The Conflict Between Algorithmic Fairness and Non-Discrimination: An Analysis of Fair Automated Hiring

Robert Lee Poe
LIDER Laboratory, Sant'Anna School of Advanced Studies
Pisa, Italy
robertlee.poe@santannapisa.it

Soumia Zohra El Mestari
Interdisciplinary Center for Security, Reliability and Trust,
University of Luxembourg
Esch-sur-Alzette, Luxembourg
soumia.elmestari@uni.lu

ABSTRACT

AI-based automated hiring systems cover a wide range of tools of varying complexity, from resume parsing tools to candidate selection models. Their close interference in economic and social life faces raising demands and investigations aiming to reduce the potential discrimination they may cause. This article covers the intersection of EU non-discrimination law and algorithmic fairness in the context of automated hiring systems. The paper analyzes the balance between equality of opportunity (formal and substantive) and equality of outcome, critiques the focus on non-conservative group fairness in machine learning, and discusses the legal implications of automated hiring systems under EU law. Additionally, it highlights often committed fallacies in relation to the process of de-biasing and advocates for a broader understanding of fairness in machine learning that aligns with EU legal standards and societal values.

CCS CONCEPTS

• **Social and professional topics → Computing / technology policy**.

KEYWORDS

Algorithmic Discrimination, Algorithmic Fairness, Fair Automated Hiring, Generalizability, Equity, Merit, Automated Decisions, Fair Machine Learning

ACM Reference Format:
Robert Lee Poe and Soumia Zohra El Mestari. 2024. The Conflict Between Algorithmic Fairness and Non-Discrimination: An Analysis of Fair Automated Hiring. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3630106.3659015

1 INTRODUCTION

Automated decision-making systems are increasingly being used to render high-impact decisions regarding human beings. Due to the many concerns about the potential societal impacts of machine learning, governments are beginning to put forward policy positions and draft regulations. In the AI Bill of Rights, the White House states that automated decisions should be designed and deployed in an equitable way [63]. In Europe, the AI Act states that automated decisions should not perpetuate historic patterns of discrimination or create new forms of discriminatory impacts [29]. To a limited extent, researchers in the field have understood that they had not happened upon an empty field (of research) but instead a garden that has been fostered, cared for, and in some cases ignored for a very long time. The garden is that of *Justice*. While perspectives from many domains on the concept of distributive social justice have been incorporated into the algorithmic fairness literature: egalitarian philosophies of distribution [7, 10, 42], socio-technical critiques of technological solutionism [18, 57], and concepts from feminist communications and data science like the myth of objectivity and meritocracy [26, 35]; the contentions between distributive justice and non-distributive justice, between comparativist and non-comparativist conceptions of discrimination, between egalitarianism and individualism, and between equality of outcome and equality of treatment have largely gone underdeveloped.

In an effort to increase viewpoint diversity and protect the fundamental rights of individuals in the EU, this article contributions are twofold. First, the article applies European Union (EU) non-discrimination law to a highly exemplative use case: algorithmic fairness in automated hiring systems. To achieve this goal, we: describe the policies of equality of opportunity (both formal and substantive) and equality of outcome, making the legally relevant distinctions therein (2); put forward and analyze the related case-law of the Court of Justice of the European Union (CJEU), extracting the legal rules (3); and examine the process of de-biasing in light of the rules surrounding the use-case, finding that real-world, unlawful discrimination is likely taking place (6).

The second contribution is a broader effort carried out concurrently with the first, to dispel common misunderstandings that lead to harmful effects (of which the unlawful discrimination of fair automated hiring is an example), we: attempt to surmise the predominant approach to algorithmic fairness through an examination of individual, causal, and group fairness—finding that group fairness is the only substantive approach to fairness that has been put forward (4); explain the trade-off between generalizable outcomes and group similar outcomes, arguing that the realization of that trade-off means that de-biasing in accordance with an independence-based fairness metric is akin to the use of quotas (5); and identify oft committed fallacies that result as a failure to realize the trade-off (5).

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659015

2 BACKGROUND

In employment decisions, equality is sought in opportunity or outcome. Equality of opportunity can be defined as formal or substantive. Formal equality of opportunity requires that applicants be assessed according to their qualifications, that those qualifications be appropriate,[1] and that the most qualified applicant receives the position [8]. Selection processes that enforce formal equality of opportunity result in inequalities in outcome between groups when the individuals of a given group are, on a whole, less qualified than another in a given field.[2] While substantive equality of opportunity requires all that formal equality of opportunity insists upon during a selection process, it is first and foremost an effort to ensure that each individual in society, regardless of their group membership, has the same opportunities to gain the prerequisite qualifications for positions so that differences between groups are minimal or nonexistent [1, 2, 6]. Equality of outcome, sometimes referred to as *equity* in social justice literature, requires group equality or similarity in results, irrespective of the differences individuals of those groups may have in terms of qualifications for a given position—generally for the purpose of providing a shortcut from opportunity to representation when substantive equality has not yet been fully realized [19].

In the European Union, *positive action* is an umbrella term used to describe soft measures like the voluntary pruning of facially neutral employment criteria that may lead to disparate impacts, mainstreaming initiatives, accommodations, the use of impact assessments, and outreach programs [17] for achieving substantive equality of opportunity for members of groups that deal with the consequences of past or present discrimination or disadvantage so that they may compete on an equal footing with others; whereas *positive discrimination* is a term used to describe strong measures for achieving equality of outcome through preferential treatment when, in a given field of employment, members of the discriminated against or disadvantaged group are not yet, on the whole, equally qualified.[3] Positive discrimination in employment decisions is controversial, and the practice has been repeatedly restrained by the Court of Justice of the European Union (CJEU) whenever employment selection processes move from the goal of ensuring formal and substantive equality of opportunity into the pursuit of equal outcomes. The CJEU case-law pertaining to positive discrimination in employment has been settled for nearly two decades, and legal scholars have repeatedly concluded that the CJEU "systematically rejects" selection processes that turn towards equality of outcome [11, 51, 59].

