applied sciences

MDPI

Review

AI Fairness in Data Management and Analytics: A Review on Challenges, Methodologies and Applications

Pu Chen [1], Linna Wu [2] and Lei Wang [1,*]

[1] College of Communication Engineering, PLA Army Engineering University, Nanjing 210007, China; puchen0127@163.com
[2] Aerospace System Engineering Shanghai, Shanghai 201109, China; wulinna1214@sina.com
* Correspondence: iponly@126.com

Abstract: This article provides a comprehensive overview of the fairness issues in artificial intelligence (AI) systems, delving into its background, definition, and development process. The article explores the fairness problem in AI through practical applications and current advances and focuses on bias analysis and fairness training as key research directions. The paper explains in detail the concept, implementation, characteristics, and use cases of each method. The paper explores strategies to reduce bias and improve fairness in AI systems, reviews challenges and solutions to real-world AI fairness applications, and proposes future research directions. In addition, this study provides an in-depth comparative analysis of the various approaches, utilizing cutting-edge research information to elucidate their different characteristics, strengths, and weaknesses. The results of the comparison provide guidance for future research. The paper concludes with an overview of existing challenges in practical applications and suggests priorities and solutions for future research. The conclusions provide insights for promoting fairness in AI systems. The information reviewed in this paper is drawn from reputable sources, including leading academic journals, prominent conference proceedings, and well-established online repositories dedicated to AI fairness. However, it is important to recognize that research nuances, sample sizes, and contextual factors may create limitations that affect the generalizability of the findings.

Keywords: AI fairness; bias analysis; data analytics

Citation: Chen, P.; Wu, L.; Wang, L. AI Fairness in Data Management and Analytics: A Review on Challenges, Methodologies and Applications. *Appl. Sci.* **2023**, *13*, 10258. https://doi.org/10.3390/app131810258

Academic Editors: Wenjie Zhang and Zhengyi Yang

Received: 29 July 2023
Revised: 31 August 2023
Accepted: 1 September 2023
Published: 13 September 2023

1. Introduction

1.1. Background

With recent advances in artificial intelligence (AI), decision making has gradually shifted from rule-based systems to machine learning-based developments (e.g., [1–3]), learning patterns from data and performing pattern recognition, inference, or prediction. Although such a new methodological trend is derived from the bias brought by human rules, this bias and unfairness are gradually permeating artificial intelligence in another form, as humans are still involved in collecting the datasets used to train machine learning in the new system [4,5].

Artificial intelligence fairness (AI Fairness) is an issue proposed in response to this status quo, which is intended to prevent different harms (or benefits) to different subgroups, thereby providing a system that both quantifies prejudice and mitigates discrimination against subgroups [6,7]. Questions about AI Fairness are practical and affect the lives around us in many ways. Some decision support systems for credit applications tend to favor certain sociodemographic groups, resulting in people living in certain areas, and people of certain ethnic backgrounds or genders having a certain selection preference for loan approval, which is difficult to make completely objective and fair [8–11]. Meanwhile, disability information is highly sensitive and cannot be shared, thus exacerbating this unfairness due to the opaqueness of the information [12]. Companies may miss out on

many potential talents due to an AI-based recruiting engine that is biased against region, gender, and ethnicity, and even cause the company's team composition to gradually become homogenized in biased elements, thereby losing the advantages of diversity [13,14].

It can be seen that the study of this issue has broad social, political, and economic significance. Once the AI misunderstands the intended task, the problem of value misalignment often ensues, and many social issues and responsibilities will arise.

1.2. Directions

Based on the works in recent years, the conceptual development of AI Fairness has focused on the following directions:

- Fairness and bias [15–17]: Introduction of widely used fairness metrics, such as disparate impact, equal opportunity, and statistical parity. Evaluation of their applicability and limitations in different contexts, contributing to a nuanced understanding of group fairness assessment.
- Algorithmic bias mitigation [18–20]: Exploration of techniques, like pre-processing, in-processing, and post-processing, to mitigate algorithmic biases. Critical analysis of their effectiveness in different scenarios, offering insights into the trade-offs between bias reduction and model performance.
- Fair representation learning [21–23]: Introduction of techniques for learning fair representations, including adversarial debiasing and adversarial learning frameworks. Investigation into their potential for producing fair and informative representations, fostering a deeper comprehension of their role in mitigating biases, understanding the true sources of disparities, aiding in the design of more targeted interventions.

Based on the above conceptual directions, we condense and analyze the methodology and technical analysis involved, and focus on the major key elements in this paper, including definition and problem formulation, bias analysis, fair training, and corresponding applications and practices.

This article undertakes a comprehensive exploration of the critical subject of fairness issues within artificial intelligence (AI) systems. The overarching scope of this survey is to provide an in-depth analysis of the multifaceted landscape of AI fairness, covering its foundational aspects, developmental trajectory, practical applications, and emerging research directions. By delving into these dimensions, the survey aims to shed light on the complex challenges linked to fairness in AI, while offering insights into potential remedies and avenues for future exploration.

1.3. Scope

This article mainly collects the research with the details of the corresponding research plan and methodological route, providing a comprehensive survey of the advancements in the domain of AI Fairness. The scope of this survey is expansive and encompasses diverse facets of AI fairness. It begins by elucidating the foundational background and definition of fairness within the realm of AI systems. Subsequently, the survey ventures into the dynamic landscape of fairness concerns, exploring practical applications and recent advancements. Of particular significance are the domains of bias analysis and fairness training, which are delved into as crucial research directions aimed at ameliorating biases and fostering equitable AI outcomes. The survey encompasses meticulous explanations of the concepts, implementations, characteristics, and practical use cases of each method, thereby providing a comprehensive understanding of their nuances.

Encompassing a retrospective spectrum, the covered literature spans from the most recent contributions (e.g., [24–26]) to the initial inception of pertinent theories, extending as far back as 1993 (e.g., [27–29]). A meticulous curation process led to the inclusion of 310 papers from an extensive pool of 1083 pieces of materials. The selection criteria entailed a thorough assessment of factors, such as the intrinsic significance, perceptible impact, novelty, ingenuity, and citation prevalence of the respective works. These works were methodically categorized and subjected to thorough examination within the manuscript.

The ensuing textual discourse encompasses not only analysis but also the deliberation and the derivation of insightful perspectives.

1.4. Contributions

This article offers an extensive and comparative analysis of diverse approaches, leveraging contemporary research to expound upon their distinct attributes, strengths, and limitations. This comparative exploration not only guides researchers but also informs practitioners, providing them with a nuanced understanding of available methods and aiding their decision making.

Additionally, the survey enriches the discourse on AI fairness by contextualizing its practical implications. By exploring strategies to mitigate bias and enhance fairness in AI systems, it bridges the gap between theoretical foundations and real-world challenges. The survey further discusses the critical subject of challenges and solutions in real-world AI fairness applications, offering insights into the current limitations and potential remedies.

Furthermore, this survey contributes by acknowledging the dynamic nature of the AI fairness landscape and the evolving nature of research and advancements. It underscores the evolving nature of the field and the limitations associated with the evidence derived from reputable sources. While these limitations stem from factors such as research nuances, sample sizes, and contextual intricacies, the survey remains committed to fostering the continuous exploration and understanding of AI fairness.

In essence, this survey encompasses a wide-ranging scope, delving into the genesis, evolution, applications, and challenges of AI fairness. Its multifaceted contributions aim to advance the understanding of fairness in AI, providing valuable insights for both academia and industry in their pursuit of equitable and unbiased AI systems.

1.5. Organization of This Article

The organization of the article is organized as follows. Section 2 introduces the background and definition of AI fairness, and Section 3 formulates the definitions and problems of fairness in AI systems. The main directions of the research of addressing AI fairness, bias analysis, and fair training are reviewed in Sections 4 and 5 with details of corresponding methodologies. Section 6 discusses the measures of migrating the bias and improving fairness in the AI system. Section 7 reviews the related issues and solutions in the practical applications of AI fairness, and corresponding future works are discussed and given. Section 8 concludes the paper.

2. Preliminary

2.1. Status Quo

Although the study of fairness in machine learning is a relatively new topic, it has attracted extensive attention. IBM launched AI Fairness 360 [30–32], which can help detect and mitigate unwanted bias in machine learning models and datasets. It provides around 70 fairness metrics to test for bias and 11 algorithms to reduce bias in datasets and models, thereby reducing software bias and improving its fairness (e.g., [33]).

Facebook has also developed the Fairness Flow tool to detect bias in AI, which works by detecting forms of statistical bias in some of Facebook's commonly used models and data labels, enabling the analysis of how certain types of AI models perform across different groups [34–36]. It defines "bias" as the systematic application of different standards to different groups of people. Given a dataset containing predictions, labels, group membership (for example, gender or age), and other information, Fairness Flow can divide the data used by the model into subsets and estimate its performance.

In 2019, Google also embedded the Fairness Indicators component in a series of AI tools it developed, resulting in tools built on top of TensorFlow model analysis that can regularly calculate and visualize fairness indicators for binary and multi-class classification [37–39].

Although the above work provides tools and theoretical analysis, due to the short research time of this problem and still in the preliminary exploratory stage, there is no

mature standard for how to quantify the risk of AI fairness and little insights in how to make decisions in the case of consensus and controversy with the commonly accepted solutions to the risks.

2.2. Review Methodology

2.2.1. Materials

Aiming at the topic along with the issues above, we collected a variety of current academic and technical materials on AI equity and conducted a synthesis study. The information synthesized in this study comes from a variety of reliable sources. These sources include recent publications in prestigious academic journals, distinguished conference proceedings, and well-established online repositories dedicated to the fairness of AI.

Our comprehensive search strategy includes systematic searches of respected academic databases, including IEEE Xplore, ScienceDirect, ACM Digital Library, Springer, Wiley Library, etc. In addition, we carefully reviewed relevant conference proceedings and authoritative organization websites to ensure research inclusiveness. In a robust and diverse collection, we conducted a meticulous review process that encompassed a wide range of sources. We extensively searched through various databases, culminating in the compilation of 1083 materials, comprising Proceedings, Miscellaneous, Articles, Tech Reports, Books, Ph.D. theses, and Collections. To filter the literature and complete the review, the review process was multifaceted and involved several key criteria.

2.2.2. Criteria

The inclusion and exclusion criteria we use in selecting sources of information are carefully thought out to ensure the relevance and quality of the studies included in our analysis. Our criteria included peer-reviewed academic articles, conference papers, and authoritative reports that explicitly address the fairness of AI in data management and analytics. During the synthesis process, we thoughtfully grouped studies based on thematic affinity, methodology, and the nature of the fairness challenges to be addressed. This grouping strategy facilitates a coherent and well-organized synthesis of the different perspectives in the literature. Our criteria for selecting materials are carefully considered and include the following factors, including the following:

- Duplication: We strive to offer diverse and original content to our audience. To avoid redundancy, we review submissions to ensure that the material we publish does not duplicate the existing content in our collection.
- Ineligible content: Our selection process also involves evaluating whether the submitted content meets our eligibility criteria, including adhering to our guidelines and standards.
- Publishing time: We value timeliness and relevance. We prioritize materials that are current and align with the most recent developments and trends in the respective field.
- Quality of publication: Ensuring the quality of the content we publish is of utmost importance. We assess the accuracy, credibility, and overall value of the material to ensure it meets our quality standards.
- Accessibility: Our goal is to make information accessible to a wide range of readers. We select materials that are well structured, clear, and easily understandable, catering to readers with varying levels of expertise.
- Similarity of content: While covering a broad spectrum of topics, we also strive for variety and distinctiveness in our content selection. We aim to present diverse perspectives and insights to enrich the reader experience.

We adopted PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) in our review, which meticulously outlines the systematic progression of the study identification, screening, eligibility, and inclusion phases, thereby increasing the reproducibility and rigor of the review process. The PRISMA procedure is shown in Figure 1 as a flowchart. After the procedure, reference lists of selected articles in the field are reviewed intensively while identifying potential studies.

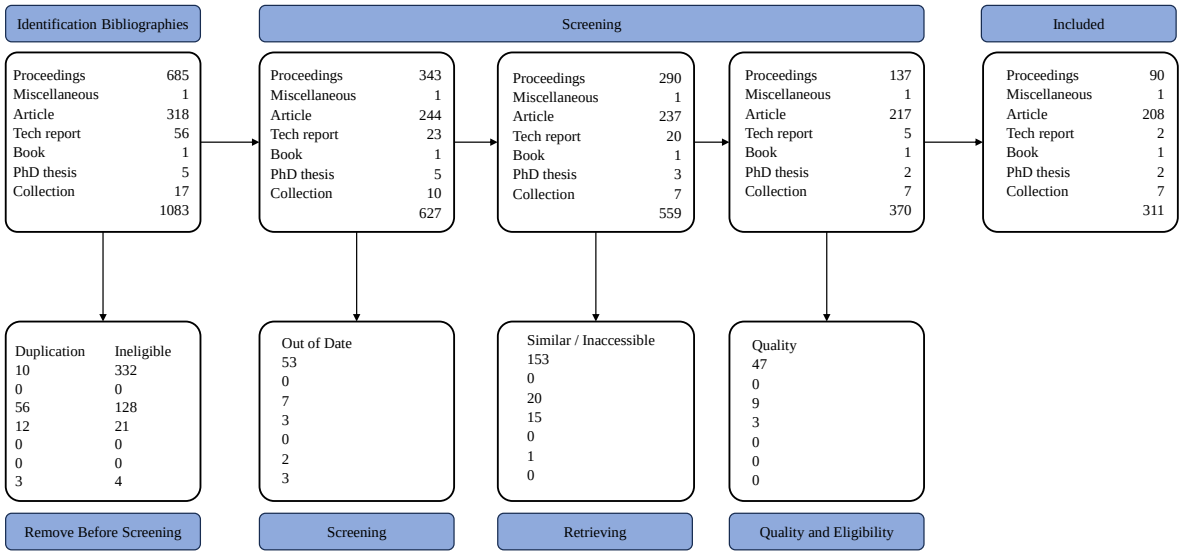

Figure 1. Procedure of preferred reporting items for systematic reviews and meta-analyses.

2.3. Limitations

It is important to recognize that while these sources have contributed significantly to our understanding, there are limitations to the evidence they provide. These limitations stem primarily from nuances in research methodology, sample size, and context, which may affect the generalizability of the conclusions drawn from individual studies. The landscape of AI fairness is dynamic, with research and advancements continually shaping our understanding of its complexities. While our current coverage might have limitations due to the rapid pace of change and ongoing research, please know that we are committed to further studying and exploring this crucial subject. Despite these inherent limitations, our review endeavors to provide a comprehensive and balanced overview of the current state of research related to the fairness of AI. We are dedicated to providing accurate and comprehensive information to readers with the notice of the need to stay engaged with emerging topics like AI fairness.

3. Definition and Problems

3.1. Definition

As AI technologies continue to permeate all aspects of society, ensuring fairness in their decision-making processes is crucial to avoid perpetuating bias and inequality. However, defining what constitutes fairness in AI is a complex and multifaceted task. So far, there are mainly seven types of definitions, including individual fairness [40,41], group fairness [42], equality of opportunity [11], disparate treatment [43], fairness through unawareness [44,45], disparate impact [46], and subgroup fairness [47].

Fairness in AI can be approached through different conceptual lenses. Individual fairness emphasizes equitable treatment for similar individuals, while group fairness aims to avoid disparate treatment based on demographic attributes. Equality of opportunity focuses on consistent predictive accuracy and error rates across various groups, regardless of outcomes.

An alternative approach is fairness through unawareness, achieved by ignoring sensitive attributes in decision making. Despite its intentions, this method might indirectly perpetuate bias present in data. Disparate impact examines whether AI systems disproportionately harm certain groups, irrespective of intent. It aims to uncover biases in outcomes, intentional or not.

To address complex interactions, subgroup fairness evaluates fairness at the intersection of multiple protected attributes, ensuring equitable experiences for diverse subgroups. These conceptions contribute to a comprehensive understanding of fairness in AI and underscore the multifaceted nature of achieving equitable outcomes.

Table 1 summarizes these approaches to fairness in AI. Individual fairness prioritizes personalized treatment, while group fairness targets demographic equity. Fairness through unawareness and equality of opportunity tackle fairness differently—ignoring attributes vs. ensuring equal chances based on qualifications. Disparate impact assesses negative effects, disparate treatment detects unequal treatment. Subgroup fairness navigates complex attribute interactions for equitable outcomes. These concepts collectively enrich the understanding of fairness in AI systems.

Table 1. Fairness definition.

