Causal Multi-Level Fairness

Vishwali Mhasawade
vishwalim@nyu.edu
Tandon School of Engineering
New York University
New York, USA

Rumi Chunara
rumi.chunara@nyu.edu
Tandon School of Engineering;
School of Global Public Health
New York University
New York, USA

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 macro (e.g. structural) levels, and it may be important to be fair to attributes at multiple levels. For example, instead of simply considering race as a causal, protected attribute of an individual, the cause may be distilled as perceived racial discrimination an individual experiences, which in turn can 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 macro-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 macro-level sensitive attributes are not accounted for, or included without accounting for their multi-level nature. Further, in the context of a real-world task of predicting income based on macro and individual-level attributes, we demonstrate an approach for mitigating unfairness, a result of multi-level sensitive attributes.

CCS CONCEPTS

• **Applied computing** → **Sociology**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

fairness; racial justice; social sciences

ACM Reference Format:
Vishwali Mhasawade and Rumi Chunara. 2021. Causal Multi-Level Fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3461702.3462587

AIES '21, May 19–21, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462587

1 INTRODUCTION

There has been much recent interest in designing algorithms that make fair predictions [3, 14, 29]. Definitions of algorithmic fairness have been summarized in detail elsewhere [40, 57]; broadly approaches to algorithmic 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 [16, 40] 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]. Historically sensitive variables such as age, gender, race have been thought of as attributes that can lead to unfairness or discrimination. Recently, the causal framework [48] has also been used to conceptualize fairness. Kusner et al. [32] introduced a causal definition of fairness, *counterfactual fairness*, which states that 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 of the individual had been different. This approach assumes prior knowledge about how data is generated and how variables affect each other, represented by a causal diagram and considers that all paths from the sensitive attribute to the outcome are unfair. Counterfactual fairness, thus, postulates that the entire effect of the sensitive attribute on the decision, along all causal paths from the sensitive attribute to the outcome is unfair. This approach is restrictive when only certain causal paths are unfair. For example, the effect of gender on income levels can be considered unfair but the effect through education attainment level is fair [9]. Path-specific counterfactual fairness has then been conceptualized, which attempts to correct the effect of the sensitive attribute on the outcomes *only* along unfair causal pathways instead of entirely removing the effect of the sensitive attribute [9]. It should be noted that throughout this work, determination of fair and unfair paths requires domain expertise and possibly discussions between policymakers, lawyers, and computer scientists.

Leveraging domain knowledge from sociology highlights that notwithstanding these clarifications of path-specific fairness, algorithmic fairness efforts in general have been limited to sensitive attributes at only the *individual-level*, missing important assertions regarding how social aspects can be influenced and created more by societal structures and cultural assumptions than by individual and psychological factors [27]. Critical Theory is an approach in sociology which motivates the consideration of macro-structural attributes, often simply called macro-properties or 'structural attributes'. These structural attributes correspond to overall organizations of society that extend beyond the immediate milieu of individual attributes yet have an impact on an individual, such as social groups, organizations, institutions, nation-states, and

their respective properties and relations [19, 35]. Accounting for a multi-level conceptualization in algorithmic fairness is consistent with recent literature which have discussed similar concerns with the conceptualization of race as a fixed, individual-level attribute [8, 22, 56, 64]. 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 macro-level phenomena associated with race, which a multi-level perspective promotes (e.g. structural racism, childhood experiences that shape the perception of race) will augment the impact of algorithmic fairness work [22]. Several different terms have been used as synonyms for macro-structural attributes, including "group-level" variables in multi-level analysis, "ecological variables", "macro-level variables", "contextual variables", "group variables", and "population variables", and we refer the reader to a full elaboration on macro-property variables in previous work by Diez-Roux [15]. In this paper, we refer to all information measured beyond an individual-level as *macro-properties* and *macro-level* attributes.

Without accounting for macro-properties, assessment of how "fair" an algorithm is may inadvertently be unfair by not accounting for important variance in attributes that are at the structural level. Indeed, today's growing availability of data that captures such effects means that leveraging the right data can be used to account for and address unexplained variance when using individual-level attributes as proxies for structural attributes. This can help work towards equity versus in-sample fairness (e.g. an algorithm fair with respect to perceived race will still be unfair to structural differences that perceived race may be a proxy for, such as neighborhood socio-economics, and including the structural factors can help to mitigate unfairness within the perceived race category) [38]. Accounting for structural factors can also clearly identify populations that input data is not representing. Indeed, many macro-properties can be considered as sensitive attributes. For example, neighborhood socioeconomic status which is a measure of the inequality in distribution of macro-level resources of education, work, and economic resources[52]. It has been shown that patients who reside in low socioeconomic neighborhoods can have significantly higher risk for readmission following hospitalization for sepsis in comparison to patients residing in higher socioeconomic neighborhoods even if all individual-level risk factors are the same, indicating that low SES is an independent factor of relevance to the outcome [20]. Thus, accounting for relevant macro-level properties while making decisions, say, about post-hospitalization care, is critical to ensure that disadvantaged patients, comprehensively accounting for all relevant sources of unfairness, receive appropriate care.

