Author version—accepted for PSA 2024 contributed paper proceedings

On Values in Fairness Optimization with Machine Learning

Heather Champion

Department of Philosophy, University of Western Ontario

Rotman Institute of Philosophy, London, Ontario

hchampi2@uwo.ca

Abstract

Statistical criteria of fairness, though controversial, bring attention to the multiobjective nature of many predictive modelling problems. In this paper, I consider how epistemic and non-epistemic values impact the design of machine learning algorithms that optimize for more than one normative goal. I focus on a major design choice between biased search strategies that directly incorporate priorities for various objectives into an optimization procedure, and unbiased search strategies that do not. I argue that both reliably generate Pareto optimal solutions such that various other values are relevant to making a rational choice between them.

1. Introduction

Philosophers and computer scientists debate whether there exist statistical criteria of fairness that can be used to assess the fairness of a machine learning (ML) model's predictions. In this paper, I bring attention to the additional values that are involved when these or other notions of fairness are integrated into an optimization procedure used to train an ML model. I focus on a major design choice in multiobjective optimization problems (MOOPs) between biased and unbiased search strategies, which are distinguished by whether they use priorities for the objectives to direct an optimization procedure. I argue that both reliably generate Pareto optimal (PO) solutions, but there are various additional, not purely epistemic, reasons to prefer one or the other.

The paper will proceed as follows. In Section 2, I explain how my analysis diverges from existing philosophical work on algorithms and values, and I specify the sense in which I consider values to impact model design choices. In Section 3, I introduce MOOPs and strategies for solving them, and I motivate why it is important to understand fairness optimization as a MOOP. In Section 4, I argue that the methodological choice between biased and unbiased search strategies is unforced by the epistemic aim of finding Pareto optimal tradeoffs. In Section 5, I discuss what additional values are relevant to this choice.

2. Values act as justificatory reasons for choices in optimization problems

Philosophers and social scientists have brought critical attention to the multitude of ways that algorithms are biased, often leading to unfairness when algorithmic predictions inform decisions. In this work, I offer a novel philosophical reflection on how algorithms implement values by focusing on the choice of what optimization method to use when incorporating fairness notions into ML design. This goes beyond existing philosophical work that highlights the relationship between values and statistical performance metric(s) including tradeoffs between statistical criteria of fairness, and data-driven practices as a whole (e.g. Fazelpour and Danks 2021). While fairness criteria can be incorporated at various stages of ML model design, including before, during, or after an optimization algorithm is run to train an ML model, my purpose here is to compare approaches that incorporate fairness notions into optimization design.

As my goal is to analyze what justifies particular design choices in optimization, I will focus on a specific type of relation between values and choices where values act as justificatory reasons for choices. This type of relation is distinguished by Ward (2021), who characterizes it as involving an appeal to values in rational arguments used to support a certain course of action. Using Ward's taxonomy: I will not attempt to answer what values *motivate* ML designers to choose various optimization strategies, which would require a separate psychological and sociological analysis. I will also not consider how values act as *causes* or *effects* of design choices; for instance, what values make designers and decision makers (DMs) more likely to formulate a decision as an ML problem or as a MOOP, or how a deployed ML model promotes certain social values and practices. Instead, I will focus on how values are (and should be) used as justificatory reasons in choices regarding the use of a particular optimization strategy. Notably, following Ward and others, my assessment focuses on pro tanto reasons for choices, rather than on what reasons fully justify a choice.

3. Optimizing for fairness in ML is a multiobjective problem

Tradeoffs arise in predictive modelling that aims to be fair in several ways, indicating that the problem of implementing almost any fairness notion in ML optimization is multiobjective. First, predictive modelling involves a cost versus benefit analysis. For example, solving a classification problem involves a search for a model that makes an optimal tradeoff between the costs and benefits of each type of predictive result (e.g. positive or negative classifications, whether true or false). In supervised ML, these predictive preferences are typically chosen by the DM and encoded as a single performance metric that may be implemented as the objective function to optimize or used to evaluate the model's final performance. Either way, during the training phase, the optimization algorithm varies the model's parameters to minimize an objective function (also called the "loss" function) that expresses the loss of the costs relative to the benefits when predicting the target variable for the population represented in the training dataset. The final trained model generates individual predictions based on an optimal risk score that represents the frequency of the target outcome given a particular set of features (in a Bayesian sense, it represents the posterior probability of the outcome given the observed features; Barocas et al. 2023, 48).

