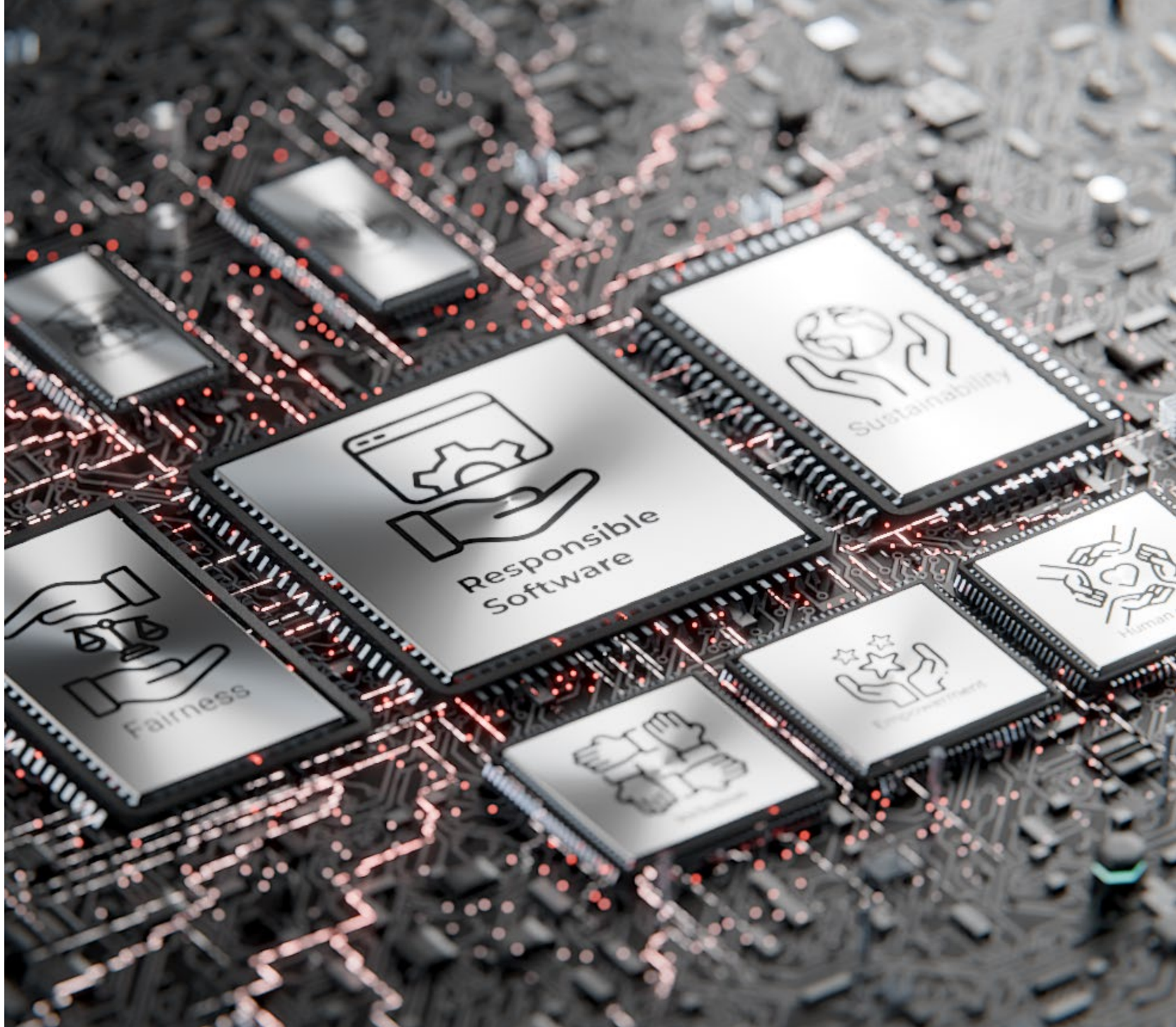
Empowerment 2 Review & Case studies
8 dec.

