Who Cares About Fairness: How Background Influences the Way Practitioners Consider Machine Learning Harms

Pablo Biedma Nunez
Supervisor: Agathe Balayn
EEMCS, Delft University of Technology, The Netherlands

24-6-2022

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

The increasing dangers of unfairness in machine learning (ML) are becoming a frequent subject of discussion, both, in academia and popular media. Recent literature focused on introducing and assessing algorithmic solutions to bias in ML. However, there is a disconnect between these solutions and practitioners' needs. By interviewing 30 ML practitioners of diverse backgrounds across 16 countries, and presenting them with a simulated use case, our study aims to investigate common fairness practices among professionals and how these are influenced by their backgrounds. The results reveal a superlative disparity among academia and industry practitioners. We also identify different practices in fairness and data exploration stages, influenced by the educational background as well as the level of experience of practitioners. Our study also finds how demographics have an impact on several aspects, such as willingness to accept and support legal actions taken against ML discrimination. In accordance with our findings, we suggest several actions that can be taken to improve fairness solutions, and we also highlight future directions for fairness research that can cause a positive impact on the way fairness is perceived by practitioners.

1 INTRODUCTION

In recent times, we have come across an increased reliance on machine learning (ML) technologies as these were incorporated into decision-making roles that embraced a discernible impact on our lives [43, 48]. Consequently, several problems have arisen. Of particular concern has been the fairness issue, omnipresent in the tech industry and even mainstream media. Countless examples such as Amazon's hiring tool discriminating against women [10, 36] or bias towards black people in criminal sentencing [2, 9, 14] prompted serious debates. However, ML harms are not limited to these prevailing examples, further instances of this problem range from bias in search engines [11, 37] to facial recognition systems performing poorly on underrepresented groups [17, 44, 46, 47, 52, 54]. The steadily growing coverage of events of this nature, at times misinterpreted, triggered a societal reaction aiming to tackle injustice in machine learning. Subsequently, institutions all over the world formulated fairness guidelines, targeting these concerns from a legislative standpoint [20, 24]. Furthermore, researchers and practitioners across disciplines also reacted to this trend by devising solutions to diminish ML harms.

These solutions are based on different fairness notions that practitioners are meant to aim for [13, 19, 27]. However, being conflicting, these definitions are far from perfect, as it is not possible to satisfy all of them simultaneously [21, 26, 29, 51]. To make it easier for ML professionals to carry out better fairness practices, some companies developed so-called fairness toolkits such as Microsoft's Fairlearn [5], IBM AIF360 [4], Google Fairness Indicators [55], and Aequitas, from The University of Chicago [42], in addition to a broad range of other captivating alternatives practitioners can opt for [1, 3, 49]. However, literature also brings to light the limitations of these. They are highly contextual, not applicable under certain models and their usability is poor [18, 22, 30, 32]. In short, the rationale behind the inability to tackle bias in ML has best been diagnosed as a disconnect between current fairness solutions and practitioners' needs [16, 23, 25, 40, 50].

We aim to fill this gap by studying practitioners' perspectives. Through 30 semi-structured interviews with machine learning specialists across 16 countries, we investigate how different backgrounds lead to different views on ML harms. We interviewed professionals from different sectors, roles, levels of experience, educational backgrounds, countries of origin, and gender, trying to answer the following question: To what extent does background influence how ML practitioners consider ML harms beyond the limited algorithmic solutions? Answering this question can contribute toward reducing bias in machine learning by gaining a broader perspective on practitioners' views. Our study reveals a remarkable difference between academia and industry practitioners, as well as how different educational backgrounds affect practices, together with the impact other demographics have on practitioners' views on fairness.

The following section of this paper will discuss related work and motivation for our research. In section 3, we will describe the adopted method together with the reasons behind choosing it. The key findings of our research will be displayed and analyzed in section 4. In section 6, we will reflect on the limitations of our study and suggest how future research can develop our work further. In chapter 5, we discuss the reproducibility of our study and its ethical aspects. Finally, our conclusion will be presented in chapter 7.

2 BACKGROUND AND MOTIVATION

Literature on algorithmic fairness has grown exponentially between the late 2010s and early 2020s. A central subset of this research has been devoted to investigating AI harm as well as its roots. These studies provide enlightening insights such as indicating that poor people are more vulnerable to ML harms, or that algorithmic biases reflect our society [31, 41]. However, machine learning does not differentiate between useful biases and discriminatory ones so studying them is not as trivial as it may seem [6]. This struggle faced by researchers in the field is best explained by Tal Zarsky who claimed that automated discrimination is more abstract, unintuitive, subtle, and intangible than regular discrimination [56].

That is why researchers shifted their focus towards the construction of fairness guidelines and tools. They did so by basing their work on mathematical definitions of fairness [19, 13, 27]. However, literature has also proven these to be misleading given their contextual nature. Let us take the concept of 'demographic parity'; according to its definition, which indicates each class should receive positive outcomes at equal rates, it is fair if we only accept the qualified applicants in one demographic and random individuals in an-

other as long as their percentages match [19]. Such scenarios, next to the fact that some of these definitions cannot be satisfied simultaneously, expose the flaws of fairness definitions many of which are being used regardless [21, 26, 29, 51].

Despite the defects of fairness metrics, literature has extensively concentrated on the creation and use of fairness toolkits based on these metrics. The toolkits are usually open source solutions that enable developers to find and mitigate bias in their AI models with a series of bias mitigators and fairness metrics [1, 3, 49]. Although companies introduced these toolkits as an attempt to further cultivate proper fairness practices, when releasing their toolkits, they already indicated their limitations explicitly. IBM indicated that their toolkit should only be used when there is an established notion of protected attribute and outcome variables and when humans collect these variables [4]. Microsoft also stated that their toolkit, Fairlearn, could not mitigate stereotyping harms, denigration harms or representation harms [5].

