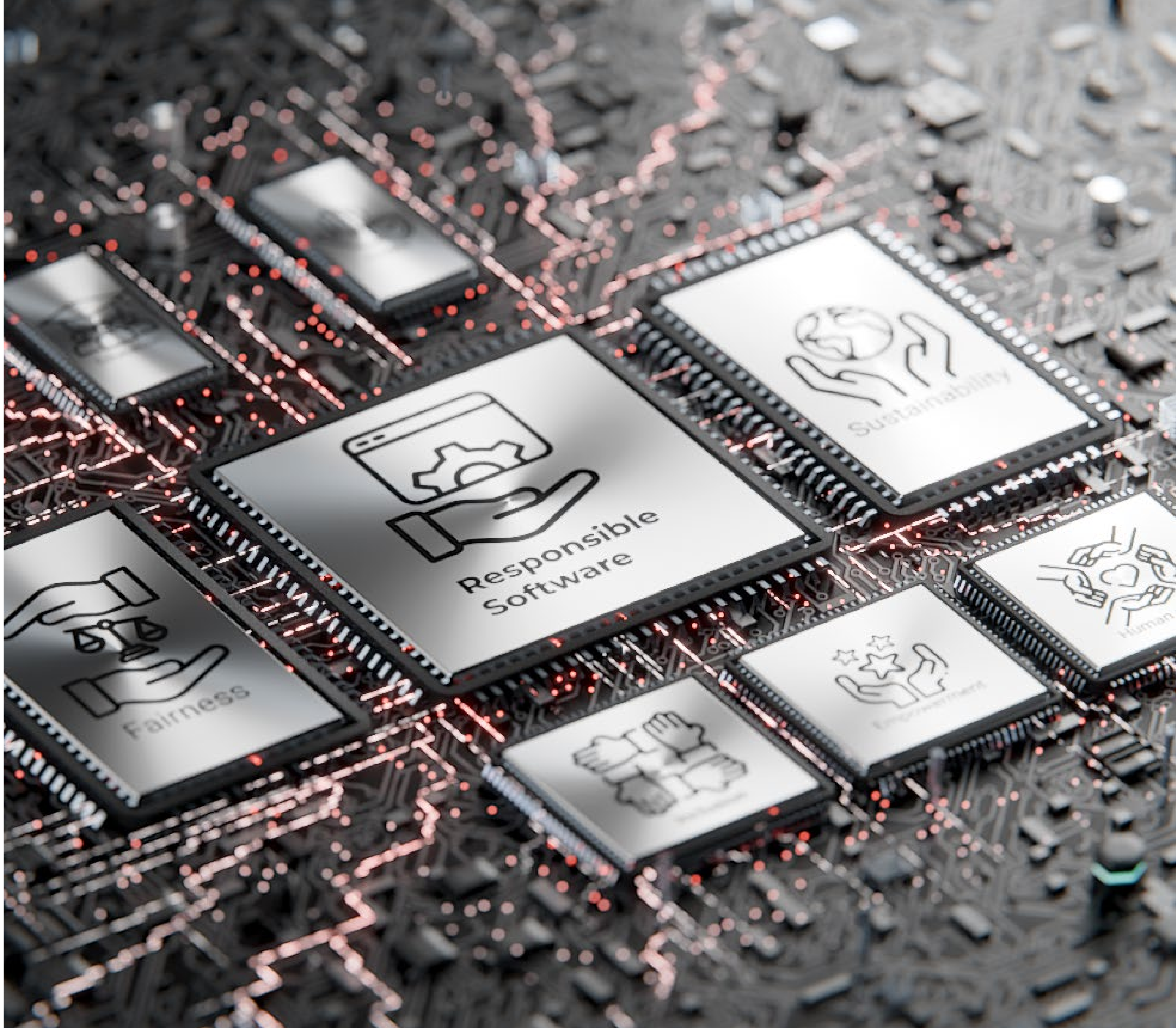
Fairness 2 Review & Case studies
14 oct.

