Practitioners and Bias in Machine Learning: A Study

ROBERT CINCA and ENRICO COSTANZA, University College London, London, United Kingdom of Great Britain and Northern Ireland
MIRCO MUSOLESI, University College London, London, United Kingdom of Great Britain and Northern Ireland and University of Bologna, Bologna, Italy

The increasing adoption of machine learning (ML) raises ethical concerns, particularly regarding bias. This study explores how ML practitioners with limited experience in bias understand and apply bias definitions, detection measures, and mitigation methods. Through a take-home task, exercises, and interviews with 22 participants, we identified five key themes: sources of bias, selecting bias metrics, detecting bias, mitigating bias, and ethical considerations. Participants faced unresolved conflicts, such as applying fairness definitions in practice, selecting context-dependent bias metrics, addressing real-world biases, balancing model performance with bias mitigation, and relying on personal perspectives over data-driven metrics. While bias mitigation techniques helped identify biases in two datasets, participants could not fully eliminate bias, citing the oversimplification of complex processes into models with limited variables. We propose designing bias detection tools that encourage practitioners to focus on the underlying assumptions and integrating bias concepts into ML practices, such as using a harmonic mean-based approach, akin to the F1 score, to balance bias and accuracy.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Machine learning**; • **Social and professional topics** → Computing education;

Additional Key Words and Phrases: ML Bias, Operationalizing Bias, machine learning, machine learning practitioners, interview study

ACM Reference format:
Robert Cinca, Enrico Costanza, and Mirco Musolesi. 2025. Practitioners and Bias in Machine Learning: A Study. *ACM Trans. Interact. Intell. Syst.* 15, 2, Article 12 (June 2025), 28 pages.
https://doi.org/10.1145/3733838

1 Introduction

Machine learning (ML) is being applied to an increasing number of diverse applications, such as loan applications, diagnosing disease, crime prevention, facial recognition, and language translation [51]. However, the expanding application of ML across various domains raises serious ethical

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/R513143/1 for the University College London Interaction Centre (UCLIC). Our study was approved by the UCLIC Ethics Committee (UCLIC_2022_004_Costanza).
Authors' Contact Information: Robert Cinca (corresponding author), University College London, London, United Kingdom of Great Britain and Northern Ireland; e-mail: robert.cinca.14@ucl.ac.uk; Enrico Costanza, University College London, London, United Kingdom of Great Britain and Northern Ireland; e-mail: e.costanza@ucl.ac.uk; Mirco Musolesi, University College London, London, United Kingdom of Great Britain and Northern Ireland and University of Bologna, Bologna, Italy; e-mail: m.musolesi@ucl.ac.uk.

ACM 2160-6463/2025/6-ART12
https://doi.org/10.1145/3733838

concerns and challenges related to bias. Certain fields, such as healthcare [75] and crime prevention [1], are particularly at risk of discrimination based on protected characteristics. Examples of biased outcomes include predicting a criminal's re-offending probability while discriminating based on race [1], how search engine results reinforce racism [57], and "automatic gender recognition algorithms" misgendering individuals [37]. Therefore, it is important to identify how ML users understand and apply bias detection and mitigation methods when modeling ML. This is a timely challenge for the HCI community, as the misapplication of ML might lead to detrimental consequences for disadvantaged groups of individuals.

We present a qualitative study aimed at understanding how novices in bias[1] can operationalize bias definitions and apply mitigation methods. The need for bias mitigation is increasingly emphasized by evolving regulatory frameworks like the EU AI Act [20] and heightened expectations for fairness in ML models. While prior research has introduced numerous fairness definitions [49, 63, 70], bias metrics [26, 35, 42], and supporting tools [8, 62, 65], the mere availability of these resources does not address bias in ML. Their effective use requires practitioners to make complex decisions. For example, previous studies have highlighted challenges such as the need for domain-specific training, enhanced communication between stakeholders, and improved organizational practices to mitigate bias [36, 47, 73, 74]. Building on this, our study provides practical insights by engaging participants with 10 fairness definitions, examining their strategies for detecting and mitigating bias across two datasets: COMPAS [1] and German Credit [25].

Twenty-two participants interested in learning and applying bias detection and mitigation techniques took part in our study. They were introduced to bias concepts through a take-home task designed to be completed within one and a half hours. The task, iteratively designed and centered around a popular browser-based coding environment, exposed participants to various methods for detecting and mitigating bias through a series of exercises. Participants demonstrated their understanding through these exercises and in a subsequent 30-minute semi-structured interview, where they elaborated on their responses and reflected on bias metrics and fairness definitions. The researchers then analyzed participant responses from both the take-home task and interviews using thematic analysis techniques to identify the challenges in operationalizing bias. Specifically, we address the following research question: how do novices in bias understand and apply a range of bias definitions, measures to detect it, and methods to mitigate it?

A thematic analysis of participant interviews and exercises identified five key themes: (1) sources of bias; (2) employing and selecting bias metrics; (3) detecting bias; (4) mitigating bias; and (5) ethical considerations. These findings revealed that participants encountered unresolved conflicts when attempting to operationalize bias, including: (1) the application of fairness definitions in practical settings, such as a loss of granularity; (2) the challenge of selecting appropriate bias metrics, with context-dependency being a key consideration; (3) addressing real-world biases; (4) balancing model performance with bias mitigation; and (5) relying on personal goals, opinions, and stereotypes rather than bias metrics and the underlying data. Despite the availability of bias mitigation techniques, participants were unable to fully eliminate bias from the COMPAS and German Credit datasets. They attributed the difficulty in mitigating bias to the oversimplification of complex real-world processes into models with a limited set of variables. Nonetheless, participants were able to reflect on effective model-building practices for bias mitigation, discussing how they would apply these strategies within their own domains while considering the sources of bias.

We offer three key contributions. Answering the research question, our *first* contribution is a series of findings that illustrate the various challenges that novices in bias experience when

[1]We refer to *novices in bias* as those interested in implementing bias detection and mitigation methods (e.g., in their own domains) and are familiar with ML and Python programming.

operationalizing a range of measures to detect and mitigate bias. *Second*, despite misconceptions and challenges that persisted, our findings indicate that participants showed an ability and creativity toward tackling bias, often proposing their own methods for mitigating biases in their work. Completing the study improved their ability to identify bias methodically and quantitatively, demonstrating the benefits of a short-term learning program. The *third* contribution is a series of implications for operationalizing bias. This involves designing bias detection tools that encourage practitioners to focus on underlying ethical principles rather than "gaming the system" by selecting bias metrics that merely make an ML model seem less biased. Another implication is to leverage the overlaps between bias and ML concepts to help integrate bias mitigation into ML practices. For instance, our findings suggest the need for a harmonic mean-based approach, similar to the F1 score, adapted from ML practices, which could be used to balance bias and accuracy.

2 Related Work

Relevant research exists on assessing ML practitioners' fairness needs within the context of fairness tools and their practical limitations. We review this work below, while also providing a background on fairness definitions and metrics employed by the study.

2.1 Assessing ML Practitioners' Fairness Needs

Holstein et al. [36] investigated the challenges and needs of industry practitioners in developing fairer ML systems. Key findings include the influence of human biases in data labeling, the impact datasets have on fairness, and the lack of fairness metrics and automated tests. The study emphasizes the need for domain-specific resources and technical tools to support fairness throughout the ML development pipeline [36]. Work by Madaio et al. [47] investigated how ML practitioners identify, assess, and mitigate bias. Through semi-structured interviews and workshops with 33 AI practitioners from various technology companies, the study highlights challenges in choosing performance metrics, identifying relevant stakeholders, and collecting suitable datasets. Varanasi and Goyal [73] conducted interviews with 23 ML practitioners to understand their difficulties in creating fair ML systems within their workplace. Key challenges include the practitioners' lack of knowledge on ML fairness principles and the conflict between different fairness definitions, given that prior work has found inherent tradeoffs between satisfying various types of fairness [9] and that there is no single definition of fairness [63].

In addition, prior work by Veale et al. [74] investigated the current approach of algorithmic fairness for ML practitioners in public sector decision-making, interviewing 27 public servants. They found key issues in current model-building practices which impact fairness, such as an over-reliance on summary statistics which masks underlying issues, and how domain experts should be modifying their ML models to account for changes in the data over time [74]. A significant amount of prior work in this area also suggests that practitioners need improved organizational processes and engagement with stakeholders to effectively conduct fairness evaluations [36, 47, 73, 74]. These studies relied on interviews, workshops, and surveys about the needs of ML practitioners to develop ML systems. Instead, our study focuses on how ML practitioners operationalize bias definitions and metrics through a series of practical exercises.

2.2 Bias Mitigation Tools and Methodologies

Regarding the detection and mitigation of unwanted algorithmic bias, several tools compile fairness metrics and definitions into bias analysis frameworks, including Aequitas [62], AI Fairness 360 [8], FairLearn [11], and Fairness Indicators [32]. These libraries allow practitioners to apply fairness definitions to their datasets and models, using the outputs to assess and mitigate bias. Additionally, other research focuses on mitigating bias from the outset by using unbiased datasets. Suggestions

for creating and documenting datasets include Factsheets [2], Datasheets for Datasets [31], and Model Cards for Model Reporting [52]. Some researchers have developed benchmark "bias-free" datasets, such as the Diversity-in-Faces dataset, which achieves statistical parity among different sensitive features [50]. While these tools and methodologies are a valuable step toward helping ML practitioners address bias in their models, their availability alone does not resolve the issue, as their effective application depends on practitioners making numerous informed decisions.

To understand how to improve adoption of these tools, Richardson et al. [59] investigated how ML practitioners use two of them: Fairness Indicators [32] and Aequitas [62]. The results revealed that fairness tools should support bias analysis across the ML model-building process and that they should allow for the customization of fairness and performance metrics. Similar research by Deng et al. [23] and Lee and Singh [43] supports these findings, with interviews and surveys indicating that fairness tools need to provide more context-specific guidance for ML practitioners throughout the entire development process. Lee and Singh's assessment of six fairness tools revealed a steep learning curve and insufficient guidance, making them challenging for practitioners unfamiliar with fairness literature. This conclusion was drawn from semi-structured interviews with fairness experts and surveys among a broader group of ML practitioners.