As time went by, researchers found even more problems with these tools, including important gaps in different stages of the ML pipeline such as data sampling or proxy analysis [30]. An overlooked gap lies in the fact that while literature claims that rows with missing values are the most relevant in terms of fairness, most toolkits simply ignore or remove these rows [32]. Some of the limitations of these toolkits relate to their usability. Research reveals the concerning difficulty of learning how to use fairness toolkits due to its poor user interface, lack of documentation, or oversimplification [30]. The remaining usability concerns revolve around transparency. To some extent, these are less easily avoidable because the 'curse of dimensionality' leads to opacity in ML debiasing tools [12]. Further reasons for a lack of transparency may include organizational or administrative secrecy as well as complexity or other preexistent structural factors [7]. Opacity leads to misuse and overtrust of any interpretability tool [25]. Therefore there is a significant danger of fairness toolkits being employed in inappropriate scenarios, or being wrongly used as a proof of total fairness of an AI system [30].

Although, as discussed throughout this section, there has been a remarkable effort put into studying the limitations of fairness metrics and toolkits, to our knowledge, not many researchers have focused on analyzing the practitioners themselves. The ones that did so, obtained valuable information. They concluded that AI students do not tend to think about ethical aspects of their models unless told to do so. Students also think that companies designing these models should be held accountable for ML harms rather than the developers themselves [33]. Literature also found that machine learning practitioners consider data collection and pre-processing the most important stages in the ML pipeline [40]. Furthermore, practitioners need solutions to be suitable for all types of models. For instance, literature found that practitioners who work on ranking, recommendation, or speech synthesis tasks, cannot easily apply the existent fairness tools [40]. On top of that, practitioners also need to be able to use such tools without having access to sensitive features. This relates to a lack of help when it comes to collecting and curating high-quality datasets that practitioners need to begin with, as well as most fairness solutions assuming access to certain demographics many practitioners lack [23]. Most of the aforementioned research involving ML practitioners arrives at the common conclusion that there is a disconnect between practitioners' needs and fairness solutions as these are driven by algorithmic methods available rather than practitioners' demands [16, 23, 25, 40, 50].

While most of these studies focus on ML practitioners' needs, our goal is to study the way these practitioners approach fairness, whether they think about it at all when building a model, and how this is affected by their background, which, to our knowledge, has not been done before. Previous studies do not consider practitioners' gender, educational background, country of origin, or other demographics. Therefore our research aims at determining the extent to which background influences how ML practitioners perceive fairness. We will also use these findings to suggest what direction current solutions should steer toward.

3 METHOD

For the most part, our research contains two different segments: an extensive literature review, and an empirical study. In the course of the literature review, we intended to gather a profound understanding of sources of harm in ML, the solutions that are currently used, and the limitations these carry. To better understand how background influences the way practitioners deal with these harms, we included an empirical element by conducting 30 semi-structured one-on-one interviews. This empirical work will be described in the following subsections.

Preparations

Our study commenced with an in-depth literature review. This investigation led to the elaboration of an extensive list of sources of harm in machine learning. This list can be found in Appendix A. Thereafter, we began to develop Jupyter notebooks with datasets that incorporated most of the sources of harm we had specified in our list. While devising these notebooks, our goal was to be able to share them with different ML experts, in order to contemplate whether these practitioners notice the potential harms, and what they would do to tackle them.

In particular, two different use cases were developed, both belonging to the medical sector. In the first one, a hospital is trying to predict whether diabetic patients will be readmitted within 30 days. The second one consists of an insurance company trying to predict high healthcare utilization. We chose these use cases because, being in the same domain, they both involve interesting yet not widely studied datasets, preventing practitioners from being already familiar with such a use case, and ensuring the validity of our study. A design brief of both use cases can be found in Appendix C. To implement them, we used two publicly available datasets, which already contained several of the problems we wanted to include, such as uneven data distribution or irrelevant attributes. The remaining sources of harm from our list, such as missing values or duplicates, were added synthetically. The resulting datasets can be found on GitHub [38]. Additionally, to also

study how practitioners interact with fairness solutions, we provided them with fairness toolkits. For those without previous experience with such tools, we elaborated a brief demo on how to use them. The toolkits provided were AIF360 [4] and Fairlearn [5], equally distributed among participants. We considered these toolkits representative enough of the full range of functionalities offered by such solutions. In Appendix C several screenshots can be found that provide an enhanced perspective on the way these notebooks look like. The complete notebooks can be found on GitHub [39].

Practitioner	Technology / Role	Interviewee ID
Student	NA	P5, P9, P10 P11, P12, P13 P18, P28
Developer	Federated Learning	P4
	Data Scientist	P6, P13, P16 P23, P26, P27
	Data Engineer	P15
	ML Engineer	P7
	IT Assistant	P17
	Fintech	P21, P22, P30
	Mobility	P24
	Recommender Systems	P25
Researcher	Computer Vision	P8, P14
	Predictive Maintenance	P1
	Ethics	P2
	Information Systems	P3
	Trustworthy AI	P10
	Medical AI	P19, P20
	Speech Recognition	P24
	Cognitive Networks	P29

Table 1: Overview of participants and their background

Participants

While selecting participants we sought to obtain a diverse subset of practitioners. Their background was defined in terms of role, sector, level of experience, educational background, country of origin, and gender. We interviewed students, developers, and researchers among others. These are involved in different technology areas ranging from recommender systems to speech recognition or computer vision. A more detailed overview of the participants can be found in Table 1, together with their interviewee ID, which will be used throughout the paper to refer to specific practitioners. Furthermore, 6 practitioners self-reported as female and 24 as male. 8 candidates come from an Asian country, 13 from European ones, 8 from North-American countries, and 1 from South America. Additionally, Figure 1 displays the educational background of the interviewees. Twenty of them [P1-P4, P10-P12, P14, P16, P18, P19, P21-P23, P25-P30] had previous experience with at least one fairness toolkit. The rest did not. Various practitioners [P5, P13, P17, P20], received ethical training as part of their bachelor's degrees. The recruitment process took place in different ways, including AI-related Slack and Discord communities, LinkedIn direct messages, word-of-mouth, and mainly, direct connections or referrals from other participants.