Automated hiring systems based on machine learning are becoming increasingly commonplace, concerns about algorithmic indirect discrimination in employment decisions are front-and-center, and the technical solutions provided by the research community often systematically deviate from the principle of equal treatment to combat disparate impacts.[4] Legal scholarship on algorithmic discrimination has predominately focused on analyzing the training data of automated systems for features that, if used, may constitute direct or indirect discrimination and the corresponding decisions of those systems for disparate or adverse outcomes [61, 67]. Indirect or "covert" discrimination is understood in contrast with direct or "overt" discrimination [66] and is fundamentally aimed at achieving substantive equality [25].[5] Indirect discrimination takes place when a neutral provision, criterion, or practice results in a disparate impact on a protected group, "unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary" [4]. Thus, using a hiring criteria that causes a disparate impact is not automatically discriminatory. Instead, such criterion are only discriminatory if the principle of proportionality is violated. The proportionality test, therefore, "opens the path for the legality of using a factor that correlates with economically or otherwise favorable traits even though the choice of that factor also leads to the unfavorable treatment of a protected group" [37].

Whether a legal analysis determining the lawfulness of using features that result in a disparate impact based on a protected attribute is performed in practice by designers of automated hiring systems is difficult to know and beyond the scope of this article. It is clear, however, that designers of these systems are aware that criteria which cause a disparate impact based on a protected attribute can potentially be deemed discriminatory [9, 31]. Their solution: require equality or similarity in employment outcomes [54, 65]. For example, authors in [54] investigated automated hiring systems and found that a number of the commercially available systems for preselection either remove or curate the training data that produce a disparate or adverse impact or modify the objective function of the learning algorithm to achieve the same result, often in accordance with the *disparate impact metric* [54].

3 DISTINGUISHING POSITIVE ACTION FROM PREFERENTIAL TREATMENT IN HIRING

There are a number of equality directives in EU law [2–5]. Each directive is an embodiment of the principle of equal treatment. Equal treatment means that there shall be no discrimination *whatsoever*, either directly or indirectly, based on the protected attribute laid out in a given directive. However, each equality directive moves from formal to substantive equality of opportunity by allowing Member States to adopt *special measures* to prevent or compensate for disadvantages linked to the protected attribute. Thus, while the exception to the individual right of equal treatment must be interpreted strictly, measures which take advantage of the derogation, while discriminatory in appearance"are in fact intended to eliminate or reduce actual instances of inequality which may exist in the reality of social life" [19–23]. For instance, in *Badeck*, the Court drew a distinction between training opportunities and employment

[1] Where appropriateness is defined in relation to moral relevance [7] or to the lawfulness of desiderata in accordance with indirect discrimination doctrine [24].

[2] *See* e.g. Plato's *Laws* discussing the notion of equality, "[W]hen equality is given to unequal things, the resultant will be unequal . . ." [53]; *see* also Hayek on Equality, Value and Merit, "From the fact that people are very different it follows that, if we treat them equally, the result must be inequality in their actual position . . ." [40].

[3] *See* [59] for a detailed explanation of the difference between positive action and positive discrimination in EU non-discrimination law; *see also* [11]. *But see* e.g. [37, 60] for the conflation of the two distinct concepts in the algorithmic fairness literature.

[4] Such practices are usually referred to as 'fairness-aware machine learning' [28].

[5] While there exists overlap between the legal analysis of [25]—interpreting a shift from formal to substantive equality in the "soul" of EU non-discrimination law based on the development of indirect discrimination doctrine—and the policy of substantive equality of opportunity cited above, the make-up and scope of substantive equality of opportunity policies, as reflected in CJEU case-law on preferential treatment, remains distinct.

opportunities [22]. The Court applied a substantive view of equality of opportunity, allowing for the reservation of training slots for the underrepresented sex, because training opportunities are precisely where the underrepresented sex can receive the qualifications required for employment positions. The *Badeck* Court also allowed, under the derogation, a national rule that guarantees that qualified women, who satisfy all the conditions required for the position, are called to interview in sectors in which they are under-represented, because such provisions do not "imply an attempt to achieve a final result." In other words, such provisions do not sacrifice equal treatment for the sake of equal outcomes.

Before moving forward, two caveats should be noted. First, the CJEU has only analyzed gender-based provisions in the employment context through the exception under Art. 2(4) of Directive 76/207 (now Directive 2006/54/EC) and Art. 141.4 of the TEC (now Article 157.4 of the TFEU). To what extent the exceptions of other equality directives will be treated similarly is a matter of academic debate [51]. Second, while the CJEU has held that direct horizontal effect can be found in the relationship between equality directives and the Charter of Fundamental Rights, that legal issue will not be discussed here [24]. As the reader moves through the following case-law, bear in mind the distinction between soft positive action measures implemented to provide substantive equality of opportunity, like the ones described above, with strong positive discrimination measures implemented to provide equality of outcome, such as the ones under review in the following cases.

In *Kalanke*, two candidates were shortlisted for the position of Section Manager at the Bremen Parks Department in Germany. Mr. Kalanke, one of the two candidates, had a diploma in horticulture and landscape gardening, had worked for the Parks Department since 1973, and had been acting as the permanent assistant to the previous Section Manager before the position was vacated; Ms. Glibmann, the other candidate, also had a diploma in landscape gardening, granted in 1983, and had worked in the Parks Department as a horticultural employee since 1975. The Parks Department management put forward Mr. Kalanke for the position, but the Staff Committee refused its consent to his promotion. The Staff Committee refused its consent for the promotion of Mr. Kalanke in accordance with the Bremen Law on Equal treatment for Men and Women in the Public Services (LGG) passed in 1990, which stated that women who have the same qualifications as men, applying for the same post, are to be given priority where women are underrepresented in the sector. Mr. Kalanke was successful in arbitration, but the Staff Committee appealed to the conciliation board where the two candidates were found to be equally qualified and priority was to be given to Ms. Glibmann. The case made its way through the labor courts, and eventually the Bundesarbeitsgericht sought a preliminary ruling from the CJEU clarifying the scope of the exception under Article 2 (4) of the Directive from the principle of equal treatment.

The Court began by stating that the purpose of Directive 76/207 is to put into effect the principle of equal treatment for men and women regarding access to employment and promotion within Member States, and that the principle of equal treatment means that there shall be no discrimination whatsoever, either directly or indirectly, on the grounds of sex. The exception under Article 2 (4) of the Directive 76/207 permits national measures which, although discriminatory in appearance, are intended to eliminate or reduce actual instances of inequality and consequently give a specific advantage to women with a view to improving their ability to compete on the labor market and to pursue a career on an equal footing with men. Since Article 2(4) is a derogation from an individual right, the Court determined that the exception must be strictly interpreted. The Court found that national rules that guarantee women *automatic* and/or *absolute* and *unconditional* priority go beyond promoting equal opportunities and overstep the limits of the exception. The Court reasoned that such measures take a shortcut from ensuring substantive equality in fact to mere equality in outcome:

> "Furthermore, in so far as it seeks to achieve equal representation of men and women in all grades and levels within a department, such a system substitutes for equality of opportunity as envisaged in Article 2(4) the result which is only to be arrived at by providing such equality of opportunity."

