Fair Enough? A Map of the Current Limitations of the Requirements to Have Fair Algorithms

Daniele Regoli[1,*,†], Alessandro Castelnovo[1,†], Nicole Inverardi[1,†], Gabriele Nanino[2] and Ilaria Penco[1,†]

[1]*Data & Artificial Intelligence Office, Intesa Sanpaolo S.p.A., Italy*
[2]*Scuola Superiore Sant'Anna, Pisa, Italy*

Abstract

In the recent years, the increase in the usage and efficiency of Artificial Intelligence and, more in general, of Automated Decision-Making systems has brought with it an increasing and welcome awareness of the risks associated with such systems. One of such risks is that of perpetuating or even amplifying bias and unjust disparities present in the data from which many of these systems learn to adjust and optimise their decisions. This awareness has prompted more and more layers of society, including policy makers, to call for fair algorithms. We believe that while many excellent and multidisciplinary research is currently being conducted, what is still fundamentally missing is the awareness that having fair algorithms is *per se* a nearly meaningless requirement, that needs to be complemented with many additional social choices to become actionable. In this work, we pinpoint and analyse a set of crucial open points that we as a society must address in order to give a concrete meaning to the increasing demand of fairness in Automated Decision-Making systems.

Keywords

Fairness, Bias, Artificial Intelligence, Machine Learning

1. Introduction

Artificial Intelligence (AI) is experiencing an epoch of unparalleled accomplishments. In particular, Machine Learning (ML) approaches to AI are fostering the development of numerous fruitful applications in an incredibly broad range of fields, including computer vision, predictive analytics, and natural language processing.

In the recent years, increasing attention is being devoted to the risks that AI and, more broadly, Automated Decision-Making (ADM), inevitably brings with it. One of such risks is that ADM systems, and ML approaches in particular, are prone to learning from data a wide range of historical and social biases, and to repeating and amplifying them at a scale that is challenging to monitor and regulate. With the term fair-AI, or fair-ML, the literature usually denotes the relatively recent area of research committed to study how to assess and correct biases involved in algorithmic decision-making processes.

Along with this raising attention from the academic world, concerns for the increasing use and diffusion of ADM systems in many industrial sectors emerged from ever broader layers of society. The issue of ethics in AI started gaining popularity, usually under umbrella terms such as "Trustworthy AI" or "Responsible AI". In AI-related contexts, several *suitcase words* began to be widely used, such as "bias", "fairness", "interpretability", etc. [1].

Against this background, in the spring of 2021 the European Commission published the "Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

*Corresponding author.

†The views and opinions expressed are those of the authors and do not necessarily reflect the views of Intesa Sanpaolo, its affiliates or its employees.

daniele.regoli@intesasanpaolo.com (D. Regoli); alessandro.castelnovo@intesasanpaolo.com (A. Castelnovo); nicole.inverardi@intesasanpaolo.com (N. Inverardi); naninogabriele@gmail.com (G. Nanino); ilaria.penco@intesasanpaolo.com (I. Penco)

0000-0003-2711-8343 (D. Regoli); 0000-0001-5234-1155 (A. Castelnovo); 0009-0006-0048-7455 (N. Inverardi); 0009-0008-0228-3733 (G. Nanino)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2025-12-16

CEUR-WS.org/Vol-4132/short14.pdf

Artificial Intelligence" [2], the very first legal framework for AI. The legislative process has concluded in June 2024 with the publication of the Artificial Intelligence Act in the Official Journal of the European Union [3]. Since the European proposal, several other countries started discussing and working on regulatory frameworks for AI systems: the Blueprint for AI and the National AI Commission Act in the U.S., the pro-innovation approach and the AI Safety Institute in the U.K., to name a few.

As a result, public and private actors that employ AI and ADM in their processes are driven to attempt to satisfy this requirement for ethics and justice. Practitioners and researches have access to several "fairness toolkits" [8, 9, 10, etc.], with the aim of assessing and mitigating unjust disparities in algorithmic outcomes. Unfortunately, the level of maturity and consistency of such toolkits appears to be still very low, and their effectiveness in addressing the overall challenge of "fairness" is debated [11].

Contribution and outline

This work aligns with a relatively recent stream of research that is focused critically on the general topic of fairness in ADM [12, 13, 14, 15, 16, 17, 18]. The critic is not on the topic *per se* —whose importance is not disputed— but rather on some usually overlooked subtleties and assumptions that often lead to overreliance and misplaced trust, which can in turn effectively lead to a deterioration of trust in ADM in the long term.

Among the generic risks of "blindly" embracing simplified recipes, we can cite the so-called *Automation Bias*, namely the propensity to place unmotivated trust on automated decisions, or —worse— the possibility of cherrypicking certain simple approaches promoting the false perception that an ADM system respects ethical values. Aivodji et al. [19] has evocatively named the latter *Fairwashing*.[1]

Even if most of the ambiguities and attention points that we here discuss have already been introduced separately by other authors, on the one hand, we try to give an overall perspective, grounding such ambiguities on few foundational intersections of the legal, ethical, and algorithmic perspectives; on the other hand, we approach the topic as a call for action, placing the focus on the fact that most of the ambiguities are a matter of decisions that are not technical in nature, but rather societal, and lie at the intersection of very diverse disciplines. In fact, the main goal of this work is to pinpoint a set of open points that constitute obstacles both for researchers in the field of AI and for practitioners and developers of ADM systems to meet the societal requirement of having fair algorithms.

