TECHNICAL REPORT
R-30-L
November, 2017

Fairness in Decision-Making – The Causal Explanation Formula

Junzhe Zhang and **Elias Bareinboim**
Purdue University
{zhang745,eb}@purdue.edu

Abstract

AI plays an increasingly prominent role in society since decisions that were once made by humans are now delegated to automated systems. These systems are currently in charge of deciding bank loans, criminals' incarceration, and the hiring of new employees, and it's not difficult to envision that they will in the future underpin most of the decisions in society. Despite the high complexity entailed by this task, there is still not much understanding of basic properties of such systems. For instance, we currently cannot detect (neither explain nor correct) whether an AI system is operating fairly (i.e., is abiding by the decision-constraints agreed by society) or it is reinforcing biases and perpetuating a preceding prejudicial practice. Issues of discrimination have been discussed extensively in legal circles, but there exists still not much understanding of the formal conditions that an automated system must adhere to be deemed fair. In this paper, we use the language of structural causality (Pearl, 2000) to fill in this gap. We start by introducing three new fine-grained measures of transmission of change from stimulus to effect called counterfactual direct (Ctf-DE), indirect (Ctf-IE), and spurious (Ctf-SE) effects. Building on these measures, we derive the causal explanation formula, which allows the AI designer to quantitatively evaluate fairness and explain the total observed disparity of decisions through different discriminatory mechanisms. We apply these results to various discrimination analysis tasks and run extensive simulations, including detection, evaluation, and optimization of decision-making under fairness constraints. We conclude studying the trade-off between different types of fairness criteria (outcome and procedural), and provide a quantitative approach to policy implementation and the design of fair decision-making systems.

Introduction

Automated systems based on artificial intelligence, machine learning, and statistics have been increasingly applied throughout a wide range of real-world decision-making scenarios, including in healthcare, law enforcement, education, and finance (Mahoney and Mohen 2007; Brennan, Dieterich, and Ehret 2009; Khandani, Kim, and Lo 2010; Sweeney 2013; Angwin et al. 2016). It is no longer far-fetched to envision a future where fully autonomous AI systems will be driving entire business decisions and, more broadly, supporting large-scale decision-making infrastructure to solve society's most challenging problems. Issues of unfairness and discrimination are pervasive when decisions are being made by humans, which, unfortunately, are not automatically solved, and can even be amplified, when machines are put in control. For instance, an AI system designed to decide incarceration of recidivist criminals may be trained with data that contains historical biases of judges that discriminated against certain races, which potentially may lead to even more discriminatory practices without much transparency or accountability. If our goal is to design systems that are ethical and fair, it is imperative to have a more refined understanding of the properties of automated decision-making in complex and uncertain scenarios.

Discrimination can be broadly partitioned into two components: *direct* and *indirect* (Council 2004). The former is concerned with settings where individuals receive less favorable treatments on the basis of a protected attribute X such as race, religion, or gender. Some extreme cases of direct discrimination include voting rights and unequal payment based on race and gender (Altonji and Blank 1999; Derfner 1973). The latter is concerned with individuals who receive treatments on the basis of inadequately justified factors that are somewhat related with (but not the same as) the protected attribute. These cases are arguably more complex to characterize, and require a more refined reasoning. One well-known example is *redlining*, where financial institutions (e.g., banks, insurance companies) deny services to residents of geographic areas in different rates, which wouldn't be necessarily a problem by itself, if not for the fact that these areas have considerably different racial and ethnic compositions. In practice, this may entail that the given institution is using the location of the applicants as a proxy to an obviously discriminatory attribute (e.g., race).

These types of discrimination (direct and indirect) are supported by two legal frameworks applied in large bodies of cases throughout the US and the EU – *disparate treatment* and *disparate impact* (Council 2004; Barocas and Selbst 2016). The disparate treatment framework enforces *procedural fairness*, namely, the equality of treatments that prohibits the use of the protected attribute in the decision process. The disparate impact framework guarantees *outcome fairness*, namely, the equality of outcomes among protected groups. Disparate impact discrimination occurs if a facially

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

neutral practice has an adverse impact on members of the protected attribute.

There is a growing literature in AI that is concerned with issues of transparency and fairness following, more or less explicitly, these two legal frameworks, including (not exhaustively) (Dwork et al. 2012; Romei and Ruggieri 2014; Mancuhan and Clifton 2014; Datta, Sen, and Zick 2016; Barocas and Selbst 2016; Hardt et al. 2016; Zhang, Wu, and Wu 2016; Kusner et al. 2017; Zafar et al. 2017b; 2017a; Chouldechova 2017; Kilbertus et al. 2017; Pleiss et al. 2017; Bonchi et al. 2017). Despite all the recent progress in the field, there is still not a clear understanding of the various metrics used to evaluate each type of discrimination individually. In practice, this translates into the current state of affairs where the fairness criterion is, almost invariably, chosen without much discussion or justification. Our goal is to fill in this gap by providing a principled approach to assist the data scientist making an informed decision about the metric used to ascertain fairness while being fully aware of the tradeoffs involved with her choice.

We build on the language of causality (Pearl 2000; Halpern 2000; Bareinboim and Pearl 2016) to express direct and indirect discrimination through the different paths connecting the protected attribute X and the outcome Y in the underlying causal diagram (see Fig. 1). For instance, direct discrimination is modeled by the direct causal path from X to Y ($X \rightarrow Y$ in Fig. 1(a)). Indirect discrimination can be further divided into two categories based on the causal mechanisms evoked to transmit change, namely: indirect *causal* discrimination, captured by indirect causal paths, i.e., one-directional paths that trace arrows pointing from X to Y except for the direct link $X \rightarrow Y$ (e.g., $X \rightarrow W \rightarrow Y$); indirect *spurious* discrimination, corresponding to all paths between X and Y but the causal ones (direct and indirect), called *spurious* paths (e.g., the *backdoor* path $X \leftarrow Z \rightarrow Y$). We will refer to these discriminatory mechanisms as *direct*, *indirect*, and *spurious* discrimination, respectively. We will soon show that no discrimination (or fairness) measure is capable of detecting and distinguishing the effects of the different discrimination mechanisms commonly found decision-making settings.

Our proposed method will decompose the observed disparities (measured as the total variation, to be defined) according to the different paths in the underlying causal (decision-making) diagram. The study of effect decomposition is indeed not new, going back to Wright's method of path analysis in linear causal models (Wright 1923; 1934) (for a survey, see (Pearl 2000, Ch. 5)). The path analysis method gained momentum in the social sciences during 1960's, becoming extremely popular in the form of the mediation formula in which the total effect of X on Y is decomposed into direct and indirect components (Baron and Kenny 1986; Bollen 1989; Duncan 1975; Fox 1980). [1] The bulk of this literature, however, required a commitment to a particular parametric form, thus falling short of providing a general method for analyzing natural and social phenomena with nonlinearities and interactions (MacKinnon 2008).

[1]Google Scholar currently counts Baron's paper, where the mediation formula appeared, as having more than 70,000 citations.

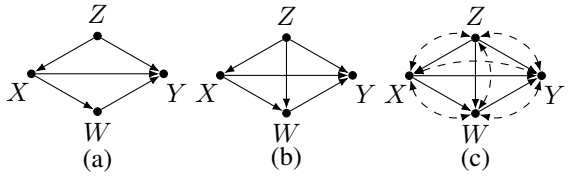


Figure 1: The causal diagrams for (a) a non-confounding model where X stands for the protected attribute, Y for the outcome, W the mediator, and Z an observed confounder; (b) the standard model where Z could also affect W; (c) the extended standard model that allows for the presence of unobserved confounders. See also footnote 3.

Pearl introduced the *causal mediation formula* for arbitrary non-parametric models, which decomposes the total effect into what was called the natural direct (NDE) and indirect (NIE) effects (Pearl 2001) (see also (Imai, Keele, and Yamamoto 2010; Imai et al. 2011; VanderWeele 2015). By and large, the literature of mediation analysis has then been focused almost exclusively on decomposing *total effects*, which we note needs to be relaxed to satisfy the requirements of fairness analysis. In particular, one is not concerned with how controlling the attribute X may bring about change in Y, but in understanding how natural variations of X affect the outcome Y, which also happens through confounding variables (e.g., redlining). Specifically, we generalize mediation analysis through the following contributions:

- We define the *counterfactual direct (Ctf-DE)*, *indirect (Ctf-IE)*, and *spurious (Ctf-SE) effects* based on the language of non-parametric structural causal models. These measures allow one, for the first time, to precisely detect and distinguish the three most natural types of discrimination, namely, direct, indirect, and spurious.
- We derive the *causal explanation formula*, which allows one to understand how the total $X - Y$ variation can be non-parametrically decomposed in terms of the counterfactual measures (Ctf-DE, IE, SE). In practice, this allows one to explain the effect of a certain discriminatory mechanism in terms of the observed disparity found in the data.
- We quantify explicitly the trade-off between the two fairness principles encountered in the literature (procedural and outcome) and study its implication in designing new policies in unfair settings. This result is relevant since the design of reparatory policies (e.g., affirmative actions) has been a more informal than principled exercise. Our results show that an even more unfair state of affairs may come about depending on how the reparatory policy is designed.

Preliminaries

Variables will be denoted by capital letters (e.g., X) and their values by small letters (x). We will use *Structural Causal Models* (SCMs) as the basic semantical framework of our analysis (Pearl 2000, Ch. 7), which is defined next.

Definition 1 (SCM (Pearl 2000))**.** A structural causal model (SCM) M is a 4-tuple $\langle U, V, F, P(U)\rangle$ where:

- U is a set of exogenous (unobserved) variables, which are determined by factors outside of the model;
- V is a set $\{V_1, \ldots, V_n\}$ of endogenous (observed) variables that are determined by variables in the model (i.e., by the variables in $U \cup V$);
- F is a set of structural functions $\{f_1, \ldots, f_n\}$ where each f_i is a process by which V_i is assigned a value $v_i \leftarrow f_i(v, u)$ in response to the current values of V and U;
- $P(u)$ is a distribution over the exogenous variables U.

Each SCM M is associated with a causal diagram G, which is a directed acyclic graph where nodes correspond to the endogenous variables (V) and the directed edges the functional relationships. Any exogenous in U is not shown explicitly in the graph, but for when it affects more than one endogenous variable. In this case, a bi-directed edge will be used to indicate the presence of the unobserved confounder (UC) affecting both variables (e.g., Fig. 1(c)).

An intervention, denoted by $do(X = x)$ (Pearl 2000, Ch. 3), represents a model manipulation where the values of a set of variables X are set fixed to x regardless of how their values are ordinarily determined (f_x). We use the counterfactual distribution $P(Y_{X=x} = y)$ to denote the causal effect of the intervention $do(X = x)$ on the outcome Y, where the counterfactual variable $Y_{X=x}$ (Y_x, for short) denotes the potential response of Y to intervention $do(X = x)$. We will consistently use the abbreviation $P(y_x)$ for the probabilities $P(Y_{X=x} = y)$, so does $P(y|x) = P(Y = y|X = x)$.

