"☑ Fairness Toolkits, A Checkbox Culture?" On the Factors that Fragment Developer Practices in Handling Algorithmic Harms

Balayn, Agathe; Yurrita, Mireia; Yang, Jie; Gadiraju, Ujwal

DOI
10.1145/3600211.3604674

Publication date
2023

Document Version
Final published version

Published in
AIES 2023 - Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society

"☑ Fairness Toolkits, A Checkbox Culture?" On the Factors that Fragment Developer Practices in Handling Algorithmic Harms

Agathe Balayn
Mireia Yurrita
a.m.a.balayn@tudelft.nl
m.yurritasemperena@tudelft.nl
Delft University of Technology
the Netherlands

Jie Yang
Ujwal Gadiraju
j.yang-3@tudelft.nl
u.k.gadiraju@tudelft.nl
Delft University of Technology
the Netherlands

ABSTRACT

Fairness toolkits are developed to support machine learning (ML) practitioners in using algorithmic fairness metrics and mitigation methods. Past studies have investigated practical challenges for toolkit usage, which are crucial to understanding how to support practitioners. However, the extent to which fairness toolkits impact practitioners' practices and enable reflexivity around algorithmic harms remains unclear (i.e., distributive unfairness beyond algorithmic fairness, and harms that are not related to the outputs of ML systems). Little is currently understood about the root factors that fragment practices when using fairness toolkits and how practitioners reflect on algorithmic harms. Yet, a deeper understanding of these facets is essential to enable the design of support tools for practitioners. To investigate the impact of toolkits on practices and identify factors that shape these practices, we carried out a qualitative study with 30 ML practitioners with varying backgrounds. Through a mixed within and between-subjects design, we tasked the practitioners with developing an ML model, and analyzed their reported practices to surface potential factors that lead to differences in practices. Interestingly, we found that fairness toolkits act as double-edge swords — with potentially positive and negative impacts on practices. Our findings showcase a plethora of human and organizational factors that play a key role in the way toolkits are envisioned and employed. These results bear implications for the design of future toolkits and educational training for practitioners and call for the creation of new policies to handle the organizational constraints faced by practitioners.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Human-centered computing** → **Empirical studies in HCI**; *User interface toolkits.*

KEYWORDS

algorithmic harms, algorithmic fairness, practices, organisational factors, human factors, fairness toolkits

AIES '23, August 08–10, 2023, Montréal, QC, Canada

ACM ISBN 979-8-4007-0231-0/23/08.
https://doi.org/10.1145/3600211.3604674

ACM Reference Format:
Agathe Balayn, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju. 2023. "☑ Fairness Toolkits, A Checkbox Culture?" On the Factors that Fragment Developer Practices in Handling Algorithmic Harms. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 08–10, 2023, Montréal, QC, Canada.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3600211.3604674

1 INTRODUCTION

It is now well-known that machine learning (ML) applications employed for decision-making might cause or reinforce distributive unfairness and other harms [3, 67, 69, 73, 100]. As a result, over the years, a great amount of theoretical research in ML has focused on conceptually understanding potential harms and on developing algorithmic methods to build ML systems that are less harmful [28, 100]. These methods, better known as algorithmic fairness metrics and unfairness mitigation methods, have lately been packaged into various *fairness toolkits* [6, 10, 23, 87, 97] to make it easier for their adoption by those who develop ML models (ML practitioners). A parallel line of research has investigated the practices of these ML practitioners, studying how they make use of proposed methods and what challenges they face. These studies are extremely important to understand how to further support practitioners.

Considering that the fairness toolkits are becoming a defacto standard means of tackling questions pertaining to algorithmic fairness[1] and potentially of teaching "ethical ML" to practitioners [13, 65], it is important to understand the extent to which practitioners rely on such toolkits, and whether and how toolkits shape their practices. Addressing this knowledge gap is a crucial step towards questioning the broad impact of fairness toolkits. A majority of past studies [24, 45, 57, 60, 77, 84, 85, 99] that have focused on the practices and challenges of practitioners in using the fairness toolkits have already identified a number of limitations of the toolkits in terms of design and technical specifications, that might hinder their adoption. However, such studies fall short in two major ways.

Fairness toolkits allow one to implement algorithmic methods for handling algorithmic unfairness. Yet, it is now well understood that these methods bear conceptual limitations [3, 43, 52, 58, 69, 88, 102]. Algorithmic unfairness observed in the outputs of an ML system is only a simplified representation of distributive unfairness in the world (what the metrics aim at quantifying), mitigation methods might themselves cause harm or not address the root causes of

[1] https://www.borealisai.com/research-blogs/industry-analysis-ai-fairness-toolkits-landscape/; https://www2.deloitte.com/de/de/pages/risk/solutions/ai-fairness-with-model-guardian.html

distributive unfairness, and other harms (beyond distributive unfairness) caused or reinforced by the use of ML systems are not accounted for by this framework (e.g., the purpose of the system itself might considered harmful, independently of the system's outputs being fair or not)[2]. None of the studies around practices and toolkits has however investigated how ML practitioners might conceive and overcome these limitations. It is especially unclear whether the toolkits narrow down practitioners' activities towards algorithmic unfairness and broader harms. These insights are necessary to envision where to focus future research efforts in terms of algorithmic harms beyond algorithmic fairness.

Additionally, prior studies do not report on differences of practices and challenges across practitioners, and the factors that cause these differences. Yet, identifying these differences, and grounding these differences into the *factors* that impact the fragmentation would allow one to identify the root causes of potential flawed practices and of certain challenges. This would allow one to envision more appropriate future solutions. In other words, explicitly looking into factors would allow one to answer the following questions: should fairness toolkits be our object of study to foster practices for handling algorithmic harms, i.e., are toolkits really the most important factor that supports and impacts practices around algorithmic harms (they would be if we would find a coherent set of practices across practitioners using a toolkit in comparison to those who do not)? Or are they only technical mediators of practices, that are impacted by deeper factors beyond the availability and design of the tool?

Hence, in this study, we ask the following two research questions: 1) How effective are toolkits in enabling practitioners to reflect about algorithmic harms and to handle them? 2) Which are the factors that affect the (in)effectiveness of toolkits in shaping practitioners' practices around algorithmic harms?

In order to answer these questions, we conduct 30 semi-structured interviews[3] with practitioners of various backgrounds. We compare practices before and after a practitioner is introduced to a fairness toolkit (within-subject experiment), and practices between practitioners who do not use a fairness toolkit to those who do (between-subject experiment), in order to understand the potential role of toolkits in shaping up practices. Besides, we further analyse qualitatively the interviews, and compare practices across practitioners, and across the two toolkits selected for this study, in order to identify potential additional factors that might impact practices.

For the participants of our study, we find that toolkits do increase awareness and use of algorithmic methods towards algorithmic fairness, do not impact considerations of algorithmic harms, yet can foster a checkbox culture with absence of reflexivity around the limitations of algorithmic fairness. More than solely toolkits, we also find that various human factors, such as types of training, and psychological and socio-demographic traits, as well as contextual factors, and especially organisational incentives, interact to shape up how practitioners make use of the toolkit, how reflexive they are around the limitations, and whether they conceive and tackle broader algorithmic harms. These factors, while they have been mentioned scatteredly across research publications that deal with perceptions of algorithmic harms [47] or the governance models of organizations around algorithmic fairness [84], had not been analyzed in detail in terms of their impact on the practices for the development of ML systems (with harms in mind). We then further discuss the implications that our findings bear when fostering reflexivity among practitioners towards avoiding algorithmic harms, e.g., in the form of design guidelines for fairness toolkits, as well as educational programs, and for further enforcing policy efforts towards making algorithmic systems less harmful.

[2]In the remaining of the paper, we use *algorithmic harms* to refer to any harm that ML systems might cause or reinforce, among which are *distributive unfairness* harms (related to the unfair ways in which resources are allocated following the recommendations made by the outputs of an ML system). We use *algorithmic unfairness* to refer to the limited conceptualisation of distributive unfairness in the lens of algorithmic metrics and methods developed by the scientific community.

[3]All our materials, resulting data, code and analysis will be shared publicly. https://osf.io/dmr82/?view_only=a00e68796f494fbb9776cf9a95fb7051

2 RELATED WORK

2.1 Fairness Toolkits for dealing with Algorithmic Unfairness

2.1.1 Algorithmic Unfairness. Each step of the machine learning (ML) lifecycle might create or reinforce *distributive unfairness* [67, 94]. Theoretical works have primarily developed *algorithmic fairness* metrics [100] that aim at measuring distributive unfairness in the outputs of the final model or in a dataset. These works also propose algorithmic unfairness mitigation methods [4, 28] that ought to improve the model's algorithmic fairness as defined by the metrics. Facing the diversity of metrics, the challenge for a practitioner is to choose the appropriate one for their task.

