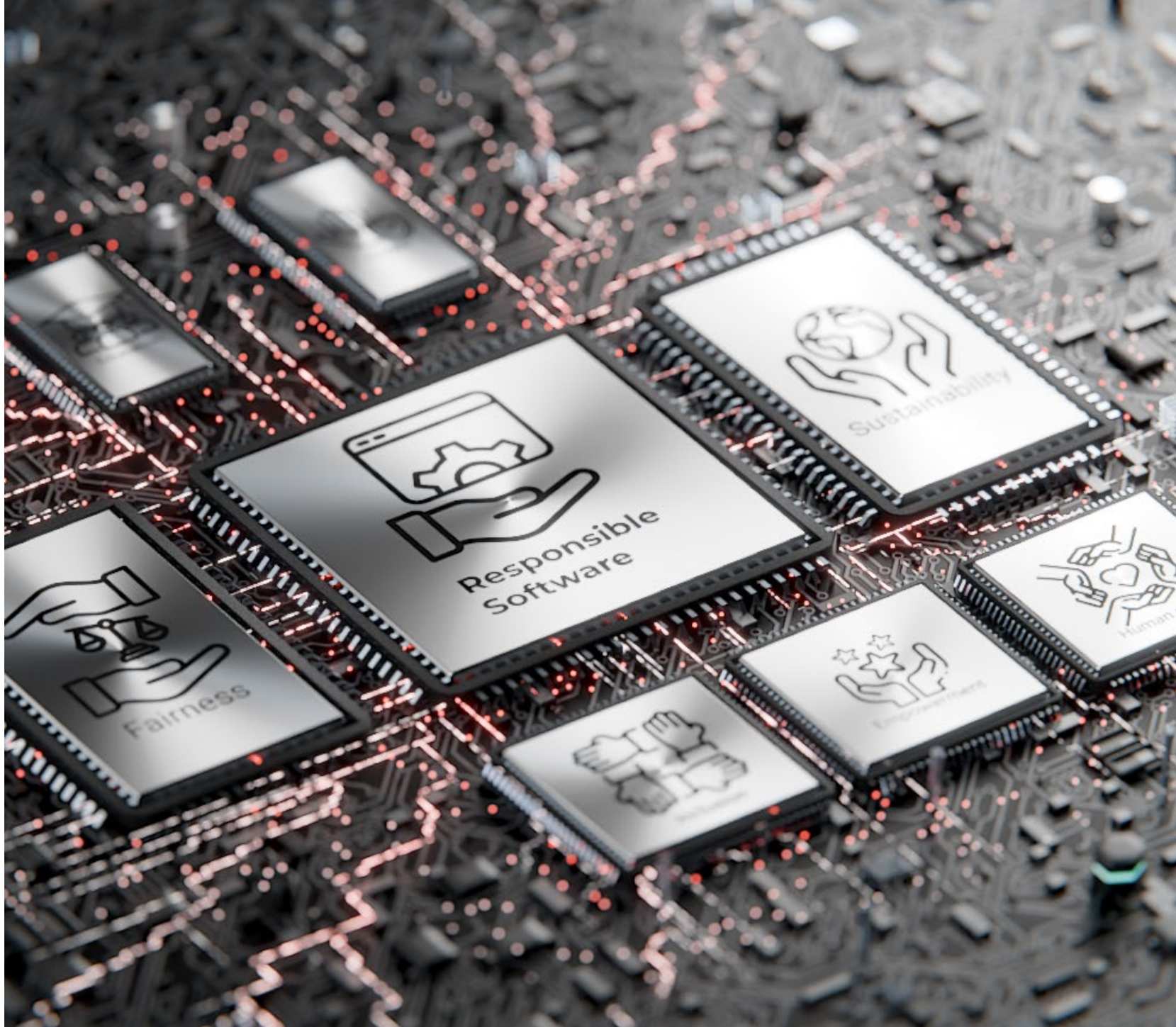
EPFL

Safety 2 Review & Case studies

30 sept.

