise.} \end{cases}$$

For this $c$, it is easy to see that by Assumption 1, $\overline{\pi}_H(0) > \overline{\pi}_C(0)$, whereas $\overline{\pi}_C(1) = \overline{\pi}_H(1)$. By Assumption 2, it follows that $\alpha(\pi_C) > \alpha(\pi_H)$, so $c$ reduces fairness. The proof for (3) is similar. □

COROLLARY 3.2. *If $\alpha(\pi_H) = 0$, then $\pi_A$ is compliance-robustly fair if and only if $\pi_A(x,a) = \pi_H(x,a)$ for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$.*

PROOF. Consider a compliance-robustly fair policy $\pi_A$, and assume to the contrary that $\pi_A(x_0, a_0) \neq \pi_H(x_0, a_0)$ for some $x_0 \in \mathcal{X}$ and $a_0 \in \mathcal{A}$. We assume that $a_0 = 1$; the case $a_0 = 0$ is similar. By Theorem 3.1, we have $\pi_A(x,1) \leq \pi_H(x,1)$ for all $x \in \mathcal{X}$, so $\pi_A(x_0, 1) < \pi_H(x_0, 1)$. Then, by Assumption 1, we have $\overline{\pi}_A(1) < \overline{\pi}_H(1)$. Also by Theorem 3.1, we have $\pi_A(x, 0) \geq \pi_H(x, 0)$ for all $x \in \mathcal{X}$, so $\overline{\pi}_A(0) \geq \overline{\pi}_H(0)$. Thus, we have

$$\overline{\pi}_A(1) < \overline{\pi}_H(1) = \overline{\pi}_H(0) \leq \overline{\pi}_A(0),$$

where the equality holds by our assumption that $\alpha(\pi_H) = 0$. Since $\overline{\pi}_A(1) \neq \overline{\pi}_A(0)$, we must have $\alpha(\pi_A) > 0 = \alpha(\pi_H)$, so by Theorem 3.1, $\overline{\pi}_A$ is not compliance-robustly fair, a contradiction. □

## A.2 Proof of Results in Section 4

To prove Theorem 4.1, we need the following lemmas. It follows by the construction of $\pi_B$ that it satisfies:

Lemma A.1. *We have $\pi_B(x, 0) \geq \pi_H(x, 0)$ and $\pi_B(x, 1) \leq \pi_H(x, 1)$ for all $x \in \mathcal{X}$.*

As we will see shortly, $\pi_B$ provides a constructive upper bound on the performance of any compliance-robustly fair policy, which will be useful for examining when the performance of $\pi_0$ exceeds that of $\pi_H$. We begin by noting that $\pi_B$ itself is compliance-robustly fair if it (in isolation) does not reduce fairness relative to $\pi_H$.

Lemma A.2. *If $\alpha(\pi_B) \leq \alpha(\pi_H)$, then $\pi_B$ is compliance-robustly fair.*

Proof. By Lemma A.1, $\pi_B$ satisfies conditions (2) and (3) in Theorem 3.1 by construction. If $\alpha(\pi_B) \leq \alpha(\pi_H)$, then (1) also holds, so by Theorem 3.1, $\pi_B$ is compliance-robustly fair. □

Furthermore, the next result shows that $\pi_B$ performs at least as well as the optimal compliance-robustly fair policy $\pi_0$.

Lemma A.3. *We have $L(\pi_0) \geq L(\pi_B)$.*

Proof. It suffices to prove that $\pi_0$ has higher deviation than $\pi_B$, in which case the claim follows by Assumption 3. We need to show that for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, we have

$$\begin{cases} \pi_0(x, a) \leq \pi_B(x, a) & \text{if } \pi_B(x, a) \leq \pi_*(x, a) \\ \pi_0(x, a) \geq \pi_B(x, a) & \text{if } \pi_B(x, a) \geq \pi_*(x, a). \end{cases} \tag{10}$$

Now, consider a point $x \in u(0)$; in this case, we have

$$\pi_0(x, 0) \geq \pi_H(x, 0) \geq \pi_*(x, 0),$$

where the first inequality follows since $\pi_0$ is compliance-robustly fair so it satisfies (2), and the second since $x \in u(0)$. Since $\pi_B(x, 0) = \pi_H(x, 0)$ for $x \in u(0)$, (10) holds. Next, consider a point $x \in \ell(1)$; in this case, we have

$$\pi_0(x, 1) \leq \pi_H(x, 1) < \pi_*(x, 1)$$

where the first inequality follows since $\pi_0$ satisfies (3), and the second since $x \in \ell(1)$. Since $\pi_B(x, 1) = \pi_H(x, 1)$ for $x \in \ell(1)$, (10) holds. Finally, if $x \notin u(0) \cup \ell(1)$, then $\pi_B(x, a) = \pi_*(x, a)$ for all $a \in \mathcal{A}$, so (10) holds. Thus, $\pi_0$ has higher deviation than $\pi_B$, so the claim follows. □

Now, we prove Theorem 4.1.

