Bias Measurement and Algorithmic Fairness in AI Systems: A Survey

www.surveyx.cn

Abstract

The survey paper explores the multifaceted domain of bias measurement, fairness metrics, algorithmic fairness, and ethical AI, emphasizing the critical need for transparency, accountability, and ethical standards in AI systems. It examines the historical context of bias and fairness, providing a comprehensive analysis of key definitions and concepts, including bias measurement, fairness metrics, algorithmic fairness, ethical AI, and responsible AI. The survey delves into statistical and computational techniques for bias measurement, highlighting the strengths and weaknesses of various fairness metrics and their application across diverse domains such as healthcare, finance, and criminal justice. It discusses ethical considerations in AI development, underscoring the importance of stakeholder engagement and multidisciplinary collaboration in promoting responsible AI practices. The paper also reviews bias mitigation strategies, categorizing them into pre-processing, in-processing, and post-processing techniques, each contributing to equitable AI outcomes. The survey concludes by synthesizing key findings and offering recommendations for future research, emphasizing the need for refined methodologies, interdisciplinary collaboration, and educational resources to advance the ethical development and deployment of AI technologies. By integrating fairness and bias mitigation strategies into the AI lifecycle, the survey highlights the potential of AI to transform global sectors responsibly while recognizing the risks of algorithmic bias that could exacerbate social inequities.

1 Introduction

1.1 Significance of Transparency and Ethical Standards

The integration of transparency and ethical standards in AI systems is crucial for fostering fairness and accountability, particularly as these technologies increasingly impact decision-making across various sectors. Transparency elucidates the decision-making pathways of AI, facilitating the identification and rectification of inherent biases, especially in high-stakes areas like healthcare and finance [1].

Ethical standards align AI systems with societal norms and values, thereby fostering trust through adherence to fairness, accountability, and privacy principles. Operationalizing these ethical principles is essential for building public trust and ensuring sustainable innovation in AI technologies [2]. Developing comprehensive metrics that operationalize accountability can bridge gaps in responsible AI practices, particularly in deploying large-scale generative AI models [3].

The interplay between algorithmic fairness and explainability underscores the necessity for responsible practices that address these dimensions [4]. A nuanced understanding of ethical standards, informed by philosophical critique, transcends simplistic definitions and is vital for advancing fairness and accountability in AI systems, enabling methodologies that minimize risks associated with unfair bias and discrimination [5, 1].

Analyzing fairness properties in machine learning systems is indispensable for promoting accountability and transparency, foundational to ethical AI [6]. As organizations increasingly adopt AI

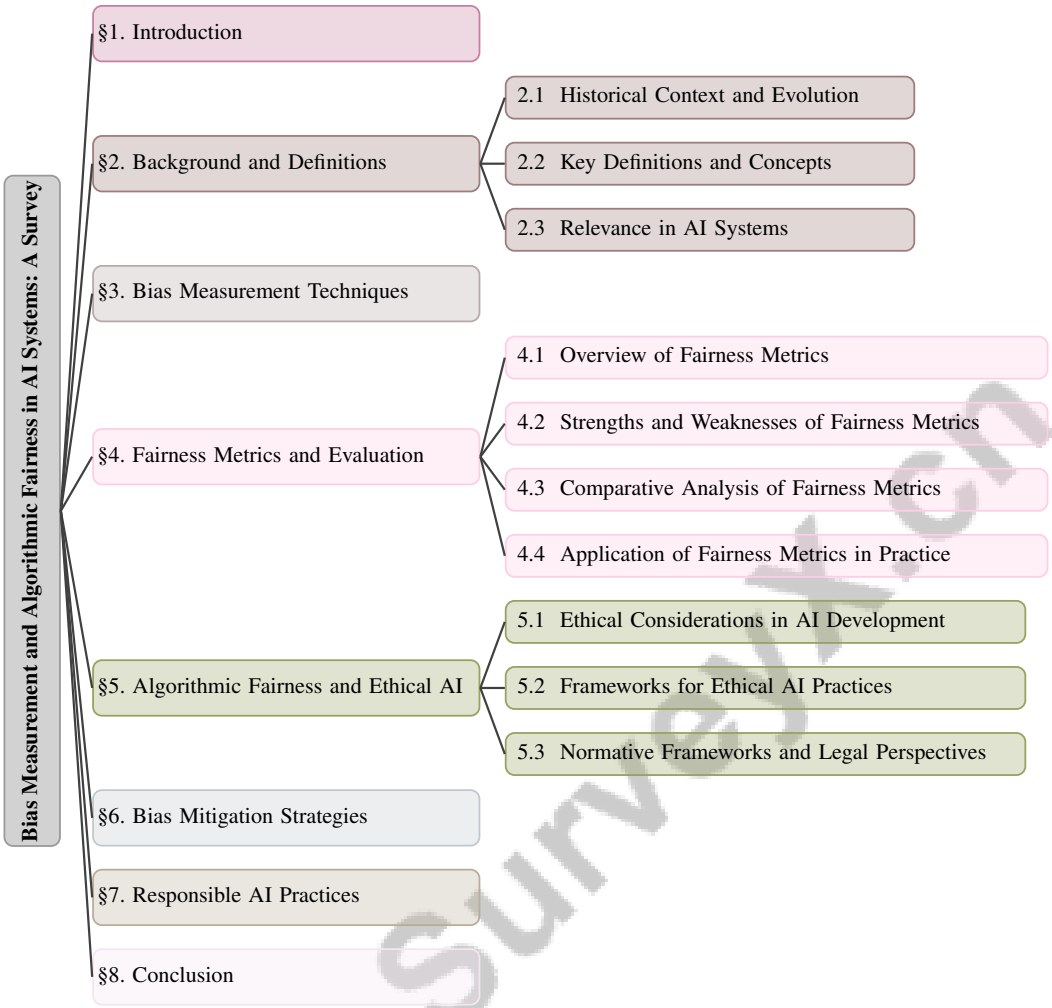


Figure 1: chapter structure

technologies, the ethical implications must be carefully considered to mitigate algorithmic bias and ensure equitable outcomes [7]. Incorporating ethical considerations into AI assessments advances responsible development and dissemination of AI technologies [8].

The shift from objective mathematical definitions of fairness to subjective perceptions that vary across individuals and contexts emphasizes the importance of transparency and ethical standards [9]. This shift is essential for addressing the complex nature of fairness in AI systems, ensuring responsible development and deployment.

1.2 Structure of the Survey

This survey provides a comprehensive analysis of bias measurement and algorithmic fairness in AI systems, focusing on practical applications and sociotechnical implications [10]. It begins with an exploration of the historical context and evolution of bias and fairness, establishing a nuanced understanding of these concepts. Following this, key definitions relevant to contemporary AI systems are outlined, framing subsequent discussions [11].

The survey transitions into examining bias measurement techniques, emphasizing statistical and computational methods while considering interdisciplinary approaches to bias detection challenges [12]. This is complemented by an analysis of fairness metrics and evaluation methods, discussing their strengths, weaknesses, and real-world applications [13]. A comparative analysis assesses the effectiveness of these metrics across various contexts [14].

Additionally, the survey delves into algorithmic fairness and ethical AI, discussing ethical frameworks that inform responsible practices. This includes context-based and society-centered approaches, alongside an examination of normative frameworks and legal perspectives [15], categorized into legal, ethical, and personal tolerances [16]. The section on bias mitigation strategies reviews pre-processing, in-processing, and post-processing techniques designed to reduce bias in AI systems [17].

In the penultimate section, responsible AI practices are scrutinized, highlighting frameworks, stakeholder engagement, and case studies demonstrating successful fairness implementations in AI [18]. The survey concludes by synthesizing key findings and offering recommendations for future research and practical applications in algorithmic fairness and bias measurement, acknowledging AI's potential to transform global health while recognizing the risks of algorithmic bias that may exacerbate social inequities.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Historical Context and Evolution

The progression of bias and fairness in AI is tied to the increased reliance on algorithmic decision-making, which has replaced human judgment in sectors like healthcare, finance, and education. This transition often introduces systematic biases, leading to unfair outcomes, especially concerning sensitive attributes such as race, gender, or socioeconomic status. In healthcare, biases in AI/ML pipelines can lead to inequitable patient outcomes, highlighting the need for fairness in healthcare algorithms. Similarly, algorithmic bias in financial services, including credit scoring and loan approvals, necessitates robust fairness metrics and frameworks [13]. In education, AI systems have been shown to perpetuate biases, resulting in inequitable outcomes for students from diverse backgrounds [19]. Historically, algorithmic fairness has been characterized by confusion and a lack of comprehensive analysis, particularly in defining and achieving fairness in decision-making systems [20]. Recent advancements aim to refine fairness definitions and develop effective mitigation strategies [21]. The widespread presence of algorithmic bias has raised ethical and social concerns, prompting systematic investigations into its impacts [22].

Algorithmic bias in business analytics underscores the need for ongoing scrutiny and improvement to ensure fair treatment based on sensitive attributes [23]. Within National Statistical Organizations (NSOs), machine learning adoption has highlighted the historical context of algorithmic fairness as these organizations strive to balance innovation with ethical considerations [24]. The evolution of fairness in AI reflects a complex interplay of technological advancement and societal values, necessitating ongoing efforts to refine fairness metrics and frameworks [25]. Achieving Trustworthy and Responsible AI, particularly regarding bias and fairness, remains a critical concern in the development of ethical AI systems [26]. Understanding factors influencing perceived fairness, including algorithm outcomes and development procedures, is essential for advancing fairness in AI [27].