We next introduce state-of-the-art discrimination measures for both procedural and outcome fairness. [2] In this paper, we will let X be the sensitive feature (whose effect we seek to assess), and Y be the outcome variable. We will let Z stand for the observed common ancestors of X and Y (called confounders), and W for all the observed intermediate variables between X and Y (called mediators); for an example, see Fig. 1(b), which we name the *standard fairness model* (for short, standard model) for its generality. Formally speaking, the standard model can fit any observational distribution over X, Y, Z, W since it implies no independence constraint.[3] We will denote by value x_1 the disadvantaged group and x_0 the advantaged one, which we will use as the baseline to measure changes of the outcome. One popular criterion for the outcome fairness is the demographic parity (Zafar et al. 2015), measured by the total variation:

Definition 2 (Total Variation (TV))**.** The total variation of event $X = x_1$ on $Y = y$ (with baseline x_0) is defined as:

$$TV_{x_0,x_1}(y) = P(y|x_1) - P(y|x_0) \tag{1}$$

[2]We will not discuss in this paper the measure known as equalized odds (EO) (Hardt et al. 2016) since it is specifically defined for supervised learning tasks. EO measures the disparate mistreatment (Zafar et al. 2017a), which is somewhat orthogonal to the disparate treatment and disparate impact frameworks.

[3]To avoid clutter in Fig. 1(b), we just depict $Z \rightarrow X$ but note that under the standard model, $Z \leftarrow -- \rightarrow X$ (Z and X are correlated) is an equally valid relationship that may be present.

The TV is nothing more than the difference between the conditional distributions of Y when (passively) observing X changing from x_0 to x_1. Another common fairness criterion is the total effect (Council 2004), which measures the difference of outcome Y while physically controlling the values of X, namely, $TE_{x_0,x_1}(y) = P(y_{x_1}) - P(y_{x_0})$. Control in this case is usually achieved through the process of randomization. Outcome fairness can also be measured using counterfactual quantities. One fundamental counterfactual metric is the effect of treatment on the treated (Pearl 2000), i.e., :

Definition 3 (Effect of Treatment On the Treated (ETT))**.** The effect of treatment on the treated of intervention $X = x_1$ on $Y = y$ (with baseline x_0) is defined as:

$$ETT_{x_0,x_1}(y) = P(y_{x_1}|x_0) - P(y|x_0) \tag{2}$$

The first factor $P(y_{x_1}|x_0)$ is a counterfactual quantity that read as "the probability of Y woud be y had X been x_1 (counterfactually), given that in the actual (factual) world $X = x_0$." (Kusner et al. 2017) studied fairness using the ETT conditioned on the sub-population $Z = z, W = w$.

Procedural fairness, which prohibits direct discrimination, is arguably the most intuitive fairness criterion found in the literature. Some even believe that it is the only valid rationale for antidiscrimination law (Barocas and Selbst 2016). In order to detect direct discrimination, one popular approach is to use the controlled direct effect (CDE), which measures the effect of X on Y while holding all the other variables W, Z fixed (also known as the *ceteris paribus* condition). Formally, $CDE_{x_0,x_1}(y_{z,w}) = P(y_{x_1,z,w}) - P(y_{x_0,z,w})$. (Datta, Sen, and Zick 2016) introduced a set of Quantitative Input Influence (QII) measures which identify the direct discrimination when parents of Y are fully observed.

(Pearl 2001) introduced natural direct (NDE) and indirect (NIE) effects to measure the effect of, respectively, the direct and indirect causal paths on the total effect of X on Y. For example, the *natural direct effect* in Fig. 1(a) is written as $NDE_{x_0,x_1}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_0})$, which measures the effect of the direct causal path $X \rightarrow Y$ and differs from the CDEs since the mediator W is set the W_{x_0}, the level that it would have naturally attained under the reference condition $X = x_0$. The definition of NDEs can be turned around and provide an operational definition for the indirect effect. The *natural indirect effect* is defined as $NIE_{x_0,x_1}(y) = P(y_{x_0,W_{x_1}}) - P(y_{x_0})$, which compares the effect of the mediator W at levels W_{x_0} and W_{x_1} on the outcome Y had X been x_0. This framework has been used as the basis for a discrimination discovery analysis under the assumption that X has no parent node in the causal diagram (no spurious discrimination) (Zhang, Wu, and Wu 2016).

Counterfactual Fairness Analysis

Despite the recent surge of interest in discrimination analysis and fairness learning in AI, two fundamental questions have rarely been discussed – (1) What discrimination's mechanism is the target of the analysis? and (2) What empirical measures would allow this mechanism to be identified and potentially controlled? In this section, we start by illustrating these points by showing how previous state-of-the-art

fairness measures can fail when kept oblivious to the discussion of the causal mechanisms that bring about discrimination in the real world. We will consider a simple example of a legal dispute over religious discrimination in hiring.

A company makes hiring decisions Y (0 for not hire, 1 for hire) and can potentially use the following attributes available in its database: (1) the religious belief X (0 for non-believer, 1 for believer), (2) the educational background Z $(0, 1$ for low, high), and (3) the location W of the applicant (0 for close to religious institutes, 1 for distant). Attributes critical for the business success can be used for determining hiring, which implies in this case that the level of education is considered a legitimate attribute to be used during the hiring process. In this society, greater levels of education curb individuals' religious participation. Also, applicants of faith tend to live closer to religious institutions due to their desire for a higher engagement in religious activities. The standard model (Fig. 1(a)) is an accurate representation of this setting. In reality, the hiring decision Y is made solely based on Z, where the company only hires applicants with higher education (i.e., the structural function $f_y(z) = z$). The company is sued by a frustrated applicant, and the court requests data from hundreds of previous cases to be analyzed. The court asks for justification from the company after observing a demographic disparity in religion composition among hired employees ($TV_{x_0,x_1}(y) = 1$, where $Y = 1$). The company argues that the disparity is mainly due to education opportunities enjoyed by the applicants (graphically, the path $X \leftarrow Z \rightarrow Y$) instead of any direct and indirect discrimination (paths $X \rightarrow Y$ and $X \rightarrow W \rightarrow Y$).

Trying to verify whether this justification could explain the observed disparity in the data, an expert witness runs a batch of discrimination tests, including the well-known measures TE, ETT, NIE, CDE, NDE, and QII. Results are shown in Fig. 2(a). To her surprise and confusion, none of these measures captured any discriminatory effect (e.g., $TE_{x_0,x_1}(y) = 0$). The natural question that arises in this case is how the presence of demographic disparity ($TV_{x_0,x_1}(y) = 1$) could be explained if none of the mechanisms came out as effective?

Remarkably, the true decision function f_y is, in general, not known a priori, and only the combination of past data (representing the previous decisions) and a measure of transmission of change from X to Y can lead to an understanding about how decisions are really being made.

In the religious example, the only effective causal mechanism is due to the spurious path ($X \leftarrow Z \rightarrow Y$), which is not covered by any discrimination measure known in the literature. These measures are able to capture only direct and indirect discrimination, which correspond to the causal paths $X \rightarrow Y$ and $X \rightarrow W \rightarrow Y$. Despite the fact that the total variation (*TV*) also encompass spurious effects transmitted along the non-causal paths, it's unable to disentangle spurious discrimination from the other types of discrimination.

We analyze properties of state-of-the-art discrimination measures and summarize them in Table. 1 (for details, see the extended technical report (Zhang and Bareinboim 2018)). A check (cross) mark represents that the measure is (is not) able to detect the discrimination in the corresponding column. A measure with only one check mark in a row is able also to distinguish the corresponding discrimination. We note that there exists no set of measures that could identify all three types of discrimination individually.

Fairness	Measures	Discrimination		
		Direct	Indirect	Spurious
Outcome	TV	✓	✓	✓
	TE	✓	✓	✗
	ETT	✓	✓	✗
	NIE	✗	✓	✗
Procedure	NDE	✓	✗	✗
	QII	✓	✗	✗
	CDE	✓	✗	✗
Our Approach	Ctf-DE	✓	✗	✗
	Ctf-IE	✗	✓	✗
	Ctf-SE	✗	✗	✓

Table 1: Summary of discrimination measures. Outcome and Procedure represent different fairness principles governed by disparate impact and treatment, respectively. Checks/crosses stand for whether the measure is able or not to detect a certain type of discrimination.

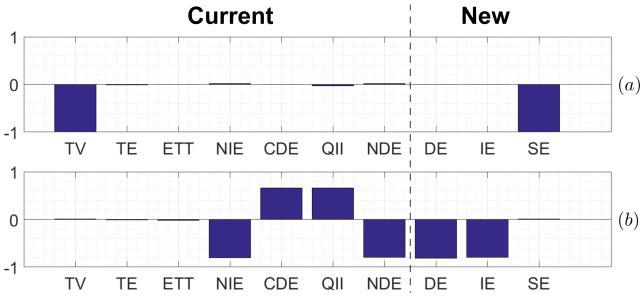


Figure 2: Simulation results of (a) the religious discrimination example (Fig. 1(a)) where state-of-art measures (TV, TE, etc) are unable to identify the mechanism behind the observed demographic disparity; (b) discrimination detection in a confounding model described in Fig. 1(c). DE, IE, SE stand for counterfactual direct, indirect, and spurious effect.

We next introduce a set of counterfactual measures that will allow for a more fine-grained explanation of the (total) observed disparity in terms of the underlying discriminatory mechanisms (direct, indirect, and spurious). We define next the *counterfactual direct effect* (Ctf-DE), which measures NDEs conditioned on the sub-population $X = x$.

Definition 4 (Counterfactual Direct Effect (Ctf-DE))**.** Given a SCM M, the counterfactual direct effect of intervention $X = x_1$ on Y (with baseline x_0) conditioned on $X = x$ is:

$$DE_{x_0,x_1}(y|x) = P(y_{x_1,W_{x_0}}|x) - P(y_{x_0}|x) \qquad (3)$$

The nested sentence $Y_{x_1,W_{x_0}} = y|X = x$ in the first factor is a more involved counterfactual compared to NDEs, and read as "the probability Y would be y had X been x_1, while W is kept at the same value that it *would have attained* had X been x_0, given that X was actually equal to x". [4] If $x = x_0$, $DE_{x_0,x_1}(y|x_0)$ measures the change in

[4]As a corollary, it is immediate to see if $DE_{x_0,x_1}(y|x) = 0$ for

probability of the outcome Y would be y (e.g., to hire) !had X been x_1, while mediators W (e.g., location) is kept at the level they would have naturally attained (had X been x_0) for the individuals that $X = x_0$.[5] Despite the apparently non-trivial reading of this nested counterfactual sentence, $DE_{x_0,x_1}(y|x)$ simply captures, mathematically, the existence of disparate treatment., which is shown below.

Property 1. *For a SCM M, if X has no direct causal path connecting Y in the causal diagram G, then $DE_{x_0,x_1}(y|x) = 0$, for any $x, y, x_0 \neq x_1$.*

In other words, if $DE_{x_0,x_1}(y|x) \neq 0$, for some values $x, y, x_0 \neq x_1$, one can conclude that the function f_y uses the sensitive feature X as an input to decide for the values of outcome Y, i.e., the existence of direct discrimination. Similarly, we could turn around the definition of Ctf-DE and formally define the counterfactual notion of indirect effect.

Definition 5 (Counterfactual Indirect Effect (Ctf-IE))**.** Given a SCM M, the counterfactual indirect effect of intervention $X = x_1$ on $Y = y$ (relative to baseline $X = x_0$) conditioned on $X = x$ is:

$$IE_{x_0,x_1}(y|x) = P(y_{x_0,W_{x_1}}|x) - P(y_{x_0}|x) \tag{4}$$

Syntactically, the definition of counterfactual IE is identical to counterfactual DEs except for the switch of x_0 and x_1 in the first term. For $x = x_0$, $IE_{x_0,x_1}(y|x_0)$ measures changes in the probability of the outcome Y would be y had X been x_0, while changing W to whatever level it would have obtained had X been x_1, in particular, for the individuals that (naturally) have $X = x_0$. The following property establishes the relationship between the existence of $IE_{x_0,x_1}(y|x)$ and indirect causal paths between X and Y.

