Closer or even farther from fairness: An assessment of whether fairness toolkits constrain practitioners with regards to algorithmic harms

Ana Maria Vasilcoiu
Supervisor: Agathe Balayn
Responsible professors: Jie Yang, Ujwal Gadiraju
EEMCS, Delft University of Technology, The Netherlands

June 19, 2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

To encourage ethical thinking in Machine Learning (ML) development, fairness researchers have created tools to assess and mitigate unfair outcomes. However, despite their efforts, algorithmic harms go beyond what the toolkits currently allow to measure. Through 30 semi-structured interviews, we investigated whether data scientists are constrained to only thinking about issues that can be tackled with these toolkits when using them in practice. The results of a comparative assessment of approaches with and without a toolkit indicate that although they can be incredibly effective, toolkits shouldn't replace educating on sources of harm and can even have hazardous consequences when improperly used. We discovered that while fairness toolkits increase practitioners' awareness of several specific sources of harm, such as questionable attributes or data sampling techniques, their greater power lies in fostering discussions about ML systems' propensity to treat individuals unfairly. On the contrary, we observed that these toolkits do not significantly help in the data documentation process, and, from observing our study participants, we also infer a risk of them blindly evaluating and optimizing for undesired outcomes as a result of choosing metrics and mitigations on unfounded or incomplete assumptions. This work supports future improvement of toolkits by providing a breakdown of perspectives around various sources of harm and reasoning about the ones that get frequently overlooked.

1 Introduction

In nowadays society, automated prediction-based decision-making tools are increasingly employed in domains such as banking and finance [1], hiring [2] and online advertising [3], and they are featured in high-stakes decisions such as in criminal sentencing [4] and medicine [5]. However, this myriad of new opportunities introduced numerous challenges, from which one of the most notable is the potential for Machine Learning (ML) systems to treat people unfairly. In recent years, there has been a growing concern around this idea that these models might result in unjustifiably different outcomes along social axes, such as gender, race or status [6].

A canonical example is the COMPAS system, a tool employed to predict recidivism risk and guide pretrial detention and release decisions. An independent journalistic inquiry found that the system perpetuated unjustifiable socioeconomic disparities, including racism against African-Americans, in the form of them having an almost double rate of being falsely labeled as future criminals when compared to the white defendants [4]. COMPAS becomes thus the archetype of social injustice in decision support systems, and one of the many examples that have granted the emergence of a new field of research, namely that of algorithmic fairness.

Although the essence of fairness can be captured as being "the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics" [7], the research literature identifies numerous different definitions for fairness, which in many situations are impossible to simultaneously satisfy [8]. This makes it a multi-faceted, highly context dependent and complex issue for which no clear, one-size-fits-all solutions exist.

Since fairness cannot be "solved", computer scientists have shown by now substantial research efforts towards the creation of metrics to measure unfairness in the outputs of ML systems, as well as algorithmic methods to assess and mitigate unfairness when detected [7] [8] [9] [10]. As the field advances, integrated toolkits are being developed for the purpose of empowering data scientists and developers with little to no experience to employ fairness considerations into the systems they are building. Worth mentioning examples of such toolkits include FairLearn [11], Google's What-if tool [12], IBM's AI Fairness 360 [13] and UChicago's Aequitas tool [14].

However, although a variety of other such toolkits already exist and new ones are constantly being developed [15], literature identifies considerable gaps between their capabilities and practitioners' needs [14], such as limited regard for practical scenarios or insufficient guidance for choosing the appropriate metrics and mitigations. The main reason for this lies in the fact that algorithmic harms largely go beyond what fairness toolkits currently allow to measure, for example biases that result from inadequate choices in model building or harms introduced by the nature, relevance and representation of data attributes. This makes it uncertain to what extent these tools actually contribute to building fairer automated systems.

In this paper, we extend on previous work trying to understand whether the toolkits being restricted to addressing only fairness related issues can negatively impact practitioners' frame of consideration of other algorithmic harms. In the current technical landscape characterized by a wide adoption of these toolkits, this is especially relevant since if this is indeed true, a significant amount of harms would go unattended with potentially hazardous consequences for the broad environment in which ML systems are being deployed. To address these concerns, we conducted 30 interviews with machine learning practitioners from different positions in the fairness expertise spectrum and found that, while toolkits lack support in several key areas, most issues are caused by a lack of educational guidance. Amongst other findings, we discovered that although toolkits have in-built support for protecting sensitive features and applying mitigation techniques, they can influence practitioners in a negative way not only because of a lack of support for interdisciplinary communication but also through insufficiently encouraging users to acquire an in-depth understanding of the underlying issues before pursuing technical solutions. This research goes further previous efforts to make fairness assessment tools usable and accessible, investigating the conceptual limitations of their use in practice and ultimately attempting to reduce bias in machine learning development.

In the remaining of this paper, we begin by reviewing the existing literature not only on the most prevalent ML harms, but also on relevant findings in the area of Fair Machine Learning in Section 2. In Section 3 we provide an in-depth description of the methodology employed, while an evaluation of the main results can be found in Section 4. We reflect on the ethical considerations of this research in Section 5. Section 6 positions the results in a broader context, discussing both limitations and areas of future work, followed by main conclusions in Section 7.

2 Background and Related Work

This section aims to not only provide the reader with an understanding around notions of harm and fairness, but to also summarize existing work on practitioners' judgement and interaction with fairness toolkits.

2.1 Algorithmic harms

To ground the discussion about fairness and the emerging technical solutions for the potential negative consequences of automated decision-making systems, it is of paramount importance to first understand the sources of harm in the ML lifecycle and how they might affect final outcomes. This is especially relevant since decisions at every stage "can lead to undesirable effects" [16], and identifying and alleviating such issues is seldomly straightforward in practice.

Literature offers several taxonomies of these sources of harm. For example, Mehrabi *et al.* [7] identify 3 principal categories, based on how they relate to the data, the algorithm or the user, while Suresh and Guttag [16] recognize 7 types of biases according to where they

intervene in the ML pipeline. Drawing from the gathered findings, this paper proposes a novel classification structured around the distinct stages of consideration practitioners undergo in their development process. The discussion in the following paragraphs will be supported by examples from a financial context, where the model's task is to predict whether a person will default on a loan or not.

Dataset & its transformations. The first such considerations revolve around the data used to train a ML model, through its collection and any consequent transformations. Undesirable properties are not only induced through sampling bias, which entails that certain segments of the population are underrepresented, but also through societal bias, where despite the data being accurate, it "records questionable aspects of the world" [6]. An example of societal bias in the lending case is basing the prediction on the individual's income, decision which can perpetuate the gender pay gap. Data shifts and concept drifts [17] are two other concepts that can play a role in this regard. Furthermore, data attributes can skew outcomes through their nature and relevance. For example, in the credit lending scenario, race shouldn't have any predictive relevance, and thus will potentially bias final outcomes if used in the model. When it comes to the nature of the attributes, we identify oversimplified representations, such as binary race, problematic since they fail to capture the whole reality.

Bias can also manifest itself in the data through sensitive features and their causal influences [10]. In the lending situation, one might exclude demographic characteristics from the prediction (e.g., race), but leave in correlated features (e.g., postal code) that can act as proxies and thus potentially lead to racially biased outcomes. Lastly, noteworthy are also population transformation strategies, such as over or under sampling, or removal of missing values, outliers and duplicates, which can negatively affect outcomes when further data collection is not a possibility.

Building of models. A different source of harm are the algorithmic design choices which have been discovered to have implications on fairness through their inadequacy to capture the data [15] as well as through amplifying performance disparities across different data groups [7] [6] [16]. Such choices include amongst others the algorithm trained and the training objective, optimization functions and model hyperparameters as well as the choice to use statistically biased estimators [18]. All these choices rarely account for a broader environment [17], and as such can adversely impact individuals and contexts outside the training dataset.

Evaluation of models. Evaluation bias is another important source of unfairness which occurs with the use of inappropriate and unrepresentative benchmarks to evaluate models [7], and which can be magnified by irrelevant or incomplete choices of metrics to report performance [16]. Literature highlights that an evaluation solely based on metrics is also problematic since these metrics bear limitations and cannot encompass all algorithmic harms. Certain assumptions, such as the idea that "decisions can be evaluated as an aggregation of separately evaluated individual decisions" [6] can also have potentially unavoidable repercussion. For example, in the loan setting, two members of the same family should not be evaluated separately since denying one a loan can have direct implications on the other's ability to repay. Discrimination based on sensitive features is another source of harm [7], similar to evaluation bias, arising from inadequate choices of protected attributes, and consequently protected groups.

Task. The last category of practitioners' considerations when building Machine Learning systems revolve around the task for which those are built. Harm is induced when there is a discrepancy between the model's intended task and the way it is deployed and used in practice [16], as well as when the task is either undesirably oversimplified or when it is disallowing novelty in the favour of reproducing historical patterns. In the loan scenario, one could argue that a decision-support system which predicts the risk of an applicant to default on a loan is

more informative and potentially less harmful than a binary decision that denies all people under a certain threshold.