2.2 Key Definitions and Concepts

Bias measurement and algorithmic fairness are crucial for developing AI systems that deliver equitable outcomes across diverse demographic groups. *Bias Measurement* involves systematically identifying and quantifying disparities in AI outcomes, crucial for preventing discrimination based on protected attributes such as race, gender, or socioeconomic status. This process requires a nuanced understanding of socio-technical complexities, particularly in sensitive domains like criminal justice, where biases can result in significant injustices [28]. Challenges in identifying and mitigating biases are often exacerbated by inherent biases in AI systems, stemming from flawed data collection methods and the necessity for diverse training data [8].

Fairness Metrics are quantitative tools used to evaluate AI systems by analyzing outcomes across different socio-demographic groups. These metrics are vital for assessing whether AI systems ensure equality of treatment or outcomes, despite varying interpretations of fairness among stakeholders [29]. Key terms such as group fairness, equal confusion fairness test, and confusion parity error are central to evaluating fairness in automated decision systems [30]. However, the lack of a widely accepted similarity metric for most machine learning tasks complicates the effective definition of individual

fairness [28]. Additionally, fairness metrics are often perceived as rigid and limited, necessitating a novel approach to evaluate fairness based on stakeholder interests [4].

Algorithmic Fairness encompasses principles and methodologies designed to ensure AI systems operate without bias, promoting equitable outcomes. This concept addresses systemic unfairness arising from algorithmic decisions influenced by sensitive attributes, which can lead to lost opportunities for disadvantaged groups [31]. Current methodologies often overlook the socially constructed nature of race, minimizing structural aspects of algorithmic unfairness [28]. Algorithmic fairness is increasingly recognized as a distinct quality dimension in frameworks like the Quality Framework for Statistical Algorithms (QF4SA) [1]. Furthermore, algorithmic fairness, bias, and ethical considerations significantly influence business decision-makers' willingness to adopt AI technologies [9].

Ethical AI involves aligning AI systems with societal norms and ethical principles, emphasizing fairness, accountability, and transparency. This alignment is crucial for translating ethical principles into actionable practices throughout the AI lifecycle, from design to deployment [1]. Key terms such as 'demographic data', 'algorithmic bias', and 'fairness' are essential for understanding challenges in AI systems [28]. Ethical AI is linked to Trustworthy Artificial Intelligence (TAI), emphasizing self-assessment techniques and evaluation methods to uphold ethical standards [3]. This survey focuses on the challenges of integrating ethical principles into AI systems, particularly the gap between technical trustworthiness and perceived trust by stakeholders [8].

Responsible AI and *Human-Centered AI* stress the importance of considering human values and societal impacts in AI development and deployment. These concepts are central to addressing entrenched biases in data and algorithms, recognizing the subjective nature of fairness across cultures [4]. Emphasizing diversity and inclusion is crucial for understanding and mitigating biases in AI technologies [1]. Addressing "data documentation debt" is vital for transparency and accountability in algorithmic fairness research [8]. Additionally, key terms such as 'fairness toolkits' and 'cognitive and behavioral motivations' are defined in the context of bias mitigation in machine learning systems [28].

These definitions and concepts form the foundation for efforts to ensure fairness and mitigate bias in AI systems, underscoring the necessity for continuous research and ethical considerations in AI development and deployment, integrating fairness and bias mitigation strategies into the AI lifecycle. The survey identifies gaps in the conceptualization and operationalization of intersectionality in AI fairness literature, highlighting the need for more inclusive and comprehensive approaches [9].

2.3 Relevance in AI Systems

The relevance of fairness, bias measurement, and ethical AI practices in AI systems is critical for ensuring alignment with societal norms and equitable outcomes across various sectors. In areas such as transaction fraud detection, addressing biases is essential, as unfair outcomes can disproportionately affect different demographic groups, necessitating careful consideration of fairness in these systems. The integration of fairness into AI is further complicated by the absence of universally accepted definitions, posing challenges to the practical implementation of fair algorithms [9].

The complexity of perceptions surrounding fairness in algorithmic decisions necessitates a nuanced understanding of these concepts within AI systems. This complexity is heightened by conflicting definitions of fairness and challenges in addressing real-world conditions due to a lack of sensitivity to context and social dynamics. Moreover, varying perceptions of fairness metrics among individuals from diverse backgrounds complicate the selection of appropriate metrics for decision-making [9].

Operationalizing ethical principles in AI systems presents significant challenges, including the intricate interplay between inherent and perceived trust, as well as the unique characteristics of AI technologies that complicate ethical implementation. These challenges underscore the need for a comprehensive approach that integrates high-level ethical guidelines with practical strategies to build trust and ensure accountability among diverse stakeholders. Fostering genuine user control over algorithmic behavior is essential to mitigate distrust and enhance the perceived value of ethical AI, thereby reinforcing public confidence in its applications [32, 33, 2, 34]. The dispersal of responsibility among multiple stakeholders and the opaque nature of AI decision-making processes further complicate the landscape. The interplay between fairness and explainability is crucial for understanding and addressing unfairness in AI systems.

Understanding diverse moral foundations can enhance user acceptance and ethical integrity in algorithmic systems, highlighting the importance of integrating these concepts into AI systems. A comprehensive assessment of AI systems that considers ethical, legal, socio-economic, and cultural issues is essential. These insights emphasize the critical need for a holistic approach to fairness in AI systems, integrating technical, ethical, and societal considerations. This approach is vital for addressing biases and discrimination identified in various AI applications, such as facial recognition and hiring algorithms, which can perpetuate systemic inequality. By focusing on enhancing data quality, developing fair algorithms, and incorporating ethical frameworks like universal human rights, we can work towards creating inclusive and equitable technological solutions that prioritize the protection of individual rights and mitigate harm to marginalized communities [22, 35, 36].

3 Bias Measurement Techniques

3.1 Statistical and Computational Bias Measurement Methods

Statistical and computational methods are pivotal for detecting and quantifying bias in AI systems, ensuring equitable outcomes across varied demographic groups. These methods, applied throughout the machine learning lifecycle, include fairness metrics like Statistical Parity Difference, Disparate Impact, and Equal Opportunity Difference, which offer structured frameworks to evaluate disparities in decision outcomes [31]. These metrics enable analysis of AI's differential impacts on protected attributes, assessing fairness at both individual and group levels.

Dynamic fairness models integrate ethical objectives with formal metrics to address fairness comprehensively, considering long-term effects and ethical implications. The Fairness-Driven Bias Identification (FDBI) method exemplifies the use of statistical techniques in bias measurement, highlighting the importance of demographic factors in legal contexts [37].

Causal reasoning frameworks are crucial for understanding bias propagation in decision-making processes. By utilizing causal models, these frameworks reveal how biases affect outcomes, facilitating the design of interventions that alter causal links between sensitive attributes and decisions. Structural causal models and counterfactual approaches aim to achieve fairness by addressing discrimination's root causes rather than correcting biased predictions post hoc. This proactive strategy evaluates potential interventions, such as policies, considering their effects on disadvantaged groups, enhancing fairness in domains like hiring, healthcare, and law enforcement [38, 39, 40, 41].

The Bias Injection Sandbox Tool (BIST) provides a simulation framework for introducing biases into machine learning data, facilitating the evaluation of fairness interventions [42]. This tool is valuable for stress-testing fairness measures and ensuring robust bias mitigation strategies. The Holistic Fairness Approach (HFA) offers a structured method for fairness assessment in production systems by cataloging costs, benefits, and explicit trade-offs [43].

The Bias Impact Assessment Framework (BIAF), akin to pharmaceutical trials, rigorously evaluates bias in AI systems, emphasizing systematic assessment methods aligned with established protocols in other fields [44]. Categorizing research into AI pipeline stages—pre-processing, in-processing, and post-processing—highlights how intersectionality is operationalized, ensuring a comprehensive bias measurement approach [26].

Developing taxonomies for fairness definitions and bias types enhances statistical and computational bias measurement techniques, structuring existing fairness research to meet ethical standards and societal expectations [45]. The Optimal Transport-based Fairness Framework (OTFF) interpolates between fairness definitions using optimal transport to adjust scores [46]. The 'Attention-Based Attribution Framework' uses an attention mechanism to quantify each attribute's effect on model outcomes, aiding fairness analysis [41].

Fairness toolkits effectively identify and mitigate bias, equipping practitioners with resources to implement fairness measures [47]. The 'Stress Testing Framework' (STF) evaluates fairness by testing machine learning systems under various conditions curated by stakeholders, emphasizing robust evaluation techniques [6]. These methodologies advance fairness in AI systems, promoting transparency, accountability, and ethical AI deployment across domains.

3.2 Challenges and Critiques

Bias measurement in AI systems faces numerous challenges that complicate achieving equitable algorithmic outcomes. A major issue is the lack of standardized metrics and processes for assessing fairness, hindering consistent evaluation across diverse AI applications [31]. This lack of standardization is compounded by the complexity of measuring fairness perceptions, influenced by subjective factors not easily quantifiable [9].

As illustrated in Figure 2, the primary challenges in AI bias measurement include standardization, legal constraints, and dataset complexities. Traditional fairness metrics often lack flexibility, failing to accommodate diverse stakeholder interests, leading to biased outcomes that inadequately reflect different demographic realities [6]. Legal constraints, particularly regarding data privacy, further limit bias measurement by restricting access to necessary demographic data for comprehensive evaluations [48]. These constraints highlight the misalignment between legal standards and AI systems' technical capabilities, complicating efforts to justify and remediate discrimination [46].