Balayn et al. [4] conducted a study with 30 participants that had practical experience with some of these tools. Using two of them (AI Fairness 360 [8] and FairLearn [11]) and subsequent interviews, the researchers identified that while these tools improve practices around algorithmic fairness, practitioners would apply metrics available through the tools and declare fairness was reached without reflecting on the appropriateness and limitations of the metrics, and without considering the tradeoff between accuracy and fairness. Our work builds on this research by identifying practical challenges that ML practitioners who are novices to bias face when detecting and mitigating biases, offering insights on how they operationalize bias, with implications for fairness tools.

2.3 Background on Fairness Definitions and Metrics

Defining, detecting, measuring, and mitigating bias in ML systems is a complex and ongoing area of research [6]. Mehrabi et al. identified 23 types of bias [49], while Srinivasan and Chandler categorized biases into 11 main types with several sub-types [67]. To address these biases, researchers have developed several techniques for measuring different types of biases, known as definitions of fairness. Our study implements some of the most popular fairness metrics and definitions as described by Mehrabi et al. [49].

The bias definitions in our study are divided into two groups: group-level fairness metrics and individual-level fairness metrics. The group-level metrics measure fairness by assessing if the metric's values are equal across different feature groups, such as male and female. The methods implemented in our study include equalized odds [35], equal opportunity [35], statistical parity [26], treatment equality [9], and fairness in relational domains [27]. Participants can compare results across different groups, such as males and females, using numerical implementations of these methods. The individual-level fairness metrics assess the effect of modifying an input feature value on the model's output, demonstrating fairness through awareness [26], counterfactual fairness [42], conditional statistical parity [21], and test fairness [16]. While the latter two can also be implemented at a group level, they are primarily demonstrated here through sensitivity analysis. This analysis shows participants how changing feature values, such as race from Asian to Hispanic, affects the model's output. It is based on the Prospector by Krause et al. [40], an interactive visual analytics system that ML practitioners can use to change the feature values and observe how the prediction responds. Additionally, fairness through unawareness [33] is considered satisfied when a model does not use any protected characteristics or their proxies. This is visually demonstrated to participants through feature correlation plots.

Table 1. Details of the ML Practitioners Who Participated in the Study, Identified by Their Participant ID, Including Their Occupation, Area of Expertise, and Sex

ID	Occupation	Specialty	Sex
01	Researcher in Metaphysics	Sound Modulations	M
02	Researcher in Healthcare	Patient EEG Signals	M
03	Researcher in Healthcare	Eye Occlusions	M
04	Undergraduate in Physics	Tailoring Advice to Individuals	M
05	Software/ML Engineer	Military Applications	M
06	Researcher in Computer Science	High Performance Computing	M
07	Masters in Advanced Computer Science	Computer Vision	M
08	Undergraduate in Economics	Economics and Statistics	F
09	ML Engineer	Non-sensitive Projects	M
10	Masters in Advanced Computer Science	Computer Vision	M
11	Masters in Advanced Computer Science	Computer Vision	M
12	ML Engineer	User Profiling	M
13	Researcher in Theoretical ML	Algorithm Performance Guarantees	F
14	Undergraduate in Economics	Economics and Statistics	M
15	Researcher in ML	Accelerating Deep Learning Models	M
16	Researcher in ML	Graph Neural Networks	M
17	Researcher in Computer Vision	Data Reconstruction	M
18	Researcher in Neural Rendering	Generating Objects and Scenes	M
19	Researcher in ML	Neural Networks	M
20	Researcher in Healthcare	Medical Imaging	M
21	Researcher in Computer Vision	Localization Robotics	M
22	Researcher in Healthcare	Rare Diseases	M

3 Study Design

We conducted a remote qualitative study to explore how novices in bias comprehend and apply various bias definitions, detection measures, and mitigation methods. The method is detailed below.

3.1 Participants

Twenty-two ML practitioners were recruited from a range of domains and specialties. Recruitment took place by distributing a call for participation by email and online forums such as Discord servers. We recruited through specific university departments' mailing lists where individuals would have knowledge of ML (Computer Science; HCI; AI-Enabled Healthcare; AI Society; Intelligent Social Systems Lab; Centre for Vision, Speech and Signal Processing) and snowball sampling through past participants' word-of-mouth referral, including industry contacts. The list of participants can be found in Table 1.

We recruited ML practitioners who were novices to bias to explore how this group could employ existing bias metrics and methods to address bias. Our goal was also to gather their reflections on the process, aiming to inform the development of computational tools and methods that can help the ML community operationalize bias. None of the participants worked in the field of fairness or had previously conducted a bias analysis of ML models. This recruitment approach acknowledges that simply providing tools and methodologies is insufficient to resolve bias in ML. Their effective use requires ML practitioners, who may have limited knowledge of fairness, to make numerous critical decisions.

Familiarity with ML and programming was essential for completing the take-home task, as participants needed to run code snippets containing ML models. All participants met the following eligibility criteria: (1) the ability to read and write simple Python code; (2) familiarity with data manipulation (e.g., experience working with Pandas); (3) experience using notebooks (e.g., Jupyter Notebook or Google Colab); and (4) some ML experience (e.g., taking a class in ML or building models that use ML). Participants had advanced educational backgrounds and were either already applying ML in their fields or planning to do so in the near future. They brought diverse expectations and motivations to the study, including a desire to better understand the risks of bias in ML and its implications for their areas of expertise. As a small incentive, and to acknowledge the time spent on the study for this specialized cohort, participants were remunerated with £25.

3.2 Apparatus

The take-home component of the study provided participants with an interactive tutorial via a browser-based coding environment, introducing them to 10 definitions of fairness and bias. They applied these definitions to detect and mitigate bias in the COMPAS dataset [1], with the option to explore the German Credit dataset [25]. These two datasets provided practical, real-world examples for participants to detect and mitigate bias. The COMPAS dataset is known for racial discrimination in predicting the risk of criminal recidivism in Broward County, Florida [1], while the German Credit dataset is known for bias based on sex [60]. There are inherent tradeoffs in satisfying different fairness criteria, making it challenging to address bias comprehensively [9].

To navigate these complexities, the interactive tutorial introduced fairness definitions and bias mitigation techniques from the literature discussed in Section 2.3 and included 12 exercises relevant to the study's research objectives. This list is not exhaustive but represents a selection of definitions covering a range of bias metrics identified by Mehrabi et al. [49], adapted for the practical nature of the study. Although implementing an alternative set of bias definitions could potentially generate a different set of participant challenges and reflections, our selected list served as a means to extract a series of insights and observations within the constraints of a one-and-a-half-hour session. Google Colaboratory (Colab) was selected as the coding environment due to its interactive notebook interface, online accessibility, no setup requirements, and the computational power it provides for running ML models directly in the browser [56].

As part of the take-home task, participants were instructed to read and then run code snippets already provided and encouraged to write their own code to conduct further analysis of the data. They were also asked to write their answers to the exercises within the notebook. The content and exercises were iteratively developed based on feedback from students and pilot participants, including experienced ML users. The exercises included seven qualitative questions and the following four quantitative exercises:

— *Q7*: For variable *statistical party delta*, and feature sex, which training type results in the lowest delta metric? Is this an improvement over the original dataset training type?
— *Q9*: For variable *equalized odds delta*, and feature value African-American, which training type results in the biggest improvement in score? If you chose this training type, how does it affect the *equalized odds delta* of the feature value Caucasian?
— *Q10*: Pick a feature value and write down which training type leads to the lowest ratio of prediction changes, and which leads to the highest ratio.
— *Q11*: Looking at the same feature value that you picked in Exercise 10, have the training types that led to the lowest and highest ratio of prediction changes changed? If so, what are they now?

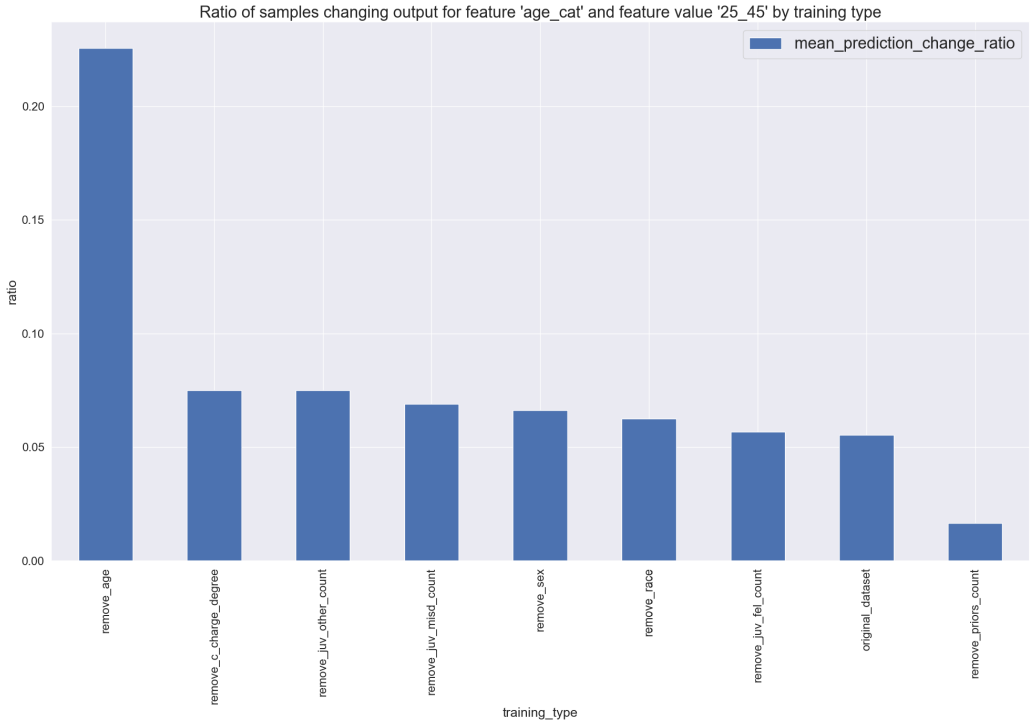

Fig. 1. An example of the ratio of samples changing output for the feature age and feature value 25–45. A reduction in bias is reflected by a decrease in the prediction change ratio. In this instance, removing the feature priors count results in the lowest level of bias.