Data Bias	Definition	Main Cause	References
Individual Fairness	Similarity at the individual level	Treat similar individuals similarly	[40,41,48,49]
Group Fairness	Equitable outcomes for demographic groups	Avoid disparities among groups	[42,50,51]
Fairness through Unawareness	Ignoring sensitive attributes	Treat individuals as if attributes are unknown	[44,45,52]
Equality of Opportunity	Equal chances for similar qualifications	Ensure equal chances for outcomes	[11,53,54]
Disparate Impact	Disproportionate negative effects	Evaluate disparities in outcomes	[6,46,55]
Disparate Treatment	Explicit unequal treatment	Detect explicit biases in treatment	[43,56,57]
Subgroup Fairness	Fairness at the intersection of multiple attributes	Consider fairness for multiple groups	[47,58]

3.2. Problems

The risks with AI systems mainly come from data accountability [24,59] and algorithm accountability [60–62]. The connotation of data accountability mainly includes data ownership, storage, use, and sharing, while algorithm accountability emphasizes determining who is responsible for the output of AI algorithms. The interplay of these two risks also raises the question of mission inclusivity [63–65], which mainly focuses on whether the AI system is effective for diverse user populations. Bias effects in machine learning are shown in Figure 2.

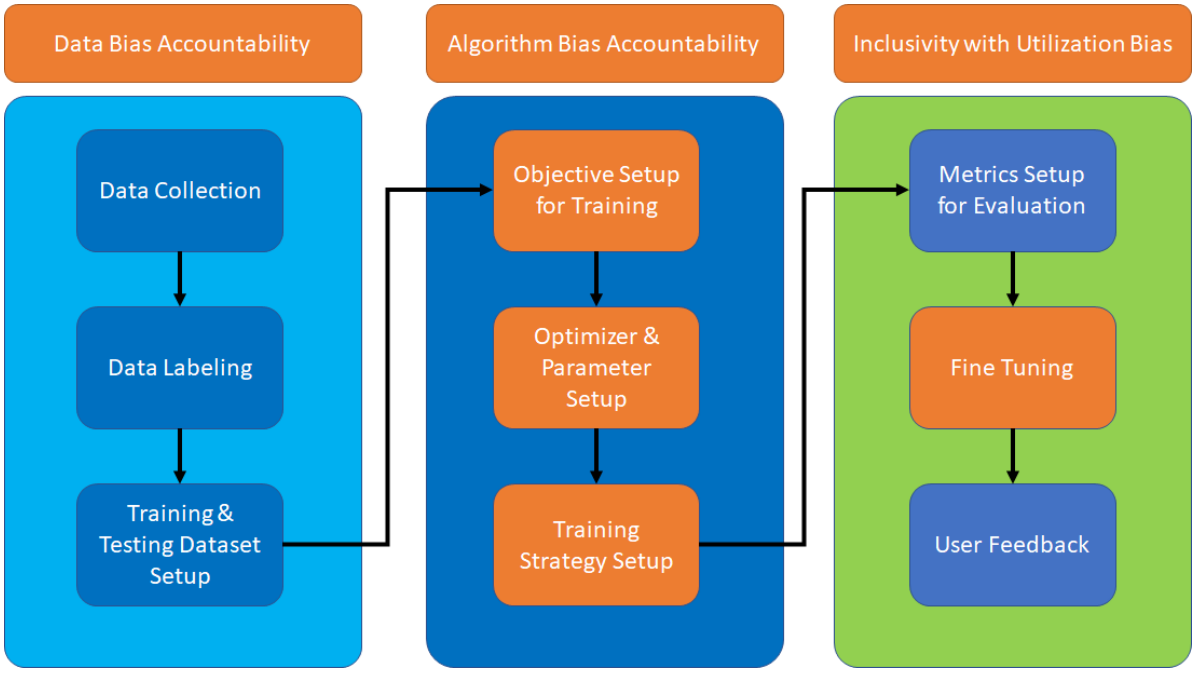

Figure 2. Bias effects in machine learning.

Biases in AI models can arise from data collection, labeling, and partitioning, affecting data integrity. Human factors during training, including optimization objectives and parameter configurations also contribute to bias. Inclusivity applications involve tuning and updates guided by user feedback, which holds substantial influence. Techniques like K-fold mitigate dataset bias, but original bias persists, highlighting the importance of robust data accountability. Subjective optimization objective design exacerbates bias effects. Addressing these issues requires comprehensive strategies for data and model development.

Feedback bias in AI model tuning can arise if participants providing feedback are disproportionately represented in specific communities or feature sets. This concentration can lead to model adjustments aligning more closely with the preferences of that group. Grouping can be based on experimental settings or attributes, aiding the analysis of cognitive and labeling differences' impact on bias patterns. Statistical features synthesized from group evaluations can influence machine learning model outcomes.

The inclusivity of intelligent computing services relies on data and algorithm accountability to mitigate bias and ensure fairness in machine learning processes. While industry practices often exclude sensitive attributes to address fairness concerns, this approach overlooks the potential influence of non-sensitive attributes in reflecting bias. Additionally, evaluating fairness using static test sets poses challenges, including potential incompleteness and inherent bias carried over from existing systems.

Moreover, the feedback loop between machine learning system outputs and inputs can perpetuate and reinforce biases, necessitating the analysis of algorithms in dynamic systems. Label noise further complicates the picture, as large datasets essential for deep network training can inadvertently incorporate incorrect labels, undermining model accuracy and performance.

Addressing these challenges, ongoing research focuses on detecting and mitigating bias while designing fair machine learning mechanisms and intelligent systems. Bias analysis and fair training emerge as critical areas, aiming to enhance the technical understanding and current status of each direction.

4. Bias Analysis

4.1. Data Bias

Data bias is a critical concern in artificial intelligence (AI) systems, as biased data can lead to unfair and discriminatory outcomes. It arises when the training data used to develop AI models are skewed, leading to biased predictions or decisions [66,67]. Biased data can perpetuate historical prejudices and result in discriminatory outcomes. There are five main types bias, including selection bias [68–71], sampling bias [25,72,73], labeling bias [26,74–77], temporal bias [78–81], aggregation bias [82–86], historical bias [52,87–89], measurement bias [4,90–92], confirmation bias, proxy bias, cultural bias, under-representation bias [93–95], and homophily bias [96–98]. Table 2 shows the comparison of the different types of data biases.

Bias in AI models can stem from various data-related sources. Selection bias arises from skewed data representation due to biased collection or incomplete sampling. Aggregation bias results when data from different sources with varying biases are combined without proper consideration. Sampling bias emerges when training data fail to represent the target population adequately. Labeling bias occurs due to errors in annotation, introducing bias into training. Measurement bias originates from inaccuracies during data collection, impacting the model's ability to learn accurately. Temporal and historical biases arise from reflecting outdated societal biases. Unconscious biases, such as cultural bias, lead to biased decisions for diverse groups. Proxy bias uses correlated proxy variables, indirectly introducing bias. Homophily bias reinforces existing patterns in prediction, potentially intensifying bias. Understanding and mitigating these biases are crucial for equitable AI systems.

Table 2. Comparison on data biases.

Data Bias	Definition	Main Cause	Impact on AI	References
Selection Bias	Certain groups are over/under-represented	Biased data collection process	AI models may not be representative, leading to biased decisions	[68–71]
Sampling Bias	Data are not a random sample	Incomplete or biased sampling	Poor generalization to new data, biased predictions	[25,72,73]
Labeling Bias	Errors in data labeling	Annotators' biases or societal stereotypes	AI models learn and perpetuate biased labels	[26,74–76]
Temporal Bias	Historical societal biases	Outdated data reflecting past biases	AI models may reinforce outdated biases	[78–81]
Aggregation Bias	Data combined from multiple sources	Differing biases in individual sources	AI models may produce skewed outcomes due to biased data	[82–85]
Historical Bias	Training data reflect past societal biases	Biases inherited from historical societal discrimination	Model may perpetuate historical biases and reinforce inequalities	[52,87–89]
Measurement Bias	Errors or inaccuracies in data collection	Data collection process introduces measurement errors	Model learns from flawed data, leading to inaccurate predictions	[4,90–92]
Confirmation Bias	Focus on specific patterns or attributes	Data collection or algorithmic bias towards specific features	Model may overlook relevant information and reinforce existing biases	[27,99–102]
Proxy Bias	Indirect reliance on sensitive attributes	Use of correlated proxy variables instead of sensitive attributes	Model indirectly relies on sensitive information, leading to biased outcomes	[42,103–105]
Cultural Bias	Data reflect cultural norms and values	Cultural influences in data collection or annotation	Model predictions may be biased for individuals from different cultural backgrounds	[72,106,107]
Under-representation Bias	Certain groups are significantly underrepresented	Low representation of certain groups in the training data	Model performance is poorer for underrepresented groups	[93–95]
Homophily Bias	Predictions based on similarity between instances	Tendency of models to make predictions based on similarity	Model may reinforce existing patterns and exacerbate biases	[96–98]

4.2. Algorithmic Bias

Algorithmic bias refers to biases inherent in the design and structure of AI models [108,109]. These biases may be unintentionally introduced during the development process, leading to unequal treatment of different groups. The main algorithmic biases include prejudice bias, sampling bias, feedback loop bias, lack of diversity bias, and automation bias.

Prejudice bias arises from biased training data, perpetuating societal stereotypes. Sampling bias stems from data misrepresentation, causing poor generalization. Feedback loop bias is a self-reinforcing cycle, where biased AI predictions lead to biased feedback. Lack of diversity bias emerges from inadequate dataset representation, affecting underrepresented groups. Automation bias involves over-reliance on AI decisions without scrutiny, potentially amplifying underlying biases.

Table 3 summarizes the comparison of different types of algorithmic bias in AI systems, highlighting their definitions and main implications. Prejudice bias originates from biased data collection, reinforcing discrimination. Sampling bias results from non-representative data, causing biased predictions. Feedback loop bias is a self-reinforcing cycle driven by biased predictions and feedback. Lack of diversity bias emerges from homogeneous

training datasets, affecting underrepresented groups. Automation bias is the uncritical reliance on AI decisions, amplifying underlying biases.

Table 3. Algorithmic bias comparison.

Algorithmic Bias	Definition	Main Cause	Impact on AI	References
Prejudice Bias	AI models trained on biased data	Biased training data and societal prejudices	Reinforces biases, leads to discriminatory outcomes	[76,110–112]
Sampling Bias	Data do not represent the target population	Incomplete or skewed sampling methods	Poor generalization, biased predictions	[85,113–115]
Feedback Loop Bias	Self-reinforcing bias cycle in AI predictions	Biased predictions influencing biased feedback	Amplifies biases, perpetuates discrimination	[116–120]
Lack of Diversity Bias	Training on limited or homogeneous datasets	Insufficient representation of diverse groups	Performs poorly for underrepresented groups	[40,121–125]
Automation Bias	Human over-reliance on AI decisions	Blind trust in AI without critical evaluation	Perpetuates biases without human intervention	[126–131]

4.3. User Interaction Bias

User interaction bias occurs when AI systems adapt their behavior based on user feedback, potentially reinforcing and amplifying existing biases [67,132]. It manifests in various forms, each contributing to biased decision making and unequal outcomes in AI systems. Table 4 summarizes the typical user interaction biases, including user feedback bias, underrepresented or biased user data bias, and automation bias.

Table 4. User interaction bias comparison.

Bias Type	Definition	Main Cause	Impact on AI	Reference
User Feedback Bias	User Feedback Bias occurs when biased user feedback or responses influence the behavior of AI systems.	Biased user feedback or responses can be influenced by users' subjective preferences, opinions, or prejudices. The AI system learns from this feedback and incorporates it into its decision-making process.	AI models may generate biased predictions and decisions based on the biased feedback, potentially leading to unequal treatment of certain user groups. User satisfaction and trust in the AI system can be affected by biased outputs.	[116–118]
Biases from Underrepresented or Biased User Data	This bias arises when the data collected from users lack diversity or contain inherent biases, which can lead to biased model predictions and decisions that disproportionately affect certain user groups.	Lack of diversity or inherent biases in user data can result from biased data collection practices, data preprocessing, or historical biases reflected in the data.	AI systems trained on biased user data may produce unfair outcomes, disproportionately impacting specific user groups. Biases in data can lead to the perpetuation and amplification of existing inequalities.	[133–135]
Automation Bias in Human–AI Interaction	Automation bias refers to biased decision making by users when utilizing AI systems, potentially influencing the AI system's outcomes and recommendations.	Automation bias can occur when users over-rely on AI recommendations without critically evaluating or verifying the results. Human trust in AI systems and the perceived authority of the AI can contribute to automation bias.	Automation bias can lead to the uncritical acceptance of AI-generated outputs, even when they are biased or inaccurate. It may result in erroneous or unfair decisions based on AI recommendations. Awareness of automation bias is crucial to avoid blindly accepting AI decisions without human oversight.	[126,128,129]

User feedback bias and bias from underrepresented or biased user data contribute to user interaction biases, influencing AI system behavior and predictions. The interaction between humans and AI, rooted in automation bias, further affects these biases. Notably, user interaction bias and algorithmic bias overlap, as biases from human–computer interaction data impact industrial intelligence models, highlighting their interconvertibility.

5. Fair Training

5.1. Fair Training Methods

In response to the above bias analysis, we hope to be able to develop an AI system without bias by conducting fair training so that we can avoid perpetuating inequalities due to discriminatory appearances caused by biases. Fairness training aims to reduce these biases and promote fair decision making. There are several fair training methods that are currently in common use, including pre-processing fairness [136], in-processing fairness [137], post-processing fairness [46], regularization-based fairness [43], counterfactual fairness [41,45,138].

Pre-processing, in-processing, and post-processing fairness techniques address bias in AI systems from different angles. Pre-processing involves modifying training data to balance group representation. In-processing modifies learning algorithms to integrate fairness. Post-processing adjusts model predictions to align with fairness goals. Additionally, regularization-based methods introduce fairness constraints in optimization, aiming to minimize disparities, while counterfactual fairness measures fairness by assessing outcome consistency for similar individuals across sensitive attributes.

Fair training techniques, as depicted in Table 5, strive to mitigate biases in AI systems by integrating fairness considerations into the training process. Through the incorporation of sensitive attributes and fairness constraints, these methods aim to diminish the impact of such attributes on model predictions, guarding against biased outcomes that could disproportionately affect marginalized groups. The challenge lies in striking a balance between fairness and accuracy, avoiding the compromise of model performance while enhancing fairness.

Table 5. Fair training method comparison.

Fair Training Method	Definition	Implementation	Key Features	References
Pre-processing Fairness	Modifying training data before feeding into the model	Re-sampling, re-weighting, data augmentation	Addresses bias at the data level	[136,139,140]
In-processing Fairness	Modifying learning algorithms or objective functions	Adversarial training, adversarial debiasing	Simultaneously optimizes for accuracy and fairness	[137,141,142]
Post-processing Fairness	Adjusting the model's predictions after training	Re-ranking, calibration	Does not require access to the model's internals	[46,143–145]
Regularization-based Fairness	Adding fairness constraints to the optimization process	Penalty terms in the loss function	Can be combined with various learning algorithms	[43,146,147]
Counterfactual Fairness	Measuring fairness based on changes in sensitive attributes	Counterfactual reasoning	Focuses on individual-level fairness	[45,148,149]

5.2. Pre-Processing Fairness

Preprocessing fairness applications involve modifying training data before feeding them into an AI model to reduce bias and promote fairness. These techniques focus on addressing biases in the data themselves, which can lead to fairer model results. Common

methods include resampling, reweighting, data augmentation, fairness-aware clustering, and synthetic oversampling techniques.

Resampling techniques, such as oversampling and undersampling, adjust data distribution to alleviate bias by equalizing group representation. Reweighting assigns higher weights to underrepresented instances during model training, reducing bias against marginalized groups. Data augmentation generates synthetic data to bolster underrepresented groups, enhancing fairness. Fairness-aware clustering ensures equitable grouping, while the Synthetic Minority Oversampling Technique (SMOTE) generates synthetic samples to balance class distribution, promoting fairness in classification tasks. These methods collectively counteract bias and enhance fairness in AI models.

In Table 6, it can be seen that re-sampling techniques handle class imbalance, reweighting adjusts data importance, data augmentation enhances diversity, attribute swapping equalizes sensitive attributes, fairness-aware clustering ensures equitable grouping, and SMOTE addresses class imbalance. By selecting and applying the appropriate preprocessing fairness method based on the specific dataset and fairness goals, AI practitioners can develop models that prioritize fairness and equitable outcomes. Continued research and experimentation with these techniques will advance the pursuit of unbiased AI applications across various domains.