Thus, the Court ruled that Art. 2(1) and (4) of the Directive 76/207 precludes national rules which *automatically* and/or *absolutely* and *unconditionally* give priority to women in sectors where they are underrepresented.

Turning next to Marschall v. Land Nordrhein-Westfalen [20]: in 1994, a teacher named Mr. Marschall applied for a promotion to an open position at a German comprehensive school. In response, Mr. Marschall was informed that, in accordance with the civil service law of the Land, a female candidate of equal suitability, competence and professional performance was to be appointed to the position because there were fewer women than men in that particular grade post in the career bracket. Mr. Marschall brought legal action. The Administrative Court of Gelsenkirchen found that the outcome of the case was dependent on the compatibility of the Land's provision with Art. 2(1) and (4) of Directive 76/207 and so a preliminary ruling was sought from the CJEU.

The Court began by distinguishing the case from *Kalanke*. Unlike in *Kalanke*, the provision in question contained a 'savings clause' that stated that where an individual male candidate had qualifications that might tilt the balance in his favor, a female candidate would not be given priority. After citing the third recital in the preamble to Recommendation 84/635/EEC on the promotion of positive action for women [1], which highlights the need for positive action to counteract prejudices that arise in the employment context due to social attitudes, behaviors, and structures, the Court agreed with the Land and other governments that, even when candidates of the opposite sex are equally qualified, male candidates tend to be promoted in preference to female candidates because of a multitude of stereotypes. Thus, ". . . the mere fact that a male candidate and a female candidate are equally qualified does not mean that they have the same chances." The Court reasoned that a national rule may be lawful under Article 2 (4) if, in each individual case:

> "it provides for male candidates who are equally as qualified as the female candidates a guarantee that the candidatures will be the subject of an *objective assessment* which will take account of all criteria specific to the individual candidates and will override the

> priority accorded to female candidates where one or more of those criteria tilts the balance in favour of the male candidate. In this respect, however, it should be remembered that those criteria must not be such as to discriminate against female candidates" [emphasis added].

Thus, the Court ruled that a national rule which, conditional on a guarantee that the candidatures will be subject to an objective assessment on an individual basis and where that objective assessment tilts in the favor of a male candidate the priority will be overridden, provides a priority to female candidates who are equally qualified, with the purpose of counteracting prejudiced tie-breaking, is compatible with Art. 2(1) and (4) of Directive 76/207.

The final case to discuss is Abrahamsson and Anderson v. Fogelqvist [21]. In 1996, eight candidates applied for a professorship at the University of Göteborg, including Ms. Abrahamsson, Ms. Destouni, Ms. Fogelqvist, and Mr. Anderson. The selection board voted twice: (1) in relation to the scientific qualifications of all candidates, Mr. Anderson received five votes and Ms. Destouni received three votes; (2) taking into account both scientific merits and a positive action provision, Ms. Destouni received six votes and Mr. Anderson two votes. The selection board proposed that Ms. Destouni be appointed, placing Mr. Anderson in second and Ms. Fogelqvist in third. Later, Ms. Destouni withdrew her application, and the Rector of the University appointed Ms. Fogelqvist to the position. The Rector stated that the difference between Mr. Anderson and Ms. Fogelqvist was not so great as to violate the requirement of objectivity in the selection process. Mr. Anderson and Ms. Abrahamsson brought legal action that eventually came before the Överklagandenämnden för Högskolan, and a preliminary ruling was requested from the CJEU.

The Court held that national rules which give a priority to candidates of an underrepresented sex who possess sufficient qualifications for a given post over a candidate of the opposite sex who would have been appointed otherwise on the basis of merit, are precluded under Article 2(1) and (4) of Directive 72/207 and Article 141(4) EC even if the difference between the candidates' qualifications are not so great as to breach the requirement of objectivity. The Court also ruled that national legislation which limits the scope of positive discrimination to a predetermined number of posts, or to posts specifically designed for that purpose, is still precluded because of the absolute and disproportionate nature of the positive discrimination practice.

Recap: in the employment context, the CJEU has stated that special measures are derogations from the principle of equal treatment and thus need be proportional. Absolute and unconditional preferences are always automatic, but not all automatic preferences are absolute and unconditional. Absolute and unconditional preferences are disproportional because such preferences make the protected attribute the key criterion when comparing candidates between sub-groups of the attribute. Automatic preferences, on the other hand, have the potential to be proportional. For instance, in the case of a tie-breaking scenario between two equally qualified candidates for the purpose of combating stereotypes. For an automatic preference to be proportional the candidates must be subject to an objective assessment, ensuring that where candidates are not equally qualified, the preference will be overridden.

4 THE TOOLS

Fairness metrics are definitions of equality formulated mathematically, and they are commonly split into three categories: group fairness, causal fairness, and individual fairness. In this section we will make use of the following notation: Y is the target variable, $\hat{Y}$ is the predicted variable, X are the features, ϵ is the threshold and S denotes the protected variables.

4.1 Individual Fairness

The mantra of individual fairness is that similar individuals should be treated similarly. The maxim of similar treatment that individual fairness embodies is an Aristotelian principle of consistency [13]. The individual fairness definition states that there should be consistency between the relevant features of two different persons and their respective outcomes in comparison to one another. More specifically, the similarity between the features of two individuals (measured as a distance) should be preserved between their respective labels.[6]

Note that the principle of consistency, defined as distance between spaces, *could* be used to detect whether there exists inconsistencies between the relevant features and ground truth of the sample, as well as when there exists inconsistencies between the sample and the outcomes. For instance, the inconsistencies could be seen as an indicator of unreliable data collection processes where data was incorrectly reported, or that the data sample is missing a set of uncollected features that could explain the current inconsistency. We emphasize, however, that the individual fairness metric itself is not concerned with determining the representativeness of the sample nor with determining how well the outcomes generalize to a target population.