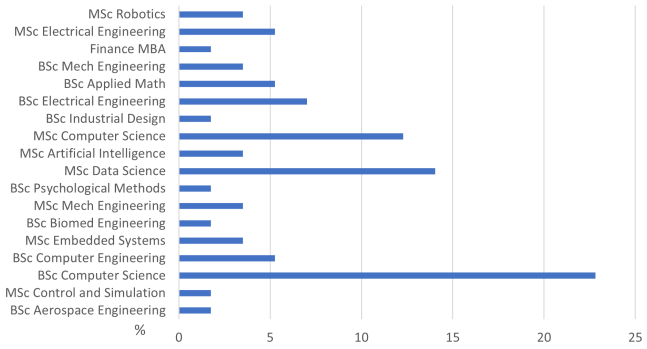

Figure 1: Most common degrees pursued by interviewees

Interviews

The interviews started with several background questions to get to know more about the participants. These consisted of demographic questions such as country of origin and gender, as well as questions about their work, educational background, and machine learning experience. Afterward, practitioners were presented with a use case. When introduced to the use cases, participants were asked to explore the dataset to determine the feasibility of the assigned task. They could do this by coding or simply thinking out loud and telling us what they would usually do in practice. In this fashion, we studied the number of potential harms practitioners detect as well as the way they deal with them, and their awareness of the consequences these actions can lead to. Finally, a few ending questions were asked, mainly regarding their view on fairness, their experience with toolkits, fairness practices at the company they work for, and general questions about bias in machine learning. A complete overview of the interview questions can be found in Appendix B.

Interviews with specialists who had previous experience with fairness toolkits lasted one hour. However, 2-hour interviews were conducted with the remaining participants. We decided it would be of more interest to use two different use cases on these practitioners. In the second use case, we would let them use a fairness toolkit for the first time. The rationale behind this choice is that this way, we could study how practices among ML developers differ when using or not a toolkit. These participants had to be shown a demo explaining how such a tool works. Appendix C contains images of these demos. Because of the additional use case and demo, we opted for a longer interview duration.

To guarantee the viability of the study, pilot interviews were conducted. These pilots induced the improvement of several aspects of the interviews, including managing the

content to ensure its feasibility under the given time constraints. Furthermore, most interviews were conducted remotely, therefore, Google Colab was used to enable participants to run the Jupyter notebooks on their machines. Microsoft Teams was also used for conducting and recording the interviews. We also used its built-in transcription tool. For the sake of analyzing the interviews, the transcript was corrected while watching the recordings. Simultaneously, notes were taken about the background of participants as well as answers to the introduction and ending questions. Furthermore, key moments were also timestamped. Depending on the stage of the interview, these moments were manually labeled as 'harm identification', 'data exploration practice', 'building of the model', 'model evaluation', and 'use of fairness toolkit'. As they were working on the use cases, some practitioners provided interesting opinions or considerations with or without being prompted by the interviewee. These were labeled as 'judgments'.

All participants were asked for consent before being recorded. These recordings were transcribed as anonymized text and then deleted within two weeks after the interviews. By signing a consent form, participants also agreed to be anonymously quoted in research outputs. They were notified that the information they provided could be used for writing an academic publication. Additionally, any personal information collected about participants that could identify them was not shared beyond the study team. They were also informed that they could refuse to answer any questions as well as withdraw from the study at any time without having to give a reason. Lastly, we would like to indicate that during this process we carefully followed the research ethics guidelines of the Delft University of Technology.

4 FINDINGS

This section aims to outline the discoveries made by virtue of our research and the conducted interviews. These findings are categorized into various subsections. For each of them, we elaborate on the implications they bring and how certain elements can be explored further by future research. Although unrelated to practitioners' backgrounds, some findings were still added for the interest of the reader.

Discrepancies in data exploration practices

Practitioners without a computer science-related background tended to look at data exploration from different perspectives. For instance, P1, whose educational background is in industrial design, mentioned the following:

> *"[My exploratory steps] would involve talking to people rather than looking at the data... See how they use similar [patient readmission prediction] systems right now if at all. [...] [I would] map out the process... "*.

This distinctive approach may bring certain advantages to people coming from less purely technical fields. However, several concerns were raised by our interviewees indicating how some data scientists don't have the right background to work in such a role. P6 stressed the following:

> *"There are lots of data scientists that are making a lot of money that are not very good at data science right now, and I've seen my [former] company give big money [to them] and they don't know basics in, like, training, splitting, or testing"*.

Regardless of their background, it is perhaps interesting to mention that most practitioners talked about the importance of a close collaboration with domain experts or clients in the exploratory stage. The only participants who did not explicitly express this need were students and researchers, with a total of 26 interviewees advocating for such a collaboration.

Literature claims that practitioners will have to invest more time in the fairness exploration stage than in the data exploration stage [40]. We found that data exploration is where practitioners spend most of their time. As P6 said, *"Not having the data cleaned up, is an issue all over the industry, where maybe 75 percent of the time in data science is spent"*. Therefore, practitioners may be reluctant to spend even more time in fairness exploration. As P14 put it: *"There is no point in doing a fairness assessment if there are no real harms tied to [the data]"*. Furthermore, fairness toolkits do not fix this problem because as P2 indicated: *"If [practitioners] use the toolkit, they use it because they are aware of potential harms but it won't help [to find] things they are not aware of"*. This may suggest that future research should be undertaken to find time-efficient ways for practitioners to perform proper fairness exploration.

Some practitioners specified that the training data should be diverse. Furthermore, practitioners also indicated that this data should accurately represent the context in which it will be used. However, one of these two things may only be possible to achieve at the cost of the other. While some participants, like P8, highlighted the importance of a balanced dataset: *"The data should be equally distributed if you want to have the same probability of giving a correct answer for each of the groups"*, participants with a longer work experience in industry [P17, P24, P25] advocated for data that represents the real world. As P17 stated: *"Data distribution depends on the environment it doesn't need to be 50-50"*. One way or another, this discrepancy is something that needs to be dealt with, and perhaps, clearer guidelines around this notion should be established in industry and academia.

