D3CODE: Disentangling Disagreements in Data across Cultures on Offensiveness Detection and Evaluation

Aida Davani
Google Research
aidamd@google.com

Mark Díaz
Google Research
markdiaz@google.com

Dylan Baker
DAIR Institute
dylan@dair-institute.org

Vinodkumar Prabhakaran
Google Research
vinodkpg@google.com

Abstract

While human annotations play a crucial role in language technologies, annotator subjectivity has long been overlooked in data collection. Recent studies that critically examine this issue are often focused on Western contexts, and solely document differences across age, gender, or racial groups. Consequently, NLP research on subjectivity have failed to consider that individuals within demographic groups may hold diverse values, which influence their perceptions beyond group norms. To effectively incorporate these considerations into NLP pipelines, we need datasets with extensive parallel annotations from a variety of social and cultural groups. In this paper we introduce the D3CODE dataset: a large-scale cross-cultural dataset of parallel annotations for offensive language in over 4.5K English sentences annotated by a pool of more than 4k annotators, balanced across gender and age, from across 21 countries, representing eight geo-cultural regions. The dataset captures annotators' moral values along six moral foundations: care, equality, proportionality, authority, loyalty, and purity. Our analyses reveal substantial regional variations in annotators' perceptions that are shaped by individual moral values, providing crucial insights for developing pluralistic, culturally sensitive NLP models.

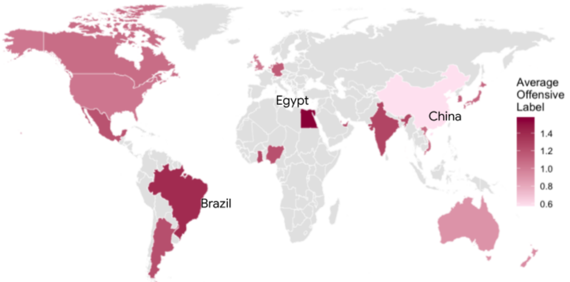


Figure 1: The distribution of labels provided from different countries. Annotators from China, Brazil, and Egypt provided significantly different labels.

1 Introduction

Designing Natural Language Processing (NLP) tools for detecting offensive or toxic text has long been an active area of research (Wulczyn et al., 2017; Founta et al., 2018). However, applying traditional NLP solutions have led to overlooking the cultural and individual factors that shape humans' varying perspectives and disagreements on what is deemed offensive (Aroyo and Welty, 2015; Waseem, 2016; Salminen et al., 2019; Uma et al., 2021; Prabhakaran et al., 2021; Akhtar et al., 2021; Wang et al., 2024). Perceiving language as offensive can depend inherently on one's beliefs and values as well as the social norms dictated by the socio-cultural context within which one's assessments are made (Eickhoff, 2018; Aroyo et al., 2019; Waseem et al., 2021; Rottger et al., 2022; Davani et al., 2023). Therefore, data curating and modeling efforts should appropriately handle such subjective factors in order to better capture and learn human perspectives about offensiveness.

As a result, recent efforts call for diversifying the rater pools as well as designing models that look beyond predicting a singular ground truth (Davani et al., 2022; Aroyo et al., 2023a). However, the efforts for diversifying annotator pools often risk reducing annotators' differences to demographic variations. In other words, subjectivity is often studied solely in relation to annotators' gender and race, within the Western context. In reality, perceptions of what is offensive extend far beyond mere differences in demographics, shaped by an individual's lived experiences, cultural background and other psychological factors (Denton et al., 2021).

Sap et al. (2022) demonstrate the impact of annotators' beliefs about racism, freedom of speech, and conservatism on their perceptions of toxicity. While individuals' systematic disagreements on

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 18511–18526
November 12-16, 2024 ©2024 Association for Computational Linguistics

notions of offensiveness reflect the complexity of their beliefs, these beliefs are often anchored in core moral values that shape their judgments. Not only do moral values influence various aspect of our cognitive processes (Greene et al., 2001), emotions (Tangney et al., 2007), and social relationships (Haidt, 2008), they also shape our judgments, motivate our behaviors, and guide our interactions. As a result, moral values, contribute significantly to our personal and cultural identity (Turiel, 2002), as we internalize societal norms and develop our moral compass (Kohlberg, 1921). Therefore, we argue that the high divergence in annotators' perceptions of offensiveness can be traced back to individuals' diverse moral values along with the cultural and social norms that dictate the boundaries of acceptable language within a society.

In this work we introduce the **D3CODE** dataset[1], built through a cross-cultural annotation effort aimed at collecting perspectives of offensiveness from 4309 participants of different age and genders across 21 countries within eight larger geo-cultural regions. Through an in-depth analysis of our dataset, we shed light on cultural and moral values that sets people apart during the annotation. We believe that this dataset can be used for assessing modeling approaches that are designed to incorporate annotators' subjective views on language, as well as for evaluating different models' cultural and moral alignment.

2 Related Work

Disagreement — even on objective tasks (Parrish et al., 2024a) — is a source of information (Jamison and Gurevych, 2015; Basile et al., 2021; Plank, 2022). Therefore, research on *perspectivism* in data (Cabitza et al., 2023) argues that treating annotators as interchangeable is ineffective when dealing with subjective language understanding tasks (Aroyo and Welty, 2013; Hovy et al., 2013; Plank et al., 2014b; Pavlick and Kwiatkowski, 2019; Dumitrache et al., 2019; Klenner et al., 2020; Díaz et al., 2022b; Weerasooriya et al., 2023a). Instead, capturing and modeling nuanced annotations and inter-annotator disagreements has been widely explored (Kairam and Heer, 2016; Founta et al., 2018; Geva et al., 2019; Chung et al., 2019; Obermeyer et al., 2019; Liu et al., 2019; Weerasooriya et al., 2020; Uma et al., 2021; Weerasooriya et al., 2023b). For instance, Rottger et al. (2022) propose a descriptive annotation paradigm for operationalizing subjectivity when surveying different beliefs.