Cécile Hardebolle

Responsible Software



Agenda for today

1. Interactive review questions on Safety 2
   (and some other topics)
2. Case studies:
   a) Edge cases
   b) The Ethics Canvas
   c) Systems Thinking (Causal Loop Diagrams)

Exam - precisions

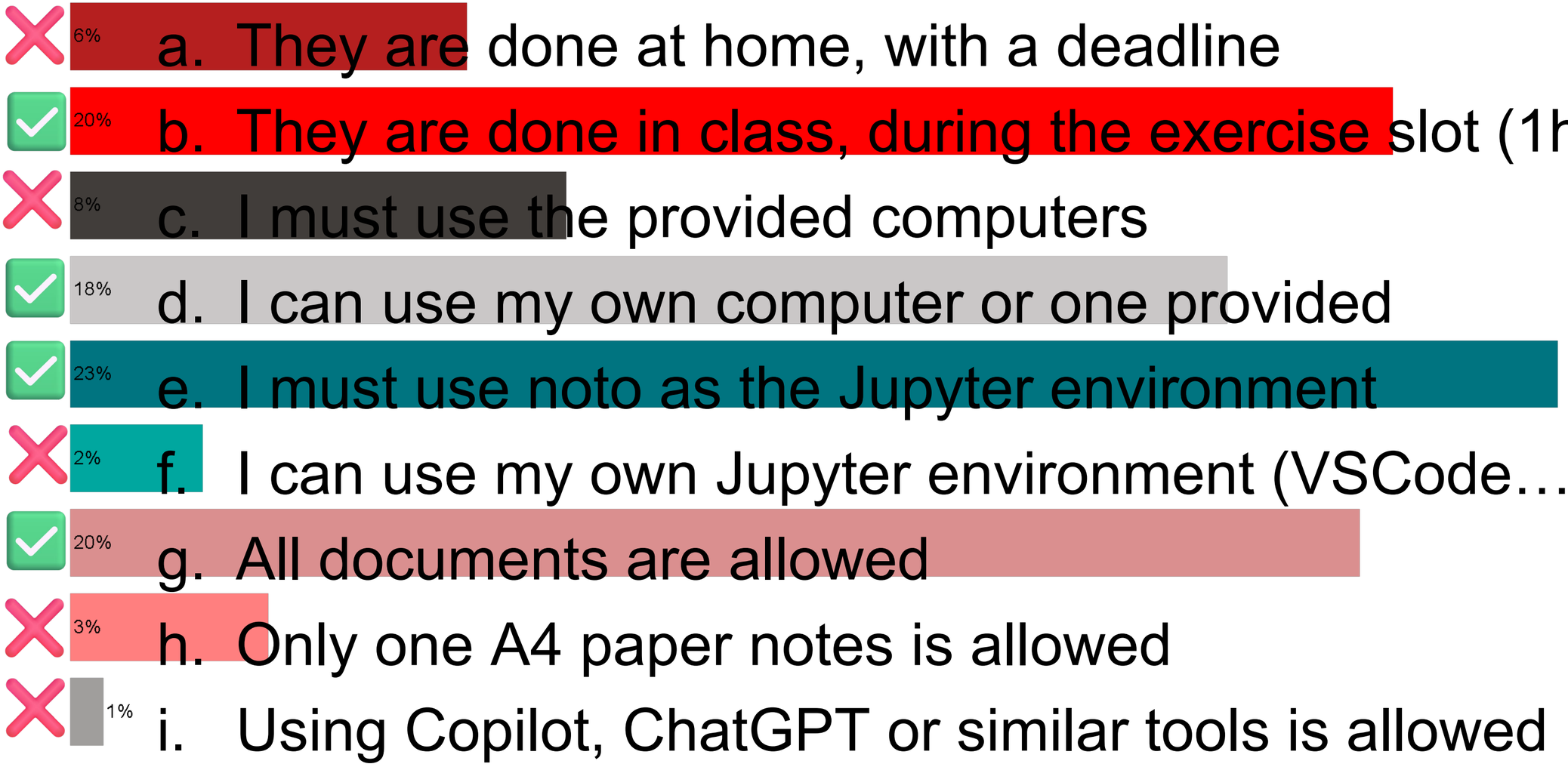
- You can have 1 sheet of paper with notes with you:
  - A4
  - Recto-verso
  - Handwritten or typed (open format)

- Blank test on October 29, 8h15-10h

Graded assignments: format

URL: ttpoll.eu
Session ID: cs290

Select **all the correct** statements about the **graded assignments:**



Graded assignments: grading

URL: ttpoll.eu
Session ID: cs290

Select all the **only correct** statement about the **graded assignments:**

❌ 8% a. Only the programming questions are graded

❌ 8% b. Only the ethical reflection questions are graded

✅ 85% c. Both are graded

Programming exercises

URL: ttpoll.eu
Session ID: cs290

We currently have very very very low attendance in the programming exercise sessions… So we are wondering **how to organize to best help you with the programming exercises.**