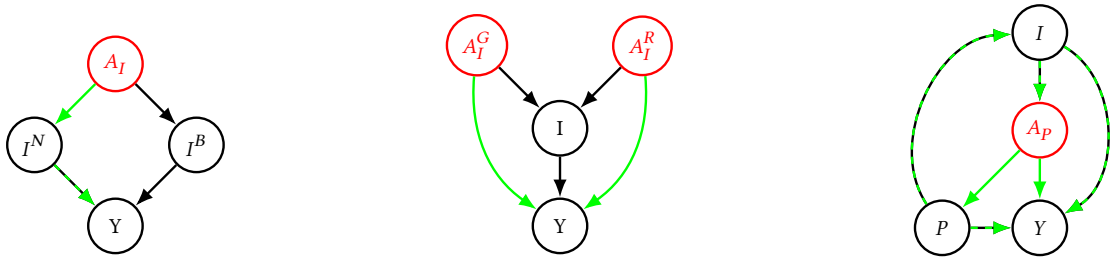
Towards this goal, we propose a novel definition of fairness called causal *multi-level fairness*, which is defined as a decision being fair towards an individual if it coincides with the one in the counterfactual world where, contrary to what is observed, the individual receives advantaged treatment at both the *macro* and *individual levels*, described by macro and individual-level sensitive attributes, respectively. This work builds upon previous work on path-specific counterfactual fairness, limited to individual-level sensitive attributes, to account for both macro 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 macro 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 [21]. 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 [26]. 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 (as is sometimes done as a proxy), we will miss the population-level factors that affect it and any particular outcome [56].

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 macro 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 [38]), it is critical to have a framework to assess these multiple sources of unfairness. Moreover, by including macro-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 macro and individual-level sensitive 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 macro-level attributes are not accounted for.
- We provide necessary conditions for achieving multi-level fairness and 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



(a) Proxy (I^N) for sensitive attributes (A_I) (b) Multiple sensitive attributes A_I^G, A_I^R (c)

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), red nodes represent sensitive attributes. 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. a biological factor) is not a proxy but also affects the outcome. 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) Example of cycles in the causal graph, where individual level attribute I, individual income, causes macro-level sensitive attribute, mean neighborhood income, A_P, which in turn affects education resources in the neighborhood, P, with Y being the outcome. Since $I \in pa_{A_P}$, this violates the necessary conditions for obtaining a fair predictor with multi-level causal fairness.

[4, 10, 18, 25, 28, 37]. 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, 26, 31–33, 43, 44, 50, 53, 62, 63]. Most of these approaches advocate for fairness by addressing an unfair causal effect of the sensitive attribute on the decision. Kusner et al. [32] 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, comprising of both individual and macro-level sensitive attributes. Inspired from research in social sciences [6, 22, 56], 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 macro-level, attributed as the macro-level sensitive attribute.

Identification of Causal Effects. While majority of the work in causal inference is on identification and estimation of the total causal effect [2, 45, 48, 49, 55], studies have also looked at identifying the causal effect along certain causal pathways [47, 54]. The most common approach for identifying the causal effect along different causal pathways is decomposing the total causal effect along direct and indirect pathways [46, 47, 54]. We leverage the approach developed by Shpitser [54] 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, [39, 58] can be used to extend our current contribution.

Path-specific causal fairness. Approaches in causal fairness such as path-specific counterfactual fairness [9, 43, 44], proxy discrimination, and unresolved discrimination [26, 59] 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 builds on the work of Chiappa [9] for mitigating unfairness by removing path-specific effects for fair predictions, doing so while accounting for both individual and macro-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 [17, 41, 60], 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 macro level, such as race and socioeconomic status [11, 30].

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) counterfactual fairness.

Causal Models: Following Pearl et al. [48] 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 $U_{pa_i} \subseteq \mathbf{U}$. Such equations relating V_i to pa_i and U_{pa_i} are known as structural equations [13]. When f_i is strictly linear, this is known as a linear structural equation model, which is represented as $V_i = \sum_{j=1}^{|pa_i|} \alpha_j \cdot pa_i[j] + U_{pa_i}$, where α is the mixing coefficient and $pa_i[j]$ denotes the jth parent of variable V_i.

Here, "pa_i" refers to the *causal parents* of V_i, i.e. the variables that affect the specific value V_i obtains. The joint distribution over all the n variables, $\Pr\left(V_1, V_2, \cdots, V_n\right)$ is given by the product of the conditional distribution of each variable, V_i, given it's causal parents pa_i as follows:

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

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 a direct cause of Y. Here, I is known as a *mediating variable* for both A_I^G and A_I^R.

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 known as a confounder, which may be unobserved, 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 $\Pr^*\left(Y|X=x\right) = \sum_Z \Pr\left(Y|X=x, Z\right)\Pr\left(Z\right)$, where $Z = \mathbf{V} \setminus \{X, Y\}$. The intervention $X = x$ results into a potential outcome, $Y_{X=x}$ (also represented as Y_x), where the distribution of the potential outcome variable is defined as $\Pr\left(Y_x\right) = \Pr^*\left(Y|X=x\right)$.

Counterfactual fairness: Counterfactual fairness is a causal notion of fairness that restricts the decisions of the algorithm to be invariant to the value assigned to the sensitive variable. Assuming A to be the sensitive attribute obtaining only two values, a and a', X as the remaining variables, and Y as the outcome of interest, counterfactual fairness is defined as follows:

Definition 3.1. *Predictor $\hat{Y}$ is **counterfactually fair** if under any context X and $A = a$,*

$$\Pr(\hat{Y}_{A\to a}(U) = y | X, A = a) = \Pr(\hat{Y}_{A\to a'}(U) = y | X, A = a) \tag{2}$$

With this background, we next discuss the problem setup.

