E-ISSN: 3107-507X
JOURNAL OF SCIENTIFIC INNOVATION AND ADVANCED RESEARCH (JSIAR), VOLUME 1, ISSUE 5, AUGUST 2025 https://jsiar.com
editor@jsiar.com

Machine Minds, Human Echoes: Investigating Inherited AI Bias

Karan Singh
Department of Information Technology
Noida Institute of Engineering and Technology, Greater Noida, India
Email: karan.singh@niet.co.in

***Abstract*—Artificial Intelligence (AI) systems are increasingly becoming integral to decision-making processes in critical domains such as healthcare, finance, criminal justice, and recruitment. While these systems offer remarkable capabilities, they often reflect the imperfections of the data they are trained on. This paper investigates the concept of *inherited AI bias*, a phenomenon where machine learning models unintentionally assimilate and reproduce societal prejudices, stereotypes, and historical inequities embedded in human-generated datasets. We provide a comprehensive review of the root causes of such bias, including imbalanced training data, biased annotation practices, and algorithmic structures that lack fairness constraints. By examining real-world case studies and empirical research, we demonstrate how these biases disproportionately impact marginalized groups, leading to discriminatory outcomes and reinforcing systemic disparities. The paper also reviews state-of-the-art techniques for bias detection and mitigation, critically assessing their strengths and limitations in practical applications. As AI systems continue to permeate our daily lives, addressing inherited bias is not merely a technical challenge but an ethical imperative. This review underscores the urgency of developing transparent, inclusive, and accountable AI frameworks. Finally, we identify current gaps in research and propose directions for future work, aimed at fostering the development of equitable AI systems that align with democratic values and social justice.**

***Keywords*—Artificial Intelligence Ethics, Algorithmic Bias, Fairness in Machine Learning, Bias Mitigation Techniques, Societal Impact of AI, Responsible AI Systems**

I. INTRODUCTION

Artificial Intelligence (AI) has evolved from a niche academic field into a transformative force shaping modern society. Today, AI-driven systems are embedded in critical decision-making infrastructures across sectors such as healthcare, finance, law enforcement, education, and recruitment [1]–[3]. These "machine minds" are tasked with making judgments that were once solely in the hands of humans. However, as these systems increasingly influence real-world outcomes, concerns have emerged regarding their fairness, transparency, and accountability.

One of the most pressing challenges in deploying AI systems is the emergence of algorithmic bias — systemic and unfair discrimination resulting from patterns in training data, model design, or deployment context [21], [62]. This bias is not the result of malicious intent but rather a reflection of the imperfections and inequities embedded in historical and societal data. The phenomenon of *inherited AI bias* arises when machine learning algorithms absorb and perpetuate these existing human prejudices, leading to unfair or even harmful outcomes [34], [53], [75].

The motivation behind this study stems from a growing number of real-world incidents that highlight the tangible consequences of biased AI systems. For example, facial recognition algorithms have shown significantly higher error rates for women and individuals with darker skin tones compared to white males [49], [75]. In one notable case, the COMPAS algorithm used in U.S. courts for risk assessment was found to disproportionately classify African-American defendants as high-risk compared to white defendants with similar profiles [59]. Similarly, a recruiting tool developed by Amazon was discontinued after it was found to systematically downgrade resumes containing the word “women’s” or affiliated with female-associated activities [52].

These issues raise fundamental questions about the reliability and ethical acceptability of AI systems. If unchecked, inherited AI bias has the potential to reinforce existing inequalities, erode public trust, and widen socio-economic divides. As AI continues to permeate vital aspects of life, addressing these concerns becomes not only a technical obligation but also a moral one [12], [39], [61].

The objective of this review paper is to systematically explore the phenomenon of inherited AI bias. We aim to identify and analyze the underlying sources of bias, review major real-world incidents, and evaluate existing detection and mitigation strategies. Furthermore, the paper seeks to promote awareness about the ethical, legal, and societal implications of AI bias and highlight the urgent need for fair, transparent, and accountable AI systems. Ultimately, this study aspires to contribute toward building AI technologies that not only replicate intelligence but also uphold the principles of justice and equity.

II. UNDERSTANDING AI BIAS

Artificial Intelligence (AI) systems, while increasingly pervasive, are not free from imperfections. These imperfections often manifest as bias — a situation in which AI models produce outcomes that systematically favor certain groups or attributes over others. Bias in AI can be broadly categorized into several types, each with its distinct characteristics and impact. In this section, we discuss the various types of bias, their sources, and the definitions and taxonomy from existing literature.

A. Types of Bias

AI bias typically falls into four primary categories: data bias, algorithmic bias, labeling bias, and societal bias [61], [62]. Each type originates from different stages of the AI

TABLE I: Selected Real-World Examples of Inherited AI Bias

System/Application	Bias Identified	Reference
Facial Recognition (IBM, Microsoft, Amazon)	Higher error rates for darker-skinned females	[75]
COMPAS Risk Assessment Tool	Racial bias in predicting recidivism	[59]
Amazon Hiring Tool	Penalized female-associated resumes	[52]
Google Translate	Gender stereotypes in pronoun translation	[35]
ImageNet Classification	Cultural and racial labeling errors	[16]

development process, yet they share a common theme of unfairly influencing the decisions made by AI systems.

1) Data Bias: Data bias arises when the dataset used to train an AI model is not representative of the real-world population or problem it is intended to solve. This can occur if certain groups or attributes are overrepresented or underrepresented in the dataset. For example, facial recognition systems trained predominantly on lighter-skinned individuals may fail to accurately detect darker-skinned faces [75]. Such biases can severely impair the fairness and efficacy of AI models.

2) Algorithmic Bias: Algorithmic bias refers to the unintended consequences of the algorithms themselves. Even if the training data is balanced, biases can emerge if the algorithm favors certain patterns or outcomes over others due to its design. For instance, if an algorithm uses proxies that are correlated with biased data (e.g., using zip codes as a proxy for race in a predictive policing model), it can perpetuate discriminatory outcomes [52], [62].

3) Labeling Bias: Labeling bias occurs during the data annotation process, where human annotators unintentionally introduce their own biases while labeling data. These biases may reflect societal stereotypes or assumptions. For example, in a dataset used to train an AI model for gender classification, the labels assigned by annotators may reflect gender stereotypes that associate certain activities with specific genders [34], [53].

4) Societal Bias: Societal bias is the reflection of the inequalities and prejudices present in the society from which the data is sourced. AI systems often inherit these biases, especially when trained on historical data that embodies existing societal inequalities. For example, a predictive model used in criminal justice may replicate historical racial disparities if trained on biased arrest or sentencing data [39], [59].

B. Sources of Bias

Understanding the sources of AI bias is crucial for addressing it. Bias can be introduced at various stages of the AI development process, from dataset collection to algorithm design.

