Can Fairness be Automated? Guidelines and Opportunities for Fairness-aware AutoML

Hilde Weerts   H.J.P.WEERTS@TUE.NL
Eindhoven University of Technology

Florian Pfisterer   PFISTERERF@GOOGLEMAIL.COM
Ludwig-Maximilians-Universität München
Munich Center for Machine Learning

Matthias Feurer   FEURERM@CS.UNI-FREIBURG.DE
Albert-Ludwigs-Universität Freiburg

Katharina Eggensperger   KATHARINA.EGGENSPERGER@UNI-TUEBINGEN.DE
Albert-Ludwigs-Universität Freiburg
University of Tübingen

Edward Bergman   BERGMANE@CS.UNI-FREIBURG.DE
Noor Awad   AWAD@CS.UNI-FREIBURG.DE
Albert-Ludwigs-Universität Freiburg

Joaquin Vanschoren   J.VANSCHOREN@TUE.NL
Mykola Pechenizkiy   M.PECHENIZKIY@TUE.NL
Eindhoven University of Technology

Bernd Bischl   BERND.BISCHL@STAT.UNI-MUENCHEN.DE
Ludwig-Maximilians-Universität München
Munich Center for Machine Learning

Frank Hutter   FH@CS.UNI-FREIBURG.DE
Albert-Ludwigs-Universität Freiburg

Abstract

The field of automated machine learning (AutoML) introduces techniques that automate parts of the development of machine learning (ML) systems, accelerating the process and reducing barriers for novices. However, decisions derived from ML models can reproduce, amplify, or even introduce unfairness in our societies, causing harm to (groups of) individuals. In response, researchers have started to propose AutoML systems that jointly optimize fairness and predictive performance to mitigate fairness-related harm. However, fairness is a complex and inherently interdisciplinary subject, and solely posing it as an optimization problem can have adverse side effects. With this work, we aim to raise awareness among developers of AutoML systems about such limitations of fairness-aware AutoML, while also calling attention to the potential of AutoML as a tool for fairness research. We present a comprehensive overview of different ways in which fairness-related harm can arise and the ensuing implications for the design of fairness-aware AutoML. We conclude that while fairness cannot be automated, fairness-aware AutoML can play an important role in the toolbox of ML practitioners. We highlight several open technical challenges for future work in this direction. Additionally, we advocate for the creation of more user-centered assistive systems designed to tackle challenges encountered in fairness work.

1. Introduction

Machine learning (ML) is a game-changing technology that has disrupted modern data-driven applications and is increasingly deployed in various applications and contexts. However, ML systems may reproduce, amplify, or even introduce unfairness in our society, causing harm to (groups of) individuals. Examples range from facial recognition systems that disproportionately fail for darker-skinned women (Buolamwini & Gebru, 2018), gender bias in automatic captions (Tatman, 2017) and resume parsing (Dastin, 2018), to the underestimation of the healthcare needs of black patients (Obermeyer et al., 2019). In response to the growing need for ML systems that align with principles of fairness, researchers have proposed numerous techniques to assess and mitigate unfairness of ML systems. Additionally, several open-source software libraries facilitate the application of these methods (Bellamy et al., 2019; Weerts et al., 2023). In practice, however, it can be challenging to incorporate fairness considerations in the design and development of ML systems, due to the following obstacles:

1. *Potential fairness-related harms are rarely prioritized* due to a lack of awareness, difficulties in anticipating fairness-related harm, or difficulties in advocating for required resources within an organization (Madaio et al., 2022).

2. *Practitioners are overwhelmed with a multitude of fairness metrics and interventions.* Choosing an appropriate metric or intervention is an active area of research and is further complicated by practical obstacles, such as the feasibility of collecting more (representative) data, the accessibility of descriptive meta-data, and the availability of cultural, contextual, and domain knowledge (Holstein et al., 2019).

3. *Once found, solutions rarely carry over to new problems.* Given the complexity and diversity of systems, contexts, and use and failure cases, fairness-aware models cannot be repurposed within a different context (Selbst et al., 2019).

The case for fairness-aware AutoML The rising field of AutoML (Hutter et al., 2019) focuses on reducing the complexity inherent to applying ML algorithms in practice by providing methods that partially automate the ML workflow. In response to fairness concerns, recent research has started to propose AutoML systems that take into account fairness objectives. By lowering the barrier to incorporate fairness considerations in an ML workflow, AutoML systems might be able to partially address several of the aforementioned problems, as they reduce the ML expertise required to build, interrogate and evaluate fairness-aware ML systems:

1. *Fairness-aware AutoML can reduce the barrier to entry.* It can substantially reduce turnaround times for developing and evaluating models along different metrics, lowering the amount of required resources, time, and technical expertise required to engage in fairness work, and enabling an iterative preference elicitation process for determining what precisely it means to be fair in a particular application.

2. *Fairness-aware AutoML can facilitate a declarative interface* for domain experts to specify *what* they would like to achieve, rather than *how* to do so. This relieves them of the necessity of staying up to date with the latest technical fairness interventions.

3. *In contrast to individual fairness-aware models, the fairness-aware AutoML process for finding such models carries over to new problems much better.* This is the entire point of AutoML: finding custom technical solutions for new problems that best satisfy the problem's particular desiderata.

However, fairness is an intricate and inherently interdisciplinary subject. Existing fairness-aware ML algorithms often operate under the assumption that fair outcomes can be achieved by optimizing for an additional fairness metric, but reality is much more complex. In particular, the scope of these algorithms is typically limited, while potential harm can originate at each step of the ML workflow. Considering fairness-aware *Auto*ML, we have a potentially more flexible and powerful tool to aid practitioners in fairness work. However, care should be taken to avoid replicating or even exacerbating known issues with existing approaches in fairness-aware ML. This raises the question: *can fairness be automated?*

Contributions   With this work, we aim to raise awareness on the opportunities and limitations of *fairness-aware AutoML*.[1] Our main contribution is a comprehensive overview of important challenges in fairness-aware ML as they apply to AutoML and the ensuing implications for the design of fairness-aware AutoML systems. Additionally, we highlight several challenges and opportunities unique to fairness-aware *Auto*ML and lay a foundation for future work toward more user-centered assistive systems. We focus on supervised learning tasks, noting that generative AI shares many of the challenges we outline but requires additional considerations. With this work, we hope to start a discussion among the research communities of fairness and AutoML to jointly work towards mitigating bias and harm in ML.

Outline   The remainder of this article is structured as follows. Section 2 covers the relevant background knowledge on algorithmic fairness and AutoML and provides a review of existing work in the domain of fairness-aware AutoML. We then cover existing challenges and their implications along two key perspectives: user input and interaction (Section 3) and the design of the AutoML system itself (Section 4). Section 5 then identifies opportunities that arise from the use of AutoML in fairness contexts. Section 6 concludes with a set of guidelines for the design & use of fairness-aware AutoML and an inventory of directions for future work.

2. Background

To better understand the role of AutoML in a fairness-aware ML workflow, we first need to determine what such a workflow looks like. Figure 1 presents a typical ML workflow consisting of five highly iterative stages:

1. *Problem understanding.*[2]   The problem is scoped and translated into an ML task including real-world success criteria, requirements, and constraints.

---

1. To highlight that AutoML is not a panacea for algorithmic fairness, and to emphasize that employing *fair AutoML* systems does not automatically ensure fair solutions, we recommend referring to AutoML systems that incorporate fairness considerations as *fairness-aware* AutoML, rather than *fair* AutoML.
2. In the original CRISP-DM process model, Shearer (2000) refer to this stage as *business* understanding. To emphasize that ethical concerns are central to responsible design, rather than an add-on to business imperatives, we refer to the stage as *problem* understanding.

2. *Data understanding and preparation.*[3] It is determined which data is required, followed by an iterative sequence of data collection, exploration, cleaning, and aggregation.

3. *Modeling.* The model selection pipeline is designed, and candidate models are evaluated against technical performance metrics and constraints, with a focus on factors such as accuracy and generalization.

4. *Evaluation.* The selected model is evaluated more broadly against real-world success criteria and requirements, including, e.g., A/B tests, user tests, or adversarial tests. Note that this evaluation goes beyond the technical evaluation performed during the *modeling* stage and explicitly considers whether the model meets real-world objectives.

5. *Deployment.* The deployment, monitoring, and maintenance are planned and executed.

In the remainder of this section, we provide an introduction to algorithmic fairness and AutoML and highlight how each of them plays a part in the ML workflow. We focus on supervised learning tasks with access to group memberships.

2.1 Algorithmic Fairness

The goal of algorithmic fairness is to ensure that the real-world outcomes of ML systems are fair with respect to legal, regulatory, or ethical notions of fairness. This is particularly relevant in the context of automated decision-making, where ML models are used to assist decision-makers or even obtain decisions automatically.

2.1.1 Fairness Metrics

Algorithmic fairness metrics typically measure the extent to which some form of equality is violated. The metrics differ primarily in terms of *what* should be equal. In the context of supervised learning, the two most prominent notions of fairness are group fairness and individual fairness. *Group fairness* is a notion of fairness that requires particular group statistics to be equal across (sub)groups defined by *sensitive features*. Sensitive features intend to measure characteristics of individuals for which disparate outcomes based on that characteristic are considered undesirable from an ethical or legal point of view. Typical examples are age, sex, disability status, ethnic origin, race, or sexual orientation. Researchers have defined various group fairness metrics that differ in terms of which group statistic should be equal, typically involving either the distribution of predicted outcomes (e.g. selection rate, average predicted value) or the predictive performance of the model (e.g. accuracy, precision, recall).

Metrics of *individual fairness* take the perspective of the individual, inspired by the Aristotelian principle that "like cases should be treated alike". Individual fairness metrics differ primarily in terms of what is considered 'similar'. Statistical interpretations quantify the similarity of two instances in terms of a similarity metric (Dwork et al., 2012) that aims to capture similarity on task-relevant characteristics (which typically excludes sensitive characteristics). Contrarily, counterfactual fairness (Kusner et al., 2017) takes a causal

---

3. In the original CRISP-DM model, data understanding and data preparation are separate stages. For ease of presentation, we have consolidated these in one step.
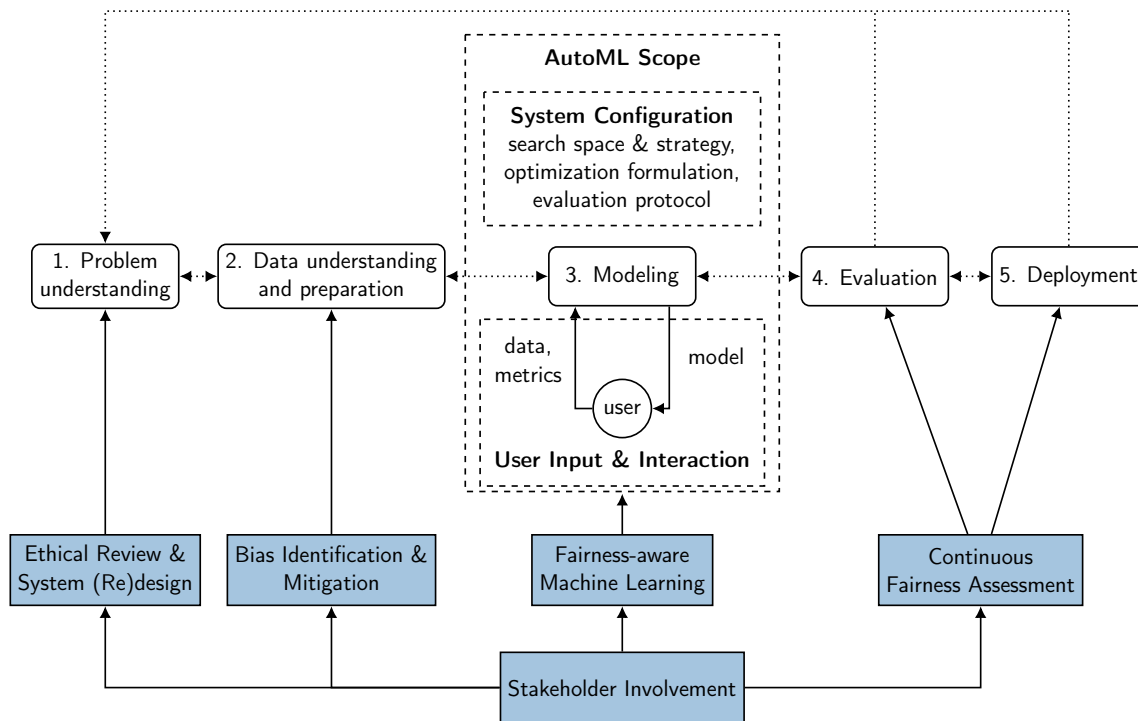
Figure 1: Example of an ML workflow adapted from the CRISP-DM (Shearer, 2000) process. Developing ML systems is an iterative process (dotted arrows) that can require frequently revisiting decisions made in previous stages. Most existing fairness-aware AutoML methods and systems address fairness analogous to fairness-aware ML techniques: the problem is formulated as an optimization task under a fairness objective or constraint. However, many important design choices are made outside of the *modeling* stage which is typically the part of the workflow that is tackled by AutoML systems. Taking fairness into account adds additional considerations (in blue) to every step of the ML workflow.

perspective, requiring similar treatment for an individual and the counterfactual of that same individual, had they belonged to a different sensitive group.

