Metaheuristic Algorithms with Applications

www.maa.reapress.com

Metaheur. Algor. Appl. Vol. 2, No. 4 (2026) 344–352.

Paper Type: Original Article

Multi-Objective Optimization in Machine Learning: Balancing Accuracy, Fairness, and Interpretability

Amir Khanjani*

Department of Computer Engineering, Ayandegan University, Tonkabon, Iran; amir.khanjani@aihe.ac.ir.

Citation:

Received: 17 ceptember 2024 *Revised: 16 december 2024* *Accepted: 02 february 2025*	Khanjani, A. (2026). Multi-objective optimization in machine learning: Balancing accuracy, fairness, and interpretability. *Metaheuristic algorithms with applications, 2*(4), 344-352.

Abstract

Machine Learning (ML) systems are frequently evaluated and optimized for predictive performance, yet real-world deployment increasingly requires simultaneous attention to fairness and interpretability. These requirements introduce objectives that can conflict in both theory and practice: Improving accuracy can exacerbate disparities across protected groups; enforcing fairness constraints can reduce utility or shift error burdens; and enhancing interpretability can restrict hypothesis classes or encourage post-hoc explanations whose fidelity is uncertain. This article treats this as a Multi-Objective Optimization (MOO) problem, emphasizing Pareto optimality, the structure of trade-offs, and decision-making on the resulting Pareto set. We review core fairness definitions and metrics, interpretability concepts and pitfalls, and MOO methods used to manage competing objectives, such as including scalarization, constrained learning, and evolutionary approaches. We then propose an evaluation and reporting framework centered on transparent visualization of trade-offs (pareto fronts and fairness-utility curves), careful metric selection (including dominance-based indicators), and documentation practices inspired by model reporting standards.

Keywords: Multi-objective optimization, Pareto optimality, Fairness, Demographic parity, Equalized odds, Calibration, Interpretability, Explainable artificial intelligence, Constrained learning, Model reporting.

1 | Introduction

Modern Machine Learning (ML) practice rarely confronts a single "best" model. In deployment, models are embedded in human contexts where costs are multidimensional: Errors have heterogeneous consequences across groups, stakeholders demand accountability and transparency, and legal or organizational requirements can impose constraints beyond predictive performance. "better accuracy" is not a sufficient design criterion. Teams face a portfolio of candidate models that trade off accuracy, fairness, and interpretability in ways that cannot be simultaneously optimized to a single optimum [1–3]. Multi-Objective Optimization (MOO)

Corresponding Author: amir.khanjani@aihe.ac.ir

https://doi.org/10.48313/maa.v2i3.55

provides a principled framework for navigating these trade-offs explicitly, rather than allowing them to be resolved implicitly through ad hoc choices.

1.1 | Context and Motivation

From a technical perspective, fairness definitions can be mutually incompatible under realistic conditions [2–6]. Calibration, Equalized Odds (EO), and Demographic Parity (DP) cannot in general be simultaneously satisfied when base rates differ across groups. Additionally, interpretability constrains the hypothesis class: Restricting models to linear or rule-based forms limits the function space that can be explored. From an organizational perspective, ML teams negotiate among stakeholders with different priorities sometimes legally mandated. The EU Artificial Intelligence (AI) act [7] imposes transparency and fairness obligations on high-risk systems, while US fair lending laws require adverse-action explanations. Without explicit multi-objective formulation, trade-offs are made implicitly often by the engineer who selects the loss function or threshold, without consulting affected communities or documenting the reasoning.

1.2 | Scope and Structure

This paper is organized as follows. Section 2 formalizes the MOO framework. Section 3 reviews fairness definitions and metrics, including key impossibility results. Section 4 addresses interpretability concepts, methods, and pitfalls. Section 5 surveys MOO methods applicable to ML, including scalarization, constrained learning, evolutionary algorithms, and multi-gradient approaches. Section 6 proposes an evaluation and reporting framework with visualization tools, metric selection guidance, and a documentation checklist, and Section 7 concludes and identifies directions for future work.

2 | Multi-Objective Optimization: Formal Framework

Standard ML training minimizes a scalar loss function $L(\theta)$ over parameters θ. However, when multiple objectives $f_1(\theta), f_2(\theta), \ldots, f_K(\theta)$ are relevant, generally no single θ^* minimizes all objectives simultaneously [4]. The MOO problem is stated as

Input:

I. Feasible parameter space Θ.

II. Objective functions f1, f2, ..., fK.

Output: Pareto optimal set P*.

Step 1. Define objective vector.

