A Systematic Review on Human Roles, Solutions, and Methodological Approaches to Address Bias in AI

AMAL HASHKY, University of Florida, United States
ERIC D. RAGAN, University of Florida, United States

People play a significant role in designing, developing, and employing artificial intelligence (AI) systems. They can consider contextual information beyond the scope of AI models, thereby influencing system outcomes. At the same time, people's choices or biases can introduce problems into the systems. This paradoxical scenario, in which people can both introduce and contribute to relieving the inherited machine bias, demands comprehensive and multidisciplinary approaches involving informed human interventions to improve systems' performances and reduce their biases. Researchers across various communities have investigated multifaceted methods to reduce and mitigate bias in AI systems. Regardless of the method, humans are always involved in the debiasing method in one way or another, emphasizing the importance of human intervention during AI systems development. In this systematic review, we analyzed 100 peer-reviewed publications from various human-computer interaction (HCI) and machine learning (ML) venues. We discuss their research efforts to minimize data bias and algorithmic bias from three angles. First, we present a comprehensive taxonomy of bias mitigation solutions, analyzing the research methodologies and standard benchmarks for evaluating these solutions, highlighting the human researcher's role in developing and evaluating solutions to address bias. Next, we identify humans' roles in alleviating biases and specify how, when, and where their involvement occurs within the AI lifecycle. Finally, we summarize the research focus and methodologies across research disciplines. Our review revealed that, while technical solutions are essential, addressing bias requires a broad perspective that integrates human oversight, ethical frameworks, and interdisciplinary collaboration.

CCS Concepts: • **General and reference → Surveys and overviews**; • **Computing methodologies → Artificial intelligence**; **Machine learning**; • **Human-centered computing**;

Additional Key Words and Phrases: Trustworthy artificial intelligence, Fairness, Robustness, Bias mitigation, Human-centered AI

ACM Reference Format:
Amal Hashky and Eric D. Ragan. 2026. A Systematic Review on Human Roles, Solutions, and Methodological Approaches to Address Bias in AI. 1, 1 (January 2026), 35 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

People today heavily rely on intelligent systems for everyday decisions, data and algorithmic biases have become a critical concern. From trivial cases like TV show recommendations through more significant scenarios such as job recruitment and loan systems to highly consequential issues related to legal systems, the unintentional consequences of embedded biases can perpetuate inequalities and hinder societal fairness. Although identifying the exact origins of these biases could be challenging, some researchers pinpointed data as a primary source, mainly through representation bias, where data may misrepresent certain groups [146, 173]. Such bias can stem from several factors, including selection and

Authors' Contact Information: Amal Hashky, University of Florida, United States, ahashky@ufl.edu; Eric D. Ragan, University of Florida, United States, eragan@ufl.edu.

Manuscript submitted to ACM

sampling biases during data collection and labeling biases in data annotation (e.g., [48, 122]). Algorithms themselves can further amplify these biases if not adequately trained and evaluated (e.g., [76, 86, 109]).

One integral aspect to consider while addressing bias is the role of human involvement in the AI lifecycle. Humans are significantly involved in designing, developing, and deploying machine learning algorithms [30]. Their decisions during each phase could be both a source of bias and a potential solution for mitigating it. For instance, ML practitioners make decisions affecting data curation and model development; it is essential that they thoroughly understand the problem requirements to implement proper data sampling and balanced data annotation strategies (e.g., [37, 43, 44]), and employ tools to counterbalance biases introduced by annotators (e.g., [64, 142, 149]). Furthermore, selecting appropriate algorithms during development can prevent amplifying hidden biases (e.g.,[60, 78, 85, 86]). Human oversight can significantly mitigate biases during deployment by overriding these systems in fully automated workflows, where decisions are instantly employed [24, 29]. Additionally, leaders' and policymakers' decisions to adjust workflows and enforce policies within the AI lifecycle are essential depending on the context of AI system deployment (e.g.,[48, 74]). In all these scenarios, human decisions can either restrain biases and increase systems' trustworthiness or amplify them, undermining their fairness toward specific groups or threatening their robustness when seeing new data.

Recognizing the need for fair outcomes and robust systems, various ethical bodies and research communities have invested significant effort in promoting trustworthy AI [161]. For example, the European Union (EU) has proposed ethical guidelines to foster the development and deployment of trustworthy AI systems [141], outlining the following seven requirements:

> *"(1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability."*

Motivated by these requirements, countless researchers from different disciplines have explored diverse solutions to mitigate bias in AI systems, each tailored to specific goals and research approaches. Throughout our review, we observed that researchers frequently use the terms *mitigate* and *reduce* bias interchangeably, despite their subtle differences in meaning in English—where *mitigate* suggests lessening the impact or severity of bias [35], and *reduce* typically implies a decrease in the extent or amount of bias [36]. To maintain consistency and align with the usage in the literature, we have adopted this convention throughout our work. Some researchers have adopted human-computer interaction (HCI) strategies, encouraging participatory design methods to establish design principles and recommendations and developing tools (e.g., [3, 55, 74, 87, 146]). In contrast, others have focused on the technical, data-driven aspects typical of the machine learning (ML) field (e.g., [25, 50, 70, 90, 154, 168]).

Addressing bias in AI systems requires active involvement and collective responsibility from humans involved. It begins with AI researchers who develop these solutions and extends to ML practitioners and other stakeholders. Motivated by this collective responsibility, we present this systematic review, adopting a comprehensive and multidisciplinary approach to exploring the state-of-the-art literature on addressing data and algorithmic bias in AI systems. While several systematic reviews address bias in AI systems, our work provides a unique contribution by adopting a human-centric perspective, unlike most existing reviews, which primarily focus on technical bias mitigation strategies. In our review, we systematically examine the roles humans play (governance, technical, operational) in reducing bias across the AI lifecycle, including a novel multidimensional classification of their roles and scopes of influence, mapping them to debiasing methods. Further, we analyze the reviewed debiasing interventions within the design, data acquisition,

modeling, and deployment cycles—a level of lifecycle granularity often missing in other surveys, which tend to focus on modeling stage interventions alone, providing an academically-comprehensive consolidated reference.

We reviewed 100 research publications from 2018 to 2024, drawing from top-tier conferences in both HCI and ML fields. This focus was chosen to manage the scope of our review while ensuring coverage of venues most relevant and highly rated for this topic. We clarify that our review is intended as a representative sample rather than a comprehensive survey of all possible relevant publications, and we acknowledge that journals and other venues indexed in databases would also be relevant to the topic. Our analysis examines the proposed solutions to address bias and the nuances of human involvement in employing these solutions. Additionally, we provide insights into the distinct motivations and approaches the ML and HCI fields take to address these challenges.

We summarize our key contributions below:

- We present a detailed taxonomy of solutions to address bias, categorizing the research methodologies and outlining standard evaluation benchmarks.
- We define humans' roles within the AI lifecycle and classify the extent of their involvement in minimizing data and algorithmic biases.
- We illustrate differences in the motivations and research methodologies employed to investigate solutions for bias across the ML and HCI fields.

By classifying existing solutions, we aim to provide researchers with a comprehensive overview, showing the breadth of solutions in state-of-the-art literature in HCI and ML disciplines and future interdisciplinary opportunities. Situating the humans' roles and contributions within these existing solutions highlights better opportunities for actively integrating humans into designing and implementing ethical AI frameworks.

The remainder of this article is organized as follows. Section 2 describes the methodology of our systematic review, including the search strategy, screening process, and coding approach. Sections 3–5 present the results of our analysis. Section 3 examines methods for mitigating bias in AI, classifying their types and detailing the methodologies for development and evaluation. Section 4 explores human roles in bias mitigation throughout the AI lifecycle, categorizing both the roles and the extent of involvement. Section 5 compares the motivations and methodologies of the ML and HCI research communities. Finally, Section 6 discusses the main challenges to address bias, offers considerations for effectively leveraging human roles, and summarizes key insights that extend beyond technical solutions.

2 Method

We adhered to a rigorous systematic literature review approach following the PRISMA 2020 statement [114]. We further used a systematic review tool, *Covidence* [27], to facilitate the reviewing process. Figure 1 shows the PRISMA flowchart of our research method. Since our inclusion criteria and coding depend on how bias is understood, we first clarify its meaning for the scope of this review. We then outline the methodology, starting with the eligibility and inclusion criteria, proceeding to the data collection strategy, and concluding with the data analysis procedures.

2.1 What is Bias?

In general terms, *bias* refers to a preference towards someone or something. Across academic disciplines, bias is defined in various ways. In psychology, it is understood as systematic cognitive distortions that deviate from rational judgment [53]. In statistics, bias represents a systematic deviation introduced when data collection or estimation methods yield

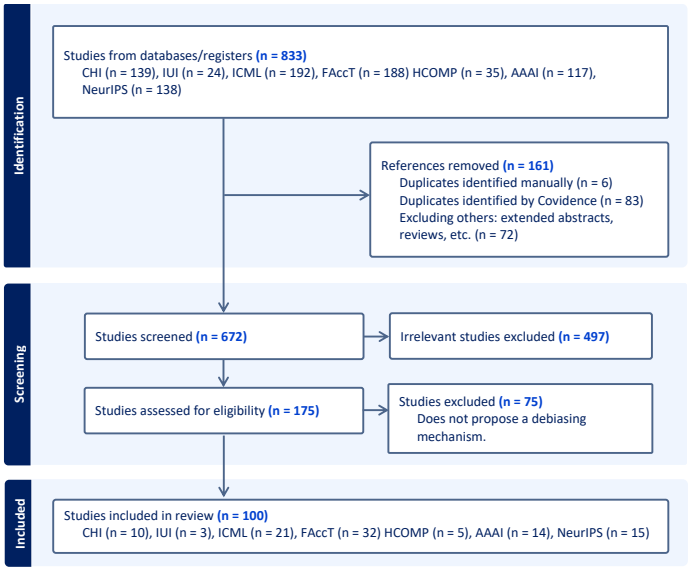

Fig. 1. An overview of our research method following *PRISMA*.

different outputs than what is expected [80]. In machine learning, Mitchell [106] defines bias as "any basis for choosing one generalization over another, other than strict consistency with the observed training instances."

For the context of our review, it is important to seek a broader definition that encompasses the various causes of bias in AI, one that integrates technical, cognitive, and societal perspectives, as well as the human role in the process. Bias in AI systems can emerge from multiple sources. At the data level, representation bias occurs when training datasets fail to accurately capture the diversity of real-world populations, resulting in the over- or under-representation of certain groups. Other sources, such as sampling, measurement, and annotation bias, may further skew datasets during collection and labeling. At the algorithmic level, model design choices and optimization strategies may unintentionally amplify hidden biases present in data, producing outputs that reinforce disparities. Even when algorithms are technically fair, the context of their deployment may introduce societal and institutional biases that reflect broader structural inequities. It is essential to acknowledge that bias is not inherently limited to technical artifacts; it often reflects human decisions and assumptions made throughout the AI lifecycle. From problem formulation and data curation to model development and deployment, human actors influence how bias manifests and how it might be mitigated. Thus, addressing bias requires not only computational solutions but also ethical frameworks, participatory practices, and governance mechanisms that take into account contextual and societal dimensions.

Various recent surveys and review articles have also classified different types of bias in AI, offering taxonomies that complement our framework (e.g., [20, 102, 104, 112, 146]). Readers seeking broader mappings of bias categories are encouraged to review their work. Bringing these strands together, for the purposes of our review, we define bias in AI as:

"Systematic and unfair deviations in data or algorithmic outcomes that disadvantage particular individuals or groups. Unlike random errors, bias reflects consistent distortions that can undermine robustness [1] *across diverse contexts, perpetuate*

[1]Here, robustness refers to the ability of AI systems to maintain consistent performance across diverse populations, contexts, and environments.

existing societal inequities, or create new forms of unfairness when embedded in automated decision-making systems. Bias is often shaped and perpetuated by human actors through decisions made during problem formulation, data collection, model design choices, and deployment".

This definition guided both our eligibility criteria and our coding framework during data analysis presented in the subsections below.

2.2 Eligibility and Inclusion Criteria

Following specific criteria to determine paper eligibility for our review, we included peer-reviewed full conference papers published between 2018 and 2024 that proposed a solution to address bias at entire or any stage of the AI lifecycle, including: 1) non-technical frameworks, design principles, and recommendations or 2) technical solutions, such as algorithmic and data-related methods, and 3) targets increasing models' fairness or robustness. At the same time, we excluded papers that addressed human cognitive biases, such as confirmation bias and anchoring bias, except those examining the impact of biased human feedback in AI systems through human-in-the-loop and interactive machine learning. We also excluded papers that solely introduced new bias detection methods or fairness auditing without presenting concrete methods to reduce bias.

2.3 Data Collection

Our discussion on bias in AI naturally suggests including papers from AI publication venues. However, one of our primary motivations is to identify human involvement in addressing bias. Therefore, we decided to include human-centered research to acquire a comprehensive understanding and diverse perspective. This multidisciplinary approach combines the deep understanding of user interaction and design principles from HCI with AI's technical solutions and data-driven methodologies, allowing us to exhaustively analyze the various human roles in bias mitigation across the AI lifecycle. This review spans various top-tier conferences in the AI and HCI communities, precisely full research articles published between 2018-2024 in four ACM Conferences: the Conference on Human Factors in Computing Systems (**CHI**), the International Conference on Machine Learning (**ICML**), Intelligent User Interfaces (**IUI**) and Fairness, Accountability, and Transparency (**FAccT**). Two AAAI conferences: Conference on Artificial Intelligence (**AAAI**) and the Human Computation and Crowdsourcing (**HCOMP**). In addition to the Conference on Neural Information Processing Systems (**NeurIPS**). This list of venues is not exhaustive and does not cover the depth of the body of research addressing bias; however, it spans a diverse collection of top-tier and middle-tier conferences that cover the scope of our review, providing a representative sample of the state-of-the-art existing research. Our sample from seven venues over the past six years reflects a substantial increase in publications in recent years, as illustrated in Figure 2.

We searched two digital libraries and two archives (*ACM Digital Library*, *Scopus*, *AAAI*, and *NeurIPS*), resulting in 833 initial articles. In our query, we used the terms (*"Bias"*, *"Data bias"* OR *"Algorithmic bias"*), and we also added (AND (*"Machine learning"* OR *"Artificial Intelligence"* OR *"AI"*)) to narrow down the research results, especially in HCI publication venues. Although one of our goals is to identify human roles, we intentionally avoided using keywords related to humans to prevent limiting our search results. From an earlier informal exploration, we observed that authors might not explicitly mention human roles, and adding specific human-related keywords to our search queries could lead to excluding relevant studies. Relying on the keywords above ensured a more comprehensive and inclusive search, capturing a wider range of papers that indirectly address human involvement through broader contexts.

After importing the initial list of 833 papers into *Covidence*, we removed 161 duplicates and excluded extended abstracts and review articles, resulting in 672 full-paper articles. Following a systematic review approach, we filtered

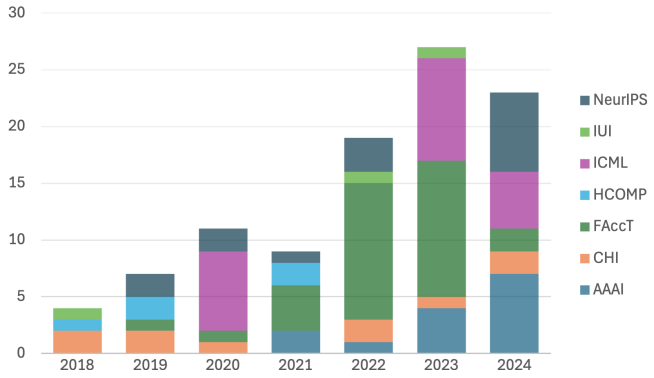

Fig. 2. Number of publications for each venue per year.

