The Landscape and Gaps in Open Source Fairness Toolkits

Michelle Seng Ah Lee
Compliant & Accountable Systems Group University of Cambridge, UK
michelle.sengah.lee@cst.cam.ac.uk

Jatinder Singh
Compliant & Accountable Systems Group University of Cambridge, UK
jatinder.singh@cst.cam.ac.uk

ABSTRACT

With the surge in literature focusing on the assessment and mitigation of unfair outcomes in algorithms, several open source 'fairness toolkits' recently emerged to make such methods widely accessible. However, little studied are the differences in approach and capabilities of existing fairness toolkits, and their fit-for-purpose in commercial contexts. Towards this, this paper identifies the gaps between the existing open source fairness toolkit capabilities and the industry practitioners' needs. Specifically, we undertake a comparative assessment of the strengths and weaknesses of six prominent open source fairness toolkits, and investigate the current landscape and gaps in fairness toolkits through an exploratory focus group, a semi-structured interview, and an anonymous survey of data science/machine learning (ML) practitioners. We identify several gaps between the toolkits' capabilities and practitioner needs, highlighting areas requiring attention and future directions towards tooling that better support 'fairness in practice.'

CCS CONCEPTS

• **Social and professional topics** → *User characteristics*; • **Software and its engineering** → **Software libraries and repositories**; • **Human-centered computing** → **Empirical studies in interaction design**; **Open source software**; *Empirical studies in collaborative and social computing*; *Visualization toolkits*; Empirical studies in visualization; Visual analytics; **User studies**; Usability testing; User interface programming; **User interface toolkits**.

KEYWORDS

fairness, bias, algorithm auditing, open source toolkits, fairness toolkits, algorithmic fairness, bias detection, bias mitigation

ACM Reference Format:
Michelle Seng Ah Lee and Jatinder Singh. 2021. The Landscape and Gaps in Open Source Fairness 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/10.1145/3411764.3445261

1 INTRODUCTION

Algorithms, especially those using machine learning (ML), are increasingly pervasive across industry and domains. As they influence decisions with significant impact on our lives, there is growing awareness and concern that the algorithmic predictions may result in unfair outcomes influenced by societal inequalities and exacerbating past discriminatory patterns.

The perceived unfairness of algorithm-assisted decisions are becoming headline news. A model used to assign a recidivism score in the criminal justice system sparked controversy when it was accused of overestimating the risk of black defendants [13]. Applecard by Goldman Sachs was investigated by a regulator after customer complaints of women receiving lower credit limits than men with the same credit standing [40]. Recruiting tools at a technology company was reportedly biased against women [10]. This perception of fairness can impact companies' profitability; for instance, an experiment found that on average, people move twice as much money away from banks that use algorithms in loan application decisions when told that they draw on proxy data for race and gender or social media data [8]. Interestingly, it was also found that people move money away from a bank when only told that it uses "advanced computing techniques," implying the general discomfort with the use of more complex algorithms [8].

The debates on fairness in systematic and algorithmic decision-making has long predated the prevalence of ML [19]. However, there is increasing scrutiny on ML models, which can present additional challenges given their often complex and opaque predictive mechanisms, and because their effects can have substantial, wide-reaching impact when deployed at scale [19].

To address the risk of unintended biases resulting in unfair algorithmic predictions, a number of private companies, research institutions, and public sector organisations have issued principles and guidelines for ethical artificial intelligence (AI). A systematic investigation of 84 such documents found "justice, fairness, and equity" is the second most prevalent principle after "transparency" [20]. New legislation in Denmark that came into effect in July 2020, introduces a mandatory requirement for companies in online space to publicly release information about their data ethics policies [15]. With increased public and regulatory scrutiny, industry and government leaders have strong incentives to mitigate the risk of unfair outcomes. The recently published AI Barometer (UK) finds that across four domains—criminal justice, financial services, health/social care, and digital/social media—"bias leading to discrimination" is found to be of both high likelihood and high impact in its risk [37]

Given these concerns, there is growing demand for technical methods that can test for potential unfairness. There is a recent proliferation of toolkits and packages for assessing fairness in algorithms, particularly where less-interpretable ML methods are used. Substantial academic literature on algorithmic fairness has concerned the development of mathematical and computational definitions of fairness (e.g. [11, 14, 26]), prompting work to explain the distinctions between them [31, 39] and the trade-offs [24] given

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

ACM ISBN 978-1-4503-8096-6/21/05.
https://doi.org/10.1145/3411764.3445261

some of these definitions are impossible to simultaneously meet. In turn, this has led to the introduction of automated bias mitigation techniques, including pre-processing methods to remove estimated bias from the data set [6, 12, 21], in-processing methods to train a model to both maximise accuracy and increase fairness [22, 46], and post-processing methods to adapt the predictions after the model build to equalise a metric of fairness between groups [14, 21, 32].

Recently, various open source 'fairness toolkits' have been developed to make these fairness methods more widely accessible to model developers. As these toolkits are to be integrated into developers' model build process, they have the potential to help improve fairness testing and mitigation at-scale across domains (if and where appropriate). On the other hand, there is a risk of these toolkits being applied to an inappropriate use case, misinterpreted without considering the assumptions or limitations of the implemented methods, and/or misused (deliberately or otherwise) as a flawed certification of an algorithm's fairness. There is currently no available general and comparative guidance on which tool is useful or appropriate for which purpose or audience. This limits the accessibility and usability of the toolkits and results in a risk that a practitioner would select a sub-optimal or inappropriate tool for their use case, or simply use the first one found without being conscious of the approach they are selecting over others.

Towards these concerns, the core contribution of this paper is the *identification of gaps between the capabilities of existing open source fairness toolkits and the requirements of practitioners*, highlighting the implications of these gaps *to inform the development of fairness-related tooling*. This is to move forward the effort to make fairness assessment accessible beyond academia and across industry by engaging practitioners to uncover their needs. Specifically, we assess the relative importance for practitioners of functionality and usability features in the context of an 'ideal' fairness toolkit. We then compare and evaluate an indicative sample of existing toolkits to summarise relevant criteria impacting toolkit selection. Finally, we identify gaps requiring urgent attention in order for toolkits to be useful for practitioners in addressing fairness issues in their real-world scenarios. Toolkits suitability designed for users (practitioners) helps better support proper use of the toolkits and accelerates their adoption. Poorly implemented or poorly explained toolkits risk engendering false confidence in flawed methodologies.

In order to identify these gaps, we use a multi-method study: an exploratory focus group, a semi-structured interview, and a survey. Our methodology was designed to identify high-level issues with practitioners with prior knowledge of algorithmic fairness challenges in their products, drilling down into the industry needs and their perceived gaps in the fairness toolkits. We validate the findings from from prior stages in each subsequent method. Details of the method are further discussed in §3.

In addition to the human subject research, we conduct comparative review of the current capabilities of six prominent open source fairness toolkits, identified as a part of the initial exploratory focus group: scikit-fairness / scikit-lego [41], IBM Fairness 360 [4], Aequitas Tool [33], Google What-if tool [44], and Fairlearn [29]. We identify and discuss the tool features validated as important to the interviewees and survey respondents, including regarding the tools' open source license, supported software, and developing organisation. We compare the coverage of fairness metrics and mitigation implementations in each toolkit.

