Toward Operationalizing Pipeline-aware ML Fairness: A Research Agenda for Developing Practical Guidelines and Tools

ANONYMOUS AUTHOR(S)

While algorithmic fairness continues to be a thriving area of research in ML, in practice, mitigating issues of bias and unfairness often gets reduced to enforcing an arbitrarily chosen fairness metric, either by enforcing fairness constraints during the optimization step, post-processing model outputs or by manipulating the training data. Recent work has called on the ML community to take a more holistic approach to tackle fairness issues by systematically investigating the multitude of design choices made through the ML pipeline and identifying effective interventions at the root cause, as opposed to the symptoms. While we share the conviction that a pipeline-based approach is the most appropriate for combating algorithmic unfairness on the ground, we believe there are currently very few methods of *operationalizing* this approach in practice. Drawing on our experience as educators and practitioners, we first demonstrate that without clear guidelines and toolkits, even individuals with specialized ML knowledge find it challenging to hypothesize how various design choices influence model behavior. We then consult the fair-ML literature to understand the progress to date toward operationalizing the pipeline-aware approach: we systematically collect and organize the prior work that attempts to detect, measure, and mitigate various sources of unfairness through the ML pipeline. We utilize this extensive categorization of previous contributions to sketch a research agenda for the community. We hope this work serves as the stepping stone toward a more comprehensive set of resources for ML researchers, practitioners, and students interested in exploring, designing, and testing pipeline-oriented approaches and guidelines to algorithmic fairness.

ACM Reference Format:
Anonymous Author(s). 2022. Toward Operationalizing Pipeline-aware ML Fairness: A Research Agenda for Developing Practical Guidelines and Tools. In *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization 2023.* ACM, New York, NY, USA, 36 pages. https://doi.org/10.1145/3531146.3533204

1 INTRODUCTION

As AI systems proliferate in socially high-stakes domains, concerns around balancing their potential for benefiting society versus harming already marginalized and underserved communities have been mounting [43, 86, 100, 225]. Through years of work by a variety of stakeholder groups, including activists, journalists, lawmakers, researchers, and AI practitioners, we now witness a growing appetite for building these systems with harm prevention in mind [146, 253]. Aside from the rising public pressure on technology companies and governments to create more equitable systems, recent legal developments, such as those from the Federal Trade Commission (FTC) [73], the White House [133, 253] and various other agencies [5] suggest an uptick in enforcement surrounding discriminatory algorithms. These recent movements give a point of leverage for technologically equipped activists and plaintiffs to challenge discriminatory algorithms and further incentivize companies to thoroughly explore the space of possible algorithms to find the least discriminatory one within those with sufficient business utility. As a result, professionals across industry, government, and nonprofits have been turning to algorithmic fairness expertise to guide their implementation of AI systems [132].

Manuscript submitted to ACM

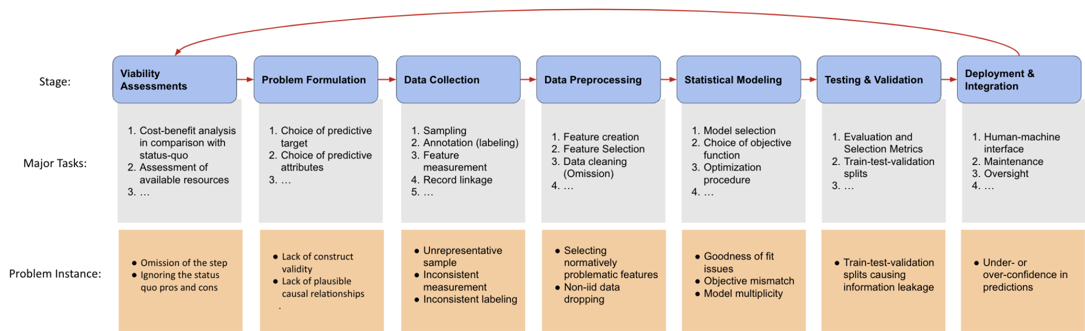

Fig. 1. A simplified view of the ML pipeline, its key stages, and instances of design choices made at each stage.

The pipeline-aware approach to algorithmic fairness. While there has been extensive work in the service of preventing algorithmic unfairness [3, 122, 151, 241, 242], mitigation in practice often gets reduced to enforcing a somewhat arbitrarily formulated fairness metric on top of a pre-developed or deployed system [196]. Practitioners often make ad-hoc mitigation choices to improve fairness metrics, such as removing sensitive attributes, changing the data distribution, enforcing fairness constraints, or post-processing model predictions. While these techniques may improve selected fairness metrics, they often have little practical impact; at worst, they can even exacerbate the same disparity metrics they aim to alleviate [181, 300]. Prior scholarship has attributed these problematic trends to the fact that the traditional approach to fairness fails to take a system-wide view of the problem. They myopically narrow mitigation to a restricted set of points along the ML pipeline (e.g., the choice of optimization function). This is despite the well-established fact that *numerous* choices there can impact the model's behavior [280]. Assessing viability/functionality of AI [246], problem formulation [236], data collection, data pre-processing, statistical modeling, testing and validation, and organizational integration are all key stages of the ML pipeline consisting of consequential design choices (see Figure 1 for an overview). By abstracting away the ML pipeline and selecting an ad-hoc mitigation strategy, the mainstream approach misses the opportunity to identify, isolate, and mitigate the underlying *sources* of unfairness, which can in turn lead to fairness-accuracy tradeoffs due to intervening at the wrong place [37], or "[hide] the real problem while creating an illusion of fairness" [8]. We join prior calls advocating for an alternate *pipeline-aware* approach to fairness [8, 280]. At a high level, this approach works as follows: given a model with undesirable fairness behavior, the ML team must search for ways in which the variety of choices made across the ML creation pipeline may have contributed to the behavior (e.g., the choice of prediction target [225]). Once plausible causes are identified, the team should evaluate whether other choices could abate the problem (e.g. changing a model's prediction target and re-training [37]).[1] This process should take place iteratively, and the model should be re-evaluated until it is deemed satisfactory, whereupon bias testing and model updates would continue throughout deployment.

The need for operationalizing the pipeline-aware approach While prior work has clearly established the benefits of the pipeline-aware view toward fairness, we contend that *conceptual awareness* of this approach alone won't be sufficient for *operationalizing* it in practice. In Section 2, we provide evidence suggesting that making informed hypotheses about the root causes of unfairness in the ML pipeline is a challenging task, even for individuals with specialized ML

[1]This description is taking an auditing perspective: when building a fair model from scratch, fairness desiderata would be described, and the practitioners would enumerate choices that can be made at each step of the ML creation pipeline and avoid choices that work against this desired behavior.

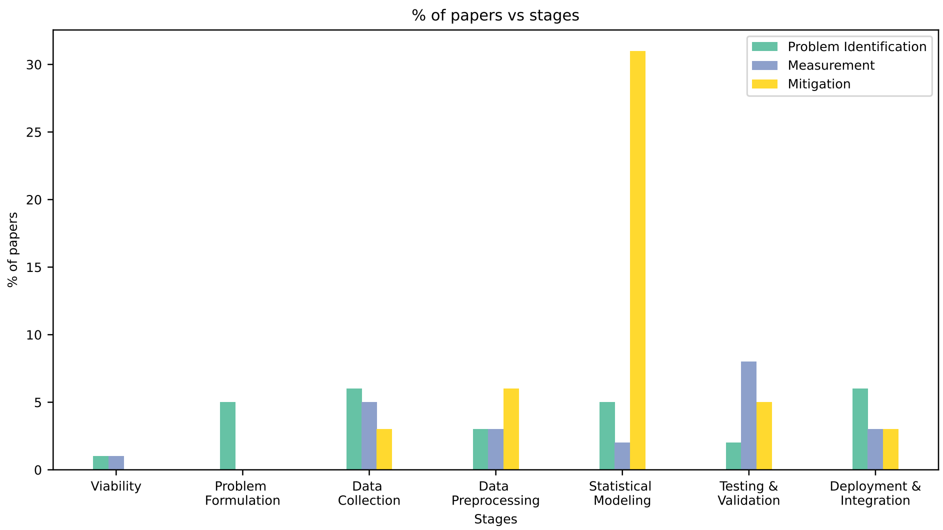

Fig. 2. The distribution of research efforts dedicated to different stages of the ML pipeline among the papers we surveyed.

knowledge and skills. For example, based on qualitative data gathered from a graduate-level fair-ML class at an R1 institution, students with significant ML background struggle to conceptualize sources of harmful model behavior and suggest appropriately chosen bias mitigation strategies. This observation is, indeed, consistent with our experience working with ML practitioners and system developers across a wide range of public policy settings. This evidence motivates our focus on "operationalization": to *find* less discriminatory models, we argue that practitioners need usable tools to measure the discrimination from, identify the underlying sources of, unfairness in the pipeline, and then match those underlying sources with appropriately designed interventions. Such tools would be instrumental to searches for "Less Discriminatory Alternative" models (LDAs)—one of the first publicized instances of which was recently conducted to replace a discriminatory credit allocation system at the private lending platform, Upstart [2]. A practical toolkit can significantly expand the space of methods currently in use [70], and offer additional points of interventions earlier in the pipeline, which can in turn ease legal concern over the use of protected attributes during training and deployment [69, 70].

A snapshot of progress toward operationalization. Having established the need for practical implementation of the pipeline-aware approach to fairness, we seek to understand the progress made so far, and ask: how far along are we toward operationalizing the pipeline-aware approach? Do we currently have useful methods and guidelines to inspect and modify the variety of design choices made throughout the ML pipeline in practice? To respond to these questions, we consult the literature on algorithmic fairness in search of methods that identify, measure, or mitigate biases arising from specific ML design choices. Through an extensive survey of the Fair-ML literature, we collect and synthesize existing pipeline-aware fairness work and map them to specific stages of the pipeline (Section 3). While we identify numerous gaps in the existing arsenal of tools, we hope the resource we have curated offers practitioners a one-stop shop for identifying potential causes of unfairness in their use cases and getting an overview of the state of the art to detect, measure, and mitigate those issues. In the near future, our goal is to turn this catalog into a comprehensive, interactive, and community-maintained set of resources and tools for Fair-ML researchers, practitioners, and students interested in exploring, designing, and testing system-level approaches to algorithmic fairness.

Sketching a path forward for the research community. Figure 2 depicts the relative effort the research community has allocated to different stages of the ML pipeline. As evident in this picture, the focus of the ML community has been largely on the statistical modeling stage, with an out-sized emphasis on mitigation strategies. This finding generalizes recent observations that "Everyone wants to do the model work, not the data work" [260]. Key stages of the pipeline, including viability assessment, problem formulation, and deployment and monitoring, have been understudied—we hypothesize due to lack of potential for novel, quantitative contributions or theoretical analysis [30]. Moreover, we observe that the majority of algorithmic mitigation techniques are left at the proof-of-concept level and are not backed by any application-grounded evaluation of usability and utility. While the HCI community has made significant progress toward documenting the challenges and needs of ML practitioners, very few contributions investigate to use of specific measures and methods to meet these needs. To amplify their impact, we call on the ML and HCI research communities to invest in collaborations aimed at identifying, assessing, and improving concrete tools and guidelines for fairness. We contend that the ML pipeline is a complex system; understanding the interactions between various design choices in this system is a challenging task yet a potentially impactful avenue for quantitative analysis. Lastly, we acknowledge that practitioners often face choices between imperfect alternatives, each leading to its unique problematic consequences—developing normative guidelines to weigh these imperfect choices against each other is a crucial avenue for collaboration between ML experts and ethicists.

2 THE NEED TO OPERATIONALIZE THE PIPELINE-AWARE APPROACH

We draw on our experience working with Machine Learning students and practitioners, and recent examples of enforcing anti-discrimination law to highlight the difficulty in, and necessity for, *operationalizing* a pipeline-aware approach to fairness. We observe that a conceptual understanding of the pipeline and the variety of choices within it are by no means sufficient to inform good practice. Even well-informed students struggle to correctly hypothesize about the underlying causes of unfairness and suggest plausible remedies. Moreover, existing mitigation techniques run the risk of violating existing legal restrictions on the use of protected attributes, so it is essential to expand the set of levers auditors have available to search for less discriminatory models.

2.1 Evidence from Regulatory Enforcement

Recent developments in attempts to regulate the design and use of ML systems have given a sense of urgency to support regulators and policymakers in these efforts [5, 72, 133, 253]. In this section, we focus on one of the first examples of enforcement of anti-discrimination law in AI systems [69] to show how a pipeline-aware approach may be particularly helpful in establishing legal liability in, and providing remedies for, discrimination in regulated AI systems. In particular, we note that when searching for less discriminatory alternatives to deployed models, a pipeline-aware approach may elide some of the potential legal problems that apply to using more traditional algorithmic fairness approaches [27, 131]. Many traditional algorithmic fairness techniques that intervene at the modeling stage often use protected attributes to change model behavior by learning new prediction thresholds or modifying the training procedure [3, 122], leading regulators to dismiss such approaches due to legal restrictions around the use of protected attributes in decision-making [27, 131]. However, pipeline-aware techniques may use protected attribute labels to *evaluate* different models and choose among them, but may not directly enforce a constraint using protected attributes, avoiding the same legal scrutiny applied to other algorithmic fairness techniques. Thus, it is imperative that we build pipeline-aware tools to empower regulators, policymakers, and advocacy groups pushing for legal requirements surrounding bias reduction in public-facing AI systems to find, and enforce the use of, less discriminatory systems.

Case Study Background: Upstart Monitorship. In 2020, the consumer finance firm Upstart, the NAACP, and Relman Colfax, a civil rights law firm, entered a legal agreement to investigate racial discrimination in Upstart's lending model due to the NAACP's concerns over the use of attributes related to educational attainment, such as the name of the college applicants attended if they had a college degree [2]. The goal of the agreement was to first to determine if there was a legally relevant difference in selection rate between Black and white applicants using Upstart's model, which they found there was [70]. Once this was established, Relman Colfax followed the disparate impact doctrine [147] to determine whether Upstart would be legally required to change their model: to do so, they search for a "less discriminatory alternative" model, or LDA, with equivalent performance but less disparate impact across racial groups. Under the disparate impact law, once discrimination is established, if such a model exists, then the company using the discriminatory model *must* replace their model with this less discriminatory alternative [147].