Cécile Hardebolle

Responsible Software



Agenda for today

1. Upcoming dates in the course
2. Interactive review questions on Fairness 2
3. Case studies:
   a) Inclusive design (from Fairness 1)
   b) Datasheets for datasets
   c) People behind the data & COMPAS
   (Harms modeling can be done as training at home)

Next dates

	Monday (SG1)	Tuesday (Computer Rooms)
14 Oct – 18 Oct	Fairness 2 cases	Graded Assignment 1
21 Oct – 25 Oct	Autumn break	
28 Oct – 1 Nov	Debriefing Graded 1	Blank Test (in **SG1**)
4 Nov – 8 Nov	Debriefing Blank Test	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 assignment & the test

Second part of the course

Date	Week	Lecture (Monday 8h15-10h)	Exercise session (Tuesday 8h15-10h)	Independent study (due before the following Monday)
09/09	1	Introduction + cases	Tutorial notebook	Introduction videos and quizzes
16/09	2	public holiday	Safety 1 notebook	Safety 1 videos and quizzes
23/09	3	Safety 1 cases	Safety 2 notebook	Safety 2 videos and quizzes
30/09	4	Safety 2 cases	Fairness 1 notebook	Fairness 1 videos and quizzes
07/10	5	Fairness 1 cases	Fairness 2 notebook	Fairness 2 videos and quizzes
14/10	6	Fairness 2 cases	Graded assignment 1	-
21/10			Autumn break	
28/10	7	Debriefing assignment	Blank test (in SG1)	-
04/11	8	Blank test debriefing	Sustainability 1 notebook	Sustainability 1 videos and quizzes
11/11	9	Sustainability 1 cases	Sustainability 2 notebook	Sustainability 2 videos and quizzes
18/11	10	Sustainability 2 cases	Empowerment 1 notebook	Empowerment 1 videos and quizzes
25/11	11	Empowerment 1 cases	Graded assignment 2	-
02/12	12	Debriefing assignment	Empowerment 2 notebook	Empowerment 2 videos and quizzes
09/12	13	Empowerment 2 cases	Conclusion + Q&A (in SG1)	Conclusion videos and quizzes
16/12	14	Final exam	-	-

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

It’s a good idea to do the Blank Test!

There are no stakes, it’s not graded, we don’t collect copies!!!

Goals =

- Get familiar with the **format** of the exam
- See what type of single choice **questions** you will get
- See how the **cases** look like
- Check how you’re doing with the **time limit** **(1h30)**

👉 Identify **where you need to improve**, so that you can better focus your revisions!

Blank Test

URL: ttpoll.eu
Session ID: cs290

I plan to do the Blank Test in class in SG1 on 29 Oct.:

88% a. Yes

4% b. No

8% c. I don’t know yet

Blank Test

URL: ttpoll.eu
Session ID: cs290

Format of the test:

58% a. I would like to get the test **printed on paper**

40% b. I prefer to get a PDF on moodle

2% c. Other

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):

- 40% a. Evaluation bias
- 44% b. Aggregation bias
- 4% c. Optimization choices
- 11% d. Deployment bias

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

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 develops a ML model to identify instances of shoplifting in retail shops. They evaluate their model on a benchmark in which actors from diverse ethnicities simulate a range of shoplifting actions.

This can lead to (select 1 answer):

✅ 61% a. Evaluation bias

❌ 6% b. Aggregation bias

(❌) 13% c. Optimization choices

(✅) 20% d. Deployment bias

- Evaluation time
- Diverse ethnicities does not guaranty fairness on other attributes (e.g. gender, etc.)
- The benchmark employs **actors** instead of real-life scenes -> does not represent the target problem [Could be also considered a form of deployment bias]
- (Using the benchmark can help identify optimization options, but it is a late stage for that)

