Journal of Artificial Intelligence Research 78 (2023) 747-780 Submitted 09/2022; published 11/2023

Generalizing Group Fairness in Machine Learning via Utilities

Jack Blandin BLANDIN1@UIC.EDU
Ian A. Kash IANKASH@UIC.EDU
University of Illinois at Chicago, Department of Computer Science, Chicago, IL 60607 USA

Abstract

Group fairness definitions such as Demographic Parity and Equal Opportunity make assumptions about the underlying decision-problem that restrict them to classification problems. Prior work has translated these definitions to other machine learning environments, such as unsupervised learning and reinforcement learning, by implementing their closest mathematical equivalent. As a result, there are numerous bespoke interpretations of these definitions. This work aims to unify the shared aspects of each of these bespoke definitions, and to this end we provide a group fairness framework that generalizes beyond just classification problems. We leverage two fairness principles that enable this generalization. First, our framework measures outcomes in terms of *utilities*, rather than predictions, and does so for both the decision-maker and the individual. Second, our framework can consider *counterfactual* outcomes, rather than just observed outcomes, thus preventing loopholes where fairness criteria are satisfied through self-fulfilling prophecies. We provide concrete examples of how our utility fairness framework avoids these assumptions and thus naturally integrates with classification, clustering, and reinforcement learning fairness problems. We also show that many of the bespoke interpretations of Demographic Parity and Equal Opportunity fit nicely as special cases of our framework.

1. Introduction

Machine learning (ML) is used to automate decision-making in settings such as hospital resource allocation (Obermeyer et al., 2019), job application screening (Raghavan et al., 2020), and criminal sentencing recommendations (Kleinberg et al., 2018). In this work, we focus on *group fairness* definitions, where an algorithm is considered fair if its results are independent of one or more protected attributes such as gender, ethnicity, or sexual-orientation. Many group fairness works focus only on classification settings (Berk et al., 2021; Chouldechova, 2017; Corbett-Davies et al., 2017; Dwork et al., 2012; Hardt et al., 2016; Kusner et al., 2017; Galhotra et al., 2017). This often conceals assumptions that do not always hold true in other contexts, resulting in definitions that are tightly coupled with a particular problem domain. In this paper we examine four such assumptions.

Assumption 1. Fair predictions have fair outcomes.

Many group fairness definitions require equal predictions between protected groups (Berk et al., 2021; Chouldechova, 2017; Corbett-Davies et al., 2017; Dwork et al., 2012; Hardt et al., 2016). For example, in the binary case with a *minority* group and a *majority* group, *Demographic Parity* considers a binary classifier to be fair if it predicts the positive class for individuals in the minority group and majority groups with equal probability. This

implicitly assumes that a positive prediction is always a good outcome for an individual. However, there are many problem domains where this is not true. For instance, Liu, Dean, Rolf, Simchowitz, and Hardt (2018) consider an algorithm that predicts whether or not a loan applicant will repay a loan, which then informs a loan-approval decision. In this scenario, a positive prediction results in a loan approval, which has a positive outcome for those who will pay back the loan, but has a *negative* outcome for those who will default on the loan. More generally, in situations where predictions impact individuals from the minority and majority groups differently, prediction-based fairness definitions may actually result in unfair outcomes (Liu et al., 2018; Creager, Madras, Pitassi, & Zemel, 2020). We refer to this as the *prediction-outcome disconnect* issue.

Assumption 2. Observed values of the target variable are independent of predictions.

Some fairness definitions depend on the observed value of the target variable as well as the prediction. For example, Equal Opportunity requires equal treatment of the qualified individuals in each group, where *qualified* refers to individuals who were observed to be in the positive class (Hardt et al., 2016). However, consider a classifier that predicts if an individual convicted of a crime will recidivate, where the prediction informs a judge's decision on whether to impose a prison sentence. It is possible that the decision of whether to assign prison time actually influences the individual's probability of being *qualified*, which corresponds to not recidivating. For example, suppose there is a group of *backlash* individuals that will only recidivate if they are sentenced to prison (Imai & Jiang, 2020). If the algorithm predicts that these individuals will recidivate, which causes the judge to sentence them to prison, these individuals will be considered *unqualified* because they will in fact recidivate. However, if the algorithm had instead predicted these backlash individuals to not recidivate, then they will not actually recidivate and will be considered qualified. Thus an algorithm can satisfy Equal Opportunity through a *self-fulfilling prophecy* by manipulating who is considered qualified (Ensign, Friedler, Neville, Scheidegger, & Venkatasubramanian, 2018; Imai & Jiang, 2020; Barocas, Hardt, & Narayanan, 2017; Kasy & Abebe, 2021).

Assumption 3. The objective is to predict some unobserved target variable.

In classification problems, the goal is to make a single prediction of some latent qualification attribute of the individual. However, this is not true in other ML environments where the decision is not necessarily a prediction of some ground-truth value, and where there may be more than one decision per individual. In sequential decision settings such as reinforcement learning (RL), the goal is to maximize a reward rather than predict a target. Additionally, there can be multiple sequential decisions made for each individual and we may wish to measure fairness across the entire sequence. Ranking problems and clustering also have differing objectives than traditional classification, and so require alternative fairness considerations. Several works have attempted to remedy this for particular environments, such as for sequential decision processes (Jabbari et al., 2017; Bower et al., 2017; Dwork et al., 2020; Emelianov et al., 2019), for ranking (Celis et al., 2017; Singh & Joachims, 2019; Zehlike et al., 2021), and for clustering (Chierichetti et al., 2017; Bera et al., 2019; Chen et al., 2019; Abbasi et al., 2021).

Assumption 4. Decisions for one individual do not impact other individuals.

Each classification prediction is independent of the predictions made for other individuals. However, this does not generalize to all of ML. In clustering, for instance, the impact of one individual's cluster assignment may depend on the cluster assignments of other individuals. For example, Abbasi et al. (2021) consider redistricting as a fair clustering problem, where fairness implies that constituents from each political party are equally represented by their assigned district. In order to measure how well a constituent is represented by their district, we need to know *who else* was assigned to their district. We term this *conjoined fairness* when the impact of a decision for one individual requires measuring the decisions made for other individuals as well.

1.1 Our Contributions

In this work, we introduce a utility-based group fairness framework that helps resolve the issues resulting from Assumptions 1-4. Most notably, our framework enables group fairness definitions to naturally extend to other ML environments, including reinforcement learning and clustering. We also define a novel interpretation of Equal Opportunity based on *mutual beneficence* that can be used even in situations where the notion of "qualification" may be less obvious. There are two principles that characterize our group fairness framework: *individual utility* (*benefit*) and *counterfactual outcomes*.

Individual Utility (Benefit) Borrowing terminology from Heidari, Ferrari, Gummadi, and Krause (2018), we introduce a variable to our framework called *benefit*, which represents the individual's utility resulting from a prediction. By measuring fairness directly in terms of benefit, our definitions enforce fair outcomes even in domains where the predictions impact individuals differently. Furthermore, since *utility* is a more universal concept than *prediction* or *target variable*, this approach makes sense in a broader range of domains where Assumptions 3 and 4 may not hold.

Counterfactual Outcomes We saw in our discussion of Assumption 2 that the standard definition of Equal Opportunity is vulnerable to self-fulfilling prophecies. In order to remedy this, we construct a more extensible Equal Opportunity definition by giving a more general interpretation of what it means to be *qualified*.[1] As we explain in Section 3, we interpret qualification as an individual where there exists a decision that will yield a good outcome for both the decision-algorithm *and the individual*. In other words, we measure qualification in terms of counterfactual utility outcomes for both the decision-algorithm and the individual. By considering counterfactual outcomes, our Equal Opportunity definition prevents self-fulfilling prophecies and is well-defined for a broader range of ML environments.

1.2 Additional Related Work

We are not the first to consider bringing utilities into fairness definitions. Liu et al. (2018) model a loan applicant's credit score as a utility function and show that adhering to common group fairness definitions can lower the credit scores of the disadvantaged groups. Heidari et al. (2018) demonstrate that optimizing for individual utility directly often results in group outcome equality. Hu and Chen (2020) characterize fair outcomes by translating group fairness differences into measures of *social welfare*. Wen, Bastani, and Topcu (2021) make use

1. We provide one possible generalization of Equal Opportunity, but there could certainly be others.

of an individual reward function to extend group fairness definitions to Markov decision processes (MDPs). Ben-Porat, Sandomirskiy, and Tennenholtz (2021) consider when individual utilities are modeled, but may be incorrect, and design fairness constraints that help the disadvantaged group despite the mismatch. Dwork, Reingold, and Rothblum (2023) characterize when the mapping of prediction probabilities to desirable outcomes differ between fair and unfair environments. Each of these works leverage individual utility to resolve fairness issues for a particular domain. Our contributions differ in that we provide definitions that generalize across many domains.

Our use of counterfactuals may seem reminiscent of the literature on causal fairness notions such as counterfactual fairness (Kusner et al., 2017; Kilbertus et al., 2017; Nabi & Shpitser, 2018; Loftus et al., 2018; Makhlouf et al., 2020). However, there the counterfactual is what decision the algorithm would make if the protected attribute were different, while for us the counterfactual is what a different choice of algorithm would do. Krishnaswamy, Jiang, Wang, Cheng, and Munagala (2021) consider counterfactual algorithm choices, but do so in order to have a baseline on how well the best classifier for a group can perform. Our use of counterfactuals is more similar to the way they are used in principal fairness (Imai & Jiang, 2020) and performative prediction (Perdomo et al., 2020; Miller et al., 2021). Mashiat, Gitiaux, Rangwala, Fowler, and Das (2022) also consider fairness in terms of outcomes over utilities, and consider counterfactuals in the form of regret. However, they focus on bridging the gap between group fairness in machine learning and fair division, and so they do not intend to generalize fairness to other settings such as RL and clustering.

Several works have tried to construct more general group fairness definitions. Creager et al. (2020) use directed acyclic graphs (DAGs) to help resolve the prediction-outcome disconnect issue, but do not generalize to RL or clustering. Alternatively, Williamson and Menon (2019) aim to generalize fairness definitions to non-binary sensitive groups and non-convex objectives by borrowing ideas from *risk measures*, but only consider loss disparity and thereby violate Assumption 1. Similarly, Jiang, Han, Fan, Yang, Mostafavi, and Hu (2021) strive to generalize Demographic Parity to continuous sensitive attributes while preserving tractable computation. Deldjoo, Anelli, Zamani, Bellogín, and Di Noia (2019) extend fairness to recommender systems in a way that enables domain knowledge injection. Tajbakhsh, Sadeghi, and Shams (2011) try to generalize the problem of cost and fairness trade-offs for cooperative data exchange. Our contributions differ from these works in that we provide a consistent paradigm for defining group fairness in machine learning, rather than extending group fairness for a specific use case.

2. Preliminaries

The group fairness definitions we study were originally developed in the context of classification. Following Hardt et al. (2016), we think of this task as predicting a target value Y based on features X and protected attribute Z where the population of individuals is represented by the joint distribution of (X, Z, Y) and the goal is to develop a classifier $\hat{Y}(X, Z)$. We typically omit the arguments to $\hat{Y}$ for brevity when they are clear. An individual is an element of $\mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$. Here $\mathcal{X}$ and $\mathcal{Y}$ are the sets of possible feature values and target values. We restrict the protected attribute space to be binary $\mathcal{Z} = \{0, 1\}$ purely for ease of

exposition.[2] We refer to individuals with $Z{=}0$ as the *minority* group, and those with $Z{=}1$ as the *majority* group. There is a loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and the objective is to find the classifier that minimizes expected loss $L(Y, \hat{Y}(X, Z))$. We refer to the tuple (X, Z, Y, L) as a supervised learning classification problem (SLCP).

While there are many group fairness definitions (Pessach & Shmueli, 2020; Verma & Rubin, 2018), we focus our exposition on two of the most important to illustrate our approach. [3]

Definition 2.1 (Classification Demographic Parity)**.** A classifier $\hat{Y}$ satisfies Classification Demographic Parity (`DemParClf`) for an SLCP (X, Z, Y, L) if

$$P(\hat{Y}{=}1 \mid Z{=}0) = P(\hat{Y}{=}1 \mid Z{=}1)\ . \tag{2.1}$$

Definition 2.2 (Classification Equal Opportunity)**.** A classifier $\hat{Y}$ satisfies Classification Equal Opportunity (`EqOppClf`) for SLCP (X, Z, Y, L) if

$$P(\hat{Y}{=}1 \mid Y{=}1, Z{=}0) = P(\hat{Y}{=}1 \mid Y{=}1, Z{=}1)\ . \tag{2.2}$$

3. Intuition for Utility-Based Fairness

`DemParClf` is defined exclusively in terms of SLCP variables. However, the concept behind Demographic Parity, that equal outcomes should be enforced across groups, may be relevant in any domain. Suppose that we instead define a more general version of Demographic Parity where we replace $\hat{Y}$ with a variable W that represents the *benefit* of the decision from the individual's perspective. Assuming W can take on a range of values, our general Demographic Parity becomes

$$P(W \geq \tau \mid Z{=}0) = P(W \geq \tau \mid Z{=}1), \tag{3.1}$$

where $\tau \in \mathbb{R}$ is some domain-specific threshold representing the minimum benefit to be considered a *good* outcome for the individual. [4] Rather than assuming that a prediction of 1 is a good outcome, as in `DemParClf`, we use W to explicitly capture the relationship between a decision and an outcome, which allows us to incorporate a variety of domain-specific aspects.

