On the Misalignment Between Legal Notions and Statistical Metrics of Intersectional Fairness

Deborah Dormah Kanubala[1], **Isabel Valera**[1,2]

[1]Saarland University
[2]Max Planck Institute for Software Systems
kanubala@cs.uni-saarland.de, ivalera@cs.uni-saarland.de

Abstract

Intersectional (un)fairness, as conceptualized in legal and social theory, emphasizes the *non-additive and structurally complex nature* of discrimination against individuals at the intersection of multiple sensitive attributes (such as race, gender, etc). Recent works have proposed statistical metrics for intersectional fairness by estimating disparities across groups of individuals sharing two or more sensitive attributes. However, it is unclear if these metrics detect *uniquely* intersectional discrimination. We therefore pose the following question, *Do current statistical intersectional metrics detect the non-additive discrimination highlighted by intersectionality theory?* More specifically, to answer this, we run controlled synthetic data experiments that explicitly allow us to control for single, multiple, intersectional, and compounded forms of discrimination. Our analyses show that current statistical metrics for intersectional fairness behave more like multi-attribute disparity measures. Specifically, they respond more strongly to additive or compounded biases than to non-additive interaction effects. While they effectively capture disparities across multiple sensitive attributes, they often fail to detect uniquely intersectional discrimination. These findings reveal a fundamental misalignment between existing intersectional fairness metrics and the legal and theoretical foundations of intersectionality. We argue that if intersectional fairness metrics are to be deemed truly intersectional, they must be explicitly designed to account for the structural, non-additive nature of intersectional discrimination.

1 Introduction

Fairness has garnered significant attention in the machine learning (ML) community, leading to the development of several fairness notions and mitigation approaches (Dwork et al. 2012; Mehrabi et al. 2021; Hardt, Price, and Srebro 2016). While substantial progress has been made, much of this work has focused on single sensitive attributes such as gender or race (Dwork et al. 2012; Chouldechova 2017; Hardt, Price, and Srebro 2016; Zafar et al. 2017). However, real-world discrimination spans multiple dimensions, as individuals often hold intersecting identities. Therefore, discrimination cannot be classified according to, or defined by, a single sensitive attribute (Uccellari 2008; Makkonen 2002; Xenidis 2018), thereby leading to what is formally described as *multidimensional discrimination*.

Broadly, multidimensional discrimination can be grouped into three main types: i) *Sequential discrimination*, which occurs when a person suffers discrimination either on the same or different grounds over a period of time (Roy, Horstmann, and Ntoutsi 2023); ii) *Multiple discrimination* occurs when individuals face separate and additive discrimination resulting from their sensitive attributes (Alvarez and Ruggieri 2025; Mercat-Bruns 2017); and iii) *Intersectional discrimination*, by contrast, arises when the interaction of different sensitive attributes leads to a unique form of discrimination that cannot be explained by simply adding discrimination across sensitive attributes (Crenshaw 2013; Council of Europe 2025). For example, Turkish women who wear a hijab face higher rejection rates in job applications, illustrating discrimination that arises specifically from the intersection of ethnicity, gender, and religion (Weichselbaumer 2020).

While these types of multidimensional discrimination have sparked conversation in the legal and social theory (Collins 2015; Uccellari 2008; Cho, Crenshaw, and McCall 2013; Makkonen 2002), the machine learning community has only recently begun to study discrimination in this area. Seminal works such as Buolamwini and Gebru (2018); Kearns et al. (2018) showed that models considered fair with respect to a single sensitive attribute can still show unfairness when considering fairness across several sensitive attributes (Chen et al. 2024). This has prompted a shift in ML fairness research toward measuring and mitigating disparities across subgroups (Kearns et al. 2018).

Nevertheless, most of this literature fails to distinguish among different forms of multidimensional discrimination, often referring to all of them as *intersectional fairness* (Foulds et al. 2020; Morina et al. 2019; Gohar and Cheng 2023). To the best of our knowledge, Alvarez and Ruggieri (2025) is one of the few works that empirically attempt to reason about different sources of discrimination, including multiple and intersectional discrimination. However, their work does not clearly articulate or isolate intersectional discrimination as non-additive effects, as conceptualised in legal theory. Complementing this, Kong (2022) provides an important philosophical analysis, highlighting three key problems in how intersectional fairness has been

operationalised in AI research.

These critiques prompt us to explicitly address this gap and test whether current metrics are sensitive to non-additive discrimination, which is the core of intersectionality as originally conceptualised in legal and social theory. This is because current metrics risk conflating intersectional discrimination with multiple discrimination.

In other words, we aim to respond to the following research question: *Do current intersectional fairness metrics detect the non-additive discrimination that defines intersectionality, or are they mainly measuring additive disparities across subgroups?* We empirically approach answering this question by creating a synthetic dataset where we have *explicit control of the different bias scenarios*, thus, i) no bias, ii) single-axis bias, iii) multiple or additive bias, iv) intersectional bias (includes only the interaction of the sensitive attributes), and v) compounded bias (includes both additive and interaction of the sensitive attributes). We formally define these distinct types of multidimensional discrimination, with particular focus on multiple and intersectional discrimination, as these are most commonly conflated in practice. In contrast, sequential discrimination can often be addressed using existing single-attribute fairness measures, applied independently at each time step (Roy, Horstmann, and Ntoutsi 2023). We then evaluate using statistical intersectional fairness metrics, including subgroup (un)fairness (Kearns et al. 2018), smoothed differential (un)fairness (Foulds et al. 2020), slift and elift (Morina et al. 2019). Our results show that while these metrics detect subgroup disparities when present, they are far more sensitive to multiple discrimination than to intersectional discrimination. In addition, the smoothed ϵ-differential fairness metric, which is originally designed to measure and mitigate intersectional unfairness, fails to distinguish additive from purely intersectional effects.

Our contributions, therefore, are that we:

1. Propose a formal definition of multidimensional discrimination, distinguishing single, multiple, intersectional, and compounded discrimination scenarios in Section 3.
2. Provide a synthetic dataset generation framework that enables controlled evaluation of intersectional fairness metrics in Section 4.
3. Conduct the first systematic empirical study of intersectional metrics against legal definitions of intersectionality in Section 5.
4. Show that existing metrics conflate additive and intersectional effects, highlighting the need for future metrics to explicitly account for intersectional discrimination in Section 5 and Section 6 respectively.

Significance. Intersectionality, as defined in legal and social theory, emphasises that individuals with overlapping sensitive attributes face unique, non-additive forms of discrimination. This form of discrimination cannot be explained by simply adding discrimination resulting from a single sensitive attribute discrimination (Crenshaw 2013; Collins and Bilge 2020). As such, intersectional fairness metrics that fail to capture these non-additive effects risk underestimating the severity of discrimination and misguiding mitigation efforts. Hence, it is critical that proposed intersectional metrics align with the structural understanding of discrimination grounded in legal theory and reflected in real-world experiences.

2 Related Work

Fairness in machine learning. The recognition that algorithmic systems can perpetuate or amplify social biases (Dwork et al. 2012; Barocas and Selbst 2016; Mehrabi et al. 2021) has led to the rapid evolution of the field in the last few years. Initial work has primarily focused on statistical notions of group fairness, including demographic parity (Dwork et al. 2012), equalised odds (Hardt, Price, and Srebro 2016), and predictive parity (Chouldechova 2017). Other works explored individual fairness, emphasising that similar individuals should receive similar outcomes (Dwork et al. 2012). While these approaches laid the foundation, they mostly considered fairness with respect to a single sensitive attribute (Zafar et al. 2017; Kanubala, Valera, and Gupta 2024). However, Kearns et al. (2018) showed that models that tend to be fair for single sensitive attributes may remain unfair across subgroups.

