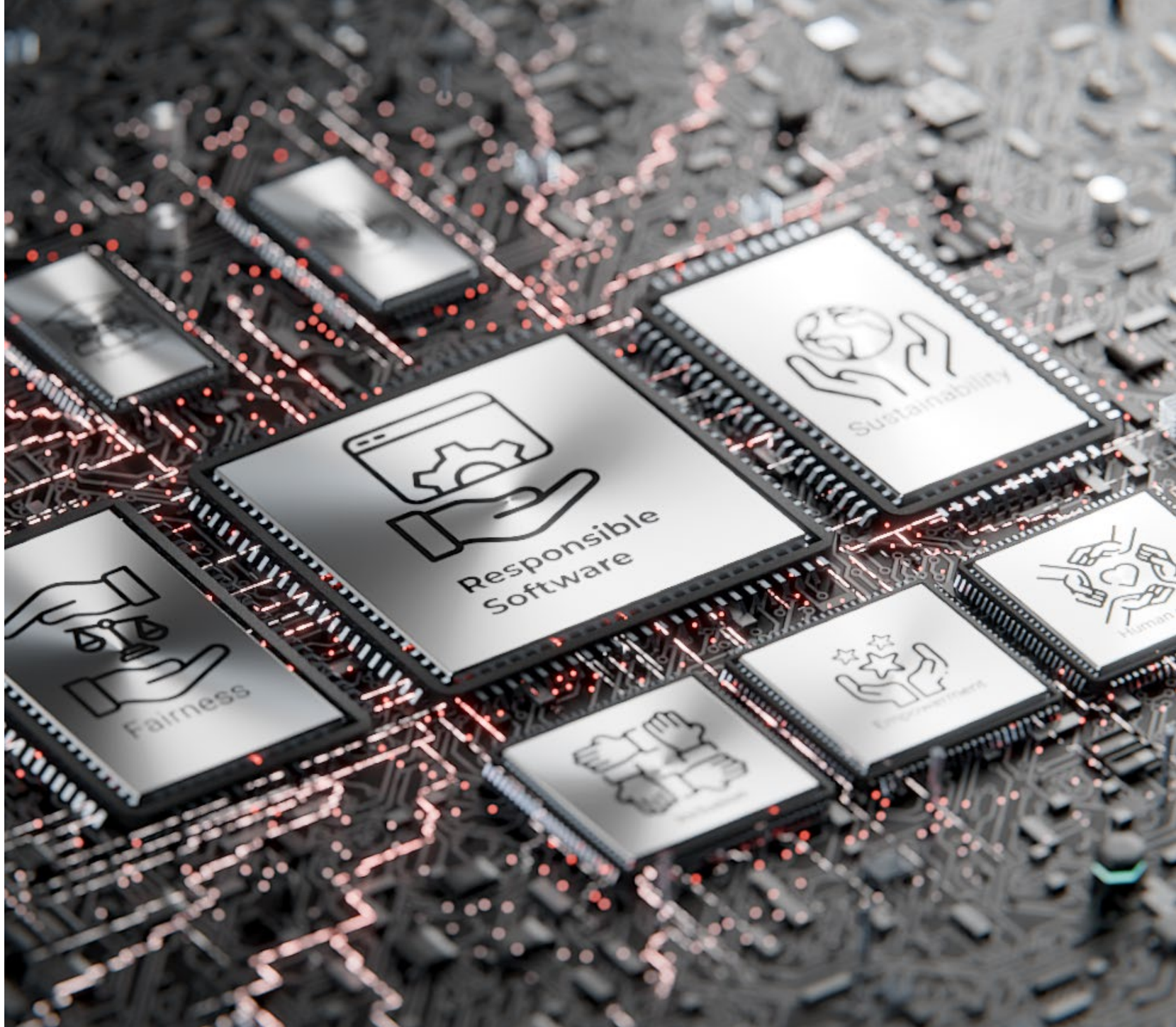
Introduction Session

15 sept.

Cécile Hardebolle

Responsible Software



Responsibility

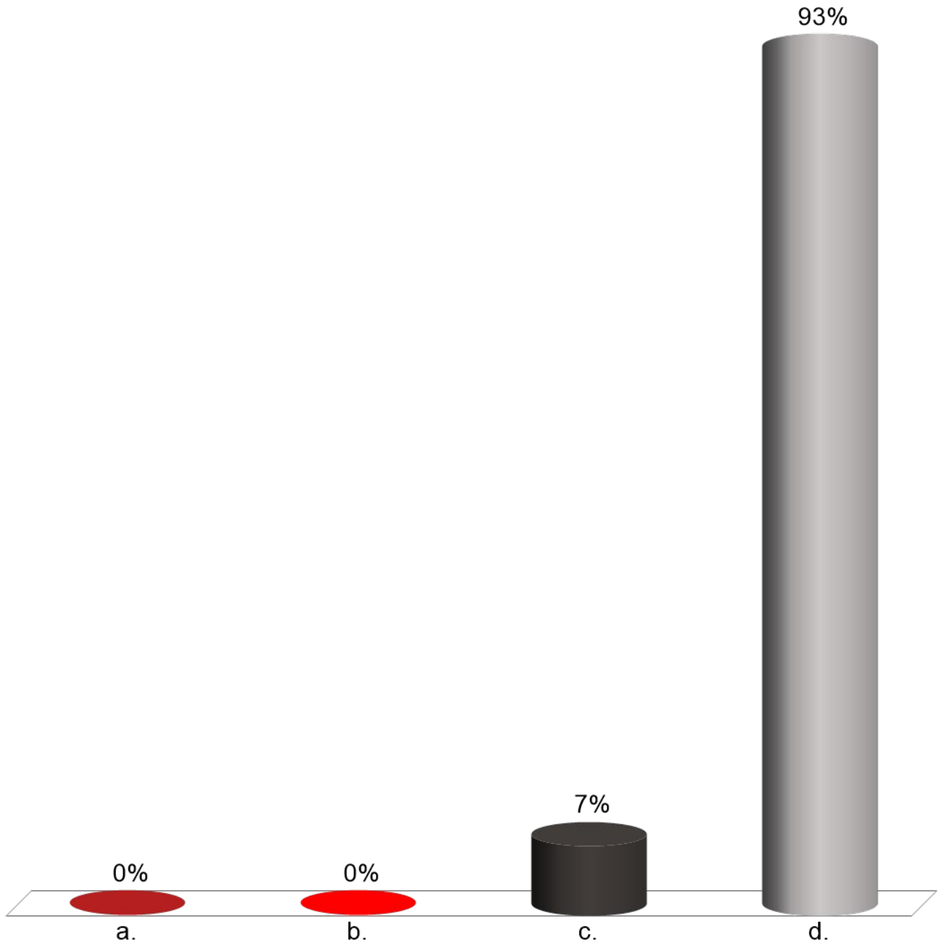
In this course, we consider that being responsible as a software engineer means:

❌ a. Making sure a liability clause is included in the software license agreement.

❌ b. Reacting rapidly to correct software bugs when they are reported.

❌ c. Being accountable for the decisions made by the development team.

✅ d. Anticipating the potential negative impacts of the software on others.



Types of issues (1/2)

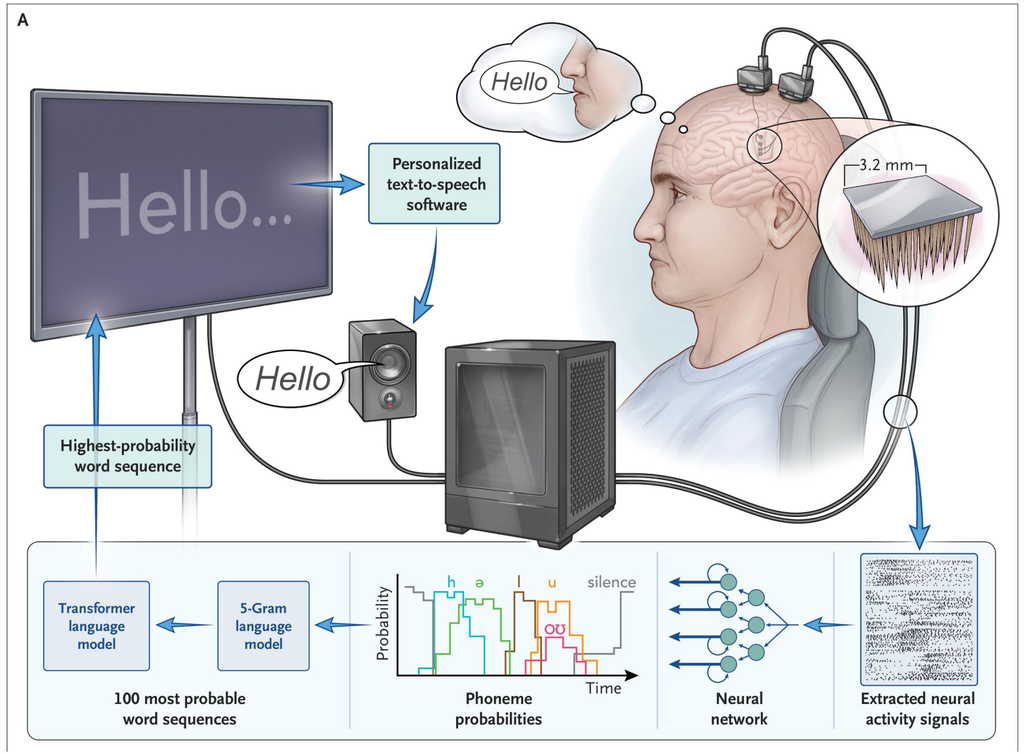
(Card et al., 2024)

Brain-to-speech software can translate neural activity associated with attempted speech into spoken words. A key challenge is ensuring that only intentional communication is captured, not private inner thoughts.

This is:

- ✅ 37% a. A technical issue
- ✅ 52% b. An ethical issue
- 11% c. An ethical dilemma

The description does not directly reflect a dilemma. But we can see a dilemma between developing the software to help people with speech impairment (inclusion) vs. not capturing their private thoughts (privacy)



Types of issues (2/2)

(Urbina et al., 2022)

A software company has developed a Machine Learning model that is able to discover new chemical compounds for medicine development. They identify that the model can also discover new chemical weapons.

This is:

- ❌ 3% a. A technical issue
- 20% b. An ethical issue
- ✅ 77% c. An ethical dilemma

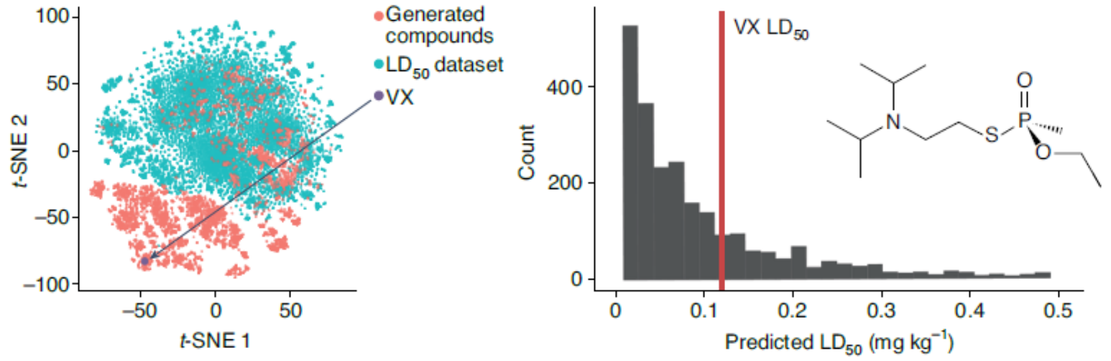
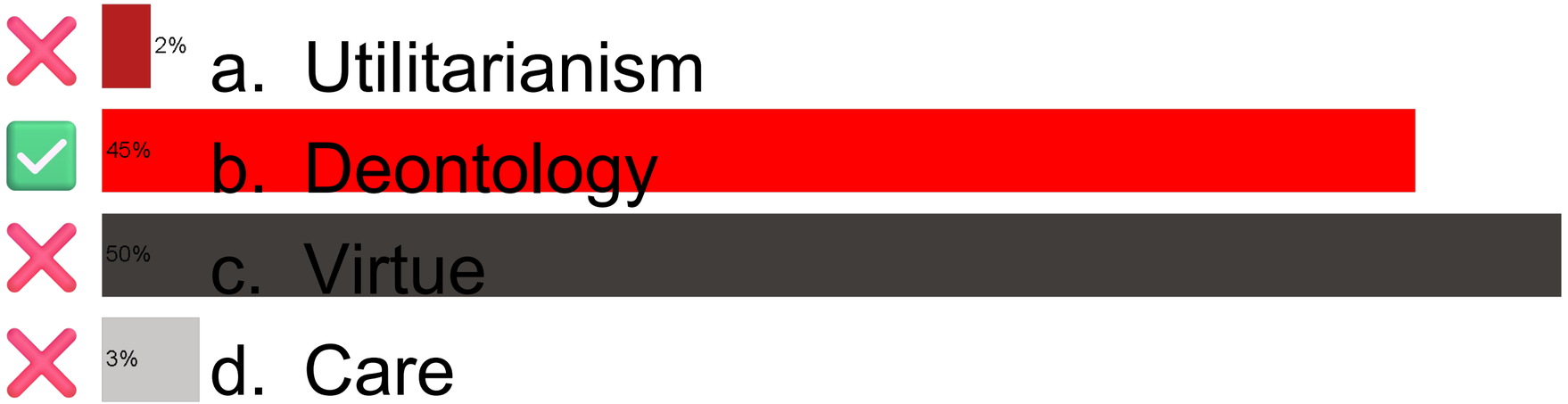


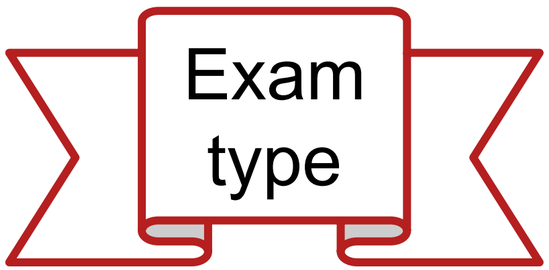
Fig. 1 | A *t*-SNE plot visualization of the LD_{50} dataset and top 2,000 MegaSyn AI-generated and predicted toxic molecules illustrating VX. Many of the molecules generated are predicted to be more toxic in vivo in the animal model than VX (histogram at right shows cut-off for VX LD_{50}). The 2D chemical structure of VX is shown on the right.

This is clearly an ethical dilemma (see next slide).
If it is an ethical dilemma then there are two underlying ethical issues, so one could argue for b potentially (but this is a bit far stretched).

Normative ethical theories (1/2)

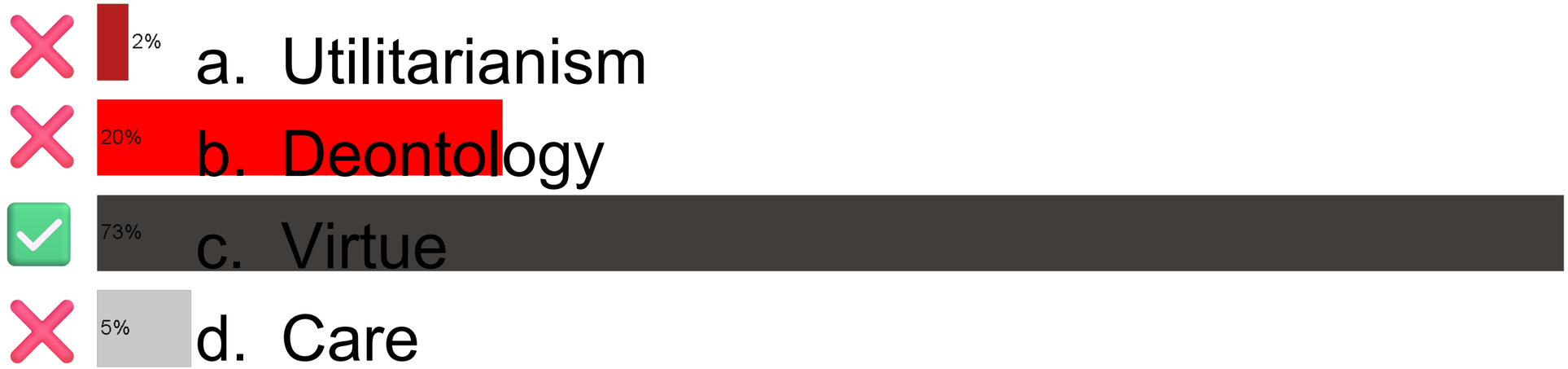
A software engineer refuses to hide a critical bug in a released product and tells you: “I believe that it is always wrong to lie, even if telling the truth might result in harm to some people.” Which ethical theory does this engineer follow?

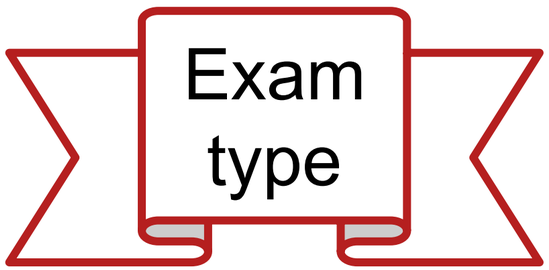




Normative ethical theories (2/2)

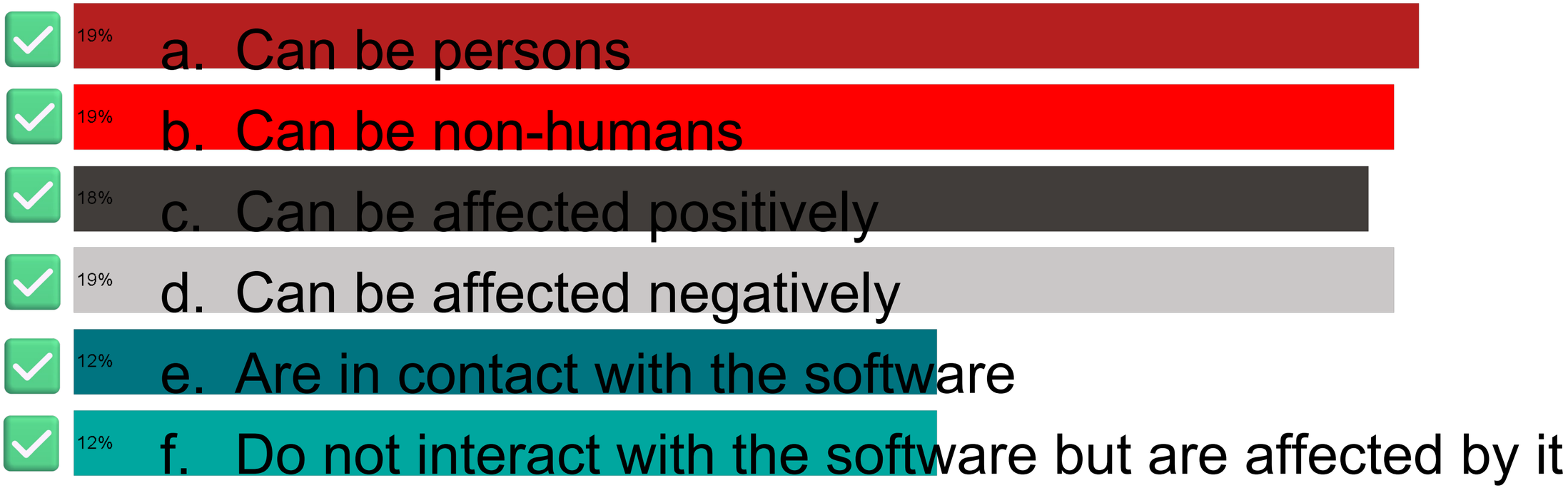
A software engineer refuses to hide a critical bug in a released product and tells you: “If I do not report this bug, I am not being a trustworthy and courageous person.” Which ethical theory does this engineer follow?





Stakeholders analysis

Which of the following statements are true about stakeholders?



