Developing and Deploying Robust Fairness-Aware AI Algorithms

CS4991 Capstone Report, 2024

Matthew Cahill
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
msc3jk@virginia.edu

ABSTRACT
Limited practices currently exist showing how to effectively prevent and manage algorithmic biases in AI system development. This challenge stems from AI algorithms trained on historical data, which often propagate societal biases that disproportionately impact historically disadvantaged populations. The solution involves a multi-faceted framework, including an AI algorithm to identify and mitigate biases, a monitoring system for emergent biases, as well as a prioritization of educating stakeholders on fairness-aware AI. The anticipated outcomes include a comprehensive framework for developing and deploying fairness-aware AI algorithms, working to improve trust and equity in future AI models and systems. Future work concerning the actual development, training, and deployment of these AI models is necessary, effectively moving to practical usage beyond theoretical frameworks.

1. INTRODUCTION

AI is quickly becoming intertwined with every facet of life, and software developers must be prepared to harness its power for societal benefit. The technical research project goals include the following structure: 1) an AI algorithm that can identify and mitigate biases in its predictions; 2) a monitoring system that can detect emergent biases in deployed models; and 3) a research project that will educate stakeholders on the vital importance and methods of ensuring fairness-aware AI.

The project's unusual constraints are as follows: First, the inner workings of fairness-aware algorithms must be transparent so their decision-making processes can be analyzed and justified. Second, any method being used to detect bias must ensure that the data of individuals is kept confidential. Third, as AI interacts with every corner of society, tailoring the training of algorithms to accurately consider unique situational biases is computationally expensive.

This project's scope will not involve producing a full-fledged AI model that accurately and efficiently removes all bias and discriminatory practices from its decision-making processes. Rather, it will focus on the theoretical framework behind developing the system and the conceptual inner workings of the algorithm training. This begins with a comprehensive review of fairness models and the current shortcomings of AI system decision-making with respect to specific demographics of historically targeted populations. The framework will cover data collection, model development, real-time monitoring and adjustment practices, and stakeholder education.

By the end of the proposed project, a tangible framework fully covering the necessary steps to achieve the development and deployment

of robust fairness-aware AI algorithms will have been produced. The impact of this research project will improve the general trust in AI systems as they will be deployed and operated equitably. Next, this newfound AI framework will be developed, trained, and deployed to further promote a safer AI future.

2. RELATED WORKS

The early workings of present-day, newly designed AI algorithms, which are trained on historical data, have already been shown to propagate and even exacerbate existing societal biases. An MIT study showed that commercial AI facial recognition tools developed by major tech companies such as IBM and Microsoft had "higher error rates for darker-skinned individuals" (Buolamwini, 2018). A similar study conducted by ProPublica highlights how certain AI models developed for use in the United States criminal justice system are "biased against African American individuals" in how they work to "detect future criminals" (Angwin, 2016). Addressing potential bias in AI system development is not only an ethical imperative, but also critical in ensuring the reliability and trustworthiness of this technological feat.

The current methods being used to address AI fairness include adversarial training, adjusting weights in training data, and post-hoc calibration (Bellamy, 2018). Tech companies deploy tools such as Google's TensorFlow Fairness Indicators to appear as though they are actively avoiding bias and discrimination in their systems. The minimal solutions that Google and other big tech AI development companies deploy are reactive, meaning they address biases once identified. While tools such as TensorFlow have a variety of fairness metrics, little to no guidance exists regarding which of these metrics are optimal. This project will address how to better combat these biases in the form of prevention and limitation rather than reaction. Structuring AI algorithmic training such that it has built-in mechanisms and strategies to eliminate biases is the foundation of this technical project. The contribution of this project's findings will work to better the sociological landscape through minimizing the damage that historically disadvantaged and vulnerable populations receive from untrustworthy AI systems.

3. PROPOSAL DESIGN

This section delves into a comprehensive design proposal centered around integrating fairness and equity into artificial intelligence systems. The approach encompasses advanced algorithms for bias detection and mitigation, continuous system monitoring, and stakeholder education.

3.1 Project Design Overview

The proposed design of the project is based on a multi-layered approach aimed at addressing the challenge of fairness in AI. This encompasses the development of an AI algorithm that is capable of identifying and mitigating biases in its predictions using a monitoring system designed to detect the earliest trace of emergent biases in deployed models. A critical component of the project design pertains to an educational program built for stakeholders with the goal of raising awareness and general understanding of the inner workings and necessity for fairness-aware AI. The overall aim of this project is to develop a scalable and adaptable framework that can be applied to various domains, ensuring that artificially intelligent systems are equitable and do not perpetuate or exacerbate existing societal biases.