The imbalanced nature of datasets, such as those in transaction fraud detection, requires specialized fairness metrics to address inherent complexities [49]. Additionally, a disconnect exists between theoretical fairness solutions and their practical implementation, often due to conflating algorithmic and human biases in current metrics [28]. This disconnect is exacerbated by challenges with causal assumptions and potential post-treatment bias, critical limitations in bias measurement [38].

Addressing AI bias challenges requires ongoing research and refinement of bias measurement frameworks, prioritizing equitable AI systems attuned to fairness complexities and diverse societal contexts. Given biases in critical areas like hiring, lending, and education, collaboration across disciplines is essential for robust mitigation strategies. These strategies should enhance data quality, develop fair algorithms, and ensure diverse perspectives in fairness research, fostering a more inclusive and responsible AI deployment approach [50, 35, 51, 22, 52].

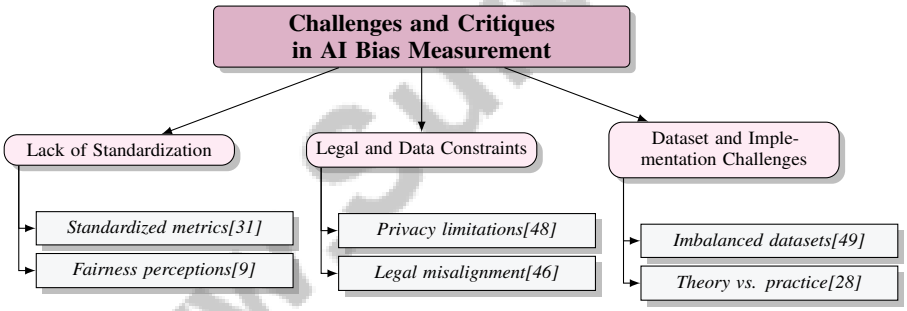


Figure 2: This figure illustrates the primary challenges in AI bias measurement, focusing on standardization, legal constraints, and dataset complexities.

3.3 Interdisciplinary Approaches to Bias Measurement

Interdisciplinary approaches are crucial for addressing AI bias's multifaceted nature, ensuring equitable technology operation across diverse contexts. These approaches draw on insights from computer science, sociology, ethics, and law to develop comprehensive bias detection and mitigation frameworks. The Dynamic Contextual Variable Evaluation (DCVE) method exemplifies this by being applicable across demographics and model types without specialized software [53]. This flexibility integrates diverse perspectives and methodologies, offering a nuanced understanding of AI bias.

Developing decision-making frameworks that customize fairness definitions marks significant progress in interdisciplinary bias measurement. These frameworks address one-size-fits-all limitations by allowing stakeholders to tailor fairness metrics to specific contexts and demographics [54]. This customization captures complex social dynamics influencing fairness perceptions, ensuring AI systems meet different communities' needs.

The FairTest framework integrates interdisciplinary methods by combining investigative primitives and metrics, allowing granular exploration and rigorous statistical bias assessments affecting specific user subpopulations [55]. By facilitating detailed bias examinations at various levels, FairTest

illustrates interdisciplinary approaches' potential to enhance bias measurement precision and effectiveness.

Interdisciplinary methods highlight integrating diverse perspectives to develop equitable and accountable AI technologies. These approaches strengthen bias assessments' rigor while aligning with societal values and ethical standards, fostering fairness in domains like education, hiring, and criminal justice, where biases can lead to systemic discrimination and inequality [22, 35, 19].

4 Fairness Metrics and Evaluation

Evaluating fairness in AI systems is essential for addressing ethical concerns in algorithmic decision-making. As AI technologies become increasingly prevalent across sectors, ensuring equitable operation across diverse demographic groups is paramount. This section delves into the foundational aspects of fairness metrics, crucial for assessing and ensuring fairness in AI applications. By examining these metrics' principles, we can appreciate their role in identifying biases and promoting equitable outcomes. The following subsection offers an overview of commonly used fairness metrics, highlighting their unique characteristics and relevance in algorithmic decision-making.

4.1 Overview of Fairness Metrics

Benchmark	Size	Domain	Task Format	Metric
FairCredit[56]	21,568	Credit Scoring	Classification	Average Odds Difference, Statistical Parity Difference
FBM[57]	81,963	Fairness IN AI	Bias Mitigation	Discrimination Measure, Fairness Score
FAT-ML[58]	36,753	Criminal Justice	Binary Classification	AUC, BA
AIF360[59]	59,842	Bias Mitigation	Binary Classification	Balanced Accuracy, Disparate Impact
FA-ML[60]	1,000	Fairness IN Machine Learning	Binary Classification	Equality of Opportunity, Disparate Impact
AIF360[61]	45,222	Bias Detection	Bias Mitigation	Statistical Parity Difference, Disparate Impact
FMB[62]	48,842	Algorithmic Fairness	Classification	Accuracy, F1-score
AIF360[63]	48,842	Income Prediction	Binary Classification	Mean Difference, BinaryLabelDatasetMetric

Table 1: The table provides a comprehensive overview of various benchmarks utilized in the evaluation of fairness metrics within AI systems. It includes details on benchmark size, domain application, task format, and the specific metrics used to assess fairness, highlighting the diversity and complexity of approaches in algorithmic decision-making. These benchmarks are instrumental in understanding the effectiveness and limitations of fairness metrics across different contexts.

Fairness metrics are vital tools for evaluating AI systems, ensuring equitable outcomes across demographic groups, and identifying biases in algorithmic decision-making. These metrics are particularly critical in sensitive applications like healthcare, credit scoring, and criminal justice, where disparities can worsen existing inequalities and erode public trust. The ethical implications of these metrics extend beyond technical performance, impacting representation and benefit distribution among affected populations [64, 65]. Common metrics include Demographic Parity, Equalized Odds, and Predictive Parity, each offering distinct fairness perspectives by assessing disparities in group outcomes.

Demographic Parity and Equalized Odds focus on whether favorable outcomes are equally distributed across groups. Demographic Parity ensures that the probability of a positive outcome is independent of demographic group, while Equalized Odds requires equal true positive and false positive rates across groups. Approaches like the Equal Confusion Fairness Test and Confusion Parity Error evaluate fairness by examining discrepancies in confusion matrices [66].

The landscape of fairness metrics is diverse, with each presenting unique trade-offs and limitations. Predictive Parity emphasizes equal positive predictive values across groups, contrasting with Demographic Parity and Equalized Odds [67]. The development of compliance-robust fairness, which guarantees equitable outcomes under adversarial compliance patterns, reflects the evolving nature of fairness evaluation [68].

Despite the proliferation of fairness metrics, consensus on their effectiveness remains elusive, as different metrics can yield conflicting results depending on context [49]. The absence of formal principles for metric selection has led to confusion among experts, necessitating ongoing refinement and diversification of fairness metrics [69]. A systematic review reveals these metrics' limitations in capturing the full spectrum of distributive justice theories [70].

Recent advancements have introduced innovative approaches like peer-induced fairness frameworks, which facilitate audits of algorithmic decision-making without requiring access to underlying models [71]. These frameworks underscore the need for continuous adaptation to the dynamic AI landscape. The effectiveness and limitations of various fairness metrics and de-biasing techniques remain a research focus, highlighting the necessity for continuous evaluation to ensure equitable AI system outcomes. Table 1 presents a detailed summary of prominent benchmarks used to evaluate fairness metrics, illustrating the breadth of applications and methodologies in the field of algorithmic fairness.

4.2 Strengths and Weaknesses of Fairness Metrics

Fairness metrics provide quantitative measures for assessing algorithmic equity across demographic groups, yet they present both strengths and limitations influencing their applicability. A significant challenge is the incompatibility and mutual exclusivity of many metrics, complicating the selection process for researchers and practitioners [72]. This complexity necessitates a nuanced understanding of metrics' trade-offs and their alignment with application goals.

Applying fairness constraints to regression models offers a structured approach to comparing biases in judicial decisions, showcasing certain metrics' strengths in specific domains [37]. However, fairness metrics' sensitivity to different bias types varies, with some inadequately detecting bias, thus limiting their effectiveness [73]. This variability underscores the importance of careful metric selection based on specific bias types and contexts.

Existing metrics often misalign with distributive justice principles, raising concerns about their sufficiency in addressing real-world inequities [70]. This misalignment suggests fairness metrics should be viewed not merely as technical tools but as instruments incorporating moral and ethical considerations for holistic fairness evaluation [74]. The approach of attributing fair decisions through interpretable feature attributions, which does not require retraining, presents advantages and limitations in balancing accuracy and fairness [41].

Additionally, national and cultural contexts influence fairness metrics' preference, indicating that metric selection must consider broader societal values and norms [30]. This cultural variability emphasizes adapting fairness metrics to specific ethical and legal frameworks, ensuring AI systems are not only fair but also culturally sensitive.

4.3 Comparative Analysis of Fairness Metrics

Comparative analysis of fairness metrics is essential for understanding their effectiveness in addressing biases across socio-demographic groups within AI systems. Traditional metrics like statistical parity, equalized odds, and predictive parity can only be simultaneously satisfied under limited conditions, particularly when base rates are similar across groups [72]. This limitation underscores the complexity of achieving fairness in algorithmic decision-making (ADM) systems and highlights the need for a nuanced approach to metric selection [75].