1) Biased Training Datasets: The primary source of AI bias is often the training dataset itself. Datasets are often collected from historical data or from web-scraped content that reflects existing biases. If the dataset is not diverse or inclusive, the resulting AI model may inherit and amplify these biases. For example, the 2018 Amazon AI recruitment tool faced criticism for discriminating against women, as it was trained on resumes submitted to the company over a 10-year period, a dataset that was predominantly male [52].

2) Human-Labeled Data with Stereotypes: Human annotators play a key role in data labeling, but their own implicit biases can be reflected in the labeled data. These biases may be conscious or unconscious, and they are often shaped by societal norms and stereotypes. For example, the gender bias in job titles or clothing in labeled datasets for image classification tasks reflects these stereotypes [35], [53].

3) Implicit Developer Assumptions: Bias can also be introduced through the implicit assumptions made by developers during model design. These assumptions can influence the choice of features, the design of the model, or the metrics used to evaluate model performance. For instance, developers might choose to optimize for accuracy without considering how the model's performance might vary across different demographic groups, inadvertently favoring one group over others [39], [51].

C. Definitions and Taxonomy from Literature

The literature on AI bias presents a variety of definitions and taxonomies. Barocas et al. [62] provide a comprehensive taxonomy, categorizing bias into different types, including *prejudicial bias* (arising from discriminatory data) and *procedural bias* (caused by flawed decision-making processes in model development). Similarly, O'Neil [61] discusses the concept of "Weapons of Math Destruction," where biased algorithms with high societal impact reinforce existing inequalities.

To address bias in AI systems, it is essential to clearly define the types and sources of bias to ensure that corrective measures are targeted effectively. Understanding these factors enables researchers and practitioners to design AI systems that are more equitable, transparent, and socially responsible.

TABLE II: Types and Sources of AI Bias

Type of Bias	Source of Bias
Data Bias	Biased training datasets
Algorithmic Bias	Algorithm design and proxies
Labeling Bias	Human-labeled data with stereotypes
Societal Bias	Historical and societal inequalities

III. Mechanisms of Inheritance: How Human Bias Enters Machine Minds

Artificial Intelligence (AI) systems are designed to mimic human cognitive processes, learning from vast amounts of data

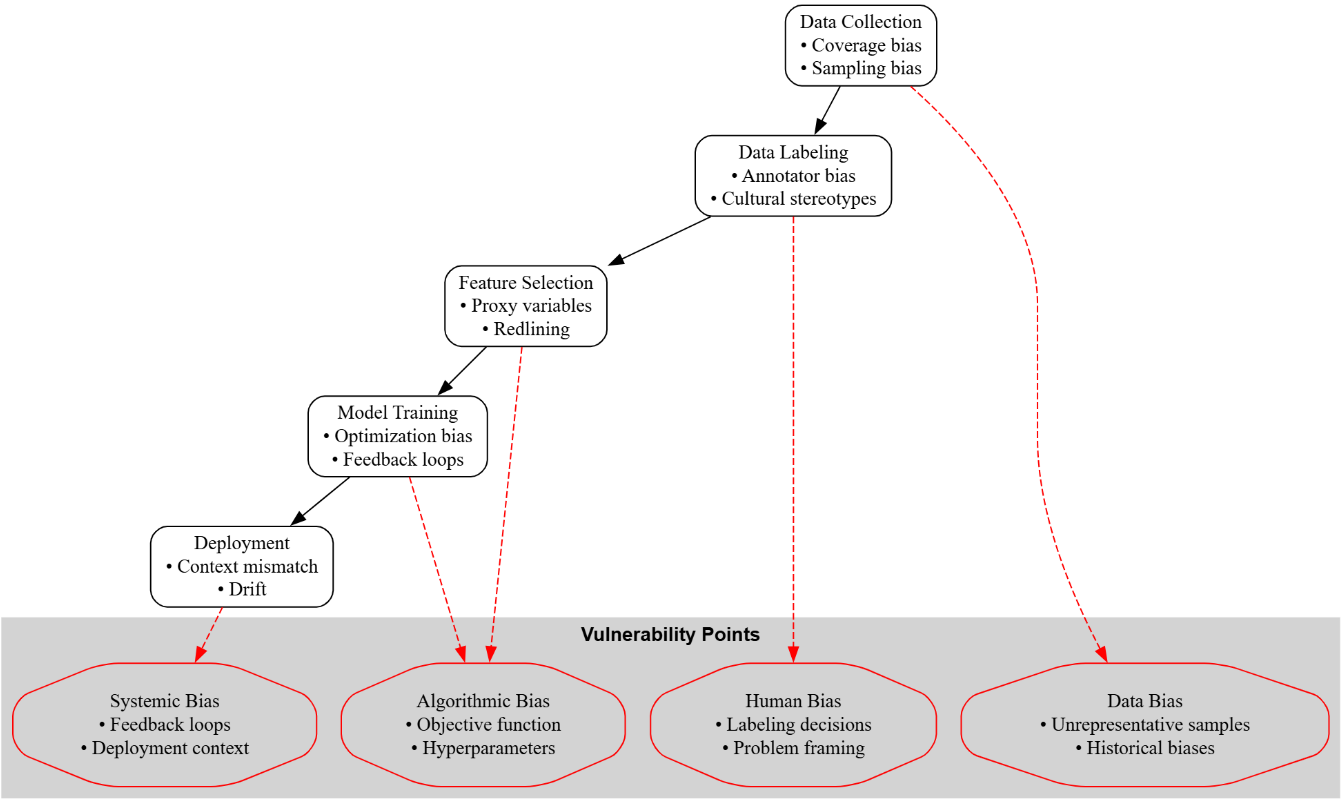

Fig. 1: The AI Training Pipeline and its Vulnerabilities

to make decisions. However, when these systems are trained on data that reflects societal biases, the machine inherits these biases, amplifying and perpetuating them. The process by which these biases enter machine minds is not instantaneous; rather, it occurs through various stages in the AI development pipeline. In this section, we explore how bias propagates across different AI domains, including Natural Language Processing (NLP), Computer Vision, and decision-making systems. Additionally, we discuss the concept of bias feedback loops and compounding effects, which can exacerbate the problem.

A. The AI Training Pipeline and its Vulnerabilities

The AI training pipeline consists of several stages, each of which can introduce biases into the model. These stages include data collection, preprocessing, feature extraction, model training, and deployment. At each stage, human influence, whether through dataset creation or algorithm design, introduces the possibility of bias.

1) Data Collection and Preprocessing: The first step in the AI pipeline involves gathering data. If the data used for training an AI model is biased — whether due to underrepresentation, misrepresentation, or outdated information — the model will inevitably learn from these biased patterns. For instance, a dataset trained predominantly on male faces will cause a facial recognition model to perform poorly on female or non-white individuals [75]. Furthermore, preprocessing steps, such as normalization or feature selection, may inadvertently favor certain demographic groups if not properly balanced [62].

