Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits

Wesley Hanwen Deng
hanwend@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Manish Nagireddy
mnagired@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Michelle Seng Ah Lee
michelle.sengah.lee@cl.cam.ac.uk
University of Cambridge
Cambridge, UK

Jatinder Singh
jatinder.singh@cl.cam.ac.uk
University of Cambridge
Cambridge, UK

Zhiwei Steven Wu
zstevenwu@cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Kenneth Holstein
kjholste@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Haiyi Zhu
haiyiz@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

Recent years have seen the development of many open-source ML fairness toolkits aimed at helping ML practitioners assess and address unfairness in their systems. However, there has been little research investigating how ML practitioners actually use these toolkits in practice. In this paper, we conducted the *first in-depth empirical exploration* of how industry practitioners (try to) work with existing fairness toolkits. In particular, we conducted think-aloud interviews to understand how participants learn about and use fairness toolkits, and explored the generality of our findings through an anonymous online survey. We identified several opportunities for fairness toolkits to better address practitioner needs and scaffold them in using toolkits effectively and responsibly. Based on these findings, we highlight implications for the design of future open-source fairness toolkits that can support practitioners in better contextualizing, communicating, and collaborating around ML fairness efforts.

ACM Reference Format:
Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3531146.3533113

1 INTRODUCTION

The real-world impacts of machine learning (ML) systems are rapidly expanding, influencing outcomes in education [18, 51], healthcare [37, 98], credit scoring [106], social media [7, 36], public services [32, 43, 53], and criminal justice [26, 38], among many other areas. A growing body of research has drawn attention to the ways these systems can, whether inadvertently or intentionally, serve to amplify existing social inequities or create new ones [12, 15, 17, 22, 48, 59, 78]. In response, recent years have seen the development of many open-source ML "fairness toolkits" intended to assist ML practitioners in assessing and addressing unfairness in the ML systems they develop [4, 13, 14, 16, 91, 102, 109]. For instance, companies such as Microsoft, Google, and IBM, have published combinations of toolkits and guidelines [2, 8, 11, 42] that incorporate fairness as part of their core values.

Despite growth in the development and dissemination of fairness toolkits, there has been little research investigating how ML practitioners *actually use* these toolkits in practice. In order to explore practitioners' perceptions and desires around open-source fairness toolkits, Lee et al. conducted interview studies and a survey to identify the gaps between the capabilities of existing fairness toolkits and the needs of industry practitioners [63]. In a similar vein, Richardson et al. conducted an interview study with twenty ML practitioners in a simulated scenario in order to generate a practitioner-oriented rubric for evaluating fair ML toolkits [89]. However, neither of these two works engaged practitioners *directly* in using a fairness toolkit within the context of a real ML task. Prior literature suggests that the design of fairness toolkits that practitioners will find usable and useful in practice is a complex problem. For example, ML practitioners often find it challenging to appropriately *formulate the problem* when translating a real-world fairness question into a form amenable to quantitative fairness assessment [52, 80, 81, 113]. When faced with a fairness-related challenge, no single developer is likely to have all of the cultural and domain knowledge relevant to understanding or addressing the issue [9, 52, 92], and the appropriateness of different fairness definitions may be socially contested [73, 74, 79, 108]. Adding an additional layer of complexity, the contextual nature of ML fairness [31, 44, 62, 64, 103] makes it particularly difficult to design general-purpose toolkits that can effectively support practitioners in assessing and addressing fairness across a wide range of ML

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533113

applications and real-world contexts. To better understand and improve the usefulness and usability of software toolkits, prior research in Human-Computer Interaction (HCI) has emphasized the importance of studying how practitioners actually attempt to use toolkits in the context of real-world tasks [75, 76]. One effective approach is conducting "think aloud" interview studies, in which participants are asked to continuously articulate their thinking while exploring and using a software toolkit [35, 77, 100, 101, 110].

In this study, we conduct think-aloud interviews and a survey to explore the following two questions: *1. How do practitioners (try to) work with existing ML fairness toolkits? 2. What opportunities exist for fairness toolkits to better support them during these phases?* We first designed a realistic ML task in which we required practitioners to build an ML model based on a real-world dataset to help allocate education resources, while thinking about the potential fairness issues present in the dataset and their model. After screening forty-one industry practitioners who responded to our recruitment survey, we invited twenty-three ML industrial practitioners to undertake the task before joining the interview study. In the end, eleven practitioners finished the ML task and completed our think-aloud interview study, in which they encountered two fairness toolkits *for the first time*, explored the toolkit APIs, and tried to use them to address fairness issues in the ML models they built, while "thinking aloud" their observations, thoughts, and confusions. We also conducted an anonymous online survey with fifty-six industry practitioners who encountered fairness issues and might have used fairness toolkits before, so as to further explore and supplement our interview observations.

Through our investigation, we found that practitioners desire better support from fairness toolkits to better contextualize ML fairness issues and help communicate often complex fairness analysis to non-technical colleagues in their work places. We also identified four distinct design requirements [75, 77] ML practitioners had when using fairness toolkits, namely, (1) the abilities to use the toolkit to learn more about ML fairness and the landscape of current ML fairness research, (2) rapidly on-boarding toolkits due to workplace time constraints, (3) the abilities to integrate the toolkits into existing ML working pipeline, and (4) using toolkits as code repositories to implement state-of-the-art or domain specific ML fairness algorithms. In addition, we surface contexts in which practitioners committed to pitfalls [75] while addressing fairness issues, as well as misused the toolkits largely due to organizational time constraints for fairness work. Informed by our findings, we highlight implications for designing future open-source fairness toolkits. More broadly, our work contributes to growing efforts from both academia and industry towards ensuring that advances in ML fairness research have positive impacts in practice.

1.1 Understanding Practitioners' Needs and Challenges around ML Fairness

A number of recent studies have investigated challenges that ML practitioners face when attempting to improve fairness in practice. For instance, through interviews and surveys with commercial product teams, Holstein et al. [52] identified many disconnects between the solutions offered by the fair ML research literature (and toolkits implementing these solutions) versus the real-world challenges faced by industry ML practitioners. Through interviews and co-design workshops, Rakova et al. [87] and Madaio et al. [66, 67] investigated the organizational challenges, tensions, and barriers that practitioners face in practice when attempting to improve fairness in ML products and services.

More recently, in response to a surge of open source ML fairness toolkits (e.g., [14, 16, 91, 109]), Lee et al. [63] undertook a comparative assessment of the strengths and weakness of six prominent open source fairness toolkits and identified gaps between these toolkits' capabilities and practitioners' needs. In addition, Richardson et al. [89] created a rubric containing two main evaluation criteria for fairness toolkits, through an interview in which practitioners reviewed analysis results generated by fairness toolkits, which were curated by researchers. While we build upon and contrast against findings from these two important prior studies in our current work, both of these studies relied on retrospective interviews and survey techniques to understand practitioners' challenges and their experiences with fairness toolkits. As a result, both studies offer valuable insights regarding the usability and design of ML fairness toolkits. However, yet to be considered are the challenges faced by practitioners when using fairness toolkits to perform a task. The current study represents the **first task-based exploration** in the literature of how ML practitioners (try to) learn about and work with fairness toolkits and their APIs.

