Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits

Brianna Richardson[*]
richardsonb@ufl.edu
Spotify

Jean Garcia-Gathright
jean@spotify.com
Spotify

Samuel F. Way
samfway@spotify.com
Spotify

Jennifer Thom
jennthom@spotify.com
Spotify

Henriette Cramer
henriette@spotify.com
Spotify

ABSTRACT

In order to support fairness-forward thinking by machine learning (ML) practitioners, fairness researchers have created toolkits that aim to transform state-of-the-art research contributions into easily-accessible APIs. Despite these efforts, recent research indicates a disconnect between the needs of practitioners and the tools offered by fairness research. By engaging 20 ML practitioners in a simulated scenario in which they utilize fairness toolkits to make critical decisions, this work aims to utilize practitioner feedback to inform recommendations for the design and creation of fair ML toolkits. Through the use of survey and interview data, our results indicate that though fair ML toolkits are incredibly impactful on users' decision-making, there is much to be desired in the design and demonstration of fairness results. To support the future development and evaluation of toolkits, this work offers a rubric that can be used to identify critical components of Fair ML toolkits.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; • **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

AI; ML; ethics; fairness; algorithmic bias; machine learning fairness; user-centric evaluation

ACM Reference Format:
Brianna Richardson, Jean Garcia-Gathright, Samuel F. Way, Jennifer Thom, and Henriette Cramer. 2021. Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3411764.3445604

1 INTRODUCTION

As new applications of machine learning (ML) emerge across industries, stakeholders and researchers alike aim to reduce the negative influence of unanticipated biases in these algorithms. News coverage has focused on several algorithms for the role they play in furthering social inequities. For instance, the ML recognition systems in autonomous vehicles have been less effective in recognizing darker skin [53]; search engine results have reinforced representation bias [42]; and even professional networking sites were discovered to recommend "male-sounding" variations of names when "female-sounding" names were searched [17]. Institutions have been quick in formulating their own guidelines for confronting algorithmic bias [28]. These guidelines focus on themes like fairness, transparency, accountability, and so on [28]. Within institutions, teams are also focusing on how to transform literature into actionable steps that can be taken by product teams [14, 22, 27, 34, 35].

Adding to these studies, the last half decade has seen a proliferation of algorithmic bias research [15]. New research communities, such as the Fairness, Accountability, and Transparency (FAccT) conference have emerged with goals of confronting algorithmic bias across industries [3]. In order to support the inclusion of fairness-forward thinking in the machine learning pipeline, applications are being created that allow non-research-based practitioners to employ state-of-the-art fairness considerations in their own work. A few examples of such toolkits include IBM's AI Fairness 360 [9], Google's Fairness Indicators [2], Microsoft's Fairlearn [39], and UChicago's Aequitas [47].

Despite these advancements in tooling, however, recent work has revealed a disconnect between the tools created by fairness experts and the needs of practitioners [27, 34, 35, 37, 50]. While several tools are available to practitioners, a lack of communication has prevented their application in high-stakes product teams. In order to encourage adoption, participation and buy-in with respect to fair ML tool development and usability must be considered.

Through analysis of prior work and practical testing of tools, we present a set of standards for evaluation of such tools in the future. By giving 20 practitioners the first-time experience of engaging with fair ML tools, this study measured the impact of these tools, collected practitioner perceptions towards them, and utilized practitioner feedback to inform recommendations for the design and creation of fair ML toolkits. Participants saw one of two fair ML toolkits, Google's Fairness Indicators or UChicago's Aequitas, and researchers were able to capture their initial reactions and thoughts. The practical needs suggested by these practitioners, in addition

[*] The first author completed this work during an internship at Spotify and is now at the University of Florida.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445604

to previous literature on tooling, were used to inform a rubric that can be used as both a guide and evaluation tool for practitioners, fairness design experts, and organizations to use in the creation and selection of fair ML in industry settings.

The contributions of this study are three-fold:

- Conduct a comprehensive literature review on user needs in FAccT tooling
- Collect user perceptions towards these tools and understand the role and the capabilities of fairness in user ML responsibilities
- Measure the level of impact of fair ML tools
- Propose a rubric that can be used as an evaluation tool and a guide for the integration of fairness tools into the workplace

2 BACKGROUND & RELATED WORK

2.1 Fair ML

Fairness in ML is defined by both mathematical and social notions of fairness and considers the distributions of both the harms and the opportunities provided by ML [15]. Fairness is defined in a multitude of ways by researchers [8, 38, 52], making the decision of which fairness methodology to adopt and which corresponding metrics to quantify highly contextual.

In order to assist practitioners in the fairness consideration process, several institutions have created toolkits that are applicable to a diverse range of models and datasets [2, 9, 20, 39, 47, 49]. Each tool differs in the means and the methods provided for enforcing fairness, the visual demonstrations of statistics, and the support provided to new fairness users, among other differences. Some tools allow for easy intersectional analysis, while others require prior data manipulation to see subgroup performance. Some tools provide interactive or colorful depictions of fairness results, while others leave the task of creating visualizations up to the user. Some tools are designed to integrate with pre-existing machine learning tools, while others are stand-alone, and the extent of any provided background information, demos, and tutorials differs substantially between toolkits.

While each tool has explicit advantages and disadvantages, to our knowledge, this is the first study engaging practitioners with fairness toolkits to evaluate the effectiveness of the tools, providing a greater understanding of the barriers to implement fair ML in practice.

2.2 The Needs of Practitioners

Several previous works have outlined the needs of ML practitioners as they relate to FAccT technologies [27, 34, 35, 37, 50]. Veale et al. [51] conducted exploratory interviews with ML practitioners across several countries and industries who were applying ML to a diverse range of practices. This work demonstrated the critical need for conversations that elicit the fairness needs of experts. Furthermore, it highlighted a disconnect between institutions and practitioners when it came to fairness [51]. Holstein et al. [27] defined critical design needs of practitioners via interviews and online surveys. Through engaging with practitioners, these researchers defined critical gaps between conversations had in industry and through fairness research literature. Madaio worked alongside practitioners to co-design a checklist of needs for the construction of organizational infrastructures of fairness [37]. In interviews with practitioners, Law et. al. [34] showed participants two bias detection prototypes with differing design strategies to understand how they impact users' insights. Their results suggest that information load and comprehensive axes are critical considerations for optimal toolkit design [34].

Several works have focused on identifying the needs of the practitioners by testing the effectiveness of FAccT technologies. Outside of fairness, but in the realm of interpretability, Kaur et al. studied the effectiveness of interpretability tools [31]. Through the use of contextual inquiry and online surveys, these researchers found that many practitioners struggled with interpretability tools and often over-trusted or misused these tools [31]. Lakkaraju & Bastani [33] tested the limits of certain explainable AI features by testing its effectiveness against black box and adversarial attacks. They found that manipulating favorable features in explanations could impact user trust by nearly 10-fold. This emphasized a need for careful consideration of how interpretations are presented to users [33]. Ribeiro et al. [46] tested the effectiveness of their XAI, LIME, by measuring the level of insight users were able to collect from engaging with the tool.

While previous studies engaged practitioners through need-finding interviews, this study allowed practitioners to actively engage with mature, publicly-released fairness toolkits.

3 METHODS

To collect practitioners' feedback on Fair ML toolkits, we conducted semi-structured, one-on-one interviews with participants who had experience assessing machine learning models. Users interacted with a Qualtrics [44] survey, a Google Colaboratory notebook [23], and Google Meet conferencing during the extent of the study.

