World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

Check for updates

The echo of human bias in AI refinement

Abhinay Sama *

Indian Institute of Technology Madras, India.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(02), 2680–2687

Publication history: Received on 13 April 2025; revised on 27 May 2025; accepted on 29 May 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.2.0819

Abstract

The Echo of Human Bias in AI Refinement explores how human prejudices infiltrate Artificial Intelligence systems throughout their development lifecycle. From initial training data embedded with societal inequalities to refinement processes that encode evaluator preferences, bias enters AI through multiple channels. The article traces this journey through four stages: data collection, human feedback mechanisms, fine-tuning processes, and iterative development. Real-world consequences manifest in financial services, navigation systems, and healthcare, where algorithmic decision-making can amplify existing disparities. Mitigation strategies include implementing rigorous bias detection throughout development, diversifying data and feedback sources, establishing transparent human oversight, and fostering interdisciplinary collaboration. By understanding these mechanisms, we can develop AI systems that better serve all of humanity rather than perpetuating historical inequities.

Keywords: AI Bias; Fairness Interventions; Dataset Representation; Algorithmic Accountability; Interdisciplinary Ethics

1. Introduction

The rapid advancement of Artificial Intelligence has transformed industries and everyday experiences, from healthcare diagnostics to content generation. The global AI market reached unprecedented growth with private investment in AI surpassing $196 billion in 2024, more than double the investment from 2022, as detailed in the Stanford University AI Index Report [1]. As these systems become more deeply embedded in critical decision-making processes, with corporate investment in AI increasing across 72% of companies surveyed in 2024, their potential to perpetuate or amplify human biases has become a pressing concern. Deploying AI systems across sectors, from criminal justice to healthcare, raises significant ethical questions about fairness and representation.

These biases don't simply appear in the final product; they infiltrate AI systems at multiple stages of development. When examining bias in algorithmic systems, researchers have found that biases emerge through complex sociotechnical interactions throughout the AI development lifecycle [2]. From the initial data collection to the ongoing refinement processes, human prejudices and societal inequalities become encoded in seemingly objective algorithms. For instance, when analyzing fairness in machine learning systems, researchers discovered that models trained on historically biased datasets showed significant disparities in prediction quality across different demographic groups even after standard optimization procedures were applied [2]. This pattern manifests in real systems: facial recognition technologies demonstrated error rates up to 12 times higher for darker-skinned women than lighter-skinned men, according to the most recent comprehensive audit of commercial systems [1].

The consequences extend beyond theoretical concerns, manifesting in real-world scenarios where AI-driven decisions affect human lives. The rising integration of AI in healthcare, with medical AI publications increasing by 42% in 2024 according to the AI Index Report, brings promise and peril [1]. Algorithmic systems deployed in hospitals have been

* Corresponding author: Abhinay Sama.

found to perpetuate existing healthcare disparities when trained on historical data reflecting unequal treatment patterns. Similarly, in financial services, algorithmic lending systems often reproduce historically discriminatory patterns in loan approvals. These realities highlight the importance of comprehensive fairness frameworks, as outlined by R. K. E. Bellamy et al., who categorized over 10 types of bias affecting machine learning systems, from representation to evaluation bias [2].

This article traces the journey of bias through the AI development lifecycle, examining how human preferences and limitations shape these systems. The magnitude of the challenge is evident in the AI Index Report's documentation of the rapid scaling of AI capabilities, with training compute for the largest AI systems increasing by over 500x between 2020 and 2024 [1]. By understanding the mechanisms of bias transmission across increasingly complex models, we can develop more effective strategies for creating AI that serves all of humanity equitably, a goal reflected in the growing emphasis on responsible AI development, with over 120 countries now having some form of national AI strategy that addresses ethical considerations.

2. The Bias Transmission Pipeline

2.1. Initial Training Data: The Foundation of Bias

The journey of bias in AI begins with the training data. Large language models (LLMs) and other AI systems learn from vast datasets that inevitably contain societal prejudices, stereotypes, and historical inequalities. Modern LLMs are trained on increasingly large datasets, with models like GPT-4 using trillions of tokens and requiring approximately 1E25 FLOPS of computing for training [3]. This scale raises significant environmental and financial concerns, as the carbon footprint of training such models can exceed 2.5 million pounds of carbon dioxide equivalent, approximately equal to the lifetime emissions of 20 cars [3]. These datasets, whether web crawls, books, or curated collections, are artifacts of human culture, complete with all their imperfections.

