



The Landscape and Gaps in Open Source Fairness Toolkits

Authors

Michelle Seng Ah Lee

Jatinder Singh

Year – 2021

- Sai Sri Lekha Thirunahari

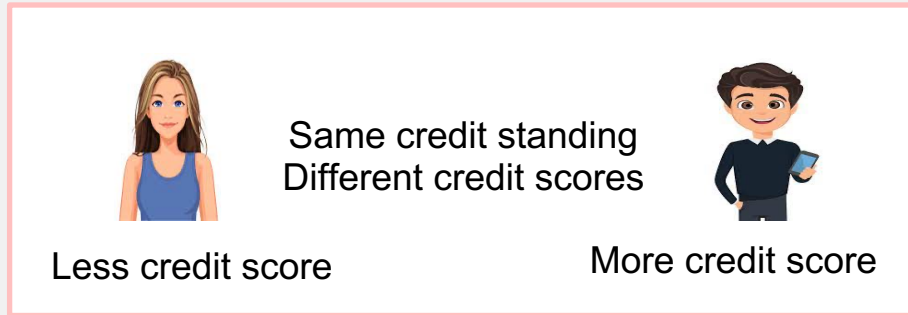


ABSTRACT

- A number of open source "fairness toolkits" have recently appeared in response to the explosion in research focusing on the evaluation and mitigation of unfair results in algorithms, making these techniques more widely available.
- This paper identifies the gaps between the existing open-source fairness toolkit capabilities and the industry practitioners' needs
- Furthermore, they point out number of flaws in the state of the art, such as compatibility problems, the demand for more complex algorithms, poorer tool integration, and inadequate documentation and support.

INTRODUCTION

- In recent years, there has been a considerable surge in the creation and use of AI and machine learning systems.
- However, it has also been discovered that these systems amplify and enhance preexisting prejudices, producing biased results.



- The authors offer a thorough analysis of the available open-source fairness toolkits, outlining their benefits, drawbacks, and contributions to the field.

INTRODUCTION

- The authors also point out a number of shortcomings in the existing landscape of open-source fairness toolkits and a number of gaps between the capabilities of the tools and the requirements of practitioners in terms of:
 - 1) functionality
 - 2) user-friendliness
 - 3) contextualisation

POPULATION TARGETED

- The data scientists, programmers, and researchers who are interested in include fairness in artificial intelligence and machine learning systems are the population targeted.



- Also, the people and organizations who are interested in incorporating fairness into their AI systems and are searching for a comprehensive examination of the state of open-source fairness toolkits as well as the shortcomings and difficulties in their creation and application.



RESEARCH QUESTIONS

1. What open-source fairness toolkits are currently available to users, and how do they differ in terms of documentation, community support, and ease of use?
2. What gaps do open-source fairness toolkits currently have, and what difficulties do practitioners have incorporating these toolkits into their machine learning workflows?
3. What suggestions are there for the future development of open-source fairness toolkits, and how can researchers and practitioners work together to make these toolkits more approachable and easier to use?

TARGET TECHNOLOGY

- The authors give a thorough analysis of the current state of open-source fairness toolkits and point out the shortcomings and difficulties in their creation and use rather than concentrating on the development of any particular technology or application.
- The authors review various open-source fairness toolkits and suggest a set of evaluation standards to spot shortcomings in the present state of the field and direct future advancement.

METHODOLOGY

The methodology consists of the following four steps:

- **exploratory focus group** - to discover well-known fairness toolkits and gain preliminary insights
- **Comparative review of the selected toolkits**- to compare the characteristics offered by each toolkit.
- **Semi-structured interviews**- with practitioners with prior experience in fairness challenges were conducted in order to better understand the characteristics of the ideal toolkit and assess how well each of the six toolkits met their needs
- **Survey**- was also conducted in order to confirm the results with a larger audience and dive deep into a few previous stages' insights.

METHODOLOGY

- Comparative review of the selected toolkits- to compare the characteristics offered by each toolkit

