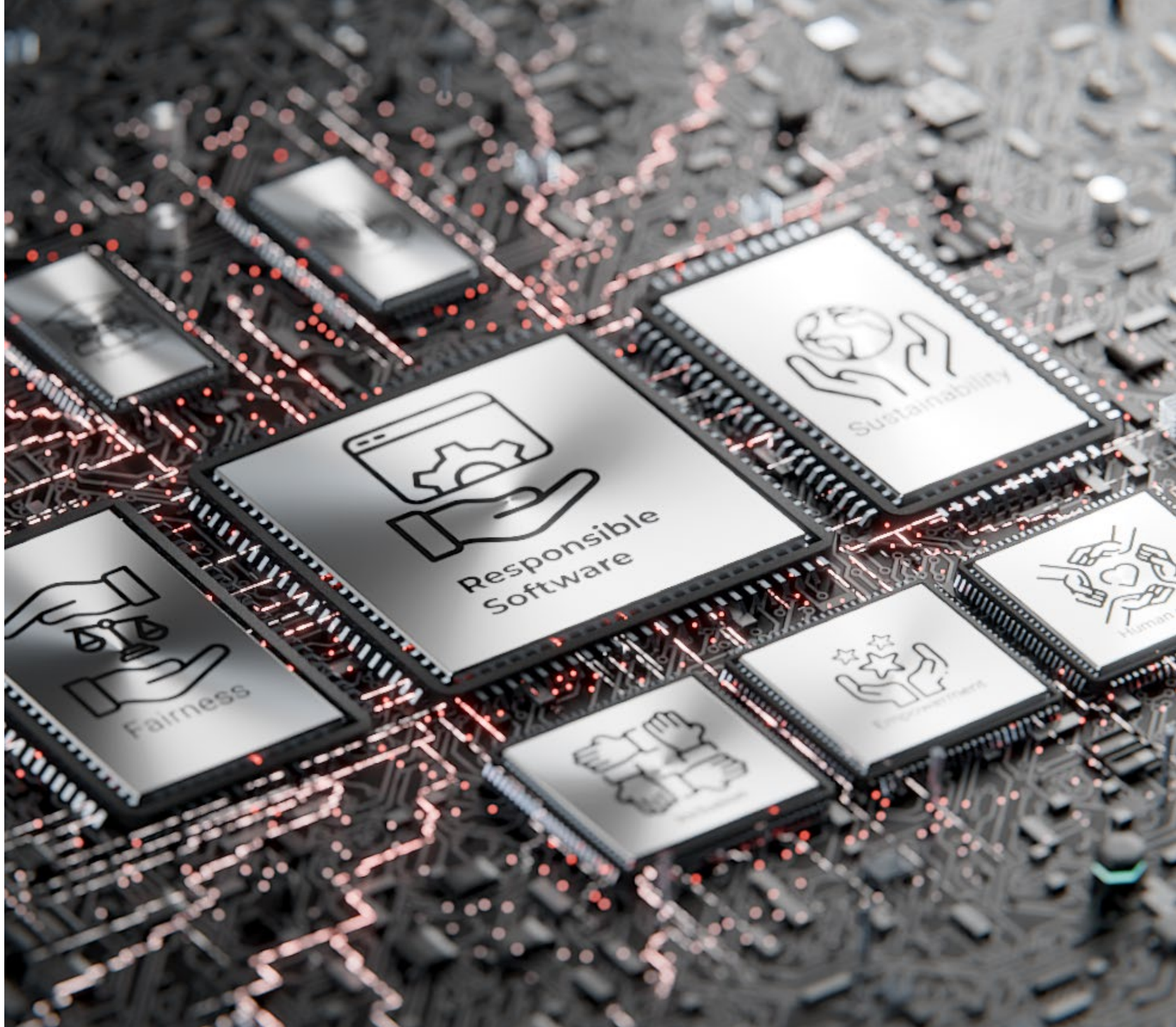
Safety 2 Review & Case studies

29 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)

Notebooks

URL: ttpoll.eu
Session ID: cs290

Where are you with your work on the notebooks?

Select the answer that best matches your situation:

66% a. I have completed all 3 notebooks (Intro, Safety 1, Safety2)