Practitioners are facing difficulty in foreseeing and attending to this multitude of highly complex algorithmic harms[17], which resulted in the need to conceptualize fairness definitions, metrics and mitigations as well as in the interest towards developing tools that can contribute to solving the issue. The following 2 subsections summarize the main advances in this field.

2.2 Fair Machine Learning and its limitations

Despite fairness being a context and application dependent concept [19], a great quantity of research has focused on attempting to define and categorize it. Verma and Rubin [8] identify statistical metrics, similarity-based measures and definitions based on causal reasoning, while others attempt to differentiate fairness definitions centered on the level at which equality is desired, namely individual, group or sub-group [7]. Not only almost each of these notions are independently discredited on their insufficiency and lack of robustness [9], but it has also been shown that they might be statistically incompatible with each other in various contexts [20].

Nonetheless, these metrics facilitate, while somewhat incomplete, a quantitative analysis of unfairness, which in turn inspired the implementation of bias mitigation algorithms. Research literature widely agrees on a classification for these techniques: preprocessing methods, aimed at balancing the dataset, in-processing or algorithm modifications that impose certain fairness constraints in the development stage of the model and post-processing approaches which consist of modifying the model's predictions such as to attain desired outcomes for different sub-populations [10] [9]. While research efforts have been devoted to perfecting these techniques, it holds that they are still not adapted to tackle bias in real-life contexts. This is largely due to their narrow algorithmic perspective, incomplete conceptualizations of discrimination and necessity to make a choice for a metric to debias the model [17] as well as their disregard of how protected classes relate to the application's broad justice aspects [21].

Encompassing metrics and mitigation algorithms, fairness toolkits are part of the vast existing solution space to algorithmic harms in ML systems [15]. Although literature contests the general methodology of pursuing fairness solutions limited to technical definitions [19], there is a high likelihood that practitioners will still resort to these toolkits in practice to discover and mitigate bias, due to their wide availability and ease of use. This is not only one of the primary reasons why a large portion of the fair ML research has focused on analyzing their shortcoming, but also a strong motivation for studies like the one documented in this paper. The next subsection dives deeper into these limitations, while also touching upon different kinds of fairness-oriented interventions.

2.3 Human factors in fair Machine Learning

The technical communities have manifested an explosion of research on open-source fairness toolkits, with a large fraction indicating a pressing disconnect between their capabilities and the needs of practitioners that intend to use them in commercial contexts [22]. Although the studied toolkits can have a huge positive impact on ethical decision-making, the design and demonstration of fairness findings leaves a lot to be desired [23]. The literature points that these toolkits not only exhibit limited regard for real-life circumstances [14], but they also lack critical components, such as recommendations for mitigation and preliminary ML pipeline stage interventions [23]. For this, Holstein *et al.* [24] identify two important areas for future research, namely support in the process of collecting and curating datasets and development

of domain-specific educational resources and tools. Insufficient guidance for choosing the appropriate metrics and algorithms based on context as well as lack of explanations that prevent both oversimplification and information overburden were also found to be prominent limitations [14]. Lastly, Deng *et al.* [25] found that "practitioners desire more support in contextualizing, communicating and collaborating around ML fairness efforts" than what is currently offered by these toolkits and that, as a consequence of their narrow scope across the Machine Learning lifecycle, individuals resort to personal knowledge to mitigate problems, which often leads to an incomplete or distorted perspective.

By conducting interviews with data scientists and machine learning engineers, these studies discovered that an effective toolkit has the power to foster an ethical and fairness oriented thinking among practitioners, whereas a poorly framed or critically lacking toolkit might instill a false sensation of reliance on defective solutions. Our research complements these findings by aiming to assess, through a similar methodology, whether it also holds that the toolkits' limited consideration of fairness issues can impact the frame of practitioners with regards to broader algorithmic harms.

On top of the studies that look into fairness toolkits, a plethora of Human Computer Interaction (HCI) research has been conducted on different types of interventions that encourage ML practitioners to notice and build a thorough understanding of ethical problems in their models. Kaur *et al.* [26] evaluate interpretability tools, and conclude on their potential to uncover issues in both datasets and models, but warn that they are oftenly misused or over-trusted. Moreover, Hohman *et al.* [27] identify an urgent need for better explanatory interfaces. Karen L. Boyd looks into an accountability intervention useful for unfamiliar dataset exploration by giving study participants a specific ML problem based on a purposefully ethically problematic dataset, technique that served as inspiration for this research [28]. And while automated ML techniques are on the rise, the literature demonstrates the indispensability of human developers in machine learning development [29], which furthers the need and impact of fairness related research.

3 Method

To understand practitioners' consideration of algorithmic harms in their practices both with and without a fairness toolkit, an empirical study was employed, part of which several semi-structured interviews were conducted. The experimental setup, recruited participants, materials used and the interview protocol are all discussed in the following subsections.

3.1 Experiments Overview

The experimental design consisted of two main study conditions corresponding to two different fairness toolkits, namely Fairlearn and AI Fairness 360, to which participants were assigned following a between-subjects methodology. Individuals interacted with one of the toolkits in a Google Colaboratory notebook in which they explored problematic datasets in an attempt to build and evaluate binary classification models. To reason about the influence of using a toolkit or not (independent variable) on practitioners' behaviour towards algorithmic harms (dependent variable), two confounding variables needed to be accounted for: individuals' degree of familiarity with the respective toolkits and the other materials used in the interviews, namely the tasks and the datasets. The first was managed by recruiting participants from different positions on the fairness toolkit experience spectrum, while the latter was tended by a well thought of design and allocation of materials.

Experience with toolkit	Practitioner	Interviewee ID
No prior experience	Master Student	P23, P24, P26, P27
	Junior Data Scientist	P21, P28
	Senior Data Scientist	P22, P25, P29, P30
Prior familiarity with Fairlearn	Researcher	P1, P2, P8
	Data scientist	P3, P4, P5, P6, P7, P9, P10
Prior familiarity with AIF360	Researcher	P12, P15, P19, P20
	Data scientist	P11, P13, P14, P16, P17, P18

Table 1: Overview of participants based on their fairness experience and ML expertise.

3.2 Participants

A total of 30 participants were recruited, in the period of April-May 2022, across varying research and industry institutions as well as different areas of technology, such as computer vision, predictive maintenance and web information systems. All of them were chosen to have either training or responsibilities in machine learning training and evaluation. A categorization of practitioners based on toolkit familiarity and ML expertise can be found in the Table 1.

Participants were recruited by means of personal network, toolkits' official Discord or Slack[1] communication channels as well as professional oriented social media platforms, such as LinkedIn[2]. The study was approved by TU Delft's HREC committee and all participants signed an Informed Consent form acknowledging the risks involved as well as agreeing to any consequent data processing.

3.3 Fairness toolkits

As previously mentioned, the two studied toolkits were Fairlearn and IBM's AI Fairness 360, chosen based on their frequent occurrence in the fairness research field as well as based on their popularity amongst industry practitioners. Additionally, authors in [14] have verified through a focus group that these tools are amongst the ones that are most likely to be found through web search that have the most suitable technical fairness-oriented techniques.

Fairlearn is an open-source toolkit that provides means for assessment and mitigation of fairness-related harms in ML systems. It covers both classification and regression tasks, and it supports a wide range of group-level fairness metrics to assess the model's implications on different groups of people identified by sensitive attributes [11]. In terms of mitigation algorithms, Fairlearn includes at least one for each of the three common types: pre-processing, in-processing and post-processing [7]. However, addressing negative impacts during the model training phase is done only by reduction approaches [30].

AIF360 is a similar toolkit, providing both detection and mitigation strategies [13]. Compared to Fairlearn, it also supports individual fairness metrics, such as sample distortion [14] but covers only classification tasks. This toolkit's portfolio of mitigation algorithms is notably larger than Fairlearn's, and while not having its own visualization functions, AIF360 provides a set of built-in common problematic datasets which can serve as basis for user instruction.

Although considerable differences exist between the two toolkits, the interview set-up and practitioner's interaction with the toolkit was designed to be fundamentally similar as to allow for a seamless comparison of results.

[1] https://discord.com/, https://slack.com/
[2] https://www.linkedin.com/

3.4 Datasets, ML tasks and models

Fairness tools were used to explore problematic datasets and sources of algorithmic harms while solving classification tasks. The participants with experience looked into a pre-processed version of the Diabetes dataset [31], where the task was to predict whether a patient will readmit within 30 days [32]. This dataset was chosen based on its rare occurrence in fairness studies and in toolkit development, with the goal of obtaining an unbiased and authentic perspective from the practitioners that interacted with this use case.

Inexperienced practitioners were interviewed over two use cases, the previously described Diabetes dataset and a simplified version of the 2015 Full Year Consolidated Data File [33], where the aim was to predict whether a patient will have a high medical services utilization rate. The latter data, while present in several tutorials of the chosen toolkits, is not greatly covered by the ethical AI field, which warranted that most study participants will not be priory acquainted with its controversial nature. Another design choice was to leave in these datasets any inherent biases as well as to manufacture additional algorithmic harms, such that if disregarded, they will be reflected in the prediction models built. A reflection upon these algorithmic harms in both use cases can be found in Appendix A.

