1

Can AI Be Fair?

Yi-Chieh Lee, Jingshu Li, Renwen Zhang, and Nan Liu

1.1 What Is Fairness? Why Does It Matter in AI?

1.1.1 What Is Fairness in AI?

Fairness has historically been a central concern across philosophical, legal, and sociological domains, but its significance takes on new dimensions in "Responsible AI." As AI systems are increasingly deployed in high-stakes domains such as healthcare, questions of equitable treatment and accountability have become critical. Bias in AI, manifesting as systematic distortions that disproportionately disadvantage certain individuals or groups, poses serious risks (Mehrabi et al., 2021). Fairness, in this context, acts as a foundational pillar of Responsible AI: without careful attention to bias and equitable outcomes, algorithmic systems risk perpetuating discrimination and undermining public trust (Caton and Haas, 2024).

The concept of fairness has been extensively discussed in multiple fields and is a critical topic in social, cultural, and institutional frameworks (Hutchinson and Mitchell, 2019). It embodies both a pursuit of legitimacy and a practical need to distribute resources and opportunities equitably. However, reaching a single unified definition of "fairness" is not straightforward; different disciplines often view and emphasize fairness from their own unique perspectives. From philosophy to sociology, from law to quantitative science, there are notable differences in how fairness is conceptualized, evaluated, and implemented.

From a philosophical standpoint, fairness encompasses both outcomes and processes: not only the results of decisions (e.g., who receives what resources or opportunities), but also the procedures and principles by which those

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

decisions are made. In ancient Greece, Aristotle proposed two forms of justice: "distributive justice," which addresses how benefits or resources should be allocated proportionally, and "corrective justice," which focuses on redressing errors or harms (von Leyden, 1985). In the twentieth century, John Rawls introduced the idea of a "veil of ignorance," causing people to develop social structures without knowing their own social status, thus underscoring the protection of the most disadvantaged (Rawls, 1991). Around the same time, Robert Nozick focused on the legitimacy and due process of individual property rights and freedoms (Nozick, 1974). Although philosophers approach fairness from diverse perspectives, they generally agree on the need to consider both appropriate support for disadvantaged groups and the legitimacy of procedures and rights.

From a sociological perspective, fairness is situated within concrete social structures and group interactions. For example, social class hierarchies in educational systems or power imbalances in racially segregated neighborhoods illustrate how social structures and everyday group interactions can influence the life chances of individuals. Sociological research often highlights macro-level power relations and institutionalized inequalities, paying special attention to the lingering effects of historical and cultural factors on certain groups (Mulligan et al., 2019), such as African American communities affected by redlining policies, or indigenous populations facing ongoing impacts of colonialism. In sociological terms, even if formal rules apply uniformly to everyone, overlooking the distinct needs of certain groups can lead to subtler forms of injustice perpetuated by institutions or technologies.

In the legal domain, fairness is closely linked to principles such as anti-discrimination and equal treatment, often manifested in prohibitions against direct or indirect discrimination. Many jurisdictions explicitly stipulate in their constitutions or civil rights laws that discrimination based on sensitive attributes, such as gender, race, religion, age, or disability, is unlawful. Legal frameworks also provide mechanisms for individuals to seek redress, such as anti-discrimination tribunals, equality commissions, and judicial review.

The rise of quantitative science has enabled more precise measurements and definitions of fairness, laying the theoretical and practical groundwork for implementing "AI fairness" in artificial intelligence systems (Hutchinson and Mitchell, 2019). These quantitative approaches provide mathematically formalized criteria that help evaluate whether algorithmic outcomes are equitable across or within groups. These and related fairness metrics will be introduced in more detail later in the chapter. By applying these quantitative metrics to identify and address bias or discrimination, it becomes easier to evaluate and mitigate potential unfairness between different groups.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

In general, AI Fairness refers to ensuring non-discriminatory, bias-free decisions and services for different population groups throughout the entire algorithm design, training, deployment, and application (Hutchinson and Mitchell, 2019). This encompasses not only the statistical balance in decision outcomes but also the soundness of how models handle individual characteristics and the explainability of these processes. From the standpoint of the AI field, integrating multiple disciplinary perspectives enables a more holistic examination of values, institutional arrangements, and technical implementation. Philosophy clarifies the concept of fairness and why it is worth pursuing; sociology shows how social structures and historical factors shape inequalities among different groups; law provides baseline rules and accountability frameworks to combat discrimination; and quantitative science gives us the tools to identify and mitigate unfairness on a technical level. During this process, conflicting priorities and trade-offs can arise due to variations in context and objectives (Narayanan, 2018; Mehrabi et al., 2021); however, it is precisely through these challenges that the notion of fairness becomes continually enriched and expanded, evolving into a shared concern that spans ethics, law, and technology.

1.1.2 Why Does Fairness Matter in AI?

As AI has been integrated into various sectors of society, *AI Fairness* is increasingly important in practice. This urgency lies in AI's penetration of areas that affect people's vital interests, such as medical diagnosis, recruitment and hiring, and financial credit. If algorithms harbor implicit bias or discrimination in these contexts, once decisions turn unfair, the consequences not only harm individual opportunities and rights but also amplify existing social inequalities. For example, using predictive models in healthcare to determine which patients should receive prioritized treatment can systematically undermine access to medical resources for certain ethnic groups or those of lower socioeconomic status if the model is biased against them (Chen et al., 2023). In recruitment and hiring, facial analysis tools embedded in video interviews have been found to rate candidates differently based on perceived gender expression or ethnicity, even when qualifications are equal (Köchling and Wehner, 2020; Hemamou et al., 2021). A similar issue may arise in financial credit: automated fraud detection systems may flag transactions from neighborhoods with higher minority populations more frequently, leading to account freezes or denials that disproportionately affect those communities (Pombal et al., 2022; Genovesi et al., 2024).