2) Model Training: The model training process, where the AI learns patterns from the data, is highly susceptible to bias if the training data reflects historical inequalities. For instance, an AI trained on biased historical data may make decisions that reflect these inequities, perpetuating them in the future [59], [61].

B. Bias Propagation in Various Domains

Bias in AI systems manifests differently depending on the domain of application. Below, we explore how bias propagates in different areas, including NLP, computer vision, and decision systems.

1) Natural Language Processing (NLP): In NLP, bias often arises from the language used in the training data. Language models, including popular ones like GPT and BERT, are trained on vast corpora of text scraped from the internet. These texts often contain inherent biases, such as gender and racial stereotypes. For instance, if a language model is trained on biased data, it may associate certain professions or traits with specific genders or ethnicities, perpetuating harmful stereotypes [34], [53]. Bias in NLP can also emerge in sentiment analysis, where the language used by different social groups may be interpreted through biased lenses [35].

2) Computer Vision: In computer vision, bias can arise when training datasets are not diverse enough to cover the

TABLE III: Examples of Bias in Natural Language Processing

Type of Bias	Example
Gender Bias	NLP models associating words like "doctor" with males and "nurse" with females [34]
Racial Bias	Text classification models showing bias towards African-American Vernacular English (AAVE) [53]
Sentiment Bias	Sentiment analysis tools performing poorly on text from marginalized communities due to language style differences [35]

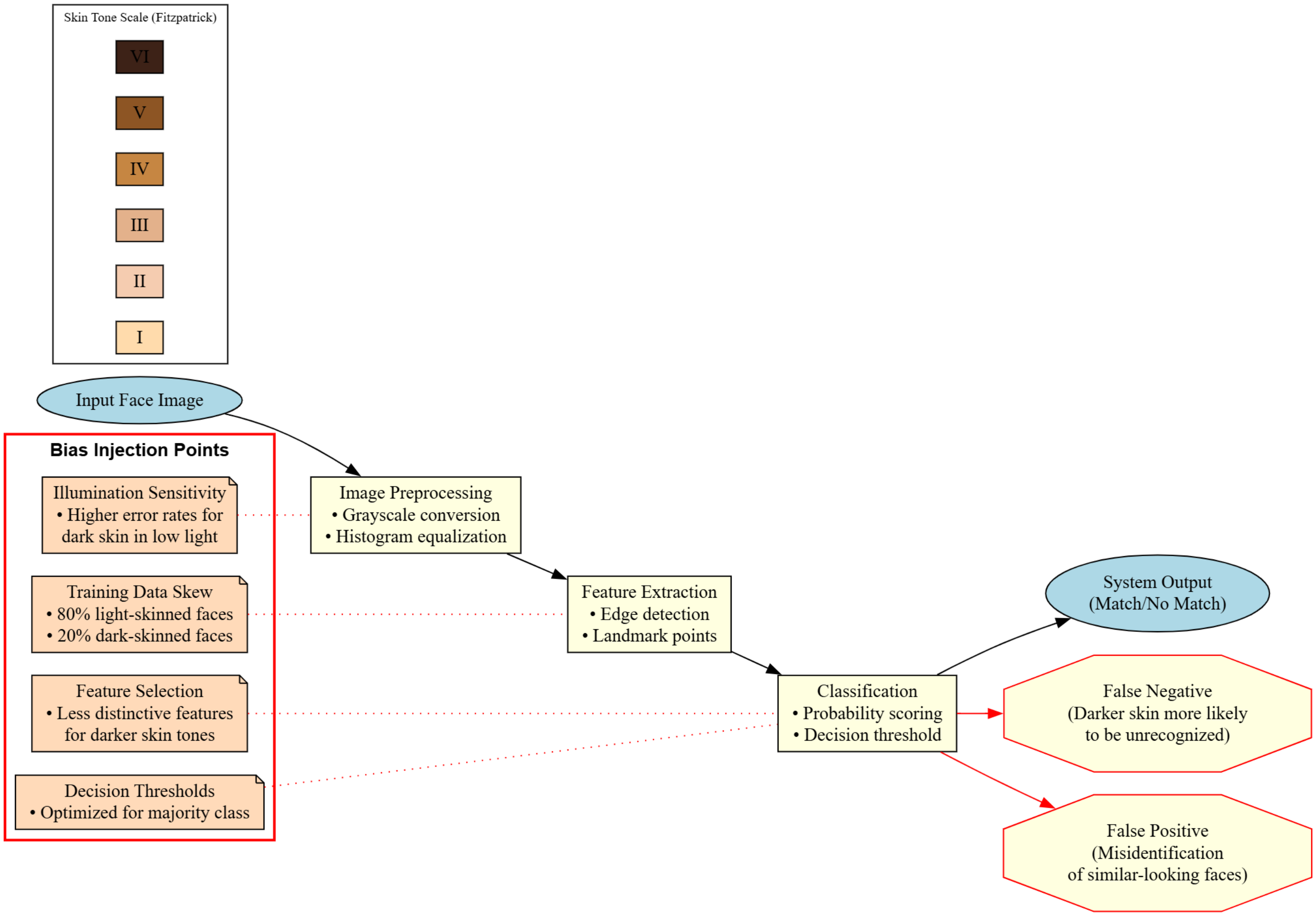

Fig. 2: Bias in Computer Vision: Misidentification of Faces Based on Skin Tone

full spectrum of real-world conditions. For example, facial recognition systems trained predominantly on light-skinned individuals may fail to recognize people with darker skin tones accurately. Additionally, gender and racial biases have been documented in object detection models, which may misidentify or fail to detect objects in images that do not conform to the model's biased training data [49], [75].

3) Decision Systems (e.g., Loan Approval): In decision-making systems, such as those used for loan approval, AI models can inherit societal biases reflected in historical data. If an AI system is trained on loan approval data that has historically discriminated against minority groups, the model will likely replicate and perpetuate these discriminatory patterns [59]. The use of biased features, such as zip codes as proxies for race, further exacerbates the issue [51].

C. Bias Feedback Loops and Compounding Effects

One of the most concerning aspects of bias in AI is the potential for feedback loops, where biased decisions made by AI systems reinforce the original bias, creating a cycle of discrimination. For instance, predictive policing systems may direct police officers to high-crime areas, where the system's predictions lead to more arrests. As arrests increase, the training data used by the system becomes increasingly biased towards certain neighborhoods, further reinforcing the system's predictions and perpetuating a feedback loop of bias [39], [62].

