Multi-dimensional Discrimination in Law and Machine Learning - A Comparative Overview

Arjun Roy*
arjun.roy@unibw.de
Institute of Computer Science, Free University of Berlin; Research Institute CODE, Bundeswehr University Munich
Germany

Jan Horstmann*
jan.horstmann@iri.uni-hannover.de
Institute for Legal Informatics, Leibniz University of Hanover
Germany

Eirini Ntoutsi
eirini.ntoutsi@unibw.de
Research Institute CODE, Bundeswehr University Munich
Germany

ABSTRACT

AI-driven decision-making can lead to discrimination against certain individuals or social groups based on protected characteristics/attributes such as race, gender, or age. The domain of fairness-aware machine learning focuses on methods and algorithms for understanding, mitigating, and accounting for bias in AI/ML models. Still, thus far, the vast majority of the proposed methods assess fairness based on a single protected attribute, e.g. only gender or race. In reality, though, human identities are multi-dimensional, and discrimination can occur based on more than one protected characteristic, leading to the so-called "multi-dimensional discrimination" or "multi-dimensional fairness" problem. While well-elaborated in legal literature, the multi-dimensionality of discrimination is less explored in the machine learning community. Recent approaches in this direction mainly follow the so-called intersectional fairness definition from the legal domain, whereas other notions like additive and sequential discrimination are less studied or not considered thus far. In this work, we overview the different definitions of multi-dimensional discrimination/fairness in the legal domain as well as how they have been transferred/ operationalized (if) in the fairness-aware machine learning domain. By juxtaposing these two domains, we draw the connections, identify the limitations, and point out open research directions.

KEYWORDS

multi-discrimination, multi-fairness, intersectional fairness, sequential fairness, additive fairness

ACM Reference Format:
Arjun Roy, Jan Horstmann, and Eirini Ntoutsi. 2023. Multi-dimensional Discrimination in Law and Machine Learning - A Comparative Overview. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3593013.3593979

*Both authors contributed equally to this research.

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3593979

1 INTRODUCTION

AI-driven decision-making has already penetrated into almost all spheres of human life, from content recommendation [57] and healthcare [66] to predictive policing [2] and autonomous driving [84], deeply affecting everyone, anywhere, anytime. In a variety of cases, discriminatory impacts of AI-driven decision-making on individuals or social groups on the basis of the *protected attributes* like gender, race, age, and others have been observed. Examples range from recidivism prediction [19], hiring [71], recommendations [69] to healthcare [9], education [41], service provision [46] and surveillance [80]. Discriminatory impacts concern both symbolic or representative equality (e.g. ads related to arrest records appearing more frequently along search results for names associated with blacks than whites [78]) and distributive equality (regarding the access to social goods, e.g. in biased hiring algorithms) [90]. The domain of fairness-aware machine learning [67], is concerned with bias and discrimination in AI systems and covers a wide range of topics from understanding bias and discrimination to methods for bias mitigation and accountability [65]. Fair ML research has also been taken up in legal scholarship, with debate as to what role statistical fairness metrics can play under anti-discrimination law [28, 30, 34, 85, 86]. However, despite this steadily growing body of research, the vast majority of proposed methods assumes that discrimination is based on a single protected attribute[1], for example, only gender or only race. We refer to this as *mono-dimensional discrimination/fairness* or *mono-discrimination/fairness.*[2] For the mono-discrimination case, several fairness definitions have been proposed, see [83] for a survey, as well as methods for mitigating mono-discrimination, e.g. [44, 47, 53].

In reality though, humans have *multi-dimensional* identities [79]. We all have a gender, racial or ethnic origin, age and sexual orientation (and more), and are *categorised* by others according to such concepts [5]. Consequently, discrimination cannot always be attributed to a single protected attribute but rather many protected attributes, for example, a *combination* of gender, race *and* age can be the basis of discrimination, leading to the problem of *multi-dimensional discrimination/fairness* [3]. Empirical evidence consistently suggests

[1]While "protected ground" (law) and "protected attribute" (ML literature) both refer to the criterion, e.g. sex, to be protected - against discriminatory treatment/impact or bias -, a 1-1 mapping between these terms cannot be assumed. This is discussed in 4.2.
[2]This is often referred to as single-axis discrimination in literature.
[3]We use the terms multi-dimensional discrimination and multi-dimensional fairness to compare anti-discrimination law and statistical fairness in ML in their respective approaches to challenges of multi-dimensional categories of identity. Fairness and non- or anti-discrimination are by no means synonymous. While discrimination as a

that discrimination in the real world is often multi-dimensional [23]. In a 2015 Eurobarometer for example, approximately a quarter of discrimination cases, as reported by the persons affected, was based on multiple grounds [24]. Multi-dimensional discrimination especially impacts ethnic minorities, as evidenced by a finding from 2010 by the European Fundamental Rights Agency (FRA) that 14% of respondents from ethnic minorities indicated feeling discriminated against on multiple grounds in the 12 months prior to the survey, with most multi-dimensional discrimination experienced by ethnic minority women [22]. Completing the picture, scholars in the field also assume that a large share of discrimination is multi-dimensional [79, 89] or even consider multi-dimensional discrimination the rule and mono-dimensional discrimination the exception [5]. Multi-dimensional discrimination also seems to play an important role in AI systems, as concrete examples of, e.g. low accuracy of facial analysis for black females [8] indicate. Targeting and profiling practices have been noted as factors possibly exacerbating the occurrence of multi-dimensional discrimination [89]. However, only in the last years the topic of multi-dimensional discrimination has caught the attention of the fairness-aware ML community [25, 26, 35, 48, 49, 72, 91, 92]. Considering multi-dimensional discrimination in ML algorithms introduces new challenges, from how to define fairness in the presence of multiple protected attributes to how to mitigate multi-dimensional discrimination. For the former, we find that the term of intersectionality from the legal domain is mainly adapted, whereas other concepts like cumulative and sequential discrimination are less used or developed. For the latter, a key challenge for ML is data scarcity as protected subgroups defined by the intersection of multiple protected attributes become smaller or even empty as the number of protected attributes increases.

The goal of this survey is to draw attention to this important topic and provide an overview of existing approaches in legal and ML literature. We intend to thereby contribute to a "legal-technical argumentation framework" [10]. Juxtaposing types of multi-dimensional discrimination from legal scholarship with operational definitions of multi-fairness in ML enables us to draw connections between the two domains, highlight differences on the conceptual level and identify directions for future research. The remainder of the paper is organised as follows: We start with an introduction of multi-dimensionality in law and a typology of multi-dimensional discrimination (Section 2) with concrete examples. Definitions of multi-dimensional discrimination in ML and associated challenges are discussed in Section 3. A critical comparison of the two domains is presented in Section 4 followed by open challenges and directions for future work in Section 5.

legal term can be found only w.r.t. a specific case, the adherence to anti-discrimination law may be said, depending on one's theoretical stance, to be one facet of fairness as a more encompassing - and very context-dependent - notion. On the inherent vagueness of fairness and its relation to the law, see [62, p. 523ff.].

2 MULTI-DIMENSIONAL DISCRIMINATION IN LAW

2.1 Theoretical background

Traditional accounts of discrimination depict it as a selective interference with the determination of outcomes based on certain, generally separate, traits of the persons affected [15]. However, critical legal scholarship has long challenged such views, especially since the coining of the term "intersectionality" as a theoretical lens pioneered by Kimberlé Crenshaw in 1989 [15]. Analysing the situation of black women, i.a. under the then prevailing anti-discrimination doctrine, Crenshaw demonstrated how it failed to address harms to this group placed at the "intersection" of race and sex discrimination by equating their discrimination to either that of blacks or women, each modelled after the more privileged members - black men and white women - of said groups [15, p. 150ff.]. Intersectionality thus popularised the notion that discrimination as a complex phenomenon is insufficiently addressed if conceptualised around separate, insular spheres of disadvantage,[4]. This served to bring forms of discrimination which involve more than one legally protected ground to the attention of legal scholarship. Importantly, structural intersectionality has worked to refocus the theory away from its reception as a mere theory of subgroup-identities and towards the dynamics of power relationships actualised through identity categories [27, p. 31]. Intersectionality thus centers power relationships [27, p.30ff.] as the issue that anti-discrimination law should address, rather than narrowly conceptualising discrimination as an (incidental or selective) interference with the process of determining outcomes [15, p.151]. Lately, some notable work has also covered possible practical implications of these theoretical frameworks for current anti-discrimination law in the EU, e.g. [27, 79, 87, 89].

In the absence of a legal definition or universally accepted terminology for discrimination involving more than one protected ground [27, 59], we choose to look at interactions between various forms of discrimination under the term *multi-dimensional discrimination*, as suggested in some of the legal literature [5, 74, 75] building on the insights of intersectionality. This term also connects to some theoretical work that aims to expand on intersectionality [42, 43] in addressing subordination, while being broad enough to capture a variety of *interactions among different protected grounds or attributes*, from one incident of discrimination involving multiple grounds to multiple discriminatory incidents over time, as an umbrella term. Lastly, it helps avoid confusion between intersectionality as a theoretical framework and intersectional discrimination in a narrow sense, set out below as a discrimination based on a nexus of intertwined grounds.

