A Framework to Assess (Dis)agreement Among Diverse Rater Groups

Anonymous ACL submission

Abstract

Human annotation plays a core role in machine learning — annotations for supervised models, safety for generative models, and human feedback for reinforcement learning, to cite a few avenues. However, the fact that many of these human annotations are inherently subjective is often overlooked. Recent work has demonstrated how ignoring rater subjectivity (typically resulting in rater disagreement) is problematic within specific tasks and for specific subgroups. Generalizable methods to harness rater disagreement and thus understand the socio-cultural leanings of subjective tasks remains an open challenge. In this paper, we propose a comprehensive disagreement analysis framework to measure *group association* in perspectives among different rater subgroups, and demonstrate its utility in assessing the extent of systematic disagreements in two datasets: (1) safety annotations of human-chatbot conversations, and (2) offensiveness annotations of social media posts, both annotated by diverse rater pools across different socio-demographic axes. Our framework (based on disagreement metrics) reveals specific rater groups that have significantly different perspectives than others on certain tasks, and helps identify demographic axes that are crucial to consider in specific task contexts.

1 Introduction

The automatic detection of unsafe, offensive or toxic text has long been an active area of research in Natural Language Processing (NLP). Originally aimed at online content moderation (Wulczyn et al., 2017; Founta et al., 2018), recently, triggered by academic and governmental calls for action (Commission, 2020; LLC, 2023), these efforts are also addressing the urgent need to equip generative language technologies with safety guardrails that prevent inadvertent generation of offensive or harmful content (Bai et al., 2022; Glaese et al., 2022).

Much of this work relies on human annotation for evaluating and training offensiveness or safety classifiers, or fine-tuning generative models. Current approaches largely overlook cultural and individual factors that shape raters' perspectives on what is safe or offensive (Aroyo and Welty, 2015; Waseem, 2016; Salminen et al., 2019; Uma et al., 2021). Systematic rater disagreements are instead circumvented by enforcing a single ground truth or using majority vote, which inadvertently marginalizes minority perspectives and further amplifies societal biases in data (Prabhakaran et al., 2021).

Recent work points to the need for greater diversity in rater pools (Thoppilan et al., 2022) and proposes ways to incorporate disagreements in the learning pipeline (Davani et al., 2022). However, incorporating rater diversity at scale is still a challenge, as there are numerous diversity axes to consider, and it is unclear which ones are relevant for particular tasks. For instance, in sentiment analysis, Prabhakaran et al. (2021) found that, while there were systematic disagreements between raters from different racial groups, there were no significant differences across gender groups. In contrast, Homan et al. (2023) found that safety annotations did not differ significantly across race/ethnicity or gender groups individually, but they do differ across intersectional race/ethnicity-gender groups. Hence, the lack of effective metrics that can capture such inter-group and intra-group cohesion at scale to determine group-level associations, is a critical issue.

In this paper, we propose a framework to measure the magnitude and strength of such systematic

diversity of perspectives among rater subgroups. Our framework combines *a suite of metrics* that measure group associations in human annotations with *a permutation tests based significance testing approach* that assesses the reliability of these associations without any independence assumptions. We apply this framework to two datasets: DICES-350 (Aroyo et al., 2023b) - 350 chatbot conversations annotated for safety by 104 raters from a diverse pool; and D3 (Davani et al., 2023b) - social media comments annotated for offensiveness by 4000 raters from 8 cultural regions, balanced across gender and age. Our framework reveals systematic disagreements along demographic lines about the safety of the conversations, and demonstrates that it picks up task-dependent group associations in an efficient and effective manner, furthering the objective of identifying meaningful diversity in perspectives in human annotations.

2 Related work

Prior work on detecting harmful language, such as toxicity (Pavlopoulos et al., 2020; Xenos et al., 2022), offensiveness (Davidson et al., 2017), and hate speech (Warner and Hirschberg, 2012; Waseem and Hovy, 2016), has led to curating datasets and developing models for social media content moderation (Wulczyn et al., 2017; Founta et al., 2018; Vidgen et al., 2019). Recent advancements in conversational AI also increased attention to ensure safety and mitigate potential harms (e.g., Solaiman and Dennison, 2021; Xu et al., 2021; Shelby et al., 2022; Si et al., 2022; Bian et al., 2023; Huang et al., 2023; Santurkar et al., 2023). The latest generation of AI-driven language technologies (OpenAI, 2022; Google, 2022, 2023; Taori et al., 2023) is based on large language models (OpenAI, 2023; Touvron et al., 2023) using reinforcement learning from human feedback (RLHF) (Christiano et al., 2023; Ouyang et al., 2022). Studies show that on human alignment tasks (of which safety is one example), rater disagreement can be as high as 40% (Ziegler et al., 2020). However, not much work has gone into developing scalable methods to deal with these high levels of rater disagreement.

Rater disagreement has a long history in NLP research as a challenge for crowd-sourced annotations and as a potential indication of human biases (Arhin et al., 2021; Mathew et al., 2021; Sahoo et al., 2022; Wich et al., 2020). Though traditionally viewed as a mark of poor quality data, disagreement is increasingly seen as an important qualitative signal in its own right, one that is present in most tasks that requires human judgement (Aroyo and Welty, 2013; Hovy et al., 2013; Plank et al., 2014; Klenner et al., 2020; Basile et al., 2021).

Empirical analyses of inter-rater disagreements put forth raters' backgrounds and experiences as the foundation of their annotations in such tasks, leading to systematic disagreements (e.g., Prabhakaran et al., 2021; Denton et al., 2021; Sap et al., 2022; Homan et al., 2023; Pei and Jurgens, 2023). For instance, raters' demographics, including first language, age, and education, can significantly impact the performance of hate speech and abusive language detectors trained on that rater's behavior (Al Kuwatly et al., 2020), and raters' stereotypes about different social groups and attitudes toward racism impact their annotations of hate speech targeting those groups and racist language (Sap et al., 2022; Davani et al., 2023a).

Therefore, a large body of work has emerged to quantify, model, and measure rater disagreement (e.g., 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). In early work, Hovy et al. (2013) introduce MACE, an unsupervised item-response model to capture raters' relative trustworthiness to more accurately aggregate annotations into a final label. Weerasooriya et al. (2020) propose predictive models for rater disagreement that take into account sampling error, a common problem in datasets with very few annotations per item. Using multi-task modeling frameworks, Fornaciari et al. (2021) add an auxiliary task to predict the soft label distribution over rater labels, Davani et al. (2022) model individual raters using a shared network to preserve their systematic disagreements until prediction, and Orlikowski et al. (2023) expand the approach by incorporating a group-specific layer to assess the benefits of socio-demographic attributes in modeling annotations.

Novel modeling efforts have further incorporated raters' demographics and other background attributes to improve the predictions (Hovy, 2015; Garten et al., 2019; Hovy and Yang, 2021), with Hung et al. (2023) demonstrating the performance improvement when predicting raters' age and gender is coupled with language modeling objectives. Our work provides a framework that anchors on intra-group and inter-group cohesion to qualify the strength of disagreements within and across groups,

and provide statistical tests to assess the reliability of these observed group-level patterns.

3 Group Associations in Annotations

Recent studies established the need to account for systematic rater disagreement in subjective tasks (Klenner et al., 2020; Basile, 2020; Prabhakaran et al., 2021; Aroyo et al., 2023a) by demonstrating socio-demographic differences in rater perceptions. However, systematic approaches to reliably assess *whether* and *how much* diversity axes impact disagreement for different tasks are still missing. To address this gap, we introduce a comprehensive analysis framework to measure statistically significant group associations within human annotations.

3.1 Terminology

Let us represent a human-annotated dataset as a collection of *items* $\mathbf{X}$ with a corresponding collection of *annotations* $\mathbf{Y}$, obtained from a collection of *raters* $\mathbf{Z}$. Each row $\mathbf{X}_i$ is an item that is annotated, and each corresponding $\mathbf{Y}_i$ captures the annotations for $\mathbf{X}_i$. The columns in $\mathbf{Y}_i$ correspond to individual raters' annotations. In other words, $\mathbf{Y}_{ij}$ represent annotations by rater $j \in \mathbf{Z}$ for item i.[1] In its simplest case, $\mathbf{Y}_{ij}$ can be a binary value, but it can be conceived as a vector capturing j's responses to different questions pertaining to i, or a one-hot encoding of j's annotation in case of categorical values. Each row $\mathbf{Z}_k$ represents a rater k and the columns of $\mathbf{Z}_k$ contain *group attributes* (e.g., demographic characteristics such as gender, race/ethnicity, and/or age associated with k). Let Π denote a set of demographic properties, e.g., $\Pi = \{\text{gender = MALE, age = GenZ}\}$. Then, let $\mathbf{Z}[\Pi] \subseteq \mathbf{Z}$ denote the subpopulation of raters satisfying that property, and let $\mathbf{Y}_{\mathbf{Z}[\Pi]}$ denote the submatrix of $\mathbf{Y}$ that captures the annotations of that subpopulation of raters according to Π.

3.2 Disagreement Analysis Framework

We aim to determine whether certain rater groups, defined in terms of their demographic attributes, systematically (and in statistically significant ways) differ from others in terms of their annotations for a given task. For this, we need to measure the (dis)agreement between raters within the group, as well as with those from outside the group.

