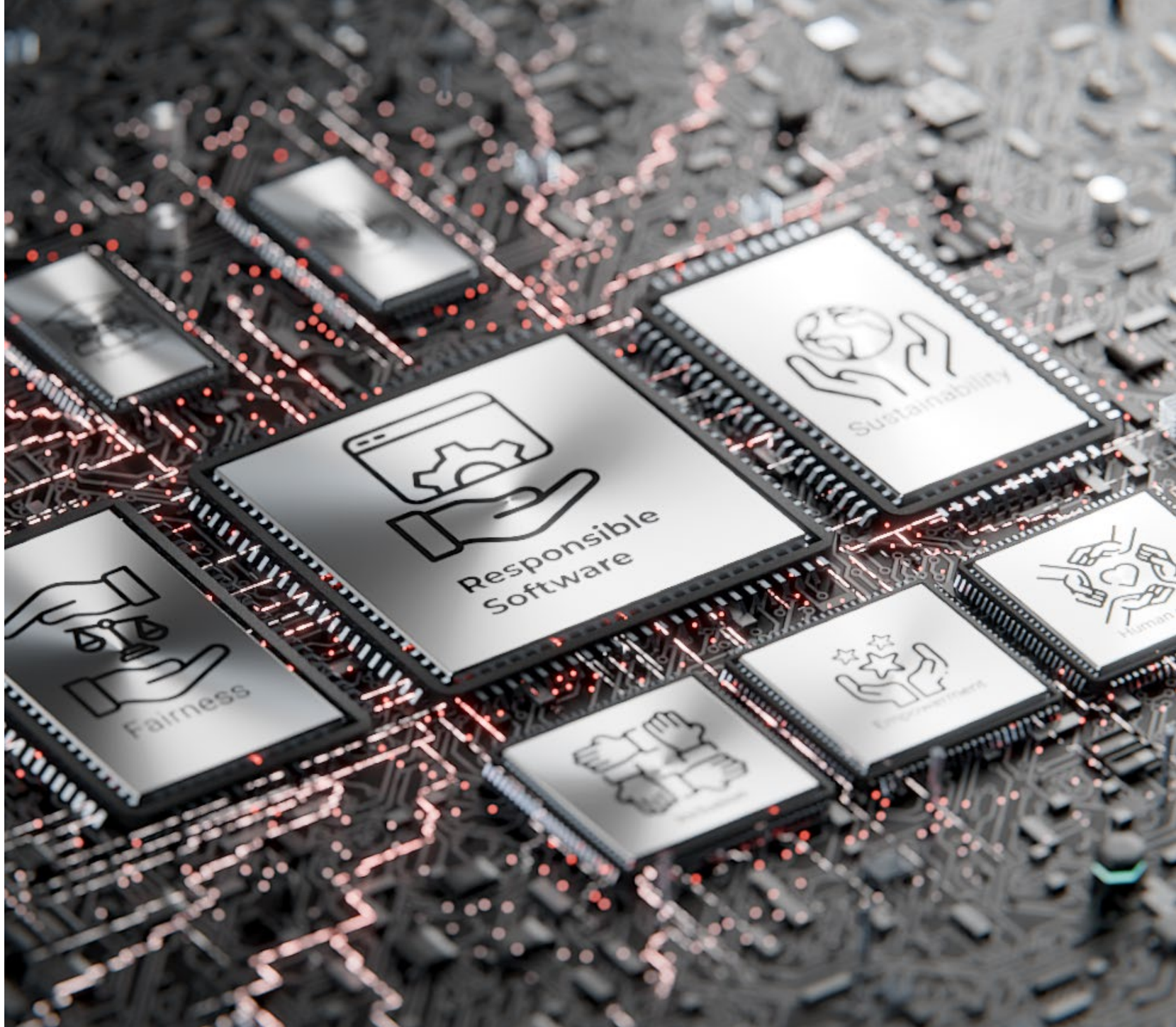
EPFL

Fairness 2 Review & Case studies

13 oct.

