Legal Taxonomies of Machine Bias: Revisiting Direct Discrimination

Reuben Binns
University of Oxford
reuben.binns@cs.ox.ac.uk

Jeremias Adams-Prassl
University of Oxford
jeremias.adams-prassl@law.ox.ac.uk

Aislinn Kelly-Lyth
University of Oxford
Aislinn.kelly-lyth@law.ox.ac.uk

ABSTRACT

Previous literature on 'fair' machine learning has appealed to legal frameworks of discrimination law to motivate a variety of discrimination and fairness metrics and de-biasing measures. Such work typically applies the US doctrine of disparate impact rather than the alternative of disparate treatment, and scholars of EU law have largely followed along similar lines, addressing algorithmic bias as a form of indirect rather than direct discrimination. In recent work, we have argued that such focus is unduly narrow in the context of European law: certain forms of algorithmic bias will constitute direct discrimination [1]. In this paper, we explore the ramifications of this argument for existing taxonomies of machine bias and algorithmic fairness, how existing fairness metrics might need to be adapted, and potentially new measures may need to be introduced. We outline how the mappings between fairness measures and discrimination definitions implied hitherto may need to be revised and revisited.

CCS CONCEPTS

• **Computing / technology policy**; • **Applied computing / Law, social and behavioral sciences / Law**;

KEYWORDS

Direct Discrimination, Disparate Treatment, Bias, Fairness, Machine Learning

ACM Reference Format:
Reuben Binns, Jeremias Adams-Prassl, and Aislinn Kelly-Lyth. 2023. Legal Taxonomies of Machine Bias: Revisiting Direct Discrimination. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3593013.3594121

1 INTRODUCTION

In response to extensive research documenting the ways in which algorithmic decision-making systems can generate outputs which systematically favour some protected groups over others, a range of measures for detecting and correcting them have been developed within the computer science literature. These often draw inspiration and motivation from the law, in particular, discrimination law. Commonly, these metrics are presented as operationalizations of disparate impact, one of two types of discrimination defined in US anti-discrimination law. Unlike disparate treatment, which focuses on a person being treated differently because of their protected characteristic (e.g. gender, race, or religion) and is near-universally illegal, disparate impact covers policies which, while facially neutral, result in worse outcomes for members of a protected class. [1] Similar taxonomies exist in other jurisdictions. In EU law,[2] there is a prima facie equivalent distinction between direct discrimination (a person is treated less favorably than a comparable other person on grounds of a protected characteristic), and indirect discrimination (where an apparently neutral provision, criterion, or practice is applied which would put persons with a protected characteristic at a particular disadvantage).

A common assumption within this literature (sometimes explicitly argued for, but more often implied) is that the concepts of disparate impact (US) and indirect discrimination (EU) are the more salient category with which to assess putative cases of algorithmic discrimination, rather than the alternative concepts of disparate treatment (US) or direct discrimination (EU). As a result, the vast majority of algorithmic fairness literature focuses on the problem of assessing and preventing discrimination from a disparate impact / indirect discrimination perspective.

In previous work, we have argued that, at least under EU law, the category of direct discrimination has a wider legal scope than typically assumed in the fair machine learning literature [1]. Drawing on a detailed review of the case law, we argued that direct discrimination captures various types of algorithmic discrimination, even in the absence of direct use of protected characteristics or discriminatory intent on the part of the decision-maker. We outlined several legal implications, in particular the possibility of alternative, potentially more powerful legal challenges to algorithmic systems. We also raised the prospect that this might trigger a re-appraisal and potential revision and expansion of existing algorithmic fairness measures to suit the specificities of direct discrimination under EU law. Given the focus on detailed legal analysis, however, we were unable to explore in greater depth the implications of our argument for technical algorithmic fairness approaches. In this paper, we turn to that task.

We begin in section 2 by outlining how prior algorithmic fairness literature has developed according to an implicit legal taxonomy

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

ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594121

[1] Griggs v Duke Power Co 401 US 424 (1971) (Griggs).
[2] We use this term to cover both European Union (EU) and UK law, as the approach in both jurisdictions is closely aligned for present purposes. Following the United Kingdom's exit from the European Union, UK courts should continue to have regard to developments in the EU equality acquis: European Union (Withdrawal) Act 2018, s 6. For an example of EU legislation drawing a distinction between direct and indirect discrimination, see 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 (recast) (2006) OJ L204/23 (the Recast Directive), art 2.

which aligns 'easy' cases with disparate treatment / direct discrimination and the computationally 'tough' cases with disparate treatment / indirect discrimination, and diagnosing why the literature has developed in this way. Section 3 briefly summarizes the legal arguments of our previous paper, including the crucial differences between US disparate treatment and European direct discrimination, and explains how two particular types of direct discrimination (inherent and subjective direct discrimination) might apply to algorithms. We then address the implications for existing and future work on technical approaches to algorithmic fairness. Section 4 outlines how data scientists might go about detecting and correcting inherent and subjective discrimination in algorithms. Having examined how the algorithmic fairness literature may need revising to account for these different types of direct discrimination, section 5 considers the converse; might existing technical approaches to discrimination in ML provide a stronger conceptual basis on which to revise or reinterpret the current, arguably unprincipled, set of distinctions drawn in EU discrimination law?

2 LEGAL TAXONOMY: SHAPING THE DEFAULT APPROACH TO FAIR ML

The literature on algorithmic fairness metrics has not generally sought comprehensively to map various types of machine bias to specific legal categories. Upon closer inspection, however, the strong influence of the law's taxonomy of discriminatory conduct on prior fair machine learning literature becomes apparent. As categories of fairness metrics and types of bias have emerged in line with legal categories which may make sense in the US discrimination law context, however, their transplant into EU equivalents does not always work [22, 28]. This, in turn, has important implications for the presumed scope of EU discrimination law as applied to algorithms. To diagnose how this happened, a very brief, highly potted history may be helpful.