Table 1. Filtering process for selected papers in seven Conferences between 2018-2024.

Filtering Process	**Publication Venues**							
	ACM CHI	*ACM IUI*	*ACM ICML*	*ACM FAccT*	*AAAI HCOMP*	*AAAI*	*NeurIPS*	*Total*
Potential papers after keyword searching	139	24	192	188	35	117	138	833
Potential papers excluding duplicates & others	77	10	191	134	18	104	138	672
Relevant papers after title and abstract screening	19	5	41	52	8	26	24	175
Included papers after full-text review	**10**	**3**	**21**	**32**	**5**	**14**	**15**	**100**

the papers through two main steps: 1) title and abstract screening and 2) full-text review. In the first step, we manually screened the titles and abstracts of all papers to identify those relevant to our review's scope, narrowing the selection to 175 articles. We then conducted a full-text review and included or excluded papers based on the eligibility criteria outlined in the previous Section 2.2, ultimately selecting 100 papers for inclusion in our review. The filtering process is summarized in Table 1.

2.4 Data Analysis

We used the *Covidence* tool for systematic review to facilitate the coding and analysis, enabling us to efficiently manage and organize the large volume of papers. Our analysis followed an adaptive iterative coding classification and a thematic approach. We conducted an exploratory review during the screening and full-text review phases to identify initial labels and possible themes consulted from previous literature reviews. Driven by our research questions and the initial labels we gathered earlier, we followed a **deductive coding** [42] —in which coding categories are developed in advance based on theory, prior literature, or predefined frameworks approach—to create a priori codes and themes; these covered different aspects of the proposed debiasing solutions including their types and evaluation methods, human intervention characteristics, such as roles and levels of involvement, and research motivations and methodologies.

While deductive coding provided a structured, quick starting point for classifying the papers, it did not capture all existing categories. Therefore, we incorporated an **inductive coding** approach [42] —where coding categories emerge directly from the data through iterative reading and interpretation—during data extraction to identify new codes and update our initial themes accordingly. Once new codes were identified, we revisited previously labeled papers for consistency in our categorization. This iterative hybrid coding approach allowed us to refine and expand the codes incrementally until all papers were appropriately categorized and themes identified. This process ensured our analysis accurately captured all elements related to our research questions, including the diverse debiasing solutions and evaluations across the AI lifecycle, the aspects of human involvement, and research motivations and methodologies.

3 Solutions and Methods for Addressing Bias in AI

Due to the complexity of identifying bias and its sources, the convoluted process of building AI systems, and the variety of individuals involved with conflicting objectives, researchers from various communities have approached bias in AI from different angles; moreover, their distinct research interest, focus, scope, and goals shape their research methods and contributions differently. The multidisciplinary nature of existing research in this area highlights the need for a comprehensive classification to showcase all the diverse solutions for tackling bias at different stages of the AI lifecycle.

As identifying human roles in mitigating bias is one of the key objectives of this review, we decided to recognize researchers' critical role in this process because their work is indispensable for developing and evaluating new debiasing methods. Although researchers' roles do not fall within the direct AI development lifecycle of a particular system, their contributions are crucial to providing various solutions for all other individuals involved. Therefore, we categorize their involvement from a research perspective under two main areas: 1) types of debiasing solutions they propose and 2) research and evaluation methodologies they adopt.

In the subsequent section (Section 4), we will analyze additional human roles with their unique contributions and expertise to apply at different phases of the AI lifecycle. Unlike researchers, these roles are directly engaged in one or more of the systems' lifecycles and collectively contribute to reducing bias through applying debiasing solutions.

In the current section, we present the main findings of our systematic review, classifying the bias mitigation solutions and situating them within the AI lifecycle. We examine each category's scope with representative examples from the literature (Section 3.1). We also summarize the evaluation methods used, mapping them to the mitigation solution categories (Section 3.2).

3.1 Types of Solutions for Algorithmic and Data Bias

Based on our review of the literature, solutions to addressing bias tend to fall into one of these categories: *1) principle and design guidelines, 2) non-algorithmic frameworks including structural, ethical, and others, 3) algorithmic solutions including algorithms and algorithmic frameworks, and 4) tools and techniques.* These solutions vary in their scope and implementation requirements. While some focus strictly on technical adjustments, others include broader structural and organizational reforms. We explain each of these solution types below. For a clearer understanding of our explanation of the types below, Table 2 classifies the papers according to the stage of the AI lifecycle where the debiasing solution is applied and Figure 3 shows the distributions of these solutions based on the AI lifecycle.

3.1.1 Principles and Design Guidelines. From our review to papers in this category, we describe *principles and design guidelines* as high-level, value-driven recommendations aimed at embedding fairness, robustness, and inclusivity into AI systems. They set the overarching ethical, legal, and societal goals while leaving flexibility in how these goals are

Table 2. Classification of papers according to the stage of the AI lifecycle where the debiasing solution is applied. As some methods span multiple stages, certain papers are listed in more than one category.

	Principles & Design Guidelines	Non-Algorithmic Frameworks	Algorithmic Solutions	Tools & Techniques
Design	[24, 29, 43, 44, 57, 65, 74, 87, 119, 130–132, 144, 153, 164]	[11, 28, 103, 116, 147]	[21, 85, 93, 109, 118, 158, 165, 167, 172]	[37]
Data Curation	[43, 44, 57, 63, 119, 132, 153]	[11, 116, 147]	[85, 93, 170]	[33, 37, 64, 142, 149]
Development	[44, 87]	[103]	[2, 4, 6, 9, 10, 13–16, 18, 21–23, 25, 26, 34, 46–48, 50, 54, 60, 62, 67, 68, 70–72, 76, 78, 85, 86, 89–92, 98, 109–111, 115, 117, 118, 120, 122, 126–129, 135, 143, 150, 151, 154–156, 158–160, 162, 163, 165–169, 171–174]	[3, 55]
Deployment	[24, 29, 43, 119]	[28, 103, 147]	-	-

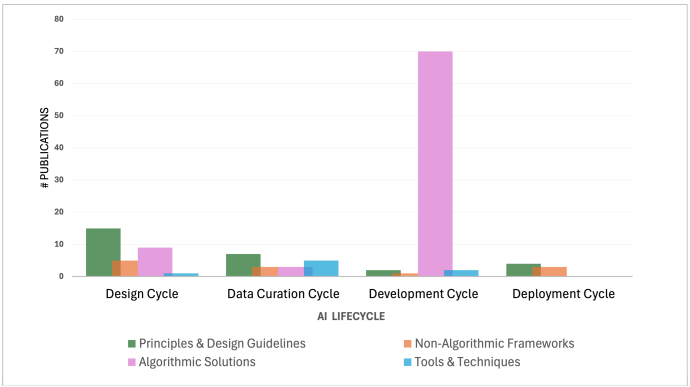

Fig. 3. The distributions of solutions to address data and algorithmic bias in the AI lifecycle

operationalized. They typically serve as high-level reminders of best practices, but do not provide a specific set of steps or instructions for practitioners to follow. Such guidelines typically serve as a conceptual guide for practitioners, influencing decisions across the AI lifecycle without prescribing a rigid process. They can be domain-specific or broadly applicable, and often require translation into more concrete, context-specific actions before implementation. This category includes 16% of the papers in our review, outlining foundational rules and design recommendations grounded in fairness and robustness standards [24, 29, 43, 44, 57, 63, 65, 74, 87, 119, 130–132, 144, 153, 164]. While often framed as theoretical constructs, these principles and guidelines are also intended to inform practice and can guide practitioners across various stages. Some researchers took a comprehensive approach to embed these standards beyond technical adjustments and addressed one or more non-technical elements of AI system design, such as legal,

ethical, and broader societal values, highlighting the need to extend the focus beyond developing fair algorithms and to consider non-technical elements of the whole AI lifecycle. For instance, instead of fixing the algorithm, Huang et al. [63] recommended addressing the biases inherent in the design process by promoting social inclusion practices and improving diversity and accessibility during data curation. Drawing inspiration from the museum sector, they proposed non-technical guidelines focused on societal values concerning practitioners, community engagement, and situational contexts. Similarly, Katell et al. [74] encouraged considering social and political contexts to reduce bias through situated interventions, such as including the community in participatory and co-design methods. Another paper [131] discussed technical and theoretical constructs that address "racial talk" when creating AI chatbots. Focusing on designing data crowdsourcing pipelines, research published by Wang et al. [153] presented recommendations to facilitate organizational and structural adjustments that consider annotators' well-being. Likewise, Sengupta et al. [132] proposed recommendations that account for biased results from skewed annotator populations.

In the above examples, researchers articulated the proposed principles and guidelines with a high level of granularity. While significant, practitioners must still translate these into specific domains and contexts to develop detailed guidelines and recommendations. For example, Freeman et al. [43] focused on improving data quality for medical imaging by drawing principles from the commodity crowdsourcing literature and mapping them to clinical needs through a collaborative, iterative design process with domain experts. Another study [44] presented nuanced guidelines for creating disability-centered datasets as a resource for developing disability-positive large language models (LLM).

Others articulated tailored technical recommendations. For instance, Levonian et al. [87] proposed detailed design guidelines for modelers and designers regarding developing interactive machine-learning systems for text annotation tasks. These guidelines aim to maximize the leverage of human input to improve the learning rate during active learning at a lower cost. Additionally, research published in [57, 65] presented domain-specific recommendations focused on eliminating racial bias in specific applications like automatic speaker recognition and visual question-answering systems.

Generally, while principles and recommendations affect decisions made during the design process of the AI lifecycle, their implications may extend to reforming institutional policies, organizational structures and operational workflows.

3.1.2 Non-Algorithmic Frameworks. Researchers have also introduced non-algorithmic frameworks that target specific bias problems and design contexts, providing practical guidance for AI development. Based on papers in this group, we interpret non-algorithmic frameworks as structured, operational tools that translate high-level principles into specific, actionable workflows or methodologies for mitigating bias in AI systems. In contrast with the previous category of *principles and guidelines*, which are aspirational and more general, these frameworks prescribe a precise sequence of steps, roles, and procedures that practitioners can directly apply. They tend to operate at a lower level of abstraction, providing concrete pathways for consistent execution, and often include mechanisms for evaluation, iteration, or stakeholder engagement. In our sample, only 5 of the 100 (5%) papers proposed some non-algorithmic framework.

For instance, research published in [11, 147] presented frameworks to guide ML practitioners during the data curation phase. Suresh et al's. [147] framework is grounded in feminism and supports iterative data collection and annotation through participatory and co-design processes. Their approach exemplifies how high-level principles can be translated into structured, operational practices for addressing bias. The approach prescribes an iterative data collection and annotation workflow that responds to observed model weaknesses and explicitly interrogates framing decisions, such as who is included or excluded in definitions of feminicide. This process defines concrete roles for practitioners, domain experts, and community stakeholders, and establishes procedures for revisiting labels, categories, and data sources over time. In addition, the framework provides actionable guidance for prioritizing marginalized groups in both data

construction and analysis and direct practitioners to focus on intersectional identities rather than statistical majorities. In another example of a non-algorithmic framework, Barbosa et al. [11] presented a framework based on ethics, focusing on managing annotation task allocations while considering human factors such as annotators' well-being. Similarly, McCradden et al. [103] presented an ethical framework supporting the integration of ML into clinical practices, from design to deployment, while promoting fair clinical operations and outcomes for patients from different social groups.

Other researchers in [116] developed a generalizable framework for auditing the data annotation pipeline. This framework considers cultural and linguistic differences across annotators, promoting inclusivity of diverse cultural, geographical, and demographic backgrounds when recruiting annotators and producing labels with uniform conceptualization. Additionally, a framework proposed by CruzCort et al. [28] advocates for employing structural changes; their framework (RISE) is established on four principles: Reformulate, Identify, Structuralize, and Expand. It takes a broader approach to consider sources of harm during problem formulation, analysis, and stakeholder identification.

3.1.3 Algorithmic Solutions. With the highest number of papers (72 out of 100, representing 72%), this category includes research proposing methods to reduce bias through algorithmic means, spanning both algorithms and algorithmic frameworks. Among these, 14 of the 72 papers discuss using algorithmic frameworks [2, 21, 67, 85, 93, 109, 118, 158, 162, 165, 167, 169, 170, 172]. We are not categorizing the papers based on these distinctions; instead, we clarify the differences below for the reader's understanding. The primary distinction lies in their scope and application: an algorithmic framework provides broad, generalized principles applicable across various models to maintain standards of fairness and robustness. Conversely, a specific bias-reducing algorithm typically targets a particular model, offering focused solution designed to minimize bias in data processing or decision-making.

Drawing from other classification schemes in previous surveys (e.g., [20, 88, 113]), we categorize these methods into **preprocessing**, **in-processing** [2], and **post-processing** approaches. Researchers have explored a wide array of techniques within each of these methods. For a detailed examination, we encourage readers to consult targeted surveys dedicated to this topic (e.g., [61, 75, 104, 139, 152]). Below, we offer a concise overview of these methods, highlighting some of the examples from our review.

Preprocessing methods. Research has consistently shown how biased datasets can adversely affect outcomes, threatening system robustness when encountering new data or diminishing fairness towards sensitive groups [104]. Typically, such bias arises from an imbalanced class representation, creating a propensity to favor the overrepresented group, often referred to as representation bias. This bias can stem from various sources, including inherited socio-technical issues in the worldviews, skewed distributions, or flawed sampling strategies [133]. One might think that the way to fix underrepresented data is to collect more data for the less representative class; however, data collection and annotation are not always possible for many reasons, such as being both time-consuming and cost-infeasible [100]. Therefore, several researchers employed different techniques to 'fixing' the data and mitigate bias before the training phase.

Of the 72 papers in the algorithmic solutions category, 19 proposed mitigating bias using preprocessing techniques. One simple option involves rebalancing the dataset through *augmentation*, creating synthetic samples to counterbalance the representation of different subgroups (e.g., [26, 68, 98, 117, 156]). *Resampling* represents another way to balance the classes representations, it involves selecting or removing a subset of the data for training the model instead of using the whole dataset, examples include research in [10, 85, 128]. Alternatively, the *reweighting* technique retain the

[2]While the term "Model Training stage" could be more descriptive, we adopt "in-processing" to remain consistent with the classification convention used in prior surveys to facilitates comparison across studies.

original data distribution but alter the significance of certain samples by assigning different weights to data records to emphasize underrepresented classes during the learning process (e.g., [4, 21, 86]).

Preprocessing methods do not require modifying the learning algorithm, making them broadly applicable across different models, especially for quantifiable data. However, they do not address biases emerging during training, such as ones related to spurious correlation [78, 109]. Also, if not carefully employed, they may increase the risk of overfitting, affecting the model's generalizability [77].

In-processing methods. Alternatively to preprocessing methods, in-processing methods focus on modifying the model to mitigate bias during training. These models continuously adjust their learning process by modifying their parameters to meet specific criteria; this dynamic feature makes them adaptable to address biases emerging during training, such as biased correlations or those caused by distribution shifts.

