Survey on Causal-based Machine Learning Fairness Notions

Karima Makhlouf
makhlouf.karima@courrier.uqam.ca
Université du Québec à Montréal
Montréal, Canada

Sami Zhioua
szhioua@hct.ac.ae
Higher Colleges of Technology
Dubai, UAE

Catuscia Palamidessi
catuscia@lix.polytechnique.fr
Inria, École Polytechnique, IPP
Paris, France

arXiv:2010.09553v3 [cs.LG] 18 Jan 2021

Abstract

Addressing the problem of fairness is crucial to safely use machine learning algorithms to support decisions with a critical impact on people's lives such as job hiring, child maltreatment, disease diagnosis, loan granting, etc. Several notions of fairness have been defined and examined in the past decade, such as, statistical parity and equalized odds. The most recent fairness notions, however, are causal-based and reflect the now widely accepted idea that using causality is necessary to appropriately address the problem of fairness. This paper examines an exhaustive list of causal-based fairness notions, in particular their applicability in real-world scenarios. As the majority of causal-based fairness notions are defined in terms of non-observable quantities (e.g. interventions and counterfactuals), their applicability depends heavily on the identifiability of those quantities from observational data. In this paper, we compile the most relevant identifiability criteria for the problem of fairness from the extensive literature on identifiability theory. These criteria are then used to decide about the applicability of causal-based fairness notions in concrete discrimination scenarios.

Keywords: Fairness, machine learning, causal-based, identifiability.

1 Introduction

The first endeavor of the research community to achieve fairness consisted in developing correlation/association-based notions, including statistical parity [Dwork et al. 2012], equalized odds [Hardt et al. 2016], predictive parity [Chouldechova 2017], and calibration [Chouldechova 2017] which primarily focus on discovering the discrepancy of statistical metrics between individuals or sub-populations. The problem of these notions is that they lack in reasoning about cause-effect relations between attributes. Moreover, the research community is realizing that fairness cannot be well assessed based only on mere correlation or association [Hardt et al. 2016; Huan et al. 2020; Khademi et al. 2019; Kusner et al. 2017; Zhang and Bareinboim 2018a,b]. A canonical example is Simpson's paradox [Pearl 2009], where the statistical conclusions drawn from the sub-populations differ from that from the whole population. On the other hand, discrimination claims usually require plaintiffs to demonstrate a causal connection between the challenged decision (e.g. hiring, firing, admission) and the sensitive feature (e.g. gender, race). It is then necessary to investigate the causal relationship between the sensitive attribute and the decision rather than the associated relationship. Various causal-based fairness notions have been recently proposed to tackle the problem of automated discrimination through causal inference lenses.

Causal-based fairness notions differ from the previous statistical fairness approaches in that they are not totally based on data but consider additional knowledge about the structure of the world, in the form of a causal model. This additional knowledge helps us understand how data is generated in the first place and how changes in variables propagate in a system. Most of these fairness notions are defined in terms of non-observable quantities such as interventions (to simulate random experiments) and counterfactuals (which consider other hypothetical worlds, in addition to the actual world). Such quantities cannot be always uniquely computed from observational data. This problem is known as identifiability and hinders significantly the usefulness of causal-based notions in practical scenarios.

There is a large body of work in the literature that addresses the problem of identifiability, typically through complex graphical criteria [Avin et al. 2005; Galles and Pearl 1995; Huang and Valtorta 2006; Malinsky et al. 2019; Pearl 2001, 2009; Shpitser and Pearl 2006, 2007, 2008; Tian and Pearl 2002; Wu et al. 2019a].

In this paper, we first provide an exhaustive list of causal-based fairness notions. Then, we compile the most relevant identifiability criteria for the specific problem of discrimination discovery. These criteria are grouped into intervention, counterfactual, direct and indirect effect, and path-specific effect identifiability results. By placing the fairness notions in the three rungs causation ladder of Pearl [Pearl and Mackenzie 2018] and using the identifiability criteria, we propose a guideline for applying causal-based fairness notions in real-scenarios.

2 The need for causality: an example

Consider the hypothetical example[1] of an automated system for deciding whether to fire a teacher at the end of the academic year. Deployed teacher evaluation systems have been suspected of bias in the past. For example, IMPACT is a teacher evaluation system used in the city of Washington DC and have been found to be unfair against teachers from

[1]Inspired by the prior convictions example in [Nabi and Shpitser 2018].

minority groups [O'Neill 2016; Quick 2015; Rhee 2019]. Assume that the system takes as input two features, namely, the location of the school where the teacher is working (C) and the initial[2] average level of the students in her class (A). The outcome is whether to fire the teacher (Y). Assume also that all 3 variables are binary with the following values: if the school is located in a high-income neighborhood, $C = 1$, otherwise (the school is located in a low-income neighborhood), $C = 0$. If the initial average score for the students assigned to the teacher is high, $A = 1$, otherwise (initial level is low), $A = 0$. Firing a teacher corresponds to $Y = 1$, while retaining her corresponds to $Y = 0$. The level of students in a given class can be influenced by several variables, but in this example, assume that it is only influenced by the location of the school; students in high-income neighborhoods are more advantaged and typically perform better in school.

Assume now that the automated decision system is suspected to be biased by the initial level of students assigned to the teacher. That is, it is claimed that the system will more likely fire teachers who have been assigned classes with low level students at the beginning of the academic year which is clearly unfair. The sensitive attribute in this case is the initial level of students assigned to the teacher (A). For concreteness, consider the prediction system that yields the following conditional probabilities:

$$
\begin{aligned}
&P(Y = 1 \mid A = 1, C = 0) = 0.02 \qquad && P(A = 1 \mid C = 0) = 0.2\\
&P(Y = 1 \mid A = 1, C = 1) = 0.0675 && P(A = 1 \mid C = 1) = 0.8\\
&P(Y = 1 \mid A = 0, C = 0) = 0.01 && P(A = 0 \mid C = 0) = 0.8\\
&P(Y = 1 \mid A = 0, C = 1) = 0.25 && P(A = 0 \mid C = 1) = 0.2
\end{aligned}
$$

and that the dataset is collected from a population where schools are located with equal proportions in high-income and low-income neighborhood, that is, $P(C = 1) = P(C = 0) = 0.5$. Assume also that the proportion of classes with a low initial average level of students is the same as the one with high average initial level of students, that is, $P(A = 1) = P(A = 0) = 0.5$. To keep the scenario simple, assume that the level of students A does not depend on any other feature except C and that the firing decision Y depends only on A and C.

A simple approach to check the fairness of the firing decision Y with respect to the sensitive attribute A is to contrast the conditional probabilities: $P(Y = 1 \mid A = 0)$ and $P(Y = 1 \mid A = 1)$ which quantify, respectively, the likelihood of firing a teacher given that she is assigned students with an initial low level versus and the likelihood of firing a teacher given that she is assigned students with an initial high level class. Such probabilities can be computed as follows:

$$
\begin{aligned}
P(Y = 1 \mid A = a) = \sum_{c \in \{0,1\}} & P(Y = 1 \mid A = a, C = c,)\\
& \times P(A = a \mid C = c)
\end{aligned}
$$

Hence,

$$
\begin{aligned}
P(Y = 1 \mid A = 1) = 0.02 \times 0.2 + 0.0675 \times 0.8 = 0.058\\
P(Y = 1 \mid A = 0) = 0.01 \times 0.8 + 0.25 \times 0.2 = 0.058
\end{aligned}
$$

As the values are equal, the rates of firing between teachers who were assigned low level students and high level students appear to be equal and hence no discrimination is detected[3]. This conclusion is flawed because it doesn't consider the mechanism by which the data is generated. In particular, the location of the school in which the teacher is working influences both the initial level of students assigned to her as well as the decision to fire or retain her. In such situations, if the data is collected in a certain way, the above conditional probabilities may end up being equal despite the presence of bias. The above example is one of such cases since the conditional probability $P(A = 0 \mid C = 0) = 0.8$ is quite extreme. Notice that teachers in low-income neighborhoods are more likely to be assigned classes where the average level of students is low, but not to that extent (80% of classes in low-income neighborhood have an average low level). In general, any statistical fairness notion which relies solely on correlation between variables, will fail to detect such bias.

To avoid such misleading conclusions, the causal relationships between variables should be considered. Figure 1 illustrates the causal relations between the three variables of the above example where the location of the school C is a confounder. Based on such causal graph, a firing decision system is fair if is as likely to fire teachers in the following two hypothetical cases: (1) when *all teachers in the population are assigned students of low level on average, and (2) when* all teachers in the population are assigned students of high level on average. This is achieved using intervention ($do()$ operator)[4] and allows to break the problematic dependence between A and C. The probabilities of firing a teacher in these two hypothetical cases are expressed as $P(Y_{A=0} = 1) = P(Y = 1 \mid do(A = 0))$ and $P(Y_{A=1} = 1) = P(Y = 1 \mid do(A = 1))$ respectively. In this simple graph, and assuming no other variable is used in the prediction, these probabilities can be computed as follows:

$$
P(Y_{A=a} = 1) = \sum_{c \in \{0,1\}} P(Y = 1 \mid A = a, C = c) \times P(C = c)
$$

Hence,

$$
\begin{aligned}
P(Y_{A=1} = 1) = 0.02 \times 0.5 + 0.0675 \times 0.5 = 0.0437\\
P(Y_{A=0} = 1) = 0.01 \times 0.5 + 0.25 \times 0.5 = 0.13
\end{aligned}
$$

The values confirm the existence of a bias against teachers which are assigned students of low level on average. These teachers are more likely to belong to minority groups.

[2]At the beginning of the academic year.

[3]This corresponds to statistical parity.

[4]Intervention and the $do()$ operator will be explained further in Section 5.1.

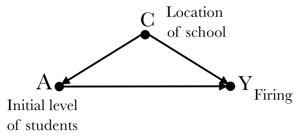


Figure 1. Causal graph of the firing example.

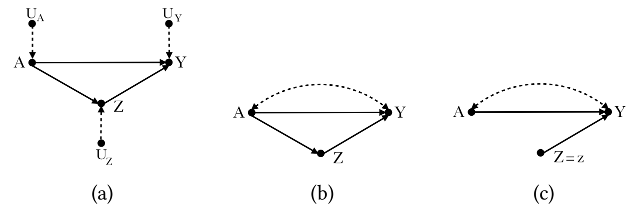


Figure 2. Markovian and semi-Markovian causal models.