Individual fairness defines *fairness* as a comparison of geometric distance between the features of two data points and the distance between the predictions assigned to these two data points. Once distance is defined, individuals can be compared, and inconsistencies (unfairness) can be rectified. However, the distance must be defined, and defining a distance presupposes prior knowledge about "fairness." In other words, the principle of consistency is *empty* [56], and so requires a substantive notion of fairness to define what makes similar cases similar (i.e. the distance). Thus, there is a circularity in the proposition that individual fairness is a definition of fairness. It may be that the principle of consistency is a necessary requirement for fairness to be achieved, but consistency or similarity alone is not sufficient to constitute an independent notion of fairness [32].

Philosophers and jurists might best understand this point by analyzing Aristotle's principle: similar individuals should be treated similarly and dissimilar individuals should be treated dissimilarly. What does it mean for individuals to be similar? There are three possible interpretations, the first two of which attempt to draw an ought from an is: (1) it might mean individuals similar in *every* respect should be treated similarly, or (2) it might mean individuals similar in *some* respect should be treated similarly [62]. Regarding

[6] Labels or target variables refer to the variable to be predicted by the machine learning algorithm.

the first, individuals cannot be identical and yet still be distinct—it is a contradiction in terms. The second interpretation leads to the absurdity that all individuals should be treated similarly because every individual is similar in some respect. The third interpretation is that of individual fairness which derives an ought from an ought: (3) individuals that are similar in some morally significant respect should be treated similarly. Hence, the principle becomes a simple tautology. "People who by a rule should be treated alike should by the rule be treated alike" [62].

Some advocates of the individual fairness approach argue that substantive notions of fairness need be defined by domain experts [28, 34], while others argue that the group fairness metrics should fill the void [32]. In any case, individual fairness should be understood as a tool to implement fairness once defined, rather than as a conception of fairness in and of itself. Having set aside individual fairness as a definition of fairness, causal fairness may be examined.

4.2 Causal Fairness

Causality based metrics define the effect of protected attributes on the decision, and thus these definitions do not rely only on the observational data but require a study of causal relationships that reflect the social and economical aspects of the data collection process. Causal fairness shares the same conception of fairness as group fairness and is only different in the sense that a different set of techniques is used to achieve this goal [15, 44, 47]. For example, the observational statistical parity measure (see equation 1), which is a group fairness definition, requires equality between the probabilities of inclusion in the positive predicted class for each protected group; while its version of causal parity(see equation 2) changes slightly this definition by introducing the notion of intervention by modifying the value of the protected attribute to a specific value and observe its effect.

$$P(\hat{y} = 1|s = 0) = P(\hat{y} = 1|s = 1) \tag{1}$$

$$P(\hat{y} = 1|do(s = 0)) = P(\hat{y} = 1|do(s = 1)) \tag{2}$$

Thus, these causal measures of fairness still link back to the idea of group similarity in outcomes; however, they reach it by introducing the causal effects that a change in the protected attribute value may cause to the decision.

4.3 Group Fairness

To build machine learning models that produce outcomes that are group similar, the first step is to define a measure or metric that reflects a notion of acceptable group dissimilarity. There exists a "zoo" of these metrics [15] that define the acceptability of group dissimilarity differently using notions of statistical independence, sufficiency, and separation. A number of surveys and reviews on the taxonomy of metrics and interventions have been published [14, 15, 49, 52].

Separation based metrics, namely equalised odds, require independence between attributes and prediction conditioned on the target Y. In other words, separation ensures that the model has the same false-positive rate and false-negative rate across groups. Taking the case law as an example, this means that an equal proportion of suitable men and women applying for the job are predicted to be suitable employees.

Sufficiency-based metrics, namely calibration and predictive parity, ensure that given the prediction, the target is independent of the group, meaning that the the prediction $\hat{Y}$ is sufficient for Y, with the same example, a sufficiency-based de-biasing algorithm will ensure an equal proportion of men and women predicted to be suitable employees are actually suitable employees. Both sufficiency and separation use the target variable and thus make an assumption about the objectivity of the target variable, for example, whether the labelling process of the data was done in an objective way and a rigorous inter annotation agreement process was done.

Independence-based fairness metrics, namely demographic parity, statistical parity, and disparate impact, are measurements of group similarity in outcomes by ensuring statistical independence between the outcomes (the predictions) and the protected features.

$$\frac{P[\hat{Y} = 1|S \neq 1]}{P[\hat{Y} = 1|S = 1]} \geq 1 - \epsilon \tag{3}$$

The mathematical formulation of disparate impact (3) is built around an independence between the joint distribution of a protected attribute S and the classification outcome $\hat{Y}$ in the case of a binary classifier. Thus, when the event $\hat{Y} = 1$, is the positive outcome, the acceptance rates of different groups must be greater than a threshold determined by a predefined term ϵ. For disparate impact, ϵ is the 80% or the 4/5 rule.

Demographic parity is a widely used metric for independence-based group fairness, especially in fairness-aware automated hiring systems. The authors of [54] found that the most common metric used by commercially available pre-selection systems is the disparate impact metric. Our use case is automated hiring, so we elaborate more on this metric throughout this paper. Independence metrics require the same positive prediction ratio across groups identified by the protected attribute. Furthermore, independence measures rely only on the distribution of features (protected and non protected) and decisions, namely on $(S, X, \hat{Y})$, thus even in the case of perfect prediction algorithm[7] the independence metrics are not necessarily satisfied. To explain, independence is satisfied in a perfect predictor only if the target is evenly distributed across all groups, which is not always the case. Therefore, independence fairness metrics do not conserve the status quo and thus are known to be "$non - conservative$" [55].

Now, simply measuring the group similarity of the outcome alone has no effect on the decision-making process. For instance, the measurement could be used to analyze whether a given feature might create a disparate impact to determine its proportionality in accordance with indirect discrimination doctrine, or to determine whether diversity goals have been met. Even the removal of *proportional* features that led to a disparate impact could be in alignment with the goal of substantive equality of opportunity. However, measuring the group similarity of outcomes (substantive equality of opportunity) is different than constraining those outcomes to be group similar (equality of outcome). Other suggestions from the algorithmic fairness literature that are in accordance with substantive equality of opportunity could include: outreach programs designed to attract talent from underrepresented groups,

[7] A perfect predictor is a predictor were the predicted labels $\hat{Y}$ are equal to the actual targets Y ($\hat{Y} = Y$)

stakeholder involvement throughout the ML pipeline (including feature selection) [48], and a diverse group composition amongst the designers of the system [46]. None of those suggestions "imply an attempt to achieve a final result" [22].