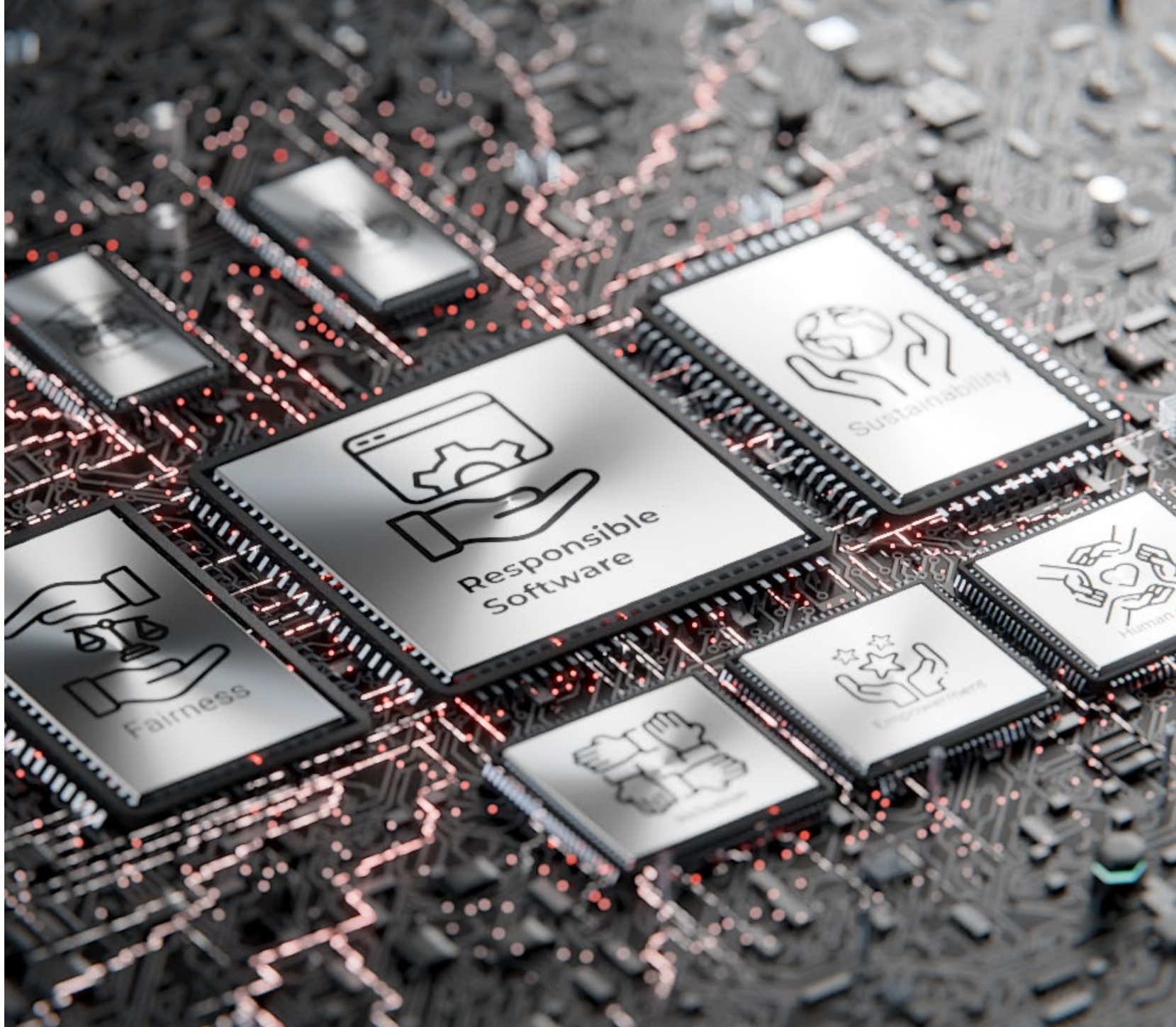
EPFL

Safety 1 Review & Case Studies

22 sept.

Cécile Hardebolle

Responsible Software

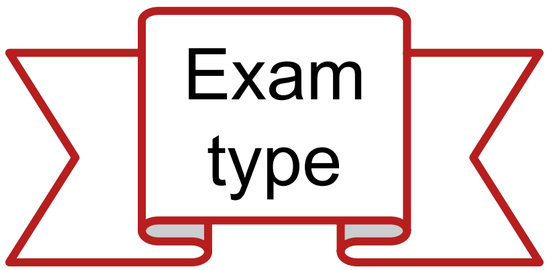


Autonomous car software - 1

The software of an autonomous car has a 10% error rate in recognizing traffic signs correctly.
We are in the presence of (select all that apply):

- ❌ 0% a. A safety threat
- ❌ 0% b. A security threat
- ✅ 0% c. A safety hazard
- ❌ 0% d. A security hazard

URL: ttpoll.eu
Session ID: cs290



Autonomous car software - 2

Stickers placed on a stop sign lead the software of an autonomous car to misclassify it as a speed limit sign.
We are in the presence of (select all that apply):

- ❌ 0% a. A safety threat
- ✅ 0% b. A security threat
- ✅ 0% c. A safety hazard
- ❌ 0% d. A security hazard

The stickers affect the system negatively
(e.g. may have been placed by bad actors)

URL: ttpoll.eu
Session ID: cs290

Worldwide “CrowdStrike” outage in 2024

This event is an example of:

- ✅ 0% a. Malfunction
- ❌ 0% b. Misuse, abuse
- ❌ 0% c. Unintended use
- ❌ 0% d. Intended use

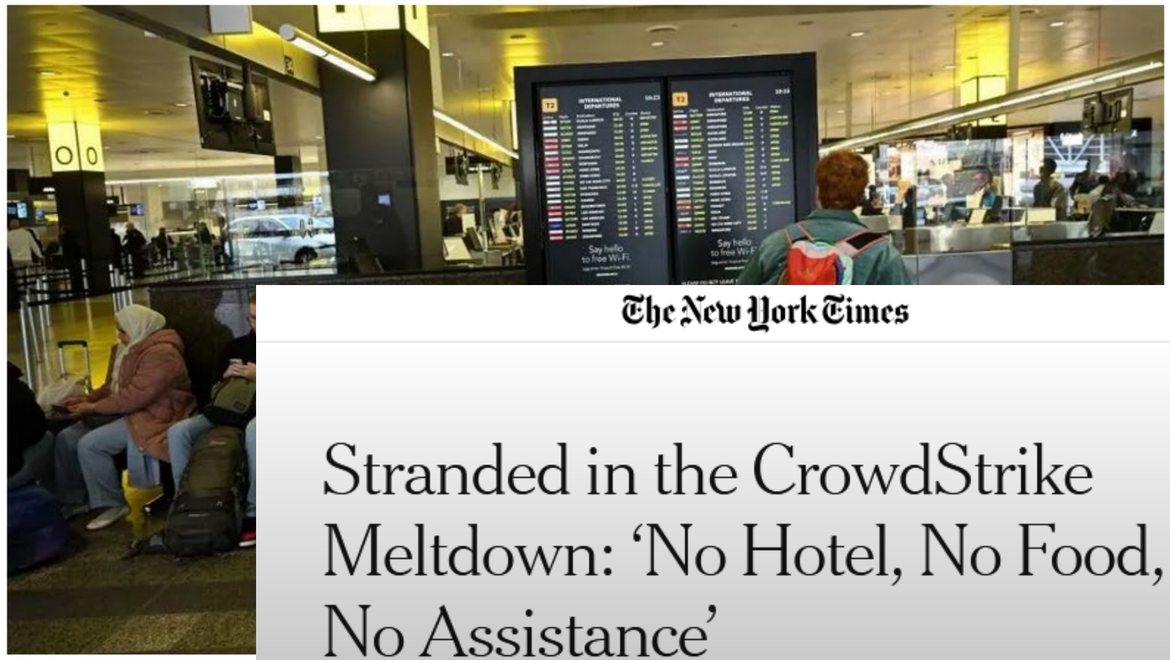
URL: ttpoll.eu
Session ID: cs290

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

❌ 0% a. Bad actors generate safety issues only

❌ 0% b. Bad actors generate security issues only

✅ 0% 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):

- 0% a. Malfunction — Yes, if we consider that a bad actor can lead a software to malfunction
- ✅ 0% b. Misuse, abuse — **Mostly misuse & abuse**
- ✅ 0% c. Unintended use
- ✅ 0% d. Intended use

Bad actors can exploit intended features (e.g. recommendation algorithms) or use products in unintended ways

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:

- ✅ 0% a. TN = actual absence of fissure, correct prediction
- ❌ 0% b. TP = actual absence of fissure, correct prediction
- ✅ 0% c. FN = actual presence of fissure, incorrect prediction
- ❌ 0% 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:

❌ 0% a. TN

0% b. TP

✅ 0% c. FN

❌ 0% d. FP

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

<u>What is the False Negative Rate (FNR)?</u>

- ❌ 0% a. 13%
- ❌ 0% b. 17%
- ✅ 0% c. 20%
- ❌ 0% d. 25%

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

Harm categories - 1

A user sees their post unfairly censored.
This harm is in the category (select one):

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

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):

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

Difficult to categorize:

- A human is harmed, we can extrapolate that there is psychological damage
- There is also financial damage for the person, but we are not allocating resources
- There is manipulation of behavior, we can say it harms social systems but it's a bit of a stretch (impact on citizenry unclear)

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):

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

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

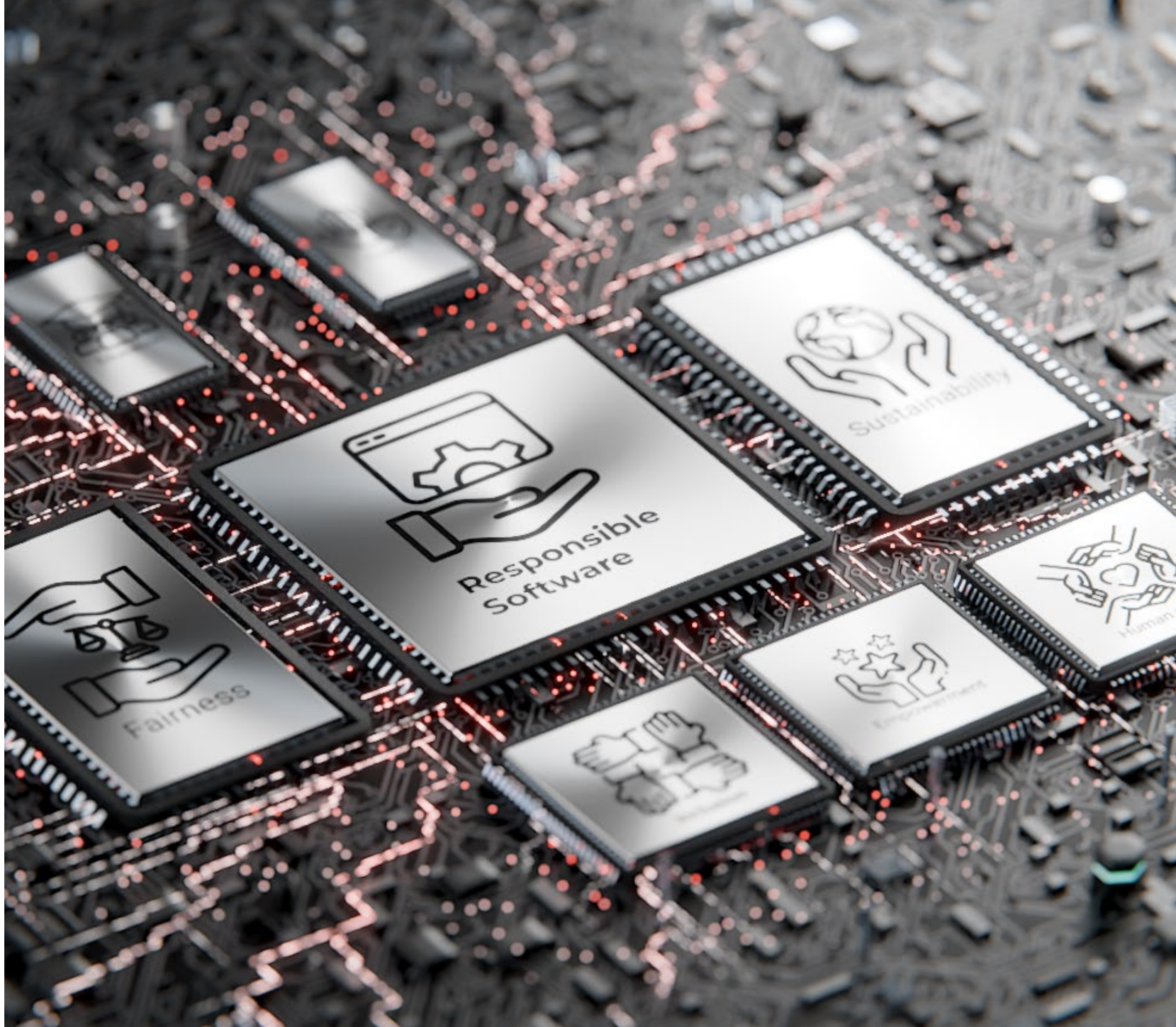
URL: ttpoll.eu
Session ID: cs290

Safety 2 Review & Case studies

29 sept.

Cécile Hardebolle

Responsible Software



Macro-level perspective

URL: ttpoll.eu
Session ID: cs290

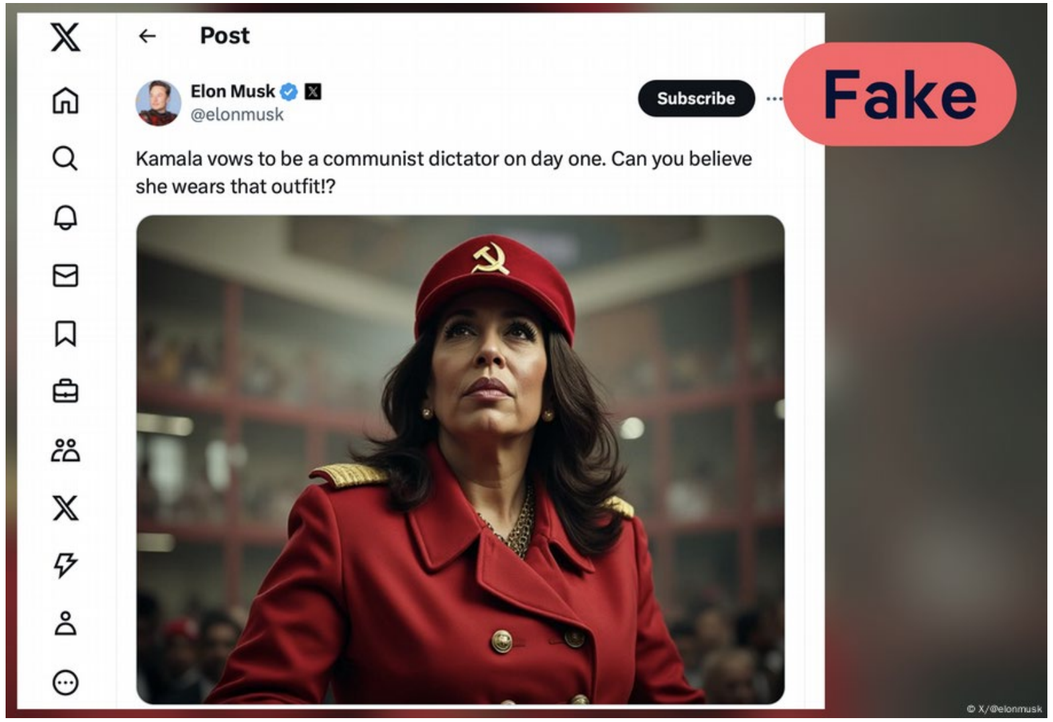
A macro-level perspective is useful (select all correct statements):

- ✅ 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

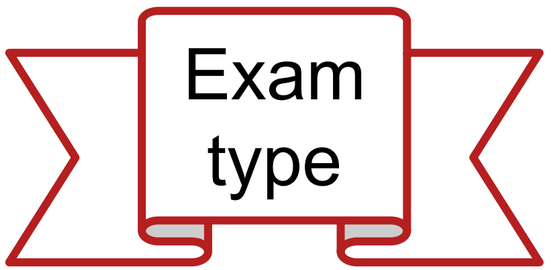
Exam type

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):

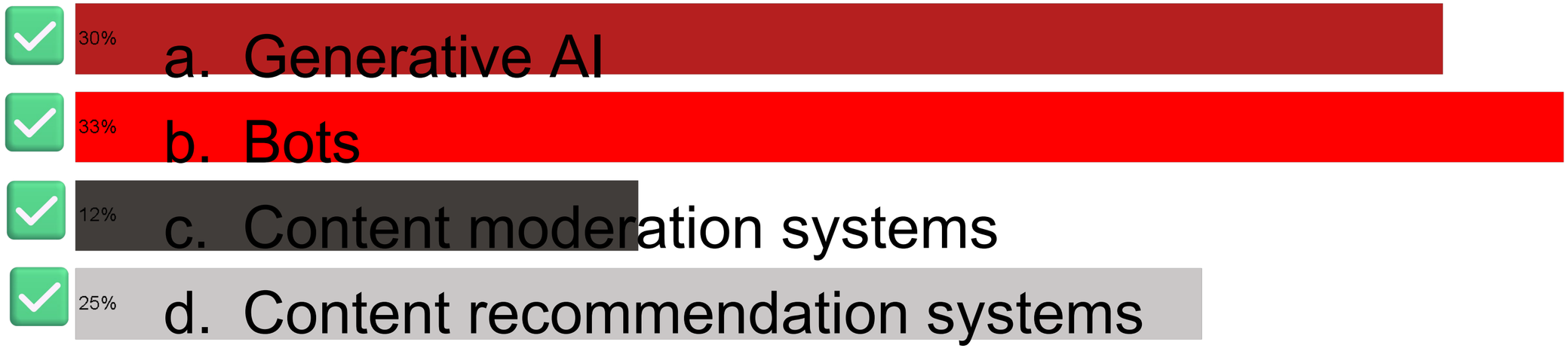
- False
- No intention to harm
- Not in the form of news (e.g. press article)



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

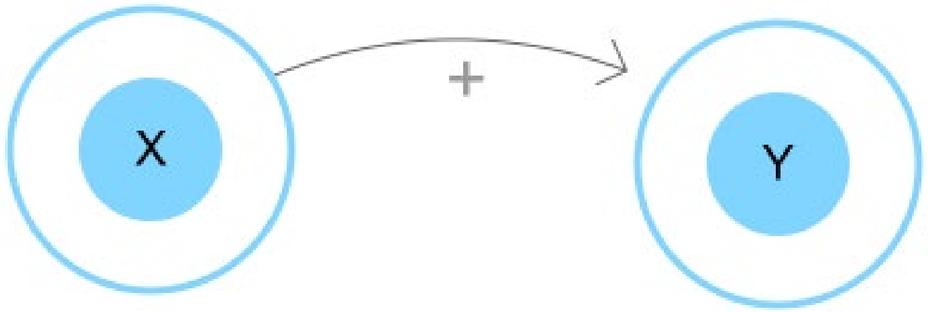
URL: ttpoll.eu
Session ID: cs290

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)

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 “+”

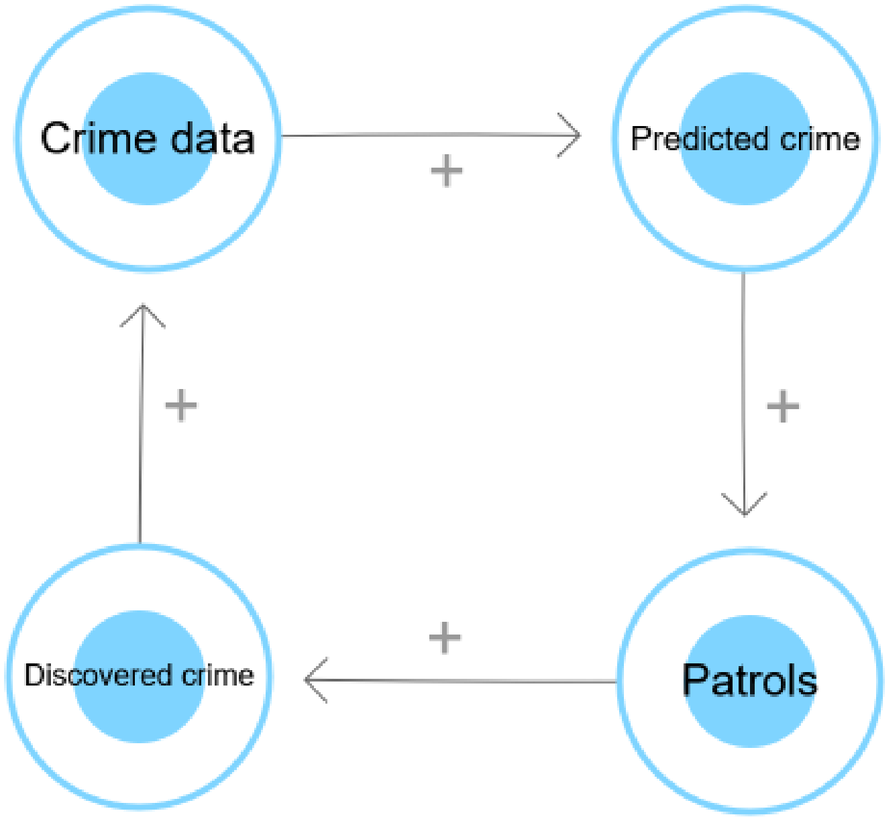
❌ 11% b. The quantity in X is added to the quantity in Y