For any solution $\theta \in \Theta$: $F(\theta) \leftarrow [f1(\theta), f2(\theta), \ldots, fK(\theta)]$.

Step 2. Define pareto dominance.

```
Function DOMINATES(a, b):
    Input: two solutions a, b ∈ Θ
    For i = 1 to K:
        if fi(a) > fi(b):
            return FALSE
    If there exists j such that fj(a) < fj(b):
        return TRUE
    Else:
        return FALSE.
```

Step 3. Construct Pareto front.

Initialize P* ← ∅

For each solution θ ∈ Θ:

dominated ← FALSE

For each solution θ' ∈ Θ:

If DOMINATES(θ', θ) == TRUE:

dominated ← TRUE

break

If dominated == FALSE:

Add θ to P*

Return P*

The key insight of the MOO framework is that it separates the optimization step identifying the Pareto front from the decision-making step selecting a preferred solution from the front. The optimization step is a technical problem that can be addressed algorithmically. The decision-making step requires human judgment, domain expertise, and stakeholder input. This separation makes trade-offs explicit and auditable. In the context of ML fairness and interpretability, the objectives might include: 1) minimizing predictive loss (e.g., cross-entropy, mean squared error), 2) minimizing a fairness violation metric (e.g., DP difference), and 3) maximizing an interpretability measure (e.g., model sparsity or a fidelity score for post-hoc explanations).

3 | Fairness: Definitions, Metrics, and Tensions

Fairness in ML has been formalized through multiple criteria, each capturing a different ethical intuition about what it means for a predictive system to treat individuals and groups equitably [8]. We review the major definitions and their known tensions.

3.1 | Group Fairness Criteria

Let A denote a binary sensitive attribute (e.g., race, gender), Y the true label, $\hat{Y}$ the predicted label, and S a score output by the classifier. The following group fairness definitions are widely used:

I. DP: $P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1)$. The positive classification rate should be equal across groups, regardless of the true label distribution.

II. EO: $P(\hat{Y} = 1 \mid Y = y, A = a)$ is equal across groups for $y \in \{0, 1\}$. Both true positive rates and false positive rates should be equal across groups [1].

III. Calibration: $P(Y = 1 \mid S = s, A = a) = s$ for all groups. Among individuals assigned score s, the proportion who are truly positive should equal s, regardless of group membership [6].

Impossibility results: A fundamental finding in the fairness literature is that, except in degenerate cases (equal base rates or perfect prediction), DP, EO, and calibration cannot all be satisfied simultaneously [2], [5]. This impossibility result is not a limitation of any particular algorithm but a mathematical consequence of the structure of the problem. It implies that any real-world system must choose among fairness criteria a normative decision that cannot be resolved by technical means alone.

3.2 | Individual and Causal Fairness

Individual fairness [9] requires that similar individuals receive similar predictions: if d(x, x') is a task-specific similarity metric over individuals, then $|f(x) - f(x')|$ should be small whenever d(x, x') is small. The practical

challenge lies in defining the metric d, which encodes substantive judgments about which differences between individuals are morally relevant.

Counterfactual fairness requires that a prediction remain unchanged under a hypothetical intervention that changes the sensitive attribute. Formally, $P(\hat{Y}_{A\leftarrow a} \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'} \mid X = x, A = a)$ for all a, a'. This requires a causal model of the data-generating process, which may be unavailable or contested.

3.3 | Fairness-Accuracy Trade-Offs

Enforcing group fairness constraints typically reduces overall predictive accuracy, because the unconstrained optimum generally violates fairness criteria when base rates differ across groups. The Pareto front between accuracy and fairness violation reveals a characteristic structure: small improvements in fairness near the unconstrained optimum are "cheap" in terms of accuracy loss, while achieving near-parity becomes increasingly expensive as the fairness constraint tightens. This diminishing-returns structure has been empirically observed across datasets and fairness criteria [10], [8] and has important implications for practice: Organizations can often achieve meaningful fairness improvements with modest accuracy costs, but eliminating the last residual disparity may require substantial sacrifices.

4 | Interpretability: Concepts, Methods, and Pitfalls

4.1 | Definitions

Lipton [11] distinguishes between transparency the degree to which a model's internal mechanisms are understandable to a human and post-hoc explainability the degree to which a separate method can provide insight into a model's behavior after training. Transparency is a property of the model itself; explainability is a property of an explanation method applied to a model.