Theorem 4.1. *Assume that $\alpha(\pi_H) \neq 0$, and that either $\pi_H(x,1) \neq \pi_*(x,1)$ for some $x \in u(1)$ or $\pi_H(x,0) \neq \pi_*(x,0)$ for some $x \in \ell(0)$. Then, we have $L(\pi_0) < L(\pi_H)$.*

Proof. If $\alpha(\pi_B) \leq \alpha(\pi_H)$, then by Lemma A.2, $\pi_B$ is compliance-robustly fair; the assumptions in the theorem statement clearly imply that $L(\pi_B) < L(\pi_H)$, so the claim follows. Otherwise, we must have $\alpha(\pi_B) > \alpha(\pi_H)$. Furthermore, Lemma A.1 implies that $\overline{\pi}_B(1) \leq \overline{\pi}_H(1)$ and $\overline{\pi}_B(0) \geq \overline{\pi}_H(0)$. Together with Assumption 2, these three conditions imply that

$$\overline{\pi}_B(1) < \overline{\pi}_B(0).$$

Intuitively, this might happen when the optimal policy satisfies $\overline{\pi}_*(1) < \overline{\pi}_*(0)$, but the human policy reverses this relationship. To compensate, we can reduce the performance of $\overline{\pi}_B$ to "shrink" the gap between $\overline{\pi}_B(1)$ and $\overline{\pi}_B(0)$. In particular, consider scaling the decisions as follows:

$$\pi_{A,\lambda}(x,a) = \begin{cases} \pi_H(x,a) & \text{if } x \in u(0) \cup \ell(1) \\ (1-\lambda)\pi_B(x,a) + \lambda\pi_H(x,a) & \text{otherwise.} \end{cases}$$

Note that $\pi_{A,0} = \pi_B$ and $\pi_{A,1} = \pi_H$. In addition, it is easy to see that $\pi_{A,\lambda}$ has strictly lower deviation than $\pi_H$ for all $\lambda \in [0,1)$ (strictness is due to Assumption 1 and our assumption on $\pi_H$ in the theorem statement). Next, by construction, for all $\lambda \in [0,1]$, we have $\pi_{A,\lambda}(x,1) \leq \pi_H(x,1)$ and $\pi_{A,\lambda}(x,0) \geq \pi_H(x,0)$. Now, consider the function

$$g(\lambda) = \overline{\pi}_{A,\lambda}(1) - \overline{\pi}_{A,\lambda}(0).$$

By the above, we have

$$g(0) = \overline{\pi}_B(1) - \overline{\pi}_B(0) \leq 0$$
$$g(1) = \overline{\pi}_H(1) - \overline{\pi}_H(0) \geq 0.$$

Thus, by the intermediate value theorem, there exists $\lambda^* \in [0,1]$ such that $g(\lambda^*) = 0$. Since

$$g(1) = \overline{\pi}_H(1) - \overline{\pi}_H(0) = \alpha(\pi_H) \neq 0,$$

we know that $\lambda^* \neq 1$, so $\lambda^* \in [0,1)$. Thus, $\pi_{A,\lambda^*}$ satisfies (1), (2), and (3), so by Theorem 3.1, it is compliance-robustly fair. In addition, since $\lambda_1^* \in [0,1)$, by the above, it has strictly lower deviation than $\pi_H$, so $L(\pi_{A,\lambda^*}) < L(\pi_H)$. Thus, we have $L(\pi_0) \leq L(\pi_{A,\lambda^*}) < L(\pi_H)$, as claimed. □

### A.3 Proof of Results in Section 5

Lemma 5.1. *Assume that $\alpha(\pi_H) \neq 0$, and that either $\pi_H(x,1) \neq \pi_*(x,1)$ for some $x \in u(1)$ or $\pi_H(x,0) \neq \pi_*(x,0)$ for some $x \in \ell(0)$. Then, there exists a compliance-robustly fair policy $\pi_A$ that is also traditionally fair ($\alpha(\pi_A) = 0$) and performance-improving ($L(\pi_A) < L(\pi_H)$) if and only if there exists a policy $\pi$ satisfying*

$$\overline{\pi}(1) \leq \overline{\pi}(0) \tag{4}$$
$$\pi(x,1) \leq \pi_B(x,1) \qquad (\forall x \in \mathcal{X}) \tag{5}$$
$$\pi(x,0) \geq \pi_B(x,0) \qquad (\forall x \in \mathcal{X}) \tag{6}$$
$$L(\pi) < L(\pi_H). \tag{7}$$

Proof. We first show that existence of $\pi_A$ implies existence of $\pi$. By Theorem 3.1, $\pi_A$ satisfies

$$\begin{aligned}
\overline{\pi}_A(1) &= \overline{\pi}_A(0) \\
\pi_H(x,0) &\leq \pi_A(x,0) && (\forall x \in \mathcal{X}) \\
\pi_A(x,1) &\leq \pi_H(x,1) && (\forall x \in \mathcal{X}).
\end{aligned}$$

Now, let

$$\pi(x,a) = \begin{cases} \max\{\pi_A(x,0), \pi_B(x,0)\} & \text{if } a = 0 \\ \min\{\pi_A(x,1), \pi_B(x,1)\} & \text{if } a = 1. \end{cases}$$