5 FROM MEASURING FAIRNESS TO DE-BIASING

The process of ensuring a final result is known as de-biasing. Before explaining how de-biasing is performed technically, it is important to understand how bias is defined and the implications of that definition. In data-driven processes like machine learning, bias is traditionally defined as a deviation from the true value of a parameter or variable [30]. In fair machine learning, bias is defined as a deviation from group similarity in outcomes [15, 16, 49, 52]. Why is this an important distinction? The distinction between these two definitions of bias illuminates the goal of the processes. Where machine learning is a historical, descriptive and predictive process, de-biasing is an ahistorical, prescriptive process. While philosophers might best understand the thrust of the point through these remarks on the distinction between is and ought, jurists might best understand by comparing the separation thesis found in legal positivism to the differentiation made here [39]. The separation thesis insists on the separation between (1) what the law *is* and (2) what the law *ought* to be. Now, when the true value of a parameter leads to group dissimilarity in outcomes, the *true value* is dubbed biased. This is due to the fact that group dissimilarities in outcomes can either be the result of: (1) group disparities existing in the target population that are reflected in a representative sample and carried into the outcomes by generalizable hypothesis assumptions (accuracies), or (2) an unrepresentative sample and/or non-generalizable assumptions that have the potential to underestimate or exaggerate group disparities (inaccuracies). De-biasing in accordance with an independence-based fairness metric is the purposeful underestimation of group disparities. In other words, decisions made on a representative sample have the potential to reflect the target population in the model outcomes, and those outcomes would have the same disparities between groups that exist in the target population. Thus to reach group similar outcomes the sample must be made unrepresentative or the hypothesis assumptions non-generalizable.

Why is it important to understand the difference between is and ought statements? Sometimes people erroneously draw conclusions about what ought to be based solely on observations of what is, without providing a justifying logical bridge. This gap forms the basis of what is commonly referred to as the is/ought fallacy [43]. When one infers an "ought" from an "is" without justification, they commit this fallacy. It represents a logical error, premised on the implicit assumption that the state of affairs necessarily dictates how it should be. Conversely, a less discussed but equally fallacious reasoning is what might be called the "ought/is" fallacy. This involves a reverse projection, where ideals about how things should be are assumed to reflect the actual state of the world. This form of reasoning often leads to a kind of wishful thinking, mistaking one's moral vision for empirical reality. For instance, if a person holds that all individuals should be treated equally (a normative statement) and, based on this belief alone, assumes that all individuals are equal (a descriptive statement), they are engaging in this reverse fallacy. This assumption, precisely the axiomatic assumption that "we are all equal," has already been noted as a common underlying axiom of algorithmic fairness [33, 34]. Such an assumption is especially absurd when one considers that the realization that "we are not all equal" is the exact motivation behind the more egalitarian strands of algorithmic fairness. Dismissing the trade-off between generalizable outcomes and group similar outcomes based on the insistence that "we are all equal" is an example of the ought/is fallacy. There are other common mistakes that result from the failure to understand the trade-off. For instance, authors in [18] argue that the conflict between "accuracy" and "fairness" is the result of framing the trade-off as an optimization problem. Their argument rests on a causal fallacy. Recognizing the "inherent conflict" between generalizable outcomes and group similar outcomes in a data setting which contains group disparities and then optimizing between those competing interests cannot be the *cause* of differences between subgroups of a target population that exist independently in that data setting.

To understand the trade-off technically, reference must be made to the trade-off between accuracy and fairness. The lower bound of that trade-off has been estimated via proof [36, 70]. And Authors in [50] have proven that in the case of a binary classifier it will be asymptotically possible to maximize both accuracy and fairness simultaneously only if the protected attribute and the target variable are perfectly independent. On the other extreme, if the protected attribute is highly correlated with the target variable then it is only possible to maximize either the accuracy or the fairness at the same time. In between those two extremes, the trade-off is determined by the strength of the correlation between the target and the protected attribute. As the [50] proof states, if the protected attribute and the target variable are perfectly independent of one another, the more generalizable the model is, the more group parity will be present. Some authors use this fact to argue that accuracy and fairness are complimentary [18, 27, 41]; even going so far as to state that the "fairness-accuracy trade-off formulation also forecloses the very reasonable possibility that accuracy is generally in accord with fairness" [18]. While it is true that under certain conditions generalizable outcomes and group similar outcomes are complimentary, the reliance on that truth to minimize the importance of the trade-off is highly misleading. Generalizable outcomes and group similar outcomes can *only* be complimentary in a data setting where no group disparities exist (necessarily defined as group parity in the context of perfect independence). If there is no group disparity in the data setting, there is no need for de-biasing. If group disparity exists in the data setting, generalizable outcomes and group similar outcomes will be uncomplimentary (i.e. the protected attributes and target variable will be correlated.) Others observe that, in practice, constraining outcomes to be group similar can sometimes increase accuracy [64]. Again, the observation is correct but can lead to a misunderstanding. When the use of a fairness constraint increases the accuracy, either the protected feature and target variable are independent (and so see the above argument) or the data sample was so unrepresentative that enforcing group similarity increased the accuracy by happenstance. And, that increase in accuracy by happenstance could never go beyond the group similarity present in the target population without decreasing the generalizability of the model.

The trade-off between generalizable outcomes and group similar outcomes is obvious. The logical conclusion of the trade-off is also obvious: where there exists the greatest need for de-biasing (i.e. data settings that contain large group disparities), data-driven processes like machine learning are most useless. In other words, as the connection between the model outcomes and the target population becomes more tenuous to become less dissimilar amongst subgroups, the use of machine learning becomes harder to justify. The more the outcome is already known (manually coded), the less need there is for a data driven approach—a script or quota could fulfill the same purpose. Thus, the trade-off presents a threat to the field. Beyond practical implications about energy sustainability and the waste of compute, why is this an important point? The use of quotas and preferential treatment for the purpose of balancing group disparities in society is not a new phenomenon, and the normative and legal questions surrounding their use have likely already been developed in a given jurisdiction. For example, quotas are directly the subject matter of the entire first half of this article. When a technology is understood, it is much easier to identify whether the use of it in a given context is lawful.

