Philosophy of Science (2025), **92**, 1372–1382
doi:10.1017/psa.2025.10136

PSA PHILOSOPHY OF SCIENCE ASSOCIATION

PHILOSOPHY OF SCIENCE

CONTRIBUTED PAPER

On Values in Fairness Optimization with Machine Learning

Heather Champion

Department of Philosophy, University of Western Ontario, London, Ontario, Canada and Rotman Institute of Philosophy, London, Ontario, Canada
Email: hchampi2@uwo.ca

(Received 25 January 2025; revised 26 March 2025; accepted 24 July 2025; first published online 02 September 2025)

Abstract

Statistical criteria of fairness, although controversial, bring attention to the multiobjective nature of many predictive modeling problems. In this article, I consider how epistemic and nonepistemic values help to justify 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 article, 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 article 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 be related to 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 Author(s), 2025. Published by Cambridge University Press on behalf of Philosophy of Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psa.2025.10136 Published online by Cambridge University Press

the epistemic aim of finding PO trade-offs. 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 trade-offs between statistical criteria of fairness, and the values pertinent to 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, my purpose here is to compare approaches that incorporate fairness notions into optimization design.

Because 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

Trade-offs arise in predictive modeling 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 modeling involves a cost-versus-benefit analysis. For example, solving a classification problem involves a search for a model that makes an optimal trade-off 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 data set. The final trained model generates individual predictions based on

https://doi.org/10.1017/psa.2025.10136 Published online by Cambridge University Press

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 trade-offs 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 trade-offs such that employing more than one criterion incorporates conflicting goals into a predictive modeling problem. Barocas et al. (2023) categorize statistical nondiscrimination criteria into three types based on 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 be independent of a sensitive attribute within each stratum of equal claim, implying error rate parity. Barocas et al. prove that formal trade-offs 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 trade-off 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 modeling 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 toward another. Thus, the challenge of achieving fairness with respect to multiple subgroups increases the number of conflicting goals in a predictive modeling problem.[1]

Philosophers and computer scientists respond to these inherent trade-offs 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 because 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

[1] Designers have proposed clever solutions to this problem, where the trade-offs between subgroups are either considered explicitly for multiple sensitive attributes (Zafar et al. 2017) or "blindly" for any subgroups of sufficient size (Martinez et al. 2021; Kearns et al. 2018).

https://doi.org/10.1017/psa.2025.10136 Published online by Cambridge University Press

compatible (Beigang 2023), or implemented as group-specific aims rather than as a summarized statistic (e.g., group distributionally robust optimization [DRO; Sagawa et al. 2020], minimax Pareto fairness [Martinez et al. 2020]). (4) The criteria should be abandoned in favor of alternative debiasing or fairness-optimization strategies, such as striving toward 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* of the proposed approaches because they involve design choices that may or may not lead to Pareto optimality (defined later), 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 trade-off 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 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 trade-off 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 trade-offs between the objectives. A MOOP can be formulated as a minimization problem with m objectives ($m2$) that map decision variables $x \in X$ in decision space X into the objective space, defined in $\mathbb{R}^m$:

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

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 the formalism of Deb et al. [2016, 151]):

1. $\mathbf{x}^{(1)}$ is no worse than $\mathbf{x}^{(2)}$ in all objectives $\left(f_j\left(\mathbf{x}^{(1)}\right) \not\succ f_j\left(\mathbf{x}^{(2)}\right)\right.$ for all $j = 1, \ldots, m$).

[2] I have outlined these responses with insight from Miconi (2017), Hedden (2021), and Beigang (2023).

https://doi.org/10.1017/psa.2025.10136 Published online by Cambridge University Press

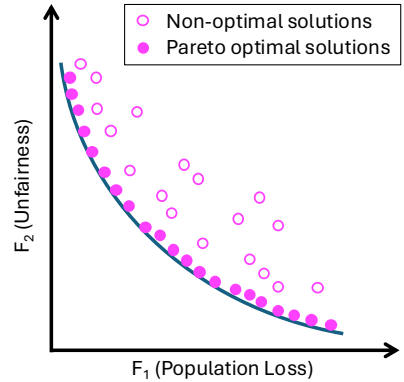


Figure 1. Trade-off solutions in objective function space when minimizing some notion of unfairness (F2) and population loss (F1).

