# Unification, Extension, and Interpretation of Group Fairness Metrics for ML-Based Decision-Making

Joachim Baumann[1,2,*,†], Corinna Hertweck[1,2,*,†], Michele Loi[3] and Christoph Heitz[2]

[1]University of Zurich, Zurich, Switzerland
[2]Zurich University of Applied Sciences, Zurich, Switzerland
[3]Polytechnic University of Milan, Milan, Italy

Abstract

Group fairness metrics are an established way of assessing the fairness of prediction-based decision-making systems. In this paper, we propose a comprehensive framework for group fairness metrics, which links them to a wide array of theories from distributive justice. Our unifying framework reveals the normative choices associated with standard group fairness metrics and allows an interpretation of their moral substance. In addition, this broader view provides a structure for the expansion of standard fairness metrics that we find in the literature. This expansion allows addressing several criticisms of standard group fairness metrics. This short paper presents the papers [1] and [2].

**Keywords**

group fairness, fairness metrics, distributive justice, consequential decision-making, machine learning

## 1. Motivation

Different measures have emerged in the algorithmic fairness literature for assessing unfairness in decision-making systems, many of which are in the category of so-called group fairness criteria. These criteria compare the decisions of a prediction-based decision-making system across socially salient groups (typically as an average over the group members). Popular group fairness criteria demand equality between conditional probabilities that can be derived from the confusion matrix, such as true positive rates parity (aka equality of opportunity [3]).

However, these standard group fairness criteria come with notable limitations: (1) Enforcing equality might yield worse results for *all* groups ("leveling down objection") [4–7]; (2) standard fairness criteria focus on an equal distribution of favorable *decisions* and not on the *consequences* of these decisions [8, 9]; (3) none of the standard group fairness criteria might be morally appropriate in a given context [10]. Several works have therefore suggested extensions of standard group fairness criteria [6, 11–14].

*EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland*
*Corresponding authors.
†These authors contributed equally.
baumann@ifi.uzh.ch (J. Baumann); corinna.hertweck@zhaw.ch (C. Hertweck); michele.loi@polimi.it (M. Loi); christoph.heitz@zhaw.ch (C. Heitz)
0000-0003-2019-4829 (J. Baumann); 0000-0002-7639-2771 (C. Hertweck); 0000-0002-7053-4724 (M. Loi); 0000-0002-6683-4150 (C. Heitz)

However, it is unclear how these extensions of group fairness criteria can be understood in a unified framework,[1] and how group fairness criteria are related to concepts of distributive justice from the philosophical literature. Our paper addresses this gap by proposing a generalized framework for group fairness based on the distributive justice literature. This framework includes all standard group fairness criteria as special cases, allows uncovering their moral assumptions, and extends them to overcome the limitations of standard fairness criteria.

## 2. Comprehensive Group Fairness Framework

Theories of distributive justice are characterized by their answers to the following questions: What is distributed? Between whom is it distributed, and which groups should be compared? And how should it be distributed? [18, 19]. We apply this framework to group fairness of ML-based decision-making systems by posing the following questions:

**Utility of decisions:** ***What is distributed?*** The utility of a decision is the amount of benefit or harm derived from receiving this decision, which is what people have (objective) reasons to desire. Focusing on utility instead of the decision as such allows us to acknowledge that, e.g., a positive decision may not always be beneficial. For example, a positive decision on a loan application may be harmful if the applicant is unable to repay the loan and ends up in debt.

**Relevant groups:** ***Between whom is it distributed?*** Group fairness is concerned with socially salient groups (e.g., defined by gender, race, or disability) as this is what theories of discrimination focus on [20]. We extend this to considering *relevant groups*, which at least have a *weak causal influence* on the prediction or outcome (or both).

**Claim differentiator:** ***Which subgroups should be compared?*** Comparing the relevant groups as such might not always be morally appropriate. For example, equality of opportunity [3] only considers individuals with $Y = 1$. In our framework, we allow for a so-called *claim differentiator*, which differentiates individuals with different claims to the utility. Different claims may be justified, e.g., by differences in deservingness, need, or merit.

**Pattern of justice:** ***How should the utility be distributed?*** A pattern of justice describes how utility should be distributed between the relevant groups. The most widely discussed patterns of justice in political philosophy are egalitarianism [21], maximin [18, 22], prioritarianism [23] and sufficientarianism [24]. All standard group fairness criteria are based on egalitarianism.

### 2.1. Generalized definition of group fairness

Taking these components together, we can formalize a fairness criterion using the expected utilities $\mathbb{E}(U)$ of the relevant groups $a \in A$ with the same claim differentiator $J = j$: $\mathbb{E}(U|J = j, A = a)$. The pattern of justice then specifies what constitutes a just distribution of $\mathbb{E}(U)$ across the relevant (sub)groups $(J = j, A = a)$, i.e., whether we should equalize the expected utilities,

[1] Unifying frameworks have been proposed by Heidari et al. [15], Loi et al. [16], Baumann and Heitz [17]. However, these attempts are restricted to the selection of one of the standard group fairness criteria, which all demand equality between different socio-demographic groups.

maximize a weighted sum of them, etc. Based on this, we propose the following generalized definition of group fairness:

**Group fairness**

Group fairness is the just distribution of utility among groups, as defined by the specification of a utility function, relevant groups, a claim differentiator, and a pattern of justice.