We build our critical analysis of fair-AI by distinguishing two broad aspects:

1. **Sensitive attributes choice**: It is not that discrimination *per se* is unjust, only discrimination with respect to some attributes that we have either to list and agree upon or to define with some reasonable criterion. (see section 2)

2. **What do we mean by unfair discrimination?**: We need to clarify what we mean by making decisions involving such attributes, and in what cases such decision-making represent an unjust discrimination. (see section 3)

Point 1, discussed in section 2, underlies several difficulties about what attributes we should monitor and what we should do with subgroups and intersections with respect to those attributes. The second point, discussed in section 3, underlies ambiguities that arise when examining the mechanisms the we should consider to be ethically unacceptable by which a sensitive attribute can influence a decision.

2. Sensitive attributes choice

2.1. When a given attribute should be *protected*?

High level non-discrimination principles can be found in several legislative frameworks. At the European level, the starting point of analysis on this topic is the EU Charter of Fundamental Rights. In particular,

[1] More precisely, Aivodji et al. [19] coined the term 'fairwashing' with respect to black-box interpretability techniques used to promote false perception of compliance with ethical values, but the concept has a straightforward extension to fair-AI techniques.

art 21,1 reads:

> *Any discrimination based on any ground such as sex, race, colour, 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.*

To give a concrete application to these principles, various directives have been adopted over time. They detail provisions for specific protected groups and/or specific domains, such as work, environment, or access to goods and services. See the EU non-discrimination website and Wachter et al. [22] for more details.

Moreover, there are domains where the EU Charter is explicitly referred to as a source of high level principles to be observed, but without any concrete details about their implementation. An example is the Consumer Credit Directive, which claims:

> *This Directive respects fundamental rights and observes the principles recognised in particular by the Charter of Fundamental Rights of the European Union. In particular, this Directive seeks to ensure full respect for the rules on protection of personal data, the right to property, non-discrimination, protection of family and professional life, and consumer protection pursuant to the Charter of Fundamental Rights of the European Union.*

The European Commission has in place a proposal of revision of the Directive on consumer credits [24] that contains the following Article 6, explicitly on non-discrimination:

> *Member States shall ensure that the conditions to be fulfilled for being granted a credit do not discriminate against consumers legally resident in the Union on the ground of their nationality or place of residence or on any ground as referred to in Article 21 of the Charter of Fundamental Rights of the European Union, when those consumers request, conclude or hold a credit agreement or crowdfunding credit services within the Union.*

This seems to suggest that all attributes referred to in the EU Charter of Fundamental Rights are to be considered protected for credit access purposes.

Something not dissimilar can be found in the US legislation. We refer to Barocas et al. [25, chap 6] and Barocas and Selbst [26] for more details.

We can summarise the above with the following:

Open Point 1 (Protected groups)**.** *Given a specific phenomenon, what are the groups of people that we should consider as* ***protected****, and with respect to which we therefore have to take care of assessing and avoiding any unjust discrimination?*

2.2. Sensitive data collection

Suppose that your team is in charge of developing a Decision System to rank applicants to a job posting in order to prioritise interviews. The data you collect come entirely from submitted CVs and the online application form, where applicants are asked questions on standard personal data (name, gender, address, date and place of birth), previous work experience, education, and skills. While developing the system, you take particular care of avoiding unjustified dependencies of the outputs on information such as gender and nationality of the applicants. You think you have done all that was possible to prevent any form of unjust discrimination, but when you present your work to your boss, she points out that you actually have no control whatsoever about discrimination with respect to a lot of other sensitive information —such as political and religious opinion— for the very trivial fact that you don't have such information to begin with. It is well-known that being blind to sensitive information is in general not enough to prevent unjust discrimination, or at least some forms of unjust discrimination [see, e.g. 25, 27, 28]. This simple example raises the following:

Open Point 2 (Sensitive data collection)**.** *Should developers of ADM systems keep track of all the sensitive attributes that they would not otherwise record,* ***for the sole purpose*** *of assessing unjust discrimination with respect to those attributes?*

2.3. The problem of aggregation

Even once we agree on what are the dimensions that contain groups of people to be protected against unjust discrimination, we still have to face the problem of how to aggregate individuals along those dimensions. The simplest example in this respect is that of age. Suppose that we are in a situation in which we agree that there should be no discrimination based on the age of the applicants. We can take the example of credit lending: in the U.S., the Equal Credit Opportunity Act forbids to discriminate "on the basis of race, color, religion, national origin, sex or marital status, or age" [Equal Credit Opportunity Act, §1691(a)(1)]. What do we actually mean by age? Should we consider separate protected groups for each year of age? Or is it enough to aggregate individuals into a handful of broader classes, e.g. < 30, $[30, 60]$, > 60? If we opt for the broader classes, how do we select the thresholds? Should we take quantiles of the age distribution of our data, or should we use some common-sense knowledge of how people are actually segmented in the society?