These concrete harms illustrate that AI fairness is not just a technical concern, but one with ethical, legal, and societal implications. In fact, the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

importance of ensuring AI fairness is closely tied to the protection of human rights and the maintenance of social stability (Teo, 2024). Rapid technological advances have, to some extent, improved the efficiency and accuracy of information processing and decision-making, but they have also brought notable ethical and legal challenges. If an AI system displays overt discrimination in judicial, educational, medical, or financial decisions, it not only violates anti-discrimination legislation but also undermines individual dignity and society's fundamental pursuit of equality. In other words, keeping AI decision-making impartial is fundamentally about safeguarding people's basic rights and societal values: individuals should have equal access to resources and opportunities, as well as the ability to question and appeal algorithmic decisions (Constantinides et al., 2024). When the public becomes aware of potential AI bias but lacks the means to understand or correct it, widespread skepticism and resistance to AI may ensue (Harper and Weber, 2022). Public trust can only be maintained by ensuring that algorithmic judgments are reliable and legitimate.

1.2 Why Is Fairness Challenging in AI? Does It Create Conflicts with Other Responsible AI Principles? What Makes It So Hard to Solve?

AI Fairness remains one of the most challenging ethical issues in AI development and deployment. One of the fundamental challenges of fairness in AI is that fairness itself is a contested concept with multiple definitions. Different disciplines and stakeholders define fairness in AI differently, leading to various mathematical and ethical interpretations. The major definitions of AI fairness can be categorized into three types: group fairness, individual fairness, and procedure fairness (Mehrabi et al., 2021).

Group fairness ensures that individuals categorized by certain attributes (e.g., race, gender, education, or social status) receive similar outcomes from an AI system. Key metrics include: (1) Demographic parity (or statistical parity), which ensures equal outcomes across groups: for example, if an automated loan approval system approves 70 percent of applicants overall, then demographic parity would require that approval rates are also close to 70 percent across different racial or gender groups; (2) Equalized odds that ensure similar error rates across different groups: such as ensuring that false negative rates (e.g., denying a loan to a qualified applicant) are equally low for both men and women; and (3) Disparate impact, which assesses if the proportion of the positive predictions is similar across groups: for example, whether

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

any group's selection rate is less than 80 percent of that of the most favored group (Pessach and Shmueli, 2022; Chen et al., 2023; Caton and Haas, 2024).

Individual fairness centers on guaranteeing that similar individuals are treated similarly, regardless of their group membership. A commonly used metric is counterfactual fairness, which ensures that an individual's outcome remains unchanged in a hypothetical scenario in which their protected attribute (e.g., gender, race) is altered (Caton and Haas, 2024; Moreau et al., 2024).

Procedural fairness, unlike group and individual fairness, focuses on the process rather than the outcome. It emphasizes transparency, explainability, and accountability in AI decision-making, ensuring that AI systems are developed and deployed in a way that upholds ethical and legal standards (Juijn et al., 2023; Genovesi et al., 2024). These definitions often conflict with each other. For example, an AI model that guarantees demographic parity may sacrifice individual fairness by giving preferential treatment to under-represented groups. The lack of a universally accepted definition of fairness makes it challenging to design AI systems that satisfy all stakeholders.

However, standardizing fairness definitions is inherently complex as each metric reflects different ethical and philosophical priorities. For example, group fairness prioritizes equity and social justice, reflecting deontological ethics that emphasize the protection of marginalized groups and rectifying historical injustice. In contrast, individual fairness is often linked to utilitarian perspectives, prioritizing overall welfare, efficiency, and predictive accuracy over group-based interventions (Verma and Rubin, 2018). These ethical tensions mean that a one-size-fits-all fairness metric is neither feasible nor desirable.

Another challenge concerns bias in data and algorithms. AI systems are trained on historical data, which often reflects societal biases (Baeza-Yates, 2016; Mehrabi et al., 2021). These biases can be explicit (e.g., discriminatory hiring practices) or implicit (e.g., under-representing certain groups in training datasets). Bias in AI can arise from multiple sources, including data collection biases (historical data may disproportionately reflect past injustices), labeling biases (human annotators may introduce subjective biases when labeling training data), and algorithmic biases (machine learning models may learn and amplify biases present in training data) (Mehrabi et al., 2021; Pessach and Shmueli, 2022). Mitigating bias in AI requires careful curation of training data and algorithmic interventions, yet it is challenging due to the deeply embedded nature of societal inequalities.

Moreover, fairness often conflicts with other principles of responsible AI, making it difficult to optimize fairness without compromising other important ethical considerations. Many AI models optimize for accuracy, but ensuring

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

fairness often requires adjusting decision thresholds, which may reduce overall predictive accuracy. For example, in predictive hiring models, adjusting for fairness may lead to hiring decisions that prioritize under-represented candidates over those with higher predicted success rates. Moreover, ensuring fairness may also compromise transparency, as transparency requires that AI decision-making processes be explainable, but some fairness-enhancing techniques, e.g., adversarial debiasing (Zhang et al., 2018; Yang et al., 2023), introduce additional complexity, making models harder to interpret. Additionally, fairness and privacy can be in tension, as ensuring fairness may require collecting sensitive demographic data (e.g., race, gender) to monitor and mitigate biases, which raises privacy concerns (Chen et al., 2023). These trade-offs highlight the difficulty of achieving fairness without compromising other ethical imperatives.

Several structural barriers contribute to the persistent difficulty of achieving fairness in AI. First, while there are multiple fairness-enhancing techniques (e.g., pre-processing, in-processing, and post-processing interventions) (Mehrabi et al., 2021), selecting the right approach depends on stakeholder priorities. However, different stakeholders, including AI developers, policymakers, businesses, and marginalized communities, may have competing interests in fairness. Companies may prioritize efficiency and profitability, whereas advocacy groups may push for equity and inclusion. The lack of consensus among stakeholders on what fairness means in specific AI applications makes it difficult to implement universally acceptable solutions. Second, AI does not operate in a vacuum; it mirrors and reinforces existing social inequalities. To achieve AI fairness, researchers must confront systemic injustices already prevalent in multiple fields, such as education, employment, healthcare, and criminal justice (Hutchinson and Mitchell, 2019). Relying solely on AI fairness interventions will not eradicate these deeply rooted disparities; comprehensive structural reforms are essential to establish equitable conditions for all groups.