Property 2. *For a SCM M, if X has no indirect causal path connecting Y in the causal diagram G, then $IE_{x_0,x_1}(y|x) = 0$, for any $y, x, x_0 \neq x_1$.*

Prop. 2 implies that indirect discrimination could be sufficiently identified by checking the condition $IE_{x_0,x_1}(y|x) \neq 0$. In words, the doubly hypothetical criterion used in Ctf-IE correctly describes the meaning of indirect discrimination.

Finally, we provide a novel operational definition to quantitatively capture spurious associations between the protected attribute X and the outcome Y.

Definition 6 (Counterfactual Spurious Effect (Ctf-SE))**.** Given a SCM M, the spurious effect of event $X = x_1$ on $Y = y$ (relative to baseline x_0) is defined as:

$$SE_{x_0,x_1}(y) = P(y_{x_0}|x_1) - P(y|x_0) \tag{5}$$

$SE_{x_0,x_1}(y)$ measures the difference in the outcome $Y = y$ had X been x_0 (written y_{x_0}) for the individuals that would naturally choose X to be x_0 versus x_1. For all settings considered in this paper, the spurious paths will be fully characterized by the back-door paths, i.e., paths between X and Y with an arrow into X (Pearl 2000, Sec. 3.3.1). We show next that Ctf-SE uncovers the spurious relations between X and Y through confounding variables (ancestors of X and Y).

all x, then $NDE_{x_0,x_1}(y) = 0$.

[5]Confounders Z remain the same regardless of interventions on X, since Z is a non-descendant node of X, $Z_x = Z$.

Property 3. *For a SCM M, if X has no back-door path connecting Y in the corresponding causal diagram G, then $SE_{x_0,x_1}(y) = 0$, for any $y, x_0 \neq x_1$.*

This guarantees that the condition $SE_{x_0,x_1}(y) \neq 0$ can be seen as a sufficient test for the existence of back-door paths connecting X and Y, i.e., the spurious discrimination.

Explaining Discrimination

After having formally defined fine-grained counterfactual measures and studied their relations with the mechanisms capable of bringing about discrimination in the world, in this section, we consider (1) how these measures are quantitatively related and (2) how they can be estimated from data.

Decomposing the Total Variation

We first note that the counterfactual SE is closely related to the ETT – i.e., SE measures differences in outcome across units that would naturally choose x_0 and x_1 had they in fact been assigned $X = x_0$, while ETT measures the difference in outcome x_1 versus x_0 for the units which would have naturally chosen $X = x_0$. If we put this together with Eq. 1 (TV), the following decomposition can be derived:

Lemma 1. *The total variation, counterfactual spurious effect, and the effect of the treatment on the treated obey the following non-parametric relationships:*

$$TV_{x_0,x_1}(y) = SE_{x_0,x_1}(y) - ETT_{x_1,x_0}(y) \tag{6}$$
$$TV_{x_0,x_1}(y) = ETT_{x_0,x_1}(y) - SE_{x_1,x_0}(y) \tag{7}$$

In words, Eq. 6 implies that the total disparity (TV) experienced by the individuals naturally attaining x_1 relative to the ones attaining x_0 equals to the disparity experienced due to the spurious discrimination *minus* the advantage the ones attaining x_1 would have gained had they been x_0. (Recall, x_0 = advantaged and x_1 = disadvantaged groups.)

In fact, the ETT of the transition from x_0 to x_1 can be further decomposed as the *difference* between the counterfactual direct effect of that transition and the counterfactual indirect effect of the reverse transition (from x_1 to x_0), i.e.:

Lemma 2. *The effect of treatment on the treatment and the counterfactual direct and indirect effects obey the following non-parametric relationships:*

$$ETT_{x_0,x_1}(y) = DE_{x_0,x_1}(y|x_0) - IE_{x_1,x_0}(y|x_0) \tag{8}$$

Lems. 1 and 2 combined lead to a general, non-parametric decomposition of the total variation, namely:

Theorem 1 (Causal Explanation Formula)**.** *The total variation, counterfactual spurious, direct, and indirect effects obey the following relationships*

$$\begin{aligned} TV_{x_0,x_1}(y) \quad &= SE_{x_0,x_1}(y) + IE_{x_0,x_1}(y|x_1) \\ &\quad - DE_{x_1,x_0}(y|x_1) \end{aligned} \tag{9}$$
$$\begin{aligned} TV_{x_0,x_1}(y) \quad &= DE_{x_0,x_1}(y|x_0) \\ &\quad - SE_{x_1,x_0}(y) - IE_{x_1,x_0}(y|x_0) \end{aligned} \tag{10}$$

Thm. 1 provides a quantitative explanation based on the underlying causal mechanisms for the disparities observed in TV. For instance, Eq. 9 explicates that the total disparity

experienced by the individuals who have naturally attained x_1 (relative to x_0) equals to the disparity experienced associated with spurious discrimination (Property. 3), *plus* the advantage it lost due to indirect discrimination (Property. 2), *minus* the advantage it would have gained without direct discrimination (Property. 1). In the religion discrimination example, if direct discrimination exists ($DE_{x_1,x_0}(y|x_1) > 0$), Eq. 9 implies that it will lower the hiring rate for people with religious beliefs. Perhaps surprisingly, this result holds non-parametrically, which means that the counterfactual effects decompose following Thm. 1 for any functional form of the underlying (generating) structural functions and for any distribution of the unobserved exogenous variables (U). Owed to their generality and ubiquity, we refer to this family of equations as the "Causal Explanation Formula" (or simply Explanation Formula).

The definitions and properties discussed so far are based on probability distributions (e.g., $P(y)$). We extend counterfactual DE, IE, and SE using expectations ($E[Y]$), denoted by $DE_{x_0,x_1}(Y|x), IE_{x_0,x_1}(Y|x)$, and $SE_{x_0,x_1}(Y)$, respectively. As a corollary, one could verify that the Explanation Formula also holds for expectation measures.

Identifying Counterfactual Measures from Data

The Causal Explanation Formula provides the precise relation between the counterfactual quantities, but it does not specify how they should be estimated from data. In this section, we study the conditions under which these counterfactual measures can be computed in practice. We start with the standard model (Fig. 1(b)) and derive a set of identification equations for the Explanation Formula when only observational data is available.

Leveraging the assumption implied by the standard model that latent confounders are independent, we can show the following general observational explanation formula:

Theorem 2 (Causal Explanation Formula (Standard model))**.** *Under the standard model, $DE_{x_0,x_1}(y|x)$, $IE_{x_0,x_1}(y|x)$, and $SE_{x_0,x_1}(y)$ can be estimated, respectively, from the observational distribution as follows:*

$$\sum_{z,w}(P(y|x_1,w,z) - P(y|x_0,w,z))P(w|x_0,z)P(z|x),$$

$$\sum_{z,w}P(y|x_0,w,z)(P(w|x_1,z) - P(w|x_0,z))P(z|x),$$

$$\sum_{z,w}P(y|x_0,w,z)P(w|x_0,z)(P(z|x_1) - P(z|x_0)).$$

These equations provide general guidance for discrimination analysis applicable to any nonlinear system, any distribution, and any type of variables. Moreover, all the quantities in Thm. 2 are expressible in terms of conditional distributions and do not involve any counterfactual, which means that they are readily estimable by any method from the observational distribution (e.g., regression, deep nets).

The Explanation Formula in the standard model is closely related to the "Mediation Formula" used for mediation analysis, which was introduced in (Pearl 2001) and is widely popular throughout the empirical sciences (see (Pearl 2012) for a survey). In fact, if no back-door paths between X and Y ($X \not\leftarrow Z$ and $Z \not\rightarrow Y$) exist, it's not difficult to see that $SE_{x_0,x_1}(y) = 0$, and the identification of $DE_{x_0,x_1}(y|x)$ and $IE_{x_0,x_1}(y|x)$ coincides with $NDE_{x_0,x_1}(y)$ and $NIE_{x_0,x_1}(y)$ in the mediation formula.

We next state the identifiability result for the explanation formula under the standard model for when the more stringent assumption that the underlying structural functions are linear is imposed (linear-standard model). [6]

Theorem 3 (Causal Explanation Formula (Linear Models))**.** *Under the assumptions of the linear-standard model, the counterfactual DE, IE, and SE of event $X = x_1$ on Y (relative to baseline x_0) can be estimated as follows:*

$$DE_{x_0,x_1}(Y|x) = \gamma_{yx.zw}(x_1 - x_0),$$
$$IE_{x_0,x_1}(Y|x) = \gamma_{yw.xz}\gamma_{wx.z}(x_1 - x_0),$$
$$SE_{x_0,x_1}(Y) = \gamma_{xz}(\gamma_{yz.xw} + \gamma_{yw.xz}\gamma_{wz.x})(x_1 - x_0),$$

where γ are the corresponding (partial) regression coefficient (e.g., $\gamma_{yx.zw}$ is the partial regression coefficient of Y on X). The causal explanation formula decomposes as:

$$TV_{x_0,x_1}(Y) = SE_{x_0,x_1}(Y) + IE_{x_0,x_1}(Y|x) + DE_{x_0,x_1}(Y|x)$$

In contrast to the non-parametric case, the outcome disparity in linear systems can be explained by the intuitively clean, and usually expected, sum of the counterfactual spurious, indirect, and direct effects.

We consider now a relaxation of the standard model to allow for unobserved confounding, see Fig. 1(c). Following the conventions in the field, latent variables are represented graphically through the dashed-bidirected arrows. We call the set of models encompassing this set of assumptions by the *extended fairness model* (for short, extended model). We present in the sequel a sufficient condition under which the corresponding effects in the explanation formula can be identified from ETT-like counterfactual distributions.

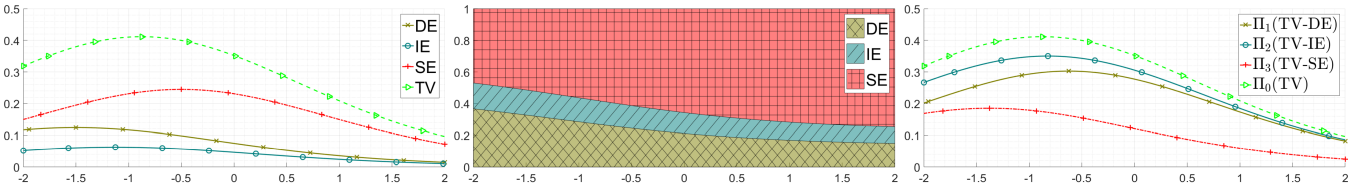
Theorem 4 (Counterfactual Identification)**.** *Under the extended model (Fig. 1(c)), if distributions $P(y_x|x')$, $P(y_{x,w}|x',w')$ and $P(w|x)$ are identifiable, then measures $SE_{x_0,x_1}(y)$, $DE_{x_0,x_1}(y|x_0)$, and $IE_{x_0,x_1}(y|x_1)$ are identifiable as well.*

The distribution $P(w|x)$ can be estimated from observational data, which is often easily available. Further, the distribution $P(y_x|x')$ and $P(y_{x,w}|x',w')$ can be estimated following the new counterfactual randomization procedure introduced in (Bareinboim, Forney, and Pearl 2015).

Applications and Simulations

We conduct experiments in different fairness tasks, including discrimination detection, explanation, and design of reparatory policies. Details of the experiments are provided in the full technical report (Zhang and Bareinboim 2018). If not stated explicitly, we assume $x_0 = 0, x_1 = 1$, and $y = 1$. We shorten the notation of direct effect and write $DE_{x_0,x_1}(y|x_0) = DE$, and similarly to IE and SE.