Several studies have investigated how ML practitioners work with algorithmic fairness metrics and mitigation methods. Topics of focus revolve around general challenges met by practitioners [22, 45, 60, 71, 74, 77, 84, 89, 99, 103], and obstacles and limitations for the application of algorithmic fairness methods. Findings outline the need to support practitioners to concretely use fairness methods, as this use is challenging due to the context dependence of methods, the current lack of guidance [45, 60], and the need for adapting methods that are incompatible with targeted tasks [45].

2.1.2 Effectiveness of Fairness Toolkits. To facilitate the adoption of algorithmic fairness metrics and mitigation methods, various companies and public institutions have built fairness toolkits. These toolkits are typically code repositories that allow for an easier implementation of the metrics and methods. Examples of these toolkits are FairLearn [10], AIF360 [6], Aequitas [87], Themis-ML [5], ML-Fairness Gym [23], TensorFlow Fairness Indicators [107], etc.

Various works [24, 57, 85] have shown through interviews the beneficial use of toolkits by practitioners for developing fair models and learning about algorithmic fairness. Yet, they also show their limitations in terms of support provided to practitioners for designing the right algorithmic fairness evaluation, noting that participants often inappropriately change their modeling task definition to fit existing tools. These works also identify obstacles to the application of the toolkits in terms of compatibility with other ML frameworks and usability, summarized into toolkit checklists that should inform the design of future toolkits. We will show that our results corroborate and complement these insights. Indeed, to the best of our knowledge, our work is the first to investigate

(or report) whether the toolkits do impact practices contrary to a situation where no toolkit would be available, whether there are differences in practices of different practitioners using a same toolkit, or whether different toolkits lead to different practices.

2.2 Fairness Toolkits for reflecting on Harms Beyond Algorithmic Unfairness

2.2.1 Algorithmic Harms. A few theoretical works have looked beyond algorithmic fairness to identify other harms of ML [3, 69]. We now present a few of these harms that are highly worthy of consideration according to the literature. Algorithmic fairness metrics and methods bear conceptual limitations, that do not allow one to comprehensively gauge the distributive unfairness they are aimed at addressing. By limiting harms to the frame of output distributions (also termed distributive justice fairness), algorithmic fairness cannot reflect the contextual factors that influence what is considered fair. For instance, it assumes that parity is always desired in the model outputs [58], it does not account for the impact one same output has on different receivers of this output [69], nor for the indirect impact on non-data subjects [52]. Looking at the process to reach algorithmic fairness (termed procedural justice), the metrics and mitigation methods do not make sure that the way in which the unfair situation is addressed is aligned with moral principles [102]. For instance, individuals or groups might see low disparate accuracy by all receiving unjustified treatment [72], or by all being treated differently (e.g., post-processing methods allocate different decision thresholds for different groups) which consists in direct discrimination [35].

Three other categories of harms have also been discussed. First, ML requires to use *datasets* whose schemas and sampling can be harmful. For instance, certain attributes and their values might be offensive [11, 108] or inappropriate [67], e.g., use of non-volitional or privacy-infringing attributes [39, 95]. Second, research questions the *desirability of the ML model* in the first place, its use for undesired applications [46, 48, 69, 70], and how it impacts structures in place [27]. Using ML for certain tasks might be questioned, for instance because it means making decisions for people by comparing them to others instead of following the principle of individual justice [9, 26], or because it reproduces historical, potentially harmful, data patterns [81]. Third, certain researchers question the *negative externalities caused by the production process* of ML applications, such as the environmental impact of data centers and model training [7, 17], the poor labor conditions of crowd workers [86, 105, 109, 111], the privacy-infringing training data [82], etc.

2.2.2 Effectiveness of Fairness Toolkits. Besides investigating the effectiveness of toolkits in enabling reflexivity around algorithmic unfairness, it is important to acknowledge the known limitations of the algorithmic fairness methods and the existence of other algorithmic harms that ML systems might pose. To the best of our knowledge, no work has investigated practices in relation to these limitations. We do not know to what extent the use of fairness toolkits —that foster the use of the algorithmic fairness methods— impacts considerations of algorithmic harms and of the limitations of algorithmic fairness (that are typically obfuscated from the toolkits). It is unclear whether fairness toolkits, that do not deal with these harms, might lead practitioners to "forget" them.

2.3 Factors Affecting the Usage of Toolkits

The effectiveness of fairness toolkits in enabling reflexive practices among ML practitioners around algorithmic unfairness and harms is conditioned by factors that shape the usage of these toolkits. Research into the characterization of these factors is still scarce. It is important to understand which factors make practitioners choose one metric or the other, and more broadly, to identify the factors that impact the decision of practitioners to try quantify unfairness, and later to mitigate it. The factors that lead a practitioner to handle broader algorithmic harms have also not been investigated in the past. Knowledge of these factors could allow one to better understand the deeper nature of the challenges faced by practitioners, and to provide more personalised support to these practitioners.

Up to now, studies have solely identified organisational factors, that are further shown to be obstacles for practitioners to develop fair models [60, 62, 84, 99]. Contrary to our work, previous studies had not accounted for human factors in their study design or in their result analysis, such as Deng et al. [24] who only reported on coarser-grain practices (e.g., they reported that the practitioners they interviewed recognize the limitations of their knowledge and wish to receive help from domain experts, but do not specify any difference across these practitioners). In our study, we find such factors, and also investigate the existence of technical ones.

3 METHODOLOGY

To characterize the effectiveness of fairness toolkits in enabling reflexive practices, and to identify the factors that might impact and fragment those practices, we adopted an empirical and qualitative approach via 30 semi-structured interviews with ML practitioners. By comparing practices within-subjects (participants are observed before and after receiving an introduction to fairness toolkits), we observe the extent to which toolkits enable or hinder reflexivity. Additionally, by comparing practices in-between subjects who bear different characteristics (e.g., background and prior experiences) and who use different toolkits, we characterize the fragmentation and delve further into the contributing factors.

3.1 Participants

We recruited our participants in the period of April-June 2022, by means of personal networks, targeted requests on social media, calls for participation on the official Discord or Slack communication channels of the toolkits, LinkedIn, and snowball sampling. The participants received no financial compensation, and their contributions were voluntary (they typically participated to learn more about algorithmic harms, and to help science progress). Our institution's ethics committee approved the study. All participants signed an informed consent form acknowledging the risks involved with participating, as well as agreeing to the interview being recorded (all interviews were conducted online), transcribed, anonymized, destroyed, and consented to the results being used in scientific publications.

A total of 30 participants were recruited across research and industry institutions, and across application domains such as healthcare, finance, and predictive maintenance (cf. supplementary material). Manual sampling was performed to make sure that all participants have responsibilities in ML model development, deployment,

or evaluation; varying levels of prior experience with ML, ranging from 2 to 15 years; and varying practical experience with algorithmic fairness and fairness toolkits (11 participants already had experience with FairLearn, and 9 with AIF360). The resulting participants differ in terms of demographic background (nationality, gender, and age), level of highest education, educational background, and type of training received around ML. Besides, participants already experienced with algorithmic fairness presented variations in terms of how they learned about the topic, the kind of experience they have had, and for how long they have worked with these issues (from 0 to 18 years).

3.2 Interview Procedure

The interviews with participants already familiar with a toolkit lasted one hour each, going through Task T1. The interviews with the other participants lasted around two hours each, through three stages (Task T1, a tutorial about one fairness toolkit, and Task T2). These three stages were designed to identify how the use of toolkits might impact practices around algorithmic harms. Comparing practices between participant groups with or without prior familiarity with the toolkits allowed us to unveil other influential factors, such as the type of training received around harms. In total, we collected 2207 minutes of recording. In Figure 1, we show the workflow of the interviews with the questions asked in each stage, for the two kinds of participants. We asked three types of questions: background experience questions (demographics, experience with ML and algorithmic fairness); reflection questions around algorithmic fairness, harms, or toolkits, and around general comments, wishes, doubts, and challenges the participants might have about their workflow or harms; and process questions to understand the reasoning behind each participant's activities during the tasks (cf. supplementary material for details on tutorial and questions).

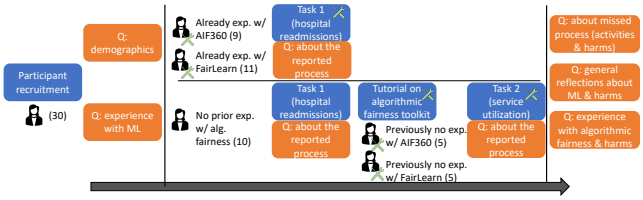

Figure 1: Interview procedure for the participants already experienced with a fairness toolkit, and for the participants who did not have any prior practical experience with algorithmic fairness. In blue: the main steps of the procedure; in orange: the questions posed in each step.