We identify several gaps between the tools' capabilities and the practitioners' needs in terms of: 1) functionality, 2) user-friendliness, and 3) contextualisation, as initially identified in the focus group and validated in the interview and survey. Our findings highlight potential directions for future toolkit development that are vital prerequisites for their widespread adoption and usage. We find that there is a steep learning curve for practitioners who are not well-versed in the academic literature on fairness, which is insufficiently addressed by the toolkit guidance documents, and is exacerbated by the user interface design that either overwhelms the user with information or oversimplifies the problem. The toolkits also have limited end-to-end coverage of the model build pipeline (design-build-deploy), with notable gaps in guidance on data sampling, proxy analysis, and mitigation strategy. We also find that the practitioners' main priority is often the toolkit's ability to adapt and integrate with their existing workflow and use case, a concern that is currently under-considered. Our discussion highlights areas requiring attention as a means for improving the support and engagement of practitioners on fairness-related challenges

2 RELATED WORK

In the technical communities, there has been an explosion of recent work introducing different mathematical formalisations of fairness (e.g. [12, 14, 26]), various technical implementations to for "debiasing" [6, 12, 21, 22, 32, 46], and the distinctions and trade-offs among the mathematically incompatible definitions [24, 28, 39]. Some interdisciplinary work with philosophy, economics, and law has shown the limitations and flawed ethical assumptions of the mathematical definitions of fairness [17, 28, 42], with some arguing that fairness as a concept is too context-specific and complex to be automated [28, 42].

HCI researchers have studied fairness with respect to public expectations and user interfaces for evalulating ML systems. Studies of public perception of algorithmic decision-making have repeatedly shown gaps between the consumer / users' perspectives on fairness and the mathematical definitions [27, 36, 45]. Researchers have studied the design and development of tools to support developers in assessing and monitoring their models and systems [2, 7, 25], but none specifically in fairness testing. Several have called for tools to support industry practitioners in developing fairer products [9, 35].

There is no work to-date, to the best of our knowledge, that reviews the features of fairness toolkits, nor of how developers perceive the importance and usability of particular features. This is likely partly attributed to their novelty, having been developed in the past two years. When released, IBM Fairness 360 briefly described past toolkits that were relatively limited in functionalities. The objective of Fairness 360 was to "unify these efforts and bring together in one open source toolkit" [4] by including the previous toolkits' features in their new product, and did not review other tools that have a different approach to fairness all together, such as Pymetrics audit-ai and Fairlearn. In contrast to IBM Fairness 360's calculation of fairness metrics on a group level, audit-ai focuses on statistical tests to assess the likelihood the disparity is due to random chance. While IBM Fairness 360 focuses on "debiasing," Fairlearn

defines fairness in relation to potential harm and explicitly avoid the term "bias." The key disparities are outlined in §4. In addition, several tools (e.g. scikit-fairness and Google What-if tool) were introduced after Fairness 360.

There are two key prior studies on high-level fairness challenges faced by practitioners. A past interview and survey of ML practitioners identified challenges they face in algorithmic fairness that they felt unresolved [18]. Another study assessed the needs in high-stakes public sector decision–making specifically with exploratory interviews [38]. Note that this is a fast-moving area, and their studies, conducted between mid-2016 to mid-2018, largely pre-date the release of open source fairness toolkits, and therefore such work only investigates the challenges practitioners face, not whether and to what extent existing tools are or could be useful. Our focus group, interviews, and surveys have reaffirmed some of the top-level themes in Holstein et al.[18]; however, our study's focus is on open source and widely-available tools (thereby providing new takeaways), while ML practitioners may also use commercial tools or those built in-house. Generally, open source toolkits can have broad impact, enabling wider adoption and access due to lower costs, accelerated experimentation and delivery, and reduced dependency on third parties [43]. Commercial tools are also challenging to assess and compare due to their proprietary nature and limited accessibility, making a gap analysis infeasible. There are also open-source considerations in procurement, requirements, and criteria, e.g. license review, frequency of updates, robustness, etc., which were not explored in previous work but we find to be important to the practitioners.

Without an understanding of the current landscape, one cannot conclude whether the challenges reported in Holstein et. al and Veale et al. continue to be truly unaddressed gaps. In addition to capturing high-level practitioner perspectives, we conduct a bottom-up review of open source toolkits by examining the code, documentation, and visualisation. In contrast to previous work, our paper aims to provide a feature summary of existing toolkits and discusses the gaps between their offering and the practitioners' requirements, specifically in open source toolkits that are best placed to have a widespread impact due to their accessibility.

3 METHODOLOGY SUMMARY

Our overall methodology entailed four steps:

(1) **Exploratory focus group** to identify prominent fairness toolkits and derive initial insights;
(2) **Comparative review of the selected toolkits** to compare the features available in each toolkit;
(3) **Semi-structured interviews** of practitioners with prior experience in fairness challenges to understand the features in an ideal toolkit and rate how well each of the six toolkits meet their needs; and
(4) **Survey** to validate the findings with a broader group and probe into a few insights from previous stages.

Our approach is structured to first derive the context and scope through the exploratory focus group and then undertake a deep-dive into the initial findings with the interviews and surveys. To get an overview of the fairness toolkits available and assess their capabilities, we started with an exploratory focus group of practitioners with an interest in the intersection between data science / ML and ethics. Then, we organised semi-structured interviews with industry practitioners with first-hand experience in fairness challenges. Finally, we designed a survey to reach a wider audience with more diverse levels of familiarity with algorithmic fairness to validate the earlier findings. The study went through our Departmental ethical review process. The questions and associated consent form for the interview and survey are included in the supplementary materials.

We will describe in more detail the methodologies, and results from the different phases of research in subsequent sections. Specifically: §4 covers the output of the exploratory focus group, which helped scope and frame the study, as well as identify initial insights; §5, the feature comparison and assessment independently conducted by the authors, which supplements the human subject research; and §6 describes the gaps identified in the interviews with supporting evidence and deep-dives from the surveys.

4 FOCUS GROUP: TOOL SELECTION & GAPS

In addition to defining research scope, the exploratory focus group identified insights to be validated in the interviews and survey, which provide a more in-depth view of the practitioner perspectives to contextualise our findings. Based on the toolkit search in the focus group, we identified six toolkits for our review which contributes by providing a systematic comparison of tool capabilities.

4.1 Focus group methodology

As the first exploratory step, we organised a focus group through a charity called DataKind UK, a community of pro bono data scientists who donate their skills to mission-driven organisations. Our aim was to get an intensive, expert-driven exploratory analysis of the current open source tool landscape in algorithmic ethics. The event, called "Ethics DataDive" brought together 18 participants for a virtual day (10am-6pm) on June 13, 2020, recruited through the charity network mailing list with several targeted invitations to individuals actively engaged in algorithmic ethics. While initially planned as an in-person event, physical meeting was not possible due to COVID-19 restrictions. Given the participants were already familiar with the relevant literature and debates, curating this group—rather than randomly sampling—allowed for more rapid and in-depth assessment of the toolkits without the need for the initial preparation or training on relevant material. The group invited by the Executive Director and Ethics Committee members based on prior active participation at DataKind UK and/or known expertise in algorithmic ethics.

The focus group consisted of 18 participants. There were 3 practitioners in the group actively investigating open source fairness toolkits, and the remaining 15 provided their input in the three presentations back to the wider group by asking questions, providing comments, and suggesting areas for investigation. The three practitioners selected were data scientists with prior experience building bespoke fairness testing models, with two working in cross-industry professional services and one in financial services.

Prior to the event, we divided volunteers into groups of 3-5 practitioners to focus on five topics in algorithmic ethics selected by the researcher following a discussion with DataKind UK Ethics Committee: 1) fairness, 2) explanation, 3) natural language processing,

4) checklists, and 5) communication strategies. We asked the participants to find toolkits in their topic area and score them on a number of metrics of functionality and user-friendliness, encouraging them to write down their assessments.

The overall objective was explained to the whole group: to explore the existing open source tools that seek to assist data science practitioners with ethical challenges. Participants were then split into the sub-groups, assigned by prioritising their stated preference collected in their registration form while maintaining a fairly even split in numbers among the groups. Two catch-up sessions occurred throughout, where each group presented their findings so far and solicited questions and comments from the wider group. The day culminated in a final presentation from each group on the key takeaways and insights.