1.2 Open Source Fairness Toolkits

ML open-source fairness toolkits intend to assist ML practitioners in assessing and (potentially) mitigating unfairness in the ML systems they develop. Fairness toolkits usually offer ready-to-use fairness metrics and mitigation algorithms [114] as their main functionalities. In short, a *fairness metric* is a quantification of unwanted bias in training data or models. A *bias mitigation algorithm* is a procedure for reducing unwanted bias in training data or models. Some popular fairness toolkits include Fairlearn [16], AIF360 [14], Aequitas [91], Themis-ML [13], What-If Tool [109], Fair-ML [4], and Fair-Test [102]. In our study, we investigated how practitioners worked with two toolkits that were identified by Lee and Singh [63] as among the most useful and well-documented: IBM's AIF360 [14] and Microsoft's Fairlearn [16]. Both toolkits contain a Application Programming Interface (API), i.e., an interface that practitioners work with in order to communicate with toolkits [77]. We briefly introduce these two toolkits below.

1.2.1 AIF360. Developed by IBM, AI Fairness 360 (AIF360)[1] is an extensible open source toolkit for detecting, understanding, and mitigating algorithmic biases [14]. With over 71 bias detection metrics and 9 bias mitigation algorithms (suited for all aspects of the ML pipeline- from pre-processing to in-processing to post-processing), AIF360 is often commended for its breadth and depth in the coverage of fairness-related topics [54, 55, 63]. Despite this, IBM notes that the toolkit should only be used in a very limited setting: allocation or risk assessment problems with well-defined protected attributes. AIF360 is also part of IBM's current effort towards "trustworthy AI," an initiative focusing on creating a holistic approach to

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

governed data and AI technology. Currently, AIF360 has 1,600 stars on GitHub and 514 forks as we are writing this paper. Also, there are 1,220 members in AIF360 slack channel, which is primarily used by engineers who ask for help regarding their syntax errors and other bugs while using the toolkit.

1.2.2 Fairlearn. Initially developed by Microsoft Research and now being maintained as a community-driven open source project, Fairlearn[2] is an open source toolkit that empowers data scientists and developers to assess and improve the fairness of their AI systems that focuses on negative impacts—specifically, allocation harms and quality-of-service harms—for groups of people. The goal of Fairlearn is to create a vibrant community and resource center that provides not only code, but also resources such as domain-specific guides for when to use different fairness metrics and bias mitigation algorithms [16]. To this end, Fairlearn boasts a detailed User Guide, which is meant to be complementary to their standard API documentation. Fairlearn has 1,200 stars on GitHub and 280 forks as we are writing this paper. The Fairlearn development team holds a weekly community call which practitioners can join through Discord, a chat and networking platform.

2 METHODS

2.1 Think-aloud Usability Evaluation Interview

2.1.1 Participants. Before beginning our study with industrial practitioners, we first recruited two university computer science students and conducted semi-structured interviews as our pilot study to help us polish and iterate upon our study protocol. For our formal interview study, we recruited practitioners working on ML products and services through a combination of purposive and snowball sampling [49]. We invited an initial set of participants through our personal connections with industry practitioners, and we then asked them to help disseminate our recruitment message through their networks. We also included a brief screening form to help us target suitable participants for our study. We specifically recruited ML practitioners who had previously encountered fairness issues in their professional work, and who had *never* used a fairness toolkit prior to joining our study. Participants had on average five years of programming experience and three years of ML experience. This enabled us to observe how practitioners who were knowledgeable about real-world fairness challenges approached learning about fairness toolkits for the first time.

Overall, forty-one industry practitioners responded to our recruitment message. Twenty-three responded to our follow-up emails, among which fourteen participants scheduled an interview time. In the end, twelve participants attended the interview study, and eleven of them completed all phases of the study. Table 1. provides details about participant demographics and their relevant experience. Specific details about their companies and working environment have been abstracted to preserve anonymity. All participants (P1 - P11, U1, U2) were compensated with a $50 Amazon gift card upon completion of the study. The study was approved by our Institutional Review Board.

[2]https://fairlearn.org/

2.1.2 Study Design. With the goal of observing how industry practitioners formulate and attempt to resolve a fairness-related problem with the aid of open-source fairness toolkits, we designed a task setting that required participants to engage with a complex real-world context. Specifically, participants were tasked with building a model to determine which students were in need of additional tutoring resources. For this study, we chose the Student Performance dataset [29], which is concerned with academic achievement in Portuguese secondary education schools. Attributes include student grades as well as demographic, social, and school-related features [30]. Since we aimed to observe participants' thought processes during the Exploratory Data Analysis (EDA) and problem formulation stages, we desired a dataset that participants would be less familiar with. Hence, we intentionally avoided more commonly used datasets such as the COMPAS Recidivism Risk Score dataset [10]. We also elected to utilize this dataset due to its reasonable size considering our study time (around 650 instances), as well as its inclusion of multiple features (both categorical and quantitative) which satisfy prevailing notions of sensitive attributes [16, 40]. For example, the features concerning parent education and family educational support might prompt practitioners into investigating how socioeconomic aspects might affect student performance and how one should treat these features while building models. These intricacies allowed us to delve deeper into the justifications behind choices that participants may make (e.g., selection of sensitive features).

2.1.3 Interview Study Protocol. The interview study consisted of a pre-interview task and a 60 minute think-aloud semi-structured interview. We now document our procedures for the full interview.

Pre-interview task: Before entering the interview, we ask participants to complete a selection of preparatory tasks in the Colab notebooks (an collaborative computational notebook) we shared with them through email. Each participant received a notebook titled with a unique number string. After setting up the coding environment, we briefly described the Student Performance dataset and introduced practitioners to the task, in which they need to predict students' future school performance to help teachers distribute tutoring resources more efficiently. The entire pre-interview required practitioners to explore the data, translate the real-world problem into a machine learning one (problem formulation), train a machine learning model, and answer some questions regarding these steps in a pre-survey. The pre-interview task took 30 - 60 minutes to finish based on participants' self-reports. We share the Colab notebook we used *through this link* to help and inspire others conducting future relevant fairness toolkits evaluations.

