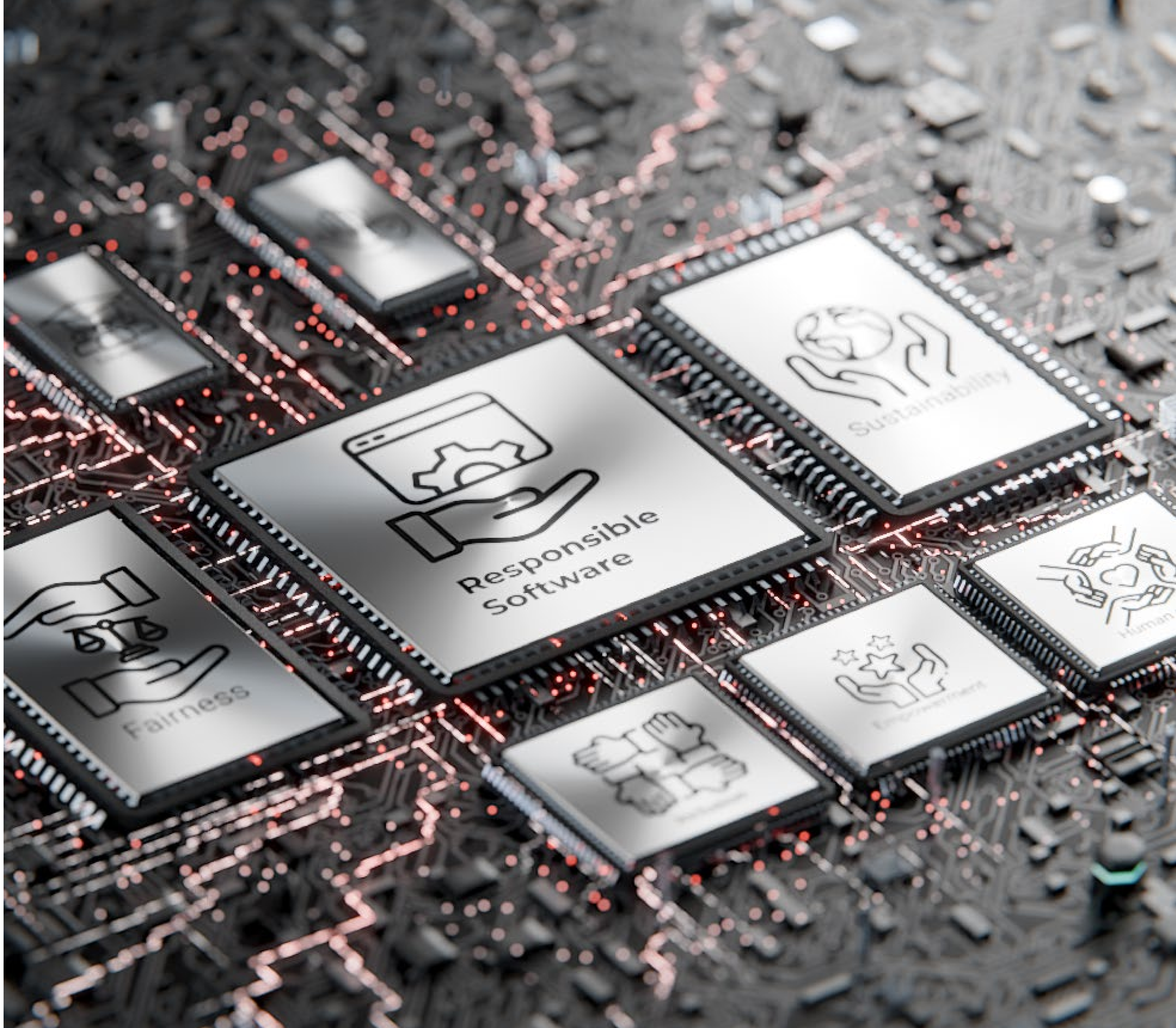
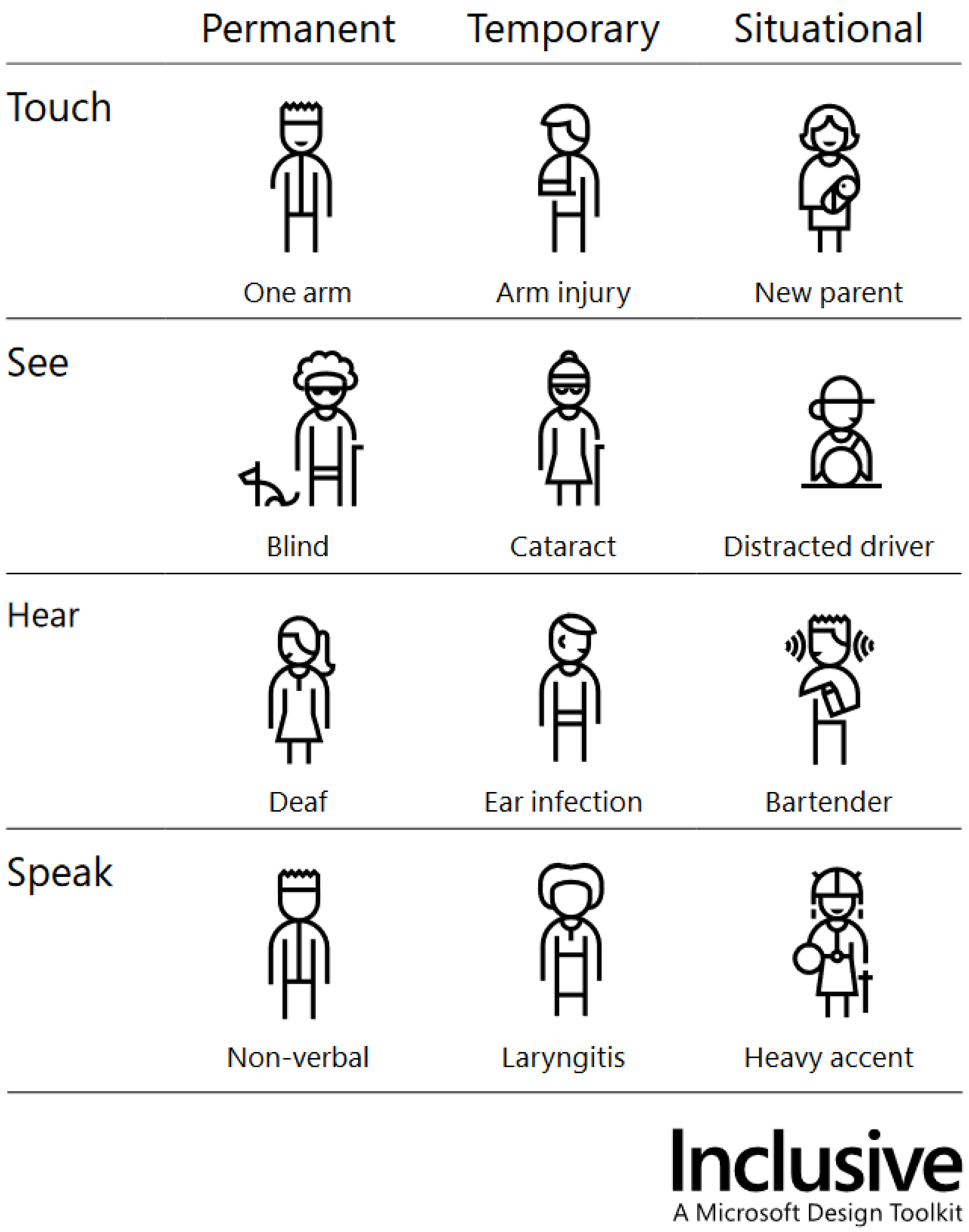
Graded 1 Debriefing
27 oct.