Nonetheless, risk scores that are optimal for the whole population may not correspond to what is optimal for various subgroups, such that striving for fairness to groups involves tradeoffs with overall predictive performance. Barocas et al. (2023) bring attention to how in general, we should not expect the costs and benefits of predictions to be equally shared across groups if the predicted target variable is not statistically independent of group membership. Here, attempts to equalize various statistical criteria of fairness across groups generally decrease overall predictive performance. The only exceptions are fairness criteria based on the notion of sufficiency, which require that the risk score is well calibrated to the predicted outcomes within groups (Barocas et al. 2023, 62). Yet even sufficiency trades off with population accuracy if models are "blinded" to group membership (e.g. for legal reasons) and the relationship between predictive features and outcomes differs across groups (Corbett-Davies et al. 2023, 16-17).

Meanwhile, different statistical notions of what constitutes fairness to groups also involve inherent tradeoffs such that employing more than one criterion incorporates conflicting goals into a predictive modelling problem. Barocas et al. (2023) categorize statistical non-discrimination criteria into three types based the notions of sufficiency, independence, and separation. Independence requires that a risk score is independent of a sensitive attribute such that the acceptance rate of a classifier is equal across groups. On the other hand, separation requires that a risk score is independent of a sensitive attribute within each stratum of equal claim, implying error rate parity. Barocas et al. prove that formal tradeoffs exist between each family of criteria when the target variable is not independent of group membership (except for degenerate solutions). Thus, there is a three-way tradeoff between the notions of sufficiency, independence, and separation such that if one is satisfied, the others cannot be (see also Chouldechova 2017, Kleinberg et al. 2016, Miconi 2017.)

Last, implementing even one statistical fairness criterion in a predictive modelling problem can benefit one demographic group while imposing significant costs for other subgroups of the population. Kearns et al. (2018) call this "fairness gerrymandering," where an appearance of equity with respect to one subgroup comes at the expense of unfairness towards another. Thus, the challenge of achieving fairness with respect to multiple subgroups increases the number of conflicting goals in a predictive modelling problem.[1]

Philosophers and computer scientists respond to these inherent tradeoffs with the following four main claims. (1) Only one statistical criterion is normatively relevant (e.g. Separation: Hellman 2020, Grant 2023. Calibration: Hedden 2021, Corbett-Davies et al. 2023). (2) Necessary statistical criteria of fairness should be assessed on a case-by-case basis, especially since it is possible to satisfy some measures from more than one family at the expense of others in those same categories (Miconi 2017). (3) The criteria could be relaxed, substantially altered to be compatible (Beigang 2023), or implemented as group-specific aims rather than as a summarized statistic (eg. Group DRO: Sagawa et al. 2020. Minimax Pareto fairness: Martinez et al. 2020). (4) The criteria should be abandoned in favour of alternative debiasing or fairness optimization strategies such as striving towards a perfect predictor (Miconi 2017 points out this option), implicit unfairness mitigation methods (see Wan et al. 2023), or direct cost–benefit analysis using real-world quantities (Corbett-Davies et al. 2023).[2]

The diversity of these responses raises the question: what might a focus on multiobjective optimization add to this debate? I propose that analyzing the values that act as justificatory reasons for choices regarding how to optimize for multiple aims is part of the normative work required for establishing the legitimacy of almost *any* approach proposed above, since they involve design choices that may or may not lead to Pareto optimality (defined below), and that might imply more or less precision in how well the solutions correspond to a DM's preferences. (The main exception is for a perfect predictor; Miconi 2017). For example, if only one statistical criterion is required for fairness in general or in a particular case, then it likely involves a tradeoff with population performance (even for sufficiency, if a "blind" model is used). The advantage of recognizing the multiobjective nature of these optimization problems is that design

