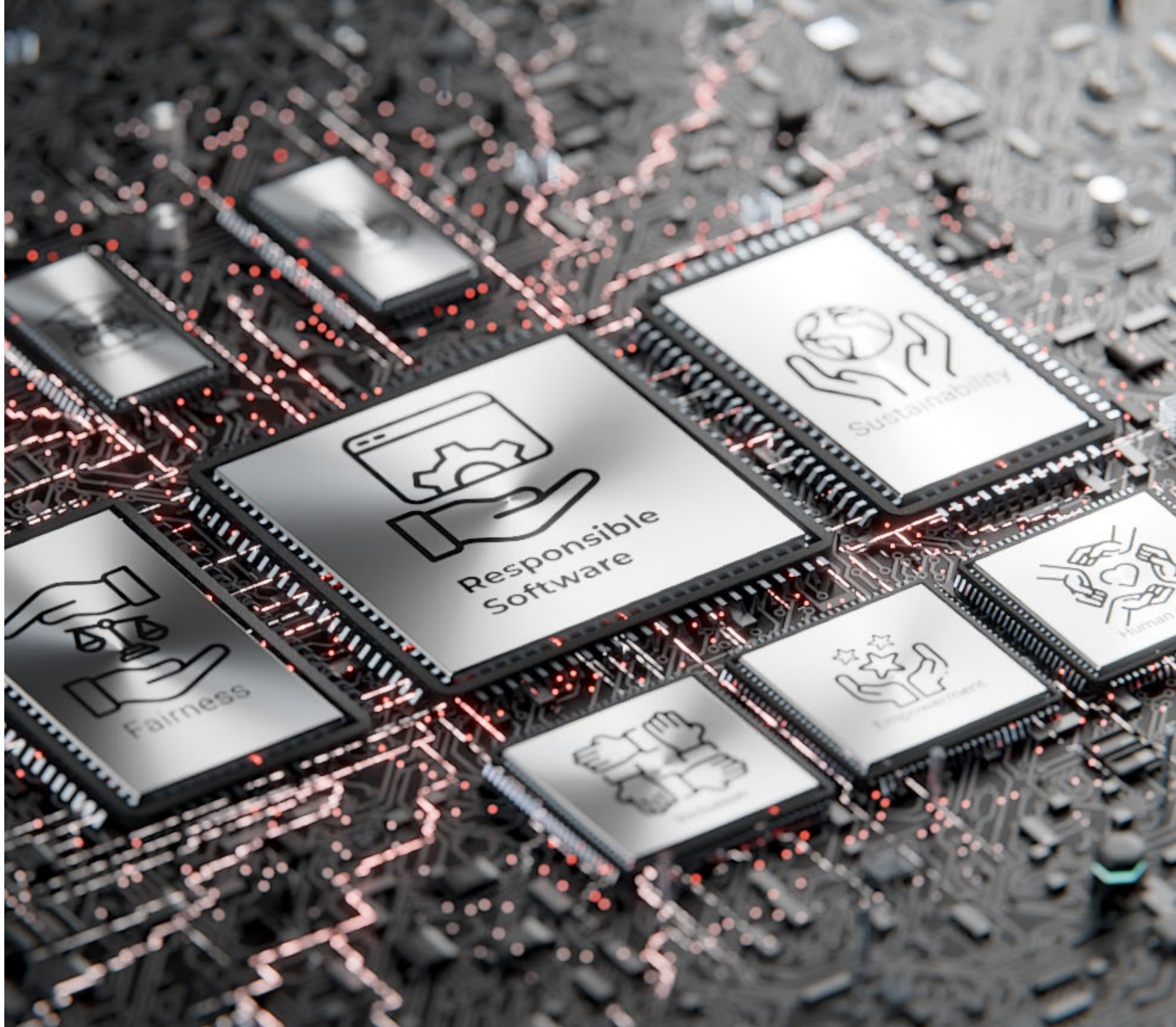
Empowerment 2 Review & Case studies
9 dec.

Cécile Hardebolle

Responsible Software



Agenda for today

1. General information and logistics
   - In-depth feedback on the course
   - Overview of results on Graded Assignment 2
   - **Logistics for the final written exam**

2. Interactive review questions on Empowerment 2

3. Case studies:
   a) Bad actors
   b) Ethical speculation (“Escape the Mirror”)
   c) Datasheets for datasets

General information & Logistics

In-depth feedback

In-depth feedback on the course

A **big thank you** already for:

- Your indicative feedback in week 5 of the semester
- The feedback you have submitted for each chapter on courseware

Now comes the time to give your **overall feedback on the course**!