[1]Note that $\mathbf{Y}_{ij}$ may be a sparse matrix if each item is labeled by only a handful of raters (which is often the case).

In-group Cohesion ($C_I(Y)$) captures how much cohesion a particular rater group has among themselves. Formally, an *in-group cohesion* metric is a mapping $C_I : 2^{\mathbf{Y}} \rightarrow \mathbb{R}$ where, for any subgroup of annotations $Y \subseteq \mathbf{Y}$, higher values of $C_I(Y)$ indicate higher levels of agreement among Y. We are interested in $C_I(\mathbf{Y}_{\mathbf{Z}[\Pi]})$, the in-group cohesion among raters who satisfy the set of demographic properties Π.

Cross-group Cohesion ($C_X(Y, Y')$) captures how much one rater group agrees with another rater group. Formally, a *cross-group cohesion* metric is a mapping $C_X : 2^{\mathbf{Y}} \times 2^{\mathbf{Y}} \rightarrow \mathbb{R}$ where, for any pair of subgroups of annotations $Y, Y' \subseteq \mathbf{Y}$, higher values of $C_X(Y, Y')$ indicate higher levels of agreement between the annotations in Y and Y'. While cross-group cohesion could be calculated for any two given subsets of annotations, we are primarily interested in $C_X(\mathbf{Y}_{\mathbf{Z}[\Pi]}, \mathbf{Y}_{\mathbf{Z}[\neg\Pi]})$, the cross-group cohesion between raters satisfying demographic properties Π and those who do not.

Group Association Index (GAI): Both *in-group* and *cross-group cohesion* are useful for assessing the strength of annotation patterns found in a demographic grouping Π. For instance, high in-group cohesion within $\mathbf{Z}[\Pi]$ and cross-group cohesion between $\mathbf{Z}[\Pi]$ and $\mathbf{Z}[\neg\Pi]$ might just mean that the task has high agreement across the board. On the other hand, $\mathbf{Z}[\Pi]$ having both low in-group and cross-group cohesion might suggest that the raters in general have a hard time agreeing with one another, regardless of the specific grouping Π. Inspired by graph-theoretic metrics for community detection in networks, such as *modularity* (Newman, 2006), we introduce a group association index that combines these two aspects into a single score:

$$GAI(\Pi) = \frac{C_I(\mathbf{Y}_{\mathbf{Z}[\Pi]})}{C_X(\mathbf{Y}_{\mathbf{Z}[\Pi]}, \mathbf{Y}_{\mathbf{Z}[\neg\Pi]})}$$

The baseline value of GAI is 1; i.e., when C_I and C_X are more or less the same, regardless of their magnitudes, the task annotation patterns have minimal or no group association with Π. When C_I is larger than C_X, the GAI values will be higher than 1, suggesting higher group association with Π for the task. On the other hand, for GAI values less than 1, raters agree more with raters outside the group than within the group, suggesting that there are potential patterns of systematic disagreement that are not captured by Π.

Diversity Sensitivity Index (DSI): GAI indicates which groups significantly differ from others. There are numerous demographic axes (e.g., gender, age, race/ethnicity, sexual orientation, etc.) along which a rater pool can be diversified. When recruiting raters, which (if any) of these should be prioritized? It helps to know whether and by how much the subgroups within any axis have a significant GAI. This is more insightful than the average GAI value. Hence, we define *diversity sensitivity index* of a task w.r.t. a demographic axis with K groups as the $\max$ of $GAI(\Pi_k)$ for $k \in [1, K]$.

3.3 Significance Testing

To ensure our diversity measurements are reliable, it is important to test their significance. Commonly used tests assume the data items are independently sampled, which doesn't hold in our case, since each annotation depends on all items with the same rater and all raters who annotated that item. So we use *permutation tests* to control for these dependencies.

Null hypothesis: For any in-group cohesion (or cross-group divergence) metric C_I (or C_X), our null hypothesis $\mathbf{H_0}$ is

> $\mathbf{H_0}$: Value of C_I (or C_X) for any (pair of) subpopulation(s) $\mathbf{Y}_{\mathbf{Z}[\Pi_1]}$ (, $\mathbf{Y}_{\mathbf{Z}[\Pi_2]}$) is independent of demographic profile(s) of member(s) of Π_1 (and Π_2).

To test H_0, we randomly shuffle the raters demographic profiles, measure the test statistic after each shuffle, and then count how many times the shuffled statistic exceeds the observed value. If the observed value is significant, then ***only a small percentage of the measurements from random groups should exceed the observed value***. Formally, p-value of C_I is defined as

$$p_{C_I}(\mathbf{Y}_{\mathbf{Z}[\Pi_1]}) =_{\text{def}} \begin{cases} \|\{s_i^* : s_i^* < C(\mathbf{Y}_{\mathbf{Z}[\Pi_1]})\}\|/N \\ \quad \text{if } C(\mathbf{Y}_{\mathbf{Z}[\Pi_1]}) < s^*_{\lfloor N/2 \rfloor}, \\ \|\{s_i^* : s_i^* > C(\mathbf{Y}_{\mathbf{Z}[\Pi_1]})\}\|/N \\ \quad \text{otherwise.} \end{cases}$$

where N is a large number and $s_1^*, \ldots, s_N^*$ are computed by the following pseudocode:

$i \leftarrow 0$
while $i < N$ **do**
  $\mathbf{Z}^* \leftarrow$ randomly permute the rows of $\mathbf{Z}$ (but fixing the indices, so that the rows map to the same annotations even though their demographics have changed)
  $i \leftarrow i + 1$
  $s_i^* \leftarrow C(\mathbf{Y}_{\mathbf{Z}^*[\Pi_1]})$
end while
reorder $s_1^*, \ldots, s_N^*$ in ascending order.

The p-value $p_{C_X}(\mathbf{Y}_{\mathbf{Z}[\Pi_1]}, \mathbf{Y}_{\mathbf{Z}[\Pi_2]})$ of C_X is defined as above, except that we replace $C_I(\mathbf{Y}_{\mathbf{Z}[\Pi_1]})$ with $C_X(\mathbf{Y}_{\mathbf{Z}[\Pi_1]}, \mathbf{Y}_{\mathbf{Z}[\Pi_2]})$ (and $C_I(\mathbf{Y}_{\mathbf{Z}^*[\Pi_1]})$ with $C_X(\mathbf{Y}_{\mathbf{Z}^*[\Pi_1]}, \mathbf{Y}_{\mathbf{Z}^*[\Pi_2]})$).

Multiple test correction: If numerous tests are conducted and the null hypothesis is true, then by the Law of Large Numbers some of them are likely to have small p-values, making them falsely appear to be significant (type I error). There is no widely accepted best practice for dealing with this problem. Some researchers advocate never using p-values for exploratory research (Hak, 2014; Trafimow and Marks, 2015) or to apply corrections such as Bonferonni (Bonferroni, 1936; Holm, 1979) against the family-wise error rate. Other researchers see those approaches as too restrictive, which can lead to important discoveries being missed (Gaus et al., 2015; Goeman and Solari, 2011; Rubin, 2017). We adopt a mixed approach and report two levels of significance: significance with no correction whatsoever and with Benjamini-Hochberg *false discovery rate (FDR)* correction (Benjamini and Hochberg, 1995).

3.4 Metrics

The concepts introduced in §3.2 are *metric-agnostic*, and the choice of metric must be justified for each experiment. Here, we describe the three kinds of metrics we use in this paper for both C_I and C_X; we compare and contrast what these metrics are sensitive to and what they reveal.

3.4.1 In-group Cohesion Metrics

IRR: We use IRR (Inter-rater reliability, particularly, Krippendorf's alpha (Krippendorff, 2004)) to measure within-group agreement while controlling for class imbalance. Krippendorf's alpha has an advantage over other IRR metrics: it can handle an arbitrary number of raters, answer options and items at one time, and it unifies and generalizes a number of other IRR metrics, including Scott's pi and Fleiss' kappa (Krippendorff, 2004). It is formulated as $1 - \frac{o_d}{e_d}$, where o_d is the mean observed disagreement between pairs of distinct raters, and e_d is the class-imbalance-controlling term. The o_d term is, effectively, hamming distance and e_d is

the expected amount of disagreement, under the assumption that each rater's responses are randomly distributed among the conversations they label (but each rater's marginal distribution of annotations is fixed), independent of the other raters' responses.

Plurality size: IRR and our many other metrics are based on counting the (dis)agreements between pairs of raters. But in practice, raters are often seen as populations whose annotations are taken as *votes*, where the most popular annotation (i.e., majority vote) becomes the gold standard response. Thus, a very natural measurement of agreement is the fraction of raters who belong to the most popular choice (similar to (Prabhakaran et al., 2021)'s approach). This metric is less sensitive to class imbalance than metrics that count pairwise disagreements. It is computed by iterating over each item, taking the argmax over the distribution of responses, and then taking its mean over all pairs.