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

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

Part 1: behavior

URL: ttpoll.eu
Session ID: cs290



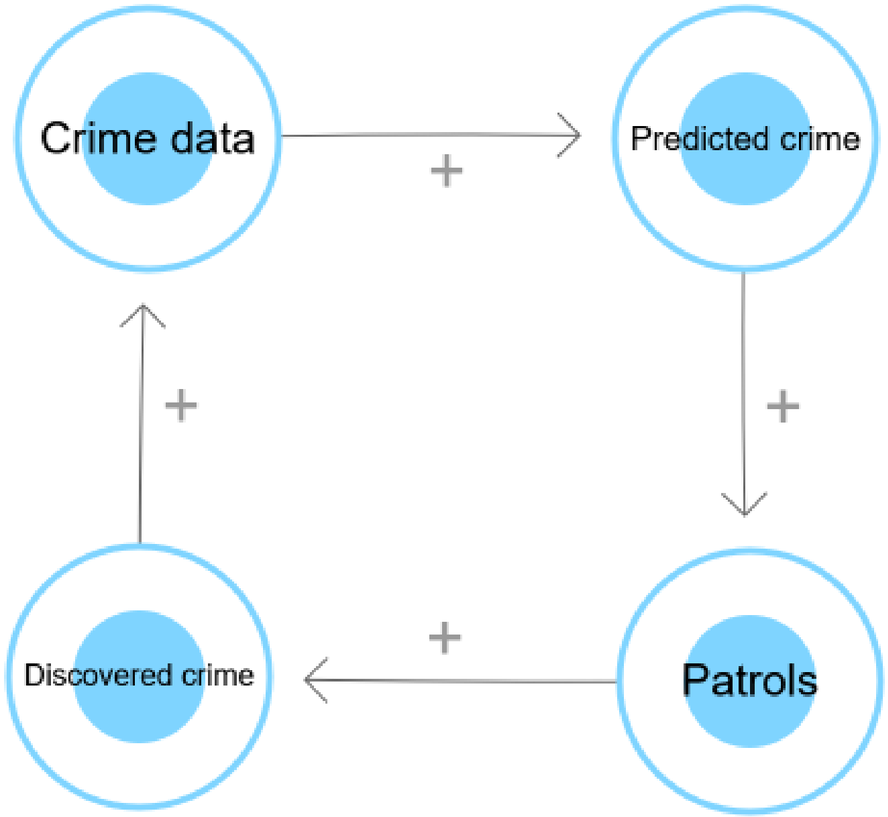
Over time, the quantities in this system will:

- ✗ 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



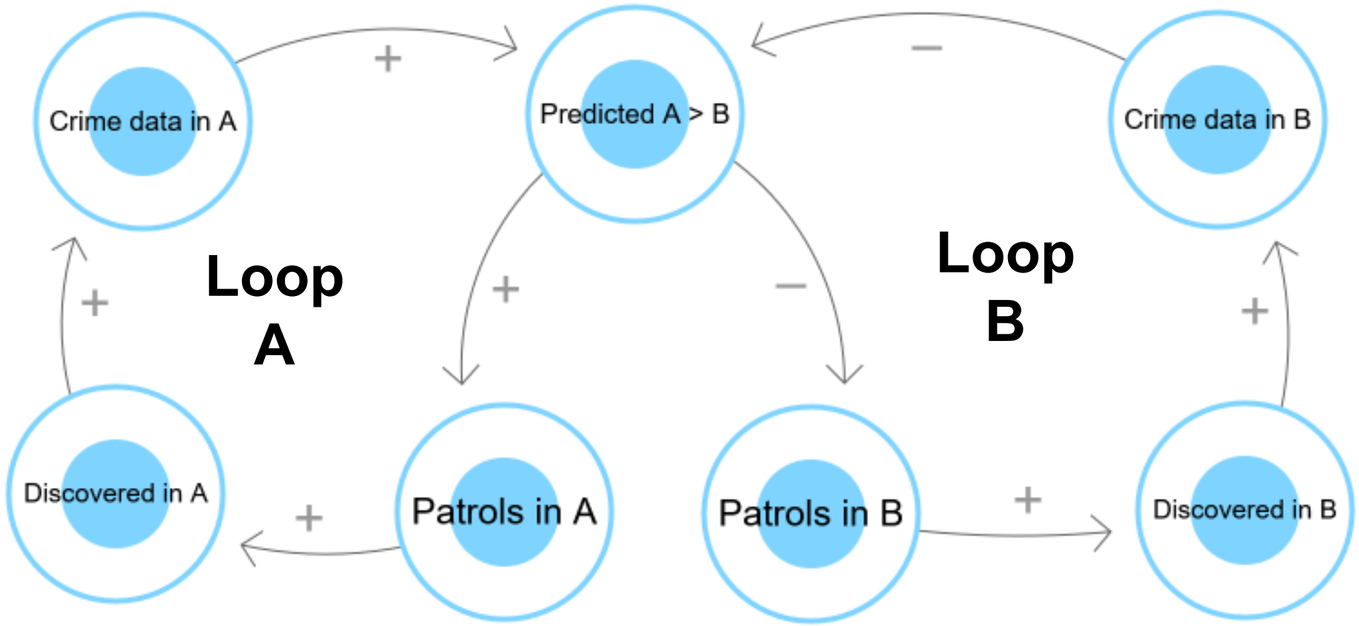
The feedback loop in this diagram is:

❌ 0% a. Balancing

✅ 100% b. Reinforcing

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 ✅

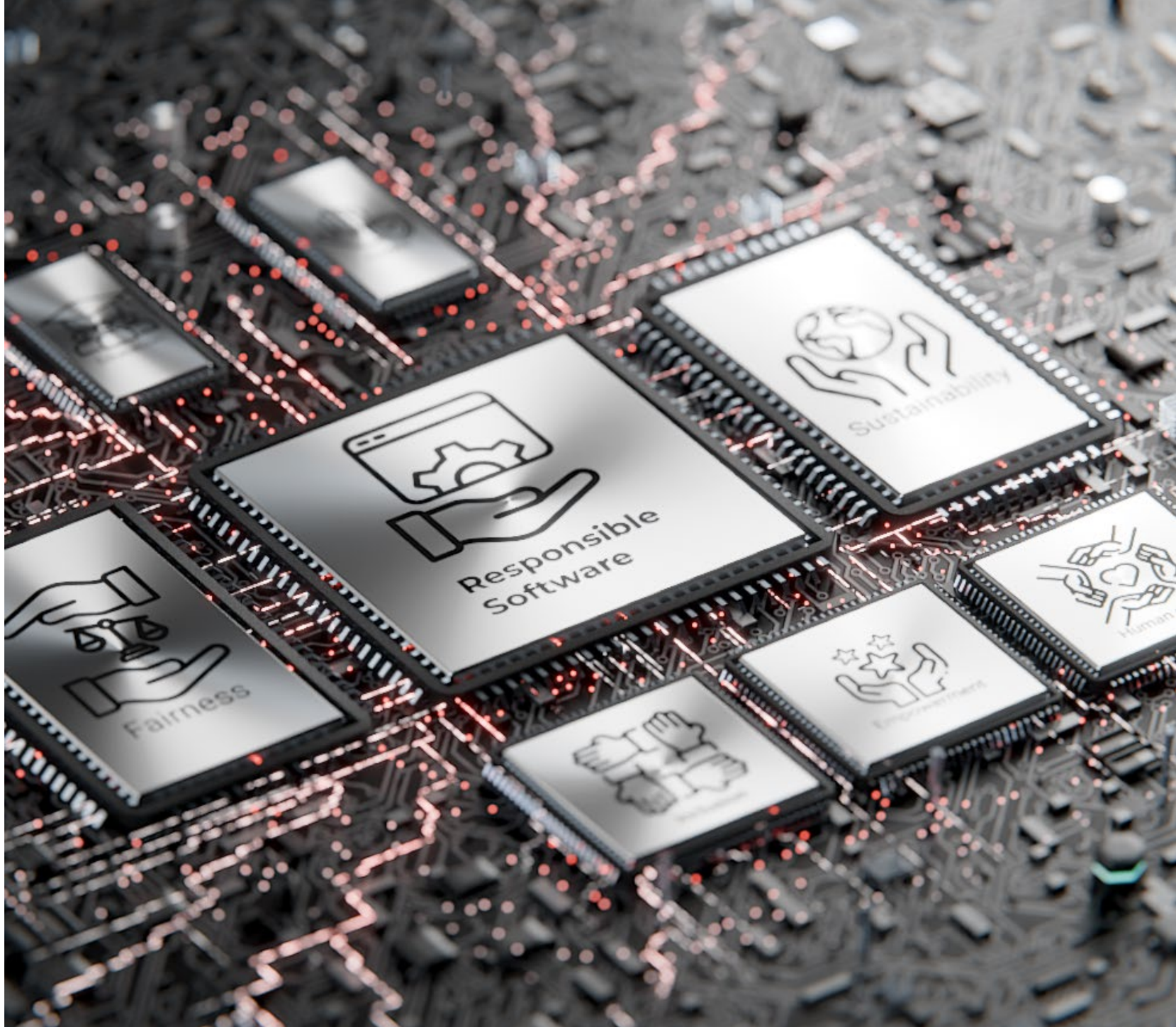
EPFL

Fairness 1 Review & Case studies

7 oct.

Cécile Hardebolle

Responsible Software



Attributes - 1

URL: ttpoll.eu
Session ID: cs290

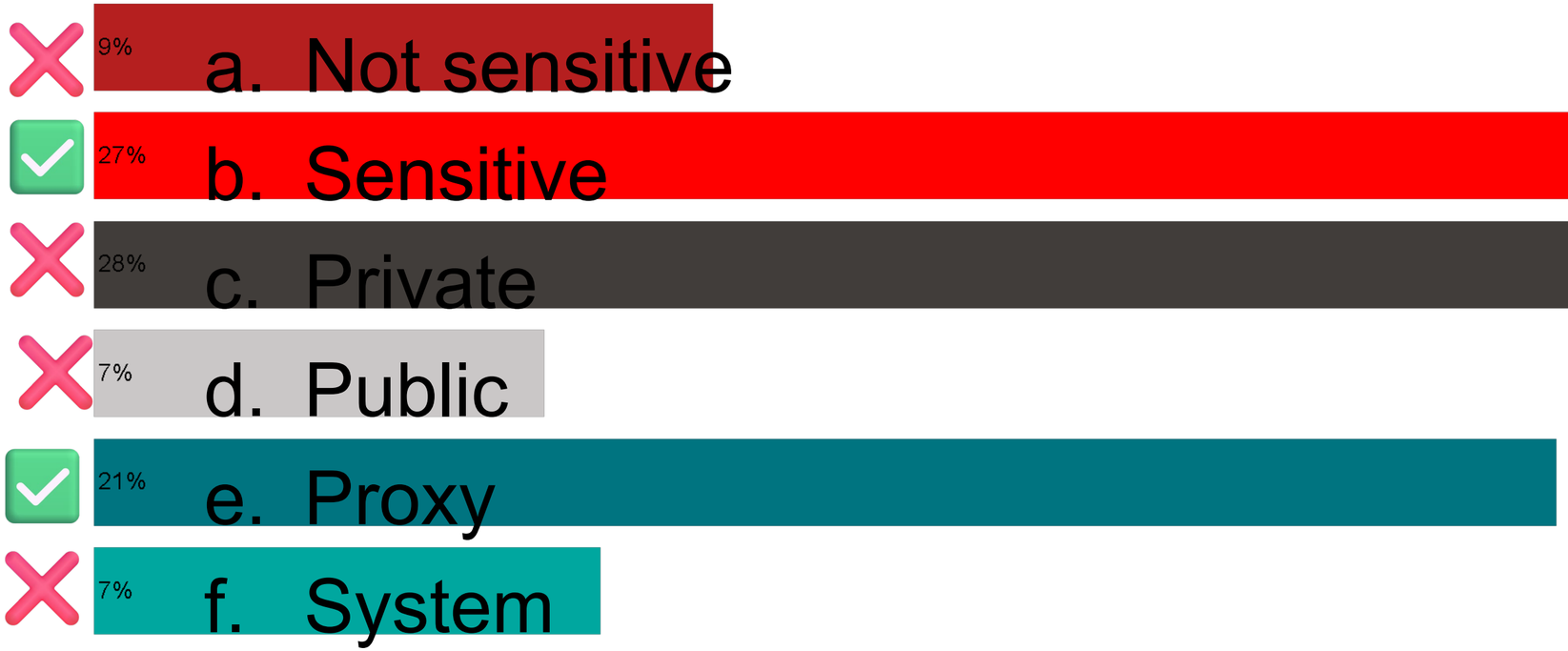
What are the characteristics of hair color as an attribute to represent people? (select all that apply)

What someone sees as “red” can be described as “auburn” by someone else, we would need to use a set of predefined categories, and some cases would be difficult to fit in (there is the same issue with skin color by the way)
-> the only way to make it “objective” would be to measure the color with colorimetry

Attributes - 2

URL: ttpoll.eu
Session ID: cs290

Let’s imagine a software that relies on SAT scores (standardized test for university admission in the US) to make recommendations of when to approve study loans.
What are the characteristics of the SAT score?



Because it is a proxy for sensitive variables, SAT score can lead to discrimination i.e. can be considered sensitive.

Article on SAT scores’ correlation with race: https://www.brookings.edu/articles/race-gaps-in-sat-scores-highlight-inequality-and-hinder-upward-mobility/

Bias - 1

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.

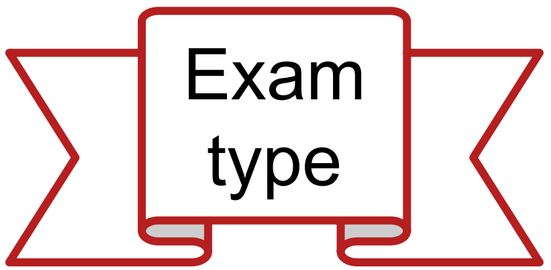
What bias will the data collected by the app probably exhibit?
(select one answer)

❌ 0% a. Confirmation bias

✅ 53% b. Representation bias

❌ 47% c. Measurement bias

❌ 0% d. Automation bias



Bias - 2

URL: ttpoll.eu
Session ID: cs290

In the new ArcFit fitness tracker, the calory burn feature uses the "metabolic equivalent of task" formula, which estimates the energy a body uses during a specific activity. The same calculation is used during walking and running.

What type of bias will the calory burn variable probably have?
(select one answer)

- ✗ 0% a. Confirmation bias
- ✗ 2% b. Representation bias
- ✓ 98% c. Measurement bias
- ✗ 0% d. Automation bias

Exam type

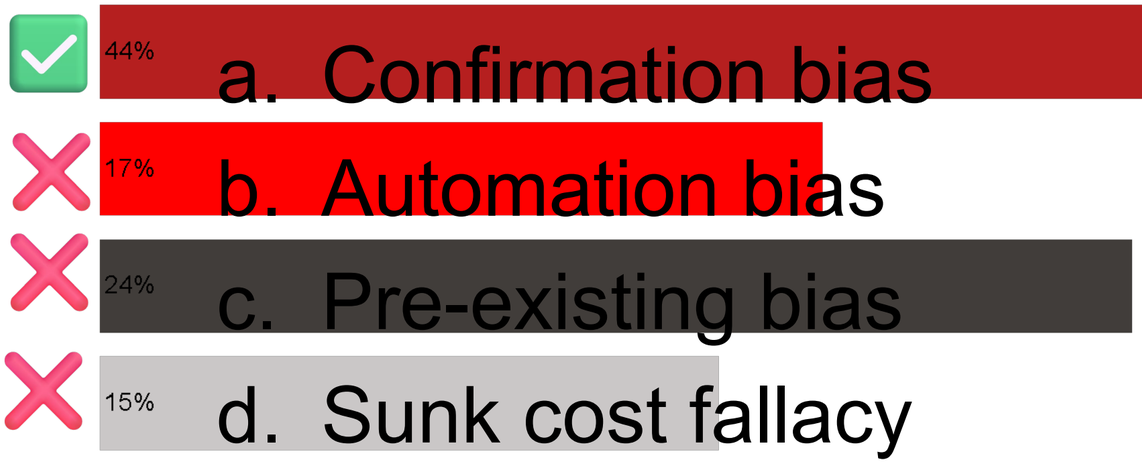
Bias - 3

URL: ttpoll.eu
Session ID: cs290

A group of computer scientists with similar background, all experts in software development, are starting a new software project in the healthcare domain.

The question has been reframed compared to the in-class session to make clearer:
- they are not using AI to develop (was unclear)
- we do not consider here biases from the data but biases from humans

What type of bias will these scientists probably have?
(select one answer)



Exam type

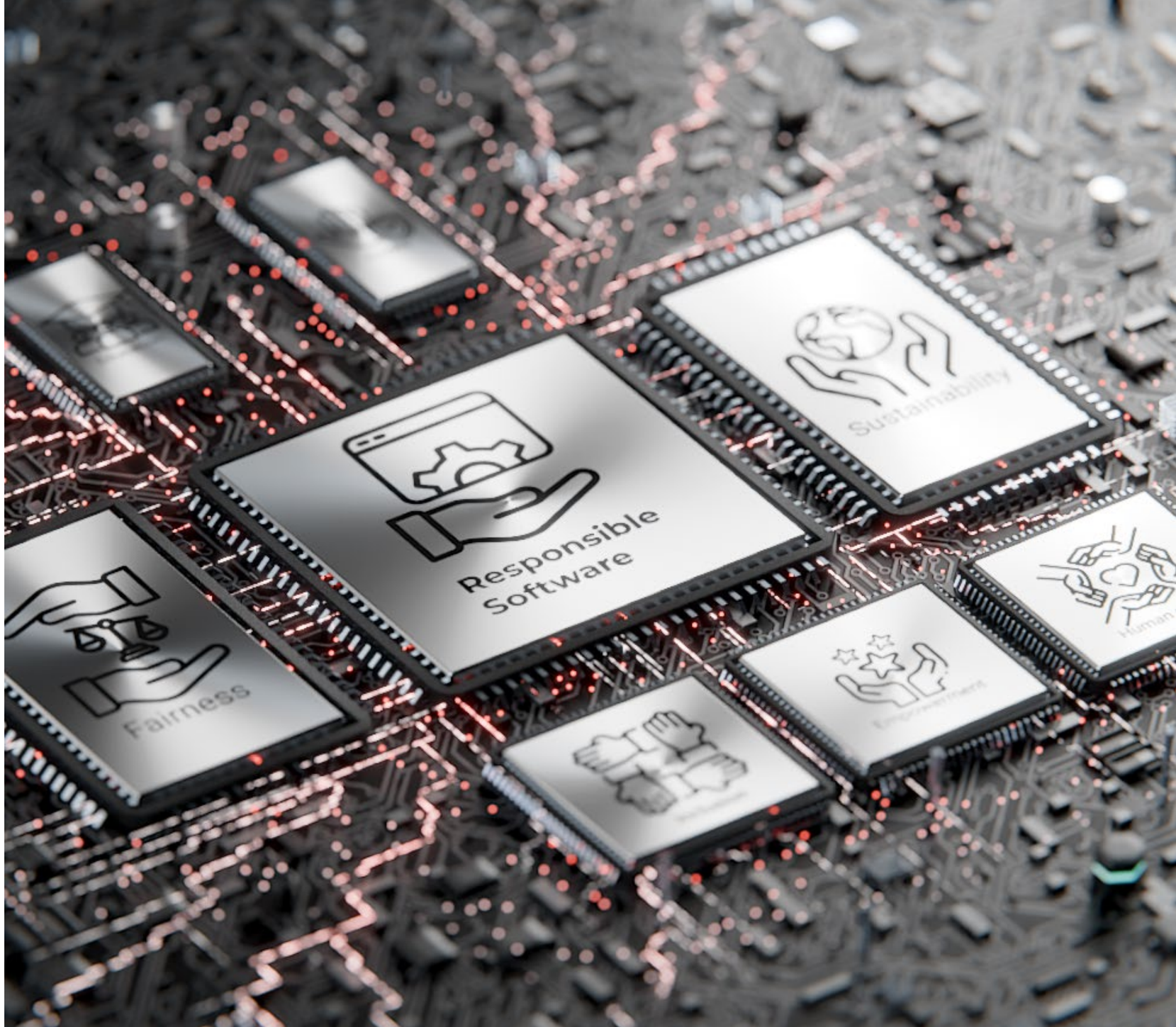
EPFL

Fairness 2 Review & Case studies

13 oct.

Cécile Hardebolle

Responsible Software



Biases in the ML lifecycle - 1

URL: ttpoll.eu
Session ID: cs290

Simpson's paradox is when the patterns observed at the level of the full sample and at the level of subgroups are opposed.

When training a ML model, Simpson's paradox can lead to (select 1 answer):

- Training time
- Pattern at aggregated level is different from patterns for subgroups

❌ 25% a. Evaluation bias

✅ 25% b. Aggregation bias

❌ 25% c. Optimization choices

❌ 25% d. Deployment bias

3.4 Aggregation Bias

Aggregation bias arises when a one-size-fits-all model is used for data in which there are underlying groups or types of examples that should be considered differently. Underlying aggregation bias is an assumption that the mapping from inputs to labels is consistent across subsets of the data. In reality, this is often not the case. A particular dataset might represent people or groups with different backgrounds, cultures or norms, and a given variable can mean something quite different across them. Aggregation bias can lead to a model that is not optimal for any group, or a model that is fit to the dominant population (e.g., if there is also representation bias).