Equality of opportunity vs Equality of outcome

When students and even industry practitioners, who were not familiar with the notion of demographic parity, were introduced to this concept for the first time, at first, they all thought it was a good indicator of fairness when we described it to them as the difference in the rate of positive outcomes between privileged and unprivileged classes. However, after we explained to them the limitations such a concept has, with real-life examples, all of them thought such a metric should not be used to estimate fairness. We explained to them, how demographic parity can potentially allow selecting only the qualified applicants in one demographic and random individuals in another as long as their percentages match [19].

This means people who are not aware of the limitations of these metrics might still use them thinking they are fair, thus

blindly causing more ML harm than reducing it. Interviewees who were already familiar with what demographic parity entails, immediately indicated their preference against the metric. P4 stressed: "Demographic parity does not indicate any sort of fairness in a model". However, we learned from one of our interviewees [P12], a master's student who interns for one of the largest fairness toolkits, that they highly rely on demographic parity: "At [my company] they think demographic parity is very important, it is like their main metric". Therefore we could argue that companies and fairness solutions should stop blindly relying on those fairness definitions that advocate for equality of outcome over equality of opportunity such as demographic parity.

Total fairness: When do practitioners stop trying?

During the interviews, we learned that most practitioners are not concerned about fairness in the first place. Of the 10 people interviewed without previous experience with fairness toolkits, none of them knew about fairness metrics. As P15 stated: *"No one looks at fairness metrics"*. After explaining how they work, we asked these participants if they would use such metrics. The same practitioner replied the following: *"If I am doing weather prediction, then no"*. Another practitioner [P9], stated the following:

> *"I am gonna be honest with you, I think in reality all that really matters is that you have a model with good performance that's scalable accurate, etc. However, if the application or if the use-case is sensitive, then you have to readjust and you have to include other metrics in your evaluation"*.

We asked interviewees whether they think absolute fairness is possible to achieve. In all the interviews, the answer was a convinced 'no'. As a follow-up question, we asked them at what point they would stop trying to reduce bias. Many participants were unsure how to answer this question. P12, a student who interns for a fairness toolkit said: *"The more I work at [my company] I am not sure what fair is anymore"*. Some practitioners saw the process as something iterative but most participants reflected on the contextual aspect of it. A participant who works as a maintainer for a fairness toolkit [P1] noted: *"We tried to move away from fairness as an optimization game to more thinking about the context"*, or as P7 put it: *"Ideal fairness depends on the area and the society and people and business itself"*. Another participant [P14] reflected on the contextual aspect of fairness and how toolkits fail to account for it: *"In many toolkits, they suggest that if disparate impact is at least 0.8, you are good to go, but you are not, this needs to be contextualized"*. Another weakness of fairness toolkits is that because different definitions of fairness may produce entirely different outcomes [13], practitioners may not know what metric to use, and toolkits do not help with this. Moreover, literature shows that most of these fairness notions include specific mathematical definitions that most users are unfamiliar with [40]. This is further worsened by the availability of several bias mitigation algorithms and even the question of whether data should be debiased in the first place.

One way or another, most interviewees see absolute fairness as something impossible to achieve, and some of them indicated that even after using all available fairness tools, you only know you are being unfair when your product has finally been deployed and some user complains online about it. This 'paradox' may make practitioners less proclive to consider ML harms beyond the limited algorithmic solutions. Regardless, the fact that many teams only focus on customer complaints rather than proactive approaches when it comes to fairness, is a clear symptom that there is a problem with fairness solutions or how these are used.

Ethical sensitivity in students and practitioners

Nora McDonald and Shimei Pan suggest there is a need for an increased presence of ethics in the computer science curriculum [33]. This claim may be correct, because, as P14 stressed: "There is a danger of giving people too many tools and not educating them on what they mean". Universities are already making a noticeable effort to increase the presence of ethics [45]. However, these efforts may not have a large impact since it will be difficult to introduce ethics in the curriculum without students just considering it 'free points' in the exams. In fact, we interviewed several computer science students who had received ethical training as part of their bachelor's degree [P5, P13, P17, P20] and we failed to find an increased ethical sensitivity in such students.

Another reason why computer science students are not sufficiently concerned about the subject may be the fact that when taught about it in class, the same problems are typically brought up: Amazon's hiring tool [10], criminal sentencing discrimination [2], and perhaps Taybot.ai [35] or some examples of bias in image recognition [54] rather than informing students about different instances of harm they can relate to. Regarding practitioners, unfortunately, a similar phenomenon occurs in fairness literature, where the introduction chapter of most papers usually refers to the same examples of bias in machine learning [6, 9, 18, 23, 33, 34, 40, 50]. Therefore practitioners' perspectives on algorithmic harms may be constrained to these instances, and that is why it is important for future research to find more efficient ways to engage students and practitioners in the ML fairness field.

Practitioners in industry vs academia

The most superlative difference in terms of the impact background had on participants, was found between practitioners who work in industry and academia. This may be because of the large difference between an academic setting and working in the industry. We found that academia tends to focus on correct practices that industry practitioners ignore in real life. P1 highlighted this difference:

> *"If I am gonna do things super nicely, I would first split [the data] into a training set, a test set, and then do the exploration on the training set and [so on]. All those nice things, but in reality, does anybody do that?"*.

Another practitioner [P16], gave an example of this crucial difference between expected practices and the industry setting: *"It could be that you collect data from really good devices, and in the real world, people use a device they bought for 20 euros from Ali Express"*.

These discrepancies make industry practitioners and researchers behave in a different manner. From one of our participants [P22], we find that data scientists in their company, unlike in academia, consider the legal consequences of their actions rather than what they deem morally appropriate in a given context. *"By regulation, this should be dropped..."*. We also find that in industry, practitioners use fairness tools mainly just to show correctness to non-technical people. P3 emphasized the following: *"Toolkits are useful to connect with business people and people who can't code"*.