3.2 AI Algorithm for Bias Identification and Mitigation

The optimal algorithm associated with this project will incorporate a design that is specifically engineering to detect and correct

biases in its predictions and output. The first step in the algorithm's design involves data preprocessing and bias detection. Statistical analysis and visualization techniques will be used to study data distributions across a variety of demographic groups. When patterns of imbalances in the analysis arise, said issues will be flagged and investigated thoroughly. The basis of the algorithm is rooted in training a machine learning model with fairness constraints integrated into the objective function. Fairness metrics including demographic parity and equalized odds will be incorporated into the way these constraints will be formulated. The inclusion of these constraints in the algorithm's foundational makeup will lead to prediction optimization in that the generated output of the system will not only be accurate, but fairness aware as well.

To both ensure that the given model generalizes well to unseen data and prevents overfitting, regularized optimization techniques will be used. These techniques will come in the form of L1 and L2 regularization. The former involves adding the absolute value of magnitude of the coefficient as a penalty term to the loss function. The latter adds the squared magnitude of the coefficient as the penalty term to the loss function. The deployment of these techniques is meant to balance the trade-off between model complexity and fairness. Following the deployment of this algorithm, it will be continuously monitored for any potential emerging bias. If new biases are detected, the model itself will be reevaluated and adjusted in a manner that allows for the correction of the training process and post-processing steps.

3.3 Educating Stakeholders

An education program will be crafted for stakeholders, which includes that of developers, policymakers, and end-users. Said program will be tailored to cover topics relating to the importance of fairness-aware AI systems, the potential risks associated with the failure to incorporate corrections to inherently biased algorithms, and the best actions that can be taken to enact positive change regarding the development and deployment of fairness-aware AI. The program will work to raise awareness of the ethical implications that come with the rapid development and deployment of artificially intelligent systems and the necessity for fairness and equity to an integral part of this technology's creation process. Guidelines and resources will be provided to encourage healthy dialogue and collaboration among stakeholders.

4. ANTICIPATED RESULTS

The anticipated result of the proposal design's implementation includes the development and deployment of a comprehensive framework for a fairness-aware AI system based on a well-trained algorithm. The outcome will be designed to effectively detect and mitigate bias in its predictions, provide an early detection system for emergent bias in deployed models, and educate stakeholders to raise awareness and understanding regarding fairness-aware AI. The algorithm itself is expected to significantly reduce biases in generated outputs, effectively ensuring fair predictions that are rooted in equity and perspective in terms of respecting historically marginalized groups. The expectations associated with the monitoring system are aligned with the creation of a continuously updating assessment that will provide real-time adjustments in biases connected to the algorithm's output. Educating stakeholders will play a critical role in the deployment of this project plan as developing a general understanding of the innerworkings of fairness-aware AI is foundationally necessary in benefitting society's most underprivileged.

The results of this project are expected to offer a unique perspective regarding AI ethics to the general field of researchers conducting work to combat harmful biases in poorly constructed AI algorithms. This research will enhance trust in AI systems after its findings are incorporated into subsequent generations of AI algorithm development.

5. CONCLUSION

This project represents a necessary step in the direction of integrating fairness and equity into the foundational makeup of AI systems. By designing a multi-faceted framework that includes an AI algorithm for identifying and mitigating biases, a monitoring system that detects emergent biases, and an educational program for stakeholders, this research can be applied to address the issue of algorithmic biases that have historically disadvantaged vulnerable populations. This project's importance is rooted in its potential to transform the development and deployment of AI algorithms, ensuring they are fair, transparent, and accountable. This research contributes valuable knowledge to the field of AI ethics, emphasizing the vitality of fairness-aware AI systems in a society that is becoming increasingly reliant on artificial intelligence to complete everyday tasks. Substantial consumer value exists in this research, as it has the potential to foster trust in AI systems to promote equity and fairness for all.

6. FUTURE WORK

The next steps of the project involve the actual development, training, and deployment of the proposed AI models based on the described theoretical framework. The preliminary versions of said models will require extensive testing to properly verify the effectiveness of the bias mitigation techniques and monitoring system. Once the educational program for stakeholders is put into place, analysis must be conducted to assess its impact on promoting fairness-aware practices in AI development. The future work associated with this research will also include scalability efforts as the framework must be able to continuously integrate additional fairness metrics when deemed necessary. Investigating the long-term effects of this system will play an important role in ensuring that this research and the AI algorithms that come with it benefit society.

REFERENCES

Angwin, J. (2016, May 23). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica.

Bellamy, R. (2018, Oct. 3) AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. IBM.

Buolamwini, J. (2018, Jan. 15). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research.