DISCRIMINATION FOR THE SAKE OF FAIRNESS

FAIRNESS BY DESIGN AND ITS LEGAL FRAMEWORK

Holly Hoch *(University of St. Gallen)*; Corinna Hertweck *(University of Zurich/Zurich University of Applied Sciences)*; Michele Loi *(University of Zurich)*; Aurelia Tamò-Larrieux *(University of Zurich/University of St. Gallen)*

ABSTRACT

As algorithms are increasingly enlisted to make critical determinations about human actors, the more frequently we see these algorithms appear in sensational headlines crying foul on discrimination. There is broad consensus among computer scientists working on this issue that such discrimination can only be avoided by intentionally collecting and consciously using sensitive information about demographic features like sex, gender, race, religion etc. Companies implementing such algorithms might, however, be wary of allowing algorithms access to such data as they fear legal repercussions, as the promoted standard has been to omit protected attributes, otherwise dubbed "fairness through unawareness". This paper asks whether such wariness is justified in light of EU data protection and anti-discrimination laws. In order to answer this question, we introduce a specific case and analyze how EU law might apply when an algorithm accesses sensitive information to make fairer predictions. We review whether such measures constitute discrimination, and for who, arriving at different conclusions based on how we define the harm of discrimination and the groups we compare. Finding that several legal claims could arise regarding the use of sensitive information, we ultimately conclude that the proffered fairness measures would be considered a positive (or affirmative) action under EU law. As such, the appropriate use of sensitive information in order to increase the fairness of an algorithm is a positive action, and not per se prohibited by EU law.

1. INTRODUCTION

Technical advances have propelled the shift from human decision-making to automated decision-making. With such advances, questions arise as to whether automated decision-making is fair in its outcomes. While social scientists have started to analyze the perceptions of individuals regarding whether automated decision-making is fair overall (or even fairer than human decision-making) (Helberger et al., 2020), recent biased decision-making systems have been flagged by data protection authorities (DPA) and the media. An example is the automated means of student grading that had occurred due to the COVID-19 pandemic which relied on historical data. In Norway, the DPA has called out such practices to be unfair according to data protection law (Norwegian DPA, 2020) while newspapers, such as the Guardian, have reported the discriminatory

impact of such practices (Naughton, 2020). Another often discussed example comes from the US courts relying on a decision support tool called Correction Offender Management Profiling for Alternative Sanction (or short COMPAS), which inhibited racial biases (Angwin and Larson, 2016).

Unfair and potentially discriminatory decisions made by algorithms put philosophers, legal scholars, and computer scientists at the forefront of an interesting problem: Is it computationally possible to teach an algorithm to assist human decisions compliant with the moral and legal requirement of avoiding wrongful discrimination?

One of the clearest discoveries stemming from the study of this problem is that most decisions with a basis in statistics *cannot avoid all forms* of discrimination. In most real-world circumstances, a decision rule based on statistical evidence will exhibit some form of discrimination. Thus, the attention and the language regarding ethics in the machine learning community has shifted recently from discrimination, here meaning prejudicial inequality, to fairness, whereby the concept of fairness is used to refer to *both* (i) the absence of certain types of biases and prejudices, and, in some approaches, (ii) to a justifiable distribution of *benefits and burdens* resulting from such decisions. Of course, both the concept of fairness and non-discrimination are closely tied, as also apparent from the legal literature. While the legal and philosophical literature has started to discuss what fairness and non-discrimination mean with respect to automated decision-making practices (Wachter et al., 2020; Bogen et al., 2019; Barocas and Selbst, 2016; Bent, 2020; Pasquale, 2015; Chandler, 2017), computer scientists have elaborated on a plurality of definitions of what fairness means, which cannot be all simultaneously achieved (Kleinberg et al., 2017).

In this article, we focus on fairness from a computer science perspective, as elaborating on the philosophical and legal concept of fairness would be beyond what is feasible in this contribution. We assume for the sake of the argument that one definition of fairness used in machine learning is correct, namely, *equality of opportunity*. Enforcing equality of opportunity requires considering the protected class (e.g. sex, race, religion) of the individual about whom a decision will be made (see, e.g. Hardt, Price and Srebro, 2016).[1]

[1] More precisely, in the technical literature one finds fairness-based technical arguments in favor of using information about the protected class not only in training the models (Zliobaite and Custers, 2016), but also in decision making (e.g. as a trigger that determines that a different classification threshold, or a different model, should be used, for different groups) (Hardt et al., 2016). A workaround apparently avoiding the present clash of fairness and anti-discrimination law would be to achieve equality of opportunity without collecting group variables from the individuals about whom a decision must be made. In order to achieve equality of opportunity between two protected groups by design (and not by fluke), this means that the training algorithm would have to be allowed to learn proxies of the protected groups. This strategy, however, faces the following objections. First, to the extent that it is a mere workaround for a strategy of achieving fairness defined in terms of group membership, the strategy is vulnerable to substantially similar legal objections. We would face the question whether it is compatible with anti-discrimination law to design a system that, intentionally or even merely predictably, exhibits such behavior which in a sense is a disguised attempt to use protected information in processing. Arguably this question can only be answered after we clear the ground on the logically prior question whether using such information directly, without involving proxies, is

What follows in this article is an analysis of the ethical and legal discourse on the matter (with the legal discourse focusing on European law), tying in both perspectives to provide a complementary picture. In this interdisciplinary article we *ask whether using protected class information, in the context of algorithmic decision making, for the sake of achieving (what we assume, for argument's sake, to be) fairer outcomes violates anti-discrimination law*. To answer this question we first answer one of the main questions of this paper, namely does European data protection law and anti-discrimination law restrict the use of sensitive data categories such as sex (i.e., protected classes of information) for the purpose of achieving fairness, or does this qualify as a circumstance under which its processing is allowed? We show that this may be legitimate if processing of this data is required for the sake of ensuring the *fairness* of the processing. We then turn to the fairness of the processing itself. To do so, we provide the relevant definitions from anti-discrimination law (Sections 2 & 4) and algorithmic fairness (Section 3). We apply these definitions to the analysis of a hypothetical case of an algorithm which implements the fairness constraint commonly found in the machine learning literature, namely Hardt et al.'s (2016) equality of opportunity. We further analyze the case presented in Section 3 providing arguments pointing to different answers to the question whether this algorithm discriminates based on sex, and, if so, against which group, and why. Finally, in Section 4, we propose our *ultima facie* interpretation of the case, involving a balance of these different perspectives.

2. LEGAL FRAMEWORK

2.1. USE OF SPECIAL CATEGORIES (I.E. PROTECTED FEATURES) OF DATA

When engineers develop algorithms for creating fairer outcomes they need to use data. Often such data will be considered personal data and even sensitive personal data (or protected features) and thus will fall under the scope of the General Data Protection Regulation (GDPR). Under the GDPR the processing of personal data revealing "racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation" (Art. 9(1) GDPR) is as a rule prohibited, yet may be processed if exhaustive legal grounds set out by the law are met (Art. 9(2) GDPR) and the fundamental principles (Art .5 GDPR) adhered to.

legally permissible, so the question we ask in the paper appears more urgent to us. Moreover, some machine learning scholars have also argued that using explicit protected group information for achieving fairness is preferable - fairness-wise - to designing an algorithm to automatically recognize proxies of such information in order to achieve the same fairness goal, e.g. in that it avoids some unwanted further distortions that appear unfair from a distinct standpoint (Lipton, Chouldechova and McAuley, 2018).

2.1.1. LEGAL GROUNDS UNDER DATA PROTECTION LAW AND POSITIVE ACTION DEVELOPMENTS

One key legal ground for the processing of protected features is explicit consent. Explicit consent can be given for one or more specified purposes prior to the data processing. As the term indicates, explicit consent cannot only be implied but must include an affirming action by the person whose data is being processed for the sake of creating fairer algorithms. However, consenting to the processing of sensitive data in order to allow for positive action may only occur in cases where information symmetries exist. In other words, if public authorities are processing sensitive data, relying on explicit consent will not be an appropriate legal ground because of the information and more importantly power asymmetries at hand (Frenzel, 2018). With respect to the processing of public authorities but also employers, they will typically (try to) rely on a legal obligation to process sensitive data. Examples of legal obligations are, in the context of employment, often labor or social benefit laws. Those laws mandate that an employer processes specific protected features for certain, specified purposes. To fall under this legal ground there must be a formal law in a member state that includes a direct obligation or right to process said data. This legal ground could become interesting if a Human Resource (HR) department uses algorithms to ensure equal opportunity.

Aside from consent or statutory obligations to process sensitive data, the processing could also be legitimized because it lies in the public interest. This ground is an opening clause that allows member states to enact their own legislation specifying the legal ground. A significant public interest includes not only averting dangers to life (physical integrity), but also dangers that would otherwise lead to irreversible damage including damages to the integrity of the legal system. A substantial public interest is given if there is a threat to the physical well-being of individuals or if other irreversible damages could occur to individuals (Frenzel, 2018). It could be argued that the allowances of positive actions within the EU (described further in Section 4) - which parallels the US system of affirmative action - could fall under the category of a public interest. The idea behind positive action is to remedy past and prevent future injustices, thus while the issue has not been adjudicated yet, attempting to achieve greater fairness is a possible public interest for a greater allowance of data processing.

Depending on the context and the extent of the processing different legal grounds may be claimed by the data controller, i.e., the entity or person determining the scope and purpose of the data processing. While it cannot be stated that all processing of sensitive data for the sake of fairness will always be covered under a legal ground, it is likely that such processing can be argued to fall under a legal ground if in line with the processing principles described below.

