Making Decisions that Reduce Discriminatory Impact

Matt J. Kusner [1 2] **Chris Russell** [1 3] **Joshua R. Loftus** [4] **Ricardo Silva** [1 5]

Abstract

As machine learning algorithms move into real-world settings, it is crucial to ensure they are aligned with societal values. There has been much work on one aspect of this, namely **the discriminatory prediction problem**: How can we reduce discrimination in *the predictions themselves*? While an important question, solutions to this problem only apply in a restricted setting, as we have full control over the predictions. Often we care about the non-discrimination of quantities *we do not have full control over*. Thus, we describe another key aspect of this challenge, **the discriminatory impact problem**: How can we reduce discrimination arising from *the real-world impact of decisions*? To address this, we describe causal methods that model the relevant parts of the real-world system in which the decisions are made. Unlike previous approaches, these models not only allow us to map the causal pathway of a single decision, but also to model the effect of *interference*–how the impact on an individual depends on decisions made about other people. Often, the goal of decision policies is to maximize a beneficial impact overall. To reduce the discrimination of these benefits, we devise a constraint inspired by recent work in counterfactual fairness (Kusner et al., 2017), and give an efficient procedure to solve the constrained optimization problem. We demonstrate our approach with an example: how to increase students taking college entrance exams in New York City public schools.

1. Introduction

Machine learning (ML) is used by companies, governments, and institutions to make life-changing decisions about individuals, such as how much to charge for insurance (Peters, 2017), how to target job ads (Yang et al., 2017), and who may likely commit a crime (Zeng et al., 2017).

However, the number of recent examples where ML algorithms have made discriminatory decisions against individuals because of their race, sex, or otherwise, poses a serious impediment to their use in the real world. For example, Google's advertisement system was more likely to show ads implying a person had been arrested when the search term was a name commonly associated with African Americans (Sweeney, 2013). In another case, algorithms that learn word embeddings produced embeddings with sexist associations such as "woman" being associated with "homemaker" (Bolukbasi et al., 2016).

In response to these and other examples, there has been much recent work aimed at quantifying and removing discrimination (Berk et al., 2017; Bolukbasi et al., 2016; Chouldechova, 2017; Dwork et al., 2012; 2018; Edwards & Storkey, 2015; Hardt et al., 2016; Kamiran & Calders, 2009; Kamishima et al., 2012; Kilbertus et al., 2017; Kleinberg et al., 2016; Kusner et al., 2017; Larson et al., 2016; Liu et al., 2018; Nabi & Shpitser, 2018; Pleiss et al., 2017; Zafar et al., 2017; Zemel et al., 2013; Zhang & Bareinboim, 2018). All of these works focus on what we call **the discriminatory prediction problem**: how to reduce discrimination of *the predictions themselves*. While important, the prediction problem isolates the problem of discrimination to the predictions: as long as we adjust them to agree with our definition of reduced discrimination we have solved the problem. Importantly, we have full control over what the predictions are. But frequently, we care about reducing the discrimination of quantities, which we call *impact*, that depend both on a decision we make *and* upon other real-world factors we cannot control. For instance, imagine a university with the power to make law school admission decisions. How do these decisions impact on: a person's salary 5 years later, whether the person graduates, or how able the person is to pay back any loans? Each of these has significant life-changing effects. Crucially, we define an impact as follows:

Definition 1. *An **impact** is a real-world event that is caused jointly by controllable (algorithmic) decisions, and other uncontrollable real-world factors (e.g., societal factors, human decision-makers) that may themselves be biased.*

[1]The Alan Turing Institute [2]University of Oxford [3]University of Surrey [4]New York University [5]University College London. Correspondence to: Matt J. Kusner <matthew.kusner@cs.ox.ac.uk>, Chris Russell <crussell@turing.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

This leads us to **the discriminatory impact problem**: how to reduce the discrimination of *the real-world impact of decisions*. As a large number of cases of discrimination are due to real-world mechanisms (e.g., income, voting, housing)[1], it is a crucial step to understand and correct for these mechanisms that alter the impact of decision-making.

Related Work. The importance of impact has recently been highlighted by the work of Liu et al. (2018). They showed how solutions to the discriminatory prediction problem may lead to worse impact, compared to a normal ML classifier. Green & Chen (2019) provide further evidence that when algorithmic risk assessments are shown to human decision-makers the final impact is fraught with unaddressed biases. These works suggest that a general framework for "algorithms-in-the-loop" are needed. The goal of this paper is to present such a general framework. Two recent works aim at specific settings where algorithms interact with uncontrollable real-world factors. The first, by Kannan et al. (2018), formulates a two-stage model where (a) applicants are admitted to college by an exam and (b) college students can be hired by an employer based on their exam, grades, and protected attributes (i.e., race, sex). They describe how to ensure the hiring impact satisfies a fairness criterion, while only being able to algorithmically control admission decisions. The second work is by Madras et al. (2018) who consider a model where some algorithmic predictions can be deferred to a black-box decision maker (e.g., a human, proprietary software). Both works describe how to address the discriminatory impact problem, however their models are highly tailored to the settings described above. Other recent works (Komiyama & Shimao, 2018; Elzayn et al., 2018; Dwork & Ilvento, 2018) consider related problems about social outcomes, allocating resources, and the effects of multiple discrimination-free predictors. However, they define discrimination purely as functions of decisions, and so do not address the impact problem. Here we present a general framework based on causal modeling to address the discriminatory impact problem. Our framework naturally generalizes to scenarios outside of those we consider in this work. Most similar to our work are Heidari et al. (2019), who use a social dynamics model to create an 'impacted dataset' which represents how individuals respond to an algorithm, and (Nabi et al., 2019), who design policies using Q-learning and value search to make certain causal paths give the same impact across different counterfactuals, similar to (Nabi & Shpitser, 2018; Kusner et al., 2017). Although the motivation of these works are similar, the algorithmic framework in the first work, and the fairness criteria and optimization techniques of the second are all very different from our work, and can be regarded as complimentary.

