Bias and Fairness in AI

Chen Yu Hang (Leader) 1211107973 Problem Statement, Literature Review, References
Vishnunanda Surain 241UC241C3 Executive Summary, Research Questions, Hypothesis and Objectives, and Flowchart of Research Activities
Vinnosh Rau A/L Samudram 1211108264 Introduction, Expected Results and Impact
Thong Yun Peng 1211107295 Research Methodology, Research Activities and Milestones

September 25, 2024

Executive Summary

As AI becomes increasingly embedded in decision-making across areas like healthcare, lending, and criminal justice, concerns about fairness and bias in these systems have grown. This research focuses on identifying where bias comes from—whether it's in the data, the models, or human involvement—and finding ways to reduce it. We will use tools such as IBM's AI Fairness 360, Google's What-If Tool, and Microsoft's Fairlearn to tackle these biases through various techniques that clean up data, adjust models during training, and refine results after training. By combining technical approaches with ethical considerations, we aim to develop AI systems that are not only accurate but also fair and just. The anticipated result is AI models that deliver fairer decisions, helping to reduce the inequalities AI can sometimes reinforce, particularly among marginalized groups.

Keywords: Bias, Fair AI, Mitigation Strategies, Ethical Technology

1 Introduction

Artificial Intelligence (AI) has become a critical component in various industries, transforming decision-making processes through automation and machine learning. AI systems are being increasingly used in sectors such as healthcare, finance, hiring, and criminal justice to streamline tasks and make complex decisions efficiently. However, significant concerns have been raised about bias and fairness in these systems. If the data used to train AI models is biased, these systems can unintentionally reinforce societal inequalities. This issue arises because AI models often rely on historical data, which may reflect existing prejudices and systemic biases. According to Kate Crawford, a leading researcher in AI ethics, "data is a reflection of our past, and if that past is discriminatory, then it will have discriminatory outcomes" (Crawford, 2021). The widespread application of AI, coupled with its potential to shape critical decisions, has prompted an increasing focus on understanding and mitigating bias within these systems.

Despite advances in AI research, the issue of bias remains a persistent challenge. Many current studies focus on technical aspects, such as creating fairness-aware algorithms and post-deployment bias detection methods. For instance, Cynthia Dwork, a pioneer in the field of algorithmic fairness, has developed methodologies to create more equitable algorithms, but she also acknowledges that "fairness is a complex and context-dependent issue that cannot be entirely solved by technology alone" (Dwork, 2012). These technical solutions are often insufficient, as they do not fully address the root causes of bias, which may stem from the data collection, pre-processing, or model design phases. Ruha Benjamin, a sociologist and scholar on race and technology, highlights the disconnect between these technical solutions and their real-world impact, particularly on marginalized communities. She argues that "without understanding the societal and structural roots of discrimination, technological fixes will continue to fall short" (Benjamin, 2019).

This research aims to bridge that gap by exploring both the technical and ethical dimensions of AI fairness. By integrating interdisciplinary approaches, it seeks to contribute to the development of AI systems that are not only efficient but also equitable.

2 Problem Statement

Our research focuses primarily on how AI algorithms compile data before making an unbiased and fair decision and ways to improve them. These algorithms, often trained on historical data, can inadvertently reflect and amplify existing societal biases, leading to unfair outcomes in areas such as hiring, lending, and law enforcement. Ensuring fairness in AI systems is imperative to curb discrimination and maintain public trust. However, identifying, mitigating, and preventing bias in AI remains a complex predicament that requires balancing technical solutions with ethical considerations. Developing robust methods to address bias and ensure fairness is essential for creating equitable and trustworthy AI systems.

3 Research Questions, Hypotheses and Objectives

Minimum of two and maximum of four for each of the research questions, hypotheses, and research objectives. You can use numbering/bullet point to list them out.

Research Questions:
1. In what ways do biases in data, algorithms, and human judgment lead to unfair results in AI systems?
2. How do biases in generative AI shape societal perceptions, and what factors contribute to this issue?
3. Are the current methods for reducing bias in AI effective at addressing inequalities?
4. Can tools like AIF360 help promote fairness in AI systems, especially in critical fields like healthcare?

Hypotheses:
1. AI models trained on biased data are more likely to perpetuate societal inequalities.
2. Techniques like data preprocessing and post-processing can meaningfully reduce bias in AI.
3. Generative AI is more prone to reinforcing stereotypes because of the biases in its input data.
4. Using fairness-focused tools like AIF360 will lead to more equitable decisions in AI systems.

