Correction

EPFL

**Teacher : Dr. Cécile Hardebolle**
**CS-290: Responsible software**
**28/10/2024**
**90 minutes**

# Student 1

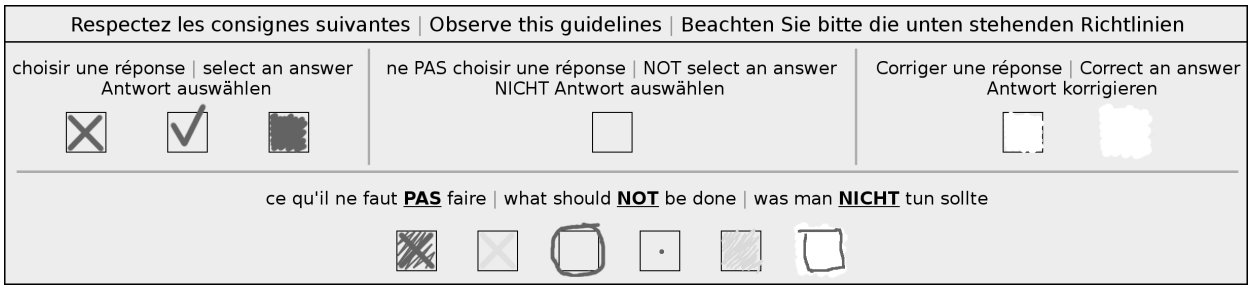
SCIPER : **999000**

**Do not turn the page before the start of the exam. This document is double-sided, has 13 pages, the last ones possibly blank. Do not unstaple.**

- Place your student card on your table.
- No paper materials other than one (1) A4 sheet of notes recto-verso is allowed to be used during the exam.
- Using a **calculator** or any electronic device is not permitted during the exam.
- **First part : single choice questions** (12 questions, 12 points)
  - for the singles choice questions, we give :

    +1 points if your answer is correct,
    0 points if you give no answer or your answer is incorrect.,
- **Second Part : true/false questions** (4 questions, 4 points)
  - for the true/false questions, we give :

    +1 points if your answer is correct,
    0 points if you give no answer or your answer is incorrect.
- **Third part : case studies** (3 questions, 20 points)
  - for each question, the number of points is noted above each question. Leave the checkbox empty.
- Use a **black or dark blue ballpen** and clearly erase with **correction fluid** if necessary.
- If a question is wrong, the teacher may decide to nullify it.



CORRECTION

## First part: single choice questions

For each question, mark the box corresponding to the single correct answer.

**Question 1**
A group of computer scientists with similar background, all experts in AI-based software development, is aware of cognitive biases. They are starting a new AI project for healthcare and they aim to minimize the impact of these biases when making design decisions.
Select the strategy they should use (best answer):

- [x] ... slow down the decision-making processes
- [ ] ... systematically include all members of their group to increase heterogeneity
- [ ] ... choose one or two of them to play the devil's advocate
- [ ] ... systematically include all members of their group to apply a participatory design method

**Question 2**
The CEO of a tech company stated in the media: "In the past, we've invested in technology to positively impact people's lives, and we have no intention of changing that strategy in the future - technology remains the best alternative."
We may interpret this as: (select the best answer)

- [x] Sunk cost fallacy
- [ ] Source cues
- [ ] System 1 thinking
- [ ] Illusory truth

**Question 3**
A start-up developed a machine learning model designed to connect people based on their personal interests. A big company has then bought the start-up and is currently using the algorithm to connect jobseekers with employers.
It is a case of...

- [x] ... deployment bias
- [ ] ... aggregation bias
- [ ] ... measurement bias
- [ ] ... intersectional bias

**Question 4**
A company has developed a complex algorithm to predict whether athletes suspected of doping actually do it. A positive result means that the algorithm classifies the athlete as at risk of doping, while a negative result means no risk of doping. The system has been used for 5 years and we have access to data about athletes that were indeed caught for doping. We found that the proportion of athletes predicted to dope amongst all predictions is higher for men rather than for women.
The fairness metric we have used is:

- [ ] error rate balance
- [ ] conditional use accuracy equality
- [ ] equal accuracy
- [x] demographic parity

