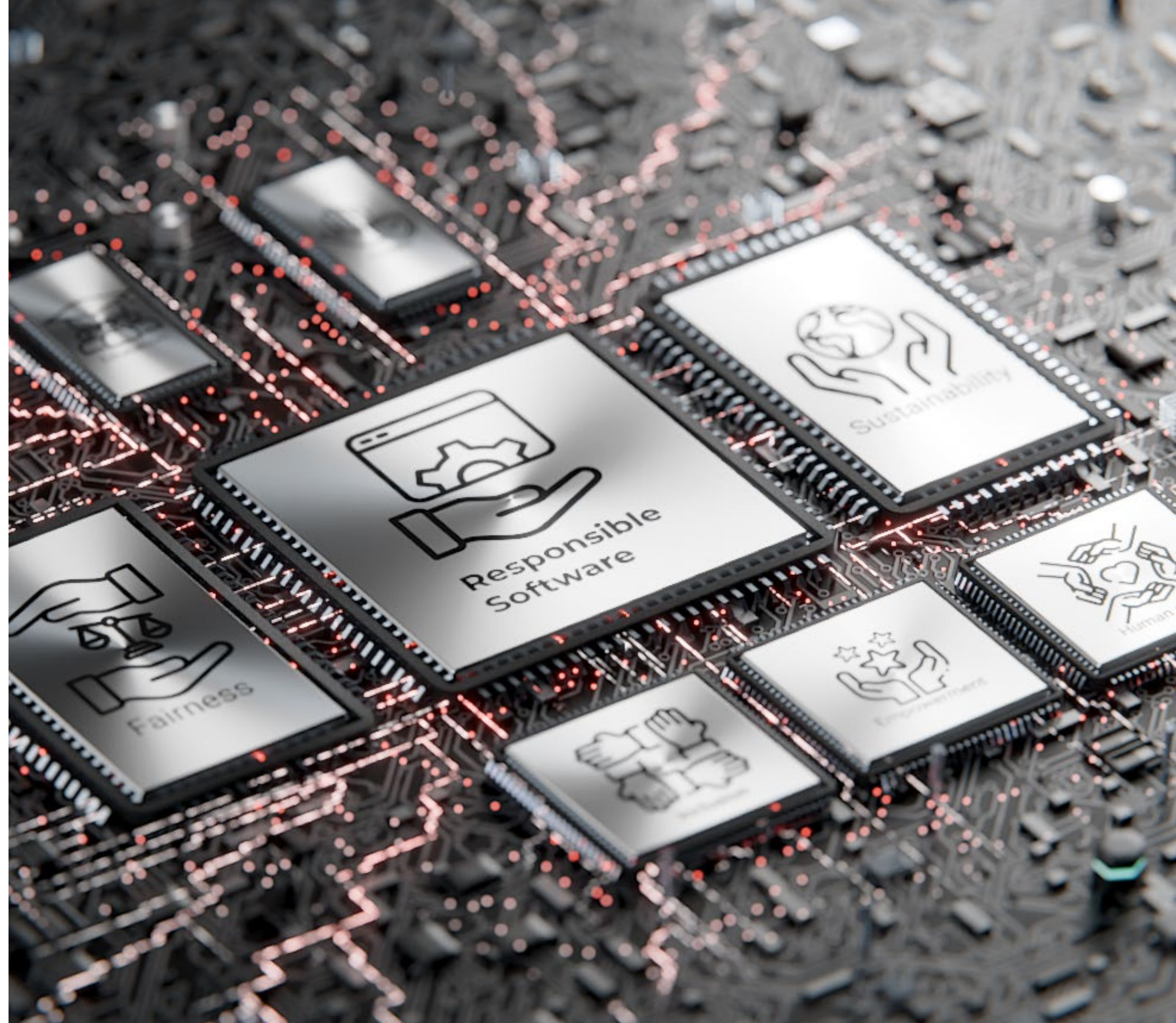


Conclusion Case studies + Q&A 10 dec.

Cécile Hardebolle

**Responsible
Software**



Agenda for today

1. Review case studies:

- a) Digital Ethics Canvas – Emotion Cancelling AI
- b) Ethics Canvas – Be My Eyes

2. Questions & Answers

3. Some interactive review questions

Case studies

Where to find the cases?

