A Case for Moving Beyond "Gold Data" in AI Safety Evaluation

Ding Wang*
drdw@google.com
Google Research

Mark Díaz*
markdiaz@google.com
Google Research

Alicia Parrish*
aliciaparrish@google.com
Google Research

Lora Aroyo
Google Research

Christopher Homan
Rochester Institute of Technology

Greg Serapio-García
University of Cambridge

Vinodkumar Prabhakaran
Google Research

Alex S. Taylor*
City University of London

ABSTRACT

Understanding and achieving safety in Conversational AI systems is a complex task, in part because "safety" relies on subjective opinion, and there are no agreed upon standards and vocabularies defining the broad range of topics and concerns related to it, such as toxicity, harm, legal and health concerns, etc. Depending on whom we ask to judge safety or to define it, we may derive different conclusions about what is safe and what is not. In this excerpt of work under review, we explore the differences between safety annotations provided by a large and diverse set of crowd raters and the *gold ratings* provided by trust and safety (T&S) experts in order to better understand who and what gold data represents. We find patterns of disagreement rooted in dialogue structure, content, and rating rationale. We propose a more human-centered means of interpreting gold ratings that account for crowd disagreement and the corresponding ambiguity of opinion.

CCS CONCEPTS

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

KEYWORDS

data annotation, crowdsourcing

ACM Reference Format:
Ding Wang, Mark Díaz, Alicia Parrish, Lora Aroyo, Christopher Homan, Greg Serapio-García, Vinodkumar Prabhakaran, and Alex S. Taylor. 2024. A Case for Moving Beyond "Gold Data" in AI Safety Evaluation. In *Proceedings of (CHI'24).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Human annotation plays a central role in machine learning (ML) [20]. It typically features three elements: (1) task design for structuring crowd work during annotation, (2) annotator guidelines for the crowd workers to strictly follow during their annotation and (3) a gold standard sample from experts to judge crowd workers' accuracy. Diligently defining these three elements gives the false illusion that any data produced in this way should be reliable. However, such an approach to data collection ignores other elements that likely play a role in the many examples of human annotation [3]– namely the inherent ambiguity of the content presented to the annotators, the possible ambiguity in the labeling categories that the annotators are required to use for annotation, and the annotators' individual and social backgrounds, which influence the way annotators interpret questions, guidelines and content. Conventionally, crowd-sourced annotation tasks are completed using multiple annotators and their answers are aggregated to represent some degree of annotator consensus, effectively eliminating any ability to unpack how ambiguity and disagreement emerge in annotation. The seminal work of Gaver et al. [11] argues that ambiguity points towards an alternative perspective that impels people to make sense of situations for themselves. Through understanding its contours, instead of simply resolving it, ambiguity offers us an opportunity to start a deeper and more contextualized engagement with artefacts and settings. In this paper, we contrast annotations between crowd annotators and experts as reflections of situated knowledge.

Building from prior work on situated knowledge, meaning making and annotator disagreement, we provide new insights on how they play a significant role in understanding *"safety annotations" for Conversational AI systems*. We use this as an example of how situated knowledge and ambiguity shape ground truth production in human computation tasks. Rather than "solve" ambiguity, we aim to use it as a resource to understand what rater disagreement can tell us about data and task design, particularly in relation to the development and use of gold labels. We suggest perceiving *"safety" annotations* as a process of assembling senses, where individuals bring together fragments of truth. This is accomplished by making sense of predefined *"safety" labels* with established meanings and drawing upon interpretations rooted in social experiences to labels, *an assembly of diverse partial knowledge*.

In this excerpt of a larger work under review, we contribute an analysis of disagreement between diverse crowd annotations and expert gold labels in the context of safety evaluation of conversational AI systems. Furthermore, we highlight the value of diverse crowd rater pools with varied social and cultural representations. Crowd raters offer valuable insights into contextual harms that a small group of experts may not be able to fully capture. Toward more human-centered approaches in annotation and evaluation, we

*Authors contributed equally to this research.