[6] In fact, the wide popularity of the mediation formula first came about under this specific set of assumptions (see also footnote 1).



(a) (b) (c)

Figure 3: (a, b) TV, ctf-DE, IE, SE (Y-axis (a)) and percentage composition map (Y-axis (b)) as a function of threshold $\gamma_0 \in [-2, 2]$ (X-axis) in a logistic model following structural model in Fig. 1; (c) Total variation (Y-axis) of three competing policies for threshold γ_0 (X-axis), including disabling $X \to Y$ (Π_1), $X \to W \to Y$ (Π_2), and $X \leftarrow Z \to Y$ (Π_3). (Zoom-in for view).

Task 1. Detecting Discrimination

We first consider a discrimination detection problem similar to (Datta, Sen, and Zick 2016). The goal of this task is to detect certain types of discrimination when only black-box access to an automated learning system is available. Clearly, the classic measures are unable to identify the three types of discrimination individually (Table. 1). Properties 1-3 support that counterfactual DE, IE, and SE constitute sufficient tests for the detection of direct, indirect, and spurious discrimination. If we compare these measures with the classic measures (based on Table 1), the counterfactual family is the only set of measures that could distinctively identify any type of discrimination. We test these results in practice.

Standard Fairness Model. We apply both classic and counterfactual SE, DE, IE measures to the religious discrimination example (Fig. 1(a)). Each measure in the explanation formula (Thm. 2) is evaluated using $5,000$ observational samples. The results (Fig. 2(a)) reveal, as discussed, that no effect is transmitted along causal paths ($DE = 0, IE = 0$), and the observed disparity $TV = -1$ is mainly attributed to the spurious effect transmitted along the corresponding backdoor path ($SE = -1$). On the other hand, all the classic measures are statistically zero ($TE = -0.014, ETT = 0, NIE = 0.02, CDE{-}0, QII = -0.33, NDE = 0.015$), failing to detect the discriminatory mechanism behind the observed disparity.

Extended Fairness Model. Consider the extended model represented in Fig. 1(c) where we apply both the classic (TV, TE, ETT, CDE, NDE, NIE, QII) and the new counterfactual measures. Counterfactual DE, IE, and SE are estimated from $5,000$ samples collected with the counterfactual randomization procedure (Thm. 4). The results (Fig. 2(b)) show that though there is no observed disparity ($TV = 0.0088$), direct ($DE = -0.8106, CDE = 0.6568, QII = 0.6650, NDE = -0.7978$) and indirect discrimination ($IE = -0.7965, IE = -0.8028$) through causal paths are still significant. Since no spurious effect is detected ($SE = 0.009$), the total effect ($TE = -0.0062$) coincides with TV and ETT (-0.0068). After all, the main conclusion of this experiment is that the prevailing practice of using the TV as the basis of the analysis is, at best, misleading. The fact that $TV = 0$ is no indication that discrimination due to direct, indirect, or spurious effects is absent.

Task 2. Explaining Discrimination

The counterfactual measures can explain how much of the observed disparity is due to their corresponding (unobserved) causal mechanism. To witness, we consider a modified logistic model similar to the one studied in (MacKinnon et al. 2007) (see Fig. 1(a)). The outcome Y is a threshold-based indicator of a linear function, such that $y = I\{\gamma_0 + \gamma_{xy}x + \gamma_{zy}z + \gamma_{wy}w + u_y\}$, where $I\{\cdot\}$ is an indicator function, U_y follows the logistic distribution, and γ_0 is the (unknown) decision threshold.

We compute analytically and plot in Fig. 3(a) $TV_{x_0,x_1}(Y)$, $-DE_{x_1,x_0}(Y|x_1)$, $IE_{x_0,x_1}(Y|x_1)$, $SE_{x_0,x_1}(Y)$ as a function of the threshold $\gamma_0 \in [-2, 2]$ (derivations are shown in the Appendix). The result agrees with the predictions of the Explanation Formula (Thm. 1) – e.g., for $\gamma_0 = 0$, $SE_{x_0,x_1}(Y) + IE_{x_0,x_1}(Y|x_1) - DE_{x_1,x_0}(Y|x_1) = 0.0462 + 0.2311 + 0.0747 = TV_{x_0,x_1}(Y) = 0.3520$. Interestingly, Fig. 3(b) shows the composition of counterfactual effects in respect to TV. The results reveal that, in this instance, the observed disparity can be mainly attributed to spurious discrimination (SE), for any $\gamma_0 \in [-2, 2]$.

Implications for Decision Analysis. When considering a more scientific perspective, the relevance of the counterfactual measures and explanation formula (Thms. 1-2) are obvious, researchers want to answer the question "how nature currently works". From a decision-making perspective, this machinery further enables us to predict how the environment will change under various policies and interventional conditions. Consider a decision-making problem relative to the standard model, where one needs to decide between three competing policies: Π_1 such that direct discrimination is eliminated (disable $X \to Y$), Π_2 where indirect discrimination is eliminated (disable $X \to W \to Y$), and Π_3 where spurious discrimination is eliminated (disable $X \leftarrow Z \to Y$). The goal is, for instance, to find the optimal policy that minimizes the total variation $TV_{x_0,x_1}(Y)$ (disparate outcome). To evaluate the new policy, we remove the measure in which the corresponding mechanism is disabled from the Explanation Formula (Thm. 1). For instance, if spurious discrimination is eliminated, the resulting observed disparity equals to $IE_{x_0,x_1}(Y|x_1) - DE_{x_1,x_0}(Y|x_1)$. The optimal policy can then be obtained by selecting one with the least resulting observed disparity.

We apply this procedure using the logistic threshold

model and evaluate policies for $\gamma_0 \in [-2, 2]$. The results in Fig. 3(c) reveal that eliminating spurious discrimination (Π_3) is the optimal policy for all instances (e.g., for $\gamma_0 = 0$, the resulting observed disparities are $\Pi_1 = 0.2773, \Pi_2 = 0.3057, \Pi_3 = 0.1209$). The conclusion supports the results in Fig. 3(b), which suggests that, in this setting, SE is the highest contributor to the observed disparity.

Task 3. Designing Reparatory Policies

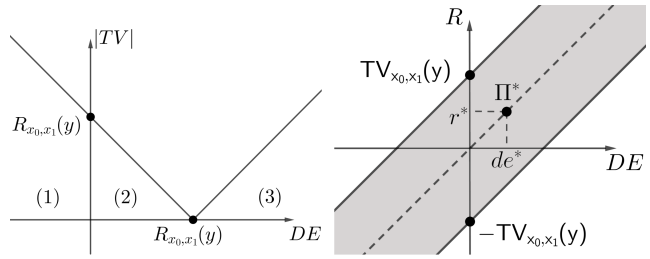
If the outcome disparity still exists after all the unjustified paths are disabled, companies and universities may be required to further control for unfairness. For instance, affirmative actions are a range of policies used to compensate previous discrimination by providing opportunities for members of the protected group. Designing these actions, by its very nature, entails a trade-off between the procedural and outcome fairness (Barocas and Selbst 2016).

In practice, an affirmative action must be "narrowly tailored" so that it minimizes the outcome disparity while not being overinclusive to introduce *reverse discrimination* (Adarand Constructors, Inc. vs. Pena 1995). Unfortunately, the definition and conditions that could allow this principle to be applied in practice are quite opaque (Ayres 1995). We provide next a quantitative interpretation of this principle based on the counterfactual measures, which in turn leads to the first operational definition of "narrow tailoring."

DE-Reparatory Policies Let the residual disparity $R_{x_0,x_1}(y) = SE_{x_0,x_1}(y) + IE_{x_0,x_1}(y|x_1)$ denote all remaining counterfactual spurious and indirect effects when all discriminatory paths are disabled. We start by considering a setting where $R_{x_0,x_1}(y)$ is fixed, and one is allowed to manipulate only $DE_{x_1,x_0}(y)$ so as to minimize the total disparity (e.g., college's admission). Suppose $R_{x_0,x_1}(y) > 0$ and that the total outcome disparity is measured by the absolute value of the total variation $|TV_{x_0,x_1}(y)|$. By Thm. 1, $|TV_{x_0,x_1}(y)| = |R_{x_0,x_1}(y) - DE_{x_1,x_0}(y|x_1)|$.

We plot in Fig. 4(a) the values of $|TV_{x_0,x_1}(y)|$ along $DE_{x_1,x_0}(y|x_1)$. The graph is divided into three regions. In the first region ($DE_{x_1,x_0}(y|x_1) < 0$), the total disparity $|TV_{x_0,x_1}(y)|$ increases as $DE_{x_1,x_0}(y|x_1) \to -\infty$. In the religious example discussed earlier, this would mean that the total disparity increases due to the direct discrimination against believers. When $DE_{x_1,x_0}(y|x_1) > R_{x_0,x_1}(y)$ (region 3), $|TV_{x_0,x_1}(y)|$ increases as $DE_{x_1,x_0}(y|x_1) \to \infty$, which represents the scenario in which the reverse discrimination occurs due to the overinclusive nature of the affirmative actions. In this case, non-believers are in disadvantage in both the treatment procedure and the outcome, since traditionally disadvantaged believers could now simply force their way in without the necessary educational background. Only in region 2 ($0 \leq DE_{x_1,x_0}(y|x_1) \leq R_{x_0,x_1}(y)$), the total disparity $|TV_{x_0,x_1}(y)|$ decreases as $DE_{x_1,x_0}(y|x_1)$ increases. In words, the affirmative action narrows the observed disparity in a controlled and rational manner. We can use this idea to provide a quantitative definition of the "narrow tailoring" principle, namely:

Definition 7. Under the assumptions that only $DE_{x_1,x_0}(y|x_1)$ is subject to change, a reparatory policy is said to be "narrowly tailored" if and only if $DE_{x_1,x_0}(y|x_1) \in [0, R_{x_0,x_1}(y)]$.



(a) (b)

Figure 4: (a) Quantitative relationship between the absolute total variation $|TV_{x_0,x_1}(y)|$ (Y-axis) and the counterfactual direct effect $DE_{x_1,x_0}(y|x_1)$ (X-axis) with the residual disparity $R_{x_0,x_1}(y)$ fixed; (b) Feasible region of $DE_{x_1,x_0}(y|x_1)$ (X-axis) and $R_{x_0,x_1}(y)$ (Y-axis) satisfying the "narrow tailoring" principle $|R_{x_0,x_1}(y) - DE_{x_1,x_0}(y|x_1)| \leq |TV_{x_0,x_1}(y)|$.

General Reparatory Policies We now consider a more involved scenario where both $DE_{x_1,x_0}(y|x_1)$ and $R_{x_0,x_1}(y)$ (IE and SE) can be changed, while not widening the total outcome disparity. In this case, the "narrow tailoring" principle can be interpreted as the constraint $|R_{x_0,x_1}(y) - DE_{x_1,x_0}(y|x_1)| \leq |TV_{x_0,x_1}(y)|$. We plot in Fig. 4(b) the feasible region of $DE_{x_1,x_0}(y|x_1)$ and $R_{x_0,x_1}(y)$ under such constraint (shadow). For any policy $\Pi^* = (de^*, r^*)$ inside the feasible region, its corresponding affirmative action must satisfy the "narrow tailoring" principle, i.e., $|r^* - de^*| \leq |TV_{x_0,x_1}(y)|$, where $|TV_{x_0,x_1}(y)|$ is the total outcome disparity before the introduction of affirmative actions. This characterization is useful for fair learning problems where the goal is to optimize a function of the counterfactual measures under constraints over outcome disparities (Calders, Kamiran, and Pechenizkiy 2009; Zemel et al. 2013).