Biases in the ML lifecycle - 2

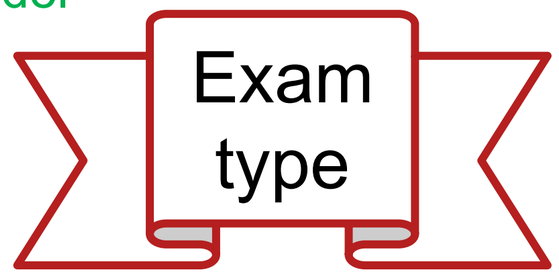
URL: ttpoll.eu
Session ID: cs290

The society RetailProtect has developed a ML model to identify instances of shoplifting in retail shops. For evaluating their model, they use a benchmark in which actors from diverse ethnicities simulate a range of shoplifting actions.

This can lead to (select 1 answer):

- ✅ 0% a. Evaluation bias
- ❌ 0% b. Aggregation bias
- ❌ 0% c. Optimization choices
- ❌ 0% d. Deployment bias

- Evaluation time
- Diverse ethnicities does not guaranty fairness on other attributes (e.g. gender, etc.)
- The benchmark employs **actors** that **simulate** shoplifting instead of real-life scenes -> actions will probably be exaggerated/different from real cases i.e. they will not evaluate correctly the performance of the model



Fairness metrics - 1

URL: ttpoll.eu
Session ID: cs290

Among the metrics below, **which can be used to assess the fairness** of a piece of software? (select all that apply)

0% a. Accuracy

0% b. False Positive Rate

0% c. False Negative Rate

0% d. False Discovery Rate

0% e. False Omission Rate

0% f. Positive Predictive Value

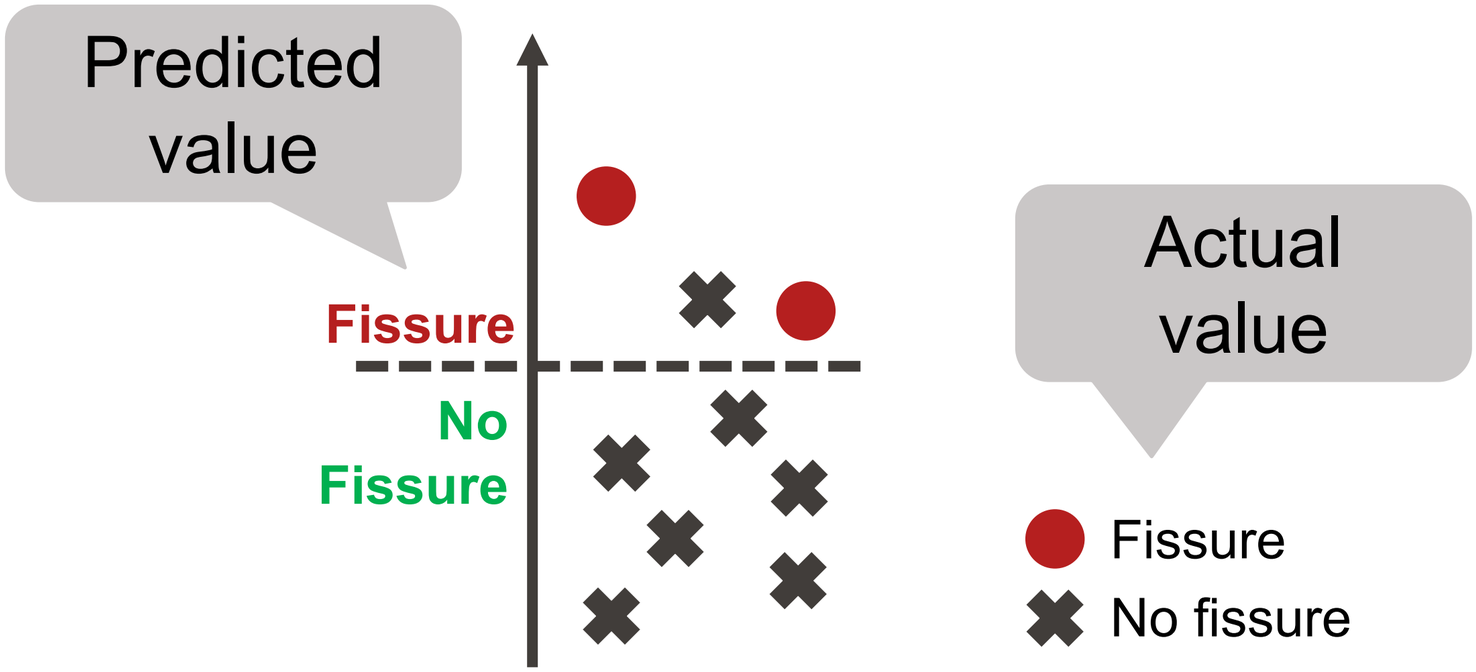
0% g. Negative Predictive Value

0% h. Proportion of positive prediction (also called acceptance rate)

All can be used as long as we compare 2 groups with it

Fairness metrics – 2

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:

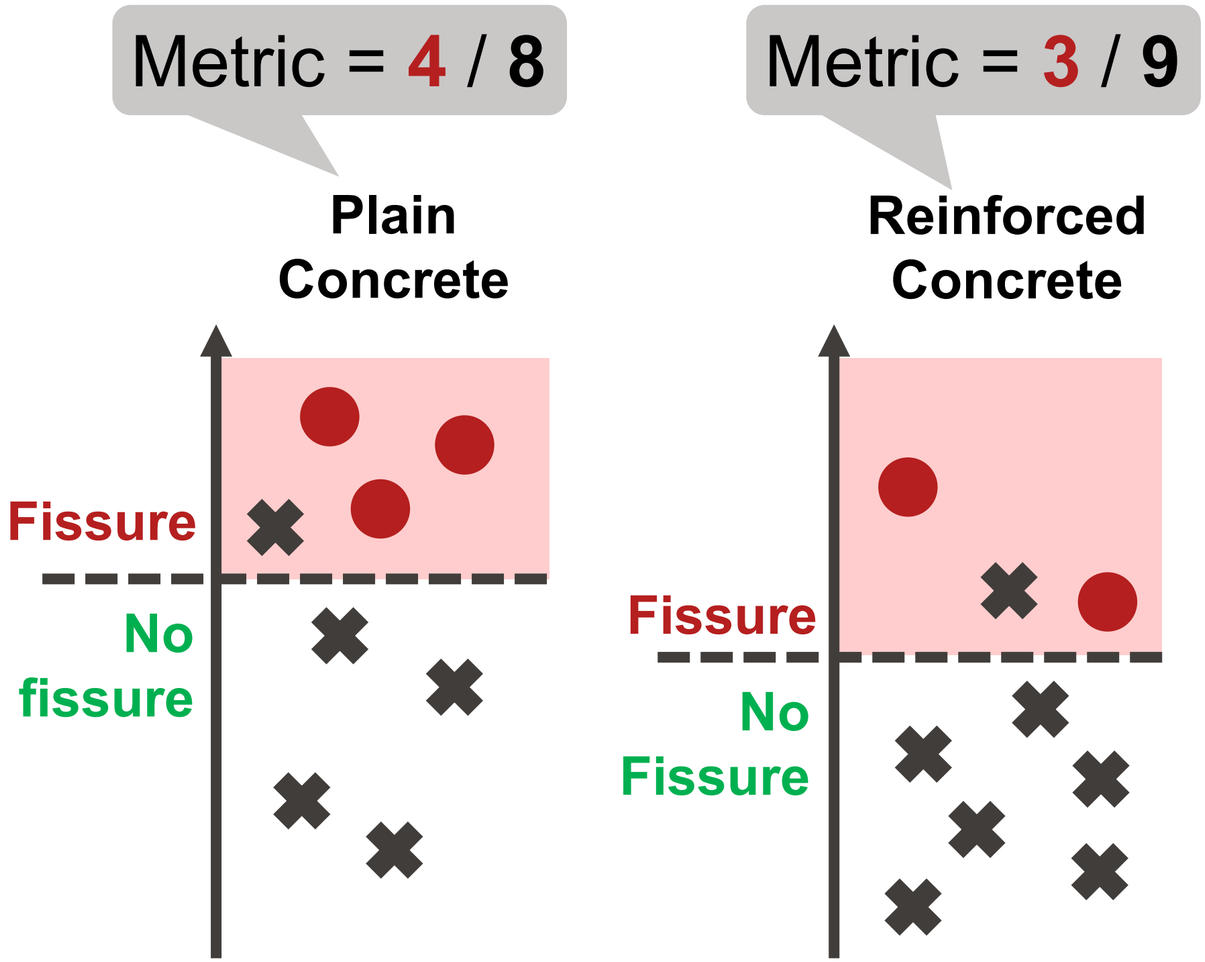


		Predicted: Fissure	Predicted: No Fissure
Actual	Fissure	2	0
Actual	No Fissure	1	6

Fairness metrics – 2a

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:



Which notion of fairness are they using?
(select 1 answer)

❌ 0%
a. Equal accuracy

❌ 0%
b. Error rate balance

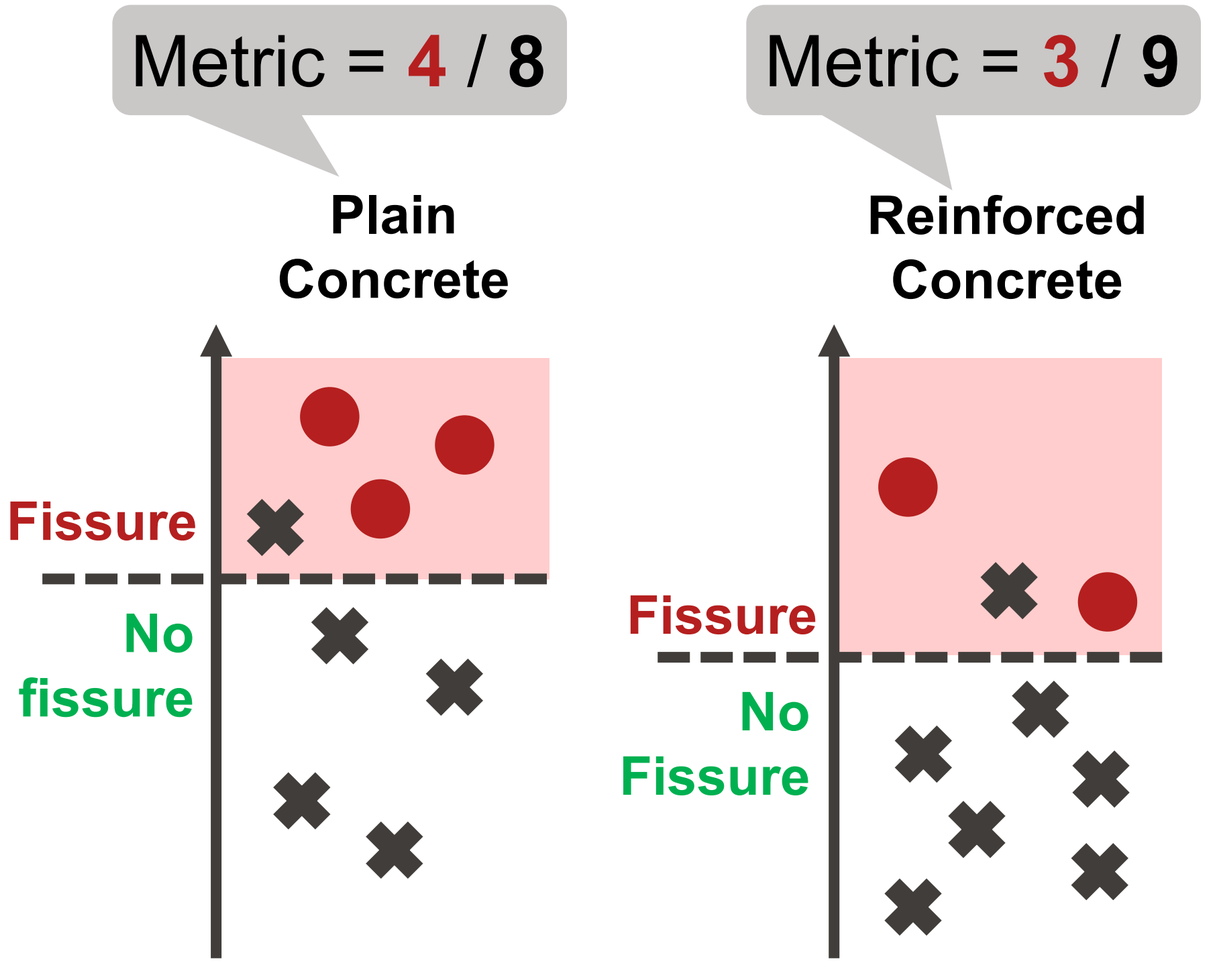
❌ 0%
c. Error parity

✅ 0%
d. Demographic parity

They compare the number of positive predictions (fissure) / total number of samples

Fairness metrics – 2b

URL: ttpoll.eu
Session ID: cs290



According to this metric, is their model fair?
(select 1 answer)

❌ 0% a. Yes

✅ 0% b. No

(✅) 0% c. Other option

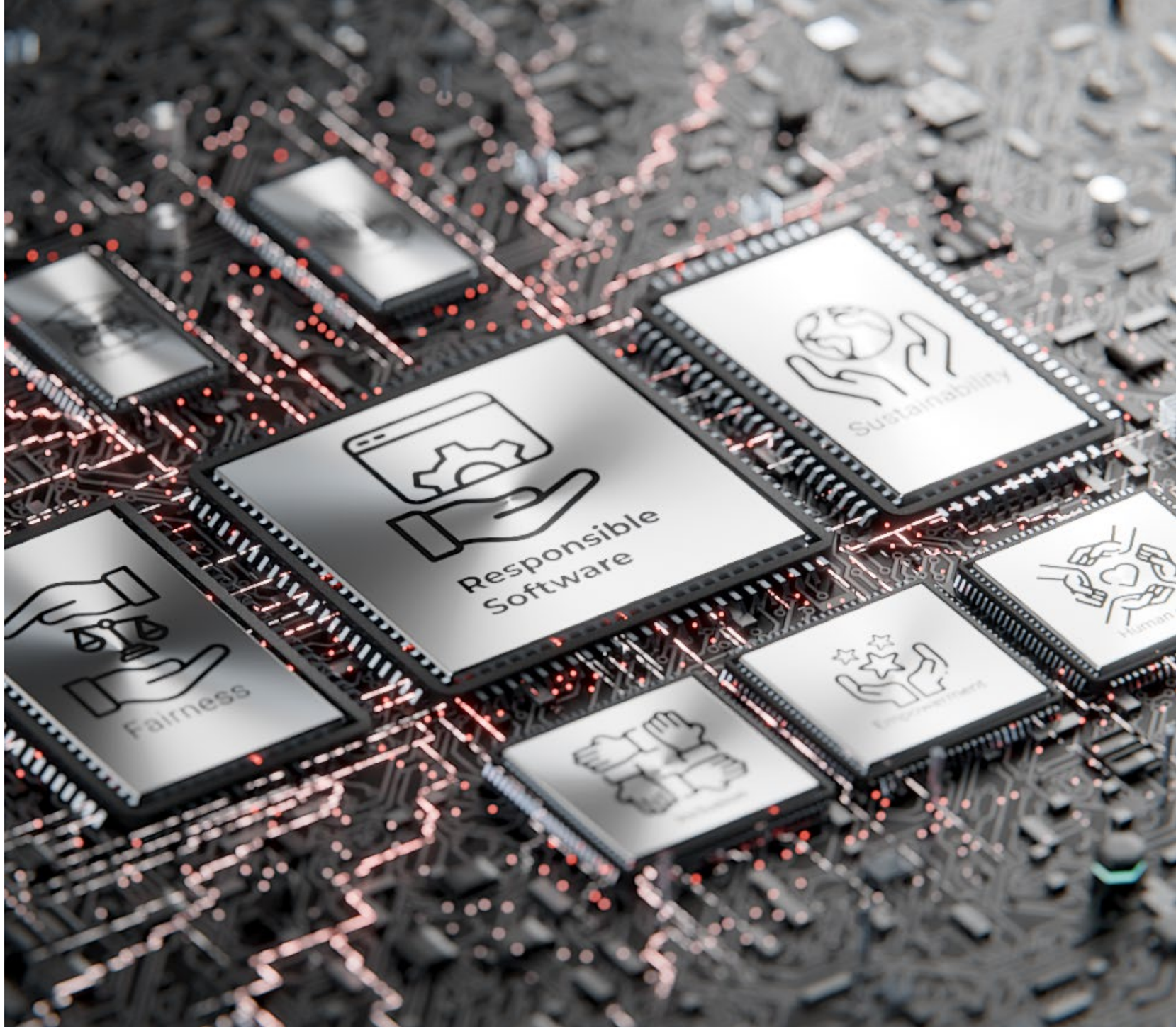
- Disparate impact ratio = 0,33 / 0,5 = 0,66
Which is far from 1 or from the tolerated 0,8
- We can question whether it is really about "fairness" in this case…

Sustainability 1 Review & Case studies

10 nov.

Cécile Hardebolle

Responsible Software

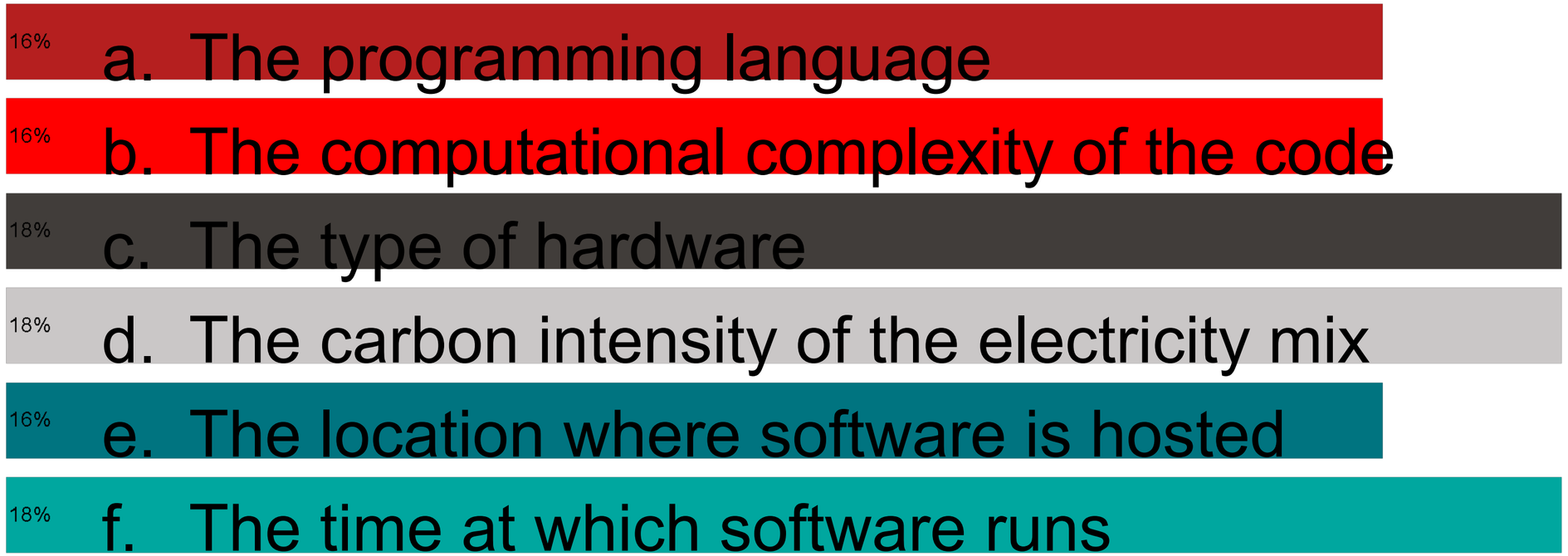


Carbon footprint factors

URL: ttpoll.eu
Session ID: cs290

What are (some of) the factors in the carbon footprint of software? (select all that apply)

All of these!



CO_2 equivalent

URL: ttpoll.eu
Session ID: cs290

An electricity production facility reports the following emissions per kWh produced:

- 250 g of carbon dioxide (CO2)
- 8 g of fossil methane (CH4)

What are the carbon emissions of the facility in g CO_2 eq / kWh (considering the GWP-100)?