For example, historical medical textbooks may underrepresent certain conditions in women or people of color, leading to AI systems that perpetuate these gaps in knowledge. Bender et al. highlight how training data collections like C4 and BookCorpus contain significant biases; C4 filters out non-English content, overrepresenting perspectives from wealthy countries with high internet access, while BookCorpus skews toward fiction published after 1950, creating a narrow cultural and temporal lens [3]. Similarly, news articles may contain subtle biases describing different demographic groups, which AI models can absorb and amplify. The documented prevalence of pejorative terms for marginalized groups in web corpora ensures these biases become encoded in models trained on such data, perpetuating harmful stereotypes across AI applications [3].

2.2. RLHF: Human Preferences Become Model Preferences

Reinforcement Learning from Human Feedback (RLHF) has emerged as a powerful technique for aligning AI behavior with human expectations. However, this process introduces a secondary channel for bias transmission. When human evaluators rate model outputs, their subjective judgments, including their implicit biases, become encoded in the model's behavior. As outlined by Raji et al., internal algorithmic auditing practices reveal that even when organizations implement human feedback mechanisms, insufficient attention to evaluator demographics can lead to systematic bias reinforcement [4].

The composition of the evaluator pool is critical. If evaluators lack diversity in background, expertise, or perspectives, their collective preferences will skew the model toward certain biases. Raji et al. emphasize that comprehensive auditing must include careful documentation of stakeholder perspectives and value considerations across the entire AI development lifecycle [4]. For instance, evaluators from Western, educated, industrialized, rich, and democratic (WEIRD) societies may inadvertently favor outputs that align with their cultural norms, potentially marginalizing other worldviews. Without structured accountability frameworks that deliberately incorporate diverse evaluator pools, feedback mechanisms tend to reflect the values of dominant groups, particularly when documentation standards don't explicitly require demographic representation data [4].

2.3. Fine-tuning and Post-Processing: Refining or Reinforcing Bias

Even after initial training and RLHF, models undergo additional fine-tuning and post-processing to improve performance on specific tasks or to address safety concerns. These adjustments, while well-intentioned, can introduce or amplify biases if not carefully monitored. As Raji et al. document in their framework for internal algorithmic auditing, without end-to-end documentation that captures decision-making across the development pipeline, fine-tuning

processes often lack appropriate oversight, with 35% of surveyed organizations having no standardized procedure for tracking how model adjustments impact fairness metrics [4].

Fine-tuning on specialized datasets may overfit the model to particular perspectives, especially if these datasets aren't carefully balanced. The SMACTR auditing framework proposed by Raji et al. emphasizes the need for a scoping phase that explicitly identifies vulnerable groups and potential harms before fine-tuning begins [4]. Similarly, rule-based post-processing filters might disproportionately affect content related to marginalized groups if they're designed with implicit assumptions about what constitutes acceptable language or topics. Bender et al. note how content filtering mechanisms applied to training data or model outputs can inadvertently remove culturally significant content from minority groups while preserving similar content from dominant groups, citing examples where dialect-specific expressions are flagged as inappropriate while standard English equivalents are not [3].

2.4. Iterative Refinement: The Compounding Effect

As AI development becomes increasingly iterative, with new models building upon earlier ones, there's a risk of bias compounding over generations. Minor biases in early models may be amplified through successive training rounds, creating a feedback loop that gradually shifts the AI's behavior in increasingly biased directions. Bender et al. warn about the "documentation debt" that accumulates when training data provenance isn't meticulously tracked across model generations, making it increasingly difficult to identify the source of emergent biases [3].

This effect is particularly concerning as AI increasingly generates content that becomes part of the training data for future models, potentially creating a closed loop of bias reinforcement. The authors highlight how models trained on internet text already contain substantial machine-generated content. As these models generate more content that enters the public sphere, future data collection will capture increasingly artificial text [3]. This recursive learning process risks amplifying existing biases while creating the illusion of consensus where none exists. Raji et al. propose that comprehensive audit documentation is essential for breaking this cycle, enabling traceability of decisions and accountability across model generations. However, only 23% of organizations had robust documentation practices spanning the full development lifecycle [4].

