DIGITAL ETHICS CANVAS

CONTEXT

The project focuses on exploring the ability to predict intelligence (ICAR) and personality traits (Big 5 + 1) . The data used comprises anonymized transcripts of 5-minute video interviews and demographic information.

SOLUTION

The solution uses machine learning models, including SVR, Neural Networks, and LightGBM, to process the textual data from transcripts.

BENEFITS

When used correctly, these algorithms can help reduce human biases in decision-making. For example, in recruitment or evaluations, someone might dislike a person and unfairly judge them as not smart or not good enough. The algorithms use objective language patterns to make fairer and more consistent assessments, avoiding personal opinions or prejudices.

WELFARE

RISK

- Can the solution be used in harmful ways, in particular with regards to vulnerable populations?
- What kind of impacts can errors from the solution have?
- What type of protection does the solution have against attacks or misuse?

MITIGATION

The model's bias against non-native English speakers could lead to unfair assessments in contexts such as application screening or cognitive ability evaluation, potentially harming non-native speakers.

Errors in predicting intelligence or personality traits could lead to misjudgments in hiring, mental health diagnostics, or personalization systems, reinforcing inequities

The "human-in-the-loop" approach ensures that decisions are not automated, mitigating risks from misuse.

FAIRNESS

RISK

- How accessible is the solution?
- What kinds of biases may affect the results ?
- Can the outcomes of the solution be different for different users or groups?
- Could the solution contribute to discrimination against people or groups?

MITIGATION

The solution is limited to native English speaker data, reducing accessibility for individuals from non-English-speaking backgrounds.

Algorithmic bias from training data composed only of native English speakers could lead to unequal outcomes for different groups.

Yes, non-native speakers may get less accurate results, creating unfairness. This happens because non-native speakers often use different words or ways of speaking compared to native English speakers.

If used improperly, the solution might reinforce discrimination against individuals whose language patterns deviate from those of the dataset.

AUTONOMY

RISK

- Can users understand how the solution works and what its limits are?
- Are users able to make choices (e.g. consent, settings) in their use of the solution and how?
- How does the solution affect user autonomy and agency?

MITIGATION

In general, users can understand the input and output of the solution, such as the data provided and the predictions made. However, how the system works internally is like a "black box," which is a downside of most of Machine Learning Algorithms.

The solution operates with human oversight, meaning users retain control over its application and can decide how to use the predictions it provides.

PRIVACY

RISK

- What data does the solution collect
- Is it collecting personal or sensitive data
- Who has access to the data?
- How is the data protected?
- Could the solution disclose / be used to disclose private information?

MITIGATION

The solution collects transcripts from 5-minute videos, which include spoken language data, and demographic information about the individuals involved.

Yes, personal data from individuals' speech could reveal sensitive details. However, the data in this project is anonymized, and we do not have access to the audio or video recordings of the individuals, which helps reduce the risk of identifying or exposing personal information.

The data is protected in compliance with the **Swiss Federal Act on Data Protection** (FADP). Additionally, all data is stored securely on local servers, with no use of foreign servers, ensuring that sensitive information is safeguarded and aligned with local data protection regulations.

SUSTAINABILITY

RISK

- What is the carbon footprint of the solution?
- What types of resources does it consume (e.g. water) -and produce (e.g. waste)?
- What type of human labor is involved?

MITIGATION

The project's carbon footprint includes computational resources for training and analyzing AI models.

The solution primarily consumes computational energy, which indirectly relies on electricity generation.

The project is undertaken by students as part of a Machine Learning course, with guidance from academic researchers.

Digital Ethics Canvas 2024 - C. Hardebolle, V. Macko, V. Ramachandran, Y. Do, N. Bartłomiejczyk, A. Holzer, P. Jermann