The number of toolkits and their diversity was the greatest for fairness; and though other four groups returned some interesting findings, this drove our research to focus on fairness toolkits. The focus group viewed that the explanation toolkits were more homogeneous than the fairness toolkits, thus a comparison and gap analysis in the former was deemed to be less insightful than in the latter. The natural language processing toolkits were concluded as too nascent and lacking in robustness to be able to implement into a developer's workflow without major modifications. The checklists and communication strategies groups reviewed diverse sets of documents but only found two of them as possibly helpful. This paper only considers fairness; the focus group's high-level observations in the other four areas are planning to be published separately.

In all, the main outputs of the focus group were two-fold: 1) specifying the scope of the toolkit assessment through the fairness toolkit search (discovery) process, and 2) identification of some initial insights and reactions from the practitioners involved.

4.2 Fairness toolkits identified

There are a number of fairness-related toolkits available. Therefore, one of the objectives of the focus group was to identify those that are most relevant and potentially useful. We took suggestions on tools to review from the focus group based on toolkits they've encountered in the past, and the suggested list was compiled and distributed to the focus group. Each member also searched online for additional candidates to review, discarding tools that were not directly related to resolving fairness concerns and/or with insufficient documentation or guidance, which would make it difficult to adapt the code or user interface for their use case. As we validate in the interview and the survey in §4, this process accords with how a typical technical team might search for an applicable fairness tools.

The six fairness toolkits were selected and reviewed through the focus group: scikit-fairness / scikit-lego [41], IBM Fairness 360 (AIF360) [4], Aequitas Tool [33], Google What-if tool [44], and Fairlearn [29]. Note, our intention was not to conduct a complete and holistic review of all available toolkits, but to systematically review a selected indicative example of fairness toolkits as a means for exploring their implications and considerations in practice. The toolkits were selected such that they 1) are likely to be found by a practitioner searching for a toolkit, 2) are open source, and 3) with relevant technical implementations of fairness-related methodologies. The focus group (as practitioners) were asked to select the tools using the same search process they would normally use, and to choose those they felt most useful and well-documented. As such, there are other toolkits that were either either discarded from the assessment in the focus group, or were not uncovered by the focus group's search efforts. Tools that were discarded from evaluation include those that contained limited technical implementations (e.g. Ethics and Algorithm Toolkit[1] and Google AI Explorables[2]) and those that were not open source (e.g. Model Guardian[3]).

Further, one theme highlighted in the focus group and supported in the interview and the survey is the participants' surprise at the toolkits' diversity. Given that toolkits significantly differ, it is important to consider a range of toolkits and their potential context of use.

4.3 Current toolkit features and evaluation

The focus group, after the initial search, divided up the toolkits among themselves for an assessment. They were given a template to fill out with example open source data sets for evaluation (see supp. material). The participants were asked to include the links to the tool and documentation, a description and an evaluation of the tool's pros and cons, with supporting screenshots from the toolkit. The assessment from the one-day exercise found that the current features of the open source toolkits include the abilities to:

- Test against standard set of fairness metrics, such as demographic parity, equal opportunity, etc.;
- Identify the most "similar" cases to compare predictions (only available in the Google What-if tool);
- Experiment with single or multiple decision thresholds for classification problems;
- Provide explanations for a prediction, e.g. feature importance, through integration with explanation toolkits;
- Produce visualisations on fairness and accuracy metrics; and
- Augment model-building process with fairness in mind (e.g. debiasing or constraint during training).

The group identified some initial high-level gaps in the open source tool landscape, which we futher validate with in-depth, semi-structured interviews. In the focus group, only between one and five people assessed each toolkit, limiting external validity due to the subjective nature of the scoring, hence using the interviews to validate the toolkit assessments.

In rating the toolkits, Fairlearn and What-if tool were considered the best overall for functionality and features. Fairlearn was commended for the easy-to-use dashboard and the coverage of fairness metrics and mitigations. Google What-if tool was seen to have the most comprehensive visualisation that can be endlessly customised. The group noted scikit-fairness may have the highest potential impact, as it integrates the most seamlessly with scikit-learn package in python frequently used for model development. The top rated for user-friendliness of the user interface was Aequitas tool, which provided a web application with step-by-step guidance that is easy to understand and relatively easier to explain to non-technical stakeholders compared to the other toolkits' outputs.

[1] http://ethicstoolkit.ai/
[2] https://pair.withgoogle.com/explorables/
[3] https://www2.deloitte.com/de/de/pages/risk/solutions/ai-fairness-with-model-guardian.html

The group found the Google What-if tool and Fairlearn to be most useful for a technical expert with pre-existing deep knowledge of fairness, given they offer options for customisation of analyses that may be challenging to navigate for those lacking a background in fairness literature. Aequitas was selected as the most useful for non-technical beginners given the easy-to-follow web application interface. For a beginner data scientist/developer, scikit-fairness was voted as the most useful due to its ease of integration into existing models built in scikit-learn, a popular package used in ML.

4.4 Identified gaps in toolkits

There were several initial gaps identified from the group's exploratory assessment.

4.4.1 Lack of consistency in toolkits' methodological approaches. A general reaction of the focus group was surprise at the significant differences in approaches between the toolkits on defining and measuring fairness. For example, audit-ai focuses on statistical tests to calculate the likelihood that the outcome disparity, e.g. test scores between men and women, is due to random chance, which none of the other toolkits explicitly offer to the user. Fairlearn specifies: "Since we define fairness in terms of harms rather than specific causes (such as societal biases), we avoid the usage of the words bias or debiasing in describing the functionality of Fairlearn".[4] On the contrary, IBM Fairness 360 focuses on implementing a large variety of debiasing techniques on a group level for classification problems: "AIF360 recommends the earliest mediation category in the pipeline that the user has permission to apply because it gives the most flexibility and opportunity to correct bias as much as possible".[5] Aequitas tool also defines fairness with respect to biases, but its main purpose is not mitigation but to produce an audit report with the intention of flagging potential unfair outcomes. Its target bias types are: "1) Biased actions or interventions that are not allocated in a way that's representative of the population; and 2) Biased outcomes through actions or interventions that are a result of your system being wrong about certain groups of people".[6] While Aequitas offers a step-by-step process to select a group-level fairness metric to produce an audit report, Google What-if tool produces interactive visualisations of individual-level explanations and inferences for the user to explore independently, rather than calculate group-level statistics.

The focus group *observed that such differences in framing reflect the toolkit designer's perspective on fairness*, which are difficult to compare without prior understanding of the different positions taken. Practitioners without a thorough understanding of fairness debates are unlikely to decipher which toolkit aligns with their goals and the significance of the design choices on their scenario.

4.4.2 Target audience. The group found that the tools assume and require a high level of expertise in both fairness literature and data science/statistics. Whereas the group was familiar with both areas, many practitioners in the broader industry would have limited background in fairness. Even for the experts in the focus group, there was a general consensus that there is insufficient guidance on the criteria under which fairness metric should be used for each use case. The visualisations were felt generally difficult to understand for non-technical users, even for tools (e.g. Google What-if tool) that claims to have a wider target audience beyond technical developers. For the tools with fairness mitigation techniques, there is limited guidance on which methods should be used and when, leaving the investigation and comparison of the academic papers introducing these methods to each practitioner.

4.4.3 Limited consideration of real-life circumstances. The focus group found a gap between real-life considerations and the academic use cases on which many of these tools are built. One focus group member noted that practitioners need examples closer to real-life use cases in which defining fairness is more complex. Indeed, IBM Fairness 360 warns, "The toolkit should only be used in a very limited setting: allocation or risk assessment problems with well-defined protected attributes in which one would like to have some sort of statistical or mathematical notion of sameness. Even then, the code and collateral contained in AIF360 is only a starting point to a broader discussion".[7] Protected or sensitive attributes refers to the feature on which the user would like to assess the fairness metric, e.g. by gender or race.