[1] Designers have proposed clever solutions to this problem, where the tradeoffs between subgroups are either considered explicitly for multiple sensitive attributes (Zafar et al. 2017) or "blindly" for any subgroups of sufficient size (Martinez 2021, Kearns et al. 2018).
[2] I have outlined these responses with insight from Miconi 2017, Hedden 2021, and Beigang 2023.

choices may be evaluated for whether they generate a PO solution: one that offers the best possible population performance for the *required* fairness (e.g. Zafar et al. 2017 consider these and contrasting cases of "business necessity"). Alternatively, one or more statistical criteria of fairness might be relaxed in order to find an intermediate compromise with population accuracy or other fairness objectives. In these cases, treating the problem as a multiobjective one enables designers to plan for a PO tradeoff solution: one without any unnecessary relaxation. Even Beigang's (2023) approach, which offers alternative, compatible, criteria of algorithmic fairness, still does not reconcile the multiple goals of distributive justice that might arise from algorithmic decision making. Meanwhile, optimizing group-specific aims (without summary statistics) is also an inherently multiobjective problem, where Pareto optimality often matters for doing no unnecessary harm (Martinez et al. 2020 raise this point) but is not guaranteed by some approaches (e.g. group DRO).

In a MOOP, the main aim is to find one or more PO solutions that represent the best possible tradeoffs between the objectives. A MOOP can be formulated as a minimization problem with m objectives ($m \geq 2$) that map decision variables $x \in X$ in decision space X into the objective space, defined in $\mathbb{R}^m$:

$$\min \quad F(x) = (f_1(x), f_2(x), \dots f_m(x))$$

A solution is PO if and only if it is not possible to improve in any one objective without degrading in at least one other. The notion of dominance distinguishes PO solutions from non-PO solutions: a solution $\mathbf{x}^{(1)}$ dominates $\mathbf{x}^{(2)}$ (equivalently, $\mathbf{x}^{(1)}$ is nondominated by $\mathbf{x}^{(2)}$) if and only if two conditions hold (following Deb et al.'s formalism; 2016, 151):

1. $\mathbf{x}^{(1)}$ is no worse than $\mathbf{x}^{(2)}$ in all objectives ($f_j\ (\mathbf{x}^{(1)}) \ntriangleright\ f_j\ (\mathbf{x}^{(2)})$ for all j=1, …, m)
2. $\mathbf{x}^{(1)}$ is strictly better than $\mathbf{x}^{(2)}$ in at least one objective ($f_{\bar{j}}\ (\mathbf{x}^{(1)}) \vartriangleleft\ f_{\bar{j}}\ (\mathbf{x}^{(2)})$) for at least one $\bar{j} \in \{1, \dots, m\}$.

The PO set is comprised of the nondominated solutions attained when the entire feasible region of the decision space is searched (Deb et al., 153). In most MOOPs, feasible solutions are defined by various constraints, including limits on the range of each parameter value that is searched. This means that in the objective function space, PO solutions mark the boundary between the feasible and infeasible regions (see Fig. 1).

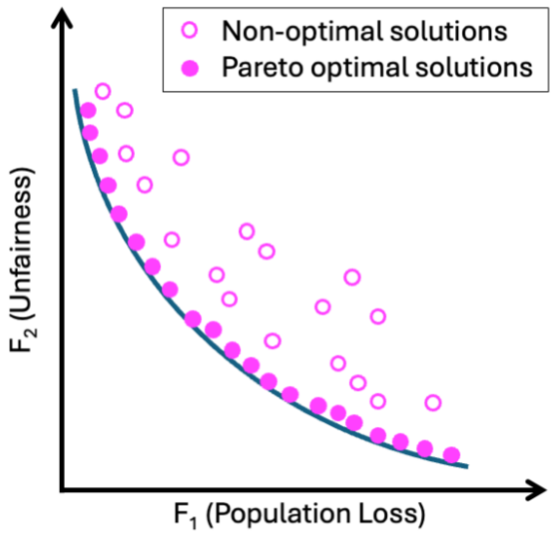

Figure 1: Tradeoff solutions in objective function space when minimizing some notion of unfairness (F2) and population loss (F1).