Once a metric is chosen, one of the three following de-biasing strategies can be adopted: (1) pre-processing the input data to remove, alter, or curate the underlying data that lead to group dissimilarities [31, 38, 69], (2) in-processing where the model is constrained to produce group similar outcomes by modifying the learning algorithm's objective functions [12, 68]; and/or (3) post-processing the output of the model, rather than changing anything about the sample or hypothesis assumptions, by using an algorithm based on a function that detects potential group dissimilarities and adjusts the labels accordingly [45]. If the chosen debiasing process requires the elimination of differences between groups based on a protected attribute, while disregarding the base-rate differences between those groups, the effect would be to give systematic, preferential treatment to one group at the expense of the other. The frequency or severity of that systematic deviation from equal treatment would depend on the strength of the correlation between the protected attribute and the target variable in the original, unmodified sample. The trade-off between an automated hiring system that seeks to achieve equality of treatment versus equality of outcome is inextricably linked to the trade-off between generalizable outcomes and group similar outcomes, where the generalizability in employment decisions is an instantiation of qualification assessment objectivity (*Marschall Test*), and group similarity in outcomes (fairness) is an instantiation of preferential treatment. Placed in this context, the "cost of fairness" [50] is the sacrifice of the individual, fundamental right to equal treatment [6].

6 ANALYSIS

To begin with the question posed at the heart of algorithmic fairness: should we address group base-rate differences through a process of de-biasing, thereby creating preferential conditions for some based on their protected attributes in order to reach an equitable distribution, or should we maintain equal treatment in the competition itself, relying instead on institutions committed to substantive equality of opportunity and positive action policies to redress factual inequalities between groups in society? In the context of hiring in the EU, that normative question has already been answered. In *Kalanke*, the CJEU put forward the primary concern and determining factor of proportionality in the employment context: whether the practice substitutes substantive equality of opportunity for the outcome that is only to be reached by the realization of factual equality in society.

In *Marschall*, the preferential treatment of the underrepresented sex was limited to tie-breaking scenarios of equally qualified candidates to counteract prejudiced tie-breaking that existed in social reality in accordance with the goal of substantive equality of opportunity. The *Marschall* "savings clause" ensured that outcome equality would not pursued—requiring the employer to subject the candidatures to an objective assessment, where the preference would be overridden if the candidate of the over-represented sex had qualifications that would tilt the balance in their favor. Unlike the practice in *Marschall*, candidates subjected to fair automated hiring processes that de-bias in accordance with an independence-based metric, are not objectively assessed in the first place, let alone given the assurance of an override. Previous to the selection process and comparison of applicants, the data sample would have been modified and/or the hypothesis assumptions trained to undervalue the qualifications of some and overvalue the qualifications of others based on their group membership. In other words, the weights of the applicants' features, without preferential treatment, are not brought to bear on the hiring decision. Thus, the integrity of a tie-breaking scenario is compromised at the outset.

In *Abrahamsson*, the CJEU considered legislation which required that a candidate for a public position belonging to the under-represented sex and possessing *sufficient qualifications* for that post must be given a preference over a candidate of the opposite sex that would have been appointed otherwise in order to achieve equal gender representation in the given field of employment. The Court found that the objectivity of the selection process could, therefore, not be precisely determined. Such a practice, the Court reasoned, would result in the selection of candidates with qualifications not equal to but inferior to those of candidates of the opposite sex, ultimately substituting the individual assessment of candidate merit for group membership. The Court also ruled that even if the scope of such a practice was limited to a predetermined number of posts, or to posts specifically designed for that purpose, it would still be precluded because of the absolute and unconditional nature of the practice. Unlike in the *Abrahamsson* case, the objectivity, or the lack thereof, of the selection process of an automated hiring system *could* be determined. A system which simply rids the outcome of group skew (defined as the quotient of between group distance and within group distance) or group dissimilarity (in accordance with a an independence-based fairness metric) and, in the process of doing so, necessarily disregards the representativness of the sample and generalizability of the model, shows its lack of an objective assessment. Further, *Abrahamsson* tells us, that creating a threshold at which candidates are qualified and then ensuring equal outcomes between groups post threshold satisfaction, would likely be precluded under an interpretation of the derogation of the relevant equality directive. For the above reasons, fair automated hiring systems that de-bias in accordance with independence-based metrics would likely be deemed unlawful due to their *automatic* preferential treatment. Such systems, certainly if used for selection processes

of public posts, are simply a high-tech evasion of law which has been settled for decades.

7 CONCLUSION

It is widely recognized that automated hiring systems must not discriminate. Often fair machine learning and the tool-set it provides is seen as the answer to creating non-discriminatory automated hiring systems. However, the fact is that the most commonly used metric in fair automated hiring systems ensures a discriminatory effect when used as a basis for a de-biasing process. If the chosen debiasing process, using a non-conservative metric, requires the elimination of differences between groups based on a protected attribute, while disregarding the base-rate differences between those groups, the effect would be to give systematic, preferential treatment to one group at the expense of the other. And the frequency or severity of that systematic deviation from equal treatment would depend on the strength of the correlation between the protected attribute and the target variable in the original, unmodified sample. While algorithmic unfairness and discrimination are often used synonymously, the importance of "accuracy" and the estimation and preservation of model generalizability should not be ignored when determining the legality of such systems. Algorithmic fairness and algorithmic non-discrimination are not one in the same, and further research into the conflicts between the two in different jurisdictions and applications is required to ensure that automated decision-making systems are just.

We suspect the reason for equality of outcome being a dominant approach on the policy side of the algorithmic fairness literature is largely due to the fact that the traditional machine learning approach has the potential to satisfy a meritocratic conception but can never satisfy an equitable conception in a data setting that contains group disparities. Our concern is that by defining fairness as equality of outcomes, the community may be leading policy-makers and regulators to believe that *fairness* is absent in automated decisions without the use of the equitable approach. We hope for an expansion in how fairness is conceived so that the literature can capture the same kind of diversity in opinion that is present in the wider societal discourse. We also hope to have shown that the rejection of the trade-off between generalizable outcomes and group similar outcomes has, in some cases, resulted in fallacious reasoning and misleading assertions. Researchers should confront the reality that group similar outcomes require the introduction of inaccuracies in a data setting where group disparities are present. The research community should be more straightforward about what is being sacrificed in the name of equal outcomes. Obfuscating the nature of that sacrifice may lead to unlawful discrimination. As was once wisely said, "There are no solutions. There are only trade-offs" [58].

