Fair Decisions Despite Imperfect Predictions

Niki Kilbertus[1,2] **Manuel Gomez-Rodriguez**[3] **Bernhard Schölkopf**[1]

Krikamol Muandet[1] **Isabel Valera**[1]

[1]MPI for Intelligent Systems, [2]University of Cambridge, [3]MPI for Software Systems

Abstract

Consequential decisions are increasingly informed by sophisticated data-driven predictive models. However, consistently learning accurate predictive models requires access to ground truth labels. Unfortunately, in practice, labels may only exist conditional on certain decisions—if a loan is denied, there is not even an option for the individual to pay back the loan. In this paper, we show that, in this *selective labels* setting, learning to predict is suboptimal in terms of both fairness and utility. To avoid this undesirable behavior, we propose to directly learn stochastic decision policies that maximize utility under fairness constraints. In the context of fair machine learning, our results suggest the need for a paradigm shift from "*learning to predict*" to "*learning to decide*". Experiments on synthetic and real-world data illustrate the favorable properties of learning to decide, in terms of both utility and fairness.

1 INTRODUCTION

The use of machine learning models to assist consequential decision making—where decisions have significant consequences for individuals—is becoming common in a variety of critical applications. For example, in pretrial release decisions, a judge may consult a learned model of the probability of recidivism to decide whether to grant bail or not. In loan decisions, a bank may decide whether or not to offer a loan based on learned estimates of the credit default probability. In fraud detection, an insurance company may flag suspicious claims based on a machine learning model's predicted probability that the claim is fraudulent. In all these scenarios, the goal of the decision maker (bank, law court, or insurance company) may be to take decisions that maximize a given utility function. In contrast, the goal of the machine learning model is solely to provide an accurate prediction of the outcome, referred to as (ground truth) label.

In this context, there has been much work on computational mechanisms to ensure that machine learning models do not disproportionately harm particular demographic groups sharing one or more sensitive attributes, e.g., race or gender (Dwork et al., 2012; Feldman et al., 2015). However, most of this work does not distinguish between decisions and label predictions and, consequently, suggests an inherent trade-off between fairness and prediction accuracy (Chouldechova, 2017; Kleinberg et al., 2017b). Only recently has the distinction been made explicit (Corbett-Davies et al., 2017; Kleinberg et al., 2017a; Mitchell et al., 2018; Valera et al., 2018). This recent line of work has shown that if a predictive model achieves perfect prediction accuracy, *deterministic threshold rules*, which derive decisions deterministically from the predictive model by thresholding, achieve maximum utility under various fairness constraints. This lends support to focusing on deterministic threshold rules and seemingly justifies using predictions and decisions interchangeably.

However, in many practical scenarios, the decision determines whether a label is realized or not—if bail (a loan) is denied, there is not even an option for the individual to reoffend (pay back the loan). This problem has been referred to by Lakkaraju et al. (2017) as *selective labels*. As a consequence, the labeled data used to train predictive models often depend on the decisions taken, which likely leads to suboptimal performance. Even worse, deterministic threshold rules using even slightly imperfect predictive models can be far from optimal (Woodworth et al., 2017). This negative result raises the following question: *Can we do better if we learn directly to decide rather than to predict?*

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

In the present work, we first articulate how the "learning to predict" approach fails in a utility maximization setting (with fairness constraints) that accommodates a variety of real-world applications, including those mentioned previously. We show that label data gathered under deterministic rules (e.g., prediction based threshold rules) are neither sufficient to improve the accuracy of the underlying predictive model, nor the utility of the decision making process. We then demonstrate how to overcome this undesirable behavior using a particular family of stochastic decision rules and introduce a simple gradient-based algorithm to learn them from data. Experiments on synthetic and real-world data illustrate our theoretical results and show that, under imperfect predictions, *learning to predict* is inferior to *learning to decide*. Code is available at github.com/nikikilbertus/fair-decisions

Related work. The work most closely related to ours analyzes the long-term effects of consequential decisions informed by data-driven predictive models on underrepresented groups (Hu & Chen, 2018; Liu et al., 2018; Mouzannar et al., 2019; Tabibian et al., 2019). However, this line of work focuses mainly on the evolution of several measures of well-being under a perfect predictive model, neglecting the data collection phase (Dimitrakakis et al., 2019; Holstein et al., 2018). In contrast, we focus on analyzing how to improve a suboptimal decision process when labels exist only for positive decisions. More broadly, our work relates to the growing literature on fairness in machine learning, which mostly attempts to match various statistics of the predictive models across protected subgroups.

We also build on previous work on counterfactual inference and policy learning (Athey & Wager, 2017; Ensign et al., 2018; Gillen et al., 2018; Heidari & Krause, 2018; Joseph et al., 2016; Jung et al., 2018; Kallus, 2018; Kallus & Zhou, 2018; Lakkaraju & Rudin, 2017). In these settings, the decision typically determines which of the potential outcomes is observed and the focus is on confounders that effect both the decision and the outcome (Rubin, 2005). In contrast, in our approach the decision determines whether there will be an outcome at all, but there is no unobserved confounding. Two notable exceptions are by Kallus & Zhou (2018) and Ensign et al. (2018), which also consider limited feedback. However, Kallus & Zhou (2018) focus on designing unbiased estimates for several fairness measures, rather than learning how to decide. Ensign et al. (2018) assume a deterministic mapping between features and labels, which allows them to reduce the problem to the apple tasting problem (Helmbold et al., 2000). Remarkably, in their deterministic setting, they also conclude that the optimal decisions should be stochastic.

Unlike in the fairness literature, where deterministic policies dominate (Corbett-Davies et al., 2017; Valera et al., 2018; Meyer et al., 2019), stochastic policies are often necessary to ensure adequate exploration (Silver et al., 2014) in contextual bandits (Dudík et al., 2011; Langford et al., 2008; Agarwal et al., 2014) and reinforcement learning (Jabbari et al., 2016; Sutton & Barto, 1998). However, the typical problem setting there differs fundamentally from ours and typically neither fairness constraints nor selective labels are taken into account. A recent notable exception is Joseph et al. (2016), initiating the study of fairness in multi-armed bandits, however, using a fairness notion orthogonal to the most popular ones (as considered in our work), and ignoring the selective labels problem.