Optimization strategies that implement multiple objective functions can be characterized according to two major design choices. The first choice depends on *when* the DM specifies preferences for how to trade off the multiple objectives. In *a priori* approaches, the DM decides before the stage of optimization and these preferences are used to obtain a single PO solution. In contrast, *a posteriori* approaches present the DM with multiple PO solutions to choose from representing different tradeoffs, after the stage of optimization.[3] Typically, a posteriori approaches aim to find a large portion of the entire PO set (Deb et al., 148). I will focus this paper on a second major design choice, which concerns whether to use a search strategy that is guided toward a particular tradeoff solution. I will use the term "biased search" for approaches that aim to find the single PO solution that corresponds to a certain set of priorities for the multiple objectives. They can be used in either an a priori or a posteriori way, depending on whether the priorities are iteratively varied to map the Pareto front (PF). On the other hand, I will call search strategies that do not assign priorities to the multiple objectives as an optimization proceeds "unbiased" (i.e. in this limited sense; see Table 1).

[3] Although some "interactive" approaches do not fit neatly within this distinction (see Deb et al. 2016, 13).

Table 1: Multiobjective fairness optimization strategies that generate Pareto optimal solutions

	A Priori	**A Posteriori**
Biased Search	• Linear scalarization (Kamishima et al. 2012) • Constrained optimization (Zafar et al. 2017) • Rawlsian minimax Pareto fairness (Martinez et al. 2020)	• Epsilon-constraint (Liu & Vicente's 2022 extension of Zafar et al. 2017) • Chebychev scalarization (Wei & Niethammer 2022)
Unbiased Search	• Constrained multigradient descent followed by automated selection of the final solution (Padh et al. 2021)	• Stochastic multigradient descent (Liu & Vicente 2022)

4. Choosing between biased and unbiased search strategies is an unforced methodological choice

Here I argue that biased and unbiased search strategies for multiobjective optimization both reliably generate PO solutions, even for ML problems that indirectly optimize the goals. This implies that choosing between these strategies is an "unforced" methodological choice (Winsberg 2012).

Biased search strategies often make use of optimization engines that offer guarantees of convergence (for convex functions), reducing the multiple objectives into a single function while incorporating priority information (e.g. as constraints on a primary objective, Zafar et al. 2017; or as weights for a scalarizing function, Kamishima 2012). For example, many deterministic algorithms are guaranteed to converge on the global optimal solution to a convex objective function (where every local minimum is a global minimum, e.g. quadratic programming, gradient-descent, Newton's method). Alternatively, stochastic optimization engines also reliably converge on global optima (of either convex or nonconvex functions). To illustrate, appropriately sized stochastic variation of parameter values can prevent premature convergence to local but not global optima.

While unbiased search strategies optimize multiple objectives simultaneously rather than reducing them, they also reliably generate PO solutions. Unbiased strategies first attempt to map the PF and then incorporate a DM's preference information (specified either a priori or a posteriori) to select a PO solution. In ML, unbiased methods map the PF by maintaining a list of nondominated solutions in various ways; for example, by comparing the list to new solutions generated at each epoch (Padh et al. 2021) or to solutions generated by multiple runs of the algorithm with randomized initial starting points (Liu & Vicente 2022). Multigradient descent algorithms follow a vector of common descent (that represents a convex combination of the objective function *gradients*) until it vanishes. The resulting "Pareto stationary" solutions can be compared in terms of dominance. In general, repeatedly vetting solutions by dominance improves the quality of the PF.

