The Perspectivist Paradigm Shift: Assumptions and Challenges of Capturing Human Labels

Eve Fleisig[1] **Su Lin Blodgett**[2] **Dan Klein**[1] **Zeerak Talat**[3]
[1]University of California Berkeley [2]Microsoft Research Montréal
[3]Mohamed Bin Zayed University of Artificial Intelligence
{efleisig,klein}@berkeley.edu
sulin.blodgett@microsoft.com z@Zeerak.org

Abstract

Longstanding data labeling practices in machine learning involve collecting and aggregating labels from multiple annotators. But what should we do when annotators disagree? Though annotator disagreement has long been seen as a problem to minimize, new *perspectivist* approaches challenge this assumption by treating disagreement as a valuable source of information. In this position paper, we examine practices and assumptions surrounding the causes of disagreement—some challenged by perspectivist approaches, and some that remain to be addressed—as well as practical and normative challenges for work operating under these assumptions. We conclude with recommendations for the data labeling pipeline and avenues for future research engaging with subjectivity and disagreement.

1 Introduction

When developing human-labeled data for machine learning (ML) tasks, labels for each example are often obtained by collecting annotations from multiple annotators, which are then aggregated to provide a single ground truth label per example. However, a line of recent work has illustrated that annotators disagree for many reasons, and that capturing this disagreement can improve model performance and calibration (Fornaciari et al., 2021; Baan et al., 2022), surface minority voices (Prabhakaran et al., 2021), and uncover task ambiguities (Balagopalan et al., 2023; Parrish et al., 2023). Researchers have begun to ask: What should we do when people disagree? How can (or should) our datasets and models account for different opinions?

We argue that this new wave of research—which, following Basile et al. (2021), we refer to as the *perspectivist turn*—constitutes a paradigm shift in data collection for ML and offers an opportunity to systematically examine the changing landscape. In this position paper, we examine practices and assumptions across papers regarding how data is collected from multiple annotators, discuss challenges raised by these approaches, and provide recommendations for rethinking data labeling when annotators disagree. We offer our own syntheses of observed practices and assumptions in natural language processing (NLP), as well as observations drawn from meta-analyses of ML research more broadly.

We first examine what has changed under this paradigm shift: we examine each paradigm's assumptions about the *causes* and *nature* of disagreement, and the *practical challenges* that arise when operating under each set of assumptions. We then explore what has not changed, identifying *normative challenges*—questions and assumptions about labeling not yet taken up in this shifting landscape. Finally, we offer recommendations for designing data labeling processes that better account for annotator disagreement, and avenues for future research. By charting these shifting assumptions and practices, we aim to surface the ways in which each paradigm succeeds, or fails, to account for the rich tapestry that disagreement can offer.

2 The Longstanding Paradigm

We characterize the *longstanding paradigm of data labeling* as work that collects labels on a data instance from annotators and aggregates them with the goal of capturing underlying ground truth labels (Snow et al., 2008; Nowak and Rüger, 2010).[1] By contrast, work in the *perspectivist paradigm* treats variation among annotator labels as a source of meaningful information (Basile et al., 2021; Plank, 2022). We first examine assumptions about the causes of disagreement and challenges faced under the longstanding paradigm.

[1]Common aggregation strategies include majority vote over labels for binary classification tasks and averaging labels for tasks that use Likert scale ratings. We group these under the collective umbrella of "averaging."

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2279–2292

2.1 What Causes Disagreement?

In this section, we examine longstanding practices and assumptions about the causes and nature of disagreement, which the perspectivist paradigm challenges. Under the longstanding paradigm, annotator disagreement is often characterized as an issue of label quality, particularly when crowdsourcing labels (Nowak and Rüger, 2010; Artstein, 2017). Disagreement is often attributed to "subjective," confusing, or inherently ambiguous tasks (Aroyo and Welty, 2014), or to low-quality (inexperienced, uninformed, or biased) annotators (Hsueh et al., 2009; Nowak and Rüger, 2010). Because spam or inconsistency is common, collecting multiple labels per example and measuring inter-annotator agreement can serve as a guarantee of data quality.

Perspectivist approaches have re-evaluated several of these practices and their underlying assumptions. Here, we discuss three such practices: attributing disagreement to bias or ineptitude, requesting labels out of context, and restricting discussion of disagreement to "subjective" tasks.

Assumption: Disagreement is due to biased or inept annotators and thus noise to eliminate. In a review of annotator diversity in data labeling, Kapania et al. (2023) find that ML practitioners "conflated...diversity with bias, viewing it...as a source of variability to be corrected or technically resolved" and attributing it to "unsatisfying work quality, or worse, questionable work ethics." Synthesizing previous work, we argue that this assumption stems from (1) a conflation between "bias" in the statistical sense and societal sense, and (2) a belief that meaningful differences of opinion only arise due to technical expertise or work quality.