[1]https://www.theatlantic.com/magazine/archive/2014/06/the-case-for-reparations/361631/

To target the impact problem we propose to use causal methods (Pearl, 2000) to model how decisions and existing discrimination cause impact. Causal models describe how different real-world quantities are related by modeling interactions via a directed acyclic graph (DAG).[2] Given this model, we describe the impact of decisions using the framework of *interventions*. Interventions allow us to model how quantities change when another is decided (*intervened on*).

In this work, we not only want to describe the impact of decisions, we want to make decisions that reduce discriminatory impact, addressing the impact problem. Often these decisions are made to maximize beneficial impact overall (or equivalently minimize harm): increase the number of students applying to college, increase families in the middle-class, increase overall access to health care, even increase profits. Inspired by work on *counterfactual fairness* (Kusner et al., 2017) we design counterfactual quantities that measure how much a decision gives beneficial impact to an individual *purely because of attributes legally protected against discrimination* (e.g., race, sex, disability status). We develop an optimization program that constrains these quantities, while maximizing the overall beneficial impact. We demonstrate our method on a real-world dataset to assess the impact of funding advanced classes on college-entrance exam-taking. Concretely our contributions are:

- We formalize the **discriminatory impact problem**, within the framework of structural causal models (SCMs) where decisions (interventions) may *interfere* with each other.
- We describe an integer program for maximizing the overall beneficial impact, such that they are not beneficial purely because of legally-protected attributes.
- We show how this IP can be encoded as a mixed-integer linear program (MILP) and demonstrate our method on allocating school funding for advanced courses in the New York City Public School District.

2. Background

Before detailing our method, we describe counterfactual fairness, causal models, causal interventions, and interference. We use upper-case letters to denote random variables, lower-case letters for scalars or functions (this will be clear from context), upper-case bold letters for matrices, and lower-case bold letters for vectors.

[2]Here we also allow interactions between individuals themselves, as decisions made about an individual may affect other related individuals. Such models are an extension of typical causal models called *interference* models (E. L. Ogburn, 2014).

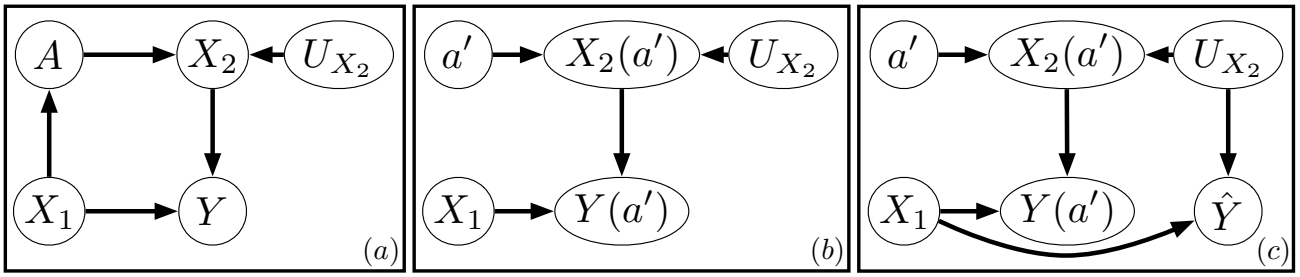

Figure 1. (a) A simple causal graph with two features X_1, X_2, protected attribute A, and outcome Y. Variables U represent hidden variables. *(b)* A counterfactual system representing the fixing A to some value a', explicitly showing new vertices where necessary: vertices "$V(a')$" are labeled "V" whenever they are not descendants of A. *(c)* The same graph, augmented by a choice of $\hat{Y}$ that does not change across counterfactual levels.

Counterfactual Fairness. Counterfactual fairness (Kusner et al., 2017) is a property of predictors based in causal models. Let A be a (set of) *protected attribute(s)* that are legally protected against discrimination (for instance in the U.S. these include: race, sex, disability status, among other things), Y a decision of interest and X a set of other features. A predictor $\hat{Y}$ of Y satisfies counterfactual fairness if it satisfies the following:

$$P(\hat{Y}(a) = y \mid A = a, X = x) \tag{1}$$
$$= P(\hat{Y}(a') = y \mid A = a, X = x),$$

for all a, a', y, x in the domains of A, Y, and X. The notation $V(a')$ refers to a *counterfactual* version of a *factual* variable V.[3] It represents the counterfactual statement "the value of V had $A = a'$ instead of the factual value". As used by (Kusner et al., 2017), counterfactuals are defined by Pearl's Structural Causal Model (SCM) framework (Pearl, 2000). This framework defines a causal model by a set of *structural equations* $V_i = g_i(pa_i, U_i)$. These equations describe how variables affect one another within a causal directed acyclic graph (DAG) $\mathcal{G}$ (pa_i are the observable parents of V_i in $\mathcal{G}$, and U_i is a (set of) parent-less unobserved latent causes of V_i). The counterfactual "world" is generated by fixing A to a', removing any edges into vertex A, and propagating the change to all descendants of A in the DAG, as shown in Figure 1 (a), (b). Any variables in the model that are not in $A \cup X$, and are not descendants of A, can be inferred given the event $\{A = a, X = x\}$, as the remaining set of equations defines a joint distribution.

