Rater Cohesion and Quality from a Vicarious Perspective

Anonymous ACL submission

Abstract

This paper discusses and contains offensive content. Human feedback is essential for building human-centered AI systems across domains where disagreement is prevalent, such as AI safety, content moderation, or sentiment analysis. Many disagreements, particularly in politically charged settings, arise because raters have opposing values or beliefs. *Vicarious annotation* is a method for breaking down disagreement by asking raters how they think others would annotate the data. In this paper, we explore the use of vicarious annotation with analytical methods for moderating rater disagreement. We employ rater-cohesion metrics to study the potential influence of political affiliations and demographic backgrounds on raters' perceptions of offense. Additionally, we utilize *CrowdTruth*'s rater quality metrics, which consider the demographics of the raters, to score the raters and their annotations. We study how the rater-quality metrics influence the in-group and cross-group rater cohesion across the personal and vicarious levels.

1 Introduction

A crucial part of many AI systems is the humans who provide feedback for learning or evaluation (Vaughan, 2018). As AI systems grow more powerful, aligning models with human values becomes even more critical. Recent work in reinforcement learning with human feedback (RLHF) (Casper et al., 2023; MacGlashan et al., 2023) highlights the gains in model performance from aligning them to human values. This RLHF research also notes the technical challenges associated with doing so.

A major challenge to eliciting human feedback is that raters frequently disagree with each other (Uma et al., 2021). Annotating political discourse is particularly challenging because disagreements are tied to human raters' values (Jost et al., 2009), making disagreement in political domains more explicit than in other annotation tasks (Yano et al., 2010; Lukin et al., 2017; Sap et al., 2022; Weerasooriya et al., 2023).

Weerasooriya et al. (2023) introduced the concept of *vicarious offense*, where human raters are asked to annotate data according to their own opinions, and also *vicariously*, e.g., on behalf of specific groups to which they *do not* belong. Such *vicarious annotations* can reveal whether a group can be trusted to represent the opinions of other groups. If the group can be trusted, then we can recruit fewer raters from the other groups and still have enough annotations to represent the population from which the raters are drawn. If the group cannot be trusted, we need to find another group that can be trusted; otherwise, the only way to obtain a representative set of annotations is to recruit from all groups.

This paper explores group coherence in vicarious annotation tasks investigating the following research questions.

RQ1 *Are some groups more coherent than others when disclosing their own perceptions of offense?*

RQ2 *How much variance is there among the cohesion levels observed when different groups predict vicarious offense for other groups?*

RQ3 *What is the impact of removing raters deemed low-quality by CrowdTruth on group cohesion?*

We address questions using metrics introduced for understanding rater cohesion (Prabhakaran et al., 2023) and CrowdTruth (Dumitrache et al., 2018), two approaches for measuring the impacts of rater disagreement. Rater cohesion metrics measure the extent to which rater disagreement is based on group membership. CrowdTruth teases disagreement apart due to differences of opinion from poor rater quality.

2 Related Work

Prior work has highlighted the prevalence of disagreement in aggregated labels for subjective NLP tasks such as toxic language detection (Binns et al., 2017; Park et al., 2018; Sap et al., 2019; Davidson et al., 2019; Al Kuwatly et al., 2020). Disagreement is often due to rater identity (race, gender, age, education, and first language) and their beliefs (political leaning) (Sap et al., 2019; Al Kuwatly et al., 2020; Larimore et al., 2021; Sap et al., 2022; Goyal et al., 2022; Pei and Jurgens, 2023; Homan et al., 2023; Weerasooriya et al., 2023; Prabhakaran et al., 2023). Studies have also highlighted the impact of rater bias on NLP datasets (Geva et al., 2019). To uncover and analyze these differences, previous work has relied on regression models and training classifiers using demographic information and comparing their predictions (Binns et al., 2017; Davidson et al., 2019; Al Kuwatly et al., 2020; Larimore et al., 2021; Goyal et al., 2022).

Recent work has advocated the use of non-aggregated (rater-level) labels (Basile et al., 2021; Prabhakaran et al., 2021; Plank, 2022; Cabitza et al., 2023) to enable an extensive treatment of this variation. To this end, Homan et al. (2023) used Bayesian multilevel models to discover intersectional effects between rater demographics and their ratings. Prabhakaran et al. (2023) proposed a framework to analyze (dis)agreement among rater subgroups. CrowdTruth (Dumitrache et al., 2018) is another framework that benefits from rater-level labels for evaluating the quality of a dataset through three dimensions: individual raters, input data items, and the overall dataset.

Our approach has one similarity to *Bayesian truth serum*, (BTS), where "impersonally informative" questions garner more honest answers (Prelec, 2004) when these answers are gathered in groups. In practice, this means using pairs of questions where one question asks for the individual's opinion and the second asks them to estimate the group distribution for this question. "BTS relies on the Bayesian assumption that people maintain a mental model of the world that is biased by their personal experiences, which leads to a belief that personally held opinions are disproportionately present amongst peers" (Frank et al., 2017). By asking raters from one political group to consider what they believe another political group thinks, we replicate the first part of BTS methodology by getting distanced, and thus more honest perceptions of how the original rater perceives a topic. Another difference is BTS works less optimally for judging subjective social posts because it requires experts to agree on a single truth, not multiple valid truths. BTS (and its variants) are used in various crowd-sourcing projects to effectively gather more honest self-reported data on non-subjective topics (Witkowski and Parkes, 2012; Faltings et al., 2014; Frank et al., 2017).

3 Methods

3.1 Vicarious annotation

Given a dataset of machine learning training or test items $\mathcal{X}$, a rater pool $\mathcal{Z}$, a subgroup Z of $\mathcal{Z}$, and a question q with response domain $\mathcal{D}$ that is asked of each item a *vicarious annotation* of $\mathcal{X}$ with respect to $(q, \mathcal{D}, Z, \mathcal{Z})$ is a matrix $\mathbf{Y}_Z$ having one row for each item in $\mathcal{X}$ and one column for each rater $z \notin Z$, and entries in $\mathcal{D}$, where the entries are responses to the question – *How would a rater in Z annotate q?*

3.2 Group Cohesion Metrics

We use the framework and the metrics proposed by Prabhakaran et al. (2023) to compare in-group cohesion and cross-group divergences. This framework utilizes permutation tests along with proposed metrics to measure the variability of judgments by diverse rater subgroups. The metrics are either *in-group* and *cross-group*. All metrics are designed so that larger values mean more cohesion.