Research Objectives:
1. To uncover the sources of bias in AI systems and understand their effects on decisions.
2. To analyze how well different strategies for reducing bias work in creating fairer outcomes.
3. To examine how biased AI affects society, particularly in areas like healthcare and law enforcement.
4. To explore how tools like AIF360 can be applied to reduce bias in AI systems used in real-world scenarios.

4 Literature Review

Based on the selected articles, bias and fairness in AI is paramount if humans are to incorporate technology further into daily activities. AI is defined as makers of intellectual machines and programs. The articles focus on the various ways data is compiled and integrated into AI algorithms to aid the decision-making procedure. Machine learning fairness addresses and eliminates algorithmic bias from machine learning models based on sensitive attributes like race, gender, sexual orientation, disability, and socioeconomic class.For instance, a dataset showed females receiving more favorable credit approvals compared to men. Bias in AI can perpetuate and even exacerbate existing inequalities, leading to discrimination against marginalized groups and limiting their access to essential services. In addition, bias in AI can also lead to new forms of discrimination, such as those dependent on skin color, ethnicity, or even physical appearance. The articles disclosed that these data and algorithms can be manipulated, altered and optimised by humans thanks to tools like IBM's AI Fairness 360, Google's What-If Tool as well as Microsoft's Fairlearn, Furthermore, researchers and practitioners have postulated various approaches to alleviate bias in AI. These approaches involve pre-processing data, model

selection, and post-processing decisions.

The collective critical analysis of the research papers has their fair share of pros and cons. As a whole, they state the different kinds of biases which plays a pivotal role in the decision-making procedure such as historical bias, aggregation bias and what not. Besides, they also offer an accessible explanation of how biases arise in AI algorithms, particularly through data selection and model training. Furthermore, the paper recommended a myriad of ways to mitigate AI machine learning which can to be conducive and pragmatic. However, they lack extensive real-world proof on the effectiveness of the proposed mitigation techniques. Additionally, they don't dive into how to apply and assess fairness constraints or use adversarial debiasing in real-world settings. All in all, further research has to be conducted as well as comprehensive data compilation before concluding.

5 Research Methodology

This section describes the steps involved in the research methodology, including the techniques, models, methods, and algorithms used, along with metrics for evaluation. The methodology integrates technical and ethical perspectives to address the issue of bias in AI systems comprehensively.

5.1 Overview of Research Methodology

The research aims to address biases in AI systems by evaluating various mitigation strategies and testing their effectiveness in real-world applications. The approach integrates technical and socio-ethical perspectives to develop a holistic solution for bias and fairness issues in AI.

5.2 Rationale for Method Choices

The selection of tools like IBM AI Fairness 360, Google What-If Tool, and Microsoft Fairlearn is based on their effectiveness in identifying and mitigating various types of biases in AI systems. These tools provide comprehensive frameworks for detecting bias and implementing fairness constraints. Their real-world applicability and the robust frameworks they offer for bias detection make them ideal for this research.

5.3 Research Steps

The methodology follows a series of structured steps aimed at identifying, mitigating, and evaluating bias in AI systems:

1. **Data Collection:** Gather datasets from reliable sources like the UCI Machine Learning Repository, Kaggle, and domain-specific datasets. Ensure data preprocessing by cleaning, normalizing, and handling missing values to maintain dataset quality and reliability. Consider privacy and ethical concerns when collecting sensitive data.

2. **Identification of Bias:** Use tools like IBM AI Fairness 360, Google What-If Tool, and Microsoft Fairlearn to detect biases in datasets and AI models, focusing on biases such as historical bias, aggregation bias, and measurement bias.

3. **Bias Mitigation Techniques:** The following bias mitigation strategies will be implemented:

    - *Pre-processing Methods*: Modify the data before training, such as by reweighting or resampling data to reduce bias.
    - *In-processing Methods*: Adjust the training algorithms directly, such as using adversarial debiasing to reduce bias during model learning.

- *Post-processing Methods*: Modify the model outputs after training, such as recalibrating predictions to ensure fairness.

4. **Model Training and Evaluation:** Train AI models using both original and mitigated datasets. Models like Logistic Regression, Random Forest, and Neural Networks will be trained. Model performance will be evaluated using metrics like accuracy, Statistical Parity Difference, and Equalized Odds to balance both performance and fairness.

5. **Comparative Analysis:** Use statistical methods such as t-tests or ANOVA to compare the performance of the bias-mitigated models against baseline models. This will help quantify the effectiveness of the mitigation techniques.

6. **Ethical Evaluation:** Collect feedback from stakeholders, including AI practitioners and affected communities, to assess the ethical implications of each bias mitigation technique. This step ensures that the chosen strategies are socially responsible, beyond just technical effectiveness.