Most research has focused on group fairness. In particular, the two most prominent group fairness metrics are *demographic parity* and *equalized odds*. We denote with random variables the predicted outcome R, the true outcome Y, and the membership of a sensitive group A.

Demographic Parity Demographic parity (Calders & Verwer, 2010) requires that the probability of predicting the positive class is independent of the sensitive class A: $P(R = 1|A = a) = P(R = 1)$, essentially requiring independence (Barocas et al., 2019) between the sensitive feature and the predicted outcome: $R \perp A$. Demographic parity is satisfied when the selection rate is equal across groups. For example, in a resume selection scenario, demographic parity holds if the proportion of selected resumes is the same for each sensitive group.

Demographic parity does not take into account the true label Y. If base rates are different across groups, i.e., $P(Y = 1|A = a) \neq P(Y = 1)$, satisfying demographic parity requires

one to make predictions that do not coincide with the observed outcomes, meaning that demographic parity rules out a perfect predictor. The metric can be classified as what Wachter et al. (2020) refers to as a bias-*transforming* metric: optimizing for this metric corresponds to changing the (observed) status quo.

Equalized Odds Equalized odds (Hardt et al., 2016) is one of the most commonly studied fairness metrics that requires an equal distribution of errors across sensitive groups. In particular, it asks for equal true and false positive rates across groups: $P(R = 1|Y = 1, A = a) = P(R = 1|Y = 1) \land P(R = 1|Y = 0, A = a) = P(R = 1|Y = 0)$. This essentially requires independence between the sensitive feature and the predicted outcome, conditional on the true outcome: $R \perp A \,|\, Y$ (Barocas et al., 2019).

As opposed to demographic parity, equalized odds does explicitly take into account Y. It is therefore what Wachter et al. (2020) refers to as a bias-*preserving* metric: optimizing for equalized odds will preserve the status quo as much as possible, implicitly assuming that any bias present in the data should be preserved.

2.1.2 FAIRNESS-AWARE MACHINE LEARNING ALGORITHMS

Interventions to mitigate fairness-related harm can take place during all stages of an ML workflow (Mitchell et al., 2021; Mehrabi et al., 2021; Raji et al., 2020; Madaio et al., 2020). Such interventions include ethical reviews of a problem definition, identifying and mitigating potential biases in the data, exhaustive fairness assessments during evaluation and deployment, and active stakeholder involvement during all stages (see Figure 1). However, the vast majority of the algorithmic fairness literature has focused on fairness-aware ML algorithms applied during the modeling stage (Holstein et al., 2019). Such technical interventions formulate fairness as an optimization task, where the goal is to achieve high predictive performance whilst also satisfying a fairness constraint. They can be roughly subdivided into three categories (Kamiran et al., 2013).

Pre-processing Pre-processing approaches can adjust the data to obscure any undesirable associations between sensitive features and a target variable (e.g., Kamiran & Calders, 2011).

Constrained learning Constrained learning techniques directly incorporate a fairness constraint in the learning algorithm, either by adapting existing learning paradigms (e.g., Calders et al., 2009; Zafar et al., 2017) or through wrapper methods (e.g., Agarwal et al., 2018).

Post-processing Post-processing approaches adjust a trained ML model, either through post-processing predictions (e.g., Hardt et al., 2016) or by adjusting the model parameters directly (e.g., Kamiran et al., 2010).

We refer to Caton and Haas (2023) for a more elaborate overview of existing approaches. There is currently little guidance on how to select fairness-aware ML techniques, their effectiveness across scenarios and in what cases their use is appropriate.

2.2 Automated Machine Learning

AutoML (Hutter et al., 2019; Escalante, 2021) is a subfield of machine learning that researches and studies methods that automate components of the ML workflow with the goals of speeding up the development of ML applications and reducing the required level of expertise for otherwise manual tasks. By lowering the barrier to apply ML methods, ML can become more accessible and allows to explore new use cases. AutoML comes in many flavours which we briefly describe in the following.

Hyperparameter optimization (HPO) The most low-level incarnation of AutoML is HPO, where the goal is to optimize the hyperparameters of an ML algorithm to minimize a user-specified cost function, such as the misclassification error (Feurer & Hutter, 2019). HPO is often used to tune the hyperparameters of deep neural networks (Bergstra et al., 2011; Snoek et al., 2012, 2015) and popular techniques are grid and random search (Bergstra & Bengio, 2012; Bouthillier & Varoquaux, 2020), Bayesian optimization (Garnett, 2022) and bandit-based methods (Jamieson & Talwalkar, 2016; Li et al., 2018).

The *Combined Algorithm Selection and Hyperparameter optimization problem* (CASH problem) extends the HPO problem to optimize an ML pipeline choosing between different ML methods (Thornton et al., 2013):

$$A^*, \boldsymbol{\lambda}^* \in \underset{A^{(j)} \in \mathcal{A}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}^{(j)}}{\arg \min} \ \frac{1}{k} \sum_{i=1}^{k} \mathcal{L}(A_{\boldsymbol{\lambda}}^{(j)}, \mathcal{D}_{train}^{(i)}, \mathcal{D}_{test}^{(i)}). \tag{1}$$

It is a hierarchical hyperparameter optimization problem with the hyperparameter space $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^{(1)} \cup \cdots \cup \boldsymbol{\Lambda}^{(l)} \cup \Lambda_r$, where $\lambda_r \in \Lambda_r = \{A^{(1)}, \ldots, A^{(l)}\}$ is a new root-level hyperparameter that selects between algorithms $A^{(1)}, \ldots, A^{(l)}$. The *search space* consists of subspaces $\boldsymbol{\Lambda}^{(j)}$ that are conditional on λ_r being instantiated to A_j. This setting is often extended to contain multiple high-level choices $r \in (1, \ldots, R)$ to construct a multi-step pipeline. The CASH problem requires specialized HPO algorithms that can deal with these hierarchical hyperparameters, for example, tree-powered Bayesian optimization methods (Hutter et al., 2011; Bergstra et al., 2011; Rakotoarison et al., 2019), or decompositions into higher and lower-level problems (Liu et al., 2020).

Neural architecture search (NAS) The recent success of deep learning methods has spurred the development of tailored methods that optimize the structure of deep neural networks (Elsken et al., 2019b; White et al., 2023). Closely related to HPO, some NAS problems can be tackled with HPO methods. Two popular approaches to NAS are gradient-based optimization via continuous relaxations of the architectures (Liu et al., 2019) and the global optimization of repeated local structures in a so-called cell search space (Real et al., 2019). The NAS and HPO problems can be combined into the joint architecture and hyperparameter search (JAHS) problem, in which both the topology of the neural network and its hyperparameters such as the learning rate and regularization hyperparameters, are optimized together (Zela et al., 2018; Zimmer et al., 2021; Hirose et al., 2021; Bansal et al., 2022).

AutoML systems AutoML systems typically extend the above-mentioned techniques with a concrete design space to automatically design ML pipelines. In its basic form, an AutoML

system for supervised learning takes as input a *dataset* and an *evaluation metric* and outputs an ML pipeline optimized for the metric. As such, the *scope* of AutoML is mostly limited to the *modeling* stage of a machine learning development workflow (Figure 1).

AutoML systems primarily differ in terms of the considered *search space* and the *search strategy* that is used to explore the search space. Popular examples are AutoWEKA (Thornton et al., 2013) that uses the Bayesian optimization algorithm SMAC (Hutter et al., 2011) to solve the CASH problem given in Equation 1, the extension Auto-sklearn (Feurer et al., 2015) that also incorporates meta-learning and explicit ensembling, and the earlier Particle Swarm Model Selection (Escalante et al., 2009) that uses a particle swarm optimization algorithm. In contrast, the TPOT AutoML system (Olson et al., 2016) uses genetic programming and can construct more general pipelines. These AutoML systems are composed of a general search space description of an ML pipeline and its hyperparameters. They then employ powerful global HPO algorithms (Feurer & Hutter, 2019) to optimize the pipeline for a dataset at hand, often using iterative optimization algorithms that suggest and refine solutions during the optimization process. However, there are also other approaches that do not rely on HPO, such as Auto-Gluon (Erickson et al., 2020) which stacks models to yield a powerful predictive pipeline. We refer to Gijsbers et al. (2022) and Feurer et al. (2022) for recent overviews of AutoML systems. In addition to these, the broader field of *automated Data Science* (AutoDS; De Bie et al. (2022)) focuses on automating additional aspects of the development process, such as exploratory data analysis, data preparation, or parts of model deployment. The focus of the current paper is on AutoML, but many of our findings are applicable to other endeavors related to automation in machine learning, including AutoDS.

2.3 AutoML and Fairness

As the interest in fairness-aware ML systems increases, researchers have started to propose AutoML systems that incorporate fairness considerations. In this section we discuss multi-objective and constrained optimization, then describe how it is used to handle fairness as an additional objective, and lastly, highlight other findings that relate to fairness and AutoML.

Multi-objective and constrained optimization   HPO methods, AutoML systems, and NAS methods have been extended to the *multi-objective optimization* case to take into account additional objectives. A related solution is given by *constrained optimization*, in which one or more objectives are optimized to meet a certain constraint value (Hernández-Lobato et al., 2016).

HPO has been used to optimize ML models to take multiple cost functions (Horn & Bischl, 2016) or resource consumption (Igel, 2005) into account. The goal of multi-objective optimization is to approximate the Pareto set representing the best possible combination of outcomes – solutions that can not be improved w.r.t. one objective without sacrificing performance w.r.t. another. The AutoML system TPOT uses multi-objective optimization to balance pipeline size and performance (Olson et al., 2016). AutoXGBoostMC proposes tuning several additional metrics, such as fairness, robustness, or interpretability metrics (Pfisterer et al., 2019). The NAS technique LEMONADE (Elsken et al., 2019a) searches for neural architectures that balance performance, latency and size. We refer to Morales-Hernández et al. (2022) and Karl et al. (2023) for reviews of multi-objective HPO and to Benmeziane et al. (2021) for a review on hardware-aware NAS, a subfield of multi-objective NAS.

Fairness as a constraint or additional objective Many fairness-aware ML algorithms consider a constrained optimization approach. Existing work on fairness-aware AutoML has extended this approach to AutoML systems. For example, Liu et al. (2020) propose a new CASH algorithm that decomposes the CASH problem into multiple simpler subproblems and allows incorporating black-box constraints along with the optimization objective. Similarly, in fairBO, Perrone et al. (2021) proposed to use standard constrained Bayesian optimization to optimize for predictive performance subject to fairness constraints, and demonstrated how Bayesian optimization can tune hyperparameters constrained for three different metrics at the same time.

Early examples of multi-objective HPO for fairness (Pfisterer et al., 2019; Chakraborty et al., 2019) use the Bayesian optimization algorithm ParEGO (Knowles, 2006) and a custom sequential model-based optimization algorithm (Nair et al., 2020), respectively. Later works propose to use multi-fidelity optimization and extend the popular Successive Halving (Karnin et al., 2013; Jamieson & Talwalkar, 2016) and Hyperband (Li et al., 2018) algorithms to multi-objective algorithms and use them to tune fairness as an auxiliary metric (Schmucker et al., 2020, 2021; Cruz et al., 2021). Most recently, Dooley et al. (2023) used multi-fidelity multi-objective joint neural architecture and hyperparameter optimization in order to find models that perform better with respect to various fairness metrics in face recognition.

Finally, *Fair AutoML* (Wu & Wang, 2021a) extended the FLAML AutoML system (Wang et al., 2021) by adding a resource allocation strategy to dynamically decide whether to evaluate new pipelines to improve performance or to mitigate bias for promising pipelines by using the exponentiated gradient technique (Agarwal et al., 2018).