Single-attribute to intersectional fairness. Recognising that fairness constraints based on single sensitive attributes can overlook discrimination at the intersection of identities, thus, Kearns et al. (2018) proposed subgroup (un)fairness. This approach requires that a model perform fairly well across all sufficiently large subgroups defined by combinations of attributes. In addition, other approaches have either proposed new intersectional metrics (Suk and Han 2023; Maheshwari et al. 2023) or a way to learn fair models given more than one sensitive attribute (Choi, Karumbaiah, and Matayoshi 2025; Němeček et al. 2025; Lin, Gupta, and Jagadish 2024). Despite these advances, intersectionality still faces challenges and amongst them is the small sample size subgroup problem (Himmelreich et al. 2024). A lot of intersectional groups have limited data, making reliable fairness assessments difficult. To address these, methods such as Foulds et al. (2020) ϵ-differential fairness and other smoothing approaches (Morina et al. 2019) have been proposed. While these methods estimate fairness metrics across small subgroups, they still largely quantify observable outcome disparities and do not explicitly distinguish between additive versus intersectional sources of bias.

Multidimensional and intersectional discrimination. In legal and social theory, discrimination is recognised as multidimensional (Uccellari 2008; Makkonen 2002), manifesting as multiple, sequential, or intersectional discrimination. Importantly, intersectionality, first theorised by Crenshaw (2013), emphasises that individuals at the intersection of multiple marginalised identities face unique, compounded discrimination that cannot be reduced to the sum of individual disadvantages (Collins and Bilge 2020). Recent ML research has begun acknowledging the multidimensional discrimination problem (Ovalle et al. 2023; Alvarez and Ruggieri 2025; Roy, Horstmann, and Ntoutsi 2023; Gohar and

Cheng 2023). However, most works either focus on detecting disparities across subgroups, without clearly distinguishing between multiple (additive) and intersectional (non-additive) effects.

Measurement problem of intersectionality. More broadly, scholars have raised concerns that fairness metrics may mismeasure the forms of discrimination they seek to detect (Binns 2018; Barocas and Selbst 2016; Corbett-Davies et al. 2023). These critiques argue that algorithmic fairness auditing often reduces structural or historical discrimination to empirical disparities, obscuring the underlying causes of discrimination. In this light, measuring outcomes between groups without identifying the true sources of the discrimination (i.e due to multiple or intersectional) might fail to capture the severity of discrimination. Our work seeks to bridge this gap by providing the first systematic empirical analysis of whether existing intersectional fairness metrics truly capture the non-additive, structural discrimination theorised in legal and social frameworks, or whether they primarily detect additive disparities across multiple sensitive attributes.

3 Definitions of Multi-Sensitive Discrimination

In this section, we formalise the definitions of different types of multidimensional discrimination with a focus on single-axis, multiple, intersectional and compounded discrimination.

3.1 Problem Setup and Notation

We consider a decision-making setting where each individual is represented by a feature vector $\{(g_1, r_1, x_1, y_1), \cdots (g_n, r_n, x_n, y_n)\}$ where $g_i \in \mathcal{A}_G$ and $r_i \in \mathcal{A}_R$ are both considered binary sensitive attributes and collectively referred to as $\mathcal{A}$. Also, we consider the case where $\mathcal{A}$ can be multidimensional, thus referring to more than one sensitive attribute. However, for simplicity in our paper, we consider that $\mathcal{A}$ consists of two attributes, namely gender and race, thus, $\mathcal{A} \subseteq \{\mathcal{A}_G, \mathcal{A}_R\}$. Additionally, $x_i \in X$ are non-sensitive attributes, and $y_i \in \{0, 1\}$ refers to a decision such as whether to grant a loan or approve bail. We assume that the sensitive attributes G and R can influence the non-sensitive features X, and consequently influence the decision Y.

Importantly, we consider the case where we can analyse discrimination not only with respect to individual sensitive attributes G or R but also at the level of their intersection. Additionally, we clarify that in our setting, Y represents actual decisions, however, we can extend these analyses to the case where the dataset could be used to train an ML model. We formalise the notion of a subgroup following the approach in Kearns et al. (2018):

Definition 1 (*Subgroups*)**.** Let $\mathcal{A} \subseteq \{A_1, A_2, \ldots, A_n\}$ be the set of n sensitive attributes. Each attribute A_i takes values from a finite set, and a_i can refer to a particular value of A_i. Let g_{a_i} denote the set of individuals for whom $A = a$. A subgroup defined by the attribute value combination $(a_1, a_2, \ldots, a_n)$ is the intersection of the corresponding groups:

$$sg_{a_1 \times a_2 \times \cdots \times a_n} = g_{a_1} \cap g_{a_2} \cap \ldots \cap g_{a_n}.$$

For example, suppose we have two sensitive attributes thus $\mathcal{A} \subseteq \{\mathcal{A}_G, \mathcal{A}_R\}$ such that $\mathcal{A}_R \in \{\textit{White}, \textit{NonWhite}\}$ and $\mathcal{A}_G \in \{\textit{Male}, \textit{Female}\}$. The subgroups formed by these attributes are:{*White Male, NonWhite Male, White Female, and NonWhite Female*}. Motivated by this, we introduce formal definitions of multiple and intersectional discrimination, which guide our analysis design.

3.2 Multidimensional Discrimination

We assume that the probability distribution of the decision outcome Y depends either directly or indirectly on the non-sensitive covariates X, as well as on the sensitive attributes G (e.g., gender) and R (e.g., race). These sensitive attributes may serve as sources of potential discrimination, whether through their direct influence on Y, or indirectly through their effect on X. The parametric form of $\mathcal{P}(y = 1)$ used in our analysis is detailed in Equation 8.

Formally, we define the marginal effect of each sensitive attribute $\mathcal{A} \in \{\mathcal{A}_G, \mathcal{A}_R\}$ on the outcome Y as the change in outcome probability across values of $\mathcal{A}$, marginalizing over all other attributes. These are computed as:

$$\Delta_{\mathcal{A}_G} = \mathcal{P}(y = 1 \mid \mathcal{A}_G = 1) - \mathcal{P}(y = 1 \mid \mathcal{A}_G = 0) \quad (1)$$

$$\Delta_{\mathcal{A}_R} = \mathcal{P}(y = 1 \mid \mathcal{A}_R = 1) - \mathcal{P}(y = 1 \mid \mathcal{A}_R = 0) \quad (2)$$

We also define the intersectional (non-additive) effect as:

$$\begin{aligned} \Delta_{\mathcal{A}_{GR}} = \mathcal{P}(y = 1 \mid \mathcal{A}_G = 1, \mathcal{A}_R = 1) \\ - \mathcal{P}(y = 1 \mid \mathcal{A}_G = 1, \mathcal{A}_R = 0) \\ - \mathcal{P}(y = 1 \mid \mathcal{A}_G = 0, \mathcal{A}_R = 1) \\ + \mathcal{P}(y = 1 \mid \mathcal{A}_G = 0, \mathcal{A}_R = 0) \end{aligned} \quad (3)$$

Here, $\Delta_{\mathcal{A}_G}$ measures the effect of changing gender on the outcome, independent of any effect from race and vice versa for $\Delta_{\mathcal{A}_R}$. Then, the interaction term, $\Delta_{\mathcal{A}_{GR}}$, captures any non-additive or interaction effect between gender and race.