Fairness metrics - 1

URL: ttpoll.eu
Session ID: cs290

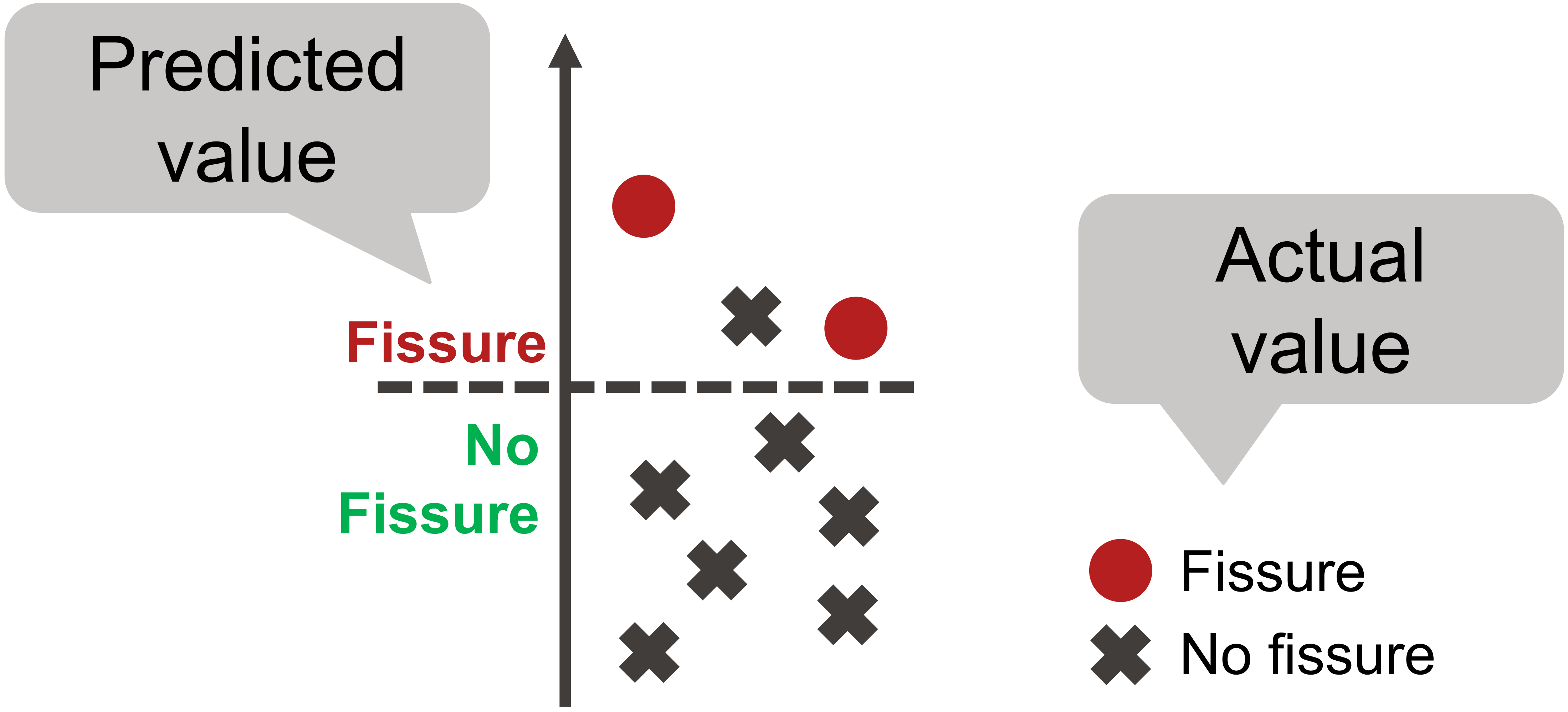
Among the metrics below, **which can be used to assess the fairness** of a piece of software? (select all that apply)

6% a. Accuracy
15% b. False Positive Rate
16% c. False Negative Rate
13% d. False Discovery Rate
15% e. False Omission Rate
12% f. Positive Predictive Value
11% g. Negative Predictive Value
12% h. Positive prediction rate (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:

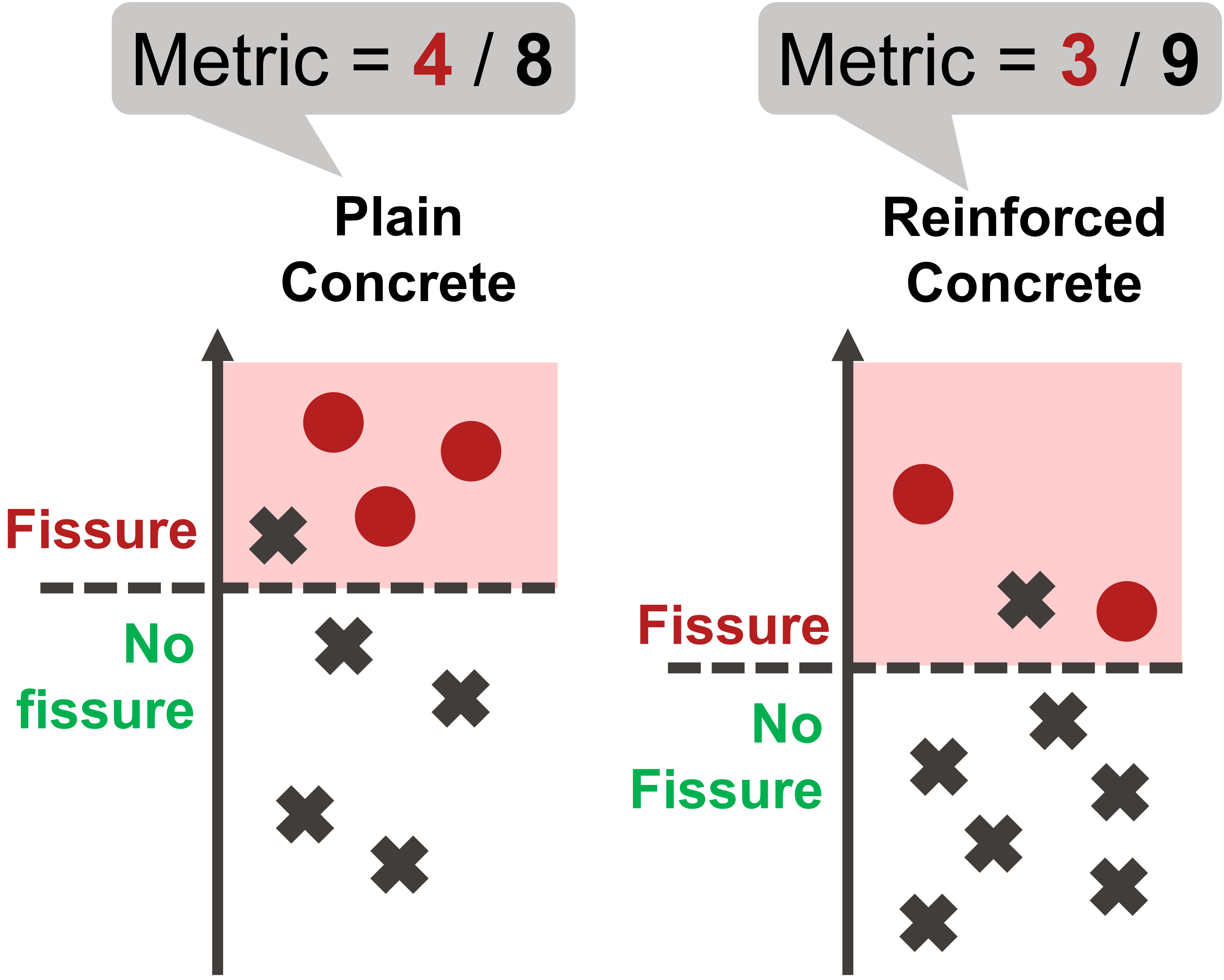


Actual \ Predicted	Fissure	No Fissure
Fissure	2	0
No Fissure	1	6

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:



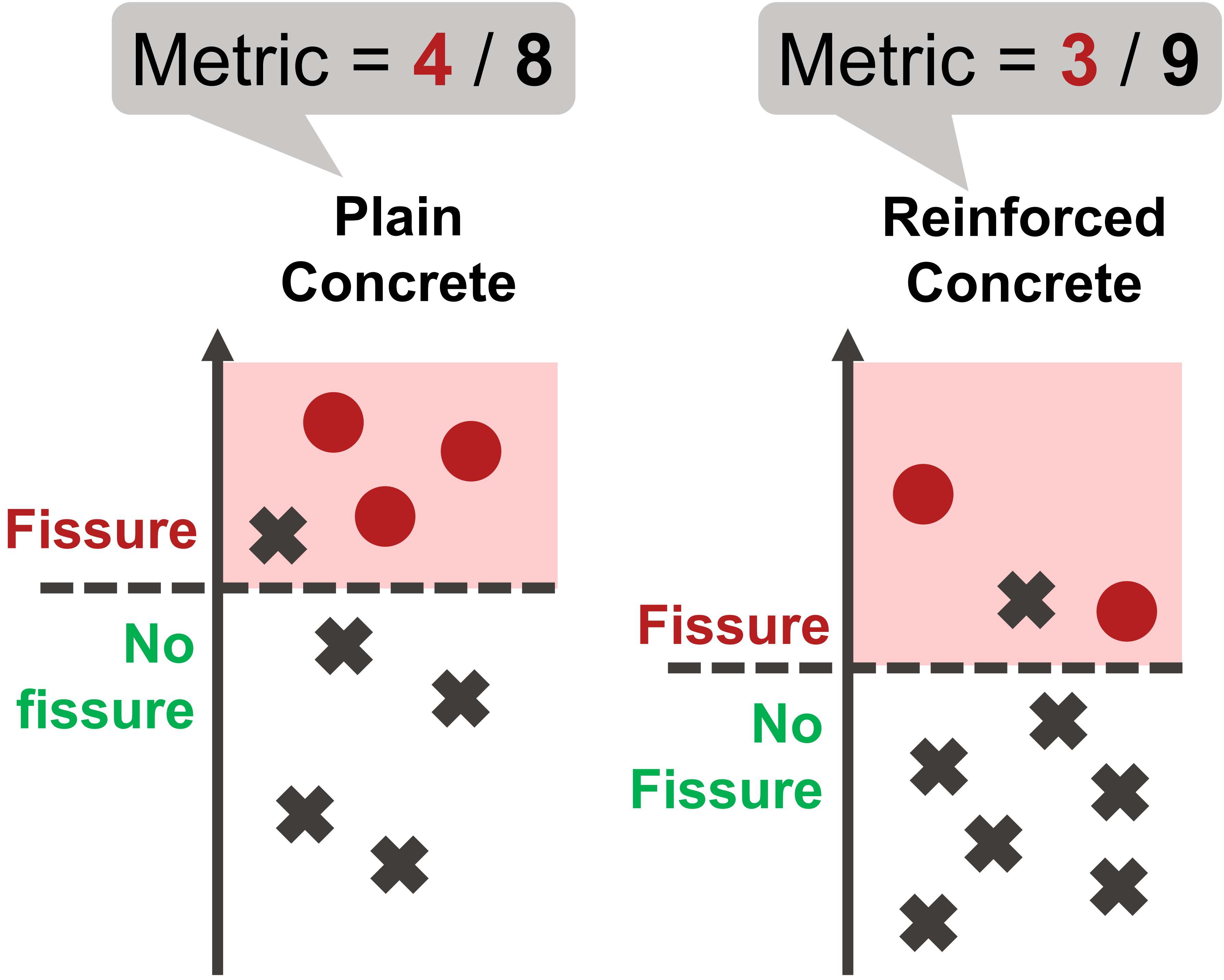
Which metric are they using? (select 1 answer)

- 13% a. Equal accuracy
- 15% b. Error rate balance
- 11% c. Error parity
- 61% 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)

- 12% a. Yes
- 61% b. No
- 27% c. Other option

- Disparate impact ratio = 0,33 / 0,5 = 0,66
Which is far from 1 or even from the tolerated 80%
- We can question whether it is really about "fairness" in this case…

Case studies

Case studies

URL: ttpoll.eu
Session ID: cs290

Regarding the case studies, I think that:

11% a. No solutions are provided

63% b. Some “proposed answers” are provided

27% c. I don’t know

Each week on Monday evening in Courseware you get “proposed answers” for all the case studies!

