Surfacing practices and limitations when building fair machine learning systems

Author: Eva Noritsyna
Supervisors: Jie Yang, Ujwal Gadiraju, Agathe Balayn

1. Research Question

To what extent do (envisioned) practices of practitioners without experience with fairness toolkits differ from those with the experience?

2. Background

- Identify and mitigate various risks and harms of using Machine Learning models in industry is an essential task. Specifically because these may produce harmful outcomes for stakeholders, including unfair or discriminatory results
- There has been substantial research into the concepts of fairness and its metrics, bias and its mitigation, and algorithmic harms and their sources.
- E.g: models propagating "structural advantages and disadvantages"[1], and opening up the possibility of "homogenity of decision making"[1]. Both of these concepts could reinforce the unfair treatment of minority groups.
- Toolkits have been created to guide practitioners to reflect on these topics and provide suggestions on algorithmic solutions to mitigate these risks
- It is not yet known how widely used and useful these toolkits are perceived as. The two toolkits that this research project will involve are the IBM AI Fairness360 and Microsoft FairLearn.

3. Method

Participants:

	None	Microsoft Fairlearn	IBM AIF360
Senior	3	3	3
Medior	3	3	3
Junior/MSc	4	4	4

Figure 1: Participant distirbution

Use cases:

1. Participants with toolkit experienceDiabetes Hospital Readmission dataset 6, with the classification task being whether the patient will readmit within 30 days
2. Participants without toolkit experience: Medical Expenditure data with the model classification task to predict whether a person would have 'high' healthcare utilization.

Open coding:
(1) identifying harm source; (2) understanding harm source;
(3) mitigating harm source; (4) identifying impacts of technique;
(5) identifying alternate approaches; (6) business factors;
(7) domain factors; and (8) task factors.

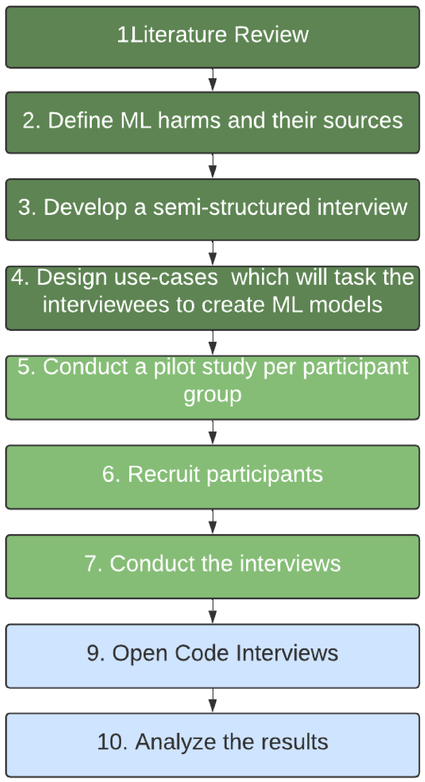

Figure 2: Method diagram

4. Results

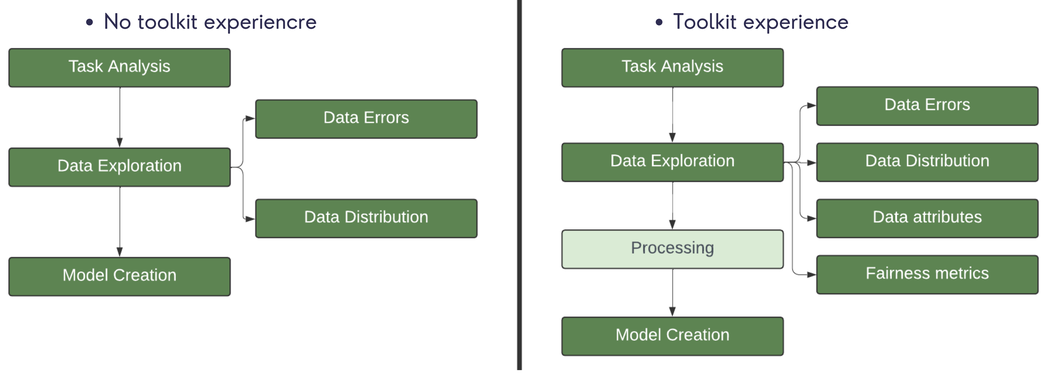

Figure 3: Typical practices workflow comparison

- **Identification of sources of harm that have an influence on the performance of the model**
- **Responses focused on the model creation rather than date exploration**
- **Mitigation techniques limited to outlier and missing values removal or imputation, correlated feature reduction and resampling**

- **Identification of sources of harm that have an influence on the performance of the model as well as fairness, specifically with unfair treatment of inderpriviledged groups**
- **Toolkit use mainly focused on fairness metric, with limited sdiscussion on fairness definition**
- **Identification was prioritised over mitigation techniques**

Business factors

- Open source
- Showcasing to stakeholders
- Integration into the pipeline

- No enforcement
- Integration into the pipeline
- Learning curve

5. Conclusion

- Goa: identify the differences in practices of practitioners with and without experience with fairness toolkits as a way to determine whether such toolkits raise the practitioners' awareness to fairness and educates the practitioner of the importance of considering fairness and bias when building machine learning systems.
- Not possible to asses with certainty whether this difference comes from the experience or from certain confounding factors.
- Suggesting that generally the experience and formal education in ethics and fairness in Machine Learning also may play a big role in the steps taken during the approach in order to identify and mitigate sources of harms while building Machine Learning models.
- In industry, experience of toolkits may be a byproduct o fthe toolkit being needed for business practices
- Some of the differences in practices between practitioners with and without experience with fairness toolkits, may be correlated with factors only relevant in industry and not in academia.

6. Limitation

- Recruitment of people with toolkit experience -> leading towards fairness
- Differences in formal education and training
- Differences in field of work; fairness centric or not
- Practitioners who were employed or associated with the development of the toolkit

1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. "Fairness and Machine Learning Limitations and Opportunities". In: 2018