A Taxonomy of Challenges to Curating Fair Datasets

Dora Zhao*
Stanford University

Morgan Klaus Scheuerman*
Sony AI

Pooja Chitre†
Arizona State University

Jerone T. A. Andrews†
Sony AI

Georgia Panagiotidou
King's College London

Shawn Walker‡
Arizona State University

Kathleen H. Pine‡
Arizona State University

Alice Xiang‡
Sony AI

Abstract

Despite extensive efforts to create *fairer* machine learning (ML) datasets, there remains a limited understanding of the practical aspects of dataset curation. Drawing from interviews with 30 ML dataset curators, we present a comprehensive taxonomy of the challenges and trade-offs encountered throughout the dataset curation lifecycle. Our findings underscore overarching issues within the broader fairness landscape that impact data curation. We conclude with recommendations aimed at fostering systemic changes to better facilitate fair dataset curation practices.

1 Introduction

Persistent concerns from academia, government, industry, and the public sphere center on the disparate impact and unfairness in machine learning (ML) [22, 28, 32, 69–71, 74, 77, 94, 143, 159]. Data is often viewed as a primary culprit, perpetuating biases and compromising fairness [36, 52, 88, 164]. In response, substantial attention has been directed towards *fair* dataset collection practices [43, 46, 62, 65, 114, 121, 135, 161, 164]. However, there remains a significant gap in understanding both the practices and practicalities of fair dataset curation.

To address this gap, we shift from theoretical, guideline-focused scholarship [3, 41, 42, 51, 63, 78, 80, 103, 117] to empirical inquiry, exploring the grounded practices of fair dataset curation. Following a well-established tradition in human-computer interaction (HCI) [67, 75, 97, 107, 125, 149], we conducted interviews with 30 dataset curators from both academia and industry who have experience curating fair vision, language, or multi-modal datasets. Through these interviews, we uncover practical challenges and trade-offs to ensuring fairness in dataset curation. Our use of qualitative methodology allowed us to surface nuanced challenges and trade-offs that regularly appear throughout the curation process and gain insights into considerations that may otherwise remain undisclosed.

We first provide three dimensions of fairness—*composition*, *process*, and *release*—that participants considered during curation. Fairness is not only a property of the final artifact—the dataset—but also a constant consideration curators must account for throughout the curation process. Through our empirical findings, we identify various challenges that obstruct different fairness goals. Building on Hutchinson et al. [78]'s conception of the dataset lifecycle, we contribute a taxonomy of challenges dataset curators encounter, addressing both dataset lifecycle-specific challenges (Section 3) and those within the broader landscape of fairness in ML (Section 4). By conducting in-depth interviews

*Joint first author

†Joint second author

‡Joint last author

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

https://doi.org/10.52202/079017-3103

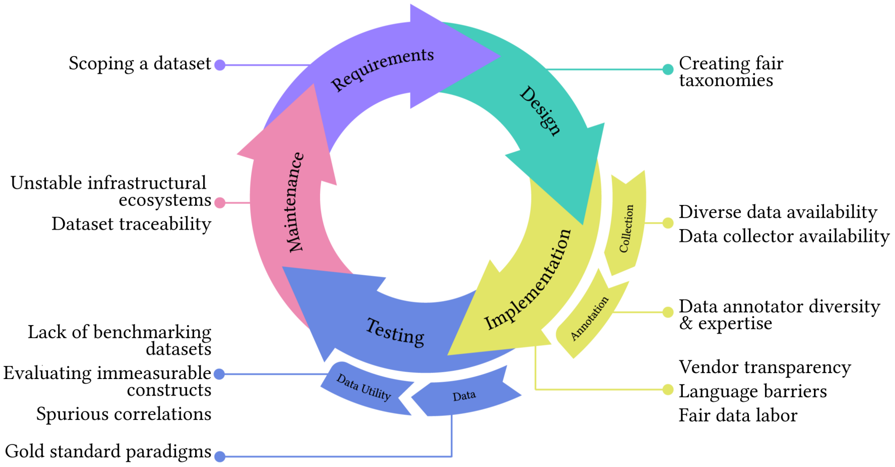

Figure 1: A circular process diagram showing how each challenge we identified maps to each phase and subphase of the dataset lifecycle.

with those engaged in fair dataset work on the ground, we provide empirical support for prior work [3, 42, 55, 78–80, 91, 93, 111], which has focused on identifying implicit challenges in the fairness literature (see Appendix B for additional background). We conclude with recommendations aimed at fostering systemic changes to better facilitate fair dataset curation practices (Section 5).

Our work aligns with existing recommendations for fair dataset curation [3, 12, 42, 51, 78, 80, 98, 102, 103, 111, 117, 131] and aims to deepen stakeholders' understanding of the specific challenges involved. By illuminating these issues, we hope to expedite more effective solutions and promote further investigation into the complexities of fairness in dataset curation.

2 Method

To understand the challenges of collecting fair datasets, we conducted 30 semi-structured interviews with ML dataset curators, each lasting between 45–60 minutes, between November 2023 and March 2024. Participants were asked to define fairness in ML datasets, describe their process for collecting fair datasets, highlight challenges encountered, and discuss any trade-offs made. Refer to Appendix A for more details, including Institutional Review Board approval.

Participants. To qualify, participants must have previously curated at least one fair ML dataset. Given the extensive discourse surrounding language and vision dataset practices [12, 16, 111], we prioritized participants in these domains. To accommodate diverse perspectives, we refrained from prescribing a specific definition of "fair." Initial recruitment was conducted through purposive sampling [142], targeting authors of public datasets, followed by outreach via social media and relevant mailing lists, with snowball sampling [142] used to expand participation.

To protect anonymity, participants are referred to as "PX", where "P" denotes "Participant" and "X" represents their identification number (e.g., P8).

Thematic Analysis. To analyze the interviews, we adopted an inductive approach [20]. We began with an initial set of codes derived from our literature review on challenges in fair data collection. Four authors independently coded the same interview to identify additional themes, refined the codebook through discussion, and repeated the process with a second interview. The remaining interviews were then equally distributed among the research team for thematic analysis.

3 Challenges During the Dataset Lifecycle

We present challenges participants encountered across the dataset lifecycle, taxonomizing them into requirements, design, implementation, evaluation, and maintenance phases (see Figure 1 and Table 3).

https://doi.org/10.52202/079017-3103

Recognizing the multi-faceted nature of *fairness*, we did not impose a specific definition during our interviews. Instead, we empowered participants to articulate their own definitions. Based on these definitions, we identified three dimensions of fairness: *composition*, which is achieved through diverse representations; *process*, which includes equitable compensation for data subjects and workers as well as recognition for curation efforts; and *release*, which emphasizes the importance of transparent and openly accessible data. The challenges we surface span all three dimensions of fairness.

3.1 Requirements

The *requirements phase* involves establishing a dataset's purpose (e.g., intended tasks such as image tagging) and defining the fairness criteria to be operationalized within the dataset (e.g., group fairness). Challenges in this phase most often manifested in the composition and process dimensions.

Scoping a dataset. Participants sought to balance fairness with utility (P8, P23, P26, P30). On the one hand, careful curation can lead to more nuanced insights compared to general-purpose datasets. As P26 explained, they would ideally "*design smaller datasets for smaller models for specific applications, nothing that is deployed on a [South Asian] scale, because that definitely won't work properly because of the [region's] geographical diversity*." Moreover, datasets containing billions of entries, such as LAION [133, 134], make oversight difficult and, as a result, may include "*unfair*" data (P18) [15]. Nonetheless, participants also had to consider utility. P13 noted ML is "*in this age of scale,*" making them "*a bit skeptical as [to] whether people are going to openly use fair datasets for training unless they're very large.*" P21 highlighted a similar tension between "*technical reasons why you need large open datasets*" and "*ethical reasons on why that shouldn't be the case.*" Fairness trade-offs pushed some (P12, P13) towards focusing on smaller evaluation datasets.

Determining fairness definitions. Nearly all participants stressed the *contextual* nature of fairness. Key factors shaping their definitions included domain (e.g., healthcare), task (e.g., sentiment analysis), and cultural context. For example, P2 highlighted the importance of cultural specificity, stating, "*you see a lot of work that talks about fairness in gender or in race. But for a [South Asian] country, race does not manifest like it manifests for America.*" Participants also made trade-offs due to the multitude of fairness definitions available [104] (Section 4.5). P19 noted that "*there's more than two dozen different fairness definitions ... used in the literature.*" This diversity necessitated sacrifices in other dimensions, as emphasized by P18, who illustrated this with the "'*no free lunch theorem*", stating, "*You can't have complete diversity with respect to, say, races,...geographies,...times of the day, and other domains. Everything is not possible. Once you clamp on one, the other one goes away.*"

3.2 Design

In the *design phase*, curators determine how to operationalize dataset requirements, including defining the dataset's taxonomy. For example, curators specify attributes for measuring fairness (e.g., skin tone) and the categories within those attributes [66, 68, 114, 145]. This phase also involves decisions on data collection and annotation methodologies (e.g., web scraping, hiring vendors). Challenges in this phase typically arose in the composition and process dimensions.

Creating fair taxonomies. Participants struggled to find a fair taxonomy under the inherent unfairness of categorization. For example, P18 devised a geographic taxonomy featuring categories for the U.S. and Asia, acknowledging that the regions "*are not homogeneous, they're very heterogenous.*" P2 also noted a theoretically ideal taxonomy is as granular as possible, but practical constraints, such as data availability (Section 3.2) and time (Section 4.3), necessitated using coarser categories. Finally, the challenge of creating a fair taxonomy was compounded by the inadequacies of existing domain taxonomies. For example, P1 and P5 pointed out that the common binary operationalization of gender in medical data erases many gender identities. Nonetheless, participants felt compelled to utilize inadequate taxonomies due to practical constraints, even if it contradicted their personal beliefs. Participants were forced to align their notions of fairness with disciplinary norms (Section 4.2).

Data availability in taxonomy design. Similar to when designing taxonomies, participants had to balance their ideal data collection methods with practical constraints. For example, P3's dataset only included Spanish and Arabic even though they *"wanted to look at other languages, but ... didn't have training data."* Participants questioned prevailing data collection paradigms, such as web scraping [3], which were seen as unethical when performed indiscriminately. For legal compliance, P25 manually collected data for two years: *“I was downloading, like clicking and clicking, because they didn’t allow me to do web scraping or didn’t have an API.*”

3.3 Implementation

The *implementation phase* marks the execution of plans from the design phase, where curators collect, annotate, and package the data into a dataset. This phase broadly encompasses two subphases: data collection and data annotation. Challenges in this phase span all three dimensions of fairness.

3.3.1 Data Collection

Data collection involves gathering relevant data to fulfill dataset requirements. Challenges during this subphase prevented participants from attaining fairness goals relevant to dataset diversity.