Definition 2 (*Single-axis discrimination*)**.** Single-axis discrimination occurs when the outcome varies with changes in a single protected attribute, while the interaction between sensitive attributes has no effect. Formally, this holds if:

$$\Delta_{\mathcal{A}_G} \neq 0 \quad \textit{or} \quad \Delta_{\mathcal{A}_R} \neq 0, \text{ but not both, and } \Delta_{\mathcal{A}_{GR}} = 0. \quad (4)$$

Example 3.1. A loan approval system is single-axis discriminatory if it shows systematically different outcomes based solely on gender or race. This form of discrimination aligns with many legal frameworks, which protect individuals or groups from discrimination based on *single* protected attribute. The work of Roy, Horstmann, and Ntoutsi (2023) describes this form of discrimination as *mono-discrimination*.

Definition 3 (*Multiple discrimination*)**.** Multiple discrimination occurs when multiple sensitive attributes independently contribute to the outcome, such that their combined effect is additive and separable, and no interaction effects are present. Formally:

$$\Delta_{\mathcal{A}_G} \neq 0 \quad \textit{and} \quad \Delta_{\mathcal{A}_R} \neq 0, \quad \textit{and} \quad \Delta_{\mathcal{A}_{GR}} = 0. \quad (5)$$

Example 3.2. Consider a loan application where an individual is denied based on two sensitive attributes, i.e religion and gender. In this case, the discrimination associated with each attribute could independently contribute to the outcome, with their effects simply adding up.

Definition 4 (*Intersectional discrimination*)**.** Intersectional discrimination occurs when the outcome is influenced by the interaction of two or more sensitive attributes, resulting in a *non-additive* form of discrimination. Intersectional discrimination indicates a *non-additive* influence of G and R on the decision, which cannot be merely separated. Formally, we define:

$$\Delta_{\mathcal{A}_{GR}} \neq 0 \quad \textit{and} \quad \Delta_{\mathcal{A}_G} = 0 \quad \textit{and} \quad \Delta_{\mathcal{A}_R} = 0 \tag{6}$$

Example 3.3. Weichselbaumer (2020) found that Turkish women were significantly less likely to be invited for an interview compared to women with German-sounding names. The discrimination was even more likely for Turkish women who wore a headscarf. Thus, suggesting that neither ethnicity nor religion separately explained discrimination, but instead the inseparable interaction of the two.

Definition 5 (*Compounded discrimination*)**.** Compounded discrimination occurs when both multiple and intersectional discrimination are present. That is, the sensitive attributes influence the outcome independently and jointly. Formally:

$$\Delta_{\mathcal{A}_G} \neq 0, \quad \text{and} \quad \Delta_{\mathcal{A}_R} \neq 0, \quad \text{and} \quad \Delta_{\mathcal{A}_{GR}} \neq 0. \tag{7}$$

Example 3.4. Compounded discrimination represents the simultaneous presence of both multiple and intersectional discrimination. For example, an individual may face disadvantages due to both their gender and race independently, while also experiencing additional harm arising from the unique combination of these attributes.

Clarifying the distinction between multiple and intersectional discrimination. Multiple and intersectional discrimination are often conflated, as they typically involve the same sensitive attributes, such as race and gender. However, the key difference lies in the discriminatory contribution of each sensitive attribute. In *multiple discrimination*, each attribute negatively affects the outcome independently. For example, someone may face bias both because of their race and because of their gender, but each source of bias can be identified on its own. This makes it possible to raise legal complaints based on each attribute separately. In contrast, *intersectional discrimination*, the sensitive attributes are inseparable, thus creating a *unique* type of discrimination. Moreover, intersectional discrimination can occur without multiple discrimination. A clear example of this case is the *Parris v. Trinity College Dublin and Others*[1] , where age and sexual orientation interactively produced a unique form of discrimination that neither sensitive attribute on its own would have caused. In addition, a number of legal laws do not recognise intersectionality as a form of discrimination that needs to be considered (Xenidis 2018; Center for Intersectional Justice [CIJ] 2019; Makkonen 2002).

[1]Parris lawsuit https://shorturl.at/FGiK8

Clarifying the distinction between intersectional and compounded discrimination. Intersectional and compounded discrimination may also appear similar, as both involve interaction effects. However, intersectional discrimination assumes neither sensitive attribute independently influences the outcome, and the discrimination arises solely from the interaction. On the other hand, compounded discrimination occurs from the independent effect on the outcome in addition to their interaction. Our interpretation of intersectional discrimination aligns with the framing of Crenshaw (2013); German Federal Anti-Discrimination Agency (2024), which defines it as a specific interaction whose effects cannot be disentangled. We also acknowledge the theoretical possibility of cases where single-axis discrimination occurs with intersectional discrimination. In such cases, we treat them as a special instance of compounded discrimination.

4 Synthetic Dataset Design Process

Having formalised the different types of discrimination in Section 3, we now describe the synthetic dataset generation process used to simulate discrimination scenarios in a controlled manner. This *controlled setting allows for a systematic analysis of intersectional fairness metrics under known sources of discrimination*. However, this might be difficult to achieve with real-world data due to the unknown sources of bias. We define a binary decision outcome Y as a non-linear function of covariates X and sensitive attributes G (e.g., gender) and R (e.g., race). These attributes act as sources of potential discrimination. Formally, we express this as:

$$\mathcal{P}(y = 1 \mid X, G, R) = \sigma\Big(f(X) + m_g(\nu_g G) + m_r(\nu_r R) + m_{gr}(\nu_{gr}(G \times R))\Big) \tag{8}$$

where:

- $\sigma(\cdot)$ is a non-linear activation function (e.g., sigmoid),
- $f(X)$ captures the influence of covariates X,
- m_g, m_r, and m_{gr} represent transformation functions over each sensitive input or their interaction,
- ν_g, ν_r, and ν_{gr} are scalar coefficients that control the overall influence of the individual and interaction effects of G and R on the decision outcome Y.

This formulation allows us to vary the values of ν_g, ν_r, and ν_{gr}, thereby enabling explicit control over whether the decision outcome is influenced by individual effects (from G or R alone), additive effects (from both G and R independently), or interaction effects (from the combination $G \times R$). This controlled setup serves as the foundation for the formal definitions of multidimensional discrimination that follow.

Feature design and structural dependencies. We model a loan approval setting involving two binary sensitive attributes *Gender* and *Race*, including three non-sensitive features, thus *Education*, *Income*, and *Savings*. The outcome variable $y_i \in \{0, 1\}$ indicates whether an individual is granted a loan or not. The structural dependencies between

variables are motivated by plausible socioeconomic relationships. Specifically:

- **Education** (E) may be influenced by sensitive attributes such as gender and race, capturing the impact of systemic inequality in access to educational opportunities.
- **Income** (I) is modelled as a function of both education and sensitive attributes, which reflects that labour market outcomes are affected by qualifications as well as sensitive attributes (Kochhar 2023).
- **Savings** (S) are assumed to depend on income, as higher income generally enables greater saving capacity.
- **Loan Amount** (L) is influenced by both income and savings. Intuitively, individuals with higher income or savings might request smaller loan amounts.
- **Loan Duration** (D) depends on the loan amount (larger loans may require longer repayment periods) and inversely on income (higher income may allow quicker repayment).