3.1 Participants

We recruited 20 participants across four research and industry institutions in July and August of 2020. All participants were individuals who had responsibilities assessing machine learning models. Participants were involved with a diverse array of technologies, including: Audio & Voice, Recommendation Systems, Bio-engineering, ML infrastructure, Ethical AI, and Natural Language Processing. 11 participants considered their work area Engineering/Product-oriented, 8 participants considered their work area Exploratory/Research-oriented, and 1 participant considered their work both research and product-oriented. Additional information about the participants can be referenced in Table 1.

Six pilot interviews were conducted prior to ensure that the level of difficulty of the tasks was appropriate considering the diverse roles being recruited for this study. Pilot interviews were also used to confirm that users would have enough time (60 minutes) to perform all tasks in the study.

Participants were recruited through internal communication channels and word-of-mouth. Invitations to participate included eligibility requirements and an invitation to learn about new and emerging fair ML toolkits. Participants for this study were, with relation to fair ML tools, mostly non-experts with high levels of interest. This collection of non-experts allowed us to gauge the the challenge presented by the novelty of the tool for first-time users.

Work Area	Role	Participant ID
Engineering/Product	ML Engagement Lead, Data Scientist, ML Engineer (4), Engineer (3), Manager (2), Data Engineer	P1, P2, [P6, P7, P11, P20], [P8, P9, P12], [P15, P19], P14
Exploratory/Research	Research Scientist (6), Data Scientist (2), Data Engineer	[P3, P4, P10, P13, P16, P18], [P15, P17], P14

Table 1: Participant Demographics. Participants self-reported work area and role along with the Participant ID that will be used for the extent of the paper.

These participants represent a pivotal, yet highly understudied, group: those without experience with ML-fairness, but who are interested in using fair toolkits to address potential negative impacts in their own ML-work. These participants are a key demographic in fair ML UX research: the practitioners who will be the first willing to engage with and advocate for fair ML tooling in their individual domains.

3.2 Fairness Toolkits

A between-subjects approach was used for this study where each participant saw one of two toolkits: either Google's Fairness Indicators toolkit [2] or UChicago's Aequitas toolkit [47] in a Google Colaboratory notebook. To select these tools used in this study, pilot tests were used to test several available tools with visual components. Of the tools that worked in the Google Colaboratory environment, we found that these two selected tools were the most representative, self-explanatory, and required the least guidance from researchers. Participants were shown the name of the toolkit, but not the institution where the toolkit came from. While no participant had experience with the toolkit that they worked with, one participant, however, did verbally notice that the Fairness Indicators toolkit was part of the TensorFlow pipeline [1] from appearance alone.

Fairness Indicators and Aequitas were the chosen toolkits for this study also for their ability to function in the study environment (Google Colaboratory) and their choice of visual aids which are influential in the fairness consideration process [9]. Fairness Indicators [2] functions as an interactive widget, where users can provide their model(s), their performance and fairness metrics of choice, and the attributes on which slicing and evaluating will take place. Users also have the choice of doing model comparisons, intersectional analysis, and performance testing across thresholds. Fairness Indicators allow users to compare at most two models and allows users complete control of the attribute slices they choose to focus on. As an interactive widget, users have the opportunity to manipulate the metrics that they see, the thresholds that they can compare, and the slices that they'd like to focus on.

Aequitas [47] also serves as a visual bias detecting toolkit. Aequitas can be accessed via three different platforms: the command line, a python package, or through their web interface. Through this toolkit, users have the option to evaluate and compare models, calculate group biases, and highlight disparities. Aequitas' python toolkit has two types of visualizations: grouped bar charts and treemaps. Users also have the option of color coding these visualizations green or red based on fairness disparity thresholds that the user can assign. Similar to Fairness Indicators, comparisons can be made across models. While these two toolkits are diverse in the types of visualizations offered and the features available, they are also very representative of the options in the fair ML space.

A randomizer was used to make sure an even distribution of participants saw each of the toolkits.

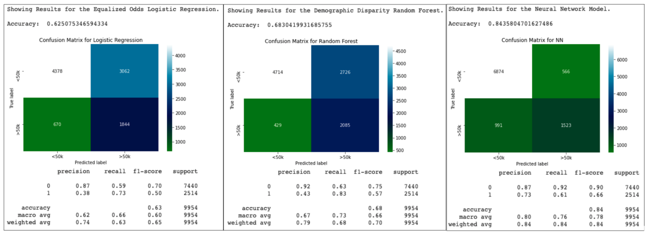
3.3 Data & Models Viewed by the Participants

Fairness tools were used to visualize and compare fairness results from three different models built from the same data. The dataset used for this study was UCI's 1994 Adult Census data [32]. This publicly-available dataset is commonly used in algorithmic bias research [4, 9, 20, 31, 36] due to its biased sampling of classes. Once the data was retrieved from the UCI page [19], gender, race, and the predictive class were separated for subsequent analysis and a few irrelevant attributes were removed. The attributes left for model generation were age, work class, education level, marital status, occupation, investment capital loss and gain, hours worked per week, and native country. These attributes were used to predict whether or not an individual makes over $50,000 a year. This dataset was chosen due to its inherent biases that - if left ignored - could be reflected in models trained with it.

Three separate models were built: a Logistic Regression (LR) Model, a Random Forest (RF) Model, and a Neural Network (NN) Model. These models were selected to provide a diverse subset of supervised learning methods and counteract any preference bias from participants. Two-thirds of the data was used for training the models while the last third was used for model evaluation. To generate fairness differences between models for participants to discover, we post-processed the models in the following ways: for Linear Regression and Random Forest, post-processing fairness modifications and balancing of training data was done on the predictors. Microsoft's FairLearn ThresholdOptimizer [39] was used to adjust these predictors to satisfy certain parity constraints. For Logistic Regression, the model was optimized to satisfy Equalized Odds, which requires that a fair classifier predict positive and negative class across groups with the same likelihood [25]. For Random Forest, the model was optimized to satisfy Demographic Parity, which states the proportion of people in each demographic group classified in the positive class should be equivalent [30]. The Neural Network did not receive any post-processing modifications. Each of the models were evaluated using the evaluation test set.

3.4 Interview Protocol

Each study took approximately one hour. Studies were conducted virtually via video conferencing. Meetings were video recorded with the participant's permission and later transcribed. Participants began with a brief Qualtrics pre-survey that allowed them to share a few details about their roles and the type of work they focus on. From there, participants began the simulated experience. They were told they would serve as the decision-maker for their team and told that a client was requesting a predictor that could determine whether or not an individual made over $50K. They were introduced to the data and shown performance output for the evaluation test set. A depiction of the results shown to the participants can be seen in Figure 1. Using the information given thus far, participants were asked their willingness to deploy each of the models to the client.



```
Showing Results for the Equalized Odds Logistic Regression.

Accuracy:  0.625075346594334

              precision    recall  f1-score   support

           0       0.87      0.59      0.70      7440
           1       0.38      0.73      0.50      2514

    accuracy                           0.63      9954
   macro avg       0.62      0.66      0.60      9954
weighted avg       0.74      0.63      0.65      9954
```

```
Showing Results for the Demographic Disparity Random Forest.

Accuracy:  0.6830419931685755

              precision    recall  f1-score   support

           0       0.92      0.63      0.75      7440
           1       0.43      0.83      0.57      2514

    accuracy                           0.68      9954
   macro avg       0.67      0.73      0.66      9954
weighted avg       0.79      0.68      0.70      9954
```

```
Showing Results for the Neural Network Model.

Accuracy:  0.8435804701627486

              precision    recall  f1-score   support

           0       0.87      0.92      0.90      7440
           1       0.73      0.61      0.66      2514

    accuracy                           0.84      9954
   macro avg       0.80      0.76      0.78      9954
weighted avg       0.84      0.84      0.84      9954
```

Figure 1: Participants were initially shown this performance output between the three models, as well as, a ROC curve that can be seen in the Appendix.