1.3 How Is Fairness Applied in Practice? What Methods or Tools Are Used?

In this section, we explore the methods used to achieve AI fairness, focusing specifically on three levels: pre-processing, in-processing, and post-processing, as shown in Figure 1.1. Overall, these methods share the common goal of reducing bias in models with respect to sensitive attributes (such as gender or race), yet each approach differs in its focus, implementation, and

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

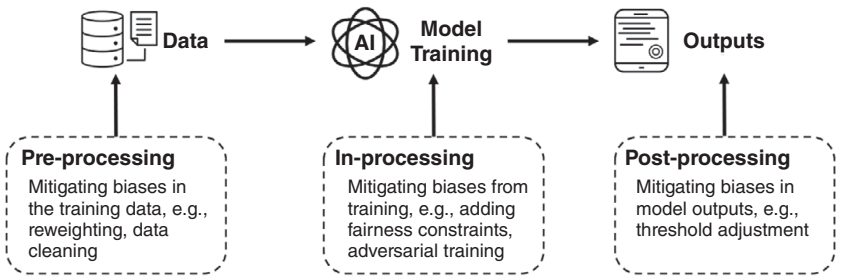


Figure 1.1 Pre-processing, in-processing, and post-processing methods

inherent limitations. The following paragraphs introduce the fundamental principles, common techniques, and concrete examples of each category and highlight the limitations encountered in practical applications, providing a comprehensive and detailed perspective to help bridge theory and practice.

Pre-processing methods primarily concentrate on adjustments made at the data level. The basic idea is to correct for biases in the training data, so that subsequent model training can be performed on a more balanced dataset. For example, reweighting techniques assign different weights to samples corresponding to various sensitive attributes to mitigate the inherent biases in the data (Kamiran and Calders, 2012; Rabonato and Berton, 2024). Although this can help correct simple disparities, such as having too few female examples in a resume screening dataset, it may not address more intricate biases hidden in the relationships between features. In practice, if the model's algorithms learn to correlate subtle patterns (e.g., certain skill keywords with particular demographic groups), simply reweighting samples may not disrupt these deep-seated, non-linear biases (subtle relationships or interactions in the data).

In addition, resampling methods, such as oversampling or undersampling, are often employed to balance the data distribution by increasing the minority group samples or reducing the majority group samples (Chawla et al., 2002; Tayebi and Garibay, 2023). While these techniques help alleviate data imbalance, oversampling can lead to overfitting of the minority class patterns, and undersampling may discard important information, thus weakening the model's generalization capability. Furthermore, some researchers propose feature transformation and debiasing techniques (e.g., a data transformation based on de-discrimination optimization goals) to reduce the impact of sensitive attributes (Feldman et al., 2015; Calmon et al., 2017). Although this method can protect data fairness to a certain degree, it may lose some valuable information during the transformation process and consequently reduce predictive performance.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

In-processing methods are directly embedded into the model training process by adjusting model parameters or introducing specific constraints to balance predictive accuracy with fairness (Liu et al., 2023). For example, fairness-constrained optimization incorporates fairness constraints directly into the objective function during training through regularization or explicit constraints to guide the learning process, ensuring the model meets additional fairness requirements while achieving original objectives (Zafar et al., 2017; Dzakpasu et al., 2024). Additionally, incorporating fairness regularization directly into the loss function has also become a common practice; this method seeks a balance between prediction error and fairness bias, yet the setting of the regularization weight is typically sensitive and often requires careful tuning depending on the dataset and task (Kamishima et al., 2011; Wan et al., 2023). Although this approach directly reduces unfairness during training, it also increases the complexity of the model training process and, in certain scenarios, may compromise prediction accuracy.

Another common technique is adversarial debiasing, which builds an adversarial learning framework that forces the main model to learn representations that do not contain sensitive information (Zhang et al., 2018; Wan et al., 2023). However, its training process often suffers from instability and complex hyperparameter tuning, and it might incur information loss that affects overall performance.

Finally, *post-processing methods* adjust the output after the model has been trained to compensate for biases arising from internal model mechanisms (Mehrabi et al., 2021). However, the limitation of post-processing methods is that they can only adjust the model's output and cannot fundamentally change the internal learning mechanisms of the model. Threshold adjustment is one commonly used technique in this category, where researchers set different decision thresholds based on the results of each group to ensure that predictive outcomes are more balanced (Wexler et al., 2019). Hardt et al. (2016) introduced the equality of opportunity and equalized odds methods as representative examples; despite their simplicity and ease of implementation, these methods can only process existing output, leaving the internal decision mechanisms untouched, and may, in some cases, reduce overall model performance. Moreover, output modification methods adjust the predicted probabilities or rankings during post-processing to meet predefined fairness criteria, though this often comes at the cost of reduced interpretability, and adjusted results might not fully reflect the true risk (Kim et al., 2019; Rabonato and Berton, 2024).

In practical applications, choosing an appropriate fairness method often depends on the specific context and requirements. In environments with large

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

and high-quality data, pre-processing methods can provide a solid fairness foundation for the model. Conversely, in emerging fields or scenarios with scarce data, incorporating fairness constraints or adversarial debiasing directly into the training process may prove more effective, as these approaches can correct biases in real-time during learning. Furthermore, in systems that are already deployed, post-processing methods are often the preferred strategy for enhancing fairness, as they require minimal modifications to the existing model and can be applied at the output stage with relative ease.

In addition, different domains and application scenarios place varying levels of emphasis on fairness. Fairness and transparency are often paramount in high-risk fields such as finance, healthcare, and the justice system; in such cases, combining multiple methods to form a multi-layered fairness assurance system is particularly appropriate. In contrast, rapidly iterating commercial applications may require a balance between fairness and efficiency. Therefore, when developing an AI fairness strategy, researchers and engineers must consider the data characteristics, the model's complexity, and the regulatory and ethical requirements of the specific industry, choosing and combining different methods flexibly to achieve the best practical outcome.

1.3.1 AI Fairness Toolkits

We introduce some commonly used open-source toolkits in this section, including their applicability and limitations (Lee and Singh, 2021; Wong et al., 2023; Sadek et al., 2024), as shown in Table 1.1. These are only a subset of fairness toolkits, and many other useful toolkits also exist.