Encoding discrimination types. We build on the causal framework proposed by Majumdar et al. (2025), which models a loan approval setting involving binary sensitive attributes. Leveraging this structure, we generate five synthetic datasets, each designed to reflect a distinct type of discrimination as defined in Section 3.2. These include single, multiple, intersectional, and compounded discrimination. Each dataset encodes a specific source of bias, allowing us to embed ground truth discrimination directly into the data. This controlled setup enables a systematic assessment of whether existing intersectional fairness metrics can accurately detect non-additive, intersectional discrimination. These datasets serve as the foundation for our empirical evaluation in Section 5. The five distinct dataset discrimination scenarios include:

- *No bias dataset.* The outcome Y is fully independent of the sensitive attributes G and R, with discrimination parameters set to zero $\nu_g = \nu_r = \nu_{gr} = 0$. This implies that no marginal effect of the sensitive attribute on the outcome:

$$\Delta_{\mathcal{A}_G} = \Delta_{\mathcal{A}_R} = \Delta_{\mathcal{A}_{GR}} = 0. \tag{9}$$

  This dataset serves as a baseline to evaluate whether intersectional fairness metrics remain calibrated in the absence of any bias.
- *Single-axis bias dataset.* In this setting, the outcome Y is influenced by only one sensitive attribute (either G or R), but not both. We simulate this by setting the value of $\nu_g > 0$, and $\nu_r = \nu_{gr} = 0$. This way of generating the data allows us to mimic discrimination resulting from a single-axis discrimination following Definition 2.
- *Multiple bias dataset.* This reflects Definition 3, where both G and R independently influence Y through additive marginal effects, but without any interaction. We simulate this setting by setting the values of $\nu_g > 0, \nu_r > 0$ and $\nu_{gr} = 0$.

Dataset	**Parameters** (ν_G, ν_R, ν_{GR})
No Bias	{0.00, 0.00, 0.00}
Single-Axis	{2.00, 0.00, 0.00}
Additive	{0.85, 0.90, 0.00}
Intersectional	{0.00, 0.00, 2.00}
Compounded	{1.00, 0.90, 1.18}

Table 1: **Discrimination parameters ν for each dataset**. Values represent coefficients for (ν_G, ν_R, ν_{GR}) which controls the effect of discrimination in the outcome Y.

- *Intersectional bias dataset.* This dataset scenario captures the form of non-additive discrimination defined in Definition 4. The outcome is influenced only by the interaction between $G \times R$, while there are no additive effects of G and R. Here, we set the value of $\nu_{gr} > 0$ and $\nu_g = \nu_r = 0$.
- *Compounded Bias Dataset.* This scenario reflects both additive and intersectional discrimination, as described in Definition 5. The interaction term $G \times R$ has a significant marginal effect on Y in addition to the additive effects of G and R. We simulate this setting by setting the values of $\nu_g > 0, \nu_r > 0$ and $\nu_{gr} > 0$.

Each dataset contains $10,000$ samples drawn once using a fixed random seed to ensure reproducibility. Table 1 provides the exact values of ν_g, ν_r, ν_{gr} for each discrimination scenario. Additional generation parameters are provided in Table 3 and the full functional forms are detailed in Equation 10 and code available at https://tinyurl.com/4s53zmw5.

Root nodes:

$$f_G : G = U_G, \quad U_G \sim \text{Bernoulli}(0.5),$$

$$f_R : R = U_R, \quad U_R \sim \text{Bernoulli}(0.5),$$

Covariates:

$$E = -0.5 + \lambda_E\big(\theta_G G + \theta_R R + \theta_{GR}(G \times R)\big) + U_E, \quad U_E \sim \mathcal{N}(0, 0.25),$$

$$I = -4 + 3E + \lambda_I\big(\beta_G G + \beta_R R + \beta_{GR}(G \times R)\big) + U_I, \quad U_I \sim \mathcal{N}(0, 4),$$

$$S = -4 + 1.5\,\mathbf{1}\{I > 0\}\,I + U_S, \quad U_S \sim \mathcal{N}(0, 5),$$

$$L = 1 - \beta_I I - \beta_S S \quad + \lambda_L\big(\rho_G G + \rho_R R + \rho_{GR}(G \times R)\big) + U_L, U_L \sim \mathcal{N}(0, 10), \quad \beta \in \{0, 0.03\},$$

$$D = -1 - \beta_I I + L + \lambda_D\big(\kappa_G G + \kappa_R R + \kappa_{GR}(G \times R)\big) + U_D, U_D \sim \mathcal{N}(0, 9),$$

Outcome:

$$Y = \mathbf{1}\Big\{15 + \alpha\big[\delta(-L - D) + 0.3(I + S + \alpha IS)\big] \geq 0.5 + \gamma\big(1 - \lambda_Y(\nu_G G + \nu_R R + \nu_{GR}(G \times R))\big)\Big\}, \tag{10}$$

Legal and social relevance. Our study uses synthetic datasets, hence allowing us to have full control and empirically investigate intersectional fairness metrics. The design of these datasets is grounded in understanding real-world challenges related to multidimensional discrimination. These dataset generation scenarios are not only methodologically distinct but also map directly onto known limitations and concerns in legal and policy contexts. Current legal frameworks, such as those in the EU and the US, often do not explicitly accommodate intersectional claims unless they can be decomposed into additive effects (Crenshaw 2013; Xenidis 2018; Center for Intersectional Justice [CIJ] 2019; Kupupika 2021). As a result, people who face discrimination only at the intersection of identities (e.g., Black women, Muslim women wearing headscarves) can often be exempted from legal recourse. This highlights the need to have metrics that can effectively decompose overall discrimination into multiple and intersectional forms. Achieving this would inform effective policy interventions but also provide clear tools to address sources of structural discrimination.

4.1 Metrics of Intersectional Fairness

In response to increasing attention to intersectional discrimination, a number of metrics have been proposed to quantitatively measure intersectional discrimination. We include a summary of these metrics in Table 2. While these metrics can capture disparities across subgroups, we *argue that they fall short of isolating the non-additive effects that define intersectionality in legal and social theory*. Specifically, most metrics quantify group-level disparities without addressing the structural, interaction-based discrimination which is at the core of intersectionality. In this section, we summarise the key metrics considered in our analyses following a taxonomy proposed by Gohar and Cheng (2023).

Minimax approaches. The minimax is based on Rawls (2001) principles of distributive justice with the overall goal of the welfare of the worst performing group (Rigobon 2023; Martinez, Bertran, and Sapiro 2020; Diana et al. 2021; Martinez et al. 2021). Ghosh, Genuit, and Reagan (2021) extends the concept of minimax to accommodate subgroups by considering a possible permutation of subgroups and estimating the ratio of the maximum and minimum from the list of subgroups. An estimated ratio greater than 0 indicates greater disparities between subgroups. The fairness metrics considered under this approach include demographic (dis)parity (Dwork et al. 2012), equality of opportunity (Hardt, Price, and Srebro 2016) and group benefit (Gartner 2020). We provide additional definitions for each intersectional metric in the appendix on github[2], and refer interested readers to comprehensive overviews of fairness notions in Mehrabi et al. (2021); Pessach and Shmueli (2022).

Differential fairness approaches. Differential fairness approaches are inspired by legal standards such as the 80% rule established in the U.S. Commission (1970) Code of Federal Regulations. Under this anti-discrimination guideline, a decision is considered potentially biased if the ratio of favourable outcomes between a disadvantaged group and an advantaged group is less than 80%. Foulds et al. (2020) extend this notion to an intersectional discrimination setting, aiming to protect individuals who belong to the intersection of multiple sensitive attributes. Rather than relying on a fixed threshold (e.g., 80%, Foulds et al. (2020) adopt a sliding-scale approach similar to differential privacy, parameterised by a tunable threshold ϵ. Building on this, Morina et al. (2019) generalises the differential fairness framework to encompass standard fairness notions such as demographic parity, true positive rate parity, etc. In practice, differential fairness measures the ratio of outcome probabilities between any pair of subgroups, with lower values of ϵ indicating greater fairness and $\epsilon > 1$ typically reflecting higher disparity. While these approaches are theoretically grounded and applicable across different prediction tasks, they primarily capture disparities between subgroup outcomes. As such, they may overlook the specific interaction effects that define intersectional discrimination in legal and social theory.

