Exercises

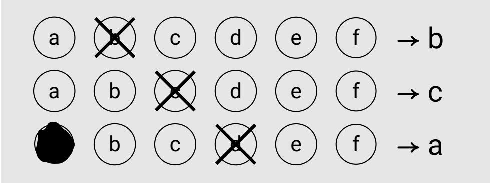
1	2	3

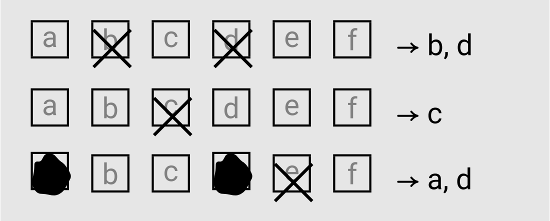
Surname, First name

____

1	1	1	1	1	1	1
2	2	2	2	2	2	2
3	3	3	3	3	3	3
4	4	4	4	4	4	4
5	5	5	5	5	5	5
6	6	6	6	6	6	6
7	7	7	7	7	7	7
8	8	8	8	8	8	8
9	9	9	9	9	9	9
0	0	0	0	0	0	0

5ARC0 Human and ethical aspects of AI
Exam Q1





Fill in your answer(s) to the multiple-choice questions as shown above (circles = one correct answer, boxes = multiple correct answers possible).

Particular Ans on paper exam instructions

- Write in a black or blue pen.
- You answer open-ended questions by using the text box. Provide your answers on the papers inside the answer box underneath a question. **If you need more space for your answers, use the extra space at the end of the exam, and clearly indicate there which question you continue answering. In the text box of the particular question, clearly state that you proceed with your answer on a different page.**
- Hand in all pages. Do not remove the staple. If you remove it anyhow, check that you hand in all pages.

Dear student,

You're about to take an exam. Write down your name and your student ID at the appropriate places above. Make sure that you enter your student ID by fully coloring the appropriate boxes. On the examination attendance card, you fill in the PDF number. You can find the correct number on the top of the first page of your exam (e.g. 1234.pdf).

Please read the following information carefully:

Date exam: 8 November 2022
Start time 09.00
End time: 12.00 (+30 minutes for time extension students)

Number of questions:
Maximum number of points/distribution of points over questions: Max 50 points

Method of determining the final grade: Actual score/ 5
Answering style: formulation, order, foundation of arguments, multiple choice:

Permitted examination aids **(delete as appropriate)**

- Scrap paper (fully blank)
- Calculator **(graphic/ basic) N/a**
- One A4 sheet of annotations N/a
- Dictionaries. If yes, please specify: English/ Dutch or English/ Other language
- Formula sheet (provided by the invigilator) N/a

Important:

- You are only permitted to visit the toilets under supervision
- Examination scripts (fully completed examination paper, stating name, student number, etc.) must always be handed in
- The house rules must be observed during the examination
- The instructions of subject experts and invigilators must be followed
- Keep your work place as clean as possible: put pencil case and breadbox away, limit snacks and drinks
- You are not permitted to share examination aids or lend them to each other
- Do not communicate with any other person by any means

During written examinations, the following actions will in any case be deemed to constitute fraud or attempted fraud:

- using another person's proof of identity/campus card (student identity card)
- having a mobile telephone or any other type of media-carrying device on your desk or in your clothes
- using, or attempting to use, unauthorized resources and aids, such as the internet, a mobile telephone, smartwatch, smart glasses etc.
- having any paper at hand other than that provided by TU/e, unless stated otherwise
- copying (in any form)
- visiting the toilet (or going outside) without permission or supervision

The final grade will be announced no later than fifteen working days after this examination took place. (first year BSc : Q4 within 5 working days and in the interim period no later than 5 working days before the 1st of September).

You can start the exam now, good luck!

Part I: Ethics of AI

2p **1a** Which of the items below are rights explicitly articulated in the Universal Declaration of Human Rights? You may choose more than one answer.

☒ Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.

☒ Everyone has the right to work, to free choice of employment, to just and favorable conditions of work and to protection against unemployment.

□ All people must be able to access the Internet; states may not unreasonably restrict an individual's access to the Internet.

□ Everyone has the the right to contraception, abortion, fertility treatment, reproductive health, and access to information about one's reproduction.

2p **1b** Which of the following is a true statement about the Belmont Report? You may choose more than one answer.

□ It contains a list of fundamental rights related to data processing.

☒ It emphasizes the importance of protecting vulnerable subjects in research.

□ It was written just after World War II in response to Nazi atrocities in the conduct of research.

☒ It was influential not just in the United States, but around the world.

2p **1c** Please paraphrase the principle of respect for persons from the Belmont Report, including two main aspects of the principle.

"Respect for persons incorporates at least two ethical convictions: first, that individuals should be treated as autonomous agents, and second, that persons with diminished autonomy are entitled to protection. The principle of respect for persons thus divides into two separate moral requirements: the requirement to acknowledge autonomy and the requirement to protect those with diminished autonomy." (Belmont Report).

