DOI:
10.9781/ijimai.2021.02.008

Document Version
Peer reviewed version

Attesting Digital Discrimination Using Norms

Natalia Criado*, Xavier Ferrer*, Jose M Such*
*King's College London, United Kingdom

***Abstract*—More and more decisions are delegated to machine learning systems. Despite initial misconceptions about machine learning (ML) systems being faultless and fair, public distrust in machine learning has been fueled in recent years by shocking examples of digital discrimination such as racist algorithms being used to inform parole decisions in the US, low-income neighborhood's targeted with high-interest loans or low credit scores, and women being undervalued by 21% in online marketing. This poses a significant challenge to the adoption of ML by companies or public sector organisations, despite ML having the potential to lead to significant reductions in cost and more efficient decisions. This has motivated technical research in the area of fair ML. However, users of ML systems do not always have the technical skills to use the fairness metrics proposed by this research and understand to what extent their algorithms can commit discrimination. To allow non-technical users to benefit from ML, simpler notions and concepts to represent and reason about digital discrimination are needed. In this paper, we use norms as an abstraction to represent different situations that may lead to digital discrimination. In particular, we formalise non-discrimination norms in the context of ML systems and propose an algorithm to check whether ML systems violate these norms.**

I. Introduction

Digital discrimination is a form of discrimination in which automated decisions taken by algorithms, increasingly based on AI techniques like machine learning, treat users unfairly, unethically, or just differently based on their personal data [1] such as income, education, gender, age, ethnicity, or religion. Digital discrimination is a serious problem [2] that is becoming even more important because an increasing number of tasks are being delegated to automated decision-making systems embedding those algorithms, such as computers, mobile devices, autonomous systems, etc. Just to give one example among many, some firms in the UK now base at least part of their decisions regarding screening or hiring candidates on automated decision-making systems[1].

Frequently the users of such machine learning (ML) systems are not technical experts and cannot assess by themselves if these algorithms are discriminatory. For example, many public organizations would like to reduce operational costs and delegate some decisions to algorithms, but at the same time need some guarantees about the ML systems not breaking anti-discrimination laws. Our approach has been precisely designed to allow non-technical users to determine if ML systems are potentially discriminatory and to make explicit under which assumptions the systems are discrimination free.

Author's copy of the manuscript accepted in the International Journal of Interactive Multimedia and Artificial Intelligence (IJIMAI). Corresponding author: Natalia Criado (email: natalia.criado@kcl.ac.uk).

[1]http://www.bbc.co.uk/news/business-36129046

This paper is organised as follows: Section II introduces background knowledge on discrimination legislation; Section III introduces our formalization of non-discrimination norms in the context of ML systems; Section IV contains our attesting algorithm; Section V illustrates the performance of our algorithm in three case studies; Section VI contains related work; and Section VII contains a discussion of the paper contribution.

II. Background

Legislation about discrimination in general, not necessarily just about digital discrimination, is varied and extensive. National and international governments and organisations have legislation that specifically prohibits discrimination; e.g., the European Convention for the Protection of Human Rights. Most of this legislation states a non-exhaustive list of criteria or protected attributes, e.g., race, gender, sexual orientation, based on which discrimination is prohibited. This means that, from a legal perspective, discrimination are usually the actions, procedures, etc., that disadvantage citizens based on their personal characteristics, which most of the time imply membership of particular groups defined by those protected attributes.

Legislation about discrimination typically distinguishes between two main types of discrimination [3]:

1) Direct discrimination. This type of discrimination, also known as disparate treatment, considers the situations in which an individual is treated differently because of their membership to a social group defined by protected attributes. This means that different social groups are being treated differently, with some of them being disadvantaged by the differences in treatment. One example of direct discrimination would be a company that has a policy of not considering candidates for hiring who are women with young children. Note, however, that direct discrimination does not necessarily mean that discrimination is explicit. Direct discrimination can be both:
    a) Explicit, as in the previous example, in which members of a particular social group, i.e., women with young children, are explicitly disadvantaged by a decision, i.e., women with your children will be treated differently and not considered for hiring.
    b) Implicit, in which the discriminated group is not explicitly mentioned or considered. Coming back to the previous example, the same company could replace the hiring policy with a new policy of not hiring candidates who have had a career break in recent years. The new policy would not explicitly

consider the relevant social group (women with children), yet it may accomplish the same exact objective, because woman with young children are statistically more likely to have had a recent career break.

2) Indirect discrimination. This type of discrimination, also known as disparate impact, considers the situations in which an apparently neutral act has a disproportionately negative effect on the members of a particular social group. This is considered discrimination, even if: i) there is no clear intention to discriminate against that particular social group, and ii) there is not any unconscious prejudice motivating the discriminatory act. For example, a company having the policy to only consider customer satisfaction scores to award promotions may have a disproportionate impact on women, as there is empirical evidence [4] suggesting that women are undervalued when evaluated for a similar objective performance, compared to their male counterparts. That is, it could be argued that the company may not have an intention to discriminate against female employees, but the promotion criteria set may effectively disadvantage them in a disproportionate way when compared to male employees.

III. Digital Discrimination Normative Model

The term digital discrimination refers to those direct or indirect discriminatory acts that are based on the automatic decisions made by an ML system. In this section, we formalise the notion of digital discrimination norms accounting for the different types of discrimination introduced in the previous section: explicit, implicit, and indirect discrimination.