Conclusion

We introduced a new family of counterfactual measures capable of distinctly capturing the most prominent causal mechanisms that can bring about discrimination in the real-world (i.e., direct, indirect, spurious). We derived the Causal Explanation Formula (for short, Explanation Formula), which allows one to understand how an observed disparity between the protected attribute and the outcome variable can be decomposed in terms of the causal mechanisms underlying the specific (and unknown) decision-making process. We provided identifiability conditions for the Explanation Formula, which delineate when (and how) the corresponding counterfactual distributions can be estimated from real data. Finally, we applied the Explanation Formula to the problem of policy evaluation, providing the first quantitative explanation of the trade-off between outcome and procedural fairness. In practice, these results allow one to design fairness-aware policies, including affirmative actions compatible with the principle of "narrow tailoring."

References

Adarand Constructors, Inc. vs. Pena. 1995. U.S. Supreme Court.

Altonji, J. G., and Blank, R. M. 1999. Race and gender in the labor market. *Handbook of labor economics* 3:3143–3259.

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

Ayres, I. 1995. Narrow tailoring. *UCLA L. Rev.* 43:1781.

Bareinboim, E., and Pearl, J. 2016. Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci.* 113:7345–7352.

Bareinboim, E.; Forney, A.; and Pearl, J. 2015. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, 1342–1350.

Barocas, S., and Selbst, A. D. 2016. Big data's disparate impact. *California Law Review* 104.

Baron, R. M., and Kenny, D. A. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology* 51(6):1173.

Bollen, K. 1989. *Structural Equations with Latent Variables*. New York: John Wiley.

Bonchi, F.; Hajian, S.; Mishra, B.; and Ramazzotti, D. 2017. Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics* 3(1):1–21.

Brennan, T.; Dieterich, W.; and Ehret, B. 2009. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior* 36(1):21–40.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09.*, 13–18. IEEE.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*.

Council, N. R. 2004. *Measuring Racial Discrimination*. Washington, DC: The National Academies Press.

Datta, A.; Sen, S.; and Zick, Y. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (SP), 2016 IEEE Symp.*, 598–617.

Derfner, A. 1973. Racial discrimination and the right to vote. *Vand. L. Rev.* 26:523.

Duncan, O. 1975. *Introduction to Structural Equation Models*. New York: Academic Press.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proc. of the 3rd Innovations in Theoretical Computer Science Conf.*, 214–226. ACM.

Fox, J. 1980. Effect analysis in structural equation models. *Sociological Methods and Research* 9(1):3–28. cites drawer.

Halpern, J. Y. 2000. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research* 12(1):317–337.

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

Imai, K.; Keele, L.; Tingley, D.; and Yamamoto, T. 2011. Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review* 105(4):765–789.

Imai, K.; Keele, L.; and Yamamoto, T. 2010. Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci.* 25(1):51–71.

Khandani, A. E.; Kim, A. J.; and Lo, A. W. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance* 34(11):2767–2787.

Kilbertus, N.; Rojas-Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. *CoRR* abs/1706.02744.

Kusner, M. J.; Loftus, J. R.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*.

MacKinnon, D. P.; Lockwood, C. M.; Brown, C. H.; Wang, W.; and Hoffman, J. M. 2007. The intermediate endpoint effect in logistic and probit regression. *Clinical Trials* 4(5):499–513.

MacKinnon, D. 2008. *An Introduction to Statistical Mediation Analysis*. New York: Lawrence Erlbaum Associates.

Mahoney, J. F., and Mohen, J. M. 2007. Method and system for loan origination and underwriting. US Patent 7,287,008.

Mancuhan, K., and Clifton, C. 2014. Combating discrimination using bayesian networks. *Artificial Intell. and Law* 22(2):211–238.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press. 2nd edition, 2009.

Pearl, J. 2001. Direct and indirect effects. In *Proc. of the 17th Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann. 411–420.

Pearl, J. 2012. The mediation formula: A guide to the assessment of causal pathways in nonlinear models. In Berzuini; Dawid; and Bernardinelli., eds., *Causality: Statistical Perspectives and Applications*, 151–179. UK: Wiley.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On fairness and calibration. *arXiv preprint arXiv:1709.02012*.

Romei, A., and Ruggieri, S. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Eng. Rev.* 29(5):582–638.

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

VanderWeele, T. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York: Oxford University Press.

Wright, S. 1923. The theory of path coefficients: A reply to Niles' criticism. *Genetics* 8:239–255.

Wright, S. 1934. The method of path coefficients. *The annals of mathematical statistics* 5(3):161–215.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2015. Learning fair classifiers. *arXiv preprint arXiv:1507.05259*.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017a. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, 1171–1180. International World Wide Web Conferences.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017b. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 325–333.

Zhang, J., and Bareinboim, E. 2018. Fairness in decision-making – the causal explanation formula. Technical Report R-30, AI Lab, Purdue University.

Zhang, L.; Wu, Y.; and Wu, X. 2016. A causal framework for discovering and removing direct and indirect discrimination. *arXiv preprint arXiv:1611.07509*.

"Fairness in Decision-Making – The Causal Explanation Formula" Supplemental Material

Appendix I. Proofs

Proofs build on three inference rules called do-calculus (Pearl, 2000, Sec. 3.4), the exclusion and independence restrictions rules of SCMs (Pearl, 2000, pp. 232), and three axioms of structural counterfactuals: composition, effectiveness, and reversibility (Pearl, 2000, Sec.7.3.1).

Proof of Property. 1. By definition,

$$\begin{aligned}&P(y_{x_1,W_{x_0}}|x) - P(y_{x_0}|x)\\&= \sum_{w,z}(P(y_{x_1,w}|x,w_{x_0},z) - P(y_{x_0}|x,w_{x_0},z))P(w_{x_0},z|x)\end{aligned}$$

By the exclusion restrictions rule, $Z = Z_x$. Thus,

$$\begin{aligned}&\sum_{w,z}(P(y_{x_1,w}|x,w_{x_0},z) - P(y_{x_0}|x,w_{x_0},z))P(w_{x_0},z|x)\\&= \sum_{w,z}(P(y_{x_1,w}|x,w_{x_0},z_{x_1,w}) - P(y_{x_0,w}|x,w_{x_0},z_{x_0,w}))\\&\cdot P(w_{x_0},z|x)\\&= \sum_{w,z}(P(y_{x_1,w,z}|x,w_{x_0},z_{x_0,w}) - P(y_{x_0,w,z}|x,w_{x_0},z_{x_0,w}))\\&\cdot P(w_{x_0},z|x)\end{aligned}$$

The last step holds by the composition axiom: for any x, w, z,

$$Z_{x,w} = z \Rightarrow Y_{x,w} = Y_{x,w,z}$$

Since X has no direct link connecting Y, $Y_{x,w,z} = Y_{w,z}$ for any x, w, z (the exclusion restrictions rule), which gives:

$$\begin{aligned}&\sum_{w,z}(P(y_{x_1,w,z}|x,w_{x_0},z_{x_0,w}) - P(y_{x_0,w,z}|x,w_{x_0},z_{x_0,w}))\\&\cdot P(w_{x_0},z|x)\\&= \sum_{w,z}(P(y_{w,z}|x,w_{x_0},z_{x_0,w}) - P(y_{w,z}|x,w_{x_0},z_{x_0,w}))\\&\cdot P(w_{x_0},z|x) = 0\end{aligned}$$ □

Proof of Property. 2. If X has no indirect causal pathway connecting Y in the causal diagram G, this means that there is no intermediate variable between X and Y, i.e., $W = \emptyset$. We thus have $P(y_{x_0,W_{x_1}}|x) = P(Y_{x_0}|x)$, which implies

$$IE_{x_0,x_1}(y) = P(y_{x_0,W_{x_1}}|x) - P(y_{x_0}|x) = 0$$ □

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Proof of Property. 3. The back-door criterion (Pearl, 2000, Sec. 11.3.2) implies $Y_x \perp\!\!\!\perp X$, i.e.,

$$P(y_{x_0}|x_1) = P(y_{x_0}|x_0) = P(y_{x_0})$$

By the composition axiom, $P(y|x_0) = P(y_{x_0}|x_0)$. Ctf-SE is thus equal to:

$$\begin{aligned}&SE_{x_0,x_1}(y) = P(y_{x_0}|x_1) - P(y|x_0)\\&= P(y_{x_0}|x_1) - P(y_{x_0}|x_0)\\&= P(y_{x_0}) - P(y_{x_0}) = 0\end{aligned}$$ □

Proof of Lemma. 1. By definition, $TV_{x_0,x_1}(y)$ can be written as:

$$\begin{aligned}&TV_{x_0,x_1}(y) = P(y|x_1) - P(y|x_0)\\&= P(y|x_1) - P(y_{x_0}|x_1) + P(y_{x_0}|x_1) - P(y|x_0)\\&= -ETT_{x_1,x_0}(y) + SE_{x_0,x_1}(y)\end{aligned}$$

Similarly, we can switch x_0 and x_1 in the decomposing term $P(y_{x_0}|x_1)$ and write $TV_{x_0,x_1}(y)$ as:

$$\begin{aligned}&TV_{x_0,x_1}(y) = P(y|x_1) - P(y|x_0)\\&= P(y|x_1) - P(y_{x_1}|x_0) + P(y_{x_1}|x_0) - P(y|x_0)\\&= -SE_{x_1,x_0}(y) + ETT_{x_0,x_1}(y)\end{aligned}$$ □

Proof of Lemma. 2. By definition, $ETT_{x_0,x_1}(y)$ is equal to

$$\begin{aligned}ETT_{x_0,x_1}(y) &= P(y_{x_1}|x_0) - P(y|x_0)\\&= P(y_{x_1}|x_0) - P(y_{x_1,W_{x_0}}|x_0)\\&+ P(y_{x_1,W_{x_0}}|x_0) - P(y_{x_0}|x_0)\\&= -IE_{x_1,x_0}(Y|x_0) + DE_{x_0,x_1}(Y|x_0)\end{aligned}$$ □

Proof of Theorem. 1. By Eq. 6 in Lem. 1,

$$TV_{x_0,x_1}(y) = SE_{x_0,x_1}(y) - ETT_{x_1,x_0}(y)$$

Replace $ETT_{x_1,x_0}(y)$ with Eq. 8 (Lem. 2),

$$TV_{x_0,x_1}(y) = SE_{x_0,x_1}(y) + IE_{x_0,x_1}(y|x_1) - DE_{x_1,x_0}(y|x_1)$$

Similarly, Eq. 7 (Lem. 1) and Eq. 8 combined give

$$\begin{aligned}&TV_{x_0,x_1}(y) = ETT_{x_0,x_1}(y) - SE_{x_1,x_0}(y)\\&= DE_{x_0,x_1}(y|x_0) - SE_{x_1,x_0}(y) - IE_{x_1,x_0}(y|x_0)\end{aligned}$$ □

Proof of Theorem. 2. We first consider the following equations, which will be useful later on in the proof

$$P(z, w|x) = P(w|x, z)P(z|x) \tag{11}$$
$$P(y_{x,z,w}) = P(y|x, z, w) \tag{12}$$
$$P(z, w_{x'}|x) = P(w|x', z)P(z|x) \tag{13}$$

Eq. 11 is licensed by Bayes rule. Eq. 12 holds by do-calculus rule 2 (Pearl, 2000, Sec. 3.4). As for Eq. 13, $P(z, w_{x'}|x)$ can be written as

$$\begin{aligned} P(z, w_{x'}|x) &= P(w_{x'}|z, x)P(z|x) \\ &= P(w_{x',z}|z, x)P(z|x) \end{aligned}$$