3 Preliminaries and Notation

Variables are denoted by capital letters. In particular, A is used for the sensitive variable (e.g., gender, race, age) and Y is used for the outcome of the automated decision system (e.g., health-care intervention, hiring, admission, releasing on parole). Small letters denote specific values of variables (e.g., $A = a'$, $W = w$). Bold capital and small letters denote a set of variables and a set of values, respectively.

A structural causal model [Pearl 2009] is a tuple $M = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ where:

- $\mathbf{U}$ is a set of exogenous variables which cannot be observed or experimented on but constitute the background knowledge behind the model.
- $\mathbf{V}$ is a set of observable variables which can be experimented on.
- $\mathbf{F}$ is a set of structural functions where each f_i is mapping $U \cup V \rightarrow V \backslash \{V_i\}$ which represents the process by which variable V_i changes in response to other variables in $U \cup V$.
- $P(\mathbf{u})$ is a probability distribution over the unobservable variables $\mathbf{U}$.

Causal assumptions between variables are captured by a causal diagram G which is a directed acyclic graph (DAG) where vertices represent variables and directed edges represent functional relationships between the variables. Directed edges can have two interpretations. A probabilistic interpretation where the edge represents a dependency among the variables such that the direction of the edge is irrelevant. A causal interpretation where the edge represents a causal influence between the corresponding variables such that the direction of the edge matters. Unobserved variables $\mathbf{U}$, which are typically not represented in the causal diagram, can be either mutually independent (Markovian model) or dependent from each others. In case the unobserved variables can be dependent and each $U_i \in \mathbf{U}$ is used in at most two functions in F, the model is called semi-Markovian. In causal diagrams of semi-Markovian models, dependent unobservable variables (unobserved confounders) are represented by a dotted bi-directed edge between observable variables. Figure 2 shows causal graphs of Markovian model (Figure 2(a)), semi-Markovian model (Figures 2(b)) and semi-Markovian model after intervening on Z (Figure 2(c)).

An intervention, noted $do(V = v)$, is a manipulation of the model that consists in fixing the value of a variable (or a set of variables) to a specific value regardless of the corresponding function f_v. Graphically, it consists in discarding all edges incident to the vertex corresponding to variable V. Figure 2(c) shows the causal diagram of the manipulated model after intervention $do(Z = z)$ denoted $M_{Z=z}$ or M_z for short. The intervention $do(V = v)$ induces a different distribution on the other variables. For example, in Figure 2(c), $do(Z = z)$ results in a different distribution on Y, namely, $P(Y|do(Z = z))$. Intuitively, while $P(Y|Z = z)$ reflects the population distribution of Y among individuals whose Z value is z, $P(Y|do(Z = z)$ reflects the population distribution of Y if *everyone in the population* had their Z value fixed at z. The obtained distribution $P(Y|do(Z = z)$ can be considered as a *counterfactual* distribution since the intervention forces Z to take a value different from the one it would take in the actual world. Such counterfactual variable is noted $Y_{Z=z}$ or Y_z for short[5]. The term counterfactual quantity is used for expressions that involve explicitly multiple worlds. In Figure 2(b), consider the expression $P(y_{a'}|Y = y, A = a) = P(y_{a'}|y, a)$. Such expression involves two worlds: an observed world where $A = a$ and $Y = y$ and a counterfactual world where $Y = y$ and $A = a'$ and it reads "the probability of $Y = y$ had A been a' given that we observed $Y = y$ and $A = a$". In the common example of job hiring, if A denotes race (a :white, a':non-white) and Y denotes the hiring decision (y:hired, y':not hired), $P(y_{a'}|y, a)$ reads "given that a white applicant has been hired, what is the probability that the same applicant is still being hired had he been non-white". Nesting counterfactuals can produce complex expressions. For example, in the relatively simple model of Figure 2(b), $P(y_{a,z_{a'}}|y'_{a'})$ reads the probability of $Y = y$ had (1) A been a' and (2) Z been z when A is a', given that an intervention $A = a'$ produced y'. This expression involves three worlds: a world where $A = a$, a world where $Z = z_{a'}$, and a world where $A = a'$. Such complex expressions are used to characterize direct, indirect, and path-specific effects.

Causal-based discrimination discovery aims at telling if the outcome of an automated decision making is fair or discriminative. Several causal-based fairness notions are defined in the literature (Section 4) and expressed in terms of joint, conditional, interventional, and counterfactual probabilities.

[5]The notations $Y_{Z\leftarrow z}$ and $Y(z)$ are used in the literature as well. $P(Y = y|do(Z = z)) = P(Y_{Z=z} = y) = P(Y_z = y) = P(y_z)$ is used to define the causal effect of z on Y.

The application of a fairness notion requires as input a dataset D and a causal graph G. While joint probabilities (e.g. $P(X = x, Y = y, Z = z)$) and conditional probabilities (e.g. $P(Y = y|X = x)$) can be trivially estimated from the dataset D, probabilities involving interventions or counterfactuals cannot always be estimated from D and G. When a probability can be estimated from observable data (D), it is said to be *identifiable*. Otherwise it is *unidentifiable*. More formally, let M_1 and M_2 be two causal models sharing the same causal graph (not including the unobservable variable $\mathbf{U}$) and the same set of probability distributions ψ, a quantity Q (e.g. intervention or counterfactual) is identifiable using ψ (noted ψ-identifiable), if the value of Q is unique and computable from ψ in any models M_1 and M_2. In other words, if there exists two models M_1 and M_2 sharing the same graph structure and the same probability distributions, but yielding different Q values, then Q is unidentifiable. Typically, the identifiability of interventional and counterfactual quantities depends on the structure of the graph, in particular, the location of unobserved confounding variables. Identifiability criteria are summarized in Section 5.

4 Causality-based Fairness notions

Without loss of generality, assume that the sensitive attribute A and the outcome Y are binary variables where $A = a_0$ denotes the privileged group (e.g. male), typically considered as the reference in characterizing discrimination, and $A = a_1$ the disadvantaged group (e.g. female).

The most common non-causal fairness notion is total variation (TV), known as statistical parity, demographic parity, or risk difference. The total variation of $A = a_1$ on the outcome $Y = y$ with reference $A = a_0$ is defined using conditional probabilities as follows:

$$TV_{a_1,a_0}(y) = P(y \mid a_1) - P(y \mid a_0) \tag{1}$$

Intuitively, $TV_{a_1,a_0}(y)$ measures the difference between the conditional distributions of Y when we (passively) observe A changing from a_0 to a_1. The main limitation of TV is purely statistical nature which makes it unable to reflect the causal relationship between A and Y, that is, it is insensitive to the mechanism by which data is generated. Total effect (TE) [Pearl 2009] is the causal version of TV and is defined in terms of experimental probabilities as follows:

$$TE_{a_1,a_0}(y) = P(y_{a_1}) - P(y_{a_0}) \tag{2}$$

TE measures the effect of the change of A from a_1 to a_0 on $Y = y$ along all the causal paths from A to Y. Intuitively, while TV reflects the difference in proportions of $Y = y$ in the current cohort, TE reflects the difference in proportions of $Y = y$ in the entire population. A more involved causal-based fairness notion considers the effect of a change in the sensitive attribute value (e.g. gender) on the outcome (e.g. probability of admission) given that we already observed the outcome for that individual. This typically involves an impossible situation which requires to go back in the past and change the sensitive attribute value. Mathematically, this can be formalized using counterfactual quantities. The simplest fairness notion using counterfactuals is the effect of treatment on the treated (ETT) [Pearl 2009] which is defined as:

$$ETT_{a_1,a_0}(y) = P(y_{a_1} \mid a_0) - P(y \mid a_0) \tag{3}$$

$P(y_{a_1} \mid a_0)$ reads the probability of $Y = y$ had A been a_1, given A had been observed to be a_0. Such probability involves two worlds: an actual world where $A = a_0$ and a counterfactual world where for the same individual $A = a_1$. Notice that $P(y_{a_0}|a_0) = P(y|a_0)$, a property called consistency [Pearl 2009].

TV, TE, and ETT fall into the framework of disparate impact [Barocas and Selbst 2016] which aims at ensuring the equality of outcomes among all groups (protected and unprotected). An alternative framework is the disparate treatment [Barocas and Selbst 2016] which seeks equality of treatment achievable through prohibiting the use of the sensitive attribute in the decision process.

Common fairness notions from the disparate treatment framework include direct effect, indirect effect, and path-specific effect. An effect can be deemed fair or unfair by an expert of the scenario at hand. Unfair effect is called discrimination. Direct discrimination is assessed using causal effect along direct edge from A to Y while indirect discrimination is measured using the causal effect along causal paths that pass through proxy attributes[6].

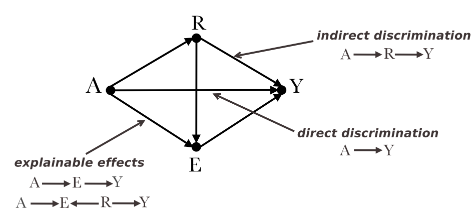


Figure 3. Teacher firing example where A: initial level of students, Y: firing decision, R: number of reported disruptive incidents in class, E: difference between final level and initial levels of students. $A \to Y$: direct discrimination. $A \to R \to Y$:indirect discrimination through the redlining variable R. $A \to E \to Y$: explainable effect of A on Y due to E.

Figure 3 presents another causal graph of the firing example (Section 1) involving a redlining variable R (e.g. the number of disruptive incidents reported in the classrooms, which is clearly dependent on the initial level of students in the class, and hence a proxy for that sensitive attribute) and an explaining variable E (e.g. the difference between the

[6]A proxy is an attribute that cannot be objectively justified if used in the decision making process. It is Known also as redlining attribute.

final level (at the end of the academic year) and the initial level, which is a legitimate feature to consider in the decision making.

Natural direct effect (NDE) [Pearl 2001] is a common notion that measures the direct discrimination and is defined as:

$$NDE_{a_1,a_0}(y) = P(y_{a_1,\mathbf{Z}_{a_0}}) - P(y_{a_0}) \tag{4}$$

Where $\mathbf{Z}$ is the set of mediator variables and $P(y_{a_1,\mathbf{Z}_{a_0}})$ is the probability of $Y = y$ had A been a_1 and had $\mathbf{Z}$ been the value it would naturally take if $A = a_0$. That is, A is set to a_1 in the single direct path $A \longrightarrow Y$ and is set to a_0 in all other indirect paths ($A \longrightarrow R \longrightarrow Y$ and $A \longrightarrow E \longrightarrow Y$).

