**“Bias in Facial Recognition”  
Project proposal and literature Review  
VIC493: Capstone Research Colloquium  
Fizzah Mansoor  
1004975743**

**Introduction**

Recent research has highlighted the vulnerabilities of modern machine learning based systems to bias, especially towards segments of society that are under-represented in training data. Machine learning (ML) systems are increasingly making decisions that impact the daily lives of individuals and society in general. (Amini, Soleimany, Schwarting, Bhatia, & Rus, 2019) In many cases, health systems rely on commercial prediction algorithms to identify and help patients with complex health needs; many of these algorithms exhibit significant racial bias, which arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. (Obermeyer, Powers, Vogeli, & Mullainathan, 2019)

This somewhat demonstrates how much can go wrong when applying machine learning technologies in high-stakes, real life situations. In this paper, we specifically address Facial Recognition Technology (FRT); **a**utomated face recognition has achieved remarkable success with the rapid developments of deep learning algorithms. Despite the improvement in the accuracy of face recognition, one topic is of significance, it has been observed that many face recognition systems have lower performance in certain demographic groups than others. Such face recognition systems are said to be *biased* in terms of demographics.

Although the increased use of this novel technology may seem exciting and convenient, machine learning systems have been found to harbor forms of bias that can maintain and often increase inequalities. FRT, like other machine learning systems, suffers from “algorithmic bias,” which is defined as systematic errors in a computer program that lead to unfair outcomes. FRT manifests bias through a substantially better identification rate for faces with lighter skin and faces that exhibit traditionally-male facial features than faces with darker skin and faces that exhibit traditionally female facial features. (Buolamwini & Gebru, 2018)

The Ethical Principles in Artificial Intelligence Team (EPAI) is committed to educating and taking action on ethical concerns within artificial intelligence and data science. We aim to tackle problems such as bias in decision systems, data privacy and surveillance, and manipulation in behavior. We plan on making AI more accessible through workshops and projects that students can collaborate on. This research paper, titiled “Bias in Facial Recognition” is part of the larger mission of the club to educate undergraduate students on the theoretical and practical applications of artificial intelligence, with an emphasis on ethical implementation of such technologies.

Facial recognition technology (FRT) utilizes software to map a person’s facial characteristics and then store the data as a face template.Algorithms or machine learning techniques are applied to a database to compare facial images or to find patterns in facial features for verification or authentication purposes.  FRT is becoming increasingly common in security surveillance systems, is also attractive for a variety of health care applications, such as diagnosing genetic disorders, monitoring patients, and providing health indicator information (related to behavior, aging, longevity, or pain experience, for example). (Martinez-Martin, 2019)

**Literature review**

In 2018, a study published by a group of researchers at Dalian University in China reported the successful training of algorithms to distinguish faces of Uyghur people, a predominantly Muslim minority ethnic group in China, from those of Korean and Tibetan ethnicity. (Wang, Zhang, Liu, Liu, & M, 2018) China had already been internationally condemned for its heavy surveillance and mass detentions of members of the Uyghur community in camps in the northwestern province of Xinjiang — which the government says are re-education centres aimed at quelling a terrorist movement. Authorities in Xinjiang have used surveillance cameras equipped with software attuned to Uyghur faces. In her book Race After Technology, Ruha Benjamin asserts that the obvious shortcomings of facial recognition technologies is not a story of ill-intentioned, sinister programmers, but still has the potential to speed up and deepen discrimination. (Benjamin, 2019)

With the widespread use of AI systems and applications in our everyday lives, it is important to take fairness issues into consideration while designing and engineering these types of systems. Such systems can be used in many sensitive environments to make important and life-changing decisions; thus, it is crucial to ensure that the decisions do not reflect discriminatory behavior toward certain groups or populations. With the commercialization of these systems, researchers are becoming aware of the biases that these applications can contain and have attempted to address them. (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2019)