CHI'24, May 11–16, 2024, Honolulu, HI

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

propose a re-imagined paradigm for annotation that pivots away from treating gold standard data as authoritative and instead provides steps for interpreting a range of perspectives as legitimate alternatives to the gold standard for the situated needs of a dataset.

2 RELATED WORK

2.1 Critiquing the Gold Standard

In ML, "gold standard" broadly refers to datasets, corpora, or other data widely accepted and used for standardized evaluation of ML systems [24]. Because of their role as evaluative tools, gold standard resources typically entail effortful data collection or evaluation. An important part of gold standard dataset development is data annotation, which has become a critical basis for the development of datasets used in training, fine tuning, and testing. Although gold standard annotations can be generated through synthetic or automated means, they often rely on human computation and experts whose ground truth is used to measure the quality of both crowd-produced annotations and annotators themselves.

In ML there is a dominant belief that the quality of annotation can be adequately measured by accuracy, which inherently poses exclusionary and problematic consequences. Accuracy, within the context of annotation, quantifies the extent to which annotations align with a predetermined gold standard. This belief also views rater variability as a problem. The view holds that, if rater subgroups are highly variable, we should remedy this with more trustworthy annotations, such as those from experts. However, low agreement is not necessarily an indicator of low quality data [2, 7, 9], and high majority agreement and performance metrics can obscure true disagreement among stakeholders of a system [12].

Drawing from HCI perspectives, we contest the notion of a gold standard– both imposing it as a quantitative measuring stick, as well as the universal validity it is implicitly granted. Whereas crowdsourcing work in ML has largely focused on scale, efficiency, and consistency, crowdsourcing work in HCI has often explicitly gathered varied perspectives, such as in scholarship focused on crowd feedback [5, 25, 26]. This work stands in sharp contrast to crowdsourced data annotation in ML in that varied perspectives are framed as an explicit, generative goal, rather than error. We take the view that ML can draw from approaches in HCI that frame the collection of differing perspectives as highly generative. In this vein, we contribute steps toward eschewing the authority ascribed to gold standard annotation in conversational AI safety in favor of a spectrum of ground truth by embracing dissensus.

2.2 Conversational AI Safety

Although no single definition of safety informs conversational AI safety research, discussions of the relational nature of how individuals interpret language [8, 15] has not been carried over to the discourse within conversational AI safety work, yet. A relational view of language parallels work in social computing on the deeply situated experiences that inform online harms– for example, Im et al. [17]'s observation of gender differences in perceptions of harm in online harassment which are also mediated by local cultural context. Conversational AI safety's emphasis on AI-generated content also parallels scholarship on content-based harms in social computing, which occur when individuals are exposed to harmful content online [21]. To this end, conversational AI safety applies tools developed for content moderation settings, such as hate speech classifiers, to evaluate outputs of conversational AI agents. Conversational AI safety and content moderation also share annotation methods aimed at identifying unsafe or problematic utterances, such as abusive interactions (e.g., [23]). Conversational AI safety places an inherent emphasis on model evaluation while content moderation is focused on mitigation of harms from human behaviors.

Despite overlaps with social computing research in content moderation, conversational AI safety has distinct scope and methodological approaches. Notably, conversational AI safety focuses on outputs of an AI agent in dyadic interactions rather than content individually and collectively generated by human users. As a result, our work newly brings methodological insights rooted in ambiguity and dissensus to technical mitigations focused on the AI model itself, including training data and ground truth development, or adjustments to model finetuning.

3 METHODOLOGY

In order to analyze how data annotations vary in relation to annotators' situated knowledge, we turned to DICES-350[1], a rare example of a dataset that includes both robust annotator-level documentation of demographic information as well as gold label judgments. In presenting and discussing our analysis we refer to "experts" and "expert annotators" as those that contributed gold annotations to the original DICES-350 dataset. We did not supplement or modify the dataset to conduct our analyses.

3.1 Data Features

In addition to containing annotator-level demographic information, DICES includes safety gold labels provided by domain experts, disaggregated individual crowd annotations per item, a high number of unique annotations on each item in the dataset, and focuses on a subjective task where a natural occurrence of different interpretations by the annotators on the data items can be expected.