CORRECTION

**Question 5** You develop a software that analyzes the weather forecast to send the population a notification in case of upcoming extreme rain (positive result).
In this context:

- [x] True Positive = rain is predicted and the prediction is correct
- [ ] True Negative = no rain is predicted, and the prediction is incorrect
- [ ] False Positive = rain is predicted and the prediction is correct
- [ ] False Negative = no rain is predicted, and the prediction is correct

**Question 6** You develop a software that analyzes the weather forecast to send the population a notification in case of upcoming extreme rain (positive result).
The False Negative Rate (FNR) is:

- [ ] The number of times rain is predicted among all times it actually didn't rained
- [ ] The number of times no rain is predicted among all times it actually didn't rained
- [x] The number of times no rain is predicted among all times it actually rained
- [ ] The number of times rain is predicted among all times it actually rained

**Question 7** A dilemma is...

- [ ] ... a situation in which you have to decide between two alternatives using a coin flip (better to leave things to chance)
- [x] ... a situation in which you have to weigh the pros and cons of each decision (and their consequences), with no decision 100% perfect or 100% unperfect
- [ ] ... a situation in which you have to weigh the pros and cons of each decision (and their consequences) and choose the one with the higher number of pros.
- [ ] ... a situation in which you should escalate the decision to your management line.

**Question 8** You work on a chatbot to provide students assistance on campus questions. At evaluation time it generates plausible nonsense with a 15% rate.
You probably face what we called an...

- [x] ... ethical issue
- [ ] ... ethical dilemma
- [ ] ... ethical blindness
- [ ] ... ethical sensitivity

Correction

**Question 9**
Here are three variables:

- Disinformation spread
- Public trust in information
- Development of disinformation software

We know that :

- As the spread of disinformation increases, the public trust in information decreases
- As the public trust in information decreases, bad actors see a growing opportunity to develop disinformation software exploiting this mistrust

In a causal loop diagram representing the dynamics between these variables, we would have (select the correct answer):

- ☐ The arrow between "Public trust in information" and "Development of disinformation software" has a positive sign and the arrow between "Development of disinformation software" and "Disinformation spread" has a negative sign.
- ■ The arrow between "Public trust in information" and "Development of disinformation software" has a negative sign and the arrow between "Development of disinformation software" and "Disinformation spread" has a positive sign.
- ☐ The arrow between "Public trust in information" and "Development of disinformation software" has a positive sign and the arrow between "Development of disinformation software" and "Disinformation spread" has a positive sign.
- ☐ The arrow between "Public trust in information" and "Development of disinformation software" has a negative sign and the arrow between "Development of disinformation software" and "Disinformation spread" has a negative sign.

**Question 10**
Fill the blanks:
If a piece of software behaves in a ___ way at first glance, but puts people of ____ at ____ , then it is a case of ____ discrimination.

- ☐ negative / several groups / an advantage / direct
- ☐ positive / identified groups / an advantage / inverse
- ■ neutral / specific groups / a disadvantage / indirect
- ☐ neutral / several groups / a disadvantage / direct
- ☐ negative / specific groups / a disadvantage / indirect

**Question 11** A bad actor launched a phishing attack on employees of Swiss public institutions to steal their login credentials. An online media outlet reported on it, with the most upvoted comments on the article criticizing the institutions for their inability to counter online threats, harming their reputation.
The harm to reputation can be classified as an impact that is:

- ☐ Direct
- ☐ Both direct and indirect
- ☐ Neither direct nor indirect
- ■ Indirect

Correction

**Question 12**

Imagine that you develop software for people from a single country. If you nonetheless envision cultural differences you are probably using the... (select the best answer)

- [x] edge case strategy
- [ ] stride strategy
- [ ] bad actor strategy
- [ ] people behind the data strategy

## Second part: true/false questions

For each question, mark the box (without erasing) TRUE if the statement is **always true** and the box FALSE if it is **not always true** (i.e., it is sometimes false).