The motivation behind (1) is that the protected attribute A should not be a cause of the prediction $\hat{Y}$ for anyone, other things (the non-descendants of A in the DAG) being equal. Informally, this translates to "we would not make a different prediction for this person had this person's protected attribute been different, given what we know about them". This is in contrast to non-causal definitions which enforce observational criteria such as $Y \perp\!\!\!\perp A \mid \hat{Y}$ (calibration (Flores et al., 2016)), or $\hat{Y} \perp\!\!\!\perp A \mid Y$ (equalized odds (Hardt et al., 2016)). As discussed by Chouldechova (2017); Kleinberg et al. (2016), in general it is not possible to enforce both conditions, particularly if $A \not\perp\!\!\!\perp Y$ (which happens if A is a cause of Y). To ensure $\hat{Y}$ is not a cause of A (neither direct or indirect), counterfactual fairness adds $\hat{Y}$ to the graph independently of A, as in Figure 1 (c), while maximizing the predictive accuracy of $\hat{Y}$. For more information about causality and fairness see the survey (Loftus et al., 2018).

In this formulation, the original decision Y might be unfair as Y is caused by protected attribute A, but we have the freedom to set our new decision $\hat{Y}$ so it is not causally affected by A. However, we often do not have the freedom to directly decide a quantity Y (i.e., an impact). Instead, we may only be able to control a decision Z that partially decides this impact. This idea of partial control of real-world quantities is formalized by *causal interventions*.

Interventions. Causal modeling defines an operation for a decision that influences an impact in the real world, called an *intervention*. Interventions are a causal primitive in the SCM framework: they describe how deciding a quantity Z affects other quantities in the causal graph.[4]

Perfect interventions are often impossible in real problems. For example, a school cannot decide an individual's post-graduation salary Y. If they could, decreasing discrimination would reduce to the discriminatory prediction problem. Instead, our goal is to consider imperfect interventions that diminish the relationship between protected attribute A and impact Y. As commonly done in the literature (Spirtes et al., 1993; Pearl, 2000; Dawid, 2002), we can represent interventions as special types of vertices in a causal graph. These vertices index particular counterfactuals. For instance, if each individual i is given a particular intervention $Z^{(i)} = z^{(i)}$, we can represent their counterfactual impacts as $Y^{(i)}(z^{(i)})$, and the corresponding causal graph will include a vertex

[3] Our notation is equivalent to that used in (Kusner et al., 2017).

[4] From now on we will use the term 'intervention' in place of the less formal term 'decision' (interventions being more general).

Z pointing to Y. This vertex represents the index of the intervention. For simplicity, we will assume that each $Z^{(i)}$ are binary, where $Z^{(i)}=0$ means no intervention is given to i (non-binary interventions are possible in our framework, the optimization is just trickier). In contrast to the original definition of counterfactual fairness which has a single counterfactual, we will also write $Y^{(i)}(a^{(i)}, z^{(i)})$ to denote the doubly-counterfactual impact for individual i with a fixed $A^{(i)} = a^{(i)}$ and intervention $Z^{(i)}=z^{(i)}$.

Interference. Because interventions applied to one individual i often affect other individual j, we consider a generalization of SCMs called *interference* models (Sobel, 2006; E. L. Ogburn, 2014). As in Aronow & Samii (2017), we are not concerned about direct causal connections between different impacts $\{Y^{(i)}, Y^{(j)}\}$. We focus exclusively on the intention-to-treat effects of interventions $\{Z^{(1)}, Z^{(2)}, \ldots, Z^{(n)}\}$ on impacts $\{Y^{(1)}, Y^{(2)}, \ldots, Y^{(n)}\}$, where n is the number of individuals. In these models, each impact $Y^{(i)}$ is now a function of the full intervention set $\mathbf{z} \equiv [z^{(1)}, z^{(2)}, \ldots, z^{(n)}]^\top$, i.e., $Y^{(i)}(a^{(i)}, \mathbf{z})$, because of possible interference. The form of interference we consider in this work is neighbor-based: a pre-defined set of "neighbors" of i, defined as $N(i) \subset \{1, 2, \ldots, n\}$, influence i. Specifically, their interventions will influence the impact of i: $Y^{(i)}$. This is represented as causal edges $\{Z^{(j)}\}_{j \in N(i)} \to Y^{(i)}$ (such edges can also be indirect).

Beneficial Impacts. Finally, alongside reducing the discrimination in impacts many decision-makers are interested in maximizing the beneficial impact across individuals: maximizing graduation rate, maximizing loan repayment, maximizing voter registration. Thus, in this work we will consider impacts Y that are beneficial: i.e., higher values are better. How can we make interventions so that not only is this overall beneficial impact maximized, but that an individual does not recieve significant benefit because of their protected attributes A? We formalize and answer this question in the next two sections.

3. The Discriminatory Impact Problem

Imagine we have a dataset of n individuals, and we have the following information about each of them: A a protected attribute (or a set of them), X real-world features that influence an impact of interest Y, and a causal graph that describes how these quantities and an intervention Z are causally related (there are many ways to discover this causal graph and we direct readers to the excellent survey Peters et al. (2017) for more details about this). We show an example graph in Figure 2. This figure describes the causal graph of two interfering individuals: red vertices correspond to individual 1, blue to individual 2. A few clarifications: i) we have described features $X^{(i)}$ as quantities that happen before the intervention and are thus not directly impacted by it. However, our framework does allow for quantities to be impacted by the intervention along the path to Y. Our experiment will describe such an example; ii) note that there are no edges from $A^{(i)}, X^{(i)}$ to Z because intervening on Z removes them (by definition); iii) for simplicity we omit edges between A and X (our framework allows for this as long as structural equations are defined, see appendix for more details), and describe direct interference between $Z^{(1)}$ and $Y^{(2)}$, and vice versa (our framework also allows for indirect interference).

