Comparison of the usage of Fairness Toolkits amongst practitioners:AIF360 and Fairlearn

Harshitaa Pandey
Supervisor(s): Agathe Baylan, Ujwal Gadiraju, Jie Yang
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

Comparison of the usage of Fairness Toolkits amongst practitioners:AIF360 and Fairlearn

Written by Harshitaa Pandey
Under the supervision of Agathe Baylan, Ujwal Gadiraju, Jie Yang
TU Delft

Abstract

Machine learning is still one of the most rapidly growing fields, and is used in a variety of different sectors such as education, healthcare, financial modeling etc(Jordan and Mitchell 2015). However, along with this demand for machine learning algorithms, there comes a need for ensuring that these algorithms are fair and contain little to no bias. Tools like Fairlearn[1] and AI Fairness 360(AIF360)[2] allows developers and data scientists to examine their codebase according to specified fairness metrics and mitigate any fairness related issues. This study aims to determine how practitioners use the separate toolkits and whether their practices are differ by the toolkit they choose to use. To do this, we conducted 29 think-aloud interviews with industry practitioners to understand how they would use Fairlearn and AIF360 in practice. The results show that fairness is a socio-technical challenge. While the toolkit does allow for participants to be structured in their approach, and raise awareness for fairness related harms, at the end of the day the fairness toolkit only provides technical help to find harms that the individual was already aware about. Based on the findings, we then suggest the design for a fairness toolkit that can help practitioners approach fairness in the most ideal manner. This toolkit would include a way to have interdisciplinary collaboration, have a larger focus on explainability, and give clear guidance to its users regarding fairness related harms.

Introduction

Machine learning has been able to transform many aspects of our life. Whether it is altering our daily life by providing access to advanced recommender systems such as Netflix(Steck 2022), or helping us manage network congestion control(Ye et al. 2018), it has become a prevalent part of our current society. However, along with the benefits, some new challenges are introduced during the process as well. This is where the concept of fairness is introduced, as algorithms which have impacts on people's lives need to be monitored to ensure that they are making their judgements in an ethical manner and function with little to no bias. Determining whether an algorithm is fair, is a difficult task as the definition of fair itself is unclear.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]https://fairlearn.org/

[2]https://aif360.mybluemix.net/

Difficulty of Defining Fairness

There are at least 21 different definitions of what fairness could potentially mean(Narayanan 2018). Each definition has different metrics to measure it, and different algorithms to mitigate it(Garg, Villasenor, and Foggo 2020). This might suggest that fairness can be simplified into one or several of the mathematical descriptions or metrics. Then, the main concern would be to satisfy said metrics. However this is not the case as fairness of an algorithm is dependent on the use-case and the stakeholders involved. This is because it is impacted by both social and technical issues(Dolata, Feuerriegel, and Schwabe 2021). Therefore in order to approach it properly, it needs to be viewed as a socio-technical issue in which both aspects need to be considered.

This socio-technical perspective can be defined as finding a way to incorporate the social component(structures,cultures,economic systems etc.) along with the technical component(software, hardware, data sources etc.)(Dolata, Feuerriegel, and Schwabe 2021) to fully understand how to define what is fair in the scenario and then to understand where the bias might lie and how to properly mitigate it.

Importance of Investigating Fairness

While defining fairness is a difficult task, accounting for fairness is an extremely important issue. As mentioned above, machine learning has been introduced in many aspects of our current society. Some of these domains in which machine learning is being used can have major repercussions on an individuals life. Examples of two domains where this might be the case are credit scoring(Dastile, Celik, and Potsane 2020), and criminal justice forecasting(Berk 2012). A well known example of when negative repercussions occurred due to there not being a proper analysis of fairness was with the tool COMPAS. This tool was supposed to assess if a participant should be discharged or not. In practice, it was found to incorrectly classify black defendants as "high risk" substantially more than the white defendants(Corbett-Davies et al. 2017). While this algorithm was intially made in the 1990's [3], it is still a point of debate to this day. Machine Learning experts are still trying to find the right defini-

[3]https://www.uclalawreview.org/injustice-ex-machina-predictive-algorithms-in-criminal-sentencing/

tion of fairness in this scenario along with the right metrics to detect it(Wang et al. 2022).

Introduction of Fairness Toolkits

Due to the difficulty of assessing and mitigating fairness, fairness toolkits were developed. These toolkits aim to aid ML practitioners in evaluating and mitigating fairness within their systems. There are more than 6 open-source toolkits which have been developed for this exact reason(Lee and Singh 2021). Due to the importance of the matter, major corporations such as IBM[4], Microsoft[5] have all been involved with developing their own toolkits. The two most most popular toolkits are Fairlearn[6] and AI Fairness 360(AIF360)[7]. Fairlearn is an open source toolkit which contains 17 bias detection metrics as well as 4 mitigation algorithm(Bird et al. 2020). AIF360 is also an open source toolkit which consists of over 71 bias detection metrics as well as 9 mitigation techniques.

Reasoning Behind the Research

Due to the importance of this topic(Hutchinson and Mitchell 2019), there is quite a bit of research done on the toolkits available on the market(Lee and Singh 2021). However, there does seem to be a lack of research on comparing how the two toolkits are being used by practitioners. The sole article which discusses this exact topic, only takes into account practitioners with no prior toolkit knowledge(**?**). Also, the majority of the data collected was from an online survey which would not clearly demonstrate how practitioners use the toolkit in practice.

The reasoning behind why this topic is important is because of the already mentioned repercussion of an unfair algorithm. No single practitioner will have the domain knowledge to conduct a full fairness analysis(Amershi et al. 2019). Since we also know that there is no single definition of fairness, it is of utmost importance to guide these practitioners to consider the "socio-technical" perspective of harms in a structured manner. Understanding how practitioners use the toolkit, will help developers in the future to design toolkits according to what is actually required from the practitioner. It is also important to ensure that these toolkits are not being designed in a way which constrains the view of fairness, as it was discussed earlier that it is a broad topic and the definition that needs to be used is dependent on many factors in the use case. Due to this reasoning, the following research question for this paper was formulated;

"To what extent are practices for practitioners who use fairness toolkits fragmented by the different fairness toolkits?"

Within this paper, we will conduct 29 think-aloud interviews to understand how practitioners use the two different toolkits. We first designed realistic use-cases with which practitioners could develop a ML model. During each interview, we talked in depth with the practitioners about how they would mitigate and detect harm in each stage of the ML pipeline. We asked them to determine which metrics and mitigation techniques they might use, and also asked them to give explanations as to why they would do so.

After the investigation, we determined that there were several aspects of the toolkit which could be changed to help the practitioners. We determined the need for interdisciplinary collaboration, a need for explainability and clear guidance from the toolkit.

Related Work

In this section, there will be a discussion regarding the previously mentioned article with a similar research question along with a more detailed description of the fairness toolkits.

Previous Research

The article which discusses the concept which we are trying to investigate in this paper was already mentioned, and it aims to find how ML practitioners use fairness toolkits. To reiterate, the main differences are that these practitioners have no prior knowledge of the toolkits and most of the data is collected through online questionnaires. The two main outcome of this research was(Deng et al. 2022);

- Practitioners need to be guided through the process of eliminating harms, and that future toolkit designer should aid users in not committing the ML fairness pitfalls.
- Toolkits need to support interactions and collaborations across all roles in the industry.

Detailed Description of Fairness Toolkits

ML toolkits consist of both bias mitigation algorithms as well as bias metrics. These metrics allow for a user to asses how much fairness related bias there might be in their ML system. The mitigation algorithms allow for the user to then decrease bias after it has been detected by the metric. The open-source nature of the current most popular toolkits, allows for the user to clearly understand how those metrics and mitigation algorithms were developed to ensure that they are using them in the proper manner.