The last step holds since: (1) by the exclusion restrictions rule, $Z = Z_{x'}$; and (2) by the composition axiom, $W_{x'} = W_{x',z}$ if $Z_{x'} = z$. Similarly, by the composition axiom, $Z = z \Rightarrow X = X_z$, which gives:

$$P(w_{x',z}|z, x)P(z|x) = P(w_{x',z}|z, x_z)P(z|x)$$

Since W is not connected with X, Z by bi-directed edges (de-confounded), by the independence restrictions rule, $W_{x,z} \perp\!\!\!\perp X_z, Z$ for any x, z. We thus have:

$$\begin{aligned} P(w_{x',z}|z, x_z)P(z|x) &= P(w_{x',z})P(z|x) \\ &= P(w|x', z)P(z|x) \end{aligned}$$

The last step holds by do-calculus rule 2. We are now ready to derived identification formulas for counterfactual DE, IE and SE in the standard model. We can write $SE_{x_0,x_1}(y)$ as following by conditioning on Z, W:

$$\begin{aligned} SE_{x_0,x_1}(y) &= P(Y_{x_0}|x_1) - P(y|x_0) \\ &= \sum_{z,w} P(y_{x_0}|x_1, z, w_{x_0})P(z, w_{x_0}|x_1) \\ &- \sum_{z,w} P(y|x_0, z, w)P(z, w|x_0) \end{aligned} \tag{14}$$

As for $P(y_{x_0}|x_1, z, w_{x_0})$, by the exclusion restrictions rule, $Z = Z_{x_0} = Z_{x_1}$. By the composition axiom,

$$Z = z \Rightarrow X = X_z, \tag{15}$$
$$Z_{x_0} = z \Rightarrow W_{x_0} = W_{x_0,z}, \tag{16}$$
$$Z_{x_0} = z, W_{x_0} = w \Rightarrow Y_{x_0} = Y_{x_0,z,w}. \tag{17}$$

We thus have:

$$P(y_{x_0}|x_1, z, w_{x_0}) = P(y_{x_0,z,w}|x_{1z}, z, w_{x_0,z}) \tag{18}$$

Since Y and X, Z, W are not connected by bi-directed edges, the independence restrictions rule gives:

$$Y_{x,z,w} \perp\!\!\!\perp X_z, Z, W_{z,w}, \tag{19}$$

for any x, z, w. This implies

$$P(y_{x_0,z,w}|x_{1z}, z, w_{x_0,z}) = P(y_{x_0,z,w}). \tag{20}$$

Replacing $P(y_{x_0}|x_1, z, w_{x_0})$ in Eq. 14 with Eq. 20 gives:

$$\begin{aligned} &\sum_{z,w} P(y_{x_0}|x_1, z, w_{x_0})P(z, w_{x_0}|x_1) \\ &- \sum_{z,w} P(y|x_0, z, w)P(z, w|x_0) \\ &= \sum_{z,w} P(y_{x_0,z,w})P(z, w_{x_0}|x_1) \\ &- \sum_{z,w} P(y|x_0, z, w)P(z, w|x_0) \end{aligned}$$

The above equation, together with Eqs. 11, 12 and 13, proves the identification formula of $SE_{x_0,x_1}(y)$:

$$SE_{x_0,x_1}(y) = \sum_{z,w} P(y|x_0, z, w)P(w|x_0, z)(P(z|x_1) - P(z|x_0)).$$

By conditioning on Z, W, we write $DE_{x_0,x_1}(y|x)$ as

$$\begin{aligned} DE_{x_0,x_1}(y|x) &= P(y_{x_1,W_{x_0}}|x) - P(y_{x_0}|x) \\ &= \sum_{z,w} P(y_{x_1,w}|x, z, w_{x_0})P(z, w_{x_0}|x_1) \\ &- \sum_{z,w} P(y_{x_0}|x, z, w_{x_0})P(z, w_{x_0}|x) \\ &= \sum_{z,w} P(y_{x_1,w}|x, z, w_{x_0})P(z, w_{x_0}|x_1) \\ &- \sum_{z,w} P(y_{x_0,w,z})P(z, w_{x_0}|x) \end{aligned} \tag{21}$$

The last step holds by Eq. 20. As for $P(y_{x_1,w}|x, z, w_{x_0})$, by the exclusion restrictions rule, $Z = Z_{x_0} = Z_{x_1,w}$. By the composition axiom, $Y_{x_1,w} = Y_{x_1,w,z}$ if $Z_{x_1,w} = z$. Together with Eqs. 15 and 16, we obtain:

$$P(y_{x_1,w}|x, z, w_{x_0}) = P(y_{x_1,w,z}|x_z, z, w_{x_0,z}). \tag{22}$$

The independence relation in Eq. 19 implies

$$P(y_{x_1,w,z}|x_z, z, w_{x_0,z}) = P(y_{x_1,w,z}). \tag{23}$$

Replacing $P(y_{x_1,w}|x, z, w_{x_0})$ in Eq. 21 with Eq. 23 gives

$$\begin{aligned} &\sum_{z,w} P(y_{x_1,w}|x, z, w_{x_0})P(z, w_{x_0}|x_1) \\ &- \sum_{z,w} P(y_{x_0,w,z})P(z, w_{x_0}|x) \\ &= \sum_{z,w} (P(y_{x_1,w,z}) - P(y_{x_0,w,z}))P(z, w_{x_0}|x) \end{aligned}$$

Together the above equation with Eqs. 11-13, we prove the identification formula of $DE_{x_0,x_1}(y|x)$:

$$\begin{aligned} DE_{x_0,x_1}(y|x) = \sum_{z,w} &(P(y|x_1, w, z) \\ &- P(y|x_0, w, z))P(w|x_0, z)P(z|x). \end{aligned}$$

Finally, $IE_{x_0,x_1}(y|x)$ equals to

$$\begin{aligned} IE_{x_0,x_1}(y|x) &= P(y_{x_0,W_{x_1}}|x) - P(y_{x_0}|x) \\ &= \sum_{z,w} P(Y_{x_0,w}|x, z, w_{x_1})P(z, w_{x_1}|x) \\ &- \sum_{z,w} P(Y_{x_0}|x, z, w_{x_0})P(z, w_{x_0}|x) \\ &= \sum_{z,w} P(Y_{x_0,z,w})(P(z, w_{x_1}|x) - P(z, w_{x_0}|x)) \end{aligned}$$

Note that in Eqs. 18, 20, 22 and 23, x, x_0, x_1 can be arbitrary values, i.e.:

$$P(y_{x_0,w}|x, z, w_{x_1}) = P(y_{x_0}|x, z, w_{x_0}) = P(y_{x_0,z,w})$$

This gives

$$IE_{x_0,x_1}(y|x) = \sum_{z,w} P(Y_{x_0,z,w})(P(z,w_{x_1}|x) - P(z,w_{x_0}|x))$$

Together the above equation with Eqs. 11-13, we prove the identification formula of $IE_{x_0,x_1}(y|x)$:

$$IE_{x_0,x_1}(y|x) = \sum_{z,w} P(y|x_0,w,z)(P(w|x_1,z) - P(w|x_0,z))P(z|x). \quad \square$$

Proof of Theorem. 3. Let us examine what Causal Explanation Formula yields when applied to the linear-standard model where

$$z = u_{xz}, \quad x = \alpha_{xz}z + u_{xz}, \quad w = \alpha_{wx}x + \alpha_{wz}z + u_w,$$
$$y = \alpha_{yx}x + \alpha_{yz}z + \alpha_{yw}w + u_y.$$

Without loss of generality, we assume u_{xz}, u_y, u_w are normal with zero mean and variance one. We consider the expectation version of Explanation Formula which replaces $P(Y|x,w,z)$ with $E[Y|x,w,z]$. Computing the conditional expectations $E[Y|x,w,z]$ and $E[W|x,z]$ gives:

$$\begin{aligned} E[Y|x,w,z] &= E[\alpha_{yx}x + \alpha_{yz}z + \alpha_{yw}w + u_y] \\ &= \alpha_{yx}x + \alpha_{yz}z + \alpha_{yw}w \\ E[W|x,z] &= E[\alpha_{wx}x + \alpha_{wz}z + u_w] \\ &= \alpha_{wx}x + \alpha_{wz}z \end{aligned}$$

Similarly, we can compute conditional expectations $E[Y|x,z]$ as following:

$$\begin{aligned} E[Y|x,z] &= E[\alpha_{yx}x + \alpha_{yz}z + \alpha_{yw}W + u_y|x,z] \\ &= \alpha_{yx}x + \alpha_{yz}z + \alpha_{yw}E[W|x,z] \\ &= \alpha_{yx}x + \alpha_{yz}z + \alpha_{yw}\alpha_{wx}x + \alpha_{yw}\alpha_{wz}z \end{aligned}$$

and these give us

$$\begin{aligned} &DE_{x_0,x_1}(Y|x) = \sum_{z,w}(\alpha_{yx}x_1 - \alpha_{yx}x_0)P(w|x_0,z)P(z|x) \\ &= \alpha_{yx}(x_1 - x_0) \\ &IE_{x_0,x_1}(Y|x) = \sum_{z,w} E[Y|x_0,w,z](P(w|x_1,z) - P(w|x_0,z))P(z|x) \\ &= \sum_{z,w}(y_0 + \alpha_{yx}x + \alpha_{yz}z + \alpha_{yw}w) \\ &\cdot (P(w|x_1,z) - P(w|x_0,z))P(z|x) \\ &= \sum_{z} \alpha_{yw}(E[W|x_1,z] - E[W|x_0,z])P(z|x) \\ &= \alpha_{yw}\alpha_{wx}(x_1 - x_0) \\ &SE_{x_0,x_1}(Y) \\ &= \sum_{z,w} E[Y|x_0,w,z]P(w|x_0,z)(P(z|x_1) - P(z|x_0)) \\ &= \sum_{z} E[Y|x_0,z](P(z|x_1) - P(z|x_0)) \\ &= \sum_{z}(y_0 + w_0 + \alpha_{yx}x + +\alpha_{yw}\alpha_{wx}x + (\alpha_{yz} + \alpha_{yw}\alpha_{wz})z) \\ &\cdot (P(z|x_1) - P(z|x_0)) \\ &= (\alpha_{yz} + \alpha_{yw}\alpha_{wz})(E[Z|x_1] - E[Z|x_0]) \\ &= \gamma_{zx}(\alpha_{yz} + \alpha_{yw}\alpha_{wz})(x_1 - x_0) \end{aligned}$$

where the regression coefficient $\gamma_{zx} = \frac{\partial}{\partial x}E[Z|x]$.

$$\begin{aligned} &TV_{x_0,x_1}(Y) = E[Y|x_1] - E[Y|x_0] \\ &= \sum_{z,w}(E[Y|x_1,w,z]P(w|x_1,z)P(z|x_1) - E[Y|x_0,w,z] \\ &\cdot P(w|x_0,z)P(z|x_0)) \\ &= (\alpha_{yx} + \alpha_{yw}\alpha_{wx})(x_1 - x_0) + (\alpha_{yz} + \alpha_{yw}\alpha_{wz}) \\ &\cdot (E[Z|x_1] - E[Z|x_0]) \\ &= (\alpha_{yx} + \alpha_{yw}\alpha_{wx} + \gamma_{zx}(\alpha_{yz} + \alpha_{yw}\alpha_{wz}))(x_1 - x_0) \end{aligned}$$