Accordingly modeling approaches were proposed to consider variations of annotator perspectives; for instance, incorporating the item-level agreement into the loss function (Uma et al., 2020; Plank et al., 2014a), leveraging annotator disagreement as an auxiliary task along with ground-truth label prediction Fornaciari et al. (2021), or employing item-level disagreements for informing model training (Leonardelli et al., 2021; Parrish et al., 2024b). However, these methods often overlook the integrity of individual labels and aggregate diverse subjectivities into a single construct (Hovy and Yang, 2021). Research has shown that providing the age or gender of the authors to text classifiers consistently and significantly improves the performance over demographic-agnostic models Hovy (2015); Hung et al. (2023). Garten et al. (2019) model users demographics embedding, and further incorporate them into language understanding tasks. Ferracane et al. (2021) add annotators' sentiment about the writers into modeling their labels. The use of multi-task modeling has been proposed as means for model annotator perspectives separately Kennedy et al. (2020); Davani et al. (2022); Hayat et al. (2022). Others (Al Kuwatly et al., 2020; Wich et al., 2020; Akhtar et al., 2021; Orlikowski et al., 2023) capture perspectives of different groups rather than single annotators. Further approaches tend to integrated annotator differences into model predictions, through personalized model tuning (Kumar et al., 2021), jury learning (Gordon et al., 2022), and training annotator embeddings (Deng et al., 2023; Mokhberian et al., 2023).

Although attending to annotators' background is gaining more attention, documenting how annotators' identity shapes their comprehension of the world and in turn language is still missing in many data curation efforts (Díaz et al., 2022b; Scheuerman et al., 2021). A number of scholars have begun to not only document annotators' identity, but also develop principled approaches for obtaining a diversity of identities and perspectives in datasets. Aroyo et al. (2023b) developed a dataset that specifically focuses on evaluating disagreement and diverse perspectives on conversational safety, and (Homan et al., 2024) leverages this same dataset to proporse a multilevel modeling approach for

[1] https://github.com/google-research-datasets/D3code

measuring annotation differences across a range of sociodemographic groups. Pei and Jurgens (2023) recruit a representative sample of annotators across sex, age, and race in the US and show the impact of annotators' background on their judgments.

The social nature of language means that sociocultural differences play meaningful roles in how individuals use language, such as offensive speech (Goyal et al., 2022; Díaz et al., 2022a). Sachdeva et al. (2022) apply Item Response Theory to capture the impact of annotators' group identity in their evaluation of harmful language. Salminen et al. (2018) and Lee et al. (2023) demonstrate statistically significant variations across countries in hate speech annotations. Sap et al. (2022) draw from social psychology research to demonstrate the impact of annotator identities and beliefs about hate speech, free speech, and racist beliefs, on their annotations of toxicity. Davani et al. (2023) investigates annotators' biases and disagreements as related to their social stereotypes.

One such nuanced factor, often not studied in AI research, is morality. Moral considerations play significant roles in how humans navigate prejudicial thoughts and behaviors (Molina et al., 2016), often manifesting in language through offensive content. The interplay between morality and group identity (Reed II and Aquino, 2003) influences many aspects of our social dynamics, including perceptions, interactions, stereotypes, and prejudices. Moreover, research in computational social science addressing harmful language reveals a concurrent occurrence of moral sentiment alongside expressions of hatred directed at social groups (Kennedy et al., 2023).

In this paper we introduce the D3CODE dataset which not only provides social factors and demographic information regarding annotators but also considers the moral values that may vary across regions and among individuals. Such information facilitates drawing connections between annotations from culturally diverse annotators, the sociocultural norms shaping their environment, and the moral values they hold.

3 D3CODE Dataset

In order to study a broad range of cultural perceptions of offensiveness, we recruited 4309 participants from 21 countries, representing eight geocultural regions, with each region represented by 2-4 countries (Table 1).[2] We discuss the reasoning behind our selection of countries and regions in more depth in Appendix A.1; however, the final selection of countries and regions was chosen to maximize cultural diversity while balancing participant access through our recruitment panel. Participants were recruited through an online survey pool, compensated in accordance to their local law, and were informed of the intended use of their responses. In order to capture the participants' perceptions of offensiveness, we asked each participant to annotate offensiveness of social media comments selected from Jigsaw datasets (Jigsaw, 2018, 2019). Furthermore, we also asked them to respond to a measurement of self-reported moral concerns, using the Moral Foundations Questionnaire (MFQ-2; Graham et al., 2013; Atari et al., 2023).[3]

		Gender			Age		
Region	**#**	**M**	**W**	**Other**	**18–30**	**30–50**	**50+**
AC.	516	306	205	5	269	168	79
ICS.	554	308	245	1	237	198	119
LA.	549	271	275	3	302	176	71
NA.	551	220	325	6	263	175	113
Oc.	517	203	307	7	161	221	135
Si.	540	280	249	11	208	228	104
SSA.	530	309	219	2	320	157	53
WE.	552	252	294	6	259	172	121

Table 1: Demographic distribution of annotators from each region, region names are shortened and represent: Arab Culture (AC.), Indian Cultural Sphere (ICS.), Latin America (LA.), North America (NA.), Oceania (Oc.), Sinosphere (Si.), Sun-Saharan Africa (SSA), and Western Europe (WE.).

3.1 Recruitment

Recruitment criteria account for various demographic attributes: (1) *Region of residence*: we recruited at least 500 participants from each of the eight regions with at least 100 participants per country, except for South Korea and Qatar where we managed to recruit only a smaller number of raters (See Table 5), (2) *Gender*: within regions, we set a maximum limit of 60% representations for Men and Women separately (for a loosely balanced representation of the two genders), while including options for selecting "non-binary / third

[2] We based the categorization of regions loosely on the UN Sustainable Development Goals groupings `https://unstats.un.org/sdgs/indicators/regional-groups` with minor modifications: combining Australia, NZ and Oceania to "Oceania", and separating North America and Europe, to facilitate easier data collection.

[3] The data card and dataset will be available upon the paper

gender," "prefer not to say," and "prefer to self identify" (with a textual input field). We recognize that collecting non-binary gender information is not safe for annotators in many countries, so we limited the specification of recruitment quota to binary genders to ensure consistency across countries. (3) *Age*: in each region at most 60% of participants are 18 to 30 years old and at least 15% are 50 years old or older. We specifically aimed to ensure adequate representation of annotators of age 50 or older, because this age group have lower engagement with crowdsourcing platforms but are equally impacted by technology advancements. Table 1 provides the final distribution of participants across different demographic groups in each region.