2 DECISIONS FROM IMPERFECT PREDICTIVE MODELS

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the feature domain, $\mathcal{S} = \{0, 1\}$ the range of sensitive attributes, and $\mathcal{Y} = \{0, 1\}$ the set of ground truth labels. We assume the standard sigma algebras on these spaces. A *decision rule* or *policy*[1] is a mapping $\pi : \mathcal{X} \times \mathcal{S} \to \mathcal{P}(\{0, 1\})$ that maps an individual's feature vector and sensitive attribute to a probability distribution over *decisions* $d \in \{0, 1\}$. We sample $\boldsymbol{x}, s$ and y from a ground truth distribution $P(\boldsymbol{x}, s, y) = P(y \,|\, \boldsymbol{x}, s)P(\boldsymbol{x}, s)$. Decisions d are sampled from a policy $d \sim \pi(d \,|\, \boldsymbol{x}, s)$, where we often write $\pi(\boldsymbol{x}, s)$ for $\pi(d \,|\, \boldsymbol{x}, s)$. The decision determines whether the label $y \sim P(y \,|\, \boldsymbol{x}, s)$ comes into existence. In loan decisions, the feature vector $\boldsymbol{x}$ may include salary, education, or credit history; the sensitive attribute s may indicate sex; a loan can be granted ($d = 1$) or denied ($d = 0$); and the label y indicates repayment ($y = 1$) or default ($y = 0$) *upon receiving a loan*.

Inspired by Corbett-Davies et al. (2017), we measure the *utility* as the expected overall profit provided by the policy with respect to the distribution P, i.e.,

$$\begin{aligned} u_P(\pi) &:= \mathbb{E}_{\boldsymbol{x}, s, y \sim P,\, d \sim \pi(\boldsymbol{x}, s)} \left[y\, d - c\, d \right] \\ &= \mathbb{E}_{\boldsymbol{x}, s \sim P} \left[\pi(d = 1 \,|\, \boldsymbol{x}, s)(P(y = 1 \,|\, \boldsymbol{x}, s) - c) \right], \end{aligned} \tag{1}$$

where $c \in (0, 1)$ may reflect economic considerations of the decision maker. For example, in a loan scenario, the utility gain is $(1 - c)$ if a loan is granted and repaid, $-c$ if a loan is granted but the individual defaults, and zero if the loan is not granted. One could think of adding a term for negative decisions of the form $g(y)(1 - d)$ for some given definition of g, however, we would not be able to compute such a term due to the selective labels, except for constant g. Therefore, without loss of generality, we assume that $g(y) = 0$ for all y, because any constant g can easily be absorbed in our framework.

[1] We use the terms *decision rule*, *decision making process* and *policy* interchangeably.

For fairness considerations, we define the *f-benefit for group* $s \in \{0,1\}$ with respect to the distribution P by

$$b_P^s(\pi) := \mathbb{E}_{\boldsymbol{x},y\sim P(\boldsymbol{x},y\,|\,s),\, d\sim\pi(\boldsymbol{x},s)}[f(d,y)],$$

with $f : \{0,1\}\times\{0,1\} \to \mathbb{R}$. Note that various common fairness criteria can be expressed as $b_P^0(\pi) = b_P^1(\pi)$ for different choices of f. For example, *demographic parity* (or no disparate impact) (Feldman et al., 2015) amounts to $f(d,y) = d$ and *equality of opportunity* (Hardt et al., 2016) amounts to $f(d,y) = d \cdot y$.

Under perfect knowledge of $P(y\,|\,\boldsymbol{x},s)$, the policy maximizing the above utility subject to the group benefit fairness constraint $b_P^0(\pi) = b_P^1(\pi)$ is a deterministic threshold rule (Corbett-Davies et al., 2017)[2]

$$\pi^*(d=1\,|\,\boldsymbol{x},s) = \mathbf{1}[P(y=1\,|\,\boldsymbol{x},s) \geq c_s], \qquad (2)$$

where we allow for group specific cost factors c_0, c_1 such that $b_P^0(\pi) = b_P^1(\pi)$. Without fairness constraints, we simply have $c_0 = c_1 = c$. However, as discussed in Woodworth et al. (2017), in practice, we typically do not have access to the true conditional distribution $P(y\,|\,\boldsymbol{x},s)$, but instead to an imperfect predictive model $Q(y\,|\,\boldsymbol{x},s)$ trained on a finite training set. Such a predictive model can similarly be used to implement a deterministic threshold rule as

$$\pi_Q(d=1\,|\,\boldsymbol{x},s) = \mathbf{1}[Q(y=1\,|\,\boldsymbol{x},s) \geq c]. \qquad (3)$$

Here, the predictor $Q(y=1\,|\,\boldsymbol{x},s) \approx P(y=1\,|\,\boldsymbol{x},s) - \delta_s$, with $\delta_s = c_s - c$, directly incorporates the fairness constraint, i.e., it is trained to maximize predictive power subject to the fairness constraint. In this context, Woodworth et al. (2017) have shown that this approach often leads to better performance than post-processing a potentially unfair predictor as proposed by Hardt et al. (2016). Unfortunately, they have also shown that, because of the mismatch between $Q(y=1\,|\,\boldsymbol{x},s)$ and $P(y=1\,|\,\boldsymbol{x},s) - \delta_s$, the resulting policy π_Q will usually still be suboptimal in terms of both utility and fairness. To make things worse, due to the selective labeling, the data points $\boldsymbol{x}, s, y$ observed under a given policy π_0 are not i.i.d. samples from the ground truth distribution $P(\boldsymbol{x},s,y)$, but instead from the weighted distribution

$$P_{\pi_0}(\boldsymbol{x},s,y) \propto P(y\,|\,\boldsymbol{x},s)\,\pi_0(d=1\,|\,\boldsymbol{x},s)\,P(\boldsymbol{x},s). \qquad (4)$$

Consequently, if π_0 is not optimal, i.e., $\pi_0 \neq \pi^*$, the necessary i.i.d. assumption for consistency results of empirical risk minimization is violated, which may also be one reason for a common observation in fairness, namely that predictive errors are often systematically larger for minority groups (Angwin et al., 2016). In the remainder, we will say that the distributions $P_{\pi_0}(\boldsymbol{x},s,y)$ and $P_{\pi_0}(\boldsymbol{x},s)$ are *induced* by the policy π_0. In the next section, we study how to learn the optimal policy, potentially subject to fairness constraints, if the data is collected from an initial faulty policy π_0.

[2]Here, $\mathbf{1}[\bullet]$ is 1 if the predicate $\bullet$ is true and 0 otherwise.