Searching for a Less Discriminatory Alternatives: Suboptimal Approaches. After establishing discriminatory model behavior, third-party investigators developed and implemented a strategy to find a less discriminatory model with similar predictive performance to Upstart's original algorithm. Notably, the third-party bias investigators discounted all algorithmic fairness techniques to search for discrimination out of hand, seemingly from concern over legal repercussions over the use of the protected attribute to influence model behavior, and a desire to "align with traditional principles gleaned from antidiscrimination jurisprudence" [69]. As the authors of the bias investigation report note:

> A range of techniques for mitigating disparities is proposed in the algorithmic fairness literature. *Some of these proposals could raise independent fair lending risks, such as the use of different models for different protected classes or the improper use of prohibited bases as predictive variables*...While [the algorithmic fairness] conversation is valuable, many "fairness" proposals do not engage or align with the established three-step disparate impact analysis reflected in case law and regulatory materials. [69]

The procedure that was taken instead was intervening at the feature selection step, and searching the feature space for a model with a subset of features that was less discriminatory. That is, the practitioners created a model for several different feature combinations, and tested the disparate impact of each one [70]. While a less discriminatory alternative was likely discovered,[2], this procedure took sufficiently long that Upstart updated its model before the investigations were completed [71]. We note that the blind, exhaustive search for a less discriminatory model at just one intervention point in the machine learning pipeline is almost certainly expensive, inefficient, and leads to suboptimal outcomes. If pipeline-aware tools had been available to better isolate sources of bias in the *entire* pipeline, beyond feature choice, it is likely they would have found a less discriminatory model with acceptable predictive performance more efficiently.

It is imperative that the approach taken in the Upstart monitorship does not become standard or accepted in the credit industry. The disparate impact framework gives advocates and model practitioners a way to challenge the use of discriminatory algorithms, and further incentivize companies to thoroughly explore the space of possible algorithms to find the least discriminatory one within those with sufficient business utility—but using ineffective methods to enforce these regulatory tools weaken their power. We suggest that in order to effectively leverage the disparate impact doctrine, we must operationalize a pipeline-aware approach to ML fairness. And, even beyond searching for LDA models, tools

[2]The model discovered was less discriminatory but also suffered performance drop, which Relman Colfax argued was within an acceptable range to be an "equivalent performance", but the exact rules around this have yet to be established—so it is unclear if this model would in fact suffice as an LDA [71].

informed by a pipeline-based approach may also aid in creating more standard and rigorous approaches to algorithmic audits more broadly.

2.2 Evidence from the Classroom

Our team has documented the challenges of instilling the pipeline-centric view of ML harms in students using traditional teaching methods (e.g., lectures containing real-world examples). The classroom activity outlined in this section was conducted by one of us at an R1 educational institution as part of a graduate-level course focused on the ethical and societal considerations around the use of ML in socially high-stakes domains. Our IRB approved the activity, and students had the option of opting out of data collection for research purposes.[3]

Population and sampling. In total, 37 students participated in the class activity. Prior completion of at least one Machine Learning course was a prerequisite for enrollment. So all participants had a non-negligible background in Machine Learning. Specifically, $\sim$ 80% characterized the familiarity with ML as intermediate, $\sim$ 14% as advances, and $\sim$ 5% as elementary. All students enrolled in the course took it as an elective. This fact implies that compared to a random sample of students with ML knowledge, our participants were likely more aware of ML harms and more motivated to address them.

Study design. Students were introduced to the ML pipeline through an approximately 45-minutes long lecture. The lecture focused on the supervised learning paradigm and broke down the supervised learning pipeline into the following five stages: problem formulation, data collection and processing, model specification, model fitting/training, and deployment in the real world. For each stage, students were presented with multiple examples of how choices at that stage can lead to harmful outcomes, such as unfairness, at the end of the pipeline (see Figure 1). They were then asked to team up with 3-4 classmates and pick a societal domain as the focus of their group activity. Examples offered to them included employment (e.g., hiring employees); education (e.g., admitting students); medicine (e.g., diagnosing skin cancer); housing (e.g., allocating limited housing units child welfare (e.g., investigating referral calls); criminal justice (e.g., pretrial sentencing); public safety (e.g., allocating patrol resources in policing e-commerce (e.g., advertising; ranking sellers on Amazon); social media (e.g., news recommendation; content moderation); and transportation (e.g., autonomous vehicles). One team suggested finance (e.g., credit lending) as their topic. Third, students were given 30 minutes to discuss the following questions about their application domain with teammates and submit their written responses individually:

- Characterizing the specific **predictive task** their team focused on.
- The **type of harm** observed.
- Their **hypotheses around the sources** of this harm in through the ML pipeline.
- Their**hypotheses around potentially effective remedies** for addressing those sources.

Findings. A thematic analysis of submitted responses revealed several challenges:

Theme 1: Specifying how a real-world problem gets translated into a predictive task was not straightforward. Table 1 overviews how each team defined their predictive tasks. For example, the finance team defined the task as *assessing the* creditworthiness *of individuals using their demographic and socio-economic data*. The public safety team defined the predictive task as *allocating police presence to* high-risk *areas*. The social media team defined the task as *predicting whether a news article is fake*. Note that notions of applicant's *creditworthiness* or neighborhood's safety *risk*,

[3]To maintain author anonymity during the review phase, we will share our IRB approval letter and if our paper is recommended for publication after the review process.

Table 1. An overview of the predictive tasks students chose to analyze through a pipeline-centric view of the ML pipeline.

Domain	Students	Predictive Task
Child Welfare	3	Predicting the risk of a child running away from their foster care within 90 days of being in the child welfare system, based on a combination of demographic and clinical characteristics, and information about them in the welfare system.
Education	4	Predicting whether an individual is cheating in an online exam based on visual and auditory cues (e.g., irregular eye/body movement, mouse movement, facial expression, and noise).
Employment	6	Predicting whether a company will hire an applicant based on the information in their resume.
Finance	3	Assessing the credit worthiness of individuals using their demographic & socio-economic data.
Housing	3	Predicting the value of a real-estate properties
Medicine	8	1) Predicting which people will/should receive an organ transplant. 2) Predicting prognosis (e.g., mortality) for comatose patients post-cardiac arrest using demographic, medical history, and medical screening data (e.g., CT scan, EEG data)
Public Safety	4	1) Allocating police presence to high risk areas; 2) Identifying suspicious individuals given a list of wanted criminals.
Social Media	4	Predicting whether a news is fake.
Transportation	2	Detecting objects (e.g., human, vehicles, road conditions) for AVs.

or *fakeness* of an article's content are not well-defined targets. Other teams did not adequately distinguish between the construct of interest and its operationalization as the target of prediction. For example, some members of the organ transplant team characterized the task as deciding *which people* should *receive an organ transplant*. In contrast, others characterized it as *predicting whether an individual would receive an organ transplant in a given hospital*. Note the difference between *"should"* and *"would"*.

Theme 2: harms were characterized in broad strokes. Some teams described harm without specifying the groups or communities that could be impacted by it and the baseline of comparison. For example, the housing team stated *price discrepancies* due to property location as the harm occurring in their domain but did not specify who could be negatively impacted by price discrepancies and in comparison with what reference group this should be considered a harm. Another example was *spread of fake news* as the harm without mentioning whom it can impact negatively and how.

Theme 3: Students had difficulty mapping the observed harm to a plausible underlying cause. For example, the child welfare group attempted to explain the harm against Black communities by noting that the feature, Race, was not quantified with sufficient granularity. The tool allowed only three racial categories: White, Black, and Other, and they hypothesized that that could be the cause of the disparity. Note that there is no plausible mechanism through which this lack of granularity might have led to disparities in child welfare risk assessments against Black communities. As another example, the finance team offered *biases of developers* as the potential source of disparity in lending practices.

Theme 4: Students had difficulty mapping their hypothesized causes to plausible remedies. For example, the Child Welfare team proposed randomly selecting instances for inclusion in the training data. The online exam proctoring group suggested further transparency (e.g., telling students what behavior results in a cheating flag) to reduce errors.

Theme 5: Students used broad-stroke language to describe causes and remedies of harm. For example, several teams referred to *biased data* as the underlying cause of harm and offered **more comprehensive data collection** as the remedy. While correct, such high-level assessments are unlikely to lead to concrete actions in practice. Another commonly proposed remedy was *human oversight* of decisions without specifying the efficiency ramifications of leaving the final call to human decision-makers.

In a session after the data collection and analysis, the instructor led a class-wide discussion in which students were encouraged to reflect on the activity and some of the gaps in the arguments presented by student teams.

3 A SURVEY OF PIPELINE-FAIRNESS

As a first step towards operationalizing the pipeline-based approach to fairness in ML, we provide a review of the ML fairness literature focused on creating a taxonomy of fairness work that locates, measures, or mitigates problems along the ML pipeline. In addition to serving as a first step towards a resource that ML practitioners can use to identify ways to diagnose and mitigate problems in their pipeline, our survey of the work already done allows us to point to the gaps in the literature that must be addressed in order to have a full understanding of how choices made along the ML pipeline translate to model behavior— and create tools which operationalize this understanding to build more effective systems.

3.1 Survey Methods and Organization.

In order to better understand the current landscape of algorithmic fairness research throughout the ML pipeline, we performed a thorough survey of the recent literature, which we classified depending upon which area of the pipeline they analyze. We gathered papers from NEURIPS, ICML, ICLR, FAccT, AIES, EAAMO, CHI, and CSCW for the past five years, i.e. 2018-2022, that contain any of the following terms in their title, abstract, or keywords: "fairness", "fair", "discrimination", "disparity", "equity".[4] In addition, we performed a series of Google Scholar searches to ensure our survey did not miss high-impact work published in other venues: One search used the keywords listed above and included papers published in any venue in the past five years with over 50 citations, through the top 50 results returned by this Google Scholar search. Additional Google Scholar searches used keywords from each step in the pipeline individually, to attempt to find papers targeted at each stage: for example, "data collection" and "fairness" and "machine learning". This results in ~1000 papers overall which fit our search criteria.

We then manually inspect each paper to understand whether and how the reported research is an instance of a pipeline-aware approach to fairness. That is, does it identify, measure, or mitigate a concrete cause of unfairness due to choices made in a specific stage of the ML pipeline. If so, we categorize the paper along two axes: what part of the pipeline it corresponds to (problem formulation, data choice, feature engineering, statistical modeling, testing and validation, or organizational realities), and whether it identifies, measures, mitigates, or provides a case study of a pipeline-based fairness problem.[5] Of our approximately 1000 papers, ~300 satisfied our criteria of being "pipeline-aware" approaches to fairness.

We present our full categorization of all the papers that we found related to pipeline-aware fairness, broken down into what stage of the pipeline they were most related to, and whether they were case study, identification, measurement, or mitigation papers, in Table 2. In our survey below, we give a sample of the space: we do not aim to be completely comprehensive, but instead, we aim to both highlight both some of the most well-known papers connected to each component of the pipeline, as well as those that offer a novel or promising perspective to understudied areas.

[4]When available: this is with the exception of NEURIPS, which we gather for years 2017-2021 due to the date of the conference relative to the time writing this work, and EAAMO, which started in 2021.

[5]By *identifying* a pipeline fairness problem, we refer to papers that point to a previously unobserved source of bias on the machine learning pipeline either through theory or through experimentation with training pipelines on common machine learning datasets; by *measuring* a pipeline fairness problem, we refer to papers that provide a generalized technique for how to identify or gauge the magnitude of a specific source of unfairness along the machine learning pipeline; by mitigating, we mean paper which develop a technique for addressing a source of bias along the AI pipeline when it arises; and by a case study we refer to an example of how a choice on the machine learning pipeline lead to unfairness in a specific application, often on an already deployed model.

3.2 Viability Assessments

3.2.1 Definition and Decisions. Viability assessments refer to a series of early investigations into whether including an ML component within the decision-making system is preferable to the status quo of decision-making; and if so, if it is *possible* to build a net-benefit ML system given available resources including data, expertise, budget, and other organizational constraints. Some decisions to make during this stage include: What are the policy goals of the decision-making problem? How can introducing an ML model promote that goal? How should the ML component be scoped? Do we have community and institutional buy-in? Is there organizational capacity to procure, build, or maintain this algorithm?

3.2.2 Case Studies and Problem Identification. To the authors' knowledge, there is very little work within the fairness literature on documenting, or detailing the bias that can arise out of, or mitigating bias from the viability assessment process. Raji et al. [246] provide extensive evidence on numerous deployed algorithmic products that simply do not work–examples of badly scoped projects [105, 224, 285]. They warn against the presumption of AI functionality, and point to several failure modes that could be remedied with a viability assessment step: such as attempting conceptually or practically impossible tasks. Wang et al. [298] point out several common flaws of predictive optimization, including the discrepancy between intervention vs. prediction, lack of construct validity, distribution shifts, and lack of contestability. Viability assessments also provide an avenue for refusal to build an ML system—though there have been discussions in the community around when to refuse to build [19], there is little published work on case studies detailing how the decision to build or not build was made. Indeed, recent work has pointed to how organizational factors—such as lack of a company's support for ethical AI approaches, or time pressure, or focusing only on client demands [237] can lead to pushing ML systems ahead without careful consideration of whether or not deploying a system in that domain is a net benefit in the given context.

3.2.3 Measurement, Mitigation, and Tools. ML practitioners have proposed guidelines for initial scoping of ML projects[6]. More recently, Wang et al. [298] have presented a rubric for assessing the *legitimacy* of predictive optimization. Coston et al. [78] propose an initial iteration of a question bank to assess the validity of a given AI design. To our knowledge, existing proposals, while promising, have not yet been evaluated in practice.