Negentropy: IRR measures pairwise agreement between raters and plurality size captures the impact of disagreement in the rating aggregation process. Another common way to measure disagreement in groups, used in polls and surveys, is to estimate the distribution of annotations associated with each item. Entropy is a common metric for measuring the randomness of a probability distribution, such as the annotations from multiple raters to a safety question about a conversation. It captures how evenly distributed the ranges of responses are. To orient all our metrics so that larger numbers mean more agreement, we report *negentropy* (Brillouin, 1953): for each conversation, we compute the entropy over the distribution of responses. Then we subtract this from the maximum value entropy can take over the response domain. For a domain with n possible responses, this number is $\ln n$. Finally, we take the mean over all conversations.

3.4.2 Cross-group Divergence Metrics

Analogous to our in-group cohesion metrics, we focus on three cross-group cohesion metrics.

XRR: *Cross-replication reliability* (Wong et al., 2021) is similar to Krippendorf's alpha, except that the pairs of raters being compared come from separate groups. Like alpha, XRR can handle arbitrary numbers of raters, answer options and items. And it also controls for class imbalance.

Voting agreement: For across-group agreement, it is equally natural, by analogy to plurality size, to compare two groups as if they were voting blocks. For each item, we compute the plurality choice for each group. To account for class imbalance, we compute Krippendorf's alpha over all conversations between the two groups, based on each group's plurality choices. Although straightforward, we are not aware of this method proposed as a group-level divergence metric.

Cross-negentropy: Cross-entropy is algorithmically similar to entropy but is computed over two distributions, not one. We define cross-negentropy in an analogous manner to negentropy.

Dataset	Items	Rater pool	Raters per item	Total annotations
DICES-350	350	104	104	582,400
D3	4554	4309	24	150,702

Table 1: DICES-350 and D3 dataset annotation stats.

4 Experiments

4.1 Data

We apply our metrics to the two datasets: **DICES-350** (Aroyo et al., 2023b),[2] and **D3** (Davani et al., 2023b). The DICES-350 dataset is a curated sample of 8K multi-turn conversation corpus generated by human agents interacting with a generative AI-chatbot (Thoppilan et al., 2022) in an adversarial setting. These conversations were then annotated for safety by a diverse rater pool. The D3 dataset contains a curated sample of social media posts from Jigsaw datasets (Jigsaw, 2019, 2018), annotated for offensiveness in text. We choose the DICES-350 and D3 datasets as they both contain fully replicated annotations from a diverse rater pool along with their demographic details, enabling our in-depth and fine-grained group-level analyses.

DICES-350 contains annotations for safety along 16 dimensions for all 350 conversation by 123 unique raters based in the US. The authors of DICES-350 aimed for an approximately equal numbers of raters in each of the 12 demographic groups (3 x 4 design) created by fully crossing age groups (GenZ, Millennial, GenX+) with race/ethnicity (Asian; Black; Latine/x; White). All raters annotated all 350 conversations. We limit our study to 104 raters after removing 19 raters who were deemed unreliable by the authors of DICES-350.

[2] https://github.com/google-research-datasets/dices-dataset/tree/main/350

DICES-350					
Race	Gender		Age		
	F	M	GenZ	Mill.	GenX+
As.	9	12	4	12	5
Bl.	16	7	13	5	5
Lat.	12	10	6	7	9
Multi.	4	9	6	2	5
Wh.	16	9	5	2	18

Table 2: DICES-350 raters in various demographic intersectional groups. Race/ethnicity information is abbreviated for space: Bl: Black; Wh: White; As: Asian; Lat: Latine; Multi: Multi-racial.

D3						
Region	Gender			Age		
	F	M	O	18-30	30-50	50+
AC.	205	306	5	269	168	79
ICS.	245	308	1	237	198	119
LA.	275	271	3	302	176	71
NA.	325	220	6	263	175	113
Oc.	307	203	7	161	221	135
Si.	249	280	11	208	228	104
SSA.	219	309	2	320	157	53
WE.	294	252	6	259	172	121

Table 3: D3 dataset raters in various intersectional groups. Region names abbreviated for space: AC: Arab Culture; ICS: Indian Cultural Sphere; LA: Latin America; NA: North America, Oc: Oceania, Si: Sinosphere; SSA: Sub-Saharan Africa, WE: Western Europe.

See Table 2 for breakdowns of the demographic groupings along race, gender, and age.

The safety annotation dimensions cover a variety of safety violations, including harmful content, unfair bias, misinformation, and political endorsements, and raters may respond *Safe*, *Unsafe*, or *Unsure*. We compute a single safety response for each rater-conversation pair by aggregating the responses into a single, overall safety response. For any conversation, if *any* of the safety annotations is *Unsafe*, then we label the entire conversation as unsafe. Otherwise, if any of the safety annotations is *Unsure*, then so is the aggregated response. Otherwise, the aggregated response is *Safe*. In other words, it only takes one reason for a conversation to be unsafe and, conversely, if a conversation is unsafe, it need only be unsafe for one reason.

D3 is similarly annotated by a diverse pool of 4k raters across 8 geo-cultural regions and 21 countries. Each item in the dataset was annotated by at least three raters in each region (∼24 annotations per item). The annotation effort aimed for capturing an approximately equal number of raters (∼450) from each region and equal ratio of representation for various demographic group across age (18 to 30, 30 to 50, and more than 50 years old) and genders (Man, Woman, and Other). See Table 3 for the breakdown of the demographic groups across different regions, gender, and age groups.

Raters were asked to label the textual items' level of offensiveness on a 5-point Likert scale, 1 being *not offensive at all* and 5 being *extremely offensive*, with the option of choosing *Unsure*. We treated a score of 3 or higher as being *Offensive*, in line with the dataset creators (Davani et al., 2023b).

4.2 Results

We report results of our analysis using IRR and XRR as the in-group and cross-group cohesion metrics in for both DICES-350 and D3 datasets in Table 4. We focus on IRR and XRR based analysis in this section, but the full results using all other metrics are presented in Tables 5, 6, and 7.

We investigate groupings along age, gender, and either race/ethnicity (DICES-350) or region (D3). For DICES-350, we also explore intersectional groups along race/ethnicity and gender (some of the intersections of age and race/ethnicity are too small to reasonably assess significance), while we explored the intersection of region with both age and gender groups in the D3 dataset. Results for all intersections and statistically significant intersections are reported in Tables 5-7 and 4, respectively.

DICES-350 results: Only race/ethnicity groupings show significant results on their own, suggesting age and gender doesn't matter. However, looking at intersectional groups, Latine women have the highest in-group cohesion (0.238), followed by White men (0.218), Latine raters (0.215), and Black women (0.213). Asian women have the lowest score (0.073), followed by White women (0.114). Latine women also have the highest cross-group cohesion (0.199), followed by Latine raters (0.189). Asian women have the lowest score (0.134), followed by White women (0.152) and White raters (0.159). White men have the highest GAI score (1.262) followed by Latine women (1.196), Latine raters (1.139), and Black women (1.130). Some groups have GAIs significantly lower than baseline; Asian women have the lowest GAI (0.540), followed by White women (0.752), suggesting that these groups have constituent subgroups that have more agreement with raters outside this group.

DICES-350				
Dimension	Group	IRR	XRR	GAI
age	gen x+	↓0.166	↓0.171	↓0.975
age	gen z	↓0.166	↓0.172	↓0.966
age	millenial	↑0.189	↑0.179	↑1.052
gender	Man	↑0.187	↑0.175	↑1.071
gender	Woman	↓0.160	↑0.175	↓0.916
race	As.	↓0.145	↓0.166	↓0.872
race	Bl.	↑0.193	↑0.181	↑1.063
race	Lat.	**↑0.215***	**↑0.189***	**↑1.139***
race	Multi.	↓0.153	↓0.168	↓0.916
race	Wh.	↓0.145	**↓0.159***	↓0.908
Statistically Significant Intersections				
race, gender	As., Woman	**↓0.073***	**↓0.134***	**↓0.540***
race, gender	Bl., Woman	**↑0.213***	↑0.188	**↑1.130***
race, gender	Lat., Woman	**↑0.238***	**↑0.199****	**↑1.196***
race, gender	Wh., Man	**↑0.218***	↓0.173	**↑1.262****
race, gender	Wh., Woman	**↓0.114***	**↓0.152***	**↓0.752***
D3				
Dimension	Group	IRR	XRR	GAI
age	(18,30)	**↑0.115****	↑0.107	**↑1.068****
age	(30,50)	**↓0.089****	↓0.104	**↓0.850****
age	50+	↑0.110	↑0.111	↑0.999
gender	Woman	↑0.110	↑0.108	↑1.024
gender	Man	↓0.105	↑0.107	↓0.976
gender	Other	↑0.209	↓0.096	**↑2.172***
region	AC.	**↑0.133****	↑0.113	**↑1.174***
region	ICS.	↓0.103	**↓0.099***	↑1.043
region	LA.	**↑0.129****	↑0.112	**↑1.152***
region	NA.	**↑0.143****	↑0.110	**↑1.307****
region	Oc.	↑0.118	↓0.103	**↑1.145***
region	Si.	**↓0.087***	**↓0.087****	↓1.002
region	SSA.	**↑0.142****	↓0.104	**↑1.361****
region	WE.	**↑0.135****	↑0.111	**↑1.222****
Statistically Significant Intersections				
region, age	ICS., (18,30)	**↓0.063****	↓0.100	**↓0.634***
region, age	ICS., (30,50)	**↓0.060***	↓0.100	**↓0.601***
region, gender	ICS., Woman	**↓0.070***	↓0.106	**↓0.655***
region, age	LA., (18,30)	**↑0.143****	↑0.118	**↑1.216***
region, gender	LA., Woman	**↑0.143****	↑0.111	**↑1.290***
region, gender	NA., Woman	**↑0.153****	↑0.116	**↑1.314****
region, age	Oc., (30,50)	↑0.112	**↓0.089****	**↑1.255***
region, gender	Oc., Woman	**↑0.133***	↑0.110	**↑1.208***
region, age	Si., (30,50)	**↓0.033****	**↓0.082****	**↓0.405****
region, age	Si., 50+	↑0.137	**↓0.061****	**↑2.225****
region, gender	Si., Woman	↓0.100	**↓0.081****	**↑1.237***
region, age	SSA., (18,30)	**↑0.146****	↓0.107	**↑1.365****
region, age	WE., (18,30)	**↑0.177****	**↑0.126****	**↑1.402****
region, gender	WE., Woman	**↑0.151****	↑0.118	**↑1.284***

