HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230083

The Equity Framework: Fairness Beyond Equalized Predictive Outcomes

KEZIAH NAGGITA [a,1] and J. CEASAR AGUMA [b]
[a] *Toyota Technological Institute at Chicago,* knaggita@ttic.edu
[b] *University of California at Irvine,* jaguma@uci.edu

Abstract. Machine Learning (ML) decision-making algorithms are now widely used in predictive decision-making, for example, to determine who to admit and give a loan. Their wide usage and consequential effects on individuals led the ML community to question and raise concerns on how the algorithms differently affect different people and communities. In this paper, we study fairness issues that arise when decision-makers use models (*proxy models*) that deviate from the models that depict the physical and social environment in which the decisions are situated (*intended models*). We also highlight the effect of *obstacles* on individual access and utilization of the models. To this end, we formulate an Equity Framework that considers equal access to the model, equal outcomes from the model, and equal utilization of the model, and consequentially achieves equity and higher social welfare than current fairness notions that aim for equality. We show how the three main aspects of the framework are connected and provide questions to guide decision-makers towards equitable decision-making. We show how failure to consider access, outcome, and utilization would exacerbate proxy gaps leading to an infinite inequity loop that reinforces structural inequities through inaccurate and incomplete ground truth curation. We, therefore, recommend a more critical look at the model design and its effect on equity and a shift towards equity achieving predictive decision-making models.

Keywords. Fairness in ML, Proxy gaps, Ethical ML

1. Introduction

Machine Learning (ML) algorithms are now heavily used in decision-making to determine who to hire, admit to a college, give a loan, among other decisions that have real-life consequences on people. Automated decision-making allows us to track decision-making and audit systems for fairness. However, since these algorithms use data already affected by systematic inequalities, they reinforce and sometimes exacerbate these inequalities (1, 2). For example, language models reveal encoded stereotypes (3, 4, 5) which can lead to unfair decision-making (6, 7). In addition, Ensign et al. (8) shows that predictive policing algorithms trained on biased historical data lead to more policing in those areas as more misleading crime data indicates a need for more policing. Consequently, disadvantaged groups in low-income communities are more likely to be incarcerated and or denied bail (9, 10). These issues have inspired the ML fairness community to define,

[1] Corresponding Author: knaggita@ttic.edu and jaguma@uci.edu

design, and implement algorithmic fairness notions and techniques to tackle the introduction, propagation, and amplification of human biases in automated decision-making systems.

Prior work on fairness in ML has focused on equalizing outcomes across groups and modeled fairness violations based on the similarity between individuals and protected groups (e.g, (11, 12)). Models focused on equalized predictive outcomes sometimes acknowledge biases in historical data (13, 14, 15) used in decision-making. However, historical structural inequalities like wealth, housing, and education inequities create different *obstacles* for people, sometimes even within the same protected groups. These obstacles present barriers that prevent full access and utilization of the models. The focus on only equalized predictive outcomes ignores and sometimes further excludes already marginalized[2] groups.

Although several ML fairness research, for example, causal and Bayesian inference (e.g., (16, 17, 18, 19)), fair decision-making (e.g., (20, 21, 22)) acknowledge obstacles individuals face, the focus is mainly on changing decision-making models, for example, to qualify obstacle-refrained individuals. We, however, argue that this only helps in the short and not the long term because unalleviated obstacles individuals face in accessing the model resurface in the model utilization, in which decision-makers evaluate individuals on how well they utilize the model as a form of feedback to the decision-makers. In addition, since decision-makers use models (*proxy models*) that deviate from the social and physical environment in which the decisions are situated (*intended models*), inequality in utilization increases as individuals are likely to face unforeseen obstacles unalleviated during decision-making. This discrepancy increases the chances of curating incomplete, inaccurate, and skewed ground truth.

Example 1.1. (Pretrial/Bail Assessment Model) First, most bail models use a model function like Public Safety Assessment (PSA) algorithm (23) to determine who to give or deny bail. The function uses features: *"age at current arrest", "current charges", "pending charges", "prior misdemeanor conviction", "prior felony conviction", "prior violent conviction", "prior failure to appear in past two years", "record of failing to appear"*, and *"record sentences to incarceration"*. An individual facing obstacles like lack of good representation might not have an invested lawyer to accurately layout/dispute current criminal charges and imposed criminal history, for example, when they add the defendant's parents' charges to their sentences (24). Another might live in a neighborhood with higher chances of rearrest, thus leading to high accumulative crime history. Often, when people are released, some have to go to halfway houses or get a probation or parole officer (PPO) residing in the same place they got convicted, and or in poor neighborhoods surrounded by crime (25, 26, 27) and a higher likelihood of false arrests due to frequent patrol (8). These obstacles disproportionately affect some groups than others, and thus leading to a high level of unequal access to the bail model. Therefore, decision-makers should ask themselves: *with this choice of predictive features: current charges, criminal history, and record of failing to appear, among others who is more(less) likely to access this model? Who is more likely to be gain(lose) by it?*

[2] We use the terms marginalized and disadvantaged interchangeably to mean individuals facing relatively high levels of obstacles that deter them from full access and utilization of the decision-making models due to historical inequities. The term advantaged refers to individuals who don't face or face few obstacles to access or utilize the decision-making models.

Second, to determine whether someone is a true or false positive, the bail model relies on whether someone appears for their trial in court. However, the proxy model used doesn't capture features that might determine whether someone reappears, for example, *"support system", "financial stability", "occupation", "living conditions"*, and *"working internet-connected mobile phone"*. However, an intended model with a model function *"likely to reappear in court"* and those features would have been a better predictor. Using the intended model function and features that depict the social and physical environment that decisions are situated helps highlight obstacles that might lead to individuals not fully utilizing the model and appearing as false positive. For example, insufficient support from PPOs due to heavy case-loads (28), non-flexible jobs with no time off, or even being too busy to remember the court date (23, 29) might be highlighted and alleviated.

Therefore, while lots of work has gone into understanding decision-making systems and making them fairer, for example, the push towards no money bail (30, 31), especially for misdemeanors, we highlight how harmful mismanaged delayed evaluation models can be. We show how the difference between proxy models and the intended models that depict the physical and social environments that the decisions are situated differently affects different people and factors tremendously in the curation of future proxy model ground truth. When ignored and misunderstood, the proxy gap makes it harder to diagnose problems accurately, find equitable solutions, and implement policies that help the most disadvantaged.