Cécile Hardebolle

Responsible Software

Happy to be back!



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

Agenda for today

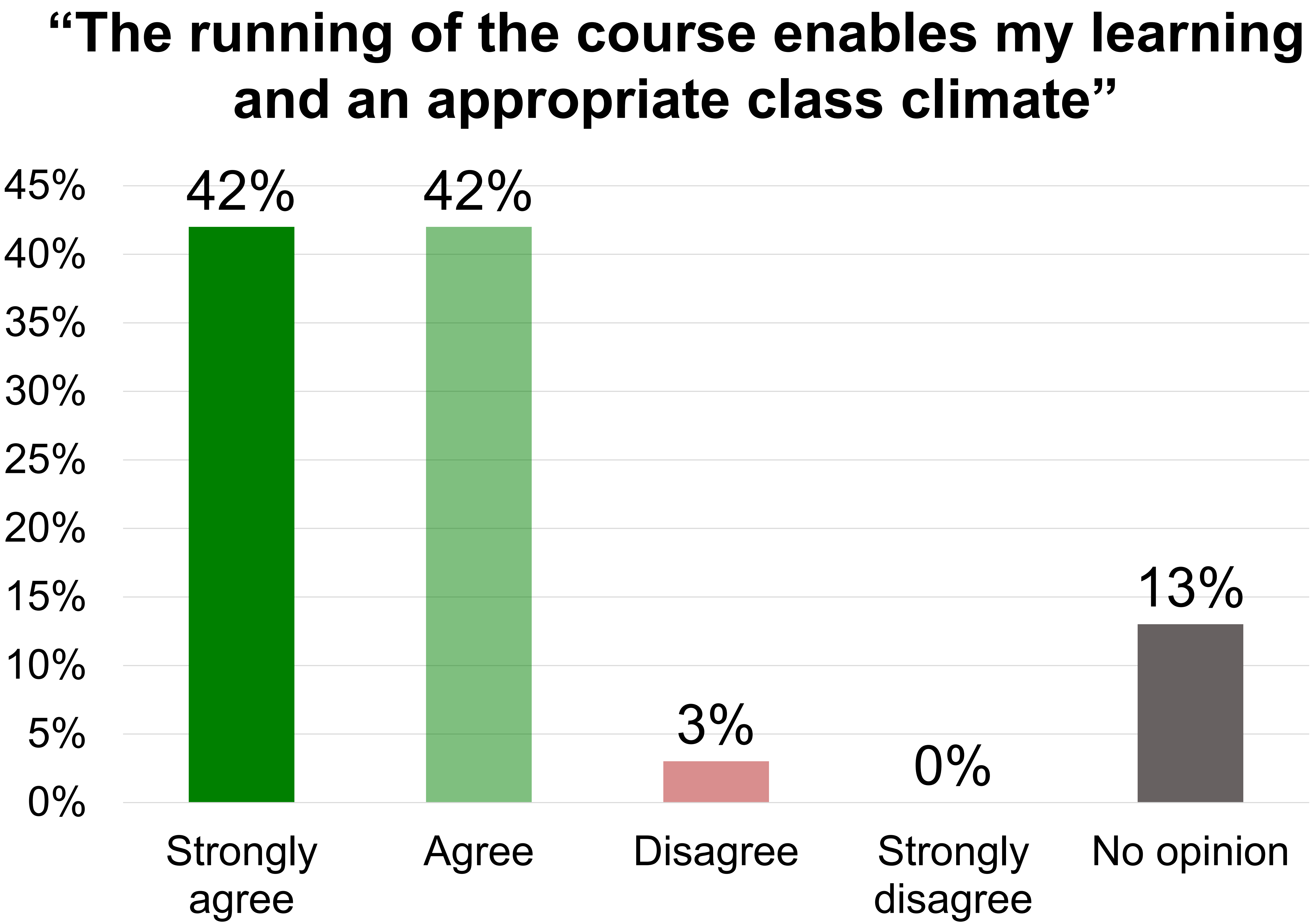
1. Debriefing of the indicative evaluation

2. Feedback on the Graded 1 assignment

Indicative evaluation

Overall indicator

Response rate:
31 / 242
= **13%**



Most frequent comments

Positive

- Really **interesting** content and relevant skills
- Good overview of **current** ethical issues with software
- Pragmatic and **practical** approach to ethics
- **Notebooks**:
  - Real world considerations
  - Cite current research

Negative

- **Flipped** classroom format
- **Workload**:
  - MOOC
  - Notebooks (long, lots of reading)
- **Overlap** with other courses
- Case studies:
  - Lack of **realism**
  - Too many similar strategies
- Rank by vote in **SpeakUp**

Measures taken

For this year:

- ■ Change SpeakUp use and debriefing
- ■ Graded case: will be on existing software instead of made-up scenarios

For next year:

- ■ Mini-project on real case with increased % of grade (instead of graded case)
- ■ Reduce number of strategies

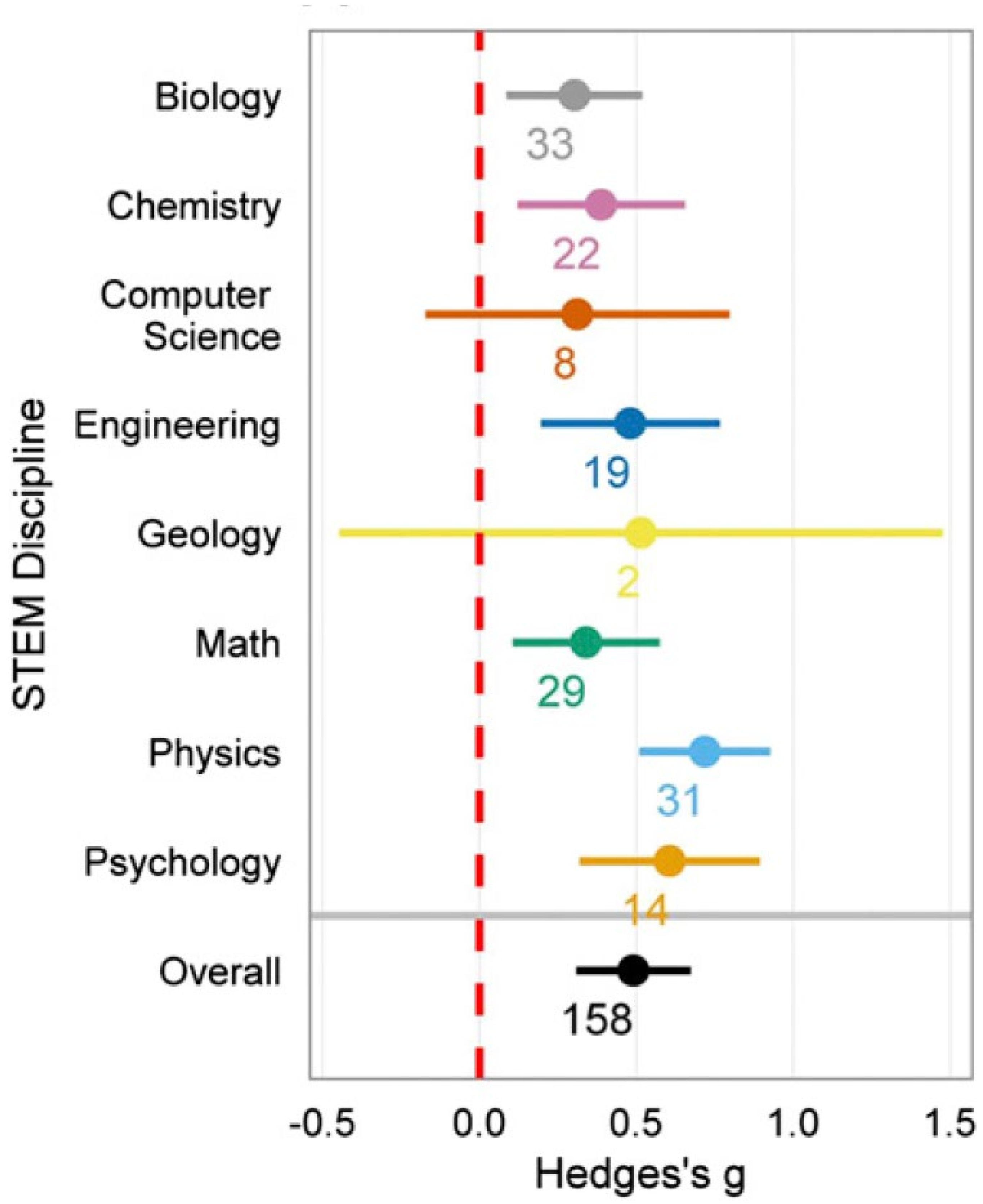
Why a flipped classroom?

