FairnessLab: A Consequence-Sensitive Bias Audit and Mitigation Toolkit

Corinna Hertweck[1,2,*,†], Joachim Baumann[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

We introduce the FairnessLab: an open-source toolkit including interactive visualizations to facilitate the development of *fair* ML-based decision-making systems. Existing bias audit tools usually just offer standard group fairness metrics, which leads to strong restrictions: neither one is morally appropriate in all contexts, and there are contexts in which none of them is morally appropriate. Building on new findings from computer science and philosophy, the FairnessLab provides a much wider range of metrics, and guides users to generate a fairness measure that is morally appropriate for a given context. Thus, the FairnessLab can be used to define context-specific fairness metrics that are sensitive to the consequences experienced by affected individuals [1, 2]. Furthermore, it includes techniques to mitigate unfairness w.r.t. a specified metric. This empowers data scientists and developers to (i) make their moral choices explicit, (ii) derive appropriate fairness metrics that are sensitive to consequences experienced by the individuals affected by the decisions, (iii) navigate the emerging tradeoffs (e.g., between efficiency and fairness of the outcomes). The source code of the FairnessLab is available at https://github.com/joebaumann/FairnessLab and a demo of the interactive web application is available at https://joebaumann.github.io/FairnessLab.

Keywords

Fairness, bias, AI audit tool, bias mitigation, trustworthy AI, ethics

1. AI Audit Tools for Algorithmic Fairness

Artificial intelligence (AI) based decision-making systems are prevalent in our society even though they are often biased against certain groups [3, 4]. As a result, people and institutions have called for audits of these systems to avoid unfair outcomes. However, most existing AI audit tools are based on just a small set of mathematically incompatible so-called group fairness criteria [5–7] – despite fairness being a highly debated and contextual concept [8, 9]. Each one of these criteria is based on several moral assumptions, which are usually not made explicit, and may or may not be met in the given context [1, 2]. Therefore, we introduce the FairnessLab: a

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

new audit tool that makes moral viewpoints explicit and allows studying their consequences both with respect to fairness and to the decision maker's goal.

2. FairnessLab

The FairnessLab is implemented as an interactive web application specifically developed for bias audits of binary decision-making systems. Similar to existing fairness audit tools (such as [10–12]), our tool allows the users to perform an audit on a loaded dataset, which represents previously taken decisions of the audited system (see Figure 3 in Appendix A.2). However, in contrast to other tools, the FairnessLab evaluates the audited system's fairness with respect to some user-generated metrics. We believe that fairness is highly contextual, which is why there is no one-size-fits-all solution to evaluate the equitability of ML-based decision systems. The FairnessLab is based on a novel theoretical approach, which allows for the definition of context-specific fairness metrics [1, 2]. In particular, it consists of a series of questions whose answers lead to a morally appropriate definition of fairness for the audited system. The theoretical approach and, thus, the FairnessLab build on the algorithmic fairness and distributive justice literature and alleviate important shortcomings of existing audit tools, which only offer standard group fairness metrics derived from the confusion matrix.

The FairnessLab compares two perspectives: (I) **Decision maker**: The people or organization designing the algorithm, deciding on its design and thereby ultimately taking the decisions in question. (II) **Decision subjects**: The people subjected to the algorithm's decisions.

The FairnessLab consists of three key components: the decision maker's score, the fairness score, and the tradeoff visualization to balance tradeoffs between the two perspectives.

(I) Decision maker's score *To what degree is the goal of the decision maker achieved?*

Creating the decision maker's score requires assessing the average benefit/harm for the decision maker [13]. This is represented by a utility function specifying each possible outcome's desirability from the perspective of the decision maker.

(II) Fairness score *To what degree is fairness towards decision subjects achieved?*

The FairnessLab's underlying framework unifies and extends standard group fairness criteria while allowing for the interpretation of the user-generated group fairness criteria. This approach is described in [1]. The FairnessLab guides stakeholders in creating a fairness metric that fits their application context. The main questions stakeholders have to answer (with detailed guidance from the FairnessLab) are: What is, ultimately, distributed? Between whom is it distributed? Which subgroups should be compared? What is a fair distribution? [14, 15].