Since its inception, fair machine learning literature has typically recognized two main ways that 'unfair' algorithms can arise, with corresponding types of technical fairness metrics. The first source of algorithmic unfairness arises where a decision-maker uses a protected characteristic as an input to an algorithm, thereby favoring or disfavoring the relevant protected group; these are implicitly seen as 'easy' cases. The second source of algorithmic unfairness arises where protected characteristics are not explicitly used but are somehow encoded in other features. This includes cases where a supposedly benign feature is known to be correlated with a protected characteristic (e.g. ZIP code and race); as well as cases where an ML algorithm might extract a high-level feature from the training data which, unknown to the data scientist building the model, also happens to be a proxy for a protected characteristic. Take the infamous example of the Amazon hiring algorithm, which scored candidates' résumés in order to select the best applicants for software engineering positions.[3] Because few previous applicants had been female, the model learned to give lower scores to resumes containing words associated with female applicants; these included the word 'women's' (as in 'women's chess club captain') and the names of women-only colleges. Such examples are regarded as 'tough' cases, and they motivate the adoption of various statistical fairness metrics which focus on the distributions of outcomes and / or errors between protected groups.

To illustrate how this implicit distinction is made in computer scientific algorithmic fairness literature, it is useful to broadly characterize the way fairness metrics and typologies of machine bias are typically defined therein. Fairness metrics are typically discussed in relation to a machine learning prediction problem. A dataset drawn from a population X, has a set of protected attributes A (e.g. sex / gender, race) and other features Z (e.g. income, loan repayments). A model M is trained using a machine learning algorithm on a dataset consisting of a sample of X, with features Z and labels Y, which might represent some outcome (typically a binary one, e.g. default or repay on a loan). M provides a mapping from features to predictions ($Z \rightarrow \hat{Y}$) and the values of those predictions are used to determine the allocation of some benefit (e.g. credit, hiring) for previously unseen individuals $x_i \in X$.

Algorithmic fairness approaches typically begin by acknowledging one very simple, but insufficient, fairness definition, which goes by various terms including: fairness through unawareness or blindness [9]; treatment parity [32]; and anti-classification [19]. This states that M should make predictions on the basis of Z alone, not A. This is the fairness definition which is typically taken to describe the 'easy' cases, where a decision maker seeks only to remove protected characteristics from use as an input to the model. Fairness through unawareness is typically only raised in fair machine learning discussions in order to be swiftly dismissed as insufficient, due to the fact that unfairness may still be possible where one or more of Z are a proxy for A [9].[4] Such proxies have typically been taken as representing those 'hard' cases. One simple approach often noted in this literature involves attempting to determine if any of Z are proxies for A, and remove them if so. However, this simple approach is typically rejected on the grounds that it would likely significantly diminish the available Z, since almost all predictively valuable variables are correlated to some extent with at least one of the protected attributes. Instead, various fairness metrics are defined in terms of outcomes and errors, conditional on membership in the protected group. These allow the underlying attributes Z to be used even if they may be proxies for A, because fairness can be applied as a constraint on the model rather than on individual features. For example, according to one such metric, demographic parity, predictions $\hat{Y}$ should be independent of protected attributes A (i.e. $Pr(\hat{Y}|\ A) = Pr(\hat{Y})$, or simply $\hat{Y} \perp A$).

The canonical literature on algorithmic bias usually implies that this division between fairness-through-unawareness (which focuses on excluding protected attributes from inputs) and other statistical metrics (which focus on the outcomes of the model and compare their impact on different protected groups), maps onto an equivalent division between two types of discrimination in US law, namely disparate treatment and disparate impact. Such papers typically acknowledge that disparate treatment only applies to algorithmic discrimination in very limited cases, specifically where the decision-maker deliberately includes protected characteristics as an input to the algorithm. Feldman et al., for example, focus on disparate impact, noting that it is "different from disparate treatment, which refers

[3] https://perma.cc/328A-UJFM

[4] For an illuminating discussion of what it means for one variable to proxy another in this context, see [26]

to *intended* or *direct* discrimination" [11]; (first emphasis original, second emphasis added). A background assumption here is that the decision-maker is aware that, by law, they are "not allowed to use [protected attributes] in making decisions, and claim[s] to use only [non-protected attributes]". Assuming such awareness means we can rule out the direct use of protected attributes which would indicate disparate treatment. The challenge thus becomes one of how to demonstrate disparate impact. Barocas and Selbst's seminal paper from 2016 similarly argued that most cases of algorithmic bias will, in US law, fall under the category of disparate impact, rather than disparate treatment; the exceptions being where protected attributes are used, or where proxies for them are used deliberately in order to disguise ('mask') discriminatory intent [2]. They point out that just as errors in data collection, mislabeled examples, or insufficiently rich features can lead to unintended disparate impact, those same errors could be made intentionally and would, absent evidence of the decision-maker's true intent, appear innocent.

Taking these examples together, we can surmise several reasons for the relative dearth of attention to direct discrimination. First, there may be an assumption that most decision-makers would already be aware of their obligation to avoid using protected characteristics directly. Furthermore, it may be assumed that those who deliberately engage in masking would not be receptive to the honest implementation of algorithmic fairness in the first place.

Second, even if cases of deliberate use of protected characteristics, or deliberate masking via proxies, were in fact frequent, the lack of focus on them might also be explained by the perception that they raise few technically interesting problems. Avoiding disparate impact through fairness-through-unawareness would simply involve examining the features used by the model to ensure no protected characteristics are used as inputs to prediction. Similarly, determining whether a decision-maker has engaged in masking with discriminatory intent is a matter of fact best established in individual cases, and in any event not amenable to algorithmic measures of fairness. As a result, an implicit distinction has thus been drawn between cases where protected characteristics are explicitly used as inputs, and cases of unintended discrimination in the absence of explicit use of protected characteristics. The former are regarded as technically easy or uninteresting cases, covered by the simple definition of 'fairness through unawareness' and dealt with through disparate treatment; while the latter are tough cases, requiring new technical fairness definitions and methods for detection and mitigation, and to be addressed through the doctrine of disparate impact (or indirect discrimination in EU law).