Table 6. Pre-processing fairness comparison.

Pre-Processing Fairness Method	Features	Pros	Cons	References
Re-sampling Techniques	Balance representation of different groups	Simple and easy to implement	May lead to loss of information and increased computation	[150–153]
Re-weighting Techniques	Assign higher weights to underrepresented groups	Does not alter the original dataset	Requires careful selection of appropriate weights	[154–159]
Data Augmentation	Generate synthetic data to increase representation	Increases the diversity of the training dataset	Synthetic data may not fully represent real-world samples	[160–163]
Fairness-aware Clustering	Cluster data points while maintaining fairness	Incorporates fairness constraints during clustering	May not guarantee perfect fairness in all clusters	[164–167]
Synthetic Minority Over-sampling Technique (SMOTE)	Generate synthetic samples for the minority class	Addresses class imbalance	May result in overfitting or noisy samples	[168–171]

5.3. In Processing Fairness

In-processing fairness refers to modifying the learning algorithm or objective function during model training to explicitly incorporate fairness constraints. These techniques aim to simultaneously optimize accuracy while reducing bias and promoting fairness. Their main approaches include adversarial training [172–176], adversarial debiasing [137,177–179], equalized odds post-processing [11,144,177,180], fair causal learning [45,181–184], and meta-fairness [163,185,186].

Table 7 summarizes the comparison of the methods above. Adversarial training and adversarial debiasing introduce an adversarial component to the learning process, aiming to minimize the influence of sensitive attributes on model predictions. These methods have been applied across tasks like natural language processing, computer vision, and recommendation systems to enhance fairness and reduce bias. Causal learning methods focus on understanding causal relationships within data and addressing confounding factors that lead to biased predictions. This approach has been implemented in domains such as healthcare and criminal justice to ensure fairer and more interpretable outcomes. Meta Fairness involves learning a fairness-aware optimization algorithm that dynamically adjusts the balance between fairness and accuracy during model training. It is particularly valuable when fairness requirements vary across user groups or over time.

Table 7. In-processing fairness comparison.

In-Processing Fairness Method	Features	Pros	Cons	References
Adversarial Training	Adversarial component to minimize bias impact	Enhances model's fairness while maintaining accuracy	Sensitive to adversarial attacks, requires additional computational resources	[172–176]
Adversarial Debiasing	Adversarial network to remove sensitive attributes	Simultaneously reduces bias and improves model's fairness	Adversarial training challenges, potential loss of predictive performance	[137,177–179]
Equalized Odds Post-processing	Adjust model predictions to ensure equalized odds	Guarantees fairness in binary classification tasks	May lead to suboptimal trade-offs between fairness and model performance	[11,144,177,180]
Causal Learning for Fairness	Focus on causal relationships to adjust for bias	Addresses confounding factors to achieve fairer predictions	Requires causal assumptions, may be limited by data availability	[45,181–184]
Meta Fairness	Learns fairness-aware optimization algorithm	Adapts fairness-accuracy trade-off to changing requirements	Complexity in learning the optimization algorithm, potential increased complexity	[163,185,186]

Adapting to different biases and trade-offs between fairness and performance, these methods provide valuable tools for equitable AI. Choosing the appropriate method hinges on factors such as the application context and bias type.

5.4. Post-Processing Fairness

Post-processing fairness methods focus on adjusting or post-processing the output of trained AI models to ensure fairness and reduce bias after the model has made its predictions. These techniques are applied after the model has made a decision to align the results with the fairness goal and mitigate any potential bias present in the predictions. Some common post-processing fairness methods include equalized odds post-processing [11,144,177,180], calibration post-processing [187–190], rejected options classification (ROC) post-processing [144,191–193], priority sampling post-processing [194–196], threshold optimization post-processing [197–200], and regularization post-processing [201–204]. Table 8 summarizes the features, pros and cons with the comparison of these methods.

Equalized odds post-processing is employed post-model training to align predictions, ensuring equal false alarm and omission rates across different groups. Calibration refines predicted probabilities to accurately reflect event likelihood. Reject option classification introduces a "reject" option to avoid biased predictions in sensitive situations. Preferential sampling post-processing reshapes training data distribution for enhanced fairness. Threshold optimization post-processing adjusts decision thresholds for a balanced fairness–accuracy trade-off. Regularization post-processing employs regularization techniques to encourage fairness during model optimization. These techniques offer ways to enhance fairness in AI predictions and are particularly useful in contexts like credit scoring and hiring decisions.

The effectiveness and suitability of post-processing fairness methods vary with the AI application and fairness goals. While valuable in certain contexts, these techniques might not entirely resolve root biases. A holistic AI fairness approach should combine pre-processing, in-processing, and post-processing methods, alongside continuous monitoring and evaluation, to ensure equitable outcomes.

Table 8. Post-processing fairness comparison.

Post-Processing Fairness Method	Features	Pros	Cons	References
Equalized Odds Post-processing	Adjust model predictions to ensure equalized odds	Ensures equalized false positive and true positive rates across groups	May lead to suboptimal trade-offs between fairness and model performance	[11,144,177,180]
Calibration Post-processing	Calibrates model's predicted probabilities	Improves fairness by aligning confidence scores with true likelihood	Calibration may not entirely remove bias from the model	[187–190]
Reject Option Classification (ROC) Post-processing	Introduces a "reject" option in classification decisions	Allows the model to abstain from predictions in high fairness concern cases	May lead to lower accuracy due to abstaining from predictions	[144,191–193]
Preferential Sampling Post-processing	Modifies the training data distribution by resampling instances	Improves fairness by adjusting the representation of different groups	May not address the root causes of bias in the model	[194–196]
Threshold Optimization Post-processing	Adjusts decision thresholds for fairness and accuracy trade-off	Allows fine-tuning of fairness and performance balance	May not fully eliminate all biases in the model	[197–200]
Regularization Post-processing	Applies fairness constraints during model training	Encourages fairness during the optimization process	Fairness constraints might impact model performance	[201–204]

5.5. Regularization Based Fairness

Regularization-based fairness methods are emerging as a promising approach to mitigate biases in machine learning models. Regularization techniques aim to enforce fairness constraints during the model training process, ensuring that the model's predictions are less influenced by sensitive attributes and promote equitable outcomes. Table 9 summarizes and compares different regularization methodologies for AI fairness, including adversarial regularization [205–208], demographic parity regularization [201,204,209–211], equalized odds regularization [201,212,213], covariate leveling regularization [214,215], and mixture density network regularization [216–218].

Table 9. Regularization-based fairness comparison.

Regularization-Based Fairness Method	Features	Pros	Cons	References
Adversarial Regularization	Introduces adversarial component	Encourages disentanglement of sensitive attributes	Computationally expensive	[205–208]
Demographic Parity Regularization	Enforces similar distributions across groups	Addresses group fairness	May lead to accuracy trade-offs	[201,204,209–211]
Equalized Odds Regularization	Ensures similar false/true positive rates	Emphasizes fairness in both rates	May lead to accuracy trade-offs	[201,212,213]
Covariate Shift Regularization	Reduces impact of biased/underrepresented subgroups	Addresses bias due to distributional differences	Sensitive to noise in the data	[214,215]
Mixture Density Network Regularization	Models uncertainty in predictions	Provides probabilistic approach to fairness regularization	Requires larger amount of data to estimate probability distributions	[216–218]

Regularization-based fairness methods introduce additional components to the model training process to mitigate bias in AI predictions. Adversarial regularization aims to minimize model dependence on sensitive attributes by introducing an adversarial component. Demographic parity regularization enforces similar prediction distributions across sensitive attribute groups. Equalized odds regularization maintains consistent false alarm and true alarm rates among these groups. Covariate leveling regularization adapts predictions to diverse data distributions. Mixture density network regularization models prediction uncertainty through probability density functions. Each approach offers distinct benefits and trade-offs in addressing bias.

5.6. Counterfactual Fairness

Counterfactual fairness is an approach that seeks to address bias in AI models by considering counterfactual scenarios, where sensitive attributes are altered while keeping other features fixed. The idea is to evaluate fairness by examining how the model's predictions would change if an individual belonged to a different demographic group, allowing for a more nuanced understanding of biases. Table 10 summarizes and compares different regularization methodologies for AI fairness, including individual fairness [40,168,196,219], equal opportunity fairness [220–222], reweighted counterfactual fairness [223–225], and oblivious training [226–228].

Table 10. Counterfactual fairness methods comparison.

Counterfactual Fairness Method	Features	Pros	Cons	References
Individual Fairness	Focuses on treating similar individuals similarly based on their features	Considers fairness at the individual level, promoting personalized fairness	Defining similarity metrics and enforcing individual fairness can be challenging	[40,196,219]
Equal Opportunity Fairness	Minimizes disparate impact on true positive rates across sensitive attribute groups	Targets fairness in favor of historically disadvantaged groups	May neglect other fairness concerns, such as false positive rates or overall accuracy	[220–222]
Equalized Odds Fairness	Aims for similar false positive and true positive rates across sensitive attribute groups	Addresses fairness in both false positives and false negatives	May lead to accuracy trade-offs between groups	[229–231]
Reweighted Counterfactual Fairness	Assigns different weights to instances based on similarity to counterfactual scenarios	Provides better fairness control by adjusting instance weights	Determining appropriate weights and balancing fairness and accuracy can be challenging	[223–225]
Oblivious Training	Trains the model to be ignorant of certain sensitive attributes during learning	Offers a simple and effective way to mitigate the impact of sensitive attributes	May result in lower model performance when sensitive attributes are relevant to the task	[226–228]

Individual fairness focuses on treating similar individuals equally despite their protected attributes. It promotes personalized fairness at the individual level, emphasizing fine-grained treatment. Equal opportunity fairness ensures similar true positive rates across different groups to prevent disparate impact in binary classification. Reweighted counterfactual fairness adjusts data weights during training to mitigate bias and can be combined with fairness-aware algorithms. Oblivious training trains models on both original and counterfactual data to promote fairness without explicit labels.

These methods address different fairness concerns, considering both individual and group fairness aspects, each with their computational and implementation considerations. Each fairness method has strengths and limitations, potentially impacting areas like model performance and interpretability. The method chosen should align with specific fairness

criteria and application contexts, as certain methods may be better suited for particular domains than others.

In medical data collection, informed consent methods are employed to clarify data usage and potential risks. Privacy techniques like anonymization protect individuals' identities. Data minimization reduces privacy risks by collecting only necessary information, though this may limit insights. Transparency communicates data collection processes, building trust while potentially raising privacy concerns. Data security measures include encryption and access controls to prevent unauthorized access. Accuracy and accountability methods involve auditing for reliable data and research outcomes. These approaches enhance data quality and accountability but may require resource allocation. Balancing these strategies is essential for ethical and effective data collection in scientific research.

6. Discussion

6.1. Fair Data Collection

To guarantee the fairness in data collection, we summarize the different methods with the comparison in Table 11 between informed consent, privacy and anonymity, privacy and anonymization, accuracy and accountability, data security, data minimization, and transparency approach.

Table 11. Fair data collection fairness methods comparison.

Method Category	Features	Pros	Cons	References
Informed Consent	Obtain explicit consent from participants	Respects individual autonomy	May lead to selection bias	[232–234]
Informed Consent	Clear explanation of data collection purpose	Builds trust with participants	Consent may not always be fully informed	[235,236]
Informed Consent	Informed of potential risks		Difficulties with complex research studies	[237–239]
Privacy and Anonymity	Data anonymization, aggregation, de-identification	Protects participant privacy	Reduced utility of anonymized data	[240,241]
Privacy and Anonymity	Prevents re-identification of individuals	Minimizes risk of data breaches	Challenges in preserving data utility	[242–244]
Data Minimization	Collect only necessary data	Reduces data collection and storage costs	Limited data for certain analyses	[28,245,246]
Data Minimization	Avoid gathering excessive/inappropriate data	Mitigates privacy risks	Potential loss of insights	[247,248]
Transparency	Clear communication of data collection process	Builds trust with data subjects	May lead to privacy concerns	[249–251]
Transparency	Information on methods and data use	Increases data sharing and collaboration	Difficulties in balancing transparency	[249–251]
Data Security	Encryption, access controls, security audits	Protects data from unauthorized access	Implementation costs	[252–254]
Data Security	Safeguards data from breaches	Prevents data manipulation and tampering	Potential usability impact	[252–254]
Accuracy and Accountability	Processes for data accuracy and accountability	Ensures reliability of data	Requires resource allocation for auditing	[24,255,256]

6.2. Regular Auditing and Monitoring

The continuous monitoring and auditing of AI systems are crucial to identify and address emerging biases throughout the AI lifecycle. Regular auditing and monitoring are crucial aspects of ensuring AI fairness in real-world applications. Table 12 summarizes

different methods for auditing and monitoring AI fairness, including disparate impact analysis, fairness-aware performance metrics, bias detection techniques, algorithmic fairness dashboards, model explanation and interpretability, and continual bias monitoring.

Table 12. Regular auditing and monitoring comparison.

Method	Features	Pros	Cons	References
Disparate Impact Analysis	Measures disparate impact ratios	Easy to implement and interpret	Only captures one aspect of fairness (impact ratios)	[6,257,258]
Fairness-aware Performance Metrics	Simultaneously evaluates accuracy and fairness	Provides a holistic view of model performance and fairness	Choice of fairness metric may not fully capture desired notions of fairness	[259–261]
Bias Detection Techniques	Identifies biases in data or model predictions	Alerts to potential fairness issues early	May require domain expertise for interpreting and addressing identified biases	[71,262,263]
Algorithmic Fairness Dashboards	Real-time visualizations and metrics for monitoring	Enables continuous fairness monitoring	Complexity in designing comprehensive dashboards	[264–266]
Model Explanation and Interpretability	Provides insights into decision-making	Facilitates understanding of model behavior and potential biases	May not fully capture complex interactions in the model, leading to limited interpretability	[267–270]
Continual Bias Monitoring	Ongoing and regular assessment	Detects and addresses emerging fairness issues over time	May require significant resources for continuous monitoring	[47,271,272]

7. AI Fairness in Practice

AI fairness has a large number of real-world applications in a variety of fields, where it is critical to ensure that machine learning models do not perpetuate or amplify bias and discrimination. The areas where the current research and application work are more focused are AI-based social infrastructure and management and business applications. Tables 13 and 14 summarize common applications and case studies with a comparison of the approaches and challenges, respectively, including education [273–277] health care [52,278–280] criminal justice and sentencing [88,281–283], hiring and recruiting [284–286], lending and credit decisions [287–292], online advertising [8,293–295], customer service and chatbots [296–303].

7.1. Social Administration

Artificial intelligence (AI) has become an integral part of all industries, changing the way decisions and processes are managed. In recent years, the concept of AI fairness has gained prominence, especially in the field of social management. AI systems are increasingly being used in areas such as criminal justice, healthcare, and education. Table 13 summarizes the typical applications, including issues, mechanisms, opportunities and challenges.

7.1.1. Health Care

AI fairness can also be applied to healthcare diagnosis and treatment recommendation systems to reduce bias and ensure fairness in healthcare delivery, as some healthcare AI systems have been found to differ in the diagnosis of certain diseases among different racial groups [52,278]. The use of a fairness-aware algorithm improves the performance of the model and provides a fairer diagnosis for all patients [279,280].

Table 13. AI fairness in social administration practices.

Application	Issues	Mechanism	Opportunities	Challenges
Health Care	Racial and gender biases in diagnosis and treatment. Unequal healthcare due to socioeconomic factors.	diversifying representative datasets. Personalized treatment plans based on individual characteristics.	Enhancing healthcare access and outcomes for all individuals. Reducing healthcare disparities.	Ensuring patient privacy and data security. Addressing biases in data collection and data sources.
Education	Bias in admissions and resource allocation. Unequal access to quality education.	Fair criteria for admissions and resource allocation. Personalized learning for individual needs. Identifying and assisting at-risk students.	Reducing educational disparities. Enhancing learning outcomes for all students.	ethical considerations regarding data privacy in educational settings. avoiding undue focus on standardized testing.
Criminal Justice and Sentencing	Racial Bias in predictive policing and sentencing. Unfair allocation of resources for crime prevention.	focus on rehabilitation with regular auditing and updating the models with transparency in decision-making.	Reducing biased arrests and sentencing. Allocating resources more efficiently.	The ethical implications of using AI in criminal justice. Ensuring model accountability and avoiding "tech-washing".

AI applications in healthcare face challenges of interpretability, trust, data privacy, and security. Privacy concerns hinder data sharing, while complex AI models lack interpretability, impacting trust. Transparent and responsible data management is needed for data sharing while protecting privacy. Explainable AI can enhance trust by making AI recommendations understandable [304]. Aligning with healthcare governance measures can ensure trustworthy AI use. Addressing these issues can revolutionize medical decision making, improve outcomes, and foster equitable and patient-centered healthcare.