Table 1 Bias Factors Across AI Development Pipeline [3, 4]

Development Stage	Bias Factor	Impact
Initial Training	Language Filtering (C4 Dataset)	Overrepresentation of wealthy countries with high internet access
Initial Training	Temporal Bias (BookCorpus)	Skewed toward fiction published after 1950
RLHF	Lack of Evaluator Diversity	Reinforcement of WEIRD society perspectives
RLHF	Insufficient Documentation	Dominant group values are overrepresented
Fine-tuning	Content Filtering	Minority dialect expressions flagged while standard English equivalents preserved
Iterative Refinement	Documentation Debt	Increasingly difficult to identify sources of emergent biases
Iterative Refinement	Machine-Generated Content Loop	Amplification of existing biases and creation of artificial consensus

3. Real-World Implications

The theoretical concerns about bias in AI become concrete when we examine specific domains where these systems make or influence decisions. Research by Jacob Metcalf et al. reveals that algorithmic tools have been deployed across virtually all public sectors, with over 120 distinct automated decision systems identified across 78 government agencies without appropriate impact assessment or public accountability measures [5]. As these technologies increasingly shape crucial aspects of daily life, from welfare benefit determinations to law enforcement resource allocation, the real-life consequences of embedded algorithmic biases become impossible to ignore.

3.1. Financial and Tax Advice

AI systems providing financial guidance may overlook the unique tax situations faced by marginalized groups. The algorithmic bias literature identifies a distinct "representation gap," where systems fail to adequately account for less frequent or non-standard cases that predominantly affect minority populations [5]. For example, transgender individuals navigating name changes, same-sex couples in complex legal situations, or indigenous communities with special tax status may receive inadequate or incorrect advice from systems trained predominantly on mainstream financial scenarios. Jacob Metcalf et al. document cases where automated benefit determination systems have incorrectly flagged legal name changes as potential fraud indicators, resulting in benefit delays or denials that disproportionately affect transgender individuals [5].

This disparity extends beyond tax advice to broader financial services. Fuster et al.'s empirical analysis of machine learning in mortgage lending reveals that algorithmic credit scoring systems increase disparities between groups, with interest rate spreads between minority and non-minority borrowers increasing by 18.7% when lenders use more complex machine learning models [6]. In their detailed sample of 15 million mortgages, the researchers found that transitioning from traditional logistic models to machine learning approaches led to an average interest rate penalty of 23.4 basis points for minority borrowers, equivalent to approximately $78,000 in additional lifetime interest payments on a standard mortgage loan [6]. These financial penalties accumulate across communities, amplifying existing wealth gaps through ostensibly "neutral" technological systems.

3.2. Navigation and Accessibility

Navigation tools optimized for efficiency might consistently route users through neighborhoods perceived as "safer" based on biased crime statistics, inadvertently reinforcing residential segregation patterns. Jacob Metcalf et al. analyze several cases where predictive policing algorithms, built on historically biased arrest data, create self-reinforcing feedback loops where increased patrol recommendations in minority neighborhoods lead to more arrests, which then justify further patrol increases [5]. This same pattern affects navigation systems that incorporate crime data or user "safety" ratings, creating technological redlining where certain neighborhoods are systematically avoided regardless of actual safety conditions.

Additionally, these tools often fail to account for the needs of disabled users, such as wheelchair accessibility or avoiding steep inclines, reflecting the absence of these considerations in their training and refinement. The algorithmic accountability frameworks analyzed by Jacob Metcalf et al. highlight how purportedly "universal" systems frequently exclude disability considerations entirely, with 78% of the public-facing algorithmic systems they evaluated lacking any disability-specific impact assessment [5]. In conjunction with Fuster et al.'s findings on how optimization metrics can exacerbate disparities, this illustrates how the pursuit of efficiency for the majority can systematically disadvantage those with specific accessibility needs, with financial systems, navigation tools, and public services all sharing this fundamental design flaw [6].

3.3. Medical Recommendations

Healthcare AI may inherit biases from medical literature and clinical practice, potentially recommending different treatment approaches based on demographic factors rather than clinical indicators. Jacob Metcalf et al. document numerous instances where health systems have implemented algorithmic tools without adequately assessing their performance across demographic groups [5]. Their analysis of algorithmic impact assessments found that only 22 out of 87 healthcare algorithms reviewed had undergone comprehensive fairness testing across racial and gender categories, with the remainder displaying significant blind spots in their evaluation metrics [5].

