A hunt for the Snark: Annotator Diversity in Data Practices

ANONYMOUS AUTHOR(S)

Diversity in datasets is a key component to building responsible AI/ML. Despite this recognition, we know little about the diversity among the annotators involved in data production. We investigated the approaches to annotator diversity through a survey with 49 AI/ML practitioners and 16 semi-structured interviews. While practitioners described nuanced understandings of annotator diversity, they rarely designed their practices to accommodate this diversity. The lack of action was explained through operational barriers, from the lack of visibility in the annotator hiring process to the conceptual difficulty in incorporating worker diversity. We argue these operational barriers and the widespread resistance to accommodating annotator diversity surface a prevailing logic in data practices—where neutrality, objectivity and 'representationalist thinking' dominate. By understanding this logic to be part of a "regime of existence", we explore alternative ways of accounting for annotator subjectivity and diversity in data practices.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**.

Additional Key Words and Phrases: data a

ACM Reference Format:

Anonymous Author(s). 2018. A hunt for the Snark: Annotator Diversity in Data Practices. In . ACM, New York, NY, USA, 25 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Machine learning depends on data to 'learn', that is, to uncover patterns, to classify, or predict potential outcomes. The process to produce human-annotated datasets is often an integral part of this AI/ML practice [38, 83]. A growing body of work is seeking to examine the diversity within datasets [28], particularly within the spaces of responsible AI research, including in discussions of fairness [9, 85] and harm mitigation [89]. Despite this effort to place *diversity* centre stage, we know little about the diversity among the annotators involved in producing datasets.

This paper reports on research to understand how diversity among annotators is conceptualised, considered and approached by data requesters and practitioners who build models with annotated data. We report results from a combination of forty-nine survey responses and sixteen in-depth, semi-structured interviews with ML practitioners (engineers, researchers, project/product managers and platform providers) whose work relied on annotated data sets. Through an analysis of these results, we seek to understand the barriers that limit a consideration of diversity in data practices.

Many of the practitioners we surveyed and interviewed did, in fact, have a sophisticated understanding of annotators' subjective decision-making, and acknowledged the risks involved in building datasets treated ostensibly as standardised across diverse groups. They showed a sensitivity to people's subjective views and how these were likely based on individual experiences and cultural norms. However, this thinking was repeatedly pushed aside when confronted with the practical work of designing annotation tasks, collecting labelled data, and developing and refining AI/ML models. What we came to see was something much deeper revolving around data annotation and production of datasets.

Practical explanations were used to justify this seeming contradiction. The overhead, costs and complexity of annotation tasks had to be minimised, and accordingly the capture of diversity among annotators was put off till later. Moreover, there were tasks that were considered less prone to subjective decision making, for example, where there would be definitive answers or where expertise overruled subjective opinion.

Examining diversity in practice revealed the logics that underlie the work across developing annotated datasets and building AI/ML models, enacting a particular idea of annotator diversity. Whether it was seen as a source of weakness in the dataset, diversity was approached as a means to achieve objectivity, while individual subjectivity was treated as an obstacle to neutral representation. That is, diversity was seen as a question of whether an objective or neutral state could be fairly and accurately represented in a dataset. Davis *et al.* [21] reveal a prevailing algorithmic idealism in algorithmic thinking which assumes a meritocratic society in which demographic disparities can be neutralised. Our research echoes this view. We found a belief that imagines a world in which data can be neutrally captured and represented, and where this neutrality is achieved by mitigating "fallible human biases on the one hand, and imperfect statistical procedures, on the other" [21, p. 2].

The work we present is informed by a distinct theoretical framing. In particular, our later discussion of the practices of data annotation builds on the idea of "regimes of existence" introduced by the sociologist Geneviève Teil [91]. Teil has used the term to examine 'terroir' in wine-making practices; as Teil recounts, terroir is a quality of wine that is assessed through aroma and taste, and discussion between colleagues. For "terroir vintners" in particular, these experiences iteratively shape the many stages of wine production. Through terroir, Teil's interest is in the tensions between objectivity and subjectivity and the ways in which the former, objectivity, "ascribes a specific regime of existence" [91, p. 480]. This is a regime where there is a "*'data' that can be discovered and whose existence unfolds independently, including from the people who live around, with or alongside [things]*" (ibid.). We find that Teil's [91] work offers us a starting point to further examine the logics which prevail in ML data practices. We argue that what is needed is not only greater diversity of annotators when producing datasets (although this would certainly help), but a shift in the orientation to data practices. Required, in short, is a change to the regime, a change to the world in which the prevailing logic discounts subjectivity.

Our paper makes three main contributions, we:

1. provide an empirical account of the status quo of annotator diversity in data practices.
2. examine and elaborate on an underlying logic that prevails in data practices that side-lines diversity-related considerations.
3. propose shifts to rethink diversity in data practices, inspired by justice-oriented and intersectional scholarship, and invite a nuanced approach to annotator diversity.

2 RELATED WORK

Below we situate our work in a body of related research, starting with the practices and discretionary choices which shape the work of data. We engage with research on the subjective interpretation involved in creating datasets, and particularly the human annotations. We then draw attention to the discourses on diversity in machine learning systems. Finally, we draw upon the concepts of representationalist thinking and regimes of existence to enable us to critically examine the overarching logics in data annotation.

2.1 The Practices of Data

Substantial work has been done in HCI/CSCW and STS to establish the ways in which data is shaped and defined through the contexts of production, as it is through the context of use or exchange [31, 90, 94]. The practices of data production are imbued with value judgements– of what is counted, what is excluded and how things are made into measurable entities– 'shaping human-computer interaction even before data reaches the computer' [76].

Past research in critical data studies has drawn particular attention to the sites of human intervention and discretionary choices which shape the work of data [63, 66, 67, 75]. Passi and Jackson [74] proposed the concept of data vision as the interplay between formal abstraction and discretion which is central to making datasets work with the chosen algorithms. These discretionary choices are situated in and range from formulating the task, choosing the training data, selecting the dataset characteristics, establishing taxonomies, post-processing the data, choosing which errors are acceptable, and communicating outcomes to stakeholders [68, 73, 76]. Muller *et al.* [68] describe how as part of this decision-making, practitioners often engaged in compromises and trade-offs when considering the quality of labels alongside the available resources. To provide a framework to consider these decisions, Cambo and Gergle [12] introduce the concepts of reflexivity and positionality from qualitative research for data science praxis to make the discretionary decisions less biased and transparent.

Others in the community have examined the work practices, experiences, and backgrounds of individuals involved in data annotation work [8, 65, 95]. More recently there is a shift towards the emerging actors that provide data annotation as a service to commissioning clients [50, 62]– who are making the work of annotation more structured and organised. Wang *et al.* [95] investigated the work of annotators involved within organised employment structures in India. They argue that data annotation is a systematic exercise of organisational structure and power and this hierarchy and control not only impacted the annotators' experiences but also their interpretations of the data [95]. They highlight how the simplistic definition of 'data quality' as accuracy rate leaves little space for the annotators' expertise, knowledge and experiences. By 'studying up' AI developers, Sambasivan and Veeraraghavan [81] demonstrated that even in settings where the data production is reliant on expert fieldworkers (*e.g.,* farmers and radiologists) and their highly situated and contextualised knowledge, practitioners still reduced field workers as data collectors, and attributed poor data quality to their work practices. Fieldworkers who produce specialised data for ML development should be seen as domain experts to have their knowledge, experience and contributions acknowledged rather than dismissed [43, 44, 60].

We contribute to this literature by focusing on the practices of incorporating annotator diversity to further our understanding of data production into the creation of ML models. Through unpacking the annotation practices (as part of the machine learning process), we show in our findings how annotator diversity does (not) fit in the current ML workflow and practice.

2.2 Subjective interpretation of data

Previous work has argued that the construction of datasets and the annotations, if any, associated with them, is about sensemaking– of "fitting the data into a frame and fitting a frame around the data" [52]. Feinberg [29] described the ways in which design figures into the infrastructure development, collection and aggregation of data [29]. Despite the use of protocols and vocabularies to enforce shared meaning and consistency, there is still a lot of room for situational contingencies and interpretive flexibility.

There is a well-established, and growing, body of work in HCOMP on examining the factors which affect the data annotations and their 'quality' (*e.g.,* incentives [87], interface [59], description [14], among others [19, 46]). In 2015, Aroyo and Welty [4] identified seven myths pervasive in the practice of data annotation, and proposed the theory of *crowd truth* with a premise that human interpretations are inherently subjective. They reject the fallacy of a single truth– assumed in many data collection efforts– of a correct interpretation for each input example [4]. Increasingly detailed guidelines typically eliminate disagreement but do not increase the quality of data, as annotators choose responses with which they may not be comfortable [4]. While the typical discourse around disagreement, both in research, and practice has been to treat disagreement as a noise [98]. Aroyo *et al.* [3] outlined the ways in which disagreement in annotation tasks arises– (1) through subjectivity inherent to the topic, (2) through subjectivity introduced by the ambiguity in the task or (3) by spam produced by bad actors.

A common thread across this body of literature is to demonstrate the role of annotators' socio-demographic factors and lived experiences on their label assessments [25, 37, 51, 58, 96]. An emerging community of researchers have turned to using disagreement as a signal for deepening their understanding of the task and the data [3, 20, 77]. As Prabhakaran *et al.* [77] demonstrate how label aggregation may introduce representational biases of individual and group perspectives, Davani *et al.* [20] propose an approach that looks beyond the use of majority vote as an aggregation method. Examining the role of annotator subjectivities and its impact on datasets is a domain of growing interest within HCOMP, FAccT and HCI. Our work extends this body of research by examining the current praxis of ML building and the practicalities which hinder nuanced diversity-related considerations in ML building.

2.3 Discourses on diversity in machine learning

Past literature on diversity in machine learning systems is within two broad areas– those who design and build the ML systems [49, 99], and the diversity considerations embedded in the data pipeline (*e.g.,* who represents and what gets represented [27, 85]. Some have called for more participatory approaches to problem formulation– by including diverse group of affected publics into the decision making process [54].

One set of discussion is around the diversity of the ML engineers, researchers and experts and in general the composition of teams creating ML systems. A growing number of voices have pointed to the lack of diversity in the AI industry [99], and emphasised the benefits from diversifying the teams as a way towards safer ML systems [22]. The motivation here is that a diverse team would lead to greater deliberation and thus more ethical AI systems [49].

Others in the community have studied the significance of diversity in the data pipeline– in the samples, who creates and is represented in the data [13]. First, there have been efforts to understand the kinds of diversity which needs to be pursued in the instances within a dataset (*e.g.,* which regions are covered, individuals of which identities are covered) and the effects of incorporating such diversity particularly on the performance of downstream ML models and applications [85]. Many researchers have pointed out the ways in which collecting more diverse data for relevant dimensions could lead to fairer ML models [9].

Fazelpour and De-Arteaga [28] contributed a taxonomy for key diversity concepts and situate them within the discussion of diversity in socio-technical machine learning systems. They emphasised the ways in which diversity-related choices and considerations are embedded throughout the design of ML systems, regardless of whether they are actively recognised. We contribute to this area of research by focusing particularly on the conceptions of diversity within annotator pools in practice, the approaches towards operationalising annotator diversity in ML projects.

2.4 Representationalist thinking

Given the complexity of the ML practice, and the contingencies that shape it, we draw on the concept of *'representation'*, influenced by scholars in the history of science and STS, such as Hacking [39], Barad [5], Goodwin [34], Haraway [41]. The key features of this thinking are:

- There are phenomena or effects in the world, awaiting discovery
- The world (and actors within it) can be observed and represented in neutral and/or objective ways
- It is possible to observe/represent features, characteristics, behaviours, etc. in isolation
- It is possible to apply these representations in wider or different contexts

Particularly important here are the systems and tools used to do this active seeing, representing and intervening, and the ways in which they cement representationalist thinking. Hacking [39, p. 186], for example, describes the scientific use of microscopes to show how 'seeing' through the instrument involves elaborate theories of light, optics, *etcetera*, as well as considerable training and skill on the part of technicians and scientists to obtain meaningful results. Despite this, scientists still speak of 'true images' obtained from microscopes and treat them as representations of a world that exists. Where the different approaches and methods yield inconsistencies and, in turn, doubts or scepticism about accuracy, a practical approach is taken; results are treated as representative of phenomena so scientists can progress with their work. The underlying logic prevails. As we progress through the findings and the discussion, we make the case that data practices and particularly the thinking surrounding annotation in datasets has parallels with this perspective on representationalist thinking.

2.5 Regimes of existence and intersectionality

As we noted in the Introduction, Teil explores the prevailing logic of neutral or objective representation through *regimes of existence* [91]. Because terroir is evidently subjective—to do with individual tasting, shared opinions and collective judgement throughout the wine-making process—it is relegated in this regime. The mechanised and scientific approach to wine production, heavily dependent on objective measures of wine quality, have a suspicious view of terroir; for the ardent critics, the inability to identify independent measures of terroir, set it out it as "groundless" [91, p. 492]. Writing about terroir, Teil suggests the combined qualities of terroir "escape scientists' objectification because they do not bend to its requirements of an apriori differentiation between product, producer, and production techniques." [91, p. 493]. In other words, terroir is not a metric they can use.

This idea of "regimes of existence" in wine tasting and production may seem a long way from diversity in data annotation. This, however, helps us to later foreground the representationalist thinking in data practices and enable us to critically examine its prevalence as an overarching logic.

It is in this regime of existence, brought into being through representationalist thinking and data practices, where we find that diversity people's subjectivities can be relegated, placed subordinate to other practical goals. This same theorising of worlds or regimes, however, presents us with opportunities for different ways forward. Like Teil, Davis, Williams and Yang [21], are disenchanted by the worlds in which scientistic and objective logics operate. In these worlds, they see an "algorithmic idealism" pervade in which fairness and equity are calculable [21, p. 3]. Overlooked or ignored is a world in which experiences are always felt in particular places and through particular bodies: a recognition that "'objectivity' is never neutral," (ibid.).

2.6 Justice-oriented approach to diversity

We draw on justice-oriented, feminist and intersectional scholarship towards diversity for inspiration [21, 61]. Intersectionality, one of the major paradigms from such scholarship, is both a normative argument and an approach for critical inquiry and practice [40]. Intersectionality emphasises the ways in which multiple social categories of difference intersect, are interrelated and mutually shape one another [15].

We turn to the three distinct approaches outlined by Leslie McCall [61] – inter-categorical, intra-categorical and anti-categorical– in dealing with the complexity of intersectionality. The *inter-categorical* complexity for intersectionality focuses on the intricate and complex relationships between multiple social groups within and across categories. It explicates the constituted inequality present among social groups. The approach of *intra-categorical* complexity, sits in the middle of the continuum between anti- and inter-categorical complexity, and maintains a critical stance towards categories while acknowledging the stable relationships that these analytical categories represent. It calls upon a need to account for the lived experiences, particularly at the points of intersection where they are most ignored. The *anti-categorical* complexity is based on the deconstruction of analytical social categories. It challenges the imposition of categorisation which renders a stable order over a heterogeneous ever-changing social reality, thus contributes to exclusion and inequality. Building on STS scholarship (such as Hacking's [39]), McCall [61] invites a greater criticality and investment in alternative imaginations of inequities in practice and of the world. Our goal here is to demonstrate that a fundamental change is needed to take diverse subjectivities seriously. However, practitioners do not tend to consider how their practices produce the phenomena they want to be represented or labelled. In other words, they resist recognising *intervening* that Hacking depicts [39].

3 METHODOLOGY

This paper defines data annotation as the sense-making practice of a given dataset, where annotators assign meaning to data using (pre-defined) labels. Previous human-centered investigations which concerning annotators' subjectivities, biases, and efficiency primarily focused on the annotators' perspectives. In this research, we 'study up' the data requesters to understand their conceptions of annotator diversity [69]. In order to gain a holistic view of the data annotation practice we employed mixed-methods research approach with a sequential explanatory design [45]. In this design, a survey aimed at getting a broader view on the ML data annotation process and diversity more specifically was conducted and analysed first. We then followed it up with semi-structured interviews to investigate, in more detail, the ways in which practitioners understand data annotator diversity and the structural and epistemic barriers. The qualitative data collection and analysis serves as the foundation of our inquiry even though the quantitative data collection came first. We describe implementation details in the following subsections.

3.1 Survey

The goal of the quantitative phase was to identify practitioners' perceptions of incorporating annotator diversity into their practice. We conducted the quantitative phase using an online questionnaire implemented in Qualtrics.

Sample. We recruited respondents through multiple channels. We emailed the survey to direct contacts and relevant mailing lists internal to our organisation. We also distributed the survey within our professional networks through Twitter and LinkedIn. We began the survey by eliciting informed consent from respondents. No personally identifiable information was recorded about the respondents. The inclusion criteria for our survey was similar to the interviews. The screener question asked, *"have you had data collected or annotated for an ML/AI project in the last 12-24 months?"*

After the screening question, we were left with n = 78 participants. However, not all respondents completed it, so we analysed only 44 of the respondents who completed at least one section. Among our survey respondents (each could select more than one role), practitioners worked in research (25), software development (14), data & applied science (7), product management (3), and user experience roles (3).

Questionnaire. Our questionnaire consisted of 22 questions in total, with a mix of multiple choice (17) and open-text questions (5). Survey takers were asked to respond to the questions in the context of the dataset annotation process for *one* of their recent projects. We began by asking respondents of their job role. The rest of the survey covered the following themes: 1) understanding their project such as the ML task and dataset curation and labelling process, 2) annotation platform selection, 3) annotator selection, 4) perceptions on annotators' subjectivities, 5) annotator information, 6) challenges in setting up annotation, and 7) ideal annotation task design. After understanding their annotation process and task type, our survey had three primary questions around their recruitment criteria (*"did you use any of the following criteria to recruit data annotators?"*, on the available information about the annotators of their dataset, and the relevant attributes of data workers which could affect the annotations. Each of these questions had socio-demographic attributes such as the annotators' age, gender; expertise; location; and more. We also ask, *"in [their] experience, to what extent does the diversity of the data annotators pool influence the dataset quality for [their] task?"* using a five point Likert scale of 1 being *"not at all influential"* to 5 being *"extremely influential"*, followed by a question on *why* they believed annotator diversity is influential to the extent that they specified.

Analysis. For the survey responses, we computed several kinds of descriptive statistics using SPSS to better understand practitioners' approaches to diversity, and particularly the kinds of diversity they see as relevant, if at all. These included the descriptives to questions presented with Likert scale response options and frequencies in questions where respondents could select multiple choices (*e.g.,* the challenges in recruiting the desired pool of annotators). We focus particularly on comparing the differences between the attributes which are used to recruit annotators, the kinds of information available to practitioners and what they would see as relevant attributes in the ideal scenario. In cases where questions were completed by a subset of the respondents, we report question-specific response rates and percentage of respondents who answered that question. Finally, we conducted a qualitative analysis to the open-ended questions following the same to the interviews (see the following section 3.2 below). We performed multiple rounds of coding at the response level in conjunction with participants' survey ratings to surface high-level themes.

3.2 Interviews

Between April and May 2022, we conducted semi-structured interviews with a total of 16 practitioners involved in the creation of annotated datasets for AI development. Each interview had structured sub-sections beginning with the participant's background in AI and in what domains or areas they have used AI. We then asked participants to walk us through a recent AI project, end to end, to learn about their typical working method and AI development process. Our interviews focused on: (1) understanding the data annotation setup; (2) selection process for annotators; (3) annotator diversity considerations; (4) ideal annotator diversity; (5) annotation documentation and reuse practices; (6) organisational structures and incentives within data annotation. Each session focused on the participants' practices, experiences and challenges with setting up data annotation tasks—particularly those that informed the annotator pool for their task.

Participant recruitment. In our sample, AI practitioners were located in, and worked primarily on projects based in US (8), India (4), UK (3), and France (1). While we interviewed practitioners working in multiple institution types, varying from large companies (8), startups (4), to academia (4), all were involved in the set-up of annotation tasks for

P#	Role	Location	Institution type
P01	Researcher	India	Large company
P02	Software Engineer	India	Large company
P03	Researcher	India	Large company
P04	Researcher	India	Large company
P05	Professor	United Kingdom	Academia
P06	Researcher	United Kingdom	Large company
P07	Program Manager	France	Large company
P08	Data Scientist	United States	Large company
P09	Researcher	United States	Academia
P10	Researcher	United Kingdom	Academia
P11	Chief Science Officer	United States	Startup
P12	Linguist	United States	Large company
P13	Researcher	United States	Academia
P14	Chief Data Officer	United States	Startup
P15	Data Scientist	United States	Startup
P16	Operations Manager	United States	Startup

Table 1. Interview participants' role, location and experience, n = 16

AI/ML projects. Many participants also contrasted their experiences working across these institution types, such as within a startup / academia and a large tech company. A majority of our participants were in research-centric roles, but a few also worked as data scientist adjacent profiles, linguists and managing operations for data annotation. Refer to table 1 for details on participant roles, locations and institution types. Many participants spoke of experiences with annotating datasets across multiple domains and AI technologies; we report the primary AI technology and domain of application at the time of the interview. The distribution of type of ML projects that our participants worked on skewed heavily towards language-based ML tasks. Among the survey respondents, twenty-two (50%) worked on language-related tasks (classification or generation), eight practitioners identified object/entity recognition as their task-type, and five worked on human evaluation of model generated data. Examples of language tasks from our interview participants include semantic parsing, translation, decontextualising sentences, harmful content detection, and more. Other types of projects represented among our interview and survey respondents include detecting anomalies in chest x-rays, segmenting rivers in images for flood forecasting, developing taxonomies of items found on an online marketplace.

We recruited participants through a combination of distribution lists, professional networks, and a third-party research recruitment agency, using snowball and purposive sampling, until we reached saturation. Our sampling strategy was to include practitioners who were involved, end to end, in setting up an annotation tasks within the last 24 months.

Interview moderation. Given the geographical spread of our participants, the interviews were conducted online using video conferencing software. We scheduled sessions based on participants' convenience and conducted all interviews in English (preferred language for the participants). During recruitment, participants were informed of the purpose of the study and researchers' affiliations. Informed written consent was obtained electronically for all interview participants and verbal consent for recording the meetings. The participants were informed that they could refuse to answer any questions or ask for the recording to be paused at any time. Each interview lasted about 50-70 minutes each. We recorded interview notes through field notes and video recordings which were transcribed verbatim subsequently. We stored all data in a private Drive folder with access limited to the research team, and deleted all personally identifiable information to protect our participants' identities. Each participant received a thank you gift card with amounts localised in consultation with regional experts (40 USD for the US, 75 USD for the UK, and 27 USD for India).

Analysis and coding. The interview transcripts were analyzed collaboratively by all authors to identify relevant themes. The analysis was inspired by and consistent with the ethnomethodological ethnographies in HCI [16, 17, 71]. Our analysis took a broadly ethnomethodological perspective. Ethnomethodologically-informed, ethnographies explicate the knowledgeable, artful ways in which workers orient to their work and how technologies and other artifacts are used as part of the methodical accomplishment of that work [10, 79]. As well as analyzing interview transcripts, we took the additional step to examine the text and visual materials shared by the participants which they used to commission their data annotation tasks. These walk-throughs which was in tacit knowledge regarding the data annotation granted us vital extra contextual understanding of the practice. Ethnomethodological analyses of work are useful in generating a granular understanding of what activities constitute 'work' in a setting, how they are accomplished in practice, who is involved in this accomplishment, what resources are drawn upon, and what skills and tools are involved in mobilizing those resources [ibid]. Through this close look at the seemingly ordinary details, our analysis seeks to unveil not just what the world looks like but how it comes to look as it does. The emphasis is, in other words, on the detail of work as understood and interpreted by the people who perform it and in our case this refers to the entire practice and process to get data annotated for ML model building.

The interview data were analysed by the authors individually and together in analysis sessions explicating a particular topic, as is typical of the ethnomethodological approach. Since we adopted the 'grounded approach' [32, 33], the techniques of constant comparison and constant iteration (*i.e.,* iterations of coding and re-coding) were used in the development of themes so as to avoid the classic problems of 'cumulation' and 'theoretical imperialism' – *"an analytically imposed reconstruction of the procedures of a setting, insufficiently sensitive to the understandings of a setting's participants."* [32] These analytic sessions allowed interesting topics to be identified and endogenous themes to emerge from the data (such as the general ML data workflow, the various approaches incorporate and dismiss diversity, and the dissonance between practice and discussion). To stay true to the grounded approach, we were extremely cautious not to impose categories external to the data to codify the data. Ethnomethodological ethnographies are valuable in informing design [79], and we used the resulting understanding of fundamental conflicts in the thinking around ML development and diverse datasets creation to inspire a set of implications that aim to address some of the challenges requesters face in incorporating diversity and therefore to steer the discussion around the diversity in data towards a more constructive and worker-centered direction.

4 FINDINGS

Our findings present the various approaches towards annotator diversity adopted by the participants and the barriers that hinder a nuanced consideration of diversity. We begin by describing the data annotation workflows to illustrate the context of data work and ML pipeline. We then describe how diversity was approached in practice and the underlying logics motivating their choices. Finally, we conclude our Findings with the various barriers to incorporating annotator diversity in data practices.

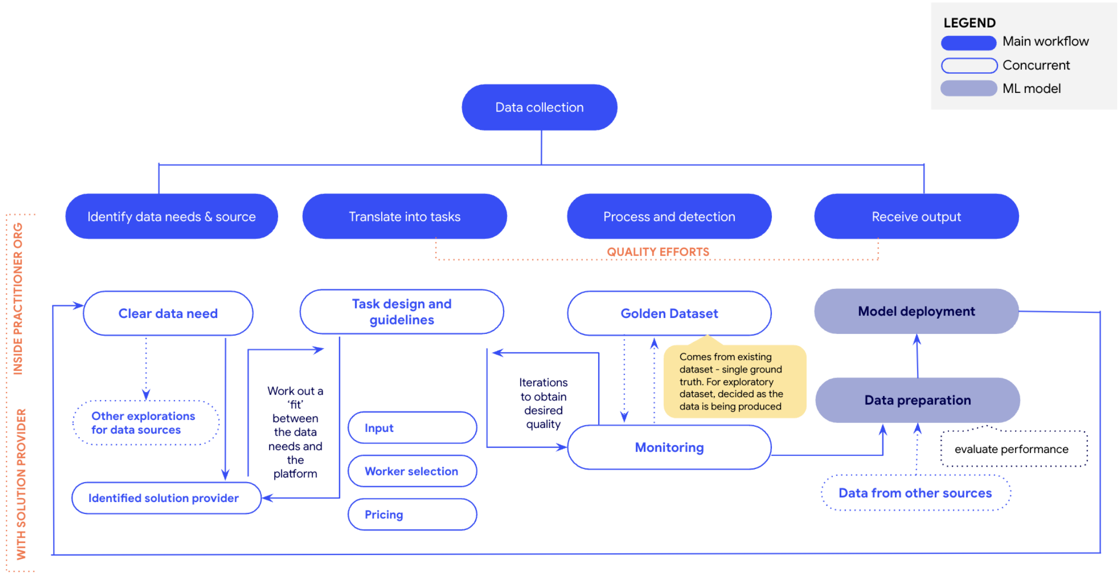
LEGEND
Main workflow
Concurrent
ML model
Data collection
Identify data needs & source
Translate into tasks
Process and detection
Receive output
QUALITY EFFORTS
INSIDE PRACTITIONER ORG
WITH SOLUTION PROVIDER
Clear data need
Other explorations for data sources
Identified solution provider
Task design and guidelines
Work out a 'fit' between the data needs and the platform
Input
Worker selection
Pricing
Golden Dataset
Iterations to obtain desired quality
Comes from existing dataset - single ground truth. For exploratory dataset, decided as the data is being produced
Monitoring
Model deployment
Data preparation
evaluate performance
Data from other sources

Fig. 1. The data annotation workflow for ML projects

4.1 Data annotation workflow and tasks

The workflow for annotation tasks began by identifying the data needs for the ML project with consideration of the downstream applications. The practitioners then solicited the platform based on their criteria (*e.g.,* cost, quality; outlined below). Participants then proceeded to design the annotation task with the platform managers' help. It started with a 'pilot' phase to test the annotation guidelines which was iteratively refined. These refinements were additional example annotations or edge cases to provide further clarifications to the platform and the annotators. While the annotators labelled the data in bulk, the practitioners would monitor the process periodically for quality, often by comparing the annotated with the 'golden dataset' which was created through 'expert' inputs or by the practitioners and verified by the models they built.

Most practitioners in our study relied on data infrastructures within their organisations for producing annotations. These were dedicated internal mediatory platforms whose primary purpose was to recruit and manage annotators and project manage tasks. We note the preference annotation platforms over annotator-facing marketplaces (*e.g.,* MTurk) among our participants (echoing Wang *et al.* [95]). Out of our forty-four survey respondents, twenty-five relied on internal annotation infrastructures, and eight outsourced to third-party vendors (*e.g.,* Appen, Scale AI). The top considerations our practitioners had for annotation infrastructure were– cost (15), timeline (16) and the quality of platform (*e.g.,* proprietary tech/UX/support etcetera) (17). Only 8 practitioners considered the diversity of data workers' on the platform as a deciding factor.

4.2 Approaches to diversity

In the following, we capture the varied perspectives on diversity, from it being seen as irrelevant to efforts made to accommodate it, even if only in partial ways. Most participants acknowledged the role of annotator subjectivities in the annotation process and some spoke of how diversity captured a balanced view and how it brought previously overlooked

Criteria	Relevant attributes	Selection criteria	Available information
Gender	n = 15	n = 4	n = 11
Geographic location (*e.g.,* country, state)	n = 20	n = 8	n = 19
Age	n = 17	n = 2	n = 8
Race/ethnic group	n = 17	n = 4	n = 8
Education level	n = 21	n = 3	n = 10
Language proficiency (*e.g.,* English, Hindi)	n = 21	n = 12	n = 15
Subject-matter expertise (*e.g.,* linguist, doctor)	n = 20	n = 5	n = 11
Political orientation (*e.g.,* liberal, conservative)	n = 7	n = 0	n = 2
Religious orientation (*e.g.,* Muslim, Christian)	n = 8	n = 0	n = 1
Sexual orientation	n = 8	n = 2	n = 2
Health, mental health, disability	n = 11	n = 1	n = 5

Table 2. The number of respondents for three questions around the attributes of data workers which could affect the annotations (relevant attributes: column 2), their criteria for selecting annotators (selecting annotators: column 3), and the kinds of information available to them about the annotators of their dataset (available information: column 4). Each of these questions had attributes (column 1) such as the annotators' age, gender; expertise; location; political orientation; and more.

sub-populations into the spotlight. However, many went on to explain their decisions not to consider annotator diversity when setting up annotation tasks. The primary focus was placed on achieving a threshold of quality. Here, quality was often measured against how closely an annotator carries out a task according to pre-defined parameters and guidelines. Below, we describe practitioners' justifications for a representationalist approach to annotator diversity and prioritising pre-defined measures of quality.

From the survey, 75% of respondents reported that diversity is *somewhat* influential to *extremely* influential to the quality of their annotated datasets (>=3 on the Likert scale). However, although most participants considered various individual attributes and characteristics of annotators as relevant to the annotation process, very few of these attributes were used to recruit them (see Table 2). For instance, in the survey, 4 out of the 15 practitioners who considered the gender of annotators as relevant criteria actually included gender while selecting annotators. Even when additional information on annotator characteristics was available, they were rarely included as part of the recruitment criteria (refer to table 2, column 'Available Information').

Additionally, through our interviews, we find a stark contrast between how participants reflected on the term 'annotator diversity' and how it was put into action. In practice, they understood it as indicative or representative of a point of view. For instance, annotators were selected for the low-resource dialect they spoke or the high flood risk areas they lived in. Language and locale were understood as a proxy for diversity or representative of it, without further reflection. Very few participants saw experiences, knowledge or expertise as facets of annotator diversity or recruited annotators for them.

In the rare cases where local knowledge and expertise were required, measurable criteria were applied to their expertise or knowledge. P6's motivation for selecting an annotator for a mapping project was to include annotators from underrepresented backgrounds into data work and *"to capture the other parts of the population"*. P9 spoke of recruiting *"a person of X identity with Y knowledge."* Both participants demonstrate a view of diversity as something represented through a category or metric rather than bound up with the experiential.

4.2.1 The pursuit of objective annotations.

In both the survey and interviews, several practitioners described why annotator diversity was not relevant, for instance, when the task was seen as objective. As a survey respondent described, *"our data had ground truth."*. Justification for this view of objectivity centred on the premise that some annotation tasks are amenable to objective decisions—that there exist a class of questions with definitive answers or types of content with definitive labels. This view was prominent for those who described their annotation tasks as a matter of *subject expertise— e.g.,* detecting anomalies in chest X-rays or linguistic corpus detection tasks. Here, participants dismissed diversity amongst their annotators as unimportant or irrelevant. P1 captured this in their experience with annotation tasks:

> *"Our primary consideration was how medically trained the annotators were and how much time they had for the annotation. So with regard to factors for diversity, I do not think that was a consideration because it was never intended to be used in the general population."*

This view extended to quality checking (where some annotation results trigger additional quality checks, for example, because of disagreement between annotators). Participants noted how, in cases of disagreement, they brought in a resolver who acted as an expert to make the final decision on an annotation. For a language understanding annotation task for voice assistants, P11 described how they dedicated 'resolvers' dealt with discrepancy, *"the resolver would pick which of those three annotations was correct. They would also have the option of picking none and doing it from scratch."* Expertise here is based on experience, with resolvers having greater years of experience in annotation.

Practitioners saw the design of tasks as an intervention point to ensure objective annotations. They spoke of training sessions and guidelines as a means for annotators to learn how to make 'correct' judgements – *i.e.,* closely follow the instructions without deviation. Participants described examples of their tasks intended to capture predetermined phenomenon (*e.g.,* river segmentation) that could be made explicit through the annotation instructions.

Detailed instructions were propagated from practitioners to annotators through layers of quality checks carried out by platform leads or team managers. These documents broke down annotation tasks into simple, repeatable sub-tasks that were *"very hard to answer in a biased way"* (P8) to reduce inconsistencies and to standardise the work for all raters. As P7 articulated, *"it is less about choosing the right raters and more about ensuring that they have that [standardised] understanding."* In effect, being 'objective' (in accordance with the practitioner's requirements) was deemed a trainable skill and the training sessions and guidelines were essential pathways to instruct the annotator to 'see' objectively.

4.2.2 The attempt to remove 'bias'.

Most practitioners recognised the complexity which annotators' diverse subjectivities introduced into the task design and the project evaluation. This complexity, however, became another justification for the participants to avoid over complicating the goal of achieving useful and testable AI/ML outcomes.

AI/ML workflows are designed to facilitate consistent evaluation across a range of source data, tasks, techniques, annotators, and so on. Controlling for these was challenging, but control over the process helped participants in comparing the performance of AI/ML models and systems, and identifying scopes for optimisation. Annotator diversity and varying subjectivities among annotators were seen as adding to the complexity that practitioners wanted to circumvent. As P6 expressed, the reason they only included questions with definitive answers in their dataset was that *"you want to have an easy way to evaluate the answers in the end."* Speaking of his specific area of work, P6 explained how information-seeking tasks are created using *"a specific span, so you can point to which span contains the answers."* In effect, this is done to limit the plausible options in an information-seeking annotation task and reduce ambiguity, especially when it comes to evaluating a model's performance. Again, the process is designed to lessen the variation in order to reduce complexity. This logic extended to the diversity of the annotators as well– annotator diversity was presented as a further source of variability and ambiguity, and thus something to be avoided to make practical progress in modelling.

This desire to control and manage ambiguity and complexity also applied to bias in annotated data. Annotator subjectivity was regularly viewed as a form of 'bias' manifested as disagreements between annotations and annotators. Describing how they understood the implications of diversity, participants frequently conflated the concept with bias. Diversity was not something to be understood but rather, like bias, a source of variability that needed to be corrected for or technically resolved. Interventions were thus designed throughout the annotation process to minimise or eliminate such disagreement due to bias, but with little to no scope to account for differences in annotations due to annotators' backgrounds and experiences.

Though many participants acknowledged the data quality (*i.e.,* the accuracy rate) and disagreement could be a result of the poorer design of either the annotation guidelines or the annotation interfaces, some attributed the disagreement and inconsistency to individual annotators' attributes. Differences were not understood in terms of annotators' diverse opinions but rather unsatisfying work quality, or worse, questionable work ethics. P3's comment is illustrative:

> *"[The] reason for disagreement could be multiple factors. They don't have the required knowledge, [the] task itself could be ambiguous [...] Second is the quality of guidelines. Third is their motivation and the quality of the work. If they are doing it without a high consideration for quality, they may not push themselves enough for high quality output and that could show up in the disagreements."*

4.2.3 The quest for neutral representation. Only a few participants experimented with incorporating annotator information in their data production and model building. They engaged with annotator diversity by recruiting annotators from a mix of identities, such as an equal gender split or multiple geographies. However, practitioners found it challenging to arrive at and prioritise a set of social categories relevant to their task. For example, P5, expressed the desire to capture a 'representation of every single person and every single dimension' in their research work on toxicity annotation and the rationale behind this attempt:

> *"Otherwise we get biased annotations, which if we do train models on that - they will amplify this bias [...] There is a disagreement based on demographic characteristics. Even if your other demographic attributes are the same, just because of the location, you might have a different perception of the data."*

The attempt was to capture differences in annotation patterns and establish a correlation between the patterns and the identities of the annotators. We note a common assumption among practitioners that collecting a wide range of worldviews can correspond to the real world, that this representation and aggregation can achieve a neutral position. This assumption also considers some knowledge can be seen from this neutral position, free from bias. They had no means of knowing the annotators' perspectives or how they arrived at their label assessments. There were little precedence given the endeavours to incorporate annotator information were still in nascent stages.

Some practitioners noted the tensions and trade-offs between representing 'everyone' *vs.* their user base. P1 spoke of procuring and having annotated 80% of their training data from the Global South, even when their models would eventually be deployed in other Global North countries such as the US due to the disparity in data regulations across the globe. Others attempted to build systems that would work well for marginalised groups, but struggled to justify, from a business standpoint, the additional resources (*e.g.,* time, budget) that these efforts required.

An important factor that hindered the early attempts at incorporating annotator diversity was that current data and ML practices do not actively support explorations of annotator diversity. When the annotators provided label assessments from 'diverse' vantage points, practitioners were not able to distinguish minority opinions from 'noise' which deviated from instructions. The annotations, potentially rich in diversity, were aggregated and distilled to eventually arrive at an agreement or an acceptable range to be useful for AI/ML modelling. Thus, the current techniques for model building are not amenable to account for ambiguity. At an individual level, the annotators were trained to adhere closely to the task instructions and to move away from their individual interpretations. At a cohort level, majority vote technique was commonly used to select the salient result. Therefore, what eventually was represented in the data, was neither diverse nor neutral.

4.3 Barriers to incorporating diversity

In our research, participants reported several barriers to accommodating annotator diversity. One, the lack of access to annotator information. Two, the limited communication between practitioners and annotators, meaning practitioners knew little about annotators beyond a 'worker-id'. Finally, the lack of actionable paths to incorporate annotators' socio-demographic information meant practitioners had little motivation to consider diversity in the first place.

4.3.1 Lack of information about and communication with annotators. Several interview participants noted that they had little information about the annotators working on their tasks. At best, practitioners knew about their annotators through website brochures or blog posts from the third-party data-labelling platforms they used for recruitment. This information was reported in aggregate and publicly available and not specific to their projects. In practice, annotation projects operated on a **'good-faith basis'**—the third-party platforms were trusted to satisfy any recruitment requirements and there were rarely opportunities to confirm if annotators met the criteria.

Among the forty-four survey respondents, 19 of them reported having access to the geographic location of their annotators, and 15 had access to the annotator's language proficiency. The other commonly available information was the education level (11), subject matter expertise (11) and the gender of the annotators (10). These were the kinds of information that the practitioners had tried to obtain about their annotators. 17 respondents did not face any challenges to obtain the information they wanted. However, 8 respondents struggled with a limited project timeline and 7 respondents cited the legal constraints in acquiring any information about their annotators. In addition to these challenges, the respondents mentioned that they had limited control over who their annotators were. 18 survey respondent even struggled with the access of annotators which they believed suitable to their task.

Having access to annotator information and incorporating it into data production also had legal ramifications. In this context, participants expressed apprehension towards collecting sensitive personal information about the annotators, such as their sexual or political orientations. This stood in tension with the need to better know annotators' backgrounds and adopting a nuanced consideration of diversity. Our interview with P8 captured the multiple barriers with respect to diversity. In managing annotation projects at a big tech company, P8 explained the challenges of recruiting a diverse group of annotators: *"you are not allowed to select people for a job based on certain characteristics. It is illegal to give a questionnaire as to their sexual orientation and select people based on certain orientations to fill up a data centre. Even in countries where it can be done - there is no way [big tech company] would expose themselves to potential PR nightmare of having an article about how [big tech company] is selecting certain sexual orientations for data annotations. "*

We see here how different structures come into play, forming barriers to recruiting diverse annotators. P8 describes how legal, ethical and corporate considerations intersect to make *"selecting annotators to be diverse... impossible"*. These complications were intensified further with a sense of moral obligation or duty. Expanding on their comments above, P8 also spoke of ensuring that annotators remained safe when they were subject to repeated annotation tasks of harmful content. Annotator well-being, and an obligation to consider it as a practitioner, became another axis to be considered if information is to be collected about annotators and any form of communication established.

Most practitioners in our interviews worked with third-party platforms to have their datasets annotated. The separation of operations between the annotation platforms and the data practitioners resulted in several difficulties. Platforms, while playing an important role in reducing the overhead on practitioners, also introduced distance and separation between practitioners and the annotators. Most annotation projects were mediated through a platform manager or team lead who facilitates the communication between practitioners and annotators. There was, barely, if ever, direct communication between those who work on the data and those who work on the models using these data. Many practitioners spoke of never 'meeting' an annotator. Speaking about their work at a big tech company, P14 talks about the communication barriers in getting the most value out of their annotation process:

> *"The thing is that it was all contracted out externally. [Big tech company] has this business deal with [annotation company]. All of [big tech company]'s tools are proprietary and internal and we have a specific interface where the moderators quickly review things. All of [the annotation company]'s workers are out in the Philippines. In the US, the majority of them are in Austin, but there is a total disconnect between the actual [big tech company] engineers and the contingent workers. "*

Geographical distance, time differences and heavily-facilitated communication enforced and amplified a separation between practitioners and annotators. To avoid the significant communication delay, the practitioners often had to resolve 'inconsistencies' in data labels among themselves. Although it was acknowledge as a poor practice, this was largely dictated by tight turnaround times and business pressures.

Challenges remained even when direct practitioner-annotator communication was possible. P5 noted the difficulties in fostering trust and collecting annotator information. They described one of their projects with MTurk where the annotators did not feel comfortable sharing personal information with the project team. The annotators wanted to understand why the information about them was collected and how it would be used. Despite P5's explanation, the annotators distrust against MTurk took over. As the interaction between the practitioners and the annotators is often facilitated through an intermediary, there is a three-way trust that needs to be established yet the practitioners being on the other end of the interface have little power to reestablish the broken trust between annotators and the platform. Thus, factors such as establishing trust with annotators, and deciding and setting up suitable pay came before considerations of bringing in diversity among annotators.

The lack of information and communication channels further aggravated the separation between the practitioners to their data annotators. Without knowing who these annotators were, it was much easier for the practitioners to consider them as interchangeable workers who carried out standardised tasks. Thus, only a few practitioners conceptualised the impact of the annotators' identities on their data and how diversity within such identities could matter.

4.3.2 Build the model first, build the model fast and build the model cheap. The status quo of driving practices in ML workflows also presents barriers to conceptualising and operationalising annotators diversity. Several practitioners noted how they had to prioritise curating larger datasets and building better-performing models over bringing in a diverse group of annotators under the pressure of short-term development timelines. Participants who worked in emerging and niche application areas—where their focus was on exploring the limits and capabilities of ML models by testing new ideas and concepts—had to prioritise 'hitting the ground running' and reaching an 'MVP' (minimal viable product) before they could consider the specifics of their annotator pool. As P11 described,

> *"Even if those things could matter, it depends on where the project is and what are the priorities—you want it to work well for everyone but at that stage you're trying to just make it work in general. [...] It takes additional resources to address smaller user groups. There's a persistent, open question on whether it was even a priority to get models working for, let's say older people or for speakers of dialect where there's not that many users because that ends up being harder to justify. It is always about trying to satisfy the needs of the largest groups of people."*

This sentiment was echoed by many who did not have the evidence for *justifying* why they should seek diversity in their annotators.

For ML practitioners, data annotation is only a small part of their ML pipeline which must fit with the other components of the workflow such as the model building. P12 described how data collection/annotation pipelines are configured to build models and not necessarily for advancing task understanding. The complexity of embracing diversity in annotators, and thus their subjectivities, stood in conflict with pipelines that are designed to arrive at definitive answers. Thus, practitioners had to work within these constraints.

Both the platform managers and practitioners work to actively reduce the cost of annotation in setting up ML data production. As has been well-documented annotation companies often recruit their data workers from countries in the Global South such as India or the Philippines where labour costs are considerably lower, while their engineering offices are in the Global North, in order to remain cost-effective (similar to [63, 66, 95]). We observed the same in our study (*e.g.,* P14). P16 who worked as an account manager with a mid-size annotation company shared their perspective from the platform side. Annotation companies operate in a hyper-competitive environment in which they typically hire workers from lower-wage countries to keep their pricing lower than competitors. Practitioners also tried to minimise annotation costs, and thus did not have much control over the annotators assigned to their projects. P6 also spoke of the connections they saw with the appropriate incentives and 'data quality', of the need to compensate annotators fairly for their time.

Practitioners spoke of the intricacies in setting up an annotation pipeline– from setting up the instruction documents and the interface, deciding the platform, and the multiple iterations it took to get things right. Though critical and necessary, the participants mentioned how the overhead created by the process could have been spent in model building. In addition to 'competing' with model building for time and resources, data annotation procurement also covered a majority of the operational costs of ML projects. It was infeasible and uneconomical to repeat the data annotation even if diversity-related concerns were found after completing the annotation process. The practitioners often had to make ends meet and in some cases resurrect the data for different purposes. Annotator diversity, that is yet to prove the value it adds to model building, was unsurprisingly side-lined as a minor figure in comparison with the more pressing priorities – how a better-performing model could be built fast with lowest cost possible.

5 DISCUSSION

> *"Feminist objectivity means quite simply situated knowledges."* —Donna Haraway [41]

Through our findings, we show how ML practitioners make sense of annotator diversity in their data practices. To a large extent the diversity among annotators is not prioritised, and their backgrounds and experiences are downplayed when set against the more pressing practical challenges of building workable AI/ML models. Practicalities such as cost control, model evaluation and product profitability cast diversity as a minor figure.

As well as this practicality-driven mindset, our findings indicate structural arrangements have a role in decisions about annotator diversity. A range of structures operate to lessen the impact and value of recruiting and involving diverse annotators. We heard about systems and training being put in place, tasks being carefully designed and tightly constrained, and datasets being iteratively developed, all so annotators could learn to see things in the 'right way' and produce uniform datasets amenable to modelling.

An overarching idea of objectivity dominated our participants' views of annotator diversity. Participants framed their decisions about diversity in terms of golden standards. The quality of annotated data and need for diversity were gauged against measurable and seemingly objective standards. Here, the annotator is treated much like Hacking's microscope (Section 2.4 [39]), as an apparatus for achieving some representation of the world. Wherever diversity is sought, it is to converge on a truth, and where there is doubt, the practicalities of progressing to model-building take precedence.

We contend that the viewpoints of annotator diversity we have captured operate within a wider *representationalist thinking*— in which both annotators and their observations can be formalised and represented in neutral ways. Again, the underlying logic that prevails allows diversity to be reduced to stable, generic categories and the differences in diverse annotator labels to be algorithmically reconciled to achieve a supposed neutrality.

Our findings also paint a picture of a practice where phenomena are being 'actively' seen and represented, and indeed intervened in [34, 39, 91]. Rather than objectively or neutrally representing the world in some passive sense, we find practitioners making concerted efforts to minimise the effects of annotators' subjectivities. Even though there is an awareness of the importance of diversity among practitioners (often expressed in sophisticated ways), data practices, altogether, serve to neutralise differences. Thus, representationalist thinking both exerts a pressure and is sustained by the practices surrounding annotation.

We draw parallels between the prevailing perspectives on data annotation and Teil's analyses of terroir in wine-making practices [91]. Terroir cannot exist, 'objectively', in the mechanised and scientific approaches to wine production. It does not align with the logics that allow the industry to operate—logics in which "apriori existence of scientific 'things' [can be] detached from their process of emergence" (ibid) or data detached from context. Like terroir in wine production, the subjectivities involved in annotating data do not have a space in data practices. In both the worlds of wine production and of data, there is an active pursuit of objectivity; practices and structures separate the observer from the observed and seek to reduce phenomena to uniform or standardised metrics. Teil's *regimes of existence* is intended to capture this active intervening in phenomena, to invite a critical perspective that helps to reveal that representionalist thinking is not neutral but dictates what counts as neutral, objective and valued, and what does not.

By posing this critique, we do not suggest the current configurations of annotator diversity are beyond repair. Our hope is to show there is an opening to *rethink* the thinking, the logics, the regime. As Teil phrases it, *"interpreting objects as distributed products—understood according to various protocols by different users and in different and endlessly renewed circumstances—enables one to restore the plurality of objects and look for local agreements between the different points of view that compose them"* [91]. To think about diversity differently, we learn from Teil, by paying attention to the protocols, users and endlessly renewed circumstances, and, in the midst of these, the local agreements and different points of view. Towards this ambition, we present three critical shifts: a rethinking of ground truth, of bias, and of diversity.

5.1 Rethinking Ground Truth

Many practitioners spoke of the conditions that prevented them from an active integration of diversity in their practice. In most cases, diversity was not the top priority among various competing considerations in ML development. We argue in our findings that the current practices followed by the practitioners present more than a practical obstacle to an alternative approach to diversity. It is common practice in ML projects (notable also among our participants) to collect multiple annotations for each instance [86, 88] and then to apply a majority voting or averaging process [53] if the annotators do not converge on the same response. This disagreement can be quite substantial [36] in various machine learning tasks (*e.g.,* toxicity detection [93], medical diagnosis [82], misinformation [101]). However, the typical ML workflow requires convergence to arrive at an outcome for downstream applications. This approach would be hard to reconcile with diversity-related considerations, where even an 'ideal' annotator diversity would eventually be reduced into a single outcome. While research indicates that such disagreement could in fact be a signal to identify issues with the task construction, many practitioners viewed disagreement as undesirable, getting in their way of achieving 'high quality data'. The common goal in setting up tasks is then to arrive at a singular 'ground truth' label to train machine learning models, minimising the importance of annotators' subjectivities.

The first step to manage these subjectivities would be to acknowledge that human annotation is fundamentally an interpretive task, and it is likely that individual annotators might arrive at a response which does not coincide with the 'majority perspective' [35]. Practitioners could consider approaches which captures the nuances in disagreements, and preserves minority perspectives [20]. Prabhakaran *et al.* [77] demonstrates how aggregating labels obfuscates socio-cultural backgrounds. Those who develop datasets could consider preserving and attaching the individual annotators' labels with each instance to enable analyses and reusability for downstream applications [77].

As humans we communicate and deliberate to resolve disagreement and conflicts, however, worker deliberation is neither appreciated nor supported when disagreements arise. In data labelling tasks, both on platform-based and within private annotation firms, the data workers frequently discuss cases of ambiguity amongst each other, often in an attempt to minimise disagreement [84, 95]. Research also indicates that annotator deliberation and exchanging justifications improves answer quality over output aggregation [26, 70]. ML practitioners often see this deliberation as tainting the dataset by collusion and such discussions are not supported or documented by current annotation tools [65, 66, 95]. In writing of objectivity's part in a regime of existence, Teil argues that the production process of terroir is a collective one [91]. This collective is not democratic, "it is a weighted collective composed of more or less competent and renowned competitors". Similar to terroir, the production of data annotations is a collective, interpretive process (ibid.). How might we then move towards seeing discussion among annotators as deliberation? Disagreements are often resolved by relying on an expert annotator or a resolver with more experience, minimising any effects of having a diverse annotator pool. Annotator deliberations complicate a notion of diversity where each annotator represents a single vote, so practitioners could consider leveraging tools which actively account for deliberative dialogues.

5.2 Rethinking Bias

Our research takes inspiration from an emerging area of research that demonstrates how individual annotator identities influence the annotation task (*e.g.,* age [25], gender [11, 97], sexual orientation [37], race/ethnicity [24, 56] and disability [1, 42]). In short, data annotation (like other data practices [90]) is shown to be situated within particular social and cultural contexts.

Despite the thoughtfulness of our participants, this situated nature of annotation work had little impact on ML or data practices. Among our practitioners, there were no substantive attempts to explore how contexts or any accompanying power relations might influence annotation and what might be done to reflect this downstream in the ML pipeline. Instead, and somewhat perversely, the expression of annotators' identities was seen as 'biasing the data'. We call for a close and critical examination of this framing of 'bias' [47]. We echo Fazelpour and De-Arteaga's [28] sentiment that diversity among the annotators of ML systems should be a justice-oriented pursuit, not (only) in a quest for better-performing models but for the potential epistemic benefits—to broaden ways of knowing. Our contribution here is to recommend a greater attention be given to both the conditions in which annotators work, and their lived experiences and aspirations [65, 95].

A range of works have provided valuable starting points for exploring this position further and the influence on practice. Miceli *et al.* [64], for instance, discuss the ways in which removing bias should not be a universal goal– but instead, how we might reflect on the underlying causes of such differences? In practice, though, we found individual annotator subjectivities were seen through the lens of 'bias mitigation'. As long as the leading belief in ML remains that biases can be caught through careful aggregation of a wide range of views, the dissonance will persist.

Providing potential ways forward, researchers have proposed documentation artefacts (*e.g.,* Datasheets for Datasets [30], Data Cards [78], Data Statements [7]) and archival artefacts of the decisions made [48] during the annotation process were proposed as a mechanism to foster deliberative accountability [75], to make explicit the tacit knowledge in data work [65, 74] and to foster fair reparation in ML building [21]. However, these artefacts are created at the tail-end of the project, after the data annotation has already been completed [80]. While these artefacts may promote greater transparency, they do not adequately address the concerns relating to annotator diversity. Documentation alone does not prevent practitioners from actively designing out the possibility of alternative viewpoints. The intervention or rethinking has to happen before the annotation starts rather than once it is finished.

Our practical recommendation is to leverage the power of communication to address the separation between practitioners and annotators, connecting them through a direct communication channel in the annotation tooling. A small number of research projects have started to explore the role for this kind of co-creation in AI fairness [100], in its broader terms, and in dataset production and curation more specifically [2, 92]. What would a worker-led, participatory approach to annotation look like in practice? Instead of practitioners imposing viewpoints on annotators, through hierarchical management and organisational structures, we invite the practitioners to explore collaborative approaches and co-creative labelling processes, together with the annotators.

5.3 Rethinking Diversity

Lastly, we would like to return to the theorising of justice-oriented intersectionality [21, 61] to rethink diversity.

Many practitioners were mindful of annotators' subjective interpretations, but the pursuit of diversity was largely seen as a way to achieve representation in a population and eventually reach a neutral outcome. Whenever diversity was considered in tasks, annotators were recruited to achieve an even spread across categories (*e.g.,* seeking proportionate distributions of gender, race, ethnicity, dialect, and more). Intersectionality provides us with a way of examining and problematising this perspective on diversity. Through the work of scholars such as McCall [61], we see how the perspective troublingly depends on social categories being static (that an individual's membership is permanent) and there being homogeneity within groups (that all group members have the same experiences). We also see groups of annotators being reduced to crude categories without deeper examination of how their identities and experience intersect, risking intensifying or amplifying exclusions or inequities.

Indeed, it is this latter point—foundational to intersectional scholarship [18]—that presents a fundamental problem for representationalist thinking and treatment of diversity recounted in our results. While this practice seeks, at best, a proportionate representation, it fails to acknowledge how the presumption of neutrality that accompanies it perpetuates precisely the kinds of biases we have just described. The approach to diversity and proportionate representation then assumes a neutral topology of categories, outside the power structures that marginalise, discriminate and exploit.

The work from Davis *et al.* [21] expands on this, explaining that such a "perspective derives from illusory cultural narratives that misalign with the world that is – a world in which discrimination is entrenched, elemental and compounding at the intersections of multiple marginalizations." [21, p. 2-3]. Davis *et al.* 's proposal for an "algorithmic reparation" is a point of departure in AI fairness scholarship, where *equity* and *reparation* (over fairness and equality) become the goals. As they conclude, the proposal is "geared towards building better systems and holding existing ones to account." [21, p. 8]

Our aim is to engage with this justice-directed, intersectional orientation in our rethinking of diversity. We take it to be a recognition of the ways those involved in AI/ML are already intervening in phenomena from positions of power and authority (not merely representing it), and should thus be accountable to the worlds being enacted [6, 39]. Set in these terms, the two recommendations we put forward suggest exploratory modes for engaging with intersectionality as a critical praxis [15]. They are intended for AI/ML practitioners to examine how the categorisation of annotators involved in producing a dataset influence a model, and where different annotators could offer alternative outcomes.

The first of our recommendations is the more modest one. It builds on McCall's notion of the *intercategorical* in intersectional thinking that speaks to the intersecting relationships of inequality present among more stable and predefined social categories (*e.g.,* class, race, ethnicity, disability, etc.). Imagined, here, is a tool that allows for exploratory visualisations, where virtual experiments might be run with different distributions of annotators across the predefined categories. This would enable practitioners to understand the impact of different annotator distributions (and intersections) on their models. Practitioners could also explore distributions and intersections that weighted marginalised identities, and inspect the impact on developing models. Such recommendations have close parallels to prior tools designed to examine data for biases [57, 72].

McCall identifies three ways of approaching intersectionality—the first, as above, examining inter-categorical complexity and two others: intra-categorical complexity and anti-categorical complexity. It is seeking to think with the second and third approaches, together, that we recommend a more radical proposal. Responding to the *intra-categorical* approach, we imagine the next steps for the tool described above would be to draw attention to the "neglected points of intersection" [61, p. 1774]. Additional considerations need to be given here to smaller subgroups that might be overlooked or appear marginal. This is to commit to a reparative approach to annotator diversity [21] where potentially overlooked identities are made salient for practitioners and given prominence, so as to have a greater influence on the AI/ML model. It might require additional time (and likely the involvement of other skill-sets *e.g.,* ethnography, participatory design and action research) to explore and make sense of the dynamic groupings and their influence on AI/ML models. Extending this further, an *anti-categorical* approach invites practitioners to take a step further, to resist the predefined categories of annotator identity but rather consider their lived experiences, organisational contexts and working conditions. Such lived experiences are beyond numeric capture, so how might they be taken into account? Our proposal, here, is to encourage a deeper analytical gaze, to promote in tools ways to question and re-examine normative groupings of identities and to look for what might lie beyond categorisation and demarcation.

6 LIMITATIONS

Our goal is not to provide an empirical account of the actual state of diversity among annotator pools, instead we focus on reporting the conceptions, approaches and logics for incorporating diversity-related considerations in ML projects. Our work can be extended by conducting research with platform providers on their practices and workflows in selecting and assigning annotators to ML tasks. We observe an attrition rate similar to prior studies that examine the practices of machine learning (*e.g.,* [23, 55]). We may have a selection bias among respondents already motivated to think about the diversity among their annotator pools, by recruiting interview and survey respondents through snowball sampling. We observe that our respondents were largely cognisant of the role of annotators' subjectivities, which may not be reflective of the general practice in ML.

All interviews and analysis were conducted over video and phone, due to the COVID-19 pandemic. As a result of travel restrictions, we were unable to include shadowing of work flows and contextual inquiry that would have otherwise been possible. However, we feel that the self-reported data practices and challenges have validity, and sufficient rigour and care was applied in covering the themes through multiple questions and solicitation of examples.

7 CONCLUSION

In this paper, we illustrate the status quo of annotator diversity in data practices through the combination of a survey and interviews with practitioners working on ML projects. In demonstrating how annotator diversity is treated as a minor figure among other competing priorities across the ML pipeline, we foreground an underlying but pervasive logic, namely *representationalist thinking*. Using findings from our survey and interviews, we elaborate on this thinking and show that it downplays the importance and value of diversity. To show how the representationalist thinking that pervades data practices might be challenged, and to invite a rethinking of diversity, we present three recommendations. Drawing inspiration from feminist scholarship, we propose (i) the rethinking of 'ground truth' in ML, proposing a move beyond 'majority voting' and towards annotator deliberation; (ii) a rethinking of 'bias', where we look beyond mitigation and instead aim to narrow the separation between ML practitioners and data annotators through direct communication and experimentation with worker-led participatory approaches; and, lastly (iii) the rethinking of annotator diversity, where we use intersectionality to shift attention away from static social categories and towards annotators' lived experiences. We call upon fellow researchers across disciplinary borders to think about and explore new approaches, and to experiment with new methods and tools, so that we centre diversity in ML data practices.

REFERENCES

[1] Jaimeen Ahn and Alice H. Oh. 2021. Mitigating Language-Dependent Ethnic Bias in BERT. In *EMNLP*.

[2] anonymous et al. in submission. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. (in submission), 1–20.

[3] Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1100–1105. https://doi.org/10.1145/3308560.3317083

[4] Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* 36, 1 (2015), 15–24.

[5] Karen Barad. 2003. Posthumanist performativity: Toward an understanding of how matter comes to matter. *Signs: Journal of women in culture and society* 28, 3 (2003), 801–831.

[6] Karen Barad. 2007. *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. duke university Press.

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

[8] Alice M Brawley and Cynthia LS Pury. 2016. Work experiences on MTurk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior* 54 (2016), 531–546.

[9] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[10] Graham Button and Wes Sharrock. 1997. *The Production of Order and the Order of Production: Possibilities for Distributed Organisations, Work and Technology in the Print Industry*. Springer Netherlands, Dordrecht, 1–16. https://doi.org/10.1007/978-94-015-7372-6_1

[11] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186. https://doi.org/10.1126/science.aal4230 arXiv:http://science.sciencemag.org/content/356/6334/183.full.pdf

[12] Scott Allen Cambo and Darren Gergle. 2022. Model Positionality and Computational Reflexivity: Promoting Reflexivity in Data Science. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 572, 19 pages. https://doi.org/10.1145/3491102.3501998

[13] L Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K Vishnoi. 2016. How to be fair and diverse? *arXiv preprint arXiv:1610.07183* (2016).

[14] Jesse Chandler, Gabriele Paolacci, and Pam Mueller. 2013. Risks and rewards of crowdsourcing marketplaces. In *Handbook of human computation*. Springer, 377–392.

[15] Patricia Hill Collins and Sirma Bilge. 2020. *Intersectionality*. John Wiley & Sons.

[16] Andrew Crabtree, Mark Rouncefield, and Peter Tolmie. 2012. *Doing design ethnography*. Springer, London.

[17] Andy Crabtree, Peter Tolmie, and Mark Rouncefield. 2013. "How many bloody examples do you want?": fieldwork and generalisation. In *Proceedings of the 13th European Conference on Computer Supported Cooperative Work ECSCW 2013*, Olav W. Bertelsen, Luigina Ciolfi, Maria Antonietta Grasso, and George Angelos Papadopoulos (Eds.). Springer Verlag, London, 1–20. https://doi.org/10.1007/978-1-4471-5346-7_1

[18] Kimberlé Crenshaw. 2013. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*. Routledge, 23–51.

[19] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)* 51, 1 (2018), 1–40.

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

[21] Jenny L Davis, Apryl Williams, and Michael W Yang. 2021. Algorithmic reparation. *Big Data & Society* 8, 2 (2021), 20539517211044808.

[22] Anne AH de Hond, Marieke M van Buchem, and Tina Hernandez-Boussard. 2022. Picture a data scientist: a call to action for increasing diversity, equity, and inclusion in the age of AI. *Journal of the American Medical Informatics Association* (2022).

[23] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. *arXiv preprint arXiv:2205.06922* (2022).

[24] Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On Measuring and Mitigating Biased Inferences of Word Embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (Apr. 2020), 7659–7666. https://doi.org/10.1609/aaai.v34i05.6267

[25] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[26] Ryan Drapeau, Lydia Chilton, Jonathan Bragg, and Daniel Weld. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 4. 32–41.

[27] Marina Drosou, HV Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. 2017. Diversity in big data: A review. *Big data* 5, 2 (2017), 73–84.

[28] Sina Fazelpour and Maria De-Arteaga. 2022. Diversity in sociotechnical machine learning systems. *Big Data & Society* 9, 1 (2022), 20539517221082027.

[29] Melanie Feinberg. 2017. A Design Perspective on Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2952–2963. https://doi.org/10.1145/3025453.3025837

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

[31] Lisa Gitelman. 2013. *Raw data is an oxymoron*. MIT press.

[32] Barney G Glaser. 1998. *Doing grounded theory: Issues and discussions*. Vol. 254. Sociology Press, Mill Valley, CA.

[33] Barney G Glaser and Anselm L Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge, London.

[34] Charles Goodwin. 1995. Seeing in depth. *Social studies of science* 25, 2 (1995), 237–274.

[35] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*. 1–19.

[36] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[37] Nitesh Goyal, Ian Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. *arXiv preprint arXiv:2205.00501* (2022).

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

[39] Ian Hacking et al. 1983. *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge university press.

[40] Ange-Marie Hancock. 2007. When multiplication doesn't equal quick addition: Examining intersectionality as a research paradigm. *Perspectives on politics* 5, 1 (2007), 63–79.

[41] Donna Haraway. 2020. Situated knowledges: The science question in feminism and the privilege of partial perspective. In *Feminist theory reader*. Routledge, 303–310.

[42] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Unintended Machine Learning Biases as Social Barriers for Persons with Disabilitiess. *SIGACCESS Access. Comput.* 125, Article 9 (mar 2020), 1 pages. https://doi.org/10.1145/3386296.3386305

[43] Lilly Irani. 2019. Justice for data janitors. In *Think in Public*. Columbia University Press, 23–40.

[44] Azra Ismail and Neha Kumar. 2018. Engaging solidarity in data collection practices for community health. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–24.

[45] Nataliya V Ivankova, John W Creswell, and Sheldon L Stick. 2006. Using mixed-methods sequential explanatory design: From theory to practice. *Field methods* 18, 1 (2006), 3–20.

[46] Farnaz Jahanbakhsh, Justin Cranshaw, Scott Counts, Walter S Lasecki, and Kori Inkpen. 2020. An experimental study of bias in platform worker ratings: The role of performance quality and gender. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[47] Florian Jaton. 2021. Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application. *Big Data & Society* 8, 1 (2021), 20539517211013569.

[48] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 306–316.

[49] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[50] Sonam Joshi. 2019. How artificial intelligence is creating jobs in India, not just stealing them | India News - Times of India. https://timesofindia.indiatimes.com/india/how-artificial-intelligence-is-creating-jobs-in-india-not-just-stealing-them/articleshow/71030863.cms. Accessed on 08/24/2021.

[51] Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1637–1648.

[52] Gary Klein, Jennifer K Phillips, Erica L Rall, and Deborah A Peluso. 2007. A data–frame theory of sensemaking. In *Expertise out of context*. Psychology Press, 118–160.

[53] Matthew Lease. 2011. On quality control and machine learning in crowdsourcing. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

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

[55] Michelle Seng Ah Lee and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

[56] Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UnQovering Stereotyping Biases via Underspecified Questions. In *EMNLP*.

[57] Sasha Luccioni, Yacine Jernite, and Meg Mitchell. 2021. Introducing the Data Measurements Tool: an Interactive Tool for Looking at Datasets. https://huggingface.co/blog/data-measurements-tool. (Accessed on 09/15/2022).

[58] Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. *arXiv preprint arXiv:2010.15149* (2020).

[59] Adam Marcus, David Karger, Samuel Madden, Robert Miller, and Sewoong Oh. 2012. Counting with the crowd. *Proceedings of the VLDB Endowment* 6, 2 (2012), 109–120.

[60] David Martin, Benjamin V Hanrahan, Jacki O'neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 224–235.

[61] Leslie McCall. 2005. The complexity of intersectionality. *Signs: Journal of women in culture and society* 30, 3 (2005), 1771–1800.

[62] Karishma Mehrotra. 2022. Human Touch. https://fiftytwo.in/story/human-touch/. (Accessed on 09/15/2022).

[63] Milagros Miceli and Julian Posada. 2022. The Data-Production Dispositif. *arXiv preprint arXiv:2205.11963* (2022).

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

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

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

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

[68] Michael Muller, Christine T. Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, and Casey Dugan. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 94, 16 pages. https://doi.org/10.1145/3411764.3445402

[69] Laura Nader. 1972. Up the anthropologist: Perspectives gained from studying up. (1972).

[70] Joaquin Navajas, Tamara Niella, Gerry Garbulsky, Bahador Bahrami, and Mariano Sigman. 2018. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour* 2, 2 (2018), 126–132.

[71] Jacki O'Neill, Stefania Castellani, Frederic Roulland, Nicolas Hairon, Cornell Juliano, and Liwei Dai. 2011. From Ethnographic Study to Mixed Reality: A Remote Collaborative Troubleshooting System. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (Hangzhou, China) *(CSCW '11)*. Association for Computing Machinery, New York, NY, USA, 225–234. https://doi.org/10.1145/1958824.1958859

[72] Google PAIR. 2022. Know Your Data. https://knowyourdata.withgoogle.com. (Accessed on 09/15/2022).

[73] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*. 39–48.

[74] Samir Passi and Steven Jackson. 2017. Data Vision: Learning to See Through Algorithmic Abstraction. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2436–2447. https://doi.org/10.1145/2998181.2998331

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

[76] Kathleen H. Pine and Max Liboiron. 2015. The Politics of Measurement and Action. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3147–3156. https://doi.org/10.1145/2702123.2702298

[77] Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On Releasing Annotator-Level Labels and Information in Datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*. 133–138.

[78] Mahima Pushkarna and Andrew Zaldivar. 2021. Data Cards: Purposeful and Transparent Documentation for Responsible AI. In *35th Conference on Neural Information Processing Systems*.

[79] David Randall, Richard Harper, and Mark Rouncefield. 2007. *Fieldwork for design: theory and practice*. Springer Science & Business Media, London.

[80] Negar Rostamzadeh, Diana Mincu, Subhrajit Roy, Andrew Smart, Lauren Wilcox, Mahima Pushkarna, Jessica Schrouff, Razvan Amironesei, Nyalleng Moorosi, and Katherine Heller. 2022. Healthsheet: Development of a Transparency Artifact for Health Datasets. *arXiv preprint arXiv:2202.13028* (2022).

[81] Nithya Sambasivan and Rajesh Veeraraghavan. 2022. The Deskilling of Domain Expertise in AI Development. In *CHI Conference on Human Factors in Computing Systems*. 1–14.

[82] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019. Understanding expert disagreement in medical data analysis through structured adjudication. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[83] Mike Schaekermann, Carrie J Cai, Abigail E Huang, and Rory Sayres. 2020. Expert discussions improve comprehension of difficult cases in medical image assessment. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[84] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–19.

[85] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536* (2017).

[86] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 614–622.

[87] Yaron Singer and Manas Mittal. 2013. Pricing mechanisms for crowdsourcing markets. In *Proceedings of the 22nd international conference on World Wide Web*. 1157–1166.

[88] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*. 254–263.

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

[90] Alex S. Taylor, Siân Lindley, Tim Regan, David Sweeney, Vasillis Vlachokyriakos, Lillie Grainger, and Jessica Lingel. 2015. Data-in-Place: Thinking through the Relations Between Data and Community. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2863–2872. https://doi.org/10.1145/2702123.2702558

[91] Geneviève Teil. 2012. No such thing as terroir? Objectivities and the regimes of existence of objects. *Science, Technology, & Human Values* 37, 5 (2012), 478–505.

[92] Lida Theodorou, Daniela Massiceti, Luisa Zintgraf, Simone Stumpf, Cecily Morrison, Edward Cutrell, Matthew Tobias Harris, and Katja Hofmann. 2021. Disability-First Dataset Creation: Lessons from Constructing a Dataset for Teachable Object Recognition with Blind and Low Vision Data Collectors. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 27, 12 pages. https://doi.org/10.1145/3441852.3471225

[93] Betty Van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. *arXiv preprint arXiv:1809.07572* (2018).

[94] Janet Vertesi and Paul Dourish. 2011. The value of data: considering the context of production in data economies. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 533–542.

[95] Ding Wang, Shantanu Prabhat, and Nithya Sambasivan. 2022. Whose AI Dream? In Search of the Aspiration in Data Annotation.. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 582, 16 pages. https://doi.org/10.1145/3491102.3502121

[96] Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*. 138–142.

[97] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *TACL* 6 (2018), 605–617. https://transacl.org/ojs/index.php/tacl/article/view/1484

[98] Peter Welinder and Pietro Perona. 2010. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 25–32.

[99] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating systems. *AI Now* (2019).

[100] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174230

[101] Xinyi Zhou and Reza Zafarani. 2018. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315* 2 (2018).