Finally, a preference for statistical fairness measures defined over impacts on groups arguably reflects a broader philosophical orientation of machine learning as a discipline. As a modelling paradigm, it is highly open, admitting a wide variety of different learning algorithms, which are primarily evaluated not based on their internal logic, complexity, or features used, but rather simply on how well they perform on the test dataset. This inculcates a desire for common performance metrics which can be applied across broadly different model types (e.g. a simple decision tree model and a complex many-layer deep learning model can both be evaluated by the same set of metrics, e.g. the F1 score). It is unsurprising, therefore, that proposed fairness measures should fit into that paradigm, focusing on the distributions of outcomes or errors across groups, especially where they can be computed using existing performance metrics. By contrast, measures which require focusing on individual variables that might be protected attributes (whether directly or proxies for them) necessitate a contextual, case-by-case analysis more suited to practitioners working on the ground, rather than abstract and generalizable measures. As a result, we argue, the fairness literature has overwhelmingly focused on these statistical measures defined over sub-groups.

In so far as disparate treatment / direct discrimination has entered into debates about algorithmic fairness, it is confined to cases of a) deliberate direct use of protected characteristics, which are 'easy' to solve, or b) deliberate use of masking to conceal discriminatory intent, proof of which lies outside the remit of computational approaches. These examples raised in the technical literature do not, however, exhaust the scope of key legal concepts. Are there other ways in which an algorithm might be directly discriminating even if protected characteristics (or deliberate proxies) are not used? The consensus of scholars working within US discrimination law appears to be 'no', because disparate treatment requires intent [15]. But what about other discrimination law frameworks? At first glance, the answer is similar. Scholars working within European discrimination law appear to have largely followed the same direction, focusing on indirect discrimination as the primary framework for challenging discriminatory algorithms [13, 29]. The basic taxonomy of EU and US anti-discrimination law is beguilingly similar, both in so far as the direct / indirect distinction appears to map onto the disparate treatment / disparate impact distinction, and also in so far as the former will nearly always be illegal, whereas the latter may be justifiable. Crucially for our purposes, however, there are some key differences regarding the scope of direct discrimination compared to disparate treatment, and in particular, the role of intent.

While these differences between US and EU discrimination law are long-acknowledged, leading scholars have nonetheless largely concurred that direct discrimination is rarely applicable to algorithms. Hacker, for example, follows Barocas and Selbst's approach, arguing that direct discrimination will be rare, arising when a protected characteristic is used as an input in a model or where an algorithm is designed to disadvantage certain protected groups [13]. The possibility that algorithms might directly discriminate even without direct use of protected characteristics is acknowledged only in certain exceptional cases. For instance, Hacker notes that an algorithm trained on data reflecting the decision-maker's unconscious bias might also constitute direct discrimination (since European case law established implicit bias as a form of direct discrimination), but nevertheless argues that "[i]n machine learning contexts, direct discrimination will be rather rare", excluding several main categories of algorithmic bias, including: sampling bias (where different protected groups are sampled from differently, leading to an unequally performing model); historical bias (where training data reflects historical prejudice); and unequal ground truth (where "capacities or risks are unevenly distributed between protected groups") [18](pp. 1151-1152). Similarly, while Wachter draws on judgments of the Court of Justice of the European Union ('CJEU') to argue that the use of "affinity profiling" – where e.g. Facebook users can be targeted based not on their race, but rather on e.g. their "interest in black culture" – might amount to direct

discrimination, her analysis is limited to the specific category of discrimination 'by association' [27]. These exceptions aside, the general consensus appeared to be that the concept of direct discrimination is "likely to be less important than that of indirect discrimination" in the context of algorithms [29].

3 DRAWING THE LINE IN EUROPEAN LAW

This section draws on legal arguments presented in previous work, which reject the consensus presented above [1]. While the detailed analysis cannot be reproduced in full, we provide an overview of the main argument and conclusions in order to illustrate how the implicit division in the algorithmic fairness literature is fundamentally at odds with EU law.

Even though the direct / indirect discrimination distinction is structurally similar to the categories of disparate treatment / impact, the dividing lines are drawn in fundamentally different way. The concept of direct discrimination is broader than disparate treatment: it applies to a wider range of algorithmic systems. Disparate treatment under US law is generally understood to require either explicit classification or evidence of discriminatory intent [41]. By contrast, EU law focuses on the reasons or grounds for a decision, rather than the purported discriminator's intention or motive: "the existence of prejudice, or an intention to discriminate, are not actually of relevance to determining whether the legal test for discrimination has been satisfied" [10]. Indeed, "[t]he dividing line between direct and indirect discrimination is emphatically not to be determined by some sort of mens rea on the part of one or more individual discriminators" [33] (70). Since unintentional discrimination can thus be 'direct' in EU law, behaviour which does not amount to disparate treatment in the US may well constitute direct discrimination in Europe. This is particularly salient for present purposes, as intentionality is often hard, if not impossible, to attribute in cases of algorithmic decision making, whereas much of human conduct is perceived through the lens of intentionality.

If not intentionality, what does distinguish direct discrimination from indirect discrimination? Two factors emerge from the case law. First, while the prohibition on direct discrimination is intended to achieve formal equality, the rules on indirect discrimination seek to advance substantive equality. Secondly, direct discrimination is reason-focussed, whereas indirect discrimination is effects-focussed. In *James v Eastleigh BC*, for example, a local council in the UK offered free swimming pool access to individuals eligible to receive the state pension. Since the pensionable age at the time was 60 for women but 65 for men, this meant that a 60 to 64-year-old woman could enter for free, while a man of the same age would have to pay. This constituted direct discrimination: there was formal inequality of treatment; the term 'pensionable age' was "no more than a convenient shorthand expression which refer[red] to the age of 60 in a woman and to the age of 65 in a man" [James, 780]. In other words, the reason for the differential treatment was sex, even if the Council had no intention to discriminate on that basis. The fact that the caselaw thus severs direct discrimination from moral blame has been criticised by leading scholars [12](pp. 39-46), but remains the position in law.