Our Contributions are summarized as follows:

1. We show how obstacles, a result of historical inequities, create the implicit and explicit barriers that deter individuals from accessing and fully utilizing models (sec.2 and sec.3). With an example, we show that it's possible to achieve equal outcomes without equal access, which further marginalizes already marginalized groups, thus reinforcing structural inequities.
2. We differentiate immediate evaluation models from delayed ones and show how the proxy gap between proxy and intended models negatively affects model utilization, future ground truth curation, and long-term fairness. We supplement our mathematical formulations with examples (sec.2 and sec.3).
3. We define the Equity Framework motivated by Almond (32) and connect all three of its aspects; equal access, equal outcomes, and equal utilization. Lastly, we provide curated questions to guide decision-makers toward equitable decision-making (sec.4).

A key insight from our work is that equal outcomes as a fairness notion is not enough to address disparities arising from the discrepancy between the proxy and intended models in delayed evaluation systems and different obstacles individuals face to access and utilize models. Generally, our model does a better job of finding inequitable aspects of the model and setting a path to more equitable decision-making than fair outcome-based predictive models. We, therefore, argue for a shift in fairness literature to consider fairness beyond predictive model outcomes. We hope our work pivots the discussion on fair predictive decision-making from equality to equity and becomes a yardstick for decision-makers to check their models for equitable decision-making.

Related Work To achieve equalized fair decision-making, researchers have proposed several metrics under the umbrella of group fairness and individual fairness. Group fairness ensures positive or negative parity in the treatment of protected groups (11, 33), and individual fairness ensures similar treatment of similar individuals per the decision-making model (34). However, historical structural inequalities like wealth, housing, and education inequities create *obstacles* which present barriers that prevent full utilization and access to models. Focusing on only equalized predictive outcomes ignores and sometimes further marginalizes already historically excluded groups.

Some researchers have highlighted the issues of access disparities among populations. For example, strategic classification (e.g, (35, 36, 37, 38, 39, 40, 41, 42)) shows how people's reactions to the decision-making model highlights the differences in access through costs people incur. Although our framing of obstacles is quite analogous to budget framing in strategic classification, we formulate obstacles more generally, and the alleviation of obstacles strictly makes things better, that is to say, the obstacle-free feature values, $\mathbf{z}$ dominate the obstacle-refrained feature values $\mathbf{x}$ and the obstacle-free label y' is not necessarily equal to the obstacle-refrained label, y. Relatedly, another body of work that highlights obstacles individuals face is causal and Bayesian inference (e.g., (16, 17, 18, 19)). However, instead of ensuring equal access, the focus is mainly on redefining decision-making models, by, for example, changing accuracy metrics, weights of different features, or features used, among other interventions to qualify obstacle-refrained individuals. In our work, we show that while redefining decision-making models might grant the obstacle-refrained individuals a foot in the door, these obstacles resurface in the evaluation phase, leading to further marginalization of the disadvantaged groups.

Just like one of the main aspects of the Equity Framework, Rawls (43) asserts that to achieve fairness, individuals with the same level of talent, ability, and willingness to use those gifts should have a fair chance at reaching desirable positions without the obstacles of their social class and background. While the concept of equity in machine learning is relatively new and mostly foreign, equity has been a goal for many reforms in justice (44), health (32), and education (45) systems. Recently, however, the concept has started taking shape and has drawn a few researchers to it. For example, Kasy and Abebe (46) takes the causal perspective to show that while predictive parity might achieve fairness across groups, it might perpetuate inequality within groups and legitimize the status quo. In doing so, Kasy and Abebe (46) brings forward new questions on the power distributions between and within groups as it relates to decision-making systems. Suresh and Guttag (47)'s major contribution is a framework for identifying biases that arise due to the historical context within which the ML development pipeline is situated. Similar to ours, their (Suresh and Guttag (47)) framework, a push towards more equitable decision-making highlights the downstream harms and how they manifest in model building, evaluation, and deployment processes.

In addition, similar to our work, Hoffmann (48), Green (49) and Davis et al. (50) highlight the failure of the current algorithmic fairness discourse in addressing the historical factors and consequential distributions that led to unfairness in the first place. We complement this critical work by formalizing the divergence from the models decision-makers use and those that depict the physical and social environment in which the decisions are situated, and providing a framework to diagnose a decision-making system for potential equity failures at different stages, before deployment (equal access), during deployment (equal outcomes), and after deployment (equal utilization). More closely related to our work is Mehrabi et al. (51), who attempts to formalize equity by equalizing the sum of historical plus future outcomes of one group to another to compensate for observed historical biases in the data. However, unlike Mehrabi et al. (51), we focus on ensuring obstacles caused by historical biases are alleviated to achieve equal access to the model. In our work, motivated by Almond (32), who defines equity as insurance that barriers that could deter some people from having full access to the available resources are alleviated, we argue that if obstacles are not alleviated, they resurface in utilization, thus widening the inequity gap.

Apart from forgetting to check for and ensure access, decision-makers more often than not use proxy models instead of intended decision-making models. This is mainly due to bounded rationality (52), past experiences and biased view of the world (53, 54, 55, 56), difficulty articulating what they want to measure (57, 58), among others. The ignorance of the discrepancy between proxy and intended models increases unequal utilization and curation of biased ground truth.

Several researchers have studied the origin, diversity, and accuracy of ground truth data used in decision-making and its effect on predictions and different populations. For example, Buolamwini and Gebru (59) showed that bias in ground truth was one of the leading causes of higher error rates on dark-skinned women than other protected groups. To better understand the origin of ground truth, Gebru et al. (60) formulated a way for researchers to ensure responsible data collection and documentation. Jacobs and Wallach (61), Jaton (62), Søgaard et al. (63), Cabitza et al. (64) and Chehdi and Cariou (65) have contested and questioned the validity of ground truth, and Aka et al. (66) proposed models of measuring bias in classification independent of ground truth. Our work adds to this body of work to show how the gap between proxy and intended model leads to faulty, inaccurate, and incomplete ground truth curation. We also provide a way to measure these proxy gaps and present a novel idea of an Equity Framework whose aim is to ensure equal model access, outcome, and utilization and facilitate efficient auditing of deployed models for equity.

2. Preliminaries

To understand the main attributes of the Equity Framework; access, outcomes, and utilization of the model, we provide an overview on obstacles and proxy gaps.

