Text, logo

Description automatically generatedLogo

Description automatically generated with medium confidence

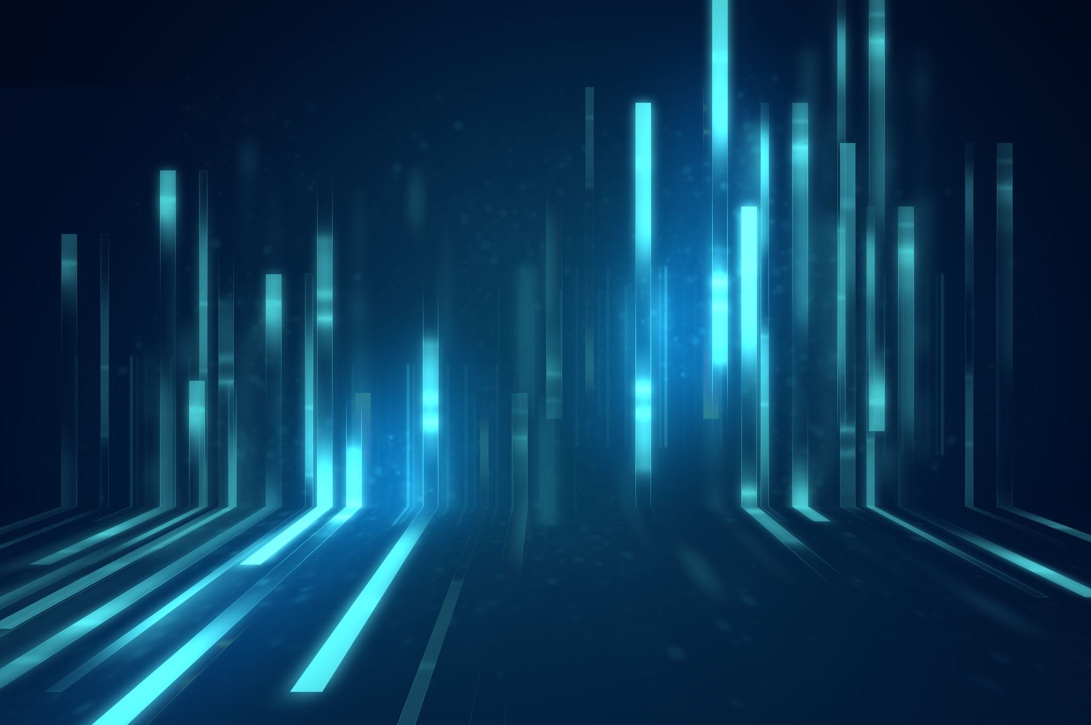
­­­

Fairness

Intermediate

Module 4

Technical Interpretations of Fairness



# Introduction

This module will provide an understanding of fairness and bias in the context of technology development

a. Understanding fairness and bias requirements with respect to data available, models to build and end-user requirements

b. Definition of common metrics/quantification methods of bias and fairness and bias mitigation techniques

c. Overview of available open-source tools for fairness/bias assessment

Apply learnings to use cases from TRI’s projects and products. The learner will be provided with Jupyter notebooks showing application of different techniques to assess fairness/bias on datasets and models performances together with multiple choice questionnaires and code snippets to modify.

­­­

**Pre-reading**

* Attacking discrimination in ML – interactive tool for exploring different types of discrimination mitigation

<http://research.google.com/bigpicture/attacking-discrimination-in-ml>

* 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>

**Session 1 reading**

* Structural bias in AI models

<https://2021.ai/structural-bias-in-ai-models>

* Brief descriptions of types of bias from TAUS, a language data network

<https://www.taus.net/resources/blog/9-types-of-data-bias-in-machine-learning>

* Codeacademy resources on types of bias

<https://www.codecademy.com/article/bias-in-data-analysis>

**Session 2 reading**

* How to tackle bias in AI

<https://2021.ai/how-to-tackle-bias-in-ai>

* How we analysed the COMPAS recidivism algorithm

<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

* Equality of opportunity in ML – article explaining one equality/fairness strategy

<https://ai.googleblog.com/2016/10/equality-of-opportunity-in-machine.html>

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

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

* A short term intervention for long term fairness – Paper that describes how strategies to improve fairness can go beyond the current situation and current model and aim to achieve a longer term fair outcome by implementing demographic parity

<https://arxiv.org/abs/1712.00064>

**Tasks**

**Activity 1**

Based on the final slides 15 and 16 of the first set of slides, write a paragraph discussing some answers to the questions on slide 16.

Some bullet pointed examples:

* What are the sources of bias?
  + Socio-economic disparities between groups
  + Social media content policies
  + Types of content available on social media (e.g. only text? Images? Videos?)
* How might these manifest in the data?
  + Social media only represents tech literate
  + Only demographics with access to social media represented
  + Interference from unknown algorithms / social media influencers
  + Data neglects things that can’t be captured by social media
  + Inconsistent geocoding
  + Social media attention/activity likely corresponds to population density
* What are the downstream risks to fairness if we apply this model? i.e. how could bias impact the outcome?
  + Might over-allocate / under-allocate resources
  + Social media users can game the system by complaining more
  + No comparison method to cross-validate
  + Facebook and Twitter algorithms skew data in ways that we aren’t fully aware of

**Activity 2**

Having read about the example of the biases in the COMPAS algorithm as part of the reading for the course, choose one of the following examples (or search for another) of bias affecting tech outputs and write a brief summary of what issue the tech was trying to solve, the biases in the data it was trained on, and how this manifested in the failure modes of that tech solution, as well as how the problem was mitigated.

* Twitter’s automated image cropping algorithm
* The PredPol algorithm
* Amazon’s recruiting engine
* IBM’s “Watson for Oncology”
* Microsoft’s online chatbot “Tay”

**Self-assessment pass/fail questions**

1) Datasets should be checked for imbalances and biases

a. Always

b. Sometimes

c. Never

2) If a dataset contains no sensitive information then we don’t need to worry about bias

a. Always

b. Sometimes

c. Never

3) As long as a process isn’t explicitly discriminating against an individual or group of people, the outcome of that process will be fair.

a. Always

b. Sometimes

c. Never

4) Data collected from the real world is free of bias

a. Always

b. Sometimes

c. Never

5) If an outcome is fair, then the steps of the process that produced that outcome are fair

a. Always

b. Sometimes

c. Never

6) Sensitive data can be inferred from non-sensitive data

a. Always

b. Sometimes

c. Never

7) Fair outcomes are achieved by ensuring everybody is treated exactly the same

a. Always

b. Sometimes

c. Never

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

a. Always

b. Sometimes

c. Never

9) Which of these is NOT a type of bias?

a. Confirmation bias

b. Selection bias

c. Historical bias

d. Diagonal bias

e. Survivorship bias

f. Reporting bias

10) A ‘lazy’ model is one that

a. Doesn’t work

b. Doesn’t treat subgroups equally

c. Doesn’t treat individuals uniquely

d. Doesn’t use all of the available data

11) 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

12) Demographic parity can help address historical biases in the long term

a. True

b. False

13) 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

14) 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 across all subgroups

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

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

15) A disadvantage of equalised odds and equal opportunity is

a. It favours lazy models

b. It only works when there are two subgroups

c. It’s expensive to implement

d. It doesn’t address historical bias in the long term

**Answers**

Qs 1) a, 2) c, 3) b, 4) c, 5) b, 6) b, 7) b, 8) c, 9) d, 10) b, 11) c, 12) a, 13) a, 14) c, 15) d Qs