- **IBM AI Fairness 360** (Bellamy et al., 2019): AI Fairness 360 (AIF360) is an open-source Python toolkit designed for both industrial and research settings, offering a wide range of fairness metrics and bias-mitigation algorithms that can be applied in data *pre-processing*, model training (*in-processing*), and *post-processing* stages. It helps developers and data scientists in production environments detect and reduce bias in their models. The toolkit can automatically analyze structured data, supports dozens of measurement methods (e.g., statistical fairness metrics and error-rate metrics), and integrates visualization or explainability techniques, allowing users to see at a glance whether their model treats different protected groups differently. AIF360 is well-suited to machine learning applications with higher social or business risk (such as credit assessment, hiring, or criminal justice) and provides an extensible framework for researchers and practitioners to quickly compare and evaluate new fairness algorithms. However, the toolkit

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

Table 1.1 *AI fairness toolkit features*

Toolkit	Setup	Models	Fairness Definition: Group	Fairness Definition: Individual Fairness	Fairness Definition: Procedural	Bias Mitigation
IBM AI Fairness 360	Python, R	Classification	✓	✓	\	AIF360 offers a variety of mitigation algorithms across three stages – pre-processing, in-processing, and post-processing – allowing developers to either correct the data before training or adjust the model during training or at the prediction output stage.
Microsoft Fairlearn	Python	Classification, Regression	✓	\	\	Fairlearn provides pre- and post-processing algorithms and a "reductions" framework (in-processing) that helps reduce unfairness by flexibly adjusting output thresholds or training weights, while maintaining overall model performance.
FairPY	Python	LLMs	✓	\	\	FairPY comes with multiple built-in debiasing mechanisms (such as Dropout Debias, NullSpace Projection, Self Debias, and Counterfactual Data Augmentation), which adjust for social biases in the outputs of large language models through retraining (in-processing) or post-processing techniques.
Aequitas	Python	Classification	✓	\	\	\
Google What-if Tool	TensorBoard, Colab, Jupyter Notebook	Classification, Regression	✓	✓	\	What-if Tool provides automatic threshold tuning (e.g., Demographic Parity, Equal Accuracy, Equal Opportunity), which sets different thresholds for different subgroups, adjusting decision boundaries to illustrate trade-offs in fairness (post-processing).

faces challenges in real-world deployment: its algorithms and metrics focus primarily on statistical notions of distributive fairness and are not comprehensive of more complex ethical or social considerations. Additionally, using the toolkit requires a certain level of familiarity with coding and data-processing workflows. For multi-model systems or real-time, interactive applications, extra integration with specific business logic and processes is needed to achieve fully comprehensive auditing and improvements.

- **Microsoft Fairlearn** (Bird et al., 2020): Fairlearn is an open-source toolkit that helps data scientists and developers assess and improve the fairness of their AI systems by focusing on group-level metrics such as allocation and quality-of-service harms (*pre-*, *in-*, *and postprocessing*). It provides an interactive visualization dashboard for evaluating how model predictions differ across sensitive groups (e.g., race, gender, age), along with several postprocessing and reduction algorithms for mitigating unfairness in classification and regression tasks. Fairlearn's key use cases involve any scenario in which a model's decisions (from job candidate screening to credit assessments) may disproportionately affect specific demographic groups, and it allows practitioners to compare trade-offs between fairness goals and standard performance metrics. However, it does not address all aspects of fairness – such as stereotyping, denigration, or broader societal challenges – nor does it guarantee a "fully debiased" model. Instead, Fairlearn is designed to be part of a holistic, sociotechnical approach that pairs these technical methods with a broader understanding of context, user needs, and policy considerations.
- **FairPy** (Viswanath and Zhang, 2023): FairPy is an open-source Python toolkit for evaluating and mitigating social biases in large pretrained language models. It offers: Bias Evaluation: Metrics like Hellinger distance, WEAT/SEAT, StereoSet, Honest, and log-likelihood to quantify language model bias toward protected attributes (gender, race, religion, etc.). Bias Mitigation: Techniques such as DiffPruning, Nullspace Projection, Counterfactual Data Augmentation, and Self-Debias to reduce bias, either by retraining the model (*in-processing*) on balanced data or by *postprocessing* its embeddings. Modular Design: "Plug-and-play" interfaces that work with popular transformers and user-defined models, plus support for custom corpora and social categories. While FairPy focuses on group-level associative biases, it does not cover all ethical complexities or ensure bias-free outputs in every context. Instead, it aims to be a practical starting point for diagnosing and reducing model biases under real-world conditions.
- **Aequitas** (Saleiro et al., 2019): Aequitas is an open-source bias and fairness audit toolkit that integrates various commonly used group-level fairness

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

metrics (such as distribution gaps and error-rate disparities) to assess performance differences across protected groups. It is available as a Python library, command-line interface, or web application, meeting both the quick bias analysis needs of data scientists during model development and the visual, user-friendly audit requirements of policymakers or non-technical stakeholders before model deployment. By focusing on group-level bias detection, Aequitas provides decision-makers with a set of statistics-based methods to answer core questions like which groups may be adversely impacted, whether such differences are statistically significant, and how the model's predictions are distributed across different populations. It is particularly well-suited for risk assessment and intervention scenarios in public policy or social service contexts such as criminal justice, public health, and social welfare. However, Aequitas does not directly cover fine-grained standards of fairness (such as individual-level fairness), and for interactive or multi-model applications, it needs to be combined with other tools and domain-specific logic to achieve a more comprehensive auditing and improvement plan.

- **Google What-if Tool** (Wexler et al., 2019): The What-If Tool (WIT) is an open-source visualization toolkit that requires no additional coding, enabling developers to analyze and debug machine learning models while also evaluating their performance under various fairness scenarios. WIT allows users to flexibly slice datasets based on different features and subgroups (e.g., gender, race) and observe fairness metrics such as prediction accuracy, false positive rate, and false negative rate across those groups in real time. It also supports independently setting classification thresholds for binary classification models to satisfy common fairness definitions like demographic parity and equal opportunity (*postprocessing*). This makes WIT especially useful for quickly comparing multiple models or testing different fairness strategies in practical settings. However, the tool also has limitations: first, WIT operates as a "black-box" analyzer, meaning it does not provide access to internal model structures. Second, it relies on user-provided data and features – if the data lacks representativeness or is biased, fairness evaluations may be compromised. Lastly, when dealing with large-scale datasets or high-dimensional features, WIT's visualization and real-time interaction capabilities may face performance challenges.

Reaching AI fairness is not just a technical challenge or engineering task; it is an ongoing process that integrates social responsibility, regulatory compliance, and technological innovation. By employing appropriate interventions at each stage (before, during, and after model training), together with the help of

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

specialized toolkits, one can significantly improve the inclusivity and equity of AI systems. Different organizations are already exploring domain-specific solutions, continuously refining assumptions and approaches to ensure that AI better serves the diverse needs of society. Only by maintaining humility in the face of data and sensitivity to bias can we enable AI to bring genuine human-centered value to healthcare, education, finance, justice, and beyond.

