"Fair Decisions, Hard and Soft" – Abstract

Those working on algorithmic fairness have offered a number of distinct criteria for fair decision-making. However, it has been recognized that a number of these criteria cannot be mutually satisfied on any model. A natural response would be to argue *for* using certain criteria and *against* others, where our preferred criteria are consistent. Here, philosophers could step in to show how certain criteria do not deliver an appropriate conception of fairness, and how others do. After all, philosophers have already long argued about our conception of fairness.

Though reasonable, we worry that it cannot succeed. It may be that these criteria capture distinct senses of fairness, or perhaps they capture other considerations of value. To allow for this, we appeal to a critical distinction in mathematical optimization, and we argue that availing ourselves of it provides for several ways of accommodating apparently inconsistent criteria.

In mathematical optimization, practitioners often distinguish between so-called 'hard constraints,' where some mathematical criterion must be satisfied with exact equality, and 'relaxed' or 'soft constraints' where violations of a criterion within a certain margin are tolerated (Boyd & Vandenberghe 2004:Sec.5.1.4). Crucially, the criteria of fairness are inconsistent only if each is treated as hard constraints on the algorithm, and inconsistency can be resolved by relaxing this assumption. So, this paper primarily concerns what considerations suggest criteria for fairness as hard or soft, and how this influences a potential algorithm for fair decision-making.

One approach suggests that the criteria for fairness should each be treated as soft constraints. This has philosophical precedence going back to Broome, who treated fairness as a consideration that can be outweighed. If the criteria are not only soft but measurable/estimable, a practitioner can choose appropriate quantitative weights for each measure of fairness (or maximum tolerable violations of each fairness constraint may be chosen) along with a measure of predictive accuracy. Then, a fair model is easily estimated using standard methods.

However, there may be reasons why certain fairness criteria should be hard constraints. We show how this comes out of accepting a connection between fairness and respect for certain rights. Rights have been argued to generate exclusionary reasons, and we show how the role played by exclusionary reasons in reasoning maps directly onto the conception of hard constraints. Given this, we suggest a view of algorithmic fairness allowing for mixed constraints, and we offer a roadmap for how to use it.

References

Adams, N. P. (forthcoming). In defense of exclusionary reasons. *Philosophical Studies*: 1-19.

Bechavod, Y. and Ligett, K. (manuscript). Penalizing Unfairness in Binary Classification. June 2017. arxiv.org, https://arxiv.org/abs/1707.00044v3.

Berk, R., *et al.* (manuscript). A Convex Framework for Fair Regression. ArXiv:1706.02409 [Cs, Stat], Jun. 2017. arXiv.org, https://arxiv.org/abs/1706.02409.

Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4):991–1013.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, USA.

Broome, J. (1991). Fairness. *Proceedings of the Aristotelian Society*, 91: 87-101.

Broome, J. (1994). Fairness versus doing the most good. *Hastings Center Report*, 24(4): 36-39.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153-163.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining (pp. 797–806). New York, NY, USA: Association for Computing Machinery.
Curtis, B. L. (2014). To be fair. *Analysis*, 74(1): 47-57.
Haas, C. (2019). The price of fairness - A framework to explore trade-offs in algorithmic fairness. *40th International Conference on Information Systems*, ICIS 2019.
Heidari, H., Ferrari, C., Gummadi, K., & Krause, A. (2018). Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31* (pp. 1265–1276). Curran Associates, Inc.
Hatzistavrou, A. (2012). Motivation, reconsideration and exclusionary reasons. *Ratio Juris*, 25(3): 318-342.
Heilmann, C. & Wintein, S. (2017). How to be fairer. *Synthese*, 194(9): 3475-3499.
Hooker, B. (2005). Fairness. *Ethical Theory and Moral Practice*, 8(4): 329-352.
Kamishima, Toshihiro, et al. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. *Machine Learning and Knowledge Discovery in Databases*, Ed. P. A. Flach et al., (pp. 35–50). Springer Link.
Kilbertus, N., Rojas-Carulla, M., Parascandolo, Hardt, G. M., Janzing, D., Scho ̈lkopf, B. and De, B. M. (2017). Avoiding Discrimination through Causal Reasoning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc.
Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In Proceedings of Innovations in Theoretical Computer Science (ITCS).
Martin, A. M. (forthcoming). Personal bonds: Directed obligations without rights. *Philosophy and Phenomenological Research*, DOI: 10.1111/phpr.12620
M. Kusner, J. Loftus, C. Russell, and R. Silva. (2017). Counterfactual Fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems* (pp. 4069-4079). Curran Associates, Inc.
Mian, E. (2002). The curious case of exclusionary reasons. *Canadian Journal of Law and Jurisprudence*, 15(1): 99-124.
Kirkpatrick, J. R. & Eastwood, N. (2015). Broome's theory of fairness and the problem of quantifying the strengths of claims. *Utilitas*, 27(1): 82-91.
Lazar, S. (2018). Limited aggregation and risk. *Philosophy and Public Affairs*, 46(2): 117-159.
Lazenby, H. (2014). Broome on fairness and lotteries. *Utilitas*, 26(4): 331-345.
Lee-Stronach, C. (2018). Moral priorities under risk. *Canadian Journal of Philosophy*, 48(6): 793-811.
Lee-Stronach, C. (forthcoming). Morality, uncertainty. *Philosophical Quarterly*.
Li, Z., *et al.* (manuscript). Kernel Dependence Regularizers and Gaussian Processes with Applications to Algorithmic Fairness. ArXiv:1911.04322 [Cs, Stat], Nov. 2019. arXiv.org, http://arxiv.org/abs/1911.04322.
Newey, C. (2016). Fairness as 'appropriate impartiality' and the problem of the self-serving bias. *Ethical Theory and Moral Practice*, 19(3): 695-709.
Paseau, A. C. & Saunders, B. (2015). Fairness and aggregation. *Utilitas*, 27(4): 460-469.
Pessach, D. and Shmueli, E.(manuscript). Algorithmic Fairness. ArXiv:2001.09784 [Cs, Stat], Jan. 2020. arXiv.org, http://arxiv.org/abs/2001.09784.
Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L., & Camps-Valls, G. (2017). Fair kernel learning. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 339–355.
Pillar, C. (2017). Treating Broome fairly. *Utilitas*, 29(2): 214-238.
Raz, J. (1999). *A theory of justice* (Rev. Ed.). Cambridge, MA: Belknap Press.
Saunders, B. (2010). Fairness between competing claims. *Res Publica*, 16(1): 41-55.
Sharadin, N. (2016). Fairness and the strengths of agents' claims. *Utilitas*, 28(3): 347-360.

Sun, M., & Gerchick, M. (2019). The Scales of (Algorithmic) Justice: Tradeoffs and Remedies. *AI Matters*, 5(2), 30–40.

Tan, J. H. W. & Bolle, F. (2006). On the relative strengths of altruism and fairness. *Theory and Decision*, 60(1): 35-67.

Tomlin, P. (2012). On fairness and claims. *Utilitas*, 24(2): 200-213.

Whiting, D. (2017). Against second-order reasons. *Noûs*, 51(2): 398-420.

Wong, P. (2020). Democratizing algorithmic fairness. *Philosophy and Technology*, 33(2): 225-244.

Yeom, S. and Tschantz M. C. (manuscript). Discriminative but Not Discriminatory: A Comparison of Fairness Definitions under Different Worldviews. Aug. 2018. arxiv.org, https://arxiv.org/abs/1808.08619v5.