3.2.1 In-group Metrics

In-group metrics measure the cohesion among raters within a group. Each metric captures a slightly different aspect of cohesion, and together, they form a robust signal of group cohesion.

- **IRR:** Inter-rater reliability (IRR) is used to measure agreement among multiple raters in a way that controls for class imbalance in the distribution of ratings over all items. Specifically, we use Krippendorff's alpha (Krippendorff, 2004), a metric that generalizes several IRR metrics by accepting an arbitrary number of raters, different levels of measurement, handling missing data, and adjusting to small sample sizes.
- **Negentropy:** Negentropy (Brillouin, 1953), unlike IRR, does not control for class imbalance, but it does account for the entire distribution of rater responses for each item. It is computed by subtracting for each item the entropy over responses from

the maximum value entropy can take. Then, we compute the mean over all the items.

- **Plurality size:** Plurality size is the fraction of raters that belong to the majority vote. Traditionally, gold standard data is based on the plurality choice for each item. However, it only measures cohesion among the most popular choice for each item; it ignores the rest of the responses, in contrast to the previous two metrics.

3.2.2 Cross-group Metrics

Cross-group metrics measure the cohesion between the raters belonging to different groups. Each of these metrics roughly corresponds to an in-group metric.

- **XRR:** Cross-replication reliability (Wong et al., 2021) is similar to IRR but is defined for raters from different groups.
- **Cross negentropy:** Cross negentropy is similar to negentropy but is computed over two distributions.
- **Voting agreement:** Voting agreement is similar to plurality size and is computed by taking the most popular response for each item for each group and then calculating Krippendorff's alpha between the two groups.

3.2.3 Group Association Index

GAI combines in-group and cross-group cohesion into a single score. We define GAI as the ratio of IRR to XRR. Thus, values higher than 1 would indicate higher in-group cohesion while values smaller than 1 indicate higher cross-group cohesion. A value of 1 indicates low or no group association.

3.3 CrowdTruth

CrowdTruth (CT) (Dumitrache et al., 2018) is a framework that connects the three dimensions: individual raters, input data items, and the dataset. These are interconnected in the CT algorithm to avoid cases where disagreement from low-quality raters can lower the overall data quality as ambiguous or vice-versa. These dimensions connected through the quality of the rater are weighted by the quality of the data items the rater has annotated and the quality of the annotations in the dataset. In this study, we calculate the CT for the entire dataset with the rater demographics and focus on the individual rater quality scores. The relevant score for the research is the worker quality score (WQS), which measures the overall agreement of one rater with other raters. We compute the metrics using a publicly available implementation[1] of CT.

4 Experiments

4.1 Data

We apply this framework to $\mathcal{D}_{voiced}$ (Weerasooriya et al., 2023), a collection of 2338 comments on YouTube videos from the official channels of three leading US cable news networks (CNN, FOX, and MSNBC) labeled by diverse raters for offensiveness. Collected over eight years spanning from 2014 to 2022, the comments, therefore, cover a variety of topics. See Tables 8–9 for more information about the dataset. We consider political leaning and gender as dimensions to compare agreement among different subgroups. This dataset is unique as it also contains labels for vicarious offenses, where raters are asked to predict offenses for others who do not share their political beliefs. We also inspect these vicarious labels of offense and compare the cohesion across subgroups.

In addition, we also apply the framework to the toxicity ratings dataset (Kumar et al., 2021) of 107,620 comments from Twitter, Reddit, and 4chan labeled for toxicity by 17,280 raters. The dataset contains demographic information about the raters and their political leaning but does not contain vicarious annotations. We sample 250 batches of 20 comments, resulting in a total of 5000 items ($\mathcal{D}_{toxicity}$ from hereafter). The sampling process ensures that each batch is annotated by the same five raters, consistent with the original dataset. We then remove raters who report multiple political affiliations. See Tables 8 and 10 for more information about the dataset.

4.2 Examples

We now present three examples from $\mathcal{D}_{voiced}$ to highlight our key findings. Section A.3 contains additional examples and further discussion.

EXP1 (MSNBC): *Send trump and his deranged softy sons to Iran!!!*

The majority vote for **EXP1** across groups is offensive. Plurality scores: Democrats: 0.83, Independents: 0.87, Republicans: 0.83, Women: 0.91, Men: 0.75, Men^{IND}: 0.75, $\text{Women}^{\text{IND}}$: 1.0, and Men^{REP}: 0.5. Findings: Even though the comment is targeted towards sons of the President, there is

[1] https://github.com/CrowdTruth/CrowdTruth-core

a portion of men as a whole and even republican leaning men finding it not offensive.

EXP2 (Fox): *Pretty soon they will start counting abortions at covid deaths. Fricking Dems*

The majority vote for **EXP2** across groups is offensive. Plurality scores: Democrats: 0.66, Independents: 1, Republicans: 0.78, Women: 0.76, Men: 1.0, and Women$^{\text{DEM}}$: 0.50. Other gender$^{\text{PP}}$ had a score of 1. Findings: In contrast to **EXP1**, **EXP2** contains a targeted attack on the Democrats, yet only 0.66 of the Democrats were cohesive that it is offensive, and women were also less cohesive in contrast to men.

EXP3 (CNN): *fuck abortion dude is literally murder bruh why do they support so fucking bad killing a child like what is wrong with those ppl smh!! dude it takes Man and a woman to create babies*

The majority vote for **EXP3** across groups is offensive. Plurality scores: Democrats: 0.66, Independents: 1, Republicans: 0.77, Women: 0.76, Men: 1, Women$^{\text{DEM}}$: 0.50, and Women$^{\text{REP}}$: 0.60. Other gender$^{\text{PP}}$ had a score of 1. Findings: men are cohesive in their overall opinion, however, women are not as strongly cohesive as men.