(Lo & Hew, 2019; Gong et al., 2024)

- Research indicates a significant and medium **positive effect on learning** in favor of flipped learning over traditional learning models (Hedge's g from 0.263 to 0.289)
- Particularly when the following activities are used:
  - A "link activity" such as a **brief review quiz at the start of class**
  - **Both group and individual activities** in class
  - **Both exercises and quizzes** after class
- Expected benefits:
  - Self-paced learning
  - More **active learning** (e.g., problem solving, quizzes, etc.)

Active Learning

(Freeman et al., 2014)

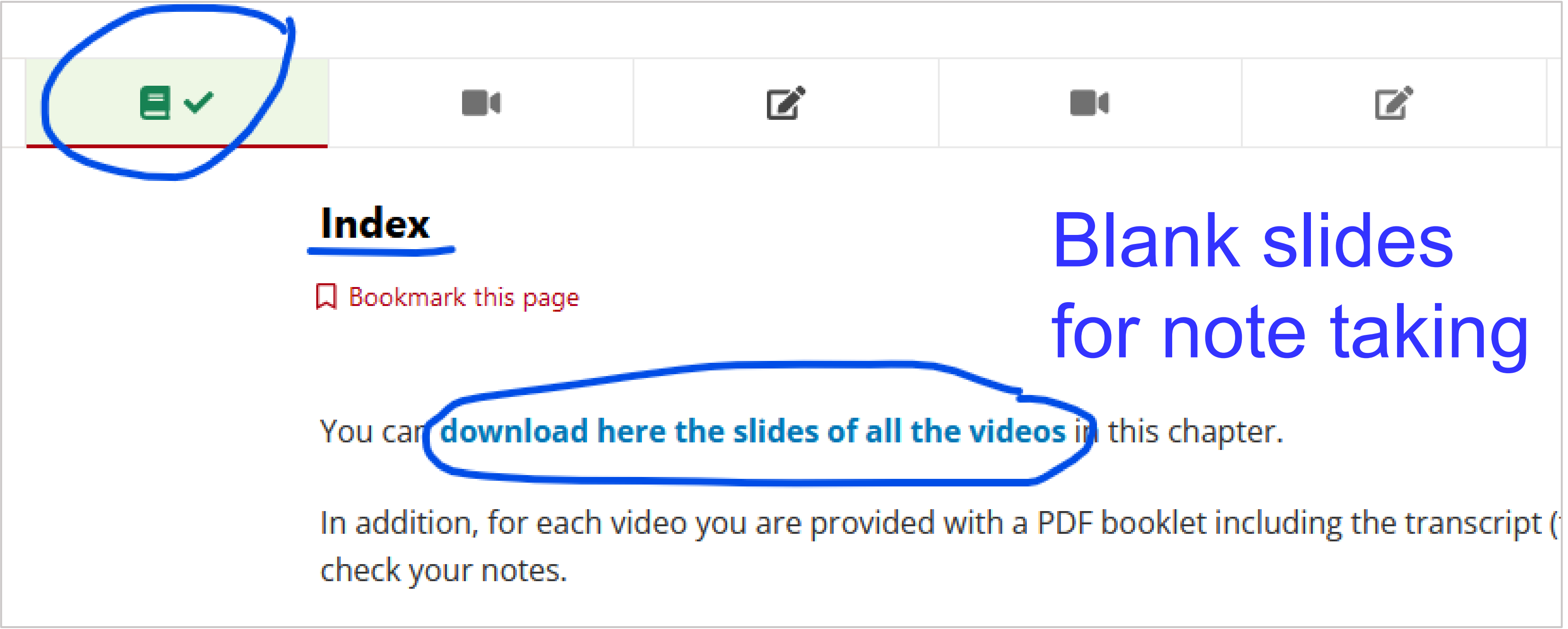


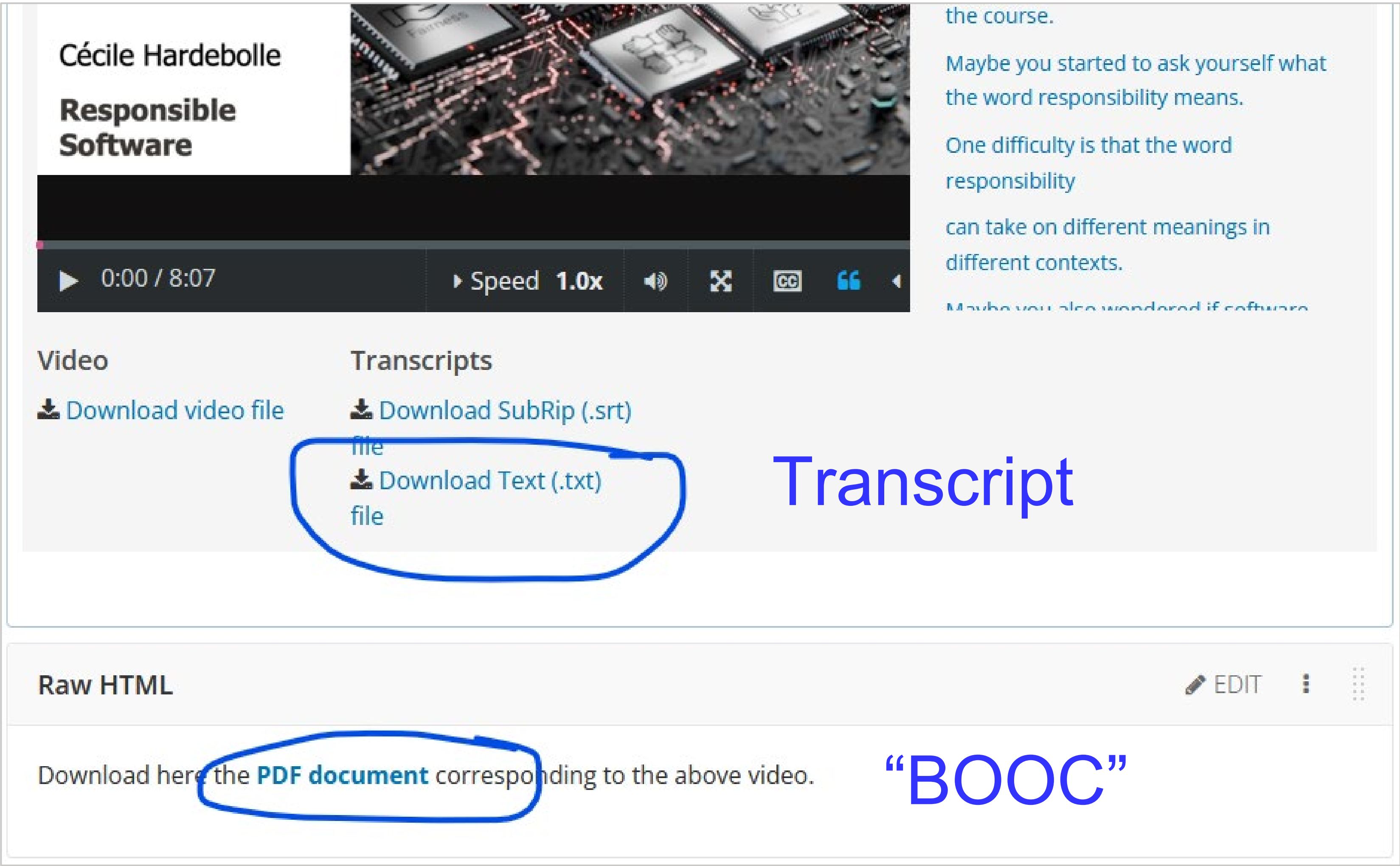
"[…] **average examination scores improved by about 6%** in active learning sections, and that students in classes with **traditional lecturing were 1.5 times more likely to fail** than were students in classes with active learning."