Tradeoff visualization *How do different decision-making systems compare with respect to the decision maker's score and the fairness score? What are the Pareto-efficient solutions?*

The first two components allow us to calculate the decision maker's score and the fairness

score for any given decision-making system if we have access to the input and output data. This, in turn, allows us to compare different decision-making systems with respect to these two variables. The `FairnessLab` provides a visualization for the decision maker's score and the fairness score for any given decision-making systems and identifies the Pareto-efficient solutions (as suggested in [9]). For a given decision-making system, the `FairnessLab` also automatically applies post-processing to compare different (upper- and lower-bound) thresholds to mitigate biases [16–18]. The tradeoff visualization allows stakeholders to discuss the ML model choices while being conscious of its fairness impact (see Figure 1 in Appendix A).

3. Conclusion

The `FairnessLab` is a tool that can be used both in developing a decision-making system and in its audit. In both cases, it leads to increased transparency and accountability: It requires stakeholders to make their assumptions (about the decision maker's utility, about fairness, and about the tradeoff of the two) explicit. This could help democratize the fairness debate: A fairness report containing all the choices made in using the `FairnessLab` as well as their justifications would allow others to scrutinize the assumptions made in developing or auditing a decision-making system. It may also help prevent ethics-washing where companies develop or audit systems using an inappropriate fairness metric.

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] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Conference on fairness, accountability and transparency, PMLR, 2018, pp. 77–91.
[4] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, ProPublica, May 23 (2016) 139–159. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[5] J. Kleinberg, S. Mullainathan, M. Raghavan, Inherent trade-offs in the fair determination of risk scores, arXiv preprint arXiv:1609.05807 (2016).
[6] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, On the (im)possibility of fairness, arXiv preprint arXiv:1609.07236 (2016).
[7] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Big data 5 (2017) 153–163.
[8] S. Barocas, M. Hardt, A. Narayanan, Fairness and machine learning, 2020. URL: http://fairmlbook.org, Incomplete Working Draft.
[9] M. Kearns, A. Roth, The ethical algorithm: The science of socially aware algorithm design, Oxford University Press, 2019.
[10] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, K. Walker, Fairlearn: A toolkit for assessing and improving fairness in AI, Technical Report, Technical Report MSR-TR-2020-32, Microsoft, May 2020., 2020.
[11] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, 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 and mitigating algorithmic bias, IBM Journal of Research and Development 63 (2019) 4:1–4:15. doi:10.1147/JRD.2019.2942287.
[12] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, R. Ghani, Aequitas: A Bias and Fairness Audit Toolkit, 2018. URL: https://arxiv.org/abs/1811.05577. doi:10.48550/ARXIV.1811.05577.
[13] C. Elkan, The Foundations of Cost-Sensitive Learning, in: Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 973–978.
[14] J. Rawls, A Theory of Justice, 2 ed., Harvard University Press, Cambridge, Massachussets, 1999.
[15] A. Sen, Equality of what?, The Tanner lecture on human values 1 (1980) 197–220.
[16] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, Algorithmic decision making and the cost of fairness, in: Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining, 2017, pp. 797–806.
[17] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, Advances in neural information processing systems 29 (2016).
[18] J. Baumann, A. Hannák, C. Heitz, Enforcing Group Fairness in Algorithmic Decision Making: Utility Maximization Under Sufficiency, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 2315–2326. URL: https://doi.org/10.1145/3531146.3534645. doi:https://doi.org/10.1145/3531146.3534645.
[19] K. Hao, J. Stray, Can you make AI fairer than a judge? Play our courtroom algorithm