2.2 Types of multi-dimensional discrimination

From this starting point, the ways in which different grounds interact can be used to distinguish between different types of discrimination. We use a common typology [27, 79], which identifies (1)

[4]Crenshaw's work advanced further important criticism towards anti-discrimination theory and practice, for example the notion of discrimination as a single, selective incident of interference with decision-making processes, the role of discriminatory intent or the "but for" approach of comparing groups. While in this work we focus on the implications for multi-dimensional protected grounds or attributes, some of these points are closely related and will arise in the discussion.

cumulative or additive discrimination (Section 2.2.1), (2) intersectional discrimination (Section 2.2.2) and (3) sequential discrimination (Section 2.2.3).

2.2.1 Cumulative discrimination. In cumulative (often also termed additive) discrimination, a disadvantage is linked to two or more grounds of discrimination, e.g. gender and race. These are, however, *conceptually separable*, meaning that one can identify distinct disadvantages linked to each involved ground which "add up" when the grounds are observed together.

Example 1. A hypothetical example of cumulative discrimination is shown in Figure 1(a), displaying the mean height of four subgroups defined on the basis of the two grounds sex and nationality. Height varies statistically between sexes and nationalities (here, we use mean height of 19-year olds in 2019 according to [63]). The impact of height requirements, e.g. for jobs in the security sector [64], thus differs depending on nationality (which in turn correlates with ethnic origin) and sex, but the disadvantage "adds up" for women of certain nationality (and correlatedly, ethnic origin). While this particular example can constitute sex discrimination under EU law [5], the dimension of potential ethnic discrimination has gone largely unexplored (see, however, [13, p. 165f.] for US law).

2.2.2 Intersectional discrimination. In intersectional discrimination, the grounds of discrimination involved are merged into one, and *cannot be separated* in the analysis without omitting the legally relevant disadvantage concerned.[6] Intersectional discrimination thus affects subgroups defined by a combination of grounds.

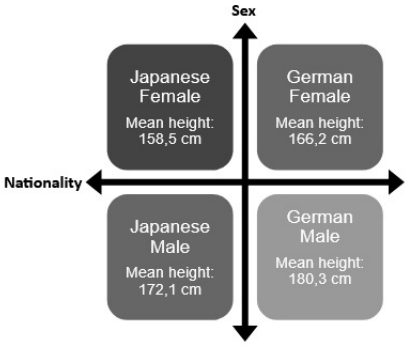
Example 2. A concrete example of intersectional discrimination is shown in Figure 1 (b). The prohibition of wearing headscarves, e.g., at the workplace (as discussed and implemented for teachers in some kindergartens in Germany[7]) specifically affects religious Muslim women, a subgroup of both women and Muslims.

2.2.3 Sequential discrimination. In sequential discrimination, discrimination occurs on the basis of the same or different grounds over several incidents in temporal sequence (see Example 3).

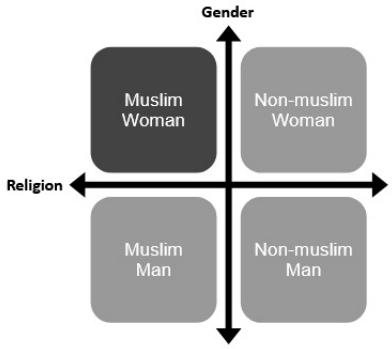
Example 3. An example can be found in Figure 2, where potential points of discrimination in a person's work life are shown. Discrimination at the earlier stages is likely to also affect the outcome at later stages.

This is the "normal case" for anti-discrimination law. To quote [27]: "This is perhaps the easiest to deal with. Each incident can be assessed on a single ground and compensation awarded accordingly". Doctrinally, anti-discrimination law usually looks at the different incidents as distinct cases, not considering possible connections or interrelations. From a societal perspective, however, it is important to recognise these cases because repeated discrimination over time can cause more severe allocative and representational harms to those affected.

Sequential incidents of discrimination can also occur in different steps in a combined process, which produces overlaps with cumulative discrimination (see Example 4). The implications of these two types of discrimination for law and ML are discussed in Section 4.3.2.



(a) Cumulative discrimination.



(b) Intersectional discrimination.

Figure 1: Cumulative and intersectional discrimination: the impact of height requirements differs according to nationality and sex, whereas the prohibition of headscarves specifically affects Muslim women. Darker shades indicate stronger impact. Sex refers to the biological property influencing height, gender to social roles influencing the wearing of a headscarf according to religious practices.

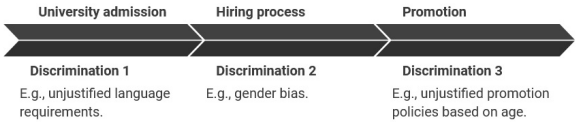


Figure 2: An example of sequential discrimination in work life based on three different grounds.

[5] ECJ, C-409/16 - *Kalliri*

[6] The entanglement of grounds has led scholars to consider intersectional discrimination "mono-dimensional" in terms of legal doctrine [87]. In our not exclusively doctrinal perspective, we nevertheless choose to include it under multi-dimensional discrimination to emphasise its resulting from a combination of multiple *grounds* which are often conceived of as separate.

[7] See, e.g. ECJ, C-804/18 - *WABE eV*. Formally, such rules often concern all "visible religious symbols" equally but due to the demographic structure of societies and specific religious expressions disproportionately affect Muslim women. This leads to the issue whether this constitutes direct or indirect discrimination [40] on the grounds of religion.

Example 4. An example of sequential discrimination in a hiring process is given in Figure 3. An old woman with disability might suffer discrimination on the first occasion due to her gender, later due to her disability and finally due to her age. Here, the final outcome of the process will be the point at the centre of legal review.

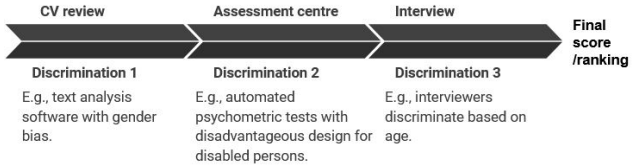


Figure 3: A recruitment process involves several steps with potential for discrimination.

2.3 Multi-dimensionality in EU law

Despite legal scholarship increasingly picking up intersectionality more generally and possible interactions of discrimination grounds in particular, this development has largely been met with hesitance in legal practice[27, 54] of EU private anti-discrimination law.[8] Discrimination under the law consists of either a *direct* discriminatory treatment or, in the case of *indirect* discrimination, the application of a seemingly neutral criterion, provision or practice leading to a particular disadvantage of a group or individual. These need to be tied to ("based on" in the language of the law) a *protected ground.*

This framework applies only to grounds of discrimination listed exhaustively in the different Directives[9]: *racial and ethnic origin*, *sex*, *religion and belief*, *disability*, *age* and *sexual orientation.* Addressing cases involving different grounds is complicated by differences in scope and justification for the different grounds of discrimination. EU law rather vaguely recognises "multiple" discrimination of women[10]. Notwithstanding this textual reference, the European Court of Justice (ECJ) has cautiously evaded the issue and on one occasion expressed resistance to establishing intersectional discrimination as a special type of discrimination [11], although some see signs of an implicit approach compatible with multiple dimensions of disadvantage in a wide interpretation of grounds in some cases [27, 89]. Hence, our discussion in this regard is mostly drawing from legal scholarship.

[8]While we introduce and illustrate multi-dimensional discrimination with a view to current EU private anti-discrimination law such an analysis can be applied to all anti-discrimination provisions. This field is mainly codified in four EU Directives: Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin; Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation; Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services; Directive 2006/54/EC of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast) and their respective implementation in member states.

[9]To that effect for Directive 2000/78/EC, ECJ, C-13/05 - *Chacón Navas* par. 56; C-306/06 - *Coleman* par. 46 and C-354/13 - *Kaltoft* par. 36.

[10]Mentioned, but not elaborated on, in recital 14 of Directive 2000/43/EC and recital 3 of Directive 2000/78/EC. EU member state laws may offer stronger protection: of note, Spain passed a law explicitly covering intersectional discrimination in 2022, Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación.

[11]Cf., as the latest examples C-808/18 - *WABE eV* par. 58 and case C-443/15 - *Parris* par. 80, for analysis see [39].

3 MULTI-DIMENSIONAL DISCRIMINATION IN AI SYSTEMS

We first introduce the typical fairness-aware learning setup and basic concepts (Section 3.1) and then we survey existing definitions of multi-discrimination in ML, organised into cumulative (Section 3.2), intersectional (Section 3.3) and sequential discrimination definitions (Section 3.4).