Between use case exploration, participants with no prior experience saw a short toolkit demo, based on two financial datasets commonly used in fairness research due to their uneven class representation, namely the German Credit Dataset [34] and the Credit-card default Dataset [35]. For individuals to understand how the toolkits intervene in the ML pipeline and how efficient they are in mitigating biases, comprehensive lists of fairness metrics were computed both on unmitigated and mitigated versions of a Logistic Regression model. The mitigation algorithms applied were ThresholdOptimizer and GridSearch for Fairlearn and Reweighting and Prejudice Remover for AIF360.

3.5 Interview protocol

The structure and the duration of each interview were decided based on the respective participant's experience. Studies with individuals with prior toolkit experience lasted roughly an hour, while people with no experience were interviewed for approximately two hours. As a convention, the two types of interviews will be henceforth named Group A and Group B. Studies were conducted online via Microsoft Teams, recorded with the participants' consent and later transcribed for the purpose of a statistical analysis. The appropriateness and the allocated time to complete the tasks were perfected through two pilot studies.

To be able to understand practitioners' awareness and actionability towards algorithmic harms, two levels of questions were designed for each such relevant harm, namely a vague indication pointing towards a larger group of harms followed by a specific indication for each harm in that group. These questions were integrated in the protocol as follows: after a few background questions, Group A was introduced to the Diabetes use case, and at the end of their exploration they were asked both Level 1 and Level 2. On the other hand, Group B was first introduced to the Medical use case, and concluded their exploration with Level 1 questions only. The next step was to walk the participants through the previously described toolkit demo, after which the Medical expenditure use case was discussed in depth. Participants were then pointed specifically towards the harms they did not mention by the Level 2 questions. A complete list of all the questions can be found in Appendix B.

At the end of each study, participants shared their past or envisioned experiences with responsible ML, addressing details such as influence of fairness in their work, personal attitudes towards impactful models as well as suggestions over the fairness toolkits.

3.6 Analysis of results

The qualitative data gathered from conducting the semi-structured interviews has been analyzed through a combination of deductive and inductive approaches. The deductive part arises from the design of the interview questions aimed at identifying strategies for harm recognition and mitigation. On top of that, a slightly modified version of a general inductive approach [36] was also applied, based on audio recordings instead of text data due to both shortcomings in semantic accuracy of the transcription tool used, and a strict study completion deadline.

The initial round of analysis started simultaneously with the first interviews conducted and consisted of deriving general categories from the evaluation aims, comprising (1) fairness related issues feasibly handled by toolkits, (2) other kinds of harms and (3) opinions of the usefulness, completeness and limitations of toolkits. These categories were iteratively refined not only through a creative process but also through a further investigation of the rest of the interviews. During this process lower-level categories were also identified, including harm identification strategies, mitigation approaches and ways of interaction with a fairness toolkit. At the same time, insightful quotes conveying essence of different categories were monitored and overlaps and redundancy amongst both types of themes were reduced. The analysis presented in the following section represents participants' perspectives, systematically vetted within an interpretative framework based on postpositivism and social constructivism assumptions [37].

In an effort to also analyze results in a quantitative way, four scales have been designed corresponding to the four different stages that constitute practitioners' behaviour towards algorithmic harms, namely awareness, understanding, identification and actionability. For each aspect, participants were rated proportional to the extent to which they manifested comprehension and thus on three levels that can be summarized as follows: not at all, partial and comprehensive.

4 Results and Evaluation

In presenting the results, we first revolve around harms that can be handled with a fairness toolkit and we continue with other potential sources of issues in Machine Learning development. Finally, we review participants' viewpoints about the usefulness, applicability and limitations of the toolkits. Appendix C contains a table that summarizes all the obtained results, in the form of specifying, for each harm, the number of participants for the four scales previously mentioned in subsection 3.6.

4.1 Fairness related harms

Data characteristics

Sensitive features and definition of protected attributes. Sensitive features were one of the first sources of harm mentioned by practitioners with toolkit experience. These individuals proved to be extremely aware of the fairness implication of defining and removing protected attributes, but mostly relied on general heuristics to identify them. They showed profound considerations in particular around demographic characteristics, such as race, gender and age, but barely any of them mentioned involving domain specialists and only 3 practitioners talked about checking regulations or guidelines around the specific use case before making a decision on whether to use these attributes in the model prediction or not. While the more popular opinion in this group was to either protect or entirely drop demographic features, 3 participants

contemplated their predictive value in the given medical context and the consequent danger of aiming for an equal prediction instead of an equitable one. Regardless of approach, all practitioners agreed that "if dropped, [you need to] ensure there is no information about them ingrained in the data" (P9, exp. F)[3].

An interesting behavior can be observed in the other group of practitioners, where 6 out of 10 came to identify the issue by themselves only after seeing the toolkit. This subgroup unanimously agreed that sensitive features should not be included in prediction, with little to no consideration of the consequences. The other 4 practitioners displayed prior knowledge of the issue and deliberately reasoned about the implications of their approaches: P19(no exp.)[4] said "even getting rid of those [the sensitive features], I don't think would make us really blind to them", while P23(no exp.) warned about there being "a lot more to it than the ones we know, which are mainly gender bias or racial bias".

Proxies and Correlations. Correlations between attributes and proxies were another two of the most frequently mentioned steps of data analysis. Most participants referred to broadly checking all correlations between the attributes in the dataset, but the ones with toolkit experience focused more on highly correlated values especially with the sensitive features. Some inexperienced practitioners came to identify this source of harm after being introduced to the toolkit, mainly due to an increased awareness around the sensitive attributes and their influences over other features. After this, their analysis became fairly similar to the other group. Only 2 participants gave clear suggestions on choosing correlation thresholds: P26 mentioned choosing 0.5 as a general threshold, while P22 proposed using statistical tests. While some people with experience talked about using techniques to decorrelate attributes, none of them mentioned making use of a toolkit for this. A considerable number of people mentioned dropping "the correlations with the target variable" (P19, no exp).

Bias mitigation

In the exploration of the first use case, mitigating bias was mentioned mostly indirectly by practitioners with no toolkit experience in their discussions around data preprocessing. Most of them were not aware of in or post processing algorithms, therefore, after the initial data preparation, their approaches focused more on experimenting with a larger variety of models and choosing the right one. Two behaviors emerged after seeing the toolkit: individuals either talked about the "need to fundamentally understand how [algorithms] behave before using them" (P19, no exp), or showed interest towards algorithms that allow for a trade-off between fairness and accuracy with little consideration on where they intervene in the pipeline, such as Fairlearn's GridSearch or AIF360's Reweighting algorithm. The majority still believed that "data is the first source of harm" (P23, no exp), and saw these mitigations as a complement to having an unbiased and representative dataset.

On the other hand, practitioners with toolkit experience exhibited profound understanding of the mitigation algorithms present in the frameworks and the fundamental differences between the stages at which they are applied. And while equipped with this knowledge, the majority still preferred data preprocessing as the focal point to attend to fairness issues: "the best place [to mitigate bias] is the data preprocessing stage because it can give indications on what [model] to use later" (P15, exp. A), "you need to look yourself into the data to see what types of biases it contains rather than using a library and looking at outcomes" (P17, exp. A). To deal with disparities in different groups, practitioners also mentioned putting people to a

[3]exp. F / exp. A = participant with prior experience with Fairlearn / AIF360

[4]no exp. = participant with no prior toolkit experience

different standard by having different decision thresholds (P1, exp. F) or giving the minority group a better rating rather than correcting for the majority group (P5, exp. F).

Some practitioners favored certain algorithms with no consideration of the additional biases they might introduce: P12(exp. A) said "with post processing in production you can just change the labels", while P5(exp. F) argued "in processing is the best option" after ruling out pre processing for data size reasons and post processing for model robustness concerns. Contrastingly, others were reluctant to use the toolkit for more than computing metrics: P1(exp. F) believed that "mitigation algorithms are not at the stage where they should be [...] changing the reality as you observe it should be a very informed change and not some thing thrown into an optimization algorithm to just randomly change the data", while P2(exp. F) mentioned "besides metrics [...] anything else is human consideration". These participants placed more emphasis on thorough assessments, efforts to comprehend what is happening and the reasons behind, discussions with domain specialists and more accurate data collection.

Evaluation of models

Choices of metrics. Practitioners largely identified two distinct evaluation goals, accuracy and fairness, and recognized the trade-off that has to be made between them for most applications. More than 90% of interviewed people mentioned computing multiple metrics to test the performance of their model and choosing them based on context. From the metrics discussed, accuracy, precision, recall and f1 score were the most frequent. Practitioners with experience also mentioned computing fairness metrics, but with a significant focus on the simpler ones, such as true positive and true negative rates, rather than the more complex ones included in the toolkits, such as equalized odds or statistical parity. Very few of them reported using toolkits for these simple metrics, with a strong preference for more familiar libraries, such as pandas or sklearn, because of their ease of use. Interestingly, participants from the Fairlearn group as well as the ones with no toolkit experience ranked accuracy above fairness, while the ones that had experience with AIF360 did the opposite. Practitioners from both categories also discussed penalizing certain errors more than others depending on their context meaning. In the scenario of hospital readmissions, that would translate for example in focusing more on minimizing the false negatives over the false positives.