(Freeman et al., 2014)

Learning more in <u>less time</u> with videos

- Use the **blank slides** provided first "page" in the Theory section (called "Index")
- Watch a **little bit slower** than normal speed to be able to **take notes <u>without stopping the video</u>** like you would do in a lecture
  - Notes should be **<u>synthetic</u>** i.e., reduced to the main points
- You can use the **transcript** or the "BOOC" (= slides + transcript) to **check your notes** (faster to read!)





Identifying what is important

Learning goals

- Define **Machine Learning (ML)** and explain conceptually how it works on a (very) simplified example
- Describe different “flavors” of ML
- Identify **sources of biases** in ML:
  - Bias in **data**
  - Bias in the **ML life cycle**: development, evaluation and deployment

Conclusion

- Machine Learning: family of techniques that **infer patterns** from **data** that can be **generalized** to unseen data
- Different “flavors” of ML
- Sensitive to **bias in data**
- Sensitive to **biases in the ML development life cycle**
  - Development
  - Evaluation
  - Deployment

Links with real world and current research

■ Theory (videos)

Intersectional group fairness

(Ghosh et al., 2021; Maheshwari et al., 2023)

Issue: group-based criteria may miss unfairness against people at the **intersection** of multiple groups

Intersectional group fairness: criteria for combined groups

⚠️ Issues:

- Number of combined groups

References

- Ghosh, A., Genuit, L., & Reagan, M. (2021). Proceedings of Machine Learning Research
- Maheshwari, G., Bellet, A., ... & Kelle Definition. In H. Bouamor, J. Pino, & K. Bali Language Processing (pp. 9018–9032). Ass main.558
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O Innovations in Theoretical Computer Scienc
- Castelnovo, A., Crupi, R., Greco, G., Regoli, fairness metrics landscape. Scientific Report
- Barocas, S., Hardt, M., & Narayanan, A. (20
- Pearl, J., & Mackenzie, D. (2018). The Book
- Caton, S., & Haas, C. (2024). Fairness in Ma https://doi.org/10.1145/3616865
- Suresh, H., & Guttag, J. (2021). A Framewor Cycle. Equity and Access in Algorithms, Med
- Pessach, D., & Shmueli, E. (2022). A Review 51:44. https://doi.org/10.1145/3494672
- Mittelstadt, B., Wachter, S., & Russell, C. (20 egalitarianism by default (SSRN Scholarly P
- Alves, G., Bernier, F., Couceiro, M., Makhlou related tensions. EURO Journal on Decision

Fair Without Leveling Down: A New Intersectional Fairness Definition

Gaurav Maheshwari, Aurélien Bellet, Pascal Denis, Mikaela Keller
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France
first_name.last_name@inria.fr

Abstract

In this work, we consider the problem of intersectional group fairness in the classification setting, where the objective is to learn discrimination-free models in the presence of several intersecting sensitive groups. First, we illustrate various shortcomings of existing fairness measures commonly used to capture intersectional fairness. Then, we propose a new definition called the α-Intersectional Fairness, which combines the absolute and the relative performance across sensitive groups and can be seen as a generalization of the notion of differential fairness. We highlight several desirable properties of the proposed definition and analyze its relation to other fairness measures. Finally, we benchmark multiple popular

Female African-Americans). For example, Buolamwini and Gebru (2018) showed that commercially available face recognition tools exhibit significantly higher error rates for darker-skinned females than for lighter-skinned males. Similar observations have been made by several studies in NLP including contextual word representation (Tan and Celis, 2019), and generative models (Kirk et al., 2021). These findings resonate with the analytical framework of *intersectionality* (Crenshaw, 1989), which argues that systems of inequality based on various attributes (like gender and race) may "intersect" to create unique effects.

To capture these effects in the context of machine learning, several intersectional fairness measures

■ Case studies

Case studies

Bookmark this page

👉 Download the **instruction sheet for the case studies** as well as this **data archive** to be used in Case 2.

You have cheatsheets for the strategies that can be useful:

- Cheatsheet for "The people behind the data"