7.1.2. Education

Artificial intelligence fairness is applied to education technology to ensure equal opportunity for students regardless of their background [273]. This is because in reality, AI-powered tutoring systems exhibit biases when assigning tasks to students [274,275]. Therefore with the incorporation of fairness awareness, the system can adjust the recommendations to treat all students fairly [276,277].

AI algorithms in college admissions and resource allocation may unintentionally perpetuate biases, impacting opportunities and diversity. Personalized learning platforms might worsen educational disparities, particularly for marginalized students. Testing and assessment bias can lead to unfair evaluations, affecting self-esteem and prospects. Future work should design fairness-conscious admissions models and AI systems optimizing fairness to mitigate bias. Transparency and accountability measures should guide AI-based educational decisions. Incorporating equity in personalized learning algorithms and diverse educational content can promote equitable support. Culturally responsive education and diverse resources can also aid in reducing bias in assessments.

7.1.3. Criminal Justice and Sentencing

Utilizing AI fairness to address bias in risk assessment tools ensures that criminal sentencing decisions are fair [88,281]. In the criminal justice system, some AI-based risk assessment tools have been found to be racially biased, resulting in harsher sentences for some minorities [282]. Implementation of fairness awareness training resulted in a significant reduction in system bias [283].

Current research predominantly addresses racial, socioeconomic, and recidivism prediction biases. Data-driven disparities may lead to biased arrests, bail decisions, and sentencing. Biased recidivism prediction algorithms misclassify groups as high risk, perpetuating unfair treatment and higher incarceration rates for marginalized groups.

Future efforts should gather diverse, representative training data to mitigate bias. Fairness-aware algorithms and AI models optimizing fairness should be developed to prevent differential treatment based on race or socioeconomic status. Transparent risk assessment models can enhance interpretability, aiding defendants and legal professionals. Regular model audits are essential to identify and rectify potential biases in model deployment.

7.2. Business

Another important application area for AI Fairness is business. More and more AI technologies are also being continuously introduced into commercial applications, and Table 14 summarizes and compares several trending widely used scenarios.

Table 14. AI fairness in business practices.

Application	Issues	Mechanism	Opportunities	Challenges
Recruiting	Bias in job ads and candidate selection. Lack of diversity in hiring.	Debiasing job descriptions, candidate screening and removing identifiable information, diversifying training data.	Increasing workforce diversity. Reducing hiring discrimination.	Balancing fairness and competence. Ensuring fairness across different demographics.
Lending and Credit Decisions	Discrimination in loan approvals. Lack of transparency in decision making.	Implementing fairness-aware algorithms, explaining model decisions, alternative data to creditworthiness.	Expanding access to credit for marginalized groups. Improving overall lending practices.	Striking a balance between fairness and risk assessment. Handling potential adversarial attacks on models.
Online Advertising	Targeting ads based on sensitive attributes. Reinforcing stereotypes through ad delivery.	Differential privacy to protect privacy, biased message screening, providing users preference controls.	Improving user experience and privacy protection. Fostering a positive brand image.	The balance between targeted ads and user privacy. Identifying and Addressing hidden biases in ad delivery.
Customer Service and Chatbots	biased responses and inappropriate interactions. Lack of understanding diverse linguistic expressions.	Training chatbots on inclusive and diverse datasets with reinforcement learning to improve interactions with feedback on bot behavior.	Enhancing user experience and customer satisfaction. Scaling customer support efficiently.	Minimizing harmful or offensive responses. Dealing with novel inputs and out-of-distribution data.

7.2.1. Hiring and Recruiting

The integration of artificial intelligence (AI) in human resource management (HRM) introduces transformative enhancements. Figure 3 shows a recruitment process supported by AI throughout. AI-driven algorithms streamline CV screening and candidate profiling, while proctored assessments ensure secure remote testing. AI optimizes interview scheduling and personalizes HR training. Behavior tracking and personality analysis provide insights into candidate dynamics, and AI aids in appraisal monitoring through performance metric analysis. These applications collectively reshape HRM practices, enhancing efficiency and informed decision making.

In the applications of AI system in HRM, artificial intelligence fairness is applied to mitigate bias in automated hiring systems, ensuring equitable and non-discriminatory candidate selection [305,306]. Many AI-driven recruitment tools have exhibited biases, favoring specific candidates and overlooking job requisites [307,308]. Utilizing AI fairness techniques rectifies these model biases, fostering impartial hiring decisions independent of attributes unrelated to job proficiency [309]. Tackling these challenges and future endeavors in AI fairness is pivotal to harnessing the potential of AI for equitable recruitment, fostering diversity and inclusivity in workforce dynamics.

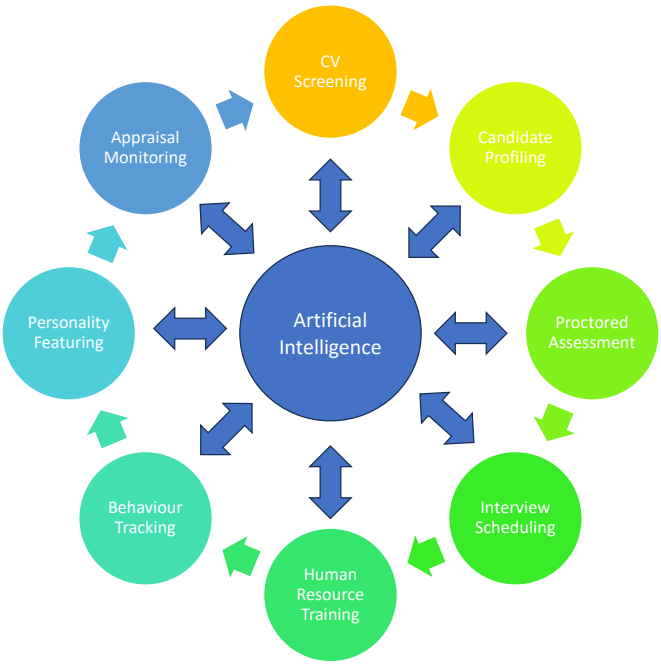

Figure 3. Tasks with artificial intelligence for hiring and recruiting in human resources.

7.2.2. Loan and Credit Decisions

AI applications in loan and credit decision making aim to improve decision accuracy, speed, and fairness while maintaining prudent risk management. As shown in Figure 4, AI applications in loan and credit decision making involve leveraging artificial intelligence techniques to assess creditworthiness, streamline lending processes, and enhance risk management. These applications use AI algorithms to analyze various data sources, such as financial records, transaction histories, and alternative data, to make more accurate and efficient lending decisions. This aids in automating and optimizing the loan approval process, reducing human bias, and increasing access to credit for underserved populations. AI assists in fraud detection, predicting default risks, and personalizing loan terms based on individual borrower profiles.

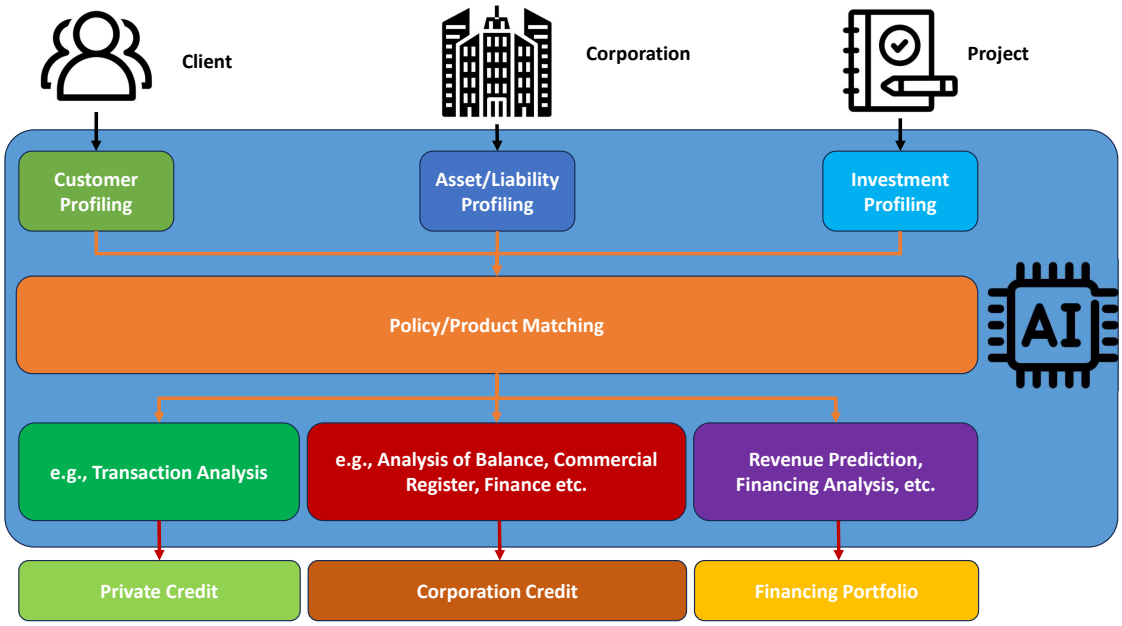

Figure 4. Tasks with artificial intelligence for hiring and recruiting in human resources.

7.2.3. Online Advertising

To counteract bias in credit-scoring models and ensure equitable access to loans and credit opportunities for all individuals, AI fairness is applied with bias migration strategies [287,288]. Certain AI-driven credit scoring models have exhibited potential bias towards specific demographic groups [288,289]. Implementing bias mitigation techniques enhances model fairness, leading to more impartial lending determinations [290–292]. Looking ahead, future efforts should focus on integrating diverse data sources like rental histories or utility payments into credit assessments while maintaining fairness. Designing AI models capable of adapting to various data distributions, including non-traditional data, can also sustain fairness and accuracy. Incorporating fairness-aware explanations into AI models offers insight into achieving equitable credit decisions with transparency. Through these applications and ongoing research, the aim is to foster inclusive and just lending practices by minimizing bias and promoting unbiased access to credit [287,288,290–292].

AI systems have a substantial role in online advertising, including targeted ad delivery, content censorship and related design. Figure 5 shows an example of AI applications according to the hierarchical taxonomy of online advertising. It can be seen that the AI support can enhance ad relevance and user experience, while also mitigating inappropriate content. However, challenges related to biases in ad targeting and content moderation necessitate the development of fairness-aware approaches to ensure equitable outcomes. In this context, AI technologies both facilitate and necessitate ongoing efforts to maintain fairness and effectiveness in online advertising practices.

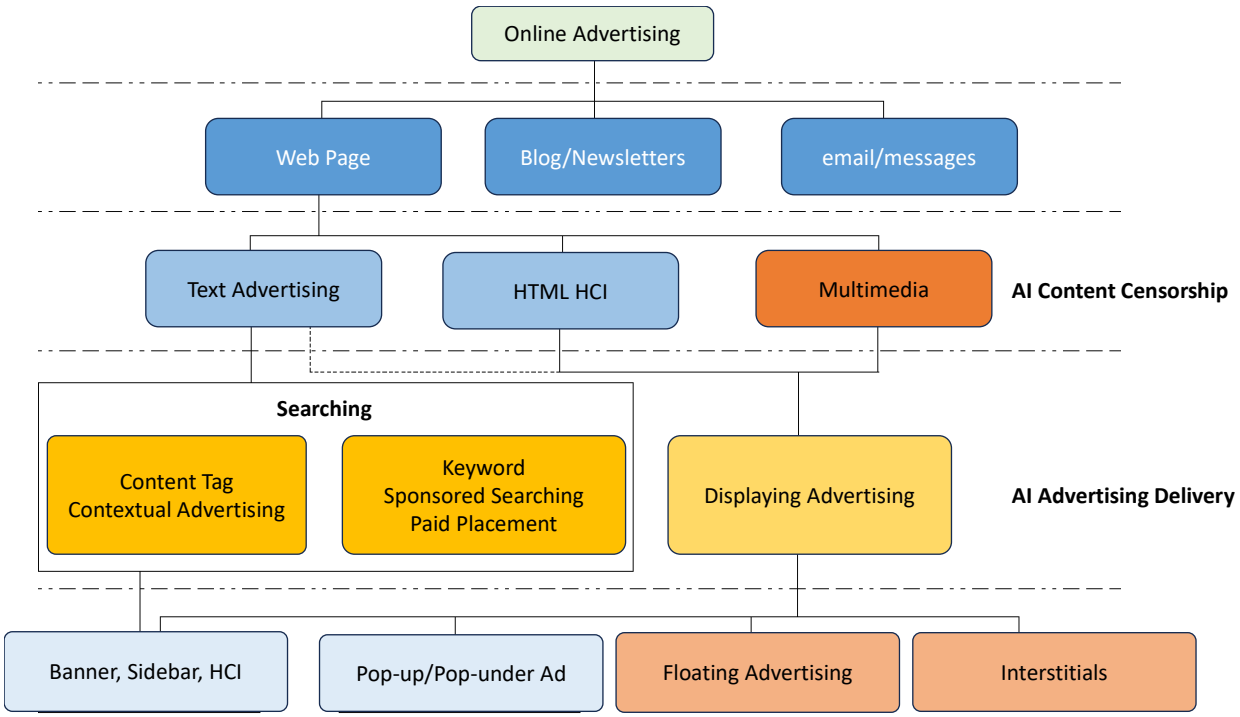

Figure 5. An example of AI support corresponding to the taxonomy of online advertising.

To address the issue above, AI fairness is used for ad targeting to avoid promoting discriminatory or biased content to users based on their attributes [293]. Some online advertising platforms have experienced problems with certain ads being disproportionately shown to users from specific demographic groups [8,294]. By incorporating fairness constraints, the platform achieved more balanced ad targeting across all users [295]. AI algorithms in advertising may unintentionally yield biased ad targeting and content due to biased training data and content generation. Future work should center on fairness-aware targeting, bias audits, diverse data, and inclusive content to ensure fairness and inclusivity in ad delivery.

7.2.4. Customer Service

Artificial intelligence is also introduced into customer service, or the customer relationship management (CRM) system [296], which enhances customer interactions and support. As shown in Figure 6, the chatbots utilize natural language processing and sentiment analysis to understand customer queries and provide accurate, timely responses. They enable automated, efficient, and personalized customer interactions, improving user experience. AI-driven chatbots handle routine inquiries, offer real-time support, and gather insights for businesses to enhance their services. This technology aims to streamline customer service operations while ensuring effective and satisfactory customer interactions.

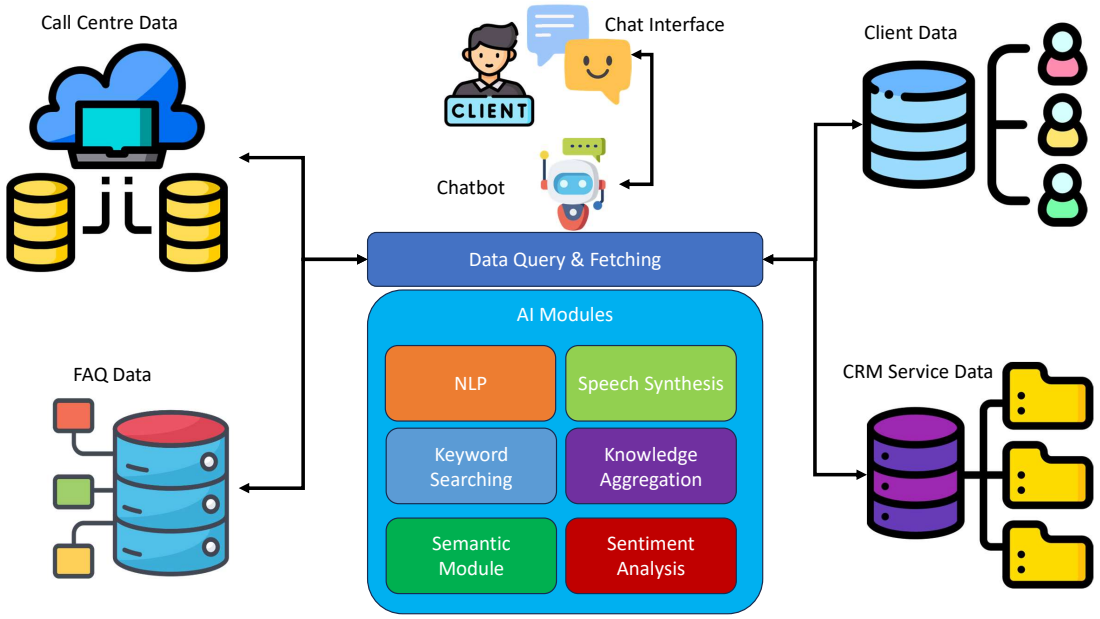

Figure 6. An example of AI support chatbot system for CRM.

