Correction

EPFL

Teacher: Dr. Cécile Hardebolle
CS-290: Responsible Software
28/10/2025

Name Firstname

SCIPER: **XXXX**

Do not turn the page before the start of the exam. This document is double-sided, has 16 pages, the last ones possibly blank. Do not unstaple.

- Place your student card on your table.
- No paper materials other than **one sheet of notes in A4 format, double-sided**, are allowed to be used during the exam.
- The use of a calculator or **any other 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 point 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 point 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).
  The number of points is noted above each question. Leave the checkboxes 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.

Respectez les consignes suivantes | Observe this guidelines | Beachten Sie bitte die unten stehenden Richtlinien		
choisir une réponse | select an answer Antwort auswählen	ne PAS choisir une réponse | NOT select an answer NICHT Antwort auswählen	Corriger une réponse | Correct an answer Antwort korrigieren
ce qu'il ne faut **PAS** faire | what should **NOT** be done | was man **NICHT** tun sollte		

CORRECTION

First part: single choice questions

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

Question 1
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:

- ☐ Representation bias
- ☐ System 1 thinking
- ☐ Implicit stereotype
- ■ Sunk cost fallacy

Question 2
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 type of fairness metric we have used is:

- ☐ Conditional use accuracy equality
- ☐ Error rate balance
- ☐ Equal accuracy
- ■ Demographic parity

Question 3
Imagine that you develop software for people from a single country. If you nonetheless envision cultural differences in this context, which strategy are you probably using?

- ■ Edge cases
- ☐ STRIDE
- ☐ The people behind the data
- ☐ Bad actors

Question 4 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:

- ■ True Positive = rain is predicted and the prediction is correct
- ☐ False Positive = rain is predicted and the prediction is correct
- ☐ True Negative = no rain is predicted, and the prediction is incorrect
- ☐ False Negative = no rain is predicted, and the prediction is correct

Correction

Question 5
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.

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

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 no rain is predicted among all times it actually didn't rain
- ☐ The number of times rain is predicted among all times it actually didn't rain
- ☐ The number of times rain is predicted among all times it actually rained
- ■ The number of times no rain is predicted among all times it actually rained

Question 7
A group of computer scientists with similar background, all experts in software development, are starting a new software project for healthcare. They are aware of cognitive biases and want to minimize the impact of these biases when making design decisions.
Which is the only strategy that could effectively help them in this context?

- ■ Use a structured approach and slow down the decision-making process
- ☐ Choose one or two of them to play the devil's advocate
- ☐ Systematically include all members of their group to increase heterogeneity
- ☐ Systematically include all members of their group to apply a participatory design method

Question 8 You work on a chatbot to provide students assistance on campus questions. When evaluating the quality of the responses it provides, you identify that the responses contain hallucinations (i.e. content that is incorrect or wrong but looks perfectly plausible) with a 15% rate.
What type of situation are you facing?

- ☐ Ethical blindness
- ■ Ethical issue
- ☐ Ethical dilemma
- ☐ 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, which arrows would we have (select only one answer)?

- [x] 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 negative 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 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 positive sign and the arrow between "Development of disinformation software" and "Disinformation spread" has a positive sign.

Question 10 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.
what type of impact is the harm to reputation as a result of the attack?

- [x] Indirect
- [ ] Direct
- [ ] Both direct and indirect
- [ ] Neither direct nor indirect

Question 11 What is a dilemma?

- [ ] 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.
- [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 decide between two alternatives using a coin flip (better to leave things to chance)
- [ ] A situation in which you should escalate the decision to your management line.

Correction

Question 12

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.

What type of bias is likely to appear?

- ☐ Aggregation bias
- ☐ Measurement bias
- ☐ Intersectional bias
- ■ Deployment bias

Second part: true/false questions

For each question, mark either the box TRUE if the statement is true or the box FALSE if the statement is false.
1 point per question.