Think-aloud semi-structured interview: During the live interview, we began by taking 5-10 minutes and asked participants to elaborate on their responses to the aforementioned questions. Then, we spent the next 30 minutes letting them explore two of the most widely known fairness toolkits [63], namely Microsoft Research's Fairlearn and IBM's AI Fairness 360 (AIF360). After participants spent roughly equal amounts of time exploring the two toolkits, we asked them to select one of the two toolkits to directly implement their fairness assessment and mitigation code in Python within the Colab notebook. Throughout the interview, we also asked participants to provide feedback on whether they could envision a

Participant ID	Academic Background	Industrial Role & Domain	Company Size	Location
P1	Computer Engineering	Researcher, Tech Company	50-249	US
P2	Computer Science	Data Scientist, Financial Services	25,000 or more	UK
P3	Computer Science	Machine Learning Engineer, Legal Tech	50-249	US
P4	N/A	Manager, Consulting Firm	5,000 - 24,000	Africa
P5	Computer Science	Applied Scientist, Tech Company	25,000 or more	US
P6	Human Rights	Researcher, Education	N/A	UK
P7	Sociology	Data Scientist, Retail	5,000 - 24,000	Sweden
P8	N/A	Data Scientist, Public Sector	25,000 or more	UK
P9	Computer Science	Data Science Manager, Oil/Gas	25,000 or more	US
P10	Business	Sr. Business Data Analyst, FinTech	25,000 or more	US
P11	Computer Science	Machine Learning Engineer, Legal Tech	5,000 - 24,000	US
U1	CS, Undergraduate	N/A	N/A	US
U2	CS, Undergraduate	N/A	N/A	US

Table 1: List of Pilot and User Study Participants.

setting where they used toolkits in practice, and the various obstacles which might be present if toolkits are to be integrated into their daily workflow. Importantly, as participants traversed through the API interfaces and implementing codes, we encouraged participants to "think aloud" [60, 105] and discuss the various information that was being displayed and how their understanding of the task (and ML fairness) was developing.

2.1.4 Data Analysis. We used an inductive thematic analysis approach [20, 27] to analyze approximately 11.5 hours of video recordings and their corresponding transcripts.[3] Two of the authors first worked together with a research assistant to conduct an open coding of the transcripts. We coded the same transcripts, discussed the code with the entire research team, then divided the rest of the transcripts and videos. We cross-compared and grouped these codes and observations into successively higher-level themes concerning the relationships between practitioners' practices and fairness toolkits' current functionalities and limitations. In Section 3, we discuss the findings identified from these codes and themes, together with implications for future fairness toolkits design. We share our initial round of thematic analysis *through this link*.

2.2 Anonymous Online Survey

We then conducted an online survey of industry practitioners to better understand the real-life drivers and obstacles to using fairness toolkits, and also to supplement the interview findings with larger sample size. The survey contained three main sections: (i) background and information questions asking participants' industry domain and relevant experience in ML, ML fairness, fairness toolkits; (ii) questions about fairness toolkits themselves; and (iii) questions about using fairness toolkits (and fairness in general) in the practitioners' workplace. All questions were optional.

This online survey was anonymous and did not ask for any directly identifying information. We emailed the survey link to direct contacts, as well as advertising it on online communities related to ML and ML fairness, e.g. those on Reddit, LinkedIn and Slack channels. We also encouraged sharing of the survey link to relevant practitioners. Of the 71 people who started the survey, 56 (79%) of the respondents completed at least one entire section, and 21(30%) completed the entire survey. Therefore, the sample size varies with each question (between 21 and 56). A similar drop-out rate was encountered by Lee et al. in their prior anonymous survey study [63]. Note that the questions would be difficult to contextualize for a respondent without a background in fairness-related challenges. This may have contributed to the drop-out rate, suggesting that few practitioners have relevant expertise or are not confident with issues of fairness. In addition, sensitive topics about organizational culture and workplace dynamics in the last survey section might also contribute to the drop-out rate. In spite of this, the opinions of this niche group are highly relevant. In addition, we were able to collect rich qualitative data from the free-text fields, in which many of the respondents discussed their experience with fairness toolkits. The anonymous survey data are available *through this link*.

[3]The interview videos and audios were recorded in line with participant consent.

3 FINDINGS

We present findings from our think-aloud interview study, divided into three main phases: preparing to evaluate fairness; exploring and learning about toolkits; and attempting to use toolkits. Across all three phases, we found that practitioners desired greater support from toolkits in contextualizing, communicating, and collaborating around ML fairness efforts. We supplement observations from think-alouds with results from our online survey. At the end of each section, we share implications for the design of future fairness toolkits.

3.1 Preparing to Evaluate Fairness

During the Exploratory Data Analysis (EDA) and problem formulation phases of our study, we observed that practitioners drew heavily upon their personal experiences to surface potential sensitive features. However, they also recognized the limitations of their own knowledge and experience, expressing desires for help from domain experts in formulating relevant and coherent fairness-related questions for a given real-world context. When formulating their

questions for fairness assessment, we observed that some participants appeared to be influenced by a toolkit's specific functionalities and limitations. We close this section by discussing implications of these observations for future toolkit designers.

3.1.1 Participants' analysis choices were heavily influenced by their personal experiences, knowledge, and beliefs. EDA is a critical step in ML development [85, 107] and an important opportunity to surface systematic errors and biases in datasets [22, 41, 93, 94]. We observed that during the EDA phase, participants often drew heavily upon personal experiences when making judgments about potentially sensitive features. For example, when explaining their concerns about the "father's job" and "mother's job" features, P5 said: *"it was less from the machine learning perspective, but more from my personal experience as a teacher before... if the parents have higher education, it gives some cognitive bias to the teacher."* We also observed that participants often relied upon general heuristics when deciding which features were potentially sensitive in a given domain. For example, nine out of eleven participants (P1, P2, P4, P5, P7, P6, P8, P9, P12) assumed that sex was a sensitive feature, believing this to be an obvious starting point. For example, as P8 said, *"the first thing is, of course, check the sex."*

Interestingly, most (seven out of nine) of the participants who assumed sex was a sensitive feature (P2, P4, P6, P7, P8, P9, P11) **attempted to mitigate biases in the ML pipeline by simply removing or ignoring the sensitive features like sex or address**. P9, for example, argued that *"I feel that sex is one of the sensitive [features]. To make the model fair, I'd rather just remove it before training (the model)."* This assumption has been discussed in prior literature as "fairness through unawareness." [34] In reality, omitting sensitive features may lead to more disparate outcomes in practice. Furthermore, none of these seven participants considered whether seemingly neutral features might be a proxies for other sensitive attributes.

3.1.2 Some participants (re)formulated the ML problem to a format for which current fairness toolkits provide more support. Through six months of ethnographic fieldwork with a corporate data science team, Passi et al. uncovered the "negotiated, not faithful, translation" of ML problem formulation which was affected by available tools and organizational resources [79]. In our study, we observed that **participants would sometimes reformulate problems based on current fairness toolkits functionalities and limitations**. For instance, after realizing that both toolkits offer better support for classification problems, five out of seven participants who had initially formulated regression problems during the pre-task phase decided to reformulate their problem as classification. For example, P2 emphasized that they would always prefer to use the example notebook toolkits offered as references whenever possible in their work. P2 then reformulated the ML problem from linear regression to *DecisionTreeClassifier*, the classifier used in Fairlearn example notebook, since *"it's easier this way to use FairLearn's example notebook as supporting material."* P4 switched to classification after realizing more comprehensive support for classification from both toolkits, adding that *"I would have just framed a classification problem if I [knew] in advance that [the] toolkit supports that more."* Survey participants expressed similar concerns through a free-text question about current toolkits limitations. For instance, one survey participant reported, *"It was hard to apply AIF360 to many of our models which are not binary classification; e.g., for regression models, there is not much guidance on if or how the toolkit should be used."* Only two participants in our think-aloud study (P5 and P9) raised concerns about reformulating the problem because of the toolkit's limitations. For example, P9 commented that *"type of model or problem should not be dependent on the functionality of the toolkits."*