We further set an exclusion criterion based on *English fluency* since our study is done on English language text; we only selected participants who self-reported a high level of proficiency in reading and writing English. We performed this study in the English language, as the most wide-spoken language across the globe, to simulate the most common data annotation settings, in which annotators (who are no necessarily English speakers) are asked to interact with and label textual data in English. Additionally, we collected participants' self-reported subjective socio-economic status (Adler et al., 2000) that may serve as a potential confound in follow-up analyses.

3.2 Annotation items

We performed this study in the English language. In order to collect textual items for participants to annotate, we selected items from Jigsaw's Toxic Comments Classification dataset (Jigsaw, 2018), and the Unintended Bias in Toxicity Classification dataset (Jigsaw, 2019). We built a dataset of N_{items} = 4554 consisting of three categories of items: (1) potentially high inter-annotator disagreement, (2) evoking moral sentiment, or (3) including language about specific social groups:

(1) Random: As the basic strategy, we randomly select 50% of the data from items that are likely to evoke disagreement. To measure disagreements on each item, we averaged the toxicity scores assigned to the item in the original dataset, ranging from 0 (lowest toxicity) to 1 (highest toxicity). Items on the two ends of the range evoke no disagreement because all annotators labeled them either as toxic or non-toxic. Therefore, we chose items with a normal distribution centered around a toxicity score of 0.5 (indicating highest disagreement) with a standard deviation of 0.2.

(2) Moral Sentiment: 10% of the dataset consists of a balanced set of items include different moral sentiments, identified through a supervised moral language tagger trained on the MFTC dataset (Hoover et al., 2020). This strategy is aimed at enabling follow up studies to investigate potential content-level correlates of disagreements, particularly as previous computational social science studies on harmful language have shown specific correlation of moral sentiment with expressions of hatred. Our tagger identified very few items with moral sentiment throughout the dataset, selecting a balances set led to a set of 500 such items.

(3) Social Group Mentions: Finally, the rest (40%) of the dataset consists of a balanced set of items that mention specific social group identities related to gender, sexual orientation, or religion (Jigsaw dataset includes manual annotation of these identity terms, which we used for our sampling). We specifically selected such items as online harmful language is largely directed at specific social groups and resonates real-world group conflicts.

3.3 Annotation task

Each participant was tasked with labeling 40 items on a 5-point Likert scale (from *not offensive at all* to *extremely offensive*). Half of the participants were provided with a note that defined *extremely offensive language* as *"profanity, strongly impolite, rude or vulgar language expressed with fighting or hurtful words in order to insult a targeted individual or group."* Other participants were expected to label items based on their own definition of offensiveness. The latter group served as a control setting of participants who are expected to lean on their individual notion of offensiveness.[4].

In case of unfamiliarity with the annotation item, participants were asked to select the option *"I do not understand this message."* Participants' reliability was tested by 5 undeniably non-offensive, control questions randomly distributed among the 40-items annotation process. Those who failed at least one quality control check were removed, and not counted against our final set of 4309 participants (refer to Appendix A.2 for test items). Each item in the final dataset was labeled by at least three

[4] We did not explicitly ask participants to provide their definition of offensiveness

participants from each region who passed the control check (a total of 24 labels). Participants were compensated at rates above the prevalent market rates for the task (which took at most 20 minutes, with a median of 13 minutes), and respecting the local regulations regarding minimum wage in their respective countries.

3.4 Moral Foundation Questionnaire

After annotation, participants were also asked to fill out the Moral Foundations Questionnaire (MFQ-2; Graham et al., 2013; Atari et al., 2023), which assesses their moral values along six different dimensions: *Care*: "avoiding emotional and physical damage to another individual," *Equality*: "equal treatment and equal outcome for individuals," *Proportionality*: "individuals getting rewarded in proportion to their merit or contribution," *Authority*: "deference toward legitimate authorities and the defense of traditions," *Loyalty*: "cooperating with ingroups and competing with outgroups," and *Purity*: "avoiding bodily and spiritual contamination and degradation" (Atari et al., 2023). We specifically rely on the MFQ-2 because it is developed and validated through extensive cross-cultural assessments of moral judgments. This characteristic makes the questionnaire a reliable tool for integrating a pluralistic definition of values into AI research. The questionnaire includes 36 statements to assess participants' priorities along each of the six foundations (see Figure 5 which shows one of the MFQ-2 questions in our survey). For instance, one MFQ-2 statement that targets the *Care* foundation is: "*Everyone should try to comfort people who are going through something hard*". We aggregate each participant's responses to compute a value between 1 to 5 to capture their moral foundations along each of these dimensions.

4 Analyses

Our analyses focus on annotators' varying perspectives and how shared social, cultural or moral attributes can help shed light on annotation behaviors. We begin by analyzing how different groups vary on expressing their lack of understanding the message by selecting the *"I don't understand this message"* option. We then study annotators' geo-cultural regions and moral values in relation to their annotations. Specifically, we consider annotator clustering either based on their similar moral values or their region of residence, and assess in-group homogeneity and out-group disagreements for clusters. The remainder of this section delves deeper into how groups of annotators from the same region or with similar moral values tend to label content differently.

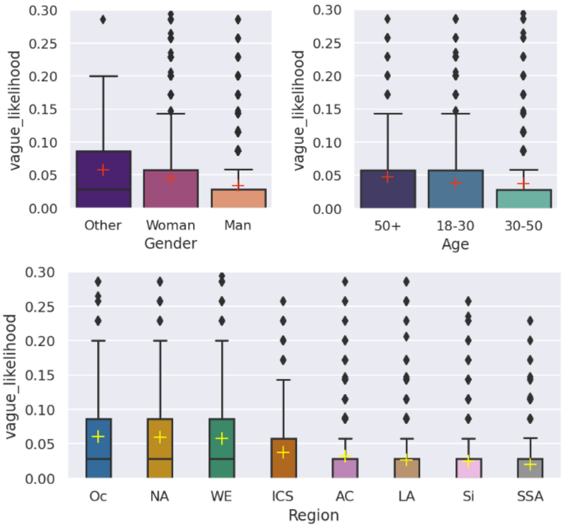


Figure 2: The likelihood of an annotator not understanding the message, grouped based on their socio-demographic information. Annotators identifying as Men, or of 50 years of old or younger are generally less likely to state they did not understand a message.