Next, the participants followed a link from the survey into one of the two Colaboratory notebooks, one for each Fairness toolkit. They were told that this notebook consists of analysis done by one of their team members and they should use these results to make a final decision on whether or not they would like to deploy these models. Each notebook is broken down into two parts. The first part, the Data Exploration Phase, allows a more in-depth picture of the data and the models, while the second part of the notebook, the Fairness Exploration phase, utilizes its respective fairness toolkit to complete fairness analysis. For each part of the notebook, users were given three questions that they must answer in the survey. These questions served multiple purposes: to compare insight and confidence when users query common data science tools like confusion matrices versus fairness toolkits and to give participants an opportunity to dig deep into the data, models, and visualizations. For each question, users provided their answer, their confidence in their answer, and can detail what type of support would strengthen their confidence. Once the six questions were complete, participants were then asked to verbalize the strengths and weaknesses they noticed for each model. From there, they were asked again their willingness to deploy each of the models. Then they completed some post-survey questions about their prior level of exposure, interests, and opinions on their respective fairness toolkit. Additional time left at the end of any session was used by the researcher to query participants about their experience, their opinions on the toolkit, and decisions they made during the study. Figure 2 provides an image depiction of the interview protocol procedures.

4 RESULTS & EVALUATION RUBRIC

In combination with the plethora of literature research studying the key features of fairness and the design needs of practitioners, the aim of the study was to support the creation of a rubric that could be used by fairness experts, practitioners, and organizational institutions alike to design, build, and evaluate the effectiveness of fair ML toolkits. Using the results from the user study and literature analysis, we created a list of needs for future fair ML assessment tools. Here we present our results along with the recommended criteria in the evaluation rubric. The rubric is composed of two parts: the first part of the rubric is about enabling analysis on ML and the second part emphasizes criteria that assists with improving the tooling. The criteria and short descriptions can be seen in Table 2.

4.1 Criteria for Supporting Fairness Analysis

Criteria on fairness receives much of its support from the abundance of algorithmic bias-based literature from the last decade. Much research has been done on how bias can influence models, which formulas can detect bias, and what methods can be used to mitigate bias [7, 11, 21, 29, 45, 50, 54]. It is critical that first and foremost, fairness toolkits are equipped with the basic functionalities emphasized by researchers when it comes to fairness. Fairness experts should work to include as many diverse options for the following criteria, while organizational units and practitioners should make sure that the options provided satisfy their needs.

4.1.1 Applicable to a diverse range of predictive tasks, data types, and models. The toolkits used in this study focus on binary/multi-class problems. Many prominent toolkits are similar in this respect. However, this is not representative of the models being used in practice. Practitioners that focus on ranking, recommendation tasks, or speech synthesis found that these tools would not be easily applicable, if at all. Furthermore, P18 mentioned that group fairness was not as important as individual fairness for their work. While there exists fairness frameworks for non-traditional, non-classification problems [11, 45, 54], the future of fairness most definitely requires toolkits versatile to a diverse range of machine learning tasks. The types and applications of machine learning are extensive and always growing [29]. While supervised learning is the most exemplified in fairness literature, it is not at all generalizable to the types of models that exists that require fairness consideration. Whether it be through expanding toolkit functionality or exemplifying this extensive functionality through demos, fairness experts must demonstrate to practitioners that their toolkits are diverse in the types of models, data and predictive tasks they are given. Fair-aware methodologies for ranking, clustering, and embedding [11, 45, 54] should also be included in these toolkits. Furthermore, there should be fairness analysis options for individuals without access to sensitive features, which was a sentiment from the practitioners in this study, as well as [27]. As new methodologies for bias detection and mitigation are proposed, fair ML toolkits should work to include a diverse set of research that covers the most diverse set of existing technologies.

4.1.2 Detecting & Mitigating bias. Fair ML tools focus on both the detection and mitigation of bias. While these two steps are separate, it is important that users are supported for both options [27]. Some participants within this study expressed that the process of detecting bias was complicated and that detection introduces more questions than it provides answers. This can be frustrating for practitioners as it forces them to search for their own solutions blindly. An optimal tool would, at the very least, provide recommendations for how users can mitigate the biases in their models.

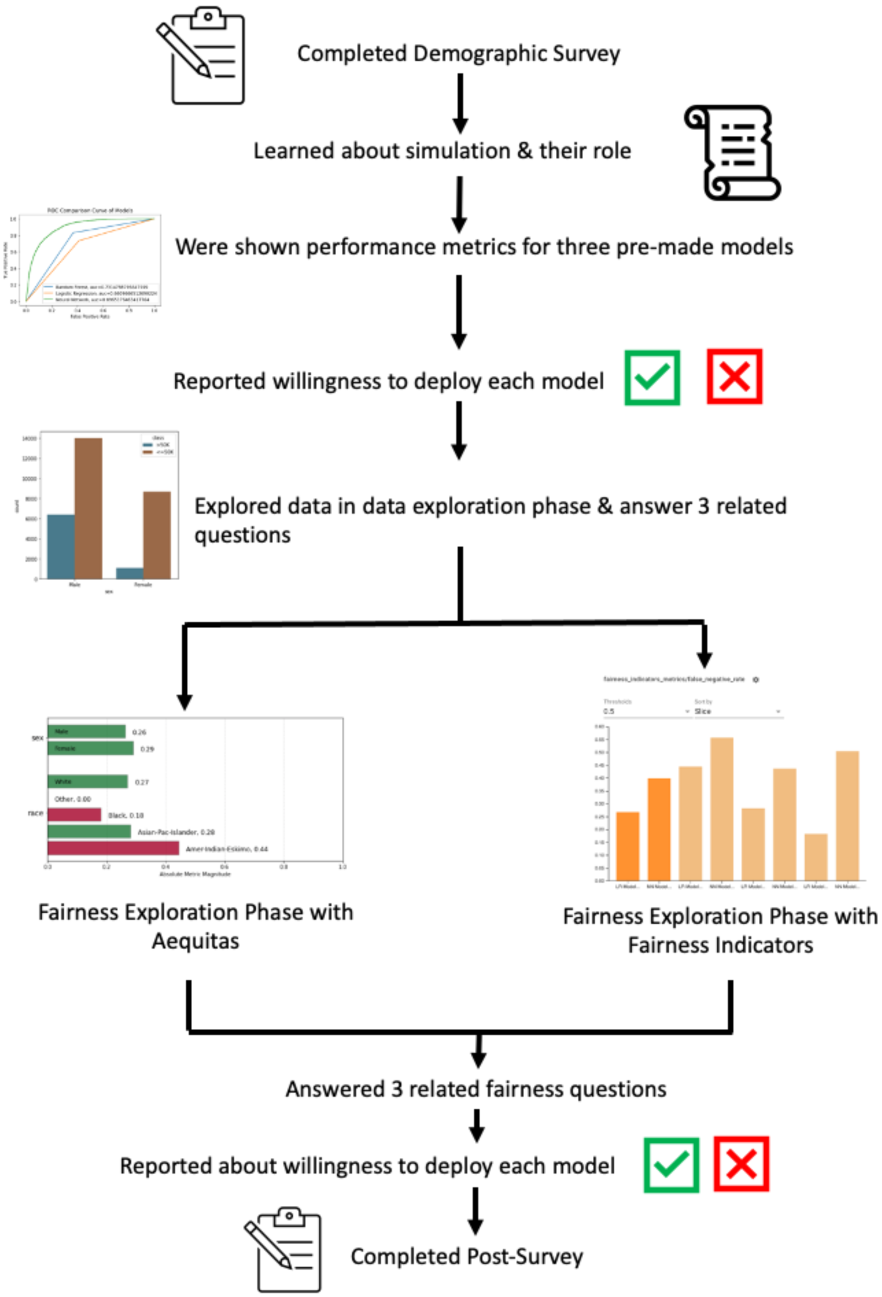

Completed Demographic Survey
Learned about simulation & their role
Were shown performance metrics for three pre-made models
Reported willingness to deploy each model
Explored data in data exploration phase & answer 3 related questions
sex
Male 0.26
Female 0.29
White 0.27
Other, 0.00
race
Black, 0.18
Asian-Pac-Islander, 0.28
Amer-Indian-Eskimo, 0.44
0.0
0.2
0.4
0.6
0.8
1.0
Absolute Metric Magnitude
Fairness Exploration Phase with Aequitas
Fairness Exploration Phase with Fairness Indicators
Answered 3 related fairness questions
Reported about willingness to deploy each model
Completed Post-Survey

Figure 2: Interview Protocol Methods. Participants underwent several stages of surveying, exploration, and query responding. The above chart depicts the flow of events each participant underwent in the hour-long session. A random and equal distribution of participants interacted with the Aequitas tool and with the Fairness Indicators tool.