game, MIT Technology Review (2019). URL: https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm.
[20] B. Green, "Fair" risk assessments: A precarious approach for criminal justice reform, in: 5th Workshop on fairness, accountability, and transparency in machine learning, 2018, pp. 1–5.
[21] B. Green, Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness, Philosophy & Technology 35 (2022) 90. URL: https://doi.org/10.1007/s13347-022-00584-6. doi:10.1007/s13347-022-00584-6.
[22] M. Bao, A. Zhou, S. Zottola, B. Brubach, S. Desmarais, A. Horowitz, K. Lum, S. Venkatasubramanian, It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks, 2021. URL: https://arxiv.org/abs/2106.05498. doi:10.48550/ARXIV.2106.05498.
[23] J. Dressel, H. Farid, The accuracy, fairness, and limits of predicting recidivism, Science Advances 4 (2018) eaao5580. URL: https://www.science.org/doi/abs/10.1126/sciadv.aao5580. doi:10.1126/sciadv.aao5580.
[24] A. Narayanan, How to recognize AI snake oil, Arthur Miller Lecture on Science and Ethics (2019).

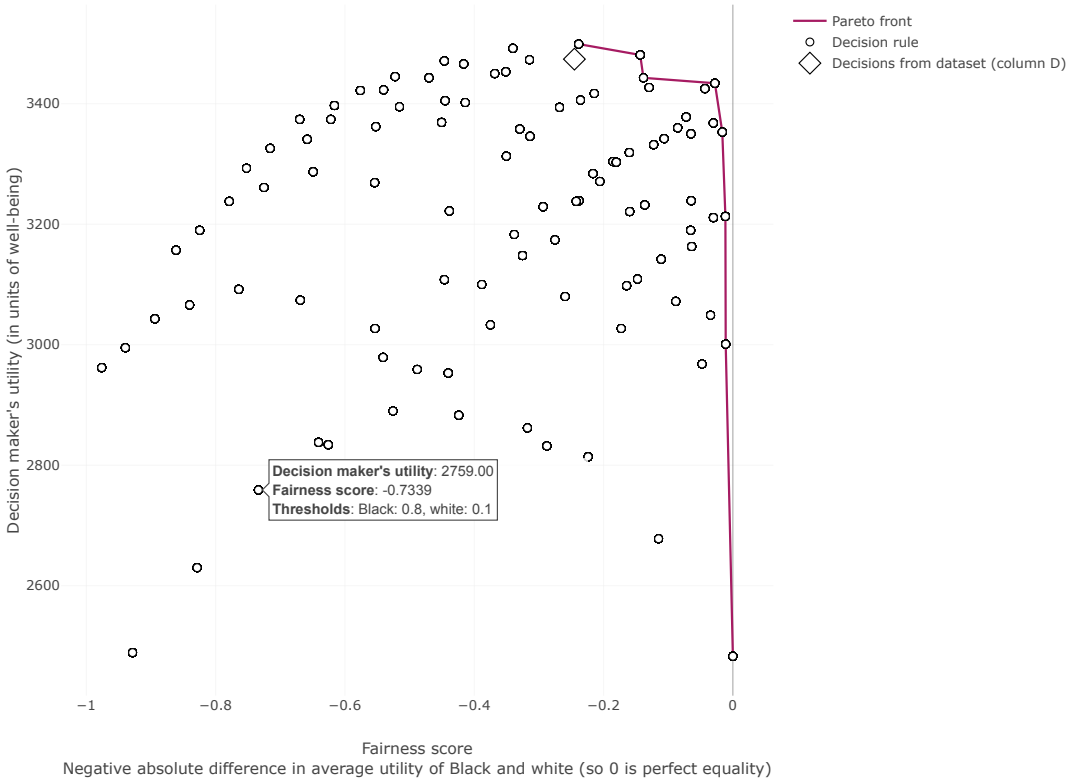
A. Web Application

Figure 1 shows two screenshots of the interactive web application. Figure 1a visualized the Pareto front (red line) of all possible decision rules (dots). This line represents all threshold combinations where (i) the fairness cannot be improved without worsening the utility of the decision maker or keeping it constant and where (ii) the utility of the decision maker cannot be improved without worsening the utility of the decision maker or keeping it constant – among the presented decision rules. Figure 1b shows the distribution of the score produced by the ML model across the specified relevant groups to compare.