The quantitative exercises required participants to use graphical outputs from the study interface to identify specific instances of bias. For instance, the graphical output in Figure 1 shows the ratio of samples that change their output for the feature value range 25–45. Removing the feature priors count results in fewer samples changing compared to the original dataset, whereas removing the feature age increases the number of changing samples relative to the original dataset. Due to the quantitative nature of these exercises, participant responses were scored using a binary system. For two-part questions, a half mark was awarded if one part was answered correctly. A summary of key concepts in the order they were presented to participants is presented below, while the full material is available in the supplementary material.

(1) Familiarization with the COMPAS dataset [1], containing 10 input features for 6,172 individuals. This dataset was selected due to its frequent use in fairness research and its well-documented biases.
(2) Participants began by computing the bias analysis metrics and generating the visualizations described in the subsequent steps. This task was performed at the outset, as some computations required several minutes to complete. The process involved training the same Neural Network algorithm across multiple ML models, each containing a different subset of the training features. While participants did not create the models themselves and could not alter the algorithm, they were able to adjust default settings for the Neural Network parameters, such as the number of hidden layers.

(3) Definition of key bias concepts, such as a group, defined as a subset of the dataset filtered by a specific feature value (e.g., male and female groups based on the sex attribute in COMPAS). This step was completed prior to any bias analysis to ensure participants understood the context and objectives of the subsequent exercises in the take-home task.
(4) Bias analysis through model evaluation, using dataframes and scatter plots to display ML and bias metrics for each group: true/false positives and negatives, equalized odds [35], equal opportunity [35], statistical parity [26], fairness in relational domains [27], treatment equality [9], accuracy, precision, recall, and F1 score. This step was crucial for participants to gain practical experience with various methods of measuring bias, using the newly learned concept of groups to partition the dataset based on feature values.
(5) Bias analysis through sensitivity analysis, where participants altered feature values and observed prediction changes, introducing additional metrics: fairness through awareness [26], counterfactual fairness [42], conditional statistical parity [21], and test fairness [16]. This step was essential to introduce the concept of individual-level fairness, demonstrating the impact of keeping all features constant while altering a protected characteristic, such as changing sex from male to female.
(6) Feature correlation plots were used to analyze bias, introducing participants to the fairness through unawareness definition of bias [33]. This step demonstrated how to mitigate bias under this definition by removing discriminatory features and identifying proxy features. It also introduced the sole model-level fairness definition included in the take-home task, along with its corresponding mitigation approach.
(7) Relative delta bias visualizations comparing metric differences for each feature value, calculated against various training types, aiding in selecting the training type that minimizes bias. Each training type was a model trained on a specific subset of features from the COMPAS and German Credit datasets. To minimize bias, a participant should select the training type with the lowest relative delta bias. With participants now familiar with methods for detecting bias, this task was designed to encourage them to reflect on strategies for mitigating bias at a group level.
(8) Absolute delta bias analysis, examining the effects of different training types on within-feature values, comparing them to the original dataset to highlight improvements or declines in metrics for specific feature values. For example, if a new training type results in higher accuracy for males, this would be considered a positive absolute delta bias. This analysis encouraged participants to explore how changes in the training type affected specific feature values using the group-level bias metrics they had learned. In practice, when bias cannot be fully eliminated, ML practitioners may focus on reducing bias for a particular group of individuals.
(9) Prediction change visualizations evaluate how the training type impacts the ratio of samples whose output changes when a given feature value is artificially modified. The least biased training type is identified as the one that causes the fewest output changes. Intuitively, if a new training type results in fewer prediction changes when the feature sex is modified from male to female, it can be said that bias has been reduced. This step assessed participants' ability to mitigate bias on individual-level bias metrics by selecting a training type that minimizes prediction changes when the value of a protected feature is artificially altered.
(10) Prediction volatility visualizations focus on variations in prediction probabilities, identifying the least biased training type by minimizing changes in probability outputs. This offers more detailed insights into output trends than the prediction change visualizations in the previous step by detecting subtle changes in output probabilities. In binary classification, we might not observe a shift from 0 to 1 or vice versa, but these small changes can still be significant.

This step offers greater detail in output trends compared to the previous one and serves as a continuation, providing deeper insights to participants while building upon the prior step.

(11) Due to time constraints, the final part of the take-home task was an optional exercise involving the exploration of the German Credit dataset [25], encouraging participants to further delve into bias concepts and mitigation techniques. Participants were specifically instructed to apply the analysis methods introduced in earlier steps to this new dataset and document at least three insights that could help reduce bias when building a model on it. This dataset, which was also discussed during the interview with all participants, is widely used by the fairness community due to its known biases and was provided to participants as an additional opportunity for reflection.

3.3 Study Procedure

The study procedure consisted of a one-and-a-half-hour take-home task detailed in Section 3.2, followed by a thirty minute semi-structured interview. These times were provided as guidelines to set participant expectations that the study was a significant commitment. However, there was no enforced time limit. To increase completion rates and to keep the gap between the interview and the take-home task to under a week (participant schedule permitting), the interview was scheduled before study material was released. Most participants completed the take-home task on the day of the interview, or in the preceding days. Participants were asked to share their answers in advance of the interview. The study was conducted in accordance with the ethical approval granted by the university ethics committee. Informed consent was obtained from all participants, ensuring they understood the nature of the study, their involvement, and how their data would be used. All participants were given the option to withdraw at any time without penalty.

In the interview, participants were asked to elaborate on their written answers to the exercises in the take-home task, specifically discussing and reflecting on metrics to detect and mitigate bias. Participants could refer to the tutorial and their written answers during the interview. To further assess participants' comprehension and reflections on analyzing biases, participants were prompted to consider potential bias issues and methods for detecting and mitigating bias in the German Credit dataset [25], regardless of whether they had explored this dataset in the take-home task. Without needing to build it, participants were asked to think aloud as they discussed how they would operationalize bias on this dataset. Finally, participants were asked a series of questions around their own experiences with ML bias at work and ethical issues around potentially biased models. The questions were designed to explore participants' written answers to the study and also as a way of getting participants to reflect on analyzing biases. The full interview guide is available in the supplementary material.

3.4 Analysis

The interviews were audio recorded and fully transcribed, and written responses to the 12 exercises in the Colab notebook were also extracted. Both the interview transcriptions and exercise answers were analyzed using inductive thematic analysis [12, 13], following the six-step data analysis process outlined by Clarke and Braun [18]. The process involved open coding conducted by the primary researcher, followed by sample checks conducted by the other researchers. Open coding enabled the exploration of the qualitative data without relying on predefined assumptions or categories [19] and is frequently employed in HCI research when participant responses are open-ended in nature [17, 22, 29]. The open coding approach involved summarizing participant statements into a set of concise, representative keywords. These codes were added to an expanding list of used codes, which was employed to help the researcher maintain consistency in the coding process.

The initial list of codes was refined down to 125 codes (full list available in the article supplementary material), based on merging codes that were misspelled or that had the same meaning. Codes that had fewer than five quotes were checked for novelty, and whether they were entirely covered by other codes. The final list of codes was deliberated among the researchers, who then independently searched for themes by identifying patterns across the codes and data, clustering the codes into groupings of four to eight categories. The initial themes were discussed and refined over seven iterations, leading to a collective decision to consolidate them into four themes. After drafting the findings, each paragraph was tagged, which prompted a reorganization of the findings into five themes, as detailed in the results. The original themes were adjusted to address imbalanced sections and to more accurately align with the research question.

4 Findings

The findings from both the interviews and exercises are presented in the following subsections, each corresponding to one of five themes. These results revealed that eight participants had some familiarity with bias in ML, while another four were aware of the existence of bias-reducing methods; however, all 12 participants had limited or no knowledge of how to address bias, hence they were considered novices to analyzing biases. Eight of the study's participants were already employing some bias prevention methods in their work, either consciously or as part of other ML practices. Our findings also reveal that before the study, seven participants had not considered bias at all, either because it had not been an issue for them, or they believed bias was not relevant to their domain. In contrast, nine participants indicated that the study would lead them to change some of their ML practices to minimize bias as much as possible in their work.

Overall, the interviews indicate that participants found the exercises interesting and engaging. All participants completed the compulsory exercises, and four participants spent more than the recommended one and a half hours to further analyze biases in the provided datasets. P02 expressed that, as an ML practitioner, handling bias is not common in practice. However, participants acknowledged the growing consensus that ML practitioners who design the models are responsible for any bias, as highlighted in emerging regulatory frameworks such as the EU AI Act [20], which mandates that developers use representative and unbiased data for model training. Participants considered this a positive step forward and suggested that bias handling should be taught as part of regular ML courses.

4.1 Sources of Bias

A recurring theme in both the interviews and exercises was participants' perceptions of the origins of bias. Table 2 highlights these sources, categorizing them into four main areas: ML models, datasets, features, and ML practitioners. The table illustrates the diverse reasons participants attributed to the presence of bias. Detailed insights into these specific sources are provided in the following subsection.

Participants identified the underlying ML models as a potential source of bias. Eight of them described the existence of an inherent algorithmic bias, described by P02 in the exercises as: "models looking for correlations rather than causation." Five participants highlighted omitted variable bias as a type of algorithmic bias, which they defined as occurring when a feature influencing the population data is omitted from the training data by the ML model, leading to lower performance on unseen test data. For example, "a face detection model may be trained only on Caucasian subjects and thus perform poorly on real data with subjects of other races" [P05]. Participants also identified sources of bias within the framework of traditional ML concepts. For instance, five participants discussed bias in the context of the bias-variance tradeoff, explaining that high bias occurs when a model overfits, performing well on the training data but failing to generalize to new, unseen data.