Fairlearn Initially developed in 2018 by MicrosoftResearch as an aid for a research paper(Agarwal et al. 2018). It provides users help with regards to not only mitigating and detecting fairness but also how to talk and write about fairness as well[8]. It also has a multitude of resources that help aid in raising awareness of the difficulty in assessing fairness and guiding users to the proper ways to do so. In fact, it also addresses the socio-technical challenge which was mentioned above[9]. The development team holds weekly discord calls in which all users are allowed to join and address any concerns/bugs.

AIF360 AIF360 was developed by IBM, and it is clear in stating that it helps users in detecting/mitigation bias in all stages of the ML pipeline(pre-processing, in-processing and post-processing). It's a comprehensive toolkit which also

[4]https://aif360.readthedocs.io/en/latest/modules/metrics.html

[5]https://github.com/microsoft/responsible-ai-toolbox

[6]https://fairlearn.org/

[7]https://aif360.mybluemix.net/

[8]https://fairlearn.org/v0.7.0/contributor$_guide/how_to_talk_about_fairness.htm$

[9]https://fairlearn.org/main/user$_guide/fairness_in_machine_learning.html$

contains a demo on how to use it as well as other example Jupyter notebooks which can be studied by users. It contains an R package alongside a Python package to cater to a wide range of practitioners. The toolkit states "should only be used in a very limited setting: allocation or risk assessment problems with well-defined protected attributes.[10]".

Method

The goal of this research is to determine whether the practices for practitioners who use fairness toolkits are fragmented by different fairness toolkits. In order to do this, we conducted 29 interviews with ML specialists with varying degrees of knowledge. The exact process we used is described in detail below.

Participant Recruitment

To recruit participants, we initially used social networks such as Linkedin, Slack and Discord to directly contact individuals we knew with some ML knowledge. These consisted of professionals both in Academia or working in the Industry with some Machine Learning knowledge. We asked these individuals what their level of expertise and domain of knowledge was, to ensure that we had a varied group of interviewees. Alongside this, we also used the snowball sampling strategy and asked the individuals to recommend other professionals who could potentially be interviewed as well. Thereafter, they were invited for an in-depth structured interview on fairness in ML.

10 participants had prior with experience in Fairlearn. Then, 9 participants with experience in AIF360 and finally 10 participants with no experience in either toolkit. Table 1 and Table 2 shows the details of the 19 individuals who were asked to participate in the interview and who had prior experience with the toolkit. This table shows some details about the participants along with their background. A description of their experience with either toolkit has been described as well which is split into three categories. An individual is given "expert" if their experience with ML is 5+ years or if they are an active contributor to the toolkit. "Proficient" is given when an individual has been a contributor in the past, or has around 2 years of ML experience. "Advanced beginner" has been assigned to the remaining participants who have prior knowledge of the toolkits but do not fit into the previous criteria. Since the most details are required from the participants with knowledge of the toolkits, the rest of the participants and their details are shown in Section 1 of the Appendix.

Interview Guide

For the interview itself, we used a think-aloud semi-structured interview approach. Before conducting the interview, we determined a set of questions we were going to ask the participants, as well as the two use-cases we would use. The exact questions are shown in Section 2 in the Appendix, and the use-cases are described in Section 2 as well. The layout of the interviews was different for participants with prior knowledge of the toolkits and the ones without. While the interview took one hour for participants with prior knowledge of the toolkits, in the case of a practitioner with no prior knowledge, the interview lasted for 2 hours. There were 19 interviews of participants with prior knowledge and 10 interviews with practitioners who had no prior knowledge of the toolkits. The process is described in more detail below.

[10]https://aif360.mybluemix.net/resourcesguidance

Individuals with Experience with Fairlearn				
	Educational Background	Location	Current Role	Experience Level with Fairlearn
P1	Industrial Engineering and Computer Science	Netherlands	Academia(Professor), and Research Engineer	Expert
P2	Accountant and Ethics	India	Bias researcher	Expert
P3	Computer Science	Canada	Academia(Student)	Experienced Beginner
P4	Computer Science	Netherlands	Academia(Student)	Proficient
P5	Computer Science	Netherlands	Academia(Student)	Proficient
P6	Computer Science	United States	Data Scientist	Expert
P7	Computer Science	United States	Academia(Student)	Experienced Beginner
P8	Computer Science	Netherlands	Principal Data Engineer	Proficient
P9	Computer Science	Canada	Data Scientist	Proficient
P10	Computer Science	United States	Data Scientist	Experienced Beginner

Figure 1: Participants with Fairlearn Experience

Individuals with Experience with AIF360				
	Educational Background	Location	Current Role	Experience Level with AIF360
P1	Computer Science	Netherlands	Academia(PhD)	Experienced Beginner
P2	Economics and Computer Science	Netherlands	Academia(PhD)	Expert
P3	Computer Science	Brazil	Data Scientist	Proficient
P4	Business and Engineering	Canada	Academia(Professor) and Investment Strategist	Proficient
P5	Electrical Engineering	India	Academia(Researcher)	Proficient
P6	Robotics	Turkey	Data Scientist	Proficient
P7	Mechanical Engineering	India	Data Scientist	Expert
P8	Computer Science	Netherlands	Academia(Student)	Expert
P9	Data Science	Netherlands	Academia(PhD)	Experienced Beginner

Figure 2: Participants with AIF360 Experience

Participants with no prior knowledge of the toolkits This interview consisted of an initial exploration of Use-case 1 by the participant. Afterwards, five participants were given a demo to Fairlearn and the remaining five were given a demo to AIF360. The process after this point was the exact same as the process for the participants with prior knowledge as described below.

Participants with prior knowledge of the toolkits This interview was separated into four sections. Even though ten participants were using Fairlearn and ten participants were using AIF360, the layout of the two interviews was the same. The first section is regarding the background questions we have for the participants. The second section is regarding the open exploration of the use-case provided. The third section is for asking them clarifying questions on the harms they mentioned, as well as having a larger discussion on the harms they did not. The fourth and final section is regarding their ending thoughts on the toolkits themselves.

Before conducting the interviews, we asked the participant if they would be willing for us to record and transcribe the interview. We also provided them with an ethics form in which we asked for their approval on us using the content from the interview. This is given in Section 3 in the Appendix. In total we collected approximately 40 hours of recorded interviews from the 29 different sessions.

Coding Process

The way in which the qualitative data was analyzed was using a mix of deductive and inductive coding processes. For the deductive coding process, we determined some baseline codes which could be used as a basis for analyzing the results. These codes were only a small part of the final codes which were collected, and were regarding the harms found in the use-case. The rest of the codes were developed using an inductive coding process. In this process, there are no predetermined codes against which you analyze your qualitative data. The codes are developed using a bottom-up approach as they're determined as you analyze the interviews(Thomas 2006).

The first step to this approach was to transcribe all the interviews from the video recordings. During this process, the initial concept of the codes which would be used was already created. After the initial codes had been created, we went through each interview thoroughly and used the tool ATLAS.ti[11] to determine where in the transcript certain codes were spoken. As inductive coding is an iterative process, we encountered other interesting codes I might want to include in the analysis as we analyzed the transcriptions. If we determined that we wanted to include the new code, we would go back to the already analyzed transcripts and check for the code there too.

The final codes we created are shown in Section 4 of the Appendix. After creating these codes, we went through the interviews a final time to pick out any insightful or important quotes from the transcriptions. After compiling these into understandable themes and patterns, the stage of showing the results was reached.