Nonetheless, I submit a few remarks on the reliability of ML-specific multiobjective optimization, as ML uses an "indirect" form of optimization (Goodfellow 2016, 268). In supervised ML, the ultimate performance of a model is judged by evaluating the objective function(s) on a specially reserved subset of the data (called the "test set") that is not used during optimization. Still, granting that the same performance measure(s) (or suitable approximations) are used for both subsets, and that standard ML assumptions hold (the training and test sets are independent and identically distributed), biased or unbiased search strategies also reliably generate PO solutions in ML.[4]

Empirical evidence also supports my claim. Liu and Vicente (2022) develop an unbiased search method for multiobjective fairness optimization based on stochastic multi-gradient descent (with the goal of using it in an a posteriori way). They compare their method to a biased search approach (Zafar et al.'s constraint-based optimization; 2017) on two convex optimization problems. The goals are to make accurate predictions of salary and to minimize disparate impact with respect to either gender or race (they use the Adult Income dataset; Kohavi 1996). For accuracy, they use a logistic regression function and for disparate impact they use a convex approximation of Calders & Verwer's (CV) score based on the decision boundary covariance (2010). The comparison requires iterating over the constraint parameter in Zafar et al.'s method and storing nondominated solutions in order to obtain a full PF with the biased search strategy (i.e. adapting it to an epsilon-constraint method: see Haimes, 1971). Liu and Vicente's (2022) results show that both methods generate high-quality PFs: when optimizing for minimal disparate impact with respect to gender, the PF-SMG algorithm dominates the EPS-fair algorithm in some regions of the tradeoff surface (their Figure 1a, 521). However, when optimizing for minimal disparate impact with respect to race, the EPS-fair algorithm dominates in some regions (Figure 2a, 522). Notably, the DM's preferences (specified either a priori or a posteriori) might require targeting a PO solution from any region of these tradeoff surfaces, where either method might dominate. Thus, neither approach offers an overall advantage in terms of Pareto dominance for these tests.[5]

This means that values are relevant to normatively assessing the choice between biased and unbiased search strategies for finding PO solutions. Particularly, in the sense articulated by

[4] But see Molnar and Freiesleben (2024) for an overview of strategies to improve the reliability of supervised ML models.

[5] Liu and Vicente (2022) also make a wider comparison of how the methods perform on convex problems using 40 datasets (522–523). They note the biased method slightly outperforms the unbiased one according to a scalar metric that measures Pareto dominance, but they observe that the PFs are quite close. Meanwhile, they are not able to use the biased method to produce PFs of sufficient quality for the comparison on nonconvex problems (524). Thus, perhaps conditioning on the problem type (convex or nonconvex) might differentiate between the epistemic advantages of biased and unbiased search strategies in attaining Pareto optimality. However, this would require a much wider analysis as additional (and sometimes orthogonal) design choices also contribute to these advantages (e.g. incorporating stochasticity).

Ward (2021), values might act as non-epistemic justificatory reasons. Winsberg (2012) brings attention to how non-epistemic values impact what he calls "unforced methodological choices" in climate modelling, where one option is not "objectively" better than another, but each presents a benefit for a different set of preferences such as preferences for inductive risks (130). While he argues it is not possible to isolate the values that have impacted the history of climate science since they hide in all its "nooks and crannies," and current climate models also rely on past modelling choices, methods for fairness optimization are not yet so "generatively entrenched" (132). Therefore, I now suggest several values that are relevant to the choice between biased and unbiased search strategies.

5. Values in multiobjective optimization

There are several differences between biased and unbiased search procedures beyond their ability to generate PO solutions that might act as justificatory reasons for choosing between them. Here I discuss three: precision in selecting a PO solution that corresponds to a certain set of preferences, computational efficiency, and dynamic adaptability. While not all of these appear to be purely epistemic, I do not attempt to classify them as epistemic or non-epistemic, or to regress to the problem of identifying values that act as reasons for choosing between them.