3.1 Basic concepts and problem formulation

We follow the typical fully supervised batch learning fairness-aware setup. Let $D = (u^{(i)}, s^{(i)}, y^{(i)}) \sim P$ be a dataset of n instances, with each instance drawn as an independent and identically distributed (i.i.d.) sample from $P(U \times S \times Y)$, where U is the subspace of *non-protected attributes* (e.g. height, weight, education, etc.), S is the subspace of *protected attributes* (e.g. race, gender, religion, etc.), and Y is the class/target attribute (e.g. loan default). For simplicity, we assume binary classification: $Y \in \{+, -\}$. The non-protected and protected attributes together define the feature space $X = U \times S$, so $x^{(i)} = (u^{(i)}, s^{(i)}), 1 \leq i \leq n$. Let S consist of k protected attributes: $S = S_1 \times S_2 \times \cdots \times S_k$. For simplicity, protected attributes are assumed to be binary: $\forall_{j=1,\cdots,k} S_j \in \{g^j, \overline{g^j}\}$, where g^j and $\overline{g^j}$ represent the *protected group* (e.g. female) and the *non-protected group* (e.g. male), respectively w.r.t. the protected attribute S_j (e.g. gender).

The intersection of different protected attributes defines the so-called *subgroups*[12]. For example, based on the binary protected attributes age, race and gender, eight different subgroups are formed including the subgroups:"young-black-women" and "old-white-men". The *collection of subgroups* is denoted by $\mathcal{SG}$ and defines as:

$$\mathcal{SG} = \{sg_m = s^1 \cap s^2 \cap \cdots \cap s^k \mid s^i \in \{g^i, \overline{g^i}\}, i = 1, \cdots k\} \quad (1)$$

Broadly, discrimination for the supervised learning set-up can be expressed in terms of differences in model performance across different subgroups; these differences can be evaluated w.r.t. one class of interest (typically, the positive class) or w.r.t. both classes. Moreover, model performance can be evaluated in terms of different conditions: just predictions or predictions given the ground truth. We use the generic notation C to denote these extra conditions.

For mono-discrimination, this broad definition can be expressed as differences in expected outcomes of the groups:

$$\mathcal{F}_{S_j} \equiv abs(P(\hat{y} \mid g^j, C) - P(\hat{y} \mid \overline{g^j}, C)) - \epsilon \quad (2)$$

where $\mathcal{F}_{S_j}$ is the *mono-discrimination* w.r.t. S_j, $\hat{y}$ is the predicted outcome, ϵ is the tolerated discrimination threshold, $abs()$ is the absolute value function, and C refers to the additional conditions w.r.t. class(es) and measure of interest. For example, Statistical Parity [20] only focuses on predictions in the positive class so $C : [\hat{y} = +]$, and Equation 2 can be re-written as: $\mathcal{F} \equiv P(\hat{y} = + \mid g) - P(\hat{y} = + \mid \overline{g})$. Equal Opportunity [32] focuses on the correct predictions in the positive class, so $C : [\hat{y} = y \mid y = +]$ and Equation 2 can be re-written as: $\mathcal{F} \equiv P(\hat{y} = y \mid y = +, g) - P(\hat{y} = y \mid y = +, \overline{g})$.

[12]We use the term *group* (*subgroup*) if a single (respectively, more than one) protected attribute(s) is used for the definition of the (sub)group.

3.2 Cumulative discrimination

Cumulative discrimination is a natural extension of mono-discrimination to the multi-discrimination case with the *conceptually isolated groups* defined separately based on each of the protected attributes. Early works on multi-discrimination [1, 93] target cumulative discrimination and formulate the problem as solving a *set of fairness constraints*, one for each protected attribute. Following the generic formulation of Equation 2, the cumulative discrimination over the protected attributes $S = S_1 \times \cdots \times S_k$ can be defined as an operation on a vector of mono-discrimination:

$$\mathcal{F}_S^{\odot} \equiv \odot(\mathcal{F}_{S_1}, \cdots, \mathcal{F}_{S_k}) \tag{3}$$

where $\mathcal{F}_{S_j} \in \mathbf{R}$ is the measured mono-discrimination w.r.t. S_j, $j = 1, \cdots, k$ using the specified mono-discrimination case (c.f., Eq. 2), and $\odot : \mathbf{R}^k \rightarrow \mathbf{R}$ is an operator e.g. $max()$, $sum()$ that defines how to "combine"/"assess" all-together the multiple mono-discrimination. More recent works [72, 91] argue that any fairness notion which aims to find the *maximum* discrimination towards any protected attribute S_j among the set of protected attributes S (i.e. using operator $max()$ for $\odot()$ in Eq 3), is equivalent to the generalised multi-dimensional discrimination formulation of Equation 3.

Although in legal practice a separate consideration of grounds in principle allows redress of cumulative discrimination (see Section 4.3.1), its application in ML comes with flaws as it targets discrimination in groups defined on single protected attributes but not in subgroups defined based on the intersection of several protected attributes. This drawback was first studied in [49] who also termed the drawback *fairness gerrymandering*. In particular, it was shown that a model trained to be individually fair w.r.t. different protected attributes can still discriminate certain subgroups defined based on the intersection of several protected attributes.

The problem can be elaborated with a hypothetical example. During a routine raid by police in some part of the world where drug trafficking is an existing major issue, some suspects (say 100) are taken into custody. Now assume that based on the protected attribute *gender* the suspected people can be divided into 60:40 *male*:*female*, and based on the protected attribute *race* the distribution is 60:40 *black*:*white*. Considering both race and gender, 4 subgroups are formed: (*White Male*, #20), (*White Female*, #20), (*Black Male*, #40), (*Black Female*, #20). Let us further assume an ML model that is deployed to classify the questioned person as either "drug trafficker" or "innocent".

Figure 4 illustrates the finer distribution of the population based on gender and race as well as the results of the ML model. As we can see, the model conditioned on the equality of positive predictions i.e Statistical Parity[13] between groups defined by gender and between groups defined by race, is fair: w.r.t. *gender* 50% of males (30 out of 60) and 50% of females (20 out of 40) are predicted as suspected *"drug trafficker"*. Likewise w.r.t *race* 50% of blacks (30 out of 60) and 50% of whites (20 out of 40) are predicted as suspected *"drug trafficker"*. Looking at the subgroups however, the model is unfair e.g. to *white females* (20 out of 20 are predicted as suspected) with 100% suspect prediction compared to *white males* (20 out of 20 are predicted as innocent) with 0% suspect prediction.

[13]The example would work with other fairness notions as well. We use statistical parity due to its simplicity.

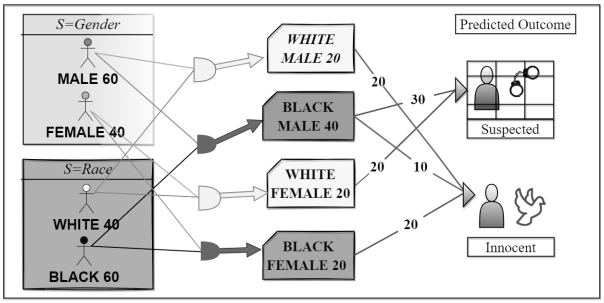


Figure 4: Prediction distribution of a hypothetical drug trafficker detection model for different population (sub)groups.

3.3 Intersectional discrimination

Intersectional discrimination looks at the subgroups defined on the intersection of multiple protected attributes (cf. also Section 2.2.2). [35] was the first to study the problem of fairness in finer subgroups, though the limitations of cumulative discrimination and the need to focus on subgroups were clearly outlined in [49]. A generic definition of intersectional discrimination for the protected attributes $S = S_1 \times \cdots \times S_k$ (adapting from the general definition of Eq. 3) can be formulated as an operation over a vector of subgroup-specific discrimination:

$$\mathcal{F}_{\mathcal{SG}}^{\odot} \equiv \odot(\mathcal{F}_{sg} | \forall sg \in \mathcal{SG}) \tag{4}$$

where $sg \in \mathcal{SG}$ is a subgroup (cf. Equation 1), $\mathcal{F}_{sg} \in \mathbf{R}$ is the measured discrimination w.r.t. sg, and $\odot : \mathbf{R}^{|\mathcal{SG}|} \rightarrow \mathbf{R}$ is an arbitrary operator, e.g. $max()$ that defines how the different subgroup discrimination should be "combined/interpreted" all-together.

Intersectional discrimination is the most vividly studied multi-dimensional discrimination type in ML literature. Different methods vary mainly w.r.t. how they define the discrimination for each subgroup, i.e. $\mathcal{F}_{sg}$, therefore, hereafter we focus on this aspect. Not all methods propose an explicit "combination" over the discrimination of the subgroups, but rather try to optimise fairness for each subgroup during discrimination mitigation [50, 58, 60, 76].

Statistical Parity Subgroup Fairness (SPSF) [49]: Kearns et al. [49] introduced the term "fairness gerrymandering" to describe the case where a classifier appears to be equitable when considering any protected attribute alone, e.g. only gender or only race, but might be unfair when looking at the intersection of different protected attributes (e.g. black women). To account for intersectional discrimination they introduced *Statistical Parity Subgroup Fairness* (*SPSF*), an extension of the *Statistical Parity* (SP) [20] definition for mono-discrimination.