4p **1d** Consider a *loan risk assessment system* developed as a machine learning application: "A supervised learning algorithm can be used quite straightforwardly to correlate previous applicants' personal data such as income, age, and home address with their eventual ability to repay loans in a timely manner. On the basis of these correlations, the system can take a new individual's personal data as input and generate an output to estimate the financial risk a bank would incur by accepting that individual's application." Identify at least three major ethical concerns that relate to such a system using concepts from Müller's article on the ethics of AI. Briefly explain why each concern arises in relation to this type of system.

Müller's article gives us three main concepts that could apply here, and one that applies to a small degree:
Privacy
Opacity of AI Systems
Bias in Decision Systems
[To a small degree] Automation puts people out of work
In the context of autonomous military robots, Müller also mentions responsibility gaps and the argument for keeping humans in the loop. This could also apply if there is not appropriate human evaluation of loan decisions or oversight.
Three points are granted for mentioning and explaining 3 of these (or another equivalent concept), and one point for giving insightful explanations that relate to the specific case.

2p **1e** The Guidelines for Trustworthy AI by the EU High Level Expert Group state that four main ethical principles are fundamental to the development, deployment and use of AI. Which of the following are among the principles mentioned? You may choose more than one answer.

□ explicability

□ privacy

☒ respect for human autonomy

□ technical robustness

2p **1f** In the debate about trust in artificial intelligence, some defend the view that it is a bad idea to talk about trust in AI. What are the two main reasons given for this view?

The two main arguments against the idea of trust in AI, as discussed in the lecture and the text of Nickel, are (1) that it is impossible to trust AI, because AI doesn't have the characteristics that we associate with trusted people such as good will and relationality, and (2) that it is bad to trust AI because it diverts attention from human agency and responsibility.

3p **1g** In Nyholm's book *Humans and Robots*, he argues that there are two main options for trying to make humans and robots/AI get along. What are they? Please give an example that illustrates the options.

Nyholm argues that we can try to make robots adjust to human behavior, or we can try to make humans adjust to robot behavior. His main example is self-driving cars, which can either adjust to imperfect human driving, or can drive "perfectly" forcing humans to change.

↳

2p **1h** In cases where an online platform has a similar purpose to an offline form of human activity, what approach to privacy protection is adopted by Nissenbaum's theory of contextual integrity? Using the example of online education or a comparable example, explain how her theory defines a privacy violation.

Nissenbaum's theory starts by identifying appropriate norms of information flow for a context. When that context exists similarly offline or online, such as in a university course, she holds that the norms should be roughly the same because the purposes and values are the same. A privacy violation can be defined by looking at the norms for the offline context: e.g., in an offline course, an instructor would be permitted to surveil students during an exam, therefore this is also permitted during an online exam conducted in students' homes. However, publicizing a recording from that exam would violate privacy in the same way as filming an on-campus exam and then publicizing the recording. Importantly, the emphasis is not on consent.

1p **1i** Which of the following statements best describes the difference between nudging and hypernudging? Choose one answer.

☒ a Nudging uses general knowledge about psychological biases and heuristics to influence people's choices, whereas hypernudging uses person-specific profiling to influence people's choices.

○ b Nudging uses general knowledge about psychological biases and heuristics to influence people's choices, whereas hypernudging is a form of manipulation.

○ c Nudging is a form of manipulation, whereas hypernudging is respectful of autonomy.

○ d Nudging is respectful of autonomy, whereas hypernudging is a form of manipulation.

3p **1j** How does Zednik define the Black Box Problem? Give an example in which an explanatory AI technique such as input heatmapping helps to mitigate the problem for a given stakeholder. Specify the stakeholder and how it mitigates the problem.

Zednik defines the Black Box problem in terms of opacity relative to the informational needs of a stakeholder: it is agent-relative and epistemic. For example, an algorithm for determining the type of cancer based on a scan image is opaque to human pathologists, if they do not concur with the diagnosis and as professionals cannot verify how the algorithm reached its conclusion. For a patient or for a hospital administrator, the opacity would be defined differently. An XAI technique such as input heatmapping can show the pathologist which parts of the image were most relevant to the diagnosis, thereby giving them some knowledge of how to attend to the image differently and mitigating the opacity. Complete answers must describe Zednik's view, not just generic definitions, and they must give a domain-specific example (e.g. banking).

1p **1k** According to Nyholm's book *Humans and Robots*, what is the main theoretical problem with how most ethicists understand the agency of robots and AI? Choose one answer.

- ○ a They understand it in terms of hierarchical agency.
- ○ b They understand it in terms of Kantian autonomy.
- ○ c They understand it in terms of collective agency.
- ☒ d They understand it in terms of individual agency.

1p **1l** Which of the following answers best completes the following argument of Schwitzgebel & Graza in order to make it *valid*?

(1) If entity A and entity B differ in how much moral consideration they deserve, there must be some relevant difference between the two entities that grounds this difference in moral status.
(2) ______________
(Conclusion) Therefore, there are possible AIs that deserve a degree of moral consideration similar to that of human beings.

a (2) Robots could be created that look visually exactly like human beings.

