Analyzing Racial and Gender Biases Perpetuated by Autonomous Vehicles From a Data Feminism Standpoint

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

Joon Kim

Spring 2022

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

Sean M. Ferguson, Department of Engineering and Society

Analyzing Racial and Gender Biases Perpetuated by Autonomous Vehicles From a Data Feminism Standpoint

Introduction

The concept of “Data Feminism” brings forth a new way of thinking about data science and data ethics informed by the ideas of intersectional feminism (D’Ignazio and Klein 2020). As we enter the age of digital automation, digitally automated technologies are gradually being introduced in aspects of everyday life, and concerningly in areas that place human life at risk. One such product is the autonomous vehicle, gradually being implemented in our roadways seemingly without regulation and without the consent of other drivers on the road. Those that favor automated technologies are influenced by the misconceptions that they eliminate human errors, and that they are inherently unbiased when it comes to decision making because they cannot think or feel emotion. The reality however is that “data-driven decision-making can be just as easily used to amplify the inequities already entrenched in public life” so long as the developers that curate the development of this neglect their existence. American sociologist Ruha Benjamin presents this emerging problem as the “New Jim Code”, likened to the Jim Crow laws passed in the United States in the 1870s, which upheld ideas of white supremacy by normalizing ideas of separate but equal principles and oppression. Today, she argues, these same discriminatory practices are creating a new digital caste system when developers fail to perceive how the historic patterns of racism can be reflected upon the data used to create automated systems to determine one’s deservedness for certain opportunities such as healthcare, housing, education and housing (Benjamin 2019, pp. 12).

Data feminism presents several principles engineers can use to analyze areas in which these sorts bias can be introduced into automated technologies, consciously and unconsciously, and calls upon developers to challenge existing unequal power structures and social hierarchies that may be perpetuating the emergence and effects of these biases. This paper applies these principles of Data Feminism to explore the ways in which biases can be introduced into these

systems, as well as shed light onto ways they can be addressed in the industry. For an application like autonomous vehicles, where passenger and pedestrian lives are held with elevated risks, it becomes crucial that the consequences brought forth by the New Jim Code are mitigated.

Background on Machine Learning Biases

Before exploring the ways in which biases can be introduced and reflected upon autonomous vehicle performance, it is important to understand the patterns in which data can adapt biases. The two most common and significant sources of biases stem from historical human biases and incomplete or unrepresentative training data.

Historical human biases are the kinds of biases that have existed for much of humanity with ideas such as racism and sexism. Machine learning algorithms, when influenced by this, may in turn end up reproducing and amplifying these biases. Kristen Lum and William Isaac explore one such instance of this in their case study on PredPol, a company that provides predictive policing softwares to assist law enforcement in identifying likely targets for police interventions. In this study, drug arrest reports from the city of Oakland's police department were found to be more concentrated around non-white and low-income populations, wheras estimates derived from data from the National Survey on Drug Use and Health suggest drug crimes to be more evenly distributed across the city (Lum, Isaac 2016). This resulted in the predictive polling algorithms that used this police report data to flag locations that were, by estimate, "already over-represented in the historical police data" (Lum, Isaac 2016). In consequence, Predpol's algorithm caused black people to be targeted roughly twice as often as white people, and other minority groups 1.5 times more often.

It is also important to note that in some cases, rather than these stereotypical views shaped by social roles and discrimination in history, these stereotypical differences can exist

through biological means and are reflected upon the differences in patterns of behavior among these groups. One study has found that men and women tend to react to stress differently, with hypertension, aggressive behavior observed generally higher in men while conditions such as depression and anxiety disorder have been found to be more prevalent among women (Verma et. al 2011). Some of these differences can be caused by biological differences, such as how some female sex hormones typically exhibit delayed containment of a stress responses. Furthermore, it has been observed that men typically exhibit more of a “fight-or-flight” response, while women exhibit a “tend-and-befriend” response. Analysis of the brain has shown that the male brain more predominantly shuts off regions associated with positive emotion and hedonic goals, causing more leading to the more aggressive behavior in the “fight-or-flight” response (Verma et. al 2011). These differences in behaviors can introduce unconscious biases to autonomous systems, where the more predominant pattern is more likely to be adapted by the algorithm.

Incomplete and under-representative training data occurs when datasets fail to equally represent all relevant demographic populations in which the autonomous technology is applied to. This results in systematically worse predictions and misclassifications for these underrepresented groups, due to the algorithm's underexposure and unfamiliarity. One study conducted by Joy Buolamwini, as part of her Gender Shades project, revealed the correlation between underrepresentation in darker skinned individuals and worse performance in facial recognition technologies. In her study, analysis of datasets used by gender classifiers and facial recognition technologies used by top companies such as Microsoft, IBM, and Adience, exposed an overwhelming overrepresentation of lighter skinned individuals, making up to 86.24% of the data. There also existed an imbalance in female vs male representation, with females making up as low as 24.6% As a result, darker females ended up becoming 32 times more likely to be misclassified than lighter skinned males. (Buolamwini 2017, pp. 32). Like facial recognition,

autonomous vehicles are also prone to misclassification if it is not trained to recognize relevant demographic groups.