The main idea is that the difference between the acceptance rate (probability of positive prediction) $\mathbf{P}(\hat{y} = + \mid sg)$ of any subgroup $sg \in SG$ from the overall acceptance rate $\mathbf{P}(\hat{y} = +)$ proportional to the relative size of the subgroup $\mathbf{P}(sg)$ in the data, must be smaller than an allowed discrimination threshold ϵ. More formally:

$$\mathcal{F}_{sg} \equiv \mathbf{P}(sg) \times abs(\mathbf{P}(\hat{y} = +) - \mathbf{P}(\hat{y} = + \mid sg)) - \epsilon \tag{5}$$

The threshold $\epsilon \in [0, 1]$ quantifies the amount of allowed discrimination towards any subgroup $sg \in \mathcal{SG}$. The relative size of the subgroup $\mathbf{P}(sg)$ allows avoiding fairness overfitting by ignoring

discrimination for small fractions of the population (i.e. subgroups which are very small in size, e.g. a singleton subgroup).

A major drawback of this definition is the possibility of high false positive rates in order to balance the acceptance rate among the different subgroups; which is a common critique of *SP* [32, 83] and stems from the fact that only predictions but not ground truth are considered. Further, the method relies upon the subgroup probability $\mathbf{P}(sg)$ estimated from the data and is therefore prone to biased data representations. The advantage of this definition lies in scenarios where a subgroup sg has very few positive instances and comparatively many negative ones. In such a case, since the relative size of the subgroup $\mathbf{P}(sg)$ is high, the discrimination in this subgroup w.r.t. the positive class is boosted despite the small number of positive instances. Such scenarios are highly likely when the number of protected attributes is large.

False Positive Subgroup Fairness (FPSF) [49]: FPSF comprises an extension of the widely used mono-discrimination notion of *Equal Opportunity* (Eq.Opps) [32] that checks equality of positive predictions between two demographic groups, assuming that people in this group qualify (i.e. the ground truth is positive). More precisely, it defines subgroup discrimination as the difference between incorrect acceptance (false positive) rate $\mathbf{P}(\hat{y} = + \mid y = -, sg)$ on a given subgroup sg and incorrect acceptance $\mathbf{P}(\hat{y} = + \mid y = -)$ on the entire population (dataset) proportional to the relative size of the negative subgroup $\mathbf{P}(y = -, sg)$, to be less than a given discrimination threshold ϵ. More formally:

$$\mathcal{F}_{sg} \equiv \mathbf{P}(y = -, sg) \times abs(\mathbf{P}(\hat{y} = + \mid y = -) - \mathbf{P}(\hat{y} = + \mid y = -, sg)) - \epsilon \tag{6}$$

By considering ground truth labels, *FPSF* overcomes the risk of high false positives rates (see critique on *SPSF*). However, like *SPSF* it relies upon distribution of the subgroups, specifically in the negative (-) class and is therefore prone to biased representations. This has been criticised in [26], where it is argued that the concept of subgroup fairness, due to the consideration of the subgroup probability $\mathbf{P}(y = -, sg)$, is affected by the population size of the subgroup $|sg|$. Thus, a discrimination towards a small subgroup sg gets unfairly overlooked.

Differential Fairness (DF) [26]: Foulds et al. [26] criticised the concept of subgroup fairness for its inability to tackle disproportionate distribution of subgroups and proposed *Differential Fairness* (DF) which extends the 80% rule of the U.S. "Equal Employment Opportunity Commission" [73] to multiple intersectional subgroups. The idea here is to restrict ratios of outcome probabilities between pairs of subgroups under a predetermined fairness threshold e^{ϵ}. More formally:

$$\mathcal{F}_{sg_j, sg_i} \equiv \frac{\max[\mathbf{P}(\hat{y} = c \mid sg_j), \mathbf{P}(\hat{y} = y \mid sg_i)]}{\min[\mathbf{P}(\hat{y} = c \mid sg_j), \mathbf{P}(\hat{y} = y \mid sg_i)]} - e^{\epsilon},\ c \in \{+, -\} \tag{7}$$

where ϵ is an admissible discrimination towards any subgroup, c is the class of interest. The value of ϵ can be set for different pairs subgroups, which can be determined using various factors such as difference in their data distribution, known historical bias, required economic utility, etc.

The authors showed that the **DF** definition closely follows data privacy definitions [51] and provides provable privacy and fairness guarantees. However, the *DF* definition is explicitly designed to extend the 80% rule [73] between any two subgroups sg_i and sg_j, which identifies disparate impact in cases where $P(y \mid sg_i)/P(y \mid sg_j) \leq 0.8$, for a disadvantaged subgroup sg_i and best performing subgroup sg_j. The definition is very closely related to the mono-discrimination definition *Statistical Parity* [20] (cf. 3.1), as it focuses only on the predicted output ignoring the ground truth.

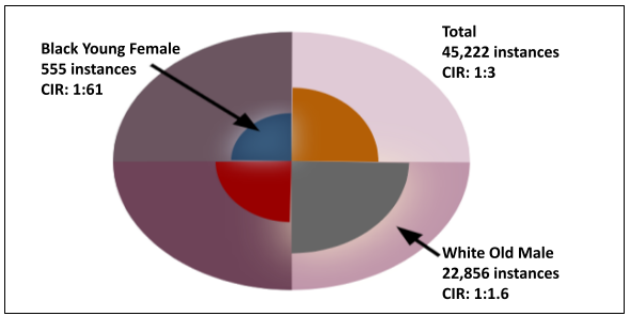


Figure 5: Illustration of the subgroup data scarcity and subgroup class imbalance problems for the Adult dataset.

Worst Case Fairness (WCF) [29]: A more recent work [29] studied the Differential Fairness (cf. Equation 7) and tried to extend the definition which generalises to match all possible mono-discrimination definitions. They formulated the discrimination definition as a *worst-case* comparison between subgroups under a given condition C, where the condition C (cf. Sec. 3.1) is applied to render the subgroup discrimination definition comparable to a specified mono-discrimination definition [20, 32, 70]. Unlike previous works[26, 49], they defined discrimination over the entire collection of subgroups $\mathcal{SG}$ as a min-max ratio of prediction probability over any subgroup. Their definition compares worst performing subgroup $(\min\{P(\hat{y}|sg, C)|sg \in \mathcal{SG}\})$ to the best $(\max\{P(\hat{y}|sg, C)|sg \in \mathcal{SG}\})$, formally defined as:

$$\mathcal{F}_{\mathcal{SG}} \equiv 1 - \frac{\min\{P(\hat{y}|sg, C)|sg \in \mathcal{SG}\}}{\max\{P(\hat{y}|sg, C)|sg \in \mathcal{SG}\}} \tag{8}$$

This definition provides the scope to have a good overall evaluation of discrimination, but it fails to provide, when needed, an in-depth information of per-subgroup discrimination independently.

Discussion: Mitigating intersectional discrimination provides the primary advantage of protecting against discrimination of finer sub-populations. However, with many protected attributes, the number of possible subgroups grows exponentially. Even assuming all protected attributes are binary, we get $2^{|S|}$ subgroups, where S is the set of protected attributes. This gives rise to a problem of data scarcity within the subgroups, which means there exist sub-populations with limited or no data, making it hard to properly (machine) learn the subgroups.

To better understand the problem, we take a look at the popular "Adult" dataset. The task is to predict whether the income of a person exceeds $50K$/year, with greater than $50K$/year being the positive class. The dataset contains $\approx 45k$ instances, we consider *race*, *gender*, and *age* as protected attributes and assume each to be binary [72], which gives us 8 subgroups in total. In Fig.5 we highlight the data distribution for some of the subgroups. The

first observation is that there is a *subgroup scarcity problem* as the subgroups have very different cardinalities. For example, the subgroup "Black Young Female" has only 555 (1.23%) instances, whereas the "White Old Male" subgroup has almost 41 times more, namely 22,856 (50.54%) instances. Besides subgroup scarcity, we observe varying positive:negative *class imbalance ratios* (CIR) for every subgroups with higher imbalance for minority subgroups. As a concrete example, the CIR for the underrepresented subgroup "Black Young Female" is 1:61 (9 positive out of a total 555 instances) which more than 38 times higher than majority subgroup "White Old Male" with 1:1.6 CIR (8.6k positives out of 22.8k instances).

