# 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 recent years, the increase in the usage and efficiency of Artificial Intelligence and, more in general, of Automated Decision-Making systems (ADM) 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. This awareness has on the one hand encouraged several scientific communities to come up with more and more appropriate ways and methods to assess, quantify, and possibly mitigate such biases and disparities. On the other hand, it 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. Namely, there is a hiatus between what the society is demanding from ADM, and what this demand actually means in real-world scenarios. In this work, we outline the key features of such a hiatus and identify 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 ADM.

## Keywords

Fairness, Bias, Artificial Intelligence, Machine Learning

The arguments raised in what follows, and detailed in a longer version of this work [1], find their place within a relatively recent stream of research that is critically focused on the general topic of fairness in ADM [2, 3, 4, 5, 6, 7, 8]. 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 over-reliance 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. [9] has evocatively named the latter *Fairwashing*.[1]

Even if most of the ambiguities and attention points that we detail in Regoli et al. [1] and briefly introduce here have already been discussed 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 identify a set of *open points* that constitute obstacles both for researchers in the field of AI

*AIxIA 2024 Discussion Papers - 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25–28, 2024*

*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)



[1]Aivodji et al. [9] 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.

**Table 1**
**Schematic summary** of current open points of fairness requirements for ADM.

| broad aspect | Open Point |
|---|---|
| Sensitive attributes choice | 1: Identification of protected groups<br>2: Sensitive data collection<br>3: Group aggregation<br>4: Intersectional bias |
| What do we mean by unfair discrimination? | 5: Direct and indirect discrimination<br>6: Objective justification<br>7: Over-reliance on observational metrics<br>8: Assumptions of causal structure |

and for practitioners and developers of ADM systems to meet the societal requirement of having fair algorithms.

In particular, we build our critical analysis of fair-AI by distinguishing two broad aspects that are roots of several ambiguities:

1. **Choice of sensitive attributes**: 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. (so-called sensitive or protected attributes)
2. **What is the true meaning of 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.

In Table 1, we summarise the key points raised throughout the paper.

## 1. Choice of sensitive attributes

In a nutshell, we claim that there is a fundamental ambiguity about what characteristics should be considered as protected for a given case and domain. High level non-discrimination principles can be found in several legislative frameworks, such as art 21,1 of the EU Charter of Fundamental Rights. But there is no clear specific legislation that organically discuss what individual characteristics should be considered *protected*. As an example, Table 2 tries to summarise the European non-discrimination laws, most of which consider specific domains of application and specific protected characteristics, thus making it unclear what to do with other potentially sensitive features and other domains. Something not dissimilar can be found in the US legislation; we refer to Barocas et al. [11, chap 6] and Barocas and Selbst [12] for more details. In general, we claim that there is no ethical or legal consensus on what are the dimensions (or the criteria to identify them) with respect to which we should assess and, eventually, mitigate for possible biases and discrimination.

This can be summarised in 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?*

In turn, this ambiguity reveals a number of additional nuanced points. For example, it is unclear whether a *fairness assessment* should be conducted with regard to all protected groups even when the ADM system is not actually collecting such protected information. Indeed, while it is fairly common to

**Table 2**
**Schematic summary** of protected categories explicitly covered by EU Directives on non-discrimination. See also the EU non-discrimination website.

| Directive | year | Domain of application | Protected categories |
|---|---|---|---|
| Race Equality Directive [Directive 2000/43/EC] | 2000 | employment, social protection, healthcare, education, access to and supply of goods and services which are available to the public | race and ethnic origin |
| Employment Directive [Directive 2000/78/EC] | 2000 | working environment | religion or belief, disability, age, sexual orientation |
| Gender Access Directive [Directive 2004/113/EC] | 2004 | access to and supply of goods and services | gender |
| Gender Equality Directive [Directive 2006/54/EC] | 2006 | employment | gender |

record information about gender or age, it is much less common to collect data about political opinions or religious belief.