Cécile Hardebolle

Responsible Software



Agenda for today

1. In-depth feedback on the course

2. Interactive review questions on Empowerment 2

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

In-depth feedback on the course

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 dashboard** & **PocketCampus app**
- Available from today (December 8) until January 11
- Space for comments!
  - Most interesting / least interesting
  - Most clear / least clear
  - Suggestions for improvement

Review questions Empowerment 2

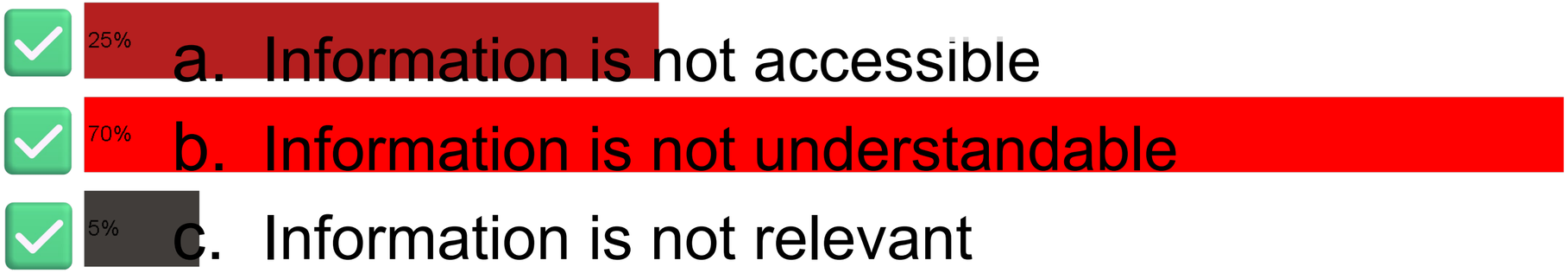
Privacy policies

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



(Sherman, 2024; Litman-Navarro, 2019)

Beer brewing dataset - 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 (composition of the data is probably the least important because it can be obtained from the data)

- ✅ 25% a. Composition of the data, including demographics
- ✅ 29% b. Description of the collection process
- ✅ 27% c. Description of the pre-processing performed
- ✅ 19% d. Description of the purposes and intended use

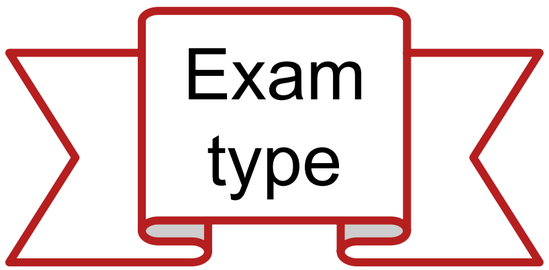
Beer brewing dataset - 2

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 have created a datasheet for your dataset.

Which of the FAIR principles do you follow by providing a datasheet?

- ❌ 8% a. Findable
- ❌ 0% b. Accessible
- ❌ 8% c. Interoperable
- ✅ 85% d. Reusable



Linear Regression Model

You have found on HuggingFace an open-source Linear Regression model that predicts the price of a house based on a range of features like lot area, construction year, number of rooms, etc. For recall, a linear regression model has the following mathematical form, where *y'* is the predicted price, x_i are the features and *b* and w_i are the final parameters of the model:

$$y' = b + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + \cdots$$

Let's imagine that you want to modify this model. How could you do it?

- ❌ 71% a. Modify the values of the parameters
- ❌ 0% b. Use a post-hoc interpretability method
- ✅ 29% c. Retrain the model with a new dataset
- ❌ 0% d. It is not possible to modify the model

- **a&c: SEE NEXT SLIDE for explanation**
- b: post-hoc interpretability methods have nothing to do with model modification (they only help interpret how the model works)
- d: the text says "open-source", so it is possible to modify the model

Note on modifying ML models

- The final parameters of a ML model represent the patterns in the data as “detected” (“learned”) by the learning algorithm
- Technically speaking, it is *possible* to modify/edit these parameters manually, however it is generally NEVER done because:
  - It is generally impossible to do it without “breaking” the model
  - Then the model does not reflect anymore the patterns learned from the data

👉 If you need to modify a ML model, you will generally retrain it with new data, modify the training procedure, etc. so that it “learns” a different pattern