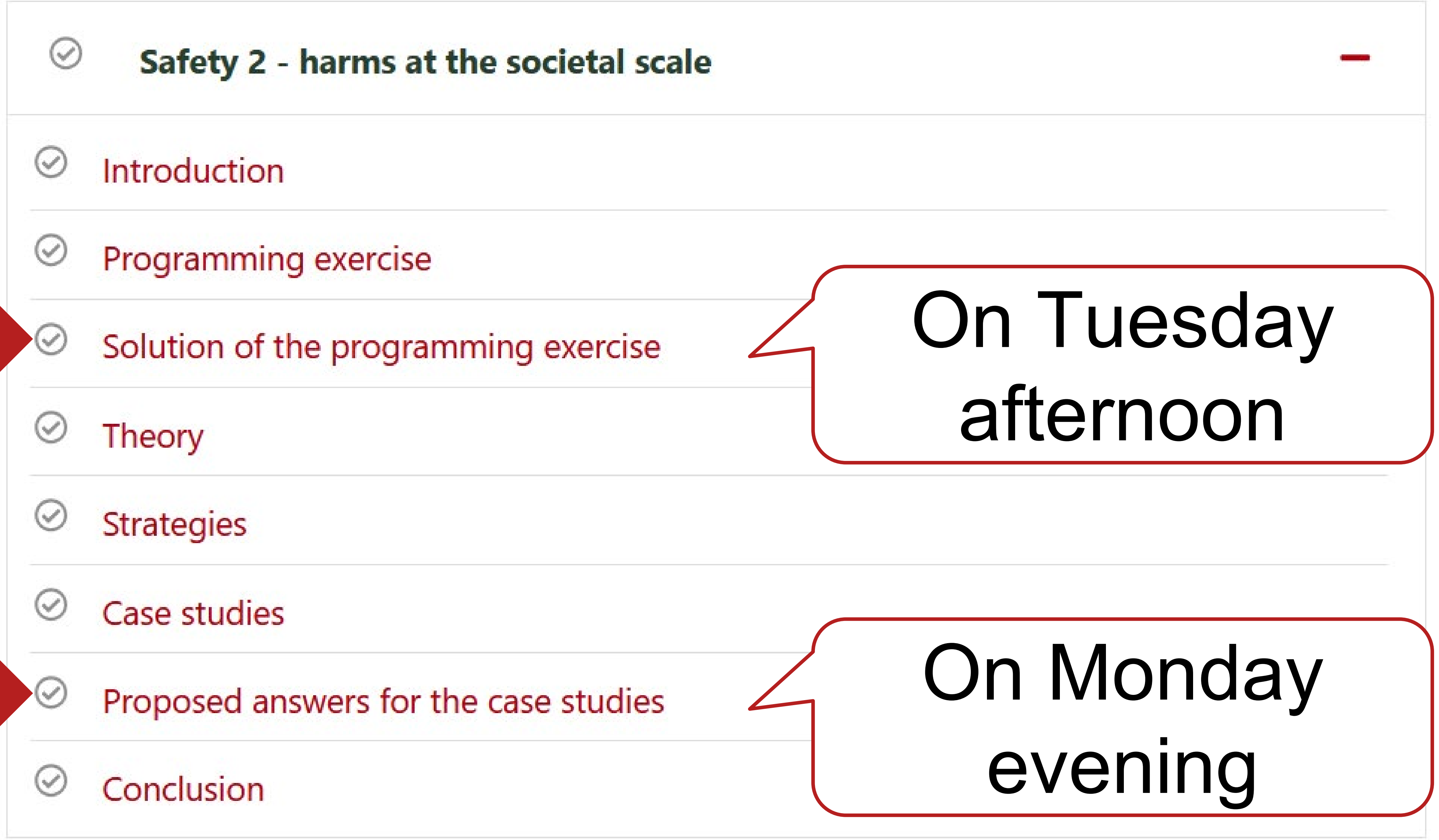
INFO DE DÉBOGAGE POUR LE PERSONNEL

Want to see some real-life cases?
Take a look at these press articles:

- People interviewed by AI for jobs face discrimination risks, Australian study warns (May 2025). The Guardian.
- Lawsuit claims discrimination by Workday's hiring tech prevented people over 40 from getting hired (May 2025). CNN Business.
- AI overwhelmingly prefers white and male job candidates in new test of resume-screening bias (October 2024). GeekWire.
- Microsoft Provided Gender Detection AI on Accident (October 2024). 404 Media.

About solutions

- Outline of the course on courseware:



On Tuesday afternoon

On Monday evening

Feedback on
Graded Notebook 1

Advice for next time

Logistics:

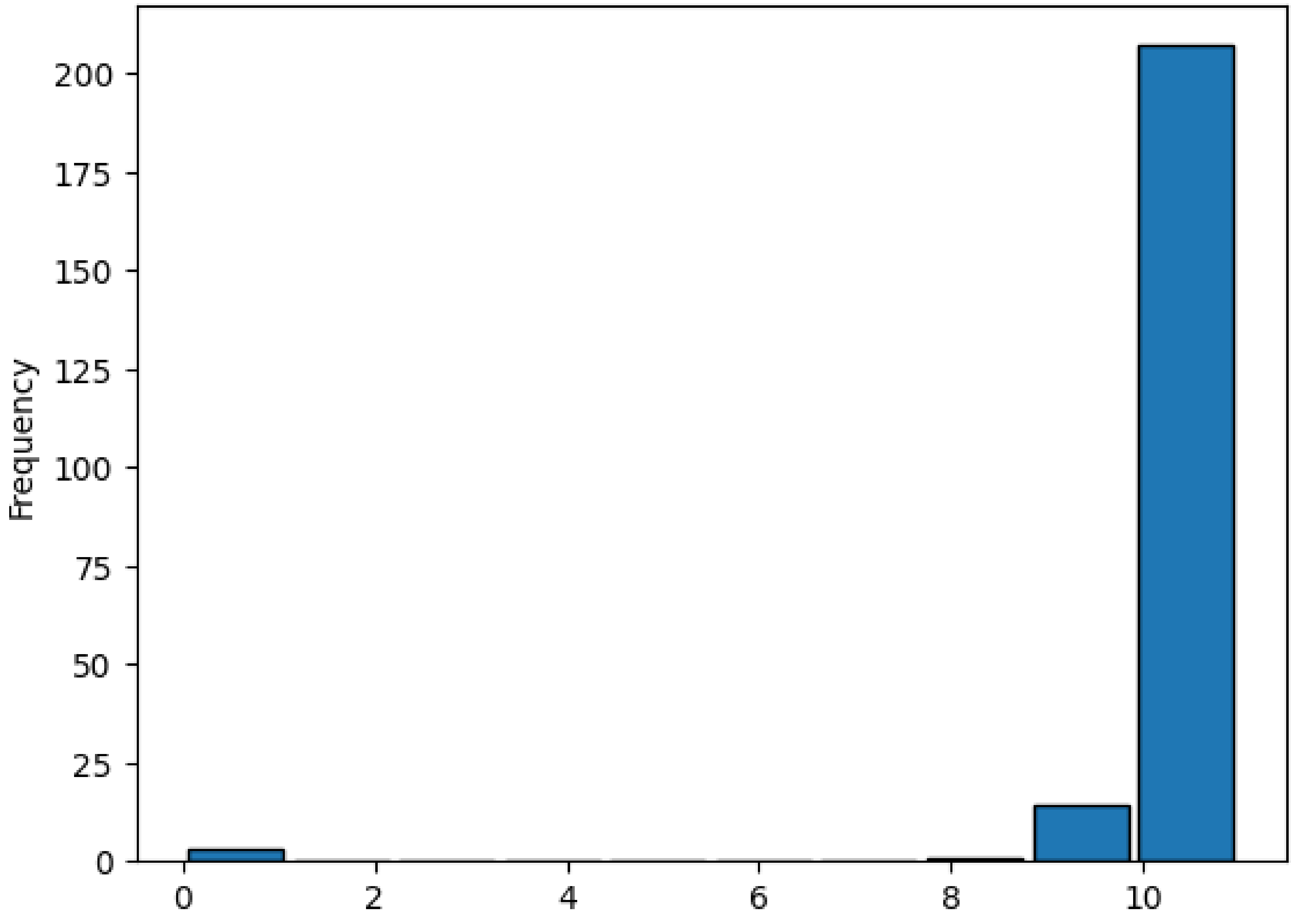
- Make sure to **submit your first work early**
  - We have been able to detect **submission issues** before the deadline
  - In case of **last-minute hitch**, there is still one submission we can grade, even if it is incomplete

- Make sure to **check your submission**

Assignment questions:

- **Read carefully the questions** and instructions
- Perform some sanity checks on your answers

Programming + SCQ questions



Maximum possible:
11 points

Mean: 10.6 points
Median: 11 points
(std: 1.3 points)

Grading of open reflection questions in progress

Questions which created more difficulty

Exercise 1

- 1.4.1 Compare the models

Exercise 2

- 2.3.1 Sex disparities: FNR
- 2.3.2 Sex disparities: DIR of the model
- 2.4.2 Young female applicants: DIR of the human decisions
- 2.5.3 Older applicants: identifying and mitigating bias

Feedback on Graded 1

Exercise 1:
Safety of an applicant selection model

1.2.1 Statistics on the data

Compute the number of people in the dataset and the number of hired applicants both by the model and by the HR experts.

Typical oversights:

- Use “hired/not hired” instead of “actual/predicted” to compute the number of *hired* applicants both by the model and by the HR experts

1.2.2 Accuracy

Compute the accuracy of the model. The accuracy value should be between 0 and 1 (not a percentage).

Typical oversights:

- Convert to percentages

1.2.3 Accuracy and safety

URL: ttpoll.eu
Session ID: cs290

Comment the result you obtain for the accuracy of this model: cite the value you obtain, explain what it means with words and conclude on the safety of the model (1 to 2 sentences).

Is an accuracy of 0,895 “good”?

1.3.1 Confusion matrix

We consider the value Hired to be the positive outcome of the model. Complete the function to compute the number of true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn) from the model predictions. To earn the point for this question, all four values must be correct, no partial credit is given.