Select all that apply to you:

41% a. I am fine with the organization as it is

23% b. The time slot does not work for me

0% c. The rooms don't work for me

35% d. I don't really need help

0% e. Other

Review questions Safety 2

Macro-level perspective

URL: ttpoll.eu
Session ID: cs290

A macro-level perspective is useful (select all correct statements):

- ✅ 27% a. When software is under design
- ✅ 15% b. After software is deployed 👉Should definitely be done before, but after ok
- ✅ 14% c. After an analysis with a meso-level perspective 👉There is no order meso/macro, could be done before
- ✅ 24% d. When considering expanding to new countries
- ✅ 20% e. When software is used by public institutions 👉Depends on the type of software. True mainly if it is software that then has an impact on population (e.g. fraud detection for social assistance)

Disinformation

URL: ttpoll.eu
Session ID: cs290

A piece of information is false but created without intention to harm.
It is (select all that apply):

False beliefs

URL: ttpoll.eu
Session ID: cs290

If you are exposed to a dis/mis-information post by Melon Husk, you are more likely to believe it because of:

❌ 5% a. System 2

(✅) 8% b. False consensus 👈 If the info is “agreed with” largely in your social circle

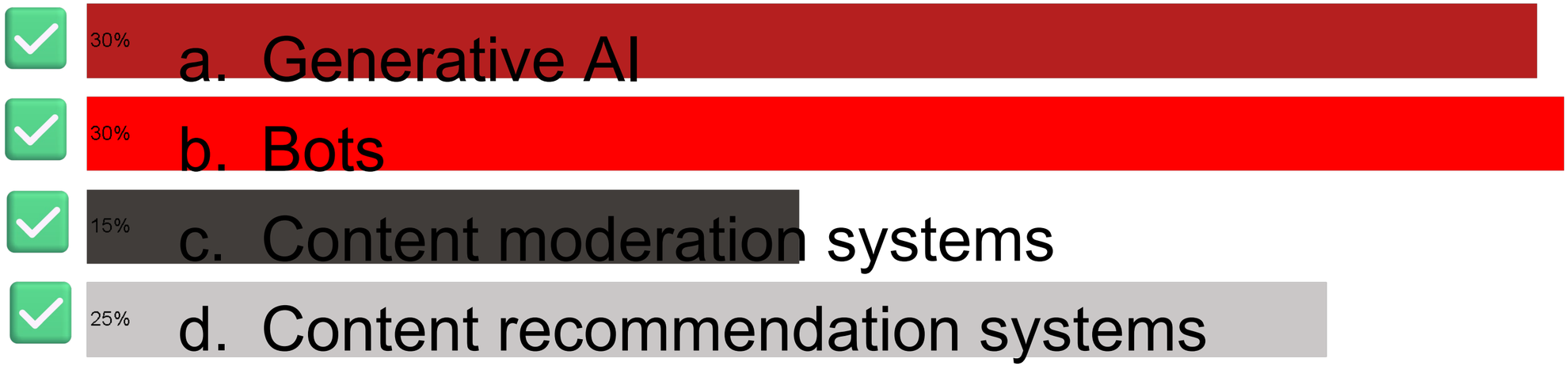
✅ 71% c. Source cues

(✅) 17% d. Illusory truth 👈 If the info is repeated a lot

Software & disinformation

URL: ttpoll.eu
Session ID: cs290

Software playing a role in disinformation can be (select all that apply):