Results

Usage of Fairlearn by experienced practitioners

The Fairlearn package itself has two components. Firstly, there are the 17 fairness related metrics it provides, and secondly there are the 4 mitigation algorithms which help the user mitigate any fairness related issues. In order to understand how each of the two components were used, we will talk about both separately. In this section, all the participants mentioned will be referenced according to their number given in Figure 1.

Before looking at the specifics of metrics and mitigation algorithms it's important to discuss some general topics which were brought up by the participants.

General

- The first general comment was that all the participants wanted to make it clear that while they were describing their process to us, they would be involving domain experts at many different stages of the cycle. For example, when P1 was asked to describe their pre-processing of the data, they said "I would work together with a domain expert on this." P6 mentioned that they would also use knowledge of the expert while figuring out which metrics are most applicable. P8 also mentioned that as a Principle Data Engineer, they would also need to involve many other stakeholders in this process such as data scientists, domain experts and people from the business.
- The second general comment is related to that as 3 out of the 10 individuals[P6,P8,P9] mentioned how this tool is important as it is easy to use for people in the business who do not have much technical knowledge. Those comments can be summed up in those quote by P6. "Having a toolkit like this allows for the business to understand Fairness in ML. [With the metrics] they can see what it consists of." What was interesting here was that the individuals who mentioned the business are all currently in the industry and they all have 2+ years of experience.
- The third general point was that 5 out of the 10 participants[P1,P2,P6,P8,P9] mentioned the fact that Fairlearn holds weekly community call in which they can address their concerns. When P2 was asked why they chose to use Fairlearn over other toolkits, they responded with "I can get help for my questions. The community is active."
- The next point to make is regarding the design choices that Fairlearn has made. P6 stated "Fairlearn's design choices are more deliberate and it shows what they suggest users to do." They also stated "I really like it[Fairlearn] because it builds up on Scikit learn, making it easy to use". This sentiment was echoed by [P1,P2,P4,P4,P7,P9].
- The last comment to make is regarding when participants chose to use Fairlearn over AIF360. There were only four participants[P1,P2,P8,P9] who claimed to have some knowledge of AIF360 as well. When P1 was asked why they chose to use Fairlearn, they responded with "Have you ever tried to import data into AIF360?". This was done in a manner which obviously insinuated that this process was tedious and unnecessary. [P2,P8,P9] all stated that the reason they would choose one over the other was how easily one would be able to integrate with their current process. This was clearly seen in this quote by P2 "They are more or less the same[Fairlearn and AIF360]. Most of this type of work is in Colab, then I would go for using Fairlearn because it's convenient and easier. If using a low-code environment, then I add AIF360 as its convenient for me to use this as a plugin".

Metrics

When it comes to the actual usage of the metrics, we saw that 6 out of the 10 participants mentioned that they would use the metrics provided by Fairlearn[P1,P2,P5,P6,P8,P9]. One participant refrained from mentioning any specific metric and said "I wouldn't have the best idea on what metric to use myself. Maybe a doctor would know best[in this use-case]." The other three individuals expressed similar views on the Fairlearn metrics which can be demonstrated with

[11] https://atlasti.com/

this quote "Yes, I could use Fairlearn capabilities, but I just use scikit-learn[12]. I'm more used to that." Interestingly, out of the 17 metrics Fairlearn provides, individuals who did choose to use Fairlearn metric used a combination of 5 of them. [P1,P2,P5,P6] all used the same combination of using FPR, TPR, FNR, and selection_rate as their metrics. What is interesting to note is that [P1,P2,P6] were classified as an "Expert" while P5 was classified as "Proficient". The other metrics which was used by both P8 and P9 was demographic_parity_difference. Again, we saw that both of these individuals was classified as "Proficient".

Mitigation Algorithms

Only 5 out of the 10 participants mentioned that they would use one of the provided mitigation algorithms[P2,P5,P6,P8,P9]. Out of the 4 bias mitigation algorithms, three of them were mentioned. While P6 did not mention any specific algorithms that they would use, they did mention the frequency with which they use these tools in this quote "I use the stuff[bias mitigation algorithms] that Fairlearn provides quite often. Especially when working with non-profit partners or other teams. The first thing we try are Fairlearn's mitigation strategies, if those don't work that's when we try other things." Three[P2,P5,P8] of the five individuals mentioned that they would use ThresholdOptimizer. The other two mentioned using Gridsearch[P6,P9]. What was interesting here was that out of these five individuals, [P5,P8,P9] were classified as "Proficient" and [P2,P8] were classified as "Expert".

While it was interesting to see what mitigation algorithms the participants chose to use, what was most interesting in this section, were the quotes from the participants who refrained from using any mitigation algorithms. While in the metrics section, all the participants who did not use any metrics from Fairlearn were still open to using them in the future, some of the participants who refrained here were very adamant that with the current state of the algorithms they would not be using them. P1 who was deemed an "Expert" stated "the Fairness Mitigation algorithms are not at the stage where I think they should be", they then went on to say "I am a fan of thorough assessment rather than blindly optimizing for something". P4 mentioned "I don't think there are that many for now, the few that are there are look nice but I know a few are lacking." Even though P8 mentioned that they would use Threshold Optimizer, they followed up with "But it's not just about ticking checkboxes. What I see in practice is that a real life dataset is different to the training dataset. I could use Threshold Optimizer but it is optimized for my training data, and when it will be used in real life it could produce really weird results and will need to be recalibrated."

Algorithmic harms The algorithmic harms were split up into four different sections:task, dataset and transformations,building of models and evaluation of models. In this section, we will give some idea of what was discussed per section but To see the accumulated algorithmic harms each participant picked up in detail please refer to this link.

Task

There didn't seem to be much of a focus of the participants in this area. 1 out of the 3 harms was picked up by 8 participants. This harm was called *Undesired Task*. Even them, 3 of the participants[P5,P6,P7] had been directly asked about it before giving an opinion on the harm.

Dataset and Transformations

There were three sub sections in this area:data attributes,data population and data errors.

In the data attributes section, *proxies* was the harm which was most referred to as it was mentioned by all of the participants. There were two harms in this section which were only discussed by two participants. The first was *oversimplified attributes* which was discussed by P1 and P6. The second was *attributes transformation* which was discussed by P1 and P5.

In the data population section, 3 out of 4 harms were extensively mentioned. However one of those harms(*Incorrect Labeling*) had 8 out of 9 participants who had be asked about the topic before they gave a reflection. The one harm which was less mentioned in this section was *Concept Drift and Covariate Shift* as only three participants picked it up[P2,P8,P9].

Interesting, the data errors section was the section was the most complete. *Duplicates* was the harm which was mentioned by the least number of participants, and all 7 participants who gave a reflection on the matter had to be directly asked to do so. All the harms were mentioned by at least 9 participants.

Building of Models

Again, this section was mentioned by quite a few participants. The choice of model was discussed by all participants besides P5. P5 said "harm is caused by data not the model".

Evaluation of Models

Once again, this section was also a focus of the conversation for many participants. 4 out of 6 harms listed here were mentioned by at least 8 participants. The two harms which were least mentioned are *Output vs Outcome* and *People outside predictions*.

Usage of AIF360 by experienced practitioners

The AIF360 package once again includes the two components mentioned above. It has 71 bias detection metrics and 9 bias mitigation algorithms. Once again to understand in detail how practitioners use both of these things, we will talk about them separately. In this section, all the participants mentioned will be referenced according to their number given in Figure 2.

General Again, before diving int metrics and mitigation algorithms, we will discuss some more general comments made by the participants.

- The first remark to make was regarding participants mentioning their preference towards using R instead of Python to do this kind of analysis. [P1,P9] were the participants who mentioned this. P1 stated simply "This kind of exploration I would do in R, mostly because I find that modelling works better in Python but I find data exploration and data visualization goes better in R", when presented with the Jupyter notebook with the use-case.