Furthermore, tools should be able to detect various forms of bias. Mehrabi et al. and Olteanu et al. define a plethora of biases that can exist in the data and models [38, 43]. Of particular importance, representation bias, sampling bias, and proxy/association bias are all biases that participants discussed wanting to pay particular attention to. While sampling and representation bias can be deduced from many fairness toolkits, proxy/association bias to identify proxy sensitive attributes is mostly missing from fairness toolkits. An optimal fairness framework would provide users explicit feedback on different forms of biases.

4.1.3 Intervening at different stages of the ML life-cycle. The effects of algorithmic bias can occur at almost any point in the ML lifecycle,

Table 2: Rubric. Considerations based on fairness, usability, and insight criteria that can be used to both select and build the fairness toolkit best for practitioners

Criteria	Explanation
Criteria for Supporting Fairness Analysis	
Applicable to a diverse range of predictive tasks	Able to assess applicable model types (binary, multi-class, continuous, clustering, ranking, embedding, etc.)
Applicable to a diverse range of data types	Able to access applicable data formats (tabular, text, image, etc.)
Can detect bias	Able to identify varying types of bias (representation bias, sample bias, proxy/association bias)
Can mitigate bias	Proposes solutions for mitigating bias in the data or in the model
Can intervene at different stages of ML lifecycle	Able to mitigate bias at applicable stages like data collection/processing, model training, and post-model evaluation
Model-Agnostic	Able to assess models of any type
Fairness Criteria-Agnostic	Able to evaluate on custom performance metrics, while also providing a diverse set of fairness metric options
Performance Criteria-Agnostic	Able to evaluate on custom performance metrics, while also providing a diverse set of fairness metric options
Provides intersectional analysis	Able to look at performance across intersections of specified subgroups
Applicable to data without sensitive features	Able to use or suggest alternative methods for fairness analysis for data that does not contain sensitive attributes
Criteria for optimal tooling	
Provides a global perspective of fairness	Users can see globally what unfairness trends are across groups
Provides a local perspective of fairness	Users can see local examples of where trends of unfairness occur
Contextualizes fairness	Contextualizes definitions and scores according to user defined labels
Diversifies explanation style	Presents performance metrics in a variety of ways to support the diversity of thinkers (tables, single-dimension graphs, comparative graphs, etc)
Provides explicit interpretation of its limitations	Provides users with a comprehensive perspective on its limitations
Provides fairness recommendations	Provides optional recommendations for acceptable levels of deviation from fairness criteria
Well-Supported by Demos & Tutorials	Users should be able to apply, with little struggle, these tools in their context
Incorporates Components from other FAccT technologies	Utilizes components from other Explainable, Interpretable AI to support fairness analysis
General ease of use guidelines	Users do not struggle with interpreting, understanding, and manipulating tool
Influential on subsequent model processing	Successfully demonstrates trends that are influential in users' next steps

from the decision of what data to collect to build the model to the locations and communities that are impacted by the finished and deployed model [7, 21, 50]. Considering that some practitioners have more flexibility at different stages of their ML pipelines, it is important that options be made available across the board.

Users in this study, as seen in previous work [27], said that intervention in the data collection and pre-processing phase was the most important to them. However, few toolkits interrogate the methods used during this phase. At the very least, recommendations should be made on the impact of pre-processing techniques, like binning and feature engineering, on fairness. Furthermore, more options should be available to allow the users to investigate the distribution of their data and easily identify sampling issues.

Currently, some toolkits do provide options for during-training and post-training modifications [9, 39], which are also pivotal features to make available to students. Fairness experts should ensure that diverse, customizable options are provided and practitioners and institutions should ensure that their selected toolkit provides the interventions critical for their work.

4.1.4 Fairness & Performance Criteria Agnostic. Despite the comparatively small body of literature centered around algorithmic bias, translations on fairness are plentiful. Narayanan identified 21 definitions for fairness [40], and with the introduction of fairness into new applications of ML, this number will continue to grow. Several authors have identified general fairness categories within fairness research [8, 38, 52]. Users must identify the fairness definitions that best satisfy their interpretations of fairness given their context.

Verma & Rubin classified fairness definitions into five categories: those that focus on predicted outcome, those that focus on predicted and actual outcome, those based on predicted probability and actual outcome, those based on similarity measures, and those based on causal reasoning [52]. Currently, most fairness metrics provided by toolkits focus on the definitions that are based on predicted and actual outcome. These metrics for some applications, however, can be counter-productive [13]. Furthermore, this widely limits the varying fairness definitions that practitioners could focus on satisfying. While an optimal tool would allow for custom fairness metrics, it would also provide by default a representative subset of metrics.

Similarly to the fairness metrics criteria, performance metrics should be customizable and diverse. A useful feature, as suggested by P2, would allow users to do fairness-performance trade-off analysis between models.

4.1.5 Provides intersectional analysis. Several participants suggested intersectional analysis, which is a feature easily accessible in Fairness Indicators and completely feasible in Aequitas. Intersectional analysis across attributes is a highly under-emphasized practice [16, 26]. Fairness toolkits should encourage and have easily accessible subgroup analyses. Furthermore, fairness experts should be diligent about the means for which subgroup performance is displayed to the user. The number of groups multiply and, therefore, the information load required is substantially higher.

4.2 Criteria for Usable Fairness Tooling

While the fairness criteria was mostly defined by fairness experts, it is critical that practitioner feedback also be incorporated into the development and evaluation of fairness toolkits. While user suggestions are still critical in the former section, this section relies heavily on user feedback to support the usability and effectiveness of the toolkit. As previous work [27, 34, 35, 37, 50] suggests, practitioner feedback is of the utmost importance for successful and meaningful fairness impact.

4.2.1 Contextualizes Fairness. One of the first tasks users have, when engaging with fair ML tooling, is their decision on an acceptable definition of fairness, which we have found to be incredibly difficult and highly contextual. When asked to re-assess their willingness to deploy, many participants vocalized the overwhelming procedure of assessing fairness. Comments like, "This is a lot to decompose" (P3), "There was more information than I thought" (P4), and "There's so many stats here" (P6) confirmed that for both tools, users were overwhelmed by the number of fairness metrics and scores presented to them. When discussing the fairness metrics presented, P2 said,

> "It was hard to know like, 'okay, what one should I be looking at?' Obviously, it's good to have them all there but it was a little hard to like parse out like, 'oh, what exactly would be the ones to isolate to know what would be the best [in terms of] fairness'."

For both Fairness Indicators and Aequitas, fairness charts are displayed across a wide variety of fairness metrics. Fairness Indicators as a widget allows for immediate control of which fairness charts can be visualized, giving users more control of the immediate information load. For Aequitas, users have control of which metrics are printed, but they must re-run scripts to change the visualizations that they see.