Direct discrimination can subdivided further into two categories: decisions made using an inherently discriminatory criterion, and decisions made through subjectively discriminatory mental processes [34](78).[5] Inherent discrimination occurs when a criterion used by a decision-maker is 'inextricably linked' to a protected characteristic. As the CJEU has recently found, even though a universally applied rule prohibiting any visible sign of political, philosophical, or religious belief in the workplace was not liable to constitute direct discrimination, a prohibition on "conspicuous, large-sized signs" could nonetheless constitute direct discrimination [35] (55, 78). The Court reiterated that "unequal treatment resulting from a rule or practice which is based on a criterion that is inextricably linked to a protected ground, in the present case religion or belief, must be regarded as being directly based on that ground" [35] (73), emphasis added).

Our previous analysis also identified how inherently discriminatory criteria can arise from the interaction between various rules and contingent historical facts, which may not individually pick out protected characteristics, but do so only when combined. We raised the hypothetical example of a formerly boys-only high school which began admitting girls in 2010. Imagine that an employer will only hire a job applicant if they are (i) a graduate of that school and (ii) were born before the year 1990. Neither of those criteria are inherently discriminatory when taken individually: some of the graduates of the school are women, and so are about half of the people born before 1990. But no woman meets both criteria when applied together. The employer's rule is therefore inherently discriminatory.

Three concrete examples demonstrate how this type of direct discrimination might arise in algorithmic systems in practice.

1) In a jurisdiction where same-sex couples cannot marry, and only same-sex couples can obtain civil partnerships, the features 'married' or 'in a civil partnership' are inextricably linked to sexual orientation [40]. Even if being married or in or civil partnership are not themselves protected characteristics (which is the case in some EU jurisdictions), these features would in this case be proxies for the protected characteristic of sexual orientation. If a credit scoring model positively or negatively correlates marriage or civil partnership with creditworthiness, this could therefore constitute direct discrimination via an inherently discriminatory criterion.

2) In natural language processing, a feature which is indissociable from a protected characteristic might be automatically extracted. Take, for example, the case of Amazon's recruitment algorithm, which gave lower scores to members of women's extra-curricular clubs and to applicants from women-only colleges. No human chose to use these features, and without the bias detection efforts on part of the data scientists involved, it is likely that no-one would have even been aware that the features were being used by the model. Even if a data scientist had examined all the features, they might not have been able to tell which were indissociable from gender without further domain knowledge (e.g. knowing whether a particular college is co-educational). Despite this, such features are indissociable proxies for gender.

[5] While the relevance of UK case law to the EU is diminished post-Brexit, these cases remain valid case law in the UK. Following the United Kingdom's exit from the European Union, UK courts should continue to have regard to developments in the EU equality acquis: European Union (Withdrawal) Act 2018, s 6.

3) More generally, machine learning methods in which features are constructed automatically out of high-dimensional, non-human interpretable data, may create features which are proxies for protected characteristics in ways which are not easily detectable. An inferred latent variable which is inextricably linked to a protected characteristic would be an example of an inherently discriminatory criterion. The link between the criterion and the protected characteristic need not be facially obvious, well-established, or enduring. If it can be shown that there is a sufficient degree of correspondence between the criterion and the protected characteristic, this in itself can be highly persuasive or even decisive in proving the criterion to be inherently discriminatory. As such, even if the decision-maker had no intention to use a protected characteristic, and was entirely unaware that a latent variable inferred by the model correlated one-to-one with a protected characteristic, they could still be engaging in direct discrimination. Since the supposed benefit of many modern ML methods is their ability to infer latent features not explicitly represented in their training data, which may include protected characteristics, such methods are more likely to result in models which could inadvertently directly discriminate in this way.

Subjectively discriminatory decision-making, on the other hand, arises when a person's protected characteristic influences the decision-maker's conscious or subconscious mental processes, such that a different outcome is reached [34](64). It was established early on that the subjective mental processes which constitute direct discrimination need not be conscious [37]. The protected characteristic does not have to have been the only or even the main cause of the result complained of; it is enough that it was *a* cause, that it "*had a significant influence on the outcome*" [37, 38]. Take the example of a recruiter who, on receiving a job application from a woman, subconsciously takes a dimmer view of it. He does not object to anything specific in the application, but multiple indicators of gender (e.g. playing a feminine-coded sport) cumulatively affect his overall impression. A female applicant who misses out on the job to a similarly qualified man would, in these circumstances, have a claim for direct discrimination. We previously argued that this translates readily into the algorithmic context. For example, the Amazon algorithm learned not only to mark down applicants from women's colleges (which would be inherent discrimination, as argued above), but also to preference applications using active verbs—words which are more commonly used by men. Like the human recruiter, the algorithm was not 'aware' that its outputs were influenced by gender, yet they were. The legal position cannot be any different because unfavourable treatment is meted out by an algorithm, rather than a human. Therefore, while not an inherently discriminatory criterion, the preferencing of active verbs is a manifestation of implicit bias, and therefore subjective discrimination.[6]

4 FAIRNESS MEASURES AND UNINTENDED DIRECT DISCRIMINATION

Based on our previous legal analysis, we argue that certain canonical cases of algorithmic fairness should, in the EU and UK context, be considered as examples of direct discrimination. This includes many of the 'tough' cases of algorithmic fairness. The implications of treating them as *direct* rather than indirect discrimination are hard to overstate: direct discrimination can only be justified in a strictly limited number of circumstances. [7] While we originally raised these legal arguments with the hope of influencing how lawyers address discriminatory algorithms, they also raise potentially novel challenges for computer scientists engaged with algorithmic fairness to apply their metrics and techniques to the 'tough' cases of potential direct discrimination. Counter to the implied division in extant algorithmic fairness literature, a number of examples from the algorithmic discrimination literature fall squarely into the existing categories of direct discrimination. In this section, we outline how the mappings between fairness metrics and discrimination law categories that have been implicitly assumed hitherto may need to be updated in light of this.

As discussed in section 2, there has been an implicit mapping in existing fair ML literature between disparate treatment (or direct discrimination in EU law terminology), and disparate impact (or indirect discrimination), to a set of mutually exclusive fairness measures. In the EU law context, this existing mapping implies that anti-classification is necessary and sufficient for avoiding direct discrimination (with the exception of deliberate masking). Various group fairness measures have been proposed as at least prima facie evidence of indirect discrimination. In particular, Wachter et al. propose 'conditional demographic disparity' (CDD) as a standard baseline statistical measurement that aligns with the European Court of Justice's 'gold standard' for assessment of prima facie (indirect) discrimination [28].