3.4 Sequential discrimination

Sequential discrimination by definition requires a sequence of events (see also Section 2.2.3). The order of the events is very important as discrimination at an earlier stage in the sequence can have a larger impact than discrimination at a later stage [7]. This topic has only recently attracted the interest of the Fair ML community. Existing works [68] are mainly aimed at long-term (future outcomes) implications [7, 16, 41, 45, 94] of discriminatory outcomes. Also, many of the works are aimed at solving sequential discrimination at an individual level, as it often relies upon feedback from the dynamic/continuous system and how an individual may act after receiving the feedback [16, 45, 94]. However, staying true to the focus of this work we will keep the discussion centred to multi-discrimination within the supervised learning setup as introduced in Sec. 3.1. However, differently from the batch learning setup of Sec. 3.1, for sequential discrimination, data arrive as a sequence of batches (each batch corresponding to a distinct event, for example, the different steps of the hiring pipeline, see Example 4). More formally, $\mathcal{D} = [D_1, D_2, \cdots, D_T]$, where $D_1 \cdots D_T$ is in a ordered sequence from 1 to T, and each $D_t \in \mathcal{D}$ consists of n_t instances such that $n_1 \geq n_2 \geq \cdots \geq n_T$. The instances of each event $D_t = (u^{(i)}, s^{(i)}, y_t^{(i)}) \sim P_t$ are drawn as i.i.d samples from the underlying distribution $P_t(U \times S \times Y)$ such that $y_t^{(i)} = + \implies \forall_{1 \leq l \leq t} y_l^{(i)} = +$, i.e. any instance (i) labelled as positive in D_t means a positive label for (i) in all the events observed before D_t (but vice-versa is not true). Following [7], for a given protected attribute $S_j \in \{g^j, \overline{g}^j j\}$ and a sequence of observed events $1, \cdots, T$, sequential fairness w.r.t. S_j can be defined as:

$$\mathcal{F}_{S_j}(T) \equiv \frac{P(\hat{y}_T \mid \overline{g}, C)}{P(\hat{y}_T \mid g, C)} - \prod_{t=0}^{T-1} \frac{1}{(1 + \mathcal{F}_{S_j}(t))} \tag{9}$$

where $\mathcal{F}_{S_j}(T)$ is sequential discrimination over the sequence T and $\hat{y}_T$ is the final predicted outcome at the end of the sequence. Notice that the definition in Eq. 9 is recursive and considers in a multiplicative form the discrimination observed from the beginning of the process till previous step $[0, \cdots, (T-1)]$. Intuitively, the multiplicative property of the definitions is designed to penalise systems for higher discrimination in an earlier step/stage of the decision process. Naturally, at the beginning of the sequence i.e. at $t = 0$, the value of $\mathcal{F}_{S_j}(0)$ is not generated by the AI system and is given externally as an input to rectify historical bias known in the society. For better understanding, let us consider the 3-stage hiring process example depicted in Fig 3. Considering gender as the protected attribute, one observes that in CV review step 80% male and 50% applicants get forwarded to the assessment (2nd) step. Now, considering $\mathcal{F}_{Gender}(0) = 0$ i.e no existing/external historical bias w.r.t gender being considered, we get $\mathcal{F}_{Gender}(1) = 0.6$ i.e a discrimination score of 0.6 after time step 1 in the pipeline. Then, even if in assessment step the system equally accepts both the groups we get $\mathcal{F}_{Gender}(2) = 0.37$, and in the interview step to have discrimination (zero) free prediction the system needs to predict with $P('accept'|female) \approx 2P('accept'|male)$, since, $\prod_{t=0}^{2} \frac{1}{1+\mathcal{F}_{Gender}(t)} \approx \frac{1}{2}$.

The extension of sequential mono-discrimination to a multi-discrimination sequential scenario could be a straightforward practice by defining discrimination $\mathcal{F}_{S_j}(T)$ w.r.t. every protected attribute S_j, for $j = 1, \cdots, k$, and then combining them using an arbitrary operator $\odot()$ as in Eq. 3. The definition presented in Eq. 9 is also applicable to more than one sub-populations [7], and thus can be generalised to a multi-discrimination definition on the intersectional subgroups $\mathcal{SG}$ (cf. Eq. 1):

$$\mathcal{F}_{sg_k}(T) \equiv \frac{P(\hat{y}_T \mid sg_m, C)}{P(\hat{y}_T \mid sg_k, C)} - \prod_{t=0}^{T-1} \frac{1}{(1 + \mathcal{F}_{sg_k}(t))} \tag{10}$$

where $\mathcal{F}_{sg_k}(T)$ defines the sequential discrimination on any underprivileged subgroup sg_k (say *black female*) compared to the most privileged subgroup sg_m (say *white male*) in the prediction $\hat{y}_{t-1}$ on the sequence of observed events till D_T.

Though no prior work has specifically addressed the sequential multi-discrimination problem, a generalisation of sequential mono-discrimination to the "multi" case seems doable. The main issue however, is still coming from data scarcity. Most of the work addressing the sequential discrimination problem relies on synthetic data generation [16, 94]. Although the ACS-PUMS dataset introduced in [17] and the Intesa Sanpaolo bank dataset used in [41] possess the temporal property which is important to address a sequential multi-discrimination problem, none of them has the pipeline information which either includes stage-wise label information [7] or change of feature/transitional information due to a decision received in the previous stage [94]. This signals a need for real-world datasets with temporal and stage-wise information that can be investigated to analyse and develop better sequential multi-fair models.

4 DISCUSSION

After reviewing the fair ML literature w.r.t. the typology from legal scholarship, some aspects bear pointing out. In addition to technical and policy-related challenges, differences in concepts of discrimination and fairness are highlighted by the disciplines' relationship with multi-dimensionality. In the fairly young interdisciplinary discourse on fairness in AI-based decision-making, law and ML have to learn how to incorporate conceptual work from their counterpart.

4.1 Caveat: Statistical fairness and anti-discrimination law

The relationship between fair ML and anti-discrimination law is evolving and far from clarified (see [86], [85] for EU law, [6] for U.S. law). Particular questions concern whether and when anti-discrimination law mandates (or forbids) the adherence to statistical fairness measures in AI-based decision making. Clearly, obligations

to apply fairness measures must be proportionate, and maximising utility (meaning predictive accuracy) for the decision-maker under fairness constraints can help ensure such proportionality [30]. But the specific balance is yet to be determined. The same applies to the choice of fairness metrics [28]. It has been demonstrated in detail that no one particular notion of fairness maps definitively to the requirements of anti-discrimination law in judicial review due to their inherent contextuality [86]. Although some authors suggest accounting for historical disadvantages with "bias transforming" fairness metrics such as statistical parity [85], such practices are not part of binding requirements under current doctrine and will most likely be tested in court. Hence, with the academic debate still ongoing and without specific case law, our considerations have to be taken with the caveat that the requirements under anti-discrimination and other fairness-related legislation cannot be said to mandate the application of a specific fairness measure. Nor do unequal outcomes as measured by fairness metrics necessarily imply discrimination in a legal sense. Research is still building a "legal-technical argumentation framework"[10] to which we aim to add the multi-dimensionality of discrimination and fairness.

4.2 Categorical grounds and attributes in law and ML

Fundamental differences can be found in the conceptualisation of protected grounds or attributes, respectively. While our reflections on categorical grounds and attributes also apply to mono-dimensional discrimination and fairness, they become particularly apparent once multi-dimensionality is introduced. Perhaps unsurprisingly, the recognition that grounds cannot be neatly separated and defined in isolation often leads to calls for an anti-discrimination law without such categories or with open-ended lists of categories [55, 56, 79] or, alternatively, to their wide, contextual or capacious interpretation[27], also for algorithmic discrimination in particular [89].

Rather than innate attributes, intersectional and multi-dimensional theories of discrimination posit that the "categories" named by grounds of discrimination are best understood as *categorisations*, embedded in a process of social ascription and categorisation. They oppose views which are critically termed "essentialism" or "essentialising" and entail the assumption that categories and questions of identity at the roots of discrimination are (only or predominantly) results of innate and static group differences. Relatedly, essentialist views tend to define groups as internally unitary or homogenous [13, 56], whereas intersectional and multidimensional theories emphasise the diversity within groups (see, e.g. [42, fn. 16]). Essentialising grounds of discrimination forces claimants under anti-discrimination law to define protected groups by a presumed difference from others and sort themselves into such a group [13, p. 168], while they may not share the traits considered defining of that group. In other words, an essentialist reading of discrimination grounds risks reifying and reproducing the categorisations it is intended to protect against. Intersectional analysis, on the other hand, emphasises that the bases for discrimination are often context-dependent social phenomena [59]. In structural intersectionality, e.g., discrimination is considered oppression along axes of social power relationships [5, 27]. Such a localisation of discrimination in the intersections of oppression is deemed central to the framework of intersectionality [52].

EU law arguably enables a non-essentialist interpretation of discrimination grounds. The term racial origin, e.g. is used in EU anti-discrimination law to combat *racial discrimination*, accompanied by an explanation that this does not imply acceptance of theories of separate human races (recital 6 of Directive 2000/43/EC) and the ECJ has applied the law to cases of "discrimination by association", implying a non-essentialist reading [14]. More specifically, [89] argues that EU anti-discrimination law can be read as implementing a (non-essentialist) dual conception of protected grounds as a recognition of social identities and as a tool to capture social hierarchies. Depending on the context, protected grounds can thus capture categorising external ascriptions and stereotypes or group affiliations and identities of the affected persons.