Table 2. Sources of Bias Identified by Participants, Grouped by Type, Reason, Count, Participants, and Stage of the Study

Source Type	Reason	Count	Participants	Stage
Model	Algorithmic bias	8	P01, P02, P04, P09, P11, P17, P20, P22	Exercises
Model	Omitted variable bias	5	P01, P08, P09, P14, P16	Interviews
Model	Bias-variance tradeoff	5	P07, P08, P10, P12, P16	Exercises
Model	Intercept in linear regression	1	P16	Exercises
Model	Bias nodes in a linear network	1	P17	Exercises
Data	Unfair, non-representative datasets	14	P01–06, P11–18	Exercises
Data	Sampling bias	11	P01, P03–05, P11, P12, P14–16, P20, P22	Interviews
Data	Prejudice bias	4	P04, P05, P09, P21	Exercises
Data	Historical bias	6	P02, P04, P09, P16, P21, P22	Interviews
Data	Data drift (price or seasonal changes)	3	P01, P05, P12	Interviews
Features	Using protected features (e.g., race)	17	P01, P02, P04–07, P09, P11–14, P16–20, P22	Exercises
Features	Disparities within a feature	14	P02, P03, P06–12, P18–22	Interviews
Features	Proxy of a protected feature	3	P01, P10, P12	Interviews
Features	Vagueness of the feature	2	P03, P06	Interviews
Practitioners	Incorrect assumptions in modeling	10	P01, P04, P06, P09, P13, P14, P16–19	Exercises
Practitioners	Personal traits and beliefs	9	P01, P04, P09, P14, P16–19, P22	Interviews

This understanding of bias aligns with the classical tradeoff between bias and variance [7], where models with high bias are typically too simplistic and unable to capture the complexities of the data. Meanwhile, two participants provided technical definitions of bias. Participant P17 described bias as "a unit in the hidden layer of a linear network" while P16 referred to it as "the intercept in a linear regression, representing the expected value when all explanatory variables are set to 0." These definitions highlight the varying ways participants conceptualized bias within traditional ML frameworks, offering different angles on its implications for model performance.

Regarding dataset issues, 14 participants highlighted unfair datasets, particularly non-representative data as a significant source of bias. They described data imbalances as a key factor contributing to unfairness: "you will probably find more training examples of certain races and if you train a model on that, it will sort of push just having that race [...] toward a prediction of '1'" [P12]. Eleven participants attributed these imbalances to issues in data collection, specifically sampling bias, which they defined as data failing to represent the overall population. This results in models that are not generalizable: "if people of certain races were more likely to be arrested in the first place as a result of existing biases, this could be reinforced by the ML algorithms" [P02]. P13 and P18 highlighted that these data collection issues result in the over-representation of specific groups in the dataset, such as African-Americans in the COMPAS dataset. Four participants referred to this over-representation as prejudice bias, which they attributed to biases inherent in the real world. Similarly, six participants described these real-world biases as historical bias, emphasizing how existing discrimination and socio-economic disparities within the data perpetuate stereotypes. Additionally, three participants raised concerns about measurement problems as a source of bias, focusing on the failure of ML practitioners to adapt their models to data drift. Examples included changes over time, such as price fluctuations or seasonal variations, which can negatively impact model performance.

Participants also discussed feature issues as a source of bias. They expected feature bias from protected characteristics in the COMPAS dataset, such as race (17 participants), sex (14 participants), and age (13 participants). Fourteen participants identified disparities within specific features as a potential source of bias, citing examples from the COMPAS dataset such as sex, juvenile felony count, misdemeanor count, and priors count. Participants also voiced concerns about the quality of features in the COMPAS and German Credit datasets, with three of them noting that some features might be proxies for others, such as the feature sex in the German Credit dataset: "you can make the argument that it's a proxy because females are more likely to be in temporary employment versus permanent employment" [P01]. Two participants mentioned that some features could be better refined, such as loan purpose in the German Credit dataset, which they considered to be too vague.

With regard to ML practitioners as a source of bias, 10 participants described bias as a systematic error in the model arising from incorrect assumptions made by practitioners during the modeling process. Examples included biases introduced through data pre-processing [P06] or preconceived notions held by practitioners [P18]. During the interview, nine participants discussed assumption bias, describing it as biases introduced by ML researchers due to their personal traits and beliefs. For instance, P16 highlighted how relying on ML libraries without fully understanding the assumptions underlying algorithms can result in bias. Meanwhile, P22 expressed their personal belief that, despite age being a sensitive feature, it is reasonable to include it when training a model on the German Credit dataset: "if I see something like a high correlation between age and target variables I think it makes sense because [...] when they get older, maybe they have more experience, they may learn how to live in the community, in a good way." Four participants noted that these issues caused by practitioners can lead to biased algorithmic design choices, resulting in biased model outputs.

In summary, participants identified four distinct main sources of bias in ML models but agreed that multiple sources are typically present: "I think it's difficult to say that bias is entirely contained within the ML approach, because bias certainly emerges just in the data itself" [P05]. For example, four participants mentioned that irregularities in the form of outliers in small datasets can cause problems for model-building and analyzing bias, or that feature imbalances might be caused by sampling bias in data collection.

4.2 Employing and Selecting Bias Metrics

This theme provides insights into the decision-making processes of participants regarding selecting different bias metrics and the practical implications of employing them. The take-home task asked participants to describe which group-level bias metrics (equalized odds, equal opportunity, statistical parity, or treatment equality) they would select to assess bias in the presented datasets. As shown in Table 3, participants did not reach a consensus on a single method for measuring bias in the COMPAS dataset, emphasizing the various advantages and limitations of different metrics. A similar pattern was observed with the German Credit dataset, where four participants who provided written responses selected equalized odds to minimize false positives. However, P04 and P12 also opted for treatment equality, with P12 additionally choosing equal opportunity.

Treatment equality emerged as the most popular bias metric, chosen by 13 participants, six of whom preferred it because it considers both false positives and false negatives. They viewed these errors as particularly harmful within the context of the COMPAS dataset, as it could result in either releasing recidivists or unnecessarily detaining individuals who would not re-offend. Four participants chose it because it yielded low values, making the model appear fairer, perhaps artificially satisfying the goal of creating a fair model. P16 chose it to prevent performance concentration on a single group.

Table 3. Bias Metrics Chosen by Participants for the COMPAS Dataset during the Take-Home Exercises, Along with Their Reasons and Corresponding Participants

Bias Metric	Reason	Count	Participants
Treatment equality	Lowest delta value	4	P04, P10, P14, P20
Treatment equality	Incorporates both false positives and false negatives	6	P02, P05, P09, P12, P17, P18
Treatment equality	Avoids performance being concentrated on a single group	1	P16
Equalized odds	Lowest delta value	2	P04, P20
Equalized odds	Penalizes false positives	2	P07, P11
Equalized odds	More stringent definition than equal opportunity	1	P03
All metrics	Base conclusions on all metrics, identify largest differences	3	P06, P12, P13
Statistical parity	Recall and F1 score are similar to statistical parity	2	P07, P22
Fairness in relational domains	Less stringent, makes sense in practice, as it sets a threshold	1	P19
Own metric	A modified version of treatment equality	1	P08
Own method	Sampling input features equally as it minimizes bias	1	P21
Accuracy	What ML practitioners aim for when training models	1	P15

Participants did not converge on a single method for measuring bias.

Equalized odds was another popular choice, selected by eight participants. Reasons for choosing it included its ability to reduce false positives (three participants), the low values observed during experimentation (two participants), and its rigorous definition, which accounts for both the odds of being correctly and incorrectly assigned a positive outcome (one participant). This is unlike equal opportunity, which considers only the odds of being correctly assigned a positive outcome. Notably, no participant chose equal opportunity as the sole metric; it was only selected by the three participants who chose to use all four metrics to draw comprehensive conclusions.

Although less frequently selected, P07 and P22 chose statistical parity because of its connection to the F1 score, a traditional ML evaluation method. P19 preferred a relaxed version of statistical parity known as fairness in relational domains, valuing its practical approach of setting a fairness threshold. Notably, during the follow-up interview, five participants questioned whether practical thresholds could be applied to the other bias metrics as well.

Participants also proposed using traditional ML evaluation methods directly, instead of bias metrics, to detect bias. Specifically, three participants stuck with traditional ML evaluation methods like accuracy [P15], or defined their own methods using ML concepts [P08, P21]. For example, P08 proposed specific ratios of false positives and negatives to measure bias impacts: "I would like to use false_positives / (true_positives + false_positives) and false_negatives / (true_negatives + false_negatives) because if any group has a higher rate of either of these, it will mean either innocent people being targeted with more suspicion because of their group, or recidivists being unsuspected because of their group" [P08].

During the interview, participants reflected on the use of bias metrics, focusing on strategies for selecting metrics, interpreting them in practical contexts, and considering the impact of group sizes. Five participants questioned whether different types of data require different bias metrics, with most participants expressing that they would use the same bias metrics for both COMPAS and German Credit. However, eight participants proposed selecting a bias metric based on the application goal, such as minimizing false positives for the COMPAS dataset, or false negatives for the German Credit dataset: "if a person who is capable of paying back the loan is less likely to get the loan because they are a certain race or age or gender, then that is probably unfair" [P08].

Participants also mentioned ML model-building considerations, like using bias metrics iteratively to improve models [P06].

Participants also noted potential issues with employing bias metrics, like selecting metrics yielding the highest fairness level (four participants), which could lead to "gaming the system." Another issue pointed out by two participants was the tradeoffs between different metrics, as they mentioned it is not possible to satisfy all fairness definitions simultaneously. They attributed this to a tradeoff between minimizing false positives and false negatives. Additionally, five participants noted that small or highly imbalanced datasets could impact the reliability of the metrics.

P08 and P09 highlighted the loss of interpretability when employing bias metrics. For equalized odds, P09 highlighted that the value for two groups might be the same, but one group could have a high false-positive rate while the other has a high true-positive rate due to the metric's harmonic mean of both false positives and true positives. For statistical parity, P09 mentioned that the summing of true and false positives leads to a loss of granularity. Regarding treatment equality, P08 explained that its construction can result in similar scores for groups with different rates of false positives and false negatives: "if a group has both greater false positives and false negatives, they will have similar scores as another group with lower false positives and false negatives, but this is still unfair" [P08]. Two participants mentioned a loss of interpretability due to subtle differences between some metrics, which made it difficult for them to grasp the distinctions, such as between equalized odds and statistical parity.

In the interview, all participants considered conventional ML model evaluation techniques within the context of mitigating bias. Seven participants used accuracy to discuss the COMPAS dataset results, while three of them noted that maximizing accuracy alone can cause bias. Six participants used other evaluation metrics like F1 score or recall, or a combination of all metrics presented in the study. Eight participants highlighted how differences in traditional metrics across groups could reveal biases without needing specialized bias metrics. For example: "based on how the models perform differently among different sub cohorts and the different ways of removing features affected the scores, this would suggest that that there would have been some biases even in the other features that don't explicitly encode biases" [P02]. This suggests that existing metrics could be repurposed to detect bias.

4.3 Detecting Bias

This theme reports on how participants implemented a selection of bias metrics and reflected on them through written responses to the study exercises and subsequent interviews. Table 4 summarizes participants' responses to the quantitative elements of the take-home task, illustrating their effectiveness in detecting bias within the datasets. Ten participants answered all four questions correctly, while nine missed points on only one question. Most errors occurred in Exercise 7, with no mistakes recorded in Exercise 10. When errors did occur, participants either failed to identify the training type that most effectively reduced bias, focused on the wrong discriminatory feature, or provided partial answers, potentially due to technical issues. Despite being an optional exercise, P04 identified bias in the German Credit dataset by examining the feature sex and found that removing features such as duration or housing reduced the level of bias.

During the interviews, participants reflected on instances of bias they detected in the COMPAS dataset. Nine participants described bias against African-Americans, observing significant disparities in bias metrics between races: "African-Americans consistently have a higher 'equalized odds / equal opportunity'" [P21]. This supports ProPublica's findings that black defendants were more likely to be misclassified as high-risk compared to white defendants [1], indicating a higher false-positive rate. Additionally, eight participants reported significant differences in bias metrics between age groups, with younger offenders scoring worse, confirming ProPublica's findings that younger

Table 4. Participants' Performance Scores on Bias Detection Exercises

Participant	Q7	Q9	Q10	Q11	Score	Score %
P01	0.5	1	1	1	3.5	87.5
P02	0	1	1	0	2	50.0
P03	0.5	1	1	1	3.5	87.5
P04	1	1	1	1	4	100.0
P05	1	1	1	1	4	100.0
P06	1	1	1	1	4	100.0
P07	1	1	1	1	4	100.0
P08	0.5	1	1	0	2.5	62.5
P09	1	0	1	1	3	75.0
P10	1	1	1	1	4	100.0
P11	1	1	1	1	4	100.0
P12	1	1	1	0	3	75.0
P13	1	1	1	1	4	100.0
P14	1	1	1	1	4	100.0
P15	0	1	1	1	3	75.0
P16	0.5	1	1	1	3.5	87.5
P17	1	1	1	1	4	100.0
P18	1	1	1	0	3	75.0
P19	1	1	1	1	4	100.0
P20	0.5	1	1	0	2.5	62.5
P21	0	1	1	1	3	75.0
P22	1	0	1	1	3	75.0
Accuracy %	75.0	90.9	100.0	77.3		

Quantitative exercises were evaluated using a binary scoring system, with half marks awarded for partially correct answers in two-part questions. Most participants answered all four questions correctly.

individuals were more likely to receive higher scores [1]. Two participants noted a data imbalance, with those aged 25–45 being over-represented. However, while ProPublica found that female defendants were more likely to receive higher re-offending scores than males, participants did not detect this. Instead, seven participants found that men were discriminated against in terms of recall: "males have almost double the recall of females, so re-offending females are half as likely to be identified" [P18].

Participants also detected bias through a feature correlation plot, finding it useful for understanding the causes of bias. Six participants noted that highly correlated input features are detrimental to the model, even if none are protected features. Five participants used feature correlations to identify bias in the COMPAS dataset, even when one feature was non-discriminatory. For example, one participant observed that non-discriminatory features correlated with specific races could lead to biased models: "the values for these features are greater than 0 mostly for people belonging to the African-American race. This could lead to a model that highly discriminates on people belonging to this race" [P10]. Another participant noted differences between African-American males and Caucasian females in F1 scores. However, eight participants were sometimes unsure about what constitutes a closely correlated proxy feature or a protected characteristic, indicating a need for

clearer guidelines. Three participants suggested that discriminatory features should be limited to those that cannot be changed, such as age, sex, and race.

4.4 Mitigating Bias

This section includes participants' reflections on the bias mitigation process as they attempted to reduce bias in the COMPAS and German Credit datasets. The results showed participants successfully reduced bias by selecting training types that minimized prediction changes for certain features in both COMPAS and German Credit.

One exercise asked participants to change the model's training type to mitigate bias. Each training type was a model trained on a specific subset of features from the COMPAS and German Credit datasets. P01 compared this mitigation method to Shapley Values, explaining in the interview that they are used in explainable AI to evaluate the importance of each feature: "if I remove race, how many label predictions change? If not many change, then you say it's not a very important feature" [P01]. To mitigate bias, participants used sensitivity analysis to measure differences between training types. This analysis involves changing the input values of potentially discriminatory features and observing changes in the model's output. For instance, changing the feature sex from male to female to see if the prediction changes. If a model is truly fair, then the output would not be influenced at all by changes in discriminatory features.

In practice, participants' goal was to select a training type where fewer predictions change, thereby partially mitigating bias. Five participants noted that the sensitivity analysis can help identify borderline cases where slight input changes affect predictions. They also saw this analysis as a way to observe drivers of predictions, by observing how specific features influence the model's output. According to five participants, it is acceptable for a model's predictions to change when input features are modified, if the changes are driven by non-discriminatory features: "with priors count, you would expect the predictions to change anyway, at least slightly, at least much more than with gender" [P04].

In the COMPAS dataset, participants examined features such as race (15 participants), age (10 participants), sex (9 participants), priors count (6 participants), juvenile other count (2 participants), juvenile misdemeanor count (1 participant), and the type of charge (1 participant). Despite the availability of bias metrics to detect bias, 16 participants relied on their intuition rather than the bias metrics to identify potentially biased features. For instance, nine participants noticed many predictions changed from *will re-offend* (a value of "1") to *will not re-offend* (a value of "0") when the feature race was changed from African-American to any other value, with four of these participants identifying training types that partially mitigated this bias. Similarly, P14 and P19 observed changes in predictions when modifying race from Caucasian to Others, with removing the feature priors count or sex resulting in fewer prediction changes.

Nine participants identified changes in the feature sex from male to female also led to many predictions changing from "1" to "0," with four of these participants finding that removing race or age led to the greatest reduction in changes. This suggests that excluding these features could decrease bias related to sex. Nine participants acknowledged potential tradeoffs in bias mitigation, noting that reducing bias for one feature might increase it for another. For example, four participants used the bias metrics to observe that removing *priors count* reduced bias for African-Americans but worsened it for Caucasians: "For equalized odds, and feature value African-American (race), removing priors count resulted in the biggest improvement in score. For feature value Caucasian (race) it negatively impacts the score" [P11]. Regarding age, five participants found that changing the age category from 25 to 45 to under 25 caused many predictions to change from "0" to "1." Three of these participants identified that the training types removing race or priors count helped reduce this effect.

Findings in Section 4.3 show how participants used the feature correlation plot to identify bias. As part of the exercises, participants used these plots to mitigate bias by removing features correlated with discriminatory ones. This approach encouraged participants to consider the fairness through unawareness definition of bias [33]. Six participants found this intuitive, as they related the definition to real-world use cases, such as companies removing labels like race and sex from their ML models. Although one participant expressed surprise that, under this definition, a model is considered fair simply by removing protected features and their proxies.

The exercise results show that 15 out of the 19 participants found priors count valuable and did not remove it despite its correlation with the output, as removing it negatively impacted the accuracy of the model. It was also not seen as a directly biased feature, though indirectly some participants argued that the feature can hold bias because certain groups are more likely to be arrested, referring to data issues as described in Section 4.1. Six participants observed a correlation between juvenile misdemeanor count and priors count and chose to remove juvenile misdemeanor count due to concerns about inferring age. Twelve participants recognized the tradeoffs in this approach, as removing features might decrease model accuracy or increase bias in other correlated features. For example: "I would not remove any of them, because removing juvenile misdemeanor count increases the sensitivity of age, removing priors count increases the sensitivity of race" [P08]. Four of these participants found it hard to justify removing even sensitive features if they are valuable for the output.

The study shifted participants' perspectives on addressing bias within their own work. As ML practitioners, they suggested that in the future, they will quantify potential bias rather than rely on a qualitative inspection of the dataset [P09, P18], conduct a sensitivity analysis on the features [P05, P09], and choose non-discriminatory features [P13]. Participants also proposed several other bias mitigation methods to tackle the bias sources described in Section 4.1 that they could use in their own domains, centering on issues with the datasets, features, model, and the practitioners themselves. These are outlined in Table 5. Four participants suggested using traditional ML techniques to mitigate bias, such as regularization, dropout, model complexity reduction, and cross-validation and that choosing certain ML algorithms over others can help reduce bias: "algorithms which fight bias will be chosen over those that don't, for example Catboost where gradient boosting is needed" [P05]. In terms of a bias mitigation tool, two participants suggested it would be useful to have automated systems to monitor bias in the data.

4.5 Ethical Considerations

Participants considered ethical ramifications when reflecting on the potential for bias in ML. Six participants expressed the view that ML applications might always contain bias due to inherent unfairness in real-world scenarios, despite implementing the bias mitigation techniques described in Section 4.4, such as by refining features in COMPAS to use location instead of race: "if you are measuring postcode and we did live in a world where ethnicities were evenly distributed geographically, theoretically only then would our models be fine" [P01].

Participants brought up ethical concerns regarding the use of ML. For instance, P18 highlighted the challenge ML practitioners face in simplifying real-world applications to fit ML models, such as the German Credit dataset missing critical insights: "if they're living with their parents, they've got an income, they're not paying much rent or whatever, then actually they could be in a really good financial situation. And the mortgage people could actually be struggling and have less money and actually be a higher risk" [P18].

Eight participants discussed the possibility that biases in ML models originate from real-world societal issues rather than the models themselves. For example, in the COMPAS dataset, two of these participants linked race and sex to crime, noting, "I believe statistically someone of an

Table 5. Bias Mitigation Proposals and Participant Counts across Various Subject Areas

Subject Area	Count	Actual Observation
Data collection	6	Collecting more data if features at risk of bias cause big changes in performance.
Data collection	1	Ensuring a diverse range of samples are used [P18].
Data collection	1	Incorporating different datasets of the same kind, rather than simply trusting one dataset [P09].
Data pre-processing	5	Resampling the output classes to make them equal. However, P21 pointed out that in practice it may be hard to balance both protected input features and the output.
Data pre-processing	1	Removing outliers, although in small datasets there may not be enough data to know if a specific sample is an outlier [P06].
Features	2	Improving features by the way they are labeled, such as measuring a person's ability to repay a credit instead of measuring risk in the German Credit dataset.
Features	1	Performing a transformation on the features in the hope that it reduces some of the biases that arise from skewed data [P08].
Features	4	Breaking features down into more granular sub-features, such as accompanying the feature race in COMPAS with other demographic data, or further splitting the age category.
Features	1	Ensuring no two features have a high correlation between each other.
Model	6	Building models that do not include discriminatory features and do not extrapolate from the data.
Model	2	One input should not be prioritized over another input.
ML practitioners	2	Discussing one's own model-building ideas with others.

Suggestions for mitigating bias were primarily focused on features, data collection, and pre-processing methods.

African-American race would be more likely to commit a crime in that respect, but that's basically a socio-economic issue [...]. I think that if you allow a model to deliver these insights then it will perpetuate the stereotype that African-Americans are people that commit crimes" [P04]. Consequently, participants suggested that bias might always be present in some ML applications due to underrepresented groups in both the training data and society. This discussion led to reflections on how differing societal views influence ML model design. For instance, in the context of the COMPAS model predicting recidivism, opinions varied on whether to prioritize individual human rights or societal safety. Two participants argued that a model could not be considered biased if the data were accurate and sufficiently sampled, implying that it is simply a reflection of bias in the world [P16, P21]. P08 suggested that the onus of ensuring that models are unbiased falls on those who design them in the first place.

Given these ethical ramifications, four participants suggested instances where ML should not be used, such as: when removing problematic features causes the accuracy to drop below an acceptable range; if bias cannot be mitigated due to missing data or features; or if the application is "very subjective," for example with the COMPAS dataset: "it's very subjective to tell whether the person

is going to commit crime or not. It's not like 100% sure that he's going to commit a crime" [P11]. Ten participants suggested that whether ML should be used despite bias concerns depends on the specific application and whether the available alternatives are more subjective. For example, P06 suggested bias might be acceptable if the application is not critical: "It may not be a problem if my model cannot recognize a white cow with black spots to a Dalmatian dog, but when it comes to matters about sex, jobs, race, and so on, this could be a real problem in case of wrong results" [P06]. This was reflected in the study: three participants found it harder to discuss and reflect on bias for the German Credit dataset because the detrimental effects were not considered as serious as with the COMPAS dataset.

Additionally, four participants mentioned the importance of bias depending on regulatory requirements. Six others emphasized avoiding bias by being meticulous throughout the model generation process, such as when choosing features, the assumptions that they make in the model, and ensuring the bias metrics yield similar scores across different groups. Meanwhile, four participants indicated that, despite gaining a greater awareness of bias sources through this study, they would not alter their approach to constructing ML models.

Participants sometimes based their reasoning on stereotypes when discussing how bias arises in the input features of the study's datasets. This was particularly evident regarding age. While considered discriminatory, four participants found its use acceptable in certain contexts. For example, P22 stated, "if I see something like a high correlation between age and target variables I think sometimes it makes sense because I think most of the people who have done something wrong, they did it when they were adolescents. But when they get older, maybe they have more experience, they [...] learn how to live in the community." Another example: "age is one thing that in this case I would argue is, well, I would keep age in credit risk models because I think inherently people become more financially free, have more money, work in higher paying jobs as they get older. So, in this case age is actually a very good indicator of your creditworthiness" [P12]. In the COMPAS dataset, eight participants thought the ML model was making judgments based on race. Five of these participants attributed this to institutional discrimination and socio-economic disadvantages. However, P16 based their reasoning on stereotypes, suggesting that cultural or habitual differences might also contribute.

5 Discussion

The findings revealed that participants, as users of ML, understood and effectively applied various bias definitions, detection measures, and mitigation methods, thus addressing the research question. This understanding was evaluated through the study exercises and interviews, where participants elaborated on their responses and reflected on best practices in model-building. While participants discussed different fairness definitions and identified bias in the COMPAS dataset, they encountered challenges during this process. Consequently, two main discussion points have been identified from the findings and are reported in the subsequent section.

5.1 Unresolved Conflicts

The findings in Section 4.1 indicate that participants had an intuitive understanding of bias formation and types but faced challenges in selecting relevant metrics, balancing model performance with bias mitigation, and relying on data over personal opinions. These challenges echo prior research suggesting most bias remains unmitigated in practice [38].

Participants experienced difficulties when selecting bias metrics, corroborating findings by Madaio et al. [47]. Section 4.2 shows that, although participants could employ various fairness definitions and bias metrics to detect bias, they were uncertain about which metric to apply in different contexts. Consequently, they often selected metrics that made the ML application appear

less biased rather than addressing the underlying fairness issues. This issue of practitioners "gaming the system" is corroborated by prior work from Veale et al. [74]. The tendency to select metrics that minimizes apparent bias highlights a broader issue in ML practices, where models are frequently chosen based on summary statistics like accuracy, obscuring important information about a model such as group sizes, and leading to the deployment of problematic models [58, 78].

The effect of employing different metrics when analyzing biases is shown in prior research on the COMPAS dataset, where ProPublica highlighted violations of equalized odds and equal opportunity fairness criteria, indicating African-American defendants had higher false-positive rates than Caucasian defendants [1]. In contrast, Northpointe, the creators of COMPAS, claimed fairness based on statistical parity [24]. Our findings in Section 4.3 show that participants differed from the bias literature in their conclusions about gender discrimination in the COMPAS dataset. This discrepancy arose as ProPublica adjusted for the same factors by balancing group sizes [1], highlighting the impact of data imbalances on bias metric applications and the conclusions drawn when analyzing biases.

Findings (e.g., in Section 4.2) indicate that bias metrics may suffer from a lack of interpretability because of their construction, where different underlying outputs can yield the same metric value. This echoes prior work which has called for ways to measure model performance beyond summary statistics that conceal important information about the model [74] and extends it to detecting bias. One suggested solution is a unified metric for algorithmic unfairness using inequality indices from economics, which accounts for group size and allows for easy comparison of algorithms at both individual and group levels [66]. This solution also addresses concerns raised by our participants (e.g., in Sections 4.1 and 4.2) regarding imbalances in group sizes.

Our findings in Section 4.3 show that participants identified biases in both the COMPAS and German Credit datasets, aligning with prior research indicating that the COMPAS dataset is biased concerning race, sex, and age [1], while the German Credit dataset is biased toward sex and low-skilled workers [60]. Despite participants' interest in fairness and the availability of bias mitigation techniques, our findings (e.g., in Section 4.4) indicate that they were unable to completely eliminate bias from the COMPAS and German Credit datasets. Participants (e.g., in Section 4.5) attributed the difficulties in mitigating bias to the simplification of complex real-world processes into models with limited variables. This may be attributed to problem selection, which prior research has shown to be the cause of failure in 87% of ML projects [69], often due to the complexity of the designed applications [39, 76]. An implication might therefore be to encourage ML practitioners to address simpler problems, reducing the impact of bias caused by sampling issues or societal factors.

Participants also found (e.g., in Sections 4.3 and 4.4) that non-discriminatory features in the COMPAS and German Credit datasets implicitly encoded real-world bias. This confused participants, with findings (e.g., in Sections 4.1 and 4.5) revealing that were unsure whether it stemmed from the way these datasets were constructed (i.e., unrepresentative data collection) or if it reflected bias in the world (i.e., representative of an unfair society). The issue of real-world bias is well-documented, with our participants echoing prior work noting racial disparities in healthcare outcomes and access [48]. Consequently, they were conflicted about the best approach to tackle this issue. Drawing on ideal theory concepts from political philosophy [72], prior research addresses this ethical concern, questioning whether ML practitioners should create models based on real-world biases, risking discrimination, or based on an idealized society, risking unrealistic models [28].

Our findings in Sections 4.2 and 4.4 show that participants struggled to balance model performance with bias mitigation, recognizing the existence of a tradeoff. This aligns with prior research demonstrating a fairness-accuracy frontier [44]. However, this contrasts with Balayn et al. [4], who found that ML practitioners did not consider the tradeoff between accuracy and fairness when

addressing biases. This difference may be attributed to the participant groups: this study included ML practitioners with an interest in fairness, whereas the previous study did not.

Meanwhile, findings in Section 4.5 indicate that participants struggled to justify mitigating bias by removing discriminatory features if it caused the model's performance to fall below an acceptable threshold. While participants were sensitive to bias, they prioritized maintaining model accuracy and were willing to use sensitive features to preserve performance, often justifying their decisions with personal opinions and stereotypes. This contrasts with Deng et al. [23], who observed that ML practitioners commonly assume that removing sensitive features like sex is necessary as it enhances fairness. However, it aligns with prior fairness research on non-ML experts by Cheng et al. [15], who found that non-ML experts were reluctant to compromise on accuracy in order to mitigate bias by achieving equalized odds across different feature groups. Additionally, this finding supports previous ML research showing a strong reliance on accuracy as a measure of model performance [40, 58, 78], extending this reliance to considerations of bias.

This section has outlined the various challenges participants, as novices in addressing bias, faced when operationalizing bias. Additionally, previous research has identified that ML practitioners often view data collection and processing as the primary challenges to fairness in ML [36, 47]. Our findings in Section 4.5 support and expand on these observations, indicating that ML practitioners perceive that bias is not an issue if the dataset is well-sampled, which can create an unwarranted sense of confidence in the ML process. These challenges and perceptions may help explain the gap between the practical use of bias metrics and current ML practices. While prior research has introduced numerous bias metrics [26, 35, 42] and tools [8, 62, 65], this section highlights that addressing bias in ML requires practitioners to make critical decisions beyond simply using these resources.

5.2 Reflections on Bias Concepts

Despite the conflicts described in Section 5.1, findings in Section 4.3 show that participants successfully identified bias in the COMPAS dataset, corroborating ProPublica's analysis [1]. Findings in Section 4.1 show that participants were intuitively aware of various sources of bias, with their explanations closely aligning with prior literature. For example, in Section 4.3, they observed a higher false-positive rate, while in Section 4.1, they pointed out that the over-representation of African-Americans could stem from biases in law enforcement practices, leading to disproportionately higher arrest rates for African-Americans. Participants also identified biased algorithmic design choices by ML practitioners as a source of bias, confirming prior research that found this to be a significant source of bias [3]. Similarly, Section 4.4 demonstrates that participants could reduce bias using the proposed sensitivity analysis method, indicating that despite conflicts, ML practitioners can reason about and operationalize bias.

Participants also proposed ways to apply bias mitigation techniques in their own ML projects. For instance, findings in Section 4.4 indicate that participants could reason about various methods applicable to their work and that the study enhanced their ability to identify bias methodically and quantitatively. Through these reflections, participants demonstrated good model-building practices and used their knowledge of ML to incorporate bias detection and mitigation techniques, showing a significant overlap between implementing these methods and current ML practices.