While the number of metrics available present an overwhelming task, the fairness metrics themselves also contribute to this information overload. Metrics available in these tools require frequent interpretation in the evaluation process. Participants unknowingly had been asked to juggle the mathematical definition of metrics they rarely worked with, the contextual scenario for interpreting these metrics, and the potential harms that could results from a high, low, or uneven metric score. In the end, this task carried too much weight, so participants decided to go one of two directions: concluding on unfairness by group disparities across all metrics or by looking at one or two fairness metrics. P4 verbalized this same process of transitioning procedures during the analysis,

> "I looked at all fairness metrics together. Spent a lot of time looking at first one, but I felt more confidence when I looked across the board."

Currently, the overload itself deterred participants from wanting to engage with these tools. P6 said, "It's unlikely I would use this tool, To be honest, it's a lot of work to use it."

There is space for much work to be done on fairness tools in reducing the information overload for users. Providing contextualization within the fairness output could be promising. Furthermore, allowing users to share and store these 'harm scenarios' frees some of the mental capacity needed to actually weigh scenarios and make decisions.

The importance of contextualization can be seen in anthropology research [5], user experience research [48], and algorithmic bias research [41]. In this situation, contextualizing fairness most similar aligns with work done in information science research, where if given additional information about the task, systems can automate contextualizing components [10]. Simple support in the form of contextualization could be incredibly beneficial to fair ML toolkits.

4.2.2 Provides a global and local perspective of fairness and diversifies explanation styles. Individuals enjoyed the visual demonstration of scores, which P4 described as "self-explanatory." One participant (P6) did mention that it would be helpful if visual components were accompanied by separate tables. While Aequitas offers this option, external tables cannot be extracted for the Fairness Indicators widget. Another participant (P2) suggested a different type of visualization where users could compare metrics against each other. For example, P2 said,

> "I think it could be interesting to look at the accuracy Fairness trade-off. So that you just have a better sense of like, what are you giving up in accuracy as the model is enforcing these Fairness metrics?"

P2 also noted that it could be beneficial if fairness definitions/equations were incorporated into the fairness widget for easy access.

The interactive component for Fairness Indicators was helpful to many participants as it allowed them to explore and compare metrics. A few participants (P1, P2, P6, P9, P16) who saw this tool wanted functionalities that would allow for more easy comparisons between metrics. P6 suggested for the placement of selected metrics,

> "Like having them in one graph would make it exponentially easier to look through, kind of like being able to discern better between the different slices and models."

In the study, the Fairness Indicators widget was also accompanied by Google's TFX [1] statistical visualization widget, which allows users to visualize distributions of the data across attributes. Many participants referenced this widget when trying to understand group fairness scores. On the other hand, Aequitas participants often mentioned the need for looking at the data more to better understand what the tool was outputting.

As Dodge et al. found in their work, different explanation styles can be seen as inherently more fair than other styles [18]. Furthermore, Arya et al. emphasized the importance of diverse explanation styles [6]. It is critical that alternative depictions of fairness scores be presented. Many existing tools currently focus on a more global outlook of fairness, identifying group patterns or trends. Local depictions of metrics can be helpful in exemplifying global trends [18, 46]. These results are also emphasized through the study. Participants exhibited signs of information overload and mistrust that could be assisted by the support of diverse explanation types for fairness outcomes.

4.2.3 Provides explicit interpretation of limitations. Previous work has been extensive when it comes to over-reliance and over-trust of FAccT technologies [27, 31, 33]. The impact of fair ML toolkits in this study is evident when comparing willingness to deploy models before and after fairness analysis as was seen in Figure 4. Logistic Regression received extensive support because the fair ML toolkits deemed it less 'unfair'. Nonetheless, it still performed, with respect to performance metrics, as poorly as it had at the start of the study. This displays that over-reliance is most definitely a concern with these technologies and, therefore, they should be accompanied by preambles that detail the limitations and dissuade over-reliance, as was shown in [37].

4.2.4 Provides fairness recommendations. For Aequitas, participants were shown fairness plots color-coded green and red where colors represented whether or not, respectively, group scores satisfied fairness determination for that metric. Many participants enjoyed this color coding as it provided recommendations for metrics or groups that might be of concern. P3, P11, and P20 all mentioned that it would be nice if equations for how those values are calculated were included in the visuals. Relatively, P16, who saw the Fairness Indicators tool, suggested that their tool include green and red color-coding for recommendations on fairness. Participants from previous work [27] also emphasized a need for support beyond fairness detection. Recommendations can come in many forms. While mitigation techniques are very helpful, text-based recommendations for handling different types of bias is also preferential to bias detection alone.

4.2.5 Well-Supported by Demos & Tutorials. Within the post-survey, participants were asked about their prior level of exposure, interests, and opinions on their respective fairness toolkit. The majority of participants self-reported low or moderate levels of familiarity with fair ML and high interests toward using fair AI. Participants also had the opportunity to share their feedback on methodologies for learning how to use a fairness toolkit and what fairness interventions are best for them and their type of work. Of particular importance, Figure 3 depicts the difference of opinion in preferred fairness learning methodologies across product-focused practitioners and research-focused practitioners. This delineation of need when it comes to learning how to use fairness methodologies should be of utmost importance to both fairness experts and institutions wishing to normalize fairness procedures.

One important consideration is how fairness is demonstrated to practitioners. For this study, the task was binary. Similarly, demos for many fairness toolkits are simplified problems. These demos seem to be influential to practitioners when it comes to assessing the capabilities of fairness toolkits. For the purpose of this study, P19 had seen the Aequitas tool, but also shared that they had attended a demo of the Fairness Indicators tool. Of interest, P19 recalled about Fairness Indicators, "I don't know how relevant it is to our domain if it is like very specific to words."

Fairness Indicators often uses the Jigsaw's Unintended Bias in Toxicity dataset [12] as its interactive case study. This presumably is the same demo that P19 saw and attributed to Fairness Indicators capabilities. This same pattern emerged as people were exposed to fair ML toolkits for the first time using the Adult Census data. P7 said about their fairness toolkit, "I feel like these are more for comparing prototypes. It seems like its for simpler models." As was depicted in Figure 3, practitioners rank highly demos and case studies as a means of learning how to use fair ML toolkits. It seems logical that they would also use these methods to understand the functionalities and capabilities of these tools as well. Therefore, fairness experts must be intentional in the amount of time and effort they place on these supportive documentation. Additionally, a diverse set of documentation must be provided to supplement the diverse needs of different practitioners (reference Figure 3.

Furthermore, participants in this study had the opportunity to rank learning methods. These results, similar to the results of Holstein et al. [27] emphasize that users require domain-specific guides exemplifying fairness analysis across applications. Participants struggled to consider how fairness might work in their field, and supported through extensive documentation, tutorials, and demos is incredibly beneficial to practitioners. Furthermore, open lines of communication between fair ML toolkits developers and practitioners is helpful, as well.

4.2.6 Incorporates Components from other FAccT technologies. For improving the tool, participants often suggested features that can be seen in other FAccT toolkits. P2 and P7 both mentioned a similar feature they thought would be beneficial. "I would like to see a tool that could simulate how you could improve a model using fairness considerations" (P7). Currently, Google's What-If widget [24] provides this exact functionality. While this study focused on bias detection, suggestions for bias mitigation are definitely useful features for any fairness toolkit.