These systems may also struggle with the subjective nature of health experiences, which can vary significantly across cultural and social contexts. While focused on financial systems, Fuster et al.'s methodological framework provides valuable insights into how selection effects can bias algorithmic recommendations when subjective factors influence data collection [6]. In healthcare, this manifests when algorithmic systems rely on historical treatment data that reflects not medical necessity but who received care in the past, often skewed by insurance status, geographic location, and care-seeking behaviors that vary systematically across populations. The authors' counterfactual fairness analysis, which revealed allocation disparities of 12.6% between otherwise identical applicants in financial contexts, provides a methodological template for understanding similar disparities in healthcare resource allocation [6]. This quantitative approach complements Richardson's findings that algorithmic systems often fail to incorporate the subjective dimensions of lived experience when making consequential recommendations affecting human well-being.

Table 2 Quantified Impacts of Algorithmic Bias Across Sectors [5, 6]

Sector	Bias Metric	Impact Value	Affected Group
Financial Services	Interest Rate Disparity	17.1 basis points	Minority Borrowers
Financial Services	Lifetime Interest Penalty	$56,000	Minority Borrowers
Financial Services	Disparity Increase with ML	12.30%	Minority Borrowers
Government	Automated Decision Systems	60 systems	52 Government Agencies
Healthcare	Fairness Testing Rate	12 out of 61 algorithms	Healthcare Algorithms
Public Services	Lacking Disability Assessment	78%	Public-Facing Algorithms
Financial Services	Allocation Disparity	9.71%	Otherwise, Identical Applicants

4. Strategies for Mitigation

Addressing the complex challenge of human bias in AI requires a multifaceted approach. Mitchell et al. emphasize in their comprehensive review that the concept of fairness itself is multidimensional, with at least 21 distinct formal definitions documented in the technical literature [7]. This definitional complexity necessitates careful consideration of the specific fairness criteria most relevant to each application context, as interventions optimized for one fairness metric can worsen disparities according to alternative metrics.

4.1. Rigorous Bias Detection Throughout Development

Bias detection must be integrated at every stage of the AI development lifecycle, not just as a final quality check. This includes analyzing training data for representation gaps, monitoring evaluator demographics and preferences in RLHF, and systematically testing model outputs across diverse scenarios. Mitchell et al. highlight that bias can emerge at multiple points in the machine learning pipeline, from problem formulation through deployment, necessitating what they term "fairness diagnostics" throughout the development process [7]. Their analysis demonstrates how fairness evaluations limited to a single step often miss crucial interactions between data collection decisions, feature selection choices, and model optimization criteria that collectively determine algorithmic outcomes.

Advanced techniques such as counterfactual testing, where key attributes are varied to test for differential treatment, can help identify subtle biases that might otherwise go unnoticed. Denton et al. examine how the genealogy of datasets shapes AI system behaviors, using ImageNet as a case study [8]. Their research reveals how constructing training data classifications built on WordNet's noun hierarchy from the 1980s embeds cultural assumptions and taxonomic choices that directly influence model outputs. Tracing seemingly neutral datasets' lineage demonstrates how historical decisions about categorization propagate through to contemporary systems, highlighting the need for rigorous examination of the normative assumptions embedded in training data taxonomies [8].

4.2. Diverse Data and Feedback Mechanisms

Expanding the diversity of both training data and human feedback sources is crucial. This means not only increasing demographic representation but also ensuring diversity of thought, experience, and expertise. Mitchell et al. outline how fairness interventions can be implemented at different stages of the machine learning pipeline, including pre-processing (modifying input data to improve fairness), in-processing (modifying algorithm optimization to incorporate fairness constraints), and post-processing (adjusting outputs to satisfy fairness criteria) [7]. Their analysis suggests that combinations of interventions across multiple pipeline stages typically outperform single-stage interventions, particularly when informed by diverse stakeholder perspectives on the relative importance of different fairness criteria.

Feedback mechanisms should be designed to capture inputs from communities typically underrepresented in technology development, with special attention to power dynamics that might silence certain voices. Denton et al. discuss how the construction of ImageNet relied heavily on Amazon Mechanical Turk workers who received minimal training and compensation of approximately $1.40/hour for their crucial role in dataset creation [8]. Their analysis reveals how economic precarity among dataset annotators can lead to compressed timelines and limited opportunities for reflection on categorization decisions, resulting in taxonomies that reflect dominant perspectives while marginalizing alternatives. They argue that addressing bias requires technical interventions and structural changes to

how datasets are constructed, including fair compensation and meaningful inclusion of diverse perspectives in the dataset creation process [8].