3.3 Problem Formulation

3.3.1 Definition and Decisions. Problem formulation [236] is the translation of a real-world problem to a prediction task: for example, turning a bank's need to select certain individuals to give loans to, or the need to reach children in danger of not graduating high school with certain resources, into a machine learning system with numerical inputs and outputs. We focus on three main decisions here: selecting a prediction target, what inputs should be used to predict that target, and the prediction universe. The selection of a prediction target translates the ultimate goal of a business or policy problem into a numerical data representation of that goal: for example, predicting "creditworthiness" by predicting the probability of missing a payment on a loan. Selection of inputs includes discussion over what features in the available data are acceptable to use to predict the target: e.g. whether it is morally acceptable to use access to a telephone as a predictor of a failure to appear in pretrial risk assessment instruments [118, 190]. The selection of the prediction universe determines who predictions are made over to solve this problem: for example, when predicting likelihood of not graduating high school, is this predicted over 7th graders, 9th graders, or 11th graders?

[6]See, e.g., http://www.datasciencepublicpolicy.org/our-work/tools-guides/data-science-project-scoping-guide/

3.3.2 Discussions and Considerations around Problem Formulation. Several works have provided helpful discussion outlining what constitutes problem formulation [236], how the process occurs and can be influenced by organizational biases, and what factors are important to consider during the problem formulation process [78, 140]. In particular, recent work has drawn on the concepts such as *validity* and *reliability* from the social science literature [78, 140] as a way to interrogate choices made during the problem formulation process. As an example, testing for validity "attempts to establish that a system does what it purports to do" [78]: e.g. as the authors note, establishing validity may be difficult in a predictive policing model that purports to predict *crime*, but in fact predicts new *arrests*, given the large body of work that points to racial disparities in arrest data [189].

3.3.3 Case Studies and Problem Identification. Several recent works have pointed to the impact of problem formulation on equity in model predictions. Obermeyer et al. [225] show that in a health care distribution algorithm meant to identify the sickest patients to recommend them for extra care, the choice to use health care costs as a prediction proxy adds to racial disparities in health care allocation. Black et al. [37] and Benami et al. [25] show that even *how* a given prediction target is formulated—e.g., in the case of Black et al., predicting tax noncompliance to select individuals for audit—as a regression problem (i.e., the dollar amount of misreported tax) or a classification problem (a binary indicator of whether tax noncompliance was over a certain amount) leads to large distributional changes in who is selected by an algorithm, thus impacting algorithmic equity. In the case of a model predicting tax noncompliance, changing from a classification to a regression problem shifts the distribution of suggested audits from a lower-income to a higher-income population. Benami et al. [25] also point to the impacts of choosing a prediction universe: the authors show that decisions of what types of permit reports to include in the data can lead to or mitigate disparate impact in environmental remediation due to a concentration of certain types of regulated facilities in areas with higher minority populations.

3.3.4 Measurement, Mitigation, and Tools. Developing protocols for assessing problems of various notions of validity (internal, external, content, predictive) of the target variable, predictive attributes, and prediction universe is a promising and necessary avenue of future work. There is preliminary work along this axis in the form of checklists and protocols, e.g. [78], but we suggest it may be possible to make a suite of quantitative tests as well. For example, access to appropriate data, ML practitioners could leverage existing model-level bias-testing frameworks [24, 258] to test for bias across a series of potential prediction tasks to inform the decision of which to choose. Tests for predictive validity simply require investigating whether the proposed target variable is predictive of other related outcomes. In fact, Obermeyer et al. [225] used such a test to uncover the racial bias behind using health care cost as a proxy for health care need, by regressing health care cost on other metrics of illness (e.g. the number of active chronic conditions a patient has), finding that there was a disparity in the correlation between health care costs and sickness in white patients versus that in Black patients. Investigating the extent to which questions of construct validity and reliability can be operationalized into a set of tests (e.g., testing correlations between potential prediction tasks, or checking for consistent model performance across two different methods of measuring the output) may be a promising area for utilizing tools and methods from social sciences.

3.4 Data Collection

3.4.1 Definition and Decisions. Data collection involves collecting or compiling data to train the model. This involves making choices—or implicitly accepting previously made choices—about how to sample, label, link, and omit data.

Some questions include—What population will we sample to build our model? How will we collect this data? What measurement device will be used?

3.4.2 Case Studies and Problem Identification. The algorithmic fairness literature is rife with examples of disparate performance and selection across demographic groups stemming from problems with the training data: for example, datasets that are unbalanced across demographic groups, i.e. have sampling bias, both in terms of sheer representation, and representation conditional on outcomes [43, 210]; have data that is disparately noisy or perturbed in some way [300]; or have labeling bias or untrustworthy labels [189].

A string of recent papers test how potential results of data collection problems—e.g. unrepresentative samples, insufficiently small data samples, differing group base rates[7]—influence machine learning model behavior from the perspective of accuracy and fairness [8, 89, 178]. Interestingly, they find that increasing dataset size, or even reducing the disparity in base rates, does not always reduce disparities in selection, false positive, or false negative rates.

However, fewer papers point to, document, measure, and show how to mitigate the aspects of the *data collection process itself* that lead to these various forms of biased datasets. Paullada et al. [239] point to failure modes in the data collection process from a high level. The Human and Computer Interaction community has done a more thorough job of considering what failure modes can occur in certain aspects of the data collection process, such as crowdsourcing data set labels with workers from platforms such as Mechanical Turk [135, 208, 262]. There are also a few instructive papers that document how sampling and label bias crept into certain real-world ML projects [201, 260]. For example, Marda and Narayan [201] show sources of historical bias, sampling bias, measurement bias, and direct discrimination as well as arbitrariness in the collection of data for the creation of Delhi's predictive policing system: for example, for measurement bias, they explain how the techniques used to map Delhi were inconsistent over the development of the tool, since the police department's ArcGIS system license expired; that there were often no formal addresses for less wealthy parts of the city thus identifying call location was largely guesswork; and that women who made calls were less likely to know their address since they often largely stay inside and are less aware of their surroundings.

While this literature does an excellent job of illuminating failure modes that can be surfaced through engaging with ML practitioners and data workers[8]—since many papers in this area are structured around interviews—they may elide more low-level technical sources of data collection problems, such as record linkage issues leading to non-IID data dropping. While they are extremely helpful in the important first step of identifying problems, they provide little direction for creating operationalizable solutions to these problems, or often even sufficiently detailed information about each system studied to be able to understand the mapping between data collection problems and model behavior. A promising area of future work is to bridge the problems identified by the HCI, CSCW, and other literatures, with the technical detail of the algorithmic fairness literature to introduce mitigation techniques for data collection harms.

3.4.3 Measurement, Mitigation, and Tools. There are myriad papers introducing methods of mitigating bias in model predictions given various data problems by making modeling changes. Common methods include data reweighting [148], data debiasing [39, 151], synthetic sampling [269, 292], and using specialized optimization functions for creating fair classifiers (according to traditional metrics) with unbalanced data [144, 177]. There are also less common data interventions, such as one paper by Liu et al. [185] which shows how to find which subgroup in the data is likely to have noisy labels, and then introduce a technique of inserting *more* noise into the labels of other subgroups in order to

[7]Which could be true or the result of measurement error
[8]i.e., individuals who actually put the data together, such as Mechanical Turkers

increase fairness and generalization, and often accuracy as well. Another interesting line of work points to how to add additional training samples to the data in order to improve fairness outcomes. [46, 53]

However, there is less work targeting mitigating bias in the data collection process itself. One well-known exception is datasheets for datasets [110], a paper that introduces a worksheet to fill out when building and distributing a new dataset in order to encourage thought around potential risks and harms as the dataset is being created (e.g. Over what timeframe was the data collected?; What mechanisms were used to collect the data?), to document potential failure and bias modes for future consumers of the data, and to document the collection, labeling, and processing steps so that if there are steps taken that may lead to bias, future users of the data can be aware. While this work is excellent, to the authors' knowledge, there is little understanding of how well this technique works in practice to prevent data collection mishaps—while some preliminary evaluation papers exist [41], there are none that evaluate its effectiveness in a real-world model building scenario. Evaluating such bias prevention techniques is imperative to understand how best to operationalize a pipeline-aware approach to fairness.

3.5 Data Preprocessing

3.5.1 Definition and Decisions. Data preprocessing refers to steps taken to make data usable by the ML model–e.g., dropping or imputing missing values, transforming (standardizing or normalizing data), as well as feature engineering, i.e. deciding how to construct features from available information for use in the model (e.g. how to encode categorical, text, image and other data as numbers), and which of the constructed features to use for prediction.

3.5.2 Case Studies and Problem Identification. There are myriad ways for biases to enter through data preprocessing decisions. Some of these, particularly those around some aspects of feature engineering, have been explored by the literature—such as creating or selecting features that are differentially informative across demographic populations [20, 99, 108], the fairness impacts of including spurious correlations predictive models [158, 303], or which choice of features among those available in the data lead to the least disparate impact [69–71].

But other entry points for bias are just beginning to be explored. For example, there is little work on understanding the fairness impacts of imputing missing data or dropping rows with missing data. Jeanselme et al. [142] show that data imputation strategies in the medical context *do* have an impact on accuracy across demographic groups, and that imputation strategies which lead to very similar overall model performance can still lead to different accuracy disparities across groups. Relatedly, Biswas et al. [32] show, among other effects of preprocessing choices, that dropping rows of a dataset with missing values instead of imputing those rows can lead to sizable differences in fairness behavior due to the changes to the training distribution. These works are an excellent first step, and pave the way for an exciting and necessary avenue for future work: developing methods for how to choose between preprocessing options given information about the data and the modeling pipeline.

Another vastly under-explored area is how data encoding–or how a feature is numerically represented—can have fairness impacts. Wan [295] presents an interesting case study of how data representation can affect bias in NLP translation models, and an interesting mitigation technique. They show that differences in translation performance between different languages may not have to do with the structure of language itself, but instead with the granularity of the representation of the language: for example, word length (longer words lead to lower performance), and how many bytes it takes to represent characters in a word. This performance disparity can be mitigated by using more granular representations of language pieces to equalize representation length across languages (e.g. subwords, different representations for characters). They also call for the creation of a new character encoding system that encodes characters outside the English alphabet more efficiently than current methods to further mitigate bias stemming from representation length. While it may be difficult to explore the impacts of different data encoding due to the likely contextualized nature of its effects, we still believe further examples of how such encodings can lead to bias are an important avenue to pursue to further understand sources of bias along the pipeline.

3.5.3 Measurement, Mitigation, and Tools. There are several methods of testing for, and mitigating, bias from the inclusion of certain *features*, but tools for isolating and removing bias from other data preprocessing steps are much more understudied. For example, Yeom et al. [322] provide methods for searching for and removing proxies for protected attributes in regression models, and Frye et al [106] introduce Asymmetric Shapley Values, a technique that can be used on a wider range of models to determine whether a feature is allowing a protected attribute to causally influence the model outcome, so that the feature can then be pruned. Additionally, the FlipTest technique [35] also produces suggestions of which features may be the source of disparate outcomes across any two populations the user may wish to compare, without attention to causality—this may be especially helpful in narrowing down the list of features to further investigate for impacts on differences in selection rate, or simply to find statistical, and not causal, discrimination. Finally, Belitz et al. [23] provide an automatic feature selection framework that only selects features that improve accuracy without reducing a user-specified notion of fairness. However, this method saw considerable drops in accuracy when selecting features in this manner. We note that even among feature-related measurement and mitigation tools, there is little work surrounding how to identify and mitigate bias from the use of features with different variances or predictive power across demographic groups.

Despite the capacity for feature engineering and imputation choices to affect model fairness, our survey failed to identify any tools that assist practitioners in mapping out the effects of their prepossessing choices from the perspective of preventing bias and suggesting mitigations. However, there may be some potential for adapting tools that have been developed for more general data exploration and preprocessing to incorporate fairness contributions as well. For instance, Breck et al. [42] describe a data validation pipeline for ML used at Google that proactively searches for data problems such as outliers and inconsistencies; and there are several data-cleaning tools that even suggest transformations according to best practices [136, 167]. Further exploring the landscape of these more general tools and their potential applicability to questions of model fairness seems to be a fruitful avenue for future work. Crucially, however, we note that we are unaware of any such systems that presently account for bias-related desiderata.

3.6 Statistical Modeling

3.6.1 Definition and Decisions. After deciding how to preprocess data, model makers must decide how they will create a model for their data and how it will be trained. Decisions here include what type of model will be used, the learning rule and loss function, regularizers, the values of hyper-parameters determining normalization and training procedures, among other decisions made continuously throughout model development [165, 172]. For example, choosing between linear, forest-based, or deep models; singular models or ensembles, choosing the architecture of deep models; what constraints to add to the loss function, how to optimize that loss function (e.g. SGD or momentum), among many other choices.

3.6.2 Case Studies and Problem Identification. The majority of algorithmic fairness papers show how to *intervene* on a model's loss function or prediction process to reduce biased behavior— but few papers point to sources of bias *stemming* from a model's loss function and other modeling decisions, and how to identify such problems. However, *every* decision at this can lead to downstream bias: for example, one could choose a learning rule that over-relies on certain features in the data, leading to skewed predictions for certain populations [173], or that over- or under-emphasizes outliers or minority populations. Model type selection has been shown to impact fairness behavior: especially the choice of high-capacity models with increased variance, as these models can be more unstable to small perturbations in training setup [34] leading to procedural fairness concerns about the nature of the decision process. A growing number of works have pointed to degradations of fairness behavior in robust models [194, 275, 311] and differentially private models [15, 282]. D'amour et al. [84] point to the importance of loss function choice in fairness behavior: they show without explicitly specifying a desired behavior–including fairness–within a model's loss function, the resulting model is unlikely to naturally display near-optimal or even acceptable behavior on that desired property.

3.6.3 Measurement, Mitigation, and Tools. Most of the focus on fairness interventions in statistical modeling is centered around changing model loss function or prediction process to enforce fairness constraints [3, 24]. However, more recently, these conventional techniques which enforce group and individual fairness metrics on top of a decision system have garnered criticism [8, 37, 280], and there has been evidence showing that enforcing such constraints can actually degrade fairness according to those same metrics [300].

