Text, logo

Description automatically generatedLogo

Description automatically generated with medium confidence

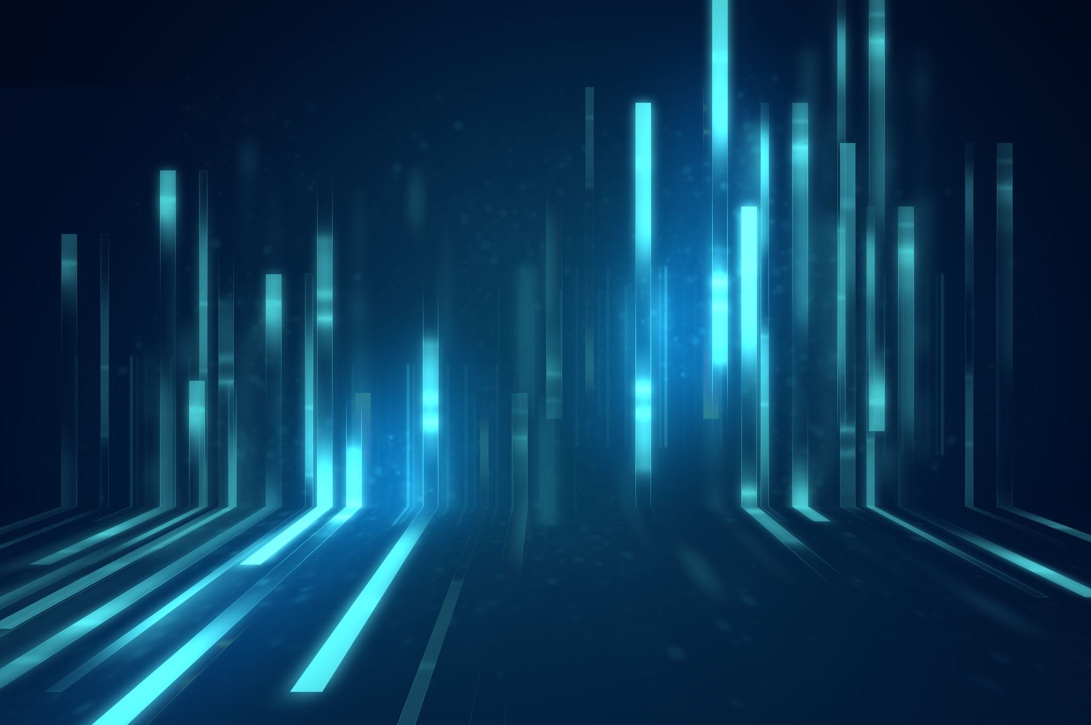
­­­

Fairness

Expert

Module 2

**Technical Interpretations of Fairness**



# Introduction

This module will expand on the intermediate course on fairness. It will provide further tools for applying learnings on fairness/bias in the development of a technological solutions.

a. Mathematical formulations, quantification and metrics of bias and fairness and mitigation techniques for different types of models

b. Performing a technical assessment of the fairness/bias quantification requirements of a problem (considering data used, model used, end-user requirements)

c. Overview of available tools for fairness/bias assessment and mitigation techniques

d. Communicating bias in the data/model to end-users, including metrics visualisation

­­­

# Pre-reading

* Guide to responsible AI practices – web page article that details some responsible AI practices, noting that these are guidelines to orient oneself towards rather than hard and fast rules on conduct because the field of fairness in AI is an active area of research

<https://ai.google/responsibilities/responsible-ai-practices/?category=fairness>

* Wired article on how Dall-E image generator generates images and the biases it struggles with

<https://www.wired.com/story/dall-e-2-ai-text-image-bias-social-media>

* Discussion on whether AI algorithms can ever be ‘fair’ with reference to COMPAS

<https://massivesci.com/articles/machine-learning-compas-racism-policing-fairness>