Fossil methane has a GWP-100 = 30
=> 250 + (8 x 30) = 490

❌ 4% a. 240 g / kWh

❌ 12% b. 258 g / kWh

✅ 65% c. 490 g / kWh

❌ 12% d. 656 g / kWh

❌ 8% e. 906 g / kWh

Power Usage Effectiveness

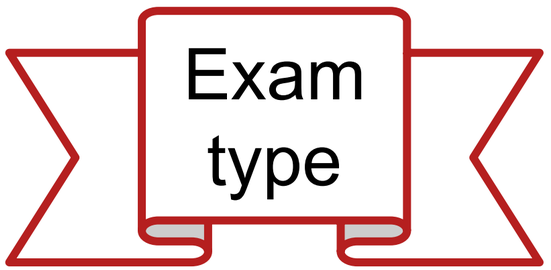
URL: ttpoll.eu
Session ID: cs290

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?

- ✗ 22% a. 0.5
- ✗ 0% b. 1
- ✗ 11% c. 1.5
- ✓ 67% d. 2

PUE = total electricity / electricity used by IT
PUE is always >= 1

Here: PUE = 8760 / (8760 * 0,5) = 2



Scopes in the GHG protocol

URL: ttpoll.eu
Session ID: cs290

For a software development company, the **electricity consumed** by software during the **development phase** falls into:

❌ 27% a. Scope 1 (direct)

✅ 40% b. Scope 2 (indirect, energy)

(✅) 13% c. Scope 3 (indirect, value chain)

(✅) 20% d. It depends

- It cannot be Scope 1 unless the software dev company produces its own electricity on site (very rare)
- If the development is done on machines hosted by the development company, the electricity used during development is bought by the company to an Electricity Provider => Scope 2
- If the development is done on a cloud platform hosted by another company, the electricity used during development is bought by a Cloud Provider, in which case it is considered coming from the value chain => Scope 3

Direct stakeholders

URL: ttpoll.eu
Session ID: cs290

Which of the following stakeholders can be considered **direct** stakeholders in the case:

27% a. Internal IT employees working on IT infrastructure
19% b. Corporate clients using the center to provide applications
8% c. Users of applications hosted by the center
24% d. Companies providing energy to the center
11% e. Local population in the area of the center
11% f. Local ecosystems in the area of the center

Direct

Direct or indirect

Indirect

The line between direct and indirect is very fine/blurred, sometimes it is hard to tell -> **argument** is important

Rebound effect

URL: ttpoll.eu
Session ID: cs290

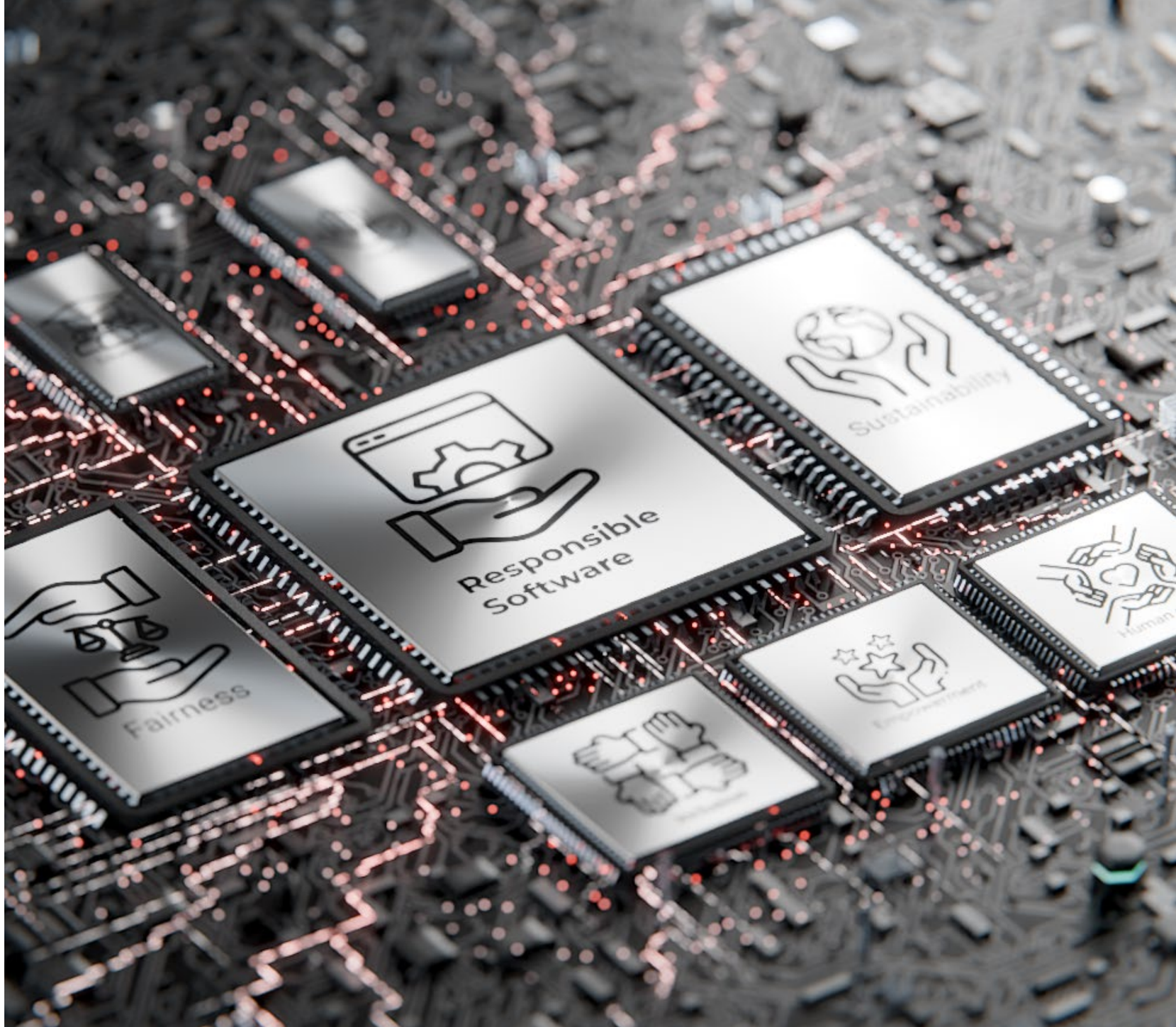
The rebound effect is when **higher energy efficiency** in a product (i.e. lower energy consumption from use) leads to an **increase of total energy consumption** because:
(select all that apply)

- ✅ 45% a. The demand for the product increases
- ❌ 0% b. The demand for the product decreases
- ✅ 39% c. The product is used more often
- ❌ 0% d. The product is used less often
- ✅ 7% e. The consumption of other products increases
- ❌ 9% f. The consumption of other products decreases

Sustainability 2
Review & Case studies
17 nov.

Cécile Hardebolle

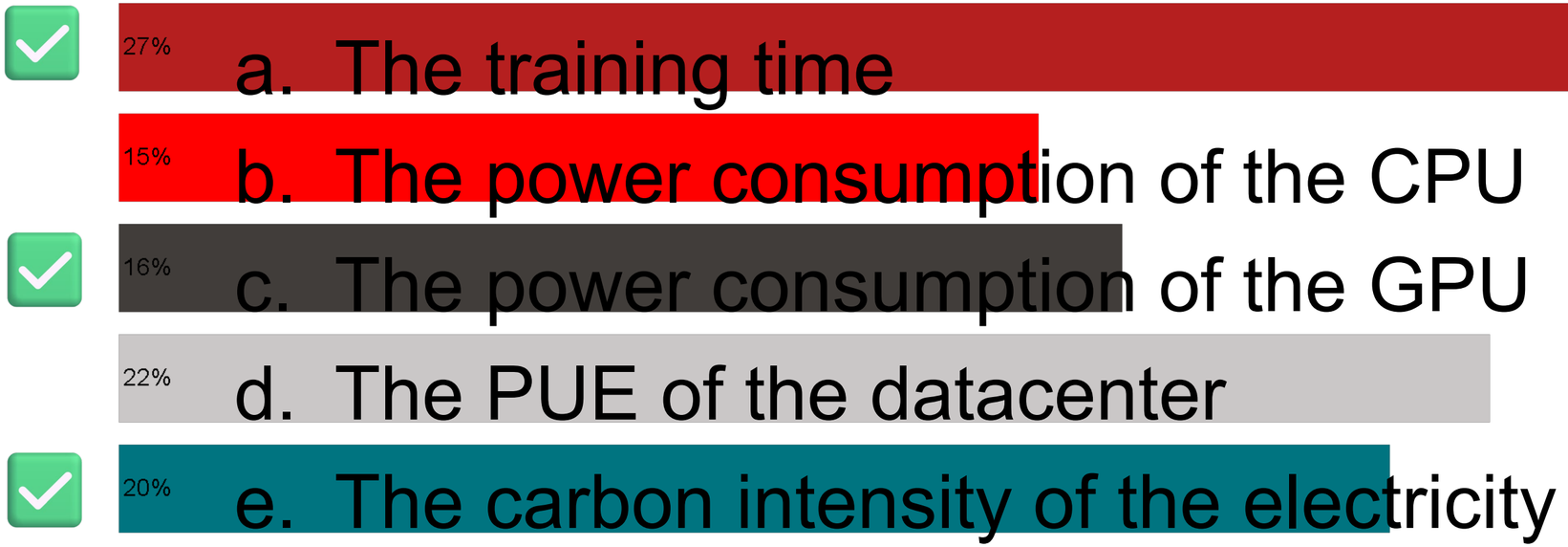
Responsible Software



The footprint of training - 1

URL: ttpoll.eu
Session ID: cs290

What are the 3 most important elements in the carbon footprint of ML training?
Rank them **by decreasing impact** (i.e. most impactful first) :



The power consumed by CPUs is usually negligible compared to GPUs

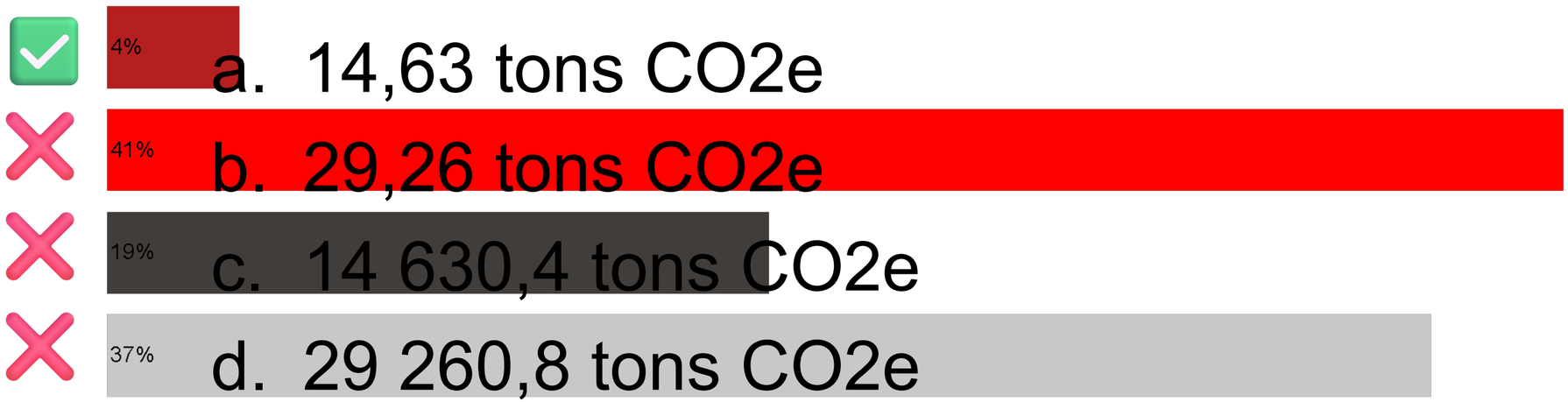
The multiplying factor from carbon intensity is usually higher than that of the PUE

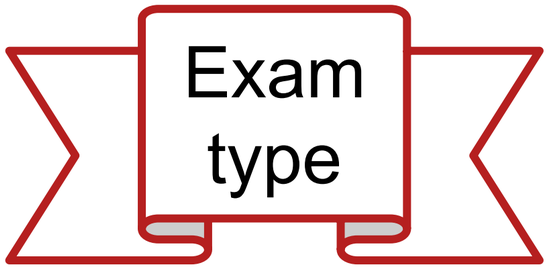
The footprint of training - 2

URL: ttpoll.eu
Session ID: cs290

Let's consider the training of the model SupChat-7B. The computing node has 2 GPUs of the model Nvidia A100 80GB, which consume 400W each. Our datacenter, which has a PUE of 1.2, is located in Germany (carbon intensity: 381g CO2e / kWh).
The training time is 80 000 hours of total GPU computation time.
What is the carbon footprint for the training of SUPMOD-7B?

Exam type but with calculations that can be done by hand (i.e. simpler than here)





Absolute time vs. GPU time

To compute energy consumption (kWh) we need:

- Instantaneous power consumption (kW)
- Time (h) ⚠️ **there are two ways to report time:**
  - Absolute time = absolute duration, where N GPUs are run in parallel
  - GPU time = total GPU use time, as if only 1 GPU were used (i.e. the number of GPUs is already factored in)

$$time_{GPU} = time_{abs} \times n_{GPUs}$$

👉Depending on how the time is reported you will use one of these formulas:

$$Enery = time_{GPU} \times power_{1GPU}$$

or

$$Enery = time_{abs} \times n_{GPUs} \times power_{1GPU}$$

The footprint of training - 2

1. The computation time is provided as “GPU time” i.e. the number of GPUs is already factored in.
2. The power consumption is given in W we need to convert to kW to be able to use the carbon intensity (g/kWh)
3. Finally we need to convert from grams to tons

$$Footprint_{training} = time_{GPU} \times power_{1GPU} \times PUE \times CarbonIntensity$$

$$Footprint_{training} = 80\,000 \times 0{,}400 \times 1{,}2 \times 381 \times {}^{1}/_{1\,000\,000}$$

$$Footprint_{training} = 14{,}63\ tons\ CO_2e$$

The footprint of inference - 1

URL: ttpoll.eu
Session ID: cs290

What are the 3 most important elements in the carbon footprint of ML inference?
Rank them **by decreasing impact** (i.e. most impactful first) :

- ✅ 33% a. The number of user queries
- ✅ 27% b. The electricity consumed per query
- 15% c. The PUE of the datacenter
- ✅ 24% d. The carbon intensity of the electricity

The footprint of inference - 2

The model SupChat-7B is now deployed in production. It is hosted on the same computing node with 2 GPUs of the model Nvidia A100 80GB, which consume 400W each. Our datacenter, which has a PUE of 1.2, is located in Germany (carbon intensity: 381g CO2e / kWh). Our model is able to serve 120 tokens per second ~~of computation time~~. It has an average of 2000 users daily and generates an average of 5000 tokens per user per day.

What is the carbon footprint of 1 day of inference?

Confusing: speed is given in absolute time!

1. What is the total GPU computation time used over 1 day (in h)?
2. What is the power consumed by the model for inference (in W)?
3. What is the total electricity consumed over 1 day (in kWh)?
4. What is the carbon footprint over 1 day (in kg CO2e)?

The footprint of inference - 2

1. The **GPU computation time** used over 1 day can be obtained from the **speed** of the model (given in absolute time), the **total number of tokens served per day** and the **number of GPUs**
   + you need to convert from seconds to hours

$$time_{GPU} = \frac{nbusers_{perday} \times nbtokens_{peruser_{perday}}}{modelspeed} \times n_{GPUs} \times \frac{1}{3600}$$

$$time_{GPU} = \frac{2000 \times 5000}{120} \times 2 \times \frac{1}{3600}$$

$$time_{GPU} = 46{,}30\ h$$

Sanity check: with 2 GPUs in parallel, 46,30 h of GPU time is taking 23,15h of absolute time i.e. a bit less than 1 day.

The footprint of inference - 2

2. Since we are working with GPU time, we need to use the instantaneous power consumption of 1 GPU (in kW):

$$Power_{inference} = 0{,}400\ kW$$

3. To get the electricity consumed we multiply GPU time with the instantaneous power consumption, we multiply by the PUE to account for the overhead electricity consumed by cooling:

$$Electricity_{inference} = time_{GPU} \times Power_{Inference} \times PUE$$

$$Electricity_{inference} = 46{,}30 \times 0{,}400 \times 1{,}2$$

$$Electricity_{inference} = 22{,}22\ kWh$$

The footprint of inference - 2

4. To get the carbon footprint in kg CO2e we multiply the electricity consumed by the carbon intensity, then we scale to kg (i.e. divide by 1000)

$$Footprint_{inference} = Electricity_{inference} \times CarbonIntensity$$

$$Footprint_{inference} = \frac{22{,}22 \times 381}{1000}$$

$$Footprint_{inference} = 8{,}47\ kg\ CO_2e$$

Total carbon footprint

URL: ttpoll.eu
Session ID: cs290

We have obtained the carbon footprint of SupChat-7B at training and at inference time. What is its total carbon footprint?

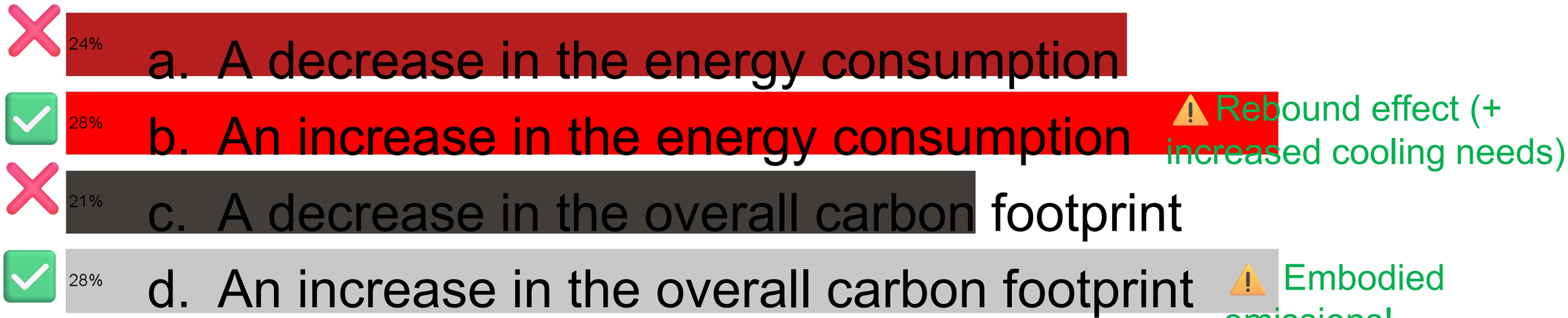
- ❌ 4% a. Training
- ❌ 0% b. Inference
- ❌ 0% c. Training x Inference
- ❌ 0% d. Inference – Training
- ✅ 89% e. Training + Inference
- ✅ 7% f. Other

We also need to add the footprint associated with embodied emissions (i.e. hardware manufacturing mainly)

Hardware renewal

URL: ttpoll.eu
Session ID: cs290

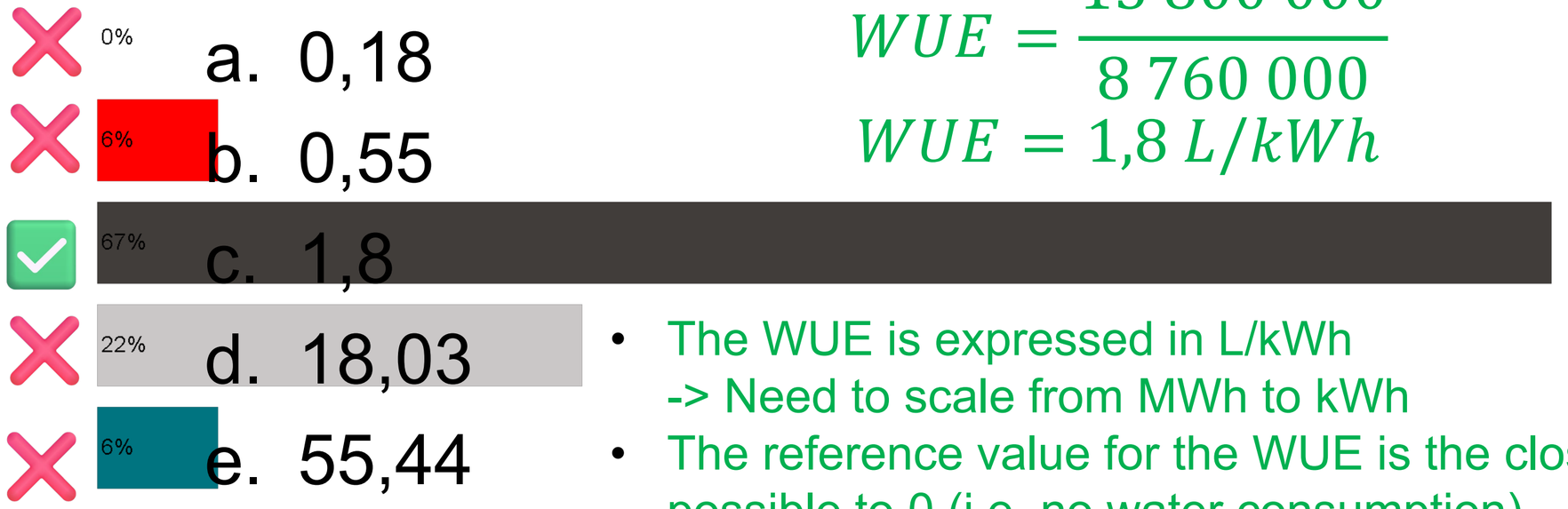
We want to optimize the energy consumption of SupChat-7B at inference time. We decide to upgrade our hardware platform and to replace our A100 GPUS with H100 GPUs. The H100 are 4 times more performant than the A100 in terms of computation speed. Their power consumption is 700W at maximum use.
What effect(s) are we likely to observe (select all that apply)?