Outside of intervening on the loss function, there are few papers that introduce mitigation techniques for bias introduced at other stages of the statistical modeling process. Notable exceptions include Islam et al [138] and Perrone et al [240], who show that hyperparameter tuning can lead to increased fairness at no cost to accuracy. Perrone et al [240] introduce a technique, Fair Baysian Optimization (FBO), which is model and fairness-definition agnostic, to select hyperparameters that optimize accuracy subject to the fairness constraint. They also experimentally demonstrate that regularization parameters are the most influential for fairness performance, and that higher regularization leads to higher fairness performance in their experiments. We hope that this technique can be used to understand further relationships between hyperparameter changes and fairness behaviors over a variety of different models, metrics, and deployment scenarios. However, as has been a common pattern, we did not discover papers that developed tools to *measure* the extent to which statistical modeling choices impacted model fairness behavior.

3.7 Testing and Validation

3.7.1 Definition and Decisions. Model testing and validation refers to the processes by which a model is determined to be performing well, both in relation to other models in the training set, but also on unseen data. Some decisions here include whether a model be evaluated only on its predictive accuracy, AUC, F1 score, or another performance metric; on fairness metrics as well (and choosing which); some notion of privacy or robustness. It also includes decisions such as what size of the dataset will be reserved for evaluation (train/test/evaluation split); what datasets the model will be evaluated on; and how many trials or k-folds the model will be evaluated on.

3.7.2 Case Studies and Problem Identification. Several papers have discussed the perils of evaluating systems on the same benchmark data sets: this can lead to overfitting to specific data sets [239] that almost guarantees distribution shift to deployment domains, meaning that results are unrepresentative for many real-world fairness applications [260], and suggests that several experimental results in the fairness literature—including those related to pipeline interventions—may be incorrect [89, 178]. Other papers still have drawn attention to specific distributional problems within commonly used fairness benchmarking datasets, such as an under-representation of older individuals [234].

Other papers have suggested issues with the metrics used to compute bias in machine learning systems themselves–e.g. Lum et al. [192] show that many of algorithmic bias metrics "are themselves statistically biased estimators of the underlying quantities they purport to represent"; others have shown that under some circumstances, such as label bias or extreme feature noise, enforcing fairness metrics can actually lead to *decreased* fairness behavior along those very same metrics in the model [300]. We believe there may be other sources of unfairness in the testing and validation part of the pipeline—such as mismatching testing and validation metrics to the application context (choosing equalized odds when a more application-specific metric would be applicable, for example [37]); but we were unable to find studies of any such problems on the ground.

3.7.3 Measurement, Mitigation, and Tools. There are several frameworks available that allow machine learning practitioners to test for the bias in their models' predictions [24, 121, 258, 306], however we note that these frameworks do not target which part of the pipeline may be adding to this bias, it only allows for bias testing to occur along the most popular fairness metrics. These frameworks allow for the most basic bias mitigation to testing and validation problems: not testing for fairness during validation at all. Several recent works have also added new or expanded datasets to be used as benchmarks in fairness contexts to prevent problems of dataset overfitting [89, 178], though fairness researchers may benefit from exploring datasets even beyond these, perhaps collaborating or borrowing data from social scientists, or exploring many of the less popular publicly available dataset such as [environmental dataset],[american community survey]. However, there were no papers that we could find during our survey which showed how to measure the extent to which testing and validation design choices influence downstream fairness behavior.

3.8 Deployment and Monitoring

3.8.1 Definition and Decisions. Deployment refers to the process of embedding a model into a larger decision system. For example, some decisions here include: how will the model be used as a component of the decision system into which it is embedded? Will the model's predictions directly become the final decision? If there is human involvement, where and how will that occur? How much discretion do humans have over adhering to model recommendations? How are model predictions communicated to decision-makers? Monitoring refers to how a model's behavior is recorded and responded to over time in order to ensure there is no degradation in performance over time, fairness or otherwise. Decisions here include: Will monitoring occur? If so, how will performance over time be measured? How and when will data drift be measured and addressed?

3.8.2 Case Studies and Problem Identification. There has been a stream of theoretical work from the algorithmic fairness literature trying to model or guarantee fairness in a joint human-ML system [156, 198]. For example, Keswani et al [156] learn a classifier and a deferral system for low-confidence outputs, with the deferral system taking into account the biases of the humans in the loop. Donahue et al.[93] develop a theoretical framework for understanding when and when not human and machine error will complement each other. While this initial largely theoretical work is promising, these techniques should be tested and compared on deployed systems—when do deferral systems work in practice, and can they lead to biases of their own? Do the human-in-the-loop design suggestions work in practice?

However, there are conflicting results as to whether human-in-the-loop systems outperform models on their own when it comes to bias. Green and Chen [116, 117] show that including Mechanical Turk workers to aid algorithmic decisions consistently decreased accuracy relative to the algorithm's performance alone, and that humans in the loop also exhibited racial bias when interacting with ML predictions. However, others [59, 87] have found that in the case of child welfare screenings, allowing call screen workers *did* reduce disparity in the screen-in rates of Black versus white children. Symmetrically, Jacobs et al. [141] show that including machine learning recommendations in selecting antidepressants for patients did not improve clinician's treatment decision performance, and in fact, interacting with incorrect recommendations—even when supplied with explanations—led to decreased performance.

3.8.3 Measurement, Mitigation, and Tools. There are a few papers providing mitigation techniques and tools for fairness monitoring and preventing distribution shifts, but we note there are no papers that appeared during our survey which provided measurement or mitigation techniques for addressing biases that arise as a result of including humans in ML decision systems—an important area of future work.

A series of recent papers have introduced methods to make models invariant to distribution shift from the perspective of accuracy as well as fairness [31, 75, 277]. However, there are several open questions in this area of research: such as, when do we want to ensure fairness criteria over data drift, and when do we want to alert the model practitioner that the distributional differences are so large that the model is not suitable for the deployment context?

Pertinently, two recent papers [9, 112] propose tools for fairness monitoring over time in deployment. Albarghouthi and Vinitsky [9] develop a technique called fairness-aware programming (implemented in Python), which allows programmers to enforce probabilistic statements over the behaviors of their functions, and get notified for violations of those statements—their framework is flexible to many behavioral desiderata even beyond fairness, and can also combine several probabilistic statements and check them simultaneously. The flexibility of the system potentially allows for a wide range of contextualized fairness desiderata to be enforced– however, it does have limitations; for example, it cannot implement notions of fairness over individual inputs such as individual fairness. Ghosh et al [112] implement a system that measures the Quantile Demographic Drift metric at run time, a notion of fairness that can shift between group and individual conceptions of fairness based on the granularity of the bins it calculates discrepancies over. Their system also offers automatic mitigation strategies (normalizing model outputs across demographic groups) and explanation techniques to understand mechanisms of bias. Additionally, Amazon's SageMaker Clarify framework allows for the tracking of a variety of traditional fairness metrics to be monitored at runtime [121].

4 RESEARCH AGENDA

Reflecting on the state of fairness literature from a pipeline-aware perspective, we offer five key insights to start operationalizing a pipeline-aware approach to fairness. Namely, we identify (1) the need to investigate and document choices made along real-world pipelines, including those related to bias mitigation; (2) the need to bridge the gap across literatures which *identify* ways the bias enters ML pipeliens on-the-ground, such as HCI, and literatures which build operational mitigation techniques, such as FairML; (3) the need to study interaction effects across decisions made along the pipeline; (4) the need to address holes in the current research–paying attention to neglected areas of the pipeline such as viability assessment, and entire modes of research such as producing *measurement* techniques to catch many of the entry points for bias identified along the pipeline; and (5) finally, the need to produce guidance on how to choose among several biased building choices.

Investigation of Real-World Pipelines The algorithmic fairness literature has a dearth of knowledge about what on-the-ground pipelines look like. Although many papers outline abstracted pipelines (including our own), the actual model creation steps taken by practitioners for a variety of real-world systems (e.g. in consumer finance, healthcare, hiring systems) are at the very least not cataloged in any centralized place. Bias entering from choices along the pipeline cannot be studied, measured, and mitigated if we do not know what choices are being made—thus mapping out actual pipeline choices is a necessary step to operationalize a pipeline-aware approach to fairness.

Action Items: (1) We encourage exploration and documentation of pipeline decisions made in a variety of machine learning pipelines in different applications; and (2) centralization of this information so that researchers can study how pipeline decisions they may have been unaware of impact fairness behavior.

Stage	Step	Problem Identification	Measurement	Mitigation
Viability Assessments	Cost/Benefit General	[246], [19], [237]	[298], [78]	
Problem Formulation	Prediction Target	[37], [225], [1], [238]		[25]
	Predictive Attributes	[118]		
	General	[249], [328], [220], [222], [236] , [63], [139], [58], [137], [189]		
Data Collection	Sampling	[43], [315], [281], [229], , [46]	[210], [323]	[269], [292], [28], [290], [271], [251]
	Annotation	[305], [300], [257], [64]	[185], [297], [30]	[317]
	Feature Measurement	[295], [230], [106]	[284]	
	Record Linkage			
	General	[57], [128], [260], [235], [148], [201]	[178], [89], [333], [233], [143], [150], [6], [39], [151]	[53], [110], [41]
Data Preprocessing	Feature Creation			
	Feature Selection	[23], [322], [35], [173]	[22], [286]	[321], [103], [124]
	Data Cleaning (Omission)		[309], [42]	[48], [330], [136], [167]
	General	[47], [199], [183], [32]	[81], [266], [62], [272], [301], [107]	[320], [228], [176], [95], [283], [231], [188], [82], [45], [52], [289], [161],
Statistical Modeling	Hypothesis Class	[33], [216],	[335]	[182]
	Optimization Function	[243], [282], [123], [85],	[84]	[310] , [256] , [10] , [83] , [256] , [251] , [252] , [267] , [68] , [61] , [250] , [187] , [180], [312], [311]
	Regularizers			[153], [264]
	Hyperparameters	[202], ,		[294], [274],[261], [138],[240]
	General	[193], [111], [186], [77], [29] [8], [15], [275]	[221], [54], [168], [296] , [194]	[319], [171], [255], [179], [18], [97], [152], [44], [204],[217], [215], [104], [329] , [163] , [270] , [119] [170] , [242] , [334] , [259] , [302] , [211] , [308] [3] , [160] , [197] , [65] , [51] , [109] , [166] , [227] , [11], [115] , [102] , [56] , [288] , [244] , [169] [94] , [126] , [40] , [293] , [325] , [313] , [205] , [218] , [50] , [277] , [98] , [155] , [324] , [79] , [114] , [245] , [184] , [265] , [174] , [67] , [21] , [91] , [212] , [291] , [125] , [4] , [307], [101] , [120] , [16] , [326] , [206],
Testing and Validation	Train-test split			
	Evaluation Metrics	[130], [96], [38], [332] , [157],	[209] , [268] , [88] , [162] , [331] , [149] , [279] , [200] [145] , [127] , [80] , [76] , [109] , [214] , [14], [287], [254], [314],	[213], [207] , [327] , [159] , [49] , [66] , [17] , [316] , [55] , [318] , [134] , [26]
	General	[226] , [273]	[36], [192], [24], [121], [239], [258], [164]	[74], [90], [60], [299],
Deployment and Integration	Human-Computer Handoff	[141] , [156], [116], [203], [92], [175], [195], [304], [93]	[59]	[13]
	Maintenance Oversight		[9],	
	General	[263] , [191] , [7], [232], [196], [117], [116], [59], [87]	[112], [113],[278], [247], [129], [31]	[219], [276], [12], [223] , [280], [198], [75]

Table 2. An taxonomy of the papers surveyed into the various sections of the ML pipeline they study, and whether they identify, measure, or mitigate a source of bias. The pink papers correspond to case studies within problem identification. Yellow colored references denote papers that correspond to more traditional approaches to fairness, i.e. imposing a fairness constraint on top of a pre-made modeling process, or introducing a new notion of fairness in the testing and evaluation section.

Disconnect between Problem Identification and Mitigation Bias problems are often discovered in disciplines on the fringe or outside of machine learning, such as human-computer interaction literature and database management [260]. Since many papers which point to problems along the AI pipeline, especially in the earlier stages, are structured around interviews—they may elide more low-level technical sources of pipeline choices leading to bias, e.g. such as record linkage issues leading to non-IID data dropping. While they are extremely helpful in the important first step of identifying problems, they provide little direction for creating operationalizable solutions to these problems, or often even sufficiently detailed information about each system studied to be able to understand the mapping between data collection problems and model behavior.

Symmetrically, many papers in the mainstream algorithmic fairness literature do not test their techniques on real decision-making systems–or even on datasets beyond Adult, German Credit, and a few others. Though case studies may often be disregarded as simply implementing old methods and thus not novel, it is crucial that we give them more attention so that we can learn about the failure modes of the techniques that we create. While there are papers about the overall problems production teams face when implementing fairness goals, e.g. [132, 196, 248, 260], these do not provide an in-depth catalogue of the successes and failures of all pipeline intervention techniques in practice. Proposed methods to intervene on the machine learning pipeline should be tested in a variety of real-world systems to see where they fail and when they are helpful in practice.

Action Items: (1) We strongly encourage testing of pipeline-based bias mitigation methods proposed to date in on-the-ground ML systems. This is necessary to uncover which perform best under various circumstances, determine their failure modes, and design methods to integrate these techniques in ML system building processes; (2) We encourage bridging the problems identified by the HCI, CSCW, and other literatures, with the technical detail of the algorithmic fairness literature to introduce mitigation techniques for data collection harms; (3) In particular, we encourage AI Fairness researchers to build off of tools for addressing generalized data and modeling pipeline issues (i.e. not specialized to fairness problems) and see how we can adopt them for algorithmic harms–for example, extending Breck et al. [42]'s data validation pipeline for ML to address fairness concerns.

