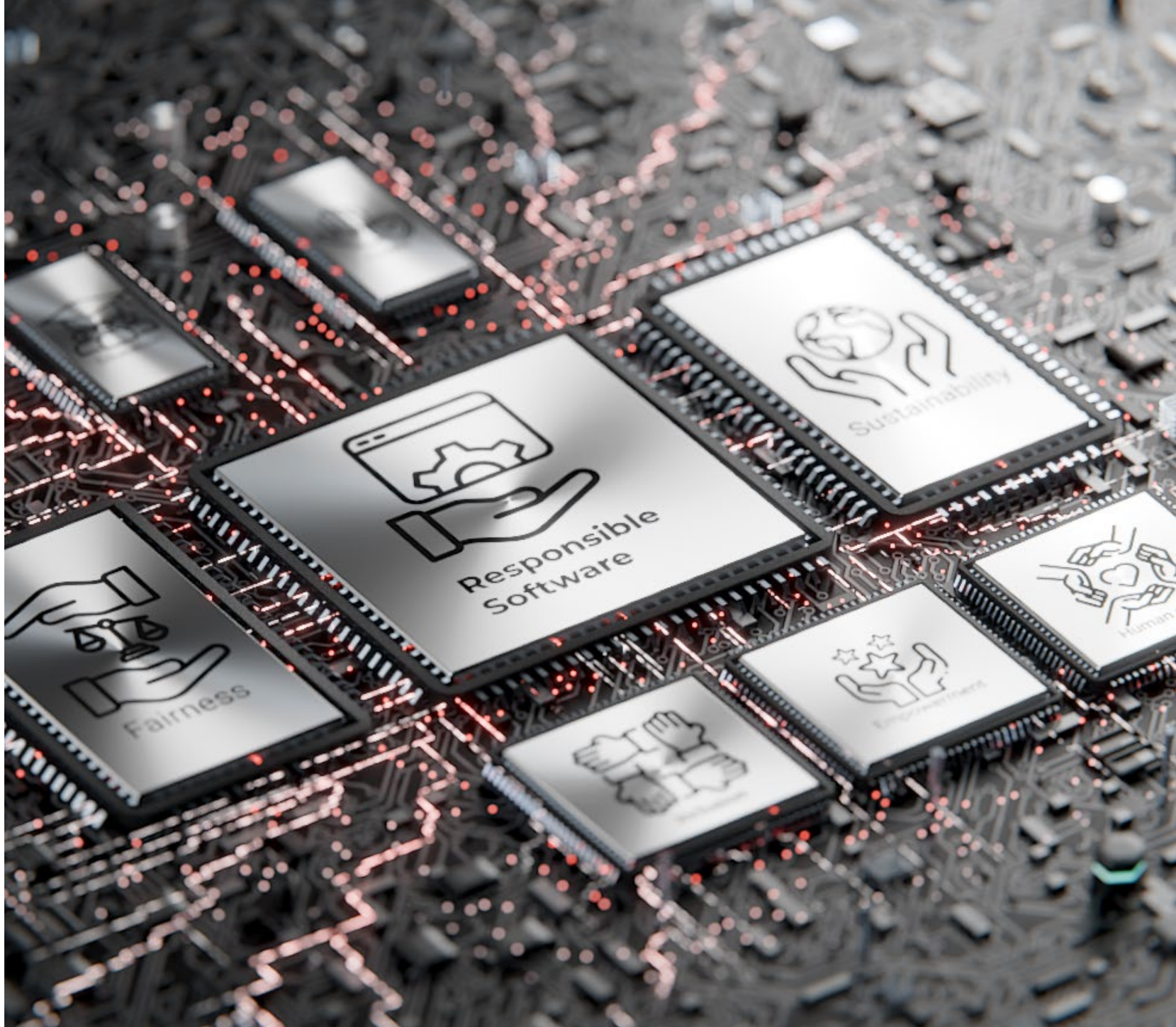
Blank Test Debriefing

4 nov.

Cécile Hardebolle

Responsible Software



Agenda for today

1. Debriefing of the blank test

2. Tips for effective revisions

Feedback on the Blank Test

Overall feedback

Thanks a lot to those who sent feedback! 🙏🙏🙏

What we will change for the exam based on your feedback:

- Single choice questions: we will remove “best answer” (?)
- T/F questions: we will remove the remove instructions “FALSE if it is not always true” (not appropriate in our context)
- We will further **clarify the instructions for the cases** + provide **indications for the length of answers** (e.g., number of sentences)

What will KEEP for the exam:

- Single choice and T/F questions will remain “positively” graded
- The overall format and length will remain similar

Single choice and T/F questions

Questions with more than 15% vote:

- 1
- 2
- 3
- 8
- 10
- 11

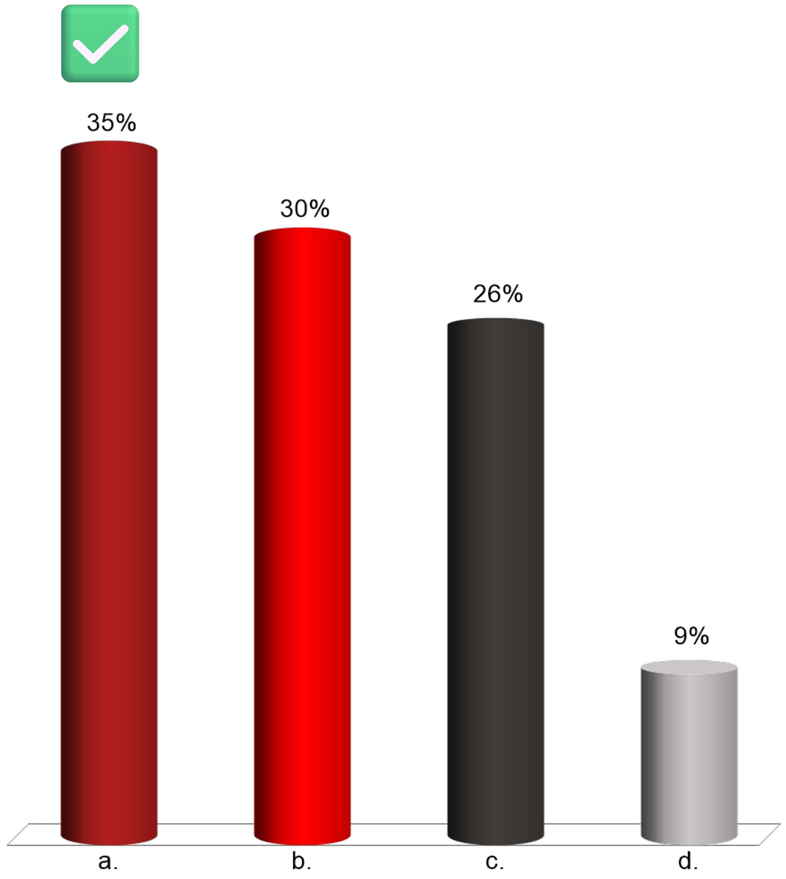
- 13
- 14
- 15
- 16

Question 1

URL: ttpoll.eu
Session ID: cs290

A group of computer scientists with similar background, all experts in AI-based software development, is aware of cognitive biases. They are starting a new AI project for healthcare and they aim to minimize the impact of these biases when making design decisions.
Select the strategy they should use:

a. Slow down the decision-making processes
b. Systematically include all members of their group to increase heterogeneity
c. Choose one or two of them to play the devil's advocate
d. Systematically include all members of their group to apply a participatory design method

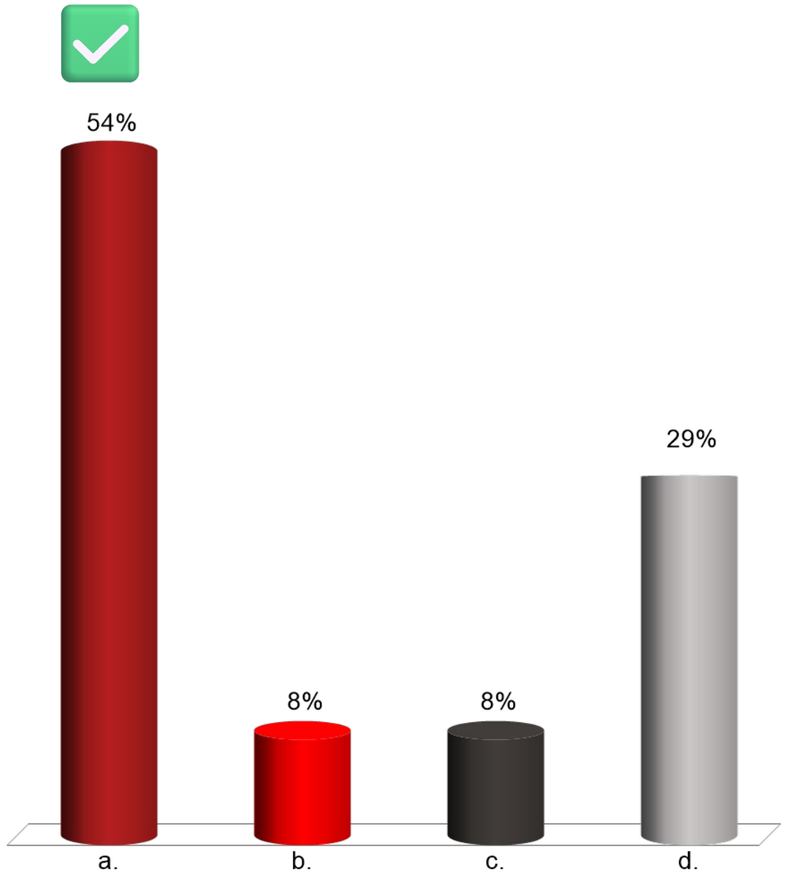


Question 2

URL: ttpoll.eu
Session ID: cs290

The CEO of a tech company stated in the media: “In the past, we've invested in technology to positively impact people's lives, and we have no intention of changing that strategy in the future - technology remains the best alternative.” We may interpret this as:

a. Sunk cost fallacy
b. Source cues
c. System 1 thinking
d. Illusory truth

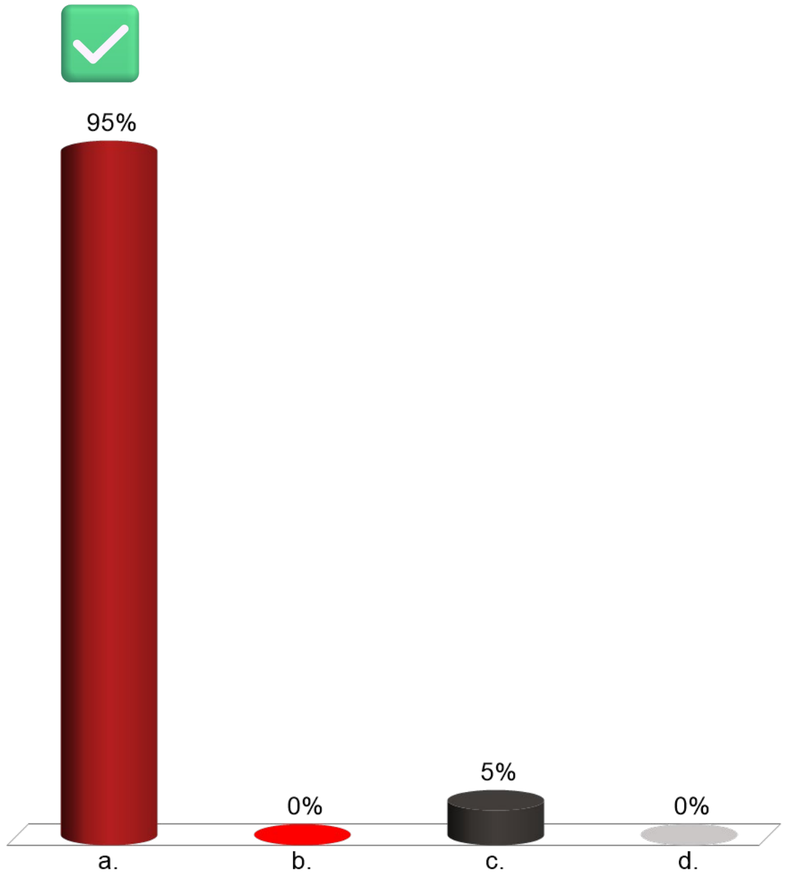


Question 3

URL: ttpoll.eu
Session ID: cs290

A start-up developed a machine learning model designed to connect people based on their personal interests. A big company has then bought the start-up and is currently using the algorithm to connect jobseekers with employers.
It is a case of …

a. … deployment bias
b. … aggregation bias
c. … measurement bias
d. … intersectional bias



Question 8

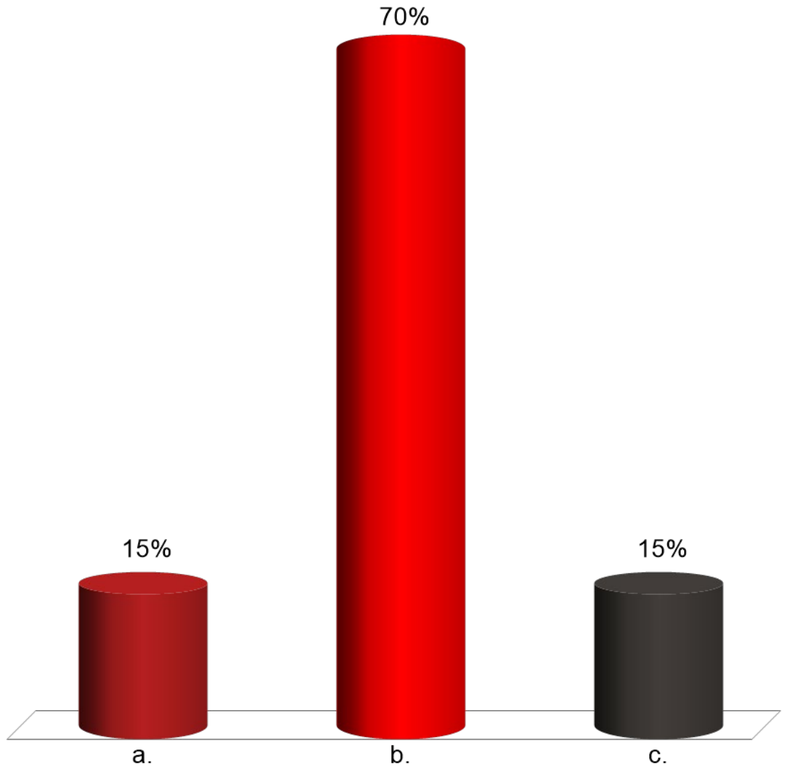
URL: ttpoll.eu
Session ID: cs290

You work on a chatbot to provide students assistance on campus questions. At evaluation time it generates plausible nonsense with a 15% rate. You probably face what we called an…

a. ... ethical issue ✅
b. ... ethical dilemma
c. ... ethical blindness
d. ... ethical sensitivity

You had most difficulty with:

a. The terms used in the question
b. The terms used in the options
c. Other



Question 10

URL: ttpoll.eu
Session ID: cs290

Fill the blanks:
If a piece of software behaves in a ___ way at first glance, but puts people of ____ at ____ , then it is a case of ____ discrimination.

0% a. negative / several groups / an advantage / direct

23% b. positive / identified groups / an advantage / inverse

✅ 73% c. neutral / specific groups / a disadvantage / indirect

5% d. neutral / several groups / a disadvantage / direct

0% e. negative / specific groups / a disadvantage / indirect

Question 11

URL: ttpoll.eu
Session ID: cs290

A bad actor launched a phishing attack on employees of Swiss public institutions to steal their login credentials. An online media outlet reported on it, with the most upvoted comments on the article criticizing the institutions for their inability to counter online threats, harming their reputation.
The harm to reputation can be classified as an impact that is:

8% a. Direct

25% b. Both direct and indirect

0% c. Neither direct nor indirect

✅ 67% d. Indirect

In the context of misuse/abuse:
- Direct impact = result from attack
- Indirect impact = secondary effect

=> Here the question is about the “harm to reputation” so indirect

Question 13

URL: ttpoll.eu
Session ID: cs290

You found a dataset with 5 variables, all self-reported by participants: eye-color, extraversion and 3 health-related variables. When analyzing the data you identify that:
- there are positive and substantial correlations among the 3 health variables
- there is a positive and substantial correlation between eye-color and extraversion
- there is no correlation between eye-color and the health variables

Eye-color is a latent variable:

Latent = not directly observable

10% a. True

✅ 90% b. False

Question 16

URL: ttpoll.eu
Session ID: cs290

You found a dataset with 5 variables, all self-reported by participants: eye-color, extraversion and 3 health-related variables. When analyzing the data you identify that:
- there are positive and substantial correlations among the 3 health variables
- there is a positive and substantial correlation between eye-color and extraversion
- there is no correlation between eye-color and the health variables