Disentangling annotator "bias." Recent work exhibits a conflation between two senses of the word *bias*: (i) a statistical sense (as in "bias-variance tradeoff"), meaning the difference between the expected value of an estimator and its actual value, and (ii) a psychological or societal sense, meaning prejudicial discrimination against a person or group (e.g., Narimanzadeh et al., 2023; Hube et al., 2019; Li et al., 2020). Kapania et al. (2023) find that practitioners "were unable to distinguish minority opinions from 'noise' that deviated from instructions." If the mean label m of a group of annotators is considered the ground truth for a data example, then an annotator whose label is far from m is statistically biased. Yet, we argue, it does not follow that the annotator must be societally biased: for example, if the annotator is a member of an affected community who knows more than other annotators about the context of the example being labeled, it may instead be the mean label that is societally biased. Since disagreement manifests as statistical bias, which is equated with societal bias, all disagreement is undesirable under this assumption.

Disentangling "expertise." Though machine learning acknowledges the value of "expert annotators"—generally people with prior training in an area, or quantifiable knowledge such as fluency in a language—Kapania et al. (2023) find that annotators are rarely recruited "based on their lived experiences, knowledge, or expertise as facets of diversity." When lived experience is not seen as a legitimate source of expertise, disagreement on that basis is more easily ascribed to "bias" than well-informed but different views. Multiple studies have indeed found that annotator opinions vary based on factors related to lived experience, including demographics, political views, and community membership (Patton et al., 2019; Larimore et al., 2021; Sap et al., 2019). These findings indicate that lived experiences shape people's judgments, and therefore that "non-experts" with different backgrounds can disagree without being "low-quality" annotators. In turn, this suggests that such disagreement ought to be treated as meaningful in its own right.

Practice: Annotators rarely receive task context. Data labeling tasks often give annotators minimal context when labeling data (Fortuna et al., 2022), thus implicitly treating such context as irrelevant to annotators' decision making. Nevertheless, context can greatly change annotator behavior; for example, in hate speech detection, giving annotators context about text authors' probable race or language variety changes annotator judgments (Sap et al., 2019), while in machine translation, detailed instructions increase annotator agreement (Popović, 2021). Information given to annotators about how labels will be used also affects their judgments, even with no change in the data being labeled: Balagopalan et al. (2023) ask annotators to do the same task framed as a factual classification or as a judgment of whether a norm was violated (e.g., whether an outfit matches a description or breaks a dress code based on the same description), and find that annotators are "less likely to say that a rule has been violated than to say that the relevant factual features...are present." This suggests that annotators account for

potential consequences that are salient to them—e.g., penalizing people for breaking a dress code. Thus, annotators' assumptions about task context—which under the longstanding paradigm have typically remained implicit—may represent an overlooked source of meaningful disagreement.

Indeed, recent work has indicated that annotators are aware of the impact of the assumptions they make on decontextualized tasks, and sometimes request more granular instructions and context. Surveying Mechanical Turk workers on the types of information that help on confusing tasks, Huang et al. (2023) find that over 50% of annotators want more context on annotations, and over two-thirds believe that knowing the purpose of the labeling task would help them.

Assumption: Disagreement is limited to "subjective" tasks. It is tempting to assume that disagreement is limited to tasks based on personal opinions, such as those that involve the quality of art or text, or those that touch on sociocultural norms, such as offensive speech detection. Yet disagreement arises even in seemingly clear-cut tasks, such as natural language inference (NLI) (Pavlick and Kwiatkowski, 2019; Jiang and de Marneffe, 2022) and semantic textual similarity (Wang et al., 2023). Geva et al. (2019) find that responses on NLI and question answering tasks vary enough by person that annotator-specific models improve downstream task performance, while Parrish et al. (2023) find that in image classification, issues such as differing names for the same objects in different regions and differing interpretations of a task (e.g., whether a picture of a bird counts as a "bird") result in disagreement. Basile et al. (2021) note other causes of annotator disagreement, such as task complexity, annotator proficiency at the task, and cognitive biases. These varied factors suggest there is no clear set of tasks that admit no subjectivity or disagreement.

2.2 Practical Challenges under the Longstanding Paradigm

Having examined this paradigm's assumptions about the causes of disagreement, in this section we accept its goal—to capture a single underlying ground truth annotation per example, ideally the broader population's opinion—at face value and examine technical challenges towards achieving it in practice. We argue that even under the longstanding assumption that capturing such a label is possible and that annotator disagreement does not reflect meaningfully differing opinions, a number of technical challenges across different stages of data labeling continue to make capturing labels difficult. Specifically, we suggest that **collected labels are not a good proxy for the stakeholder population's views**, and that **diverse recruitment is not enough**, because even uniform sampling of the annotator pool with aggregated labels inaccurately models the broader population for several reasons:

Unrepresentative annotator pools. The demographics of crowdworking platforms such as Mechanical Turk are not representative of most populations of interest (including system users, affected stakeholders, or even the population of the regions from which crowdworkers are recruited). For example, U.S. Mechanical Turk workers are disproportionately white and young compared to the general U.S. population (Pew Research Center, 2016).

Sample error. When small numbers of annotators are recruited relative to the population size, the average of their ratings is likely to be farther from the average of the full population (Narimanzadeh et al., 2023; Geva et al., 2019). This effect is exacerbated when few annotators annotate each data item, making it less likely that an annotator with relevant background is assigned to annotate a particular item. Moreover, under current crowdsourcing practices, there is often no limit on how many annotations one person may do, resulting in datasets that may reflect only the opinions of the most prolific annotators (Geva et al., 2019).