- Online form on moodle
- Available from today (December 9) until January 12
- Space for comments!
  - Most interesting / least interesting
  - Most clear / least clear
  - Suggestions for improvement

General information & Logistics

Final exam

Final exam - Logistics

URL: ttpoll.eu
Session ID: cs290

Select all the **correct** statements about the **final exam:**

1% a. It is in the winter exam session

26% b. It is on the last week of term

0% c. It includes programming

26% d. It includes case studies

22% e. It includes SCQ/TF questions on the videos

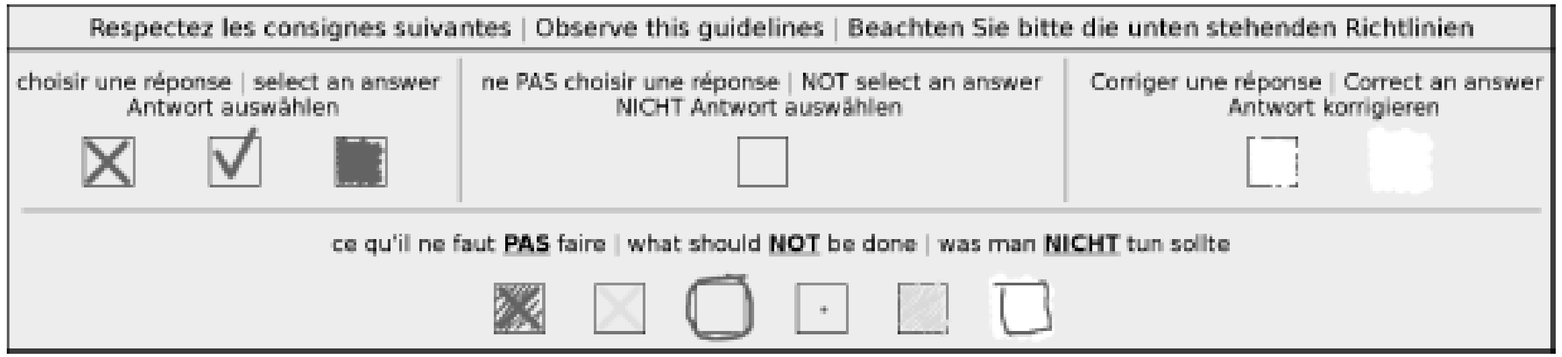
0% f. All documents are allowed

25% g. Only one A4 paper of notes is allowed

Final exam – Rules (reminder)

- The exam is on **paper** and includes:
  single choice question, true/false questions and case studies
- No electronic devices allowed
- No documents allowed except **one (1) sheet of paper**:
  size A4, recto-verso, free format (printed/handwritten, no restriction)

- Use a **black** or **dark blue ballpen**
- Follow instructions for selecting and erasing properly:
  - Marked = selected
  - Blanked = not selected



Final exam – Logistics

- You are **assigned a seat**, communicated on moodle
  If you see an issue with assigned seats, please contact me!
- Make sure to display your **camipro card** on your table

- The exam **starts at <u>8h15</u>** and you have **<u>1h30</u> to work**,
  except special arrangements
  - The exam copy is on your table – you MUST wait until 8h15 to open it
  - When indicated (normally 9h45), you MUST put your pen down and wait while we collect copies

- First 30 minutes: late arrival possible, no early departure
- Last 15 minutes: no early departure

Final exam – Formatting answers

Considering the following extract of the harms modeling table, describe what should go in the different cells:

- For cells A, B, C and E: describe 1 harm that corresponds to the category
- For cell D: indicate the corresponding harm category

Make sure to identify your answers with the corresponding letters [1 point / answer].

Category	**Type of harm**	**Social assistant chatbot**
Humans	Physical injury	**A)**
Allocation of Resources	Opportunity loss	**B)**
Human Rights	Liberty loss	**C)**
	D)	Most intimate feelings are now "public"
Social System Harms	Social	