2.1. Obstacles

In this paper, we define *obstacles* as the implicit and explicit barriers that deter individuals from effectively interacting with the decision-making model. We assume that the decision-maker uses features as inputs to their models, and barriers implicitly and explicitly affect these inputs. An individual with feature representation $\mathbf{z} \in \mathbb{R}^d$, will instead have feature representation $\mathbf{x} \in \mathbb{R}^d$ due to obstacles faced when interacting with the decision-making model. For example, a work-study student with knowledge of test scores factoring greatly in their performance is less likely to prepare adequately or attend enough discussions and will have feature representation $\mathbf{x}$ instead of $\mathbf{z}$ due to lack of ample time. The student doesn't show up to the model as they could have,

had their obstacles been alleviated. We say that **z** **dominates** **x**, that is $\mathbf{z} \succ \mathbf{x}$ because, $\forall i \in [d], \mathbf{z}_i \geq \mathbf{x}_i$ and $\exists i \in [d], \mathbf{z}_i > \mathbf{x}_i$.
Mathematically, we represent obstacles as: $\mathscr{O}(\mathbf{x},\mathbf{z}) = \langle \alpha, \mathbf{z}-\mathbf{x} \rangle \in \mathbb{R}_{\geq 0}$ where $\alpha \in \mathbb{R}^d$ is how much the implicit and explicit barriers constrain an individual from effectively interacting with the decision-making model. To improve an individual's access to the model means alleviating these obstacles, $\mathscr{O}(\mathbf{x},\mathbf{z})$, such that the likelihood, $P(\mathscr{Y}|\mathscr{X})$ of effective interaction with the model increases $P(\mathscr{Y}=1|\mathscr{X}=\mathbf{z}) \geq P(\mathscr{Y}=1|\mathscr{X}=\mathbf{x})$.

The assumption that alleviating the obstacles leads to **z** dominating **x** and improves the likelihood of desirable prediction might not capture all the complexities of inequities effects on individuals and relationships of features to labels. Future works could explore a more generalized formulation. Additionally, within any decision-making system, there could be several obstacles to the access and utilization of the model, most specific to the decision to be made. We assume a finite number of obstacles that directly affect access to and utilization of any given decision-making system. We imagine to be obstacles a decision-maker has full knowledge of and can create policies to alleviate said obstacles.

2.2. *Proxy gap*

In some ML predictive decision-making systems, evaluation is immediate and straightforward, e.g., after predicting a cat from the image, the ML decision-maker only has to compare the picture with the prediction to know if it's a true or false positive. However, in most systems, evaluation is not immediate and often done with a separate model, e.g., in tasks like predicting loan worthiness, the decision-maker only knows who is a true or false positive after the decisions take form in the social and physical environment. We therefore, define the *proxy model* as the model decision-makers use for making decisions and the *intended model* as the evaluation model that depicts the physical and social environment in which the decisions take form. Because these two models often diverge, we measure the gap between them. In this work, we define the *proxy gap* as the discrepancy between the proxy model and the intended model. We consider three kinds of proxy gaps: *feature proxy gap* measures the discrepancy between proxy features and intended features, *label proxy gap* measures the discrepancy between proxy label function[3] and the intended label function, and *obstacle proxy gap* measures the discrepancy between access obstacles and utilization obstacles.

3. Attributes of the Equity Framework

In this section we give a detailed description of the attributes of the Equity Framework —equal access, outcome, and utilization—, their formulations and how they are related.

3.1. *Equal Access*

If a decision-making model, f, is such that individuals face obstacles, $\mathscr{O}_f(\mathbf{x},\mathbf{z})$, that hinder their effective interaction with the model, it becomes an equal access model when

[3] For language simplicity, we sometimes use "label" instead of "label function" or "model function". The term (decision-making) model defines a decision-making system that takes in features and obstacles, has a set of policies to alleviate the obstacles individuals face to access and utilize the model, and has a class of label functions that output decisions.

obstacles all individuals face are alleviated. The model's policy, Φ, ensures that the obstacles each of the individual faces are reduced to 0. Therefore, Φ is defined as,

$$\Phi(\mathscr{O}_f(\mathbf{x},\mathbf{z}),\delta) = \max\{\mathscr{O}_f(\mathbf{x},\mathbf{z}) - \delta, 0\}$$

Here $\delta \geq 0$ is the decision-maker's resource budget for alleviating obstacles individuals face, where $\mathscr{O}_f(\mathbf{x},\mathbf{z}) \ll \delta$ implies a surplus of resources.

Definition 3.1. (Model access) Given a model, f, assume each individual, i, faces obstacles $\mathscr{O}_{f,i}$ to access the model. An individual achieves full access to the model if they either face no obstacles to access the model, $\mathscr{O}_{f,i}(\mathbf{x}_{f,i},\mathbf{z}_{f,i}) = 0$ or if the model checks for and alleviates all the obstacles an individual, i ,faces, $\Phi_f(\mathscr{O}_f(\mathbf{x}_{f,i},\mathbf{z}_{f,i}),\delta) = 0$. Model access based on whether individuals have full access to the model is defined as,

$$\Psi(f) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{\Phi_f(\mathscr{O}_{f,i}(\mathbf{x}_{f,i},\mathbf{z}_{f,i}),\delta)=0\}}}{n}$$

Therefore, if $\Psi(f) = 1$, the model, f, has **equal access**. Additionally, during decision-making, the individual (removed i to avoid notation clutter) reveals themselves as $\mathbf{x}_f^{\text{rev}}$ and the decision-maker sees them as y_f^{rev}, where,

$$(\mathbf{x}_f^{\text{rev}}, y_f^{\text{rev}}) = \begin{cases} (\mathbf{z}_f, y_f^{'}), \text{ if } \mathscr{O}_f(\mathbf{x}_f,\mathbf{z}_f) = 0 \text{ or } \Phi_f(\mathscr{O}_f(\mathbf{x}_f,\mathbf{z}_f),\delta) = 0 \\ (\mathbf{x}_f, y_f), \text{ o/w} \end{cases}$$

Example 3.1. Individuals in different domains face different obstacles to access the models in those domains. For example, participant 8 in a subject study conducted by Mercer et al. (67) to assess the impact of term-time employment on students, said "I get less time to focus on my assignments and to do my reading and prepare for my lectures." Because they simultaneously work and study, they don't perform as they should have if they had a scholarship or financial aid (policy).

3.2. Equal Outcomes

Fairness in predictive outcomes has been well covered in the algorithmic fairness literature. In this paper we adopt equalized odds (EO) (68) fairness metric as a measure of model outcomes. Other parity measures for example, demographic parity (69), accuracy parity (70) can be used.