Open Point 3 (Group aggregation)**.** *The specific identification of most attributes that are commonly considered protected depends on alternative ways of aggregating individuals: what strategy should developers follow to choose the proper aggregation when assessing unjust discrimination?*

This same problem arises for almost all the potentially protected attributes, think e.g. of profession, place of birth, political opinions, disability, ethnicity.

It is important to note that, at least for some of the potentially protected characteristics, there are concerns and discussions about the prospect of placing people in rigid and exclusive categories [30]. For instance, multiracial individuals come from various racial groupings. Indeed, at least on a biological/genetic level, race and ethnicity are now seen as extremely fluid and nuanced ideas rather than simple categorical attributes. Gender and sexual orientation are subject to very comparable criticism.

2.4. How shall we deal with several sensitive attributes *at the same time*?

Most of the literature on fairness in ML deals with a single sensitive dimension, usually assumed to be binary. Even if this is hardly the case in real-world scenarios, it is true that, when facing with multiple sensitive variables, one could simply repeat the assessment for each of the variables, separately.

However, this may hide forms of *intersectional bias* [31, 32]. Namely, the fact that the decision system may have an equal impact on men and women and an equal impact on black individuals and white individuals but still present significant disparities between black women and white men.

This issue arises whenever we assess discrimination with respect to groups: achieving some form of parity at the group level may hide disparities *within* the group, e.g. at the intersection of multiple sensitive characteristics. Kearns et al. [33] call this phenomenon *fairness gerrymandering*. It is true that this drawback might be easily resolved, in principle, by conducting a fairness evaluation on all potential sensitive group intersections. Unfortunately, this is not a very practical solution, since the number of subgroups to take into account increases exponentially as more dimensions are added, and similarly, the number of data samples in each subgroup rapidly decreases, making it extremely difficult to draw any reliable statistical conclusion [32, 34].

Roy et al. [32] further underline that the multi-dimensional aspect of discrimination exacerbates the already mentioned issues of aggregation and categorisation of individuals into groups, and is of course very much connected to the choice of protected groups.

Open Point 4 (Intersectional bias)**.** *Is it fair enough to evaluate unjust discrimination with respect to* ***sensitive attributes separately****? If not, which* ***combinations of sensitive characteristics*** *should we give priority to (given that we cannot realistically hope to assess all possible combinations)?*

3. What do we mean by unfair discrimination?

Clarifying what we mean in concrete terms when we state that a particular group of individuals is discriminated against is probably the single most important piece that is still missing. Quantitative

studies to assess the presence of unjust discrimination have naturally turned mainly to legal literature and anti-discrimination legislative frameworks to come up with concrete mechanisms underlying unfair practices in decision-making.

The legal debate on the topic, as well as the legislative provisions, is highly diversified, and in various respects still very much open. Moreover, the concept of discrimination has deep roots in both political and moral philosophy, with arguments that go beyond the focus on how to fairly compare people belonging to different social groups, and rather discuss on what grounds individuals should be compared at all, with debates going as far as discussing merit and desert as true drivers of individual success.

In light of this, it should come as little surprise that the need for fair algorithms has a variety of nuances and complexities, many originating from very fundamental and hotly contested conceptions, which we can barely hope to be met at all.

In the following section, we summarise the basic legal notions that have guided most of the literature on quantitative assessment of unjust discrimination of protected groups in ADM.

3.1. Direct and indirect discrimination

The key distinction when analysing the concept of discrimination in legal situations is arguably that between *direct discrimination* and *indirect discrimination* [35]. Generally speaking, an indirect discrimination occurs when the decision depends on a sensitive characteristic through other variables. An example can provide a straightforward intuition. Suppose that you have to develop a decision system to support credit lending based on available applicants' characteristics. Among these characteristics there are, e.g., the level of income, the requested amount, the record of past loans, and the gender attribute. The system will show a direct discrimination with respect to gender if it is aware of the gender attribute, and uses it to estimate the optimal outcome. For example, an ML model that is given access to historical data on loan applications and their (non-)repayment, may learn a statistical association between gender and loan repayments, thus effectively assigning a bonus/malus weight to the gender attribute of the applicant. Conversely, if the system is not aware of the gender attribute, or —more precisely— if it does not assign any contribution to the gender attribute itself, the system is not showing a direct discrimination. However, it may nevertheless undergo indirect discrimination with respect to gender since it makes decisions on the basis of the level of income, which is associated with gender. In the last case, the system shows an indirect discrimination with respect to gender, mediated by the level of income.

The direct vs. indirect distinction for discrimination is definitely not the only one proposed in the literature or present in legislative frameworks. This approach is mainly found in the European Union and U.K. laws —see, e.g., Directive 2006/54/EC and Directive 2000/43/EC, and also [38]— while U.S. anti-discrimination laws rely on a similar distinction between *disparate treatment* and *disparate impact*, the former representing forms of explicit discrimination while the latter encompassing a broader set of indirect mechanisms [see, e.g., 26, 25].