3.3 Materials

Use-Cases. We chose two use-cases, the first one involving the prediction of *hospital readmissions* within 30 days for individual patients [93], referred to as Task T1, and the other involving the prediction of low or high *medical services utilization* [42], referred to as Task T2. Using these tasks instead of discussing the participants' own use-cases was important to be able to rigorously compare practices over the same case and to surface the factors that impact practices for a same use-case. We pre-processed the two corresponding datasets for them to have similar characteristics (number of attributes and of records), and to be prone to similar harms (cf. supplementary material). By employing comparable domains and datasets without re-using the exact same use-case for the two tasks of the interviews, we aimed to minimize learning effects. We chose the domain of healthcare because it is prone to various harms, requires expertise to be handled correctly (i.e., we could check whether the participants mentioned the limits of their knowledge [24]), several corresponding datasets were available, and these are not the most frequent use-cases in the algorithmic fairness literature which allows us to minimize the confounding effect of familiarity with the domain of application. Our choice also allows us to mimic a realistic situation, where oftentimes, practitioners have to develop or deploy models without having extensive expertise in the domain of application. In such cases, practitioners' decisions might lead to harms, that fairness toolkits are meant to empower practitioners to reflect about.

Tasks, Toolkits, and Notebooks. For each task, we shared a Google Colab notebook with the participants, which included a design brief with one of the two datasets pre-loaded. The design brief mentioned that a hospital (or an insurance company) wanted to optimize their cost and services (or their prices), and therefore wanted to investigate whether ML could help them predict readmissions (or utilization, respectively). The institution tasked the participant to investigate this feasibility possibly using the dataset they had collected, and to report on their findings by speaking outloud. Along the investigation, when participants mentioned some code-based exploration, we shared corresponding code snippets prepared before the interviews to speed up the process.

For the interviews with practitioners who had used a fairness toolkit in the past or with the ones we introduced to a toolkit, we loaded a specific toolkit (FairLearn [10], or IBM AIF360 [6]) into the notebook, that they were most familiar with. We consider these toolkits because they contain a large number of functionalities around algorithmic fairness; they are the most studied toolkits in research [24, 57] and appear to be popular among practitioners. Cf. supplementary material for details about our interview materials.

Analysis of the Transcripts. We analysed the transcripts using a combination of inductive and deductive coding. The first author identified the segments discussing the main themes we wished to discuss (e.g., the harms, their conceptions, identification, and handling, and toolkit use), and coded any other emerging themes (e.g., other factors that practitioners trade-off when developing ML models) in collaboration with four other researchers. Then, the author in discussion with the other authors, reconciled redundant codes. Finally, this first author studied each of these codes based on their associated participants. While we cannot certainly identify which factors cause observed variations in terms of conceptions and practices based on our qualitative study, certain practitioners explicitly mentioned potential factors that we report. We also explore quantitative differences based on the background information we have about the practitioners (yet, all the factors are impacting practices in different ways, that we cannot explore within our study).

4 RESULTS

4.1 On the Effectiveness of Toolkits

In terms of algorithmic unfairness, practitioners reported the toolkits to be extremely useful for them to quantify and mitigate unfairness, what was confirmed by our observations. Yet, we also identify drawbacks of the toolkits for distributive unfairness, that we describe next. In terms of algorithmic harms beyond distributive unfairness, we did not note any evidence of positive or negative impact of the toolkits on practitioners' considerations and practices.

4.1.1 Effectiveness of Toolkits. Among toolkit-inexperienced practitioners, toolkits fostered a positive shift in practices around algorithmic fairness between task T1 and their introduction in task T2. Before being introduced to the toolkits (T1), it was not natural for the practitioners to reflect about algorithmic fairness. After our tutorial (T2), they began discussing potential unfairness caused by the outputs of their models and trade-offs between different fairness metrics and with accuracy, to judge which model is satisfactory (even if superficially on occasion). They also started envisioning approaches to mitigate the potential issues with the outputs. Hence, toolkits, for these practitioners, represent a means to foster awareness around distributive unfairness and its causes. *P19: "Just seeing how it worked, made me realize that it's not only about the dataset, but there's bias everywhere."* It also represents a means to learn about existing solutions to mitigate unfairness, and a prompt to start actively tackling the issue (being readily-available code repositories, toolkits lower the entry-barrier to the problem). *P17: "If it's quick and easy, run a quick check. 'Oh, there is something there I didn't think of. I need to explore that.' I could see that happening."*

As for toolkit-experienced practitioners, they primarily use toolkits to speed-up their processes around algorithmic fairness, and to foster communication with other stakeholders. *P11: "I talk to business people and this is how they can connect to this topic from the technical side because they can't code or anything."*

4.1.2 Undesirable Consequences of Toolkits: Reducing Harms to Algorithmic Fairness. Despite their perceived utility, toolkits can be misleading, and create a gateway to a narrow view on distributive justice. 6 out of 10 participants who were inexperienced with fairness, 4 out of 9 relatively more experienced ones, and 2 out of 11 very experienced ones took the toolkits at face value. They applied all fairness metrics available through the toolkits without considering their meaning and appropriateness, declared a model satisfying if certain values of (often arbitrarily picked) fairness metrics were reached (sometimes operating a non-informed balance between accuracy and fairness metrics) without reflecting on their limitations. *P13: "With the use of toolkit, I don't think my view changed. [Before having the toolkit,] I already believed in what the techniques could do. So if the toolkit correctly implements techniques, I have faith in it."*

55% of practitioners who were more experienced with fairness explicitly expressed concerns surrounding the toolkits. Toolkits might narrow down critical thinking around what is measured in relation to distributive fairness and be misleading, limit reflections on broader socio-technical concepts, and foster techno-solutionism triggered by the development of unfairness mitigation methods. *P22: "You cannot rely on the toolkit. You need to understand the problem and the domain knowledge. I can easily see these toolkits like before metrics like precision, recall were just thrown at random without knowing the actual meaning. Things like statistical parity difference, as they become more common, I can see them being misused because a lot of people don't even know their definitions. It's easy for people to misinterpret them."* Practitioners also felt that toolkits encode biases in their setup. *P23: "These libraries can introduce some biases that you are not aware of, so you don't need to put all the chances on those libraries, you should look into data yourself to see what type of bias data contains."* All in all, toolkits might illegitimately serve as a checkbox. *P3: "Fairness for many companies is just a small checkbox, and sometimes people put their mark without any question. I hope there will be a time when they understand that fairness is not about code and just picking up one toolbox. [..] The toolkits would constrain your view if you're using them blindly."* This is in direct contradiction with the way a few participants perceive the toolkit as an opportunity to realize and convey the complexity of the distributive justice problem *P21: "The recurring theme of our conversation is that fairness is difficult, and this realisation is what the toolkits achieve. They give a large variety of options to make fair models, but their biggest positive impact is helping practitioners realize that this is not a topic where we just do the same five steps and we have a fair model, but it's something that requires a lot of consideration."* This is evidence that beyond the toolkit itself, there are additional factors that impact practices –we discuss them next.

4.1.3 Technical Factors: Differences across Toolkits. We do not find any notable difference in the conceptions of harms between practitioners who used different toolkits, irrespective of their experience with fairness. While in practice some functionalities (metrics and mitigation methods) are only supported by one of the toolkits, this did not appear to be a major obstacle to the practitioners, who seemed to use other methods when needed (some practitioners also mentioned having to design novel methods to tackle their problems). This could however potentially be dangerous for beginner practitioners who learn about algorithmic fairness solely through the toolkits, and may revert to sub-optimal metrics and methods.

Practitioners did mention factors that impact the adoption of toolkits: compatibility with existing frameworks and code, frequency of maintenance and open source nature, ease of adoption and learning curve, transparent implementation and documentation, amount of functionalities and adaptability to various use-cases, and socio-technical questions the toolkits foster (cf. supplementary material for details about these factors and the others we identify). Interestingly, these mainly refer to non-functional requirements. While practitioners agree on these requirements, the evaluation of the satisfaction of a requirement for a toolkit was sometimes contradictory across practitioners when choosing one toolkit over the other (oftentimes, practitioners did not know both toolkits, but used similar arguments for explaining the choice of one over the other), e.g., they mentioned choosing AIF360 or FairLearn both because of their compatibility with existing coding frameworks.

4.2 Human Factors

Finding out that the toolkits are not the only factor that substantially fragments practices, we turn to the human factors and the specificities of each practitioner to understand observed variations.