Extraversion is a latent variable:

Question 14

URL: ttpoll.eu
Session ID: cs290

You found a dataset with 5 variables, all self-reported by participants: eye-color, extraversion and 3 health-related variables. When analyzing the data you identify that:
- there are positive and substantial correlations among the 3 health variables
- there is a positive and substantial correlation between eye-color and extraversion
- there is no correlation between eye-color and the health variables

Eye-color is a proxy for health:

No correlation between eye-color and the health variables

9% a. True

✅ 91% b. False

Question 15

URL: ttpoll.eu
Session ID: cs290

You found a dataset with 5 variables, all self-reported by participants: eye-color, extraversion and 3 health-related variables. When analyzing the data you identify that:
- there are positive and substantial correlations among the 3 health variables
- there is a positive and substantial correlation between eye-color and extraversion
- there is no correlation between eye-color and the health variables

Eye-color is a sensitive attribute:

✅ 89% a. True

11% b. False

Considered sensitive:
- May lead to identification e.g., if combined with other attributes
- May lead to indirect discrimination

Your feedback on case studies

Question 18 (Case 2 Value Analysis):

- We underestimated points relatively to the length of the answers,
  5 points were probably not enough,
  7 or 8 points would have been more appropriate
- Lack of clarity on the instructions regarding stakeholders vs. values
- Lack of clarity on the format of the answer (table or text)

Question 19 (Case 3 UDIP):

- We overestimated points, 10 points were probably too much,
  8 points would have been more appropriate

All case studies questions received more than 15% vote for debriefing!

Question 17 / Case 1: Harm modeling

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 detriment	**E)**

See posts on SpeakUp

Post your ideas:
https://speakup.epfl.ch
Room key: **80844**

Question 17 / Case 1: Harm modeling

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 detriment	**E)**

See posts on SpeakUp

Post your ideas:
https://speakup.epfl.ch
Room key: **38348**

Question 17 / Case 1: Harm modeling

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 detriment	**E)**

See posts on SpeakUp

Post your ideas:
https://speakup.epfl.ch
Room key: **96072**

Question 17 / Case 1: Harm modeling

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 detriment	**E)**

See posts on SpeakUp

Post your ideas:
https://speakup.epfl.ch
Room key: **88201**

Question 18 / Case 2: Value analysis

Clarification of the instructions:

Goal = identify **2 value-based benefits** and **2 value-based harms**

For that:

- You needed to identify some stakeholders (>= 1)
- Then fill the table or report the content with letters

  👉 result in a table with 4 lines or 4 times the text for A+B+C+D

Stakeholder	Key Value	Benefits	Harms	Justification
Stakeholder: **(A)**	Value name and description: **(B)**	Benefit or Harm: **(C)**		It's a value-based benefit/harm for this stakeholder because: **(D)**

Question 18 / Case 2: Value analysis

Which **stakeholders** did you identify in the case?

- 1 post / stakeholder
- briefly describe their role / type (a few words)

See posts on SpeakUp

If your stakeholder has already been posted vote for it: 👍

Post your ideas:
https://speakup.epfl.ch
Room key: **62453**

Question 18 / Case 2: Value analysis

Which **value-based benefits** did you identify?

- 1 post / value
- name the value
- briefly describe why it's a benefit (1-2 sentences)

See posts on SpeakUp

If your benefit has already been posted
vote for it: 👍

Post your ideas:
https://speakup.epfl.ch
Room key: **31719**

Question 18 / Case 2: Value analysis

Which **value-based harms** did you identify?

- 1 post / value
- name the value
- briefly describe why it's a harm (1-2 sentences)

If your harm has already been posted
vote for it: 👍

See posts on SpeakUp

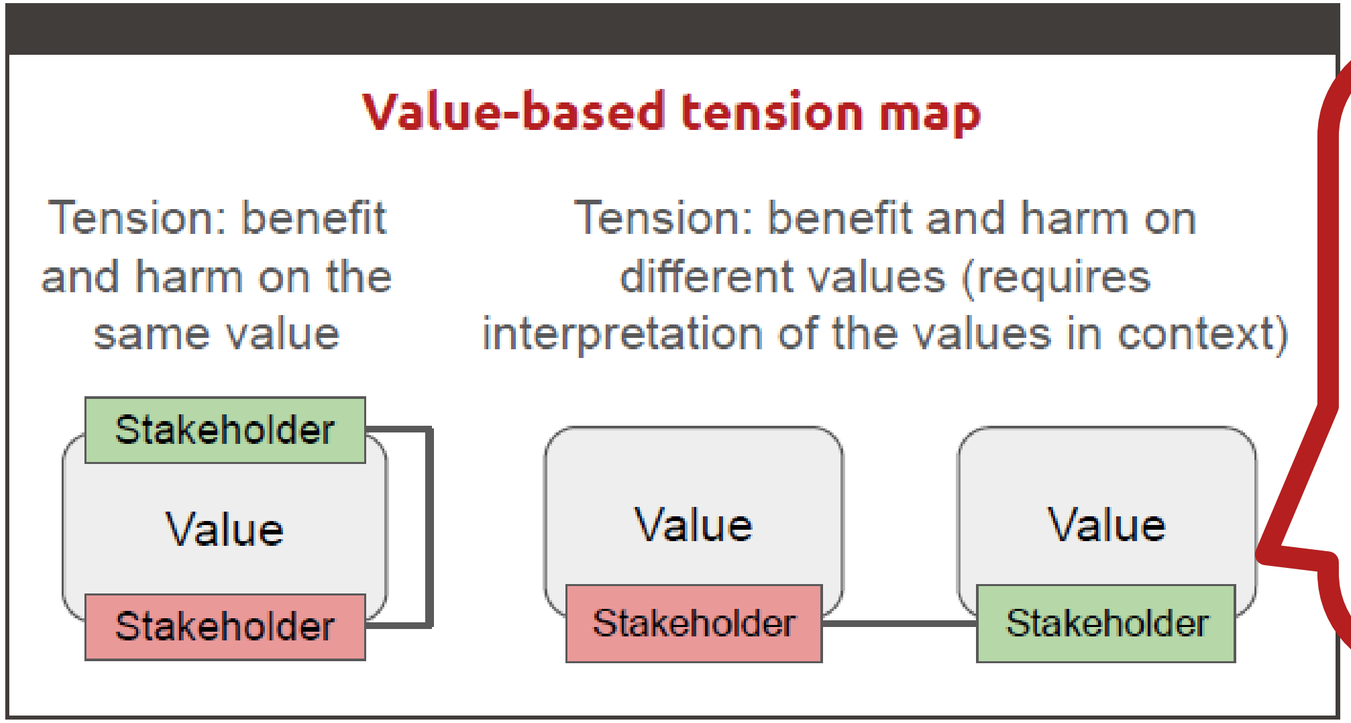
Post your ideas:
https://speakup.epfl.ch
Room key: **83726**