One major suggestion from users demonstrates features that you might see in an explainability toolkit. Several participants (P7, P8, P10, P11, P19) suggested that additional feature analysis be a part of the fairness consideration process. Since sensitive attributes can be reflected in other attributes, participants wanted to see what

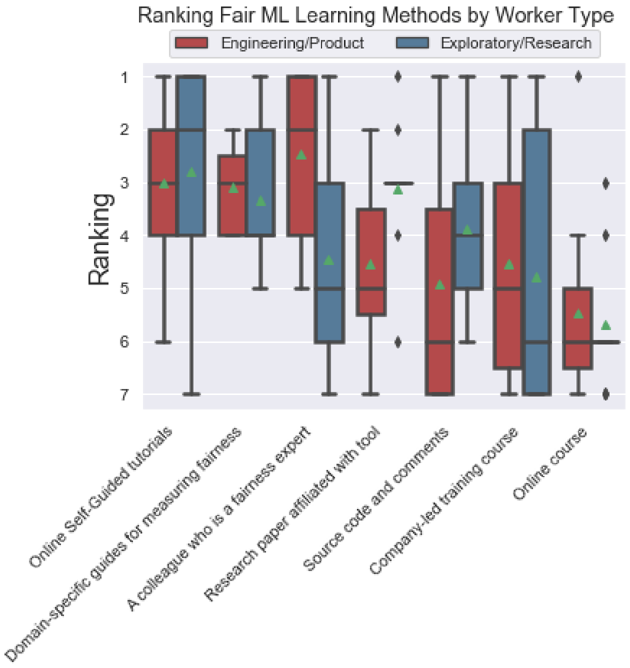


Figure 3: Ranking of fair ML Learning tools between product-focused practitioners and research-focused practitioners depicts that there are delineations between the needs of these groups even when it comes to learning about fair ML toolkits. Black diamonds demonstrate outlier points and green triangles mark the mean for each group.

features are guiding performance. A tool like LIME [46] would integrate well for this type of feature.

A small group of participants verbalized feelings of mistrust towards these tools. P10 even stated,

> "I'm not super like convinced by these plots to be fair. Like I feel it raises- I mean it's maybe interesting you know just to kind of get a feel for you know how things are stacking up against each other. But I feel it raises more questions than it really answers but maybe that's the [point]."

For fairness experts to garner the trust they need from practitioners, as much support should be provided to the toolkit. As suggested above, support in the form of help icons with definitions and equations and explanations could provide support to practitioners to better understand the information being shown. Furthermore, the implementation of explainability or interpretability could allow users to become more trusting of the results. Dodge et al. [18] found that intersecting fairness with explainability impacted user's perceptions on fairness outcomes, with some explanation types being more influential than others.

4.2.7 *General ease of use guidelines.* Users were asked 6 questions that gauged their understanding of the data, the models, and the fairness toolkits. For these questions, Wilcoxon test comparisons across the phases and Mann-Whitney test comparisons across the two different fairness toolkits were analyzed. There were no statistically significant differences between outcomes with respect to correctness, confidence, and timeliness ($p<0.05$), however this is not conclusive in light of the limited sample size. Comparisons between toolkits can be seen in Table 3. When comparing time between the Data Exploration Phase and the Fairness Exploration Phase, there was a significant difference in time taken by users to answer questions (Mann-Whitney, $U=21.0$, $p<0.01$), with the latter phase taking substantially longer than the former.

These results depict that participants were able to successfully gain basic insight with high confidence. Understandably, when it came to engaging with new fairness visualizations and metrics for the first time, users took more time to come to conclusions. However, with the high confidence scores across the tools, participants felt as though they were using and reading these tools successfully.

One participant with experience with fairness toolkits (P16), noted that having seen Fairness Indicators for the first time as a fairness expert, they would presume that this would be a difficult analysis for those without experience. P16 noted,

> "I know about these rates but I felt that there was not a lot of components for educating the user."

Despite fair ML having been a relatively specialized set of technologies, designed and used by a somewhat specific group of people, their broader adoption will still require an increased focus on usability of the tools that implement them. With the lack of enforcement of fairness considerations, many users are volunteering to use fair ML and could, therefore, be easily dissuaded from doing so by poor design decisions. In this study, a few participants were discouraged from using fair ML when features did not work as expected or interfaces were too difficult to use. Some participants even discussed preferring to do the analysis on their own over using readily available tools because of the difficulty in navigating or understanding output from the tools.

	Fairness Indicators	Aequitas
Average Correctness (0-1)	0.85	0.93
Average Confidence (1-7)	5.93	5.90
Average Time Taken (in seconds)	297.00 s	309.10 s

Table 3: Between the two toolkits that participants saw, there was no significant difference for the fairness phase between the correctness in answers, the confidence in answers, and the time taken to use each tool.

4.2.8 Influential on Model Processing. Participants were asked both before and after visualizing fairness analysis about their willingness to deploy each of the three models. Figure 4 depicts the change in opinion for each users across each model. Likelihood to deploy was measured on a scale from -3 to +3, with negative -3 being extremely unlikely to deploy and +3 being extremely likely to deploy. When Wilcoxon signed-rank tests were done compare willingness to deploy for each model before and after fairness analysis, outcome was significant for both LR (t=9.0, p<0.01) and NN (t=12.5, p<0.01). The average change in willingness for LR, RF, and NN models were 1.20, 0.55, and -1.75, respectively.

Participants were generally vocal about the insight brought by using the fairness toolkits and the fairness analysis process. P1 noticed that their willingness to deploy was heavily impacted by the fairness analysis.

> "[I]f I was going to be looking at this as somebody who's shipping things out, it would definitely make me think twice, because originally I talked about my neural net having potential to be shipped out, but it's obviously not as great for some groups versus others." (P1)

Several other participants also took notice of how the tool so clearly changed their perspective. Furthermore, the comparative group performance analysis was new to some. One participant (P5) verbally noted,

> "These graphs facilitate my thinking about [fairness] and comparing these models. [...] I wouldn't have thought of these comparisons without seeing this graph."

The impact of fairness analysis was evident in the change in perspective. While participants were understandably never too enthusiastic to launch models they were unfamiliar with, fairness analysis did provide additional information that was influential on their decision to deploy. Despite how performance for the LR model was substantially lower than for the NN model, when participants served as the decision-maker in these scenarios most all of them sacrificed traditional performance for 'fairer' outcomes.

A central goal behind any FAccT tool is to inspire change in the motivations, means, and results of machine learning algorithms. Many of the components of this portion of the Insight and Usability section work towards guaranteeing that, if necessary, practitioners counteract any bias they might encounter. Furthermore, fair ML toolkits should be influential on practitioners and stakeholders. P7 mentioned that their respective tool would be useful for convincing stakeholders of fairness considerations,

> "I think this has potential as a tool that you can demonstrate to the business side [for] why fairness considerations are not degrading product performance."

Fairness experts should focus on being interpretable and influential for both practitioners and their lead decision-makers who may not be ML-experts. If, for any reason, change does not occur, attention must be placed on whether the tool is effectively performing its responsibilities.

5 CONCLUSION

Many of the practitioners in our study vocalized that this was their first experience considering fairness in a systematic way. For those with similarly formatted data and access to sensitive attributes, the importance and the accessibility of fairness toolkits was made apparent. However, a large number of participants (P2, P3, P4, P6, P8, P9, P11, P18, P19) voiced their concerns on whether or not fairness was applicable to their work. Participants mulled over what fairness would even mean in the context of their domain. While some participants concluded that fairness (defined via group performance) was not relevant to their work, others began to formulate what fairness might look like and consider the impact of their work from a new angle. This discussion of how to define fairness in context, determine the impact of unfairness, and decide how to respond to it is critical in team settings and the use of fair ML tools in some capacity encourages practitioners to define fairness in their own domain. Furthermore, the diversity of fairness issues detected and the solutions offered by practitioners suggests that this discussion might be the most pivotal first step in incorporating fairness. The blindspots of practitioners still exist [27] and the easiest solution is one that is through collaboration. When comments were consolidated, participants in this study were able to expose more bias and propose more solutions than currently any one toolkit can provide. Therefore, the importance of fairness education for ML practitioners and fairness discussion within ML teams cannot be over-emphasized.

