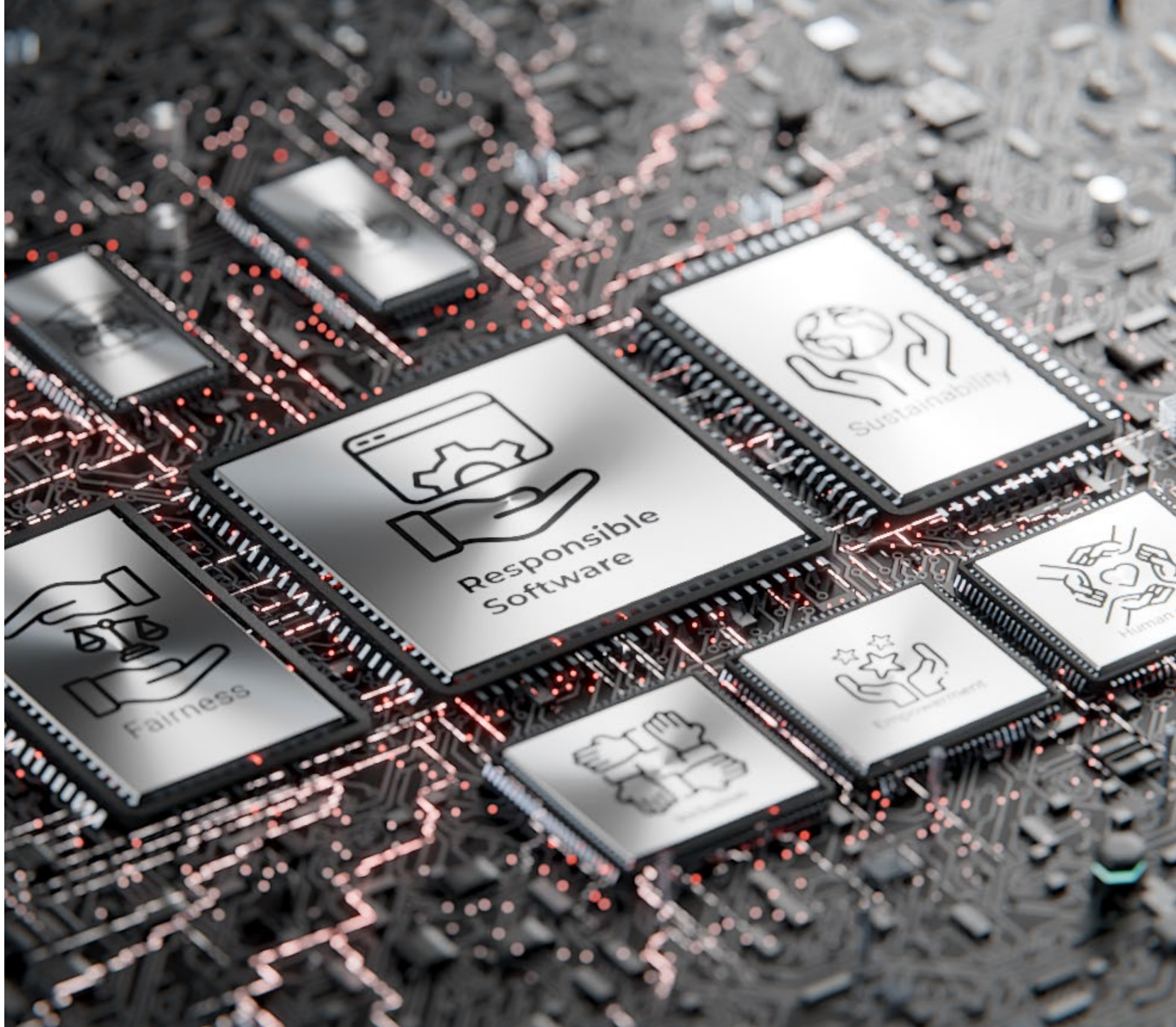
Conclusion
Case studies
+ Q&A
15 dec.

Cécile Hardebolle

Responsible Software



Agenda for today

1. Next dates
2. Exam logistics

3. Review case studies:
   a) Ethics Canvas – Be My Eyes
   b) Digital Ethics Canvas – Emotion Cancelling AI

4. Q&A: your questions

5. Some interactive review questions

Next dates

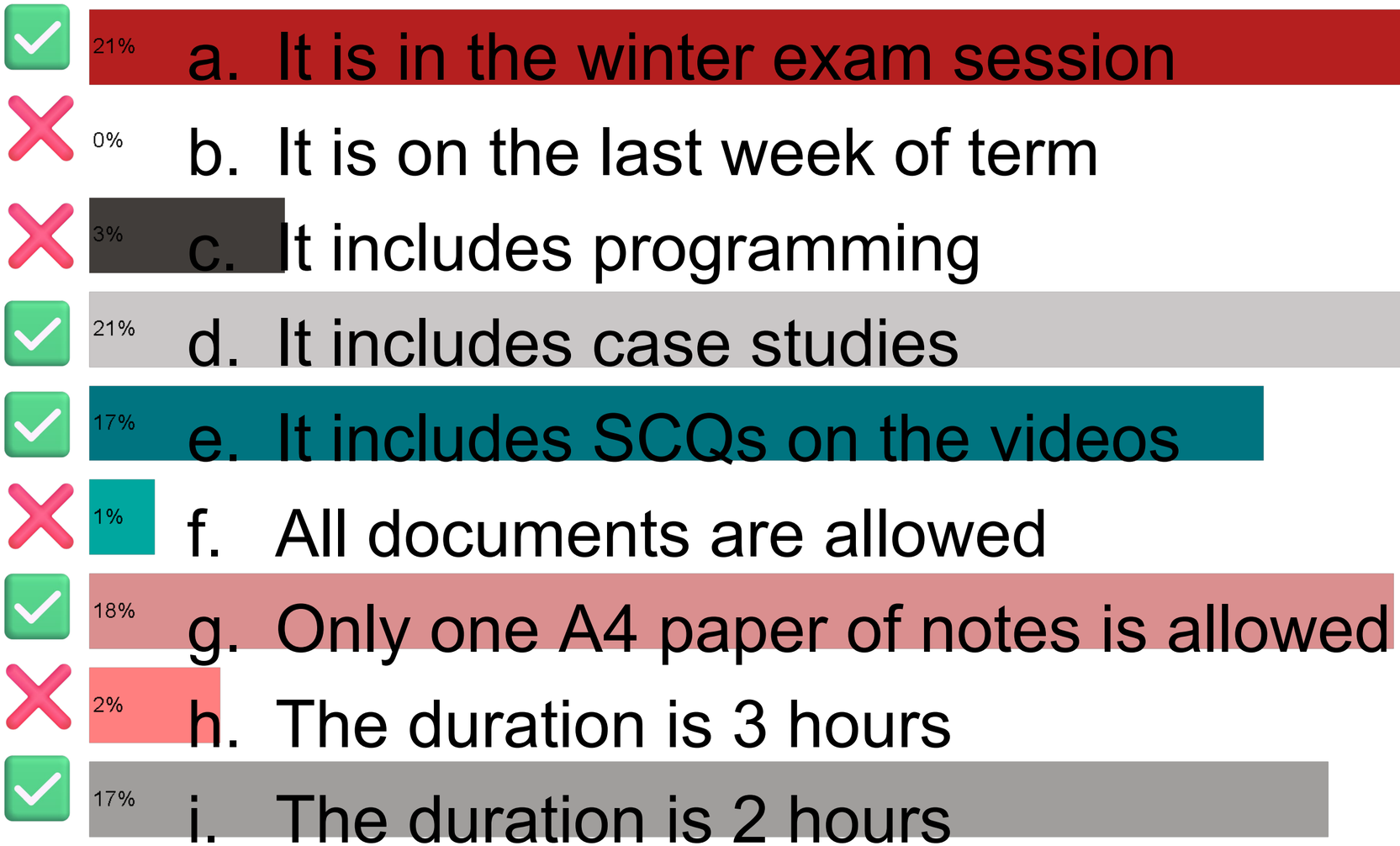
- There is **no exercise session tomorrow**.