[12]https://scikit-learn.org/stable/

- The second comment was about how domain experts are mentioned throughout the exploration. Once again, all participants mentioned this aspect. Specifically, many stated that they would involve medical experts during the data exploration stage to verify if the data they had made sense.
- The third comment was about how participants felt that the toolkit was intended for a larger group of stakeholders including the business who are not familiar with ML. This is seen in the comment from P1 saying "Toolkits are really helpful for people who might not be super technical,I sometimes talk to business people from IBM and this is how they connect to the topic from a technical side".
- The last comment to make was regarding how many individuals chose to speak about how "comprehensive" AIF360 is. P5 stated directly "AIF360 is comprehensive and it accounts for a lot of things". The same sentiment was shared by [P2,P3,P5].

Metrics

4 out of the 9 participants[P2,P3,P5,P8] mentioned that they would use the metrics provided by AIF360. All 4 of the participants mentioned using statistical_parity_difference. 3 out of the 4 participants[P2,P3,P5] also mentioned using desperate_impact. There were only two more metrics which were mentioned and they were both by P3. This participant mentioned using equal opportunity difference along with average odds difference as well. What should also be mentioned here was the discussion regarding where to use metrics. P2 a participants labeled as "Expert" mentioned wanting to use metrics for assessing both the pre-processing and post-processing stage. P3 a participant labeled as "Proficient" also stated something similar and said "I would apply metrics in all stages: pre-processing,in-processing and post-processing". Even participants who chose to refrain from using the AIF360 metrics mentioned the importance of metrics at different stages. P6 said that they think "metrics computed before model training are the most important" and P9 also mirrored this sentiment.

Mitigation Algorithms

Once again 4 out of the 9 participants[P2,P3,P5,P8] mentioned that they would use the metrics provided by the toolkit. Immediately, we can see that the participants who said that they would use the mitigation algorithms provided, are also the same ones who mentioned that they would use the metrics provided. Even when looking at it in more detail, [P2,P3,P5] still mentioned the same mitigation algorithms as they all said they would use Reweighing and mentioned the importance of bias mitigation in the pre-processing stage. While P5 also shared this sentiment by saying "Pre-processing of the data is where I would intervene the most[when it comes to bias mitigation]". The algorithm mentioned there was PrejudiceRemover which is an in-processing technique. The other algorithms mentioned were DisparateImpactRemover by P2 which is another pre-processing technique. P8 mentioned using AdversarialDebiasing which is a in-processing technique. The participants who chose to not use a bias mitigation algorithm listed similar reasoning which can seen in this quote from P6, "I know AIF360 has tools like Reweighing but I'm not sure how effective they are. Maybe they're introducing bias to the situation". P9 stated that they would do their own PCA analysis and use SHAP values as well which allow them to explain the decisions that their model makes.

Algorithmic harms The algorithmic harms were split up into four different sections:task, dataset and transformations,building of models and evaluation of models. In this section, we will give some idea of what was discussed per section but To see the accumulated algorithmic harms each participant picked up in detail please refer to this link.

Task

There were 2 out of 3 aspects which were mentioned here. The only harm which wasn't discussed by any participant was *Oversimplified objective labels*. *Undesired task* was mentioned by [P1,P3,P4,P5] and *Task that only reproduces historical data patterns* was mentioned by all participants besides P1.

Dataset and Transformations

There were three sub sections in this area:data attributes,data population and data errors.

In the data attributes section, the only harm which wasn't mentioned was *Attributes Transformation*. On the other hand, *Sensitive/Protected Features* was mentioned by all of the participants. *Irrelevant Attributes* was also mentioned by 6 out of 9 participants[P3,P4,P5,P6,P7,P8].

In the data population section, *Over/under representation* was mentioned by all participants. *Concept Drift and Covariate Shift* was mentioned by no participants. [P5,P6] were also the only participants to mention *Incorrect Labels* and only did this when explicitly asked about them.

The data errors section was not a prominent section of discussion. *Duplicates*,*Removal of missing values/outliers/duplicates*,*Outliers* were all discussed by less than 3 participants each. The only aspect discussed in detail here was *Missing Data* which was discussed by 6 out of 9 participants[P1,P3,P4,P5,P6,P7].

Building of Models

The two most discussed harms here were *Choices[of algorithm]* which was discussed by 5 out of 9 participants[P1,P4,P5,P6,P7] and *Bias Mitigation*, which was discussed by 6 out of 9 participants[P1,P2,P3,P5,P6,P7]. The rest of the harms were discussed by less than two participants.

Evaluation of Models

This was another section which was not extensively discussed. The most awareness was about *Incomplete/irrelevant choices of protected attributes* as 5 out of 9 participants[P1,P3,P5,P7,P8] mentioned it. *Parity only* also had 3 out of 9 participants[P1,P5,P6] who discussed it. However, 2 out of those 3 participants[P5,P6] were directly asked about the topic. The rest of the harms has 2 or less participants who mentioned them.

Comparison between how experienced practitioners use Fairlearn and AIF360

When comparing the usage of the two toolkits, there were lots of similarities and differences. These can be seen in the table below;

	Section	
Similarity	General	Users of both toolkits emphasized the importance of involving domain experts and other stakeholders while de-biasing an algorithm.
		Users of both toolkits felt that the "business"(individuals with no prior ML knowledge), could really benefit from a toolkit, as it allows for them to have an understanding of what fairness means with regards to ML.
		Users of both toolkits felt that the toolkit that they had most experience with, was the most easily integratable into their working environment.
	Algorithmic Harms	Building of Models and Evaluation of models was an area which users of both toolkits mentioned extensively.
		All the users of both toolkits also mentioned sensitive/protected features.
Differences	General	AIF360 users mentioned having the need to use R in explorations like this
		Fairlearn users mentioned how their development team holds weekly calls and having this kind of support is the reason they would use this toolkit.
		Fairlearn users described the toolkit to be more deliberate with their design choices and what they recommended users to use.
	Metrics	60% of participants used metrics from Fairlearn. They used FPR,TPR,FNR, selection_rate and demographic_parity_difference as their metrics.
		44.4% of participants mentioned using metrics from AIF360. They used statistical_parity_difference,desperate_impact,equal_oppurtunity_difference,and average_odds_difference
	Mitigation Algorithms	50% used Fairlearn mitigation algorithms. They used ThresholdOptimizer(post-processing) and Gridsearch(In-processing).
		44.4% used AIF360 mitigation algorithms. They used Reweighing(pre-processing), DisparateImpactRemover(pre-processing),PrejudiceRemover(in-processing), and AdversarialDebiasing(in-processing).
	Algorithmic Harms	Users of Fairlearn had a much more detailed analysis of the data. Harms related to data errors were discussed much more than with users of AIF360 as less than 3 participants mentioned data errors with AIF360, while 7 mentioned data errors with Fairlearn.
		There seems to be more analysis on the task with AIF360. 2 out of 3 aspects were mentioned while Fairlearn users only mentioned 1 out of 3.

Figure 3: Comparison of toolkits

Usage of Fairlearn by inexperienced practitioners

In this section, we will discuss how five individuals who had no prior experience of Fairlearn used the toolkit. These practitioners will be referenced according to their number given in Section 1.1 of the Appendix.

Metrics

The metrics were used by 3 out of 5 participants[P1,P3,P4]. P1 stated "Metrics are cool. Demographic Parity is interesting and it looks like it is easy to use". While P3 did not mention any specific metric, they stated "I would definitely use these[metrics]. It's useful from the explainability perspective to have these metrics to assess the influence of features and parameters." P4 simply stated that they would use FPR and FNR as metrics.