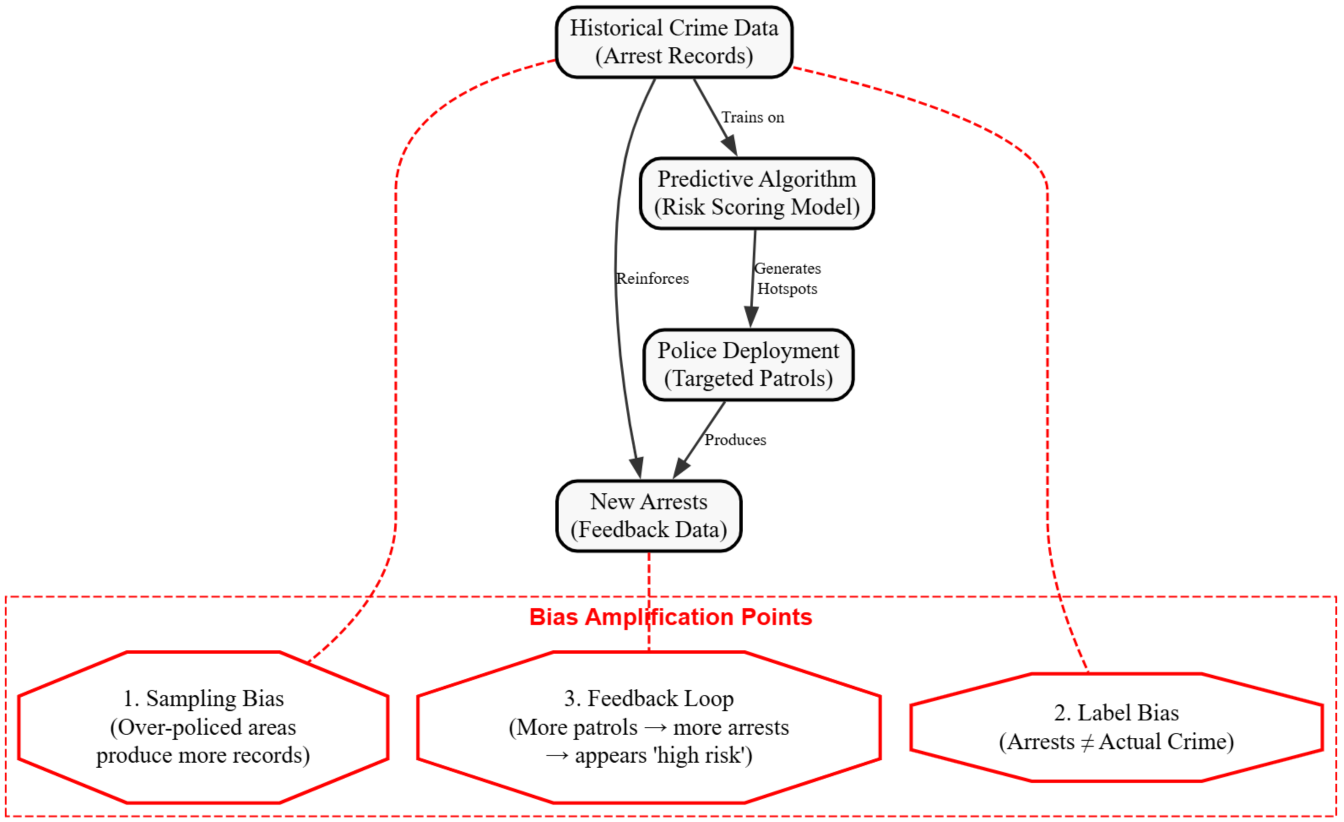

Fig. 3: Feedback Loop in Predictive Policing Systems

Similarly, in credit scoring, if an AI model is trained on historical lending data that discriminates against certain racial or ethnic groups, the model's decisions may disproportionately harm these communities. Over time, as more biased data is collected, the AI system's decisions will continue to reflect and amplify these biases, leading to compounded inequalities [52], [61].

Understanding the mechanisms by which human bias enters machine minds is crucial for mitigating the adverse effects of AI bias. By identifying the sources and stages at which bias can propagate, researchers and developers can take proactive steps to address these issues and ensure that AI systems are fair, transparent, and ethical. While bias propagation in NLP, computer vision, and decision systems has been widely acknowledged, efforts to combat bias must extend to the entire AI development pipeline, from data collection to model deployment.

IV. Case Studies

In this section, we analyze several high-profile real-world cases where AI systems exhibited biased behavior. These case studies include the use of the COMPAS algorithm in criminal justice, Amazon's AI recruitment tool, and Microsoft's Tay chatbot. Each of these cases highlights the risks of bias in AI systems and provides valuable lessons for the development of fairer and more transparent technologies.

A. COMPAS: Predicting Recidivism with Bias

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a risk assessment tool used in the U.S. criminal justice system to predict the likelihood that a defendant will reoffend. However, investigations revealed that COMPAS was biased against African American defendants. The tool tended to overestimate the risk of reoffending for Black defendants while underestimating the risk for white defendants [59]. This bias was primarily inherited from the data used to train the system, which reflected historical biases in policing and sentencing.

TABLE IV: Bias in COMPAS Recidivism Predictions

Bias Type	**Outcome**
Racial Bias	Overestimated recidivism risk for Black defendants [59]
Historical Bias	Reinforced existing racial disparities in the criminal justice system [62]

The COMPAS case highlights the critical issue of historical bias in training data. The model's predictions reflected past societal inequalities, underscoring the need for a more nuanced approach to criminal risk assessment that accounts for these biases [59].