2.1.2. COMPLIANCE WITH DATA PROTECTION PRINCIPLES

Aside from the legal grounds, the processing of sensitive data still needs to comply with the data protection principles. In other words, even if explicit consent is given, the principles of fairness, transparency, purpose limitation, data minimization, accuracy, storage limitation, and security must be adhered to (Art. 5 GDPR). Importantly, the processing must be lawful, transparent and

fair (Art. 5(1)(a) GDPR). *Lawful* can be understood narrowly or broadly: Narrowly, lawfulness can be fulfilled by demonstrating the compliance with a legal ground (see above), more broadly, lawfulness requires compliance with all laws, including anti-discrimination law (see below). The principle of *transparency* requires data controllers to provide specified information to data subjects when collecting sensitive data (Art. 12-14 GDPR) in an easily accessible and understandable manner. These information requirements include providing information on the purposes of processing (e.g., the purpose of making algorithms fairer). Linked to both, the principle of lawfulness and the principle of transparency, is the principle of *fairness*.

On the one hand, the principle of fairness is linked to the principle of lawfulness as fairness must also be understood as procedural fairness where the data processing abides by the procedural rules set out by data protection law (Malgieri, 2020). On the other hand, the principle of fairness is also inherently linked to the concept of transparency, as assessing fairness involves not only analyzing the (1) effects of the processing on an individual or group of individuals but also understanding their (2) expectations, which in turn is linked to how transparent the purposes of the processing were made in the first place (ICO, 2017). The latter criteria requires a combination of transparency practices and limiting the purpose of processing protected categories of data. For instance, in the mentioned Norwegian grading case (see Section 1 Introduction) one key argument against awarding grades based on inferential analysis was that it did not adhere to the reasonable expectations of the students (Norwegian DPA, 2020)[2]. The former criteria looks in particular at whether unwelcomed and unplanned negative impacts are the result of the processing. While not all unwelcomed effects are necessarily unfair, discriminatory effects are considered unfair (WP29, WP 251, 2018; CNIL, 2018).

2.2. CONCEPTS OF DISCRIMINATION

In legal terms, 'discrimination' refers to an adverse act committed against a legally protected individual or group. Under EU law, the primary types of discrimination are direct and indirect. In brief, direct discrimination is defined by adverse treatment based on a protected characteristic, (sexual orientation, religion, etc.) while indirect discrimination illustrates more underlying social issues, describing a situation in which an "apparently neutral provision, criterion or practice" impacts a protected group disproportionately compared to others (*all EU Non-Discrimination Directives include*). Though direct discrimination is generally more obvious and therefore more

[2] Wording: "Presumably, students would expect that a grade is awarded based on their demonstrable academic achievements and that the grade would reflect the work they have put in as well as the knowledge and skills they have attained. Even though exams were cancelled this year, we do not see how the current situation would alter their expectations in that regard. This expectation is reasonable as grades will be used by educational institutions as a measurement of the individual student's academic level in the context of admissions to higher education. Potential employers will likely regard grades in the same way. Through the use of "school context" and "historical data", the IBO has utilised information concerning how other students at the same school have performed historically to adjust the final grades. This means that the IBO expects students to foresee and accept that their grade is influenced by a factor completely outside their control, which does not have any connection to their demonstrable and individual academic achievements, and that relates to other people. This assumption is unjustified" (cf. https://www.datatilsynet.no/en/news/2020/the-norwegian-dpa-intends-to-order-rectification-of-ib-grades/)

commonly brought forth in legal claims, the continued and rising use of algorithms already has and likely will continue to develop a number of indirect discrimination claims as a reflection of societal values and decision making-processes, for example in risk modelling (see Bartlett, 2019).

Discrimination always implies disadvantage for one group: the group discriminated against. A first step in bringing a claim is to define the group that has been disadvantaged. In direct discrimination, this is a simple task: the rule, practice, or action alleged to be discriminatory must explicitly refer to a protected characteristic. For indirect discrimination, defining the disadvantaged group is more complicated: an "apparently neutral provision, criterion or practice" must be shown to significantly disadvantage a legally protected group, while the offending behavior simultaneously does not explicitly address the disadvantaged group. Protected groups, or "classes", are statutorily created. (Art. 2(b) of Council Directive 2000/43/EC (Racial Equality Directive)). For example, Article 21 of the EU Charter of Fundamental Rights establishes that "[a]ny discrimination based on any ground such as sex, race, color, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited." National laws may also further explicitly define protected classes, leading to varied approaches and protections across the EU.

However, proof of discrimination is not always a coup de grâce for the accused, as there are instances of "allowable" discrimination. Direct discrimination must be explicitly legally allowed, for example on the basis of a "genuine occupational requirement', whereas indirect discrimination can be justified if a legitimate aim is pursued and the measure is necessary and proportionate (for example, see *Kalanke,*1995; *Marschall*, 1997; and *Abrahamsson*, 2000). Discussed further in Section 4, these concepts of "allowable" discrimination based on legitimate pursued aims and mitigating factors become particularly critical when analysing situations in which an algorithm is trained or designed to intentionally discriminate based on protected class information for the purpose of achieving greater fairness for the disadvantaged class(es).

Most importantly, when bringing forth a claim for discrimination there must be a harm suffered, the disadvantage. Harm is defined as "either less favorable treatment (direct discrimination), particular disadvantage (indirect discrimination) or the violation of the dignity of a person (harassment and segregation)" (Farkas and O'Farrell, 2015, 41). Primarily, anti-discrimination law is concerned with the interference or violation of one's rights. In terms of damage, this harm may be material, (for example, monetary), or immaterial (emotional distress) (Farkas and O'Farrell, 2015; *Prosecutor v. Ferdinand Nahimana, Jean-Bosco Barayagwiza and Hassan Ngeze*). In order to bring forth a discrimination claim direct or indirect, a claimant must meet three key evidential requirements to establish a prima facie case: (1) a particular harm has or is likely to occur; (2) the harm impacts or is likely to impact a protected group and, finally, (3) the harm has a disproportionately (negative) impact on the protected group when compared with another group in a similar situation ("comparators") (Farkas and O'Farrell, 2015, Handbook, 2018).

In order to identify discrimination, courts refer to a comparator to assess whether other persons or groups in a similar situation have experienced the harm in question, and to what effect (Recast Directive, Art. 2(1)(a); Race and Framework Directive, Art. 2(2)(a)). A comparator consists of a person or group in a comparably similar circumstance, with the main differentiator being the absence of the protected ground from one group. Defining a comparator may be difficult in real life situations and hypothetical comparators may be used if one does not otherwise exist. For these hypotheticals the key factor lies in creating a comparator that factors out the protected ground(s) invoked. (Epstein and Masters, 2011). For example, in the case we analyze in this paper the protected class is "women" and the comparator "men". This is a straightforward example used to illustrate the larger thematic issues, however, it is not only the protected class aspect that comes into question, the comparator must be in a similar position as the individual(s) alleging mistreatment. Depending on the claim and qualities of the parties this could mean similar education, skills, financial situation, or another similarity that shows the discrimination was based on the protected characteristic and not another reason.

3. FAIRNESS BY DESIGN

3.1. FAIRNESS AS EQUALITY OF OPPORTUNITY

Who gets treated with a potentially life-saving new drug, who should be released from prison or who is granted a loan are all questions that are now frequently answered by algorithms. Such algorithms can follow simple, fixed rules. A stock trading algorithm, for example, may sell or buy stocks once their price falls below a specific value. However, algorithms in practical usage are typically far more complex. Instead of the rules being explicitly encoded, the algorithm learns its own rules from data. This is what is referred to as *machine learning*. In credit lending, for example, the algorithm may have access to a bank's credit lending history for the past ten years, including the credit history, income, recurring debts, account balance, assets etc. of their loan-holding customers. Data regarding whether customers have defaulted or paid back their previous loans is also provided. The algorithm's task is to learn to differentiate between the customers who pay back their loan and those who default, based on the financial data and histories provided. The resulting machine learning model could then be employed to predict how likely a new credit applicant is to pay back their loan based on the given data (credit history, income etc.).

This data used to train a machine learning algorithm, however, might also include protected attributes, such as age, sex or race. As the usage of these features appears to infringe on anti-discrimination law, the first impulse for creating a "fair" machine learning algorithm might be the *anti-classification* approach (also known as "fairness through unawareness"): disregarding any protected attributes. However, this cannot guarantee that the algorithm will not discriminate based on protected attributes - it might simply do so through *proxy variables* instead, as protected attributes are oftentimes redundantly encoded in the other data. Thus, even if the protected attribute

is left out of the dataset, if enough other data is available, it is likely that the protected attribute can be inferred (Datta et al., 2017; Dwork et al., 2011; Wachter et al., 2019). Consequently, an algorithm not having access to protected attributes might still discriminate against these protected attributes through proxy variables.

As this anti-classification approach has been shown to be insufficient, the mitigation of algorithmic biases has become an increasingly important topic in the machine learning literature in recent years. The current consensus among both machine learning and legal scholars is that ignoring protected attributes is often not sufficient in order to ensure fairness (Corbett-Davies and Goel, 2018; Kleinberg et al. 2018; Žliobaitė and Custers, 2016). In fact, data protection law that regulates the processing of protected attributes (referred to as sensitive personal data) provides legal grounds for the use of said data to ensure fairer processes (see above Section 2). What fairness amounts to, however, is less clear. Some researchers argue that there is no single correct fairness measure, but that fairness is highly *contextual* (Friedler et al., 2016; Hardt et al., 2016; Wachter et al., 2019) and "*task-specific*" (Dwork et al., 2011).

One commonly cited fairness measure is that of *equality of opportunity* as proposed by Hardt et al. (2016). Going forward, our legal analysis will focus on this fairness measure and discuss the proposal from a legal perspective. (We shall always write "equality of opportunity" to mean the specific conception of it by Hardt, Price and Srebro and not the more general concept in political thought.) To explain the measure, let us consider the example of credit lending, in which a machine learning algorithm tries to predict whether a person applying for a loan should be granted the loan. The basic idea of the equality of opportunity measure is that *every person who would pay back their loan and would thus deserve to be given a loan has the same probability of obtaining the loan* - independent of any, possibly protected, group membership. If we consider two distinct demographic groups, such as female applicants and male applicants, this means that the same share of women who would pay back their loan and men who would pay back their loan are given a loan. If, for example, 200 women apply who would pay back their loan and 100 of them are granted their loan, then of 100 male applicants who would pay back their loan, about 50 should receive their loan. This measure, i.e., how likely a non-defaulting person is to receive a loan, is referred to as the *true positive rate*. "Equality of opportunity" is therefore sometimes referred to as "*true positive rate parity*".