4.3 Results

Tables 1, 2, 3, 5, 6, and 7 show results for in-group and cross-group cohesion for the personal and vicarious offense. Significant results (see Section A.2 for details) are indicated in bold at the $p = 0.05$ significance level, ↓ indicates the result is less than expected under the null hypothesis and ↑ indicates the result is greater than expected.

4.3.1 Group Cohesion for Personal Offense

Table 1 shows results on $\mathcal{D}_{voiced}$ for personal-level offense by political leaning and gender. For each group, we report in-group metrics (indicated by ∩'s) for the group and cross-group metrics (indicated by ⊗'s) between the group and all raters not in the group. For political leaning, only Independents show uniformly higher in-group and cross-group agreement than other groups. Democrats have significantly lower cross-group agreement but mixed in-group agreement results. They also have the highest GAI score. Republicans have no significant results, but all agreement metrics and GAI are lower than median random scores.

For gender, men and women have significantly lower XRR scores than random groups. Intersectional groups of political leaning and gender show some noteworthy differences from the single-variable groups and some extreme values, particularly among women. Both Republican and Democratic women have lower-than-expected in- and cross-group scores. Particularly notable is that the voting agreement for Democratic women is 0.085, compared to 0.321 for Democrats and 0.415 for all women, and is by far the lowest among all groups, single or intersectional. Republican women, by contrast, have the highest GAI score (1.434) among all groups. Independent women generally have higher than expected agreement scores, with the highest Negentropy, Cross Negentropy, and Plurality size of all groups by a substantial margin.

Table 2 shows results on $\mathcal{D}_{toxicity}$ for personal-level offense by political leaning and gender. Republicans show significantly lower IRR, XRR, and voting agreement scores than random groups. Democrats show higher-than-expected in-group scores and lower-than-expected cross-group scores. Independents have higher-than-expected in-group scores. Intersectional groups of political leaning and gender reveal some noteworthy differences. Republican women have higher-than-expected in-group scores with the highest Negentropy (0.664) and Plurality size (0.980) among all groups. In contrast, Republican men have lower-than-expected in- and cross-group scores. Particularly notable is that Republican men have the lowest scores among all groups for all cross-group metrics. They also have the lowest IRR (0.155) among all groups. Democrat men have lower-than-expected cross-group scores with the lowest GAI (0.635). Independent men have higher-than-expected in-group scores with the highest IRR (0.395) and GAI (1.506) among all groups, while Independent women have higher-than-expected cross-group scores with the highest voting agreement (0.356) among all groups.

Compared to Table 1, Democrats, Republicans, and men remain consistent with lower-than-expected cross-group scores. Independents remain consistent with higher-than-expected in-group scores and Republican men with lower-than-expected in-group scores. Women flip from their lower-than-expected in-group score to a higher-than-expected score. Democrat men flip from their higher-than-expected cross-group score to a lower-than-expected score.

Group	IRR ∩	XRR ⊗	Negentropy ∩	Cross ⊗ Negentropy	Plurality ∩ Size	Voting ⊗ Agreement	GAI
Dem	↑0.176	↓0.148	↓0.323	↓**0.257**	↓0.815	↓**0.321**	↑1.193
Rep	↓0.139	↓0.154	↓0.326	↓0.282	↓0.824	↓0.394	↓0.902
Ind	↑0.208	↑0.178	↑**0.433**	↑0.310	↑**0.872**	↑**0.488**	↑1.171
Men	↑0.178	↓**0.149**	↑0.338	↓0.301	↑0.834	↓0.414	↑1.196
Women	↓0.156	↓**0.150**	↓0.308	↑0.317	↓0.817	↓0.410	↑1.044
Dem, Men	↑0.204	↑0.177	↑0.484	↑0.310	↑0.884	↑0.335	↑1.152
Dem, Women	↑0.167	↓0.161	↓0.391	↓**0.179**	↓**0.826**	↓**0.085**	↑1.042
Rep, Men	↓0.108	↓0.150	↓0.421	↑0.280	↓0.853	↓0.308	↓0.725
Rep, Women	↑0.170	↓**0.118**	↓0.410	↓0.206	↓0.851	↓0.215	↑**1.434**
Ind, Men	↑0.203	↑0.184	↓0.457	↓0.249	↓0.868	↑0.277	↑1.103
Ind, Women	↓0.154	↓0.149	↑**0.567**	↑**0.375**	↑**0.925**	↑0.377	↑1.029

Table 1: Results of in-group and cross-group cohesion metrics on $\mathcal{D}_{voiced}$. ∩ stands for in-group metric and ⊗ stands for cross-group metric. Significant results are indicated in bold at the $p = 0.05$ significance level, ↓ indicates the result is less than expected under the null hypothesis, and ↑ indicates the result is greater than expected.

Group	IRR ∩	XRR ⊗	Negentropy ∩	Cross ⊗ Negentropy	Plurality ∩ Size	Voting ⊗ Agreement	GAI
Dem	↑0.283	↓0.258	↑0.545	↓0.494	↑0.905	↓0.287	↑1.097
Rep	↓**0.185**	↓**0.237**	↑0.596	↓0.450	↑0.933	↓**0.233**	↓0.783
Ind	↑0.292	↓0.266	↑0.610	↓0.444	↑0.942	↑0.306	↑1.097
Men	↓0.235	↓0.259	↑0.527	↓0.506	↑0.897	↓0.289	↓0.905
Women	↑0.283	↓0.251	↑0.502	↑0.538	↑0.879	↓0.277	↑1.128
Dem, Men	↓**0.157**	↓0.247	↑0.635	↓0.422	↑0.959	↓0.276	↓0.635
Dem, Women	↑0.303	↑**0.299**	↓0.602	↓0.430	↓0.937	↑0.334	↑1.013
Rep, Men	↓0.155	↓**0.221**	↓0.639	↓**0.379**	↓0.961	↓**0.223**	↓0.703
Rep, Women	↑0.287	↓0.240	↑**0.664**	↑0.424	↑**0.980**	↓**0.251**	↑1.199
Ind, Men	↑0.395	↓0.262	↑0.654	↑**0.446**	↑0.972	↓0.266	↑1.506
Ind, Women	↓0.220	↑0.282	↑0.648	↑0.419	↑0.967	↑0.356	↓0.781

Table 2: Results of in-group and cross-group cohesion metrics on $\mathcal{D}_{toxicity}$. ∩ stands for in-group metric and ⊗ stands for cross-group metric. Significant results are indicated in bold at the $p = 0.05$ significance level, ↓ indicates the result is less than expected under the null hypothesis and ↑ indicates the result is greater than expected.