## 3. Conclusion

None of the standard group fairness criteria is morally appropriate in all contexts, and there are even contexts in which none is morally appropriate. To overcome these limitations, we propose a new framework that *extends* the currently discussed approaches of group fairness. Our framework is also a *unification* in that it includes all standard measures of group fairness as special cases. This allows us to uncover the implicit moral assumptions to better *interpret* each of them.

## Acknowledgments

We thank the other members of our project and colleagues (Eleonora Viganò, Ulrich Leicht-Deobald, Serhiy Kandul, Markus Christen, Anikó Hannák, Nicolò Pagan, Stefania Ionescu, Aleksandra Urman, Leonore Röseler, Azza Bouleimen, and Egwuchukwu Ani) for their continuous feedback on the framework presented in this paper. We also thank participants of our algorithmic fairness workshop at the Applied Machine Learning Days (AMLD) at École polytechnique fédérale de Lausanne (EPFL) in Switzerland and the participants of the course "Informatics, Ethics and Society" at the University of Zurich for critical discussions. This work was supported by the National Research Programme "Digital Transformation" (NRP 77) of the Swiss National Science Foundation (SNSF) — grant number 187473 — and by Innosuisse — grant number 44692.1 IP-SBM. Michele Loi was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 898322.

## References

[1] J. Baumann, C. Hertweck, M. Loi, C. Heitz, Distributive justice as the foundational premise of fair ml: Unification, extension, and interpretation of group fairness metrics (2023). URL: http://arxiv.org/abs/2206.02897. arXiv:2206.02897.

[2] C. Hertweck, J. Baumann, M. Loi, E. Viganò, C. Heitz, A justice-based framework for the analysis of algorithmic fairness-utility trade-offs (2023). URL: http://arxiv.org/abs/2206.02891. arXiv:2206.02891.

[3] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, arXiv preprint arXiv:1610.02413 (2016).

[4] D. Parfit, Equality or priority, The Lindley lecture, Department of Philosophy, University of Kansas, 1995.
[5] R. Crisp, Equality, Priority, and Compassion 113 (2003) 745–763. URL: https://doi.org/10.1086/373954. doi:10.1086/373954.
[6] L. Hu, Y. Chen, Fair classification and social welfare, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 535–545.
[7] S. Holm, Egalitarianism and Algorithmic Fairness, Philosophy & Technology 36 (2023) 6. URL: https://doi.org/10.1007/s13347-023-00607-w. doi:10.1007/s13347-023-00607-w.
[8] R. Binns, Fairness in machine learning: Lessons from political philosophy, 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, New York, NY, USA, 2018, pp. 149–159. URL: http://proceedings.mlr.press/v81/binns18a.html.
[9] J. Finocchiaro, R. Maio, F. Monachou, G. K. Patro, M. Raghavan, A.-A. Stoica, S. Tsirtsis, Bridging machine learning and mechanism design towards algorithmic fairness, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 489–503.
[10] M. Kuppler, C. Kern, R. L. Bach, F. Kreuter, Distributive justice and fairness metrics in automated decision-making: How much overlap is there?, 2021. arXiv:2105.01441.
[11] O. Ben-Porat, F. Sandomirskiy, M. Tennenholtz, Protecting the protected group: Circumventing harmful fairness, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 5176–5184.
[12] S. Hossain, A. Mladenovic, N. Shah, Designing fairly fair classifiers via economic fairness notions, in: Proceedings of The Web Conference 2020, 2020, pp. 1559–1569.
[13] N. Martinez, M. Bertran, G. Sapiro, Minimax pareto fairness: A multi objective perspective, in: International Conference on Machine Learning, PMLR, 2020, pp. 6755–6764.
[14] E. Diana, W. Gill, M. Kearns, K. Kenthapadi, A. Roth, Minimax group fairness: Algorithms and experiments, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 66–76.
[15] H. Heidari, M. Loi, K. P. Gummadi, A. Krause, A moral framework for understanding fair ml through economic models of equality of opportunity, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 181–190.
[16] M. Loi, A. Herlitz, H. Heidari, A philosophical theory of fairness for prediction-based decisions, Available at SSRN 3450300 (2019).
[17] J. Baumann, C. Heitz, Group Fairness in Prediction-Based Decision Making: From Moral Assessment to Implementation, in: 2022 9th Swiss Conference on Data Science (SDS), 2022, pp. 19–25. doi:10.1109/SDS54800.2022.00011.
[18] J. Rawls, A Theory of Justice, 2 ed., Harvard University Press, Cambridge, Massachussets, 1999.
[19] A. Sen, Equality of what?, The Tanner lecture on human values 1 (1980) 197–220.
[20] A. Altman, Discrimination, in: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Winter 2020 ed., Metaphysics Research Lab, Stanford University, 2020.
[21] R. Arneson, Egalitarianism, in: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Summer 2013 ed., Metaphysics Research Lab, Stanford University, 2013.
[22] J. Rawls, Justice as fairness: A restatement, Harvard University Press, 2001.

[23] N. Holtug, Prioritarianism, in: Oxford Research Encyclopedia of Politics, 2017.
[24] L. Shields, Sufficientarianism, Philosophy Compass 15 (2020) e12704. URL: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/phc3.12704. doi:https://doi.org/10.1111/phc3.12704.