Inherently interpretable models include linear models, logistic regression, decision trees, rule lists, and generalized additive models. Their structure permits direct inspection of how inputs map to outputs. Post-hoc explanation methods include Local Interpretable Model-agnostic Explanations (LIME) [12], Shapley Additive Explanations (SHAP) [13], attention visualization, saliency maps, and counterfactual explanations. These methods approximate or summarize the behavior of complex models without altering the model itself.

4.2 | Interpretability-Accuracy Trade-Off

The conventional wisdom holds that simpler, more interpretable models sacrifice predictive accuracy relative to complex black-box models. However, Rudin [3] argues compellingly that for structured and tabular data, inherently interpretable models can match or closely approach the performance of black-box models when sufficient effort is invested in feature engineering and model selection. The accuracy gap is often smaller than assumed, particularly in domains where the underlying relationships are approximately linear or additive.

In contrast, for unstructured data (images, natural language, audio), deep neural networks with millions or billions of parameters currently dominate, and interpretable alternatives lag significantly. In these domains, post-hoc explanation methods are the primary tool for understanding model behavior, despite their known limitations.

4.3 | Pitfalls of Post-Hoc Explanation

Several recent findings raise concerns about the reliability of post-hoc explanations:

I. Instability and manipulability: Slack et al. [14] demonstrated that adversarial classifiers can be constructed to produce arbitrary LIME and SHAP explanations while maintaining discriminatory behavior. This shows that post-hoc explanations can be "gamed" to hide unfairness.

II. Attention as explanation: Jain and Wallace [15] showed that attention weights in neural networks do not reliably indicate feature importance. Alternative attention distributions can produce identical predictions, undermining the use of attention as an explanation mechanism.

III. Interpretability theater: There is a risk that post-hoc explanations serve primarily to satisfy regulatory checkboxes or stakeholder expectations without providing genuine understanding of model behavior. Organizations may deploy explanations that appear transparent but do not actually support meaningful oversight.

These findings reinforce Rudin's [3] argument for preferring inherently interpretable models in high-stakes applications where possible, and for exercising caution when relying on post-hoc methods.

5 | Multi-Objective Optimization Methods for Machine Learning

This section surveys the principal MOO methods applicable to ML, organized by algorithmic paradigm.

5.1 | Scalarization

Scalarization converts a multi-objective problem into a single-objective problem by combining objectives into a scalar function.

Weighted sum: Minimize $\sum_k w_k f_k(\theta)$, where $w_k \geq 0$ and $\sum w_k = 1$. By varying the weight vector, different Pareto-optimal solutions can be recovered. However, linear scalarization can only find solutions on the convex hull of the Pareto front [4]. Solutions in non-convex regions are inaccessible.

Chebyshev scalarization: Minimize $\max_k w_k |f_k(\theta) - z_k^*|$, where z^* is a reference point (e.g., the ideal point). This approach can find solutions on non-convex Pareto regions and provides full coverage of the front.

Epsilon-constraint method: Minimize $f_1(\theta)$ subject to $f_k(\theta) \leq \varepsilon_k$ for $k = 2, ..., K$. By systematically varying the ε bounds, the Pareto front can be traced. This method also achieves full Pareto coverage.

5.2 | Constrained Learning

A natural approach for fairness-aware ML is to treat fairness as a constraint rather than an objective: minimize $L(\theta)$ subject to fairness_metric$(\theta) \leq \varepsilon$. This formulation directly controls the maximum allowable fairness violation while optimizing predictive performance within that bound.

Lagrangian relaxation introduces dual variables λ to convert the constrained problem into an unconstrained saddle-point problem: $\min_\theta \max_{\lambda \geq 0} L(\theta) + \lambda.$ (fairness_metric$(\theta) - \varepsilon$). Practical implementations include projected gradient descent and the exponentiated gradient method of Agarwal et al. [10], which reduces fair classification to a sequence of cost-sensitive classification problems. The exponentiated gradient approach is particularly attractive because it can wrap any base classifier and provides convergence guarantees.

5.3 | Evolutionary and Population-Based Methods

Evolutionary multi-objective algorithms maintain a population of solutions and evolve them toward the Pareto front over successive generations.

NSGA-II uses non-dominated sorting to rank solutions by dominance level and crowding distance to maintain diversity along the front [16]. It is the most widely used evolutionary MOO algorithm and has been applied to hyperparameter optimization, architecture search, and fair model selection.

MOEA/D decomposes the multi-objective problem into a set of scalar subproblems using weight vectors and solves them simultaneously, exploiting neighborhood structure [17]. It is particularly effective for problems with many objectives (> 3).