An ML system can be defined by a set of input features $\mathcal{I} = \{I_1, ...I_m\}$, where each feature I_i takes values from a discrete domain D_{I_i}; and an output feature O, which also takes values from a discrete domain D_O.[2] Note that, in this paper, we are interested in ML systems where the input may contain, directly or indirectly, personal information about individuals in order to attest discrimination. For this reason, the set of *protected* features is also defined; i.e., $\mathcal{P} = \{P_1, .., P_n\}$, where each protected feature $P_i \in \mathcal{P}$ takes values from a discrete domain D_{P_i}. It may be that protected features are part of the input directly used by an ML system, but it is not necessary, e.g., as we will see later, protected features could be strongly *associated* with the inputs even if not directly used as inputs.

The decisions of an ML system can be represented as a dataset DS formed by tuples $(p_1, ...p_n, i_1, ...i_m, o)$, where each tuple represents a previous decision made by the ML system about a particular individual with protected attributes $p_1, ...p_n$, input attributes $i_1, ...i_m$, and algorithm outcome[3] o. In particular, each $p_i \in D_{P_i}$, $i_i \in D_{I_i}$ and $o \in D_O$.

[2]For simplicity we assume domains are discrete, but this is without loss of generality, as any continuous domain can be discretized.

[3]Note that it is possible to consider discrimination in an algorithm by considering the ground-truth labels as well. See Appendix VIII-B for more details about this particular type of discrimination, which in some cases is known as disparate mistreatment [5].

In the following, we provide a formalization of non-discrimination norms for ML systems and define how domain knowledge can be represented using norm exceptions. These normative notions are illustrated with an example.

A. Digital Discrimination Norms

As aforementioned, in the legislation around the world, we find the following types of discrimination: direct (also known as disparate treatment), which further classifies into explicit and implicit; and indirect (disparate impact) [6]. Next, we contextualise these notions in the context of digital discrimination and we formally represent them as computational norms using deontic logic[4]. These deontic norms express anti discrimination rules of behaviour for ML systems using concepts and terminology easily understood by non-technical users.

1) Direct Discrimination: Direct Discrimination is the unequal behavior toward someone because of a protected characteristic. We consider the two types of direct discrimination identified in previous literature, as discussed in Section II: explicit and implicit discrimination.

a) Explicit Discrimination.: In terms of ML systems, this type of discrimination is equivalent to having some of the protected attributes considered in the systems' input. Norms preventing explicit discrimination can be formalised as prohibitions to include protected attributes in the input of the system as follows:

$$\forall P_i \in \mathcal{P} : \mathbf{F}(P_i \in \mathcal{I})$$

The set of all explicit discrimination norms is denoted by N_E and has a size of $|\mathcal{P}|$.

b) Implicit Discrimination. : This type of discrimination can be formalised as a situation where the values of a set of input attributes of an ML system correlate with the value of one or more of the protected attributes. Therefore, norms preventing implicit discrimination can be formalised as follows:

$$\forall P_i \in \mathcal{P} : \mathbf{F}(P_i \text{ is a function of } \mathcal{I})$$

Note that P_i is a function of $\mathcal{I}$ is defined in terms of a process to detect associations, correlations or dependencies between attributes (Section VI provides more details about techniques and metrics that can be used for this). Also note that the set of all implicit discrimination norms is denoted by N_I and has a size of $|\mathcal{P}|$.

Remark 1: If an explicit discrimination norm for a protected feature P_i is violated, then the implicit discrimination norm for P_i is also violated. The inverse inference, however, does not hold.

2) Indirect Discrimination: Indirect Discrimination (disparate impact) refers to decisions that adversely affect one group of people of a protected characteristic more than another. This equals to state that for a particular protected attribute value $p \in D_{P_i}$, the probability of a given outcome

[4]For simplicity, we don't consider compound discrimination in the main part of this paper. For a definition of compound discrimination norms see Appendix VIII-A.

$o \in D_o$ is x times lower than that of the values of the same protected attribute P with the highest probability. Formally, we can define the norm prohibiting indirect discrimination as:

$$\forall P_i \in \mathcal{P}, \forall p \in D_{P_i}, \forall o \in D_O : \mathbf{F}(P_i \downarrow_o^p)$$

where $P_i \downarrow_o^p$ denotes:

$$Pr(O = o|P_i = p) < x \times \max_{\forall p' \in D_{P_i} \setminus \{p\}} Pr(O = o|P_i = p')$$

with $Pr(O = o|P_i = p)$ standing for the probability that the outcome o is given to an individual with protected attribute p. Therefore, the norm states that it is forbidden that for a given group, characterised by having p as the value for the protected feature P_i, the probability of an outcome o is x times lower than the probability of the same outcome o for all the alternative groups, which are characterised by having the other values for P_i (i.e., $D_{P_i} \setminus \{p\}$).

Note that different methods can be used to estimate this probability. In Section VI, we provide a review of the different techniques that may be used. Also note that the value $x \in [0, 1]$ is a constant representing the extent of the disproportion allowed in a particular domain[5].

The set of all disparate impact norms is denoted by N_D and has a size of $|\mathcal{P}| \times \overline{D_\mathcal{P}} \times |D_O|$, where $\overline{D_\mathcal{P}}$ denotes the average number of values belonging to the domain of protected attributes. That is, there is one disparate impact norm per each group, characterised by having a particular value for a given protected feature, and each possible outcome.

3) Norm Violations: Based on the definitions above, the full set of anti-discrimination norms considered is represented as a collection denoted by $N = (N_E, N_I, N_D)$, where N_E, N_I, N_D are as defined above, representing norms against explicit, implicit and indirect discrimination.

Whenever any of the norms in N are violated, there may then be a case of discrimination. However, some of these violations could be considered inconsequential, as we describe next, or there may also be domain-dependent exceptions (as defined later on in Section III-B).