The survey explores the application of the four-fifths rule and disparate impact metrics, revealing significant discrepancies in their interpretations across contexts [76]. These discrepancies indicate that while some methods may reduce bias in specific situations, they often fail to address underlying power structures perpetuating inequality [77]. This insight emphasizes the importance of considering broader systemic factors when evaluating fairness metrics' effectiveness.

Traditional fairness metrics tend to obscure inequalities within combined subgroups, highlighting the necessity for intersectional approaches that capture intersectional group fairness's nuanced realities [78]. Intersectional approaches offer a comprehensive understanding of how different socio-demographic factors intersect to influence fairness outcomes, providing a more accurate assessment of biases in AI systems.

The analysis includes a comparison of various explanation methods, differing in effectiveness, approach, and outcomes [4]. These differences underscore the importance of selecting appropriate explanation methods aligning with specific fairness objectives and contexts.

Comparative analysis of fairness metrics underscores the necessity for ongoing adaptation and refinement of these tools, as the intricate relationships among various fairness measures and their compatibility must be thoroughly understood to effectively tackle fairness's multifaceted challenges in AI systems. This understanding is crucial for algorithm designers and users, enabling informed choices about the most suitable fairness metrics for specific applications and objectives, particularly in light of diverse biases in predictive algorithms [60, 72]. By considering different metrics' limitations and strengths, researchers and practitioners can better tailor their approaches to promote equitable outcomes across diverse applications.

4.4 Application of Fairness Metrics in Practice

Fairness metrics are crucial for AI systems' practical evaluation and implementation across various real-world applications. In practice, these metrics assess and mitigate biases, ensuring AI systems deliver equitable outcomes across diverse demographic groups. A key challenge in applying fairness metrics is the inherent incompatibility of many metrics when base rates are unequal across groups, necessitating careful selection tailored to specific contexts and objectives [72].

In healthcare, fairness metrics assess and ensure equitable distribution of medical resources and treatments among diverse patient populations, addressing potential biases in AI algorithms that could lead to disparities in health outcomes. These metrics facilitate AI systems' evaluation, promoting accountability and transparency in clinical practice, while emphasizing diverse data importance and collaborative efforts among healthcare professionals, researchers, and policymakers to mitigate biases and enhance fairness [79, 80, 58, 81, 82]. Metrics like Equalized Odds and Predictive Parity ensure diagnostic algorithms do not disproportionately favor or disadvantage any demographic group, promoting equitable healthcare outcomes.

In financial services, fairness metrics assess credit scoring models' fairness and loan approval processes. By employing metrics like Demographic Parity, financial institutions can identify and address biases that may lead to discriminatory lending practices. This application highlights fairness metrics' critical role in enhancing trust and accountability within financial systems, facilitating stakeholder engagement and collective decision-making. The EARN Fairness framework enables stakeholders to articulate individual preferences and negotiate fairness metrics collaboratively, addressing bias complexities in AI models and promoting equitable outcomes. Evaluating fairness in transaction fraud detection models underscores tailored metrics' necessity considering unique fraud data challenges, revealing significant biases affecting service quality and fraud protection. A nuanced, stakeholder-centered approach to fairness is essential for fostering inclusive and trustworthy financial systems [83, 49, 62].

In criminal justice, fairness metrics evaluate predictive policing and risk assessment algorithms. By employing various fairness metrics, we can systematically identify and mitigate biases in legal data and hiring practices, ensuring these technologies do not reinforce systemic inequalities or perpetuate discrimination against historically marginalized groups, ultimately leading to more equitable outcomes in justice and resource allocation [84, 65, 13, 70, 37].

The application of fairness metrics extends to employment, where they evaluate hiring algorithms and performance evaluation systems. By implementing various fairness metrics in recruitment processes, organizations can systematically identify and mitigate biases from AI-driven decision-making, fostering a more equitable hiring environment that enhances workplace diversity and inclusion. This proactive approach ensures candidates are evaluated fairly, regardless of socio-demographic characteristics, contributing to improved organizational culture and performance [83, 85, 13, 86, 58].

Successful application of fairness metrics requires a nuanced understanding of the specific context and objectives of the AI system. This process entails selecting metrics aligning with desired fairness outcomes and considering the application domain's unique characteristics, including historical discrimination embedded in the data and varying perceptions of fairness across cultural contexts. By employing a principled framework for metric elicitation, practitioners can tailor performance and fairness metrics to the task, context, and population involved. Understanding correlations among different fairness metrics can help identify a representative set effectively addressing fairness nuances

in diverse decision-making scenarios [69, 87, 30, 62]. Through careful implementation, fairness metrics can guide the development and deployment of AI systems that are fair, accountable, and aligned with societal values.

5 Algorithmic Fairness and Ethical AI

The study of algorithmic fairness and ethical AI requires a thorough understanding of the principles guiding the development of AI technologies. It is essential to address fairness, transparency, and accountability as core elements in the responsible creation and deployment of AI systems. The following subsection will explore the integration of ethical considerations throughout the AI lifecycle, focusing on stakeholder engagement and the contextual nature of fairness perceptions.

5.1 Ethical Considerations in AI Development

Incorporating ethical considerations into AI development involves a comprehensive approach that emphasizes fairness, transparency, and accountability to navigate the complexities posed by AI technologies. A pivotal element is engaging stakeholders in defining fairness, as existing metrics often fail to capture the diverse and contextual nature of fairness perceptions [6]. This underscores the need for stakeholder involvement throughout the AI lifecycle to embed ethical principles from design to deployment.

Understanding algorithmic biases, especially through explainability, highlights the necessity of ethical considerations in AI development [4]. Differentiating between algorithmic and human biases is critical, particularly in two-sided marketplaces where fairness dynamics are complex [29]. The 'Responsible AI by Design' methodology exemplifies integrating ethical principles into AI development, focusing on minimizing bias and discrimination risks [1].

Research on AI accountability emphasizes the need for robust governance frameworks and metrics to guide responsible AI practices [3]. The FairCompass method, which incorporates a human-in-the-loop approach, enhances the application of fairness metrics, aligning AI systems with ethical standards [28]. This approach is vital for identifying and mitigating biases that could lead to unfair decisions against marginalized groups [31].

Cultural variability in fairness perceptions further emphasizes the importance of context in decision-making [30]. This sensitivity ensures AI systems are technically sound and socially responsive. However, significant challenges in ethical AI development include the absence of established frameworks for integrating ethical considerations into AI assessments and the focus on technical metrics over broader societal implications [8].

Recognizing that fairness perceptions vary significantly based on demographic factors and service nature complicates the ethical landscape of AI development [9]. This variability necessitates a flexible approach to ethical AI, considering the diverse needs and circumstances of different demographic groups. Prioritizing ethical principles throughout the AI lifecycle ensures AI systems promote equitable outcomes and reflect society's diverse values and expectations.

5.2 Frameworks for Ethical AI Practices

Ethical AI frameworks are crucial for ensuring AI systems align with societal values and ethical standards. These frameworks integrate ethical considerations throughout the AI lifecycle, promoting fairness, transparency, and accountability. Benjamins et al. propose a structured approach, including principles, training, tools, and governance models tailored for ethical AI practices [1]. This approach highlights the necessity of embedding ethical principles at every stage of AI development.

Zhu et al.'s survey compares various ethical frameworks and algorithms, assessing their effectiveness in ensuring trustworthiness and operationalizing ethical principles [2]. This analysis offers insights into existing frameworks' strengths and limitations, guiding the development of robust ethical guidelines.

A multidisciplinary and multi-stakeholder perspective is vital for advancing responsible AI, as noted by Scantamburlo et al. [8]. This perspective calls for cooperation among academia, industry, and the public to ensure comprehensive and ethical AI assessments.

Hannan et al. emphasize integrating subjective fairness perceptions into algorithmic decision-making [9]. This integration acknowledges the complex interplay of factors influencing fairness perceptions, underscoring the importance of diverse perspectives in ethical AI practices.

Critical data studies and Science and Technology Studies (STS) advocate for a nuanced understanding of algorithmic impacts [88]. This perspective encourages interdisciplinary collaboration to address AI fairness's multifaceted nature.

Future research should focus on integrating intersectionality as a fundamental aspect of AI fairness, promoting interdisciplinary collaboration and critical reflexivity [89]. Additionally, integrating socio-technical aspects into fairness evaluations is crucial, especially in developing tailored methods for transaction fraud detection [49].

Categorizing existing research into frameworks based on Responsible AI (RAI) principles and the Software Development Life Cycle (SDLC) phases reveals gaps and overlaps in the current landscape [90]. This categorization highlights the need for comprehensive frameworks addressing all AI development and deployment phases, ensuring ethical considerations are consistently integrated.

Ethical AI frameworks guide developing and deploying AI technologies that are fair, transparent, and accountable. By prioritizing stakeholder engagement and embedding ethical considerations throughout the AI development lifecycle, these frameworks ensure AI systems meet technical specifications, resonate with societal values, and uphold ethical standards. This approach addresses the need for comprehensive guidance across all Software Development Life Cycle (SDLC) phases, fostering trust and accountability in AI technologies [2, 91, 92, 8, 90].