We could also define fairness in terms of equal expected benefits as

$$\mathbb{E}[W \mid Z{=}0] = \mathbb{E}[W \mid Z{=}1]. \tag{3.2}$$

2. In Section 5.4 we consider three protected groups to show how our approach generalizes beyond just two groups.
3. Appendix A details how other fairness metrics are implemented in our framework.
4. For simplicity, our formulation assumes individuals have cardinal utilities, so that whether an outcome is good can be determined by comparing utility to a fixed threshold. However, a similar definition could be given in terms of ordinal preferences since all we really need is to be able to calculate the probability that an individual receives an outcome which is considered good.

This eliminates the need for thresholds, but may be susceptible to outliers dominating the fairness signal. While both definitions are viable, we focus on the threshold form (Equation 3.1) for consistency and for easy comparison to the classification form (Equation 2.1) which more closely resembles the threshold form.

We can also modify `EqOppClf` to use W instead of $\hat{Y}$: $P(W \geq \tau \mid Y{=}1, Z{=}0) = P(W \geq \tau \mid Y{=}1, Z{=}1)$. However, this definition is still using the SLCP variable Y. In order to extend this definition to environments outside of classification, we need to inspect the intuition for Equal Opportunity, which is that the probability that a *qualified* individual receives a beneficial outcome is independent of the individual's protected attribute. The part of the definition referring to the *beneficial outcome* is already covered by the benefit concept, so we only need to modify the definition to allow *qualified* to also to extend to other settings. We develop intuition for what it means to be qualified by considering some examples of Equal Opportunity:

- The probability that a *skilled* job candidate is hired is independent of their protected attribute.
- The probability that a *straight-A* student is admitted to a university is independent of their protected attribute.

Thus, qualified individuals are those where there exists a decision with a mutually beneficial outcome for both the individual and the decision-maker:

- The beneficial outcome for a job applicant is to be hired. If hired, a skilled job candidate will also benefit the employer since they will be competent at their job.
- The beneficial outcome for a student is to be admitted to the university. If admitted, a straight-A student will benefit the university by enhancing the university's reputation.

Therefore, our general Equal Opportunity interpretation is: *For the subset of individuals where there exists an outcome that will benefit both the individual and the decision-maker, the probability that a beneficial individual outcome occurring is independent of the individual's protected attribute.* We can represent this notion of mutual beneficence in equation form as

$$P(W \geq \tau \mid \Gamma{=}1, Z{=}0) = P(W \geq \tau \mid \Gamma{=}1, Z{=}1) \tag{3.3}$$

where Γ is an indicator random variable with $\Gamma = 1$ when the decision-algorithm can produce an outcome that is beneficial for both the individual and the decision-maker. The benefit to the individual is captured by W. We can similarly capture the impact on the decision-maker with a *cost function* C, thus Γ becomes:[5]

$$\Gamma = \begin{cases} 1 \text{ if } \exists \hat{Y}' : W_{\hat{Y}'} \geq \tau \wedge C_{\hat{Y}'} \leq \rho \\ 0 \text{ otherwise .} \end{cases} \tag{3.4}$$

Here $W_{\hat{Y}'}$ and $C_{\hat{Y}'}$ are the benefit and cost, respectively, produced by predictor $\hat{Y}'$; and ρ is similar to τ but for the cost instead of benefit.We can check that Equation 3.3 generalizes

5. We use *cost* over *utility* due to the convention of minimizing loss functions.

well by applying it to the Section 1 recidivism example where `EqOppClf` allows for self-fulfilling prophecies: *For the subset of individuals (inmates) where there exists an outcome that will benefit both the individual (no prison) and the decision-maker (no recidivism), the probability that a beneficial individual outcome (no prison) occurring is independent of the individual's protected attribute.* We see that our more general Equal Opportunity resolves the self-fulfilling prophecy issue by conditioning on individuals who *could have been qualified*. Thus, the qualified individuals are those that will not recidivate if they do not receive prison time. In other words, our more general interpretation conditions on *counterfactually* qualified individuals.

Our counterfactual utility framework embraces a notion shared by Mashiat et al. (2022) that "what is fair" may be contingent on "what is possible". For instance, in an effort to bridge the gap between fairness in machine learning and fair division, Mashiat et al. compare two different forms of outcome-based fairness that depend on counterfactuals—the first being an improvement over a counterfactual baseline (e.g. if there was no fairness intervention), and the second with respect to the best possible outcome. We posit that the definition of a "good" outcome should be context-dependent, and so we allow for numerous interpretations by abstracting W and τ. This allows for fairness definitions to be extended to new domains beyond classification, while also allowing for more nuanced definitions such as those of Mashiat et al.

4. Utility Fairness Problem Definitions

We now provide our formal model for defining benefit and counterfactual qualification. We do so in an abstraction that we term a *Utility Fairness Problem* (UFP). UFPs generalize the classification definitions from Section 2 to other ML environments such as RL and clustering.

4.1 Utility Fairness Problem

In a UFP, a decision-maker selects a decision-algorithm m which has somehow been selected from a class of such algorithms M. An individual is an outcome of random variable (I, Z). $I \in \mathcal{I}$ represents the individual's non-sensitive attributes that are relevant for determining a decision's impact on the decision-maker or on the individual themselves. E.g. in university admissions, I may include the applicant's GPA since it is may be a proxy for post-graduation success which impacts the university's reputation; I may also include the applicant's family income level since a rejection may have greater impact for applicants with less options to choose from. $Z \in \mathcal{Z} = \{0, 1\}$ captures the individual's protected attribute. The decision-maker has a *cost function* $C : (\mathcal{I} \times \{0, 1\}) \times M \to \mathbb{R}$ which maps an individual's relevant non-sensitive attributes, sensitive attribute, and a decision-algorithm to the expected cost. We capture the cost associated with a given decision-algorithm m as a random variable $C_m : \mathcal{I} \times \{0, 1\} \to \mathbb{R}$. The impact of m on an individual is captured by the *benefit function* $W : (\mathcal{I} \times \{0, 1\}) \times M \to \mathbb{R}$ which is identical to the cost function except that it maps to expected benefit instead. Similar to the cost function, W depends on the individual's attributes and the decision-algorithm, so we represent the benefit associated with a given decision-algorithm m as a random variable $W_m : \mathcal{I} \times \{0, 1\} \to \mathbb{R}$. Two threshold constants τ and ρ are required where τ represents the minimum benefit needed for the outcome to

be considered *good* from the individual's perspective, and ρ represents the maximum cost needed for the outcome to be considered *good* from the decision-algorithm's perspective.[6] In summary, an UFP is compactly represented by a 7-tuple $(I, Z, M, W, C, \tau, \rho)$. Fairness definitions are then characterized by comparisons of benefit W, cost C, thresholds τ and ρ, and protected attribute Z.

4.2 Utility Fairness Definitions

We can now formally define our Utility Demographic Parity and Utility Equal Opportunity. As discussed in Section 3, multiple variants of the definition are possible based on different ways of summarizing the distribution of benefits of each group.

Definition 4.1 (Utility Demographic Parity)**.** Given UFP $(I, Z, M, W, C, \tau, \rho)$, a decision-algorithm $m \in M$ satisfies Utility Demographic Parity (`DemParUtil`) for threshold τ if

$$P(W_m \geq \tau \mid Z{=}0) = P(W_m \geq \tau \mid Z{=}1) \; . \tag{4.1}$$

Alternatively, m satisfies Expectation Utility Demographic Parity `DemParExpUtil` if

$$\mathbb{E}[W_m \mid Z{=}0] = \mathbb{E}[W_m \mid Z{=}1] \; . \tag{4.2}$$

Definition 4.2 (Utility Equal Opportunity)**.** Given UFP $(I, Z, M, W, C, \tau, \rho)$, a decision-algorithm $m \in M$ satisfies Utility Equal Opportunity (`EqOppUtil`) if

$$P(W_m \geq \tau \mid \Gamma{=}1, Z{=}0) = P(W_m \geq \tau \mid \Gamma{=}1, Z{=}1) \tag{4.3}$$

where Γ is an indicator variable with

$$\Gamma = \begin{cases} 1 \;\; \text{if} \;\; \exists m' \in M : W_{m'} \geq \tau \wedge C_{m'} \leq \rho \\ 0 \;\; \text{otherwise} \end{cases} \tag{4.4}$$

Alternatively, m satisfies Expectation Utility Equal Opportunity `EqOppExpUtil` if

$$\mathbb{E}[W_m \mid \Gamma{=}1, Z{=}0] = \mathbb{E}[W_m \mid \Gamma{=}1, Z{=}1] \tag{4.5}$$

where Γ is defined as in Equation 4.4.

For brevity, we only provide explicit UFP definitions for Demographic Parity and Equal Opportunity. We defer to Appendix A the UFP definitions for other common group fairness definitions, including *Predictive Parity* (Chouldechova, 2017), *Equalized Odds* (Hardt et al., 2016), *Conditional Demographic Parity* (Corbett-Davies et al., 2017), *Predictive Equality* (Chouldechova, 2017), *Conditional Use Accuracy Equality* (Berk et al., 2021), *Overall Accuracy Equality* (Berk et al., 2021), *Treatment Equality* (Berk et al., 2021), and *Test Fairness* (Chouldechova, 2017).

6. In Section 5 we consider a number of examples which show how these thresholds can be chosen based based on domain-specific considerations.

4.3 Utility Fairness Definitions Applied to Binary Classification

Here we show how to apply our generalized definitions to a specific environment: binary classification, and show that they reduce to their original classification counterparts.

A binary classification problem is an SLCP (X, Z, Y, L). For concreteness we assume L is the zero-one loss. We can construct a corresponding UFP with non-sensitive individual attributes $I = (X \times Y)$, decision algorithms space $M = \hat{\mathcal{Y}} = (X \times \{0,1\}) \times \{0,1\}$ the set of all possible classifiers, and cost function $C = L$. A positive prediction $\hat{Y} = 1$ always implies a good outcome for the individual. This corresponds to a benefit function $W = \hat{Y}$ with minimum threshold $\tau = 1$. We set the maximum cost threshold to $\rho = 0$ so that a good outcome from the decision-maker's perspective reflects a correct prediction ($L = 0$). The individual's target Y is not influenced by the prediction $\hat{Y}$, thus the parameterized benefit $W_m = W = \hat{Y}$ and $C_m = C = L$. A binary classifier $\hat{Y} : \mathcal{X} \times \{0,1\} \to \{0,1\}$ under this problem formulation satisfies `DemParUtil` if

$$P(\hat{Y} \geq 1 \mid Z{=}0) = P(\hat{Y} \geq 1 \mid Z{=}1)\ . \tag{4.6}$$

Because this is a binary classification problem, $\hat{Y} \geq 1$ is equivalent to $\hat{Y} = 1$, which makes Equation 4.6 equivalent to `DemParClf` (Equation 2.1). Similarly, a classifier $\hat{Y}$ satisfies `EqOppUtil` if

$$P(\hat{Y} \geq 1 \mid \Gamma{=}1, Z{=}0) = P(\hat{Y} \geq 1 \mid \Gamma{=}1, Z{=}1) \tag{4.7}$$

where Γ is an indicator variable with

$$\Gamma = \begin{cases} 1 & \text{if } \exists \hat{Y}' \in \hat{\mathcal{Y}} : \hat{Y}' \geq 1 \wedge L(Y, \hat{Y}') \leq 0 \\ 0 & \text{otherwise}\ . \end{cases} \tag{4.8}$$

If $\Gamma{=}1$ in Equation 4.8, then it must be that $Y{=}1$, which makes Equation 4.7 equivalent to traditional Equal Opportunity (Equation 2.2). Since Equation 4.6 reduces to Equation 2.1, and Equation 4.7 reduces to 2.2, `DemParClf` and `EqOppClf` are special cases of `DemParUtil` and `EqOppUtil`, respectively.