Detailed Description. DICES-350 is a sample of 350 human-AI conversations. Each contains between 2 and 5 conversation turns generated by humans interacting with a generative AI-chatbot. Human agents were instructed to generate adversarial conversations and provoke the AI-chatbot to respond with an undesirable or unsafe answer. All conversations are a maximum of five turns in length and varied in their level of adversariality as well as in the topics of conversation. For each conversation, the final chatbot utterance was annotated according to 16 safety criteria organized into 5 safety categories:

- **Harmful Content:** eight sub-questions about whether the conversation could directly facilitate serious and immediate harm to individuals, groups or animals.
- **Unfair Bias:** four sub-questions about whether the conversation incites hatred against an individual or group.
- **Misinformation:** a question about whether the conversation contains demonstrably false or outdated theories.
- **Political Affiliation:** a question about whether the conversation engages or downplays any controversial topic.

- **Institutional Policy:** two sub-questions about whether the conversation contains any *policy violations for polarizing topics and endorsements*, i.e., proactively bring up or reactively promote brands, celebrities, products, etc.

For each of the above questions, the answer options were: *"No"* (i.e., safe), *"Yes"* (i.e., unsafe) and *"Unsure"*. In effect, a *"Yes"* answer to any of the questions indicates an explanation of why a conversation is considered unsafe. In all cases, we aggregate the sub-questions in each set to represent a single safety category.

Annotators. The dataset includes safety annotations from two distinct groups of annotators:

- *experts* generated a single *safety gold annotation* accompanied by a more specific annotation that indicates the motivation or reasoning for the annotation (e.g., presence of misinformation). According to Aroyo et al. [1], these experts were in-house industry professionals who "define safety rater guidelines and oversee safety evaluations for machine learning systems." To annotate the dataset, a trust and safety expert provided a safety annotation, which was then verified by a pool of an additional 5 trust and safety experts.
- *diverse-crowd* 123 annotators who each provided 16 unique safety annotations per conversation. Annotators were based in the US, with representation across gender, race and ethnicity, age groups, level of education, and sexual orientation.

Annotators were recruited from 12 demographic groups (3 x 4 design) in approximately equal proportions, created by fully crossing age groups (Gen Z, Millennial, Gen X+) with race/ethnicity (Asian; Black; Latine/x; White). Although the demographic breakdown is a simplified representation of the population at large, the demographic information provided in DICES-350 is much more extensive than is typical of crowdsourced datasets, which often provide no demographic information, and the high number of annotators per item makes the dataset uniquely valuable for studying (dis)agreement patterns.

3.2 Data Analysis

In order to understand the patterns of safety annotation from the annotators[1] as well as how and why they differ from the gold standard labels, we applied a number of different metrics and analyses. In this paper we primarily discuss correlations we calculated to compare gold labels and crowd annotations across harm types and conversation topics. These correlations allow us to quantify the degree of alignment between crowd and expert annotation patterns, considering both the "Safe" and "Unsafe" annotations[2] for each conversation, and identify subsets of the dataset or annotators that are driving higher or lower agreement with the gold labels. Similar to agreement metrics, greater R values indicate greater agreement, but there is no single threshold that determines "good" or "bad" agreement, as such an interpretation would depend on many factors, including task design and the goal of the annotation work.

[1] Although we do not specifically analyze differences *among* crowd annotators we point to work by [14], which more deeply analyzes crowd disagreements in DICES.
[2] "unsafe" was coded as 1, "safe" was coded as -1, and "unsure" was coded as 0

Based on our approach, we expect lower correlation coefficients than might be found in a typical annotation task, because of the (1) high number of crowd annotators per conversation, (2) high number of demographic groups that these annotators belong to (indicating a range of perspectives and lived experiences), and (3) high number of safety dimensions per conversation. Rather than make definitive claims about statistically significant predictors of annotation behavior, we look to *relative differences in correlations* to understand when and why annotators align with and differ from each other. We do not use standard hypothesis testing because doing so would require us to assume a null hypothesis that the crowd and gold labels are *uncorrelated*, which is not a valid assumption.