4.3.2 Group Cohesion for Vicarious Offense

First, looking at in-group cohesion of vicarious predictions (e.g., Republican → Democrat) shown in Table 3 versus self-ratings from the predicting group (e.g., Republican) shown in Table 1, Republicans are more cohesive by all in-group metrics when predicting vicariously, rather than for themselves. This is also true for Democrats, but only when predicting for Independents. When predicting for Republicans in-group cohesion numbers are mixed. Independents show very similar results to Democrats.

Next, looking at in-group cohesion of vicarious predictions (e.g., Republican → Democrat, Table 3) to self-ratings from the target group (e.g., Democrats, 1), Independents have greater in-group cohesion when predicting for either Republicans or Democrats than for themselves. Both Republicans and Democrats show inconclusive results.

Finally, Table 3 compares self-ratings from the target group (e.g., Democrats) to vicarious predictions (e.g., Republican → Democrat). Independents have a higher cohesion with Democrats when predicting vicarious offense for Democrats than Republicans predicting vicarious offense for them by all cross-group metrics. Independents also have higher cohesion with Republicans when predicting vicariously for Republicans than Democrats predicting vicariously for Republicans by all cross-group metrics. Particularly noteworthy is that the voting agreement for Independents predicting vicariously for Republicans is significantly higher (0.354) as compared to Democrats predicting vicariously for Republicans which is the lowest among all groups (0.247). Republicans have a higher cohesion with Independents when predicting vicariously for them than Democrats predicting vicariously for Independents by all cross-group metrics.

4.4 CrowdTruth Evaluation

As introduced earlier, CrowdTruth's triangle of disagreement is dependent on the raters, data item/unit, and dataset/task. We focus on the worker quality score (WQS, ranging from a minimum of 0 to 1) for this study. The WQS measures the overall agreement of one rater over other raters and favors raters

Group	IRR ∩	XRR ⊗	Negentropy ∩	Cross ⊗ Negentropy	Plurality ∩ Size	Voting ⊗ Agreement	GAI
Rep → Dem (v Dem)	↓0.164	↓0.140	↓**0.386**	↓0.330	↓**0.855**	↑0.330	↓1.169
Ind → Dem (v Dem)	↑0.220	↑0.183	↓0.460	↓0.349	↓0.886	↑0.353	↓1.201
Dem → Rep (v Rep)	↑0.172	↓0.127	↓**0.300**	↓**0.259**	↓**0.797**	↑0.247	↑1.350
Ind → Rep (v Rep)	↑0.188	↑0.163	↑0.425	↑0.343	↑0.863	↑**0.354**	↓1.153
Dem → Ind (v Ind)	↑0.143	↓0.145	↓**0.328**	↑**0.416**	↓**0.815**	↓0.268	↑0.982
Rep → Ind (v Ind)	↑0.141	↑0.171	↓0.347	↑**0.421**	↓0.832	↑0.323	↓0.824

Table 3: Results of vicarious alignment on $\mathcal{D}_{voiced}$. ∩ stands for in-group metric and ⊗ stands for cross-group metric. Significant results are indicated in bold at the $p = 0.05$ significance level, ↓ indicates the result is less than expected under the null hypothesis and ↑ indicates the result is greater than expected.

	$\mathcal{D}_{voiced}$	$\mathcal{D}_{toxicity}$
Annotations	2250	360
Data items	1405	360
Democrats	8	3
Republicans	9	12
Independents	8	2

Table 4: Data impacted after CrowdTruth filtering.

that agree with others.

Figure 1 shows the distribution of WQS from CrowdTruth. We filter out raters with a WQS below 0.1 and re-run our cohesion metrics on the remaining dataset. We identified WQS 0.1 as a cut-off based on the distribution of the scores.

Table 4 shows the data impacted after filtering out raters deemed low-quality using CrowdTruth. Nearly identical numbers of Republican-, Democrat-, and Independent-leaning raters were removed (8–9 each) from $\mathcal{D}_{voiced}$. More Republican-leaning raters (12) were removed from $\mathcal{D}_{toxicity}$ than Democrat- and Independent-leaning raters (3 and 2, respectively).

4.4.1 Results of Group Cohesion for Personal Offense after CrowdTruth Filtering

Table 5 shows the results on $\mathcal{D}_{voiced}$ for in-group and cross-group cohesion metrics for the dimensions of political leaning and gender after CrowdTruth filtering. Compared to Table 1, overall, nearly all in- and cross-group metrics increase after applying CrowdTruth. The most noteworthy exception is that Independent women have lower IRR and GAI scores. The number of significant results added and removed between tables after CrowdTruth filtering are all the same (seven). This number is approximately the same as the expected false positive rate at $p = 0.05$, taking both Tables 1 and 5 into account. This result illustrates why we should not read the p-values here as a measure of statistical significance *per se*, but rather as a relative measure of the likelihood the effects are due to a true difference in the underlying population from (a) random group(s) of the same size(s).

Figure 1: Distribution of CrowdTruth worker quality score (WQS) for each rater in the datasets. We use the WQS to filter out lower-rated raters from the datasets.

Table 6 shows the results on $\mathcal{D}_{toxicity}$ for in-group and cross-group cohesion metrics for the dimensions of political leaning and gender after CrowdTruth filtering. Compared to Table 2, nearly all in- and cross-group metrics increase after applying CrowdTruth. Particularly notable is that Republicans and Republican men have higher-than-expected in-group scores and lower-than-expected cross-group scores.