Question 18 / Case 2: Value analysis

Value based tension map:

Value tensions

Situations that **oppose** a value-based **benefit** with a value-based **harm**, either **between stakeholders** or **for a single stakeholder**.



⚠️ **Colors at the exam:** replace with mentions “benefit:” and “harm:”

- Map with the values that you presented in the table/text
- Explanation of 1 tension

Question 19 / Case 3: Strategies

As the ethics referee of the team, you are asked to anticipate potential consequences of the deployment of the platform in terms of safety and fairness.
Name one strategy seen in the course that you can apply for this task.

👉 **You need to have in mind (or on your A4 paper) the different strategies we have seen in the course and identify situations where they are adapted**

Question 19 / Case 3: Strategies

List the strategies we have seen in class
(try without looking at your notes!)

- 1 post / strategy
- name the strategy

If a strategy has already been posted
vote for it: 👍

Post your ideas:
https://speakup.epfl.ch
Room key: **51619**

Question 19 / Case 3: Strategies

Here are **4 categories of strategies**
👉 for each category list all the strategies that correspond (duplicates possible):

- Investigating stakeholders
- Eliciting values
- Anticipating impacts
- Thinking in systems

Strategies by category

Category	Strategies
Investigating stakeholders	• Stakeholder analysis • Bad actors • People behind the data • Ethics Canvas
Eliciting values	• Analyzing values
Anticipating impacts	• Ethical speculation • Bad actors modeling • STRIDE • Harm modeling • Edge cases • Causal loop diagrams • Inclusive design • Analyzing values • People behind the data • Datasheets for datasets • Ethics Canvas
Thinking in systems	• Causal loop diagrams
Making decisions	*NOT SEEN YET!*

Question 19 / Case 3: Strategies

As the ethics referee of the team, you are asked to **anticipate potential consequences** of the deployment of the platform in terms of safety and fairness.
Name one strategy seen in the course that you can apply for this task.

Category	Strategies
Anticipating impacts	• Ethical speculation • Bad actors modeling • STRIDE • ~~Harm modeling~~ • Edge cases • Causal loop diagrams • Inclusive design • Analyzing values • People behind the data • Datasheets for datasets • Ethics Canvas

Question 19 / Case 3: Strategies

As the ethics referee of the team, you are asked to **anticipate potential consequences** of the deployment of the platform in terms of safety and fairness.

Explain the strategy:
(a) Justify why this strategy is appropriate for this task
(2 sentences)
(b) Describe briefly how to apply this strategy:

- Goal
- When to use it
- Steps / phases / questions

Question 19 / Case 3: Strategies

Which **safety issue** did you identify?

- 1 post / safety issue
- briefly describe the issue

If your safety issue has already been posted vote for it: 👍

Post your ideas:
https://speakup.epfl.ch
Room key: **20158**

Question 19 / Case 3: Strategies

Which **fairness issue** did you identify?