Traditionally, machine learning algorithms relied on reliable labels from experts to build predictions. More recently however, algorithms have been receiving data from the general population in the form of labeling, annotations, etc. The result is that algorithms are subject to bias that is born from ingesting unchecked information, such as biased samples and biased labels. Furthermore, people and algorithms are increasingly engaged in interactive processes wherein neither the human nor the algorithms receive unbiased data. (Sun, Nasraoui, & Shafto, 2020) Algorithms can also make biased predictions, leading to what is now known as algorithmic bias. On the other hand, human’s reaction to the output of machine learning methods with algorithmic bias worsen the situations by making decision based on biased information, which will probably be consumed by algorithms later.

**In 2020, Gong et al. addressed the problem of bias in automated face recognition and demographic attribute estimation algorithms, where errors are lower on certain cohorts belonging to specific demographic groups, by presenting a novel de-biasing adversarial network called DebFace. DebFace learns to extract disentangled feature representations for both unbiased face recognition and demographics estimation. The network consists of one identity classifier and three demographic classifiers (for gender, age and race) that are trained to distinguish identity and demographic attributes respectively. Adversarial learning (a machine learning** technique that attempts to fool models by supplying deceptive input**) is adopted to minimize correlation among feature factors to dampen bias influence from other factors. This approach was successful in reducing bias in face recognition.** (Gong, Liu, & Jain, 2020)

**A similar approach was undertaken by Yucer et al in a 2020 study; researchers** proposed a novel adversarial derived data augmentation methodology that aimed to enable dataset balance at a per-subject level via the use of image-to-image transformation for the transfer of sensitive racial characteristic facial features. This approach was successful in automatically constructing a synthesised dataset by transforming facial images across varying racial domains, while still preserving identity-related features, such that racially dependant features subsequently become irrelevant within the determination of subject identity. The proposed technique had a positive impact on recognition performance for (racial) minority groups within an originally imbalanced training dataset by reducing the pre-race variance in performance. (Yucer, Akçay, Al-Moubayed, & Breckon, 2020).

The second model we will be examining is the Learned Latent Structure algorithm which works by mitigating the hidden, and potentially unknown, biases within training data. This algorithm fuses the original learning task with a variational auto-encoder to learn the latent structure within the dataset and then adaptively uses the learned latent distributions to re-weight the importance of certain data points while training. This model demonstrated increased overall performance as well as decreased categorical bias with our de-biasing approach.. (Amini, Soleimany, Schwarting, Bhatia, & Rus, 2019)

The training dataset is fed through the encoder network, which provides an estimate of the latent distribution. The model is programmed to increase the relative frequency of rare data points by increased sampling of under-represented regions of the latent space. Training on the debiased data batch now forces the classifier into a choice of parameters that work better in rare cases without strong deterioration of performance for common training examples. Most importantly, the de-biasing is not manually specified beforehand but instead based on learned latent variables. (Amini, Soleimany, Schwarting, Bhatia, & Rus, 2019) **For this research paper, we hypothesize that the learned latent structure algorithm is marginally better suited as a de-biasing algorithm for FRTs than DebFace because adversarial learning models may skew the performance in unexpected ways, even if they manage to mitigate algorithmic bias.**

**Techniques**

Owing to the relative inexperience of all students associated with this research paper, we will be using open source labs posted by Aamini on their github account in the case of the Learned Latent Structure de-biasing method (<https://github.com/aamini/introtodeeplearning/tree/master/>), and the code uploaded by Gong on his github account for the DebFace de-biasing method at <https://github.com/gongsixue/DebFace>. Both algorithms will be evaluated on the Pilot Parliaments Benchmark (PPB), a dataset specifically designed to evaluate biases in computer vision systems.