- Now is the time to give your **overall feedback on the course!**
  - Online form on moodle dashboard & PocketCampus app
  - Available until **January 11**
  - Space for comments!
    - Most interesting / least interesting
    - Most clear / least clear
    - Suggestions for improvement

Exam logistics

About the final exam

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



URL: ttpoll.eu
Session ID: cs290

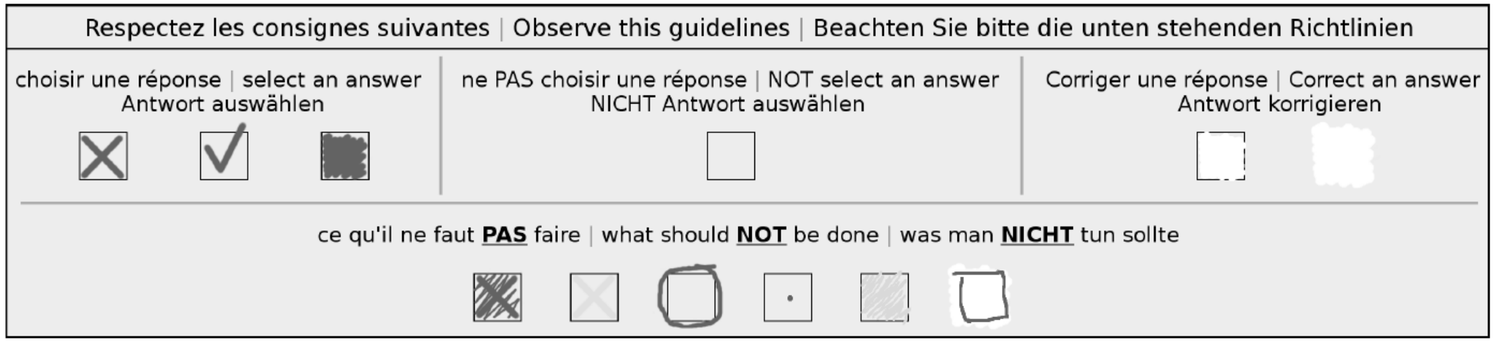
Exam rules (reminder)

Date: 21 January, 9h15
Duration: 2h, except arrangements

- 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
  - Whitened = not selected



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** **or** **ID** on your table

- The exam **starts at 9h15** and you have **2h to work**,
  except special arrangements
  - The exam copy is on your table – you MUST wait until 9h15 to open it
  - When indicated (normally 11h15), 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

Formatting answers

Task:

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

- [4 x 1 point] For cells A, B, C and E: describe 1 harm that corresponds to the category (1-2 sentences for each harm)
- [1 point] For cell D: indicate the corresponding harm category

Make sure to identify your answers with the corresponding letters (no need to reproduce the table).

Category	**Type of harm**	**Description of harms**
Humans	Physical injury	**A)**
Allocation of Resources	Opportunity loss	**B)**
Human Rights	Liberty loss	**C)**
	D)	Most int
Social System Harms	Social detriment	**E)**

Proposed answer

(A) If the chatbot gives users unsafe or negligent advice (such as suggesting meeting up with a stranger in a remote location or promoting risky activities), they may be exposed to physical injuries. They may tend to follow advice that expose them to risk because they consider the chatbot to be an "expert".

(B) In companies, using the chatbot to predict employees' emotions in the context of promotion or reward decisions may lead to some people being disadvantaged because they don't fit the underlying model (which may also be biased). Their access to opportunities would be unfairly restricted.

Case studies

Where to find the cases?

1. Go to **courseware**

2. Find the **case studies** for today: **Conclusion**

3. Download the **instruction sheet**

+ From previous chapters, you will need:

- Digital Ethics Canvas (7 – Empowerment 1)
- Ethics Canvas (2 – Safety 2)

Ethics Canvas
(review from Safety 2)

Instructions

- Read the software description
  (you can also take a look at the referenced website)
- Fill out the Ethics Canvas:
  - **Stage 1: Identify relevant stakeholders**
    ➡️ fill out blocks 1 and 2
  - **Stage 2: Identify ethical impacts**
    ➡️ fill out blocks 3, 4, 5, 6, 7 and 8
  - **Stage 3: Discuss remedial actions**
    ➡️ fill out block 9

“Be My Eyes”

👉 1 post = 1 stakeholder

Post your ideas:
https://speakup.epfl.ch
Room key: **50113**

👉 1 post = 1 ethical impact

Post your ideas:
https://speakup.epfl.ch
Room key: **01710**

“Be My Eyes”: remedial actions

👉 1 post = 1 remedial action (“what can we do”)

Post your ideas:
https://speakup.epfl.ch
Room key: **65533**

Digital Ethics Canvas
(review from Empowerment 1)

Instructions

- Read the software description
  (you can also take a look at the referenced news article)
- Fill out the Digital Ethics Canvas:
  - Context & Solution
  - **Benefits**: list 3 benefits (think about a range of stakeholders)
  - **Risks**:
    - For each of the 5 lenses identify and **describe 1 risk**
    - Select 1 risk and evaluate its **overall level**:
      - Severity of **impacts**
      - **Probability** to happen

Probability \ Severity	low	mid	high
low	low	low	mid
mid	low	mid	high
high	mid	high	high

  - **Mitigation**: for each risk, identify a corresponding mitigation measure

“Emotion Cancelling AI”

👉 1 post = 1 benefit

Post your ideas:
https://speakup.epfl.ch
Room key: **57217**

👉 1 post = 1 risk + ethical lens

Post your ideas:
https://speakup.epfl.ch
Room key: **38467**

Risks:
Make sure to explain how the risk relates to the corresponding ethical lens (e.g. if you put a risk into “Fairness”, it must be clear what is unfair or biased).

Evaluating the level of risk - 1

URL: ttpoll.eu
Session ID: cs290

Consider the following Privacy risk: **“Identifying customer emotions can lead to the disclosure of information the customers might consider private”**. How would you evaluate the level of this risk in terms of probability and severity of impacts? (select 2 options: 1 for probability, 1 for severity)

a. Probability: low
b. Probability: medium
c. Probability: high
d. Severity: low
e. Severity: medium
f. Severity: high

- Do not mix probability (“when?”) with severity (“what?”)
- Make sure you know how to get the overall level of risk using the risk matrix

Evaluating the level of risk - 1

URL: ttpoll.eu
Session ID: cs290

Consider the following Autonomy risk: **“The system’s real-time alterations may reduce employee's ability to rely on their own judgment in emotionally charged situations”**. How would you evaluate the level of this risk in terms of probability and severity of impacts? (select 2 options: 1 for probability, 1 for severity)

a. Probability: low
b. Probability: medium
c. Probability: high
d. Severity: low
e. Severity: medium
f. Severity: high

“Emotion Cancelling AI”: mitigation

Consider the following Privacy risk: **“Identifying customer emotions can lead to the disclosure of information the customers might consider private” [HIGH RISK]**

Which mitigation options could help reduce the risk?

👉 1 post = 1 mitigation option

Post your ideas:
https://speakup.epfl.ch
Room key: **02297**

Comparison!

Ethics Canvas

Project Title:

Date:

Individuals affected

Identify the types or categories of individuals affected by the product or service, such as men/women, user/non- user, age-category, etc.