**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?*

Moreover, there is ambiguity regarding the very definition of (protected) *group*. For instance, age can be aggregated in multiple ways, and assessing fairness with respect to different aggregations can potentially lead to very different results. On a more abstract level, there are concerns and discussions about the prospect of placing people in rigid and exclusive categories [13]. 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.

**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?*

Finally, even if there are some proposals in the literature on how to fix this, there is still no consensus on how to deal with the exponentially growing number of subgroups to take into account when considering the intersection of several protected attributes at the same time [14, 15]:

**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)?*

## 2. What's the true meaning of unfair discrimination?

We believe that the crucial fuzziness around fair-AI lies in the fact that there is no consensus regarding what does it mean to unfairly discriminate a group of people with respect to others. Both legislative and ethical literature make the distinction between **direct and indirect discrimination**,[2] the former indicating an explicit use of protected characteristics to make a decision, while the latter being a discrimination through characteristics somehow associated to protected ones, but not protected *per se*. A possible example of indirect gender discrimination may be that of using income as a variable to make decisions on loan approvals, given that income is a variable typically correlated to gender.

[2]Direct vs. indirect discrimination is more common in EU legislative frameworks — see, e.g., Directive 2006/54/EC and Directive 2000/43/EC, and also [21] — while U.S. anti-discrimination laws rely on a similar distinction between *disparate treatment* and *disparate impact* [see, e.g., 12, 11].

However useful, this distinction raises a set of open problems, the first reflected in the fact that it is not always possible to avoid both direct and indirect discrimination, and that some kind of balance should be somehow tolerated or even desirable (think, e.g., of *affirmative action*, which is indeed a strategy to remove indirect discrimination through explicit — i.e., direct — discrimination):

**Open Point 5 (Direct vs Indirect discrimination).** *When evaluating unjust discrimination, should developers of ADM systems take into account all the potential direct and indirect ways by which sensitive characteristics may have affected the outcome? Is it acceptable to* ***engage in direct discrimination in order to prevent indirect discrimination****?*

In particular, the concept of indirect discrimination is complex and hides many subtleties. In fact, most legislative frameworks actually admit that some characteristics can be legitimately used to make decisions, *even when associated with protected attributes*, since they represent a "business need" (such as the income in the previous example on loan approval).[3]

**Open Point 6 (Legitimate Business needs).** *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****?*

In this respect, we would like to point out that a possible way out for Open Point 6 about legitimate business needs, would be that of identifying *ex-ante* a set of variables as the only *legitimate* features to be used in particularly delicate domains (such as job recruiting). Namely, this can be seen as the counterpoint of Fairness Through Unawareness, or *Blindness* [see, e.g. 23]. Blindness consists in building ADM systems that are not exposed directly to sensitive attributes. While this strategy prevents the possibility of direct discrimination, it leaves room for indirect discrimination through the use of variables associated to sensitive characteristics. Given this context, rather than offering a list of unusable attributes, we could offer a list of attributes that are the *only allowed* for a particular domain/case, assuming that we consider those attributes *relevant* for the task at hand and will therefore accept any discrepancies that may result from the association of such variables with sensitive attributes. [4]

In order to capture different concepts of (direct and indirect) discrimination, the fair-AI literature has developed a wide range of observational metrics [25, 26]. However, on the one hand it has been proved that most of such metrics are mutually incompatible [26], and on the other hand metrics that are solely observational are blind to the real underlying mechanism and only provide a static picture of an often very intricate phenomenon. For a given use-case, choosing only one observational metric is, at best, exceedingly challenging and most likely simplistic. 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., 27, 28, 29, 30]. 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 [31], suggesting that focusing too much on a single metric maybe counter-productive after all. Furthermore, employing such metrics creates more uncertainty about the numerical threshold that indicates the true existence of unfairness, or even about the specific form of metric that should be used (e.g., taking ratios vs. differences of significant quantities — see, e.g., Ruggieri et al. [2]).

**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. A clear connection between quantitative metrics and unjust discrimination is still missing.*

[3] See, e.g., Directive 2006/54/EC that explicitly refers to "objective justification", or US Civil Rights Act that talks about "business necessities".