On the other hand, natural indirect effect (NIE) [Pearl 2001] measures the indirect effect of A on Y and is defined as:

$$NIE_{a_1,a_0}(y) = P(y_{a_0,\mathbf{Z}_{a_1}}) - P(y_{a_0}) \tag{5}$$

The problem with NIE is that it does not distinguish between the fair (explainable) and unfair (indirect discrimination) effects. Path-specific effect [Pearl 2009] is a more nuanced measure that characterizes the causal effect in terms of specific paths.

Given a path set π, the π-specific effect is defined as:

$$PSE^{\pi}_{a_1,a_0}(y) = P(y_{a_1|_{\pi},a_0|_{\overline{\pi}}}) - P(y_{a_0}) \tag{6}$$

where $P(y_{a_1|_{\pi},a_0|_{\overline{\pi}}})$ is the probability of $Y = y$ in the counterfactual situation where the effect of A on Y with the intervention (a_1) is transmitted along π, while the effect of A on Y without the intervention (a_0) is transmitted along paths not in π (denoted by: $\overline{\pi}$).

4.1 No unresolved discrimination

No unresolved discrimination [Kilbertus et al. 2017] is a fairness notion that falls into the disparate treatment framework and focuses on the indirect causal effects from A to Y. No unresolved discrimination is satisfied when no directed path from A to Y is allowed, except via a resolving (explaining) variable E. A resolving variable is any variable in a causal graph that is influenced by the sensitive attribute in a manner that it is accepted as nondiscriminatory. Figure 4 presents two alternative causal graphs for the teacher firing example. The graph at the left exhibits unresolved discrimination along the heavy paths: $A \longrightarrow R \longrightarrow Y$ and $A \longrightarrow Y$. By contrast, the graph at the right does not exhibit any unresolved discrimination as the effect of A on Y is justified by the resolved variable E: $A \longrightarrow E \longrightarrow Y$.

4.2 No proxy discrimination

Similarly to no unresolved discrimination, no proxy discrimination [Kilbertus et al. 2017] focuses on indirect discrimination. A causal graph exhibits potential proxy discrimination if there exists a path from the protected attribute A to the outcome Y that is blocked by a proxy/redlining variable R. It is called proxy because it is used to decide about the outcome Y while it is a descendent of A which is significantly

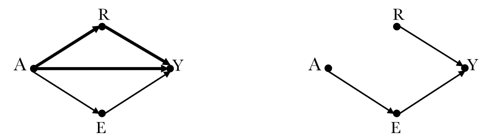


Figure 4. Two alternative graphs for the teacher firing example. Y exhibits unresolved discrimination in the left graph (along the heavy paths), but not in the right one.

correlated with it in such a way that using the proxy in the decision has almost the same impact as using A directly. An outcome variable Y exhibits no proxy discrimination if the equality:

$$P(Y \mid do(R = r)) = P(Y \mid do(R = r')) \quad \forall\, r, r' \in dom(R) \tag{7}$$

holds for any potential proxy R.

Figure 5 shows two similar causal graphs for the same teacher firing example. The causal graph at the left presents a potential proxy discrimination via the path: $A \longrightarrow R \longrightarrow Y$. However, the graph at the right is free of proxy discrimination as the edge between A and its proxy R has been removed due to the intervention on R ($R = r$).

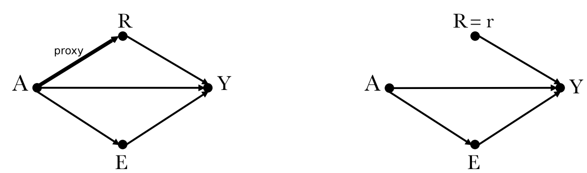


Figure 5. The graph at the left exhibits a potential proxy discrimination (along the heavy edge between A and R) but not in the right one.

4.3 Fair on Average Causal Effect (FACE) and Fair on Average Causal Effect on the Treated (FACT)

Khademi et al. [Khademi et al. 2019] defined two group-based fairness notions based on the potential outcome framework [Imbens and R 2015] called FACE and FACT. FACE considers the average causal effect of the sensitive attribute A on the outcome Y at a population level while FACT focuses on the same effect but on the sub-population/group level. Let $Y_i^{(a)}$ be the potential outcome of a data point i had A been a. Using FACE, an outcome Y is said to be fair, on average over all individuals (data points) in the population, with respect to A, iff:

$$\mathbf{E}[Y_i^{(a_1)} - Y_i^{(a_0)}] = 0 \tag{8}$$

where $\mathbf{E}[.]$ is the expectation of a random variable over all data inputs. This is equivalent to the expected value of the $TE_{a_1,a_0}(Y)$ (Eq. 2) in the causal model of Pearl [Pearl 2009]. Using FACT, an outcome Y is said to be fair with respect to the sensitive attribute A, on average over individuals of the group $A = a_0$, iff:

$$\mathbf{E}[Y_i^{(a_1)} - Y_i^{(a_0)} | A^i = a_0] = 0 \quad (9)$$

This is equivalent to the expected value of $ETT_{a_1,a_0}(Y)$ (Eq. 3). Under a set of assumptions (Section 3 in [Khademi et al. 2019]), these two fairness measures can be estimated empirically without addressing the identifiability problem. Khademi et al. estimated FACE using Inverse Probability Weighting (IPW) [Robins et al. 2000] and FACT using matching methods [Stuart 2010].

4.4 Counterfactual fairness

Counterfactual fairness [Kusner et al. 2017] is a fine-grained variant of ETT conditioned on all attributes. That is, an outcome Y is counterfactually fair if under any assignment of values $\mathbf{X} = \mathbf{x}$,

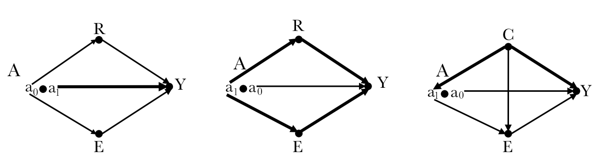
$$P(y_{a_1} \mid \mathbf{X} = \mathbf{x}, A = a_0) = P(y_{a_0} \mid \mathbf{X} = \mathbf{x}, A = a_0) \quad (10)$$

where $\mathbf{X} = \mathbf{V} \backslash \{A, Y\}$ is the set of all remaining variables. Since conditioning is done on all remaining variable $\mathbf{X}$, counterfactual fairness is an individual notion. According to Eq. 10, counterfactual fairness is satisfied if the probability distribution of the outcome Y is the same in the actual and counterfactual worlds, for every possible individual.

Kusner et al. [Kusner et al. 2017] tested counterfactual fairness by generating, for every individual in the population, another sample with counterfactual sensitive value. Then, they compared the density functions of the actual samples with the counterfactual samples. To be fair, a predictor should produce outcome values where actual and counterfactual density plots are identical.

4.5 Counterfactual Effects

By conditioning on the sensitive attribute $A = a$, Zhang and Bareinboim [Zhang and Bareinboim 2018b] defined two variants of NDE (Eq. 4 and NIE (Eq. 5) which focus on the direct and indirect effect for a specific group. In addition, they characterize a third type of effect, spurious, which considers the back-door paths between A and Y, that is, paths with an arrow into A. For instance, in Figure 6, the observed disparities between the protected and unprotected groups can be decomposed into direct ($A \longrightarrow Y$), indirect ($A \longrightarrow R \longrightarrow Y$ and $A \longrightarrow E \longrightarrow Y$) and spurious ($A \longleftarrow C \longrightarrow Y$) effects.



(a) Direct effect. (b) Indirect effect. (c) Spurious effect.

Figure 6. Direct, indirect, and spurious effects.

The three effects are defined as follows:

$$DE_{a_1,a_0}(y|a) = P(y_{a_1,\mathbf{Z}_{a_0}}|a) - P(y_{a_0}|a) \quad (11)$$
$$IE_{a_1,a_0}(y|a) = P(y_{a_0,\mathbf{Z}_{a_1}}|a) - P(y_{a_0}|a) \quad (12)$$
$$SE_{a_1,a_0}(y) = P(y_{a_0}|a_1) - P(y|a_0) \quad (13)$$

where in Eq. 11 and 12, a can be a_0 or a_1. Considering the simple job hiring example and focusing on the unprotected group ($A = a_1$ e.g. females), $DE_{a_1,a_0}(y|a_1)$ measures the change in the probability of Y (e.g. hiring) had A been a_1 (e.g. female), while mediators $\mathbf{Z}$ are kept at the level they would take had A been a_0 (e.g. male), in particular for the individuals $A = a_1$ (e.g. female). $SE_{a_1,a_0}(y)$ reads the change in the probability of hiring Y had A been a_0 (e.g. female) for the individuals that would naturally possess $A = a_0$ versus a_1. In Figure 6(c), the heavy path represents the spurious effect between A and Y through the confounder C. If there is no such back-door path, $SE_{a_1,a_0}(y)$ would be zero.