- 1 post / fairness issue
- briefly describe the issue

If your fairness issue has already been posted vote for it: 👍

Post your ideas:
https://speakup.epfl.ch
Room key: **00774**

Tips for effective revisions

Revision techniques

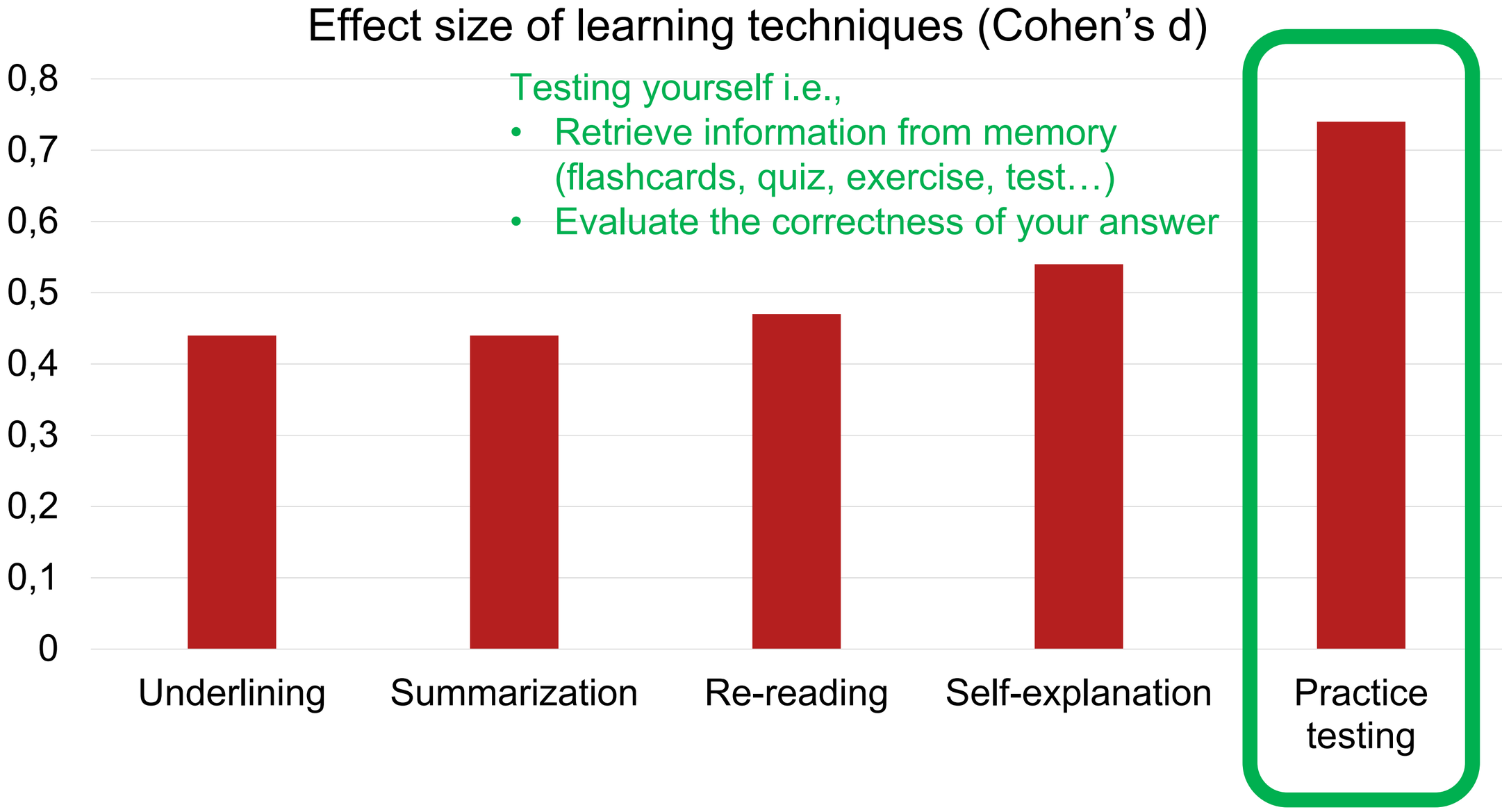
URL: ttpoll.eu
Session ID: cs290

Select the technique(s) that you generally use for your revisions:

a. Re-read notes/course documents
b. Highlight important points in notes/course documents
c. Create summaries from notes and/or course documents
d. Create summaries from memory (without looking at notes)
e. Use flashcards / self-made questions
f. Re-read the solutions of exercises / case studies
g. Re-do exercises / case studies then check the solutions
h. Other