3.1.3 Participants were conscious about the need for expert guidance to inform analysis choices at the EDA phase. We observed that **participants were eager for guidance from domain experts and other relevant stakeholders at the EDA phase**. For instance, P3 mentioned that they wanted to consult dataset builders about *"how the data was collected and how the features [were] being defined"* P1, P8, P10, and P11 all wished to present their data analysis results to domain experts in order to understand *"implications of technical concepts under different social context[s]"* (P1). For example, while explaining their EDA analysis results, P11 emphasized that they would *"definitely chat with education experts and legal experts even before the modeling."*

When identifying potential sensitive features, P7 pointed out that *"race is not common to have [as a dataset feature] in Sweden,"* noting that they would want to *"consult with domain experts before deciding which features could potentially risk introducing bias."* P4 mentioned the need to discuss *"which features might generate what type of bias with domain experts in the Portuguese education system"*. From our survey, of the twenty-seven respondents who answered the question regarding which experts they would engage on a fairness issue, twenty-one (78%) had *"legal and regulatory experts"* as one of their selections. Twenty (74%) selected *"business domain experts,"* and fourteen (52%) selected *"reputational risk experts."*

3.1.4 ***Implications.***

- **Broaden the scope of fairness toolkits to across the ML development lifecycle**. A large body of recent FAccT literature has highlighted the importance of exploring and analyzing datasets to surface potential biases [22, 33, 41, 71, 83, 93, 94]. However, current fairness toolkits offer little support for early stages of the ML development lifecycle, such as the EDA and problem formulation stages [28, 64, 64, 70, 89]. For instance, HCI researchers have explored the design of interactive interfaces like Facets [1], FairVis [25] and FairSight [5] to help users explore datasets and discover potential biases. To support the EDA phase, fairness toolkits should consider including similar interfaces in their designs. In addition, Boyd et al. found that context documents like *Datasheets* [41] could better scaffold an ML practitioner's process of issue discovery, understanding, and ethical decision-making around ML training datasets[19]. Future fairness toolkits could include instructions and educational materials on creating and reviewing *Datasheets* [41], and similar documentations like "Dataset Nutrition Labels" [50] and *Model Cards* [72] to better inform practitioners' data explorations.

- **Design fairness toolkits to facilitate interdisciplinary conversations and collaborations**. In our study, participants expressed desires for guidance from domain experts and other relevant stakeholders, to better understand the social and cultural context in which a given dataset or ML system is situated. Therefore, we suggest that future fairness toolkits might be explicitly designed as *social computing systems* that help to facilitate conversations between different toolkit users with diverse backgrounds and knowledge (e.g,. by connecting different toolkit users for peer support on an ad hoc basis). For example, toolkit users with technical backgrounds but lack certain expertise in law or gender study could seek help from relevant domain experts through the potential social computing functions offered by the future open-sourced fairness toolkits. Fairness toolkits could also introduce interactive, deliberation-driven design activities to encourage critical reflections and facilitate interdisciplinary conversations around ML fairness issues (cf. [96, 111]).
- **Show practitioners both patterns and anti-patterns for toolkit use**. Helping users recognize, diagnose, and recover from errors is essential to the design of usable APIs and toolkits [77]. In our study, a large proportion of participants committed to the "fairness through unawareness" trap [34]. Prior work suggests that timely, clear warnings and error messages can be effective in helping users avoid common pitfalls [77]. In the case of fairness toolkits, one possible design opportunity to help users avoid various "fairness traps" is to display contextual warning messages that target common pitfalls. In addition, in many domains like medicine [46], aviation [23], and structural engineering [39], checklists are used to support task completion, guide decision making, and prompt critical conversations among stakeholders [65, 67]. Well-designed fairness checklists may help practitioners avoid oversimplifying or forcing an ML problem formulation due to toolkits constraints [67].

3.2 Exploring and Learning about Toolkits

Through the "exploring and learning" phase of our study, we identified four major design requirements for fairness toolkits among participants, discussed below. Based on our observations, we discuss future implications for the design of fairness toolkits that can serve the needs of a diverse group of potential users.

3.2.1 Some participants wanted to be able to use toolkits as educational tools. . Instead of directly applying fairness toolkits to their current projects to address fairness issues, some participants were most interested in using toolkits to **learn more about ML fairness concepts and terminologies.** P2 said that for a fairness toolkit, *"ML fairness is something new... not directly [related] to my work and I just got into it maybe a few months ago... the most important thing for me is the [explanations of metrics] instead of the mitigation [code]."* While going through both APIs, P6 commented that, in their future work, they would like to use AIF360 as a starting point to learn more about different fairness metric definitions and relevant academic papers, in addition to their use cases. P10 pointed out that, since *"there is a lack of [resources] to learn more about fairness in [my] company"*, they appreciated that AIF360 offered a broad view of state-of-the-art ML fairness techniques, and included a designated reference section for convenience. Our survey results further supported this finding. For instance, one survey respondent commented on the lack of organizational process or *"official training and awareness raising"* around fairness issues. In addition, for the question: "What are some of the following options that most likely be your reason(s) to explore an open source fairness toolkit?," we had thirty-two out of thirty-nine (82%) selected "learn more about ML fairness concepts and terminologies", and twenty-seven (69%) selected "learn more about the typical process of dealing with ML fairness." Out of thirty-one survey respondents for whom these questions applied, only five (16%) said they have a defined process in their organisation for fairness, while fifteen (48%) said they do not have such a process, and eleven (35%) were not sure. Out of twenty-four survey respondents who specified whether their organization would provide sufficient time and resources to address a fairness-related concern, responses were nearly evenly split, with ten (48%) responding "No", and eleven (52%) responding "Yes." The drop-off of thirteen respondents at this stage of the survey may be due to the sensitive nature of this question.

3.2.2 Some participants wanted rapid onboarding to fairness toolkits due to workplace time constraints. . Some participants preferred to **learn the API functionality and on-board toolkits as quickly as possible**. Even before opening any toolkit's website, P3 shared, *"first, I will see if there is any quick tutorial I can go through."* When exploring Fairlearn, this participant then proceeded directly to the "Quickstart." Similarly, when exploring and comparing the two APIs, P4 and P8 revealed through their think-alouds that they were focusing on finding an existing notebook to follow so that they could begin using the toolkit as quickly as possible.

While it is possible that participants were incentivized to find a quick solution in order to complete our study in the allotted time, when asked, practitioners suggested that they would do the same in their day to day work because, as P4 said, *"there is always a time constraint in the real work."* Our survey results further supported that same time pressure existed in practitioners' daily work. Of twenty-four respondents who answered the final question of the survey, fourteen (58%) agreed that they would look for the fastest available solution when encountering fairness issues in their work. We further expand upon this point in Section 3.3.3.