Interestingly, by considering these fine grained variants of NDE and NIE, it is possible to decompose $TV_{a_1,a_0}(y)$ (Eq. 1 and $ETT_{a_1,a_0}(y)$ (Eq. 3) as follows:

$$TV_{a_1,a_0}(y) = SE_{a_1,a_0}(y) + IE_{a_1,a_0}(y|a_1) - DE_{a_0,a_1}(y|a_1) \quad (14)$$
$$ETT_{a_1,a_0}(y) = DE_{a_1,a_0}(y|a_0) - IE_{a_0,a_1}(y|a_0) \quad (15)$$

In linear models, the three types of effects sum up to the total variation:

$$TV_{a_1,a_0}(y) = SE_{a_1,a_0}(y) + IE_{a_1,a_0}(y|a_1) + DE_{a_1,a_0}(y|a) \quad (16)$$

4.6 Counterfactual Error Rates

Equalized odds is an important statistical fairness notion which requires equality of error rates (TPR and FPR) across sub-populations, that is,

$$ER_{a_1,a_0}(\hat{y}|y) = P(\hat{y} \mid a_1, y) - P(\hat{y} \mid a_0, y) = 0 \quad (17)$$

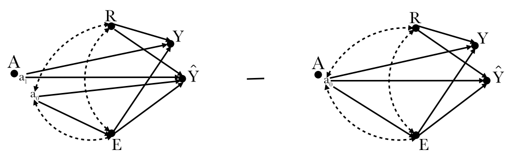
where $\hat{y}$ denotes the prediction while y denotes the true outcome. The problem of this statistical notion is the difficulty to identify the causes behind the discrimination if any. Zhang and Bareinboim [Zhang and Bareinboim 2018a] decompose equalized odds (Eq. 17) using three counterfactual measures corresponding to the direct, indirect and spurious effects of A on $\hat{Y}$. The three measures are counterfactual direct error rate, counterfactual indirect error rate, and counterfactual spurious error rate. Let $\hat{y} = f(\hat{\mathbf{pa}})$ be a classifier where $\hat{\mathbf{PA}}$ is the set of input features (parent variables of $\hat{Y}$) for the classifier. The counterfactual error rates for a sub-population a, y (with prediction $\hat{y} \neq y$) are defined as:

$$ER^d_{a_1,a_0}(\hat{y} \mid a, y) = P(\hat{y}_{a_1,y,(\hat{\mathbf{PA}} \backslash A)_{a_0,y}} \mid a, y) - P(\hat{y}_{a_0,y} \mid a, y) \quad (18)$$

$$ER^i_{a_1,a_0}(\hat{y} \mid a, y) = P(\hat{y}_{a_0,y,(\hat{\mathbf{PA}} \backslash A)_{a_1,y}} \mid a, y) - P(\hat{y}_{a_0,y} \mid a, y) \quad (19)$$

$$ER^s_{a_1,a_0}(\hat{y} \mid y) = P(\hat{y}_{a_0,y} \mid a_1, y) - P(\hat{y}_{a_0,y} \mid a_0, y) \quad (20)$$

For example, the counterfactual direct error rate (Eq. 18) measures the error rate (disparity between the true and the predicted outcome) in terms of the direct effects of the sensitive attribute A on the prediction $\hat{Y}$. In the simple job hiring example, considering the rejected females sub-population ($a = a_1$ and y = rejected), it reads: for a rejected female candidate, how would the prediction $\hat{Y}$ change had the candidate been a female (A been a_1), while keeping all the other features $\hat{\mathbf{PA}} \setminus A$ at the level that they would attain had "she was male", compared to the prediction $\hat{Y}$ she would receive had "she was male" and being rejected?



(a) $P(\hat{y}_{a_1,y,E_{a_0,y},R}) \mid a_0, y)$ (b) $P(\hat{y} \mid a_0, y)$

Figure 7. Illustration of the counterfactual direct error rate in the job hiring example. There are unobserved confounders between A, R and E because the definition is conditioning on a collider (Y).

Figure 7 illustrates how counterfactual direct error rate is applied. Thus, the input of A to the direct path $A \longrightarrow \hat{Y}$ is changed from a_0 (baseline) to a_1, while keeping the value of A to the other variables (R and E) fixed at the baseline level (a_0, y). Since the direct path $A \longrightarrow \hat{Y}$ is the only difference between models of Figure 7 (a-b), the change in $\hat{Y}$ measures the direct effect of A on $\hat{Y}$. The variable $E_{a_0,y} = E$ since E is a non-descendant node of A and Y.

Interestingly, the statistical equalized odd error rate (Eq. 17) can be decomposed in terms of the three above causal-based error rates:

$$ER_{a_1,a_0}(\hat{y} \mid y) = ER^d_{a_1,a_0}(\hat{y} \mid a_0, y) - ER^i_{a_0,a_1}(\hat{y} \mid a_0, y) - ER^s_{a_0,a_1}(\hat{y} \mid y) \tag{21}$$

4.7 Individual direct discrimination

Individual direct discrimination [Zhang et al. 2016] aims to discover the direct discrimination at the individual level. It is based on situation testing [Bendick 2007], a legally grounded technique for analyzing the discrimination at an individual level. It consists in comparing the individual with similar individuals from both groups (protected and unprotected). That is, for an individual i in question, find the k other individuals which are the most similar to i in the group $A = a_0$ and k similar individuals from the group $A = a_1$. The first set is denoted as S^+ while the second as $\mathbf{S}^-$. The target individual is considered as discriminated if the difference observed between the rate of positive decisions in $\mathbf{S}^-$ and $\mathbf{S}^+$ is higher than a predefined threshold τ (typically 5%).

Causal inference is used to define the distance function $d(i, i')$ required to select the elements of $\mathbf{S}^-$ and $\mathbf{S}^+$. First, only attributes that are direct causes of the outcome should considered in the computation of the distance. That is, based on the causal graph, $\mathbf{Q} = Pa(Y) \setminus \{A\}$ denotes the set of variables that should be used in the distance function. Second, the causal effect of each of the selected attributes ($Q_k \in \mathbf{Q}$) on the the outcome should be considered in the function definition. In particular, for each variable Q_k, $CE(q_k, q'_k)$ measures the causal effect on the outcome when the value of Q_k changes from q_k to q'_k and is defined as:

$$CE(q_k, q'_k) = P(y_{\mathbf{q}}) - P(y_{q'_k,\mathbf{q}\setminus\{q_k\}}) \tag{22}$$

where $(P(y_{\mathbf{q}}))$ is the effect of the intervention that forces the set $\mathbf{Q}$ to take the set of values $\mathbf{q}$, and $(P(y_{q'_k,\mathbf{q}\setminus\{q_k\}}))$ is the effect of the intervention that forces Q_k to take value q'_k and other attributes in $\mathbf{Q}$ to take the same values as $\mathbf{q}$.

4.8 Non-Discrimination Criterion

Non-discrimination criterion [Zhang et al. 2017a] is a group fairness notion that aims to discover and to quantify direct discrimination. This notion is based on the identification of meaningful partitions of attributes that can be used to provide quantitative assessment for discrimination. A partition includes a subset of non-protected attributes (called *block set*) which blocks the causal effect from A to Y while a group is specified by a value assignment to these attributes. Given a block set $\mathbf{Q}$, the discrimination between protected and unprotected groups is assessed by computing the risk difference [Romei and Ruggieri 2011]:

$$\mid \ \Delta P|_{\mathbf{q}} \mid = \mid P(y \mid a_1, \mathbf{q}) - Pr(y \mid a_0, \mathbf{q}) \mid \tag{23}$$

where $\mathbf{q}$ is a value assignment for the block set $\mathbf{Q}$ and the absolute value to consider both positive and negative discriminations. If the risk difference is less than τ for all combination of values of all block sets, no direct discrimination is reported. Eq. 23 holds for each value assignment q of each block set $\mathbf{Q}$ ($\mathbf{Q} = Par(Y) \setminus \{A\}$). This notion is similar to the individual direct discrimination except that instead of using the set $\mathbf{Q}$ as an input to measure the similarity between individuals, it is used to define meaningful partitions to assess discrimination among groups.

4.9 PC-Fairness

Path-Specific Counterfactual Fairness has been first introduced by Chiappa [Chiappa 2019]. Wu et al. [Wu et al. 2019b] used the same term to denote a general fairness formalization that covers various causal-based fairness notions. That is, by differently tuning its parameters, it is possible to match several of the causal-based fairness notions mentioned above. Given a factual condition $\mathbf{O} = \mathbf{o}$ where $\mathbf{O} \subseteq \{A, \mathbf{V}, Y\}$ and a causal path set π, an outcome Y achieves the PC-fairness iff:

$$PCE^{\pi}_{a_1,a_0}(y \mid \mathbf{o}) = 0 \tag{24}$$

where $PCE^{\pi}_{a_1,a_0}(y \mid \mathbf{o}) = P(y_{a_1|\pi,a_0|\overline{\pi}} \mid \mathbf{o}) - P(y_{a_0} \mid \mathbf{o})$

Intuitively, $PCE^{\pi}_{a_1,a_0}(y \mid \mathbf{o})$ represents the path-specific counterfactual effect of the value change of A from a_0 to a_1 on Y through the specific causal path set π (with reference a_0) and given the factual observation $\mathbf{o}$.

Most of the aforementioned causal-based fairness notions can be expressed as special cases of PC-Fairness. For instance, if the set π includes all possible paths and $\mathbf{O} = \mathbf{V} \setminus \{Y\}$, PC-fairness corresponds to counterfactual fairness (Eq. (10)). If the set π includes all possible paths and $\mathbf{O} = \emptyset$, PC-fairness corresponds to total effect (Eq. 2). Chiappa and Isaac discuss some of these notions along with PC-fairness in [Chiappa and Isaac 2018] and [Chiappa et al. 2020].

4.10 Equality of Effort

Equality of effort [Huan et al. 2020] fairness notion identifies discrimination by assessing how much effort is needed by the disadvantaged individual/group to reach a certain level of outcome. A treatment variable T (considered as a legitimate variable) is selected and used to address the question: "to what extend this treatment variable T should change to make the individual (or a group of individuals) achieve a certain outcome level?". Hence, this notion focuses on whether the effort to reach a certain outcome level is the same for the protected and unprotected groups. Considering the simple job hiring example, the education level E is a good choice for the treatment variable. Two equality of effort notions are defined based on the potential outcome framework [Imbens and Rubin 2015], individual γ-Equal effort and system γ-Equal effort. Let $Y_i^{(t)}$ be the potential outcome for individual i had T been t and $\mathbf{E}[Y_i^{(t)}]$ be the expected outcome for individual i. As $Y_i^{(t)}$ is not observable, situation testing is used to estimate it in a similar way as individual direct discrimination (Section 4.7). Let $\mathbf{S}^+$ and $\mathbf{S}^-$ be the two sets of similar individuals with $A = a_0$ and $A = a_1$, respectively, and $\mathbf{E}[Y_{\mathbf{S}^+}^{(t)}]$ be the expected outcome under treatment t for the subgroup $\mathbf{S}^+$. The minimal effort needed to achieve γ-level of outcome variable within the subgroup $\mathbf{S}^+$ is defined as:

$$\Psi_{\mathbf{S}^+}(\gamma) = \underset{t \in T}{argmin} \, \{\mathbf{E}[Y_{\mathbf{S}^+}^{(t)}] \geq \gamma\} \tag{25}$$

Individual γ-Equal effort is satisfied for individual i if:

$$\Psi_{\mathbf{S}^+}(\gamma) = \Psi_{\mathbf{S}^-}(\gamma) \tag{26}$$

System γ-Equal effort is satisfied for a sub-population (e.g. $A = a_1$) if:

$$\Psi_{\mathbf{D}^+}(\gamma) = \Psi_{\mathbf{D}^-}(\gamma) \tag{27}$$

where $\mathbf{D}^+$ and $\mathbf{D}^-$ are the subsets of the entire dataset with sensitive attributes a_0 and a_1, respectively.

