Ensuring Transparency and Reducing Bias in AI

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Malek Thabet

Spring 2025

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Pedro A. P. Francisco, Department of Engineering and Society

Introduction:

Artificial Intelligence (AI) is increasingly shaping industries, with 77% of companies using or exploring AI and 83% prioritizing it in their plans (Prestianni, 2024). Technologies such as machine learning (ML) and natural language processing (NLP) are at the forefront of business solutions, transforming how organizations interact with and manage information. Whether AI is being used as an internal tool for companies to sift through thousands of resumes to identify key candidates or algorithms classifying noisy data to create more digestible problems for engineers, AI is not only enhancing efficiency but also massively accelerating data analysis and processing (Fisher, 2024). As this technology rapidly evolves, its integration across industries becomes more widespread, raising critical questions about fairness and the potential for unintended and unaccounted bias. It is reasonable to assume AI operates neutrally, yet real world applications reveal systemic biases that can have far reaching consequences. AI driven decisions in crucial sectors, such as criminal justice and healthcare, can inadvertently reinforce societal inequalities. For example, the COMPAS tool, used to assess the risk of reoffending, has been shown to misclassify risk levels, disproportionately affecting Black defendants (Angwin, 2016). In the healthcare sector, an algorithm prioritized White patients over Black patients with similar risk profiles, exacerbating existing disparities (Obermeyer, 2019). Across industries, there is a growing number of cases where AI systems demonstrate unintended biases with significant consequences. These instances emphasize the urgent need to address algorithmic bias to prevent discrimination in life altering decisions.

This algorithmic bias arises from multiple factors, including skewed training data, flawed model design, and a lack of diversity in AI development teams (Awan, 2023). Without comprehensive frameworks, biased AI systems will persist because of these factors, perpetuating unfair use and fueling public distrust in emerging technology. Furthermore, even addressing algorithmic bias in AI involves a complex network of actors across both public and private sectors. Public sector actors, including government agencies, regulatory bodies and policymakers, play a critical role in establishing and enforcing standards that promote fair use. In the private sector, companies and industry leaders are responsible for integrating these considerations into the implementation of their products and services. The collaboration between these actors is essential to developing comprehensive solutions that address algorithmic bias.

Given these challenges, regulatory bodies must evolve at a pace that matches technological advancements to ensure fairness and transparency without stifling innovation. Establishing structured, adaptable guidelines is essential for promoting ethical AI development. These frameworks should be designed to hold both public and private sector entities accountable. This paper will examine existing regulatory frameworks and proposed solutions for mitigating AI bias, drawing from government reports, academic research, and industry recommendations. By identifying gaps in current policies, this paper will propose a comprehensive structure that reduces bias while promoting transparency in AI systems.

Background / Motivation:

This research seeks to explore the pervasive issue of algorithmic bias in AI systems and its impact on fairness across various industries. As AI continues to play an increasingly central role in business and societal decision-making, addressing the inherent biases within these systems is crucial to ensure equitable outcomes. Understanding how these biases manifest and the real-world consequences they have is essential for reducing bias and promoting transparency in AI development.

Algorithmic bias typically arises from one or more of the following factors: flawed training data, model design, or development processes. Skewed training data is one of the primary causes of algorithmic bias, where historical or societal biases are embedded in the data that AI systems are trained on. Model design flaws, such as the improper selection of features or weighting of parameters, can also lead to biased outcomes. Additionally, the lack of diversity within AI development teams, whether socially, politically or socially, can influence the way algorithms are designed (Awan, 2023). With this, algorithmic bias can theoretically manifest in two forms, direct and indirect. Direct bias occurs when an algorithm explicitly reflects or amplifies a prejudice, whether the data itself is inherently discriminatory, or whether a developer is making a conscious decision. Indirect bias will refer to biases that emerge indirectly due to the way the algorithm processes data, inadvertent oversight from the development team, etc. Effective strategies to identify and mitigate bias must be built on a thorough understanding of the underlying factors (data, model design, development process) and the ways in which they manifest within AI systems (direct and indirect).