3.2.3 Participants emphasized the importance of being able to integrate toolkits into their existing ML working pipelines. Validating findings from Lee et al [63], most (ten out of eleven) participants in our interview study emphasized that, as a precondition for adoption in their workplaces, **fairness toolkits must be easily integrated into ML practitioners' existing workflows.** On this note, eight participants (P1, P2, P3, P4, P7, P8, P10, P11) all mentioned the resemblance of Fairlearn to scikit-learn [84] (an open-source, BSD-licensed machine learning python library widely used by ML community) in terms of API classes and functions. Since scikit-learn was their go-to library to build and evaluate ML models, participants believed that this similarity could help them incorporate toolkits into their current ML pipeline. To illustrate, when learning and comparing two toolkit APIs, P3 commented that *"the first aspect that I am considering is how easily it can be directly concatenated*

in my current project." P3 appreciated that Fairlearn developers *"introduce this tool with very standard scikit-learn [syntax]. It will have more people interested in using this tool because it's already very aligned with what I'm comfortable using."*

3.2.4 Some participants wanted to use fairness toolkits as code repositories to build their own tool. Lastly, a few participants expressed desires to **understand the implementation details behind methods in a fairness toolkit, as a starting point to build their own tools.** When asked how they might use Fairlearn or AIF360 in practice, P1 and P11 both entered toolkits' GitHub pages to see whether it would be possible to use the toolkits' current implementations as references for implementing their own algorithms or toolkits. For instance, P11 needed domain-specific fairness assessment and mitigation methods for a recommendation system, which current toolkits did not include. Comparing the GitHub pages of two toolkits, P1 noted that AIF360's GitHub *"clearly listed out supported fairness metrics and mitigation algorithms."* Similarly, P11 felt that AIF360 *"has a more organized codebase to start with."* Among the nineteen survey respondents who said they would rather build or extend their own tools, fourteen (79%) of them selected the *"need to understand the low-level implementation"* as one of their reasons for not using an out-of-the-box tool.

3.2.5 ***Implications.*** Our study surfaced that practitioners seek to use fairness toolkits for diverse purposes, leading to different expectations for what fairness toolkits should offer. Here, we suggest possible directions for fairness toolkits to support practitioners' diverse needs.

- **Support practitioner learning within fairness toolkits**. Current toolkits are mainly designed to be used by practitioners for problem-solving [89]. Our findings suggest that some practitioners might look to fairness toolkits as convenient sites to learn about unfamiliar ML fairness concepts. Future fairness toolkits might be explicitly designed as learning tools, for example with designated pages or interactive modules that introduce ML fairness concepts, procedures, and best practices. Toolkits might proactively direct practitioners to these pages both when they first begin using a toolkit, and at critical points throughout their use of a toolkit. Support for such "just-in-time" learning could help alleviate challenges identified in prior studies on fairness toolkits, in which practitioners struggled to pinpoint the resources they needed the most [63, 89].
- **Better support practitioners in incorporating toolkits into their existing ML pipelines**. Echoing findings from Lee et al., our findings highlight the importance of supporting practitioners in more easily integrating fairness toolkits into their existing ML workflows [63]. One possible strategy, noted by participants in our study, is to align syntax and function nomenclature used in fairness toolkits with that of existing popular programming languages (e.g., python [104], R [86]) and software libraries (e.g., scikit-learn [84], Pandas [68], PyTorch [82], TensorFlow [3]) that are being widely used by data scientists and ML practitioners.
- **Adapt to different time constraints and scaffold responsible use of fairness toolkits**. ML practitioners are often operating under time pressure, with minimal or no organizational processes in place to support fairness work, and with little to no training around ML fairness [52, 66, 67, 69–71, 88]. In light of these pressures and constraints, fairness toolkits face a difficult design challenge. They must be carefully designed to keep the time barrier to entry low enough that practitioners will be able to use these tools in their work, but without lowering the barrier to an extent that promotes irresponsible use. Approaches such as adding contextual interface warnings or including built-in checklists, discussed in 3.1.4, may be helpful in achieving this goal. In addition, for practitioners who desire detailed and thorough examples or extensible code, toolkits should provide well-contextualized "worked examples" [89], and aim to provide a flexible and clearly documented codebase.

3.3 Attempting to Use the Toolkits

In this section, we document findings from observing how practitioners actually attempt to assess and mitigate a fairness-related problem through concrete usage of the toolkits. In doing so, we aim to understand obstacles that may hinder effective use of fairness toolkits in practice.

3.3.1 Participants desired better support for communication around fairness issues across different roles in an organization. Although current toolkits target individual developers, participants wanted to see **more functionality to support collaboration among multiple, diverse roles, including non-engineers**. After using the toolkits, we asked participants to explain their rationale for whether or not they would actually use a fairness toolkit in their own workflow. Before delving into the patterns we observed, we note that, at the time of our interviews, the Fairlearn toolkit supported a dashboard which many participants saw in example notebooks and tutorial documentation. Although this widget is no longer being developed by the Fairlearn team, we observed many valuable takeaways from participants' use of this Jupyter notebook widget. Several participants (P1, P3, P6, P8, P9, P11) commented on the usefulness of the dashboard visualizations. For instance, before seeing the dashboard P6 said that the toolkit *"was doing something to the data, but I can't really see what it's doing so that inherently makes me feel uncomfortable."* Then, when presented with the Dashboard, they immediately felt more at ease with the idea of using Fairlearn. Along these lines, P3 and P6 noted that the dashboard allowed them to easily scan for potential fairness issues, which would be helpful under time constraints.

Participants connected the need for dashboard-like functionality to a broader **need for visualizations in order to explain fairness toolkits to non-engineering roles within an organization.** For example, P1 and P2 noted that the dashboard makes it simple to convey the toolkit's message to a larger audience and therefore motivate others to further engage with the toolkit. P2 and P9 reaffirmed this by claiming respectively that *"a simple notebook format and compelling visualizations are needed for [organizational] leadership to adopt the toolkits"* and *"from a business standpoint, the [Fairlearn] Dashboard is more helpful than Jupyter notebook."* From our survey, when asked to what extent certain factors influence their decision on whether to perform fairness testing, of the

twenty-four people who responded, all but two people (92%) said "stakeholder demand" is important. This is supported by survey respondents who emphasized the role of external pressure, e.g. *"if [concerns are] raised by superiors"* or *"if there is a significant fairness issue, we would request more time and budget from the stakeholders."* Another survey respondent noted that *"often managers and software developers are faced with more complex realities"* and most out-of-box techniques that toolkits offer do not keep the *"use cases, regulatory landscape, and real-world deployment in mind"*. Therefore, to encourage broader use, a respondent from our survey commented that they expected fairness toolkits to *"provide relevant communication material (e.g. reports that can be circulated internally for the purpose of improving the software development process and/or scorecards that can be shared with external stakeholders like customers and investors)."*

3.3.2 Participants desired more actionable guidance. Fairness-related nomenclature in toolkits' documentation was often unintuitive to participants. For example, regarding bias mitigation algorithms, P7 claimed that *"[Fairlearn's] CorrelationRemover is more intuitive as opposed to [AIF's] DisparateImpactRemover."* Participants generally had a specific need in mind when delving into the documentation. P10 made the point that *"toolkits need to do a better job of explaining [fairness metrics] and why users would want to use them"*. For instance, P9 explained that it's important to *"situate users"* before delving into the intricacies of a fairness metric or bias mitigation algorithm. This is supported by a respondent from our survey, who said *"I think most fairness toolkits are tailored for data scientists with overly complicated metrics. This makes it hard to explain the metrics to the business stakeholders which most of the time are looking for something that is intuitive and meaningful."* As noted in the previous subsection, cross-functional communication is pivotal for fair ML efforts to have an impact in real organizational settings. To support this, we observed that participants often desired **more guidance and support in contextualizing toolkit functionalities or outputs, beyond the level of documentation provided by standard software packages.**