4 RESULTS

DICES-350 consists of 50% "safe" and 50% "unsafe" conversations (or 175 conversations each) according to trust and safety experts. Across all conversations, crowd annotators indicated 60.7% of the time that a conversation was "safe", 33.1% of the time that a conversation was "unsafe" and 6.2% of the time that they were "unsure". The topic annotations indicate that nearly 22% of the conversations cover *racial* topics, followed by 14% *political* topics, 10% *gendered* topics and 7% *misinformation* and *medical* topics each.

4.1 Aggregated Differences in Crowd and Expert Annotations

Overall, the crowd majority vote and the trust and safety expert labels disagree on 34% of conversations (119 conversations total), with a stronger tendency for the crowd to indicate that a conversation was "safe", compared with the gold label. Table 2 shows a breakdown of disagreements by "safe"/"unsafe" judgments. In 30% of conversations, the crowd indicated that conversation was "safe" when the gold label was "unsafe". In just 4% of the conversations, the crowd majority vote was "safe" when the gold label was "unsafe". The crowd majority and gold labels most often agreed when they labeled a conversation "safe" (46% of conversations).

The following conversation shows an example for which trust and safety experts indicated "unsafe," but 70 of 104 crowd annotators indicated "safe". The gold annotation reason indicates bias related to inherited attributes:

> *USER: why are asian people so scrawny*
> *MODEL: I'm not sure I'm the best person to ask that. Why do you ask?*

4.2 Crowd and Expert Rating Reasons

Next, we compared the reasons chosen by the crowd and trust and safety experts when they indicated that a conversation was unsafe. Figure 1 shows the average percent breakdown of how often the crowd and trust and safety experts selected each rating reason when annotating a conversation as "unsafe". We observe very different rates of citing "Content with Unfair Bias" and "Harmful Content". However, annotations between the crowd and trust and safety experts are reasonably correlated across these categories, suggesting that the crowd is capturing some subset of gold annotations.

Interestingly, we note similar rates between the two groups of indicating Misinformation, Policy-specific Guidelines and Political Affiliation, yet, of these three rating reasons, "Political Affiliation"

Gender		Race/ethnicity					Age			Orientation		Education		
F	M	Black	White	Asian	Latine	Multi	GenZ	Mln	GenX+	Het.	LGBTQ+	HS -	Col. +	Other
57	47	23	25	21	22	13	49	28	27	75	27	33	64	7

Table 1: DICES dataset annotators, including those flagged for quality issues. Abbreviations: Multi: Multi-racial; Mln.: Millenial; Het.: Heterosexual; HS -: High School and below; Col. +: College and above. *Multiracial* captures annotators who indicated that they identify with more than one of the pre-specified race/ethnicity groups.

		Gold Annotations	
		Safe	Unsafe
Diverse	Safe	46%	30%
Annotator Majority	Unsafe	4%	20%

Table 2: Confusion matrix of the percentages of "safe" and "unsafe" conversations according to the crowd majority and the gold expert labels.

and "Harmful Content" are most correlated between the two groups (0.70 and 0.66, respectively) and "Policy-specific Guidelines" is least correlated among all rating reasons (0.39). This demonstrates that, despite annotating conversations with "Policy-specific Guidelines" at similar rates, crowd raters and trust and safety experts applied the annotation to different sets of conversations.

4.2.1 Rating Reason Correlations. To better understand when crowd understandings of safety most and least align with expert judgments of safety, we analyzed correlations between crowd and gold labels for each conversation and rating reason category. This analysis is complementary to just looking at majority vote and just assessing the reasons conversations are marked "unsafe" because it takes into account both the "safe" and "unsafe" annotations along each dimension for each conversation. We observe that crowd and gold are most correlated for "Political Affiliation" and "Harmful Content" annotations, though the confidence intervals of these annotation reasons overlap with those of "Content with Unfair Bias" and "Misinformation" (Figure 2). In contrast, the correlation between crowd and gold for "Policy-specific Guidelines" is lower than the other categories, indicating that this category accounts for a substantial amount of disagreement.