5. Applications

In Section 1 and 3 we discussed the motivation and intuition for our utility fairness framework. In Section 4 we formally defined the UFP framework and showed that the original classification forms of Demographic Parity and Equal Opportunity are special cases of the utility definitions. In this section we provide several examples demonstrating how UFPs are useful in practice. In each example, one or more of the four assumptions detailed in Section 1 are violated, making classification fairness definitions poor characterizations of their intended fairness measures. We demonstrate how UFPs help facilitate a more appropriate fairness definition in each case.

5.1 Prediction-Outcome Disconnect with German Credit Dataset

Here we provide an experimental analysis on an environment where classification fairness metrics fail to appropriately measure fairness due to Assumption 1. In order to ensure

that our analysis is consistent with other group fairness works, we leverage the *fairness-comparison* benchmark of Friedler et al. for data preprocessing, algorithm implementation, and fairness measurement calculations (Friedler et al., 2019).[7]

Dataset We consider the loan application scenario described by the German Credit Dataset (Dua & Graff, 2017), which consists of 1,000 loan application records. Each record in the dataset consists of 20 attributes about a loan applicant, including a binary label indicating whether the applicant is a good or bad credit risk. Following convention (Friedler et al., 2019), we consider the credit risk label as our prediction target Y, with $Y = 1$ corresponding to good-risk applicants and $Y = 0$ corresponding to bad-risk applicants. The classification objective is to correctly predict which applicants are good-risk, and which are bad-risk. In addition to the credit risk label, the dataset consists of other financial attributes about the applicant such as the number of open credit lines, credit history, as well as demographic information such as age and sex. For this experiment, we consider the applicant's sex the protected attribute, with $Z = 0$ corresponding to female applicants and $Z = 1$ corresponding to male applicants. The dataset also provides a payoff matrix representing the downstream "cost" of each prediction error, where we assume that loans are granted to applicants predicted to be good-risk, and loans are rejected for applicants predicted to be bad-risk. We can use this payoff matrix to define our UFP *cost function* C:

$$C(Y, Z, \hat{Y}) = \begin{cases} 0 & \text{if } \hat{Y} = Y \\ 1 & \text{if } \hat{Y} = 0 \wedge Y = 1 \\ 5 & \text{if } \hat{Y} = 1 \wedge Y = 0 \end{cases} \tag{5.1}$$

where $\hat{Y}$ is the binary prediction with $\hat{Y} = 1$ if the classifier predicts a good-risk applicant and $\hat{Y} = 0$ if it predicts bad-risk. Equation 5.1 implies zero cost for granting loans to good-risk applicants or rejecting bad risk applicants, a cost of 1 if a good-risk applicant is rejected for a loan, and a cost of 5 if a loan is granted to a bad-risk applicant. The dataset thus articulates that it is much worse for an applicant to fail to repay a loan (i.e. defaulting) than it is to reject an applicant that would have repaid a loan.

Algorithms We evaluate six of the binary classification algorithms studied in (Friedler et al., 2019), with a mix of algorithms that optimize for accuracy only and algorithms that optimize for both accuracy and fairness. The first two algorithms are standard classification techniques that only optimize for accuracy: Decision Tree (`DT`) and Support Vector Machine (`SVM`). The next three algorithms are Feldman Decision Tree (`Feld-DT`), Feldman SVM (`Feld-SVM`), and Feldman Logistic Regression (`Feld-LR`) (Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 2015), each of which optimizes for both accuracy and fairness by preprocessing techniques that modify the input attributes (X) to have equal marginal distributions based on the subsets of that attribute with a given sensitive value. The final algorithm, `ZafarFair` (Zafar, Valera, Rogriguez, & Gummadi, 2017), adds a fairness constraint to the accuracy optimization.

7. We make some modifications to the framework, but these are only for extension purposes, such as for new fairness measurements. The full code repository to reproduce the results in this paper is available at https://github.com/jackblandin/group-fairness-in-machine-learning-via-utilities.

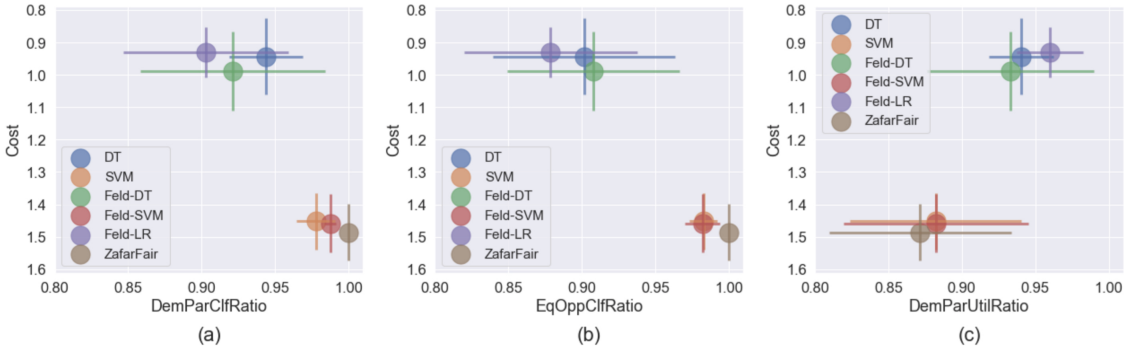


Figure 1: Using the German Credit dataset, six binary classification algorithms are evaluated for efficiency (i.e. *cost* in our UFP framework), shown on the y-axis, and fairness, shown on the x-axis. The axes are constructed so that more efficient (lower cost) algorithms are higher-up and more fair algorithms are further right.

Fairness Measures We evaluate each of the six algorithms on three different fairness measures. The first two measures quantify the extent to which `DemParClf` and `EqOppClf` are achieved:

$$\texttt{DemParClfRatio} = \min\left(\frac{P(\hat{Y}=1 \mid Z=0)}{P(\hat{Y}=1 \mid Z=1)}, \frac{P(\hat{Y}=1 \mid Z=1)}{P(\hat{Y}=1 \mid Z=0)},\right), \tag{5.2}$$

$$\texttt{EqOppClfRatio} = \min\left(\frac{P(\hat{Y}=1 \mid Y=1, Z=0)}{P(\hat{Y}=1 \mid Y=1, Z=1)}, \frac{P(\hat{Y}=1 \mid Y=1, Z=1)}{P(\hat{Y}=1 \mid Y=1, Z=0)}\right). \tag{5.3}$$

For the third measure, we introduce a utility-based measure that quantifies the extent to which `DemParUtil` is achieved:

$$\texttt{DemParUtilRatio} = \min\left(\frac{P(W \geq \tau \mid Z=0)}{P(W \geq \tau \mid Z=1)}, \frac{P(W \geq \tau \mid Z=1)}{P(W \geq \tau \mid Z=0)}\right) \tag{5.4}$$

where the benefit function W is equal to the (negative) cost function defined in Equation 5.1,[8] and $\tau = -1$ so that $W \geq \tau$ indicates the applicant did not default. We selected this threshold since it best separates the extreme benefit values that occur from applicant defaults.

Results We execute and measure each algorithm using 10-fold cross-validation. For each performance measurement, we report the average value as well as the 10th and 90th percentiles. The results are shown in Figure 1. A practitioner using classification fairness definitions, as represented in Figures 1a and 1b, may conclude that a trade-off needs to be made between cost and fairness. On the other hand, a practitioner using the utility-based fairness definitions, as in Figure 1c, would correctly conclude that fairness is positively correlated with cost, and so no trade-off is necessary. The discrepancy occurs because our

8. Since the applicant and lender (prediction algorithm) share the same incentives, we can reasonably assume that they share the same utility functions. Section 5.2 discusses an example where the benefit and cost functions diverge.

utility fairness measures are able to properly weight the magnitude of impact on the applicant for a false positive (default) versus a false negative (rejecting a qualified applicant) through appropriate instantiation of the benefit function. I.e. an applicant defaulting on a loan is considered more severe than a qualified applicant not receiving a loan, so UFP fairness measures weight applicant defaults higher when measuring fairness. More generally, fairness definitions composed using our UFP framework can encode the fairness impact for different outcomes through the benefit function, thereby mitigating the prediction-outcome disconnect issue.

5.2 Self-Fulfilling Prophecies

Here we show how to apply utility-based fairness to a setting with feedback loops, and show how `EqOppUtil`'s counterfactual interpretation of qualification prevents self-fulfilling prophecies. We use the recidivism prediction example posed by Imai and Jiang (Imai & Jiang, 2020) where a binary classifier predicts whether an inmate convicted of a crime will recidivate, which informs a judge's decision $\hat{Y}$ of whether to detain ($\hat{Y} = 0$) or release ($\hat{Y} = 1$) the inmate. The target variable Y corresponds to whether or not the inmate will recidivate, with $Y = 0$ indicating recidivism. This problem differs from typical classification since Y is influenced by $\hat{Y}$. When decisions influence the observed target variable, it is helpful to visualize the dataset by *principal strata* (Frangakis & Rubin, 2002) where each principal stratum characterizes how an individual would be affected by the decision $\hat{Y}$ with respect to the variable of interest Y. Since this is a binary classification problem with binary decisions and binary targets, we have a total of four principal strata. We assign labels to each stratum according to their behavior in Table 5.1. For example, an individual in the `Backlash` stratum will recidivate if they are detained, so $P(Y{=}1 \mid \hat{Y}{=}0) = 0$, but will not recidivate if released, so $P(Y{=}1 \mid \hat{Y}{=}1) = 1$.

To model inmates who always prefer to be released, we can take the benefit function to be a binary function with $W = 1$ when the inmate is released and $W = 0$ when detained. Similarly, to model a judge (decision-maker) who always prefers outcomes where the inmate does not recidivate, we can set the cost function to $C = 1$ when the inmate recidivates and $C = 0$ when they do not. Following a similar fairness criteria of that posed by (Imai & Jiang, 2020), we want to ensure that inmates who will not recidivate if released are released with equal probability for each protected group. Therefore, we set the benefit threshold $\tau = 1$ so that a good outcome from an inmate's perspective is when they are released. Similarly, we set the cost threshold $\rho = 0$ such that a good outcome from the judge's perspective is when an inmate does not recidivate.

`EqOppClf` considers an individual as qualified if the value of Y is observed to be 1. This means that an inmate is qualified if they do not recidivate, which corresponds to inmates that (a) are in the `Safe` stratum, (b) are in the `Backlash` stratum and are released, or (c) are in the `Preventable` stratum and are detained. Therefore, even if the minority and majority inmate populations are identical in every way other than their protected attribute, a classifier could satisfy `EqOppClf` while having different release rates for inmates who would not recidivate. For example, a decision-maker could get away with detaining more safe minority inmates than the majority simply by releasing more preventable inmates. We can

		$P(Y=1\|\hat{Y}=1)=0$	$P(Y=1\|\hat{Y}=1)=1$
		Dangerous	Backlash
$P(Y=1\|\hat{Y}=0)=0$	Detained	`Unq`	`CfUtil`
	Released	`Unq`	`Clf`, `CfUtil`
		Preventable	Safe
$P(Y=1\|\hat{Y}=0)=1$	Detained	`Clf`	`Clf`, `CfUtil`
	Released	`Unq`	`Clf`, `CfUtil`

Table 5.1: The four principal strata for the recidivism prediction problem. `Clf` cells correspond to qualified inmates according to `EqOppClf`. `CfUtil` are qualified according to `EqOppUtil`. `Unq` are unqualified according to both.

see how this works by inspecting an equivalent form of `EqOppClf`:

$$\frac{\|(\hat{Y}=1 \cap Y=1 \cap Z=0)\|}{\|(Y=1 \cap Z=0)\|} = \frac{\|(\hat{Y}=1 \cap Y=1 \cap Z=1)\|}{\|(Y=1 \cap Z=1)\|} \ .$$

Detaining more `Safe` minority inmates reduces the numerator for the minority, but this can be offset by releasing more `Preventable` inmates which causes the minority denominator to also decrease. The classifier thus attains "fairness" through a self-fulfilling prophecy by manipulating who is considered "qualified".

Alternatively, `EqOppUtil` has a prediction-independent definition of "qualification", and so it cannot be satisfied through qualification manipulation. Referencing Equation 4.4, an individual is qualified under `EqOppUtil` with $\tau = 1$ and $\rho = 0$ if $\exists \hat{Y}' \in \hat{\mathcal{Y}} : W_{\hat{Y}'} \geq 1 \wedge C_{\hat{Y}'} \leq 0$. In other words, an inmate is considered qualified if there exists a classifier that will produce $W \geq 1$ and $C \leq 0$, which is only possible for individuals who will not recidivate when released. Thus, according to `EqOppUtil`, an inmate is qualified if they are in the `Safe` or `Backlash` stratum, regardless of if they are detained or released.