1. Go to **moodle**
 2. Find the **link to 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)

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

		Impact			
Probability	X	🕒	🕒	🕒	🕒
	🕒	Green	Green	Yellow	Yellow
	🕒	Green	Yellow	Red	Red
	🕒	Yellow	Red	Red	Red

Green: Low risk
Yellow: Medium risk
Red: High risk

“Emotion Cancelling AI”

👉 1 post = 1 benefit

Post your ideas:

<https://speakup.epfl.ch>

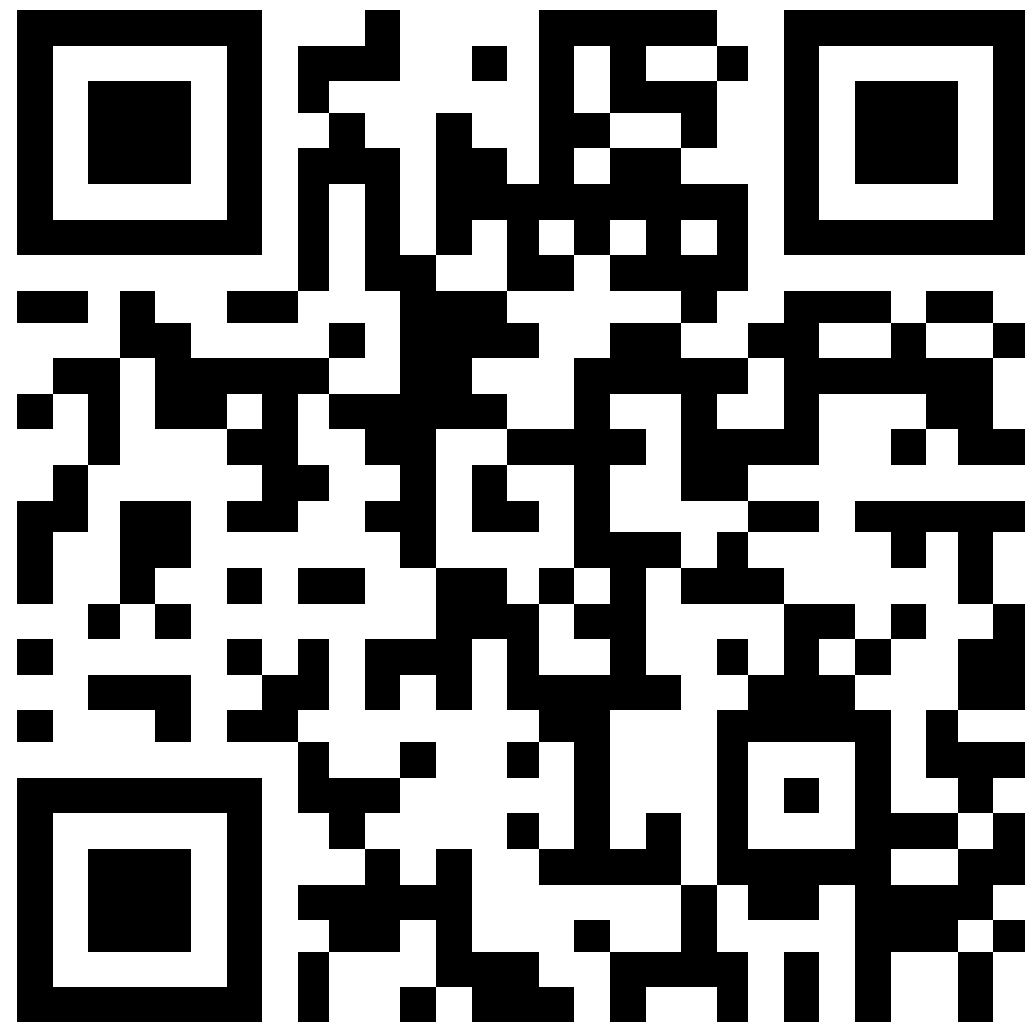
Room key: **57217**

👉 1 post = 1 risk + ethical lens

Post your ideas:

<https://speakup.epfl.ch>

Room key: **38467**



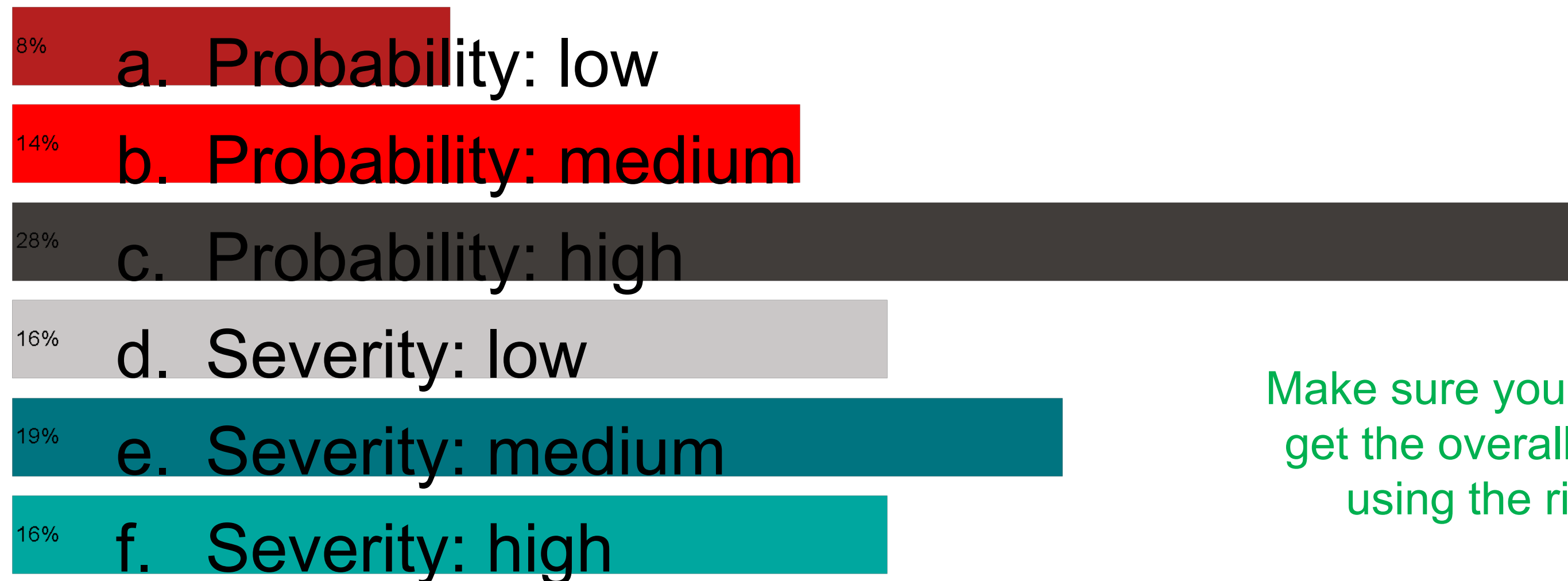
Risks:
Make sure to explain how the risk relates to the corresponding ethical lens (e.g. if you put a risk into “Fairness”, it must be clear what is unfair or biased).



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)



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)

3%

a. Probability: low

20%

b. Probability: medium

28%

c. Probability: high

7%

d. Severity: low

25%

e. Severity: medium

18%

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

Make sure to explain how
your proposal helps reduce
the corresponding risk

Post your ideas:

<https://speakup.epfl.ch>

Room key: **02297**



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**



Comparison!

Ethics Canvas

Project Title:

Date:

Ethics Canvas v1.8 - [ethicscanvas.org](#) © ADAPT Centre & Trinity College Dublin & Dublin City University, 2017.

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	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	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	Relations Discuss problematic differences in individual behaviour such as differences in habits, time-schedules, choice of activities, etc  4	Group Conflicts Discuss the impact on the relationships between the groups identified, e.g. employers and unions  6	

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

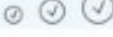
RISK

- 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





This work is licensed under [CC BY-NC 4.0](#)

Digital Ethics Canvas 2024 - C. Norddahl, V. Panko, V. Ramachandran, T. Su, R. Bartholomew, A. Adair, P. Jermann

Q&A

Attributes: Sensitive or Protected?

“Are any sensitive or protected attributes defined in law?

Can we have an exact definition of both? In the course we defined sensitive attributes as those that can have ethical implications, and protected that can harm.”

■ **“Sensitive”: can have ethical implications (privacy, fairness...)**

■ **“Protected”:**

- = Sensitive
- = Protected by law ⚠ depends on geographical location
+ data protection law & non-discrimination law
- = Should not be used / should be treated specifically to prevent harm
in software design, in particular ML

GDPR (EU) – Personal data

What is personal data?

Personal data is any information that relates to an identified or identifiable living individual (data subject). Different pieces of information, which together can lead to the identification of a particular person, may also be considered personal data.

Personal data that has been de-identified, encrypted or **pseudonymised** but can be used to re-identify a person remains personal data and falls within the scope of the General Data Protection Regulation (GDPR), the EU's main data protection law.

Examples of personal data

- a name and surname
- a home address
- an email address such as '[name.surname@company.com](#) 
- an Internet Protocol (IP) address
- an identification card number
- a cookie ID
- the advertising identifier of your phone
- data held by a hospital or doctor, which could be a symbol that uniquely identifies a person

Data protection rights

Under the GDPR, individuals have several rights over their personal data.