Gaining knowledge through fairness-aware AutoML Additionally, Perrone et al. (2021) thoroughly study the impact of tuned hyperparameters on a fairness metric and find that larger regularization often leads to fairer results. Moreover, they find that tuning standard ML algorithms leads to comparable or even better performance than tuning fairness-aware ML algorithms, such as support vector machines that incorporate a constraint on a fairness metric during model training (Donini et al., 2018). Similarly, Dooley et al. (2023) demonstrate that by only optimizing the neural architecture and hyperparameters of deep neural networks, it is possible to yield better combinations of fairness metric and predictive performance than by relying on existing bias mitigation approaches, and that fairer neural architectures exhibit a reduced linear separability of protected attributes. Finally, Cruz et al. (2021) find that targeted hyperparameter optimization can improve fairness metrics at a small cost in predictive performance. Moreover, they also find that different ML algorithms occupy distinct regions of the fairness-predictive performance space, suggesting a relationship between specific model classes and fairness.

In conclusion, existing work on fairness-aware AutoML has combined fairness-aware ML techniques proposed in algorithmic fairness research, as well as the choice of neural architectures and hyperparameters, with multi-objective or constrained optimization. This offers several potential advantages over existing techniques, e.g. in terms of flexibility. However, as the research field of algorithmic fairness has evolved to include more (inter)disciplinary perspectives, it has become clear that formulating fairness as an optimization task is not always effective in achieving fairer real-world outcomes. This raises the question: how should

we design fairness-aware AutoML systems to achieve fair outcomes? In the remainder of this work, we will dive deeper into the limitations and opportunities of fairness-aware AutoML from two perspectives: the user's inputs into the system (Section 3) and the AutoML system's design (Section 4).

3. Effect of the User Inputs and User Interactions on Fairness

AutoML - and by extension fairness-aware AutoML - traditionally mostly addresses the *modeling* stage of an ML workflow. However, fairness-related harm can arise at each stage of the ML development process. Therefore, the way in which an AutoML system is integrated into a workflow plays an important part in achieving fair real-world outcomes. In this section, we set out how the user's input (data, metrics, and constraints) affects fairness and what this implies for fairness-aware AutoML.

3.1 Data

Data is one of the main user inputs to an AutoML system. Coincidentally, biases in datasets are typically regarded as one of the primary sources for fairness-related harm, which makes bias identification and mitigation crucial components of a fairness-aware ML workflow (Figure 1). There are many different types of bias that can be present in datasets. We will limit our overview to three biases that are particularly important for notions of group fairness and refer the interested reader to Mitchell et al. (2021) and Mehrabi et al. (2021) for a more exhaustive overview.

Historical Bias *Historical bias* refers to social biases that are encoded in a dataset and can be reproduced by an ML model (Suresh & Guttag, 2021). In particular, a dataset may reflect historical injustices that exist in our societies (Bao et al., 2021). For example, if historically people from lower socioeconomic backgrounds have had fewer opportunities to receive a high-quality education, they may be less suitable for jobs where education is essential, resulting in lower hiring rates for these groups. Unaccounted for, an ML model is likely to reproduce this type of bias and predict fewer positives for historically marginalized groups.

Measurement Bias Datasets embed a plethora of subjective design choices, including what is measured and how. *Measurement bias* occurs when measurements do not accurately or adequately measure the concept we intended to measure. When measurement bias is related to sensitive group membership, it can be a source of unfairness. For example, recidivism risk assessment models are sometimes trained on arrest records, which reflect only a subset of all criminal behavior and might be affected by biased policing practices. As a result, arrest records can be a biased measurement of true criminal activity. Importantly, these issues cannot be observed from the data alone and must be inferred from the context. As a result, meaningfully addressing measurement bias will require careful consideration of the data generation and collection processes (Jacobs & Wallach, 2021).

Representation Bias   A dataset may suffer from *representation bias*,[4] meaning that the dataset contains too few instances of a sensitive group to make accurate predictions. This issue is most prevalent when the data distribution differs substantially between groups. For example, Buolamwini and Gebru (2018) have shown that facial recognition systems failed disproportionately for pictures of darker-skinned women, a group consistently underrepresented in facial recognition datasets. Even if the number of instances is similar, features included in a dataset may be less informative for some groups compared to others (Barocas & Selbst, 2016).

Implications   While it is impossible to identify all types of bias from data alone, future fairness-aware AutoML systems can include several safeguards and diagnostic tools that support practitioners. In particular, systems can flag potential issues, such as disparate base rates (pointing to historical or measurement bias) and small (sub)groups (pointing to representation bias). This can then be used to trigger a user interaction. Furthermore, future work could explore incorporating approaches that facilitate careful modeling of historical bias and measurement bias through causal or statistical modeling, assisted by AutoML methods. Additionally, future systems could incorporate requests for meta-data in the user interface (e.g., in the form of datasheets; Gebru et al., 2021) that help practitioners identify potential issues with their data and facilitate documentation for audits.

3.2 Selecting Fairness Metrics

Another important user input of an AutoML system is the metric(s) for which the ML model is optimized. As seen in Section 2.3, fairness-aware AutoML systems typically use a multi-objective or constrained learning approach to incorporate quantitative fairness metrics in the learning process. An underlying assumption of this approach is that all relevant fairness and performance metrics can be clearly identified and adequately reflect all relevant ethical considerations (ethical review and system (re-)design in Figure 1). This assumption is unlikely to hold in practice for several reasons.

Identifying relevant groups is challenging   Addressing fairness requires identifying socially relevant groups that are at risk of harm. It is often hard to anticipate in advance for which subgroups the system might fail (Chen et al., 2019; Wachter et al., 2021), or to even identify them in advance (Ruggieri et al., 2023), resulting in a reactive strategy rather than an anticipatory approach (Shankar et al., 2022). Moreover, commonly studied sensitive characteristics, such as race and gender, are social constructs that are complex to measure (Hanna et al., 2020; Jacobs & Wallach, 2021); this is further complicated when we consider intersectionality (Chen et al., 2019; Buolamwini & Gebru, 2018) and privacy concerns.

Choosing a fairness metric is challenging   Determining which fairness metric is most appropriate for a given use case is a highly non-trivial question and still an active area of research (Hellman, 2020; Wachter et al., 2021; Hertweck et al., 2021; Hedden, 2021; Weerts

---

4. Despite apparent similarities, representation bias is not the same as selection bias in statistical analysis. Selection bias refers to a failure to achieve proper randomization in selection samples, while representation bias refers to the minimum number of samples that are required to make accurate predictions for a subgroup.

et al., 2022). This is further complicated by the fact that many fairness constraints cannot be satisfied simultaneously (Kleinberg et al., 2017; Chouldechova, 2017) as each fairness metric corresponds to a different set of empirical and normative assumptions (Weerts et al., 2022). For example, arguments for or against demographic parity can depend both on the causes of disparities (e.g., measurement bias or historical bias) and the consequences (e.g., the cost of false positives).

Fairness metrics are simplified notions of fairness Every quantitative fairness metric is necessarily a simplification, lacking aspects of the substantive nature of fairness long debated by philosophers, legal scholars, and sociologists (Selbst et al., 2019; Jacobs & Wallach, 2021; Schwöbel & Remmers, 2022; Watkins et al., 2022). While simplified notions of fairness can potentially help to *assess* potential fairness-related harms, it can be challenging to anticipate all relevant side effects of (technical) interventions that *enforce* them. In particular, group fairness metrics are predominantly parity-based, meaning they enforce equality of statistics, such as error rates, across sensitive groups. Beyond parity, group fairness metrics do not set specific constraints on the distribution of predictions. For example, it is theoretically possible to achieve equal false positive rates by increasing the false positive rate of the better-off group such that it is equal to that of the worst-off group, which in most cases will not make the worst-off group any better off. The under-specification of fairness metrics can be especially problematic when we consider bias-*transforming* metrics such as demographic parity that require deviating from the status quo. Underspecification of the world we would like to see can do more harm than good. Borrowing an example from Dwork et al. (2012), it is possible to increase the selection rate for female applicants by inviting the least qualified female applicants for an interview. While this intervention would satisfy demographic parity, it is unlikely that any of the unqualified applicants would actually make it to the next round. Although extreme "solutions" like this are unlikely to end up on the Pareto front of a fairness-aware AutoML solution, more subtle variations of this issue can be hard to detect. Without careful modeling of measurement bias and historical bias, simply enforcing demographic parity – while a very important measure for equitable outcomes – can have undesirable side effects that can even harm the groups the intervention was designed to protect (Liu et al., 2018; Weerts et al., 2022).

Implications On the one hand, flexibility in defining fairness metrics is crucial to address various fairness concerns and most current AutoML systems are indeed agnostic to the metric(s) to be optimized.[5] On the other hand, practitioners may not always anticipate adverse side effects of such (hand-crafted) fairness metrics (Liu et al., 2018; Weerts et al., 2022) or lack available implementations, e.g. for settings other than binary classification. This is not to say that practitioners should refrain from attempting to quantify fairness at all, which would simply lead to a continuation of current practice. Moreover, we also think that these are not reasons against fairness-aware *Auto*ML, as all of the above problems of choosing an appropriate fairness metric similarly apply to standard ML.

---

5. Currently, only the fairness extension of FLAML (Wu & Wang, 2021a) is limited in this regard due to its use of the exponentiated gradient to obtain fairer models, which is only defined for a limited number of metrics. However, this is to a large extent a limitation of the implementation rather than the methodology.

Instead, practitioners need to appreciate the complexities of quantifying fairness, solicit input from relevant stakeholders, exercise caution in optimizing for fairness metrics, and proactively monitor the system for unforeseen side effects. Indeed, it is our hope that the faster development cycle that fairness-aware *Auto*ML affords will facilitate more iterations on these complex issues. While the use of limited fairness metrics will often be better than ignoring fairness entirely, developers of fairness-aware AutoML systems need to steer clear of "fairness-washing" and instead design their system in a way that encourages users to thoroughly scrutinize their models. Going beyond reporting fairness metric scores, more comprehensive evaluations could include grounding metrics in real-world quantities (Corbett-Davies & Goel, 2018), the use of disaggregated evaluations (Barocas et al., 2021) and integration of interpretability tools (Molnar et al., 2022). Additionally, we observe a tension between guiding practitioners in their choice of fairness metric, whilst avoiding oversimplification of ethical concerns. While AutoML systems are well equipped to optimize for bias-preserving notions of fairness that emphasize equal predictive performance across groups, particular care should be taken with optimization for bias-transforming metrics.

3.3 Use of the System

How a fairness-aware AutoML system and its output are used plays an important part in achieving fair real-world outcomes.

Automation Bias   One possible advantage of fairness-aware *Auto*ML over more traditional fairness-aware ML techniques is that fairness-aware AutoML allows practitioners to spend more time on other tasks, such as careful selection of training data and the right fairness metrics for the application at hand. However, as with any technology, the introduction of fairness-aware AutoML into the toolbox of ML practitioners can have unintended consequences. In particular, there is the risk of *automation bias*: the tendency to over-trust suggestions from automated systems, even in the presence of contradictory information (Skitka et al., 1999). For example, in the context of interpretability, Kaur et al. (2020) find that ML practitioners may take the output of ML interpretability tools at face value, rather than use them as a starting point for further inquiry and understanding of the ML model. In the case of fairness-aware AutoML, automation bias could cause practitioners to more easily adopt an ML model suggested by the system compared to hand-crafted models, without thoroughly scrutinizing the proposed solution. In other words, practitioners may expect that fairness-aware AutoML will not just output the *best* model, but also optimizes for the *right* objective almost automatically. As explained above, it is challenging to identify all relevant side effects of technical interventions in advance, which makes the thorough evaluation of models crucial. This is further complicated by organizational dynamics, which may favor the rapid deployment of models (Madaio et al., 2020). Additionally, there is the risk of over-reliance on the automated system: if a fairness-aware AutoML system is not designed to tackle an issue (e.g., intersections of multiple sensitive features), a user might be tempted to simply ignore the problem, as not doing so means they need to resort to a manual solution.[6] Finally, an iterative model development stage can lead to insights into the dataset and problem at hand that could be lost as more steps are automated.

---

6. Of course, issues related to the lack of availability of particular capabilities are not limited to AutoML systems and also apply to (fairness-aware) ML more generally.

Role of (Fairness-Aware) (Auto)ML Potential fairness-related harm depends not only on the model itself but also on the sociotechnical system the model is a part of (Selbst et al., 2019) and technical interventions during the modeling stage can only hope to address parts of the problem.