3.3.3 Participants frequently copied code directly from toy examples provided by the API. Although participants mentioned the need for domain expert guidance in tailoring a fairness analysis to a particular problem, during the code writing stage, most of our participants **directly copy-pasted code that they found in tutorial notebooks or toolkit documentation**. Only *one* participant (P7) attempted to reason through possible implications of their choices of particular fairness metrics or sensitive features. Although copy-pasting code from online examples is a common practice among developers when attempting to integrate a new software package into a working pipeline [58], doing so uncritically in the context of fair ML work may be particularly dangerous, given that fairness is highly context-dependent [31, 44, 95, 103].

3.3.4 ***Implications.***

- **Toolkits should provide use-specific and context-specific guidance.** Meaningful assessments are inherently contextual [28]. Use-specific guidance is therefore important for helping practitioners place the toolkit's offerings within the broader context of concerns, while encouraging 'cross-functional' collaborations where appropriate. For instance, a toolkit's documentation could indicate its relation to a specific legal or regulatory regime (e.g. GDPR, CCPA), which in turn can prompt interactions with colleagues in legal or compliance. Similarly, toolkits might contain context-specific guidance highlighting the considerations of various real-world applications. For example, toolkit developers could provide resources on the specific social and cultural contexts for a technical tutorial or example notebook. By doing so, toolkits could encourage practitioners to think more critically about the non-technical context in their approach to fairness-related issues. Offering context-specific guidance could also help practitioners avoid the issues with directly copy-pasting code from toolkits to their specific applications.
- **Opportunities for toolkits to support cross-functional collaboration and organizational buy-in.** Our results point to the need for toolkits to facilitate fairness related conversations and collaborations in a cross-functional setting, where not all team members will necessarily have much knowledge around either ML or ML fairness. It may be helpful, for example, for future toolkits to support such communication by generating visualizations and reports that are tailored for presentation to non-engineers. A large body of work in FAccT and HCI has surfaced how organizational factors (e.g., organizational cultures and incentives) impact ML fairness efforts in practice [52, 66, 67, 69, 71, 88]. Explicitly designing toolkits for use by *teams* and *organizations*, not just engineers, may help to address some of the challenges practitioners face in getting organizational buy-in and pushing fairness work forward.

4 DISCUSSION AND FUTURE DIRECTIONS

4.1 Fostering interdisciplinary communication and collaboration

To better address fair ML issues, prior FAccT research has highlighted the need for contextualizing technical ML fairness work through cross-domain collaborations [67, 81] and bringing in the lived experiences of real-world stakeholders, especially those from marginalized communities [52]. Our findings have shown that practitioners desire more support in contextualizing, communicating, and collaborating around ML fairness efforts, throughout the ML development lifecycle. From toolkit developers' perspectives, designing and creating a useful, context-specific example notebook also requires the help of domain experts. However, cross-functional collaboration in building potential solutions for ML fairness is intrinsically challenging due to background knowledge gaps [61, 112], conflicting values [99], ambiguous goals [31], and organizational barriers [66, 88]. As current fairness toolkits are mainly developed by and for practitioners with technical backgrounds in ML and software development, it is our hope that future fairness toolkit developers will explore a broader possibility space for the role that toolkits could play in ML fairness practice. For example, fairness toolkits could be explicitly designed to foster communication and collaboration across diverse roles within an organization, or could support practitioners in connecting with other stakeholders with

relevant domain knowledge or lived experiences outside their current organizations.

Designing toolkits to better facilitate meaningful interdisciplinary communication and collaboration beyond technical fair ML work could also help practitioners avoid the "solutionism trap" discussed by Selbst et al., which refers to the pitfalls of presuming a fairness-related issue can be solved through technical intervention alone [95]. Our think-aloud study engaged participants in learning about and using the fairness toolkits of today, which are currently designed to support relatively narrow, technical fair ML interventions. As a result, we centred on a technical approach for tackling a socio-technical problem (educational resource allocation), without leaving much space for participants to critically question whether an ML approach is appropriate in the first place. In practice, it is important to consider a broader space of potential remedies to fairness concerns. Future fairness toolkits might be designed to scaffold developers through broader reflection processes, for example by prompting them to ask "should we build ML in the first place?" or "could we solve this fairness issue solely through technical intervention?".

In Section 3.1.4, we briefly discussed opportunities for community building, building upon the social channels (e.g., community Slack or Discord channels) that existing toolkits already provide. Future work is needed to explore ways to build sustainable, diverse, interdisciplinary communities of practice. Toolkit developers interested in pursuing this vision may be able to draw lessons from other open-source community-building efforts [21, 47], or from recent work exploring ways to support collective algorithm auditing [6, 22, 24, 97]. In short, we advocate for future fairness toolkits to position themselves as socio-technical systems that enable more collaborative approaches to ML fairness practice.

4.2 Future directions for evaluating fairness toolkits

With more fairness toolkits being developed and deployed by research institutions and private companies, how might we support more effective and responsible use of fairness toolkits in the future? Beyond the design implications presented above, we hope that our work will inspire future empirical evaluations of practitioners' use of fairness toolkits, to empirically inform toolkit designs. Future work should aim to engage practitioners from more diverse regions [92], domains [52, 79, 108], and organizational contexts [67, 81], given that practices and challenges may vary significantly across each of these dimensions [57, 66].

As we observed in our study, with little to no training around ML fairness, several practitioners fell prey to the "fairness through unawareness" [34] trap (Section 3.1.1). Moreover, practitioners could seek for the most convenient solutions, rather than the appropriate ones, with minimal or no organizational processes in place to support fairness work and workplace time constraint(Section 3.1.3, Section 3.3.3), failed to consider the procedural, contextual nature of ML fairness [73, 74, 79, 95, 108]. Future toolkit designers should prevent practitioners from committing to other common ML fairness pitfalls discussed by prior research, for example, "fairness gerrymandering," [56] "solutionism trap," [95] and "formalism trap" [95]. To this end, toolkit designers should empirically study how practitioners apply fairness toolkits to more complex tasks with a wider range of datasets.

Our study investigates individual practitioners' use of fairness toolkits. However, as discussed in Section 3.3 when tackling complex, multi-faceted fairness issues in real-world settings, fairness toolkits need to support interactions and collaborations across diverse roles, including non-engineers (Section 3.3.4). Future work is needed to explore how *teams* in organizations *collectively* use fairness toolkits on real-world tasks. This could enable insight into the ways team dynamics might add additional frictions to toolkit use [90]. Only through longer-term ethnographic studies on in-house use of fairness toolkits can toolkit developers fully understand toolkits' uses, limitations, and potential impacts. Finally, we encourage fairness toolkit developers and researchers to not only use findings from such studies to iterate on toolkits' designs, but also to publish key empirical findings in the format of white papers, blog posts, and toolkit documentation, in the interest of communicating toolkit usage patterns and anti-patterns (cf. [45]).