Inclusive Design
(from Fairness 1)

Inclusive Design (Fairness 1)

Have you done the **Inclusive Design case** from Fairness 1?

Documents

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

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

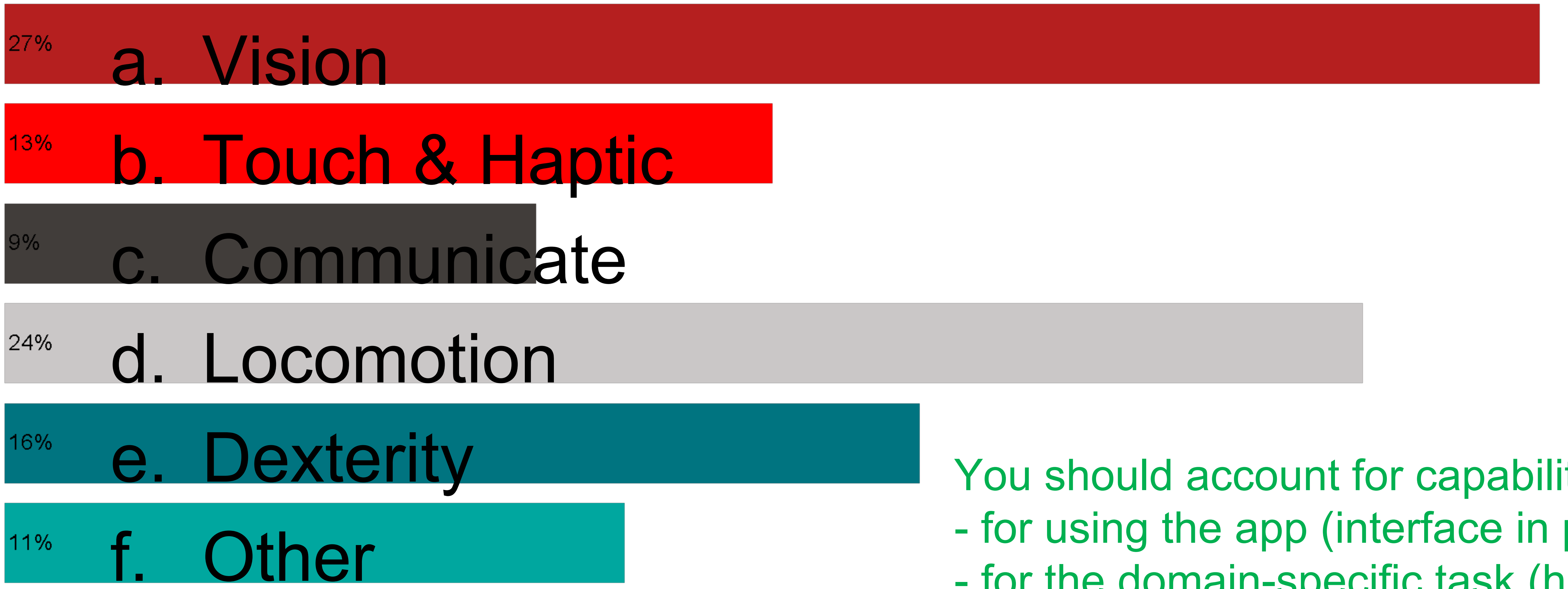
Instructions

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)

