Summative Assignment

Module code and title	COMP2261 Artificial Intelligence
Academic year	2023-24
Coursework title	Bias in AI coursework
Coursework credits	5 credits
% of module's final mark	25%
Lecturer	Shauna Concannon
Submission date*	Tuesday, March 19, 2024 14:00
Estimated hours of work	10 hours
Submission method	TurnitIn submission point on Ultra

Additional coursework files	*Assignment brief only*
Required submission items and formats	*One file as PDF*

* This is the deadline for all submissions except where an approved extension is in place.

Late submissions received within 5 working days of the deadline will be capped at 40%.

Late submissions received later than 5 days after the deadline will receive a mark of 0.

It is your responsibility to check that your submission has uploaded successfully and obtain a submission receipt.

Your work must be done by yourself (or your group, if there is an assigned groupwork component) and comply with the university rules about plagiarism and collusion. Students suspected of plagiarism, either of published or unpublished sources, including the work of other students, or of collusion will be dealt with according to University guidelines (https://www.dur.ac.uk/learningandteaching.handbook/6/2/4/).

Ethics & Bias in AI Module Assignment

Assignment introduction

In this piece of work you will produce an Ethical Impact Assessment of a piece of technology that utilises AI and develop a set of recommendations in response to your findings. You will use Value Sensitive Design to inform your approach.

Assignment format and submission

You should submit your Ethical impact assessment and recommendations report (1500 words MAX) via Turnitin. **The deadline for your submissions is 2pm on Thursday 19 March 2024.**

Please note that all submissions will be subject to plagiarism and collusion checks. If you use any generative AI tools you must clearly reference these.

Marking overview

1a) Introduction	15
1b) Ethical impact assessment	35
1c) Recommendations	45
Writing skills, clarity of the document and consistent formatting, referencing, presentation	5
TOTAL MARKS	**100**

Assignment descriptions and tasks

While working for a start-up, you are leading a machine learning project that involves designing an algorithm to perform an emotion recognition task on images of human faces. You have to produce an ethical impact assessment and a set of recommendations for the project before work commences. Specifically, you are required to analyse the task of emotion recognition using Value Sensitive Design and provide a set of recommendations that makes use of existing guidelines and toolkits.

Your report should include three sections:

- Introduction
- Ethical impact assessment informed by Value Sensitive Design
- Recommendations

Details on what should be included in each section are detailed below.

1a) Introduction (15 MARKS)
In the introduction you will introduce the task, convey to colleagues in the company why ethical impact assessments are useful/necessary and summarise the Value Sensitive Design approach. You should anticipate that the system you produce could be deployed by various different organisations or be made available to others via open sourcing.

In this section you should:

- Provide a description of what Value Sensitive Design is, providing arguments for why it is appropriate or beneficial for this task.
- Provide a description of the machine learning task you are evaluating, the data sources that will be used to develop the system, the types of predictions or outputs it will produce.
- Provide details of potential use cases of the system and any immediately apparent ethical issues that should be considered in advance of development.

1b): Ethical impact assessment informed by Value Sensitive Design (35 MARKS)

For the ethical impact assessment you need to identify the key stakeholders and values relevant to the project.

You can present the following information in a table:

- Identify key stakeholders and differentiate between direct and indirect stakeholders
- Detail which values are relevant to different stakeholder groups in this project
- Assess and describe key risks and potential harms for each stakeholder group

Example table:

Stakeholder	Values	Potential risks / harms
Include here the stakeholder, describe the role if not immediately obvious, and whether they are an indirect or direct *e.g. Doctor, system end-user (direct)*	***List key values, plus the motivation for why they are relevant*** *e.g. Accuracy - professional integrity will be impacted system's performance*	***What are the potential risks from this stakeholder groups perspective*** *e.g Risk of misdiagnosis could impair care the doctor is able to provide, harm the doctors reputation or result in disciplinary action*

Additionally, you should include a short-written analysis of the key value conflicts.

- Explain how you identified / selected the human values and provide any additional context on why they are relevant to the particular stakeholders.
- Provide a summary of the value conflict analysis. Identify and describe any potential value conflicts and provide an argument for how these may be resolved, or which group's interests should be prioritised.
- Articulate any additional considerations or actions. Outline any additional empirical or technical investigations you would need to carry out to complete your assessment. Can you identify any potential areas that might pose an issue (you can draw on similar or related examples from the literature or case studies where appropriate)?

1c): Recommendations & Considerations (45 MARKS)

Write a set of recommendations. This should be informed by section b). You can include guidance on **data collection and preparation, task design, or task deployment**. Your recommendations should include **both technical and non-technical** components. You can draw on and utilise existing AI ethics frameworks and bias mitigation toolkits (be sure to explicitly reference these, including online sources).
You should take a critical and reflective approach, noting potential strengths and limitations. Your recommendations do not need to solve every issue, however, they should highlight key areas that you have selected as most relevant and could include some of the following:

Motivation for recommendations

- Describe how the findings from your **Ethical impact assessment** inform what additional information you will need to gather or what approach you propose.

Datasets

- Articulate the key considerations relating to data preparation and data quality.
- Assess what data is required for the task. Are existing datasets available? If so, do they have any limitations? If not, what type of data will you collect?
- Describe the data preparation / collection / annotation / documentation that is required.
- Highlight any processes that should be put in place to guide / monitor this?

Risk and bias mitigation measures

- Provide 2-3 examples of **specific tools, techniques and methods** that could be adopted to mitigate **or minimize harm** to individuals or groups of individuals; **improve the safety, fairness or equity** of the system and/or **promote responsible and ethical development**
- You can draw on techniques covered in the lectures, practical sessions, course reading or your individual inquiry.
- Articulate how the techniques selected connect to challenges raised in the ethical impact section and be explicit about the ways in which they serve to address this.

Critical assessment and limitations

- You should also include a critical assessment of the recommendations. You should **explicitly highlight any limitations** associated with the proposed methods.
    - Evaluate how suitable they are for this particular task.
    - Specify what the selected tools/techniques *can* and ***can't*** do.
- Reflect on the challenges you foresee that are not easily remediable or that existing tools / approaches may not adequately address?
- Identify any areas that require further action, research or evaluation before definitive recommendations can be made. State this together with any suggestions for how to approach this.