Sources of unfair real-world outcomes are not limited to the direct inputs and outputs of a fairness-aware AutoML system, but also include whether and how a practitioner incorporates the system in their workflow. For example, the way in which a problem is formulated during the *problem understanding* stage is implicitly shaped by practitioners' assumptions (Madaio et al., 2020). While this is not problematic by and of itself, practitioners may not be able to identify and assess their own biases (Raji et al., 2020). Even prioritizing some problems over others constitutes a value judgment. A municipality may use ML to detect fraud in welfare benefit applications, but could also use ML to identify citizens who are eligible for welfare but have not yet applied. Which project is pursued affects how stakeholders will be impacted. Value judgments become even more explicit during the design of requirements and key performance indicators, which dictate what the system is optimized for. At the other end of the workflow, during *deployment*, decision-makers may interact with the ML model in a way that introduces fairness-related harm. For example, in pretrial risk assessment, not the predictions of the model but the final decision of the judge determines the real-world consequences. If fairness is a requirement, it is therefore not sufficient to consider only the output of the model, but also how the predictions are used by decision-makers (Selbst et al., 2019). Another example is the risk of reinforcing feedback loops. Most fairness assessments during the *modeling* stage are performed at a single point in time. However, once deployed, decision-making systems can shape which data is collected in the future and thus can reinforce existing or even introduce bias (Ensign et al., 2018; Liu et al., 2018; Cooper & Abrams, 2021; Schwöbel & Remmers, 2022).

The introduction of fairness-aware ML, and by extension fairness-aware AutoML, could incentivize ML practitioners to disregard other, potentially more effective, options in the decision space. For example, when data is taken as a given input, this negates the possibility to collect more or better data. Similarly, inquiry into the data and its provenance might lead to insights that improve the quality of a resulting model, e.g. detecting data leakage or relevant missing features. Additionally, a non-technical intervention might be more appropriate. For example, rather than decreasing the false positive rate in recidivism risk assessments through technical interventions, false positives could be made less costly by changing the real-world consequences from sending a defendant to jail to enrolling them in a community service program (Corbett-Davies et al., 2017).

Implications In order to avoid falling prey to the adverse effects of automation, AutoML system developers could consider approaches introduced in the field of human-computer interaction. For example, to avoid users simply accepting the default solutions, principles of *seamful design* (Kaur et al., 2022) can be used to implement intentional friction that promotes critical reflection. In this way, fairness-aware AutoML users can be encouraged to move from a passive role to the more active, engaged role that is required for fairness work. Additionally, we urge developers of fairness-aware AutoML to present their system as what it is: an important tool amongst many others in the toolbox of a fairness-aware ML practitioner that can help them get the most out of the many possible methods for the modelling stage.

In this spirit, developers should see AutoML as a tool that primarily *supports* practitioners when they engage in data work, rather than a tool that *replaces* them (Wang et al., 2019). Beyond communication, recognizing the narrow scope of current AutoML systems also opens up paths for future work. For example, we envision AutoML systems that provide user prompts or documentation that encourages practitioners to consider alternative interventions. More radically, future work could expand the frame of AutoML to include more nuanced modeling of the sociotechnical system of which the ML model is a part, such as potential feedback loops, user interactions, or causal models.

4. Effect of the AutoML System Design on Fairness

Several design decisions of a fairness-aware AutoML system shape its resulting models and for which tasks it will perform well. To illuminate the implications of those decisions, we discuss choices made for different components of the system including the *search space*, *formulation of the optimization problem* and *evaluation protocol*.

4.1 Search Space

By designing the search space of an AutoML system, we make implicit decisions about what kind of algorithms and what kind of processing steps a system should consider. This allows *codifying established best practices*, e.g. by relying on processing steps and model classes that demonstrably lead to improved results.

This includes aspects such as feature selection and transformation methods, fairness-aware machine learning techniques, ML algorithms, and subsequent stacking or ensembling. To provide an example, the space of possible processing steps could be constrained to pre-processing steps and ML algorithms that yield easily interpretable ML pipelines, lending themselves to a better analysis by domain experts. While desirable, this may exclude pipelines that better reflect the real-world complexities captured in a dataset. Sources of harm in this stage include *aggregation bias* (Suresh & Guttag, 2021), which occurs when a learning algorithm is unable to capture the data distributions of distinct groups. Similarly, without considering fairness, extensive regularization may result in a model that only captures the data distribution of a majority group, but fails for minority groups.

Decisions made at this stage can also shape the capabilities of the resulting system that are not captured by fairness metrics. To provide an example, including fairness-aware machine learning techniques that return randomized predictions (e.g., Hardt et al., 2016; Agarwal et al., 2018) could lead to stochastic predictions in the final model. While this may lead to high performance on fairness metrics in expectation, randomization is an undesirable outcome in many real-world applications (Weerts et al., 2022).

Implications Components included in an AutoML system shape the quality of the resulting models and the system's applicability. Focusing an AutoML system's search space on inherently interpretable models might simplify and improve the reliability of model audits, but this might lead to models that are not sufficiently complex resulting in *aggregation bias*. Similarly, allowing the user to include or exclude different components can lead to desirable or undesirable properties in the resulting models which should be reflected in the system's API design. For example, a fairness-aware AutoML system could provide

users flexibility in deciding whether to use techniques that lead to randomized predictions. An important research direction considers uncovering the advantages, disadvantages, and implicit normative assumptions of particular (fairness-aware) ML algorithms across multiple dimensions, including specific understandings of fairness and interpretability.

4.2 Optimization Problem Formulation

While previous work has used both constrained and multi-objective optimization to take fairness considerations into account, little attention has been given to the advantages and disadvantages that come with the two different paradigms, neither in the context of fairness-aware AutoML nor in the general context of optimization.[7]

Imposing a fairness constraint simplifies subsequent model selection procedures compared to the multi-objective scenario, as it allows for selecting the 'best' model satisfying a previously specified constraint. This strongly simplifies the optimization problem (Perrone et al., 2021) and communication of results. On the other hand, by imposing constraints, we implicitly rephrase the goal of model building to *find the best model that is still ethically/legally allowed*, ignoring the possibility to employ models that strike more favorable trade-offs. Additionally, how such constraints should be set is a non-trivial question – particularly in advance. Previous work has often cited the *four-fifths* rule (e.g., Feldman et al., 2015) as an example of such a constraint, but this rule only applies to a very narrow domain of US labor law and translating such legal requirements into fairness metrics requires multiple abstractions that likely invalidate resulting measurements (Watkins et al., 2022).[8] Similarly, EU anti-discrimination law is designed to be context-sensitive and reliant on interpretation (Wachter et al., 2021), making it challenging to set any hard constraints in advance.

We visualize the multi-objective and constrained optimization perspective w.r.t. minimization in Figure 2. A multi-objective method (see Figure 2a) returns all solutions that are located on the Pareto-front (dashed line), i.e., solutions where no other more favorable trade-off is available. A constrained optimization method (see Figure 2b) instead disregards all solutions that violate a chosen constraint (here unfairness < 0.075) depicted in red and returns the point with the lowest error from all solutions satisfying the constraint.

Implications Employing a multi-objective perspective allows for a better understanding of all available trade-offs, the effect of models and model hyperparameters, and which decisions impact fairness. This is particularly important since the fairness-aware (Auto)ML workflow is iterative in nature, due to the complexities of quantifying fairness and soliciting input from relevant stakeholders discussed in Section 3.2. Knowledge of the entire set of Pareto-optimal models is much more valuable in this iterative process than the single solution that a constrained optimization solution would yield. Furthermore, should constraints change in the future, results from multi-objective optimization might allow for selecting a different Pareto-optimal model while the constrained optimization approach might require full retraining of the entire AutoML system. Further research into combining multi-objective

---

7. We note that it is also possible to combine both paradigms into constrained multi-objective optimization, however, we are not aware of existing work that compares this approach to the constrained optimization and multi-objective optimization paradigms.

8. More generally, we argue that fairness metrics and constraints loosely derived from legal texts should not be equated with legal fairness principles.

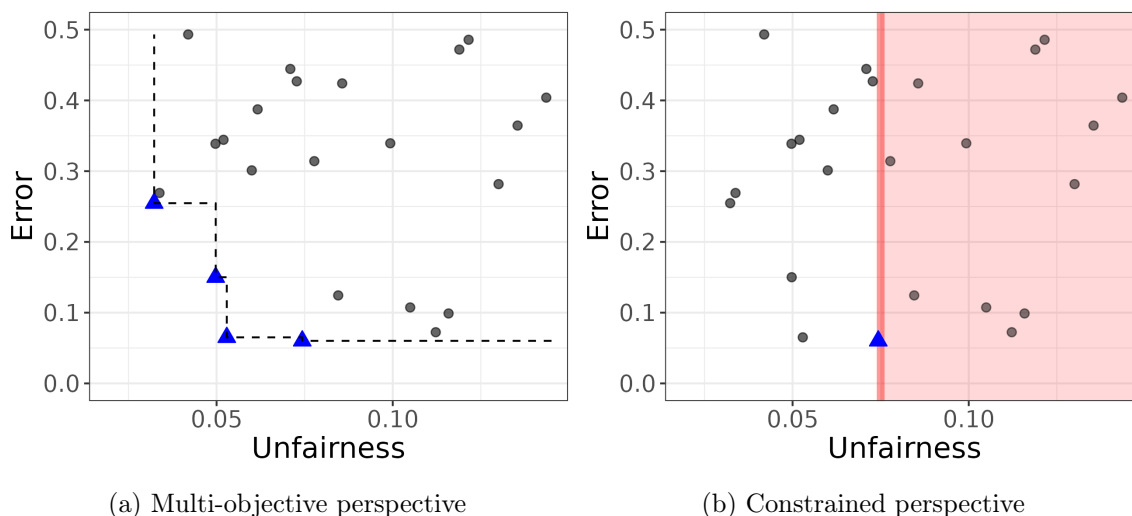(a) Multi-objective perspective
(b) Constrained perspective

Figure 2: Multi-objective perspective (left) and constrained perspective (right) showing exemplary unfairness and error values of found ML models (black dots). We depict optimal solutions returned by the minimization procedure as blue triangles, the approximation of the Pareto-front with a dashed line and the constraint (unfairness > 0.075) as a red area.

and constrained optimization could lead to improved approximations of relevant segments of the Pareto-front without wasting compute budget on ethically questionable solutions that yield high performance.

4.3 Evaluation Protocol

Another set of problems arises from inadequate evaluation protocols. Most technical metrics applied during model selection only yield noisy estimates of the true generalization error due to the limited amount of data available and stochasticity of algorithms, soft- and hardware (Bouthillier et al., 2021). The typical evaluation protocol uses a *train-valid-test* split (Raschka, 2018), where *train* is used for training a model, *valid* is used for model selection, i.e. selecting a final model from the AutoML system, and *test* is held out for a final evaluation, i.e. estimating the performance on unseen data or comparing AutoML systems.[9]

While the development of evaluation protocols for ML is challenging in general, the challenge is exacerbated when we consider group fairness metrics that require the estimation of group statistics (e.g., selection rate). Sensitive groups are often small, resulting in noisy estimates - especially when multiple sensitive characteristics are considered. Additionally, group fairness metrics summarize the comparison of those group statistics across groups (e.g., by taking the maximum difference), resulting in statistically biased upwards metrics that could possibly exaggerate disparities (Lum et al., 2022).

Robust estimation is further complicated as evaluation often concerns multiple metrics (e.g., both a fairness constraint and a predictive performance objective). Estimating the performance of a given fairness-aware AutoML system requires an evaluation protocol that

---

9. We note that instead of a single *train-valid* split, one can also use other protocols, e.g. *cross-validation*.

ensures that the reported quantities (e.g. predictive performance or fairness metrics) are robustly estimated (Agrawal et al., 2020). The typical *train-valid-test* protocol does not transfer to multiple objectives or additional constraints, where the performance of solutions is typically assessed as the hypervolume covered by non-dominated solutions – the Pareto-optimal set (Zitzler et al., 2003). Since metrics are noisy, candidate solutions chosen by the AutoML system on *validation* data might no longer satisfy constraints or be Pareto-optimal on the test set or when deployed in practice (Feurer et al., 2023).