The implications of algorithmic bias are far-reaching, affecting various sectors where AI is integrated into decision-making processes. One of the most widely discussed cases is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) tool used in the U.S. criminal justice system. This AI system is designed to predict the likelihood of recidivism among offenders, but investigations revealed that the tool disproportionately flagged Black defendants as high-risk, while White defendants were often classified as lower risk (Angwin, 2016). This racial bias in predictive tools has the potential to perpetuate discriminatory practices in sentencing and parole decisions, leading to unfair outcomes for some communities.

In the private sector, AI systems have also been shown to perpetuate existing disparities. A widely cited study found that an algorithm used to prioritize patients for care disproportionately favored White patients over Black patients, despite both groups having similar levels of medical risk. This discrepancy arose from the fact that the algorithm relied on historical healthcare spending data, which reflected systemic inequities in healthcare access and treatment (Obermeyer, 2019).

Lastly, AI systems have also been introduced in hiring processes, where resume-screening tools have been found to exhibit bias against female candidates. In 2018, Amazon publicly apologized for its internal recruiting tool that was found to improperly favor male candidates for roles in software and technology. Because the model was trained on historical data from a predominantly male dominated field, it penalized resumes that included the word "women's" and even downgraded the graduates of two candidates from an all-women's college (Dastin, 2019).

Illustrated by these three examples, the consequences of algorithmic bias extend far beyond simple technical errors. The broader societal consequences of these biases are profound, affecting everything from job opportunities to healthcare access. In the worst-case scenarios, biased AI systems can promote discrimination and inequality. Just as importantly, it deters trust in technology that can revolutionize all aspects of life. Given the increasing reliance on AI for critical decision-making, addressing these biases is essential not only for fairness but also for safeguarding human trust within the technology.

Addressing algorithmic bias in AI is an urgent and essential task, given the technology's pervasive role in shaping decision-making across industries both public and private. By conducting this research, the aim is to offer a comprehensive analysis of the causes and consequences of algorithmic bias, as well as proposed solutions to mitigate its effects in whatever shape it takes. This work is critical because biased AI systems have the potential to worsen existing inequalities, while reducing public interest in revolutionizing technology. By exploring the regulatory frameworks that currently govern AI and identifying gaps in those regulations, this research seeks to contribute to the development of more effective policies and practices that promote fairness and transparency in AI systems against all forms of bias. Ultimately, the goal is to provide a framework that not only addresses bias but also encourages innovation to help limit these popular case studies of algorithmic bias and lack of transparency.

Methodology:

This paper adopts the Responsible Research and Innovation (RRI) framework to examine the societal impacts of algorithmic bias in AI systems. RRI's focus on ethics, inclusivity, and foresight enables a nuanced understanding of how AI technologies can intentionally and unintentionally create social inequalities through bias (Iordanou et al., 2025). To establish a framework of principles to guide and ensure fairness and transparency in the public and private sectors, public information will be gathered and analyzed from the perspective of an RRI lens.

A qualitative approach will be employed, combining government reports, academic literature, and industry documentation. By analyzing sources such as the article "Regulating AI: Opportunities to Combat Algorithmic Bias and Technological Redlining", the research will investigate the intersection of AI and regulatory bodies. Most importantly, the article reviews the specifics of former President Biden's executive order on AI. Specifically, the executive order, however now rescinded, outlines several general strategies to promote safety and security of AI. Most notably, the order mandates evaluating how federal agencies collect information to understand risks arising from AI, guidance to landlords and federal contractors on combating bias, facilitating training assistance for federal departments, and more (Lorch, 2023). This article will be instrumental in framing the public regulatory perspective.

For general guidance for private companies, this paper will examine industry efforts, drawing on the analysis provided in the Harvard Business Review (Manyika, 2022). This work highlights the prevalence of biases in AI systems and discusses strategies for correcting these biases. Outlined within the article are seven elements, ranging from improving data quality, incorporating diverse perspectives, and increasing transparency in AI. These seven elements can be reviewed and analyzed from the perspective of RRI, ensuring stakeholder engagement to ensure their relevance. However, while these seven elements will offer an appropriate starting point, they lack actionable guidelines and tangible suggestions for minimizing bias and promoting transparency.