The two individuals who did not mention using any metrics had different reasoning for doing so. P2 stated "I would need to look at the mathematical equations and understand fundamentally what is going on before using any functionality." P5 mentioned that they "use the stuff[metrics] they are used to from Scikit-learn".

Mitigation Algorithms

When it came to the mitigation algorithms used, 2[P4,P5] out of the 5 participants mentioned they would use it. Interestingly, both of them mentioned using Gridsearch. P5 mentioned "Pre-processing is the most important part. That's the moment you can introduce or mitigate a lot of bias". However, they did not mention CorrelationRemover, which is Fairlearn's pre-processing tool.

The reasoning behind why 3 participants withheld from using the mitigation techniques all had to do with distrust and inexperience. P1 stated "Someone, somewhere decided what to include in this toolkit. But fairness is subjective. I would not rely on the tools provided here". P2 once again stated that in order for them to use these algorithms they needed to look at the equations they were based on and understand what they are doing better. P3 stated "What I would be interested in would be to reverse engineer these decisions that the toolkit makes. I think explainability is of higher value".

Algorithmic Harms In this section, there will be discussion on the harms which were picked up by each participant and how it changed after using the Fairlearn.

There were two participants[P1,P4] who had a difference in the harms they picked up before and after the toolkit was introduced. For P1, there were quite a few changes that were made after they were introduced to use the toolkit. Firstly, their awareness of sensitive features was increased as they previously had to be asked to define the sensitive attributes but in this use-case they defined them on their own. They also dropped missing values and sensitive features which was not done before. The difference for P4 was in the data pre-processing stage. This quote helps sum up why this was the case "I became aware that bias mitigation should happen in all stages of the pipeline".

[P2,P3,P5] did not have any differences to their approach on harms before and after the toolkit was introduced. For P2,this could have been due to the mistrust that they had with the toolkit, as when asked about what they thought of the toolkit, they answered with "Tools like this probably make it[finding/mitigating harms] more idiot proof but in data science, the devil is in the details". However, when asked directly if their perspective on harms changed, they answered with "I learned how easy it was, and that made my outlook on ethics in data science change". The reasoning behind no difference for P3 can be understood by this quote "Tools didn't help in thinking of the fact that there might be bias, but having some quantitative measure of the difference helped". For P5, the reasoning might be understood from the quote "Someone who is aware of the problem of fairness/bias will do anything to mitigate it, not just use the tool just to I'm done after using it". This quote seems to suggest that the individual thought they had a clear understanding of the "problem" before hand and did not think a tool helped them boraden their understanding.

Usage of AIF360 by inexperienced practitioners

In this section, we will discuss how five individuals who had no prior knowledge of AIF360 used the toolkit. These practitioners will be referenced according to their number given in Section 1.2 of the Appendix.

Metrics

4 out of the 5 individuals[P1,P2,P3,P4] stated that they would use the metrics provided from AIF360. The only participant who refrained from using it said "I should, but no one[in the industry] looks at fairness metrics unfortunately". They then went on to say "I guess I would use the metrics depending on the domain. Choosing the right one requires theoretical knowledge which I do not have". Interestingly, the exact opposite statement was given by 3 participants[P1,P2,P3] who all stated how this toolkit would work well in practice, insinuating that there was a need to do so. This is seen in a statement by P1 saying "I do not think there are any limitations to this toolkit. I think it will work well in practice". [P1,P2,P3] also used the same metric in the interview which was DemographicParity. While P1 solely relied on this metric and said "Demographic parity is a good assessment of fairness", both P2 and P3 used other fairness metrics as well. P2 used disparate_impact_ratio, and both P3 and P4 would use statistical_parity_difference. P4 also made it clear that they would "rely on metrics indeed and make sure to select the right ones."

Mitigation Algorithms 3 out of the 5 participants[P2,P4,P5] mentioned using at least one of the bias mitigation algorithms provided by AIF360. Interestingly, the only participant who refrained from using the metrics[P5] stated that they would "use all of them[when asked about which mitigation algorithm they would use]". They mentioned wanting to have bias mitigation in all three stages of the ML pipeline(Pre-processing, in-processing and post-processing). P2 and P4 chose a pre-processing bias mitigation technique called Reweighing. P2 said "this technique really stood out for me. I will definitely try and use it in my next project." P3 had some reservations to using the bias mitigation algorithms as they did not want the accuracy to be affected. P1 simply did not remember any of the algorithms from the toolkit.

Algorithmic Harms In this section, there will be discussion on the harms which were picked up by each participant and how it changed after using the AIF360.

1 out of 5 participants[P1] had a clear change in the amount of harms which were picked up before and after using the toolkit. The change was in the protected attributes, as they would now also define all demographics as protected. While this was the only clear difference, when asking the participants if their perspective changed, P5 said "Yes, it changed. Before the toolkit, I did not know about algorithmic harms and the metrics used for them".

[P2,P3,P4] all had similar reasoning behind why their perspective didn't change. While all three did have some positives to say about the toolkit, they also mentioned that their perspective was not changed as they already had a good understanding of algorithmic harms.

Comparison between how practitioners with no prior toolkit experience use Fairlearn and AIF360

When comparing the usage of the two toolkits, there were lots of similarities and differences. These can be seen in the table below;

	Section	
Similarity	General	New users of both toolkits emphasized the need for explainability and transparency in the toolkits.
		New users of both toolkits mentioned how important bias mitigation was in all stages of the ML pipeline.
	Metrics	New users of both toolkits used DemographicParity as a metric that they would use for fairness analysis.
	Algorithmic Harms	New users of both toolkits defined sensitive/protected attributes differently after being introduced to the toolkit.
Differences	Metrics	60% of new users of Fairlearn mentioned they would use a metric. They used DemographicParity, FPR and FNR.
		80% of new users of AIF360 mentioned that they would use a metric. They used DemographicParity,disparate_impact_ratio,statistical_parity_difference,
	Mitigation Algorithms	40% of new users of Fairlearn mentioned that they would use mitigation algorithms.They mentioned Gridsearch.
		60% of new users of AIF360 mentioned that they would use mitigation algorithms. They used Reweighing.
	Algorithmic Harms	40% of Fairlearn users had a higher number of harms picked up after being introduced to the toolkit, while AIF360 only had 20%.

Figure 4: Comparison of toolkits(Practitioners with no prior experience)

Discussion

In this section, we will discuss the results and try and understand how practitioners actually use these toolkits and what they would want from them in the future.

A toolkit which allows for collaboration

This point was brought up in Figure 3. Here, we can see that practitioners with experience in either toolkit mention this want of collaboration. This goes back to the idea of fairness being a socio-technical challenge(Dolata, Feuerriegel, and Schwabe 2021), and every individual having a different view on fairness due to a multitude of reasons. These reasons could be their field of expertise, their cultural background(Sambasivan et al. 2021) and a variety of other factors. The domains in which machine learning is being used is vast(Fabris et al. 2022), and having the right group of working together to find any fairness related harms is crucial.

This point is mirrored in several articles. In this, the two most relevant articles are mentioned. (Holstein et al. 2019). The first article states that allowing for interdisciplinary communication and collaboration allows practitioners to escape the "solutionism trap"(Deng et al. 2022). The solutionism trap is referring how practitioners often fall into the trap of relying on the technology to solve their fairness related harms(Selbst et al. 2019a). However, achieving fairness goes back to the idea of it needing to be a combination of social and technical insights.