Representation of demographic groups on algorithmic training and pedestrian recognition performance

One task an autonomous vehicle must be able to perform is to be able to identify pedestrians by training upon large amounts of images consisting of pedestrians and objects. This data becomes easily prone to biases against underrepresented demographic groups. These biases can stem from existing sociocultural differences between demographic populations, and can be easily neglected from the data without an awareness and understanding of these differences. For example, “women in the United Arab Emirates are socially stigmatized against photographing their faces, skewing the availability of this type of data to be lower than the true distribution” (Jo and Gebru 2020, pp. 309). The importance of addressing the mere existence of these sociocultural differences is further emphasized by the variations in volumes of demographic populations between regions and countries. In one study on the performance of algorithms of the Caltech Pedestrian Detection Benchmark, a statistically significant bias against children was found on 100% of the top 24 performing methods. This bias is correlated with the “overfitting of the distribution of appearance of pedestrians in the dataset, which is itself biased towards adults” (Brandão 2019, pp. 3). Furthermore, a higher difference on miss rates were found on datasets from INRIA, which is based in France, such that 18% of the population was aged 0-14 years old at the time of the study, leading to a lower chance of children to be included within a random sampling of data (Brandão 2019, pp. 3). Thus, from this study we can see how demographic groups can become more at risk than others by autonomous vehicles, which is exacerbated by the belief that some demographic groups such as children should be more worthy of being prioritized for protection. To mitigate the possibility for these algorithms to be influenced by such biases, it is important to understand the context in

which data comes from and what the data is composed of and ensure equal representation of demographic groups. Data collection should be enforced for equal representation of all demographic populations in order to maintain applicability in the realistically diverse social environments these vehicles will operate in for any social context.

Much like humans, machine-learning algorithms are able to classify demographic groups according to certain stereotypical features after sufficient training. These stereotypical features include but are not limited to skin tone, stature, and the color, length, and texture of hair. Determining fair and equal image classification algorithms have remained a challenging problem for machine learning algorithms for a long period of time, with facial and entity recognition tests conducted by the National Institute of Standards and Technology (NIST) determining that these algorithms have “a harder time recognizing people with darker skin” and “consistently found that they perform less well for women than men” (Simonite 2019). This problem is no different for pedestrian recognition algorithms used by autonomous vehicles, which must use these physical features in order to distinguish pedestrians from objects. This task has remained historically challenging to address all demographic populations. Some algorithms have used rhythmic features or motion patterns that are unique to humans, where the pedestrian’s feet or legs must be visible to extract the features (Zhao, Thorpe 2000, pp. 149), but this algorithm may fail for those that have mobility impairments and thus must rely on devices such as wheelchairs as a means of moving around. With algorithms such as these, we can see how biases can be introduced within algorithmic design when they fail to consider all applicable demographic groups. It is important that these gender and demographical stereotypical differences are carefully considered and accounted for in the data used to train the algorithms, and that the algorithms are tuned accordingly based on these differences. However, determining how exactly to tune the algorithms to account for this is another difficult challenge in and of itself.

Algorithm biases from stereotypical behaviors

Unconscious biases may also be introduced to machine learning datasets and algorithmic decisions from the ways in which different gender characteristics may be reflected upon the data. According to data collected from the Insurance Institute for Highway Safety, many more male drivers in the U.S have been killed than female drivers (NHTSA). This may be correlated to gender stereotypical behaviors. Research findings have shown that “male drivers are more likely – compared to female drivers – to exhibit various patterns of aggressive driving, such as cutting another vehicle, honking the horn, or exhibiting road rage” (Fountas et al 2019, pp. 2), and such characteristics may be in turn adopted to some extent by algorithms trained upon it. Emerging research which aims to train autonomous vehicles to exhibit more human driving behavior should be weary of these stereotypical differences. In one specific study, researchers have utilized data features such as “deceleration rates, stopping/slowing speeds, stopping/slowing durations and acceleration rates while participants drove specific routes in Los Angeles.” (Tavassoli et al 2019, pp. 1). These characteristics reflected upon collected data may have profound impacts especially on neural network algorithms used by autonomous vehicles, where “even the slightest bias present into the initial set of data points will be ‘ingrained’, or ‘encoded’, in resulting neural network models” (Cheong et al 2015, pp. 21). This goes to suggest that an imbalance in representation as well as an algorithmic design that fails to account for these characteristic differences lead to biases in the driving techniques these autonomous vehicles end up utilizing. With a more heavy representation of male driving patterns in data, autonomous vehicles may learn to adapt more aggressive techniques that may push the general consumer base more at risk for accidents.