[4] Notice that this a notion of *process fairness*, and is similar, in spirit, to the "feature-apriori fairness" introduced by Grgic-Hlaca et al. [24].

One of the problems of observational metrics is that they are blind to the causal structure of the underlying phenomenon, with the risk of attributing to bias and discrimination what is instead spurious correlation. Indeed, is true that using a causality-aware approach makes it possible to transparently disentangle between direct, indirect, and spurious effects, as brilliantly showcased by Plečko and Bareinboim [32].

However, identifying the causal structure of a given use case is far from straightforward. Moreover, there is an open philosophical debate regarding the very notion of human attributes as *causes* [33]. Therefore, even if we welcome the causal analysis of the underlying phenomenon in order to better assess for the presence of bias and discrimination, we would like to raise the attention on the following point:

**Open Point 8 (Assumptions of causal structure).** *Causal tools require strong, often unverifiable, assumptions. Downstream consequences of wrong assumptions can lead to wrong or even harmful actions. Therefore, particular care must be taken when relying on such tools.*

## 3. Conclusions

In this work, we support the view that the understanding of the landscape of unjust discrimination in ADM, despite the impressive work done in the last decade or so by the fair-AI community, is not yet mature enough to be "put into practice". In particular, there are gaps at the intersection between the mathematical and statistical tools developed by statisticians and AI researchers, the legal non-discrimination provisions, and the ethical and social notions of fairness. We believe these gaps are deep, and not so easy to bridge, in part precisely because they lie at the boundaries between quite different worlds.

We believe that the requirement to develop fair algorithms is still too vague and that, in order to be put in place, we have to clarify and be aware of a set of open points, most of which are societal in nature rather than technical. Given the strong multidisciplinary content, this challenge will be best addressed and discussed when researchers and practitioners from the various fields involved (e.g., statisticians, AI experts, ethicists, and legal experts) collaborate toward the goal, potentially creating a shared vocabulary and set of working notions. In fact, this is the spirit that also guided this work.

As a final remark, notice that there are of course other problematic aspects of unjust discrimination in ADM systems that are somehow out of the scope of this work and would have required a much broader analysis. We can cite, e.g., how to face the challenge of bias in models with unstructured data such as images and text, especially in generative AI models such as the modern Large Language Models; or more technological problems, such as those raised by the modularity of AI systems, that are usually composed of several steps and components, making it complicated to clarify how fairness issues may propagate through the process.

## 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.

## References

[1] 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.

[2] 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.
[3] N. A. Saxena, W. Zhang, C. Shahabi, Missed Opportunities in Fair AI, 2023, pp. 961–964. doi:10.1137/1.9781611977653.ch110.
[4] 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.
[5] 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.
[6] 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.
[7] 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.
[8] 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.
[9] 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.
[10] 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.
[11] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, fairmlbook.org, 2019. http://www.fairmlbook.org.
[12] S. Barocas, A. D. Selbst, Big data's disparate impact, California law review (2016) 671–732. doi:10.15779/Z38BG31.
[13] 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.
[14] 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.
[15] 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.
[16] 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.
[17] 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.
[18] The Council of The European Union, Council Directive 2000/78/EC of 27 November 2000 es-

tablishing a general framework for equal treatment in employment and occupation, 2000. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32000L0078.

[19] The Council of The European Union, Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services, 2004. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32004L0113.

[20] 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.

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

[22] 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.

[23] 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.

[24] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, A. Weller, The case for process fairness in learning: Feature selection for fair decision making, in: NIPS symposium on machine learning and the law, volume 1, Barcelona, Spain, 2016, p. 11. URL: https://www.mlandthelaw.org/papers/grgic.pdf.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] D. Plečko, E. Bareinboim, Causal fairness analysis: A causal toolkit for fair machine learning, Foundations and Trends® in Machine Learning 17 (2024) 304–589. doi:10.1561/2200000106.

[33] L. Hu, I. Kohler-Hausmann, What's sex got to do with fair machine learning?, arXiv preprint arXiv:2006.01770 (2020). doi:10.48550/arXiv.2006.01770.