First, perhaps counterintuitively, *unbiased* methods appear to have an overall advantage in offering more precise control over solution preferences. One reason for this is that it is somewhat difficult to achieve a dense and well-spread PF with a biased method. For weight-based methods, an evenly distributed set of weights does not necessarily generate an even spread of PO solutions (Deb et al. 2016, 160). Meanwhile, constraint-based methods require advanced knowledge of the PF to achieve a sufficient resolution, including what upper bounds to use for the constraints and what step size to vary them with. Liu & Vicente's (2022) results demonstrate this difficulty: their unbiased method produces a more dense and well-spread PF compared to the epsilon-constraint method (522-523). Also, the overall advantage of unbiased methods in generating higher resolution PFs applies even when preferences are specified a priori. For example, if constraints are set too narrow, no feasible solution will be obtained, and if they are broad enough to admit more than one PO solution, a highly resolved PF affords better fine-grained solution control. In addition, while it is possible in principle with some biased methods to discover nonconvex regions of a PF (e.g. epsilon-constraint, Chebyshev scalarization), the challenge of tuning the weights or step-size of the constraints means that it is more difficult than with unbiased methods. For this reason, designers of biased methods often choose to use convex approximations of the objective functions (e.g. Zafar et al. 2017).

Second, biased and unbiased methods differ in the efficiency at which they generate one or more PO solutions. Since biased methods execute single objective optimization routines, they are most efficient for a priori problems that aim to generate a single PO solution. However, they require iterative optimizations for solving a posteriori problems, and so they may be inefficient for problems with many objectives (such as fairness optimization, see Section 3). Also, if the objective functions are nonconvex, biased methods require iterative exploration of the PF to discover any nonconvex regions, which also makes them less efficient overall. On the other hand, while unbiased methods might be slightly less efficient at generating single PO solutions, they outperform biased methods in efficiently mapping the PF (Liu & Vicente also measure this explicitly; 2022). Thus, different measures of computational efficiency are relevant to choosing between biased and unbiased methods.

Third, unbiased methods have the advantage that the PF may be dynamically adapted as new data becomes available. Liu & Vicente (2022) highlight that this is an important feature of their design, and they present results that simulate a streaming scenario: the PFs they compute on successive batches of the training data adaptively converge to the final PF computed for the whole training dataset (527, 535). Optimizing the multiple objectives simultaneously and stochastically sampling the data makes this efficient; it does not require restarting a series of biased searches to compute a new PF. Instead, new solutions computed with fresh data can be directly compared to previous solutions in terms of dominance. The ability to handle streaming data is important because in realistic socioeconomic systems, PFs evolve dynamically.

In sum, while neither biased nor unbiased search methods offer an overall advantage in terms of Pareto dominance, biased searches might be preferred for efficiently computing a single PO solution, and unbiased methods for precise control over tradeoff preferences, efficient PF mapping, and dynamic PO prediction.

6. Conclusion

I emphasized that since in general, implementing fairness notions in ML optimization is a multiobjective problem, it is important to normatively assess the design choices involved in various methods that incorporate more than one objective in optimization. I argued that choosing between biased and unbiased search strategies is not forced by the epistemic aim of finding nondominated, PO solutions. Instead, values such as computational efficiency and preferences regarding various types of predictive risks (precision in tradeoff control, robustness to distribution shifts) impact this choice.

6.1. Future Work

My analysis is relevant to future ethical reflection on the design of fair algorithms. For instance, work that assesses the normative significance and priority that should be given to various values such as the ones I have highlighted here. In particular, the importance of obtaining Pareto optimality must be ethically grounded, and here I briefly highlight a couple of possible approaches stemming from Grant et al.'s (2025) proposal regarding duties of evidential consideration owed to decision subjects. First, the accuracy of a predictive model relative to other available predictive models is a morally salient feature for showing due consideration to the substantive claims of decision subjects (11). Thus, if an optimization method that does not reliably generate PO solutions is chosen (e.g. group DRO; Sagawa et al. 2020), a decision subject may not have been shown due consideration since better methods exist for safeguarding against unnecessary loss of accuracy (i.e. loss that is not necessary for gain in some other normative goal). Second, optimizing for population accuracy might also raise duties to decision subjects to safeguard against unnecessary differential accuracy between groups or unnecessary loss of other kinds of fairness. These duties would safeguard against insensitivity to the *multidimensional nature* of the salient moral claims, rather than establish general priorities for various normative goals.