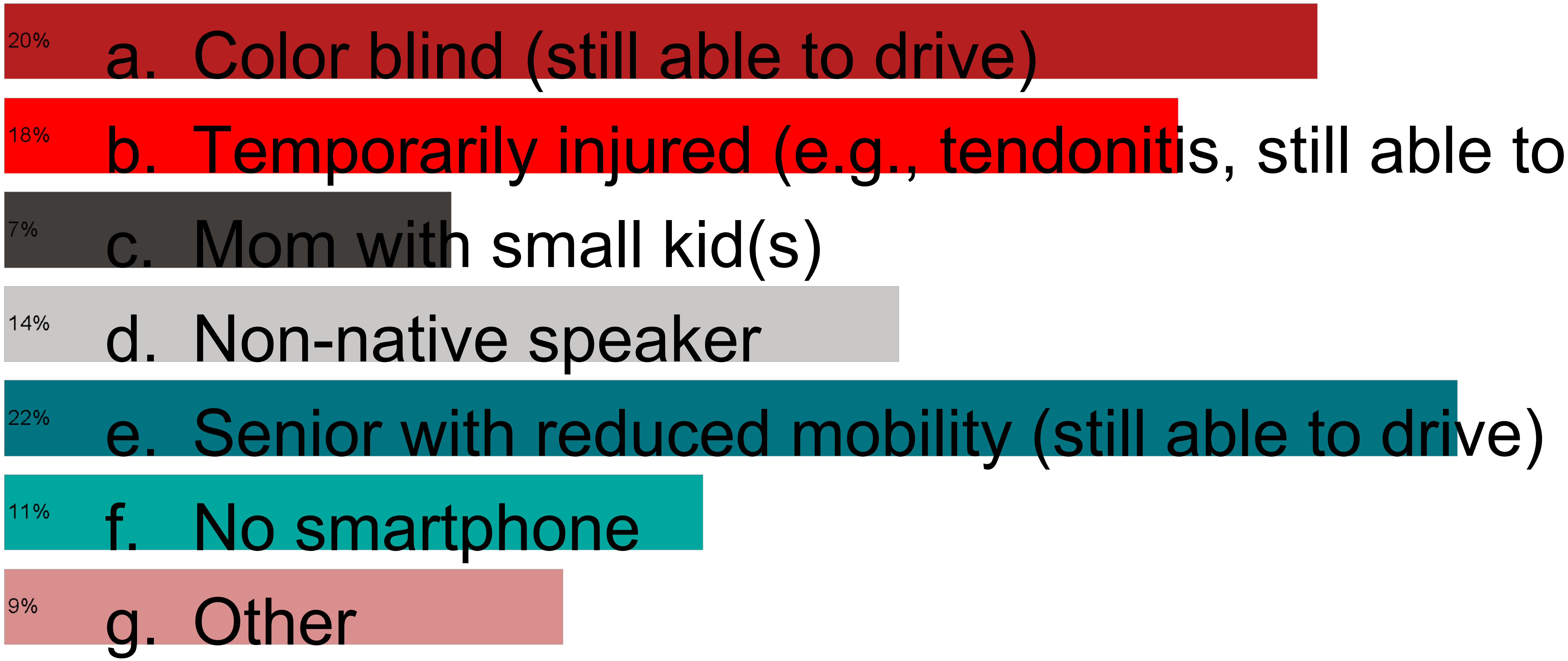


You should account for capabilities:
- for using the app (interface in particular)
- for the domain-specific task (here for parking a car)

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)



