Algorithmic Bias in Facial Recognition: Exploring Racial and Gender Disparities through CNN Models

Gender Shades and George Floyd's Death: Raising Public Awareness of Algorithmic Bias

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Claire Yoon

May 9, 2023

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.

ADVISORS

Kathryn A. Neeley, Department of Engineering and Society

Rosanne Vrugtman, Department of Computer Science

Algorithmic Biases Towards Gender and Races in Facial Recognition Technology

Overview:

The topic of the project focuses on Facial Recognition Technology. This technology relies on massive data sets to identify faces and accurately detect an individual's identity. While it provides convenience, the technology exhibits algorithmic biases that result in misidentification for certain gender and racial groups. I plan to investigate the technology and explore ways that could minimize these undesirable consequences. To do this, I will adopt Batya Friedman's Value Sensitive Design theory, as it considers all factors not only for stakeholders influenced by the technology and it focuses on fairness and equity, making it the most suitable methodology for my project.

Positionality:

Having spent my formative years in South Korea before relocating to the United States, my country played a significant role in shaping my values and philosophies. The population in Korea is predominantly homogeneous; therefore, I never considered myself part of a minority group. However, everything changed after moving to the US. I now find myself belonging to a minority group. This sentiment is not solely a result of my ethnicity; my enthusiasm for engineering, though ultimately undeterred, has also been challenged by certain societal pressures. As I progressed through my engineering courses, I started to realize that there were fewer and fewer female students in each of my classes.

As a female engineer, I believe my personal experiences and the disparities I have encountered make me better understand my research topic. I worked on a project involving Convolutional Neural Networks (CNN) algorithms for image classification, examining algorithmic biases with respect to gender and race in FRT. This technology has misidentified people of color over white people, which has led to undesirable societal consequences. My project interests are significantly influenced by my experiences. Had I remained in a majority group in my country or not pursued to be a female engineer, I might not have been interested in this topic. As this project holds personal significance, I believe I can provide a thorough explanation of the subject matter.

Problematization:

The issue I am concentrating on pertains to algorithmic biases within Facial Recognition Technology, a subject that has garnered considerable controversy due to its implications for racial discrimination. Studies have revealed persistent inaccuracies in algorithms designed for facial detection when applied to people of color. Moreover, these algorithms have demonstrated reduced accuracy for female subjects compared to male subjects. The most critical concern surrounding this technology arises when it is employed for surveillance purposes to identify criminal suspects. Numerous instances have been documented wherein people of color were apprehended due to misidentification. Consequently, comprehending this problem is of paramount importance for the success of this project.

Primary Question:

How should we address algorithmic bias in Facial Recognition Technology for equitable performance across diverse populations?

Projected Outcomes:

In my project, I aim to investigate the algorithmic bias in Facial Recognition Technology and research strategies to minimize misidentification rates. My work seeks to challenge dominant paradigms and resolve unfairness in technology. The projected outcomes of this project would benefit specific gender and racial groups by suggesting solutional ideas to reduce false identifications, which have previously led to the wrongful arrest of innocent individuals, particularly people of color. This project will ultimately serve as a benchmark for fairness assessment during the development cycle and refinement of algorithms.

Technical Project Description:

This project, conducted as part of the Data Science department's course, involved collaboration with two group members. We developed Convolutional Neural Network (CNN) models to investigate the accuracy of the models in certain groups based on gender and race. CNN, a class of deep learning neural networks, is primarily designed for analyzing and processing images, with applications in computer vision tasks.

The project aimed at examining algorithmic biases within Facial Recognition Technology (FRT), which required a thorough understanding of the technology itself. FRT utilizes vast datasets to train systems in verifying individual faces. This process involves capturing, analyzing, and comparing patterns based on a person's unique facial features. The versatility and convenience of FRT have led to its widespread adoption across a multitude of industries, enabling the resolution of various challenges that were once prevalent a decade ago. FRT has been transformative in numerous fields, with applications ranging from securing sensitive information on mobile devices to locating missing individuals. Additionally, it has proven invaluable in identifying criminals, thereby contributing to public safety. By examining and addressing the algorithmic biases in FRT, the project aimed to research reducing algorithmic bias toward certain groups for ultimately enhancing the technology's effectiveness and fairness.