5 CONCLUSION

In this study, we conducted the first empirical exploration of how industry practitioners (try to) work with fairness toolkits in practice. Through our think-aloud methods and accompanying anonymous survey, we found that practitioners needed support from toolkits in order to help them better contextualize fairness issues, as well as to assist them in fostering communication and collaboration with non-technical peers in their organizational settings. Additionally, we discovered numerous design implications for future developers of toolkits that seek to address complex, socio-technical problems. Beyond this, we hope our findings provide guidance for the creation of interdisciplinary communities, dedicated to providing a holistic space in order to combat fairness-related issues.

ACKNOWLEDGMENTS

We thank all industry practitioners who signed up for our interview study, replied to our recruitment emails, and finished the think-aloud study. We also thank anonymous industry practitioners who filled out our online survey. We would like to express our gratitude to Michael Madaio, Miro Dudik, Roman Lutz, Hilde Weerts, and Alex Cabrera for their feedback on different stages of this work. Finally, we thank our anonymous reviewers for their thoughtful feedback that help us further improve this work. This work was supported by the National Science Foundation (NSF) under Award No. IIS-2001851, CNS-1952085, IIS-2000782, the NSF Program on Fairness in AI in collaboration with Amazon under Award No. IIS-1939606, the award from Jacob Foundation for CERES network, the Carnegie Mellon University Block Center for Technology and Society Award No. 53680.1.5007718, Aviva and the UK Engineering and Physical Science Research Council (EP/R033501/1, EP/P024394/1).

REFERENCES

[1] 2017. Facets - visualizations for ML datasets. arXiv:1810.01943 https://pair-code.github.io/facets/
[2] 2021. People AI Guidebook. (2021). https://pair.withgoogle.com/guidebook/
[3] Martín Abadi and Ashish Agarwal et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[4] Julius A Adebayo et al. 2016. *FairML: ToolBox for diagnosing bias in predictive modeling*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[5] Yongsu Ahn and Yu-Ru Lin. 2019. Fairsight: Visual analytics for fairness in decision making. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1086–1095.

[6] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[7] Oscar Alvarado and Annika Waern. 2018. Towards algorithmic experience: Initial efforts for social media contexts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[8] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. *Guidelines for Human-AI Interaction*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[9] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 249–260.

[10] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016), 2016.

[11] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

[12] Joshua Asplund, Motahhare Eslami, Hari Sundaram, Christian Sandvig, and Karrie Karahalios. 2020. Auditing race and gender discrimination in online housing markets. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 24–35.

[13] Niels Bantilan. 2018. Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. *Journal of Technology in Human Services* 36, 1 (2018), 15–30.

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

[15] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[16] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI.* Technical Report MSR-TR-2020-32. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[17] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016), 4349–4357.

[18] Nigel Bosch, Sidney K D'Mello, Ryan S Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2016. Detecting student emotions in computer-enabled classrooms.. In *IJCAI*. 4125–4129.

[19] Karen Boyd. 2021. Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021), 1 – 27.

[20] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[21] Sue Lacey Bryant, Andrea Forte, and Amy Bruckman. 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In *GROUP '05*.

[22] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[23] Barbara Katrina Burian. 2006. Design Guidance for Emergency and Abnormal Checklists in Aviation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50 (2006), 106 – 110.

[24] Ángel Alexander Cabrera, Abraham J Druck, Jason I Hong, and Adam Perer. 2021. Discovering and Validating AI Errors With Crowdsourced Failure Reports. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–22.

[25] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FairVis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 46–56.

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

[27] Victoria Clarke and Virginia Braun. 2014. Thematic analysis. In *Encyclopedia of critical psychology*. Springer, 1947–1952.

[28] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2021. Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 598–609. https://doi.org/10.1145/3442188.3445921

[29] Paulo Cortez. [n. d.]. *Student Performance Dataset*. https://archive.ics.uci.edu/ml/datasets/Student+Performance

[30] Paulo Cortez and Alice Maria Gonçalves Silva. 2008. Using data mining to predict secondary school student performance. (2008).

[31] Sophia T. Dasch, Vincent Rice, Venkat R. Lakshminarayanan, Taiwo A. Togun, C. Malik Boykin, and Sarah M. Brown. 2020. Opportunities for a More Interdisciplinary Approach to Perceptions of Fairness in Machine Learning.

[32] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. *A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376638

[33] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society* 8, 2 (2021), 20539517211035955.

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

[35] Brian Ellis, Jeffrey Stylos, and Brad Myers. 2007. The factory pattern in API design: A usability evaluation. In *29th International Conference on Software Engineering (ICSE'07)*. IEEE, 302–312.

[36] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. " I always assumed that I wasn't really that close to [her]" Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 153–162.

[37] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118.

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

[39] Lincoln H. Forbes and Syed M. Ahmed. 2010. Modern Construction : Lean Project Delivery and Integrated Practices.

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

[41] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[42] Soumya Ghosh, Q Vera Liao, Karthikeyan Natesan Ramamurthy, Jiri Navratil, Prasanna Sattigeri, Kush R Varshney, and Yunfeng Zhang. 2021. Uncertainty Quantification 360: A Holistic Toolkit for Quantifying and Communicating the Uncertainty of AI. *arXiv preprint arXiv:2106.01410* (2021).

[43] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019). https://doi.org/10.1145/3359152

[44] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 903–912. https://doi.org/10.1145/3178876.3186138

[45] Philip Guo. 2021. Ten Million Users and Ten Years Later: Python Tutor's Design Guidelines for Building Scalable and Sustainable Research Software in Academia. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 1235–1251.

[46] Brigette M. Hales and Peter J. Pronovost. 2006. The checklist–a tool for error management and performance improvement. *Journal of critical care* 21 3 (2006), 231–5.

[47] Aaron Halfaker, R Stuart Geiger, Jonathan T Morgan, and John Riedl. 2013. The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist* 57, 5 (2013), 664–688.

[48] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.

[49] Douglas D. Heckathorn. 2011. Comment: Snowball versus Respondent-Driven Sampling. *Sociological Methodology* 41, 1 (2011), 355–366. https://doi.org/10.1111/j.1467-9531.2011.01244.x arXiv:https://doi.org/10.1111/j.1467-9531.2011.01244.x PMID: 22228916.

[50] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *ArXiv* abs/1805.03677 (2018).

[51] Kenneth Holstein and Vincent Aleven. 2021. Designing for human-AI complementarity in K-12 education. *arXiv preprint arXiv:2104.01266* (2021).

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

[53] Naja Holten Møller, Irina Shklovski, and Thomas T. Hildebrandt. 2020. Shifting Concepts of Value: Designing Algorithmic Decision-Support Systems for Public Services. *NordiCHI* (2020), 1–12. https://doi.org/10.1145/3419249.3420149

[54] Knut T Hufthammer, Tor H Aasheim, Sølve Ånneland, Håvard Brynjulfsen, and Marija Slavkovik. 2020. Bias mitigation with AIF360: A comparative study. In *Norsk IKT-konferanse for forskning og utdanning*.