Our findings (e.g., in Sections 4.2 and 4.3) indicate that participants recognized a tradeoff between different bias metrics when detecting bias, attributing this to the balance between minimizing false positives and false negatives. This supports prior research, which also identified such tradeoffs between bias metrics [49, 66], and extends these insights to the context of empirical testing by ML practitioners. Additionally, our findings reveal that participants did not universally adopt a single method for measuring bias, instead noting the limitations of each metric. For example,

metrics like treatment equality, which aim to balance minimizing false positives and false negatives, were criticized for sacrificing granularity. This contrasts with the work of Balayn et al. [4], who found that ML practitioners lacking a strong interest in fairness often applied bias metrics without critically reflecting on their suitability or limitations. This finding also contrasts with prior research on non-ML experts, where Srivastava et al. [68] found that lay users preferred the adoption of statistical parity, possibly because its simpler mathematical construction made it the most intuitive representation of fairness.

Our participants, as ML experts relatively new to bias considerations, suggested (e.g., in Section 4.2) choosing bias metrics that prioritized minimizing either false positives or false negatives, depending on the objectives of the ML application. This aligns with previous research on the precision-recall tradeoff encountered by ML practitioners [14], which is based on confusion matrix elements, and expands this understanding to bias metrics. However, this finding contrasts with the work of Haider et al. [34], which found that non-ML experts' preferences for fairness definitions were highly subjective and shaped by their individual backgrounds.

Although the study was designed to assess participants' understanding of bias concepts through the short exercises and interviews, the findings indicate that even a short-term learning program enabled novices to develop a strong grasp of bias detection and mitigation. For instance, the results in Section 4.1 demonstrate that the study helped participants recognize the risk of introducing their own biases when designing ML models. This supports calls by prior research [36, 73] for more resources to help ML practitioners enhance fairness and extends their findings by demonstrating the practical benefits. It also aligns with findings by Lee and Singh [43], suggesting that better guidance may help overcome the steep learning curve of existing fairness tools and complements previous research indicating that short training sessions can significantly improve learners' mental models and system satisfaction [41], extending these findings to the analysis of biases.

6 Implications

Sections 4.3 and 5.2 show that participants successfully detected bias and were able to partially mitigate it in a practical setting. These results imply that ML practitioners could benefit from more encouragement to use bias detection and mitigation methods. Despite numerous fairness definitions, bias metrics, and packaged tools available (e.g., [8, 62, 65]), the availability of these tools and methodologies alone does not resolve bias in ML, as their effective use requires practitioners to make numerous critical decisions. Prior work on how ML practitioners use fairness tools indicate that existing fairness tools need to better align with the needs of ML practitioners [23, 43, 59]. Our study supports this call and expands on it by offering a series of implications for operationalizing bias, including insights into the design of bias mitigation strategies and the integration of bias concepts with ML practices, as detailed below.

6.1 Designing Bias Mitigation Strategies and Tools

A more consistent application of fairness principles is needed across diverse contexts and applications. Our findings (e.g., in Sections 4.3 and 4.4) reveal that participants used various techniques to detect and mitigate bias, often arriving at differing conclusions about the data. These discrepancies mirror the differences between our participants' findings, ProPublica, and Northpointe, as discussed in Section 5.1, and align with prior work highlighting tensions between different fairness interpretations [73]. This finding supports calls from previous research on bias assessment tools [23, 43, 59] to standardize bias metrics and provide transparent benchmarks for assessing dataset fairness. However, such standardization may be impractical due to the lack of a universally agreed-upon definition of fairness [63] and the wide variety of bias types and sources of discrimination [49].

In addition, Sections 4.2 and 5.1 revealed that novices in bias often struggled to select appropriate bias metrics, expressing uncertainty about which metrics to use and whether their choice should vary across ML applications. This finding aligns with Nakao et al. [54], who observed that both non-ML and ML experts unfamiliar with fairness concepts sought additional explanations of bias metrics to understand their appropriate use. These results suggest that simply providing customizable fairness and performance metrics, as proposed by Richardson et al. [59], could overwhelm novices in bias analysis. Instead, practitioners would benefit from guidance in selecting the most suitable bias metrics for their specific contexts. For instance, Smith et al. [64] recommended that practitioners first define their fairness objectives and then use a curated menu of choices to identify relevant metrics. Similarly, Mitchell et al. [53] proposed emphasizing a deeper understanding of foundational concepts and assumptions to guide metric selection.

However, other researchers have advocated prioritizing ethical fundamentals over metric-driven approaches [10, 30]. For instance, Binns [10] has called into question key tradeoffs, both between individual and group fairness, as well as among group fairness metrics, arguing that ML practitioners should instead focus on core assumptions, such as the purpose of the model and the data-gathering process. Our findings (e.g., in Section 4.2) and discussions (e.g., in Section 5.1) support this perspective. While participants reflected on potential sources of bias, we observed instances where they "gamed the system" by selecting metrics that made the ML application appear less biased, rather than addressing the underlying fairness issues.

Section 4.4 implies the need to automate parts of the bias mitigation process, especially for high-dimensional datasets and models, as novices may be overwhelmed by excessive information. This aligns with prior research on the complexity of ML algorithms, which can make even explainable models like linear models and decision trees difficult to interpret due to human cognitive limits when dealing with high-dimensional data [46], extending this challenge to analyzing biases.

6.2 Leveraging Overlaps between Bias Concepts and ML Practices

Our findings indicate the potential to integrate bias concepts into ML practices, apply existing ML practices to measure bias, or adapt concepts from ML for bias. As discussed in Section 5.1, a key implication for designing bias detection methods for novices is the need to quantify the bias-accuracy tradeoff. In ML, the F1 score, a widely adopted metric that balances precision and recall [14, 71], provides a useful framework. We suggest that a similar harmonic mean-based approach, adapted from ML practices, could be applied to balance bias and accuracy. Since ML users are already familiar with the F1 score, leveraging this existing knowledge could help novices better understand a new metric for evaluating the bias-accuracy tradeoff.

Sections 4.1, 4.4, and 5.2 highlight a notable overlap between bias concepts and standard ML practices like data handling and feature refinement, which could help novices in grasping bias concepts. This finding complements previous research that showed ML practitioners often deprioritize unfamiliar values [73]. By integrating bias concepts with familiar ML practices, we expect that ML practitioners' engagement with issues of fairness to increase. Participants noted that typical ML tasks involving the training dataset, such as correct data sampling, balancing datasets, and collecting relevant features, can also help mitigate bias. This finding aligns with prior research by Holstein et al. [36], which found that ML practitioners consider standard dataset-related tasks, like collecting more training data, essential for addressing fairness issues. Therefore, leveraging these overlaps in designing bias mitigation strategies and tools can better equip ML practitioners to understand and apply them effectively.

The findings (e.g., in Section 4.2) indicate that traditional evaluation metrics familiar to ML practitioners, despite their limitations, can be used to measure group-level bias and highlight disparities between groups. This approach can foster awareness of bias detection while leveraging

established ML practices. These metrics could serve as iterative feedback loops for addressing bias, building on prior research on iterative ML model design processes [77] and extending them to bias mitigation strategies. However, this approach should be used with caution: prior work on non-ML experts has shown that while adjusting feature weights through iterative feedback loops could improve the fairness of a ML model, it sometimes resulted in worsening fairness [55].

There is also an opportunity to design visualizations explaining bias metrics, as highlighted by findings in Section 4.3, confirming prior work that visualizing model evaluations leads to better understanding and increased trust [45], with participants in our study effectively using bar charts to visualize and identify bias.

7 Limitations and Future Work

Our participants, as ML experts with a high level of education, were likely better equipped to identify and address bias within the study. For instance, prior research by Saha et al. [61] on a general audience of non-ML experts found that education strongly predicts comprehension of fairness concepts. Despite this, our findings (e.g., in Sections 4.1, 4.2, and 4.3) revealed several unresolved conflicts, highlighting the challenges ML experts who are new to bias face when attempting to operationalize various measures for detecting and mitigating bias. Moreover, the take-home task was designed to evaluate participants across a range of computational methods applicable in the ML community. We believe this participant group, with expertise spanning areas such as medical imaging, military applications, and robotics, reflects individuals well-suited for such an evaluation. They demonstrated both a willingness and a realistic capability to apply fairness concepts within the context of a one-and-a-half-hour take-home task and potentially beyond, in their own professional work.

Although the COMPAS dataset is widely used by the fairness community, prior research has pointed out issues with using it to assess algorithmic fairness through bias metrics [5], noting that the dataset contains errors in how the data were measured and collected, therefore not making it a representative "real-world" dataset. For the purposes of our study, these measurement issues provided participants with opportunities to reflect on both the dataset and the metrics. For example, Bao et al. [5] emphasized how historical structures have perpetuated injustices. Our findings (e.g., in Section 4.1) reveal that participants recognized and reflected on this issue, noting how discrimination and socio-economic disparities in data collection have led to the over-representation of certain groups in the COMPAS dataset.

Finally, the primary aim of our research was to expose ML practitioners who are novices in bias to a range of fairness definitions and bias metrics. However, the implications (e.g., in Section 6.1) align with prior research emphasizing the importance of prioritizing ethical fundamentals over metric-driven approaches to bias mitigation [10, 30]. Future research could build on this by focusing on the key assumptions and normative goals ML practitioners encounter when addressing and operationalizing bias.

8 Conclusion

This article presented a qualitative study designed to understand how ML practitioners who are novices in bias can operationalize definitions of bias and apply mitigation methods. The study included 22 participants who engaged with a series of methods to detect and mitigate bias through an interactive take-home tutorial, completed various exercises, and were subsequently interviewed. Our findings revealed that participants encountered a variety of unresolved conflicts, including selecting relevant bias metrics, tradeoffs between different bias metrics, and a tension between relying on their own opinions and stereotypes rather than on the data and metrics. However, despite these challenges, participants demonstrated a sensitivity to bias by correctly identifying biases in

two datasets used by the fairness community while reflecting on incorporating bias mitigation methods within their own ML processes. Finally, starting from these findings, we identified a series of implications for operationalizing bias, such as leveraging the overlaps between bias concepts and ML practices and for the effective design of bias mitigation tools. As the number of domains where ML is being applied increases, it becomes increasingly important for ML users to identify and mitigate bias. This is a timely challenge for the HCI community, as the misapplication of ML might lead to detrimental consequences for disadvantaged individuals or groups.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. *Machine bias. ProPublica* 23, 1 (2016), 139–159. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K. Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

[3] Ricardo Baeza-Yates. 2018. Bias on the web. *Communications of the ACM* 61, 6 (2018), 54–61.

