Christian-Albrechts-Universität zu Kiel
Technische Fakultät

Creation and Evaluation of a Self-Learning Course for Data Hazard Labels

Bachelor Thesis

Joost Krüger

17.07.2024

Table of Contents

Introduction

- Creation of a self learing course using LiaScript
- Aimed at university students
- Intersection between ethics and data science
- Course can be used as supplementary material for lectures

Data Hazards

Created and maintained by the Data Hazard Project

Goal: Create a tool to help initiate ethics discussion on data science

Data hazard labels:

- **Similar to real Hazard labels**
- **Currently 11 labels for data science and 5 for synthetic biology**
- **Can be assigned to any projects using data science**
- **Not strict true or false**
- **Multiple labels can apply**

Data Hazards

Generic Data Hazard

Reinforces Existing Biases

Automates Decisionmaking

Ranks or Classifies People

Danger of Misuse

May cause Harm

Not present:

- **High Environmental Cost**
- **Risk to privacy**
- **Lacks Informed Consent**
- **Difficult to understand**

Current Prototype

Data Hazard Labels

- **Reinforces Existing Biases**
  - Definition
  - Examples
  - Prevention of Bias
  - Videos
  - Quiz
- **Ranks or classifies people**
  - Definition
  - Examples
  - Prevention of improper ranking and classification
  - Videos
  - Quiz
- **Automates decision making**
  - Definition
  - Examples
  - Precautions for automated decision making

Definition

This label indicates that data, algorithms, or software could lead to unfair treatment of individuals or certain groups. There are various reasons why this hazard might arise.

One source of bias can be the input data itself. If the data used contains biases and these are not corrected, the bias will be perpetuated. For example, if historical data reflects societal biases, algorithms trained on this data will likely replicate those biases.

Additionally, the design of an algorithm can introduce bias. If an algorithm gives more importance to certain characteristics, it might favor some individuals or groups over others. This can happen even if the intention is not to discriminate.

Societal biases can also infiltrate data and algorithms, reflecting and reinforcing existing stereotypes. These biases are often unintentional and unwanted, but if left unaddressed, they can have negative and unforeseen consequences. It is crucial to recognize and mitigate these biases to ensure fairness and equity in the use of data and technology.

Current Prototype

Examples

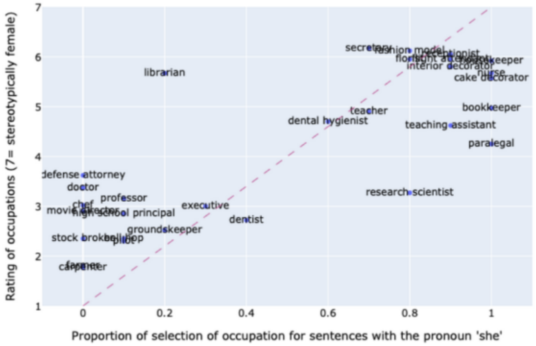
Input data

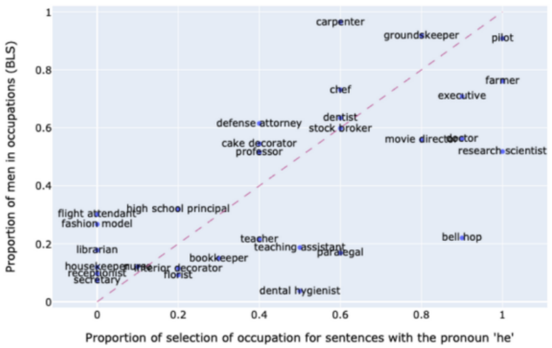
An algorithm that uses historic employment data that comes to the conclusion that men are more suited to managerial position, as historically men were favoured or even the only allowed candidates for such positions.

Societal Bias

Natural Language processing data can reinforce sexist biases due to a bias in training data. This could mean that a model evaluates certain jobs such as secretary or caretaker as intrinsically linked to women.

Such cases were studied and both natural and large language models were found perpetuate stereotypes. Since these models are used more, great care should be taken when working with such cases and active measures taken to prevent the spread of such stereotyping. Such cases prove furthermore that