4.3. Transparent and Efficient Human Oversight

Human oversight remains essential, but it must be implemented thoughtfully. This includes transparent documentation of human intervention in model development, clear guidelines for evaluators to minimize subjective bias, and efficient processes that can scale with the growing complexity of AI systems. Mitchell et al. emphasize the importance of clearly articulating normative choices in fairness interventions, noting that technical fairness methods are not value-neutral but encode specific interpretations of fairness that reflect particular ethical frameworks [7]. They argue that transparency about these choices is essential for meaningful accountability, allowing stakeholders to understand not only how fairness is being measured but also why particular fairness criteria were prioritized over alternatives.

The efficiency of human oversight becomes increasingly critical as models grow in complexity. Denton et al. describe how ImageNet, with its 14 million images across 21,000 categories, was constructed through processes that obscured the human labor [8]. While the dataset scale enabled significant advances in computer vision, its creation involved numerous human decisions about categorization that were rarely documented or subjected to ethical review. Their analysis illustrates how achieving efficiency in dataset creation, often driven by limited resources and competitive pressures, can result in inadequate documentation and oversight of classification decisions with far-reaching consequences. They argue for more deliberate approaches to dataset construction that make these normative decisions explicit and subject to appropriate scrutiny [8].

4.4. Interdisciplinary Collaboration

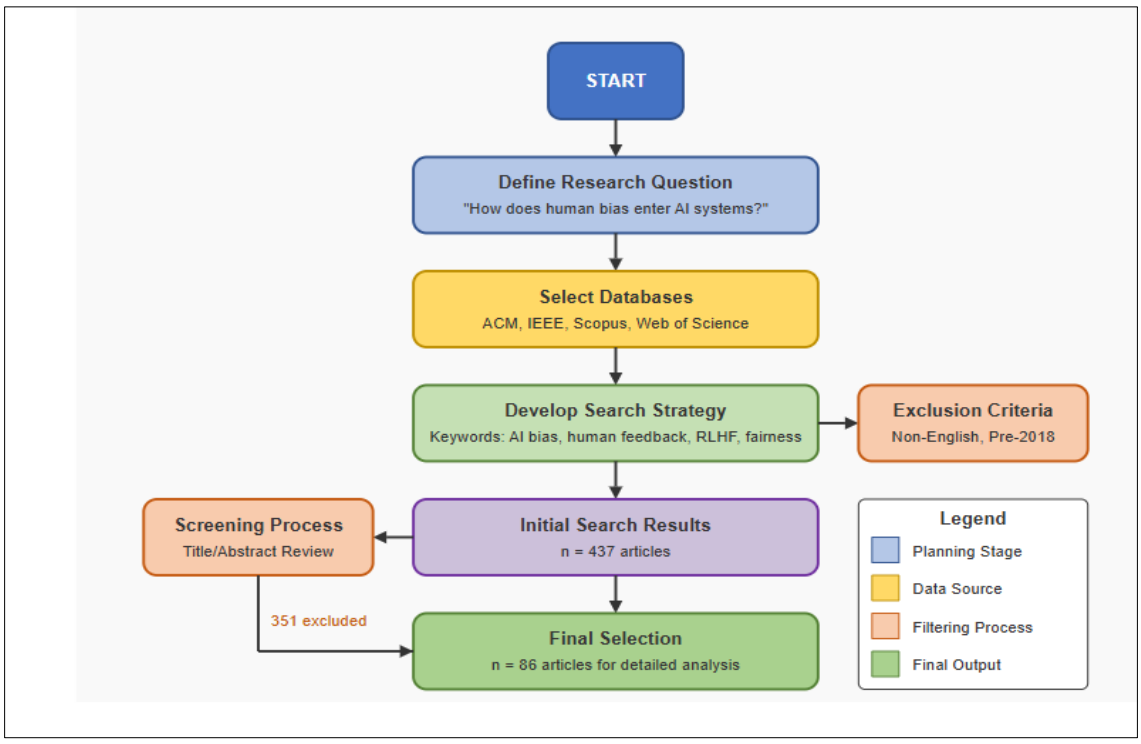

Figure 1 Mapping the Flow of Human Bias in AI Systems [7, 8]

Perhaps most importantly, addressing bias in AI requires collaboration across disciplines. Ethicists can help identify potential concerns, subject matter experts can provide domain-specific context, and AI engineers can implement technical solutions. This collaboration should extend beyond academic boundaries to include community stakeholders affected by AI systems. Mitchell et al. emphasize that algorithmic fairness is inherently a technical and sociolegal challenge, requiring engagement with disciplines ranging from computer science and statistics to sociology, law, and ethics [7]. They note that fairness interventions developed solely within technical frameworks often fail to address the broader systemic factors that give rise to disparities, highlighting the necessity of interdisciplinary approaches that can contextualize technical solutions within social and institutional realities.

