Text, logo

Description automatically generatedLogo

Description automatically generated with medium confidence

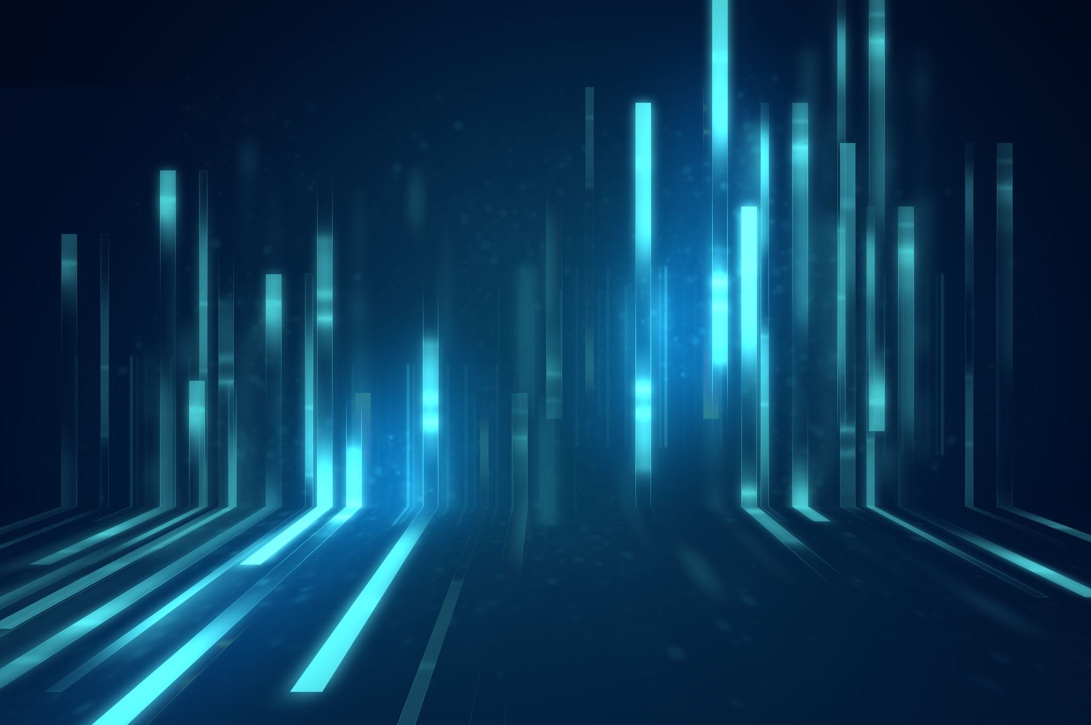
­­­

Fairness

Expert

Module 2

**Technical Interpretations of Fairness – ex 1**



# 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.

Here we include the first example scenario as an exercise to identify potential sources of bias in a practical setting.

­­­

# Scenario

**Disaster relief**

A severe weather event has hit the coast. Many areas have been flooded, and other areas damaged by strong winds. There are reports of people facing the loss of their homes to the damage. The Emergency Relief Agency is considering the use of social media data to assess the damage done and distribute resources to affected areas, including food and shelter, emergency services, vehicles and other resources to restore power and communications.

**Goal**

Accurately assess damage and send appropriate relief resources

**Data**

Social media posts (Twitter, Facebook etc, some of which may be contain geographical data within and nearby the affected area)

**Actions**

Assess and allocate relief efforts

# Questions - Sources of Bias

* What are some potential sources of bias in the underlying data?
* What are the risks to fairness in downstream applications and deployment of the model described?

# Questions – Bias Metrics

* How would you describe a false positive in this problem to a policymaker or business owner? What’s the potential harm/cost of one?
* How would you describe a false negative to a policymaker or business owner? What’s the potential harm/cost of one?
* What confusion matrix metric (e.g., FPR, FNR, TPR, FDR, etc.) would you choose to focus on in terms of equity for this case?

# Example answers

**What are some potential sources of bias in the underlying data?**

Sample bias:

Missing people who don’t use social media

* Old people don’t use social media.
* Lack of access to internet for some places affected / may not be able to tweet/social media
* Data only from FB / twitter users
* Data missing geocoding (posts about the disaster without knowing where from)

How active people of impacted area are on social platforms

People who are posting might be biased to people of a specific ethnicity

Retweets and reposts

People more likely to post about damage/negative outcomes than positive

How complete is the dataset that FB and Twitter are able to share? Data sharing or access restrictions as well as users privacy settings might interfere with data collection

What is the social and demographic distribution in the area – twitter demographics (younger, professional, tech literate) aren’t necessarily representative.

Reliability: Some posted contents may not be true

Geo-information can be wrong (with high noise)

Confirmation bias: we will only detect signals that can be expressed through social media, anything that can’t be captured by social media activity will be ignored

**What are the risks to fairness in downstream applications and deployment of the model described?**

* Might over-allocate resources based on the biases. “Rich gets richer”
* Might under-allocate resources
* Social media users can game the system by complaining more
* Do we have a comparison method to check against?
* Facebook and Twitter algorithms skew data in ways that we aren’t fully aware of
* Maybe places that are more densely populated get prioritized b/c more social media users

­­­