This example shows how self-fulfilling prophecies can be avoided by considering counterfactuals in qualification. Appendix B.1 has a fully worked example.

5.3 Fairness in Reinforcement Learning

In this section we provide an example of how our counterfactual utility definitions extend to RL, a domain that violates Assumption 3. We provide formalisms for MDPs and then discuss how to construct the corresponding UFP. We defer a fully worked RL example to Appendix B.2.

Definition 5.1. A *Markov Decision Process* (MDP) is a 6-tuple $\{S, \mathcal{A}, T, R, \gamma, \mu\}$ where S is a set of states; A is a set of actions; $T : S \times \mathcal{A} \to \Delta S$ is a mapping of state-action pairs to a distribution over new states: $T(s_t|s_{t-1}, a_{t-1})$; $R : S \times \mathcal{A} \to \mathbb{R}$ is the reward function, which maps a state-action pair to a real number; $\gamma \in [0, 1]$ is the discount factor; and μ is the initial state probability distribution. A typical goal is to find a policy $\pi^* \in \Pi$ that maximizes the expected discounted reward.

When constructing the UFP, some parameters can be inferred from the MDP directly, while others need to be defined according to the problem domain and desired fairness

criteria. UFP parameters I, Z, M, and C can be inferred from the MDP as follows. The MDP state $s \in S$ corresponds to an individual's unprotected attributes $\tilde{s} \in \tilde{S}$ (i.e. $I = \tilde{s}$) and protected attribute Z (i.e. $s = \{\tilde{s}, Z\}$ and $S = \tilde{S} \times \mathcal{Z}$). Therefore, the initial state s_0 represents an individual in the first timestep $s_0 = \{\tilde{s}_0, Z\}$. The individual's unprotected attributes $\tilde{s}$ can change over subsequent timesteps, and do so according to the transition function $s_t \sim T(s_{t-1}, \pi(s_{t-1}))$ with $s_t = \{\tilde{s}_t, Z\}$. We assume that the individual's protected attribute Z does not change throughout an episode. $M = \Pi$, the set of policies. Following RL reward convention, the cost function C_π is the negative expected cumulative sum of rewards after executing the policy $\pi \in \Pi$:

$$C_\pi = -\mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t))\big] \ . \tag{5.5}$$

We can construct the benefit function W similarly by defining it as the expected cumulative sum of a domain-defined *benefit contribution function* $w : S \times \mathcal{A} \to \mathbb{R}$:

$$W_\pi = \mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t w(s_t, \pi(s_t))\big] \tag{5.6}$$

Thus, in order to define $W(s, \pi)$, and therefore $W_\pi(s)$, we only need to define $w(s, a)$. As we can see, the benefit contribution function w shares the same signature as the reward function R, and can be thought of as the *individual's* reward function.[9] The remaining UFP parameters τ and ρ can be assigned to implement the desired fairness criteria.

5.4 Fairness in Clustering with Chicago Ward Redistricting

In this section we apply our framework to a clustering setting where we evaluate the fairness of two competing ward redistricting maps for the city of Chicago. We demonstrate how a practitioner may decide on selecting the appropriate benefit function and τ parameterization in a real-world scenario. We also show how our framework extends to situations with more than two protected groups.

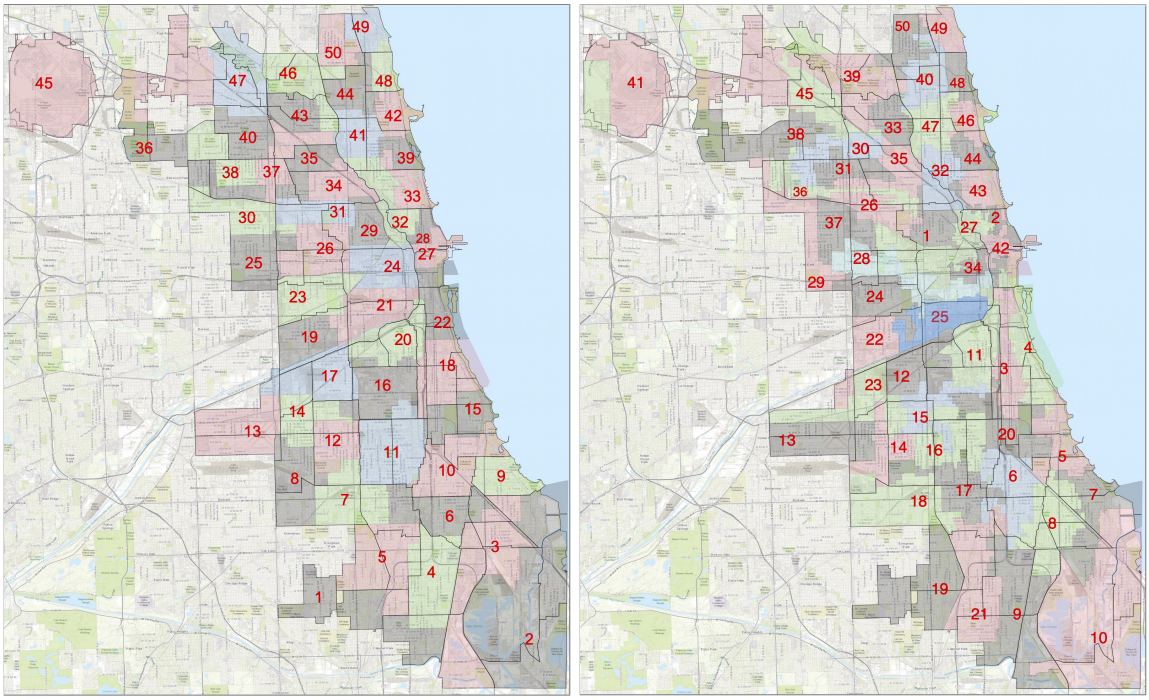
We consider the task of ward redistricting for the city of Chicago. Figure 2 shows two different ward maps proposals for 2,746,388 constituents and 50 wards. Figure 2a shows the People's Map which was drawn by the Chicago Advisory Redistricting Commission—a committee of 13 individuals tasked with producing a map that empowered historically marginalized communities (City of Chicago – Office of the City Clerk, 2023a). The People's Map was rejected by the Chicago city council in favor of the Rules Committee Map (Figure 2b), which went into effect in May of 2023 (City of Chicago – Office of the City Clerk, 2023b).

We model this redistricting setting as a clustering problem where the goal is to segment a geographic region into $\mathcal{K} = 50$ wards so that each of the $n = 2,746,388$ constituents are assigned to a single ward. Formally, we define a ward map J as a mapping of n constituents $(\bar{X}, \bar{Z})$ to an n-length vector of ward assignments, where $\bar{Z}$ and $\bar{X}$ are n-length vectors of protected and unprotected individual attributes, respectively. The unprotected attributes may correspond to any non-sensitive attributes such as geographic location, although we

9. This is similar in concept to *individual* utility function in (Wen et al., 2021) and the *benefit function* in (Heidari et al., 2018).

do not explicitly need them for the fairness computations in this example. The protected attribute Z corresponds to race. Since Chicago has significant populations of black, hispanic, and white constituents, we allow the protected attribute $Z \in \mathcal{Z}$ to take on three values: $\mathcal{Z} = \{\texttt{black}, \texttt{hispanic}, \texttt{white}\}$. We can easily modify our Section 4 definitions to account for this generalization of Z. For instance, we can define `DemParUtil` (Equation 4.1) as:

$$P(W_m \geq \tau \mid Z{=}z^i) = P(W_m \geq \tau \mid Z{=}z^j) \quad \forall\, z^i, z^j \in \mathcal{Z}\,. \tag{5.7}$$



(a) The People's Map. (b) The Rules Committee Map.

Figure 2: Two different 2023 Chicago ward map proposals. Images from Chicago Advisory Redistricting Commission (2023).

Next we evaluate the fairness of each map using our utility framework. We start by considering one common clustering fairness definition, *balanced clustering*, which we can implement as `DemParUtil`.

Balanced Clustering *Balanced clustering* strives to distribute each protected group evenly across all clusters (Chierichetti et al., 2017) . We can represent balanced clustering as `DemParUtil` by having an individual's benefit be proportional to the *balance* of their assigned cluster. That is, for an individual i sampled from $(\bar{X}, \bar{Z})$ and assigned to cluster k by clustering J, we define the individual's benefit as the ratio of the least populated protected group in cluster k relative to the rest of the population in cluster k:

$$\begin{aligned} W_J &= \texttt{balance}(k) \\ &= \min_{z \in \mathcal{Z}} \left[\frac{\sum_{j=0}^{j=n-1} \left[1 \mid J(X^j, Z^j) = k \wedge Z^i = Z^j\right]}{\sum_{j=0}^{j=n-1} \left[1 \mid J(X^j, Z^j) = k \wedge Z^i \neq Z^j\right]} \right]. \end{aligned} \tag{5.8}$$

In this case, τ represents the minimum balance to be considered a "well-balanced" cluster. To determine the appropriate value for τ, we need to consider the ward redistricting domain—a ward map satisfies `DemParUtil` if the probability that a constituent is assigned to a τ-balanced ward is independent of their protected attribute. So we need to decide what minimum balance we want to assign to τ.

We use data from the 2020 US Census (United States Census Bureau, 2021a) to obtain the Citizen Voting Age Population (CVAP) split by race.[10] Figure 3 shows the results of evaluating $P(W_J \geq \tau | Z = z) \, \forall z \in \mathcal{Z}$ for the two ward maps for different values of τ. We can think of τ on the x-axis as the minimum balance for a cluster to be considered "well-balanced", and the y-axis as the probability that a constituent is assigned to a "well-balanced" ward.

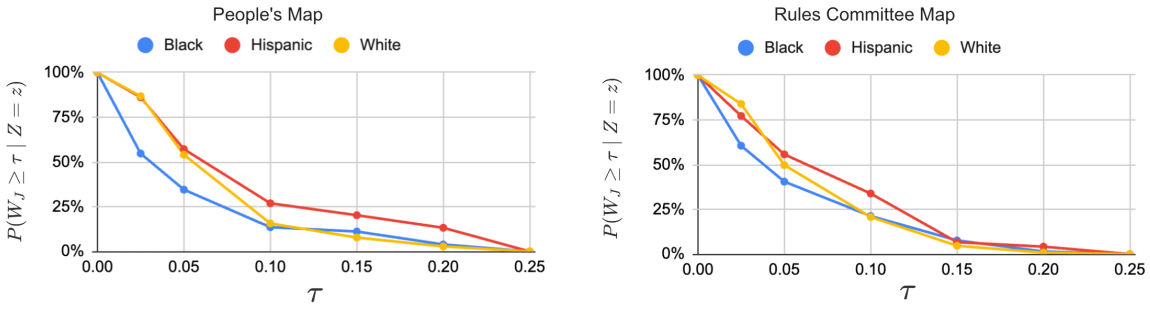


Figure 3: Group clustering results for the two Chicago ward map proposals. The y-axis shows the probability that a constituent with race z is assigned to a "well-balanced" ward, where "well-balanced" is defined by τ on the x-axis.

Figure 3 only has nonzero probabilities when $\tau \leq .20$. So if we set $\tau > .20$, then no ward would be considered well-balanced. This begs the question——is being assigned to a ward with a balance of .20 a desirable outcome for a constituent? Reflecting on this question, it is difficult to come up with a reason why a more balanced ward implies a better outcome for a constituent. In other words, *balance* does not appear to be an appropriate benefit function for constituents. If the goal is to ensure that each protected group has a proportional representation in government, as appears to be the goal of the Rule's Committee Map (Mercado, 2023),[11] then being assigned to a well-balanced ward is actually a *bad* outcome—balanced wards can result in a form of gerrymandering where a single group has a slight majority in most or all of the wards. A benefit function that encourages fair group representation in government more closely resembles *representative clustering*.