Consider a binary classification setting and population distribution, $\mathscr{D}$ of size n. Each individual with protected group membership $grp \in \mathscr{G} = \{0,1\}$ reveals features $\mathbf{x}^{\text{rev}} \in \mathscr{X}^{\text{rev}} \subseteq \mathbb{R}^d$, and has a label $y^{\text{rev}} \in \mathscr{Y}^{\text{rev}} = \{0,1\}$. To make a prediction, a decision maker employs a model function, $h : \mathscr{X}^{\text{rev}} \rightarrow \mathscr{Y}^{\text{rev}}$ that generalizes with minimal accuracy loss, $\min_{h\in\mathscr{H}} L(h,\mathscr{D})$ where $\mathscr{H}$ is the set of all possible model functions and $L(h,\mathscr{D})$ is the loss function. Throughout this paper, for simplification and WLOG, we assume that there is one protected feature. We define EO violation as $[P(h(\mathbf{x}^{\text{rev}}) = 1|y^{\text{rev}} = 1,\ grp = 0) - P(h(\mathbf{x}^{\text{rev}}) = 1|y^{\text{rev}} = 1,\ grp = 1)] + [P(h(\mathbf{x}^{\text{rev}}) = 1|y^{\text{rev}} = 0,\ grp = 0) - P(h(\mathbf{x}^{\text{rev}}) = 1|y^{\text{rev}} = 0,\ grp = 1)]$

Definition 3.2. (Model outcomes) We define the outcomes of the model, f as $\Omega(f) =$ EO violation. If the EO violation $= 0$, the model outcomes, $\Omega(f) = 1$, indicating that the model achieves **equal outcomes**.

Example 3.2. An example of a model with unequal outcomes is the Propublica model in which the model function in the Prater-Borden case (71) was "particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants" (71).

3.2.1. Model outcomes and access

To illustrate how model access relates to model outcomes, let's assume we have two individuals, a and b, and two models: an equal access model, f', and an unequal access model, f'', whose label functions respectively are, $h'(\mathbf{x}^{\text{rev}}) = 1$, if $\|\mathbf{x}^{\text{rev}}\| \geq 5.5$, and 0 otherwise and $h''(\mathbf{x}^{\text{rev}}) = 1$, if $\|\mathbf{x}^{\text{rev}}\| \geq 6$, and 0 otherwise. Person a faced with obstacles to access f' and f'' and b not faced with obstacles to access both models.

With the unequal access model, a reveals $\mathbf{x}_a = (5,0)$, and is predicted as $h''(\mathbf{x}_a) = 0$, and with the equal access model, a reveals $\mathbf{z}_a = (6,0)$ and $h'(\mathbf{z}_a) = 1$. On the other hand, b doesn't face obstacles to present themselves to both models, and their feature values $\mathbf{x}_b = \mathbf{z}_b = (6,0)$. Person b is predicted as $h'(\mathbf{z}_b) = 1$ and $h''(\mathbf{x}_b) = 1$. From the example, since both model functions rely on information they receive, they achieve equal outcomes even though f'' doesn't ensure equal access.

The example shows that for any equal access model f', there is an unequal access and equal outcome model f'', such that; $\Psi(f'') < \Psi(f')$ and $\Omega(f'') = \Omega(f')$. Therefore, with equality fairness notions like equalized odds, the model function because it relies on received data can appear to achieve fairness when it's actually exacerbating inequity through unequal access. This setting further marginalizes disadvantaged groups, and the problem is decision-makers don't even realize the gravity of the error. We, therefore, urge the ML fairness community to check for equal access before equal outcomes.

3.3. Equal Utilization

Setting Assume a given proxy model, P, with a model function, $h_P \in \mathscr{H}_P : \mathscr{X}_P^{\text{rev}} \to \mathscr{Y}_P^{\text{rev}}$, is trained on proxy features $\mathbf{x}_P^{\text{rev}} \in \mathscr{X}_P^{\text{rev}} \subseteq \mathbb{R}^d$ and proxy target labels $y_P^{\text{rev}} \in \mathscr{Y}_P^{\text{rev}} = \{0,1\}$. Assume the trained model function achieves maximal accuracy with minimal loss, $\min_{h_P \in \mathscr{H}_P} L(h_P, \mathscr{D}_P)$, where $L(\cdot)$ is the loss function. The best trained proxy model function, h_P, is then tested in the wild on new arriving individuals. The arriving individuals have similar proxy feature variables as those considered in training, $\mathscr{X}_{PT}^{\text{rev}}$. Individuals are faced by access obstacles $\mathscr{O}_P : \mathscr{X}_P \times \mathscr{X}_P \to \mathbb{R}_{\geq 0}$, and depending on whether the proxy model has equal access or not determines $\mathbf{x}_{PT}^{\text{rev}} \in \mathscr{X}_{PT}^{\text{rev}}$ as described in Section 3.1. The trained proxy model, h_P, then predicts them as $y_{PT} \in \mathscr{Y}_{PT}^{\text{rev}}$.

Let all the individual qualified by the proxy model be in $\mathscr{B}$, such that, $\forall i \in \{1,\cdots,m\}$, $y_{PT,i}^{\text{rev}} = 1$, where $m = |\mathscr{B}|$. To evaluate full utilization, we assume a pre-existing intended model, T with trained model function $h_T \in \mathscr{H}_T : \mathscr{X}_T^{\text{rev}} \to \mathscr{Y}_T^{\text{rev}}$, trained on intended features $\mathscr{X}_T^{\text{rev}} \subseteq \mathbb{R}^D$ and intended target labels $\mathscr{Y}_T^{\text{rev}} = \{0,1\}$. To evaluate how well individuals in $\mathscr{B}$ utilize the model, decision-makers have to wait for the decisions to take form in the physical and social environment. In attempting to utilize the model, individuals might face utilization obstacles $\mathscr{O}_T : \mathscr{X}_T \times \mathscr{X}_T \to \mathbb{R}_{\geq 0}$. Depending

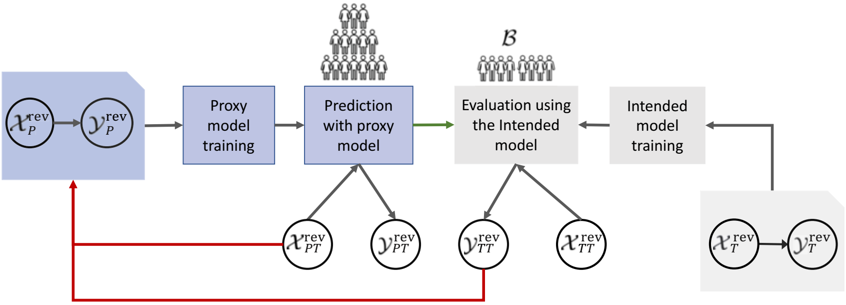
$\mathcal{X}_P^{\text{rev}}$
$\mathcal{Y}_P^{\text{rev}}$
Proxy model training
Prediction with proxy model
$\mathcal{B}$
Evaluation using the Intended model
Intended model training
$\mathcal{X}_{PT}^{\text{rev}}$
$\mathcal{Y}_{PT}^{\text{rev}}$
$\mathcal{Y}_{TT}^{\text{rev}}$
$\mathcal{X}_{TT}^{\text{rev}}$
$\mathcal{X}_T^{\text{rev}}$
$\mathcal{Y}_T^{\text{rev}}$