The rights of individuals

- Right to be informed
- Right of access
- Right to rectification
- Right to erasure
- Right to restriction of processing
- Right to data portability
- Right to object
- Rights in relation to automated decision-making and profiling

https://commission.europa.eu/law/law-topic/data-protection_en

GDPR (EU) – Sensitive data

What personal data is considered sensitive?

Answer

The following personal data is considered 'sensitive' and is subject to specific processing conditions:

- personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs;
- trade-union membership;
- genetic data, biometric data processed solely to identify a human being;
- health-related data;
- data concerning a person's sex life or sexual orientation.

References

- [Article 4\(13\), \(14\) and \(15\)](#) and [Article 9](#) and [Recitals \(51\) to \(56\) of the GDPR](#)

https://commission.europa.eu/law/law-topic/data-protection/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data_en

Under what conditions can my company/organisation process sensitive data?

Answer

Your company/organisation can only process [sensitive data](#) if one of the following conditions is met:

- the **explicit consent** of the individual was obtained (a law may rule out this option in certain cases);
- an **EU or national law or a collective agreement**, requires your company/organisation to process the data to comply with its obligations and rights, and those of the individuals, in the fields of employment, social security and social protection law;
- the **vital interests** of the person, or of a person physically or legally incapable of giving consent, are at stake
- you are a **foundation, association or other not-for-profit body** with a political, philosophical, religious or trade union aim, processing data about its **members** or about people in regular contact with the organisation;
- the personal data was **manifestly made public** by the individual;
- the data is required for the establishment, exercise or defence of **legal claims**
- the data is processed for reasons of **substantial public interest** on the basis of EU or national law;
- the data is processed for the purposes of **preventive or occupational medicine**, assessment of the working capacity of the employee, medical diagnosis, the provision of health or social care or treatment, or the management of health or social care systems and services on the basis of EU or national law, or on the basis of a contract as a health professional;
- the data is processed for reasons of **public interest in the field of public health** on the basis of EU or national law;
- the data is processed for **archiving, scientific or historical research purposes** or statistical purposes on the basis of EU or national law.

EU Charter of Fundamental Rights

←

TITLE III
Equality

→

←

Article 21 - Non-discrimination

→

Article 21 - Non-discrimination

1. Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.

2. Within the scope of application of the Treaties and without prejudice to any of their specific provisions, any discrimination on grounds of nationality shall be prohibited.

EXPLANATIONS

CASE LAW REFERENCES

NATIONAL CONSTITUTIONAL LAW

EU LAW

INTERNATIONAL LAW

PRODUCTS

Review questions **“Whole Course”**

Ethical sensitivity refers to a person's... (select 1 answer)

0%

a. ... willingness to acknowledge their mistakes



82%

b. ... capability to understand the impact of a situation on others

5%

c. ... ability to act to benefit others even at their own expense

13%

d. ... ability to enforce ethical values in a team

An ethical dilemma is a situation where... (select 1 answer)

0%

a. ... there is a solution that is better than the others

5%

b. ... it is possible to find an ethical compromise

5%

c. ... a good trade-off between solutions is possible

90%

d. ... all solutions conflict with an ethical value



Autonomous car software

URL: ttpoll.eu
Session ID: cs290

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

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

18%

a. A safety threat

24%

b. A security threat



41%

c. A safety hazard

16%

d. A security hazard

The “confusing” matrix

URL: ttpoll.eu
Session ID: cs290

We use software to detect fissures in concrete walls before they become visible to the naked eye.

A positive result means presence of fissure.

Select all the correct statements:



44%

a. TN = actual absence of fissure, correct prediction

8%

b. TP = actual absence of fissure, correct prediction



41%

c. FN = actual presence of fissure, incorrect prediction

8%

d. FP = actual presence of fissure, incorrect prediction

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):

- ☒ 59% a. Misinformation
- ☐ 11% b. Disinformation
- ☐ 9% c. Malinformation
- ☒ 20% d. Fake news

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 (select 1 answer):

- 9%

a. System 2
- 6%

b. False consensus
- ✓

83%

c. Source cues
- 3%

d. Illusory truth

Attributes

URL: ttpoll.eu
Session ID: cs290

Hair color as an attribute to represent people is:
(select all that apply)

- ☒ 0% a. A sensitive attribute
- ☒ 0% b. An observed variable
- ☐ 0% c. A latent variable
- ☐ 0% d. An objective representation of people
- ☒ 0% e. A subjective representation of people

Bias

URL: ttpoll.eu
Session ID: cs290

The city of Lozhann decides to deploy a smartphone app that allows residents to report potholes throughout the city to help with the identification of repair needs.