Another harm identified in the literature review stage was the consideration of demographic parity alone when evaluating or mitigating bias. The fact that most practitioners did not mention optimizing for this metric, although it's one of the most prevalent in the toolkits' libraries, shows that they are aware that this metric makes strong assumptions that are not always applicable and that it should be investigated only as an indication. P27 gave a thoughtful example from a college admission context: "if you have 95% men or women then you probably have a problem, but you shouldn't strive to have exactly 50% of each either". Lastly, the issue of a too large dependency on metrics was only briefly touched upon by practitioners mentioning trade-offs based on what is more relevant for the context, with the majority of them not being able to identify evaluation alternatives to computing the chosen metrics.

Broader environment. Here we discuss three broader algorithmic harms: output versus outcome, harms for individuals or environment outside the dataset and lack of consideration for people not directly subject to the model predictions. While a large number of practitioners talked about involving domain specialists and designing support systems, instead of fully automating decisions, none of them mentioned looking too far outside the context of their models. When asked, several practitioners with no toolkit experience acknowledged "the responsibility of the developer" (P24, no exp.), but no concrete solutions were given.

4.2 Other types of harms

Task

When it comes to the task itself, 20 participants had some minimum consideration of whether it makes sense and whether it should be automated. Some reported thinking about "where the model would fit in the current process" (P1, exp. F) and mentioned that "a [ML] system should only be used to inform decisions, and not fully automate them" (P25, no exp). Others also thought about "impacts of wrong decisions" (P9, exp. F) and privacy issues (P3 exp. F; P24, no exp.), but the common denominator in more than a half of the interviews was the idea of closely involving domain-specialists in the initial design phase. There were no significant differences in awareness between people with and without toolkit experience.

Data attributes

Irrelevant/Questionable attributes & their removal. Participants with prior toolkit experience exhibited a deeper understanding towards the relevance of the data attributes and stressed the importance of understanding the relation between each attribute and the prediction target more frequently than participants with no experience. Half of these individuals identified all questionable attributes that were synthetically added to the datasets, such as marital status or region, and talked about involving a domain expert in their assessment, while some also mentioned ideas for additional information that could be useful in the context under discussion. P9(exp. F) identified "throwing in all the data and checking which features had the most predictive power" as an inferior approach to spending the necessary time to "understand each variable and what kind of influence it could have", while P27's (no exp) approach would have been exactly removing the last 10 or 20 percent of the least predictive features. However, there were 5 individuals with no experience that also briefly mentioned selecting attributes based on context and whose opinions stayed constant before and after using the toolkit.

Oversimplified attributes. Questions about oversimplified attributes were not asked explicitly in the interviews with the people with toolkit experience and only two participants unassisted identified that the binary representations of race and gender could be a harm. When asked, all individuals with no experience manifested actionability towards this issue, proposing either collection of data that is more representative to "society's requests" (P18, no exp.) or clearly informing stakeholders when deploying the model that "it only works for represented samples" (P26, no exp). Questions were asked after seeing the toolkit, so their behavior can either be attributed to an increased awareness because of the toolkit or to a prior inattentiveness. Interestingly enough, P27(no exp) identified the issue by themselves after interacting with the toolkit and extensively talked about it from a fairness perspective.

Data population

Incorrect or biased labels. No participant discussed issues around the source of the data labels before being prompted by the interviewer, but almost all of them identified the annotators' lack of context knowledge and the potential bias they might introduce as necessary considerations to be made. Only 2 experienced people discussed the invisible worker issue [38] by mentioning looking into the sources of the labels from a "[people] exploitative point of view" (P1, exp. F).

Data representation & sampling. Checking the data representation and performing any consequent data transformations were one of the first things practitioners, from both categories, mentioned during their use case exploration. Almost all participants exhibited substantial understanding towards how unfairness stems from an uneven distribution in terms of not only the sensitive attributes but also the target label. Different practitioners had conflicting opinions, namely some aimed for an "even representation" (P5 exp. F; P22, no exp.) of the sensitive features, while the majoritar rest thought the data should be representative of reality, even if it is not perfectly balanced. While several individuals, almost evenly split across the two groups, only mentioned checking the distribution of the data with no notable mitigation suggestions, the larger proportion displayed consideration towards not only valuable practices but also the fairness implications of these practices.

The majority of interviewees expressed concerns towards undersampling techniques because of a loss of data, but participants that had experience with AIF360 offered a more in depth reasoning, an example being P13 that mentioned an additional loss of "nuances that may be important to differentiate [between] labels". A more remarkable difference in approach can be observed however in the alternatives proposed to undersampling: participants with toolkit experience majoritarily preferred adjusting decision thresholds after training or weighting the errors in underrepresented classes more, while participants with no experience talked about choosing a model that can work well with imbalanced data, with P27(no exp.) mentioning the power of anomaly detection algorithms. On the opposite side of the spectrum, several individuals mentioned oversampling approaches, however none of them had any consideration of the potential harms these techniques may introduce.

Lastly, after walking through the toolkit tutorial, P18(no exp.) changed their opinion from aiming towards a dataset that is representative to reality to having a balanced one, without questioning any further implications. This poses a serious threat since most toolkit demos available online apply this data transformation without raising sufficient awareness.

Handling of data errors

Missing data. 9 out of 10 practitioners with no experience identified, by themselves and for both use cases, missing values as a requisite in the initial data analysis phase, with almost all agreeing with going for "replacement rather than removal if possible" (P23, no exp.), however without considerations of potentially introducing harms. Some of them discussed imputing numerical attributes and experimenting with the categorical ones, while others mentioned building and comparing separate models, with and without the features that have missing data. When it comes to the other group of participants, a quarter of them either did not discuss the issue or did not exhibit enough actionability when asked. Out of the rest, some mentioned adapting practices based on "how much is missing" (P7, exp. F) as well as the potential problems posed by imputation approaches, such as introducing "artificial qualities in the data" (P15, exp. A). Only P15 mentioned algorithms in the toolkit that can "account for missing points".

Outliers. Very few participants directly mentioned this kind of data error by themselves, but in turn talked about plotting distribution of continuous variables as one of the first things they would do. There were no considerable differences between participants with and without experience and neither between with and without toolkit for the latter. Insightful opinions about handling outliers came from both types of individuals: P6(exp. F) mentioned that "you need to contextualize outliers and anomalies the way you contextualize fairness - even if you have a statistically large or small value, there might be a reason for it" and P19(no exp.)

warned that although removing outliers can increase accuracy, "the model is not going to work well for those [data points]". Interviewees also mentioned manual checking and exclusion of outliers that clearly steam from wrong data collection.

Duplicates. The issue with duplicate entries has not been discussed with any individual that had experience with AIF360. Only two practitioners, from the no toolkit experience group, identified the harm on their own, while all other participants discussed it after the interviewer's specific indication. Besides the possibility of the toolkit constraining the frame of consideration of this harm, this finding can be attributed to two other reasons: (1) some participants (exp. F) were priorly acquainted with the datasets and (2) some assumed medical records would be unique despite the absence of a personal identifiable attribute. There was a general agreement around the practice of removing only actual duplicates after a thorough check as to avoid "model overfitting on them" (P24, no exp.).

Building of models

Choices. Out of the 30 interviewees, only 6, majoritarily with prior toolkit experience, didn't identify the choices made during the model building phase as a significant source of harm, either because of a lack of knowledge or stemming from a belief that "harm is caused by the data, not the model" (P5, exp. F). When it comes to the other participants, all of them touched upon harms introduced by model building decisions during their own use case exploration, however with a significant difference in consideration between the people with and without experience with the toolkit. The first category stressed that "feature importance" (P6 exp. F; P11, P15, exp. A) and "the choice of training features" (P4, exp. F) are the two main sources of bias in this case, while the latter group discussed choices of hyperparameters, their optimization and choices of thresholds. P2(exp. F) also stressed that "model choices need to be subjective to the bias mitigation you want to do". Two commonalities can be observed in the practices of practitioners with different fairness experience, namely not only the idea that the choice of the model should depend on the data size and distribution but also that, in a lot of contexts, "explainability is of higher value than accuracy" (P20, no exp.).

Environmental impact. Most participants exhibited awareness and understanding towards the environmental impact of model training only when asked explicitly, which can partially be attributed to the nature of the explored use cases not posing significant issues towards this harm. The matter was not discussed in 9 interviews while 3 participants reported that they would not normally look into aspects like the carbon emissions of their models. 2 participants identified a trade off between the negative impact of the model's training and the purpose that it serves, one of them mentioning the example of an agricultural drone that can help save energy consumption. The rest of the participants manifested great actionability by deliberating whether largely complex models are even needed in the first place: P15(exp. A) said "if a simple algorithms does your prediction quite well, that should be used [instead of a big model]", while P10(exp. F) talked about "making sure the amount of data we are requesting doesn't have any side effect on the environment".