Representative Clustering Rather than trying to evenly distribute groups across clusters, *representative clustering* tries to group similar individuals together (Abbasi et al., 2021). We can implement representative clustering by defining the benefit of a constituent's

10. All data points used are provided in Tables C.1-C.2 in Appendix C.
11. The People's Map was interested in equitable representation in government, but they also wanted to preserve neighborhoods as much as possible (Chicago Advisory Redistricting Commission, 2023).

ward assignment as the proportion of constituents in their assigned ward who share the same protected attribute (race) as the constituent. Formally, the benefit for constituent i assigned to ward k is

$$W_J = \frac{\sum_{j=0}^{j=n-1} \left[1 \mid J(X^j, Z^j) = k \wedge Z^i = Z^j \right]}{\sum_{j=0}^{j=n-1} \left[1 \mid J(X^j, Z^j) = k \right]} \,. \tag{5.9}$$

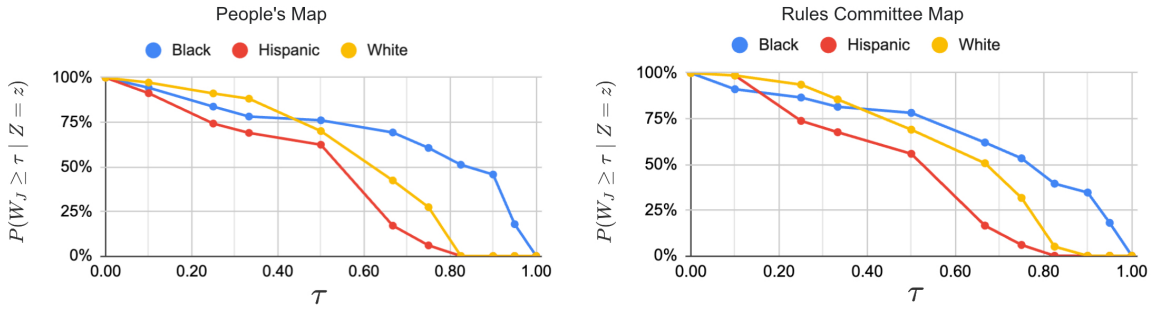


Figure 4: Representative clustering results for the two Chicago ward map proposals. The y-axis shows the probability that a constituent with race z is assigned to a ward with at least a τ proportion of z constituents. E.g. the blue circle at $\tau = .5$ shows the probability that a Black constituent will be assigned to a ward with a Black majority.

With this benefit function, a ward map satisfies `DemParUtil` if the probability that a constituent is assigned to a ward with at least τ proportion of constituents with their same race. As before, we need to select a proper value for τ. Figure 4 shows the `DemParUtil` results for various values of τ. One intuitive choice is to set $\tau = .5$ so that a constituent's "good" outcome is when they are assigned to a ward where their race represents the majority. In that case, the People's Map appears to be more fair than the Rule's Committee Map since the three data points at $\tau = .5$ are closer together.

Once again, however, we need to check that our benefit function and τ parameterization properly characterize a "good" outcome for a constituent. Is it really a good outcome if the constituent is in the majority of their ward, but the majority is insufficient to secure an *electoral* majority? For instance, if we consider voter turnout rates from the 2020 United States Presidential Election: 62.6% for blacks, 53.7% for hispanics, and 70.9% for whites (United States Census Bureau, 2021b), we immediately realize that a slight CVAP majority for blacks or hispanics may not be enough to secure an electoral majority. Therefore, we need to modify the definition for a "good" outcome to actually reflect the outcome we desire: constituents from each race should have an equal probability of being assigned to a ward where they are likely to have an electoral majority. We can also adjust the benefit function to reflect the updated definition for a "good" outcome. Setting $V(Z) \to \mathbb{R} \in [0, 1]$ as the voter turnout rate for protected group Z, we modify the benefit function to be proportional to the number of constituents in their protected group who are expected to vote:

$$W_J = \frac{\sum_{j=0}^{j=n-1} \left[V(Z^j) \mid J(X^j, Z^j) = k \wedge Z^i = Z^j \right]}{\sum_{j=0}^{j=n-1} \left[V(Z^j) \mid J(X^j, Z^j) = k \right]} \,. \tag{5.10}$$

Figure 5 shows the results of using the Equation 5.10 benefit function with the 2020 United States Presidential Election voter turnout rates. Once we factor in these voter

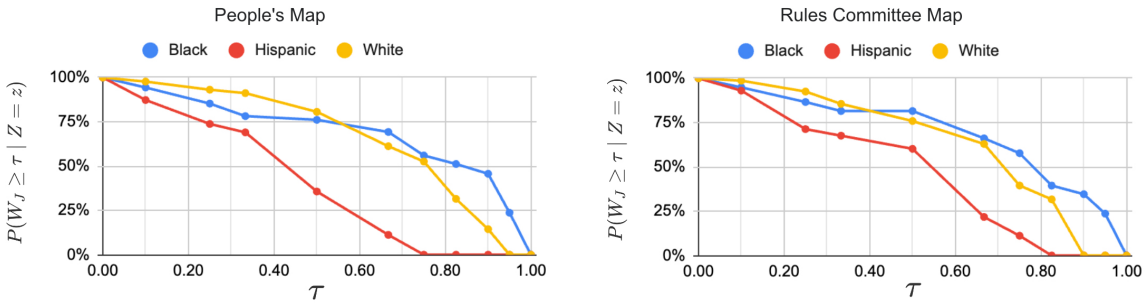


Figure 5: Representative clustering results for the two Chicago ward map proposals with voter turnout factored into the benefit function.

turnout rates, it is clear that the hispanic population has a much lower probability of achieving the desired outcome than the black and white population.

Our goal in this example is not to declare which of the two maps are more fair, nor is it to declare which benefit function is most appropriate. For instance, we did not incorporate measures of neighborhood preservation into the benefit function, which was one of the primary goals of the People's Map. Rather, we aim to illustrate how even a well-intentioned practitioner can unknowingly select policies that do not adhere to their fairness principles when they do not properly consider the *outcomes* they are striving to achieve. Using our framework, the need to choose W and τ encourages practitioners to articulate what they believe to be the ideal outcome for individuals. This leads to more productive debates on which fairness principles are relevant and how they can be achieved.

6. Discussion

We conclude by discussing three practical considerations. First, there may be situations where counterfactual outcomes need to be considered but the causal structure between decisions and observed outcomes is unknown. In this case, we can leverage techniques from the causal inference literature (Rubin, 2005) to estimate the causal structure, or use off-policy evaluation techniques (Bang & Robins, 2005; Creager et al., 2020) in order to estimate counterfactual outcomes without explicitly learning the causal structure. Second, our analysis focuses on expanding the range of settings where group fairness definitions can be applied. There are additional issues that are orthogonal to our goal, including how best to satisfy fairness during learning and how to trade off between fairness and utility. Third, practitioners may disagree on the most appropriate benefit function or appropriate threshold values to be considered "good" outcomes. We actually see benefit and threshold disagreements as a feature of our framework since, unlike prediction-based fairness metrics, our definitions naturally decouple discussions of "how individuals are impacted" from "what is fair", thereby focusing debates on the actual points of disagreement.

Acknowledgments

This material is based upon work supported by the NSF Program on Fairness in AI in Collaboration with Amazon under Award No. 1939743. Any opinion, findings, and conclu-

sions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or Amazon.

Appendix A. Extension to Other Group Fairness Definitions.

In addition to Demographic Parity and Equal Opportunity, we can extend our framework to implement various other group fairness definitions. In this section we translate common group fairness definitions from their original classification (`Clf`) form into their utility-based (`UFP`) representations.

Predictive Parity *Predictive Parity* (Chouldechova, 2017) is essentially the inverse of Equal Opportunity, which requires that the probability that an individual predicted to be positive actually belongs to the positive class is equal for both groups:

$$P(Y{=}1 \mid \hat{Y}{=}1, Z{=}0) = P(Y{=}1 \mid \hat{Y}{=}1, Z{=}1) \ ; \tag{Clf}$$

$$P(\Gamma{=}1 \mid W \geq \tau, Z{=}0) = P(\Gamma{=}1 \mid W \geq \tau, Z{=}1) \ . \tag{UFP}$$

Equalized Odds Similar to Equal Opportunity, *Equalized Odds* (Hardt et al., 2016) requires both the true positive rates ($P(\hat{Y} = 1|Y = 1)$) and false positive rates ($P(\hat{Y} = 1|Y = 0)$) to be equal:

$$P(\hat{Y}{=}1 \mid Y{=}y, Z{=}0) = P(\hat{Y}{=}1 \mid Y{=}y, Z{=}0) \ ; \tag{Clf}$$

$$\begin{array}{l} \big(P(W_m \geq \tau \mid \Gamma{=}1, Z{=}0) = P(W_m \geq \tau \mid \Gamma{=}1, Z{=}1)\big) \wedge \\ \big(P(W_m \geq \tau \mid \Gamma{=}0, Z{=}0) = P(W_m \geq \tau \mid \Gamma{=}0, Z{=}1)\big) \ . \end{array} \tag{UFP}$$

Conditional Demographic Parity *Conditional Demographic Parity* (Corbett-Davies et al., 2017) extends Demographic Parity (Definition 2.1) by allowing one or more legitimate attributes L to impact the outcome of the decision:

$$P(\hat{Y}{=}1 \mid L{=}l, Z{=}0) = P(\hat{Y}{=}1 \mid L{=}l, Z{=}1) \ ; \tag{Clf}$$

$$P(W_m \geq \tau \mid L{=}l, Z{=}0) = P(W_m \geq \tau \mid L{=}l, Z{=}1) \ . \tag{UFP}$$

for some l. Here L is playing a similar role as Γ in Equation 4.4, as it requires equal benefit for some subset of the general population.

Predictive Equality *Predictive Equality* (Chouldechova, 2017) is satisfied if individuals in the negative class have equal probabilities of receiving a positive prediction for each protected group:

$$P(\hat{Y}{=}1 \mid Y{=}0, Z{=}0) = P(\hat{Y}{=}1 \mid Y{=}0, Z{=}1) \ ; \tag{Clf}$$

$$P(W_m \geq \tau \mid \Gamma{=}0, Z{=}0) = P(W_m \geq \tau \mid \Gamma{=}0, Z{=}1) \ . \tag{UFP}$$

Conditional Use Accuracy Equality *Conditional Use Accuracy Equality* (Berk et al., 2021) requires the probability for individuals with positive predictions to belong to the positive class to be equal for both protected groups, *and* the probability for individuals with negative predictions to belong to the negative class to be equal for both protected groups:

$$\begin{aligned}&\big(P(Y{=}1 \mid \hat{Y}{=}1, Z{=}0) = P(Y{=}1 \mid \hat{Y}{=}1, Z{=}1)\big)\\ \wedge&\big(P(Y{=}0 \mid \hat{Y}{=}0, Z{=}0) = P(Y{=}0 \mid \hat{Y}{=}0, Z{=}1)\big)\ ;\end{aligned} \tag{Clf}$$

$$\begin{aligned}&\big(P(\Gamma{=}1 \mid W_m \geq \tau, Z{=}0) = P(\Gamma{=}1 \mid W_m \geq \tau, Z{=}1)\big)\\ \wedge&\big(P(\Gamma{=}0 \mid W_m < \tau, Z{=}0) = P(\Gamma{=}0 \mid W_m < \tau, Z{=}1)\big)\ .\end{aligned} \tag{UFP}$$

Overall Accuracy Equality *Overall Accuracy Equality* (Berk et al., 2021) requires the probability that an individual is assigned to their true class to be equal for both protected groups:

$$P(\hat{Y}{=}Y, Z{=}0) = P(\hat{Y}{=}Y, Z{=}1)\ ; \tag{Clf}$$

$$\begin{aligned}&\big(P(W_m \geq \tau, \Gamma{=}1, Z{=}0) = P(W_m \geq \tau, \Gamma{=}1, Z{=}1)\big)\\ \wedge&\big(P(W_m < \tau, \Gamma{=}0, Z{=}0) = P(W_m < \tau, \Gamma{=}0, Z{=}1)\big)\ .\end{aligned} \tag{UFP}$$

We interpret $Y = 1$ as $\Gamma = 1$ and $\hat{Y} = 1$ as $W_m \geq \tau$. Similarly, we interpret $Y = 0$ as $\Gamma = 0$ and $\hat{Y} = 0$ as $W_m \geq \tau$.

Treatment Equality *Treatment Equality* (Berk et al., 2021) requires an equal ratio of false negatives ($P(\hat{Y}{=}0|Y{=}1)$) and false positives ($P(\hat{Y}{=}1|Y{=}0)$) for each protected group:

$$\frac{P(\hat{Y}{=}0 \mid Y{=}1, Z{=}0)}{P(\hat{Y}{=}1 \mid Y{=}0, Z{=}0)} = \frac{P(\hat{Y}{=}0 \mid Y{=}1, Z{=}1)}{P(\hat{Y}{=}1 \mid Y{=}0, Z{=}1)}\ ; \tag{Clf}$$

$$\frac{P(W_m < \tau \mid \Gamma{=}1, Z{=}0)}{P(W_m \geq \tau \mid \Gamma{=}0, Z{=}0)} = \frac{P(W_m < \tau \mid \Gamma{=}1, Z{=}1)}{P(W_m \geq \tau \mid \Gamma{=}0, Z{=}1)}\ . \tag{UFP}$$

Test Fairness *Test Fairness* (Chouldechova, 2017) applies to classifiers that predict a probability S rather than a binary class Y. A classifier satisfies Test Fairness if, for any predicted probability S, individuals in each protected group have equal probability of being in the positive class:

$$P(Y{=}1 \mid S{=}s, Z{=}0) = P(Y{=}1 \mid S{=}s, Z{=}1)\ \forall s \in [0, 1]\ \ ; \tag{Clf}$$

$$P(\Gamma{=}1 \mid W_m{=}w, Z{=}0) = P(\Gamma{=}1 \mid W_m{=}w, Z{=}1)\ \forall w \in \mathbb{R}\ . \tag{UFP}$$

Appendix B. Fully Worked Examples

B.1 Self-Fulfilling Prophecies

Here we provide a more complete example of the recidivism prediction problem from Section 5.2, which illustrates how Assumption 2 allows for self-fulfilling prophecies with `EqOppClf`

but not with `EqOppUtil`. Additionally, we show that *principal fairness* (Imai & Jiang, 2020), which also prevents self-fulfilling prophecies, is a special case of our framework.