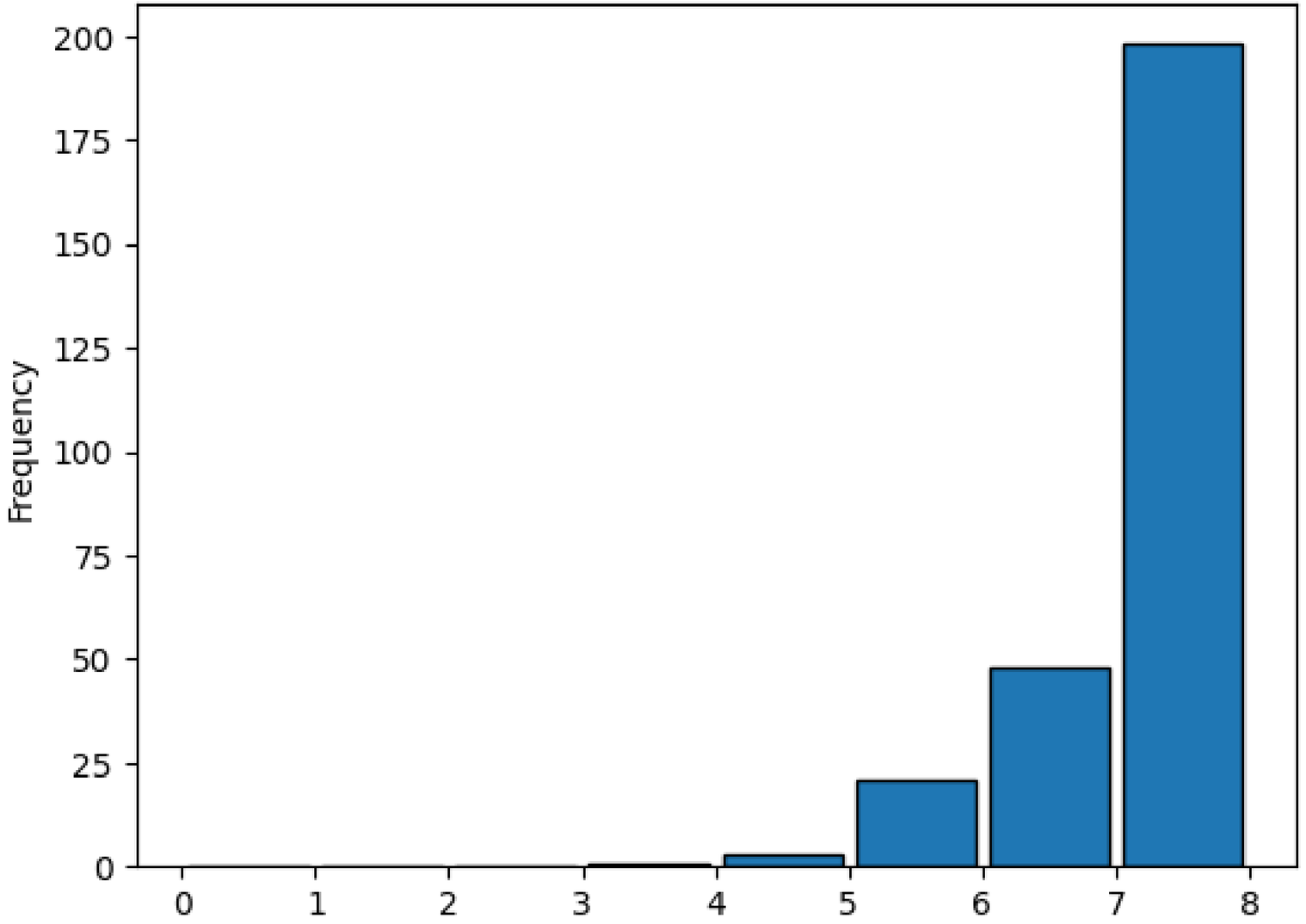
(A) The chatbot could recommend aggressive behaviors or inappropriate conduct (e.g. a teacher asking for help to deal with an unruly pupil: the chatbot could advise inappropriate gestures instead of adequate mediation options, for example)

(B) Making a decision based on the chatbot's advice could lead people to miss some opportunities for social connection (e.g. random encounters). Biases in the chatbot could lead to some people being excluded from social relationships or from services or o˙ers.