(a) The image is a world map highlighting the prevalence of COVID-19 cases across different countries.[34]

(b) Research Paper Selection Process[91]

Figure 3: Examples of Frameworks for Ethical AI Practices

As shown in Figure 3, the examples illustrate the importance of algorithmic fairness and ethical AI through two visual frameworks. The first image, a world map, shows COVID-19 cases across countries, emphasizing ethical considerations in data visualization to ensure clear and unbiased communication. The second image, a flowchart, outlines the research paper selection process, highlighting a systematic approach to assess AI ethics research comprehensively. These frameworks exemplify the critical role of ethical guidelines in AI development, emphasizing transparency, accountability, and fairness in data-driven decision-making [34, 91].

5.3 Normative Frameworks and Legal Perspectives

Normative frameworks and legal perspectives are essential for advancing algorithmic fairness, providing structured approaches to embedding ethical considerations into AI systems. Frameworks categorizing fairness definitions, such as demographic parity, equal opportunity, and counterfactual fairness, highlight their legal implications and the need for nuanced legal interpretations of fairness interventions. These frameworks harmonize fairness metrics with legal standards, ensuring AI systems adhere to non-discrimination laws and ethical norms. The EARN Fairness framework, for example, facilitates stakeholder engagement in defining fairness metrics, incorporating diverse perspectives. Similarly, a decision-theoretic framework based on UK anti-discrimination law introduces the 'conditional estimation parity' metric, aligning with legal principles to address algorithmic discrimination in high-stakes decisions. Additionally, ongoing research emphasizes the importance of ethical considerations and legal frameworks in educational AI applications, advocating for interdisciplinary

approaches to mitigate biases. These efforts aim to create AI systems that meet legal requirements and promote equity and inclusivity across contexts [83, 93, 35, 94].

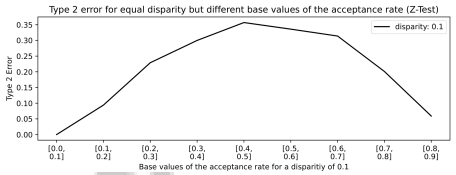
The constructed nature of race, as articulated through historical evidence and theoretical arguments, significantly shapes algorithmic fairness, emphasizing the need for frameworks acknowledging socio-cultural dimensions of fairness [95]. This perspective aligns with the necessity for normative frameworks integrating moral dimensions identified in Moral Foundations Theory, encompassing care, fairness, loyalty, authority, sanctity, and liberty [96]. These moral dimensions provide a comprehensive foundation for developing fairness metrics that are ethically robust and socially responsive.

Framing ethical arguments about algorithmic bias is crucial for influencing managerial decisions regarding AI technologies, highlighting the importance of normative frameworks in guiding ethical AI practices [7]. Understanding factors influencing perceived fairness, such as individual outcomes, transparency, and demographic differences, is central to developing frameworks addressing societal expectations and legal standards [27].

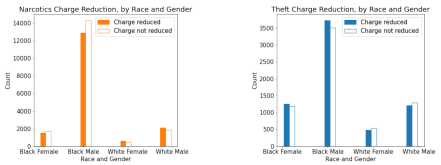
Despite various ethical frameworks, significant gaps remain in translating high-level ethical principles into practical AI applications [2]. These gaps underscore the need for comprehensive frameworks providing practical tools and support across all Software Development Life Cycle (SDLC) phases, ensuring AI systems are developed and deployed ethically and legally [90].

Comparative analysis of existing frameworks and metrics for AI accountability reveals strengths and weaknesses, emphasizing the need for adaptable, context-sensitive frameworks [3]. Additionally, the implications of different fairness notions on algorithmic design highlight the critical role of normative frameworks in guiding the ethical development and deployment of AI systems [29].

Integrating normative frameworks and legal perspectives is crucial for ensuring AI systems are technically sound and aligned with societal values and legal standards. Addressing the intricate relationship among ethical, legal, and social factors facilitates creating AI systems prioritizing fairness, accountability, and ethical responsibility. They provide essential guidelines and principles for practitioners throughout various Software Development Life Cycle (SDLC) stages, although many existing frameworks focus on the Requirements Elicitation phase. This highlights a significant gap in comprehensive support for all stakeholders involved in Responsible AI (RAI) development and deployment, particularly in mitigating biases and ensuring equitable outcomes across sectors such as hiring, lending, and criminal justice. Integrating a human rights-based approach further emphasizes centering discussions around AI's impacts on individuals and communities, ensuring users' rights and well-being are prioritized in design and implementation processes [36, 91, 92, 90, 22].



(a) Type 2 error for equal disparity but different base values of the acceptance rate (Z-Test)[97]



(b) Comparison of Narcotics Charge Reduction by Race and Gender and Theft Charge Reduction by Race and Gender[37]

Figure 4: Examples of Normative Frameworks and Legal Perspectives

As shown in Figure 4, the integration of normative frameworks and legal perspectives is crucial for understanding and addressing biases in decision-making systems. The examples illustrate these challenges. The first example, a line graph, examines Type 2 error rates associated with equal disparity across different acceptance rate base values, offering insights into statistical variances within algorithmic assessments. This analysis, grounded in a Z-Test, highlights how slight changes in acceptance rate parameters can influence error rates and affect fairness outcomes. The second example presents a comparative analysis through bar charts, focusing on narcotics and theft charge reduction by race and gender. This visual comparison underscores disparities in legal outcomes, revealing significant differences in charge reduction rates across demographic groups. Together, these examples highlight the importance of integrating rigorous statistical analysis and equitable legal standards to foster fair and unbiased AI systems [97, 37].

6 Bias Mitigation Strategies

Addressing bias in AI systems requires a comprehensive exploration of mitigation strategies throughout the AI lifecycle. These strategies are categorized into pre-processing, in-processing, and post-processing methods, each playing a critical role in promoting fairness and equity. As illustrated in Figure 5, the hierarchical structure of these bias mitigation strategies emphasizes the significance of each category. Pre-processing techniques ensure dataset integrity, focusing on dataset fairness and representativeness before analysis and model training. In-processing methods refine model training, while post-processing adjusts outputs to ensure equitable results. The figure also highlights essential components such as stakeholder engagement, fairness constraints, and the need for continuous strategy assessment, all of which are vital for ensuring ethical AI development and deployment.

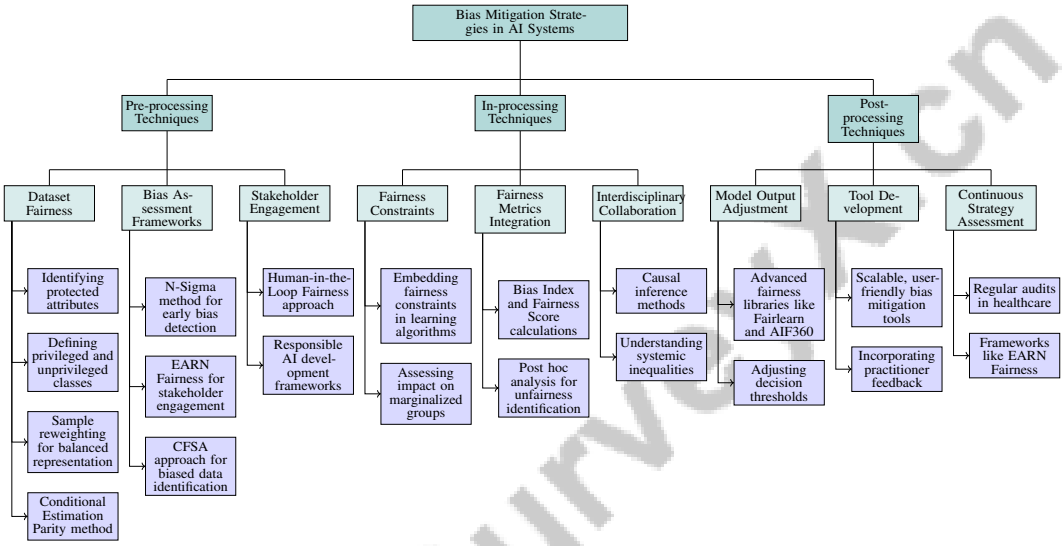


Figure 5: This figure illustrates the hierarchical structure of bias mitigation strategies in AI systems, categorized into pre-processing, in-processing, and post-processing techniques. Each category encompasses specific methods and frameworks aimed at promoting fairness and equity throughout the AI lifecycle. The figure highlights the importance of dataset fairness, stakeholder engagement, fairness constraints, and continuous strategy assessment in ensuring ethical AI development and deployment.

6.1 Pre-processing Techniques

Pre-processing techniques are crucial for ensuring fair datasets in AI systems, rectifying inherent biases to promote equitable outcomes. Identifying protected attributes and defining privileged and unprivileged classes is essential for understanding socio-demographic landscapes and implementing bias mitigation strategies [31]. Sample reweighting adjusts training data distribution to ensure balanced demographic representation, reducing bias and enhancing fairness [26]. The Conditional Estimation Parity method addresses estimation error differences between groups with protected attributes, aligning with anti-discrimination laws.

The N-Sigma method assesses bias before final decisions, mitigating potential biases at the outset [98]. Frameworks like EARN Fairness engage stakeholders in fairness metric discussions before data processing, aligning with societal values [83]. The CFSA approach, using techniques like biased example identification and Counterfactual Bias List construction, emphasizes addressing biased data before model training [99]. Incorporating stakeholder feedback, as seen in the Human-in-the-Loop Fairness approach, improves fairness metrics, highlighting diverse perspectives' importance in developing fair AI systems [18]. Fairness considerations should permeate the entire data processing pipeline, ensuring comprehensive bias mitigation [24].