4 PROBLEM SETUP

Let us denote all the variables associated with the system being modelled as $\mathbf{V} := \{A_I, A_P, \mathbf{I}, \mathbf{P}, Y\}$, where A_I and A_P are the sensitive attributes at the individual-level and the macro-level respectively[1], $\mathbf{I}$ is a non empty set of individual level variables, $\mathbf{P}$ is a non empty set of macro-level variables, and Y is an outcome of interest. We assume access to a causal graph, $\mathcal{G}$, consisting of $\mathbf{V}$ which accurately represents the data-generating process. Our goal is to obtain a classifier, $\hat{Y}_{\text{fair}}$, which is fair with respect to both A_P and A_I. While previous approaches in algorithmic fairness have only considered A_I as a sensitive attribute, we present a novel approach to also counter any discrimination resulting from A_P. While macro-level attributes, $\mathbf{P}$ and A_P, may seem like regular nodes in the causal graph, the difference is that based on them being macro-properties, their effects on the outcome may be mediated via the individual level attributes, $\mathbf{I}$ and A_I, making it nontrivial to mitigate unfairness due to these macro-properties. We highlight that this paper presents, to our knowledge, the first approach for incorporating unfairness due to macro-properties into algorithmic fairness. In order to further help guide the reader regarding differences in macro and individual-level variables, we clarify the types of macro-level attributes which are considered in our setting. Per the extant literature [15], macro-level attributes can be categorized into two types:

- **Category 1**: non-aggregate group properties like nationality, existence of certain types of regulations, population density, degree of income inequality in a community, political regime, legal status of women which do not include individual properties in their computation
- **Category 2**: aggregate properties such as neighborhood income and median household income, which include individual properties (here individual income) in their computation

Thus, individual variables influence aggregate macro-properties (Category 2) but not non-aggregate macro-properties (Category 1). This categorization of macro-properties helps us to identify the necessary conditions for obtaining a fair classifier with respect to both individual and macro-level sensitive attributes. In parallel, a main assumption in causal fairness and causal inference settings is that we have access to a causal graph and that the graph does not include any bidirected edges or cycles, [9, 32, 44] which are assumptions we maintain for the causal multi-level fairness setting. This assumption also clarifies that we should limit our consideration of macro-level attributes to Category 1 and not consider aggregate information

[1]Variables will be denoted by uppercase letters, V, values by lowercase letters, v, and sets by bold letters, $\mathbf{V}$.

like mean neighborhood income as macro-level sensitive variables since individual-level information has a causal influence on mean neighborhood income, i.e. $I \rightarrow A_P$. Certainly, an increase in the individual income increases mean neighborhood income, which can result in feedback loops in the causal graph which are disallowed. Figure 1c represents such a case involving cycles which the current framework does not support. Further, we assume that the structural equation model representing the functional form underlying the causal graph is linear in nature, i.e. a node is a linear combination of its causal parents. Future work should extend the analysis to non-linear relationships and non-parametric forms. We also assume that we are aware about the fair and unfair causal paths for the multi-level sensitive attributes, again consistent with previous work [9]. Next, we present an example to realize the importance of macro-level sensitive attributes.

4.1 An Illustrative Example

In prior sections we have motivated the need to consider macro-level attributes. We now describe multi-level sensitive attributes and provide an illustrative example of why considering macro-level sensitive attributes in algorithmic fairness efforts can be important. Consider neighborhood socioeconomic status (SES) as A_P. Neighborhood SES can be a cause of population-level factors such as resources available. Neighborhood SES can also affect individual-level attributes, A_I such as perception of race[2] (which is protected), by shaping the cultural and social experiences of an individual [12] as well as non-protected attributes, I (e.g. body-mass index, given that neighborhood SES can shape the types of resources available and norms [36]). Furthermore, research has described how macro attributes such as neighborhood SES can affect health behaviors like smoking directly as well as through individual factors such as perceived racial discrimination. Thus, given that this macro-attribute shapes the resources and circumstances of an individual (A_I can be influenced by A_P), it could be desirable for algorithms which are being used to decide fair allocation of health services or resources to predicted smokers, to account for this macro-property. Not being fair to this property lacks consideration for variance within the individual attributes that it influences. In other words, accounting for only individual-level attributes such as perceived racial discrimination can still lead to unfairness without considering that people within the same individual-level category could have different resources or opportunities available to them.

A general framework for macro/individual attribute relationships and possible unfair paths in presented in Figure 2. 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) [17, 41, 60, 61]. While this assumption of independence may hold for individual-level sensitive attributes like age and gender, it fails when we consider sensitive attributes at both the individual and macro level. It is important to understand the relationship between individual and macro-level attributes in order to delineate the path-specific effects that may lead to biased predictions of Y.

[2]We note that race is a social construct and henceforth use "perceived racial discrimination" here consistent with best practice [1].

Next, we describe how these interactions can lead to unfairness, by introducing multi-level path-specific fairness accounting for both individual and macro-level sensitive attributes.

5 MULTI-LEVEL PATH-SPECIFIC FAIRNESS

We begin the discussion about multi-level path-specific fairness by first introducing path-specific effects, discussing the necessary conditions to identify the effects with multiple sensitive attributes, and finally describe an approach to obtain fair predictions with multi-level path specific effects.

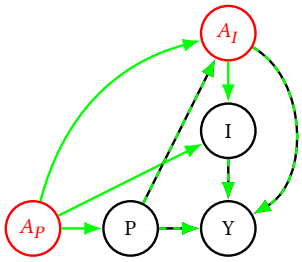


Figure 2: Multi-level sensitive attributes A_I, A_P. Macro-level variables, A_P (e.g. neighborhood SES), P (e.g. other zipcode level factors), and individual-level ones, A_I (e.g. perceived racial discrimination), I, affect the outcome Y (e.g. a health behavior).

5.1 Path-specific effects

While counterfactual fairness assumes that the entire effect of the sensitive variables on the outcomes is problematic and constructs the predictor with only the non-descendants of the sensitive attribute [32], it is a fairly restrictive scenario especially when the sensitive attributes can affect the outcome along both *fair* and *unfair* pathways as shown in Figure 2. Thus, disregarding the effect only along unfair paths can help in preserving the fair individual information regarding sensitive attributes [9]. Indeed, domain experts and policymakers can guide these decisions about identifying fair and unfair pathways. We are interested only in correcting the effect of the sensitive attribute on the outcome along the unfair paths. For this we turn to path-specific effects, which isolate the causal effect of a treatment on an outcome *only* along a certain causal pathway [45]. In order to obtain the causal effect only along a certain pathway, the idea that path-specific effects can be formulated as nested counterfactuals is leveraged [54]. This is done by considering that along the causal path of interest, say, $A_I \rightarrow I \rightarrow Y$, the variable, A_I, propagates the observed value, for example, a_i in Figure 2, and along other pathways, $A_I \rightarrow Y$, the counterfactual value, a_i' is propagated, assuming that $A_I = \{a_i, a_i'\}$. In this case, the effect of any mediator between A_I and Y is evaluated according to the corresponding value of A_I propagated along the specific causal path, the mediator I obtains the value $I(a_i)$. Here, the primary object of interest is the potential outcome variable, $Y(a_i)$, which represents the outcome if, possibly contrary to fact, A_I were externally set to value a_i. Given the values of a_i, a_i', comparison of $Y(a_i')$ and $Y(a_i)$ in expectation: $\mathbb{E}[Y(a_i)] - \mathbb{E}[Y(a_i')]$ would allow us to quantify the path-specific causal effect, PSE, of A_I on Y.

Using this concept of path-specific effect, we next state path-specific counterfactual fairness, defined by Chiappa [9].

Definition 5.1. *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 (denoted by π) was set to the counterfactual value.*

$$\Pr(\hat{Y}_{A\to a_\pi}(U) = y|X, A = a) = \Pr(\hat{Y}_{A\to a'_\pi}(U) = y|X, A = a) \quad (3)$$

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, i.e. $\hat{Y}_{\text{fair}} = \hat{Y} - \text{PSE}$. However, a key limitation of the approach presented by Chiappa [9] is that only the path-specific counterfactual fairness with respect to the sensitive attributes at the individual level is considered. The presence of sensitive attributes at the macro-level that affect properties at the individual level makes it non-trivial to estimate the path-specific effect consisting of both macro and individual-level sensitive attributes. Following, we describe the identifiability conditions for obtaining path-specific counterfactual fairness with respect to multi-level sensitive attributes by estimating multi-level path-specific effects.

Identification of multi-level path-specific effects.

Proposition 1. *In the absence of any unmeasured confounding between A_P and A_I and $\{A_I, \boldsymbol{I}\} \notin pa_{A_P}$, 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, where A_P is a non-aggregate macro-property not involving any individual-level computation. Next, we discuss the interaction between A_P and A_I, where the macro-level sensitive attribute is a causal parent of the individual-level sensitive attribute, $A_P \in \{pa_{A_I}\}$.

Figure 2 presents such an exemplar case of multi-level interactions where macro-level sensitive attributes can cause individual level variables, with the data generating process as described in Figure 3a. 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. 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) $A_I \to Y$. With prior assumption that $A_I \to Y$ and $P \to Y$ are not discriminatory causal paths, we evaluate the path-specific effect along $\cdots \to I \to Y$, where $\cdots \to I$ repesents all paths into I . Overall the path-specific effect can be calculated as follows:

$$\text{PSE}^{\cdots\to I\to Y}_{A_P\leftarrow a_p, A_I\leftarrow a_i} = \mathbb{E}\left[Y\left(a'_i, P\left(a'_p\right), I\left(a_i, a_p\right)\right)\right]. \quad (4)$$

The path-specific effect along $\cdots \to I \to Y$ 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 the counterfactual and 0 in the original. The counterfactual distribution can be estimated as follows:

$$\begin{aligned} &Y\left(a'_i, P\left(a'_p\right), I\left(a_i, a_p\right)\right) \\ &= \int_{P,I} \Pr\left(Y|a'_i, I, P\right) \Pr\left(I|a_i, a_p\right) \Pr\left(P|a'_p\right). \end{aligned} \quad (5)$$

We obtain the mean of this distribution as follows:

$$\begin{aligned} &\mathbb{E}[Y|I, P, a'_i] \\ &= \theta^y + \theta^y_i\theta^i + \theta^y_p\theta^p + \theta^y_{a_i}a'_i + \theta^y_i\theta^i_{a_i}a_i + \theta^y_i\theta^i_{a_p}a_p + \theta^y_p\theta^p_{a_p}a'_p. \end{aligned} \quad (6)$$

$\text{PSE}^{\cdots\to I\to Y}_{A_P\leftarrow a_p, A_I\leftarrow a_i}$ is then obtained by comparing the observed and counterfactual value of the mean of the counterfactual distribution,

$$\begin{aligned} \mathbb{E}[Y|a_p, a_i] - \mathbb{E}[Y|a'_p, a'_i] &= \theta^y_{a_i}\left(a'_i - a'_i\right) + \theta^y_i\theta^i_{a_i}\left(a_i - a'_i\right) \\ &\quad + \theta^y_i\theta^i_{a_p}\left(a_p - a'_p\right) + \theta^y_p\theta^p_{a_p}\left(a'_p - a'_p\right) \\ &= \theta^y_i\theta^i_{a_i}\left(a_i - a'_i\right) + \theta^y_i\theta^i_{a_p}\left(a_p - a'_p\right). \end{aligned} \quad (7)$$

Substituting $a_i = 1, a_p = 1, a'_i = 0, a'_p = 0$ in Equation 7 we obtain the path-specific effect as a function of the parameters θ[3]:

$$\text{PSE}^{\cdots\to I\to Y}_{A_P\leftarrow a_p, A_I\leftarrow a_i} = \theta^y_i\theta^i_{a_i} + \theta^y_i\theta^i_{a_p}. \quad (8)$$

By the same approach, path-specific effects for individual and macro-level sensitive attributes as well as the multi-level path specific effects for all the causal paths in Figure 2, are also evaluated. Estimating the multi-level path-specific effects helps in constructing a fair predictor which is described in the following section.

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 $\hat{Y}_{\text{fair}} = \mathbb{E}\left[f_Y\left(\theta^Y\right)\right] - \beta * \text{PSE}_\pi$

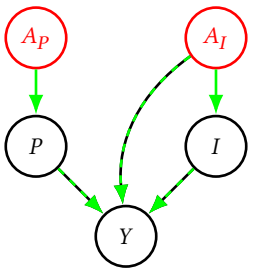
5.2 Fair predictions with multi-level path-specific effects.

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

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

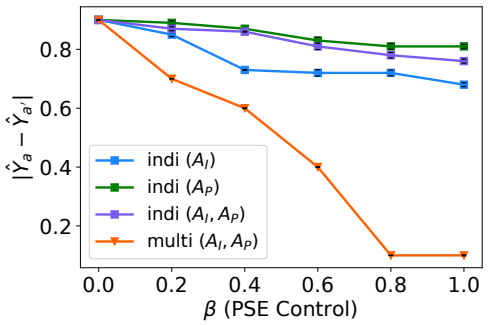
$$A_P \sim \text{Bernoulli}\,(p = 0.5)\,,$$
$$P = \theta^{p} + \theta^{p}_{a_p} A_P + \epsilon_p,$$
$$\gamma = \theta^{a_i} + \theta^{a_i}_{a_p} A_P + \theta^{a_i}_{p} P + \epsilon_{a_i},$$
$$A_I = \text{Bernoulli}(\sigma(\gamma)),$$
$$I = \theta^{i} + \theta^{i}_{a_p} A_P + \theta^{i}_{a_i} A_I + \epsilon_i,$$
$$Y = \theta^{y} + \theta^{y}_{p} P + \theta^{y}_{i} I + \theta^{y}_{a_i} A_I + \epsilon_y$$
$$\theta^{p}, \theta^{a_i}, \theta^{i}, \theta^{y} = (0.2, 0.2, 0.2, 0.2)$$
$$\theta^{p}_{a_p}, \theta^{a_i}_{a_p}, \theta^{a_i}_{p}, \theta^{i}_{a_p}, \theta^{i}_{a_i} \sim U\,(0.2, 0.7)$$
$$\theta^{y}_{p}, \theta^{y}_{i}, \theta^{y}_{a_i} \sim U\,(0.3, 0.95)$$
$$\epsilon_p, \epsilon_{a_i}, \epsilon_i, \epsilon_y \sim \mathcal{N}\,(0, 1)$$
$$\sigma = 1/(1 + \exp\,(-x))$$

(a)



A_P	$\theta^{y}_{p}\theta^{p}_{a_p}$
A_P, A_I	$\theta^{y}_{p}\theta^{p}_{a_p} + \theta^{y}_{i}\theta^{i}_{a_i}$

(b)



(c)

Figure 3: Experimental details regarding multi-level example (Figure 2), a) Data generating process, b) Resulting causal graph if A_P is considered as an individual-level sensitive attribute instead of macro-level sensitive attribute as presented in Figure 2, $A_P \notin pa_{A_I}$, resulting path-specific effects with A_P as individual-level sensitive attribute, and c) Path-specific unfairness controlling for discriminatory effects of just A_I (blue), A_P considered as an individual-level sensitive attribute instead of macro-level sensitive attribute (green), both A_I, A_P considered as individual-level sensitive attributes (purple), and A_I, A_P considered as multi-level sensitive attributes (orange), over 10 runs.

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 $\cdots \rightarrow I \rightarrow Y$, $\text{PSE}^{\cdots \rightarrow I \rightarrow Y}_{A_P \leftarrow a_p, A_I \leftarrow a_i} = \theta^{y}_{i}\left(\theta^{i}_{a_i} + \theta^{i}_{a_p}\right)$, we can control for the discrimination at test time as follows:

$$\hat{y}^{n}_{\text{fair}} = \theta^{y} + \theta^{y}_{p} p^{n} + \theta^{y}_{i} i^{n} + \theta^{y}_{a} a^{n}_{i} - \beta\left(\theta^{y}_{i}\left(\theta^{i}_{a_i} + \theta^{i}_{a_p}\right)\right) \quad (9)$$

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. We study the effect of the control over different path-specific effects on the performance metric of the classifier. The steps are outlined in Algorithm 1. We highlight that an accurate causal graph and a linear functional form are key assumptions of this approach.

6 EXPERIMENTS

Our experiments in this section serve to empirically assess and demonstrate residual unfairness when correcting for path-specific effects of multiple sensitive attributes. We consider a synthetic setting demonstrated in Figure 2 and real-world setting for income prediction presented in Figure 4a; in both cases the structural causal model is known a priori as well as the unfair pathways. 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 path-specific counterfactual distribution, referred to as *counterfactual distribution*[4]. 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 signifies residual unfairness.

6.1 Synthetic setting

Here we generate data according to the structural equation model presented in Figure 3a with linear relationships with the exception that we consider Y to be binary, using a sigmoid function [5]. This represents the data-generating process in accordance with the multi-level setting presented in Figure 2. The simulation parameters, θ, are non-negative to ensure a positive path-specific effect, i.e. an unfair effect. Furthermore, the parameters are standardized between 0 and 1 to allow for computational efficiency. The range of the parameters in the uniform distribution is simulated to ensure a realistic setting where the counterfactual distribution is diverse, and the differences between the original and counterfactual distribution is significant to evaluate multi-level fairness. If there are little differences in the original and the counterfactual distribution, path-specific effects become negligible resulting in no significant unfairness issues. In total, 10 simulations with $N = 2000$ samples in each are performed. We train a linear regression model for I, P, and a logistic regression

[4] Path-specific counterfactual distribution is obtained by altering the original value to the counterfactual value only along the unfair paths (described by green in the causal graph), and retaining the original value along the fair pathways.

[5] The analysis of the path-specific effects remains similar to the one discussed in Section 5, where the effects are calculated on the odds ratio scale as Nabi and Shpitser [44].

model for A_I, Y to learn the parameters (θ). We then calculate three path-specific effects as follows: 1) PSE_{A_P,A_I} accounting for the effect of both A_P and A_I, 2) PSE_{A_I} accounting for the effect of only A_I, and 3) PSE_{A_P} accounting for the effect of only A_P. Details about this calculation are presented in Equation 7.

We observe that controlling for the multi-level path-specific effects of both A_P and A_I leads to little drop in the model performance (accuracy). We also assess the counterfactual fairness of the resulting model after removing the unfair multi-level path-specific effects. This is done by training a classifier for obtaining fair predictors, $\hat{y}_{a_i,a_p}$, on the actual and counterfactual data respectively, where the counterfactual data is obtained by altering the values of the sensitive variables, A_P and A_I. This is done by first learning the parameters, θ from the original data and then altering the values of the sensitive attributes, A_P and A_I and obtaining the values of P, I and Y from these counterfactual values of A_P, A_I and θ. We assess the counterfactual fairness of the resulting models while accounting for unfairness due to both A_P, A_I and either A_I or A_P, in Figure 3c. 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. In order to assess the advantage of considering the multi-level nature of the data, we compare with the model presented in Figure 3b, where we treat the macro-level sensitive attribute, A_P as another individual-level sensitive attribute, A_I, i.e. $A_P \notin pa_{A_I}$. 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. Similarly, the *green* line represents the setting presented in Figure 3b, where A_P, is treated as an individual-level sensitive attribute (multi-level nature of its action removed), and *purple* line represents the case where both A_I and A_P are treated as individual-level sensitive attributes. 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. Thus, accounting for individual-level path-specific effects solely (blue, green, and purple lines) does not result in counterfactually fair predictions as can be seen from the high difference (0.7- 0.9). Moreover, 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 [34]. 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, 44] 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 and Shpitser [44], we do not include race or capital gain and loss variables. Here, A represents the individual-level protected attribute sex, C is 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 nationality, C, could be considered as a non-aggregate macro-level sensitive attribute. This is because it may affect social influences and/or resources available at the macro-level. For example, nationality affects both education levels and martial status of individuals. Moreover, we also note that marital status, M, can also be considered as an individual-level sensitive attribute, as we may not want to discriminate individuals based on their marital status [24, 51]. Thus, C, a macro-level sensitive attribute affects M, an individual-level sensitive attribute and the relationship between C and other variables in the graph mimics 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 nationality, in order to understand the effect that nationality may have. We highlight that macro-properties such as nationality are included in datasets like UCI adult, however, correcting for any unfairness due to macro-level sensitive attributes has not been considered in previous fairness studies. We also emphasize that there may be other datasets that better express the need for macro and individual-level sensitive attributes, however we consider this dataset/setting in order to be consistent with 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 C as the macro-level sensitive attribute and A, M as the individual-level sensitive attributes.

To first evaluate the path-specific effect of a single individual-level sensitive attribute as done in prior work [9], we evaluate the effect of only A (the individual level sensitive attribute considered in Chiappa [9]) along $A \to Y$ and $A \to M \to \cdots \to Y$ to be 3.716. We obtain a fair prediction as follows:

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

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 macro (C) and individual-level sensitive attributes (A, M)[6]. In the counterfactual world, the individual receives advantaged macro and individual level treatment. Since, C is mostly represented by the value *39*, and A is represented by 2 levels 0 and 1, worst to best, we consider the baseline values to be $c = 39$ and $a = 1$. We compute the path-specific effect of C on Y via $C \to M \to \cdots \to Y$, $C \to L \to \cdots \to Y$, $A \to M \to \cdots \to Y$, and $A \to Y$ in the baseline scenario to be 10.65. As we do not completely remove the effect of C; only along certain paths as described above, we still retain important information about nationality, C, an informative feature. After correcting for the unfair effect of C, A, and M, the accuracy is 77.92% ($\beta = 1$),

[6]We use pre-processed data from [42].

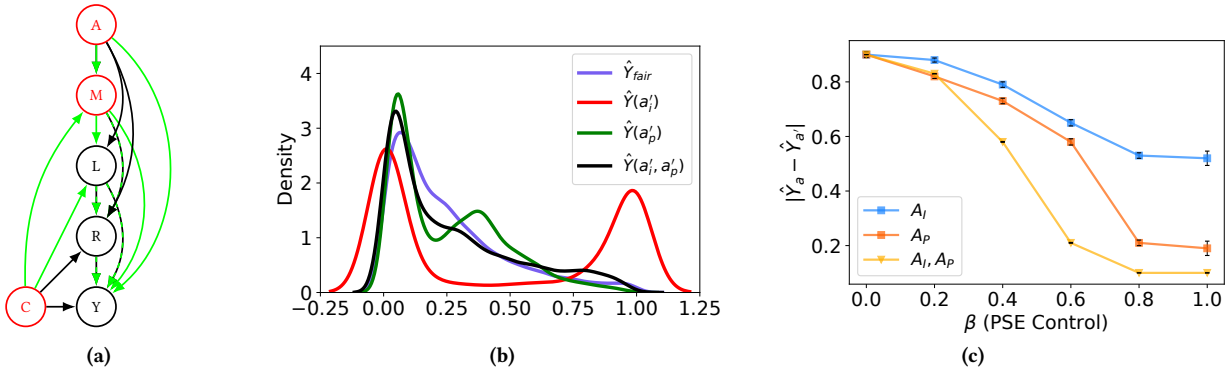


Figure 4: a) Causal graph for the UCI Adult dataset [9], A and M represent the individual-level protected attributes, sex and marital status, respectively, C is age and nationality, L is the level of education, R corresponds to working class, occupation, and hours per week, Y is the income class, unfair paths are represented in green b) Density of $\hat{Y}$, c) path-specific unfairness, $|\hat{Y}_a - \hat{Y}_{a'}|$ controlling for the effects of just A_I (blue), A_P (orange) and both A_I, A_P (yellow).

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, M and C.

Next, we study the residual unfairness in a series of counterfactual settings which account for unfairness at individual, macro-level sensitive attributes and their combination. Results are presented in Figure 4b. If density of two plots coincide, then they are counterfactually fair, i.e. there is no difference in prediction based on unfair pathways. The *purple* plot represents the fair model, $\hat{Y}_{\text{fair}}$, trained on the original data that corrects for the multi-level path-specific effects of $\{A, M, C\}$. 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. We next assess fairness of models that account for varying fair/unfair pathways and variable types. First, we generate individual-level counterfactual data by only altering the individual-level sensitive attributes A and M along the unfair pathways, and train a model on the individual-level counterfactual data, we call this $\hat{Y}\left(a_i'\right)$. The *red* curve represents the density plot of $\hat{Y}\left(a_i'\right)$ after controlling for the individual-level path-specific effect at $\beta = 1$. Similarly, we generate macro-level counterfactual data by only altering the macro-level sensitive attribute, C, along the unfair paths, $C \rightarrow M \rightarrow \cdots, C \rightarrow L \rightarrow \cdots$, and train a model on this macro-level counterfactual data, which is represented by $\hat{Y}(a_p')$ (*green* curve). At last, we generate individual and macro-level counterfactual data, by altering A, M and C along all unfair paths, represented by green in Figure 4a. In this case, the predictor, $\hat{Y}(a_i', a_p')$, is represented by the *black* curve, As can be seen from the general overlap of the *purple* and *black* curves, controlling for path-specific effects of both individual and macro-level sensitive attributes generates counterfactually fair models, while only controlling for the individual or macro variables doesn't provide the same density estimate as a fair model. Similarly, the residual unfairness is presented in Figure 4c, and controlling for the multi-level path-specific effects (yellow curve) results in a counterfactual fair model as the difference, $|\hat{Y}_a - \hat{Y}_{a'}|$ approaches zero. 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 macro-level attributes.

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 macro-level described by the macro-level sensitive attribute. A framework like this can be used to assess unfairness at each level, and identify the places for intervention that would reduce unfairness best (e.g. via macro-level policies versus individual attributes). As we show here, leveraging datasets that represent more than individual attributes can improve fairness, as often individual level attributes have variance with respect to macro-properties, or are proxies for macro-properties which may be the critical sources of unfairness [38]. This is becoming possible given the large number of open data sets representing macro-attributes that are available, relevant to health, economics, and many other settings. Through our experiments, we illustrate the importance of accounting for macro-level sensitive attributes by exhibiting residual unfairness if they are not accounted for. Importantly, we show this residual unfairness persists even in cases when the same information and variables are considered, but without accounting for the multi-level nature of their interaction.

Finally, we demonstrate a method for achieving fair outcomes by removing unfair path-specific effects with respect to both individual and macro-level sensitive attributes. While a clear understanding of the macro-level sensitive attributes is essential for multi-level fairness, approaches to learning the latent representation of the

macro-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.

ACKNOWLEDGMENTS

We acknowledge funding from the National Science Foundation, award number 1845487. We also thank Harvineet Singh for helpful discussions.