Another well known distinction is the one made within the framework of the theory of *equality of opportunity*, i.e. the idea supporting the need to 'level the playing field' for all individuals: different degrees of successes and failures of individuals are fairly distributed only when they "play" on a field without "slopes" that may advantage some with respect to others. Experts identify a wide range of possibilities, from a *formal equality of opportunity*, proposing the avoidance of explicit discrimination of protected groups, to forms of more and more *substantive equality of opportunity*, supporting the avoidance of more subtle and indirect mechanisms of discrimination (and sometimes equated to the *equality of outcome*) [38, 39]. Unfortunately, as it is the case for most of the concepts in discrimination theory, the level of consensus about what equality of opportunity actually means in real-world scenarios is incredibly low. Quoting Bevir [40]:

> *There is widespread agreement that equality of opportunity is a good thing, even a constituent of a just society, but very little consensus on what it requires. Defenders of equality of opportunity suppose that it requires people to be able to compete on equal terms, on a "level playing field,"*

> *but they disagree over what it means to do so. They believe that equality of opportunity is compatible with, and indeed justifies, inequalities of outcome of some sort, but there is considerable disagreement over precisely what degree and kind of inequalities it justifies and how it does so.*

Against this background, while it is widely recognised that direct discrimination is hardly acceptable, the debate on indirect discrimination is much more nuanced, as we shall also discuss later on. Moreover, two factors, among others, contribute to make the distinction at times quite loose and difficult to maintain: 1. a system may not explicitly use sensitive characteristics, but still produce an indirect discrimination through variables that are *per se* not sensitive, but very much associated to sensitive information; 2. oftentimes, one way to avoid indirect discrimination is to provide favourable conditions to the otherwise unfavoured group. Unfortunately, using dual standards on the basis of sensitive characteristics is precisely what direct discrimination means.

Factor 1 is the well known problem of *proxy variables*, that may be exploited —possibly with an explicit discriminatory intent— in order to obtain a discriminatory outcome while maintaining a formal absence of direct discrimination. Ingold and Spoer [41] describe the situation of the Amazon same-day delivery service in 6 metropolitan areas in the U.S.: they document a strong correlation between neighbourhoods not reached by the same-day delivery service and the presence of a significant majority of black residents. The most striking example is that of Boston, where the very central neighbourhood of Roxbury (with a 59 percent presence of black residents) was not granted the same-day delivery service, while all the areas surrounding Roxbury are eligible for the service.[2] Amazon declared that demographics of neighbourhoods played no role in their choices. In cases like this, it is indeed difficult to disentangle direct and indirect forms of discrimination, and agree on what we should and should not tolerate as fair practices.

The second factor is an example of the so-called *affirmative* (or *positive*) *action*, i.e. providing more favourable conditions to groups of individuals traditionally discriminated against. The other side of the coin is known as *reverse discrimination*, i.e. maintaining dual standards not justified by task-related skills and characteristics [42].

As noted previously, the landscape is complex and extremely diverse, and a thorough discussion is well beyond the scope of this work. However, we want to make clear that there is no consensus in the scientific communities on what mechanisms actually constitute an unjust discrimination, particularly regarding indirect forms and their potential tension with the direct ones.

In light of this, we state the following:

Open Point 5 (Direct and indirect discrimination)**.** *When assessing unfair discrimination, how should developers of ADM systems decide which direct or indirect influences of sensitive characteristics on the outcome to address, given that it is generally impossible to eliminate all of them at once? Is it acceptable to* ***engage in direct discrimination in order to prevent indirect discrimination****?*

3.2. What characteristics represent a source of objective justification for discrimination?

Open Point 5 is largely intertwined with the discussion on how to deal with indirect discrimination, once we agree that also indirect discrimination may be unacceptable in some cases.

The focus of the debate is sometimes presented as the identification of 'business necessities'. With this term, one loosely refers to the possible presence of characteristics that are fundamentally relevant to the target estimation, therefore their use to perform ADM is justifiable even when such characteristics bear some dependence on sensitive dimensions. The problem is how to decide when a variable is eligible to represent a business necessity.

[2] As explicitly declared in [41], Amazon later extended its same-day delivery service to the entire metropolitan area of Boston, as well as to that of New York and Chicago.

Legislative frameworks make explicit reference to this possibility. For instance the Directive 2006/54/EC on non-discrimination of men and women in matters of employment and occupation, has provisions of objective justification already in the very definition of indirect discrimination (art 2,1(b))

> *'indirect discrimination': where an apparently neutral provision, criterion or practice would put persons of one sex at a particular disadvantage compared with persons of the other sex, unless that provision, criterion or practice is objectively justified by a legitimate aim, and the means of achieving that aim are appropriate and necessary*

The key ingredients to have an objective justification for indirect discrimination are here identified to be a "legitimate aim", and "appropriate and necessary means" to achieve it.

Similar arguments can be found in U.S. laws. The US Civil Rights Act on non-discrimination in the workplace, among the criteria in support of the presence of unjust discrimination, talks about the case in which "the respondent fails to demonstrate that the challenged practice is job related for the position in question and consistent with business necessity". Here "job-relatedness" and "business necessity" are cited as legitimate reasons to have indirect discrimination.