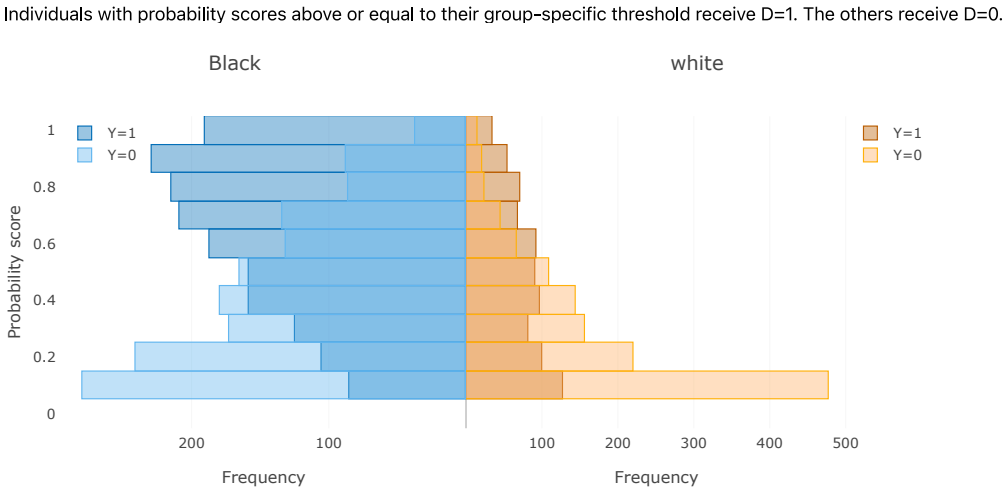
A.1. Running an Audit Using the `FairnessLab`

The fairness audit is performed by following these steps:

- Upload a dataset.
- Define fairness for the given context by specifying one's normative preferences regarding six value-laden questions (see Figure 2):
    1. **Utility of the decision maker:** How should we assess the benefit/harm that the decision maker derives from the decisions?
    2. **Utility of the decision subjects:** How should we assess the benefit/harm that the decision subjects derive from the decisions?
    3. **Relevant groups:** What groups of people are affected unequally by decision-making systems because being a member of a group is a (direct or indirect) cause of inequality? These could, for example, be groups defined by race, gender, disability status, sexual orientation, etc.



(a) Pareto front: dots represent possible post-processing decision rules.



(b) Group-specific score distributions.

Figure 1: Screen shot from the interactive web application for the COMPAS example.

4. **Claim differentiator:** By virtue of which features can individuals morally demand equal consideration by the decision maker?
5. **Pattern of justice:** Should the goal of justice be equality or some other distribution (e.g., maximizing the expectations of the worst-off group)?
6. **Tradeoff decision:** How strongly should fairness be pursued if it comes into conflict with the utility of the decision maker?

- Based on these configurations:
  - The decisions specified in the input dataset are audited, i.e., their fairness is quantified.
  - A menu of options is presented to evaluate and derive optimal decision rules for a certain degree of fairness.

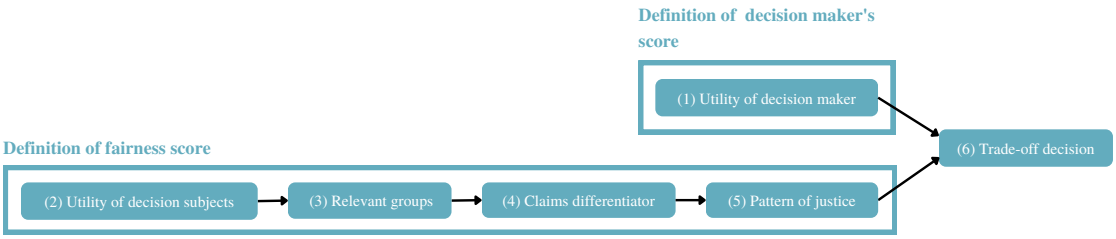


Figure 2: The six steps of the underlying framework from [1, 2] and their connections.

A.2. Comparison With Existing AI Audit Tools

Compared to existing bias audit tools, the set-up of the `FairnessLab` is very similar, as it also analyzes a given dataset for bias w.r.t. specified groups. However, the way fairness can be defined using the `FairnessLab` is conceptually different and, in addition to this, it outputs not only a bias report but also offers insights into existing tradeoffs and alternatives – see Figure 3.

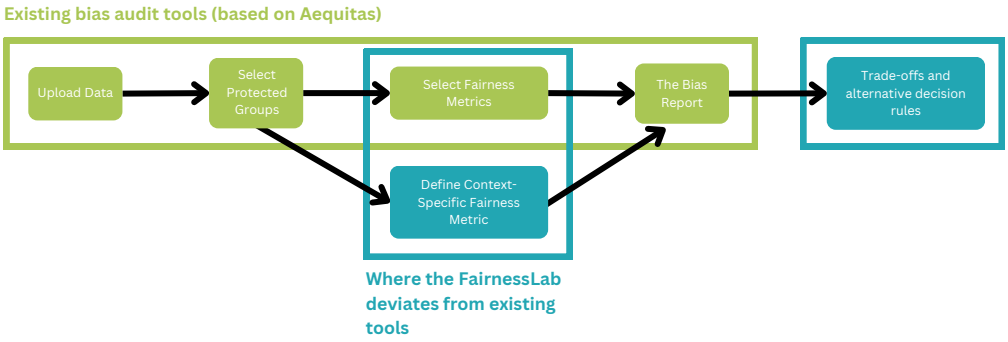


Figure 3: Comparison of the `FairnessLab` to existing bias audit tools (such as Aequitas [12]).

B. Reproducible Audit Example

We showcase the `FairnessLab` by auditing the COMPAS algorithm, which is used in parts of the US criminal justice system. Using the `FairnessLab`, we replicate existing analyses of this algorithm [4, 19] and provide new insights. Surprisingly, we find a way to make "better" decisions from the predictions given by the audited tool both with respect to bias and efficiency. This shows that previous audits that have relied on fairness metrics from the existing literature fall short. Minority groups have to be favored more than suggested by previous analyses in order to lessen the bias of the algorithm. This example audit is publically available at https://github.com/joebaumann/FairnessLab/blob/main/demo/COMPAS_audit.pdf.

C. Ethics Statement

Note that group-specific thresholds cannot be said to make a tool like COMPAS "fair": The systemic racism embedded in the US criminal justice system cannot be "fixed" by a risk assessment tool that has been audited for bias – deeper reforms are necessary [20, 21]. A tool used to decide who to detain may actually reinforce existing structures and get in the way of such deeper reforms. A better use of a predictive tool could be in rehabilitation efforts as highlighted by [22]. Note that a change of how the tool is used would also change the audit as the decision to allow someone to participate in a rehabilitation program would result in different utilities for decision subjects than the decision to imprison them. More generally, tools like COMPAS do not only have a fairness issue – their low accuracy also raises questions about their deployment [23]. As [24] points out, predicting social outcomes is extremely difficult or even impossible, so tools trying to do that are "fundamentally dubious" [24, p. 9].

Our audit is thus in no way meant to legitimize the usage of risk assessment systems in the criminal justice system. Rather, it is meant to highlight one of the shortcomings of previous audits: The `FairnessLab` allows for a reevaluation of the assumptions hidden in existing audits and for new audits that make use of context-specific user-generated fairness criteria.