However, the various examples of *directly* discriminatory algorithms described above suggest the need for more nuance in this implicit mapping. Can existing fairness metrics, hitherto deployed to address indirect discrimination, be re-used for direct discrimination? What other technical approaches might need to be developed or applied? In this section we begin by examining these questions in light of the two categories of direct discrimination outlined above: inherent and subjective direct discrimination, before moving on to consider how other types of unintended direct discrimination, beyond the inherent and subjective, might be dealt with in terms of algorithmic fairness.

4.1 'Inherently Discriminatory' Algorithms

How might a data scientist assess whether a model is inherently discriminatory? The possibility of indissociable proxies suggest the need for different kinds of metrics and analysis to detect them. Consider a hypothetical example which is similar to the context of *James v Eastleigh BC* [36]. As part of a public health campaign, a local government is attempting to select members of the public to receive free entry to sports leisure facilities. They use a machine learning model designed to identify those whose health will benefit most from the intervention. They use (what they believe to be) non-protected features Z, including a 'receives_state_pension' variable, and a 'm_health_clinic' variable which records if they are registered at a particular type of health clinic. How might they uncover any

[6] This raises difficult legal issues of causation, i.e. the scope of 'because of' when considering upstream discrimination; see section 5.2 below.

[7] For more discussion of this and the prospects for legal challenges against directly discriminating algorithms, see [1].

variables (or combinations of variables) which may be indissociable from a protected characteristic?

To address the simplest cases of indissociable proxies, the data scientist might take each of the Z variables and test if they are correlated with any of the protected characteristics. They might produce a correlation matrix comparing the Z variables against each of the protected attributes A. Looking at the various cells in the 'm_health_clinic' matrix, the data scientist observes a perfect correlation with the protected characteristic of pregnancy status. Unknown to the data scientist prior to the analysis, the 'm_health_clinic' variable refers specifically to maternal health centres which only serve birthing people. Therefore, the data scientist can now conclude that the model is directly discriminating, because 'm_health_clinic' is indissociable from pregnancy status. While this is a simple example, which a data scientist with better knowledge of the meaning of the variable in question might have been able to anticipate without even conducting the analysis, it at least demonstrates how other less obvious indissociable proxies might be identified.

Turning to the P column in the correlation matrix, the data scientist would find that P is correlated with the age variable A. Age is a protected characteristic, so a model which uses P is likely to present prima facie evidence of indirect discrimination by age (which might be justifiable as a proportionate means of a legitimate aim in the context of public health). However, since the correlation is not perfect, the data scientist might conclude that this is at most indirect discrimination, but is not direct discrimination. However, this would be a mistake. As shown in the example of *James* mentioned above, indissociable proxies need not perfectly correlate to the protected characteristic in all cases. The state pension in that context was given to men from age 65 and women from age 60; on this ground, the court found that free entry to the swimming pool for people of state pension age was held to be inherently discriminatory. The court did not conclude that because some men do receive the state pension, and some women do not, that P is not inherently discriminatory. As Campbell and Smith point out, there was no exact correlation between the adverse treatment (having to pay) and the protected group (men), because some women did have to pay to enter the swimming pool [7](pp. 269-270). Rather, the criterion in question only discriminated against men of a certain age (namely, those between 60-65). In other words, it is not necessary that a criterion perfectly captures all members of a certain group; it is enough that there is some subset, defined in terms of other characteristics, within which one protected group is treated worse than the other.

It would therefore be sufficient for indissociability if a proxy is perfect for only a particular range of values within that proxy. A correlation matrix which only compares the whole distribution of A against Z will therefore not be sufficient for this purpose, since it presents the correlation between A and the putative indissociable proxy Z overall; the range-specific indissociability of the proxy will thus be 'hidden' behind the overall figure. A simple remedy would be to split the dataset by the values of each attribute A, and produce a correlation matrix for each value. For instance, one correlation matrix would be produced for each sex / gender; another for each sexual orientation; etc. This would reveal to the data scientist that when disaggregated by sex / gender, P and A are perfectly correlated, showing that P is an indissociable proxy for sex / gender.

There are likely many alternative, potentially more efficient ways of detecting indissociability within a set of protected attributes (and sub-ranges within them). However, any such approach would face additional challenges. First, the number of comparisons might need to be quite extensive. There are nine protected characteristics in EU equality law; each of them can take multiple values. Some have a relatively limited number of values, e.g. 'marital status' but others have an indeterminate and potentially large set of possible values, e.g. 'religion or belief'. Each of these values would need to be compared against one another for every feature in Z. Second, only some of the protected characteristics can be modelled as categorical variables: age, for example, is continuous. Such cases raise questions about how to partition and compare different values against each other (a more general problem for fairness metrics which assume protected classes to be categorical). Finally, correlation analysis would need to account for cases where the relationship is not linear. There may also be a need for model-specific measures to capture how inherent discrimination may arise for particular non-linear model types. For instance, in decision tree models, any sub-node, or multiple sub-nodes within a decision path, which splits perfectly between values of A, would arguably constitute an inherently discriminatory decision tree model.

Putting these difficulties to one side, there is a further complication relating to an ambiguity in the law, namely: how strong does the relation between the proxy and the protected characteristic need to be? The case law has given rise to divergent answers. UK courts have held that the correspondence has to be 100%, i.e. all individuals who fail to meet the criterion share the particular protected characteristic, while the CJEU has taken a more flexible approach, holding that the harms of the criterion need not fall exactly along lines of protected characteristics. [8] This 'indissociability' has been more qualitative in nature; it might be enough that a model only uses an imperfect (but highly correlated) proxy for a protected characteristic. The UK approach, on the other hand, suggests that a criterion which acts as a proxy for sex may be used by a decision-maker if 99% of those impacted by it are women; but if the proportion rises to 100%, then the rule is unjustifiable direct discrimination. On the other hand, as illustrated in *James*, UK courts have held inherent discrimination can occur even when a proxy is only perfectly correlated with a protected characteristic within a limited range (i.e. being a state pensioner within the age of 60-65 is perfectly correlated with sex, even though being a state pensioner in general is imperfectly correlated with sex). Wherever the line is drawn between inherent direct discrimination and mere indirect discrimination, if it is based merely on a degree of correlation, then it may seem arbitrary; as Collins & Khaitan argue, this is hardly a principled distinction [8] (p 20). While such legal ambiguities cannot be resolved here, either interpretation is in principle amenable to statistical analysis of the kind described above; either enforcing the 100% correlation constraint, or relaxing it to whatever threshold may be legally applicable in a given context.

