CLOSER OR EVEN FARTHER FROM FAIRNESS:

AN ASSESSMENT OF WHETHER FAIRNESS TOOLKITS CONSTRAIN PRACTITIONERS WITH REGARDS TO ALGORITHMIC HARMS

AUTHORS

Ana Maria Vasilcoiu (A.M.Vasilcoiu@student.tudelft.nl)
Supervisor: Agathe Balayn
Responsible professors: Jie Yang, Ujwal Gadiraju

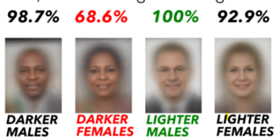
AFFILIATION

EEMCS, Delft University of Technology,
The Netherlands

TU Delft Delft University of Technology

1. INTRODUCTION

- Increased use of automated decision making models in both day-to-day activities and also high-stakes domains
- New challenge = potential for ML systems to treat people unfairly
  - E.g., COMPAS [1], Amazon's hiring tool, facial recognition algorithms



- Numerous different definitions for fairness, impossible to simultaneously satisfy
  => highly-complex and multi-faceted problem that cannot be "solved"
- Fairness toolkits = metrics to measure unfairness in outputs + algorithmic methods to mitigate it when detected [2][3][4]
- Numerous examples (Fairlearn & IBM's AIF360) & new ones constantly being developed

2. BACKGROUND / PROBLEM

- *Limitations of metrics*: insufficient & lacking robustness [4], incompatible in various contexts
- *Limitations of mitigations*: narrow algorithmic perspective, incomplete conceptualizations of discrimination, necessity to choose a metric to debias for [5], disregard of broader justice aspects
- *Gaps between toolkits capabilities and practitioners' needs*:
  - Limited regard to real-life situations [6]
  - Insufficient guidance & educational support [6]
  - Algorithmic harms that go beyond what they currently allow to measure

3. RESEARCH QUESTION

To what extent do toolkits constrain the frame of practitioners with regards to algorithmic harms?

Objectives:

- Identify general limitations of toolkits
- Discover practices for assessment & mitigation of harms (comparison with and without a toolkit)
- Check potential for missed or unattended harms when using a toolkit

4. METHODOLOGY

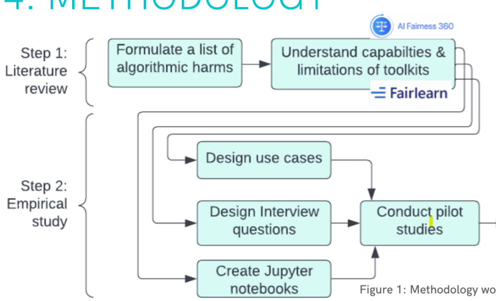


Figure 1: Methodology workflow

Participants (total = 30):

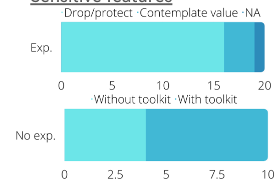
Experience with toolkit	ML expertise	Number
No prior experience	Master Student	4
	Junior Data Scientist	2
	Senior Data Scientist	4
Prior familiarity with Fairlearn	Researcher	3
	Data scientist	7
Prior familiarity with AIF360	Researcher	4
	Data scientist	6

Protocol: think aloud while exploring problematic datasets

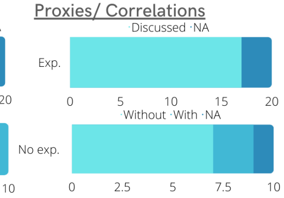
5. RESULTS

Fairness related harms

Sensitive features



Proxies/ Correlations

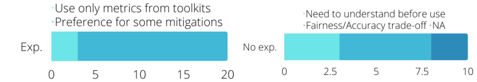


Choices of (fairness) metrics

- compute multiple metrics & choose based on context
- parity only - awareness from all participants
- too large dependency on metrics & their limitations barely addressed (no alternatives given)

Bias mitigation

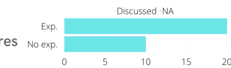
- data-preprocessing preferred by both groups



Other types of harms

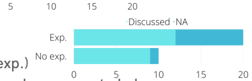
Data representation

- even representation of sensitive features VS. representative of reality

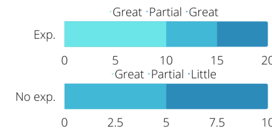


Data over/under sampling

- concerns: loss of data (no exp.) & loss of nuances (exp.)
- alternatives: adjust thresholds & weight errors in underrepresented class more (exp.) & choice of model that works with imbalanced data (no exp.)



Irrelevant attributes



Oversimplified attributes

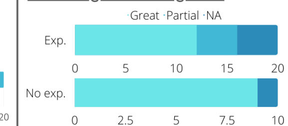
- only 2/20 people with experience unassistedly found issues with binary race/gender
- proposed solutions: collection of more representative data & informing stakeholders that model works only for represented groups

Broader harms:

(same for both types of participants & for both with and without toolkit)

- Task - involve domain specialists + minimum consideration of whether task makes sense & whether it should be automated
- Environmental impact & Harms in a broader environment - great understanding when prompted, but little to no actionability

Handling of missing data



- no exp: replacement > removal
- exp: adapt on how much is missing, concerns around introducing artificial qualities by imputation

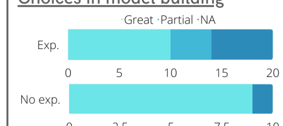
Handling of outliers

- mentioned indirectly, manual check & understand reasons

Handling of duplicates

- remove only actual duplicates after thorough check

Choices in model building



- no exp: hyperparameter optimization, thresholds
- exp: feature importance, choice of training features
- all: explainability > accuracy, model should depend on data size & distribution

Incorrect labels

- when asked, all identified annotator's lack of context knowledge and personal bias as potential problems
- invisible worker discussed only by 2 people with experience

6. DISCUSSION

Harms for which toolkits resulted in greater understanding:

- irrelevant attributes
- sources of labels
- undersampling techniques

Harms for which future intervention is required:

- sensitive features & protected attributes
- bias mitigation algorithms
- handling of missing data
- choices in model building

Recommendations for future toolkits:

- more actionable guidance on assessment & mitigation data characteristics harms (e.g., instruction materials for DataSheets)
- promote interdisciplinary collaborations between stakeholders from different backgrounds

7. LIMITATIONS

- Limited number of toolkits analyzed
- Sample size & participation bias
- Limited number and narrow scope of models & tasks

8. CONCLUSIONS

- Toolkits can lead to a disregard of insufficiently covered harms, but can engender proactiveness towards multiple other issues
- Confirmed importance of thorough design and evaluation of toolkits & need to educate practitioners prior to using them

RELATED LITERATURE

[1] Surya Mattu Julia Angwin, Jeff Larson. Machine bias: There's software used across the country to predict future criminals. and itâs biased against blacks. 2019
[2] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galtyan. A survey on bias and fairness in machine learning. ACM Comput. Surv., 54(6), jul 2021.
[3] Sahil Verma and Julia Rubin. Fairness definitions explained. In 2018 ieee/acm international workshop on software fairness (fairware), pages 1-7. IEEE, 2018.
[4] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In Proceedings of the conference on fairness, accountability, and transparency, pages 329-338, 2019
[5] Agathe Balayn and Seda Gürses. Beyond debiasing: Regulating ai and its inequalities. EDRi Report. https://edri. org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online. pdf, 2021.
[6] Michelle Seng Ah Lee and Jat Singh. The landscape and gaps in open source fairness toolkits. In Proceedings of the 2021 CHI conference on human factors in computingsystems, pages 1-13, 2021.