Interaction Effects Most fairness-related papers focus on *one issue* in the pipeline: they present a bias problem, then often give a mitigation method, implicitly assuming that the identified source of bias is the only one of interest to the model practitioner, and the suggested intervention will not lead to other effects on the model, including other forms of bias. That is, much of the literature fails to provide insight into how different biases and mitigation techniques along the pipeline interact. Does solving each issue in isolation work, or does a pipeline-aware approach to fairness need to engage with interaction effects to work in practice? There are a few papers that show the impacts of the intersection between multiple sources of bias has on the effectiveness of interventions–e.g., Li et al. [181] point to how active sampling techniques to address sampling bias can make models *more* unfair when label bias is present. However, any set of choices on the machine learning pipeline may have interaction effects, suggesting many paths for exploration. Beyond studying interaction effects, there are very few papers[9] which engage with the entire pipeline: taking fairness into account when making every modeling decision, and testing along the way—which is ideally the end goal of a pipeline-based approach to fairness.

Action Items: We recommend (1) the investigation of interaction effects of various sources of bias, *and* mitigation strategies to target various sources of bias, across the ML pipeline, as well as (2) the development of tools to allow for such exploration.[10]

Holes in Research: Measurement Methods and Others Within the few papers that do discuss pipeline-aware approaches to fairness, most are problem identification or mitigation techniques— only 17%[11] of the papers that we identified are actually providing techniques to measure whether or not a given pipeline choice will lead to downstream unfairness ex-ante. But measurement is key because it may allow us to decide when or when not to take an action. How can we effectively measure algorithmic harms? Relatedly, there are many proposed solutions of how to solve a problem once it has already happened–e.g. unrepresentative data, etc.—but how can we develop tests to *prevent* choices

[9]Though there are some, e.g. [154, 281]
[10]For example, a prototype of such a tool for specific parts of the pipeline was recently introduced by Akpinar et al [8].
[11]We calculate this by taking all of the measurement papers identified (67) and subtracting the number that simply introduce new fairness metrics (15), and dividing this number (52) by the total number of pipeline fairness papers identified.

that lead to algorithmic harms? For example, can we predict beforehand whether and how fairness problems will result from the way in which a model is integrated into a given decision structure?

Additionally, we find that research on *how* bias can enter machine learning decision systems via choices made along the pipeline is unevenly distributed across the pipeline. There are some areas of the pipeline that are well-studied, such as data collection; but the majority of the pipeline has light-to-no coverage in the fairness literature: e.g. viability assessments, problem formulation, and large parts of organizational integration. These areas must be prioritized for searching for fairness failure modes and mitigation techniques.

Finally, the majority of the research that has been done measures the effect of various pipeline choices (e.g. feature selection, data preprocessing) on demographic disparity, TPR/FPR/EO, or accuracy disparity. How can we map the pipeline in a general enough fashion that we may be able to understand how more contextualized notions of fairness are impacted by pipeline decisions?

Action Items (1) We encourage building methods for *measuring* the harms introduced by decisions along the machine learning pipeline; (2) Exploring pipeline decisions that have received little attention; and (3) exploring the fairness effects of pipeline decisions beyond the most common metrics.

Guidance on Choosing Among Imperfect Design Alternatives ML practitioners often face choices between imperfect/biased alternatives. It will almost always be the case that with adequate effort, they can produce fairer, but not perfectly fair models. Any choice a designer makes will likely lead to some form of bias, some more and some less problematic in the given decision-making domain. However, there is little guidance in the literature on deciding when one alternative is better than another in light of all contextual considerations, when certain biases are conditionally acceptable, and how the answers to these questions depend on the context.

Action Items: (1) Developing normative guidelines to weigh a variety of imperfect/biased design choices against each other is a crucial avenue for collaboration between ML experts and ethicists; (2) Exploring how different types of bias are currently traded off in practice when they are discovered; and (3) Understanding if there are any relationships or couplings of different sources or forms of bias that often go together, or have opposing relationships.

5 CONCLUSION

As a whole, this works aims to aid researchers in *operationalizing* a pipeline-approach to fairness in machine learning by providing a detailed account of the sources of, measurements for, and mitigations of bias already discovered along the pipeline in the prior literature, as well as pointing to specific areas of inquiry that are required to fully understand the mapping from model building to model behavior.

REFERENCES

[1] Uk: Automated benefits system failing people in need, Oct 2020. URL https://www.hrw.org/news/2020/09/29/uk-automated-benefits-system-failing-people-need.

[2] Naacp legal defense and educational fund and student borrower protection center announce fair lending testing agreement with upstart network. https://protectborrowers.org/naacpldf-sbpc-upstart-agreement/, 2020.

[3] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/agarwal18a.html.

[4] Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 120–129. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/agarwal19d.html.

[5] NIST AI. Artificial intelligence risk management framework (ai rmf 1.0). 2023.

[6] Nil-Jana Akpinar, Maria De-Arteaga, and Alexandra Chouldechova. The effect of differential victim crime reporting on predictive policing systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 838–849, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445877. URL https://doi.org/10.1145/3442188.3445877.

[7] Nil-Jana Akpinar, Cyrus DiCiccio, Preetam Nandy, and Kinjal Basu. Long-term dynamics of fairness intervention in connection recommender systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 22–35, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534173. URL https://doi.org/10.1145/3514094.3534173.

[8] Nil-Jana Akpinar, Manish Nagireddy, Logan Stapleton, Hao-Fei Cheng, Haiyi Zhu, Steven Wu, and Hoda Heidari. A sandbox tool to bias (stress)-test fairness algorithms. *arXiv preprint arXiv:2204.10233*, 2022.

[9] Aws Albarghouthi and Samuel Vinitsky. Fairness-aware programming. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 211–219, 2019.

[10] Junaid Ali, Muhammad Bilal Zafar, Adish Singla, and Krishna P. Gummadi. Loss-aversively fair classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 211–218, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314266. URL https://doi.org/10.1145/3306618.3314266.

[11] Junaid Ali, Preethi Lahoti, and Krishna P. Gummadi. Accounting for model uncertainty in algorithmic discrimination. page 336–345, 2021. doi: 10.1145/3461702.3462630. URL https://doi.org/10.1145/3461702.3462630.

[12] Abdulaziz A. Almuzaini, Chidansh A. Bhatt, David M. Pennock, and Vivek K. Singh. Abcinml: Anticipatory bias correction in machine learning applications. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1552–1560, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533211. URL https://doi.org/10.1145/3531146.3533211.

[13] Ariful Islam Anik and Andrea Bunt. Data-centric explanations: Explaining training data of machine learning systems to promote transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445736. URL https://doi.org/10.1145/3411764.3445736.

[14] Pranjal Awasthi, Alex Beutel, Matth¨aus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 206–214, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445884. URL https://doi.org/10.1145/3442188.3445884.

[15] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

[16] Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. Rényi fair inference. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkgsUJrtDB.

[17] Maria-Florina F Balcan, Travis Dick, Ritesh Noothigattu, and Ariel D Procaccia. Envy-free classification. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2019/file/e94550c93cd70fe748e6982b3439ad3b-Paper.pdf.

[18] Mislav Balunovic, Anian Ruoss, and Martin Vechev. Fair normalizing flows. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=BrFIKuxrZE.

[19] Solon Barocas, Asia J Biega, Benjamin Fish, Jędrzej Niklas, and Luke Stark. When not to design, build, or deploy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 695–695, 2020.

[20] Alexander Bartik and Scott Nelson. Deleting a signal: Evidence from pre-employment credit checks. 2016.

[21] Yahav Bechavod, Christopher Jung, and Steven Z. Wu. Metric-free individual fairness in online learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11214–11225. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/80b618ebcac7aa97a6dac2ba65cb7e36-Paper.pdf.

[22] Clara Belitz, Lan Jiang, and Nigel Bosch. Automating procedurally fair feature selection in machine learning. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 379–389, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462585. URL https://doi.org/10.1145/3461702.3462585.

[23] Clara Belitz, Lan Jiang, and Nigel Bosch. Automating procedurally fair feature selection in machine learning. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 379–389, 2021.

[24] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, abs/1810.01943, 2018. URL http://arxiv.bs/1810.0194.

[25] Elinor Benami, Reid Whitaker, Vincent La, Hongjin Lin, Brandon R Anderson, and Daniel E Ho. The distributive effects of risk prediction in environmental compliance: Algorithmic design, environmental justice, and public policy. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 90–105, 2021.

[26] Henry C Bendekgey and Erik Sudderth. Scalable and stable surrogates for flexible classifiers with fairness constraints. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 30023–30036. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2021/file/fc2e6a440b94f64831840137698021e1-Paper.pdf.

[27] Jason R Bent. Is algorithmic affirmative action legal. *Geo. LJ*, 108:803, 2019.

[28] Elena Beretta, Antonio Vetrò, Bruno Lepri, and Juan Carlos De Martin. Detecting discriminatory risk through data annotation based on bayesian inferences. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 794–804, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445940. URL https://doi.org/10.1145/3442188.3445940.

[29] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H. Chi. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 453–459, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314234. URL https://doi.org/10.1145/3306618.3314234.

[30] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The values encoded in machine learning research. *arXiv preprint arXiv:2106.15590*, 2021.

[31] Arpita Biswas and Suvam Mukherjee. Ensuring fairness under prior probability shifts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 414–424, 2021.

[32] Sumon Biswas and Hridesh Rajan. Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 981–993, 2021.

[33] Emily Black and Matt Fredrikson. Leave-one-out unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 285–295, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445894. URL https://doi.org/10.1145/3442188.3445894.

[34] Emily Black and Matt Fredrikson. Leave-one-out unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 285–295, 2021.

[35] Emily Black, Samuel Yeom, and Matt Fredrikson. Fliptest: Fairness auditing via optimal transport. *CoRR*, abs/1906.09218, 2019. URL http://arxiv.bs/1906.0921.

[36] Emily Black, Samuel Yeom, and Matt Fredrikson. Fliptest: Fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 111–121, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372845. URL https://doi.org/10.1145/3351095.3372845.

[37] Emily Black, Hadi Elzayn, Alexandra Chouldechova, Jacob Goldin, and Daniel Ho. Algorithmic fairness and vertical equity: Income fairness with irs tax audit models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1479–1503, 2022.

[38] Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 850–863, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533149. URL https://doi.org/10.1145/3531146.3533149.

[39] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, 2016.

[40] Amanda Bower, Hamid Eftekhari, Mikhail Yurochkin, and Yuekai Sun. Individually fair rankings. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=71zCSP%5FHuBN.

[41] Karen L Boyd. Datasheets for datasets help ml engineers notice and understand ethical issues in training data. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–27, 2021.

[42] Eric Breck, Neoklis Polyzotis, Sudip Roy, Steven Whang, and Martin Zinkevich. Data validation for machine learning. In *MLSys*, 2019.

[43] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

[44] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. Balanced neighborhoods for multi-sided fairness in recommendation. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 202–214. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/burke18a.html.

[45] Maarten Buyl and Tijl De Bie. DeBayes: a Bayesian method for debiasing network embeddings. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1220–1229. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/buyl20a.html.

[46] William Cai, Ro Encarnacion, Bobbie Chern, Sam Corbett-Davies, Miranda Bogen, Stevie Bergman, and Sharad Goel. Adaptive sampling strategies to construct equitable training datasets. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1467–1478, 2022.

[47] Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R. Banaji. Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 156–170, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534162. URL https://doi.org/10.1145/3514094.3534162.

[48] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf.

[49] Ran Canetti, Aloni Cohen, Nishanth Dikkala, Govind Ramnarayan, Sarah Scheffler, and Adam Smith. From soft classifiers to hard decisions: How fair can we be? In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 309–318, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287561. URL https://doi.org/10.1145/3287560.3287561.

[50] Semih Cayci, Swati Gupta, and Atilla Eryilmaz. Group-fair online allocation in continuous time. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13750–13761. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/9ec0cfdc84044494e10582436e013e64-Paper.pdf.

[51] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019.

[52] L. Elisa Celis, Vijay Keswani, and Nisheeth Vishnoi. Data preprocessing to mitigate bias: A maximum entropy based approach. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1349–1359. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/celis20a.html.

[53] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002*, 2018.

[54] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 339–348, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287594. URL https://doi.org/10.1145/3287560.3287594.

[55] Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. Classification under misspecification: Halfspaces, generalized linear models, and evolvability. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8391–8403. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2020/file/5f8b73c0d4b1bf60dd7173b660b87c29-Paper.pdf.

[56] Violet (Xinying) Chen and J. N. Hooker. A just approach balancing rawlsian leximax fairness and utilitarianism. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 221–227, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375844. URL https://doi.org/10.1145/3375627.3375844.

[57] Yuan Chen, Wenbo Fei, Qinxia Wang, Donglin Zeng, and Yuanjia Wang. Dynamic covid risk assessment accounting for community virus exposure from a spatial-temporal transmission model. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27747–27760. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/e97a4f04ef1b914f6a1698caa364f693-Paper.pdf.

[58] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. Soliciting stakeholders' fairness notions in child maltreatment predictive systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445308. URL https://doi.org/10.1145/3411764.3445308.

[59] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. How child welfare workers reduce racial disparities in algorithmic decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3501831. URL https://doi.org/10.1145/3491102.3501831.

[60] Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=N6JECD-PI5w.

[61] Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using kernel density estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15088–15099. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2020/file/ac3870fcad1cfc367825cda0101eee62-Paper.pdf.

[62] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1887–1898. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/choi20a.html.

[63] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 134–148. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/chouldechova18a.html.

[64] Evgenia Christoforou, Pinar Barlas, and Jahna Otterbacher. It's about time: A view of crowdsourced data before and during the pandemic. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445317. URL https://doi.org/10.1145/3411764.3445317.

[65] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=DNl5s5BXeBn.

[66] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2019/file/ba51e6158bcaf80fd0d834950251e693-Paper.pdf.

[67] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression via plug-in estimator and recalibration with statistical guarantees. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19137–19148. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2020/file/ddd808772c035aed516d42ad3559be5f-Paper.pdf.

[68] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with wasserstein barycenters. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7321–7331. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2020/file/51cdbd2611e844ece5d80878eb770436-Paper.pdf.

[69] Relman Colfax. ir lending monitorship of upstart network's lending model: Initial report of the independent monitor. https://www.relmanlaw.com/media/cases/1088_Upstart%20Initial%20Report%20-%20Final.pdf, 2021.