This type of fairness is unlikely to be achieved without any intervention (Hardt et al., 2016; Liu et al., 2019). In credit lending, the predictor would typically deliver a score based on which the bank then decides whether a loan should be granted or not. The anti-classification approach would set a single threshold. Every person - regardless of their sex - whose score is at least as high as the threshold gets the loan, every person below the threshold is denied the loan. This threshold could be chosen to maximize profits: The lower the score, the more likely people will default on their loan, but the higher the score, the more profitable credits the bank then misses out on by eliminating loan-paying candidates. It is thus the bank's job to find the trade-off that maximizes their profit.

However, when we pick a single threshold, we cannot guarantee that the true positive rates are equal across groups (Hardt et al., 2016). Generally speaking, the lower the threshold is, the more people receive a loan. Consequently, the true positive rate increases as more people who would pay back their loan actually receive a loan. The higher the threshold is, the fewer people will receive a loan. Consequently, the true positive rate decreases as more people who would pay back their loan do not get a loan. While these trends hold across groups, the true positive rates do not change at the same speed across groups. In order to achieve equality of opportunity, Hardt et al. argue for a different threshold for each group.

So how do we find the thresholds that lead to the same true positive rate for each group? For the sake of this example, let us assume that we can achieve every possible true positive rate (between 0 and 100%) simply by changing the threshold. Let us assume that this is also true when we consider the demographic groups individually, meaning that for each demographic group, there is a threshold that allows us to achieve a given true positive rate. The simplest way to find this threshold would be to try every possible threshold and observe the resulting true positive rates. As already discussed, the threshold to achieve a given true positive rate might be different for every group. *Achieving true positive rate parity, i.e., equality of opportunity, might thus mean setting different thresholds for each group*.

If we can achieve parity for every true positive rate, the only question left is what true positive rate we should pick. To answer this question, we calculate the profit associated with every true positive rate. This is done in a similar fashion as when selecting a single threshold: We simply have to weigh the cost of granting defaulting loans against the loss of denying profitable credits. Finally, we pick the true positive rate which delivers the highest profit. This way we end up with (typically different) thresholds for each group with maximum profit subject to the constraint guaranteeing equal opportunity. We will refer to this changed decision making process as the equality of opportunity algorithm, or "*EQOP algorithm*" for short.

3.2. PRIMA FACIE INTERPRETATIONS OF EQUALITY OF OPPORTUNITY IN TERMS OF DISCRIMINATION

In this section, we discuss a single case that utilizes the EQOP algorithm to achieve equality of opportunity and view the results through various lenses with respect to the legal and ethical ramifications for the two groups.

We consider the example presented here as a fairly common situation occurring in practice. Assume that a bank trains a model whose aim is to predict how likely a loan applicant is to pay back a loan they have applied for. Based on these criteria and the ensuing score of probability, it then has to be decided who receives a loan. To automate this process, the bank implements a threshold. Every applicant whose repayment-probability is higher than this threshold receives their loan. The question then is: What should the threshold be? Should an applicant who is more likely to repay than to default (i.e. has a repayment probability of more than 50%) receive a loan?

Probably not: Naturally, the bank wants to maximize their profits and since a defaulting loan costs more than a repaid loan benefits the bank, they probably want to be more cautious and only grant loans with notably higher repayment probabilities. They thus calculate the expected win and loss for each threshold and pick the one that maximizes their profit. The bank does not want to (and legally is not allowed to) discriminate based on sex, so they neither consider sex in the training of the model nor when setting the threshold even though it might be more profitable to do so. This is the anti-classification strategy previously mentioned.

Suppose that this model is less likely to give loans to women. This is an instance of indirect discrimination as sex information is not used, so the credit lending algorithm appears to be neutral, but ends up disproportionally affecting women. Several reasons are imaginable for this. For our example, we will assume that the women in the original dataset are less likely to repay their loans than the men in the dataset (e.g. 50% of the women pay back their loan, but 60% of the men do), so it is not surprising that the algorithm will give a lower score to the average woman than to the average man. Simply because the rates are different in the original dataset, the model will likely show such differences, too, and even amplify them. This is the case because even though the model does not consider sex, it picks up on the differences between the sexes through proxy variables, such as the hours worked per week. If women work fewer hours, the algorithm may find this correlation to be helpful to distinguish between defaulting and repaying applicants. When this correlation is exploited, existing differences may be amplified. This amplification is what we see in Figure 1: Men have a higher probability than women of paying back their loan, but the difference between the two groups is greater for the probability predictions.

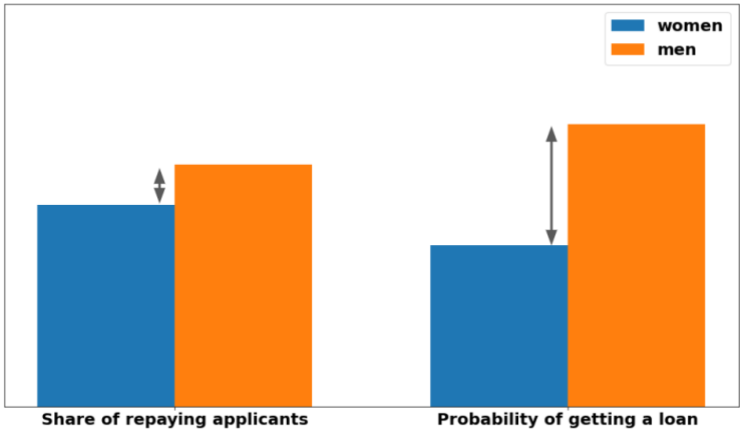

Figure 1: Comparison of the share of repaying applicants (women: 50%, men: 60%) to the probability of getting a loan (women: 40%, men: 70%) under the anti-classification approach. The difference between the probability of getting a loan between women and men is bigger than the difference between the share of repaying applicants. Note: the provided figure numbers throughout the text are purely illustrative for the given hypotheticals and not based on a particular dataset.

Let us now suppose that the software is audited by an independent non-governmental organization, which is able to demonstrate the sex disparity in lending rates by generating a large number of fake loan applications and collecting the bank's automated responses. As the bank sees this as problematic for their reputation, they consult an AI ethicist to help find a solution to the problem. By studying the predictions produced by the algorithm on historical data, they are able to not only reproduce the finding that women are generally less likely to receive a loan, but find that this disparity persists even when looking at only the qualified applicants: Compared to men who repay their loans, women who repay their loans are less likely to be predicted to repay. Again, let us assume for the sake of our example, that the ethicist's advice is to equalize this latter metric by implementing Hardt et al.'s equality of opportunity constraint, which - so the ethicist argues - adequately captures what fairness requires in such and similar cases. Currently, the loan approval rate of men who repay their loan is not equal to that of women who repay their loan. Instead, almost all men who repay receive a loan while far fewer women who repay receive one. If the credit scoring function should stay the same, the threshold must be changed. As the bank's data scientist clarifies, this requires the bank to pick group-specific thresholds: One for men and one for women. Again, use of an EQOP algorithm, which the bank now uses, picks the most profitable setting among the thresholds that fulfill the equality of opportunity measure. As it turns out, through this change, the threshold has become higher for men, but lower for women. Women-who-repay and men-who-repay are now accepted at the same rate. While women generally are still accepted at a lower rate, the difference in the approval rates is not as stark as with the single anti-classification threshold.

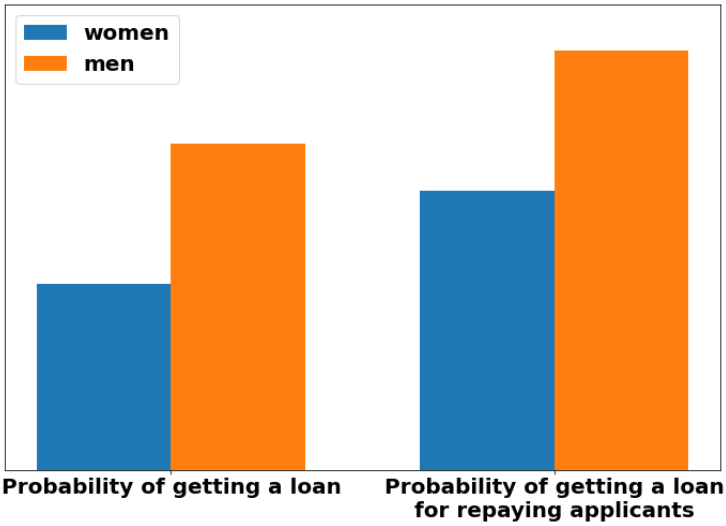

Figure 2: Comparison of the probability of getting a loan (women: 40%, men: 70%) to the probability of getting a loan as a repaying applicant (women: 60%, men: 90%) under the anti-classification approach. Women are less likely to receive a loan both in general and when only looking at repaying applicants.

We will now explore how different interpretations of this fairness-by-design scenario could be analyzed in pursuit of a legal anti-discrimination claim. The list of interpretations is neither exhaustive nor do all these interpretations pertain to every imaginable perspective. The goal is to demonstrate that the usage of protected attributes in machine learning can be viewed from different perspectives, which give rise to varying legal questions.

In order to make it easier for the reader to follow the different arguments and the assumptions they rely on, we summarize these relations in Table 1, below.

Table 1. Normative assumptions about harm and comparators and their implications for the assessment of direct discrimination by EQOP

Harm of discrimination defined as:	**Comparison group defined as:**	**Has a group potentially suffered from discrimination?**	**Location in essay**
Being denied a loan	Women vs. men (asking loans)	Yes, women (indirect)	3.2.1 - *1st argument*
Being denied a loan	Non defaulting men vs. non-defaulting women	No	3.2.1 - *2nd argument*
Being denied a loan one would have obtained if sex information had been ignored	Women vs. men (asking loans)	Yes, men (direct)	3.2.2 - *1st argument*
Being denied a loan	Women vs. men with the same risk score	Yes, men (direct)	3.2.2 - *2nd argument*

3.2.1. HOW DOES EQOP IMPACT WOMEN?

In this section, we consider the question of whether EQOP constitutes discrimination against women. By looking at the situation from two angles, we first make the claim that first glance, the case seems to constitute indirect discrimination against women. Secondly, we argue that EQOP is actually not discriminatory, neither against women nor men. The two arguments are lined out in Figure 3 below.