REFERENCES

[1] Maurianne Ed Adams, Lee Anne Ed Bell, and Pat Ed Griffin. 2007. *Teaching for diversity and social justice*. Routledge/Taylor & Francis Group.
[2] 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.
[3] 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).
[4] 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.
[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] Amy B Dailey, Stanislav V Kasl, Theodore R Holford, Tene T Lewis, and Beth A Jones. 2010. Neighborhood-and individual-level socioeconomic variation in perceptions of racial discrimination. *Ethnicity & health* 15, 2 (2010), 145–163.
[13] Giuseppe De Arcangelis. 1993. Structural Equations with Latent Variables.
[14] 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).
[15] Ana V Diez-Roux. 1998. Bringing context back into epidemiology: variables and fallacies in multilevel analysis. *American journal of public health* 88, 2 (1998), 216–222.
[16] 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.
[17] 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.
[18] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016).
[19] Brian Furze and Pauline Savy. 2014. *Sociology in today's world*. Cengage AU.
[20] Panagis Galiatsatos, Amber Follin, Fahid Alghanim, Melissa Sherry, Carol Sylvester, Yamisi Daniel, Arjun Chanmugam, Jennifer Townsend, Suchi Saria, Amy J Kind, et al. 2020. The Association Between Neighborhood Socioeconomic Disadvantage and Readmissions for Patients Hospitalized With Sepsis. *Critical care medicine* 48, 6 (2020), 808–814.
[21] Andrew Gelman. 2006. Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics* 48, 3 (2006), 432–435.
[22] 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.
[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] Courtney G Joslin. 2015. Marital status discrimination 2.0. *BUL Rev.* 95 (2015), 805.
[25] Maximilian Kasy and Rediet Abebe. 2021. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 576–586.
[26] 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.
[27] Joe L Kincheloe and Peter McLaren. 2011. Rethinking critical theory and qualitative research. In *Key works in critical pedagogy*. Brill Sense, 285–326.
[28] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).
[29] 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.
[30] 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.
[31] Matt Kusner, Chris Russell, Joshua Loftus, and Ricardo Silva. 2019. Making decisions that reduce discriminatory impacts. In *International Conference on Machine Learning*. 3591–3600.
[32] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in neural information processing systems*. 4066–4076.
[33] Matt J Kusner, Chris Russell, Joshua R Loftus, and Ricardo Silva. 2018. Causal Interventions for Fairness. *arXiv preprint arXiv:1806.02380* (2018).
[34] Moshe Lichman et al. 2013. UCI machine learning repository.
[35] William Little, Ron McGivern, and Nathan Kerins. 2016. *Introduction to Sociology-2nd Canadian Edition*. BC Campus.
[36] Charu Mathur, Darin J Erickson, Melissa H Stigler, Jean L Forster, and John R Finnegan Jr. 2013. Individual and neighborhood socioeconomic status effects on adolescent smoking: a multilevel cohort-sequential latent growth analysis. *American Journal of Public Health* 103, 3 (2013), 543–548.
[37] 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.
[38] Vishwali Mhasawade, Yuan Zhao, and Rumi Chunara. 2020. Machine Learning in Population and Public Health. *arXiv preprint arXiv:2008.07278* (2020).
[39] 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.
[40] Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 2021. Fairness in Risk Assessment Instruments: Post-Processing to Achieve Counterfactual Equalized Odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3442188.3445902
[41] 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).
[42] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. 2019. Learning optimal fair policies. *Proceedings of machine learning research* 97 (2019), 4674.
[43] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. 2019. Optimal training of fair predictive models. *arXiv preprint arXiv:1910.04109* (2019).
[44] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[45] 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.
[46] Judea Pearl. 2012. The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prevention science* 13, 4 (2012), 426–436.
[47] Judea Pearl. 2013. Direct and indirect effects. *arXiv preprint arXiv:1301.2300* (2013).
[48] Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress* (2000).
[49] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference*. The MIT Press.
[50] Bilal Qureshi, Faisal Kamiran, Asim Karim, and Salvatore Ruggieri. 2016. Causal discrimination discovery through propensity score analysis.
[51] Cordelia W Reimers. 1985. Cultural differences in labor force participation among married women. *The American Economic Review* 75, 2 (1985), 251–255.
[52] Catherine E Ross and John Mirowsky. 2008. Neighborhood socioeconomic status and health: context or composition? *City & Community* 7, 2 (2008), 163–179.
[53] 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.

[54] Ilya Shpitser. 2013. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive science* 37, 6 (2013), 1011–1035.
[55] Jin Tian and Judea Pearl. 2002. A General Identification Condition for Causal Effects. In *Eighteenth National Conference on Artificial Intelligence*. American Association for Artificial Intelligence.
[56] 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.
[57] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.
[58] Yixin Wang and David M Blei. 2019. Multiple causes: A causal graphical view. *arXiv preprint arXiv:1905.12793* (2019).
[59] 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.
[60] Forest Yang, Mouhamadou Cisse, and Sanmi Koyejo. 2020. Fairness with Overlapping Groups; a Probabilistic Perspective. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.).
[61] Ke Yang, Joshua R Loftus, and Julia Stoyanovich. 2020. Causal intersectionality for fair ranking. *arXiv preprint arXiv:2006.08688* (2020).
[62] Junzhe Zhang and Elias Bareinboim. 2018. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*. 3671–3681.
[63] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *Proceedings of the... AAAI Conference on Artificial Intelligence*.
[64] 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.