ML, on the other hand, utilises the categories in question as *features* or *attributes* and often needs to assume their stability. At first glance, ML methodology therefore seems to align more smoothly with an essentialist conception of discrimination. In fair ML, discrimination grounds appear as protected *attributes* or *groups*. Scholarship has emphasised how race is often assumed to be fixed and mono-dimensional, even in work on algorithmic fairness [31]. But because an essentialist concept of discrimination grounds can end up reproducing the same group differences and hierarchies that anti-discrimination law and fair ML aim to mitigate, critical scholarship is increasingly calling for a shift of focus in fair ML from protected attributes to structural oppression [52] or social hierarchy [36]. Relatedly, work in fair ML research explicitly referencing intersectional theories (such as [49]) has been criticised for missing the lessons of intersectionality, pursuing intersectional fairness only by splitting protected groups further into subgroups [36, 52]. In this regard, Iyola Solanke's concise observation that "[w]hilst attributes may be innate, stigmas are produced" [77], could be an important lesson for fair ML research.

It is, at the least, a challenge to operationalise the contextuality of anti-discrimination laws in ML [86], and multi-dimensional discrimination, which is only insufficiently reflected in law itself, exacerbates this challenge. Nevertheless, scholarship has already highlighted potential methodological routes beyond fixed attributes for ML [31] and for law [11, 89]. The aforementioned criticism also demonstrates that the fairness community extends beyond the design and application of fairness definitions and may be able to integrate to scholarship calling for a wider perspective on the role of algorithms in oppression, e.g. [52]. A re-conceptualisation of application of protected attributes can mean looking at the whole socio-technical process of the introduction of AI-based decision-making in a given environment, including awareness of assumptions about – or constructions of – target variables, desired properties and ground truth as well as the selection and construction of protected attributes and respective labels. While some domains particularly require objectivity and stability in class labels (consider, e.g. skin tone differences in cancer screening), in other contexts, labels for protected attributes may be appropriately obtained from

[14] See ECJ, C-303/06 - *Coleman*, C-83/14 - *CHEZ*; for analysis, see [4, 89]. Note that at the same time, essentialist interpretations persist, as demonstrated for German jurisprudence by [56] or [33, p. 27-29] and it can be questioned whether categories in anti-discrimination law always tend towards essentialism [13].

affected persons themselves where personal identity or group affiliation matters but from external sources where stereotypes need to be countered. Such a process, informed by social sciences, would go some way towards implementing the dual understanding of protected grounds in anti-discrimination law.

4.3 Conceptualisation of types of multi-dimensional discrimination

The terminology regarding different types of discrimination differs; the matter has even been described as a "lexical battlefield" [88]. Some scholars caution against defining seemingly clear-cut types of multi-dimensional discrimination, calling it a "dangerously simplifying complication" [5]. Below, we highlight some complexities within the typology here applied.

4.3.1 Cumulative and intersectional fairness. The criticism regarding a typology of multi-dimensional discrimination becomes apparent with the the distinction between cumulative and intersectional discrimination: intersectionality as a theory is capable of capturing both without the need to define a clear distinction, as the respective disadvantages are manifestations of the underlying intersecting power relationships. In the affected persons' experience, too, the involved grounds are likely to be present at the same time and inseparably. Consequentially, some doubt the usefulness cumulative discrimination as a separate type [75]. As apparent from Figure 1, cumulative and intersectional discrimination both concern how an individual or group is affected when focusing on *subgroups* defined by two or more discrimination grounds or protected attributes. This common property can be observed in discussions about adequate compensation: A court finding of discrimination based on only one of the involved grounds may not reflect the discriminated individual's experience[11] and the extent of the injustice suffered[79]. Some jurisdictions award higher compensation in cases of "multiple discrimination" [12, 27]. Whether higher compensation should be awarded is subject to debate (see [75] for a brief overview) and higher compensation for both cumulative and intersectional discrimination may be motivated by the argument that individuals in subgroups are more vulnerable in many respects[75, 79].

However, despite the somewhat blurry line between the two [5], in intersectional discrimination intersecting grounds of discrimination are so intertwined that they practically constitute a single criterion applied for differentiation. The disadvantage of a discriminated subgroup in intersectional discrimination cannot be explained by "adding up" disadvantages of two or more groups defined by only one protected ground. Precisely this disadvantageous impact on a subgroup leads to challenges in legal protection that do not pertain to cumulative discrimination: While intersectional and cumulative discrimination both may in some cases be captured by invoking only one of the involved grounds in court or invoking grounds separately [27], such a strategy is likely to result in complete review for cumulative discrimination (even if potentially understating a claimant's alleged disadvantage), but not for intersectional discrimination. The ECJ cases mentioned in 2.3 and literature [11, 15, 55] have demonstrated that intersectional discrimination tends to elude judicial analysis altogether when the involved grounds are separated. Moreover, differences between cumulative and intersectional discrimination can be highly relevant in doctrine for potential grounds for justification [87].

ML approaches to intersectional fairness, by focusing on subgroup fairness, also start out by considering different protected attributes fundamentally intertwined. This makes sense because intersectional discrimination involves the specific challenge of data scarcity for subgroups discussed in 3.3. On the other hand, the common focus of cumulative and intersectional discrimination on subgroups can play an important role, as Sections 3.2 and 3.3 demonstrate. Ensuring statistical fairness - by any measure - for subgroups will also do so for groups, thus mitigating intersectional *and* cumulative disparities. Approaches relying only on the notion of cumulative discrimination have the drawbacks highlighted in 3.2, but may be helpful when data scarcity renders a focus on intersections all but practically impossible.

4.3.2 Cumulative and sequential fairness. In sequential discrimination each of the multiple incidents of discrimination in a temporal order may involve any other type of discrimination (monodimensional, cumulative or intersectional). Importantly, there is a potential overlap with cumulative discrimination. Most legal tools against discrimination, such as liability, are retrospective and require a "discriminatory treatment" or "particular disadvantage" in an area covered by anti-discrimination law (employment and, to a lesser degree, supply of goods and services). Legal redress is thus often restricted to a treatment, criterion, provision or practice with a direct and tangible *economic impact* on the affected person [15]. In hiring, e.g. the rejection of a candidate will typically be at the centre of analysis. But applicants usually undergo several steps in a recruitment process, e.g. CV review, assessment centre tasks or psychometric measurement and interviews as shown in Figure 3. Each of these practices bears potential for discrimination [16]. If different steps disadvantage a candidate based on different discrimination grounds, the final decision may present itself as cumulative discrimination.

In ML, however, cases involving cumulative discrimination from a legal perspective may be addressed as a special type of sequential discrimination from an *engineering perspective*: the process can be segmented and each step addressed according to the sequence. Such use-case are sometimes referred to as *fair pipelines* [68]. Research has shown that under certain conditions interventions at one stage can even propagate through the whole process [7]. Due to the presence of a well-defined task and outcome (e.g. fill an open position) such cases offer more concrete options for interventions for fairness than the "typical" sequential discrimination from legal scholarship (cf. Fig. 2), where the effects of discrimination at one decision point on subsequent decisions are hard to determine. Sequential scenarios are well-known and highly important for ML: learning a model is typically the result of a multi-step process,

[15] This applies especially to jurisdictions relying on individual claimants to enforce anti-discrimination law. In C-54/07 - *Feryn*, par. 21-28, however, the ECJ ruled (notably following a type of class action under Dutch law) that an identifiable affected individual was not required for a finding of discrimination if an employer publicly declares their intention to discriminate in hiring.

[16] For examples of potential biases, see https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G for automated CV review and https://www.technologyreview.com/2021/02/11/1017955/auditors-testing-ai-hiring-algorithms-bias-big-questions-remain/ for automated assessment center applications.

from data selection to pre-processing, cleaning, model selection and evaluation. Bias can penetrate in each of these steps; e.g. w.r.t. dataset selection a strong bias towards certain demographics has been shown in visual datasets [21] or, w.r.t. pre-processing, it has been shown that the encoding method for categorical protected attributes can lead to biased models [61]. Addressing such cases as sequential discrimination seems a promising route for fair ML.

5 OUTLOOK

We have introduced the multi-dimensionality of discrimination, taken from legal scholarship, as a not yet fully explored foundational problem of fairness in AI-based decision-making. Looking more closely at different types of discrimination, we can learn how to better address them in decision-making processes. Our review of ML research has shown that the field has begun to address some of the issues raised by multi-dimensionality, predominantly focussing on intersectionally fair algorithms. We have also pointed out obstacles to a common understanding of protected grounds and attributes, and highlighted concerns that essentialist approaches insufficiently reflect multi-dimensionality. These findings raise questions that legal and computer science scholarship have just begun to explore. We can only point to a few of these:

Sources and definition of protected attributes: Research reflecting on protected attributes is needed to fully appreciate the multi-dimensionality of discrimination. This begins with the question *which* protected attributes to use in a given context: Should these be limited to the grounds covered by anti-discrimination law or address further disadvantages? This choice is even more challenging when multiple jurisdictions are involved [10]. More work is needed on the expansion of fairness frameworks to more protected grounds. Subsequently, protected attributes need to be defined and data labelled accordingly. This process needs to be informed by other disciplines, especially social sciences and law, where cases are available.