Implications An ideal test set for fairness assessments is (1) free of measurement bias (see Section 3.1) and errors, and (2) contains sufficient data for each subgroup to accurately estimate group statistics. When subpopulations are very small, it may be infeasible to collect sufficient data through random sampling. In such cases, a weighted sampling approach in which small subgroups are oversampled may be more appropriate. However, care should be taken to not overburden already marginalized groups and to ensure that any further interpretation of the results takes note of sampling bias. If additional data collection is infeasible, stratification along the sensitive feature should be used to ensure that the test set approximately preserves the representation of each sensitive group in the train and test set - an approach commonly used for imbalanced classification problems. Future work should focus on developing evaluation protocols that explicitly take into account the uncertainty of estimates of fairness metrics. Furthermore, the assessment of multiple objectives or constraints requires the use of more robust evaluation protocols that are adapted to this setting (e.g., Feurer et al., 2023).

4.4 Benchmark-driven Development

Benchmarks for studying, developing and comparing algorithms have been widely used to track the state-of-the-art in subfields of ML research. Benchmarks define typical tasks that should be solved by ML systems, providing a definition of each task's relevant properties, such as included data or evaluation metric(s). In its simplest form, a benchmark task for AutoML is characterized by a dataset (with a predefined set of features and target variable) and a predictive performance metric (including an evaluation protocol). The design of AutoML systems is often largely benchmark driven – systems are developed to compete on standardized test suites (Gijsbers et al., 2022) or in AutoML competitions (Guyon et al., 2019). This has the benefit that new system components are immediately tested for their empirical performance and only included if they provide substantial benefits. Furthermore, this allows for thoroughly studying algorithms, objective comparisons, and visibility and reproducibility of research progress. Translating benchmark results into real-world improvements requires the availability of and access to benchmarks that adequately reflect the contextual complexity and technical challenges of relevant real-world tasks (Raji et al., 2021). We identify several reasons why this assumption may not be applicable in the context of fairness research.

Existing benchmarks are decontextualized Several benchmarks have been critiqued for a lack of similarity to real-world applications (Langley, 1996; Saitta & Neri, 1998; Wagstaff, 2012). The construction of fairness benchmarks often seems to be guided by availability, rather than by a careful abstraction of a real-world problem. Especially in the context of fairness, these characteristics lead to issues around the validity of the collected data (Ding

et al., 2021; Bao et al., 2021; Grömping, 2019) as well as a disconnect from real-world applications (Bao et al., 2021). As a result, progress on fairness benchmarks is unlikely to reflect progress on the real-world outcomes that motivate fairness research. For example, Bao et al. (2021) describe how the use of ProPublica's COMPAS dataset (Angwin et al., 2016) as a benchmark is misleading, as performing well on this task cannot be tied to real-world impact in the field of criminal justice. Using benchmarks to solely chase state-of-the-art performance, researchers risk losing sight of the context of the data - which is crucial to advance social outcomes such as fairness.

Small set of benchmarks   Many subfields in ML research focus on empirical performance improvements on a very small set of benchmark tasks, raising concerns regarding overfitting (Recht et al., 2019) and a lack of contribution to scientific understanding (Hooker, 1995; Dehghani et al., 2021). Fairness-aware ML methods are no exception and have typically been evaluated only across a small set of benchmark tasks (cf. Table 1 in the Supplementary material), strongly limiting the conclusions that can be drawn from such experiments on a meta-level.

Existing benchmark tasks are limited to the modeling stage   The high emphasis on mitigation algorithms in algorithmic fairness research has resulted in a set of benchmark datasets (cf. Table 1 in the Supplementary material) that only reflect a very small aspect of typical fairness-aware ML workflows since AutoML typically focuses on the modeling stage (cf. Figure 1). This can result in a blind spot for specific capabilities, such as evaluating long-term impact or handling intersectional sensitive groups, that are not tested in a typical fairness benchmark, which can trickle down to the users of the AutoML system.

Implications   If fairness-aware AutoML system development is guided by benchmarks, the datasets making up such benchmarks should be *contextualized*, *reflect real-world challenges*, and *sufficiently large*. We urge AutoML and fairness researchers to jointly develop contextualized benchmark tasks that reflect the challenges that practitioners face in practice when working on fairness-related tasks. Defining such benchmark tasks requires a realistic application scenario along with problem constraints as well as relevant fairness and performance metric(s) tailored to real-world outcomes. In many domains, data collection will require a great deal of effort that may take years, including (but not limited to) collaborating with domain experts, intense stakeholder management, and resolving incomplete and complex database systems. Another important reason for the lack of datasets is concerns around data privacy, since fairness-related ML tasks typically comprise sensitive data. Synthetically generated data has been successfully used in other fields and could also provide a useful source here. This would enable researchers to thoroughly study the performance of algorithms w.r.t. to specific problem characteristics. However, even "realistic" synthetic datasets may not accurately reflect the real-world context compared to their "real-world" counterparts and care should be taken in the interpretation of the results. Additionally, data privacy issues could be tackled on an organizational level, but given the diversity of data protection regulations across jurisdictions, specific recommendations are beyond the scope of this work. If constructed well, we believe that benchmarks could inspire the development of fairness-aware AutoML systems with more diverse capabilities and highlight complexities and problems that occur in practice.

While quantitative benchmarks have widely been used to demonstrate a method's superiority, we argue that especially in the context of fairness-aware AutoML, we need to move beyond numerical performance. In practice, users might often prefer simpler solutions over complex methods as they simplify debugging and retraining. Moreover, performance gains from more complex models often do not translate to benefits during deployment (Shankar et al., 2022). In particular, we believe future work should shift in focus from "horse race" analyses to a better understanding of when and why a particular system works and – perhaps more importantly – when it does not. Rather than pursuing incremental improvements in numerical performance, we believe that evaluation approaches designed to diagnose potential failure modes (Raji et al., 2021) and controlled experiments motivated by specific hypotheses (Hooker, 1995) will be crucial for advancing fairness research.

Until better benchmark tasks become available, we would like to highlight that we cannot endorse using existing benchmark tasks to claim progress in fairness research. If researchers use existing benchmark tasks to assess the strengths and weaknesses of multi-objective or constrained optimization algorithms, we recommend (1) a clear statement that multi-objective / constrained optimization work does not constitute fairness research. We also recommend (2) emphasizing work on more sensible benchmarks and (3) relying on very targeted problem scenarios within a clearly defined application context, clearly stating involved biases and resulting harms.

5. Opportunities for Fairness-aware AutoML

Provided that the use of (fairness-aware) ML is justifiable, ML pipelines should be constructed in the best possible manner. Existing work points towards several opportunities for fairness-aware AutoML to contribute to this goal.

Fairness-aware AutoML allows integrating best practices in the modeling stage. Practitioners often lack knowledge on how to integrate fairness toolkits into their workflow (Holstein et al., 2019; Deng et al., 2022) and may face difficulties in proper evaluation of fairness metrics (Agrawal et al., 2020). AutoML systems can implement relevant search spaces, mitigation techniques, and evaluation protocols into a configurable pipeline that is optimized automatically, relieving the user from staying up-to-date with the technical literature. Moreover, AutoML systems can codify best practices that ensure that model selection is done correctly, preventing mistakes arising from, for example, wrongly coded evaluation protocols or undetected train-test leakage.

Fairness-aware AutoML may outperform traditional fairness-aware ML techniques. Previous work has shown that (1) even an arguably simple constrained Bayesian optimization of a standard ML algorithm is en-par with solutions discovered by fairness-aware ML algorithms (Perrone et al., 2021), with the advantage of being much more flexible, and that (2) the joint multi-objective optimization of deep neural architectures and their hyperparameters can also outperform other bias mitigation techniques in the context of face recognition (Dooley et al., 2023). Importantly, standard fairness-aware ML algorithms, such as a fairness-aware support vector machine (Donini et al., 2018) require hyperparameter tuning, too, which is an additional argument for employing fairness-aware AutoML techniques. We are not aware of a thorough comparison of manual tuning with a constrained or

multi-objective hyperparameter optimization algorithm, but by extrapolating results from the single-objective setting (Bergstra & Bengio, 2012; Snoek et al., 2012; Chen et al., 2018; Henderson et al., 2018; Zhang et al., 2021) into a constrained or multi-objective setting, we believe that fairness-aware AutoML is perfectly posed to support practitioners to reliably and efficiently tune hyperparameters.

Fairness-aware AutoML may fit more easily into existing workflows.   By building upon standard HPO rather than hand-crafted ML algorithms that were altered to take fairness into account, AutoML solutions are typically easier to fit into existing workflows (Cruz et al., 2021) and can be used together with different off-the-shelf machine learning models. For example, Perrone et al. (2021) demonstrate that their method works for random forests (Breimann, 2001), XGBoost (Chen & Guestrin, 2016), neural networks, and linear models. Similarly, Cruz et al. (2021) have jointly optimized random forests, decision trees, LightGBM (Ke et al., 2017), logistic regression, neural networks, and decision trees wrapped with the exponentiated gradient method (Agarwal et al., 2018) – also showing that each of the models leads to a different trade-off between the performance and the fairness metric. In addition (and in contrast to most fairness-aware ML strategies), fairness-aware AutoML systems are typically agnostic to the fairness metric at hand (Perrone et al., 2021; Chakraborty et al., 2019; Cruz et al., 2021; Wu & Wang, 2021a) and can be used for arbitrary models while not relying on a specific model class being adapted (Perrone et al., 2021). As such, fairness-aware AutoML has the potential to address the portability trap described by Selbst et al. (2019): while individual models and even individual bias mitigation techniques do not generalize from one usage context to another, the AutoML process of *searching* for the Pareto front of optimal tradeoffs of user-defined objectives in a particular application generalizes across usage contexts. Moreover, algorithmic solutions are sometimes deployed without considering fairness issues at all, due to a lack of awareness, knowledge, or expertise (Lum & Isaac, 2016; Buolamwini & Gebru, 2018). While technical interventions should not be regarded as the sole tool for addressing unfairness, enriching AutoML can lower the barrier for practitioners and domain experts to incorporate fairness considerations. Relieving the user from some of the technical complexities of ML pipeline design can allow them to spend more time on aspects where a human in the loop is essential.

Fairness-aware AutoML can be a useful tool to improve understanding of metrics, data, & models   Machine learning can be a useful tool not just during model development, but also during data exploration. In particular, ML can be used to investigate relationships in the data that lead to potential fairness-related harm (Wachter et al., 2020). AutoML can be a useful tool to quickly build a variety of ML models with high predictive performance. By evaluating models w.r.t. a wide variety of (fairness) metrics, practitioners can gain insight into potential biases contained in the data and labels. While current AutoML systems typically do not facilitate this type of analysis, we envision future systems that allow a practitioner to investigate which features a model typically relies on or whether an (unconstrained) ML pipeline leads to performance disparities. Such knowledge can then be used to, e.g. inform policy decisions that address fairness-related harms at their core. By using a multi-objective AutoML system for exploration, a user can also learn about the trade-offs between different objectives. In contrast to comparing solutions obtained by

manual inspection, a multi-objective AutoML system directly explores the Pareto front of optimal trade-offs.

Fairness-aware AutoML solutions lend themselves to further inspection of the learning process. Such inspections have traditionally been used to better understand the relationship between the characteristics of an ML task, the parameters of learners, and predictive performance (Jones et al., 1998; Hutter et al., 2013, 2014; Fawcett & Hoos, 2016; Golovin et al., 2017; Biedenkapp et al., 2017, 2018; van Rijn & Hutter, 2018; Probst et al., 2019; Weerts et al., 2020; Moosbauer et al., 2021). Similar analyses can be performed in fairness-aware ML. For example, Perrone et al. (2021) used hyperparameter importance analysis to get further insights into the learners, finding that regularization parameters are particularly important to meet fairness constraints. These insights are readily available as output of an AutoML system, having generated meta-data during the interaction of pipeline optimization and the data at hand.

6. Conclusions

Fairness-aware AutoML systems have the potential to overcome several challenges ML practitioners face in fairness work. By simplifying and accelerating the integration of fairness considerations into ML solutions, these systems can allow practitioners to allocate more time to tasks where a human-in-the loop is essential. Moreover, fairness-aware AutoML systems can be a useful tool to rapidly improve understanding of the input data and machine learning process. However, the effectiveness of fairness-aware AutoML systems in mitigating real-world harm depends on various factors, including not only the design of the system, but also the way in which practitioners utilize it. Furthermore, while issues related to fairness can emerge in every stage of the ML workflow, AutoML is currently primarily involved with the modeling stage, limiting its role in fairness-aware ML workflows. This raises the question: how can we develop fairness-aware AutoML systems in a way that acknowledges the complexity of algorithmic fairness throughout the entire workflow? In this concluding section, we propose several guidelines for research and development of fairness-aware AutoML and set out directions for future work.

Guidelines for Fairness-aware AutoML   It is important to be aware of the consequences of making technological solutions available. To assist researchers and developers interested in fairness-aware AutoML, we have formulated the following guidelines.