We found further disparities between industry and academia regarding the use of toolkits. The main one concerns mitigation algorithms. Two practitioners [P16, P21] explicitly stated they would only use mitigation algorithms in research but never in production. P1, who, like P16, works for a fairness toolkit, mentioned the following:

> *"This is gonna be very controversial but [...] I think, right now, honestly, fairness mitigation algorithms are not really at the stage where they should be routinely applied in practice"*.

The same participant also suggested that, if you need mitigation approaches to be accurate or fair, you should always question whether machine learning makes sense to use in that context. P1 also explained that while using such algorithms, you blindly optimize for something. This is, in part, because, as P25 emphasized: *"[Mitigation algorithms] are a little bit difficult to understand. I've tried reading the papers a few times"*. The bias mitigation process was described as a shot in the dark by multiple participants. This forces practitioners to blindly search for their own solutions to improve fairness scores. As P9 stressed:

> *"It is really important to understand how to compute these, and how to draw the line between what is actually bias and what isn't, which is hard because this is subjective [...] It would be nice to be able to track down or reverse engineer this because having black box models outputting scores may be not enough for deciding if the decision is fair or not"*.

This relates to the problem of explainability which was described as a key concern by several interviewees working in the industry. As P12 noted: *"Explainability may be more important [than fairness]"*. Literature has shown that lack of explainability may lead to practitioners ignoring an outcome when their initial judgment differs from it [28]. When these people have, or think they, have a strong background in their domain, the likelihood of this happening increases. Furthermore, these practitioners tend to be unwilling to listen to unfamiliar viewpoints [8]. This problem is becoming increasingly relevant as the 'curse of dimensionality' leads to opacity in ML debiasing tools [12]. Therefore further work may be needed to reinforce explainability in fairness solutions.

Apart from these concerns revolving around debiasing, industry practitioners find further limitations in current fairness solutions that researchers do not necessarily face. First of all, as one of the participants highlighted [P10], most of the research on fairness is done in binary classification tasks. The industry practitioners we interviewed, focus on a more varied series of applications than interviewees in academia. Therefore, their needs when it comes to dealing with ML harms go beyond the currently available algorithmic solutions as these applications cannot be represented by the mathematical definitions fairness metrics use [23]. As P24 put it:

> *"[The toolkit] is very academic at this point. It does have a lot of metrics, and it is very useful, but for work, or commercial settings, there are many other things to account for"*.

To summarize, while practitioners who work in academia may be more sensitive to ML harms due to dealing with them more often, industry practitioners face more limitations as they work in real-world scenarios. Further work needs to be done in the fairness field to address these limitations.

Responsibility in responsible AI

We find that several companies, follow the 'right' fairness practices merely to avoid negative consequences legally, economically, or through damages to the image of the company. As P16 noted: *"Fairness for many companies is just a small checkbox"*. This cynical behavior is extended to the practitioners who work for such companies. When we introduced our use-case to P22 and asked about their approach, this interviewee pointed out the procedure they are instructed to follow at their company: *"By regulation, this should be dropped..."*.

As with P22, we find that participants who come from North American countries or work for North American companies are more willing to accept and support legal actions taken against discrimination in AI than European participants. This may be a reflection of the amount of work undertaken in the field in America as opposed to lesser efforts by European law [53].

We also asked participants who they thought should be accountable for fairness concerns. The responses were diverse. Several participants [P17, P23, P26] claimed that developers should be responsible. P23 expressed it clearly: *"Fairness is my own responsibility"*. Other practitioners [P5, P13, P21, P24] mentioned that companies, c-levels, product owners, or someone else other than themselves should be liable for such matters. For instance, P21 said: *"The c-level people should be responsible for [responsible ML questions]"*.

One way or another, accountability in responsible AI, as well as decision-making support, should be studied further to be able to tackle responsible ML matters more efficiently. Furthermore, it may be beneficial to devise separate tools and approaches that differ from each other depending on practitioners' backgrounds rather than employing universal solutions.

5 LIMITATIONS AND FUTURE WORK

While this study considered how background affects sensitivity to ML harms, it is important to keep in mind that ethical sensitivity is a trait and also a skill that can be taught or developed [15]. Therefore, it can be present regardless of your background and there exist risks of false correlation when studying how this background affects your moral standards when it comes to judging ML harms. Our study is also

limited by a small sample size of practitioners who do not represent the full range of ML professionals. This can also lead to a risk of false correlation.

Further limitations to this study correspond to the toolkits we presented practitioners with, the use cases we devised as well as participation bias. We consider the information collected and the participants recruited adequate to conduct our study and reach the presented conclusion. However, we could have collected further demographics of participants or recruited a more diverse group. Therefore, future work can focus on a larger subset of practitioners with even more diverse backgrounds. More techniques to obtain valuable insights can be used such as focus groups or questionnaires. Additionally, these studies could potentially involve a wider variety of fairness solutions, using more than just two toolkits.

Our findings can be used for future research to try and improve fairness solutions as well as produce new ones. Future work could focus on studying in-depth each of the concerns brought up throughout the paper such as finding time-efficient ways to perform fairness exploration, establishing clearer guidelines regarding data distribution, studying ways to efficiently include ethics in the computer science curriculum, increasing explainability of fairness solutions, defining accountability in responsible ML, and increasing decision-making support in the field. Regarding background, our study found several disparities among practitioners, and therefore future research may explore the possibility of developing different approaches or tools for different backgrounds.

Our research is partly motivated by a need for avoiding misinterpretation of fairness and its metrics. We urge future studies to take this into account by exploring the contextual nature of fairness definitions and avoiding charging these politically.

6 RESPONSIBLE RESEARCH

As human research is a core component of this study, several ethical considerations are applicable. While conducting our study, we consciously followed the rules and regulations of the Delft University of Technology Human Research Committee. All interviewees read and signed a consent form. Participants authorized being recorded. These recordings were transcribed as anonymized text and then deleted within two weeks after the interviews. Participants also agreed to be anonymously quoted in research outputs. They were notified that the information they provided could be used for writing an academic publication. Additionally, any personal information collected about participants that could identify them was not shared beyond the study team. They were also informed that they could refuse to answer any questions as well as withdraw from the study at any time without having to give a reason. Lastly, we would like to indicate that during this process we carefully followed the research ethics guidelines of the Delft University of Technology as well as the Netherlands Code of Conduct for Research Integrity.