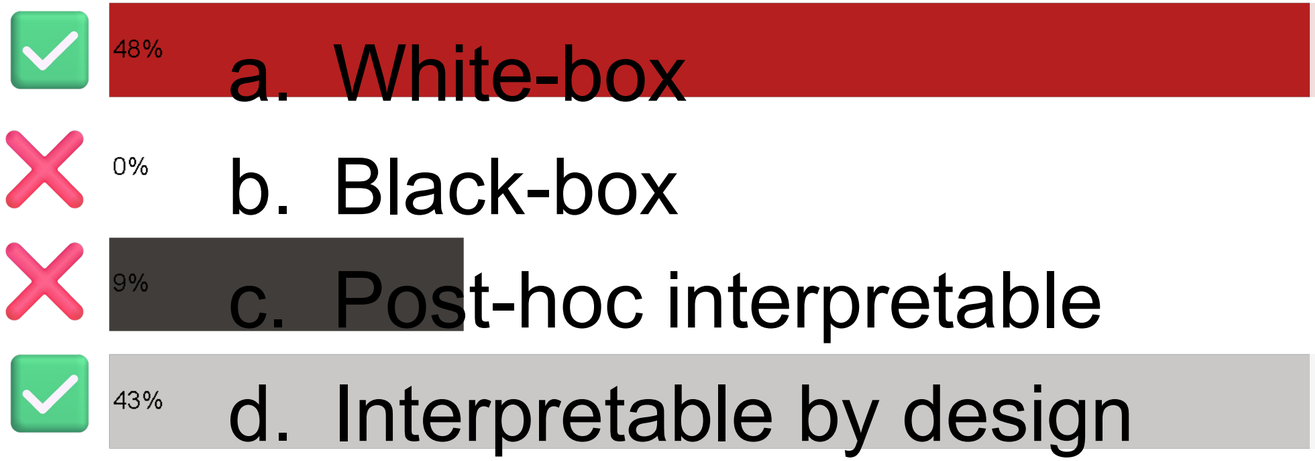
⚠️ In Fairness 1, the university admission software is **NOT** a ML model, it is a “classic” algorithm designed by hand (not by learning from data), which is why we CAN modify the “parameters” [the goal was to show you that unfairness is not specific to ML, it happens with classic algorithms too]

Logistic Regression Model

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 all that apply):



"White-box" models are interpretable by design (i.e. the two terms are synonyms)

Note: a post-hoc interpretability method CAN be used on a white-box model, however this is generally not done because it is unnecessary in most cases (since white-box models are interpretable by definition) and it does not bring any advantage (since trust in post-hoc methods is generally lower because they are external to the model)

COMPAS

URL: ttpoll.eu
Session ID: cs290

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

46% a. The design documentation

0% b. The user documentation

23% c. The training code

15% d. The training dataset

0% e. A post-hoc interpretability method

✅ 15% 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 **courseware**

2. Find 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 strategy category
+ 1 harmful action with negative impact

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

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

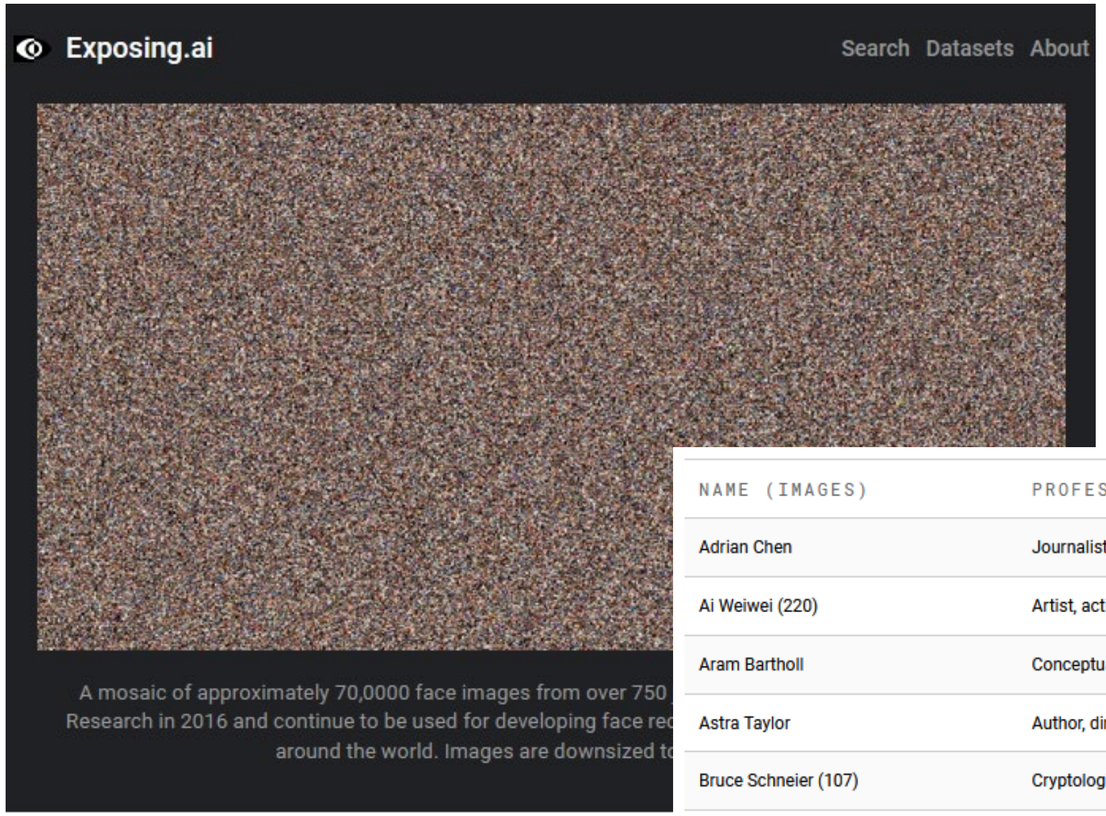
Post your ideas:
https://speakup.epfl.ch
Room key: **64393**