Responsible AI development frameworks emphasize early stakeholder engagement to ensure fairness and accountability [68]. Ensuring datasets are representative and bias-free contributes to equitable

AI systems aligned with societal norms. Future research should focus on developing inclusive, context-aware methodologies for operationalizing race in algorithmic systems.

6.2 In-processing Techniques

In-processing techniques mitigate bias during model training by embedding fairness constraints into learning algorithms, ensuring equitable outcomes across demographic groups. Assessing AI systems' impact on marginalized groups and analyzing fairness metrics during development is essential for identifying potential biases [15]. Ensuring estimation errors do not disproportionately disadvantage protected groups is crucial for minimizing discriminatory effects in automated decision-making [93]. This aligns legal frameworks with algorithmic fairness definitions, particularly in complex AI system feedback loops [100]. Integrating fairness metrics into the training dataset evaluation process, such as Bias Index and Fairness Score calculations, provides a structured means of assessing fairness during learning [31].

Post hoc analysis to identify unfairness sources can also be an in-processing technique, modifying learning to address identified biases [66]. Causal inference methods evaluate fairness metrics, highlighting their effectiveness in various bias scenarios [73]. Refining fairness metrics and developing effective bias mitigation strategies is emphasized in contexts like transaction fraud detection, where biases have significant real-world implications [49]. Challenges remain in engaging with intersectionality's social contexts, as current studies often lack depth in understanding systemic inequalities perpetuated by AI [89]. This underscores the importance of interdisciplinary collaboration in developing in-processing techniques sensitive to complex social dynamics [101].

6.3 Post-processing Techniques

Post-processing techniques reduce bias in AI systems by adjusting model outputs to promote fairness among demographic groups. This is crucial given biased AI predictions' ethical implications, which can reinforce systemic inequalities in sectors like healthcare, finance, and law enforcement. Advanced fairness libraries like Fairlearn and AIF360 enable systematic bias analysis and mitigation, ensuring AI outcomes do not disproportionately disadvantage specific groups [22, 50, 102, 59]. These strategies address biases persisting despite pre-processing and in-processing interventions, providing additional fairness assurance.

Developing scalable, user-friendly tools for implementing bias mitigation strategies is critical for ongoing fairness evaluations in AI systems. These tools are vital for correcting labeling bias and synthesizing new examples as needed [88]. Such interventions are particularly important in AI-driven recruitment, where continuous research and adaptation maintain fair AI practices [88]. Post-processing techniques involve adjusting decision thresholds or altering predicted outcomes to align with fairness criteria, informed by fairness metrics evaluating model output bias. In healthcare, regular audits and robust bias mitigation strategies ensure AI systems remain fair and responsive to evolving demographic needs [88].

Incorporating practitioner feedback refines post-processing tools and frameworks. Integrating Unbiasedness (UB) measures into software enables real-time bias analysis, facilitating continuous refinement and application across datasets [88]. This ensures AI systems are technically sound and aligned with ethical standards and societal values. Post-processing techniques are integral to ethical AI system deployment. Continuous strategy assessment and enhancement enable developers to navigate bias and fairness challenges in AI, fostering equitable outcomes in applications like education, hiring, and healthcare. This ongoing refinement addresses algorithmic biases perpetuating discrimination against specific demographics. Frameworks like EARN Fairness facilitate stakeholder engagement in defining fairness metrics, while tools like Fairlearn and AIF360 provide actionable bias mitigation insights. A collaborative, interdisciplinary approach ensures AI technologies promote fairness and inclusivity [83, 50, 13, 35, 102].

7 Responsible AI Practices

7.1 Frameworks and Guidelines for Responsible AI

Frameworks and guidelines for responsible AI are essential for aligning AI systems with ethical standards and societal values. These frameworks emphasize integrating ethical considerations throughout the AI lifecycle to promote fairness, transparency, and accountability [103]. Engaging diverse stakeholders, including practitioners, legal experts, and affected communities, ensures AI systems are equitable and accountable. Embedding fairness throughout the machine learning lifecycle prevents the perpetuation of biases and addresses historical inequalities [52, 101]. Multidisciplinary collaboration among educators, researchers, and developers is crucial for establishing comprehensive frameworks across various domains, such as education [104]. It amalgamates diverse perspectives to tackle ethical challenges posed by AI technologies [92]. Future research should focus on creating standardized frameworks that incorporate technical and socio-ethical considerations for a holistic approach to responsible AI development [105]. Engaging stakeholders and fostering interdisciplinary collaboration effectively guide the creation of fair, accountable AI systems aligned with societal values.

7.2 Stakeholder Engagement and Multidisciplinary Collaboration

Effective responsible AI practices depend on active stakeholder engagement and multidisciplinary collaboration. Diverse perspectives from stakeholders—including practitioners, legal experts, affected communities, and policymakers—are crucial for overseeing algorithmic systems and addressing various needs and concerns [106]. Stakeholder engagement helps recognize and mitigate potential risks and biases in AI systems, ensuring they are technically robust and socially responsible. This involvement enhances transparency and accountability by incorporating insights from non-technical stakeholders, informing AI design, implementation, and evaluation. Integrating stakeholder perspectives aligns AI applications with societal values, resulting in responsible and inclusive outcomes [18, 103, 83, 8]. Multidisciplinary collaboration synthesizes insights from computer science, sociology, ethics, and law to address governance, fairness, transparency, and public engagement issues, strengthening the ethical framework guiding AI design and deployment [103, 92, 91, 8]. Initiatives like the Observatory on Society and Artificial Intelligence (OSAI) exemplify efforts to promote a comprehensive understanding of AI's implications, advocating for a holistic approach encompassing ethical, legal, social, and cultural dimensions. This collaboration is essential for tackling AI's complex challenges, allowing experts to develop innovative solutions that enhance fairness, reduce bias, and ensure alignment with societal values.

7.3 Case Studies and Practical Applications

Real-world case studies illustrate the practical implementation of responsible AI practices, highlighting successes and challenges in deploying ethical AI systems. These case studies analyze how fairness, transparency, and accountability are implemented in AI technologies across various sectors, showcasing challenges, mitigation strategies, and ethical implications of biases [19, 91, 35, 22, 102]. In healthcare, fairness metrics promote equitable access to medical resources and treatments, identifying and mitigating biases in AI algorithms to ensure fairer health outcomes [80, 35, 81, 102, 82]. Metrics such as Equalized Odds and Predictive Parity enable healthcare providers to evaluate and adjust diagnostic algorithms, preventing biases and enhancing trust in AI systems. In finance, AI systems assess fairness in credit scoring models and loan approvals, using metrics like Demographic Parity to identify and address biases, ensuring compliance with regulatory frameworks and promoting equitable access to credit [56, 107, 94]. This application underscores responsible AI's role in fostering trust and accountability within financial systems. The criminal justice system benefits from responsible AI by using fairness metrics to evaluate predictive policing and risk assessment algorithms, identifying biases and promoting equitable justice outcomes [108, 37]. In employment, AI-driven recruitment and performance evaluation systems implement fairness metrics to address biases, ensuring equitable candidate evaluation and promoting diversity [83, 85, 13, 86]. These case studies demonstrate the benefits of integrating fairness, transparency, and accountability into AI systems, ensuring they respect individual rights while navigating governance and societal implications [32, 91, 109, 110, 34].

8 Conclusion

8.1 Future Directions and Recommendations

Advancing algorithmic fairness and bias measurement requires a focus on refining methodologies and tools that ensure ethical AI development and deployment. Enhancing the 'Responsible AI by Design' framework, which prioritizes the integration of standardized fairness metrics and the development of automated ethical AI tools, is crucial for embedding ethical considerations throughout AI systems' lifecycles. Empirical validation of these metrics is necessary to confirm their applicability across various contexts, necessitating the adaptation of existing catalogs to cover a wider range of applications. This includes the creation of sophisticated scoring mechanisms to evaluate fairness more effectively. Domain-specific tools, such as FairCompass, should be refined to operationalize fairness in diverse sectors.

Addressing ethical considerations in the selection of protected attributes and refining fairness metrics is essential for advancing algorithmic fairness. Expanding research to include diverse countries and scenarios, along with investigating cultural influences on fairness metric preferences, will enhance our understanding of fairness across different cultural and demographic contexts. This approach will foster a comprehensive understanding of fairness perceptions globally.

Strengthening interdisciplinary collaboration and developing educational resources are vital for increasing awareness of responsible AI among stakeholders. By fostering collaboration between academia, industry, and the public sector, future research can develop concrete ethical AI practices informed by diverse perspectives and expertise. Investigating demographic differences in fairness perceptions and testing these across varied populations will enrich our understanding of fairness in AI systems. This research will provide valuable insights into how different groups perceive and experience fairness, guiding the development of more inclusive and equitable AI technologies.

References

[1] Richard Benjamins, Alberto Barbado, and Daniel Sierra. Responsible ai by design in practice, 2019.

[2] Liming Zhu, Xiwei Xu, Qinghua Lu, Guido Governatori, and Jon Whittle. Ai and ethics – operationalising responsible ai, 2021.

[3] Boming Xia, Qinghua Lu, Liming Zhu, Sung Une Lee, Yue Liu, and Zhenchang Xing. Towards a responsible ai metrics catalogue: A collection of metrics for ai accountability, 2024.

[4] Christos Fragkathoulas, Vasiliki Papanikou, Danae Pla Karidi, and Evaggelia Pitoura. On explaining unfairness: An overview, 2024.

[5] Henry Cerbone. Providing a philosophical critique and guidance of fairness metrics, 2021.