In this paper, we define inconsequential norm violations as those violations which can be considered trivial, since they have little effect on the decisions made by the ML system. Importantly, inconsequential violations are anyway worth considering, as they may be an indicator of bad practices (e.g., considering disability status of students in university admissions may be immoral even if that information is ultimately not influencing much the decision).

Remark 2: If an explicit discrimination norm for a protected feature P_i is violated and no indirect discrimination norm for P_i is violated, then the violation is inconsequential as the protected feature P_i is not affecting significantly the decision-making process. If an implicit discrimination norm for protected feature P_i is violated and no indirect discrimination norm for P_i is violated, then the violation is inconsequential as the protected feature P_i is not affecting significantly the decision-making process.

[5]For example, the US *fourth-fifth rule* from the Equal Employment Opportunity Commission (1978) states a job selection rate for the protected group of less than 4/5 of the selection rate for the unprotected group [7].

B. Norm Exceptions

The previous section formalises the general definition of anti-discrimination norms. In general, when these norms are violated there is a potential case of digital discrimination. However, there are domains in which the violation of these norms is justifiable, and hence not result in discrimination. To allow for such type of domain knowledge to be explicitly represented and accounted for, we use the notion of domain permission norms, which define exceptions to the general anti-discrimination norms.

1) Exceptions to Direct Discrimination Norms:

a) Exception to Violate Explicit Norms. : This refers to the cases where permission to use protected attributes in decision making may be justified. For example, legislation does not usually consider discriminatory to use religion as a criteria for hiring a religion teacher at a school. An explicit permission to use a protected attribute $P_i \in \mathcal{P}$ can be defined as follows:

$$\mathbf{P}(P_i \in I)$$

The set of all exceptions to explicit discrimination norms is denoted by E_E.

b) Exception to Violate Implicit Norms.: This refers to the cases where permission to allow for correlations between a protected attribute and input attributes is justified. For example, for some particular jobs, (e.g., firefighters) the candidates may need to demonstrate physical strength, which is correlated with gender. In such cases, it may be lawful to consider the results of fitness tests in hiring decisions. This allowed correlation between a protected attribute $P_i \in \mathcal{P}$ and a subset of the input attributes $I \subset \mathcal{I}$ can be represented as a permission norm as follows:

$$\mathbf{P}(P_i \text{ is a function of } I)$$

The set of all exceptions to implicit discrimination norms is denoted by E_I.

Remark 3: An exception to an explicit discrimination norm about protected attribute P_i entails an exception for the implicit discrimination norm related to P_i and all input attributes. The inverse relationship does not hold.

2) Exceptions to Indirect Discrimination Norms: This refers to the cases where permission to treat different groups disparately may be explainable. For example, on average, women Uber drivers are paid less than men drivers [8], but that could be explained by factors such as driver experience, time and location of rides, etc. An exception to allow for a significant difference on an outcome $o \in D_o$ for a particular protected group $p \in D_{P_i}$ where $P_i \in \mathcal{P}$ can be formalised as follows:

$$\mathbf{P}(P_i \downarrow_o^p)$$

The set of all exceptions to indirect discrimination norms is denoted by E_D.

Remark 4: An exception to an explicit discrimination norm about protected attribute P_i does not entail an exception to any indirect discrimination norms for P_i. An exception to an implicit discrimination norm about protected attribute P_i does not entail an exception to any indirect discrimination norms for P_i.

There may be cases in which it is lawful to consider protected attributes in the decision-making process, either explicitly or implicitly, as long as that information is not used to disproportionately disadvantage the members of a certain group; e.g., positive discrimination practices allows the use of gender and race information to increase the number of employees from minority groups in a company or business, which are known to have been discriminated against in the past. In this case there is an exception to an explicit discrimination norm about gender and race, as long as that information is not used to adversely affect any group; e.g., gender information can be used by the ML system as long as all genders do not have disproportional probabilities to obtain the different outcomes.

Domain exceptions to discrimination norms are represented as a collection denoted by $E = (E_E, E_I, E_D)$.

C. Example: Credit Risk Assessment

To illustrate the different types of norms and exceptions let us consider an example of a decision making system that classifies individuals as high or low risk in a credit risk assessment scenario.

The attributes used to describe invididuals are:

$$\mathcal{I} = \{Age, Job, Salary\}$$

where and $Age \in \{[20,30],[30,40],...\}$, $Job \in \{Unemployed, Unskilled, ...\}$, and $Salary \in \{[0,20k],[20k,30k],...\}$. According to common discrimination law, protected attributes are defined as:

$$\mathcal{P} = \{Gender, Age\}$$

where $Gender \in \{Male, Female\}$. The output variable is:

$$O = Risk$$

where $Risk \in \{High, Low\}$.

In this example the following norms are generated considering protected attributes:

$$\mathbf{F}(Gender \in \mathcal{I}), \mathbf{F}(Age \in \mathcal{I}),$$
$$\mathbf{F}(Gender \text{ is a function of } \mathcal{I}), \mathbf{F}(Age \text{ is a function of } \mathcal{I})$$
$$\mathbf{F}(Gender \downarrow_{High}^{Male}), \mathbf{F}(Gender \downarrow_{Low}^{Male}),$$
$$\mathbf{F}(Gender \downarrow_{High}^{Female}), \mathbf{F}(Gender \downarrow_{Low}^{Female}),$$
$$\mathbf{F}(Age \downarrow_{High}^{[20,30]}), \mathbf{F}(Age \downarrow_{Low}^{[20,30]}), ...$$
$$..., \mathbf{F}(Age \downarrow_{High}^{[70,80]}), \mathbf{F}(Age \downarrow_{Low}^{[70,80]}),$$