Barocas and Selbst [26] discuss in details the evolution and oscillation of the interpretation of such terminology by the U.S. Supreme Court, somehow emphasising the ambiguity of it. The point is that all these concepts are inherently fuzzy. One may interpret a characteristic to be job related if it correlates with job performance, and similarly job performance may be considered a legitimate aim.

Albach and Wright [44] present the results of extensive surveys (2157 Amazon Mechanical Turk workers) on different domains to understand what are the perceived drivers of fairness in Machine Learning applications. Interestingly, they find that 'relevance' and 'increases accuracy' are the 2 dimensions that explain most of the perceived fairness of a given variable.

These results seem to support the loose interpretation of business necessities as any variable that can increase predictive performance. But notice that this interpretation strongly undermines the very concept of indirect discrimination: if all the variables that correlate with the outcome can be used legitimately irrespective of their dependence on sensitive characteristics, then we hardly have indirect discrimination at all.

Against this background, practitioners who need to assess and possibly mitigate any unjust discrimination when developing an ADM system face the following:

Open Point 6 (Objective justification)**.** *What qualities should an attribute have, if any, to be **eligible for use** in automatic judgements, even if it serves as a basis for indirect discrimination?*

3.3. The risk of over-reliance on fairness metrics

Sometimes, the various forms of unjust discrimination that we have encountered so far are associated with specific *observational metrics*. With this term, we mean quantities that can be computed by having access to samples of the joint distribution $P(X, A, Y, S)$, where the variable A represents the protected attribute, X denotes the set of other (non-sensitive) characteristics, Y is the target variable, and S is the model outcome.

A discussion of fairness metrics is out of scope in this study —there is plenty of literature on the topic [see, e.g., 27, 45, and references therein]— we here just recall the main classes of observational metrics [25]: *independence* ($S \perp\!\!\!\perp A$); *conditional independence* ($S \perp\!\!\!\perp A \mid X'$, with $X' \subseteq X$); *separation* ($S \perp\!\!\!\perp A \mid Y$); *sufficiency* ($Y \perp\!\!\!\perp A \mid S$).

Independence is usually associated with disparate impact: the fact that the model outcome bears no dependence on the sensitive attribute, directly translates into a substantial parity in the outcomes. In the simple case of binary classifications, where $S \in \{0, 1\}$,independence is indeed equivalent to what is known as *Demographic Parity* (or *Statistical Parity*), namely the equality of acceptance rate among all demographic groups along the sensitive dimension.

The introduction of business needs is sometimes associated with conditional independence and sometimes with separation. The idea behind these associations is the following: if we agree that

some variables (that we collectively label with $X' \subseteq X$) are acceptable grounds to make decisions irrespective of their dependence on sensitive attributes (i.e., they represent business necessities), then the natural quantity to inspect is the independence of the outcome on sensitive attributes *stratified* by those variables. Namely, if we agree that the use of the level of income for credit decisions is justified by business needs, then we may require to have equality of acceptance rate between, say, male and female applicants *within the same level of income*. In other words, we are requiring that the only gender disparity we are willing to tolerate is the one justified by the level of income. In a seminal paper on Conditional Demographic Parity, Wachter et al. [22] suggest the use of this type of metric in relation to the EU non-discrimination legal framework.

Similarly, in the case of separation metrics, we allow only disparities justified by the target variable: e.g., we tolerate gender disparities as long as they are justified by actual disparities in repayment rates. This is in line with the previous discussion on business necessities: if business needs can be justified on grounds of evidence of predictive performance, then metrics of the separation class are indeed more appropriate.

While these arguments are valid, they unfortunately conceal certain possible weaknesses that we will cover in the reminder of the section.

The label problem Many of the statistical metrics (namely, those of the separation and sufficiency classes) rely on the comparison between model outcomes (S) and target labels (Y). Unfortunately, relying on labels is delicate for (at least) two types of drawbacks. On the one hand, we have a problem of actual *access* to labels: e.g., when granting loans, there is no information about applicants that do not receive loans in the first place (namely, access to Y is possible only for applicants with positive outcome S). Moreover, even when we do have access to labels, oftentimes there may be a significant delay, as it is the case for repayment of loans, or for job performance in recruiting decision-making: in such scenarios we can monitor labels (and thus fairness metrics based on those labels) only considerably *ex-post*. On the other hand, there may be an issue of *label bias* [46]: this happens when seemingly effective proxies for ground truth are chosen as target, but there are mechanisms (often involved and ignored by the developer and/or the decision-maker) by which those proxies embed forms of bias.