1.4 What Is an Example of Fairness in AI Done Well? What Is an Example of Fairness in AI Gone Wrong?

As AI technology proliferates in many fields, ensuring model fairness and preventing algorithmic discrimination have become critical social concerns. On the surface, AI decisions are based on "objective" data and complex algorithms, yet much of the data can carry existing social biases. Once implemented on a scale in automated decision-making, these biases can exacerbate the inequitable treatment of certain groups.

- In *healthcare*, if the training data come primarily from economically developed regions or a single demographic population, a diagnostic model may struggle to accurately identify conditions in patients from other regions or ethnic backgrounds, which can lead to serious misdiagnoses (Li et al., 2023).
- In *human resources*, automated hiring systems that rely on historical data, where certain genders or educational backgrounds were favored, may inadvertently learn and reinforce these biases (Köchling and Wehner, 2020; Quiñonero-Candela et al., 2023).
- In *financial services*, credit assessment and risk control models are particularly susceptible to social biases. For example, if certain groups have historically weaker credit scores, machine learning models can automatically classify them as high-risk, dramatically increasing denial rates (Stevens et al., 2020).
- In *public safety and criminal justice*, predictive tools are used to gauge the likelihood of recidivism. However, if training data reflect differing arrest patterns or uneven intensity of policing for certain regions or demographic groups, the model may interpret these as indicators of criminality (Dressel and Farid, 2018).
- In *education*, there have been experimental AI applications for admissions, course selection, and prediction of academic performance. If the training set includes predominantly students from prestigious high schools, wealthy

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

families, or major cities, the system can fail to account for students from rural areas or lower-income households (Marcinkowski et al., 2020).

Next, we will explore several real-world cases to illustrate both the successes and failures of AI fairness in practice, as well as the challenges and potential solutions.

1.4.1 Example of AI Fairness Done Well

The following cases can be characterized as examples of AI systems working well in fairness. These examples share several characteristics: (1) fairness was explicitly treated as a design objective, (2) interdisciplinary collaboration was involved, and (3) systems were evaluated using real-world data with fairness metrics. They demonstrate how fairness can be incorporated at both technical and organizational levels.

Promoting Equitable Ventilator Allocation through AI

A notable example of fairness for AI comes from the context of healthcare resource allocation, particularly during crises when resources such as ventilators were in short supply. In a collaborative effort between researchers at Princeton University and Northwestern University, the study proposed a deep reinforcement learning approach to optimize ventilator allocation for critically ill patients (Li et al., 2024). Their system leverages an advanced transformer-based architecture to comprehensively analyze patients' health trajectories and the interactions between individuals. Most importantly, fairness is explicitly incorporated as a key objective in the decision-making process. Compared to traditional rule-based or score-based protocols, this AI model can automatically learn more reasonable and human-centered allocation strategies while reducing biases related to age, race, or other demographic factors.

Beyond its technical innovation, the study also had a significant social impact. When evaluating real-world clinical data, the AI system improved overall patient survival rates and substantially reduced disparities in resource access among different demographic groups. For example, while some existing protocols tended to favor white or younger patients, the new method of Li et al. (2024) significantly improved resource accessibility for under-represented minorities. By training the model with fairness as a core value, the approach helps address long-standing structural inequities in healthcare systems and improves access for vulnerable populations. This research highlights the potential of AI to tackle real-world societal challenges and offers an inspiring model for building fair and sustainable service allocation systems.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

LinkedIn's AI Fairness Framework

LinkedIn, a professional networking platform, not only connects hundreds of millions of professionals worldwide but also plays a key role in hiring, career development, and knowledge sharing. As the platform has grown, the algorithms behind it have become increasingly important in shaping users' access to opportunities. Without careful design and oversight, these algorithms risk reinforcing existing social inequalities, making it more difficult for minority groups to access professional or networking opportunities. To address this challenge, LinkedIn has made AI fairness one of its core long-term product principles and developed a systematic fairness framework (Quiñonero-Candela et al., 2023). This framework is built on two key pillars: "equal treatment" and a "holistic fairness and inclusion strategy," aiming to reduce algorithmic bias across groups while preserving privacy.

This framework has been successfully applied to several features of the platform. For example, in the "People Also Viewed" feature, LinkedIn found that women users with comparable levels of activity were receiving significantly lower model scores than their men counterparts, resulting in less visibility in recommendations. The team first attempted to address this gap by refining model architecture and feature engineering. When these standard approaches fell short, they cautiously introduced gender as a post-processing signal to fine-tune the outcomes. The results showed that this strategy effectively reduced gender disparities and improved overall connection success rates. This case demonstrates LinkedIn's practical implementation of AI fairness principles and highlights how technology can support greater diversity and reduce structural inequalities. Through continuous monitoring and cross-functional collaboration, LinkedIn is steadily building a more inclusive and equitable professional platform. Its approach also offers a replicable model for other large-scale platforms and tech companies, helping to advance fairness in algorithmic systems across the industry.

1.4.2 Example of AI Fairness Gone Wrong

The following cases illustrate the consequences of neglecting fairness considerations in high-stakes applications. Common characteristics include: (1) reliance on biased or unrepresentative data, (2) lack of transparency or accountability mechanisms, and (3) failure to anticipate a disparate impact on marginalized groups. These cases underscore how algorithmic systems, when implemented without fairness safeguards, can perpetuate or even exacerbate existing inequalities.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

COMPAS in Criminal Justice

One of the most infamous examples of AI fairness failure is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) risk assessment tool (Northpointe, 2012), used in the US criminal justice system to predict recidivism (the likelihood of an offender reoffending). An investigation by ProPublica found that COMPAS disproportionately labeled black defendants as high-risk, even when their actual recidivism rates were lower than white defendants who were labeled low-risk (Angwin et al., 2022). The algorithm also induced the opposite error, significantly underestimating future crimes among Caucasians (false-negative rate; FNR). The proprietary nature of the algorithm made it difficult to scrutinize or challenge its predictions. The algorithm prioritized predictive accuracy over fairness, resulting in systematically unfair results. COMPAS illustrates how AI fairness failures can reinforce systemic discrimination, particularly when AI is used in high-stakes decision-making without transparency and rigorous fairness testing.