4.2.1 Experience in Algorithmic Harms. As already mentioned, the amount of prior experience with algorithmic fairness (which includes experience with fairness toolkits) seem to impact practices on average. Relatively inexperienced practitioners typically think of fewer harms and reflect on issues with less critical attitude, and more often solely relying on their intuition, than the more experienced practitioners. Most participants who are just entering the realm of distributive fairness through a toolkit are not very critical about algorithmic fairness. *P20: "Using it this way seems to be one of the best ways, taking into account what I knew before, and what I learned today about the toolkit."* They become more critical if they accumulate more practical experience and knowledge by further exploring the toolkits' guidelines. Hence, more than the mere amount of experience, the type of prior experience with algorithmic fairness is a factor that seems to strongly impact practices. For instance, practices among the most experienced practitioners do vary, with some also relying solely on sometimes flawed intuitions (e.g., removing samples with missing values always improves the ML model performance), while others systematically involved external sources of information and rigorous computations (e.g., other stakeholders, laws, guidelines, business) and potentially make use of statistical tests.

4.2.2 Ways of Learning about Algorithmic Harms.

Types of Interactions with Others. The practitioners who displayed a more critical attitude discussed having learned about distributive fairness through interactions with various stakeholders. For instance, half of the participants who have learned about the metrics primarily through the code and 70% of the inexperienced participants who only briefly learned about the metrics during our interview discussed observing all metrics without reflecting on their meaning, while all the ones who have had more interactions with the research community (7 participants) or other interdisciplinary teams (3 participants) judged choices based on use-cases. These interactions (discussions, workshops, and conferences) often involve colleagues, clients, or researchers in AI ethics that highlight potential limitations and critical attitude to keep, or illustrate the subjectivity of the topic. *P3: "We invited one developer of FairLearn to run workshops. Her message was clear: you can ingrain fairness in code, but if you don't understand what you're doing, you will be in the world where we are already."* Similarly to previous results showing that discussions can positively impact fairness considerations [66, 79], the participants we introduced to the toolkits also mentioned the benefits of our discussion (to make them conscious of potential harms and of the limitations of their own, often non-critical practices), more than the one of the toolkits. *P20: "[Do you feel like your perspective on algorithmic harms changed after after seeing the toolkit?] Yes, I mean more after this discussion altogether. I personally wouldn't have taken some of them into account myself if I weren't pointed in the right direction by your questions."* Our participants reflected about the choice of fairness metrics and mitigation methods, once we explicitly prompted them about specific use-cases and actual meaning of different choices. *P28: "You also mentioned proxy. And I realized that just protecting some variables doesn't mean that you have removed completely that bias."*

Types of Courses. Other practitioners learn about various harms and algorithmic fairness by reading literature (e.g., P9 mentions the diagram from the Algorithmic Justice League) or by following courses on ML in general, on AI ethics, or on ethics of technology. The way the course is taught seems to impact practices, as one practitioner discussed having been trained through use-cases and was able to identify a number of harms, while four others mentioned a few ML ethics courses with toolkits introduced during the courses but did not reflect on any harm during the interview.

Importance of the Design of the Learning Material. While practitioners learn and develop their experience with ML and algorithmic harms via various means, leading to various practices, they also seem to interpret differently the same material, sometimes leading to misconceptions. While we discuss in a later subsection relevant human factors, we emphasize here the importance of the framing of the materials around harms. For instance, certain initiatives, although having a legitimate aim —warning against issues or proposing relevant approaches— sometimes had the inverse effects, and narrowed down the view of the practitioners towards related harms. This was especially the case for the recent "data first" approach advertised by different research communities [2], that led certain practitioners not to understand that model design might also create algorithmic unfairness; *P22 "I talk about the data quality first like Dr. Andrew Ng says. Data-driven ML is becoming very prominent."* Similarly, P9, P16, P23 learned about model energy-consumption issues by reading the "Stochastic Parrot" paper [7], leading them to acknowledge these issues solely for large language models, but not for other types of simpler ML models.

Next to the framing of harms, the vocabulary employed (e.g., "bias", "sensitive feature", "protected attribute") also revealed to be a source of confusion and flawed practices. For instance, certain fairness-inexperienced practitioners only conceived "biases" as statistical skews without relations to, e.g., sensitive attributes or harms *P30 "with medical instruments, for a specific machine, there is some specific noise in the data. If you know which machine measured the blood pressure, then you know the bias in the data."* Some expert practitioners even warned about issues with loaded terms.

4.2.3 Disciplinary Experience.

ML Experiences. The amount of experience with ML also seem to be an impacting factor for practices around algorithmic harms. We observed that practitioners who have longer experience with ML (independently of having experience or not with algorithmic harms) reflect about more harms, more in-depth, and often envision more diverse mitigation methods than less experienced practitioners. For instance, three of those practitioners without experience around fairness were able to envision potential harms from the model design, and naturally evaluated the model based on subgroups of population without knowing the concept of equalized odd, whereas practitioners relatively inexperienced in ML with some algorithmic fairness training often did not account for this. Three participants who had extensive experience with data science but were inexperienced with fairness and three mildly experienced ones were also more critical about the toolkits. *P18: "You always need to question existing tools and practices to be able to improve and innovate."*

Experiences with other Fields. Three practitioners who have not only studied ML or data science emphasized the potential benefits of their background: a participant trained as an ethicist; another trained in industrial design *P1: "This is my industrial engineering background talking. Let's map out the process to see, if we would be using a model, where it would fit in the current process and what requirements might be there? Is this supposed to be a fully automated system? How are people going to use this system? [..] For that, I talk to people. Can you imagine yourself saying that? [sarcastic remark about computer scientists]"*; and a last one in sociology *P29: "that's why they hired me: someone who's both good on the computer science side and on this sociology side."* These participants indeed identified more relevant harms and presented a more critical attitude towards their own activities, reinforcing the importance of involving multiple stakeholders with a diversity of backgrounds when the ML practitioners themselves do not have the relevant education.

4.2.4 Personal Factors. As we hinted at earlier, practitioners might behave differently even when presenting similar prior training and experience, within similar contexts. This hints at the existence of additional human factors that impact practices. Especially, non-volitional, socio-demographic factors were explicitly reported by practitioners as drivers of certain practices, such as gender, nationality, and culture that impact their ways of perceiving harms. Belonging to a minority might also change the lived experiences and efforts put onto harm mitigation. *P13: "I felt my obligation because I participate in many unprivileged classes. So I would like another person to do it for me."*

Although not always directly observable via our interviews, other factors (e.g., psychology traits, abilities, and the resulting personal interests) appeared to be at play. For instance, when asking the practitioners to envision potential limitations of fairness metrics and mitigation methods, many of them could neither envision any conceptual one, nor see the potential risks of distribution shifts (that is a more technical and well-known topic –mentioned by only 20% of the participants). Similarly, when we prompted the participants to reflect broadly about their approaches, many did not envision or acknowledge any potential limitation. Yet, some participants showed more reflexivity, accurately recognized being biased and having to make subjective, uninformed choices, and acknowledged the complexity and subjectivity of the choices they make. *P20: "I'm sure that there is a possibility to create bias if I create features based on my interpretation of the data or what I think in my subconscious about people that get ill."* A few (also recognized not really knowing the potential impact but potentially keeping the benefice of the doubt. *P4: "For hyperparameters like learning rate, I can't see the connection with how it might harm people because it just influences accuracy. But I'm hesitant to say it doesn't affect it at all because you never know with these things, so you should always be cautious."*

4.3 Contextual Factors

Along the interviews, practitioners also mentioned a number of organisational factors that represent obstacles or impetus towards handling questions of algorithmic harms.

4.3.1 Incentives and Support. Several participants discussed monetary incentives (financial compensation) and non-monetary incentives and opportunities (possibility to get dedicated time for investigating harms), or the lack thereof, provided by their organization, that impact their considerations and actions. *P14: "the challenge is that, from a legality compliance and the organization perspectives, the appreciation should be there for you to spend the time."* Several participants mentioned engaging in volunteer work in their organization, in order to setup trainings and tools for tackling harms, or directly investigate harms for their own ML projects.

Others also reported on the material support (or the lack thereof) provided to them to facilitate tackling algorithmic harms. They especially mentioned the access to convenient tools (such as the fairness toolkits), and education around the topic (e.g., via the participation to workshops and seminars ordered by the organisation). Human support was also reported, especially the facilitation of the access to various relevant stakeholders (e.g., domain experts, decision-subjects, researchers) who might be able to give indication on the existence of potential harms and the way to solve them.