- **Clearly state assumptions and limitations.** When talking about fairness-aware AutoML and solutions generated by such systems, prominently state assumptions and limitations. It must be clear for what use cases the system is suitable, how it was tested, and in which scenarios it should not be used. In particular, clearly state that satisfying fairness metrics cannot guarantee a fair system (Corbett-Davies & Goel, 2018) and may even lead to adverse effects (Cooper & Abrams, 2021).

- **Support users in identifying sources of fairness-related harm.** Sources of fairness-related harm can often be mitigated more effectively through interventions beyond the modeling stage of an ML workflow, such as better data collection, an

alternative problem formulation, and non-technical interventions related to the use of the system. As such, it is important to help users identify potential issues. One direction for this could be informing users of potential problems via fairness checklists (Madaio et al., 2020) or inviting explicit reflection on data and model lineage, e.g., by pre-populating or simplifying the creation of *Data sheets* (Gebru et al., 2021) and *Model cards* (Mitchell et al., 2019).

- **Incorporate principles of seamful design.** Automation involves the risk that users overly trust solutions without employing the required amount of scrutiny of data and resulting models. This might result in sources of bias not accounted for during model development, evaluation, and deployment and, thus, could ultimately lead to adverse effects for sensitive groups. Systems should therefore take great care w.r.t. how solutions are presented and which conclusions can be drawn from reported metrics. To counteract automation bias, we encourage system developers to consider existing practices in human-computer interaction, such as adding intentional friction through seamful design (Kaur et al., 2022), which encourages users to reflect on explicit and implicit design choices.

- **Support users in statistically sound fairness evaluation.** Assessing the fairness of a model requires that quantities are estimated based on a sufficiently large and representative sample of the population under consideration while ensuring that the data can support such conclusions. If not caught, errors in the evaluation could lead to blind spots for subgroups not sufficiently represented in the data or premature conclusions that are unsupported by the available data. AutoML can and should codify best practices to prevent user errors during evaluation. In particular, we recommend warning users if the results of a fairness evaluation are not supported by sufficient data.

- **Account for inherent limitations of fairness metrics.** It is challenging to capture nuanced understandings of fairness in quantitative fairness metrics. As a result, optimization for a particular fairness constraint or objective can have undesirable side effects. To avoid these issues, fairness-aware AutoML should support users in comprehensive evaluations beyond simple metrics. Additionally, we discourage solutions that enforce bias-transforming metrics such as demographic parity without careful modeling of the social context and data collection processes that motivate these metrics.

- **Ensure well-substantiated system design.** The search space is an important design choice of a fairness-aware AutoML system. It defines the applicability and quality of models an AutoML system outputs, as well as the trade-offs between desiderata that will be explored during learning. We recommend carefully documenting the design decisions made during the development process and explaining how they incorporate user requirements.

- **Evaluate the system against contextualized benchmarks.** If new fairness-aware AutoML systems are designed, it is important that those systems not only solve an oversimplified problem but actually assist the user in achieving fairer outcomes. Existing benchmarks are often oversimplified and do not reflect the real-world requirements of a fairness-aware AutoML system. Hence, novel contextualized benchmarks should

be created to analyze to what extent fairness-aware AutoML systems meet these requirements.

- **Support users in performing quick iterations.** The fairness-aware (Auto)ML workflow is necessarily iterative. We thus encourage the development of fast and interactive fairness-aware AutoML systems that allow for rapid iterations.

Opportunities Having discussed perspective of users and developers of fairness-aware AutoML systems in Sections 3 and 4, as well as the opportunities for fairness-aware AutoML systems in Section 5, we now briefly summarize the potential benefit of fairness-aware AutoML for each stakeholder to open up a dialogue for progressing towards fairness-aware AutoML:

- **AutoML researcher.** AutoML systems can have a positive real-world impact if they support users in building fairer models. Directly applicable research directions are improving evaluation and constrained multi-objective optimization. Moreover, we suggest extending the limited knowledge on how users interact with AutoML systems (Xin et al., 2021) and developing interactive systems that support them best in doing so (Wang et al., 2019; Crisan & Fiore-Gartland, 2021).

- **Fairness researcher.** Fairness-aware AutoML stands and falls with the quality of available fairness metrics (Ruggieri et al., 2023) and available testbeds to develop fairness-aware AutoML systems. Therefore, we expect further research in contextual benchmark problems in combination with less abstract and more tangible fairness criteria will be a fruitful direction for future research.

- **ML practitioner.** We urge users of ML to familiarize themselves with the dangers of applying ML in sensitive applications, for example by following the references provided in this paper. If designing an ML intervention is the best solution to a problem, we propose relying on principled techniques for building ML models and suggest fairness-aware AutoML techniques.

Future Work We posit that fairness-aware AutoML systems can play an important role in typical fairness-aware ML workflows by tackling model selection, hyperparameter optimization, and model evaluation and identify several promising directions for future work towards this end. An important research topic is the design of evaluation protocols that can properly handle multiple objectives or constraints, as well as noisy metrics in a statistically sound way. Furthermore, while most current work on fairness-aware ML approaches the problem from a constrained optimization perspective, much less work has explored the promising multi-objective optimization scenario. Moreover, incorporating aspects of the *evaluation* stage into the AutoML loop, by e.g., modeling latent preferences expressed during evaluation or making it easier to incorporate other real-world objectives into the optimization loop, is a promising avenue for further research. We also believe that AutoML and hyperparameter analysis can be a useful tool for fairness researchers as well as AutoML researchers to improve understanding of the (fairness-aware) learning process and inform guidance on the suitability of particular approaches. Additionally, we hope to encourage the joint development of more realistic benchmarks. Furthermore, the present paper focuses primarily on settings with

structured data, and we are aware of only one publication in the realm of fairness-aware AutoML that tackles image data (Dooley et al., 2023). While many of the discussed challenges are also relevant to settings with unstructured data, we expect several unique challenges related to bias and unfairness of models trained on unstructured data such as images and text, for example, missing meta-data on group membership. We believe that exploring these challenges and opportunities is an interesting direction for future work. Similarly, generative AI shares many of the challenges we outlined, but will require additional considerations that warrant further research. Finally, we want to argue in favor of AutoML systems that are more interactive in nature: instead of monolithic one-solution-fits-all systems, future work should go into *assistive* systems, that, e.g., point out possible problems in the data, or guide the user through fairness audits (Landers & Behrend, 2022; Madaio et al., 2020). We believe that an interdisciplinary approach bridging AutoML and human-computer interaction research will be crucial for designing effective user interaction interfaces and counteracting potential automation bias.

Revisiting the original question that motivated this work, we conclude that *fairness cannot be automated*. Instead, an iterative, context-sensitive process guided by human domain experts is essential to arrive at fair outcomes. Nevertheless, in cases where technical interventions are appropriate, *fairness-aware AutoML systems can lower the barrier to incorporating fairness considerations* in the ML workflow and support users through the *integration of best practices and state-of-the-art approaches*, without the need to follow the latest technical literature. However, we emphasize that basic data science skills remain crucial for correctly framing the problem and continuously monitoring performance. In summary, even with a humble attitude towards the role of AutoML, we believe fairness-aware AutoML to be an important research direction that is likely to have a substantive impact.

Acknowledgement

Funded by the European Union

Appendix A. Datasets Used in Fairness-aware AutoML Research

We collect datasets used for empirical comparisons in the fairness-aware AutoML works we cite in order to highlight the limited scope and quantity of datasets used during benchmarks. fairness-aware AutoML systems are typically evaluated on only 1 to 4 datasets, typically employing either a fairness metric and a performance metric (A-E) or a combined metric that evaluates performance given a fairness constraint is satisfied (F-I). While additional datasets have been collected and investigated in the context of fairness benchmarking (Quy et al., 2022; Fabris et al., 2022), they have not been used to evaluate fairness-aware AutoML systems. In addition, benchmarks based on such datasets might suffer from the same problems, such as limited scope and lack of representativity discussed in Section 4.4.

Table 1: Fairness-related datasets used in prior research on fairness-aware AutoML.

Paper	Adult	German Credit	Compas	Donors Choice	AOF (private)	Default Risk (Kaggle)	Bank	MEPS
A Pfisterer et al. (2019)	✓							
B Schmucker et al. (2020)		✓	✓					
C Schmucker et al. (2021)	✓							
D Chakraborty et al. (2019)	✓	✓	✓					
E Cruz et al. (2021)	✓		✓	✓	✓			
F Liu et al. (2020)						✓		
G Perrone et al. (2021)	✓	✓	✓					
H Wu and Wang (2021a)	✓		✓				✓	
I Wu and Wang (2021b)	✓		✓				✓	✓

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. In Dy, J., & Krause, A. (Eds.), *Proceedings of the 35th International Conference on Machine Learning (ICML'18)*, Vol. 80, pp. 60–69. Proceedings of Machine Learning Research.

Agrawal, A., Pfisterer, F., Bischl, B., Chen, J., Sood, S., Shah, S., Buet-Golfouse, F., Mateen, B., & Vollmer, S. (2020). Debiasing classifiers: is reality at variance with expectation?. *arXiv, 2011.02407 [cs.LG]*.

Angwin, J., Larson, J., Mattu, S., & Kichner, L. (2016). Machine Bias.. ProPublica.

Bansal, A., Stoll, D., Janowski, M., Zela, A., & Hutter, F. (2022). JAHS-Bench-201: A Foundation For Research On Joint Architecture And Hyperparameter Search. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., Lum, K., & Venkatasubramanian, S. (2021). It's COMPASlicated: The Messy Relationship between

RAI Datasets and Algorithmic Fairness Benchmarks. In Vanschoren, J., & Yeung, S. (Eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Curran Associates.

Barocas, S., Guo, A., Kamar, E., Krones, J., Morris, M., Wortman Vaughan, J., Wadsworth, W., & Wallach, H. (2021). Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. In Fourcade, M., Kuipers, B., Lazar, S., & Mulligan, D. (Eds.), *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES'21)*.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. fairml-book.org. `http://www.fairmlbook.org`.

Barocas, S., & Selbst, A. (2016). Big data's disparate impact. *California Law Review, 104*, 671.

Bellamy, R., Dey, K., Hind, M., Hoffman, S., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development, 63*(4/5), 4:1–15.

Benmeziane, H., El Maghraoui, K., Ouarnoughi, H., Niar, S., Wistuba, M., & Wang, N. (2021). A Comprehensive Survey on Hardware-Aware Neural Architecture Search. *arxiv, 2101.09336 [cs.LG]*.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., & Weinberger, K. (Eds.), *Proceedings of the 24th International Conference on Advances in Neural Information Processing Systems (NeurIPS'11)*, pp. 2546–2554. Curran Associates.

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research, 13*, 281–305.

Biedenkapp, A., Lindauer, M., Eggensperger, K., Fawcett, C., Hoos, H., & Hutter, F. (2017). Efficient Parameter Importance Analysis via Ablation with Surrogates. In S.Singh, & Markovitch, S. (Eds.), *Proceedings of the Thirty-First Conference on Artificial Intelligence (AAAI'17)*, pp. 773–779. AAAI Press.

Biedenkapp, A., Marben, J., Lindauer, M., & Hutter, F. (2018). CAVE: Configuration Assessment, Visualization and Evaluation. In Battiti, R., Brunato, M., Kotsireas, I., & Pardalos, P. (Eds.), *Proceedings of the International Conference on Learning and Intelligent Optimization (LION)*, Lecture Notes in Computer Science. Springer.

Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Mohammadi Sepahvand, N., Raff, E., Madan, K., Voleti, V., Ebrahimi Kahou, S., Michalski, V., Arbel, T., Pal, C., Varoquaux, G., & Vincent, P. (2021). Accounting for Variance in Machine Learning Benchmarks. In Smola, A., Dimakis, A., & Stoica, I. (Eds.), *Proceedings of Machine Learning and Systems 3*, Vol. 3, pp. 747–769.

Bouthillier, X., & Varoquaux, G. (2020). Survey of machine-learning experimental methods at NeurIPS2019 and ICLR2020. Research report hal-02447823, Inria Saclay Ile de France.

Breimann, L. (2001). Random Forests. *Machine Learning Journal, 45*, 5–32.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Barocas, S. (Ed.), *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT'18)*, pp. 77–91. ACM Press.

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building Classifiers with Independency Constraints. In Saygin, Y., Yu, J., Kargupta, H., Wang, W., Ranka, S., Yu, P., & Wu, X. (Eds.), *ICDM Workshops 2009, IEEE International Conference on Data Mining Workshops, Miami, Florida, USA, 6 December 2009*, pp. 13–18. IEEE Computer Society.

Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery, 21*(2), 277–292.

Caton, S., & Haas, C. (2023). Fairness in Machine Learning: A Survey. *ACM Comput. Surv.*, .

Chakraborty, J., Xia, T., Fahid, F., & Menzies, T. (2019). Software Engineering for Fairness: A Case Study with Hyperparameter Optimization. In *Proceedings of the 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE.

Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. In d. boyd, & Morgenstern, J. (Eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT'19)*, pp. 339–348. ACM Press.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Krishnapuram, B., Shah, M., Smola, A., Aggarwal, C., Shen, D., & Rastogi, R. (Eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, pp. 785–794. ACM Press.

Chen, Y., Huang, A., Wang, Z., Antonoglou, I., Schrittwieser, J., Silver, D., & de Freitas, N. (2018). Bayesian Optimization in AlphaGo. *arxiv, 1812.06855 [cs.LG]*.

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data, 5*(2), 153–163.

Cooper, A., & Abrams, E. (2021). Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research. In Fourcade, M., Kuipers, B., Lazar, S., & Mulligan, D. (Eds.), *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES'21)*, pp. 46–54.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv, 1808.00023 [cs.CY]*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic Decision Making and the Cost of Fairness. In Matwin, S., Yu, S., & Farooq, F. (Eds.), *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*, pp. 797—-806. ACM Press.

Crisan, A., & Fiore-Gartland, B. (2021). Fits and starts: Enterprise use of AutoML and the role of humans in the loop. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15.

Cruz, A., Saleiro, P., Belem, C., Soares, C., & Bizarro, P. (2021). Promoting Fairness through Hyperparameter Optimization. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'21)*, pp. 1036–1041. IEEE.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G`.

De Bie, T., De Raedt, L., Hernández-Orallo, J., Hoos, H., Smyth, P., & Williams, C. (2022). Automating Data Science. *Communications of the ACM*, *65*(3), 76–87.

Dehghani, M., Tay, Y., Gritsenko, A., Zhao, Z., Houlsby, N., Diaz, F., Metzler, D., & Vinyals, O. (2021). The Benchmark Lottery. *arXiv, 2107.07002 [cs.LG]*.

Deng, W., Nagireddy, M., Lee, M., Singh, J., Wu, Z., Holstein, K., & Zhu, H. (2022). Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In Isbell, C., Lazar, S., Oh, A., & Xiang, A. (Eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT'22)*, pp. 473—-484. ACM Press.

Ding, F., Hardt, M., Miller, J., & Schmidt, L. (2021). Retiring Adult: New Datasets for Fair Machine Learning. In Ranzato, M., Beygelzimer, A., Nguyen, K., Liang, P., Vaughan, J., & Dauphin, Y. (Eds.), *Proceedings of the 34th International Conference on Advances in Neural Information Processing Systems (NeurIPS'21)*. Curran Associates.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., & Pontil, M. (2018). Empirical Risk Minimization Under Fairness Constraints. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates.

Dooley, S., Sukthanker, R., Dickerson, J., White, C., Hutter, F., & Goldblum, M. (2023). Rethinking Bias Mitigation: Fairer Architectures Make for Fairer Face Recognition. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226.

Elsken, T., Metzen, J., & Hutter, F. (2019a). Efficient Multi-Objective Neural Architecture Search via Lamarckian Evolution. In *Proceedings of the International Conference on Learning Representations (ICLR'19)*. Published online: `iclr.cc`.

Elsken, T., Metzen, J., & Hutter, F. (2019b). Neural Architecture Search. In Hutter et al. (2019), chap. 3, pp. 63–77. Available for free at http://automl.org/book.

Ensign, D., Friedler, S., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. In Barocas, S. (Ed.), *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT'18)*, pp. 160–171. ACM Press.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., & Smola, A. (2020). AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. *arxiv, 2003.06505 [stat.ML]*.

Escalante, H. (2021). Automated Machine Learning–A Brief Review at the End of the Early Years. In Pillay, N., & Qu, R. (Eds.), *Automated Design of Machine Learning and Search Algorithms*, pp. 11–28. Springer.

Escalante, H., Montes, M., & Sucar, E. (2009). Particle Swarm Model Selection. *Journal of Machine Learning Research*, *10*, 405–440.

Fabris, A., Messina, S., Silvello, G., & Susto, G. (2022). Algorithmic Fairness Datasets: The Story so Far. *Data Mining and Knowledge Discovery*, *36*(6), 2074–2152.

Fawcett, C., & Hoos, H. (2016). Analysing differences between algorithm configurations through ablation. *Journal of Heuristics*, *22*(4), 431–458.

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. In Cao, L., & Zhang, C. (Eds.), *Proceedings of the 21nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*, pp. 259–268. ACM Press.

Feurer, M., Eggensperger, K., Bergman, E., Pfisterer, F., Bischl, B., & Hutter, F. (2023). Mind the Gap: Measuring Generalization Performance Across Multiple Objectives. In Crémilleux, B., , Hess, S., & Nijssen, S. (Eds.), *Advances in Intelligent Data Analysis XXI*, pp. 130–142 Cham. Springer Nature Switzerland.

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., & Hutter, F. (2022). Auto-Sklearn 2.0: Hands-free AutoML via Meta-Learning. *Journal of Machine Learning Research*, *23*(261), 1 – 61.

Feurer, M., & Hutter, F. (2019). Hyperparameter Optimization. In Hutter et al. (2019), chap. 1, pp. 3 – 38. Available for free at http://automl.org/book.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and Robust Automated Machine Learning. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., & Garnett, R. (Eds.), *Proceedings of the 28th International Conference on Advances in Neural Information Processing Systems (NeurIPS'15)*, pp. 2962–2970. Curran Associates.

Garnett, R. (2022). *Bayesian Optimization*. Cambridge University Press. in preparation.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92.

Gijsbers, P., Bueno, M., Coors, S., LeDell, E., Poirier, S., Thomas, J., Bischl, B., & Vanschoren, J. (2022). AMLB: an AutoML Benchmark. *arXiv, 2207.12560 [cs.LG]*.

Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J., & Sculley, D. (2017). Google Vizier: A Service for Black-Box Optimization. In Matwin, S., Yu, S., & Farooq, F. (Eds.), *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*, pp. 1487–1495. ACM Press.

Grömping, U. (2019). South German credit data: Correcting a widely used data set. *TechReport*, *4*, 2019.

Guyon, I., Sun-Hosoya, L., Boullé, M., Escalante, H., Escalera, S., Liu, Z., Jajetic, D., Ray, B., Saeed, M., Sebag, M., Statnikov, A., Tu, W., & Viegas, E. (2019). Analysis of the AutoML Challenge Series 2015-2018. In Hutter et al. (2019), chap. 10, pp. 177–219. Available for free at http://automl.org/book.

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a Critical Race Methodology in Algorithmic Fairness. In Castillo, C., & Hildebrandt, M. (Eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT'20)*, pp. 501—512. ACM Press.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., von Luxburg, U., Guyon, I., & Garnett, R. (Eds.), *Proceedings of the 29th International Conference on Advances in Neural Information Processing Systems (NeurIPS'16)*, pp. 3323–3331. Curran Associates.

Hedden, B. (2021). On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs*, *49*(2).

Hellman, D. (2020). Measuring algorithmic fairness. *Virginia Law Review*, *106*(4), 811–866.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. In McIlraith, S., & Weinberger, K. (Eds.), *Proceedings of the Thirty-Second Conference on Artificial Intelligence (AAAI'18)*. AAAI Press.

Hernández-Lobato, J., Gelbart, M., Adams, R., Hoffman, M., & Ghahramani, Z. (2016). A general framework for constrained Bayesian optimization using information-based search. *Journal of Machine Learning Research*, *17*(1), 5549–5601.

Hertweck, C., Heitz, C., & Loi, M. (2021). On the Moral Justification of Statistical Parity. In Elish, M., Isaac, W., & Zemel, R. (Eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT'21)*, pp. 747—-757. ACM Press.

Hirose, Y., Yoshinari, N., & Shirakawa, S. (2021). NAS-HPO-Bench-II: A Benchmark Dataset on Joint Optimization of Convolutional Neural Network Architecture and Training Hyperparameters. In Balasubramanian, V., & Tsang, I. (Eds.), *Proceedings of The 13th Asian Conference on Machine Learning*, Vol. 157 of *Proceedings of Machine Learning Research*, pp. 1349–1364. PMLR.

Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*, p. 1–16. ACM Press.

Hooker, J. (1995). Testing heuristics: We have it all wrong. *Journal of Heuristics*, *1*(1), 33–42.

Horn, D., & Bischl, B. (2016). Multi-objective parameter configuration of machine learning algorithms using model-based optimization. In Likas, A. (Ed.), *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–8. IEEE.

Hutter, F., Hoos, H., & Leyton-Brown, K. (2011). Sequential Model-Based Optimization for General Algorithm Configuration. In Coello, C. (Ed.), *Proceedings of the Fifth International Conference on Learning and Intelligent Optimization (LION'11)*, Vol. 6683 of *Lecture Notes in Computer Science*, pp. 507–523. Springer.

Hutter, F., Hoos, H., & Leyton-Brown, K. (2014). An Efficient Approach for Assessing Hyperparameter Importance. In Xing, E., & Jebara, T. (Eds.), *Proceedings of the 31th International Conference on Machine Learning, (ICML'14)*, pp. 754–762. Omnipress.

Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Identifying Key Algorithm Parameters and Instance Features using Forward Selection. In Nicosia, G., & Pardalos, P. (Eds.), *Proceedings of the 7th International Conference on Learning and Optimization (LION-7)*, Lecture Notes in Computer Science. Springer Berlin Heidelberg.

Hutter, F., Kotthoff, L., & Vanschoren, J. (Eds.). (2019). *Automated Machine Learning: Methods, Systems, Challenges*. Springer. Available for free at http://automl.org/book.

Igel, C. (2005). Multi-objective Model Selection for Support Vector Machines. In Coello, C., Aguirre, A., & Zitzler, E. (Eds.), *Evolutionary Multi-Criterion Optimization*, pp. 534–546. Springer.

Jacobs, A., & Wallach, H. (2021). Measurement and fairness. In Elish, M., Isaac, W., & Zemel, R. (Eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT'21)*, pp. 375–385. ACM Press.

Jamieson, K., & Talwalkar, A. (2016). Non-stochastic Best Arm Identification and Hyperparameter Optimization. In Gretton, A., & Robert, C. (Eds.), *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS'16)*, Vol. 51. Proceedings of Machine Learning Research.

Jones, D., Schonlau, M., & Welch, W. (1998). Efficient Global Optimization of Expensive Black Box Functions. *Journal of Global Optimization*, *13*, 455–492.

Kamiran, F., & Calders, T. (2011). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems, 33*(1), 1–33.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination Aware Decision Tree Learning. In Webb, G., Liu, B., Zhang, C., Gunopulos, D., & Wu, X. (Eds.), *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*, pp. 869–874. IEEE Computer Society.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2013). Techniques for Discrimination-Free Predictive Models. In Custers, B., Calders, T., Schermer, B., & Zarsky, T. (Eds.), *Discrimination and Privacy in the Information Society - Data Mining and Profiling in Large Databases*, Vol. 3 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*, pp. 223–239. Springer.

Karl, F., Pielok, T., Moosbauer, J., Pfisterer, F., Coors, S., Binder, M., Schneider, L., Thomas, J., Richter, J., Lang, M., Garrido-Merchán, E., Branke, J., & Bischl, B. (2023). Multi-Objective Hyperparameter Optimization in Machine Learning – An Overview. *ACM Trans. Evol. Learn. Optim.*, .

Karnin, Z., Koren, T., & Somekh, O. (2013). Almost Optimal Exploration in Multi-Armed Bandits. In Dasgupta, S., & McAllester, D. (Eds.), *Proceedings of the 30th International Conference on Machine Learning (ICML'13)*, pp. 1238–1246. Omnipress.

Kaur, H., Adar, E., Gilbert, E., & Lampe, C. (2022). Sensible AI: Re-Imagining Interpretability and Explainability Using Sensemaking Theory. In Isbell, C., Lazar, S., Oh, A., & Xiang, A. (Eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT'22)*, p. 702–714. ACM Press.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Vaughan, J. (2020). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In Bernhaupt, R., Mueller, F., Verweij, D., & Andres, J. (Eds.), *Proceedings of the Conference on Human Factors in Computing Systems (CHI'20)*, pp. 1–14. ACM Press.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R. (Eds.), *Proceedings of the 30th International Conference on Advances in Neural Information Processing Systems (NeurIPS'17)*. Curran Associates.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. In Papadimitriou, C. (Ed.), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, Vol. 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 43:1–43:23 Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Knowles, J. (2006). ParEGO: a hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation, 10*(1), 50–66.

Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R. (Eds.), *Proceedings of the 30th International Conference on Advances in Neural Information Processing Systems (NeurIPS'17)*. Curran Associates.

Landers, R., & Behrend, T. (2022). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist, 1*.

Langley, P. (1996). Relevance and Insight in Experimental Studies. *IEEE Expert Online, 11*, 11–12.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2018). Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research, 18*(185), 1–52.

Liu, H., Simonyan, K., & Yang, Y. (2019). DARTS: Differentiable Architecture Search. In *Proceedings of the International Conference on Learning Representations (ICLR'19)*. Published online: `iclr.cc`.

Liu, L., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed Impact of Fair Machine Learning. In Dy, J., & Krause, A. (Eds.), *Proceedings of the 35th International Conference on Machine Learning (ICML'18)*, Vol. 80, pp. 3150–3158. Proceedings of Machine Learning Research.

Liu, S., Ram, P., Vijaykeerthy, D., Bouneffouf, D., Bramble, G., Samulowitz, H., Wang, D., Conn, A., & Gray, A. (2020). An ADMM Based Framework for AutoML Pipeline Configuration. In Rossi, F., Conitzer, V., & Sha, F. (Eds.), *Proceedings of the Thirty-Fourth Conference on Artificial Intelligence (AAAI'20)*, pp. 4892–4899. Association for the Advancement of Artificial Intelligence, AAAI Press.

Lum, K., & Isaac, W. (2016). To predict and serve?. *Significance, 13*(5), 14–19.

Lum, K., Zhang, Y., & Bower, A. (2022). De-Biasing "Bias" Measurement. In Isbell, C., Lazar, S., Oh, A., & Xiang, A. (Eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT'22)*, p. 379–389. ACM Press.

Madaio, M., Egede, L., Subramonyam, H., Wortman Vaughan, J., & Wallach, H. (2022). Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW1), 1–26.

Madaio, M., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In Bernhaupt, R., Mueller, F., Verweij, D., & Andres, J. (Eds.), *Proceedings of the Conference on Human Factors in Computing Systems (CHI'20)*, pp. 1–14. ACM Press.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys, 54*(6).

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I., & Gebru, T. (2019). Model cards for model reporting. In d. boyd, & Morgenstern, J. (Eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT'19)*, pp. 220–229. ACM Press.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*, 141–163.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C., Casalicchio, G., Grosse-Wentrup, M., & Bischl, B. (2022). General pitfalls of model-agnostic interpretation methods for machine learning models. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pp. 39–68. Springer.

Moosbauer, J., Herbinger, J., Casalicchio, G., Lindauer, M., & Bischl, B. (2021). Explaining Hyperparameter Optimization via Partial Dependence Plots. In Ranzato, M., Beygelzimer, A., Nguyen, K., Liang, P., Vaughan, J., & Dauphin, Y. (Eds.), *Proceedings of the 34th International Conference on Advances in Neural Information Processing Systems (NeurIPS'21)*, pp. 2280–2291. Curran Associates.

Morales-Hernández, A., Nieuwenhuyse, I. V., & Gonzalez, S. (2022). A survey on multi-objective hyperparameter optimization algorithms for Machine Learning. *Artificial Intelligence Review*, .

Nair, V., Yu, Z., Menzies, T., Siegmund, N., & Apel, S. (2020). Finding Faster Configurations Using FLASH. *IEEE Transactions on Software Engineering*, *46*(7), 794–811.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*, 447–453.

Olson, R., Bartley, N., Urbanowicz, R., & Moore, J. (2016). Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. In Friedrich, T. (Ed.), *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'16)*, pp. 485–492. ACM Press.

Perrone, V., Donini, M., Zafar, M., Schmucker, R., Kenthapadi, K., & Archambeau, C. (2021). Fair Bayesian Optimization. In Fourcade, M., Kuipers, B., Lazar, S., & Mulligan, D. (Eds.), *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES'21)*, pp. 854—-863.

Pfisterer, F., Coors, S., Thomas, J., & Bischl, B. (2019). Multi-Objective Automatic Machine Learning with AutoxgboostMC. *arXiv*, *1908.10796 [stat.ML]*.

Probst, P., Boulesteix, A., & Bischl, B. (2019). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *Journal of Machine Learning Research*, *20*(53), 1–32.

Quy, T., Roy, A., Iosifidis, V., Zhang, W., & Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery*, *12*(3), e1452.

Raji, I., Denton, E., Bender, E., Hanna, A., & Paullada, A. (2021). AI and the Everything in the Whole Wide World Benchmark. In Vanschoren, J., & Yeung, S. (Eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Curran Associates.

Raji, I., Smart, A., White, R., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In Castillo, C., & Hildebrandt, M. (Eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT'20)*. ACM Press.

Rakotoarison, H., Schoenauer, M., & Sebag, M. (2019). Automated Machine Learning with Monte-Carlo Tree Search. In Kraus, S. (Ed.), *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19)*, pp. 3296–3303.

Raschka, S. (2018). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *arxiv, 1811.12808 [stat.ML]*.

Real, E., Aggarwal, A., Huang, Y., & Le, Q. (2019). Aging Evolution for Image Classifier Architecture Search. In Hentenryck, P. V., & Zhou, Z. (Eds.), *Proceedings of the Thirty-Third Conference on Artificial Intelligence (AAAI'19)*. AAAI Press.

Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do ImageNet Classifiers Generalize to ImageNet?. In Chaudhuri, K., & Salakhutdinov, R. (Eds.), *Proceedings of the 36th International Conference on Machine Learning (ICML'19)*, Vol. 97, pp. 5389–5400. Proceedings of Machine Learning Research.

Ruggieri, S., Alvarez, J., Pugnana, A., State, L., & Turini, F. (2023). Can We Trust Fair-AI?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, pp. 15421–15430.

Saitta, L., & Neri, F. (1998). Learning in the "Real World". *Machine Learning*, *30*(2–3), 133–163.

Schmucker, R., Donini, M., Perrone, V., Zafar, M., & Archambeau, C. (2020). Multi-Objective Multi-Fidelity Hyperparameter Optimization with Application to Fairness. In Calandra, R., Clune, J., Grant, E., Schwarz, J., Vanschoren, J., Visin, F., & Wang, J. (Eds.), *NeurIPS 2020 Workshop on Meta-Learning*.

Schmucker, R., Donini, M., Zafar, M., Salinas, D., & Archambeau, C. (2021). Multi-objective Asynchronous Successive Halving. *arXiv, 2106.12639 [stat.ML]*.

Schwöbel, P., & Remmers, P. (2022). The Long Arc of Fairness: Formalisations and Ethical Discourse. In Isbell, C., Lazar, S., Oh, A., & Xiang, A. (Eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT'22)*. ACM Press.

Selbst, A., boyd, d., Friedler, S., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In d. boyd, & Morgenstern, J. (Eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT'19)*, pp. 59–68. ACM Press.

Shankar, S., Garcia, R., Hellerstein, J., & Parameswaran, A. (2022). Operationalizing Machine Learning: An Interview Study. *arXiv, 2209.09125 [cs.SE]*.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing, 5*(4), 13–22.

Skitka, L., Mosier, K., & Burdick, M. (1999). Does automation bias decision-making?. *International Journal of Human-Computer Studies, 51*(5), 991–1006.

Snoek, J., Larochelle, H., & Adams, R. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. In Bartlett, P., Pereira, F., Burges, C., Bottou, L., & Weinberger, K. (Eds.), *Proceedings of the 25th International Conference on Advances in Neural Information Processing Systems (NeurIPS'12)*, pp. 2960–2968. Curran Associates.

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M., Prabhat, & Adams, R. (2015). Scalable Bayesian Optimization Using Deep Neural Networks. In Bach, F., & Blei, D. (Eds.), *Proceedings of the 32nd International Conference on Machine Learning (ICML'15)*, Vol. 37, pp. 2171–2180. Omnipress.

Suresh, H., & Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21)*. ACM Press.

Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pp. 53–59.

Thornton, C., Hutter, F., Hoos, H., & Leyton-Brown, K. (2013). Auto-WEKA: combined selection and Hyperparameter Optimization of classification algorithms. In Dhillon, I., Koren, Y., Ghani, R., Senator, T., Bradley, P., Parekh, R., He, J., Grossman, R., & Uthurusamy, R. (Eds.), *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13)*, pp. 847–855. ACM Press.

van Rijn, J., & Hutter, F. (2018). Hyperparameter Importance Across Datasets. In Guo, Y., & Farooq, F. (Eds.), *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'18)*, pp. 2367–2376. ACM Press.

Wachter, S., Mittelstadt, B., & Russell, C. (2020). Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. *West Virginia Law Review, 123*.

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review, 41*, 105567.

Wagstaff, K. (2012). Machine Learning that Matters. *arXiv, 1206.4656 [cs.LG]*.

Wang, C., Wu, Q., Weimer, M., & Zhu, E. (2021). FLAML: A Fast and Lightweight AutoML Library. In Smola, A., Dimakis, A., & Stoica, I. (Eds.), *Proceedings of Machine Learning and Systems 3*, Vol. 3, pp. 434–447.

Wang, D., Weisz, J., Muller, M., Ram, P., Geyer, W., Dugan, C., Tausczik, Y., Samulowitz, H., & Gray, A. (2019). Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proc. ACM Hum.-Comput. Interact.*, *3*.

Watkins, E., McKenna, M., & Chen, J. (2022). The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness. *arXiv, 2202.09519 [cs.CY]*.

Weerts, H., Müller, A., & Vanschoren, J. (2020). Importance of tuning hyperparameters of machine learning algorithms. *arXiv, 2007.07588 [stat.ML]*.

Weerts, H., Royakkers, L., & Pechenizkiy, M. (2022). Does the End Justify the Means? On the Moral Justification of Fairness-Aware Machine Learning. *arXiv, 2202.08536 [cs.LG]*.

Weerts, H., Dudík, M., Edgar, R., Jalali, A., Lutz, R., & Madaio, M. (2023). Fairlearn: Assessing and Improving Fairness of AI Systems. *Journal of Machine Learning Research*, *24*(257), 1–8.

White, C., Safari, M., Sukthanker, R., Ru, B., Elsken, T., Zela, A., Dey, D., & Hutter, F. (2023). Neural Architecture Search: Insights from 1000 Papers. *arXiv, 2301.08727.pdf [cs.LG]*.

Wu, Q., & Wang, C. (2021a). Fair AutoML. *arXiv, 2111.06495v1 [cs.LG]*.

Wu, Q., & Wang, C. (2021b). FairAutoML: Embracing Unfairness Mitigation in AutoML. *arXiv, 2111.06495v2 [cs.LG]*.

Xin, D., Wu, E., Lee, D., Salehi, N., & Parameswaran, A. (2021). Whither AutoML? Understanding the Role of Automation in Machine Learning Workflows. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21.

Zafar, M., Valera, I., Rogriguez, M., & Gummadi, K. (2017). Fairness Constraints: Mechanisms for Fair Classification. In Singh, A., & Zhu, J. (Eds.), *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS'17)*, Vol. 54, pp. 962–970. Proceedings of Machine Learning Research.

Zela, A., Klein, A., Falkner, S., & Hutter, F. (2018). Towards Automated Deep Learning: Efficient Joint Neural Architecture and Hyperparameter Search. In Garnett, R., Hutter, F., Vanschoren, J., Brazdil, P., Caruana, R., Giraud-Carrier, C., Guyon, I., & Kégl, B. (Eds.), *ICML workshop on Automated Machine Learning (AutoML workshop 2018)*.

Zhang, B., Rajan, R., Pineda, L., Lambert, N., Biedenkapp, A., Chua, K., Hutter, F., & Calandra, R. (2021). On the Importance of Hyperparameter Optimization for Model-based Reinforcement Learning. In Banerjee, A., & Fukumizu, K. (Eds.), *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS'21)*, pp. 4015–4023. Proceedings of Machine Learning Research.

Zimmer, L., Lindauer, M., & Hutter, F. (2021). Auto-PyTorch Tabular: Multi-Fidelity MetaLearning for Efficient and Robust AutoDL. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*, 3079 – 3090.

Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C., & Fonseca, V. (2003). Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on Evolutionary Computation, 7*, 117–132.