Honesty, scrupulousness, transparency, independence, and responsibility were values we always kept in mind while performing our research. Regarding reproducibility, our study can be replicated by following the method outlined in the third chapter. Due to the diverse background of the participants recruited and the availability of the tools used, this study can be reproduced anywhere in the world by using a dataset that encompasses similar use cases and includes the harms we listed in our list of harms which can be found in Appendix A. For the sake of reproducibility, interview questions can also be found in Appendix B. The exact datasets and notebooks used can also be found on GitHub [38, 39].

7 CONCLUSION

Our study provides key insights into the way background influences how ML practitioners perceive fairness. Several of these findings will be epitomized in this section.

Among other discoveries, our study suggests how alternative educational backgrounds may bring advantages in fairness considerations while posing the risk of practitioners not being sufficiently prepared for a data science role. Both sides should be taken into account when hiring the right candidates for such positions. Similarly, we illustrate how depending on the level of experience of practitioners, there are some discrepancies with regard to how data should represent the context in which it is used and how to deal with historical bias. For better fairness practices regarding these factors, clearer guidelines should be established.

We failed to find an increased ethical sensitivity among students who received ethics training as part of their computer science curriculum. We show how this may be improved by presenting students and practitioners with a wider range of instances of harm in ethics courses and literature.

This paper also offers a view into how researchers in academia focus on aspects that practitioners in the industry usually ignore. These practitioners appear to be more concerned about legal consequences and regulations of the company they work for. We also show how, unlike researchers, industry practitioners mostly use fairness toolkits just to show correctness and communicate with business or non-technical people. These differences and limitations should be considered when building and assessing fairness solutions.

Although independent of their background, we saw how practitioners stand against demographic parity and associating fairness with equity. Therefore we argued that fairness solutions should not include those fairness definitions that advocate for equality of outcome over equality of opportunity.

It is crucial to understand how practitioners act to improve current fairness solutions. Furthermore, our findings suggest that we may need different approaches or tools that differ according to practitioners' backgrounds. We encourage further research to address this need together with all the concerns outlined throughout the paper. We long for this effort to cause a positive impact on the way fairness solutions are developed, and hopefully, the way fairness is perceived by practitioners.

8 ACKNOWLEDGEMENTS

Firstly, the authors would like to thank the participants of the interviews for devoting an important part of their busy schedule to us. We had the opportunity to interview a talented

group of people who turned out to be genuinely interested in the work we were conducting. Therefore, we would also like to express our appreciation to everyone who helped us recruit these candidates. We also wish to acknowledge our supervisor Agathe Balayn for her frequent and helpful feedback and for always taking the time to extensively answer our questions about the research. Furthermore, we would also like to express our profound gratitude to the responsible professors Jie Yang and Ujwal Gadiraju, and the Delft University of Technology. We are extremely thankful for their support, guidance, and professionalism. Lastly, we would also like to thank our research teammates Ana-Maria Vasilcoiu, Eva Noritsyna, and Harshitaa Pandey

REFERENCES

[1] Julius Adebayo. Fairml: Toolbox for diagnosing bias in predictive modeling. *Massachusetts Institute of Technology*, 2016.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. Propublica, May 2016.

[3] Niels Bantilan. Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. *Journal of Technology in Human Services*, 36:15 – 30, 2017.

[4] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *IBM*, 2018.

[5] Sarah Bird, Miroslav 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*, September 2020.

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

[7] Jenna Burrell. How the machine 'thinks: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3, January 2016.

[8] Paul R. Carlile. Transferring, translating, and transforming: An integrative framework for managing knowledge across boundaries. *Organ. Sci.*, 15:555–568, 2004.

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

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

[11] Matt Day. How linkedin's search engine may reflect a gender bias. Seattle Times, 2016.

[12] Pedro Domingos. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books, New York, 1 edition, 2015.

[13] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA, 2015. Association for Computing Machinery.

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

[15] Samantha Fowler, Dana Zeidler, and Troy Sadler. Moral sensitivity in the context of socioscientific issues in high school science students. *International Journal of Science Education*, 31:279–296, 01 2009.

[16] Jean I. Garcia-Gathright, Aaron Springer, and Henriette Cramer. Assessing and addressing algorithmic bias - but before we get there. *ArXiv*, 2018.

[17] Joseph Goldstein and Ali Watkins. She was arrested at 14. then her photo went to a facial recognition database. The New York Times, 2019.

[18] Ben Green and Lily Hu. The myth in the methodology: Towards a recontextualization of fairness in machine learning. *Harvard University.*, 2018.

[19] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *NIPS*, 2016.

[20] Rasmus Hauch. Denmark introduces mandatory legislation for ai and data ethics. 2021.AI, 2020.

[21] Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. A moral framework for understanding of fair ml through economic models of equality of opportunity. *ArXiv*, 2018.

[22] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. Intelligent tutors as teachers' aides: Exploring teacher needs for real-time analytics in blended classrooms. in proceedings of the seventh international learning analytics and knowledge conference. *ACM*, pages 257–266, 2017.

[23] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudík, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? *ArXiv*, December 2018.

[24] Anna Jobin, Marcello Ienca, and Efy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, September 2019.

[25] 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. *ArXiv*, 2020.

[26] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *ArXiv*, September 2016.

[27] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[28] Sarah Lebovitz, Hila Lifshitz-Assaf, and Natalia Levina. To engage or not to engage with ai for critical judgments: How professionals deal with opacity when using ai for medical diagnosis. *Organization Science*, 33, 01 2022.

[29] Michelle Lee, Luciano Floridi, and Jat Singh. Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. *AI and Ethics*, 1, 11 2021.

[30] Michelle Lee and Jat Singh. The landscape and gaps in open source fairness toolkits. *ACM*, 2021.