General information & Logistics

Graded 2

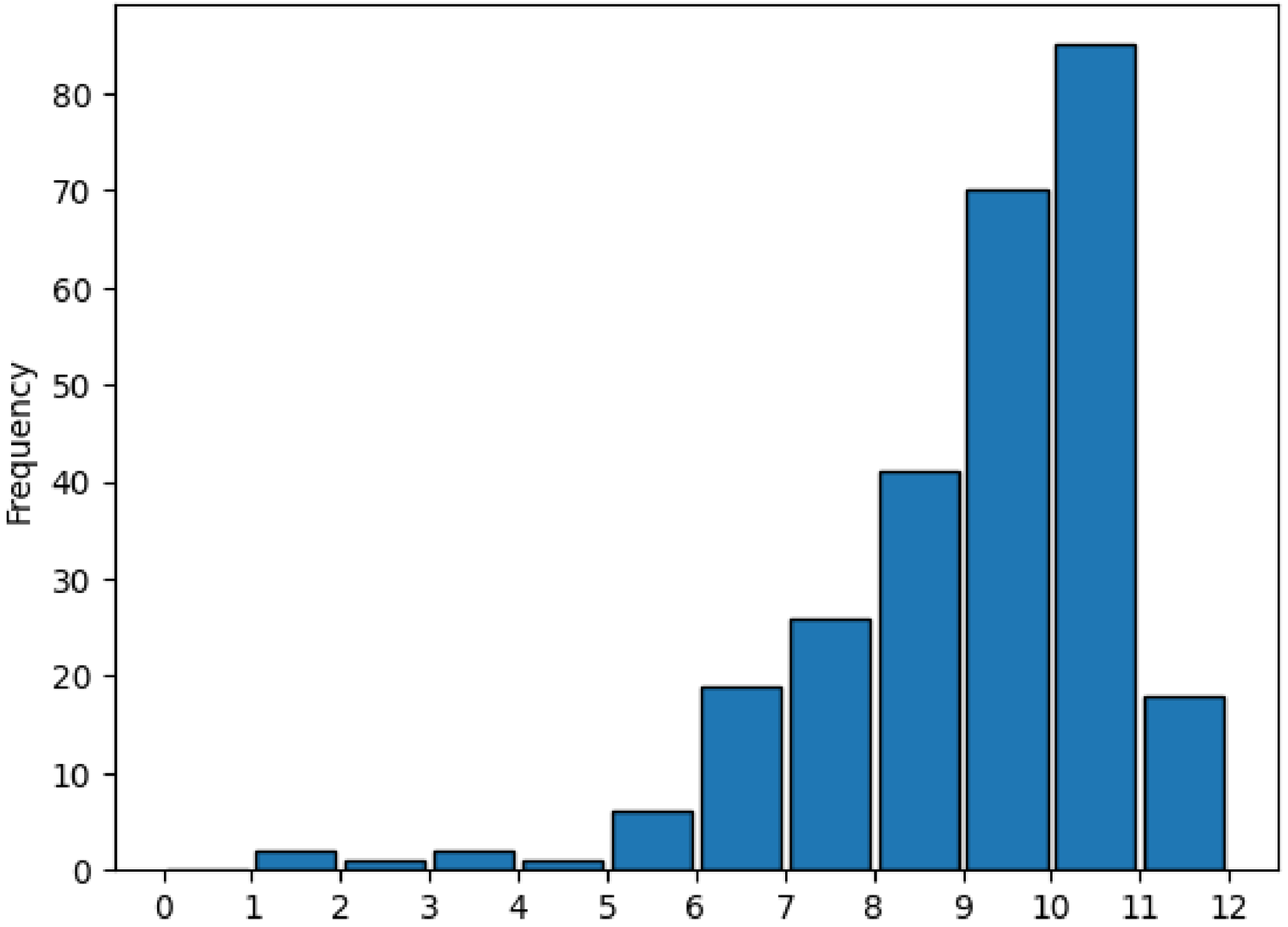
Graded 2 - Programming questions



Maximum possible:
8 points

Mean: 7.2 points
Median: 7.4 points
(std: 0.8 points)