5.4 Evaluation Metrics

The evaluation of bias mitigation techniques will rely on the following metrics:

- **Statistical Parity Difference:** Measures the difference in positive outcome rates between protected and unprotected groups.
- **Equalized Odds:** Assesses whether false positive and false negative rates are equally distributed across different groups.
- **Accuracy:** Measures overall model performance but will be considered alongside fairness metrics to avoid over-optimization at the expense of fairness.
- **False Positive Rate Disparity:** Examines disparities in false positive rates across different demographic groups.
- **Equal Opportunity Difference:** Evaluates whether the true positive rate is equal across different groups.

5.5 Potential Limitations and Mitigation

While bias mitigation techniques are promising, they come with limitations:

- **Computational Costs:** Some methods, particularly in-processing techniques like adversarial debiasing, are computationally expensive. This will be mitigated by selecting models and algorithms optimized for efficiency.
- **Data Distortion:** Pre-processing techniques such as reweighting may distort the original dataset. Care will be taken to ensure minimal alteration while maximizing fairness.
- **Real-world Applicability:** Some techniques may work well in controlled environments but may face challenges in real-world deployment. To address this, real-world testing and stakeholder feedback will be integrated into the methodology.

5.6 Visual Representation of Methodology

The research methodology will be summarized in a flowchart, detailing the sequence from data collection to ethical evaluation. This visual representation will clarify the workflow and inter-dependencies between steps.

5.7 Ethical Considerations

An essential aspect of this research is the ethical evaluation of bias mitigation strategies. Stakeholder feedback, including input from marginalized communities, will be collected through surveys or interviews. This ensures that the AI systems are not only technically sound but also align with ethical standards such as fairness, transparency, and accountability.

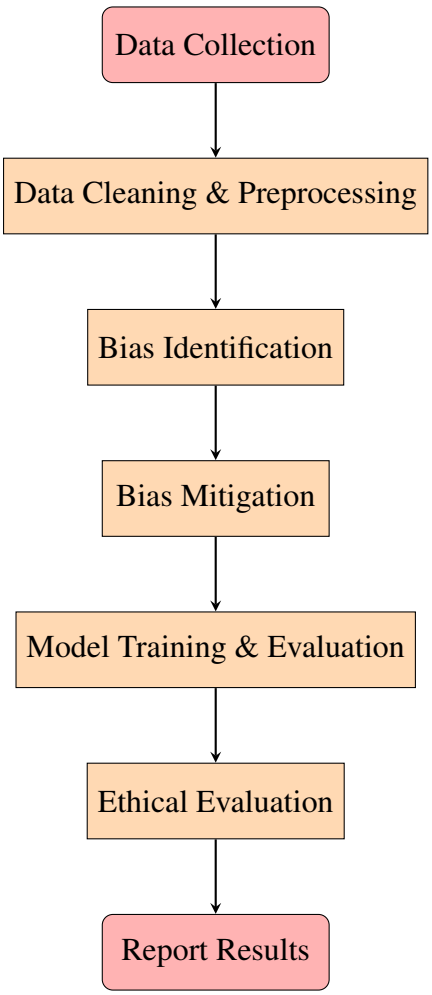
Table 1: Comparison of Bias Mitigation Techniques

Technique	**Description**	**Advantages**	**Disadvantages**	**Example Application**
Pre-processing	Modifies data before training, e.g., reweighting samples	Easy to implement	Can distort original data	Healthcare data rebalancing
In-processing	Changes the algorithm during training, e.g., adversarial debiasing	Directly addresses bias during learning	Computationally intensive	Fair neural network training
Post-processing	Adjusts predictions after training	Doesn't alter the model	Less control over decision-making	Credit approval recalibration

6 Research Activities and Milestones

This section describes the research activities and their timeline. The flowchart illustrates the sequence of activities, while the Gantt chart provides a clear timeline for completing each task within 7 weeks.