2. $\mathbf{x}^{(1)}$ is strictly better than $\mathbf{x}^{(2)}$ in at least one objective $\left(f_{\bar{j}}\left(\mathbf{x}^{(1)}\right) \lhd f_{\bar{j}}\left(\mathbf{x}^{(2)}\right)\right)$ for at least

one $\bar{j} \in \{1, \ldots, m\}$.

The PO set comprises the nondominated solutions attained when the entire feasible region of the decision space is searched (Deb et al. 2016, 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).

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 trade-offs, after the stage of optimization.[3] Typically, a posteriori approaches aim to find a large portion of the entire PO set (Deb et al. 2016, 148). I will focus this article on a second major design choice, which concerns whether to use a search strategy that is guided toward a particular trade-off 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 an 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).

https://doi.org/10.1017/psa.2025.10136 Published online by Cambridge University Press

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 and Vicente's 2022 extension of Zafar et al. 2017) • Chebyshev scalarization (Wei and Niethammer 2022)
Unbiased Search	• Constrained multigradient descent followed by automated selection of the final solution (Padh et al. 2021)	• Stochastic multigradient descent (Liu and 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 et al. 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 and 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 because ML uses an "indirect" form of optimization (Goodfellow et al. 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

https://doi.org/10.1017/psa.2025.10136 Published online by Cambridge University Press

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 multigradient descent (with the goal of using it in an a posteriori way). They compare their method to a biased search approach (the constraint-based optimization of Zafar et al. [2017]) on two convex optimization problems. The goals are to make accurate predictions of salary and minimize disparate impact with respect to either gender or race (they use the Adult Income data set; Kohavi and Becker 1996). For accuracy, they use a logistic regression function, and for disparate impact, they use a convex approximation of Calders and Verwer's (2010; CV) score based on the decision boundary covariance. 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 Pareto front–stochastic multigradient (PF-SMG) algorithm dominates the epsilon-constraint (EPS)-fair algorithm in some regions of the trade-off surface (their fig. 1a, 521). However, when optimizing for minimal disparate impact with respect to race, the EPS-fair algorithm dominates in some regions (their fig. 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 trade-off 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 Ward (2021), values might act as nonepistemic justificatory reasons. Winsberg (2012) brings attention to how nonepistemic values affect what he calls "unforced methodological choices" in climate modeling, 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 that it is not possible to isolate the values that have affected the history of climate science because they hide in all its "nooks and crannies," and current climate models also rely on past modeling choices, methods for fairness optimization are not yet so

[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 data sets (522–23). They note that 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 because additional (and sometimes orthogonal) design choices also contribute to these advantages (e.g., incorporating stochasticity).

https://doi.org/10.1017/psa.2025.10136 Published online by Cambridge University Press

"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 nonepistemic 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 and Vicente's (2022) results demonstrate this difficulty: Their unbiased method produces a more dense and well-spread PF compared with the epsilon-constraint method (522–23). 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 narrowly, 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 with which they generate one or more PO solutions. Because 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, so they may be inefficient for problems with many objectives (e.g., fairness optimization; see sec. 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 and Vicente [2022] also measure this explicitly). 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 and 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

https://doi.org/10.1017/psa.2025.10136 Published online by Cambridge University Press

converge to the final PF computed for the whole training data set (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 with 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 might be preferred for precise control over trade-off preferences, efficient PF mapping, and dynamic PO prediction.

6. Conclusion

I emphasized that because implementing fairness notions in ML optimization is generally 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 trade-off control, robustness to distribution shifts) help justify 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 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 because 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 establishing 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 trade-offs, 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

https://doi.org/10.1017/psa.2025.10136 Published online by Cambridge University Press

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 nonextreme*).

Acknowledgments. Thanks to support from the Northeastern Artificial Intelligence and Data Ethics (AIDE) Summer Training Program 2023 and National Science Foundation Award 2147220. This article 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). I would also like to thank Chris Smeenk and Kathleen A. Creel for helpful discussions.

References

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

Beigang, Fabian. 2023. "Reconciling Algorithmic Fairness Criteria." *Philosophy & Public Affairs* 51 (2):166–90. https://doi.org/10.1111/papa.12233.

Branke, Jürgen, Kalyanmoy Deb, Henning Dierolf, and Matthias Osswald. 2004. "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: Springer. https://doi.org/10.1007/978-3-540-30217-9_73.