In addition, in this example, there are also several exceptions to the norms as follows:

$$\mathbf{P}(Age \in \mathcal{I})$$
$$\mathbf{P}(Gender \text{ is a function of } \{Salary\}),$$
$$\mathbf{P}(Age \downarrow_{High}^{[20,30]}), \mathbf{P}(Age \downarrow_{Low}^{[20,30]}), ...$$
$$..., \mathbf{P}(Age \downarrow_{High}^{[70,80]}), \mathbf{P}(Age \downarrow_{Low}^{[70,80]}),$$

In particular, it may be considered lawful to use age in the credit risk assessment, as it is common practice to use age to estimate health risks, insurance, unemployment rates, etc. By Remark 3, it is implicitly permitted that age is a function of input attributes. The pay gap phenomenon also explains a degree of correlation between salary and gender. In this case, however, the use of salary for credit risk assessment may be considered lawful (i.e., salary has not been used as a way to discriminate women, but as a way to determine the capability of individuals to pay a credit back). Finally, it is considered permitted to allow age to have a significant impact on credit risk assessment decisions and any age groups to be discriminated on this basis.

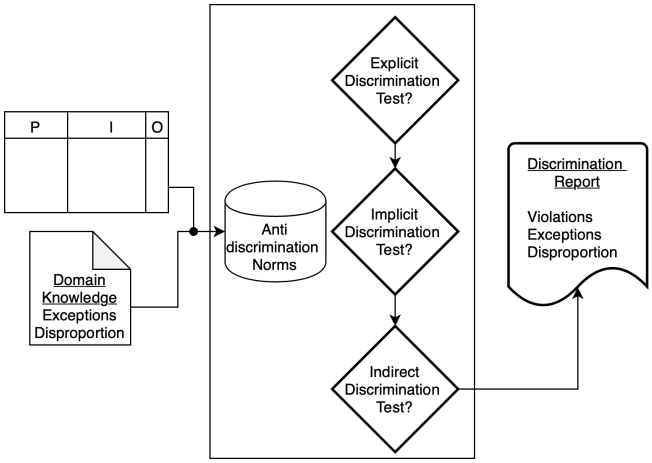

Fig. 1. Overview of the Attesting Process

IV. Digital Discrimination Attesting Process

The digital discrimination attesting process, which is depicted with all its steps in Figure 1, takes as input a decision dataset and the domain exceptions defined by the user, and it returns a discrimination report with information about any potential discrimination cases (i.e., the minimal list of norm violations) and the assumptions made in the attesting process (i.e., the list of exceptions provided by the user and the allowed disproportion ratio)[6].

The attesting algorithm (see Algorithm 1) starts by generating the list of discrimination norms based on the input, protected and output attributes (line 7), and then, it checks compliance with the different types of norms.

a) Explicit direct discrimination: The algorithm starts by checking compliance with explicit direct discrimination norms (lines 8-15). In particular, for each protected attribute it checks if there is an permission norm allowing the ML system to use it as input (line 9). If not, it checks if the explicit discrimination norm is violated, which is the same as checking for set membership. For each explicit norm that is violated, a new inconsequential violation is added (line 11); later on the algorithm will confirm if this violation is actually inconsequential or not. Finally, the implicit norm related to that protected attribute is removed (line 12). Note that our goal is to produce the minimal set of violations and, by Remark III-A1b, the explicit norm is more general.

[6]Note the purpose of our paper is to allow non-technical users to attest whether ML systems discriminate. We do not focus on the mitigation of discrimination when found. For examples of the growing research field on mitigating discrimination see [9], [10], [11], [12].

b) Implicit direct discrimination: The algorithm checks for implicit direct discrimination in lines 16-24. For each implicit norm, the algorithm checks if there is an exception to an explicit norm for the same protected attribute (as stated in Remark 3). If not, the algorithm checks if the norm is violated using the dataset DS as a representative sample[7]. An implicit norm is violated when there is a subset of input attributes determining the value of a protected attribute. If the norm is violated, the algorithm checks for a permission norm allowing for that particular violation. In particular, the algorithm checks if there is an exception for that set of input attributes, or a subset of it, determining the protected attribute (lines 18-19). Again, if the norm is finally considered to be violated, a new inconsequential violation is created (line 21). As before, the algorithm will determine later on if that violation is actually inconsequential or not.

c) Indirect discrimination: The algorithm checks for indirect discrimination in lines 25-39. The algorithm starts by checking for each indirect norm whether there is an exception to it (line 26). If there is not, it checks if the indirect norm is violated (line 27). To determine if an indirect norm is violated the dataset DS is used as a representative sample to calculate probabilities associated to each outcome and protected group [8]. For each indirect norm that is violated, a new violation is created (line 28). As stated in Remark 2, if there are inconsequential violations of explicit norms related to that protected attribute, these are converted into consequential ones (lines 29-32). The violation of the indirect norm associated to a protected attribute demonstrates that decisions are having a disproportionate impact based on that protected attribute. Similarly, if there are inconsequential violations of implicit norms related to that protected attribute, these are converted into consequential ones (lines 33-36).

d) Discrimination Report: Finally, the algorithm outputs the list of inconsequential and consequential violations found. Note that the discrimination report will contain not only the information about norm violations (if any), but also the information about the exceptions considered in its analysis and the level of allowed disproportion specified by the user.

e) Complexity: The complexity of the algorithm to attest digital discrimination is determined by the size of the biggest norm set (or exception set). In this case, the complexity is given by $\mathbf{O}(|\mathcal{P}| \times \overline{D_{\mathcal{P}}} \times |D_O|)$. This assumes that the norm violation checks are performed offline and can be retrieved in constant time. Section VI discusses different methods to check compliance of implicit and indirect norms (note checking compliance of explicit norms equates to checking set membership).