Parameters $\alpha_{yx}, \alpha_{yz}, \alpha_{yw}, \alpha_{wx}, \alpha_{wz}$ can be identified with partial regression coefficients as following:

$$\alpha_{yx} = \gamma_{yx.zw}, \quad \alpha_{yz} = \gamma_{yz.xw}, \quad \alpha_{yw} = \gamma_{yw.xz},$$
$$\alpha_{wx} = \gamma_{wx.z}, \quad \alpha_{wz} = \gamma_{wz.x}.$$

We thus obtain identification formulas for counterfactual DE, IE and SE in the linear-standard model. In particular, Causal Explanation Formula produces the standard, additive relation in the linear-standard model, i.e.,

$$TV_{x_0,x_1}(Y) = SE_{x_0,x_1}(Y) + IE_{x_0,x_1}(Y|x) + DE_{x_0,x_1}(Y|x) \quad \square$$

Proof of Theorem. 4. Note that $SE_{x_0,x_1}(y)$, $DE_{x_0,x_1}(y|x_0)$ and $IE_{x_0,x_1}(y|x_1)$ consist of two types of quantities: $P(y_x|x')$ and $P(y_{x,W_{x'}}|x')$. The former is identifiable, and we now consider the latter:

$$\begin{aligned} P(y_{x,W_{x'}}|x') &= \sum_{w} P(y_{x,w}|x',w_{x'})P(w_{x'}|x') \\ &= \sum_{w} P(y_{x,w}|x',w)P(w|x') \end{aligned}$$

where $P(y_{x,w}|x',w)$ and $P(w|x')$ are both identifiable. Therefore, the nested counterfactual $P(y_{x,W_{x'}}|x')$ is also identifiable. $\square$

Appendix II. Analysis of Current Methods

In this section, we will provide the detailed analysis of results presented in Table. 1. Specifically, we study capabilities and limitations of state-of-art discrimination measures. For each measure, we formally analyze what types of discrimination it can (cannot) distinctly identify. If not stated, we assume $x_0 = 0, x_1 = 1$.

Total Variation (TV). Recall that the total variation of event $X = x_1$ on $Y = y$ (with baseline x_0) is defined as $TV_{x_0,x_1}(y) = P(y|x_1) - P(y|x_0)$. TV measures all paths (causal and non-causal) connecting from the protected attribute X to the outcome Y, formally,

Lemma 3. *For a SCM M, if X has no path connecting Y in the causal diagram G, then $TV_{x_0,x_1}(y) = 0$, for any $y, x_0 \neq x_1$.*

Lem. 3 is implied by the soundness of *d-separation* (Koller and Friedman, 2009, Sec. 3.3.2). This lemma says that the condition $TV_{x_0,x_1}(y) \neq 0$ is a sufficient test for the existence of paths connecting X and Y. However, this path could be either direct, indirect or spurious. In other words, $TV_{x_0,x_1}(y) \neq 0$ could only detect, but not distinctly identify underlying discriminatory mechanisms.

Total Effect (TE). The total effect measures the causal effect of intervention $X = x_1$ on $Y = y$ (with baseline x_0), namely, $TE_{x_0,x_1}(y) = P(y_{x_1}) - P(y_{x_0})$. TE measures effect transmitted along causal paths connecting from X to Y. Formally,

Lemma 4. *For a SCM M, if X has no causal path connecting Y in the causal diagram G, then $TE_{x_0,x_1}(y) = 0$, for any $y, x_0 \neq x_1$.*

Proof. If X has no causal path connecting Y, then X is a non-descendant of Y. This implies that $Y_x = Y$ (Halpern, 2000). Thus,

$$TE_{x_0,x_1}(y) = P(y_{x_1}) - P(y_{x_0}) = P(y) - P(y) = 0 \,\square$$

Lem. 4 says that one can test the existence of causal paths between X and Y by checking whether $TE_{x_0,x_1}(y) \neq 0$. However, this condition fails to capture the existence of back-door paths, i.e., the spurious discrimination.

Lemma 5. *There exists a SCM M where X and Y are not connected by any causal path, but $TE_{x_0,x_1}(y) \neq 0$ for some $y, x_0 \neq x_1$.*

Proof. We can prove this lemma by constructing a such SCM M, where X, Y, U are binary variables in $\{0, 1\}$, $P(U = 0) = 0.9$. Values of y are decided by function $y = x \oplus u$ ($\oplus$ stands for the *"xor"* operator), and X follows a uniform distribution. In this model, X and Y are only connected by the direct causal path $X \to Y$. However, for $y = 1$, $TE_{x_0,x_1}(y)$ is equal to

$$\begin{aligned} TE_{x_0,x_1}(y) &= P(y_{x_1}) - P(y_{x_0}) \\ &= P(U = 0) - P(U = 1) = 0.8 \end{aligned}$$

which is not zero. $\square$

ETT and Counterfactual Fairness. Recall that the effect of treatment on the treated (ETT) of event $X = x_1$ on $Y = y$ is defined as

$$ETT_{x_0,x_1}(y) = P(y_{x_1}|x_0) - P(y|x_0)$$

Kusner et al. (2017) defined the counterfactual fairness measure using ETT conditioned on the sub-population $Z = z, W = w$, namely

Definition 8. The counterfactual fairness measure of intervention $X = x_1$ on Y (with baseline x_0) conditioned on $Z = z, W = w$ is defined as:

$$ETT_{x_0,x_1}(y|z, w) = P(y_{x_1}|x_0, z, w) - P(y|x_0, z, w)$$

Kusner et al. (2017) showed that the counterfactual fairness measures effects transmitted along causal paths from X to Y. We here provide an alternative proof.

Lemma 6. *For a SCM M, if X has no causal path connecting Y in the causal diagram G, then $TE_{x_0,x_1}(y) = 0$, for any $y, x_0 \neq x_1$.*

Proof. If X has no causal path connecting Y, then X is a non-descendant of Y. This implies that $Y_x = Y$ (Halpern, 2000). Thus,

$$\begin{aligned} &ETT_{x_0,x_1}(y|z, w) = P(y_{x_1}|x_0, z, w) - P(y|x_0, z, w) \\ &= P(y|x_0, z, w) - P(y|x_0, z, w) = 0 \end{aligned} \qquad \square$$

Lem. 6 can be seen as a sufficient test ($ETT_{x_0,x_1}(y|z, w) \neq 0$) for the existence of causal paths between X and Y. However, the counterfactual fairness measure fails to capture the existence of spurious discrimination.

Lemma 7. *There exists a SCM M where X and Y are not connected by any back-door path, but $ETT_{x_0,x_1}(y|z, w) \neq 0$ for some y, z, w and $x_0 \neq x_1$.*

Proof. We can prove this lemma with the same SCM M constructed in the proof of Lem. 5. In this example, we have $Z = W = \emptyset$, and there exists no back-door path between X and Y. For $y = 1$, we have:

$$ETT_{x_0,x_1}(y|z, w) = ETT_{x_0,x_1}(y) = TE_{x_0,x_1}(y) = 0.8$$

which is not zero. $\square$

As a corollary, it is immediate to see that the condition $ETT_{x_0,x_1}(y) \neq 0$ is also oblivious to the existence of spurious discrimination.

Natural Direct and Indirect Effect. Pearl (2001) introduced natural direct (NDE) and indirect (NIE) effects to measure the direct and indirect causal effect of X on Y. Formally,

Definition 9 (Natural Diret Effect(NDE))**.** Given a SCM M, the natural direct effect of intervention $X = x_1$ on $Y = y$ (relative to baseline x_0) is defined as:

$$NDE_{x_0,x_1}(Y = y) = P(y_{x_1,W_{x_0}}) - P(y_{x_0})$$

Definition 10 (Natural Indirect Effect(NDE))**.** Given a SCM M, the natural indirect effect of intervention $X = x_1$ on $Y = y$ (relative to baseline $X = x_0$) is:

$$NIE_{x_0,x_1}(y) = P(y_{x_0,W_{x_1}}) - P(y_{x_0}) \tag{24}$$

In fact, NDE and NIE measure, respectively, effects associated with direct and indirect causal paths from X to Y.

Lemma 8. *For a SCM M, if X has no direct causal path connecting Y in the causal diagram G, then $NDE_{x_0,x_1}(y) = 0$, for any $y, x_0 \neq x_1$.*

Proof. By definition,

$$\begin{aligned} &NDE_{x_0,x_1}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_0}) \\ &= \sum_{z,w} P(y_{x_1,w}|z, w_{x_0}) - P(y_{x_0}|z, w_{x_0})P(z, w_{x_0}) \\ &= \sum_{z,w} P(y_{x_1,z,w}|z, w_{x_0}) - P(y_{x_0,z,w}|z, w_{x_0})P(z, w_{x_0}) \end{aligned}$$

In this last step, $Z = Z_x = Z_{x,w}$ for any x by the exclusion restrictions rule. $Y_x = Y_{x,w} = Y_{x,z,w}$ if $Z = Z_x = Z_{x,w} = z$ and $W_x = w$ for any x, z, w (the composition axiom). Since X has no direct causal path connecting Y, by the exclusion restrictions rule, $Y_{x,z,w} = Y_{z,w}$ for any x, z, w. We thus have

$$\begin{aligned} &\sum_{z,w} P(y_{x_1,z,w}|z, w_{x_0}) - P(y_{x_0,z,w}|z, w_{x_0}) \\ &= \sum_{z,w} P(y_{z,w}|z, w_{x_0}) - P(y_{z,w}|z, w_{x_0}) = 0 \end{aligned} \qquad \square$$

Lemma 9. *For a SCM M, if X has no indirect causal path connecting Y in the causal diagram G, then $NIE_{x_0,x_1}(y) = 0$, for any $y, x_0 \neq x_1$.*

Proof. Since X has no indirect causal path connecting Y, the mediators $W = \emptyset$. We can obtain

$$NIE_{x_0,x_1}(y) = P(y_{x_0,W_{x_1}}) - P(y_{x_0})$$
$$= P(y_{x_0}) - P(y_{x_0}) = 0$$

□

Lems. 8 and 9 suggest that one could distinctly identify direct discrimination and indirect discrimination by checking, respectively, the condition $NDE_{x_0,x_1}(y) \neq 0$ and $NIE_{x_0,x_1}(y) \neq 0$. We next show that direct (indirect) discrimination is the only type of discrimination detected by NDE (NIE).

Lemma 10. *There exists a SCM M where X and Y are not connected by any indirect causal or back-door path, but $NDEx_0, x_1(y) \neq 0$ for for some $y, x_0 \neq x_1$.*

Proof. Consider the SCM M constructed in the proof of Lem. 5. In this model, X and Y are only connected by the direct causal path $X \to Y$. However, for $y = 1$, NDE is equal to:

$$NDE_{x_0,x_1}(y) = TE_{x_0,x_1}(y) = 0.8$$

which is not zero. □

Lemma 11. *There exists a SCM M where X and Y are not connected by any direct causal or back-door path, but $NIEx_0, x_1(y) \neq 0$ for some $y, x_0 \neq x_1$.*

Proof. Consider a SCM M, where X, Y, W, U are binary variables in $\{0, 1\}$, $P(U = 0) = 0.9$. Values of Y are decided by function $y = w \oplus u$ ($\oplus$ stands for the *"xor"* operator), values of W are decided by $w = x$ and X follows a uniform distribution. In this model, X and Y are only connected by the indirect causal path $X \to W \to Y$. However, for $y = 1$, NIE is equal to:

$$NIE_{x_0,x_1}(y) = P(y_{x_0,W_{x_1}}) - P(y_{x_0})$$
$$= P(U = 0) - P(U = 1) = 0.8$$

which is not zero. □

Controlled Direct Effect (CDE). The controlled direct effect (CDE) measures the sensitivity of Y to (interventional) variations of X while physically holding all the other observed variables W fixed. Formally

Definition 11 (CDE Fairness)**.** Given a SCM M, for any z, w, the controlled direct effect of intervention $X = x_1$ on $Y = y$ (relative to baseline x_0) is defined as:

$$CDE_{x_0,x_1}(y_{z,w}) = P(y_{x_1,z,w}) - P(y_{x_0,z,w})$$

We use CDEs to define a qualitative measure to capture the presence of direct causal path $X \to Y$, i.e., direct discrimination.