We use the recidivism prediction example posed by Imai and Jiang (Imai & Jiang, 2020) where a binary classifier predicts whether an inmate convicted of a crime will recidivate. The target variable Y corresponds to whether or not the inmate will recidivate, with $Y = 0$ indicating recidivism. This problem differs from typical classification since Y is influenced by $\hat{Y}$. When decisions influence the observed target variable, it is helpful to visualize the dataset by *principal strata* (Frangakis & Rubin, 2002) where each principal stratum characterizes how an individual would be affected by the decision $\hat{Y}$ with respect to the variable of interest Y. Since this is a binary classification problem with binary decisions and binary targets, we have a total of four principal strata. We assign labels to each stratum according to their behavior in Table 5.1. For example, an individual in the `Backlash` stratum will recidivate if they are detained, so $P(Y{=}1 \mid \hat{Y}{=}0) = 0$, but will not recidivate if released, so $P(Y{=}1 \mid \hat{Y}{=}1) = 1$.To model inmates who always prefer to be released, we can take the benefit function to be a binary function with $W = 1$ when the inmate is released and $W = 0$ when detained. Similarly, to model a judge (decision-maker) who always prefers outcomes where the inmate does not recidivate, we can set the cost function to $C = 1$ when the inmate recidivates and $C = 0$ when they do not.

Following a similar fairness criteria of that posed by (Imai & Jiang, 2020), we want to ensure that inmates who will not recidivate if released are released with equal probability for each protected group. Therefore, we set the benefit threshold $\tau = 1$ so that a good outcome from an inmate's perspective is when they are released. Similarly, we set the cost threshold $\rho = 0$ such that a good outcome from the judge's perspective is when an inmate does not recidivate.

We wish to evaluate the fairness of a classifier $\hat{Y}^\dagger$ that produces the results shown in Table B.1. We compare three different fairness definitions when evaluating $\hat{Y}^\dagger$: `EqOppClf`, `EqOppUtil`, and principal fairness.

An individual is qualified under `EqOppClf` if Y is observed to be 1. This means that an inmate is qualified if they do not recidivate, which corresponds to inmates that (a) are in the `Safe` stratum, (b) are in the `Backlash` stratum and are released, or (c) are in the `Preventable` stratum and are detained. Thus, `EqOppClf` becomes:

$$\begin{aligned} P\big(\hat{Y}{=}1 \;\big|\; Z{=}0, &(Y^P = \text{Backlash} \wedge \hat{Y}{=}1) \\ &\vee (Y^P = \text{Preventable} \wedge \hat{Y}{=}0) \\ &\vee (Y^P = \text{Safe})\big) \\ = P\big(\hat{Y}{=}1 \;\big|\; Z{=}1, &(Y^P = \text{Backlash} \wedge \hat{Y}{=}1) \\ &\vee (Y^P = \text{Preventable} \wedge \hat{Y}{=}0) \\ &\vee (Y^P = \text{Safe})\big)\,. \end{aligned}$$

Therefore, even if the minority and majority inmate populations are identical in every way other than their protected attribute, a classifier could satisfy `EqOppClf` while having different release rates for inmates who would not recidivate. This can be done through a self-fulfilling prophecy where the classifier manipulates who is considered "qualified". $\hat{Y}^\dagger$ accomplishes this by detaining more `Backlash` minority inmates and releasing more

`Preventable` minority inmates (Table B.1), while still satisfying `EqOppClf`:

$$\frac{20+160}{20+80+40+160} \stackrel{?}{=} \frac{20+160}{20+80+40+160}$$
$$\frac{3}{5} = \frac{3}{5} .$$

$\hat{Y}^{\dagger}$ causes two-thirds of the minority ($Z=0$) Backlash inmates to recidivate by detaining them, thus rendering them unqualified according to `EqOppClf`. [12] Since $\hat{Y}^{\dagger}$ detains only half of the majority ($Z=1$) Backlash inmates, this results in a larger proportion of minority inmates who were rendered unqualified through detainment. This results in a self-fulfilling prophecy since $\hat{Y}^{\dagger}$ satisfies `EqOppClf` by biasing the selection of qualified inmates rather than by making fair decisions.

Conversely, `EqOppUtil` does not allow for self-fulfilling prophecies since it has a prediction-independent definition of "qualification". Referencing Equation 4.4, an individual is qualified under `EqOppUtil` with $\tau = 1$ and $\rho = 0$ if $\exists \hat{Y}' \in \hat{\mathcal{Y}} : W_{\hat{Y}'} \geq 1 \wedge C_{\hat{Y}'} \leq 0$. In other words, an inmate is considered qualified if there exists a classifier that will produce $W \geq 1$ and $C \leq 0$, which is only possible for individuals who will not recidivate when released. Thus, according to `EqOppUtil`, an inmate is qualified if they are in the `Safe` or `Backlash` stratum, regardless of if they are detained or released. `EqOppUtil` is then evaluated as:

$$P\big(\hat{Y}{=}1 \;\big|\; Z{=}0, Y^P \in \{\text{Safe}, \text{Backlash}\}\big)$$
$$\stackrel{?}{=} P\big(\hat{Y}{=}1 \;\big|\; Z{=}1, Y^P \in \{\text{Safe}, \text{Backlash}\}\big)$$

$$\frac{20+160}{40+20+40+160} \stackrel{?}{=} \frac{20+160}{20+20+40+160}$$
$$.692 < .750 .$$

As expected, the proportion of qualified minority inmates who were released (.692) is less than that of majority inmates (.750), which means that $\hat{Y}^{\dagger}$ does not satisfy `EqOppUtil`. Contrasted against `EqOppClf` which requires equal release rates for those *observed* to not recidivate, `EqOppUtil` accounts for counterfactuals by requiring equal release rates for those who would not recidivate *if released*. By considering counterfactuals, `EqOppUtil` ensures fairness is not satisfied through self-fulfilling prophecies.

Although it is a stricter set of requirements than Equal Opportunity, principal fairness (Imai & Jiang, 2020) also aims to prevent self-fulfilling prophecies by requiring equal release rates for each principal stratum. To demonstrate the robustness of our UFP model, we will implement principal fairness as a UFP instantiation. If there are p principal strata, principal fairness is defined as a conjunction of p constraints:

$$P(W_C \geq \tau \mid Z{=}0, \Gamma^i{=}1) = P(W_C \geq \tau \mid Z{=}1, \Gamma^i{=}1) \quad \forall i \in \{0, ..., p-1\} \tag{B.1}$$

12. Similarly, the detained Preventable inmates were "manipulated" into not recidivating.

$Z=0$			$Z=1$		
	Dangerous	Backlash		Dangerous	Backlash
Detained	120	40	Detained	80	20
Released	30	20	Released	20	20
	Preventable	Safe		Preventable	Safe
Detained	80	40	Detained	80	40
Released	10	160	Released	80	160

Table B.1: A numerical illustration of the results of the predictions from classifier $\hat{Y}^\dagger$ on 1,000 inmates, separated by protected attribute and principal stratum. Each cell represents the number of inmates in the principal stratum and protected group who were detained ($\hat{Y}=0$) and released ($\hat{Y}=1$). The table is partially reproduced from Imai and Jiang's example (Imai & Jiang, 2020) which represents the results of a classifier that satisfies `EqOppClf`, `EqOppUtil`, and principal fairness. However, we modified the $Z=0$ numerical results to demonstrate a scenario where `EqOppClf` is satisfied, but `EqOppUtil` and principal fairness are not. See Table 5.1 for the definitions of each principal stratum.

where $\Gamma^i=1$ if the individual is in the i^{th} principal stratum. For the recidivism prediction problem, $W_C=W=\hat{Y}$, so Equation B.1 corresponds to

$$P(\hat{Y}\geq 1 \mid Z{=}0, \text{Danger}) = P(\hat{Y}\geq 1 \mid Z{=}1, \text{Danger}) \tag{B.2a}$$
$$\wedge\, P(\hat{Y}\geq 1 \mid Z{=}0, \text{Backlash}) = P(\hat{Y}\geq 1 \mid Z{=}1, \text{Backlash}) \tag{B.2b}$$
$$\wedge\, P(\hat{Y}\geq 1 \mid Z{=}0, \text{Prevent}) = P(\hat{Y}\geq 1 \mid Z{=}1, \text{Prevent}) \tag{B.2c}$$
$$\wedge\, P(\hat{Y}\geq 1 \mid Z{=}0, \text{Safe}) = P(\hat{Y}\geq 1 \mid Z{=}1, \text{Safe})\,. \tag{B.2d}$$

Principal fairness is not satisfied by $\hat{Y}^\dagger$ since the release rates of the Backlash (B.2b) and Preventable (B.2c) strata are unequal between protected groups. I.e. for Backlash:

$$\begin{aligned} &P(\hat{Y}\geq 1 \mid Z{=}0, Y^P = \text{Backlash}) \\ &\stackrel{?}{=} P(\hat{Y}\geq 1 \mid Z{=}1, Y^P = \text{Backlash}) \\ &\frac{20}{40+20} \stackrel{?}{=} \frac{20}{20+20} \\ &\frac{1}{3} \neq \frac{1}{2}\,. \end{aligned}$$

Generally, we prefer `EqOppUtil` over principal fairness since the latter requires equality in strata that may be irrelevant to fairness (e.g. requiring equal release rates in the Dangerous strata is not relevant since releasing a Dangerous inmate is an undesirable outcome). However, as demonstrated, principal fairness is well defined within our UFP model.

B.2 Fairness in Reinforcement Learning

Here we continue the discussion from Section 5.3 on applying our framework to reinforcement learning, and focus the discussion on a two-stage loan application MDP. We apply

Equations 5.5 and 5.6 to Equation 4.3 and 4.4 in order to define `EqOppUtil` in the RL setting. We then compare `EqOppUtil` to that of (Wen et al., 2021) who also provide an RL-translation of Equal Opportunity using an individual utility function, but do not fully consider counterfactual scenarios.[13]

As a motivating example, we consider a two-stage loan application decision process represented as an MDP, where a loan applicant applies for loans in two sequential timesteps, as shown in Figure 6. The decision-maker corresponds to the lender, who is represented by a policy which can either grant or reject the applicant's loan application in each timestep. There are two types of applicants. The first type, *prime*, will pay back a loan with 70% probability in the first timestep, and 80% in the second timestep. The second type, *subprime*, will pay back a loan with 60% probability in the first, and 70% in the second. Applicants in the minority group are twice as likely to be subprime as prime, whereas applicants in the majority group are twice as likely to be prime as subprime. The MDP state includes the applicant's behavior type (prime or subprime), protected attribute (minority or majority), and the timestep (0 or 1). The reward function R is defined so that the lender benefits when a loan is repaid, loses when a loan is defaulted on, and is indifferent when a loan is rejected. Table B.2 provides the full definition of R.

UFP parameters I, Z, C, and M can be inferred from the MDP as described in Section 5.3, so we only need to define W, τ, and ρ. Next, we need to define the benefit contribution function w, from which we can construct the benefit function W. Similar to the single-stage loan example in Section 5.1, we define the the benefit contribution function w so that an applicant benefits when they repay a loan, loses when they default on a loan, and is indifferent when rejected. The full definition of w is defined in Table B.2.

In order to define the remaining UFP parameters τ and ρ, we first need to establish our fairness objective. Building on the *no unnecessary harm* principle (Ustun, Liu, & Parkes, 2019; Martinez, Bertran, & Sapiro, 2020), we aim to to ensure that the lender does not cause significant harm to one protected group more than the other, unless doing so avoids severe harm to the lender. [14] We consider significant harm for the applicant to be when they default twice, when they are rejected twice, or when they are rejected once and default once. In other words, we consider a policy to be causing an applicant significant harm unless at least one loan is granted and repaid. This corresponds to a benefit threshold of $\tau = 1$. From the lender's perspective, we consider significant harm to be when the applicant defaults at least once, which corresponds to a cost threshold of $\rho = -4$. Because our fairness objective considers both the applicant's benefit and lender's cost, we will want to use Equal Opportunity as our fairness definition.