Algorithmic Bias in the Housing Rental Market

SafeRent is an AI-driven service that is used in the US housing rental market to assess the risk and creditworthiness of the tenant. On the surface, it offers efficiency to landlords by automating the screening of applications through a "safety score." However, increasing scrutiny has revealed that this seemingly neutral system may embed hidden biases (Bhuiyan, 2024). Applicants from marginalized groups, such as people of color or low-income backgrounds, reported being unfairly scored, even when they had stable jobs and a good credit history. The system's dependence on factors such as ZIP code, past housing records, and financial behavior, all of which can correlate with race and social class in the United States, means that it can systematically disadvantage vulnerable populations. Applicants had no way to understand, challenge, or appeal their scores, as the scoring process operates as a "black box" with little transparency or recourse.

The case of SafeRent has triggered broader concerns about fairness, transparency, and accountability in AI systems used for critical life decisions. As algorithms increasingly shape access to housing, jobs, healthcare, and education, even subtle biases can reinforce and magnify structural inequalities. In SafeRent's case, affected individuals may face repeated rejections, limiting their access to better neighborhoods, schools, or opportunities, and contributing to long-term social stratification. Public backlash sparked renewed calls for algorithmic transparency, asking who is impacted, how decisions are made, and what safeguards exist. It also accelerated regulatory conversations around

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

AI governance. The SafeRent example underscores a key lesson: Without proactive fairness measures and mechanisms for accountability, AI can easily become a gatekeeper that deepens inequality, rather than a tool that helps reduce it.

Lessons Learned

Together, these examples illustrate the dual potential of AI in shaping a more equitable or more divided society. When fairness is treated as a core design goal, as seen in ventilator allocation and LinkedIn cases, AI can actively help correct systemic imbalances and improve access for historically disadvantaged groups. On the other hand, when fairness is overlooked or sacrificed for efficiency, as demonstrated by the COMPAS and SafeRent examples, AI systems risk entrenching and amplifying existing inequalities, often without visibility or recourse for those affected. These cases underscore the importance of embedding fairness principles throughout the lifecycle of AI development, from data collection and model design to deployment and monitoring. They also highlight the urgent need for governance policies that ensure transparency, accountability, and redress. Ultimately, advancing AI fairness is not only a technical challenge but a social imperative – one that requires interdisciplinary collaboration, ongoing reflection, and a commitment to equity in both intention and outcome.

1.5 What Is Missing to Make Fairness Work in AI? What Needs to Change?

Achieving fairness in AI requires systemic changes in how AI is designed, developed, deployed, and regulated. Although progress has been made, several critical gaps remain. To make fairness work in AI, we need improvements in regulation, transparency, accountability, and interdisciplinary collaboration.

1. **Contextualized Fairness Definitions and Metrics:** A key challenge in AI fairness is the lack of universally accepted definitions and metrics, leading to inconsistencies across industries. Standardizing fairness is difficult because different metrics reflect competing ethical and philosophical priorities (Binns, 2021). A one-size-fits-all approach is neither feasible nor desirable, as fairness must be evaluated within its specific social and ethical context. For example, fairness in criminal justice may require addressing structural racism and historical over-policing, while fairness in credit lending may need to navigate between equitable access and financial risk

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

minimization. To address this, scholars increasingly call for contextualized, domain-specific fairness frameworks (Holstein et al., 2019; Mitchell et al., 2021). These frameworks emphasize that fairness must be defined and operationalized in ways that are sensitive to the domain of application, the impacted communities, and the goals and constraints of the specific AI system. In this view, fairness is not simply a technical property of algorithms but a normative judgment that must be negotiated among stakeholders.

2. **Intersectional Fairness Models:** Moreover, bias is often subtle, multidimensional, and intersectional, affecting individuals based on multiple overlapping attributes rather than single categories, which standard detection methods may miss (Crenshaw, 1989; Kearns et al., 2019). AI models should be evaluated not just on single demographic attributes but on combinations of attributes (e.g., Black women, Latino individuals with disabilities, etc.), such as the seminal paper on Gender Shades (Boulamwini and Gebru, 2018). Fairness audits should incorporate real-world intersectional case studies rather than relying on abstract fairness metrics alone. Conducting multi-attribute bias testing has the potential to uncover such hidden disparities before deployment. Adding to this challenge, most fairness metrics assume that people can be neatly categorized into fixed groups (e.g., male/female, white/non-white) (Barocas et al., 2023). Although these simplified groupings make mathematical modeling more tractable, they do not reflect the complexity and fluidity of real-world identities. In reality, sensitive attributes can be non-binary (e.g., gender diversity beyond male/female), context-dependent (e.g., fairness in hiring may differ from that in healthcare), or interconnected (e.g., racial bias may intersect with socioeconomic status). Current algorithms struggle to handle this complexity, and thus may produce outputs that appear fair when evaluated using standard metrics, but nonetheless entrench or exacerbate existing inequalities (Hanna et al., 2020). Addressing these limitations requires a shift toward more flexible, adaptive, and socially informed approaches to fairness, including: (1) multi-attribute fairness auditing that examines intersections of multiple identity categories; (2) causal modeling techniques to differentiate between correlation and structural bias; and (3) participatory design processes that include input from affected communities to define what fairness means in a given context. We argue that moving beyond fixed-group assumptions is essential for building AI systems that are genuinely inclusive and aligned with the complexities of human identity and lived experience.
3. **More Transparent and Interpretable AI Models:** A significant barrier to achieving fairness in AI lies in the opacity of machine learning models,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

particularly black-box models such as deep neural networks (Pasquale, 2015; Burrell, 2016). Without transparency, it is difficult to detect or correct unfair biases. The rise of explainable AI (XAI) aims to address this issue by designing methods that make the behavior of the model more transparent, interpretable, and accountable (Hoffman et al., 2018; Molnar, 2020; Dwivedi et al., 2023). Interpretability is particularly crucial in high-stakes domains like criminal justice, finance, and healthcare, where opaque decisions can have life-changing consequences. Researchers have developed a variety of XAI techniques, including feature attribution methods (e.g., SHAP, LIME), counterfactual explanations, and rule-based models that provide simplified representations of complex models (Molnar, 2020). These tools help uncover how models use input to generate predictions and can highlight whether unfair biases are influencing outcomes. Developers should also integrate open-source fairness and interpretability tools into their workflows, such as IBM's AI Fairness 360, Google's What-If Tool, and Microsoft's Fairlearn. However, these tools remain underutilized in real-world deployments, often seen as add-ons rather than integral parts of the AI lifecycle. For fairness and transparency to become standard practice, such tools must be incorporated more seamlessly into model development, testing, and monitoring pipelines.