[6] David Lopez-Paz, Diane Bouchacourt, Levent Sagun, and Nicolas Usunier. Measuring and signing fairness as performance under multiple stakeholder distributions, 2022.

[7] Bo Cowgill, Fabrizio Dell'Acqua, and Sandra Matz. The managerial effects of algorithmic fairness activism, 2020.

[8] Teresa Scantamburlo, Atia Cortés, and Marie Schacht. Progressing towards responsible ai, 2020.

[9] Jacqueline Hannan, Huei-Yen Winnie Chen, and Kenneth Joseph. Who gets what, according to whom? an analysis of fairness perceptions in service allocation, 2021.

[10] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems, 2023.

[11] Boris Ruf and Marcin Detyniecki. Towards the right kind of fairness in ai, 2021.

[12] Tiago Palma Pagano, Rafael Bessa Loureiro, Fernanda Vitória Nascimento Lisboa, Gustavo Oliveira Ramos Cruz, Rodrigo Matos Peixoto, Guilherme Aragão de Sousa Guimarães, Lucas Lisboa dos Santos, Maira Matos Araujo, Marco Cruz, Ewerton Lopes Silva de Oliveira, et al. Bias and unfairness in machine learning models: a systematic literature review. *arXiv preprint arXiv:2202.08176*, 2022.

[13] Alessandro Fabris, Nina Baranowska, Matthew J. Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J. Biega. Fairness and bias in algorithmic hiring: a multidisciplinary survey, 2024.

[14] Luca Deck, Astrid Schomäcker, Timo Speith, Jakob Schöffer, Lena Kästner, and Niklas Kühl. Mapping the potential of explainable ai for fairness along the ai lifecycle, 2024.

[15] David Leslie, Cami Rincon, Morgan Briggs, Antonella Perini, Smera Jayadeva, Ann Borda, SJ Bennett, Christopher Burr, Mhairi Aitken, Michael Katell, et al. Ai fairness in practice. *arXiv preprint arXiv:2403.14636*, 2024.

[16] Renqiang Luo, Tao Tang, Feng Xia, Jiaying Liu, Chengpei Xu, Leo Yu Zhang, Wei Xiang, and Chengqi Zhang. Algorithmic fairness: A tolerance perspective, 2024.

[17] Farzaneh Dehghani, Mahsa Dibaji, Fahim Anzum, Lily Dey, Alican Basdemir, Sayeh Bayat, Jean-Christophe Boucher, Steve Drew, Sarah Elaine Eaton, Richard Frayne, Gouri Ginde, Ashley Harris, Yani Ioannou, Catherine Lebel, John Lysack, Leslie Salgado Arzuaga, Emma Stanley, Roberto Souza, Ronnie de Souza Santos, Lana Wells, Tyler Williamson, Matthias Wilms, Zaman Wahid, Mark Ungrin, Marina Gavrilova, and Mariana Bento. Trustworthy and responsible ai for human-centric autonomous decision-making systems, 2024.

[18] Evdoxia Taka, Yuri Nakao, Ryosuke Sonoda, Takuya Yokota, Lin Luo, and Simone Stumpf. Human-in-the-loop fairness: Integrating stakeholder feedback to incorporate fairness perspectives in responsible ai, 2024.

[19] Emily Barnes and James Hutson. Navigating the ethical terrain of ai in higher education: Strategies for mitigating bias and promoting fairness. In *Forum for Education Studies*, volume 2, 2024.

[20] Jella Pfeiffer, Julia Gutschow, Christian Haas, Florian Möslein, Oliver Maspfuhl, Frederik Borgers, and Suzana Alpsancar. Algorithmic fairness in ai: an interdisciplinary view. *Business & Information Systems Engineering*, 65(2):209–222, 2023.

[21] Samer B. Nashed, Justin Svegliato, and Su Lin Blodgett. Fairness and sequential decision making: Limits, lessons, and opportunities, 2023.

[22] Jeff Shuford. Examining ethical aspects of ai: addressing bias and equity in the discipline. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, 3(1):262–280, 2024.

[23] Maria De-Arteaga, Stefan Feuerriegel, and Maytal Saar-Tsechansky. Algorithmic fairness in business analytics: Directions for research and practice, 2022.

[24] Patrick Oliver Schenk and Christoph Kern. Connecting algorithmic fairness to quality dimensions in machine learning in official statistics and survey production, 2024.

[25] Hilde J. P. Weerts. An introduction to algorithmic fairness, 2021.

[26] Hong Qin, Jude Kong, Wandi Ding, Ramneek Ahluwalia, Christo El Morr, Zeynep Engin, Jake Okechukwu Effoduh, Rebecca Hwa, Serena Jingchuan Guo, Laleh Seyyed-Kalantari, Sylvia Kiwuwa Muyingo, Candace Makeda Moore, Ravi Parikh, Reva Schwartz, Dongxiao Zhu, Xiaoqian Wang, and Yiye Zhang. Towards trustworthy artificial intelligence for equitable global health, 2023.

[27] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences, 2020.

[28] Jessica Liu, Huaming Chen, Jun Shen, and Kim-Kwang Raymond Choo. Faircompass: Operationalising fairness in machine learning, 2023.

[29] YinYin Yu and Guillaume Saint-Jacques. Choosing an algorithmic fairness metric for an online marketplace: Detecting and quantifying algorithmic bias on linkedin, 2022.

[30] Yuya Sasaki, Sohei Tokuno, Haruka Maeda, and Osamu Sakura. Evaluating fairness metrics across borders from human perceptions, 2024.

[31] Avinash Agarwal, Harsh Agarwal, and Nihaarika Agarwal. Fairness score and process standardization: framework for fairness certification in artificial intelligence systems. *AI and Ethics*, 3(1):267–279, 2023.

[32] Peter Cihon, Moritz J. Kleinaltenkamp, Jonas Schuett, and Seth D. Baum. Ai certification: Advancing ethical practice by reducing information asymmetries, 2021.

[33] Thomas Krendl Gilbert, Megan Welle Brozek, and Andrew Brozek. Beyond bias and compliance: Towards individual agency and plurality of ethics in ai, 2023.

[34] Anna Jobin, Marcello Ienca, and Effy Vayena. Artificial intelligence: the global landscape of ethics guidelines, 2019.

[35] Sribala Vidyadhari Chinta, Zichong Wang, Zhipeng Yin, Nhat Hoang, Matthew Gonzalez, Tai Le Quy, and Wenbin Zhang. Fairaied: Navigating fairness, bias, and ethics in educational ai applications, 2024.

[36] Vinodkumar Prabhakaran, Margaret Mitchell, Timnit Gebru, and Iason Gabriel. A human rights-based approach to responsible ai, 2022.

[37] Jackson Sargent and Melanie Weber. Identifying biases in legal data: An algorithmic fairness perspective, 2021.

[38] Aida Rahmattalabi and Alice Xiang. Promises and challenges of causality for ethical machine learning, 2022.

[39] Matt J. Kusner, Chris Russell, Joshua R. Loftus, and Ricardo Silva. Causal interventions for fairness, 2018.

[40] Bhavya Ghai and Klaus Mueller. D-bias: A causality-based human-in-the-loop system for tackling algorithmic bias, 2022.

[41] Ninareh Mehrabi, Umang Gupta, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. Attributing fair decisions with attention interventions, 2021.

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

[43] Chloé Bakalar, Renata Barreto, Stevie Bergman, Miranda Bogen, Bobbie Chern, Sam Corbett-Davies, Melissa Hall, Isabel Kloumann, Michelle Lam, Joaquin Quiñonero Candela, Manish Raghavan, Joshua Simons, Jonathan Tannen, Edmund Tong, Kate Vredenburgh, and Jiejing Zhao. Fairness on the ground: Applying algorithmic fairness approaches to production systems, 2021.

[44] Original research.

[45] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. Algorithmic fairness datasets: the story so far, 2022.

[46] Philip Hacker, Emil Wiedemann, and Meike Zehlike. Towards a flexible framework for algorithmic fairness, 2020.

[47] Gianmario Voria, Stefano Lambiase, Maria Concetta Schiavone, Gemma Catolino, and Fabio Palomba. From expectation to habit: Why do software practitioners adopt fairness toolkits?, 2024.

[48] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. "what we can't measure, we can't understand": Challenges to demographic data procurement in the pursuit of fairness, 2021.

[49] Parameswaran Kamalaruban, Yulu Pi, Stuart Burrell, Eleanor Drage, Piotr Skalski, Jason Wong, and David Sutton. Evaluating fairness in transaction fraud models: Fairness metrics, bias audits, and challenges, 2024.

[50] Tahsin Alamgir Kheya, Mohamed Reda Bouadjenek, and Sunil Aryal. The pursuit of fairness in artificial intelligence models: A survey, 2024.

[51] Jennifer Chien, A. Stevie Bergman, Kevin R. McKee, Nenad Tomasev, Vinodkumar Prabhakaran, Rida Qadri, Nahema Marchal, and William Isaac. (unfair) norms in fairness research: A meta-analysis, 2024.

[52] Brianna Richardson and Juan E Gilbert. A framework for fairness: A systematic review of existing fair ai solutions. *arXiv preprint arXiv:2112.05700*, 2021.

[53] Kristian Lum, Yunfeng Zhang, and Amanda Bower. De-biasing "bias" measurement, 2022.

[54] Boris Ruf, Chaouki Boutharouite, and Marcin Detyniecki. Getting fairness right: Towards a toolbox for practitioners, 2020.