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 sensitive attribute

■ TRUE ☐ FALSE

Question 15 Eye-color is a proxy for health

☐ TRUE ■ FALSE

Question 16 Extraversion is a latent variable

■ TRUE ☐ FALSE

Third part: case Studies

Answer in the empty space provided. Use the extra pages at the end if you need more space.
Your answer should be concise but make your reasoning clear and your argument should be explicitly justified.
Leave the check-boxes empty, they are used for grading.

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

☐0 ☐.5 ☐1 ☐.5 ☐2 ☐.5 ☐3 ☐.5 ☐4 ☐.5 ■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:

- [4 x 1 point] For cells A, B, C and E: describe 1 harm that corresponds to the category (1-2 sentences for each harm)
- [1 point] For cell D: indicate the corresponding harm category

Make sure to identify your answers with the corresponding letters (no need to reproduce the table).

Category	**Type of harm**	**Description of harms**
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)**

Correction

Proposed answer

(A) If the chatbot gives users unsafe or negligent advice (such as suggesting meeting up with a stranger in a remote location or promoting risky activities), they may be exposed to physical injuries. They may tend to follow advice that expose them to risk because they consider the chatbot to be an "expert".

(B) In companies, using the chatbot to predict employees' emotions in the context of promotion or reward decisions may lead to some people being disadvantaged because they don't fit the underlying model (which may also be biased). Their access to opportunities would be unfairly restricted.

(C) The chatbot manipulates the emotions and decisions of users. This could be used to nudge them toward certain partners, products, or political opinions based on commercial or ideological interests. This would reduce the autonomy and independent judgment of users, affecting their freedom.

(D) Privacy loss

(E) Overreliance on the chatbot may weaken the social skills and empathy of humans on the long term, reducing their ability to navigate real emotional complexity. Over time, people could start to struggle to handle misunderstandings or emotional nuances in real relationships without the help of the chatbot, leading to an overall deskilling in society.

CORRECTION

Question 18: Case 2: Values analysis - Personalized deals *This question is worth 7 points.*

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.
We provide the following stakeholders:

- Lydia, the webstore manager.
- Hari, a customer who browses and purchases products on the webshop.

Follow the 2 steps below:

1 [5 points] Identify 2 values from stakeholders that are supported by the software (= 2 value-based benefits) and 2 values from stakeholders that are opposed by the software (= 2 value-based harms), i.e., 4 values in total.
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) [not graded] Name the stakeholder, who must be one of the 2 stakholders mentioned above

(B) [4 x 0.5 points] 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) [4 x 0.25 points] Indicate if the value is supported (value-based benefit) or harmed (value-based harm) for this stakeholder

(D) [4 x 0.5 points] Justify why it is supported / harmed by the software

Make sure to identify your answers with the corresponding letters (no need to reproduce the table).
A list of Schwarts's values is provided in appendix 3.1.

2 [2 points] Draw a value-based tension map showing at least 1 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

Table 1: Schwartz' Value Table - Source: Schwartz et al. (2012).

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

Correction

Proposed answer

1) Value-based benefits/harms

(i)

(A) Stakeholder: Lydia, the webstore 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 for this stakeholder because: If influence is successful, it means more client's buying products and greater sales.

(ii)

(A) Stakeholder: Hari, the customer who browses and purchases products on the webshop
(B) Value name and description: Stimulation — For Hari, this value means enjoying exciting and fresh experiences in online shopping, as the product interface is tailored to him and his tastes.
(C) Benefit or Harm: Benefit
(D) It's a value-based benefit because: The personalized offers make the shopping experience dynamic and tailored to his interests, giving him more novelty and satisfaction in discovering products.