13. They also provide an RL-translation of demographic parity. As this does not involve qualification, it is equivalent to `DemParUtil` in this setting. They do not examine the ability of this approach to extend to non-RL settings.

14. Our meaning of the *no unnecessary harm* principle is slightly different from other works. For example, (Ustun et al., 2019; Martinez et al., 2020) use it to mean that one protected group's benefit should not decrease unless it increases the benefit of another protected group. Here, we use it to mean that the probability difference of causing negative benefit between the protected groups should not increase unless doing so decreases the cost for the decision-maker. Intuitively, other works consider the principle to mean Pareto optimality across all of the protected groups, whereas we consider it as Pareto optimality across all protected groups plus the decision-maker.

State s			Action a	Outcome	w	R	Probability
Applicant Type	Z	Timestep					
Prime	*	0	Grant	Repaid	+2	+3	.7
				Defaulted	-1	0	.3
Subprime	*	0	Grant	Repaid	+2	+3	.6
				Defaulted	-1	0	.4
Prime	*	1	Grant	Repaid	+2	+3	.8
				Defaulted	-1	0	.2
Subprime	*	1	Grant	Repaid	+2	+3	.7
				Defaulted	-1	0	.3
*	*	*	Reject	Rejected	0	+2	1

Table B.2: Joint distributions for the benefit contributions w and rewards R for the two-stage loan MDP example. The * character serves as a wildcard and represents "any value". As an example of interpreting this table, the first row is interpreted as follows: a Prime applicant in the initial timestep that is granted a loan will repay the loan with 70% probability, which yields $w = +2$ and $R = +3$, and will default on the loan with 30% probability, yielding $w = -1$ and $R = 0$. These values are also shown within the MDP diagram in Figure 6.

Suppose we are given the policy π^{Prime} that assigns loans to all prime applicants and rejects loans to all subprime applicants, and we wish to evaluate if π^{Prime} satisfies our fairness objective. First, we will evaluate Wen, Bastani, and Topcu's MDP translation of Equal Opportunity (Wen et al., 2021), hereafter referred to as `EqOppMDPStatic`. `EqOppMDPStatic` requires the cumulative expected individual rewards (benefit) to be equal for qualified individuals in both protected groups:

$$\mathbb{E}(W_\pi \mid p_0 \geq \alpha, Z{=}0) = \mathbb{E}(W_\pi \mid p_0 \geq \alpha, Z{=}1)$$

where p_0 is the individual's probability of repaying the loan in the first timestep, and α is some qualification threshold. We can apply `EqOppMDPStatic` to our two-stage loan MDP example by selecting the qualification threshold α, which we set to $\alpha = 2/3$ since the optimal policy grants loans to applicants with a repayment probability of at least $2/3$. This means prime applicants ($p_0 = .7$) are qualified under `EqOppMDPStatic` while subprime applicants ($p_0 = .6$) are not. Therefore, the policy π^{Prime} that grants loans to all prime applicants and rejects all loans to subprime applicants is fair according to `EqOppMDPStatic` since

$$\mathbb{E}(W_{\pi^{\text{Prime}}} \mid \text{Prime}, Z{=}0) = \mathbb{E}(W_{\pi^{\text{Prime}}} \mid \text{Prime}, Z{=}1)\ .$$

However, subprime applicants are as likely to repay a loan in the second timestep as prime applicants are in the first. Certainly the lender would prefer to grant them loans, so it seems unfair to say that subprime applicants are forever unqualified just because they are initially beneath the qualification threshold. Instead, we want our fairness definition to be able to understand that qualification may be a moving target, and that the applicants' repayment probability in later timesteps should also be considered.

`EqOppUtil`, on the other hand, considers an applicant to be qualified if there exists a policy that will result in good outcomes for both the applicant and the lender:

$$P(W_{\pi^{\text{Prime}}} \geq \tau \mid \Gamma{=}1, Z{=}0) = P(W_{\pi^{\text{Prime}}} \geq \tau \mid \Gamma{=}1, Z{=}1) \tag{B.3}$$

where Γ is an indicator variable representing qualified individuals:

$$\Gamma = \begin{cases} 1 \;\text{ if }\; \exists \pi' \in \Pi : W_{\pi'} \geq \tau \wedge C_{\pi'} \leq \rho \\ 0 \;\text{ otherwise .} \end{cases}$$

This means that qualification under `EqOppUtil` is determined by the applicant's repayment probability across both timesteps, rather than just the initial timestep as in `EqOppMDPStatic`. When we initialized our two-stage loan UFP parameters, we set the threshold parameters τ and ρ such that a good outcome, for both the applicant and lender, is when at least one loan is granted and repaid. Table B.3 shows that there exists a policy π^{Fair} that will, in expectation, yield such benefit and cost values for every applicant. Therefore, all applicants are considered qualified (i.e. $\Gamma = 1$ for all applicants). Using the values shown in Table B.4, we see that the benefit for Prime applicants is $W_{\pi^{\text{Prime}}} = 2.5$, which is greater than τ. On the other hand, the benefit for Subprime applicants is $W_{\pi^{\text{Prime}}} = 0$ which is less than τ. Therefore, `EqOppUtil` evaluates as:

$$\begin{gathered} P(W_{\pi^{\text{Prime}}} \geq 1 \mid Z{=}0) \stackrel{?}{=} P(W_{\pi^{\text{Prime}}} \geq 1 \mid Z{=}1) \\ P(\text{Prime} \mid Z = 0) \stackrel{?}{=} P(\text{Prime} \mid Z = 1) \\ .34 \neq .66 \,. \end{gathered}$$

Under π^{Prime}, the probability that a minority applicant will have benefit above τ is .34. This is because the only way to have benefit above τ is to be Prime, and the probability of a minority applicant being Prime is .34. Since a majority applicant has a higher probability of having benefit above τ (.66), π^{Prime} does not satisfy `EqOppUtil`.

Relative to `EqOppMDPStatic`, `EqOppUtil` better aligns with fairness intuition as it deems the prime-only policy π^{Prime} unfair since it results in a lower probability of at least one successful loan repayment for minority applicants than majority applicants. More generally, `EqOppUtil` is a more robust interpretation of Equal Opportunity since it naturally allows qualification to be defined across multiple timesteps.

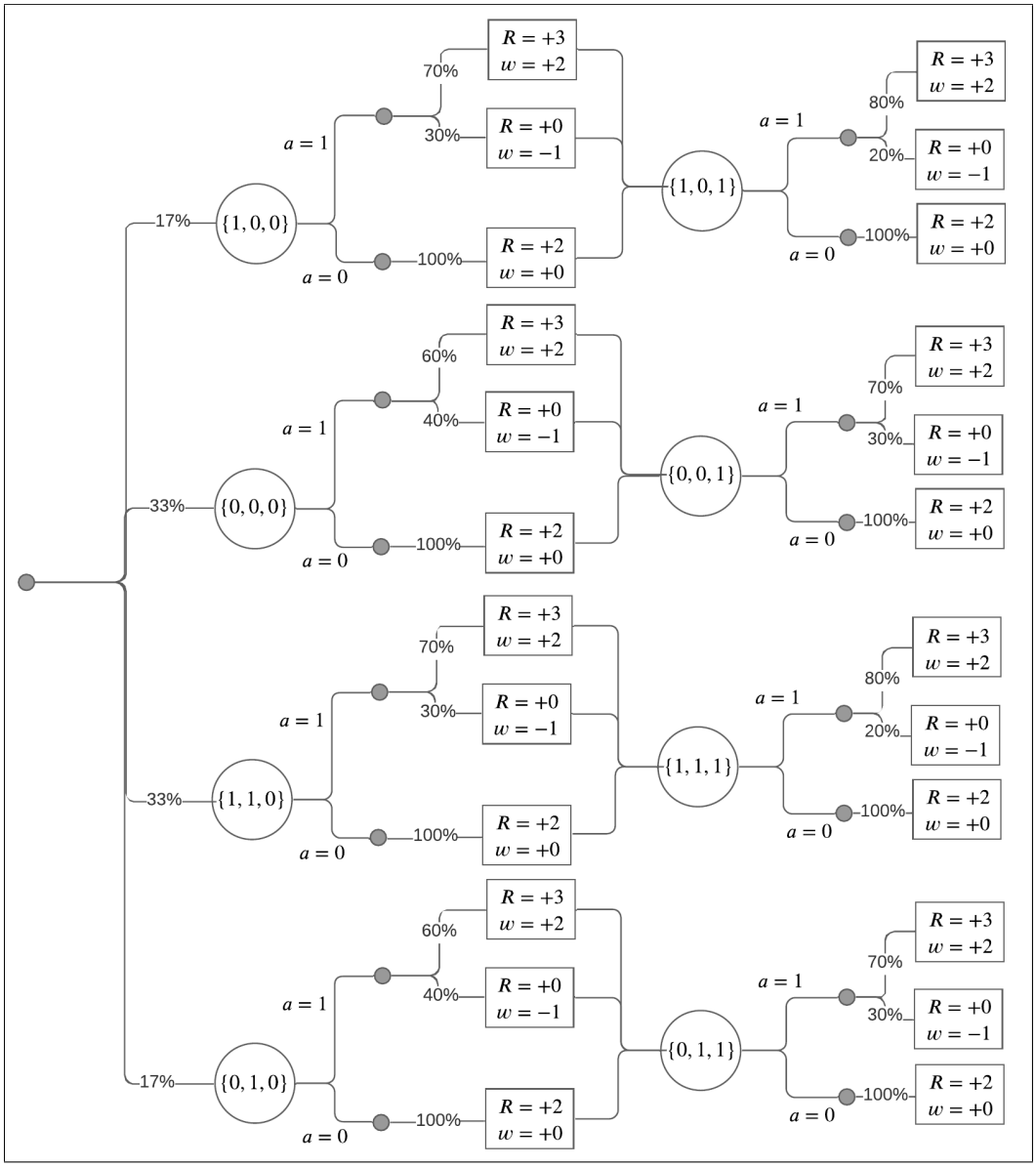


Figure 6: Two-stage loan application MDP with two actions and eight total states. Benefit contributions $w(s, a)$ are displayed alongside the rewards $R(s, a)$. States (large circles) have 3 parameters, with the first element indicating if the applicant is prime (1) or subprime (0); the second element is the binary protected attribute Z with $Z = 0$ indicating the minority and $Z = 1$ the majority; and the third element is the loan timestep with zero indicating the first timestep and 1 indicating the second timestep. The action $a = 1$ corresponds to the lender granting the applicant a loan, and $a = 0$ corresponds to the lender rejecting the applicant. The four left-most percentages represent the initial state distribution μ; e.g. an applicant sampled from μ has a 17% probability of being a minority, prime applicant. The remaining percentages represent the joint probabilities of the benefit contributions w and rewards R occurring, given the selected action.

Applicant Type	Z	1st Outcome	2nd Outcome	$\sum w$	$-\sum R$	Probability	$W_\pi = \mathbb{E}[\sum w]$	$C_\pi = -\mathbb{E}[\sum R]$
Prime	0	Repaid	Repaid	+4	-6	(.7)(.8) = .56	2.5	-4.5
		Repaid	Defaulted	+1	-3	(.7)(.2) = .14		
		Defaulted	Repaid	+1	-3	(.3)(.8) = .24		
		Defaulted	Defaulted	-2	0	(.3)(.2) = .06		
Subprime	0	Rejected	Repaid	+2	-5	(1)(.7) = .70	1.1	-4.1
		Rejected	Defaulted	-1	-2	(1)(.3) = .30		
Prime	1	Repaid	Repaid	+4	-6	(.7)(.8) = .56	2.5	-4.5
		Repaid	Defaulted	+1	-3	(.7)(.2) = .14		
		Defaulted	Repaid	+1	-3	(.3)(.8) = .24		
		Defaulted	Defaulted	-2	0	(.3)(.2) = .06		
Subprime	1	Rejected	Repaid	+2	-5	(1)(.7) = .70	1.1	-4.1
		Rejected	Defaulted	-1	-2	(1)(.3) = .30		

Table B.3: The above calculations correspond to the outcomes produced by the policy π^{Fair}, which is the policy that rejects Subprime applicants in the first timestep, and grants loans otherwise. Since all applicants have $W_\pi \geq \tau$ and $C_\pi \leq \rho$, all applicants are qualified (i.e. $\Gamma = 1$) when evaluating *any* policy for `EqOppUtil`.