Group	IRR ∩	XRR ⊗	Negentropy ∩	Cross ⊗ Negentropy	Plurality ∩ Size	Voting ⊗ Agreement	GAI
Dem	↑0.238	↓0.197	↓0.403	↓0.349	↓0.855	**↓0.367**	**↑1.203**
Rep	↓0.167	↓0.193	↓0.376	↑0.381	↓0.851	↓0.473	↓0.864
Ind	↑0.251	↑0.215	**↑0.487**	↑0.383	**↑0.898**	↑0.537	↑1.165
Men	↑0.213	**↓0.187**	↑0.387	↓0.384	↑0.861	↓0.493	↑1.141
Women	↓0.202	**↓0.187**	↓0.379	↑0.384	↓0.854	↓0.482	↑1.085
Dem, Men	↑0.204	↑0.205	↓0.484	↑0.359	↓0.884	↓0.340	↑0.993
Dem, Women	**↑0.305**	↑0.222	↓0.507	↓0.302	↓0.892	**↓0.206**	**↑1.373**
Rep, Men	↓0.148	↓0.197	↑0.481	↑0.371	↑0.885	↑0.371	↓0.750
Rep, Women	↓0.175	**↓0.154**	**↓0.433**	↓0.299	**↓0.864**	↓0.272	↑1.142
Ind, Men	↑0.284	↑0.241	↑0.537	↓0.348	↑0.910	↑0.349	↑1.182
Ind, Women	**↓0.110**	↓0.174	↑0.572	**↑0.423**	↑0.930	↑0.393	**↓0.631**
Δ	0.047	0.041	0.053	0.083	0.029	0.060	0.130

Table 5: Results of in-group and cross-group cohesion metrics on $\mathcal{D}_{voiced}$ after CrowdTruth filtering. ∩ stands for in-group metric and ⊗ stands for cross-group metric. Significant results are indicated in bold at the $p = 0.05$ significance level, ↓ indicates the result is less than expected under the null hypothesis, and ↑ indicates the result is greater than expected. Orange indicates the result is significant before applying CT, Cyan indicates the result is significant after applying CT, and Green indicates the result is significant before and after applying CT. Δ is the mean absolute difference of metric scores before and after applying CT.

Group	IRR ∩	XRR ⊗	Negentropy ∩	Cross ⊗ Negentropy	Plurality ∩ Size	Voting ⊗ Agreement	GAI
Dem	↓0.293	↓0.289	↑0.548	↑0.518	↑0.907	↓0.306	↑1.013
Rep	↑0.307	**↓0.275**	**↑0.618**	**↓0.448**	↑0.948	**↓0.272**	↑1.119
Ind	↓0.294	↓0.292	↑0.611	↓0.460	↑0.943	↓0.325	↑1.009
Men	↓0.291	↓0.291	↑0.544	↓0.515	↑0.906	↓0.309	↑1.003
Women	↓0.297	↓0.285	↓0.508	↑0.552	↓0.883	↓0.303	↑1.042
Dem, Men	**↓0.157**	↓0.284	↑0.634	↑0.447	↓0.958	↓0.298	**↓0.552**
Dem, Women	↑0.303	↑0.318	↓0.601	↓0.449	↓0.937	↑0.344	↓0.953
Rep, Men	↑0.372	↓0.274	↑0.661	**↓0.389**	↑0.977	↓0.287	↑1.361
Rep, Women	↑0.357	↓0.277	**↑0.668**	↑0.439	**↑0.983**	↓0.282	↑1.287
Ind, Men	↑0.395	↓0.297	↑0.653	**↑0.458**	↑0.971	↓0.305	↑1.330
Ind, Women	↓0.220	↓0.294	↑0.648	↑0.442	↓0.967	↑0.354	↓0.747
Δ	0.044662	0.032115	0.007043	0.015255	0.004648	0.026464	0.162877

Table 6: Results of in-group and cross-group cohesion metrics on $\mathcal{D}_{toxicity}$ after CrowdTruth filtering. ∩ stands for in-group metric and ⊗ stands for cross-group metric. Significant results are indicated in bold at the $p = 0.05$ significance level, ↓ indicates the result is less than expected under the null hypothesis, and ↑ indicates the result is greater than expected. Orange indicates the result is significant before applying CT, Cyan indicates the result is significant after applying CT, and Green indicates the result is significant before and after applying CT. Δ is the mean absolute difference of metric scores before and after applying CT.

4.4.2 Results of Group Cohesion for Vicarious Offense after CrowdTruth Filtering

Table 7 shows the results for in-group and cross-group cohesion metrics for vicarious predictions after CrowdTruth filtering. Compared to Table 3, overall, nearly all in- and cross-group metrics increase after applying CrowdTruth except for some minor variation in GAI.

5 Discussion

Regarding RQ1, the major takeaways are that, of the political groups, Independents are the most cohesive, both with themselves and with others. Democrats are the least cohesive with others. Republicans are the least internally cohesive.

Because Independents split their votes between Democrats and Republicans in most elections, we were not surprised by their relatively high level of cohesion with other groups. However, their internal cohesion was not as commonsensical. Perhaps it was due to the Democrats and Republicans containing both extremist and more moderate members, who tend to agree on inoffensive content but whose extreme members are more readily "triggered" by moderately offensive content. And perhaps Independents contain fewer extreme members. Regarding intersections between gender and political leaning, considering women raters mostly amplifies the results seen by political leaning. Independent women have the highest cohesion among

Group	IRR ∩	XRR ⊗	Negentropy ∩	Cross ⊗ Negentropy	Plurality ∩ Size	Voting ⊗ Agreement	GAI
Rep → Dem (v Dem)	↓0.181	↓0.176	**↓0.419**	↓0.411	**↓0.871**	↓0.331	↓1.027
Ind → Dem (v Dem)	↑0.252	**↑0.231**	↓0.502	↓0.423	↑0.906	**↑0.418**	↓1.091
Dem → Rep (v Rep)	↑0.230	↓0.166	**↓0.376**	**↓0.346**	**↓0.840**	↑0.283	↑1.389
Ind → Rep (v Rep)	↑0.215	↑0.191	↑0.470	↓0.402	↑0.887	**↑0.393**	↓1.123
Dem → Ind (v Ind)	↑0.203	↑0.200	↓0.413	**↑0.487**	↓0.860	↑0.353	↑1.016
Rep → Ind (v Ind)	↓0.164	↑0.200	**↓0.393**	**↑0.486**	↓0.857	↑0.372	↓0.821
Δ	0.036	0.039	0.055	0.073	0.029	0.046	0.060