Not discussed

The list of harms produced as a result of the prior literature study contained a few more sources of harm for which the gathered data has not been deemed enough to draw relevant conclusions. These algorithmic harms are (1) feature engineering, (2) general data or model transformations after applying bias mitigation techniques and (3) concept drifts and covariate shifts that characterize data.

4.3 Opinions of toolkits

Besides observing practitioners' interaction with algorithmic harms by having them explore use cases, we also asked them directly how they feel about generally using fairness toolkits and whether they think these toolkits changed their considerations in any way.

Almost all of them perceived toolkits as a useful addition to their normal practices, however for different reasons, depending on their general fairness knowledge. From the group with no toolkit expertise, the least experienced individuals reported that the toolkit helped them identify problems they did not consider before, such as sensitive or oversimplified attributes, while the rest, together with the group with toolkit knowledge, had other motivations. Noteworthy examples were "a more appropriate alignment" between datasets characteristics and interpretations (P2, exp. F), a more structured unfairness investigation and mitigation process (P14, exp. A; P22, no exp.) as well as an efficient way to "quantify uncertainty" (P6, exp F; P20, no exp.). These individuals also mentioned that interpretations don't come from the toolkits themselves and that using a toolkit shouldn't replace having a good understanding of the problem and a comprehensive domain knowledge.

On the contrary, practitioners, especially the ones acquainted with the workings and limitations of toolkits, also perceived them to be constraining in some ways. They identified a potential for misuse in the case of "using them blindly" (P9, exp. F) or in case developers "don't know the definitions [of fairness metrics]" (P16, exp. A). They also stressed that the toolkits are "fairly academic at this point" (P15, exp. A) and that there is a "big learning curve" (P12, exp. A). Moreover, while enthusiastic about picking up a new skill, practitioners with no experience also drew attention towards the need of a critical view when using these tools in practice.

5 Responsible Research

Having presented the main results and prior to discussing how they fit into the broader context of fairness research, it is imperative to carefully deliberate the ethical facets of all aspects that constitute the documented study. To do this, we will examine how the presented work adheres to five principles formulated in the Netherlands Code of Conduct for Research integrity [39].

Honesty. Conformity with the first principle of honesty is uphold by an accurate and truthful reporting of the results gathered in both research phases, namely the literature study and the conducted experiments. No unfounded claims are being made and only the direct data outcomes are being documented without any evidence manipulation. .

Scrupulousness. Scrupulousness has been achieved by solely undertaking scientific practices in all stages of the research, from meticulously designing and conducting interviews to systematically disseminating and reporting the results. Furthermore, data is consistently documented from an objective stance and no personal bias interferes with its validity and accuracy.

Transparency. The work also adheres to the principle of transparency as it chronicles a complete and detailed overview of all the steps that have been taken without omitting any relevant information. As most other scientific studies, the documented research is subject to reproducibility barriers, such as complexity, technological change or human error [40], which served as primary motivation not only for providing a complete set of instructions describing the methodology employed but also for attempting to present the data without excessive prior manipulation. This has ultimately ensured that the work is not only verifiable but also reproducible.

Independence. Independence manifests itself in the scientific considerations that have guided this research. The reasoning behind the choice of the two fairness toolkits used in the interviews has been thoroughly explained, highlighting that there were no commercial, political or other non-scholarly influences behind this decision. Moreover, the author has not refrained from reporting all relevant results, including the ones that might conflict with the interests of the people promoting the toolkits.

Responsibility. The existing literature gap motivated why the current research is scientifically relevant, ultimately attempting to provide guidance towards a better consideration and mitigation of fairness related issues in Machine Learning development. The work has accounted for legitimate interests of the human test subjects, including and not only limited to consequent data processing and utilization within and outside the scope of this study.

6 Discussion

Based on the knowledge gathered from previous research investigating the use of fairness toolkits in practice, we formulated, in the beginning of our study, the following hypothesis: these tools might narrow down the scope of harms data scientists address in their considerations when building and deploying ML systems. Through conducting an empirical study and a comparative assessment of practices with and without a toolkit, we managed to acquire not only meaningful insights into the general effectiveness and shortcomings of fairness toolkits, but also areas of future research for practitioners support towards attending and mitigating unfairness issues in their models.

Findings and Implications

Our findings show that, while being true that toolkits slightly shift practitioners' focus towards fairness related issues included in these libraries, they hold more advantages than pose problems when employed in the development pipeline. For a large majority of the harms analyzed, we have observed no significant differences in understanding and approach between practitioners with and without toolkit experience as well as between toolkit assisted and unassisted use case exploration. Nonetheless, there are not only some areas that require future intervention but also areas in which the toolkits resulted in greater understanding and actionability.

Future intervention

Sensitive features. When it comes to the sensitive features, we observed that the popular opinion amongst the experienced group was to either remove or protect them, without little to no consideration of further implications. This is similar to what authors found in [25], namely that practitioners oftenly draw upon personal experiences or general heuristics to surface potential sensitivities and thus fall prey to the "fairness through unawareness" trap [41]. This practice is somewhat encouraged by the toolkits, with their seamless support of defining protected attributes in the model building and mitigation stages, therefore, in this sense, toolkits are capable of narrowing down perspectives to predefined ways of handling sensitivities in datasets. However, a large part of the inexperienced practitioners manifested awareness of this issue only after being introduced to the toolkit, so we conclude that the problem does not lie in the use of the toolkit, but in insufficient guidance and educational support.

Bias mitigation. Half the focus in fairness toolkits are bias mitigation algorithms, becoming thus a pressing issue not only for the current research but also for future studies. We

explored the potential of their misuse and incorrect application, and discovered that some practitioners are willing to utilize certain algorithms, without a thorough understanding of their inner workings, in the case that they render better outcomes in terms of the computed fairness metrics. These individuals came mostly from an academic setting, while the industry data scientists either pointed towards the indispensability of an in-depth comprehension or showed interest only towards algorithms that explore a trade-off between fairness and accuracy. We believe the latter group's attitude was influenced by their extensive past work experiences, with almost always being constrained to achieve certain accuracy targets when developing systems in production settings. If correctly used, mitigation solutions, in the form of algorithms that can be applied in the three critical points in the pipeline, can be very powerful, thus, from the mentioned findings, we conclude again on a shortage of user guidance.

Missing data. While a large number of practitioners with fairness expertise reasoned in great detail about missing data and potential implications of imputing values, a quarter of them failed to gauge this issue before being prompted and only one participant seemed to be aware of toolkit's algorithms that account for missing values. On the other hand, in the group of inexperienced practitioners, there was a single case of not exploring the matter in question, and considerations didn't change in the second use case exploration. Toolkits' documentations, especially of the ones investigated in this study, clearly specify how handling of NA values is being done and even warn about the potential dangers of both removal and imputation practices. Therefore, on top of a lacking support, we also conclude on a potential overreliance and overtrust in these kinds of technical solutions.

Choices of models. Choices made during model building, such as algorithm, training objective or optimization of hyperparameters can also affect the model's fairness. We observed that practitioners with no toolkit experience were more attentive towards this, while the other group focused on the data as the primary source of bias and the focal point for intervention. And while this approach is not wrong, overlooking implications during the model building stage can render all other efforts futile. With this, we point towards a need to place more emphasis on the idea that toolkits should ideally complement and not replace standard practices, and consequently towards the broader issue of practitioners' educational support.

Greater understanding due to toolkit use

The investigated toolkits improved practitioners' considerations around the following harms:

- **Irrelevant attributes:** where practitioners with toolkit experience mentioned more in depth looking for questionable features, closely involving domain specialists and choosing training features based on relevance and relation to target.
- **Source of labels:** where the invisible worker issue was touched upon only by practitioners with toolkit experience.
- **Undersampling techniques:** where, while everyone was concerned about the loss of data, only individuals with toolkit experience questioned the loss of nuances or important details that can enrich the predictions.

Limitations

Limitations of this study include participation bias, the sample size and the limited number of toolkits investigated. To increase the generalizability of the results, future work could look into a larger subset of participants from a wider variety of backgrounds in terms of experience in the field of fair ML and field of work. Furthermore, the study only focused on the exploration of a small number of models and tasks, with which some participants had prior experience. Due to these use cases being restrictive by nature as well as due to the fact that the toolkit tutorials

only covered a small subset of the provided functionality, not all results could be attributed to the use of a fairness toolkit, which was ultimately what the study tried to investigate.

Future directions

In line with the need for support and educational guidance towards ethical sensitivity as well as with our interviewees' wish to involve domain specialists and legal experts in their processes, we promote the importance of some of the recommendations already formulated in existing literature. We believe that the scope of future fairness toolkits should be broadened to not only include instruction materials to create and review DataSheets [28], but to also promote interdisciplinary collaboration between stakeholders from different backgrounds [25].