1

Behaviour

Discuss problematic changes to individual behaviour that may be prompted by the application e.g. differences in habits, time-schedules, choice of activities, people behaving more individualistic or collectivist, people behaving more or less materialistic.

3

Relations

Discuss problematic differences in individual behaviour such as differences in habits, time-schedules, choice of activities, etc

4

What can we do?

Select the four most important Ethical impacts you discussed. Identify ways of solving these impacts by changing your project's product/service design, organisation.Or by providing recommendations for its use or spelling out more clearly to users the values driving the design

9

Worldviews

Discuss how the general perception of somebody's role in society can be affected by the project,

5

Group Conflicts

Discuss the impact on the relationships between the groups identified, e.g. employers and unions

6

Groups affected

Identify the collectives or communities, e.g. groups or organisations, that can be affected by your product or service, such as environmental and religious groups, unions, professional bodies, competing companies and government agencies, considering any interest they might have in the effects of the product or service.

2

Product or Service Failure

Discuss the potential negative impact of your product or service failing to operate as intended,eg technical or human error, financial failure/ receivership/acquisition, security breach, data loss, etc.

7

Problematic Use of Resources

Discuss possible negative impacts of the consumption of resources of your project, e.g. climate impacts, privacy impacts, employment impacts etc.

8

DIGITAL ETHICS CANVAS

CONTEXT | SOLUTION | BENEFITS

WELFARE

RISK

- Can the solution be used in harmful ways, in particular with regards to vulnerable populations?
- What kind of impacts can errors from the solution have?
- What type of protection does the solution have against attacks or misuse?

MITIGATION

FAIRNESS

RISK

- How accessible is the solution?
- What kinds of biases may affect the results ?
- Can the outcomes of the solution be different for different users or groups?
- Could the solution contribute to discrimination against people or groups?

MITIGATION

AUTONOMY

RISK

- Can users understand how the solution works and what its limits are?
- Are users able to make choices (e.g. consent, settings) in their use of the solution and how?
- How does the solution affect user autonomy and agency?

MITIGATION

PRIVACY

RISK

- What data does the solution collect
- Is it collecting personal or sensitive data
- Who has access to the data?
- How is the data protected?
- Could the solution disclose / be used to disclose private information?

MITIGATION

SUSTAINABILITY

RISK

- What is the carbon footprint of the solution?
- What types of resources does it consume (e.g. water) -and produce (e.g. waste)?
- What type of human labor is involved?

MITIGATION

Digital Ethics Canvas 2024 – C. Vanhebelle, V. Mackx, V. Ramachandran, T. Du, N. Bartholomäyczyk, A. Holzer, P. Jermann

Q&A

Question – Sustainability

“For the sustainability section: If we take into account the energy mix every time when computing the carbon footprint of a service, doesn't this create a very problematic framework where "good" countries get to have all their digital services hosted on their soil, but countries with a less renewable energy mix "should not"? Many of the countries that have a clean energy mix today, used to rely just as much on coal/oil, as developing countries do now. How can we address this historical imbalance correctly when analysing the carbon footprint?”

- Very good question!
- Both an issue (increased energy consumption, stress on infrastructures & pop) and an advantage (“sovereign” services)
- Multi-faceted problem: benefit-risk analysis, ethical frameworks

Question – Revision strategies 1

- “Would it be a good idea to rewatch all the videos if we have a "short memory" (as in not remembering all the details of each topic) , and what would be the best recommended way to revise for this subject.”
- “do you think that working in groups for example making the sheetshit is a good idea”
- “Do you consider flashCard a good way to prepare ? “

Cf. tips for effective revisions
+ groups can be helpful in addition to individual work

Question – Revision strategies 2

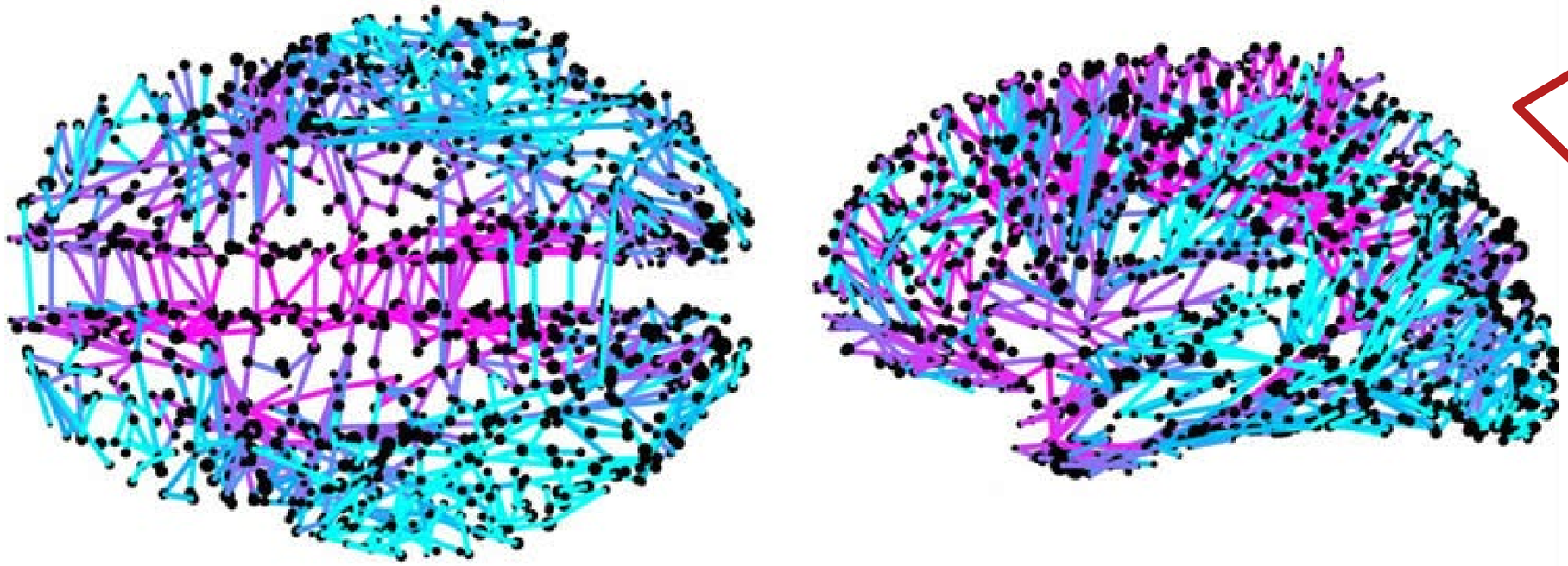
“Will we have access to more material to practice for the exam ?
Because right now we only have last year's mock exam...
If that's the only resource you're offering us, how do you think we can best prepare ? (For example, would it be appropriate to ask an LLM to generate questions similar to those in the mock exam ?)”