By construction, $\pi$ satisfies (5) and (6). Furthermore, we have

$$\overline{\pi}(1) \leq \overline{\pi}_A(1) = \overline{\pi}_A(0) \leq \overline{\pi}(0),$$

where the first inequality follows since $\pi(x,1) \leq \pi_A(x,1)$ and the second since $\pi$ satisfies $\pi(x,0) \geq \pi_A(x,0)$. Thus, $\pi$ satisfies (4). Finally, to show that $L(\pi) < L(\pi_H)$, it suffices to show that $\pi$ has lower or equal deviation compared to $\pi_A$, since this implies that $L(\pi) \leq L(\pi_A) < L(\pi_H)$. To this end, recall that for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, we have $\pi_B(x,a) \in \{\pi_H(x,a), \pi_*(x,a)\}$. If $\pi_A(x,a) \neq \pi_H(x,a)$ and $\pi(x,a) \neq \pi_A(x,a)$, then we must $\pi(x,a) = \pi_B(x,a)$, so $\pi(x,a) \in \{\pi_H(x,a), \pi_*(x,a)\}$. In this case, we cannot have $\pi(x,a) = \pi_H(x,a)$, since either $a = 0$ and $\pi(x,0) \geq \pi_A(x,0) > \pi_H(x,0)$, or $a = 1$ and $\pi(x,1) \leq \pi_A(x,1) < \pi_H(x,1)$. Thus, we must have $\pi(x,a) = \pi_*(x,a)$. In general, it follows that $\pi(x,a) \in \{\pi_A(x,a), \pi_*(x,a)\}$, which straightforwardly implies that $\pi$ has lower or equal deviation compared to $\pi_A$. The claim follows.

Next, we prove that existence of $\pi$ implies the existence of $\pi_A$. First, if $\overline{\pi}_B(1) \leq \overline{\pi}_B(0)$, then the result follows from the proof of Theorem 4.1, which shows that if $\overline{\pi}_B(1) \leq \overline{\pi}_B(0)$, then there exists a compliance-robustly fair policy $\pi$ such that $\alpha(\pi) = 0$. Thus, it suffices to consider the case $\overline{\pi}_B(1) > \overline{\pi}_B(0)$. In this case, by (5) and (6), we have

$$\begin{aligned}
\pi(x,1) &\leq \min\{\pi_*(x,1), \pi_H(x,1)\} = \pi_B(x,1) \leq \pi_*(x,1) \\
\pi(x,0) &\geq \max\{\pi_*(x,0), \pi_H(x,0)\} = \pi_B(x,0) \geq \pi_*(x,0).
\end{aligned}$$

Thus, $\pi_B$ has lower or equal deviation compared to $\pi$, so $L(\pi_B) \leq L(\pi) < L(\pi_H)$. Consider

$$\pi_{A,\lambda}(x,a) = \lambda\pi(x,a) + (1-\lambda)\pi_B(x,a),$$

where $\lambda \in [0,1]$. Note that $\pi_{A,0} = \pi_B$ and $\pi_{A,1} = \pi$. It is easy to see that $\pi_{A,\lambda}$ has lower or equal deviation compared to $\pi$, so $L(\pi_{A,\lambda}) \leq L(\pi) < L(\pi_H)$ for all $\lambda$. Now, define

$$g(\lambda) = \overline{\pi}_{A,\lambda}(1) - \overline{\pi}_{A,\lambda}(0),$$

so

$$\begin{aligned}
g(0) &= \overline{\pi}_B(1) - \overline{\pi}_B(0) > 0 \\
g(1) &= \overline{\pi}(1) - \overline{\pi}(0) < 0.
\end{aligned}$$

By the intermediate value theorem, there exists $\lambda^* \in (0,1)$ such that $g(\lambda^*) = 0$. Then, we have $\alpha(\pi_{A,\lambda^*}) = 0$ and $L(\pi_{A,\lambda^*}) < L(\pi_H)$. It also directly follows from Theorem 3.1

that $\pi_{A,\lambda^*}$ is compliance-robustly fair. Thus, $\pi_{A,\lambda^*}$ satisfies our desiderata, so the claim follows. □

Proposition 5.3. *There exists $\mathcal{X}$, $\mathbb{P}$, $L$, and $\pi_H$ satisfying $\alpha(\pi_H) \neq 0$ and $\pi_H \neq \pi^*$ such that for any policy $\pi$, $\pi$ cannot simultaneously satisfy all of the following: (i) $\pi \in \Pi_{fair}$, (ii) $\alpha(\pi) = 0$, and (iii) $L(\pi) \leq L(\pi_H)$.*

Proof. Let $\mathcal{X} = \{1\}$ be singleton; thus, we can omit it from our notation. Let

$$\mathbb{P}(a, y) = \begin{cases} \frac{1}{2}(1-\epsilon) & \text{if } a = 1 \wedge y = 1 \\ \frac{1}{2}\epsilon & \text{if } a = 1 \wedge y = 0 \\ \frac{1}{2}\epsilon & \text{if } a = 0 \wedge y = 1 \\ \frac{1}{2}(1-\epsilon) & \text{if } a = 0 \wedge y = 0 \end{cases}$$

for any $\epsilon \in (0, 1/7]$. Let the loss be

$$L(\pi) = \mathbb{E}[(\pi(a) - y)^2] = \frac{1}{2}\left[(1-\epsilon)(\pi(1)-1)^2 + \epsilon\pi(1)^2 + \epsilon(\pi(0)-1)^2 + (1-\epsilon)\pi(0)^2\right].$$