Subgroup-based approaches. Subgroup fairness was proposed by Kearns et al. (2018) as a constraint for learning a rich collection of exponentially many subgroups. However, it can also serve as a measuring metric to detect calibration disparities across subgroups. While useful for ensuring predictive consistency, they may not reflect structural disadvantage or detect intersectional effects. In particular, this approach considers individual protected attributes into different subgroups and ensures fairness within these subgroups.

5 Empirical Analysis of Intersectional Fairness Metrics

In this section, we empirically analyze how well intersectional fairness metrics align with legal and social theories of intersectionality. For simplicity, we used two binary sensitive attributes, thus $Gender(G) \in \{Male : 0, Female : 1\}$ and $Race(R) \in \{White : 0, Black : 1\}$, whereas the decision outcome Y can take values $\{Good : 1, Bad : 0\}$. We define a protected group as a subgroup at a disadvantage. For example, in a hiring scenario, we identify individuals who belong to the subgroup of $\{Female, Black\}$ as the protected group. Building on our controlled datasets introduced in Section 4, we structure our analysis around a set of research questions (RQ). Additionally, in presenting all the figures, we apply minimax scaling to standardize the intersectional metrics values for comparability across datasets. Ideally, standardising the metrics within the datasets would have been preferred. However, this is not feasible here, as each metric returns a single value per dataset. Instead, we scale all values to the 0 and 1 range across datasets. Further details on the min-max scaling procedure are provided on the GitHub repository of the project.

5.1 RQ1: Do Intersectional Fairness Metrics Detect Bias When It Is Not Present?

We begin our empirical analyses with the no-bias dataset, which serves as our baseline where the outcome is inde-

[2]Github code: https://tinyurl.com/4s53zmw5

Approaches	Metric	Definition	Applies
MiniMax Approaches	Demographic disparity	$1 - \frac{\min \mathcal{P}(y=1 \mid A \in sg_{a_i})}{\max \mathcal{P}(y=1 \mid A \in sg_{a_j})}$	Data
	Equality of opportunity	$1 - \frac{\min \mathcal{P}(\hat{y}=1 \mid y=1, A \in sg_{a_i})}{\max \mathcal{P}(\hat{y}=1 \mid y=1, A \in sg_{a_j})}$	Classifier
	Group benefit	$1 - \frac{\min \mathcal{P}(\hat{y}=1 \mid A \in sg_{a_i})}{\max \mathcal{P}(y=1 \mid A \in sg_{a_j})}$	Classifier
Differential fairness Approaches	Smoothed ϵ- differential unfairness	$e^{-\epsilon} \leq \frac{\mathcal{P}(\hat{y}=1 \mid A \in sg_{a_i})}{\mathcal{P}(\hat{y}=1 \mid A \in sg_{a_j})} \leq e^{\epsilon}$	Classifier
	True positive rate disparity (TPR)	$e^{-\epsilon} \leq \frac{\mathcal{P}(\hat{y}=1 \mid y=1, A \in sg_{a_i})}{\mathcal{P}(\hat{y}=1 \mid y=1, A \in sg_{a_j})} \leq e^{\epsilon}$	Classifier
	False positive rate disparity (TPR)	$e^{-\epsilon} \leq \frac{\mathcal{P}(\hat{y}=1 \mid y=0, A \in sg_{a_i})}{\mathcal{P}(\hat{y}=1 \mid y=0, A \in sg_{a_j})} \leq e^{\epsilon}$	Classifier
	elift	$e^{-\epsilon} \leq \frac{\mathcal{P}(y=1 \mid A \in sg_{a_i})}{\mathcal{P}(y=1)} \leq e^{\epsilon}$	Data
	slift	$e^{-\epsilon} \leq \frac{\mathcal{P}(y=1 \mid A \in sg_{a_i})}{\mathcal{P}(y=1 \mid A \in sg_{a_j})} \leq e^{\epsilon}$	Data
Subgroup-based Approaches	Subgroup unfairness	$\lvert\mathcal{P}(y=1) - \mathcal{P}(y=1 \mid A \in sg_{a_i})\rvert \cdot \mathcal{P}(A \in sg_{a_i}) \leq \gamma$	Data

Table 2: **Summary of intersectional fairness metrics**. Metrics are organized by approach, with definitions and an indication of whether they are applicable to the decision outcome Y, or the classifier predictions $\hat{Y}$. Some metrics require access to both the classifier predictions and ground truth labels. Smoothed ϵ-differential unfairness and slift are conceptually similar, but we apply the former to the classifier predictions and the latter to the decision outcome Y.

pendent of the sensitive attributes. This allows us to analyze whether metrics designed for intersectionality falsely detect discrimination when discrimination is absent. A metric detecting discrimination when not present could lead to false conclusions and undermine its trust in auditing or policy interventions. To explore this, we use four of the intersectional metrics i.e demographic disparity, elift, slift and subgroup metrics), which have been designed to be used in detecting discrimination on the decision outcome. Ideally, metrics should return low or zero values when discrimination is not present. As shown in Figure 1, all metrics correctly report values near zero after min-max normalization. This indicates that these metrics are well calibrated to avoid bias detection when no disparities exist.

★ All selected four intersectional metrics return near-zero values on the no-bias dataset. This confirms these metrics do not falsely measure discrimination when it is absent.

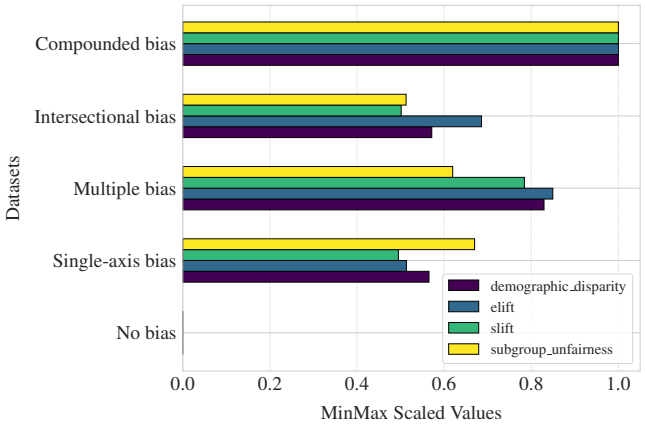

Figure 1: **Evaluation of intersectional fairness metrics.** Normalized values of intersectional fairness metrics (using min-max scaling) across different discrimination scenarios in the synthetic datasets.