**Lesson 1 reading**

* Statice resource on types of bias that can interfere with machine learning - <https://www.statice.ai/post/data-bias-types>
* Google machine learning ‘types of bias’ - <https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias>
* Google machine learning ‘evaluating bias’ - <https://developers.google.com/machine-learning/crash-course/fairness/identifying-bias>
* Fairness and explanation in AI informed decision making - <https://www.mdpi.com/2504-4990/4/2/26>

(pdf link: <https://www.mdpi.com/2504-4990/4/2/26/pdf>)

**Lesson 2 reading**

* Fairness definitions explained – paper that rigorously discusses mathematical conceptions of fairness

<https://vsahil.github.io/files/papers/Fairness_Definitions_Explained_2018.pdf>

* On formalising fairness in ML – paper detailing how to formalise fairness mathematically

<https://arxiv.org/pdf/1710.03184.pdf>

* Fairness constraints – paper that relaxes a fairness criterion to develop a more practical fairness heuristic

<https://arxiv.org/abs/1507.05259>

* Impossibility theorem of ML fairness – describes how different types of fairness are statistically incompatible, necessitating trade-offs

<https://arxiv.org/abs/2007.06024>

**Self-assessment pass/fail questions**

1) Fairness is guaranteed when data doesn’t contain any sensitive information

a. Always

b. Sometimes

c. Never

2) Datasets should be audited for bias before we begin any data science task

a. Always

b. Sometimes

c. Never

3) Ensuring all individuals receive exactly the same treatment is the way to achieve fair outcomes

a. Always

b. Sometimes

c. Never

4) Ensuring all similar individuals receive similar treatment is the way to achieve fairness

a. Always

b. Sometimes

c. Never

5) The way to achieve fairness in a model depends on the use case

a. Always

b. Sometimes

c. Never

6) We can obtain fair outcomes from unfair data

a. Always

b. Sometimes

c. Never

7) False positives are more harmful than false negatives

a. Always

b. Sometimes

c. Never

8) It is possible to satisfy all definitions of fairness simultaneously

a. Always

b. Sometimes

c. Never

9) The precision of a classifier is calculated by

a. True positives / (true positives + true Negatives)

b. True positives / true negatives

c. True positives / total positives

d. True positives / (true positives + false negatives)

e. True positives / number of predictions

10) Demographic parity means

a. Making sure the most disadvantaged subgroup has more positive outcomes than the least disadvantaged

b. Treating all subgroups the same

c. All subgroups have the same probability of a positive outcome

d. The model has the same precision for all subgroups

11) Equalised odds means

a. The sensitivity and specificity is the same across all subgroups

b. Everybody the same chance of a positive outcome

c. Ignoring sensitive characteristics

d. Probability of a true positive and a true negative are equal

12) In the context of model bias in a classifier, equal opportunity means

a. There is a chance of a positive outcome for every subgroup

b. Just the specificity of the model is the same for all subgroups

c. Just the sensitivity of the model is the same for all subgroups

d. The chance of a true positive is the same for all subgroups

13) Which of these is NOT a bias mitigation strategy?

a. Post-hoc mitigation (post-processing)

b. Resampling (pre-processing)

c. Rebalancing (out-processing)

d. Regularisation (in-processing)

14) A disadvantage of optimising for demographic parity is

a. It rewards ‘lazy’ models

b. It rejects the optimal classifier

c. It leads to short term low model precision

d. All of the above

15) In the context of model selection, the ‘Pareto front’ is

a. The collection of models that have the highest accuracy

b. The collection of models that minimise disparity the most

c. The collection of models that represent the best trade-offs between precision and fairness

d. The collection of models that represent worst-case scenarios

­­­

**Answers**

Qs 1) C, 2) A, 3) B, 4) B, 5) A, 6) B, 7) B, 8) C, 9) C, 10) C, 11) A, 12) C, 13) C, 14) D, 15) C Qs

