Responsible AI Assignment 2 Description

1. **Introduction**

2. **Description**

3. **Deliverables**

4. **Requirements**

1. Introduction

Requirements and guidelines

The objective of this assignment is to analyze an existing system, or a system to-be-designed where AI is used to solve a specific problem in a domain of interest.

This assignment should be done in groups of two-three students.

The groups should be communicated with the course examinator by 31.03.2024.

2. Description

Select an existing system or a system to-be-designed that uses AI.

The aim of this assignment is to analyze an existing system, or a system to-be-designed where AI is used to tackle a certain problem in a domain of interest (e.g., work, hobby)**. This implies identifying ethical issues (e.g., transparency, trust, fairness, security, privacy) and suggesting technical solutions that deal with them (e.g., value alignment, governance, explainability) to make the AI system more responsible.**

Establish resources to be used for the system selected.

In this assignment you will either select one of the topics suggested to you or you will define a topic by yourself. A list with suggested topics is provided to you on yOUlearn in the Assignment 2 section.

Important to mention that the nature of this assignment is not per se or just technical as it needs to reflect firstly the ethical reasoning about a technical artefact and secondly its technical design.

In case that you define your own topic considering the following aspects:

- **The case containing the system** to be analyzed is concrete and its description is available (e.g., online, personal documentation).
- **The stakeholders involved/impacted** in this case are known and indicated.

Bibliography: for this assignment, you can use one or more the following resources discussing the intersection between ethical aspects, methodologies, and AI:

- **Value Sensitive Design methodology** available here: Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. *Early engagement and new technologies: Opening up the laboratory*, 55-95.
  Optional: in case you are interested to read about the twenty years of history of using VSD in different projects, you can read the following article: Winkler, T., & Spiekermann, S. (2021). Twenty years of value sensitive design: a review of methodological practices in VSD projects. *Ethics and Information Technology*, *23*, 17-21.
  Optional: in case you are interested in reading a recent book on Value Sensitive Design / Value Based Design, you can read the following book: Friedman, B., & Hendry, D. G. (2019). *Value*

sensitive design: Shaping technology with moral imagination. Mit Press.

Important: The Optional resources here provided are not required for making this assignment.

- **Value alignment** available here: Vamplew, P., Dazeley, R., Foale, C., Firmin, S., & Mummery, J. (2018). Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology*, *20*, 27-40.
- **Connection between ethical principles, values, and AI system development phases available here:** Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, *26*(4), 2141-2168.
- **AI ethics principles** available here: Khan, A. A., Badshah, S., Liang, P., Waseem, M., Khan, B., Ahmad, A., ... & Akbar, M. A. (2022, June). Ethics of AI: A systematic literature review of principles and challenges. In *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering 2022* (pp. 383-392).
- **AI ethical principles, values, and impact** available here: UNESCO (2021). Recommendations on the ethics of Artificial Intelligence.
- **AI Impact Assessment** available here: EGP (2018). Artificial Intelligence Impact Assessment.
- **Ethics Canvas** available here: The Open Ethics Canvas.

  (Optional: this is a framework introduced by Kalra, A. (2020). Artificial Intelligence Ethics Canvas)
- **Explainable AI** available here: Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115 and

  Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38.
- **Design practices example for Explainable AI** available here: Liao, Q. V., Gruen, D., & Miller, S. (2020, April). Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).
- **Stakeholders and governance perspective and responses on ethical AI issues** available here: Stahl, B. C., Antoniou, J., Ryan, M., Macnish, K., & Jiya, T. (2022). Organisational responses to the ethical issues of artificial intelligence. *AI & SOCIETY*, *37*(1), 23-37.
- **Example: Ethical concerns for military swarm robots** available here: Wasilow, S., & Thorpe, J. B. (2019). Artificial intelligence, robotics, ethics, and the military: A Canadian perspective. *AI Magazine*, *40*(1), 37-48.
- **Example: Ethics and human control for drones** available here: Steen, M., van Diggelen, J., Timan, T., & van der Stap, N. (2022).

Meaningful human control of drones: exploring human–machine teaming, informed by four different ethical perspectives. *AI and Ethics*, 1-13.