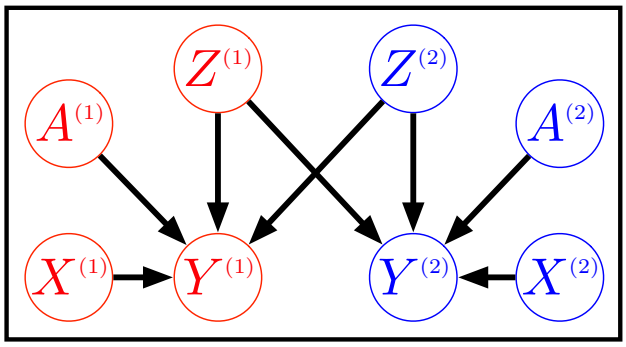

Figure 2. An example causal diagram with interventions Z, impacts Y, real-world features X, and protected attribute A. Here, individuals $1, 2$ interfere with each other: their interventions may alter the impact of each other.

3.1. An Example

For more intuition about the discriminatory impact problem we describe a real-world example of housing relocation subsidies in the green box on the following page.

3.2. Learning Impacts

Before addressing discrimination we will start by learning how to maximize beneficial impact Y. As Y is a random variable we propose to maximize a summary of Y: its expected value $\mathbb{E}[Y]$. Then our goal is to assign interventions $\mathbf{z}$ to maximize the sum of expected benefits. As in the example, it is often unreasonable to assume we can assign everyone an intervention. Thus, the maximization is subject to a maximum budget b which we formalize as follows:

$$\max_{\mathbf{z} \in \{0,1\}^n} \sum_{i=1}^{n} \mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}], \quad (2)$$

$$s.t., \sum_{i=1}^{n} z^{(i)} \leq b$$

where $a^{(i)}$, $x^{(i)}$ are the factual realizations of $A^{(i)}$ and $X^{(i)}$. Recall that this conditional expectation is given by a causal model with interference. We note that it is always well-

Example: Housing Relocation Subsidies

Consider two individuals that live in the same neighborhood, such that $\{A^{(1)}, A^{(2)}\}$ are their races, $\{X^{(1)}, X^{(2)}\}$ are their professional qualifications, $\{Y^{(1)}, Y^{(2)}\}$ are their annual incomes in 5 years, and $\{Z^{(1)}, Z^{(2)}\}$, are interventions: if $Z^{(i)}=1$ person i gets a subsidy to move to a neighborhood with better transport links. Figure 2 shows a causal graph for this scenario: $A^{(i)}$ and $X^{(i)}$ have effects on $Y^{(i)}$, as does the intervention $Z^{(i)}$. For a moment, imagine there is no interference between the intervention of one individual $Z^{(i)}$ and the impact of the other individual $Y^{(j)}$ (i.e., the crossing arrows are removed in Figure 2).

Imagine that US Department of Housing and Urban Development only has the budget to grant an intervention to one individual. Imagine that both individuals are nearly identical: they have the same professional qualifications $\{X^{(1)}, X^{(2)}\}$ but different races $\{A^{(1)}, A^{(2)}\}$. Individual 1 is a member of a majority race and is privileged because of it. Specifically, given the intervention $Z^{(1)}=1$ their impact $Y^{(1)}([z^{(1)}=1, z^{(2)}=0])=\$100,000$ is larger than that of individual 2 if they had received the intervention $Y^{(2)}([z^{(1)}=0, z^{(2)}=1])=\$50,000$.

Now consider that there is also interference: if one individual i receives a subsidy, their moving out causes others to move out and property prices to fall in their old neighborhood. This negatively affects the impact $Y^{(j)}$ of individual j who did not get the intervention. With this interference we have:

$$Y^{(1)}([z^{(1)}=1, z^{(2)}=0])=\$100,000 \qquad Y^{(2)}([z^{(1)}=1, z^{(2)}=0])=\$10,000$$
$$Y^{(1)}([z^{(1)}=0, z^{(2)}=1])=\$60,000 \qquad Y^{(2)}([z^{(1)}=0, z^{(2)}=1])=\$50,000$$

Even though these individuals have identical qualifications $\{X^{(1)}, X^{(2)}\}$ the benefits they receive are different: individual 1 has larger benefit $Y^{(1)}$ than individual 2, $Y^{(2)}$, in all cases. In fact, the difference seems purely based on race: based on their similarity, it seems that if individual 1 had the race of individual 2 their benefit would go down, whereas in the reverse case, the benefit of individual 2 would go up. How can we ensure that interventions are beneficial overall while limiting the benefit that is due purely to protected attributes such as race?

defined, regardless of the neighborhood of each individual (Arbour et al., 2016; Aronow & Samii, 2017).

In the housing example, if we make interventions purely to maximize overall benefit, then it doesn't matter if we give the intervention to individual 1 or 2, the overall benefit is $\$110,000$. However, giving the intervention to individual 1 severely harms individual 2 just because of their race. In the next section we describe a method to not only maximize overall benefit, but to constrain the amount of individual benefit that is due to race (or any protected A).

4. Our Solution

To bound the impact due to discrimination, we propose constraints on *counterfactual privilege*:

$$\begin{aligned}&\mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)}=a^{(i)}, X^{(i)}=\mathbf{x}^{(i)}] \\ -&\mathbb{E}[Y^{(i)}(a', \mathbf{z}) \mid A^{(i)}=a^{(i)}, X^{(i)}=\mathbf{x}^{(i)}] < \tau,\end{aligned} \tag{3}$$

for some $\tau \geq 0$, all a' in the domain of A, and $i \in \{1, \ldots, n\}$. The first term of the constraint is the *actual* benefit received by individual i for interventions $\mathbf{z}$. The second term is the *counterfactual* benefit they would have received had they had attribute $A=a'$. The intuition here is that these constraints prevent interventions that allow an individual i to *gain more than τ units in expected benefit $Y^{(i)}$ due $a^{(i)}$*.