[4] Agathe Balayn, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju .2023. "Fairness Toolkits, A Checkbox Culture?" On the factors that fragment developer practices in handling algorithmic harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 482–495.

[5] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2021. It's COMPASlicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, Vol. 1. MIT Press. Retrieved from https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/92cc227532d17e56e07902b254dfad10-Paper-round1.pdf

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. Fairness and Machine Learning. fairmlbook.org, Retrieved from http://www.fairmlbook.org

[7] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences* 116, 32 (2019), 15849–15854.

[8] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4:1–4:15. DOI: https://doi.org/10.1147/JRD.2019.2942287

[9] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44.

[10] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, 514–524.

[11] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A Toolkit for Assessing and Improving Fairness in AI. *Technical Report MSR-TR-2020-32. Microsoft*. Retrieved from https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[12] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. DOI: https://doi.org/10.1191/1478088706QP063OA

[13] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597.

[14] Michael Buckland and Fredric Gey. 1994. The relationship between recall and precision. *Journal of the American Society for Information Science* 45, 1 (1994), 12–19.

[15] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. 2021. Soliciting stakeholders' fairness notions in child maltreatment predictive systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, 1–17.

[16] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.

[17] Robert Cinca, Enrico Costanza, Mirco Musolesi, and Muna Alebri. 2025. "What are they not telling me?" Learning machine learning: Understanding the challenges for novices. *International Journal of Human-Computer Studies* 196 (2025), 103438. DOI: https://doi.org/10.1016/j.ijhcs.2024.103438

[18] Victoria Clarke and Virginia Braun. 2013. Teaching thematic analysis: Overcoming challenges and developing strategies for effective learning. *The Psychologist* 26, 2 (2013), 120–123.

[19] Tom Cole and Marco Gillies. 2022. More than a bit of coding:(un-) grounded (non-) theory in HCI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems Extended Abstracts*. ACM, 1–11.
[20] European Commission. 2024. Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Retrieved form https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689
[21] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 797–806.
[22] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. 2012. Social coding in GitHub: Transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. ACM, 1277–1286.
[23] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring how machine learning practitioners (try to) use fairness toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 473–484.
[24] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc* 1 (2016), 1–39.
[25] Dheeru Dua and Casey Graff. 2017. UCI machine learning repository. Retrieved from http://archive.ics.uci.edu/ml
[26] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 214–226.
[27] Golnoosh Farnadi, Behrouz Babaki, and Lise Getoor. 2018. Fairness in relational domains. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 108–114.
[28] Sina Fazelpour and Zachary C. Lipton. 2020. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AAAI, 57–63.
[29] Isabella Ferreira, Jinghui Cheng, and Bram Adams. 2021. The "shut the f** k up" phenomenon: Characterizing incivility in open source code review discussions. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–35.
[30] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM* 64, 4 (2021), 136–143.
[31] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. arXiv:1803.09010. Retrieved from https://arxiv.org/abs/1803.09010
[32] Google. 2020. Tensorfow's fairness evaluation and visualization toolkit. Retrieved from https://github.com/tensorflow/fairness-indicators
[33] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *Proceedings of the NIPS Symposium on Machine Learning and the Law*, Vol. 1. ACM, 2.
[34] Chowdhury Mohammad Rakin Haider, Christopher Clifton, and Ming Yin. 2024. Do crowdsourced fairness preferences correlate with risk perceptions?. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. ACM, 304–324.
[35] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, Vol. 29, 3315–3323.
[36] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–16.
[37] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction 2*, CSCW (2018), 1–22.
[38] Kristin M. Kostick-Quenet, I. Glenn Cohen, Sara Gerke, Bernard Lo, James Antaki, Faezah Movahedi, Hasna Njah, Lauren Schoen, Jerry E. Estep, and J. S. Blumenthal-Barby. 2022. Mitigating racial bias in machine learning. *Journal of Law, Medicine & Ethics* 50, 1 (2022), 92–100.
[39] Cassie Kozyrkov. 2018. Advice for finding AI use cases. Retrieved from https://hackernoon.com/imagine-a-drunk-island-advice-for-finding-ai-use-cases-8d47495d4c3f
[40] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 5686–5697. DOI: https://doi.org/10.1145/2858036.2858529
[41] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more?: The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, 1–10. DOI: https://doi.org/10.1145/2207676.2207678

[42] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*. Curran Associates Inc., 4069–4079. DOI: https://doi.org/10.5555/3294996.3295162

[43] Michelle Seng,Ah Lee, and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13.

[44] Annie Liang, Jay Lu, and Xiaosheng Mu. 2021. Algorithm design: A fairness-accuracy frontier. arXiv:2112.09975. Retrieved from https://doi.org/10.48550/arXiv.2112.09975

[45] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, 2119–2128. DOI: https://doi.org/10.1145/1518701.1519023

[46] Zachary C. Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57. DOI: https://doi.org/10.1145/3236386.3241340

[47] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the fairness of AI systems: AI practitioners' processes, challenges, and needs for support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–26.

[48] Ivy W. Maina, Tanisha D. Belton, Sara Ginzberg, Ajit Singh, and Tiffani J. Johnson. 2018. A decade of studying implicit racial/ethnic bias in healthcare providers using the implicit association test. *Social Science & Medicine* 199 (2018), 219–229.

[49] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[50] Michele Merler, Nalini Ratha, Rogerio S. Feris, and John R. Smith. 2019. Diversity in faces. arXiv:1901.10436. Retrieved from https://arxiv.org/abs/1901.10436

[51] Melanie Mitchell. 2019. *Artificial Intelligence: A Guide for Thinking Humans*. Penguin, United Kingdom.

[52] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 220–229.

[53] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8, 1 (2021), 141–163.

[54] Yuri Nakao, Lorenzo Strappelli, Simone Stumpf, Aisha Naseer, Daniele Regoli, and Giulia Del Gamba. 2023. Towards responsible AI: A design space exploration of human-centered artificial intelligence user interfaces to investigate fairness. *International Journal of Human–Computer Interaction* 39, 9 (2023), 1762–1788.

[55] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. 2022. Toward involving end-users in interactive human-in-the-loop AI fairness. *ACM Transactions on Interactive Intelligent Systems* 12, 3 (2022), 1–30.

[56] Mark J. Nelson and Amy K. Hoover. 2020. Notes on using Google Colaboratory in AI education. In *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education*. ACM, 533–534.

[57] Safiya U. Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

[58] Donghao Ren, Saleema Amershi, Bongshin Lee, Jina Suh, and Jason D. Williams. 2016. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2016), 61–70. DOI: https://doi.org/10.1109/TVCG.2016.2598828

[59] Brianna Richardson, Jean Garcia-Gathright, Samuel F. Way, Jennifer Thom, and Henriette Cramer. 2021. 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*. ACM, 1–13.

[60] Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. 2010. Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data* 4, 2 (2010), 1–40.

[61] Debjani Saha, Candice Schumann, Duncan Mcelfresh, John Dickerson, Michelle Mazurek, and Michael Tschantz. 2020. Measuring non-expert comprehension of machine learning fairness metrics. In *Proceedings of the International Conference on Machine Learning*. PMLR, 8377–8387.

[62] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. 2018. *Aequitas: A bias and fairness audit toolkit*. arXiv:1811.05577. Retrieved from https://doi.org/10.48550/ARXIV.1811.05577

[63] Nripsuta Ani Saxena. 2019. Perceptions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 537–538.

[64] Jessie J. Smith, Lex Beattie, and Henriette Cramer. 2023. Scoping fairness objectives and identifying fairness metrics for recommender systems: The practitioners' perspective. In *Proceedings of the ACM Web Conference 2023*. ACM, 3648–3659.

[65] Kacper Sokol, Alexander Hepburn, Rafael Poyiadzi, Matthew Clifford, Raul Santos-Rodriguez, and Peter Flach. 2020. FAT forensics: A python toolbox for implementing and deploying fairness, accountability and transparency algorithms in predictive systems. *Journal of Open Source Software* 5, 49 (2020), 1904.

[66] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. ACM, 2239–2248.

[67] Ramya Srinivasan and Ajay Chander. 2021. Biases in AI systems: A survey for practitioners. *Queue* 19, 2 (2021), 45–64.

[68] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2459–2468.

[69] VentureBeat Staff. 2019. Why do 87% of data science projects never make it into production. Retrieved from https://venturebeat.com/ai/why-do-87-of-data-science-projects-never-make-it-into-production/

[70] Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, 1–9.

[71] Abdel Aziz Taha and Allan Hanbury. 2015. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging* 15, 1 (2015), 1–28. DOI: https://doi.org/10.1186/s12880-015-0068-x

[72] Laura Valentini. 2012. Ideal vs. non-ideal theory: A conceptual map. *Philosophy Compass* 7, 9 (2012), 654–664.

[73] Rama Adithya Varanasi and Nitesh Goyal. 2023. "It is currently hodgepodge": Examining AI/ML practitioners' challenges during co-production of responsible AI values. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, 1–17.

[74] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, 1–14. DOI: https://doi.org/10.1145/3173574.3174156

[75] Kerstin N. Vokinger, Stefan Feuerriegel, and Aaron S. Kesselheim. 2021. Mitigating bias in machine learning for medicine. *Communications Medicine* 1, 1 (2021), 1–3.

[76] Joyce Weiner. 2020. Why AI/data science projects fail: How to avoid project pitfalls. *Synthesis Lectures on Computation and Analytics* 1, 1 (2020), i–77.

[77] Doris Xin, Litian Ma, Jialin Liu, Stephen Macke, Shuchen Song, and Aditya Parameswaran. 2018. Accelerating human-in-the-loop machine learning: Challenges and opportunities. In *Proceedings of the Second Workshop on Data Management for End-to-End Machine Learning*. ACM, 1–4.

[78] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding interactive machine learning tool design in how non-experts actually build models. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, 573–584. DOI: https://doi.org/10.1145/3196709.3196729

Received 24 June 2024; revised 26 March 2025; accepted 9 April 2025