More actionable strategies will be drawn upon from the findings of the "Addressing Bias in AI" report, which examines real-world case studies. This resource will be pivotal in identifying regulatory gaps and shortcomings in current AI fairness frameworks by combining it with the RRI framework. The article discusses the guidelines, techniques, and tools that must be adopted in response to emerging legislation and trends in AI ethics. In the guidelines is an emphasis on integrating ethics throughout the AI lifecycle, and “ensuring inclusive access to and participation in the development of AI”. As for techniques and tools, the outline stresses the necessity for internal and external audits of AI development processes, which will be initially challenging for businesses to accept with external audits seemingly a threat to proprietary technology. Additionally, tools for constructing fair AI range on the scale of technological complexity, but most important is the need for explainable AI (Matthews & Murphy, 2023).

These guidelines, techniques and tools can be compared against current regulatory orders, such as President Biden's mentioned above, and outlines in academic literature.

The integration of RRI with these diverse resources allows for a comprehensive analysis of AI regulation, providing insights into how these frameworks can better account for the ethical, social, and technological challenges posed by AI. The comparative analysis will highlight both successful approaches and gaps, enabling the development of a more inclusive, reflexive, and anticipatory regulatory environment for AI.

Literature Review:

This literature review aims to explore the current state of knowledge on algorithmic bias, the factors that seemingly contribute to it, and existing regulatory frameworks designed to address these issues.

Algorithmic bias stems from several factors, including biased data, flawed model design, and a lack of diversity in the team's responsible for developing AI systems. Awan (2023) offers a comprehensive overview of how algorithmic bias arises in AI systems, outlining the role of biased training data. AI algorithms are trained on historical data that may contain embedded biases reflecting societal inequalities. Awan's work details various types of bias, including pre-processing bias, confirmation bias, exclusion bias, and model bias. These biases may arrive from data cleaning processes, training data development, or favoring certain outputs, indicating that bias can arise at truly any stage of AI development. These biases complicate efforts to ensure fairness and equity in AI systems.

GeeksforGeeks (2024) also identifies types of biases that can emerge throughout the lifecycle of AI systems: sampling bias, algorithmic bias, and confirmation bias. While confirmation bias has been previously discussed, this article introduces additional biases that can arise during development and deployment. Sampling bias occurs when the data used to train an AI system is not representative of the population it is intended to serve. Additionally, algorithmic bias arises when the design or structure of the algorithm itself inadvertently leads to discriminatory outcomes, which may occur without direct intention or intervention from the developers. As highlighted in these two articles, the development of AI is an intricate and highly cautious process, where each critical decision carries the potential to undermine the integrity of the entire project.

When these projects are not approached with the necessary attention, they inevitably lead to models that not only underperform but also compromise user trust and outcomes. Obermeyer (2019) demonstrate how an algorithm that relied on historical healthcare spending data reflected systemic disparities in healthcare access, a byproduct of sampling bias.

In addition to biased data, the lack of diversity within AI development teams is a critical issue. Hao (2020) discusses the "White Guy Problem" in AI, highlighting the significant underrepresentation of women and minorities in AI development roles. This lack of diversity among AI researchers, engineers, and developers contributes to the perpetuation of biases in AI systems. For instance, as noted by Hao, women make up only 18% of authors at leading AI conferences, and Black workers constitute just 2.5% of Google’s workforce. This demographic homogeneity can lead to the creation of systems that reflect the perspectives and biases of a narrow group, subjecting systems to unintended biases because of alike backgrounds.

Recognizing the importance of addressing algorithmic bias, scholars and policymakers have proposed various frameworks for regulating AI and promoting fairness. Lorch (2023) examines the regulatory approaches taken by the U.S. government, particularly focusing on former President Biden’s executive order on AI. While repealed by the current administration, the order mandated federal agencies to evaluate AI risks, provide guidance on combating bias, and establish training programs for federal employees. It represents a significant step toward addressing algorithmic bias at the public sector level. Lorch argues that such efforts can help mitigate technological redlining and promote fairness, especially in sectors like housing and criminal justice.