Consider what this means for the housing example for $\tau = \$0$. Because individuals 1 and 2 are identical except for their race A, they are reasonable approximations to counterfactual versions of each other. Thus, if the intervention is given to individual 1 the left-hand side of eq. (3) equals $Y^{(1)}([z^{(1)}=1, z^{(2)}=0]) - Y^{(2)}([z^{(1)}=1, z^{(2)}=0])=\$90,000$, which doesn't satisfy the constraint ($\tau = \$0$). If however, the intervention is given to individual 2 we have $Y^{(2)}([z^{(1)}=0, z^{(2)}=1]) - Y^{(1)}([z^{(1)}=0, z^{(2)}=1])=-\$10,000$ which does satisfy the constraint. Thus, this constraint ensures interventions create impacts that aren't due to A.

Comparing with a counterfactual is inspired by counterfactual fairness eq. (1). To be more similar to eq. (1) we could have bounded the *absolute difference* in the above equation. We intentionally did not do this for the following reason: it penalizes individuals who would have a *better impact had their race been different*. Thus, it harms already-disadvantaged individuals. In the above example, the second intervention would now not have satisfied the constraint as $\$10,000 \not< \tau$. As our goal is to improve impacts for already-disadvantaged individuals, we use the constraint in eq. (3).

A Formulation with Fewer Assumptions. One downside to the constraint in eq. (3) is that in general it requires

full-knowledge of the specific form of all structural equations.[5] This is because if some feature X_k is a descendant of A, then in general $X_k(a) \neq X_k(a')$. However, assuming we know the structural equations is usually a very strong assumption. To avoid this, we propose to consider X that are not descendants of A (as shown in Figure 2) and fit a model that will not require any structural equation except for $\mathbb{E}[Y]$. Thus we propose a variation of the above constraint:

$$\mathbb{E}_{\mathcal{M}_\prec}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)}_\prec = \mathbf{x}^{(i)}_\prec] \qquad (4)$$
$$\underbrace{-\mathbb{E}_{\mathcal{M}_\prec}[Y^{(i)}(a', \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)}_\prec = \mathbf{x}^{(i)}_\prec]}_{c_{ia'}} < \tau,$$

where $X^{(i)}_\prec$ is the subset of $X^{(i)}$ that are non-descendants of $A^{(i)}$ in the causal graph, and $\mathcal{M}_\prec$ is a causal model that omits any observed descendants of A except for Y. (note that A and $X_\prec$ can still non-linearly interact to cause Y, as in our experiments). Without eq. (4), in general, one requires assumptions that cannot be tested even with randomized controlled trials (Loftus et al., 2018; Kusner et al., 2017). In contrast, the objective function (2) and constraints (4) can in principle be estimated by experiments. Note that the objective function (2) can use all information in $X^{(i)}$, since there is no need to propagate counterfactual values of $A^{(i)}$. Hence, we use two structural equations for the impact Y: one with $X^{(i)}$, and one with $X^{(i)}_\prec$. The full constrained optimization problem is therefore:

$$\max_{\mathbf{z}\in\{0,1\}^n} \sum_{i=1}^{n} \mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}] \quad (5)$$
$$s.t., \sum_{i=1}^{n} z^{(i)} \leq b$$
$$c_{ia'} \leq \tau \quad \forall a' \in \mathcal{A},\ i \in \{1, \ldots, n\},$$

where $\mathcal{A}$ is the domain of A and $\tau \geq 0$. We stress that using non-descendants of A is not necessary for our formulation. In the appendix we describe a setup that uses structural equations with arrows from A to X.

4.1. The Optimization Framework

We propose a procedure to solve eq. (5) optimally. As eq. (5) is NP-hard, our procedure will run in exponential time in the worst case. However, in practice it runs extremely fast (see Figure 6). Our formulation is general enough to accommodate any functional form for the structural equation for Y. To do so, we reformulate eq. (5) as a mixed-integer-linear-program (MILP). To avoid fractional solutions from the MILP for intervention set $\mathbf{z}$, we use integer constraints to enforce that each intervention $z^{(i)}$ is binary in the final solution. Recall that for each individual i there are a set of neighbor individuals $N(i)$ whose interventions interfere on their impact $Y^{(i)}$. Specifically, we let $N(i)$ be the nearest K neighbors. Let these interventions be called $\mathbf{z}^{N(i)}$. We begin by introducing a fixed auxiliary matrix $\mathbf{E} \in \{0,1\}^{(2^K, K)}$. Each row $\mathbf{e}_j$ corresponds to one of the possible values that $\mathbf{z}^{N(i)}$ can take (i.e, all possible K-length binary vectors).

Additionally we introduce a matrix $\mathbf{H} \in [0,1]^{(n, 2^K)}$ where each row $\mathbf{h}_i$ indicates for individual i, which of the 2^K possible neighbor interferences affect $Y^{(i)}$ (i.e., each row is a 1-hot vector). We will optimize $\mathbf{H}$ jointly with $\mathbf{z}$. This allows us to rewrite the objective of eq. (5) as:

$$\sum_{i=1}^{n} \sum_{j=1}^{2^k} h_{ij} \underbrace{\mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}^{N(i)} = \mathbf{e}_j) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}]}_{\xi^{ij}(a^{(i)})}.$$