Then, it is easy to check that so that

$$\pi_*(a) = \begin{cases} 1-\epsilon & \text{if } a = 1 \\ \epsilon & \text{if } a = 0. \end{cases}$$

In addition, suppose that the human policy is

$$\pi_H(a) = \begin{cases} 1-\epsilon & \text{if } a = 1 \\ \epsilon/2 & \text{if } a = 0. \end{cases}$$

In this case, $\pi_B = \pi_*$, and $\alpha(\pi_B) = \alpha(\pi_*) < \alpha(\pi_H)$, so by Theorem 4.1, $\pi_B$ is compliance-robustly fair; in addition, it strictly improves performance, though it is itself unfair. Thus, $\Pi_{\text{fair}} \neq \emptyset$.

Next, we show that for any compliance-robustly fair policy $\pi$, if $\alpha(\pi) = 0$, then $L(\pi) \geq L(\pi_H)$. Since $\mathcal{X}$ is singleton, we have $\overline{\pi}(a) = \pi(a)$, so $\alpha(\pi) = 0$ implies $\pi(0) = \pi(1)$. Thus, it suffices to consider a policy $\pi(0) = \pi(1) = \beta$. For any such policy, the loss is

$$L(\pi) = \frac{1}{2}\left[(1-\epsilon)(\beta-1)^2 + \epsilon\beta^2 + \epsilon(\beta-1)^2 + (1-\epsilon)\beta^2\right] = \frac{1}{2}\left[(\beta-1)^2 + \beta^2\right],$$

which is minimized when $\beta = 1/2$, in which case $L(\pi) = 1/4$. In contrast, we have

$$\begin{aligned} L(\pi_H) &= \frac{1}{2}\left[(1-\epsilon)(\epsilon)^2 + \epsilon(1-\epsilon)^2 + \epsilon(\epsilon/2-1)^2 + (1-\epsilon)(\epsilon/2)^2\right] \\ &= \frac{1}{2}\left[(\epsilon/2)^2 + 2\epsilon(1-\epsilon)\right). \end{aligned}$$

It is easy to verify that when $\epsilon \in (0, \frac{1}{7}]$, we have $L(\pi_H) < \frac{1}{4} \leq L(\pi)$.

□

### A.4 Compliance Issues for General Fairness Conditions

We define a general class of fairness criteria, subsuming demographic parity [Calders et al., 2009, Zliobaite, 2015] and equalized odds [Chen et al., 2023, Hardt et al., 2016]. We then show that, under this general class, fair policies are not necessarily compliance-robustly fair. Thus, in all cases, one must optimize separately for performance-improving compliance-robustly fair policies (as illustrated in Algorithm 1).

We define a *fairness criterion* as a function that takes a policy as input and outputs a value representing how fair the policy is. For example, $\alpha(\pi) = |\overline{\pi}(1) - \overline{\pi}(0)|$ quantifies fairness under the equality of opportunity criterion.

*Definition A.4.* A *fairness criterion* is a function $\varphi : \Pi \to \mathbb{R}_{\geq 0}$, where $\Pi$ is the space of all policies. Given $\varphi$, we say a policy $\pi \in \Pi$ is *fairer* than another policy $\pi' \in \Pi$ if $\varphi(\pi) < \varphi(\pi')$.

Next, we extend our concept of a compliance-robustly fair policy to general fairness criteria.

*Definition A.5.* Given a human policy $\pi_H$ and an algorithmic policy $\pi_A$, we say that $\pi_A$ is *compliance-robustly fair* with respect to $\pi_H$ if for every compliance function $c$, the resulting human-AI policy $\pi_C$ satisfies $\varphi(\pi_C) \leq \varphi(\pi_H)$.

The following assumption characterizes the class of fairness criteria that are susceptible to selective compliance issues. That is, if a fairness criterion satisfies the assumption, there is tension between traditional fairness and compliance-robust fairness.

Assumption 4. *Given a fairness condition $\varphi$, there exist policies $\pi_{low}$ and $\pi_{high}$, and a compliance function $c_0$, such that (i) we have*

$$\varphi(\pi_{low}) < \varphi(\pi_{high}),$$

*(ii) the human-AI policy*

$$\pi_C(x, a) = \begin{cases} \pi_{low}(x, a) & \textit{if } c_0(x, a) = 1 \\ \pi_{high}(x, a) & \textit{otherwise}, \end{cases}$$

*satisfies*

$$\varphi(\pi_C) < \varphi(\pi_{high}),$$

*and (iii) the human-AI policy*

$$\pi'_C(x, a) = \begin{cases} \pi_{high}(x, a) & \textit{if } c_0(x, a) = 1 \\ \pi_{low}(x, a) & \textit{otherwise}. \end{cases}$$