Current Prototype

Data Hazard Labels

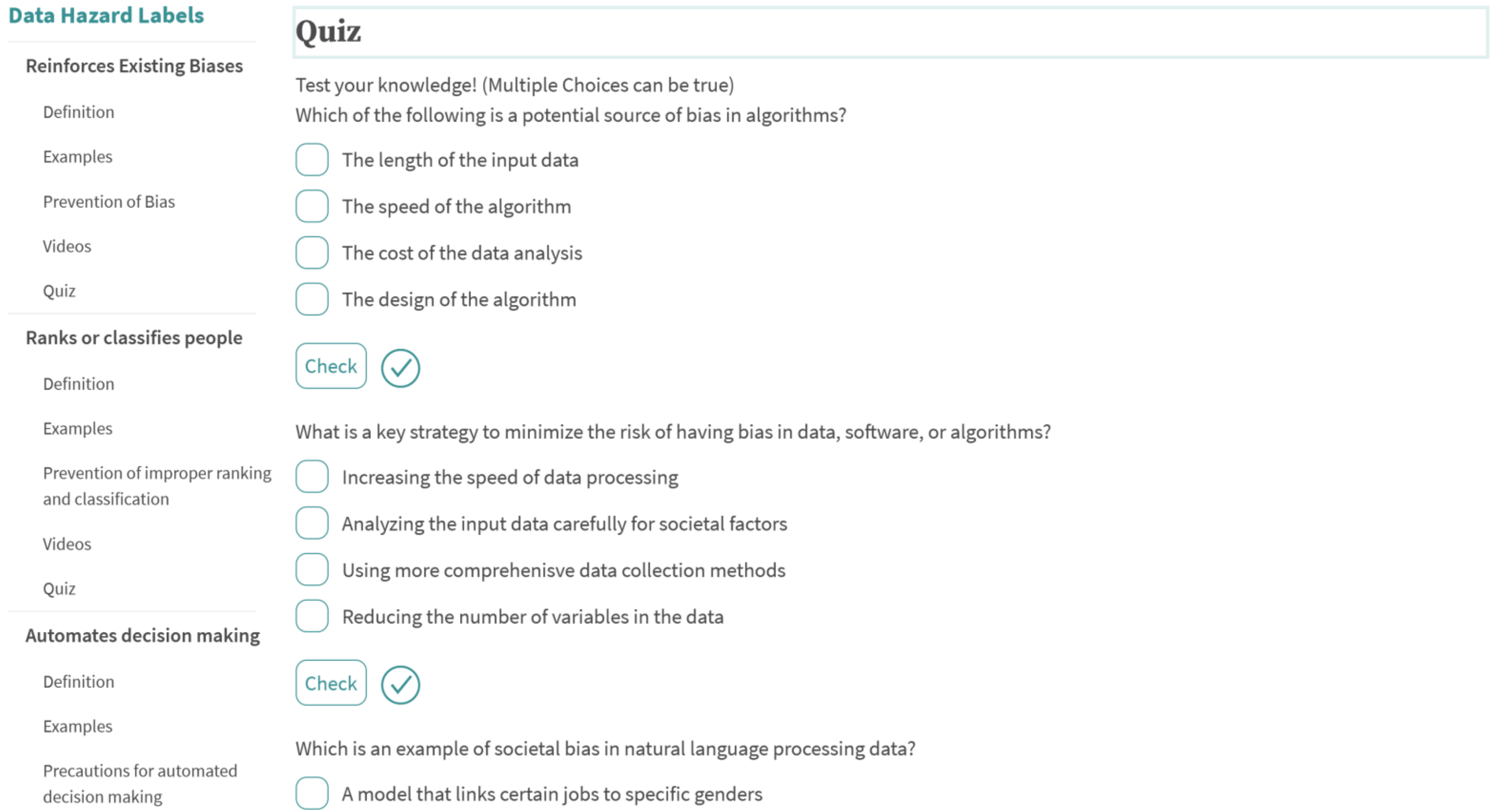
- **Reinforces Existing Biases**
  - Definition
  - Examples
  - Prevention of Bias
  - Videos
  - Quiz
- **Ranks or classifies people**
  - Definition
  - Examples
  - Prevention of improper ranking and classification
  - Videos
  - Quiz
- **Automates decision making**
  - Definition
  - Examples
  - Precautions for automated decision making

Videos

Documentary: Coded Bias (https://www.imdb.com/title/tt11394170/)

Die Schwachstelle von KI – wir Menschen | ZDF Magazin Royale (10.09.2021) 11:22-12:22

Current Prototype



Current Prototype

Search

Videos
Quiz
Danger of misuse
Definition
Examples
Prevention of misuse
Videos
Quiz
May cause direct harm
Definition
Examples
Prevention of harm
Videos
Quiz
Generic Data Hazard
How to apply data hazard labels?
Sources and additional reading

How to apply data hazard labels?

The application of data hazard labels is not meant as an apply and be done with it concept. The labels are meant to open up discussion and help with reflection on responsible ethical use and potential consequences of data science. Ethics in data science is a complex topic so there is no need for a group to come to a definitive conclusion on which labels do or do not apply.

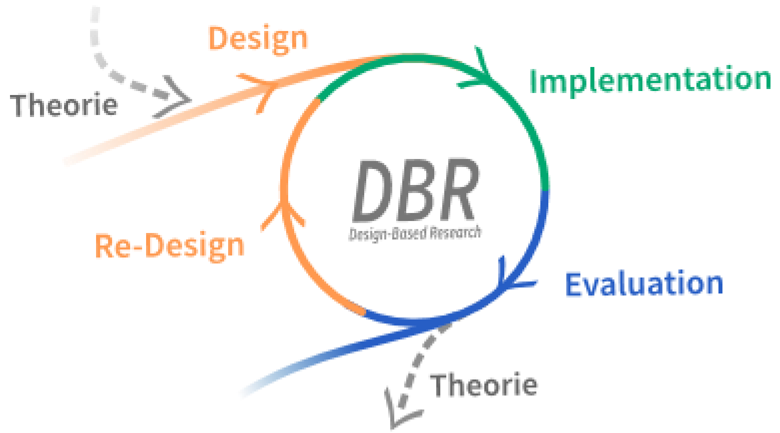
The labels also allow external people to participate in discussions on the ethical impact a technology or algorithm might have which can be tough to get into otherwise. Researchers working in data science might not have the neccesary training or knowledge to be able to consider all these possible cases, which is why the data hazard project was started in the first place.

So if you are working on a project using data science, take a minute to think which labels could apply to your project and encourage others to do the same to get a well rounded perspective on the possible outcomes!

← 33 →

Evaluation and Interviews

- Next step is interviews
- Interviewing experts from relevant fields
- Getting a feedback on design and content
- Use feedback to improve learning course



Questions

Thank you for your attention!

Christian-Albrechts-Universität zu Kiel

Quellen

- The data hazard project (https://datahazards.com/index.html)
- Design based research icon (https://commons.wikimedia.org/wiki/File:DBR_german_colour.svg)