Cécile Hardebolle

Responsible Software



Agenda for today

1. Upcoming dates in the course

2. Interactive review questions on Fairness 2

3. Case studies:
   a) Datasheets for datasets
   b) People behind the data & COMPAS
   c) (Harms modeling can be done as training at home)

Next dates

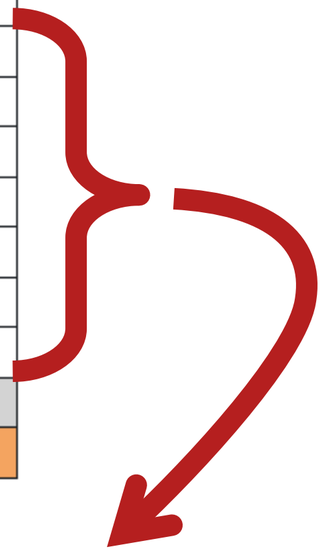
	Monday	Tuesday
13 Oct – 17 Oct	Fairness 2 cases	Graded notebook 1
20 Oct – 24 Oct	Autumn break	
27 Oct – 30 Oct	Debriefing Graded 1	Mock Test
3 Nov – 7 Nov	Debriefing Mock Test (in CO3)	Sustainability 1 notebook

"Debriefing" =

- I will give a global **feedback** to the class
- We will work together through the **most difficult exercises**
- We will discuss your **questions** on the notebook & the mock test

Second part of the course

Date	Week	Lecture (Monday 15h15-17h) in STCC Cloud C	Exercise session (Tuesday 10h15-12h)	Independent study (due before the following Monday)
08/09	1	Getting started	Introduction notebook	Introduction videos and quizzes
15/09	2	Introduction cases (**in CO3**)	Safety 1 notebook	Safety 1 videos and quizzes
22/09	3	public holiday	Safety 2 notebook	Safety 2 videos and quizzes
29/09	4	Safety 2 cases	Fairness 1 notebook	Fairness 1 videos and quizzes
06/10	5	Fairness 1 cases (**in CO3**)	Fairness 2 notebook	Fairness 2 videos and quizzes
13/10	6	Fairness 2 cases	Graded notebook 1	-
20/10			Autumn break	
27/10	7	Graded 1 debriefing	Mock test	-
03/11	8	Mock test debriefing (**in CO3**)	Sustainability 1 notebook	Sustainability 1 videos and quizzes
10/11	9	Sustainability 1 cases	Sustainability 2 notebook	Sustainability 2 videos and quizzes
17/11	10	Sustainability 2 cases	Empowerment 1 notebook	Empowerment 1 videos and quizzes
24/11	11	Empowerment 1 cases	Graded notebook 2	-
01/12	12	Graded 2 debriefing	Empowerment 2 notebook	Empowerment 2 videos and quizzes
08/12	13	Empowerment 2 cases	Graded case	Conclusion videos and quizzes
15/12	14	Conclusion cases	Conclusion review	-
Revisions				
TBD	Exams	Written exam		



- There will be **fewer** videos
- We will **practice again** with a good number of the strategies

Review questions Fairness 2

Biases in the ML lifecycle - 1

URL: ttpoll.eu
Session ID: cs290

Simpson's paradox is when the patterns observed at the level of the full sample and at the level of subgroups are opposed.

When training a ML model, Simpson's paradox can lead to (select 1 answer):

- Training time
- Pattern at aggregated level is different from patterns for subgroups

❌ 25% a. Evaluation bias

✅ 25% b. Aggregation bias

❌ 25% c. Optimization choices

❌ 25% d. Deployment bias

3.4 Aggregation Bias

Aggregation bias arises when a one-size-fits-all model is used for data in which there are underlying groups or types of examples that should be considered differently. Underlying aggregation bias is an assumption that the mapping from inputs to labels is consistent across subsets of the data. In reality, this is often not the case. A particular dataset might represent people or groups with different backgrounds, cultures or norms, and a given variable can mean something quite different across them. Aggregation bias can lead to a model that is not optimal for any group, or a model that is fit to the dominant population (e.g., if there is also representation bias).