Water Usage Effectiveness

URL: ttpoll.eu
Session ID: cs290

The datacenter hosting SupChat-7B consumes an average of 1 MW.
This means annually a total of 8 760 MWh of electricity.
It consumes approximately 15.8 million liters of water each year.
What is the WUE of the datacenter (onsite only)?

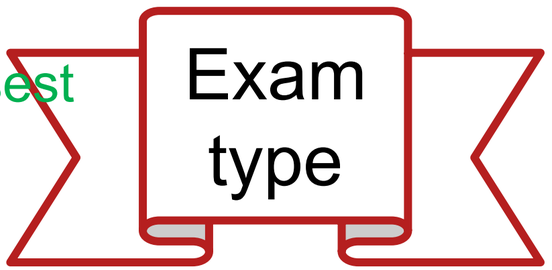


$$WUE = \frac{15\,800\,000}{8\,760\,000}$$

$$WUE = 1{,}8\ L/kWh$$

Exam type but with calculations that can be done by hand (i.e. simpler than here)

- The WUE is expressed in L/kWh
  -> Need to scale from MWh to kWh
- The reference value for the WUE is the closest possible to 0 (i.e. no water consumption)



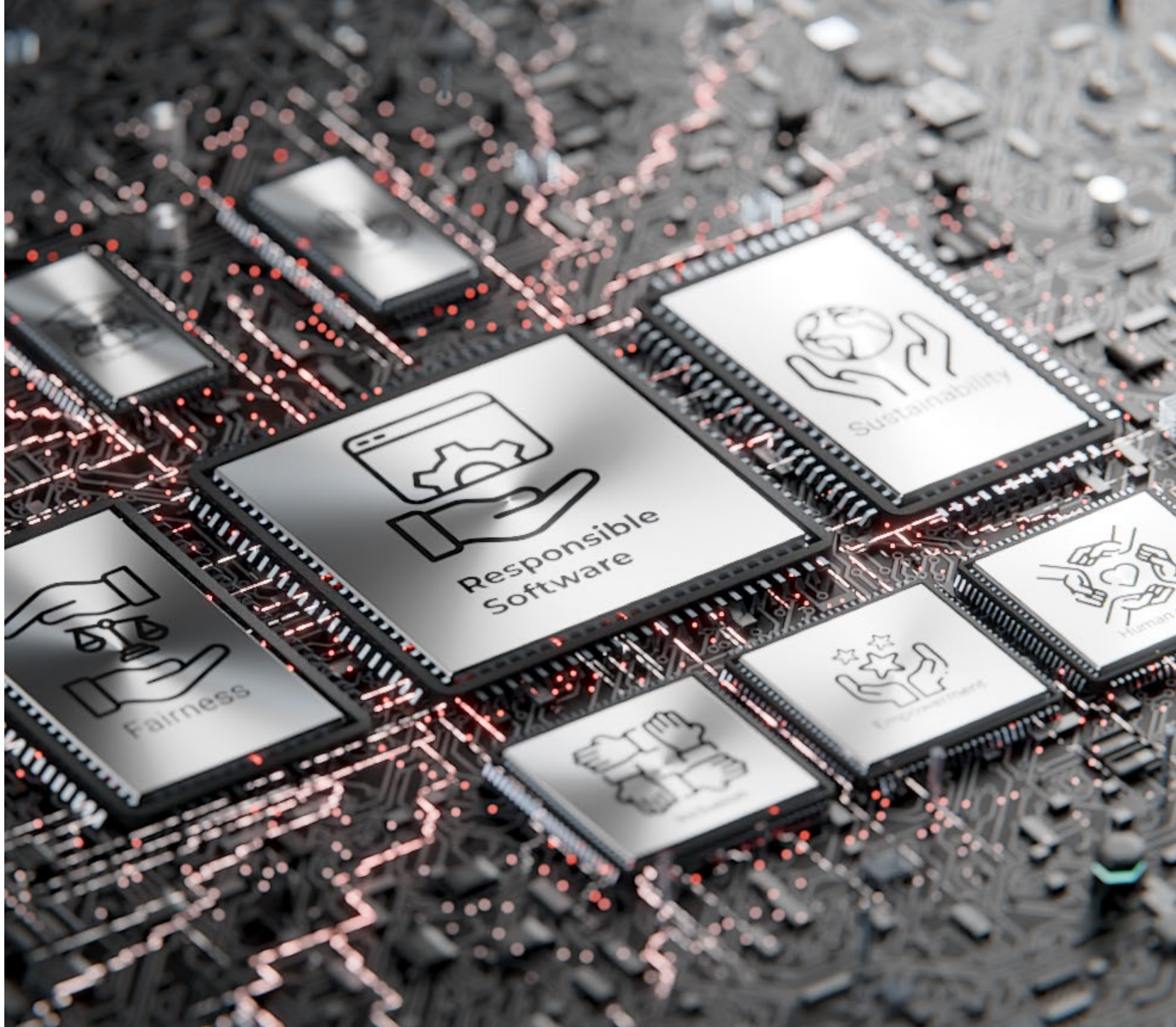
https://dgtlinfra.com/data-center-water-usage/

Empowerment 1 Review & Case studies

24 nov.

Cécile Hardebolle

Responsible Software



Meditation app

URL: ttpoll.eu
Session ID: cs290

ZenPath is an app dedicated to mental well-being that offers guided meditation sessions online. To reduce user dropout, they decide to display a popup after a user skips two sessions where the “Resume Today!” button is preselected.
What type of nudging technique is most likely used here?

❌ 42% a. Opt-in

/!\ Opt-in = button NOT preselected

❌ 5% b. Social proof

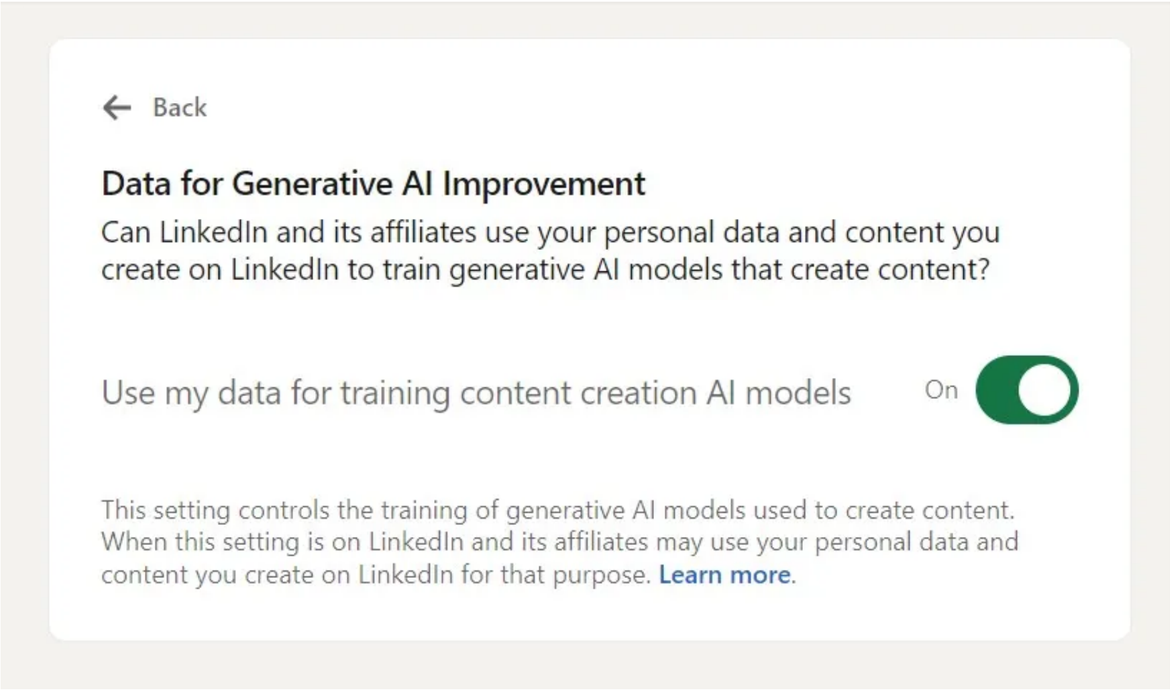
❌ 0% c. Scarcity

✅ 53% d. Default

Exam type

Use of data

URL: ttpoll.eu
Session ID: cs290



This is one of the settings on LinkedIn in the USA, set to its default value.
What is the most likely outcome?

❌ 0% a. Most users will turn the setting off

❌ 0% b. Most users will turn the setting on

✅ 100% c. Most users will let the setting as is

❌ 0% d. Most users will change the setting

Navigation app

URL: ttpoll.eu
Session ID: cs290

In an effort towards more sustainability, the itinerary search in Noodle Maps now returns 2 itinerary options in the following order:

1) the most fuel-efficient but longest itinerary
2) the shortest but least fuel-efficient itinerary

What are the characteristics of this nudge? (select all that apply)

- ✓ 21% a. Takes advantage of System 1
- 9% b. Takes advantage of System 2

Does not really push users to reflect, but relies on the effect of order

- ✓ 34% c. Transparent to the user
- 2% d. Covert

Depends on implementation, but can be said to be visible to the users

- ✓ 32% e. Ethically fine
- 2% f. Ethically problematic

3 criteria: autonomy, transparency, welfare - this example can be thought to be fine, some criticisms relate to interfering with autonomy + benefit to community vs. individual user

Deceptive patterns vs nudges

URL: ttpoll.eu
Session ID: cs290

Which of the following are characteristics **shared** by nudges and **deceptive patterns**? (select all that apply)

- ✓ 22% a. They modify the choice architecture
- ✓ 16% b. They make users do things they didn't originally mean to
- ✓ 26% c. They take advantage of how humans make decisions
- ✓ 25% d. They intentionally bias user behavior
- 0% e. They restrict choices
- 0% f. They benefit users
- 10% g. They benefit another party
- 1% h. They make users lose track of time

Shared characteristics (item b can be discussed…)

Characteristics of either nudges or deceptive patterns. [Items g and f can lead to confusion and should be reframed]

E-commerce platform

URL: ttpoll.eu
Session ID: cs290

The e-commerce platform Shine would like to implement new features to improve the experience of its various categories of users. Here is the list of envisaged features.
Which of them best matches the definition of a deceptive pattern?

❌ 17% a. Personalize style recommendations based on past browsing

❌ 8% b. Display user-provided past purchase data to recommend sizes

✅ 75% c. Register users to a ShineClub membership trial on checkout

❌ 0% d. Provide downloadable QR codes for the free return of items

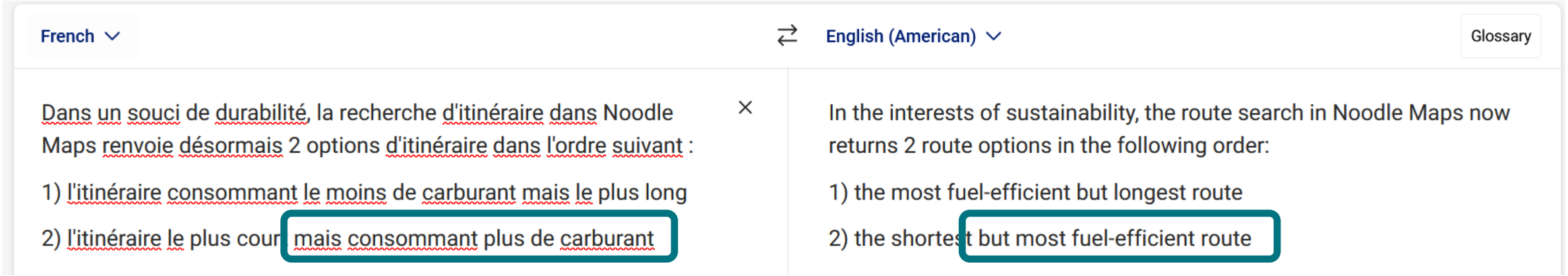
Exam type

https://www.theguardian.com/business/2025/jun/05/shein-fashion-consumer-watchdog-complaint-eu-alleged-dark-practices

Translation

URL: ttpoll.eu
Session ID: cs290

Consider the following translation. What is the issue here?



❌ 0% a. Parity error

❌ 56% b. Factuality error

❌ 0% c. Measurement error

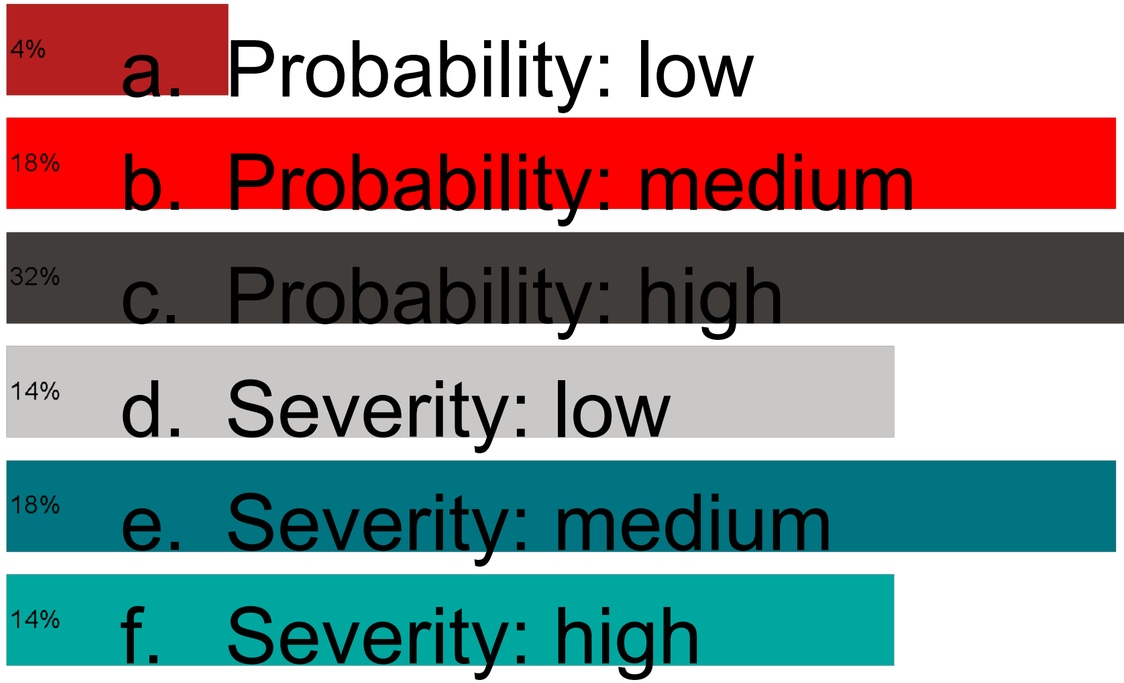
✅ 44% d. Faithfulness error

The response is erroneous compared to the input (prompt).
(Here since “Noodle Maps” does not exist, it cannot really be argued that the error relates to a known fact i.e. it is not a Factuality Error)

Evaluating the level of risk - 1

URL: ttpoll.eu
Session ID: cs290

Consider the following Privacy risk: **“Tracks personal app usage”**
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)



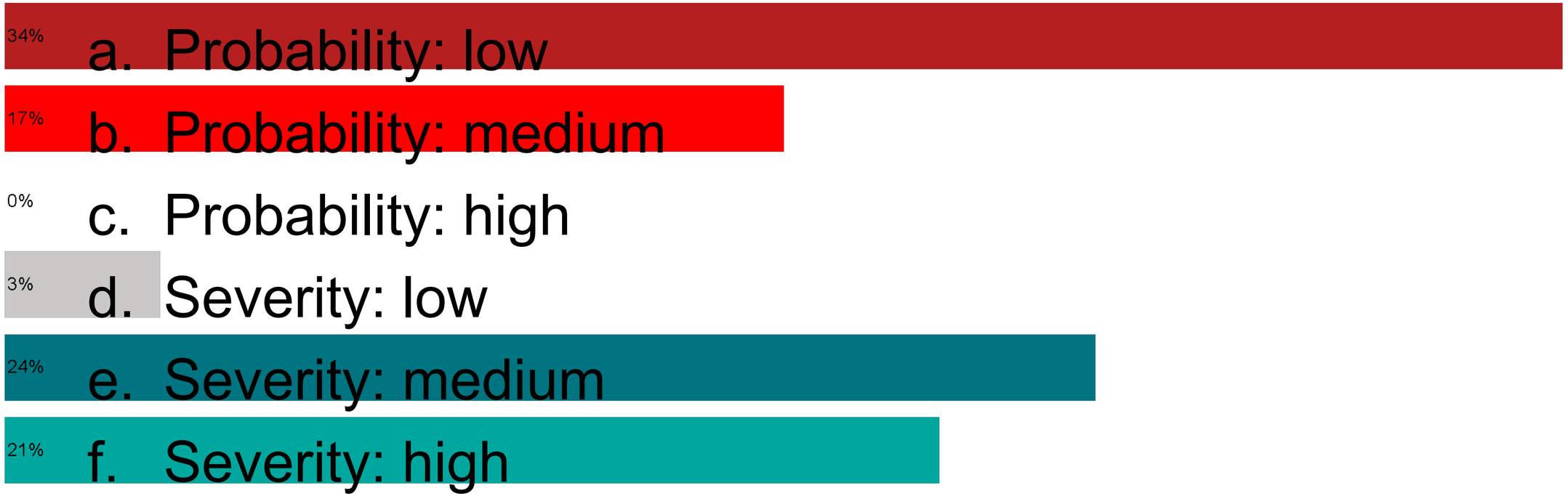
Qualitative evaluation: you need to provide a **justification** to support your evaluation of the probability/severity (including hypotheses you make on how the app is implemented), such as:

- Probability High: the app relies on tracking, so it necessarily is going to happen
- Severity High: tracking means collecting behavioral data over time, which can be considered sensitive (may disclose personal info)

Evaluating the level of risk - 2

URL: ttpoll.eu
Session ID: cs290

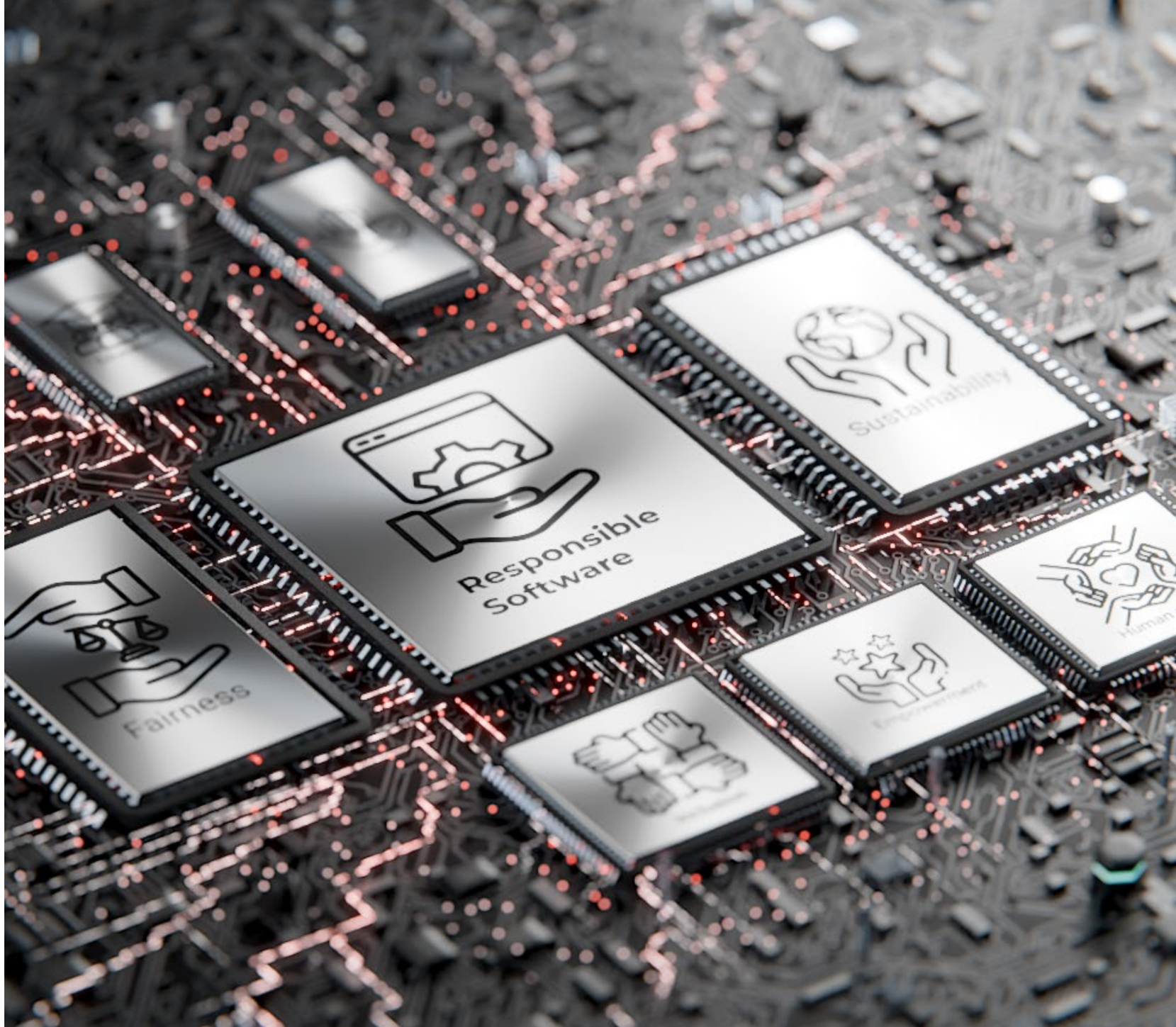
Consider the following Welfare risk: **“Excessive reminders could lead to stress or anxiety”**. How would you evaluate the level of this risk in terms of probability and severity of impacts?
(select 2 options: 1 for probability, 1 for severity)



Empowerment 2 Review & Case studies
8 dec.

Cécile Hardebolle

Responsible Software



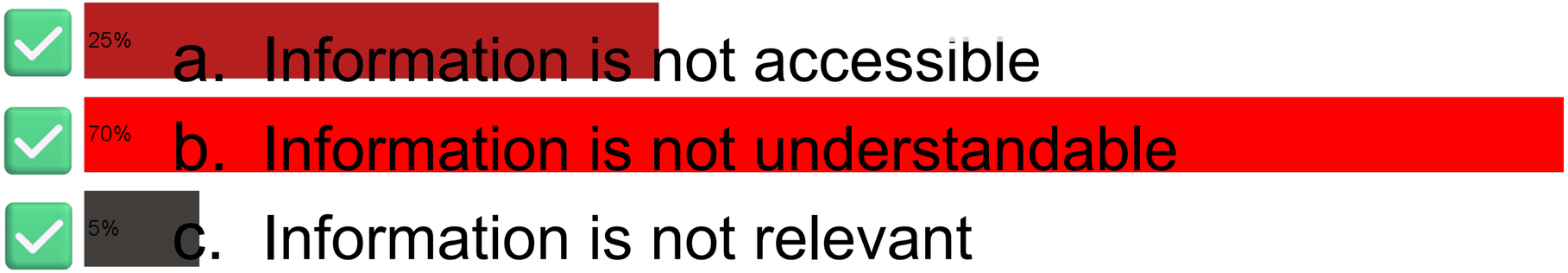
Privacy policies

URL: ttpoll.eu
Session ID: cs290

Several studies have shown that the privacy policies of many online platforms and websites are extremely long (several thousand of words, taking in the 20 minutes to read on average), use legalistic terminology and are hard to navigate.

This can be said to be a transparency issue because (select all that apply):

All of these can be argued:
- Hard to navigate = accessibility issue
- Legalistic vocab = understandability issue
- Extremely long = relevance issue



(Sherman, 2024; Litman-Navarro, 2019)

Beer brewing dataset - 1

URL: ttpoll.eu
Session ID: cs290

One of the results of your Bachelor thesis is a very cool dataset which contains tasting profiles and consumer reviews for 3197 unique beers from 934 different breweries. This dataset can be used to train machine learning models for sentiment analysis and classification tasks.
You want to make the dataset public.
For ensuring transparency you should also publish with it:
(select all that apply):

All of these (composition of the data is probably the least important because it can be obtained from the data)

- ✅ 25% a. Composition of the data, including demographics
- ✅ 29% b. Description of the collection process
- ✅ 27% c. Description of the pre-processing performed
- ✅ 19% d. Description of the purposes and intended use

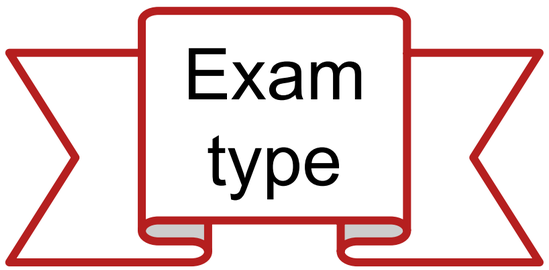
Beer brewing dataset - 2

URL: ttpoll.eu
Session ID: cs290

One of the results of your Bachelor thesis is a very cool dataset which contains tasting profiles and consumer reviews for 3197 unique beers from 934 different breweries. This dataset can be used to train machine learning models for sentiment analysis and classification tasks.
You have created a datasheet for your dataset.

Which of the FAIR principles do you follow by providing a datasheet?

- ❌ 8% a. Findable
- ❌ 0% b. Accessible
- ❌ 8% c. Interoperable
- ✅ 85% d. Reusable



Linear Regression Model

You have found on HuggingFace an open-source Linear Regression model that predicts the price of a house based on a range of features like lot area, construction year, number of rooms, etc. For recall, a linear regression model has the following mathematical form, where *y'* is the predicted price, x_i are the features and *b* and w_i are the final parameters of the model:

$$y' = b + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + \cdots$$

Let's imagine that you want to modify this model. How could you do it?

- ❌ 71% a. Modify the values of the parameters
- ❌ 0% b. Use a post-hoc interpretability method
- ✅ 29% c. Retrain the model with a new dataset
- ❌ 0% d. It is not possible to modify the model

- **a&c: SEE NEXT SLIDE for explanation**
- b: post-hoc interpretability methods have nothing to do with model modification (they only help interpret how the model works)
- d: the text says "open-source", so it is possible to modify the model

Note on modifying ML models

- The final parameters of a ML model represent the patterns in the data as “detected” (“learned”) by the learning algorithm
- Technically speaking, it is *possible* to modify/edit these parameters manually, however it is generally NEVER done because:
  - It is generally impossible to do it without “breaking” the model
  - Then the model does not reflect anymore the patterns learned from the data

👉 If you need to modify a ML model, you will generally retrain it with new data, modify the training procedure, etc. so that it “learns” a different pattern

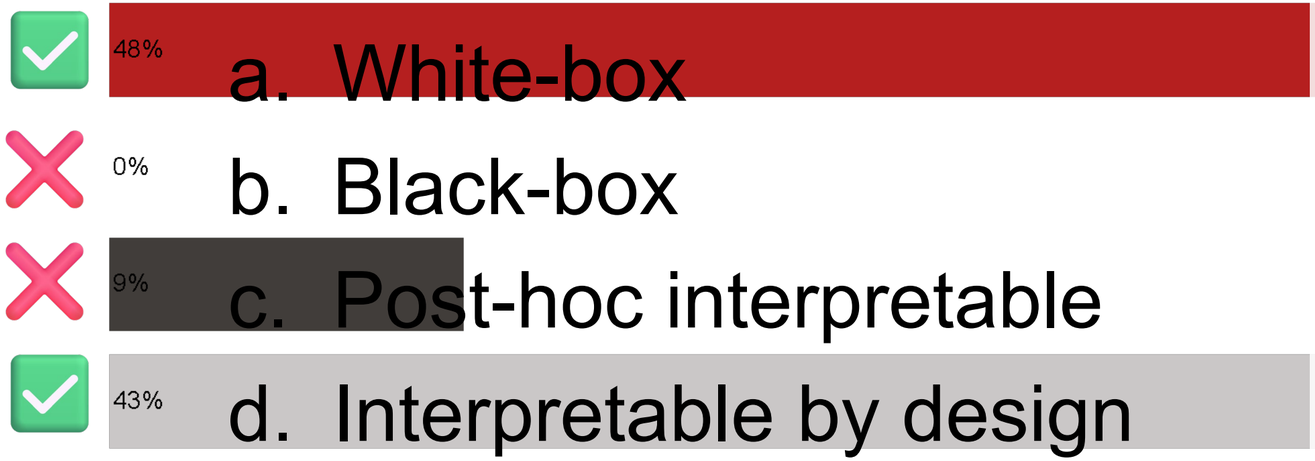
⚠️ In Fairness 1, the university admission software is **NOT** a ML model, it is a “classic” algorithm designed by hand (not by learning from data), which is why we CAN modify the “parameters” [the goal was to show you that unfairness is not specific to ML, it happens with classic algorithms too]

Logistic Regression Model

URL: ttpoll.eu
Session ID: cs290

In the Fairness 2 notebook you have created a Logistic Regression model on the ProPublica dataset to try to reproduce how the COMPAS software predicts the risk of recidivism.

The Logistic Regression model you have created can be said to be (select all that apply):



"White-box" models are interpretable by design (i.e. the two terms are synonyms)

Note: a post-hoc interpretability method CAN be used on a white-box model, however this is generally not done because it is unnecessary in most cases (since white-box models are interpretable by definition) and it does not bring any advantage (since trust in post-hoc methods is generally lower because they are external to the model)

COMPAS

URL: ttpoll.eu
Session ID: cs290

To have transparency on the ML model behind the COMPAS software would mean to have access to:

46% a. The design documentation

0% b. The user documentation

23% c. The training code

15% d. The training dataset

0% e. A post-hoc interpretability method

✅ 15% f. It depends

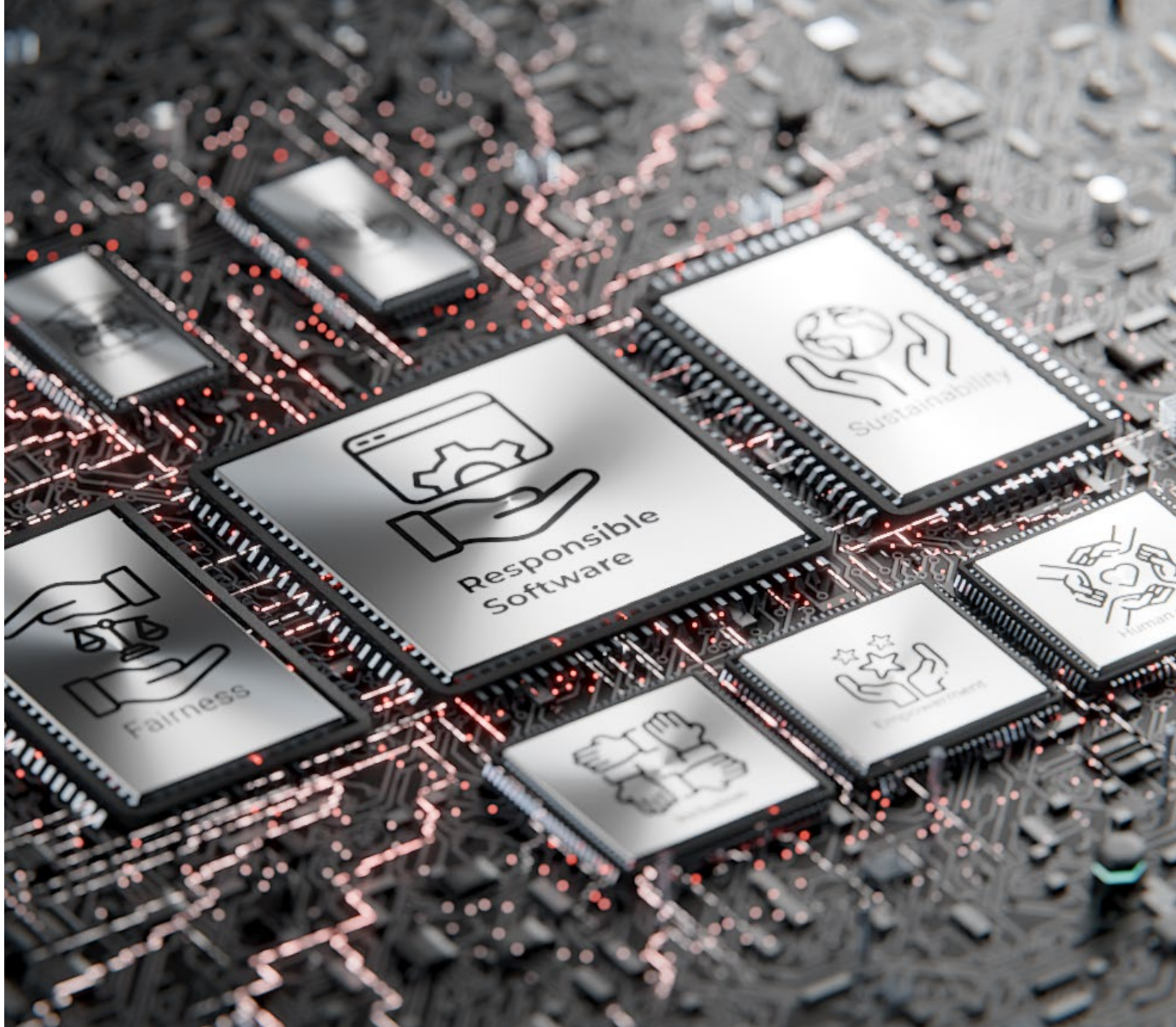
It depends on the stakeholder considered
Transparency = “the degree to which stakeholders can answer their questions by using the information they obtain about a software system during its life cycle”

-> All these options could be used potentially!

Conclusion
Case studies
+ Q&A
15 dec.

Cécile Hardebolle

Responsible Software



Review questions “Whole Course”

Ethical sensitivity

New

URL: ttpoll.eu
Session ID: cs290

What is ethical sensitivity?

0% a. The ability to predict all technical outcomes before deployment

✅ % b. The capability to identify the impact of a situation on others

0% c. The ability to act to benefit others even at your own expense

18% d. The capacity to account for all ethical values simultaneously

Chemical discovery

URL: ttpoll.eu
Session ID: cs290

A software company has developed a Machine Learning model that is able to discover new chemical compounds for medicine development. They identify that the model can also discover new chemical weapons.

What type issue is this?

0% a. A technical issue

22% b. An ethical issue

✅ 78% c. An ethical dilemma

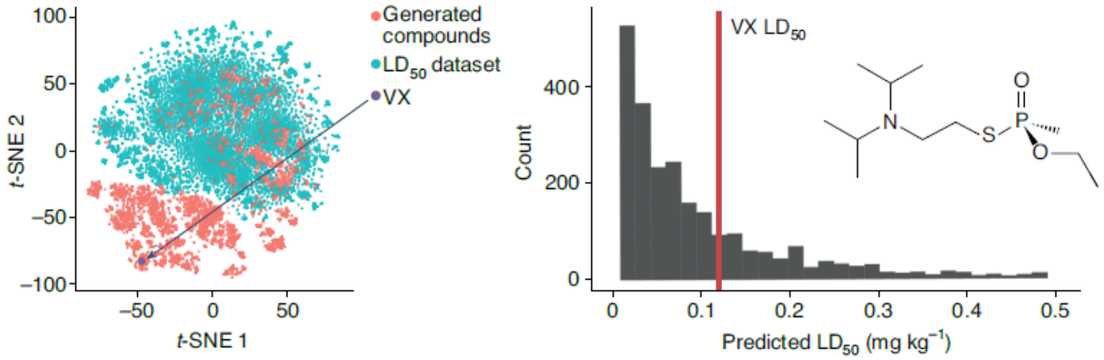


Fig. 1 | A *t*-SNE plot visualization of the LD_{50} dataset and top 2,000 MegaSyn AI-generated and predicted toxic molecules illustrating VX. Many of the molecules generated are predicted to be more toxic in vivo in the animal model than VX (histogram at right shows cut-off for VX LD_{50}). The 2D chemical structure of VX is shown on the right.

(Urbina et al., 2022)

Vulnerabilities

New

URL: ttpoll.eu
Session ID: cs290

A software engineer decides to postpone the launch of a new feature due to the late discovery of a security vulnerability and justifies: “The new feature would bring us some short-term benefits but would have serious negative consequences for all of our customers, our aim must be the greatest good for the greatest number.”
Which ethical theory does this engineer follow?

Exam type

Food delivery

New

URL: ttpoll.eu
Session ID: cs290

An online food delivery app experiences a data breach where customer payment details are stolen.
What type of risks are represented in this situation?

0% a. Safety risks from misdiagnosed food allergies

4% b. Safety risks from incorrect delivery scheduling

13% c. Sociotechnical risks in app-driver communication

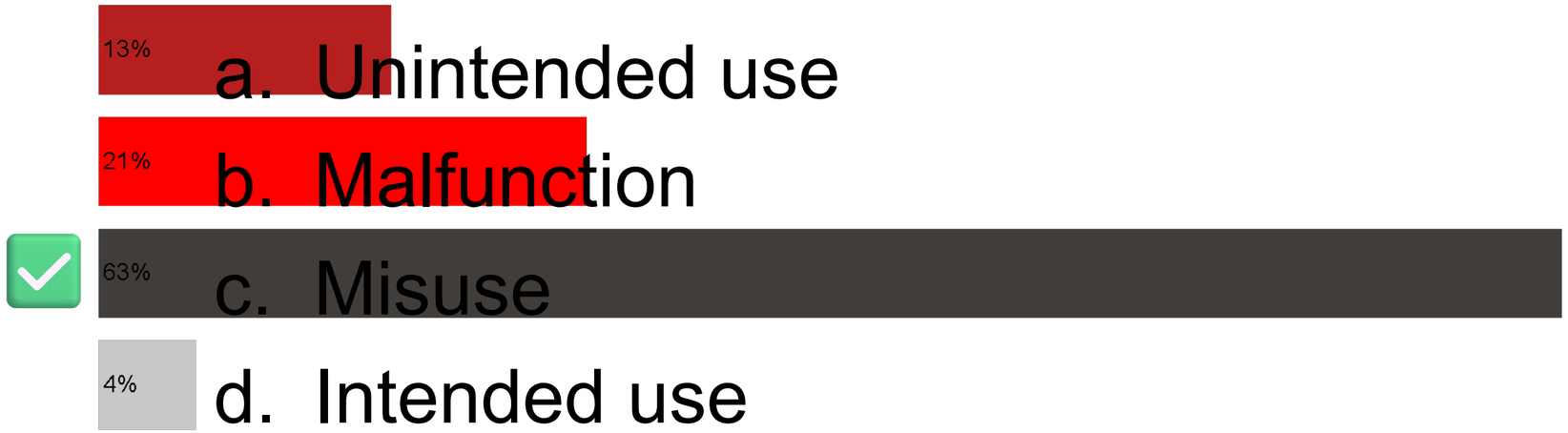
✅ 83% d. Security risks from unauthorized system access

Hospital

New

URL: ttpoll.eu
Session ID: cs290

Patient records in a hospital have been encrypted by cybercriminals who demand payment to restore access, causing emergency services to halt and delay critical care for patients.
Which harm scenario does this represent?



Exam type

Fissures in concrete

URL: ttpoll.eu
Session ID: cs290

The company SuperCrack has developed a model to detect fissures in concrete walls before they become visible to the naked eye. A positive result means presence of fissure.
Which of the statements below is correct?

87% a. TN = actual absence of fissure, correct prediction

4% b. TN = actual presence of fissure, incorrect prediction

0% c. TP = actual presence of fissure, incorrect prediction

9% d. TP = actual absence of fissure, correct prediction

Exam type

Contagious disease

New

URL: ttpoll.eu
Session ID: cs290

A rapid test for a contagious disease (infected = positive result) shows a high number of false negatives.
What are the consequences of false negatives in terms of safety?

0% a. Healthy people continue their daily activities as normal.

4% b. Healthy people receive unnecessary quarantine.

0% c. Infected individuals receive the appropriate medication.

✅ 96% d. Infected individuals spread the disease unknowingly.

Exam type

Political campaign

New

URL: ttpoll.eu
Session ID: cs290

A whistleblower releases authentic internal documents from the campaign of a political party with the goal of damaging the party's public image for the upcoming election.
What type of information is this?