In the category of algorithmic solutions, 48 out of 72 papers proposed the reduction of bias via in-processing techniques. Some of the most prominent techniques fall under adjusted learning where the learning procedure is changed to mitigate bias (e.g., [71, 78, 90, 109, 118, 120, 154, 168]). Others researchers employed adversarial learning where two models are trained simultaneously: a classifier that predicts outcomes and an adversary model that learns to exploit fairness issues. These models compete against each other, with the adversarial model challenging the classifier to improve its fairness, resulting in enhanced overall performance (e.g., [25, 50, 127, 167]). Additionally, some methods leverage unlabeled data to learn fair representations (e.g., [122, 168]), while others focus on fairness in embedding learning for networks and graphs (e.g., [16, 76]). However, these methods are complex to implement and can reduce the overall accuracy, raising concerns about balancing the trade-off between accuracy and fairness. Additionally, they require access to the model architecture components which is not always possible.

Post-processing methods. These methods are designed to address bias post-model training and are particularly useful when previous measures to mitigate bias during data curation and model training fall short or when there are limitations on accessing data or model components to perform a pre-processing or in-processing techniques. With only 11 papers [34, 92, 110, 126, 129, 155, 158–160, 169, 172], this set of techniques represents the least explored in our sample.

Within this group, several studies employ optimization-based techniques. For example, DiCiccio et al. [34] apply threshold optimization to adjust the decision threshold to satisfy fairness criteria, while Nandy et al. [110] and Wang et al. [155] use ranking optimization to reorder recommendations in line with fairness objectives. Alternative approaches, such as those reported by researchers in these papers [92, 158–160, 169], implement constraints-based approaches typically on loss or objective functions to enforce fairness. A similar approach presented by Zhao et al. [172] uses constraint-based approaches for explanation fairness. A third type of interventions, exemplified in references [126, 129], utilizes influence-based approaches. Sattigeri et al. [129] modify the predictions at the instance level by dropping training points after estimating the statistical influence scores. While the methods introduced by Richardson et al. [126] involve resampling data points by applying one or more of the following: removing, relabeling, adding, or duplicating samples to modify the training data distribution. Although these are data-level interventions, which are more commonly used for pre-processing, the interventions in this specific case are guided by the identification of problematic training points after a model has already been trained. For this reason, we consider the use of data interventions in Richardson et al. [126] to fall under post-processing.

The efficiency of post-processing methods highly depends on the specific AI application and the desired fairness outcomes. Although these techniques can be beneficial in some situations, they may not fully address the underlying root cause for biases.

3.1.4 Tools and Techniques. This category of bias mitigation solutions includes all other techniques that do not align with the methods previously discussed. It contains methods with techniques requiring simple adjustments to the workflow or interactive tools that facilitate human input at one phase of the AI lifecycle. Our review includes seven papers in this category. Two of these studies presented interactive systems to reduce bias during system development [3, 55]. These two systems facilitate human control during the training and fine-tuning phases by manipulating specific model components, thereby reducing biases. Ahn et al. [3] identifies "blind spots" in the decision space of a model by visualizing the associations of concepts with their target classes, allowing practitioners to identify spurious associations and evaluate mitigation strategies to correct them. Similarly, the system in [55] guides the attention of a deep neural network by allowing users to modify the attention maps in real time, directing the model to focus on relevant and unbiased features.

The remaining five papers focus on supporting ML practitioners during the data curation phase. These studies examine straightforward techniques and tools to mitigate cognitive biases introduced by crowdsourcers and systematic biases inherent in the tools used. For instance, Draws et al. [37] proposed a practical and proactive approach, adapted from psychology, to review crowdsourcing tasks and identify prevalent cognitive biases before initiating the data collection process. This approach, with its immediate applicability, allows data crowdsourcing designers to adjust these tasks as necessary to mitigate bias beforehand. It can also be used retrospectively on collected datasets to identify the type of biases, aiding in the decision of effective mitigation interventions.

In another study by Hube et al. [64], the researchers proposed two proactive strategies based on social projection and promoting self-awareness to significantly reduce workers' biases during annotation tasks. Both strategies require minor modifications to the annotation tool by adding simple messages to the task description and periodic reminders throughout the task. These additions are designed to guide workers' thought processes in specific ways based on social projection; the first prompts workers to consider how their peers might label the same content. In contrast, the second one encourages them to reflect on the topic's controversial nature and consider how their personal views might bias their judgments. Additionally, Diaz et al. [33] studied age-related bias in sentiment analysis; they suggested a simple technique to mitigate underrepresentation bias by isolating age-related data in the training corpora. This approach helps specify the origins of output bias and evaluates the impact of specific data manipulations on reducing bias before model training.

Focusing on data aggregation methods, Thebault-Spieker et al. [149], suggested that aggregating judgments from heterogeneous workers in political content moderation may significantly mitigate political biases. While Song et al. [142] also proposed an aggregation technique for image segmentation tools, suggesting employing different tools for different workers and then aggregating the results to reduce systematic biases.

3.2 Methodologies for Developing and Evaluating Bias Mitigation Solutions

3.2.1 Research Methodologies. This section focuses on the second category of our classification of the researcher's role in developing bias mitigation methods: the research methodologies adopted by researchers. Exploring the vast array of diverse methodologies within one category of debiasing solutions and across the various solutions is essential for understanding how these solutions were formulated, facilitating reproducibility, and identifying best research practices. It also opens up potential areas of refinement and expansion in research within this field.

Our analysis revealed six research methods across all solution categories outlined in the previous section. These methods include **user research, user-based experiments, data-based experiments, case studies, conceptual analysis** and **theoretical analysis**. The diversity of these methods mirrors the variety of solutions we presented in

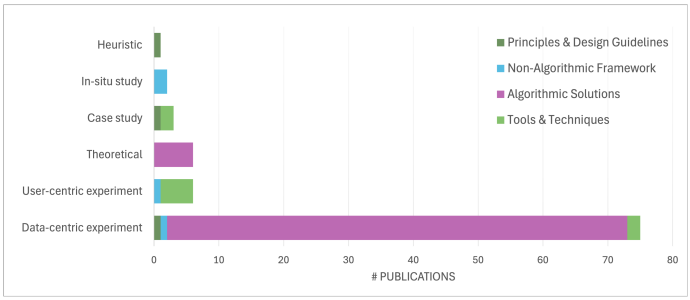

Fig. 4. The various evaluation methods utilized across different categories of bias mitigation solutions.

the previous section and highlights the need for conducting multiple research approaches due to the complexity of addressing bias.

From our analysis, **data-based experimentation** emerges as the most commonly employed methodology across the four types of solutions, with a clear dominance in *algorithmic solutions* (71 out of the 100 papers), suggesting that the development of algorithmic approaches to bias mitigation heavily relies on quantitative data investigations, with the exception of one paper [135] that relied on **theoretical analysis** approach.

While **user-based experiments** (10 papers) and **case studies** (9 papers) may be numerically less prevalent, they were employed across three debiasing solutions categories: *1) principles and design guidelines, 2) non-algorithmic frameworks, and 3) tools and techniques*, excluding *algorithmic solutions*. **user-based experiments** were primarily employed to develop bias mitigation methods under the *tools and techniques* category, reflecting the importance of involving users directly in shaping these methods. In contrast, **case studies** were used the most in creating *principles and design guidelines*, indicating the need for conducting an in-depth study and detailed analysis of real-world scenarios to generate these guidelines. **User research** (5 papers) methods such as surveys, interviews, and focus groups were only employed to develop *principles and design guidelines*, which suggests the significant value of collecting user data for creating this type of solutions.

Finally, **conceptual analysis** (3 papers) was used the least and mainly in developing two solution categories: *1) principles and design guidelines and 2) non-algorithmic frameworks*, which signify the role of the theoretical approach in establishing foundational bias mitigation strategies.

3.2.2 Evaluation Methodologies. Additionally, we analyzed the methods employed by researchers to evaluate the robustness, fairness or both of their proposed solution. We categorized them into the following categories: **data-centric experiments, user-centric experiments, theoretical, case studies, in-situ studies**, and **heuristic**. While this categorization may overlap with the one presented regarding research methodologies, the previous one considered the overall employed methodologies for research, while this one focuses on identifying the employed method for evaluating the debiasing solution only.

Figure 4 emphasizes the use of data-centric experiments in evaluating *Algorithmic Solutions*. In this approach, developers test the proposed algorithms on their **benchmark datasets** to compare their **robustness, fairness**, or both, against other algorithms (baselines) using a quantifiable measure. This process involves specifying a benchmark dataset and a quantifiable metric(s) related to robustness or fairness.

Table 3. Overview of the most common benchmark datasets used for bias mitigation research. The table shows examples from our review sample. We list datasets that were used by at least two publications or more.

Benchmark Datasets	Data Type	Publication References	Count
Adult (Census Income) [12]	Tabular	[4, 14, 21, 22, 47, 67, 68, 70, 90, 118, 127–129, 154, 160, 165, 170]	17
COMPAS [7]	Tabular	[21, 22, 67, 68, 90, 122, 126–128, 158, 165, 167, 170, 172, 173]	15
German Credit [59]	Tabular	[14, 47, 68, 90, 118, 122, 154, 158, 170, 172]	10
CIFAR [82]	Images	[50, 54, 62, 71, 72, 109, 111, 115, 168, 171]	10
MNIST [32]	Images	[9, 13, 46, 50, 60, 71, 86, 109, 115, 167]	10
ImageNet [31]	Images	[9, 60, 62, 71, 78, 85, 111, 171, 174]	9
CelebA [95]	Images	[13, 26, 60, 62, 64, 67, 78]	7
Law School [157]	Tabular	[4, 118, 154, 159, 170]	5
Communities and Crime [124]	Tabular	[118, 127, 159]	3
Bank [108]	Tabular	[22, 156, 165]	3
Default [66]	Tabular	[14, 165]	2
MEPS [17]	Tabular	[68, 122]	2

Table 4. Overview of the most common predictive metrics in evaluating machine learning models' performance with some examples from our surveyed papers.

Predictive Metric	Description	Mathematical Description	Publication References
Accuracy	overall proportion of correct predictions	$\frac{TP+TN}{TP+TN+FP+FN}$	[6, 9, 21, 48, 62, 109, 117, 120, 127, 128, 135, 154–156, 160, 163, 168, 171, 172]
Balanced Accuracy	average of sensitivity (true positive rate) and specificity (true negative rate) for each class	$\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right)$	[68, 129, 151]
Precision	proportion of correctly identified positives among all predicted positives	$\frac{TP}{TP+FP}$	[10, 149]
Recall	proportion of true positives correctly detected	$\frac{TP}{TP+FN}$	[10, 89, 149, 156]
F_1	harmonic mean of precision and recall	$\frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$	[10, 91, 118, 150, 151, 158, 163, 172]
AUC–ROC	probability that a classifier ranks a positive example higher than a negative one	-	[4, 16, 23, 89, 117, 151, 165, 167, 172]

Notation: TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives

Benchmark datasets are publicly available resources widely used by the ML research community to train, test, and compare algorithms under consistent conditions [79]. Table 3 presents the most commonly used benchmark datasets, along with examples from our surveyed papers. As described earlier, the evaluation process involves measuring the models' robustness, fairness, or both. **Robustness** refers to the stability of predictive metrics under various conditions [140], while **fairness** ensures that model performance is distributed equally across subpopulations [104]. Both robustness and fairness are constructs or abstract qualities that developers cannot observe directly, but they must measure them

Table 5. Fairness Metrics in Machine Learning with identified examples from our review sample. To improve practical interpretability, the metrics are defined using confusion-matrix components rather than probability-based notation, following practitioner-oriented formulations in prior work [45].

Fairness Metric	Mathematical Description	Publication References
Statistical (Demographic) Parity	proportion of positive predictions should be equal across different groups $PPR_{g_0} = PPR_{g_1}, \quad \text{where } PPR_g = \frac{TP_g+FP_g}{TP_g+FP_g+TN_g+FN_g}$	[4, 6, 16, 22, 28, 62, 72, 90, 128, 129, 135, 156, 159, 160, 163, 170]
Equal Opportunity	true positive rates should be equal across different groups $TPR_{g_0} = TPR_{g_1}$	[4, 6, 16, 28, 62, 67, 135, 158, 163, 170]
Equalized Odds	both true positive and false positive rates should be equal across different groups $TPR_{g_0} = TPR_{g_1} \quad \text{and} \quad FPR_{g_0} = FPR_{g_1}$	[22, 67, 72, 90, 128, 129, 158]
Average Odds Difference	measures the average difference in false positive rates and true positive rates between groups $\frac{1}{2}\left[(TPR_{g_0} - TPR_{g_1}) + (FPR_{g_0} - FPR_{g_1})\right]$	[68, 90, 158]
Predictive Parity	positive predictive value (precision) should be equal across different groups $PPV_{g_0} = PPV_{g_1}$	[34]
Predictive Equality	false positive rates should be equal across different groups $FPR_{g_0} = FPR_{g_1}$	[4]

Notation: $TPR_g = \frac{TP_g}{TP_g+FN_g}$, $FPR_g = \frac{FP_g}{FP_g+TN_g}$, and $PPV_g = \frac{TP_g}{TP_g+FP_g}$ for sensitive group g. TP, FP, FN, and TN denote true positives, false positives, false negatives, and true negatives.

through quantitative metrics. However, these constructs extend beyond raw metrics such as the number of correct predictions or whether different groups receive similar outcomes; instead, they demand a suite of chosen metrics to capture their full complexity. Typically, developers evaluate the performance of the models through one of the following predictive metrics: *accuracy*, *precision*, *recall*, F_1 score, and the *Area Under the Receiver Operating Characteristic Curve* ($AUC-ROC$). Table 4 summarizes the most commonly used predictive performance metrics, including their definitions, mathematical formulations, and key references from our surveyed papers. When robustness is of interest, these same metrics are applied under perturbed conditions—such as noisy inputs (e.g., [154, 160]), adversarial examples (e.g., [127, 135, 171]), or domain shifts (e.g, [6, 21])—to quantify the stability of performance across scenarios. While such metrics indicate how well a model achieves its intended task, they do not fully capture models' fairness; therefore, developers must use other metrics designed to detect disparities in outcomes between demographic groups. Examples include *Statistical Parity (Demographic Parity), equalized odds, equal opportunity*, and *predictive parity*. Table 5 shows the most common fairness metrics with their descriptions and some examples from our surveyed papers. For further details of fairness metrics, we refer readers to the survey by Caton and Haas [20].

To further clarify the evaluation process, take this example from Bahng et al. [9]. Their research evaluated the proposed debiasing approach on *ImageNet* (benchmark dataset) using ResNet-18 as the backbone model. They trained a

de-biased representation by encouraging statistical independence from a biased representation. Then they measured the model *accuracy* (a performance metric) on the validation set, allowing them to assess whether the mitigation method preserved predictive performance while reducing dataset bias. In another example, Zhang et al. [170] proposed a solution leveraging confident learning to mitigate label bias. Their model is constructed based on a simple neural network using ReLU activation functions. They evaluated it on several benchmark datasets, including *Adult, COMPAS, Default*, and *Law*. The researchers used *accuracy* as a predictive measure for model performance. To measure fairness on the other hand, they used *demographic parity distance (DOP), difference in equal opportunity (DOD)*, and *p%*, a resemblance measure for the *demographic parity distance* metric.

While data-centric experiments are dominant for evaluating *Algorithmic Solutions*, Figure 4 shows that user-centric experiments and case studies are most common for evaluating *tools and techniques*, demonstrating the importance of understanding user interaction and real-world applications to effectively employ these types of solutions to address bias. In contrast, user-centric experiments and case studies are dominant methods for evaluating tools and techniques, demonstrating the importance of understanding user interaction and real-world applications to effectively employ these types of solution to address bias.