In addition, the group noted that all products require protected features as input, when in reality, practitioners may not have access to individual-level demographics. For example, to test for demographic disparity between white and black customers, information on each person's race is required, which may not be collected by the organisation. The tools' focus on simple problems with binary classification in the examples and demos but comparatively limited focus on regression challenges with multiple, potentially interacting protected features that could, for example, result in intersectional discrimination.

4.5 Identification of next steps

The focus group derived the list of toolkits forming the basis for further investigation, in a manner akin to a general practitioner's search process. We next independently investigated these to compile their comparative features and capabilities (§5). The interview, discussed in §6, was divided into questions about the interviewee, importance of a list of features in their ideal fairness toolkit, and their assessment of whether each toolkit addresses each feature. These features were informed by the focus group and investigation, and divided into customisation, functionality, and user-friendliness, highlighting the key themes discussed in §4.4.1 (functionality), §4.4.2 (user-friendliness), and §4.4.3 (customisation).

5 FEATURE COMPARISON

For the toolkits identified in the focus group, we conducted an assessment of each toolkit and compiled a feature comparison table, displayed in Figure 1. This exercise is to provide a systematic comparison across a range of toolkits, both to help give detail of the different considerations relevant to fairness tools and to assist practitioners to identify the toolkit that meets their criteria. For future iterations of fairness tooling, the key themes on the aspects and features that are important to the practitioner will guide their

[4] https://fairlearn.github.io/user_guide/fairness_in_machine_learning.html#fairness-of-ai-systems
[5] http://aif360.mybluemix.net/resources#guidance
[6] http://www.datasciencepublicpolicy.org/projects/aequitas/
[7] http://aif360.mybluemix.net/resources#guidance

evolution. The differences in the toolkits' approaches will be discussed to point out the potential issues and considerations from a practitioner's point of view, which is later validated in the interview and survey. This section will report on the key feature differences among the six toolkits, informed by the focus group discussions on what criteria a practitioner seeks in a fairness toolkit. This exercise forms the basis for the interviews and surveys, which will validate the importance of these features.

We first studied each toolkit to gather relevant information about its functionality. Figure 1 contains the list of toolkits and the types of models covered: regression problems, classification problems (binary only or multi-class), and/or problems with multi-class protected features. A subset of the toolkits handle regression (predicting a continuous variable, e.g. income) as well as classification (predicting a discrete variable, e.g. loan approved or denied). Some toolkits can only handle binary protected/sensitive features (e.g. male vs. female), while others support multi-class features (e.g. age or racial groups). As will be discussed in the next section, practitioners search for tools that are compatible with their model, and if working on a regression problem, two of these toolkits can be ruled out immediately. Figure 1 also contains the fairness metrics and mitigation techniques supported by the tool. The most comprehensive of them is IBM Fairness 360 with more than 70 metrics, although its focus is on binary classification problems with some multi-class classification support and no support for regression.

We observed that one potential point of confusion is the differences in terminologies and definitions for the same metric. For example, equal opportunity difference is synonymous with false negative rate difference, and equal odds tests for both false positive and false negative rate disparities [39].

Most of these tools are also focused on group-level fairness metrics, while only Google What-if tool has a focus on individual-level fairness. IBM Fairness 360 supports some individual fairness metrics (sample distortion).

The variety of fairness metrics renders it especially challenging for the user to know what metric is appropriate for each use case. The toolkits have different approaches to guiding users on which metrics is appropriate for any use case. This theme will be further explored in the analysis of interview and survey results.

Most of these packages are built for integration with python, with one tool (IBM Fairness 360) with R support. All of them except Google What-if tool is built to allow for analysis on-premise, i.e. without uploading data into an external environment. What-if tool requires data upload, and its website specifies:

> "WIT [What-if tool] uses pre-trained models and runs entirely in the browser. We don't store, collect or share datasets loaded into the What-if tool. If using the tool inside TensorBoard, then access to that TensorBoard instance can be controlled through the authorized_groups command-line flag to TensorBoard. Anyone with access to the TensorBoard instance will be able to see data from the datasets that the instance has permissions to load from disk. If using WIT inside of colab, access to the data is controlled by the colab kernel, outside the scope of WIT [44]."

Similarly, Aequitas tool, while it has a desktop version available, also has a web-based application through which a user can upload a data set, with the caption:

> "Data you upload is used to generate the audit report. While the data is deleted, we host the audit report in perpetuity. If your data is private and sensitive, we encourage you to use the desktop version of the audit tool [33]"

The open source licenses in each of the toolkits' Github repository are either MIT or Apache 2.0 with the exception of Aequitas Tool, which has customised its own license. Aequitas Tool license appears broadly permissive, and the key restriction being that the copyright notice must be included in any future adaptations of the code, and UChicago accepts no liability and provides the code without warranty.[8] All of these tool contributors are based primarily in the United States, with one academic organisation and four private entities. Only scikit-fairness is built completely through the open source platform without any corporate sponsorship or involvement. The release date of the toolkits are in 2018 except scikit-fairness (2019).

These feature comparisons that we conducted and compiled were made available to the interviewees along with a select number of screenshots, standardised to include: metric calculation, guidance of metric selection, and visualisation. The interviewees found the feature comparison useful, as it gives a summary of relevant information they would otherwise be searching for re each tool.

6 INTERVIEW AND SURVEY FINDINGS

This section presents the key findings regarding the gaps in existing toolkits with regards to the needs of practitioners as revealed by the interview and the survey. As the purpose of the survey was to validate the findings in the interviews with a wider audience, we will present the insight from the interview, followed by any supporting evidence from the survey.

6.1 Interview method

We conducted semi-structured practitioner interviews, with questions constructed from the themes that emerged in the focus group. The interviews consisted of three parts: 1) demographics, 2) features of an ideal fairness toolkit, and 3) review of existing toolkits.

The first part of the interview collected information on the interviewee's role, domain, and technology area and asked about a product in which a fairness testing toolkit would have been useful. The second part asked the interviewee to rate from "Extremely important" to "Not at all important" features that were identified in the focus group, i.e. relating to customisation, functionality, and user-friendliness. The toolkit evaluation (part 3) involved the same features the interviewees rated in an ideal toolkit with the addition of the System Usability Scale survey, a well-defined and tested metric for usability [3]. The lead researcher then walked through the key features, metric calculation, guidance, and visualisation of each of the six toolkits (included in supplementary materials) providing links for the interviewee to explore. The screenshots taken in feature comparison (§5) were used for the walk-through, with the template included in the supplementary materials. Any questions

[8] For full license wording, see: https://github.com/dssg/aequitas/blob/master/LICENSE

						Models covered					Group fairness							Individual			
Tool	Setup	Open source user license	Release date	Organization	Open for anyone to contribute code?	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	Other fairness metrics	Bias mitigation
Scikit-fairness / scikit-lego	python (sklearn)	MIT	2019-03-31	N/A	✓	✓	✓	X	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, Reweighing, 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	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

Figure 1: Open source toolkit feature summary table

by participants that required the interviewer to voice an opinion (e.g. "what do you think of this functionality") was deflected with a response that "the purpose of this interview is to get your perspective, and I am happy to share the results and my opinion after the research project is complete." The collection of responses was facilitated by Qualtrics survey software,[9] which the interviewee used to fill out their ratings, but the interviewees were encouraged to verbally voice any thoughts and feedback throughout the process.

