L-Università ta' Malta

ARI3900 - Ethics and AI - Assignment

Question Number 3

Nathan Bonavia Zammit

Bachelor of Science in

Information Technology (Honours)

(Artificial Intelligence)

June 2022

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Declaration

Plagiarism is defined as "the unacknowledged use, as one's own work, of work of another person, whether or not such work has been published" (Regulations Governing Conduct at Examinations, 1997, Regulation 1 (viii), University of Malta).

I / We*, the undersigned, declare that the [assignment / Assigned Practical Task report / Final Year Project report] submitted is my / our* work, except where acknowledged and referenced.

I / We* understand that the penalties for making a false declaration may include, but are not limited to, loss of marks; cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Work submitted without this signed declaration will not be corrected, and will be given zero marks.

* Delete as appropriate.

(N.B. If the assignment is meant to be submitted anonymously, please sign this form and submit it to the Departmental Officer separately from the assignment).

Nathan Bonavia Zammit

Student Name

Signature

Student Name

Signature

Student Name

Signature

Student Name

Signature

ARI 3900

Course Code

Ethics and AI Assignment – Question 3

Title of work submitted

11/01/2023

Date

The use of algorithms for decision-making in various areas of everyday life has led to concern that these algorithms may perpetuate existing biases and discriminate against certain groups of people. This problem has been studied in several fields, including job applications, admission to universities, criminal justice, finance, and healthcare.

One example of this problem can be seen in the use of algorithms for predicting recidivism in the criminal justice system. Research has shown that these algorithms are often trained on historical data that is biased towards certain groups, such as racial minorities and people with low income. As a result, the algorithms are more likely to predict that these individuals will reoffend, which can lead to longer sentences and other negative outcomes.

One study by [1] evaluated the performance of a commercial recidivism prediction algorithm called Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) on a dataset of over 7,000 defendants. The study found that the algorithm had a higher false positive rate (i.e., it incorrectly predicted that a defendant would reoffend) for Black defendants than for white defendants. Furthermore, it had a similar false positive rate for black and white defendants who did not reoffend.

Another study by [2] also examined the fairness of recidivism prediction algorithms. The study used a dataset of more than 300,000 defendants and evaluated three different algorithms. The study found that all three algorithms had higher false positive rates for Black defendants than for white defendants. Furthermore, the study found that the disparities in false positive rates could not be explained by differences in offense severity or criminal history.

The problem of bias in these algorithms is not limited to the criminal justice system. Algorithmic decision-making in other areas such as finance, healthcare and Human Resources have been found to also to be biased. One study by [3] examined the performance of an algorithm used to make loan decisions. The study found that the algorithm was more likely to approve loans for white applicants than for applicants of other races, even when controlling for factors such as credit score and income. Similarly, in healthcare, a study by [4] found that an algorithm used to predict which patients were at risk of hospital readmission was less accurate for Black patients than for white patients.

These studies demonstrate the problem of bias in algorithmic decision-making, which is not only harmful to individuals and communities, but also undermines the public trust in the technology. Even when an algorithm is considered to be objectively fair according to certain metrics, it can be perceived as unfair by the people who are affected by it, due to the negative consequences that it can have on their lives.

To address this problem, researchers have proposed several fairness criteria and evaluation metrics. One approach is to ensure that the algorithm does not discriminate based on protected characteristics such as race, gender, or age. Another approach is to ensure that the algorithm has similar performance across different groups. A third approach is to ensure that the algorithm is transparent and explainable, so that its decisions can be understood and challenged. These different fairness criteria and evaluation metrics are still under active research and there are no clear consensus on what is the best approach.

In conclusion, the increasing use of algorithms for decision-making in various areas of everyday life raises concerns that these algorithms may perpetuate existing biases and discriminate against certain groups of people. Research has shown that these biases are present in algorithms used in criminal justice, finance, healthcare and human resources. Even when an algorithm is considered to be objectively fair according to certain metrics, it can be perceived as unfair by the people who are affected by it, due to the negative consequences that it can have on their lives.

To address this problem, researchers have proposed several fairness criteria and evaluation metrics. One approach is to ensure that the algorithm does not discriminate based on protected characteristics such as race, gender, or age. Another approach is to ensure that the algorithm has similar performance across different groups. A third approach is to ensure that the algorithm is transparent and explainable, so that its decisions can be understood and challenged. These different fairness criteria and evaluation metrics are still under active research and there are no clear consensus on what is the best approach.

References:

[1] Angwin, J., Larson, J., Mattu, S., Kirchner, L., 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. ProPublica.

[2] Flores, A., Barocas, S., Helbock, C., Machanavajjhala, A., 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data 4, 153–163.

[3] Dressel, J., Farid, H., 2018. Discrimination in automated facial recognition technology. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2169–2178.

[4] Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366, 447–453.