4 Human Role in Bias Mitigation Throughout the AI Lifecycle

Although AI systems have unparalleled capabilities for analyzing data and extracting relationships, they are prone to producing discriminative decisions due to existing biases in data and algorithms—biases often introduced by their creators during data collection and system development [75]. Interestingly, humans can outperform AI in specific scenarios by considering nuances and contextual information beyond machine learning capabilities. This human ability signifies the need for an informed human role in actively debiasing machine algorithms and their decisions (e.g., [24, 64, 74]).

Often, the research papers did not explicitly articulate the specific human roles involved in debiasing, but the roles could be inferred them from contextual cues in the papers, such as descriptions of the debiasing mechanism, or assumptions stated by the authors regarding their responsibilities. For example, descriptions of model adjustments often implied technical roles, while discussions of policy, oversight, or deployment practices suggested governance or operational roles. Our analysis defined humans' roles based on the requirements of the proposed debiasing solution. It also identifies their intervention points within the AI lifecycle based on when the solutions are employed. This review extends beyond defining their role to determining the depth of human involvement and interaction with AI systems in minimizing bias. Understanding how, when, and where humans contribute to reducing bias helps to identify gaps and limitations in current practices and frameworks and propose more effective strategies for integrating human expertise into designing and employing ethical frameworks for AI workflows.

We begin by categorizing the different human actors involved and outlining their specific functions (Section 4.1), followed by an assessment of the varying levels of their involvement and decision-making authority (Section 4.2). We then examine debiasing solutions explicitly designed for interactive machine learning, highlighting how human input shapes model behavior and bias reduction (Section 4.3).

4.1 Who is the Human?

Translating an AI system from a design to an application requires the collaborative efforts of individuals with diverse skills and capabilities. Typically, building an AI system goes through an iterative cycle of design, data curation, development, and deployment, requiring the efforts of individuals with computer-related expertise, such as data and AI

scientists and ML engineers [30]. However, our analysis suggests that improving the trustworthiness of AI systems extends beyond traditional roles; it demands further contributions from additional individuals. Moreover, some must participate across several AI lifecycle stages to ensure system robustness and fairness. Reviewing the proposed debiasing methods led us to categorize individuals into distinct roles; these include AI/ML practitioners such as data scientists and ML engineers, data annotators, domain experts, policymakers, and end-users.

Most of the papers in our review (97%) proposed solutions requiring **AI/ML practitioners'** efforts. Their role begin in the design cycle, focusing on problem definition and formation, and identifying design and data requirements. Improving systems' trustworthiness also starts in this stage, requiring data scientists to apply guidelines, principles, and design recommendations supporting that goal. For example, Levonian et al. [87], targeting model designers and developers, proposed non-random sampling guidelines to improve the robustness of text classifiers. In other work in evaluating automated speaker recognition bias, Hutiri et al. [65] recommended that developers carefully select the error metrics when evaluating these models across subgroups. Katell et al. [74] urges ML practitioners not to rely solely on technical interventions but also to include community-based methods in designing equitable algorithmic systems for situated contexts. In this paper, the authors discuss the Algorithmic Equity Toolkit as an example for engaging local community groups, advocacy campaigns, and policy stakeholders in the co-design process.

During the data curation cycle, data scientists define data requirements and design data collection pipelines. Several papers in our review proposed frameworks and design recommendations that demand practitioners' attention when establishing these requirements and pipelines, ensuring curating balanced datasets that satisfy fairness and robustness objectives. For instance, in an effort to rehumanize the crowdsourcing process, Barbosa et al. [11] investigated the efficiency of a crowdsourcing framework that manages workers' sample distribution by considering their demographics and other criteria to mitigate their bias in the data. Another study [132] urged to consider models' contextual information when designing data crowdsource pipelines. Regarding data annotations, Hirota et al. [57] called on practitioners to reduce gender and racial bias by considering ethical aspects when designing data annotation processes for creating visual question-answering (VQA) datasets. Similarly, Pang et al. [116] provided practitioners with a generalized framework emphasizing global inclusivity when recruiting labelers and ensuring consistent labeling when auditing their data annotation processes.

Several other papers presented technical interventions for ML engineers to mitigate bias in the data. Thebault-Spieker et al. [149] investigated a data aggregation technique that reduces political bias by combining samples from heterogeneous labelers. Another example recommended a social projection technique targeting annotators' judgments to reduce biases during subjective annotation tasks [64] .

Ensuring systems' trustworthiness mandates selecting appropriate algorithms, constraints, and objectives during the development cycle; this is where the data scientists' role becomes inevitable as they bridge the gap between theoretical fairness frameworks and practical implementation. Our review includes several papers laying out various algorithmic solutions under AI/ML practitioners' hands for diverse cases, such as: balancing data distribution before training using pre-processing methods (e.g., [4, 26, 68, 117], across-groups equitable learning using in-processing methods (e.g., [6, 14, 47, 62, 155, 167], and other post-processing methods to meet fairness objectives (e.g., [34, 94, 110].

Additionally, we identified studies advocating for employing non-algorithmic frameworks that impose fairness constraints beyond the traditional algorithmic approaches. These include structural frameworks [147], ethical frameworks [103], and a framework guided by feminist principles [28]. While these frameworks are non-technical, AI/ML practitioners participate in reducing bias by integrating these frameworks during system development and deployment.

Other studies focused on the critical roles of **data annotators'** in reducing bias. 9% of the proposed solutions involve annotators' effort to help reduce bias during data annotation. These studies range from including them in the participatory design process [43] to promoting global inclusivity during recruitment [116] and specifying annotators' characteristics, work organization, and labor conditions to cultivate and encourage better environments for workers [153]. Additionally, some research employed social projections and political self-awareness approach to reduce cognitive biases among workers during data annotation [64], while others created new interactive methods to facilitate image annotation with simple clicks [55]. Other studies published in [142, 149] mitigated data bias through label aggregation techniques.

Research introducing non-algorithmic frameworks to address bias necessitates careful consideration and active involvement from **policymakers** (16%). Their role is to review and refine these frameworks and enact legal regulations at the institutional and organizational levels to employ them effectively. These frameworks cover a broad spectrum, including structural interventions [28, 147], medical ethics and justice-based theories [103], ethical consideration for workers in data crowdsourcing [11] and promoting social inclusion, diversity and accessibility during data collection [63]. In another direction, Katell et al. [74] presented a case study for an Algorithmic Equity Toolkit to understand the discrimination dimensions beyond algorithmic bias and hold policymakers accountable for interventions that must consider historical, political, and institutional contexts in which systems are situated. Other researchers called for inclusive policy frameworks supporting data collection centered around people with disabilities [44]. Meanwhile, Wang et al. [153] called for systematic and structural changes in work practices that prioritize annotators' well-being and benefits over those of annotation companies.

Our analysis includes minimal research on debiasing methods that involve roles beyond AI/ML practitioners, annotators, and policymakers such as, domain experts and end users. Only five papers (5%) highlighted **domain experts'** role in reducing bias through participatory design processes and collaborative decision-making in AI systems. Their domain-specific knowledge supports tailoring AI systems to particular real-world applications and situated contexts. They contribute substantially during problem conceptualization, data collection, model evaluation, and system post-deployment. For instance, Suresh et al. [147] guided by data feminism, described a participatory process involving activists co-designing datasets and ML models. Related to the medical field, McCradden et al. [103] presented an ethical decision-making framework, JustEFAB, to address bias in clinical tools, guided by medical ethics and social justice principles and reviewed by multiple stakeholders, including clinicians. Similarly, Freeman et al. [43] proposed principles and guidelines for an iterative participatory data collection process for medical imaging, combining input from medical experts across various medical subdomains. Advocating for non-technical interventions, recent research published in [74, 153] emphasized the importance of active participation from all stakeholders and addressing bias within its specific contexts. Meanwhile, Cheng et al. [24] proposed design implications to enhance the collaboration between social workers and AI, particularly concerning racial disparities in child welfare contexts.

In our sample, only 4 out of the 100 studies closely examined the role of **end users** in reducing AI bias in high-risk decision-making contexts. Two studies focusing on social workers' assessment of child maltreatment [24, 29] emphasized the significance of maintaining human agency over the machine and highlighted the risks of full automation. Another study by Peng et al. [119] recommended decoupling bias sources to reduce gender bias in hiring. This research demonstrated that decision-makers gender and the distribution of genders within a profession significantly impact hiring decisions. In another study, researchers recommended improving Visual Question Answering (VQA) datasets by allowing dataset users to report ethical problems with the data [57].

4.2 Level of Human Involvement in the Debiasing Intervention

As shown in Section 4.1, reducing bias to satisfy robustness and fairness objectives requires various skilled and experienced individuals throughout the entire AI lifecycle, from design stages to deployment. We defined their roles based on the requirements and scope of the debiasing intervention method; for instance, algorithmic solutions demand the expertise of AI/ML practitioners. We also captured several aspects of their roles to determine the depth of their involvement, such as the frequency of their input, the timing of their participation throughout the AI lifecycle, the complexity of their decisions, and the impact of their actions. Defining the depth of involvement provides an understanding of stakeholders' varying degrees of influence and responsibility in reducing bias during the AI development process. Thus, by attempting to define levels of involvement, our review can pinpoint areas lacking participation in some of these roles and suggest a more inclusive, comprehensive approach to debiasing involving individuals beyond the technical aspects. It is important to note that these levels are not entirely disjoint, as roles can overlap and evolve across different stages depending on the type of debiasing solution. To summarize differences in roles, our analysis led us to classify three levels of human involvement depth, which we describe as low, medium, and high, which we explain next.

4.2.1 Low-Level Involvement. This level describes scenarios where human interaction with the system is limited to one-time or infrequent interventions (e.g., [14, 16, 18, 34, 46, 47, 50, 62, 71, 76, 86, 91, 117, 120, 122]) Generally, they occur when performing data preprocessing, model training, and fine-tuning scenarios. The influence of human actions extends across all system outcomes but is relatively minimal as the system mainly operates autonomously. For example, employing data preprocessing techniques aimed at balancing datasets and reducing bias by aggregation [142, 149] and resampling [33]; once applied by ML practitioners, these preprocessing steps do not require ongoing human intervention as the model processes the data independently during training. Another example is using an improved k-means clustering algorithm to achieve fair clustering *(Fair-Lloyd)* [47]. Implemented by ML engineers during the development cycle, this approach can significantly enhance the system's fairness. However, human interaction is limited to the algorithm's initial implementation, after which the system operates autonomously. While crucial, this one-time, task-specific involvement does not require continuous intervention by ML practitioners, reflecting their low involvement level during these intervention scenarios.

4.2.2 Medium-Level Involvement. This level covers cases requiring regular human input, such as employing algorithmic frameworks that promote fairness and robustness [21, 85, 93, 109, 118, 158, 165, 167, 172]. These cases may occur during several development phases and commonly involve AI/ML practitioners, such as data scientists and ML engineers, whose actions moderately influence operational outcomes. At this level, understanding the system's requirements is essential for making decisions that align with the framework standards. For instance, this understanding is crucial to deciding the proper framework, inclusive and representative data collection pipelines, setting reasonable model parameters, and choosing evaluation metrics.

During the data curation phase, scientists might iteratively refine datasets to ensure a balanced representation of different demographic groups; for example, Liu et al. [93], in an effort to reduce the human efforts and disagreement in labeling tasks, developed an algorithmic framework for learning labels distribution using few labels per item. Employing this framework impacts the design of the data curation and development process, requiring the following from practitioners: collecting and aggregating diverse and representative human-annotated labels per item, designing and implementing label distribution methods, iteratively refining datasets to maintain balanced label distributions, and continuously evaluating and tuning model performance to enhance reliability and fairness.

In the development stage, algorithmic frameworks require ML engineers to adjust algorithms to align with framework guidelines and standards continuously. For instance, Wu et al. [158] presented a fairness-aware PUL *(FairPUL)* algorithmic framework for fair classification, a post-processing model-agnostic method based on positive and unlabeled learning *(PUL)* to utilize unlabeled data. The *FairPUL* framework, involves several steps where data scientists must actively engage with the model and the data, including estimating specific parameters that require a detailed understanding of the validation data and the model's behavior, then computing critical terms using both labeled and unlabeled datasets to minimize unfairness in the data, and finally computing the final model parameters for the optimal fair classifier which require a comprehensive understanding of the underlying mathematical framework of the used model.

Our analysis identified only algorithmic frameworks at this level of involvement, likely due to their extensive nature. These frameworks guide algorithms' development, implementation, and evaluation, as demonstrated in the *FairPUL* example. As a result, such interventions require iterative and regular human involvement based on the problem definition, context, and models' mathematical characteristics.

4.2.3 High-Level Involvement. This level includes methods characterized by continuous human interactions and high-impact decisions. We classify two types of solutions under this category: 1) principles and design guidelines (e.g., [43, 44, 63, 119, 130, 132, 144, 153, 164]) and 2) non-algorithmic frameworks (e.g., [11, 103, 116, 146]). Addressing bias by following specific principles and guidelines and adopting ethical and structural frameworks heavily rely on humans. These methods require humans to make subtle design judgments and ethical and fairness considerations to align AI system outcomes with established standards. They also have a multi-faceted impact during several phases of the AI lifecycle, including institutional policies, organizational practices, model outcomes, and the roles of other stakeholders, such as annotators and end users. For example, Peng et al. [119] encouraged practitioners to consider global inclusivity while recruiting annotators to increase data heterogeneity and evaluate the data labels for consistency using their auditing framework. This framework requires data scientists to adjust data collection pipelines and perform ongoing analyses and adjustments based on the guidelines. The impact extends to operational changes for data crowdsourcing companies, annotators' job opportunities and task natures, ML engineers' assignments, and overall system outcomes. In another case, Huang et al. [63] encouraged practitioners to adopt the following principles to promote social inclusion in curated data: 1) embrace practices for cultural humility to reduce their own biases, 2) consider situational contexts of where the models are to be adopted, and 3) engage different communities to match their needs. Adopting these principles requires the continuous engagement of various stakeholders, such as data scientists, ML engineers, policymakers, and end-users, during different phases; for example, cultural humility practices need to be formally set by an authoritative figure and may require developing tools to help practitioners to self-reflect and adjust. At the same time, engaging the community would require allocating resources and setting policies to facilitate their engagement.

Ensuring human agency in human-AI collaborative decision-making also falls under this category. As recommended by several studies [24, 64, 119], this approach requires continuous review of AI decisions and consideration of contextual contexts beyond machine capabilities. Humans can override system decisions when found to be discriminative, increasing fairness across groups. Adopting these recommendations can significantly impact systems' outcomes, but success requires policymakers and regulators to set policies that change workflows to ensure human oversight.

Human interventions through interactive tools and systems are also classified as high-level due to the continuous and active engagement required to manage and mitigate bias. For instance, He et al. [55] developed an interactive system that explains deep neural network *(DNN)* decisions by visualizing the contributing features; the system empowers practitioners to actively engage with the model by highlighting regions of interest to guide the model's attention.

Similarly, other researchers presented an interactive visual analysis system that allows scientists to reduce system errors by identifying missing associations between concepts and target classes and evaluating mitigation methods to reduce bias [3]. Both examples require practitioners to monitor and engage with the model continuously, interpret visualizations and complex data relationships, make complex decisions, and perform iterative model adjustments and evaluations. Utilizing interactive systems to effectively improve systems' trustworthiness involves collaboration among stakeholders, including AI/ML practitioners and domain experts, mainly when applied within specific domains. Additionally, policymakers and regulators must set policies and roles to facilitate changes in AI workflows.

4.3 Human Interaction

Recent research in the HCI community has focused on incorporating interactive methods to improve AI systems through human collaboration. This area, commonly known as Interactive Machine Learning (IML), as described by Dudley and Kristensson [38], is an interaction paradigm that describes a wide range of methods to leverage user feedback through ongoing and active engagement, thereby improving AI performance.