You found a dataset with 5 variables, all self-reported by participants: eye-color, extraversion and 3 health-related variables. When analyzing the data you identify that:

- there are positive and substantial correlations among the 3 health variables
- there is a positive and substantial correlation between eye-color and extraversion
- there is no correlation between eye-color and the health variables

**Question 13** Eye-color is a latent variable

☐ TRUE ■ FALSE

**Question 14** Eye-color is a proxy for health

☐ TRUE ■ FALSE

**Question 15** Eye-color is a sensitive attribute

■ TRUE ☐ FALSE

**Question 16** Extraversion is a latent variable

■ TRUE ☐ FALSE

## Third part: case Studies

Answer in the empty space below. Your answer should be carefully justified, and all the steps of your argument should be discussed in details. Leave the check-boxes empty, they are used for the grading.

**Question 17: Case 1: Harms modeling - Social assistant chatbot** *This question is worth 5 points.*

☐0 ☐1 ☐2 ☐3 ☐4 ■5

**Scenario:**

In the realm of technological innovation, a revolutionary social-assistant chatbot emerges, designed to offer guidance on relationships. This cutting-edge human-centered AI, inspired by Snapchat's AI chatbot, aims to become an indispensable part of people's lives. Sarah and James are two individuals with contrasting lives. James, a young artist, craves genuine connections with like-minded people, while Sarah, a young consultant, struggles to balance her career with her personal life. Sarah and James turn to this chatbot for relationship advice. Its sophisticated algorithm analyzes their preferences, communication styles, and social behaviors to offer tailored suggestions for interactions. In addition to helping them identify others' emotions, it provides them with conversation starters and even helps plan memorable dates. As the chatbot gains traction and spreads throughout society, it becomes an integral part of society's social, economic, and political landscape. It reshapes how people approach dating and relationships, influencing not only their personal lives but also impacting the dating industry, advertising strategies, and even political campaign tactics. In addition, companies rely on the chatbot to predict the emotions of their staff and their clients to maximize their benefits. Yet, there are those who remain skeptical of the chatbot's far-reaching influence. Some individuals, wary of data privacy concerns and the potential for manipulation, opt to abstain from using the technology. They seek more traditional avenues for forming connections, believing in the value of genuine human interactions and the potential risks that come with relying on AI for personal advice.

**Task:**

Considering the following extract of the harms modeling table, describe what should go in the different cells:

- For cells A, B, C and E: describe 1 harm that corresponds to the category
- For cell D: indicate the corresponding harm category

Make sure to identify your answers with the corresponding letters [1 point / answer].

| **Category** | **Type of harm** | **Social assistant chatbot** |
|---|---|---|
| Humans | Physical injury | **A)** |
| Allocation of Resources | Opportunity loss | **B)** |
| Human Rights | Liberty loss | **C)** |
| | **D)** | Most intimate feelings are now "public" |
| Social System Harms | Social detriment | **E)** |

## Solution

(A) The chatbot could recommend aggressive behaviors or inappropriate conduct (e.g. a teacher asking for help to deal with an unruly pupil: the chatbot could advise inappropriate gestures instead of adequate mediation options, for example)

(B) Making a decision based on the chatbot's advice could lead people to miss some opportunities for social connection (e.g. random encounters). Biases in the chatbot could lead to some people being excluded from social relationships or from services or o˙ers.

(C) We can consider the case where the chatbot is used by companies to train employees in the emotional reactions expected to interact effectively with customers. Employees who best produce (and fake) the expected emotional reactions receive a bonus. The company encourages and discourages certain types of behavior, which increases conformity , decreases natural behavior, and limits the emergence of genuine human emotions.

(D) Privacy loss

(E) The chatbot could adapt its recommendations based on the socio-economic status of the user and then reinforce social discrimination. It could reinforce stereotypes (having recommendations based on identity factors). Over-relying on the chatbot's advice could prevent people from learning the skills needed to live in society.

CORRECTION

**Question 18: Case 2: Values analysis - Personalized deals** *This question is worth 5 points.*

☐ 0 ☐ .5 ☐ 1 ☐ .5 ■ 2

**Scenario:**

A webshop manager wants to offer interesting deals to the shop's customers, and thinks that it would be best to offer personalized deals to each one of them. As the customers provide their email address when registering, the manager creates the following script: for each user, the script finds some account linked to the mail address (Facebook, YouTube, Amazon, Retail stores, etc.) and buys the data related to that user. With that data, a personalized offer containing deals adapted to the centers of interest of the user is sent directly by email.