4. **Stronger Regulations and Accountability Mechanisms:** Even as AI's societal impact grows, formal regulations specifically addressing AI fairness are in their infancy. ISO 24028:2020 covers AI fairness as part of a broader effort to guide responsible development (Constantinides et al., 2024). However, few jurisdictions have comprehensive AI fairness laws. The European Union is pioneering an AI Act with a risk-based framework (banning some harmful uses and imposing strict rules on high-risk AI) (Petro, 2024). In contrast, the United States has no binding federal AI law, relying instead on a patchwork of existing statutes (e.g., anti-discrimination laws like the Fair Housing Act or Equal Credit Opportunity Act) that only scratch the surface of AI's unique risks. This lack of overarching regulation creates compliance gaps and uncertainty. Without clear policies, many companies deprioritize fairness in favor of speed, efficiency, or profit. Stronger regulations are needed to ensure that AI fairness is not just a recommendation but a legal requirement. Governments must introduce AI fairness laws that mandate fairness audits, bias testing, and transparency reports. Organizations should establish AI ethics committees that hold developers accountable for fairness issues. AI fairness violations should carry legal and financial consequences, similar to data privacy violations under the GDPR law.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

5. **Interdisciplinary Collaboration:** AI fairness is not solely a technical issue; it lies at the intersection of technology, law, sociology, and philosophy. Yet, AI development is often dominated by engineers and data scientists, without input from domain experts. Therefore, cross-disciplinary efforts are essential to make real progress. Teams that address AI fairness should include not only data scientists but also lawyers, ethicists, social scientists, and domain experts relevant to AI applications. Governments and NGOs can facilitate interdisciplinary forums and working groups. For example, the EU's High-Level Expert Group on AI included legal scholars, philosophers, and industry practitioners who jointly formulated ethical AI guidelines (European Commission, 2024). Universities and AI research institutes should offer interdisciplinary training on AI fairness, ensuring that developers understand both the technical and ethical dimensions. Companies that deploy AI in high-risk areas (e.g., hiring, healthcare) should partner with civil rights organizations and advocacy groups to address fairness concerns. These collaborations can generate holistic solutions to improve AI fairness.

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. 2022. Machine bias. In K. Martin (ed.), *Ethics of Data and Analytics* (New York: Auerbach Publications), pp. 254–264.

Baeza-Yates, R. 2016. Data and algorithmic bias in the web. *Proceedings of the 8th ACM Conference on Web Science*. (Hannover: Association for Computing Machinery), p. 1.

Barocas, S., Hardt, M., and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities* (Cambridge, MA: MIT Press).

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., et al. 2019. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *IBM Journal of Research and Development*, **63**(4/5), 4:1–4:15.

Bhuiyan, J. 2024 (Dec.). She didn't get an apartment because of an AI-generated score – and sued to help others avoid the same fate. *The Guardian*.

Binns, R. 2021 (Mar.). Fairness in machine learning: Lessons from political philosophy. *arXiv:1712.03586* [cs].

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., et al. 2020 (May). Fairlearn: A toolkit for assessing and improving fairness in AI. *Tech. rept. MSR-TR-2020-32*. Microsoft.

Boulamwini, J., and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, **81**, 77–91.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

Burrell, J. 2016. How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society*, **3**(1), 2053951715622512.

Calmon, F. P, Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. 2017. Optimized pre-processing for discrimination prevention. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (New York: Association for Computing Machinery).

Caton, S., and Haas, C. 2024. Fairness in machine learning: A survey. *ACM Computing Surveys*, **56**(7), 1–38.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, **16**(June), 321–357.

Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Lu, M. Y., et al. 2023. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, **7**(6), 719–742.

Constantinides, M., Bogucka, E., Quercia, D., Kallio, S., and Tahaei, M. 2024. RAI guidelines: Method for generating responsible AI guidelines grounded in regulations and usable by (non-) technical roles. *Proceedings of the ACM on Human-Computer Interaction*, **8**(CSCW2), 1–28.

Crenshaw, K. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, **1989**(1), article 8. http://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8In (last accessed February 17, 2026).

Dressel, J., and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, **4**(1), eaao5580.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., et al. 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, **55**(9), 1–33.

Dzakpasu, D. Q., Liu, J., Li, J., and Liu, L. 2024. Integrating fair representation learning with fairness regularization for intersectional group fairness. *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Association for Computing Machinery), pp. 560–569.

European Commission. 2024 (Oct.). *High-level expert group on artificial intelligence*.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney: Association for Computing Machinery), pp. 259–268.

Genovesi, S., Mönig, J. M., Schmitz, A., Poretschkin, M., Akila, M., Kahdan, M., et al. 2024. Standardizing fairness-evaluation procedures: Interdisciplinary insights on machine learning algorithms in creditworthiness assessments for small personal loans. *AI and Ethics*, **4**(2), 537–553.

Hanna, A., Denton, E., Smart, A., and Smith-Loud, J. 2020. Towards a critical race methodology in algorithmic fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Springer), pp. 501–512.

Hardt, M., Price, E., and Srebro, N. 2016 (Oct.). Equality of opportunity in supervised learning. *arXiv:1610.02413* [cs].

Harper, S. B., and Weber, E. S. 2022. Fiduciary responsibility: Facilitating public trust in automated decision making. *Journal of Social Computing*, **3**(4), 345–362.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

Hemamou, L., Guillon, A., Martin, J.-C., and Clavel, C. 2021. Don't judge me by my face: An indirect adversarial approach to remove sensitive information from multimodal neural representation in asynchronous job video interviews. *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)* (New York: IEEE), pp. 1–8.

Hoffman, R. R, Mueller, S. T, Klein, G., and Litman, J. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Holstein, K., Wortman J. V., Daumé, H., Dudik, M., and Wallach, H. 2019. Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow: Association for Computing Machinery), pp. 1–16.