Interviewees were recruited through direct contact and snowball (or chain-referral) sampling, a method often used in qualitative sociological research in which a population is difficult to reach: this method starts with a convenience sample of initial subject and recruiting their networks in subsequent sampling [16]. The potential biases in this sampling process is discussed in §7. We emailed contacts with industry experience we knew to be interested in ML fairness and asked them to connect us to any colleagues or acquaintances working on algorithms with potential fairness concerns. Because the interview required practitioners to view and critique snapshots from existing toolkits, we specifically recruited those already familiar with the terminologies, concepts, and academic debates around fairness. Snowball sampling was selected because of the scarcity of industry practitioners who have had prior hands-on experience with fairness challenges in their algorithms. Because we asked interviewees for potentially commercially sensitive and confidential information about their company's internal products, each interviewee was given a consent form outlining the purpose of this project and assuring them that no identifying information about the individual or the company will be shared.

We had 15 interviewees, each from a different company. As per the consent form, the respondents and their companies will not be identified or linked to any demographic information to preserve their confidentiality, but we will refer to each of them by their ID (i.e. I1-I15). The interviewees worked in technology (4 - I3, I5, I6, I8), professional services (2 - I1, I12), retail banking (2 - I4, I15), insurance (2 - I10, I13), financial technology (1 - I2), marketing (1 - I14), media (1 - I7), public sector (1 - I11), and academia (1 - I9). All interviewees were based in the U.K. except one from Canada and one from the U.S. All interviewees completed the structured questionnaire except one (I8), in which we were out of time after covering 4 of the 6 toolkits. Fairlearn and audit-ai ratings therefore have n = 14.

Permission was obtained to record the interviews for transcription (with identifications removed), with the recording subsequently deleted. Interviews lasted between 60 and 90 minutes and conducted over a video conference call application. We asked whether they have faced fairness-related challenges in their past products or models and proceeded when they confirmed this as a confirmation of their eligibility for the interview. We transcribed and tagged key themes in the interviews through affinity diagramming, generating codes for topics and grouping them into themes.

Given the time constraints of the interview (60-90min), it was impractical to get each interviewee to comprehensively assess each toolkit by reading all relevant documentation and trying it out on a different data set. Therefore, their views may be limited by what

[9] https://www.qualtrics.com

was presented as snapshots of the tools. A few interviewees (I1, I3, I4, I10, and I14) had prior experience with a subset of the toolkits, which allowed them to comment more extensively on their features. Despite this, we saw that the interviewees were able to provide valuable and consistent insights into the importance of each feature in their ideal toolkit and how well each tool appears to meet their needs.

6.2 Survey method

To validate our findings from the interviews with a larger sample size and a broader practitioner population, we conducted an anonymous online survey following the similar structure to the interview, using Qualtrics survey software. The survey did not ask for any identifying information, e.g. name, company, or contact details.

We emailed the survey link to direct contacts, as well as advertising it on online communities, e.g. meet-up, Facebook, and LinkedIn groups, related to data science and analytics. We also encouraged sharing of the survey link to anyone working in data science and analytics. Of the 71 people who started the survey, 41 (57.7%) of the respondents completed at least one section and 26 (36.6%) completed the entire survey. Additional demographics and summary statistics are in the supplementary materials. While we were looking to widen our reach to practitioners with varying levels of knowledge and experience in fairness, the questions, especially those on what features would be important to test a model for fairness, could be difficult to contextualise if the prospective respondent has no background nor reference point regarding fairness-related challenges. This might have contributed to the drop-out rate and limited the potential sample size, which indicates that fairness is an area of interest to many practitioners but for which few have relevant expertise. With the rising scrutiny on fairness issues in industry, as this challenge becomes pervasive across industries and generalist (non-specialist) data scientists, we note that designing a toolkit for an average practitioner will become increasingly important.

We structured the survey into the following sections: 1) importance of various features in a fairness toolkit, with questions around the respondent's requirements and preferences, 2) feedback on fairness toolkit visualisations, 3) feedback on a fairness toolkit guidance, 4) familiarity with fairness literature, and 5) demographic information. In Section 1, we asked for ratings on various features from "Extremely important" to "Not at all important", with probing questions on trends identified in the interviews. Next, using the Qualtrics randomiser, we randomly selected and presented two visualisations from the toolkits from screenshots taken for the interview template (included in supplementary materials) and asked for feedback. We then randomly selected one snapshot of the guidance documentation (also from the interview template) from the toolkits to ask for feedback, as well as asking the respondents to rank features of guidance documentation in the order of importance, also derived from the interviews. This randomisation was in order to keep the survey length manageable and reduce drop-off rates. Given the survey was for qualitative expansion on themes previously explored, we believed this would add depth without hindering sample size. We asked the user a number of questions to assess their familiarity with algorithmic and ML fairness literature and ended the survey with the demographic information collection, including a question about any specific past challenges related to fairness in their work. All questions were optional.

While the aim of the survey was not a comprehensive comparison of the toolkits, it validated some of the insights in the interview, especially the relative importance of each feature and any "deal-breakers" or properties that would make a toolkit ineligible for use by the practitioner. The free-text answers, especially on the helpfulness of a randomly selected subgroup of toolkit screenshots brought additional perspectives, given the survey respondents often had less prior experience in fairness-related model development challenges than the interviewees. The summary statistics aggregated for all respondents are included in the supplementary materials.

6.3 Key findings

Four crucial gaps emerged from in both the interview and survey:

(1) the steep learning curve to use the toolkits,
(2) lack of a tailored user experience that avoids both information overload and oversimplification,
(3) limited coverage of fairness considerations in the end-to-end model development lifecycle (i.e. beyond the model build and testing stage), and
(4) limited ability to adapt and integrate the tools to "plug and play" with the existing workflow.

The first two are considerations on user-friendliness, echoing the initial findings from the exploratory focus group in *§4.4.2 Target audience*. The third discusses functionality, validating findings from *§4.4.1 Lack of consistency in toolkits' methodological approaches*. The fourth gap addresses contextualisation and customisation, which was raised in *§4.4.3 Limited consideration of real-life circumstances*. We will now elaborate on each of these gaps in turn.

6.4 User-friendliness

6.4.1 Steep learning curve required to use the toolkits and limited guidance on metric selection. One of the recurring themes in the focus group and interviews was the difficulty in understanding the fairness considerations for practitioners without prior background on the topic. 40% of interviewees rated the "Guidance for users unfamiliar with fairness academic literature" as "Extremely important" with another 13% rating it as "Very important."

Many interviewees commented on the complexity of the fairness-related challenges. I1 claimed to have taken "two months if not more" to learn about fairness testing, noting that one would have to do "at least weeks of reading." I11 said that "the choice of fairness measures as well as their corresponding trade-offs will depend on the context. In certain areas, these will be less clear and even less obvious to practitioners."

In contrast with the perceived importance for guidance, the guidance for users unfamiliar with fairness literature rated an average of 2.87-3.67 out of 5. I3 reviewed the toolkits as not having sufficiently "easy explanations about the concepts and metrics [that are] not in academic language and style." I1, I5, and I6 alluded to the possibility of "fairness gerrymandering"[5, 23] with practitioners selecting the metric based on which metric the model is able to meet among the metrics available. I3 also noted that there is no way to understand if a functionality or metric is "widely accepted."

Table 1 contains the average System Usability Scale (SUS) score out of 100 and its standard deviation. SUS provides a standardised measurement to compare the toolkits to supplement the topic-specific questions, as the toolkits aim at both developers and higher-level practitioners (§4.4.2) and can inform non-technical stakeholders. A systematic survey of the SUS found "products which are at least passable have SUS scores above 70, with better products scoring in the high 70s to upper 80s" [3], scores we see are beyond that for the fairness toolkits. In the SUS survey, almost half of the interviewees agreed or strongly agreed with: "I needed to learn a lot of things before I could get going with this [fairness toolkit] system." Given the interviewees were specifically sampled such that they were very familiar with algorithmic fairness issues, this may be underestimating the learning curve required for the wider practitioner population.