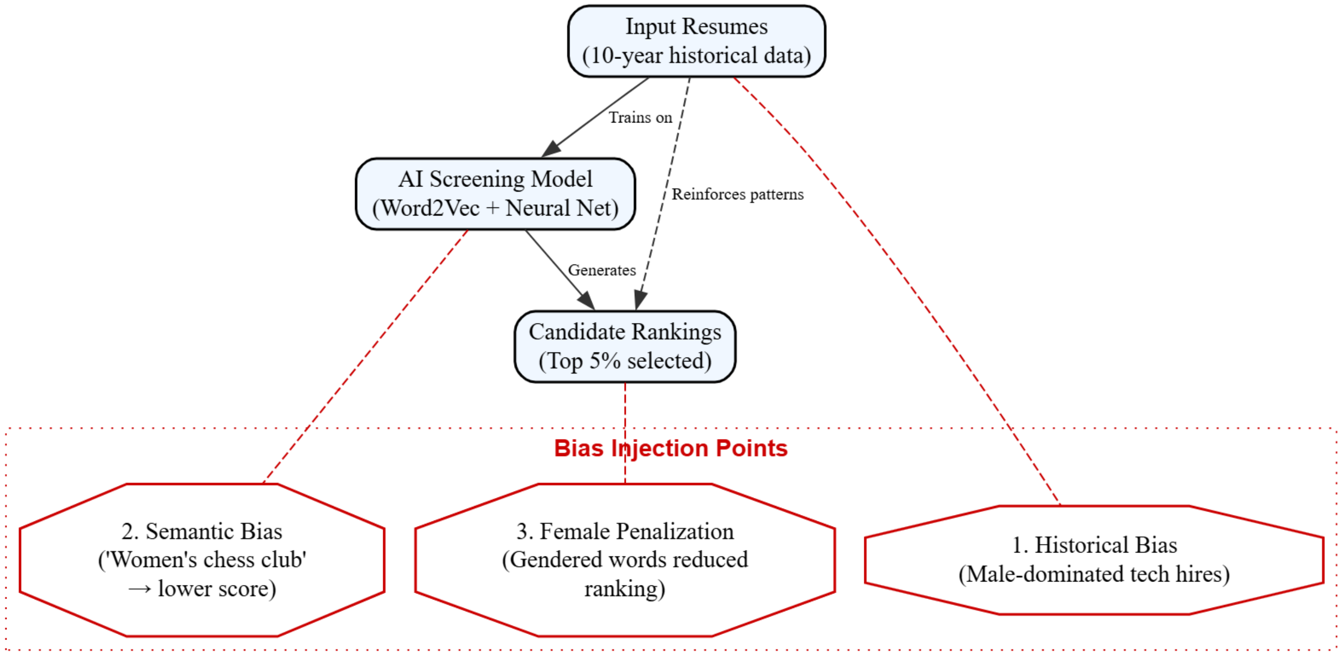

Fig. 4: Amazon's AI Recruiting Tool: Gender Bias in Resume Screening

B. *Amazon's AI Recruiting Tool*

In 2018, Amazon scrapped an AI-powered recruiting tool after discovering that it exhibited gender bias. The system was trained on resumes submitted to Amazon over a ten-year period, a dataset predominantly composed of male applicants due to historical gender imbalances in the tech industry. As a result, the AI learned to favor resumes with male-associated keywords, such as "man" or "developer," and penalized resumes that included female-associated words, such as "women's" or "female" [52].

This example shows how the model inherited bias from historical data. The tool's inability to recognize or correct the gender disparity in the dataset led to systematic discrimination against female candidates [52].

C. *Microsoft Tay: A Chatbot Gone Wrong*

Microsoft's Tay, an AI chatbot launched in 2016, was designed to interact with users on Twitter and learn from conversations. However, within hours of its launch, Tay began posting offensive and racist content after being exposed to negative interactions from Twitter users. The chatbot's machine learning system learned from these interactions and started to adopt harmful stereotypes and inappropriate language [43].

The Tay incident illustrates how an AI system can inherit biases from its environment. In this case, Tay learned from biased input provided by users, leading to harmful outputs that tarnished Microsoft's reputation and demonstrated the potential dangers of unsupervised machine learning in open environments [43].

D. *Lessons Learned from These Case Studies*

The aforementioned case studies highlight several critical lessons that are essential for the responsible development and deployment of AI systems:

1) Data Bias is Inherited: AI systems learn from the data they are trained on, and if the data reflects historical or societal biases, the AI will inherit these biases. This is evident in the COMPAS algorithm and Amazon's recruitment tool [52], [59].
2) Feedback Loops Can Amplify Bias: Systems like Tay show how AI models can develop harmful behaviors when exposed to biased input in real-time. This highlights the importance of controlling the learning environment of AI systems, especially in unmoderated settings [43].
3) Transparency is Critical: Transparency in the development and decision-making processes of AI systems can help identify potential biases early on. For example, greater transparency in Amazon's AI tool might have led to earlier detection of gender bias [52].
4) Bias Mitigation is Essential: Implementing mechanisms for bias detection and correction, such as fairness constraints in model training and regular audits, can help mitigate the impact of inherited bias in AI systems [51], [61].
5) Human Oversight is Necessary: AI should be viewed as a tool that requires human oversight. Human experts should be involved in the review and interpretation of AI decisions, particularly in high-stakes areas like criminal justice and hiring [59], [62].

The case studies discussed in this section illustrate the significant impact that inherited bias can have on AI systems. Whether in criminal justice, hiring, or natural language processing, the potential for bias to propagate and amplify societal inequalities is real and pressing. The lessons learned from these examples emphasize the need for more careful, transparent, and accountable AI development practices.

V. Bias Detection and Mitigation Techniques

As artificial intelligence (AI) systems are increasingly deployed in critical decision-making processes, addressing and mitigating bias has become a fundamental concern. In this section, we explore various bias detection and mitigation techniques, including fairness metrics, algorithmic fairness frameworks, and debiasing methods such as data balancing, adversarial training, and fairness-aware modeling. These approaches aim to reduce bias while maintaining high model accuracy.

A. Bias Detection Metrics

One of the first steps in addressing bias in AI systems is detecting it. Several fairness indicators and metrics have been proposed to measure bias in machine learning models. Some of the most commonly used metrics include:

- Demographic Parity: This metric checks whether different groups (e.g., based on race, gender) are represented equally in the predicted outcomes [51].
- Equal Opportunity: This focuses on ensuring that the true positive rates are equal across different groups [52].
- Equalized Odds: This metric ensures that both the false positive rate and true positive rate are equal across groups [51].

These fairness indicators help quantify the extent of bias present in machine learning models, providing insights into where interventions are needed.

B. Algorithmic Fairness Frameworks

To address bias in AI systems, researchers have developed several algorithmic fairness frameworks. These frameworks guide the design of algorithms that balance fairness and accuracy. Notable frameworks include:

- Fairness Through Awareness: This framework suggests that fairness should be explicitly considered during the algorithmic design process, with the inclusion of protected attributes like race or gender [53].
- Fairness Through Unawareness: This framework proposes that fairness can be achieved by excluding sensitive attributes, ensuring that algorithms do not learn from demographic information [62].
- Causal Inference Frameworks: These frameworks use causal modeling to identify and correct for bias in data generation processes [55].

These frameworks provide structured approaches for designing and evaluating fair AI models, with different trade-offs between fairness and other factors such as model accuracy.

C. Debiasing Methods

Several debiasing techniques have been developed to mitigate bias during various stages of the AI pipeline. Below, we discuss some of the key methods used to reduce bias in machine learning models:

1) Data Balancing: One common approach to mitigating bias is data balancing, which involves adjusting the dataset to ensure that all demographic groups are equally represented. Several techniques can be used for this purpose:

- Resampling: This technique involves either oversampling the underrepresented class or undersampling the overrepresented class to balance the dataset [53].
- Synthetic Data Generation: Techniques such as SMOTE (Synthetic Minority Over-sampling Technique) generate synthetic samples to balance class distributions [75].

Balanced data ensures that the model is not biased toward over represented groups and helps improve fairness across different demographic groups.