[70] Relman Colfax. ir lending monitorship of upstart network's lending model: Second report of the independent monitor, 2021.

[71] Relman Colfax. Fair lending monitorship of upstart network's lending model: Third report of the independent monitor. https://www.relmanlaw.com/media/cases/1333_PUBLIC%20Upstart%20Monitorship%203rd%20Report%20FINAL.pdf, 2022.

[72] Federal Trade Commission. Commercial surveillance and data security rulemaking, 2022. URL https://www.ftc.gov/legal-library/browse/federal-register-notices/commercial-surveillance-data-security-rulemaking.

[73] Federal Trade Commission. Trade regulation rule on commercial surveillance and data security, Proposed 08/22/2022.

[74] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 91–98, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314236. URL https://doi.org/10.1145/3306618.3314236.

[75] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 91–98, 2019.

[76] Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 582–593, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372851. URL https://doi.org/10.1145/3351095.3372851.

[77] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing fairness over the set of good models under selective labels. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2144–2155. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/coston21a.html.

[78] Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. A validity perspective on evaluating the justified use of data-driven decision-making algorithms. *arXiv preprint arXiv:2206.14983*, 2022.

[79] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1397–1405. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/cotter19b.html.

[80] Cyrus Cousins. An axiomatic theory of provably-fair welfare-centric machine learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16610–16621. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/8b0bb3eff8c1e5bf7f206125959921d7-Paper.pdf.

[81] Cyrus Cousins. Uncertainty and the social planner's problem: Why sample complexity matters. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2004–2015, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533243. URL https://doi.org/10.1145/3531146.3533243.

[82] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1436–1445. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/creager19a.html.

[83] Sean Current, Yuntian He, Saket Gurukar, and Srinivasan Parthasarathy. Fairegm: Fair link prediction and recommendation via emulated graph modification. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394772. doi: 10.1145/3551624.3555287. URL https://doi.org/10.1145/3551624.3555287.

[84] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

[85] Ishita Dasgupta, Erin Grant, and Tom Griffiths. Distinguishing rule and exemplar-based generalization in learning systems. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 4816–4830. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/dasgupta22b.html.

[86] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, 2018.

[87] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[88] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 66–76, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462523. URL https://doi.org/10.1145/3461702.3462523.

[89] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

[90] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278729. URL https://doi.org/10.1145/3278721.3278729.

[91] Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. Two-sided fairness in rankings via lorenz dominance. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8596–8608. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/48259990138bc03361556fb3f94c5d45-Paper.pdf.

[92] Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1639–1656, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533221. URL https://doi.org/10.1145/3531146.3533221.

[93] Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1639–1656, 2022.

[94] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. 31, 2018. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2018/file/83cdcec08fbf90370fcf53bdd56604ff-Paper.pdf.

[95] Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. Fairness via representation neutralization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12091–12103. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/64ff7983a47d331b13a81156e2f4d29d-Paper.pdf.

[96] Natalie Dullerud, Karsten Roth, Kimia Hamidieh, Nicolas Papernot, and Marzyeh Ghassemi. Is fairness only metric deep? evaluating and addressing subgroup gaps in deep metric learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=js62%5FxuLDDv.

[97] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 119–133. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/dwork18a.html.

[98] Hadi Elzayn, Shahin Jabbari, Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, and Zachary Schutzman. Fair algorithms for learning in allocation problems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 170–179, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287571. URL https://doi.org/10.1145/3287560.3287571.

[99] Vitalii Emelianov, Nicolas Gast, Krishna P Gummadi, and Patrick Loiseau. On fair selection in the presence of implicit and differential variance. *Artificial Intelligence*, 302:103609, 2022.

[100] Equivant. Practitioner's guide to COMPAS core. http://quivant.com/wp-content/uploas/Practitioners-Guide-to-COMPAS-Core-040419.pdf, 2019.

[101] Golnoosh Farnadi, Behrouz Babaki, and Lise Getoor. Fairness in relational domains. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 108–114, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278733. URL https://doi.org/10.1145/3278721.3278733.

[102] Benjamin Fish and Luke Stark. It's not fairness, and it's not fair: The failure of distributional equality and the promise of relational equality in complete-information hiring games. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394772. doi: 10.1145/3551624.3555296. URL https://doi.org/10.1145/3551624.3555296.

[103] Hortense Fong, Vineet Kumar, Anay Mehrotra, and Nisheeth K. Vishnoi. Fairness for auc via feature augmentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 610, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533126. URL https://doi.org/10.1145/3531146.3533126.

[104] Adam Foster, Arpi Vezer, Craig A. Glastonbury, Paidi Creed, Samer Abujudeh, and Aaron Sim. Contrastive mixture of posteriors for counterfactual inference, data integration and fairness. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 6578–6621. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/foster22a.html.

[105] Karoline Freeman, Julia Geppert, Chris Stinton, Daniel Todkill, Samantha Johnson, Aileen Clarke, and Sian Taylor-Phillips. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *bmj*, 374, 2021.

[106] Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33:1229–1239, 2020.

[107] Georgi Ganev, Bristena Oprisanu, and Emiliano De Cristofaro. Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 6944–6959. PMLR, 17–23 Jul 2022.

URL https://proceedings.mlr.press/v162/ganev22a.html.

[108] Nikhil Garg, Hannah Li, and Faidra Monachou. Dropping standardized testing for admissions trades off information and access. *arXiv preprint arXiv:2010.04396*, 2020.

[109] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 219–226, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3317950. URL https://doi.org/10.1145/3306618.3317950.

[110] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal III Daumé, and Kate Crawford. Datasheets for datasets. arxiv. *arXiv preprint arXiv:1803.09010*, 2018.

[111] Azin Ghazimatin, Matthaus Kleindessner, Chris Russell, Ziawasch Abedjan, and Jacek Golebiowski. Measuring fairness of rankings under noisy sensitive information. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2263–2279, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3534641. URL https://doi.org/10.1145/3531146.3534641.

[112] Avijit Ghosh, Aalok Shanbhag, and Christo Wilson. Faircanary: Rapid continuous explainable fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 307–316, 2022.

[113] Stephen Giguere, Blossom Metevier, Yuriy Brun, Philip S. Thomas, Scott Niekum, and Bruno Castro da Silva. Fairness guarantees under demographic shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=wbPObLm6ueA.

[114] Naman Goel, Mohammad Yaghini, and Boi Faltings. Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 116, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278722. URL https://doi.org/10.1145/3278721.3278722.

[115] Przemyslaw A. Grabowicz, Nicholas Perello, and Aarshee Mishra. Marrying fairness and explainability in supervised learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1905–1916, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533236. URL https://doi.org/10.1145/3531146.3533236.

[116] Ben Green and Yiling Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 90–99, 2019.

[117] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

[118] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[119] Matan Halevy, Camille Harris, Amy Bruckman, Diyi Yang, and Ayanna Howard. Mitigating racial biases in toxic language detection with an equity-based ensemble framework. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385534. doi: 10.1145/3465416.3483299. URL https://doi.org/10.1145/3465416.3483299.

[120] Sariel Har-Peled and Sepideh Mahabadi. Near neighbor: Who is the fairest of them all? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2019/file/742141ceda6b8f6786609d31c8ef129f-Paper.pdf.

[121] Michaela Hardt, Xiaoguang Chen, Xiaoyi Cheng, Michele Donini, Jason Gelman, Satish Gollaprolu, John He, Pedro Larroy, Xinyu Liu, Nick McCarthy, et al. Amazon sagemaker clarify: Machine learning bias detection and explainability in the cloud. *arXiv preprint arXiv:2109.03285*, 2021.

[122] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 2016.

[123] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/hashimoto18a.html.

[124] Yuzi He, Keith Burghardt, and Kristina Lerman. A geometric solution to fair representations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 279–285, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375864. URL https://doi.org/10.1145/3375627.3375864.

[125] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

[126] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2018/file/be3159ad04564bfb90db9e32851ebf9c-Paper.pdf.

[127] Jonathan Herington. Measuring fairness in an unfair world. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 286–292, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375854. URL https://doi.org/10.1145/3375627.3375854.

[128] Danula Hettiachchi, Mark Sanderson, Jorge Goncalves, Simo Hosio, Gabriella Kazai, Matthew Lease, Mike Schaekermann, and Emine Yilmaz. Investigating and mitigating biases in crowdsourced data. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '21, page 331–334, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384797. doi:

10.1145/3462204.3481729. URL https://doi.org/10.1145/3462204.3481729.

[129] Fabian Hinder, André Artelt, and Barbara Hammer. Towards non-parametric drift detection via dynamic adapting window independence drift detection (DAWIDD). In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4249–4259. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/hinder20a.html.

[130] Gaurush Hiranandani, Harikrishna Narasimhan, and Sanmi Koyejo. Fair performance metric elicitation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11083–11095. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/7ec2442aa04c157590b2fa1a7d093a33-Paper.pdf.

[131] Daniel E Ho and Alice Xiang. Affirmative algorithms: The legal grounds for fairness as awareness. *U. Chi. L. Rev. Online*, page 134, 2020.

[132] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.

[133] White House. Blueprint for an ai bill of rights: Making automated systems work for the american people, 2022.

[134] Feihu Huang, Xidong Wu, and Heng Huang. Efficient mirror descent ascent methods for nonsmooth minimax problems. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10431–10443. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2021/file/56503192b14190d3826780d47c0d3bf3-Paper.pdf.

[135] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. CHI '19, 2019.

[136] Nick Hynes, D. Sculley, and Michael Terry. The data linter: Lightweight automated sanity checking for ml data sets. 2017. URL http://learningsys.org/nips17/assets/papers/paper%5F19.pdf.

[137] Christina Ilvento, Meena Jagadeesan, and Shuchi Chawla. Multi-category fairness in sponsored search auctions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 348–358, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372848. URL https://doi.org/10.1145/3351095.3372848.

[138] Rashidul Islam, Shimei Pan, and James R Foulds. Can we obtain fairness for free? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 586–596, 2021.

[139] Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 375–385, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445901. URL https://doi.org/10.1145/3442188.3445901.

[140] Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385, 2021.

[141] Maia Jacobs, Melanie F Pradier, Thomas H McCoy Jr, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry*, 11(1):108, 2021.

[142] Vincent Jeanselme, Maria De-Arteaga, Zhe Zhang, Jessica Barrett, and Brian Tom. Imputation strategies under clinical presence: Impact on algorithmic fairness. In *Machine Learning for Health*, pages 12–34. PMLR, 2022.

[143] Disi Ji, Padhraic Smyth, and Mark Steyvers. Can i trust my fairness metric? assessing fairness with unlabeled data and bayesian inference. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18600–18612. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d83de59e10227072a9c034ce10029c39-Paper.pdf.

[144] Weijie Jiang and Zachary A Pardos. Towards equity and algorithmic fairness in student grade prediction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 608–617, 2021.

[145] Zhimeng Jiang, Xiaotian Han, Chao Fan, Fan Yang, Ali Mostafavi, and Xia Hu. Generalized demographic parity for group fairness. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=YigKlMJwjye.

[146] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.

[147] United States Dept. Of Justice. Title vi legal manual, section vii: Proving discrimination – disparate impact, Accessed 2023. URL https://www.justice.gov/crt/fcs/T6Manual7#:~:text=)%3B%20Gaston%20Cty.-,v.,results%20in%20racial%20discrimination.%E2%80%9D%20H.R.

[148] Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2439–2448. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kallus18a.html.

[149] Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2019/file/73e0f7487b8e5297182c5a711d20bf26-Paper.pdf.

[150] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 110, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3373154. URL https://doi.org/10.1145/3351095.3373154.

[151] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33 (1):1–33, 2012.

[152] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Recommendation independence. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*,

pages 187–201. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/kamishima18a.html.

[153] Sai Srinivas Kancheti, Abbavaram Gowtham Reddy, Vineeth N Balasubramanian, and Amit Sharma. Matching learned causal effects of neural networks with domain priors. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10676–10696. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/kancheti22a.html.

[154] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and P. M. Krafft. Toward situated interventions for algorithmic equity: Lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 45–55, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372874. URL https://doi.org/10.1145/3351095.3372874.

[155] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kearns18a.html.

[156] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 154–165, 2021.

[157] Mohammad Mahdi Khalili, Xueru Zhang, and Mahed Abroshan. Fair sequential selection using supervised learning models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 28144–28155. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/ed277964a8959e72a0d987e598dfbe72-Paper.pdf.

[158] Fereshte Khani and Percy Liang. Removing spurious features can hurt accuracy and affect groups disproportionately. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 196–205, 2021.

[159] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2017/file/f5f8590cd58a54e94377e6ae2eded4d9-Paper.pdf.

[160] Niki Kilbertus, Adria Gascon, Matt Kusner, Michael Veale, Krishna Gummadi, and Adrian Weller. Blind justice: Fairness with encrypted sensitive attributes. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2630–2639. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kilbertus18a.html.

[161] Dongha Kim, Kunwoong Kim, Insung Kong, Ilsang Ohn, and Yongdai Kim. Learning fair representation with a parametric integral probability metric. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11074–11101. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/kim22b.html.

[162] Michael Kim, Omer Reingold, and Guy Rothblum. Fairness through computationally-bounded awareness. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2018/file/c8dfece5cc68249206e4690fc4737a8d-Paper.pdf.

[163] Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 247–254, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314287. URL https://doi.org/10.1145/3306618.3314287.

[164] Michael P. Kim, Aleksandra Korolova, Guy N. Rothblum, and Gal Yona. Preference-informed fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 546, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3373155. URL https://doi.org/10.1145/3351095.3373155.

[165] Pauline T. Kim. Race-aware algorithms: Fairness, nondiscrimination and affirmative action. *California Law Review*, 110, 2022.

[166] Junpei Komiyama, Akiko Takeda, Junya Honda, and Hajime Shimao. Nonconvex optimization for regression with fairness constraints. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2737–2746. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/komiyama18a.html.

[167] Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J Franklin, and Ken Goldberg. Activeclean: Interactive data cleaning for statistical modeling. *Proceedings of the VLDB Endowment*, 9(12):948–959, 2016.