3 FROM DETERMINISTIC TO STOCHASTIC POLICIES

Consider a class of policies Π, within which we want to maximize utility, as defined in eq. (1) subject to the group benefit fairness constraint $b_P^0(\pi) = b_P^1(\pi)$. We formulate this as an unconstrained optimization with an additional penalty term, namely to maximize

$$v_P(\pi) := u_P(\pi) - \frac{\lambda}{2}(b_P^0(\pi) - b_P^1(\pi))^2 \qquad (5)$$

over $\pi \in \Pi$ under the assumption that we do not have access to samples from the ground truth distribution $P(\boldsymbol{x},s,y)$, which $u_P(\pi)$ and $b_P^s(\pi)$ depend on. Instead, we only have access to samples from a distribution $P_{\pi_0}(\boldsymbol{x},s,y)$ induced by a given initial policy π_0 as in eq. (4). We first analyze this problem for deterministic threshold rules, before considering general deterministic policies, and finally also general stochastic policies.

3.1 Deterministic policies

First, assume the initial policy π_0 is a given deterministic threshold rule and Π is the set of all deterministic threshold rules, which means that each $\pi \in \Pi$ (and π_0) is of the form eq. (3) for some predictive model $Q(y\,|\,\boldsymbol{x},s)$. Given a hypothesis class of predictive models $\mathcal{Q}$, we reformulate eq. (5) to maximize

$$v_P(\pi_Q) := u_P(\pi_Q) - \frac{\lambda}{2}(b_P^0(\pi_Q) - b_P^1(\pi_Q))^2 \qquad (6)$$

over $Q \in \mathcal{Q}$, where the utility and the benefits for $s \in \{0,1\}$ are simply $u_P(\pi_Q) = \mathbb{E}_{\boldsymbol{x},s,y\sim P}[\mathbf{1}[Q(y=1\,|\,\boldsymbol{x},s) \geq c](y-c)]$ and $b_P^s(\pi_Q) = \mathbb{E}_{\boldsymbol{x},s,y\sim P}[f(\mathbf{1}[Q(y=1\,|\,\boldsymbol{x},s) \geq c], y)]$. Note that eq. (5) has a unique optimum π^*. Therefore, if $\pi^* \in \Pi$ (the set of all deterministic threshold rules), eq. (6) will also reach this optimum if $\mathcal{Q}$ is rich enough. However, the optimal predictor Q^* may not be unique, because the utility and the benefits are not sensitive to the precise values of $Q(y\,|\,\boldsymbol{x},s)$ above or below c.

If we only have access to samples from the distribution P_{π_0} induced by some $\pi_0 \neq \pi^*$, we may choose to simply learn a predictive model $Q_0^* \in \mathcal{Q}$ that empirically maximizes the objective $v_{P_{\pi_0}}(\pi_Q)$, where the utility and the benefits are computed with respect to the induced distribution P_{π_0}. However, the following negative result shows that, under mild conditions, Q_0^* leads to a suboptimal deterministic threshold rule.[3]

[3]All proofs can be found in appendix A.

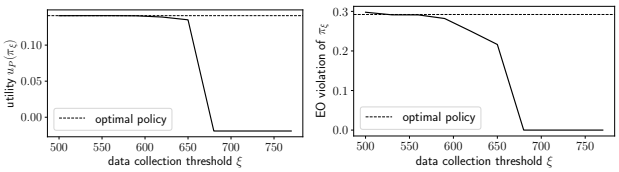

Figure 1: We show the utility and violation of equal opportunity of threshold decision rules $\pi^{(\xi_0)}$ learned from data collected with an initial threshold of ξ_0. Harsh data collection policies (i.e., large ξ_0)—while achieving equal opportunity—render the learned policies useless in terms of utility.

Proposition 1. *If there exists a subset $\mathcal{V} \subset \mathcal{X} \times \mathcal{S}$ of positive measure under P such that $P(y = 1 \mid \mathcal{V}) \geq c$ and $P_{\pi_0}(y = 1 \mid \mathcal{V}) < c$, then there exists a maximum $Q_0^* \in \mathcal{Q}$ of $v_{P_{\pi_0}}$ such that $v_P(\pi_{Q_0^*}) < v_P(\pi_{Q^*})$.*

Lending example. We briefly illustrate this result in a lending example based on FICO credit score data as described in Hardt et al. (2016). Such single feature scenarios are highly relevant for score-based decision support systems where full training data and the functional form of the score are often not available (e.g., also for pretrial risk assessment). For any score that is strictly monotonic in the true success rate, the optimal policy is simply to threshold the score. This lends additional support to score-based systems.

Here, we can generate new scores for a given group via inverse transform sampling from the known cumulative distribution functions. We consider 80% white and 20% black applicants. A hypothetical new bank that has access to FICO scores $x \in \mathcal{X} := \{300, \ldots, 820\}$, but not to the corresponding repayment probabilities may expect to be profitable if at least 70% of granted loans are repaid, i.e., $c = 0.7$. A risk-averse lender may initially choose a high score threshold $\xi \in \mathcal{X}$ and employ the decision rule $\mathbf{1}[x > \xi]$. After collecting repayment data $\mathcal{D}^{(\xi)} := \{(x_i, y_i)\}_{i=1}^n$ with this initial threshold, they learn a model $Q_\xi(y = 1 \mid x)$ and then decide based on $\pi_\xi(d = 1 \mid x) = \mathbf{1}[Q_\xi(y = 1 \mid x) > c]$. In Figure 1 we show how the initial data collection threshold ξ affects utility and fairness of the resulting predictive model-based decision rule. Conservatively high initial thresholds of $\xi \geq 650$ lead to essentially useless decisions π_ξ, because of imperfect prediction models regardless of how much data was collected. More lenient initial policies can result in near optimal decisions with improved fairness compared to the maximum utility policy for the given cost c (dashed). Details of this motivating example can be found in appendix B.

Impossibility results. Supplementing the result in Proposition 1, we will now prove that—in certain situations—a sequence of deterministic threshold rules, where each threshold rule is of the form of eq. (3) and its associated predictive model is trained using the data gathered through the deployment of previous threshold rules, fails to recover the optimal policy despite it being in the hypothesis class. To this end, we consider a *sequential policy learning task*, which is given by a tuple $(\pi_0, \Pi', \mathcal{A})$, where: a) $\Pi' \subset \Pi$ is the hypothesis class of policies, b) $\pi_0 \in \Pi'$ is the initial policy, and c) $\mathcal{A} : \Pi' \times \bigcup_{i=1}^{\infty} (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^i \to \Pi'$ is an update rule. The update rule $\mathcal{A}$ takes an existing policy π_t and a dataset $\mathcal{D} \in (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^n$ and produces an updated policy π_{t+1}, which typically aims to improve the policy in terms of the objective function $v_P(\pi)$ in eq. (5). In our setting, the dataset $\mathcal{D}$ is collected by deploying previous policies, i.e., from a mixture of the distributions $P_{\pi_\tau}(\boldsymbol{x}, s, y)$ with $\tau \leq t$.