*satisfies*

$$\varphi(\pi'_C) < \varphi(\pi_{high}).$$

In this assumption, condition (i) says that according to $\varphi$, the policies $\pi_{\text{low}}$ and $\pi_{\text{high}}$ are increasingly unfair. Then, condition (ii) says that if $\pi_{\text{high}}$ is the human policy and $\pi_{\text{low}}$ is the AI policy, then the resulting human-AI policy under the compliance function $c_0$ is strictly fairer than the human policy $\pi_{\text{high}}$. Finally, condition (iii) says that if $\pi_{\text{low}}$ is the human policy and $\pi_{\text{high}}$ is the AI policy, then the resulting human-AI policy under $c_0$ is again strictly more fair than $\pi_{\text{high}}$. In fact, condition (i) is not necessary, but we include it since it adds intuition—the human-AI policy can be thought of as moving closer to $\pi_{\text{low}}$ from $\pi_{\text{high}}$ in both cases.

Intuitively, conditions (ii) and (iii) say that there exist two policies $\pi_{\text{low}}$ and $\pi_{\text{high}}$ with different fairness levels such that either of the two human-AI policies formed by combining them has fairness strictly less than $\pi_{\text{high}}$. These conditions are met by a wide range of algorithmic fairness definitions; later in this section, we will show that two widely-used fairness definitions—demographic parity and equalized odds—satisfy it.

Next, we show that any fairness definition satisfying Assumption 4 is vulnerable to the selective compliance problem. This result demonstrates the pervasive nature of the selective compliance problem; as a result, there exists an inherent tension between traditional fairness and compliance-robust fairness for a broad class of fairness definitions.

Theorem A.6. *For any fairness condition $\varphi$ satisfying Assumption 4, there exists a human policy $\pi_H$ and an algorithmic policy $\pi_A$ such that $\varphi(\pi_A) \leq \varphi(\pi_H)$ but $\pi_A$ is not compliance-robustly fair for $\pi_H$.*

Proof. We show that it is always possible to construct a human-AI policy $\pi_C$ that is less fair than the human-alone policy $\pi_H$ under Assumption 4, even though the AI policy $\pi_A$ is fairer than the human-alone policy $\pi_H$.

Let $\pi_{\text{low}}$, $\pi_{\text{high}}$, and $c_0$ be as defined in Assumption 4, and consider the policy

$$\pi_1(x, a) = \begin{cases} \pi_{\text{high}}(x, a) & \text{if } c_0(x, a) = 1 \\ \pi_{\text{low}}(x, a) & \text{otherwise.} \end{cases}$$

By Assumption 4, $\varphi(\pi_1) < \varphi(\pi_{\text{high}})$. Also, consider the policy

$$\pi_2(x, a) = \begin{cases} \pi_{\text{low}}(x, a) & \text{if } c_0(x, a) = 1 \\ \pi_{\text{high}}(x, a) & \text{otherwise.} \end{cases}$$

By Assumption 4, $\varphi(\pi_2) < \varphi(\pi_{\text{high}})$. Now, if $\varphi(\pi_1) \leq \varphi(\pi_2)$, then consider

$$\pi_C^{(1)}(x, a) = \begin{cases} \pi_1(x, a) & \text{if } c_0(x, a) = 1 \\ \pi_2(x, a) & \text{otherwise.} \end{cases}$$

Note that $\pi_C^{(1)} = \pi_{\text{high}}$ since $\pi_C^{(1)}(x, a) = \pi_1(x, a) = \pi_{\text{high}}(x, a)$ if $c_0(x, a) = 1$ and $\pi_C^{(1)}(x, a) = \pi_2(x, a) = \pi_{\text{high}}(x, a)$ otherwise. Thus, $\varphi(\pi_2) < \varphi(\pi_{\text{high}}) = \varphi(\pi_C^{(1)})$. Taking $\pi_A = \pi_1$ and $\pi_H = \pi_2$, we have $\varphi(\pi_1) \leq \varphi(\pi_2)$, but $\pi_1$ is not compliance-robustly fair for $\pi_2$ because $\varphi(\pi_C^{(1)}) > \varphi(\pi_2)$.

Otherwise, we have $\varphi(\pi_1) > \varphi(\pi_2)$. Let

$$\pi_C^{(2)}(x, a) = \begin{cases} \pi_2(x, a) & \text{if } \tilde{c}_0(x, a) = 1 \\ \pi_1(x, a) & \text{otherwise.} \end{cases}$$

where

$$\tilde{c}_0(x,a) = 1 - c_0(x,a).$$

Similar to before, we have $\pi_C^{(2)} = \pi_{\text{high}}$. Thus, $\varphi(\pi_1) < \varphi(\pi_{\text{high}}) = \varphi(\pi_C^{(2)})$. Taking $\pi_A = \pi_2$ and $\pi_H = \pi_1$, we have $\varphi(\pi_2) < \varphi(\pi_1)$, but $\pi_2$ is not compliance-robustly fair for $\pi_1$ because $\varphi(\pi_C^{(2)}) > \varphi(\pi_1)$. □