And while our study did not surface sizable differences in approaches to data documentation with and without a fairness toolkit, it reiterated the widely known fact that currently these approaches are largely ad-hoc and short-sighted [42]. We stand by the opinion that future toolkits have the potential to close this gap if they are to provide "more actionable guidance on how the characteristics of datasets might result in harms and how these harms might be mitigated" [42], such as support for assessing the relevance and problematic nature of the data attributes for specific contexts.

Finally, our research proves its usefulness in a broader context by attempting to mitigate concerns around the future of the data science profession. By investigating potential downsides of employing fairness toolkits in production and paving the way towards more appropriate technical solutions to fairness and data documentation, the study alleviates distress around the prospect of a hostile takeover by ill-suited automated tools, in a world where "AutoAI is the future of data science" [43].

7 Conclusion

Through an empirical exploration of practices around algorithmic harms, we aimed to identify whether fairness toolkits can generally constrain fairness considerations and lead to overlooking certain harms. By engaging with practitioners, we discovered that our initial hypothesis was only partially true: while toolkits can lead to a disregard towards insufficiently covered sources of harm, they also have the potential to engender proactiveness towards a multitude of other issues. We have found that future intervention is needed to increase the coverage of the toolkits across the Machine Learning lifecycle, such as to include better support for selecting sensitive features, applying mitigation algorithms as well as for making appropriate choices during model building. However, we also found that fairness toolkits generally increase practitioners' awareness around the socio-technical implications of their work. This study confirmed the importance of a thorough design and evaluation of toolkits as well as the need to rigorously educate practitioners on sources of unfairness in their models prior to making use of such a tool in practice. Beyond this, we hope our findings instill interest towards further studies that can integrate the results into ways to improve the usability and applicability of future toolkits.

References

[1] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance*, 77(1):5–47, 2022.

[2] Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pages 1389–1398, 2018.

[3] Latanya Sweeney. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54, 2013.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications, 2016.

[5] Rahul C Deo. Machine learning in medicine. *Circulation*, 132(20):1920–1930, 2015.

[6] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021.

[7] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[8] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

[9] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019.

[10] N Ntoutsi, P Fafalios, U Gadiraju, V Iosifidis, W Nejdl, ME Vidal, S Ruggieri, F Turini, S Papadopoulos, E Krasanakis, et al. Bias in data-driven ai systemsâan introductory survey. arxiv 2020. *arXiv preprint arXiv:2001.09762*.

[11] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

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

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

[14] Michelle Seng Ah Lee and Jat Singh. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–13, 2021.

[15] Brianna Richardson and Juan E Gilbert. A framework for fairness: A systematic review of existing fair ai solutions. *arXiv preprint arXiv:2112.05700*, 2021.

[16] Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization*, pages 1–9. 2021.

[17] Agathe Balayn and Seda Gürses. Beyond debiasing: Regulating ai and its inequalities. *EDRi Report. https://edri. org/wp-content/uploads/2021/09/EDRi_ Beyond-Debiasing-Report_ Online. pdf*, 2021.

[18] David Danks and Alex John London. Algorithmic bias in autonomous systems. In *IJCAI*, volume 17, pages 4691–4697, 2017.

[19] Ben Green and Lily Hu. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the machine learning: the debates workshop*, 2018.

[20] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[21] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pages 149–159. PMLR, 2018.

[22] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[23] Brianna Richardson, Jean Garcia-Gathright, Samuel F Way, Jennifer Thom, and Henriette Cramer. Towards fairness in practice: A practitioner-oriented rubric for evaluating fair ml toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[24] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.

[25] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. Exploring how machine learning practitioners (try to) use fairness toolkits. *arXiv preprint arXiv:2205.06922*, 2022.

[26] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

[27] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13, 2019.

[28] Karen L Boyd. Datasheets for datasets help ml engineers notice and understand ethical issues in training data. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–27, 2021.

[29] Doris Xin, Eva Yiwei Wu, Doris Jung-Lin Lee, Niloufar Salehi, and Aditya Parameswaran. Whither automl? understanding the role of automation in machine learning workflows. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[30] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[31] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

[32] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1452, 2022.

[33] Agency for Healthcare Research & Quality. Meps hc-181: 2015 full year consolidated data file. 2017.

[34] Professor Dr. Hans Hofmann. German credit data set. `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`, 2000.

[35] I-Cheng Yeh. Default of credit card clients data set. `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`, 2009.

[36] David R Thomas. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, 27(2):237–246, 2006.

[37] JW Creswell and CN Poth. Philosophical assumptions and interpretive frameworks. *Qualitative inquiry and research design: choosing among five approaches. Los Angeles: Sage Publications*, pages 15–41, 2013.

[38] Will Douglas Heaven. Ai needs to face up to its invisible-worker problem. `https://www.technologyreview.com/2020/12/11/1014081/ai-machine-learning-crowd-gig-worker-problem-amazon-mechanical-turk`, December 2020.

[39] K Algra, L Bouter, A Hol, J van Kreveld, D Andriessen, C Bijleveld, R D'Alessandro, J Dankelman, and P Werkhoven. Netherlands code of conduct for research integrity 2018, 2018.

[40] Jesse M Alston and Jessica A Rick. A beginnerâs guide to conducting reproducible research. *Bulletin of the Ecological Society of America*, 102(2):1–14, 2021.

[41] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[42] Amy Heger, Elizabeth B Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata. *arXiv preprint arXiv:2206.02923*, 2022.

[43] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

A Appendix A - Algorithmic harms in interview use cases

Harms in Diabetes dataset:

- Task: The model's task can be considered to have oversimplified objective labels since patients are either classified as readmitted within 30 days or not. Because of the high-stakes context, the task has important ethical implications since it directly affects people (both the hospital the subjects to the modelâs predictions).
- Irrelevant / Questionable attributes: Marital status can be considered irrelevant to predicting readmission. Questionable attributes include, amongst others, region.
- Oversimplified attributes: Age is represented in 3 groups: over 60, 30-60, 30 or younger. Gender is encoded as binary.
- Sensitive attributes: Race, gender, age can be considered sensitive.
- Proxies / Correlations: Region is significantly correlated with medicare and serves as a proxy for the race attribute.
- Data representation Target label imbalance: RACE is represented as follows: 74% Caucasian. 20% AfricanAmerican and the rest divided by 4 other categories. By age, people over 60 are the most represented (60%). The ratio in terms of the target attribute is 88%:12%.
- Over/under sampling: Balancing the dataset in terms of the target attribute or the uneven represented attributes (race, age) severely reduces the data size and furthers data imbalance in other attributes.
- Handling of missing data: There are 21.832 out of 101.766 missing values in the weight attribute, and 49.949 in the medical_specialty attribute.
- Handling of outliers: Most outliers are found in 2 numerical attributes: number_of_lab _procedures and number_of_medications.
- Handling of (near)duplicates: There are no identical duplicates in this dataset. Since the records donât contain personal identifiable information (eg. patient number), it is impossible to say if near-duplicates belong to the same or to different patients.

Harms in Medical expenditure dataset:

- Task: The modelâs task can be considered unethical or undesired since prices for insurance are computed based on a simplification of estimating low/high utilization (people above value 10 in terms of utilization will pay a double price). Moreover, based on the choice of attributes used for modeling, the task is subject to reproducing historical data patterns without allowing for novelty.
- Irrelevant / Questionable attributes: Marital status could be considered irrelevant for healthcare utilization prediction. There are several other questionable attributes such as military service information.
- Oversimplified attributes: RACE is encoded as binary (white and non-white), such as there is no difference being made between races grouped under non-white.

- Sensitive attributes: RACE, SEX and AGE can be considered sensitive.
- Proxies / Correlations: RACE is significantly correlated with region, poverty status and health insurance coverage.
- Data representation Target label imbalance: RACE is skewed towards white people (66%:33%). In terms of AGE, people between 0 and 20 are the most represented, while 80+ are severely underrepresented. The ratio in terms of the target attribute is 80%:20%.
- Over/under sampling: Balancing the dataset in terms of the target attribute reduces the data size with 60% and furthers data imbalance in other attributes.
- Handling of missing data: 790 values missing in the WEIGHT attribute (out of 4400 total records). They all belong to âfemalesâ data points, therefore removing the records with missing values creates an imbalance in terms of the SEX attribute.
- Handling of outliers: Most outliers can be found in terms of the PERWT15F attribute (which is calculated based on region, status, race, sex, age and poverty status). Removing all outliers would significantly decrease the data size and lead to information loss.
- Handling of (near)duplicates: There are 870 duplicates out of 4400 total records. Their removal, as in the case of missing data and outliers, would have a significant impact on dataset size.

B Appendix B - Interview Questions

Background Questions:

- Demographic
  - Where are you from?
  - What is your gender?
  - What is your educational background?
- Experience with machine learning
  - Students:
    - What is your experience with machine learning?
    - Do you have any work experience in ML or data-science?
  - Practitioners:
    - Do you work in academia or industry? What is your role?
    - What is your technology area? What kind of domain have you worked with now and in the past? (e.g. banking, healthcare, etc.)
    - For how long have you been working with machine learning/data engineering?

