SIT799 Human Aligned Artificial Intelligence

Distinction Task 3.1: Mini literature review on bias and discrimination in AI

Overview

During week 3, you have been introduced to: what is Bias and discrimination in AI; Examples of AI bias and discrimination; A set of possible causes that lead AI algorithms to learn unhealthy stereotypes; Some solutions to fight bias and discrimination in AI. To better understand bias and discrimination in AI, in this assignment, you are asked to do a mini literature review.

A literature review lets us know what research has come before us, and what is still to come. The goal of a literature review is to survey ideas and developments in a field done at the start a research project to:

- Understand the state-of-the-art in a particular area;
- get an understanding of the important ideas and background knowledge of an area
- identify gaps or opportunities for further research;
- compare work critically, e.g., to argue why a certain piece of research is original or better compared to other.

A literature review is not just paragraphs of summaries of papers, it is an attempt to link, connect, compare and contrast prior research in a field to:

- provides a roadmap of an area;
- presents a new perspective on a given area;
- provides a classification (or taxonomy) of the published research in a given area;
- provides a chronological overview of how an area has developed.

To complete this assignment, you need to refer back to Week 3 lecture material.

Submission Details

Write a mini **800 words** literature review on bias and discrimination in AI, describing the work from research papers as follows:

- an introduction to the application domain;
- key concepts and ideas, obtained from the papers;
- summarize, contrast, and compare the research done across the papers (noting similarities and differences in problem being addressed, approach taken, outcomes, technology used, pros and cons, etc) including experiments done (if any), results and limitations, advantages and disadvantages; trace the development of ideas across the papers;
- a conclusion of key ideas and any ideas for future work.

Note that a literature review typically has many papers reviewed, depending on the topic of review, and would aim to cover as many papers in the area as possible, but this exercise is to provide a general overview on discrimination and bias in AI from different perspectives (i.e., as represented by the reviewed papers).

Possible sources of research papers are these electronic databases:

- IEEExplore digital library
- Google scholar
- ACM Digital library

Constraint

The submitted report should not exceed **800 words** in length with at least Review at least 10 papers reviewed. The report should have a **high-quality writing style**.