*Demographic parity.* Now, we show that demographic parity satisfies Assumption 4, implying that it suffers from compliance-related problems. In particular, redefine the following:

$$\overline{\pi}(a) = \sum_{x \in \mathcal{X}} \pi(x,a)\mathbb{P}(x \mid a),$$

so demographic parity is given by $\alpha_D(\pi) = |\bar{\pi}(1) - \bar{\pi}(0)|$. We need to establish a setting for which the two policies and the compliance function in Assumption 4 exist. Let $\mathcal{X} = \{1\}$ be singleton; then, we can omit it from our notation. Next, we construct $\pi_{\text{high}}$ and $\pi_{\text{low}}$ as follows:

$$\pi_{\text{low}}(a) = \begin{cases} \frac{1}{2} + \epsilon & \text{if } a = 1 \\ \frac{1}{2} - \epsilon & \text{if } a = 0 \end{cases}$$

$$\pi_{\text{high}}(a) = \begin{cases} \frac{1}{2} + 3\epsilon & \text{if } a = 1 \\ \frac{1}{2} - 2\epsilon & \text{if } a = 0, \end{cases}$$

where $\epsilon \in (1/6, 1/4)$. Also, consider the compliance function:

$$c_0(a) = \begin{cases} 1 & \text{if } a = 1 \\ 0 & \text{if } a = 0, \end{cases}$$

which implies $\pi_C$ and $\pi_C'$ are as follows:

$$\pi_C(a) = \begin{cases} \frac{1}{2} + \epsilon & \text{if } a = 1 \\ \frac{1}{2} - 2\epsilon & \text{if } a = 0 \end{cases}$$

$$\pi_C'(a) = \begin{cases} \frac{1}{2} + 3\epsilon & \text{if } a = 1 \\ \frac{1}{2} - \epsilon & \text{if } a = 0 \end{cases}$$

With these definitions, it is easy to see that Assumption 4 is satisfied.

*Equalized Odds.* The case of equalized odds is similar to that of equal opportunities. Redefine the following:

$$\overline{\pi}(a,y) = \sum_{x \in \mathcal{X}} \pi(x,a)\mathbb{P}(x|a,y).$$

Then, equalized odds can be defined as follows:

$$\varphi(\pi) = \sup_{y \in \{0,1\}} |\overline{\pi}(1,y) - \overline{\pi}(0,y)|.$$

As before, consider $\mathcal{X} = \{1\}$ be singleton, then we can omit it from our notation. Note that $\overline{\pi}(1, y) = \pi(1)$ and $\overline{\pi}(0, y) = \pi(0)$. Next, we define $\pi_{\text{high}}$ and $\pi_{\text{low}}$ as follows:

$$\pi_{\text{low}}(a) = \begin{cases} \frac{1}{2} + \epsilon & \text{if } a = 1 \\ \frac{1}{2} - \epsilon & \text{if } a = 0 \end{cases}$$

$$\pi_{\text{high}}(a) = \begin{cases} \frac{1}{2} + 3\epsilon & \text{if } a = 1 \\ \frac{1}{2} - 2\epsilon & \text{if } a = 0, \end{cases}$$

where $\epsilon \in (1/6, 1/4)$. Also, consider the compliance function:

$$c_0(a) = \begin{cases} 1 & \text{if } a = 1 \\ 0 & \text{if } a = 0, \end{cases}$$

which implies that $\pi_C$ and $\pi'_C$ are as follows:

$$\pi_C(a) = \begin{cases} \frac{1}{2} + \epsilon & \text{if } a = 1 \\ \frac{1}{2} - 2\epsilon & \text{if } a = 0 \end{cases}$$

$$\pi'_C(a) = \begin{cases} \frac{1}{2} + 3\epsilon & \text{if } a = 1 \\ \frac{1}{2} - \epsilon & \text{if } a = 0 \end{cases}$$

Again, it is easy to see that Assumption 4 is satisfied.

## A.5 Experimental Details

*Sample Selection.* Following the setup of Stevenson and Doleac [2024], we restrict the sample to defendants that are eligible for the non-violent risk assessment tool, which is our population of interest—we select defendants that (1) committed a drug, larceny, or fraud offense, (2) do not have a history of violent offenses, and (3) are considered for a prison or jail sentence. Then, we augment the criminal sentence records by merging it with defendants' demographic information obtained from the Virginia Court Data website. We also restrict to Non-Hispanic White and Black defendants.

*Defendant Covariates.* Following Stevenson and Doleac [2024], we use "Defendant Sex", "Defendant Age", "Defendant Race", "Defendant in Youthful Offender Program", "Charge Type", "Mandatory Minimum Sentence", "Recommend Prison", "First Offender", "Recommended Sentence Length", and "Primary Offenses".

*Estimating the judges' policies.* We leverage guidelines-recommended sentences to infer judges' perceived recidivism risk. The guidelines provide judges with a range of suitable sentences (e.g., 6 months to 2 years), the midpoint of which is defined as the "guidelines-recommended sentence." Following Stevenson and Doleac [2024], we consider the judge to perceive an offender to have a low recidivism risk if (1) the guideline-recommended sentence is prison (more than 12 months), but the judge assigns a sentence of 6 months or less of jail time, or (2) the guideline-recommended sentence is jail (less or equal to 12 months), but the judge assigns a sentence of zero (i.e., not incarcerated at all).