Biases in the ML lifecycle - 2

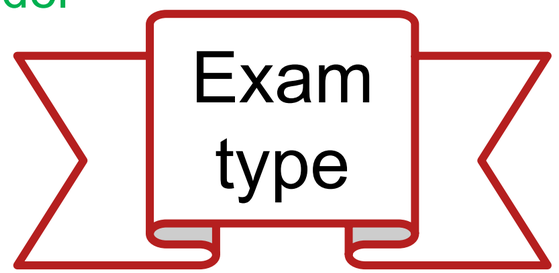
URL: ttpoll.eu
Session ID: cs290

The society RetailProtect has developed a ML model to identify instances of shoplifting in retail shops. For evaluating their model, they use a benchmark in which actors from diverse ethnicities simulate a range of shoplifting actions.

This can lead to (select 1 answer):

- ✅ 0% a. Evaluation bias
- ❌ 0% b. Aggregation bias
- ❌ 0% c. Optimization choices
- ❌ 0% d. Deployment bias

- Evaluation time
- Diverse ethnicities does not guaranty fairness on other attributes (e.g. gender, etc.)
- The benchmark employs **actors** that **simulate** shoplifting instead of real-life scenes -> actions will probably be exaggerated/different from real cases i.e. they will not evaluate correctly the performance of the model



Fairness metrics - 1

URL: ttpoll.eu
Session ID: cs290

Among the metrics below, **which can be used to assess the fairness** of a piece of software? (select all that apply)

0% a. Accuracy

0% b. False Positive Rate

0% c. False Negative Rate

0% d. False Discovery Rate

0% e. False Omission Rate

0% f. Positive Predictive Value

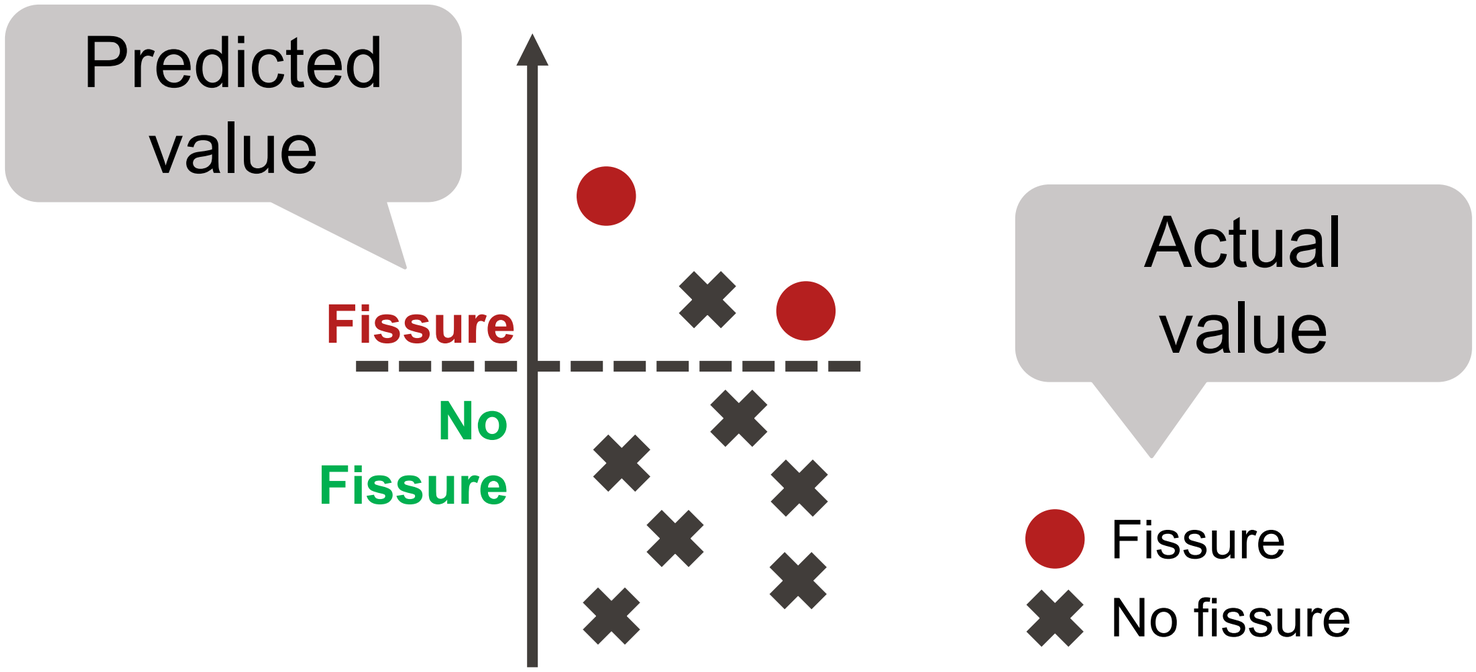
0% g. Negative Predictive Value

0% h. Proportion of positive prediction (also called acceptance rate)