In terms of regulation from industry leaders and private institutions, Matthews and Murphy (2023) discuss the importance of AI explainability, which allows stakeholders to understand how AI systems make decisions. They emphasize that transparent AI systems enable users to better assess the fairness of decisions, particularly in high-stakes areas. The report stresses the need for regular audits and third-party evaluations to ensure that AI systems operate fairly and without the many biases mentioned earlier. They believe integration of explainability techniques into AI development processes will promote accountability and trust. Manyika (2022) offers guidelines to reduce bias in AI systems, emphasizing better data quality, diverse teams, and increased transparency. He advocates for internal audits and stakeholder engagement. To help promote this transparency, Mitchell (2019) introduces "model cards," which provide detailed information about a model's performance across demographics, helping users understand potential biases.

The literature on algorithms and their bias reveals the multifaceted nature of the problem, as well as the urgent need for regulatory frameworks. Regulators and industry leaders have identified several issues related to the cause of AI bias, largely like the issues outlined before. These regulators and leaders have proposed various regulatory frameworks, guidelines, tools, and techniques, including government mandates, internal audits, and AI explainability documentation. Drawing on the RRI approach and the frameworks and underlying issues discussed in the preceding literature, it is essential to establish a comprehensive and adaptable set of guidelines.

Discussion / Results:

The evidence collected in the previous section reveals that the persistence of bias in AI systems stems from several factors including skewed data, flawed model design, and a lack of diversity within development teams. As concerns grow over the misuse of AI in both public and private sectors, questions around transparency between developers and users are becoming more urgent. An RRI approach offers a lens through which to examine AI development, aiming to foster equitability and openness in decision-making processes. By adopting this, the results emphasize the need for ethics, inclusivity, and foresight in AI governance. Both public and private sectors need to engage in collaborative efforts that prioritize transparency and inclusivity, thus ensuring that AI technologies contribute positively to society while minimizing harmful biases.

This framework is grounded in key principles derived from the RRI approach and existing literature, ensuring the mitigation of bias in model development, the creation of explainable documentation for these models, and the enforcement of these guidelines to uphold these standards throughout their lifecycle.

Firstly, developers must improve data quality as is essential. This involves ensuring that training data for any tool used externally or for external purposes is representative and free from embedded societal biases, as well as regularly auditing data to detect and correct any disparities that may arise. Data must be conducted by external teams who are not directly involved in the development process. These auditors may be independent enough to avoid conflicts of interest, or internal to protect proprietary information. Regardless, they must be familiar with the system to effectively assess its fairness and accuracy while not being overly influenced by prior knowledge. As much as possible, businesses most anticipate these audits and ensure the audibility of their projects and tools if third parties are necessary. Lastly, with much of the auditing framework currently revolving around the conclusion of technology, it is essential that these are performed sequentially throughout the entirety of the AI lifecycle.

Secondly, increasing diverse perspectives during AI development is crucial. A more diverse team will help identify potential biases and ensure that a broader range of perspectives is considered during the development process. Although recent rulings and policymakers nationwide have rejected diversity mandates, there are still effective strategies that can be implemented. Blind resume screenings, diverse interview panels, and inclusive language will all help promote fairness within the workplace itself. Furthermore, it is necessary to invest in educational opportunities that help make a more diverse AI pipeline (AI4ALL, Partnership on AI, the Inclusive AI Foundation). To further promote this diversity at the early stages of development, encourage stakeholder engagement and participatory design during the design process of the AI lifecycle.

Thirdly, transparency in AI decision-making processes is vital. Tools like "model cards" must be utilized to provide detailed information about how AI systems perform. This is a framework aimed at promoting transparency in AI decision-making processes. Model cards are concise documents accompanying trained models, providing benchmarked evaluations across various conditions, such as different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, Fitzpatrick skin type) and their intersections. They also disclose the context in which the models are intended to be used, the performance evaluation procedures, and other relevant details. By offering documentation to create explainable AI, model cards help users understand the limitations and potential biases of AI systems. This framework, which primarily applies to human-centered machine learning models, can be adapted for any trained model. This documentation represents a significant step toward the responsible democratization of machine learning and AI technologies, increasing transparency about how well these systems work and encouraging others to include similar evaluations when releasing models (Mitchell, 2019).