25% a. Misinformation

8% b. Disinformation

✅ 67% c. Malinformation

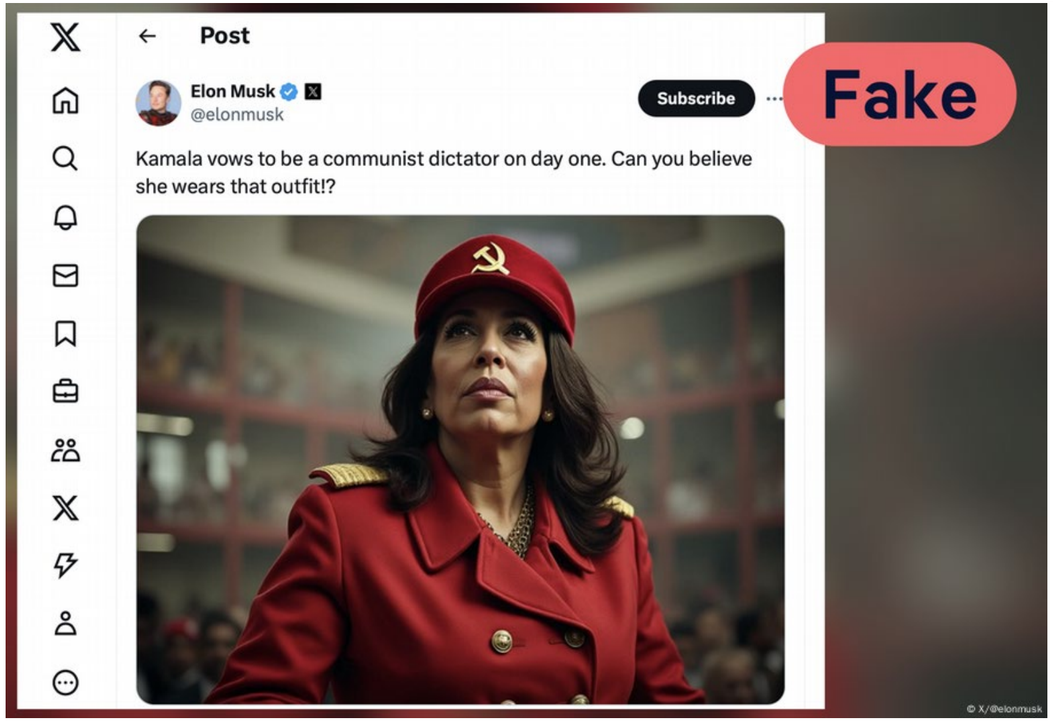
0% d. Fake news

Exam type

Posts on Twitter

URL: ttpoll.eu
Session ID: cs290

One dis-/mis-information post by Elon Musk appears in your Twitter timeline.
Why would you be more likely to believe it than other posts?



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

0% a. System 2

30% b. Illusory truth

✅ 65% c. Source cues

5% d. Prebunking

Exam type

Loans

New

URL: ttpoll.eu
Session ID: cs290

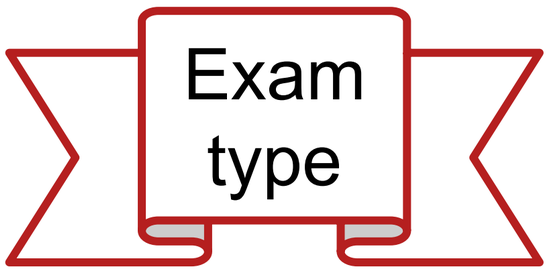
A ML model for loan approval consistently denies loans to applicants from rural neighborhoods. The model has been trained on data from the bank covering all the loan decisions taken in the last 5 years for all the neighborhoods served by the bank.
Which type of bias is most likely present in the data from this scenario?

11% a. Sampling bias

26% b. Representation bias

0% c. Measurement bias

✅ 63% d. Preexisting bias



Shoplifting

URL: ttpoll.eu
Session ID: cs290

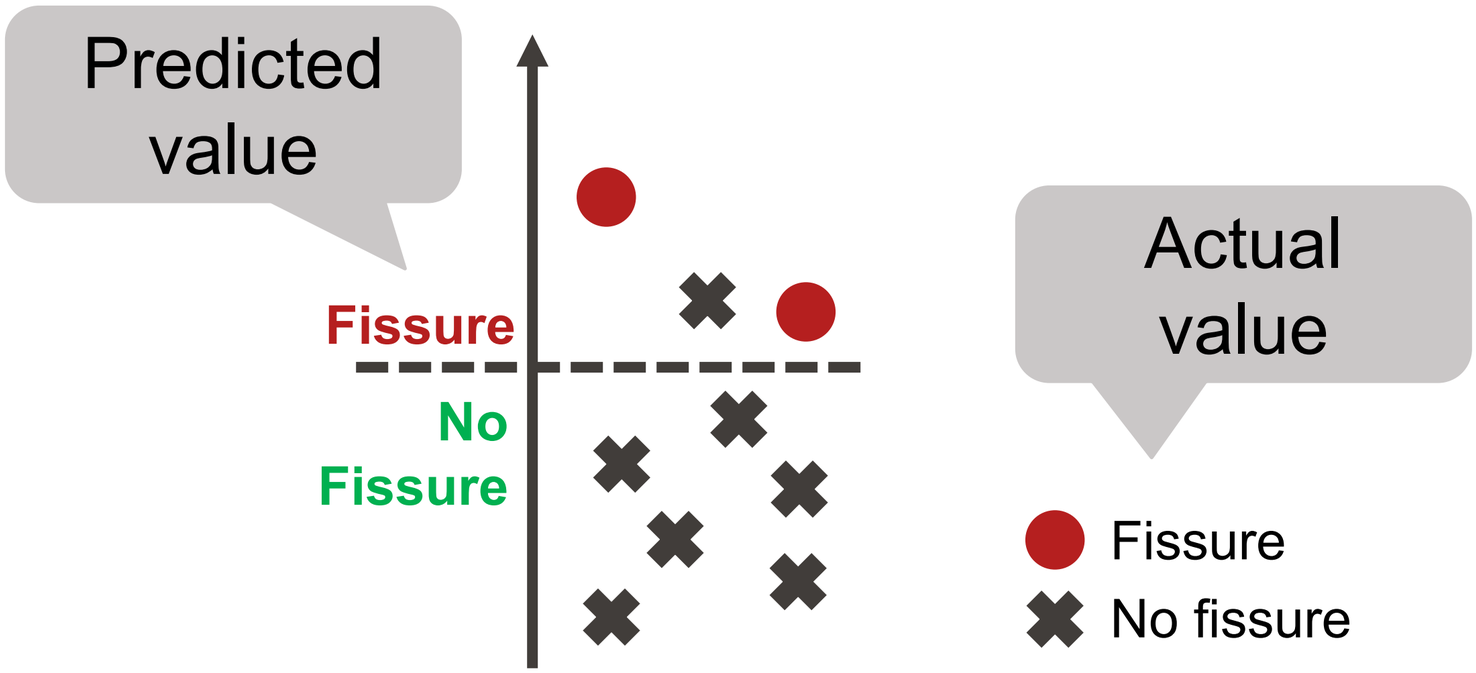
The society RetailProtect develops a ML model to identify instances of shoplifting in retail shops. They evaluate their model on a benchmark in which actors from diverse ethnicities simulate a range of shoplifting actions. They plan to deploy soon in shops.

What type of bias is present in this scenario?

- ✅ 0% a. Evaluation bias
- 0% b. Aggregation bias
- 0% c. Optimization bias
- 0% d. Deployment bias

Fissures in concrete (again)

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:

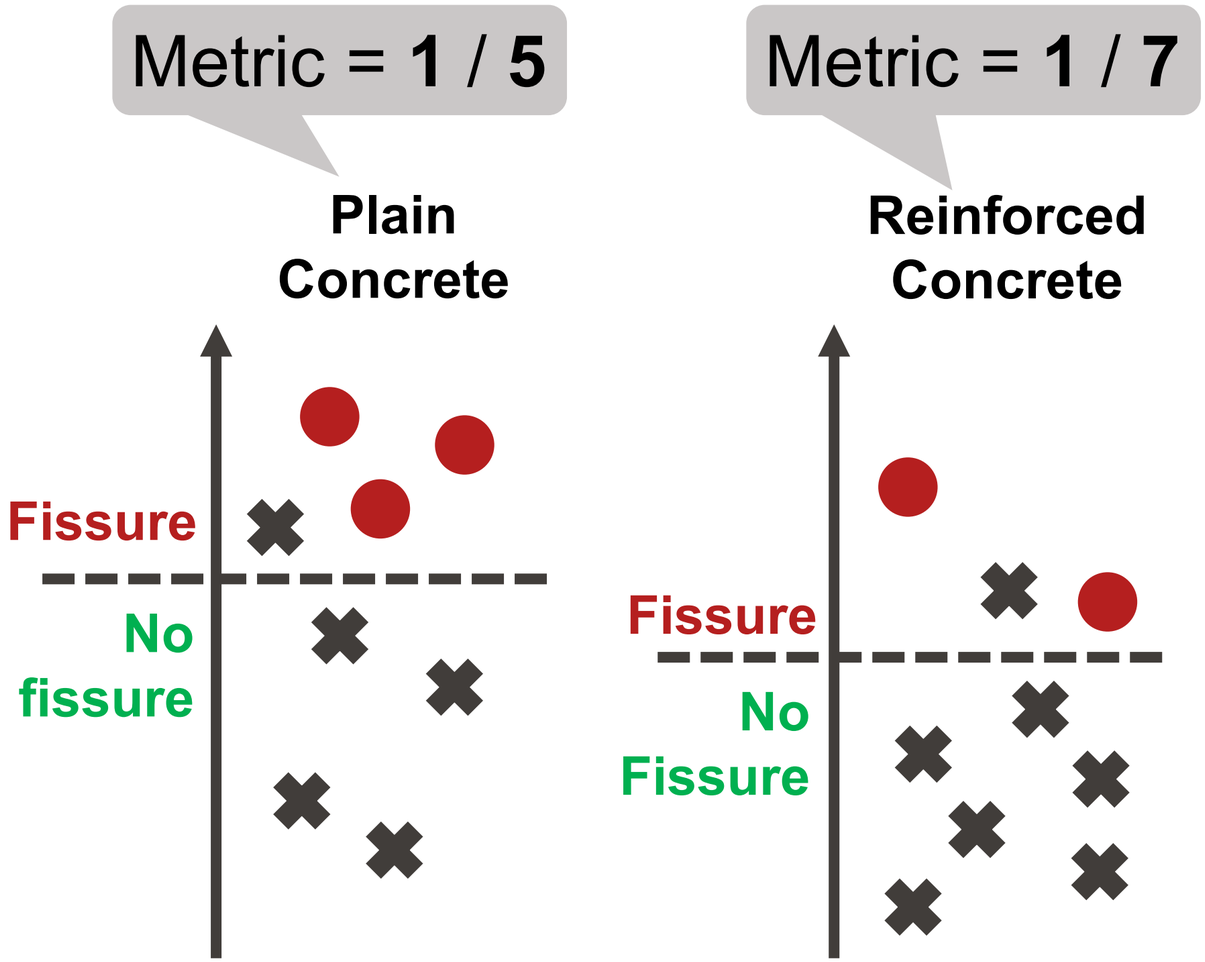


		Predicted: Fissure	Predicted: No Fissure
Actual	Fissure	2	0
Actual	No Fissure	1	6

Fissures in concrete (again) New

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:



Which metric are they using? (select 1 answer)

0% a. Accuracy

0% b. FNR

0% ✅ c. FPR

0% d. Positive prediction rate

University admissions

New

URL: ttpoll.eu
Session ID: cs290

A model trained to help screen applications to university has an accuracy of 97% and the false positive rate (FPR) is 5% for group X and 6% for group Y. However, the Disparate Impact Ratio is 0,613 with group X having a higher admission rate.
What is most likely happening in this situation?

0% a. Differences in the FNR are causing the low DIR

0% b. The DIR indicates a higher error rate for group Y

0% c. The applicants from group X have stronger profiles

✅ 0% d. Group Y has a lower rate of actual positive labels

Seen in the Graded Notebook 2: inconsistency between fairness metrics indicate that the data used as "ground truth" i.e. actual data is biased, groups having dissimilar rates of positive labels (= pre-existing bias)
[This is what is called the "impossibility result"]

Datacenter cooling

URL: ttpoll.eu
Session ID: cs290

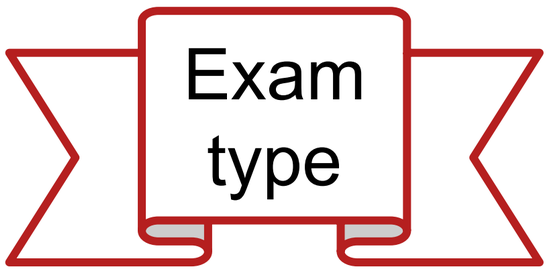
The GreenDC datacenter consumes an average of 1 MW.
This means annually a total of 8 760 MWh of electricity.
50% of this electricity is used to power the IT equipment.
What is the PUE of GreenDC?

0% a. 0.5

0% b. 1

0% c. 1.5

✅ 0% d. 2



Datacenter water

New

URL: ttpoll.eu
Session ID: cs290

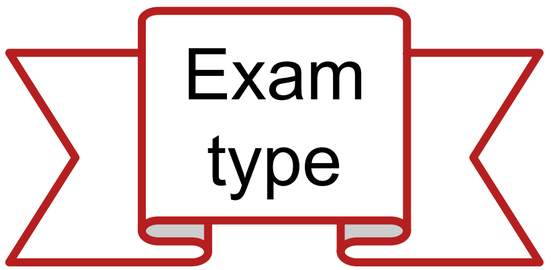
The TitanCore datacenter consumes a total of 24 000 MWh of electricity annually. It consumes approximately 16 million liters of water each year.
What is the WUE of the datacenter (onsite only)?

0% a. 0,000667

0% b. 0,667

✅ 0% c. 1,5

0% d. 1500



LLM training

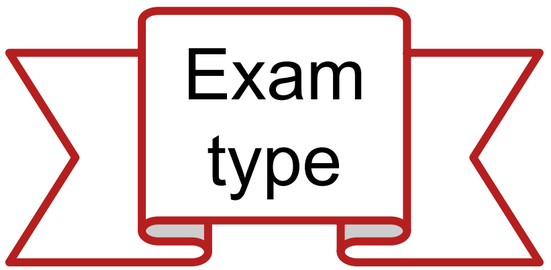
New

URL: ttpoll.eu
Session ID: cs290

The training of the LLM “BreezeTalk” took 3 months using 100% of the resources available on a 10-server cluster.
Each server has an embodied footprint of 1200 kg CO2e and a 3-year lifespan.
What share of embodied footprint should be allocated to BreezeTalk (training only), in kg CO2e?

a. 100
✅ b. 1000
c. 3000
d. 12000

M_s = M_h x time_share x resource_share
M_s = (10 x 1200) x (3/12 / 3) x (100 / 100)
M_s = 12000 x 1/12

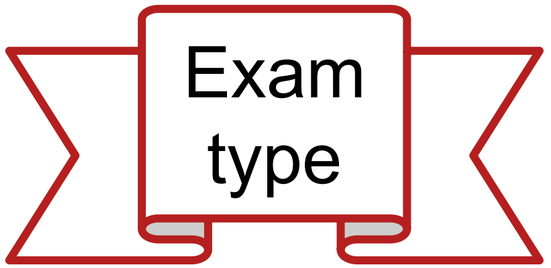


Meditation app

URL: ttpoll.eu
Session ID: cs290

ZenPath is an app dedicated to mental well-being that offers guided meditation sessions online. To reduce user dropout, they decide to display a popup after a user skips two sessions where the “Resume Today!” button is preselected.
What type of nudging technique is most likely used here?

0% a. Opt-in
0% b. Social proof
0% c. Scarcity
✅ d. Default



E-commerce platform

URL: ttpoll.eu
Session ID: cs290

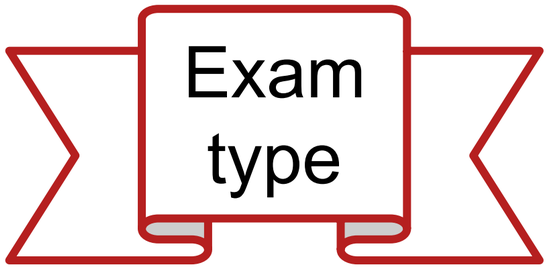
The e-commerce platform Shine would like to implement new features to improve the experience of its various categories of users. Here is the list of envisaged features.
Which of them best matches the definition of a deceptive pattern?

0% a. Personalize style recommendations based on past browsing

0% b. Display user-provided past purchase data to recommend sizes

✅ c. Register users to a ShineClub membership trial on checkout

0% d. Provide downloadable QR codes for the free return of items



Beer brewing dataset

URL: ttpoll.eu
Session ID: cs290

One of the results of your Bachelor thesis is a very cool dataset which contains tasting profiles and consumer reviews for 3197 unique beers from 934 different breweries. This dataset can be used to train machine learning models for sentiment analysis and classification tasks.
You have created a datasheet for your dataset.

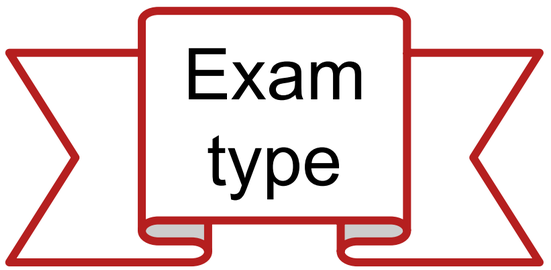
Which of the FAIR principles do you follow by providing a datasheet?

0% a. Findable

0% b. Accessible

0% c. Interoperable

✅ 0% d. Reusable

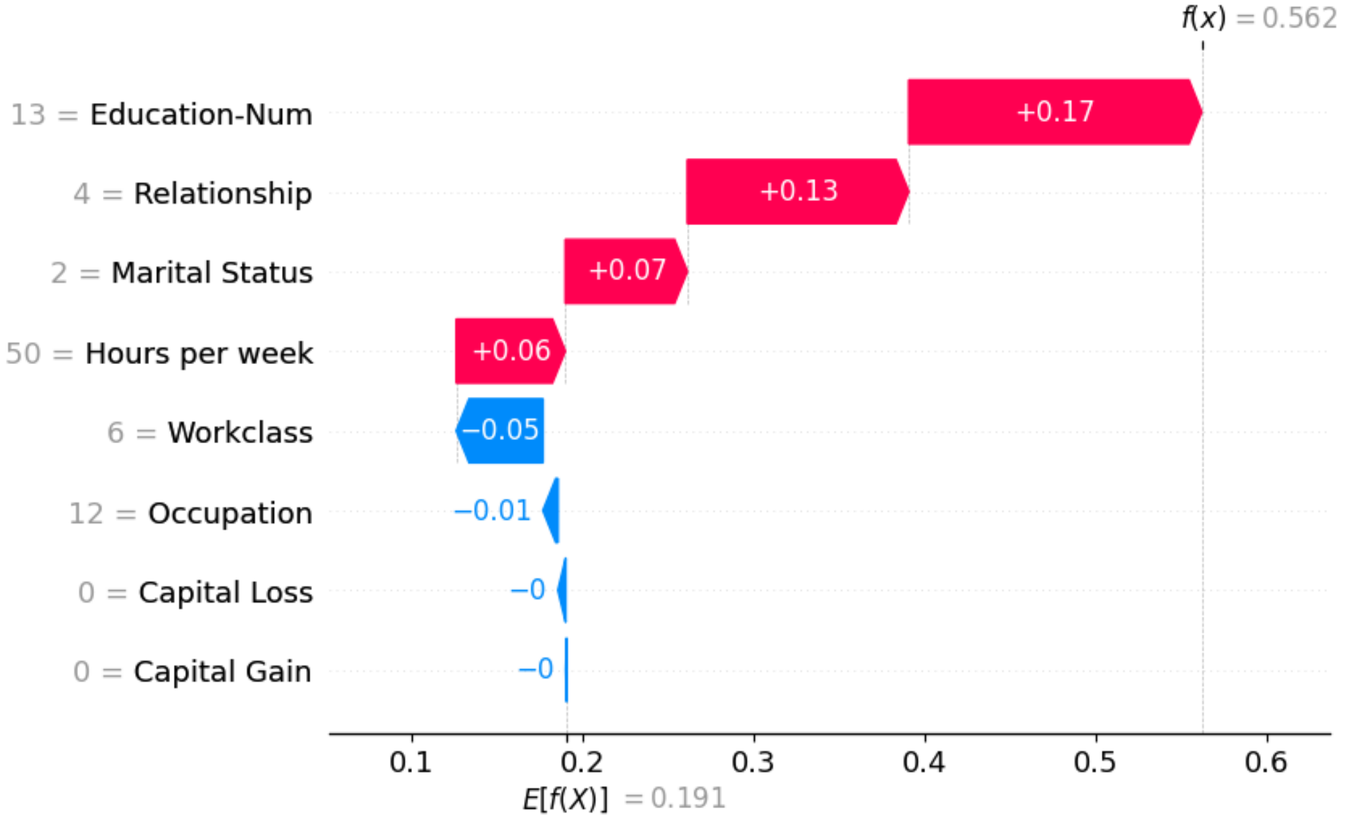


Loans (again)

New

URL: ttpoll.eu
Session ID: cs290

The plot on the right displays the SHAP values obtained for the prediction generated by our ML model for customer 1113.
What does this plot represent in terms of interpretability method?



✅ a. A local explanation

0% b. A global explanation

0% c. A feature importance analysis

0% d. A feature correlation analysis