- Please use the mock exam from this year
- Other resources available:
  - All **case studies** with proposed answers
  - All in-class interactive **questions (exam type)** + *new questions today*
- Use of LLM to generate exam-type questions: I do not recommend because of hallucination risk + (in)correct distractors

Tips for effective revisions

Recall = *reconstruct*

(Mišić et al., 2015)



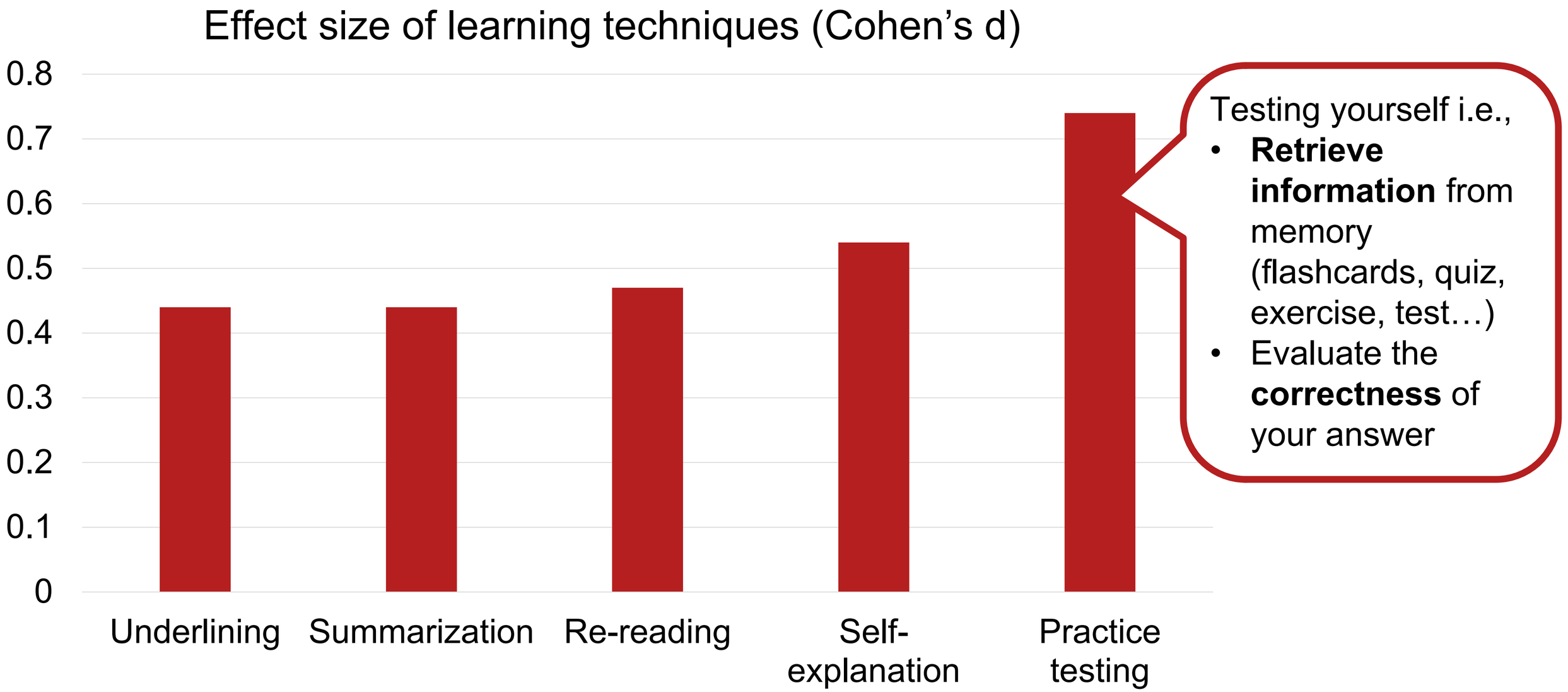
Times at which different connections in the brain are used to spread information

Information is NOT “stored” into memory like in a “drawer”, instead it is **reconstructed** from a **network of connections**

👉 **You should practice reconstruction**

Learning techniques

(Donoghue & Hattie, 2021)



Recommended revision techniques

a. Create and use **flashcards**
b. Re-do the in-class **interactive questions**
c. Recall course content **from memory** without cues/prompts
d. Re-do **case studies** then check the solutions
e. Re-do the **mock test** in exam-like conditions (limited time)
f. Prepare your **A4 sheet of notes**, focusing on structure

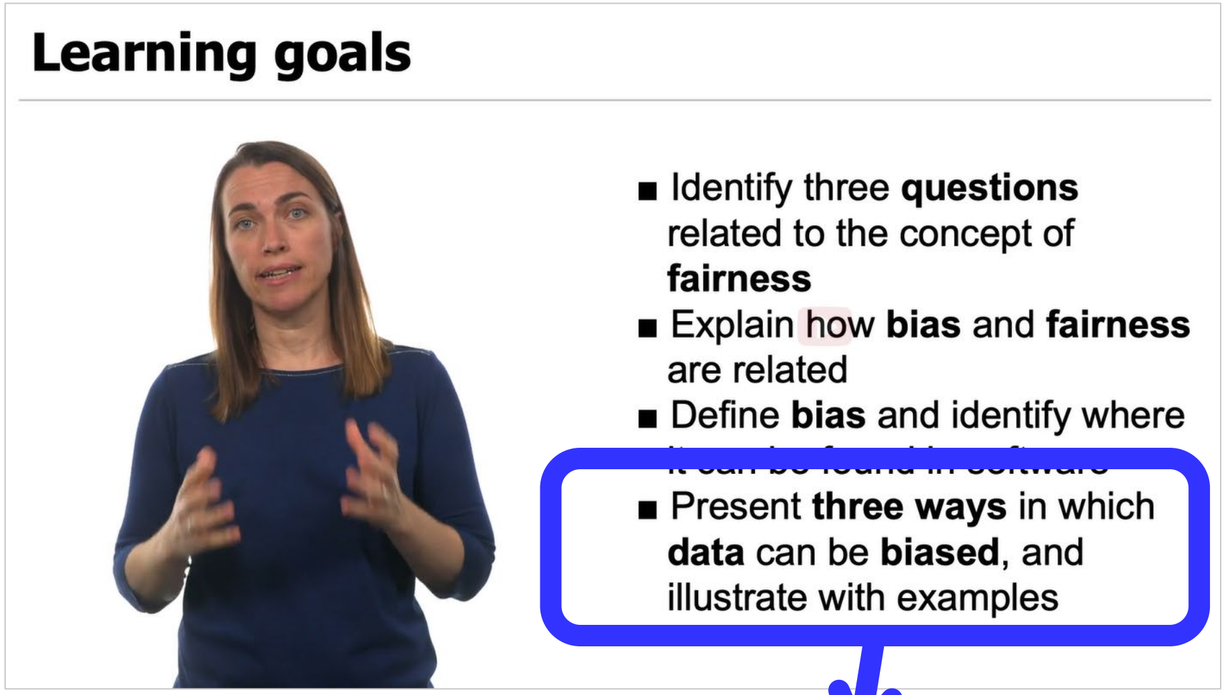
Flashcards

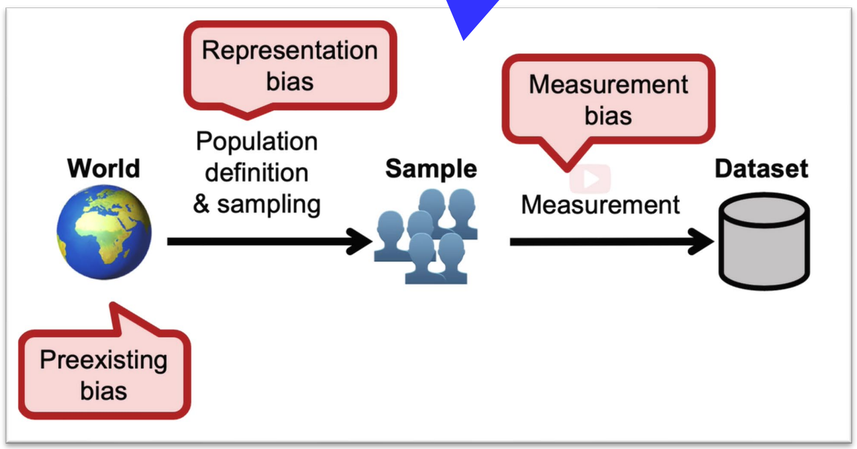
1 card:

- Recto = goal or question
- Verso = answer

Content to create cards:

- **learning goals** at the beginning of the videos
- **in-class interactive questions** (answers removed)
  ⚠️ questions NOT tagged
  “exam type” may have multiple correct answers etc.