Also, the choice between a priori and a posteriori approaches should be normatively assessed. If a DM has clear requirements for tradeoffs, it seems to justify the use of an a priori approach. However, if a DM is unable to precisely specify preferences, a posteriori approaches provide useful information for making ethical judgments: they can be used to identify "knee" solutions, where the curve of the PF changes rapidly between two objectives such that a small improvement in one objective causes a large degradation in at least one other. Branke et al. (2004) argue that in the absence of preference information, knee solutions are most likely to be interesting to a DM because other nearby solutions involve a large loss in at least one objective (i.e. knee solutions are *locally non-extreme*).

References

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, MA: MIT Press, 2023.

Beigang, Fabian. “Reconciling Algorithmic Fairness Criteria.” *Philosophy & Public Affairs* 51, no. 2 (April 2023): 166–90. https://doi.org/10.1111/papa.12233.

Branke, Jürgen, Kalyanmoy Deb, Henning Dierolf, and Matthias Osswald. “Finding Knees in Multi-Objective Optimization.” In *Parallel Problem Solving from Nature - PPSN VIII*, edited by Xin Yao, Edmund K. Burke, José A. Lozano, Jim Smith, Juan Julián Merelo-Guervós, John A. Bullinaria, Jonathan E. Rowe, Peter Tiňo, Ata Kabán, and Hans-Paul Schwefel, 722–31. Berlin, Heidelberg: Springer, 2004. https://doi.org/10.1007/978-3-540-30217-9_73.

Calders, Toon, and Sicco Verwer. “Three Naive Bayes Approaches for Discrimination-Free Classification.” *Data Mining and Knowledge Discovery* 21, no. 2 (September 2010): 277–92. https://doi.org/10.1007/s10618-010-0190-x.

Chouldechova, Alexandra. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” arXiv, February 28, 2017. https://doi.org/10.1089/big.2016.0047.

Corbett-Davies, Sam, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. “The Measure and Mismeasure of Fairness.” arXiv, August 14, 2023. https://doi.org/10.48550/arXiv.1808.00023.

Deb, Kalyanmoy, Karthik Sindhya, and Jussi Hakanen. “Multi-Objective Optimization.” In *Decision Sciences*. 145-183: CRC Press, 2016.

Fazelpour, Sina, and David Danks. “Algorithmic Bias: Senses, Sources, Solutions.” *Philosophy Compass* 16, no. 8 (June 12, 2021): e12760. https://doi.org/10.1111/phc3.12760.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Grant, David Gray. “Equalized Odds Is a Requirement of Algorithmic Fairness.” *Synthese* 201, no. 3 (March 8, 2023): 101. https://doi.org/10.1007/s11229-023-04054-0.

Grant, David Gray, Jeff Behrends, and John Basl. “What We Owe to Decision-Subjects: Beyond Transparency and Explanation in Automated Decision-Making.” *Philosophical Studies* 182, no. 1 (January 1, 2025): 55–85. https://doi.org/10.1007/s11098-023-02013-6.

Haimes, YV. “On a Bicriterion Formulation of the Problems of Integrated System Identification and System Optimization.” *IEEE Transactions on Systems, Man, and Cybernetics* SMC-1, no. 3 (July 1971): 296–97. https://doi.org/10.1109/TSMC.1971.4308298.

Hedden, Brian. “On Statistical Criteria of Algorithmic Fairness.” *Philosophy & Public Affairs* 49, no. 2 (March 1, 2021): 209–31. https://doi.org/10.1111/papa.12189.

Hellman, Deborah. “Measuring Algorithmic Fairness.” *Virginia Law Review* 106, no. 4 (June 2020): 811–66. https://www.jstor.org/stable/27074708.

Kamishima, Toshihiro, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. "Fairness-Aware Classifier with Prejudice Remover Regularizer." In *Machine Learning and Knowledge Discovery in Databases*, edited by Peter A. Flach, Tijl De Bie, and Nello Cristianini, 35–50. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2012. https://doi.org/10.1007/978-3-642-33486-3_3.