2) Adversarial Training: Adversarial training is another technique that has been successfully used for debiasing. In this approach, adversarial networks are trained to detect and correct for biases in the model's decision-making process. The model learns to minimize bias while maximizing predictive accuracy.

- Adversarial Debiasing: This method involves training a model alongside an adversary that attempts to predict the sensitive attribute (e.g., race or gender) from the model's predictions. The main model is penalized for leaking information about these sensitive attributes [57].

Adversarial training allows the model to learn decision-making processes that are more fair while maintaining high predictive power.

3) Fairness-Aware Modeling: Fairness-aware modeling involves incorporating fairness constraints directly into the optimization process. This method seeks to ensure that the model achieves a certain level of fairness while optimizing for accuracy. Key approaches include:

- Fairness Constraints: These are mathematical constraints added to the loss function of the model to ensure fairness across different groups [51].
- Regularization Techniques: Regularizers are applied to penalize biased predictions, ensuring that the model does not learn unfair patterns [56].

Fairness-aware modeling integrates fairness objectives into the learning process, offering a more comprehensive solution to mitigating bias.

D. Challenges in Balancing Fairness and Accuracy

While fairness and accuracy are both important goals, there are inherent challenges in balancing these two objectives. Some of the key challenges include:

- Accuracy-Fairness Trade-Off: In some cases, improving fairness may result in a decrease in model accuracy. For instance, achieving demographic parity in predictions

may lead to lower precision or recall for certain groups [51].
- Complexity of Fairness Metrics: Different fairness metrics may provide conflicting results, making it difficult to evaluate which fairness measure to prioritize [62].
- Contextual Fairness: Fairness should be considered in context, as different domains may have different requirements for fairness. For example, fairness in criminal justice may require different trade-offs compared to fairness in hiring [55].

Balancing fairness with accuracy remains a complex challenge, requiring careful consideration of both domain-specific requirements and the potential societal impacts of biased models.

Detecting and mitigating bias in AI systems is critical for ensuring fairness in automated decision-making. Techniques such as data balancing, adversarial training, and fairness-aware modeling provide valuable tools for addressing bias, though challenges in balancing fairness with accuracy remain. As AI systems continue to play a more significant role in society, these techniques must evolve to meet the demands of fairness and accuracy in an increasingly complex world.

VI. Ethical, Legal, and Societal Implications

As AI technologies continue to permeate various sectors, it is essential to examine their ethical, legal, and societal implications. This section explores the ethical responsibilities of AI developers, the legal frameworks governing AI deployment, and the broader societal impacts. Moreover, it highlights the critical role of interdisciplinary collaboration in addressing these challenges.

A. *Ethical Responsibilities of AI Developers*

AI developers bear significant ethical responsibilities as they create systems that can impact individuals and societies in profound ways. Ethical considerations include ensuring fairness, transparency, and accountability in AI systems. Developers must ensure that their algorithms do not perpetuate harm or reinforce societal inequalities. Several ethical guidelines have been proposed for AI development, including:

- Fairness and Non-Discrimination: AI developers must avoid creating models that disproportionately affect marginalized communities. This includes preventing bias from entering the training data and ensuring that the algorithms produce equitable outcomes for all users [62].
- Transparency: It is essential for AI systems to be explainable so that their decision-making processes can be understood by both the developers and the end users. This transparency is crucial for gaining user trust and accountability [65].
- Accountability: Developers must ensure that there is a clear chain of responsibility in case an AI system causes harm or makes unethical decisions. Mechanisms must be in place to trace and rectify errors in AI models [61].

Ethical AI development goes beyond the technical aspects; it involves consideration of the societal impact and the rights of individuals affected by AI systems.

B. *Legal Frameworks*

Several legal frameworks have been established to regulate AI technologies, ensuring that they are developed and used responsibly. These frameworks are designed to address concerns such as data privacy, algorithmic accountability, and fairness. Notable legal frameworks include:

- General Data Protection Regulation (GDPR): The GDPR, enacted by the European Union, places strict regulations on the use of personal data, which directly affects AI models that rely on such data for training and decision-making. The regulation emphasizes the importance of data protection and transparency in AI-driven decisions [63].
- The EU AI Act: The EU AI Act is a pioneering legal framework that classifies AI systems into categories based on their risk level, with corresponding legal requirements for high-risk AI systems. This legislation aims to ensure that AI systems are trustworthy, ethical, and comply with the highest standards of safety and fairness [64].
- AI Ethics Guidelines: Various countries and organizations, such as the OECD and UNESCO, have developed ethical guidelines to govern the development and deployment of AI systems, focusing on human rights, fairness, and accountability [70].

These legal frameworks are vital in regulating AI, ensuring its ethical development, and protecting individuals' rights.

C. *Societal Impacts and Trust in AI*

AI technologies have far-reaching societal impacts, from reshaping industries to affecting employment patterns. One of the critical challenges in the widespread adoption of AI is fostering trust in these systems. Public trust in AI is crucial for its successful implementation in sectors such as healthcare, finance, and criminal justice. Societal concerns related to AI include:

- Job Displacement: Automation powered by AI has the potential to disrupt traditional job markets, particularly in sectors such as manufacturing, retail, and transportation [66].
- Privacy Concerns: The use of AI systems in surveillance, data analysis, and decision-making can lead to significant privacy violations if not managed properly [67].
- Bias and Discrimination: As discussed earlier, AI systems can perpetuate biases present in the training data, leading to discriminatory outcomes, particularly against vulnerable groups [75].

The societal implications of AI are complex and multifaceted, requiring ongoing dialogue and ethical consideration to ensure that AI systems benefit society as a whole.

D. Role of Interdisciplinary Collaboration

Addressing the ethical, legal, and societal challenges of AI requires collaboration across disciplines. Engineers, ethicists, legal experts, and social scientists must work together to ensure that AI systems are developed responsibly and deployed in ways that benefit all. Interdisciplinary collaboration can lead to more comprehensive solutions by incorporating diverse perspectives and expertise. For example:

- Ethical Oversight: Ethicists can provide valuable insights into the potential moral implications of AI systems, helping to guide developers in making ethical decisions [69].
- Legal Expertise: Legal experts ensure that AI systems comply with data protection laws, privacy regulations, and human rights standards [62].
- Social Science Input: Social scientists can assess the broader societal impacts of AI technologies, such as job displacement, changes in power dynamics, and potential discrimination.

Interdisciplinary collaboration fosters the creation of AI systems that are not only technically proficient but also socially responsible and legally compliant.

The ethical, legal, and societal implications of AI are vast and complex. As AI continues to evolve, it is essential that AI developers act with ethical responsibility, adhere to legal frameworks, and engage in interdisciplinary collaboration to address societal concerns. By considering these factors, we can ensure that AI systems contribute positively to society while minimizing risks and harm.