[8] For further discussion of these doctrinal differences, see [1].

Furthermore, there is a related question of how many examples would need to be observed before we conclude that a given variable or sub-range of a variable is indissociable from a protected characteristic. Are observations of the distributions of protected characteristics of individuals from the training data sufficient evidence, or do we need unbiased figures for the entire population (e.g. from census data)? What if there are differences between training / test data and deployment? How could we distinguish a subspace which is robustly indissociable from one which only spuriously appears to be so due to unrepresentative training data? The examples raised earlier are plausibly robust in the appropriate ways; in *James*, the proxy variable of statutory retirement perfectly predicts sex / gender because of a government policy; in the previously boys-only school example, the interaction between age and graduation are reliably associated with gender because of the school's history. But if there are many subspaces to test and many different protected groups (including some with relatively few members), some apparently exclusive subspaces will eventually be found by chance, prompting the potential need for multiple-comparison correction measures. Such questions are similar to those raised in relation to robustness of fairness measures under covariate drift [23], and will be particularly acute under a strict one-to-one interpretation of inherent discrimination as implied by some UK case law; under the more relaxed interpretation developed by the CJEU, a subspace would not have to be guaranteed to exclude people with a given protected characteristic entirely to qualify as inherently discriminatory.

4.2 Subjectively Discriminatory Algorithms

Subjective discrimination seems less amenable to any form of fairness measure based solely on analysis of the model or the training set alone. This is because no amount of data or model analysis allows one to see into the minds of the potentially subjectively biased human labellers who generated the labels in the training data to see what was in fact influencing them subconsciously.

This doesn't mean that subjectively discriminatory algorithms are entirely undetectable; after all, in purely human cases of subjective discrimination, the courts could not see inside the human minds of those responsible. Appropriate evidentiary standards would therefore have to be developed and adapted to the algorithmic context. One could imagine that the statistical fairness methods developed hitherto might be useful as prima facie evidence of potential subjective discrimination (similar to the proposal to use conditional demographic disparity as prima facie evidence of indirect discrimination [28]), although further evidence would be needed to distinguish between cases where the labelling process is driven by unconscious biases from those in which broader upstream social structures are the cause of disparities in the training data.

5 COULD TECHNICAL DEFINITIONS CHALLENGE LEGAL DISTINCTIONS?

Thus far, we have argued that the mapping of the EU legal concepts of direct and indirect discrimination onto measures of algorithmic fairness has become somewhat misguided as a result of the focus on US law in technical fairness definitions. In this section, we consider the converse: might debates about fair machine learning provide a stronger conceptual basis on which to revise or reinterpret the current, arguably unprincipled, set of distinctions drawn in EU discrimination law?

5.1 Challenges to the Legal Taxonomy of Direct Discrimination

Previous sections examined how two specific types of direct discrimination – inherent and subjective direct discrimination – might apply to algorithmic bias. However, there are some cases of algorithmic discrimination which do not appear to fit well with either category, but for which it may seem intuitively correct to say that some individuals are treated worse 'because of' their protected characteristics. [9] One prominent example is facial recognition systems which perform worse on darker-skinned women, as studied in the Gender Shades project [6]. Do they fail black women 'because of' their protected characteristics, in the sense required for a direct discrimination case?

Note first that, on the face of it, neither subjective nor inherent discrimination applies here. Unlike the sexist human recruiter case, we cannot point to biased mental processes of human labellers of the training data used to build the facial recognition system, so subjective discrimination seems inapplicable. Inherent discrimination also seems an unlikely fit, as it is unclear what features would be acting as proxies indissociable from the intersection of race and gender. There would be no feature or combination of features in the model that could be straightforwardly identified as a proxy for race. Indeed, to attempt to identify such features could be seen as implying a problematic conception of race as a set of physiognomic features [24, 31]. As such, it would be hard to make out an inherent discrimination case.

However, despite not fitting either of the extant categories of direct discrimination, it still seems correct to say that in some sense, these facial recognition systems treated darker-skinned black women worse at least in part *because of* their race and gender. Such a claim need not imply that protected characteristics like race provide a singular causal explanation, nor even that they can be causally modelled as such, although they may be. We note here that there is debate within the causal inference literature, and the philosophy of social science, as to whether categories like race and gender can be meaningfully captured in causal models [5, 14, 16]. We do not aim to comment on this debate here; suffice to say that whatever ontological status the law affords or assumes about protected characteristics, we contend that it ought to be able to capture the straightforward intuition that the relatively poor performance of the systems evaluated in the Gender Shades project is in the relevant sense *because of* the intersection of race and gender.

One might object that those systems treated darker-skinned women worse not because of the intersection of race and gender but rather because of the unrepresentative training data. This latter kind of explanation – which appeals not to the protected characteristic of the subject of the decision, but rather to the unrepresentative

[9] While we use here the phrase 'because of', found in UK legislation implementing EU law, the alternative phrase 'on the grounds of' often appears in the latter; see e.g. the EU The Race Equality Directive 2000/43/EC.

distribution of protected groups in the training data – might be marshalled as a defence against an accusation of direct discrimination. But the latter explanation is not a mutually exclusive alternative to the former. Indeed both factors are necessary to explain the worse treatment faced by darker-skinned women; it is the combination of the unrepresentative training data *and* the protected characteristics of those receiving worse treatment which cause these particular outcomes. This suggests that facial recognition systems exhibiting gender and racial biases in this way therefore ought to fall within the scope of direct discrimination.