4.3.2 Procedural Obligations. Procedural obligations were also reported by participants, as wishes to foster algorithmic harm considerations. In terms of requirements or guidelines for the ML system to be built, they reported that, oftentimes, the organisation did not specify any harm-related requirement, and that certain requirements would come in opposition to the mitigation of harms (due to existing impossibility results; limited access to data, e.g., due to cost, etc.) —a clear hindrance towards harm mitigation. For instance, P16 and P19 described that their decision to develop a system is based primarily on the system's usefulness (time and cost saved) for the business that requires it, leaving out questions about harms towards data subjects *P16: "It's appropriate and relevant for the business. They want to save money or to reduce time of work."* Subjective norms (the vision that the society might have on the organisation, or the belief that the organisation has on the way of handling harms of other organisations) also played a role in the establishment of requirements by the organisation. In certain cases, it made the organisation push the practitioners towards investigating harms, while in other cases it refrained them to do so —for instance, P13 mentioned that if the public knew about a certain harm mitigation approach, they would not accept the ML system deployment *P13: "[talking about post-processing methods that flip certain model outputs] They imply a bias in the process. It would be a problem for the company to say that they are doing this: if I am a company and I am saying publicly that I am imputing bias on my model, how would society react to it?"*

Next to inexistent, ambiguous, or contradictory requirements, the allocation of responsibilities towards harms was described as structurally unclear for the practitioners. Very few practitioners mentioned clear allocation of responsibilities by their organisation (e.g., existence of an ethics committee). This represented one more challenge for the practitioners, as that did not necessarily provide them with the needed power to make choices towards harm mitigation. Particularly, participants often discussed that they can strive to make harms transparent within their projects, but that the model requesters have the final say in deployment decisions.

4.4 Interactions between Factors

Here, we provide a short description of the main interactions we identified between factors, that reveal the importance of psychological traits and other human factors, and reinforce the need to account for the entangled nature of these factors.

4.4.1 Perceived or Actual Responsibility. We described that organizational factors might leave responsibility around harms ambiguous. In such situation, different practitioners react differently (hinting again at the importance of human factors): they perceive their responsibility differently, and engage to different extents in activities that are not promoted by the organizations in order to tackle harms. Certain practitioners argued that as data scientists that know the most about the system, they are the ones responsible for identifying and reporting harms (if not also for making decisions on system requirements and deployment) *P17: "It needs to be the responsibility of the developer, or have a developer that is some sort of fairness compliance person, that's doing some peer reviews of code, because once you get to the developers' boss, they don't know code."*; that the model requesters are the ones deciding for any requirement; that the C-level and managers should be responsible to incentivise the engineers and to make choices where practitioners do not have knowledge *P19: "As much as I would probably want to, I don't think I have all the necessary background for that."*; or that a committee within the organization should be responsible as it would gather more diverse expertise *P16: "We have a committee of ethics. If we have any questions, we can go there to understand their opinion, it will not be the decision of one person but a collective decision."*

4.4.2 Obstacles and Efforts. We mentioned that practitioners might lack resources (e.g., access to relevant stakeholders) and knowledge to tackle harms. In such cases, we identify different attitudes towards the challenge. While it is well-known that collaboration in the ML lifecycle is often needed for the practitioners [24, 51, 80, 110], prior work and our study both show that tackling questions around algorithmic harms is still predominantly the job of ML practitioners alone. Except for certain highly-ML experienced practitioners, most of them did not mention putting proactive extensive effort into reaching out to relevant stakeholders. In terms of knowledge, many of the participants who admitted lacking knowledge to identify or mitigate harms, concluded by reporting that they consequently do not put effort into acting on harms. *P10: "I am slightly aware of it but I wouldn't be able to say how to make changes towards that. I don't have any experience."* Instead, others mentioned searching into research papers to identify appropriate methods. For instance, P15, P18, P24, P27 proposed to look into research that trades-off model size (assuming a smaller model would be less energy-consuming) and accuracy performance to reduce environmental impact. Some practitioners explained potentially having a higher propensity to put effort onto fairness challenges because they have research experience, and hence can search within publications for relevant methods *P7: "I'm interested in research. When you try to apply these tools, that is connecting the academic world to the business side."* Similarly, when participants mentioned that no method exists yet to tackle a harm, certain would attempt to create a new one, while others would wait for research to progress.

5 DISCUSSION & IMPLICATIONS

5.1 The Renewed Importance of Factors

5.1.1 Summary of our Findings. In our study, we found that a complex set of interdependent human and organisational factors interact, and result in diverse practices of machine learning (ML) practitioners around algorithmic harms. For instance, we identified that, overall, practitioners who have little experience with ML and have not received practical and critical training around algorithmic fairness often stop at the application of a few fairness metrics and mitigation methods. The more experienced practitioners and those with an interdisciplinary background present a more critical attitude, attempt to go beyond what fairness toolkits permit (e.g., by envisioning non-algorithmic ways to avoid algorithmic unfairness), especially when they had opportunities to discuss these topics with experts. Next to these prior experiences, organizational constraints and incentives also represent drivers or obstacles towards deeply tackling harms, that, in interaction with psychological and sociodemographic traits, result in a diversity of trade-offs made between algorithmic harms and other business considerations.

While it is natural that such types of factors impact practices in the context of ML model development and algorithmic harms, no investigation of such factors had been performed. This study provides a first qualitative investigation that bear broad implications, and whose output validity should be later investigated through quantitative studies. As toolkits cannot serve as straightforward recipes for the practitioners, practitioners should also be supported in exercising due diligence. We argue that this should go through the development of better means for knowledge dissemination and training, the design of supportive materials and new organizational processes, and the consideration of organizational factors.

5.1.2 A Lukewarm Perspective on Toolkits. Our results bring evidence confirming the results of prior works on the use of various documentation and code toolkits, that have shown that these toolkits can indeed support ML practitioners in finding more algorithmic harms than without a toolkit [16, 24]. Yet, our results also bring more nuance to the benefits of toolkits, and show the risks of using those. These nuances had not been demonstrated in prior, empirical works on toolkit practices, as they did not focus on the impact of toolkits on algorithmic harms, but only on the correct implementation of algorithmic fairness methods. Our results also provide empirical evidence for prior broader works that argued against the techno-solutionism of algorithmic fairness [34], demonstrated the potential dangers of ethics washing [8], and more broadly warned against automating ML processes, e.g., through AutoML [106].

Prior work [24] had not discussed major differences in usage of different fairness toolkits. We corroborate such findings. Besides, the factors we find practitioners mentioning as important for selecting a toolkit are well aligned with the insights of prior works on the use of these toolkits [24, 57, 85]. These works have developed, among others, rubrics for the design of better toolkits, including similar functionalities (compatibility with various models, inclusion of diverse fairness metrics, guidance along the entire ML lifecycle, facilitating interdisciplinary conversations, etc.) and non-functional requirements (e.g., learning curve, compatibility with common coding frameworks, etc.). We especially echo the recommendations

they make to better guide practitioners along socio-technical considerations [104], in order to avoid the pitfalls emphasized by our participants. These prior works however had not discussed the contradictory evaluation of toolkits by practitioners, that we found in our interviews, and that would merit further investigation.

5.1.3 The Importance of Human Factors. Although prior works have sparsely investigated human factors that impact attitudes towards algorithmic fairness, we find a number of prior results that align with ours, and hint at the validity of our results. While these studies do not investigate ML practitioners specifically (but computer science students, or decision subjects), they are still relatable, as perceptions of fairness impact follow-up practices towards harms. Besides, our work expands on these prior results in that it looks at a broader range of harms, and at different types of individuals.

- *Toolkit.* A few works [24, 57] show the potential usefulness of toolkits and their current practical limitations. No study mentions potential negative impact that we identified.
- *Experience.* Kleanthous et al. [50] identified the impact that the level of computer science education has in understanding fairness issues along an ML pipeline, that we also identified. Yet, no study reveals the importance of the type of educational background and the type of prior ML experience and fairness training.
- *Socio-demographic factors.* Quantitative studies [47, 79] have shown the impact of gender on students' considerations of ML fairness, privacy, and non-maleficence. Prior work has also shown the effect of gender and race on judgements of fairness metrics [37, 41]. While this is not a result we could explore due to the imbalanced distribution of participants we had, all our female participants also displayed a critical attitude towards their practices and acknowledged various harms, whereas the results were more disparate across male participants.
- *Non-volitional factors.* Others [38, 66] found that non-volitional factors, e.g., political views and experiences with identity-based vulnerability, are relevant. Our results also hinted at the importance of non-volitional factors, as multiple practitioners referred to their personal interest in the topic, or being part of discriminated minorities, as motivating factors.