V. Case studies

In this section, we illustrate the performance of our digital discrimination attesting algorithm using three well-known

[7]Different statistical methods can be used to determine if there is a correlation between input attributes and protected attributes. Refer to Section VI for more details.

[8]Different statistical methods can be used to determine the probability of obtaining an outcome value for a particular protected group. Refer to Section VI for more details.

Algorithm 1: Digital Discrimination Attesting

1 **DiscriminationAttesting** $(\mathcal{P}, \mathcal{I}, O, DS, E, x)$
inputs : A set of protect attributes $\mathcal{P}$
A set of input attributes $\mathcal{I}$
An output attribute O
A dataset DS
A collection of exceptions (E_E, E_I, E_D)
$x \in [0, 1]$ a constant representing the disproportion allowed
output: A collection of violated norms (V_E, V_I, V_D)
A collection of norms that have been violated inconsequentially (I_D, I_I)
2 $V_E \leftarrow \emptyset$
3 $V_I \leftarrow \emptyset$
4 $V_D \leftarrow \emptyset$
5 $I_E \leftarrow \emptyset$
6 $I_I \leftarrow \emptyset$
7 $(N_E, N_I, N_D) \leftarrow GenerateNorms(\mathcal{P}, \mathcal{I}, O)$
// Attesting Explicit Discrimination
8 **foreach** $\mathbf{F}(P_i \in \mathcal{I}) \in N_E$ **do**
9 **if** $\nexists \mathbf{P}(P_i \in \mathcal{I}) \in E_E$ **then**
10 **if** $P_i \in \mathcal{I}$ **then**
11 $I_E \leftarrow I_E \cup \{\mathbf{F}(P_i \in \mathcal{I})\}$
12 $N_I \leftarrow N_I \setminus \{\mathbf{F}(P_i$ is a function of $\mathcal{I})\}$
13 **end**
14 **end**
15 **end**
// Attesting Implicit Discrimination
16 **foreach** $\mathbf{F}(P_i$ *is a function of* $\mathcal{I}) \in N_I$ **do**
17 **if** $\nexists \mathbf{P}(P_i \in \mathcal{I}) \in E_E$ **then**
18 **foreach** $I \subseteq \mathcal{I} : I$ *is the minimal set such that* P_i *is a function of* I **do**
19 **if** $\nexists \mathbf{P}(P_i$ *is a function of* $I') : I \subseteq I'$ **then**
20 $I_I \leftarrow I_I \cup \{\mathbf{F}(P_i$ is a function of $\mathcal{I})\}$
21 **end**
22 **end**
23 **end**
24 **end**
// Attesting Indirect Discrimination
25 **foreach** $\mathbf{F}(P_i \downarrow_o^p) \in N_D$ **do**
26 **if** $\neg\exists \mathbf{P}(P_i \downarrow_o^p) \in E_D$ **then**
27 **if** $\exists p' \in D_{P_i} : \frac{Pr(O=o|P_i=p)}{Pr(O=o|P_i=p')} < x$ **then**
28 $V_D \leftarrow V_D \cup \{\mathbf{F}(P_i \downarrow_o^p)\}$
29 **if** $\mathbf{F}(P_i \in \mathcal{I}) \in I_E$ **then**
30 $I_E \leftarrow I_E \setminus \{\mathbf{F}(P_i \in \mathcal{I})\}$
31 $V_E \leftarrow V_E \cup \{\mathbf{F}(P_i \in \mathcal{I})\}$
32 **end**
33 **if** $\mathbf{F}(P_i$ *is a function of* $\mathcal{I}) \in I_I$ **then**
34 $I_I \leftarrow I_I \setminus \{\mathbf{F}(P_i$ is a function of $\mathcal{I})\}$
35 $V_I \leftarrow V_I \cup \{\mathbf{F}(P_i$ is a function of $\mathcal{I})\}$
36 **end**
37 **end**
38 **end**
39 **end**
40 $V \leftarrow (V_E, V_I, V_D)$
41 $I \leftarrow (I_E, I_I)$
42 **return** *V,I*

datasets: the German dataset[9], the Adult dataset[10], and the COMPAS Recidivism dataset[11].

In our implementation[12], we have used the *sklearn* library for normalised mutual information [13] to detect violations of implicit discrimination norms. The normalised mutual information (NMI) is a measure of the mutual dependence between the two variables that quantifies the "amount of information" obtained about one random variable through observing the other random variable. The NMI returns 0 when there is no mutual information between the variables tested, and 1 when there exist a perfect correlation. In the implementation, the minimum coefficient for mutual information can be configured; we used a minimum threshold of 0.6 in the experiments below as indicative of a strong correlation between input and protected attributes. To detect indirect discrimination we have set to 0.8 the allowed disproportion ratio, inspired by the US *fourth-fifth* rule from the Equal Employment Opportunity Commission (1978), a threshold commonly used to detect disparate impact in domains like employee selection procedures[13]; and we have calculated the probabilities using the frequencies in the dataset as a representative sample. Also, due to the small size of the datasets used in the case studies, we have used the Chi-Squared Test [14] to determine those violations of indirect discrimination norms that are statistically significant (p-value < 0.05). To discretise numeric values, we have used quantile discretisation, which is a well-known method for discretising continuous variables in ML [15].

A. Adult Dataset

The Adult dataset uses 14 attributes to determine if a given person makes over 50K a year. The attributes include education, work class, age, sex, race, and occupation, among others. The dataset contains 48842 instances.