Typical oversights:

- Inversion of FN and FP values

1.3.2 False positives

URL: ttpoll.eu
Session ID: cs290

What is the meaning of a false positive in this context?

0% a. The applicant is not hired by HR experts and not recommended by the model

88% b. The applicant is not hired by HR experts but recommended by the model

0% c. The applicant is hired by HR experts and recommended by the model

13% d. The applicant is hired by HR experts but not recommended by the model

1.3.3 Consequences of false positives

Describe one possible consequence of a false positive in this context (1 sentence)?

- ≠ stakeholders
- ≠ use scenarios

👉 1 post / consequence

We will NOT use the vote system this time

Post your ideas:
https://speakup.epfl.ch
Room key: **38133**

1.3.2 False negatives

URL: ttpoll.eu
Session ID: cs290

What is the meaning of a false negative in this context?

4% a. The applicant is not hired by HR experts and not recommended by the model

0% b. The applicant is not hired by HR experts but recommended by the model

4% c. The applicant is hired by HR experts and recommended by the model

92% d. The applicant is hired by HR experts but not recommended by the model

1.3.4 Consequences of false negatives

Describe one possible consequence of a false negative in this context (1 sentence)?

👉 1 post / consequence

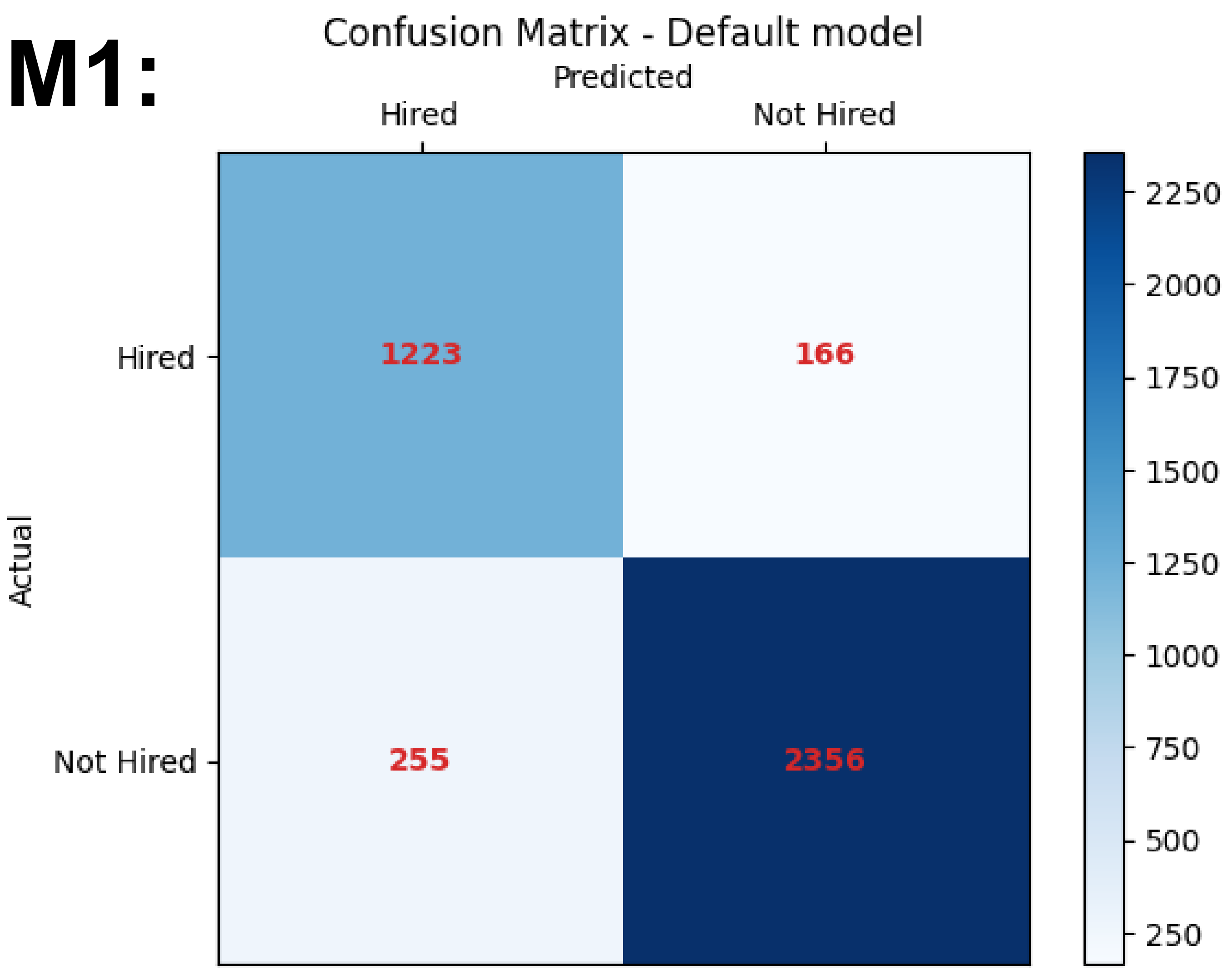
Post your ideas:
https://speakup.epfl.ch
Room key: **91321**

1.4.1 Model comparison

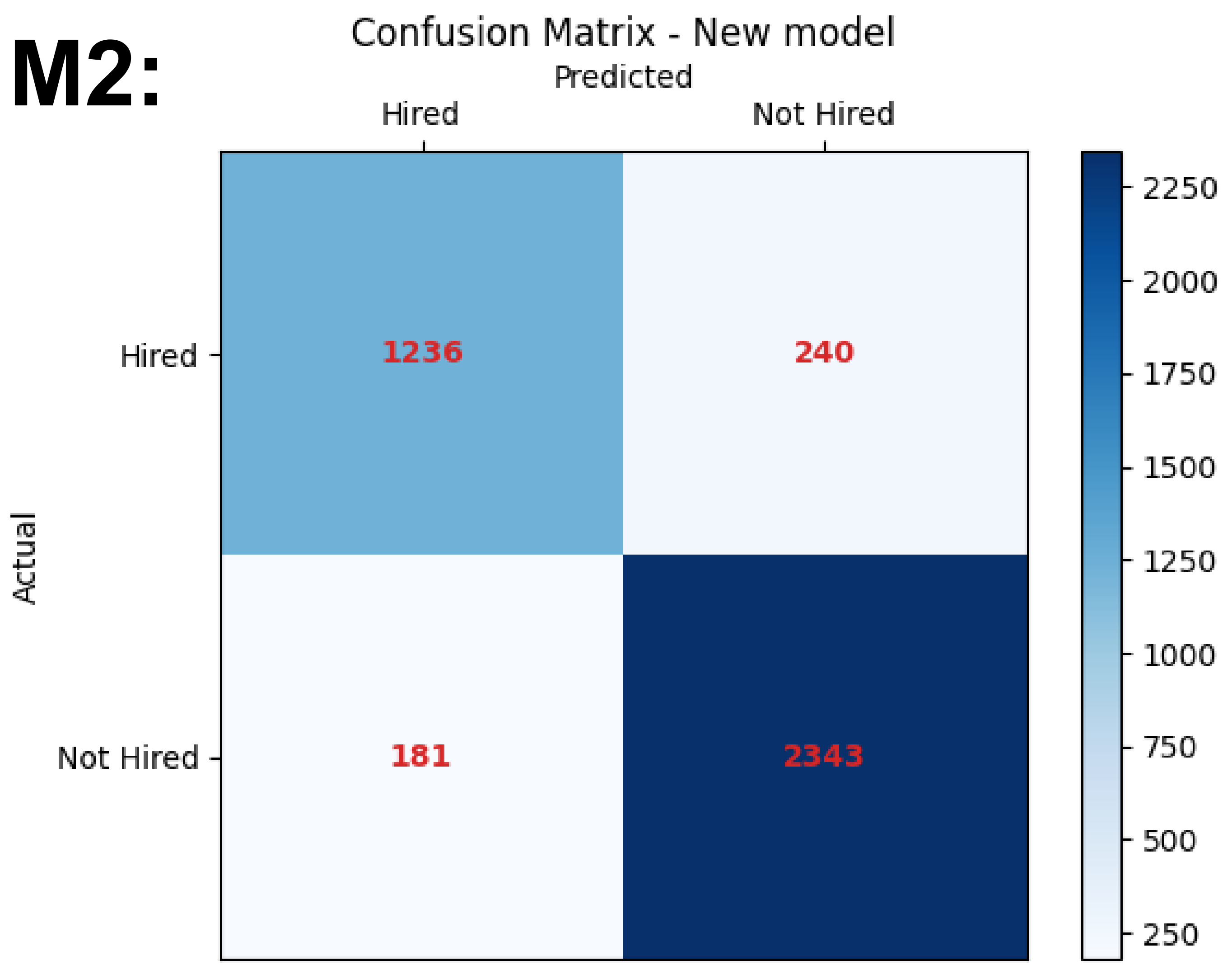
BrightPath Consulting is currently recruiting a large number of consultants to create a new department, but it operates in a tight labor market where qualified applicants are scarce. The priority is to hire as many good applicants as possible.