All can be used as long as we compare 2 groups with it

Fairness metrics – 2

The company SuperCrack has developed a model to detect fissures in concrete before they become visible.
They evaluate their model against a benchmark.
The results look like this:

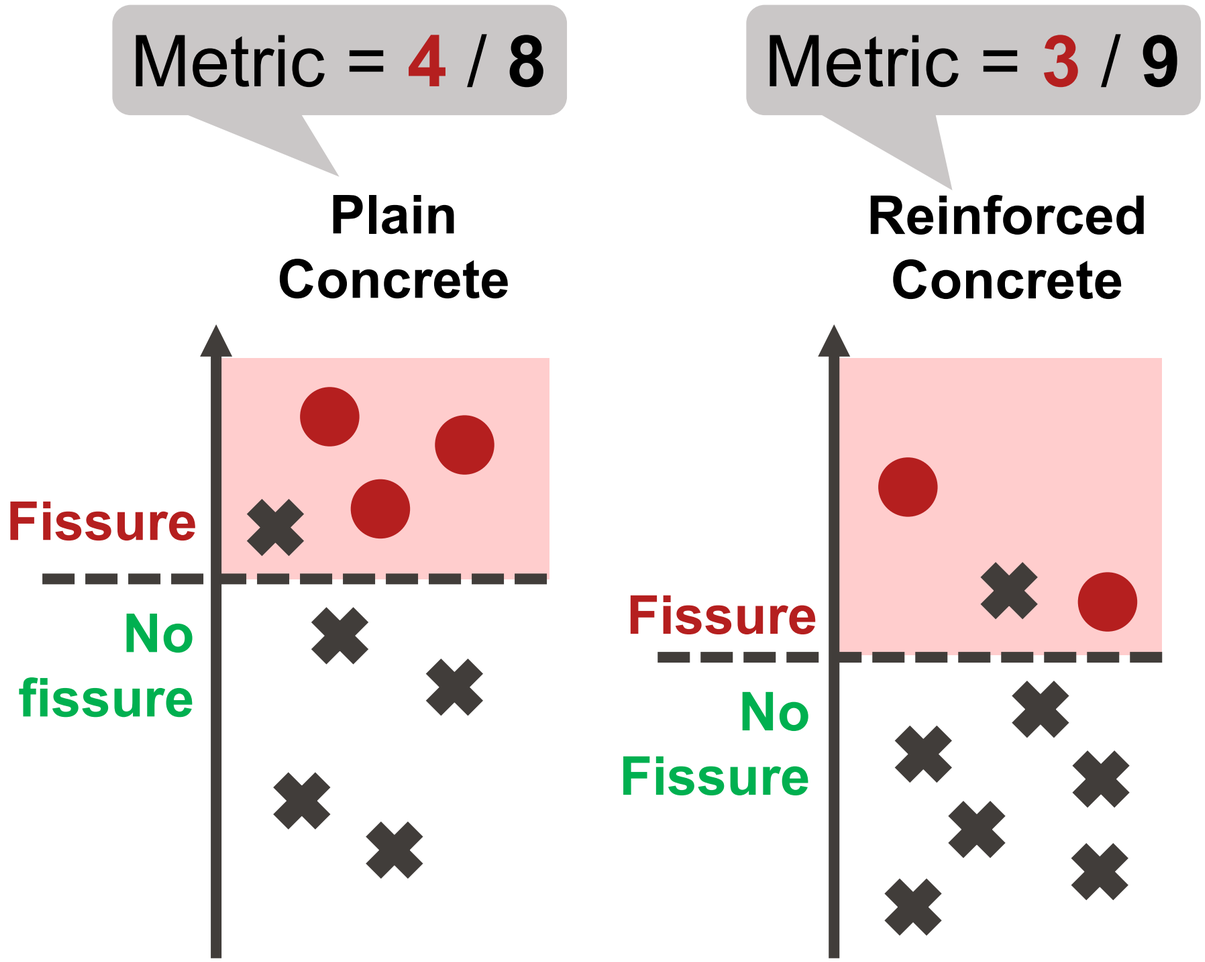


		Predicted: Fissure	Predicted: No Fissure
Actual	Fissure	2	0
Actual	No Fissure	1	6