The incompatibility problem It is well-known that statistical metrics are generally not mutually compatible [47, 25]. For example, one cannot hope to have an ADM system presenting no Demographic Disparities and, at the same time, having equal error rates for different (sensitive) groups of individuals. This is true for all the classes of metrics listed above, a part for degenerate cases. This evidence implies that, when using such metrics, one must make choices, that are ultimately associated with the perspective one has on unjust discrimination. Some works have proposed guidelines —usually in the form of decision trees or diagrams— to help finding the most appropriate statistical metric given domain-specific constraints [see, e.g., 48, 49, 50, 51]. However, as the authors of such works clearly acknowledge, the process of following the proposed decision diagrams is itself complicated, involving necessarily multi-disciplinary competencies, and in any case they warn not to take these diagrams too categorically or as a set of well-established prescriptions. To make things even more blurry, if it is true that having perfect parity with respect to different metric classes is mathematically impossible, allowing a limited level of disparity may be attainable with multiple metrics at the same time [52], suggesting that focusing too much on a single metric maybe counter-productive after all.

The threshold problem Once practitioners have computed the values of observational metrics, they face the problem of "deciding" whether the numbers found are enough to raise a warning or not. A 10% difference in acceptance rate between male and female is fair enough? What about using the ratio of acceptance rates instead of the difference? Borrowing from Ruggieri et al. [12] "The apparently innocuous choice between the algebraic operators (1) or (2) [difference vs ratio], however, has an enormous impact on how decisions are affected." The threshold problem is as trivial as it sounds: statistical metrics are real-valued measures of (some notion of) disparities, they are not a binary trigger.

Some attempts to face this issue can be found in the U.S. context. One is the well-known *four-fifths rule* [53], stating that an acceptance rate (in recruiting setting) of the 'unfavoured' group less than 80% of the acceptance rate of the 'favoured' group constitutes an evidence of disparate impact.

Lack of causality Another drawback of purely observational metrics is that they are blind with respect to the underlying generation mechanisms: it is well known that statistical correlations among variables may have very different natures. They may be due to a direct causation of one variable on another; to causation mediated by other variables; or may be spurious, i.e. due to some common factor causing both of them; or may even be spuriously created when conditioning on a common effect [54, 55].

Notice, incidentally, that while on the one hand causality is an important ingredient to make the fairness assessment more reliable, on the other hand it does not come without ambiguities and attention points: we refer to a longer (preprint) version of this analysis for details on the topic of causality in fair-AI and its associated open points [56].

A vivid critic of the "metric" approach to fairness can be found in [16], where the authors face the fair-AI problem from the perspective of Science and Technology Studies. In particular, they talk about a *Formalism Trap*: fairness and discrimination are complex concepts, that contain many subtle dimensions, among which a *procedural*, a *contextual*, and a *contestable* dimension, that cannot be grasped by relatively simple statistical quantities. Of course, this does not imply that statistical measures are altogether wrong or useless, but only that they must be utilised as tools, to be set in a wider context.

Indeed, survey studies about fairness *perception* in ML show a very diverse and nuanced landscape, where perspectives on fairness are associated with socio-demographic factors of the participants [57], and consensus on fairness notions is very low [58, 59].

Overall, considering together all these aspects about observational metrics, we can state the following:

Open Point 7 (Over-reliance on observational metrics)**.** *Purely observational fairness metrics should be taken with a grain of salt. At best, they can be used* ***as a means for a deeper reasoning*** *on the mechanisms underlying a phenomenon, rather than a final word on the presence or lack of unjust discrimination.*

4. Conclusion

This work highlights that the main weaknesses in addressing non-discrimination in ADM systems lie in the gap between abstract fairness principles and their practical application, and while the literature has proposed remedies — ranging from fairness metrics to causal frameworks — we argue these remain limited in scope. We suggest that long-term progress requires interventions beyond technical fixes, particularly cultural measures such as education and training, and structural incentives aimed at addressing inequities ex-ante, since downstream algorithmic adjustments often prove inadequate, especially for indirect discrimination. More broadly, we emphasize the need for a multidisciplinary and multi-stakeholder framework involving ethicists, sociologists, legal experts, policymakers, providers, and users, to ensure trustworthy governance, as data scientists alone cannot address ethical and societal dimensions. Unlike narrow procedural recommendations, which remain immature and often overly reliant on statistical metrics, we see these "soft interventions" as a necessary foundation for future developments. Finally, we propose exploring the idea of setting domain-specific lists of "legitimate" variables as an alternative to both Fairness Through unawareness strategies and reliance on fairness metrics, though this would require policy consensus and ongoing monitoring to mitigate residual risks.

5. Acknowledgments

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. Grant Agreement no. 101120763 - TANGO.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Z. C. Lipton, J. Steinhardt, Troubling trends in machine learning scholarship, arXiv preprint arXiv:1807.03341 (2018).

[2] The European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

[3] The European Parliament and the Council of The European Union, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence, 2024. URL: http://data.europa.eu/eli/reg/2024/1689/oj.

[4] White House Office of Science and Technology Policy , Blueprint for an AI Bill of Rights: Making Automated Systems Work For The American People, 2022. https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

[5] US Congress, National AI Commission Act (2023). URL: https://www.congress.gov/bill/118th-congress/house-bill/4223?s=1&r=1.

[6] Department for Science, Innovation & Technology, A pro-innovation approach to AI regulation, 2023. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach.