Free recall

- Take a **blank sheet** of paper
- **Note down** everything you remember on:
  - A video
  - A module
  - A chapter
- Then **check** against your notes

Review questions “Whole Course”

Ethical sensitivity

New

URL: ttpoll.eu
Session ID: cs290

What is ethical sensitivity?

0% a. The ability to predict all technical outcomes before deployment

✅ % b. The capability to identify the impact of a situation on others

0% c. The ability to act to benefit others even at your own expense

18% d. The capacity to account for all ethical values simultaneously

Chemical discovery

URL: ttpoll.eu
Session ID: cs290

A software company has developed a Machine Learning model that is able to discover new chemical compounds for medicine development. They identify that the model can also discover new chemical weapons.

What type issue is this?

0% a. A technical issue

22% b. An ethical issue

✅ 78% c. An ethical dilemma

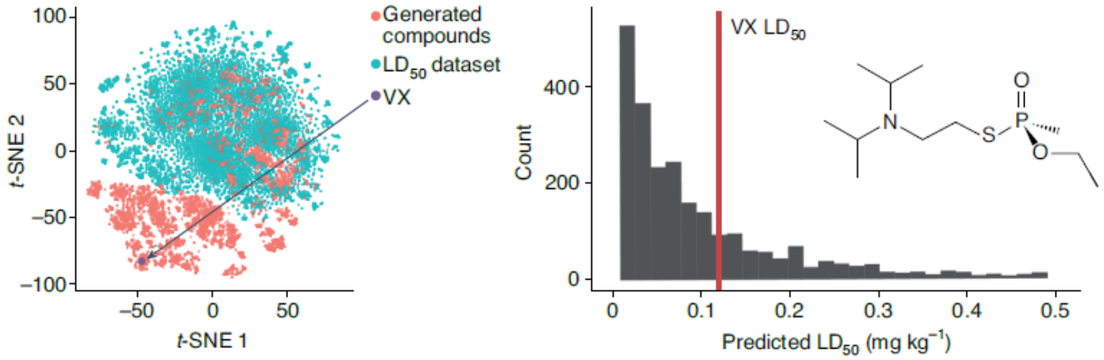


Fig. 1 | A *t*-SNE plot visualization of the LD_{50} dataset and top 2,000 MegaSyn AI-generated and predicted toxic molecules illustrating VX. Many of the molecules generated are predicted to be more toxic in vivo in the animal model than VX (histogram at right shows cut-off for VX LD_{50}). The 2D chemical structure of VX is shown on the right.

(Urbina et al., 2022)

Vulnerabilities

New

URL: ttpoll.eu
Session ID: cs290

A software engineer decides to postpone the launch of a new feature due to the late discovery of a security vulnerability and justifies: “The new feature would bring us some short-term benefits but would have serious negative consequences for all of our customers, our aim must be the greatest good for the greatest number.”
Which ethical theory does this engineer follow?

Exam type

Food delivery

New

URL: ttpoll.eu
Session ID: cs290

An online food delivery app experiences a data breach where customer payment details are stolen.
What type of risks are represented in this situation?

0% a. Safety risks from misdiagnosed food allergies

4% b. Safety risks from incorrect delivery scheduling

13% c. Sociotechnical risks in app-driver communication

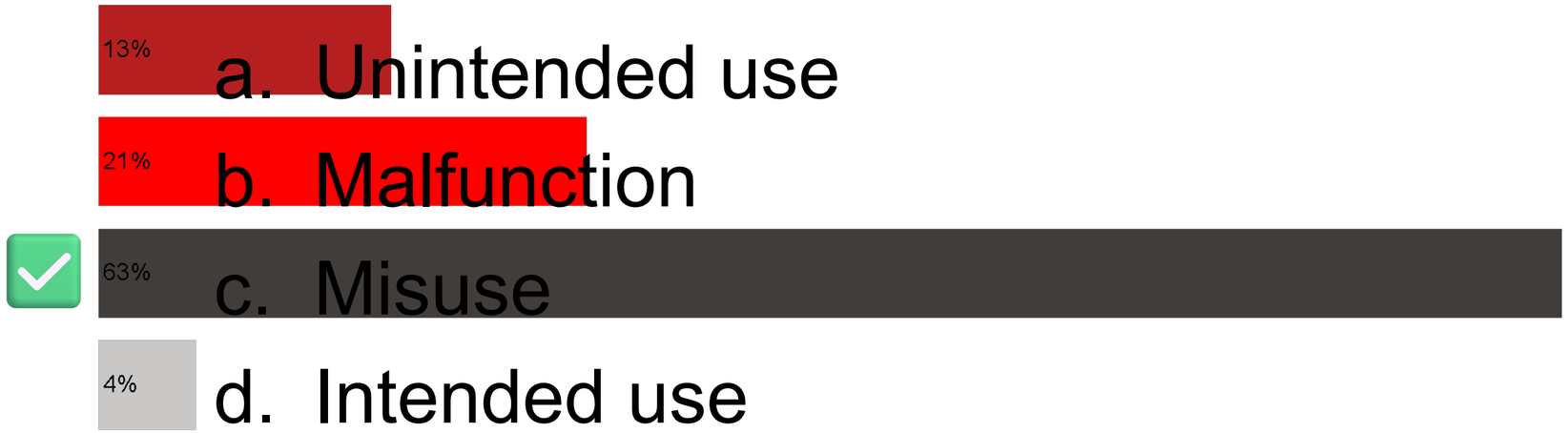
✅ 83% d. Security risks from unauthorized system access

Hospital

New

URL: ttpoll.eu
Session ID: cs290

Patient records in a hospital have been encrypted by cybercriminals who demand payment to restore access, causing emergency services to halt and delay critical care for patients.
Which harm scenario does this represent?



Exam type

Fissures in concrete

URL: ttpoll.eu
Session ID: cs290

The company SuperCrack has developed a model to detect fissures in concrete walls before they become visible to the naked eye. A positive result means presence of fissure.
Which of the statements below is correct?

87% a. TN = actual absence of fissure, correct prediction

4% b. TN = actual presence of fissure, incorrect prediction

0% c. TP = actual presence of fissure, incorrect prediction

9% d. TP = actual absence of fissure, correct prediction

Exam type

Contagious disease

New

URL: ttpoll.eu
Session ID: cs290

A rapid test for a contagious disease (infected = positive result) shows a high number of false negatives.
What are the consequences of false negatives in terms of safety?

0% a. Healthy people continue their daily activities as normal.

4% b. Healthy people receive unnecessary quarantine.

0% c. Infected individuals receive the appropriate medication.

✅ 96% d. Infected individuals spread the disease unknowingly.

Exam type

Political campaign

New

URL: ttpoll.eu
Session ID: cs290

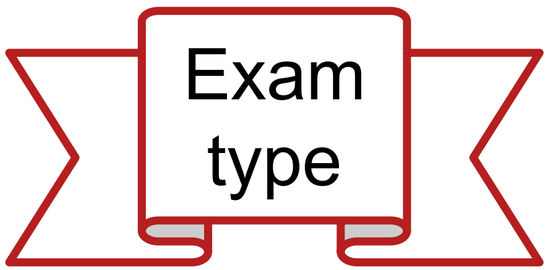
A whistleblower releases authentic internal documents from the campaign of a political party with the goal of damaging the party's public image for the upcoming election.
What type of information is this?