App Type	Z	1st Outcome	2nd Outcome	$\sum w$	Γ	Probability	W_π	$P(\text{AppType} \mid Z)$	$P(W_\pi \geq \tau \mid Z, \Gamma{=}1)$
Prime	0	Repaid	Repaid	+4	1	(.7)(.8) = .56	2.5	.34	.34
		Repaid	Defaulted	+1	1	(.7)(.2) = .14			
		Defaulted	Repaid	+1	1	(.3)(.8) = .24			
		Defaulted	Defaulted	-2	1	(.3)(.2) = .06			
Subprime	0	Rejected	Rejected	0	1	(1)(1) = 1	0.0	.66	
Prime	1	Repaid	Repaid	+4	1	(.7)(.8) = .56	2.5	.66	.66
		Repaid	Defaulted	+1	1	(.7)(.2) = .14			
		Defaulted	Repaid	+1	1	(.3)(.8) = .24			
		Defaulted	Defaulted	-2	1	(.3)(.2) = .06			
Subprime	1	Rejected	Rejected	0	1	(1)(1) = 1	0.0	.34	

Table B.4: Calculations for the probabilities of each possible outcome under policy π^{Prime}. E.g. under π^{Prime}, the probability that a $Z = 0$ Prime applicant will repay loans in both timesteps, thus resulting in +4 benefit, is .56. These calculations are used to produce the right-most column, which is used to evaluate `EqOppUtil`. Since $.34 \neq .66$, π^{Prime} does not satisfy `EqOppUtil`.

Appendix C. Additional Tables

Ward	CVAP	Black CVAP	Hispanic CVAP	White CVAP
1	40771	52.14%	4.19%	42.82%
2	39919	28.41%	52.78%	18.34%
3	40855	93.67%	3.38%	2.69%
4	40856	95.58%	2.45%	1.30%
5	41571	75.64%	2.22%	21.58%
6	43965	96.72%	0.80%	1.97%
7	40990	90.37%	5.65%	2.39%
8	38956	18.88%	59.18%	20.47%
9	42447	94.66%	1.51%	3.18%
10	41144	93.19%	2.10%	3.47%
11	42205	95.47%	3.04%	1.34%
12	40112	35.70%	57.23%	6.10%
13	41557	2.02%	41.58%	55.09%
14	38830	2.02%	75.78%	19.76%
15	47934	53.25%	4.73%	34.30%
16	40920	19.51%	52.41%	22.37%
17	38301	7.12%	68.11%	13.41%
18	43413	86.74%	2.31%	7.75%
19	42094	24.44%	68.45%	6.93%
20	43281	3.58%	21.11%	35.04%
21	42594	14.73%	51.38%	27.14%
22	51182	26.57%	7.06%	53.42%
23	41529	79.12%	11.28%	7.91%
24	47477	21.74%	7.50%	57.48%
25	39942	91.25%	4.92%	3.51%
26	44031	68.25%	14.14%	16.15%
27	53557	4.26%	5.51%	77.98%
28	51540	6.38%	5.17%	79.39%
29	48816	5.74%	13.63%	74.28%
30	41144	70.79%	19.28%	8.96%
31	43737	10.53%	50.54%	35.63%
32	46475	11.77%	7.69%	74.12%
33	49855	5.34%	5.64%	82.03%
34	44598	4.84%	25.78%	62.98%
35	44684	4.16%	52.63%	38.09%
36	43997	2.96%	50%	42.06%
37	43924	4.29%	57.12%	32.75%
38	43230	3.28%	50.12%	41.95%
39	52945	4.35%	6.82%	82.44%
40	41776	2.44%	37.41%	53.96%
41	44359	2.67%	10.41%	80.22%
42	51865	15.72%	9%	64.69%
43	44071	6.15%	32.82%	46.21%
44	45034	6.56%	13.68%	67.30%
45	45268	1.25%	11.53%	81.59%
46	41362	3.94%	18.93%	55.99%
47	42482	1.36%	19.36%	71.16%
48	49778	14.00%	10.57%	65.84%
49	47391	27.68%	14.21%	52.71%
50	42451	11.63%	13.41%	48.21%

Table C.1: CVAP data from the 2020 US Census (United States Census Bureau, 2021a) used to evaluate the fairness of the **People's Map** for Chicago's 2023 redistricting ward map proposal (City of Chicago – Office of the City Clerk, 2023b).

Ward	CVAP	Black CVAP	Hispanic CVAP	White CVAP
9	40325	94.59%	2.52%	2.21%
10	38281	26.24%	54.22%	19.10%
19	40513	28.46%	5.66%	64.77%
21	43187	96.96%	1.13%	1.52%
8	42336	97.49%	0.61%	1.58%
7	41013	91.71%	4.90%	3.08%
18	43203	65.17%	22.16%	10.46%
6	41059	96.96%	1.32%	1.20%
17	39973	93.52%	4.40%	1.82%
14	38121	16.17%	69.85%	12.43%
5	43910	64.63%	3.69%	25.85%
16	38420	86.26%	11.52%	1.68%
13	41253	2.09%	51.28%	45.58%
20	39232	78.01%	7.66%	11.04%
15	39521	19.42%	66.10%	12.45%
23	39036	1.94%	61.31%	34.33%
12	38678	2.40%	67.41%	15.69%
3	45585	59.49%	4.23%	28.16%
22	37982	12.47%	78.06%	8.82%
4	44404	61.03%	4.47%	27.02%
11	43048	3.47%	16.78%	39.75%
25	42510	9.85%	58.40%	25.85%
24	38911	81.77%	12.23%	5.43%
28	40438	74.16%	7.00%	14.26%
34	48652	11.01%	7.34%	67.59%
42	50156	4.90%	5.84%	77.43%
27	46716	48.46%	10.35%	36.40%
37	40105	79.99%	17.48%	2.04%
29	42138	69.00%	14.59%	15.03%
1	50297	5.91%	17.58%	70.71%
2	53324	5.61%	4.76%	81.14%
26	43651	11.64%	63.18%	22.52%
43	50972	5.11%	5.41%	81.53%
36	44843	5.76%	53.88%	35.94%
32	46560	4.55%	10.63%	78.34%
35	47303	3.57%	42%	48.99%
31	44826	4.28%	61.24%	30.30%
30	44979	3.07%	48.72%	43.11%
44	52301	3.81%	5.93%	83.78%
38	46397	2.43%	24.22%	67.13%
47	47236	3.30%	9.50%	80.18%
46	51848	13.80%	10%	68.56%
33	46241	5.67%	37.10%	43.15%
45	45243	1.90%	22.46%	68.20%
39	44803	3.59%	18.28%	59.65%
41	43865	1.15%	12.24%	80.54%
48	50596	14.65%	10.52%	63.58%
40	46671	8.53%	15.46%	58.39%
50	41672	10.85%	13.74%	51.02%
49	48718	26.95%	13.05%	55.13%

Table C.2: CVAP data from the 2020 US Census (United States Census Bureau, 2021a) used to evaluate the fairness of the **Rules Committee Map** for Chicago's 2023 redistricting ward map proposal (City of Chicago – Office of the City Clerk, 2023b).

References

Abbasi, M., Bhaskara, A., & Venkatasubramanian, S. (2021). Fair clustering via equitable group representations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 504–514.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962–973.

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, *1*, 2.

Ben-Porat, O., Sandomirskiy, F., & Tennenholtz, M. (2021). Protecting the protected group: Circumventing harmful fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp. 5176–5184.

Bera, S., Chakrabarty, D., Flores, N., & Negahbani, M. (2019). Fair algorithms for clustering. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, *50*(1), 3–44.

Bower, A., Kitchen, S. N., Niss, L., Strauss, M. J., Vargas, A., & Venkatasubramanian, S. (2017). Fair pipelines..

Celis, L. E., Straszak, D., & Vishnoi, N. K. (2017). Ranking with fairness constraints..

Chen, X., Fain, B., Lyu, L., & Munagala, K. (2019). Proportionally fair clustering. In *International Conference on Machine Learning*, pp. 1032–1041. PMLR.

Chicago Advisory Redistricting Commission (2023). Chicago advisory redistricting commission.. Accessed: 2023-07-21.

Chierichetti, F., Kumar, R., Lattanzi, S., & Vassilvitskii, S. (2017). Fair clustering through fairlets. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

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

City of Chicago – Office of the City Clerk (2023a). Record number o2021-5299.. Accessed: 2023-07-21.

City of Chicago – Office of the City Clerk (2023b). Record number o2022-1318.. Accessed: 2023-07-21.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). 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.

Creager, E., Madras, D., Pitassi, T., & Zemel, R. (2020). Causal modeling for fairness in dynamical systems. In *International Conference on Machine Learning*, pp. 2185–2195. PMLR.

Deldjoo, Y., Anelli, V. W., Zamani, H., Bellogín, A., & Di Noia, T. (2019). Recommender systems fairness evaluation via generalized cross entropy..

Dua, D., & Graff, C. (2017). Uci machine learning repository..

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226.

Dwork, C., Ilvento, C., & Jagadeesan, M. (2020). Individual fairness in pipelines. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Dwork, C., Reingold, O., & Rothblum, G. N. (2023). From the real towards the ideal: Risk prediction in a better world. In *4th Symposium on Foundations of Responsible Computing (FORC 2023)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Emelianov, V., Arvanitakis, G., Gast, N., Gummadi, K., & Loiseau, P. (2019). The price of local fairness in multistage selection. In *IJCAI-2019-Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 5836–5842. International Joint Conferences on Artificial Intelligence Organization.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*, pp. 160–171. PMLR.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*(1), 21–29.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 329–338.

Galhotra, S., Brun, Y., & Meliou, A. (2017). Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pp. 498–510.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 29. Curran Associates, Inc.

Heidari, H., Ferrari, C., Gummadi, K., & Krause, A. (2018). Fairness behind a veil of ignorance: A welfare analysis for automated decision making. *Advances in Neural Information Processing Systems*, *31*.

Hu, L., & Chen, Y. (2020). Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 535–545.

Imai, K., & Jiang, Z. (2020). Principal fairness for human and algorithmic decision-making..

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., & Roth, A. (2017). Fairness in reinforcement learning. In *International Conference on Machine Learning*, pp. 1617–1626. PMLR.

Jiang, Z., Han, X., Fan, C., Yang, F., Mostafavi, A., & Hu, X. (2021). Generalized demographic parity for group fairness. In *International Conference on Learning Representations*.

Kasy, M., & Abebe, R. (2021). Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 576–586.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 656–666.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, *133*(1), 237–293.

Krishnaswamy, A., Jiang, Z., Wang, K., Cheng, Y., & Munagala, K. (2021). Fair for all: Best-effort fairness guarantees for classification. In *International Conference on Artificial Intelligence and Statistics*, pp. 3259–3267. PMLR.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR.

Loftus, J. R., Russell, C., Kusner, M. J., & Silva, R. (2018). Causal reasoning for algorithmic fairness..

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2020). Survey on causal-based machine learning fairness notions..

Martinez, N., Bertran, M., & Sapiro, G. (2020). Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pp. 6755–6764. PMLR.

Mashiat, T., Gitiaux, X., Rangwala, H., Fowler, P., & Das, S. (2022). Trade-offs between group fairness metrics in societal resource allocation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1095–1105.

Mercado, M. (2023). Chicago's controversial new ward map approved with 16 black, 14 latino wards..

Miller, J., Perdomo, J. C., & Zrnic, T. (2021). Outside the echo chamber: Optimizing the performative risk..

Nabi, R., & Shpitser, I. (2018). Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., & Hardt, M. (2020). Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR.

Pessach, D., & Shmueli, E. (2020). Algorithmic fairness..

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 469–481.

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

Singh, A., & Joachims, T. (2019). Policy learning for fairness in ranking..

Tajbakhsh, S. E., Sadeghi, P., & Shams, R. (2011). A generalized model for cost and fairness analysis in coded cooperative data exchange. In *2011 International Symposium on Networking Coding*, pp. 1–6. IEEE.

United States Census Bureau (2021a). Decennial census p.l. 94-171 redistricting data.. Accessed: 2023-07-23.

United States Census Bureau (2021b). Voting and registration in the election of november 2020.. Accessed: 2023-07-22.

Ustun, B., Liu, Y., & Parkes, D. (2019). Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pp. 6373–6382. PMLR.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pp. 1–7. IEEE.

Wen, M., Bastani, O., & Topcu, U. (2021). Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics*, pp. 1144–1152. PMLR.

Williamson, R., & Menon, A. (2019). Fairness risk measures. In *International Conference on Machine Learning*, pp. 6786–6797. PMLR.

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970. PMLR.

Zehlike, M., Yang, K., & Stoyanovich, J. (2021). Fairness in ranking: A survey..