[7] gov.uk, Prime Minister launches new AI Safety Institute, gov.uk: Press Release (2023). URL: https://www.gov.uk/government/news/prime-minister-launches-new-ai-safety-institute.

[8] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018. URL: https://arxiv.org/abs/1810.01943.

[9] H. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, M. Madaio, Fairlearn: Assessing and improving fairness of ai systems, 2023. arXiv:2303.16626.

[10] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson, The what-if tool: Interactive probing of machine learning models, IEEE Transactions on Visualization and Computer Graphics 26 (2020) 56–65. doi:10.1109/TVCG.2019.2934619.

[11] M. S. A. Lee, J. Singh, The landscape and gaps in open source fairness toolkits, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, Association for Computing Machinery, New York, NY, USA, 2021. doi:10.1145/3411764.3445261.

[12] S. Ruggieri, J. M. Alvarez, A. Pugnana, L. State, F. Turini, Can We Trust Fair-AI?, Proceedings of the AAAI Conference on Artificial Intelligence 37 (2023) 15421–15430. doi:10.1609/aaai.v37i13.26798.

[13] N. A. Saxena, W. Zhang, C. Shahabi, Missed Opportunities in Fair AI, 2023, pp. 961–964. doi:10.1137/1.9781611977653.ch110.

[14] M. Buyl, T. De Bie, Inherent limitations of ai fairness, Commun. ACM 67 (2024) 48–55. URL: https://doi.org/10.1145/3624700. doi:10.1145/3624700.

[15] M. Dolata, S. Feuerriegel, G. Schwabe, A sociotechnical view of algorithmic fairness, Information Systems Journal 32 (2022) 754–818. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/isj.12370. doi:https://doi.org/10.1111/isj.12370. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/isj.12370.

[16] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, J. Vertesi, Fairness and abstraction in sociotechnical systems, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 59–68. doi:10.1145/3287560.3287598.

[17] A. F. Cooper, E. Abrams, N. NA, Emergent unfairness in algorithmic fairness-accuracy trade-off research, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 46–54. doi:10.1145/3461702.3462519.

[18] A. L. Hoffmann, Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse, Information, Communication & Society 22 (2019) 900–915. doi:10.1080/1369118X.2019.1573912.

[19] U. Aivodji, H. Arai, O. Fortineau, S. Gambs, S. Hara, A. Tapp, Fairwashing: the risk of rationalization, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 161–170. URL: https://proceedings.mlr.press/v97/aivodji19a.html.

[20] The European Parliament, the Council and the Commission, Charter of Fundamental Rights of the European Union, 2012. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A12012P%2FTXT.

[21] The European Commission, Directorate-General for Communication, Non-discrimination, –. URL: https://commission.europa.eu/aid-development-cooperation-fundamental-rights/your-rights-eu/know-your-rights/equality/non-discrimination_en, accessed: 2023-08-22.

[22] S. Wachter, B. Mittelstadt, C. Russell, Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai, Computer Law & Security Review 41 (2021) 105567. doi:https://doi.org/10.1016/j.clsr.2021.105567.

[23] The European Parliament and the Council of The European Union, Directive 2008/48/EC of the European Parliament and of the Council of 23 April 2008 on credit agreements for consumers and repealing Council Directive 87/102/EEC, 2008. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32008L0048.

[24] The European Commission, Proposal for a Directive of the European Parliament and of the Council on consumer credits, 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0347.

[25] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, fairmlbook.org, 2019. http://www.fairmlbook.org.

[26] S. Barocas, A. D. Selbst, Big data's disparate impact, California law review (2016) 671–732. doi:10.15779/Z38BG31.

[27] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, A. C. Cosentini, A clarification of the nuances in the fairness metrics landscape, Scientific Reports 12 (2022) 4209. doi:10.1038/s41598-022-07939-1.

[28] M. J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual fairness, Advances in neural information processing systems 30 (2017). URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/1271a7029c9df08643b631b02cf9e116-Abstract.html.

[29] U.S. Government Publishing Office, Equal Credit Opportunity Act of 1974, 1974. URL: https://www.govinfo.gov/content/pkg/USCODE-2011-title15/html/USCODE-2011-title15-chap41-subchapIV.htm, United States Code, Title 15. Chapter 41, Subchapter IV.

[30] C. Lu, J. Kay, K. McKee, Subverting machines, fluctuating identities: Re-learning human categorization, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 1005–1015. doi:10.1145/3531146.3533161.

[31] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: S. A. Friedler, C. Wilson (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency, volume 81 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 77–91. URL: https://proceedings.mlr.press/v81/buolamwini18a.html.

[32] A. Roy, J. Horstmann, E. Ntoutsi, Multi-dimensional discrimination in law and machine learning - a comparative overview, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 89–100. doi:10.1145/3593013.3593979.

[33] M. Kearns, S. Neel, A. Roth, Z. S. Wu, Preventing fairness gerrymandering: Auditing and learning for subgroup fairness, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 2564–2572. URL: https://proceedings.mlr.press/v80/kearns18a.html.

[34] Y. Kong, Are "intersectionally fair" ai algorithms really fair to women of color? a philosophical analysis, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 485–494. doi:10.1145/3531146.3533114.