The future of fair ML heavily depends on the participation and feedback of practitioners. This feedback informs design needs and ensures fairness tools are used and used effectively. While the fairness needs can be clearly supported by literature, the assurance that tools are usable and allow successful insight must be validated by the practitioners. By engaging with practitioners, this work was able to confirm the importance and impact of fair ML toolkits, elicit the fairness background of potential users, identify methods best suited for fairness learning, and identify need areas in fair ML design. Through feedback from practitioners and support from previous literature, this work introduced a rubric that summarizes critical components necessary for any complete fair ML toolkit.

From this research, we have developed suggestions for institutions, fairness experts, and practitioners. First, institutions should be actively involved in the fairness process. They should incentivize fairness analysis, remain knowledgeable on fairness research,

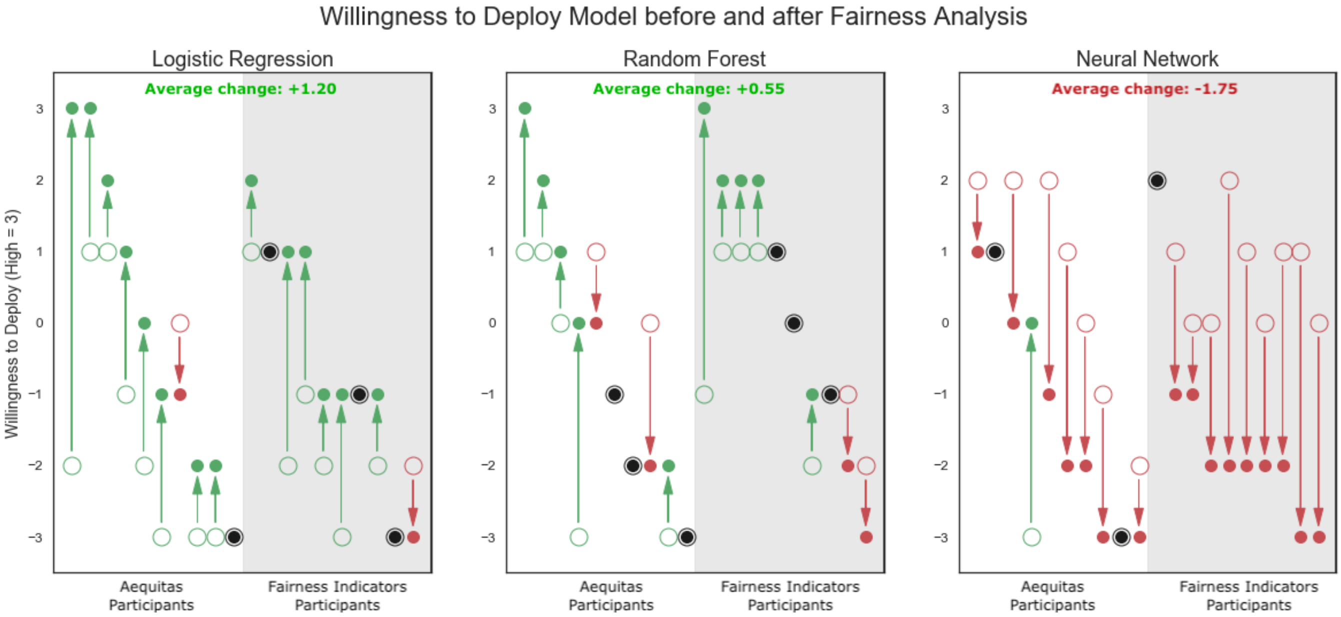


Figure 4: Participants' significant change in willingness to deploy before and after fairness analysis depicts the influential impact that fair ML toolkits has on decision-making. In particular, the Neural Network, which had no fairness processing done on it was the most favored before fairness analysis and the least favored after analysis. Empty circles represent participants starting willingness, while filled circles represent their final willingness to deploy. Green is used to depict participants with an increased willingness to deploy, red depicts a decrease, and black demonstrates no change.

and provide infrastructure that supports organizational fairness guidelines. Furthermore, they should provide their practitioners with a diverse set of fairness supports that allows for successful learning and application of fair AI. Institutions should ensure that practitioners have all the tools necessary, including access to fairness experts and domain-specific guides to fairness research. This rubric can be used by the actors within institutions in selecting the toolkit(s) that best fit the needs of their practitioners.

Second, fairness experts could fill the existing gap in the development of a more complete 'One Stop Shop' toolkit for fairness. This could be best demonstrated as an open-source fairness toolkit where the community of fairness experts could incorporate their contributions while also providing practitioners an easy means of critiquing. Fairness experts should diversify their toolkits to consider the wide variety of tasks that exists across the ML space. In addition, critical fairness components like recommendations for mitigation and providing fairness interventions for early ML stages are currently missing from many toolkits. Furthermore, several features that could guide the fairness decision-making process and reduce information overload for practitioners are missing.

Third, practitioners should also be active participants in implementing fairness into their work spaces. They should be diligent in learning and selecting the tool that best fits their work. Practitioners interested in fair ML should advocate for its implementation in their teams and projects. This support can be incentived by toolkit creators and institutions. Furthermore, the fairness research community is still fairly new and feedback and support is encouraged for most all technologies. This open line of communication should be made apparent by toolkit creators and used often between practitioners and fairness experts.

This work exemplifies what we believe could be the beginning of a new era of fair ML. As can be seen in the design of high-stakes products, fair ML should undergo iterative design procedures that actively engage practitioners and stakeholders. Future work will focus on a wider array of practitioners, especially those with medial to low interest in fairness. Furthermore, the results of this work should be used to inform a more complete set of fair ML that can be incorporated into high-stake product and research teams.

5.1 Limitations.

Limitations of the study include the sample size, participation bias, and available tooling. Future work can focus on a larger subset of practitioners from a more diverse background with respect to institution affiliation and interests in fair ML. Furthermore, without social distancing restrictions, different techniques, such as participatory design or in-person group discussions about fairness, could be used to further unpack how practitioners would use tools in different environments. Furthermore, this study did not do a comparative analysis of tool features, so future work would focus on deciphering the strengths and weaknesses of toolkit features. Finally, the current study methodology did not focus on isolating differing translations of fairness. Future work would interrogate how tooling impacts an individual's definition of fairness in context.

ACKNOWLEDGMENTS

Firstly, we want to send a huge thanks to our participants for taking the time to participate in this study. We had such an enthusiastic group of participants who were incredibly excited about the future of this work. Furthermore, we want to send thanks to all the individuals who helped recruit and spread the word. Lastly, we want to thank a few of our awesome team members who helped extensively along the way, including: Nediyana Daskalova, Ziang Xiao, and Lex Beattie.

REFERENCES

[1] 2020. TensorFlow Extended (TFX) | ML Production Pipelines. https://www.tensorflow.org/tfx

[2] 2020. Tensorflow's Fairness Evaluation and Visualization Toolkit. https://github.com/tensorflow/fairness-indicators

[3] ACM. 2020. ACM FAccT. https://facctconference.org/

[4] Aniya Aggarwal, Seema Nagar, and Diptikalyan Saha. 2019. Black Box Fairness Testing of Machine Learning Models. 11 (2019). https://doi.org/10.1145/3338906.3338937

[5] Rikke Sand Andersen and Mette Bech Risør. 2014. The importance of contextualization. Anthropological reflections on descriptive analysis, its limitations and implications. *Anthropology and Medicine* 21, 3 (sep 2014), 345–356. https://doi.org/10.1080/13648470.2013.876355

[6] Vijay Arya, Rachel K E Bellamy, Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R Varshney, Dennis Wei, and Yunfeng Zhang. [n.d.]. *One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques.* Technical Report. arXiv:1909.03012v2 http://aix360.