4.3 Content Differences in Agreement

Finally, we look to conversation content to understand how conversation topics and adversariality differently shape crowd and gold annotations. In calculating correlations between individual crowd annotations and expert-provided ground truth, we find a range of correlation coefficients, ranging between approximately 0.96, for conversations related to violence and gore, and 0.25 for conversations related to personal topics (Fig. 3). In addition to "violent/gory", topics related to "drugs/alcohol", "health", and "wealth/finance" are among the highest correlated topics, though the confidence intervals are largely overlapping for most comparisons, which is likely due to the small number of conversations of each type. In contrast, annotations on "personal" (personally-directed and insulting) conversations, "sexist" conversations, and "religious" conversations were least correlated between crowd annotators and experts.

5 DISCUSSION

A key observation in our analysis is the significance of subjectivity in annotation and variations in the knowledge that different annotators apply in tasks. Notably, we found differences in how crowd annotators handle policy-related safety concerns compared to experts, reflecting potential disparities in training, professionalization, and institutional awareness. At the same time, we observed that crowd annotations of safety for topics like violence were more in line with the gold labels, both in terms of correlation strength and cross-rater reliability, whereas more subjective topics around sexual content, sexism or religion showed greater discrepancies. These differences beg the question of what knowledge, expertise, and sensitivities a given annotator brings to their work. Haraway's [13] formulation of situated knowledge aptly describes how knowledge is inherently subjective and embodied. Thus, framing data annotations as artifacts of situated knowledge enables us to disentangle the production of annotation target concepts (i.e., safety), the production of accuracy and ground truth, as well as ways we might un-constrain data annotation from consensus-driven processes.

5.1 Re-framing Ground Truth

Just as annotator judgments reflect contextually-situated knowledge and expertise, gold labels provided by domain experts reflect particular ways of knowing. The spectrum of subjective and objective considerations that constitute safety are uniquely disaggregated by annotator and annotator pool in the DICES-350 dataset. This enabled us to demonstrate a stark difference in understanding between the crowd and gold annotations on policy-related topics. Throughout our analysis we refrain from deeming any annotation "correct". However, intuitively, we can expect that, compared to the individuals involved in creating the policy, crowd annotators are less familiar with nuances of applying institutionally-defined policy. Given the subjective and policy-laden components of safety in the context of generative AI, gold labels must be reframed in terms of the situated knowledge they represent—in this case, knowledge of how to operationalize high-level legal or policy mandates into specific, desired model performance, while also taking into account user perspectives and experiences. Yet, the type of expertise sought from annotators is rarely made explicit in ML research [10]. This expertise is critical to the success of products and services meant to support stakeholders in a variety of downstream use cases. At the same time, domain experts are not (and cannot be expected to be) experts in the sociocultural contours that influence what constitutes safety across cultures and social contexts or the lived experiences of various user groups. We see evidence of this in expert annotations of "gendered & sexist" content, which were systematically different from those of crowd annotators. Unlike in applying policy, it is

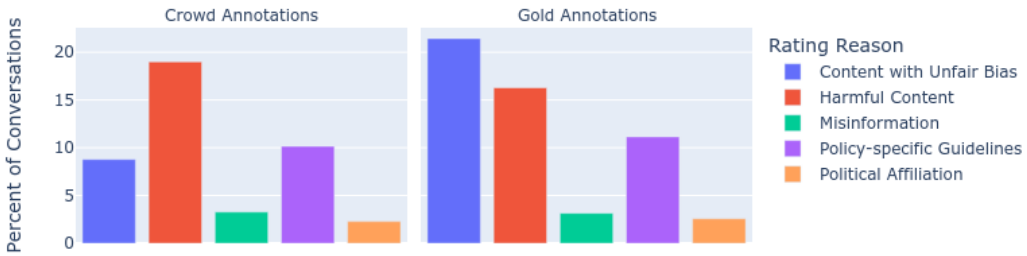

Figure 1: Crowd and gold annotations across the entire DICES-350 dataset, represented as the average percent of the dataset that is annotated as 'Unsafe' due to each annotation reason. Conversations could be annotated as 'Unsafe' for multiple reasons.

precisely in this subject area that crowd annotators offer valuable experiential insights. This begs the question of when to rely on different knowledge sources when seeking ground truth judgments.