Aggregation treating minority opinions as noise results in miscalibrated models. Narimanzadeh et al. (2023) note that majority voting always discards data from "minority raters holding less popular opinions" and moves the estimated mean further from the true population mean by non-randomly discarding ratings. Aggregated judgments have disproportionately high agreement with white annotators (Prabhakaran et al., 2021), reflecting the fact that aggregated labels typically reflect the opinions of groups with higher representation and minimize the representation of minority opinions. As a result, downstream models are often miscalibrated with respect to diversity of opinions between annotators (Baan et al., 2022).

2.3 Normative Challenges

While the perspectivist literature has identified and challenged a number of longstanding assumptions

about disagreement, several longstanding assumptions remain only partially addressed even in perspectivist work.

Sometimes, there is no ground truth. The existing paradigm of data labeling implicitly imagines annotation as a process of uncovering the single "ground truth" label for the data, using annotators as noisy approximators. However, findings across a range of tasks suggest that there is often no such ground truth. This may occur because the task is underspecified (e.g., the intent of the data labeling process is not clear enough to the annotators to eliminate all ambiguity); the fact that disagreement occurs even in tasks not usually seen as "subjective" highlights the difficulty of removing all potential ambiguity. Alternatively, it may occur because reasonable people who fully understand the intent of the annotation could have different opinions, leaving the "ground truth" undefined.

Averaging labels loses information about a population's values. Averaging opinions, e.g., via majority vote, has a millennia-old history as a way of democratically aggregating views on an issue (Boegehold, 1963). However, naively averaging data labels encounters serious issues in practice. People are not equally well-informed or culturally grounded for all tasks, nor do they face equal consequences from model decisions. Expertise—including less quantifiable factors such as lived experience and sociocultural background—is key for many tasks, particularly when a task affects a particular community. Yet averaged labels ignore such considerations, resulting in lower-quality datasets that may disregard those who are most affected.

3 The Perspectivist Turn

Perspectivist efforts argue that longstanding approaches are insufficient when (1) annotators frequently disagree in ways that are important to capture, and (2) even with diverse annotator recruitment, aggregate labels often fail to adequately represent the true population's opinions. Approaches in the perspectivist turn include training with annotators' individual labels or pertinent details about the annotators and explicitly modeling individual annotators' behavior (e.g., Davani et al., 2022; Gordon et al., 2022; Plepi et al., 2022; Sachdeva et al., 2022); training with probability distributions over labels (Fornaciari et al., 2021; Uma et al., 2020); calibrating to variance between annotators (Baan et al., 2022); collecting labels from many annotators (Nie et al., 2020; Aroyo et al., 2023); and investigating causes of disagreement (e.g., Goyal et al., 2022; Larimore et al., 2021; Pei and Jurgens, 2023).[2] Here, we explore how perspectivist approaches conceptualize the causes and nature of disagreement, as well as emerging practical and normative challenges.

3.1 Rethinking Causes of Disagreement

Perspectivist approaches have challenged many, but not all, of the longstanding assumptions described in Section 2.1. In this section, we chart how these approaches reconceptualize disagreement.

Perspectivist approaches recognize that annotator demographics and lived experiences can result in disagreement. Recent studies have examined demographic factors that lead to disagreement, such as race, gender, and age, as well as cultural factors such as education, political affiliation, and native language proficiency (e.g., Goyal et al., 2022; Thorn Jakobsen et al., 2023; Al Kuwatly et al., 2020; Wan et al., 2023; Pei and Jurgens, 2023), with a view toward ensuring that the opinions of people from different backgrounds are represented.

Nevertheless, differences between demographic groups only partly explain disagreement. While this work has been important in better understanding where and how disagreement arises, these methods often assume that disagreement can be well-characterized by demographic factors alone. However, recent work suggests that non-demographic factors are more probable sources of disagreement than some demographic factors across multiple tasks. Many demographic factors do not appear to be good predictors of disagreement across all tasks; Orlikowski et al. (2023) find that modeling gender, age, education, and sexual orientation in isolation do not predict disagreement effectively on a hate speech task, Biester et al. (2022) find no significant differences based on gender across multiple tasks, and Fleisig et al. (2023) find that while race is an important factor in predicting disagreement on hate speech detection, factors such as gender and education are not.

Conversely, factors beyond demographics often cause differences in opinion. These may be task-specific; for example, social media usage and

[2] See Plank (2022) and Cabitza et al. (2023) for discussions of the range of perspectivist work.

opinions on whether online toxic content is a problem greatly help to predict labels on hate speech detection (Fleisig et al., 2023). Other key factors lie outside the scope of what perspectivist work has considered. For example, Miceli and Posada (2022) describe "errors" by Venezuelan image labelers due to differences between English and Spanish, since translations of some words refer to slightly different set of objects. Such issues suggest that a wide range of experiences and perspectives not well-captured by demographics may help to explain systematic disagreement between annotators, but only some of these have been explored.

Regardless of the predictive power of demographics, understanding the opinions of stakeholders from a range of demographic backgrounds is a key contribution of perspectivist work: both because it is important that people from a range of different backgrounds be heard even if they often agree, and because views on more specific topics can vary along demographic axes even if they are not relevant for every item in a dataset. However, widening the scope of potential causes of disagreement would deepen our understanding of why disagreement occurs, improve modeling of annotator behavior, and help to target annotator recruitment to axes that cause disagreement for specific tasks.