[31] Mary Madden, Michele Estrin Gilman, Karen E. C. Levy, and Alice E. Marwick. Privacy, poverty and big data: A matrix of vulnerabilities for poor americans. *Empirical Studies eJournal*, pages 53–125, 2017.

[32] Fernando Martínez-Plumed, Cesar Ferri, David Nieves, and José Hernández-Orallo. Missing the missing values: The ugly duckling of fairness in machine learning. *Wiley Periodicals LLC*, 2021.

[33] Nora McDonald and Shimei Pan. Intersectional ai: A study of how information science students think about ethics and their impact. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), October 2020.

[34] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 08 2019.

[35] Angela Moscaritolo. Microsoft puts tay chatbot in time out after racist tweets, Mar 2016.

[36] BBC News. Amazon scrapped 'sexist ai' tool. BBC, 2018.

[37] Safya Umoja Noble. Algorithms of oppression: How search engines reinforce racism. *NYU Press*, 2018.

[38] Pablo Biedma Nunez. Datasets. https://github.com/pablobiedma/datasets, 2022.

[39] Pablo Biedma Nunez. Notebooks. https://github.com/pablobiedma/notebooks, 2022.

[40] 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. *ACM*, 2021.

[41] Rashida Richardson, Jason Schultz, and Kate Crawford. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online*, 2019.

[42] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness. *ArXiv*, 2018.

[43] Salim Sheikh. *Understanding the role of artificial intelligence and its future social impact*. Advances in human and social aspects of technology book series. IGI Global, 2020.

[44] Natasha Singer. Amazon's facial recognition wrongly identifies 28 lawmakers. The New York Times, 2018.

[45] Natasha Singer. Tech's ethical 'dark side': Harvard, stanford and others want to address it. The New York Times, 2018.

[46] Natasha Singer. Amazon is pushing facial technology that a study says could be biased. The New York Times, 2019.

[47] Jacob Snow. Amazon's face recognition falsely matched 28 members of congress with mugshots. ACLU, 2018.

[48] Michael Cheng-Tek Tai. The impact of artificial intelligence on human society and bioethics. *Tzu Chi Med. J.*, 32(4):339–343, October 2020.

[49] Florian Tramèr, Vaggelis Atlidakis, Roxana Geambasu, Daniel J. Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Fairtest: Discovering unwarranted associations in data-driven applications. *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–416, 2017.

[50] Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decisionmaking. *ArXiv*, 2018.

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

[52] James Vincent. Gender and racial bias found in amazon's facial recognition technology (again). The Verge, 2019.

[53] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *SSRN Electronic Journal*, 01 2020.

[54] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. *ArXiv*, February 2019.

[55] Catherina Xu and Tulsee Doshi. Fairness indicators: Scalable infrastructure for fair ml systems. Google Research, 2020.

[56] Tal Zarsky. The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, 41(1):118–132, 2016.

A LIST OF SOURCES OF HARMS

A.1 Task

1. Undesired task
2. Task that only reproduces historical data patterns (without allowing for novelty)

A.2 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**
   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/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

A.3 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")
3. Other issues:
   a. Environmental Impact of model training
   b. Invisible worker
   c. Potential harms for individuals/environment outside the dataset.

A.4 Evaluation of models

1. Measurement Bias
2. Incomplete/irrelevant choices of protected attributes, and protected groups
3. Incomplete/irrelevant choices of fairness metrics (Trade-offs between metrics)
4. 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
5. (No) consideration of harms caused to people that are not directly subject to the model predictions (e.g. if someone is not given a loan, their family will also have problems.)

B INTERVIEW QUESTIONS

B.1 Background Questions

1. Demographics:
   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.2 Introduce Use cases

- Use case 1 (Diabetes):

Management of hyperglycemia in hospitalized patients has a significant bearing on outcomes, 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 healthcare costs and negatively influence hospitals' reputations. In this context, predicting readmissions in the early stages becomes very important since it allows prompting great attention to patients with a 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:

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.

Can you speak out loud to explain to us what you would do to answer the question? [of course, you don't necessarily have to do everything you would do in practice, you can also simply tell us about your plans]

- Use case 2 (Medical expenditure):

An insurance company has tasked you to develop a healthcare utilization scoring model that they will be able to employ when deciding the price of insurance for individuals. The model classification task is to predict whether a person would have 'high' healthcare utilization. To complete the task, the company has provided you with the 2015 Consolidated Medical Expenditure data.

As in the previous use case that you have seen, I am asking you again to speak out loud while trying to explore the use case to answer this question. You can of course use any tool that you would typically use, but you are encouraged to also make use of the toolkit I just presented to you.

After the participants' use-case exploration:

- What do you conclude from your exploration?
- *Use-case 1:*
  - Do you think the hospital can automate its task? If not, why?
  - If yes, what would they still need to do?
  - What would be some difficulties the hospital might face?
  - And especially, do you think the hospital can use this dataset to determine if a patient would need to be readmitted? Why, why not? What other dataset?
- *Use case 2:*
  - Do you think the insurance company can use this model to determine whether a person would have "high" healthcare utilization?
  - What would be some difficulties they might face?
  - What would they need to do further?
- What would be your concerns?
- What interrogations did you have when going through the process?
- What kind of challenges did you encounter?
- Were there things you would have liked to do but could not do 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?

B.3 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?

B.4 Questions about harms they don't mention

(See the list of harms).

B.5 Ending Questions

Practices at the company (if applicable)

a. To what extent do you use human monitoring versus automation in your company to reduce ML harms? Does this work?

b. Did you ever face a trade-off between fairness and accuracy in your work?

c. 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

d. Have you ever been confronted with ML models that can have a strong impact on certain stakeholders?

   i. If yes, what kind of impact was it?
   ii. How did you deal with that?
   iii. Did you proceed differently from other ML models?
   iv. What kind of models were they? for what kind of task and domain?

e. People now start talking about "responsible AI". Have you heard about that? What does that mean for you?

   i. How have you learned about that? Training? Self-learning? . . . ?
   ii. How much would you say you know about that? Can you give some examples of what you know? if they know some things, you can try to ask some more specific questions, like "ML fairness", what is it for you?
   iii. To what extent is this important in/ relevant to your work?

f. Responsibility: 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?