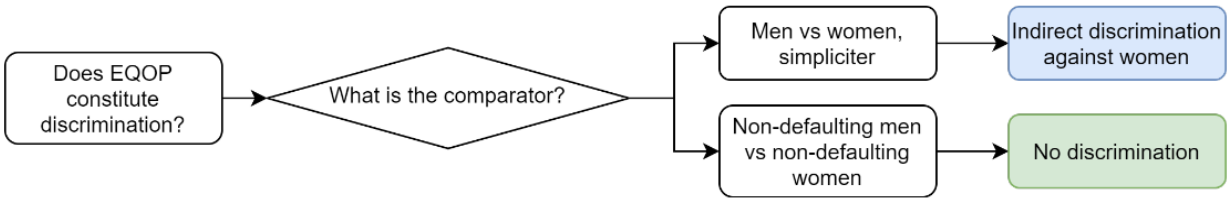

Figure 3: Graphical representation of the arguments in 3.2.1. Whether the EQOP approach may be considered as discriminatory depends on the comparators.

Let us first begin with the first legal question: namely, does EQOP discriminate directly against women? By our legal definition, direct discrimination occurs when:

(1) a particular harm has or is likely to occur impacting a protected group;

(2) the harm impacts a protected group *because of* their protected characteristic, which amounts to showing that the harm having a disproportionately (negative) impact on the protected group can be evidenced by comparing with another group in a similar situation ("comparator"), which differs by the protected characteristic.

It turns out that (1) is satisfied because women disproportionately being denied a loan is a form of separate treatment. What is critical is the reasoning and evidence basis behind the claim in (2), the claim that the harm occurs because of sex or more precisely (since an applicant's sex does not cause any decision *alone*) because of a causal disposition of the algorithm to treat a client less favorably than a *relevant comparator* of a different sex. This, as it turns out, depends critically on how one interprets the causal role of sex information in the algorithm and of how the comparator class is defined.

As hypothesized, EQOP still lends more frequently to men than women with all test data, even if the use of the two thresholds makes this effect less pronounced than the initial ("biased") algorithm used by the bank (see Figure 4). If one asks whether the algorithm is disposed to treat a client less favorable than a relevant comparator of the different sex, and "a relevant comparator" is defined simply as another client asking for a loan (i.e. without specifying any other feature of this client), then one must conclude that EQOP treats women less favorably than men, so women are discriminated against. Remember that, in our stipulation, EQOP mitigates the inequality of loan acceptance by sex somewhat (not by design, but as a collateral effect of achieving equality of opportunity), but not entirely.[3] *Prima facie*, it thus seems like we ought to answer (2) in the affirmative and conclude that EQOP discriminates against women applying for loans (relative to

[3] This is possible, for example, if women more frequently default than men. If so the rate of lending to *non-defaulting men and women* can be equal (as required by equality of opportunity), while lending propensity compared by reference to sex only (without looking at the distinction between defaulting and non-defaulting clients) will be unequal. Thus, this case cannot be considered a failure of the design approach adopted to achieve the intended fairness standard.

men) since EQOP, and that this discrimination is direct since EQOP uses intentionally uses sex information to decide a lending decision.

> *First argument:* EQOP amounts to direct discrimination against women because it uses sex information and women are less likely to receive a loan.

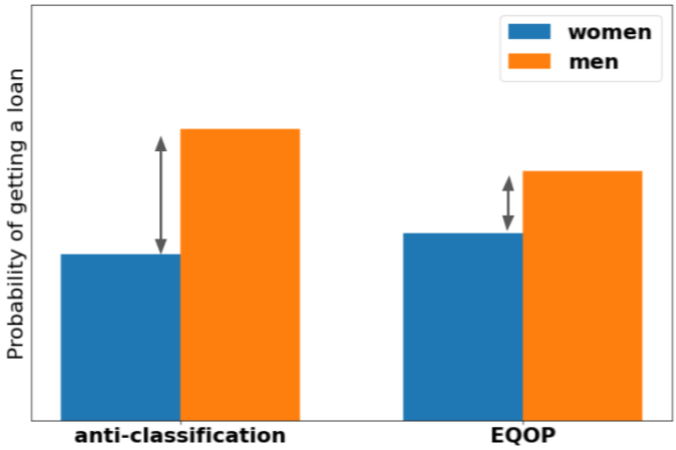

Figure 4: Comparison of the probability of getting a loan under the anti-classification (women: 40%, men: 70%) and EQOP approach (women: 45%, men: 60%). Under both approaches, women are on average less likely to receive a loan than men. However, the difference between the likelihood of receiving a loan between women and men is smaller under the EQOP approach.

However, *ultima facie*, the case for the direct discrimination against women does not hold if we consider, also independently from each other, two further elements: the *fine-grained* causal role of sex information, and a fairness-based definition of the salient comparison group. As we shall see, either issue can be analysed from a plurality of different perspectives, of which we discuss further in (Section 4).

First, we should consider not only the fact that sex information is used by EQOP, but how it uses it. As stated in (2), inequality does not only have to occur, but it has to occur *because of* a protected characteristic. Even though the EQOP algorithm utilizes sex information to determine sex-specific thresholds, this use is not worsening the inequality against women in our case. To see this, we must compare the inequality produced by the EQOP algorithm to the inequality that existed in the original algorithm, where sex information was ignored. This disparity has not been amplified, but instead reduced, by the EQOP algorithm. It is true that the algorithm still is more likely to lend to men than women, but the *net* effect of taking sex into account is to mitigate this inequality. The net effect of taking sex into account can be understood as the difference in the gap between men's and women's loan approval rates between the original model (designed to maximize utility, unconstrained) and the EQOP model (designed to maximize utility, subject to the EQOP fairness constraint). While the threshold is not per se sex-neutral, in that it was lowered for women, the effect of sex-information *per se* - i.e. the causal contribution of lowering of the threshold – is not to further disadvantage *women* relative to men. This justifies the rejection of the idea that EQOP

discriminates ***directly*** *against women.* Therefore, the disparity that we observe in the acceptance rate between men's and women's applications under the EQOP algorithm is thus still (we contend) - just like the anti-classification algorithm - a case of indirect discrimination against women and not a case of direct discrimination. (There still remains an argument, however, that EQOP discriminates directly *against men*, and we shall consider arguments in support of this conclusion.)

Second, even ignoring the previously specified causality argument, one may object that the above *prima facie* argument for direct discrimination *against women* fails to identify the *morally relevant* comparator, i.e. it does not compare women being denied loans to men "under similar circumstances", in the sense that is relevant for assessing the fairness of a leading decision. It may be maintained that in assessing whether women are discriminated against by EQOP one ought to compare women with men who are similar in the morally relevant sense. Clearly, *non-defaulting women* (women who repay their debts with the bank) do not deserve to be treated differently from *non-defaulting* men. However, it is not morally objectionable if an algorithm treats *defaulting* women differently from *non-defaulting* men. We can test EQOP with historical data, used as statistical test data, to measure how EQOP behaves with respect to clients of different sexes classified, in data science language, by their "true label", i.e. as defaulting or not. As evident from Figure 5, *non-defaulting women* and *non-defaulting men* have the same probability of receiving loans with the EQOP algorithm (whereas non-defaulting women had lower chances of getting a loan than non-defaulting men when judged by the original anti-classification algorithm). Thus, if one judges EQOP based on this comparison it turns out that criterion (2) for discrimination is not met: while the harm of being denied the loan affects both female and male non-defaulting clients (i.e. the predictions are not perfectly accurate), it affects non-defaulting women and men exactly to the same degree. Thus, EQOP does not have a disproportionately (negative) impact on women relative to men *in the (arguably) morally relevant similar circumstances*. Notice that this is not a coincidence in our hypothetical case: achieving this equality in probability is what EQOP is designed to achieve. The algorithm modifies the thresholds until frequencies of loans received by non-defaulting clients (in the test data) are the same on average for the two sex groups. Thus, if comparators are chosen appropriately, one ought to conclude that EQOP does not discriminate according to sex, either directly or indirectly.

This brings us to our second fairness claim:

> *Second argument:* EQOP does not amount to discrimination (neither indirect nor direct) because non-defaulting women and non-defaulting men have the same chances of getting a loan.

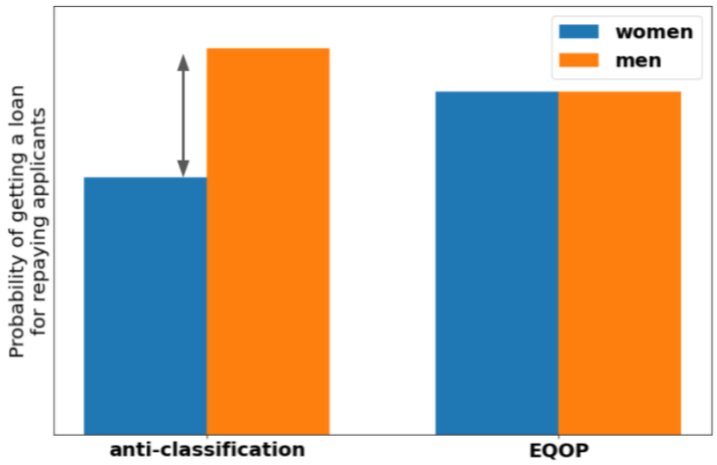

Figure 5: Comparison of the probability of getting a loan as a repaying applicant under the anti-classification (women: 60%, men: 90%) and EQOP approach (women: 80%, men: 80%). Under the anti-classification approach, repaying women are less likely to receive a loan than repaying men. Under the EQOP approach, repaying women and repaying men are equally likely to receive a loan.