Figure 1.: The explicit proxy decision-making model (light blue) is trained on ground truth and tested in the wild. All positively classified individuals from the proxy model are evaluated for utilization using the implicit intended model (light grey). Decision-makers ignorant of the proxy gap, use the utilization feedback $\mathscr{Y}_{TT}^{\text{rev}}$ and proxy features $\mathscr{X}_{PT}^{\text{rev}}$ to curate ground truth (red lines).

on whether the intended model alleviates utilization obstacles or not determines how individuals in $\mathscr{B}$ reveal themselves to the intended model as described in Section 3.1. Individuals in $\mathscr{B}$ reveal themselves as $\mathbf{x}_{TT}^{\text{rev}} \in \mathscr{X}_{TT}^{\text{rev}}$ whose variables are similar to $\mathscr{X}_{T}^{\text{rev}}$ and the intended model function then predicts them as $y_{TT} \in \mathscr{Y}_{TT}^{\text{rev}}$.

Definition 3.3. (Model utilization) A proxy model qualified individual, $i \in [m]$, achieves full model utilization if and only if they remain qualified in the physical and social environment in which the model takes form. That is to say, $y_{PT,i}^{\text{rev}} = y_{TT,i}^{\text{rev}} = 1$. Model utilization $\zeta(f)$, is then defined as

$$\zeta(f) = \frac{\sum_{i=1}^{m} \mathbb{1}_{\{y_{PT,i}^{\text{rev}} = y_{TT,i}^{\text{rev}}\}}}{m}$$

A model, f, therefore has **equal utilization** if $\zeta(f) = 1$.

Example 3.3. First, Turiel and Aste (72) separated loan-worthy from non-loan worth individuals. They then used the intended model to determine if loan-worthy individuals $\forall i \in \{1, \cdots, m\}$, $y_{PT,i}^{\text{rev}} = 1$ were true positive (paid back the loan) or false positive (defaulted on the loan). Analysis showed that $15\% - 20\%$ didn't fully utilize the model, that is to say, they defaulted on the loan and their y_{TT}^{rev} was 0.

Ground truth curation Other than higher chances of unequal utilization, one of the biggest spillover effects of ignorance of the discrepancy between intended model and proxy model in delayed evaluation models is incorrect and incomplete ground truth curation. Decision-makers oblivious to the intended model and it's influence on how individuals predicted positive by the proxy model utilize the model, use the proxy model features $\mathscr{X}_{PT}^{\text{rev}}$ and intended model labels $\mathscr{Y}_{TT}^{\text{rev}}$ to curate ground truth as shown with red lines in Figure 1. Because of this, the next trained proxy model is trained on incomplete ground truth that falsely portrays obstacle restrained individuals, doesn't capture true feedback from the model deployment in the wild, and consequently leads to inequities.

3.3.1. Feature, label and obstacle proxy gaps

For readability, we eliminate the superscript rev from the definitions of the proxy gaps.

Feature proxy gap The feature proxy gap, $\Gamma_X(\mathscr{X}_P, \mathscr{X}_T) \in \{0,1\}^D$, where D is the number of intended features, defines how different the proxy model features variables are from the intended model feature variables.

$$\text{Therefore, } \forall i \in [D],\ \Gamma_X(\mathscr{X}_P, \mathscr{X}_T)_i = \begin{cases} 0, \text{ if for } i, \exists j : \mathscr{X}_T[i] \equiv \mathscr{X}_P[j] \\ 1, \text{ o/w} \end{cases}$$

When $\Gamma_X(\mathscr{X}_P, \mathscr{X}_T) = (0, \cdots, 0) \in \mathbb{R}^D = \mathbf{0}_D$, the feature proxy gap is non-existent.

Example 3.4. (High feature proxy gap) Assume the proxy model function uses features $\{a,b,c\}$ and intended model function uses featured $\{u,v,w\}$. Now, if feature u is such that none of the features $\{a,b,c\}$ is exactly similar to u, then $\Gamma_X(\mathscr{X}_P, \mathscr{X}_T)_1 = 1$. Similarly, if the same applies to features v and w, then the feature proxy gap $\Gamma_X(\mathscr{X}_P, \mathscr{X}_T) = (1,1,1)$, indicating that none of the features used in the proxy model is exactly similar to those in the intended model. For example, assume a bail model determining whether or not defendant reappears in court uses proxy model features "*criminal history*", "*current crime*", and "*age at arrest*" and the intended model features "*job*", "*support system*", "*financial stability*" will have a feature proxy gap $\Gamma_X(\mathscr{X}_P, \mathscr{X}_T) = (1,1,1)$.

Label proxy gap To compute the label proxy gap we first compute the feature importance scores for each model's features. Assume h_P and h_T are the same class of functions, for example both are logistic regression models. Let ω_P be the feature importance scores of proxy model function h_P, and let ω_T be the feature importance scores of intended model function h_T. The label proxy gap is then $\Gamma_L(h_P, h_T) \in \mathbb{R}^D$, where D is the number of intended features. Therefore; $\forall i \in [D]$,

$$\Gamma_L(h_P, h_T)_i = \begin{cases} \omega_{T,i} - \omega_{P,j}, \text{ if for } i, \exists j : \mathscr{X}_T[i] \equiv \mathscr{X}_P[j] \ \& \ \text{sign}(\omega_{T,i}) = \text{sign}(\omega_{P,j}) \\ \omega_{T,i}, \text{ o/w} \end{cases}$$

When $\Gamma_L(h_P, h_T) = (0, \cdots, 0) \in \mathbb{R}^D = \mathbf{0}_D$, then the label proxy gap is non-existent. While the difference between feature importance scores might be the best indication of the discrepancy of the value of the different features to the proxy and intended models, our formulation might not generalize well to all model settings.

Claim 3.1. If $\Gamma_X(\mathscr{X}_P, \mathscr{X}_T) \neq \mathbf{0}_D$, then $\Gamma_L(h_P, h_T) \neq \mathbf{0}_D$ and $\Gamma_X(\mathscr{X}_P, \mathscr{X}_T) = \mathbf{0}_D$ does not imply $\Gamma_L(h_P, h_T) = \mathbf{0}_D$.