To introduce useful notation and terminology, note that any deterministic threshold policy π is fully characterized by the sets $W_d(\pi) := \{(\boldsymbol{x}, s) \mid \pi(\boldsymbol{x}, s) = d\}$ for $d \in \{0, 1\}$, i.e., we can partition the space $\mathcal{X} \times \mathcal{S} = W_0(\pi) \cup W_1(\pi)$ into negative and positive decisions. Then, we say an update rule is *non-exploring on $\mathcal{D}$* iff $W_0(\mathcal{A}(\pi, \mathcal{D})) \subset W_0(\pi)$. Intuitively, this means that no individual who has received a negative decision under the old policy π would receive a positive decision under the new policy $\mathcal{A}(\pi, \mathcal{D}))$. Remarkably, common learning algorithms for classification, such as gradient boosted trees are *error based*, i.e., they only change the decision function when they make errors on the training set. As a result, they lead to non-exploring update rules on $\mathcal{D}$ whenever they achieve zero error.

Proposition 2. *Let $(\pi_0, \Pi', \mathcal{A})$ be a sequential policy learning task, where $\Pi' \subset \Pi$ are deterministic threshold policies based on a class of predictive models, and let the initial policy be more strict than the optimal one, i.e., $W_0(\pi_0) \supsetneq W_0(\pi^*)$. If $\mathcal{A}$ is non-exploring on any i.i.d. sample $\mathcal{D} \sim P_{\pi_t}(\boldsymbol{x}, s, y)$ with probability at least $1 - \delta_t$ for all $t \in \mathbb{N}$, then $Pr[\pi_T \neq \pi^*] > 1 - \sum_{t=0}^{T} \delta_t$ for any $T \in \mathbb{N}$.*

We can thus conclude that, for error based learning algorithms under no fairness constraints, learning within deterministic threshold policies is guaranteed to fail. Even though the optimal policy lies within the set of deterministic threshold policies, it cannot easily be approximated within this set starting from a suboptimal predictive model. We illustrate this fact in appendix C.

Corollary 3. *A deterministic threshold policy $\pi \neq \pi^*$ with $\Pr[\pi(\boldsymbol{x}, s) \neq y] = 0$ under P will never converge to π^* under an error based learning algorithm for the underlying predictive model.*

While we have focused on deterministic threshold rules, our results readily generalize to *all* deterministic policies. An arbitrary deterministic policy π can always be written as a threshold rule π_Q as in eq. (3) with

$Q(y = 1 \mid \boldsymbol{x}, s) = \mathbf{1}[\pi(d = 1 \mid \boldsymbol{x}, s) = 1]$. To conclude, if we can only observe the outcomes of previous decisions taken by a deterministic initial policy π_0, these outcomes may be insufficient to find the (fair) deterministic decision rule that maximizes utility.

3.2 Stochastic policies

To overcome the undesirable behavior exhibited by deterministic policies discussed in the previous section, one could use a fully randomized initial policy, where $\pi_0(d = 1 \mid \boldsymbol{x}, s) = 1/2$ for all $\boldsymbol{x}, s$. It readily follows from eq. (4) that then $P_{\pi_0} = P$. Hence, if the hypothesis class of predictive models $\mathcal{Q}$ is rich enough, we could learn the optimal policy π^* from data gathered under π_0. In practice, fully randomized initial policies are unacceptable in terms of utility or unethical—it would entail releasing defendants by a coin flip. Fortunately, we will show next that full randomization is not required to learn the optimal policy. We only need to choose an initial policy π_0 such that $\pi_0(d = 1 \mid \boldsymbol{x}, s) > 0$ on any measurable subset of $\mathcal{X} \times \mathcal{S}$ with positive probability under P, a requirement that is more acceptable for the decision maker in terms of initial utility. We refer to any policy with this property as an *exploring* policy.[4] For an exploring policy π_0, we can compute the utility in eq. (1) and the group benefits for $s \in \{0, 1\}$ via inverse propensity score weighting

$$\begin{aligned} u_{P_{\pi_0}}(\pi, \pi_0) &:= \mathbb{E}_{\substack{\boldsymbol{x},s,y\sim P_{\pi_0} \\ d\sim\pi(\boldsymbol{x},s)}}\Big[\frac{d(y-c)}{\pi_0(d=1\mid \boldsymbol{x},s)}\Big], \\ b^s_{P_{\pi_0}}(\pi, \pi_0) &:= \mathbb{E}_{\substack{\boldsymbol{x},s,y\sim P_{\pi_0} \\ d\sim\pi(\boldsymbol{x},s)}}\Big[\frac{f(d,y)}{\pi_0(d=1\mid \boldsymbol{x},s)}\Big]. \end{aligned} \tag{7}$$

Crucially, even though $u_P(\pi) = u_{P_{\pi_0}}(\pi, \pi_0)$ and $b^s_P(\pi) = b^s_{P_{\pi_0}}(\pi, \pi_0)$, the expectations are with respect to the induced distribution $P_{\pi_0}(\boldsymbol{x}, s, y)$, yielding the following positive result.

Proposition 4. *Let Π be the set of exploring policies and $\pi_0 \in \Pi \setminus \{\pi^*\}$. Then, the optimal objective value is*

$$v(\pi^*) = \sup_{\pi\in\Pi\setminus\{\pi^*\}} \Big\{ u_{P_{\pi_0}}(\pi, \pi_0) - \frac{\lambda}{2}(b^0_{P_{\pi_0}}(\pi, \pi_0) - b^1_{P_{\pi_0}}(\pi, \pi_0))^2 \Big\}.$$

This shows that—unlike within deterministic threshold models—within exploring policies we can learn the optimal policy using only data from an induced distribution. Finally, we would like to highlight that not all exploring policies may be (equally) acceptable to society. For example, in lending scenarios without fairness constraints (i.e., $\lambda = 0$), it may appear wasteful to deny a loan with probability greater than zero to individuals who are believed to repay by the current model. In those cases, one may like to consider exploring policies that, given sufficient evidence, decide $d = 1$ deterministically, i.e., $\pi_0(d = 1 \mid \boldsymbol{x}, s) = 1$ for some values of $\boldsymbol{x}, s$. Other settings, like the criminal justice system, may call for a more general discussion about the ethics of non-deterministic decision making.

[4] π is exploring, iff the true distribution P is absolutely continuous w.r.t. the induced distribution P_π. This means the data collection distribution must not ignore regions where the true distribution puts mass. This condition does not strictly require randomness, but could be achieved by a pre-determined process, e.g., "$d = 1$ every n-th time".