[35] A. Altman, Discrimination, in: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Winter 2020 ed., Metaphysics Research Lab, Stanford University, 2020.

[36] The European Parliament and the Council of The European Union, Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast), 2006. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32006L0054.

[37] The Council of The European Union, Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, 2000. URL: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32000L0043:en:HTML.

[38] C. Barnard, B. Hepple, Substantive equality, The Cambridge Law Journal 59 (2000) 562–585. doi:10.1017/S0008197300000246.

[39] G. Elford, Equality of Opportunity, in: E. N. Zalta, U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy, Fall 2023 ed., Metaphysics Research Lab, Stanford University, 2023.

[40] M. Bevir, Equality of Opportunity, in: Encyclopedia of Political Theory, SAGE Publications, Inc., 2010. doi:10.4135/9781412958660.

[41] D. I. Ingold, S. Spoer, Amazon doesn't consider the race of its customers. should it?, Bloomberg (2016). URL: https://www.bloomberg.com/graphics/2016-amazon-same-day.

[42] Z. Lipton, J. McAuley, A. Chouldechova, Does mitigating ml's impact disparity require treatment disparity?, Advances in neural information processing systems 31 (2018). URL: https://proceedings.neurips.cc/paper_files/paper/2018/hash/8e0384779e58ce2af40eb365b318cc32-Abstract.html.

[43] U.S. Government Publishing Office, Civil Rights Act of 1964, 1964. URL: https://www.govinfo.gov/app/details/COMPS-342, Public Law 88–352; 78 Stat. 241, as Amended Through P.L. 114–95, Enacted December 10, 2015.

[44] M. Albach, J. R. Wright, The role of accuracy in algorithmic process fairness across multiple domains, in: Proceedings of the 22nd ACM Conference on Economics and Computation, EC '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 29–49. doi:10.1145/3465456.3467620.

[45] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, K. Lum, Algorithmic fairness: Choices, assumptions, and definitions, Annual Review of Statistics and Its Application 8 (2021) 141–163. doi:10.1146/annurev-statistics-042720-125902.

[46] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science 366 (2019) 447–453. doi:10.1126/science.aax2342.

[47] J. Kleinberg, S. Mullainathan, M. Raghavan, Inherent trade-offs in the fair determination of risk scores, 2016. arXiv:1609.05807.

[48] K. Makhlouf, S. Zhioua, C. Palamidessi, On the applicability of machine learning fairness notions, SIGKDD Explor. Newsl. 23 (2021) 14–23. doi:10.1145/3468507.3468511.

[49] K. Makhlouf, S. Zhioua, C. Palamidessi, Machine learning fairness notions: Bridging the gap with real-world applications, Information Processing & Management 58 (2021) 102642. doi:https://doi.org/10.1016/j.ipm.2021.102642.

[50] M. A. Haeri, K. Hartmann, J. Sirsch, G. Wenzelburger, K. A. Zweig, Promises and pitfalls of algorithm use by state authorities, Philosophy & Technology 35 (2022) 33. doi:10.1007/s13347-022-00528-0.

[51] J. J. Smith, L. Beattie, H. Cramer, Scoping fairness objectives and identifying fairness metrics

for recommender systems: The practitioners' perspective, in: Proceedings of the ACM Web Conference 2023, WWW '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 3648–3659. doi:10.1145/3543507.3583204.

[52] A. Bell, L. Bynum, N. Drushchak, T. Zakharchenko, L. Rosenblatt, J. Stoyanovich, The possibility of fairness: Revisiting the impossibility theorem in practice, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 400–422. doi:10.1145/3593013.3594007.

[53] Equal Employment Opportunity Commission, Uniform Guidelines on Employment Selection Procedures, 2015. URL: https://www.govinfo.gov/content/pkg/CFR-2011-title29-vol4/xml/CFR-2011-title29-vol4-part1607.xml, 29 Code of Federal Regulation §1607.4(D).

[54] J. Peters, D. Janzing, B. Schölkopf, Elements of Causal Inference, The MIT Press, Cambridge, 2017.

[55] B. Neal, Introduction to causal inference, Course Lecture Notes (draft) (2020).

[56] D. Regoli, A. Castelnovo, N. Inverardi, G. Nanino, I. Penco, Fair enough? a map of the current limitations of the requirements to have fair algorithms, 2024. URL: https://arxiv.org/abs/2311.12435. arXiv:2311.12435.

[57] N. Grgić-Hlača, G. Lima, A. Weller, E. M. Redmiles, Dimensions of diversity in human perceptions of algorithmic fairness, in: Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO '22, Association for Computing Machinery, New York, NY, USA, 2022. doi:10.1145/3551624.3555306.

[58] C. Starke, J. Baleis, B. Keller, F. Marcinkowski, Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature, Big Data & Society 9 (2022) 20539517221115189. doi:10.1177/20539517221115189.

[59] G. Harrison, J. Hanson, C. Jacinto, J. Ramirez, B. Ur, An empirical study on the perceived fairness of realistic, imperfect machine learning models, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 392–402. doi:10.1145/3351095.3372831.