Instructions

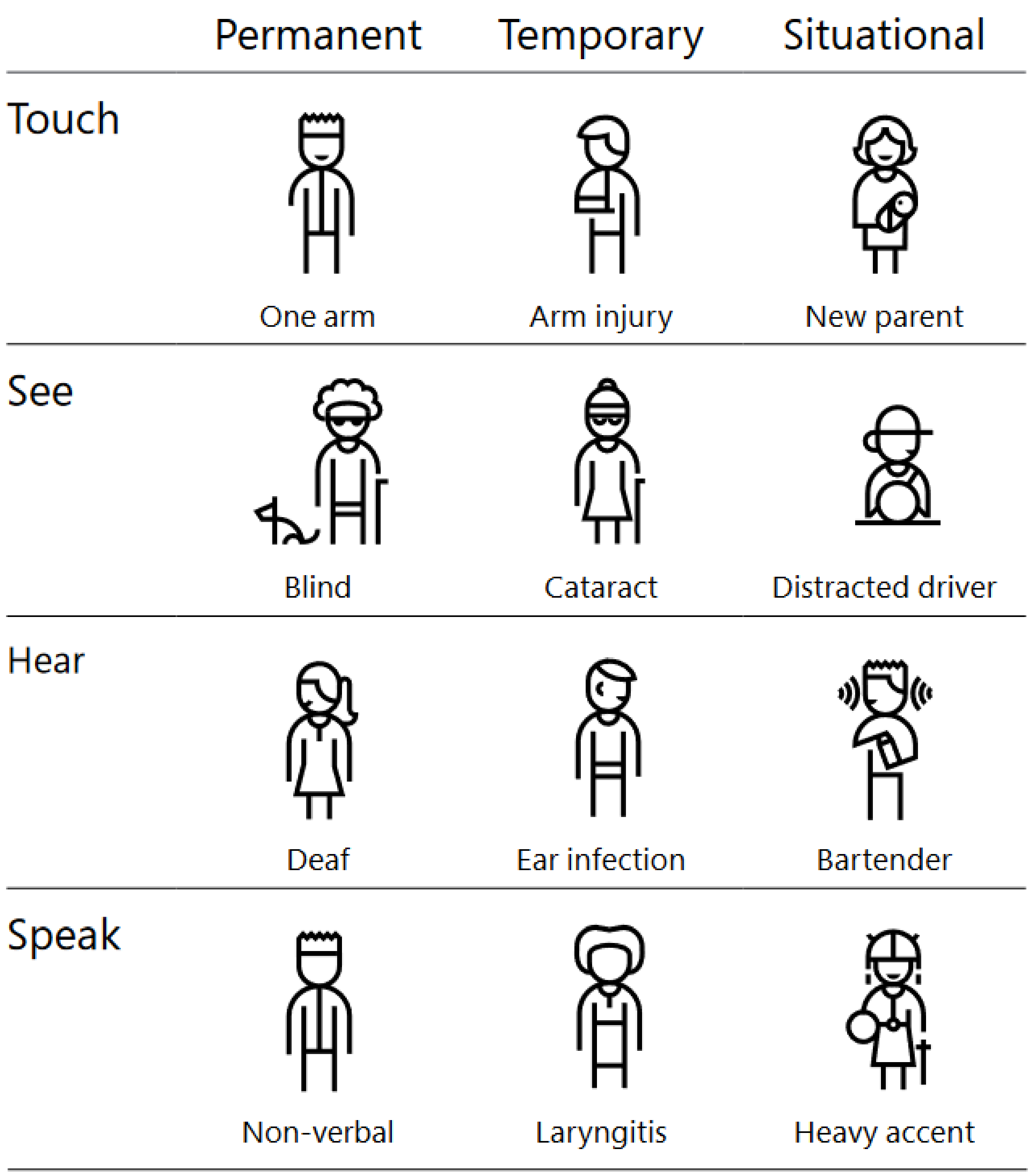
Apply the inclusive design strategy:

- 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: **<u>Fairness 2</u>**
   👉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 <u>range of stakeholders</u>, 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?

25% a. Missing consent from image authors
11% b. Missing consent from the photo-hosting website
31% c. Possible re-identification or inference of private info
29% d. Potential offensive content in the images
4% 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?

- 40% a. Unclear population represented by the dataset
- 29% b. No information about subgroups representation
- 24% c. Potential biased error rates in alignment + cropping process
- 7% 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)

41% a. Model unfit for the aimed population

52% b. Differential error rates for subgroups

7% 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 or any other software)

Apply the "People behind the data" strategy:

- Stage 02: read the questionnaire, select a few variables of interest
- Stage 03: select 2 rows in the dataset, based on one demographic attribute of your choice
  👉 combine information from the questionnaire and from the data to **imagine the profile and stories of the people 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 (SG1)	Tuesday (Computer Rooms)
14 Oct – 18 Oct	Fairness 2 cases	Graded Assignment 1
21 Oct – 25 Oct	Autumn break	
28 Oct – 1 Nov	Debriefing Graded 1	Blank Test (in **SG1**)
4 Nov – 8 Nov	Debriefing Blank Test	Sustainability 1 notebook