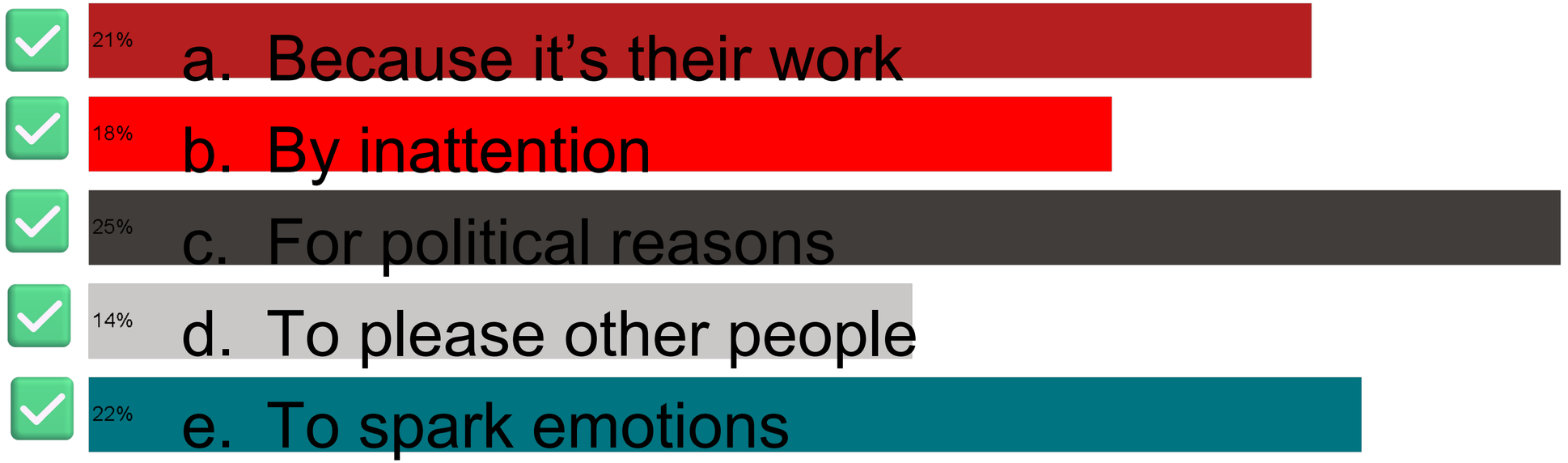


+ Other types of software (e.g. photo edition etc.)

Humans & disinformation

URL: ttpoll.eu
Session ID: cs290

Humans playing a role in disinformation do it (select all that apply):



Humans play a role both as **producers** and as **receivers** (re-emitters)

Case studies

Where to find the cases?

1. Go to **moodle**

2. Find the **link to the case studies** for today: **Safety 2**
   👉this link will send you to courseware
   (where you can find all the course material)

3. Download:
   - The **instruction sheet**
   - The 3 cheatsheets:
     Edge Cases, Ethical Canvas, Causal Loop Diagrams

Edge cases

Edge cases

- Edge case = **problem** or **situation** that occurs only at an **extreme** (maximum or minimum) **operating parameter**
- Commonly used in software testing to account for boundary conditions
- We’re “hijacking” it to account for macro level effects 😉

Instructions

Individually, read the scenario, then apply Edge Case Analysis

👉 For each “edge case” category in the cheatsheet:

1. Identify the **issues it could cause** for the software
2. Describe which **change you could make** to improve it

Share with your neighbor:

- Did you identify the same issues?

Post your edge cases

Add an *anonymous* comment to the **thread “Post your edge cases” on Ed Discussion** with:

- The edge case category considered
- A short description of the **issue you have identified** (1line)
- A short description of the **design change you would make** (1 line)