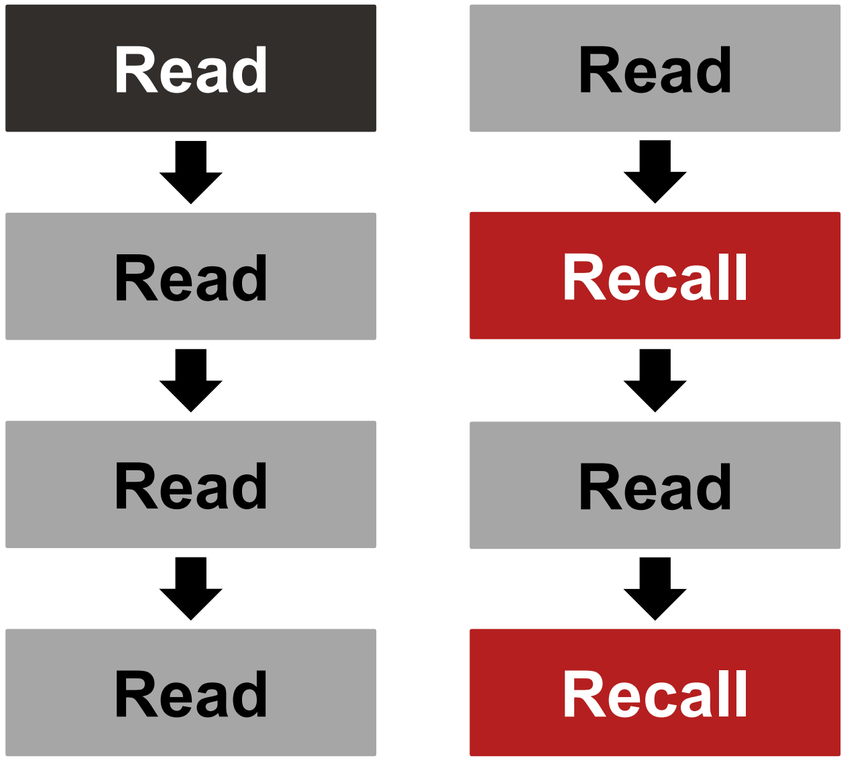
Learning techniques

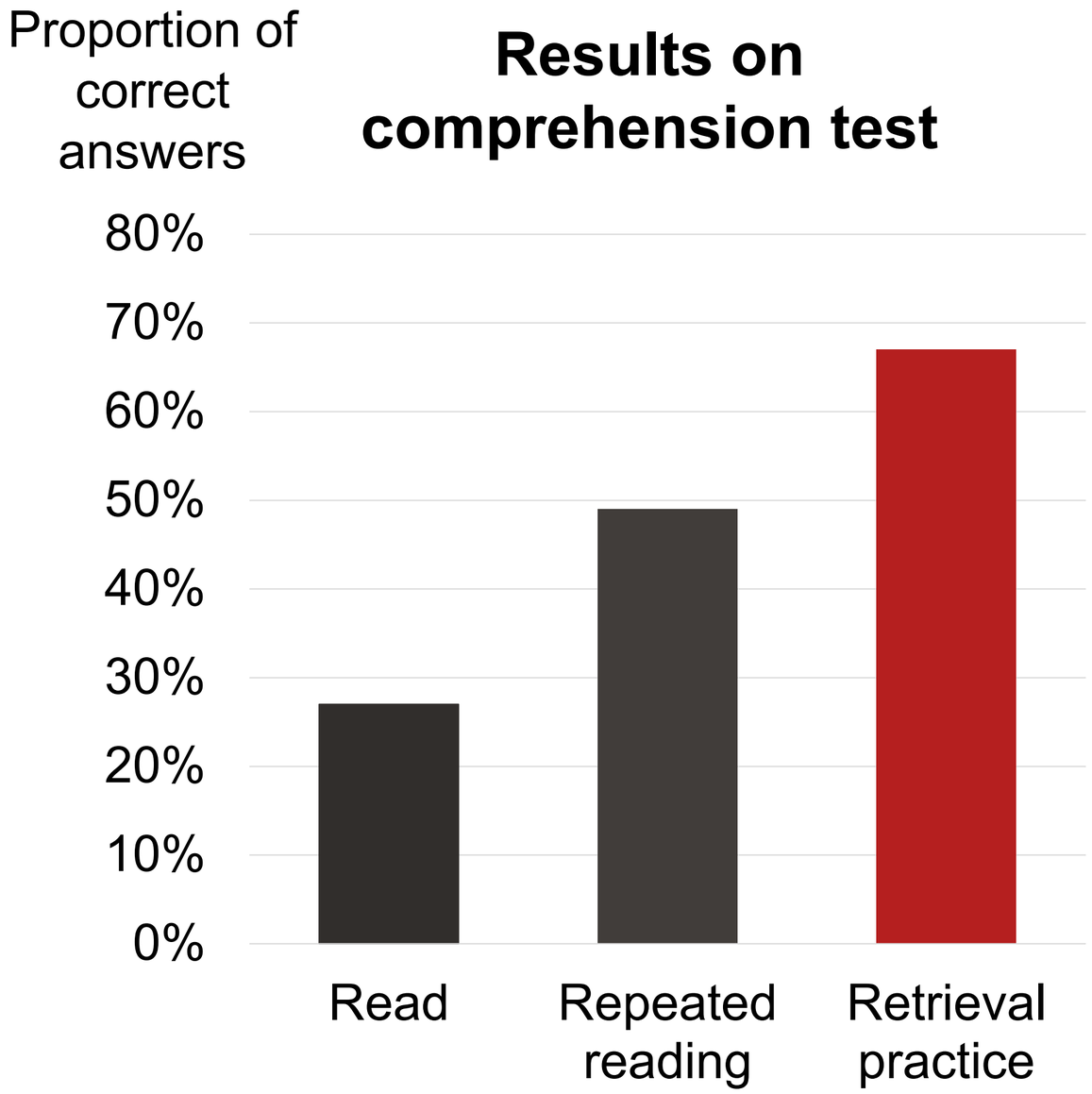
(Donoghue & Hattie, 2021)



Retrieval practice

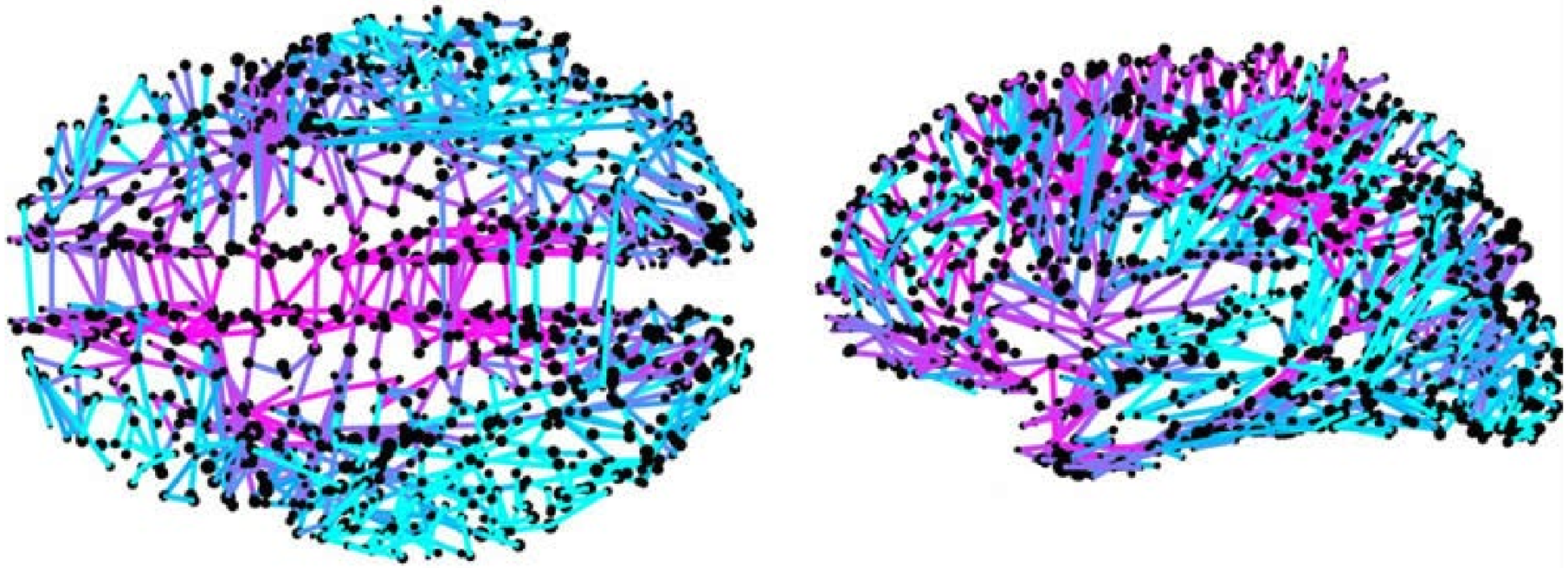
(Karpicke and Blunt, 2011)