Fairness metrics – 2a

URL: ttpoll.eu
Session ID: cs290

They want to know whether their model performs equally well for plain concrete and for reinforced concrete. Here are the results:



Which notion of fairness are they using?
(select 1 answer)

❌ 0%
a. Equal accuracy

❌ 0%
b. Error rate balance

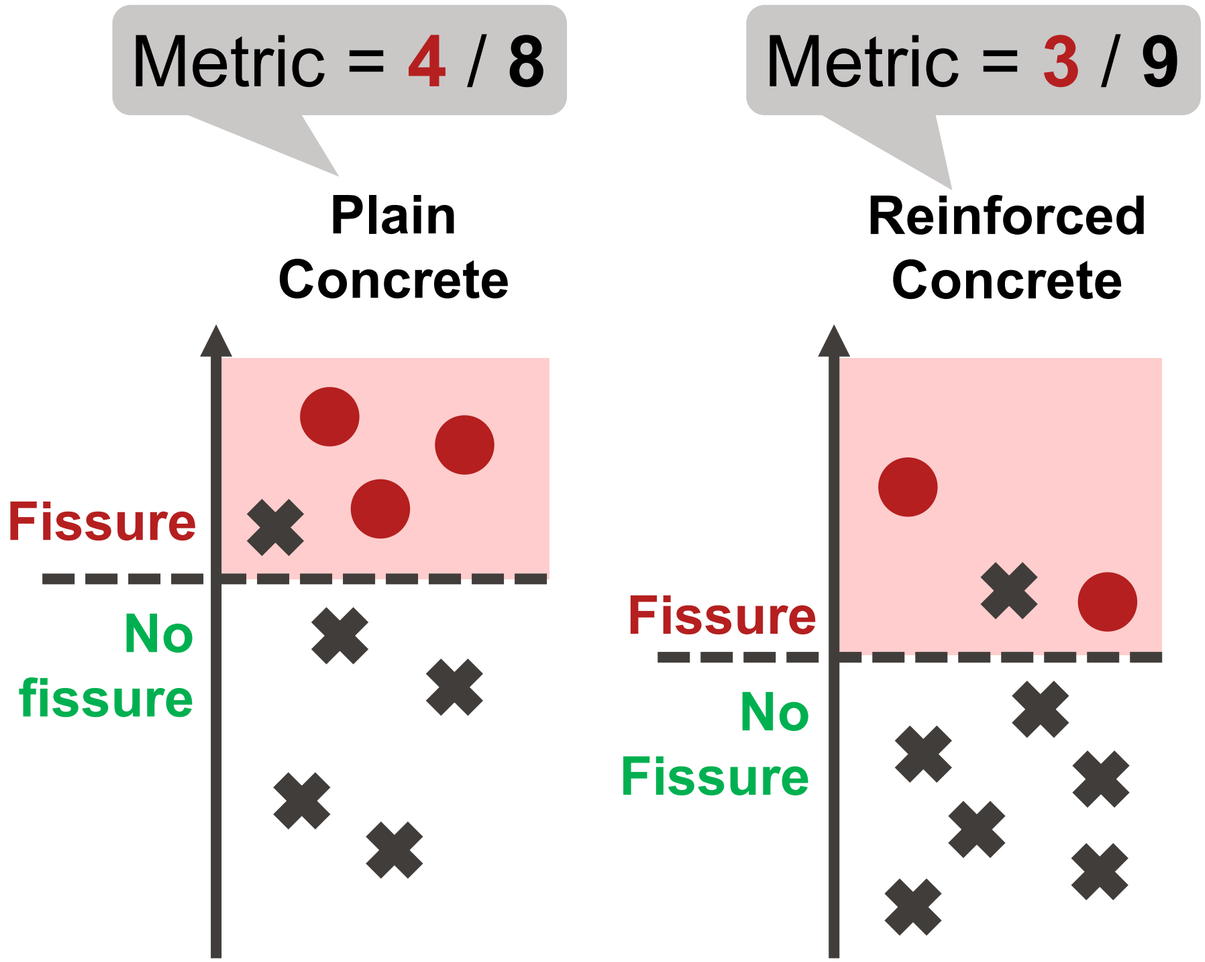
❌ 0%
c. Error parity

✅ 0%
d. Demographic parity

They compare the number of positive predictions (fissure) / total number of samples

