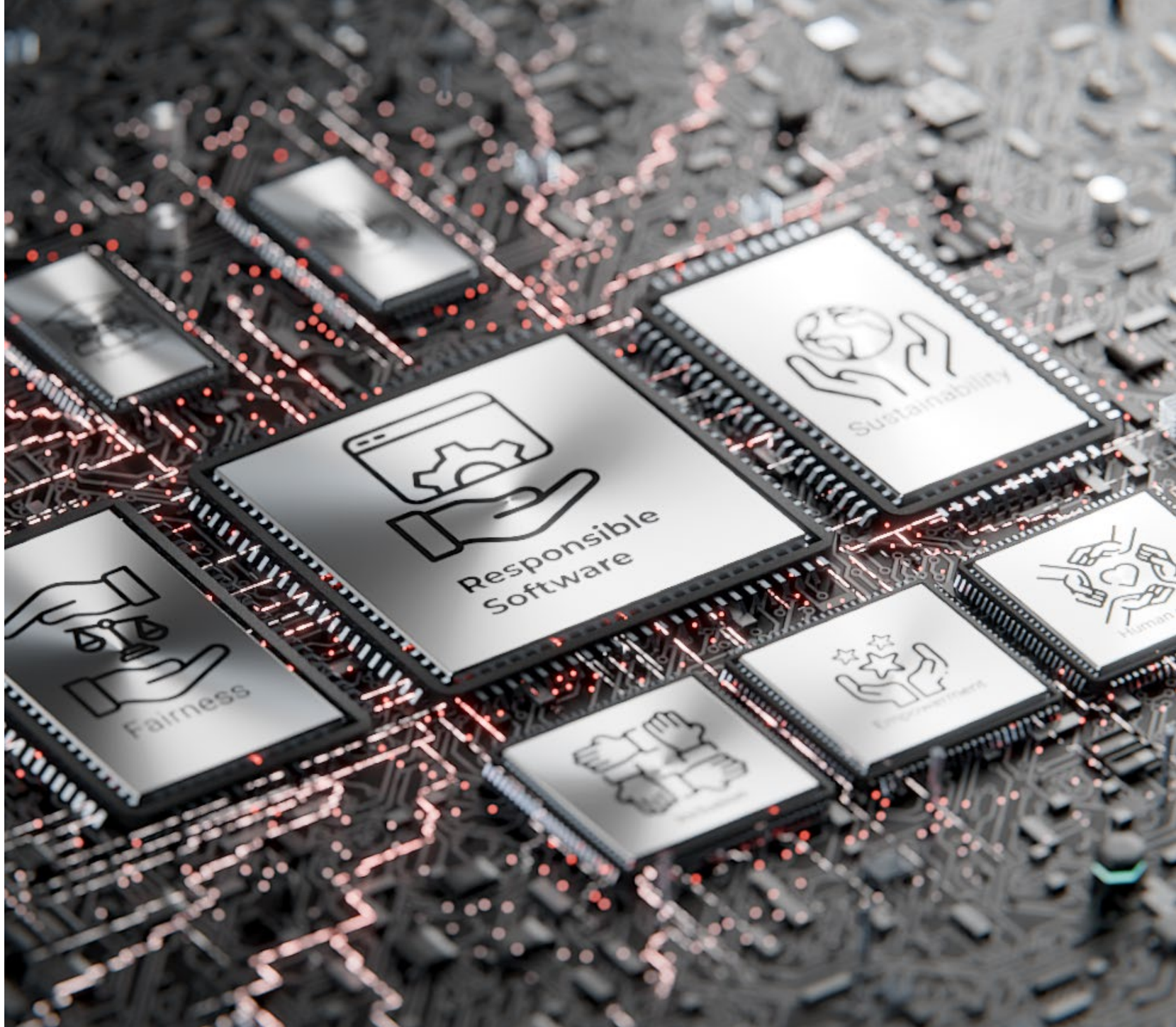
EPFL

# Safety 1 Review & Case Studies

**23 sept.**

Cécile Hardebolle

## Responsible Software



# Agenda for today

1. Interactive review questions on Safety 1
   (and some other topics)
2. Case studies:
   a) Bad actors
   b) STRIDE
   c) Harm modeling

# Logistics

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

- ❌ 9% a. It is in the winter exam session
- ✅ 21% b. It is on the last week of term
- ❌ 6% c. It includes programming
- ✅ 24% d. It includes case studies
- ✅ 14% e. It includes MCQs on the videos
- ❌ 2% f. All documents are allowed
- ✅ 23% g. Only one A4 paper notes is allowed

URL: ttpoll.eu
Session ID: cs290

# Workload

So far, **in addition to contact hours**, my **weekly workload** for the course is around:


URL: ttpoll.eu
Session ID: cs290

# Autonomous car software - 1

The software of an autonomous car fails to recognize traffic signs correctly.