Preliminary Literature Review & Findings:

A key study conducted by Joy Buolamwini in 2018, a computer scientist and founder of the Algorithmic Justice League, titled "Gender Shades" at the MIT Media Lab, examined the performance of facial recognition systems developed by IBM, Microsoft, and Face++. She sought to determine their accuracy in identifying gender across different demographic groups. The study revealed significant disparities in the algorithms' accuracy when classifying faces by gender, with higher error rates for darker-skinned and female faces. In particular, darker-skinned females experienced the highest misidentification rates, unmasking a distinct bias in the algorithms towards lighter-skinned and male individuals (Buolamwini, 2018). Buolamwini's research has underscored the necessity to confront algorithmic bias in FRT, propelling further research to cultivate more inclusive and equitable systems, while concurrently elevating consciousness of the broader ethical implications of AI and machine learning technologies within society.

Simultaneously, other AI researchers Ramya Srinivasan and Ajay Chander have concentrated on the more expansive issue of bias in AI systems. They accentuate the significance of educating machine learning developers about biases that may arise in the AI pipeline and champion strategies to mitigate them. Their taxonomy of biases offers practical guidelines for limiting and testing for bias throughout the AI pipeline stages, bridging the gap between research and practice (Srinivasan & Chander, 2021). These studies encountered challenges in

implementing research ideas in real-world scenarios, but they contributed to enhancing awareness of transparent AI and encouraging practical skills to address algorithmic bias in AI/ML models.

Building upon these approaches, future work can focus on refining the guidelines provided by these researchers, ensuring that FRT is developed more responsibly.

STS Project Proposal:

Science, Technology, and Society (STS) is an interdisciplinary field that investigates the relationships between scientific development, technological innovations, and their impacts on society, culture, and values. My project can be considered an STS project because the project examines the impact on society and seeks to resolve the unpleasable result by developing strategies for reducing the bias toward certain gender and races.

In approaching the issue of algorithmic bias in FRT, the project aligns with several ecosystems of knowledge in STS. The problem that I am approaching is race and gender studies because the outcomes derived from algorithmic bias in FRT exhibit racial and gender discrimination. To understand the implications of biased ML systems, we need to consider ethics and values as a significant role. There is a primary author that investigated it. According to *Weapons of Math Destruction* by Cathy O'Neil, an American data scientist, highlights the dangers of opaque and unaccountable algorithms (O'Neil, 2016). To make ML systems transparent, she explains how models should be defined, what attributes of the models need to be contained, and what we should work on with the models. She emphasizes that models should be continuously updated and have statistical rigor. This would give me a deeper understanding of algorithmic bias and how responsible and ethical development should be conducted.

For this project, I will be using Batya Friedman's Value Sensitive Design methodology (Friedman et al., 2017). I believe that this framework is particularly fit into my research topic since it accounts for human values and provides suitable methods to examine the topic during the design process. To accomplish the project, I anticipate utilizing the Diverse Voices method (Magassa et al., 2017). Reviewing the method thoroughly, I will gather existing knowledge on algorithmic bias, its impact, and various approaches to identify and address potential biases and blind spots in technology design. I also expect that the method helps to understand stakeholder values to translate into technical design decisions as well as examine the dominant narrative to

address algorithmic bias in FRT. This method would ensure that FRT models need to be transparent and equitable to all users by incorporating their values and ethical considerations. By employing the Diverse Voices method, the project will gather diverse perspectives and data sources.

Barriers & Boons

One potential blind spot might be my positionality as a researcher with limited experience and knowledge in the field. The STS framework I will use is an unfamiliar area. So, I will look for guidance from experienced researchers who have worked on the relevant topics. I will also set realistic goals and prioritize the most crucial factors of the project to accomplish it within a limited time.

References

Friedman, B., Hendry, D. F., & Borning, A. (2017). A Survey of Value Sensitive Design Methods. *Foundations and Trends in Human-computer Interaction*, *11*(2), 63–125. https://doi.org/10.1561/1100000015

Magassa, L., Young, M., & Friedman, B. (2017). *Diverse Voices: a How-to Guide for Facilitating Inclusiveness in Tech Policy*. The Tech Policy Lab at the University of Washington. http://techpolicylab.uw.edu/wp-content/uploads/2017/10/TPL_Diverse_Voices_How-To_Guide_2017.pdf

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. https://ci.nii.ac.jp/ncid/BB22310261

Srinivasan, R., & Chander, A. (2021). Biases in AI systems. *Communications of the ACM*, *64*(8), 44–49. https://doi.org/10.1145/3464903

TED. (2017, March 29). *How I'm fighting bias in algorithms | Joy Buolamwini* [Video]. YouTube. https://www.youtube.com/watch?v=UG_X_7g63rY