After participants use-case exploration:

- What do you conclude from your exploration?
- Do you think the task can be automated? What would be some difficulties the hospital/insurance company might face?
- What would be your concerns? What interrogations did you have when going through the process?
- Were there things you would have liked to do but could not due to time, or due to the tools available being limited? If there were, could you elaborate on that now, what would you have done and why?
- Particularly, the hospital/insurance company has recently heard of questions around responsible AI / fairness in AI, now popular in industry. Would this application be concerned by these questions? Why / why not?

Questions about harms:

- Task
  - Level 1: Could you tell me if there is any reflection about the topic of responsible AI you would have around the modelâs task? Prior to looking into the data and potential models, are there other things you typically do / look into / need to know?
  - Level 2: In your process, would you / has it happened that you had to assess / decide by yourself about the appropriateness of the task? What about the relevance of the task to machine learning? What about the meaningfulness of the labels the model should return for each data point? If so, what would you look at in these 2 use cases, and how?

- Data attributes
  - Level 1: Is there any other reflection/discussion you might typically have around the attributes of the dataset?
  - Level 2:
    * Have you ever heard of irrelevant or incomplete attributes? Do you think these are applicable to these cases? If yes, why didnât you mention them before?
    * What about oversimplified or sensitive attributes?
    * Would you consider any other things when doing attribute transformation, such as definition/removal of protected attributes?
    * How would you detect all of these issues in general/ in these use cases? How would you approach solving them?
- Data population
  - Level 1: How about data representation/distribution? Is there any other analysis you typically make?
  - Level 2:
    * What would you consider when thinking about labeling the data or the labels associated with the data samples?
    * Do you think the data is representative to the real population in this case(s)? Is this something you would usually consider? And if yes, what would you think about? If not, why is it not important?
    * Do you think that population transformation (through over/under sampling) has any implication? Is this something you would consider?
    * If they donât explain when answering previous questions: How would you detect these issues regarding the data population? How would you approach solving them?
- Data errors
  - Level 1: What would you say about the quality of the dataset in this case(s)? Are there any things you would improve/solve before building a model?
  - Level 2: For each harm in this category (missing data / outliers / duplicates)
    * Are you familiar with this harm?
    * How would you detect it?
    * How would you approach solving it?
- Building of models
  - Level 1: When building a ML model for such a task, are there any considerations you would take besides trying to achieve an efficient and accurate model?
  - Level 2:
    * Choices: Do you think that your choices (choices in the algorithm, training objective, hyperparameters, post processing of outputs etc.) can have any harmful implications?

    * Bias mitigation: How would you deal with bias mitigation? In which stage would you intervene the most (data preprocessing, model selection, model evaluation etc.)? Why? How? And are you satisfied once you did that?
    * Model transformation: If you were to do additional transformations after applying a bias mitigation method and/or training one model, transformations either in the data or the model, are there any things you would revisit? Do you think this is important?
    * Other issues:
        · Do you think environmental protection strategies apply to training a ML model? If so, how? Would you be willing to make compromises in terms of the dataset size to minimize things such as energy consumption or carbon emissions?
        · Many ML tasks involve labeling data, would you look into the source of these labels? Can you identify any issues with using platforms like Amazon Mechanical Turk?
        · Do you think it is your responsibility to look for implications beyond the context of the model you are developing, say in a broader environment? If so, what would you consider?

- Evaluation of models
    - Level 1: You mentioned X when talking about the evaluation of the model. How do you usually choose your approach for this? What are usually your goals when evaluating a model? Are there things you try to avoid by employing a certain approach?
    - Level 2
        * Incomplete/irrelevant choices of protected attributes: Do you usually look into defining protected attributes and protected groups? If so, how do you do it? What would you consider for this in the context of the given use case(s)? Do you think your choices can impede your results? Is it possible that these choices are incomplete or not always relevant?
        * Incomplete/irrelevant choices of fairness metrics: Would you use fairness metrics to evaluate your model? How would you choose fairness metrics for a specific use case? What would you choose to use for this use case(s)? Why? For your evaluation, do you rely only on metrics or would you have any other considerations? Why ? What kind of considerations?
        * Too large dependence on metrics:Would you make any trade-offs between the different metrics that you know/that exist? Would you change anything in the metrics you use and optimize for? If so, what? Are there any limitations to the fairness metrics that you think should be addressed otherwise, e.g., legally or economically?
        * Output vs outcome: Would you consider how different people might be affected by the same output? Would you take into consideration a distinction between the output and the outcome of your model? If you think that would be important, how could you imagine doing that? Would it be challenging?
        * Parity only: Do you think demographic parity is a good enough indicator of a fair model? Why/why not?

    * People outside predictions: Besides the people directly subject to the model predictions, are there any other stakeholders that might be affected? How and why would you consider those?

Ending questions:

- Practices at the company (if applicable)
  - To what extent do you use human monitoring versus automation in your company to reduce ML harms? Does this work?
  - Did you ever face a trade-off between fairness and accuracy in your work?
  - Do you think absolute fairness is possible to achieve? At what point do you stop trying to reduce bias?
- Experience with responsible ML / ML fairness
  - Have you ever been confronted with ML models that can have a strong impact on certain stakeholders?
  - People now start talking about âresponsible AIâ. Have you heard about that? What does that mean for you?
  - Who do you think is responsible for tackling responsible ML questions? How does it work at your company? Are some stakeholders tasked to look into it? who?
  - What is your experience with Microsoft Fairlearn/AI Fairness 360?
  - Did you have any machine learning ethics / responsible machine learning training? At uni? At a company? Somewhere else? Please elaborate. How did you learn about these topics / to use the toolkit?
- Questions about toolkit
  - What do you think of the toolkit?
  - How complete do you think the toolkit is? Can you rely solely on that?
  - What do you think about ML toolkits and their effectiveness? What do you think about the metrics they provide? What problems do you find in them?
- Toolkit and change of perspective
  - After starting to use the fairness toolkit did your perspective on algorithmic harms change? If so, how?
  - Do you feel that, with the toolkit, your fairness consideration is limited to only the problems/metrics shown by the toolkit? Or on the contrary, do you feel like the toolkit helped you identify problems you wouldn't have considered otherwise?
- Toolkit comparison
  - If they have experience using both Fairlearn and AI Fairness 360, what are the differences between the two?
  - Was there a reason why you chose to learn how to use one toolkit over another?
  - When trying to explore the harms in a dataset, do you choose a specific toolkit to help you do so? If so, why would you choose one toolkit over another?

C Appendix C - Summary of findings

	Awareness	Understanding	Identification	Actionability
Task 1. Fairlearn	G - 8/10 NA - 2/10	G- 8/10 NA - 2/10	G - 8/10 NA - 2/10	G - 5/10 S - 3/10 NA - 2/10
2. AIF360	G - 4/10 S - 3/10 NA - 3/10	G - 4/10 S - 3/10 NA - 3/10	G - 4/10 S - 3/10 NA - 3/10	G - 3/10 S - 4/10 NA - 3/10
3. No experience	G - 8/10 NA - 2/10	G - 8/10 NA - 2/10	G - 8/10 NA - 2/10	G - 7/10 S - 1/10 NA - 2/10
Irrelevant/Questionable attributes 1. Fairlearn	G - 9/10 NA - 1/10	G - 9/10 NA - 1/10	G - 9/10 NA - 1/10	G - 5/10 S - 4/10 NA - 1/10
2. AIF360	G - 5/10 S - 2/10 NA - 3/10	G - 5/10 S - 2/10 NA - 3/10	G - 5/10 S - 2/10 NA - 3/10	G - 4/10 S - 2/10 NA - 4/10
3. No experience	G - 9/10 NA - 1/10	G - 9/10 NA - 1/10	G - 6/10 S - 4/10	G - 5/10 NA - 5/10
Oversimplified attributes 1. Fairlearn	G - 1/10 S - 1/10 NA - 8/10	G - 1/10 S - 1/10 NA - 8/10	G - 1/10 S - 1/10 NA - 8/10	G - 1/10 S - 1/10 NA - 8/10
2. AIF360	G - 1/10 NA - 9/10	G - 1/10 NA - 9/10	G - 1/10 NA - 9/10	G - 1/10 NA - 9/10
3. No experience	G - 10/10 (7 after level 2, 2 by themselves both use cases, 1 by themselves after toolkit)	G - 10/10 (7 after level 2, 2 by themselves both use cases, 1 by themselves after toolkit)	G - 10/10 (7 after level 2, 2 by themselves both use cases, 1 by themselves after toolkit)	G - 8/10 S - 1/10 NA - 1/10
Sensitive attributes 1. Fairlearn	G - 10/10	G - 10/10	G - 10/10	G - 10/10
2. AIF360	G - 9/10 NA - 1/10	G - 9/10 NA - 1/10	G - 9/10 NA - 1/10	G - 9/10 NA - 1/10
3. No	G - 10/10 (5 after	G - 10/10 (5 after	G - 10/10 (5 after	G - 10/10 (5 after