5.2 RQ2: Do Metrics Respond Differently to Multiple vs. Intersectional Bias?

A central goal of this empirical analysis is to examine whether intersectional fairness metrics are sensitive to the distinction between multiple and intersectional discrimination. As explained in Section 4, multiple discrimination arises from the additive effects of individual attributes, while intersectional discrimination emerges from their interaction, producing a unique and inseparable form of discrimination. Figure 1, reveals a clear trend where metrics report higher values for multiple bias datasets than for the intersectional bias dataset, despite both encoding meaningful discrimination. In particular, we observe a chronological increase in discrimination as we move from $Single > intersectional > Multiple > Compounded$ dataset. This difference suggests that most of these metrics are sensitive

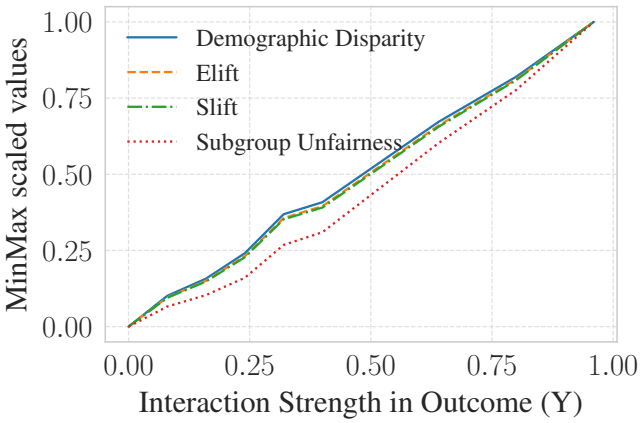

Figure 2: **Sensitivity of interactions in the intersectionality Data**. All metrics show an increase in discrimination as interaction increases.

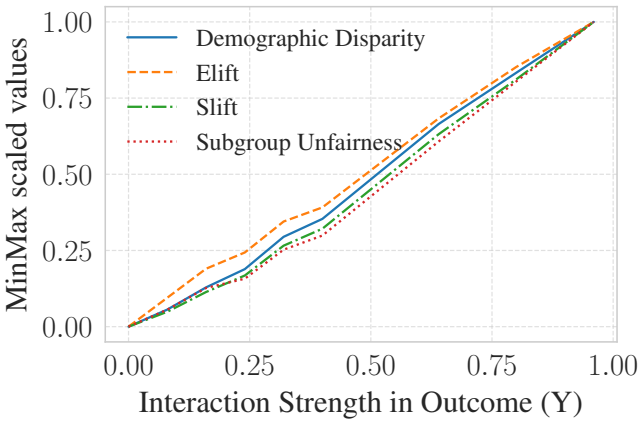

Figure 3: **Sensitivity of multiple discrimination dataset**. By varying the strength of additive discrimination, metrics show a steady increase as the discrimination strength increases.

to additive disparities than to non-additive, intersectional effects. This discrepancy reveals a key limitation that current metrics largely treat intersectional fairness as an empirical disparity problem rather than a structural one. That is, they measure observable outcome gaps between groups, but not whether those gaps arise from independent attribute effects or from their interaction. This aligns with critiques from intersectionality scholars who argue that merely disaggregating by group overlooks the compounded and interlocking nature of discrimination at the intersection of identities (Crenshaw 2013; Xenidis 2018; Center for Intersectional Justice [CIJ] 2019). In particular, this pattern persists even in metrics inspired by legal notions of fairness, such as slift and elift (Morina et al. 2019; Foulds et al. 2020). Although designed to protect intersectional subgroups, these metrics depend on comparing ratios of positive outcomes and are not explicitly designed to separate additive from non-additive effects. As a result, intersectional discrimination may be underestimated or even missed entirely in fairness audits.

★ Intersectional fairness metrics are more sensitive to additive than to non-additive discrimination, suggesting a misalignment with intersectionality's emphasis on non-additive, structurally embedded discrimination.

5.3 RQ3: Do Metrics Respond Differently to Intersectional vs. Compounded Bias?

This question helps us reflect on the real-world cases where individuals may experience both independent and interlocking disadvantages. For example, Black women may face discrimination in hiring not only for being Black or for being women, but also from the compounded effect of both identities. As shown in Figure 1, all metrics report their highest values under the compounded bias scenario. This aligns with the dataset design, where both ν_G, ν_R, and ν_{GR} are non-zero, producing both additive and interactional disparities. However, this also raises a challenge, where it is unclear whether these metrics respond primarily to the additive terms, the interaction term, or both. In fairness auditing, particularly those aimed at identifying structurally embedded discrimination metrics that cannot disentangle additive from non-additive contributions, may overestimate fairness when only intersectional bias is present, or underestimate structural risk when additive components dominate. This highlights the need for metrics that not only quantify disparity but also identify its source.

★ While metrics respond strongly to compounded bias, they conflate additive and intersectional effects. This limits their ability to isolate intersectional discrimination and risk misguided policy interventions.

5.4 RQ4: Do Intersectional Fairness Metrics Increase With Stronger Interaction Effects?

To assess whether existing metrics are responsive to intersectional discrimination, we conduct two sensitivity analyses. First, in the multiple bias dataset, we vary the strength of the discrimination from the two independent sensitive attributes while keeping their interaction as $\nu_{GR} = 0$. Second, in a dataset with only intersectional discrimination, we vary the strength of the interaction term ν_{GR} while holding the additive terms at zero. Figure 2 and Figure 3 show the results for sensitivity for intersectional and multiple bias settings, respectively. Here, we observe a similar trend for the different metrics in both cases. For example, all metrics show a steady increase as the strength of discrimination increases. This indicates that current metrics are responsive to general outcome disparities regardless of whether those arise from additive or interactional sources. However, this observed increase does not imply that the metrics distinguish between additive and non-additive discrimination. Rather, they conflate the two, as the same metric score can result from fun-

Parameter	Value	Description
Sample size	10,000	Number of generated samples
Random seed	42	For reproducibility
Gender prob.	0.5	Probability of gender
Race prob.	0.5	Probability of race
β_S	0.6	Savings effect on loan (L)
β_I	0.8	Income effect on L and D
γ	0.10	Bias scaling in outcome (Y)
δ	0.45	Treatment weight in Y
η	2.0	Noise std. in Y

Table 3: Parameter settings for synthetic dataset generation.

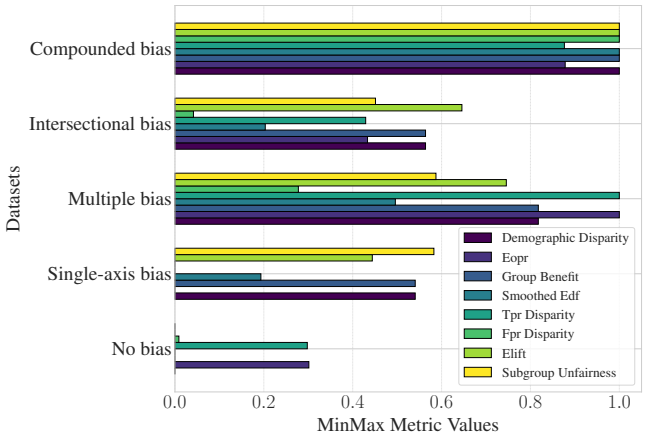

Figure 4: **Evaluation of intersectional metrics on trained classifier**. Intersectional metrics produce higher discrimination values when additive discrimination is present.

damentally different underlying forms of bias.

★ Intersectional fairness metrics increase with both additive and interaction strength, indicating that they capture disparities without distinguishing their structural origin.