- **Example: Guidelines for Human-AI interaction** available here: Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019, May). Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1-13).
- **Example: Ethical concerns and strategies in healthcare** available here: Li, F., Ruijs, N., & Lu, Y. (2022). Ethics & AI: A Systematic Review on Ethical Concerns and Related Strategies for Designing with AI in Healthcare. *AI*, *4*(1), 28-53.

These resources present both the background for assessing ethical principles, values, impact and perspectives of stakeholders, ethical impact assessment, the use of the Ethics Canvas framework that can be used by being instantiated for this assignment, and Explainable AI, plus a series of examples from different societal domains.

Write a design report around 2500 words.

Write a design report of around 2500 words describing the existing/to-be-designed system following the requirements presented below.

Reading: read the materials corresponding to the topic selected.

Communication: communicate with the teaching team using the Discussion forum in yOUlearn and by e-mail.

Time: the expected time needed for this assignment is about 30 hours.

Deadline on 16.06.2024

Deadline: submit this assignment by 16.06.2024.

This assignment is 40% of the course grade.

Grading: this assignment represents 40% of the final grade associated with this course. To pass this assignment the grade should be higher or equal to 5.5.

3. Deliverables

The deliverables of this assignment are as follows:

- The design report of around 2500 words (exclusive references with resources used).
- List with resources used in case they are available online or an archive with the used documents.
- Optional: just in case that the system is designed/implemented: Ethics Canvas instantiation for this design assignment or the design architecture, source code as a Jupyter Notebook, and dataset(s) used, or any other additional materials used.

Report structure

4. Requirements

The design report should be structured as follows:

- **Personal Information:** provide the following information about yourself and the course:
  - Course name and ID.
  - Name and student ID.
  - Title of the presentation.

- **Introduction:** introduce the context of this assignment, aim of this system (either existing or to-be-designed), and how is AI used.
  - **Introduce the context** where the system exists/should exist.
  - **Discuss the aim/main functionality of the system and how is AI used/should be designed to be used.**
  - **Briefly mention who are the stakeholders involved/impacted by using this system.** Consider here the following flow:

Not a responsible AI system.

Action of AI system -> Impact on Stakeholders => deal with the Impact on ethical norms and values

This implies that there is an ethical problem in the *Action of AI system* component => responsibility controls should be applied.

This section aims to provide an introductory perspective and prepare the ethical issue(s) that will be further explained.

- **Ethical Issue Identification:** discuss the ethical issue(s) (see here ethical principles and values), assess the impact produced, and the stakeholders involved/impacted.
  - **Impact assessment**: type of impact and stakeholders impacted.
  - **Ethical issue(s) analysis**: analyze the ethical issue(s) identified. In case that you consider analyzing a solution which has its source code available, you can directly refer to it and provide a link/attachment to/with it.

This section provides a list with ethical issues(s) and values fringed and stakeholders affected.

- **Responsible Solution Design:** once the ethical issue(s) is/are identified and the impact and stakeholders are known, **one or more design directions are considered to making the AI system responsible: value alignment, Explainable AI, governance etc.**
  - **Design Introduction**: introduce the type of method you plan to consider applying corresponding controls and transform the existing system into a responsible AI system.

- **Design Description:** discuss the responsible AI methods applied using one or more of the instruments of choice (e.g., just text, design architecture, short implementation etc.).

e AI system became responsible.

This section aims to fix the first component of the flow presented below, e.g., the Action of AI systems. It does that by proposing controls that should be considered since the design of AI systems. Hence, the flow becomes:

Respect to ethical norms and values through their integration in the design of the AI system -> Responsible Action of AI system -> Positive/Desired Impact on Stakeholders => Protection of ethical norms and values

This implies that responsibility controls are properly applied => the AI system has become a Responsible AI system.

- **Discussion and Conclusions:** reflect on the design discussed/proposed and provide concluding remarks for the issue(s) tackled in the AI system considered.
  - **Discuss the solution proposed** illustrating the controls applied to make the AI system responsible.
  - **Discuss the impact change to relevant stakeholders** impacted using the AI system considered.
  - **Concluding remarks based on personal reflection and future perspectives** in this domain.

Good luck!