Calders, Toon, and Sicco Verwer. (2010). "Three Naive Bayes Approaches for Discrimination-Free Classification." *Data Mining and Knowledge Discovery* 21 (2):277–92. https://doi.org/10.1007/s10618-010-0190-x.

Chouldechova, Alexandra. (2017). "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *arXiv*. https://doi.org/10.1089/big.2016.0047.

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

Deb, Kalyanmoy, Karthik Sindhya, and Jussi Hakanen. 2016. "Multi-Objective Optimization." In *Decision Sciences, edited by Raghu Nandan Sengupta, Aparna Gupta, and Joydeep Dutta*, 145–83. Boca Raton, FL: CRC Press.

Fazelpour, Sina, and David Danks. 2021. "Algorithmic Bias: Senses, Sources, Solutions." *Philosophy Compass* 16 (8):e12760. https://doi.org/10.1111/phc3.12760.

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

Grant, David Gray. 2023. "Equalized Odds Is a Requirement of Algorithmic Fairness." *Synthese* 201 (3):101. https://doi.org/10.1007/s11229-023-04054-0.

Grant, David Gray, Jeff Behrends, and John Basl. 2025. "What We Owe to Decision-Subjects: Beyond Transparency and Explanation in Automated Decision-Making." *Philosophical Studies* 182 (1):55–85. https://doi.org/10.1007/s11098-023-02013-6.

Haimes, Yv. 1971. "On a Bicriterion Formulation of the Problems of Integrated System Identification and System Optimization." *IEEE Transactions on Systems, Man, and Cybernetics* SMC-1 (3):296–97. https://doi.org/10.1109/TSMC.1971.4308298.

Hedden, Brian. 2021. "On Statistical Criteria of Algorithmic Fairness." *Philosophy & Public Affairs* 49 (2):209–31. https://doi.org/10.1111/papa.12189.

Hellman, Deborah. 2020. "Measuring Algorithmic Fairness." *Virginia Law Review* 106 (4):811–66. https://www.jstor.org/stable/27074708.

Kamishima, Toshihiro, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. "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. Berlin: Springer. https://doi.org/10.1007/978-3-642-33486-3_3.

Kearns, Michael, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness." In *Proceedings of the 35th*

https://doi.org/10.1017/psa.2025.10136 Published online by Cambridge University Press

International Conference on Machine Learning, 2564–72. Cambridge, MA: MLR Press. https://proceedings.mlr.press/v80/kearns18a.html.

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

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

Liu, Suyun, and Luis Nunes Vicente. 2022. "Accuracy and Fairness Trade-Offs in Machine Learning: A Stochastic Multi-Objective Approach." *Computational Management Science* 19 (3):513–37. https://doi.org/10.1007/s10287-022-00425-z.

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

Martinez, Natalia L., Martin A. Bertran, Afroditi Papadaki, Miguel Rodrigues, and Guillermo Sapiro. 2021. "Blind Pareto Fairness and Subgroup Robustness." In *Proceedings of the 38th International Conference on Machine Learning*, 7492–501. Cambridge, MA: MLR Press. https://proceedings.mlr.press/v139/martinez21a.html.

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

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

Padh, Kirtan, Diego Antognini, Emma Lejal-Glaude, Boi Faltings, and Claudiu Musat. 2021. "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. Cambridge, MA: MLR Press. https://proceedings.mlr.press/v161/padh21a.html.

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

Wan, Mingyang, Daochen Zha, Ninghao Liu, and Na Zou. 2023. "In-Processing Modeling Techniques for Machine Learning Fairness: A Survey." *ACM Transactions on Knowledge Discovery from Data* 17 (3):1–27. https://doi.org/10.1145/3551390.

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

Wei, Susan, and Marc Niethammer. 2022. "The Fairness-Accuracy Pareto Front." *Statistical Analysis and Data Mining: An ASA Data Science Journal* 15 (3):287–302. https://doi.org/10.1002/sam.11560.

Winsberg, Eric. 2012. "Values and Uncertainties in the Predictions of Global Climate Models." *Kennedy Institute of Ethics Journal* 22 (2):111–37. https://doi.org/10.1353/ken.2012.0008.

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

Cite this article: Champion, Heather. 2025. "On Values in Fairness Optimization with Machine Learning." *Philosophy of Science* 92 (5):1372–1382. https://doi.org/10.1017/psa.2025.10136

https://doi.org/10.1017/psa.2025.10136 Published online by Cambridge University Press