4.11 Interventional and justifiable fairness

Interventional fairness [Salimi et al. 2019] is a group-level fairness that can be seen as a strong version of total effect (Eq. 2). Instead of intervening only on the sensitive attribute A, interventional fairness intervenes on all remaining variables. Let $\mathbf{K}$ be a subset of $\mathbf{V}$ excluding A and Y, that is, $\mathbf{K} \subseteq \mathbf{V} - \{A, Y\}$. A predicting algorithm is $\mathbf{K}$-fair if for any assignment of values $\mathbf{K} = \mathbf{k}$ and outcome $Y = y$:

$$P(y_{a_0,\mathbf{k}}) = P(y_{a_0,\mathbf{k}}) \tag{28}$$

A predicting algorithm is interventionally fair if it is $\mathbf{K}$-fair for every set of variables $\mathbf{K}$. Jusifitiable fairness is a relaxation of interventional fairness achieved by classifying the variables as admissible (denoted as $\mathbf{E}$) or inadmissible (denoted as $\mathbf{R}$) which correspond, respectively, to explainable and proxy/redlining variables as defined in the beginning of Section 4. A predicting algorithm is justifiably fair if it is $\mathbf{K}$-fair with respect to only supersets of $\mathbf{E}$, that is, $\mathbf{K} \supseteq \mathbf{E}$.

Graphically, if all directed paths from the sensitive attribute A to the outcome Y go through an admissible attribute in $\mathbf{E}$, then the algorithm is justifiably fair. Notice that in case $\mathbf{E} = \emptyset$, justifiable fairness coincides with interventional fairness.

4.12 Individual equalized counterfactual odds

Individual equalized counterfactual odds [Pfohl et al. 2019] has been primarily proposed to assess fairness in the context of clinical decision making. This notion is an extended version of counterfactual fairness (Section 4.4) and equalized odds [Hardt et al. 2016]. A predictor satisfies this notion iff:

$$P(\hat{y}_{a_1} \mid \mathbf{X} = \mathbf{x}, y_{a_1}, A = a_0) = P(\hat{y}_{a_0} \mid \mathbf{X} = \mathbf{x}, y_{a_0}, A = a_0) \tag{29}$$

In other words, Eq. 29 implies that the predictor must be counterfactually fair given the outcome Y matching the counterfactual outcome y_{a_0}. Thus, in addition to requiring predictions to be the same across factual/counterfactual samples (counterfactual fairness desiderata), those samples must share the same value of the actual outcome Y (equalized odds desiderata) as well.

5 Identifiability

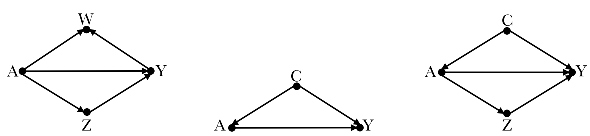
The identifiability of causal quantities has been extensively studied in the literature: causal effect (intervention) identifiability [Galles and Pearl 1995; Huang and Valtorta 2006; Pearl 2009; Shpitser and Pearl 2006, 2008; Tian 2004; Tian and Pearl 2002; Tian and Shpitser 2003], counterfactual identifiability [Shpitser 2013; Shpitser and Pearl 2007, 2008; Wu et al. 2019a], direct/indirect effects [Pearl 2001] and path-specific effect identifiability [Avin et al. 2005; Malinsky et al. 2019; Shpitser 2013; Zhang and Wu 2017; Zhang et al. 2017b]. This section summarizes the main identifiability conditions as they relate to the specific problem of discrimination discovery.

5.1 Identifiability of causal effect (intervention)

The causal effect of a cause variable X on an effect variable Y is computed using $P(Y_x) = P(Y|do(X = x))$, the distribution of Y after the intervention $X = x$. In discrimination setup, the cause is typically the sensitive attribute A. A basic case where identifiability can be avoided altogether is when it is possible to perform experiments by intervening on the sensitive attribute A. When this is possible, randomized controlled trial (RCT) [Fisher 1992] can be used to estimate the causal effect. RCT consists in randomly assigning subjects (e.g. individuals) to treatments (e.g. gender), then comparing the outcome Y of all treatment groups. However, in the context of machine learning fairness, RCT is often not an option as experiments can be too costly to implement or physically impossible to carry out (e.g. changing the gender of a job applicant).

In Markovian models (no unobserved confounding), the causal effect is always identifiable (Corollary 3.2.6 in [Pearl 2009]). The simplest case is when there is no confounding between A and Y (Figure 8(a)). In that case, the causal effect matches the conditional probability regardless of any mediator:

$$P(y_a) = P(y|do(a)) = P(y|a) \tag{30}$$



(a) Simple Markovian model with a collider (W) and a mediator (Z).

(b) Simple Markovian model with a confounder (C).

(c) Simple Markovian model with a confounder (C) and a mediator (Z).

Figure 8. Simple causal graphs

In presence of an observable confounder (Figure 8(b)), $P(y_a)$ is identifiable by adjusting on the confounder:

$$P(y_a) = \sum_C P(y|a,c)\ P(c) \tag{31}$$

where the summation is on values c in the domain (sample space) of C denoted as $dom(C)$. Eq. 31 is called the back-door formula[7]. In Makovian models, the causal effect $P(y_a)$ can also be computed using the truncation formula:

$$P(y_a) = \sum_{\substack{\mathbf{V}\setminus\{A,Y\}\\ Y=y}} \prod_{V\in\mathbf{V}\setminus\{A\}} P(v|\mathbf{Pa}_V) \tag{32}$$

where $\mathbf{Pa}_V$ denotes the parent variables of V.

Using the back-door formula (31) or the truncated formula (32) on Figure 8(c) produces the same result:

$$\begin{aligned} P(y_a) &= \sum_C \sum_Z P(y|a,z,c)\ P(z|a)\ P(c) \\ &= \sum_C P(y|a,c)\ P(c) \end{aligned}$$

while the joint probability

$$P(y,a,c,z) = P(y|a,c,z)\ P(z|a)\ P(a|c)\ P(c)$$

and the conditional probability

$$P(y|a) = \sum_C \sum_Z P(y|a,c,z)\ P(c,z|a)$$

.

For semi-Markovian models, identifiability of $P(y_a)$ is not guaranteed. In case it is identifiable, Pearl [Pearl 2009] proposes a do-calculus composed of three rules allowing to turn interventional probabilities to observational ones:

1. $P(y_a|\mathbf{z},w) = P(y_a|\mathbf{z})$ provided that the set of variables $\mathbf{Z}$ blocks all backdoor paths from W to Y after all arrows leading to A have been deleted.
2. $P(y_a|\mathbf{z}) = P(y|a,\mathbf{z})$ provided that the set of variables $\mathbf{Z}$ blocks all backdoor paths from A to Y.
3. $P(y_a) = P(y)$ provided that there are no causal paths between A and Y.

do-calculus has been proven to be sound and complete in the identification of interventional distributions [Huang and Valtorta 2006]. For example, $P(y_a)$ is identifiable in Figure 9(d). By applying the chain rule following the topological order: $W_2 < A < W_1 < W_3 < Y$, we get:

$$\begin{aligned} P(y_a) = &\sum_{w_1 w_2 w_3} P(y|do(a),w_1,w_2,w_3)\ P(w_1|do(a),w_2)\ P(w_2) \\ &\qquad \times P(w_3|w_2,w_1,do(a)) \qquad (33)\\ = &\sum_{w_1 w_2} P(y|do(a),w_1,w_2)\ P(w_1|do(a),w_2)\ P(w_2) \qquad (34)\\ = &\sum_{w_1 w_2} P(y|do(a),w_2)\ P(w_1|a,w_2)\ P(w_2) \qquad (35)\\ = &\sum_{w_1 w_2}\sum_{a'} P(y|a',w_2,do(w_1))\ P(a'|do(w_1),w_2) \\ &\qquad \times P(w_1|a,w_2)\ P(w_2) \qquad (36)\\ = &\sum_{w'_1}\sum_{w'_2}\sum_{a'} P(y|w'_1,w'_2,a')\ P(a'|w'_2)\ P(w'_1|w'_2,a)\ P(w'_2) \\ &\qquad (37) \end{aligned}$$

Note that w_3 is omitted from (34) since it is considered latent [Tikka and Karvanen 2017a]. Applying Rule 2 followed by Rule 3 to the first term in (34) yields to $P(y|do(a),w_2)$ (35). Likewise, applying Rule 2 to the second term in (34) leads to $P(w_1|a,w_2)$. Thus, the original problem reduces to identifying the term $P(y|do(a),w_2)$ in (35). Here we cannot apply Rule 2 to exchange $do(a)$ with a because $G_{\underline{A}}$ (graph obtained by removing all emanating arrows from A) contains a backdoor path from A to Y. Thus, to block that path, we need to condition and to sum over all values of A as shown in Eq. (36) ($\sum_{a'}\ P(y|a',w_2,do(w_1))\ P(a'|do(w_1),w_2)$. Now, applying Rule 2 to $P(y|a',w_2,do(w_1))$ and Rule 3 to $P(a'|do(w_1),w_2)$ and adding the other terms results in the final expression in (37).

In [Pearl 2009] Section 3.5, Pearl gives other examples of using the do-calculus for the identifiability of semi-Markovian models.

The problem of do-calculus is the difficulty to determine the correct order of application of the rules. Using the wrong

[7] Called also adjustment formula or stratification.

order may hinder the identifiability or produce a very complex expression [Tikka and Karvanen 2017b]. To address this issue, several contributions in the identifiability literature focused on defining graphical patterns and mapping them to simple and concise intervention-free expressions [Shpitser and Pearl 2006; Tian 2004; Tian and Pearl 2002; Tian and Shpitser 2003].

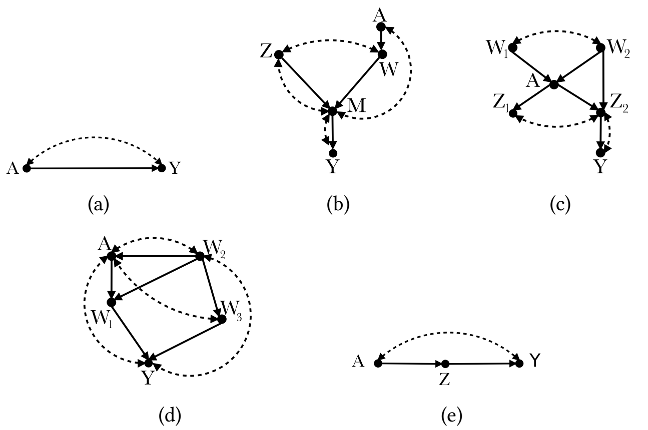


Figure 9. Figure 9(a) presents the“bow" graph, Figure 9(b) illustrates the structure of a c-tree, Figure 9(c) shows a semi-Markovian model where $P(y_a)$ is observable, Figure 9(d) presents a semi-Markovian model where $P(y_a)$ is identifiable and Figure 9(e) illustrates a simple example of the front-door criterion.

The simplest case where $P(y_a)$ is identifiable in a semi-Markovian model is when the sensitive attribute A is not involved in any confounding. That is, there is no bi-directional edge connected to A. This matches Theorem 3.2.5 in [Pearl 2009] which states that if all parents of a cause variable A are observable, the causal effect of that variable is identifiable. Figure 9(c) shows a graph diagram of a semi-Markovian model where $P(y_a)$ is identifiable. If such criterion is satisfied, the causal effect can be computed as follows [Tian and Pearl 2002]:

$$P(y_a) = \sum_{\mathbf{pa}_a} P(y|a, \mathbf{pa}_a)\ P(\mathbf{pa}_a) \tag{38}$$

where $\mathbf{pa}_a$ is the set of values of the parents of A. For instance, applying Eq. 38 to the graphs shown in Figures 8(b) and 8(c) yields to the same result obtained in Eq. 31.

A more complex criterion of identifiability of $P(y_a)$ is when the sensitive attribute A is involved in confounding, but no child of A is involved in confounding. Figure 9(d) shows a graph satisfying this criterion. Tian and Pearl [Tian and Pearl 2002] show that the causal effect of A on all other variables $(\mathbf{V}\backslash\{A\})$ denoted as $P(\mathbf{v}_a) = P_a(\mathbf{v})$ is identifiable and is given by Theorem 2 [Tian and Pearl 2002]:

$$P_a(\mathbf{v}) = \left(\prod_{i|V_i\in \mathbf{ch_A}} P(v_i|\mathbf{pa}_i)\right) \sum_{a'\in dom(A)} \frac{P(v)}{\prod_{i|V_i\in \mathbf{ch_A}} P(v_i|\mathbf{pa}_i)} \tag{39}$$

where ch_A is the set of the children of the node A while $\mathbf{pa_i}$ is the set of values of the parents of the variable V_i. This result cannot be used since the aim is to assess the effect of the sensitive attribute A on a single variable (the outcome Y) and hence Eq. 39 should be adapted. To illustrate that, consider the example of the causal graph in Figure 9(d). Applying Eq. 39 to that graph leads to:

$$P_a(w_1, w_2, w_3, y) = P(w_1|\ a, w_2) \sum_{a'\in dom(A)} \frac{P(w_1, w_2, w_3, y)}{P(w_1|\ a', w_2)}) \tag{40}$$

$$= P(w_1|\ a, w_2) \sum_{a'\in dom(A)} P(y, w_3|\ a', w_1, w_2) P(a', w_2) \tag{41}$$

where Eq. 41 is obtained by applying the Bayes' rule. Adapting Eq. 39 in order to compute $P(y_a)$ (causal effect on the single variable Y) requires summing over the possible values of variables W_1, W_2 and W_3 as follows. Starting from Eq. 41, summing over W_1 gives:

$$P_a(w_2, w_3, y) = \sum_{w_1'\in dom(W_1)} P(w_1'|\ a, w_2) \sum_{a'\in dom(A)} P(y, w_3|\ a', w_1', w_2) \times P(a', w_2)$$

Similarly, summing over W_2 and W_3 (and omitting $dom()$ for conciseness), leads to:

$$P(y_a) = \sum_{w_1'} \sum_{w_2'} \sum_{a'} P(y|w_1', w_2', a')\ P(a'|w_2')\ P(w_1'|w_2', a)\ P(w_2') \tag{42}$$

Note that w_3 is omitted from Eq. 42 for the same reason mentioned above.

All the above criteria can be generalized to the case where the sensitive attribute is not connected to any of its children through a confounding path. In such case, c-component factorization can be used. A c-component is a set of vertices in the graph such that every pair of vertices are connected by a confounding edge. The idea of c-component factorization is to decompose the identification problem into smaller sub-problems, that is, a disjoint set of c-components in order to calculate $P(y_a)$. For example, in Figure 9(c), there are three c-components: $\{\{W1, W2\}, \{A\}, \{Z1, Z2, Y\}\}$. Hence, as long as there is no confounding path connecting A to any of its direct children, $P(y_a)$ is identifiable. C-component factorization is used in the ID algorithm [Shpitser and Pearl 2008] which is proven to be complete for causal effect identification. However, as the do-calculus approach, the algorithm produces typically long and complex expressions [Tikka 2018].

In case there is an unobservable confounding between the sensitive attribute A and the outcome Y, all the above criteria will fail. However, $P(y_a)$ can still be identifiable using the front-door criterion. This criterion is satisfied in Figure 9(e) and consists in having a mediator variable Z such that:

- there are no backdoor paths from A to Z
- all backdoor paths from Z to Y are blocked by A.

A backdoor path from A to Z is any path starting at A with a backward edge $\leftarrow$ into A (e.g. $A \leftarrow \ldots Z$). If such criterion is satisfied, $P(y_a)$ can be computed as follows:

$$\begin{aligned} P(y_a) &= \sum_Z P(y|do(z))\ P(z|do(a)) \\ &= \sum_Z P(y|z,a)\ P(a)\ P(z|a) \end{aligned} \tag{43}$$

Shpitser and Pearl proved that all the unidentifiable cases of the causal effect $P(y_a)$ boil down to a general graphical structure called the hedge criterion. Based on this criterion, they designed a complete identifiability algorithm called *ID* which outputs the expression of $P(y_a)$ if it is identifiable, or the reason of the unidentifiability, otherwise.

The simplest graph in which the causal effect between A and Y is not identifiable is the"bow" graph (Figure 9(a)). This simple unidentifiability criterion can be generalized to a more complex graphs called c-tree. A c-tree is a graph that is at the same time a tree[8] and a c-component. Figure 9(b) shows an example of a c-tree. If the causal graph is a c-tree rooted in the outcome variable Y, $P(y_a)$ is unidentifiable [Shpitser and Pearl 2008].

5.2 Identifiability of counterfactuals

Most of causality-based fairness notions listed in Section 4 are defined in terms of counterfactual quantities. Hence, the applicability of those notions depends heavily on the identifiability of the counterfactuals composing them. If a fairness notion is relying on a counterfactual quantity which is proven to be unidentifiable in the causal model at hand, the notion cannot be used to compute the level of discrimination[9]. In Markovian, as well as semi-Markovian models, if all parameters of the causal model are known (including $P(\mathbf{u})$), any counterfactual is identifiable and can be computed using the three steps abduction, action, and prediction (Theorem 7.1.7 in [Pearl 2009]). Let $P_* = \{P_{\mathbf{x}} | \mathbf{X} \subseteq \mathbf{V},\ \mathbf{x}\ \textit{a value assignment of}\ \mathbf{X}\}$ be the set of all interventional distributions in a given causal model. While the identifiability of interventional proabilities $P(y_a)$ is characterized based on observational probabilities $P(\mathbf{v})$, in this section, the identifiability of counterfactuals is characterized in terms of interventional probabilities P_*. Then, combining these results with the criteria of the previous section, a counterfactual can, in turn, be identified using observational probabilities $P(\mathbf{v})$.

[8]Notice that, in this paper, the direction of the arrows between nodes is reversed compared to the usual tree structure.

[9]But it can be used to find bounds on the actual value.

Given a causal graph G of a Markovian model and a counterfactual expression $\gamma = v_x|e$ with e some arbitrary set of evidence, identifying and computing $P(\gamma)$ requires to construct a counterfactual graph which combines parallel worlds. Every world is represented by a model M_x corresponding to each subscript in the counterfactual expression. For example, given the causal graph in Figure 1 and the counterfactual expression $y_{a_1}|a_0$, the resulting counterfactual graph is shown in Figure 10(d). The counterfactual graph should be "reduced" by merging together vertices that share the same causal mechanism (**make-cg** algorithm in [Shpitser and Pearl 2008] automates this procedure). The resulting counterfactual graph can be considered as a typical causal graph for a larger causal model. Consequently, all the graphical criteria listed in Section 5.1 for the identifiability of causal effects apply on the counterfactual graph to identify counterfactual quantities, in particular, the c-component factorization of the counterfactual graph [Shpitser and Pearl 2007]. ID^* and IDC^* algorithms [Shpitser and Pearl 2008] automate the identifiability and computation of counterfactuals based on all the above criteria.

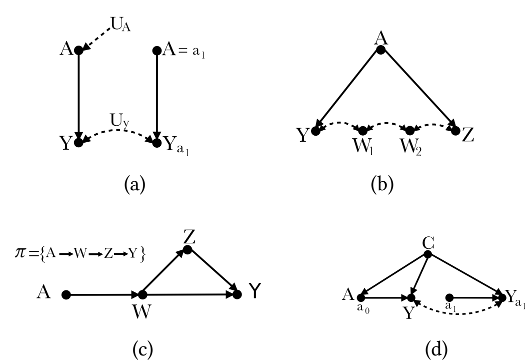


Figure 10. Causal graphs.

The simplest unidentifiable counterfactual quantity is $P(y_{a'}, y'_a)$ which is called the probability of necessity and sufficiency. The corresponding counterfactual graph is the W-graph that has the same structure as to Figure 10(a). This simple criterion can be generalized to the zig-zag graph (Figure 10(b)) where the counterfactual $P(y_a, w_1, w_2, z')$ is not identifiable.

