# A hunt for the Snark: Annotator Diversity in Data Practices

Shivani Kapania
kapania@google.com
Google Research
Bengaluru, India

Alex S. Taylor
alex.taylor@city.ac.uk
City, University of London
London, UK

Ding Wang
drdw@google.com
Google Research
Atlanta, USA

## ABSTRACT

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 16 semi-structured interviews and a survey with 44 AI/ML practitioners. While practitioners described nuanced understandings of annotator diversity, they rarely designed dataset production to account for diversity in the annotation process. 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 that such 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**.

## KEYWORDS

data annotation, data work, machine learning, ML datasets, diversity, annotator diversity, data production

**ACM Reference Format:**
Shivani Kapania, Alex S. Taylor, and Ding Wang. 2023. A hunt for the Snark: Annotator Diversity in Data Practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3544548.3580645

## 1 INTRODUCTION

A growing body of work examines diversity within datasets [28], particularly within the spaces of responsible AI research, including in discussions of fairness [7, 85] and harm mitigation [89]. Despite efforts to place data diversity centre stage, we know little about the role of diversity in producing human-annotated datasets. Data annotators perform critical sense-making labours of assigning meaning to data through labels, work that is crucial for building and evaluating many machine learning (ML) systems that exist today [38, 80, 83]. This paper reports on research seeking to understand how diversity among annotators is conceptualised and operationalised by practitioners who build models with annotated data. We draw on a combination of 44 survey responses and 16 in-depth, semi-structured interviews with ML practitioners (engineers, researchers, project/product managers). Through an analysis of these results, we also aim 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 standardised datasets representing 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. For instance, practitioners prioritised reducing the costs and complexity of annotation tasks, and accordingly, the capture of diversity among annotators was a non-essential and low-priority consideration.

Amidst this pragmatic approach to diversity, we came to see the importance of deeper knowledge-making or epistemic questions around the production of datasets. Examining diversity *in practice* revealed the logics that underlie the development of annotated datasets and building of AI/ML models. Set against such logics, diversity was seen as a question of whether an objective or neutral state could be fairly and accurately represented. Davis *et al.* [19] reveal an *algorithmic idealism* which assumes a meritocratic society in which demographic disparities can be neutralised. In our research, we found echoes of this view. We found datasets and models are built in a world where 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" [19, p. 2].

Seeking to critically examine such an orientation, our work is informed by a distinct theoretical framing. Specifically, our later discussion of the practices of data annotation builds on the idea of 'regimes of existence' introduced by sociologist Geneviève Teil [91]. Teil has used the term to examine 'terroir' in wine-making practices; as Teil recounts, terroir is a property of wine that is assessed through aroma and taste, and discussion between vintners. 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.).



*CHI '23, April 23–28, 2023, Hamburg, Germany*

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3580645

We find that Teil's [91] work offers us a starting point to further examine the logics prevailing 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 epistemic 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 operationalisation of annotator diversity in data practices.
(2) examine the underlying logic in data practices that relegates diversity-related considerations.
(3) propose shifts to rethink diversity in data practices, inspired by justice-oriented and intersectional scholarship, and invite a deeper examination of annotator diversity.

## 2 RELATED WORK

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

Prior scholarship on diversity in machine learning systems engages with two broad areas– those who design and build ML systems [49, 100], and the diversity considerations embedded in the data pipeline (*e.g.,* who represents and what gets represented [27, 85]). In the first of these two areas, we see engagement with discourses around the diversity of ML engineers, researchers, and subject-matter experts, and the composition of teams creating ML systems. Recent research points to the lack of diversity in the AI industry [100], and emphasises the benefits from diversifying teams as a way towards safer ML systems [20]. The implication here is that diverse teams lead to greater deliberation and thus more ethical systems [49]. In the second area–on diversity in the data pipeline– researchers have focused on who creates and is represented in the data [11]. Here, there have been efforts to understand the kinds of diversity in the instances within a dataset (*e.g.,* which regions or individual identities are covered) and the effects of accounting for such diversity on the performance of downstream ML models [85]. Researchers have argued that collecting more diverse data for relevant dimensions could lead to fairer ML models [7]. Fazelpour and De-Arteaga [28] emphasised the ways in which diversity-related considerations are embedded throughout the design of ML systems, regardless of whether they are actively recognised [28]. We contribute to this area of research by focusing on the conceptions of diversity within data annotator pools.

### 2.1 The Practices of Dataset Production

Substantial work in HCI, Science & Technology Studies (STS), and Fairness, Accountability and Transparency (FAccT) has established the ways in which data is shaped and defined through the contexts of production, just 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 [75].

Prior research in critical data studies has drawn particular attention to the sites of human intervention and discretionary choices that shape the work of data [61, 64, 66, 74]. Passi and Jackson [73] 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 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 [67, 72, 75]. Muller *et al.* [67] described how as part of this decision-making, practitioners engaged in compromises and trade-offs when considering the quality of labels alongside the available resources. To provide a framework to understand these choices, Cambo and Gergle [10] introduced the concepts of reflexivity and positionality for data science praxis to make the discretionary decisions more transparent.

Others in the community have examined the work practices, experiences, and backgrounds of individuals involved in data annotation work [6, 63, 95]. There has been a recent shift towards emerging actors that provide data annotation as a service, and as a consequence make the work of annotation more structured and organised [50, 60]. 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 power, and the hierarchical structure and control not only impacts the annotators' experiences but also their interpretations of the data [95]. They highlighted how a simplistic definition of 'data quality' as accuracy rate leaves little space for the annotators' expertise, knowledge and experiences.

Sambasivan and Veeraraghavan [81] demonstrated that even in settings where data production is reliant on expert fieldworkers (*e.g.,* farmers and radiologists) and their highly situated knowledge, practitioners still reduced field workers to data collectors, and attributed poor data quality to their work practices. Researchers have recognised that 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, 58]. We contribute to this literature by focusing on the practices of incorporating annotator diversity to further our understanding of data production for the creation of AI systems. Through unpacking the data annotation practices (as part of the machine learning process), we show in our findings how annotator diversity does not easily align with current ML workflow or practices, and the ways in which annotators' unique perspectives are under-valued and often actively minimised.

### 2.2 Factors affecting human-annotated data

There is a well-established, and growing, body of work that examines the factors affecting data annotations and their 'quality' (*e.g.,* incentives [87], interface [57], description [12], among others [17, 46]). In 2015, Aroyo and Welty [3] identified seven myths pervasive in the practice of data annotation, and proposed the theory of *crowd truth* based on the 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 [3]. 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 [3]. The typical discourse around disagreement, both in research, and practice has been to treat disagreement as noise [98].

A common thread across this body of literature demonstrates the role annotators' of socio-demographic backgrounds and lived experiences on their label assessments [23, 37, 51, 56, 96]. Researchers have turned to disagreement as a signal for deepening their understanding of the task and the data [2, 18, 76]. Prabhakaran *et al.* [76] demonstrate how label aggregation may introduce representational biases of individual and group perspectives and Davani *et al.* [18] 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 Human Computation (HCOMP), FAccT and HCI. Our work extends this body of research by examining the current praxis of ML model building and the practicalities which hinder nuanced diversity-related considerations.

## 2.3 Critical and analytic orientation

Below, we set out a critical and analytic orientation that helped us respond to current research and examine our study's results. This orientation has emerged through our readings of three threads of theory drawn from STS and feminist, intersectional scholarship– namely representationalist thinking, regimes of existence, and intersectionality.

*2.3.1 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 such as Hacking [39], Barad [4], Goodwin [34], Haraway [41]. The key features of this representationalist 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, *etc.*, as well as considerable training 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. 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 representationalist thinking.

*2.3.2 Regimes of existence.* Teil explores the prevailing logic of neutral or objective representation through *regimes of existence* [91]. Because terroir is evidently subjective—to do with individual tastes, 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].

This idea of "regimes of existence" in wine tasting and production may seem a long way from diversity in data annotation. Later, however, it will help us foreground the representationalist thinking in data practices and enable a critical examination of its prevalence as an overarching logic. It is through this regime of existence, brought into being through representationalist thinking and data practices, where we see how diversity and 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 alternative ways forward. Like Teil, Davis and their colleagues [19] 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 [19, 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".

*2.3.3 Intersectionality.* We draw on justice-oriented feminist and intersectional theory for inspiration towards annotator diversity [19, 59]. 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 [13].

We turn to the three distinct approaches outlined by Leslie McCall [59] – 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 contributing to exclusion and inequality. Building on STS and feminist scholarship, McCall [59] invites a greater criticality and investment in alternative imaginations of inequities in practice.

# 3 METHODOLOGY

Previous human-centered investigations concerning annotators' subjectivities, biases, and efficiency have primarily focused on annotators' perspectives [23, 63]. In this research, we focused on AI practitioners in order to understand their conceptions of annotator diversity [68]. To gain a holistic view of data annotation practices, we employed a mixed-methods approach with a sequential explanatory design [45]. This involved conducting a survey with the aim of eliciting broad-brushed and higher-level perspectives from a wide audience. The survey was followed by semi-structured interviews to investigate, in more detail, the ways in which practitioners apply their understanding of annotator diversity in practice. The qualitative data was used to elaborate and explain the survey results (*e.g.,* the rationale behind data annotation task design), and served as the foundation for our inquiry. We present our approaches to the survey and interviews in the following subsections.

## 3.1 Survey

The goal of the survey was to identify practitioners' perceptions of incorporating annotator diversity into their practice. We conducted the survey using an online questionnaire implemented in Qualtrics, and analysed responses from 44 respondents.

**Participant recruitment.** We recruited survey participants through multiple channels: advertising on social networks such as Twitter and LinkedIn and emailing direct contacts and mailing lists internal to our organisation. We began the survey by eliciting informed consent from respondents. No personally identifiable information was recorded about the respondents in accordance with our organisation's research privacy and ethics guidelines. The inclusion criteria for our survey was similar to the interviews, where we recruited practitioners who have collected or annotated data for an AI/ML project in the last 12-24 months. After the screening question, we were left with n = 78 participants. However, not all respondents completed all sections in the survey; thus we analysed a total of 44 responses (from those 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). Respondents were asked to answer questions by referring to a data annotation process they had recently been involved in. We began by asking respondents for 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. 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.** We computed a range of descriptive statistics using SPSS to better understand practitioners' approaches to diversity, and particularly the kinds of diversity they view as relevant, if at all. These included descriptives to questions presented with Likert scale response options and multiple choice answers (*e.g.,* the challenges in recruiting the desired pool of annotators). We focused on comparing the differences between the attributes which were used to recruit annotators, the kinds of information available to practitioners and what they would see as relevant attributes in the ideal scenario. Each of these questions had socio-demographic attributes such as the annotators' age, gender; expertise; location; and more. 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. We include direct quotes from our survey respondents in the Findings with the prefix 'S#', to differentiate them from our interview participants that were prefixed with 'P#'.

## 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 participants describing a recent project where they sought data annotations, end to end, to learn about their typical working method and AI development process. Our interviews focused on similar themes as the survey questionnaire, with additional questions on annotation documentation and reuse practices, and organisational structures and incentives within data annotation. Each session focused on the participants' practices, experiences and challenges with setting up data annotation tasks.

**Participant recruitment.** 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. 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 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; however, we report the primary AI technology and domain of application at the time of the interview.

**Interview moderation.** Given the geographical spread of our participants, the interviews were conducted online using video conferencing software. We scheduled sessions based on participants'

| 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**

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 analysis was inspired by and consistent with the ethnomethodological ethnographies in HCI [14, 15, 70]. 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 [8, 78]. In addition to analyzing interview transcripts, we examined 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 mobilising 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 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 annotate data 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' [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 cautious not to impose categories external to the data to codify the data. Ethnomethodological ethnographies are valuable in informing design [78], 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 AI practitioners face in incorporating diversity and therefore to steer the discussion around the diversity in data towards a more constructive justice-oriented direction.

# 4 FINDINGS

We begin by describing a typical data annotation workflow to illustrate the contexts of data work within the ML pipeline (section 4.1). Our study identified three approaches towards diversity in practice and the underlying logics motivating our participants' choices (section 4.2). We then present the barriers limiting a nuanced consideration of annotator diversity (section 4.3).

## 4.1 Data Annotation Workflow and Tasks

The first step in a typical data annotation workflow was identifying the data needs for the ML projects while considering the downstream applications (depicted in figure 1). The practitioners then selected the annotation service/platform and proceeded to design the annotation task with the help of the platform managers, starting with a pilot phase to test and iteratively improve the annotation guidelines. These improvements often included additional examples or edge cases to provide clearer instructions or clarifications to the annotators. While the annotators labelled the data in bulk, the practitioners monitored the process regularly to ensure the quality of the annotations. The monitoring often involved comparing the annotated data with a 'golden dataset' created by experts or the practitioners, and verified by the machine learning models they built.

In our study, most practitioners relied on internal data infrastructures to produce annotations. These platforms were used to recruit and manage annotators and facilitate annotation tasks. Our survey results showed a preference for these internal platforms over external annotator-facing marketplaces such as MTurk (echoing Wang *et al.* [95]). Of the 44 survey respondents, 25 used internal infrastructures and 8 outsourced to third-party vendors like Appen and Scale AI. The top factors influencing platform choice were cost (15), timeline (16), and platform quality (17) (reflected through proprietary tech, UX, support). Only 8 practitioners considered the diversity of data workers on the platform as a deciding factor.

The distribution of type of ML projects that our participants worked on skewed heavily towards language-based ML tasks. Among the survey respondents, 22 (50%) worked on language-related tasks

(classification or generation), eight practitioners identified 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, de-contextualising 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.

## 4.2 Approaches to Annotator Diversity

Below, we capture the varied perspectives to annotator diversity among the participants in our study, from some who considered diversity as irrelevant, to others who made efforts to accommodate it, even if only in partial ways. Most practitioners acknowledged the role of annotator subjectivities in the annotation process; some emphasised the importance of diversity in achieving a balanced view and bringing previously overlooked sub-populations into consideration. However, many went on to explain their decisions to not consider annotator diversity in setting up annotation tasks. Their primary focus was on achieving a threshold of quality, which was often measured against how closely an annotator followed pre-defined parameters and guidelines.

According to the survey results, a significant majority of respondents (75%) considered diversity to be a *somewhat* to *extremely* influential factor (>=3 on the Likert scale) in the quality of their annotated datasets. However, despite the perceived importance of various individual attributes and characteristics of annotators, very few of these factors were utilised in the recruitment process (see table 2). For instance, only 4 out of 15 practitioners who considered gender to be a relevant criterion included it in their selection criteria. Furthermore, even when additional information on annotator characteristics was available, it was rarely utilised in the recruitment process (table 2, column 'Available Information'). These findings suggest a potential disconnect between the perceived importance of diversity in annotation and actual recruitment practices.

Additionally, our interviews revealed significant contrast between participants' reflections on the concept of annotator diversity and its actual implementation. Annotator diversity was often understood as representative of a particular perspective or point of view. For example, annotators were selected based on the low-resource dialect they spoke or the high flood risk areas they lived in, using language and location as proxies for diversity. Very few participants actively recruited annotators based on their lived experiences, knowledge, or expertise as facets of diversity. In the rare cases where local knowledge and expertise were considered important, measurable criteria were applied to assess the annotators' expertise or knowledge. For instance, P6's motivation for selecting an annotator for a mapping project was to include individuals from underrepresented backgrounds and *"capture the other parts of the population"*. Similarly, P9 spoke of recruiting *"a person of X identity with Y knowledge"* to ensure diversity. Overall, interview participants demonstrated a view of diversity through the lens of categories and metrics, rather than tied to experiential knowledge and expertise. Next, we discuss practitioners' justifications for taking a representationalist approach to annotator diversity and prioritising measures of quality.

| 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 survey respondents for three questions: One, the attributes of data workers which could affect the annotations; Two, respondents' criteria for selecting annotators, and third, the kinds of information available to practitioners about the annotators of their dataset.**

*4.2.1 The pursuit of objective annotations.* Many practitioners invoked the domain and specific nature of their annotation task (*e.g.,* text style transfer) to justify de-prioritising annotator diversity. In both the survey and interviews, several practitioners stated that annotator diversity was not necessary when the task was considered objective. One survey respondent stated, *"our data had ground truth,"* to suggest how some annotation tasks can be objectively assessed. A belief in the objectivity of certain annotation tasks was often based on the notion that some questions have definitive answers and certain content can be definitively labelled. Annotator diversity was dismissed as irrelevant, particularly by those who described their annotation tasks as requiring subject expertise, such as detecting anomalies in chest X-rays or linguistic corpus detection tasks. P1 highlighted this perspective 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."*

We observed similar practices in contexts of quality checking, when some annotation results triggered additional quality checks. In cases of disagreement between annotators, resolvers (acting as experts) were brought in for final decision-making on the correct annotation. For a language understanding annotation task for voice assistants, P11 described how resolvers would handle discrepancies, either by choosing one of the existing annotations as correct or creating a new one from scratch. The resolvers' expertise was often determined by professional experience (typically greater number of years of experience in the field of annotation).

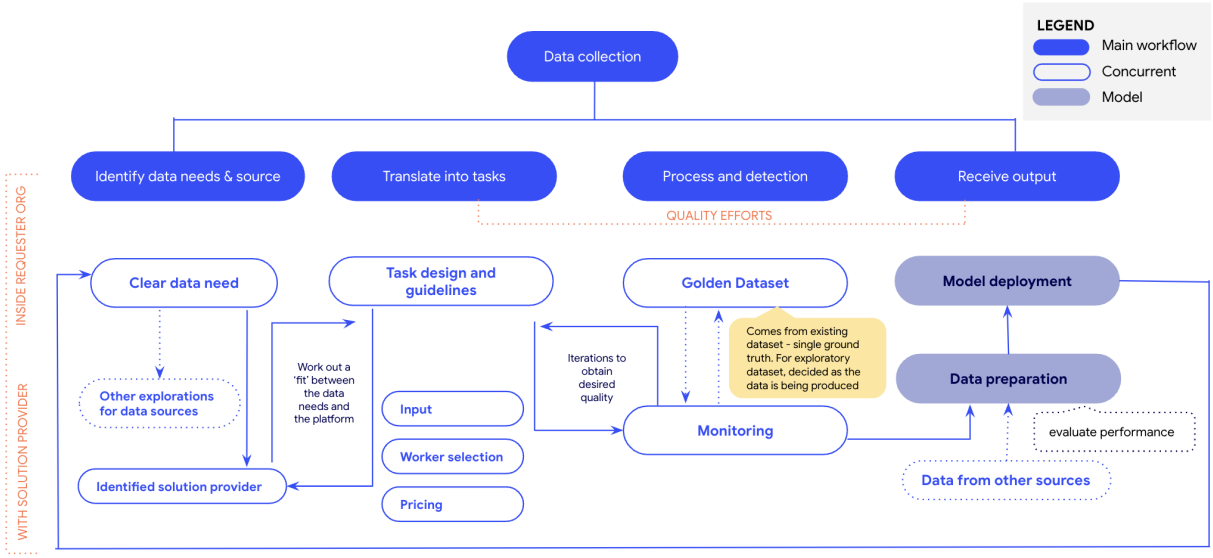


**Figure 1: The data annotation workflow for ML projects**

Practitioners considered the design of tasks as an effective intervention point to ensure objective annotations. They used training sessions and guidelines to teach annotators how to make correct judgements by closely following the instructions without deviation. Participants provided examples of their tasks (*e.g.,* river segmentation) which were intended to capture predetermined phenomena that could be made explicit through the annotation instructions. Detailed instructions were passed from practitioners to annotators through layers of quality checks conducted by platform leads or team managers. Instruction documents broke down annotation tasks into simple, repeatable sub-tasks that were *"very hard to answer in a biased way"* (P8), all as efforts 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" was considered a trainable skill and the training sessions and guidelines were essential for instructing annotators to "see" objectively.

*4.2.2 The attempt to remove bias.* Most practitioners recognised the complexity in accounting for annotators' diverse subjectivities, and used that as a justification to avoid over-complicating the goal of achieving useful and testable AI/ML outcomes. AI/ML workflows were designed to facilitate consistent evaluation across a range of source data, tasks, techniques, and annotators. This control over the development process helped participants compare the performance of AI/ML models and identify areas for optimisation. For example, P6 included questions with definitive answers in their dataset *"to have an easy way to evaluate the answers in the end."* Speaking of their 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, practitioners enacted mechanisms to circumvent complexity by limiting the plausible options in an annotation task and reduce ambiguity for evaluating a model's performance.

The desire to control ambiguity and complexity in data annotations extended to addressing annotator subjectivities, which were regularly framed as a form of 'bias' manifested in disagreements between annotators. In discussions about the implications of diversity, participants frequently conflated the concept of diversity with bias, viewing it not as something to be understood, but rather as a source of variability to be corrected or technically resolved. Practitioners made concerted efforts to minimise the effects of annotator diversity in order to make practical progress in modelling. To account for potential biases and differences in annotators' backgrounds and experiences, interventions were carefully implemented throughout the annotation process in order to eliminate disagreement and reduce annotator bias.

Many practitioners acknowledged that the data quality (*i.e.,* typically measured the accuracy rate) and disagreement could be a result of the flaws in the design of the annotation guidelines or annotation interfaces. However, a few participants 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 of this perspective: *"[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 handful of participants took the extra step of incorporating annotator information into their data production and model building processes. In an effort to engage with annotator diversity, they recruited annotators from

various backgrounds, such as a balanced gender ratio and multiple geographic locations. However, these practitioners struggled to determine and prioritise a set of relevant social categories for their specific task and domain. For example, P5, expressed the desire to capture a *'representation of every single person and every single dimension'* in their research on toxicity annotation and the rationale behind this attempt: *"otherwise we get biased annotations, and if we 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."* Seeking representative annotations was closely intertwined with efforts to eliminate bias and the pursuit of objectivity.

Practitioners often described how collecting a diverse range of perspectives can accurately reflect the real world, and that this representation and aggregation can achieve a neutral stance for building machine learning models. Participants who were attuned to the effects of diversity often attempted to capture differences in annotation patterns and establish a correlation between the patterns and the identities of the annotators. Many also noted the tension between representing the diverse global population and catering to their specific user base. P1 discussed their medical image annotation project where they procured and annotated 80% of their training data from the Global South, despite the eventual deployment of their models in the Global North due to varying data regulations. A few participants attempted to build systems that would effectively serve marginalised groups, but struggled to justify the additional resources (*e.g.,* time and budget) needed for these efforts from a business perspective.

The current state of data and machine learning practices did not actively support explorations of annotator diversity, limiting early attempts to incorporate diverse annotator perspectives into AI/ML models. When annotators provided label assessments from diverse viewpoints, practitioners were unable to distinguish minority opinions from 'noise' that 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. However, at an individual level, annotators were trained to adhere closely to task instructions and to move away from their individual interpretations. At a cohort level, the majority vote technique was commonly used to select the salient result, resulting in data that is neither diverse nor neutral.

## 4.3 Barriers to Incorporating Diversity

In our study, participants identified several challenges to accommodating annotator diversity. Firstly, participants reported a lack of access to information about annotators, hindering their ability to understand and account for annotators' unique perspectives and backgrounds. Additionally, the limited communication and collaboration between practitioners and annotators resulted in practitioners having minimal knowledge of annotators beyond their worker identification (worker-ID). Lastly, the lack of clear and actionable pathways to incorporate annotator socio-demographic information into the development and evaluation process further diminished the motivation for practitioners to prioritise diversity among data annotators.

*4.3.1 Lack of information about annotators.* Several practitioners expressed their lack of knowledge about the annotators working on their tasks. Often, the only information they had about these annotators came from website brochures or blog posts from the third-party data-labelling platforms they used for recruitment. This information was not specific to their projects, but rather was reported in aggregate and publicly available. In practice, annotation projects were run on a "good-faith basis" where the third-party platforms were trusted to satisfy the annotator recruitment requirements, and there were rarely any opportunities to confirm if annotators met the specified criteria.

Out of the 44 survey respondents, 19 reported having access to the annotators' geographic location and 15 had access to their language proficiency. The most commonly available annotator information was education level (11), followed by subject matter expertise (11) and gender (10). While 17 respondents did not face any challenges in obtaining the information they required, many others faced limited project timelines (8) and legal constraints (7), that limited practitioners' ability to better understand their annotators' background. Additionally, 18 respondents reported difficulties in accessing suitable annotators for their tasks. Overall, the respondents expressed a lack of control over the selection of their annotators.

While understanding annotators' backgrounds and considering diversity was important for improving the data production process, the incorporation of annotator information into data production had legal implications as well. Practitioners were wary of collecting sensitive personal information about annotators, such as their sexual or political orientations. This created a tension between protecting annotator privacy and gaining a more nuanced understanding of their backgrounds. 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 center. Even in countries where it can be done, there is no way [big tech company] would expose themselves to a potential PR nightmare of an article about how [big tech company] is selecting certain sexual orientations for data annotations. "*

A range of structures, such as legal, ethical, and corporate considerations, created obstacles to recruiting diverse annotators. As P8 noted, these considerations intersect to make *"selecting annotators to be diverse... impossible"*. Additionally, for P8, to effectively collect information about annotators, it was crucial to consider the well-being of annotators as a fundamental practice. The moral obligation to protect annotators from potential harm during repeated annotation tasks involving harmful content added another layer of complexity to the process.

*4.3.2 Separation of operations.* Most practitioners in our interviews relied on third-party platforms to annotate their datasets, and this separation of operations between the annotation platforms and the practitioners led to several challenges. Annotation platforms, while helpful in reducing the workload for practitioners, introduced a disconnect between the practitioners and the annotators. Most

annotation projects were mediated by a platform manager or team lead who facilitated the communication between the practitioners and annotators. As a result, direct communication between these groups was rare, if it happened at all. For example, P14 outlined the communication barriers in getting the most value out of their annotation process at a large tech company:

> *"The thing is that it was all contracted out externally. 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. "*

Challenges still existed even when direct communication between practitioners and annotators was possible. P5 discussed the difficulties in building trust and gathering information from annotators. In one project using MTurk, the annotators were uncomfortable with sharing personal information with P5's project team. The annotators wanted to understand why their information was being collected and how it would be used, but despite the explanation from P5, the annotators' distrust of MTurk (and consequently P5's team) persisted. Due to the use of intermediaries in facilitating interactions between practitioners and annotators, a three-way trust needs to be established, but the practitioners had limited power to rebuild trust between annotators and the platform. Therefore, factors such as establishing trust with annotators and determining appropriate pay came before considerations of diversity among annotators.

Geographical distance, time differences, and heavily-facilitated communication enforced and amplified the separation between practitioners and annotators. In order to avoid significant delays in communication, practitioners often had to resolve 'inconsistencies' in data labels on their own. While this was acknowledged as poor practice, it was largely dictated by tight turnaround times and business pressures. The lack of information and communication channels further exacerbated the separation between practitioners and their data annotators. Without knowing their annotators or having the opportunity to meet them, practitioners were more likely to consider them as interchangeable workers carrying out standardised tasks. As a result, only a few practitioners conceptualised the impact of annotator identities on their data and the importance of diversity within these identities.

*4.3.3 Competing priorities in machine learning development.* The status quo of driving practices in ML workflow presented challenges to conceptualising and operationalising annotators' diversity. Under the pressure of short-term development timelines, several practitioners noted how they had to prioritise curating larger datasets and building better-performing models over including a diverse group of annotators. Many participants worked in emerging and niche application areas, with a focus on exploring the limits and capabilities of ML models by testing new ideas and concepts. They had to prioritise *'hitting the ground running'* and reaching an 'MVP' (minimal viable product) before considering the specifics of their annotator pool. P11 articulated a sentiment echoed by many participants who struggled to find evidence for *justifying* annotator diversity: *"Even if [annotator diversity] could matter, it depends on where the project is— you want [the product] to work well for everyone but at an early stage you're just trying to make the product work. [...] It takes additional resources to address smaller user groups. There's a persistent question on whether it was even a priority to get models working for older people or speakers of dialect where there's not 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."*

For ML practitioners, data annotation was a part of the ML pipeline that must seamlessly integrate with the other components of the workflow, such as model building. P12 highlighted how data collection and annotation pipelines were often configured to support model building rather than advancing task understanding. The complexity of incorporating diverse annotator subjectivities stood in conflict with machine learning pipelines designed to produce definitive answers.

Both the platform managers and practitioners actively worked towards reducing the cost of annotation in setting up ML data production. As previously 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 [61, 64, 95]). We observed similar patterns in our study (*e.g.,* P14). According to P16 (platform account manager with a mid-size annotation company), annotation companies operated in a highly competitive environment where they typically hired workers from lower-wage countries to offer pricing lower than competitors. Practitioners also focused on minimising annotation costs, and thus, had limited control over which annotators are assigned to their projects. However, P6 noted the connection between appropriate incentives and 'data quality', positioning fair compensation as an obstacle only to high-quality data.

Practitioners discussed the complexities in setting up an annotation pipeline, including creating instruction documents, choosing a platform, and refining the process through multiple iterations. However, they also noted that the setup of annotation tasks was effort-intensive, taking away from valuable time and resources that could be spent on model building. In addition to 'competing' with model building for time and resources, data annotation procurement accounted for a significant portion of the operational costs of ML projects, making it difficult to justify repeating the process for any diversity-related concerns that surface post-hoc. Annotator diversity, which has yet to be proven to add value to model building, was often overlooked in favour of more pressing priorities, such as building a better-performing model quickly and cost-effectively.

## 5 DISCUSSION

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

We contend that practitioner viewpoints of annotator diversity operate within a wider *representationalist thinking*— in which both annotators and their observations are formalised and meant to be represented in neutral ways. An overarching idea of objectivity dominated our participants' views of annotator diversity. The quality of annotated data and need for diversity were gauged against

measurable and ostensibly objective standards. Here, the annotator is treated much like Hacking's microscope (Section 2.3 [39]), as an apparatus for achieving a representation of the world. The prevailing logic 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. In closing, we argue that including and prioritising historically marginalized perspectives is at odds with this idea of representativeness. We call on practitioners to be accountable for groups they represent and push to the margins in the search for a neutral representation.

Our findings paint a picture of practices where phenomena are being 'actively' seen and represented, and intervened in [34, 39, 91]. Rather than objectively or neutrally representing the world in some passive sense, we find practitioners make 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. We learned about systems and training being put in place, tasks being carefully designed and tightly constrained, and datasets being iteratively developed, all so annotators could see data in the 'right way' and produce datasets amenable to modelling and evaluation. 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 the context of its production. Like terroir in large-scale wine production, the subjectivities involved in annotation are rarely accounted for 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* captures this active intervening in phenomena; it invites 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. To think about diversity differently, we learn from Teil, is to *"[interpret] objects as distributed products—understood according to various protocols by different users and in different and endlessly renewed circumstances—enabling one to restore the plurality of objects and look for local agreements between the different points of view that compose them"* [91]. Towards this ambition, we present three critical shifts: a rethinking of ground truth, of bias, and of diversity.

## 5.1 Rethinking Ground Truth

Through our findings, we demonstrate the ways in which the diversity among annotators is deprioritised among various competing considerations, and their backgrounds and experiences are downplayed when set against the practical challenges of building AI/ML models. Practicalities such as cost control, model evaluation and product profitability cast diversity as a minor figure. We argue that as well as this practicality-driven mindset, the typical machine learning workflow limits a nuanced operationalisation of annotator diversity by requiring convergence. Annotation tasks are set up to arrive at a 'ground truth' label for training machine learning models, minimising the importance of annotators' subjectivities. The process of arriving at consensus, as the political theorist Chantal Mouffe describes, acts as a stabilisation of power and fundamentally entails a form of exclusion [65]. To account for diversity then is to explicitly design for conflicting interpretations in data annotation. Yet, practitioners' current approaches are in tension with diversity-related considerations, where enacting an 'ideal' annotator diversity is to eventually reduce subjective discretion and choice into a single outcome.

It is common practice in ML projects (notable among our participants) to collect multiple annotations for each labelled instance [86, 88] and then to apply a majority voting or averaging process [52] if the annotators do not converge. This disagreement can be quite substantial [36] in various machine learning tasks (*e.g.,* toxicity detection [93], medical diagnosis [82], misinformation [102]). While research indicates that such disagreement could, in fact, act as a signal for identifying issues with the task construction [3], many practitioners viewed disagreement as undesirable, impeding production of 'high quality data'. 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. Additionally, in data labelling tasks, both on platform-based and within private annotation firms, the data workers frequently discuss cases of ambiguity amongst themselves, 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, 69]. ML practitioners can view discussions among annotators as tainting the dataset by collusion, and such conversations about labels are neither supported nor documented by current annotation tools [63, 64, 95].

The first step to manage annotator subjectivities would be to acknowledge that data production is fundamentally a collective and interpretive task, and there are likely cases where individual annotators will not conform with the 'majority perspective' [35]. Here, practitioners would do well to consider approaches that capture the nuances in disagreements and preserve minority perspectives [18]. Prabhakaran *et al.* [76] demonstrated how aggregating labels obfuscates socio-cultural backgrounds. Those who develop datasets should consider preserving and attaching individual annotators' labels with each instance to enable analyses and reusability for downstream applications [76]. Practitioners could turn towards a design of data production processes that embraces a *"commitment to discovering and inventing ways to express and enable productive dissent and contestation" [24]. Annotation tools and processes that support a multiplicity of voices offer a productive starting point for engaging with the diverse annotator positions in a pluralist society.*

## 5.2 Rethinking Bias

Our research takes inspiration from an emerging area of scholarship that demonstrates how individual annotator identities influence

the annotation task (*e.g.,* age [23], gender [9, 97], sexual orientation [37], race/ethnicity [22, 54] and disability [1, 42]). In short, data annotation (like other data practices [90]) is situated within particular social and cultural contexts. Among our participants, there were no substantive attempts to explore how contexts and any accompanying power relations might influence annotation or efforts 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 to both the conditions in which annotators work, and their lived experiences and aspirations [63, 95].

A range of scholarship has provided valuable starting points for exploring this position further and the influence on practice. Miceli *et al.* [62], for instance, discuss the ways in which removing annotator bias (as a reflection of poor quality) should not be a universal goal. Instead, how might we reflect on the underlying causes of such differences? In practice, we found individual annotator subjectivities were seen through the lens of 'bias mitigation'. As long as the prevailing assumption in the field of machine learning is that biases can be identified, corrected and neutralised through aggregation of multiple perspectives, the dissonance between research and practice will persist.

Providing potential ways forward, researchers have proposed documentation artefacts (*e.g.,* Datasheets for Datasets [30], Data Cards [77], Data Statements [5]) and archival artefacts of the decisions made [48] during the annotation process were proposed as a mechanism to foster deliberative accountability [74], to make explicit the tacit knowledge in data work [63, 73] and to foster fair reparation in ML building [19]. However, these artifacts are often created at the tail-end of the project, after the data annotation has been completed [79]. While documentation artifacts can promote transparency, they do not effectively address the representationalist thinking that underlies data practices and as a consequence the impoverished considerations of annotator diversity. Documentation alone does not prevent practitioners from intentionally excluding conflicting viewpoints. To intervene and account for diverse annotator perspectives, approaches must take account of differences and subjectivities from the outset.

Our findings suggest opportunities for addressing the separation between practitioners and annotators by supporting direct communication within the annotation tool. A small number of research projects have started to explore the role of co-creation in AI fairness [101], and in dataset production and curation [29, 92]. What might a participatory approach to annotation look like in practice? Instead of practitioners imposing viewpoints on annotators through hierarchical organisational structures, we invite practitioners to explore collaborative approaches and co-created labelling setups (*e.g.,* co-production of annotation artifacts), together with the annotators.

## 5.3 Rethinking Diversity

Lastly, we would like to return to the theorising of justice-oriented intersectionality [19, 59] to rethink diversity. Many practitioners were mindful of annotators' subjective interpretations, but the pursuit of diversity was largely seen as a way to achieve representativeness in a population. When diversity was considered in tasks, annotators were recruited to seek a proportionate distribution across one or more categories (*e.g.,* gender, race, ethnicity, dialect). Intersectionality provides us with a way of examining and problematising this perspective on diversity. Through the work of scholars such as McCall [59] and Hancock [40], we see how the perspective troublingly depends on assumptions of static social categories (that an individual's membership is permanent) and 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 experiences intersect, risking intensifying exclusions or inequities.

Indeed, it is this latter point—foundational to intersectional scholarship [16]—that presents a fundamental problem for representationalist thinking and the 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 perpetuates precisely the kinds of biases we have just described. The approach to diversity and proportionate representation assumes a neutral topology of categories, outside the power structures that marginalise, discriminate and exploit. The work from Davis *et al.* [19] 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." [19, 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" [19, p. 8].

Our aim is to engage with this justice-directed, intersectional orientation. 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 thus should be accountable to the worlds being enacted [25]. Set in these terms, we put forward recommendations for exploratory modes of engaging with intersectionality as a critical praxis [13]. These suggestions are intended to help AI/ML practitioners examine how annotators involved in dataset production influence downstream models, 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 pre-defined social categories (*e.g.,* class, race, ethnicity, disability, etcetera). Imagined, here, is a tool that allows for exploratory visualisations, where virtual experiments might be run with different distributions of annotators across the pre-defined 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 [55, 71].

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" [59, 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 [19] 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 AND FUTURE WORK

We may have a selection bias for respondents already motivated to consider the diversity among their annotator pools, by recruiting interview and survey respondents through snowball sampling. We observe that many participants were cognisant of the role of annotators' subjectivities, which may not be reflective of the general practice in ML. A recognition of annotators' subjectivities might be influenced by the education background, professional experience, culture norms and workflows of the practitioner. Future studies might consider drawing on generative frameworks such as 'community of practice' [99] that offer a lens to examine the diversity of annotators in machine learning praxis.

Our goal is not to provide an empirical account of the actual state of diversity among annotator pools, instead we focus on reporting the approaches and logics for 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 a survey drop-off rate (from n = 78 to n = 44), similar to prior studies that examine the practices of machine learning (*e.g.,* [21, 53]) through research with practitioners. As a result of COVID-19 restrictions, we were unable to include shadowing of workflows and contextual inquiry that would have otherwise been possible. However, self-reported data practices and challenges have validity, and we applied sufficient rigour and care to cover 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*, that 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 and intersectional scholarship, we propose (i) the rethinking of 'ground truth' in ML, proposing a move beyond 'majority voting' and towards enabling 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 invite researchers across disciplinary borders to explore new approaches and to experiment with new methods and tools, so that we can centre diversity in ML data practices.

## REFERENCES


[1] Jaimeen Ahn and Alice Oh. 2021. Mitigating Language-Dependent Ethnic Bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 533–549. https://doi.org/10.18653/v1/2021.emnlp-main.42

[2] 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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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

[9] 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

[10] 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

[11] L. Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K. Vishnoi. 2016. How to be Fair and Diverse? *arXiv e-prints*, Article arXiv:1610.07183 (Oct. 2016),

5 pages. https://doi.org/10.48550/arXiv.1610.07183 arXiv:1610.07183 [cs.LG]
[12] Jesse Chandler, Gabriele Paolacci, and Pam Mueller. 2013. *Risks and Rewards of Crowdsourcing Marketplaces.* Springer New York, New York, NY, 377–392. https://doi.org/10.1007/978-1-4614-8806-4_30
[13] Patricia Hill Collins and Sirma Bilge. 2020. *Intersectionality.* John Wiley & Sons.
[14] Andrew Crabtree, Mark Rouncefield, and Peter Tolmie. 2012. *Doing design ethnography.* Springer, London.
[15] 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
[16] 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, New York, NY, USA, 23–51.
[17] 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.
[18] 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.
[19] Jenny L Davis, Apryl Williams, and Michael W Yang. 2021. Algorithmic reparation. *Big Data & Society* 8, 2 (2021), 20539517211044808.
[20] Anne A H 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* 29, 12 (09 2022), 2178–2181. https://doi.org/10.1093/jamia/ocac156 arXiv:https://academic.oup.com/jamia/article-pdf/29/12/2178/47027843/ocac156.pdf
[21] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22).* Association for Computing Machinery, New York, NY, USA, 473–484. https://doi.org/10.1145/3531146.3533113
[22] 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
[23] Mark Diaz, 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* (Montreal QC, Canada) *(CHI '18).* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173986
[24] C. Disalvo. 2015. *Adversarial Design.* MIT Press, Cambridge, MA, USA. https://books.google.co.in/books?id=gdcGEAAAQBAJ
[25] Lynn Dombrowski, Ellie Harmon, and Sarah Fox. 2016. Social Justice-Oriented Interaction Design: Outlining Key Design Strategies and Commitments. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (Brisbane, QLD, Australia) *(DIS '16).* Association for Computing Machinery, New York, NY, USA, 656–671. https://doi.org/10.1145/2901790.2901861
[26] Ryan Drapeau, Lydia Chilton, Jonathan Bragg, and Daniel Weld. 2016. MicroTalk: Using Argumentation to Improve Crowdsourcing Accuracy. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 4, 1, 32–41. https://doi.org/10.1609/hcomp.v4i1.13270
[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] Vinitha Gadiraju, Kane Shaun, Sunipa Dev, Alex Taylor, Ding Wang, Robin Brewer, and Emily Denton. In submission. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. , 20 pages.
[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] L. Gitelman. 2013. *Raw Data Is an Oxymoron.* MIT Press, Cambridge, MA, USA. https://books.google.co.in/books?id=oARaHF4D6h0C
[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 *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 115, 19 pages. https://doi.org/10.1145/3491102.3502004
[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* (Yokohama, Japan) *(CHI '21).* Association for Computing Machinery, New York, NY, USA, Article 388, 14 pages. https://doi.org/10.1145/3411764.3445423
[37] Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 363 (nov 2022), 28 pages. https://doi.org/10.1145/3555088
[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. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. Vol. 14. Feminist Studies, Inc., 575–599. http://www.jstor.org/stable/3178066
[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, New York, NY, USA, 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* (Honolulu, HI, USA) *(CHI '20).* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376860
[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* (Barcelona, Spain) *(FAT* '20).* Association for Computing Machinery, New York, NY, USA, 306–316. https://doi.org/10.1145/3351095.3372829
[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* (San Francisco, California, USA) *(CSCW '16).* Association for Computing Machinery, New York, NY, USA, 1637–1648. https://doi.org/10.1145/2818048.2820016
[52] Matthew Lease. 2011. On Quality Control and Machine Learning in Crowdsourcing. In *Proceedings of the 11th AAAI Conference on Human Computation (AAAIWS'11-11).* AAAI Press, Palo Alto, CA, USA, 97–102.
[53] 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* (Yokohama, Japan) *(CHI '21).* Association for Computing Machinery, New York, NY, USA, Article 699, 13 pages. https://doi.org/10.1145/3411764.3445261
[54] Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing Stereotyping Biases via Underspecified Questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020.* Association for Computational Linguistics, Online, 3475–3489. https://doi.org/10.18653/v1/2020.findings-emnlp.311

[55] 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).

[56] Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting Stance in Media On Global Warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3296–3315. https://doi.org/10.18653/v1/2020.findings-emnlp.296

[57] 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.

[58] 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* (Baltimore, Maryland, USA) *(CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 224–235. https://doi.org/10.1145/2531602.2531663

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

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

[61] Milagros Miceli and Julian Posada. 2022. The Data-Production Dispositif. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 460 (nov 2022), 37 pages. https://doi.org/10.1145/3555561

[62] 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.

[63] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 115 (oct 2020), 25 pages. https://doi.org/10.1145/3415186

[64] 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* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 161–172. https://doi.org/10.1145/3442188.3445880

[65] Chantal Mouffe. 1999. Deliberative democracy or agonistic pluralism? *Social research* (1999), 745–758.

[66] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300356

[67] 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

[68] Laura Nader. 1972. Up the anthropologist: Perspectives gained from studying up.

[69] 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.

[70] 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

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

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

[73] 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

[74] 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.

[75] 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

[76] 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*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 133–138. https://doi.org/10.18653/v1/2021.law-1.14

[77] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1776–1826. https://doi.org/10.1145/3531146.3533231

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

[79] 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. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1943–1961. https://doi.org/10.1145/3531146.3533239

[80] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 39, 15 pages. https://doi.org/10.1145/3411764.3445518

[81] Nithya Sambasivan and Rajesh Veeraraghavan. 2022. The Deskilling of Domain Expertise in AI Development. 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 587, 14 pages. https://doi.org/10.1145/3491102.3517578

[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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376290

[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.

[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* (Las Vegas, Nevada, USA) *(KDD '08)*. Association for Computing Machinery, New York, NY, USA, 614–622. https://doi.org/10.1145/1401890.1401965

[87] Yaron Singer and Manas Mittal. 2013. Pricing Mechanisms for Crowdsourcing Markets. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13)*. Association for Computing Machinery, New York, NY, USA, 1157–1166. https://doi.org/10.1145/2488388.2488489

[88] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew 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*. Association for Computational Linguistics, Honolulu, Hawaii, 254–263. https://aclanthology.org/D08-1027

[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*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3465416.3483305

[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. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, Brussels, Belgium, 33–42. https://doi.org/10.18653/v1/W18-5105

[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* (Hangzhou, China) *(CSCW '11)*. Association for Computing Machinery, New York, NY, USA, 533–542. https://doi.org/10.1145/1958824.1958906

[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*. Association for Computational Linguistics, Austin, Texas, 138–142. https://doi.org/10.18653/v1/W16-5618

[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, San Francisco, CA, USA, 25–32. https://doi.org/10.1109/CVPRW.2010.5543189

[99] Etienne Wenger et al. 1998. Communities of practice: Learning as a social system. *Systems thinker* 9, 5 (1998), 2–3.

[100] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating systems.

[101] 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

[102] Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surveys* 53, 5, Article 109 (sep 2020), 40 pages. https://doi.org/10.1145/3395046