To summarize, we can again turn to Figure 3, which shows how the choice of comparators impacts the verdict on whether EQOP constitutes discrimination or not. If judging harm requires identifying a comparator, then intentionally choosing equality of opportunity as the fairness view implemented by the algorithm amounts to a precise choice of a comparator. The fairness idea implemented by EQOP is to treat people *with the same actual favored outcome* in a (statistically speaking) similar manner, irrespective of their protected characteristic. This is analogous, for example, to requiring the same rate of favorable parole decisions for all prisoners who turn out to be non-recidivists, irrespective of their race.[4] If we all agreed that this is what fairness requires, we have a moral ground for choosing a relevant comparator of a different protected group. We should then discard as irrelevant the fact that the algorithm is favorably disposed towards lending to men *in general*, compared to women *in general*, who apply for a loan, and only consider the fact that it is *not* favorably disposed towards *either* sex when comparing *non-defaulting* clients for that type of loan. Whether this criterion is satisfied can be determined ex post, or before implementing EQOP if sufficient reliable historical data are available in the test phase.

3.2.2. DOES EQOP DISCRIMINATE DIRECTLY AGAINST MEN?

Let us return to the idea that considers men applying for a loan to be the pertinent comparison group to women applying for a loan. In Section 3.2.1, we have argued that under this interpretation of the comparison groups, EQOP is potentially guilty of direct discrimination. Moreover, we have argued that EQOP is, at first glance, guilty of direct discrimination against women, not against

[4] This was the fairness criterion implicitly adopted by ProPublica as a normative premise of its claim that the COMPAS algorithm exhibited "machine bias" against blacks.

men, because women are disproportionately likely to be negatively impacted (i.e. denied loans) when compared to men.

We shall now argue that this interpretation of direct discrimination may be turned on its head in the particular case we have hypothesized: EQOP is guilty of direct discrimination, but, in fact, the group discriminated against are men applying for a loan, not women applying for a loan.

We discuss two further *prima facie* arguments for this. Each argument considers different features of EQOP that are all instantiated by our imaginary equality of opportunity algorithm trained with the data and applied in the context of the hypothetical case at hand. As each of these features may occur independently from each other in a case where the algorithm has been trained with different data (and it must make decisions about different populations, e.g. in a society in which women are as rich as men on average), there may be hypothetical or real cases where not all the arguments highlighted here are relevant. When we write that these features are independent, we mean that it is not logically or mathematically necessary for them to be related in the way of a dilemma, as assumed for this case. This means that it is not a universal feature of all profit-maximizing algorithms that a threshold for women and men are respectively lower and higher than the anti-classification (sex-blind) threshold. This is a scenario we have imagined to occur, a feature of the hypothetical case under examination, as the story provided in the introduction clarifies. However, in our analysis we deal with the more interesting case in which equality of opportunity can *only* be satisfied with two distinct thresholds, where one threshold is lower and the other higher than the single anti-classification threshold.

The first argument, to which we now turn, makes the same assumption about the relevant *comparator group* as the first argument of 3.2.1, namely we still compare female and male clients *simpliciter*, as opposed to (what would be arguably fair in an equality of opportunity perspective) *non-defaulting* women with *non-defaulting* men.

> *First argument:* EQOP amounts to direct discrimination against men because if EQOP had (explicitly and intentionally) ignored sex information, some men denied loans by EQOP would have obtained loans, and some women given loans by EQOP would not have received them.

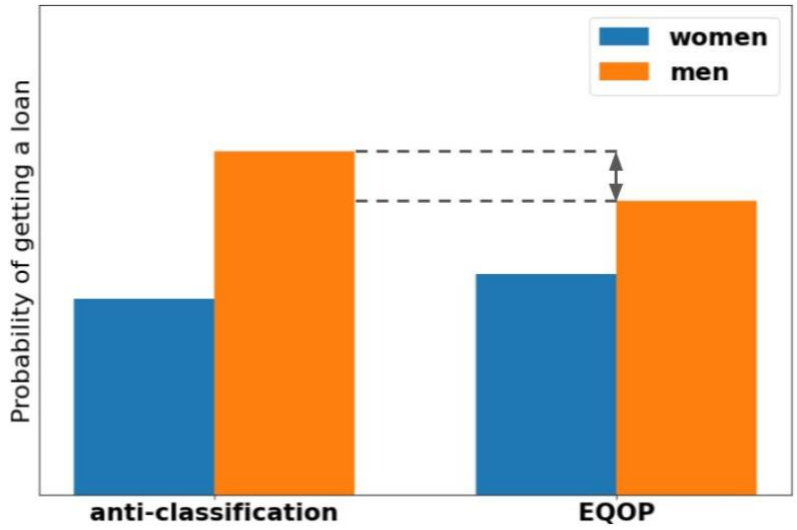

Figure 6: Comparison of the probability of getting a loan under the anti-classification (women: 40%, men: 70%) and EQOP approach (women: 45%, men: 60%). Under the EQOP approach, men are less likely to receive a loan than under the anti-classification approach.

In this comparison, we do not conceptualize harm as the harm of being treated unequally. On the contrary, we focus on how the use of sex information affects the probability of receiving a loan for members of one group, in this case, males. As Figure 6 shows, we are dealing with a (realistic) example in which the EQOP algorithm ends up approving lending to fewer men than the original algorithm[5]. So, the harm *for men* in this case is the harm of receiving worse treatment *if* sex information is considered compared when it is not considered. We could also describe this as the causal effect of sex on the decision *for men*. This effect appears to be harmful for at least some men since the probability of men to be given loans is lowered.

When assessing if the use of protected information (e.g., sex) is *harmful* for persons of one particular group, we need to select a relevant baseline for harm. The relevant comparison[6], in this case, is the initial algorithm developed by the bank, which supposedly maximizes the utility for the bank (e.g. profit), which had been criticized for being *indirectly* discriminatory and unfair. Given this assumption, the answer to the question "does EQOP discriminate against a sex, and

[5] This case seems realistic as in order to achieve EQOP, the threshold for women has to be decreased and/or the threshold for men has to be increased. If the threshold for women is decreased too much, the bank approves too many defaulting applicants (false positives), which are costly. Not approving a repaying applicant (false negative) is not as costly - the bank just does not get to profit from giving out a repaying loan, but at least it does not lose the entire loan). Thus, there is an incentive for mostly equalizing the rates by increasing the threshold for men.

[6] A "relevant comparison" is not the comparator in the sense of Figure 3. In this paper (and thus in Figure 3), we use the term "comparator", to mean a a comparison between members of *distinct* groups, e.g. men vs. women, or sub-groups of these (demographic) groups (such as non-defaulting men vs. non-defaulting women). All accounts of harm involve a comparison but it can also be a "condition comparison", i.e., a comparison of the same group under two alternative conditions. For example, harm can be assessed by comparing the average group outcome (e.g. loans for men) in an algorithm in which sex information is not processed vs. the average group outcome in an algorithm in which sex information is processed. This is a direct example of the kind of comparison we discuss in this section.

which one is it?" is that EQOP discriminates directly against *men*, not women. The reasoning behind this claim is that because the threshold for men is higher than the single threshold in the sex-blind algorithm, the EQOP algorithm denies some men loans that the sex-blind (single threshold) algorithm would have given to them. Thus, (2) is satisfied: the harm from using sex in the algorithm[7] impacts a particular group, namely men, that would not have been impacted if they had been judged by the original sex-blind threshold. Moreover, the evidentiary basis (2) is satisfied: the harm in question impacts the two groups in a disproportionate manner. In fact, in this hypothetical case, it is not true of any woman, that the EQOP algorithm denies them a loan that the sex-blind algorithm would have assigned. This follows from the fact that women's probability of being given a loan by the EQOP algorithm is higher than the probability of being given a loan by the original, sex-blind algorithm. The first claim is thus that EQOP discriminates directly against men which is illustrated in Figure 7.

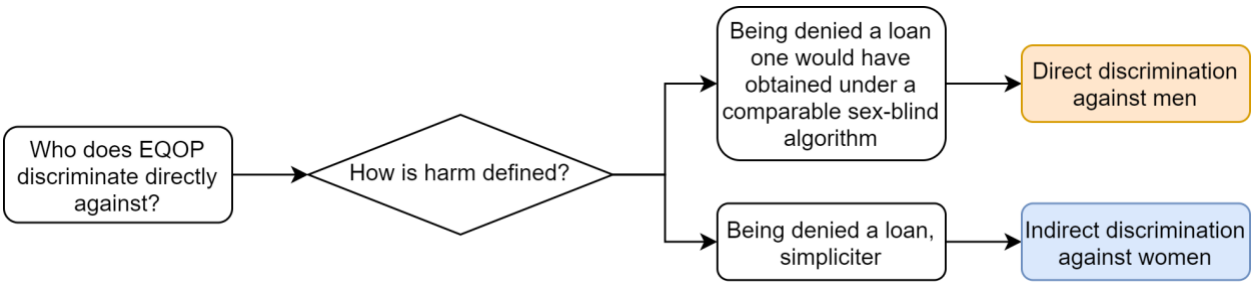

Figure 7: Graphical representation of the first argument in 3.2.2. If the definition of comparators is "men vs women, simpliciter", then whether the EQOP approach is seen as direct discrimination against women or men depends on the definition of harm.

> *Second argument:* EQOP amounts to direct discrimination against men because men applying for loans are more likely to be rejected compared to *women with the same risk score*.

[7] Where harm means: being denied a loan that one would have obtained if sex information had been ignored.