5.5 RQ5: Do Trained Classifiers Amplify or Preserve Bias Captured by Intersectional Fairness Metrics?

While the previous RQs focused on measuring discrimination in the observed data, typically, we are interested in learning models that can help us automate the decision-making processes. Therefore, it is critical to assess whether trained classifiers preserve, amplify, or reduce the biases captured by intersectional fairness metrics. To investigate this, we trained a logistic regression classifier using 75% of each dataset for training and then tested on the remaining 25%. As shown in Table 4, the model achieves strong predictive performance across all datasets, confirming that any disparities observed in predictions are not due to poor model fit. We used eight different intersectional fairness metrics, which allow us to estimate discrimination in the classifier predictions. Figure 4 presents these results across the different dataset scenarios discussed in Section 4. Notably, some metrics, such as demographic disparity and FPR Disparity, report non-zero discrimination even in the no-bias dataset, suggesting potential false positives. Moreover, in the multiple and compounded bias scenarios, most discrimination scores yield higher scores than in the original data. These results indicate that the classifier amplifies the underlying discrimination. These findings reveal that even when trained on data with modest or no apparent bias, classifiers can learn and propagate patterns of disparity, particularly when the bias arises from additive effects.

★ Classifiers can amplify bias even when trained on minimally biased data. Fairness metrics applied to predictions often show stronger disparities, underscoring the need for post-training fairness evaluation.

6 Discussion

Our findings provide new insights into the measurement of intersectional fairness. In this section, we reflect on the implications of our analysis for fairness auditing, highlight current limitations, and suggest directions for future research.

Implications for fairness auditing. Our analysis empirically confirms what intersectionality theorists have long argued, that additive group-level disparity metrics, even when applied across intersectional subgroups, fail to capture the non-additive structural discrimination that intersectionality describes (Crenshaw 2013; Collins and Bilge 2020; Xenidis 2018). The inability of current metrics to distinguish between multiple discrimination (additive) and intersectional discrimination (intersectional) raises an important concern for fairness auditing. Fairness audits that rely on these metrics may systematically underestimate harm for the most structurally marginalised groups, especially in settings where discrimination arises only at the intersection of attributes (Andrews, Shah, and Cheng 2023; Alvarez and Ruggieri 2025). This calls for caution when fairness practitioners report fairness metrics for intersectional groups. Reporting subgroup disparities without considering the source of the disparities risks misdiagnosing discrimination and misguided policy mitigation efforts.

What should intersectional fairness measure. Our work provides a formal taxonomy of multidimensional discrimination, defining single-axis, multiple, intersectional, and compounded discrimination in a way that enables precise measurement. In doing so, we raise a fundamental conceptual question: what should intersectional fairness metrics actually aim to capture? A common point of confusion is whether multiple and compounded discrimination are merely special cases of intersectional discrimination. Legally and theoretically, however, clear distinctions have been established (Makkonen 2002; Crenshaw 2013; Xenidis 2018). Our formalization in Section 3.2 aims to reinforce these distinctions: multiple discrimination reflects separa-

ble, additive effects of individual attributes, while intersectional discrimination captures non-additive discrimination that emerges only at the intersection of sensitive attributes. This distinction then leads to a deeper normative challenge: should intersectional fairness metrics aim to detect subgroup disparities more broadly, or should they specifically isolate the structural discrimination that arises from interactions between sensitive attributes? Most existing metrics equate intersectional fairness with subgroup parity (Morina et al. 2019; Kong 2022; Gohar and Cheng 2023). In doing so, they implicitly treat intersectionality as a general measure of disparity. However, this framing diverges from legal and social understandings, which emphasize that overlapping systems of oppression produce qualitatively distinct experiences of discrimination (Crenshaw 2013). For example, the German Federal Anti-Discrimination Agency defines intersectional discrimination as the "specific interaction or overlapping of different characteristics of discrimination. . . [that] cannot be separated anymore" (German Federal Anti-Discrimination Agency 2024). Such definitions highlight a deeper structural account of harm that cannot be captured by simple disaggregation or additive modeling. We argue that future work must go beyond measuring disparities and instead focus on decomposing them. The ability to distinguish between additive and non-additive discrimination is not merely a technical refinement it carries significant legal and policy implications.

Toward causal decomposition approaches. One key observation from our study is that almost all current intersectional metrics are disparity-based and lack a causal grounding. This aligns with recent calls to bridge the gap between observational fairness metrics and structural causal reasoning (Barocas and Selbst 2016; Corbett-Davies et al. 2023; Kusner et al. 2017). Our results suggest that causal decomposition methods that are designed to separate direct, indirect, and interaction effects between attributes and outcomes could offer a more faithful measurement of intersectional discrimination (Kusner et al. 2017; Chiappa 2019). By explicitly modelling how sensitive attributes interact to influence decisions, causal frameworks may help address the very misalignment we identify between legal intersectionality and statistical fairness auditing. We see this as a promising direction for future work.

Limitations and scope for further research. Our experiments are conducted using synthetic data with fully controlled discrimination structures. While this allows for clean identification of multiple vs. interactional effects, real-world datasets are more complex and noisy. As such, further work is needed to test whether these patterns generalise to high-dimensional real-world data. Additionally, we focus on binary sensitive attributes (e.g., gender and race). Future research should investigate whether our findings hold under multi-valued or continuous sensitive attributes.

7 Conclusion

This work highlights a critical gap between the legal-social understanding of intersectionality and its operationalisation in machine learning fairness metrics. Our controlled experiments demonstrate that existing metrics, while effective at detecting subgroup disparities, remain largely insensitive to non-additive interaction effects, which importantly defines a characteristic of intersectional discrimination. This finding highlights the risk of mismeasurement in fairness auditing and raises an urgent need for more structurally aware evaluation frameworks. While these metrics remain valuable, they must be complemented with approaches capable of decomposing and isolating intersectional discrimination. Causal decomposition methods present promising directions for addressing this challenge. We believe that this work prompts the fairness community to rethink the design of intersectional fairness metrics, moving beyond disparity detection toward methods that can capture the full structural complexity of discrimination. Accurately diagnosing the nature of discrimination is essential for any mitigation strategy to be effective.

Dataset	Accuracy	F1-Score	AUC
No Bias	**0.946**	0.941	**0.987**
Single Bias	0.936	0.942	0.983
Multiple Bias	0.936	0.949	0.985
Intersectional Bias	0.943	0.943	0.986
Compound Bias	0.936	**0.950**	0.986

Table 4: **Evaluation of classifier**. Performance of trained classifier on test data (75% train, 25% test split).

★ Our work shows that current statistical fairness auditing tools are not sufficient for detecting the types of structural harms intersectionality theory describes. We propose causal decomposition and clearer normative definitions as urgent research directions to bridge this misalignment.

Acknowledgements

We thank Augustine Denteh for invaluable guidance and great insight on the setting of multidimensional discrimination, as well as the PL Group for their valuable support. This work has been funded by the European Union ERC-2021-STG, SAML, project 101040177 https://cordis.europa.eu/project/id/101040177. However, the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible.

References

Alvarez, J. M.; and Ruggieri, S. 2025. Counterfactual Situation Testing: From Single to Multidimensional Discrimination. *arXiv preprint arXiv:2502.01267*.

Andrews, K. S.; Shah, B.; and Cheng, L. 2023. Intersectionality and Testimonial Injustice in Medical Records. *arXiv preprint arXiv:2306.13675*.

Barocas, S.; and Selbst, A. D. 2016. Big data's Disparate Impact. *California Law Review*, 104: 671.

Binns, R. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Conference on Fairness, Accountability and Transparency*, 149–159. PMLR.

Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional accuracy disparities in Commercial Gender classification. In *Conference on Fairness, Accountability and Transparency*, 77–91. PMLR.

Center for Intersectional Justice [CIJ]. 2019. Intersectional discrimination in Europe: Relevance, Challenges and Ways Forward. Commissioned by The European Network Against Racism (ENAR).

Chen, Z.; Zhang, J. M.; Sarro, F.; and Harman, M. 2024. Fairness Improvement with Multiple Protected Attributes: How Far Are We? In *IEEE/ACM International Conference on Software Engineering*, 1–13.