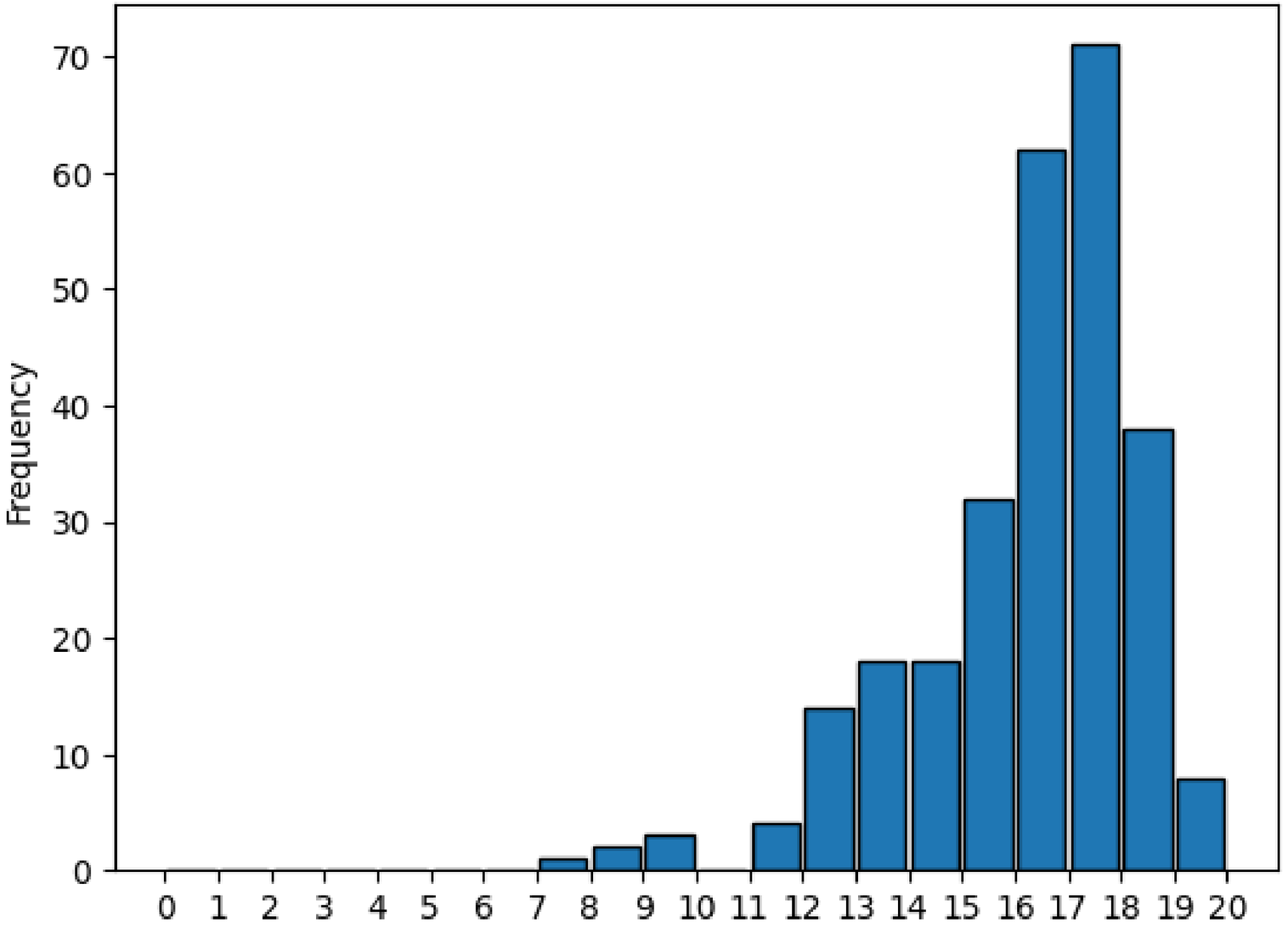
Graded 2 - Open reflection questions



Maximum possible:
12 points

Mean: 9 points
Median: 9.5 points
(std: 1.7 points)

Graded 2 - Overall distribution of points



Maximum possible:
20 points

Mean: 16.2 points
Median: 16.7 points
(std: 2.1 points)

Your questions

We have been **overwhelmed** with questions!

Before posting your question, you must:

- For code questions: check the notebook solution
- For open questions: check the slides Debriefing Graded 2

We are doing our best to answer in a reasonable delay.
Please bear with us while we process the messages.
Thank you for your understanding.

Review questions Empowerment 2

Privacy policies

URL: ttpoll.eu
Session ID: cs290

Several studies have shown that the privacy policies of many online platforms and websites are extremely long (several thousand of words, taking in the 20 minutes to read on average), use legalistic terminology and are hard to navigate.

This can be said to be a transparency issue because (select all that apply):

All of these can be argued:
- Hard to navigate = accessibility issue
- Legalistic vocab = understandability issue
- Extremely long = relevance issue

- ✅ 7% a. Information is not accessible
- ✅ 82% b. Information is not understandable
- ✅ 11% c. Information is not relevant

(Sherman, 2024; Litman-Navarro, 2019)

Transparency and datasets - 1

URL: ttpoll.eu
Session ID: cs290

One of the results of your Bachelor thesis is a very cool dataset which contains tasting profiles and consumer reviews for 3197 unique beers from 934 different breweries. This dataset can be used to train machine learning models for sentiment analysis and classification tasks.
You want to make the dataset public.
For ensuring transparency you should also publish with it:
(select all that apply):

All of these (descriptive stats is probably the least important because it can be obtained from the data)

- ☑ 16% a. Descriptive statistics
- ☑ 24% b. Composition of the data, including demographics of people
- ☑ 23% c. Description of the collection process
- ☑ 20% d. Description of the pre-processing performed
- ☑ 17% e. Description of the purposes and intended use

Transparency and datasets - 2

URL: ttpoll.eu
Session ID: cs290

One of the results of your Bachelor thesis is a very cool dataset which contains tasting profiles and consumer reviews for 3197 unique beers from 934 different breweries. This dataset can be used to train machine learning models for sentiment analysis and classification tasks.
You want to make the dataset public.