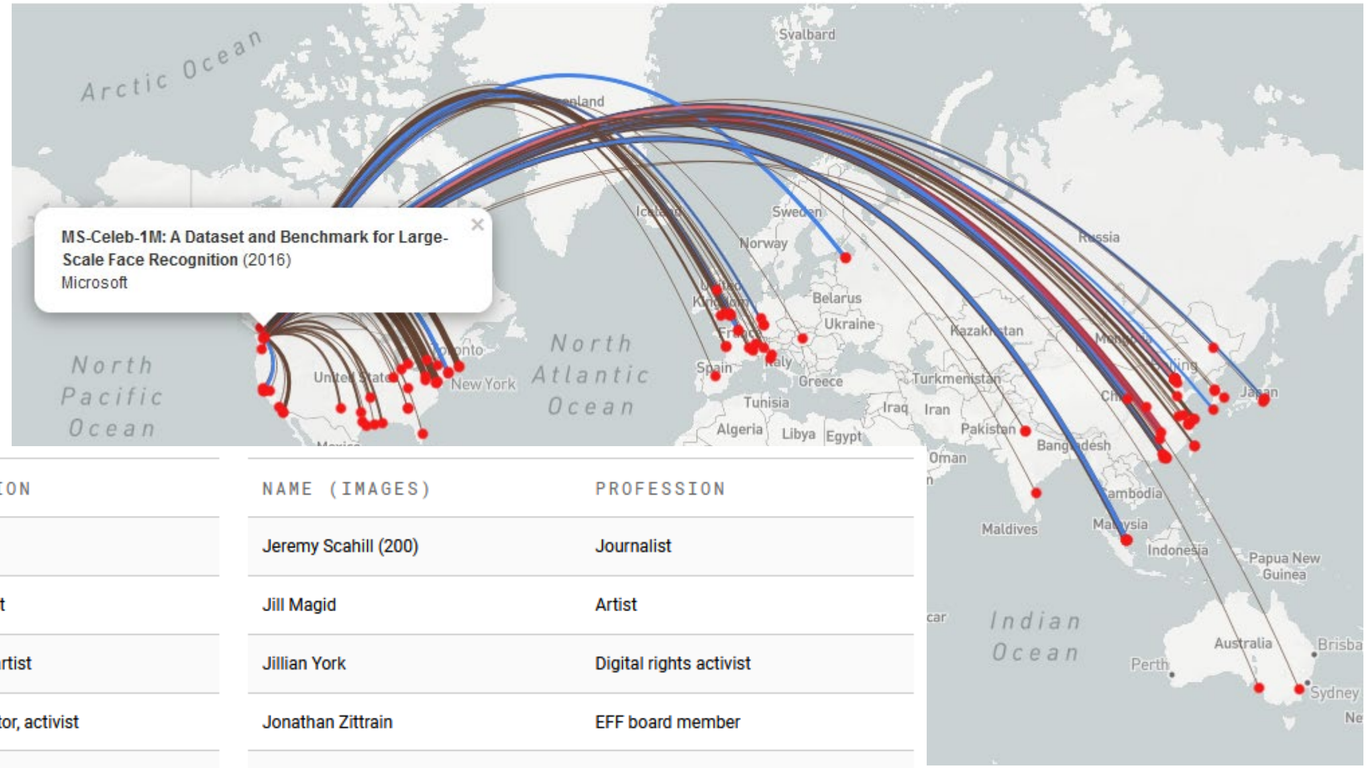
Datasheet

- 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

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

Part I – The dark and pessimistic story

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

Related to (dis-)empowerment

(Dis-)Empowerment: Ethical Speculation

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

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

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

Part II – The happy ending

1. Develop 1 or 2 sentences that describe:
   - Immediate harms
   - Future consequences
2. Imagine a happy ending for your character
   - What can the creators of the product do? Or other stakeholders?
   - What could we do now to prevent future harm?

What’s next?

Next dates

	Monday (STCC Cloud C)	Tuesday (INF1 & CO5)
8 Dec – 14 Dec	Empowerment 2 cases	Graded Case
15 Dec – 21 Dec	Conclusion cases + Q&A	---

Q&A: you can ask any question you want on the course content, the exam, etc. before 23h59 on Dec. 14:

👉 1 post = 1 question

Post your ideas:
https://speakup.epfl.ch
Room key: **53228**

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/