**Task:**

Your overall task is to perform an analysis of the values and value tensions involved for the different stakeholders in the case. Follow the 2 steps below:

1 [4 points] Consider some stakeholders in the case and identify 2 values which are supported by the software (= 2 value-based benefits) and 2 values that are opposed by the software (= 2 value-based harms).
Consider the value-based benefit/harm table template below and describe what would go in each cell for each of the values you identified:

(A) Describe the stakeholder

(B) Name the value (you should use the names in Appendix 3.1) and explain in your own words what the value means for this stakeholder

(C) Indicate if the value is supported (value-based benefit) or harmed (value-based harm) for this stakeholder

(D) Justify why it is supported / harmed by the software

Make sure to identify your answers with the corresponding letters.
A list of Schwarts's values is provided in appendix 3.1.

2 [1 point] Draw a value-based tension map showing a value tension and provide an explanation of the tension.

| **Stakeholder** | **Key Value** | **Benefits** | **Harms** | **Justification** |
|---|---|---|---|---|
| Stakeholder: **(A)** | Value name and description: **(B)** | Benefit or Harm: **(C)** | | It's a value-based benefit/harm for this stakeholder because: **(D)** |

**Appendix 3.1**

| | | |
|---|---|---|
| **Self-enhancement** | Power Resources | Power through control of material and social resources |
| | Power Dominance | Power through exercising control over people |
| | Achievement | Personal success through demonstrating competence according to social standards |
| | Hedonism | Pleasure and sensuous gratification for oneself |
| **Openness to change** | Stimulation | Excitement, novelty, and challenge in life |
| | Self-direction Action | The freedom to determine one's own actions |
| | Self-direction Thought | The freedom to cultivate one's own ideas and abilities |
| **Self-transcendence** | Universalism Tolerance | Acceptance and understanding of those who are different from oneself |
| | Universalism Concern | Commitment to equality, justice, and protection for all people |
| | Universalism Nature | Preservation of the natural environment |
| | Humility | Recognizing one's insignificance in the larger scheme of things |
| | Benevolence Dependability | Being a reliable and trustworthy member of the in-group |
| | Benevolence Caring | Devotion to the welfare of in-group members |
| **Conservation** | Tradition | Maintaining and preserving cultural, family, or religious traditions |
| | Conformity Interpersonal | Avoidance of upsetting or harming other people |
| | Conformity Rules | Compliance with rules, laws, and formal obligations |
| | Security Societal | Safety and stability in the wider society |
| | Security Personal | Safety in one's immediate environment |
| | Face | Security and power through maintaining one's public image and avoiding humiliation |

Table 1: Source: Schwartz et al. (2012).

# Correction

**1) Value-based benefits/harms**

(i)

(A) Stakeholder: Lydia, the store manager
(B) Value name and description: Power Dominance — For the store manager this value means control over its customers, the ability to directly influence their behaviors.
(C) Benefit or Harm: Benefit
(D) It's a value-based benefit/harm for this stakeholder because: If influence is successful, it means more client's buying her products and greater sales.

(ii)

(A) Stakeholder: Hari, the customer not caring about his personal data
(B) Value name and description: Hedonism — Hari buys a lot of things on the Internet, he wants the buying process to be as quick as possible so that he can spend more time using products than buying them.
(C) Benefit or Harm: Benefit
(D) It's a value-based benefit/harm for this stakeholder because: product recommendations are tailored to Hari's taste, he doesn't have to search anything on the website, he only selects recommended products making his customer experience very enjoyable.

(iii)

(A) Stakeholder: Lydia, the store manager
(B) Value name and description: Conformity Interpersonal — For the store manager this value means avoiding upsetting customers who care about transparency about how their data are used.
(C) Benefit or Harm: Harm
(D) It's a value-based benefit/harm for this stakeholder because: If :

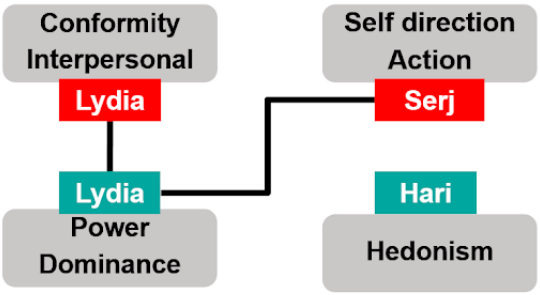
- the store manager doesn't communicate properly to customers how she is able to propose targeted products (= using the data from other companies)
- customers have doubts about the company's use of their personal data (e.g. as soon as they log on to their customer account for the first time they have specific products suggestions corresponding to prior buyings on other platforms)

This could upset customers into not buying.