The document you would need to attach to the dataset for best transparency is called:

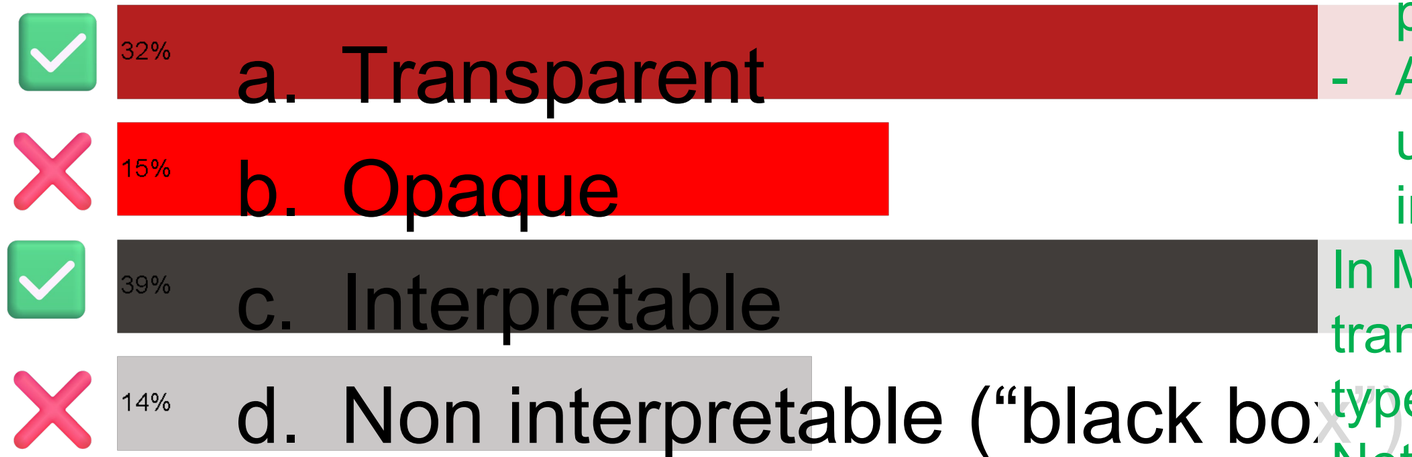
- ❌ 5% a. A dataframe
- ✅ 84% b. A datasheet
- ❌ 8% c. A database
- ❌ 3% d. A statement of reasons

Transparency and ML - 1

URL: ttpoll.eu
Session ID: cs290

In the Fairness 2 notebook you have created a Logistic Regression model on the ProPublica dataset to try to reproduce how the COMPAS software predicts the risk of recidivism.

The logistic regression model you have created can be said to be (select 2 options):



- Stakeholder considered = you = developer
- Access to code, training data and parameters -> transparency
- Ability to make sense of parameters and understand how the model works -> interpretable

In ML, the “understandable” criteria of transparency cannot be obtained for some types of models (mainly Deep Neural Networks) + this criteria is specifically called “interpretability”

Transparency and ML - 2

URL: ttpoll.eu
Session ID: cs290

To have transparency on the ML model behind the COMPAS software would mean to have access to:

24% a. The design documentation
8% b. The user documentation
0% c. The code
11% d. The training dataset
5% e. A post-hoc interpretability method
✅ 51% f. It depends

It depends on the stakeholder considered (Transparency = “the degree to which stakeholders can answer their questions by using the information they obtain about a software system during its life cycle”).

-> All these options could be used potentially.

Case studies

Where to find the cases?

1. Go to **moodle**

2. Find the **link to the case studies** for today: **Empowerment 2**

3. Download the **instruction sheet**

+ From previous chapters, you will need:

- Bad actors (1 – Safety 1)
- Ethical speculation “Escape the Mirror” (0 – Introduction)

Bad Actors
(review from Safety 1)

Instructions

- Read the context description

- Review the **5 categories** in the Bad Actors strategy:
  Money, Politics, Entertainment, Self-Interest, Ideas
  - Which **harmful actions** could be taken?
  - What would be the **potential impacts** on stakeholders?

(Dis-)Empowerment: Bad Actors

Which harmful actions could be taken?
What would be the potential impacts on stakeholders?

👉 1 post = 1 harmful action & its negative impact
+ corresponding strategy category

See answers on SpeakUp (check the comments on the posts to see feedback)

Post your ideas:
https://speakup.epfl.ch
Room key: **98075**

Ethical Speculation
(review from Intro)

Instructions

Imagine an episode of “Escape the Mirror” (our version of “Black Mirror”) where **the main character is disempowered because of software** (e.g., deceived, manipulated, left without recourse…).