Table 7: Results of vicarious alignment on $\mathcal{D}_{voiced}$ after CrowdTruth filtering. ∩ stands for in-group metric and ⊗ stands for cross-group metric. Significant results are indicated in bold at the $p = 0.05$ significance level, ↓ indicates the result is less than expected under the null hypothesis, and ↑ indicates the result is greater than expected. Orange indicates the result is significant before applying CT, Cyan indicates the result is significant after applying CT, and Green indicates the result is significant before and after applying CT. Δ is the mean absolute difference of metric scores before and after applying CT.

themselves as well as with other groups. Democrat women have lower cohesion among themselves and with other groups. Our results suggest that women are driving disagreement among Democrats and agreement among Independents.

Regarding RQ2, Independents have higher cohesion with Democrats and Republicans while predicting vicariously for them. Republicans have a higher cohesion with Independents while predicting vicarious offense. Democrats, again, appear to be the most isolated, this time in terms of their ability to predict what other groups find offensive.

For RQ3, it is not surprising that using CrowdTruth to remove raters leads to higher cohesion scores; the CrowdTruth quality metrics depend on within-group agreement levels. One notable exception is Independent women. In terms of p-value, XRR is the most stable metric. CrowdTruth filtering seems to particularly benefit IRR and XRR, and hurt the Negentropy metrics the most.

5.1 Implications for Data Collection

Since Independents are the most cohesive, one might conclude that with a limited budget, it would make sense to have slightly more Independents than other raters, because of their high cross-group cohesiveness. However, one must weigh this utilitarian conclusion against the cold reality that offensive content is sometimes directed at marginal groups by other marginal groups in such a way as to be unnoticeable by most people. And so, particularly in settings where such *dog whistling* behavior is likely, this is likely the wrong action to take (Mendelsohn et al., 2023).

The relatively low level of cohesion between Democrats and other raters suggests that some minimum amount of the budget should be allocated to Democrats, because the other raters do not represent their beliefs. However, the relatively low level of cohesion between Democrats predicting vicariously for others suggests that they should not get too much budget. The relatively low levels of in-group cohesion for Women Democrats suggest they should have a larger substantial portion of the Democratic budget than men because there is more variance in their responses.

A critical open question remains: how many raters do we need in each group for each item to have confidence that our annotations fairly represent the target audience? This is an important question that we believe has yet to get the attention it truly deserves.

6 Conclusion

Our investigation into the dynamics of rater cohesion in politically charged content moderation settings, through the lens of self and *vicarious annotation*, gender, and political affiliations, reveals valuable insights into the challenges of building inclusive and human-centered AI systems. Our findings reveal notable disparities in cohesion levels, highlighting the influence of gender and political affiliation. For instance, Independent women and Democrat women show significantly different patterns of cohesion both within their groups and with other groups. We also note that Independents show higher vicarious cohesion with other groups. This finding opens up a strategic avenue for more efficient rater recruitment, implying that Independents can effectively approximate the viewpoints of Democrats and Republicans. Consequently, *vicarious annotation* emerges as a valuable tool for optimizing rater recruitment, ensuring diverse representation under resource constraints.

7 Limitations

While our study computes subgroup cohesion metrics along two critical demographic dimensions (gender and political leaning), the findings may not be generalizable to other demographics such as education level, cultural background, and economic status. Future studies should employ the proposed framework to investigate the level of cohesion among raters belonging to other important demographic subgroups. Another limitation of this work is the simplification of political ideologies into three groups: Democrats, Republicans, and Independents. This, however, may not capture the full spectrum of political beliefs and identities. For instance, a rater can be socially Republican but fiscally Liberal. A more granular analysis that considers the multidimensional nature of political ideologies could reveal more intricate patterns of cohesion.

CrowdTruth is inherently an algorithm designed to dissolve disagreements. By filtering out lower-scored raters, we can remove disagreements, resulting in a more agreeable dataset.

Ethics Statement

The datasets utilized in this study consist of a human-annotated compilation of publicly accessible YouTube, Twitter, Reddit, and 4chan comments, as introduced by Kumar et al. (2021) and Weerasooriya et al. (2023). These datasets do not reveal any identifiable information about the raters. The authors of the original datasets claimed they consulted their institutional review board to ensure the safety of raters during the data collection. We are aware that in some previous studies, raters had raised concerns about the impact of mental trauma when annotating for safety for ChatGPT (Hao and Seetharaman, 2023) and social networks (Wexler, 2023). However, no such concerns were reported by the authors of these datasets.

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, Online. Association for Computational Linguistics.

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

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II 9*, pages 405–415. Springer.

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

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *arXiv preprint*. ArXiv:2307.15217 [cs].

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. *arXiv preprint*. ArXiv:1808.06080 [cs].

Boi Faltings, Radu Jurca, Pearl Pu, and Bao Duy Tran. 2014. Incentives to counter bias in human computation. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 2, pages 59–66.

Morgan R. Frank, Manuel Cebrian, Galen Pickard, and Iyad Rahwan. 2017. Validating bayesian truth serum in large-scale online human experiments. *PLOS ONE*, 12(5):1–13.

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. *arXiv preprint arXiv:1908.07898*.

Jelle J. Goeman and Aldo Solari. 2011. Multiple Testing for Exploratory Research. *Statistical Science*, 26(4):584 – 597.

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.

Karen Hao and Deepa Seetharaman. 2023. Cleaning Up ChatGPT Takes Heavy Toll on Human Workers. *Wall Street Journal*.

Christopher M Homan, Greg Serapio-Garcia, Lora Aroyo, Mark Diaz, Alicia Parrish, Vinodkumar Prabhakaran, Alex S Taylor, and Ding Wang. 2023. Intersectionality in conversational ai safety: How bayesian multilevel models help understand diverse perceptions of safety. *arXiv preprint arXiv:2306.11530*.