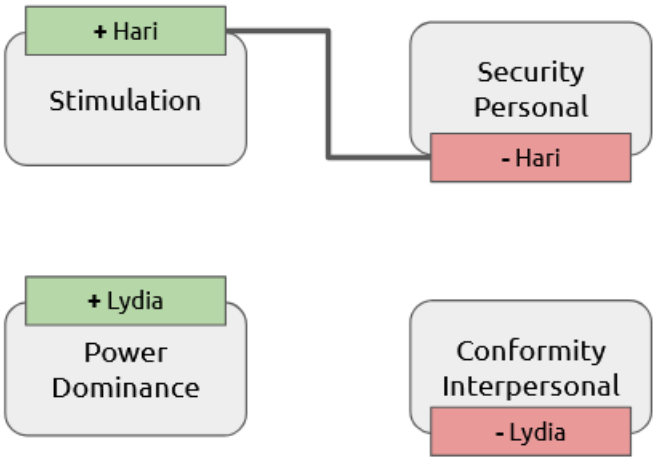
(iii)

(A) Stakeholder: Lydia, the webstore 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 harm for this stakeholder because: If the store manager does not properly communicate to customers how she is able to offer them targeted products (using data from other companies) and if customers have doubts about the company's use of their personal data, this could discourage them from buying.

(iv)

(A) Stakeholder: Hari, the customer who browses and purchases products on the webshop
(B) Value name and description: Security personal — For Hari, this value means protecting his private life and making decisions without hidden influence.
(C) Benefit or Harm: Harm
(D) It's a value-based harm for this stakeholder because: The recommendation system manipulates his purchasing behavior without his awareness by using its personal data from other websites, compromising his privacy and independence in decision-making.

2) Value-based tension map



Explanation for the Stimulation <-> Security Personal tension:
Some customers, such as Hari, appreciate the excitement and novelty of receiving personalized offers, which enhances their shopping experience. On the other hand, they may be concerned about the risks associated with their personal data and loss of trust in the webshop, knowing that they are used for business purposes. The software thus creates tension between offering a stimulating experience and ensuring data protection and customer security.

CORRECTION

Question 19: Case 3: Universal Digital Identity Platform *This question is worth 8 points.*

Scenario:

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 (1 sentence).

   (b) Describe briefly how to apply this strategy (2-3 sentences).

3 Describe the result of applying the strategy:

   (a) [2 points] Present one safety issue you identify in the case (2-3 sentences)

   (b) [2 points] Present one fairness issue you identify in the case (2-3 sentences)

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

Correction

Proposed answer #1:

1 We can use the Edge Cases strategy.

2 (a) The Edge Cases strategy is well suited for the exercise, as we are being asked to anticipate the potential ethical consequences of deploying the platform, and the Edge Cases strategy is used to analyze the software (consequences, possibilities for improvement) at the macro level perspective, i.e., on a global scale, which is what interests us in this case.

(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 happens if the UDIP is 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 any 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.

Proposed answer #2:

1 We can use the Ethics Canvas.

2 (a) The Ethics Canvas is well suited for this exercise as it provides a structured way to identify and mitigate ethical impacts, their dependencies, and to determine the stakeholders involved at the individual, meso, and macro levels, which is particularly relevant for this scenario.

(b) We use this strategy in the early stages of software development in order to systematically identify stakeholders, potential ethical impacts, and mitigation solutions. The strategy consists of 3 stages:

- Stage 1: identifying the relevant stakeholders (Who is affected by UDIP?)
- Stage 2: identifying the ethical impacts for these people and groups (What are the impacts of UDIP?)
- Stage 3: discussing remedial actions to address these impacts (How can we address the impacts of UDIP?)

3 (a) Block "Product or service failure" of the canvas: The platform relies on biometric data such as facial recognition and fingerprints for identification and authentication. This means that every person is permanently registered and monitored by the platform. If this data is not sufficiently protected, malicious actors could exploit it for harmful purposes. This is a security issue that leads to a safety issue.

(b) Block "Groups affected" of the canvas: Facial recognition is affected by algorithmic biases linked to ethnicity, which would lead to indirect discrimination for some of the affected groups. Furthermore, individuals "outside the system," such as refugees or those unwilling to share their data, would be systematically marginalized and deprived of essential services like banking, healthcare, and education.

Correction

Correction