ACKNOWLEDGMENTS

The research presented in this paper has received funding from the European Union's funded project LeADS under Grant Agreement no. 956562. We would like to give special thanks to Gabriele Lenzini, Jean-Michel Loubes, and Maciej Zuziak for their advice and feedback throughout the writing process.

REFERENCES

[1] 1984. 84/635/EEC: Council recommendation of 13 December 1984 on the promotion of positive action for women. Official Journal of the European Communities. , 34–35 pages. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:31984H0635

[2] 2000. Directive 2000/43/EC of the European Parliament and of the Council of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin. Official Journal of the European Communities. , 22–26 pages. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32000L0043

[3] 2000. Directive 2000/78/EC of the European Parliament and of the Council of 27 November 2000 establishing a general framework for equal treatment in employment and occupation. Official Journal of the European Communities. , 16–22 pages. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32000L0078

[4] 2004. Council 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. Official Journal of the European Union. , 37–43 pages. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32004L0113

[5] 2006. Directive 2006/54/EC of the European Parliament and of the Council 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. Official Journal of the European Union. , 23–36 pages. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32006L0054

[6] 2012. Charter of Fundamental Rights of the European Union. Official Journal of the European Union. , 391–407 pages. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A12012P%2FTXT

[7] Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. 2022. Towards Substantive Conceptions of Algorithmic Fairness: Normative Guidance from Equal Opportunity Doctrines. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3551624.3555303

[8] Richard Arneson. 2015. Equality of Opportunity. In *The Stanford Encyclopedia of Philosophy* (summer 2015 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2015/entries/equal-opportunity/

[9] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 3 (2016), 671–732. https://www.jstor.org/stable/24758720 Publisher: California Law Review, Inc..

[10] Joachim Baumann, Corinna Hertweck, Michele Loi, and Christoph Heitz. 2022. Distributive Justice as the Foundational Premise of Fair ML: Unification, Extension, and Interpretation of Group Fairness Metrics. (2022). https://doi.org/10.48550/ARXIV.2206.02897 Publisher: arXiv Version Number: 2.

[11] M Bell and European Commission. 2007. Putting Equality into Practice: What role for positive action? *Office for Official Publications of the European Communities* (2007).

[12] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A Convex Framework for Fair Regression. (2017). https://doi.org/10.48550/ARXIV.1706.02409 Publisher: arXiv Version Number: 1.

[13] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 514–524. https://doi.org/10.1145/3351095.3372864

[14] Alycia N. Carey and Xintao Wu. 2023. The statistical fairness field guide: perspectives from social and formal sciences. *AI and Ethics* 3, 1 (Feb. 2023), 1–23. https://doi.org/10.1007/s43681-022-00183-3

[15] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports* 12, 1 (March 2022), 4209. https://doi.org/10.1038/s41598-022-07939-1 Number: 1 Publisher: Nature Publishing Group.

[16] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory?. In *Advances in Neural Information Processing Systems*, Vol. 31. Curran Associates, Inc. https://papers.nips.cc/paper/2018/hash/1f1baa5b8edac74eb4eaa329f14a0361-Abstract.html

[17] European Commission, Social Affairs Directorate-General for Employment, Inclusion, and M De Vos. 2007. *Beyond formal equality : positive action under directives 2000/43/EC and 2000/78/EC.* Publications Office.

[18] A. Feder Cooper, Ellen Abrams, and NA NA. 2021. Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 46–54. https://doi.org/10.1145/3461702.3462519

[19] Court of Justice of the European Union. 1995. Judgment of the Court of 17 October 1995. - Eckhard Kalanke v Freie Hansestadt Bremen. - Reference for a preliminary ruling: Bundesarbeitsgericht - Germany. *European Court reports* 1995

(1995), I–03051. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:61993CJ0450

[20] Court of Justice of the European Union. 1997. Judgment of the Court of 11 November 1997. - Hellmut Marschall v Land Nordrhein-Westfalen. - Reference for a preliminary ruling: Verwaltungsgericht Gelsenkirchen - Germany. *European Court reports* 1997 (1997), I–06363. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:61995CJ0409

[21] Court of Justice of the European Union. 2000. Judgment of the Court (Fifth Chamber) of 6 July 2000. - Katarina Abrahamsson and Leif Anderson v Elisabet Fogelqvist. - Reference for a preliminary ruling: Överklagandenämnden för Högskolan - Sweden. *European Court reports* 2000 (2000), I–05539. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:61998CJ0407

[22] Court of Justice of the European Union. 2000. Judgment of the Court of 28 March 2000. - Georg Badeck and Others, interveners: Hessische Ministerpräsident and Landesanwalt beim Staatsgerichtshof des Landes Hessen. - Reference for a preliminary ruling: Staatsgerichtshof des Landes Hessen - Germany. *European Court reports* 2000 (2000), I–01875. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:61997CJ0158

[23] Court of Justice of the European Union. 2002. Judgment of the Court of 19 March 2002. - H. Lommers v Minister van Landbouw, Natuurbeheer en Visserij. - Reference for a preliminary ruling: Centrale Raad van Beroep - Netherlands. *European Court reports* 2002 (2002), I–02891. https://curia.europa.eu/juris/liste.jsf?language=en&num=C-476/99

[24] Mirjam de Mol. 2011. The Novel Approach of the CJEU on the Horizontal Direct Effect of the EU Principle of Non-Discrimination: (Unbridled) Expansionism of EU Law? *Maastricht Journal of European and Comparative Law* 18, 1-2 (March 2011), 109–135. https://doi.org/10.1177/1023263X1101800106 Publisher: SAGE Publications Ltd.

[25] Marc De Vos. 2020. The European Court of Justice and the march towards substantive equality in European Union anti-discrimination law. *International Journal of Discrimination and the Law* 20, 1 (March 2020), 62–87. https://doi.org/10.1177/1358229120927947 Publisher: SAGE Publications Ltd.

[26] Catherine D'Ignazio and Lauren Klein. 2020. 6. The Numbers Don't Speak for Themselves. *Data Feminism* (March 2020). https://data-feminism.mitpress.mit.edu/pub/czq9dfs5/release/3

[27] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush R. Varshney. 2020. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. JMLR.org, 2803–2813.

[28] 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 (ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[29] European Commission. 2021. 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. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[30] Brian Everitt. 2006. *The Cambridge Dictionary of Statistics* (3rd ed ed.). Cambridge University Press, Cambridge, UK. http://site.ebrary.com/id/10150287 OCLC: 161828328.