However, some customer service chatbots show biased responses to users who speak certain language dialects [297–299]. Aiming at this, AI fairness is used in chatbot design to avoid biased responses or inappropriate behavior towards users [300,301]. After implementing fairness checks, the chatbot provides culturally sensitive and fair interactions [302,303].

The current challenge in the customer service bot domain pertains to mitigating uncertain biased and inappropriate responses. Chatbots often unknowingly offer biased or offensive replies due to training data exposure, leading to customer dissatisfaction and reputational harm. Additionally, limitations in comprehending diverse linguistic expressions hinder accurate responses to various language forms, including slang. Further, inadequacies in addressing sensitive topics and emotional responses lead to inappropriate interactions in some customer service bots.

In prospective research, countering the aforementioned concerns requires embedding bias detection and mitigation mechanisms to identify and address biased language and responses in chatbot interactions. Mitigating biased replies can be achieved by adopting inclusive training data representing diverse user demographics and employing natural language processing techniques to enhance language comprehension. Continuous learning is essential for customer service bots to adapt and comprehend various language styles through user interactions.

8. Conclusions

This article introduces the study of fairness in artificial intelligence, detailing the background and definition of this concept. The article introduces the development process of the fairness problem in AI systems from the perspectives of practical applications and the current state of development, and reviews and discusses the main research directions

for solving the fairness problem in AI—bias analysis and fairness training—respectively. In the course of the review, the ideas and implementations of each method are explained in detail, and their respective characteristics and occasions of use are compared. The article also explores measures to reduce bias and improve fairness in AI systems, reviews relevant problems and solutions in practical applications of AI fairness, and discusses possible future research directions. On the basis of the theoretical foundations and methodology of AI fairness, the paper also explores scenarios and application examples in practical applications, thus contributing to the current discussion on fair and unbiased AI systems.

This paper also provides an in-depth comparison of the characteristics, advantages, and disadvantages of each of the different methods, based on the collation of the state-of-the-art research. The results of the comparison will provide advisory support for future research and development. At the same time, this paper also summarizes some of the existing problems in existing applications and proposes some focuses and solution ideas for future research work. These summaries will provide ideas for the further development of fairness in future AI systems.

The information synthesized in this study comes from a variety of reliable sources. These sources include recent publications in prestigious academic journals, distinguished conference proceedings, and well-established online repositories dedicated to the fairness of AI. It is important to recognize that while these sources have contributed significantly to our understanding, there are limitations to the evidence they provide. These limitations stem primarily from nuances in the research methodology, sample size, and context, which may affect the generalizability of conclusions drawn from individual studies. The landscape of AI fairness is dynamic, with research and advancements continually shaping our understanding of its complexities. While our current coverage might have limitations due to the rapid pace of change and ongoing research, please know that we are committed to further studying and exploring this crucial subject.

Author Contributions: The idea of this paper came from P.C., who also arranged the division of labor for this paper. Our work focuses are as follows: P.C. is responsible for introducing the background and definition of AI fairness; reviewing the main research directions of addressing AI fairness, including the related methods of bias analysis and fair training. L.W. (Lei Wang) is responsible for explaining the fairness problem and evaluation methods in AI systems, as well as the measures to reduce bias and improve fairness. L.W. (Linna Wu) is responsible for evaluating the related issues and solutions in practical applications. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Angerschmid, A.; Zhou, J.; Theuermann, K.; Chen, F.; Holzinger, A. Fairness and explanation in AI-informed decision making. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 556–579. [CrossRef]
2. Kratsch, W.; Manderscheid, J.; Röglinger, M.; Seyfried, J. Machine learning in business process monitoring: A comparison of deep learning and classical approaches used for outcome prediction. *Bus. Inf. Syst. Eng.* **2021**, *63*, 261–276. [CrossRef]
3. Kraus, M.; Feuerriegel, S.; Oztekin, A. Deep learning in business analytics and operations research: Models, applications and managerial implications. *Eur. J. Oper. Res.* **2020**, *281*, 628–641. [CrossRef]
4. Varona, D.; Suárez, J.L. Discrimination, bias, fairness, and trustworthy AI. *Appl. Sci.* **2022**, *12*, 5826. [CrossRef]
5. Saghiri, A.M.; Vahidipour, S.M.; Jabbarpour, M.R.; Sookhak, M.; Forestiero, A. A survey of Artificial Intelligence challenges: Analyzing the definitions, relationships, and evolutions. *Appl. Sci.* **2022**, *12*, 4054. [CrossRef]
6. Barocas, S.; Selbst, A.D. Big data's disparate impact. *Calif. Law Rev.* **2016**, *104*, 671–732. [CrossRef]
7. Corsello, A.; Santangelo, A. May Artificial Intelligence Influence Future Pediatric Research?—The Case of ChatGPT. *Children* **2023**, *10*, 757. [CrossRef] [PubMed]
8. Von Zahn, M.; Feuerriegel, S.; Kuehl, N. The cost of fairness in AI: Evidence from e-commerce. *Bus. Inf. Syst. Eng.* **2021**, *64*, 335–348. [CrossRef]
9. Liu, L.T.; Dean, S.; Rolf, E.; Simchowitz, M.; Hardt, M. Delayed impact of fair machine learning. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 3150–3158.
10. Cathy, O. *How Big Data Increases Inequality and Threatens Democracy*; Crown Publishing Group: New York, NY, USA, 2016.
11. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 3323–3331.

12. Trewin, S. AI fairness for people with disabilities: Point of view. *arXiv* **2018**, arXiv:1811.10670.
13. Kodiyan, A.A. An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool. *Researchgate Prepr.* **2019**, 1–19.
14. Righetti, L.; Madhavan, R.; Chatila, R. Unintended consequences of biased robotic and Artificial Intelligence systems [ethical, legal, and societal issues]. *IEEE Robot. Autom. Mag.* **2019**, *26*, 11–13. [CrossRef]
15. Garg, P.; Villasenor, J.; Foggo, V. Fairness metrics: A comparative analysis. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 3662–3666.
16. Mehrotra, A.; Sachs, J.; Celis, L.E. Revisiting Group Fairness Metrics: The Effect of Networks. *Proc. Acm Hum. Comput. Interact.* **2022**, *6*, 1–29. [CrossRef]
17. Ezzeldin, Y.H.; Yan, S.; He, C.; Ferrara, E.; Avestimehr, A.S. Fairfed: Enabling group fairness in federated learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 7494–7502.
18. Hooker, S. Moving beyond "algorithmic bias is a data problem". *Patterns* **2021**, *2*, 100241. [CrossRef]
19. Amini, A.; Soleimany, A.P.; Schwarting, W.; Bhatia, S.N.; Rus, D. Uncovering and mitigating algorithmic bias through learned latent structure. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019; pp. 289–295.
20. Yang, J.; Soltan, A.A.; Eyre, D.W.; Yang, Y.; Clifton, D.A. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digit. Med.* **2023**, *6*, 55. [CrossRef]
21. Li, S. Towards Trustworthy Representation Learning. In Proceedings of the 2023 SIAM International Conference on Data Mining (SDM), Minneapolis, MN, USA, 27–29 April 2023; SIAM: Philadelphia, PA, USA, 2023; pp. 957–960.
22. Creager, E.; Madras, D.; Jacobsen, J.H.; Weis, M.; Swersky, K.; Pitassi, T.; Zemel, R. Flexibly fair representation learning by disentanglement. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 10–15 June 2019; pp. 1436–1445.
23. McNamara, D.; Ong, C.S.; Williamson, R.C. Costs and benefits of fair representation learning. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019; pp. 263–270.
24. Sahlgren, O. The politics and reciprocal (re) configuration of accountability and fairness in data-driven education. *Learn. Media Technol.* **2023**, *48*, 95–108. [CrossRef]
25. Ravishankar, P.; Mo, Q.; McFowland III, E.; Neill, D.B. Provable Detection of Propagating Sampling Bias in Prediction Models. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 9562–9569. [CrossRef]
26. Park, J.; Ellezhuthil, R.D.; Isaac, J.; Mergerson, C.; Feldman, L.; Singh, V. Misinformation Detection Algorithms and Fairness across Political Ideologies: The Impact of Article Level Labeling. In Proceedings of the 15th ACM Web Science Conference 2023, Austin, TX, USA, 30 April–1 May 2023; pp. 107–116.
27. Friedrich, J. Primary error detection and minimization (PEDMIN) strategies in social cognition: A reinterpretation of confirmation bias phenomena. *Psychol. Rev.* **1993**, *100*, 298. [CrossRef] [PubMed]
28. Frincke, D.; Tobin, D.; McConnell, J.; Marconi, J.; Polla, D. A framework for cooperative intrusion detection. In Proceedings of the 21st NIST-NCSC National Information Systems Security Conference, Arlington, VA, USA, 2–8 April 1998; pp. 361–373.
29. Estivill-Castro, V.; Brankovic, L. Data swapping: Balancing privacy against precision in mining for logic rules. In *International Conference on Data Warehousing and Knowledge Discovery*; Springer: Berlin/Heidelberg, Germany, 1999; pp. 389–398.
30. Bellamy, R.K.; Dey, K.; Hind, M.; Hoffman, S.C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv* **2018**, arXiv:1810.01943.
31. Zhang, Y.; Bellamy, R.K.; Singh, M.; Liao, Q.V. Introduction to AI fairness. In Proceedings of the Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2020; pp. 1–4.
32. Mahoney, T.; Varshney, K.; Hind, M. *AI Fairness*; O'Reilly Media Incorporated: Sebastopol, CA, USA, 2020.
33. Mosteiro, P.; Kuiper, J.; Masthoff, J.; Scheepers, F.; Spruit, M. Bias discovery in machine learning models for mental health. *Information* **2022**, *13*, 237. [CrossRef]
34. Wing, J.M. Trustworthy AI. *Commun. ACM* **2021**, *64*, 64–71. [CrossRef]
35. Percy, C.; Dragicevic, S.; Sarkar, S.; d'Avila Garcez, A. Accountability in AI: From principles to industry-specific accreditation. *AI Commun.* **2021**, *34*, 181–196. [CrossRef]
36. Benjamins, R.; Barbado, A.; Sierra, D. Responsible AI by design in practice. *arXiv* **2019**, arXiv:1909.12838.
37. Dignum, V. The myth of complete AI-fairness. In Proceedings of the Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual, 15–18 June 2021; Springer: Cham, Switzerland, 2021; pp. 3–8.
38. Silberg, J.; Manyika, J. *Notes from the AI Frontier: Tackling Bias in AI (and in Humans)*; McKinsey Global Institute: San Francisco, CA, USA, 2019; Volume 1.
39. Bird, S.; Kenthapadi, K.; Kiciman, E.; Mitchell, M. Fairness-aware machine learning: Practical challenges and lessons learned. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Melbourne, Australia, 11–15 February 2019; pp. 834–835.
40. Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; Zemel, R. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Cambridge, MA, USA, 8–10 January 2012; pp. 214–226.

41. Islam, R.; Keya, K.N.; Pan, S.; Sarwate, A.D.; Foulds, J.R. Differential Fairness: An Intersectional Framework for Fair AI. *Entropy* **2023**, *25*, 660. [CrossRef] [PubMed]
42. Barocas, S.; Hardt, M.; Narayanan, A. Fairness in machine learning. *Nips Tutor.* **2017**, *1*, 2017.
43. Zafar, M.B.; Valera, I.; Rogriguez, M.G.; Gummadi, K.P. Fairness constraints: Mechanisms for fair classification. In Proceedings of the Artificial Intelligence and Statistics, PMLR, Ft. Lauderdale, FL, USA, 20–22 April 2017; pp. 962–970.
44. Cornacchia, G.; Anelli, V.W.; Biancofiore, G.M.; Narducci, F.; Pomo, C.; Ragone, A.; Di Sciascio, E. Auditing fairness under unawareness through counterfactual reasoning. *Inf. Process. Manag.* **2023**, *60*, 103224. [CrossRef]
45. Kusner, M.J.; Loftus, J.; Russell, C.; Silva, R. Counterfactual fairness. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
46. Feldman, M.; Friedler, S.A.; Moeller, J.; Scheidegger, C.; Venkatasubramanian, S. Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 11–14 August 2015; pp. 259–268.
47. Kearns, M.; Neel, S.; Roth, A.; Wu, Z.S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 2564–2572.
48. Fleisher, W. What's fair about individual fairness? In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 19–21 May 2021; pp. 480–490.
49. Mukherjee, D.; Yurochkin, M.; Banerjee, M.; Sun, Y. Two simple ways to learn individual fairness metrics from data. In Proceedings of the International Conference on Machine Learning, PMLR, Copenhagen, Denmark, 16–19 December 2020; pp. 7097–7107.
50. Dwork, C.; Ilvento, C. Group fairness under composition. In Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT* 2018), New York, NY, USA, 23–24 February 2018; Volume 3.
51. Binns, R. On the apparent conflict between individual and group fairness. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 514–524.
52. Chen, R.J.; Wang, J.J.; Williamson, D.F.; Chen, T.Y.; Lipkova, J.; Lu, M.Y.; Sahai, S.; Mahmood, F. Algorithmic fairness in Artificial Intelligence for medicine and healthcare. *Nat. Biomed. Eng.* **2023**, *7*, 719–742. [CrossRef]
53. Sloan, R.H.; Warner, R. Beyond bias: Artificial Intelligence and social justice. *Va. Law Technol.* **2020**, *24*, 1. [CrossRef]
54. Feuerriegel, S.; Dolata, M.; Schwabe, G. Fair AI: Challenges and opportunities. *Bus. Inf. Syst. Eng.* **2020**, *62*, 379–384. [CrossRef]
55. Bing, L.; Pettit, B.; Slavinski, I. Incomparable punishments: How economic inequality contributes to the disparate impact of legal fines and fees. *RSF Russell Sage Found. J. Soc. Sci.* **2022**, *8*, 118–136. [CrossRef] [PubMed]
56. Wang, L.; Zhu, H. How are ML-Based Online Content Moderation Systems Actually Used? Studying Community Size, Local Activity, and Disparate Treatment. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; pp. 824–838.
57. Tom, D.; Computing, D. *Eliminating Disparate Treatment in Modeling Default of Credit Card Clients*; Technical Report; Center for Open Science: Charlottesville, VA, USA, 2023.
58. Shui, C.; Xu, G.; Chen, Q.; Li, J.; Ling, C.X.; Arbel, T.; Wang, B.; Gagné, C. On learning fairness and accuracy on multiple subgroups. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 34121–34135.
59. Mayernik, M.S. Open data: Accountability and transparency. *Big Data Soc.* **2017**, *4*, 2053951717718853. [CrossRef]
60. Zhou, N.; Zhang, Z.; Nair, V.N.; Singhal, H.; Chen, J.; Sudjianto, A. Bias, Fairness, and Accountability with AI and ML Algorithms. *arXiv* **2021**, arXiv:2105.06558.
61. Shin, D. User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability. *J. Broadcast. Electron. Media* **2020**, *64*, 541–565. [CrossRef]
62. Sokol, K.; Hepburn, A.; Poyiadzi, R.; Clifford, M.; Santos-Rodriguez, R.; Flach, P. Fat forensics: A python toolbox for implementing and deploying fairness, accountability and transparency algorithms in predictive systems. *arXiv* **2022**, arXiv:2209.03805.
63. Gevaert, C.M.; Carman, M.; Rosman, B.; Georgiadou, Y.; Soden, R. Fairness and accountability of AI in disaster risk management: Opportunities and challenges. *Patterns* **2021**, *2*, 100363. [CrossRef]
64. Morris, M.R. AI and accessibility. *Commun. ACM* **2020**, *63*, 35–37. [CrossRef]
65. Israni, S.T.; Matheny, M.E.; Matlow, R.; Whicher, D. Equity, inclusivity, and innovative digital technologies to improve adolescent and young adult health. *J. Adolesc. Health* **2020**, *67*, S4–S6. [CrossRef]
66. Ntoutsi, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, V.; Nejdl, W.; Vidal, M.E.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; et al. Bias in data-driven Artificial Intelligence systems—An introductory survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1356. [CrossRef]
67. Baeza-Yates, R. Bias on the web. *Commun. ACM* **2018**, *61*, 54–61. [CrossRef]
68. Pessach, D.; Shmueli, E. Improving fairness of Artificial Intelligence algorithms in Privileged-Group Selection Bias data settings. *Expert Syst. Appl.* **2021**, *185*, 115667. [CrossRef]
69. Wang, Y.; Singh, L. Analyzing the impact of missing values and selection bias on fairness. *Int. J. Data Sci. Anal.* **2021**, *12*, 101–119. [CrossRef]
70. Russell, G.; Mandy, W.; Elliott, D.; White, R.; Pittwood, T.; Ford, T. Selection bias on intellectual ability in autism research: A cross-sectional review and meta-analysis. *Mol. Autism* **2019**, *10*, 1–10. [CrossRef]
71. Bolukbasi, T.; Chang, K.W.; Zou, J.Y.; Saligrama, V.; Kalai, A.T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Adv. Neural Inf. Process. Syst.* **2016**, *29*.

72. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–35. [CrossRef]
73. Torralba, A.; Efros, A.A. Unbiased look at dataset bias. In Proceedings of the CVPR, Colorado Springs, CO, USA, 20–25 June 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 1521–1528.
74. Liao, Y.; Naghizadeh, P. The impacts of labeling biases on fairness criteria. In Proceedings of the 10th International Conference on Learning Representations, ICLR, Virtually, 25–29 April 2022; pp. 25–29.
75. Paulus, J.K.; Kent, D.M. Predictably unequal: Understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ Digit. Med.* **2020**, *3*, 99. [CrossRef]
76. Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; Chang, K.W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv* **2017**, arXiv:1707.09457.
77. Yang, N.; Yuan, D.; Liu, C.Z.; Deng, Y.; Bao, W. FedIL: Federated Incremental Learning from Decentralized Unlabeled Data with Convergence Analysis. *arXiv* **2023**, arXiv:2302.11823.
78. Tripathi, S.; Musiolik, T.H. Fairness and ethics in Artificial Intelligence-based medical imaging. In *Research Anthology on Improving Medical Imaging Techniques for Analysis and Intervention*; IGI Global: Hershey, PA, USA, 2023; pp. 79–90.
79. Mashhadi, A.; Kyllo, A.; Parizi, R.M. Fairness in Federated Learning for Spatial-Temporal Applications. *arXiv* **2022**, arXiv:2201.06598.
80. Zhao, D.; Yu, G.; Xu, P.; Luo, M. Equivalence between dropout and data augmentation: A mathematical check. *Neural Netw.* **2019**, *115*, 82–89. [CrossRef] [PubMed]
81. Chun, J.S.; Brockner, J.; De Cremer, D. How temporal and social comparisons in performance evaluation affect fairness perceptions. *Organ. Behav. Hum. Decis. Process.* **2018**, *145*, 1–15. [CrossRef]
82. Asiedu, M.N.; Dieng, A.; Oppong, A.; Nagawa, M.; Koyejo, S.; Heller, K. Globalizing Fairness Attributes in Machine Learning: A Case Study on Health in Africa. *arXiv* **2023**, arXiv:2304.02190.
83. Hutiri, W.T.; Ding, A.Y. Bias in automated speaker recognition. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; pp. 230–247.
84. Makhlouf, K.; Zhioua, S.; Palamidessi, C. Machine learning fairness notions: Bridging the gap with real-world applications. *Inf. Process. Manag.* **2021**, *58*, 102642. [CrossRef]
85. Kallus, N.; Zhou, A. Residual unfairness in fair machine learning from prejudiced data. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 2439–2448.
86. Yang, N.; Yuan, D.; Zhang, Y.; Deng, Y.; Bao, W. Asynchronous Semi-Supervised Federated Learning with Provable Convergence in Edge Computing. *IEEE Netw.* **2022**, *36*, 136–143. [CrossRef]
87. So, W.; Lohia, P.; Pimplikar, R.; Hosoi, A.; D'Ignazio, C. Beyond Fairness: Reparative Algorithms to Address Historical Injustices of Housing Discrimination in the US. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; pp. 988–1004.
88. Alikhademi, K.; Drobina, E.; Prioleau, D.; Richardson, B.; Purves, D.; Gilbert, J.E. A review of predictive policing from the perspective of fairness. *Artif. Intell. Law* **2022**, *30*, 1–17. [CrossRef]
89. Rajkomar, A.; Hardt, M.; Howell, M.D.; Corrado, G.; Chin, M.H. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.* **2018**, *169*, 866–872. [CrossRef]
90. Woo, S.E.; LeBreton, J.M.; Keith, M.G.; Tay, L. Bias, fairness, and validity in graduate-school admissions: A psychometric perspective. *Perspect. Psychol. Sci.* **2023**, *18*, 3–31. [CrossRef]
91. Weerts, H.; Pfisterer, F.; Feurer, M.; Eggensperger, K.; Bergman, E.; Awad, N.; Vanschoren, J.; Pechenizkiy, M.; Bischl, B.; Hutter, F. Can Fairness be Automated? Guidelines and Opportunities for Fairness-aware AutoML. *arXiv* **2023**, arXiv:2303.08485.
92. Hauer, K.E.; Park, Y.S.; Bullock, J.L.; Tekian, A. "My Assessments Are Biased!" Measurement and Sociocultural Approaches to Achieve Fairness in Assessment in Medical Education. *Acad. Med. J. Assoc. Am. Med. Coll.* **2023**, *online ahead of print*.
93. Chen, Y.; Mahoney, C.; Grasso, I.; Wali, E.; Matthews, A.; Middleton, T.; Njie, M.; Matthews, J. Gender bias and under-representation in natural language processing across human languages. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 19–21 May 2021; pp. 24–34.
94. Chai, J.; Wang, X. Fairness with adaptive weights. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 2853–2866.
95. Zhou, Q.; Mareček, J.; Shorten, R. Fairness in Forecasting of Observations of Linear Dynamical Systems. *J. Artif. Intell. Res.* **2023**, *76*, 1247–1280. [CrossRef]
96. Spinelli, I.; Scardapane, S.; Hussain, A.; Uncini, A. Fairdrop: Biased edge dropout for enhancing fairness in graph representation learning. *IEEE Trans. Artif. Intell.* **2021**, *3*, 344–354. [CrossRef]
97. Yu, C.; Liao, W. Professionalism and homophily bias: A study of Airbnb stay choice and review positivity. *Int. J. Hosp. Manag.* **2023**, *110*, 103433. [CrossRef]
98. Lerchenmueller, M.; Hoisl, K.; Schmallenbach, L. Homophily, biased attention, and the gender gap in science. In *Academy of Management Proceedings*; Academy of Management Briarcliff Manor: New York, NY, USA, 2019; Volume 2019, p. 14784.
99. Vogrin, M.; Wood, G.; Schmickl, T. Confirmation Bias as a Mechanism to Focus Attention Enhances Signal Detection. *J. Artif. Soc. Soc. Simul.* **2023**, *26*, 2. [CrossRef]

100. Kulkarni, A.; Shivananda, A.; Manure, A. Actions, Biases, and Human-in-the-Loop. In *Introduction to Prescriptive AI: A Primer for Decision Intelligence Solutioning with Python*; Springer: Berkeley, CA, USA, 2023; pp. 125–142.
101. Gwebu, K.L.; Wang, J.; Zifla, E. Can warnings curb the spread of fake news? The interplay between warning, trust and confirmation bias. *Behav. Inf. Technol.* **2022**, *41*, 3552–3573. [CrossRef]
102. Miller, A.C. Confronting confirmation bias: Giving truth a fighting chance in the information age. *Soc. Educ.* **2016**, *80*, 276–279.
103. Ghazimatin, A.; Kleindessner, M.; Russell, C.; Abedjan, Z.; Golebiowski, J. Measuring fairness of rankings under noisy sensitive information. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; pp. 2263–2279.
104. Warner, R.; Sloan, R.H. Making Artificial Intelligence transparent: Fairness and the problem of proxy variables. *Crim. Justice Ethics* **2021**, *40*, 23–39. [CrossRef]
105. Mazilu, L.; Paton, N.W.; Konstantinou, N.; Fernandes, A.A. Fairness in data wrangling. In Proceedings of the 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI), Las Vegas, NV, USA, 11–13 August 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 341–348.
106. Caliskan, A.; Bryson, J.J.; Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **2017**, *356*, 183–186. [CrossRef] [PubMed]
107. Helms, J.E. Fairness is not validity or cultural bias in racial-group assessment: A quantitative perspective. *Am. Psychol.* **2006**, *61*, 845. [CrossRef]
108. Danks, D.; London, A.J. Algorithmic Bias in Autonomous Systems. *Ijcai* **2017**, *17*, 4691–4697.
109. Kordzadeh, N.; Ghasemaghaei, M. Algorithmic bias: Review, synthesis, and future research directions. *Eur. J. Inf. Syst.* **2022**, *31*, 388–409. [CrossRef]
110. Shen, X.; Plested, J.; Caldwell, S.; Gedeon, T. Exploring biases and prejudice of facial synthesis via semantic latent space. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–8.
111. Garcia, M. Racist in the Machine. *World Policy J.* **2016**, *33*, 111–117. [CrossRef]
112. Heffernan, T. Sexism, racism, prejudice, and bias: A literature review and synthesis of research surrounding student evaluations of courses and teaching. *Assess. Eval. High. Educ.* **2022**, *47*, 144–154. [CrossRef]
113. Prabhu, A.; Dognin, C.; Singh, M. Sampling bias in deep active classification: An empirical study. *arXiv* **2019**, arXiv:1909.09389.
114. Cortes, C.; Mohri, M. Domain adaptation and sample bias correction theory and algorithm for regression. *Theor. Comput. Sci.* **2014**, *519*, 103–126. [CrossRef]
115. Griffith, G.J.; Morris, T.T.; Tudball, M.J.; Herbert, A.; Mancano, G.; Pike, L.; Sharp, G.C.; Sterne, J.; Palmer, T.M.; Davey Smith, G.; et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat. Commun.* **2020**, *11*, 5749. [CrossRef]
116. Kleinberg, J.; Mullainathan, S.; Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv* **2016**, arXiv:1609.05807.
117. Mansoury, M.; Abdollahpouri, H.; Pechenizkiy, M.; Mobasher, B.; Burke, R. Feedback loop and bias amplification in recommender systems. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual, 19–23 October 2020; pp. 2145–2148.
118. Pan, W.; Cui, S.; Wen, H.; Chen, K.; Zhang, C.; Wang, F. Correcting the user feedback-loop bias for recommendation systems. *arXiv* **2021**, arXiv:2109.06037.
119. Taori, R.; Hashimoto, T. Data feedback loops: Model-driven amplification of dataset biases. In Proceedings of the International Conference on Machine Learning, PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 33883–33920.
120. Vokinger, K.N.; Feuerriegel, S.; Kesselheim, A.S. Mitigating bias in machine learning for medicine. *Commun. Med.* **2021**, *1*, 25. [CrossRef]
121. Kuhlman, C.; Jackson, L.; Chunara, R. No computation without representation: Avoiding data and algorithm biases through diversity. *arXiv* **2020**, arXiv:2002.11836.
122. Raub, M. Bots, bias and big data: Artificial Intelligence, algorithmic bias and disparate impact liability in hiring practices. *Ark. L. Rev.* **2018**, *71*, 529.
123. Norori, N.; Hu, Q.; Aellen, F.M.; Faraci, F.D.; Tzovara, A. Addressing bias in big data and AI for health care: A call for open science. *Patterns* **2021**, *2*, 100347. [CrossRef]
124. Kafai, Y.; Proctor, C.; Lui, D. From theory bias to theory dialogue: Embracing cognitive, situated, and critical framings of computational thinking in K-12 CS education. *ACM Inroads* **2020**, *11*, 44–53. [CrossRef]
125. Celi, L.A.; Cellini, J.; Charpignon, M.L.; Dee, E.C.; Dernoncourt, F.; Eber, R.; Mitchell, W.G.; Moukheiber, L.; Schirmer, J.; Situ, J.; et al. Sources of bias in Artificial Intelligence that perpetuate healthcare disparities—A global review. *PLoS Digit. Health* **2022**, *1*, e0000022. [CrossRef] [PubMed]
126. Schemmer, M.; Kühl, N.; Benz, C.; Satzger, G. On the influence of explainable AI on automation bias. *arXiv* **2022**, arXiv:2204.08859.
127. Alon-Barkat, S.; Busuioc, M. Human–AI interactions in public sector decision making: "Automation bias" and "selective adherence" to algorithmic advice. *J. Public Adm. Res. Theory* **2023**, *33*, 153–169. [CrossRef]
128. Jones-Jang, S.M.; Park, Y.J. How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *J. Comput. Mediat. Commun.* **2023**, *28*, zmac029. [CrossRef]

129. Strauß, S. Deep automation bias: How to tackle a wicked problem of ai? *Big Data Cogn. Comput.* **2021**, *5*, 18. [CrossRef]
130. Raisch, S.; Krakowski, S. Artificial Intelligence and management: The automation–augmentation paradox. *Acad. Manag. Rev.* **2021**, *46*, 192–210. [CrossRef]
131. Lyons, J.B.; Guznov, S.Y. Individual differences in human–machine trust: A multi-study look at the perfect automation schema. *Theor. Issues Ergon. Sci.* **2019**, *20*, 440–458. [CrossRef]
132. Nakao, Y.; Stumpf, S.; Ahmed, S.; Naseer, A.; Strappelli, L. Toward involving end-users in interactive human-in-the-loop AI fairness. *ACM Trans. Interact. Intell. Syst. (TiiS)* **2022**, *12*, 1–30. [CrossRef]
133. Yarger, L.; Cobb Payton, F.; Neupane, B. Algorithmic equity in the hiring of underrepresented IT job candidates. *Online Inf. Rev.* **2020**, *44*, 383–395. [CrossRef]
134. Zhou, Y.; Kantarcioglu, M.; Clifton, C. On Improving Fairness of AI Models with Synthetic Minority Oversampling Techniques. In Proceedings of the 2023 SIAM International Conference on Data Mining (SDM), Minneapolis, MN, USA, 27–29 April 2023; SIAM: Philadelphia, PA, USA, 2023; pp. 874–882.
135. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *366*, 447–453. [CrossRef]
136. Calmon, F.; Wei, D.; Vinzamuri, B.; Natesan Ramamurthy, K.; Varshney, K.R. Optimized pre-processing for discrimination prevention. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
137. Zhang, B.H.; Lemoine, B.; Mitchell, M. Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018; pp. 335–340.
138. Chiappa, S. Path-specific counterfactual fairness. *AAAI Conf. Artif. Intell.* **2019**, *33*, 7801–7808. [CrossRef]
139. Sun, Y.; Haghighat, F.; Fung, B.C. Trade-off between accuracy and fairness of data-driven building and indoor environment models: A comparative study of pre-processing methods. *Energy* **2022**, *239*, 122273. [CrossRef]
140. Sun, Y.; Fung, B.C.; Haghighat, F. The generalizability of pre-processing techniques on the accuracy and fairness of data-driven building models: A case study. *Energy Build.* **2022**, *268*, 112204. [CrossRef]
141. Wan, M.; Zha, D.; Liu, N.; Zou, N. In-processing modeling techniques for machine learning fairness: A survey. *ACM Trans. Knowl. Discov. Data* **2023**, *17*, 1–27. [CrossRef]
142. Sun, Y.; Fung, B.C.; Haghighat, F. In-Processing fairness improvement methods for regression Data-Driven building Models: Achieving uniform energy prediction. *Energy Build.* **2022**, *277*, 112565. [CrossRef]
143. Petersen, F.; Mukherjee, D.; Sun, Y.; Yurochkin, M. Post-processing for individual fairness. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 25944–25955.
144. Lohia, P.K.; Ramamurthy, K.N.; Bhide, M.; Saha, D.; Varshney, K.R.; Puri, R. Bias mitigation post-processing for individual and group fairness. In Proceedings of the Icassp 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 2847–2851.
145. Putzel, P.; Lee, S. Blackbox post-processing for multiclass fairness. *arXiv* **2022**, arXiv:2201.04461.
146. Jung, S.; Park, T.; Chun, S.; Moon, T. Re-weighting Based Group Fairness Regularization via Classwise Robust Optimization. *arXiv* **2023**, arXiv:2303.00442.
147. Lal, G.R.; Geyik, S.C.; Kenthapadi, K. Fairness-aware online personalization. *arXiv* **2020**, arXiv:2007.15270.
148. Wu, Y.; Zhang, L.; Wu, X. Counterfactual fairness: Unidentification, bound and algorithm. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019.
149. Cheong, J.; Kalkan, S.; Gunes, H. Counterfactual fairness for facial expression recognition. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 245–261.
150. Wang, X.; Li, B.; Su, X.; Peng, H.; Wang, L.; Lu, C.; Wang, C. Autonomous dispatch trajectory planning on flight deck: A search-resampling-optimization framework. *Eng. Appl. Artif. Intell.* **2023**, *119*, 105792. [CrossRef]
151. Xie, S.M.; Santurkar, S.; Ma, T.; Liang, P. Data selection for language models via importance resampling. *arXiv* **2023**, arXiv:2302.03169.
152. Khushi, M.; Shaukat, K.; Alam, T.M.; Hameed, I.A.; Uddin, S.; Luo, S.; Yang, X.; Reyes, M.C. A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access* **2021**, *9*, 109960–109975. [CrossRef]
153. Ghorbani, R.; Ghousi, R. Comparing different resampling methods in predicting students' performance using machine learning techniques. *IEEE Access* **2020**, *8*, 67899–67911. [CrossRef]
154. He, E.; Xie, Y.; Liu, L.; Chen, W.; Jin, Z.; Jia, X. Physics Guided Neural Networks for Time-Aware Fairness: An Application in Crop Yield Prediction. *AAAI Conf. Artif. Intell.* **2023**, *37*, 14223–14231. [CrossRef]
155. Wang, S.; Wang, B.; Zhang, Z.; Heidari, A.A.; Chen, H. Class-aware sample reweighting optimal transport for multi-source domain adaptation. *Neurocomputing* **2023**, *523*, 213–223. [CrossRef]
156. Song, P.; Li, P.; Dai, L.; Wang, T.; Chen, Z. Boosting R-CNN: Reweighting R-CNN samples by RPN's error for underwater object detection. *Neurocomputing* **2023**, *530*, 150–164. [CrossRef]
157. Jin, M.; Ju, C.J.T.; Chen, Z.; Liu, Y.C.; Droppo, J.; Stolcke, A. Adversarial reweighting for speaker verification fairness. *arXiv* **2022**, arXiv:2207.07776.
158. Kieninger, S.; Donati, L.; Keller, B.G. Dynamical reweighting methods for Markov models. *Curr. Opin. Struct. Biol.* **2020**, *61*, 124–131. [CrossRef]