Fairness metrics – 2b

URL: ttpoll.eu
Session ID: cs290



According to this metric, is their model fair?
(select 1 answer)

❌ 0% a. Yes

✅ 0% b. No

(✅) 0% c. Other option

- Disparate impact ratio = 0,33 / 0,5 = 0,66
Which is far from 1 or from the tolerated 0,8
- We can question whether it is really about "fairness" in this case…

Case studies

Inclusive Design
(from Fairness 1)

Documents

You need the following documents from **Fairness 1**:

- The **instruction sheet**
- The **Inclusive Design cheatsheet**

Instructions

Read the scenario
With your neighbor, **think about how you would design the app** (feel free to draw sketches, etc.)

Apply the inclusive design strategy:

- Stage 01: Identify the **capabilities** required from users
- Stage 02: Identify **“Non-Average” Users** (NAUs)
- Stage 03: Identify any additional capabilities and non-users and/or minorities

Capabilities

Which capabilities have you identified for your app? (select all that apply)

0% a. Vision

0% b. Touch & Haptic

0% c. Communicate

0% d. Locomotion

0% e. Dexterity

0% f. Other

You should account for capabilities:
- for using the app (interface in particular)
- for the domain-specific task (here for parking a car and accessing city locations) to make the logic of your app inclusive

“Non-Average” users (NAUs)

Which “non-average” users have you identified for your app? (select all that apply)

0% a. Color blind (still able to drive)

0% b. Temporarily injured (e.g., broken elbow, passenger)

0% c. Mom with small kid(s)

0% d. Non-native speaker

0% e. Senior with reduced mobility (still able to drive)

0% f. No smartphone

0% g. Other

Instructions

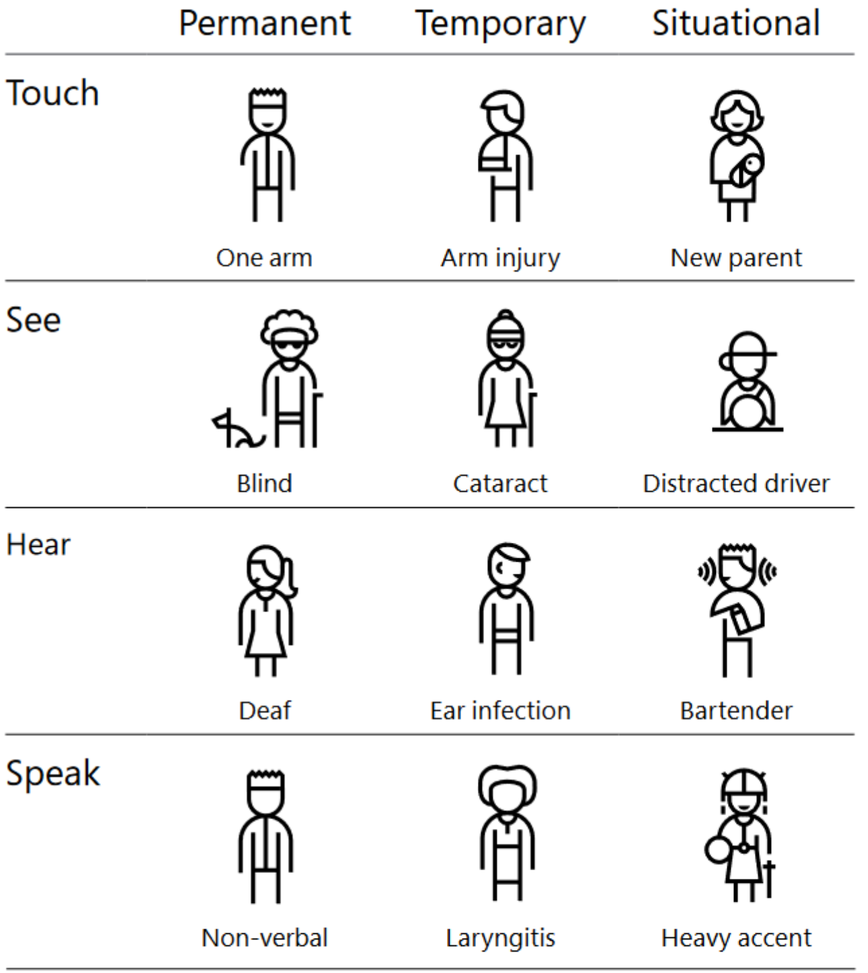
Apply the inclusive design strategy:

- Stage 04: Propose changes to your design that would improve its inclusivity

Overall debriefing of the strategy

There's a great diversity of people out there!!!

- Some choices in design and features can **make software unusable** for some people
- It may not be possible to be inclusive for everyone
- But making software more inclusive usually **benefits everyone**



Inclusive
A Microsoft Design Toolkit

https://inclusive.microsoft.design/tools-and-activities/InclusiveActivityCards.pdf

Datasheets for Datasets

Where to find the cases?

1. Go to **moodle**

2. Find the **link to the case studies** for today: **Fairness 2**
   👉this link will send you to courseware
   (where you can find all the course material)

3. Download:
   - The **instruction sheet**
   - 1 cheatsheet: People Behind The Data

Instructions

Read the datasheet and, **thinking about a range of stakeholders, try to spot:**

1. One **safety** issue
2. One **fairness** issue

If you were to use this dataset for **training a machine learning model able to identify faces**, which type of ethical issue(s) could manifest in the model?

Safety-related issues

URL: ttpoll.eu
Session ID: cs290

Which safety-related issues did you identify?

0% a. Missing consent from image authors
0% b. Missing consent from the photo-hosting website
0% c. Possible re-identification or inference of private info
0% d. Potential offensive content in the images
0% e. Other

All of these are safety issues with this dataset as documented in the provided datasheet

Fairness-related issues

URL: ttpoll.eu
Session ID: cs290

Which fairness-related issues did you identify?

0% a. Unclear population represented by the dataset
0% b. No information about subgroups representation
0% c. Potential biased error rates in alignment + cropping process
0% d. Other

All of these are fairness issues with this dataset as documented in the provided datasheet

Issues in resulting ML model

URL: ttpoll.eu
Session ID: cs290

If you were to use this dataset for training a machine learning model able to **identify faces**, which type of ethical issue(s) **could** manifest in the model? (select all that apply)

0% a. Model unfit for the aimed population
0% b. Differential error rates for subgroups
0% c. Other

All of these are issues that could manifest in a ML model trained on this data

Overall debriefing of the strategy

Data scientists and Machine Learning engineers who **use a datasheet** when thinking about a ML problem **identify ethical issues**:

- Earlier
- More often

It’s not super shiny or exciting, but it seems to help!

Boyd, K. L. (2021). Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 438:1-438:27. https://doi.org/10.1145/3479582

People Behind The Data

Instructions

Documents you have (Stage 01):

- Raw COMPAS questionnaire
- Dataset provided by ProPublica (real people!)
  (download it and open it with Excel)

Apply the “People behind the data” strategy:

- Stage 02: read the questionnaire, select 5 variables of interest
- Stage 03: select 1 row in the dataset, based on characteristics of your choice

  👉 combine information from the questionnaire and from the data to **imagine the profile and story of one person behind the data**

Reflect

Answer the following questions:

- What have you learned about the data based on your exploration?
- Which potential harmful impacts could using this data generate?
- What would be your next steps: would you use these data? What other possibilities would you have?

Overall debriefing of the strategy

When working with data, we can easily forget that there are people behind the numbers…

This strategy helps you practice with:

- Empathy
- Storytelling

What’s next?

	Monday	Tuesday
13 Oct – 17 Oct	Fairness 2 cases	Graded notebook 1
20 Oct – 24 Oct	Autumn break	
27 Oct – 30 Oct	Debriefing Graded 1	Mock Test
3 Nov – 7 Nov	Debriefing Mock Test (in CO3)	Sustainability 1 notebook