25% a. Misinformation

8% b. Disinformation

✅ 67% c. Malinformation

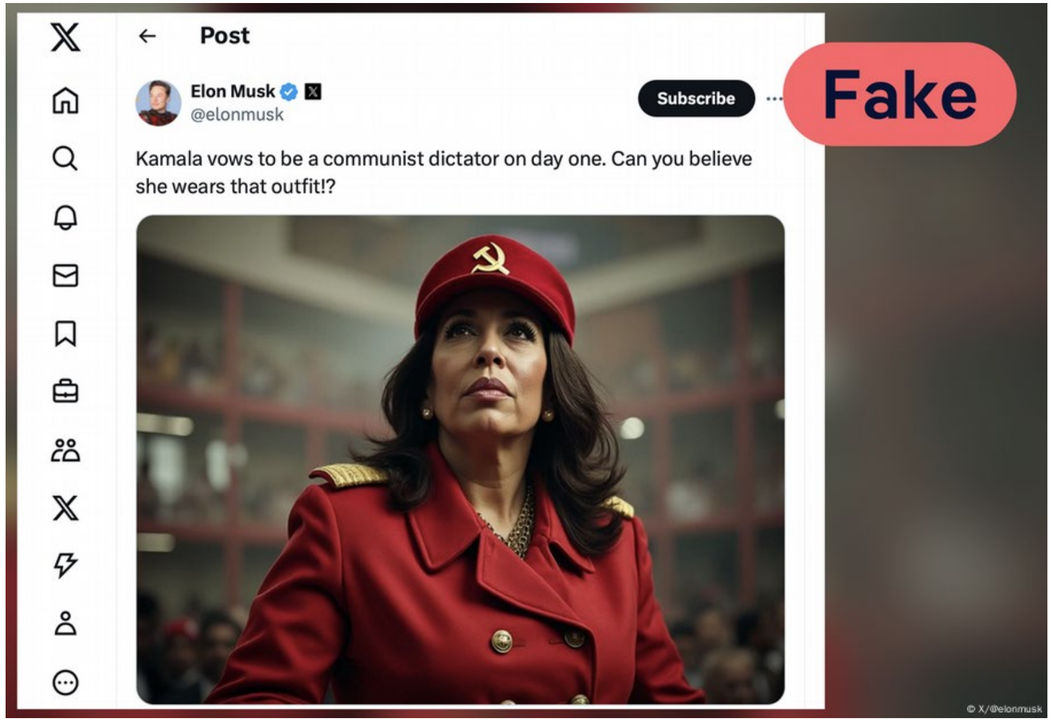
0% d. Fake news



Posts on Twitter

URL: ttpoll.eu
Session ID: cs290

One dis-/mis-information post by Elon Musk appears in your Twitter timeline.
Why would you be more likely to believe it than other posts?



Fact check: Elon Musk spreads US election lies. (2024, February 11). Dw.Com. https://www.dw.com/en/fact-check-how-elon-musk-is-spreading-us-election-lies/a-70663408

0% a. System 2

30% b. Illusory truth

✅ 65% c. Source cues

5% d. Prebunking

Exam type

Loans

New

URL: ttpoll.eu
Session ID: cs290

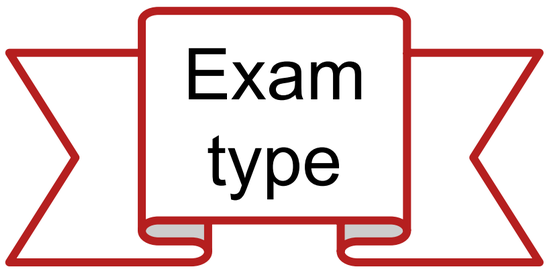
A ML model for loan approval consistently denies loans to applicants from rural neighborhoods. The model has been trained on data from the bank covering all the loan decisions taken in the last 5 years for all the neighborhoods served by the bank.
Which type of bias is most likely present in the data from this scenario?

11% a. Sampling bias

26% b. Representation bias

0% c. Measurement bias

✅ 63% d. Preexisting bias



Shoplifting

URL: ttpoll.eu
Session ID: cs290

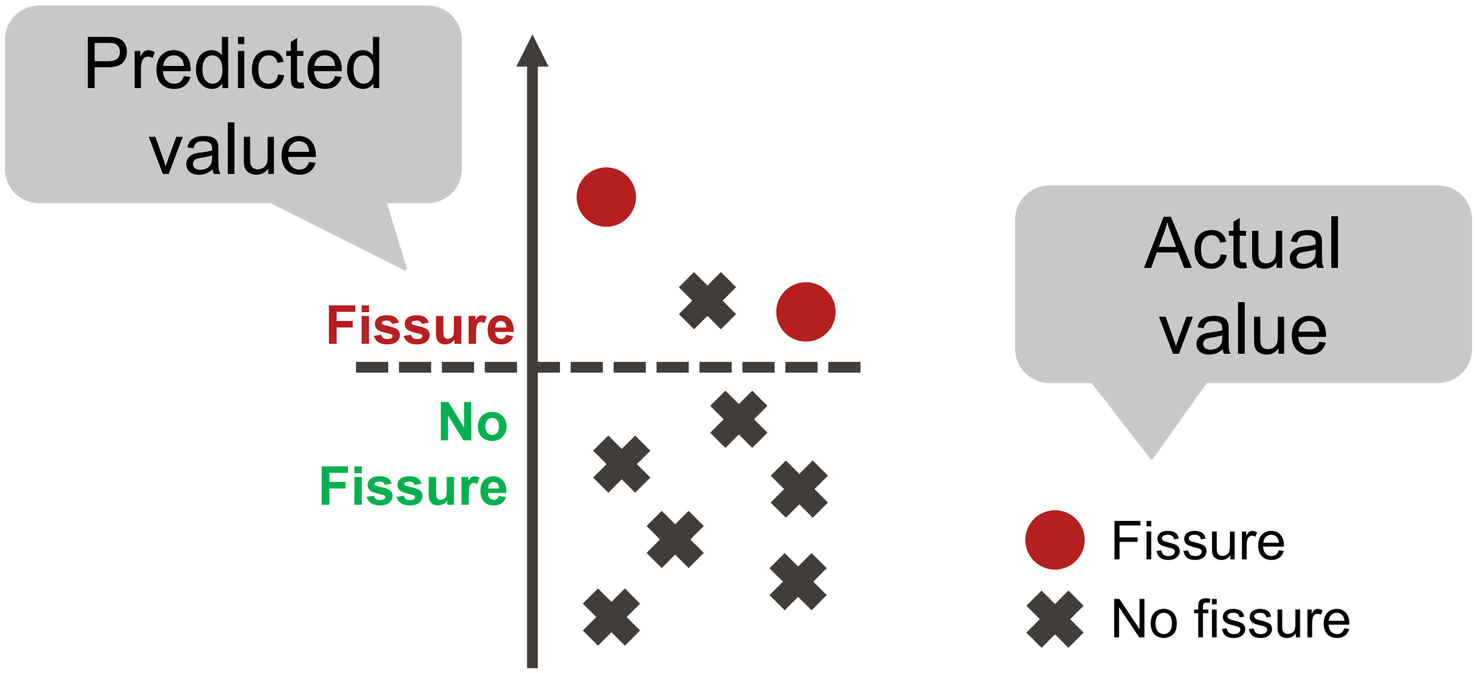
The society RetailProtect develops a ML model to identify instances of shoplifting in retail shops. They evaluate their model on a benchmark in which actors from diverse ethnicities simulate a range of shoplifting actions. They plan to deploy soon in shops.

What type of bias is present in this scenario?