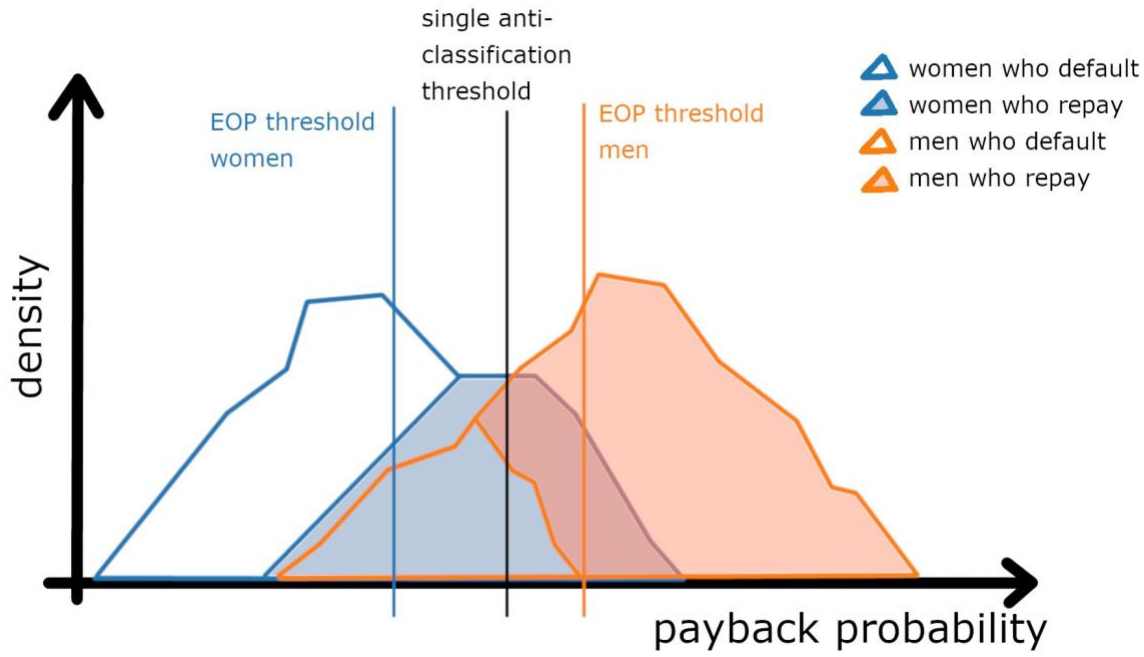

Figure 8: Comparison of the thresholds under the anti-classification approach (black) and the EQOP approach (blue for women, orange for men). The curves show an exemplary distribution of payback probability scores for women and men. Based on this distribution, we can see that men are generally assigned a higher repayment probability than women. We can also see that the threshold for men is higher under the EQOP approach while it is lower for women compared to the anti-classification threshold. This choice of thresholds ensures the equalization of the true positive rates (TPRs) as described in Section 3.1. A group's TPR can be understood visually as the share of its shaded area that lies above its group-specific threshold. This share would be unequal if the single anti-classification threshold was picked, but it is equalized through the group-specific thresholds. As described in Section 3.1, there are several possible threshold pairs that achieve an equalization of the TPRs - we assume that this pair maximizes the profits of the bank which is why these thresholds were picked.

As already mentioned, this argument assumes that the relevant harm of direct discrimination in EQOP is being denied a loan one wants (as opposed to being denied a loan *because* of how sex information is intentionally used in the decision). However, this argument assumes that the relevant comparison group of a woman applying for a loan should be *a man applying for a loan with the same risk score.* This interpretation gains its plausibility when risk scores of men and women are adequately calibrated scores, expressing as a probability the confidence of the bank that a prospective client will repay a loan. If the bank is equally confident that women and men are equally likely to repay a loan if and only if men and women have the same risk score, then it is at least plausible that men are a relevant comparison group for women *who have the same risk score*, not for women *generally* (and conversely). For example, suppose that a risk score is a number from 0 to 1 corresponding to the probability that a client will repay a loan. Assume that

when this score is, for example, 0.7, this corresponds to a probability of repayment of 0.7, irrespective of the sex of the client, and given all the information reasonably available and considered by the bank. Suppose also that this is the case for all risk scores (e.g. 0.2, 0.8 etc.) produced by the model used to award loans. This makes it plausible that women with a 0.2 score ought to be compared with men with a 0.2 score and not, say, with men with a 0.8 score. Then, it is easy to show that conditions of (2) are satisfied in the case we described if the relevant comparison group for women who are denied a loan is seen in the group of men with the same risk score, and conversely.

First, let us show that (2) is satisfied, namely the harm impacts or is likely to impact a protected group. Clearly, men are a protected group, and they remain a protected group when they are men applying for loans and denied a loan.

Second, let us show that the evidentiary standard of (2) is satisfied, namely, the harm of being denied a loan is likely to harm men disproportionately compared to women in a similar situation, where "being in a similar situation" is interpreted as having the same risk score.

This is shown in Figure 8, which we will further illustrate through an example. Suppose that the threshold for men is 80% (orange line in Figure 8) and that the threshold is at 60% for women (blue line in Figure 8) respectively. That is, men are only given loans when they have a repayment probability of at least 80% while women are given loans with a repayment probability of at least 60%. Consider now all men with a repayment probability between 60% and 80%. The EQOP algorithm, using two different thresholds for men and women, denies a loan to all men in this group. However, women with the same predicted repayment probabilities (so between 60% and 80%) are granted loans. Thus, men are harmed compared to women in a similar situation, and this happens by virtue of EQOP using sex information to differentiate the decision for the two groups. Intuitively, one may say that EQOP holds men to a higher standard than women. According to this argument, EQOP discriminates directly based on sex, and men, not women, are the group discriminated against.

Summing up, in the case under examination, it can be argued that even with the same definition of the harm of discrimination (harm as a disposition of the algorithm to treat one less favorably than a relevant comparator group of the different sex), whether it is women or men who are discriminated directly against depends on our choice for the relevant comparison group, which is debatable on moral grounds. A different definition of harm (being denied a loan one would have obtained if sex information had been ignored) supports a distinct argument that the discriminated group is that of men, not women, which takes as salient the fact that repayment thresholds for women and men are respectively lower and higher than those used in an analogous, but sex-blind, algorithm. The results of this section are summarized in Figure 9.

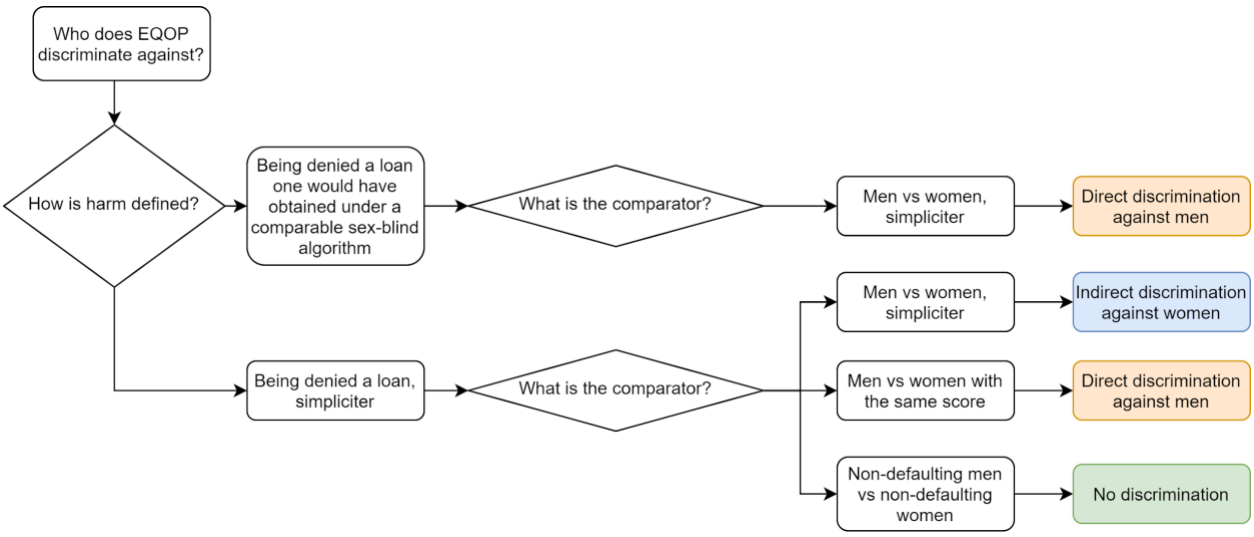

Figure 9: Graphical representation of all arguments in 3.2. The question of whether and who EQOP directly discriminates against depends on the answers to the questions: "How is harm defined?" and "What is the comparator?"

4. DISCUSSION: ULTIMA FACIE JUDGMENT

Returning to the legal ground for such application, the EQOP algorithm is a permissible means of positive action ultimately achieving greater fairness. We first explain the use from the perspective of the positive, advantageous, effects on women, and then from the alternative position of, potentially disadvantaged, men. Our analysis solidifies the legitimacy of the algorithm as a positive action under EU law and dispels claims of positive discrimination. (In this article we utilize the term "positive action" to describe measures taken to redress previous injustice and promote equality, however, there are many different terms used nationally and internationally to describe such anti-discrimination measures. While multiple other terms can be and are employed in practice, throughout the EU "positive action" or "positive measures" are the most widely used and are thus used here.)

4.1. USE OF THE EQOP ALGORITHM IS A POSITIVE ACTION

Looking to the implementation of the EQOP algorithm, we can objectively see a direct impact of achieving greater fairness for women. However, the legal analysis of this positive action must take into account the potential for other legal regimes to apply to the situation, for example, data protection, as well as the ramifications for the "victims". We examine the legality of the EQOP algorithm from an anti-discrimination perspective as well as under current data protection schemes, ultimately concluding it is a permissible special action.