Recall = *reconstruct*

(Mišić et al., 2015)



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

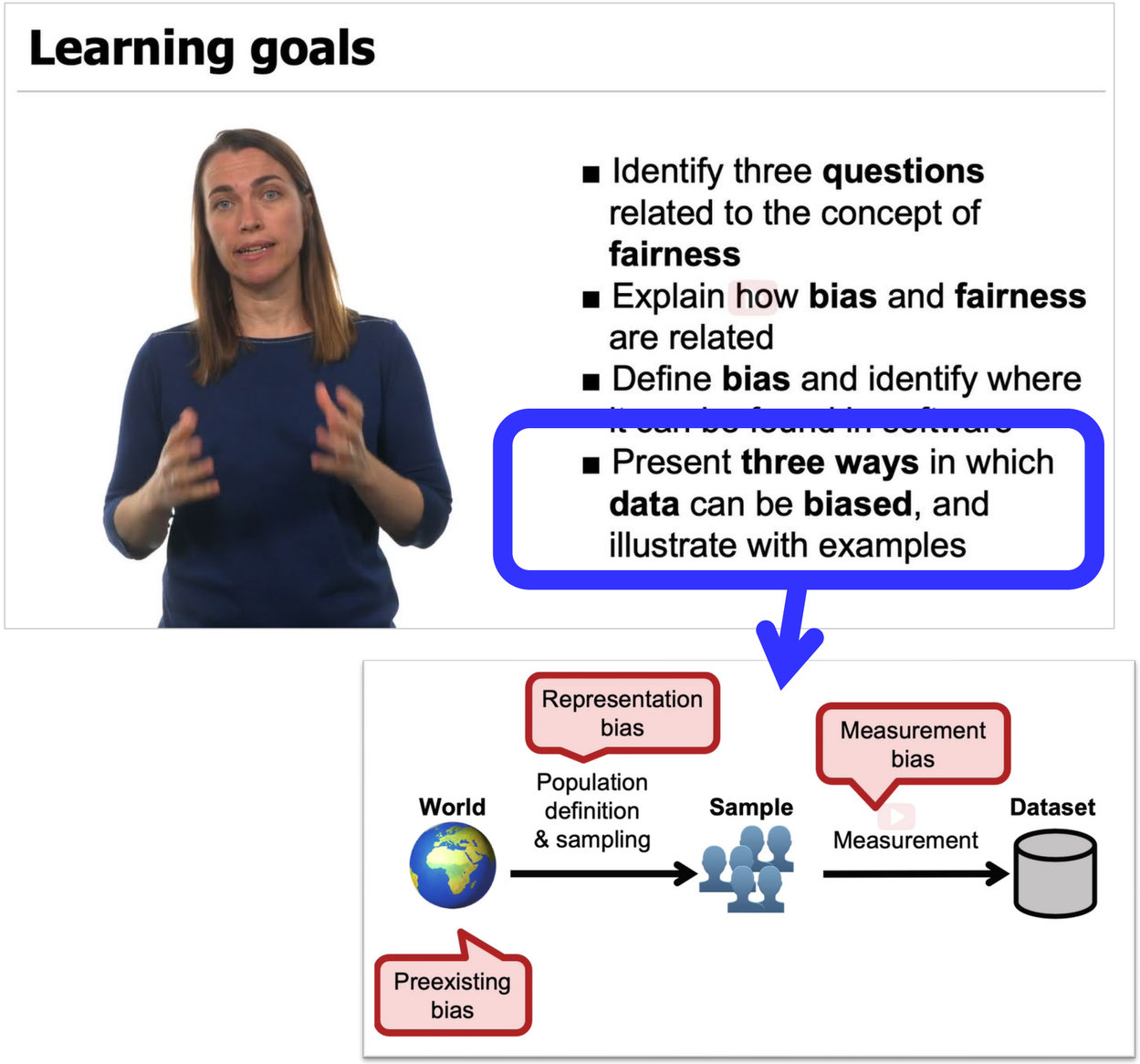
Information is not “stored” into memory like in a “drawer”, instead it is *reconstructed* from a network of connections

=> You should practice *reconstruction*

Flashcards

Use the **learning goals** at the beginning of the videos to create **flashcards**

- 1 goal = 1 card (or more)
- Recto = 1 goal / question
- Verso = answer



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

Other suggestions

- **Re-do the case studies without the “proposed answers”**
  Then check with the “proposed answers”
- **Re-do the blank test without the “solution”**
  + **time yourself**: practice strategy and speed!
- Reuse the quizzes done in class
  ⚠️ Watchout they are not designed like exam questions since they can have multiple correct answers
  -> you can transform them into single choice questions
- Prepare your allowed A4 sheet of notes for the exam
  (pay attention to structure + synthetization)

What’s next?

We start Sustainability 1!

Tomorrow, Tuesday 5: notebook on the carbon footprint of algorithms

+ I will be **available for answering individual questions regarding the Graded Assignment 1**

👉 from **8h15 to 10h in CM 1 111**

By Monday 11:

- Watch **videos 5.1 to 5.4** + do the **quizzes**
- Finish the notebook
  (and any other leftover from previous weeks)