3.2 Emerging Practical Challenges

Perspectivist approaches have re-evaluated many longstanding assumptions in data annotation regarding the origins and value of disagreement. However, its new ambitions to engage with the full spectrum of human perspectives bring new challenges regarding data quality, data ethics, institutional pressures, and personalization.

Assessing data quality while capturing disagreement is difficult but critical. A major motivating factor for aggregating multiple annotators' labels is the concern over spam and inattentive or inept annotators, resulting in much research focused on maximizing agreement as a metric of data quality (see Section 2.1). The tension between preserving all annotator opinions and removing "noise" means that perspectivist approaches will face limited use unless alternative methods are developed to maintain data quality without discarding disagreement. Promising examples of these methods include de Marneffe et al. (2019), which uses clear-cut control samples for which the authors are willing to assume that no disagreement could reasonably occur. Deng et al. (2023, Appendix B) collect a variety of quality checks from previous work that, besides inter-annotator agreement, include completion time (Diaz et al., 2018), correlation between similar labels (Demszky et al., 2020), and briefing or training annotators (Akhtar et al., 2021).

Evaluation still relies primarily on majority-vote labels. Plank (2022) notes that a majority of perspectivist papers still evaluate against averaged "gold" labels, which undercuts the potential utility of perspectivist methods. We argue that the continued evaluation via averaging is a symptom of deeper problem: even if we model diverse annotator opinions, models typically produce a single output or classification, and we lack metrics for the quality of that single output besides its similarity to the gold aggregated label. That is, despite the more detailed and diverse data gathered from perspectivist work, the community lacks methods to evaluate models using that data (though see Section 4 for approaches beginning to explore such methods).

Collecting more detailed data requires considering impacts on data subjects. Collecting the opinions of minoritized populations could constitute an undue burden on minoritized groups, especially if the data collection does not result in a commensurate benefit in terms of quality of service for that group. There is also a tradeoff between the richness of collected data and preserving privacy of group members. Potential ways forward include learning from less data so fewer data points are needed, using privacy-preserving machine learning methods (Xu et al., 2021), and engaging with community-led methods for preserving data ownership, such as indigenous data sovereignty (Kukutai and Taylor, 2016).

Participatory approaches conflict with institutional pressures. Institutional pressures hinder efforts to collect more representative and complex data, particularly when it comes to meaningfully involving participants. Researchers face pressures to collect data quickly, not better. By contrast, participatory approaches aim to build mutual, reciprocal relationships; grapple explicitly with power dynamics between researchers and participants, as well as between participants; engage with specific contexts of use; and rethink what is on the table for participants—for example, extending beyond data collection to problem formulation

or evaluation (Delgado et al., 2023). Thus, calls to increase participation may underestimate the extent to which institutional factors discourage such approaches. As a result, lowering boundaries to participation through platforms or methods of data labeling that improve communication and empower participants is key, as well as pressuring institutions to incentivize slower, more thoughtful, and more context-specific (rather than maximally portable (Selbst et al., 2019)) data collection.

3.3 Emerging Normative Challenges

Perspectivist work exposes longstanding assumptions regarding ground truth and the merits of aggregation, but some assumptions still remain implicit in perspectivist approaches. We delineate normative challenges that perspectivist work still faces regarding majority-vote labels, the bounds of acceptable disagreement, and researcher positionality.

Perspectivist approaches do not always explicitly take a normative stance. Machine learning researchers often do not take explicit stances on what systems ought or ought not to do, under the assumption that research is or should be neutral and does not reflect social values or researcher perspectives (Birhane et al., 2022; Santy et al., 2023). But as emerging perspectivist efforts aim to engage with the full spectrum of human perspectives, researchers and practitioners will need to grapple explicitly with challenging normative questions—does the problem formulation admit a correct answer, and (if there is one) whose perspectives form the basis for that answer? Are some perspectives prioritized, or they are all weighed equally?

Engaging explicitly with these questions is especially critical because not doing so may leave important assumptions implicit and therefore unavailable for discussion (Blodgett et al., 2020), or even cause its own harm (Talat et al., 2021). For example, in the absence of explicit definitions of hate speech, research may instead rely on aggregation of crowdsourced perspectives to decide what constitutes hate speech. But such an aggregation may in fact unjustly neglect the views of minoritized groups (Thylstrup and Talat, 2020).

We therefore see discussion of these normative questions as essential as the perspectivist literature continues to develop. If, as we suggest in Section 3.2, the community ought to be developing the technical machinery to model and evaluate beyond majority vote labels, then as a prerequisite, the community must explore what it wants that machinery to model and evaluate.

Bounds of "acceptable" disagreement typically remain implicit. Rottger et al. (2022) distinguish between a descriptive annotation paradigm, in which annotators are encouraged to provide subjective opinions without researcher influence, and a prescriptive one, in which annotators are encouraged to be "objective" and adhere to strict guidelines. This dichotomy can aid researchers in deciding whether disagreement on a data labeling task should serve as a signal that the task is underspecified or as valuable information to preserve.