While the studies above align with our work, other studies seem contradicting. Some studies have not found impact of socio-demographic or other human factors on the perception of different fairness metrics [22, 37, 91], and the results of other studies are contradicting each other in terms of fairness perceptions, as detailed in [41]. For example, Wang et al. [101] identified that people with higher computer literacy perceive algorithmic decision-making fairer than what people with lower levels of literacy perceive, and that age, gender, race, and education level do not have a significant impact. Contrary to these findings, others [47, 79] pointed to the impact of gender, and our work showed the variability in perceptions of fairness among all our participants who were highly computer literate. We argue that these contradictions are due to the absence of detailed investigation of the impact of the human factors we identified, or to the lack of relevant intersectional considerations across factors.

5.1.4 Contextual Factors: Obstacles or Vectors. Our study identified various clashing constraints and objectives that practitioners have to take into account during the ML lifecycle. Some of these points have already been highlighted in previous empirical works, such as the conflict between business goals (e.g., the system should work for a majority of cases but not necessarily for edge cases to have a competitive advantage) and practitioners' goals (making sure to have high accuracy on all kinds of population) [61, 75, 78], or the lack of organisational support [84] (time and cost allocated, development of tools and guidelines, etc.), that result in individual efforts instead of organizational processes. Other factors had not been discussed until now to the best of our knowledge, in the context of practices for handling algorithmic harms.

5.2 Reflexivity via Renewed Experiences

Facing the importance of various factors, one should take those into account in the future development of support structures for ML practitioners to tackle algorithmic harms. Support should be personalised to the relevant types of practitioners we identified.

5.2.1 Guidelines for the Design of Toolkits. While fairness toolkits mildly contribute to enacting reflexive practices around algorithmic harms, they still represent an almost inevitable medium for algorithmic fairness. They appear as double-edge swords according to our results. This is where the danger of breeding a "*Checkbox Culture*" can manifest among practitioners with respect to handling algorithmic harms. Our work especially shows the need for pointers to relevant activities and resources within toolkits [56], while emphasizing the complexity of the problem and its context-dependence. Toolkits should also be adapted to the type of stakeholders that use them, based on their prior training, experiences, and other human factors, showing pop-up warnings, enforcing attention checks towards harms, allowing for different functionalities, or proposing trainings before using the toolkits. This will be a challenge as existing warnings in FairLearn [10] do not seem to always be considered by the practitioners. Besides, we need to make sure the toolkits do not become new checkboxes, but instead foster critical thinking.

5.2.2 Due Diligence through Education.

Topical Education. Since our results highlighted the importance of the type of training and experience practitioners have received about ML and harms, we join prior studies in advocating for more education of ML practitioners [24, 51, 89]. Many works [12, 14, 19, 29, 31, 44, 49, 65, 83] have discussed ways to provide a responsible AI education to developers, and we recommend to refer to their insights (e.g., modular approaches to responsible AI education for easy integration into courses, including events reported in news articles). We also recommend to rely on insights from farther domains such as data science teaching [32, 54, 92] (perhaps even more worrying than our results, low-ML-experienced practitioners also failed into well-known, non-harm-related traps, such as not reflecting on the limitation of accuracy as a performance metric), ethics and HCI [20, 25, 30], or even ethics of long-established fields such as medicine [21], which have tackled tangential questions. We emphasize the importance of accounting for the breadth of the topic (only Garrett et al. [31] noticed the absence of certain harms like environmental impact from existing courses), its complexity, and the importance to raise awareness about the issues and to train on tackling them.

Change of Attitudes. Next to teaching about algorithmic harms, it is important to develop the moral sensitivity [14], the critical attitude, and the reflexivity of future practitioners [68][4], in this highly-subjective context (Green and Viljoen [36] talk about an algorithmic realism approach, acknowledging the contextual, porous, and political nature of these harms and objectives) where no easy solution to algorithmic harm can be prescribed. Three concrete mediums of good practices surfaced from our interviews: discussions with diverse stakeholders to develop awareness around the subjectivity of the problem, warnings to develop a critical attitude towards existing theories and tools, and use-cases to experience potential challenges in the responsible use of tools. These should be incorporated in the trainings. We envision that trainings using close-to-real-world use-cases, starting from the beginning of the ML lifecycle (problem formulation) to the end (deployment and monitoring), with various stakeholders to interact with, and varying degrees of challenges (e.g., having all harm-related and other constraints explicit or proactively identifying them), could be beneficial. Markus and al. [64] insist on accounting for organisational dynamics in such trainings.

Terminological Considerations in Education Material. The terminological confusions we identified align with prior works [72] that highlight disciplinary confusions in the task of making a model fair, and works that studied the impact of terminological choices [53] on one's perceptions of an ML system. Mulligan et al. [72] promote the value of shared vocabularies and reconciling taxonomies that facilitate discussions. We echo these recommendations and the ones of P29 who suggested to move away from loaded terms towards more specific words, e.g., characterizing the type of bias in relation to the harm it creates, arguing that these materials should not only contain definitions such as it is currently done [59], but should also make concepts clear to the extent of pointing out to the different related theories behind them.

5.2.3 Acknowledging Contextual Factors. While these factors are often unspoken in the research community, they have to be accounted for by practitioners, as they are inherently in tension with handling algorithmic harms, but most practitioners currently face the dilemmas alone. We argue that the research community and policy makers should account for these factors further, and support —sometimes empower— practitioners in the decisions they have to make along the ML pipeline. Interdisciplinary research is needed to understand how to prioritize tackling the different harms (beyond distributive fairness), accounting for realistic trade-offs that have to be made across stakeholders and acknowledging practical constraints. Relevant directions are the understanding of preferences of stakeholders beyond well-studied preferences across fairness metrics [37, 41], the development of frameworks to uncover and negotiate preferences between stakeholders [18, 55, 96], and the creation of guidance for practitioners to navigate the trade-offs.

Knowledge and due diligence are not enough when practitioners do not receive structural incentives. P18 mentioned *"Practice is different from the ethical goals of the world. I had an interview. I said it's important to recommend people music that is worthwhile listening to. The manager told me these are idealistic thoughts, not how the real world operates, this company is all about revenue. So fairness at a company level, it depends on the culture and ethics of the people."* Hence, we join [84] in the idea of developing organizational processes to foster the development of good practices: the design of guidelines [63], e.g., for identifying responsibilities and appropriate requirements, the facilitation of interdisciplinary collaborations [83, 104], and the establishment of structural incentives and principles such as slowness [76]. Development of regulations, that explicitly account for organisational obstacles (e.g., making sure some employees of an organization are well-equipped to investigate algorithmic harms, have time dedicated for it) could also incentivise these organizations [33, 90, 98].

[4]Miceli et al. [68] refer to Bourdieu's notion of reflexivity [15] that would apply to ML practices "an analytical tool to sensitize researchers to "the social and intellectual unconscious" that condition their thoughts and practices in research, and is, therefore, an integral part of and a "necessary prerequisite" for scientific inquiry".

5.3 Rigorously Investigating the Factors

The factors we identified should be quantitatively explored in the future to validate our results (identified conceptions for each harm could serve as dependent variables). This would inform the design of trainings and supportive tools (e.g., the categories of individuals to tailor them to), and the constitution of ML development teams, accounting for the perceptions and abilities of each member. We foresee challenges in the design of a rigorous experimental setup: difficulties to quantify human factors, need to account for interactions between them, and need for specific scales around each harm, their different perceptions, and mitigation approaches. Apparent contradictions among results of prior works seem to be due to subtle differences in what is measured, who is the experiment subject, and potential interactions between multiple factors, which are differences that one should aim at controlling in future studies.

Existing research could be used to overcome these challenges. A measurement has been developed to quantitatively measure undergraduate student's attitudes towards the ethics of AI [47], that could be useful to evaluate how these factors are impactful. Yet, one should first complete this instrument to account for the types of harms that are currently left out from the instrument and for which we identified a variability of conceptions, and not only for attitudes towards harms but also towards their mitigation. The insights and methods from social psychology studies about human processes of taking actions, such as the theory of reasoned action or the theory of planned behavior [1, 40], could also be adapted to further analyse results, as they hint at a diversity of factors and their co-existence, for action taking. We already see correspondences, for instance in the subjective norms and perceived control mentioned by these theories, and that our interviewed practitioners also discussed, e.g., when mentioning the image ML ethics give to an organization.

6 LIMITATIONS

While we strived for recruiting a diversity of participants in terms of demographics, experience with ML and fairness, we could not obtain a significant sample for combined categories. Impossibility came from the relatively small amount of practitioners tackling these issues in the world (e.g., few practitioners could be found working regularly with the AIF360 toolkit), the duration of our interviews, and the controversial character of the topic. Yet, since several of our observations are corroborated with previous studies, one

can suppose some generalisability of our results. This also indicates future challenges in quantitatively investigating the factors.