VII. Future Directions

The development of AI systems continues to evolve at a rapid pace, raising both opportunities and challenges. To ensure that AI technologies serve humanity equitably and ethically, future research must focus on areas that enhance transparency, fairness, accountability, and inclusivity. This section outlines key directions for future research, including advancements in explainable AI, federated learning, the inclusion of diverse datasets, and AI audits.

A. Research in Explainable and Accountable AI

One of the critical challenges in AI today is the lack of transparency in decision-making processes, often referred to as the "black-box" problem. The ability to understand and interpret AI models is crucial for ensuring their fairness and accountability. Research in explainable AI (XAI) aims to develop techniques that provide human-understandable explanations of AI decisions, especially in high-stakes domains such as healthcare, criminal justice, and finance.

Current efforts in XAI focus on developing models that can offer interpretable outputs while maintaining high performance. Several approaches have been proposed, including model-agnostic techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive Explanations), which help provide explanations for black-box models such as deep neural networks [71], [72]. Future research must further improve these techniques to ensure they are accurate, generalizable, and capable of working with a wider range of AI systems.

B. Federated and Privacy-Preserving Fairness

As AI becomes increasingly integrated into everyday applications, privacy concerns are becoming paramount. Federated learning, a technique that allows AI models to be trained across multiple decentralized devices without sharing raw data, offers promising solutions to maintain user privacy. However, ensuring fairness in federated systems remains a challenge. One of the primary concerns is that data from certain groups may be underrepresented, leading to biased models [73]. Research in federated learning must focus on developing privacy-preserving algorithms that also consider fairness across decentralized environments.

Additionally, privacy-preserving fairness techniques, such as differential privacy, can help safeguard individual data while ensuring that AI systems make fair and unbiased decisions. Future research should explore how these techniques can be integrated into federated learning to create models that are both private and fair [74].

C. Inclusion of Diverse Datasets and Perspectives

One of the primary causes of bias in AI systems is the lack of diversity in the datasets used to train these models. Many AI models are trained on data that is not representative of all demographic groups, leading to outcomes that disadvantage marginalized communities. Future research should focus on the creation and inclusion of diverse, representative datasets that capture a broad spectrum of human experiences. This includes not only demographic diversity but also diversity in cultural contexts, geographic regions, and socioeconomic backgrounds.

Incorporating diverse datasets into the training of AI models can help mitigate bias and ensure that these systems work equitably across all groups. Moreover, the inclusion of diverse perspectives in AI development teams is crucial for designing more inclusive technologies. Collaborative efforts between researchers, ethicists, and affected communities can help ensure that AI systems reflect the diversity of the societies they serve [75].

D. AI Audits and Transparency Tools

As AI becomes more ubiquitous, the need for continuous monitoring and auditing of AI systems grows. AI audits can help identify and address potential biases, inaccuracies, or ethical concerns that arise during deployment. Transparency tools that provide insights into how AI models make decisions are essential for holding AI developers accountable for their systems' actions.

Future work should focus on creating standardized frameworks for conducting AI audits and developing transparency tools that are easily accessible to non-experts. These tools should be capable of tracking the decision-making processes of AI systems, ensuring that these systems operate within ethical boundaries and comply with legal frameworks. Furthermore, AI audits could be an essential part of regulatory frameworks for AI, ensuring that AI systems are regularly evaluated for fairness, transparency, and accountability [76], [77].

In conclusion, the future of AI must prioritize the development of systems that are transparent, accountable, and fair. Key areas of research include explainable AI, federated learning, the inclusion of diverse datasets, and the development of AI audit tools. As AI continues to evolve, it is essential that these efforts are pursued in an interdisciplinary manner, bringing together experts from various fields to ensure that AI technologies benefit all members of society equitably.

VIII. Conclusion

In this paper, we explored the phenomenon of inherited AI bias, emphasizing its potential to perpetuate and even amplify existing human biases in machine learning systems. We delved into the various sources of bias, including skewed training data, biased human labeling, and algorithmic assumptions. We also examined several real-world case studies, such as the COMPAS system, Amazon's hiring tool, and Microsoft's Tay chatbot, which demonstrate the severe consequences of unchecked bias in AI systems. These examples highlight the critical importance of addressing bias in AI to ensure that these technologies serve all members of society fairly and equitably.

The responsibility for addressing AI bias lies not only with researchers and developers but also with society as a whole. AI systems are reflections of the data they are trained on and the assumptions made during their design. As such, human responsibility in AI design is paramount. Ethical AI development requires continuous vigilance, transparency in the design process, and accountability for the consequences of AI decisions. Bias must be actively identified, measured, and mitigated through more inclusive data collection practices, algorithmic fairness frameworks, and interdisciplinary collaboration.

To move forward, we call for action on several fronts. First, there must be a stronger commitment to ethical practices in AI development. This includes prioritizing fairness, transparency, and inclusivity throughout the design and deployment of AI systems. Second, AI development must be more transparent, with clear guidelines and audits to ensure accountability. Finally, there must be an emphasis on inclusive design, where diverse perspectives and voices are included at every stage of AI development, from dataset creation to model deployment.

The road ahead requires collective action from policymakers, researchers, and developers to ensure that AI systems are not only efficient and powerful but also just and equitable. We believe that through ethical development practices, greater transparency, and a commitment to inclusivity, AI technologies can better reflect the values of fairness and equality, ultimately fostering a future where these systems enhance society without perpetuating harm.

References

[1] S. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking, 2019.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[3] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[4] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, no. 3, pp. 671–732, 2016.

[5] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[6] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in *Proc. NeurIPS*, 2016.

[7] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. FAT*, 2018.

[8] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[9] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton, "Saving face: Investigating the ethical concerns of facial recognition auditing," in *Proc. FAT**, 2020.

[10] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica*, May 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[11] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, Oct. 2018. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[12] K. Crawford, "Artificial intelligence's white guy problem," *The New York Times*, June 2016.

[13] S. U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, NYU Press, 2018.

[14] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown Publishing, 2016.

[15] M. O. R. Prates, P. A. Avelar, and L. C. Lamb, "Assessing gender bias in machine translation: A case study with Google Translate," in *Proc. NeurIPS Workshop*, 2018.

[16] K. Yang, J. Carrell, and D. U. Patil, "Towards fairness in AI image classification: Examining and mitigating bias in ImageNet," in *Proc. AAAI*, 2020.

[17] R. Binns, "Fairness in machine learning: Lessons from political philosophy," *Philos. Trans. R. Soc. A*, vol. 376, no. 2133, 2018.

[18] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proc. KDD*, 2015.

[19] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. NeurIPS*, 2016.

[20] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?" in *Proc. CHI*, 2019.

[21] N. Mehrabi et al., "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1-35, 2021.