Falling within a number of legal regimes, the most important of which being international human rights law(s), EU regulations explicitly allow for positive actions to atone for historical

disadvantages, "[w]ith a view to ensuring full equality in practice, the principle of equal treatment *shall not prevent* any Member State from maintaining or adopting *specific measures to prevent or compensate for disadvantages linked to [a protected ground*]" (Equal Treatment Directive, Art. 7 (1))). As the directive is written in a negation, (common language regarding special measures), meaning special actions are *not prohibited* versus encouraged or required, questions regarding the obligation to implement special actions still exist in a legal grey area. Further adding to the lack of clarity, the EU Charter of Fundamental Rights extends the *necessity* for special *protections* to apply to a number of classes based on sex, age, and the disabled. It may be arguable that a special action is a form of protection, however, for our purposes it seems clear that - at a minimum - use of a special action to prevent continued discrimination against women in the economic sphere, e.g. not obtaining a bank loan due to algorithmic discrimination, would likely be permissible. Discrimination here being that due to a mere statistical fact, equally qualified women and men (i.e., non-defaulting women and men) are classified as being at higher or lower risk of defaulting and, for that reason, some of them are not given a loan. This is a legitimate prediction that can be made, given sufficiently reliable data, for any classifier that violates equality of opportunity in the above defined sense. This discrimination can only be avoided by designing algorithms that achieve EQOP on purpose.

Beyond the allowance of special actions, the type and use of such measures are strictly defined. The Handbook on European non-discrimination law explains "such measures should be appropriate to the situation to be remedied, legitimate and necessary in a democratic society. Furthermore, they should respect the principles of fairness and proportionality, be temporary and they shall not be continued after the objectives for which they have been taken have been achieved." Regarding women specifically, the UN Committee on the Elimination of Discrimination Against Women notes such measures conceivably include "preferential treatment; targeted recruitment, hiring and promotion; numerical goals connected with time frames; and quota systems".

Generally, positive actions are intended to be short-term exceptional actions to promote individuals who normally experience discrimination (Handbook, 2018). Because of their nature, they are viewed as an exception to anti-discrimination law since their effects favor one group (the historically disadvantaged) over others (For example, see *Kalanke*, 1995; *Marschall*, 1997; and *Abrahamsson*, 2000). Thus, courts generally accept that different treatment can or will occur, but views such disparate treatment as justifiable for correction of past (mis)treatment. This is important when considering the perspective of the newly disadvantaged group. While the ultimate goal is to develop a measure to mitigate a specific form of historical discrimination, and improve the share of loans across groups, it is always necessary to consider the impact of such measures in total.

Applying the EU's strict parameters for a high-level analysis, we can see that the EQOP algorithm falls squarely within the definition of permissible positive action. First, there are a variety of legal regimes that aim to promote the use of such positive actions to achieve greater fairness for women

based on the protected ground of sex, thus fairness between men and women is a legitimate and appropriate objective. Second, the EQOP algorithm is a proportional measure in that, as explained above, while the changed threshold(s) help promote women applicants, overall women still suffer disadvantages. Men, on the other hand, are only minimally harmed in comparison to the original algorithm[8], only in comparison to women who benefit from an altered (lowered) threshold. Thus, while EQOP helps mitigate the inequality of loan acceptance by sex, only appropriate risk scores are granted a loan. This represents a tailored and proportional approach. Third, the time frame for which such an algorithm is in use may be monitored, but as the original algorithm was inherently discriminatory, there can be no return to such a model. While another method of approval may be developed in the future to better suit the bank's needs[9], it seems apparent that use of a more-fair algorithm should remain in place, against the risk of continued and systematic discrimination against women.

There are innumerable situations where an EQOP algorithm could be implemented to achieve greater fairness as an anti-discrimination measure. The first considerations for determining whether such a positive action would be permissible is whether it is: i) an appropriate and legitimate aim; ii) proportional (including a fairness analysis) and, iii) occurring for only as long as the situation requires (*Glatzel*, 2014; Handbook, 2018). Such measures still would have to pass muster under national laws as well as any other legal regimes covering the affected field. Again, whether States have any obligation to take positive action measures is still unclear and developing. In our case, while use of the EQOP algorithm to achieve greater fairness may not be pre-emptively required, use of the original discriminatory algorithm would, eventually, run afoul of anti-discrimination laws and require correction. Finally, even if a positive action meets the aforementioned criteria, further analysis is required to examine the effects of such measures and the other legal regimes which may have overlapping regulation. While implementation by state may vary, use of the EQOP algorithm would likely be permitted under the broad EU positive action affirming regulations.

Though several legal regimes may have a nexus to a positive action and require review, one key law for our EQOP algorithm, and the most likely to occur in practice, is that of data protection law. As mentioned in Section 2.1 data protection law does not prohibit the processing of sensitive data, but it is required to ensure that the processing adheres to an adequate legal ground enumerated, in the EU, by the GDPR. The adequacy of a legal ground will depend on the concrete circumstances of the case and on where the data processing occurs as the GDPR does not fully harmonize data protection law but leaves room for specific provisions within the jurisdiction of Member States. While explicit consent is often relied upon as a legal basis, it is likely that in the case of the employment of EQOP algorithms men will be less likely to consent to such practices,

[8] The minimal harm is a feature of the example we use in the case study, but not a mathematical implication of achieving EQOP. In practice, men could be harmed more, less, not at all or even profit in that more men receive a loan. It seems, however, likeliest that men will be somewhat harmed due to reasons outlined in footnote 5.

[9] This could be done, for example, by retraining the model based on more recent data assuming the difference in the probability of repayment between men and women becomes lower over time.

as it could raise their fears that the processing will lead to a relative disadvantage to them. However, this assumption may not hold stand, especially not if the phrasing of the consent form elicits that the use of the EQOP algorithm benefits overall fairness (who would say no to fairer outcomes if asked to consent?). However, if we assume that too many men opt out of the EQOP algorithm due to fears that they will be treated less favorably, the algorithm cannot perform properly. In such a scenario, it would be preferable to rely on another legal ground, especially a legal obligation to employ EQOP algorithm if proven that it improves the overall fairness of a distribution of – in our example – loans. Here too, Member States could even include a legal obligation to ensure positive action.

Aside from the legal ground the fundamental principles of processing of personal data must be adhered to (Section 2.1.2). It will be important for the development of EQOP algorithms to transparently explain how the algorithm works and how sensitive data are being processed. Transparency under the GDPR contains a rather narrow, focused prospective and retrospective informational perspective (Felzman et al., 2019). Prospectively, information must be provided to the subjects of the EQOP algorithm with respect on what information about them is being processed and for what purposes. The retrospective aspect comes into play, when automatic decision-making is being enabled (meaning that data is being processed and decisions with an impact on an individual are being reached without human involvement). In other words, in our context, how loans are in the end distributed and by whom will influence whether and to what extent the latter transparency requirement must be fulfilled with.

4.2. USE OF THE EQOP ALGORITHM DOES NOT CONSTITUTE ILLEGAL DISCRIMINATION AGAINST MEN

While clearly beneficial in achieving greater fairness for women, contrastingly, implementing the EQOP algorithm appears to have a negative impact on men, as discussed in Section 3.2.2. Often considered "reverse discrimination", such actions may be viewed as an infringement on the rights of the now disadvantaged group (here, men). Many factors may contribute to such views. In our case, men may claim disadvantage based on the differing thresholds between women and men, giving preferential treatment to women. This measure applies different treatment leading to discrimination against men, in that the threshold for men to obtain a loan is higher (more difficult) than for women. However, this type of discrimination is permissible under certain circumstances.

When remedying for past discrimination or injustice, "equal opportunity" will in-fact preference some at the expense of others. As mentioned, *positive action is considered discriminatory* for the victimized group, however, it is considered an *exception* as it is justified as a means to correct and prevent disadvantage. The UN explains such measures are explicitly for "eliminating existing inequality as well as preventing future imbalances" (CERD, 2009). Further, the European Court of Human Rights has illuminated the field, noting a special measure "is discriminatory if it has no objective and reasonable justification; in other words, if it does not pursue a *legitimate aim*…"

(emphasis added in italics). The Court goes on to note that a special measure may also be discriminatory "if there is not a reasonable relationship of *proportionality* between the means employed and the aim sought to be realised." (*Abdulaziz*, 1985; *Burden*, 2008; *Guberina*, 2016). Thus, while there is agreement that often a form of discrimination may occur, it is exempt due to a (often silent) balancing analysis regarding the legitimate aim of the measure and the proportionality of means employed.

From an anti-discrimination perspective, we can look to the harm(s) suffered to ensure that the positive action is a legitimate aim with proportional means and effects. Here, we are concerned with the historic economic disadvantages suffered by women, a legitimate cause for remediation. Use of the bank's original algorithm had a discriminatory impact on women, which led to the bank's interest in remedying such impact. Legitimate aims have included various means of eliminating historical inequalities, for example disadvantages to women in the labor sector (see, for example, *Kalanke*, 1995; *Marschall*, 1997). Here, we can see the social and economic importance tied to the bank's aims. While there is no explicit criteria for what constitutes a "legitimate aim" it is important to consider that these cases often include (and seek to promote) *historically disadvantaged groups*. It is not efficient or possible to alleviate every form of "discrimination" or disadvantage as there are social and ethical considerations to be balanced (see also Foran, 2019). Just because disadvantage occurs, it does not always rise to the level of an actionable claim or require special measures. It is important to note that for a special action to be upheld, the goal sought must rise to the level of requiring interference - interference that may in turn cause disadvantage to others. Thus, the legitimate aim criteria is utilized as a broad protection to ensure a tailor made and common-sense approach to affirmative actions. Ensuring proportionality is a second such safeguard. As noted, such measures are strictly construed and must be *proportional* (See *Burden*, 2008; *Guberina*, 2016; and *Abrahamsson*, 2000). For example, when the positive action still allows for individual exceptional cases to merit further review (*Marschall*, 1997). In some cases, discretion may be built into a positive action to allow for individualized outcomes or goals, achieving a proportional approach. Proportionality likewise is strictly construed in our bank loan example, in that not all loans are given to women, nor are women provided an egregious advantage (contrast *Abrahamsson*, 2000). When utilizing the EQOP algorithm women still obtain loans less frequently than men (possibly evidencing continued discrimination), however, we can say the playing field has been modestly altered in that the opportunity to obtain a loan becomes more fair for the historically disadvantaged group. Where harm may be evidenced is when men with a low score do not obtain loans while women with the same, or lower score do. While there is differential treatment, in essence, the primary effect is to promote more women with an appropriate risk score into the category of successful applicants, as the men with a non-successful low score likely would have been denied under the original algorithm (with exceptions).