Many data labeling tasks combine descriptive and prescriptive practices. Task-specific bounds of acceptability often define when variation should be preserved: a painting of a bird might be reasonably labeled "painting" or "bird," but not "cat." Setting these bounds is particularly fraught for tasks involving social norms, such as hate speech detection. Understanding where to set guidelines, and where to permit variation, is task-specific and difficult. For example, there is widespread disagreement over how to operationalize "toxicity" and "alignment," concepts whose bounds often go unstated despite being central to major "subjective" tasks (Thylstrup and Talat, 2020; Kirk et al., 2023). However, without explicitly setting such bounds, we encounter the problems faced under the majority-vote paradigm: opinions defined nebulously by aggregation result in normative boundaries that are hard to pinpoint, let alone contest. These boundaries may thus be difficult to change even when they are demonstrably unfair.

Personalization may not resolve issues of disagreement. Increasingly powerful language models present the possibility of personalizing models to individual users rather than using a single model to satisfy many different preferences (Plepi et al., 2022; Flek, 2020). We argue that personalization alters issues related to disagreement but does not necessarily solve them. While some types of personalization are beneficial (e.g., targeting a scientific explanation to students at different levels), others could perpetuate harms (e.g., supporting misinformation that a user believes). Personalization does not bypass normative issues, but rather changes the structure of the problem: the difficult decision becomes whether and when personalization is appropriate. Here, the community might draw on work in

recommender systems, in which personalization is a primary goal and persistent concerns arise about its appropriate scope and potential harms (Ekstrand et al., 2012; Wang et al., 2022; Li et al., 2023).

4 Recommendations for Practical Challenges

Annotator disagreement carries implications for all stages of the data labeling process. We provide recommendations for each of these stages:

Before data labeling begins. If prescriptive decisions are made about acceptable bounds of disagreement when designing a data labeling process, these decisions should be made explicitly. In addition, consider potential axes of disagreement for the specific task at hand, such as linguistic or sociocultural differences, ambiguous labels, or differences of opinion. Considering these normative questions—who or what is the data collection for—and potential sources of disagreement before beginning data labeling can help to design the process so that important differences in opinion are captured, and sources of confusion are minimized.

Recruitment. Depending on the extent to which the collected data should reflect the opinions of all stakeholders or focus on experts, different best practices for annotator recruitment apply. If the objective is to reflect the views of a particular population, such as potential users, it is crucial to recruit a representative sample of that group. This may sometimes require additional recruitment efforts to account for different demographics' uneven participation in crowdsourcing. In addition, rather than filtering out "noisy" annotators based on whether they disagree with others, alternative filtering strategies such as checking intra-annotator agreement (Abercrombie et al., 2023) or doing multiple rounds of qualification tasks before the main task (Zhang et al., 2023) can help to reduce spam without discarding minority opinions.

An intermediate approach might consider stratifying the recruited sample of annotators based on important axes of disagreement (e.g., different countries where a model will be used) to upsample groups that might otherwise be underrepresented. In addition, for tasks involving different types of expertise (e.g., system summarizing medical or legal documents, or a language model giving advice to specific communities), consider allocating annotators to items based on their expertise.

Other considerations apply regardless of the recruited population. Recruiting a large annotator pool helps mitigate sample error, and capping annotations per annotator can prevent a dataset from primarily reflecting the views of a few annotators. When modeling disagreement, consider collecting annotator data specifically about factors likely to cause disagreement for the task at hand.

Data labeling design. Given annotators' frequent concerns over a lack of task context, and the effects of task context on annotator judgments, it is key to give annotators more context when labeling data. This includes what the data will be used for (e.g., for what task, for which users) and potential effects of system decisions (e.g., whether the system will be used in a punitive way). Furthermore, use disagreement as a signal to prompt reflection and iteration on the data labeling process. For example, disagreement can signal confusing instructions or an insufficiently rich space of potential labels. In cases where ambiguities could cause disagreement (e.g., whether pictures of birds count as birds), or where annotators might provide labels not foreseen by task designers (e.g., Sheppard et al., 2023), provide ways for annotators to indicate uncertainty, such as an "unsure/unclear" option, and ways to give open-ended feedback so that the task can be clarified or expanded.

Dataset documentation. Details on the data labeling process can help future stakeholders to understand factors that might have affected annotator judgments. Previous data documentation work has recommended including information such as annotator demographics and labeling task instructions (McMillan-Major et al., 2024) or the original task for which data was collected (Gebru et al., 2021). Expanding on this work, we recommend also documenting (i) annotator selection procedures, including the number of annotators and restrictions on participation, (ii) the distribution of items labeled per annotator, and (iii) any annotator filtering used. Future dataset users can also benefit from describing normative bounds imposed on the data labeling process and rationales for discarding any data. Providing non-aggregated individual labels when possible also helps to avoid information loss from aggregation.

Model design and evaluation. Different model objectives aside from accuracy on predicting aggregate labels, such as measuring KL divergence

between predictions and the distribution of annotator labels, or calibrating to the distribution of annotator opinions, allow disagreement to be accounted for during training. During evaluation, potential alternatives to using averaged "gold" labels include measuring distributional similarity, e.g., with KL divergence, cosine similarity between lists of outputs, or a correlation coefficient (Nie et al., 2020; Dumitrache et al., 2019; Zhou et al., 2022), evaluating accuracy at modeling individual annotators (Davani et al., 2022; Resnick et al., 2021), and measuring model calibration to population uncertainty (Baan et al., 2022). Evaluator disagreement is also a useful signal: if evaluators disagree over the quality of a model output, this information can help to pinpoint model weaknesses or reveal instances of disparate quality of service for different subgroups.

5 Recommendations for Normative Challenges

In this section, we discuss potential avenues for research aiming to engage annotator disagreement.

Replace implicit normative decisions with explicit ones. Majority-vote aggregation captures the average view of the aggregated population with every annotator weighted equally. By contrast, data labeling tasks where some people are clearly better-informed (e.g., doctors in medical domains, or speakers of a language for translation) have an implicit "expert-driven" framing, in which only some views are solicited. This includes considering lived experience as a form of expertise, which can prove critical to successful annotation. We can imagine a spectrum of practices ranging from "democratic" to "expert-driven," with different points along this spectrum suited to different situations. For a task requiring medical knowledge, it would be unreasonable to use labelers with no medical training; when setting community norms, all community members' views are important. Each data labeling task requires choosing a point on this spectrum. Making this decision explicitly, rather than defaulting to majority vote, can help to create decision rules that are easier to define and contest.

Draw on parallel problems from other disciplines. The broader questions of how to capture a population's views, and make decisions based on them, has a long history across a range of traditions involving stakeholder participation, from social choice theory and mechanism design (Arrow, 1977; Feldman and Serrano, 2006) to value-sensitive and participatory design (Friedman, 1996; Muller and Kuhn, 1993):

Science and technology studies. Bowker and Star's (2000) analysis of the political and social dimensions of classification highlights the importance of retrievability, the process of retaining the voices of people conducting classification for systems to maintain "maximum political flexibility." As perspectivist methods examine ways to retain the opinions of individual labelers, this line of work can help to understand how individual voices can be lost, merged, or preserved in systems that draw on them; and understand how we can, as Bowker and Star (2000) note, "reflect new institutional arrangements or personal trajectories." Bowker and Star, as well as Winner (1980) and Agre (2014), illustrate how technological artifacts embed and reproduce social and political values; drawing on Douglas, Lepawsky (2019) and Scheuerman et al. (2021) investigate institutional factors and perspectives of researchers or industry leaders who describe subjectivity as a problem to minimize. Together, this literature contextualizes subjectivity and disagreement, and emphasizes the need for critical reflection on practices and assumptions in technological development.

Elsewhere, critiques of machine learning, including approaches to fairness and ethics, can offer opportunities for perspectivist efforts to reflect on assumptions surrounding representation and inclusion. For example, Hoffmann (2020) complicates the notion of inclusion in dataset design by pointing out that such inclusion can forestall calls for more radical change, while Stevens and Keyes (2021) similarly observe that more representative datasets for e.g., facial recognition do not address the more fundamental problem of surveillance.

Philosophy of mind. Literature related to why, despite our understanding of the physical processes involved, we still lack a full understanding of where subjective feelings come from and why they differ, such as discussion of qualia (Lewis, 1930; Jackson, 1982; Chalmers, 1997, among others), can provide a starting point for discussion of differences of opinion that are not easily situated in terms of the annotator's background.

Voting and social choice. Many issues regarding optimal data labeling resemble issues regarding ideal voting mechanisms, with different constraints. In electoral settings, the full population's opinions

may be solicited, and single decisions based on their choices have widespread effects (e.g., electing an official who makes decisions in many policy areas). During annotation, by contrast, it is often infeasible to solicit all stakeholder opinions, and different aspects of model outputs can be decided independently (e.g., decisions on coherence or offensiveness of model outputs). However, overarching themes of how to aggregate preferences while maximizing stakeholder satisfaction and welfare (Arrow, 1977; Sen, 2018) could provide useful lessons for perspectivist work.

Pragmatics. The community could take inspiration from the notion of a "common ground" in pragmatics, wherein conversational participants communicate based on a shared understanding of the world. This shared understanding is based on factors that include demographic attributes, but also factors such as the specific speech situation, the participants' professions, online communities, languages spoken, and imagined audiences (Clark and Carlson, 1982; Goffman, 1976). Annotation of text functions like a communicative situation in which the annotator interprets language while making assumptions about the speaker, purpose, and audience of the text based on their own background, and a wide variety of factors in their background may be relevant based on the text. Focusing on content moderation, Thylstrup and Talat (2020) draw on Hall et al. (1997) to describe how these assumptions become embedded in datasets: data annotators serve as intermediaries who read on behalf of the intended recipient and often interpret text differently from the specific intended reading that the sender meant to encode, with the result that systems based on those labels encode the intermediary position instead of that of the sender or intended recipient. Understanding the range of factors that influence annotator interpretation could disentangle more latent factors behind disagreement in data collection.

Participatory design. Elsewhere, participatory traditions that interrogate power dynamics in order for "non-expert stakeholders to provide direct input on technology design" (Delgado et al., 2023) can offer practitioners valuable insights for navigating disagreement and reflecting on assumptions about disagreement embedded in their practices (Friedman, 1996; Muller and Kuhn, 1993).

Take advantage of nuanced output spaces to meet diverse stakeholder needs. The existence of disagreement does not rule out the possibility of building systems that produce single outputs affecting a whole population. Edenberg and Wood (2023) note that "any society that protects freedom of thought and expression" experiences "continued disagreement about key normative questions," but we still "find fair terms of social cooperation without requiring everyone to agree." Even when providing a single output is unavoidable, treating preferences non-unidimensionally can help to arrive at single outputs that better serve more people. Systems that provide for a broad range of potential outputs, including generative models, can help to consider different, non-contradictory values that seem to result in disagreeing preferences. For example, if one annotator prefers a language model output that is non-discriminatory and another prefers one that is concise, they might disagree on their preferences between two outputs, but a non-discriminatory and concise output could satisfy both annotators. Revealing that preferences are not unidimensional and exploring the resulting space of potential outputs opens ways to generate greater consensus.

6 Conclusion

Assuming that tasks have a ground truth, using majority-vote aggregation, and avoiding a normative stance have long been common practices in data labeling. However, a growing perspectivist literature is recognizing that datasets and models must be designed to account for the full spectrum of human perspectives. We argue that perspectivist approaches can accomplish their goals more fully by considering causes of disagreement beyond demographics, addressing tensions with data quality and research pressures, and reasoning explicitly about normative considerations.

Limitations and Ethical Considerations

Our position paper aims to provide an analysis of key questions regarding longstanding and emeerging paradigms of data collection, but it is not a comprehensive meta-analysis or literature review; thus, we acknowledge that some relevant work may have been overlooked because we have not comprehensively searched for all papers related to these issues. Overlooking some work carries the risk of narrowing the set of potential perspectives that are considered in future research based on the avenues we discuss.

Acknowledgments

Thank you to members of the Berkeley NLP and Algorithms, Data, and Society groups for their feedback, particularly Deborah Raji and Nicholas Tomlin. Thank you as well to the anonymous reviewers for their helpful suggestions.

References

Gavin Abercrombie, Verena Rieser, and Dirk Hovy. 2023. Consistency is Key: Disentangling Label Variation in Natural Language Processing with Intra-Annotator Agreement. ArXiv:2301.10684 [cs].

Philip E Agre. 2014. Toward a critical technical practice: Lessons learned in trying to reform AI. In *Social science, technical systems, and cooperative work*, pages 131–157. Psychology Press.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Lora Aroyo, Alex S. Taylor, Mark Diaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-Garcia, Vinodkumar Prabhakaran, and Ding Wang. 2023. DICES dataset: Diversity in conversational ai evaluation for safety.

Lora Aroyo and Chris Welty. 2014. The Three Sides of CrowdTruth. *Human Computation*, 1(1).

Kenneth J Arrow. 1977. *Social Choice and Individual Values*, 2 edition. Cowles Foundation Monographs. Yale University Press, New Haven, CT.

Ron Artstein. 2017. Inter-annotator agreement. *Handbook of linguistic annotation*, pages 297–313.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aparna Balagopalan, David Madras, David H. Yang, Dylan Hadfield-Menell, Gillian K. Hadfield, and Marzyeh Ghassemi. 2023. Judging facts, judging norms: Training machine learning models to judge humans requires a modified approach to labeling data. *Science Advances*, 9(19):eabq0701. Publisher: American Association for the Advancement of Science.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. 2022. Analyzing the effects of annotator gender across NLP tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 10–19, Marseille, France. European Language Resources Association.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. ACM.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Alan L. Boegehold. 1963. Toward a study of athenian voting procedure. *Hesperia: The Journal of the American School of Classical Studies at Athens*, 32(4):366–374.

Geoffrey C Bowker and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. MIT Press, London, England.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

David J Chalmers. 1997. *The conscious mind*. Philosophy of Mind. Oxford University Press, New York, NY.

Herbert H. Clark and Thomas B. Carlson. 1982. Hearers and speech acts. *Language*, 58(2):332.

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*, 10:92–110.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA. Association for Computing Machinery.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

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*, CHI '18, page 1–14, New York, NY, USA. Association for Computing Machinery.

Mary Douglas. 1978. *Purity and danger: an analysis of the concepts of pollution and taboo*, repr edition. Routledge, London. OCLC: 248038797.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2164–2170, Minneapolis, Minnesota. Association for Computational Linguistics.

Elizabeth Edenberg and Alexandra Wood. 2023. Disambiguating Algorithmic Bias: From Neutrality to Justice. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pages 691–704, New York, NY, USA. Association for Computing Machinery.

Michael D Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2012. Fairness in recommender systems. In *Recommender Systems Handbook*, pages 679–707. Springer.

Allan Feldman and Roberto Serrano. 2006. *Welfare economics and social choice theory*, 2 edition. Springer, New York, NY.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Lucie Flek. 2020. Returning the N to NLP: Towards Contextually Personalized Classification Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838, Online. Association for Computational Linguistics.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond Black & White: Leveraging Annotator Disagreement via Soft-Label Multi-Task Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Paula Fortuna, Monica Dominguez, Leo Wanner, and Zeerak Talat. 2022. Directions for NLP practices applied to online hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11794–11805, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Batya Friedman. 1996. Value-sensitive design. *interactions*, 3(6):16–23.

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

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Erving Goffman. 1976. Replies and responses. *Language in Society*, 5(3):257–313.

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*, CHI '22, pages 1–19, New York, NY, USA. Association for Computing Machinery.

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).

Stuart Hall et al. 1997. The spectacle of the other. *Representation: Cultural representations and signifying practices*, 7.