4.1 Lack of Understanding

We start our analyses by investigating the patterns of annotators not understanding the provided text. While recent modeling efforts have shown the practical ways in which annotators' ambiguity or confidence can help inform the model. However, in many data annotation efforts, annotators' lack of understanding is either not captured or discarded. We ask whether specific groups of annotators are more likely to not understand the annotation item, and as a result, their responses are more likely to be discarded. We compared annotators with different demographics (along Gender, Age, and Region) on how likely they are to select the "I don't understand" answer (Figure 2). All further studies of the paper relies on the dataset after removing these answers.

Gender: When grouping annotators based on their gender or age, Men are overall less probable to state lack of understanding (M = .03, SD = .07), compared to Women (M = .05, SD = .08, $p <$ .001), and other genders (M = 0.06, SD = .07, $p =$ .03). However, Women and other genders did not differently select this label ($p =$ .34).

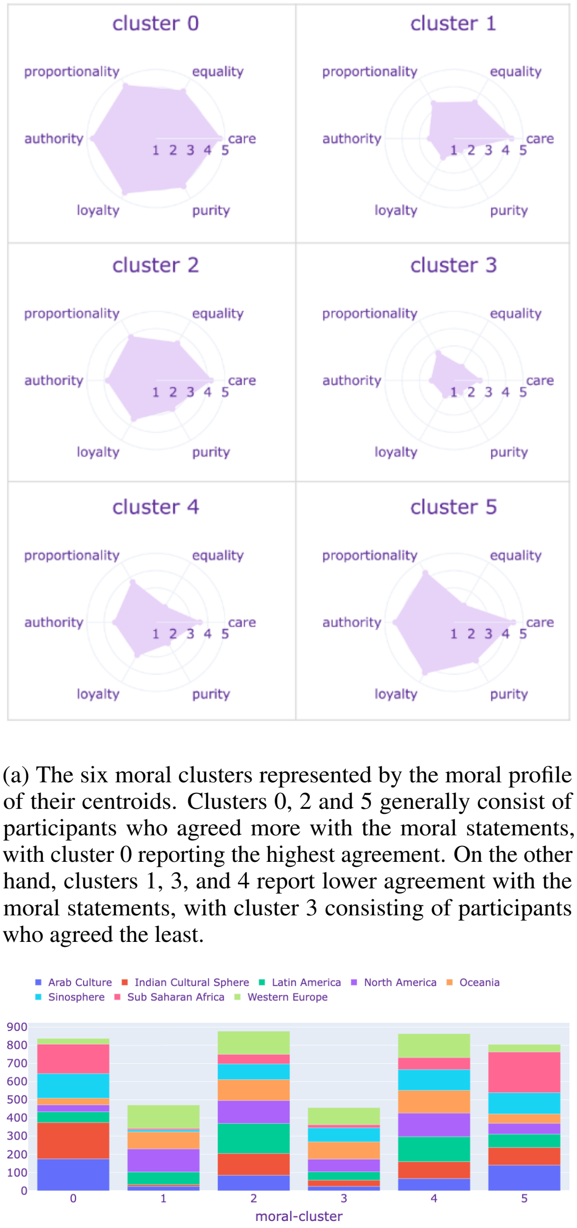
Age: Participants who were aged 50 or more were more likely to state lack of understanding (M = .05, SD = .09), compared to 30–50 year-old (M = .04, SD = .08, $p < .01$), and 18–30 year-old (M = .04, SD = .07, $p < .01$). The difference of the latter two groups was insignificant ($p = .85$)

Region: We further looked into the regional differences in not understanding the answers; a pairwise Tukey test shows that annotators from Oceania (M = 0.06, SD = 0.1), North America (M = 0.06, SD = 0.09), and Western Europe (M = 0.06, SD = 0.09) were all significantly more probably to state lack of understanding compared to Indian Cultural Sphere (M = 0.04, SD = 0.08), Arab Culture (M = 0.03, SD = 0.06), Latin America (M = 0.03, SD = 0.06), Sinosphere (M = 0.02, SD = 0.07), and Sub Saharan Africa (M = 0.02, SD = 0.05) with all p values lower than .05.

It is important to note that distinguishing between lack of understanding due to annotator limitations versus inherent ambiguity in the data is crucial for robust analysis and dataset curation. While our current data does not provide a reliable way to make this distinction, future work could explore strategies that combine annotator responses with text analysis techniques to identify data points that are objectively vague or difficult to understand.

4.2 Morally Aligned Annotators

To systematically study annotators' perspectives with regard to varying moral values we first cluster annotators into groups with high internal moral similarity. We used an unsupervised data-driven approach for K-nearest neighbors clustering with an Elbow method for determining the optimal number of clusters (see Appendix A.4). Figure 3a represents the resulting six clusters by the average moral values of their members. Figure 3b represents the distribution of annotators from different regions across the six moral clusters. As shown by the plots, regions have varying presence in the moral clusters; cluster 0 consists of annotators who agreed most with all dimensions of the moral foundations questionnaire, most participants in this cluster are from Indian Cultural Sphere, Sub Saharan Africa and Arab Culture. On the other hand, cluster 3 includes annotators who agreed the least with MFQ-2 values along most dimensions; while this cluster has the fewest annotators, most of them were from Western Europe, Oceania, and Sinosphere, in our data. Other 4 clusters each have their specific distribution of moral values across the axes, that show the most prevalent moral values in the annotator pool.



(a) The six moral clusters represented by the moral profile of their centroids. Clusters 0, 2 and 5 generally consist of participants who agreed more with the moral statements, with cluster 0 reporting the highest agreement. On the other hand, clusters 1, 3, and 4 report lower agreement with the moral statements, with cluster 3 consisting of participants who agreed the least.

(b) Distribution of participants from different regions across different moral clusters. Variances of regional presence are noticeable in several cases, e.g., cluster 0 mostly consists of participants from Indian Cultural Sphere, Arab Culture, and Sub-Saharan Africa.

Figure 3

In general, while our results replicated previous evidence of regional differences in specific moral values (Atari et al., 2023) (e.g., more collective cultures such as Arab Culture score higher on values such as Purity and Authority), our analysis also demonstrates that cultural differences are not enough to completely explain annotators' moral preferences, as none of the clusters perfectly align with cultural regions.