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
All	*60.43*	*14.84*

Table 1: Toolkit System Usability Survey Scores

Of the survey respondents, 16% classified themselves as "extremely familiar" with the current academic debates in algorithmic fairness, with 32% "very familiar" and 52% "somewhat familiar." No one responded that they were "not at all familiar" with the fairness literature. For those who responded to the free-text field on whether a randomly selected toolkit guidance layout was helpful, they described it as "quite dense," "too much text," and "might be better broken down in a Q&A format."

One guidance material that had an overall strong positive feedback was the Aequitas Tool (Figure 2), which contains a decision tree to assist a user in selecting a metric. However, several commented that while the structure was strong, the wording was difficult to understand with no associated definitions or guidance. Some also expressed concern it oversimplified the criteria for selecting a metric. Google What-if tool's associated blog post that features six experts arguing and disagreeing over fairness definitions was also commended for its colloquialism and its easy-to-understand representation of the conflicts between the fairness definitions. Fairlearn's dashboard was also seen as easy to follow due to its step-by-step walkthrough of the fairness testing process.

6.4.2 Information overload vs. over-simplification of complex results. In reviewing the visualisation and guidance, interviewees and survey respondents were often split on their preference on the level of detail provided. A survey respondent notes that "good information design carries people well through the anxiety of overwhelming data." Some respondents found the amount of information provided prohibitively complex, calling it "quite dense" and "a bit of overload." Several commented that the guidance was too long: "for hands-on technical people who are picking this up on a whim and wanting to use it quickly, they want half the text." For What-if tool in particular, the number of options on the screen was overwhelming for some interviewees. One survey respondent said an ideal toolkit should

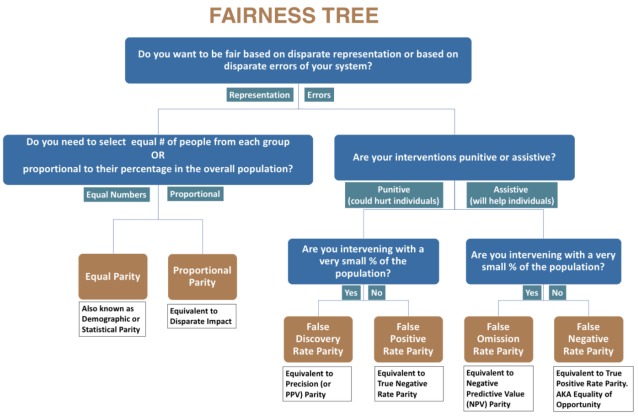

Figure 2: Aequitas guidance

have "an easy and intuitive user interface with transparent and clear back-end code." Another said, "reducing complexity would lead to higher transparency, and that's crucial."

By contrast, others had a strong preference in favour of the detailed interface. One survey respondent said of Fairlearn dashboard that "this makes everything look clear-cut, which it really isn't 'in the wild.' " Another interviewee wanted more detail in the guidance for Fairness 360, saying "I feel confident technically using the system, but am I confident in a sense of trusting what it tells me? I don't think so" (I3). There was resistance to the idea of over-simplifying fairness, which many saw as a complex concept.

Given there may be differences in the level of detail each user requires for his or her purpose, an ideal toolkit should have both (i) a number of options on the user interface that allows the user to deep-dive and slice and dice the analyses, and (ii) an easy-to-use interface that guides the user step-by-step. The former was rated at 3.81/5 and at least "Very important" by 69.2% of the survey respondents; the latter had a lower average importance score (3.15/5) but still at least "Very important" by 42.3% of the survey respondents. The interviewees rated a "well-structured user interface" as an average importance of 3.33 on a scale of 5.

6.4.3 Need for "translation" for a non-technical audience: As well as being challenging for data science practitioners with no fairness background, the toolkits were overwhelmingly rated as challenging for a non-technical user, especially in producing visualisations, guidance, and user interface that can be navigated by those without a background in math, statistics, and computer science.

It was reported that "for all toolkits, apart from Aequitas, you need somebody technical to run the analysis" and then "translate the findings" to the non-technical stakeholders (I5). Speaking of the various visualisations, an interviewee said "there's no way a non-technical person could understand this." This results in a gap between the analysis done by the practitioners and what can be understood by the business function. While a toolkit's main target audience may be the technical developer, over half of the survey respondents 80.8% rated the interpretability of results and visualisations to a non-technical audience as either "Very" or "Extremely important." In rating the visualisations, the survey respondents said they are "likely to be problematic," "more difficult," and "impossible"

to understand for a non-technical audience. One survey respondent said the guidance text emphasising the mathematical definitions was "only useful because I have a background in statistics."

6.4.4 Accessibility of toolkit search process. As suggested in the focus group, we find that almost all interviewees (14 out of 15) claim to "use a search engine and iterate through the results until one that meets their criteria are found, and no further search is conducted." Only one interviewee reported to "comprehensively search for all available toolkits to compare the strengths and weaknesses before selecting the optimal tool." Two interviewees would additionally ask colleagues for advice and search for other work that encountered similar issues. This finding serves to highlight an additional contribution of our work in comparing the features of the six toolkits. The feature comparison chart aims to provide the practitioners with sufficient information about some of the prominent fairness toolkits, as selected through a similar search and discovery process, in order to help them identify the one they need for their use case.

All survey respondents were asked whether they had used any of the six toolkits (multiple selection allowed) and whether there were any other toolkits they were familiar with that were not listed. Only one respondent said they knew of another toolkit: FAT Forensics [34], released in late 2019 resulting from a collaboration between the Uni of Bristol and Thales. Overall, that there were no other toolkits the practitioners were familiar with suggests that our landscape coverage was sufficiently representative for exploring the issues.

6.5 Toolkit features

6.5.1 Limited coverage of the model pipeline. Echoing the findings in [18], the interviewees emphasised the apparent focus of the toolkits on the model building and evaluation process as compared with the remaining model lifecycle. According to I1, "Each section of the model building pipeline is important – testing your training data, representation, model output, proxy variables, etc... no tool has an end-to-end 'this is what is going on in your system.' " In the survey, 75% respondents answered the "coverage of different model build pipeline" to be at least "Very important." A survey respondent specifically pointed this out as a limitation of toolkits they previously used, saying "most of the toolkits tend to straddle both [auditing / mitigation and data exploration] which presents challenges... so it is not as useful as a part of ML pipeline."

Some gaps specifically mentioned were checking whether the data set is representative of the broader population and whether there were features acting as proxies of protected features, e.g. postcode for race or occupation for gender. I2 claimed a major gap was in the lack of benchmarking data sets or a reference point for whether there is a selection bias in the data collection process. I10 suggested there should be a way to understand which input features are potentially acting as proxies for protected features, "especially when the feature engineering has been done by a human" (I10). A survey respondent also noted "the analysis needs to explore the idea of proxies, something we do manually today."

6.5.2 Limited information on possible mitigation strategies. There was a strongly mixed amount of enthusiasm for tools that offered "debiasing" pre-processing, in-processing, and post-processing implementations. Several interviewees (I1, I3, I5, I6, I10, I11, I13) were skeptical of these techniques. I3 claimed that these methods are "dangerous because it looks simple but doesn't solve any problem. It's like a gimmick, like training a constrained classifier, but it doesn't solve the underlying issues of bias you may have." I1 said it "doesn't solve the bias at the root." One interviewee (I10) claimed some of the bias mitigation tools could be inconsistent with anti-discrimination laws, especially any that explicitly use a protected feature (e.g. race) to give preferential treatment, and suggested that the mitigation strategy should depend on the context, which "may not always be a technical solution." On the other hand, several other interviewees (I2, I4, I8, I9, I12) viewed these implementations favourably. I12 noted that some tools' "lack of mitigating action leaves a huge knowledge gap for data scientists to fill."