Table 4: 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 $C_X = \text{XRR}$ and $C_I = \text{IRR}$.

The DSI metric looks at what is the highest GAI for each diversity axis (including intersectional axes) we consider. In the DICES-350, we observe the higher DSI for the intersectional axis of gender and race (1.262 for White men), followed by race considered alone (1.139 for Latine raters). These numbers suggest that it is crucial to prioritize recruiting raters with a diverse representation along race and gender, while diversifying along age may be less crucial based on our results for this task. Note that, although unlikely, applying our framework along other intersectional axes including age may reveal other group associations.

D3 results: Here, 18-to-30-year-old Western Europeans have the highest IRR (0.177), followed by North American women (0.153) and Western European women (0.151). Lowest scores are reported for 30-to-50-year-old raters from Sinosphere (0.033) and Indian Cultural Sphere (0.060), followed by 18-to-30-year-old (0.063), and women (0.070) groups from Indian Cultural Sphere. 18-to-30-year-old Western Europeans also have the highest XRR (0.126) followed by non-significant scores for Western European women (0.118) and North American women (0.116). Lowest XRR is reported for 50+-year-old raters of Sinosphere (0.061), followed by Sinosphere women (0.081) and 30-to-50-year-old Sinosphere raters (0.082), all significant after BH corrections. In terms of GAI scores, 50+-year-old raters of Sinosphere (2.225), and raters identifying with non-binary genders (2.172) report the highest GAI, followed by 18-to-30-year-old groups in Western European (1.402) and Sub-Saharan Africa (1.365); all significant after BH correction. Notably, unlike the DICES-350, different age and region groups have significantly high GAI scores; 18-to-30-year-old (1.068), North America (1.307), Sub Saharan Africa (1.361), and Western Europe (1.222). Interestingly, intersectional results demonstrate that while women in general did not report high GAI, subgroups of women in different regions show more in-group agreement.

We observe the highest DSI for the intersectional axis of region and age (Sinosphere, 50+) at 2.225, followed by a high DSI for gender (Other) at 2.172. This shows the importance of prioritizing raters from non-binary gender groups and specific subgroups along region and age to capture important diverse perspectives in assessing offense.

5 Discussion

Our framework provides a means to assess the cohesion and strength of group associations along different axes of diversity that matter for a given task, identifying different groups, including intersectional groups, that are relevant for specific tasks.

Task specific insights: Our analysis provides insights about specific rater groups for each task. For instance, in the conversational safety task (DICES-350), White men having the highest and Asian women the lowest in-group cohesion. Interestingly, White women and Asian men had opposite cohesion trends from their alter-genders. This suggests that men are driving the high cohesion observed in White raters, and that women and men counteract each other in the weak effects observed in Asian raters overall. High coherence among White men is due to their strong tendency to prefer *Safe* to *Unsafe* annotations by a nearly $3:1$ ratio. On the other hand, for the offense annotation task (D3), most regional groups show significant group associations. Notably, Indian cultural sphere and Sinosphere shows no significant in-group cohesion (nor GAI), although 50+ groups within Sinosphere show high in-group cohesion. Age is a notable factor across board, both individually and within intersectional groups, suggesting the need for diversification of rater pools around age groups.

Flexibility of group granularity: Our analysis is generic enough that it can be applied groups defined by any subset of demographic characteristics, enabling it to easily reveal intersectional group associations. For instance, although age and gender groups revealed no association for safety, intersectional analysis revealed that gender plays a substantial role in driving race-level group tendencies.

Flexibility of metrics: Our framework is extensible to any (comparable) underlying in-group cohesion and cross-group divergence metrics. We observe that the values across our metrics vary (see Table 5 & 6); IRR numbers are relatively low (around 0.2) while other metrics report much higher agreements. These disparities may point to potential overcompensation for class imbalance ($2:1$ for *safe* to *unsafe*) in the IRR metric. IRR is typically used to compare small groups of raters. With larger groups of raters there are quadratically more pairs of raters, and the high dimensionality of the response vectors (350 responses per rater) means that all pairs can potentially be very different from each other: there is both more space to disagree and more disagreements to count. Negentropy and plurality size are less sensitive to these effects, since they are both based on the distributions of all raters, not on the pairwise relationships between all raters. Future work should look into which metrics may be more suitable in specific task and data settings (e.g., number of raters, replication factor, etc.).

Versatility across dataset characteristics: The two datasets we applied our framework to differ not only on the underlying tasks, but also on other dataset characteristics/structure. DICES-350 contains fully parallel annotations (i.e., all 104 annotators annotated all 350 items), whereas D3 contains batches of annotations where sets of 35 items contain fully parallel annotations from 24+ raters. These differences did not hinder the applicability of the analysis framework. In fact, the D3 analysis provides a potential pathway where such highly parallel annotations by broadly diverse rater pools could be performed in early phases, that can then inform more streamlined data collection through curated rater pools representing selected diversity axes based on this analysis, essentially saving cost while ensuring diversity in data.

Exploratory Analysis: Our approach also illustrate the usefulness of significance testing to exploratory analysis. We see the role of significance testing in exploratory research as a compass that provides perspective in light of conflicting results that lack inherent scales for interpretation. While they impose a hefty computational burden, the permutation tests control for joint dependencies in the data between raters and conversations that simpler tests do not. However, we believe the extra computational effort is well worth the trouble, especially in informing rater recruitment decisions.

6 Conclusion

We introduced an analytical framework to measure diversity in annotations among rater subgroups, to better understand the socio-cultural leanings of subjective tasks. We proposed a group association index that combines in-group and cross-group cohesion, along with statistical significance using permutation tests. Applying this framework to two datasets of safety annotations, we demonstrated how it reveals systematic disagreements across various intersectional subgroups. Our work contributes to the efforts on bringing in diverse perspectives in data in an efficient and effective manner, furthering the goal of robust socio-technical evaluations of AI models. Furthermore, our framework provides actionable insights for practitioners to help prioritize demographic axes when diversifying rater pools. Future work will investigate how the framework may enable dynamic data collection that can adapt to emergent group associations among raters across different types of content and tasks.

7 Limitations

We acknowledge that the demographic breakdown in both datasets is a simplified representation of the population at large. We assume this was done to facilitate recruitment of raters in each group and to allow for less complexity in analysing intersecting groups. However, our analysis framework was applied on two independent datasets with different rater pools, demographic breakdowns and data collection designs, which points to its generalizability. Provided more granular demographic data, we are confident the frameworks can be readily applied.

We recognize that further research is needed to extend such analysis to other intersectional groups that we have not been investigated in this paper. For example, we believe that further slicing the ethnicity, native languages and age groups is likely to reveal further insights and provide additional evidence of systematic differences between different groupings of raters. Due to page limit this paper focuses on introducing the disagreement analysis framework, and provide initial analysis to demonstrate its utility in revealing significant group associations along socio-demographic lines.

Finally, we recognize more work is needed to distinguish *good* from *bad* disagreement. We focused on revealing statistically significant cohesion within groups (and lack of it across groups), which may weed out noisy disagreements. However, more work is needed to disentangle disagreements that are important to retain in the interest of retaining diverse perspectives, vs. those that are undesirable from a practitioners' perspective (e.g., lack of training in a particular rater platform/pool).

While the use of significance tests in exploratory analysis is controversial (Balluerka et al., 2005), there is usually a degree of arbitrariness in their use, for instance, in the choice of level (e.g., $p = 0.05$, in our case), if nothing else. In the case of exploratory research such as ours, one must be careful not to abuse significance testing. For instance, we deliberately held back on a deeper exploration of intersectionality to reduce the risk of p-hacking (see discussion in § 3.3, 4.2). We also note that we have many more significant results at the $p = 0.05$ level than chance would predict. There is also arbitrariness in the metrics used. For instance, there isn't uniform agreement on how to interpret well-established metrics such as Krippendorf's alpha.

8 Statement of Ethics

According to the DICES-350 and D3 datasets authors all the demographics data is self-declared. Raters were presented a consent form before signing up for both studies to inform them about the gathering of personal demographics and that the content to be rated is adversarial (i.e., would possibly contain offensive content). All demographics questions had the option "Prefer not to answer". All data was collected in anonymized way after the data collection tasks were completed by the raters. Raters were allowed to quit the study at any time.

References

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.

Kofi Arhin, Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Moninder Singh. 2021. Ground-truth, whose truth? – examining the challenges with annotating toxic text datasets.

Lora Aroyo, Mark Diaz, Christopher Homan, Vinodkumar Prabhakaran, Alex Taylor, and Ding Wang. 2023a. The reasonable effectiveness of diverse evaluation data. *arXiv preprint arXiv:2301.09406*.

Lora Aroyo, Alex S. Taylor, Mark Díaz, Christopher Michael Homan, Alicia Parrish, Greg Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023b. Dices dataset: Diversity in conversational AI evaluation for safety.

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.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Nekane Balluerka, Juana Gómez, and Dolores Hidalgo. 2005. The controversy over null hypothesis significance testing revisited. *Methodology*, 1(2):55–70.

Valerio Basile. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. *CEUR Workshop*.

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.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Ning Bian, Peilin Liu, Xianpei Han, Hongyu Lin, Yaojie Lu, Ben He, and Le Sun. 2023. A drop of ink may make a million think: The spread of false information in large language models. *arXiv preprint arXiv:2305.04812*.

Carlo Bonferroni. 1936. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62.

Leon Brillouin. 1953. The negentropy principle of information. *Journal of Applied Physics*, 24(9):1152–1163.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences.

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.

European Commission. 2020. The digital services act: Ensuring a safe and accountable online environment. *The Digital Services Act: Ensuring a safe and accountable online environment*.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023a. 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.

Aida Mostafazadeh Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2023b. Disentangling perceptions of offensiveness: Cultural and moral correlates.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

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*.

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.

Wilhelm Gaus, B Mayer, and R Muche. 2015. Interpretation of statistical significance-exploratory versus confirmative testing in clinical trials, epidemiological studies, meta-analyses and toxicological screening (using Ginkgo biloba as an example). *Clinical & Experimental Pharmacology*, 5(4):182–187.

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.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando,

Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements.

Jelle J. Goeman and Aldo Solari. 2011. Multiple testing for exploratory research. *Statistical Science*, 26(4):584 – 597.

Google. 2022. Pathways language model (PaLM): Scaling to 540 billion parameters for breakthrough performance.

Google. 2023. PaLM 2 technical report.

Tony Hak. 2014. After statistics reform: Should we still teach significance testing? *ERIM Report Series Reference No. ERS-2014-001-ORG*.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Christopher 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.

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.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 294–297, New York, NY, USA. Association for Computing Machinery.

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*.

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*.

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

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

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

MultiMedia LLC. 2023. FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Mark EJ Newman. 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582.

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

OpenAI. 2022. Introducing ChatGPT.

OpenAI. 2023. GPT-4 technical report.

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)*, Toronto, Canada. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? *arXiv preprint arXiv:2006.00998*.

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, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *ACL*.

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.

Mark Rubin. 2017. Do p values lose their meaning in exploratory analyses? it depends how you define the familywise error rate. *Review of General Psychology*, 21(3):269–275.

Nihar Sahoo, Himanshu Gupta, and Pushpak Bhattacharyya. 2022. Detecting unintended social bias in toxic language datasets.

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.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*.

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.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, Ajung Moon, Negar Rostamzadeh, Paul Nicholas, YILLA-AKBARI N'MAH, Jess Gallegos, Andrew Smart, and GURLEEN VIRK. 2022. Identifying sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. *arXiv preprint arXiv:2210.05791*.

Wai Man Si, Michael Backes, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Savvas Zannettou, and Yang Zhang. 2022. Why so toxic? measuring and triggering toxic behavior in open-domain chatbots. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, page 2659–2673, New York, NY, USA. Association for Computing Machinery.

Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. LaMDA: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models.

David Trafimow and Michael Marks. 2015. Editorial. *Basic and Applied Social Psychology*, 37(1):1–2.

Alexandra N 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.

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.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

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 and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

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

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.

Alexandros Xenos, John Pavlopoulos, Ion Androutsopoulos, Lucas Dixon, Jeffrey Sorensen, and Léo Laugier. 2022. Toxicity detection sensitive to conversational context. *First Monday*.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Recipes for safety in open-domain chatbots.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences.

A Appendix

Figures 1–8 report, for each metric and demographic group, the score of the metric as a horizontal black line and, subimposed beneath each horizontal line, a histogram of the metric scores under the permutation sampling determined by our null hypothesis. Result are significant when the horizontal is at the extreme end of the histograms. Histograms are also color-coded by the significance of the results they support: red histograms indicate that the result is significant at the $p = 0.05$ level, but only before adjusting for the false positive rate (FPR); green indicates significance at the $p = 0.05$ level, even after FPR adjustment. Given the exploratory nature of the work, both kinds of significance are meaningful and merit attention. But we can feel more confident that the FPR adjusted results are likely more robust and repeatable.

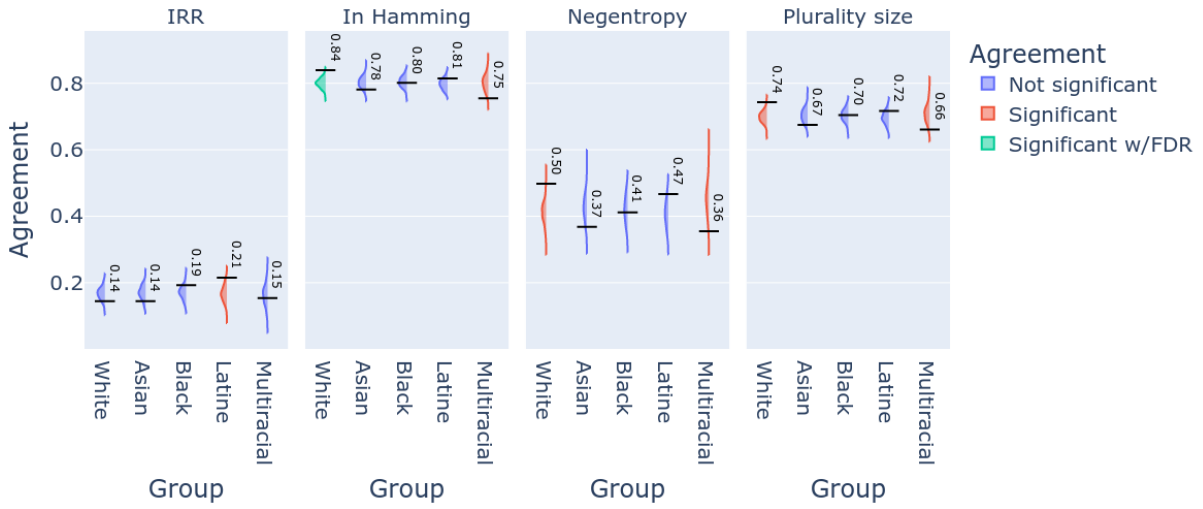

Figure 1: Within-group agreement metrics, by race/ethnicity. Negentropy and plurality size indicate that White raters have significantly more, and Multiracial significantly less, agreement than other races/ethnicities. IRR indicates that Latine raters have significantly more agreement than other races/ethnicities

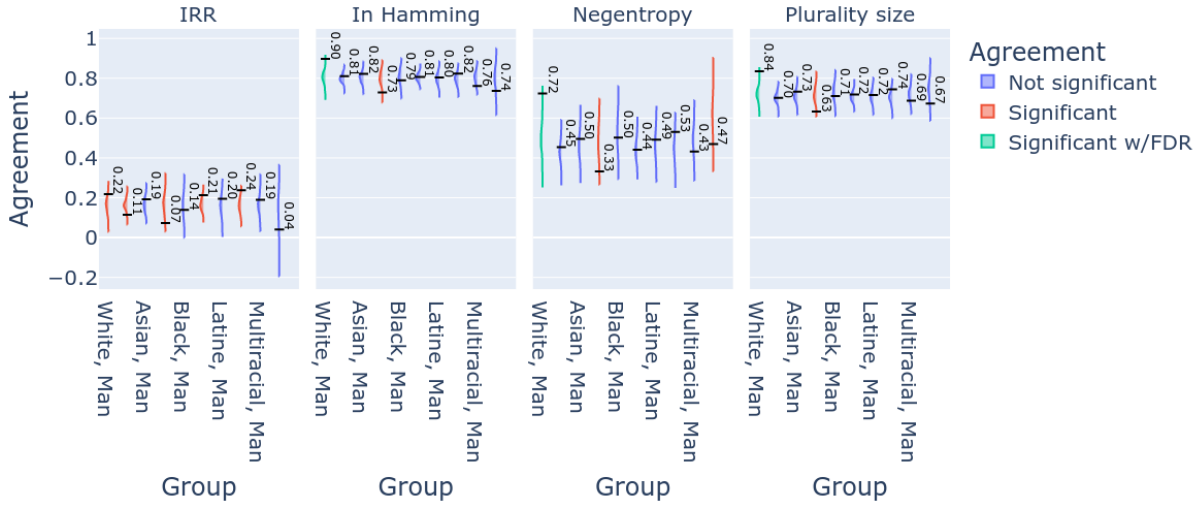

Figure 2: Within-group agreement metrics, by race/ethnicity and gender. Histograms represent the distribution of agreement values under the null hypothesis. Black horizontal bars represent the observed values. These results show that white men have significantly less agreement than other groups, according to negentropy and plurality size, neither of which control for class imbalance. IRR shows that with controlling for class imbalance between *safe* and *unsafe* annotations, the amount of agreement is more moderate. Asian women show nearly the opposite results, with less agreement than other groups unless class imbalance is controlled.

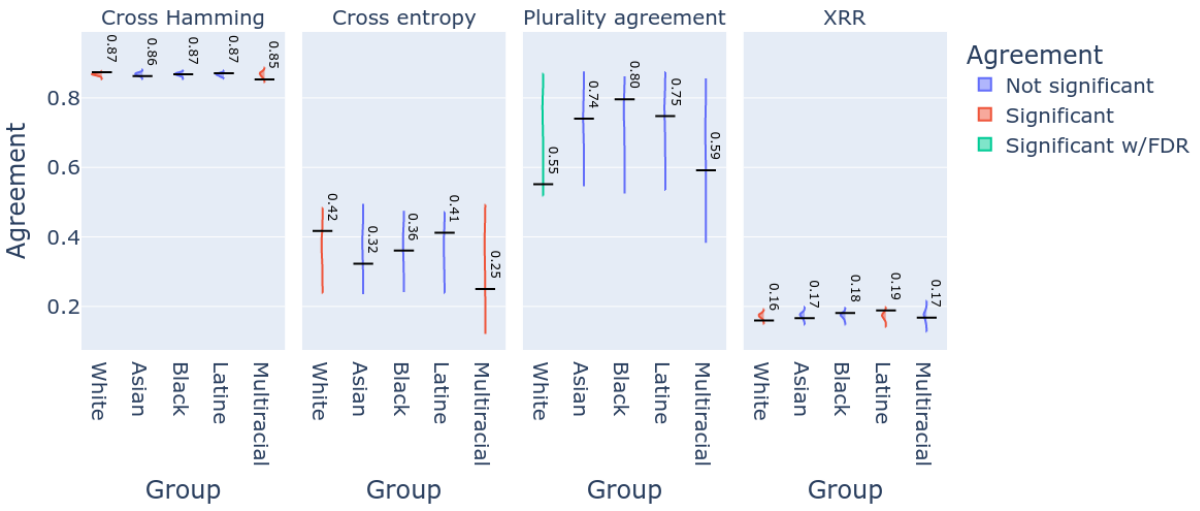

Figure 3: Across-group agreement metrics, by race/ethnicity. Histograms represent the distribution of agreement values under the null hypothesis. Black horizontal bars represent the observed values. White and multiracial voters show less overall agreement with others. Latine voters show more agreement with others.

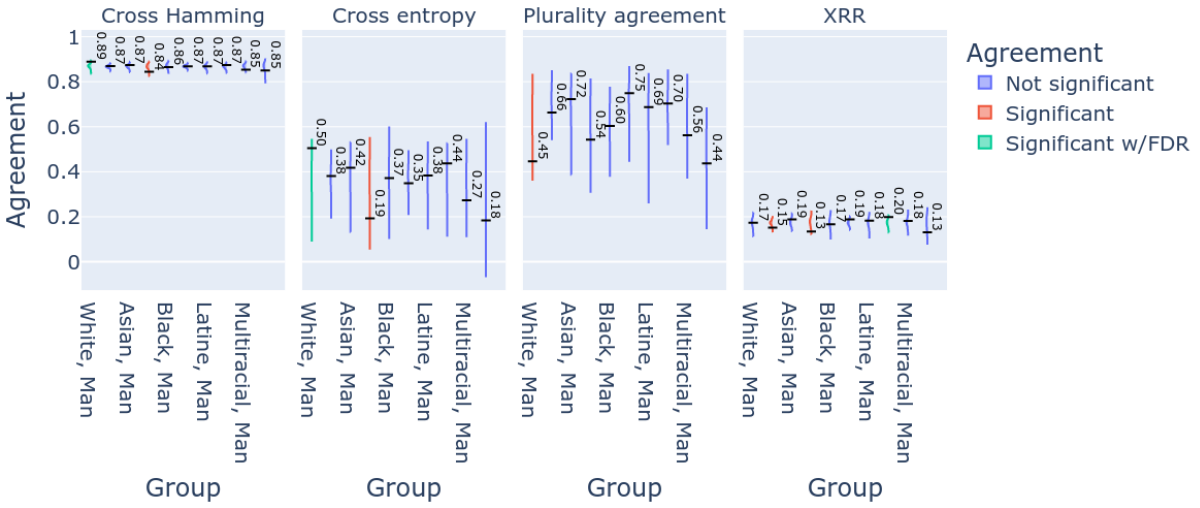

Figure 4: Across-group agreement metrics, by race/ethnicity and gender. Histograms represent the distribution of agreement values under the null hypothesis. Black horizontal bars represent the observed values. Here, white men show signs of significantly low plurality agreement. With other groups. Yet safety agreement is significantly high (though will a small effect size). This seeming disparity is due to the high class imbalance within safety reasons and white men's tendency to favor *safe annotations*. And so for specific safety reasons they appear more agreeable. However, when these reasons are aggregated into an overall safety score, differences between which men and other groups reveal themselves.

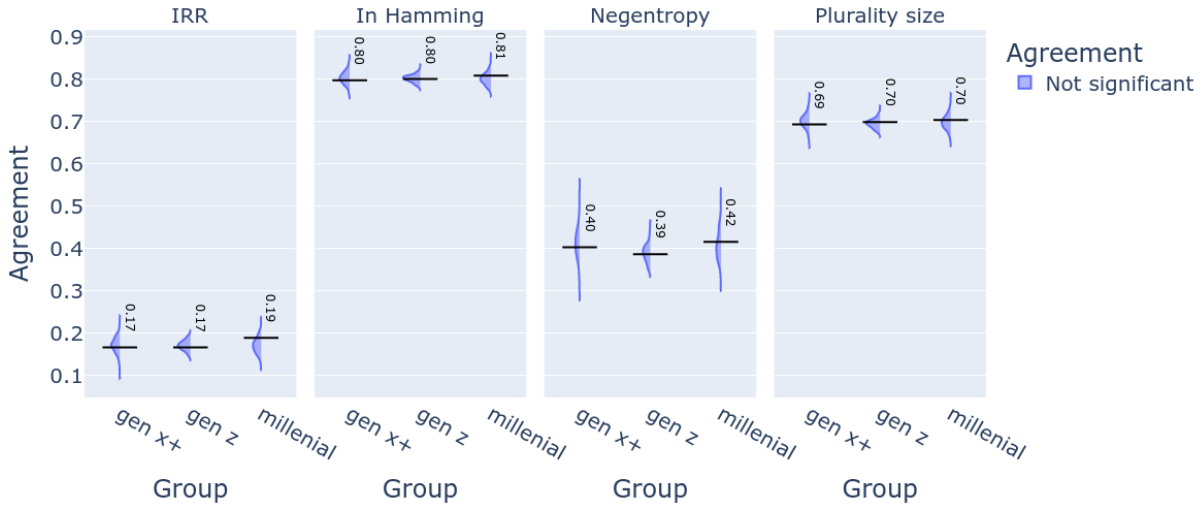

Figure 5: Within-group agreement metrics, by age. Histograms represent the distribution of agreement values under the null hypothesis. Black horizontal bars represent the observed values. None of these groups show significant amounts of difference in disagreement.

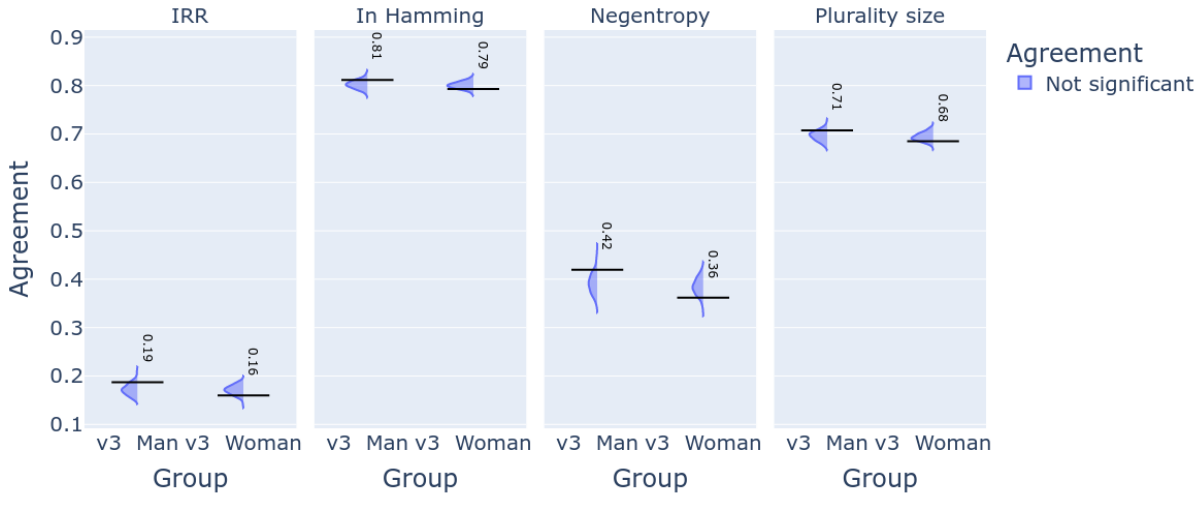

Figure 6: Within-group agreement metrics, by gender. Histograms represent the distribution of agreement values under the null hypothesis. Black horizontal bars represent the observed values. None of these groups show significant amounts of difference in disagreement.

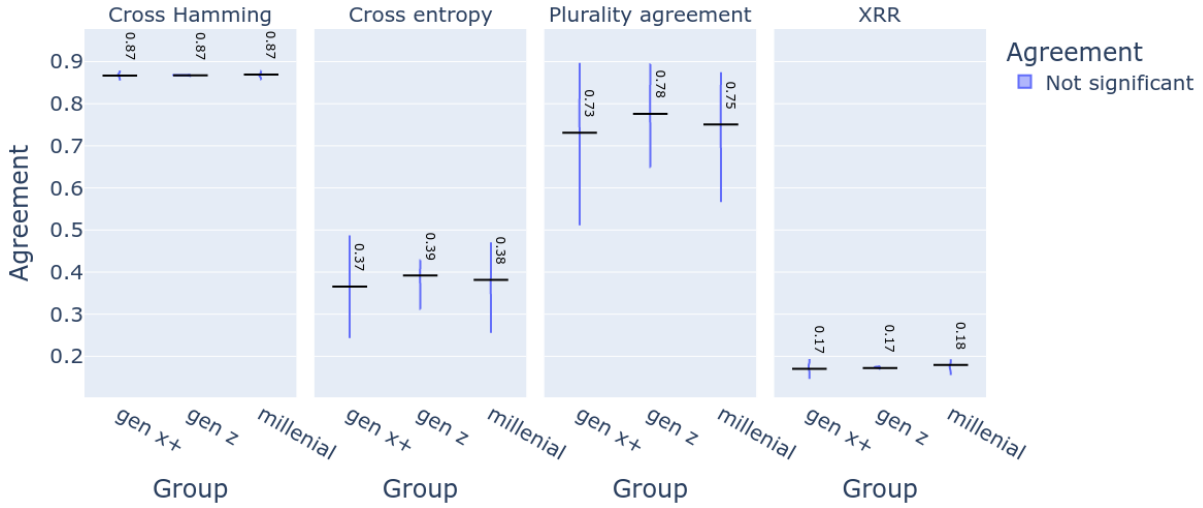

Figure 7: Across-group agreement metrics, by age. Histograms represent the distribution of agreement values under the null hypothesis. Black horizontal bars represent the observed values.

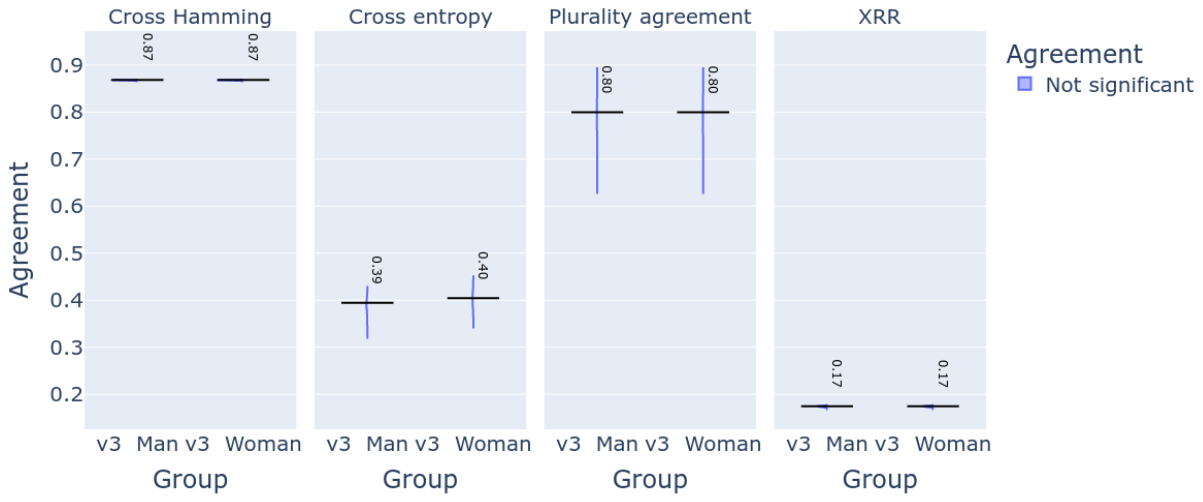

Figure 8: Across-group agreement metrics, by gender. Histograms represent the distribution of agreement values under the null hypothesis. Black horizontal bars represent the observed values.

	Dimension	Group	IRR	XRR	Negentropy	Cross Negentropy	Plurality size	Plurality agreement	GAI
0	[age]	gen x+	↓0.166	↓0.171	↓0.402	↓0.365	↓0.693	↓0.731	↓0.975
1	[age]	gen z	↓0.166	↓0.172	↓0.386	↑0.392	↑0.698	↓0.776	↓0.966
2	[age]	millenial	↑0.189	↑0.179	↑0.415	↑0.381	↑0.703	↓0.751	↑1.052
3	[gender]	Man	↑0.187	↑0.175	↑0.419	↑0.394	↑0.707	↑0.800	↑1.071
4	[gender]	Woman	↓0.160	↑0.175	↓0.362	↑0.404	↓0.685	↑0.800	↓0.916
5	[race]	Asian	↓0.145	↓0.166	↓0.368	↓0.323	↓0.675	↓0.740	↓0.872
6	[race]	Black	↑0.193	↑0.181	↓0.411	↓0.361	↑0.705	↑0.796	↑1.063
7	[race]	Latine	↑**0.215***	↑**0.189***	↑0.467	↑0.412	↑0.716	↓0.747	↑**1.139***
8	[race]	Multiracial	↓0.153	↓0.168	↓**0.355***	↓**0.250***	↓**0.661***	↓0.592	↓0.916
9	[race]	White	↓0.145	↓**0.159***	↑**0.498***	↑**0.417***	↑**0.744***	↓**0.552****	↓0.908
10	[race, gender]	Asian, Man	↑0.193	↑0.188	↑0.495	↑0.417	↑0.733	↑0.722	↑1.024
11	[race, gender]	Asian, Woman	↓**0.073***	↓**0.134***	↓**0.332***	↓**0.193***	↓**0.633***	↓0.543	↓**0.540***
12	[race, gender]	Black, Man	↓0.139	↓0.167	↓0.502	↓0.371	↓0.710	↓0.604	↓0.831
13	[race, gender]	Black, Woman	↑**0.213***	↑0.188	↑0.441	↓0.349	↑0.718	↑0.749	↑**1.130***
14	[race, gender]	Latine, Man	↑0.195	↑0.183	↑0.491	↑0.383	↑0.716	↑0.687	↑1.069
15	[race, gender]	Latine, Woman	↑**0.238***	↑**0.199****	↑0.530	↑0.437	↑0.745	↑0.704	↑**1.196***
16	[race, gender]	Multiracial, Man	↑0.190	↑0.182	↓0.432	↓0.273	↓0.688	↓0.562	↑1.043
17	[race, gender]	Multiracial, Woman	↓0.041	↓0.131	↓**0.470***	↓0.184	↓0.674	↓0.438	↓0.312
18	[race, gender]	White, Man	↑**0.218***	↓0.173	↑**0.724****	↑**0.505****	↑**0.835****	↓**0.446***	↑**1.262****
19	[race, gender]	White, Woman	↓**0.114***	↓**0.152***	↑0.454	↑0.381	↓0.702	↓0.663	↓**0.752***

Table 5: Results for in-group and cross-group cohesion, and GAI for demographic and intersectional groups within **DICES-350**. Significant results are in **bold**. A single asterisk (*) means the result is significant at the $p = 0.05$ level. A double asterisk (**) means the results are significant after Benjamini-Hochberg correction. A $\downarrow$ means that the result is less than expected under the null hypothesis. A $\uparrow$ means the result is greater. We report GAI based on C_X = XRR and C_I = IRR. The DSI results are based on variable that minimized each dimension, and they are as follows. Age: 1.052 (millennial), gender: 1.071 (men), race/ethnicity: 1.139 (Latine raters), (gender, race/ethnicity): 1.262 (White men).