[✓ b] (2) There are possible AIs that do not differ in relevant respects from human beings.

c (2) It is cruel to treat a robot badly in the same way that it is cruel to treat a human badly.

d (2) Currently existing AIs are relevantly different from human beings.

Part II: Humans and AI

Section mostly not relevant to 2023-2024

1p **2a** Which aspect of an explanation would a human/user (not a data scientist) value the least important:

- (a) An accurate and complete explanation
- (b) An explanation that tells me what I need to get a different outcome
- (c) An explanation that focuses on abnormal causes
- (d) An explanation that gives a single good reason

2p **2b** Amershi et al. (2014) discuss Crayons (Fails and Olsen, 2003), and early interactive machine learning tool.
Describe the tool and discuss how Crayons is different from classical model building and updating as discussed by Amershi.

2p **2c** Use the type of AI technology incorporated in digital cameras to explain Shneiderman's (2020) more recent vision on the discussion of human control versus automation.

not 1D thinking (more automation means less user control). A digital camera provides a lot of automation and does the job without any control quite well, but does allow the user to control settings quickly and effortless

2p **2d** Amershi at al. (2014) states “Users are people, not oracles”. What was meant by this and does the Crayons system adheres to this guideline? Discuss why.

Not relevant to the 2023-2024 course

3p **2e** LIME is an XAI tool. It can be used to explain, for instance to a customer, why he got (or not) a loan, and it works even for black box models. Describe how LIME works and discuss one drawback of this type of tool with respect to framing towards the decision class.

LIME is an algorithm that can explain the predictions of any classier or regressor in a faithful way, by approximating it locally with an interpretable model. (Model agnostic, local fidelity, intepretable)
It automatically frames the results towards the decision class, so if it predicts you do not get a loan, it might show you negative factors that predict why you do not get a loan,or in ther words factors that contribute towards getting is (double negation). People would like these type of explanations to be framed towards the positive class (getting a loan) and doing that made lime output easier to understand in the study discussed in class.

3p **2f** In the lecture on Digital Nudging and Personalization, research was discussed on partitioning and ordering effects on Healthcare insurance decisions (Dellaert et al. 2022 study 1). Discuss based on that study why choice architectures cannot be avoided, and how/when partitioning can both increase and decrease the decision quality (in terms of costs of insurance).

2p **2g** In the evaluation of the Rasch-based energy recommender system by Starke et al. (2017), The authors measured both behavioral data (clicks/choices) and subjective measures. Among the subjective measures, there were more perceptual measures (perceived effort/support) and experience measures (satisfaction). Discuss based on the user-centric evaluation framework what the benefits are from measuring in this way, compared to just looking at the choices people made with the system.

↳

2p **2h** In the research on genre exploration by Liang and Willemsen (2022), it was found that people with high musical expertise were differently affected by nudges than people with low expertise. Discuss the results and relate this to the differences in long and short term preferences for music between these two groups.

1p **2i** In his lecture on Intelligence Augmentation, Wijnand IJsselsteijn discussed the use of “cognitive tools” to support and enhance our thinking process. Which term is closely associated with actions performed to uncover information that is hidden or hard to compute mentally?

- ⓐ Epistemic action
- ⓑ Cognitive structuring
- ⓒ Pragmatic action
- ⓓ Mental transformation

4p **2j** idjourney is an independent research lab that produces a proprietary artificial intelligence program that creates images from textual descriptions, similar to OpenAI’s DALL-E and the open-source Stable Diffusion. In the lecture by Pei-Ying Lin and Ralf Schmidt, and their corresponding paper, they discussed community responses (based on Discord and Twitter data) to the use of Midjourney as a form of creative AI. In their lecture, Pei-Ying and Ralf highlighted some ethical issues when using Midjourney, which you subsequently were asked to discuss in groups.

Based on the lecture, your readings, and your group discussion, **please name and discuss 3 separate ethical issues** related to generative AI art, in particular Midjourney. For each issue, please highlight the stakeholders involved, and assess the likelihood of this issue occuring in the real world, and its potential impact on direct and indirect stakeholders.

1p **2k** In both the lecture by Wijnand IJsselsteijn and the essay by Carter and Nielsen, Intelligence Augmentation (IA) is discussed.

a) What is the main conceptual difference between artificial intelligence and intelligence augmentation?

2p **2l** b) In the conclusion of their essay, Carter and Nielsen put interface design at center stage. They write “It is conventional wisdom that AI will change how we interact with computers. Unfortunately, many in the AI community greatly underestimate the depth of interface design, often regarding it as a simple problem, mostly about making things pretty or easy-to-use.” Describe the alternative view of interface design that Carter and Nielsen propose, and why this matters to IA.

Backup space

3 If you need more space for your answers, use this extra space and clearly indicate here which question you continue answering. In the text box of the particular question, clearly state that you proceed with your answer in this space.

↳

⤵