Note that we introduce a sum over all possible $\mathbf{z}^{N(i)}$ and use $\mathbf{H}$ to indicate which element of this sum is non-zero. We can rewrite the constraints in a similar way. To ensure that each row $\mathbf{h}_i$ agrees with the actual $\mathbf{z}^{N(i)}$ we enforce the following constraints: $\mathbb{I}[\mathbf{e}_j = 1] h_{ij} \leq \mathbf{z}^{N(i)}$ and $\mathbb{I}[\mathbf{e}_j = 0] h_{ij} \leq 1 - \mathbf{z}^{N(i)}$, where $\mathbb{I}$ is the indicator function that operates on each element of a vector. The first constraint ensures that the non-zero entries of $\mathbf{e}_j$ are consistent with $\mathbf{z}^{N(i)}$ via h_{ij}, and the second ensures the zero entries agree. Finally, to ensure that each row of $\mathbf{H}$ is 1-hot we introduce the constraint $\sum_{j=1}^{2^K} h_{ij} = 1$ for all i. This yields the following optimization program:

$$\max_{\substack{\mathbf{z}\in\{0,1\}^n \\ \mathbf{H}\in[0,1]^{(n,2^K)}}} \sum_{i=1}^{n} \sum_{j=1}^{2^K} h_{ij} \xi^{ij}(a^{(i)}) \qquad (6)$$
$$s.t., \sum_{j=1}^{2^K} h_{i,j} \left[\xi^{ij}_\prec(a^{(i)}) - \xi^{ij}_\prec(a')\right] < \tau, \quad \forall a', i$$
$$\mathbb{I}[\mathbf{e}_j = 1] h_{ij} \leq \mathbf{z}^{N(i)}, \qquad \forall i, j$$
$$\mathbb{I}[\mathbf{e}_j = 0] h_{ij} \leq 1 - \mathbf{z}^{N(i)}, \qquad \forall i, j$$
$$\sum_{j=1}^{2^K} h_{ij} = 1, \qquad \forall i$$
$$\sum_{i=1}^{n} z^{(i)} \leq b.$$

Here $\xi^{ij}_\prec(a^{(i)})$ means the expectation is conditioned on $X^{(i)}_\prec$ as in eq. (4), and $\xi^{ij}_\prec(a')$ means we're taking the expectation of counterfactual $Y^{(i)}(a', \mathbf{z}^{N(i)} = \mathbf{e}_j)$.

Other Fairness Constraints. Our formulation eq. (5) and our optimization framework eq. (6) is general enough to

[5] Depending on the graph, it is possible to identify the functionals without the structural equations (Nabi & Shpitser, 2018).

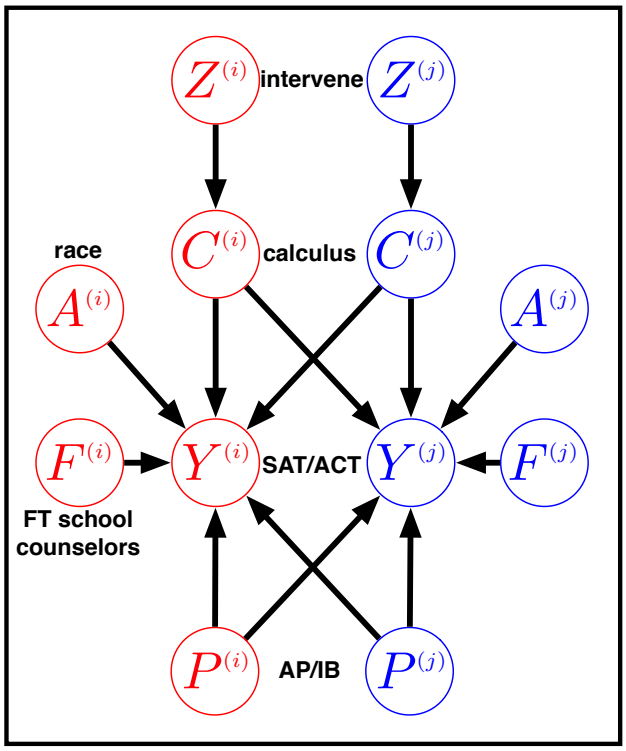

Figure 3. The model for the NYC school dataset.

handle any fairness constraint that can be phrased as an (in)equality. In the appendix we detail how our framework can handle for example (a) parity constraints, and (b) optimizing purely for minority outcomes.

5. Results

We now demonstrate our technique on a real-world dataset.

Dataset. We compiled a dataset on 345 high schools from the New York City Public School District, largely from the Civil Rights Data Collection (CRDC)[6]. The CRDC collects data on U.S. public primary and secondary schools to ensure that the U.S. Department of Education's financial assistance does not discriminate 'on the basis of race, color, national origin, sex, and disability.' This dataset contains the distribution of race (A), *Full-time Counselors* (F): the number of full-time counselors employed (fractional values indicate part-time work), *AP/IB* (P): if the school offers Advanced Placement or International Baccalaureate classes, *Calculus* (C): whether the school offers Calculus courses, and *SAT/ACT-taking* (Y): the percent of students who take the college entrance examinations, the SAT and/or the ACT.

Setup. In this experiment, we imagine that the U.S. Department of Education wishes to intervene to offer financial assistance to schools to hire a Calculus teacher, a class that is commonly taken in the U.S. at a college level. The goal is to increase the number of students that are likely to attend college, as measured by the fraction of students taking the entrance examinations (via *SAT/ACT-taking*). It is reasonable to assume that this intervention is exact. Specifically, if the intervention is given to school i, i.e., $Z^{(i)}=1$, then we assume that the school offers Calculus, i.e., $C^{(i)}=1$. Without considerating discrimination, the Department would simply assign interventions to maximize the total expected percent of students taking the SAT/ACT until they reach their allocation budget B. However, to ensure we allocate interventions to schools that will benefit *independent of their societal privilege due to race* we will learn a model using the discrimination-reducing constraints described in eq. (5). We begin by formulating a causal model that describes the relationships between the variables.