6.6 Contextualisation

6.6.1 Limited adaptability of existing toolkits to a customised use case. The strongest consensus regarding the ideal fairness toolkit was the importance of the "ability to adapt to a context-specific use case and data," with all responses either 5/5 or 4/5 and an average of 4.7. Similarly, in the survey, all except one respondent rated this as "Extremely" or "Very important." The existing toolkits were rated on the same criteria as an average of 3.24 out of 5, with PyMetrics audit-ai scoring the lowest at 2.71 and IBM Fairness 360 the highest at 3.73, with several interviewees noting that additional work would be needed for the toolkits to be applicable to their use cases.

Because audit-ai was built tailored to the U.S. employment guidelines for internal use as an algorithmic hiring company, many found it to be inapplicable to their own use cases. For example, the "4/5ths rule," i.e. the guidance that the lowest-passing group has to be within 4/5ths of the pass rate of the highest-passing group, may not be an appropriate threshold in other domains. However, their unique approach of statistical testing to calculate the likelihood that the disparity is due to random chance was lauded by a few interviewees (I5, I6) and survey respondents as important, compared with other tools reporting the outcome disparity (e.g. re false positive rate) without confidence intervals or statistical significance.

IBM Fairness 360 was rated highly for having "a lot of useful code"; however, one interviewee (I9) who had extensively used the tool noted that a lot of their tool is "hard-coded to their data and their use case, so it was a matter of how much extra work is needed." I13 critiqued the tools for having relatively little focus on regression problems compared to simple binary regression problems. For his work in insurance pricing, 95% of his work involves regression problems with multi-class protected features, thus several of these toolkits are not applicable; referring back to Figure 1, only What-if tool and Fairlearn has coverage of these model types. This issue was also highlighted in the focus group (§4.4.3) as a major concern on the toolkit's adaptability.

6.6.2 Challenges in integrating the toolkit into an existing model pipeline. Another point of consensus was the importance of the ease of integration of a toolkit into the model building workflow and pipeline. This was rated as "Extremely important" for 60% of interviewees and "Very important" for the remaining 40%. Among the survey respondents, 85% rated it as at least "Very important."

However, the toolkits were rated an average of 3.24 in their ease of integration, with the lowest score at 2.47 for Google What-if tool and the highest score at 3.93 for Scikit-fairness. This is due to the significant differences in how they integrate with existing workflows. Google What-if tool's primary aim is to visualise the data such that the model developer could explore potential biases, separately to model development; scikit-fairness was built to embed fairness testing and mitigation directly into the model build.

As discussed briefly in §4.1, several interviewees criticised Google's What-if tool and the Aequitas web application for the requirement to upload the data, noting this would face challenges from their organisations on whether this is GDPR-compliant and in adherence to relevant privacy policies. This partially contributed to the low score of Google's What-if tool, especially given the visualisation required a setup in Tensorboard or Jupyter notebook. One survey respondent said any toolkit with any processing off-premise, even if the data set is not stored, "would need a very large amount of governance and security validation to be allowed to be used with corporate data." This sentiment was repeated for several survey respondents, with many listing any solution that is not completely on the local computer as a 'deal-breaker.'

Several interviewees (I1, I3, I4, I10, I11, I13, I15) claimed that having to upload their data sets, even if it is not stored, could immediate disqualify the tool for usage due to organisation's policies. Only one interviewee (I14) said that this was not an issue because the company has a pre-arranged partnership agreement with Google.

Scikit-fairness received a high score because most of the interviewees were python users; however, two interviewees (I1 and I11) commented that it is only easy to integrate into scikit-learn and gives no flexibility if working with any other package. I11 said for her organisation that often does not use python or R, many of these packages would require an integration layer to work with their existing ML models. However, this seems to be a limited issue among those surveyed. When asked "do you use tools that easily integrate with packages built in python and R," all responded yes.

7 DISCUSSION AND FUTURE DIRECTION

We studied the comparative features of six existing fairness toolkits and identified several key gaps in their capabilities in meeting practitioners' needs. While algorithmic fairness represents a high-profile discussion, and is increasingly an area of concern across industry, among the survey respondents, only 48% considered themselves at least "very familiar" with existing fairness literature. To test one's familiarity with issues of fairness, we asked which two fairness definitions are generally incompatible, and 44% selected the correct answer: equal odds and positive predictive parity, whose incompatibility was a well-cited example in the U.S. criminal recidivism scoring model and academically proven [24]. 36% gave the incorrect answer of equal opportunity and equal odds; equal opportunity is a subset of equal odds, meaning that if equal odds is satisfied, it implies equal opportunity is satisfied [14]. The remaining 20% responded they are not sure. Future work could explore an average practitioners' familiarity with fairness issues in a more representative sample.

The high drop-out rate in the survey (42.3%), i.e. those who start the survey but abandon it after reading the questions on fairness toolkits, also suggests that the prospective respondents may be interested in fairness considerations but do not have the relevant understanding of the topic. While the resulting low sample size limits the external validity of the findings and the ability to conduct more in-depth statistical tests, the key findings persist through the focus group, interviews, and surveys. The methodology, such as reporting of percentages, scales, with references to specific interviewees, is consistent with past human subject research on fairness [18, 38].

For the focus group and interview, practitioners with expertise in fairness were purposefully recruited and sampled; therefore, the results are only representative of those with pre-existing understanding of the typical fairness challenges. However, the fact that both these stages found gaps and limitations, especially in user-friendliness and interpretability of the toolkits and their guidance, suggests that the learning curve may actually be much steeper for an average practitioner with more limited exposure to fairness metrics. While the nature of these toolkits may evolve over time, our findings on practitioner needs and the high-level perceived gaps would provide important signposts for future development.

Again, our aim was not an exhaustive comparison of the entire range of fairness toolkits, but rather, we selected examples that were indicative of the landscape in order to elaborate general issues and concerns. Our survey showed no one except for one respondent had used a toolkit not on our list, suggesting that there were no glaring gaps in our landscape assessment. Given the method of uncovering toolkits was through search engines, an approach confirmed in our study as consistent with what occurs in practice, these six are those that practitioners are likely to come across in their search processes. Therefore, even if there is a toolkit that closes some of these gaps, its limited awareness and accessibility is still an obstacle to its adoption. More generally, however, many of the issues are broad, general, and some concerns context-specific, meaning many of the findings will remain relevant even in-spite of improved toolkit iterations.

It was clear that a user interface with a one-size-fits-all tailoring toward practitioners with prior understanding of fairness limits the accessibility of these toolkits. Different users have varying preferences and needs from their interface. A key example of this is the high standard deviation in the survey ranking of the importance of mathematical definitions in a toolkit guidance (mean: 4.04/8, standard deviation: 2.79) and the ranking of the importance of visualisations that are helpful for a non-technical audience (mean: 4.48/8, standard deviation: 2.57). As flagged in the interview, some practitioners with a background in statistics may want a detailed mathematical definition, while those looking for a quick proof-of-concept may want a simple user interface for business stakeholders' review. To validate this, there was no strong consensus in the ranking of usefulness in guidance features (average ranking from 4.04 to 5.12) except in two cases: the importance of the explanation of the intuition behind a definition ($\bar{X}$: 2.77, S_X: 1.82) and the non-importance of the relevant legal context ($\bar{X}$: 6.62, S_x: 1.81). One survey respondent explained that the legal context is "better dealt with by a more appropriate (not data) person." Future work could explore how the roles and responsibilities of relevant stakeholders in business (e.g. risk practitioners, lawyers, and business product owner) may be able to interpret and provide relevant context for the toolkit's application. Future work could deep-dive on specific

human-computer interaction considerations, e.g. API usability studies [1, 30, 47] considering developers' perspective may be relevant in assessing the specific technical strengths and limitations of some of the toolkits. The guidance and design of the tool, along with its functionality, could affect the user's interpretation of toolkit outputs, potentially raising the risk that the user could be misled with over-simplified explanations to be overconfident in a model's fairness or confounded by its complexity and pushed to abandon the toolkit.