The second article also discusses the need of having several individuals "in the loop" when mitigating bias from the system(Holstein et al. 2019). It discusses how individuals themselves carry their own bias which could influence what

they consider as fair. This exact sentiment was also seen in our research as participants themselves stated that they think fairness is subjective, and that toolkits like these would also carry their own bias from whoever implemented them.

A toolkit which incorporates explainability at every step

Explainability was a topic which was brought up mostly by the practitioners with no prior knowledge of either toolkit and can be seen in Figure 4. We can see here that in order for these practitioners to start relying on toolkits like this, they need to understand how each algorithm is being implemented and why to justify using it when they are making their own model. Practitioners clearly showed this sentiment when they refrained from using any metrics/mitigation algorithms and stated that in order for them to do so, they need to understand what the mathematical equation is.

Once again, the importance of explainability is seen in other related works(Baniecki et al. 2020a). However, what is even more interesting is that there are already packages that are being made to allow for fairness and explainability to exist in conjunction with each other(Baniecki et al. 2020b). This Python package focuses on combining explainability and fairness into one package and showing it all in a user-friendly and visual manner.

A toolkit which provides clear guidance

Another aspect which was brought up in Figure 3 was how Fairlearn was preferred by some participants due to clear design choices and guidance. We've already seen the complexity of having to define Fairness(Narayanan 2018). However, the bigger concern is not that someone is cautious in their fairness approach but that an individual is overconfident with having completed the "de-biasing process. This is when there is a need for having clear guidance and considerations from the toolkit that you are using. This can be seen in Fairlearn as it provides clear guidelines on how to approach and discuss fairness[13]. This is done while also maintaining that several established rules in Fairness such as the four-fifths rule[(Jones et al. 2020)] might not be valid. There was a study conducted in which 8 different toolkits were compared to see which ones were suggesting this rule. The only toolkit which did not do so was Fairlearn(Watkins, McKenna, and Chen 2022).

Future of Toolkits From this discussion, we can start to formulate what the ideal toolkit would include. The first and main point would be to have a toolkit which allows for interdisciplinary collaboration. The second point would be to allow for the toolkit to integrate explainability in every single section whether pre-processing, in-processing or post-processing. Due to the domain specific definitions of fairness, the most important aspect is to make sure that all of your assumptions when doing the fairness analysis can be analyzed by another individual. This would also be a point of interdisciplinary interaction, as after the algorithm/-model was clearly documented it could be sent around to a larger group of individuals to help mitigate bias even further. Lastly, there is a need for a toolkit to provide clear guidance to the users. There are lots of pitfalls when it comes to checking for fairness in an algorithm. Some of which include "solutionism trap"(Selbst et al. 2019b), "formalism trap"(Selbst et al. 2019b) and "gerrymandering"(Kearns et al. 2018). Guiding your users and helping them avoid these pitfalls is of key importance.

Limitations

Here we will list some limitations of the research that was conducted;

- Some of our participants had prior knowledge of the use-case that was being used. This occurred for 2 participants who had experience with Fairlearn as there was already existing demos on the use-case that we chose to use. This could have affected the harms they mentioned, and the analysis they conducted. However, one thing to be said here is that these were also the individuals who were labelled as "Expert" and were already mentioning all of the algorithmic harms.
- Another aspect to consider was that the participants with prior experience of either toolkit had agreed to do this interview with the knowledge that they would be talking about fairness. This might have skewed some of their answers as they would know that I would be looking for fairness related answers.
- Some of the participants with experience who agreed to do the interview had/were contributing to either Fairlearn or AIF360. Of course, these participants would be less likely to comment on the flaws of their respective toolkit, but in practice we actually noticed that these individuals were very open to constructively analyzing the toolkit.
- Another thing to mention was that the practitioners who were given a demo of the toolkit, were immediately afterwards asked to do the use-case preparation. They were not given time to look through the toolkit documentation themselves and were reliant on what the demo consisted of. The ideal would have been to have a break between the demo and use-case analysis but the meeting was already 2 hours and we were unsure of whether practitioners would agree to more than that.
- The last thing to consider was that not all of the participants were open to coding during the interview. This did lead to some gaps as we did not know exactly how they would have chosen to implement it in practice.

Responsible Research

In this section, a discussion will be held around responsible research practices and how this research maintained those said practices. The topics mentioned are following the standard introduced in the Netherlands Code of Conduct for responsible research[14] Specifically, I will be mentioning the ethical implications that arise from conducting interviews, maintaining data integrity and reproducibility and finally the topic of plagiarism.

[13] https://fairlearn.org/v0.7.0/contributor$_guide/how_to_talk_about_fairness.html$

[14] https://www.universiteitenvannederland.nl/files/documents/Netherlands

Ethical Implications of interviews

Since a major constituent of this project was to conduct the thirty interviews, it was of utmost importance that we considered the ethical implications that these interviews might have had. Firstly, we ensured that each participant was voluntarily recruited and were given a form asking for consent prior to holding the interview itself. This form consisted of;

1. Defining the purpose of the study
2. Explaining the task given in the interview
3. Benefits and risks of participating
4. Procedures for withdrawal
5. Collection and use of personal information
6. Research data and data retention period

As you can see above, we took great care in informing the participants what they would be asked to do along with explaining how we would be treating the data collected from them. We also made sure to include that they were allowed to opt out of the interview at any time, and they simply had to inform us if they wanted any information to be deleted from the recording/transcription.

Another aspect which is relevant here was achieving universalism when it comes to evaluating the research results. Since I did not have a set criteria that I would be evaluating the interviews against, I made sure to clearly explain the method I followed, along with an explanation of why I chose to evaluate the data in that manner.

Data Integrity

Data integrity is a broad concept which refers to several concepts such as data manipulation,fabrication and trimming. When it comes to manipulation of the data or "cherry-picking" , the most applicable of the three concerns above was data trimming. The way I chose to address this concern was being very transparent about the fact that I was picking selective quotes from over 40 hours of interviews. I was also very transparent about the method that I followed to reach these selective quotes as well. If this had not been sensitive data, I would have also liked to include all the transcriptions of each interview as well.

Reproducibility

When it comes to the reproducibility of this research, it is difficult to assess as the majority of the research is dependent on the participants chosen for the interviews. To help with reproducibility I was transparent with mentioning the domain and experience level of every individual chosen in this research. However, it could be that an individual would have a similar domain and experience level with a different outlook on fairness.

While I have tried to be as descriptive with my method as possible, I also provided my contact details in case of an inquiry or recommendations that a reader might have. This is also a publicly available paper and both Fairlearn and AIF360 are open source as well.

Plagiarism

Ensuring that the research conducted is properly used and cited is extremely important Not only is it important because the right author should be given credit but also because it's necessary to leave a trail fr readers in case they might be skeptical of where you are getting your information from. To make sure that this was the case, I cited all the words,texts,processes,results,and arguments in my paper.

Conclusion

This study aimed to understand how practitioners would use Fairlearn and AIF360 in practice. After conducting 29 interviews with the participants we analyzed the data per toolkit to come up with any reoccurring patterns. Afterwards, we used that analysis to understand what was needed from a fairness toolkit to help inform future developers on how to make a toolkit which could support the users in the most idealistic way.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*, 60–69. PMLR.

Amershi, S.; Weld, D.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P. N.; Inkpen, K.; et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, 1–13.

Baniecki, H.; Kretowicz, W.; Piatyszek, P.; Wisniewski, J.; and Biecek, P. 2020a. dalex: Responsible machine learning with interactive explainability and fairness in python. *arXiv preprint arXiv:2012.14406*.

Baniecki, H.; Kretowicz, W.; Piatyszek, P.; Wisniewski, J.; and Biecek, P. 2020b. dalex: Responsible machine learning with interactive explainability and fairness in python. *arXiv preprint arXiv:2012.14406*.