Let us assume that the gender, age, native country and race are protected and that the other attributes are the inputs of a ML system.

$$\mathcal{I} = \{workclass, education, education\_num, occupation, capital\_gain, capital\_loss, hours\_per\_week, fnlwgt\}$$

Note attribute $education\_num$ represents the number of education years, and $fnlwgt$ represents the number of people the census believes the entry represents.

$$\mathcal{P} = \{age, gender, native\_country, relationship, marital\_status, race\}$$

$$O = income$$

where $income = \{<= 50k, > 50k\}$. In this case age is related to experience and seniority so it is considered lawful to use age to discriminate:

$$\mathbf{P}(age \downarrow_{<=50k}^{[0,16)}), \mathbf{P}(age \downarrow_{>50k}^{[0,16)}),$$

...

$$\mathbf{P}(age \downarrow_{<=50k}^{[75,99)}), \mathbf{P}(age \downarrow_{>50k}^{[75,99)})$$

After executing our algorithm several violations of indirect discrimination norms are detected. For example:

$$\mathbf{F}\ gender \downarrow_{>50k}^{female}$$

$$\mathbf{F}\ race \downarrow_{>50k}^{black}$$

$$\mathbf{F}\ native\_country \downarrow_{>50k}^{Nicaragua}$$

$$\mathbf{F}\ marital\_status \downarrow_{<=50k}^{Married-civ-spouse}$$

The first violation above indicates that females have a disproportionate lower probability of being classified as making more than 50k when compared with males. In particular, the dataset contains 21790 male instances out of which 6662 are classified as high income (i.e., the probability of income greater than 50k for male is 30%), whereas only 1179 female records out of 10771 are classified as high income (i.e., the probability of income greater than 50k for female is 11%). In this case $11\% < 0.8 \times 30\%$ and it is considered disproportionate. The other violations above indicate that black people and nicaraguans have a disproportionate lower probability of being classified as making more than 50k when compared with other groups, in accordance with previous reports of discrimination in the dataset [16]. On the contrary, married people are significantly less likely of being classified as making less than 50k. Found violations are associated with particular values of gender, native country, relationship and marital-status attributes. This indicates that the decision making process may have a disparate impact on people belonging to particular protected groups.

B. German Credit Dataset

The German dataset contains information about people who ask for a credit. Each person is classified as good or bad credit risks. This is the inspiration for the small example contained in section III-C. In particular, the full dataset uses 20 attributes to represent each person, which include information like age, employment status, gender and personal status of the applicant; and the duration, amount and purpose of the credit. The dataset contains 1000 instances.

Let's us assume an ML system where age, personal status and sex, and being a foreign worker are considered protected attributes, and the rest of the features in the German dataset are considered inputs:

$$\mathcal{I} = \{job, housing, savings, .., amount, duration, purpose\}$$

$$\mathcal{P} = \{age, personal\_status\_and\_sex, foreign\_worker\}$$

$$O = risk$$

where $risk = \{high, low\}$. In this case, it is considered lawful to use age to discriminate credit risks as people are less likely to repay credits as they become older, hence, we consider age as an exception:

$$\mathbf{P}(age \downarrow_{good}^{[0,16)}), \mathbf{P}(age \downarrow_{bad}^{[0,16)}),$$

[9]https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[10]https://archive.ics.uci.edu/ml/datasets/adult

[11]https://github.com/propublica/compas-analysis/

[12]Available on Github at https://github.com/xfold/NormativeApproachToDiscrimination

[13]http://www.uniformguidelines.com

...

$$\mathbf{P}(age \downarrow_{good}^{[75,99)}), \mathbf{P}(age \downarrow_{bad}^{[75,99)})$$

After executing our algorithm, the following violation is detected:

$$\mathbf{F}(foreign\_worker \downarrow_{good}^{yes})$$

The violation means that foreign workers have a disproportionate low probability of being considered a good credit risk.

C. COMPAS Recidivism Dataset

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm is a popular commercial algorithm used by judges and parole officers for scoring criminal defendant's likelihood of re-offending (recidivism), and is increasingly being used in pretrial and sentencing. The dataset has been widely used to study automatic decision systems related with recidivism [17], and it was found to be strongly biased against blacks [18]. The original dataset contains 28 columns which correspond to the variables used by the COMPAS algorithm to make its predictions, including data regarding sex, ethnicity, and marital status (among others), together with the final assessment made by the algorithm, the estimated recidivism score.

In the analysis below, we focus on *pretrial* instances and assessments about the *risk of recidivism* and *risk of violence*, following the analysis performed by ProPublica[14], considering sex and ethnicity as protected variables:

$$\mathcal{I} = \{marital\_status, legal\_status, \ldots\}$$

$$\mathcal{P} = \{sex, ethnicity\}$$

$$O = recidivism\_score$$

where $recidivism\_score = \{low, medium, high\}$. After executing our algorithm, the following violation is detected:

$$\mathbf{F}(ethnicity \downarrow_{low}^{African-American})$$

The violation means that African-Americans have a disproportionate low probability of being considered with a low recidivism score when compared with other sub-populations, coinciding with the results reported in [18]. The reported bias becomes especially apparent when comparing African-American with Caucasian ethnicities, with African-Americans being consistently tagged by the COMPAS algorithm with *higher* and *medium* recidivism scores way more frequently than the Caucasian sub-population.

[14]*https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm*

VI. Related work

Recent research has addressed the problem of discrimination and bias in machine learning. These novel tools are most of the time aimed at technical users capable of interpreting different statistical results, programming, etc. Our algorithm is, on the contrary, aimed at non-technical users (albeit they may be domain experts). The notion of norm and exception is a suitable abstraction to represent the results these statistical analysis to non-technical users. For example, IBM's AI Fairness 360 Open Source Toolkit[15] and Google's What-if-tool[16], are probably two of the most comprehensive toolkits offering a great choice of bias metrics. However, its intended audience are technical users with previous knowledge of machine learning and statistics. Indeed, there are a large number of fairness metrics that may be appropriate for a given application [6], [19]. Also it is difficult for non-technical users to represent domain knowledge in a way that it can be taken into account by the metrics.