[For participants with experience]

g. What is your experience with Microsoft Fairlearn / AI Fairness 360?

   i. Why did you start working with this toolkit?
   ii. For how long have you been working with it?
   iii. How/when do you use it?

h. 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?

i. How would you describe your knowledge about responsible AI in general? And about fairness concepts more specifically?

Toolkit and change of perspective:

[For participants with experience]

j. After starting to use the fairness toolkit did your perspective on algorithmic harms change? If so, how?

k. 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?

[For participants with NO experience]

l. After using the fairness toolkit did your perspective on algorithmic harms change? If so, how?

m. Do you feel that, with the toolkit, your fairness consideration was 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:

n. If they have experience using both Fairlearn and AI Fairness 360, what are the differences between the two?

o. Was there a reason why you chose to learn how to use one toolkit over another?

p. 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 NOTEBOOKS

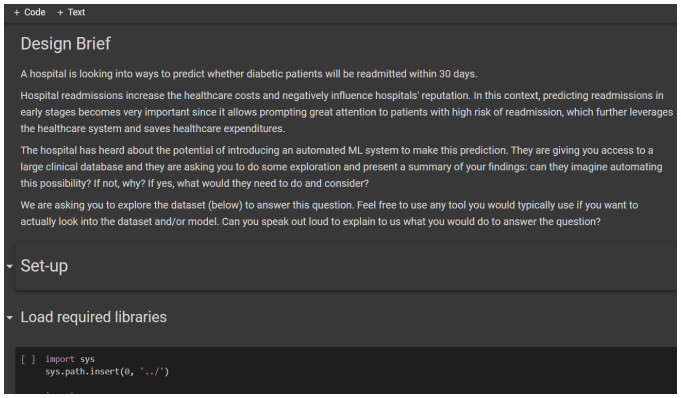

Figure 2: Design brief of first use case

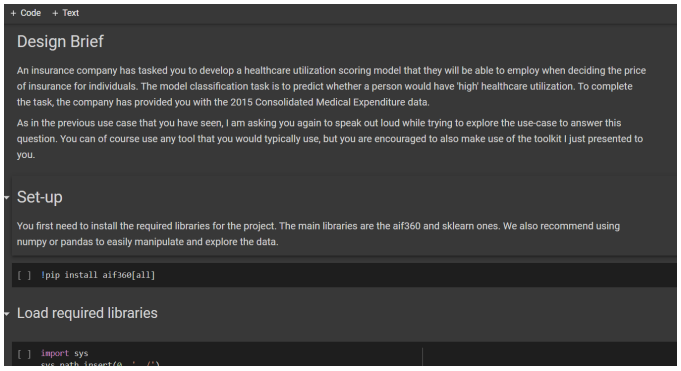

Figure 3: Design brief of second use case

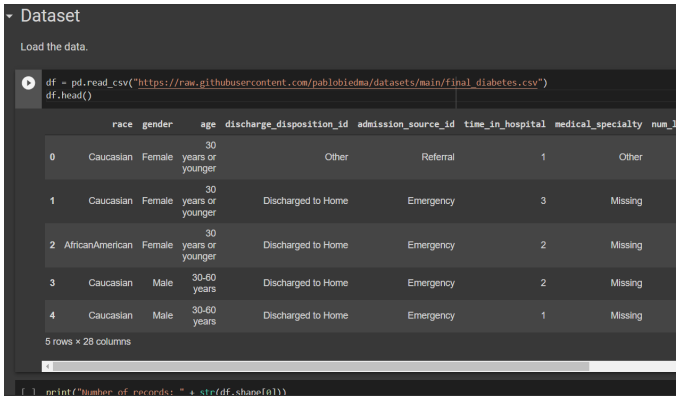

Figure 4: Overview of diabetes dataset

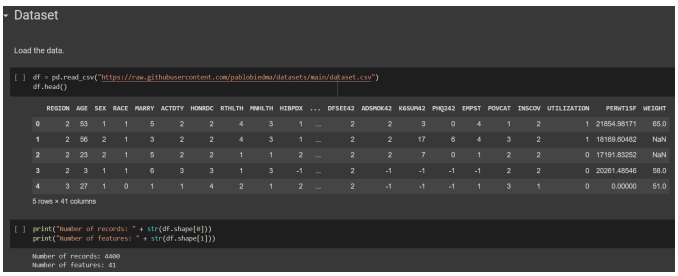

Figure 5: Overview of medical expenditure dataset

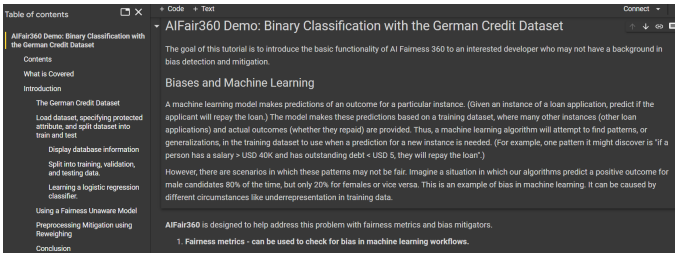

Figure 6: Overview of AIF360 demo

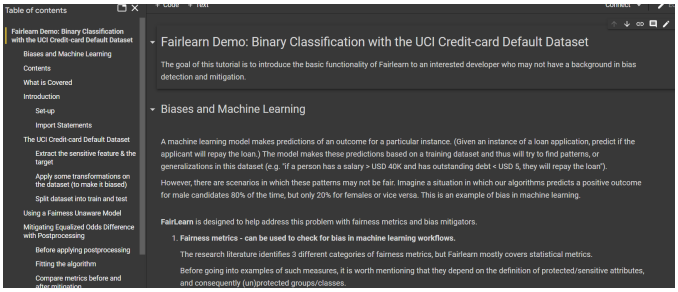

Figure 7: Overview of Fairlearn demo