The data collected by the app will probably exhibit:
(select all that apply)

- ☒ 14% a. Preexisting bias
- ☐ 6% b. Confirmation bias
- ☒ 38% c. Representation bias
- ☒ 41% d. Measurement bias
- ☐ 1% e. Automation bias



Biases in the ML lifecycle

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.

This is a case of (select 1 answer):



55%

a. Evaluation bias

33%

b. Aggregation bias

0%

c. Optimization bias

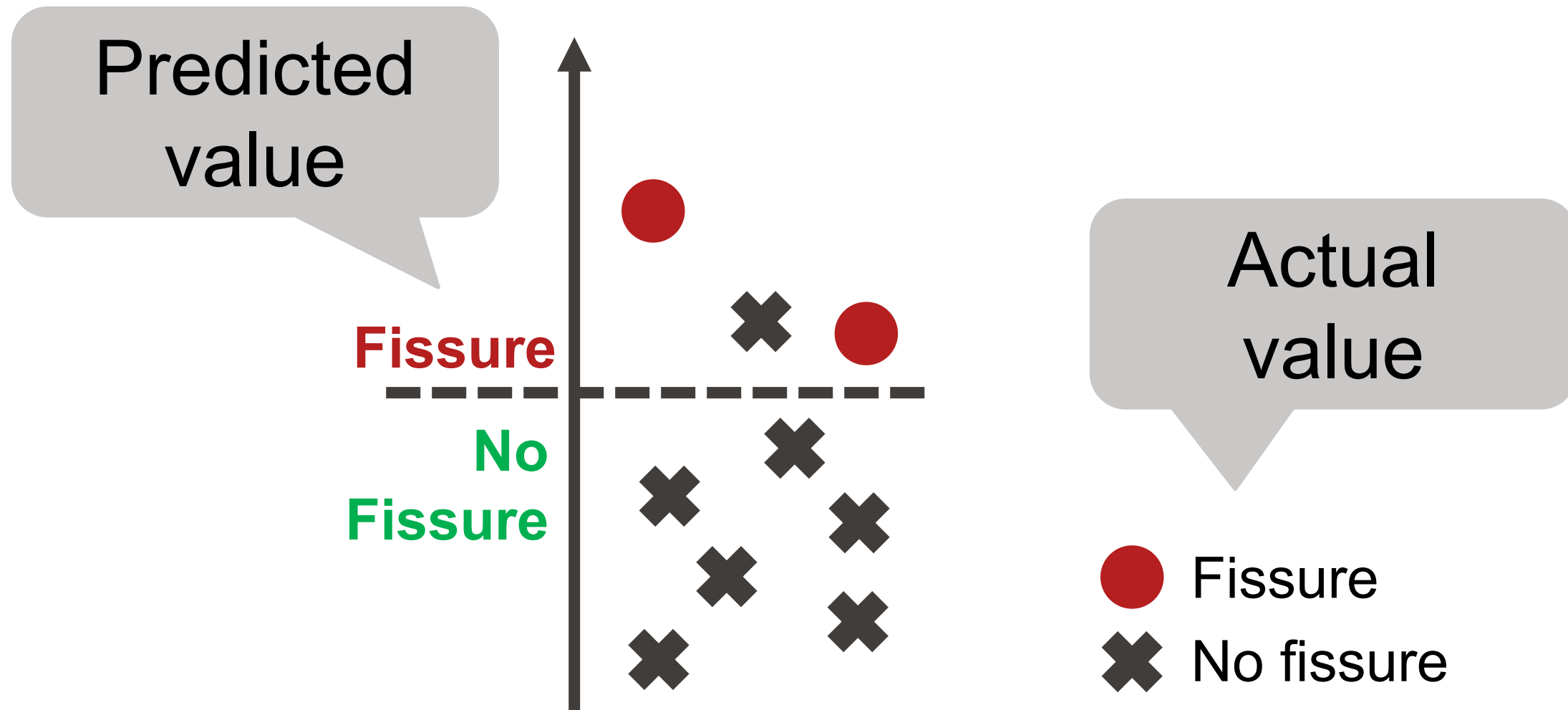
12%

d. Deployment bias

Evaluation: a benchmark
with **actors** does not
correspond to the target
application context (real
shops)

Fairness metrics

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	No Fissure
Actual	Fissure	2	0
	No Fissure	1	6