This outcome would require the judicial recognition of a new, additional category of direct discrimination – an option open to courts in their interpretation of broadly worded statutes, the meaning of which can develop long after their initial framing [3](section 14.1). A new type of direct discrimination better suited to particular kinds of algorithmic discrimination, such as those exhibited in facial recognition models, might also give rise to the need for a set of robust and established technical measures for proving it.

5.2 Defining the Scope of 'Because of'

Examples of implicit bias leading to a discriminatory model raise further questions about how to interpret the scope of the 'because of' / 'on the grounds of' language key to the concept of direct discrimination. If we grant, for the purposes of applying the legal category, that a model directly discriminates when it treats people with a protected characteristic worse because it was trained on data labelled by an implicitly biased human, what other sources of bias might also count as direct discrimination? It seems plausible that a similar account of the influence of protected characteristics on model outputs could in theory be applied to additional examples of historic bias. For instance, gender or racial stereotypes may be reproduced in search engine results in part because of the impact of gender and race on the behaviours of those who upload and search for content on search engines. If we can attribute at least a significant part of a model's biased outputs to the impact of protected characteristics on the social processes which reproduce those characteristics, can we say that direct discrimination arises, whether through an implicitly biased labeller or via historic societal interactions with search engines? Put differently, how far back in a causal chain must a protected characteristic be to be too remote to be attributed as a significant cause of the algorithmic bias?

As outlined in our prior work [1], the courts have sought to avoid overly legalistic approaches to answering this question [*Nagarajan*], but some limits can be identified. Take, for example, a reference provided by a discriminatory previous employer, which has been tainted by the ex-employer's bias against its employee. The recipient prospective employer decides not to hire the applicant on the basis of the bad reference. In some sense, the failure to hire is 'because of' the applicant's protected characteristic, but the courts have held that the prospective employer will not be liable for direct discrimination in these circumstances [39]. Still, it is unclear under what circumstances a third party breaks the causal chain that would otherwise make a decision-maker liable. While the source of the bias in *Reynolds* [39](i.e. the previous employer) is one step removed from the decision-maker (the prospective employer), the same was arguably true in *James* [36]: age only operated via the definition of pensionable age, and that definition (like the bad reference) was provided by a 'third party' of sorts (in this case, the legislature). Imagine a workforce evaluation model trained on a biased sample which doesn't include sufficient representation of a protected group due to historical exclusion from the industry. In this case, the cause of the sampling bias is arguably the explicit or implicit bias of previous decision-makers, which has in turn caused the model to treat new workers from the protected group worse. An 'upstream' process, in which members of a protected class were discriminated against, has led to a downstream process in which other members of that same class are discriminated against. If we grant (as argued above) that training data labelled by biased hiring managers can result in directly discriminatory recruitment algorithms, why wouldn't direct discrimination also apply where labelling decisions were made further upstream, for example by a third-party vendor of AI services for recruitment? The disadvantage experienced by applicants would be no less 'because of' the protected characteristic in these circumstances, and as we argued in section 4.2, automating a biased process should not insulate the decision-maker from liability. On the *Reynolds* approach, however, it might appear that a biased algorithmic output which originates from external bias – like the bad reference – is merely 'tainted information', the use of which cannot create liability. The law therefore lacks a clear means of measuring the degree of distance in a causal chain, and a means for deciding where along that chain a cause ceases to qualify as 'direct enough' for a case of direct discrimination.

These ambiguities can only be satisfactorily resolved in the courts or through new law. However, some technical and empirical work in fair machine learning might provide a useful guide in the absence of legal certainty. First, the field has already provided various taxonomies of algorithmic bias which carefully distinguish different sources of bias at different stages upstream of the decision-making process. For instance, Mitchell et. al characterise various forms of bias, including 'societal bias' which concerns objectionable social structures and past injustices (which may lie entirely upstream of the decision maker), nonrepresentative sampling and measurement error (which occur around the point of data collection), model bias and evaluation bias (which are located around the choice of model) [21]. Such typologies of bias could provide a useful heuristic for courts seeking to understand where to draw the line between direct and indirect discrimination.

Second, there is a burgeoning literature which seeks to measure algorithmic fairness not in terms of a single decision-making point, but rather in terms of sequential interactions, interventions, and causal processes which led up to a final decision outcome. For instance, work on 'fairness in pipelines' examines how compound decision-making processes can result in unfairness [4]. Similarly, causal models of fairness attempt to create graphs which model the causal relationships between variables that account for disparities between protected classes [17, 18, 30]. [10] Empirically modelling pipelines and causal relationships in this way could help trace and measure the contribution of upstream decisions to downstream

[10] As mentioned above, there are debates as to whether protected characteristics can themselves meaningfully be represented as nodes in such graphs [5, 14, 16]; but even if they cannot, it may still be possible to causally model the social structures which give rise to disparities between protected groups [20], in ways that are informative for litigators seeking to establish a case of direct discrimination.

unfairness, enabling litigators to more clearly attribute direct discrimination, whether it arises from biased labellers of training data, historical exclusion, modelling decisions, or some other process. In other areas of law, the issue of causation has received significant academic and judicial attention [25]. The question of upstream causal impact is not new—but the algorithmic context brings renewed urgency to finding coherent solutions when it comes to discrimination.

6 CONCLUSION

Given the inherently interdisciplinary nature of research on fairness and discrimination in algorithmic decision-making, and the significant challenges of translating between law and computer science, it would be understandable for computer scientists to focus their efforts on implementing a select range of key legal definitions. But to date, this selection has been almost exclusively focused on one legal framework and sub-type of discrimination within it, namely the US disparate impact doctrine. This has in our view skewed the development of algorithmic discrimination research in computer science and law in an unhelpfully narrow direction, and perhaps also had a similarly narrowing effect on approaches to litigation and compliance in practice. A wider view, adopting direct discrimination as an additional lens through which to understand algorithmic discrimination, would have serious practical reverberations, as a range of existing practices could no longer be legally deployed. More fundamentally, it opens up opportunities in both computer science and legal research to question existing taxonomies and justificatory regimes in working towards a coherent understanding of the legal treatment of machine bias.