[168] Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda G"urses. Pots: Protective optimization technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 177–188, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372853. URL https://doi.org/10.1145/3351095.3372853.

[169] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

[170] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 728–740. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2020/file/07fc15c9d169ee48573edd749d25945d-Paper.pdf.

[171] Alex Lamy, Ziyuan Zhong, Aditya K Menon, and Nakul Verma. Noise-tolerant fair classification. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/8d5e957f297893487bd98fa830fa6413-Paper.pdf.

[172] David Lehr and Paul Ohm. Playing with the data: what legal scholars should learn about machine learning. *UCDL Rev.*, 51:653, 2017.

[173] Klas Leino, Matt Fredrikson, Emily Black, Shayak Sen, and Anupam Datta. Feature-wise bias amplification. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1ecm2C9K7.

[174] Liu Leqi, Adarsh Prasad, and Pradeep K Ravikumar. On human-aligned risk minimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2019/file/cd6b73b67c77edeaff94e24b961119dd-Paper.pdf.

[175] Weiwen Leung, Zheng Zhang, Daviti Jibuti, Jinhao Zhao, Maximilian Klein, Casey Pierce, Lionel Robert, and Haiyi Zhu. Race, gender and beauty: The effect of information provision on online hiring biases. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–11, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376874. URL https://doi.org/10.1145/3313831.3376874.

[176] Mingchen Li, Xuechen Zhang, Christos Thrampoulidis, Jiasi Chen, and Samet Oymak. Autobalance: Optimized loss functions for imbalanced data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3163–3177. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/191f8f858acda435ae0daf994e2a72c2-Paper.pdf.

[177] Mingchen Li, Xuechen Zhang, Christos Thrampoulidis, Jiasi Chen, and Samet Oymak. Autobalance: Optimized loss functions for imbalanced data. *Advances in Neural Information Processing Systems*, 34:3163–3177, 2021.

[178] Nianyun Li, Naman Goel, and Elliott Ash. Data-centric factors in algorithmic fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 396–410, 2022.

[179] Peizhao Li, Yifei Wang, Han Zhao, Pengyu Hong, and Hongfu Liu. On dyadic fairness: Exploring and mitigating bias in graph connections. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xgGS6PmzNq6.

[180] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=K5YasWXZT3O.

[181] Yunyi Li, Maria De-Arteaga, and Maytal Saar-Tsechansky. When more data lead us astray: Active data acquisition in the presence of label bias. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 133–146, 2022.

[182] Zhuoyan Li, Zhuoran Lu, and Ming Yin. Towards better detection of biased language with scarce, noisy, and biased annotations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 411–423, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534142. URL https://doi.org/10.1145/3514094.3534142.

[183] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6565–6576. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/liang21a.html.

[184] David Liu, Zohair Shafi, William Fleisher, Tina Eliassi-Rad, and Scott Alfeld. Rawlsnet: Altering bayesian networks to encode rawlsian fair equality of opportunity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 745–755, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462618. URL https://doi.org/10.1145/3461702.3462618.

[185] Yang Liu and Jialu Wang. Can less be more? when increasing-to-balancing label noise rates considered beneficial. *Advances in Neural Information Processing Systems*, 34:17467–17479, 2021.

[186] Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. Too relaxed to be fair. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6360–6369. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/lohaus20a.html.

[187] Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2017/file/48ab2f9b45957ab574cf005eb8a76760-Paper.pdf.

[188] Lydia R. Lucchesi, Petra M. Kuhnert, Jenny L. Davis, and Lexing Xie. Smallset timelines: A visual representation of data preprocessing decisions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1136–1153, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533175. URL https://doi.org/10.1145/3531146.3533175.

[189] Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

[190] Kristian Lum and Tarak Shah. Measures of fairness for new york city's supervised release risk assessment tool. *Human Rights Data Analytics Group*, page 21, 2019.

[191] Kristian Lum, Chesa Boudin, and Megan Price. The impact of overbooking on a pre-trial risk assessment tool. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 482–491, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372846. URL https://doi.org/10.1145/3351095.3372846.

[192] Kristian Lum, Yunfeng Zhang, and Amanda Bower. De-biasing "bias" measurement. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 379–389, 2022.

[193] Jiaqi Ma, Junwei Deng, and Qiaozhu Mei. Subgroup generalization and fairness of graph neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1048–1061. Curran

Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/08425b881bcde94a383cd258cea331be-Paper.pdf.

[194] Xinsong Ma, Zekai Wang, and Weiwei Liu. On the tradeoff between robustness and fairness. In *Advances in Neural Information Processing Systems*, 2022.

[195] Zilin Ma and Krzysztof Z. Gajos. Not just a preference: Reducing biased decision-making on dating websites. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517587. URL https://doi.org/10.1145/3491102.3517587.

[196] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[197] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3384–3393. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/madras18a.html.

[198] David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, 31, 2018.

[199] Subha Maity, Debarghya Mukherjee, Mikhail Yurochkin, and Yuekai Sun. Does enforcing fairness mitigate biases caused by subpopulation shift? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25773–25784. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/d800149d2f947ad4d64f34668f8b20f6-Paper.pdf.

[200] Subha Maity, Songkai Xue, Mikhail Yurochkin, and Yuekai Sun. Statistical inference for individual fairness. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=z9k8BWL-%5F2u.

[201] Vidushi Marda and Shivangi Narayan. Data in new delhi's predictive policing system. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 317–324, 2020.

[202] Jeanna Neefe Matthews, Graham Northup, Isabella Grasso, Stephen Lorenz, Marzieh Babaeianjelodar, Hunter Bashaw, Sumona Mondal, Abigail Matthews, Mariama Njie, and Jessica Goldthwaite. When trusted black boxes don't agree: Incentivizing iterative improvement and accountability in critical software systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 102–108, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375807. URL https://doi.org/10.1145/3375627.3375807.

[203] Bryce McLaughlin, Jann Spiess, and Talia Gillis. On the fairness of machine-assisted human decisions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 890, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533152. URL https://doi.org/10.1145/3531146.3533152.

[204] Anay Mehrotra and L. Elisa Celis. Mitigating bias in set selection with noisy protected attributes. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 237–248, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445887. URL https://doi.org/10.1145/3442188.3445887.

[205] Blossom Metevier, Stephen Giguere, Sarah Brockman, Ari Kobren, Yuriy Brun, Emma Brunskill, and Philip S. Thomas. Offline contextual bandits with high probability fairness guarantees. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/d69768b3da745b77e82cdbddcc8bac98-Paper.pdf.

[206] Anna Meyer, Aws Albarghouthi, and Loris D' Antoni. Certifying robustness to programmable data bias in decision trees. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26276–26288. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/dcf531edc9b229acfe0f4b87e1e278dd-Paper.pdf.

[207] Vishwali Mhasawade and Rumi Chunara. Causal multi-level fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 784–794, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462587. URL https://doi.org/10.1145/3461702.3462587.

[208] Milagros Miceli, Martin Schuessler, and Tianling Yang. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *CSCW*, 2020.

[209] Filip Michalsky. Fairness criteria for face recognition applications. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 527–528, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314308. URL https://doi.org/10.1145/3306618.3314308.

[210] Alan Mishler. Modeling risk and achieving algorithmic fairness using potential outcomes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 555–556, 2019.

[211] Alan Mishler and Edward H. Kennedy. Fade: Fair double ensemble learning for observable and counterfactual outcomes. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1053, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533167. URL https://doi.org/10.1145/3531146.3533167.

[212] Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 386–400, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445902. URL https://doi.org/10.1145/3442188.3445902.

[213] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*,

volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/415185ea244ea2b2bedeb0449b926802-Paper.pdf.

[214] Ece Çiğdem Mutlu, Niloofar Yousefi, and Ozlem Ozmen Garibay. Contrastive counterfactual fairness in algorithmic decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 499–507, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534143. URL https://doi.org/10.1145/3514094.3534143.

[215] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4674–4682. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/nabi19a.html.

[216] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P. Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 466–477, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445910. URL https://doi.org/10.1145/3442188.3445910.

[217] Preetam Nandy, Cyrus DiCiccio, Divya Venugopalan, Heloise Logan, Kinjal Basu, and Noureddine El Karoui. Achieving fairness via post-processing in web-scale recommender systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 715–725, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533136. URL https://doi.org/10.1145/3531146.3533136.

[218] Harikrishna Narasimhan, Andrew Cotter, Yichen Zhou, Serena Wang, and Wenshuo Guo. Approximate heavily-constrained learning with lagrange multiplier models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8693–8703. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/62db9e3397c76207a687c360e0243317-Paper.pdf.

[219] Milad Nasr and Michael Carl Tschantz. Bidding strategies with gender nondiscrimination constraints for online ad auctions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 337–347, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3375783. URL https://doi.org/10.1145/3351095.3375783.

[220] Maria Conchita A. Navarro and Orit Shaer. Re-imagining systems in the realm of immigration in higher education through participatory design. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW'22 Companion, page 76–79, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391900. doi: 10.1145/3500868.3559457. URL https://doi.org/10.1145/3500868.3559457.

[221] Aviv Navon, Aviv Shamsian, Ethan Fetaya, and Gal Chechik. Learning the pareto front with hypernetworks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=NjF772F4ZZR.

[222] Sarah Nikkhah, Akash Uday Rode, Priyanjali Mittal, Neha K. Kulkarni, Salonee Nadkarni, Emily L. Mueller, and Andrew D. Miller. "i feel like i need to split myself in half": Using role theory to design for parents as caregiving teams in the children's hospital. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW'22 Companion, page 115–120, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391900. doi: 10.1145/3500868.3559466. URL https://doi.org/10.1145/3500868.3559466.

[223] Alejandro Noriega-Campero, Michiel A. Bakker, Bernardo Garcia-Bulle, and Alex 'Sandy' Pentland. Active fairness in algorithmic decision making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 77–83, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314277. URL https://doi.org/10.1145/3306618.3314277.

[224] L Oakden-Rayner, J Dunnmon, G Carniero, and C Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. arxiv, 2019.

[225] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342. URL https://www.science.org/doi/abs/10.1126/science.aax2342.

[226] Simon Olofsson, Marc Deisenroth, and Ruth Misener. Design of experiments for model discrimination hybridising analytical and data-driven approaches. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3908–3917. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/olofsson18a.html.

[227] Luca Oneto, Michele Doninini, Amon Elders, and Massimiliano Pontil. Taking advantage of multitask learning for fair classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 227–237, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314255. URL https://doi.org/10.1145/3306618.3314255.

[228] Luca Oneto, Michele Donini, Giulia Luise, Carlo Ciliberto, Andreas Maurer, and Massimiliano Pontil. Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15360–15370. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/af9c0e0c1dee63e5acad8b7ed1a5be96-Paper.pdf.

[229] Jaspar Pahl, Ines Rieger, Anna M"oller, Thomas Wittenberg, and Ute Schmid. Female, white, 27? bias evaluation on data and algorithms for affect recognition in faces. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 973–987, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533159. URL https://doi.org/10.1145/3531146.3533159.

[230] Akshat Pandey and Aylin Caliskan. Disparate impact of artificial intelligence bias in ridehailing economy's price discrimination algorithms. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 822–833, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462561. URL https://doi.org/10.1145/3461702.3462561.

[231] Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 446–457, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372843. URL https://doi.org/10.1145/3351095.3372843.

[232] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. Designing fair ai in human resource management: Understanding tensions surrounding algorithmic evaluation and envisioning stakeholder-centered solutions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517672. URL https://doi.org/10.1145/3491102.3517672.

[233] Joon Sung Park, Michael S. Bernstein, Robin N. Brewer, Ece Kamar, and Meredith Ringel Morris. Understanding the representation and representativeness of age in ai data sets. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 834–842, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462590. URL https://doi.org/10.1145/3461702.3462590.

[234] Joon Sung Park, Michael S Bernstein, Robin N Brewer, Ece Kamar, and Meredith Ringel Morris. Understanding the representation and representativeness of age in ai data sets. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 834–842, 2021.

[235] Joon Sung Park, Danielle Bragg, Ece Kamar, and Meredith Ringel Morris. Designing an online infrastructure for collecting ai data from people with disabilities. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 52–63, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445870. URL https://doi.org/10.1145/3442188.3445870.

[236] Samir Passi and Solon Barocas. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 39–48, 2019.

[237] Samir Passi and Phoebe Sengers. Making data science systems work. *Big Data & Society*, 7(2):2053951720939605, 2020.

[238] Ioannis Pastaltzidis, Nikolaos Dimitriou, Katherine Quezada-Tavarez, Stergios Aidinlis, Thomas Marquenie, Agata Gurzawska, and Dimitrios Tzovaras. Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2302–2314, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3534644. URL https://doi.org/10.1145/3531146.3534644.

[239] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

[240] Valerio Perrone, Michele Donini, Muhammad Bilal Zafar, Robin Schmucker, Krishnaram Kenthapadi, and Cédric Archambeau. Fair bayesian optimization. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 854–863, 2021.

[241] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[242] Felix Petersen, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin. Post-processing for individual fairness. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25944–25955. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2021/file/d9fea4ca7e4a74c318ec27c1deb0796c-Paper.pdf.

[243] David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 189–199, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372872. URL https://doi.org/10.1145/3351095.3372872.

[244] Novi Quadrianto and Viktoriia Sharmanska. Recycling privileged learning and distribution matching for fairness. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2017/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

[245] Edward Raff, Jared Sylvester, and Steven Mills. Fair forests: Regularized tree induction to minimize model bias. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 243–250, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278742. URL https://doi.org/10.1145/3278721.3278742.

[246] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. The fallacy of ai functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 959–972, 2022.

[247] Lydia Reader, Pegah Nokhiz, Cathleen Power, Neal Patwari, Suresh Venkatasubramanian, and Sorelle Friedler. Models for understanding and quantifying feedback in societal systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1765–1775, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533230. URL https://doi.org/10.1145/3531146.3533230.

[248] Brianna Richardson, Jean Garcia-Gathright, Samuel F Way, Jennifer Thom, and Henriette Cramer. Towards fairness in practice: A practitioner-oriented rubric for evaluating fair ml toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[249] Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. Modeling assumptions clash with the real world: Transparency, equity, and community challenges for student assignment algorithms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445748. URL https://doi.org/10.1145/3411764.3445748.