Proof. From the definitions of the feature proxy gap, $\Gamma_X(\mathscr{X}_P, \mathscr{X}_T)$, and label proxy gap, $\Gamma_L(h_P, h_T)$, we can see that if the $\Gamma_X(\mathscr{X}_P, \mathscr{X}_T) \neq \mathbf{0}_D$, then the label proxy gap, $\Gamma_L(h_P, h_T) \neq \mathbf{0}_D$ since $\exists i$, such that $\Gamma_L(h_P, h_T)_i = \omega_{T,i}$.

If the feature proxy gap, $\Gamma_X(\mathscr{X}_P, \mathscr{X}_T) = \mathbf{0}_D$, then assuming the feature importance scores of the proxy model ω_P and intended model ω_T are equal, then the label proxy label $\Gamma_L(h_P, h_T) = \mathbf{0}_D$. However, if the feature importance scores are not equal, then $\forall i \in [D]$, $\Gamma_L(h_P, h_T)_i = \omega_{T,i} - \omega_{P,j}$.

Obstacle gap Access obstacles are those alleviated in accessing the proxy model, and utilization obstacles are those alleviated in utilizing the intended model. The obstacle gap looks at how different utilization obstacles are from the access obstacles. The bigger the feature proxy gap, the higher the obstacle gap. Additionally, since determining model utilization is implicit, utilization obstacles are less likely to be alleviated as the decision-maker might not be aware of the extent of the effect of their existence. Therefore an increase in the obstacle gap likely increases unequal utilization.

4. Equity Modelling

In a one-shot setting, the decision-maker makes a decision and then verifies the correctness of those decisions. In immediate evaluation models, the decision-maker immediately evaluates the correctness of their decision, for example, the prediction of a dog could be verified by crosschecking the prediction with the dog image. In delayed evaluation, the decision-maker waits until the decision takes form in the physical and social environment to verify their decision.

The Equity Framework draws attention to some of the unrealized consequences of delayed evaluation in which positively qualified individuals are evaluated by different models. It's crucial to assess model utilization which evaluates whether an individual is a true or false positive, and effects of alleviation of access and utilization obstacles, and model outcomes. To best help decision-makers plan for equitable decision-making, we provide a summary of questions in table Table 1 that decision-makers should consider when creating an equitable delayed evaluation decision-making model.

5. Conclusion

In this work, we highlighted the main components —equal model access, outcome, and utilization— of achieving equitable decision-making. We argue that since decision-makers are likely to choose models that deviate from those that depict the social and physical environment decisions are situated, relying solely on existing outcome-focused fairness metrics without alleviating access and utilization obstacles is harmful as it leads to wrong, biased and incomplete ground truth and consequently, a cycle of inequalities feeding and reinforcing historical biases.

While our formulations focus on strict equal outcomes, equal access, and equal utilization, we note that this might be hard to achieve in practice. Additionally, to mathematically define obstacles and how they hinder access and utilization, we assumed that the effects of obstacles individuals face could only be seen directly in feature values. Future works could explore varied and robust identification and alleviation of obstacles. Although formally defining equitable decision-making is complex, our work attempts to draw concepts to make the definition more concrete and lay a foundation for defining equity in ML for predictive decision-making. We hope to spark a conversation about what fairness should entail. That decision-makers try to ensure individuals have equal access and outcomes from the models and fully utilize the models. We believe the process is expensive and non-trivial, and therefore hope that the Equity Framework helps decision-makers have checkpoints for what their model does or doesn't achieve en route to equitable decision-making.

Guiding questions for an equitable decision-making model	
Task	**Questions**
Proxy model diagnosis and or selection	1) What class of model functions would be the best to use, $\mathscr{H}_P$? 2) What features will be the most predictive? 3) Given a selection of features, $\mathscr{X}_P$, what access obstacles, $\mathscr{O}_P$, will individuals face to access this model? 4) Who is most likely to face the most (fewest) access obstacles? 5) Which policy can alleviate the obstacles individuals face? Do I have the policy, Φ_P, to alleviate the obstacles? What is the model access Ψ_P? 6) How well does the model function perform on individuals in different groups, $\Omega(P)$? 7) How would a change in features affect accuracy, obstacles faced, and model access Ψ_P ? 8) How well does this model reflect the physical and social environment in which decisions take form? 9) Does the chosen model achieve equal access or optimal access threshold score?
Evaluation model diagnosis and or selection	1) What class of evaluation model functions would be the best to use, $\mathscr{H}_T$? 2) What are features I am I using for evaluation? What is the feature proxy gap, $\Gamma_X(\mathscr{X}_P, \mathscr{X}_T)$? 3) What is the label proxy gap, $\Gamma_L(h_P, h_T)$? 4) Given a selection of evaluation features, $\mathscr{X}_T$, what utilization obstacles will individuals face to utilize the model? 5) What is the obstacle gap? 6) Who is most likely to face the most (fewest) utilization obstacles? 7) If I changed the features, would that increase (decrease) the accuracy of evaluation results and increase (decrease) utilization obstacles faced? 8) Which policy can alleviate the utilization obstacles individuals face? Do I have the policy, Φ_T, to alleviate the utilization obstacles? 9) How well does the evaluation model function perform on individuals in different groups? 10) Does the chosen model achieve equal utilization or optimal utilization threshold score?
Ground truth diagnosis and or curation	1) Given the obstacle gap, label proxy gap, $\Gamma_L(h_P, h_T)$, and feature proxy gap, $\Gamma_X(\mathscr{X}_P, \mathscr{X}_T)$, should I use proxy or evaluation model features/labels or both? 2) If I choose these features/labels, given utilization, $\zeta(P)$, access, $\Psi(P)$, and outcome, $\Omega(P)$, who is most likely to be misrepresented in the new ground truth? Do I exhaustively capture obstacles individuals face?

Table 1.: Questions decision-makers should, at the minimum, ask themselves

Acknowledgements

We thank Samni Koyejo for the immensely helpful feedback on the manuscript. We also thank the reviewers for the thoughtful and insightful suggestions that helped improve this paper. Keziah Naggita was supported in part by the National Science Foundation under grant CCF-1815011 and by the Simons Foundation under the Simons Collaboration on the Theory of Algorithmic Fairness. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding sources.

References

[1] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023, 2018. URL http://arxiv.org/abs/1808.00023.

[2] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 149–159. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/binns18a.html.