[7] Tobias Baer. 2019. *Understand, Manage, and Prevent Algorithmic Bias: A Guide for Business Users and Data Scientists.* Apress.

[8] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and machine learning.* fairmlbook.org. https://fairmlbook.org/index.html

[9] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *Advances in Neural Information Processing Systems* 2017-Decem, Nips (oct 2018), 5681–5690. arXiv:1810.01943 http://arxiv.org/abs/1810.01943

[10] Jay Budzik and Kristian Hammond. 1999. Watson: Anticipating and Contextualizing Information Needs. In *Proceedings of the Sixty-second Annual Meeting of the American Society for Information Science.*

[11] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. 2017. Fair Clustering Through Fairlets. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS).* Long Beach, CA, 5029–5037.

[12] Civil Comments. 2019. Jigsaw Unintended Bias in Toxicity Classification . https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

[13] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. (jul 2018). arXiv:1808.00023 http://arxiv.org/abs/1808.00023

[14] Henriette Cramer, Sravana Reddy, Romain Takeo Bouyer, Jean Garcia-Gathright, and Aaron Springer. 2019. Translation, tracks & Data: An algorithmic bias effort in practice. In *Conference on Human Factors in Computing Systems - Proceedings.* Association for Computing Machinery. https://doi.org/10.1145/3290607.3299057

[15] Kate Crawford. 2017. The Trouble with Bias. https://www.youtube.com/watch?v=fMym{_}BKWQzk

[16] Kimberle Crenshaw. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics . *University of Chicago Legal Forum* 1989, 1 (1989), 139–167. http://chicagounbound.uchicago.edu/uclfhttp://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8

[17] Matt Day. 2016. How LinkedIn's search engine may reflect a gender bias. https://www.seattletimes.com/business/microsoft/how-linkedins-search-engine-may-reflect-a-bias/

[18] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. *International Conference on Intelligent User Interfaces, Proceedings IUI* Part F147615 (jan 2019), 275–285. https://doi.org/10.1145/3301275.3302310 arXiv:1901.07694

[19] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[20] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2018. A comparative study of fairness-enhancing interventions in machine learning. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* (feb 2018), 329–338. arXiv:1802.04422 http://arxiv.org/abs/1802.04422

[21] Salvador García, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez, and Francisco Herrera. 2016. Big data preprocessing: methods and prospects. *Big Data Analytics* 1, 1 (2016), 9.

[22] Jean Garcia-Gathright, Aaron Springer, and Henriette Cramer. 2018. Assessing and Addressing Algorithmic Bias - But Before We Get There. In *Proceedings ofthe AAAI2018 Spring Symposium: Designing the User Experience ofArtificial Intelligence.* arXiv:1809.03332 http://arxiv.org/abs/1809.03332

[23] Google. 2020. Google Colaboratory.

[24] Google. 2020. What-If Tool. https://pair-code.github.io/what-if-tool/get-started/

[25] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems.* 3315–3323.

[26] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22, 7 (2019), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

[27] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudík, and Hanna Wallach. 2018. Improving fairness in machine learning systems: What do industry practitioners need? *Conference on Human Factors in Computing Systems - Proceedings* (dec 2018). https://doi.org/10.1145/3290605.3300830 arXiv:1812.05239

[28] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (sep 2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[29] M. I. Jordan and T. M. Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. , 255–260 pages. https://doi.org/10.1126/science.aaa8415

[30] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (oct 2012), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[31] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3313831.3376219

[32] Ronny Kohavi and Barry Becker. 1996. Adult Data Set. http://archive.ics.uci.edu/ml/datasets/Adult

[33] Himabindu Lakkaraju and Osbert Bastani. 2019. "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (nov 2019), 79–85. arXiv:1911.06473 http://arxiv.org/abs/1911.06473

[34] Po-Ming Law, Sana Malik, Fan Du, and Moumita Sinha. [n.d.]. *The Impact of Presentation Style on Human-In-The-Loop Detection of Algorithmic Bias.* Technical Report. arXiv:2004.12388v3 https://youtu.be/8ZqCKxsbMHg

[35] Po-Ming Law, Sana Malik, Fan Du, and Moumita Sinha. 2020. Designing Tools for Semi-Automated Detection of Machine Learning Biases: An Interview Study. In *Proceedings of the CHI 2020 Workshop on Detection and Design for Cognitive Biases in People and Computing Systems.*

[36] Pranay K. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, and Ruchir Puri. 2018. Bias Mitigation Post-processing for Individual and Group Fairness. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2019-May (dec 2018), 2847–2851. arXiv:1812.06135 http://arxiv.org/abs/1812.06135

[37] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *CHI Conference on Human Factors in Computing Systems.* ACM, Honolulu. https://doi.org/10.1145/3313831.3376445

[38] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. (2019). arXiv:1908.09635v2

[39] Microsoft. 2020. Fairlearn. https://fairlearn.github.io/

[40] Arvind Narayanan. 2018. 21 fairness definitions and their politics. https://fairmlbook.org/tutorial2.html,https://www.youtube.com/embed/jIXIuYdnyyk

[41] David T. Newman, Nathanael J. Fast, and Derek J. Harmon. 2020. When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes* 160 (sep 2020), 149–167. https://doi.org/10.1016/j.obhdp.2020.03.008

[42] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism* (first ed.). NYU Press. https://doi.org/10.2307/j.ctt1pwt9w5

[43] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data* 2, 13 (jul 2019), 13. https://doi.org/10.3389/fdata.2019.00013

[44] Qualtrics. 2020. Qualtrics. https://www.qualtrics.com

[45] Tahleen Rahman, Bartlomiej Surma, Michael Backes, and Yang Zhang. 2019. FairWalk: Towards fair graph embedding. In *IJCAI International Joint Conference*

on Artificial Intelligence, Vol. 2019-August. International Joint Conferences on Artificial Intelligence, 3289–3295. https://doi.org/10.24963/ijcai.2019/456

[46] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. [n.d.]. *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*. Technical Report. arXiv:1602.04938v1

[47] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. 2018. Aequitas: A Bias and Fairness Audit Toolkit. (nov 2018). arXiv:1811.05577 http://arxiv.org/abs/1811.05577

[48] Marcus Specht, Andreas Lorenz, and Andreas Zimmermann. 2006. An architecture for contextualized learning experiences. In *Proceedings - Sixth International Conference on Advanced Learning Technologies, ICALT 2006*, Vol. 2006. 169–173. https://doi.org/10.1109/icalt.2006.1652397

[49] Sriram Vasudevan, Cyrus DiCiccio, and Kinjal Basu. 2020. Addressing bias in large-scale AI applications: The LinkedIn Fairness Toolkit. https://engineering.linkedin.com/blog/2020/lift-addressing-bias-in-large-scale-ai-applications

[50] Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4, 2 (2017). https://doi.org/10.1177/2053951717743530

[51] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (feb 2018). https://doi.org/10.1145/3173574.3174014 arXiv:1802.01029

[52] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. *IEEE/ACM International Workshop on Software Fairness* 18 (2018). https://doi.org/10.1145/3194770.3194776

[53] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive Inequity in Object Detection. (feb 2019). arXiv:1902.11097 http://arxiv.org/abs/1902.11097

[54] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A fair top-k ranking algorithm. In *International Conference on Information and Knowledge Management, Proceedings*, Vol. Part F131841. Association for Computing Machinery, New York, NY, USA, 1569–1578. https://doi.org/10.1145/3132847.3132938