In the context of addressing bias in AI, interactive tools enable practitioners to visualize data, interpret model outputs, make informed adjustments, and directly manipulate data or model components, thereby reducing data and algorithmic bias. Only a few studies in our review investigated adjusting data and model components through interactive methods. For example, Ahn et al. [3] developed an interactive visual analysis system to engage ML practitioners throughout the development cycle. This system enables users to identify bias by inspecting the association between data and target classes and manipulating these associations to evaluate different mitigation strategies. Similarly, He et al. [55] created an interactive system that allows users to control the attention of a deep neural network by annotating regions of interest in an image with simple clicks. Both studies demonstrated that giving users direct control over data and model components during the training and fine-tuning phases empirically reduced errors and improved system performance. In another study, the researchers investigated the effectiveness of using a multi-tool approach over a single tool to reduce systematic errors introduced by the tools during image segmentation [142]. With four interactive methods to segment images, including basic trace, pin-placing, drag and drop, and flood fill, they found that assigning different methods to different workers for the same task and aggregating their answers can significantly reduce dataset bias. Additionally, Levonian et al. [87] proposed a set of design guidelines to improve interactive interfaces for text annotations. They empirically investigated ways to enable users to provide feedback to the model during active learning. Through full-text search, users could seed the classifier with the required samples, increasing the learning rate and reducing human costs.

5 Classifying Differences in Motivations and Methodologies: ML vs. HCI

Due to the breadth of this topic, we decided to include research from human-centered AI and intelligent systems communities, providing us with a comprehensive look at the state-of-the-art research on this topic. While both communities aim to reduce bias, each approaches the challenge from different perspectives due to their distinct focuses. Human-centered AI (HCAI) research prioritizes designing AI systems that maintain human agency, explain its decisions and align with their needs by understanding their interactions and behaviors throughout their experience with AI systems [8, 138]. In contrast, machine learning research focuses on developing algorithms and frameworks to improve the performance and reliability of AI systems [5, 99].

This section focuses on identifying the research motivations (Section 5.1) and methodologies (Section 5.2) across the seven publication venues included in our review.

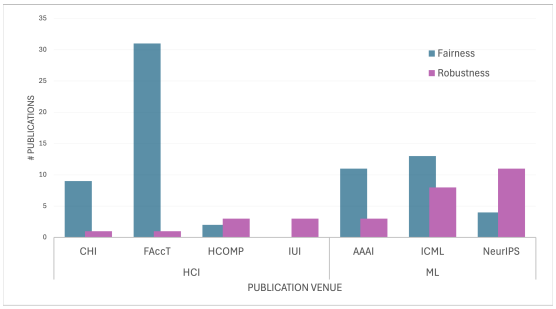

Fig. 5. The distribution of research publications focused on the principles of fairness and robustness across academic venues included in our review.

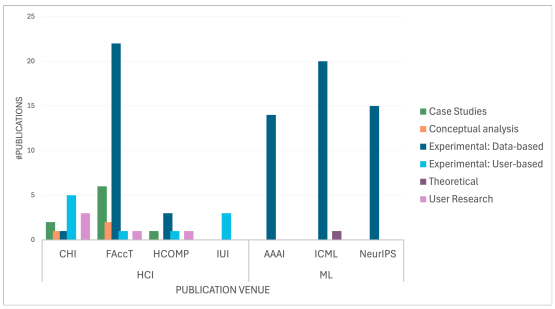

Fig. 6. The distribution of research methodologies employed across academic venues included in our review.

5.1 Targeted Trustworthy Principle

Influenced by the EU requirements (mentioned in Section 1) for creating trustworthy AI [141], our analysis revealed two key research motivations for minimizing bias. The first is increasing robustness, which aims to enhance the model's performance in real-world scenarios when faced with unexpected new data, whether it is due to population and correlation shift (e.g., [6, 128]) or adversarial inputs [127], thereby reducing its generalization error (e.g., [78, 85]). The second motivation is increasing fairness, which focuses on delivering equal opportunities across different individual and group levels (e.g., [68, 129]). Addressing data and algorithmic bias have a direct effect on on the outcomes in both of these cases. Figure 5 shows the distribution of research publications focused on the principles of fairness and robustness across the academic venues included in our review.

We classify papers into one of the above categories based on the following criteria: 1) explicit references to robustness or fairness as their goal to minimize bias or 2) the evaluation metrics used to assess the proposed bias minimization strategies. Typically, strategies aimed at enhancing robustness utilize performance-related metrics such as model accuracy (e.g., [62, 87, 132, 149, 160]). In contrast, strategies focused on fairness often involve one or more fairness metrics, such as group or individual fairness measures (e.g., [68, 76, 90, 110, 129]).

Among the papers focusing on increasing robustness, 73% were published in ML venues. In contrast, 60% of the papers dedicated to enhancing fairness were found in HCI venues. These results indicate that the machine learning community tends to prioritize improving model performance, while the HCI community places greater emphasis on reducing social and ethical biases to enhance fairness across individuals and groups.

5.2 Identified Research Methodologies

The scope of different research fields mandates different research methodologies. Figure 6 shows the distribution of research methodologies employed across academic venues included in our review. Our analysis revealed that all three ML venues conduct empirical research through data-based experimental studies on benchmark datasets (e.g., [3, 9, 10, 16, 37, 91, 110, 160]), highlighting the reliance on empirical data for conducting computational experiments. As defined in [84], these studies typically aim to establish the superiority of a new method over an existing one, considering the new method as the independent variable. The dependent variable is some measure of performance such as: *Accuracy* (e.g., [21, 34, 48, 109, 120, 154]), *F1 score* (e.g., [91, 118, 158]), *ROC AUC* (e.g.,[16, 89, 117, 165, 167]).

In contrast, the HCI venues (CHI, FAccT, HCOMP, IUI) illustrate a broader mix of empirical research methodologies. Data-based and user-based experiments, case studies, and user research are common methods across CHI, FAccT, and HCOMP, reflecting their interdisciplinary approach. Other theoretical methods, such as conceptual analysis, are also found in FAccT and CHI. Meanwhile, IUI only illustrated empirical methodologies through user-based experiments.

Focsing on human-centered methods, our analysis shows that FAccT, CHI and HCOMP stand out as having diverse approaches. In CHI, user experiments are the most prevalent methodology, followed by a mix of retrospective analysis, observations, and interviews. In FAccT, participatory and co-design methods, surveys, and observations are the most common, with little emphasis on user experiments. HCOMP shows a balanced distribution of interviews, pilot studies, and user experiments.

6 Discussion

We reviewed the state-of-the-art literature on addressing bias from various ML and HCI venues. Figure 7 provides a comprehensive overview of our classification scheme of the wide range of solutions in addressing bias, the different roles of people, their depth of involvement during the debiasing process, and the research methodologies and motivation across different ML and HCI venues, referencing examples from the publications in our sample, which we extensively discussed in the preceding sections.

In this section, we highlight some of the main challenges to addressing bias identified from the reviewed papers (Section 6.1), considerations to effectively leverage human roles for addressing bias (Section 6.2), and summarize key insights from the data that extend beyond technical solutions (Section 6.3).

6.1 Overcoming Bias: Challenges in Existing Solutions and Methods

The classification of solutions for addressing bias in AI, as presented in Section 3, provides a comprehensive overview, showing the breadth of solutions in state-of-the-art literature in HCI and ML fields. However, the breadth of these solutions also highlights several significant challenges: the complex character of bias, the lack of real-world evaluations and the limited interdisciplinary research in this area.

Bias is Complex. Bias can originate from various sources and manifest at different stages of the AI lifecycle, including data collection processes [37, 85, 93], algorithmic design [69, 91, 117], and deployment conditions [24, 29, 119]. Each phase introduces unique challenges and requires distinct strategies for bias mitigation. Consequently, we see diverse solutions from the design phase to deployment in our review, each addressing different aspects of bias, such as systems' robustness (e.g., [9, 71, 89]), social fairness (e.g, [33, 64, 65]), ethical standards (e.g., [16, 103, 117]), structural changes and regulations (e.g., [28, 147]) requiring different interventions and human roles.

Moreover, identifying and measuring bias are inherently difficult processes. Bias is not always apparent and can be deeply embedded in a data or model. Even when bias is detected, determining the appropriate debiasing mechanism is complex, particularly for already deployed systems. Implementing changes to mitigate bias in such systems might require significant modifications to existing workflows, which can incur substantial human, financial, and time costs. While it is generally more feasible to address bias starting from the design cycle, where interventions can be integrated into the foundational stages of AI development, this proactive approach might still require identifying potential biases and their sources before they manifest. This task is complex and necessitates a thorough understanding of the data, the domain, and the societal context in which the AI system will operate.

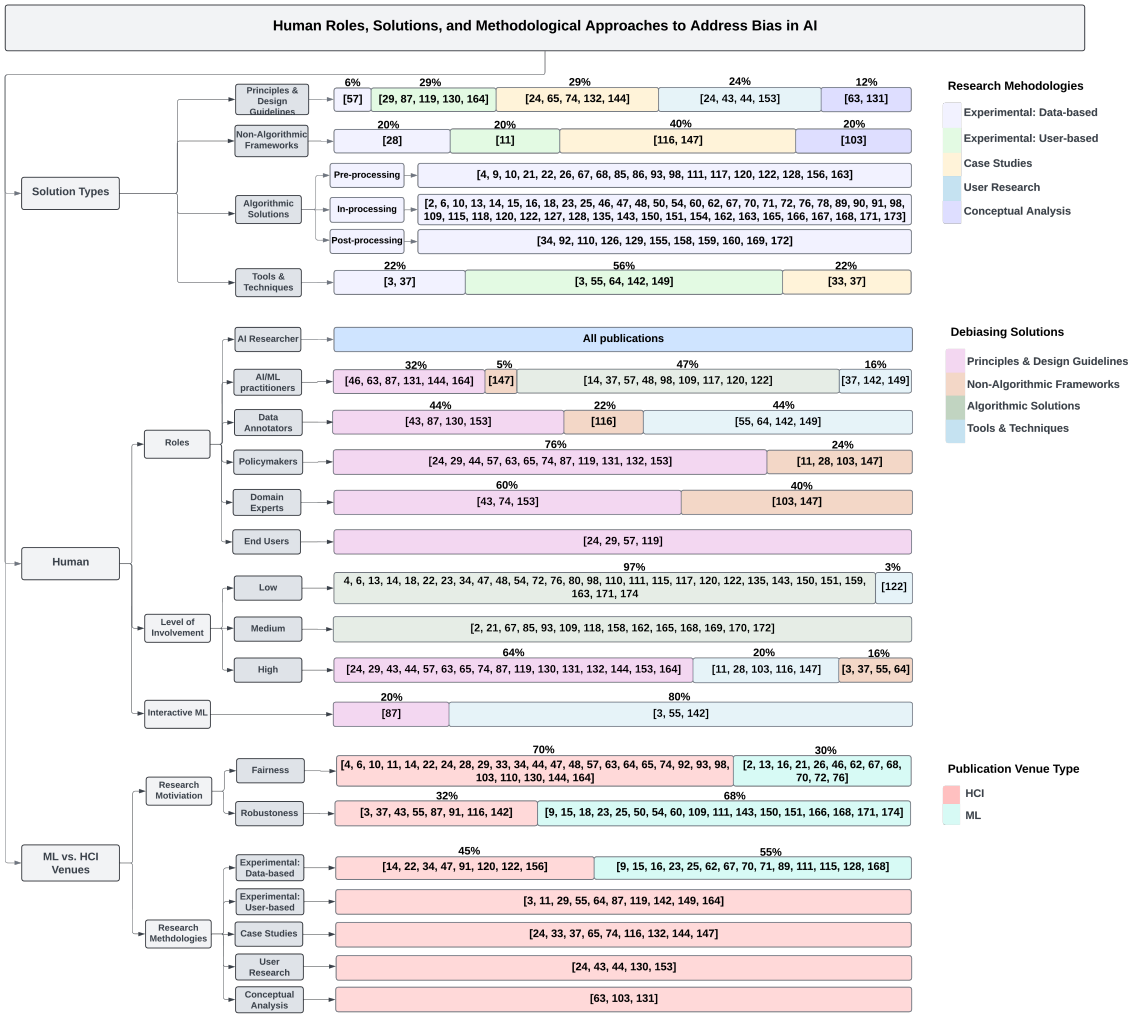

Fig. 7. Our classification scheme of the wide range of solutions in addressing bias, the different roles of people, their depth of involvement during the debiasing process, and the research methodologies and motivation across different ML and HCI venues, referencing examples from the publications in our sample. Note that the sample sizes across categories are different.

Lack of Real-World Evaluations. From our analysis, we found that researchers primarily conducted data-based experiments on real-world and synthetic datasets to assess the performance of algorithmic solutions; on the other hand, user-based experiments were conducted to evaluate tools and techniques. These controlled experiments, while helpful, primarily fail to capture the full scope of real-world complexities. Only 6% of the papers in our sample employed other methods of evaluation, including theoretical, case studies, in-situ, and heuristic.

Out of these methods, in-situ evaluations can provide the most accurate assessments of the proposed solutions to address bias. Only two papers (2% of the sample) reported employing in-situ evaluations. Such evaluations are applied in real-world scenarios involving all relevant stakeholders under realistic conditions, where numerous variables and

unexpected challenges can influence outcomes. These evaluations can help reveal the gaps between solutions' theoretical and practical utility. Additionally, involving multiple stakeholders—including practitioners, domain experts, end-users, and impacted communities—can provide a holistic view of the solution's effectiveness and exposes other critical issues such as ethical implications, user acceptance, scalability, and unintended consequences. However, conducting in-situ evaluations is challenging. The logistical efforts and time required to implement these solutions in real-world settings are substantial, involving coordination among various stakeholders and entities. Ethical and privacy concerns also arise, mainly when dealing with end-user data, requiring precautions to protect individuals' privacy.

Further, translating research findings into practical solutions involves overcoming technical, organizational, and regulatory limitations. Challenges might also emerge due to incompatibility with existing infrastructure, organizational resistance to change, and compliance with legal and ethical standards, which can discourage researchers from evaluating their solutions in real-world settings.

Moreover, the fast pace of research in this field adds another layer of complexity. Researchers are often under pressure to publish their findings quickly, which can discourage thorough, time-consuming evaluations in favor of more immediate, less comprehensive assessments. This competitive environment can lead to a focus on incremental advances rather than comprehensive, long-term solutions.

Limited Interdisciplinary Collaboration. While the variety of solution types in our sample highlights the multidisciplinary nature of bias mitigation research, it also reveals significant gaps in interdisciplinary collaboration. The machine learning community primarily focuses on algorithmic approaches, which align with the technical goals of the field. However, the absence of non-algorithmic solutions may suggest open opportunities to enhance the effectiveness of bias mitigation by incorporating more human-centered perspectives.

Papers from the HCI community offered a wider variety of solutions, as HCI is, by nature, a highly interdisciplinary subfield in computing. Still, the converse in approaches seen in ML conferences was generally found in some HCI venues, with algorithmic approaches being less common or sometimes lacking (e.g., CHI and IUI). Although, the FAccT conference stands out, featuring all four types of solutions. Most solutions presented in FAccT papers (total = 32) were algorithmic (62%), 13% focused on non-algorithmic solutions, 22% presented principles and guidelines, and 3% focused on tools and techniques. This diverse mix of solutions reflects a strong emphasis within the FAccT community on integrating multiple perspectives and fostering an interdisciplinary approach. This trend is likely driven by the conference's explicit focus on fairness, accountability, and transparency, naturally encouraging broader, more inclusive discussions. HCOMP also showed a variety of these four solution types, but this may not be fully representative due to the small number of papers included (5 out of 100). All together, the results of solution types across venues demonstrates the need and value of interdisciplinary venues, and they also motivate potential benefits of adopting interdisciplinary methods within any particular community.