experience	toolkit, 5 by themselves for both cases)	toolkit, 5 by themselves for both cases)	toolkit, 5 by themselves for both cases)	toolkit, 5 by themselves for both cases)
Proxies / Correlations 1. Fairlearn	G - 10/10 (3 after level 2, 7 by themselves)	G - 10/10 (3 after level 2, 7 by themselves)	G - 10/10 (3 after level 2, 7 by themselves)	G - 10/10 (3 after level 2, 7 by themselves)
2. AIF360	G - 6/10 S - 1/10 NA - 3/10	G - 6/10 S - 1/10 NA - 3/10	G - 6/10 S - 1/10 NA - 3/10	G - 3/10 S - 4/10 NA - 3/10
3. No experience	G - 9/10 (2 after toolkit, 7 by themselves for both cases) NA - 1/10	G - 9/10 (2 after toolkit, 7 by themselves for both cases) NA - 1/10	G - 9/10 (2 after toolkit, 7 by themselves for both cases) NA - 1/10	G - 9/10 (2 after toolkit, 7 by themselves for both cases) NA - 1/10
Incorrect labels 1. Fairlearn	G - 9/10 (all after level 2) NA - 1/10	G - 9/10 (all after level 2) NA - 1/10	G - 9/10 (all after level 2) NA - 1/10	G - 7/10 (all after level 2) NA - 3/10
2. AIF360	G - 3/10 (all after level 2) NA - 7/10	G - 3/10 (all after level 2) NA - 7/10	G - 3/10 (all after level 2) NA - 7/10	G - 3/10 (all after level 2) NA - 7/10
3. No experience	G - 7/10 (all after level 2) NA - 3/10	G - 7/10 (all after level 2) NA - 3/10	G - 7/10 (all after level 2) NA - 3/10	G - 7/10 (all after level 2) NA - 3/10
Over/under representation 1. Fairlearn	G - 8/10 (by themselves) S - 2/10 (after level 2)	G - 8/10 (by themselves) S - 2/10 (after level 2)	G - 8/10 (by themselves) S - 2/10 (after level 2)	G - 8/10 (by themselves) S - 2/10 (after level 2)
2. AIF360	G - 9/10 (by themselves) S - 1/10 (after level 2)	G - 9/10 (by themselves) S - 1/10 (after level 2)	G - 9/10 (by themselves) S - 1/10 (after level 2)	G - 9/10 (by themselves) S - 1/10 (after level 2)
3. No experience	G - 10/10 (by themselves)	G - 10/10 (by themselves)	G - 10/10 (by themselves)	G - 10/10 (by themselves)
Over/under sampling	G - 6/10 (by	G - 6/10 (by	G - 6/10 (by	G - 6/10 (by

1. Fairlearn	themselves) NA - 4/10	themselves) NA - 4/10	themselves) NA - 4/10	themselves) NA - 4/10
2. AIF360	G - 6/10 (by themselves) NA - 4/10	G - 6/10 (by themselves) NA - 4/10	G - 6/10 (by themselves) NA - 4/10	G - 6/10 (by themselves) NA - 4/10
3. No experience	G - 9/10 (by themselves) NA - 1/10	G - 9/10 (by themselves) NA - 1/10	G - 9/10 (by themselves) NA - 1/10	G - 9/10 (by themselves) NA - 1/10
Handling of Missing data 1. Fairlearn	G - 8/10 (3 after level 2, 5 by themselves) S - 2/10	G - 8/10 (3 after level 2, 5 by themselves) S - 2/10	G - 8/10 (3 after level 2, 5 by themselves) S - 2/10	G - 6/10 S - 2/10 NA - 2/10
2. AIF360	G - 6/10 S - 2/10 NA - 2/10	G - 6/10 S - 2/10 NA - 2/10	G - 6/10 S - 2/10 NA - 2/10	G - 6/10 S - 2/10 NA - 2/10
3. No experience	G - 10/10 (9 by themselves, 1 after level 2)	G - 10/10 (9 by themselves, 1 after level 2)	G - 10/10 (9 by themselves, 1 after level 2)	G - 9/10 NA - 1/10
Handling of Outliers 1. Fairlearn	G - 10/10 (6 after level 2, 4 by themselves)	G - 10/10 (6 after level 2, 4 by themselves)	G - 10/10 (6 after level 2, 4 by themselves)	G - 7/10 S - 2/10 NA - 1/10
2. AIF360	G - 3/10 S - 3/10 NA - 4/10	G - 3/10 S - 3/10 NA - 4/10	G - 3/10 S - 3/10 NA - 4/10	G - 3/10 S - 3/10 NA - 4/10
3. No experience	G - 9/10 (5 after level 2, 1 after level 1, 3 by themselves) NA - 1/10	G - 9/10 (5 after level 2, 1 after level 1, 3 by themselves) NA - 1/10	G - 9/10 (5 after level 2, 1 after level 1, 3 by themselves) NA - 1/10	G - 9/10 (5 after level 2, 1 after level 1, 3 by themselves) NA - 1/10
Handling of Duplicates 1. Fairlearn	G - 7/10 (after level 2) NA - 3/10	G - 7/10 (after level 2) NA - 3/10	G - 7/10 (after level 2) NA - 3/10	G - 7/10 (after level 2) NA - 3/10
2. AIF360	NA - 10/10	NA - 10/10	NA - 10/10	NA - 10/10
3. No experience	G - 10/10 (7 after level 2, 3 by themselves for both cases)	G - 10/10 (7 after level 2,3 by themselves for both cases)	G - 10/10 (7 after level 2,3 by themselves for both cases)	G - 9/10 NA - 1/10

Choices in model building 1. Fairlearn	G - 8/10 (2 after level 2, 6 by themselves) NA - 2/10	G - 8/10 (2 after level 2, 6 by themselves) NA - 2/10	G - 8/10 (2 after level 2, 6 by themselves) NA - 2/10	G - 6/10 S - 1/10 NA - 3/10
2. AIF360	G - 7/10 NA - 3/10	G - 7/10 NA - 3/10	G - 7/10 NA - 3/10	G - 4/10 S - 3/10 NA - 3/10
3. No experience	G - 9/10 (by themselves for both cases) NA - 1/10	G - 9/10 (by themselves for both cases) NA - 1/10	G - 9/10 (by themselves for both cases) NA - 1/10	G - 9/10 (by themselves for both cases) NA - 1/10
Bias mitigation 1. Fairlearn	G - 8/10 NA - 2/10	G - 8/10 NA - 2/10	G - 8/10 NA - 2/10	G - 8/10 NA - 2/10
2. AIF360	G - 7/10 S - 2/10 NA - 1/10	G - 7/10 S - 2/10 NA - 1/10	G - 7/10 S - 2/10 NA - 1/10	G - 7/10 S - 2/10 NA - 1/10
3. No experience	G - 6/10 S - 3/10 NA - 1/10	G - 6/10 S - 3/10 NA - 1/10	G - 6/10 S - 3/10 NA - 1/10	G - 6/10 S - 3/10 NA - 1/10
Environmental impact 1. Fairlearn	G - 7/10 (after level 2) NA - 3/10	G - 7/10 (after level 2) NA - 3/10	G - 7/10 (after level 2) NA - 3/10	G - 7/10 (after level 2) NA - 3/10
2. AIF360	G - 5/10 (after level 2) NA - 5/10	G - 5/10 (after level 2) NA - 5/10	G - 5/10 (after level 2) NA - 5/10	G - 5/10 (after level 2) NA - 5/10
3. No experience	G - 5/10 (after level 2) NA - 5/10	G - 5/10 (after level 2) NA - 5/10	G - 5/10 (after level 2) NA - 5/10	G - 5/10 (after level 2) NA - 5/10
Harms in broader environment 1. Fairlearn	G - 7/10 (after level 2) NA - 3/10	G - 7/10 (after level 2) NA - 3/10	G - 7/10 (after level 2) NA - 3/10	S - 4/10 (after level 2) NA - 6/10
2. AIF360	G - 3/10 (after level 2) NA - 7/10	G - 3/10 (after level 2) NA - 7/10	G - 3/10 (after level 2) NA - 7/10	S - 2/10 (after level 2) NA - 8/10
3. No experience	G - 8/10 (after level 2)	G - 8/10 (after level 2)	G - 8/10 (after level 2)	S - 7/10 (after level 2)

	NA - 2/10	NA - 2/10	NA - 2/10	NA - 3/10
Choices of metrics + Too large dependency on metrics + Parity only 1. Fairlearn	G - 8/10 S - 2/10	G - 8/10 S - 2/10	G - 8/10 S - 2/10	G - 8/10 S - 2/10
2. AIF360	G - 7/10 S - 3/10	G - 7/10 S - 3/10	G - 7/10 S - 3/10	G - 7/10 S - 3/10
3. No experience	G - 10/10	G - 10/10	G - 10/10	G - 10/10 (increased after toolkit)

G = great; S = slightly; NA = not discussed / not understood