Hutchinson, B., and Mitchell, M.. 2019. 50 Years of test (un)fairness: Lessons for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT*'19* (New York: Association for Computing Machinery), pp. 49–58.

Juijn, G., Stoimenova, N., Reis, J., and Nguyen, D. 2023. Perceived algorithmic fairness using organizational justice theory: An empirical case study on algorithmic hiring. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal: Association for Computing Machinery), pp. 775–785.

Kamiran, F., and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, **33**(1), 1–33.

Kamishima, T., Akaho, S., and Sakuma, J. 2011. Fairness-aware learning through regularization approach. *2011 IEEE 11th International Conference on Data Mining Workshops* (Vancouver: IEEE), pp. 643–650.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. 2019. An empirical study of rich subgroup fairness for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery), pp. 100–109.

Kim, M. P., Ghorbani, A., and Zou, J. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu: Association for Computing Machinery), pp. 247–254.

Köchling, A., and Wehner, M. C. 2020. Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, **13**(3), 795–848.

Lee, M. S. A., and Singh, J. 2021. The landscape and gaps in open source fairness toolkits. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama: Association for Computing Machinery), pp. 1–13.

von Leyden, W. 1985. *Aristotle on Equality and Justice: His Political Argument* (London: Palgrave Macmillan).

Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E., and Mohr, D. C. 2023. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *NPJ Digital Medicine*, **6**(1), 236.

Li, Y., Mao, C., Huang, K., Wang, H., Yu, Z., Wang, M., et al., 2024 (Aug.). Deep reinforcement learning for efficient and fair allocation of health care resources. *arXiv:2309.08560* [cs].

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

Liu, M., Ning, Y., Teixayavong, S., Mertens, M., Xu, J., Ting, D. S. W., et al. 2023. A translational perspective towards clinical AI fairness. *NPJ Digital Medicine*, **6**(1), 172.

Marcinkowski, F., Kieslich, K., Starke, C., and Lünich, M.. 2020. Implications of AI (un-)fairness in higher education admissions: The effects of perceived AI (un-) fairness on exit, voice and organizational reputation. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona: Association for Computing Machinery), pp. 122–130.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, **54**(6), 115:1–115:35.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, **8**(1), 141–163.

Molnar, C. 2020. Interpretable machine learning. Lulu.com.

Moreau, T., Sinatra, R., and Sekara, V. 2024. Failing our youngest: On the biases, pitfalls, and risks in a decision support algorithm used for child protection. *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro: Association for Computing Machinery), pp. 290–300.

Mulligan, D. K., Kroll, J. A., Kohli, N., and Wong, R. Y. 2019. This thing called fairness: Disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human–Computer Interaction*, **3**(CSCW), 119:1–119:36.

Narayanan, A. 2018. Fairness definitions and their politics. *Tutorial presented at the Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery).

Northpointe. 2012. *Practitioners Guide to COMPAS* (Northpointe).

Nozick, R. 1974. *Anarchy, State, and Utopia*. (Hoboken, NJ: John Wiley & Sons).

Pasquale, F. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, MA: Harvard University Press).

Pessach, D., and Shmueli, E. 2022. A review on fairness in machine learning. *ACM Computing Surveys*, **55**(3), 51:1–51:44.

Petro, T. 2024 (Mar.). The risks of fragmented privacy and AI regulations. *Directors & Boards Newsletter*.

Pombal, J., Cruz, A. F, Bravo, J., Saleiro, P., Figueiredo, M. A. T., and Bizarro, P. 2022. Understanding unfairness in fraud detection through model and data bias interactions. *arXiv preprint arXiv:2207.06273*.

Quiñonero-Candela, J., Wu, Y., Hsu, B., Jain, S., Ramos, J., Adams, J., et al. 2023 (June). Disentangling and operationalizing AI fairness at LinkedIn. *2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1213–1228. arXiv:2306.00025 [cs].

Rabonato, R. T., and Berton, L. 2024. A systematic review of fairness in machine learning. *AI and Ethics*, **5**, 1–12.

Rawls, J. 1991. Justice as fairness: Political not metaphysical. In J. A. Corlett (ed.), *Equality and Liberty* (London: Palgrave Macmillan), pp. 145–173.

Sadek, M., Constantinides, M., Quercia, D., and Mougenot, C. 2024. Guidelines for integrating value sensitive design in responsible AI toolkits. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (New York: Association for Computing Machinery), pp. 1–20.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., et al. 2019 (Apr.). Aequitas: A bias and fairness audit toolkit. *arXiv:1811.05577* [cs].

Stevens, A., Deruyck, P., van Veldhoven, Z., and Vanthienen, J. 2020. Explainability and fairness in machine learning: Improve fair end-to-end lending for Kiva. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. (Canberra: IEEE), pp. 1241–1248.

Tayebi, A., and Garibay, O. O. 2023. Improving fairness via deep ensemble framework using preprocessing interventions. *International Conference on Human-Computer Interaction* (Cham: Springer), pp. 477–489.

Teo, S. A. 2024. Artificial intelligence and its "slow violence"to human rights. *AI and Ethics*, **5**, 1–16.

Verma, S., and Rubin, J. 2018. Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness. FairWare '18* (New York: Association for Computing Machinery), pp. 1–7.

Viswanath, H., and Zhang, T. 2023 (Feb.). FairPy: A toolkit for evaluation of social biases and their mitigation in large language models. *arXiv:2302.05508* [cs].

Wan, M., Zha, D., Liu, N., and Zou, N. 2023. In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data*, **17**(3), 1–27.

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viegas, F., and Wilson, J. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 1–1. arXiv:1907.04135 [cs].

Wong, R. Y, Madaio, M. A, and Merrill, N. 2023. Seeing like a toolkit: How toolkits envision the work of AI ethics. *Proceedings of the ACM on Human-Computer Interaction*, **7**(CSCW1), 1–27.

Yang, J., Soltan, A. A. S., Eyre, D. W., Yang, Y., and Clifton, D. A. 2023. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digital Medicine*, **6**(1), 55.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *Proceedings of the 26th International Conference on World Wide Web* (Perth: International World Wide Web Conferences Steering Committee), pp. 1171–1180.

Zhang, B. H., Lemoine, B., and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans: Association for Computing Machinery), pp. 335–340.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 01:13:12, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009625180.002