159. Zhou, X.; Lin, Y.; Pi, R.; Zhang, W.; Xu, R.; Cui, P.; Zhang, T. Model agnostic sample reweighting for out-of-distribution learning. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 27203–27221.
160. Khalifa, N.E.; Loey, M.; Mirjalili, S. A comprehensive survey of recent trends in deep learning for digital images augmentation. *Artif. Intell. Rev.* **2022**, *55*, 2351–2377. [CrossRef]
161. Pastaltzidis, I.; Dimitriou, N.; Quezada-Tavarez, K.; Aidinlis, S.; Marquenie, T.; Gurzawska, A.; Tzovaras, D. 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, Seoul, Republic of Korea, 21–24 June 2022; pp. 2302–2314.
162. Kose, O.D.; Shen, Y. Fair node representation learning via adaptive data augmentation. *arXiv* **2022**, arXiv:2201.08549.
163. Zhang, Y.; Sang, J. Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 4346–4354.
164. Zheng, L.; Zhu, Y.; He, J. Fairness-aware Multi-view Clustering. In Proceedings of the 2023 SIAM International Conference on Data Mining (SDM), Minneapolis, MN, USA, 27–29 April 2023; SIAM: Philadelphia, PA, USA, 2023; pp. 856–864.
165. Le Quy, T.; Friege, G.; Ntoutsi, E. A Review of Clustering Models in Educational Data Science Toward Fairness-Aware Learning. In *Educational Data Science: Essentials, Approaches, and Tendencies: Proactive Education based on Empirical Big Data Evidence*; Springer: Singapore, 2023; pp. 43–94.
166. Chierichetti, F.; Kumar, R.; Lattanzi, S.; Vassilvitskii, S. Fair clustering through fairlets. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
167. Kamishima, T.; Akaho, S.; Asoh, H.; Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, 24–28 September 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 35–50.
168. Chakraborty, J.; Majumder, S.; Menzies, T. Bias in machine learning software: Why? how? what to do? In Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Athens, Greece, 23–28 August 2021; pp. 429–440.
169. Blagus, R.; Lusa, L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinform.* **2013**, *14*, 106. [CrossRef] [PubMed]
170. Blagus, R.; Lusa, L. Evaluation of smote for high-dimensional class-imbalanced microarray data. In Proceedings of the 2012 11th International Conference on Machine Learning and Applications, Boca Raton, FL, USA, 12–15 December 2012; IEEE: Piscataway, NJ, USA, 2012; Volume 2, pp. 89–94.
171. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
172. Zhao, W.; Alwidian, S.; Mahmoud, Q.H. Adversarial Training Methods for Deep Learning: A Systematic Review. *Algorithms* **2022**, *15*, 283. [CrossRef]
173. Bai, T.; Luo, J.; Zhao, J.; Wen, B.; Wang, Q. Recent advances in adversarial training for adversarial robustness. *arXiv* **2021**, arXiv:2102.01356.
174. Wong, E.; Rice, L.; Kolter, J.Z. Fast is better than free: Revisiting adversarial training. *arXiv* **2020**, arXiv:2001.03994.
175. Andriushchenko, M.; Flammarion, N. Understanding and improving fast adversarial training. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 16048–16059.
176. Shafahi, A.; Najibi, M.; Ghiasi, M.A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L.S.; Taylor, G.; Goldstein, T. Adversarial training for free! *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
177. Lim, J.; Kim, Y.; Kim, B.; Ahn, C.; Shin, J.; Yang, E.; Han, S. BiasAdv: Bias-Adversarial Augmentation for Model Debiasing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 3832–3841.
178. Hong, J.; Zhu, Z.; Yu, S.; Wang, Z.; Dodge, H.H.; Zhou, J. Federated adversarial debiasing for fair and transferable representations. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Singapore, 14–18 August 2021; pp. 617–627.
179. Darlow, L.; Jastrzębski, S.; Storkey, A. Latent adversarial debiasing: Mitigating collider bias in deep neural networks. *arXiv* **2020**, arXiv:2011.11486.
180. Mishler, A.; Kennedy, E.H.; Chouldechova, A. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event/Toronto, ON, Canada, 3–10 March 2021; pp. 386–400.
181. Roy, S.; Salimi, B. Causal inference in data analysis with applications to fairness and explanations. In *Reasoning Web. Causality, Explanations and Declarative Knowledge: 18th International Summer School 2022, Berlin, Germany, 27–30 September 2022*; Springer: Cham, Switzerland, 2023; pp. 105–131.
182. Madras, D.; Creager, E.; Pitassi, T.; Zemel, R. Fairness through causal awareness: Learning causal latent-variable models for biased data. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 349–358.
183. Loftus, J.R.; Russell, C.; Kusner, M.J.; Silva, R. Causal reasoning for algorithmic fairness. *arXiv* **2018**, arXiv:1805.05859.
184. Hinnefeld, J.H.; Cooman, P.; Mammo, N.; Deese, R. Evaluating fairness metrics in the presence of dataset bias. *arXiv* **2018**, arXiv:1809.09245.

185. Modén, M.U.; Lundin, J.; Tallvid, M.; Ponti, M. Involving teachers in meta-design of AI to ensure situated fairness. *Proceedings* **2022**, *1613*, 0073.
186. Zhao, C.; Li, C.; Li, J.; Chen, F. Fair meta-learning for few-shot classification. In Proceedings of the 2020 IEEE International Conference on Knowledge Graph (ICKG), Nanjing, China, 9–11 August 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 275–282.
187. Hsu, B.; Chen, X.; Han, Y.; Namkoong, H.; Basu, K. An Operational Perspective to Fairness Interventions: Where and How to Intervene. *arXiv* **2023**, arXiv:2302.01574.
188. Salvador, T.; Cairns, S.; Voleti, V.; Marshall, N.; Oberman, A. Faircal: Fairness calibration for face verification. *arXiv* **2021**, arXiv:2106.03761.
189. Noriega-Campero, A.; Bakker, M.A.; Garcia-Bulle, B.; Pentland, A. Active fairness in algorithmic decision making. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019; pp. 77–83.
190. Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; Weinberger, K.Q. On fairness and calibration. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
191. Tahir, A.; Cheng, L.; Liu, H. Fairness through Aleatoric Uncertainty. *arXiv* **2023**, arXiv:2304.03646.
192. Tubella, A.A.; Barsotti, F.; Koçer, R.G.; Mendez, J.A. Ethical implications of fairness interventions: What might be hidden behind engineering choices? *Ethics Inf. Technol.* **2022**, *24*, 12. [CrossRef]
193. Kamishima, T.; Akaho, S.; Asoh, H.; Sakuma, J. Model-based and actual independence for fairness-aware classification. *Data Min. Knowl. Discov.* **2018**, *32*, 258–286. [CrossRef]
194. Kasmi, M.L. Machine Learning Fairness in Finance: An Application to Credit Scoring. Ph.D. Thesis, Tilburg University, Tilburg, The Netherlands, 2021.
195. Zhang, T.; Zhu, T.; Li, J.; Han, M.; Zhou, W.; Philip, S.Y. Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 1763–1774. [CrossRef]
196. Caton, S.; Haas, C. Fairness in machine learning: A survey. *arXiv* **2020**, arXiv:2010.04053.
197. Small, E.A.; Sokol, K.; Manning, D.; Salim, F.D.; Chan, J. Equalised Odds is not Equal Individual Odds: Post-processing for Group and Individual Fairness. *arXiv* **2023**, arXiv:2304.09779.
198. Jang, T.; Shi, P.; Wang, X. Group-aware threshold adaptation for fair classification. *AAAI Conf. Artif. Intell.* **2022**, *36*, 6988–6995. [CrossRef]
199. Nguyen, D.; Gupta, S.; Rana, S.; Shilton, A.; Venkatesh, S. Fairness improvement for black-box classifiers with Gaussian process. *Inf. Sci.* **2021**, *576*, 542–556. [CrossRef]
200. Iosifidis, V.; Fetahu, B.; Ntoutsi, E. Fae: A fairness-aware ensemble framework. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1375–1380.
201. Zhong, M.; Tandon, R. Learning Fair Classifiers via Min-Max F-divergence Regularization. *arXiv* **2023**, arXiv:2306.16552.
202. Nandy, P.; Diciccio, C.; Venugopalan, D.; Logan, H.; Basu, K.; El Karoui, N. Achieving Fairness via Post-Processing in Web-Scale Recommender Systems. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; pp. 715–725.
203. Boratto, L.; Fenu, G.; Marras, M. Interplay between upsampling and regularization for provider fairness in recommender systems. *User Model. User Adapt. Interact.* **2021**, *31*, 421–455. [CrossRef]
204. Yao, S.; Huang, B. Beyond parity: Fairness objectives for collaborative filtering. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
205. Yu, B.; Wu, J.; Ma, J.; Zhu, Z. Tangent-normal adversarial regularization for semi-supervised learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10676–10684.
206. Sato, M.; Suzuki, J.; Kiyono, S. Effective adversarial regularization for neural machine translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 204–210.
207. Nasr, M.; Shokri, R.; Houmansadr, A. Machine learning with membership privacy using adversarial regularization. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, Toronto, ON, Canada, 15–19 October 2018; pp. 634–646.
208. Mertikopoulos, P.; Papadimitriou, C.; Piliouras, G. Cycles in adversarial regularized learning. In Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, LA, USA, 7–10 January 2018; SIAM: Philadelphia, PA, USA, 2018; pp. 2703–2717.
209. Du, M.; Yang, F.; Zou, N.; Hu, X. Fairness in deep learning: A computational perspective. *IEEE Intell. Syst.* **2020**, *36*, 25–34. [CrossRef]
210. Horesh, Y.; Haas, N.; Mishraky, E.; Resheff, Y.S.; Meir Lador, S. Paired-consistency: An example-based model-agnostic approach to fairness regularization in machine learning. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, 16–20 September 2019; Springer: Cham, Switzerland, 2020; pp. 590–604.
211. Lohaus, M.; Kleindessner, M.; Kenthapadi, K.; Locatello, F.; Russell, C. Are Two Heads the Same as One? Identifying Disparate Treatment in Fair Neural Networks. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 16548–16562.
212. Romano, Y.; Bates, S.; Candes, E. Achieving equalized odds by resampling sensitive attributes. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 361–371.
213. Cho, J.; Hwang, G.; Suh, C. A fair classifier using mutual information. In Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT), Los Angeles, CA, USA, 21–26 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 2521–2526.

214. Wieling, M.; Nerbonne, J.; Baayen, R.H. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE* **2011**, *6*, e23613. [CrossRef]
215. Bhanot, K.; Qi, M.; Erickson, J.S.; Guyon, I.; Bennett, K.P. The problem of fairness in synthetic healthcare data. *Entropy* **2021**, *23*, 1165. [CrossRef]
216. Brusaferri, A.; Matteucci, M.; Spinelli, S.; Vitali, A. Probabilistic electric load forecasting through Bayesian mixture density networks. *Appl. Energy* **2022**, *309*, 118341. [CrossRef]
217. Errica, F.; Bacciu, D.; Micheli, A. Graph mixture density networks. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 3025–3035.
218. Makansi, O.; Ilg, E.; Cicek, O.; Brox, T. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7144–7153.
219. John, P.G.; Vijaykeerthy, D.; Saha, D. Verifying individual fairness in machine learning models. In Proceedings of the Conference on Uncertainty in Artificial Intelligence, PMLR, Virtual, 3–6 August 2020; pp. 749–758.
220. Han, X.; Baldwin, T.; Cohn, T. Towards equal opportunity fairness through adversarial learning. *arXiv* **2022**, arXiv:2203.06317.
221. Shen, A.; Han, X.; Cohn, T.; Baldwin, T.; Frermann, L. Optimising equal opportunity fairness in model training. *arXiv* **2022**, arXiv:2205.02393.
222. Verma, S.; Rubin, J. Fairness definitions explained. In Proceedings of the International Workshop on Software Fairness, Gothenburg, Sweden, 29 May 2018; pp. 1–7.
223. Balashankar, A.; Wang, X.; Packer, B.; Thain, N.; Chi, E.; Beutel, A. Can we improve model robustness through secondary attribute counterfactuals? In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Virtual, 7–11 November 2021; pp. 4701–4712.
224. Dong, Z.; Zhu, H.; Cheng, P.; Feng, X.; Cai, G.; He, X.; Xu, J.; Wen, J. Counterfactual learning for recommender system. In Proceedings of the 14th ACM Conference on Recommender Systems, Virtual Event, Brazil, 22–26 September 2020; pp. 568–569.
225. Veitch, V.; D'Amour, A.; Yadlowsky, S.; Eisenstein, J. Counterfactual invariance to spurious correlations in text classification. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 16196–16208.
226. Chang, Y.C.; Lu, C.J. Oblivious polynomial evaluation and oblivious neural learning. In Proceedings of the Advances in Cryptology—ASIACRYPT 2001: 7th International Conference on the Theory and Application of Cryptology and Information Security Gold Coast, Australia, 9–13 December 2001; Springer: Berlin/Heidelberg, Germany, 2001; pp. 369–384.
227. Meister, M.; Sheikholeslami, S.; Andersson, R.; Ormenisan, A.A.; Dowling, J. Towards distribution transparency for supervised ML with oblivious training functions. In Proceedings of the Workshop MLOps Syst, Austin, TX, USA, 2–4 March 2020; pp. 1–3.
228. Liu, J.; Juuti, M.; Lu, Y.; Asokan, N. Oblivious neural network predictions via minionn transformations. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, Dallas, TX, USA, 30 October–3 November 2017; pp. 619–631.
229. Goel, N.; Yaghini, M.; Faltings, B. Non-discriminatory machine learning through convex fairness criteria. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018; p. 116.
230. Makhlouf, K.; Zhioua, S.; Palamidessi, C. Survey on causal-based machine learning fairness notions. *arXiv* **2020**, arXiv:2010.09553.
231. Gölz, P.; Kahng, A.; Procaccia, A.D. Paradoxes in fair machine learning. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
232. Ferryman, K.; Pitcan, M. *Fairness in Precision Medicine*; Data and Society Research Institute: New York, NY, USA, 2018.
233. Dempsey, W.; Foster, I.; Fraser, S.; Kesselman, C. Sharing begins at home: How continuous and ubiquitous FAIRness can enhance research productivity and data reuse. *Harv. Data Sci. Rev.* **2022**, *4*, 10–11. [CrossRef]
234. Durand, C.M.; Segev, D.; Sugarman, J. Realizing HOPE: The ethics of organ transplantation from HIV-positive donors. *Ann. Intern. Med.* **2016**, *165*, 138–142. [CrossRef]
235. Rubinstein, Y.R.; McInnes, P. NIH/NCATS/GRDR® Common Data Elements: A leading force for standardized data collection. *Contemp. Clin. Trials* **2015**, *42*, 78–80. [CrossRef]
236. Frick, K.D. Micro-costing quantity data collection methods. *Med. Care* **2009**, *47*, S76. [CrossRef] [PubMed]
237. Rothstein, M.A. Informed consent for secondary research under the new NIH data sharing policy. *J. Law Med. Ethics* **2021**, *49*, 489–494. [CrossRef] [PubMed]
238. Greely, H.T.; Grady, C.; Ramos, K.M.; Chiong, W.; Eberwine, J.; Farahany, N.A.; Johnson, L.S.M.; Hyman, B.T.; Hyman, S.E.; Rommelfanger, K.S.; et al. Neuroethics guiding principles for the NIH BRAIN initiative. *J. Neurosci.* **2018**, *38*, 10586. [CrossRef] [PubMed]
239. Nijhawan, L.P.; Janodia, M.D.; Muddukrishna, B.; Bhat, K.M.; Bairy, K.L.; Udupa, N.; Musmade, P.B. Informed consent: Issues and challenges. *J. Adv. Pharm. Technol. Res.* **2013**, *4*, 134.
240. Elliot, M.; Mackey, E.; O'Hara, K.; Tudor, C. *The Anonymisation Decision-Making Framework*; UKAN: Manchester, UK, 2016; p. 171.
241. Rosner, G. De-Identification as Public Policy. *J. Data Prot. Priv.* **2019**, *3*, 1–18.
242. Moretón, A.; Jaramillo, A. Anonymisation and re-identification risk for voice data. *Eur. Data Prot. L. Rev.* **2021**, *7*, 274. [CrossRef]
243. Rumbold, J.M.; Pierscionek, B.K. A critique of the regulation of data science in healthcare research in the European Union. *BMC Med. Ethics* **2017**, *18*, 27. [CrossRef] [PubMed]
244. Stalla-Bourdillon, S.; Knight, A. Anonymous data v. personal data-false debate: An EU perspective on anonymization, pseudonymization and personal data. *Wis. Int'l LJ* **2016**, *34*, 284.