The iterative, multi-phase nature of the AI lifecycle generally further highlights the necessity of interdisciplinary collaboration. Building and evaluating practical and robust solutions to address bias requires a holistic approach that considers the entire AI lifecycle rather than focusing on a single phase; it requires insights and specialized knowledge from various researchers at every stage of the AI lifecycle. Moreover, interdisciplinary collaborations enrich the research process by incorporating diverse methodologies and viewpoints, leading to innovative solutions. For example, combining the technical aspects of ML with the ethical and social insights from HCI can result in balanced and contextually aware approaches to bias mitigation. Addressing bias in AI systems is a complex, multifaceted challenge that requires the

combined efforts of researchers from various fields. Together, researchers from different fields can develop technically efficient, socially responsible, and ethically robust solutions for creating trustworthy AI systems.

6.2 Leveraging Humans: Considerations for Effective Involvement

Human-Centered Approach. Our classification of human roles reveals two distinct standpoints on their involvement in mitigating bias:

(1) *Human contribution to reducing bias through research activities (outside the AI lifecycle)*: We see this rooted in the human-centered AI space [19]. It includes methods that place humans at the center of the design process[3] to enhance their performance in making trustworthy AI systems. It also includes participatory design methods involving various stakeholders through design and evaluation phases [136, 137].
(2) *Human control over bias in real-world applications (within the AI lifecycle)*: This point focuses on human-in-the-loop methods or interactive machine learning [38, 40], where humans maintain control over bias within real-world AI applications. This approach ensures ongoing human oversight and intervention during the deployment and operation of AI systems, emphasizing the crucial role of human judgment in mitigating bias.

The above acknowledges the significant responsibility of engaging various stakeholders outside and within the AI lifecycle. Several researchers, particularly from the HCI community, including the European Union (EU) and HCAI institutes from UC Berkeley and MIT, advocated for human involvement in creating ethical and trustworthy AI [161]. Typically, in research, the primary contributors to addressing the bias problem are the researchers themselves. However, adopting a human-centered approach allows other stakeholders to contribute, bringing in domain knowledge, situational awareness, and contextual details [161] that the researchers may lack. Human-centered methods were evident only in papers proposing non-algorithmic solutions, including principles and design guidelines, non-algorithmic frameworks, and tools and techniques (e.g., [24, 103, 147]). This observation aligns with the HCI community's focus, where human-centred methods are foundational to their research. Applying a user-centric approach to designing and building real-world applications begins with defining business requirements, followed by researching users to understand their needs before starting the design process [73]. Several papers adopting this approach included domain experts, data annotators, and end-users as active participants in various research phases. These individuals contributed through surveys, interviews, case studies, and user experiments, providing valuable insights that helped researchers better understand and address user needs (e.g., [43, 119, 147]).

Additionally, EU guidelines advocated for human-centered approach because it supports maintaining human agency and control over full autonomy resulting in developing and evaluating fair and robust systems [141]. Engaging users and other stakeholders throughout the AI lifecycle ensures that the systems are designed with a comprehensive understanding of the contextual and ethical implications, leading to more effective and trustworthy AI solutions.

As we mentioned earlier in Section 4, we often had to infer the roles of various stakeholders when the research did not explicitly mention them, which clearly indicates users' lack of involvement in research. Moreover, the absence of an explicit characterization of these roles suggests a limited consideration of human control over these solutions outside and within the AI lifecycle; this is especially evident in the papers proposing algorithmic solutions. However, these solutions primarily require higher involvement from AI/ML practitioners; researchers must consider wider dimensions and broader perspectives when applying them in real-world contexts, including involving other stakeholders.

[3]Also referred to as *user-centered* or *user-centric*

Explainable AI as an Enabling Tool for Human Oversight. While this review focuses on methods that explicitly aim to address bias, *eXplainable AI* (XAI) techniques play an important complementary role in bias-related workflows, as illustrated in He et al. [55] and Ahn et al. [3] from our review sample. Despite the relevance and potential benefits of explanation for human understanding of algorithmic biases, the limited presence of XAI in our review may highlight opportunities for further exploration at the intersection of XAI with bias mitigation. Techniques for explanation vary greatly in both scope and function to aid human understanding or intervention [1, 107, 121]. Some of the most common approaches provide local explanations, such as Local Interpretable Model-agnostic Explanations (LIME) [125] and SHapley Additive exPlanations (SHAP) [97], which explain individual predictions by highlighting the most contributing features to those predictions. Others offer global explanations that summarize overall model behavior or feature importance across a dataset (e.g., [56, 81, 96]). Explainability methods may also be model-agnostic, applicable across different architectures, or model-specific, leveraging internal model structures such as attention mechanisms or gradients [1, 49].

However, XAI techniques do not, on their own, directly reduce bias. Instead, they primarily support interpretation [101], bias detection [105], and human-in-the-loop oversight [145] by making model behavior more transparent to different human roles identified in our review, including developers, domain experts, and policymakers. In practice, explanations may often guide follow-up actions aiming to address bias at different stages of the AI lifecycle, such as adjusting model design choices during development cycle (e.g., [3, 55]) or influencing systems' decisions at deployment time (e.g., [58, 101]). Other techniques enable direct human interaction with models, an approach often known as *explanatory interactive learning* [148]. In this approach, explanations support active human involvement, allowing users to reflect on how a model is making decisions and whether its reasoning aligns with their expectations. Instead of modifying model parameters directly, users can provide feedback by adjusting explanations to indicate how they think the model should behave differently. The model can then be updated to better align its explanations, and in turn its behavior, with the user's mental model. Prior research has demonstrated such methods relevant to addressing bias, with examples including the use of interactive explanations for debugging model problems [83, 134].

Additionally, XAI tools can also inform policy and governance decisions in regulated domains [51], supporting organizational and institutional oversight by human actors such as policymakers and domain experts. For example, explanations are increasingly required to justify automated decisions in sensitive applications such as loan applications [52, 123] and hiring systems [39, 41]. While XAI tools can make models more transparent and accountable, they cannot ensure bias is reduced without human intervention. Meaningful mitigation still depends on humans' interpretation and response to these explanations. Thus, XAI's role is best understood as an enabling tool for governance and human oversight, not a standalone bias-reducing solution.

Collective Responsibility. People can play a crucial role in managing bias as the creators and operators of AI systems. Their decisions at each phase of the AI lifecycle can significantly influence system outcomes [75]. Individuals' diverse knowledge, experience, and personal differences highlight the need for informed, organized, and active human involvement outside and within the AI lifecycle to manage sources of bias effectively at different phases [141]. Recognizing the collective responsibility to address bias in AI, we emphasize the need for a participatory approach. This participatory approach, which includes stakeholders outside and within the AI lifecycle, is crucial for maintaining human superiority and agency over the machine and requires a clear understanding of various stakeholders' roles, their level of involvement, and the impact of their decisions highlighting the dimensions of their single and collective responsibilities in managing bias in AI systems.

6.3 Insights from the Data: Exploring Broader Dimensions

Beyond Technical Solutions. Our research on solutions to address bias in AI has resulted in a broad range of strategies covering the entire AI lifecycle. These strategies, ranging from principles and design guidelines to non-algorithmic frameworks, tools, and techniques, are not limited to the modeling phase but cover the entire AI lifecycle. While our classification may not contain all solutions proposed in the literature, it provides an overview of the solutions covering the entire lifecycle, distinguishing our contribution from other survey papers that focus primarily on the modeling phase which typically categorize debiasing methods into pre-processing, in-processing, and post-processing, all within the development phase (e.g., [20, 104, 113, 133]). We still included some examples of this classification in our survey for a comprehensive overview. By including these categories, we aim to bridge the traditional modeling-focused approaches with a broader perspective, incorporating multiple stages of AI design and deployment from various research disciplines.

The Need for Ethical and Policy Frameworks. Technical solutions to bias can only go so far without clear ethical and regulatory guidelines. Our review identified multiple papers emphasizing the importance of ethical standards, structural changes, and policy interventions. These frameworks provide necessary governance to ensure that AI systems operate fairly and responsibly, mainly when deployed in high-stakes environments such as healthcare, child welfare, and law enforcement systems. [24, 43, 103, 147] Ensuring that AI systems adhere to ethical guidelines is not just a recommendation but a crucial dimension in mitigating bias and fostering trustworthiness in AI.

Addressing Human Roles Factors. Bias in AI systems is often a result of human decision-making, a factor that plays a critical and integral role in designing, developing, and deploying these systems [75]. Data selection, annotation practices, and model evaluation can introduce human biases, making it essential to consider human oversight and intervention throughout different points of the AI lifecycle as part of the solution. Our classification integrates human role factors, often neglected in purely technical frameworks. By defining the role of humans and the depth of their involvement as presented in 4, we aim to highlight humans' crucial and valued involvement throughout the AI lifecycle, from design to implementation and oversight (e.g., [43, 74, 87, 103]).

Bridging the Gap Between Research and Practice. Although many proposed solutions show promise in controlled research environments, there is often a gap between theoretical approaches and real-world applications. This gap stems from various challenges, including scalability, system integration, and organizational resistance to change. Addressing these gaps requires a more interdisciplinary approach that involves collaboration between technical researchers, domain experts, and policymakers. Our survey highlights the need for broader dialogue between disciplines to ensure that solutions can be successfully translated into practical applications.

Acknowledgments

The authors would like to thank Dr. Vincent Bindschaedler and Dr. Daisy Zhe Wang, as well as the anonymous reviewers, for their helpful comments on earlier versions of this manuscript. This work was supported in part by NSF award 1900767 and by DARPA under HR00112390063.

References

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[2] Nimesh Agrawal, Anuj Kumar Sirohi, Sandeep Kumar, and Jayadeva. 2024. No Prejudice! Fair Federated Graph Neural Networks for Personalized Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 10 (2024), 10775–10783. https://doi.org/10.1609/aaai.v38i10.28950

[3] Yongsu Ahn, Yu-Ru Lin, Panpan Xu, and Zeng Dai. 2023. ESCAPE: Countering Systematic Errors from Machine's Blind Spots via Interactive Visual Analysis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3544548.3581373

[4] Abdulaziz A. Almuzaini, Chidansh A. Bhatt, David M. Pennock, and Vivek K. Singh. 2022. ABCinML: Anticipatory Bias Correction in Machine Learning Applications. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1552–1560. https://doi.org/10.1145/3531146.3533211

[5] Ethem Alpaydin. 2021. *Machine learning*. MIT press.

[6] Jose M. Alvarez, Kristen M. Scott, Bettina Berendt, and Salvatore Ruggieri. 2023. Domain Adaptive Decision Trees: Implications for Accuracy and Fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 423–433. https://doi.org/10.1145/3593013.3594008

[7] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine Bias. In *Ethics of Data and Analytics*. Auerbach Publications.

[8] Jan Auernhammer. 2020. Human-centered AI: The role of Human-centered Design Research in the development of AI. (2020).

[9] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. 2020. Learning de-biased representations with biased representations, Vol. 119. JMLR.org, 528–539.

[10] Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2021. Differential Tweetment: Mitigating Racial Dialect Bias in Harmful Tweet Detection. Association for Computing Machinery, 116–128. https://doi.org/10.1145/3442188.3445875

[11] Natã M. Barbosa and Monchu Chen. 2019. Rehumanized Crowdsourcing: A Labeling Framework Addressing Bias and Ethics in Machine Learning. Association for Computing Machinery, 1–12. https://doi.org/10.1145/3290605.3300773

[12] Ronny Kohavi Barry Becker. 1996. Adult. https://doi.org/10.24432/C5XW20

[13] Abhipsa Basu, Saswat Subhajyoti Mallick, and R. Venkatesh Babu. 2024. Mitigating Biases in Blackbox Feature Extractors for Image Classification Tasks, Vol. 37. Curran Associates, Inc., 106411–106439.

[14] Francois Buet-Golfouse and Islam Utyagulov. 2022. Towards Fair Unsupervised Learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1399–1409. https://doi.org/10.1145/3531146.3533197

[15] Alexander Bukharin, Tianyi Liu, Shengjie Wang, Simiao Zuo, Weihao Gao, Wen Yan, and Tuo Zhao. 2023. Machine learning force fields with data cost aware training. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23, Vol. 202)*. JMLR.org, Honolulu, Hawaii, USA, 3219–3232.

[16] Maarten Buyl and Tijl De Bie. 2020. DeBayes: a Bayesian method for debiasing network embeddings, Vol. 119. JMLR.org, 1220–1229.

[17] Julie Bynum. 2021. Medical Expenditure Panel Survey (MEPS) for Dementia Researchers, 2015-2019. https://doi.org/10.3886/E154381V1

[18] Remi Cadene, Corentin Dancette, Hedi Ben younes, Matthieu Cord, and Devi Parikh. 2019. RUBi: Reducing Unimodal Biases for Visual Question Answering, Vol. 32. Curran Associates, Inc.

[19] Tara Capel and Margot Brereton. 2023. What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–23. https://doi.org/10.1145/3544548.3580959

[20] Simon Caton and Christian Haas. 2023. Fairness in Machine Learning: A Survey. *Comput. Surveys* (Aug. 2023). https://doi.org/10.1145/3616865

[21] Elisa L. Celis, Vijay Keswani, and Nisheeth K. Vishnoi. 2020. Data preprocessing to mitigate bias: a maximum entropy based approach, Vol. 119. JMLR.org, 1349–1359.

[22] Eunice Chan, Zhining Liu, Ruizhong Qiu, Yuheng Zhang, Ross Maciejewski, and Hanghang Tong. 2024. Group Fairness via Group Consensus. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1788–1808. https://dl.acm.org/doi/10.1145/3630106.3659006

[23] Trenton Chang and Jenna Wiens. 2024. From biased selective labels to pseudo-labels. In *Proceedings of the 41st International Conference on Machine Learning*. 6286–6324. https://dl.acm.org/doi/10.5555/3692070.3692313

[24] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3491102.3501831

[25] Jinwoo Choi, Chen Gao, Joseph C. E. Messou, and Jia-Bin Huang. 2019. Why Can' t I Dance in the Mall? Learning to Mitigate Scene Bias in Action Recognition, Vol. 32. Curran Associates, Inc.

[26] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. 2020. Fair generative modeling via weak supervision, Vol. 119. JMLR.org, 1887–1898.

[27] Covidence.org. n.d.. Covidence Systematic Review Tool. *Covidence.org*. Retrieved August 12, 2025, from https://www.covidence.org/.

[28] Efrén Cruz Cortés, Sarah Rajtmajer, and Debashis Ghosh. 2022. Locality of Technical Objects and the Role of Structural Interventions for Systemic Change. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2327–2341. https://doi.org/10.1145/3531146.3534646

[29] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376638

[30] Daswin De Silva, Rashmika Nawaratne, Jacek Ruminski, Aleksander Malinowski, and Milos Manic. 2022. Human System Interaction in Review: Advancing the Artificial Intelligence Transformation. In *2022 15th International Conference on Human System Interaction (HSI)*. 1–5. https://doi.org/10.1109/HSI55341.2022.9869473

[31] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[32] Li Deng. 2012. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Processing Magazine* 29, 6 (Nov. 2012), 141–142. https://doi.org/10.1109/MSP.2012.2211477

[33] Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing Age-Related Bias in Sentiment Analysis. Association for Computing Machinery, 1–14. https://doi.org/10.1145/3173574.3173986

[34] Cyrus DiCiccio, Brian Hsu, Yinyin Yu, Preetam Nandy, and Kinjal Basu. 2023. Detection and Mitigation of Algorithmic Bias via Predictive Parity. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1801–1816. https://doi.org/10.1145/3593013.3594117

[35] Dictionary.com. n.d.. Mitigate. *Dictionary.com*. Retrieved September 20, 2024, from https://www.dictionary.com/browse/mitigate.