Fairness metrics

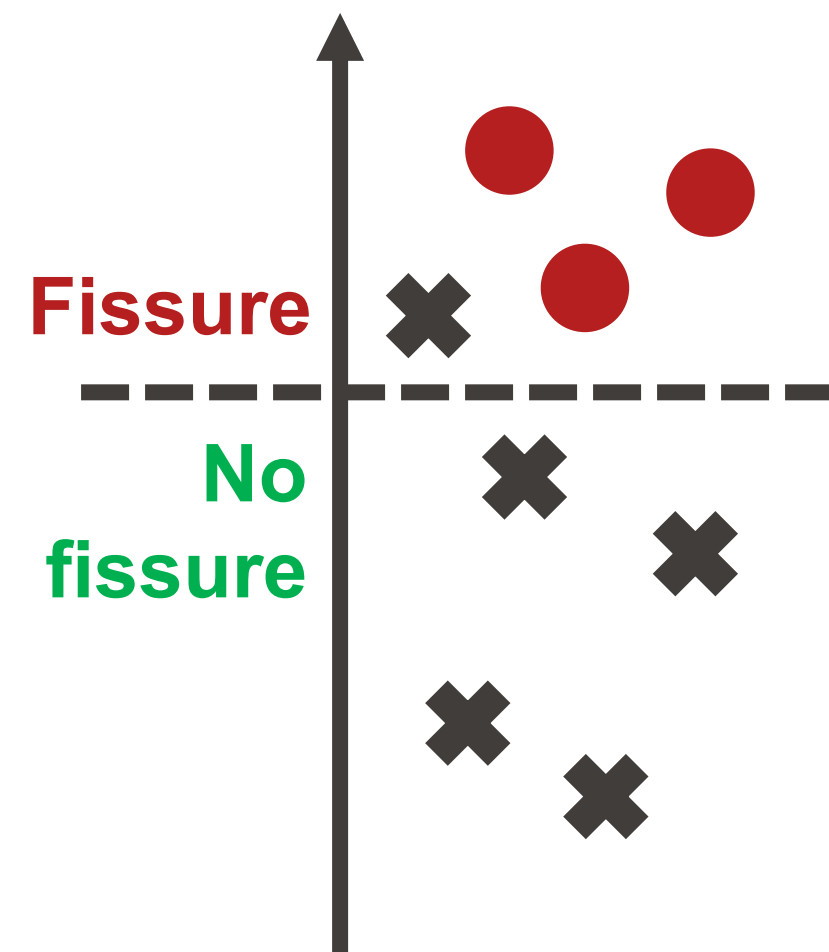
New question

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:

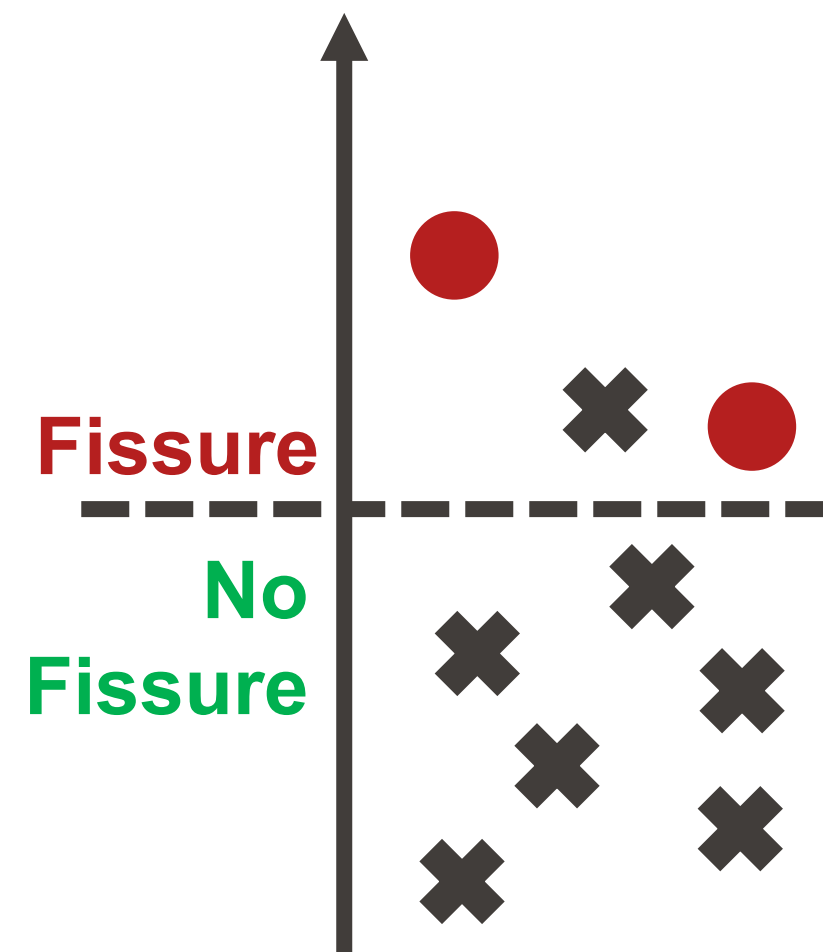
Metric = 1 / 5

Plain
Concrete

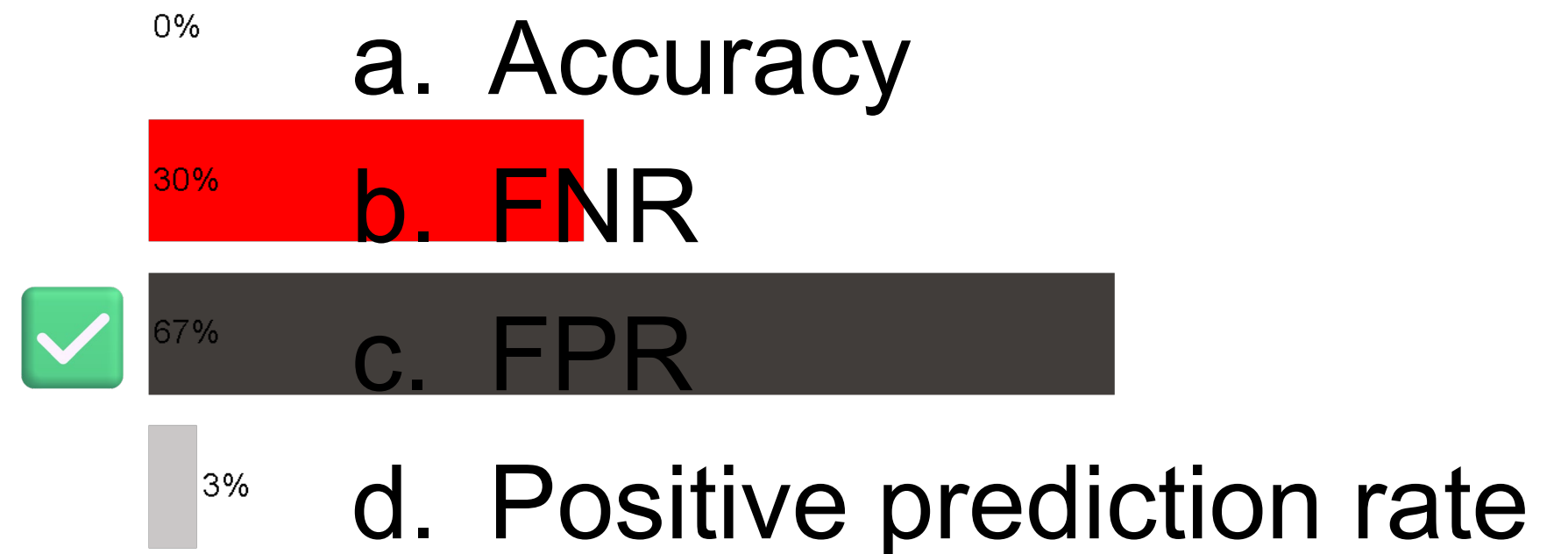


Metric = 1 / 7

Reinforced
Concrete



Which metric are they using? (select 1 answer)



Power Usage Effectiveness

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



The causes of water consumption from a datacenter which are included in the metric WUE_{source} are (select all that apply):

0%

a. Material mining

0%

b. Hardware manufacturing

0%

c. Concrete production



0%

d. Electricity production



0%

e. Cooling

Nudges

URL: ttpoll.eu
Session ID: cs290

Which of the following are examples of digital (software) nudges?
(select all that apply)

- ☐ 0% a. Automatic redirection to another website.
- ☐ 0% b. Automatic newsletter subscription as stated in usage policy.
- ☒ 0% c. Default value in online form
- ☐ 0% d. Notice about strictly necessary cookies
- ☒ 0% e. Notice about the behavior of other people

Translation

URL: ttpoll.eu
Session ID: cs290

Consider the following translation. What is the issue here?