Kearns, Michael, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness." In *Proceedings of the 35th International Conference on Machine Learning*, 2564–72. PMLR, 2018. https://proceedings.mlr.press/v80/kearns18a.html.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores." arXiv, November 17, 2016. https://doi.org/10.48550/arXiv.1609.05807.

Kohavi, Ronny, and Barry Becker. "Adult Income Dataset (UCI Machine Learning Repository)," 1996. https://archive.ics.uci.edu/ml/datasets/adult.

Liu, Suyun, and Luis Nunes Vicente. "Accuracy and Fairness Trade-Offs in Machine Learning: A Stochastic Multi-Objective Approach." *Computational Management Science* 19, no. 3 (July 1, 2022): 513–37. https://doi.org/10.1007/s10287-022-00425-z.

Martinez, Natalia, Martin Bertran, and Guillermo Sapiro. "Minimax Pareto Fairness: A Multi Objective Perspective." In *Proceedings of the 37th International Conference on Machine Learning*, 6755–64. PMLR, 2020. https://proceedings.mlr.press/v119/martinez20a.html.

Martinez, Natalia L., Martin A. Bertran, Afroditi Papadaki, Miguel Rodrigues, and Guillermo Sapiro. "Blind Pareto Fairness and Subgroup Robustness." In *Proceedings of the 38th International Conference on Machine Learning*, 7492–7501. PMLR, 2021. https://proceedings.mlr.press/v139/martinez21a.html.

Miconi, Thomas. "The Impossibility of 'Fairness': A Generalized Impossibility Result for Decisions." arXiv, September 11, 2017. https://doi.org/10.48550/arXiv.1707.01195.

Molnar, Christoph, and Timo Freiesleben. *Supervised Machine Learning for Science: How to Stop Worrying and Love Your Black Box*, 2024. https://ml-science-book.com/.

Padh, Kirtan, Diego Antognini, Emma Lejal-Glaude, Boi Faltings, and Claudiu Musat. "Addressing Fairness in Classification with a Model-Agnostic Multi-Objective Algorithm." In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, 600–609. PMLR, 2021. https://proceedings.mlr.press/v161/padh21a.html.

Sagawa, Shiori, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. "Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization." arXiv, April 2, 2020. https://doi.org/10.48550/arXiv.1911.08731.

Wan, Mingyang, Daochen Zha, Ninghao Liu, and Na Zou. "In-Processing Modeling Techniques for Machine Learning Fairness: A Survey." *ACM Transactions on Knowledge Discovery from Data* 17, no. 3 (March 20, 2023): 1–27. https://doi.org/10.1145/3551390.

Ward, Zina B. "On Value-Laden Science." *Studies in History and Philosophy of Science Part A* 85 (February 2021): 54–62. https://doi.org/10.1016/j.shpsa.2020.09.006.

Wei, Susan, and Marc Niethammer. "The Fairness-Accuracy Pareto Front." *Statistical Analysis and Data Mining: An ASA Data Science Journal* 15, no. 3 (June 2022): 287–302. https://doi.org/10.1002/sam.11560.

Winsberg, Eric. "Values and Uncertainties in the Predictions of Global Climate Models." *Kennedy Institute of Ethics Journal* 22, no. 2 (June 2012): 111–37. https://doi.org/10.1353/ken.2012.0008.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. "Fairness Constraints: Mechanisms for Fair Classification." arXiv, March 23, 2017. https://doi.org/10.48550/arXiv.1507.05259.

Funding

Thanks to support from the Northeastern AIDE Summer Training Program 2023 and NSF Award # 2147220. This paper also draws on research supported by the Social Sciences and Humanities Research Council of Canada. Ce article s'appuie sur des recherches financées par le Conseil de recherches en sciences humaines du Canada.

Acknowledgements

Thanks to helpful comments from participants at the Northeastern AIDE Summer Training Program 2023. I would also like to thank Chris Smeenk and Kathleen A. Creel for helpful discussions.