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

Fissures in concrete (again)

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:

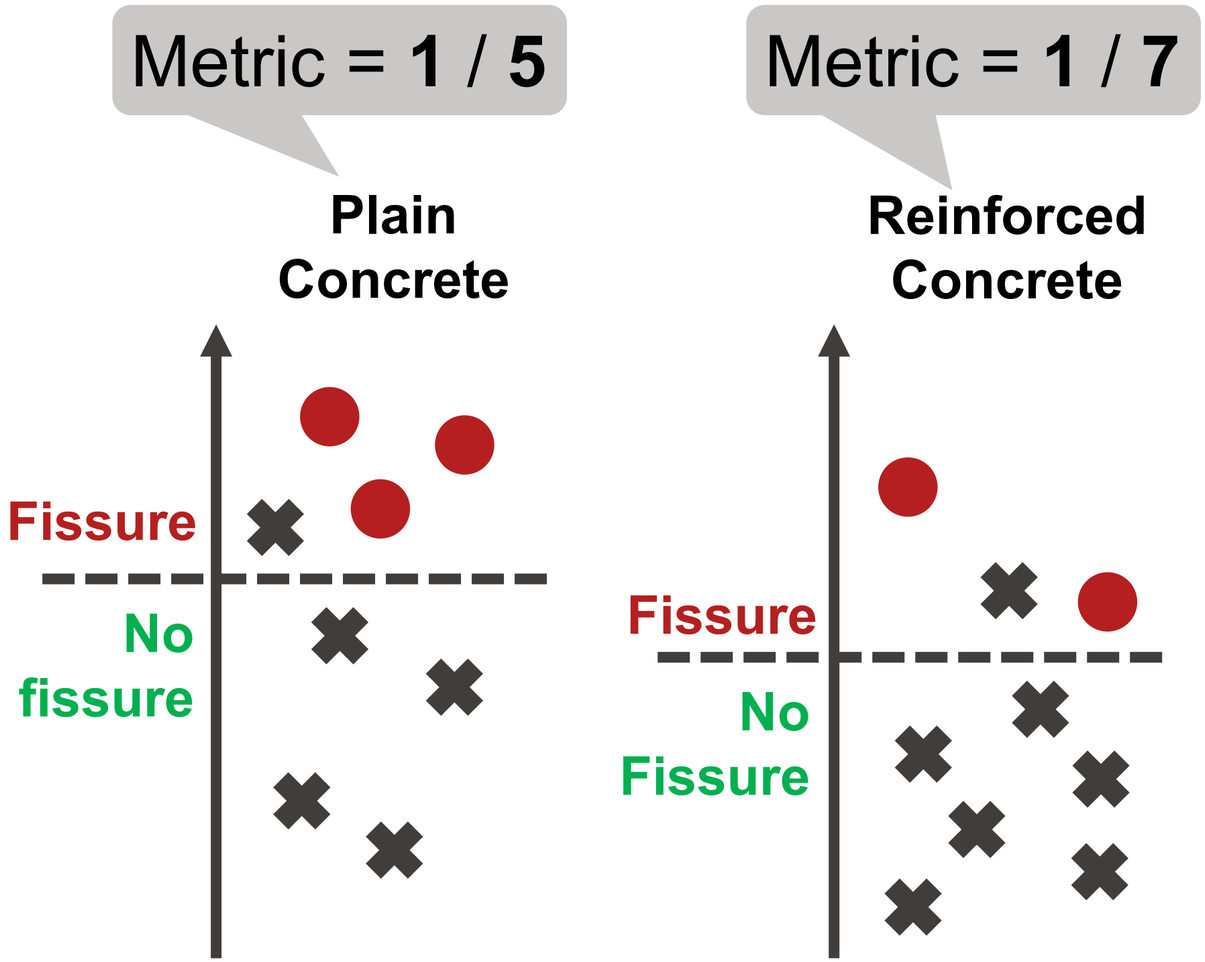


		Predicted: Fissure	Predicted: No Fissure
Actual	Fissure	2	0
Actual	No Fissure	1	6

Fissures in concrete (again) New

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)

0% a. Accuracy

0% b. FNR

0% ✅ c. FPR

0% d. Positive prediction rate

University admissions

New

URL: ttpoll.eu
Session ID: cs290

A model trained to help screen applications to university has an accuracy of 97% and the false positive rate (FPR) is 5% for group X and 6% for group Y. However, the Disparate Impact Ratio is 0,613 with group X having a higher admission rate.
What is most likely happening in this situation?

0% a. Differences in the FNR are causing the low DIR

0% b. The DIR indicates a higher error rate for group Y

0% c. The applicants from group X have stronger profiles

✅ 0% d. Group Y has a lower rate of actual positive labels

Seen in the Graded Notebook 2: inconsistency between fairness metrics indicate that the data used as "ground truth" i.e. actual data is biased, groups having dissimilar rates of positive labels (= pre-existing bias)
[This is what is called the "impossibility result"]

Datacenter cooling

URL: ttpoll.eu
Session ID: cs290

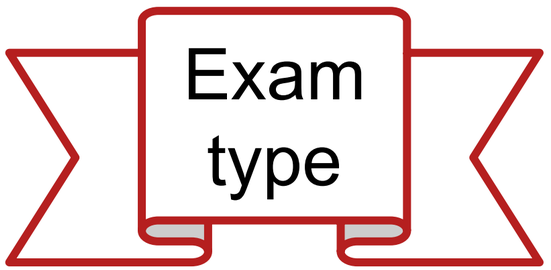
The GreenDC datacenter consumes an average of 1 MW.
This means annually a total of 8 760 MWh of electricity.
50% of this electricity is used to power the IT equipment.
What is the PUE of GreenDC?

0% a. 0.5

0% b. 1

0% c. 1.5

✅ 0% d. 2



Datacenter water

URL: ttpoll.eu
Session ID: cs290

The TitanCore datacenter consumes a total of 24 000 MWh of electricity annually. It consumes approximately 16 million liters of water each year.
What is the WUE of the datacenter (onsite only)?

0% a. 0,000667

✅ 0% b. 0,667

0% c. 1,5

0% d. 1500

$$WUE = \frac{16\,000\,000}{24\,000\,000}$$

$$WUE = 0{,}667\ L/kWh$$

Exam type

LLM training

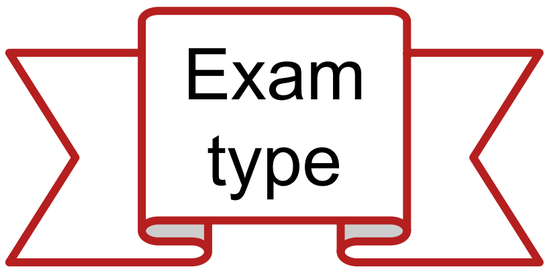
New

URL: ttpoll.eu
Session ID: cs290

The training of the LLM “BreezeTalk” took 3 months using 100% of the resources available on a 10-server cluster.
Each server has an embodied footprint of 1200 kg CO2e and a 3-year lifespan.
What share of embodied footprint should be allocated to BreezeTalk (training only), in kg CO2e?

a. 100
✅ b. 1000
c. 3000
d. 12000

M_s = M_h x time_share x resource_share
M_s = (10 x 1200) x (3/12 / 3) x (100 / 100)
M_s = 12000 x 1/12

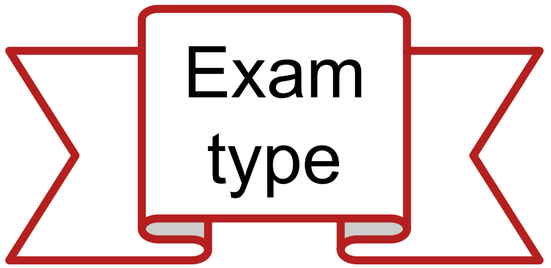


Meditation app

URL: ttpoll.eu
Session ID: cs290

ZenPath is an app dedicated to mental well-being that offers guided meditation sessions online. To reduce user dropout, they decide to display a popup after a user skips two sessions where the “Resume Today!” button is preselected.
What type of nudging technique is most likely used here?