Flowchart of Research Activities

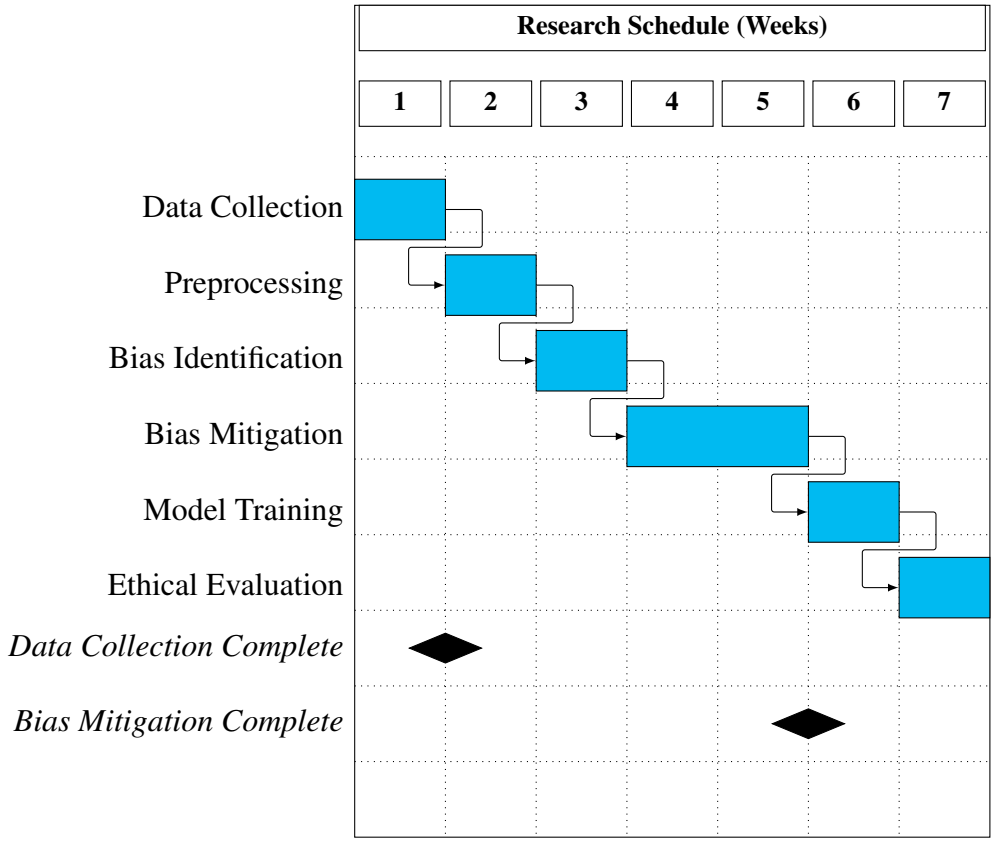


Explanations of the Research Activities

1. **Data Collection:** Gather data from reliable sources to ensure it is diverse and representative.
2. **Data Cleaning & Preprocessing:** Remove errors and inconsistencies to prepare the data for model training.
3. **Bias Identification:** Use tools like IBM AI Fairness 360 to detect biases in data and models.
4. **Bias Mitigation:** Reduce bias by adjusting data, algorithms, or model outputs.
5. **Model Training & Evaluation:** Train the AI model using refined data and evaluate it for performance and fairness.
6. **Ethical Evaluation:** Gather feedback from stakeholders to ensure the model is fair and ethically sound.
7. **Reporting Results:** Share findings on bias mitigation, model performance, and ethical considerations.

Gantt Chart for Research Timeline

The Gantt chart below outlines the timeline for the project, spanning 7 weeks. Each task is mapped to a specific week, and milestones indicate key achievements.



7 Expected Results and Impact

Improved Understanding of Bias Origins: The analysis will uncover the various sources of bias in AI systems, including those embedded in the data, algorithms, and human decision-making processes. This will lead to a deeper comprehension of how biases manifest in real-world AI applications.
Effectiveness of Mitigation Strategies: The study will provide a comparative analysis of different bias mitigation techniques (pre-processing, in-processing, and post-processing), evaluating their effectiveness in enhancing fairness across AI models in critical areas such as healthcare and criminal justice.
Role of Fairness Tools: By applying tools like IBM AI Fairness 360 and Microsoft Fairlearn, the research will demonstrate the practicality and limitations of these tools in addressing bias in real-world scenarios. This will offer actionable insights into how these tools can be integrated into AI development workflows to ensure fairness.

References

- Lory Seraydarian (2023), Bias and Fairness in AI Algorithms, plat.ai/blog/bias-and-fairness-in-ai-algorithms/
- Luca CM Melchionna (2023), Bias and Fairness in Artificial Intelligence, nysba.org/bias-and-fairness-in-artificial-intelligence/
- Emilio Ferrara (2023), Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies, arxiv.org/pdf/2304.07683
- Jpn J Radiol (2024), Fairness of artificial intelligence in healthcare: review and recommendations, www.ncbi.nlm.nih.gov/pmc/articles/PMC10764412/
- Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.
- Dwork, C. (2012). *Fairness Through Awareness*. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12).
- Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press.