Closely related to our work is [20], where the authors proposed to infer classification rules from a given dataset and to detect those classification rules that can cause direct and indirect discrimination. They also allow for domain knowledge, expressed as rules, to be taken into account. Despite the similarities with this work, our proposal has two additional, potential benefits: it doesn't assume that meaningful rules can be inferred, note that it may be impossible to infer rules from complex decision-making algorithms; and it hides to the user the complexities of the analysis process using the notions of norms and exceptions.

a) Implicit Discrimination. : Tramèr et al. [21] developed a methodology and toolkit combining different metrics for discovering associations, or proxies, between attributes. In particular, they studied different metrics that can be used to analyse the relationship between protected attributes and input attributes such as the Pearson correlation, which only works for scalar attributes linearly related; and Mutual Information, which can be applied to categorical attributes.

b) Indirect Discrimination. : There have been many different metrics proposed to measure indirect discrimination both in the raw data used for training as well as the decisions made by the systems. We refer the reader to [22] for an extensive survey in the topic. This survey also discusses other traditional statistical measures that could be applied to measure discrimination. In particular, the authors classify the metrics into: statistical tests, which are used to compute and calculate whether there is discrimination in a dataset; absolute measures, which are used to calculate the magnitude of the discrimination present in a dataset; conditional measures, which are used to capture the extent to which the differences between groups are due to protected attributes or other characteristics of individuals; and structural measures, which are used to identify, for each individual in the dataset, whether they are being discriminated. Next, we also give some more detailed examples of work on indirect discrimination.

[15]https://aif360.mybluemix.net

[16]https://pair-code.github.io/what-if-tool/

In [23], the authors proposed metrics to determine the extent of influence that the inputs to an automated decision-making system may have on its outputs. Although this paper is not intended to detect indirect discrimination per se, the measures the authors of the paper propose have the potential to increase the transparency of decisions made by opaque machine learning algorithms. This, in turn, may provide useful information for the detection of discrimination[24]. Other works have also attempted to propose metrics to capture discrimination in particular applications of Machine Learning. For instance, one example is the work that has attempted to detect discrimination in the applications of ML to Natural Language processing [25], [26], [27]. In these works, the approach followed is to explore the relationships between the words learned by the ML model to detect whether particular words or meanings are more associated to particular individuals based on their personal characteristics.

In addition to the work on detecting discrimination, there is also work focusing on making ML models fairer to start with. For instance, in [28], they test for fairness based on a similarity measure between individuals. For fairness to hold, the distance between the distributions of outputs for individuals should at most be the distance between the two individuals as estimated by means of the similarity metric. In [29], the authors first gather human judgments about the different protected features in the context of two real-world scenarios using Amazon Mechanical Turk. Using the set of *human-assessed* protected features, they compare the accuracy of different classifiers to test the trade-off between process fairness and output accuracy. In [30], they assume fairness can be attested by means of a directed causal graph, in which attributes are presented as nodes joined by edges which, by means of equations, represent the relations between attributes. Finally, the set of violations presented in our approach could also be extended with recent advances in explainable AI. One example is the post-hoc approach of Local Interpretable Model-Agnostic Explanations (LIME), which makes use of adversarial learning to generate counterfactual explanations [31].

VII. Conclusion

Digital discrimination is becoming a significant problem as more decisions are delegated to ML systems. Indeed, recent legislation and citizen initiatives are demanding more transparency about the way in which decisions are made using their data. In response to that, several metrics and tools have been proposed to analyse biases in ML systems. However, these tools often require expert ML or statistical knowledge that many users of ML systems do not necessarily possess.

In this paper, we proposed to use normative notions as an abstraction that may be more easily understood by non-technical users; simplifying the representation of the potential discrimination risks and the input of domain knowledge. Our digital discrimination attesting algorithm not only checks if ML systems are potentially discriminatory but also makes explicit under which assumptions these systems are discrimination free.

As future work, we plan to: i) investigate different metrics to be used in the attesting algorithm and to identify the most usable ones; ii) conduct user studies to further refine the way in which norms could be accessed and influenced by non-technical users to help them understand discrimination risks; and iii) explore interfaces to allow non-technical users to easily introduce exceptions and explanations to communicate the algorithm outputs to these users.

Acknowledgment

The research reported in this article was funded by the Engineering and Physical Sciences Research Council (EPSRC) under grant EP/R033188/1. This research is part of the cross-disciplinary project Discovering and Attesting Digital Discrimination (DADD) – visit our project website for further details: https://dadd-project.org.

References

[1] N. Criado and J. Such, "Digital discrimination," in *Algorithmic Regulation*. Oxford University Press, 2019.

[2] C. O'neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[3] A. Altman, "Discrimination," 2011.

[4] P. Koskinen Sandberg, "Intertwining gender inequalities and gender-neutral legitimacy in job evaluation and performance-related pay," *Gender, Work & Organization*, vol. 24, no. 2, pp. 156–170, 2017.

[5] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1171–1180.

[6] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 2018, pp. 1–7.

[7] S. Barocas and A. Selbst, "Big Data's Disparate Impact," *California law review*, vol. 104, no. 1, pp. 671–729, 2016. [Online]. Available: https://ssrn.com/abstract=2477899

[8] C. Cook, R. Diamond, J. Hall, J. A. List, and P. Oyer, "The gender earnings gap in the gig economy: Evidence from over a million rideshare drivers," National Bureau of Economic Research, Tech. Rep., 2018.