Lemma 12. *For a SCM M, if X has no direct causal path connecting Y in the causal diagram G, then $CDE_{x_0,x_1}(y_{z,w}) = 0$, for any y, z, w and $x_0 \neq x_1$.*

Proof. Since X has no direct causal path connecting Y, by the exclusion restrictions rule (Pearl, 2000, Sec. 7.3.2), $Y_{x,z,w} = Y_{z,w}$ for any x, z, w. We thus have

$$CDE_{x_0,x_1}(y_{z,w}) = P(Y_{x_1,z,w}) - P(y_{x_0,z,w})$$
$$= P(y_{z,w}) - P(y_{z,w}) = 0$$

□

Lem. 14 implies that the condition $CDE_{x_0,x_1}(Y_{z,w}) \neq 0$ is a sufficient test for identifying direct discrimination. However, we next show this condition is unable to detect indirect and spurious discrimination.

Lemma 13. *There exists a SCM M where X and Y are not connected by any indirect causal or back-door path, but $CDEx_0, x_1(y) \neq 0$ for for some $y, x_0 \neq x_1$.*

Proof. Consider the SCM M constructed in the proof of Lem. 5.In this model, X and Y are only connected by the direct causal path $X \to Y$. However, for $y = 1$, CDE is equal to:

$$CDE_{x_0,x_1}(y) = TE_{x_0,x_1}(y) = 0.8$$

which is not zero. □

Average QII Datta, Sen, and Zick (2016) introduced the average QII measure to capture the degree of direct influence of input X on outcome Y – the expected change in Y induced by two independent stochastic interventions $do(W \sim P(w))$ and $do(X \sim P(x))$. [1] $P(w)$ and $P(x)$ are marginals of the observational distribution $P(x, y, w)$, i.e.:

Definition 12 (Average QII)**.** The average QII of X on Y is defined as:

$$QII_X(Y) = E[Y] - E[Y_{X \sim P(x), Z, W \sim P(z,w)}]$$

QII can be used to construct a sufficient test for detecting the existence of direct discrimination when all parents of Y are observed.

Lemma 14. *For a SCM M, if X has no direct causal path connecting Y in the causal diagram G and all parents of Y are observed, then $QII_X(Y) = 0$.*

Proof. By definition,

$$QII_X(Y) = E[Y] - E[Y_{X \sim P(x), Z, W \sim P(z,w)}]$$
$$= \sum_{x,z,w} E[Y|x, z, w]P(x, z, w) - E[Y_{x,z,w}]P(x)P(z, w)$$

Since all parents of Y are observed, $E[Y|x, z, w] = E[Y_{z,w}]$, which gives:

$$\sum_{x,z,w} E[Y|x, z, w]P(x, z, w) - E[Y_{x,z,w}]P(x)P(z, w)$$
$$= \sum_{x,z,w} E[Y_{x,z,w}](P(x, z, w) - P(x)P(z, w))$$

[1] Values of X are sampled from the distribution $P(x)$.

Since X has no direct causal path connecting Y, by the exclusion restrictions rule, $Y_{x,z,w} = Y_{z,w}$ for any x, z, w. We thus have:

$$\begin{aligned}
&\sum_{x,z,w} E[Y_{x,z,w}](P(x,z,w) - P(x)P(z,w)) \\
&= \sum_{x,z,w} E[Y_{z,w}](P(x,z,w) - P(x)P(z,w)) \\
&= \sum_{z,w} E[Y_{z,w}](\sum_{x} P(x,z,w) - \sum_{x} P(x)P(z,w)) \\
&= \sum_{z,w} E[Y_{z,w}](P(z,w) - P(z,w)) = 0 \qquad \square
\end{aligned}$$

However, in general settings (e.g., the extended model), the condition $QII_X(Y) \neq 0$ is not a sufficient test of identifying any type of discrimination.

Lemma 15. *There exists a SCM M where X and Y are not connected by any path, but $QII_X(Y) \neq 0$.*

Proof. Consider a SCM M where X, Y, Z, W, U are binary variables in $\{0, 1\}$ and U follows a uniform distribution. Values of Z are decided by the function $z = u$ and values of Y are decided by the function $y = z \oplus u$. X, W follows arbitrary independent distributions $P(x), P(w)$. Values of Y are equal to:

$$y = z \oplus u = u \oplus u = 0,$$

which implies that $E[Y] = 0$. Since $P(u)$ is a uniform distribution and $z = u$, the marginal $P(z)$ is also an uniform distribution. Since Z is the only parent of Y, the exclusion restrictions rule gives $Y_{x,z,w} = Y_z$. The quantity $E[Y_{X\sim P(x),Z,W\sim P(z,w)}]$ is thus equal to

$$\begin{aligned}
E[Y_{X\sim P(x),Z,W\sim P(z,w)}] &= \sum_{x,z,w} E[Y_{x,z,w}]P(x)P(z,w) \\
= \sum_{z} E[Y_z] \sum_{x,w} P(x)P(z,w) &= \sum_{z} E[Y_z]P(z)
\end{aligned}$$

Values of Y_z are decided by the function $y = z \oplus u$, where z and u follow an independent uniform distribution respectively. Thus,

$$\sum_{z} E[Y_z]P(z) = \sum_{z,u} I\{z \oplus u = 1\}P(z)P(u) = 0.5$$

where $I\{\cdot\}$ is an indicator function. By definition, $QII_X(Y)$ is equal to

$$QII_X(Y) = E[Y] - E[Y_{X\sim P(x),Z,W\sim P(z,w)}] = -0.5$$

which is not zero. $\square$

Appendix III. Parametrizations

In this section, we provide full parametrizations for simulations in the paper.

Discrimination Detection

Standard Fairness Model. Recall that in the religious discrimination example, a company makes hiring decisions Y and can potentially use the following attributes that are available in its database: 1) the religious belief X, 2) the educational background Z, and 3) the location W of the applicant. Fig. 1(a) is the causal model for this setting. X, Y, Z, W are binary variables taking values in $\{0, 1\}$. Z following a uniform distribution such that $P(Z = 0) = 0.5$. Values of X are decided by function $x = \neg z$, and values of W are decided by function $w = x$. The hiring decision Y is made solely based on Z, s.t., $y = f_y(z) = z$. Noting that education is critical for business success, the same is considered a legitimate reason for hiring (path $X \leftarrow Z \rightarrow Y$ is justified).

Extended Fairness Model. Consider a instance M of Extended fairness model, where $X, Y, Z, W, U_1, U_2 \in \{0, 1\}$. U_1, U_2 are exogenous variables following distributions $P(U_1 = 0) = 0.9$ and $P(U_2 = 0) = 0.1$. Values of X, Y, Z, W are decided by, respectively, functions

$$\begin{aligned}
x &= u_1, \quad y = x \oplus z \oplus w \oplus u_1 \oplus u_2, \\
z &= u_1, \quad w = x \oplus z \oplus u_1,
\end{aligned}$$

where $\oplus$ stand for the *"xor"* operator.

Discrimination Explanation

For simulations in this section, we consider a logistic model similar to the one treated in (MacKinnon et al., 2007). Let us retain the linear model in Eqs. with one modification: the outcome will be a threshold-based indicator of the linear outcome Y in the linear-standard model. Formally, we regard

$$Y^* = \gamma_{xy}x + \gamma_{zy}z + \gamma_{wy}w + u_y$$

as a latent variable, and define the outcome Y as

$$y = I\{\gamma_0 + \gamma_{xy}x + \gamma_{zy}z + \gamma_{wy}w + u_y\},$$

where $I\{\cdot\}$ is an indicator function, and γ_0 is some unknown threshold level. For simulations in the paper, we use parameters $\gamma_{xy} = \gamma_{wy} = \gamma_{zy} = \gamma_{xw} = \gamma_{zx} = 0.5$. If U_y follows the logistic distribution,

$$P(U_y < u) = L(u) \triangleq \frac{1}{1 + e^{-u}}.$$

Thus, $P(Y = 1|x, w, z)$ attains the form

$$\begin{aligned}
P(Y = 1|x, w, z) &= \frac{1}{1 + e^{-(\gamma_0 + \gamma_{xy}x + \gamma_{wy}w + \gamma_{zy}z)}} \\
&= L(\gamma_0 + \gamma_{xy}x + \gamma_{wy}w + \gamma_{zy}z)
\end{aligned}$$

We assume that U_z, U_w are normal with zero mean and infinitesimal $\sigma_z << 1, \sigma_w << 1$.

Given this logistic model, we will now compute the TV and counterfactual DE, IE and SE associated with the transition from $x_0 = 0$ to $x_1 = 1$. From the Causal Explanation

Formula in the standard model (Thm. 2), we have:

$$
\begin{aligned}
&DE_{x_1,x_0}(Y|x_1) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[L(\gamma_0+\gamma_{wy}w+\gamma_{zy}z)-L(\gamma_0+\gamma_{xy}+\gamma_{wy}w+\gamma_{zy}z)] \\
&\cdot f_{W|X}(w|x_1)f_{Z|X}(z|x_1)dwdz \\
&= L(\gamma_0+\gamma_{wy}+\gamma_{zy})-L(\gamma_0+\gamma_{xy}+\gamma_{wy}+\gamma_{zy})+0(\sigma_z^2+\sigma_w^2) \\
&IE_{x_0,x_1}(Y|x_0) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[L(\gamma_0+\gamma_{wy}w+\gamma_{zy}z)] \\
&\cdot [f_{W|X}(w|x_1)-f_{W|X}(w|x_1)]f_{Z|X}(z|x_1)dwdz \\
&= L(\gamma_0+\gamma_{wy}+\gamma_{zy})-L(\gamma_0+\gamma_{zy})+0(\sigma_z^2+\sigma_w^2) \\
&SE_{x_0,x_1}(Y) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[L(\gamma_0+\gamma_{wy}w+\gamma_{zy}z)] \\
&\cdot f_{W|X}(w|x_0)[f_{Z|X}(z|x_1)-f_{Z|X}(z|x_0)]dwdz \\
&= L(\gamma_0+\gamma_{zy})-L(\gamma_0)+0(\sigma_z^2+\sigma_w^2)
\end{aligned}
$$

where $0(\sigma_z^2+\sigma_w^2)\to 0$ as $\sigma_z,\sigma_w\to 0$.

References

Datta, A.; Sen, S.; and Zick, Y. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (SP), 2016 IEEE Symp.*, 598–617.

Halpern, J. Y. 2000. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research* 12(1):317–337.

Koller, D., and Friedman, N. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Kusner, M. J.; Loftus, J. R.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*.

MacKinnon, D. P.; Lockwood, C. M.; Brown, C. H.; Wang, W.; and Hoffman, J. M. 2007. The intermediate endpoint effect in logistic and probit regression. *Clinical Trials* 4(5):499–513.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press. 2nd edition, 2009.

Pearl, J. 2001. Direct and indirect effects. In *Proc. of the 17th Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann. 411–420.