While an affected group may still fall victim to a form of disparate treatment or discrimination, it is the goal of reparation for historic injustice, and more importantly, the elimination of future inequality that, on the whole, balance the harm(s) suffered. Differential treatment can and will

occur through the nature of positive measures, however, justice is served through balancing the goals of a legitimate non-discrimination measure with proportional objectives and outcomes. In our example, the indirect discrimination against women induced by the original anti-classification algorithm is partially reduced through the implementation of the EQOP algorithm. We argue that this reduction is proportional if we assume that whether a person deserves to be granted a loan primarily depends on whether they will repay the loan or not. Insofar, all people who would repay their loan are equally deserving of it. In this case, systematic differences in the rate at which deserving women and deserving men are granted loans are unjust. It would thus be proportional to (attempt to) equalize these rates - even if this means that fewer qualified men will receive a loan in exchange for an equalization of the acceptance rate of qualified men and women. It is possible to enact safeguards in that exceptional candidates may be reviewed or successful outside of the thresholds, (for example in *Marschall,* 1997*).* Positive actions are often primarily concerned with mitigation of discrimination in total, outweighing the cost to few by the benefit to many. A balancing evaluation and case by case analysis is imperative.

Finally, in each instance we must consider balancing the interests at hand. In our example the primary interests include data protection and individual privacy, economic costs (for example, minimizing the occurrence of the bank granting loans to undeserving applicants) as well as the social and ethical considerations of achieving greater fairness between the sexes. In the balancing analysis that must occur at the level of the bank deciding to implement such measures, or later by a court if a claim were to arise, there are underlying ethical considerations to fully realize the social values and benefits implemented through a type of affirmative action or special measure. While not explicitly defined under EU law, the silence allows for more interpretation and greater possibilities for possible answers.

While this paper has focused on a high-level analysis of the possibilities under EU laws (subject to national and sector-specific laws), it is worth noting the interesting legal possibilities (or impossibilities) internationally. Future research should include the compatibility of such measures via sector-specific and more localized analyses. Every case of equality of opportunity needs to be evaluated relative to the amount of disparate impact it causes. In some cases equality of opportunity may not be justifiable given the current legal frameworks as described (for instance, if they fail to meet legitimate aim or proportionality requirements).

5. CONCLUSION

This paper analyses how data collection, and specifically, collection of sensitive information, may be permissible under EU law in order to build predictive algorithms that explicitly take fairness into account. First, we presented the laws pertinent to this case: anti-discrimination and data protection law. Next, we considered a hypothetical, yet realistic, case of a bank that determines the approval of loan applications through a predictive model. A first version of this model does not take (sensitive) sex information into account, while the second version of the model intentionally utilizes this information in order to equalize the share of qualified women and men who receive a

loan. We offer different interpretations of the updated model, while acknowledging that such a case could give rise to numerous interpretations. Our findings evidence prima facie arguments for the occurrence of indirect discrimination against women as well as possible claims of direct discrimination against men. However, the most salient interpretation of the updated model is that, even considering its impact on men, it is a permissible positive action and is not actionable as discriminatory due to the legitimate and proportional equalization of acceptance rates across qualified men and women. We weigh these claims with respect to current EU anti-discrimination and data protection laws and find that using sensitive information to build fairer predictive models in similar circumstances is likely permissible under EU law as a positive action.

REFERENCES

Angwin, J., and Larson, J. (2016, May 23). Machine Bias. *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Article 29 Working Party, 'Guidelines on Automated individual decision-making and Profiling for the purposes of the Regulation 2016/679' (WP 251, rev.01, 6 February 2018), <https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053> (last accessed 30 June 2020)

Barocas, S., and Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev., 104*, 671.

Bartlett, R., Morse, A., Stanton, R., and Wallace, N. (2019). *Consumer-lending discrimination in the FinTech era* (No. w25943). National Bureau of Economic Research.

Bent, J. (2020). "Is Algorithmic Affirmative Action Legal?" *The Georgetown Law Journal*, 108(4), 804-853.

Bogen, M., Rieke, A., & Ahmed, S. (2020). *Awareness in Practice: Tensions in Access to Sensitive Attribute Data for Antidiscrimination*. 9. https://arxiv.org/abs/1912.06171

Burden v. the United Kingdom [GC], No. 13378/05, 29 April 2008

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society,* 1–12. https://doi.org/10.1177/2053951715622512

Chandler, A. (2017). The Racist Algorithm?, *Mich. L. Rev.* 115(6), 1023-1045.

Commission Nationale de l'Informatique et des Libertés (CNIL), 'Algorithms and artificial intelligence: CNIL's report on the ethical issues' 25 May 2018 <https://www.cnil.fr/en/algorithms-and-artificial-intelligence-cnils-report-ethical-issues>(last accessed 30 June 2020)

Corbett-Davies, S., and Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Cormen, T. C., Leiserson, C. E., Riverst, R. L., and Stein, C. (2009). *Introduction to algorithms* (3rd ed.). Cambridge, MA: MIT Press.

Datta, A., Fredrikson, M., Ko, G., Mardziel, P., and Sen, S. (2017). Proxy non-discrimination in data-driven systems. *arXiv preprint arXiv:1707.08120*.

Directive 2006/54/EC of the European Parliament and of the Council on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast) (2006)

Council Directive 2000/43/EC implementing the principle of equal treatment between persons irrespective of racial or ethnic origin (2000).

Council Directive 2000/78/EC establishing a general framework for equal treatment in employment and occupation (2000)

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226).

ECJ, Kalanke v. Freie Hansestadt Bremen, Case C-450/93 [1995] ECR I-3051, 17 October 1995.

ECJ, Marschall v. Land Nordrhein-Westfalen, Case C-409/95 [1997] ECR I-6363, 11 November 1997.

ECJ, Abrahamsson and Leif Anderson v. Elisabet Fogelqvist, Case C-407/98 [2000] ECR I-5539, 6 July 2000.

ECtHR, Guberina v. Croatia, No. 23682/13, 22 March 2016.

ECtHR, Abdulaziz, Cabales and Balkandali v. The United Kingdom, 15/1983/71/107-109, 24 April 1985

Epstein, P. and Masters, D. (2011). Direct Discrimination: A Practical Guide to Comparators. London: Cloisters Chambers.

European Union Agency for Fundamental Rights and Council of Europe (2018). The Handbook on European non-discrimination law. *Available at:* https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-handbook-non-discrimination-law-2018_en.pdf

Farkas, L., and O'Farrell, O. (2015). *Reversing the burden of proof: Practical dilemmas at the European and national level*. Publications Office of the European Union.

Felzmann, H., Villaronga, E. F., Lutz, C., and Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, *6*(1), 2053951719860542.

Foran, M. P. (2019). Discrimination as an Individual Wrong. *Oxford Journal of Legal Studies*, *39*(4), 901-929.

Frenzel, (2018) Commentary on Article 9 GDPR, in Paal, Pauly, DS-GVO Kommentierung, 2nd edition.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2016). On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315-3323).

Helberger, N., Araujo, T., and de Vreese, C. H. (2020). Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. *Computer Law & Security Review*, *39*, 105456. https://doi.org/10.1016/j.clsr.2020.105456.

Information Commissioner's Office, 'Big data, artificial intelligence, machine learning, and data protection' (2017) available at:
https://ico.org.uk/media/fororganisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf
(last accessed 30 June 2020)

Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. (2018, May). Algorithmic fairness. In *Aea papers and proceedings* (Vol. 108, pp. 22-27).

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. https://doi.org/10.4230/LIPIcs.ITCS.2017.43.

Lipton, Z., McAuley, J., and Chouldechova, A. (2018). Does mitigating ML's impact disparity require treatment disparity?. In *Advances in neural information processing systems* (pp. 8125-8135).

Liu, L. T., Simchowitz, M., and Hardt, M. (2019, May). The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning* (pp. 4051-4060). PMLR.

Malgieri, G. (2020, January). The concept of fairness in the GDPR: a linguistic and contextual interpretation. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 154-166). DOI: https://doi.org/10.1145/3351095.3372868

Norwegian DPA (2020). The Norwegian DPA intends to order rectification of IB grades. Retrieved from https://www.datatilsynet.no/en/news/2020/the-norwegian-dpa-intends-to-order-rectification-of-ib-grades (last accessed August 30 2020)

Naughton, J. (2020, September 6). From viral conspiracies to exam fiascos, algorithms come with serious side effects. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2020/sep/06/from-viral-conspiracies-to-exam-fiascos-algorithms-come-with-serious-side-effects

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge and London: Harvard University Press.

Sánchez-Monedero, J., Dencik, L., and Edwards, L. (2020). What does it mean to "solve" the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 458–468. DOI: https://doi.org/10.1145/3351095.3372849

Consolidated versions of the Treaty on European Union and the Treaty on the Functioning of the European Union (TFEU) [2016] OJ C202/1.

UN, Committee on the Elimination of Discrimination Against Women (CEDAW) (2004), General Recommendation No. 25: Art. 4, para. 1, of the Convention (temporary special measures), UN Doc. A/59/38 (SUPP), 18 March 2004, para. 22.

UN, Committee on the Elimination of Racial Discrimination (CERD)(2009), General Recommendation 32: The Meaning and Scope of Special Measures in the International Convention on the Elimination of All Forms of Racial Discrimination, UN Doc. CERD/C/GC/32, 24 September 2009.

UN, International Criminal Tribunal for Rwanda, Prosecutor v. Ferdinand Nahimana, Jean-Bosco
Barayagwiza and Hassan Ngeze, Case No. ICTR-99-52-T.

Wachter, S., Mittelstadt, B., and Russell, C. (2020). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Available at SSRN*.: https://ssrn.com/abstract=3547922

Wolfgang Glatzel v. Freistaat Bayern, CJEU, Case C-356/12 [2014] 22 May 2014.

Žliobaitė, I., and Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, *24*(2), 183-201.