Berk, R. 2012. *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media.

Bird, S.; Dudík, M.; Edgar, R.; Horn, B.; Lutz, R.; Milan, V.; Sameki, M.; Wallach, H.; and Walker, K. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

Dastile, X.; Celik, T.; and Potsane, M. 2020. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91: 106263.

Deng, W. H.; Nagireddy, M.; Lee, M. S. A.; Singh, J.; Wu, Z. S.; Holstein, K.; and Zhu, H. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. *arXiv preprint arXiv:2205.06922*.

Dolata, M.; Feuerriegel, S.; and Schwabe, G. 2021. A sociotechnical view of algorithmic fairness. *Information Systems Journal*.

Fabris, A.; Messina, S.; Silvello, G.; and Susto, G. A. 2022. Algorithmic Fairness Datasets: the Story so Far. *arXiv preprint arXiv:2202.01711*.

Garg, P.; Villasenor, J.; and Foggo, V. 2020. Fairness Metrics: A Comparative Analysis. In *2020 IEEE International Conference on Big Data (Big Data)*, 3662–3666.

Holstein, K.; Wortman Vaughan, J.; Daumé III, H.; Dudik, M.; and Wallach, H. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–16.

Hutchinson, B.; and Mitchell, M. 2019. 50 Years of Test (Un)Fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 49–58. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Jones, G. P.; Hickey, J. M.; Di Stefano, P. G.; Dhanjal, C.; Stoddart, L. C.; and Vasileiou, V. 2020. Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms. *arXiv preprint arXiv:2010.03986*.

Jordan, M. I.; and Mitchell, T. M. 2015. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245): 255–260.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, 2564–2572. PMLR.

Lee, M. S. A.; and Singh, J. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–13.

Narayanan, A. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, volume 1170, 3.

Sambasivan, N.; Arnesen, E.; Hutchinson, B.; Doshi, T.; and Prabhakaran, V. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 315–328.

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019a. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, 59–68.

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019b. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, 59–68.

Steck, H. 2022. Netflix research: Machine Learning.

Thomas, D. R. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*, 27(2): 237–246.

Wang, C.; Han, B.; Patel, B.; and Rudin, C. 2022. In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *Journal of Quantitative Criminology*, 1–63.

Watkins, E. A.; McKenna, M.; and Chen, J. 2022. The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness. *arXiv preprint arXiv:2202.09519*.

Ye, H.; Liang, L.; Ye Li, G.; Kim, J.; Lu, L.; and Wu, M. 2018. Machine Learning for Vehicular Networks: Recent Advances and Application Examples. *IEEE Vehicular Technology Magazine*, 13(2): 94–101.

Appendix

1 Participants

1.1 Participants with no prior experience of Fairlearn

Individuals with no experience with Fairlearn			
	Educational Background	Location	Current Role
P1	Computer Science Engineer	Germany	Data Scientist
P2	Mechanical Engineering	United States	Senior Data Scientist
P3	Data Science	Netherlands	Academia(Student)
P4	Computer Science	Netherlands	Academia(Student)
P5	Computer Science and AI	Romania	Academia(Student)

Figure 1: Participants with no prior Fairlearn Experience

1.2 Participants with no prior experience of AIF360

Individuals with no experience with AIF360			
	Educational Background	Location	Current Role
P1	Computer Science	Netherlands	Academia(Student)
P2	Data Science	Spain	Data Scientist
P3	Bioinformatics	Netherlands	Academia(Student)
P4	AI/ML	Netherlands	Academia(Student)
P5	Electronics	India	Data Scientist

Figure 2: Participants with no prior AIF360 Experience

2 Interview Questions

The entire structure of the interview can be found here.

A.Background Questions (you can refrain from answering any questions if it makes you uncomfortable)

1. Demographic:
   a. Where are you from?
   b. What is your gender?
   c. What is your educational background?
2. Experience with machine learning
   a. Students
      i. What is your experience with machine learning?
      ii. Do you have any work experience in ML or data-science?
   b. Practitioners
      i. Do you work in academia or industry?
      ii. What is your role?
      iii. What is your technology area? (NLP, Recommender Systems,Chatbots, Vision etc.) What kind of task? (regression, classification, ...) What kind of domain have you worked with now and in the past? (e.g. banking, healthcare, etc.)
      iv. For how long have you been working with machine learning/data engineering?

B.Introduce Use cases

- **Use case 1 (Diabetes)**

  Management of hyperglycemia in hospitalized patients has a significant bearing on outcome, in terms of both morbidity and mortality. However, there are few national assessments of diabetes care during hospitalization which could serve as a baseline for change. In this context, a hospital is looking into ways to predict whether diabetic patients will be readmitted within 30 days.

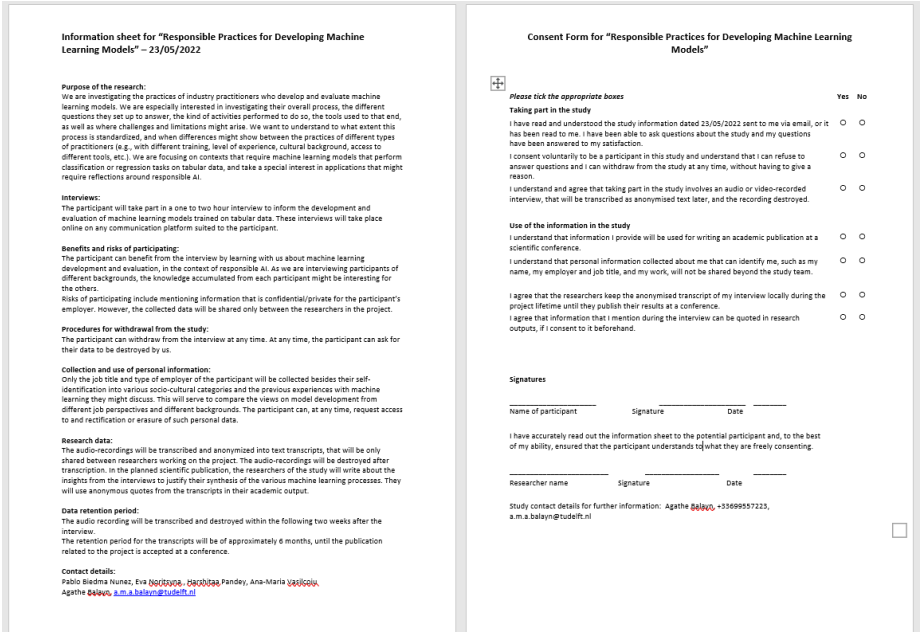
  Hospital readmissions increase the healthcare costs and negatively influence hospitals' reputation. In this context, predicting readmissions in early stages becomes very important since it allows prompting great attention to patients with high risk of readmission, which further leverages the healthcare system and saves healthcare expenditures.

  The hospital has heard about the potential of introducing an automated ML system to make this prediction. They are giving you access to a large clinical database and they are asking you to do some exploration and present a summary of your findings: can they imagine automating this possibility? If not, why? If yes, what would they need to do and consider?

  Task description (experience + toolkit & no experience + no toolkit):
  We are asking you to explore the use-case to answer this question. Feel free to use any tool you would typically use if you want to actually look into the dataset and/or model. We can provide a Jupyter notebook in which both the dataset and the toolkit are loaded.