Lastly, in the U.S., the Federal Trade Commission (FTC) and the Equal Employment Opportunity Commission (EEOC) must be at the forefront in overseeing and regulating algorithmic bias, each with distinct responsibilities. The FTC, under Section 5 of the FTC Act, can address biased algorithms by treating them as unfair or deceptive practices, particularly when they cause substantial harm to consumers or mislead them. In doing so, the FTC encourages the importance of testing, transparency, and inclusive data sets to prevent legal challenges. The EEOC, on the other hand, focuses on ensuring that algorithms do not facilitate employment discrimination under Title VII of the Civil Rights Act, specifically targeting both disparate treatment and disparate impact. The EEOC’s "four-fifths rule" provides a framework for assessing biased outcomes in employment decisions, and their recent initiatives highlight the need for fairness in AI-driven hiring tools.

As the legal landscape surrounding AI continues to evolve, it will eventually become more defined and concrete. When this occurs, the Department of Justice (DOJ) is likely to take on the responsibility of overseeing AI-related malpractice, ensuring fairness and transparency across technology’s applications. To effectively carry out this crucial role, it is essential that the DOJ establishes and develops a dedicated, prepared AI-focused unit to enforce these emerging challenges and guidelines.

Conclusion:

Addressing algorithmic bias in AI systems is an urgent and necessary task to ensure fairness and transparency across industries. The pervasive nature of AI in critical decision-making processes, such as criminal justice, healthcare, and hiring, underscores the significant consequences of biased algorithms. Algorithmic bias stems from a variety of issues that must be mitigated to build a comprehensive framework. It is crucial to adopt a holistic approach that involves improving data quality, increasing diversity in development teams and throughout the AI lifecycle, ensuring transparency in AI decision-making through explainable AI, and integrating robust regulatory frameworks. Government agencies like the FTC and EEOC, under the current legal landscape, will play pivotal roles in enforcing fairness in AI. Furthermore, the DOJ must develop a dedicated AI unit to address and monitor malpractice as the legal landscape evolves to be best prepared to handle this rapidly evolving technology. By drawing on frameworks like RRI, this research aims to contribute to the development of comprehensive, adaptable guidelines that promote fairness while fostering innovation. Ultimately, addressing algorithmic bias by adopting this outlined framework will not only safeguard against discrimination but also help restore public trust in AI technologies, ensuring their positive impact on society.

References:

1. Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016, May 23). *Machine bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
2. Awan, A. A. (2023, July 17). *What is algorithmic bias?*. DataCamp. https://www.datacamp.com/blog/what-is-algorithmic-bias
3. Dastin, J. (2018, October 10). *Insight - Amazon scraps secret AI recruiting tool that showed bias against women.* https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/
4. Fisher, M. (2024, October 15). *AI for Business Automation: Transforming Efficiency, cost savings, and customer experience*. Models AI. https://modelsai.org/blog/ai-for-business-automation
5. GeeksforGeeks. (2024, April 24). *Fairness and bias in Artificial Intelligence*. https://www.geeksforgeeks.org/fairness-and-bias-in-artificial-intelligence/
6. Iordanou, K., & Antoniou, J. (2022, December 8). *Research* |. UCLan. https://clok.uclan.ac.uk/44322/
7. Lorch, K. (2023). *Regulating AI: Opportunities to combat algorithmic bias and technological redlining | Journal of Public and International Affairs*. Princeton University. https://jpia.princeton.edu/news/regulating-ai-opportunities-combat-algorithmic-bias-and-technological-redlining
8. Manyika, J., Silberg, J., & Presten, B. (2022, November 17). *What do we do about the biases in ai?*. Harvard Business Review. https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

9. Matthews, V., & Murphy, M. (2023). *Reuters: Addressing Bias in Artificial Intelligence*.
10. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3287560.3287596
11. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342
12. Prestianni, T. (2024, November 7). *131 AI statistics and trends for (2024)*. National University. https://www.nu.edu/blog/ai-statistics-trends/