Fairness trade-offs between (sub-)groups: With an increasing number of protected attributes, including subgroups, it becomes more likely that increasing fairness for one attribute limits or decreases fairness for others. Beyond a balancing of rights [28], the law does not provide clear guidance for such scenarios. A factual "hierarchy" of discrimination grounds exists w.r.t. scope and strictness of protection [37, 38], but it seems unlikely that it was intended by the lawmakers to be applied to direct trade-offs (cf. the explanatory memorandum to Directive 2000/78/EC[14]). Doing so could incite a "battle of oppressions" [37] on who is more deserving of protection. While this is a known issue, it may become more pressing when directly laid open by the seeming precision of statistical fairness measures. Future work should address these questions.

Data scarcity for intersectional subgroups: On a practical level, methods to assess fairness under the condition of data scarcity (see, e.g. [82] for tentative ideas) are important to detect and address cumulative and, especially, intersectional forms of discrimination as the collection of more data on protected attributes, including subgroups, meets various challenges[3]. In the legal domain, much comes down to the question of enabling the collection and use of strictly regulated sensitive data (cf. art. 9 GDPR) for fair ML while ensuring data privacy [81]. Art. 10 (5) of the planned EU AI Act [17], allowing the use of sensitive personal data for bias monitoring, detection and correction may be a step towards a new balance. Data scarcity could be also addressed in other ways, for example by generating synthetic data of desired characteristics [18]. However, describing the characteristics of the targeted subgroups is not an easy task and might introduce subgroup biases and prejudices and lead to both allocative and representational harms. Even if data is available, statistical tests applied in ECJ jurisprudence are not always suitable for identifying discrimination of small minority groups [86], which is a particular problem for intersectional discrimination. Thus, work on suitable statistical tests for discrimination under the law is needed.

Sequential scenarios: ML models are the result of complex pipelines with several components and decisions affecting the resulting models. As bias and discrimination w.r.t. a single or more protected attributes can arise at any stage of the pipeline, it is important to take into account the discriminatory effects of these components in the overall pipeline and address them holistically rather than in isolation in order to improve the overall utility of the model.

ACKNOWLEDGMENTS

This research work is funded by Volkswagen Foundation under the call "Artificial Intelligence and the Society of the Future" project "Bias and Discrimination in Big Data and Algorithmic Processing - BIAS" and by the European Union under the Horizon Europe MAMMOth project, Grant Agreement ID: 101070285. We especially thank our colleagues in the projects for invaluable discussions and helpful suggestions.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*. PMLR, 60–69.

[2] Kiana Alikhademi, Emma Drobina, Diandra Prioleau, Brianna Richardson, Duncan Purves, and Juan E Gilbert. 2021. A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law* (2021), 1–17. https://doi.org/10.1007/s10506-021-09286-4

[3] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 249–260. https://doi.org/10.1145/3442188.3445888

[4] Shreya Atrey. 2017. Redefining frontiers of EU discrimination law. *Public Law* (2017), 185–195.

[5] Susanne Baer, Melanie Bittner, and Anna Lena Göttsche. 2010. Mehrdimensionale Diskriminierung – Begriffe, Theorien und juristische Analyse. Teilexpertise, erstellt im Auftrag der Antidiskriminierungsstelle des Bundes. https://www.antidiskriminierungsstelle.de/SharedDocs/downloads/DE/publikationen/Expertisen/expertise_mehrdimensionale_diskriminierung_jur_analyse.pdf?__blob=publicationFile&v=3

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

[7] Amanda Bower, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alexander Vargas, and Suresh Venkatasubramanian. 2017. Fair Pipelines. https://doi.org/10.48550/ARXIV.1707.00391

[8] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine*

[17] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative Acts, COM(2021) 206 final, 21st April 2021.

Learning Research, Vol. 81), Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[9] Eli M Cahan, Tina Hernandez-Boussard, Sonoo Thadaney-Israni, and Daniel L Rubin. 2019. Putting the data before the algorithm in big data addressing personalized healthcare. *NPJ Digital Medicine* 2, 1 (2019), 1–6. https://doi.org/10.1038/s41746-019-0157-2

[10] Laura Carmichael, Sophie Stalla-Bourdillon, and Steffen Staab. 2016. Data Mining and Automated Discrimination: A Mixed Legal/Technical Perspective. *IEEE Intelligent Systems* 31, 6 (2016), 51–55. https://doi.org/10.1109/MIS.2016.96

[11] Victoria Chege. 2012. The European Union anti-discrimination directives and European Union equality law: the case of multi-dimensional discrimination. *ERA Forum* 13, 2 (01 Aug 2012), 275–293. https://doi.org/10.1007/s12027-012-0260-1

[12] Isabelle Chopin and Catharina Germaine. 2021. *A comparative analysis of non-discrimination law in Europe 2020: the 27 EU Member States, Albania, Iceland, Liechtenstein, Montenegro, North Macedonia, Norway, Serbia, Turkey and the United Kingdom compared.* Technical Report. European Commission Directorate-General for Justice and Consumers. https://doi.org/doi/10.2838/72272

[13] Sujit Choudhry. 2000. Distribution vs. Recognition: The Case of Anti-Discrimination Laws. *George Mason Law Review* 9 (2000), 145–178.

[14] European Commission. 1999. Proposal for a Council Directive establishing a general framework for equal treatment in employment and occupation.

[15] Kimberle Crenshaw. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum* (1989), 139–167. https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8

[16] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 525–534.

[17] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems* 34 (2021), 6478–6490.

[18] Luiz Henrique dos Santos Fernandes, Kate Smith-Miles, and Ana Carolina Lorena. 2022. Generating Diverse Clustering Datasets with Targeted Characteristics. In *Intelligent Systems*, João Carlos Xavier-Junior and Ricardo Araújo Rios (Eds.). Springer International Publishing, Cham, 398–412.

[19] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580. https://doi.org/10.1126/sciadv.aao5580

[20] 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.

[21] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. 2022. A survey on bias in visual datasets. *Computer Vision and Image Understanding* 223 (2022), 103552.

[22] European Union Agency for Fundamental Rights. 2010. *Data in Focus Report: Multiple Discrimination.* Technical Report. https://fra.europa.eu/sites/default/files/fra_uploads/1454-EU_MIDIS_DiF5-multiple-discrimination_EN.pdf

[23] European Union Agency for Fundamental Rights. 2022. *Equality in the EU 20 Years on from the initial Implementation of the Equality Directives.* Technical Report. European Union Agency for Fundamental Rights. https://doi.org/10.2811/275515

[24] European Commission Directorate-General for Justice and Consumers. 2015. *Special Eurobarometer 437: Discrimination in the EU in 2015.* Technical Report. https://doi.org/doi/10.2838/499763

[25] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. Bayesian Modeling of Intersectional Fairness: The Variance of Bias. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 424–432.

[26] 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.

[27] Sandra Fredman. 2016. Intersectional discrimination in EU gender equality and non-discrimination law. *Brussels: European Commission* (2016).

[28] Caroline Gentgen-Barg. 2022. Fairnessmetriken bei algorithmischen Entscheidungen aus juristischer Perspektive. In *Daten, Plattformen und KI als Dreiklang unserer Zeit - Tagungsband der Herbstakademie 2022 der Deutschen Stiftung für Recht und Informatik*, Christian Heinze (Ed.). 543–560.

[29] Avijit Ghosh, Lea Genuit, and Mary Reagan. 2021. Characterizing intersectional group fairness with worst-case comparisons. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion.* PMLR, 22–34.

[30] Philipp Hacker. 2018. Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review* 55 (2018), 1143–1185. https://doi.org/10.54648/cola2018095

[31] 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* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 501–512. https://doi.org/10.1145/3351095.3372826

[32] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NIPS*.

[33] Felix Hartmann. 2019. Diskriminierung durch Antidiskriminierungsrecht? Möglichkeiten und Grenzen eines postkategorialen Diskriminierungsschutzes in der Europäischen Union. *Europäische Zeitschrift für Arbeitsrecht* 12 (2019), 24–44.

[34] Marc P. Hauer, Johannes Kevekordes, and Maryam Amir Haeri. 2021. Legal perspective on possible fairness measures – A legal discussion using the example of hiring decisions. *Computer Law & Security Review* 42 (2021), 105583. https://doi.org/10.1016/j.clsr.2021.105583

[35] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 1939–1948. https://proceedings.mlr.press/v80/hebert-johnson18a.html

[36] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22, 7 (2019), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

[37] Elisabeth Holzleithner. 2005. Mainstreaming Equality: Dis/Entangling Grounds of Discrimination. *Transnational Law & Contemporary Problems* 14 (2005), 927–957.

[38] Erica Howard. 2006. The case for a considered Hierarchy of Discrimination Grounds in EU Law. *Maastricht Journal of European and Comparative Law* 13 (2006), 445–470. https://doi.org/10.1177/1023263X0601300404

[39] Erica Howard. 2018. EU anti-discrimination law: Has the CJEU stopped moving forward? *International Journal of Discrimination and the Law* 18 (2018), 60–81.

[40] Erica Howard. 2021. German headscarf cases at the ECJ: a glimmer of hope? Blog entry, European Law Blog. https://europeanlawblog.eu/2021/07/26/german-headscarf-cases-at-the-ecj-a-glimmer-of-hope/ accessed 6 February 2023.