25% b. I have completed 2 notebooks

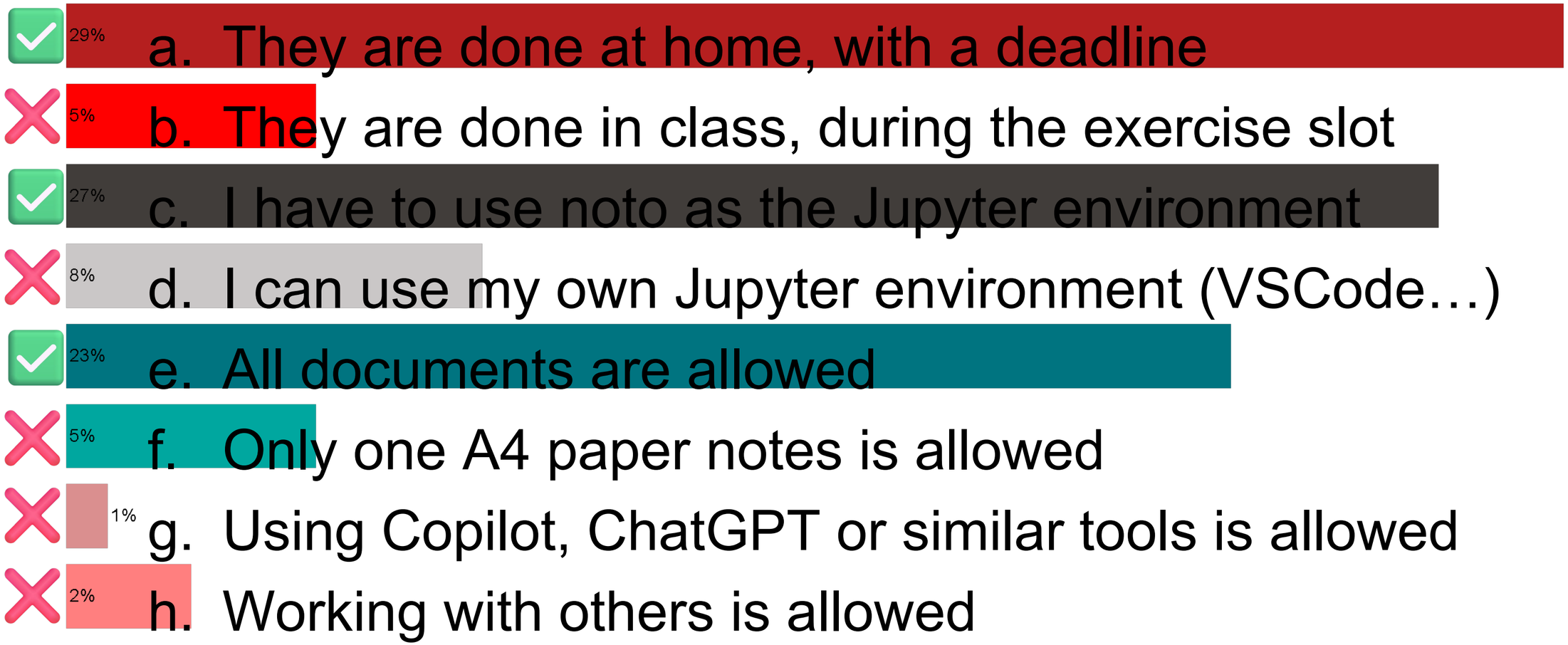
2% c. I have completed 1 notebook

7% d. I have not yet worked on the notebooks

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:**

- ✗ 16% a. Only the programming questions are graded
- ✗ 7% b. Only the ethical reflection questions are graded
- ✓ 77% c. Both are graded

Graded assignment 1

	Dates
Release	October 10
Submission	October 14 at 23h59

Grading
8% of total grade
Coding questions + reflection questions

Topics
Safety 1 & 2
Fairness 1 & 2

- Reminder of the rules:
  - No GenAI
  - No group work
  - Use noto
- Support session with assistants on Tuesday, October 14, 10h-12h

  👉 They do not have the solution!
  They can help you debug or help with tech issues
  They can help you submit your work on moodle

Review questions Safety 2

Macro-level perspective

URL: ttpoll.eu
Session ID: cs290

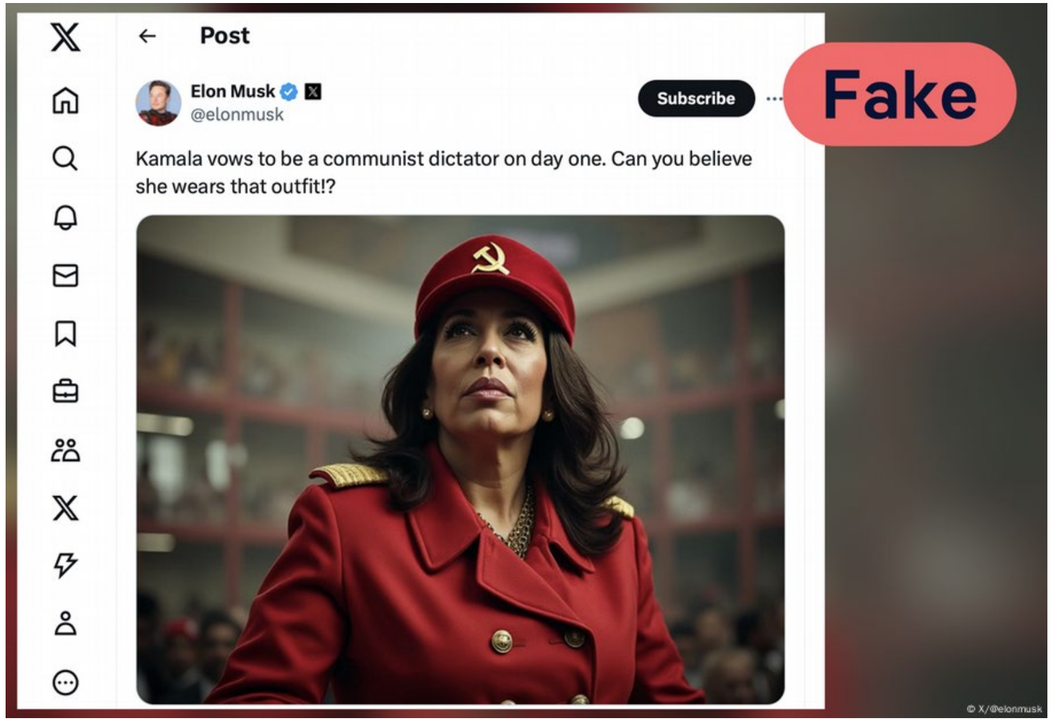
A macro-level perspective is useful (select all correct statements):