We are in the presence of (select all that apply):

❌ 17% a. A safety threat

16% b. A security threat

Yes, if we consider it can be exploited to damage the system

✅ 58% c. A safety hazard

❌ 9% d. A security hazard

URL: ttpoll.eu
Session ID: cs290

# Autonomous car software - 2

When weather conditions include rain, the software of an autonomous car fails to recognize traffic signs correctly.
We are in the presence of (select all that apply):

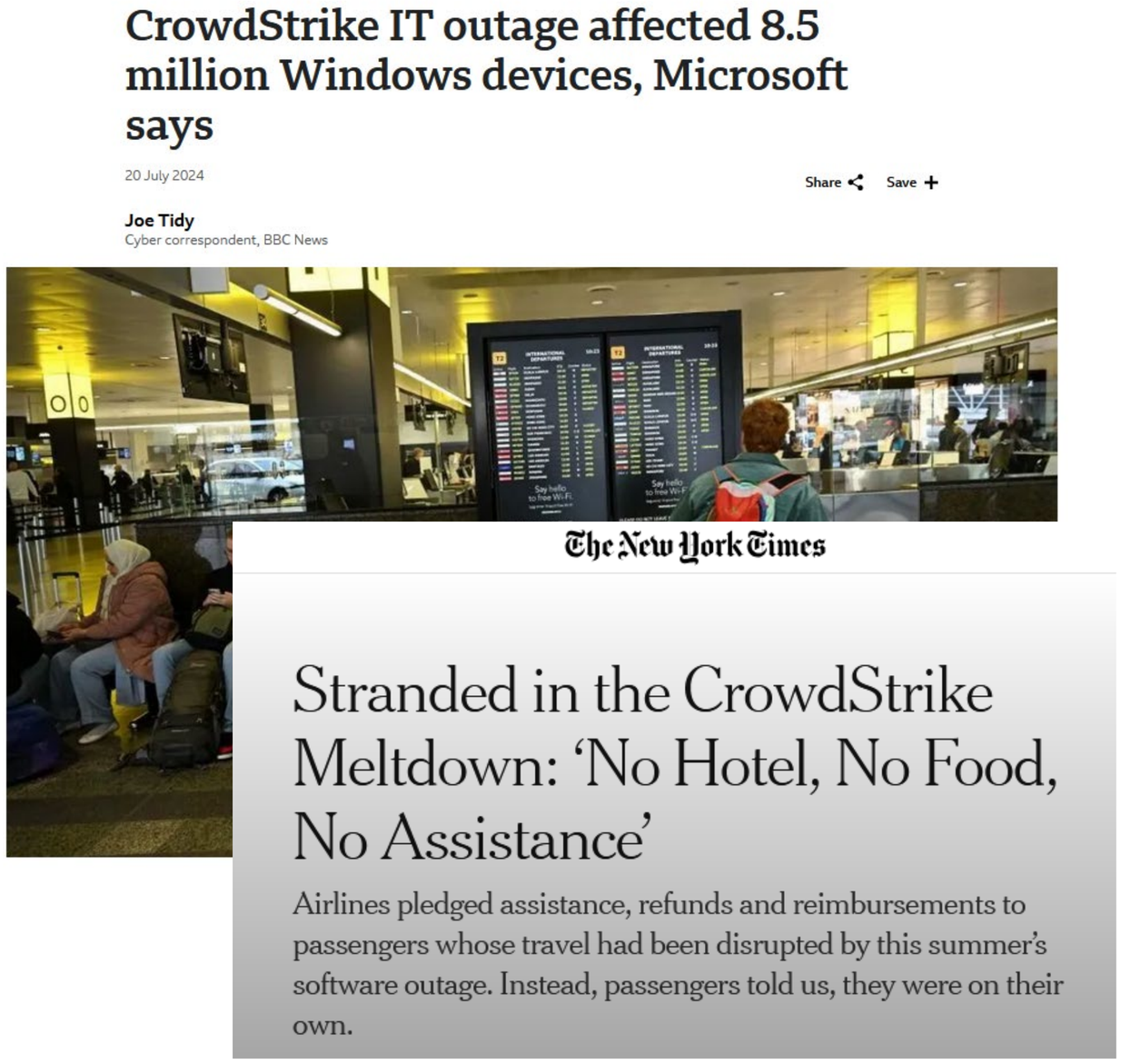

URL: ttpoll.eu
Session ID: cs290

# Worldwide “CrowdStrike” outage in 2024

This event is an example of:

- ✅ 88% a. Malfunction
- 6% b. Misuse, abuse
- 5% c. Unintended use
- 1% d. Intended use

URL: ttpoll.eu
Session ID: cs290



CrowdStrike IT outage affected 8.5 million Windows devices, Microsoft says

20 July 2024

Share Save

Joe Tidy
Cyber correspondent, BBC News

The New York Times

Stranded in the CrowdStrike Meltdown: ‘No Hotel, No Food, No Assistance’

Airlines pledged assistance, refunds and reimbursements to passengers whose travel had been disrupted by this summer’s software outage. Instead, passengers told us, they were on their own.

# Bad actors, safety and security

9% a. Bad actors generate safety issues only

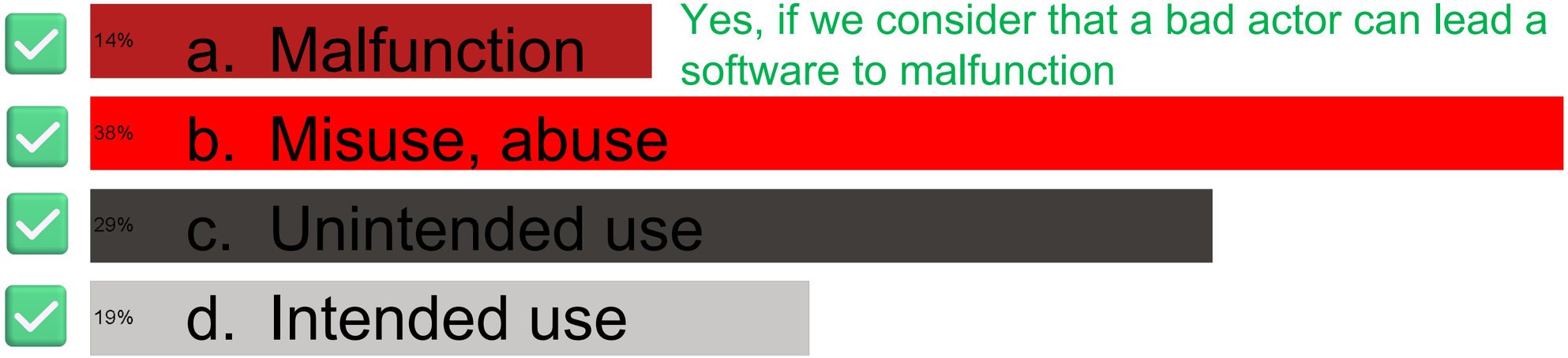
18% b. Bad actors generate security issues only

✅ 72% c. Bad actors generate both security and safety issues

URL: ttpoll.eu
Session ID: cs290

# Bad actors and the 4 scenarios

Bad actors can be involved in (select all that apply):




Yes, if we consider that a bad actor can lead a software to malfunction

URL: ttpoll.eu
Session ID: cs290

# The “confusing” matrix - 1

We use software to detect fissures in concrete walls before they become visible to the naked eye.
A positive result means presence of fissure.

URL: ttpoll.eu
Session ID: cs290

Select all the correct statements:

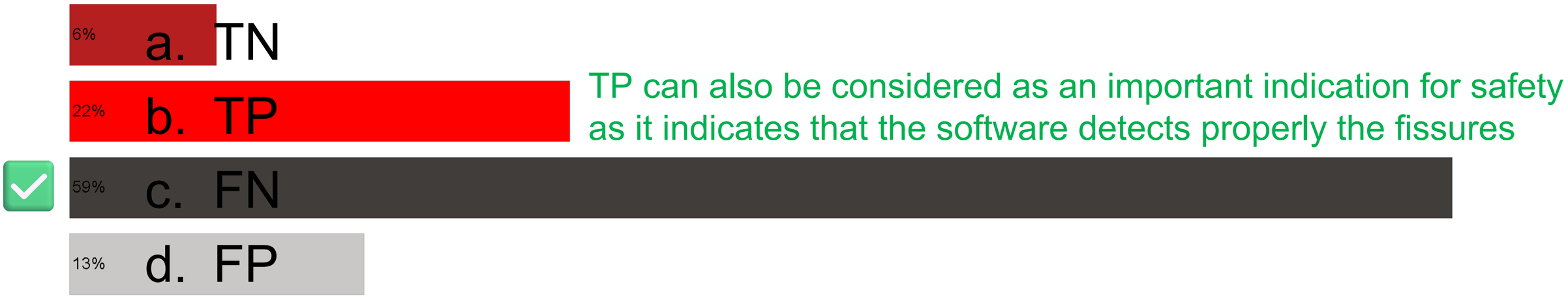
- ✅ 42% a. TN = actual absence of fissure, correct prediction
- ❌ 12% b. TP = actual absence of fissure, correct prediction
- ✅ 34% c. FN = actual presence of fissure, incorrect prediction
- ❌ 12% d. FP = actual presence of fissure, incorrect prediction

# The “confusing” matrix - 2

We use software to detect fissures in concrete walls before they become visible to the naked eye.
A positive result means presence of fissure.

URL: ttpoll.eu
Session ID: cs290

From a safety perspective, the indicator we should pay most attention to is:




TP can also be considered as an important indication for safety as it indicates that the software detects properly the fissures

# The “confusing” matrix - 3

We use software to detect fissures in concrete walls before they become visible to the naked eye.
A positive result means presence of fissure.
Here is the confusion matrix you get 👉

| | | Predicted: Fissure | Predicted: No Fissure |
|---|---|---|---|
| Actual | Fissure | 60 | 15 |
| Actual | No Fissure | 20 | 100 |

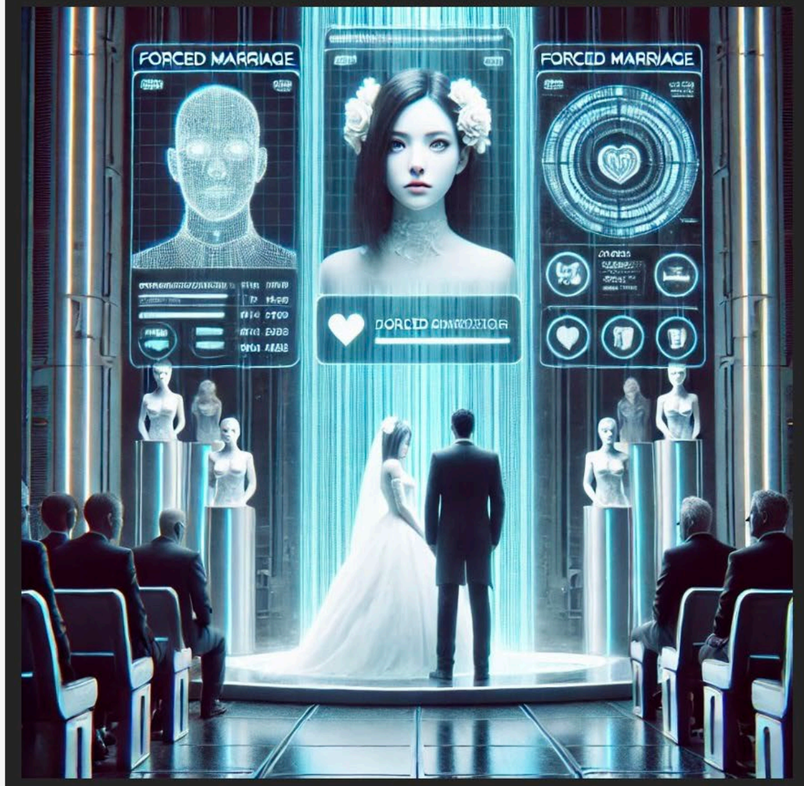
What is the False Negative Rate (FNR)?