0% a. Opt-in
0% b. Social proof
0% c. Scarcity
✅ d. Default



E-commerce platform

URL: ttpoll.eu
Session ID: cs290

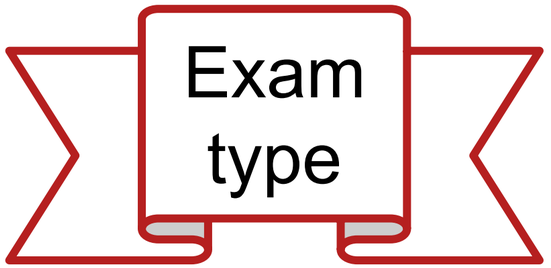
The e-commerce platform Shine would like to implement new features to improve the experience of its various categories of users. Here is the list of envisaged features.
Which of them best matches the definition of a deceptive pattern?

0% a. Personalize style recommendations based on past browsing

0% b. Display user-provided past purchase data to recommend sizes

✅ c. Register users to a ShineClub membership trial on checkout

0% d. Provide downloadable QR codes for the free return of items



Beer brewing dataset

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.

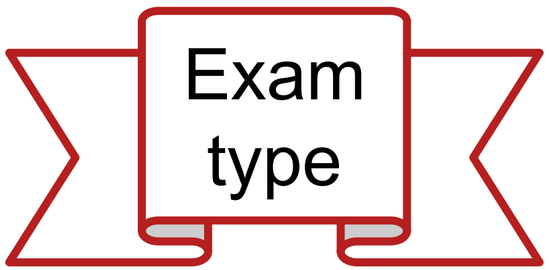
Which of the FAIR principles do you follow by providing a datasheet?

0% a. Findable

0% b. Accessible

0% c. Interoperable

✅ 0% d. Reusable

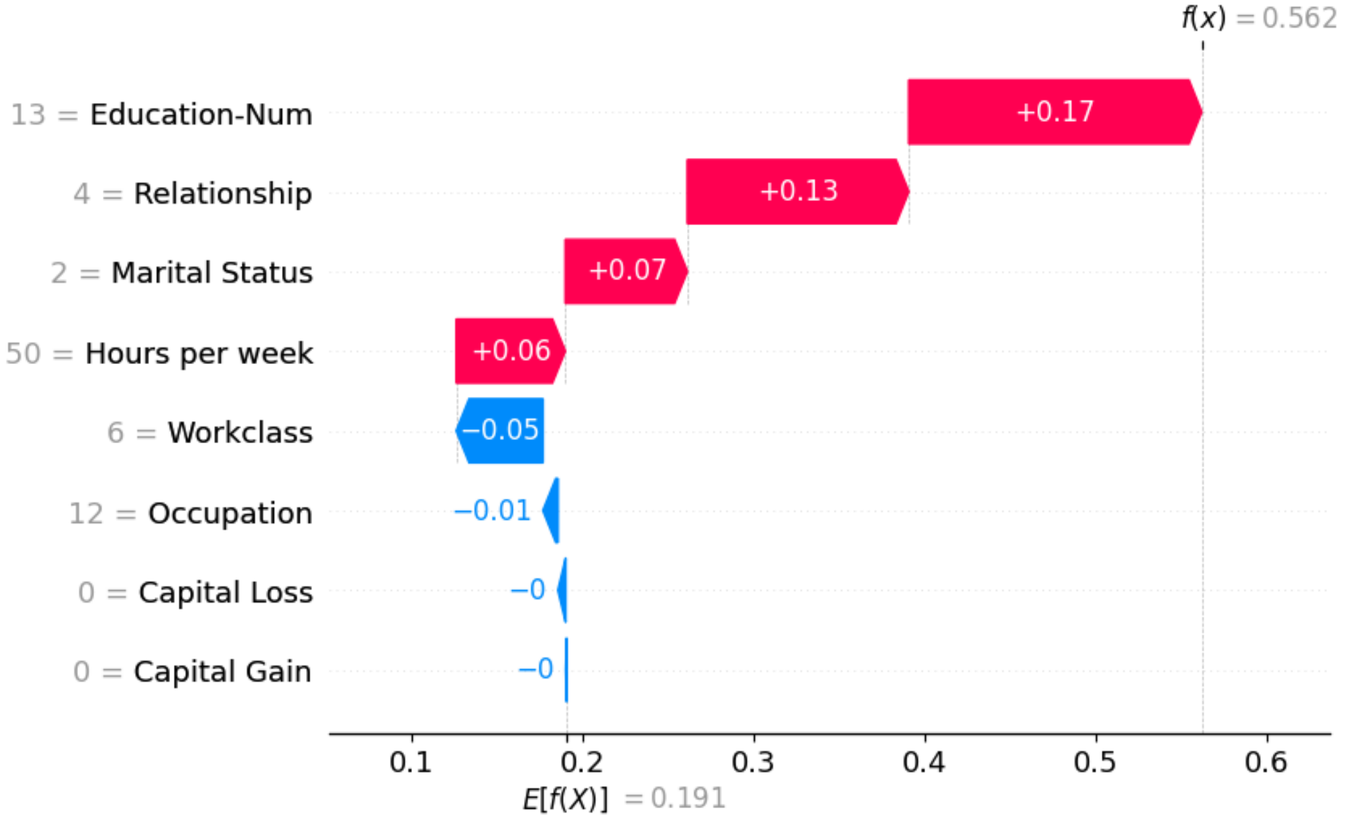


Loans (again)

New

URL: ttpoll.eu
Session ID: cs290

The plot on the right displays the SHAP values obtained for the prediction generated by our ML model for customer 1113.
What does this plot represent in terms of interpretability method?



✅ a. A local explanation

0% b. A global explanation

0% c. A feature importance analysis

0% d. A feature correlation analysis

Conclusion

Responsible engineering of software

"The **way a technology is designed** determines its possibilities, which can, for *better* or for *worse*, **have consequences for human wellbeing**."
(Roeser, 2012)

"Computing professionals' actions change the world. To **act responsibly**, they should **reflect upon the wider impacts** of their work, consistently **supporting the public good**."
(ACM, 2018)

Making engineering design **decisions** responsibly:

1. With a goal to **do good**
2. While **preventing *avoidable* negative impacts**
3. Taking **people**, other systems, **social structures** and our **planet** into account

Ethical decision-making

(adapted and simplified from: Schwartz, 2016; Rest, 1986)

Recognize

Evaluate

Act

Experience + Knowledge + Habits

Problem to solve, with **programming**

Ethical & technical **theory + strategies**

Application of strategies to cases

We MUST do better than that 👇

Meta and OpenAI have spawned a wave of AI sex companions—and some of them are children

The uncensored AI economy is booming, giving rise to hard legal and ethical questions.

BY **BEN WEISS** AND **ALEXANDRA STERNLICHT**

January 8, 2024 at 3:00 PM GMT+1

KLAATU BARADA NIKTO

Your AI clone could target your family, but there's a simple defense

The FBI now recommends choosing a secret password to thwart AI voice clones from tricking people.

BENJ EDWARDS – 6 DÉC. 2024 20:22 | 124

Universal credit

Revealed: bias found in AI system used to detect UK benefits fraud

Exclusive: Age, disability, marital status and nationality influence decisions to investigate claims, prompting fears of 'hurt first, fix later' approach

Robert Booth *UK technology editor*

Fri 6 Dec 2024 06.00 CET

And now what?

- You have the power to make some change!
- Don’t get fooled by the hype and the shiny useless/harmful software trends…
- A lot of questions seen in the course need more research!

Thank you for attending this course, good luck for the exams and all the best for all your projects! ✨ ✨ ✨

References

- https://fortune.com/longform/meta-openai-uncensored-ai-companions-child-pornography/
- https://arstechnica.com/ai/2024/12/your-ai-clone-could-target-your-family-but-theres-a-simple-defense/
- https://www.theguardian.com/society/2024/dec/06/revealed-bias-found-in-ai-system-used-to-detect-uk-benefits