- ✅ 30% a. When software is under design
- ✅ 13% b. After software is deployed 👉Should definitely be done before, but after ok
- ✅ 12% 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
- ✅ 21% 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)

False beliefs

URL: ttpoll.eu
Session ID: cs290

One dis-/mis-information post by Elon Musk appears in your Twitter timeline.
You are more likely to believe it because of (choose one):



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

- ❌ 2% a. System 2
- ❌ 32% b. Illusory truth
- ✅ 66% c. Source cues
- ❌ 0% d. Prebunking

Exam type

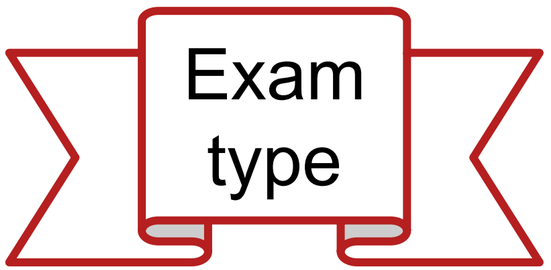
Dis/Mis-information

URL: ttpoll.eu
Session ID: cs290

Your friend tells you:
“Eating carrots will drastically improve your night vision.”
This is (choose one):

- ✅ 65% a. Misinformation
- ❌ 7% b. Disinformation
- ❌ 14% c. Malinformation
- ❌ 14% d. Fake news

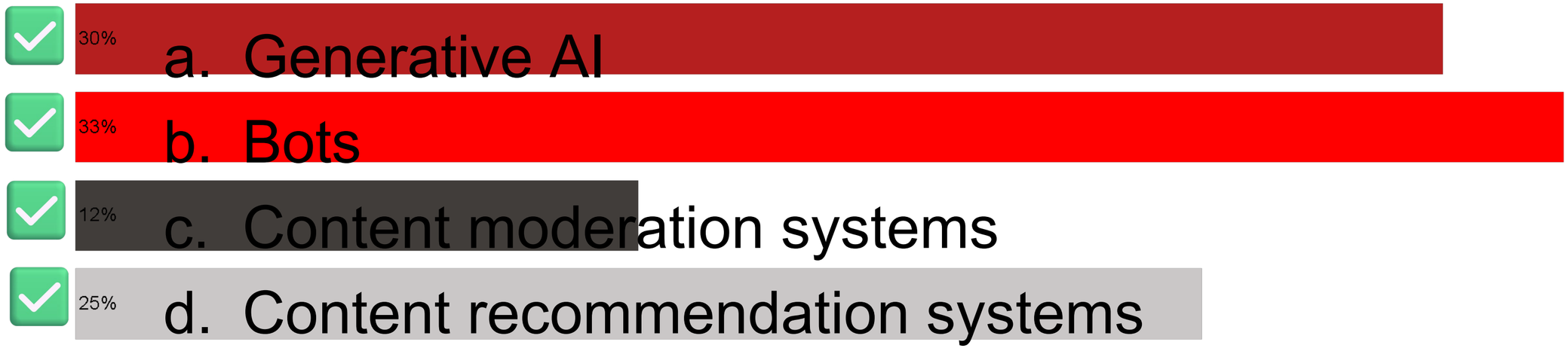
- False
- No intention to harm
- Not in the form of news (e.g. press article)



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, intentionally or not)

Case studies

Where to find the cases?

1. Go to **courseware**
2. Find **the case studies** for today: **Safety 2**
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 at an **extreme value** (e.g. maximum or minimum) of an **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
- Apply Edge Case Analysis: for each “edge case” category, identify the **issues it could cause** for the software
  - Global Reach
  - Mass Adoption
  - Longevity

Share with your neighbor:

- Did you identify the same issues?

Post your edge cases

Which **edge cases** did you identify?

👉 1 post / edge case

- Name of the **category**
- Brief **description**

⚠️ If your edge case has already been posted, add a vote 👍

There are great answers on SpeakUp, **have a look!**

Overall:

- Mention the category name (Global Reach / Mass Adoption / Longevity)
- Make sure the case described matches with the category chosen

Post your ideas:

https://speakup.epfl.ch

Room key: **78844**

Mitigation options

Choose 1 edge case:
How could you mitigate the consequences by making changes in the design of the app?

Do the same with all edge cases!