- 33% a. 13%
- 8% b. 17%
- ✅ 55% c. 20%
- 4% d. 25%

FNR = FN / Actual P
= FN / (TP + FN)
= 15 / 75
= 20%

# Case studies

# Ethical speculation

| Algorithm's Bride | |
|---|---|
| In 2042, society is controlled by the Match algorithm, which assigns everyone a spouse by the age of 30. A user, nearing her 30th birthday, is given a match by the system but refuses to submit, rejecting a life dictated by data. As she resists, the algorithm starts stripping away her privileges healthcare, employment, even social connections, isolating her from society. With her 30th birthday approaching, she must decide marry the algorithm's choice or face a life as an outcast, erased from society for defying the system. |  |

# Where to find the cases?

1. Go to **moodle**
2. Find the **link to the case studies** for today
   👉this link will send you to courseware
   (where you can find all the course material)
3. Download:
   - The instruction sheet
   - The 3 cheatsheets
   - The template

# Bad actors

# Instructions

**Remember the notebook case: the social media “Catter”**

**Individually, identify the potential bad actors:**

1. Use the motivation categories to brainstorm a range of **harmful actions** that could be performed by **bad actors**
2. Identify the **impacts** for users and for the platform

**Share with your neighbor:**

- Did you identify the same bad actors?
- Can you agree on a final list?

# Which among these are bad actors in Catter?

Select all that apply:

37% a. Sassy posts a fake picture of Dogs invading Purrville

37% b. KitKat re-posts covert Dog propaganda

2% c. Felix promotes cat-biscuits that his cousin cooks

24% d. Tuna posts a series of angry replies to Catnip's post

Bad actor = risk of harm + intention

URL: ttpoll.eu
Session ID: cs290

# Post your bad actors

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

- A short description of the **bad actor type / motivation** (1line)
- A short description of the **security / safety impacts** (1 line)

⚠️ If your bad actor has already been posted, add a “like” / heart to it

# Overall debriefing of the strategy

- **Motivation categories overlap** and it is not always clear how to classify some types of actions
- The goal is to identify a **range of possible scenarios** that could create **threats** for your system (security) and **hazards** for your users (safety)

👉 **Use as a help for brainstorming**

# STRIDE

# Instructions

**Individually:**

1. Match each proposition to a threat category from STRIDE
2. Describe a countermeasure to prevent / mitigate the issue

**Share with your neighbor:**

- Compare your matching
- Discuss your countermeasures

# Debriefing

1.
    1. Information Disclosure (I)
2.
    2. Tampering (T) + Elevation of privilege
3.
    3. Spoofing (S)
4.
    4. Denial of Service (D)
5.
    5. Repudiation (R)

# Why do we do this?

- Goal = identifying how bad actors can generate **security + safety issues** that **lead to harm** in order to anticipate and prevent it

# Harm modeling

# Harm categories - 1

A user sees their post unfairly censored.

This harm is in the category (select one):


URL: ttpoll.eu
Session ID: cs290

# Harm categories - 2

A fitness app leaks GPS location data on social media.
This harm is in the category (select one):

| | |
|---|---|
| 3% | a. Physical injury |
| 2% | b. Emotional or psychological injury |
| 0% | c. Opportunity loss |
| 0% | d. Economic loss |
| 0% | e. Dignity loss |
| 0% | f. Liberty loss |
| 95% | g. Privacy loss |
| 0% | h. Environmental impact |
| 0% | i. Manipulation |
| 0% | j. Social detriment |

URL: ttpoll.eu
Session ID: cs290

# Harm categories - 3

Online ads lead a compulsive shopper to additional purchases.
This harm is in the category (select one):