[31] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

[32] Will Fleisher. 2021. What's Fair about Individual Fairness?. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 480–490. https://doi.org/10.1145/3461702.3462621

[33] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. https://doi.org/10.48550/arXiv.1609.07236 arXiv:1609.07236 [cs, stat].

[34] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (March 2021), 136–143. https://doi.org/10.1145/3433949

[35] Timnit Gebru. 2020. Race and Gender. In *The Oxford Handbook of Ethics of AI*, Markus D. Dubber, Frank Pasquale, and Sunit Das (Eds.). Oxford University Press, 0. https://doi.org/10.1093/oxfordhb/9780190067397.013.16

[36] Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet. 2020. Projection to Fairness in Statistical Learning. https://doi.org/10.48550/arXiv.2005.11720 arXiv:2005.11720 [cs, math, stat].

[37] Philipp Hacker. 2018. Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law. https://papers.ssrn.com/abstract=3164973

[38] Sara Hajian and Josep Domingo-Ferrer. 2013. A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. *IEEE Transactions on Knowledge and Data Engineering* 25, 7 (July 2013), 1445–1459. https://doi.org/10.1109/TKDE.2012.72 Conference Name: IEEE Transactions on Knowledge and Data Engineering.

[39] H. L. A. Hart. 1957. Positivism and the Separation of Law and Morals. *Harvard Law Review* 71, 4 (1957), 593–629. https://heinonline.org/HOL/P?h=hein.journals/hlr71&i=625

[40] Friedrich August Hayek. 1976. *The Constitution of Liberty*. Routledge & Kegan Paul. Google-Books-ID: CMXanAEACAAJ.

[41] Deborah Hellman. 2019. Measuring Algorithmic Fairness. *Virgina Law Review* 106, 4 (July 2019), 56. https://virginialawreview.org/articles/measuring-algorithmic-fairness/

[42] Sune Holm. 2023. Egalitarianism and Algorithmic Fairness. *Philosophy & Technology* 36, 1 (Jan. 2023), 6. https://doi.org/10.1007/s13347-023-00607-w

[43] David Hume. 1888. *A Treatise of Human Nature*. Oxford: The Clarendon Press. https://oll.libertyfund.org/title/bigge-a-treatise-of-human-nature

[44] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2017/hash/f5f8590cd58a54e94377e6ae2eded4d9-Abstract.html

[45] Michael P. Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 247–254. https://doi.org/10.1145/3306618.3314287

[46] Caitlin Kuhlman, Latifa Jackson, and Rumi Chunara. 2020. No Computation without Representation: Avoiding Data and Algorithm Biases through Diversity. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)*. Association for Computing Machinery, New York, NY, USA, 3593. https://doi.org/10.1145/3394486.3411074

[47] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

[48] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 182:1–182:26. https://doi.org/10.1145/3359284

[49] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (July 2021), 115:1–115:35. https://doi.org/10.1145/3457607

[50] Aditya Krishna Menon and Robert C. Williamson. 2018. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 107–118. https://proceedings.mlr.press/v81/menon18a.html ISSN: 2640-3498.

[51] Colm O'Cinneide. 2006. Positive Action and the Limits of Existing Law. *Maastricht Journal of European and Comparative Law* 13, 3 (Sept. 2006), 351–364. https://doi.org/10.1177/1023263X0601300307 Publisher: SAGE Publications Ltd.

[52] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *Comput. Surveys* 55, 3 (Feb. 2022), 51:1–51:44. https://doi.org/10.1145/3494672

[53] Plato, Harold North Fowler, W. R. M. Lamb, Robert Gregg Bury, and Paul Shorey. 1914. *Plato in twelve volumes: with an English translation*. W. Heinemann ; Harvard University Press, London, Cambridge. OCLC: 25431534.

[54] 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 (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 469–481. https://doi.org/10.1145/3351095.3372828

[55] Tim Räz. 2021. Group fairness: Independence revisited. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 129–137.

[56] Frederick Schauer. 2018. On Treating Unlike Cases Alike. *Constitutional Commentary* 33 (May 2018), 13. https://papers.ssrn.com/abstract=3183939

[57] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[58] Thomas Sowell. 1987. *A Conflict of Visions*. W. Morrow. Google-Books-ID: Fp22AAAAIAAJ.

[59] Jozefien Van Caeneghem. 2019. *Legal Aspects of Ethnic Data Collection and Positive Action: The Roma Minority in Europe*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-23668-7

[60] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. *SSRN Electronic Journal* (2021). https://doi.org/10.2139/

ssrn.3792772

[61] 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 (July 2021), 105567. https://doi.org/10.1016/j.clsr.2021.105567

[62] Peter Westen. 1982. The Empty Idea of Equality. *Harvard Law Review* 95, 3 (1982), 537–596. https://doi.org/10.2307/1340593 Publisher: The Harvard Law Review Association.

[63] White House Office of Science and Technology Policy. 2022. Blueprint for an AI Bill of Rights. The White House. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[64] Michael Wick, swetasudha panda, and Jean-Baptiste Tristan. 2019. Unlocking Fairness: a Trade-off Revisited. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/hash/373e4c5d8edfa8b74fd4b6791d0cf6dc-Abstract.html

[65] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 666–677. https://doi.org/10.1145/3442188.3445928

[66] Jan Wouters and Michal Ovádek. 2021. Equality and Non-discrimination Law in the EU. In *The European Union and Human Rights: Analysis, Cases, and Materials*, Jan Wouters and Michal Ovádek (Eds.). Oxford University Press, 0. https://doi.org/10.1093/oso/9780198814177.003.0007

[67] Raphaële Xenidis and Linda Senden. 2019. EU Non-Discrimination Law in the Era of Artificial Intelligence: Mapping the Challenges of Algorithmic Discrimination. https://papers.ssrn.com/abstract=3529524

[68] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Vol. 54. arXiv, Fort Lauderdale, Florida, USA., 12. https://doi.org/10.48550/arXiv.1507.05259 arXiv:1507.05259 [cs, stat].

[69] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. Achieving Non-Discrimination in Data Release. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. Association for Computing Machinery, New York, NY, USA, 1335–1344. https://doi.org/10.1145/3097983.3098167

[70] Han Zhao and Geoffrey J. Gordon. 2022. Inherent tradeoffs in learning fair representations. *The Journal of Machine Learning Research* 23, 1 (Jan. 2022), 57:2527–57:2552.