[41] Qian Hu and Huzefa Rangwala. 2020. Towards Fair Educational Data Mining: A Case Study on Detecting At-Risk Students. *International Educational Data Mining Society* (2020).

[42] Darren Lenard Hutchinson. 1997. Out Yet Unseen: A Racial Critique of Gay and Lesbian Legal Theory and Political Discourse. *Connecticut Law Review* (1997), 561–645.

[43] Darren Lenard Hutchinson. 2001. Identity crisis: Intersectionality, multidimensionality, and the development of an adequate theory of subordination. *Michigan Journal of Race & Law* (2001), 285–317.

[44] Vasileios Iosifidis and Eirini Ntoutsi. 2019. Adafair: Cumulative fairness adaptive boosting. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 781–790.

[45] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in reinforcement learning. In *International conference on machine learning*. PMLR, 1617–1626.

[46] Joni R Jackson. 2018. Algorithmic bias. *Journal of Leadership, Accountability and Ethics* 15, 4 (2018), 55–65.

[47] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *2010 IEEE international conference on data mining*. IEEE, 869–874.

[48] Jian Kang, Tiankai Xie, Xintao Wu, Ross Maciejewski, and Hanghang Tong. 2021. MultiFair: Multi-Group Fairness in Machine Learning. *arXiv preprint arXiv:2105.11069* (2021).

[49] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*. PMLR, 2564–2572.

[50] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 100–109.

[51] Daniel Kifer and Ashwin Machanavajjhala. 2014. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)* 39, 1 (2014), 1–36.

[52] Youjin Kong. 2022. Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 485–494. https://doi.org/10.1145/3531146.3533114

[53] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*. 853–862.

[54] MariaCaterina La Barbera and Marta Cruells López. 2011. Toward the Implementation of Intersectionality in the European Multilevel Legal Praxis: B. S. v. Spain. *Law & Society Review* (2011), 1167–1201.

[55] Ulrike Lembke and Doris Liebscher. 2014. Postkategoriales Antidiskriminierungsrecht? - Oder: Wie kommen Konzepte der Intersektionalität in die Rechtsdogmatik? In *Intersektionelle Benachteiligung und Diskriminierung*, I. Philipp, S.and Meier, V. Apostolovski, K. Starl, and K. Schmidlechner (Eds.). 261–289.

[56] Doris Liebscher, Tarek Naguib, Tino Plümecke, and Juana Remus. 2012. Wege aus der Essentialismusfalle: Überlegungen zu einem postkategorialen Antidiskriminierungsrecht. *Kritische Justiz* 45 (2012), 204–218. Issue 2. https://doi.org/doi.org/10.5771/0023-4834-2012-2-204

[57] Weiwen Liu, Feng Liu, Ruiming Tang, Ben Liao, Guangyong Chen, and Pheng Ann Heng. 2020. Balancing between accuracy and fairness for interactive recommendation with reinforcement learning. In *Pacific-asia conference on knowledge discovery and data mining*. Springer, 155–167.

[58] Jiaqi Ma, Junwei Deng, and Qiaozhu Mei. 2021. Subgroup generalization and fairness of graph neural networks. *Advances in Neural Information Processing Systems* 34 (2021), 1048–1061.

[59] Timo Makkonen. 2002. *Multiple, compound and intersectional discrimination: bringing the experiences of the most marginalized to the fore*. Aabo Akademi University. Institute for Human Rights.

[60] Natalia L Martinez, Martin A Bertran, Afroditi Papadaki, Miguel Rodrigues, and Guillermo Sapiro. 2021. Blind pareto fairness and subgroup robustness. In *International Conference on Machine Learning*. PMLR, 7492–7501.

[61] Carlos Mougan, Jose M Alvarez, Gourab K Patro, Salvatore Ruggieri, and Steffen Staab. 2022. Fairness implications of encoding protected categorical attributes. *arXiv preprint arXiv:2201.11358* (2022).

[62] Thomas B. Nachbar. 2021. Algorithmic Fairness, algorithmic Discrimination. *Florida State University Law Review* 48 (2021), 509–558.

[63] NDC-RisC. 2019. Mean height. Online resource. https://ncdrisc.org/height-mean-ranking.html accessed 6 February 2023.

[64] Rick Noack. 29 June 2018. Want to be a Police Officer in Germany? Don't be short or have narrow hips. *The Washington Post* (29 June 2018). https://www.washingtonpost.com/news/worldviews/wp/2018/06/29/want-to-be-a-police-officer-in-germany-dont-be-short-or-have-narrow-hips/ accessed 6 February 2023.

[65] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.

[66] Arjun Panesar. 2019. *Machine learning and AI for healthcare*. Springer.

[67] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 560–568.

[68] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *ACM Comput. Surv.* 55, 3, Article 51 (feb 2022), 44 pages. https://doi.org/10.1145/3494672

[69] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2021. Fairness in rankings and recommendations: an overview. *The VLDB Journal* (2021), 1–28.

[70] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *arXiv preprint arXiv:1709.02012* (2017).

[71] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 469–481.

[72] Arjun Roy, Vasileios Iosifidis, and Eirini Ntoutsi. 2022. Multi-fairness under class-imbalance. In *International Conference on Discovery Science*. Springer.

[73] Ronald B Rubin. 1978. The Uniform Guidelines on Employee Selection Procedures: Compromises and Controversies. *Cath. UL Rev.* 28 (1978), 605.

[74] Dagmar Schiek. 2005. Broadening the Scope and the Norms of EU Gender Equality Law: Towards a multidimensional Conception of Equality Law. *Maastricht Journal of European and Comparative Law* 12 (2005), 427–466. https://doi.org/10.1177/1023263X0501200405

[75] Dagmar Schiek. 2009. From European Union non-discrimination law towards multidimensional equality law for Europe. In *European Union Non-Discrimination Law: Comparative Perspectives on multidimensional Equality Law*, Dagmar Schiek and Viktoria Chege (Eds.). 3–27.

[76] Changjian Shui, Gezheng Xu, CHEN Qi, Jiaqi Li, Charles Ling, Tal Arbel, Boyu Wang, and Christian Gagné. 2022. On Learning Fairness and Accuracy on Multiple Subgroups. In *Advances in Neural Information Processing Systems*.

[77] Iyiola Solanke. 2009. Putting Race and Gender Together: A New Approach To Intersectionality. *The Modern Law Review* 72, 5 (2009), 723–749. https://doi.org/10.1111/j.1468-2230.2009.00765.x

[78] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery: Google ads, black names and white names, racial discrimination, and click advertising. *ACM ueue* 11 (2013), 1–19.

[79] Paola Uccellari. 2008. Multiple Discrimination: How Law can Reflect Reality. *The Equal Rights Review* 1 (2008), 24–49.

[80] Lachlan Urquhart and Diana Miranda. 2022. Policing faces: the present and future of intelligent facial surveillance. *Information & Communications Technology Law* 31, 2 (2022), 194–219.

[81] Marvin van Bekkum and Frederik Zuiderveen Borgesius. 2023. Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? *Computer Law & Security Review* 48 (2023), 105770. https://doi.org/10.1016/j.clsr.2022.105770

[82] Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4, 2 (2017), 2053951717743530. https://doi.org/10.1177/2053951717743530 arXiv:https://doi.org/10.1177/2053951717743530

[83] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 1–7.

[84] Trevor Vidano, Francis Assadian, and Nihal Gulati. 2023. Artificially Intelligent Active Safety Systems. In *AI-enabled Technologies for Autonomous and Connected Vehicles*. Springer, 213–254.

[85] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. *West Virginia Law Review* 123 (2021), 735–790.

[86] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review* 41 (2021), 105567. https://doi.org/10.1016/j.clsr.2021.105567

[87] Nils Weinberg. 2020. Ansätze zur Dogmatik der intersektionalen Benachteiligung. *Europäische Zeitschrift für Arbeitsrecht* 13 (2020), 60–77.

[88] Raphaële Xenidis. 2018. Multiple Discrimination in EU Anti-Discrimination Law: Towards redressing complex Inequality? In *EU Anti-Discrimination Law beyond Gender*, U. Belavusau and K. Henrard (Eds.). 41–74.

[89] Raphaële Xenidis. 2020. Tuning EU equality law to algorithmic discrimination: Three pathways to resilience. *Maastricht Journal of European and Comparative Law* 27 (2020), 736–758.

[90] Raphaële Xenidis and Linda Senden. 2020. EU non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination. In *General Principles of EU law and the EU Digital Order*, Ulf Bernitz, Xavier Groussot, Jaan Paju, and Sybe A. de Vries (Eds.). 151–182.

[91] Forest Yang, Mouhamadou Cisse, and Oluwasanmi O Koyejo. 2020. Fairness with overlapping groups; a probabilistic perspective. *Advances in Neural Information Processing Systems* 33 (2020).

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

[93] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. 2019. Fairness Constraints: A Flexible Approach for Fair Classification. *J. Mach. Learn. Res.* 20, 75 (2019), 1–42.

[94] Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. 2020. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems* 33 (2020), 18457–18469.