0% a. Physical injury

3% b. Emotional or psychological injury

0% c. Opportunity loss

49% d. Economic loss

0% e. Dignity loss

0% f. Liberty loss

0% g. Privacy loss

0% h. Environmental impact

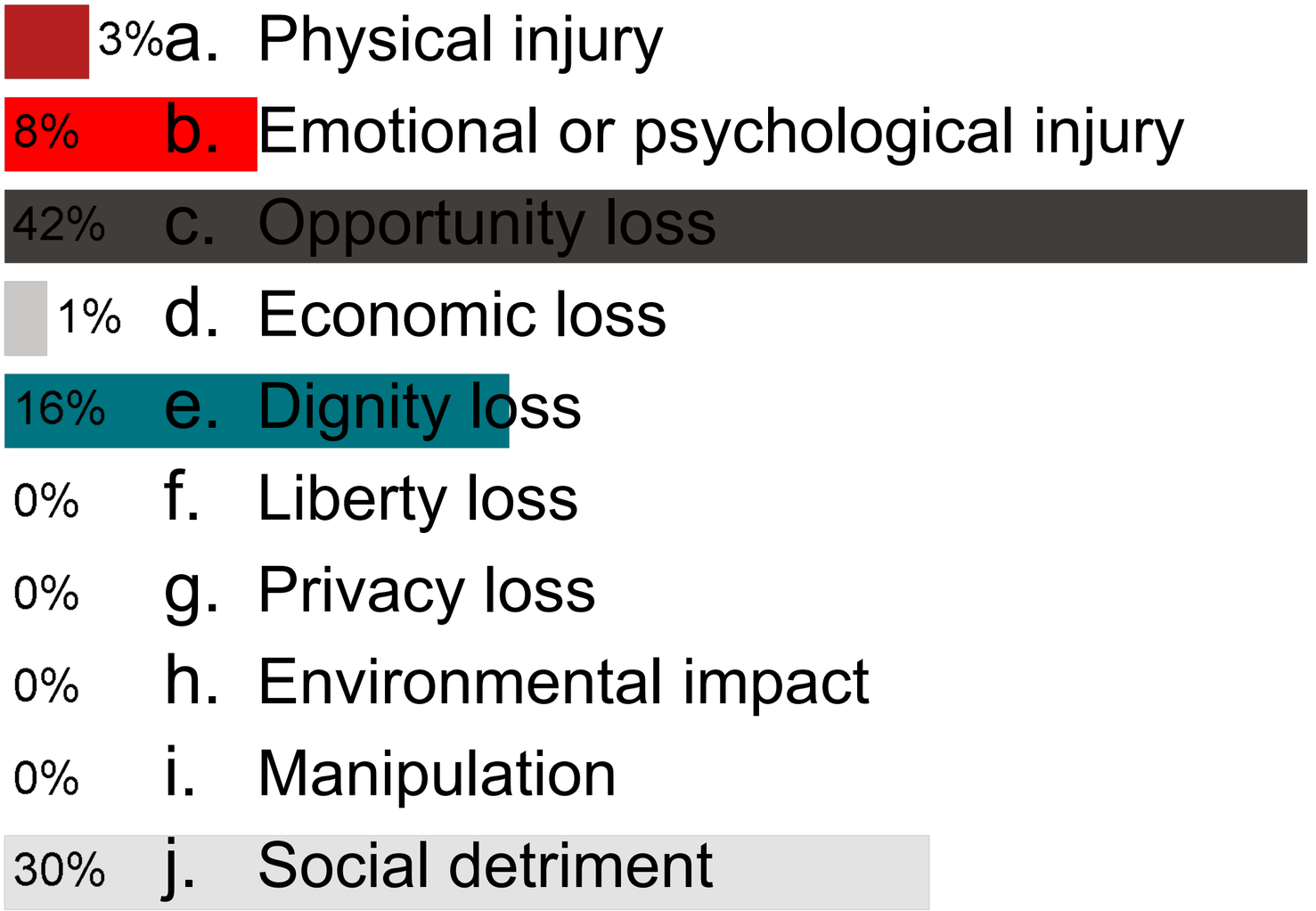
48% i. Manipulation

0% j. Social detriment

URL: ttpoll.eu
Session ID: cs290

# Harm categories - 4

A recruitment software indirectly discriminates based on people's name. This harm is in the category (select one):




URL: ttpoll.eu
Session ID: cs290

# Harm categories - 5

The results of an image search engine for “Nurse” show only women. This harm is in the category (select one):

0% a. Physical injury

3% b. Emotional or psychological injury

0% c. Opportunity loss

0% d. Economic loss

7% e. Dignity loss

0% f. Liberty loss

0% g. Privacy loss

1% h. Environmental impact

9% i. Manipulation

80% j. Social detriment

URL: ttpoll.eu
Session ID: cs290

# Instructions

**Read the <u>“Smart home technologies” scenario (2nd one)</u>**
*Use the “*Affect-Display Textile Garment*” scenario for later practice.*

**Individually, fill out the harm table from the template**
Use the description of the categories from the cheatsheet

**Share with your neighbor:**
- Did you identify the same harms?
- Did you classify harms in the same way?

# Overall debriefing of the strategy

It may not always be evident to classify some types of harms:

- Some harms may **fall into several categories**
- Some categories of harms **overlap**

**Not all categories apply to each case!**

We should do this **for different scenarios!**
(including not creating the product)

# What’s next?

# We start Safety 2!

Tomorrow, Tuesday 24: notebook on content recommendation

By Monday 30:

- Watch **videos 2.1 to 2.5** + do the **quizzes**
- Finish the notebook
  (and any other leftover from previous weeks)