[55] Brittany Johnson, Jesse Bartola, Rico Angell, Katherine Keith, Sam Witty, Stephen J Giguere, and Yuriy Brun. 2020. Fairkit, Fairkit, on the Wall, Who's the Fairest of Them All? Supporting Data Scientists in Training Fair Models. *arXiv preprint arXiv:2012.09951* (2020).

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

[57] Jon Kleinberg. 2018. Inherent trade-offs in algorithmic fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*. 40–40.

[58] Andrew J Ko, Robin Abraham, Laura Beckwith, Alan Blackwell, Margaret Burnett, Martin Erwig, Chris Scaffidi, Joseph Lawrance, Henry Lieberman, Brad Myers, et al. 2011. The state of the art in end-user software engineering. *ACM Computing Surveys (CSUR)* 43, 3 (2011), 1–44.

[59] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689.

[60] Julia Kupis, Sydney Johnson, Gregory M. Hallihan, and Dana Lee Olstad. 2019. Assessing the Usability of the Automated Self-Administered Dietary Assessment Tool (ASA24) among Low-Income Adults. *Nutrients* 11 (2019).

[61] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–35.

[62] Michelle Seng Ah Lee, Luciano Floridi, and Jatinder Singh. 2021. Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. *AI and Ethics* 1, 4 (2021), 529–544.

[63] Michelle Seng Ah Lee and Jat Singh. 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.

[64] Michelle Seng Ah Lee and Jatinder Singh. 2021. *Risk Identification Questionnaire for Detecting Unintended Bias in the Machine Learning Development Lifecycle*. Association for Computing Machinery, New York, NY, USA, 704–714. https://doi.org/10.1145/3461702.3462572

[65] Lorelei A Lingard, Sherry Espin, Barbara Rubin, Sarah Whyte, Marcela Colmenares, G. Ross Baker, Diane Doran, Ethan D. Grober, Beverley A. Orser, John Bohnen, and Richard Reznick. 2005. Getting teams to talk: development and pilot implementation of a checklist to promote interprofessional communication in the OR. *Quality and Safety in Health Care* 14 (2005), 340 – 346.

[66] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *arXiv preprint arXiv:2112.05675* (2021).

[67] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[68] Wes McKinney et al. 2011. pandas: a foundational Python library for data analysis and statistics. *Python for high performance and scientific computing* 14, 9 (2011), 1–9.

[69] Jacob Metcalf, Emanuel Moss, et al. 2019. Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics. *Social Research: An International Quarterly* 86, 2 (2019), 449–476.

[70] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–14.

[71] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25.

[72] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[73] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867* (2018).

[74] Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. 2019. This thing called fairness: disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–36.

[75] Lauren Murphy, Mary Beth Kery, Oluwatosin Alliyu, Andrew Peter Macvean, and Brad A. Myers. 2018. API Designers in the Field: Design Practices and Challenges for Creating Usable APIs. *2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (2018), 249–258.

[76] Brad A. Myers, Amy J. Ko, Thomas D. LaToza, and YoungSeok Yoon. 2016. Programmers Are Users Too: Human-Centered Methods for Improving Programming Tools. *Computer* 49 (2016), 44–52.

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

[78] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[79] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 39–48.

[80] Samir Passi and Steven Jackson. 2017. Data vision: Learning to see through algorithmic abstraction. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 2436–2447.

[81] Samir Passi and Steven J Jackson. 2018. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–28.

[82] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[83] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336.

[84] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[85] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. 2018. Data lifecycle challenges in production machine learning: a survey. *ACM SIGMOD Record* 47, 2 (2018), 17–28.

[86] R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[87] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2020. Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for shifting Organizational Practices. *arXiv preprint arXiv:2006.12358* (2020).

[88] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[89] Brianna Richardson, Jean Garcia-Gathright, Samuel F Way, Jennifer Thom, and Henriette Cramer. 2021. Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[90] Niloufar Salehi, Lilly C Irani, Michael S Bernstein, Ali Alkhatib, Eva Ogbe, and Kristy Milland. 2015. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 1621–1630.

[91] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018).

[92] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 315–328.

[93] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Kumar Paritosh, and Lora Mois Aroyo. 2021. "Everyone wants to

do the model work, not the data work": Data Cascades in High-Stakes AI.
[94] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37.
[95] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 59–68.
[96] Hong Shen, Wesley H Deng, Aditi Chattopadhyay, Zhiwei Steven Wu, Xu Wang, and Haiyi Zhu. 2021. Value Cards: An Educational Toolkit for Teaching Social Impacts of Machine Learning through Deliberation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 850–861.
[97] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 433 (oct 2021), 29 pages. https://doi.org/10.1145/3479577
[98] Korsuk Sirinukunwattana, Shan e Ahmed Raza, Yee-Wah Tsang, David R. J. Snead, Ian A. Cree, and Nasir M. Rajpoot. 2016. Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. *IEEE transactions on medical imaging* 35 5 (2016), 1196–1206.
[99] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019).
[100] Jeffrey Stylos and Brad A Myers. 2008. The implications of method placement on API learnability. In *Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering*. 105–112.
[101] Joshua Sushine, James D Herbsleb, and Jonathan Aldrich. 2015. Searching the state space: A qualitative study of API protocol usability. In *2015 IEEE 23rd International Conference on Program Comprehension*. IEEE, 82–93.
[102] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2017. FairTest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 401–416.
[103] Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
[104] Guido Van Rossum and Fred L Drake Jr. 1995. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.
[105] Maarten van Someren, Yvonne Barnard, and Jacobijn A. C. Sandberg. 1994. The think aloud method: a practical approach to modelling cognitive processes. *Knowledge Based Systems* (1994).
[106] Neil Vigdor. 2019. Apple card investigated after gender discrimination complaints. *The New York Times* (2019).
[107] Dakuo Wang, Q Vera Liao, Yunfeng Zhang, Udayan Khurana, Horst Samulowitz, Soya Park, Michael Muller, and Lisa Amini. 2021. How Much Automation Does a Data Scientist Want? *arXiv preprint arXiv:2101.03970* (2021).
[108] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[109] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.
[110] Chamila Wijayarathna and Nalin AG Arachchilage. 2019. An empirical usability analysis of the google authentication api. In *Proceedings of the Evaluation and Assessment on Software Engineering*. 268–274.
[111] Richmond Y Wong and Tonya Nguyen. 2021. Timelines: A World-Building Activity for Values Advocacy. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
[112] Bowen Yu, Ye Yuan, Loren G. Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. Keeping Designers in the Loop: Communicating Inherent Algorithmic Trade-offs Across Multiple Objectives. *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (2020).
[113] Amy X. Zhang, Michael Muller, and Dakuo Wang. 2020. How do Data Science Workers Collaborate? Roles, Workflows, and Tools. *Proc. ACM Hum.-Comput. Interact.* CSCW (Oct. 2020).
[114] Z Zhong. 2018. A Tutorial on Fairness in Machine Learning. https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb