Harmful Impacts of ML: Empirically Triangulating the Concerns and Practices of Developers

AGATHE BALAYN, Microsoft Research, USA
UJWAL GADIRAJU, Delft University of Technology, The Netherlands

Machine learning (ML) models used in decision-making tasks are known to bear harmful impacts. To tackle such impact, researchers have focused on developing tools to mitigate algorithmic fairness issues and to support ML developers in their algorithmic fairness-centered practices. Yet, little has been triangulated about the concerns and practices of ML developers towards the broader impact of ML that arises from complex questions of distributive unfairness and unsustainable pillars underlying ML models (e.g., opaque task formulation, inappropriate datasets, energy-intensive infrastructures). In this qualitative study, we conducted 30 semi-structured interviews using a convenience sampling of developers with varying educational backgrounds and varying experience with ML and algorithmic fairness. We surface (mis)conceptions and (questionable) practices around harms and their mitigation. Our study reveals no standard across developers' concerns and practices, and tensions developers face when attempting to curb the undesirable impacts of ML models. These insights triangulate prior results on algorithmic fairness and shed light on various unsolved theoretical, design, methodological, and governance challenges. Our findings constitute a vital step forward to support developers and our broader community in navigating this growing, increasingly ubiquitous, footprint of ML.

Keywords: empirical study, practitioners' concerns, practices, responsible AI

Reference Format:
Agathe Balayn and Ujwal Gadiraju. 2025. Harmful Impacts of ML: Empirically Triangulating the Concerns and Practices of Developers. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 37 pages.

1 Introduction

The potential harmful impact that developing machine learning (ML) models and using them to conduct decision-making tasks can have is now well-established. Substantial scholarship has conceptually examined the harm that ML can cause, be it system-level questions of problematic design decisions [12, 98] or questionable usages of the ML models [57], or algorithm-level questions, e.g., of inappropriate datasets [43], or of unfair model outputs [75]. To address such harmful impact, the HCI community [23, 50] has established the necessity to support ML developers, i.e., those who participate in the design of datasets or ML algorithms, in their ML system development work. They are often the first stakeholders who can act on ML harm through the various design choices they make.

Authors' Contact Information: Agathe Balayn, Microsoft Research, New York, USA, balaynagathe@microsoft.com; Ujwal Gadiraju, Delft University of Technology, Delft, The Netherlands, u.k.gadiraju@tudelft.nl.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

Supporting ML developers in handling the harmful impact of ML requires first understanding their perceptions of ML harm and the challenges they face in tackling them. Prior works have employed various approaches, including direct inquiries prompting ML developers to articulate their challenges [50, 115], and investigations of their practices [69], such as their use of fairness toolkits to mitigate unfair model outputs [23, 95]. However, three notable research gaps remain. From an epistemological standpoint, these works predominantly focus on challenges linked to addressing unfair model outputs, neglecting other potential harmful impacts of ML. Methodologically, there is a need to adopt a more holistic lens on ML developers' relation to ML's harmful impact. Gaining insights into ML developers' conceptualizations of the harmful impact of ML in conjunction with their practices, should offer a renewed understanding of ML developers, compared to directly prompting them about ML fairness or their usage of ML fairness toolkits. Note that no work has separately investigated the broad concerns of ML developers either. Finally, we acknowledge the practical challenges inherent in studying ML developers. A relatively small and skewed subset of the ML developer population has been examined until now, with an emphasis on ML developers with some practical experience with mitigating the harmful impact of ML [69] or those compelled to use certain tools without prior experience [23]. Moreover, the practices of this subset of developers might have evolved greatly over the past four years given the rapidly evolving nature of the ML field. Therefore, we argue that triangulation efforts are essential to update and expand upon prior insights. Addressing these research gaps is crucial to supporting ML developers in tackling the harmful impact of ML. Thus, we ask — **How do ML developers conceive and handle the harmful impact of ML?**

To answer this question and address the above methodological considerations, we adopted an approach that complements prior works. Inspired by Deng et al., we conducted a think-aloud study followed by interviews with developers ($N = 30$). We recruited ML developers corresponding to varying demographic and educational backgrounds and varying levels of experience with ML. Differently from prior works, we first tasked developers with investigating an ML problem and observed their concerns and practices around ML's harmful impact without pre-specifying any harm or any tool. Only during the semi-structured interviews, we then questioned them about various potential ML harmful impacts and foreseen challenges.

We found a layered set of concerns ML developers express about their ML systems and a set of activities they perform to tackle these concerns. Across ML developers, we also found fragmented conceptions and prioritization of harms, and fragmented goals and practices towards handling these harms, with potentially limited or flawed considerations among them. These results corroborate prior findings around algorithmic fairness, and provide a novel and extensive understanding of other ML concerns and the connected practices. Where some developers are satisfied technically trading off accuracy with fairness and other factors, others recognize the complexity of the socio-technical issue and acknowledge diverse unsolvable tensions. This calls for various theoretical and empirical investigations and design efforts, to guide developers in their design choices towards building ML models with controlled harmful impact. This also sheds light on deeper questions around the methodologies our community has employed to understand and support practitioners, and on the central stage it has given to ML developers.

2 Background and Related Literature

2.1 Conceptual Understanding of the Harmful Impact of ML Systems

2.1.1 *The Various Types of Harmful Impact*

Conceptual works [5, 15, 75, 105, 119] have investigated the harmful impact that ML can have. At a system level, harm can arise from the use and production of the ML system. Previous research has questioned the *desirability of using an ML model*, its use for undesired applications [51, 57, 75, 76], and how it impacts the current structures in place [33]. For instance, using ML might be questionable in situations where novelty is desirable because ML only allows the reproduction of historical and potentially harmful data patterns [90] with recent developments in generative AI notwithstanding. Researchers have also questioned the *negative externalities caused by the production process* of ML applications, such as the environmental impact of model training [12, 18], the labor conditions of data workers [98, 123, 125, 129], the privacy-infringing data used for training [91], etc.

At an algorithm level, researchers typically emphasize concerns around the training and test datasets, and around the outputs of the ML model. ML requires to use *datasets* whose *schemas and sampling* can be harmful. For instance, certain attributes might be inappropriate [72], e.g., use of non-volitional or privacy-infringing attributes [42, 111], they might neglect the complexity of the concept they ought to represent (e.g., the race attribute [43]), or force populations in non-adapted categories (e.g., binary gender) [99]. The dataset distribution, despite a correct dataset schema, might present biases [75, 78, 121], e.g., excluding certain populations. The social impact of wrong outputs of ML models has also been categorized into various taxonomies depending on the context of the use of the models [8, 29, 56, 105], e.g., representational or allocative harms, stereotyping, demeaning, or reifying social groups, etc. As we focus on decision-making systems for resource allocation, we now delve deeper into distributive unfairness, i.e., unfair outputs of an ML model that can cause allocative harms, among others.

2.1.2 *Zooming-in on Distributive Unfairness*

Increasingly, the research community interested in circumventing the harm of ML has focused on technical issues of distributive unfairness [31]. Researchers have developed diverse algorithmic fairness metrics [116] that aim at measuring distributive unfairness in the outputs of the model or in a dataset, unfairness mitigation methods [6, 37] that ought to improve the model algorithmic unfairness as defined by the metrics, and fairness code toolkits [16] to support ML developers in adopting these metrics and methods. Critical works have shown the conceptual limitations of such efforts: there is a gap between algorithmic unfairness and actual harm caused by ML systems in practice. Algorithmic fairness metrics cannot reflect the contextual factors that influence what is considered distributively unfair: for instance, they wrongly assume that parity is always desired in the system outputs [67], do not account for the impact one same output has on different decision-subjects [75], while also not accounting for indirect impact on non-data subjects [61]. Besides, looking at the process to reach algorithmic fairness (procedural justice), the mitigation methods do not ensure that how the unfair situation is addressed is aligned with moral principles [118] and tackles the structural causes of unfairness might remain [31, 79]: for instance, a model can reach low disparate accuracy by treating all individuals or groups unjustifiably [79], or differently (e.g., post-processing method allocate different decision thresholds for different groups) which consists in direct discrimination [40].

2.2 Practices Against Machine Learning Harmful Impact

2.2.1 Concerns for ML Harmful Impact

To the best of our knowledge, only two works [4, 35] brush upon the perceptions that ML developers have of distributive unfairness. Friedle [35] shows the difficulty ML developers have in pinning down a relevant distributive fairness notion, while Ashkotrab et al. [4] show the impact that various visualisations of algorithmic fairness and accuracy have on the choice ML developers make of models to be deployed. No work has studied ML developers' concerns around any other harm. The closest to ML developers are Widder et al. [120] who investigate the ethical concerns of *general software developers* (military, privacy, advertising, surveillance), and Kleanthous et al. [58] who identify that computer science students express different concerns around the fairness of the outputs of the ML models and the appropriateness of the training dataset (concern also identified once for ML developers in one sentence [23]). The majority of works on harm perceptions focuses instead on fairness perceptions of decision-subjects or the public [44, 45, 63, 100, 107, 112, 117, 126], and sometimes on considerations such as privacy and maleficence [55, 87]. In this work, we leverage conceptual works on harms to investigate ML developers' concerns.

2.2.2 Existing Practices Around ML Harmful Impact

More works investigate practices of ML developers around algorithmic fairness, yet none investigates their practices with regard to other ML harm –what we do here. These works discuss challenges faced by ML developers in assessing and mitigating algorithmic unfairness in their own contexts [50, 69, 102, 115] and usages and limitations of algorithmic fairness toolkits [7, 23, 65, 95], and sometimes describe the first steps to assess fairness [35, 69]. The end goal of our work also consists in identifying ML developers' challenges and support opportunities, and represents both an effort of triangulation in doing so and of complementing existing insights, as we explained in Section 1. Particularly, note that none of these works explicitly present the complete sequence of steps ML developers follow to interpret diverse concerns, assess them, and mitigate them —often, the studies directly prompt ML developers to discuss their perceived challenges. We fill this gap through our work, as knowledge of the developers' workflow can help us identify a more complete set of challenges (by avoiding developers' blind spots) and a wider set of research opportunities for supporting them.

3 Method

Interview Procedure. To identify the nuances through which ML developers conceive and handle ML's harmful impact, we adopted an empirical and qualitative approach via 30 semi-structured interviews. We provided an ML model development task to our participants before asking further questions, to first observe their "raw" concerns and practices, providing us data to identify potential limitations and challenges they consider immediately. After going through the task, we asked three types of questions: *background experience questions* (demographics, experience with ML and algorithmic fairness); *reflection questions* around the harmful impact of the given task and of the ML model they developed, and around general wishes, doubts, and challenges the participants might have about their workflow; and *process questions* to understand the reasoning behind each participant's activities during the tasks, especially about harm-concerns these activities might raise. The interviews lasted around one hour on average.

Participants. We recruited our participants using personal networks, targeted requests on social media, calls for participation on the official Discord or Slack communication channels of fairness toolkits, and snowball sampling.

The participants received no financial compensation, and their contributions were fully voluntary (they were motivated by their desire to discuss and reflect on their ML practices with academic researchers). 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.

30 participants were recruited across research and industry institutions, and application domains such as healthcare, finance, and predictive maintenance. Manual sampling was performed to make sure that all participants had (a) responsibilities in ML model development, deployment, or evaluation; (b) varying levels of prior experience with ML, ranging from 2 to 15 years; and (c) varying experience with algorithmic fairness. The participants differed in terms of demographics (nationality, gender, and age) and educational background (highest qualification). 26 of the participants are from Europe, which enabled us to investigate whether concerns and practices reported in past research (that primarily involved participants from North America) also apply to the european context.

Tasks. We chose one existing ML model development use-case, involving the prediction of *hospital readmissions* within 30 days for individual patients [108]. We pre-processed the dataset to simulate potential harm. We chose the domain of healthcare because the increasing use of ML systems makes it prone to various harms, it requires expertise to be handled correctly, and several corresponding datasets are available. This represents realistic scenarios where ML developers often have to develop models without having extensive expertise in the domain of application [101] —only 4 out of the 30 participants reported having some healthcare knowledge, among which only one had more extensive, practical experience. Moreover, since these are not the most frequent use-cases in the algorithmic fairness literature, we could maximize the potential for each participant to be investigating them for the first time.

We shared a Google Colab notebook with the participants, which included a design brief with the pre-loaded dataset. If they discussed it, we helped them to load a fairness toolkit into the notebook (FairLearn [16] or IBM AIF360 [11]). The design brief mentioned that a hospital wanted to optimize their cost and services, and therefore wanted to investigate whether ML could help them predict readmissions. The institution tasked the participant to investigate this feasibility using the available dataset and to report on their findings by speaking out loud.

Analysis of the Transcripts. We analysed the transcripts with a reflexive thematic analysis approach, using a combination of inductive and deductive coding. The first author identified the segments reporting on the main themes we wished to discuss (e.g., concerns around ML harmful impact, identification, and handling steps), and coded emerging themes (e.g., factors that developers trade-off when developing ML models). Then, they identified the response declinations of each participant for each higher-level theme (e.g., choice of fairness metrics based on expert advice, or applying all of them). Later, in discussion with the other authors, the first author reconciled incoherent or redundant codes, and identified additional transversal themes (e.g., prioritization of harms or requirements). Finally, based on our preliminary analysis of the conceptual literature about harm, we critically reflected on the codes to identify flaws in participants' perceptions or approaches while accounting for the subjectivity of the knowledge built on the topic, where such information was available (note that there is not a single, correct, approach to perceive or tackle harms). This process resulted in 276 codes. Further details about the interview process, participants, materials and questions, and the resulting codes, are included in Appendix B.

4 Results & Discussion

We present the findings of our study following the stages of the ML developers' workflow.

4.1 Shaping Concerns around ML Harm: Disparate Reflections and Conceptions

4.1.1 A Rich Set of Concerns

We identified three conceptual layers of concerns that developers expressed during our tasks. The first layer corresponds to the general themes developers think about to shape their concerns: distributive unfairness, harmful dataset, system desirability, development process. These are the macro-categories of harms described in Section 2.1.1. In the remainder of this paper, we color-code developers' considerations based on these macro-categories, and underline them based on the layer they belong to (from no line to two lines). For instance, P28 referred both to the desirability of the system and to its development process: *P28 "We need to look at the bigger picture to see if our work is ethical. That can go for the carbon footprint, the sustainability, the impact this may have on the labour market, and in warfare."* The meaningfulness and utility of the ML prediction task were also questioned; *P1 "Think whether the problem was formulated in a way that makes sense, for example why is 30 days the cut off? Is there something specific about these dates or was it just chosen out of the data?"*

The second layer corresponds to fine-grained categories of concerns identified per macro-category, that display the richness and diversity of reflections that the developers had about their system. For instance, in terms of problematic data schema, participants discussed the desirability of features, the sensitivity of features, or the meaningfulness of their encoding (e.g., which values are encoded, how they are aggregated, etc.). *P20 "White and non-white... From the start, it's a bad feature. People who are not white are also different between them. This should have been a categorical feature with all the races possible."* About the desirability of an ML system, developers discussed the potentially problematic goal of the system itself, the appropriateness of using ML towards this goal, and the appropriateness of the subsequent ML task formulation toward reaching the goal. A number of their concerns haven't been discussed in-depth in the conceptual literature in the past. For example, several developers questioned which modes of human-ML collaboration the system should be designed for to be considered acceptable, and suggested that although ML can serve to remove human biases, one should remain cautious when using the outputs of an ML system and ensure human oversight; *P27 "It should be a doctor and in addition, this model. I don't think we should just believe the output of the model, but things should be used hand in hand with an expert."* This shared control is often discussed in the context of accuracy [9] but less for ML harmful impact such as distributive fairness. In terms of the development process, the developers reflected upon the labor conditions of the crowd workers they might employ, the environmental impact of training and deploying models, and privacy issues. P6 also discussed their concern for equally sharing resources (e.g., GPU clusters) across an organization; *P6 "This was a university cluster that we shared with others. I didn't want to hog the whole cluster for myself."*

The third layer corresponds to complementary conceptions of the second layer's concerns. Again, not all considerations are discussed within the conceptual literature. For instance, while several works have argued that one should consider the ethicality of the goal an ML system is built for [57, 76], research has not yet discussed how certain practical concerns might relate to harm considerations. P3 argued that one should not employ ML in a system in contexts where the system has to be updated at a fast pace to avoid certain harms: ML-based systems are not flexible enough for urgent updates, as ML developers shy away from modifying them; *P3 "Everybody is afraid* *of changing something "if you change this, it breaks this". So we usually start with: what is the problem that you try to solve? could it be solved by simple query, by business rules, or statistical model? If not, by machine learning? It's not about amplifying the buzz and having AI everywhere. It's about the real value of using it."* We describe developers' concerns exhaustively in Appendix Table 4 and 5.

This layered set of diverse concerns overlaps and extends those discussed in prior conceptual literature. For instance, these concerns reflect well certain traps Selbst et al. [104] conceptualized from a critical analysis of the technical literature on algorithmic fairness (especially the ripple effect, formalism, framing, and solutionism traps). While such types of concerns had not been studied empirically in the past with ML developers, other empirical findings [120] on the ethical concerns of software engineers overlap with ours, in terms of questions of privacy, environment, inequalities, and labor conditions.

4.1.2 A High Diversity of Concerns Across Developers

ML developers showed diversity in the breadth and depth of their concerns. They touched upon different categories and sub-categories of concerns. For example, certain developers did not mention any concern at all before being explicitly prompted about potential harm –arguing that they are not used to such reflections–, while others reflected on a large diversity of harm. Many participants mentioned concerns around privacy infringement in training data, yet, at the deeper level, when prompted for more details, most of them envisioned issues specifically with either consent for data use or with data anonymisation. Similarly, several participants engaged in critical reflections about the appropriateness of the data schema, but they did not all focus on the same aspects, be it the completeness of the set of attributes, the meaningfulness of each attribute, and of their encoding. Overall, we find a low frequency at which distributive fairness is raised as a concern, either due to a lack of awareness of the potentially harmful impact of ML model outputs or a lack of understanding of the sources of unfair model outputs or due to subjectivity and the developers simply not considering outputs as being potentially harmful.

Disagreement manifested in the third layer of concerns, which presents opposing considerations. For instance, in terms of the goals of the system, participants recognized that different stakeholders might have varying goals in mind for the ML system, and showed partiality in putting forward one stakeholder's goals over the others; such as P16 declaring the system desirable as soon as it benefits the organization that deploys it (*P16 "It's appropriate for the business. They want to save money or to reduce time of the workers."*), while P17 insisted on not developing such system, arguing against the morality of the goal towards society (*P17 "That's a big problem. Everybody as they get older, they have more health costs, so that'd be price gauging, the hot button issue of building based on pre-existing conditions. For health insurance, that's unethical."*). In terms of feature sensitivity, developers disagreed on the exceptions making a sensitive feature not harmful, e.g., exception as soon as the feature is related to the target label, or if it is volitional *and* related to it. Even when developers agreed on the sensitive features, they did not envision the same use of these features for the system to not be harmful. Some mentioned that such features should not be used in any case, whereas others proposed exceptions, e.g., when the model does not attribute high-importance weights or when its output does not display disparities across them. The subjectivity also manifested around questions of distributive fairness. Participants mentioned different conceptions of the ideal output distribution, that can be attributed to different moral assumptions and theories in political philosophy [13]. For instance, they referred either to notions of predictive parity or to notions of statistical parity that reflect different cases of equality of opportunity [47].

4.1.3 Potential Flaws in the Concerns

Certain considerations around potential ML harmful impact are questionable per existing research. For example, in terms of algorithmic fairness, certain sensitive features are protected by law in certain contexts and certain output distributions are demanded, yet certain developers were not aware of these questions. For instance, some participants, even when prompted, could not envision any potential harm in the systems' outputs: e.g., P25 envisioned that model features might be problematic, but not model outputs *P25 "In terms of building the model, considering fairness? Didn't we consider all of these things already? we removed all the features, stuff like that. The next step after cleaning everything is model building."* Similarly, research [48, 67] has shown the limitations of considerations of parity in output distributions, that were only envisioned by three developers. Besides, 30% of developers posited that a data distribution representative of the real world will always lead to training a fair, non-harmful, model (and that "debiasing" a dataset is not desirable) as one should not distort the way the world is (WYSIWYG –What You See Is What You Get [36]) –as opposed to another vision of fairness arguing for the importance of accounting for existing historical biases (WAE –We Are All Equal) in data [72], a vision shared by 63% of our participants who expressed the need for changing current data distributions to mitigate algorithmic unfairness. Some participants also explained that in the absence of more research and because of their own lack of knowledge around ML environmental impact, they would consider the issue does not exist or is not severe. *P8 'There are better ways than reducing model training to improve the environment."*

The other questionable considerations revolved around the understanding developers had of potential sources of harm, where limited understanding resulted in participants missing the potential for harm of certain ML design choices. *P2 "I don't think that giving a parameter a certain value can lead to harmful implications. I think it's mostly caused by the data, not really by the model."* Especially, prior work [32, 52, 71, 101, 103, 121] has highlighted a wide spectrum of challenges surrounding some of the data and model activities of the ML lifecycle, that can impact algorithmic unfairness and other data-related harms. In the interviews, developers discussed such activities and others that they perform — e.g., data processing, data cleaning, crowdsourcing-based data labeling. However, most developers did not envision any harm that these activities might cause or reinforce despite discussing algorithmic fairness issues in general (cf. Appendix Table 11, 12). Potential negative implications of more well-known issues such as distribution shifts between deployment and training data, be it in terms of accuracy (more familiar) or algorithmic unfairness [94] did not emerge. Only 3% to 10% of the developers acknowledged potential harms from these activities (e.g., P5 for data outliers, P21 for missing values, and P1, P29, P30 for other preprocessing activities), mentioning skews to the datasets that the activities might cause, which would lead to algorithmic unfairness in the outputs and/or silencing certain populations in the dataset. Note however that certain envisioned connections between the activities of the ML lifecycle, the ML task design, and harms went beyond what is discussed in the literature. For instance, prior work [103] has discussed **processing of data errors** as an activity that can impact algorithmic fairness. Yet, P29 suggested thinking beyond technological causes for algorithmic unfairness, to the meaning for the data subjects and the design of the system beyond the algorithm. *"In Southern California where there's a large Hispanic population, when testing a model to allocate poverty benefits to low-income individuals, they found that Hispanic applicants were rejected at higher rates, just because these applicants aren't fluent in English [mentions **data outliers**]. They have trouble with the application form. So the solution to make this system fair was just to **offer the form in Spanish**, you don't do anything with the model."*

4.2 Setting Concern-Based Goals In Context: Goal Diversity due to Envisioned Tensions

4.2.1 Envisioning Tensions

A recurring theme along the developer's workflow is tensions: developers trade-off various factors while considering potential harms. Some tensions emerge when conceptualising when to consider something harmful (e.g., the opposed desirability of the system for the system provider and for the society). Others are discussed when deciding whether to handle a category of harm (e.g., how severe the harm is compared to system objectives), and how to handle such harm (e.g., mitigating distributive unfairness by collecting more data might be privacy infringing). We identify four types of tensions (see Appendix Table 7), many of which are not accounted for by most ML developers. While most of these tensions had not been discussed in prior empirical works about ML practices and harms, they resonate with the frequent negotiations that data scientists have to conduct in their common workflows [84].

Developers take into account the **requirements concerning the ML model capabilities**. For instance, P7 envisioned a direct trade-off between data and algorithm choices to uphold system requirements, and harms related to the development process; *P7 "We had a company involved in paper recycling. We definitely had to make sure that the amount of data that we are requesting or any other client request wouldn't have any side effect on the environment."* Certain participants do not realize such tension, such as P2 who first chose a type of algorithm to build an ML model focusing on explainability power, and only later considered algorithmic fairness without questioning the initial choice, assuming choice independence between explainability and fairness [10]; *P2 "I first check a lot of different classification models. And check which one has the highest AUC value. Then I choose the model, but if there is a more explainable model that just lacks a bit of accuracy, then I would choose that one."*

Developers also account for **system infrastructure requirements** (e.g., computational power for training), again with or without realizing the impact on potential harms. For instance, P3 and P29 both discussed that different model sizes might be adapted to working with different computational infrastructures because of the computation power they require and that working with these different models also entails more or less complicated deployment and maintenance processes. However, neither one of them realized the impact of the model size on, e.g., algorithmic fairness or environmental impact; *P3 "The simpler is the model, the easier it will be to deploy, the easier it will be to monitor, and the easier will be to retrain"* On the contrary, P15 worried that although one might want to use smaller models and less computational power to reduce the energy consumption of model training, it was not possible as they would not be able to achieve the same accuracy levels.

Developers also have to bend to **external constraints** to develop their systems, such as constraints on the data available to train and test the system, due to factors such as the feasibility and cost of collecting new data. A few of the developers directly perceived such constraints as obstacles to building fairer models; *P1 "In machine learning, you will often see that people choose a target label based on what happens to be available or what's easy to get rather than when you think about more statistical inference and stuff like that, then it's typically much more well thought out. Many of the issues with fairness can come from mismeasurement."* Few developers also raised challenges related to the **time** they are given to develop their systems, and the inability to handle harms in this time, such as *P22 "[talking about algorithmic fairness] Everybody has deadlines and this is going to add to the work. But it is important in the long run."*

Finally, seven developers posited that addressing certain harms is **inherently in tension** with other harms. For instance, within a harm category, in the vein of fairness impossibility results [59], P21 discussed the impossibility of simultaneously satisfying several fairness metrics; *P21 "optimizing for one type of fairness will suddenly make another type of fairness worse. if I optimize for fairness between individuals, it's possible that the fairness between groups will suffer, but also even one level lower, if I optimize for predictive parity, it's possible that the disparate impact will suffer."* Other participants discussed tensions across categories of harm. For instance, P9 envisioned that making a system fairer would require collecting more data, which could be privacy-infringing, and certain participants' conceptions of harms were contextual and extremely relative, as they considered the environmental impact of model training non-harmful as long as the ML system was desirable for society or that it would somehow allow to save some energy somewhere, while others solely saw the potential for harm.

Beyond not always being aware of the tensions, note that developers sometimes hold invalid conceptions around these tensions. For instance, nine developers envisioned the acontextual existence of a fairness-accuracy trade-off [22, 27, 70], especially because they did not reflect on data biases that might render measures of accuracy invalid. One developer considered a feature harmful to be used by the model but argued for not dropping it, believing they would not be able to monitor for output bias (incorrect as the training and test sets can be different). Few prior works have studied these tensions and potential misconceptions quantitatively.

4.2.2 Prioritizing Amidst Tensions

Because of the tensions, developers have to prioritize certain objectives or harms. These priorities differ across developers. For instance, while some developers reported being ready to use smaller models and datasets resulting in less accurate models in order to reduce the environmental or labor impact of model training, others judged model performance as the highest priority to optimize the model. Their prioritization was mainly informed by how important and severe they considered each harm individually, and relatively (when they perceived a tension, such as *P21 "This boils down to making a rational choice of what are we actually trying to optimize at the early stages? And keeping in mind that making some sort of fairness metric better, it can still negatively influence other metrics."*), the feasibility and effort needed to address the harm, and various cost-benefit trade-offs (e.g., utilitarian view vs. libertarian view) such as *P18 "This would not really be of my concern as in having to include, for sex, maybe 20 categorical options. Because at the end of the day, we're not doing politics, we're trying to solve a problem"*. Often, prioritization was found to be context-dependent, as demonstrated by P8 when discussing the trade-off between the environmental impact of ML systems and the desirability of these systems; P8 *"It's not something that's on the top of my mind in the case of a model for a hospital. But for models being made for creating new images, like creating artwork, you could think is that worth it? There's a fine line in between the hospital and artwork."*

4.2.3 Defining Various Goals

Developers who consider it important to handle a concern do not all take upon the same goals. Most adopted goals to mitigate the harmful impact. Yet, others did not because of other priorities and tensions, or the lack of (awareness of) methods for mitigation. For instance, a few developers discussed the impossibility of addressing subjectivity in labels; *P5 "As far as I have a reasonable comfort on the quality of data, I'll go ahead. There's no end point to understanding data annotation, there will always be bias."* Beside the pragmatic decision not to address a harmful impact, certain developers mentioned keeping track of the harm (e.g., when a population is silenced if the corresponding records are erased from the data) as a memo to carefully use the system, and sometimes to design work-arounds the harm, e.g., by having human decisions for the non-supported populations. *P21 "I would see* *whether we have any important outliers in the data. What could be a problem is: say you know that five people in this big dataset of 100000 records spent in hospital 100 days and all the others spent less than 20. Then the question would be whether the model that I built is at all applicable to such people. Probably not, so maybe it's best to remove records that seem to have very strong outliers. And have that caveat that the model shouldn't be applied in some very rare cases."* The last solution that three developers proposed is not deploying the system, or making the harm transparent for the decision maker to take such an executive decision; *P1 "if you need the mitigation approaches for the model to be accurate or have a good selection rate, you should question whether ML makes sense to use in this scenario." P6 "I would have this conversation with the hospital. I could only say where we're confident and where we're not."* We refer the readers to Appendix Table 6 to obtain more details on the ways harms are prioritized and how their handling is operationalized.

Beyond the binary decision of addressing harm, developers discuss the extent to which the severity of a harmful impact has to be decreased to be satisfied with the ML system. The thresholds of satisfaction and the rationale for establishing such thresholds differ across developers. They either relied on the judgment of other stakeholders (e.g., data subjects, model requesters, or domain experts *P6 "What is an acceptable difference in performance is a difficult question, and that's something you want to talk to all the stakeholders about."*), on comparisons with prior algorithmic or human baselines, or on their intuition (P27: *"In an ideal scenario, you want the system to be fully fair and accurate, but if you increase one, you decrease the other. So we want to cut in half the burrito, like an optimal trade-off. And that's context-dependent. If fairness is important, for example you have to classify felonies with race, then you shift to fairness, but if fairness is a low priority in the context, then you shift more to accuracy."*)

4.3 Acting on the Concerns: Plurality of Operationalisation Practices

4.3.1 A More Complex Workflow for Handling ML Harmful Impact

From our analysis, eight activities that ML developers perform specifically to handle harm emerged, in addition to the typical ML lifecycle activities that can impact harm. These are 1) understanding the allocation of responsibilities and power relations within their organization to identify potential obligations or obstacles to tackle harms; 2) envisioning the potentially harmful impact of the project; 3) identifying tensions between the potentially harmful impact of their ML system and other aspects of the systems; 4) prioritizing harms and setting up realistic goals for each harm; 5) identifying, adapting/developing, and applying algorithmic unfairness metrics and mitigation methods; 6) identifying, developing, and applying strategies to account for the other harms ML models foster; 7) actively warning the stakeholders empowered to deploy the ML model about the harms; and 8) working to develop reusable toolkits and responsible AI processes within their organization (often voluntarily). Not all developers performed each step, e.g., as they would not necessarily consider something to be harmful (subjectivity), nor realize the existence of tensions, or they would not have the opportunity or responsibility (nor would they take this responsibility) to handle harms. Certain activities also occur in different orders, sometimes iteratively, e.g., 5) and 6) are often performed simultaneously, and potentially serve to update on 3) and 4). That the process of handling harms of ML systems consists of multiple steps, in addition to the traditional ML lifecycle, is typically not accounted for by any prior research on ML workflows [19, 60, 77, 88, 127]. Only the idea of negotiating goals (4) has been made explicit in the past [84], and the one of understanding power relations (1) has been hinted at [7, 69]. We now discuss 5) and 6) in more depth as they are crucial to ML harm practices.

4.3.2 A Diversity of Approaches for Handling ML Harmful Impact

Developers adopt various strategies to actively handle harms outside the distributive fairness category. They might bring additional constraints onto the development process (e.g., on the dataset size, schema, or computational power, such as *P15 "We have 20000 GPUs and it gives a very high human-level accuracy. On the flip side, if you have this much power budget, how do you obtain this same accuracy within any alternative algorithm with much less compute power?"*), or engage in additional data engineering and model engineering efforts (e.g., deletion or re-collection of data in relation to privacy). They also sometimes envision restructuring the learning task and the broader system design and interactions with users, in the case of the desirability of the ML system.

As for distributive fairness, developers employ various approaches to quantify and tackle it. For instance, they considered one or multiple fairness metrics simultaneously, often selected among either group performance or group distribution, but sometimes among individual fairness (causal fairness metrics were only mentioned by one developer); *P2 "because this model will work in hospital with patients where fairness is important, we check all the group fairness metrics of FairLearn."* Similarly, for mitigating unfairness, they either proposed various manual or semi-synthetic transformations of the dataset, or applied different fairness mitigation methods across the three existing categories of methods. While most approaches revolve around data and algorithmic changes related to mitigation methods from the literature, some system-design-level transformations are also proposed that are not extensively discussed in the literature. For instance, P28 brought the need to develop a different, more usable, interface for the decision subjects to enter their data (avoiding dataset under-representation from minority individuals not familiar with the technology or input language), five developers proposed to leave out under-represented populations from the dataset and model, and five others modeled a new learning task; *P6 "We actually have enough data that we might be able to train separate models. So you might not even use the normal FairLearn strategy, which is to train one model that works well across populations."* Three participants also talked about envisioned non-technical solutions to harm identified by assessing algorithmic fairness *P29 "If you find some disparity, what does that mean in the real world? What is the intervention you take? If you don't understand the harm, you can't take an intervention to stop the harm. That part is very important because there are plenty of cases where there's an intervention that isn't technical."* Appendix Table 8, 9 and 10 lists the ways in which distributive unfairness is identified and mitigated.

4.3.3 A Diversity of Critical Reflections around Handling ML Harmful Impact

Some approaches employed might not be appropriate, either because they do not have the intended effects stated by the developer, or because they can cause new harm in certain contexts. For instance, in order to reach algorithmic fairness, three developers proposed to simply drop the sensitive attribute that presents unequal distributions, overlooking the limitations of "fairness through unawareness" [28] and especially the existence of proxy attributes that might skew a model. With regard to the issues that we had injected in the ML systems, 30% of developers did not realize the need for data sampling transformations to reach algorithmic fairness, nor the limitations associated with having too few data samples for certain categories of population. Other developers decided to aggregate data of different underrepresented groups to create a more equally-distributed dataset in comparison to the majority group, without envisioning that relevant differences between these groups might prevent algorithmic fairness [34]. Finally, other developers filtered out under-represented populations to reach parity across smaller numbers of groups, which can lead to harm for the silenced groups —what most did not realize. Concerning algorithmic fairness, this result corroborates prior empirical works, e.g., the problematic belief in fairness through unawareness [23], and empirically validates prior conceptual work, e.g., for the various harmful forgetting practices such as data silences and the flawed WYSIATI ("What You See Is All There Is") assumption conceptualized by Muller et al. [78].

A majority of developers did not engage in reflective practices around epistemic or practical limitations of their workflow. The limitations identified by those who did matched the ones brought up by the conceptual literature. For instance, they talked about the limitations of fairness metrics in accounting for individual differences when receiving wrong outputs [75] or accounting for the impact of the systems on non-decision-subjects stakeholders [61]. For fairness mitigation, they discussed that some approaches might not be considered ethical [118] –P1 *"One thing that people very commonly do is use different decision thresholds. The ones that I was talking about earlier for different groups, and that's a very easy way to get different selection rates, but what does it imply in practice? What this really means is that you literally put people to a different standard. And then whether that's justifiable or not, it really depends on the scenario."*–, or that they reflect techno-solutionist trends where the solution allows to reach parity in numbers but does not solve the societal cause of the problem [31]. *P2 "Demographic parity: making the decisions equal for everyone. It depends a lot on the way you do this. You can positively discriminate to get these outcomes, and it differs by use case if this is fair. You can also make the model work less good for the majority group and then it would be demographic parity. I wouldn't consider that fair."* In the face of such limitations, the developers were often at a loss in knowing how to react.

These results validate and corroborate empirical works. Especially, certain participants present misconceptions towards certain fairness metrics [4, 21], and follow various, potentially flawed, rationales for selecting metrics and protected attributes [23, 35, 69, 93, 97]. Our results also extend these works. They elucidate developers' perceptions of the gap between algorithmic fairness and distributive fairness —only a few developers acknowledge it.

5 Implications

5.1 Supporting Practitioners in Every Step of their Workflow

5.1.1 Supporting ML Developers

The multitude of misconceptions and mis-handlings around various harms beyond distributive fairness show the necessity to investigate how to support ML developers, and better understand where these issues stem from [7]. While it might be tempting to standardize harm-related considerations and practices, similarly to prior attempts at standardizing ML processes (e.g., MLOps [2, 110]) or algorithmic fairness [1]), it would be infeasible facing the rich nature of considerations identified, nor desirable due to the subjectivity of the problem. Instead, accounting for the general lack of recognition from ML developers that their work extends beyond a purely technical task to a social-technical one, we argue that the research community should first invest efforts into changing the mindsets of ML developers, and particularly foster contextualisation and reflexivity activities [20, 73], which are not commonplace. Insights from prior work on reflexivity outside ML could be used for this purpose [25, 30].

Short-term, we should equip developers with actionable tools to tackle the various harm-related steps of the ML lifecycle. Drawing upon the insights in this work, educational materials could delve into potential harms of ML and practical tensions, substantiated with specific facts and figures to avoid misconceptions, as well as lists of approaches to handle harms and warnings about mis-handlings, serving as best practices and anti-patterns. Prior works expressing recommendations to ML developers and researchers, e.g., to circumvent potential "fairness" traps [104] could also be leveraged to build such materials. Practical tools could also guide developers in their workflows. Contrary to prior raw toolkits [16] centered around algorithmic fairness questions and existing technical solutions, we argue that practical tools should be designed with specific steps of the workflow we identified in mind. That would unambiguously fulfill needs of developers and avoid difficulties they face to adopt existing tools, e.g., not knowing *when* to use fairness toolkits [23]. To the best of our knowledge, few tools are directed towards ML developers for the steps we identified, whereas identifying potential harmful impact, or eliciting tensions and defining priorities is always important. Existing tools could also be adapted to account for broad harmful impact, be it fairness tools, e.g., via warning messages or checkboxes in order to probe reflections, or other ML tools, e.g., risk assessment [97] or requirement elicitation frameworks with explicit fields around harmful impact. In any case, the tools should not neglect the diversity of harmful impact concerns we uncovered, and the interdependence of the practices to handle each of them.

5.1.2 ML Developers or Other Practitioners?

Recent debates discuss whether ML developers are the right individuals to address the socio-technical problems of ML (the myth of ML developers as “ethical unicorn” [92]). Our findings echo these debates. The concerns of ML developers go beyond any computer science training, e.g., warfare or economic implications of unfair system outputs, which translates into the misconceptions we identified. Besides, along the harm-related steps, various participants expressed the need to consult non-ML experts or resources, e.g., to decide whether their ML system might cause specific harm. To the best of our knowledge, there is no thorough argumentation suggesting ML developers as best suited to make the decisions they currently take in each step. It is impossible and not necessarily desirable to expect ML developers to make meaningful decisions—this displaces decisions on subjective topics from an ensemble of domain experts in the context that an ML system is deployed onto a single technical expert.

These considerations open up various questions. On the one hand, we should investigate how to foster collaborations with domain experts and stakeholders all along the harm-related steps of ML developers. Collaboration should be for the ML developers to receive the help needed or for them to supply useful information to appropriate non-developer parties with decision-making powers. Existing works have already identified the need for collaborations with ML developers in other contexts [23, 60, 69, 88, 109, 109, 115, 122, 127], or developed tools for various collaborative purposes [64], and their insights could be leveraged for questions of harms. Works around ML transparency via documentation [3, 17, 26, 38, 46, 49, 53, 73, 74] could also be adapted to log information relevant to the steps we uncovered, e.g., matrices of tensions identified between factors and harms, and justifications for the resulting prioritization. On the other side, who the relevant stakeholders to involve are and what powers they should have remains an open research question. HCI scholarship has started to broaden its scope from ML developers to involving UX designers in ML workflows [24, 109], and even considering broader organizational context [93, 113]. Designing new roles specialized in ML harms is also a new trend [96]. Further conceptual and empirical work is required to understand the pros and cons of involving various stakeholders, and their concomitant challenges.

5.2 Expanding Conceptual Research on Harms, A Tool to Reflect on Practices

To corroborate and extend prior empirical works, we extensively leveraged prior conceptual works, and our findings in turn have the potential to inform such efforts. Such prior conceptual works focus on algorithmic fairness, formalize issues around flawed assumptions made by developers or researchers [41, 66, 86, 104, 115], discuss the underlying philosophical theories of different algorithmic fairness tools [13, 31, 36, 62, 118], demonstrate results about tensions [14, 54, 59, 89, 106, 124], and analyse sources of unfairness [32, 71, 103, 121]. These works represent rigorous frameworks for us to critically analyze the conceptions and practices of our participants, e.g., when they discussed apparent but sometimes invalid trade-offs between group and individual fairness metrics [14] or distributive and procedural fairness [42] or between accuracy and algorithmic fairness [54].

These works do not yet characterize every conception, prioritization, and handling approach we identified, especially around harms beyond distributive questions. Future work could investigate each finding independently, e.g., by conducting empirical studies, theoretical proof-based works, or conceptual reflections, to better understand their ins and outs. Particularly, our results outline a multitude of unspoken factors in the research community, e.g., conflicting ML performance, infrastructure requirements, or external data constraints (except the conflicting business/developer goals [69, 83, 85], and lack of metrics and mitigation methods for certain contexts [50]). As these factors are inherently in tension with harms, they unavoidably have to be accounted for by developers.

5.3 Revisiting the Methodologies Employed in Empirical ML Scholarship

Study Design. We adopted a design that shifts from prior works [23, 50] to question prior assumptions, moving away from a study around the use of a tool (we only use notebooks and toolkits as probes to investigate current practices) and away from directly prompting for challenges and specifying harm, to a study around general practices leaving open the concerns. This enabled us to uncover new limitations and challenges in the practices of ML developers, leading to new research implications, especially showing that fairness toolkits might not be a solution in cases where ML developers do not hold meaningful reflections around distributive fairness. We notice an interesting parallel between the predominant techno-solutionist approach to solving distributive fairness via the limited concept of algorithmic fairness, and the HCI trend of developing fairness toolkits and studying challenges with algorithmic fairness conceptualizations without examining the needs first. While these prior works have been essential first steps towards supporting ML developers, some challenges previously identified, e.g., with fairness toolkits [23], could have been avoided by conducting formative studies around ML practices. Hence, introducing more diverse need-finding methodologies from the HCI community [39, 128] could help our community ground future research endeavors in the needs of practitioners.

Limitations. Although we are among the first to explore methodological shifts and holistically analyse concerns and related practices in the context of ML's harmful impact, we should not be the last. Our experimental setup bears limitations that might hinder the generalisability of our findings. While we strived to recruit a diverse set of participants in terms of demographics and experience with ML, it was not possible to obtain a larger sample for each category. Several of our observations, however, corroborate findings from previous studies, hinting at their validity. Yet, focusing on other domains –especially participants' own use cases within their particular organizational context–, and on less-represented segments of the population using targeted recruitment methods would be important in the future. Finally, we acknowledge our own unavoidable subjectivity in identifying and characterizing potential harms and misconceptions, calling for further efforts of triangulation.

6 Conclusion

Our study represents a testimony of the constant socio-technical negotiations [84] needed to build a machine learning model. Our results echo previous studies on algorithmic fairness and contribute to the effort of triangulation of results in HCI research [68] for ML. We also complement prior works with new evidence of the complex and potentially worrying state of ML practices around broader harms, building a deeper and more comprehensive understanding of the (mis)conceptions and (mis)handling around algorithmic harms. This raises theoretical, design, methodological, and governance challenges to ultimately guide practitioners in curbing the impact of ML models. We believe that transdisciplinary efforts are needed to tackle these challenges.

Acknowledgments

We thank all the participants of our studies, without whom this work would not have been possible. We also thank Pablo Biedma Núñez, Eva Noritsyna, Harshitaa Pandey, and Ana-Maria Vasilcoiu for contributing to the elaboration of the research set-up and performing a preliminary analysis on the data presented in this paper. These contributions can be found in their Bachelor theses [80–82, 114].

References

[1] Avinash Agarwal, Harsh Agarwal, and Nihaarika Agarwal. 2022. Fairness Score and process standardization: framework for fairness certification in artificial intelligence systems. *AI and Ethics* (2022), 1–13.

[2] Sridhar Alla and Suman Kalyan Adari. 2021. What is mlops? In *Beginning MLOps with MLFlow*. Springer, 79–124.

[3] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[4] Zahra Ashktorab, Benjamin Hoover, Mayank Agarwal, Casey Dugan, Werner Geyer, Hao Bang Yang, and Mikhail Yurochkin. 2023. Fairness Evaluation in Text Classification: Machine Learning Practitioner Perspectives of Individual and Group Fairness. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[5] 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).

[6] 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.

[7] 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 *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 482–495.

[8] Jack Bandy. 2021. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction* 5, CSCW1 (2021), 1–34.

[9] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

[10] Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. 2022. It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 248–266.

[11] 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.

[12] 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

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

[14] 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.

[15] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. 2023. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 396–410.

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

[17] 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.

[18] 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

[19] Ángel Alexander Cabrera, Marco Tulio Ribeiro, Bongshin Lee, Robert Deline, Adam Perer, and Steven M Drucker. 2023. What did my AI learn? how data scientists make sense of model behavior. *ACM Transactions on Computer-Human Interaction* 30, 1 (2023), 1–27.

[20] Scott Allen Cambo and Darren Gergle. 2022. Model Positionality and Computational Reflexivity: Promoting Reflexivity in Data Science. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[21] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. 2021. Soliciting stakeholders' fairness notions in child maltreatment predictive systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[22] A Feder Cooper, Ellen Abrams, and Na Na. 2021. Emergent unfairness in algorithmic fairness-accuracy trade-off research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 46–54.

[23] 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).

[24] Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. 2023. Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 705–716.

[25] Kimberly Do, Rock Yuren Pang, Jiachen Jiang, and Katharina Reinecke. 2023. "That's important, but..." How Computer Science Researchers Anticipate Unintended Consequences of Their Research Innovations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[26] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 9. 48–59.

[27] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. 2020. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International Conference on Machine Learning*. PMLR, 2803–2813.

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

[29] Elizabeth Edenberg and Alexandra Wood. 2023. An Epistemic Lens on Algorithmic Fairness. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–10.

[30] Salma Elsayed-Ali, Sara E Berger, Vagner Figueredo De Santana, and Juana Catalina Becerra Sandoval. 2023. Responsible & Inclusive Cards: An Online Card Tool to Promote Critical Reflection in Technology Industry Work Practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.

[31] 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.

[32] Martínez-Plumed Fernando, Ferri Cèsar, Nieves David, and Hernández-Orallo José. 2021. Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems* 36, 7 (2021), 3217–3258.

[33] 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

[34] Jessie Finocchiaro, Roland Maio, Faidra Monachou, Gourab K Patro, Manish Raghavan, Ana-Andreea Stoica, and Stratis Tsirtsis. 2021. Bridging machine learning and mechanism design towards algorithmic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 489–503.

[35] Cosima Friedle. 2022. How Can Fairness Tools Impact the Understanding of Fairness and the Processes Within a Machine Learning Development Team? *Junior Management Science* 7, 5 (2022), 1289–1300.

[36] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (2021), 136–143.

[37] 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.

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

[39] Colin M Gray, Erik Stolterman, and Martin A Siegel. 2014. Reprioritizing the relationship between HCI research and practice: bubble-up and trickle-down effects. In *Proceedings of the 2014 conference on Designing interactive systems*. 725–734.

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

[41] Nina Grgić-Hlača, Christoph Engel, and Krishna P Gummadi. 2019. Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[42] 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

[43] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 501–512.

[44] Jacqueline Hannan, Huei-Yen Winnie Chen, and Kenneth Joseph. 2021. Who gets what, according to whom? an analysis of fairness perceptions in service allocation. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 555–565.

[45] 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.

[46] Amy Heger, Elizabeth B Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. *arXiv preprint arXiv:2206.02923* (2022).

[47] Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. 2019. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the conference on fairness, accountability, and transparency*. 181–190.

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

[49] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. The dataset nutrition label. *Data Protection and Privacy, Volume 12: Data Protection and Democracy* 12 (2020), 1.

[50] 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

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

[52] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[53] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 560–575.

[54] Rashidul Islam, Shimei Pan, and James R Foulds. 2021. Can We Obtain Fairness For Free?. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 586–596.

[55] 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.

[56] Jared Katzman, Solon Barocas, Su Lin Blodgett, Kristen Laird, Morgan Klaus Scheuerman, and Hanna Wallach. 2023. Representational Harms in Image Tagging. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence (2023)*, Vol. 5.

[57] 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

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

[59] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, Vol. 67. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 43.

[60] 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.

[61] 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.

[62] Derek Leben. 2020. Normative principles for evaluating fairness in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 86–92.

[63] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.

[64] 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.

[65] 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.

[66] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the impact of automated suggestions on decision making: Domain experts mediate model errors but take less initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[67] 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.

[68] Wendy E Mackay and Anne-Laure Fayard. 1997. HCI, natural science and design: a framework for triangulation across disciplines. In *Proceedings of the 2nd conference on Designing interactive systems: processes, practices, methods, and techniques*. 223–234.

[69] 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.

[70] Subha Maity, Debarghya Mukherjee, Mikhail Yurochkin, and Yuekai Sun. 2020. There is no trade-off: enforcing fairness can improve accuracy. *stat* 1050 (2020), 6.

[71] Fernando Martínez-Plumed, Cèsar Ferri, David Nieves, and José Hernández-Orallo. 2019. Fairness and missing values. *arXiv preprint arXiv:1905.12728* (2019).

[72] 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.

[73] 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.

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

[75] 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.

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

[77] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[78] Michael Muller and Angelika Strohmayer. 2022. Forgetting Practices in the Data Sciences. 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 323, 19 pages. https://doi.org/10.1145/3491102.3517644

[79] 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.

[80] Eva Noritsyna. 2022. Surfacing Differences in Practices When Building Fair Machine Learning Systems with Fairness Toolkits: an Empirical Study. https://resolver.tudelft.nl/uuid:05e20f01-b1a2-4542-80f8-7e9ba12ab097

[81] Pablo Biedma Nunez. 2022. Who Cares About Fairness: How Background Influences the Way Practitioners Consider Machine Learning Harms. https://resolver.tudelft.nl/uuid:80206eee-fd99-4511-ba08-95716e3f1cf7

[82] Harshitaa Pandey. 2022. Comparison of the usage of Fairness Toolkits amongst practitioners:AIF360 and Fairlearn. https://resolver.tudelft.nl/uuid:4ef11035-2f60-436f-85f9-7b9bed73b66d

[83] 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.

[84] 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

[85] 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

[86] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. 2019. What you see is what you get? the impact of representation criteria on human bias in hiring. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 125–134.

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

[88] 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.

[89] David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. 2020. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 189–199.

[90] 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

[91] 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

[92] 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.

[93] 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

[94] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D Ziebart. 2021. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 9419–9427.

[95] 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.

[96] Shalaleh Rismani and AJung Moon. 2023. What does it mean to be an AI Ethicist: An ontology of existing roles. *AIES* (2023).

[97] Shalaleh Rismani, Renee Shelby, Andrew Smart, Edgar Jatho, Joshua Kroll, AJung Moon, and Negar Rostamzadeh. 2023. From plane crashes to algorithmic harm: applicability of safety engineering frameworks for responsible ML. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[98] 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

[99] Bonnie Ruberg and Spencer Ruelos. 2020. Data for queer lives: How LGBTQ gender and sexuality identities challenge norms of demographics. *Big Data & Society* 7, 1 (2020), 2053951720933286.

[100] Debjani Saha, Candice Schumann, Duncan Mcelfresh, John Dickerson, Michelle Mazurek, and Michael Tschantz. 2020. Measuring non-expert comprehension of machine learning fairness metrics. In *International Conference on Machine Learning*. PMLR, 8377–8387.

[101] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[102] 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).

[103] Sebastian Schelter, Yuxuan He, Jatin Khilnani, and Julia Stoyanovich. 2020. FairPrep: Promoting Data to a First-Class Citizen in Studies on Fairness-Enhancing Interventions. In *Proceedings of the 23nd International Conference on Extending Database Technology, EDBT, 2020*.

[104] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[105] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 723–741.

[106] Reza Shokri, Martin Strobel, and Yair Zick. 2021. On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 231–241.

[107] 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.

[108] 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).

[109] Hariharan Subramonyam, Jane Im, Colleen Seifert, and Eytan Adar. 2022. Solving Separation-of-Concerns Problems in Collaborative Design of Human-AI Systems through Leaky Abstractions. 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 481, 21 pages. https://doi.org/10.1145/3491102.3517537

[110] Damian A Tamburri. 2020. Sustainable mlops: Trends and challenges. In *2020 22nd international symposium on symbolic and numeric algorithms for scientific computing (SYNASC)*. IEEE, 17–23.

[111] 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. https://doi.org/10.1145/3287560.3287566

[112] Niels Van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. 2021. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[113] Rama Adithya Varanasi and Nitesh Goyal. 2023. "It is currently hodgepodge": Examining AI/ML Practitioners' Challenges during Co-production of Responsible AI Values. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[114] Ana Maria Vasilcoiu. 2022. Closer or even farther from fairness: An assessment of whether fairness toolkits constrain practitioners with regards to algorithmic harms. https://resolver.tudelft.nl/uuid:a3e99758-3d1b-402a-bd33-2ab9fd871ce2

[115] 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, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox (Eds.). ACM, 440. https://doi.org/10.1145/3173574.3174014

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

[117] 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.

[118] 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).

[119] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.

[120] David Gray Widder, Derrick Zhen, Laura Dabbish, and James Herbsleb. 2023. It's about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 467–479.

[121] 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

[122] 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).

[123] 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

[124] Depeng Xu, Shuhan Yuan, and Xintao Wu. 2019. Achieving differential privacy and fairness in logistic regression. In *Companion proceedings of The 2019 world wide web conference*. 594–599.

[125] 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

[126] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effects of Explanations, Human Oversight, and Contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[127] 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.

[128] John Zimmerman and Jodi Forlizzi. 2014. Research through design in HCI. In *Ways of Knowing in HCI*. Springer, 167–189.

[129] 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