On equal demographic representation for fair performance

As previously discussed, autonomous vehicle algorithms can easily become biased in pedestrian recognition tasks when the datasets they use become unbalanced in terms of representation across all relevant demographic groups. Previous research has shown the favorable effects of a dataset with a balanced race composition on fair algorithmic performance

in facial recognition technologies. In one specific dataset called Fairface created by Karkk¨ainen, Joo (2019), incremental increases were made to a dataset of faces, with estimates of demographic compositions of each country were made between iterations. Adaptive adjustments were made accordingly with measures such as excluding U.S and European countries in later iterations such that the dataset was not dominated by the White race (pp. 1549). Crowdsourced labeling, a separate model to evaluate these labels using ground truths, and manual verification were used to further refine this data. (pp. 1551). Consequently, when compared with other datasets, the model trained using Fairface ended up outperforming all other models as well as producing more consistent results for race, gender, and age classification (pp. 1552). Impressively, the Fairface model achieved a “less than 1% accuracy discrepancy between male ↔ female and White ↔ non-White for gender classification” with other models resulting in a gender gap as high as 32% (pp. 1553). From these results, we can see how Fairface sets an example for how datasets can be carefully tended and tuned in order to achieve equal representation and more fair performance.

Achieving more fair performance with the tuning of algorithmic design

Fairface demonstrated that more unbiased performance can be achieved when datasets are designed with an elevated focus on equal representation, but alternatively, the same effect is able to be achieved when the algorithms themselves are designed and tuned in a way such a more heavily represented group is prevented from affecting performance. In one study that analyzed the potential sources for predictive inequity in object vs pedestrian detection algorithms, it was concluded that sources of this predictive inequity most likely stemmed from the fact that, classified through the Fitzpatrick skin scale, lighter skinned individuals were represented around three times more than darker skinned individuals (Wilson et al 2019, pp. 8). For this study, a solution was as simple as a reweighting of loss functions for their respective groups, and was able to achieve improved performance on the darker skinned pedestrians without sacrificing performance on the lighter skinned pedestrians (Wilson et al 2019, pp. 9).

Thus, this study provides a compelling example of how biased performance does not always solely stem from unequal representation, but can reside inside and be addressed in the design of the algorithm itself. However, it is important to note that effective algorithmic tuning cannot be achieved without a proper understanding and awareness of these biases in the first place, which is why we must work towards implementing a more educated and diverse workforce equipped with the principles of data feminism.

Need for higher diversity in existing power structures

Although biases may be addressed in systematic ways, the challenge of addressing historical biases that can lead to such flaws in design requires us to examine existing power structures inside the industry that creates these autonomous vehicles. According to the U.S Equal Employment Opportunity Commission, the high tech industry has historically employed a larger share of whites (making up to 68.5%) and men (making up to 64%). The lean towards this particular demographic is problematic because it can obscure the viewpoints from these potentially marginalized demographics to identify points of stereotypical differences that can be reflected upon the data. What can be left are superficial decisions enacted upon ideas of these historical biases. Ruha Benjamin touches upon the experience of an ex apple employee working on Siri's voice recognition system in which they suggested accounting for African American English to their boss, only to be met with a response that "Apple products are for the premium market" (Benjamin 2019 pp. 19). Similar ideas could exist in autonomous vehicle development teams, as the self-driving car is considered by many as a luxury since its recent introduction to the consumer market..

The truth of this matter however, is that more diverse workplaces often impede inclinations of groupthink, which can in turn decrease factual errors on top of mitigating racial biases. In one study, jury panels were constructed with varying compositions of white and black jurors, which were then tasked with deciding whether a black defendant against white victims was guilty. The more diverse group ended up utilizing more facts related to the case while

making fewer factual errors regarding the evidence (Sommers 2006, pp. 605). In addition, more diverse groups were found to bring up topics of racism more frequently, and seeked additional perspectives, such as the child witness to the assault from this particular case.

Although not directly tied to the machine learning industry, the observations from this study can help us imagine how a more diverse body of developers can strengthen the mechanisms for designing and tuning autonomous vehicle algorithms. With more diverse developers, there will be a greater awareness of demographic differences stemming from the viewpoints of these workers, and more diverse perspectives from various groups of stakeholders may also be sought and emphasized. Easily overlooked gender stereotypical behaviors can be brought up into discussion from first hand experiences of female workers, and experiences of marginalized groups effected by biased artificial intelligence. This can bring into light areas of consideration to avoid echoing the same negative experiences. The inclusion of this higher diverse work population may as a whole encourage and educate previously apathetic workers on more multifaceted ways of thinking among previously apathetic workers. Consequently, a more diverse workforce wil shape and guide the industry with more insightful knowledge guided upon data humanistic principles to tackle and prevent biases.