Diverse data availability. Similar to concerns raised regarding dataset taxonomies (Section 3.2), participants raised concerns about data availability for creating a fair dataset. For example, P28 described how sexist stereotypes permeate web data, such as *“women [being] associated with nurse more often than men.”* Additionally, P18 encountered difficulties sourcing web data from *“Middle Eastern”* and “*African countries”* but found “*lots and lots and lots of images from India, Japan and [the] U.S., which are like the three most dominant geographies in uploading pictures.*” Participants also lamented the inaccessibility of specialized or proprietary data, such as medical records or data from private companies, which could significantly improve the creation of fair datasets. P4 stated that *“because people don’t own large e-commerce platforms or social media platforms, or whatever, we just kind of have to deal with things that we can gather from existing systems.”*

Interestingly, synthetic data, sometimes presented as a potential solution to biased data [6, 120, 141], was met with skepticism as it could perpetuate stereotypes or inadequately represent underrepresented groups [156]. As P19 pointed out, *“You might address some of the missing data points [with synthetic data] but at the end of the day it’s still the same underlying data distribution, right?”*

Data collector availability. Many participants associated fairness with geographically diverse data. For example, P22 expressed how they would *“proactively sample more data from underrepresented regions.”* Yet, actualizing this objective proved challenging, as P12 highlighted the difficulty in *“get[ting] hold of people ... from very, very small regions.”* Infrastructure hurdles, such as limited internet and mobile phone access, further complicated the process [3, 80]. Equipping data collectors with necessary equipment is costly (Section 4.3) and logistically challenging, as *“you might have to give people smartphones to start and you’d also need more labor on the ground ... who are working in these different regions to come together and do this”* (P12).

3.3.2 Data Annotation

Data annotation involves labeling data with attributes specified during design. Participants faced challenges recruiting annotators who had requisite expertise or came from diverse backgrounds. Upholding fair labor practices (Section 3.3.3) during annotation also presented challenges.

Data annotator diversity and expertise. The interpretation and application of annotation categories can vary based on an annotator’s perspective [2, 7, 25, 81]. P22 described finding annotators for labelling building styles across different geographies: *“You give this same image to a local labeler who is in that culture, who is an expert in, you know, their architecture ... then you get a much better label.”* Yet, participants had difficulty hiring annotators that met their desired aims. While P2 highlighted the value of diverse annotator backgrounds or beliefs to ensure annotations reflected a wide range of experiences, accessing diverse annotators was challenging, *“because some of the attributes of [annotators’] personal lives might even be illegal to ask about in a particular country.”* Participants also confronted challenges in recruiting annotators with specialized expertise. For example, despite offering “*$75 or $100 per hour,*” P1 faced difficulties finding and incentivizing medical experts to annotate radiology data. Annotators who lack diversity or expertise in data concepts may lead to issues with data quality, including inaccuracies [76], biases [44, 48, 127], and overly homogeneous annotations [48, 115]. Notably, P13 highlighted that crowdsourced annotators regularly embed gender biases into datasets such that *“researchers [need] to make sure that annotators represent everyone because [if] not, you’re just gonna have a skewed pool of annotations as well.”*

3.3.3 Implementation Processes

Participants expressed challenges not only with dataset content but also with the *implementation* of data collection and annotation. We provide three main considerations discussed by participants.

Vendor transparency. Collaborating with data vendors introduced transparency challenges, hindering fairness efforts. First, as prior research documented [128], vendors may prohibit access to data worker identities, such as demographic details or location (as described by P2 in Section 3.3.2). Thus, it is impossible to evaluate potential biases or expertise linked to identity characteristics, such as how an annotator's cultural identity may influence their engagement with data concepts. Second, participants had little oversight into worker compensation or encountered communication restrictions imposed by vendors. As P12 said, *"I think [pay] was fair in terms of [being] calibrated across different countries ... but we weren't able to get exact numbers, because that was confidential."* P6 described how vendor platform design inhibited direct collection of feedback from data workers, impeding efforts to improve fairness in dataset creation and labor conditions (e.g., [101, 102]) (Section 4.5).

Language barriers. Curating fair datasets often involves collecting geographically diverse data, which may require data workers proficient in languages different from those of curators. Language barriers can hinder effective communication, necessitating fairness concepts established in the design phase (Section 3.2) to be accurately translated into the workers' native languages. Improper translations can result in misinterpreted labels or instructions and may even lead to contract breaches, particularly concerning subject consent. Addressing language barriers often involves resorting to translation services, which may be constrained by cost (Section 4.3) or introduce its own fairness concerns. Further, participants had to ensure translations accurately reflected their intentions, but as P3 noted, *"We relied on our translators to come up with those sorts of decisions in terms of Spanish."*

Fair data labor. Several participants (P6, P11, P12, P14, P16, P24, P28) expressed concern about engaging in fair labor practices when working with data workers, but systemic organizational (Section 4.3) and regulatory (Section 4.4) issues made achieving these standards difficult.

3.4 Evaluation

The *evaluation phase* involves assessing data quality and testing dataset utility. Challenges in this phase can result in homogeneous annotations, benchmarking difficulties, and spurious correlations, most often affecting the composition and release of a dataset.

3.4.1 Assessing Data Quality

Assessing data quality entails validating and refining the data and its annotations to ensure clarity and consistency with project requirements. (Re)alignment of data and annotations with the guidelines from the design phase is often referred to as *quality assurance*.

Gold standard paradigms. Participants often sought to capture a diversity of perspectives across annotators. Thus, prevailing practices for validation and cleaning, such as majority voting and annotator agreement metrics, may be unsuitable. As P24 emphasized, majority voting can "*squash or stifle diverse opinions when it comes to subjective tasks*." When disagreement is integral to the objective, annotator agreement metrics become inappropriate, making it difficult to "validate" annotation quality. Gold standard paradigms are intrinsically tied to disciplinary challenges (Section 4.2); if submitting a publication involving dataset creation, reviewers might still call for annotator agreement metrics and believe the quality of the data is poor if agreement is low.

Similarly, common practices used to clean or filter data can perpetuate dominant cultural beliefs. Data that might appear noisy or incorrect can hold significance for certain communities. P14 explained how quality filters resulted in "*get[ting] rid of vernacular that's not perfect English but is maybe like African-American vernacular or like Hispanic-American vernacular, and that also introduces bias and lowers the diversity of the dataset.*" This echoes prior work [9] which found that standard data filters might disproportionately exclude content from already marginalized groups.

3.4.2 Evaluating Data Utility

To ensure dataset utility, curators must evaluate its effectiveness, often through *requirements testing* to confirm its suitability for the intended purpose. Participants aimed to align the dataset with fairness definitions and mitigate any potential biases present in the data.

Lack of benchmarking datasets. Curators often seek to benchmark their datasets to showcase their utility. However, since many participants aimed to create unprecedented fair datasets to address existing gaps, this norm posed a challenge as comparable datasets were non-existent. Reflecting on the struggles with a novel geodiverse dataset, P12 explained, "*We couldn't measure it unless we had a dataset that actually was fair. Since we don't have a dataset that is fair..., you are arguing in circles.*" Furthermore, even if comparable datasets exist, they may harbor fairness issues of their own.

Evaluating immeasurable constructs. Evaluating whether a dataset aligns with fairness definitions presupposes that fairness is a construct amenable to measurement. While some participants offered quantifiable indicators of fairness, such as demographic diversity, others argued that fairness defies quantification. P14 criticized measurement-oriented perspectives, stating, "*They also assume that fairness can be measured, can be evaluated, and can be improved. And I think that all of this is a more positivist mindset.* " Even with a definition in mind, testing may feel incomplete. As P28 said, "*Even when you provide a way to measure fairness, you're probably overlooking something.*"

Spurious correlations. Several participants (P6, P23, P28) aimed to avoid introducing spurious correlations that affected the fairness of the dataset's composition [24, 53, 82]. While these correlations may not be "*connected with any demographic or social variable*" (P23), they can still influence downstream models and result in biased decisions. However, as recent research [100] has revealed, spurious correlations with demographic attributes are ubiquitous. Thus, enumerating and removing all possible correlations is virtually impossible.

3.5 Maintenance

In the *maintenance phase*, curators must consider both how their dataset is released and strategies for ensuring its ongoing utility over time. Challenges at this stage often linked back to participant concerns around fairness in dataset release (i.e., ensuring the data is transparent and openly accessible).

Unstable infrastructural ecosystems. Digital data is intrinsically impermanent. Some participants (P1, P8, P30) emphasized the risk of data instances disappearing due to broken links or shifts in platform popularity or ownership, as observed with platforms like Twitter. Therefore, curators must then not only monitor for missing data but also find suitable replacements that match the original dataset's distribution. This can be particularly burdensome when the data was expensive (Section 4.3) or difficult to collect (Section 3.3.1). As data goes missing, datasets can become unbalanced and thus "unfair," demonstrating how fairness issues with data release are linked to concerns about composition.

Dataset traceability mechanisms. The challenge of dataset stewardship is exacerbated by inadequate traceability mechanisms [112, 132]. Participants underscored their inability to track users and usage patterns of their datasets. One commonly used proxy is citations in academic papers, but it was hard to *"distinguish citations that use the data versus citations that use the broader idea of the paper*" (P2). This is concerning, especially if fair datasets are repurposed in unintended ways. While prior works [3, 112, 132] have suggested data usage policies to mitigate such risks, enforcing them becomes impractical when curators are unaware of actual data users.

4 Challenges Overarching the Broader Landscape of Fairness

The dataset curation process is influenced by the environments in which curators operate, meaning their decisions are not made in isolation. Many challenges span all phases of the lifecycle, shaping the broader landscape of dataset fairness. We identified five levels within this landscape, where challenges may emerge from one or more levels, affecting dataset curation at every phase of the dataset lifecycle (see Figure 2 and Table 4).

https://doi.org/10.52202/079017-3103

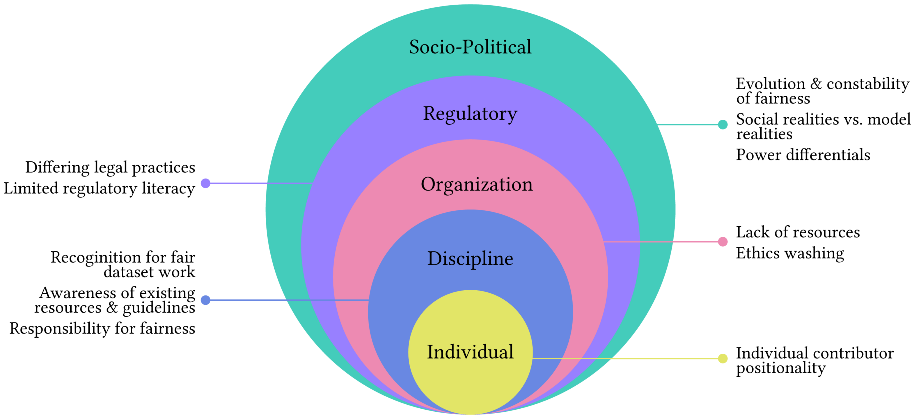
Socio-Political
Regulatory
Organization
Discipline
Individual
Evolution & constability of fairness
Social realities vs. model realities
Power differentials
Differing legal practices
Limited regulatory literacy
Lack of resources
Ethics washing
Recoginition for fair dataset work
Awareness of existing resources & guidelines
Responsibility for fairness
Individual contributor positionality

Figure 2: A social ecological [21] representation of challenges in each layer in the overarching landscape of fairness. A social ecological model shows how each layer is nested but interconnected.

4.1 Individual Level

The *individual level* of the dataset curation landscape refers to the contributors of fair datasets, such as data curators, data subjects, and data workers.

Individual contributor positionality. Decisions made by contributors were inevitably influenced by their own unique perspectives [126, 129]. As P24 said, *"There's this stuff we swim in that we don't really realize is even there."* Despite recognizing this influence, assessing its tangible impact on the dataset remained elusive. Addressing and diversifying contributor positionality is further complicated by other challenges within the dataset curation landscape, such as cost and power differentials. Positionality was evident in instances where participants felt they had to make trade-offs during processes like designing taxonomies that may erase others' experiences. P27 encouraged reflecting on personal values: "*Is this [research] actually in line with your life philosophy? Was it in line with your gender, with your sexuality... If it's not, would you still want to be doing this?*"

4.2 Discipline Level

The *discipline level* of the dataset curation landscape centers on the norms and practices governing specific academic disciplines, particularly ML [14, 45, 118, 131].

Recognition for fair dataset work. Despite the growing demand for data in ML, according to participants, fair dataset curation efforts were not seen as significant contributions to the field. P11 described a *"lack of general disciplinary value of datasets as contributions."* While some major conferences like NeurIPS [148] have introduced dataset tracks, few venues prioritize dataset-focused work. This lack of appreciation discourages efforts to ensure dataset stability and longevity [131].

Incentive mechanisms. Incentives in ML do not align well with the costs of fair dataset curation. According to P11, there's *"just [a] total lack of resources and time to actually deeply engage with labeling and sourcing those labels and getting people who are representative of those labels to be the data workers."* Participants echoed well-documented observations that model work is valued over data work [124, 125, 131], with P21 stating that *"data is kind of a second-class citizen in ML research."* Consequently, P25 felt *"people are [not] seriously talking about fairness ... people are still just get[ting] whatever [data] they get to do their research, or publish, or whatever."*

Awareness of existing resources and guidelines. Participants had limited awareness of existing guidance for fair dataset curation. This lack of awareness may be attributed to some of these resources (e.g., [42, 78, 80, 117]) being disseminated outside of traditional ML venues (e.g., NeurIPS, *CL, ICML, CVPR). As P29 admitted, *"I don't remember any explicit guidelines that I've stumbled through for fair dataset collection. Honestly!"* Promoting interdisciplinary awareness of fairness efforts among those primarily involved in ML is challenging due to highly disciplinary norms that prioritize novelty in ML methods over discussions on fair dataset curation.

https://doi.org/10.52202/079017-3103

Responsibility for fairness. The burden of responsibility for fairness weighs most heavily on individuals aware of fairness concerns in ML. Participants echoed findings from prior research [14] that document how fairness is not a top priority for many ML researchers. For example, P25 said that "*[in] the team I work with... I never heard them talking about [how] the dataset has to be fair.*" In P25's experience, the norm was to cursorily engage with fairness issues without substantive changes to research practices. Given the lower prioritization of fairness in ML, the onus falls on individual researchers who "*have a strong sense of justice and fairness*" (P24) or are part of fairness-oriented communities to elevate these concerns. However, this commitment often lacks external recognition and may hinder resource allocation and research progress. Participants recognized that collecting fair data is more challenging and resource-intensive compared to conventional methods: "*If you want to build a fair dataset, maybe the most efficient way to do that is to scrape the web, but getting really diverse data in an ethical way is really hard and really expensive*" (P11).

4.3 Organization Level

The *organization level* refers to the organizations where individuals conduct fair dataset curation work, which could vary in size or nature, such as academic or industry settings.

Lack of resources. Insufficient resources were a significant challenge across all phases of the dataset lifecycle. As P1 declared: '*'Money?! (laughs) If you have money, you can have a very high quality of data."* Fair data collection methods are costly, especially concerning data quality and annotation, which often require hiring experts. Convincing funders or stakeholders of the value of investing in fair datasets proved difficult, as noted by P24: *"It's hard to convince somebody to spend thousands and thousands to collect [a] dataset of recordings."* Moreover, participants aimed to compensate data subjects and workers fairly, *"not just the minimum wages that many times academia gives"* (P29). Longterm maintenance costs added to the financial burden, with difficulties in securing ongoing funding. P1 stated no academic or industry organization *"[wants] to spend another millions of money every year ... to maintain those products."*

Ethics washing. Participants disapproved of organizations that superficially promote fair ML but fail to meaningfully integrate fairness into their practices [151]. According to P16, the *"language of fairness is simply external lip service [that] ultimately boils down to looking at the maximization of other imperatives, such as economic ones."* Resource constraints exacerbate this issue, leading organizations to prioritize efficiency and cost-effectiveness over fairness. As P22 noted, *"A lot of big companies do responsible AI shenanigans ... for marketing ... And then a new shiny thing comes down the road, and then they join that instead."* When fairness is valued primarily for its marketing appeal rather than its impact on product development, it is not prioritized for monetary or labor investment.

4.4 Regulatory Level

The *regulatory level* concerns laws and policies governing dataset curation and use. Participants expressed anxieties about violating regulations they were not necessarily equipped to fully understand.

Differing legal practices. Contextual laws and regulations posed a challenge for participants. P2 described how *"laws in America or laws in Europe ... might not be directly applicable to a [South Asian] country that has a very different societal situation."* Contextually contingent laws and policies further complicated efforts to obtain data from diverse, underrepresented populations (Section 3.3.1).

Legal risk. Throughout the dataset lifecycle, participants faced the looming risk of unintentionally violating laws and regulations, potentially leading to breaches of privacy, labor, or data ownership laws. Instances of inadvertent violations are not uncommon, as highlighted by participants' experiences with web scraping practices. For example, P21 was aware that *"people discovered links to child pornography"* in a widely used benchmark dataset [16, 144]. In another instance, P5 described working on a clinical dataset only to learn that releasing it was *"not possible because it's not consistent with the privacy laws in France."* To mitigate these risks, some participants adopted highly cautious practices, such as exclusively collecting royalty-free or Creative Commons images, and storing only image URLs to avoid any copyright violations. However, these strategies can result in dataset instability, as observed by P8, who faced issues with broken URLs.

Limited regulatory literacy. Insufficient understanding about navigating the law intensified concerns about legal risk. P8 described it as *"a big learning curve to understand what we were allowed to store*

https://doi.org/10.52202/079017-3103

and what we weren't." As a result, P8 consulted an intellectual property lawyer. However, depending on the other constraints dataset curators are under, such as discipline (Section 4.2) or organization (Section 4.3) level constraints, hiring legal counsel may be untenable.

4.5 Socio-Political Level

The *socio-political level* covers the shifting social and political contexts around fairness in which curators operate. These challenges can be conceptualized as thorny, fluid, and arguably insoluble.

Evolution and contestability of fairness. According to P3, fairness will *"always be up for debate,"* making it *"sort of impossible for there to be like a gold standard."* Fairness is subjectively perceived, influenced by individual contexts, experiences, and beliefs [129]. This subjectivity fuels ongoing scholarly debates [42, 89, 137]; it also fueled diverse perspectives among participants. As P30 pointed out, *"There are people from the audience who say that we have a good definition [of fairness], and there are some people who say that we have a terrible definition. And there's no way to make everyone happy."* The absence of a universally accepted definition complicated participants' efforts to operationalize fairness in dataset curation. Further, existing guidelines may not suit every notion of fairness, leading to divergent curation methodologies. As P14 highlighted, *"It's kind of like a philosophical question ... while the quantitative method says that fairness can be achieved, contrast it to qualitative that we are just trying to understand the experience here."* Beyond disagreements about what fairness means (or should mean), participants also noted that current definitions are not stable. As P16 put it, fairness *"should be a notion that is able to evolve within society, and certain forms of injustice that were not considered injustice[s] in the past now are ... there might be other evolution towards the future that we currently do not incorporate in our definition of fairness, and we need to account for that."* This perpetual evolution presents challenges for dataset curators. They must decide whether to regularly update datasets or retract them as definitions evolve. However, both approaches have limitations in addressing the continued use of previously released datasets [93, 112].

Social realities versus model realities. P8 described how the real world is different than "*what's experimentally valid and testable.*" Due to the complexity of the real world, certain groups inevitably remain underrepresented, misrepresented, or overlooked entirely despite best efforts. For example, P12 mentioned that while they wanted to collect images from underrepresented countries, data collector availability constrained their options (Section 3.3.1). Participants also questioned whether balanced representation was even the best approach. As P1 pointed out, *"The problem is when you actually apply such a model to the real world, the real world is imbalanced, right?"* This echoes the classic trade-off between fairness and accuracy in algorithmic fairness work [31]. Curators must wrestle not only with the impossible task of how to best account for every human experience in a dataset, but also whether or not they should be.

Power differentials. Power imbalances contribute to fairness issues during the curation process that are not visible in the dataset's composition. Participants noted how more elite institutions and companies dominate efforts to create fair datasets, largely owing to their access to resources (Section 4.3). Similar to findings from prior work [85], P21 described how most public datasets are not used, with the majority of *"the datasets that get used in ML research [being] created by a very, very small elite cadre of ... academic institutions that have close affiliations with top industry researchers."* Similarly, P16 felt it was problematic that the *"most important tools"* remain in the hands of a few companies, *"yet they are given the freedom to define what is fair, and their definition is used, and then the safeguards that do exist might not always align or ensure protection."*. Thinking on a geopolitical scale, P2 noted that *"the field of algorithmic fairness has been dominated by the Western perspective."* This imbalanced representation exacerbates other challenges previously outlined, including those at the implementation, disciplinary, and organizational levels.

Power differentials also permeate relationships between dataset curators and other stakeholders, including data subjects and workers [152]. For example, P6 described how curators have complete oversight over worker compensation: *"So many platforms don't actually ensure that you're fairly compensating workers. And it's really up to the individual researchers which is a crazy system that sets absolutely the wrong incentives."* P10 compared the impulse to collect data cost-effectively, at the expense of data subjects, as *"a particular kind of colonial impulse, like, this is just up for grabs."* Similarly, curator decisions have profound implications downstream. P22 described the difficulty of *"fighting"* clients who do not prioritize model performance on heavily under-resourced populations, given they are not central to business incentives: *"It's like, '99% of my customer[s] will be fine, why*

https://doi.org/10.52202/079017-3103

do I need to care about that last 1%?" Overall, dataset curation was seen as *"a very unfair process, no matter how you do it ... unless you're going to literally tackle society"* (P8).

5 Recommendations for Enabling Fair Dataset Curation

Finally, we highlight recommendations across the three dimensions of fairness for facilitating fair dataset curation. We focus on top-down efforts, reflecting the need for systemic changes rather than relying solely on individual contributions. See Appendix D for additional recommendations.

Composition. To better enable fair dataset *composition*, we encourage interventions for more flexible and robust data practices. For example, at the design phase (Section 3.2), flexible taxonomies can facilitate different operationalizations rather than forcing curators to use only one taxonomy (e.g., protected attributes can include self-reported and third-party labels). At the discipline level (Section 4.2), we advocate for more communication across academic communities. Papers published outside traditional ML venues (e.g., CHI, FAccT, CSCW) have provided guidance on data curation, such as annotation practices [30, 42, 83, 155] or considerations on taxonomies [7, 64, 84, 160].

Process. A change in the fair dataset curation *process* requires not only norm-setting within fairness communities, but also legal and policy interventions. For example, at the implementation phase (Section 3.3.3), participants were concerned about labor rights for data workers. As a discipline, we should have norms about compensating workers, at least at the local minimum wage, for their labor and support efforts to introduce policies that offer codified protection for data workers. Furthermore, at the regulatory level (Section 4.4), rather than expecting curators to develop legal expertise, we advocate for the creation of accessible resources on legal practices regarding dataset collection.

Release. We encourage interventions that allow for fairness post-*release*. For example, at the maintenance phase (Section 3.5), efforts to build tools and policies to enable better dataset traceability could alleviate concerns with dataset misuse. Additionally, at the organization level (Section 4.3), funding entities should invest in maintenance, rather than solely focusing on modeling research. Monetarily valuing long-term maintenance plans as research contributions may help shift perspectives about revision, maintenance, and use policies at the discipline level (Section 4.2).

6 Discussion and Conclusion

Our qualitative data reflects the experiences of our participants, and while we identify shared themes, these challenges may not be universally applicable or entirely representative. Despite efforts to recruit diverse dataset curators, our sample is skewed toward curators from North America and Europe, reflecting the Western-centric nature of ML and fairness research [85, 138]. Given the challenges raised around creating culturally contextualized datasets and navigating power dynamics across regions, future work should aim to include more geographically diverse voices, especially from the Global South, for deeper, more nuanced insights.

Despite these limitations, our study offers an important foundation for addressing the practical challenges in fair dataset curation. Through interviews with dataset curators engaged in fair dataset work, we developed a taxonomy of challenges across the dataset lifecycle and the broader fairness landscape. Participants navigated complex trade-offs between ideal fairness goals and practical constraints such as data availability, resources, and time. While we acknowledge the limitations of our methodology, taxonomizing these challenges is a crucial first step in developing long-lasting solutions to support fair dataset curation.

Addressing these challenges will require effort not only from individual dataset curators but also systemic changes at organizational, disciplinary, and regulatory levels. Beyond providing dataset curators with grounded evidence to support their efforts in building fair datasets, our taxonomy offers stakeholders a pathway to address each challenge individually and opens avenues for further, more targeted investigations into the many challenges of curating fair datasets.

Acknowledgments and Disclosure of Funding

This work was funded by Sony Research.

https://doi.org/10.52202/079017-3103

References

[1] Announcing the NeurIPS Code of Ethics 2013; NeurIPS Blog — blog.neurips.cc. https://blog.neurips.cc/2023/04/20/announcing-the-neurips-code-of-ethics/. [Accessed 14-08-2023].

[2] Jerone T. A. Andrews, Przemyslaw Joniak, and Alice Xiang. A view from somewhere: Human-centric face representations. In *International Conference on Learning Representations (ICLR)*, 2023.

[3] Jerone T. A. Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. Ethical considerations for responsible data curation. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2023.

[4] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

[5] Yuki M Asano, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. Pass: An imagenet replacement for self-supervised pretraining without humans. In *f*, 2021.

[6] Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. Digiface-1m: 1 million digital face images for face recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[7] Teanna Barrett, Quanze Chen, and Amy Zhang. Skin deep: Investigating subjectivity in skin tone annotations for computer vision benchmark datasets. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

[8] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics (TACL)*, 6:587–604, December 2018. doi: 10.1162/tacl_a_00041.

[9] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

[10] Sebastian Benthall and Bruce D Haynes. Racial categories in machine learning. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2019.

[11] Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. Introducing the neurips 2021 paper checklist. 2021. URL https://blog.neurips.cc/2021/03/26/introducing-the-neurips-2021-paper-checklist/.

[12] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[13] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

[14] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The values encoded in machine learning research. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[15] Abeba Birhane, Vinay Prabhu, Sang Han, and Vishnu Naresh Boddeti. On hate scaling laws for data-swamps. *arXiv preprint arXiv:2306.13141*, 2023.

[16] Abeba Birhane, Sanghyun Han, Vishnu Boddeti, Sasha Luccioni, et al. Into the laion's den: Investigating hate in multimodal datasets. *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2024.

https://doi.org/10.52202/079017-3103

[17] Borhane Blili-Hamelin and Leif Hancox-Li. Making intelligence: Ethical values in iq and ml benchmarks. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

[18] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of “bias” in nlp. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[19] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[20] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, January 2006. ISSN 1478-0887. doi: 10.1191/1478088706qp063oa.

[21] Urie Bronfenbrenner et al. Ecological models of human development. *International encyclopedia of education*, 3(2):37–43, 1994.

[22] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2018.

[23] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[24] Cristian S Calude and Giuseppe Longo. The deluge of spurious correlations in big data. *Foundations of science*, 22:595–612, 2017.

[25] Scott Allen Cambo and Darren Gergle. Model positionality and computational reflexivity: Promoting reflexivity in data science. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2022.

[26] Alan Chan, Chinasa T Okolo, Zachary Terner, and Angelina Wang. The limits of global inclusion in AI development. In *AAAI Conference on Artificial Intelligence - Workshop on Reframing Diversity in AI*, 2021.

[27] Alan Chan, Herbie Bradley, and Nitarshan Rajkumar. Reclaiming the digital commons: A public data trust for training data. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2023.

[28] Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[29] Kasia S Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. The dataset nutrition label (2nd gen): Leveraging context to mitigate harms in artificial intelligence. *arXiv preprint arXiv:2201.03954*, 2022.

[30] Katherine M Collins, Umang Bhatt, and Adrian Weller. Eliciting and learning with soft labels from every annotator. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2022.

[31] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

[32] Council of Europe. Inclusion and anti-discrimination: AI & discrimination. https://www.coe.int/en/web/inclusion-and-antidiscrimination/ai-and-discrimination, n.d. Accessed November 24, 2022.

[33] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics (TACL)*, 10:92–110, 2022.

[34] Sebastião Vieira de Freitas Netto, Marcos Felipe Falcão Sobral, Ana Regina Bezerra Ribeiro, and Gleibson Robert da Luz Soares. Concepts and forms of greenwashing: A systematic review. *Environmental Sciences Europe*, 32:1–12, 2020.

[35] Paul De Hert and Vagelis Papakonstantinou. The new general data protection regulation: Still a sound system for the protection of individuals? *Computer law & security review*, 32(2): 179–194, 2016.

[36] Taher Dehkharghanian, Azam Asilian Bidgoli, Abtin Riasatian, Pooria Mazaheri, Clinton JV Campbell, Liron Pantanowitz, HR Tizhoosh, and Shahryar Rahnamayan. Biased data, biased AI: deep networks predict the acquisition site of tcga images. *Diagnostic pathology*, 18(1):67, 2023.

[37] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[38] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. Exploring how machine learning practitioners (try to) use fairness toolkits. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[39] Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. Understanding practices, challenges, and opportunities for user-engaged algorithm auditing in industry practice. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

[40] Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. Investigating practices and opportunities for cross-functional collaboration around AI fairness in industry practice. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

[41] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399*, 2020.

[42] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[43] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

[44] Yi Ding, Jacob You, Tonja-Katrin Machulla, Jennifer Jacobs, Pradeep Sen, and Tobias Höllerer. Impact of annotator demographics on sentiment dataset labeling. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 2022.

[45] Ravit Dotan and Smitha Milli. Value-laden disciplinary shifts in machine learning. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

[46] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6):2074–2152, 2022.

[47] Ruth R Faden and Tom L Beauchamp. *A history and theory of informed consent*. Oxford University Press, 1986.

[48] Shaoyang Fan, Pinar Barlas, Evgenia Christoforou, Jahna Otterbacher, Shazia Sadiq, and Gianluca Demartini. Socio-economic diversity in human annotations. In *ACM Web Science Conference (WebSci)*, 2022.

[49] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79, 2024.

[50] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. Uncurated image-text datasets: Shedding light on demographic bias. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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

[52] R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

[53] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[54] Apoorva Gondimalla, Varshinee Sreekanth, Govind Joshi, Whitney Nelson, Eunsol Choi, Stephen C Slota, Sherri R Greenberg, Kenneth R Fleischmann, and Min Kyung Lee. Aligning data with the goals of an organization and its workers: Designing data labeling for social service case notes. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2024.

[55] Google PAIR. Google pair. people + ai guidebook. https://pair.withgoogle.com/guidebook, 2019. Accessed February 1, 2023.

[56] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2021.

[57] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2022.

[58] Colin M Gray, Ike Obi, Shruthi Sai Chivukula, Ziqing Li, Thomas V Carlock, Matthew S Will, Anne C Pivonka, Janna Johns, Brookley Rigsbee, Ambika R Menon, et al. Building an ethics-focused action plan: Roles, process moves, and trajectories. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2024.

[59] Mary L Gray and Siddharth Suri. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books, 2019.

[60] Barbara J Grosz, David Gray Grant, Kate Vredenburgh, Jeff Behrends, Lily Hu, Alison Simmons, and Jim Waldo. Embedded ethics: integrating ethics across cs education. *Communications of the ACM*, 62(8):54–61, 2019.

[61] Michael M. Grynbaum and Ryan Mac. The times sues openAI and microsoft over a.i. use of copyrighted work. *New York Times*, 2023. URL https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

[62] Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. Facet: Fairness in computer vision evaluation benchmark. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[63] Margot Hanley, Apoorv Khandelwal, Hadar Averbuch-Elor, Noah Snavely, and Helen Nissenbaum. An ethical highlighter for people-centric dataset creation. In *Advances in Neural Information Processing Systems (NeurIPS) - Navigating the Broader Impacts of AI Research Workshop*, 2020.

https://doi.org/10.52202/079017-3103

[64] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

[65] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. Casual conversations: A dataset for measuring fairness in ai. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021.

[66] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. Towards measuring fairness in ai: the casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021.

[67] Amy K Heger, Liz B Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–29, 2022.

[68] Courtney M Heldreth, Ellis P Monk, Alan T Clark, Candice Schumann, Xango Eyee, and Susanna Ricco. Which skin tone measures are the most inclusive? an investigation of skin tone measures for artificial intelligence. *ACM Journal on Responsible Computing*, 1(1):1–21, 2024.

[69] Alex Hern. Flickr faces complaints over 'offensive' auto-tagging for photos, May 2015. URL https://www.theguardian.com/technology/2015/may/20/flickr-complaints-offensive-auto-tagging-photos.

[70] Alex Hern. Google's solution to accidental algorithmic racism: Ban gorillas, January 2018. URL https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people.

[71] Alex Hern. Twitter apologises for 'racist' image-cropping algorithm, September 2020. URL https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm.

[72] Kashmir Hill and Aaron Krolik. How photos of your kids are powering surveillance technology. *New York Times*, 2019.

[73] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Gender and racial bias in visual question answering datasets. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[74] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. Dialect prejudice predicts AI decisions about people's character, employability, and criminality. *arXiv preprint arXiv:2403.00742*, 2024.

[75] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2019.

[76] Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. Data quality from crowdsourcing: a study of annotation selection criteria. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL) - Workshop on Active Learning for Natural Language Processing*, 2009.

[77] Andrew Hundt, William Agnew, Vicky Zeng, Severin Kacianka, and Matthew Gombolay. Robots enact malignant stereotypes. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[78] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

[79] IBM. Design for ai. https://www.ibm.com/design/ai, 2019. Accessed February 1, 2023.

https://doi.org/10.52202/079017-3103

[80] Eun Seo Jo and Timnit Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *ACM Conference on Fairness, Accountability and Transparency (FAccT)*, 2020.

[81] Shivani Kapania, Alex S Taylor, and Ding Wang. A hunt for the snark: Annotator diversity in data practices. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2023.

[82] Jared Katzman, Angelina Wang, Morgan Scheuerman, Su Lin Blodgett, Kristen Laird, Hanna Wallach, and Solon Barocas. Taxonomizing and measuring representational harms: A look at image tagging. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[83] Gunay Kazimzade and Milagros Miceli. Biased priorities, biased outcomes: three recommendations for ethics-oriented data annotation practices. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2020.

[84] Zaid Khan and Yun Fu. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

[85] Bernard Koch, Emily Denton, Alex Hanna, and Jacob Gates Foster. Reduced, reused and recycled: The life of a dataset in machine learning research. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

[86] Tzu-Sheng Kuo, Aaron Lee Halfaker, Zirui Cheng, Jiwoo Kim, Meng-Hsin Wu, Tongshuang Wu, Kenneth Holstein, and Haiyi Zhu. Wikibench: Community-driven data curation for AI evaluation on wikipedia. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2024.

[87] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision (IJCV)*, 128(7):1956–1981, 2020.

[88] Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori B Hashimoto. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2023.

[89] Sarah Lebovitz, Natalia Levina, and Hila Lifshitz-Assaf. Is AI ground truth really true? the dangers of training and evaluating AI tools based on experts' know-what. *MIS Q.*, 45, 2021.

[90] Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[91] Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, Li Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*, 4(8):669–677, 2022.

[92] Alexandra Luccioni and Joseph Viviano. What's in the box? An analysis of undesirable content in the common crawl corpus. In *International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (IJCNLP)*, 2021.

[93] Alexandra Sasha Luccioni, Frances Corry, Hamsini Sridharan, Mike Ananny, Jason Schultz, and Kate Crawford. A framework for deprecating datasets: Standardizing documentation, identification, and communication. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[94] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[95] Davide Luzzini, Federico Caniato, and Gianluca Spina. Designing vendor evaluation systems: An empirical analysis. *Journal of Purchasing and Supply Management*, 20(2):113–129, 2014.

[96] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. Assessing the fairness of ai systems: Ai practitioners' processes, challenges, and needs for support. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1): 1–26, 2022.

[97] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2020.

[98] Angelina McMillan-Major, Emily M Bender, and Batya Friedman. Data statements: From technical concept to community practice. *ACM Journal on Responsible Computing*, 1(1):1–17, 2024.

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

[100] Nicole Meister, Dora Zhao, Angelina Wang, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. Gender artifacts in visual datasets. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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

[102] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. Documenting computer vision datasets: An invitation to reflexive data practices. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

[103] Milagros Miceli, Julian Posada, and Tianling Yang. Studying up machine learning data: Why talk about bias when we mean power? *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–14, 2022.

[104] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application*, 8:141–163, 2021.

[105] Brent Daniel Mittelstadt and Luciano Floridi. The ethics of big data: current and foreseeable issues in biomedical contexts. *The ethics of biomedical big data*, pages 445–480, 2016.

[106] Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. Towards AI accountability infrastructure: Gaps and opportunities in AI audit tooling. *arXiv preprint arXiv:2402.17861*, 2024.

[107] Will Orr and Kate Crawford. Building better datasets: Seven recommendations for responsible design from dataset creators. *Journal of Data-centric Machine Learning Research*, 1:1–21, 2024.

[108] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. Pre-registration: Why and how. *Journal of Consumer Psychology*, 31(1):151–162, 2021.

[109] Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Andrews, Rebecca Bourke, Alice Xiang, and Allison Koenecke. Augmented datasheets for speech datasets and ethical decision-making. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

[110] Joon Sung Park, Danielle Bragg, Ece Kamar, and Meredith Ringel Morris. Designing an online infrastructure for collecting AI data from people with disabilities. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

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

[112] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

[113] Eugenia Politou, Efthimios Alepis, and Constantinos Patsakis. Forgetting personal data and revoking consent under the gdpr: Challenges and proposed solutions. *Journal of cybersecurity*, 4(1):tyy001, 2018.

[114] Bilal Porgali, Vítor Albiero, Jordan Ryda, Cristian Canton Ferrer, and Caner Hazirbas. The casual conversations v2 dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[115] Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Díaz. On releasing annotator-level labels and information in datasets. In *Linguistic Annotation Workshop (LAW) and Designing Meaning Representations (DMR) Workshop*, 2021.

[116] Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Alicia Parrish, Alex Taylor, Mark Díaz, and Ding Wang. A framework to assess (dis) agreement among diverse rater groups. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2024.

[117] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[118] Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. AI and the everything in the whole wide world benchmark. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

[119] Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. You can't sit with us: exclusionary pedagogy in ai ethics education. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

[120] Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[121] Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2024.

[122] William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2022.

[123] Norma RA Romm. Interdisciplinary practice as reflexivity. *Systemic Practice and Action Research*, 11:63–77, 1998.

[124] Marco Rondina, Antonio Vetrò, and Juan Carlos De Martin. Completeness of datasets documentation on ML/AI repositories: An empirical investigation. In *EPIA Conference on Artificial Intelligence*, 2023.

[125] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2021.

[126] Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. NLPositionality: Characterizing design biases of datasets and models. In *Annual Meeting of the Association for Computational Linguistics (ACL).*

[127] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2022.

[128] Morgan Klaus Scheuerman. In the walled garden: Challenges and opportunities for research on the practices of the AI tech industry. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.

[129] Morgan Klaus Scheuerman and Jed R. Brubaker. Products of positionality: How tech workers shape identity concepts in computer vision. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2024.

[130] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW1):1–35, 2020.

[131] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37, 2021.

[132] Morgan Klaus Scheuerman, Katy Weathington, Tarun Mugunthan, Emily Denton, and Casey Fiesler. From human to data to dataset: Mapping the traceability of human subjects in computer vision datasets. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–33, 2023.

[133] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *Advances in Neural Information Processing Systems (NeurIPS) - Workshop on Data-centric AI*, 2021.

[134] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2022.

[135] Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Rebecca Pantofaru. A step toward more inclusive people annotations for fairness. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2021.

[136] Carsten Schwemmer, Carly Knight, Emily D Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W Lockhart. Diagnosing gender bias in image recognition systems. *Socius*, 6:2378023120967171, 2020.

[137] Shilad Sen, Margaret E Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko, Samuel Naden, Jesse Russell, Zixiao Wang, and Brent Hecht. Turkers, scholars," arafat" and" peace" cultural communities and algorithmic gold standards. In *ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW)*, 2015.

[138] Ali Akbar Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. Weird faccts: How western, educated, industrialized, rich, and democratic is facct? In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

[139] Hong Shen, Leijie Wang, Wesley H Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. The model card authoring toolkit: Toward community-centered, deliberation-driven AI design. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[140] Andrew Smart, Ding Wang, Ellis Monk, Mark Díaz, Atoosa Kasirzadeh, Erin Van Liemt, and Sonja Schmer-Galunder. Discipline and label: A weird genealogy and social theory of data annotation. *arXiv preprint arXiv:2402.06811*, 2024.

[141] Brandon Abreu Smith, Miguel Farinha, Siobhan Mackenzie Hall, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. Balancing the picture: Debiasing vision-language datasets with synthetic contrast sets. In *Advances in Neural Information Processing Systems (NeurIPS) - Workshop on Synthetic Data Generation with Generative AI*, 2023.

[142] Hamed Taherdoost. Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research. *SSRN Electronic Journal*, April 2018. doi: 10.2139/ssrn.3205035.

[143] Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*, 2023.

[144] David Thiel. Identifying and Eliminating CSAM in Generative ML Training Data and Models. 2023. doi: 10.25740/kh752sm9123. URL https://purl.stanford.edu/kh752sm9123.

[145] William Thong, Przemyslaw Joniak, and Alice Xiang. Beyond skin tone: A multidimensional measure of apparent skin color. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[146] Nenad Tomasev, Kevin R McKee, Jackie Kay, and Shakir Mohamed. Fairness for unobserved characteristics: Insights from technological impacts on queer communities. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2021.

[147] Carlos Toxtli, Siddharth Suri, and Saiph Savage. Quantifying the invisible labor in crowd work. *Proceedings of the ACM on human-computer interaction*, 5(CSCW2):1–26, 2021.

[148] Joaquin Vanschoren and Serena Yeung. Announcing the neurips 2021 datasets and benchmarks track. *Medium*, 2021. URL https://neuripsconf.medium.com/announcing-the-neurips-2021-datasets-and-benchmarks-track-644e27c1e66c.

[149] Rama Adithya Varanasi and Nitesh Goyal. "It is currently hodgepodge": Examining AI/ML practitioners' challenges during co-production of responsible AI values. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2023.

[150] Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[151] Angelina Wang, Teresa Datta, and John P Dickerson. Strategies for increasing corporate responsible AI prioritization. *arXiv preprint arXiv:2405.03855*, 2024.

[152] Ding Wang, Shantanu Prabhat, and Nithya Sambasivan. Whose AI dream? In search of the aspiration in data annotation. In *ACM CHI Conference on Human Factors in Computing Systems*, 2022.

[153] Ge Wang, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. Challenges and opportunities in translating ethical AI principles into practice for children. *Nature Machine Intelligence*, pages 1–6, 2024.

[154] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[155] Mark E Whiting, Grant Hugh, and Michael S Bernstein. Fair work: Crowd work minimum wage with one line of code. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, volume 7, 2019.

[156] Cedric Deslandes Whitney and Justin Norman. Real risks of fake data: Synthetic data, diversity-washing and consent circumvention. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.

[157] Lauren Wilcox, Robin Brewer, and Fernando Diaz. AI consent futures: A case study on voice data collection with clinicians. *Proceedings of the ACM on Human-Computer Interaction*, 7 (CSCW2):1–30, 2023.

[158] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[159] World Health Organization and others. Ethics and governance of artificial intelligence for health: Who guidance. 2021.

[160] Alice Xiang. Mirror, mirror, on the wall, who's the fairest of them all? *Dædalus*, 153(1): 250–267, 2024.

[161] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

[162] Yu Yang, Aayush Gupta, Jianwei Feng, Prateek Singhal, Vivek Yadav, Yue Wu, Pradeep Natarajan, Varsha Hedau, and Jungseock Joo. Enhancing fairness in face detection in computer vision systems by demographic bias mitigation. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2022.

[163] Rui-Jie Yew and Alice Xiang. Regulating facial processing technologies: Tensions between legal and technical considerations in the application of illinois bipa. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[164] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[165] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

https://doi.org/10.52202/079017-3103

A Methods

A.1 Participant Recruitment

We interviewed 30 ML dataset collectors, refining our protocol through two pilot tests before recruitment. Out of 204 individuals contacted, we received 95 no responses, 51 declines, and 28 who did not meet the criteria. Recruitment concluded with 30 participants, reaching thematic saturation. As shown in Tables 1 and 2, participants represent diverse backgrounds and experiences, with a predominant presence from academia. Compensation consisted of a $75 Amazon gift card, or the equivalent in the participant's local currency.

Type	***Count***
Role	Graduate student (13), Post-Doctorate Researcher (6), Faculty (4), Researcher [Industry] (3), Researcher [Institute-based] (2), Other (2)
Setting	University (23), Industry (4), Academic Research Institute (2), Think Tank (1)
Modality	Language (16), Vision (9), Multi-modal (5), Tabular (3)
Location	Northern America (19), Southern Europe (3), Western Europe (3), Northern Europe (2), Latin American & the Caribbean (1), Western Africa (1), Southern Asia (1)

Table 1: Summary statistics of participant demographics. The locations are coded at the region level according to the United Nations geoscheme. Since some participants had experience collecting datasets in more than one modality, the counts in this row exceed 30.

A.2 Participant Anonymity

At the beginning of the interview, participants were asked to provide their informed consent. They were given the option to opt-out of the interview and also told they have the right to withdraw from the study at any time. Participants were also asked for permission to record the study over Zoom. For data protection, each interview was transcribed from the Zoom recording and identifying details—including but not limited to names, institutions, and dataset names—were redacted from the interview transcript before the coding process. To preserve participant anonymity, participant recruiting and interviews were conducted only by members of the research team from Arizona State University. Only the redacted interviews were shared with other members of the research team for analysis.

A.3 Thematic Analysis

We also provide additional details on our thematic analysis protocol. After establishing an initial codebook of themes, the research team (N=4) independently coded one of the interviews. We then reconvened and synchronously discussed how we coded the interviews and analyzed where we differed when applying codes. After this initial coding round, we again independently coded a second interview and repeated the same process of discussing any disagreements amongst the team before creating a finalized codebook. Only after reaching agreement on the definitions and applications of codes did we split up the remaining interviews amongst the team members.

To identify themes from the code, we had each member of the research team first generate themes, with supporting quotations, they observed in the interviews. Then, the research team met synchronously over four sessions to discuss and distill these observations into the higher-level themes discussed in the paper.

Finally, to ensure thorough consideration, we drew on a diverse range of expertise by following contemporary interdisciplinary practices [119, 123]. Our team consists of researchers, practitioners, and lawyers with backgrounds in HCI, ML, CV, algorithmic fairness, health sciences and policy, data visualization, and social and behavioral science. With varied ethnic, cultural, and gender backgrounds, we bring together extensive experience in dataset design, model training, and the development of ethical guidelines.

Participants		
Participant ID	***Organization Type***	***Dataset Focus***
P1	Academia	Language
P2	Academia	Language
P3	Academia	Language
P4	Academia	Other
P5	Academia	Language
P6	Industry	Language
P7	Industry	Language
P8	Academia	Multi-modal
P9	Academia	Language
P10	Academia	Language
P11	Academia	Vision
P12	Academia	Vision
P13	Industry	Vision
P14	University	Vision, Language, Other
P15	University	Vision
P16	Academia	Other
P17	Academia	Multi-modal
P18	Academia	Vision
P19	Academia	Other
P20	Academia	Vision
P21	Academia	Language
P22	Industry	Language, Vision
P23	Academia	Language, Multi-modal
P24	Academia	Language
P25	Academia	Language
P26	Academia	Vision
P27	Academia	Language
P28	Academia	Multi-modal
P29	Academia	Language
P30	Academia	Multi-modal

Table 2: A of participants we interviewed for this study. Organization type refers to whether participants were in academia or industry. Dataset focus refers to the type of data participants collected for their dataset. “Vision” refers to visual data such as images and/or videos. “Language” refers to natural language data, such as textual data and/or spoken language data. “Multi-modal” refers to datasets which included both vision and language data. “Other” refers to datasets that fall outside of this schema, such as tabular datasets.

https://doi.org/10.52202/079017-3103

A.4 Interview Protocol

We provide the protocol used to guide the semi-structured interview process conducted with participants. The interview questions were designed based on considerations around fair dataset curation that had been raised in the existing literature. Depending on the answers that the participants provided, the interviewers asked relevant follow-up questions. The questions are as follows:

- Please briefly describe your current role and responsibilities. What way(s) does your current role interface with dataset collection for machine learning?
- What is the role of machine learning in your organization?
- What types of data do you collect to train and/or evaluate ML algorithms? What are the sources of this data?
- Do you have any processes or are you currently developing any processes to ensure the fairness of data collected and used to train and/or evaluate ML algorithms?
- How does your organization define "fairness" of datasets? Do you have a formal, codified definition of fairness?
- How did your organization decide on the definition for fairness? Which factors influence this?
- How do you ensure collection of fair datasets to train and/or evaluate ML algorithms? Or fairness when repurposing collected datasets?
- Can you walk me through the process of making data collection and or data sets fair, as you do and experience it?
- Which best practices did you employ to ensure the collection or making of fair datasets?
- Which factors, in your experience, influence the making/collection of fair datasets?
- What challenges did you experience during the process of making/collecting datasets?
- How did you handle those challenges?
  - What were some workarounds/ solutions?
  - If you cannot recall any challenges, what about the process made it relatively smooth / why do you think there were not challenges?
  - Were any parts easier or more difficult than expected?
- Thinking back to the process of making or collecting datasets, I'd like you to tell me a story about a time when you experienced any trade-off related to fairness of the dataset during that process — meaning, you had to sacrifice something to increase the fairness of the dataset, or you sacrificed fairness to achieve something else.
- What challenges has your organization had in maintaining fairness in your datasets?
- Since collecting fair datasets, have you released any of these datasets?
- Thinking beyond your specific domain, what items should be included in more general guidelines for the creation and maintenance of fair datasets to train and/or evaluate ML algorithms? Are there any gaps in our current practices?
- Do you have any comments or other points to make? Is there anything we did not cover in the interview which you would like to talk about?
- Do you have any suggestions/advice about who we should talk to next?

B Background

Concerns over the disparate impacts or unjust outcomes associated with machine learning (ML) continue to persist [22, 28, 74, 94, 143]. One of the central concerns underscoring the pursuit of fair ML remains the datasets used to develop ML systems [12, 13, 22, 92, 111]. Yet obtaining fair and ethically-sourced datasets remains a challenge. Data is often perceived as the scourge of ML models and a source of for downstream biases [36, 52, 88, 164]. Here, we provide background on prior scholarship documenting the current issues with dataset curation, as well as work focused on improving those practices.

Issues with existing dataset curation practices. Poor training data can lead to representational harms [22, 50, 73], such as stereotyping [19, 23, 49, 136], spurious correlations [100, 154, 165], and poor performance or total erasure of certain populations [22, 158, 164]. Poor evaluation data means harmful model outcomes may be overlooked or missed, especially as they cascade into various (often unintended) domains [125]. Beyond data's impact on models directly, ML datasets are increasingly scrutinized for violating the ethical values of privacy and consent [3, 35, 105, 111, 113], reinforcing disputable social constructs [10, 17, 64, 84, 130], including highly offensive content [12, 15, 164], and exploiting vulnerable populations for both data and annotations [59, 140, 152].

Practices for collecting large-scale data, such as web scraping, have consistently failed to meet many legal standards at the local and national level, violating copyright laws [61], biometric laws [72, 163], and even including child exploitation content [16, 144]. The difficulty of authoring and maintaining a comprehensively "fair" dataset is exacerbated by differential definitions of fairness and how to measure it (or whether it can be measured at all) [4, 99, 146]. Current approaches to dataset documentation also obscure the inherently collaborative work that dataset authors must engage in and negotiate [101, 102].

Improving dataset curation practices. Given the vast and varied issues with ML datasets, there has been a extensive line of work focused on improving dataset collection practices. These efforts have evolved substantially beyond *ante hoc* calls for more transparent and robust documentation of existing datasets [8, 29, 42, 51, 98, 109, 117], such as datasheets, which often result in a *ante hoc* approach, thus failing to capture decisions and trade-offs which might have occurred prior to and during data collection. Thus, scholars are attempting to provide frameworks at different levels of granularity of considering the responsibility of dataset authors leading to frameworks or design guidelines for both *pre hoc* and *per hoc* dataset curation [3, 54, 58, 107, 110, 131, 157].

For example, at a higher level, Scheuerman et al. [131] proposed a value-centric framework that centers values like positional expertise and contextually-relevant annotations. Andrews et al. [3] released a comprehensive set of considerations for responsibly curating human-centric computer vision datasets, covering topics like consent, human diversity, and subject revocation. Recent work from Orr and Crawford [107] distilled seven recommendations from interviews with 18 dataset curators. Their work highlights high-level themes such as advocating for more dataset auditing, ensuring participant privacy, and encouraging more documentation. Scholars are also increasingly providing highly contextual and specific guidance for collecting data on certain subgroups and vulnerable populations, such as children [72, 153, 157], who are increasingly ending up in large web-scraped datasets [16, 144].

The scholarship focused on providing considerations and guidance for ethical dataset curation has been invaluable. However, how authors actually approach curating fair datasets is still opaque—especially given documented gaps between guidance and practice. Prior work has uncovered numerous barriers to incorporating fairness into practice [38, 67, 75, 96, 97, 151], including misalignments between available toolkits and product needs [38], organizational trade-offs that make auditing methods less effective [39, 106], and difficulty negotiating expectations across roles [40, 96]. Yet literature on the challenges to creating fair datasets currently lacks a holistic framing of fairness that involves not only the composition of datasets, but the practices of producing and maintaining them. Identifying the challenges currently facing dataset curators focused on creating fair datasets is crucial to enabling fairer dataset curation in both industry and academic settings.

C Additional Figures and Tables

To illustrate each of the challenges we identified, in Table 3, we provide an example quotation from our interviews. We similarly map out the overarching landscape of fairness challenges using a social ecological model [21] and provide detailed examples (see Figure 2 and Table 4).

Taxonomy of Challenges to Creating Fair Datasets

Phase		***Challenge(s)***	***Definition***	***Example***
Requirements		Scoping a dataset	Determining the size and scope of the dataset and its taxonomy while remaining true to fairness goals	*"If the face images are on a billion scale, there's no way we can identify each of this person in real world to reach out to them ask if they are okay with it."* (P18)
		Determining fairness definitions	Deciding which definition of fairness to adopt and which not to	*"Other work does not explicitly define fairness and that assumption can change the way you look at something. And so, being explicit about your intentions and about your working definitions and the inclusions and the exclusions of the scope of work, ... is getting attention now more."* (P2)
Design		Creating fair taxonomies	Establishing a system of classification that aligns with fairness definitions, despite classifications being inherently imperfect	*"There are power relationships within what the model represents and what it represents is hegemony and it represents rigidity and classification, and it does not represent all of that queerness."* (P27)
		Data availability in taxonomy design	Designing taxonomy with knowledge of data (un)availability in mind	*"The basic challenge is actually the availability of the data ... So, you need to set up boundaries in your research. I discuss the certain limitation on this issue in [our] paper. We don't have information on gender because the healthcare system does not adopt non-binary gender attributes."* (P1)
Implementation		Vendor transparency	Working with data vendors can hinder fairness efforts	*"Communication is difficult on platforms like [ANONYMIZED VENDOR PLATFORM]. It's a little bit easier on other platforms, but that has definitely been a blocker. It is like easy communication just doesn't exist."* (P6)
		Language barriers	Navigating language barriers between curators and data workers and/or data	*"So, we relied on our translators to come up with those sorts of decisions in terms in terms of Spanish. We did. Some of us didn't know Spanish, not natively and fluently."* (P3)
		Fair data labor	Ensuring data workers are treated and compensated fairly while navigating resource and regulatory constraints	*"Especially, because so many platforms don't actually ensure that you're fairly compensating workers. And it's really up to the individual researchers which is a crazy system that sets absolutely the wrong incentives."* (P6)
	Data Collection	Diverse data availability	Difficulties collecting sufficiently diverse or representative data during the data collection process	*"Because when I look at images of stoves on the Internet. . . .. The reason those images are up on the Internet is because they satisfy something other than someone's going to use to train a machine learning model. So, people put them up because they think it's something new or exciting, or something different in some way. So, a lot of stoves that are there for ImageNet are basically like product images from stores that they sell. So, the stoves always look very clean and new and things like that."* (P12)
		Data collector availability	Identifying data collectors who can collect underrepresented data	*"Because it very much depends on where we can get hold of people, right? And by that, I mean where it is up and has its workforce. And also how much money [does it cost] because it becomes more expensive as you're trying to get a lot of people from very, very small regions."* (P12)
	Data Annotation	Data annotator diversity and expertise	Identifying data annotators with situated expertise	*"The challenges in actually getting diverse annotators. Especially for like, let's say, there have been recent studies, or there has been one paper that talks about how the views of a person or a kind of how the views of the person affect the annotations that they do for hate speech. So, like people with certain social, certain political viewpoints, might annotate something as hate while others might not. And so how do you get diversity in your annotators? Because some of the attributes of their personal lives might even be illegal to ask about in a particular country ... And once you have it, how do you contextualize their annotations to their lived experiences?"* (P2)

(Continued on next page...)

Phase		***Challenge(s)***	***Definition***	***Example***
Evaluation	***Data Quality***	Gold standard paradigms	Models for assessing dataset quality can promote “unfairness”	*“There is no ground truth to that question.It can vary from a person’s lived experiences to the next. So, it is inherently a subjective question. So, we did not want to squash those annotations down to a majority quote.”* (P2)
	Data Utility	Lack of benchmarking datasets	Comparable benchmark datasets for which to evaluate new fair datasets are not available	*“There’s a lot of pressure to do well benchmark data sets. And so, there’s a risk of them being used overused because you need to show that you did well on the data set that everyone recognizes, even if it might not be the most appropriate.”* (P21)
		Evaluating immeasurable constructs	Proving dataset quality when fairness constructs are not quantifiable	*“I think quantitative methods almost always assume that fairness can be achieved in some way and they also often assume that there is already a robust definition of fairness that we’ve conceptualized and that we can use to test our systems. They also assume that fairness can be measured, can be evaluated and can be improved. And I think that all of this is a more positivist mindset.”* (P14)
		Spurious correlations	Accounting for and controlling spurious correlations	*“So, then, what happens is there is a geography bias which is being incurred in this data sets implicitly, which is not really explicit. I’m gonna train the models on this. The models just exaggerate the bias and when this model is deployed on, say, android phones, or software or laptops, or anything, the consumers are worldwide, right?”* (P18)
Maintenance		Unstable infrastructural ecosystems	Data in datasets may go missing or become deprecated, resulting in fairness issues	*“I think maintenance is more going to be a matter of making sure that when links become deprecated, we maintain the same principles of trying to find a diverse range of images to replace it.”* (P8)
		Dataset traceability mechanisms	Inability to track dataset usage or prevent misuses	*“ there have been cases where a researcher reached out to me and said, ‘Hey, I tried this with your data set. I’m getting like these confusing results. Can we talk?’ And then I find out they’re using it in a way that wasn’t intended.”* (P6)

Table 3: A table describing each of the challenges throughout the phases of the dataset lifecycle.

Phase		***Challenge(s)***	***Definition***	***Example***
Overarching	***Individual***	Individual contributor positionality	Every contributor to a dataset has their own positionality, including biases	*“Even the idea of the perspectivism ... Most obviously in my work is the research questions, and then the way it informs the direction of research, and even possibly down to the way we qualify how good a data set and how interesting a dataset is!”* (P24)
	Discipline	Recognition for fair dataset work	Datasets are undervalued in machine learning	*“The right way is also rewarding people for doing it the right way right like the idea that you should be able to publish a data set and that be a valuable contribution, because in machine learning, it’s an extremely valuable contribution. And yet it’s not something that is valued.”* (P21)
		Awareness of existing resources and guidelines	Curators are unaware of existing resources for fair datasets or how to apply them	*“If I recall like, I don’t like remember any explicit guidelines that I’ve stumbled through for fair data set collection. Honestly!”* (P29)
		Responsibility for fairness	Those with an awareness about fairness issues feel a responsibility to do fairness work, while those who are not aware are excused from fairness work	*“In general, I will say the motivation is having fairness because you have this responsibility of understanding and improving transparency and improving general oversight on what we deploy.”* (P28)
	Organization	Lack of resources	Fair dataset work is not given resources in the form of time, money, personnel, tools, etc.	*“Research is driven by building bigger and bigger models and that is increasingly, punitively expensive. From a resource standpoint, from a money standpoint, from an environmental standpoint. And data has, in general, been undervalued in machine learning.* (P21)
		Ethics washing	Fairness is treated as a marketing tactic rather than necessary	*“One of the big reasons a lot of big companies do responsible AI shenanigans is for marketing ... then a new shiny thing comes down the road and then they join that instead.”* (P22)
	Regulatory	Differing legal practices	Laws, regulations, and policies governing fairness differ by context	*“Laws in America or laws in Europe ... might not be directly applicable to a country like [in South Asia] that has very different societal situation.”* (P2)
		Limited regulatory literacy	Dataset curators are not equipped to understand the regulatory landscape	*“It was a big learning curve to understand what we were allowed to store and what we weren’t in terms of the legal sense. So, it was a challenge to us personally, because we didn’t have experience. So, we consulted with an IP lawyer to get insight into that, but really just making sure that what we were presenting and storing was legal.”* (P8)
	Socio-Political	Evolution and contestability of fairness	Perspectives and policies on fairness evolve over time, constantly evolving the landscape of what a fair dataset is	*“The question [of] whether fairness should be defined through a singular definition within a specific instrument is tricky because . . . It should be a notion that is able to evolve within society, and certain forms of injustice that were not considered injustice in in the past now are. If we looked at the position of members of the LGBTQIA+ community, it was criminalized. Racism was also accepted. Now we clearly say it’s not so. There might be other evolution towards the future that we currently do not incorporate in our definition of fairness, and we need to account for that.”* (P16)
		Social realities versus model realities	The “real” world is inherently complex and multi-faceted, but machine learning datasets (and downstream models) require more simplistic approaches	*“Benchmark[s] which are made for fairness ... still have a very structured, kind of neutral way of portraying things like race or gender that don’t actually engage with the socio-historical meaning of that.* (P11)
		Power differentials	Different institutions (e.g., industry vs. academia; elite universities vs. R3s), actors (e.g., data curators vs data workers), and regions (e.g., the West vs the Rest) have different power to shape fairness concepts and practices	*“I mean you hear of data coming from these marginalized regions but then this is centralization process with one institution getting credit for it and the reputations of other countries not sharing that credits and some not reaping benefits of it. So, there’s especially in countries that are poorer, there’s then less incentive for them to actually contribute to datasets.”* (P8)

Table 4: A table describing each of the challenges overarching the broader landscape of fairness

D Detailed Recommendations for Enabling Fair Dataset Curation

Recommendations are aimed at diverse stakeholders influencing fair dataset curation, including—but not limited to—individual contributors, academic institutions and venues, industrial organizations, policymakers, and the affected public. Unlike in the main body of the text, where we describe the challenges with the dataset lifecycle first and the challenges with the overarching landscape of fairness second, here, we present considerations with the overarching landscape foremost. We also begin with the highest level of the dataset landscape, the *socio-political level*, rather than the lowest, the *individual level*. Our goal is to underscore how top-down changes can have broader impacts downstream on individual data curators and the dataset lifecycle. We advocate for more systemic changes rather than placing the onus of fairness onto individuals. The following recommendations in Appendices D.1 and D.2 are examples. We imagine there are many more interventions which would be effective in improving fair dataset curation.

D.1 Recommendations Overarching the Broader Landscape of Fairness

D.1.1 Socio-Political Level

Evolution and contestability of fairness. As conceptualizations of fairness inevitably change, curators should aim to keep datasets up-to-date. For example, we recommend that curators revise and amend datasets to comply with new conceptualizations of fairness. For example, Yang et al. [162] obfuscated faces in ImageNet after release as an effort to mitigate concerns about data subject privacy. Furthermore, when datasets cannot be aligned with new standards, norms, laws, or policies surrounding fairness, they ought to be deprecated and no longer used. Curators can refer to Luccioni et al. [93]'s framework for retracting and deprecating datasets to better understand this process.

We also recommend that data curators clearly document the decisions that were made about contextually and temporally relevant definitions of fairness. Thus, even if the original curator cannot afford to update the dataset, others can continue to maintain its documentation pointing toward new research showing the issues with past fairness operationalizations.

Social realities versus model realities. We recommend dataset curators engage with affected communities to understand the needs and potential impacts datasets and downstream models have on the lives of real people. This includes situating data curation decisions in the experiences and perspectives of affected communities. For example, Kuo et al. [86] introduced WikiBench, a system for creating community-driven evaluation dataset on Wikipedia. Using WikiBench, community members can work together to select, label, and discuss instances for an evaluation dataset. Adopting a more participatory and bottom-up approach allows dataset curators to ensure that they are capturing the concepts most relevant to impacted communities.

Power differentials. First, we recommend incentivizing dataset curation with fairness perspectives outside of the West and Global North [26, 138]. For program committees or conference chairs, potential actions can include having special tracks for these datasets or offering travel scholarship for researchers to the conference. We advocate for approaches that empower researchers from the Global South to create their own datasets.

Another power differential participants discussed was between researchers and data subjects or annotators. To address this, we urge curators to center the agency and consent of data subjects as well as the expertise of data workers. Rather than treating data workers as "ghost workers" [59], curators should ensure that data workers are meaningfully involved throughout the data curation process and thought of as contributors rather than solely as a labor source.

D.1.2 Regulatory Level

To help minimize legal risk, our first recommendation is for the the discipline to develop ethical review processes to assess for potential legal implications of dataset collection. Venues, such as NeurIPS, have instituted impact statements and paper checklists for submitted works [11]. We recommend that this reviews extends to include legal risks. We advocate for this discipline-wide approach as it can defray potential concerns regarding resource mismatches when it comes to consulting legal counsel. By developing a standardized procedure for legal compliance across datasets, it also reduces the burden on individual curators who may have limited regulatory literacy.

Nonetheless, we still recommend that individual curators pay particular care when collecting data containing people or about people. One alternative here, such as that taken by Asano et al. [5] and Ramaswamy et al. [121], is to ensure there are no people in the dataset. Of course, there is still a need for human-centric datasets. In this case, we urge curators to recognize that using royalty-free or Creative Commons licenses does not absolve the data of potential ethical or legal issues regarding privacy or consent [3]. Instead, curators ought to obtain informed consent from data subjects following well-established protocols from human subjects research [47].

D.1.3 Organization Level

Ethics washing. Participants were disillusioned by organizations that treated fairness as "lip service" and engaging in the practices of ethics-washing. Echoing Wang et al. [151], we recommend institutional efforts to keep organizations liable for the ethical AI promises that they make. In addition to relying on individual contributions from researchers and journalists, having watchdog organizations monitor for ethics-washing. This recommendation draws from existing practices of monitoring companies for "greenwashing", or manipulative promises from companies that they are engaging in environmentally friendly actions [34].

D.1.4 Discipline Level

Lack of recognition and incentives. Since 2021, there have been efforts to introduce more dataset-focused tracks, such as the Datasets and Benchmarks track at NeurIPS [1] or the Journal of Data-Centric Machine Learning Research (DMLR). We recommend building on this trend and encouraging more dataset-focused tracks, including some that have specific sub-areas dedicated to fairness-oriented datasets. This can help address the lack of recognition and incentives for fair dataset work.

Responsibility for fairness. Fairness-oriented changes ought to be widely adopted amongst ML dataset creators, not only those who may be more "fairness' or "justice" oriented. To encourage this shift, we recommend adopting more educational training on these subjects. Universities can include fairness and ethics into computer science courses. An example of this are the Embedded EthiCS programs at universities which encourage students to think critically about the technology they are learning about in their computer science courses [60]. Beyond university courses, AI ethics review processes can also mandate certifications that researchers must complete prior to getting approval similar to the trainings that researchers must complete before receiving IRB approval.

D.1.5 Individual Level

Individual contributor positionality. Contributor biases are inevitable. Our recommendations here focus not on removing all individual biases but rather on encouraging curators to get multiple perspectives and reflect on what biases they may be bringing prior to data collection. One recommendation is to institute a "pre-registration" system similar to what social scientists have in place [108]. Pre-registration requires social scientists to publicly state their hypotheses, methods, data collection process, and analysis plans prior to beginning their experiment. Filling out a pre-registration prior to data collection could encourage curators to think through design biases and justify the choices they have made in a transparent and standardized manner.

D.2 Recommendations During the Dataset Lifecycle

D.2.1 Requirements

Determining fairness definitions. Participants considered fairness to be highly contextual. To ensure that the definitions of fairness match those of impacted communities, we recommend that curators solicit and incorporate community feedback into the design and evaluation of fairness criteria [18]. This can help ensure that the dataset reflects the needs and values of diverse populations. As an example for how this can be done, curators can look to works such as Shen et al. [139] which aimed to involve community members in deliberative processes for defining AI systems. Similar participatory processes can be adapted for determining fairness definitions in datasets.

D.2.2 Design

Creating fair taxonomies. When designing a label taxonomy, we encourage curators to evaluate trade-offs associated with adopting coarser categories, such as loss of granularity versus feasibility and practicality. Data curators should report both their ideal data collection scenario and the actual approach taken. This information is useful not only from a transparency perspective but also for other researchers who may face similar issues in the future.

In addition, these taxonomies should be designed with scalability in mind. Curators should make provisions to ensure the taxonomy is flexible enough to incorporate new data if collected. For example, the OpenImages dataset [87] has had several new versions and additions since its initial release, including Schumann et al. [135]'s new demographic annotations which are aimed to aid with fairness research.

D.2.3 Implementation

Vendor transparency. As third-party vendors offer an alternative path for data collection, we recommend curators prioritize transparency both in negotiations with these vendors and when reporting their results. In negotiations with data vendors, curators should prioritize transparency clauses in the contract. For example, curators should advocate for transparency in data worker identities and compensation handling. This can help to ensure that they have access to necessary information for evaluating dataset fairness. During the collection process, data vendors should be held accountable for transparency practices through regular monitoring and evaluation. This could involve conducting audits or assessments to ensure compliance with transparency agreements and guidelines.

To reduce the burden on individual curators, there should be a discipline-wide effort to evaluate and benchmark data vendors based on transparency and ethical data collection practices. From management studies, there is a line of work on vendor evaluation systems and vendor scorecards that can be adapted for third-party data curation services [95].

Language barriers. When faced with language barriers, the data curation team should ensure that they have members who have an understanding of the data collection project's context, goals, and data requirements such that they can provide more contextually appropriate translations. If this is not possible, we recommend establishing partnerships with local community organizations or language schools to access language resources at reduced costs.

Fair data labor. To ensure fair data labor practices, we recommend curators create clear guidelines and protocols for hiring, training, and evaluating data workers to promote fairness and prevent exploitation. Following prior works [3, 26, 147, 155], we also advocate for transparent and equitable compensation structures for data workers. When possible, curators should provide opportunities for professional development and advancement for data workers.

Diverse data availability. Curators should consider using alternative data sources beyond web data, such as community-driven platforms or public repositories, to supplement dataset diversity. To support this, organizations should invest in creating public data trusts [27] or data consortia [80] as an alternative source for large-scale data.

Data collector availability. To address a lack of data collector availability, we recommend curators form partnerships with universities, organizations (e.g., NGOs, non-profits), or community groups, operating in underrepresented regions. These partnerships can help the recruitment of data collectors from the target regions, leveraging existing networks and/or local expertise to overcome challenges. For example, Rojas et al. [122] partnered with Gapminder and individual photographers to collect geographically diverse images for the DollarStreet dataset.

Data annotator diversity and expertise. When recruiting data annotators, curators should research and understand which personal attributes are legally protected and cannot be asked about during the hiring process. Further, they should be cognizant of cultural nuances. For example, disclosing sexuality can potentially endanger workers. Thus, rather than directly asking sensitive personal attributes, curators can utilize alternative methods for assessing annotator qualifications and suitability for the project. This is especially helpful when annotators may not want to disclose certain attributes.

Finally, curators should offer training and resources to annotators to help them understand the cultural significance of the data they are annotating.

D.2.4 Evaluation

Gold standard paradigms. Dataset curators can adopt evaluation methods that embrace diverse perspectives rather than only using consensus-based methods, which may only showcase the viewpoint of the majority. Works from both machine learning [33, 90] and human-computer interaction [56, 57] have encourage using a multiplicity of annotations, which can showcase disagreement, rather than using majority voting. For example, a curator may capture a diversity of annotations from each annotator, with qualitative explanations as to why the annotator chose each label. Prior works [116, 150] have also provided frameworks for quantifying disagreement across diverse groups of annotators that can be used as an alternative measure to consensus-based approaches.

Evaluating immeasurable constructs. When it comes to evaluating immeasurable constructs, curators can supplement quantitative metrics with qualitative approaches. This could include interviews with data workers to better understand their point of view and reveal any potential biases or ethical issues that arose during the collection process. Furthermore, as Miceli et al. [102] advocate for in their work, there should be more reflexivity in the data collection process. Concretely, refereed publications should require more critical reflection on the limitations and trade-offs of the dataset by the curators.

D.2.5 Maintenance

Unstable infrastructural ecosystems. To manage unstable infrastructural ecosystems, we recommend building standardized methods for checking the availability of data sources and creating protocols to replace instances if they have been deprecated. For example, P8 mentioned developing automated scripts that would periodically check whether their dataset instances were still available. Rather than waiting for dataset users to notify curators that certain instances are no longer available, this allows for proactive maintenance.

Going hand-in-hand with this, once curators are aware that certain instances are deprecated, they should have a plan for replacing them in a way that maintains the overall composition of the dataset. This can be challenging, especially for datasets where compositional fairness is prioritized. We recommend that dataset curators create a protocol for identifying alternative data sources that match the distribution and characteristics of the original dataset at the design phase. For example, dataset curators can keep a portion of collected data as "backup" that they can use to replace instances that are deprecated or removed.

Dataset traceability mechanisms. One challenge curators faced was tracing how their dataset was used after release. Often they relied on citation metrics as a proxy; however, it was difficult to disambiguate whether the citation meant the authors were using the dataset or referring to concepts in the paper. As an alternative, we recommend curators require users to register or authenticate their identity before accessing datasets, enabling better tracking and accountability. For example, ImageNet [37] requires users to sign in to their platform before downloading data. Another option can be to use permanent digital identifiers, such as DOIs, which is already a standard for some journals such as *Nature*.[4] Similarly, curators can use centralized data repositories (e.g., Hugging Face, Kaggle, Zenodo, Harvard Dataverse, Mendeley data).

E Broader Impacts

Our work focuses on understanding the challenges with fair dataset collection by conducting on-the-ground interviews with dataset curators. We provide a taxonomy of challenges that curators face throughout the dataset lifecycle and an exploration into the broader landscape of challenges curators face. For dataset curators, this work provides valuable insight into the nuance and trade-offs related to dataset creation that may not appear in publications. By formalizing this otherwise tacit knowledge, we hope to make the process of fair dataset collection more accessible for curators. More broadly, we intend for our work to have an impact on machine learning as a discipline. We seek

[4] https://www.nature.com/ncomms/editorial-policies/reporting-standards

https://doi.org/10.52202/079017-3103

to emphasize the importance of dataset curators' labor, which often is undervalued [124, 125, 131]. Furthermore, we provide an extensive set of recommendations that can be implemented by either individual contributors or from the top-down. By using these recommendations and the challenges we have surfaced, we hope to help facilitate better fair dataset curation practices within the machine learning community.

F Author Contributions

D.Z. and J.T.A.A. conceived of the idea for the project in this paper. D.Z., M.S., P.C., J.T.A.A., G.P., S.W., and K.P. were involved in discussing the themes of the overarching project. J.T.A.A. and A.X. acquired the financial support for the project. J.T.A.A., S.W., K.P, and A.X. provided oversight and leadership to the research team working on the project.

D.Z., J.T.A.A., P.C., S.W., and K.P. were involved in developing the interview protocol. P.C., S.W., and K.P. recruited participants and conducted the interviews. P.C. transcribed and redacted the interviews.

D.Z., M.S., J.T.A.A., P.C., G.P., S.W., and K.P. were involved in developing the thematic codebook from the interviews. D.Z., M.S., J.T.A.A., and K.P. conducted analysis of the interviews, applying themes from the codebook.

D.Z. and M.S. drafted the manuscript. J.T.A.A., G.P., S.W., and K.P. reviewed and commented on the manuscript. M.S., P.C., and D.Z. created the figures and tables in the manuscript.

https://doi.org/10.52202/079017-3103