[55] Florian Tramèr, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Fairtest: Discovering unwarranted associations in data-driven applications, 2016.

[56] Darie Moldovan. Algorithmic decision making methods for fair credit scoring, 2023.

[57] Manh Khoi Duong and Stefan Conrad. Measuring and mitigating bias for tabular datasets with multiple protected attributes, 2024.

[58] Marius Miron, Songül Tolan, Emilia Gómez, and Carlos Castillo. Addressing multiple metrics of group fairness in data-driven decision making, 2020.

[59] Christina Hastings Blow, Lijun Qian, Camille Gibson, Pamela Obiomon, and Xishuang Dong. Comprehensive validation on reweighting samples for bias mitigation via aif360, 2023.

[60] Gareth P. Jones, James M. Hickey, Pietro G. Di Stefano, Charanpal Dhanjal, Laura C. Stoddart, and Vlasios Vasileiou. Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms, 2020.

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

[62] Suvodeep Majumder, Joymallya Chakraborty, Gina R. Bai, Kathryn T. Stolee, and Tim Menzies. Fair enough: Searching for sufficient measures of fairness, 2022.

[63] Trisha Mahoney, Kush Varshney, and Michael Hind. *AI fairness*. O'Reilly Media, Incorporated, 2020.

[64] Debjani Saha, Candice Schumann, Duncan C. McElfresh, John P. Dickerson, Michelle L. Mazurek, and Michael Carl Tschantz. Measuring non-expert comprehension of machine learning fairness metrics, 2020.

[65] Terrence Neumann, Maria De-Arteaga, and Sina Fazelpour. Justice in misinformation detection systems: An analysis of algorithms, stakeholders, and potential harms, 2022.

[66] Furkan Gursoy and Ioannis A. Kakadiaris. Equal confusion fairness: Measuring group-based disparities in automated decision systems, 2023.

[67] Zeyu Tang, Jiji Zhang, and Kun Zhang. What-is and how-to for fairness in machine learning: A survey, reflection, and perspective, 2023.

[68] Haosen Ge, Hamsa Bastani, and Osbert Bastani. Rethinking algorithmic fairness for human-ai collaboration, 2025.

[69] Gaurush Hiranandani, Harikrishna Narasimhan, and Oluwasanmi Koyejo. Fair performance metric elicitation, 2020.

[70] Matthias Kuppler, Christoph Kern, Ruben L. Bach, and Frauke Kreuter. Distributive justice and fairness metrics in automated decision-making: How much overlap is there?, 2021.

[71] Dana Pessach and Erez Shmueli. Algorithmic fairness, 2020.

[72] Pratyush Garg, John Villasenor, and Virginia Foggo. Fairness metrics: A comparative analysis, 2020.

[73] J. Henry Hinnefeld, Peter Cooman, Nat Mammo, and Rupert Deese. Evaluating fairness metrics in the presence of dataset bias, 2018.

[74] Hilde Weerts, Lambèr Royakkers, and Mykola Pechenizkiy. Are there exceptions to goodhart's law? on the moral justification of fairness-aware machine learning, 2024.

[75] Corinna Hertweck and Christoph Heitz. A systematic approach to group fairness in automated decision making, 2021.

[76] Elizabeth Anne Watkins, Michael McKenna, and Jiahao Chen. The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness, 2022.

[77] Susan Leavy, Barry O'Sullivan, and Eugenia Siapera. Data, power and bias in artificial intelligence, 2020.

[78] Avijit Ghosh, Lea Genuit, and Mary Reagan. Characterizing intersectional group fairness with worst-case comparisons, 2022.

[79] Carol J McCall, Dave DeCaprio, and Joseph Gartner. The measurement and mitigation of algorithmic bias and unfairness in healthcare ai models developed for the cms ai health outcomes challenge. *medRxiv*, pages 2022–09, 2022.

[80] Sribala Vidyadhari Chinta, Zichong Wang, Xingyu Zhang, Thang Doan Viet, Ayesha Kashif, Monique Antoinette Smith, and Wenbin Zhang. Ai-driven healthcare: A survey on ensuring fairness and mitigating bias. *arXiv preprint arXiv:2407.19655*, 2024.

[81] Daiju Ueda, Taichi Kakinuma, Shohei Fujita, Koji Kamagata, Yasutaka Fushimi, Rintaro Ito, Yusuke Matsui, Taiki Nozaki, Takeshi Nakaura, Noriyuki Fujima, et al. Fairness of artificial intelligence in healthcare: review and recommendations. *Japanese Journal of Radiology*, 42(1):3–15, 2024.

[82] Ibomoiye Domor Mienye, Theo G Swart, and George Obaido. Fairness metrics in ai healthcare applications: a review. In *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 284–289. IEEE, 2024.

[83] Lin Luo, Yuri Nakao, Mathieu Chollet, Hiroya Inakoshi, and Simone Stumpf. Earn fairness: Explaining, asking, reviewing, and negotiating artificial intelligence fairness metrics among stakeholders, 2025.

[84] Amina A. Abdu, Irene V. Pasquetto, and Abigail Z. Jacobs. An empirical analysis of racial categories in the algorithmic fairness literature, 2023.

[85] Dena F Mujtaba and Nihar R Mahapatra. Fairness in ai-driven recruitment: Challenges, metrics, methods, and future directions. *arXiv preprint arXiv:2405.19699*, 2024.

[86] Dena F. Mujtaba and Nihar R. Mahapatra. Fairness in ai-driven recruitment: Challenges, metrics, methods, and future directions, 2024.

[87] Hadis Anahideh, Nazanin Nezami, and Abolfazl Asudeh. Finding representative group fairness metrics using correlation estimations, 2022.

[88] Andrés Domínguez Hernández and Vassilis Galanos. A toolkit of dilemmas: Beyond debiasing and fairness formulas for responsible ai/ml, 2023.

[89] Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang. Factoring the matrix of domination: A critical review and reimagination of intersectionality in ai fairness, 2023.

[90] Vita Santa Barletta, Danilo Caivano, Domenico Gigante, and Azzurra Ragone. A rapid review of responsible ai frameworks: How to guide the development of ethical ai, 2023.

[91] Mohammad Tahaei, Marios Constantinides, Daniele Quercia, and Michael Muller. A systematic literature review of human-centered, ethical, and responsible ai, 2023.

[92] Conrad Sanderson, Qinghua Lu, David Douglas, Xiwei Xu, Liming Zhu, and Jon Whittle. Towards implementing responsible ai, 2023.

[93] Holli Sargeant and Måns Magnusson. Formalising anti-discrimination law in automated decision systems, 2025.

[94] Lisa Koutsoviti Koumeri, Magali Legast, Yasaman Yousefi, Koen Vanhoof, Axel Legay, and Christoph Schommer. Compatibility of fairness metrics with eu non-discrimination laws: Demographic parity conditional demographic disparity, 2023.

[95] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness, 2019.

[96] Kimi Wenzel, Geoff Kaufman, and Laura Dabbish. Beyond fairness: Alternative moral dimensions for assessing algorithms and designing systems, 2023.

[97] Luca Deck, Jan-Laurin Müller, Conradin Braun, Domenique Zipperling, and Niklas Kühl. Implications of the ai act for non-discrimination law and algorithmic fairness, 2024.

[98] Daniel DeAlcala, Ignacio Serna, Aythami Morales, Julian Fierrez, and Javier Ortega-Garcia. Measuring bias in ai models: an statistical approach introducing n-sigma. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1167–1172. IEEE, 2023.

[99] Zichong Wang, Yang Zhou, Israat Haque, David Lo, and Wenbin Zhang. Towards fair machine learning software: Understanding and addressing model bias through counterfactual thinking, 2024.

[100] Giorgos Giannopoulos, Maria Psalla, Loukas Kavouras, Dimitris Sacharidis, Jakub Marecek, German M Matilla, and Ioannis Emiris. Fairness in ai: challenges in bridging the gap between algorithms and law, 2024.

[101] Hilde Weerts, Raphaële Xenidis, Fabien Tarissan, Henrik Palmer Olsen, and Mykola Pechenizkiy. The neutrality fallacy: When algorithmic fairness interventions are (not) positive action, 2024.

[102] Ahmed Rashed, Abdelkrim Kallich, and Mohamed Eltayeb. Analyzing fairness of computer vision and natural language processing models, 2024.

[103] Muneera Bano, Didar Zowghi, Pip Shea, and Georgina Ibarra. Investigating responsible ai for scientific research: An empirical study, 2023.

[104] The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges.

[105] Abdoul Jalil Djiberou Mahamadou and Artem A. Trotsyuk. Revisiting technical bias mitigation strategies, 2024.

[106] Nina Grgić-Hlača, Gabriel Lima, Adrian Weller, and Elissa M. Redmiles. Dimensions of diversity in human perceptions of algorithmic fairness, 2022.

[107] Nengfeng Zhou, Zach Zhang, Vijayan N. Nair, Harsh Singhal, Jie Chen, and Agus Sudjianto. Bias, fairness, and accountability with ai and ml algorithms, 2021.

[108] Alexandra Chouldechova and Max G'Sell. Fairer and more accurate, but for whom?, 2017.

[109] Petar Radanliev and Omar Santos. Ethics and responsible ai deployment, 2023.

[110] Esther Taiwo, Ahmed Akinsola, Edward Tella, Kolade Makinde, and Mayowa Akinwande. A review of the ethics of artificial intelligence and its applications in the united states, 2023.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.