Given this context, which of the two models would you advise to use: the new model M2 or the previous model M1? Justify your choice (3 sentences).

M1:



M2:



Feedback on Graded 1

<u>Exercise 2:</u>
Fairness in hiring

2.2.1 Statistics on the data

Compute the mean of the Age and Experience columns for male and female applicants separately.

Typical oversights:

- Forgotten columns in the resulting dataframe
- Typos

2.3.1 Sex disparities: FNR

Compute the false negative rate (FNR) for male and female applicants. You can reuse the function you implemented in the previous exercise (see question 1.3.1).

Typical **errors**:

- Wrong definition of FNR instead of FN / (FN + TP)
  - FN / df_male.shape[0]
  - FN / (TP+FP)
  - FN / (TP+FP+TN+FN)

Review:
- Video 4.2 “Group Fairness”
- Notebook Fairness 2

2.3.2 Sex disparities: DIR

Bias from **the model**

Bias from **humans**

- What is the goal of:
  - Analyzing the DIR on the **Predicted** column?
  - Analyzing the DIR on the **Actual** column?

	A	B	C	D	E	F
1	Sex	Age	Education	Experience	Predicted	Actual
2	Female	36	MSc	8	Not Hired	Not Hired
3	Male	22	PhD	4	Hired	Hired
4	Female	31	MSc	8	Not Hired	Not Hired
5	Female	37	HS	5	Not Hired	Not Hired
6	Male	38	BSc	7	Hired	Hired
7	Male	57	MSc	14	Not Hired	Not Hired
8	Male	36	BSc	8	Not Hired	Not Hired
9	Female	32	BSc	10	Not Hired	Not Hired
10	Male	22	BSc	4	Hired	Hired
11	Male	40	MSc	12	Hired	Hired
12	Female	32	MSc	5	Not Hired	Not Hired
13	Male	31	BSc	11	Hired	Hired
14	Female	44	BSc	8	Not Hired	Not Hired
15	Female	28	BSc	6	Hired	Hired

2.3.2 Sex disparities: DIR of the model

Compute the disparate impact ratio between male and female applicants. Use the column Predicted to base your analysis on. This will ensure we focus on the model and not the human decisions. Consider the group with the highest selection rate as the privileged group.

Typical oversights:

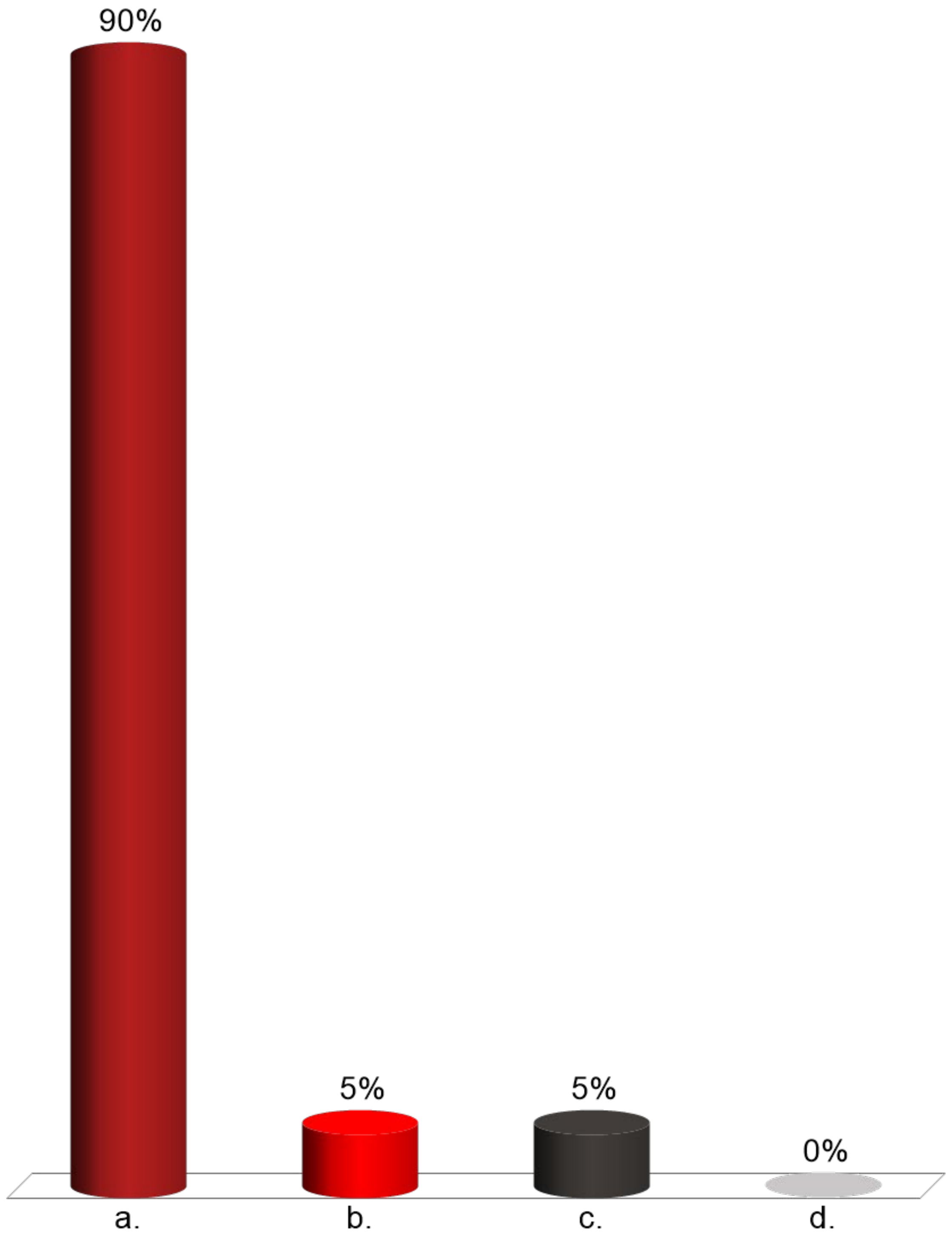
- Computation over the Actual column
- Wrong privileged group
- Typos (names of dataframes, name of variable)

2.3.3 FNR and DIR

URL: ttpoll.eu
Session ID: cs290

Which statement best distinguishes the False Negative Rate (FNR) from the Disparate Impact Ratio (DIR)?

a. The FNR measures the proportion of actually hired applicants incorrectly rejected by the model, while the DIR compares the recommendation rates between different demographic groups.
b. The FNR compares error rates between groups, while the DIR evaluates the difference in accuracy between different demographic groups.
c. The FNR measures the proportion of applicants recommended by the model who are actually not hired, while the DIR evaluates the difference in accuracy between different demographic groups.
d. The FNR evaluates the difference in accuracy between different demographic groups, while the DIR compares error rates between groups.