Tool	Setup	Open source user license	Release date	Organization	Open for anyone to contribute code?	Models covered				Group fairness				Individual		Other fairness metrics	Bias mitigation	
						Regression Classification (binary/outcome)	Multi-class outcome	Handles multi-class protected feature?	Demographic parity (statistical parity)	Equal opportunity / True positive parity / False positive error rate balance	Equal odds (True positive and false positive parity)	Disparate impact	Discovery rate	Omission rate	Counterfactual fairness			Sample distortion metrics
Scikit-fairness / scikit-lego	python (sklearn)	MIT	2019-03-31	N/A	✓	✓	✓	X	X	✓	✓	X	X	X	X	X	N/A	Pre-processing: information filter
IBM Fairness 360	python 3.5+, R	Apache 2.0	2018-06-01	IBM	✓	X	✓	✓	✓	✓	✓	✓	✓	✓	X	✓	Generalized Entropy Index Differential Fairness and Bias Amplification (full list here: https://aif360.readthedocs.io/en/latest/modules/generated/aif360.metrics.ClassificationMetric.html)	Optimized Preprocessing, Disparate Impact Remover, Equalized Odds Post-processing, Reweighting, Reject Option Classification, Prejudice Remover Regularizer, Calibrated Equalized Odds Postprocessing, Learning Fair Representations, Adversarial Debiasing, Meta-Algorithm for Fair Classification, Rich Subgroup Fairness
Aequitas tool	python 3.6+	Custom	2018-02-13	UChicago	✓	X	✓	X	✓	✓	✓	✓	X	X	✓	X	N/A	N/A
Google What-if tool	Tensorboard / Jupyter or Colab notebook	Apache 2.0	2018-09-11	Google	✓	✓	✓	✓	✓	✓	✓	X	X	X	X	✓	X	Group thresholds Threshold optimization based on fairness constraints
PyMetrics audit-ai	python	MIT	2018-05-18	PyMetrics	X	✓	✓	X	X	X	X	X	✓	X	X	X	Statistical tests to determine chance the disparity is due to random chance (ANOVA, 4/5th, fisher, z-test, bayes factor, chi squared sim beta ratio, classifier posterior_probabilities)	N/A
Fairlearn	python	MIT	2018-05-15	Microsoft	✓	✓	✓	X	✓	✓	✓	X	X	X	X	X	Group max / min / summary	Exponentiated Gradient, GridSearch, Threshold Optimizer

GAPS IDENTIFIED

The interviews and surveys revealed some key gaps, which are detailed below in three parts on toolkit features, contextualization, and user-friendliness.

Gaps: User-friendliness

- Steep learning curve required to use the toolkits and limited guidance on metric selection

Toolkits	Average SUS	StdDev SUS
Aequitas Tool	61.33	15.78
Fairlearn	65.71	12.99
Google What-if tool	60.33	17.14
IBM Fairness 360	54.50	13.89
PyMetrics Audit AI	58.04	10.29
Scikit-fairness	62.83	17.32
<i>All</i>	<i>60.43</i>	<i>14.84</i>

- Information overload vs. over-simplification of complex results
- Need for “translation” for a non-technical audience
- Accessibility of toolkit search process

GAPS IDENTIFIED

Gaps: Toolkit features

- Limited coverage of the model pipeline.
- Limited information on possible mitigation strategies.

Gaps: Contextualisation

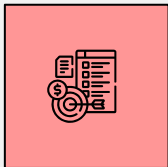
- Limited adaptability of existing toolkits to a customised use case.
- Challenges in integrating the toolkit into an existing model pipeline.



EVALUATION OF STRENGTHS

- The authors' extensive assessment of the literature to determine the most widely applied open-source fairness toolkits and their usage of a set of uniform criteria to assess and compare the toolkits are the methodologies' strongest points.
- The toolkits are beneficial for a variety of fairness-related activities because the paper outlines a wide range of approaches employed in them, such as bias mitigation, bias detection, and fairness measurements.

WEAKNESSES



limited Scope

only regarding open-source fairness toolkits



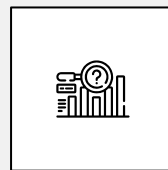
Evaluation Criteria

Some of these standards might not apply to all toolkits.



Selection Bias

Depending on the popularity of the toolkits



Lack of Empirical Evaluation

includes no empirical analysis of the toolkits



MAJOR FINDINGS

1

Wide range of toolkits

2

Lack of consistency

3

Potential improvement

4

ML frameworks integration

DISCUSSION POINTS

Importance of fairness in machine learning:

Tools are required to make sure that algorithms don't reinforce prejudice or discrimination.

Open-source toolkits and community contributions:

Fairness in machine learning can be used more widely through democratizing its use and increasing its accessibility with the use of open-source toolkits

Need for empirical evaluation:

Emphasizes the demand for additional empirical research that assess the performance of fairness toolkits in practical contexts.