John T. Jost, Christopher M. Federico, and Jaime L. Napier. 2009. Political Ideology: Its Structure, Functions, and Elective Affinities. *Annual Review of Psychology*, 60(1):307–337. _eprint: https://doi.org/10.1146/annurev.psych.60.110707.163600.

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

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.

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

Stephanie M. Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument Strength is in the Eye of the Beholder: Audience Effects in Persuasion. *arXiv preprint*. ArXiv:1708.09085 [cs].

James MacGlashan, Mark K. Ho, Robert Loftin, Bei Peng, Guan Wang, David Roberts, Matthew E. Taylor, and Michael L. Littman. 2023. Interactive Learning from Policy-Dependent Human Feedback. *arXiv preprint*. ArXiv:1701.06049 [cs].

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15162–15180, Toronto, Canada. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Drazen Prelec. 2004. A bayesian truth serum for subjective data. *Science (New York, N.Y.)*, 306:462–6.

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

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.

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.

Jennifer Wortman Vaughan. 2018. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *Journal of Machine Learning Research*, 18(193):1–46.

	$\mathcal{D}_{voiced}$	$\mathcal{D}_{toxicity}$
Items	2338	5000
Raters	726	803
Raters per item (min, median, max)	(5, 19, 60)	(1, 3, 5)
Annotations	45725	16380

Table 8: Dataset annotation statistics

	Dem	Rep	Ind	Total
Men	126	146	113	385
Women	118	115	103	336
NA	3	1	1	5
Total	247	262	217	726

Table 9: $\mathcal{D}_{voiced}$ raters in political leaning X gender intersectional groups

	Dem	Rep	Ind	Total
Men	147	114	108	369
Women	201	116	111	428
NA	4	2	0	6
Total	352	232	219	803

Table 10: $\mathcal{D}_{toxicity}$ raters in political leaning X gender intersectional groups

Tharindu Weerasooriya, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher Homan, and Ashiqur KhudaBukhsh. 2023. Vicarious offense and noise audit of offensive speech classifiers: Unifying human and machine disagreement on what is offensive. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11648–11668, Singapore. Association for Computational Linguistics.

Alexandra Wexler. 2023. Lawsuits by Moderators of Violent Online Content Pose Threat to Big Tech. *Wall Street Journal*.

Jens Witkowski and David Parkes. 2012. A robust bayesian truth serum for small populations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 26(1):1492–1498.

Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication reliability - an empirical approach to interpreting inter-rater reliability. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7053–7065, Online. Association for Computational Linguistics.

T. Yano, P. Resnik, and Noah A. Smith. 2010. Shedding (a Thousand Points of) Light on Biased Language.

A Appendix

A.1 Dataset Statistics

Tables 8–10 show the annotation statistics and the raters in political leaning X gender intersectional groups for $\mathcal{D}_{voiced}$ and $\mathcal{D}_{toxicity}$.

A.2 Significance Testing

Following Prabhakaran et al. (2023), we utilize *null hypothesis significance tests (NHSTs)* to test the significance of our results. For any cohesion metric our null hypothesis $\mathbf{H_{null}}$ is that the value of the in-group or cross-group metric for any subgroup is independent of the political leaning and demographics of the raters.

We conduct permutation tests by randomly shuffling the political leaning and demographics of the raters, measuring the shuffled test statistic, and counting the number of times the shuffled statistic exceeds (lags) the observed value. For us, p-value is the fraction of times we observe a test statistic that is more extreme than the observed value. We use 1000 trials for our experiments.

Given the large number of tests we conduct, the value of NHST for us is not in the tests *per se*, and so we do not correct for the false discovery rate. Rather, we follow Goeman and Solari (2011) and others, who advocate using p-values in exploratory settings as a concise way of measuring the *relative* significance of some results versus others.

A.3 Metric-Guided Heuristic Study

We sampled examples from $\mathcal{D}_{voiced}$ based on the cohesion metrics to understand the strengths and limitations of this work. We discussed EXP1 through EXP3, highlighting how each targeted demographic group shows the least cohesion. This theme exists for cases such as EXP5 and EXP6.

EXP5 is an example where the comment is not attributing "Republic" to the Republicans but as a call to action to galvanize fellow Americans to vote for gun control and take action. Only 0.71 of the Republicans agree with this perspective, and women are even less cohesive.

EXP7 is a case where the comment is offensive to the Democrat-leaning voters. However, they are the least cohesive out of the three political leanings. This further supports the observation of impacted/called-out demographic groups being less cohesive.

EXP1 (MSNBC): *Send trump and his deranged softy sons to Iran!!!*

The majority vote for **EXP1** across groups is offensive. Plurality scores: Democrats: 0.83, Independents: 0.87, Republicans: 0.83, Women: 0.91, Men: 0.75, Men^{IND}: 0.75, Women^{IND}: 1.0, and Men^{REP}: 0.5.

EXP2 (Fox): *Pretty soon they will start counting abortions at covid deaths . Fricking Dems*

The majority vote for **EXP2** across groups is offensive. Plurality scores: Democrats: 0.66, Independents: 1, Republicans: 0.78, Women: 0.76, Men: 1.0, and Women^{DEM}: 0.50. Other gender^{PP} had a score of 1.

EXP3 (CNN): *fuck abortion dude is literally murder bruh why do they support so fucking bad killing a child like what is wrong with those ppl smh!! dude it takes Man and a woman to create babies*

The majority vote for **EXP3** across groups is offensive. Plurality scores: Democrats: 0.66, Independents: 1, Republicans: 0.77, Women: 0.76, Men: 1, Women^{DEM}: 0.50, and Women^{REP}: 0.60. Other gender^{PP} had a score of 1.

EXP4 (MSNBC): *I love me some liberal tears! Let's Go Brandon!!!*

The majority vote for **EXP4** across groups is offensive. Plurality scores: Democrats: 0.50, Independents: 0.85, Republicans: 0.57, Women: 0.50, Men: 0.87, and Women^{REP}: 0.75. Other gender^{PP} had a score of 1.