Chiappa, S. 2019. Path-Specific Counterfactual Fairness. In *AAAI Conference on Artificial Intelligence*, volume 33, 7801–7808.

Cho, S.; Crenshaw, K. W.; and McCall, L. 2013. Toward A Field of Intersectionality Studies: Theory, Applications, and Praxis. *Signs: Journal of Women in Culture and Society*, 38(4): 785–810.

Choi, J.; Karumbaiah, S.; and Matayoshi, J. 2025. Bias or Insufficient Sample Size? Improving Reliable Estimation of Algorithmic Bias for Minority Groups. In *Proceedings of International Learning Analytics and Knowledge Conference*, 547–557.

Chouldechova, A. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2): 153–163.

Collins, P. H. 2015. Intersectionality's Definitional Dilemmas. *Annual Review of Sociology*, 41(1): 1–20.

Collins, P. H.; and Bilge, S. 2020. *Intersectionality*. John Wiley & Sons.

Commission, E. E. O. 1970. Guidelines on Employee Selection Procedures. *Federal Register*, 35(149): 12333–12336.

Corbett-Davies, S.; Gaebler, J. D.; Nilforoshan, H.; Shroff, R.; and Goel, S. 2023. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312): 1–117.

Council of Europe. 2025. Intersectionality and Multiple Discrimination. https://www.coe.int/en/web/gender-matters/intersectionality-and-multiple-discrimination. Accessed: February 21, 2025.

Crenshaw, K. 2013. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. In *Feminist Legal Theories*, 23–51. Routledge.

Diana, E.; Gill, W.; Kearns, M.; Kenthapadi, K.; and Roth, A. 2021. Minimax Group Fairness: Algorithms and Experiments. In *AAAI/ACM Conference on AI, Ethics, and Society*, 66–76.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness Through Awareness. In *Innovations in Theoretical Computer Science Conference*, 214–226.

Foulds, J. R.; Islam, R.; Keya, K. N.; and Pan, S. 2020. An Intersectional Definition of Fairness. In *IEEE International Conference on Data Engineering (ICDE)*, 1918–1921. IEEE.

Gartner, J. 2020. A New Metric for Quantifying Machine Learning Fairness in Healthcare. Accessed: May 5, 2025.

German Federal Anti-Discrimination Agency. 2024. Forms of Discrimination. Accessed: May 5, 2025.

Ghosh, A.; Genuit, L.; and Reagan, M. 2021. Characterizing Intersectional Group Fairness with Worst-Case Comparisons. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, 22–34. PMLR.

Gohar, U.; and Cheng, L. 2023. A Survey on Intersectional Fairness in Machine Learning: Notions, Mitigation, and Challenges. *arXiv preprint arXiv:2305.06969*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, 29.

Himmelreich, J.; Hsu, A.; Lum, K.; and Veomett, E. 2024. The Intersectionality Problem for Algorithmic Fairness. *arXiv preprint arXiv:2411.02569*.

Kanubala, D. D.; Valera, I.; and Gupta, K. 2024. Fairness Beyond Binary Decisions: A Case Study on German Credit. *European Workshop on Algorithmic Fairness*.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *International Conference on Machine Learning*, 2564–2572. PMLR.

Kochhar, R. 2023. The Enduring Grip of the Gender Pay Gap. https://www.pewresearch.org/social-trends/2023/03/01/the-enduring-grip-of-the-gender-pay-gap/. Pew Research Center.

Kong, Y. 2022. Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In *ACM Conference on Fairness, Accountability, and Transparency*, 485–494.

Kupupika, T. 2021. Shaping our freedom dreams: Reclaiming intersectionality through Black feminist legal theory. *Virginia Law Review Online*, 107: 27.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. *Advances in Neural Information Processing Systems*, 30.

Lin, Y.; Gupta, S.; and Jagadish, H. 2024. Mitigating Subgroup Unfairness in Machine Learning Classifiers: A Data-Driven Approach. In *IEEE International Conference on Data Engineering (ICDE)*, 2151–2163. IEEE.

Maheshwari, G.; Bellet, A.; Denis, P.; and Keller, M. 2023. Fair Without Leveling Down: A New Intersectional Fairness Definition. *arXiv preprint arXiv:2305.12495*.

Majumdar, A.; Kanubala, D. D.; Gupta, K.; and Valera, I. 2025. A Causal Framework to Measure and Mitigate Non-binary Treatment Discrimination. *arXiv preprint arXiv:2503.22454*.

Makkonen, T. 2002. Multiple, Compound and Intersectional Discrimination: Bringing the Experiences of the Most

Marginalized to the Fore. Åbo Akademi University, Institute for Human Rights.

Martinez, N.; Bertran, M.; and Sapiro, G. 2020. Minimax Pareto Fairness: A Multi Objective Perspective. In *International Conference on Machine Learning*, 6755–6764. PMLR.

Martinez, N. L.; Bertran, M. A.; Papadaki, A.; Rodrigues, M.; and Sapiro, G. 2021. Blind Pareto Fairness and Subgroup Robustness. In *International Conference on Machine Learning*, 7492–7501. PMLR.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6): 1–35.

Mercat-Bruns, M. 2017. Multiple Discrimination and Intersectionality: Issues of Equality and Liberty. *International Social Science Journal*, 67(223-224): 43–54.

Morina, G.; Oliinyk, V.; Waton, J.; Marusic, I.; and Georgatzis, K. 2019. Auditing and Achieving Intersectional Fairness in Classification Problems. *arXiv preprint arXiv:1911.01468*.

Němeček, J.; Kozdoba, M.; Kryvoviaz, I.; Pevnỳ, T.; and Mareček, J. 2025. Bias Detection via Maximum Subgroup Discrepancy. *arXiv preprint arXiv:2502.02221*.

Ovalle, A.; Subramonian, A.; Gautam, V.; Gee, G.; and Chang, K.-W. 2023. Factoring The Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness. In *AAAI/ACM Conference on AI, Ethics, and Society*, 496–511.

Pessach, D.; and Shmueli, E. 2022. A Review on Fairness in Machine Learning. *ACM Computing Surveys*, 55(3): 1–44.

Rawls, J. 2001. *Justice as Fairness: A Restatement*. Harvard University Press.

Rigobon, D. E. 2023. From Utilitarian to Rawlsian Designs for Algorithmic Fairness. *arXiv preprint arXiv:2302.03567*.

Roy, A.; Horstmann, J.; and Ntoutsi, E. 2023. Multidimensional Discrimination in Law and Machine Learning-A Comparative Overview. In *ACM Conference on Fairness, Accountability, and Transparency*, 89–100.

Suk, Y.; and Han, K. C. T. 2023. Evaluating Intersectional Fairness in Algorithmic Decision Making using Intersectional Differential Algorithmic Functioning. *Journal of Educational and Behavioral Statistics*, 10769986241269820.

Uccellari, P. 2008. Multiple Discrimination: How Law Can Reflect Reality. *The Equal Rights Review*, 1: 24–49.

Weichselbaumer, D. 2020. Multiple Discrimination Against Female Immigrants Wearing Headscarves. *ILR Review*, 73(3): 600–627.

Xenidis, R. 2018. Multiple Discrimination in EU Anti-Discrimination Law: Towards Redressing Complex Inequality? In Belavusau, U.; and Henrard, K., eds., *EU Anti-Discrimination Law Beyond Gender*, 41–74. Oxford; Portland: Hart Publishing.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment. In *International Conference on World Wide Web*, 1171–1180.