Discussion

Through the research conducted in this paper, we have explored the ways in which representational and historical biases can unfairly skew autonomous vehicle performance against marginalized groups. The Caltech Pedestrian Detection Benchmark study not only tells us how we need to be weary about how datasets can become representationally biased due the fact that some demographic groups are more prevalent in the population than others, but also how demographic compositions that vary across regions can also contribute to this. Sociocultural differences among settings also skew availability in data. These points of insight are strongly suggestive of the idea that datasets and algorithms should not rely on any single source of this data because it is not guaranteed that it will be optimal in terms of applicability for

all different contexts with varying demographics. Like data feminism suggests, we should consistently analyze the context in which data comes from to ensure that discriminative performance is not a result of unbalanced representation. The Fairface study outlines the positive effects when data is corrected with an equal representation on all relevant populations to mitigate this sort of representative bias.

It has also been explored how biases can be introduced from differences in physical features as well as demographic behaviors among groups, in which the design of the algorithm and how it processes this data may be more at fault. In addition, it is not guaranteed that developers will always be able to collect data with enough samples from all populations due to sociocultural differences, such as how women in the United Arab Emirates are stigmatized against photographing their faces. The study conducted by Wilson et al. exemplifies how these algorithms can be tuned to weigh more importance in categorizations for biased or underrepresented groups. However, in the case of autonomous vehicles being influenced by male and female driving patterns presents a challenge that can exist when it comes to these algorithms. For systems that use driving behaviors, should we intentionally unbalance the data with more female drivers such that the vehicle learns to adopt these more careful driving patterns? Or how about the aggression of male driving patterns that can perhaps be viewed more favorably to those that want to minimize driving time? Perhaps this is a question that requires further discussion between the autonomous vehicle industry and its audience.

Lastly, biases can be introduced from the existing power structures of an industry primarily dominated by Caucasian males. This is a problem that cannot be solved systematically with the tuning of datasets and algorithms, but rather will require reevaluation of rooted employment practices and encouragement of these underrepresented groups in tech. Certain movements today such as “Women in Tech” are attempting to close the gender gap and help women embrace technology with measures such as involving world leaders and providing various means of support, but it will also require efforts from these tech industries aswell to push

towards measures of inclusion, as many women report experiencing discriminatory practices within the hiring process and in the workplace. I am hopeful that future generations of engineers may be better educated by schools on ideas of inclusion, and with the ideas of data feminism, and that more women are empowered to challenge the status quo in the tech industry.

References

1. *Diversity in high tech*. U.S. Equal Employment Opportunity Commission. (n.d.). from https://www.eeoc.gov/special-report/diversity-high-tech
2. Lum, K., & Isaac, W. (2016). To predict and serve?. *Significance*, *13*(5), 14-19.
3. Kärkkäinen, K., & Joo, J. (2019). Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*.
4. Jo, E. S., & Gebru, T. (2020, January). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 306-316).
5. Brandao, M. (2019). Age and gender bias in pedestrian detection algorithms. arXiv preprint arXiv:1906.10490.
6. Verma, R., Balhara, Y. P., & Gupta, C. S. (2011). Gender differences in stress response: Role of developmental and biological determinants. *Industrial psychiatry journal*, *20*(1), 4–10. https://doi.org/10.4103/0972-6748.98407
7. Wilson, B., Hoffman, J., & Morgenstern, J. (2019). Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*.
8. Simonite, T. (2019, July 22). *The best algorithms still struggle to recognize Black Faces*. Wired.https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/.
9. Cheong, M., Lederman, R., McLoughney, A., Njoto, S., & Wirth, A. (2020). Ethical implications of AI bias as a result of workforce gender imbalance. University of Melbourne.
10. Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy

disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

11. Fountas, G., Pantangi, S. S., Hulme, K. F., & Anastasopoulos, P. C. (2019). The effects of driver fatigue, gender, and distracted driving on perceived and observed aggressive driving behavior: A correlated grouped random parameters bivariate probit approach. *Analytic methods in accident research*, *22*, 100091.

12. *Automated vehicles for safety*. NHTSA. (n.d.). https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety.

13. Leavy, S. (2018, May). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st international workshop on gender equality in software engineering* (pp. 14-16).

14. Tavassoli, A., Cymbalist, N., Dunning, A., & Krauss, D. (2019). Learning from human naturalistic driving behavior at stop signs for autonomous vehicles. *Learning*, *1*, 1021.

15. Buolamwini, J. A. (2017). *Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers* (Doctoral dissertation, Massachusetts Institute of Technology).