*Human-AI Collaborative Policy Construction.* We first estimate the optimal performance-maximizing policy $\pi_A^*$, our compliance-robustly fair policy $\pi_A^{\text{robust}}$, and the performance-maximizing traditionally fair (i.e., satisfying Equality of Opportunity) policy $\pi_A^{\text{trad fair}}$ for each judge (based on their estimated $\pi_H$) using data prior to the deployment of the risk assessment tool.

To learn $\pi_A^*$ and $\pi_A^{\text{trad fair}}$, we require the true outcome $y$ for defendants. Consistent with VCSC's definitions, we label any defendant that receives another felony conviction within a three-year window after their release as a recidivist. We then train a gradient boosted decision tree [Ke et al., 2017] to predict whether a defendant is a recidivist based on the same observed defendant covariates, yielding $\pi_A^*$. For $\pi_A^{\text{trad fair}}$, we use the methods proposed by Weerts et al. [2023] to enforce the Equality of Opportunity fairness constraint.

Then, we have the following four policies:

$$\pi_C^{\text{actual}}(x, a) = \begin{cases} \pi_A^{\text{actual}}(x, a) & \text{if } c(x, a) = 1 \\ \pi_H(x, a) & \text{otherwise.} \end{cases}$$

$$\pi_C^{\text{robust}}(x, a) = \begin{cases} \pi_A^{\text{robust}}(x, a) & \text{if } c(x, a) = 1 \\ \pi_H(x, a) & \text{otherwise.} \end{cases}$$

$$\pi_C^*(x, a) = \begin{cases} \pi_A^*(x, a) & \text{if } c(x, a) = 1 \\ \pi_H(x, a) & \text{otherwise.} \end{cases}$$

$$\pi_C^{\text{trad fair}}(x, a) = \begin{cases} \pi_A^{\text{trad fair}}(x, a) & \text{if } c(x, a) = 1 \\ \pi_H(x, a) & \text{otherwise,} \end{cases}$$

*Metrics.* To evaluate performance, we compute the average loss as follows:

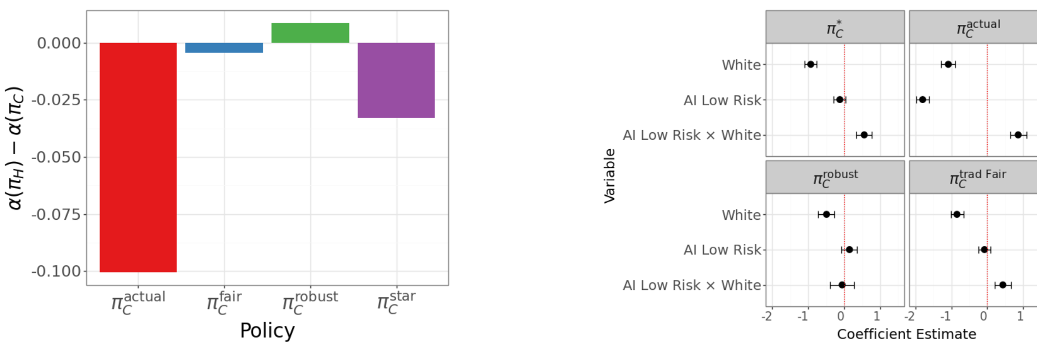
$$L(\pi_H) - L(\pi_C) = \frac{1}{N}\sum_{i=1}^{N} \ell(\pi_H(x_i, a_i), y_i) - \frac{1}{N}\sum_{i=1}^{N} \ell(\pi_C(x_i, a_i), y_i),$$

where $N$ is the number of samples in our evaluation dataset. The outcome $y_i$ indicates whether the defendant $i$ in fact recidivates (which we observe in the data). Note that a positive difference in average loss indicates an improvement in performance over the judges' policy. Similarly, we evaluate fairness using the following metric:

$$\alpha(\pi_H) - \alpha(\pi_C),$$

where $\alpha(\pi)$ is the slack in group fairness for $\pi$. In this case, a positive difference indicates that $\pi_C$ improves equity over the judges' policy.

*Problematic Compliance Patterns.* We focus on judges who exhibit fairness deterioration, as shown in Figure 2—i.e., $\alpha(\pi_H) - \alpha(\pi_C) < 0$ across $\pi_A^*$, $\pi_A^{\text{actual}}$, and $\pi_A^{\text{trad fair}}$— and select the bottom 50%, yielding 11 judges. From these judges, we derive a single "problematic compliance function", $c_{\text{problem}}$, by averaging individual judges' compliance functions. Similarly, we compute the "average human-alone policy," $\pi_H^{\text{Ave}}$, by averaging their individual human-alone policies.