4 LEARNING EXPLORING POLICIES

In this section, we exemplify Proposition 4 via a simple, yet practical, gradient-based algorithm to find the solution to eq. (5) within a (differentiable) parameterized class of exploring policies $\Pi(\Theta)$ using data gathered by a given, already deployed, exploring policy π_0. While our algorithm works for any differentiable class of exploring policies, here we consider two examples of exploring policy classes in particular. First, the *logistic policy*, which is given by $\pi_{\boldsymbol{\theta}}(d = 1 \mid \boldsymbol{x}, s) = \sigma(\boldsymbol{\phi}(\boldsymbol{x}, s)^\top \boldsymbol{\theta}) \in (0, 1)$, where $\sigma(a) := \frac{1}{1+\exp(-a)}$ is the logistic function, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$ are the model parameters, and $\boldsymbol{\phi} : \mathbb{R}^d \times \{0, 1\} \to \mathbb{R}^m$ is a fixed feature map. Second, the *semi-logistic policy*, which deterministically approves examples believed to contribute positively to the utility by the current model and only explores stochastically on the remaining ones, i.e., $\tilde{\pi}_{\boldsymbol{\theta}}(d = 1 \mid \boldsymbol{x}, s) = \mathbf{1}\left[\boldsymbol{\phi}(\boldsymbol{x}, s)^\top \boldsymbol{\theta} \geq 0\right] + \mathbf{1}\left[\boldsymbol{\phi}(\boldsymbol{x}, s)^\top \boldsymbol{\theta} < 0\right] \sigma\left(\boldsymbol{\phi}(\boldsymbol{x}, s)^\top \boldsymbol{\theta}\right)$.

We use stochastic gradient ascent (SGA) (Kiefer et al., 1952) to learn the parameters of the new policy, i.e., $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \alpha_i \nabla_{\boldsymbol{\theta}} v_P(\pi_{\boldsymbol{\theta}})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_i}$, where $\nabla_{\boldsymbol{\theta}} v_P(\pi_{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{\theta}} u_P(\pi_{\boldsymbol{\theta}}) - \lambda(b_0(\pi_{\boldsymbol{\theta}}) - b_1(\pi_{\boldsymbol{\theta}}))(\nabla_{\boldsymbol{\theta}} b_0(\pi_{\boldsymbol{\theta}}) - \nabla_{\boldsymbol{\theta}} b_1(\pi_{\boldsymbol{\theta}}))$, and $\alpha_i > 0$ is the learning rate at step $i \in \mathbb{N}$. With the reweighting from eq. (7) and the log-derivative trick (Williams, 1992), we can compute the gradient of the utility and the benefits as

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} u_P(\pi_{\boldsymbol{\theta}}) &= \mathbb{E}_{\substack{\boldsymbol{x},s,y\sim P_{\pi_0} \\ d\sim\pi_{\boldsymbol{\theta}}(\boldsymbol{x},s)}}\Big[\frac{d\,(y-c)\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}}{\pi_0(d=1\mid \boldsymbol{x},s)}\Big], \\ \nabla_{\boldsymbol{\theta}} b^s_P(\pi_{\boldsymbol{\theta}}) &= \mathbb{E}_{\substack{\boldsymbol{x},s,y\sim P_{\pi_0} \\ d\sim\pi_{\boldsymbol{\theta}}(\boldsymbol{x},s)}}\Big[\frac{f(d,y)\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}}{\pi_0(d=1\mid \boldsymbol{x},s)}\Big], \end{aligned} \tag{8}$$

where $\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}} := \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(d \mid \boldsymbol{x}, s)$ is the score function (Hyvärinen, 2005). Thus, our implementation resembles a REINFORCE algorithm with horizon one. A detailed derivation of the of the score functions and respective gradients can be found in appendix D.

Unfortunately, the above procedure has two main drawbacks. First, it may require an abundance of data from P_{π_0}, which can be unacceptable in terms of utility if π_0 is far from optimal. Second, if $\pi_0(d = 1 \mid \boldsymbol{x}, s)$ is small in a region where $\pi_{\boldsymbol{\theta}}$ often takes positive decisions, one may expect that an empirical estimate of the above gra-

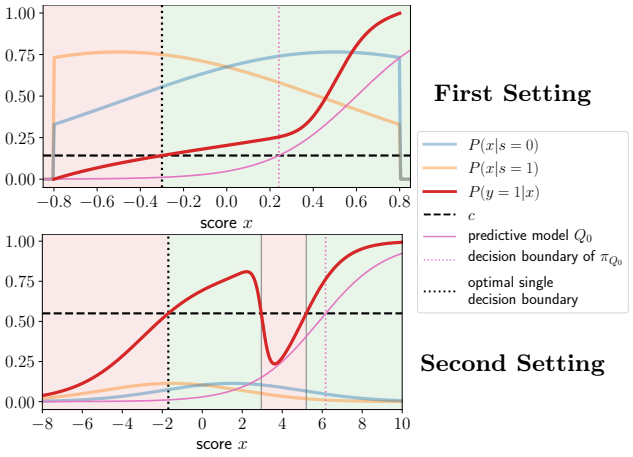

Figure 2: Two synthetic settings. In red, we show $P(y=1\,|\,x)$, where the score x is drawn from different distributions for the two groups (blue/orange). For given c (black, dashed), the optimal policy decides $d=1$ ($d=0$) in the shaded green (red) regions. The vertical black, dotted line shows the best policy achievable with a single threshold on x. In pink, we show a possible imperfect logistic predictive model and its corresponding (suboptimal) threshold in x.

dient will have high variance, due to similar arguments as in weighted inverse propensity scoring (Sutton & Barto, 1998). On the other hand, in most practical applications updating the model after every single decision is impractical. Typically, a fixed model will be deployed for a certain period, before it is updated.

To overcome these drawbacks, we build two types of sequences of policies $\{\pi_{\boldsymbol{\theta}_t}\}_{t=0}^{T}$: a) the *iterative sequence* $\pi_{t+1} := \mathcal{A}(\pi_t, \mathcal{D}^t)$ with $\mathcal{D}^t \sim P_{\pi_t}(\boldsymbol{x}, s, y)$, where only data gathered by the immediately previous policy are used to update the current policy; and b) the *aggregated sequence* $\pi_{t+1} := \mathcal{A}(\pi_t, \bigcup_{i=0}^{t} \mathcal{D}^i)$ with $\mathcal{D}^i \sim P_{\pi_i}(\boldsymbol{x}, s, y)$, where data gathered by all previous policies are used to update the current policy. The overall training procedure is shown in Algorithm 1. Note that in UpdatePolicy the input data $\mathcal{D}$ was collected under $\pi_{\boldsymbol{\theta}'}$. The decisions $d \sim \pi_{\boldsymbol{\theta}^{(j)}}$ are just sampled to implement SGA. The function Minibatch$(\mathcal{D}, B)$ samples a minibatch of size B from the dataset $\mathcal{D}$ and InitializePolicy() initializes the policy parameters. In appendix D we show in detail how to compute gradient estimates $\nabla_{\boldsymbol{\theta}} v(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(j)}}$. In Algorithm 1 we learn each policy π_t only using data from the previous policy π_{t-1}. This may readily be generalized to a mix of various previous policies $\pi_{t'}$ in eq. (9). Averaging multiple gradient estimators for several $t' < t$ is again an unbiased gradient estimator. To reduce variance, in practice one may consider recent policies $\pi_{t'}$ most similar to π_t.