Anna Lauren Hoffmann. 2020. Terms of inclusion: Data, discourse, violence. *New Media & Society*, 23(12):3539–3556.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pages 27–35.

Olivia Huang, Eve Fleisig, and Dan Klein. 2023. Incorporating worker perspectives into mturk annotation practices for NLP.

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

Frank Jackson. 1982. Epiphenomenal qualia. *The Philosophical Quarterly (1950-)*, 32(127):127–136.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

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, New York, NY, USA. Association for Computing Machinery.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2023. The empty signifier problem: Towards clearer paradigms for operationalising "alignment" in large language models.

Tahu Kukutai and John Taylor, editors. 2016. *Indigenous Data Sovereignty: Toward an agenda*, volume 38. ANU Press.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Josh Lepawsky. 2019. No insides on the outsides. *Discard Studies*.

Clarence Irving Lewis. 1930. Mind and the world-order. *International Journal of Ethics*, 40(4):550–556.

Yanying Li, Haipei Sun, and Wendy Hui Wang. 2020. Towards fair truth discovery from biased crowdsourced answers. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 599–607, New York, NY, USA. Association for Computing Machinery.

Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. 2023. Fairness in recommendation: Foundations, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 14(5):1–48.

Angelina McMillan-Major, Emily M. Bender, and Batya Friedman. 2024. Data statements: From technical concept to community practice. *ACM J. Responsib. Comput.*, 1(1).

Milagros Miceli and Julian Posada. 2022. The data-production dispositif. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Communications of the ACM*, 36(6):24–28.

Hasti Narimanzadeh, Arash Badie-Modiri, Iuliia G. Smirnova, and Ted Hsuan Yun Chen. 2023. Crowdsourcing subjective annotations using pairwise comparisons reduces bias and error compared to the majority-vote method. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.

Alicia Parrish, Sarah Laszlo, and Lora Aroyo. 2023. Is a picture of a bird a bird: Policy recommendations for dealing with ambiguity in machine vision models. ArXiv:2306.15777 [cs].

Desmond Upton Patton, Philipp Blandfort, William R. Frey, Michael B. Gaskell, and Svebor Karaman. 2019. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. In *Hawaii International Conference on System Sciences*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Jiaxin Pei and David Jurgens. 2023. When Do Annotator Demographics Matter? Measuring the Influence of Annotator Demographics with the POPQUORN Dataset.

Pew Research Center. 2016. Research in the crowdsourcing age, a case study.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maja Popović. 2021. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243, Online. Association for Computational Linguistics.

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*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paul Resnick, Yuqing Kong, Grant Schoenebeck, and Tim Weninger. 2021. Survey equivalence: A procedure for measuring classifier accuracy against human labels. *CoRR*, abs/2106.01254.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? disciplinary values in computer vision dataset development. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, New York, NY, USA. Association for Computing Machinery.

Amartya Sen. 2018. *Collective choice and social welfare*. Harvard University Press.

Brooklyn Sheppard, Anna Richter, Allison Cohen, Elizabeth Allyn Smith, Tamara Kneese, Carolyne Pelletier, Ioana Baldini, and Yue Dong. 2023. Subtle misogyny detection and mitigation: An expert-annotated dataset. In *Socially Responsible Language Modelling Research (SoLaR) Workshop*.

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*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Nikki Stevens and Os Keyes. 2021. Seeing infrastructure: race, facial recognition and the politics of data. *Cultural Studies*, 35(4-5):833–853.

Zeerak Talat, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied Machine Learning: On the Illusion of Objectivity in NLP.

Terne Sasha Thorn Jakobsen, Laura Cabello, and Anders Søgaard. 2023. Being right for whose right reasons? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1033–1054, Toronto, Canada. Association for Computational Linguistics.

Nanna Thylstrup and Zeerak Talat. 2020. Detecting 'Dirt' and 'Toxicity': Rethinking Content Moderation as Pollution Behaviour. *SSRN Electronic Journal*.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A Case for Soft Loss Functions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8:173–177.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14523–14530.

Yifan Wang, Weizhi Ma, M. Zhang, Yiqun Liu, and Shaoping Ma. 2022. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41:1 – 43.

Yuxia Wang, Shimin Tao, Ning Xie, Hao Yang, Timothy Baldwin, and Karin Verspoor. 2023. Collective Human Opinions in Semantic Textual Similarity. *Transactions of the Association for Computational Linguistics*, 11:997–1013.

Langdon Winner. 1980. Do artifacts have politics? *Daedalus*, 109(1):121–136.

Runhua Xu, Nathalie Baracaldo, and James B. D. Joshi. 2021. Privacy-preserving machine learning: Methods, challenges and directions. *ArXiv*, abs/2108.04417.

Lining Zhang, Simon Mille, Yufang Hou, Daniel Deutsch, Elizabeth Clark, Yixin Liu, Saad Mahamood, Sebastian Gehrmann, Miruna Clinciu, Khyathi Raghavi Chandu, and João Sedoc. 2023. A needle in a haystack: An analysis of high-agreement workers on MTurk for summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14944–14982, Toronto, Canada. Association for Computational Linguistics.

Xiang Zhou, Yixin Nie, and Mohit Bansal. 2022. Distributed NLI: Learning to predict human opinion distributions for language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 972–987, Dublin, Ireland. Association for Computational Linguistics.