[250] Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. Sample selection for fair and robust training. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 815–827. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2021/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf.

[251] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YNnpaAKeCfx.

[252] Yaniv Romano, Stephen Bates, and Emmanuel Candes. Achieving equalized odds by resampling sensitive attributes. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 361–371. Curran Associates, Inc.,

2020. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2020/file/03593ce517feac573fdaafa6dcedef61-Paper.pdf.

[253] BRIEFING ROOM. Executive order on advancing racial equity and support for underserved communities through the federal government. 2021.

[254] Jonathan Roth, Guillaume Saint-Jacques, and YinYin Yu. An outcome test of discrimination for ranked lists. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 350–356, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533102. URL https://doi.org/10.1145/3531146.3533102.

[255] Anian Ruoss, Mislav Balunovic, Marc Fischer, and Martin Vechev. Learning certified individually fair representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7584–7596. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/55d491cf951b1b920900684d71419282-Paper.pdf.

[256] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: Integrating different counterfactual assumptions in fairness. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2017/file/1271a7029c9df08643b631b02cf9e116-Paper.pdf.

[257] Pratik S. Sachdeva, Renata Barreto, Claudia von Vacano, and Chris J. Kennedy. Assessing annotator identity sensitivity via item response theory: A case study in a hate speech corpus. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1585–1603, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533216. URL https://doi.org/10.1145/3531146.3533216.

[258] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

[259] Tiago Salvador, Stephanie Cairns, Vikram Voleti, Noah Marshall, and Adam M Oberman. Faircal: Fairness calibration for face verification. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nRj0NcmSuxb.

[260] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[261] Robin Schmucker, Michele Donini, Valerio Perrone, and Cédric Archambeau. Multi-objective multi-fidelity hyperparameter optimization with application to fairness. In *NeurIPS 2020 Workshop on Meta-learning*, 2020. URL https://www.amazon.science/publications/multi-objective-multi-fidelity-hyperparameter-optimization-with-application-to-fairness.

[262] Shilad Sen, Margaret E. Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko, Samuel Naden, Jesse Russell, Zixiao (Ken) Wang, and Brent Hecht. Turkers, scholars, "arafat" and "peace": Cultural communities and algorithmic gold standards. CSCW, 2015.

[263] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. ""the human body is a black box"": Supporting clinical decision-making with deep learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 99–109, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372827. URL https://doi.org/10.1145/3351095.3372827.

[264] Abhin Shah, Yuheng Bu, Joshua K Lee, Subhro Das, Rameswar Panda, Prasanna Sattigeri, and Gregory W Wornell. Selective regression under fairness criteria. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19598–19615. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/shah22a.html.

[265] Kulin Shah, Pooja Gupta, Amit Deshpande, and Chiranjib Bhattacharyya. Rawlsian fair adaptation of deep learning classifiers. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 936–945, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462592. URL https://doi.org/10.1145/3461702.3462592.

[266] Amr Sharaf, Hal Daume III, and Renkun Ni. Promoting fairness in learned models by learning to active learn under parity constraints. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2149–2156, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3534632. URL https://doi.org/10.1145/3531146.3534632.

[267] Saeed Sharifi-Malvajerdi, Michael Kearns, and Aaron Roth. Average individual fairness: Algorithms, generalization and experiments. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2019/file/0e1feae55e360ff05fef58199b3fa521-Paper.pdf.

[268] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 166–172, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375812. URL https://doi.org/10.1145/3375627.3375812.

[269] Shubham Sharma, Yunfeng Zhang, Jesús M Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R Varshney. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 358–364, 2020.

[270] Shubham Sharma, Alan H. Gee, David Paydarfar, and Joydeep Ghosh. Fair-n: Fair and robust neural networks for structured data. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 946–955, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462559. URL https://doi.org/10.1145/3461702.3462559.

[271] Shubhanshu Shekhar, Greg Fields, Mohammad Ghavamzadeh, and Tara Javidi. Adaptive sampling for minimax fair classification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24535–24544. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2021/file/cd7c230fc5deb01ff5f7b1be1acef9cf-Paper.pdf.

[272] Changjian Shui, Qi Chen, Jiaqi Li, Boyu Wang, and Christian Gagné. Fair representation learning through implicit path alignment. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20156–20175. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/shui22a.html.

[273] Nian Si, Karthyek Murthy, Jose Blanchet, and Viet Anh Nguyen. Testing group fairness via optimal transport projections. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9649–9659. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/si21a.html.

[274] Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. Getfair: Generalized fairness tuning of classification models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 289–299, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533094. URL https://doi.org/10.1145/3531146.3533094.

[275] Harvineet Singh. Fair, robust, and data-efficient machine learning in healthcare. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 914, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3539552. URL https://doi.org/10.1145/3514094.3539552.

[276] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 3–13, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445865. URL https://doi.org/10.1145/3442188.3445865.

[277] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 3–13, 2021.

[278] Dylan Slack, Sorelle A. Friedler, and Emile Givental. Fairness warnings and fair-maml: Learning fairly with minimal data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 200–209, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372839. URL https://doi.org/10.1145/3351095.3372839.

[279] Gavin Smith, Roberto Mansilla, and James Goulding. Model class reliance for random forests. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22305–22315. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2020/file/fd512441a1a791770a6fa573d688bff5-Paper.pdf.

[280] Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization*, pages 1–9. 2021.

[281] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxen, Angeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. Towards intersectional feminist and participatory ml: A case study in supporting feminicide counterdata collection. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 667–678, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533132. URL https://doi.org/10.1145/3531146.3533132.

[282] Vinith M Suriyakumar, Nicolas Papernot, Anna Goldenberg, and Marzyeh Ghassemi. Chasing your long tails: Differentially private prediction in health care settings. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 723–734, 2021.

[283] Chris Sweeney and Maryam Najafian. Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 359–368, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372837. URL https://doi.org/10.1145/3351095.3372837.

[284] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 305–311, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314270. URL https://doi.org/10.1145/3306618.3314270.

[285] Maia Szalavitz. The pain was unbearable. so why did doctors turn her away. *Wired*, 2021.

[286] Yi Chern Tan and L. Elisa Celis. Assessing social and intersectional biases in contextualized word representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/201d546992726352471cfea6b0df0a48-Paper.pdf.

[287] Bahar Taskesen, Jose Blanchet, Daniel Kuhn, and Viet Anh Nguyen. A statistical test for probabilistic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 648–665, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445927. URL https://doi.org/10.1145/3442188.3445927.

[288] Oliver Thomas, Miri Zilka, Adrian Weller, and Novi Quadrianto. An algorithmic framework for positive action. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385534. doi: 10.1145/3465416.3483303. URL https://doi.org/10.1145/3465416.3483303.

[289] Frederik Tr"auble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Sch"olkopf, and Stefan Bauer. On disentangled representations learned from correlated data. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10401–10412. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/trauble21a.html.

[290] Yao-Hung Hubert Tsai, Tianqin Li, Martin Q. Ma, Han Zhao, Kun Zhang, Louis-Philippe Morency, and Ruslan Salakhutdinov. Conditional contrastive learning with kernel. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=AAJLBoGt0XM.

[291] Nicolas Usunier, Virginie Do, and Elvis Dohmatob. Fast online ranking with fairness of exposure. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2157–2167, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi:

10.1145/3531146.3534633. URL https://doi.org/10.1145/3531146.3534633.

[292] Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela van der Schaar. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, 34:22221–22233, 2021.

[293] Alexander Vargo, Fan Zhang, Mikhail Yurochkin, and Yuekai Sun. Individually fair gradient boosting. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=JBAa9we1AL.

[294] Praveen Venkatesh, Sanghamitra Dutta, Neil Mehta, and Pulkit Grover. Can information flows suggest targets for interventions in neural circuits? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3149–3162. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/18de4beb01f6a17b6e1dfb9813ba6045-Paper.pdf.

[295] Ada Wan. Fairness in representation for multilingual NLP: Insights from controlled experiments on conditional language modeling. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=-llS6TiOew.

[296] Angelina Wang and Olga Russakovsky. Directional bias amplification. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10882–10893. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/wang21t.html.

[297] Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. Measuring representational harms in image captioning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 324–335, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533099. URL https://doi.org/10.1145/3531146.3533099.

[298] Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *Available at SSRN*, 2022.

[299] Hao Wang, Berk Ustun, and Flavio Calmon. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6618–6627. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/wang19l.html.

[300] Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 526–536, 2021.

[301] Jialu Wang, Xin Eric Wang, and Yang Liu. Understanding instance-level impact of fairness constraints. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23114–23130. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wang22ac.html.

[302] Xiuling Wang and Wendy Hui Wang. Providing item-side individual fairness for deep recommender systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 117–127, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533079. URL https://doi.org/10.1145/3531146.3533079.

[303] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020.

[304] Elizabeth Anne Watkins. Took a pic and got declined, vexed and perplexed: Facial recognition in algorithmic management. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, CSCW '20 Companion, page 177–182, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380591. doi: 10.1145/3406865.3418383. URL https://doi.org/10.1145/3406865.3418383.

[305] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. *arXiv preprint arXiv:2110.12088*, 2021.

[306] James Wexler, Mahima Pushkarna, Sara Robinson, Tolga Bolukbasi, and Andrew Zaldivar. Probing ml models for fairness with the what-if tool and shap: Hands-on tutorial. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 705, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3375662. URL https://doi.org/10.1145/3351095.3375662.

[307] Michael Wick, swetasudha panda, and Jean-Baptiste Tristan. Unlocking fairness: a trade-off revisited. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/373e4c5d8edfa8b74fd4b6791d0cf6dc-Paper.pdf.

[308] Ziwei Wu and Jingrui He. Fairness-aware model-agnostic positive and unlabeled learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1698–1708, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533225. URL https://doi.org/10.1145/3531146.3533225.

[309] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8cb22bdd0b7ba1ab13d742e22eed8da2-Paper.pdf.

[310] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11492–11501. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/xu21b.html.

[311] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning*, pages 11492–11501. PMLR, 2021.

[312] Xinyi Xu, Lingjuan Lyu, Xingjun Ma, Chenglin Miao, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Gradient driven rewards to guarantee fairness in collaborative machine learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in*

Neural Information Processing Systems, volume 34, pages 16104–16117. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/8682cc30db9c025ecd3fee433f8ab54c-Paper.pdf.

[313] Yilun Xu, Hao He, Tianxiao Shen, and Tommi S. Jaakkola. Controlling directions orthogonal to a classifier. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=DIjCrlsu6Z.

[314] Tom Yan and Chicheng Zhang. Active fairness auditing. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 24929–24962. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/yan22c.html.

[315] Eddie Yang and Margaret E. Roberts. Censorship of online encyclopedias: Implications for nlp models. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 537–548, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445916. URL https://doi.org/10.1145/3442188.3445916.

[316] Forest Yang, Mouhamadou Cisse, and Sanmi Koyejo. Fairness with overlapping groups; a probabilistic perspective. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4067–4078. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2020/file/29c0605a3bab4229e46723f89cf59d83-Paper.pdf.

[317] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 547–558, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3375709. URL https://doi.org/10.1145/3351095.3375709.

[318] Wanqian Yang, Lars Lorch, Moritz Graule, Himabindu Lakkaraju, and Finale Doshi-Velez. Incorporating interpretable output constraints in bayesian neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12721–12731. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2020/file/95c7dfc5538e1ce71301cf92a9a96bd0-Paper.pdf.

[319] Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, and Xiang Ren. Refining language models with compositional explanations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8954–8967. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/4b26dc4663ccf960c8538d595d0a1d3a-Paper.pdf.

[320] Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/e6384711491713d29bc63fc5eeb5ba4f-Paper.pdf.

[321] Samuel Yeom, Anupam Datta, and Matt Fredrikson. Hunting for discriminatory proxies in linear regression models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/6cd9313ed34ef58bad3fdd504355e72c-Paper.pdf.

[322] Samuel Yeom, Anupam Datta, and Matt Fredrikson. Hunting for discriminatory proxies in linear regression models. In *Advances in Neural Information Processing Systems*, pages 4568–4578, 2018.

[323] William Yik, Limnanthes Serafini, Timothy Lindsey, and George D Montañez. Identifying bias in data using two-distribution hypothesis tests. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 831–844, 2022.

[324] Gal Yona and Guy Rothblum. Probably approximately metric-fair learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5680–5688. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/yona18a.html.

[325] Mikhail Yurochkin and Yuekai Sun. Sensei: Sensitive set invariance for enforcing individual fairness. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=DktZb97%5FFx.

[326] Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ml models with sensitive subspace robustness. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1gdkxHFDH.

[327] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2017/file/82161242827b703e6acf9c726942a1e4-Paper.pdf.

[328] Angie Zhang, Alexander Boltz, Chun Wei Wang, and Min Kyung Lee. Algorithmic management reimagined for workers and by workers: Centering worker well-being in gig work. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3501866. URL https://doi.org/10.1145/3491102.3501866.

[329] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278779. URL https://doi.org/10.1145/3278721.3278779.

[330] Hongjing Zhang and Ian Davidson. Towards fair deep anomaly detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 138–148, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445878. URL https://doi.org/10.1145/3442188.3445878.

[331] Junzhe Zhang and Elias Bareinboim. Equality of opportunity in classification: A causal approach. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

URL https://proceedings.neurips.cc/paper%5Ffiles/paper/2018/file/ff1418e8cc993fe8abcfe3ce2003e5c5-Paper.pdf.

[332] Marilyn Zhang. Affirmative algorithms: Relational equality as algorithmic fairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 495–507, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533115. URL https://doi.org/10.1145/3531146.3533115.

[333] Yiliang Zhang and Qi Long. Assessing fairness in the presence of missing data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16007–16019. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/85dca1d270f7f9aef00c9d372f114482-Paper.pdf.

[334] Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. Conditional learning of fair representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hkekl0NFPr.

[335] Zhaowei Zhu, Tianyi Luo, and Yang Liu. The rich get richer: Disparate impact of semi-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=DXPftn5kjQK.