Causal Model. The structure of the causal model we propose is shown in Figure 3 (a subset of the graph is shown for schools i and j). Recall that technically $Z^{(i)}$ does not directly cause observable variables. $C^{(i)}$ is hidden to the extent that its value is only observable after the action takes place. All variables directly affect the impact $Y^{(i)}$ (SAT/ACT-taking). Frequently schools will allow students from nearby schools to take classes that are not offered at their own school. Thus we model both the Calculus class variables C and the AP/IB class variables P as affecting the impact of students at neighboring schools. Specifically, we propose the following structural equations for Y with interference:

$$\begin{aligned}\mathbb{E}[Y^{(i)}(\mathbf{a},\mathbf{z}) \mid A^{(i)}=\mathbf{a}^{(i)}, P^{(i)}=p^{(i)}, F^{(i)}=f^{(i)}] = \\ \boldsymbol{\alpha}^\top \mathbf{a} \max_{\substack{j\in N(i)\\ s.t., z^{(j)}=1}} s(i,j) C^{(j)}(\mathbf{z}) \\ + \boldsymbol{\beta}^\top \mathbf{a} \max_{\substack{j\in N(i)\\ s.t., z^{(j)}=1}} s(i,j) p^{(j)} \\ + \boldsymbol{\gamma}^\top \mathbf{a} f^{(i)} + \boldsymbol{\theta}^\top \mathbf{a}.\end{aligned} \tag{7}$$

To simplify notation we let $N(i)$ refer to the nearby schools of school i *and also* i. This way the max terms are also able to select i (if $z^{(i)}=1$). Further, $C^{(j)}(\mathbf{z}) = z^{(j)}$ and $s(i,j)$ is the similarity of schools i and j. We construct both $N(i)$ and $s(i,j)$ using GIS coordinates for each school in our dataset[7]: $N(i)$ is the nearest $K=5$ schools to school i and $s(i,j)$ is the inverse distance in GIS coordinate space. The vector $\mathbf{a}^{(i)}$ is the proportion of $(black, Hispanic, white)$ students at school i. We fit the parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}$ via maximum likelihood using an observed dataset $\{c^{(i)}, \mathbf{a}^{(i)}, p^{(i)}, f^{(i)}, y^{(i)}\}_{i=1}^n$. For counterfactuals a' our goal is to judge the largest impact due to race, so we consider the extreme counterfactuals: where

[6] https://ocrdata.ed.gov/

[7] https://data.cityofnewyork.us/Education/School-Point-Locations/jfju-ynrr

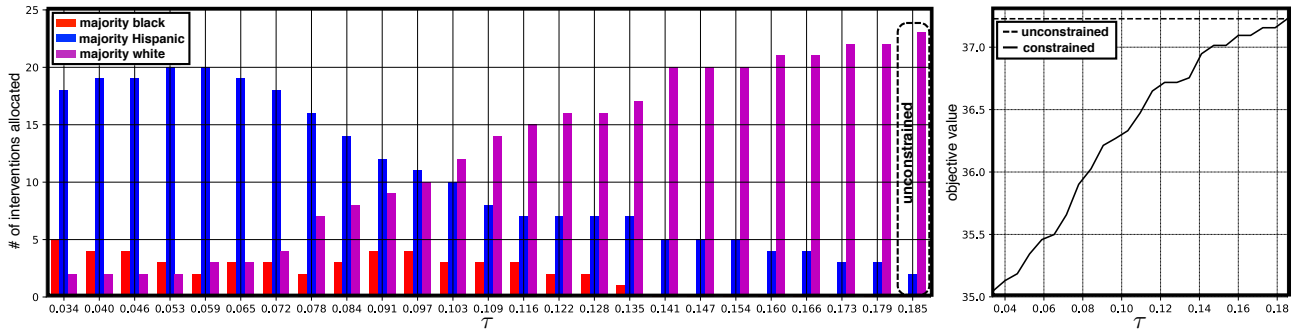

Figure 4. The resulting interventions for the NYC school dataset with and without discrimination-reducing constraints. See text for details.

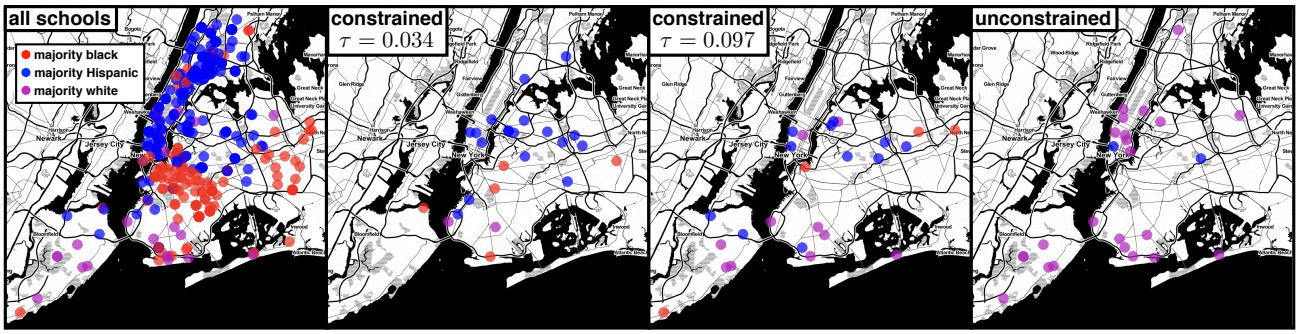

Figure 5. The left-most plot shows the locations of the 345 New York City High Schools, and their majority race. The remaining plots show the allocations of interventions for each policy.

each school consists of students of a single race, either $(black, Hispanic, white)$ students. Thus, to plug these three counterfactuals into eq. (7), we encode them as one-hot vectors, e.g., $\mathbf{a}' = [1, 0, 0]$ signifies the majority black school counterfactual.