The importance of community participation cannot be overstated. Denton et al. argue that dataset development should be recognized as a form of infrastructure building with significant public impact, warranting the same level of oversight and stakeholder engagement as other critical infrastructure projects [8]. Their examination of ImageNet reveals how taxonomic decisions, including categories for "bad person" or stereotypical classifications of people by appearance, encode social hierarchies that subsequently influence AI systems trained on these data. They advocate for community-engaged approaches to dataset creation that recognize how classification systems embed values and potentially reify social categories in ways that can harm marginalized groups. By bringing together diverse expertise, including from communities historically excluded from technology development, AI systems can better navigate the complex ethical terrain of categorization and representation [8].

5. Conclusion

The echo of human bias in AI systems presents a profound ethical challenge beyond technical considerations. As Artificial Intelligence becomes more deeply integrated into society, the four primary channels through which bias infiltrates these systems demand our urgent attention. The training data foundation, with its embedded societal prejudices, language filtering biases, and temporal limitations, shapes AI systems from their inception. Human feedback mechanisms, particularly through RLHF, introduce secondary channels for bias transmission when evaluator pools lack diversity or when documentation standards fail to require demographic representation. Fine-tuning and post-processing stages often lack appropriate oversight, with many organizations having no standardized procedures for tracking fairness impacts as models are refined. The iterative nature of modern AI development creates potential for bias amplification through successive generations, especially as machine-generated content increasingly enters training datasets.

The real-world implications of these biases manifest across critical sectors. In financial services, algorithmic credit scoring systems demonstrably increase disparities between demographic groups, leading to significant financial penalties for minority borrowers. Navigation tools inadvertently reinforce residential segregation patterns while systematically ignoring the needs of disabled users. Healthcare algorithms frequently lack comprehensive fairness testing, leading to treatment recommendations that reflect historical inequities rather than clinical necessity. These concrete consequences underscore how bias in AI can perpetuate and amplify existing social inequalities.

Addressing these challenges requires a multifaceted mitigation strategy. Rigorous bias detection must be integrated throughout the development lifecycle, with fairness diagnostics applied from problem formulation through deployment. Diverse data and feedback mechanisms are essential, particularly to incorporating perspectives from underrepresented communities. Transparent human oversight with clear documentation of normative choices in fairness interventions allows for meaningful accountability. Most importantly, interdisciplinary collaboration that brings together ethicists, domain experts, engineers, and affected communities can help contextualize technical solutions within social realities.

We gain crucial insight into creating more equitable technology by recognizing that AI systems inevitably reflect their creators' values, assumptions, and biases. The path forward requires vigilance in monitoring, humility in recognizing the limitations of purely technical approaches, and an unwavering commitment to equity in AI development. Through deliberate attention to bias at each stage of the AI lifecycle, we can strive to create systems that serve all of humanity, not just those whose voices have historically shaped technological development.

References

[1] Jack Clark and Ray Perrault, "Artificial Intelligence Index Report 2023," Stanford University Human-Centered Artificial Intelligence, 2023. [Online]. Available: https://hai-production.s3.amazonaws.com/files/hai_ai-index-report_2023.pdf

[2] R. K. E. Bellamy et al., "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," IEEE Xplore, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8843908

[3] Emily M. Bender, et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" ACM Digital Library, 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3442188.3445922

[4] Inioluwa Deborah Raji et al., "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing," ACM Digital Library, 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3351095.3372873

[5] Jacob Metcalf, et al., "Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts," ACM Digital Library, 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3442188.3445935

[6] Andreas Fuster, et al., "Predictably Unequal? The Effects of Machine Learning on Credit Markets," The Journal of Finance, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/epdf/10.1111/jofi.13090

[7] Shira Mitchell et al., "Algorithmic Fairness: Choices, Assumptions, and Definitions," Annual Review of Statistics and Its Application, 2021. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-042720-125902

[8] Emily Denton et al., "On the genealogy of machine learning datasets: A critical history of ImageNet," Big Data & Society, 2021. [Online]. Available: https://journals.sagepub.com/doi/10.1177/20539517211035955