[9] S. Hajian and J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining," *IEEE transactions on knowledge and data engineering*, vol. 25, no. 7, pp. 1445–1459, 2012.

[10] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, 2017, pp. 3992–4001.

[11] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

[12] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[13] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[14] W. G. Cochran, "The $\chi 2$ test of goodness of fit," *The Annals of Mathematical Statistics*, pp. 315–345, 1952.

[15] F. Freese, *Elementary statistical methods for foresters*. US Department of Agriculture, 1967, no. 317.

[16] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović *et al.*, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.

[17] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[18] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias risk assessments in criminal sentencing," *ProPublica, May*, vol. 23, 2016.

[19] X. Ferrer, T. van Nuenen, J. M. Such, M. Coté, and N. Criado, "Bias and Discrimination in AI: a cross-disciplinary perspective," *IEEE Technology and Society Magazine (forthcoming)*, 2020.

[20] D. Pedreschi, S. Ruggieri, and F. Turini, "Integrating induction and deduction for finding evidence of discrimination," in *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, 2009, p. 157–166.
[21] F. Tramer, V. Atlidakis, R. Geambasu, D. Hsu, J.-P. Hubaux, M. Humbert, A. Juels, and H. Lin, "Fairtest: Discovering unwarranted associations in data-driven applications," in *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2017, pp. 401–416.
[22] I. Zliobaite, "A survey on measuring indirect discrimination in machine learning," *arXiv preprint arXiv:1511.00148*, 2015.
[23] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016, pp. 598–617.
[24] T. van Nuenen, X. Ferrer, J. M. Such, and M. Cote, "Transparency for whom? assessing discriminatory artificial intelligence," *Computer*, vol. 53, no. 11, pp. 36–44, 2020.
[25] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.
[26] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes," *PNAS 2018*, vol. 115, no. 16, pp. E3635–E3644, 2018.
[27] X. Ferrer, T. van Nuenen, J. M. Such, and N. Criado, "Discovering and categorising language biases in reddit," in *International AAAI Conference on Web and Social Media (ICWSM 2021) (forthcoming)*, 2020.
[28] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *ITCS 2012*. ACM, 2012, pp. 214–226.
[29] N. Grgić-Hlača, M. Zafar, K. Gummadi, and A. Weller, "Beyond Distributive Fairness in Algorithmic Decision Making," *AAAI*, pp. 51–60, 2018. [Online]. Available: https://people.mpi-sws.org/~nghlaca/papers/fair_feature_selection.pdf
[30] N. Kilbertus, M. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," in *NIPS'17*, 2017, pp. 656–666.
[31] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, and G. Klein Macrocognition, "DARPA XAI Literature Review p. Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI Prepared by Task Area 2 Institute for Human and Machine Cognition," no. February 2019, 2019. [Online]. Available: https://arxiv.org/pdf/1902.01876.pdf

VIII. Appendix

A. Compound Discrimination

Compound discrimination is discrimination based on a combination of protected attributes. In that case of compound discrimination the previous discrimination norms are rewritten as follows:

- Direct.
  - Explicit. There is no need to change the definition of explicit discrimination norms to account for compound discrimination, since the prohibition to include a set of protected attributes in the input can be represented by a set of explicit norms referring to each individual protected attribute.
  - Implicit. There is no need to change the definition of implicit discrimination norms to account for compound discrimination, since the prohibition to have a set of protected attributes as a function of input attributes can be represented by a set of implicit norms referring to each individual protected attribute.
- Indirect (disparate impact). In this case the norms need to represent that for a particular combination of protected attribute values $p_1, ..., p_k$, where each $p_i \in P_i$; the probability of a given outcome $o \in D_o$ is x times lower than for values of the same protected attributes with the highest probability:

$$\forall \{P_1, ..., P_k\} \subseteq \mathcal{P}, (p_1, ..., p_k) \in D_{P_1} \times ... \times D_{P_k}, o \in D_O : \mathbf{F}(\{P_1, ..., P_k\} \downarrow_o^{(p_1,...,p_k)})$$

where $\{P_1, ..., P_k\} \downarrow_o^{(p_1,...,p_k)}$ denotes:

$$Pr(O = o | P_1 = p_1, ..., P_k = p_k) < x \times \max_{\forall \{(p'_1,...,p'_k)\} \in D_{P_1} \times ... \times D_{P_k}} Pr(O = o | P_1 = p'_1, ..., P_k = p'_k)$$

B. Discrimination in Classification Process

In this paper we have focused on digital discrimination; i.e., discriminatory acts facilitated by the automatic decisions made by a ML system. However, it is possible to consider the discrimination in the algorithm itself. This is also know as disparate mistreatment [5]. In those cases it is necessary consider not only the outcome of the algorithm but also the ground-truth labels for the individuals, denoted by G. In those cases, it could be possible to formalise that for no particular value of a protected attribute the ML system can perform significantly worse than for the others groups. This equals to state that for a particular protected attribute value $p \in D_{P_i}$, the probability of the ML assigning the correct outcome $(O = g)$ is x times lower than that of the values of the same protected attribute P with the highest probability. Formally, we can define the norm prohibiting disparate treatment as:

$$\forall P_i \in \mathcal{P}, p \in D_{P_i}, g \in D_G : \mathbf{F}(P_i \uparrow_g^p)$$

where $P_i \uparrow_g^p$ represents:

$$Pr(O = g | P_i = p, G = g) < x \times \max_{\forall p' \in D_{P_i}} Pr(O = g | P_i = p', G = g)$$

$Pr(O = g | P_i = p, G = g)$ stands for probability that the algorithm outcome O is equal to the ground-truth label g for an individual with protected attribute $P_i = p$.