**Owing to the limited scope of this paper, there will be significant uncertainties in the results. It is not yet known which model will perform better than the other, and we may not be able to pinpoint the reasons for one model’s success over the other. We aim to address this by carefully working through the mathematical basis for each de-biasing algorithm, observing how each model behaves in response to changes in training data, and tracking where each model falls short.**

**Expected outcomes**

The goals of this research are multifaceted, and aim to address both the ethical implications, as well as the technological basis of bias in facial learning algorithms. The ethics focused part of this paper includes reflecting on the societal impact of bias in facial recognition (including how the use of such technologies by law enforcement could disproportionately affect people of colour), investigating potential causes of biased models and the flawed data collection methods that construct the databases for such algorithms (Uighur concentration camps in China, police recognition technology in the US, etc). The technical aspect of this paper will focus on researching and implementing other de-biasing methods and comparing their performances when trained on the same dataset.

**This project is still in its early stages, and a more complete literature review is underway. We hope to scratch the surface of why certain models work better than others by directly comparing the efficiency of two separate de-biasing methods. The paper will feature a comprehensive look at the underlying mechanism of de-biasing models, what specific systemic errors they target, and how they can be further improved. The learned latent structure model and DebFace are both relatively recent inventions, and their cross examination on similar data sets will reveal the cost of choosing less powerful de-biasing models or forgoing de-biasing altogether, as well as the effect that homogenous, unfair databases have on outputs of machine learning models.**

**Motivation**

**I work with a dedicated team of undergraduates who have little experience with formal research, but are committed to doing their best to contribute to a field that needs more young minds now than ever before. Toronto is a hub of artificial intelligence research, and there needs to be a greater emphasis on approaching research on this topic that pointedly focuses on the ethical implications of implementing these fast-growing technologies on a large scale.**

**Since the field is still relatively new, there is a lack of comprehensive studies and surveys that actively compare results from different de-biasing methods. De-biasing is a powerful tool in reducing algorithmic biases that negatively impact already vulnerable communities and further legitimize the racism and sexism already endemic in public institutions. We hope that this paper contributes towards the research already being done towards creating more efficient de-biasing methods, and also reveal the underlying mechanism and efficiency of two different approaches to de-biasing.**

# Bibliography

Amini, A., Soleimany, A. P., Schwarting, W., Bhatia, S. N., & Rus, D. (2019). Uncovering and Mitigating Algorithmic Bias through Learned. *AAAI/ACM Conference on AI, Ethics, and Society* (pp. 289-295). New York, NY: Association for Computing Machinery.

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code.* Polity.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Conference on Fairness, Accountability, and Transparency*, (pp. 77-91).

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* *, 542*, 115-118.

Gong, S., Liu, X., & Jain, A. K. (2020). Jointly De-Biasing Face Recognition and Demographic Attribute Estimation. *European Conference on Computer Vision* (pp. 330-347). Springer.

Martinez-Martin, N. (2019). What Are Important Ethical Implications of Using Facial Recognition Technology in Health Care? *AMA J Ethics* *, 21* (2), 180-187.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *arXiv: Learning* .

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* , 447-453.

Shen, X., Plested, J., Caldwell, S., & Gedeon, T. (2021). Exploring Biases and Prejudice of Facial Synthesis via Semantic Latent Space. *International Joint Conference on Neural Networks (IJCNN)*, (pp. 1-8).

Sun, W., Nasraoui, O., & Shafto, P. (2020). Evolution and impact of bias in human and machine learning algorithm interaction. *PLOS ONE* *, 15(8)*.

Wang, C., Zhang, Q., Liu, W., Liu, Y., & M, L. (2018). Facial feature discovery for ethnicity recognition. *WIREs Data Mining Knowledge Discovery* *, 9*.

Yucer, S., Akçay, S., Al-Moubayed, N., & Breckon, T. P. (2020). Exploring Racial Bias within Face Recognition via per-subject Adversarially-Enabled Data Augmentation. *Seyma Yucer; Samet Akçay; Noura Al-Moubayed; Toby P. Breckon* .