ACKNOWLEDGMENTS

Adams-Prassl and Kelly-Lyth gratefully acknowledge funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement No 947806). Reuben Binns is supported by the Oxford Martin School EWADA Programme.

REFERENCES

[1] Adams-Prassl, Jeremias, Reuben Binns, and Aislinn Kelly-Lyth. "Directly discriminatory algorithms." The Modern Law Review 86.1 (2022): 144-175.
[2] Barocas, Solon and Andrew Selbst. 2016. Big Data's Disparate Impact. California Law Review, 104, 671-732.
[3] Bailey and Norbury(eds). 2020. Bennion on Statutory Interpretation. (8th ed). LexisNexis.
[4] Bower, Amanda, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alexander Vargas, and Suresh Venkatasubramanian. "Fair Pipelines." arXiv e-prints (2017): arXiv-1707
[5] Bright, Liam Kofi, Daniel Malinsky, and Morgan Thompson. 2016. Causally interpreting intersectionality theory. Philosophy of Science, 83(1), 60-81
[6] Buolamwini, Joy and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of Machine Learning Research, 81, 1-15
[7] Campbell, Colin, and Dale Smith. 2020. The Grounding Requirement for Direct Discrimination. The Law Quarterly Review, 136, 285-283
[8] Collins, Hugh, and Tarunabh Khaitan. 2018. Indirect Discrimination Law: Controversies and Critical Questions. In Collins and Khaitan (Eds.), Foundations of Indirect Discrimination Law. Bloomsbury Publishing
[9] Dwork, Cynthia, *et al.*. 2012. Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference, 214-226.
[10] European Union Agency for Fundamental Rights ('FRA') and Council of Europe. 2018. Handbook on European non-discrimination law. Luxembourg: Publications Office of the European Union.
[11] Feldman, Michael, *et al.* 2015. Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 259-268.
[12] Fredman. 2018. Direct and Indirect Discrimination: Is There Still a Divide? In Collins and Khaitan (Eds.), Foundations of Indirect Discrimination Law. Bloomsbury Publishing.
[13] Hacker, Philipp. 2018. Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. Common Market Law Review, 55(4), 1143–1185.
[14] Kasirzadeh, Atoosa, and Andrew Smart. 2021. The use and misuse of counterfactuals in ethical machine learning. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 228-236.
[15] Kleinberg, Jon, *et al.* 2018. Discrimination in the Age of Algorithms. Journal of Legal Analysis, 10, 113-174.
[16] Hu, Lily and Issa Kohler-Hausmann. 2020. What's Sex Got To Do With Fair Machine Learning? In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 513.
[17] Kilbertus, Niki, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. "Avoiding discrimination through causal reasoning." Advances in neural information processing systems 30 (2017).
[18] Kusner, Matt J., *et al.* 2017. Counterfactual fairness. arXiv preprint arXiv:1703.06856 (2017).
[19] Lipton, Zachary, Julian McAuley and Alexandra Chouldechova. 2018. Does mitigating ML's impact disparity require treatment disparity? Advances in neural information processing systems 31.
[20] Malinsky, Daniel. "Intervening on structure." Synthese 195, no. 5 (2018): 2295-2312.
[21] Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. "Algorithmic fairness: Choices, assumptions, and definitions." Annual Review of Statistics and Its Application 8 (2021): 141-163.
[22] Sánchez-Monedero, Javier, Lina Dencik, and Lilian Edwards. "What does it mean to'solve'the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems." Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020.
[23] Singh, Harvineet, *et al.* 2021. Fairness violations and mitigation under covariate shift'. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 3-13.
[24] Stark, Luke, and Jevan Hutson. 2021. Physiognomic Artificial Intelligence. Available at SSRN 3927300 (2021).
[25] Steel, Sandy. 2015. Proof of Causation in Tort Law. Cambridge University Press.
[26] Tschantz, Michael Carl. "What is Proxy Discrimination?." 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022.
[27] Wachter, Sandra. 2020. Affinity Profiling and Discrimination by Association in Online Behavioral Advertising. Berkeley Technology Law Journal, 35, 367-430.
[28] Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI." Computer Law & Security Review 41 (2021): 105567.
[29] Xenidis, Raphaele and Linda Senden. 2020. EU Non-Discrimination Law in the Era of Artificial Intelligence: Mapping the Challenges of Algorithmic Discrimination. In Ulf Bernitz *et al* (Eds.), General Principles of EU Law and the EU Digital Order. Wolters Kluwer.
[30] Yang, Ke, Joshua R. Loftus, and Julia Stoyanovich. 2020. Causal intersectionality for fair ranking. arXiv preprint arXiv:2006.08688 (2020)
[31] y Arcas, Blaise Agüera, Margaret Mitchell, and Alexander Todorov. "Physiognomy's new clothes." Medium (6 May 2017), online:< https://medium. com/@blaisea/physiognomys-new-clothesf2d4b59fdd6a 52 (2017).
[32] Zafar, Muhammad Bilal, *et al.* "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment." Proceedings of the 26th international conference on world wide web. 2017
[33] Patmalniece v Secretary of State for Work and Pensions [2011]UKSC 11, [2011]WLR 783
[34] R (on the application of E) v JFS Governing Body [2009]UKSC 15, [2010]2 AC 728.
[35] Joined Cases C-804/18 and C-341/19 IX v WABE eV and MH Müller Handels GmbH v MJ ECLI:EU:C:2021:594
[36] James v Eastleigh BC [1990]2 AC 751, [1990]3 WLR 55
[37] Nagarajan v London Regional Transport and others [2000]1 AC 501, [1999]3 WLR 425
[38] O'Neill v Governors of St Thomas More Roman Catholic Voluntary Aided Upper School [1997 ICR 33, [1996]IRLR 372.
[39] Reynolds v CLFIS Ltd [2015]EWCA Civ 439, [2015]ICR 1010
[40] Case C-267/06 Tadao Maruko v Versorgungsanstalt der deutschen Bühnen ECLI:EU:C:2008:179.
[41] Kentucky Retirement Systems v EEOC 554 US 135 (2008)