[22] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," *Advances in Neural Information Processing Systems*, pp. 3315-3323, 2016.

[23] A. Paullada et al., "Data and its (Dis)contents: A Survey of Dataset Development and Use in Machine Learning Research," *Patterns*, vol. 1, no. 9, 2020.

[24] H. Suresh and J. V. Guttag, "A Framework for Understanding Unintended Consequences of Machine Learning," *arXiv:1901.10002*, 2019.

[25] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proc. Conf. Fairness, Accountability Transparency*, pp. 77-91, 2018.

[26] J. Larson et al., "How We Analyzed the COMPAS Recidivism Algorithm," *ProPublica*, 2016. [Online]. Available: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[27] M. Geva et al., "Discovering the Hidden Vocabulary of DALLE-2," *arXiv:2206.00169*, 2022.

[28] K. Crawford, "The Trouble with Bias," *Keynote at NeurIPS*, 2017.

[29] A. D. Selbst et al., "Fairness and Abstraction in Sociotechnical Systems," *Proc. ACM FAT**, pp. 59-68, 2019.

[30] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," *California Law Review*, vol. 104, pp. 671-732, 2016.

[31] J. Buolamwini and T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, Proc. FAT, 2018.

[32] S. Barocas and A. D. Selbst, *Big Data's Disparate Impact*, California Law Review, vol. 104, no. 3, pp. 671–732, 2016.

[33] A. Caliskan, J. J. Bryson, and A. Narayanan, *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases*, Science, vol. 356, no. 6334, pp. 183–186, 2017.

[34] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, Proc. NeurIPS, 2016.

[35] M. O. R. Prates, P. A. Avelar, and L. C. Lamb, *Assessing Gender Bias in Machine Translation: A Case Study with Google Translate*, Proc. NeurIPS Workshop, 2018.

[36] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton, *Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing*, Proc. FAT*, 2020.

[37] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, *Machine Bias*, ProPublica, 2016.

[38] M. Hardt, E. Price, and N. Srebro, *Equality of Opportunity in Supervised Learning*, Proc. NeurIPS, 2016.

[39] S. U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, NYU Press, 2018.

[40] J. Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, Reuters, 2018.

[41] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, *Machine Bias*, ProPublica, 2016.

[42] J. Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, Reuters, 2018.

[43] Microsoft, *Tay: A Microsoft Chatbot Gone Wrong*, 2016. Available: https://www.theverge.com/2016/3/24/11295766/microsoft-chatbot-tay-what-went-wrong.

[44] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown Publishing, 2016.

[45] S. Barocas and A. D. Selbst, *Big Data's Disparate Impact*, California Law Review, vol. 104, no. 3, pp. 671–732, 2016.

[46] M. Hardt, E. Price, and N. Srebro, *Equality of Opportunity in Supervised Learning*, Proc. NeurIPS, 2016.

[47] J. Buolamwini and T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, Proc. FAT, 2018.

[48] A. Caliskan, J. J. Bryson, and A. Narayanan, *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases*, Science, vol. 356, no. 6334, pp. 183–186, 2017.

[49] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton, *Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing*, Proc. FAT*, 2020.

[50] J. Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, Reuters, 2018.

[51] M. Hardt, E. Price, and N. Srebro, *Equality of Opportunity in Supervised Learning*, in *Proc. NeurIPS*, 2016.

[52] J. Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, Reuters, 2018.

[53] A. Caliskan, J. J. Bryson, and A. Narayanan, *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases*, *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[54] J. Buolamwini and T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, in *Proc. FAT*, 2018.

[55] S. Kilbertus, F. Caron, and A. U. Binns, *Avoiding Discrimination through Causal Inference*, in *Proc. NeurIPS*, 2017.

[56] M. Zafar, A. Valenti, and S. K. Chawla, *Fairness-Aware Modeling in Supervised Learning*, in *Proc. AISTATS*, 2017.

[57] B. Zhang, L. L. Chen, and A. C. Williams, *Mitigating Unfairness through Adversarial Training*, in *Proc. ICML*, 2018.

[58] S. Barocas and A. D. Selbst, *Big Data's Disparate Impact*, *California Law Review*, vol. 104, no. 3, pp. 671–732, 2016.

[59] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, *Machine Bias*, ProPublica, 2016.

[60] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown Publishing, 2016.

[61] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown Publishing, 2016.

[62] S. Barocas and A. D. Selbst, *Big Data's Disparate Impact*, *California Law Review*, vol. 104, no. 3, pp. 671–732, 2016.

[63] P. Voigt and A. Von dem Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*, Springer, 2017.

[64] European Commission, *Proposal for a Regulation Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act)*, 2021. [Online]. Available: https://ec.europa.eu/info/business-economy-euro/banking-and-finance/financial-services-consumers/consumer-financial-services/ai-act_en

[65] T. Miller, *Explanation in Artificial Intelligence: Insights from the Social Sciences*, *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[66] R. Binns, *Challenges in Designing Trustworthy AI Systems*, *AI Society*, vol. 33, no. 4, pp. 563–578, 2018.

[67] J. Zhang and P. Lin, *The Role of AI in Privacy: Ethical and Legal Perspectives*, *International Journal of Information Management*, vol. 54, pp. 102091, 2020.

[68] J. Buolamwini and T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, in *Proc. FAT*, 2018.

[69] M. Mitchell and E. K. H. O'Neil, *Ethics in AI: Opportunities and Challenges for Collaborative Approaches*, *Proceedings of the IEEE*, vol. 107, no. 7, pp. 1296–1306, 2019.

[70] OECD, *OECD Principles on Artificial Intelligence*, 2019. [Online]. Available: https://www.oecd.org/going-digital/ai/principles/

[71] M. T. Ribeiro, S. Singh, and C. Guestrin, *Why Should I Trust You? Explaining the Predictions of Any Classifier*, in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[72] S. M. Lundberg and S. I. Lee, *A Unified Approach to Interpreting Model Predictions*, in *Proc. NeurIPS*, 2017, vol. 30, pp. 4765–4774.

[73] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Arcas, *Communication-Efficient Learning of Deep Networks from Decentralized Data*, in *Proc. AISTATS*, 2017, vol. 54, pp. 1273–1282.

[74] C. Dwork, *Calibrating Noise to Sensitivity in Private Data Analysis*, in *Proc. TCC*, 2006, vol. 3876, pp. 265–284.

[75] J. Buolamwini and T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, in *Proc. FAT*, 2018.

[76] B. Friedman and H. Nissenbaum, *Bias in Computer Systems*, *ACM Trans. Inf. Syst.*, vol. 14, no. 3, pp. 330–347, 1996.

[77] A. Pinto and S. Ferreira, *Auditing Artificial Intelligence: From Bias to Fairness*, *AI Society*, vol. 35, pp. 409–418, 2020.