Resources

Other Fairness Libraries of Note

- Aequitas
- AIF360
- Awesome Fairness in AI
- Dalex
- Fairlearn
- Fairness Comparison
- FAT Forensics
- ML Fairness Gym
- Themis ML

Recorded References

Crawford, K. (2017, December). The Trouble with Bias [Conference presentation]. NeurIPS 2017, Long Beach, CA. https://youtu.be/fMym_BKWQzk

Stucchio, C. (2018, October). AI Ethics, Impossibility Theorems and Tradeoffs [Conference presentation]. Crunch Data Conference 2018, Budapest, Hungary. https://www.youtube.com/watch?v=Zn7oWIhFffs

Additional Resources and Tutorials

Zhong, Z. (2018). "A Tutorial on Fairness in Machine Learning". Towards Data Science. https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

Cortez, V. (2019). ["How to define fairness to detect and prevent discriminatory outcomes in Machine Learning"](https://towardsdatascience.com/how-to-define-fairness-to-detect-and-prevent-discriminatory-outcomes-in-machine-learning-ef23fd408ef2#:~:text = Demographic%20Parity%20states%20that%20the,%E2%80%9Cbeing%20shown%20the%20ad%E2%80%9D). Towards Data Science. https://towardsdatascience.com/how-to-define-fairness-to-detect-and-prevent-discriminatory-outcomes-in-machine-learning-ef23fd408ef2#:~:text =

Demographic%20Parity%20states%20that%20the,%E2%80%9Cbeing%20shown%20the%20ad%E2%80%9D

Google People + AI Research (PAIR). PAIR Explorables: Measuring Fairness. https://pair.withgoogle.com/explorables/measuring-fairness/

Academic References

Agniel D, Kohane IS, & Weber GM (2018). Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. Bmj, 361. Retrieved from https://www.bmj.com/content/361/bmj.k1479

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. In International Conference on Machine Learning (pp. 60-69). PMLR. Available through arXiv preprint:1803.02453.

Agarwal, A., Dudik, M., & Wu, Z. S. (2019, May). Fair regression: Quantitative definitions and reduction-based algorithms. In International Conference on Machine Learning (pp. 120-129). PMLR. Available through https://arxiv.org/pdf/1905.12843.pdf

Bantilan N (2018). Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. Journal of Technology in Human Services, 36(1), 15-30. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/15228835.2017.1416512

Barocas S, & Selbst AD (2016). Big data's disparate impact. California Law Review, 104, 671. Retrieved from http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf

Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, ... & Nagar S (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv Preprint. arXiv:1810.01943.. See Also AIF360 Documentation

Bird S, Dudík M, Wallach H, & Walker K (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft Research. Retrieved from https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_whitepaper.pdf. See Also FairLearn Reference.

Dwork C, Hardt M, Pitassi T, Reingold O, & Zemel R (2012, January). Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214-226). Retrieved from https://arxiv.org/pdf/1104.3913.pdf

Equal Employment Opportunity Commission, & Civil Service Commission, Department of Labor & Department of Justice (1978). Uniform guidelines on employee selection procedures. Federal Register, 43(166), 38290-38315. Retrieved from http://uniformguidelines.com/uniformguidelines.html#18

Hardt M, Price E, & Srebro N (2016). Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323). Retrieved from http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf

Healthcare Cost and Utilization Project (HCUP) (2017, March). HCUP CCS. Agency for Healthcare Research and Quality, Rockville, MD. Retrieved from www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, & Mark RG (2016). Scientific Data. MIMIC-III, a freely accessible critical care database. DOI: 10.1038/sdata.2016.35. Retrieved from http://www.nature.com/articles/sdata201635

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018, July). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In International Conference on Machine Learning (pp. 2564-2572). PMLR. Available through http://proceedings.mlr.press/v80/kearns18a.html

Kim M, Reingol O, & Rothblum G (2018). Fairness through computationally-bounded awareness. In Advances in Neural Information Processing Systems (pp. 4842-4852). Retrieved from https://arxiv.org/pdf/1803.03239.pdf

National Association for the Advancement of Colored People (NAACP) (2012). Criminal Justice Fact Sheet. NAACP. Retrieved from http://www.naacp.org/pages/criminal-justice-fact-sheet.

Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review, 29(5), 582-638. Retrieved from https://www.cambridge.org/core/journals/knowledge-engineering-review/article/multidisciplinary-survey-on-discrimination-analysis/D69E925AC96CDEC643C18A07F2A326D7

Russell C, Kusner MJ, Loftus J, & Silva R (2017). When worlds collide: integrating different counterfactual assumptions in fairness. In Advances in Neural Information Processing Systems (pp. 6414-6423). Retrieved from https://papers.nips.cc/paper/7220-when-worlds-collide-integrating-different-counterfactual-assumptions-in-fairness.pdf

Shorrocks AF (1980). The class of additively decomposable inequality measures. Econometrica: Journal of the Econometric Society, 613-625. Retrieved from http://www.vcharite.univ-mrs.fr/PP/lubrano/atelier/shorrocks1980.pdf

Speicher T, Heidari H, Grgic-Hlaca N, Gummadi KP, Singla A, Weller A, & Zafar M B (2018, July). A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2239-2248). Retrieved from https://arxiv.org/pdf/1807.00787.pdf

Zemel R, Wu Y, Swersky K, Pitassi T, & Dwork C (2013, February). Learning fair representations. International Conference on Machine Learning (pp. 325-333). Retrieved from http://proceedings.mlr.press/v28/zemel13.pdf

Zafar MB, Valera I, Gomez Rodriguez, M, & Gummadi KP (2017, April). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th international conference on world wide web (pp. 1171-1180). https://arxiv.org/pdf/1610.08452.pdf