[36] Dictionary.com. n.d.. Reduce. *Dictionary.com*. Retrieved September 20, 2024, from https://www.dictionary.com/browse/reduce.

[37] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A Checklist to Combat Cognitive Biases in Crowdsourcing, Vol. 9. 48–59. https://doi.org/10.1609/hcomp.v9i1.18939

[38] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (June 2018), 1–37. https://doi.org/10.1145/3185517

[39] Stephen Fabeyo. 2025. Explainable AI in employment decision-making: a systematic review of transparency methods in hiring algorithms. *Issues in Information Systems* 26, 3 (2025).

[40] Jerry Alan Fails and Dan R. Olsen. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, Miami Florida USA, 39–45. https://doi.org/10.1145/604045.604056

[41] Nadine Y. Fares, Samuel A. Moore, and Manar Jammal. 2025. Mitigating Bias in AI Recruitment: Leveraging LIME for Fair and Transparent Hiring Models. In *Intelligent Systems and Applications*, Kohei Arai (Ed.). Vol. 1553. Springer Nature Switzerland, Cham, 448–464. https://doi.org/10.1007/978-3-031-99958-1_29

[42] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods* 5, 1 (March 2006), 80–92. https://doi.org/10.1177/160940690600500107

[43] Beverly Freeman, Naama Hammel, Sonia Phene, Abigail Huang, R. Ackermann, Olga Kanzheleva, Miles Hutson, Caitlin Taggart, Quang Duong, and Rory Sayres. 2021. Iterative Quality Control Strategies for Expert Medical Image Labeling. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9 (2021), 60–71. https://doi.org/10.1609/hcomp.v9i1.18940

[44] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 205–216. https://doi.org/10.1145/3593013.3593989

[45] Pratyush Garg, John Villasenor, and Virginia Foggo. 2020. Fairness Metrics: A Comparative Analysis. In *2020 IEEE International Conference on Big Data (Big Data)*. 3662–3666. https://doi.org/10.1109/BigData50022.2020.9378025

[46] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. 2020. Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 3197–3208. https://proceedings.neurips.cc/paper_files/paper/2020/hash/20d749bc05f47d2bd3026ce457dcfd8e-Abstract.html

[47] Mehrdad Ghadiri, Samira Samadi, and Santosh Vempala. 2021. Socially Fair k-Means Clustering. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 438–448. https://doi.org/10.1145/3442188.3445906

[48] Luke Guerdan, Amanda Coston, Kenneth Holstein, and Zhiwei Steven Wu. 2023. Counterfactual Prediction Under Outcome Measurement Error. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1584–1598. https://doi.org/10.1145/3593013.3594101

[49] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5 (Sept. 2019), 1–42. https://doi.org/10.1145/3236009

[50] Yongxin Guo, Xiaoying Tang, and Tao Lin. 2023. FedBR: improving federated learning on heterogeneous data via local learning bias reduction. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23, Vol. 202)*. JMLR.org, Honolulu, Hawaii, USA, 12034–12054.

[51] Nikhil Gupta. 2025. Explainable AI for Regulatory Compliance in Financial and Healthcare Sectors: A comprehensive review. *International Journal of Advances in Engineering and Management* 7 (March 2025), 489. https://doi.org/10.35629/5252-0703489494

[52] Elowen Hartley and Li Kevin. 2025. Explainable AI for Loan Approval Decisions in FinTech Platforms. *Journal of Computer Science and Software Applications* 5, 6 (2025). https://www.mfacademia.org/index.php/jcssa/article/view/231

[53] Martie G. Haselton, Daniel Nettle, and Damian R. Murray. [n. d.]. The Evolution of Cognitive Bias. In *The Handbook of Evolutionary Psychology*. John Wiley & Sons, Ltd, 1–20. https://doi.org/10.1002/9781119125563.evpsych241

[54] Xilin He, Jingyu Hu, Qinliang Lin, Cheng Luo, Weicheng Xie, Siyang Song, Muhammad Haris Khan, and Linlin Shen. 2024. Towards Combating Frequency Simplicity-biased Learning for Domain Generalization, Vol. 37. Curran Associates, Inc., 31078–31102.

[55] Yi He, Xi Yang, Chia-Ming Chang, Haoran Xie, and Takeo Igarashi. 2023. Efficient Human-in-the-loop System for Guiding DNNs Attention. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. Association for Computing Machinery, New York, NY, USA, 294–306. https://doi.org/10.1145/3581641.3584074

[56] Andreas Henelius, Kai Puolamäki, Henrik Boström, Lars Asker, and Panagiotis Papapetrou. 2014. A peek into the black box: exploring classifiers by randomization. *Data Mining and Knowledge Discovery* 28, 5 (Sept. 2014), 1503–1529. https://doi.org/10.1007/s10618-014-0368-8

[57] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022. Gender and Racial Bias in Visual Question Answering Datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1280–1292. https://doi.org/10.1145/3531146.3533184

[58] Lennart Hofeditz, Sünje Clausen, Alexander Rieß, Milad Mirbabaie, and Stefan Stieglitz. 2022. Applying XAI to an AI-based system for candidate management to mitigate bias and discrimination in hiring. *Electronic Markets* 32, 4 (Dec. 2022). https://doi.org/10.1007/s12525-022-00600-9

[59] Hans Hofmann. 1994. Statlog (German Credit Data). https://doi.org/10.24432/C5NC77

[60] Youngkyu Hong and Eunho Yang. 2021. Unbiased Classification through Bias-Contrastive and Bias-Balanced Learning. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., 26449–26461.

[61] Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. 2023. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *ACM Journal on Responsible Computing* (Nov. 2023). https://doi.org/10.1145/3631326

[62] Ramtin Hosseini, Li Zhang, Bhanu Garg, and Pengtao Xie. 2023. Fair and accurate decision making through group-aware learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23, Vol. 202)*. JMLR.org, Honolulu, Hawaii, USA, 13254–13269.

[63] Han-Yin Huang and Cynthia C. S. Liem. 2022. Social Inclusion in Curated Contexts: Insights from Museum Practices. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 300–309. https://doi.org/10.1145/3531146.3533095

[64] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments. Association for Computing Machinery, 1–12. https://doi.org/10.1145/3290605.3300637

[65] Wiebke Toussaint Hutiri and Aaron Yi Ding. 2022. Bias in Automated Speaker Recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3531146.3533089

[66] I-Cheng Yeh. 2009. Default of Credit Card Clients. https://doi.org/10.24432/C55S3H

[67] Taeuk Jang, Xiaoqian Wang, and Heng Huang. 2024. Adversarial Fairness Network. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 20 (2024), 22159–22166. https://doi.org/10.1609/aaai.v38i20.30220

[68] Taeuk Jang, Feng Zheng, and Xiaoqian Wang. 2021. Constructing a Fair Classifier with Generated Fair Data. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 9 (May 2021), 7908–7916. https://doi.org/10.1609/aaai.v35i9.16965

[69] Erik Jones and Jacob Steinhardt. 2022. Capturing Failures of Large Language Models via Human Cognitive Biases. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., 11785–11799.

[70] Matthew Jones, Huy Lê Nguyê~n, and Thy Nguyen. 2020. Fair k-centers via maximum matching, Vol. 119. JMLR.org, 4940–4949.

[71] Yeonsung Jung, Hajin Shim, June Yong Yang, and Eunho Yang. 2023. Fighting fire with fire: contrastive debiasing without bias-free data via generative bias-transformation. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23, Vol. 202)*. JMLR.org, Honolulu, Hawaii, USA, 15435–15450.

[72] Yeonsung Jung, Jaeyun Song, June Yong Yang, Jin-Hwa Kim, Sung-Yub Kim, and Eunho Yang. 2024. A Simple Remedy for Dataset Bias via Self-Influence: A Mislabeled Sample Perspective, Vol. 37. Curran Associates, Inc., 43632–43662.

[73] Eija Kaasinen, Tiina Kymäläinen, Marketta Niemelä, Thomas Olsson, Minni Kanerva, and Veikko Ikonen. 2013. A User-Centric View of Intelligent Environments: User Expectations, User Experience and User Role in Building Intelligent Environments. *Computers* 2, 1 (March 2013), 1–33. https://doi.org/10.3390/computers2010001

[74] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and P. M. Krafft. 2020. Toward situated interventions for algorithmic equity: lessons from the field. Association for Computing Machinery, 45–55. https://doi.org/10.1145/3351095.3372874

[75] Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. 2022. Trustworthy Artificial Intelligence: A Review. *Comput. Surveys* 55, 2 (Jan. 2022), 39:1–39:38. https://doi.org/10.1145/3491209

[76] Ahmad Khajehnejad, Moein Khajehnejad, Mahmoudreza Babaei, Krishna P. Gummadi, Adrian Weller, and Baharan Mirzasoleiman. 2022. CrossWalk: Fairness-Enhanced Node Representation Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 11 (June 2022), 11963–11970. https://doi.org/10.1609/aaai.v36i11.21454

[77] Cherry Khosla and Baljit Singh Saini. 2020. Enhancing Performance of Deep Learning Models with different Data Augmentation Techniques: A Survey. In *2020 International Conference on Intelligent Engineering and Management (ICIEM)*. 79–85. https://doi.org/10.1109/ICIEM48762.2020.9160048

[78] Nayeong Kim, Sehyun Hwang, Sungsoo Ahn, Jaesik Park, and Suha Kwak. 2022. Learning Debiased Classifier with Biased Committee. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., 18403–18415.

[79] Bernard Koch, Emily Denton, Alex Hanna, and Jacob G. Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. https://doi.org/10.48550/arXiv.2112.01716

[80] Ioannis Kosmidis. 2014. Bias in parametric estimation: reduction and useful side-effects. *WIREs Computational Statistics* 6, 3 (May 2014), 185–196. https://doi.org/10.1002/wics.1296

[81] Sanjay Krishnan and Eugene Wu. 2017. PALM: Machine Learning Explanations For Iterative Debugging. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*. ACM, Chicago IL USA, 1–6. https://doi.org/10.1145/3077257.3077271

[82] Alex Krizhevsky. 2009. Learning Multiple Layers of Features from Tiny Images. (2009).

[83] Todd Kulesza, Simone Stumpf, Margaret Burnett, Weng-Keen Wong, Yann Riche, Travis Moore, Ian Oberst, Amber Shinsel, and Kevin McIntosh. 2010. Explanatory debugging: Supporting end-user debugging of machine-learned programs. In *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*. IEEE, 41–48.

[84] Pat Langley. 2000. Crafting Papers on Machine Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML '00)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1207–1216.

[85] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases, Vol. 119. JMLR.org, 1078–1088.

[86] Jungsoo Lee, Jeonghoon Park, Daeyoung Kim, Juyoung Lee, Edward Choi, and Jaegul Choo. 2023. Revisiting the Importance of Amplifying Bias for Debiasing. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 12 (June 2023), 14974–14981. https://doi.org/10.1609/aaai.v37i12.26748

[87] Zachary Levonian, Chia-Jung Lee, Vanessa Murdock, and F. Maxwell Harper. 2022. Trade-offs in Sampling and Search for Early-stage Interactive Text Classification. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 566–583. https://doi.org/10.1145/3490099.3511134

[88] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023. Trustworthy AI: From Principles to Practices. *Comput. Surveys* 55, 9 (Jan. 2023), 177:1–177:46. https://doi.org/10.1145/3555803

[89] Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, Peng Wu, and Peng Cui. 2023. Propensity matters: measuring and enhancing balancing for recommendation. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23, Vol. 202)*. JMLR.org, Honolulu, Hawaii, USA.

[90] Xuran Li, Peng Wu, and Jing Su. 2023. Accurate Fairness: Improving Individual Fairness without Trading Accuracy. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 12 (June 2023), 14312–14320. https://doi.org/10.1609/aaai.v37i12.26674

[91] Christopher Lin, Mausam Mausam, and Daniel Weld. 2018. Active learning with unbalanced classes and example-generation queries. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 6. 98–107. https://ojs.aaai.org/index.php/HCOMP/article/view/13334

[92] David Liu, Virginie Do, Nicolas Usunier, and Maximilian Nickel. 2023. Group fairness without demographics using social networks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1432–1449. https://doi.org/10.1145/3593013.3594091

[93] Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher M. Homan. 2019. Learning to Predict Population-Level Label Distributions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (Oct. 2019), 68–76. https://doi.org/10.1609/hcomp.v7i1.5286

[94] Yuhan Liu, Ananda Theertha Suresh, Wennan Zhu, Peter Kairouz, and Marco Gruteser. 2023. Algorithms for bounding contribution for histogram estimation under user-level privacyAlgorithms for bounding contribution for histogram estimation under user-level privacy. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23, Vol. 202)*. JMLR.org, Honolulu, Hawaii, USA, 21969–21996.

[95] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 3730–3738. https://doi.org/10.1109/ICCV.2015.425

[96] Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, Beijing China, 150–158. https://doi.org/10.1145/2339530.2339556

[97] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

[98] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2019. Fairness through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data. Association for Computing Machinery, 349–358. https://doi.org/10.1145/3287560.3287564

[99] Batta Mahesh. 2020. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]* 9, 1 (2020), 381–386.

[100] Rafid Mahmood, James Lucas, Jose M. Alvarez, Sanja Fidler, and Marc Law. 2022. Optimizing Data Collection for Machine Learning. *Advances in Neural Information Processing Systems* 35 (Dec. 2022), 29915–29928.

[101] Avleen Malhi, Samanta Knapic, and Kary Främling. 2020. Explainable Agents for Less Bias in Human-Agent Decision Making. *Explainable, Transparent Autonomous Agents and Multi-Agent Systems* 12175 (June 2020), 129–146. https://doi.org/10.1007/978-3-030-51924-7_8

[102] Konstantinos Mavrogiorgos, Athanasios Kiourtis, Argyro Mavrogiorgou, Andreas Menychtas, and Dimosthenis Kyriazis. [n. d.]. Bias in Machine Learning: A Literature Review. 14, 19 ([n. d.]), 8860. https://doi.org/10.3390/app14198860

[103] Melissa Mccradden, Oluwadara Odusi, Shalmali Joshi, Ismail Akrout, Kagiso Ndlovu, Ben Glocker, Gabriel Maicas, Xiaoxuan Liu, Mjaye Mazwi, Tee Garnett, Lauren Oakden-Rayner, Myrtede Alfred, Irvine Sihlahla, Oswa Shafei, and Anna Goldenberg. 2023. What's fair is… fair? Presenting JustEFAB, an ethical framework for operationalizing medical ethics and social justice in the integration of clinical machine learning: JustEFAB. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1505–1519. https://doi.org/10.1145/3593013.3594096

[104] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (July 2021), 115:1–115:35. https://doi.org/10.1145/3457607

[105] Agnieszka Mikołajczyk, Michał Grochowski, and Arkadiusz Kwasigroch. 2021. Towards Explainable Classifiers Using the Counterfactual Approach - Global Explanations for Discovering Bias in Data. *Journal of Artificial Intelligence and Soft Computing Research* 11, 1 (Jan. 2021), 51–67. https://doi.org/10.2478/jaiscr-2021-0004

[106] Tom M. Mitchell. 1980. The need for biases in learning generalizations (Technical Report No. CBM-TR-117).

[107] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11, 3-4 (2021), 1–45.

[108] Sérgio Moro, Paulo Cortez, and Paulo Rita. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems* 62 (June 2014), 22–31. https://doi.org/10.1016/j.dss.2014.03.001

[109] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. 2020. Learning from Failure: De-biasing Classifier from Biased Classifier, Vol. 33. Curran Associates, Inc., 20673–20684. https://proceedings.neurips.cc/paper/2020/file/eddc3427c5d77843c2253f1e799fe933-Paper.pdf

[110] Preetam Nandy, Cyrus DiCiccio, Divya Venugopalan, Heloise Logan, Kinjal Basu, and Noureddine El Karoui. 2022. Achieving Fairness via Post-Processing in Web-Scale Recommender Systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 715–725. https://doi.org/10.1145/3531146.3533136

[111] Tuan Hai Dang Nguyen, Paymon Haddad, Eric Gan, and Baharan Mirzasoleiman. 2024. Changing the Training Data Distribution to Reduce Simplicity Bias Improves In-distribution Generalization, Vol. 37. Curran Associates, Inc., 68854–68896.

[112] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. [n. d.]. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. https://doi.org/10.2139/ssrn.2886526

[113] Kalia Orphanou, Jahna Otterbacher, Styliani Kleanthous, Khuyagbaatar Batsuren, Fausto Giunchiglia, Veronika Bogina, Avital Shulner Tal, Alan Hartman, and Tsvi Kuflik. 2022. Mitigating Bias in Algorithmic Systems—A Fish-eye View. *Comput. Surveys* 55, 5 (Dec. 2022), 87:1–87:37. https://doi.org/10.1145/3527152

[114] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372 (March 2021), n71. https://doi.org/10.1136/bmj.n71

[115] Zibin Pan, Chi Li, Fangchen Yu, Shuyi Wang, Haijin Wang, Xiaoying Tang, and Junhua Zhao. 2024. FedLF: Layer-Wise Fair Federated Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 13 (2024), 14527–14535. https://doi.org/10.1609/aaai.v38i13.29368

[116] Rock Yuren Pang, Jack Cenatempo, Franklyn Graham, Bridgette Kuehn, Maddy Whisenant, Portia Botchway, Katie Stone Perez, and Allison Koenecke. 2023. Auditing Cross-Cultural Consistency of Human-Annotated Labels for Recommendation Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1531–1552. https://doi.org/10.1145/3593013.3594098

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

[118] Jaakko Peltonen, Wen Xu, Timo Nummenmaa, and Jyrki Nummenmaa. 2023. Fair neighbor embedding. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23, Vol. 202)*. JMLR.org, Honolulu, Hawaii, USA, 27564–27584.

[119] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. 2019. What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (2019), 125–134. https://doi.org/10.1609/hcomp.v7i1.5281

[120] Raphael Poulain, Mirza Farhan Bin Tarek, and Rahmatollah Beheshti. 2023. Improving Fairness in AI Models on Electronic Health Records: The Case for Federated Learning Methods. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1599–1608. https://doi.org/10.1145/3593013.3594102

[121] Muhammad Raees, Inge Meijerink, Ioanna Lykourentzou, Vassilis-Javed Khan, and Konstantinos Papangelis. 2024. From explainable to interactive AI: A literature review on current trends in human-AI interaction. *International Journal of Human-Computer Studies* 189 (2024), 103301.

[122] Miriam Rateike, Ayan Majumdar, Olga Mineeva, Krishna P. Gummadi, and Isabel Valera. 2022. Don't Throw it Away! The Utility of Unlabeled Data in Fair Decision Making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1421–1433. https://doi.org/10.1145/3531146.3533199

[123] Vishnu Ravi, Vineet Srivastava, Maninder Singh, Ravi Burila, Nikhil Kassetty, Padma Vardhineedi, Venkata Pasam, Nuzhat Prova, Islam Prova, and Indrajit De. 2025. *Explainable AI (XAI) for Credit Scoring and Loan Approvals*. https://doi.org/10.13140/RG.2.2.35960.15368

[124] Michael Redmond and Alok Baveja. 2002. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research* 141, 3 (Sept. 2002), 660–678. https://doi.org/10.1016/S0377-2217(01)00264-8

[125] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[126] Brianna Richardson, Prasanna Sattigeri, Dennis Wei, Karthikeyan Natesan Ramamurthy, Kush Varshney, Amit Dhurandhar, and Juan E. Gilbert. 2023. Add-Remove-or-Relabel: Practitioner-Friendly Bias Mitigation via Influential Fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 736–752. https://doi.org/10.1145/3593013.3594039

[127] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. 2020. FR-train: a mutual information-based approach to fair and robust training, Vol. 119. JMLR.org, 8147–8157.

[128] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. 2023. Improving fair training under correlation shifts. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23, Vol. 202)*. JMLR.org, Honolulu, Hawaii, USA, 29179–29209.

[129] Prasanna Sattigeri, Soumya Ghosh, Inkit Padhi, Pierre Dognin, and Kush R Varshney. 2022. Fair Infinitesimal Jackknife: Mitigating the Influence of Biased Training Data Points Without Refitting. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., 35894–35906. https://proceedings.neurips.cc/paper_files/paper/2022/hash/e94481b99473c83b2e79d91c64eb37d1-Abstract-Conference.html

[130] Morgan Klaus Scheuerman and Jed R. Brubaker. 2024. Products of Positionality: How Tech Workers Shape Identity Concepts in Computer Vision. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18. https://dl.acm.org/doi/10.1145/3613904.3641890

[131] Ari Schlesinger, Kenton P. O'Hara, and Alex S. Taylor. 2018. Let's Talk About Race: Identity, Chatbots, and AI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173889

[132] Nandana Sengupta, Ashwini Vaidya, and James Evans. 2023. In her Shoes: Gendered Labelling in Crowdsourced Safety Perceptions Data from India. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 183–192. https://doi.org/10.1145/3593013.3593987

[133] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. 2023. Representation Bias in Data: A Survey on Identification and Resolution Techniques. *Comput. Surveys* 55, 13s (July 2023), 293:1–293:39. https://doi.org/10.1145/3588433

[134] Reza Shahriari, Yichi Yang, Danish Nisar Ahmed Tamboli, Michael Perez, Yuheng Zha, Jinyu Hou, Mingkai Deng, Eric D Ragan, Jaime Ruiz, Daisy Zhe Wang, et al. 2025. MuCHEx: A Multimodal Conversational Debugging Tool for Interactive Visual Exploration of Hierarchical Object Classification. *IEEE Computer Graphics and Applications* (2025).

[135] Mohit Sharma and Amit Jayant Deshpande. 2024. How far can fairness constraints help recover from biased data? In *Proceedings of the 41st International Conference on Machine Learning*. 44515–44544.

[136] Ben Shneiderman. 2020. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems* 10, 4 (Dec. 2020), 1–31. https://doi.org/10.1145/3419764

[137] Ben Shneiderman. 2020. Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction* 12, 3 (2020).

[138] Ben Shneiderman. 2022. *Human-centered AI.* Oxford University Press.

[139] Sunzida Siddique, Mohd Ariful Haque, Roy George, Kishor Datta Gupta, Debashis Gupta, and Md Jobair Hossain Faruk. 2024. Survey on Machine Learning Biases and Mitigation Techniques. *Digital* 4, 1 (March 2024), 1–68. https://doi.org/10.3390/digital4010001

[140] Edward Small, Wei Shao, Zeliang Zhang, Peihan Liu, Jeffrey Chan, Kacper Sokol, and Flora Salim. 2024. How Robust is your Fair Model? Exploring the Robustness of Diverse Fairness Strategies. https://doi.org/10.48550/arXiv.2207.04581

[141] Nathalie A. Smuha. 2019. The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence. *Computer Law Review International* 20, 4 (Aug. 2019), 97–106. https://doi.org/10.9785/cri-2019-200402

[142] Jean Y. Song, Raymond Fok, Alan Lundgard, Fan Yang, Juho Kim, and Walter S. Lasecki. 2018. Two Tools are Better Than One: Tool Diversity as a Means of Improving Aggregate Crowd Performance. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. Association for Computing Machinery, New York, NY, USA, 559–570. https://doi.org/10.1145/3172944.3172948

[143] Xiang Song, Yuhang He, Songlin Dong, and Yihong Gong. 2024. Non-exemplar Domain Incremental Object Detection via Learning Domain Bias. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 13 (2024), 15056–15065. https://doi.org/10.1609/aaai.v38i13.29427

[144] Ramya Srinivasan. 2024. To See or Not to See: Understanding the Tensions of Algorithmic Curation for Visual Arts. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 444–455. https://dl.acm.org/doi/10.1145/3630106.3658917

[145] Muhammad Suffian and Alessandro Bogliolo. 2022. Investigation and mitigation of bias in explainable AI. In *CEUR Workshop Proceedings*, Vol. 3319. 89–94. https://ceur-ws.org/Vol-3319/paper9.pdf

[146] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3465416.3483305

[147] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxen, Angeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 667–678. https://doi.org/10.1145/3531146.3533132

[148] Stefano Teso and Kristian Kersting. 2019. Explanatory Interactive Machine Learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Honolulu HI USA, 239–245. https://doi.org/10.1145/3306618.3314293

[149] Jacob Thebault-Spieker, Sukrit Venkatagiri, Naomi Mine, and Kurt Luther. 2023. Diverse Perspectives Can Mitigate Political Bias in Crowdsourced Content Moderation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1280–1291. https://doi.org/10.1145/3593013.3594080

[150] Zichen Tian, Zhaozheng Chen, and Qianru Sun. 2024. Learning De-Biased Representations for Remote-Sensing Imagery, Vol. 37. Curran Associates, Inc., 57970–57992. https://proceedings.neurips.cc/paper_files/paper/2024/hash/6a8e10164a90d5c3660c3949289f969a-Abstract-Conference.html

[151] Yun-Da Tsai, Cayon Liow, Yin Sheng Siang, and Shou-De Lin. 2024. Toward More Generalized Malicious URL Detection Models. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 19 (2024), 21628–21636. https://doi.org/10.1609/aaai.v38i19.30161

[152] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. 2023. In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Transactions on Knowledge Discovery from Data* 17, 3 (March 2023), 35:1–35:27. https://doi.org/10.1145/3551390

[153] Ding Wang, Shantanu Prabhat, and Nithya Sambasivan. 2022. Whose AI Dream? In search of the aspiration in data annotation.. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3491102.3502121

[154] Jialu Wang, Yang Liu, and Caleb Levy. 2021. Fair Classification with Group-Dependent Label Noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 526–536. https://doi.org/10.1145/3442188.3445915

[155] Xiuling Wang and Wendy Hui Wang. 2022. Providing Item-side Individual Fairness for Deep Recommender Systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 117–127. https://doi.org/10.1145/3531146.3533079

[156] Zichong Wang, Nripsuta Saxena, Tongjia Yu, Sneha Karki, Tyler Zetty, Israat Haque, Shan Zhou, Dukka Kc, Ian Stockwell, Xuyu Wang, Albert Bifet, and Wenbin Zhang. 2023. Preventing Discriminatory Decision-making in Evolving Data Streams. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 149–159. https://doi.org/10.1145/3593013.3593984

[157] Linda F. Wightman. 1998. *LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series.* Technical Report.

[158] Ziwei Wu and Jingrui He. 2022. Fairness-aware Model-agnostic Positive and Unlabeled Learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1698–1708. https://doi.org/10.1145/3531146.3533225

[159] Ruicheng Xian, Qiaobo Li, Gautam Kamath, and Han Zhao. 2024. Differentially private post-processing for fair regression. In *Proceedings of the 41st International Conference on Machine Learning*. 54212–54235. https://dl.acm.org/doi/10.5555/3692070.3694296

[160] Ruicheng Xian, Lang Yin, and Han Zhao. 2023. Fair and optimal classification via post-processing. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23, Vol. 202)*. JMLR.org, Honolulu, Hawaii, USA, 37977–38012.

[161] Wei Xu. 2019. Toward human-centered AI: a perspective from human-computer interaction. *Interactions* 26, 4 (June 2019), 42–46. https://doi.org/10.1145/3328485

[162] Yuancheng Xu, Chenghao Deng, Yanchao Sun, Ruijie Zheng, Xiyao Wang, Jieyu Zhao, and Furong Huang. 2024. Adapting static fairness to sequential decision-making. In *Proceedings of the 41st International Conference on Machine Learning*. 54962–54982. https://dl.acm.org/doi/10.5555/3692070.3694332

[163] Cheng Yang, Jixi Liu, Yunhe Yan, and Chuan Shi. 2024. FairSIN: Achieving Fairness in Graph Neural Networks through Sensitive Information Neutralization. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 8 (2024), 9241–9249. https://doi.org/10.1609/aaai.v38i8.28776

[164] Mingzhe Yang, Hiromi Arai, Naomi Yamashita, and Yukino Baba. 2024. Fair Machine Guidance to Enhance Fair Decision Making in Biased People. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18. https://dl.acm.org/doi/10.1145/3613904.3642627

[165] Zhenhuan Yang, Yan Lok Ko, Kush R. Varshney, and Yiming Ying. 2023. Minimax AUC Fairness: Efficient Algorithm with Provable Convergence. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 10 (June 2023), 11909–11917. https://doi.org/10.1609/aaai.v37i10.26405

[166] An Zhang, Wenchang Ma, Xiang Wang, and Tat-Seng Chua. 2022. Incorporating Bias-aware Margins into Contrastive Loss for Collaborative Filtering. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., 7866–7878.

[167] Hongjing Zhang and Ian Davidson. 2021. Towards Fair Deep Anomaly Detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 138–148. https://doi.org/10.1145/3442188.3445878

[168] Jie Zhang, Xiaosong Ma, Song Guo, and Wenchao Xu. 2023. Towards unbiased training in federated open-world semi-supervised learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23, Vol. 202)*. JMLR.org, Honolulu, Hawaii, USA, 41498–41509.

[169] Lujing Zhang, Aaron Roth, and Linjun Zhang. 2024. Fair risk control. In *Proceedings of the 41st International Conference on Machine Learning*. 59783–59805. https://dl.acm.org/doi/10.5555/3692070.3694541

[170] Yixuan Zhang, Boyu Li, Zenan Ling, and Feng Zhou. 2024. Mitigating Label Bias in Machine Learning: Fairness through Confident Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 15 (2024), 16917–16925. https://doi.org/10.1609/aaai.v38i15.29634

[171] Shiji Zhao, Ranjie Duan, Xizhe Wang, and Xingxing Wei. 2024. Improving Adversarial Robust Fairness via Anti-Bias Soft Label Distillation, Vol. 37. Curran Associates, Inc., 89125–89149.

[172] Yuying Zhao, Yu Wang, and Tyler Derr. 2023. Fairness and Explainability: Bridging the Gap towards Fair Model Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 9 (June 2023), 11363–11371. https://doi.org/10.1609/aaai.v37i9.26344

[173] Quan Zhou, Jakub Marecek, and Robert N. Shorten. 2021. Fairness in Forecasting and Learning Linear Dynamical Systems. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 12 (2021), 11134–11142. https://doi.org/10.1609/aaai.v35i12.17328

[174] Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang. 2024. Enhancing Zero-Shot Vision Models by Label-Free Prompt Distribution Learning and Bias Correcting, Vol. 37. Curran Associates, Inc., 2001–2025.