	Dimension	Group	IRR	XRR	Negentropy	Cross Negentropy	Plurality size	Plurality agreement	GAI
0	[age]	(18,30)	↑**0.115****	↑0.107	↑**0.631****	↓0.297	↓0.405	↓**0.689****	↑**1.068****
1	[age]	(30,50)	↓**0.089****	↓0.104	↓**0.571****	↑0.340	↓**0.377***	↑**0.720****	↓**0.850****
2	[age]	50+	↑0.110	↑0.111	↓0.480	↑0.389	↑0.356	↑0.754	↑0.999
3	[gender]	Woman	↑0.110	↑0.108	↑**0.634****	↓**0.267****	↑0.424	↓**0.692****	↑1.024
4	[gender]	Man	↓0.105	↑0.107	↓**0.612****	↑**0.307****	↑0.423	↑**0.702****	↓0.976
5	[gender]	Other	↑0.209	↓0.096	↓0.030	↓0.605	↑0.192	↑0.978	↑**2.172***
6	[region]	Arab Culture	↑**0.133****	↑0.113	↑**0.452****	↓0.413	↓0.272	↓**0.759****	↑**1.174***
7	[region]	Indian Cultural Sphere	↓0.103	↓**0.099***	↑**0.457****	↓0.418	↓0.280	↓**0.760****	↑1.043
8	[region]	Latin America	↑**0.129****	↑0.112	↑**0.449****	↓**0.400***	↑0.317	↓**0.764****	↑**1.152***
9	[region]	North America	↑**0.143****	↑0.110	↑**0.443****	↓**0.393****	↑0.316	↓0.772	↑**1.307****
10	[region]	Oceania	↑0.118	↓0.103	↓**0.372****	↓0.411	↑0.303	↑**0.797****	↑**1.145***
11	[region]	Sinosphere	↓**0.087***	↓**0.087****	↓0.405	↓**0.381****	↓**0.223****	↑0.788	↓1.002
12	[region]	Sub Saharan Africa	↑**0.142****	↓0.104	↑0.418	↓**0.385****	↓**0.262***	↓0.777	↑**1.361****
13	[region]	Western Europe	↑**0.135****	↑0.111	↑**0.448****	↓**0.383****	↑**0.356****	↓0.768	↑**1.222****

Table 6: Results for in-group and cross-group cohesion, and GAI for demographic groups of **D3** raters. Significant results are in **bold**: * for significance at $p < 0.05$, ** for significance after Benjamini-Hochberg correction. A single asterisk (*) means significant at the $p = 0.05$ level. A double asterisk (**) means the results are significant after Benjamini-Hochberg correction. A $\downarrow$ (or $\uparrow$) 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.

	Dimension	Group	IRR	XRR	Negentropy	Cross Negentropy	Plurality size	Plurality agreement	GAI
0	[region, age]	AC., (18,30)	↑0.119	↑0.111	↑0.268	↑0.477	↓**0.207***	↓0.836	↑1.070
1	[region, age]	AC., (30,50)	↑0.116	↑0.112	↓0.184	↓0.481	↓0.226	↓0.886	↑1.040
2	[region, age]	AC., 50+	↑**0.190***	↑**0.179****	↓0.080	↑**0.610****	↑0.228	↑0.947	↑1.060
3	[region, gender]	AC., Man	↑0.129	↑0.109	↓0.284	↑**0.489****	↓0.227	↓0.828	↑1.185
4	[region, gender]	AC., Woman	↑0.125	↑0.117	↓0.198	↑0.488	↓0.202	↑0.875	↑1.064
5	[region, age]	ICS., (18,30)	↓**0.063****	↓0.100	↑0.246	↑0.485	↓0.223	↓0.849	↓**0.634***
6	[region, age]	ICS., (30,50)	↓**0.060***	↓0.100	↑0.215	↓0.482	↑0.236	↓0.868	↓**0.601***
7	[region, age]	ICS., 50+	↓0.063	↓0.103	↓0.121	↓0.513	↑0.246	↑0.922	↓0.614
8	[region, gender]	ICS., Man	↓0.093	↓0.098	↓0.284	↓0.455	↓0.241	↓0.831	↓0.953
9	[region, gender]	ICS., Woman	↓**0.070***	↓0.106	↓0.233	↑0.475	↓**0.197****	↑0.860	↓**0.655***
10	[region, age]	LA., (18,30)	↑**0.143****	↑0.118	↓0.278	↑0.475	↓0.248	↑0.837	↑**1.216***
11	[region, age]	LA., (30,50)	↓0.069	↓**0.092***	↑**0.227****	↑0.514	↓0.209	↓**0.864***	↓0.747
12	[region, age]	LA., 50+	↑0.158	↑0.136	↑0.096	↑0.583	↑0.235	↓0.933	↑1.157
13	[region, gender]	LA., Man	↑0.118	↓0.108	↓0.259	↑0.477	↓0.228	↓0.842	↑1.096
14	[region, gender]	LA., Woman	↑**0.143****	↑0.111	↓0.251	↑0.473	↓0.241	↑0.849	↑**1.290***
15	[region, age]	NA., (18,30)	↑**0.150****	↑**0.124****	↑0.272	↑0.472	↑0.250	↓**0.829****	↑1.215
16	[region, age]	NA., (30,50)	↑0.105	↓0.102	↓0.173	↓0.471	↑0.249	↑**0.898***	↑1.024
17	[region, age]	NA., 50+	↓0.099	↓0.098	↑0.139	↓0.519	↓0.210	↓0.911	↑1.016
18	[region, gender]	NA., Man	↑0.113	↑0.112	↓**0.188****	↓**0.454***	↑**0.278***	↑**0.885****	↑1.005
19	[region, gender]	NA., Woman	↑**0.153****	↑0.116	↑0.299	↓0.449	↓0.239	↓0.825	↑**1.314****
20	[region, age]	Oc., (18,30)	↑0.113	↑0.121	↓0.155	↑0.510	↑0.230	↑0.900	↑0.932
21	[region, age]	Oc., (30,50)	↑0.112	↓**0.089****	↓**0.173****	↓**0.455***	↓0.218	↑**0.900****	↑**1.255***
22	[region, age]	Oc., 50+	↓0.081	↑0.115	↓0.140	↓**0.481***	↑**0.286****	↑0.914	↓0.699
23	[region, gender]	Oc., Man	↓0.090	↓**0.091***	↓**0.170****	↓**0.448****	↓0.219	↑**0.899****	↑0.988
24	[region, gender]	Oc., Woman	↑**0.133***	↑0.110	↓**0.252****	↑0.464	↑0.266	↑**0.853****	↑**1.208***
25	[region, age]	Si., (18,30)	↑0.112	↓0.108	↓0.190	↓**0.456****	↓0.217	↑0.883	↑1.029
26	[region, age]	Si., (30,50)	↓**0.033****	↓**0.082****	↓**0.209***	↓**0.423****	↓**0.175****	↑0.873	↓**0.405****
27	[region, age]	Si., 50+	↑0.137	↓**0.061****	↓**0.071****	↓**0.478****	↓**0.152****	↑**0.954****	↑**2.225****
28	[region, gender]	Si., Man	↓0.093	↓**0.091****	↓0.260	↓**0.426****	↓**0.190****	↑0.843	↑1.022
29	[region, gender]	Si., Woman	↓0.100	↓**0.081****	↓**0.196****	↓**0.413****	↓**0.168****	↑**0.883****	↑**1.237***
30	[region, age]	SSA., (18,30)	↑**0.146****	↓0.107	↓0.280	↑0.462	↓**0.222***	↑0.834	↑**1.365****
31	[region, age]	SSA., (30,50)	↑0.135	↑0.119	↓0.160	↓0.485	↓0.218	↑0.900	↑1.137
32	[region, age]	SSA., 50+	↑0.163	↑0.125	↑0.079	↓0.592	↑0.208	↓0.950	↑1.299
33	[region, gender]	SSA., Man	↑**0.132***	↑**0.119***	↓0.286	↓**0.435***	↑0.268	↓0.829	↑1.104
34	[region, gender]	SSA., Woman	↑0.119	↑0.109	↓0.213	↓0.470	↓0.233	↑0.870	↑1.093
35	[region, age]	WE., (18,30)	↑**0.177****	↑**0.126****	↓0.246	↑0.469	↑**0.285***	↓0.849	↑**1.402****
36	[region, age]	WE., (30,50)	↓0.085	↓**0.093***	↓0.173	↓0.487	↓0.205	↑0.896	↓0.923
37	[region, age]	WE., 50+	↑0.117	↓0.104	↑0.152	↑0.545	↑0.220	↓0.905	↑1.120
38	[region, gender]	WE., Man	↑0.116	↓0.106	↓**0.214****	↓**0.443****	↑0.257	↑**0.874****	↑1.096
39	[region, gender]	WE., Woman	↑**0.151****	↑0.118	↑0.292	↓0.452	↓0.243	↓**0.825***	↑**1.284***

Table 7: Results for in-group and cross-group cohesion, and GAI for intersectional demographic groups within **D3**. Significant results are in **bold**: * for significance at $p < 0.05$, ** for significance after Benjamini-Hochberg correction. A single asterisk (*) means significant at the $p = 0.05$ level. A double asterisk (**) means the results are significant after Benjamini-Hochberg correction. A $\downarrow$ (or $\uparrow$) means that the result is less (or greater) than expected under the null hypothesis. GAI results based on $C_X = \text{XRR}$ and $C_I = \text{IRR}$.