Pearl [Pearl 2009] proves two results about the identifiability of counterfactuals. First, for linear causal models (i.e. the functions $\mathbf{F}$ are linear), any counterfactual is experimentally (using P_*) identifiable whenever the model parameters are identified. Second, in linear causal models, if some of the model parameters are unknown, any counterfactual of the form $\mathbf{E}(Y_a|e)$ where e is some arbitrary set of evidence, is identifiable provided that $\mathbf{E}(y_a)$ is identifiable. Finally, there is no single necessary and sufficient criterion for the identifiability of counterfactuals in semi-Markovian models [Avin et al. 2005].

To illustrate the computation of a counterfactual probability, consider the teacher firing example of Figure 1 and the counterfactual probability $P(y_{a_1}|a_0)$ which reads the probability of firing a teacher who is assigned a class with a high initial level of students (a_0) had she been assigned a class with a low initial level of students (a_1). Applying **make-cg** algorithm based on this counterfactual quantity produces the counterfactual graph in Figure 10(d) which combines two worlds: the actual world where the teacher has normally $A = a_0$ and the counterfactual world where *the same* teacher is assigned $A = a_1$. Both variables C are reduced to a single variable and Y and Y_{a_1} are connected by an unobservable confounder. The counterfactual graph is composed of three c-components $\{C\}, \{A\}, \{Y, Y_{a_1}\}$. Applying algorithm IDC^* [Shpitser and Pearl 2008] results in:

$$P(y_{a_1}|a_0) = \frac{\sum_{y,c} \; Q(c)\; Q(a_0)\; Q(y, y_{a_1})}{P(a_0)} \tag{44}$$

where $Q(\mathbf{v}) = P(\mathbf{v}|pa(\mathbf{V}))$ in the counterfactual graph. Hence,

$$\begin{aligned}
P(y_{a_1}|a_0) &= \frac{\sum_{y,c} \; P(c)\; P(a_0|c)\; P(y, y_{a_1}|c)}{P(a_0)} \\
&= \frac{\sum_{c} \; P(c)\; P(a_0|c)\; P(y_{a_1}|c)}{P(a_0)} && (45)\\
&= \frac{\sum_{c} \; P(c)\; P(a_0|c)\; P(y|a_1, c)}{P(a_0)} && (46)\\
&= \frac{0.5 \times 0.8 \times 0.25 + 0.5 \times 0.2 \times 0.01}{0.5} \\
&= 0.202
\end{aligned}$$

y in Eq. (45) is cancelled by summation while $P(y_{a_1}|c)$ in the same equation is transformed into $P(y|a_1, c)$ in Eq. (46) using Rule 2 of the *do*-calculus.

5.3 Identifiability of direct and indirect effects

In Markovian models, the average natural direct effect NDE and the average natural indirect effect NIE are always identifiable (from observational data) and can be computed as follows [Pearl 2001]:

$$NDE_{a_1,a_0}(Y) = \sum_{\mathbf{s}} \sum_{\mathbf{z}} \Big(\mathbf{E}[Y|a_1, \mathbf{z}] - \mathbf{E}[Y|a_0, \mathbf{z}]\Big) P(\mathbf{z}|a_0, \mathbf{s}) P(\mathbf{s}) \tag{47}$$

$$NIE_{a_1,a_0}(Y) = \sum_{\mathbf{s}} \sum_{\mathbf{z}} \mathbf{E}[Y|a_0, \mathbf{z}] \; \Big(P(\mathbf{z}|a_1, \mathbf{s}) - P(\mathbf{z}|a_0, \mathbf{s})\Big) \; P(\mathbf{s}) \tag{48}$$

where $\mathbf{Z}$ is a set of mediator variables and $\mathbf{S}$ is any set of variables satisfying the back-door criterion between the sensitive variable A and the mediator variables $\mathbf{Z}$, that is, *(i)* no variable in $\mathbf{S}$ is a descendant of A and *(ii)* $\mathbf{S}$ blocks all back-door paths between A and $\mathbf{Z}$. A simpler formulation can be used in case there is no confounding between A and $\mathbf{Z}$, where the need for $\mathbf{S}$ is dropped altogether:

$$NDE_{a_1,a_0}(Y) = \sum_{\mathbf{z}} \Big(\mathbf{E}[Y|a_1, \mathbf{z}] - \mathbf{E}[Y|a_0, \mathbf{z}]\Big) \; P(\mathbf{z}|a_0) \tag{49}$$

$$NIE_{a_1,a_0}(Y) = \sum_{\mathbf{z}} \mathbf{E}[Y|a_0, \mathbf{z}] \; \Big(P(\mathbf{z}|a_1) - P(\mathbf{z}|a_0)\Big) \tag{50}$$

In semi-Markovian models, NDE and NIE are not generally identifiable, even if we have the luxury to perform any experiment using RCT, because of the nested counterfactuals $P(Y_{a_1}, \mathbf{Z}_{a_0})$ and $P(Y_{a_0}, \mathbf{Z}_{a_1})$ in Eq. 4 and Eq. 5, respectively. Nevertheless, these quantities are identifiable *from experimental data* provided that there is a set of variables $\mathbf{W}$ which are parents of the outcome variable Y but non-descendents of A and Z such that $Y_{a_0,z} \perp\!\!\!\perp Z_{a_0}|\mathbf{W}$ (reads: $Y_{a_0,z}$ and Z_{a_0} are independent conditional of $\mathbf{W}$). This condition can be easily checked from the causal graph as follows: $\mathbf{W}$ d-separates Y and $\mathbf{Z}$ in the graph formed by deleting all arrows emanating from A and $\mathbf{Z}$, denoted simply as $(Y \perp\!\!\!\perp \mathbf{Z}|\mathbf{W})_{G_{\underline{AZ}}}$.

If such graphical condition is satisfied, NDE and NIE can be computed from experimental quantities as follows:

$$NDE_{a_1,a_0}(Y) = \sum_{\mathbf{z},\mathbf{w}} \Big(\mathbf{E}[Y_{a_1,\mathbf{z}}|\mathbf{w}] - \mathbf{E}[Y_{a_0,\mathbf{z}}|\mathbf{w}]\Big) \; P(\mathbf{Z}_{a_0} = \mathbf{z}|\mathbf{w}) \; P(\mathbf{w}) \tag{51}$$

$$NIE_{a_1,a_0}(Y) = \sum_{\mathbf{z},\mathbf{w}} \mathbf{E}[Y_{a_0,\mathbf{z}}|\mathbf{w}] \Big(P(\mathbf{Z}_{a_1} = \mathbf{z}|\mathbf{w}) - P(\mathbf{Z}_{a_0} = \mathbf{z}|\mathbf{w})\Big) P(\mathbf{w}) \tag{52}$$

5.4 Identifiability of path-specific effects

The identifiability of $PSE_{\pi}(a_1, a_0)$ in Markovian models depends on whether $P(y|do(a_1|_{\pi}, a_0|_{\bar{\pi}}))$ is identifiable. Avin et al. [Avin et al. 2005] gave a single necessary and sufficient criterion for the identifiability of $P(y|do(a_1|_{\pi}, a_0|_{\bar{\pi}}))$ in Markovian models called recanting witness criterion. This criterion holds when there is a vertex W along the causal path π that is connected to Y through another causal path not in π. For instance, Figure 10(c) satisfies the recanting witness criterion when $\pi = A \rightarrow W \rightarrow Z \rightarrow Y$ with W as witness. The corresponding graph structure is called "kite" graph. When this criterion is satisfied, $P(y|do(a_1|_{\pi}, a_0|_{\bar{\pi}}))$ is not identifiable, and consequently, $PSE_{\pi}(a_1, a_0)$ is not identifiable. Shpitser [Shpitser 2013] generalizes this criterion to semi-Markovian models known as recanting district criterion.

6 Applicability

In his book, *The Book of Why* [Pearl and Mackenzie 2018], Pearl describes a causation ladder with three rungs: statistical observations (seeing), intervention (doing), and counterfactual (imagining). In this section, all causal-based fairness notions (Section 4) are placed in the causation ladder which will help us address the problem of their applicability in real-scenarios. The causation ladder is structured in such a way that quantity at a certain rung can be identified in terms of quantities at the rung just below it. As a consequence, the higher the rung, the more challenging the problem of identifiability, and hence the less applicable a fairness notion defined at that rung.

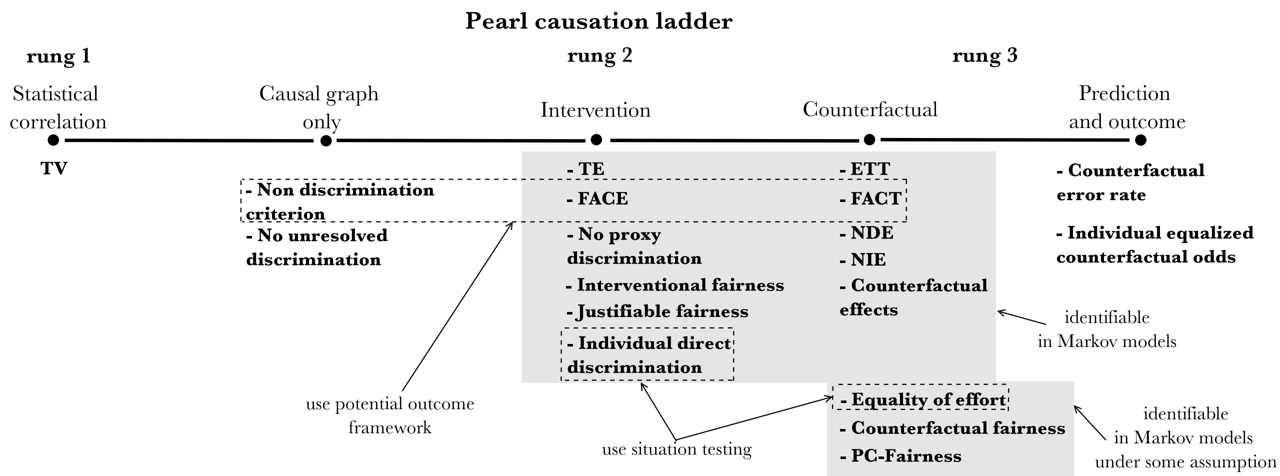


Figure 11. Classification of causal-based fairness notions according to Pearl causation ladder [Pearl et al. 2016]

.

The diagram in Figure 11 shows the causation ladder and indicates at which rung every causal-based fairness notion stands. TV which is the only non-causal fairness notion covered in this paper is at rung 1. It is always applicable provided that a set of observations (dataset) is available. No unresolved and non-discrimination criteria are placed midway between rungs 1 and 2 as they are applicable provided that the causal graph is available along the dataset. Non-discrimination criterion, however, requires the Markov property to be applicable because causal dependence through unobservable paths cannot be blocked. It also has an exponential complexity since it considers all combination of values of the parent variables of the outcome Y. A relaxation is described by the authors [Zhang et al. 2017a] but the notion remains computationally intractable.

Fairness notions at rung 2 (TE, $FACE$, No-proxy discrimination, interventional and justifiable fairness, and individual direct discrimination) are applicable in any scenario where either experiments (RCT) are possible or hypothetical interventions are identifiable. As mentioned in Section 5.1, in Markovian models any intervention probability is identifiable from observational data. Hence, these fairness notions are always applicable in Markovian models, except for individual direct discrimination which assumes in addition that the sensitive attribute A has no parent [Zhang et al. 2016]. In semi-Markovian models, the applicability of these rung 2 notions depends on the identifiability of the intervention terms used in their respective definitions. For instance, for individual direct discrimination, the term in question is $CE(q_k, q'_k)$ in Eq. 22. For the particular case of FACE notion, even if the intervention quantity $\mathbf{E}[Y_i^{(a_1)}]$ in Eq. 8 is not identifiable, it can be still estimated empirically provided the assumptions in [Khademi et al. 2019] in Section 3 hold.

The bulk of causal-based fairness notions are defined in terms of counterfactual quantities and hence are placed in rung 3 of the causation ladder. In Figure 11, the counterfactual notions are ranked from top to bottom according to their degree of applicability. For instance, counterfactual effects are placed on top of counterfactual fairness to indicate that the former is applicable in more scenarios than the latter. In Markovian models, the top 5 notions (ETT, FACT, NDE, NIE, and counterfactual effects) are always identifiable and hence applicable. That is, specific formula are already available to compute each counterfactual term used in their definitions.

For instance, given a Markovian model, the three counterfactual effects (Section 4.5) can be computed from observational data as follows:

$$DE_{a_1,a_0}(y|a) = \sum_{\mathbf{Z},\mathbf{W}}(P(y \mid a_1, \mathbf{z}, \mathbf{w}) - P(y \mid a_0, \mathbf{z}, \mathbf{w}))P(\mathbf{z} \mid a_0, \mathbf{w}) \times P(\mathbf{w} \mid a)$$

$$IE_{a_1,a_0}(y|a) = \sum_{\mathbf{Z},\mathbf{W}} P(y \mid a_0, \mathbf{z}, \mathbf{w})P(\mathbf{z} \mid a_1, \mathbf{w}) - P(\mathbf{z} \mid a_0, \mathbf{w})) \times P(\mathbf{w} \mid a)$$

$$SE_{a_1,a_0}(y) = \sum_{\mathbf{Z},\mathbf{W}} P(y \mid a_0, \mathbf{z}, \mathbf{w})P(\mathbf{z} \mid a_0, \mathbf{w})(P(\mathbf{w} \mid a_1) - P(\mathbf{w} \mid a_0)$$

where $\mathbf{Z}$ and $\mathbf{W}$ are sets of mediator and confounder variables, respectively.

In Markovian models, the identifiability of counterfactual fairness and individual equality of effort [Huan et al. 2020] depends on the identifiability of the term $P(y_{a_1}|\mathbf{X} = \mathbf{x}, A = a_0)$ which is only identifiable if $\mathbf{X}$ does not contain any variable which is at the same time descendant of A and ancestor of Y, that is, $\mathbf{X} \cap \mathbf{B} = \emptyset$ where $\mathbf{B} = An(Y) \cap De(A)$[Wu et al. 2019a]. PC-Fairness is applicable provided that the model is

Markovian and the recanting witness criterion is not satisfied. In semi-Markovian models, unless all model parameters are known (including $P(\mathbf{u})$)[10], the identifiability of rung 3 fairness notions depends on the criteria discussed in Section 5.2, which rarely hold in practice.

Finally, counterfactual error rate and individual equalized counterfactual odds are special cases of rung 3 fairness notions as they are the only notions that condition on the true outcome Y to assess the fairness of the prediction $\hat{Y}$ (Eq. 18,19, 20, and 29). Such conditioning has an important implication on identifiability since Y is a collider, and conditioning on a collider creates a dependence between the previous variables [Pearl 2009]. This leads to unobservable confounding between the causes of Y. Hence, even if the causal model is Markovian, applying both notions turns it into a semi-Markovian model. Zhang and Bareinboim [Zhang and Bareinboim 2018a] define an identifiability criterion for counterfactual error rate in Markovian models called explanation criterion.

7 Conclusion

Applying causal-based fairness notions in practical scenarios is significantly hindered by the problem of identifiability. Based on an extensive body of work on identifiability theory, we summarized the most relevant identifiability criteria to the problem of discrimination discovery. These criteria are then used to characterize the situations where causal-based fairness notions can be used in practical scenarios.

The main objective of this paper is to bridge the gap between the practical scenarios of automated (and generally unintentional) discrimination discovery and the mostly theoretical tackling of the problem in the literature. This effort can be of particular interest to clinicians, civil right activists, anti-discrimination law enforcement agencies, and practitioners in fields where automated decision making systems are increasingly used.

8 Acknowledgements

The work of Catuscia Palamidessi was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme. Grant agreement № 835294.

References

Chen Avin, Ilya Shpitser, and Judea Pearl. 2005. Identifiability of path-specific effects. (2005).

Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

Marc Bendick. 2007. Situation testing for employment discrimination in the United States of America. *Horizons stratégiques* 3 (2007), 17–39.

Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7801–7808.

Silvia Chiappa and William S Isaac. 2018. A causal Bayesian networks viewpoint on fairness. In *IFIP International Summer School on Privacy and Identity Management*. Springer, 3–20.

Silvia Chiappa, Ray Jiang, Tom Stepleton, Aldo Pacchiano, Heinrich Jiang, and John Aslanides. 2020. A General Approach to Fairness with Optimal Transport.. In *AAAI*. 3633–3640.

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

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

Ronald Aylmer Fisher. 1992. Statistical methods for research workers. In *Breakthroughs in statistics*. Springer, 66–70.

David Galles and Judea Pearl. 1995. Testing identifiability of causal effects. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. 185–195.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

Wen Huan, Yongkai Wu, Lu Zhang, and Xintao Wu. 2020. Fairness through Equality of Effort. In *Companion Proceedings of the Web Conference 2020*. 743–751.

Yimin Huang and Marco Valtorta. 2006. Identifiability in causal bayesian networks: A sound and complete algorithm. In *Proceedings of the national conference on artificial intelligence*, Vol. 21. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 1149.

Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. 2019. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*. 2907–2914.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*. 656–666.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in neural information processing systems*. 4066–4076.

Daniel Malinsky, Ilya Shpitser, and Thomas Richardson. 2019. A potential outcomes calculus for identifying conditional path-specific effects. *Proceedings of machine learning research* 89 (2019), 3080.

Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, Vol. 2018. NIH Public Access, 1931.

Catherine O'Neill. 2016. Weapons of math destruction. *How Big Data Increases Inequality and Threatens Democracy* (2016).

Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. 411–420.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic Books.

Stephen R Pfohl, Tony Duan, Daisy Yi Ding, and Nigam H Shah. 2019. Counterfactual Reasoning for Fair Clinical Risk Prediction. In *Machine Learning for Healthcare Conference*. 325–358.

Kimberly Quick. 2015. The Unfair Effects of IMPACT on Teachers with the Toughest Jobs. *The Century Foundation* (2015).

Michelle Rhee. 2019. IMPACT: The DCPS Evaluation and Feedback System for School-Based Personnel.

James M Robins, Miguel Angel Hernán, and Babette Brumback. 2000. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology* 11, 5 (2000), 551.

[10]In that case, it is possible to use the three steps abduction, action, and prediction [Pearl 2009].

Andrea Romei and Salvatore Ruggieri. 2011. A multidisciplinary survey on discrimination analysis.

Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*. 793–810.

Ilya Shpitser. 2013. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive science* 37, 6 (2013), 1011–1035.

Ilya Shpitser and Judea Pearl. 2006. Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. 437–444.

Ilya Shpitser and Judea Pearl. 2007. What counterfactuals can be tested. In *23rd Conference on Uncertainty in Artificial Intelligence, UAI 2007*. 352–359.

Ilya Shpitser and Judea Pearl. 2008. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research* 9, Sep (2008), 1941–1979.

Elizabeth A Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* 25, 1 (2010), 1.

Jin Tian. 2004. Identifying linear causal effects. In *AAAI*. 104–111.

Jin Tian and Judea Pearl. 2002. A general identification condition for causal effects. In *Aaai/iaai*. 567–573.

Jin Tian and Ilya Shpitser. 2003. On the identification of causal effects. (2003).

Santtu Tikka. 2018. Improving identification algorithms in causal inference. *Report/University of Jyväskylä, Department of Mathematics and Statistics* 168 (2018).

Santtu Tikka and Juha Karvanen. 2017a. Enhancing identification of causal effects by pruning. *The Journal of Machine Learning Research* 18, 1 (2017), 7072–7094.

Santtu Tikka and Juha Karvanen. 2017b. Simplifying probabilistic expressions in causal inference. *The Journal of Machine Learning Research* 18, 1 (2017), 1203–1232.

Yongkai Wu, Lu Zhang, and Xintao Wu. 2019a. Counterfactual Fairness: Unidentification, Bound and Algorithm.. In *IJCAI*. 1438–1444.

Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. 2019b. Pc-fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*. 3404–3414.

Junzhe Zhang and Elias Bareinboim. 2018a. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*. 3671–3681.

Junzhe Zhang and Elias Bareinboim. 2018b. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Lu Zhang and Xintao Wu. 2017. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics* 4, 1 (2017), 1–16.

Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. Situation Testing-Based Discrimination Discovery: A Causal Inference Approach.. In *IJCAI*, Vol. 16. 2718–2724.

Lu Zhang, Yongkai Wu, and Xintao Wu. 2017a. Achieving non-discrimination in data release. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1335–1344.

Lu Zhang, Yongkai Wu, and Xintao Wu. 2017b. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 3929–3935.