Inspiration = list of topics or news articles

- Write down **a short pitch** which focuses on 1 main character
- Identify the **ethical issue(s)** flagged by your story, such as:
  - Opaque biased algorithm
  - Emotional manipulation
  - Political deception
  - Creating dependency
  - …

Related to (dis-)empowerment

(Dis-)Empowerment: Ethical Speculation

👉 1 post = 1 short pitch
+ corresponding ethical issue(s)

See answers on SpeakUp (check the comments on the posts to see feedback)

Post your ideas:
https://speakup.epfl.ch
Room key: **72959**

Datasheets
(review from Fairness 2)

Instructions

Context:

- ML task = identifying people from an image
- Dataset = MS-Celeb-1M

Instructions

- Read the **summary** we provide from the original research article
- Fill out the **datasheet** (some parts are already filled out)
- Highlight **2 ethical problems** with this dataset

(Dis-)Empowerment: Datasets

👉 1 post = 1 ethical issue with the dataset

See answers on SpeakUp (check the comments on the posts to see feedback)

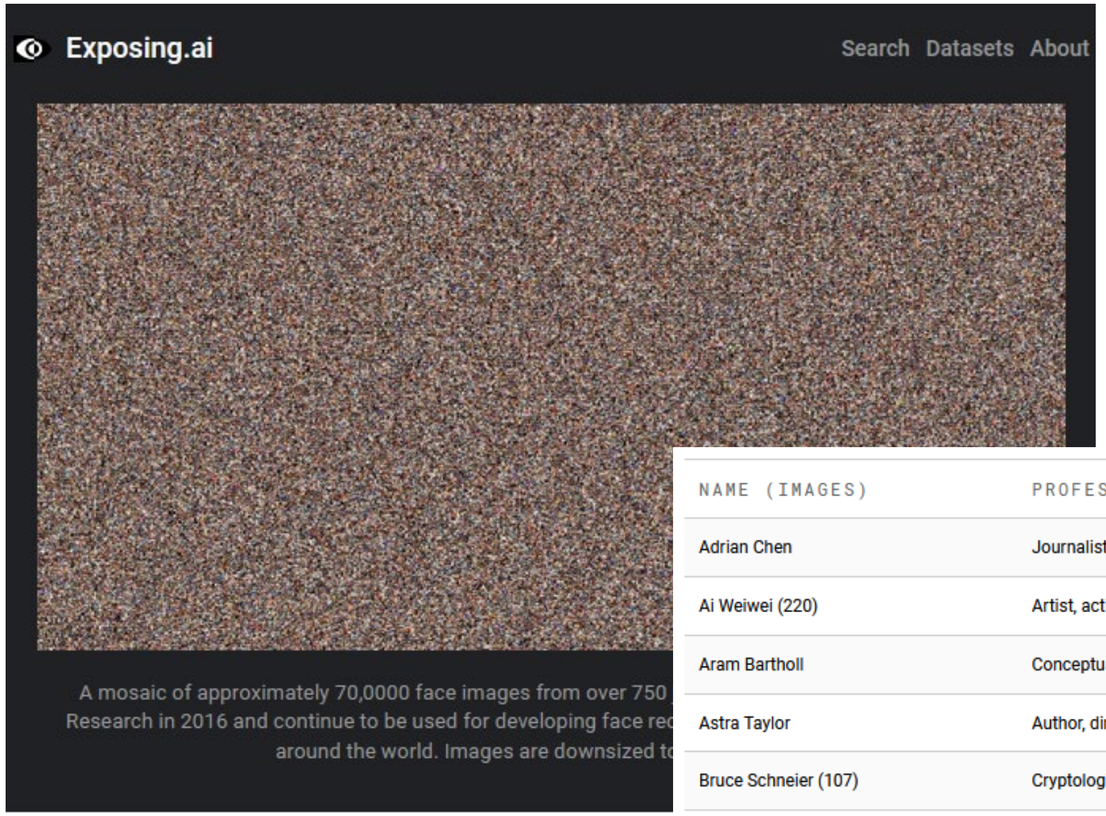
Post your ideas:
https://speakup.epfl.ch
Room key: **64393**