245. Ilavsky, J. Nika: Software for two-dimensional data reduction. *J. Appl. Crystallogr.* **2012**, *45*, 324–328. [CrossRef]
246. Fietzke, J.; Liebetrau, V.; Günther, D.; Gürs, K.; Hametner, K.; Zumholz, K.; Hansteen, T.; Eisenhauer, A. An alternative data acquisition and evaluation strategy for improved isotope ratio precision using LA-MC-ICP-MS applied to stable and radiogenic strontium isotopes in carbonates. *J. Anal. At. Spectrom.* **2008**, *23*, 955–961. [CrossRef]
247. Gwynne, S. *Conventions in the Collection and Use of Human Performance Data*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2010; pp. 10–928.
248. Buckleton, J.S.; Bright, J.A.; Cheng, K.; Budowle, B.; Coble, M.D. NIST interlaboratory studies involving DNA mixtures (MIX13): A modern analysis. *Forensic Sci. Int. Genet.* **2018**, *37*, 172–179. [CrossRef] [PubMed]
249. Sydes, M.R.; Johnson, A.L.; Meredith, S.K.; Rauchenberger, M.; South, A.; Parmar, M.K. Sharing data from clinical trials: The rationale for a controlled access approach. *Trials* **2015**, *16*, 104. [CrossRef] [PubMed]
250. Abdul Razack, H.I.; Aranjani, J.M.; Mathew, S.T. Clinical trial transparency regulations: Implications to various scholarly publishing stakeholders. *Sci. Public Policy* **2022**, *49*, 951–961. [CrossRef]
251. Alemayehu, D.; Anziano, R.J.; Levenstein, M. Perspectives on clinical trial data transparency and disclosure. *Contemp. Clin. Trials* **2014**, *39*, 28–33. [CrossRef]
252. Force, J.T.; Initiative, T. Security and privacy controls for federal information systems and organizations. *NIST Spec. Publ.* **2013**, *800*, 8–13.
253. Plans, B.E.A. Assessing security and privacy controls in federal information systems and organizations. *NIST Spec. Publ.* **2014**, *800*, 53A.
254. Dempsey, K.; Witte, G.; Rike, D. *Summary of NIST SP 800-53, Revision 4: Security and Privacy Controls for Federal Information Systems and Organizations*; Technical Report; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2014.
255. Passi, S.; Jackson, S.J. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proc. ACM Hum. Comput. Interact.* **2018**, *2*, 1–28. [CrossRef]
256. Hutt, E.; Polikoff, M.S. Toward a framework for public accountability in education reform. *Educ. Res.* **2020**, *49*, 503–511. [CrossRef]
257. Carle, S.D. A social movement history of Title VII Disparate Impact analysis. *Fla. L. Rev.* **2011**, *63*, 251. [CrossRef]
258. Griffith, D.; McKinney, B. Using Disparate Impact Analysis to Develop Anti-Racist Policies: An Application to Coronavirus Liability Waivers. *J. High. Educ. Manag.* **2021**, *36*, 104–116.
259. Liu, S.; Ge, Y.; Xu, S.; Zhang, Y.; Marian, A. Fairness-aware federated matrix factorization. In Proceedings of the 16th ACM Conference on Recommender Systems, Seattle, WA, USA, 18–22 September 2022; pp. 168–178.
260. Gao, R.; Ge, Y.; Shah, C. FAIR: Fairness-aware information retrieval evaluation. *J. Assoc. Inf. Sci. Technol.* **2022**, *73*, 1461–1473. [CrossRef]
261. Zhang, W.; Ntoutsi, E. Faht: An adaptive fairness-aware decision tree classifier. *arXiv* **2019**, arXiv:1907.07237.
262. Serna, I.; DeAlcala, D.; Morales, A.; Fierrez, J.; Ortega-Garcia, J. IFBiD: Inference-free bias detection. *arXiv* **2021**, arXiv:2109.04374.
263. Li, B.; Peng, H.; Sainju, R.; Yang, J.; Yang, L.; Liang, Y.; Jiang, W.; Wang, B.; Liu, H.; Ding, C. Detecting gender bias in transformer-based models: A case study on BERT. *arXiv* **2021**, arXiv:2110.15733.
264. Constantin, R.; Dück, M.; Alexandrov, A.; Matošević, P.; Keidar, D.; El-Assady, M. How Do Algorithmic Fairness Metrics Align with Human Judgement? A Mixed-Initiative System for Contextualized Fairness Assessment. In Proceedings of the 2022 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX), Oklahoma City, OK, USA, 16 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–7.
265. Goel, Z. Algorithmic Fairness Final Report.
266. Bird, S.; Dudík, M.; Edgar, R.; Horn, B.; Lutz, R.; Milan, V.; Sameki, M.; Wallach, H.; Walker, K. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft Tech. Rep.* **2020**.
267. Jethani, N.; Sudarshan, M.; Aphinyanaphongs, Y.; Ranganath, R. Have We Learned to Explain?: How Interpretability Methods Can Learn to Encode Predictions in their Interpretations. In Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, Virtual, 13–15 April 2021; pp. 1459–1467.
268. Stiglic, G.; Kocbek, P.; Fijacko, N.; Zitnik, M.; Verbert, K.; Cilar, L. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1379. [CrossRef]
269. Moraffah, R.; Karami, M.; Guo, R.; Raglin, A.; Liu, H. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explor. Newsl.* **2020**, *22*, 18–33. [CrossRef]
270. Jacovi, A.; Swayamdipta, S.; Ravfogel, S.; Elazar, Y.; Choi, Y.; Goldberg, Y. Contrastive explanations for model interpretability. *arXiv* **2021**, arXiv:2103.01378.
271. Jeffries, A.C.; Wallace, L.; Coutts, A.J.; McLaren, S.J.; McCall, A.; Impellizzeri, F.M. Athlete-reported outcome measures for monitoring training responses: A systematic review of risk of bias and measurement property quality according to the COSMIN guidelines. *Int. J. Sport. Physiol. Perform.* **2020**, *15*, 1203–1215. [CrossRef] [PubMed]
272. Oliveira-Rodrigues, C.; Correia, A.M.; Valente, R.; Gil, Á.; Gandra, M.; Liberal, M.; Rosso, M.; Pierce, G.; Sousa-Pinto, I. Assessing data bias in visual surveys from a cetacean monitoring programme. *Sci. Data* **2022**, *9*, 682. [CrossRef]
273. Memarian, B.; Doleck, T. Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI), and higher education: A systematic review. *Comput. Educ. Artif. Intell.* **2023**, *5*, 100152. [CrossRef]

274. Marcinkowski, F.; Kieslich, K.; Starke, C.; Lünich, M. Implications of AI (un-) fairness in higher education admissions: The effects of perceived AI (un-) fairness on exit, voice and organizational reputation. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 122–130.
275. Kizilcec, R.F.; Lee, H. Algorithmic fairness in education. In *The Ethics of Artificial Intelligence in Education*; Routledge: Boca Raton, FL, USA, 2022; pp. 174–202.
276. Mashhadi, A.; Zolyomi, A.; Quedado, J. A Case Study of Integrating Fairness Visualization Tools in Machine Learning Education. In Proceedings of the CHI Conference on Human Factors in Computing Systems Extended Abstracts, New Orleans, LA, USA, 29 April–5 May 2022; pp. 1–7.
277. Fenu, G.; Galici, R.; Marras, M. Experts' view on challenges and needs for fairness in artificial intelligence for education. In *International Conference on Artificial Intelligence in Education*; Springer: Cham, Switzerland, 2022; pp. 243–255.
278. Chen, R.J.; Chen, T.Y.; Lipkova, J.; Wang, J.J.; Williamson, D.F.; Lu, M.Y.; Sahai, S.; Mahmood, F. Algorithm fairness in ai for medicine and healthcare. *arXiv* **2021**, arXiv:2110.00603.
279. Gichoya, J.W.; McCoy, L.G.; Celi, L.A.; Ghassemi, M. Equity in essence: A call for operationalising fairness in machine learning for healthcare. *BMJ Health Care Inform.* **2021**, *28*, e100289. [CrossRef]
280. Johnson, K.B.; Wei, W.Q.; Weeraratne, D.; Frisse, M.E.; Misulis, K.; Rhee, K.; Zhao, J.; Snowdon, J.L. Precision medicine, AI, and the future of personalized health care. *Clin. Transl. Sci.* **2021**, *14*, 86–93. [CrossRef]
281. Chiao, V. Fairness, accountability and transparency: Notes on algorithmic decision-making in criminal justice. *Int. J. Law Context* **2019**, *15*, 126–139. [CrossRef]
282. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine bias. In *Ethics of Data and Analytics*; Auerbach Publications: Boca Raton, FL, USA, 2022; pp. 254–264.
283. Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res.* **2021**, *50*, 3–44. [CrossRef]
284. Mujtaba, D.F.; Mahapatra, N.R. Ethical considerations in AI-based recruitment. In Proceedings of the 2019 IEEE International Symposium on Technology and Society (ISTAS), Medford, MA, USA, 15–16 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–7.
285. Hunkenschroer, A.L.; Luetge, C. Ethics of AI-enabled recruiting and selection: A review and research agenda. *J. Bus. Ethics* **2022**, *178*, 977–1007. [CrossRef]
286. Nugent, S.E.; Scott-Parker, S. Recruitment AI has a Disability Problem: Anticipating and mitigating unfair automated hiring decisions. In *Towards Trustworthy Artificial Intelligent Systems*; Springer: Cham, Swizeraland, 2022; pp. 85–96.
287. Hurlin, C.; Pérignon, C.; Saurin, S. The fairness of credit scoring models. *arXiv* **2022**, arXiv:2205.10200.
288. Gemalmaz, M.A.; Yin, M. Understanding Decision Subjects' Fairness Perceptions and Retention in Repeated Interactions with AI-Based Decision Systems. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, Oxford, UK, 19–21 May 2021; pp. 295–306.
289. Genovesi, S.; Mönig, J.M.; Schmitz, A.; Poretschkin, M.; Akila, M.; Kahdan, M.; Kleiner, R.; Krieger, L.; Zimmermann, A. Standardizing fairness-evaluation procedures: Interdisciplinary insights on machine learning algorithms in creditworthiness assessments for small personal loans. *AI Ethics* **2023**, 1–17. [CrossRef]
290. Hiller, J.S. Fairness in the eyes of the beholder: Ai; fairness; and alternative credit scoring. *W. Va. L. Rev.* **2020**, *123*, 907.
291. Kumar, I.E.; Hines, K.E.; Dickerson, J.P. Equalizing credit opportunity in algorithms: Aligning algorithmic fairness research with us fair lending regulation. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, Oxford, UK, 19–21 May 2021; pp. 357–368.
292. Moldovan, D. Algorithmic decision making methods for fair credit scoring. *IEEE Access* **2023**, *11*, 59729–59743. [CrossRef]
293. Rodgers, W.; Nguyen, T. Advertising benefits from ethical Artificial Intelligence algorithmic purchase decision pathways. *J. Bus. Ethics* **2022**, *178*, 1043–1061. [CrossRef]
294. Yuan, D. Artificial Intelligence, Fairness and Productivity. Ph.D. Thesis, University of Pittsburgh, Pittsburgh, PA, USA, 2023.
295. Bateni, A.; Chan, M.C.; Eitel-Porter, R. AI fairness: From principles to practice. *arXiv* **2022**, arXiv:2207.09833.
296. Rossi, F. Building trust in Artificial Intelligence. *J. Int. Aff.* **2018**, *72*, 127–134.
297. Bang, J.; Kim, S.; Nam, J.W.; Yang, D.G. Ethical chatbot design for reducing negative effects of biased data and unethical conversations. In Proceedings of the 2021 International Conference on Platform Technology and Service (PlatCon), Jeju, Republic of Korea, 23–25 August 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–5.
298. Følstad, A.; Araujo, T.; Law, E.L.C.; Brandtzaeg, P.B.; Papadopoulos, S.; Reis, L.; Baez, M.; Laban, G.; McAllister, P.; Ischen, C.; et al. Future directions for chatbot research: An interdisciplinary research agenda. *Computing* **2021**, *103*, 2915–2942. [CrossRef]
299. Lewicki, K.; Lee, M.S.A.; Cobbe, J.; Singh, J. Out of Context: Investigating the Bias and Fairness Concerns of "Artificial Intelligence as a Service". In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; pp. 1–17.
300. Chen, Q.; Lu, Y.; Gong, Y.; Xiong, J. Can AI chatbots help retain customers? Impact of AI service quality on customer loyalty. *Internet Res.* **2023**. [CrossRef]
301. Chen, Y.; Jensen, S.; Albert, L.J.; Gupta, S.; Lee, T. Artificial Intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. *Inf. Syst. Front.* **2023**, *25*, 161–182. [CrossRef]

302. Simbeck, K. FAccT-Check on AI regulation: Systematic Evaluation of AI Regulation on the Example of the Legislation on the Use of AI in the Public Sector in the German Federal State of Schleswig-Holstein. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; pp. 89–96.
303. Srivastava, B.; Rossi, F.; Usmani, S.; Bernagozzi, M. Personalized chatbot trustworthiness ratings. *IEEE Trans. Technol. Soc.* **2020**, *1*, 184–192. [CrossRef]
304. Hulsen, T. Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare. *AI* **2023**, *4*, 652–666. [CrossRef]
305. Chen, Z. Collaboration among recruiters and Artificial Intelligence: Removing human prejudices in employment. *Cogn. Technol. Work.* **2023**, *25*, 135–149. [CrossRef] [PubMed]
306. Rieskamp, J.; Hofeditz, L.; Mirbabaie, M.; Stieglitz, S. Approaches to improve fairness when deploying ai-based algorithms in hiring—Using a systematic literature review to guide future research. In Proceedings of the 56th Hawaii International Conference on System Sciences, HICSS 2023, Maui, HI, USA, 3–6 January 2023.
307. Hunkenschroer, A.L.; Kriebitz, A. Is AI recruiting (un) ethical? A human rights perspective on the use of AI for hiring. *AI Ethics* **2023**, *3*, 199–213. [CrossRef] [PubMed]
308. Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of Data and Analytics*; Auerbach Publications: Boca Raton, FL, USA, 2022; pp. 296–299.
309. Hunkenschroer, A.L.; Lütge, C. How to improve fairness perceptions of AI in hiring: The crucial role of positioning and sensitization. *AI Ethics J.* **2021**, *2*, 1–19. [CrossRef]