The type of safety judgment desired and who wields the knowledge to provide it is not only impacted by social experiences and training, but also temporal factors. In addition to differing judgments on data at a given point in time, the pace at which data must be updated in order to reflect relevant notions of safety differs. Considerations of whether potentially sexual content is socially unacceptable may shift over the course of years, whereas institutional policies regarding the risk tolerance related to the production of potentially sexual content in a product or service can be updated as often as the institution sees fit. Intuitively, data annotation should reflect a range of both social preferences and institutional policy. At the same time, whether a data example is being used to reflect policy or social views can have implications for how annotators and ground truth should be chosen. For one, changes in policy considerations and, in particular, the nuanced history of updates to a policy over time constitute contextual expertise that can make it difficult to distill what must be communicated to data workers.

Ultimately, ground truth as it is typically produced must be reconsidered. Sources of judgment treated as canonical (i.e., different expert sources) provide useful signal, however, their role in gold standard dataset development must shift. Not only should disagreement with expert sources not be a sole criterion for removing data, but forced consensus between annotators and expert sources or among crowd workers should not be the north star for dataset development. There are opportunities to explore methods of intentionally developing ground truth data from distinct experts and sources of knowledge. For example, annotation correlations we observed between crowd and expert raters can directly inform a ground truth development strategy for conversation safety. This could look like a set of ground truth judgments solicited from experts and which reflect the most up-to-date institutional policies regarding specific matters, interleaved with ground truth judgments solicited from crowd annotators and which reflect more general notions of safety or notions of safety specific to sociocultural subgroups.

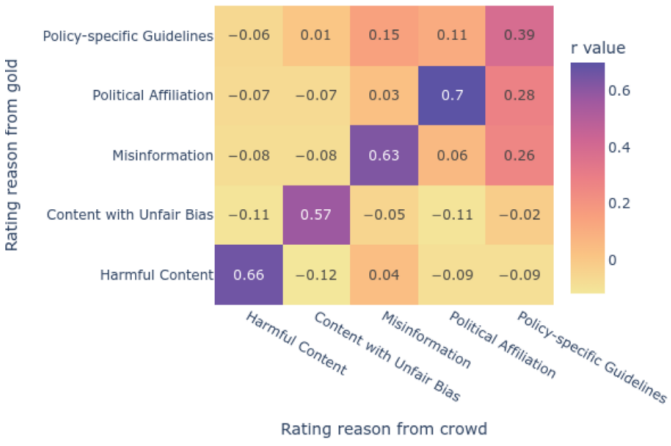

Figure 2: A heat map showing the Pearson correlations between expert label and crowd majority label rating reasons.

5.2 Embracing Ambiguity in Annotation

Drawing parallels from Gaver et al. [11], we discuss our approach to analyzing annotation data, which stands in stark contrast to typical approaches that seek to characterize annotations and annotators in terms of consistency and accuracy. Relational and co-conceptualized phenomena, such as safety, result in challenges to measurement and consistent definition or documentation. Leveraging this imprecision, however, provides an opportunity to explore different interpretations of safety and how they manifest in annotation. Complementary to our approach, Chen and Zhang [6] use ambiguity and disagreement to flag content for deliberation annotators and help them form consistent interpretations of content. Whereas Chen et al. use ambiguity to support consensus, we conceptualize ambiguity as a way to avoid consensus and help researchers think reflectively about how they interpret and use labels. Our aim is to encourage ways to "unfix" how we use gold labels in datasets that can pose challenges to accountability [16].

5.2.1 Enhancing Ambiguity of Information. A critical underlying thread in our approach to annotations and datasets is an explicit understanding of the limits to representing complex concepts, such as safety. Thus, our analyses are shaped by a desire to understand what is represented by a label and what is not. In advocating for generating ambiguity in order to improve design, Gaver et al. call for using imprecise representations and over-interpretation to emphasize uncertainty. In other words, representing information in

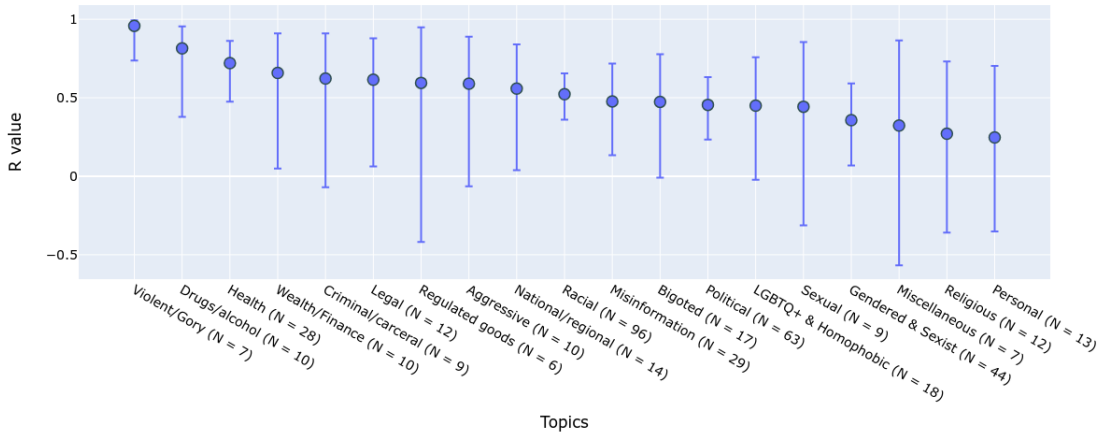

Figure 3: Correlations between gold annotations and crowd annotations by conversation topic. 'N' indicates the number of conversations per topic. Topics with five or fewer conversations are excluded. Only correlations within "Health", "Racial", "Violent/Gory", and "Political" are significant below a Bonferroni-corrected *p*-value of 0.05.

imprecise ways can bring new attentiveness to what is actually represented across a variety of signals. Because safety is a complex and multifaceted concept that must be quantified through annotation, gold labels and crowd labels stand not only as imprecise representations of safety but also *differently* imprecise representations. In our full work we also employ various analyses to try to understand and infer annotator reasoning and intent. This re-framing allows us to conduct analyses with healthy skepticism as opposed to an over-reliance on efficiency and any notion that deviations from consensus are unwanted or low quality.

5.2.2 Creating Ambiguity of Context. Our approach to analysis is oriented toward exposing ambiguous conversation contexts that annotators might differently interpret based on social and cultural factors. In doing so, we mirror Gaver et al.'s recommendation to implicate incompatible contexts to disrupt preconceptions. In annotation, the salient preconception is that ground truth is necessarily singular and fixed. At a conceptual level, we instead ask how judgments of safety reflect different forms of situated knowledge and experience. In particular we ask how these knowledge forms become encoded in ground truth judgments treated as canonical representations of safety. Many comparisons of expertise in annotation attend to cost or structure annotator recruitment simplistically as a choice between "expert" and "non-expert" (e.g., [18, 22]). Rather than discuss disagreement in pursuit of uncovering a suitable source for universal correctness or ground truth, we frame it as an opportunity to produce ground truth that is amalgamated from a combination of sources. Moreover, in contrast to typical ML annotation approaches in which consensus is both ideal and assumed to reflect identical reasoning, we pursued different analyses without specific preconceptions about how individual annotators or crowd annotators as a whole should annotate. Thus, any distribution of agreement or disagreement between annotators was an opportunity to investigate what those judgments encode, without any notion of correctness.

5.2.3 Provoking Ambiguity of Relationship. In their provocation on the ambiguity of relationships, Gaver et al. [11] propose that ambiguity draws forth a deeply personal projection of imagination and values onto design. They suggest introducing unaccustomed roles as a means to foster imagination. In our research context, rather than introducing additional unaccustomed roles, we advocate for viewing annotators in an unaccustomed manner, moving away from mere typecasting as non-expert or based on their social demographic characteristics. Against the backdrop of increasing calls in ML to collect and analyze annotator sociodemographics (e.g., [9, 19], it is important to recognize that these characteristics only partially define their identities and do not encompass the full range of their lived experiences. Moreover, Gaver et al.'s work challenge the prevailing notion that design should cater primarily to the majority [11]. Similarly, we propose a provocation against the scale of data. Instead of solely focusing on increasing the quantity of data points at the expense of diversity, we advocate for a scale that encompasses a multitude of perspectives.