Figure 3: Interview Guide

3 Consent Form for Interviews



Information sheet for "Responsible Practices for Developing Machine Learning Models" – 23/05/2022

Purpose of the research:
We are investigating the practices of industry practitioners who develop and evaluate machine learning models. We are especially interested in investigating their overall process, the different questions they set up to answer, the kind of activities performed to do so, the tools used to that end, as well as where challenges and limitations might arise. We want to understand to what extent this process is standardized, and when differences might show between the practices of different types of practitioners (e.g., with different training, level of experience, cultural background, access to different tools, etc.). We are focusing on contexts that require machine learning models that perform classification or regression tasks on tabular data, and take a special interest in applications that might require reflections around responsible AI.

Interviews:
The participant will take part in a one to two hour interview to inform the development and evaluation of machine learning models trained on tabular data. These interviews will take place online on any communication platform suited to the participant.

Benefits and risks of participating:
The participant can benefit from the interview by learning with us about machine learning development and evaluation, in the context of responsible AI. As we are interviewing participants of different backgrounds, the knowledge accumulated from each participant might be interesting for the others.
Risks of participating include mentioning information that is confidential/private for the participant's employer. However, the collected data will be shared only between the researchers in the project.

Procedures for withdrawal from the study:
The participant can withdraw from the interview at any time. At any time, the participant can ask for their data to be destroyed by us.

Collection and use of personal information:
Only the job title and type of employer of the participant will be collected besides their self-identification into various socio-cultural categories and the previous experiences with machine learning they might discuss. This will serve to compare the views on model development from different job perspectives and different backgrounds. The participant can, at any time, request access to and rectification or erasure of such personal data.

Research data:
The audio-recordings will be transcribed and anonymized into text transcripts, that will be only shared between researchers working on the project. The audio-recordings will be destroyed after transcription. In the planned scientific publication, the researchers of the study will write about the insights from the interviews to justify their synthesis of the various machine learning processes. They will use anonymous quotes from the transcripts in their academic output.

Data retention period:
The audio recording will be transcribed and destroyed within the following two weeks after the interview.
The retention period for the transcripts will be of approximately 6 months, until the publication related to the project is accepted at a conference.

Contact details:
Pablo Bledma Nunez, Eva Nordtvoe , Harshitaa Pandey, Ana-Maria Vasilcoiu,
Agathe Balayn, a.m.a.balayn@tudelft.nl

Consent Form for "Responsible Practices for Developing Machine Learning Models"

Please tick the appropriate boxes	Yes	No
Taking part in the study		
I have read and understood the study information dated 23/05/2022 sent to me via email, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	○	○
I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	○	○
I understand and agree that taking part in the study involves an audio or video-recorded interview, that will be transcribed as anonymised text later, and the recording destroyed.	○	○
Use of the information in the study		
I understand that information I provide will be used for writing an academic publication at a scientific conference.	○	○
I understand that personal information collected about me that can identify me, such as my name, my employer and job title, and my work, will not be shared beyond the study team.	○	○
I agree that the researchers keep the anonymised transcript of my interview locally during the project lifetime until they publish their results at a conference.	○	○
I agree that information that I mention during the interview can be quoted in research outputs, if I consent to it beforehand.	○	○

Signatures

_____ _____ _____
Name of participant Signature Date

I have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

_____ _____ _____
Researcher name Signature Date

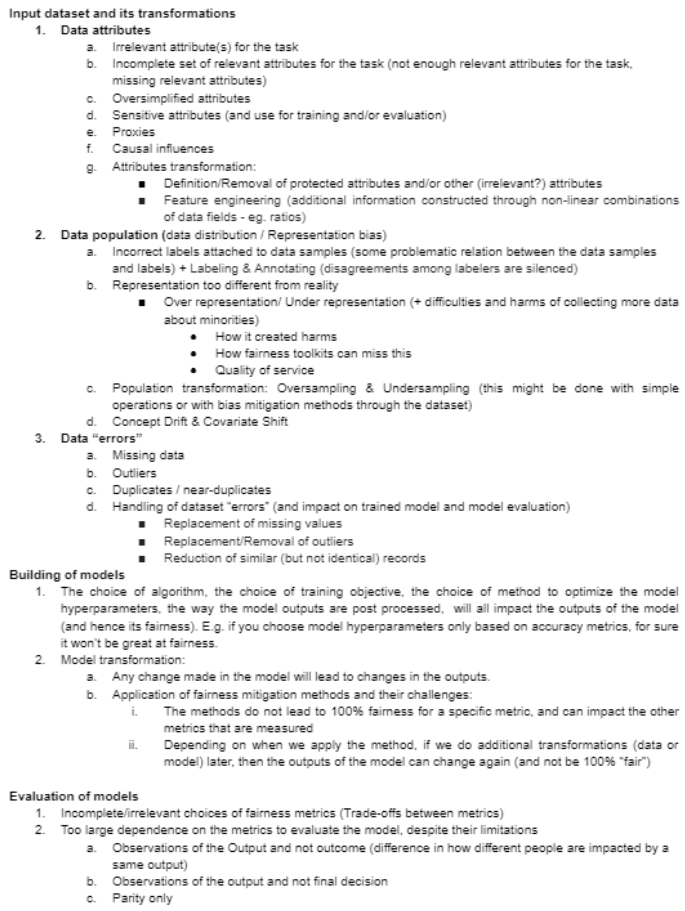
Study contact details for further information: Agathe Balayn, +33699557223, a.m.a.balayn@tudelft.nl

Figure 4: Consent form for interview

4 Codes for analysis

The codes which were used to analyze the interviews can be found on this link.

5 Algorithmic Harms



Input dataset and its transformations

1. **Data attributes**
   a. Irrelevant attribute(s) for the task
   b. Incomplete set of relevant attributes for the task (not enough relevant attributes for the task, missing relevant attributes)
   c. Oversimplified attributes
   d. Sensitive attributes (and use for training and/or evaluation)
   e. Proxies
   f. Causal influences
   g. Attributes transformation:
      - Definition/Removal of protected attributes and/or other (irrelevant?) attributes
      - Feature engineering (additional information constructed through non-linear combinations of data fields - eg. ratios)
2. **Data population (**data distribution / Representation bias)
   a. Incorrect labels attached to data samples (some problematic relation between the data samples and labels) + Labeling & Annotating (disagreements among labelers are silenced)
   b. Representation too different from reality
      - Over representation/ Under representation (+ difficulties and harms of collecting more data about minorities)
        - How it created harms
        - How fairness toolkits can miss this
        - Quality of service
   c. Population transformation: Oversampling & Undersampling (this might be done with simple operations or with bias mitigation methods through the dataset)
   d. Concept Drift & Covariate Shift
3. **Data "errors"**
   a. Missing data
   b. Outliers
   c. Duplicates / near-duplicates
   d. Handling of dataset "errors" (and impact on trained model and model evaluation)
      - Replacement of missing values
      - Replacement/Removal of outliers
      - Reduction of similar (but not identical) records

Building of models

1. The choice of algorithm, the choice of training objective, the choice of method to optimize the model hyperparameters, the way the model outputs are post processed, will all impact the outputs of the model (and hence its fairness). E.g. if you choose model hyperparameters only based on accuracy metrics, for sure it won't be great at fairness.
2. Model transformation:
   a. Any change made in the model will lead to changes in the outputs.
   b. Application of fairness mitigation methods and their challenges:
      i. The methods do not lead to 100% fairness for a specific metric, and can impact the other metrics that are measured
      ii. Depending on when we apply the method, if we do additional transformations (data or model) later, then the outputs of the model can change again (and not be 100% "fair")

Evaluation of models

1. Incomplete/irrelevant choices of fairness metrics (Trade-offs between metrics)
2. Too large dependence on the metrics to evaluate the model, despite their limitations
   a. Observations of the Output and not outcome (difference in how different people are impacted by a same output)
   b. Observations of the output and not final decision
   c. Parity only

Figure 5: Algorithmic Harms