Results. To evaluate the effect on SAT/ACT-taking when allocating Calculus courses we start with the null allocation vector $\mathbf{z} = 0$ (i.e., no school has a Calculus course). We then solve the optimization problem in eq. (5) (using the MILP framework in Section 3.3) with the structural equation for Y in eq. (7), and a budget b of 25 schools. We use the Python interface to the Gurobi optimization package to solve the MILP[8]. The results of our model is shown in Figure 4. The left plot shows the number of interventions allocated to schools by the majority race of each school. The right plot shows the objective value achieved by the constrained and unconstrained models. On the far right of the left plot is the unconstrained allocation. In this case, all interventions but 2 are given to majority white schools. When τ is small both majority black and Hispanic schools receive allocations, indicating that these schools benefit the least from their race. As τ is increased, Hispanic school allocations increase, then decrease as majority white schools are allocated.

Figure 5 shows how each policy allocates these interventions on a map of New York City. The constrained policy ($\tau = 0.034$, the first set of bars in Figure 4) assigns interventions to majority Hispanic and majority black schools that have high utility because of things not due to race. As τ is increased ($\tau = 0.097$, the eleventh set of bars in Figure 4) the allocation includes more majority white schools, less Hispanic schools, and roughly the same number of majority black schools, with more allocations in Staten Island. The unconstrained policy assigns interventions to schools in lower Manhattan and Brooklyn, and all allocations except two are to white schools. See the appendix for results on the run-time of the MILP (under 5 minutes for all settings) and different fairness constraints.

6. Conclusion

In this paper we describe the discriminatory impact problem, a problem that has gained much recent attention, but for which no general solution exists. We argue that causal models are a perfect tool to model how impact is affected by decisions and real-world factors. We then propose a solution to the problem: an optimization problem with counterfactual constraints from a causal model of the impact. We give an efficient procedure for solving this optimization problem and demonstrate it on a course allocation problem for New York City schools. We believe this is just the tip of the iceberg; there are many possibilities for future work around designing new constraints, optimization procedures, and causal models, while reducing necessary assumptions.

[8] https://github.com/mkusner/reducing_discriminatory_impact

Acknowledgments

This work was supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1. CR acknowledges additional support under the EPSRC Platform Grant EP/P022529/1.

References

Arbour, D., Garant, D., and Jensen, D. Inferring network effects in relational data. *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 715–724, 2016.

Aronow, P. M. and Samii, C. Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics*, 11:1912–1947, 2017.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *arXiv preprint:1703.09207*, 2017.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pp. 4349–4357, 2016.

Chiappa, S. and Gillam, T. Path-specific counterfactual fairness. *arXiv:1802.08139*, 2018.

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

Dawid, A. P. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70:161–189, 2002.

Dwork, C. and Ilvento, C. Fairness under composition. *arXiv preprint arXiv:1806.06122*, 2018.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, pp. 214–226. ACM, 2012.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. D. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pp. 119–133, 2018.

E. L. Ogburn, T. J. V. Causal diagrams for interference. *Statistical Science*, 29:559–578, 2014.

Edwards, H. and Storkey, A. Censoring representations with an adversary. *arXiv preprint:1511.05897*, 2015.

Elzayn, H., Jabbari, S., Jung, C., Kearns, M., Neel, S., Roth, A., and Schutzman, Z. Fair algorithms for learning in allocation problems. *arXiv preprint arXiv:1808.10549*, 2018.

Flores, A. W., Bechtel, K., and Lowenkamp, C. T. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016.

Green, B. and Chen, Y. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 90–99. ACM, 2019.

Hardt, M., Price, E., Srebro, N., et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Heidari, H., Nanda, V., and Gummadi, K. P. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. In *ICML*, 2019.

Kamiran, F. and Calders, T. Classifying without discriminating. In *International Conference on Computer, Control and Communication*, pp. 1–6. IEEE, 2009.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50. Springer, 2012.

Kannan, S., Roth, A., and Ziani, J. Downstream effects of affirmative action. *arXiv preprint arXiv:1808.09004*, 2018.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pp. 656–666, 2017.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint:1609.05807*, 2016.

Komiyama, J. and Shimao, H. Comparing fairness criteria based on social outcome. *arXiv preprint arXiv:1806.05112*, 2018.

Kusner, M., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30:4066–4076, 2017.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9, 2016.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*, 2018.

Loftus, J., Russell, C., Kusner, M., and Silva, R. Causal reasoning for algorithmic fairness. *arxiv:1805.05859*, 2018.

Madras, D., Pitassi, T., and Zemel, R. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, pp. 6147–6157, 2018.

Nabi, R. and Shpitser, I. Fair inference on outcomes. *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Nabi, R., Malinsky, D., and Shpitser, I. Learning optimal fair policies. In *ICML*, 2019.

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

Peters, G. W. Statistical machine learning and data analytic methods for risk and insurance. 2017.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5684–5693, 2017.

Russell, C., Kusner, M., Loftus, J., and Silva, R. When worlds collide: integrating different counterfactual assumptons in fairness. *Advances in Neural Information Processing Systems*, 30:6417–6426, 2017.

Sobel, M. What do randomized studies of housing mobility demonstrate? *Journal of the American Statistical Association*, 101:1398–1407, 2006.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction and Search*. Lecture Notes in Statistics 81. Springer, 1993.

Sweeney, L. Discrimination in online ad delivery. *Queue*, 11(3):10, 2013.

Yang, S., Korayem, M., AlJadda, K., Grainger, T., and Natarajan, S. Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive statistical relational learning approach. *Knowledge-Based Systems*, 136:37–45, 2017.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *World Wide Web Conference*, pp. 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

Zeng, J., Ustun, B., and Rudin, C. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3):689–722, 2017.

Zhang, J. and Bareinboim, E. Fairness in decision-making: The causal explanation formula. In *AAAI Conference on Artificial Intelligence*, 2018.