Our weighted sampling closely relates to the concept

Algorithm 1 ConsequentialLearning: train a sequence of policies $\pi_{\boldsymbol{\theta}_t}$ of increasing $v_P(\pi_{\boldsymbol{\theta}_t})$.

Require: Cost c, time steps T, decisions N, iterations M, minibatch size B, penalty λ, learning rate α.

$\boldsymbol{\theta}_0 \leftarrow$ InitializePolicy()
for $t = 0, \ldots, T-1$ **do** ▷ time steps
  $\mathcal{D}^t \leftarrow$ CollectData$(\boldsymbol{\theta}_t, N)$
  $\boldsymbol{\theta}_{t+1} \leftarrow$ UpdatePolicy$(\boldsymbol{\theta}_t, \mathcal{D}^t, M, B, \alpha)$
return $\{\pi_{\boldsymbol{\theta}_t}\}_{t=0}^{T}$

function CollectData$(\boldsymbol{\theta}, N)$
  $\mathcal{D} \leftarrow \emptyset$
  for $i = 1, \ldots, N$ **do** ▷ N decisions
    $(\boldsymbol{x}_i, s_i) \sim P(\boldsymbol{x}, s)$ and $d_i \sim \pi_{\boldsymbol{\theta}}(\boldsymbol{x}_i, s_i)$
    if $d_i = 1$ **then** ▷ positive decision
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\boldsymbol{x}_i, s_i, y_i)\}$ with $y_i \sim P(y\,|\,\boldsymbol{x}_i, s_i)$
  return $\mathcal{D}$ ▷ data observed under $\pi_{\boldsymbol{\theta}}$

function UpdatePolicy$(\boldsymbol{\theta}', \mathcal{D}, M, B, \alpha)$
  $\boldsymbol{\theta}^{(0)} \leftarrow \boldsymbol{\theta}'$
  for $j = 1, \ldots, M$ **do** ▷ iterations
    $\mathcal{D}^{(j)} \leftarrow$ Minibatch$(\mathcal{D}, B)$ ▷ sample minibatch
    $\nabla \leftarrow 0$, $n_j \leftarrow 0$
    for $(\boldsymbol{x}, s, y) \in \mathcal{D}^{(j)}$ **do** ▷ accumulate gradients
      $d \sim \pi_{\boldsymbol{\theta}^{(j)}}(\boldsymbol{x}, s)$
      if $d = 1$ **then**
        $n_j \leftarrow n_j + 1$
        $\nabla \leftarrow \nabla + \nabla_{\boldsymbol{\theta}} v(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(j)}}$
    $\boldsymbol{\theta}^{(j+1)} \leftarrow \boldsymbol{\theta}^{(j)} + \alpha \frac{\nabla}{n_j}$
  return $\boldsymbol{\theta}^{M}$

of weighted inverse propensity scoring (wIPS), commonly used in counterfactual learning Bottou et al. (2013); Swaminathan & Joachims (2015a), off-policy reinforcement learning Sutton & Barto (1998), and contextual bandits Langford et al. (2008). However, a key difference is that, in wIPS, the labels y are always observed, which we elaborate on in appendix D. Despite this difference, we believe that recent advances in variance reduction for wIPS such as clipped-wIPS Bottou et al. (2013), self-normalized estimators Swaminathan & Joachims (2015b), or doubly robust estimators Dudík et al. (2011) may be applicable to our setting. This is left for future work. Finally, we opt for the simple SGA approach on (semi-)logistic policies over, e.g., contextual bandits algorithms, because it provides a direct and fairer comparison with commonly used prediction based decision policies (e.g., logistic regression), also often trained via SGA.

5 EXPERIMENTS

In our experiments, we learn a sequence of policies $\{\pi_{\boldsymbol{\theta}_t}\}_{t=1}^{T}$ using the following strategies:

Optimal: decisions are taken by the optimal deterministic threshold rule π^* given by eq. (2), i.e., $\pi_t = \pi^*$ for all t. It can only be computed when the ground truth conditional $P(y\,|\,\boldsymbol{x}, s)$ is known.

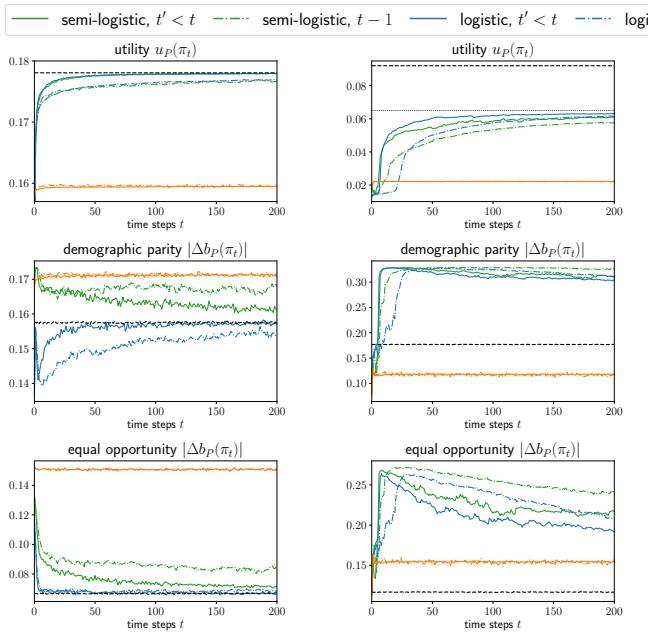

Figure 3: Utility, demographic parity and equality of opportunity in the synthetic settings of Figure 2 (first setting left, second setting right).

Deterministic: decisions are taken by deterministic threshold policies $\pi_t = \pi_{Q_t}$, where Q_t are logistic models maximizing label likelihood trained either in an iterative or aggregate sequence.
Logistic: decisions are taken by logistic policies $\pi_t = \pi_{\boldsymbol{\theta}_t}$ trained via Algorithm 1 either in an iterative or aggregate sequence.
Semi-logistic: decisions are taken by semi-logistic policies $\tilde{\pi}_t = \tilde{\pi}_{\boldsymbol{\theta}_t}$ trained via Algorithm 1 either in an iterative or aggregate sequence.

It is crucial that while each of the above methods decides over the same set of proposed $\{(\boldsymbol{x}_i, s_i)\}_{i=1}^N$ at each time step t, depending on their decisions, they may collect labels for differing subsets and thus receive different amounts of new training data. During learning, we record the following metrics:[5]

Utility: the utility $u_P(\pi_t)$ achieved by the current policy π_t estimated empirically on a held-out dataset, the *test set*, sampled i.i.d. from the ground truth distribution $P(\boldsymbol{x}, s, y)$. This is the utility that the decision maker would obtain if they deployed the current policy π_t at large in the population.
Fairness: the difference in group benefits between sensitive groups $|\Delta b_P(\pi)| = |b_P^0(\pi) - b_P^1(\pi)|$ for both disparate impact $(f(d, y) = d)$ and equal opportunity $(f(d, y) = d \cdot y)$. A policy satisfies the chosen fairness criterion iff $|\Delta b_P(\pi)| = 0$. Again, we estimate fairness empirically on the test set and thus measure the level of fairness π_t would achieve in the entire population.

[5]For readability we only show medians over 30 runs. Figures with 25 and 75 percentiles are in appendix E.

5.1 Experiments on synthetic data

We assume that there is a single non-sensitive feature $x \in \mathbb{R}$ per individual—similar to the lending example in Section 3—and a sensitive attribute $s \in \{0, 1\}$. While $P(x \mid s = 0) \neq P(x \mid s = 1)$, in our experiments the policies only take x as input, and *not* the sensitive attribute, which is only used for the fairness constraints. We consider two different settings, illustrated in Figure 2, where $s \sim \text{Ber}(0.5)$ and the distributions over x differ for the two groups, see appendix E. In the first setting, the conditional probability $P(y = 1 \mid x)$ is strictly monotonic in the score and does not depend on s, but is not well calibrated, i.e., not directly proportional to x. In the second setting, the conditional probability $P(y = 1 \mid x)$ crosses the cost threshold c multiple times resulting in two disjoint intervals for which the optimal decision is $d = 1$ (green areas).

Figure 3 summarizes the results for $\lambda = 0$, i.e., without fairness constraints. Our method outperforms prediction based deterministic threshold rules in terms of utility in both settings. This can be easily understood from the evolution of policies illustrated in Figure 9 in appendix E. In the first setting, exploring policies locate the optimal decision boundary, whereas the deterministic threshold rules get stuck, even though $P(y = 1 \mid x)$ is monotonic in x. In the second setting, our methods explore more and eventually identify the best single threshold at the black vertical dotted line in Figure 2. In contrast, non-exploring deterministic threshold rules converge to a suboptimal threshold at $x \approx 5$, ignoring the left green region.

In the first setting, we also observe that the suboptimal predictive models amplify unfairness beyond the levels exhibited by the optimal policy both in terms of demographic parity and equality of opportunity. For our approach, levels of unfairness are comparable to or even below those of the optimal policy. The second setting shows that depending on the ground truth distribution, higher utility can be directly linked to larger fairness violations. In such cases, our approach allows to explicitly control for fairness. Results on utility, demographic parity and equality of opportunity under fairness constraints with different λ are shown in Figures 10 and 11 in appendix E. In essence, λ trades off utility and fairness violations to the point of perfect fairness in the ground truth distribution.

5.2 Experiments on real data

Here, we use the COMPAS recidivism dataset compiled by ProPublica Angwin et al. (2016), which comprises of information about criminal offenders screened through the COMPAS tool in Broward County, Florida during 2013-2014. For each offender, the dataset contains a set of demographic features, the criminal history,

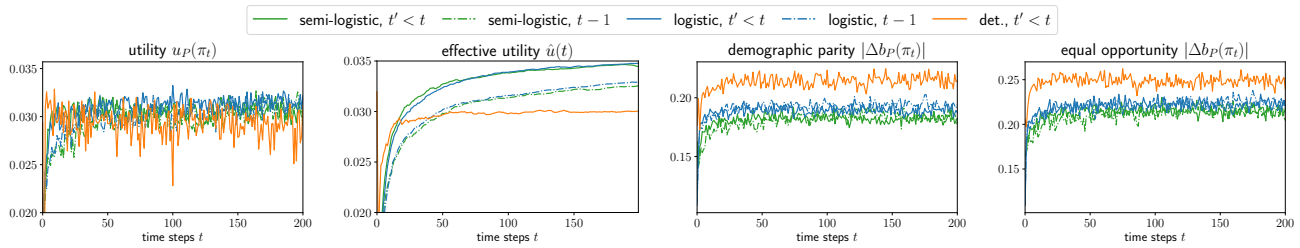

Figure 4: Training progress on COMPAS data for $\lambda = 0$, i.e., without fairness constraints.

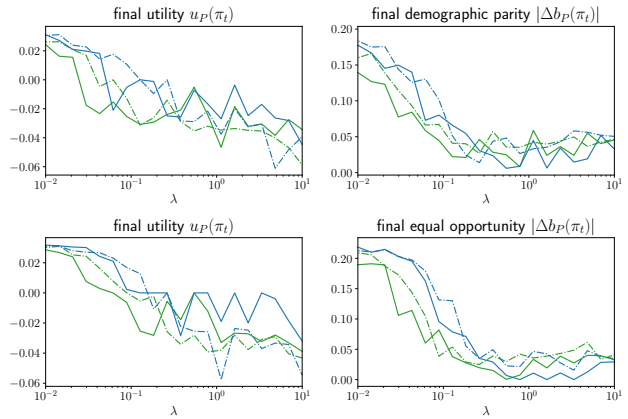

Figure 5: Fairness evaluation on COMPAS data for the final ($t = 200$) policy as a function of λ for demographic parity (top) and for equal opportunity (bottom).

and the risk score assigned by COMPAS. Moreover, ProPublica collected whether or not these individuals were rearrested within two years of the screening. In our experiments, $s \in \{0, 1\}$ indicates whether individuals were identified "white", y indicates rearrest, and $d \sim \pi(\boldsymbol{x}, s)$ determines whether an individual is let out on parole. Again, s is not used as an input. We use 80% of the data for training, where at each step t, we sample (with replacement) N individuals, and the remaining 20% as a held-out set to evaluate each learned policy in the population of interest.