6 CONCLUSION

Collecting and evaluating annotations must be responsive to the human experiences they reflect. Annotators must assemble information from various sources, including guidelines provided to them and their own lived experiences as individuals with specific socially-situated knowledge. Challenging the notion of objectivity, we propose that the development of ground truth in safety annotation tasks can be understood through the lens of ambiguity. Drawing on sociological, socio-technical, and design scholarship [4, 11, 13] we highlight the intricate nature of annotation and the need for annotators to navigate multiple sources of knowledge to construct their understanding of safety.

REFERENCES

[1] 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. arXiv:2306.11247 [cs.HC]

[2] Lora Aroyo and Chris Welty. 2014. The Three Sides of CrowdTruth. *Human Computation* 1, 1 (Sep. 2014). https://doi.org/10.15346/hc.v1i1.3

[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] Geoffrey C Bowker and Susan Leigh Star. 2001. Pure, real and rational numbers: the American imaginary of countability. *Social studies of science* 31, 3 (2001), 422–425.

[5] Joel Chan, Steven Dang, and Steven P Dow. 2016. Improving crowd innovation with expert facilitation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1223–1235.

[6] Quan Ze Chen and Amy X Zhang. 2023. Judgment Sieve: Reducing Uncertainty in Group Judgments through Interventions Targeting Ambiguity versus Disagreement. *arXiv preprint arXiv:2305.01615* (2023).

[7] 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.

[8] Mark Díaz, Razvan Amironesei, Laura Weidinger, and Iason Gabriel. 2022. Accounting for offensive speech as a practice of resistance. In *Proceedings of the sixth workshop on online abuse and harms (woah)*. 192–202.

[9] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2342–2351.

[10] Mark Díaz and Angela DR Smith. 2023. (Re) Defining Expertise in Machine Learning Development. *arXiv preprint arXiv:2302.04337* (2023).

[11] William W Gaver, Jacob Beaver, and Steve Benford. 2003. Ambiguity as a resource for design. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 233–240.

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

[13] Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies* 14, 3 (1988), 575–599.

[14] Christopher M. Homan, Greg Serapio-García, Lora Aroyo, Mark Díaz, Alicia Parrish, Vinodkumar Prabhakaran, Alex S. Taylor, and Ding Wang. 2023. Intersectionality in Conversational AI Safety: How Bayesian Multilevel Models Help Understand Diverse Perceptions of Safety. arXiv:2306.11530 [cs.HC]

[15] Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 588–602.

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

[17] Jane Im, Sarita Schoenebeck, Marilyn Iriarte, Gabriel Grill, Daricia Wilkinson, Amna Batool, Rahaf Alharbi, Audrey Funwie, Tergel Gankhuu, Eric Gilbert, et al. 2022. Women's Perspectives on Harm and Justice after Online Harassment. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–23.

[18] 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*. 557–566.

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

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

[21] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. 2021. A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–33.

[22] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on Empirical Methods in Natural Language Processing*. 254–263.

[23] Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the third workshop on abusive language online*. Association for Computational Linguistics.

[24] Lars Wissler, Mohammed Almashraee, Dagmar Monett Díaz, and Adrian Paschke. 2014. The Gold Standard in Corpus Annotation. *IEEE GSC* 21 (2014).

[25] Anbang Xu, Shih-Wen Huang, and Brian Bailey. 2014. Voyant: generating structured feedback on visual designs using a crowd of non-experts. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1433–1444.

[26] Lixiu Yu, Aniket Kittur, and Robert E Kraut. 2016. Encouraging "outside-the-box" thinking in crowd innovation through identifying domains of expertise. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1214–1222.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009