Evolutionary methods have the advantage of naturally producing diverse, well-distributed Pareto fronts in a single run. Their primary limitation is computational cost: Each generation requires evaluating the full population, which can be prohibitive for large ML models requiring expensive training procedures.

Table 1. Comparison of MOO approaches.

Method	Pareto Coverage	Computational Cost	Ease of Integration
Weighted sum scalarization	Partial (convex regions only)	Low	High
Chebyshev scalarization	Full	Moderate	Moderate
Epsilon-constraint	Full	Moderate	Moderate
Lagrangian relaxation	Partial	Low–moderate	High
NSGA-II	Full	High	Low
MOEA/D	Full	High	Low

5.4 | Multi-Gradient Methods

Multiple-Gradient Descent Algorithm (MGDA) finds a single descent direction that reduces all objectives simultaneously by solving a minimum-norm problem in the convex hull of individual objective gradients [18]. When such a common descent direction exists, MGDA provides an efficient, gradient-based path toward the Pareto front without requiring weight specification.

Pareto MTL extends this idea to multi-task learning [19]. It generates a set of well-distributed Pareto-optimal solutions by decomposing the multi-objective problem into constrained subproblems, each targeting a different region of the Pareto front. The result is a diverse set of models that span the trade-off surface, enabling practitioners to select the solution that best matches their priorities.

6 | Evaluation and Reporting Framework

A multi-objective approach to ML is only as useful as the framework used to evaluate, visualize, and communicate its results. We propose a structured evaluation and reporting framework inspired by model cards [20] and datasheets for datasets [21].

6.1 | Visualization

Effective visualization is essential for communicating trade-offs to diverse stakeholders. We recommend the following visualization tools:

I. Pareto front plots (2D and 3D): Scatter plots of objective values for all candidate solutions, with Pareto-optimal solutions highlighted. These are the primary tool for communicating the trade-off structure.

II. Fairness-utility curves: Plots of accuracy (or other utility metrics) against fairness violation for a sweep of constraint thresholds, illustrating the cost of fairness.

III. Radar/spider charts: Multi-metric comparison of selected models on a common scale, useful for presenting multiple objectives simultaneously.

IV. Parallel coordinate plots: For high-dimensional objective spaces (>3 objectives), parallel coordinate plots allow visualization of the relationships among all objectives across the solution set.

6.2 | Metric Selection

Evaluating the quality of a Pareto front approximation requires specialized indicators:

I. Dominance-based indicators: The hypervolume indicator measures the volume of objective space dominated by the Pareto front relative to a reference point. It is the only unary indicator that is strictly monotone with respect to Pareto dominance. The epsilon-indicator measures the minimum factor by which every point in a reference set must be multiplied to be dominated by the approximation set.

II. Distribution indicators: Spacing measures the uniformity of the distribution of solutions along the front. Spread measures the extent of the front, ensuring solutions cover the full range of trade-offs.

III. Per-group metrics: Beyond aggregate performance, report all relevant metrics disaggregated by protected group: true positive rate, false positive rate, positive predictive value, and calibration error for each group.

6.3 | Reporting Checklist

We propose the following checklist for reporting multi-objective ML results, designed to promote transparency, reproducibility, and accountability:

Table 2. MOO reporting checklist.

Category	Item	Description
Objectives	Objective definitions	Formal definitions of each objective function, including mathematical formulation and units
Objectives	Trade-off structure	Characterization of conflicts between objectives, including known impossibility results
Data	Protected attributes	List of sensitive attributes and their distributions in training and evaluation data
Data	Data splits	Training/validation/test split strategy, including stratification and temporal considerations
Method	MOO algorithm	Algorithm used and justification for its selection over alternatives
Method	Hyperparameters	All MOO-specific hyperparameters (weights, constraint thresholds, population sizes, etc.)
Results	Pareto front	Visualization and tabulation of Pareto-optimal solutions with objective values
Results	Selected solution	Final model choice from the Pareto set and rationale for selection
Results	Fairness metrics	Per-group performance metrics for the selected model and key alternatives
Limitations	Known gaps	Acknowledged limitations, failure modes, and conditions under which results may not generalize

6.4 | Decision-Making on the Pareto Set

Once the Pareto front has been identified, a decision-maker must select a specific solution for deployment. Several structured approaches exist:

I. Knee-point selection: Identify the "knee" of the Pareto front the point of maximum marginal rate of substitution, where small improvements in one objective require increasingly large sacrifices in another. This represents a natural compromise.