2.3.4 Sex disparities

URL: ttpoll.eu
Session ID: cs290

Based on the two metrics you have computed above, identify whether the model is discriminating by sex or not: for each metric, cite the value you obtain and how you interpret it then conclude (2 to 3 sentences).

False negative rate (Male): 0.120
False negative rate (Female): 0.119
Disparate impact ratio: 0.918

Is the model discriminating by sex or not?

14% a. Yes
82% b. No
5% c. Other

2.4.1 Young female/male applicants

Isolate the subset of the dataframe containing male and female between 26 years old exclusive and 34 years old exclusive. You need to have the following columns at the end: Sex, Age, Education, Experience, Predicted, Actual.

Very few mistakes

2.4.2 Young f/m: DIR of human decisions

Calculate the disparate impact ratio (DIR) between young male and young female applicants based on the Actual column. This will ensure we focus on the human decisions and not the model output. Consider the group with the highest selection rate as the privileged group. You get 0.25 point for each correct proportion and 0.5 for the correct DIR.

Typical oversights:

- Computation over the Predicted column
- Typos (names of dataframes, name of variable)

2.4.3 Fairness in HR experts

URL: ttpoll.eu
Session ID: cs290

Based on the metric you have computed above, identify whether the human decisions from HR experts are fair for the young female applicants compared to the young male applicants of the same age group: cite the value you obtain for the metric and explain how you interpret it (1 sentence).

DIR of the HR decisions (female / male): 0.622

Are human decisions fair to young female applicants?

0% a. Yes

100% b. No

0% c. Other

2.5.1 Age disparities: FNR

URL: ttpoll.eu
Session ID: cs290

Based on the false negative rates provided above, would you say that the model is fair with respect to age? If no, which age category is put at a disadvantage?

FNR for older applicants is 0.048
FNR for younger applicants is 0.121

Is the model fair with respect to age?

0% a. Yes
95% b. No
5% c. Other

2.5.1 Age disparities: FNR

Based on the false negative rates provided above, would you say that the model is fair with respect to age? If no, which age category is put at a disadvantage?

FNR for older applicants is 0.048
FNR for younger applicants is 0.121

Who is discriminated?

2.5.2 Age disparities: DIR

URL: ttpoll.eu
Session ID: cs290

Based on the disparate impact ratio provided above, would you say that the model is fair with respect to age? If no, which age category is put at a disadvantage?

DIR of the model (older / younger): 0.733

Is the model fair with respect to age?

0% a. Yes
100% b. No
0% c. Other

2.5.2 Age disparities: DIR

Based on the disparate impact ratio provided above, would you say that the model is fair with respect to age? If no, which age category is put at a disadvantage?

DIR of the model (older / younger): 0.733

Who is discriminated?

2.5.3 Identifying & mitigating bias

What could explain the results you obtained on the FNR and the DIR?

- Use the code cell below to make the exploration or calculations you want for this question, the code is not graded.
- Provide your answer in the raw cell below: identify the issue (1 sentence) and justify your answer by citing two relevant metrics of your choice (2 sentences, including numerical values).
  Only the text answer you provide below is graded.

- Very low rate of completely correct answers…

2.5.3 Identifying & mitigating bias

Regarding the Fairness 2 notebook, on the COMPAS case:

16% a. I have completed all sections, including the optional ones
64% b. I have completed all sections except the optional ones
4% c. I have partially completed the notebook
16% d. I have not started the notebook

URL: ttpoll.eu
Session ID: cs290

ProPublica vs. NorthPointe

Why do they obtain **opposite results?**

■ ProPublica: FPR and FNR

	Caucasian	African American
False Positive Rate (FPR) (wrongly predicted recidivist among *not* recidivists)	21%	**41%**
False Negative Rate (FNR) (wrongly predicted *not* recidivist among recidivists)	**50%**	29%

→ Discrimination against African American group

■ NorthPoint: 1-PPV and 1-NPV

	Caucasian	African American
1 - Positive Predictive Value (1-PPV=FDR) (*not* recidivists among the predicted recidivists)	**38%**	31%
1 - Negative Predictive Value (1-NPV=FOR) (recidivists among the predicted *not* recidivists)	31%	**39%**

→ Discrimination against Caucasian group

ProPublica vs. NorthPointe

- Statistics from the data:

	Caucasian	**African American**
Number of defendants	2103	**3175**

	Caucasian	**African American**
Actual recidivists (P)	42%	**56%**
Actual not recidivists (N)	**58%**	44%

Post your ideas:
https://speakup.epfl.ch
Room key: **68969**

- **What types of biases are present?**

👉 1 post / bias

The impossibility results

(Chouldechova, 2017; Barocas et al., 2023; Mitchell et al., 2021)

Mathematical **impossibilities**:

- Among error parity measures
- With demographic parity

In a world that is unfair and with classifiers that cannot be perfect, **not all group fairness criteria can hold at the same time**

👉 We have to **choose which view of fairness is most important** in a specific context

"The impossibility results aren't some kind of artifact of statistical decision making; they simply **reveal moral dilemmas**."

(Barocas et al. 2023)

2.5.4 Solutions to bias

Propose one (technical) solution that could be used to mitigate the bias in the model (1 sentence). Complete the cell below.

👉 1 post / solution

Post your ideas:
https://speakup.epfl.ch
Room key: **15585**

Other info on Graded 1

Group work & GenAI

- We have noticed a number of **specific errors** that seem to be **shared over several notebooks** (questions 2.3.1, 2.3.2 and 2.4.2)
  - Wrong definitions for FNR
  - Specific types of typos in DIR computations

- We have noticed specific python code and verbose answers which are **typical of the use of Generative AI Tools**

👉 We will check if there are repeated issues in Graded Notebook 2 for the persons involved

Graded 1: grade release

- Release of **score** (over 26) in the coming days
- Corresponding **grades** will be released afterwards
  - Computation:
    - If points <= 13, grade = 1 + 3/13 * points
    - If points > 13, grade = 2 + 2/13 * points
  - Then grade x 0,08 to get the proportion of final grade
    - A grade of 6 gives 0,48 in the final grade

- Have questions on your score?

  👉 Post on Ed:
  - **Private** message
  - In Graded Assignments > Graded notebook 1

What’s next?

Mock test on Tuesday 28, 10h15-12h

- A PDF file will be available for download on moodle
- Assistants will provide support in rooms INF 1 and CO 5
  - We recommend that you **do the test in exam-like conditions** and that you **time yourself** so that you know how you are performing in terms of speed.
  - It is a good opportunity to test your A4 sheet of notes (reminder: only 1 sheet of notes, A4 double-sided, can be handwritten or printed).

<u>**By Thursday 30 October**</u>, we ask you to fill out a **survey on the mock test** to get your feedback and prepare the debriefing that we will do on Monday 3 November.

It’s a good idea to do the Mock 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 **time**

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

Mock test

URL: ttpoll.eu
Session ID: cs290

Do you plan to come in INF1 and CO5 tomorrow 10h15-12h?
(e.g. to do the test or to ask questions)

Next dates

	Monday	Tuesday
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