French ▾

↔ English (American) ▾

Glossary

Dans un souci de durabilité, la recherche d'itinéraire dans Noodle Maps renvoie désormais 2 options d'itinéraire dans l'ordre suivant :
1) l'itinéraire consommant le moins de carburant mais le plus long
2) l'itinéraire le plus court mais consommant plus de carburant

×

In the interests of sustainability, the route search in Noodle Maps now returns 2 route options in the following order:
1) the most fuel-efficient but longest route
2) the shortest but most fuel-efficient route

0%

a. Parity error

0%

b. Factuality error

0%

c. Measurement error



0%

d. Faithfulness error

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)

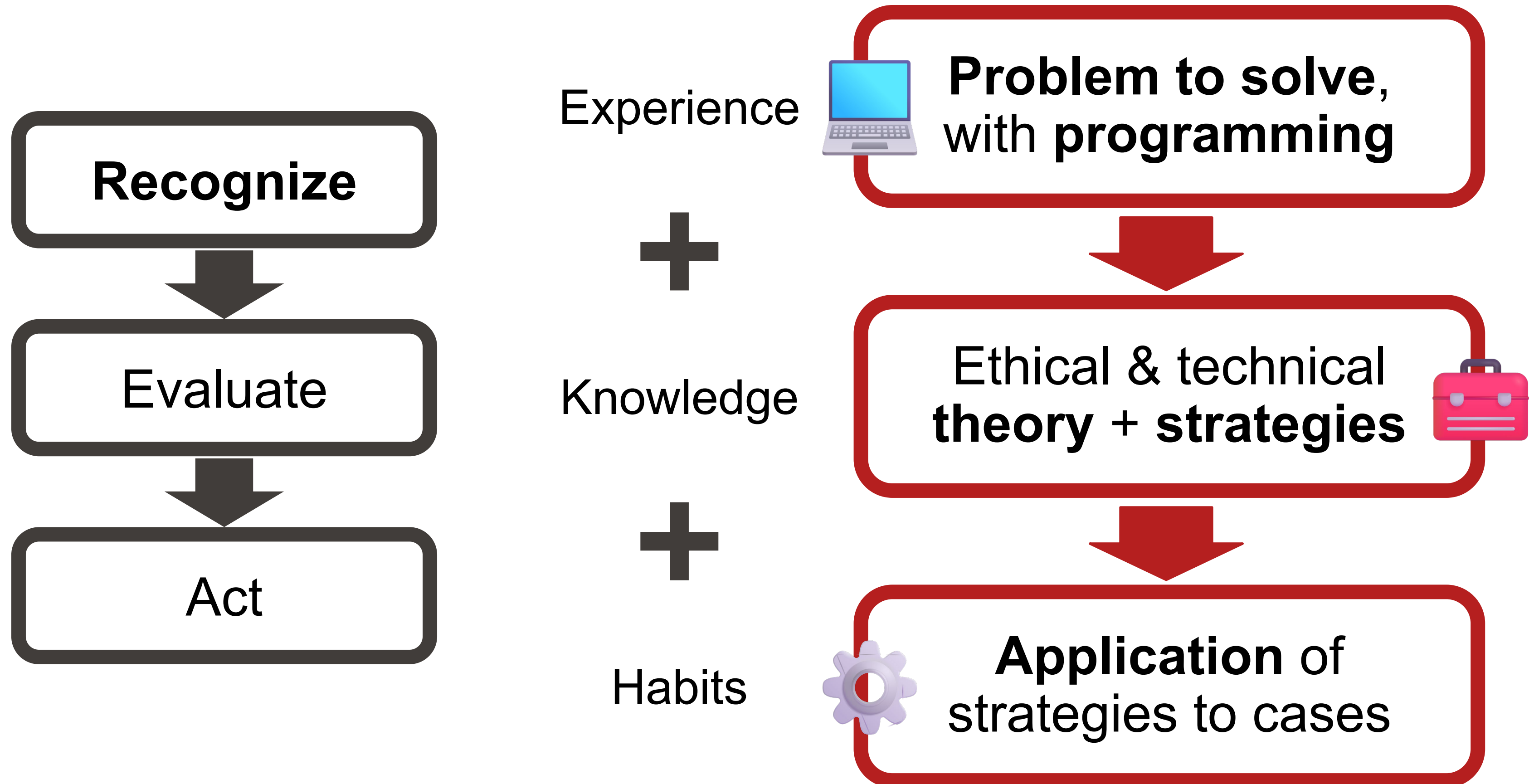
“Computing professionals’ actions change the world. To **act responsibly**, they should **reflect upon the wider impacts** of their work, consistently **supporting the public good.**”
(ACM, 2018)

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)



We can 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
exam 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>