Dimension	Group	IRR	XRR	GAI
Region	AC.	↑0.13**	↑0.11	↑1.17*
	ICS.	↓0.10	↓0.10*	↑1.04
	LA.	↑0.13**	↑0.11	↑1.15*
	NA.	↑0.14**	↑0.11	↑1.31**
	Oc.	↑0.12	↓0.10	↑1.15*
	Si.	↓0.09*	↓0.09**	↓1.00
	SSA.	↑0.14**	↓0.10	↑1.36**
	WE.	↑0.14**	↑0.11	↑1.22**
Moral Cluster	0	↑0.12*	↑0.12**	↑1.05
	1	↑0.12	↑0.11	↑1.04
	2	↑0.18**	↑0.12**	↑1.46**
	3	↓0.07**	↓0.10**	↓0.75**
	4	↑0.11	↑0.11	↑1.00
	5	↓0.09*	↓0.09**	↓0.97

Table 2: Results for in-group and cross-group cohesion, and GAI. Significant results are in **bold**: * for significance at $p < 0.05$, ** for significance after Benjamini-Hochberg correction. A ↓ (or ↑) means that the result is less (or greater) than expected under the null hypothesis. GAI results based on C_X = XRR and C_I = IRR.

4.3 Disagreement among Groups

Additionally, we explore the homogeneity of annotations within various clusters of annotators. We specifically compare moral clusters' homogeneity with the alternative clustering approach that considers annotators of the same region to have similar perceptions. We considered region as an alternative means for clustering annotators because collected annotations tend to vary significantly across regions and countries (the distribution of ratings collected from different countries is provided in Figure 6). Inspired by Prabhakaran et al. (2023), we use the Group Association Index (GAI) metric which provides a measurement of perspective diversities within annotator groups. In other words, for each specific group of annotators, GAI provides the ratio of an in-group measurement of agreement to a cross-group measurement of cohesion. In our specific case, we measure in-group agreement through Inter-Rater Reliability (IRR), and cross-group cohesion through Cross-Replication Reliability (XRR; Wong et al., 2021). The GAI metric is then defined as the ratio to IRR to XRR, and a value more than 1 reports higher internal vs. external cohesion.

Table 2 includes the results for six moral clusters and eight geo-cultural groups of annotators. In all 14 cases, we assessed the agreement between each specific group and the combined responses of all other annotators. While the highest GAI score is achieved by one of the moral cluster (cluster 2, with low agreement with purity values and moderated moral values on all other axes), moral cluster in general have high variation in their homogeneity. On the other hand, regional clusters are generally more distinct in their perspectives.

4.4 Disagreement on Categories of Content

We further analyze the various types of content that annotators may label as offensive. As outlined in Section 3, annotated items are chosen using three strategies: random selection, morality-based selection, and social identity-based selection. Figure 4 shows that annotators tend to have varying degrees of disagreement (calculated as the standard deviation of labels assigned to the item) when labeling items selected based on different strategies. As the plot shows, items that mention specific social identity groups evoke highest levels of disagreement (Mean = .47, SD = .06), significantly higher than items with moral sentiment (Mean = .31, SD = .16) and the randomly selected items (Mean = .41, SD = .10), both with p < .001. It is important to note that our randomly selected items were deliberately chosen from those with high disagreement in the original Jigsaw dataset. Our analysis indicates that items mentioning social identity groups tend to evoke even more disagreement.

In addition to disagreement between annotators, items can be labeled differently by various groups of annotators. The aggregated labels from each region demonstrates how recruiting annotators from specific regions could lead to having thoroughly different final dataset. Table 3 represents items with high cross-region disagreement.

5 Discussion

Research on safety considerations of large language models has mostly focused on evaluations of model harms through crowdsourced benchmarks (Srivastava et al., 2023; Wang et al., 2022). However, while annotators from different regions are shown to have different perspectives regarding this task (Salminen et al., 2018; Lee et al., 2023), current benchmarks fail to represent the cultural and individual variations in human moral judgements about generated language and model outputs. They also lack comprehensive understanding of human values and cultural norms that drive diversity of perspectives in annotations. This work presents a cross-cultural experiment with participants across various cultural and demographic backgrounds.

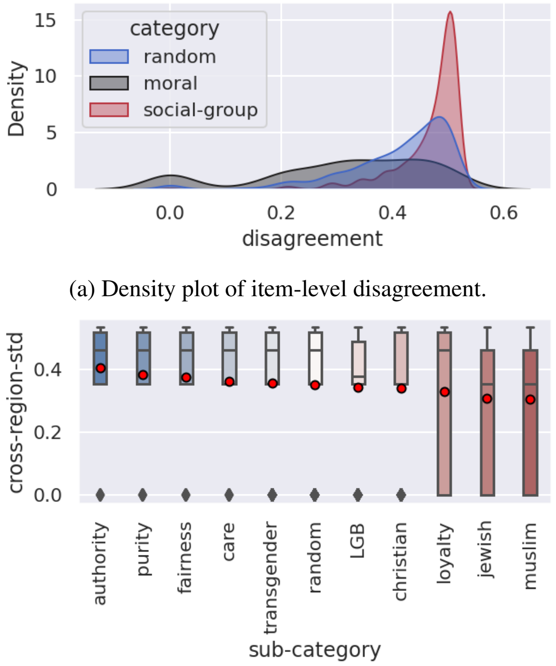
	Offensive		
Item	Yes	No	Category
Transgender athletes..... Transgender students..... This is what feminism gets you in society. This was unheard of only a short 20 years ago.	AC, LA, NA, Oc	ICS, Si, SSA, WE	transgender
Does pointing out that a growing majority of Americans support adultery (abortion, same sex marriage, even theft), change God's law an iota?	AC, ICS, LA, NA	Oc, Si, SSA, WE	LGB
The women came out because he is a hypocrite, a child predator. Come on trumpers, defend the child predator Christian, lol	ICS, LA, Oc, SSA	AC, NA, Si, WE	christian

Table 3: Highest cross-regional disagreement.

Our dataset captures valuable insights into human perceptions on offensive language, revealing demographic differences in annotation certainty, and regional, as well as moral psychological variations in perceiving offensiveness.

Our first analyses captures how participants with different demographic background might express their unfamiliarity with the annotation. In general, annotators not identifying as Men and annotators aged 50 and above are more likely to select the "I don't understand" option. Moreover, annotators from Oceania, North America, and Western Europe were significantly more probably to state that they did not understand the message compared to Indian Cultural Sphere, Arab Culture, Latin America, Sinosphere, and Sub Saharan Africa. Therefore, it is important to note how this kind of uncertainty in annotating might occurred disproportionately for different groups.

Our dataset also represent different categories of content within a well-known machine learning corpus, with annotators having varying levels of disagreement for labeling content from different categories. While items with moral sentiment are the least likely to evoke disagreement, items mentioning specific social groups are more likely to have a varying range of annotation. This finding replicates previous findings on how group perception and stereotypes affect harm perception targeting different social groups, in a cross-cultural context. Consequently, these findings underscore the need for further research into social dynamics within diverse cultural contexts to better understand and mitigate harmful risks of language technologies.



(a) Density plot of item-level disagreement.

(b) Cross-regional disagreements (standard deviation of majority votes from all regions) or each sub-category.

Figure 4: Items related to social groups (christian, transgender, Jewish, Muslim and LGB) generally evoke more disagreement compared to random items.

Our dataset and findings demonstrate the potential for incorporating diverse cultural and individual moral perspectives to enhance fairness in LLMs. By moving beyond traditional socio-demographic groupings and understanding how individual moral values shape perceptions, we provide a pathway for defining model alignment strategies that capture a broader range of human values, thus contributing to the approaches to address representation biases by challenging value alignment frameworks that prioritize normative cultural or societal values.

6 Conclusion

We introduce the D3CODE dataset, which captures the results of a cross-cultural annotation experiment for understanding disagreements on perceiving offensiveness in language. Our findings reveal significant demographic and regional variations in perceptions of offensive language, underlining the necessity of incorporating diverse perspectives into reinforcement learning with human feedback. Additionally, the dataset showcases differences in annotation certainty and disagreement levels across various content categories, particularly concerning mentions of specific social groups. These findings underscore the imperative for further research into social dynamics within diverse cultural contexts to mitigate the risks associated with harmful language in language technologies and promote fairness and inclusivity in digital interactions.

Limitations

In our work, we focus on moral foundations as a way to measure differences in values across groups; however, values can be measured in other ways, including other psychological questionnaires such as the Schwartz's value survey (Lindeman and Verkasalo, 2005) or the World Value survey (Inglehart et al., 2000) as well as through interviews, case studies and ethnography. Importantly, while our annotator sample represents diverse cultural perspectives, the items in our dataset are in English, which may explain the different rates of "I don't know" responses observed across regions. Moreover, English data likely features lower representation of certain content, such as offensive content about social groups, celebrations, or politics specific to certain regions and languages. In addition, to preserve our ability to compare data cross-culturally, we focused on demographic categories that are broadly recognized. As a result, we did not conduct analyses of demographic differences that are specific to particular cultural regions, such as caste, and we did not collect highly sensitive demographic information, such as sexual orientation. We acknowledge that salient social categories can differ greatly across geo-cultural reasons, therefore our selection of categories should not be considered exhaustive. Finally, our selection of countries within each cultural region was informed by access feasibility via our data collection platform, which may have introduced unexpected sampling biases.

Our clustering approach is unsupervised and data-driven, with the primary goal of identifying distinct groups of annotators that behave similarly. While not all clusters have immediate intuitive interpretations, they each represent a group of annotators who share similar moral values, differentiating them from other groups. We avoided overinterpreting the patterns in all clusters but we do acknowledge that further qualitative exploration of these clusters is needed to extract more insight about the annotators in each group.

We chose language proficiency as a criteria since most widely used NLP-based content moderation tools tend to focus on English, and English is a language spoken across diverse cultures across the globe that facilitates such a study. We acknowledge that as a result our participant pool may not be a good population representation across regions. However, our primary aim is not to comprehensively capture regional moral differences (a question addressed in social psychology research, e.g., Atari et al. (2023)). Instead, we focus on demonstrating how biases creep into the ML pipeline, as a result of existing crowdsourcing efforts for English content/data annotations relying on English speakers without accounting for the cultural differences in countries where English is not the first language. According to your helpful feedback, we will discuss the motivation of requiring English proficiency, and acknowledge the limitations it entails.

Ethics Statement

In this work, we collected and modeled annotator responses primarily to demonstrate geocultural differences. Our results and approaches are not meant to be used to define user preferences or platform policies. For example, a subgroup's higher or lower tendency to identify content as offensive does not necessarily mean that content moderation policies should differ for that group. In addition, our work does not advocate for treating any particular cultural group's labels as more "correct" than those of another cultural group.

Acknowledgements

We would like to express our gratitude to Alicia Parrish, and Kathy Meier-Hellstern for their support and insightful feedback on this paper. We appreciate Blake Schritter and James Stewart for their support throughout the data collection process.

References

Nancy E Adler, Elissa S Epel, Grace Castellazzo, and Jeannette R Ickovics. 2000. Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy, white women. *Health psychology*, 19(6):586.

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. *arXiv preprint arXiv:2106.15896*.

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, Mark Diaz, Christopher Homan, Vinodkumar Prabhakaran, Alex Taylor, and Ding Wang. 2023a. The reasonable effectiveness of diverse evaluation data.

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In *Companion proceedings of the 2019 world wide web conference*, pages 1100–1105.

Lora Aroyo, Alex S Taylor, Mark Díaz, Christopher M Homan, Alicia Parrish, Greg Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023b. DICES dataset: Diversity in conversational ai evaluation for safety. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013(2013).

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

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2023. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 125.

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*, Online. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

John Joon Young Chung, Jean Y Song, Sindhu Kutty, Sungsoo Hong, Juho Kim, and Walter S Lasecki. 2019. Efficient elicitation approaches to estimate collective crowd answers. *CSCW*, pages 1–25.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

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.

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.

Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554*.

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

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

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.

Carsten Eickhoff. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 162–170.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2021. Did they answer? Subjective acts and intents in conversational discourse. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1644, 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.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Justin Garten, Brendan Kennedy, Joe Hoover, Kenji Sagae, and Morteza Dehghani. 2019. Incorporating demographic embeddings into language understanding. *Cognitive science*, 43(1):e12701.

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.

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*, pages 1–19.

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. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

Joshua D Greene, R Brian Sommerville, Leigh E Nystrom, John M Darley, and Jonathan D Cohen. 2001. An fmri investigation of emotional engagement in moral judgment. *Science*, 293(5537):2105–2108.

Jonathan Haidt. 2008. Morality. *Perspectives on psychological science*, 3(1):65–72.

Hassan Hayat, Carles Ventura, and Agata Lapedriza. 2022. Modeling subjective affect annotations with multi-task learning. *Sensors*, 22(14):5245.

Christopher Homan, Gregory Serapio-Garcia, Lora Aroyo, Mark Diaz, Alicia Parrish, Vinodkumar Prabhakaran, Alex Taylor, and Ding Wang. 2024. Intersectionality in AI safety: Using multilevel models to understand diverse perceptions of safety in conversational AI. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 131–141, Torino, Italia. ELRA and ICCL.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 752–762.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

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*, pages 588–602.

Chia-Chien Hung, Anne Lauscher, Dirk Hovy, Simone Paolo Ponzetto, and Goran Glavaš. 2023. Can demographic factors improve text classification? revisiting demographic adaptation in the age of transformers. In *Findings of the 2023 Association for Computational Linguistics*.

Ronald Inglehart, Miguel Basanez, Jaime Diez-Medrano, Loek Halman, and Ruud Luijkx. 2000. World values surveys and european values surveys, 1981-1984, 1990-1993, and 1995-1997. *Ann Arbor-Michigan, Institute for Social Research, ICPSR version*.

Emily Jamison and Iryna Gurevych. 2015. Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297, Lisbon, Portugal. Association for Computational Linguistics.

Jigsaw. 2018. Toxic comment classification challenge. Accessed: 2021-05-01.

Jigsaw. 2019. Unintended bias in toxicity classification. Accessed: 2021-05-01.

Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *CSCW*.

Brendan Kennedy, Preni Golazizian, Jackson Trager, Mohammad Atari, Joe Hoover, Aida Mostafazadeh Davani, and Morteza Dehghani. 2023. The (moral) language of hate. *PNAS nexus*, 2(7):pgad210.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Manfred Klenner, Anne Göhring, and Michael Amsler. 2020. Harmonization sometimes harms. *CEUR Workshops Proc.*

Lawrence Kohlberg. 1921. *The philosophy of moral development: Moral stages and the idea of justice*, volume 1. San Francisco: harper & row.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 299–318.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Juho Kim, and Alice Oh. 2023. Crehate: Cross-cultural re-annotation of english hate speech dataset. *arXiv preprint arXiv:2308.16705*.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marjaana Lindeman and Markku Verkasalo. 2005. Measuring values with the short schwartz's value survey. *Journal of personality assessment*, 85(2):170–178.

Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher M. Homan. 2019. Learning to predict population-level label distributions. In *HCOMP*.

Negar Mokhberian, Myrl G Marmarelis, Frederic R Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2023. Capturing perspectives of crowdsourced annotators in subjective learning tasks. *arXiv preprint arXiv:2311.09743*.

Ludwin E Molina, Linda R Tropp, and Chris Goode. 2016. Reflections on prejudice and intergroup relations. *Current Opinion in Psychology*, 11:120–124. Intergroup relations.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*.

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, Susan Hao, Sarah Laszlo, and Lora Aroyo. 2024a. Is a picture of a bird a bird? a mixed-methods approach to understanding diverse human perspectives and ambiguity in machine vision models. pages 1–18.

Alicia Parrish, Vinodkumar Prabhakaran, Lora Aroyo, Mark Díaz, Christopher M. Homan, Greg Serapio-García, Alex S. Taylor, and Ding Wang. 2024b. Diversity-aware annotation for conversational AI safety. In *Proceedings of Safety4ConvAI: The Third Workshop on Safety for Conversational AI @ LREC-COLING 2024*, pages 8–15, Torino, Italia. ELRA and ICCL.

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. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. pages 10671–10682.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014a. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014b. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Alicia Parrish, Alex Taylor, Mark Díaz, and Ding Wang. 2023. A framework to assess (dis)agreement among diverse rater groups. *arXiv preprint arXiv:2311.05074*.

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.

Americus Reed II and Karl F Aquino. 2003. Moral identity and the expanding circle of moral regard toward out-groups. *Journal of personality and social psychology*, 84(6):1270.

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 S Sachdeva, Renata Barreto, Claudia von Vacano, and Chris J Kennedy. 2022. Assessing annotator identity sensitivity via item response theory: A case study in a hate speech corpus. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1585–1603.

Joni Salminen, Hind Almerekhi, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J Jansen. 2019. Online hate ratings vary by extremes: A statistical analysis. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 213–217.

Joni Salminen, Fabio Veronesi, Hind Almerekhi, Soon-Gvo Jung, and Bernard J Jansen. 2018. Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings. In *2018 fifth international conference on social networks analysis, management and security (snams)*, pages 88–94. IEEE.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

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

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

June Price Tangney, Jeff Stuewig, and Debra J Mashek. 2007. Moral emotions and moral behavior. *Annu. Rev. Psychol.*, 58:345–372.

Elliot Turiel. 2002. *The culture of morality: Social development, context, and conflict*. Cambridge University Press.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 173–177.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2022. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models.

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.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? Annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied machine learning: On the illusion of objectivity in nlp.

Tharindu Cyril Weerasooriya, Tong Liu, and Christopher M. Homan. 2020. Neighborhood-based pooling for population-level label distribution learning. In *ECAI*.

Tharindu Cyril Weerasooriya, Sarah Luger, Saloni Poddar, Ashiqur KhudaBukhsh, and Christopher Homan. 2023a. Subjective crowd disagreements for subjective data: Uncovering meaningful CrowdOpinion with population-level learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 950–966, Toronto, Canada. Association for Computational Linguistics.

Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. 2023b. Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with DisCo. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4679–4695, Toronto, Canada. Association for Computational Linguistics.

Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020. Investigating annotator bias with a graph-based approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 191–199, Online. Association for Computational Linguistics.

Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication reliability–an empirical approach to interpreting inter-rater reliability. *arXiv preprint arXiv:2106.07393*.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

A Appendix

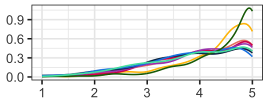
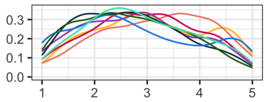
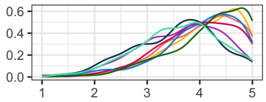
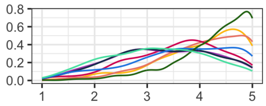
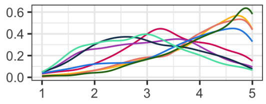
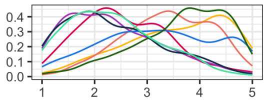
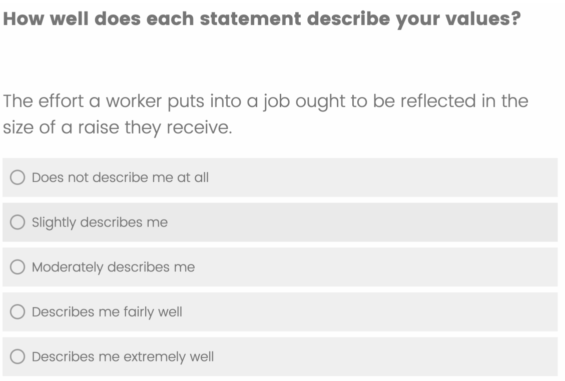
Foundation		F(7, 4287)
Care		34.48*
Equality		13.37*
Propor.		51.24*
Authority		102.25*
Loyalty		158.30*
Purity		203.55*
Arab Culture Indian Cultural Sphere Latin America North America Oceania Sinosphere Subsaharan Africa Western Europe		

Table 4: Distribution of the moral foundations scores and the results of one-way ANOVA analysis conducted for each moral foundation across regions. *means the p-value of the analysis is lower than .001



How well does each statement describe your values?

The effort a worker puts into a job ought to be reflected in the size of a raise they receive.

○ Does not describe me at all

○ Slightly describes me

○ Moderately describes me

○ Describes me fairly well

○ Describes me extremely well

Figure 5: Sample of MFQ-2 questions in our survey

A.1 Regions and Countries

Our selected list of geo-cultural regions and countries within regions is not meant to be exhaustive, rather just to make sure that our study is done on a set of countries with diverse cultural histories. Each region listed has countries and sub-regions that have distinct cultural practices, and it is wrong to assume that the country we choose would comprehensively represent that region. Similarly, the countries listed are meant as likely places to collect data from, based on familiarity with previous data collection efforts, which potentially reflect the power structures existing within those regions. Also, each country is rarely a monolith in terms of culture (e.g., India has diverse subcultures, Australia being characterized as a "Western" culture erases the vibrant Australian Aboriginal culture). Collected data also reflects the local disparities in who tends to be in the social strata that maximally overlaps with the data-collection workforce in respective regions, and subcultures they represent.

A.2 Test Items

We used participants' response to 5 test items to evaluate their understanding of offensiveness. These items (listed below) were unanimously labeled as *not offensive at all* by authors and were used as an attention check to remove participants who annotated them otherwise.

- Could you explain your point in a little more detail please.
- The NY Times yesterday had a rebuttal by the Iranian UN ambassador.
- Ditto to everything said here.
- Just looked at the link. Impressive and missed opportunity.
- Don't be so hard on yourself. Your life will go on.

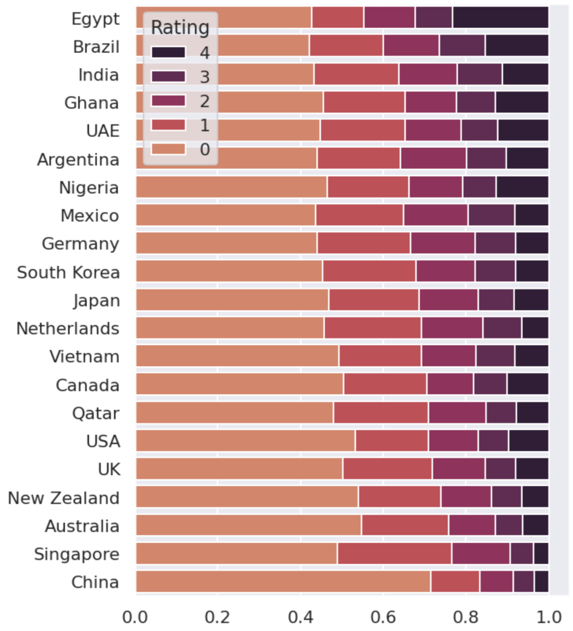


Figure 6: Distribution of the different labels provided by annotators of different countries. The y-axis is sorted based on the average offensive label captured in each country.

Region	Country
Arab Culture	Egypt, Qatar, UAE
Indian Cultural Sphere	India, Singapore
Latin America	Argentina, Brazil, Mexico
North America	Canada, USA
Oceania	Australia, New Zealand
Sinosphere	China, Japan, South Korea, Vietnam
Sub-Saharan Africa	Ghana, Nigeria
Western Europe	Germany, Netherlands, UK

Table 5: List of regions and countries within them in our dataset.

A.3 Data Cleaning

We selected thresholds for the amount of time needed to finish the survey and removed annotators who performed the task either quicker or slower than the expectation. Annotators with similar answers to all items were also removed from the data.

A.4 Moral clusters

Figure 7 shows the plot of “distortions” that led to us selecting 6 as the optimal number of moral clusters: for each potential value of k (number of clusters), the distortion (average sum of squared distance between each data point to the centroid) is calculated. Distortion measures how tightly grouped the data points are within each cluster. Lower distortion indicates better clustering. In Figure 7, we can observe that $k = 6$ is the point where adding more clusters does not significantly decrease the distortion. In other words, it strikes a balance between maximizing the distinctness of clusters and minimizing the complexity of the model.

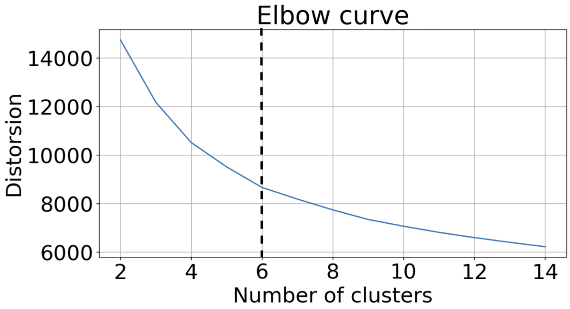


Figure 7: The distortion value captured for different options for number of moral clusters.