(iv)

(A) Stakeholder: Serj, the customer unaware of the data resale market
(B) Value name and description: Self direction Action — For Serj, this value means making his own choices in life, including when buying products and services.
(C) Benefit or Harm: Harm
(D) It's a value-based benefit/harm for this stakeholder because: Serj's buying behavior is influenced by buying algorithms that he is not aware of.

**2) Value-based tension map**




**Explanation for the Self direction Action <-> Power Dominance tension:**
As a store manager, Lydia wants to have as much impact as possible on her customers' buying behavior so that her business is profitable. To do this, she uses a recommendation system to trigger purchases more effectively, based on previous purchases made on other platforms. This could create an unwanted influence on customers wishing to decide on their own actions.

**Question 19: Case 3: Universal Digital Identity Platform** *This question is worth 10 points.*

0 1 2 3 4 5 6 7 8 9 10

**Scenario:**

Consider this made up very futuristic case study: You are part of an international team tasked with developing a Universal Digital Identity Platform (UDIP). This platform is intended to be the ultimate authentication system and replace the paradigm of having one account for each service we use. UDIP provides every individual a unique digital identity, which can be used globally for accessing various services such as banking, healthcare, education, and government services. The platform will utilize biometric data, including facial recognition and fingerprints, to ensure secure and accurate identification and authentication. The aim is to streamline access to services, reduce fraud, and enhance global connectivity. Governments and private companies worldwide are eager to adopt this system to improve efficiency and security. The platform has the potential to become a foundational technology, potentially affecting billions of people.

**Task:**

As the ethics referee of the team, you are asked to anticipate potential consequences of the deployment of the platform in terms of safety and fairness. Follow the 3 steps below:

1 [1 point] Name one strategy seen in the course that you can apply for this task. Warning: you cannot use "Harm Modeling" for this case.

2 [3 points] Explain the strategy:

   (a) Justify why this strategy is appropriate for this task.

   (b) Describe briefly how to apply this strategy.

3 Describe the result of applying the strategy.

   (a) [3 points] Present one safety issue you identify in the case

   (b) [3 points] Present one fairness issue you identify in the case

   Specify any assumption you make (that is not clearly stated in the scenario) about the system and its stakeholders.

# Correction

**Solution:**

1 We can use Edge Cases (note that Ethical speculation and Ethics Canvas are other possible answers).

2 (a) The Edge Cases strategy is well suited for the exercise since we are being asked to anticipate the potential ethical consequences of deploying the platform, and this strategy aims precisely to foresee the potential consequences and opportunities to improve software by analyzing edge/extreme cases.

(b) To apply this strategy we have to consider 3 cases and answer 3 corresponding assessment questions:
- What happens if the UDIP reach global success, in other words if a diversity of people around the world are using it (global reach case)?
- What happens if the UDIP is adopted by a huge amount of people (mass adoption case)?
- What would happen if the UDIP was used for a long period of time (longevity case)?

We would need to conduct this analysis at every stage of the project, in particular in the initial design stage, in order to anticipate issues with scaling up and extending our user base internationally.

3 (a) In the case of "mass adoption", we assume that all critical services such as banking, education, emergency services, etc. would use this system. So a failure in the system is very problematic for its users because it would prevent them from accessing primary services. This is a safety issue caused by the overreliance on the system without appropriate backup plans thought ahead.

(b) In the case of "global reach", one could imagine that the system is adopted by several countries in order to allow their apps or services to be used by all people around the world. This can cause a fairness issue, because some countries would not have the infrastructure or the ressources to permit its residents to use this system, and at the same time some systems would only work with this authentication system. Therefore we would increase the gap between developed countries and the rest of the world.