[3] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018. ISSN 0027-8424. . URL https://www.pnas.org/content/115/16/E3635.

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

[5] Sudeep Bhatia. The semantic representation of prejudice and stereotypes. *Cognition*, 164:46–60, 03 2017. . URL https://doi.org/10.1016/j.cognition.2017.03.016.

[6] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Jan 2019. . URL http://dx.doi.org/10.1145/3287560.3287572.

[7] Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013, September 2004. . URL https://www.aeaweb.org/articles?id=10.1257/0002828042002561.

[8] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In Sorelle A. Friedler and Christo Wilson, editors, *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of

Proceedings of Machine Learning Research, pages 160–171. PMLR, 2018. URL http://proceedings.mlr.press/v81/ensign18a.html.

[9] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. . https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, May 2016. Accessed: 2021-03-08.

[10] David Arnold, Will Dobbie, and Crystal S. Yang. Racial bias in bail decisions. *Quarterly Journal of Economics*, 133(4):1885–1932, 2018. URL https://academic.oup.com/qje/article/133/4/1885/5025665.

[11] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807, 2016. URL http://arxiv.org/abs/1609.05807.

[12] Cynthia Dwork and Christina Ilvento. Fairness Under Composition. *arXiv e-prints*, June 2018. URL https://ui.adsabs.harvard.edu/abs/2018arXiv180606122D.

[13] Harini Suresh and John V. Guttag. A framework for understanding unintended consequences of machine learning. *CoRR*, abs/1901.10002, 2019. URL http://arxiv.org/abs/1901.10002.

[14] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019. ISSN 2624-909X. . URL https://www.frontiersin.org/article/10.3389/fdata.2019.00013.

[15] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), July 2021. ISSN 0360-0300. . URL https://doi.org/10.1145/3457607.

[16] Niki Kilbertus, Philip J. Ball, Matt J. Kusner, Adrian Weller, and Ricardo Silva. The sensitivity of counterfactual fairness to unmeasured confounding. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 616–626. PMLR, 22–25 Jul 2020. URL https://proceedings.mlr.press/v115/kilbertus20a.html.

[17] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 349–358, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. . URL https://doi.org/10.1145/3287560.3287564.

[18] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *CoRR*, abs/2010.09553, 2020. URL https://arxiv.org/abs/2010.09553.

[19] David Liu, Zohair Shafi, William Fleisher, Tina Eliassi-Rad, and Scott Alfeld. RAWLSNET: altering bayesian networks to encode rawlsian fair equality of opportunity. *CoRR*, abs/2104.03909, 2021. URL https://arxiv.org/abs/2104.03909.

[20] L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. Interventions for ranking in the presence of implicit bias. *CoRR*, abs/2001.08767, 2020. URL https://arxiv.org/abs/2001.08767.

[21] Jon Kleinberg and Manish Raghavan. Selection Problems in the Presence of Implicit Bias. In Anna R. Karlin, editor, *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 33:1–33:17, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-060-6. . URL http://drops.dagstuhl.de/opus/volltexte/2018/8323.

[22] Vitalii Emelianov, Nicolas Gast, Krishna P. Gummadi, and Patrick Loiseau. On fair selection in the presence of implicit variance. In *Proceedings of the 21st ACM Conference on Economics and Computation*, EC '20, page 649–675, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379755. . URL https://doi.org/10.1145/3391403.3399482.

[23] Harv. L. Rev. 1125 Note. Bail Reform and Risk Assessment: The Cautionary Tale of Federal Sentencing, February 2018. URL https://harvardlawreview.org/2018/02/bail-reform-and-risk-assessment-the-cautionary-tale-of-federal-sentencing/. Accessed: 2021-06-08.

[24] Noe George Gutierrez. Reducing recidivism: People on parole and probation, 07 2020. URL https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?article=2245&context=etd. Accessed: 2021-06-08.

[25] Derek Gilna. When Halfway Houses Pose Full-Time Problems , January 2015. URL https://www.prisonlegalnews.org/news/2015/jan/10/when-halfway-houses-pose-full-time-problems/. Accessed: 2021-06-08.

[26] Sam Dolnick. Pennsylvania Study Finds Halfway Houses Don't Reduce Recidivism, 03 2013. URL https://www.nytimes.com/2013/03/25/nyregion/pennsylvania-study-finds-halfway-houses-dont-reduce-recidivism.html?partner=rss&emc=rss&smid=tw-nytimes. Accessed: 2021-06-08.

[27] Lauren Sukin. When Jail Is The Better Option: The Failure of Halfway Houses, 06 2015. URL https://tcf.org/content/commentary/when-jail-is-the-better-option-the-failure-of-halfway-houses/?session=1&agreed=1. Accessed: 2021-06-08.

[28] Jason A. Okonofua, Kimia Saadatian, Joseph Ocampo, Michael Ruiz, and Perfecta Delgado Oxholm. A scalable empathic supervision intervention to mitigate recidivism from probation and parole. *Proceedings of the National Academy of Sciences*, 118(14), 2021. ISSN 0027-8424. . URL https://www.pnas.org/content/118/14/e2018036118.

[29] Thomas Joanna and Fox Aubrey. Notifications:A Summary of the Research and Best Practices for Building Effective Reminder Systems, 03 2021. URL https://www.nycja.org/assets/downloads/Court-Notification-Report-DRAFT-NN-3-8-final.pdf. Accessed: 2021-06-08.

[30] Yevgeniy P. Pislar and Rachel Puleo. Proposition 25: Replace Cash Bail with Risk Assessment Referendum Referendum, January 2020. URL https://scholarlycommons.pacific.edu/cgi/viewcontent.cgi?article=1103&context=california-initiative-review. Accessed: 2021-06-08.

[31] Stephanie Wykstra. Bail reform, which could save millions of unconvicted people from jail, explained, 10 2018. URL https://www.vox.com/future-perfect/2018/10/17/17955306/bail-reform-criminal-justice-inequality. Accessed: 2021-06-08.

[32] Palo Almond. An analysis of equity and its application to health visiting. *Journal of advanced nursing*, 37:598–606, 04 2002. . URL https://pubmed.ncbi.nlm.nih.gov/11879424/.

[33] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *CoRR*, abs/1703.00056, 2017. URL http://arxiv.org/abs/1703.00056.

[34] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311151. . URL https://doi.org/10.1145/2090236.2090255.

[35] John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6917–6926. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/miller20b.html.

[36] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, page 111–122, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340571. . URL https://doi.org/10.1145/2840728.2840730.

[37] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, EC '18, page 55–70, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450358293. . URL https://doi.org/10.1145/3219166.3219193.