(a) Fairness of four human-AI policies under the observed problematic compliance function

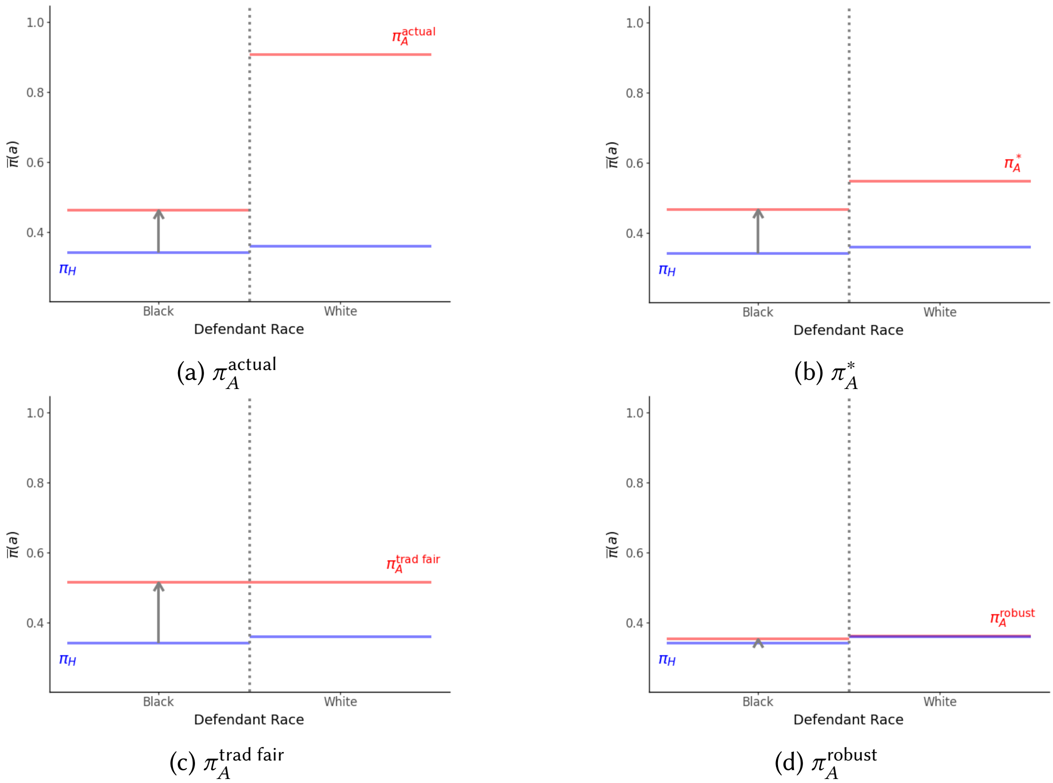
(b) Regression coefficients from regressing judges' compliance decisions on defendants' race and AI's recommendation

Fig. 3. In the left panel, we show the fairness comparison of $\pi_C^{\text{actual}}$, $\pi_C^{\text{robust}}$, $\pi_C^*$ and $\pi_C^{\text{trad fair}}$ for a problematic compliance function. In the right panel, we present the regression coefficients from the regression specification in Section 6.2. The error bars are 95% bootstrapped confidence intervals.

Using the problematic compliance function $c_{\text{problem}}$ and the average human-alone policy $\pi_H^{\text{Ave}}$, we simulate the four human-AI policies: $\pi_C^*$, $\pi_C^{\text{robust}}$, $\pi_C^{\text{actual}}$, and $\pi_C^{\text{trad fair}}$. We compare their fairness and present the results in Figure 3a. The vertical axis represents the unfairness level, $\alpha(\pi_H) - \alpha(\pi_C)$. A negative value indicates that the human-AI policy $\pi_C$ is less fair than the human-alone policy $\pi_H$. Indeed, all human-AI policies, except the compliance-robustly fair policy, reduce fairness.

In the regression presented in Section 6.2, the parameter $\beta_3$ captures our quantity of interest—a positive value indicates that judges comply more often for White defendants when the algorithmic recommendation is more lenient than their independent decisions, suggesting that racial disparities are exacerbated under algorithmic advice. We run this regression for each of the four human-AI policies. As shown in Figure 3b, the estimated $\beta_3$ is positive and statistically significant for all human-AI policies ($\pi_C^{\text{actual}}$, $\pi_C^{\text{trad fair}}$, and $\pi_C^*$), indicating that judges' compliance behaviors exacerbate existing racial biases under these policies. In contrast, our compliance-robustly fair policy ($\pi_A^{\text{robust}}$) effectively guards against such problematic compliance behaviors.

As discussed in Section 4, the algorithmic policy cannot further advantage the advantaged group (in this case, Whites) than the human-alone policy without risking increased disparities for problematic compliance patterns. We identify a defendant subgroup that experiences the most significant fairness deterioration under our "problematic compliance function"—specifically, 40-50 year old males who are not first offenders, are charged with drug-related offenses, and are recommended prison time based on VCSC guidelines. In Figure 4, we illustrate the compliance implications for different algorithmic advice strategies. Our compliance-robust policy is the only one that preserves overall fairness by imitating the human policy for the advantaged subgroup, as in the construction of $\pi_B$.




(a) $\pi_A^{\text{actual}}$

(b) $\pi_A^{*}$

(c) $\pi_A^{\text{trad fair}}$

(d) $\pi_A^{\text{robust}}$

Fig. 4. Policy depictions for a defendant subgroup that evokes increased unfairness: [male, 40-50 years old, guideline recommends prison, non-first offender, drug-related offenses]. $\pi_A^{\text{robust}}$ preserves fairness by imitating the human policy.