We first summarize the results for $\lambda = 0$, i.e., without fairness constraints in Figure 4. A slight initial utility advantage of the deterministic threshold rule is quickly overcome by our exploring policies. This is best seen when looking at *effective utility*, the average utility accumulated by the decision maker on training data up to time t, for which our strategies dominate after $t = 50$. Hence, early exploration not only pays off to eventually be able to take better decisions, but also reaps higher profit during training. Moreover, all strategies based on exploring policies consistently achieve lower violations of both fairness metrics than the deterministic threshold rules. In summary, even without fairness constraints, i.e., in a pure utility maximization setting, exploring policies achieve higher utility and simultaneously reduce unfairness compared to deterministic threshold rules.

In Figure 5, we show how utility and demographic parity (equal opportunity) of the final policy $\pi_{t=200}$ change as a function of λ when constraining demographic parity (equal opportunity). As expected, while we are able to achieve low demographic parity (equal opportunity), this comes with a drop in utility. All remaining metrics under both constraints are shown in Figure 13 in appendix E. Finally, two remarks are in order. First, for real-world data we cannot evaluate the optimal policy and do not expect it to reside in our model class. However, even when logistic models do not perfectly capture the conditional $P(y = 1 \mid \boldsymbol{x})$, our comparisons here are "fair" in that all strategies have equal modeling capacity. Second, we take the COMPAS dataset as our (empirical) ground truth distribution even though it likely also suffered from selective labels. To learn about the real distribution underlying the dataset, we would need to actually deploy our strategy.

6 DISCUSSION

In this work, we have analyzed consequential decision making using imperfect predictive models, which are learned from data gathered by potentially biased historical decisions. First, we have articulated how this approach fails to optimize utility when starting with a faulty deterministic policy. Next, we have presented how directly learning to decide with exploring policies avoids this failure mode while respecting common fairness constraints. Finally, we have introduced and evaluated a simple, yet practical gradient-based algorithm to learn fair exploring policies.

Unlike most previous work on fairness in machine learning, which phrases decision making directly as a prediction problem, we argue for a shift from "learning to predict" to "learning to decide". Not only does this lead to improved fairness in this context, but it also establishes connections to other areas such as counterfactual inference, reinforcement learning and contextual bandits. Within reinforcement learning, it would be interesting to move beyond a static distribution P by incorporating feedback from decisions or non-static externalities. Moreover, since we have shown how shifting focus from learning predictions to learning decisions requires exploration, we hope to stimulate future research on how to explore ethically in different domains.

Acknowledgments

We thank Floyd Kretschmar for useful discussions regarding the implementation.

References

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1638–1646, Bejing, China, 2014. PMLR.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias: There is software used across the country to predict future criminals. and it is biased against blacks. *ProPublica, May*, 23, 2016.

Athey, S. and Wager, S. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

Bottou, L., Peters, J., nonero Candela, J. Q., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806. ACM, 2017.

Dimitrakakis, C., Liu, Y., Parkes, D., and Radanovic, G. Bayesian fairness. In *AAAI*, 2019.

Dudík, M., Langford, J., and Li, L. Doubly Robust Policy Evaluation and Learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104. Omnipress, 2011.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pp. 214–226. ACM, 2012.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. Decision making with limited feedback: Error bounds for recidivism prediction and predictive policing. *JMLR*, 2018.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268, 2015.

Gillen, S., Jung, C., Kearns, M., and Roth, A. Online learning with an unknown fairness metric. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 2605–2614. Curran Associates, Inc., 2018.

Hardt, M., Price, E., Srebro, N., et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.

Heidari, H. and Krause, A. Preventing disparate treatment in sequential decision making. In *IJCAI*, pp. 2248–2254, 2018.

Helmbold, D. P., Littlestone, N., and Long, P. M. Apple tasting. *Information and Computation*, 161(2):85–139, 2000.

Holstein, K., Vaughan, J. W., Daumé III, H., Dudík, M., and Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? *arXiv preprint arXiv:1812.05239*, 2018.

Hu, L. and Chen, Y. A short-term intervention for long-term fairness in the labor market. In *World Wide Web Conference*, WWW '18, pp. 1389–1398, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee.

Hyvärinen, A. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. Fairness in reinforcement learning. *arXiv preprint arXiv:1611.03071*, 2016.

Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pp. 325–333, 2016.

Jung, J., Shroff, R., Feller, A., and Goel, S. Algorithmic decision making in the presence of unmeasured confounding. *arXiv preprint arXiv:1805.01868*, 2018.

Kallus, N. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, pp. 8909–8920, 2018.

Kallus, N. and Zhou, A. Residual unfairness in fair machine learning from prejudiced data. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2439–2448, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Kiefer, J., Wolfowitz, J., et al. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. Human decisions and machine predictions. *The quarterly journal of economics*, 133 (1):237–293, 2017a.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent Trade-Offs in the Fair Determination of Risk Scores. In Papadimitriou, C. H. (ed.), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017b. ISBN 978-3-95977-029-3.

Lakkaraju, H. and Rudin, C. Learning Cost-Effective and Interpretable Treatment Regimes. In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 166–175, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., and Mullainathan, S. 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*, pp. 275–284. ACM, 2017.

Langford, J., Strehl, A., and Wortman, J. Exploration scavenging. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 528–535, New York, NY, USA, 2008. ACM.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3150–3158, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Meyer, M. N., Heck, P. R., Holtzman, G. S., Anderson, S. M., Cai, W., Watts, D. J., and Chabris, C. F. Objecting to experiments that compare two unobjectionable policies or treatments. *Proceedings of the National Academy of Sciences*, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1820701116.

Mitchell, S., Potash, E., and Barocas, S. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

Mouzannar, H., Ohannessian, M. I., and Srebro, N. From fair decision making to social equality. In *FAT*, 2019.

Pearl, J. *Causality*. Cambridge University Press, 2009.

Rubin, D. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100 (469):322–331, 2005.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 387–395. JMLR.org, 2014.

Sutton, R. S. and Barto, A. G. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.

Swaminathan, A. and Joachims, T. Counterfactual risk minimization: Learning from logged bandit feedback. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 814–823. JMLR.org, 2015a.

Swaminathan, A. and Joachims, T. The self-normalized estimator for counterfactual learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pp. 3231–3239, Cambridge, MA, USA, 2015b. MIT Press.

Tabibian, B., Gomez, V., De, A., Schoelkopf, B., and Gomez-Rodriguez, M. Consequential ranking algorithms and long-term welfare. *arXiv preprint arXiv:1905.05305*, 2019.

Valera, I., Singla, A., and Gomez-Rodriguez, M. Enhancing the accuracy and fairness of human decision making. In *Neural Information Processing Systems*, 2018.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. In Kale, S. and Shamir, O. (eds.), *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1920–1953, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.