Ethical implications:

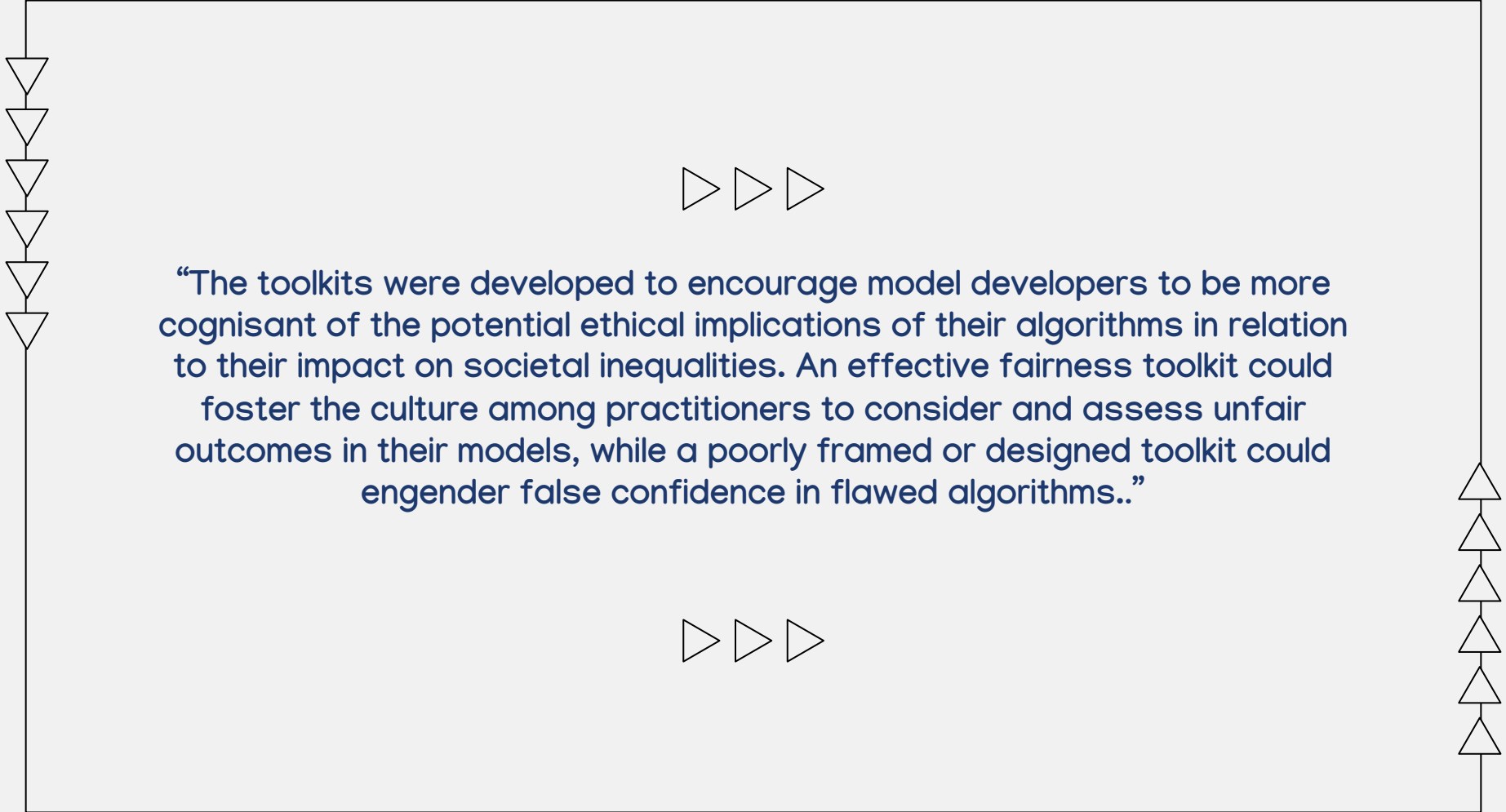
Fairness toolkits can assist in alleviating these ethical concerns and enabling more moral and socially responsible machine learning.





PERSONAL THOUGHTS AND CONCLUSION

- An alternative to the techniques used in the paper could be to evaluate the toolkits in greater detail, including benchmarking their performance on other datasets.
- Surveying industry professionals and decision-makers to learn about their objectives and priorities with relation to fairness in machine learning is another possible possibility.

A decorative border surrounds the text. On the left side, there is a vertical line with six downward-pointing triangles. On the right side, there is a vertical line with six upward-pointing triangles. In the center of the page, there are two sets of three right-pointing triangles, one above and one below the main text block.

“The toolkits were developed to encourage model developers to be more cognisant of the potential ethical implications of their algorithms in relation to their impact on societal inequalities. An effective fairness toolkit could foster the culture among practitioners to consider and assess unfair outcomes in their models, while a poorly framed or designed toolkit could engender false confidence in flawed algorithms..”



**THANK
YOU**