It is also important to consider whether the toolkits with necessarily reductionist definitions of fairness are appropriate and beneficial from a societal standpoint. Several academics have objected to the "automation" of fairness assessments because these tools fail to consider the socio-technical system, the nuanced philosophical and ethical debates, and the legal context of what it means to be fair [28, 42]. For IBM Fairness 360, in answering whether the tool should be used at all, the guidance warns that the tool applies to limited settings and is intended as a starting point for wider discussion [4]. The practitioners in the interview and survey were generally positive in their reaction to the notion of a fairness toolkit to help navigate an extremely complex issue, but several expressed concern for "fairness gerrymandering," (or "ethics washing") [5, 23] or selecting the metric based on which ones were satisfied, and for the false confidence the toolkits may give to the model developer based on an incomplete or partial assessment of fairness. Future work could examine in-depth the disclaimers and limitations described for each of the toolkits and whether they align to the academic understanding of suitability of each implemented method.

8 CONCLUSION

Fairness toolkits are a fairly recent phenomenon, particularly in the last two years, and several interviewees were surprised to learn about their availability and diversity. Only 54% survey respondents had used any open source fairness toolkit before, despite our sampling of groups with likely exposure to fairness-related concerns. With the growing attention on issues of fairness, it is important that any fairness toolkits are accessible, usable and fit for purpose.

Our paper contributes a gap analysis and the associated findings regarding practitioner needs and the features of available open source fairness toolkits. With a focus group, semi-structured interviews, and surveys, we identified key themes of practitioner requirements that require more attention. This analysis can help inform future tool development in order to bridge the gap between the introduction of methodologies in academia and their applicability in real-life industry contexts.

We also provide a feature summary with relevant characteristics of each of the six selected toolkits, which can help facilitate a practitioners' toolkit search and evaluation. We have found that the toolkits are diverse in their approaches and do not simply reflect different implementations of the same fairness methodologies. Given that (as our results indicate) many practitioners look for a tool until they find one that meets their needs, a comparative review of the toolkits would help practitioners understand the toolkits' offerings and aid their selection process. We will therefore share this table of features (Fig. 1) on GitHub, along with the other workstream outcomes of the focus group (Ethics DataDive) hosted by DataKind UK, in order to allow for others to comment on and update what is available in the open source landscape. It is our hope that this will become a reference point and a repository of information on the practitioners' guide to various ethics toolkits.

Our results suggest that industry practitioners are still struggling with finding a way to identify and mitigate potential unfairness in their models and systems. Only by keeping close to the practitioners' requirements and preferences can the open source developers ensure widespread adoption of their toolkits. 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. Future development of toolkits should remain vigilant to ensure their adoption is aligned to the over-arching goal: to ensure our algorithms reflect our ethical values of non-discrimination and fairness.

ACKNOWLEDGMENTS

We would like to thank all the industry practitioners who gave their time to participate in this research project. Special thanks to DataKind UK and the Ethics Committee for hosting and helping organise the virtual Ethics DataDive. The Compliant & Accountable Systems Group acknowledges the financial support of the UK Engineering & Physical Sciences Research Council, Aviva and Microsoft through the Microsoft Cloud Computing Research Centre.

REFERENCES

[1] Yasemin Acar, Michael Backes, Sascha Fahl, Simson Garfinkel, Doowon Kim, Michelle L Mazurek, and Christian Stransky. 2017. Comparing the usability of cryptographic APIs. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, IEEE, 154–171.

[2] Saleema Amershi, Max Chickering, Steven M Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 337–346.

[3] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.

[4] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, A Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[5] Elettra Bietti. 2020. From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 210–219.

[6] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*. 3992–4001.

[7] Nan-Chen Chen, Jina Suh, Johan Verwey, Gonzalo Ramos, Steven Drucker, and Patrice Simard. 2018. AnchorViz: Facilitating classifier error discovery through interactive semantic data exploration. In *23rd International Conference on Intelligent User Interfaces*. 269–280.

[8] Aisling Ni Chonaire and Jannna Ter Meer. 2020. The perception of fairness of algorithms and proxy information in financial services: A report for the Centre for Data Ethics and Innovation. *The Behavioural Insights Team* (2020).

[9] Kate Crawford. 2017. Artificial intelligence with very real biases. *The Wall Street Journal. Retrieved from https://www. wsj. com/articles/artificial-intelligencewith-very-real-biases-1508252717* (2017).

[10] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, October 2018.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in*

theoretical computer science conference. ACM, 214–226.

[12] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.

[13] Avi Feller, Emma Pierson, Sam Corbett-Davies, and Sharad Goel. 2016. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *The Washington Post* (2016).

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

[15] Rasmus Hauch. 2020. Denmark introduces mandatory legislation for AI and data ethics. *2021.ai* (2020). https://2021.ai/denmark-introduces-mandatory-legislation-ai-data-ethics/

[16] Douglas D Heckathorn. 2011. Comment: Snowball versus respondent-driven sampling. *Sociological methodology* 41, 1 (2011), 355–366.

[17] Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. 2018. A moral framework for understanding of fair ml through economic models of equality of opportunity. *arXiv preprint arXiv:1809.03400* (2018).

[18] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.

[19] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un) fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 49–58.

[20] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[21] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*. IEEE, 924–929.

[22] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[23] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*. PMLR, 2564–2572.

[24] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[25] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[26] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. *arXiv e-prints*, Article arXiv:1703.06856 (March 2017), arXiv:1703.06856 pages. arXiv:1703.06856 [stat.ML]

[27] Min Kyung Lee and Su Baykal. 2017. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1035–1048.

[28] Michelle Seng Ah Lee, Luciano Floridi, and Jatinder Singh. 2020. From fairness metrics to key ethics indicators (KEIs): a context-aware approach to algorithmic ethics in an unequal society. *Available on SSRN* (2020). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3679975

[29] Microsoft and contributors. 2019. Fairlearn. https://fairlearn.github.io/

[30] Brad A Myers and Jeffrey Stylos. 2016. Improving API usability. *Commun. ACM* 59, 6 (2016), 62–69.

[31] Arvind Narayanan. 2018. *Tutorial: 21 Definitions of Fairness and their Politics*. YouTube. https://www.youtube.com/watch?v=jIXIuYdnyyk

[32] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.

[33] 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. *arXiv preprint arXiv:1811.05577* (2018).

[34] Kacper Sokol, Raul Santos-Rodriguez, and Peter Flach. 2019. FAT Forensics: A Python toolbox for algorithmic fairness, accountability and transparency. *arXiv preprint arXiv:1909.05167* (2019).

[35] Aaron Springer, Jean Garcia-Gathright, and Henriette Cramer. 2018. Assessing and Addressing Algorithmic Bias-But Before We Get There.... In *AAAI Spring Symposia*.

[36] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2459–2468.

[37] Roger Taylor. 2020. AI Barometer Report. *Centre for Data Ethics and Innovation* (2020). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/894170/CDEI_AI_Barometer.pdf

[38] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[39] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.

[40] Neil Vigdor. 2019. Apple card investigated after gender discrimination complaints. *The New York Times* (2019).

[41] Matthijs Vincent and ManyOthers. 2019. scikit-fairness. https://github.com/koaning/scikit-fairness

[42] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2020. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Available at SSRN* (2020).

[43] Stephen Walli, Dave Gynn, and Bruno von Rotz. 2005. The growth of open source software in organizations. *Publication Report. Optaros Inc* (2005).

[44] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.

[45] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[46] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.

[47] Minhaz Zibran. 2008. What makes APIs difficult to use. *International Journal of Computer Science and Network Security (IJCSNS)* 8, 4 (2008), 255–261.