⚠️ If your edge case has already been posted, add a “like” / heart

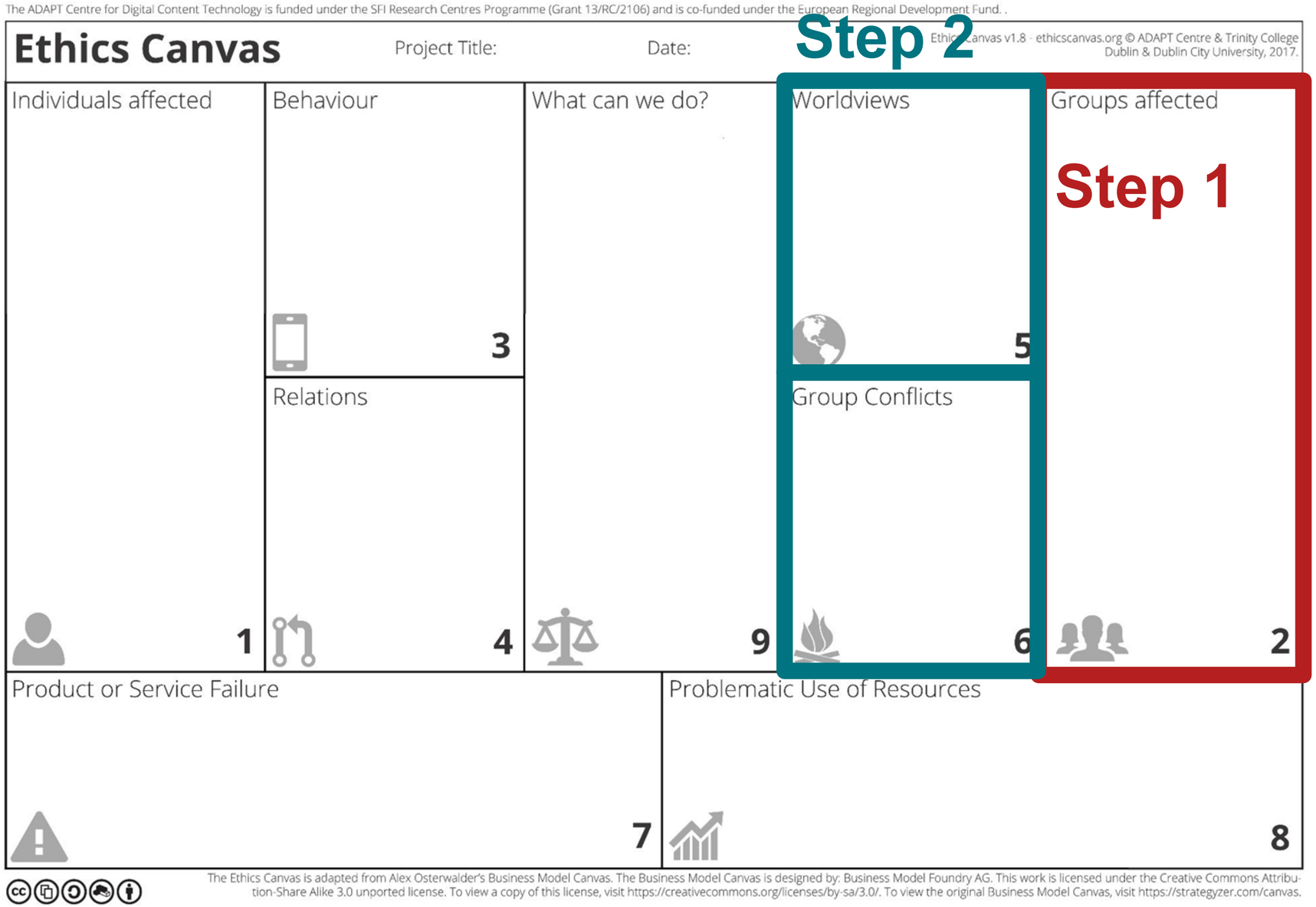
The Ethics Canvas

Instructions

Individually:

- Read the scenario
- Step 1:
  groups affected
- Step 2:
  - Worldviews
  - Group Conflicts

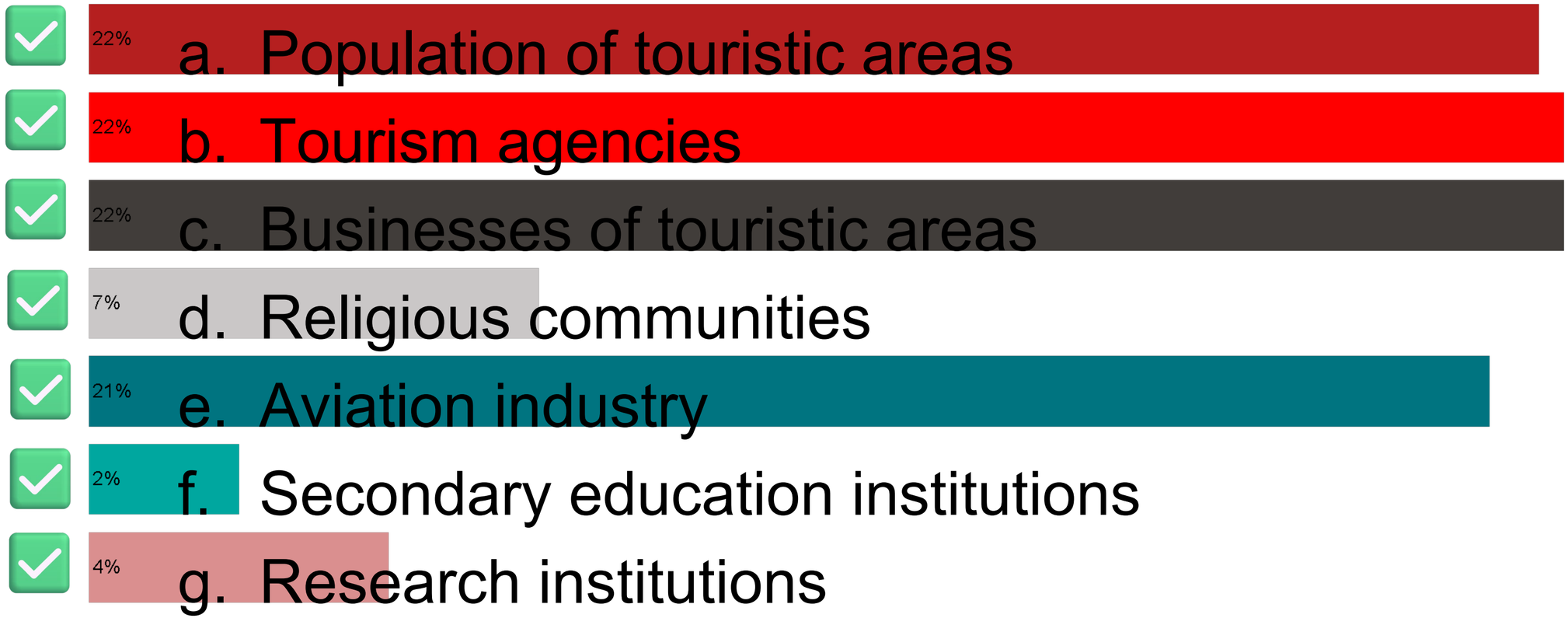
Compare with your neighbor!



Groups affected

URL: ttpoll.eu
Session ID: cs290

Which groups can be affected in the VR case (select all that apply)?



Impacts

URL: ttpoll.eu
Session ID: cs290

The types of impacts you found are:

12% a. Only negative

0% b. Only positive

✅ 88% c. Both negative and positive

Overall debriefing of the strategy

The Ethics Canvas is a general strategy that combines:

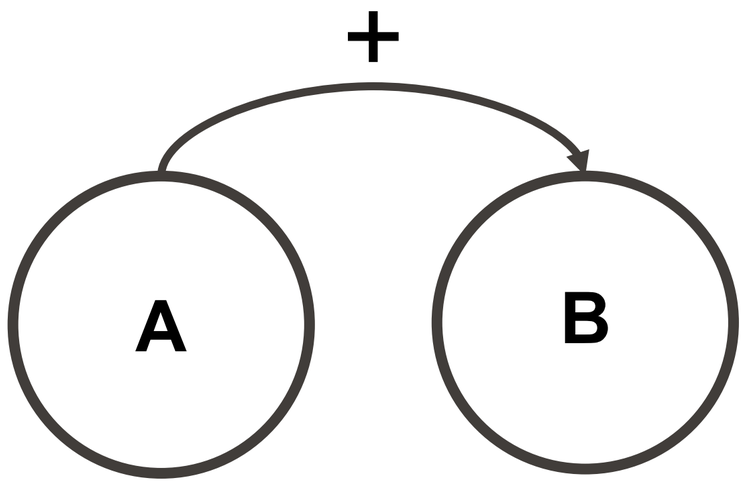
- User and stakeholder analysis
- Impact anticipation

It can help to think about impacts at the micro and **macro levels, both positive and negative**

Systems Thinking
(Causal Loop Diagrams)

Causal Loop Diagrams

URL: ttpoll.eu
Session ID: cs290



The arrow with label “+” means:

❌ 8% a. There’s a transition from state A to state B on token “+”

❌ 8% b. The quantity in A is added to the quantity in B

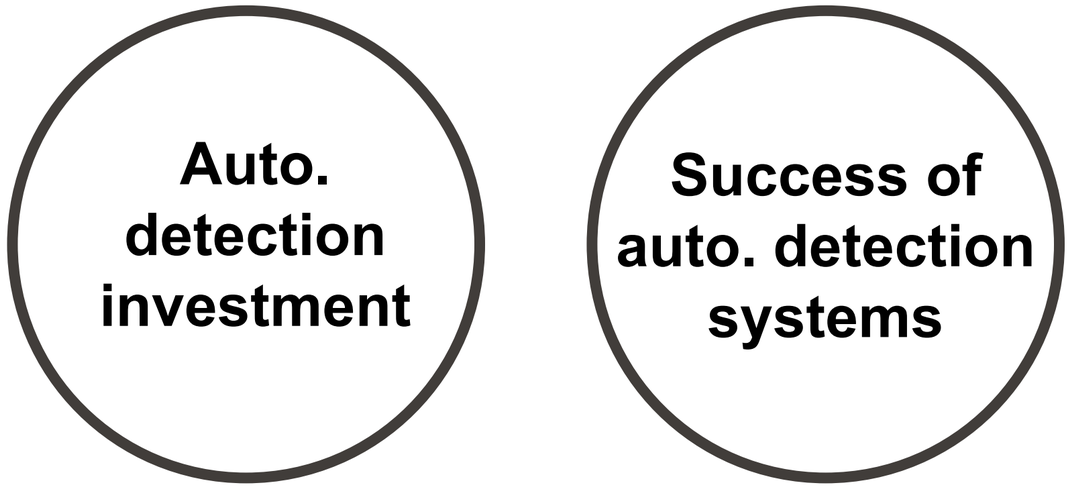
❌ 18% c. A and B change in a positive direction

✅ 65% d. B changes in the same direction as A

Causal Loop Diagrams - 1

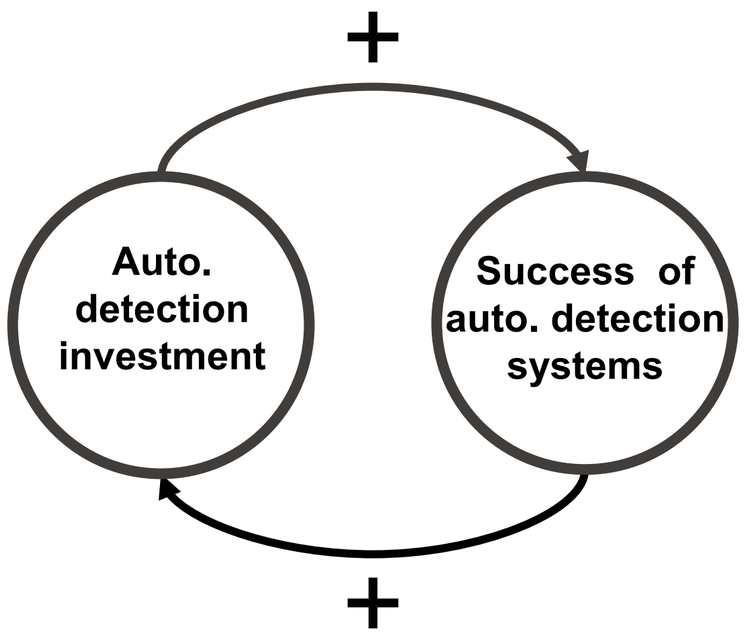
Key variables:

- Investment of resources in automatic detection systems to detect/filter fake news
- Success of automatic detection systems to detect/filter fake news



Causal Loop Diagrams – 1a

URL: ttpoll.eu
Session ID: cs290



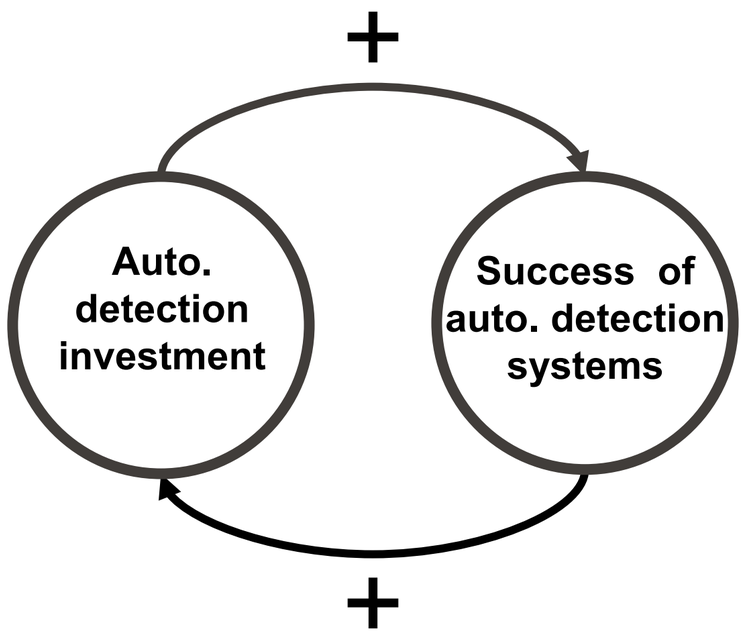
Over time, the quantities in this system will:

- ✗ 2% a. Stabilize
- ✗ 7% b. Increase
- ✗ 0% c. Decrease
- ✓ 91% d. It depends

Simulation: https://go.epfl.ch/cs290-cld-1

Causal Loop Diagrams – 1b

URL: ttpoll.eu
Session ID: cs290



The loop in this diagram is:

❌ 4% a. Balancing

✅ 96% b. Reinforcing

Instructions

Individually, read the context and scenario, then <u>work on Part 1</u>
(Part 2 can be used for revisions)

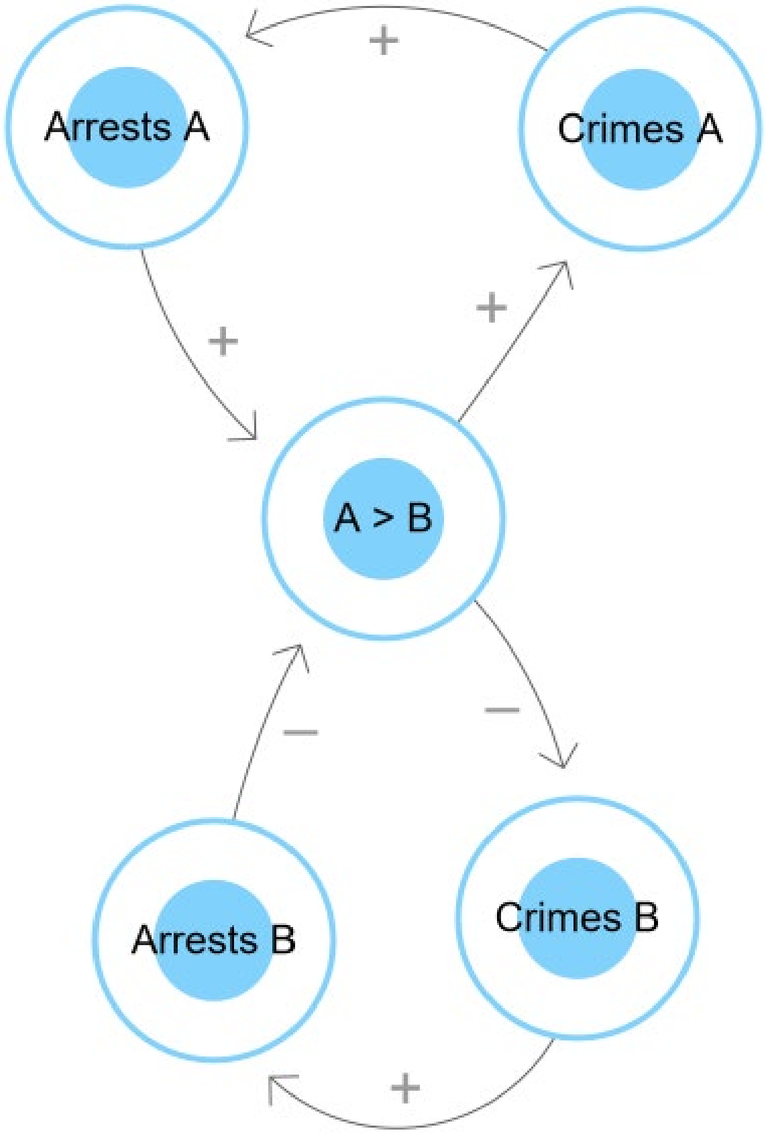
1. Describe with words the behavior of each loop
2. Indicate if the loops are balancing or reinforcing
3. Describe the consequences of using such a system

Compare with your neighbor:

- Do you agree on the types of loops?
- What consequences do you identify?

Predictive policing

URL: ttpoll.eu
Session ID: cs290



Over time, the quantities in this system will:

- ✗ 10% a. Stabilize
- ✗ 10% b. Increase
- ✗ 3% c. Decrease
- ✓ 77% d. It depends

LOOPY

Very useful online tool to simulate causal loop diagrams:
https://ncase.me/loopy/

Try it out!

What’s next?

We start Fairness 1!

Tomorrow, Tuesday 1: notebook on university admissions

By Monday 7:

- Watch **videos 3.1 to 3.5** + do the **quizzes**
- Finish the notebook
  (and any other leftover from previous weeks)