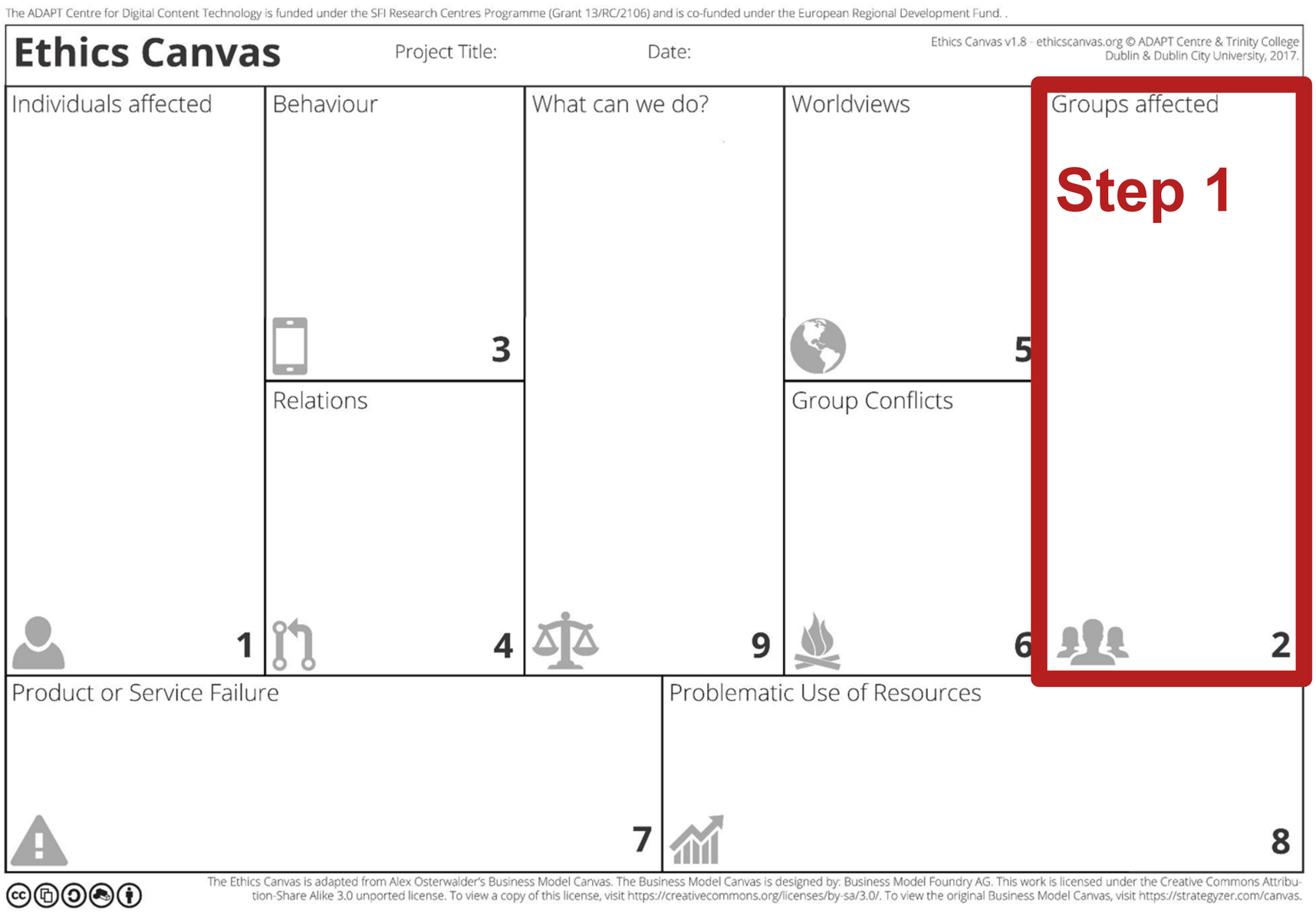
That’s how you can create better software, that make a real difference!

The Ethics Canvas

Instructions

Individually:

- Read the scenario
- Step 1:
  groups affected



Post your groups

Which **affected groups** did you identify?
👉 1 post / group

Check the answers on Speak Up + lots of other groups possible (cultural minorities, religious groups, etc.)

⚠️ If your group has already been posted, add a vote 👍

Post your ideas:
https://speakup.epfl.ch
Room key: **25756**

Debriefing of Step 1

One difficulty with this step of the canvas is to distinguish between:

- **Individuals affected (block 1)**
  - 👉 Consider types of individuals who are directly/indirectly affected
    [Micro level perspective!]

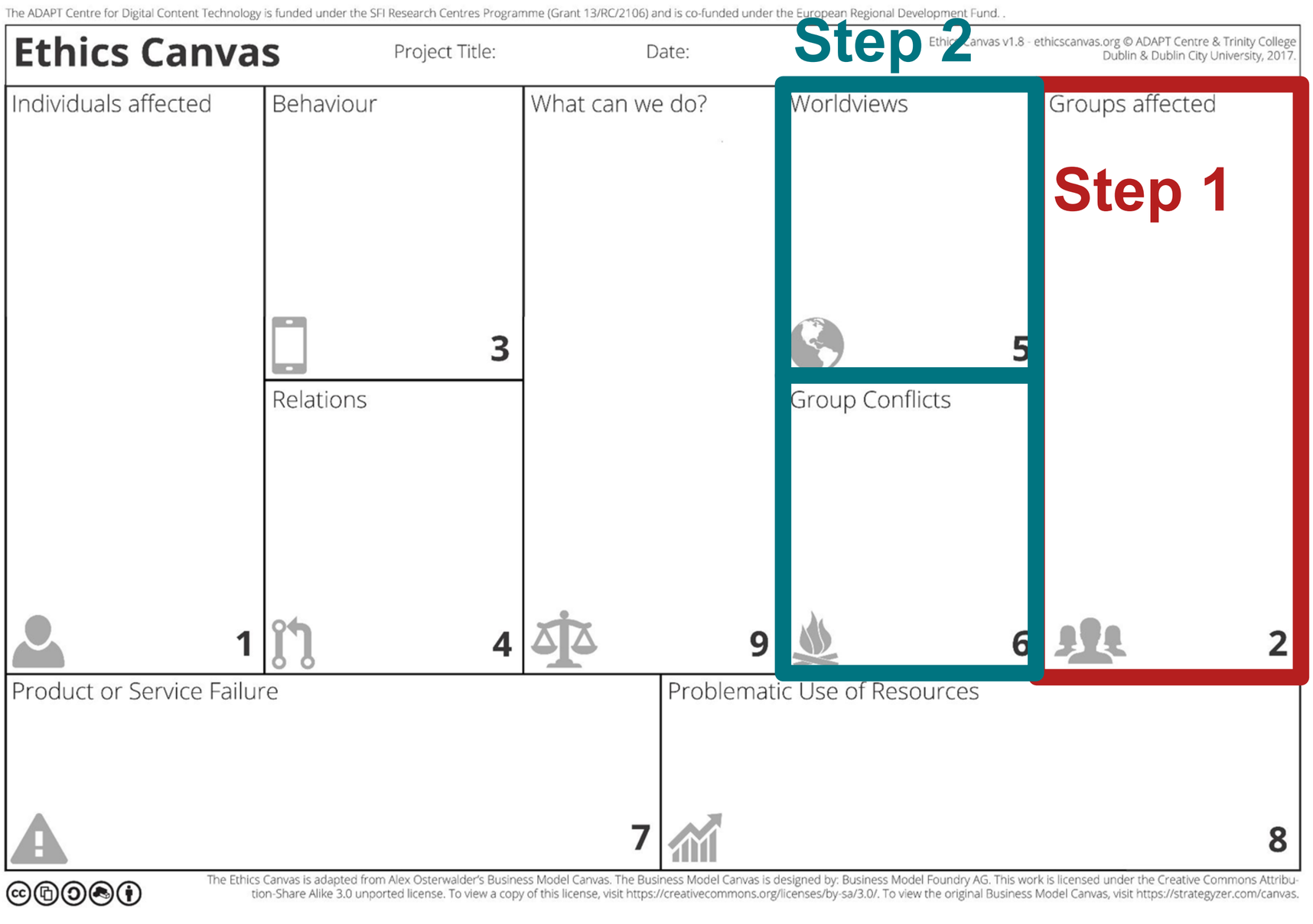
- **Groups affected (block 2)**
  - 👉 Consider communities, institutions, organizations…
    [Meso/macro level perspective!]

Instructions

Individually:

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

Compare with your neighbor!



Worldviews & group conflicts

URL: ttpoll.eu
Session ID: cs290

The types of impacts you found are:

27% a. Only negative

6% b. Only positive

✅ 67% c. Both negative and positive

Overall debriefing of the strategy

The Ethics Canvas is a general strategy that combines:

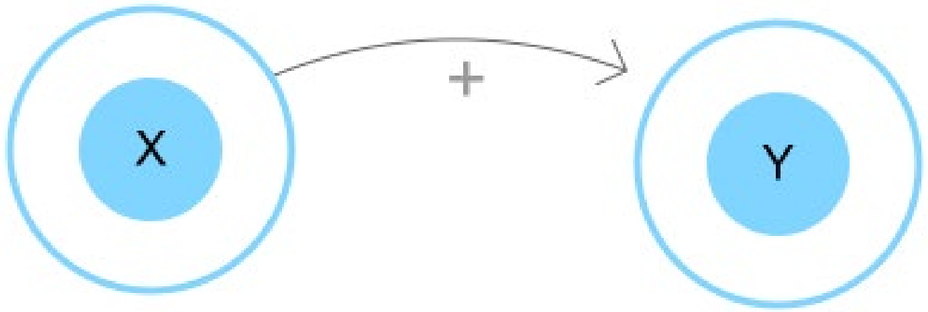
- User and stakeholder analysis
- Impact anticipation

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:

❌ 9% a. There’s a transition from state X to state Y on token “+”

❌ 11% b. The quantity in X is added to the quantity in Y

❌ 14% c. X and Y both change in an increasing direction

✅ 66% d. Y changes in the same direction as X

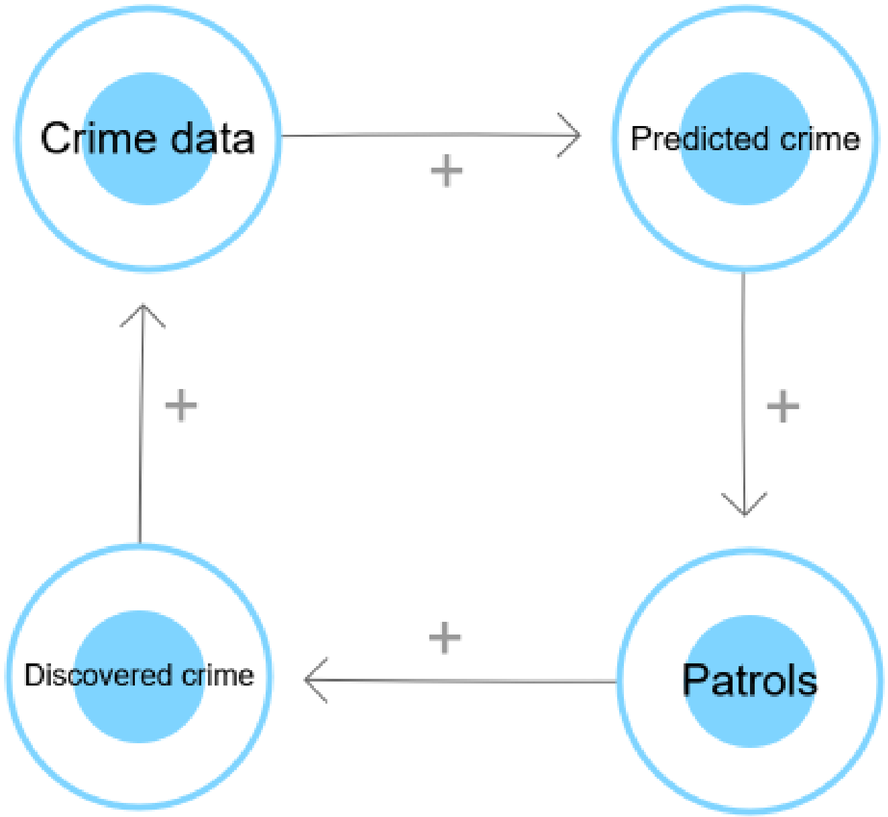
Instructions

Individually, read the context and scenario,
Then <u>work on Part 1:</u>

1. Describe with words how variables influence each other
2. Indicate the nature of the feedback loop

Part 1: behavior

URL: ttpoll.eu
Session ID: cs290



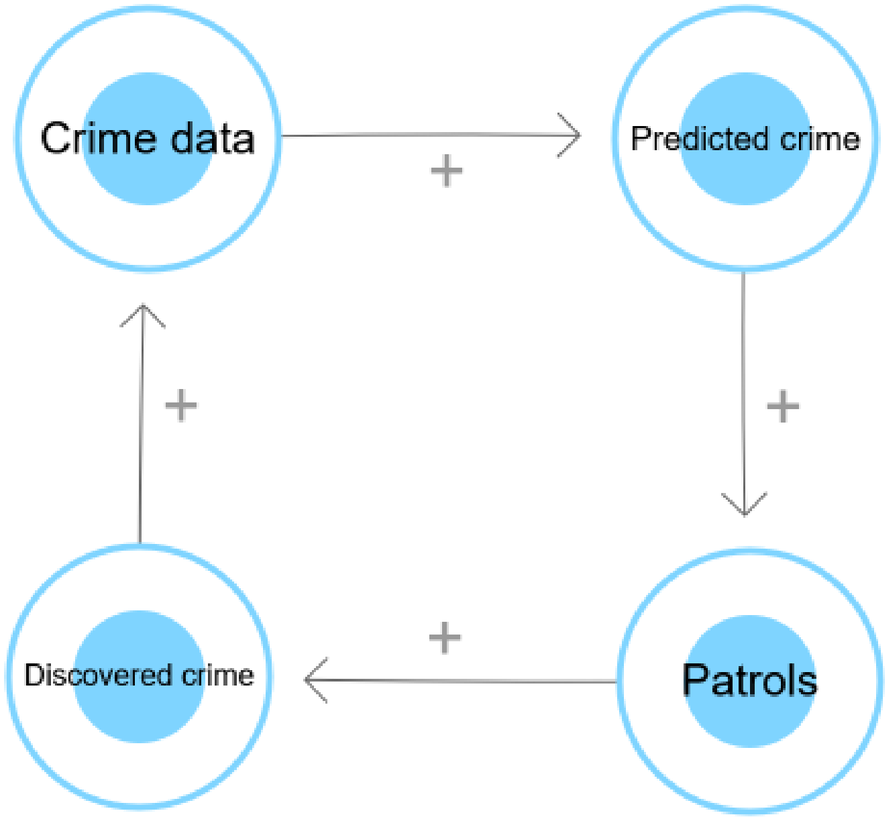
Over time, the quantities in this system will:

- ✗ 3% a. Stabilize
- ✗ 44% b. Increase
- ✗ 0% c. Decrease
- ✓ 54% d. It depends

The first variable to change will determine whether the quantities will increase or decrease.

Part 1: type of feedback loop

URL: ttpoll.eu
Session ID: cs290



The feedback loop in this diagram is:

❌ 0% a. Balancing

✅ 100% b. Reinforcing

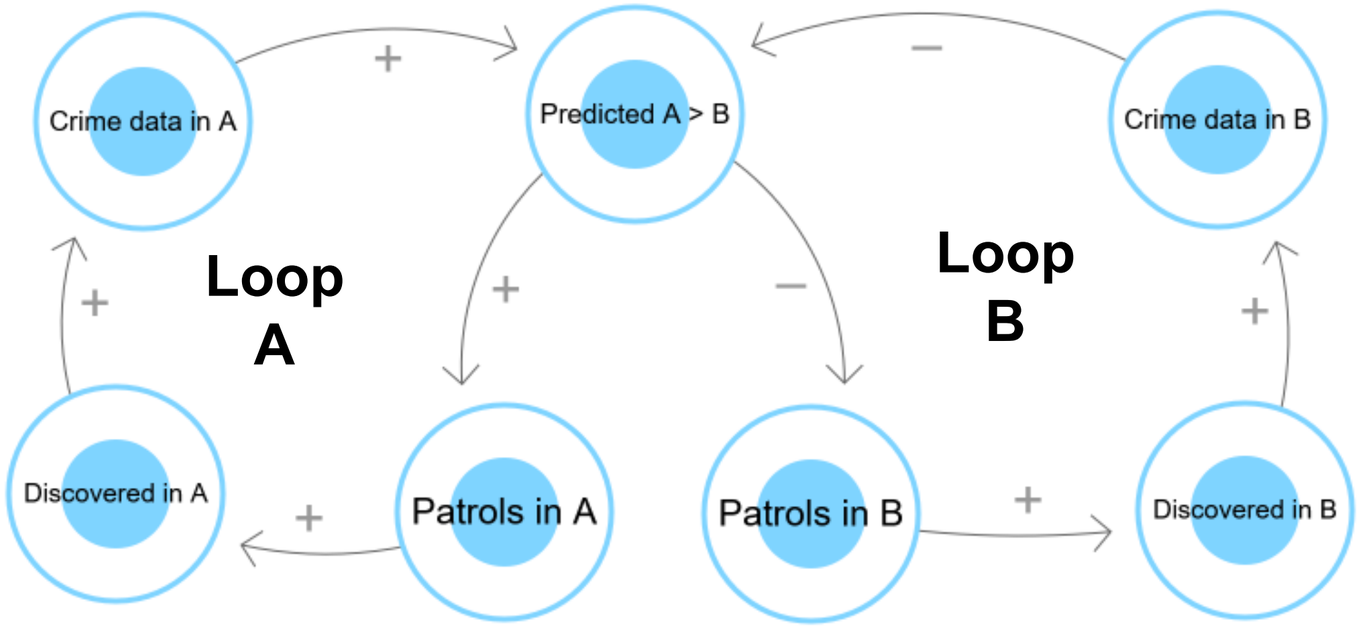
Instructions

Work on Part 2:

1. Describe with words how variables influence each other
2. Indicate the nature of the two feedback loops
3. Identify the long-term effects of deploying the predictive policing system

Part 2: types of feedback loops

URL: ttpoll.eu
Session ID: cs290



What is the type of loops A and B? (select **2** answers):

4% a. Loop A is balancing ❌

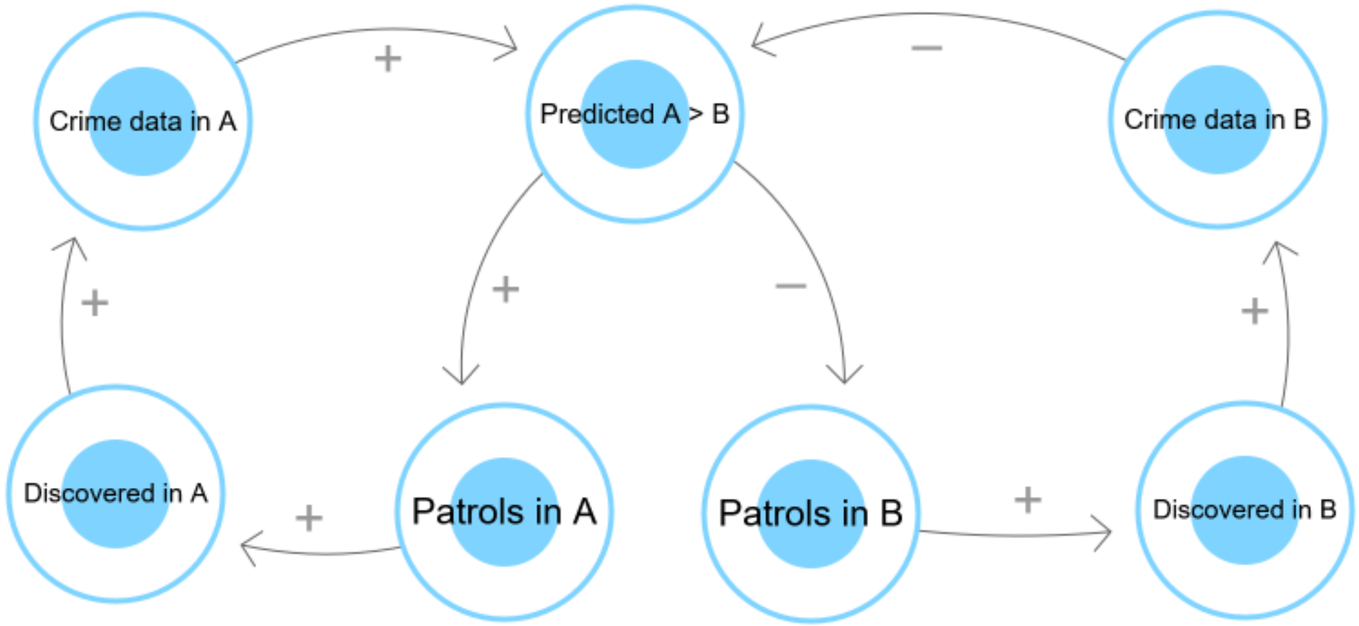
46% b. Loop A is reinforcing ✅

26% c. Loop B is balancing ❌

24% d. Loop B is reinforcing ✅

Part 2: behavior

URL: ttpoll.eu
Session ID: cs290



Over time, the introduction of the system will lead to:

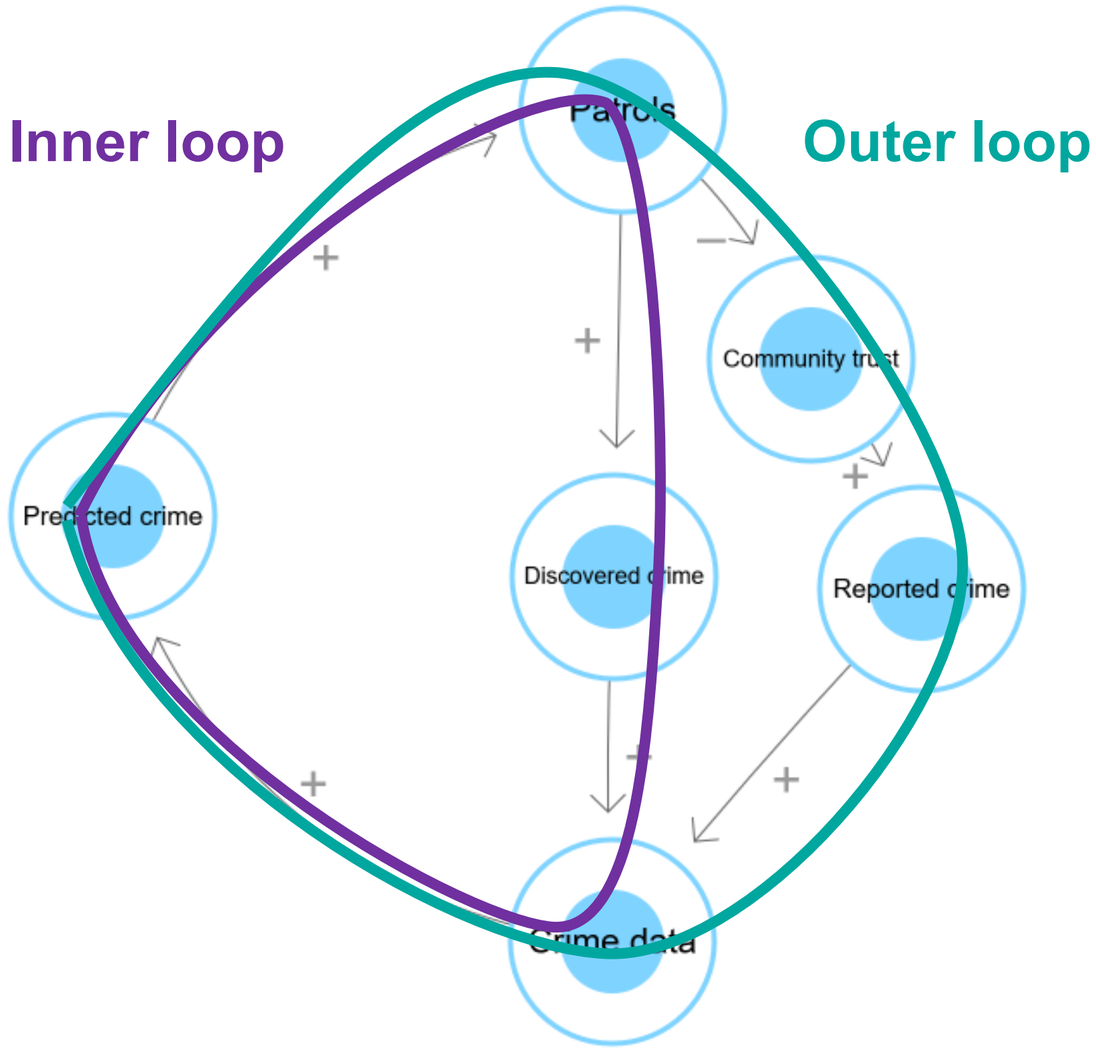
a. Balanced policing in A & B ❌
b. Over-policing in A ✅
c. Over-policing in B ❌
d. We cannot know ❌

Because the text indicates that more data has been collected in zone A historically, Crime data in A can be considered to increase which will lead to over-policing in A
/!\ We are talking about observed crime here, not true crime.

Instructions

Work on Part 3 at home:

1. Describe with words how variables influence each other
2. Indicate the nature of the two feedback loops



Debriefing

- For now, you are learning to **read** causal loop diagrams
- Later (in Sustainability 2) you will learn how to **draw** such diagrams

What’s next?

Computer rooms

We have very low attendance during the exercise sessions
👉 We are reducing the number of rooms with assistants

Room	Assistants
INF 1 (≈ 100 seats, no computers)	Eugène Bergeron Camille Lannoye Yingxuan You
CO 5 (≈ 40 seats)	Othmane Housni Elyes Trabelsi

(except during the Graded Assignement support sessions,
where we will open more rooms)

We start Fairness 1!

Tomorrow, Tuesday 30: notebook on fairness in university admissions

By Monday 6:

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

On Monday 6:

- Interactive questions on the theory
- Work on the **case studies together in class**