II. Stakeholder preference elicitation: Systematically gather preferences from relevant stakeholders (e.g., domain experts, affected communities, compliance officers) and use these to define weights or aspiration levels.

III. Lexicographic ordering: Rank objectives by priority and optimize sequentially: first optimize the most important objective, then optimize the second within the set of solutions that are optimal for the first, and so on.

IV. Satisficing thresholds: Define minimum acceptable levels for each objective and select from solutions that meet all thresholds. If multiple solutions satisfy all constraints, use a secondary criterion (e.g., robustness, simplicity) to select among them.

7 | Conclusion

MOO provides the principled framework needed to manage the inherent trade-offs among accuracy, fairness, and interpretability in ML systems. In this paper, we have reviewed the formal MOO framework, including Pareto dominance and optimality; fairness definitions and their impossibility results, which demonstrate that

normative choices among criteria are unavoidable; interpretability concepts and the pitfalls of post-hoc explanation methods; and practical MOO methods ranging from scalarization and constrained learning to evolutionary algorithms and multi-gradient approaches.

Our proposed evaluation and reporting framework emphasizes transparency through Pareto front visualization, careful metric selection using dominance-based and distributional indicators, and standardized documentation modeled on model cards and datasheets. By making trade-offs explicit and auditable, this framework supports more responsible and informed decision-making in ML system design and deployment.

Future work should address several open challenges: 1) scalability of evolutionary and population-based methods to large foundation models with billions of parameters; 2) integration of MOO with fine-tuning and adaptation of pre-trained foundation models; 3) dynamic and online MOO for deployment-time adaptation as data distributions shift; 4) causal fairness formulations that go beyond observational criteria to reason about the mechanisms of discrimination; and 5) the development of regulatory compliance frameworks that operationalize legal requirements (such as the EU AI Act) through multi-objective formulations with verifiable guarantees.

References

[1] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *30th conference on neural information processing systems (NIPS 2016)* (PP. 1-9). Neural Information Processing Systems Foundation. https://proceedings.neurips.cc/paper_files/paper/2016/file/6a9659feb1216f14f7384ba499518b38-Paper.pdf

[2] Kleinberg, J. M., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *Information technology convergence and services* (pp. 43:1-43:23). Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[3] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[4] Marler, R. T., & Arora, J. S. (2004). Survey of multi-objective optimization methods for engineering. *Structural and multidisciplinary optimization*, *26*(6), 369–395. https://doi.org/10.1007/s00158-003-0368-6

[5] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

[6] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in neural information processing systems* (pp. 5680–5689). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/b8b9c74ac526fffbeb2d39ab038d1cd7-Abstract.html

[7] EU AI Act. (2024). *Regulation (EU) 2024/2847 of the european parliament and of the council*. http://hctinsight.com/webzine/webzine/202501/file/ce/ce5.pdf

[8] Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press. https://books.google.com/books?id=HuGwEAAAQBAJ

[9] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214–226). Association for Computing Machinery. https://doi.org/10.1145/2090236.2090255

[10] Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 60–69). PMLR. https://proceedings.mlr.press/v80/agarwal18a.html

[11] Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31-57. https://doi.org/10.1145/3233231

[12] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you? Explaining the predictions of any classifier. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

[13] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777). Curran Associates Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[14] Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post HOC explanation methods. *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 180–186). Association for Computing Machinery. https://doi.org/10.1145/3375627.3375830

[15] Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies (long and short papers)*. Association for Computational Linguistics. (pp. 3543-3556). https://doi.org/10.18653/v1/N19-1357

[16] Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation*, *6*(2), 182–197. https://doi.org/10.1109/4235.996017

[17] Zhang, Q., & Li, H. (2008). MOEA/D: A multiobjective evolutionary algorithm based on decomposition. *Evolutionary computation, IEEE transactions on*, *11*, 712–731. https://doi.org/10.1109/TEVC.2007.892759

[18] Mahapatra, D., & Rajan, V. (2020). Multi-task learning with user preferences: Gradient descent with controlled ascent in pareto optimization. *Proceedings of the 37th international conference on machine learning* (pp. 6597–6607). Journal of Machine Learning Research (JMLR). https://doi.org/10.5555/3524938.3525550

[19] Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, S. K. (2019). Pareto multi-task learning. *Advances in neural information processing systems* (pp. 12037–12047). Neural Information Processing Systems Foundation (NeurIPS Foundation). https://proceedings.neurips.cc/paper_files/paper/2019/hash/685bfde03eb646c27ed565881917c71c-Abstract.html

[20] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., … & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220–229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

[21] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://dx.doi.org/10.1145/3458723