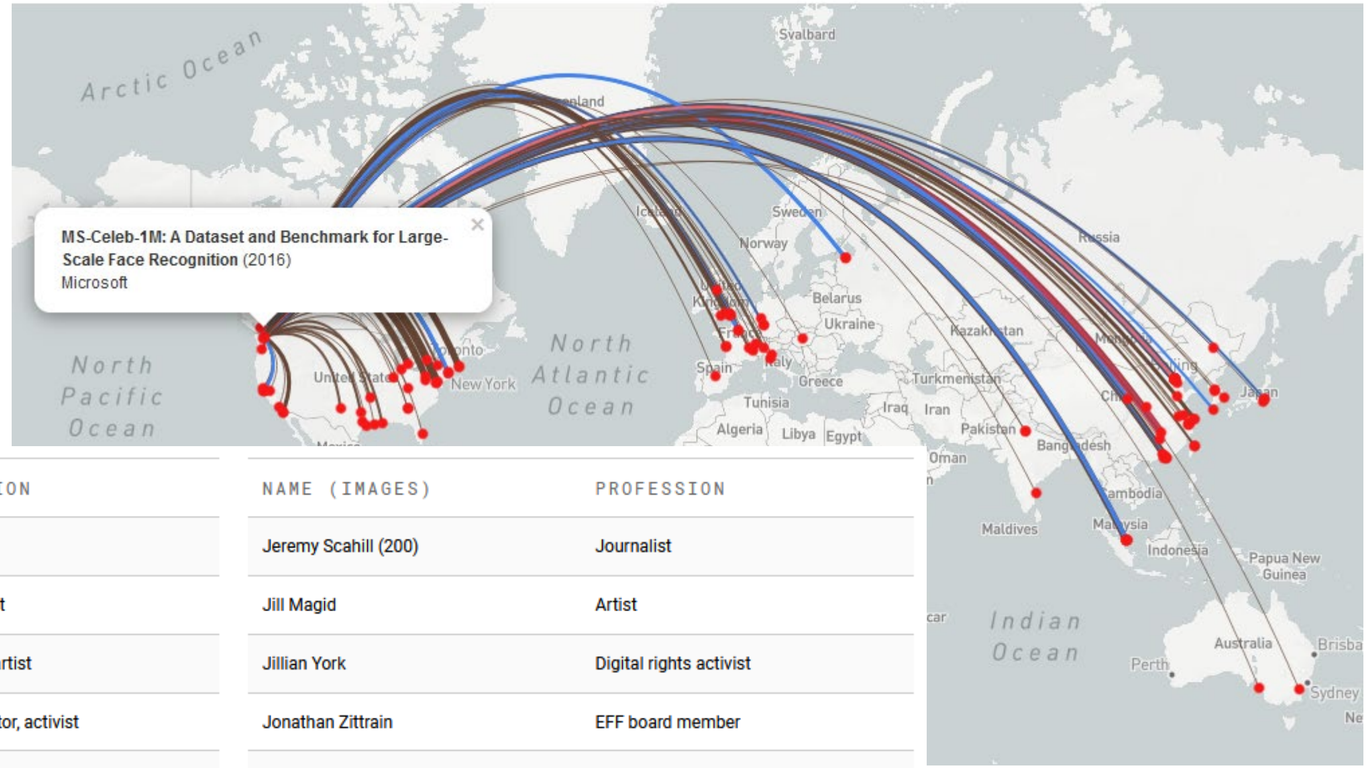
Datasheet

- Representativeness
- Confidentiality
- Problematic content (NSFW)
- Identification of people
- Sensitive data
- Acquisition of data, consent

“Exposing.ai”: MS-Celeb-1M

(Harvey & Laplace, 2021)





MS-Celeb-1M (M

NAME (IMAGES)	PROFESSION
Adrian Chen	Journalist
Ai Weiwei (220)	Artist, activist
Aram Bartholl	Conceptual artist
Astra Taylor	Author, director, activist
Bruce Schneier (107)	Cryptologist
Cory Doctorow (104)	Blogger, journalist
danah boyd	Data & Society founder
Edward Felten	Former FTC Chief Technologist
Evgeny Morozov (108)	Tech writer, researcher
Glenn Greenwald (86)	Journalist, author
Hito Steyerl	Artist, writer
James Risen	Journalist

NAME (IMAGES)	PROFESSION
Jeremy Scahill (200)	Journalist
Jill Magid	Artist
Jillian York	Digital rights activist
Jonathan Zittrain	EFF board member
Julie Brill	Former FTC Commissioner
Kim Zetter	Journalist, author
Laura Poitras (104)	Filmmaker
Luke DuBois	Artist
Michael Anti	Political blogger
Manal al-Sharif (101)	Women's rights activist
Shoshana Zuboff	Author, academic
Trevor Paglen	Artist, researcher

What’s next?

Tomorrow at 8h15 in SG1

- **Review cases**: use general strategies **on real software**!
  - Digital Ethics Canvas – Emotion Cancelling AI
  - Ethics Canvas – Be My Eyes

- **Your questions!**

References

- Litman-Navarro, K. (2019, June 12). We Read 150 Privacy Policies. They Were an Incomprehensible Disaster. The New York Times.https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html
- Sherman, J. (2024, November 25). Meta’s Privacy Policies: Designed Badly, by Design? | TechPolicy.Press. Tech Policy Press. https://techpolicy.press/metas-privacy-policies-designed-badly-by-design
- Harvey, A., & Laplace, J. (2021). Exposing.ai. Exposing.Ai. https://exposing.ai/