Due to time considerations, practitioners could not extensively explore the toolkits beyond our tutorial. Letting them familiarize themselves further with algorithmic fairness before conducting task T2, would possibly provide a few different results on the impact of experience and toolkits on practices as practices evolve long-term. For instance, FairLearn provides warnings about algorithmic harms that the participants did not see during the interviews, but that could change their attitudes. Yet, the interviews with practitioners experienced with toolkits allowed us to somewhat control for this, and did not show related differences.

Finally, our participants were not placed into a specific organization and did not have access to different stakeholders. While this was useful for us to fairly compare practices across participants, we foresee the importance of further studies, e.g., with the practitioners' own projects, to identify additional factors.

7 CONCLUSION

Our study led to an extended characterization of the complex, intertwined, factors (toolkits, human, and organizational) impacting the differences of conceptions and practices about algorithmic harms that surface across ML practitioners. These results do not only align with prior works that surfaced a few factors in relation to algorithmic fairness, but also extend and complement these works with information around a more comprehensive consideration of algorithmic harms. Particularly, we found that the use of fairness toolkits does not necessarily lead to its envisioned impact, and can at times promote a checkbox culture, if it is not accompanied by a distinction of the background and prior training the user of the toolkit received, as well as of the pressures their organisations puts on them. In summary, our study constitutes a strong testimony that ML practitioners are not as much "ethical unicorns" [83] (i.e., practitioners who ensure a comprehensive handling of algorithmic harms of the ML systems they work on), than *subjective unicorns encaged in an organization*. Such findings bear strong implications for future research opportunities around the refinement of the toolkits and of educational programs, accounting for these human factors, and for potential regulations to address organizational concerns.

ACKNOWLEDGMENTS

This work was partially supported by the HyperEdge Sensing project funded by Cognizant. We would like to thank all the participants of our studies, without whom this work would not have been possible. Besides, we would like to thank Pablo Biedma Nũnez, Eva Noritsyna, Harshitaa Pandey, and Ana-Maria Vasilcoiu, who participated in interviewing participants.

REFERENCES

[1] Icek Ajzen. 1991. The theory of planned behavior. *Organizational behavior and human decision processes* 50, 2 (1991), 179–211.
[2] Lora Aroyo, Matthew Lease, Praveen Paritosh, and Mike Schaekermann. 2022. Data excellence for AI: why should you care? *Interactions* 29, 2 (2022), 66–69.
[3] Agathe Balayn and Seda Gürses. 2021. Beyond Debiasing: Regulating AI and its inequalities. *EDRi Report. https://edri. org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online. pdf* (2021).
[4] Agathe Balayn, Christoph Lofi, and Geert-Jan Houben. 2021. Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal* 30, 5 (2021), 739–768.
[5] 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.
[6] 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.
[7] 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 *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, Madeleine Clare Elish, William Isaac, and Richard S. Zemel (Eds.). ACM, 610–623. https://doi.org/10.1145/3442188.3445922
[8] Elettra Bietti. 2020. From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 210–219.
[9] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 514–524.
[10] 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. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).
[11] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*. IEEE, 1536–1546. https://doi.org/10.1109/WACV48630.2021.00158
[12] Veronika Bogina, Alan Hartman, Tsvi Kuflik, and Avital Shulner-Tal. 2021. Educating Software and AI Stakeholders About Algorithmic Fairness, Accountability, Transparency and Ethics. *International Journal of Artificial Intelligence in Education* (2021), 1–26.
[13] Veronika Bogina, Alan Hartman, Tsvi Kuflik, and Avital Shulner-Tal. 2022. Educating software and AI stakeholders about algorithmic fairness, accountability, transparency and ethics. *International Journal of Artificial Intelligence in Education* 32, 3 (2022), 808–833.
[14] Jason Borenstein and Ayanna Howard. 2021. Emerging challenges in AI and the need for AI ethics education. *AI and Ethics* 1, 1 (2021), 61–65.
[15] Pierre Bourdieu and Loïc JD Wacquant. 1992. *An invitation to reflexive sociology*. University of Chicago press.
[16] Karen L 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, CSCW2 (2021), 1–27.
[17] Benedetta Brevini. 2020. Black boxes, not green: Mythologizing artificial intelligence and omitting the environment. *Big Data & Society* 7, 2 (2020), 2053951720935141. https://doi.org/10.1177/2053951720935141 arXiv:https://doi.org/10.1177/2053951720935141
[18] Jenna Burrell, Zoe Kahn, Anne Jonas, and Daniel Griffin. 2019. When users control the algorithms: values expressed in practices on twitter. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–20.
[19] Emanuelle Burton, Judy Goldsmith, and Nicholas Mattei. 2015. Teaching AI Ethics Using Science Fiction.. In *Aaai workshop: Ai and ethics*. Citeseer.
[20] Emanuelle Burton, Judy Goldsmith, Nicholas Mattei, Cory Siler, and Sara-Jo Swiatek. 2023. *Computing and Technology Ethics: Engaging through Science Fiction*. MIT Press.
[21] Alastair V Campbell, Jacqueline Chin, and Teck-Chuan Voo. 2007. How can we know that ethics education produces ethical doctors? *Medical teacher* 29, 5 (2007), 431–436.
[22] Bo Cowgill, Fabrizio Dell'Acqua, Samuel Deng, Daniel Hsu, Nakul Verma, and Augustin Chaintreau. 2020. Biased programmers? Or biased data? A field experiment in operationalizing AI ethics. In *Proceedings of the 21st ACM Conference on Economics and Computation*. 679–681.
[23] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 525–534.
[24] 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. *FAccT* (2022).
[25] Eva Eriksson, Elisabet M Nilsson, Anne-Marie Hansen, and Tilde Bekker. 2022. Teaching for Values in Human–Computer Interaction. *Frontiers in Computer Science* 4 (2022).
[26] Sina Fazelpour and Zachary C Lipton. 2020. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 57–63.

[27] Tobias Fiebig, Seda F. Gürses, Carlos Hernandez Gañán, Erna Kotkamp, Fernando Kuipers, Martina Lindorfer, Menghua Prisse, and Taritha Sari. 2021. Heads in the Clouds: Measuring the Implications of Universities Migrating to Public Clouds. *CoRR* abs/2104.09462 (2021). arXiv:2104.09462 https://arxiv.org/abs/2104.09462

[28] 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.

[29] Heidi Furey and Fred Martin. 2019. AI education matters: a modular approach to AI ethics education. *AI Matters* 4, 4 (2019), 13–15.

[30] Ajit G. Pillai, A Baki Kocaballi, Tuck Wah Leong, Rafael A. Calvo, Nassim Parvin, Katie Shilton, Jenny Waycott, Casey Fiesler, John C. Havens, and Naseem Ahmadpour. 2021. Co-designing resources for ethics education in HCI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–5.

[31] Natalie Garrett, Nathan Beard, and Casey Fiesler. 2020. More Than" If Time Allows" The Role of Ethics in AI Education. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 272–278.

[32] Yolanda Gil. 2016. Teaching big data analytics skills with intelligent workflow systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[33] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* 38, 3 (2017), 50–57.

[34] Ben Green. 2021. The contestation of tech ethics: A sociotechnical approach to technology ethics in practice. *Journal of Social Computing* 2, 3 (2021), 209–225.

[35] Ben Green. 2021. Escaping the" Impossibility of Fairness": From Formal to Substantive Algorithmic Fairness. *arXiv preprint arXiv:2107.04642* (2021).

[36] Ben Green and Salomé Viljoen. 2020. Algorithmic realism: expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 19–31.

[37] Nina Grgić-Hlača, Gabriel Lima, Adrian Weller, and Elissa M Redmiles. 2022. Dimensions of Diversity in Human Perceptions of Algorithmic Fairness. In *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, 1–12.

[38] Nina Grgić-Hlača, Adrian Weller, and Elissa M Redmiles. 2020. Dimensions of diversity in human perceptions of algorithmic fairness. *arXiv preprint arXiv:2005.00808* (2020).

[39] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 51–60. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16523

[40] Jerold L Hale, Brian J Householder, and Kathryn L Greene. 2002. The theory of reasoned action. *The persuasion handbook: Developments in theory and practice* 14, 2002 (2002), 259–286.

[41] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 392–402.

[42] MEPS HC. 2017. 181: 2015 Full Year Consolidated Data File. *Agency for Healthcare Research and Quality* (2017).

[43] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22, 7 (2019), 900–915.

[44] Anna Lauren Hoffmann and Katherine Alejandra Cross. 2021. Teaching data ethics: Foundations and possibilities from engineering and computer science ethics education. (2021).

[45] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, and Hanna M. 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, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos (Eds.). ACM, 600. https://doi.org/10.1145/3290605.3300830

[46] Lotte Houwing. 2020. Stop the Creep of Biometric Surveillance Technology. *Eur. Data Prot. L. Rev.* 6 (2020), 174.

[47] Yeonju Jang, Seongyune Choi, and Hyeoncheol Kim. 2022. Development and validation of an instrument to measure undergraduate students' attitudes toward the ethics of artificial intelligence (AT-EAI) and analysis of its difference by gender and experience of AI education. *Education and Information Technologies* (2022), 1–33.

[48] Os Keyes, Jevan A. Hutson, and Meredith Durbin. 2019. A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, Regan L. Mandryk, Stephen A. Brewster, Mark Hancock, Geraldine Fitzpatrick, Anna L. Cox, Vassilis Kostakos, and Mark Perry (Eds.). ACM. https://doi.org/10.1145/3290607.3310433

[49] Sountongnoma Martial Anicet Kiemde and Ahmed Dooguy Kora. 2021. Towards an ethics of AI in Africa: rule of education. *AI and Ethics* (2021), 1–6.

[50] Styliani Kleanthous, Maria Kasinidou, Pınar Barlas, and Jahna Otterbacher. 2022. Perception of fairness in algorithmic decisions: Future developers' perspective. *Patterns* 3, 1 (2022), 100380.

[51] Sean Kross and Philip Guo. 2021. Orienting, framing, bridging, magic, and counseling: How data scientists navigate the outer loop of client collaborations in industry and academia. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.

[52] Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2020. POTs: protective optimization technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 177–188.

[53] Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J König, and Nina Grgić-Hlača. 2022. "Look! it'sa computer program! it's an algorithm! it's ai!": does terminology affect human perceptions and evaluations of algorithmic decision-making systems?. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–28.

[54] Niklas Lavesson. 2010. Learning machine learning: a case study. *IEEE Transactions on Education* 53, 4 (2010), 672–676.

[55] 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.

[56] 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.

[57] 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.

[58] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*. PMLR, 3150–3158.

[59] Estevez Almenzar M, Fernandez Llorca D, Gomez Gutierrez E, and Martinez Plumed F. 2022. *Glossary of human-centric artificial intelligence.* Scientific analysis or review, Technical guidance KJ-NA-31113-EN-N (online). Luxembourg (Luxembourg). https://doi.org/10.2760/860665(online)

[60] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 52 (apr 2022), 26 pages. https://doi.org/10.1145/3512899

[61] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–26.

[62] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376445

[63] 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.

[64] M Lynne Markus, Marco Marabelli, and Christina Zhu. 2019. POETs and quants: Ethics education for data scientists and managers. *Marco and Zhu, Xiaolin (Christina), POETs and Quants: Ethics Education for Data Scientists and Managers (November 19, 2019)* (2019).

[65] Afra Mashhadi, Annuska Zolyomi, and Jay Quedado. 2022. A Case Study of Integrating Fairness Visualization Tools in Machine Learning Education. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[66] Nora McDonald and Shimei Pan. 2020. Intersectional AI: A Study of How Information Science Students Think about Ethics and Their Impact. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–19.

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

[68] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting computer vision datasets: an invitation to reflexive data practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 161–172.

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

[70] Petra Molnar. 2021. Technological Testing Grounds and Surveillance Sandboxes: Migration and Border Technology at the Frontiers. *Fletcher F. World Aff.* 45 (2021), 109.

[71] Jessica Morley, Libby Kinsey, Anat Elhalal, Francesca Garcia, Marta Ziosi, and Luciano Floridi. 2021. Operationalising AI ethics: barriers, enablers and next steps. *AI & SOCIETY* (2021), 1–13.

[72] 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.

[73] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.

[74] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 2022. Designing Fair AI in Human Resource Management: Understanding Tensions Surrounding Algorithmic Evaluation and Envisioning Stakeholder-Centered Solutions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 51, 22 pages.

[75] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 2022. Designing Fair AI in Human Resource Management: Understanding Tensions Surrounding Algorithmic Evaluation and Envisioning Stakeholder-Centered Solutions. In *CHI Conference on Human Factors in Computing Systems*. 1–22.

[76] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–15.

[77] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3287560.3287567

[78] Samir Passi and Phoebe Sengers. 2020. Making data science systems work. *Big Data & Society* 7, 2 (2020), 2053951720939605. https://doi.org/10.1177/2053951720939605 arXiv:https://doi.org/10.1177/2053951720939605

[79] Emma Pierson. 2017. Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124* (2017).

[80] David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. 2021. How ai developers overcome communication challenges in a multidisciplinary team: A case study. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.

[81] Manish Raghavan, Solon Barocas, Jon M. Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna (Eds.). ACM, 469–481. https://doi.org/10.1145/3351095.3372828

[82] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*, Annette N. Markham, Julia Powles, Toby Walsh, and Anne L. Washington (Eds.). ACM, 145–151. https://doi.org/10.1145/3375627.3375820

[83] Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. 2021. You can't sit with us: Exclusionary pedagogy in ai ethics education. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 515–525.

[84] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where Responsible AI Meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 7 (apr 2021), 23 pages. https://doi.org/10.1145/3449081

[85] 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.

[86] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Extended Abstracts Volume, Atlanta, Georgia, USA, April 10-15, 2010*, Elizabeth D. Mynatt, Don Schoner, Geraldine Fitzpatrick, Scott E. Hudson, W. Keith Edwards, and Tom Rodden (Eds.). ACM, 2863–2872. https://doi.org/10.1145/1753846.1753873

[87] 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).

[88] Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. 2020. What does it mean to'solve'the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 458–468.

[89] Conrad Sanderson, David Douglas, Qinghua Lu, Emma Schleiger, Jon Whittle, Justine Lacey, Glenn Newnham, Stefan Hajkowicz, Cathy Robinson, and David Hansen. 2021. AI ethics principles in practice: Perspectives of designers and developers. *arXiv preprint arXiv:2112.07467* (2021).

[90] Nathalie A Smuha. 2019. The EU approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International* 20, 4 (2019), 97–106.

[91] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2459–2468.

[92] Thilo Stadelmann, Julian Keuzenkamp, Helmut Grabner, and Christoph Würsch. 2021. The AI-atlas: didactics for teaching AI and machine learning on-site, online, and hybrid. *Education Sciences* 11, 7 (2021), 318.

[93] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. 2014. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international* 2014 (2014).

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

[95] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, danah boyd and Jamie H. Morgenstern (Eds.). ACM, 10–19.

[96] Niels Van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M Kelly, and Vassilis Kostakos. 2019. Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[97] Sriram Vasudevan and Krishnaram Kenthapadi. 2020. Lift: A scalable framework for measuring fairness in ml applications. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2773–2780.

[98] Michael Veale and Frederik Zuiderveen Borgesius. 2021. Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International* 22, 4 (2021), 97–112.

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

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

[101] 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.

[102] Hilde Weerts, Lambèr Royakkers, and Mykola Pechenizkiy. 2022. Does the End Justify the Means? On the Moral Justification of Fairness-Aware Machine Learning. *arXiv preprint arXiv:2202.08536* (2022).

[103] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, Madeleine Clare Elish, William Isaac, and Richard S. Zemel (Eds.). ACM, 666–677. https://doi.org/10.1145/3442188.3445928

[104] Richmond Y Wong, Michael A Madaio, and Nick Merrill. 2022. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. *arXiv preprint arXiv:2202.08792* (2022).

[105] Huichuan Xia, Yang Wang, Yun Huang, and Anuj Shah. 2017. "Our Privacy Needs to be Protected at All Costs": Crowd Workers' Privacy Experiences on Amazon Mechanical Turk. *Proc. ACM Hum. Comput. Interact.* 1, CSCW (2017), 113:1–113:22. https://doi.org/10.1145/3134748

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

[107] Catherina Xu, Christina Greer, Manasi N Joshi, and Tulsee Doshi. 2020. Fairness Indicators Demo: Scalable Infrastructure for Fair ML Systems. (2020).

[108] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna (Eds.). ACM, 547–558. https://doi.org/10.1145/3351095.3375709

[109] Ming Yin, Siddharth Suri, and Mary L. Gray. 2018. Running Out of Time: The Impact and Value of Flexibility in On-Demand Crowdwork. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox (Eds.). ACM, 430. https://doi.org/10.1145/3173574.3174004

[110] Amy X Zhang, Michael Muller, and Dakuo Wang. 2020. How do data science workers collaborate? roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

[111] Kathryn Zyskowski, Meredith Ringel Morris, Jeffrey P. Bigham, Mary L. Gray, and Shaun K. Kane. 2015. Accessible Crowdwork?: Understanding the Value in and Challenge of Microtask Employment for People with Disabilities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015*, Dan Cosley, Andrea Forte, Luigina Ciolfi, and David McDonald (Eds.). ACM, 1682–1693. https://doi.org/10.1145/2675133.2675158