Causal Multi-Level Fairness

Vishwali Mhasawade
vishwalim@nyu.edu
New York University

Rumi Chunara
rumi.chunara@nyu.edu
New York University

ABSTRACT

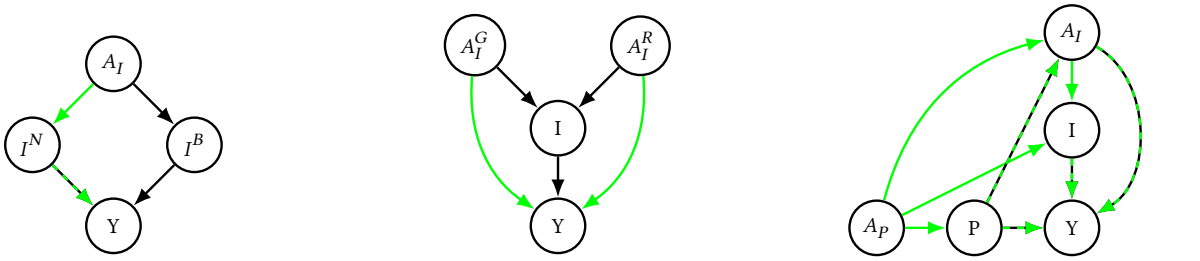
Algorithmic systems are known to impact marginalized groups severely, and more so, if all sources of bias are not considered. While work in algorithmic fairness to-date has primarily focused on addressing discrimination due to individually linked attributes, social science research elucidates how some properties we link to individuals can be conceptualized as having causes at population (e.g. structural/social) levels and it may be important to be fair to attributes at multiple levels. For example, instead of simply considering race as a protected attribute of an individual, it can be thought of as the perceived race of an individual which in turn may be affected by neighborhood-level factors. This multi-level conceptualization is relevant to questions of fairness, as it may not only be important to take into account if the individual belonged to another demographic group, but also if the individual received advantaged treatment at the population-level. In this paper, we formalize the problem of multi-level fairness using tools from causal inference in a manner that allows one to assess and account for effects of sensitive attributes at multiple levels. We show importance of the problem by illustrating residual unfairness if population-level sensitive attributes are not accounted for. Further, in the context of a real-world task of predicting income based on population and individual-level attributes, we demonstrate an approach for mitigating unfairness due to multi-level sensitive attributes.

1 INTRODUCTION

There has been much recent interest in designing algorithms that make fair predictions [2, 13, 27]. Definitions of fairness have been summarized in detail elsewhere [36, 50]; broadly approaches to fairness in machine learning can be divided into two kinds: group fairness, which ensures some form of statistical parity for members of different protected groups [15, 36] and individual notions of fairness which aim to ensure that people who are 'similar' with respect to the classification task receive similar outcomes [5, 6, 23]. Recently, the causal framework [43] has been used to conceptualize fairness. In one such approach, counterfactual fairness has been conceptualized as a setting in which a sensitive attribute might affect the decision along both fair and unfair pathways [30]. In such a scenario, a decision is fair toward an individual if it coincides with the one that would have been taken in a counterfactual world in which the sensitive attribute along the unfair pathways were different. Path-specific counterfactual fairness has then been conceptualized, which attempts to correct the observations corresponding to variables that are descendants of the sensitive attribute along unfair causal pathways [9]. Overall, algorithmic fairness efforts generally consider fairness with respect to *one* sensitive attribute, which is commonly an attribute ascribed to the individual. However, many factors at the individual level are also influenced by structural or social processes (hereafter referred to as "population-level") and thus fairness can be decomposed into a portion at the individual level, and a portion at a population-level.

For example, the problem of potential sex bias in Berkeley's graduate admission is a well-cited situation wherein there was a higher rate of admission for male applicants overall, but when examined by department, a slight bias toward female applicants [4, 43]. This example indicates a fair pathway for the influence of sex on admission (through department choice), and an unfair pathway (if the college treated male and female applicants with the same qualifications and applying to the same departments differently because of sex) [9, 43]. Reasons for the fair pathway are ascribed to social factors. Social factors can likewise be decomposed into fair and unfair components with respect to the specific outcome (e.g. those that shunt women toward fields of graduate study that are more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects as unfair, and those that shape what types of food women versus men eat may not directly affect sex in relation to admission) [4, 18]. Related to this decomposition of effects, recent literature in critical race theory, sociology, epidemiology and algorithmic fairness have discussed similar concerns with the conceptualization of race as a fixed, individual-level attribute [8, 21, 49, 57]. Analytically framing race as an individual-level attribute ignores systemic racism and other factors which are the mechanisms by which race has consequences [8]. Instead, accounting for the structural and institutional-level phenomena associated with race (e.g. structural racism) will augment the impact of algorithmic fairness work [21].

Towards this goal, here we propose a novel definition of fairness called causal *multi-level fairness*, which is defined as a decision being fair towards an individual if it coincides with the one in the counterfactual world where, contrary to what is observed, the individual receives advantaged treatment at the population and individual levels, described by population and individual-level sensitive attributes. This work extends previous work on individual-level path-specific counterfactual fairness to account for both population and individual-level discrimination. Multi-level sensitive attributes are a specific case of multiple sensitive attributes, in which we specifically consider a sensitive attribute at a population level that influences one at an individual level. This is an important and broad class of settings; such multi-level interactions are also considered in multi-level modelling in statistics [19]. Moreover, this work addresses a different problem than the setting of proxy variables, in which the aim is to eliminate influence of the proxy on outcome prediction [25]. Indeed, for variables such as race, detailed analyses from epidemiology have articulated how concepts such as racial inequality can be decomposed into the portion that would be eliminated by equalizing adult socioeconomic status across racial groups, and a portion of the inequality that would remain even if adult socioeconomic status across racial groups were equalized (i.e. racism). Thus if we simply ascribe the race variable to the individual



(a) Proxy (I^N) for sensitive attributes (A_I) **(b) Multiple sensitive attributes A_I^G, A_I^R** **(c) Multi-level sensitive attributes A_I, A_P**

Figure 1: Causal graphs to describe and distinguish different general sensitive attribute settings. Green arrows denote discriminatory causal paths and dashed green-black arrows from any node R to S denote discrimination only due to the portion of the effect from a green arrow into R and not from R itself (fair paths defined via a priori knowledge). a) Individual-level variable I^N (e.g. name) acts as a proxy for the individual sensitive attribute A_I (e.g. perceived race) and affects outcome Y (e.g. health outcomes), I^B (e.g. biological factors) is not a proxy but also affects the outcomes b) Multiple sensitive attributes at the individual level, A_I^G (e.g. gender) and A_I^R (e.g. perceived race) affect individual variables, I, and health outcome, Y, c) Population level variables, A_P (e.g. neighborhood SES), P (e.g. zipcode), and individual-level ones, A_I (e.g. perception of race), I (e.g. biological factors), affect the outcome Y (e.g. health outcomes). Setting c) is described in more detail in Section 4.

(as is sometimes done as a proxy), we will miss the population-level factors that affect it and any particular outcome [49].

Explicitly accounting for multi-level factors enables us to decompose concepts such as racial inequality and approach questions such as: what would the outcome be if there was a different treatment on a population-level attribute such as neighborhood socioeconomic status? This is an important step in auditing not just the sources of bias that can lead to discriminatory outcomes but also in identifying and assessing the level of impact of different causal pathways that contribute to unfairness. Given that in some subject areas, the most effective interventions are at the population level (e.g. in health the largest opportunity for decreasing incidence due to several diseases lie with the social determinants of health such as availability of resources versus individual-level attributes [34]), it is critical to have a framework to assess these multiple sources of unfairness. Moreover, by including population-level factors and their influence on individual ones, we engage themes of inequality and power by specifying attributes outside of an individual's control.

In sum, our work is one step towards integrating perspectives that articulate the multi-dimensionality of sensitive attributes (for example race, and other social constructs) into algorithmic approaches. We do this by bringing focus to social processes which often affect individual attributes, and developing a framework to account for multiple casual pathways of unfairness amongst them. Our specific contributions are:

- We formalize multi-level causal systems, which include potentially fair and unfair path effects at both population and individual-level variables. Such a framework enables the algorithmicist to conceptualize systems that include the systemic factors that shape outcomes.
- Using the above framework, we demonstrate that residual fairness can result if population-level attributes are not accounted for.
- We demonstrate an approach for mitigating multi-level unfairness while retaining model performance.

2 RELATED WORK

Causal Fairness. Several statistical fairness criteria have been introduced over the last decade, to ensure models are fair with respect to group fairness metrics or individual fairness metrics. However, further discussions have highlighted that several of the fairness metrics cannot be concurrently satisfied on the same data [3, 10, 17, 24, 26, 33]. In light of this, causal approaches to fairness have been recently developed to provide a more intuitive reasoning corresponding to domain specifics of the applications [7, 9, 25, 29–31, 38, 39, 45, 46, 55, 56]. Most of these approaches advocate for fairness by addressing an unfair causal effect of the sensitive attribute on the decision. Kusner et al. [30] have introduced an individual-level causal definition of fairness, known as counterfactual fairness. The intuition is that the decision is fair if it coincides with the one that would have been taken in a counterfactual world in which the individual would be identified by a different sensitive attribute. For example, a hiring decision is counterfactually fair if the individual identified by the gender *male* would have also been offered the job had the individual identified as *female*. In sum, these efforts develop concepts of fairness with respect to individual level sensitive attributes, while here we develop fairness accounting for multi-level sensitive attributes. Inspired from research in social sciences [6, 21, 49], the work here extends the idea to multi-level sensitive attributes. In our considered setting, a decision is fair not only if the individual identified with a different individual-level sensitive attribute but also if the individual received advantaged treatment at the population-level, attributed as the population-level sensitive attribute.

Identification of Causal Effects. While majority of the work in causal inference is on identification and estimation of total causal effects [1, 40, 43, 44, 48], studies have also looked at identifying the causal effects along certain causal pathways [42, 47]. The most common approach for identifying the causal effects along different causal pathways is decomposing the total causal effect along direct

and indirect pathways [41, 42, 47]. We leverage the approach developed by Shpitser [47] on how causal effects of multiple variables along a single causal pathway can be identified. While we assume no unmeasured confounding for the analysis in this work, research in causal estimation in the presence of unmeasured confounding, [35, 51] can be used to extend our current contribution.

Path-specific causal fairness. Approaches in causal fairness such as path-specific counterfactual fairness [9, 38, 39], proxy discrimination, and unresolved discrimination [25, 52] have aimed to understand the effect of sensitive attributes (i.e. variables that correspond to gender, race, disability, or other protected attributes of individuals) on outcomes directly as well as indirectly to identify the causal pathways (represented using a causal diagram) that result into discriminatory predictions of outcomes based on the sensitive attribute. Generally such approaches focus on sensitive attributes at the individual-level and require an understanding of the discriminatory causal pathways for mitigating causal discrimination along the specified pathways. Our approach echoes that of Chiappa [9] for mitigating unfairness by removing path-specific effects for fair predictions, and we do so while accounting for both individual and population-level unfairness.

Intersectional fairness. There is recent focus on identifying the impact of multiple sensitive attributes (intersectionality) on model predictions. There are several works that have been developed for this setting [16, 37, 53], however these approaches do not take into account the different causal interactions between sensitive attributes themselves which can be important in identifying bias due to intersectionality. In particular, this includes intersectional attributes at the individual and population level, such as race and socioeconomic status [11, 28].

3 BACKGROUND

We begin by introducing the tools needed to outline multi-level fairness, namely, (1) causal models, (2) their graphical definition, (3) causal effects, and (4) causal effects in multi-level systems consisting of individual and population variables.

Causal Models: Following Pearl et al. [43] we define a causal model as a triple of sets $(\mathbf{U}, \mathbf{V}, F)$ such that:

- $\mathbf{U}$ are a set of latent variables, which are not caused by any of the observed variables in $\mathbf{V}$.
- F is set of functions for each $V_i \in \mathbf{V}$, such that $V_i = f_i\left(pa_i, U_{pa_i}\right)$, $pa_i \subseteq \mathbf{V} \setminus V_i$ and $U_i \subseteq \mathbf{U}$. Such equations relating V_i to pa_i and U_i are known as structural equations [12].

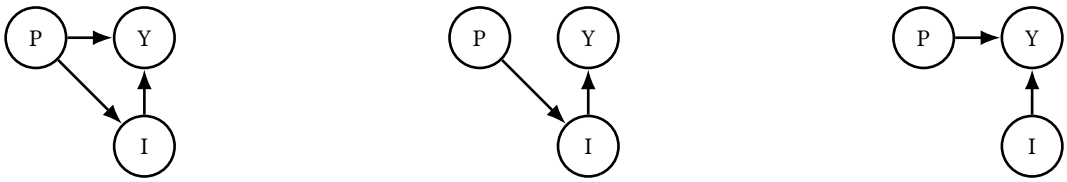
Here, "pa_i" refers to the causal parents of V_i, the variables that affect the value V_i obtains. We provide a graphical interpretation of pa_i in detail below. The structural equations relating V_i to pa_i are important to define the joint distribution over all the variables given by the product of the conditional distribution of each variable, V_i given it's causal parents pa_i, represented as $\Pr\left(V_i \mid pa_i\right)$. The joint distribution is given by:

$$\Pr\left(\mathbf{V}\right) = \prod_i \Pr\left(V_i \mid pa_i\right). \tag{1}$$

Graphical definition of Causal Models: While structural equations define the relations between variables we can graphically represent the causal relationships between random variables using Graphical Causal Models (GCMs) [9]. The nodes in a GCM represent the random variables of interest, V, and the edges represent the causal and statistical relationships between them. Here, we restrict our analysis to Directed Acyclic Graphs (DAGs) where there are no cycles, i.e., a node cannot have an edge both originating from and terminating at itself. Furthermore, a node Y is known as a child of another node X if there is an edge between the two that originates at X and terminates on Y. Here, X is called a direct cause of Y. If Y is a descendant of X, with some other variable Z along the path from X to Y, $X \to \cdots \to Z \to \cdots Y$, then Z is known as a *mediator* variable and X remains as a potential cause of Y. For example, in Figure 1b, A_I^G, A_I^R, I, Y are the random variables of interest. A_I^G, A_I^R are direct causes of I and Y while I is also a direct cause of Y. Here, I is a mediating variable for both A_I^G and A_I^R, decomposing the total average causal effect on Y into direct ($A_I^G \to Y, A_I^R \to Y$) and indirect effects via I ($A_I^G \to I \to Y, A_I^R \to I \to Y$). In this case, the paths from A_I^G and A_I^R to Y through I are fair, as commonly represented in fairness problem [9, 38].

Causal effects: The causal effect of X on Y is the information propagated by X towards Y via the causal directed paths, $X \to \cdots \to Y$. This is equal to the conditional distribution of Y given X if there are no bidirected paths between X and Y. A bidirected path between X and Y, $X \leftarrow \cdots \to Y$ represents confounding; that is some causal variable, confounder affecting both X and Y. If confounders are present then the causal effect can be estimated by intervening upon X. This means that we externally set the value of X to the desired value x and remove any edges that terminate at X, since manually setting X to x would inhibit any causation by pa_x. After intervening the causal effect of X on Y is given by $\Pr^*\left(Y \mid X = x\right) = \sum_Z \Pr\left(Y \mid X = x, Z\right) \Pr\left(Z\right)$, where $Z = \mathrm{V} \setminus \{X, Y\}$. The intervention $X = x$ results into a potential outcome, $Y_{X=x}$, where the distribution of the potential outcome variable is defined as $\Pr\left(Y_x\right) = \Pr^*\left(Y \mid X = x\right)$.

Causal effects in multi-level systems We now look at how the idea of causal effects can be extended to multi-level systems. Multi-level systems comprise of interaction between variables at multiple levels [19]. When considering systems that describe data from people, the lowest level, say, Level 1 represents individual-level factors that correspond to specific individual-level attributes. Level 1 observations about individuals can be grouped together into mutually exclusive categories at "Level 2". These Level 2 groups can correspond to, for example, neighborhoods, schools or hospitals. Here, we refer to information at Level 2 as population-level information because information at Level 2 has the ability to affect more than one individual. For example, the population-level variable of neighborhood socio-economic status (which can be computed as mean income of all neighborhood residents) affects resources available



(a) $I = f_I(\mathrm{P}), Y = f_Y(\mathrm{I}, \mathrm{P})$. **(b)** $I = f_I(\mathrm{P}), Y = f_Y(\mathrm{I})$. **(c)** $Y = f_Y(\mathrm{I}, \mathrm{P})$.

Figure 2: Interactions between population variables P, individual variables I, and outcome Y, structural equations corresponding to the specific causal graphs are outlined for each of the structure in a), b), and c).

for communities altogether rather than just a single individual. We represent variables at Level 1, individual-level variables, as I[1]. Population-level variables (Level 2) are represented by P. To examine the possible cases of interaction between these variables to understand the causal effect of P on Y, and determine if the causal effect of P on Y can be identified in all scenarios, let us consider the simplistic task of predicting a health outcome, denoted by Y from population-level variables P and individual-level variables I. Interactions between P, I, and Y can be represented with causal graphs as shown in Figure 2. Cases (1), (2), and (3) represent all possible interactions between P, I, and Y, under no unmeasured confounding. Although scenarios 2 and 3 are subsets of 1, we elucidate all of them for clarity. We demonstrate how the causal effect of P on Y can be estimated in these three scenarios depending on the causal paths between P and Y, where the causal effect of P on Y can vary due to the changes in the causal structure.

Case 1. Both population and individual level factors affect health outcomes through different causal paths. Figure 2a presents such a case where both individual-level factors and health outcome are affected by population-level factors. Recalling our definition of causal models, here $U = P, \mathbf{V} = \{I, Y\}$. P propagates its true value p along the paths $P \rightarrow Y$ and $P \rightarrow I \rightarrow Y$. Therefore, we can estimate the causal effect of P on Y as follows:

$$\Pr(Y \mid P = p) = \int_I \Pr(Y \mid I, p) \Pr(I \mid p). \tag{2}$$

Case 2. Population-level factors affect health outcomes indirectly by only affecting individual-level factors, and having no direct effect on the outcome. This scenario is represented in Figure 2b. In this setting the effect of P on Y is propagated only via I, i.e., via $P \rightarrow I \rightarrow Y$. Thus the causal effect of P on Y is only an indirect effect, calculated as:

$$\Pr(Y \mid P = p) = \int_I \Pr(Y \mid I) \Pr(I \mid p). \tag{3}$$

Case 3. Population and individual-level variables do not interact, but independently affect the outcome, as represented in Figure 2c. In this case $\Pr(I \mid P = p) = \Pr(I)$ since $I \perp\!\!\!\perp P$. The causal effect of P on Y simplies to:

$$\Pr(Y \mid P = p) = \int_I \Pr(Y \mid I, p) \Pr(I). \tag{4}$$

[1]Variables will be denoted by uppercase letters, V, values by lowercase letters, v, and vectors by bold letters, V.

We now develop upon an understanding of how these interactions can lead to unfairness, by introducing multi-level sensitive attributes and their path specific effects.

4 MULTI-LEVEL SENSITIVE ATTRIBUTES

To begin discussion of multi-level sensitive attributes, we first highlight that it is critical to identify the presence of sensitive variables (those having the potential of leading to discrimination) at both individual and population level. While previous work in fairness has considered sensitive attributes such as age, gender, and race at the individual level, there are several examples of population-level attributes highlighted in the epidemiology, computer science and sociology literature, that have influence on individual-level attributes (e.g. neighborhood SES, urban/rural status, social norms, etc.), which one may want to be fair with respect to [20, 21, 49, 57]. Here, we represent sensitive attributes at the individual level by A_I and those at the population level by A_P. It is important to note that by the nature of individual and population-level attributes, A_I can be influenced by A_P. For example, consider neighborhood socioeconomic status (SES) as A_P, as we may want to be fair to a person's context. Neighborhood SES can be a direct cause of other population-level factors such as structural racism, and also can affect individual-level attributes that are protected (A_I) (e.g. perception of race) and otherwise (I) (e.g. body-mass index, given that neighborhood SES can shape the types of food/physical activity resources available [14, 22]). Thus we illustrate a general framework for population/individual attribute relationships and possible unfair paths in Figure 1c. In illustrating the relationship between A_P and A_I (green arrow) we make distinction between the multi-level fairness setting and literature on intersectionality in fairness, which has considered multiple sensitive attributes that are independent of each other (Figure 1b) [16, 37, 53, 54]. While this assumption of independence may hold for individual-level sensitive attributes like age and race, it fails when we consider sensitive attributes at both the individual and population level. It is important to understand the relationship between individual and population-level attributes in order to delineate the path-specific effects that may lead to biased predictions of Y, in the following section.

5 MULTI-LEVEL PATH-SPECIFIC FAIRNESS

Path-specific effects. We have demonstrated causal effects of all possible interactions of population and individual variables on outcomes (Figure 2). However, we are often interested in identifying the causal effect of a treatment on an outcome only along a certain

causal pathway, instead of identifying the average causal effect. This is known as a *path-specific effect* and corresponds to the effect of treatment on the outcome *only* via a specific path of interest [40]. For example, the path specific effect of X on Y along path π is defined by evaluating the outcome Y_x that would be obtained if X propagates the observed value x only along π and x', the counterfactual value along all the other paths. In this case the effect of any mediator between X and Y is evaluated according to the corresponding value of X propagated along the specific causal path.

Path-specific counterfactual fairness is defined as a decision being fair towards an individual if it coincides with the one that would have been made in a counterfactual world in which the sensitive attribute along the unfair pathways was set to the counterfactual value [9]. Estimating path-specific counterfactual fairness requires a prior understanding of the discriminatory causal paths. Chiappa [9] further propose that in order to obtain a fair decision, the path-specific effect of the sensitive attribute along the discriminatory causal paths is removed from the model predictions. In order to do this, the idea that path-specific effects can be formulated as nested counterfactuals is leveraged [47]. This is done by considering that along the causal path of interest, the variable propagates the observed value, for example, a_p in Figure 1c, and along other pathways, the counterfactual value, a'_p is propagated. For example, if we consider A_P to be binary with two values 0 and 1, one of them is considered to be the baseline representing advantaged treatment while the other represents discrimination. Here, the primary object of interest is the potential outcome variable, $Y(a_p)$, which represents the outcome if, possibly contrary to fact, A_P were set to value a_p. Given the values of a_p, a'_p, comparison of $Y(a'_p)$ and $Y(a_p)$ in expectation: $\mathbb{E}[Y(a_p)] - \mathbb{E}[Y(a'_p)]$ would allow us to quantify the path-specific causal effect of A_P on Y. However, a key limitation of Chiappa [9] is that only the path-specific counterfactual fairness with respect to the sensitive attributes at the individual level are considered. The presence of sensitive attributes at the population-level that affect properties at the individual level makes it non-trivial to estimate the path-specific effect consisting of both population and individual-level sensitive attributes. Following, we develop path-specific counterfactual fairness with respect to multi-level sensitive attributes by estimating *multi-level path-specific effects*. To reiterate, we define *multi-level fairness* as a decision being fair towards an individual if it coincides with the one in the counterfactual world where contrary to what is observed, the individual receives advantaged treatment at both the population and individual-levels, described by population and individual-level sensitive attributes.

5.1 Identification of multi-level path-specific effects.

Proposition 1. *In the absence of any unmeasured confounding between A_P and A_I, the multi-level path-specific effects of both A_P and A_I on Y are identifiable.*

Proposition 1 follows from the possible structural interactions between A_P and A_I. We illustrate how the path-specific effects can be identified using the scenario from Figure 1c which presents such an exemplar case of multi-level interactions where population-level sensitive attributes can cause individual level variables. Overall the path-specific effect (PSE) can be identified as :

$$\begin{aligned}&\Pr\left(Y \mid A_P = a_p, A_I = a_i\right)\\ &= \int_{P,I} \Pr\left(Y \mid P, I, a_p, a_i\right) \Pr\left(I \mid P, a_p, a_i\right) \Pr\left(P \mid I, a_i, a_p\right)\end{aligned} \tag{5}$$

where the conditionals vary based on the specific causal path under analysis and the sensitive attributes intervened upon. For example, let us analyze the multi-level path-specific effect of just $A_P = a_p$ without considering the effect of A_I via the causal path $A_P \to P \to A_I \to I \to Y$:

$$\text{PSE}^{A_P \to P \to A_I \to I \to Y}_{A_P \leftarrow a_p} = \mathbb{E}[Y(P(a_p), A_I(a'_p, P), I(a'_p, A_I))] - \mathbb{E}[Y(a'_p)]. \tag{6}$$

However, the combined effect of $A_P \leftarrow a_p$ and $A_I \leftarrow a_i$ on Y via path $A_P \to P \to A_I \to I \to Y$ on Y is different than the one given in Equation 6 and follows as:

$$\text{PSE}^{A_P \to P \to A_I \to I \to Y}_{A_I \leftarrow a_i, A_P \leftarrow a_p} = \mathbb{E}[Y(P(a_p), a'_i, I(a'_p, a_i,))] - \mathbb{E}[Y(a'_p, a'_i)]. \tag{7}$$

Thus, the path-specific effect along the same path can differ based on which sensitive variables were intervened upon. This makes it crucial to analyze the effect of the population-level sensitive attributes as well as the individual-level sensitive attributes along a discriminatory causal pathway.

Now let us now try to evaluate the multi-level path-specific effects for the setting represented in Figure 1c and assuming linear interactions between all the variables. The data generating process is as follows:

$$\begin{aligned}A_P &= \text{Bernoulli}\,(\pi)\,,\\ P &= \theta^p + \theta^p_{a_p} A_P + \epsilon_p,\\ A_I &= \text{logit}\left(\theta^{a_i} + \theta^{a_i}_{a_p} A_P + \theta^{a_i}_p P + \epsilon_{a_i}\right),\\ I &= \theta^i + \theta^i_{a_p} A_P + \theta^i_{a_i} A_I + \epsilon_i,\\ Y &= \theta^y + \theta^y_p P + \theta^y_i I + \theta^y_{a_i} A_I + \epsilon_y.\end{aligned} \tag{8}$$

A_P and A_I are binary variables where we assume a'_p and a'_i to represent the baseline values denoting advantaged treatments. The rest of the variables P, I and Y are continuous and follow a linear relationship between the parents and the specific variables. $\epsilon_p, \epsilon_{a_i}, \epsilon_i, \epsilon_y$ are unobserved zero-mean Gaussian terms with variances $\sigma^2_p, \sigma^2_{a_i}, \sigma^2_i, \sigma^2_y$ respectively.

For this specific model we obtain the multi-level path-specific effects consisting of both A_P and A_I. The population level sensitive attribute, A_P affects Y via multiple paths along 1) $A_P \to P \to Y$, 2) $A_P \to I \to Y$, 3) $A_P \to A_I \to Y$, and 4) $A_P \to A_I \to I \to Y$; the individual level sensitive attribute A_I affects Y along two causal paths, 1) $A_I \to I \to Y$, and 2) $A_I \to Y$. If we assume that $A_I \to Y$ and $P \to Y$ are not discriminatory causal paths then we need to assess the path-specific effect along $\cdots \to I \cdots Y$. In order to assess the potential discriminatory factors along $\cdots \to I \to Y$ we look at the path-specific effects along $A_P \to I \to Y$ and $A_I \to I \to Y$ by considering the effects of 1) both A_P, A_I, 2) only A_P, and 3) only A_I.

1. (A_P, A_I.) Multi-level path-specific effect of both A_P and A_I is given as follows:

$$\text{PSE}^{A_P \to I \to Y, A_I \to I \to Y}_{A_P \leftarrow a_p, A_I \leftarrow a_i} = \mathbb{E}\left[Y\left(a_i', P\left(a_p'\right), I\left(a_i, a_p\right)\right)\right]. \quad (9)$$

The path-specific effect along $A_P \to I \to Y$ and $A_I \to I \to Y$ is computed by comparison of the counterfactual variable $Y\left(a_i', P\left(a_p'\right), I\left(a_i, a_p\right)\right)$, where a_p, a_i are set to the baseline value of 1 in one counterfactual and 0 in another. The counterfactual distribution can be estimated as follows:

$$\int_{P,I} \Pr\left(Y \mid a_i', I, P\right) \Pr\left(I \mid a_i, a_p\right) \Pr\left(P \mid a_p'\right). \quad (10)$$

We obtain the mean of this distribution as follows:

$$\mathbb{E}[Y \mid I, P, ai'] = \theta^y + \theta^y_i \theta^i + \theta^y_p \theta^p + \theta^y_{a_i} a_i' + \theta^y_i \theta^i_{a_i} a_i + \theta^y_i \theta^i_{a_p} a_p + \theta^y_p \theta^p_{a_p} a_p'. \quad (11)$$

We then obtain the path-specific effect along $A_P \to I \to Y$ and $A_I \to I \to Y$, $\mathbb{E}[Y \mid a_p, a_i] - \mathbb{E}[Y \mid a_p', a_i']$ by comparing the observed and counterfactual value of the mean of the counterfactual distribution. This results into the following:

$$\theta^y_{a_i}\left(a_i' - a_i'\right) + \theta^y_i \theta^i_{a_i}\left(a_i - a_i'\right) + \theta^y_i \theta^i_{a_p}\left(a_p - a_p'\right) + \theta^y_p \theta^p_{a_p}\left(a_p' - a_p'\right) = \theta^y_i \theta^i_{a_i}\left(a_i - a_i'\right) + \theta^y_i \theta^i_{a_p}\left(a_p - a_p'\right). \quad (12)$$

Substituting $a_i = 1, a_p = 1, a_i' = 0, a_p' = 0$ in Equation 12 we obtain the path-specific effect as a funtion of the parameters θ:

$$\text{PSE}^{A_P \to I \to Y, A_I \to I \to Y}_{A_P \leftarrow a_p, A_I \leftarrow a_i} = \theta^y_i \theta^i_{a_i} + \theta^y_i \theta^i_{a_p}. \quad (13)$$

2. (A_P). We can use the same technique for evaluating the path-specific effect of $A_P = a_p$ along $A_P \to I \to Y$, $A_I \to I \to Y$ without intervening upon the individual level sensitive attribute A_I. The counterfactual distribution in this case is as follows:

$$\int_{P,I,A_I} \Pr\left(Y \mid A_I, I, P\right) \Pr\left(I \mid A_I, a_p\right) \Pr\left(P \mid a_p'\right) \Pr\left(A_I \mid P, a_p'\right). \quad (14)$$

Following the procedure from Equation 11, we obtain the path-specific effect along $A_P \to I \to Y, A_I \to I \to Y$ as[2]:

$$\text{PSE}^{A_P \to I \to Y, A_I \to I \to Y}_{A_p \leftarrow a_p} = \theta^y_i \theta^i_{a_p}. \quad (15)$$

3. (A_I). We also assess the effect of only intervening on the individual-level attributes via paths $A_I \to I \to Y$ and $A_P \to I \to Y$. The counterfactual distribution in this case looks as follows:

$$\int_{P,I,A_P} \Pr\left(Y \mid a_i, I, P\right) \Pr\left(I \mid a_i, A_P\right) \Pr\left(P \mid A_P\right) \Pr\left(A_P\right) \quad (16)$$

The mean distribution of which is:

$$\theta^y + \theta^y_i \theta^i + \theta^y_p \theta^p + \theta^y_{a_i} a_i' + \theta^y_i \theta^i_{a_i} a_i + \theta^y_i \theta^i_{a_p} a_p' + \theta^y_p \theta^p_{a_p} a_p'. \quad (17)$$

[2]Plug-in estimators can be used to compute the parameters needed for estimating PSE.

The path-specific effect along $A_P \to I \to Y$ and $A_I \to I \to Y$ can thus be calculated as:

$$\theta^y_{a_i}\left(a_i' - a_i'\right) + \theta^y_i \theta^i_{a_i}\left(a_i - a_i'\right) + \theta^y_i \theta^i_{a_p}\left(a_p' - a_p'\right) + \theta^y_p \theta^p_{a_p}\left(a_p' - a_p'\right) \quad (18)$$

$$= \theta^y_i \theta^i_{a_i}\left(a_i - a_i'\right). \quad (19)$$

After substituting $a_i = 1$ and $a_i' = 0$ this results in: $PSE^{A_I \to I \to Y, A_P \to I \to Y}_{A_i \leftarrow a_i} = \theta^y_i \theta^i_{a_i}$.

Therefore, the path-specific effect representing the unfairness along a path changes, based on which sensitive attributes were intervened upon, affecting the values of a_i and a_p in the counterfactual distribution. By the same approach, path-specific effects for individual and population-level sensitive attributes as well as the multi-level path specific effects for all the causal paths in Figure 1c, are evaluated and reported in Table 1.

5.2 Fair predictions with multi-level path-specific effects.

Fair outcomes can be estimated by removing the unfair path-specific effect from the prediction (essentially making a correction on all the descendants of the sensitive attributes), $\hat{Y}$ by simply subtracting it, i.e. $\hat{Y}_{\text{fair}} = \hat{Y} - \text{PSE}$. As done in Chiappa [9], removing such unfair information at test time ensures that the resulting decision coincides with the one that would have been taken in a counterfactual world in which the sensitive attribute along the unfair pathways were set to the baseline. We leverage this same idea for multi-level causal fairness by generating fair predictions, $\hat{y}_{\text{fair}}$ by controlling for the path-specific effects of multi-level sensitive attributes. For example, consider the PSE resulting from intervening upon both A_P and A_I via the path $\cdots \to I \to Y$, $\text{PSE}^{A_P \to I \to Y, A_I \to I \to Y}_{A_P \leftarrow a_p, A_I \leftarrow a_i} = \theta^y_i\left(\theta^i_{a_i} + \theta^i_{a_p}\right)$, we can control for the discrimination at test time as follows:

$$\hat{y}^n_{\text{fair}} = \theta^y + \theta^y_p p^n + \theta^y_i i^n + \theta^y_a a^n_i - \beta\left(\theta^y_i\left(\theta^i_{a_i} + \theta^i_{\theta^i_{a_p}}\right)\right) \quad (20)$$

where β accounts for the control over the path-specific effect. For our analysis β ranges from 0 to 1, where 0 denotes that we do not account for any path-specific effect while 1 denotes that we remove the entire path-specific effect. Intermediate values allow for only partial removal of the path-specific effect. The effect of the control over different path-specific effects on the performance metric of the classifier for the example from Figure 1c is shown in Figure 3a. The algorithm for implementing this approach is presented below.

Algorithm 1 Causal Multi-level Fairness

Input: Causal Graph $\mathcal{G}$ consisting of nodes V, data $\mathcal{D}$ over V, discriminatory causal pathways π, β.
Output: Fair predictor $\hat{Y}_{\text{fair}}$, model parameters θ
for $V_i \in \mathbf{V}$ **do**
  Estimate $\hat{\theta}^{V_i} \leftarrow \arg\min_\theta \sum_{k \in \mathcal{D}} l\left(V_i^{(k)}, f_i\left(pa_i^{(k)}\right)\right)$
end for
Calculate path-specific effects, PSE along π using $\{\theta^{\mathbf{V}}\}$
return $\hat{Y}_{\text{fair}} = \mathbb{E}\left[f_Y\left(\theta^Y\right)\right] - \beta * \text{PSE}_\pi$

Sensitive Attribute	Causal Path(s)	PSE
A_P	$A_P \to P \to Y$	$\theta_p^y \cdot \theta_{a_p}^p$
	$A_P \to I \to Y$	$\theta_i^y \cdot \theta_{a_p}^i$
	$A_P \to A_I \to I \to Y$	$\theta_i^y \cdot \theta_{a_i}^i \cdot \theta_{a_p}^{a_i}$
	$A_P \to A_I \to Y$	$\theta_{a_i}^y \cdot \theta_{a_p}^{a_i}$
	$A_P \to A_I \to Y$	$\theta_{a_i}^y \cdot \theta_{a_p}^{a_i}$
	$A_P \to \cdots \to I \to Y$	$\theta_i^y \left(\theta_{a_p}^i + \theta_{a_i}^i \cdot \theta_{a_p}^{a_i}\right)$
	$A_P \to \cdots \to Y$	$\theta_p^y \cdot \theta_{a_p}^p + \theta_i^y \left(\theta_{a_p}^i + \theta_{a_i}^i \cdot \theta_{a_p}^{a_i}\right) + \theta_{a_i}^y \cdot \theta_{a_p}^{a_i}$
A_I	$A_I \to Y$	$\theta_{a_i}^y$
	$A_I \to I \to Y$	$\theta_i^y \cdot \theta_{a_i}^i$
	$A_I \to \cdots \to Y$	$\theta_{a_i}^y + \theta_i^y \cdot \theta_{a_i}^i$
A_P, A_I	$A_P \to P \to Y, A_I \to Y$	$\theta_p^y \cdot \theta_{a_p}^p + \theta_{a_i}^y$
	$A_P \to P \to Y, A_I \to I \to Y$	$\theta_p^y \cdot \theta_{a_p}^p + \theta_i^y \cdot \theta_{a_i}^i$
	$A_P \to P \to Y, A_I \to I \to Y, A_I \to Y$	$\theta_p^y \cdot \theta_{a_p}^p + \theta_i^y \cdot \theta_{a_i}^i + \theta_{a_i}^y$
	$A_P \to I \to Y, A_I \to I \to Y$	$\theta_i^y \left(\theta_{a_i}^i + \theta_{a_p}^i\right)$
	$A_P \to I \to Y, A_I \to Y$	$\theta_{a_i}^y + \theta_i^y \cdot \theta_{a_p}^i$
	$A_P \to I \to Y, A_I \to \cdots \to Y$	$\theta_{a_i}^y + \theta_i^y \left(\theta_{a_i}^i + \theta_{a_p}^i\right)$
	$A_P \to \cdots \to Y, A_I \to \cdots \to Y$	$\theta_{a_i}^y + \theta_i^y \left(\theta_{a_i}^i + \theta_{a_p}^i\right) + \theta_i^y \cdot \theta_{a_p}^i$

Table 1: Path-specific effects along different causal paths, each obtained by same procedure as Eq. 12.

6 EXPERIMENTS

Our experiments in the this section serve to empirically assess and demonstrate residual unfairness when correcting for individual or population-level versus individual and population-level path specific effects. We consider a synthetic setting demonstrated in Figure 1c and real-world setting for income prediction presented in Figure 4a; in both cases the structural causal model is known a priori. In each case, we first learn the parameters of the model (Θ) from the observed data based on the known underlying causal graph, $\mathcal{G}$ as illustrated in Algorithm 1. We then draw observed and counterfactual samples from the distribution θ by sampling from the counterfactual individual and population-level sensitive attributes, $A_I = a_i'$ and $A_P = a_p'$, respectively. We estimate fair predictors for both the observed and counterfactual data using Algorithm 1. In case of no resultant unfairness, distributions of the outcome estimate for both observed and counterfactual data should coincide. Density plots that do not coincide is evidence of residual unfairness.

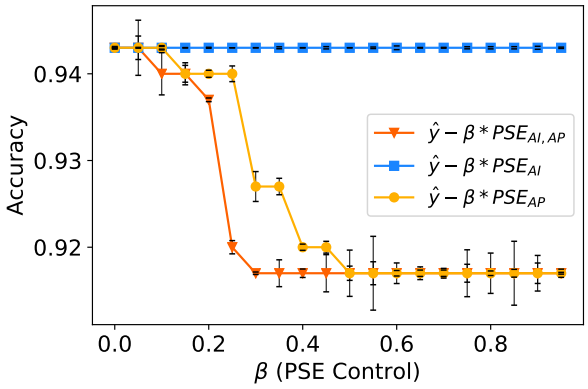
6.1 Synthetic setting

Here we consider the data-generating process presented in Equation 8 where $\theta^p = 0.2, \theta_{a_p}^p = 0.9, \theta^{a_i} = 0.2, \theta_{a_p}^{a_i} = 0.7, \theta_p^{a_i} = 0.05, \theta^i = 0.3, \theta_{a_p}^i = 0.2, \theta_{a_p}^i = 0.9, \theta^y = 0.2, \theta_p^y = 0.7, \theta_i^y = 0.75, \theta_{a_i}^y = 0.68$, with the exception that we consider Y to be a binary variable $\left(Y = \text{logit}\left(\theta^y + \theta_p^y P + \theta_i^y I + \theta_{a_i}^y A_I + \epsilon_y\right)\right)$. We first generate data according to the structural equations from Equation 8 and then train a linear regression model for I, P and a logistic regression model for A_I, Y to learn the parameters (θ). We then calculate multi-level path-specific effects of both A_P and A_I based on Table 1. Resulting effects are summarized in Table 2.

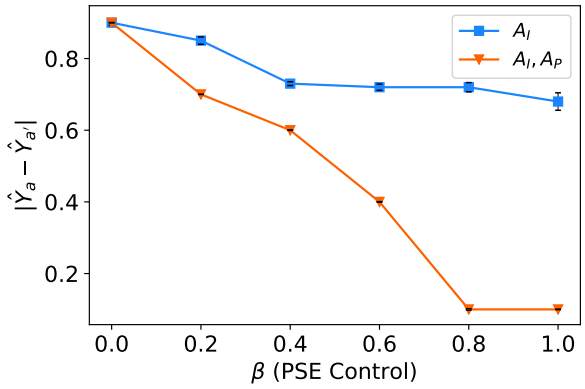
Causal Path(s)	PSE
$A_I \to Y$	0.077
$A_I \to I \to Y$	0.007
$A_I \to \cdots \to Y$	9.794
$A_P \to P \to Y, A_I \to Y$	0.101
$A_P \to P \to Y, A_I \to I \to Y$	9.742
$A_P \to P \to Y, A_I \to I \to Y, A_I \to Y$	9.819
$A_P \to I \to Y, A_I \to I \to Y$	17.491
$A_P \to I \to Y, A_I \to Y$	7.851
$A_P \to I \to Y, A_I \to \cdots \to Y$	17.569
$A_P \to \cdots \to Y, A_I \to \cdots \to Y$	25.343

Table 2: Path-specific effects along different causal paths for different sensitive attributes, procedure similar to the one outlined in Equation 12.

We observe that controlling for the multi-level path-specific effects leads to little drop in the model performance (accuracy) shown in Figure 3a. The *blue* line represents the performance with respect to the control parameter (β) for PSE while only accounting for the path-specific effect of A_I, the *orange* line represents performance while controlling for the path-specific effect of the both A_P and A_I, and the *yellow* line represents the accuracy while controlling



(a) Model performance for different fair predictors, $\hat{y}_{\text{fair}} = \hat{y} - \beta * \text{PSE}$ accounting for PSE due to only A_I, A_P, and both A_I, A_P.



(b) Path-specific unfairness controlling for discriminatory effects of just A_I (blue) and both A_I, A_P (orange), over 10 runs.

Figure 3: Results for synthetic experiments for the setting presented in Figure 1c.

for the effect of A_P alone, all along the path $\cdots \rightarrow I \rightarrow Y$. Since the PSE for A_I is small (0.007) there is little effect on the model performance shown by the steady nature of the *blue* line.

We also assess the counterfactual fairness of the resulting model after removing the unfair multi-level path-specific effects. We assess the counterfactual fairness of the resulting models while accounting for unfairness due to both A_P, A_I and only A_I (Figure 3b). The Y axis shows the average difference between observed and counterfactual predictions, $|\hat{Y}_a - \hat{Y}_{a'}|$ where a is the observed value of the sensitive attribute and a' is the corresponding counterfactual value. This difference is analyzed for varying control of the path-specific effect (PSE), β along the X-axis. The *blue* line represents the difference while accounting only for the individual-level path specific effect A_I while the *orange* line represents the multi-level path-specific effect of both A_P and A_I. For counterfactually fair models we expect the difference to be close to zero since controlling for the path-specific effects cancels out the counterfactual change. We observe that this happens only while accounting for the multi-level path-specific effects and not just individual-level path specific effect. Accounting for just the individual-level path-specific effect (blue line) does not result in counterfactually fair predictions as can be seen from the high difference (0.7-0.9). Thus, correcting for the unfair multi-level path-specific effects of both A_P and A_I results in counterfactually fair predictions with a slight drop in accuracy from 94.5% to 91.5%.

6.2 UCI Adult Dataset

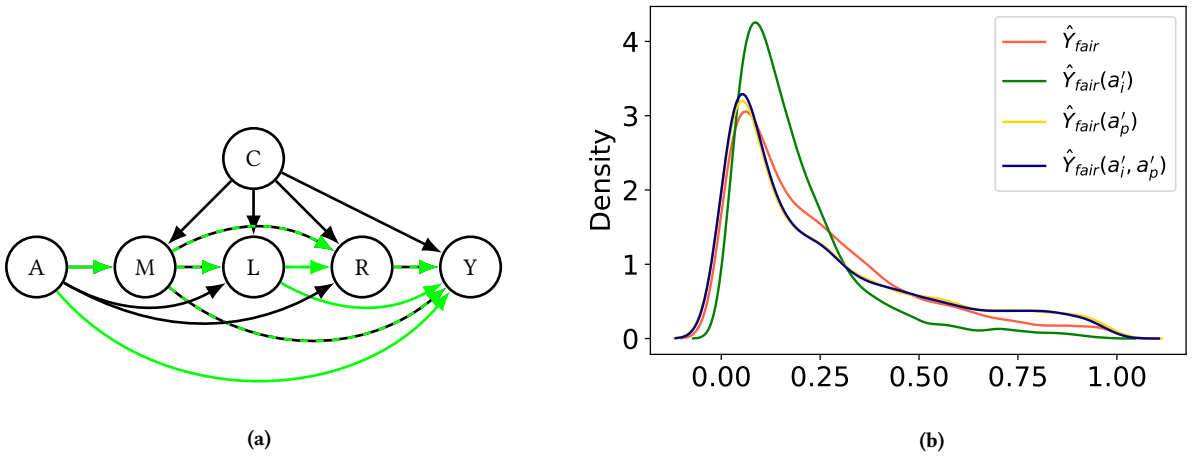
We demonstrate real-world evaluation of the proposed approach on the Adult dataset from the UCI repository [32]. The UCI Adult dataset is amenable for demonstration of our method, given that it was used in Chiappa [9] to assess path-specific counterfactual fairness, and because there are variables oriented in a manner that create a potential multi-level scenario. We consider this well-studied setting [9, 38] wherein the goal is to predict whether an individual's income is above or below $50,000. The data consist of age, working class, education level, marital status, occupation, relationship, race, gender, capital gain and loss, working hours, and nationality variables for 48842 individuals. The causal graph is represented in Figure 4a. Consistent with Chiappa [9] and Nabi et al. [38], we do not include race or capital gain and loss variables. A represents the individual-level protected attribute sex, C is age and nationality, M is marital status, L is the level of education, R corresponds to working class, occupation, and hours per week, Y is the income class. We make an observation that education level, L, could, in certain situations, be considered a population-level sensitive attribute. This is because it may be affected by social influences and/or resources available at the neighborhood level. Indeed, the relationship between L and other variables in the graph mimic that of a multi-level scenario. Moreover, it may be reasonable to consider a setting in which it is desirable to predict income, while also being fair to education, in order to understand the effect that education level may have, and show the importance of intervening on education at a population-level. Thus, we emphasize that there may be other scenarios (as described in above sections) and data that better express the need for population and individual-level sensitive attributes, however we consider this dataset/setting in order to be consistent to previous work in causal fairness which has used the same dataset, though have only analyzed unfairness due to an individual-level sensitive attribute [9]. Paths in Figure 4a are thus defined based on mapping the figure from Chiappa [9] to the multi-level framework in Figure 1b with L as the population-level sensitive attribute and A as the individual-level sensitive attribute.

To first evaluate the path-specific effect of just the individual-level sensitive attribute, we evaluate the effect of only A (the individual level sensitive attribute) along $A \rightarrow Y$ and $A \rightarrow M \rightarrow \cdots \rightarrow Y$ to be 3.716. We obtain a fair prediction as follows:

$$\hat{y}_{\text{fair}} = \text{logit}\left(\theta^y + \theta^y_c C + \theta^y_m \left(M - \theta^m_a\right) + \theta^y_l \left(L - \theta^l_m \theta^m_a\right)\right)$$

by removing the path specific effect of A along a priori known discriminatory paths. The fair accuracy is 78.87%, consistent with Chiappa [9].

We next evaluate the path-specific effect of both the population (L) and individual-level sensitive attributes (A). In the counterfactual world, the individual receives the advantaged population and



(a) **(b)**

Figure 4: a) Causal graph for the UCI Adult dataset [9], b) Density of $\hat{Y}$; *orange* curve represents outcome density for a model trained on observed data with multi-level unfairness correction, *green curve* represents a counterfactual model accounting for only individual-level path specific effect, A, *yellow* represents a counterfactual model accounting for just the population-level, L path-specific effect, and *blue* represents the counterfactual model accounting for multi-level path-specific effects of both the individual and population level sensitive attributes, A, L.

individual level treatment. Since, L is represented by 6 levels[3] from 1 to 6, and A is represented by 2 levels 0 and 1, worst to best, we consider the baseline values to be $l = 6$ and $a = 1$. We compute the path-specific effect of L on Y via $L \rightarrow \cdots \rightarrow Y, A \rightarrow M \rightarrow \cdots \rightarrow Y$, and $A \rightarrow Y$ in the baseline scenario to be 10.65. After correcting for the unfair effect of L and A, the accuracy is 77.83% ($\beta = 1$), in comparison with an accuracy of 78.87% from removing the unfair effect of A alone. Thus, there is not significant drop in the model performance after correcting for the unfair effects of both A and L.

Next we study the residual unfairness in predictions while accounting for only unfair effects of individual, or only population-level sensitive attributes. The results are presented in Figure 4b. If densities of the predicted outcome $\hat{Y}$ coincide with the counterfactual model then the two models are counterfactually fair. The *orange* plot represents the fair predictions for a model trained on the observed data after correcting for multi-level unfairness of both A and L. In the process of learning the fair predictions, we learn the parameters for the model represented in Figure 4a as illustrated in algorithm 1 and then generate counterfactual samples from the observed distribution. First we generate individual-level counterfactual data by only altering the individual-level sensitive attribute *sex* ($a' = 1$) and train a model on the individual-level counterfactual data, we call this $\hat{Y}_{\text{fair}}(a_i')$. The *green* curve represents the density plot of $\hat{Y}_{\text{fair}}(a_i')$ after controlling for the individual-level path-specific effect at $\beta = 1$. Similarly, *yellow* and *purple* plots represent the density plots controlling for the path-specific effect of just the population-level sensitive attribute *level of education* ($l' = 6$), $\hat{Y}_{\text{fair}}\left(a_p'\right)$, and both individual and population-level sensitive attributes ($a' = 1, l' = 6$), $\hat{Y}_{\text{fair}}\left(a_i', a_p'\right)$, respectively. As can be seen from the overlap of the *orange* and *blue* curves, controlling for path-specific effects of both individual and population-level sensitive attributes generates counterfactually fair models. Thus, correcting the multi-level path specific effect does not marginally drop the model performance, while ensuring that predictions are fair with respect to individual and population-level attributes.

[3]We use pre-processed data from [38].

7 CONCLUSION

Our work extends algorithmic fairness to account for the multi-level and socially-constructed nature of forces that shape unfairness. In doing so, we first articulate a new definition of fairness, multi-level fairness. Multi-level fairness articulates a decision being fair towards the individual if it coincides with the one in the counterfactual world where, contrary to what is observed, the individual identifies with the advantaged sensitive attribute at the individual-level and also receives advantaged treatment at the population-level described by the population-level sensitive attribute. We illustrate the importance of accounting for population-level sensitive attributes by exhibiting residual unfairness if they are not accounted for.
Finally, we demonstrate a method for achieving fair outcomes by removing unfair path-specific effects with respect to both individual and population-level sensitive attributes. This works represents a step towards more comprehensively accounting for the multi-level unfairness, which has been identified as an important challenge in research across many fields. While a clear understanding of the population-level sensitive attributes is essential for multi-level fairness, approaches to learning the latent representation of the population-level sensitive attributes in their absence from individual-level factors could be important future work. As an initial framework, we consider path-specific effects for linear models here; we also endeavor to extend this to fewer assumptions on the functional form in the future.

REFERENCES

[1] C Avin, I Shpitser, and J Pearl. 2005. Identifiability of Path Specific Effects, In Proceedings of International Joint Conference on Artificial Intelligence. *Edinburgh, Scotland* 357 (2005), 363.

[2] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The Problem with Bias: From Allocative to Representational Harms in Machine Learning'. *Special Interest Group for Computing, Information and Society (SIGCIS)* 2 (2017).

[3] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018), 0049124118782533.

[4] Peter J Bickel, Eugene A Hammel, and J William O'Connell. 1975. Sex bias in graduate admissions: Data from Berkeley. *Science* 187, 4175 (1975), 398–404.

[5] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*. 405–414.

[6] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 514–524.

[7] Francesco Bonchi, Sara Hajian, Bud Mishra, and Daniele Ramazzotti. 2017. Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics* 3, 1 (2017), 1–21.

[8] Rhea W Boyd, Edwin G Lindo, Lachelle D Weeks, and Monica R McLemore. 2020. On racism: a new standard for publishing on racial health inequities. *Health Affairs Blog* 10 (2020).

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

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

[11] Sheryl L Coley, Tracy R Nichols, Kelly L Rulison, Robert E Aronson, Shelly L Brown-Jeffy, and Sharon D Morrison. 2015. Race, socioeconomic status, and age: exploring intersections in preterm birth disparities among teen mothers. *International journal of population research* 2015 (2015).

[12] Giuseppe De Arcangelis. 1993. Structural Equations with Latent Variables.

[13] Nicholas Diakopoulos, Sorelle Friedler, Marcelo Arenas, Solon Barocas, Michael Hay, Bill Howe, Hosagrahar Visvesvaraya Jagadish, Kris Unsworth, Arnaud Sahuguet, Suresh Venkatasubramanian, et al. 2017. Principles for accountable algorithms and a social impact statement for algorithms. *FAT/ML* (2017).

[14] Dustin T Duncan, Marcia C Castro, Steven L Gortmaker, Jared Aldstadt, Steven J Melly, and Gary G Bennett. 2012. Racial differences in the built environment—body mass index relationship? A geospatial analysis of adolescents in urban neighborhoods. *International journal of health geographics* 11, 1 (2012), 11.

[15] 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.

[16] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 1918–1921.

[17] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016).

[18] Tanis Furst, Margaret Connors, Carole A Bisogni, Jeffery Sobal, and Laura Winter Falk. 1996. Food choice: a conceptual model of the process. *Appetite* 26, 3 (1996), 247–266.

[19] Andrew Gelman. 2006. Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics* 48, 3 (2006), 432–435.

[20] Susan A Hall, Jay S Kaufman, and Thomas C Ricketts. 2006. Defining urban and rural areas in US epidemiologic studies. *Journal of urban health* 83, 2 (2006), 162–175.

[21] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 501–512.

[22] Jana A Hirsch, Kari A Moore, Tonatiuh Barrientos-Gutierrez, Shannon J Brines, Melissa A Zagorski, Daniel A Rodriguez, and Ana V Diez Roux. 2014. Built environment change and change in BMI and waist circumference: Multi-ethnic S tudy of A therosclerosis. *Obesity* 22, 11 (2014), 2450–2457.

[23] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2016. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559* 1, 2 (2016).

[24] Maximilian Kasy and Rediet Abebe. 2020. *Fairness, equality, and power in algorithmic decision making*. Technical Report. Working paper.

[25] 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.

[26] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[27] Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. 2016. Accountable algorithms. *U. Pa. L. Rev.* 165 (2016), 633.

[28] Yi-Lung Kuo, Alex Casillas, Kate E Walton, Jason D Way, and Joann L Moore. 2020. The intersectionality of race/ethnicity and socioeconomic status on social and emotional skills. *Journal of Research in Personality* 84 (2020), 103905.

[29] Matt Kusner, Chris Russell, Joshua Loftus, and Ricardo Silva. 2019. Making decisions that reduce discriminatory impacts. In *International Conference on Machine Learning*. 3591–3600.

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

[31] Matt J Kusner, Chris Russell, Joshua R Loftus, and Ricardo Silva. 2018. Causal Interventions for Fairness. *arXiv preprint arXiv:1806.02380* (2018).

[32] Moshe Lichman et al. 2013. UCI machine learning repository.

[33] Melissa D McCradden, Shalmali Joshi, Mjaye Mazwi, and James A Anderson. 2020. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health* 2, 5 (2020), e221–e223.

[34] Vishwali Mhasawade, Yuan Zhao, and Rumi Chunara. 2020. Machine Learning in Population and Public Health. *arXiv preprint arXiv:2008.07278* (2020).

[35] Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. 2018. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika* 105, 4 (2018), 987–993.

[36] Alan Mishler and Edward H Kennedy. 2020. Fairness in Risk Assessment Instruments: Post-Processing to Achieve Counterfactual Equalized Odds. *arXiv preprint arXiv:2009.02841* (2020).

[37] Giulio Morina, Viktoriia Oliinyk, Julian Waton, Ines Marusic, and Konstantinos Georgatzis. 2019. Auditing and Achieving Intersectional Fairness in Classification Problems. *arXiv preprint arXiv:1911.01468* (2019).

[38] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. 2019. Learning optimal fair policies. *Proceedings of machine learning research* 97 (2019), 4674.

[39] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. 2019. Optimal training of fair predictive models. *arXiv preprint arXiv:1910.04109* (2019).

[40] Judea Pearl. 2001. Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (Seattle, Washington) *(UAI'01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 411–420.

[41] Judea Pearl. 2012. The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prevention science* 13, 4 (2012), 426–436.

[42] Judea Pearl. 2013. Direct and indirect effects. *arXiv preprint arXiv:1301.2300* (2013).

[43] Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress* (2000).

[44] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference*. The MIT Press.

[45] Bilal Qureshi, Faisal Kamiran, Asim Karim, and Salvatore Ruggieri. 2016. Causal discrimination discovery through propensity score analysis.

[46] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. 2017. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*. 6414–6423.

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

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

[49] Tyler J VanderWeele and Whitney R Robinson. 2014. On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology (Cambridge, Mass.)* 25, 4 (2014), 473.

[50] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.

[51] Yixin Wang and David M Blei. 2019. Multiple causes: A causal graphical view. *arXiv preprint arXiv:1905.12793* (2019).

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

[53] Forest Yang, Moustapha Cisse, and Sanmi Koyejo. 2020. Fairness with Overlapping Groups. *arXiv preprint arXiv:2006.13485* (2020).

[54] Ke Yang, Joshua R Loftus, and Julia Stoyanovich. 2020. Causal intersectionality for fair ranking. *arXiv preprint arXiv:2006.08688* (2020).

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

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

[57] Tukufu Zuberi. 2000. Deracializing social statistics: Problems in the quantification of race. *The Annals of the American Academy of Political and Social Science* 568, 1 (2000), 172–185.