[38] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 259–268, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6125-5. . URL http://doi.acm.org/10.1145/3287560.3287597.

[39] Saba Ahmadi, Hedyeh Beyhaghi, Avrim Blum, and Keziah Naggita. The strategic perceptron. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, page 6–25, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385541. URL https://doi.org/10.1145/3465456.3467629.

[40] Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 230–239, New York, NY, USA, 2019. Association for Computing Machinery. . URL https://doi.org/10.1145/3287560.3287576.

[41] Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, page 825–844, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367929. . URL https://doi.org/10.1145/3328526.3329584.

[42] Alex Frankel and Navin Kartik. Improving Information from Manipulable Data. *Journal of the European Economic Association*, 06 2021. ISSN 1542-4766. . URL https://doi.org/10.1093/jeea/jvab017.

[43] John Rawls. *A Theory of Justice*. Harvard University Press, 1971. ISBN 9780674880108. URL http://www.jstor.org/stable/j.ctvjf9z6v.

[44] Sean Nicholson-Crotty, Zachary Birchmeier, and David Valentine. Exploring the Impact of School Discipline on Racial Disproportion in the Juvenile Justice System. *Social Science Quarterly*, 90(4):1003–1018, December 2009. URL https://ideas.repec.org/a/bla/socsci/v90y2009i4p1003-1018.html.

[45] Stephen Gorard and Emma Smith. An international comparison of equity in education systems. *Comparative Education*, 40(1):15–28, 2004. . URL https://doi.org/10.1080/0305006042000184863.

[46] Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 576–586, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. . URL https://doi.org/10.1145/3442188.3445919.

[47] Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385534. . URL https://doi.org/10.1145/3465416.3483305.

[48] Anna Lauren Hoffmann. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22:900–915, 2019. URL https://doi.org/10.1080/1369118X.2019.1573912.

[49] Ben Green. Impossibility of what? formal and substantive equality in algorithmic fairness. *CoRR*, abs/2107.04642, 2021. URL https://arxiv.org/abs/2107.04642.

[50] Jenny L. Davis, Apryl Williams, and Michael W. Yang. Algorithmic reparation. *Big Data & Society*, 8(2):20539517211044808, 2021. . URL https://doi.org/10.1177/20539517211044808.

[51] Ninareh Mehrabi, Yuzhong Huang, and Fred Morstatter. Statistical equity: A fairness classification objective. *CoRR*, abs/2005.07293, 2020. URL https://arxiv.org/abs/2005.07293.

[52] Herbert A. Simon. *Bounded Rationality*, pages 15–18. Palgrave Macmillan UK, London, 1990. ISBN 978-1-349-20568-4. . URL https://doi.org/10.1007/978-1-349-20568-4_5.

[53] E. Juliusson, Niklas Karlsson, and Tommy Gärling. Weighing the past and future in decision making. *European Journal of Cognitive Psychology - EUR J COGN PSYCHOL*, 17:561–575, 07 2005. . URL https://doi.org/10.1080/09541440440000159.

[54] K. Stanovich and R. F. West. On the relative independence of thinking biases and cognitive ability. *Journal of personality and social psychology*, 94 4:672–95, 2008. . URL https://pubmed.ncbi.nlm.nih.gov/18361678/.

[55] W. Bruine de Bruin, Andrew M Parker, and B. Fischhoff. Individual differences in adult decision-making competence. *Journal of personality and social psychology*, 92 5:938–56, 2007. . URL https://pubmed.ncbi.nlm.nih.gov/17484614/.

[56] Melissa Acevedo and Joachim I. Krueger. Two egocentric sources of the decision to vote: The voter's illusion and the belief in personal relevance. *Political Psychology*, 25(1):115–134, 2004. ISSN 0162895X, 14679221. URL http://www.jstor.org/stable/3792526.

[57] Robert J. Bloomfield. What counts and what gets counted (2nd edition), January 2017. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2899141.

[58] Gregory W. Fischer, Nirmala Damodaran, Kathryn B. Laskey, and David Lincoln. Preferences for proxy attributes. *Management Science*, 33(2):198–214, 1987. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/2631637.

[59] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

[60] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, nov 2021. ISSN 0001-0782. . URL https://doi.org/10.1145/3458723.

[61] Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 375–385, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. . URL https://doi.org/10.1145/3442188.3445901.

[62] Florian Jaton. Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application. *Big Data & Society*, 8(1): 20539517211013569, 2021. . URL https://doi.org/10.1177/20539517211013569.

[63] Anders Søgaard, Barbara Plank, and Dirk Hovy. Selection bias, label bias, and bias in ground truth. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 11–13, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C14-3005.

[64] Federico Cabitza, Andrea Campagner, Domenico Albano, Alberto Aliprandi, Alberto Bruno, Vito Chianca, Angelo Corazza, Francesco Di Pietto, Angelo Gambino, Salvatore Gitto, Carmelo Messina, Davide Orlandi, Luigi Pedone, Marcello Zappia, and Luca Maria Sconfienza. The elephant in the machine: Proposing a new metric of data reliability and its application to a medical case to assess classification reliability. *Applied Sciences*, 10(11), 2020. ISSN 2076-3417. . URL https://www.mdpi.com/2076-3417/10/11/4014.

[65] Kacem Chehdi and Claude Cariou. Learning or assessment of classification algorithms relying on biased ground truth data: what interest? *Journal of Applied Remote Sensing*, 13(3):1 – 26, 2019. . URL https://doi.org/10.1117/1.JRS.13.034522.

[66] Osman Aka, Ken Burke, Alex Bauerle, Christina Greer, and Margaret Mitchell. Measuring model biases in the absence of ground truth. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, page 327–335, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. URL https://doi.org/10.1145/3461702.3462557.

[67] Jenny Mercer, James Clay, and Leanne Etheridge. Experiencing term-time employment as a non-traditional aged university student: a welsh study. *Research in Post-Compulsory Education*, 21:181–195, 07 2016. .

[68] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819. URL http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf.

[69] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642. . URL https://doi.org/10.1145/2783258.2783311.

[70] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021. . URL https://doi.org/10.1177/0049124118782533.

[71] Angwin Julia, Larson Jeff, Mattu Surya, and Kirchner Lauren. There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica.org*, 06 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[72] Jeremy D. Turiel and Tomaso Aste. P2p loan acceptance and default prediction with artificial intelligence. *Entrepreneurship & Finance eJournal*, July 2019. URL https://ssrn.com/abstract=3417122orhttp://dx.doi.org/10.2139/ssrn.3417122.