EXP5 (MSNBC): *Absolutely useless posts! Here are the facts: URL These corrupt politicians and lobbyist need to go! Get out and vote for safer gun laws in November for Senate Seat and in 2020!! WE the Republic are in charge -not these clowns! Remember, it could be your loved one next!!*

The majority vote for **EXP5** across groups except for Democrat-leaning women is offensive. Plurality scores: Democrats: 0.57, Independents: 1, Republicans: 0.71, Women: 0.57, Men: 0.87, Women^{DEM}: 0.60, and Men^{REP}: 0.83. Other gender^{PP} had a score of 1.

EXP6 (FOX): *If a person kills a pregnant mother, do they not get charged for a double homicide?*

The majority vote for **EXP6** across groups is offensive. Plurality scores: Democrats: 0.60, Independents: 1, Republicans: 0.80, Women: 0.50, Men: 1, Women^{DEM}: 0.66, and Women^{REP}: 0.50. Other gender^{PP} had a score of 1.

EXP7 (MSNBC): *heres a great idea why dont the rich and stupid old biden get rid of the guns on there body guards first before they try telling Americans to give up there weapons lets go brandan stupid dems beed to be thrown in prison they are traiters to this country*

The majority vote for **EXP7** across groups is offensive. Plurality scores; Democrats: 0.66, Independents: 1, Republicans: 0.80, Women: 0.50, Men: 1, and Women^{DEM}: 0.80. Other gender^{PP} had a score of 1.

A.4 Experimental Code

The code to reproduce results will be released upon acceptance of the paper. The anonymized version of the code is available on `https://anonymous.4open.science/r/emnlp_2592/`.

A.5 Median permutation test scores

Group	IRR ∩	XRR ⊗	Negentropy ∩	Cross ⊗ Negentropy	Plurality ∩ Size	Voting ⊗ Agreement	GAI
Dem	0.159	0.162	0.354	0.287	0.833	0.414	0.984
Rep	0.160	0.162	0.347	0.289	0.832	0.424	0.983
Ind	0.160	0.162	0.372	0.284	0.839	0.393	0.989
Man	0.159	0.163	0.316	0.315	0.824	0.456	0.980
Woman	0.161	0.163	0.325	0.303	0.826	0.456	0.991
Dem, Man	0.155	0.163	0.445	0.271	0.864	0.286	0.961
Dem, Woman	0.158	0.162	0.458	0.271	0.870	0.274	0.983
Rep, Man	0.157	0.163	0.422	0.272	0.856	0.310	0.969
Rep, Woman	0.155	0.162	0.461	0.271	0.871	0.271	0.967
Ind, Man	0.158	0.162	0.465	0.272	0.874	0.271	0.988
Ind, Woman	0.160	0.162	0.481	0.271	0.880	0.259	0.994

Table 11: Median permutation scores for in-group and cross-group cohesion on $\mathcal{D}_{voiced}$

Group	IRR ∩	XRR ⊗	Negentropy ∩	Cross ⊗ Negentropy	Plurality ∩ Size	Voting ⊗ Agreement	GAI
Dem	0.270	0.271	0.532	0.501	0.897	0.298	0.997
Rep	0.271	0.271	0.590	0.454	0.931	0.301	1.001
Ind	0.269	0.272	0.596	0.449	0.935	0.304	0.991
Men	0.270	0.270	0.526	0.507	0.894	0.299	1.000
Women	0.270	0.270	0.498	0.532	0.878	0.299	0.996
Dem, Men	0.270	0.272	0.629	0.426	0.956	0.304	0.997
Dem, Women	0.270	0.271	0.603	0.444	0.939	0.303	0.997
Rep, Men	0.275	0.272	0.644	0.415	0.966	0.307	1.001
Rep, Women	0.270	0.272	0.644	0.415	0.966	0.305	1.004
Ind, Men	0.271	0.270	0.647	0.412	0.968	0.304	1.003
Ind, Women	0.272	0.271	0.646	0.413	0.967	0.306	1.008

Table 12: Median permutation scores for in-group and cross-group cohesion on $\mathcal{D}_{toxicity}$

Group	IRR ∩	XRR ⊗	Negentropy ∩	Cross ⊗ Negentropy	Plurality ∩ Size	Voting ⊗ Agreement	GAI
Rep → Dem (v Dem)	0.184	0.150	0.458	0.370	0.881	0.296	1.201
Ind → Dem (v Dem)	0.184	0.153	0.480	0.369	0.888	0.283	1.203
Dem → Rep (v Rep)	0.168	0.136	0.402	0.331	0.846	0.227	1.223
Ind → Rep (v Rep)	0.166	0.137	0.422	0.329	0.854	0.212	1.222
Dem → Ind (v Ind)	0.137	0.147	0.392	0.357	0.844	0.268	0.922
Rep → Ind (v Ind)	0.137	0.148	0.386	0.358	0.842	0.271	0.921

Table 13: Median permutation scores for vicarious alignment on $\mathcal{D}_{voiced}$

Group	IRR ∩	XRR ⊗	Negentropy ∩	Cross ⊗ Negentropy	Plurality ∩ Size	Voting ⊗ Agreement	GAI
Dem	0.202	0.204	0.414	0.367	0.864	0.471	0.988
Rep	0.201	0.203	0.409	0.371	0.864	0.484	0.985
Ind	0.202	0.203	0.427	0.363	0.868	0.449	0.990
Man	0.201	0.203	0.377	0.394	0.856	0.512	0.988
Woman	0.203	0.203	0.389	0.383	0.859	0.512	1.001
Dem, Man	0.199	0.202	0.488	0.350	0.889	0.342	0.980
Dem, Woman	0.202	0.202	0.507	0.349	0.897	0.328	0.996
Rep, Man	0.197	0.202	0.474	0.356	0.884	0.370	0.977
Rep, Woman	0.198	0.201	0.506	0.351	0.897	0.329	0.978
Ind, Man	0.198	0.200	0.510	0.349	0.897	0.322	0.983
Ind, Woman	0.201	0.202	0.517	0.347	0.901	0.315	0.989

Table 14: Median permutation scores for in-group and cross-group cohesion on $\mathcal{D}_{voiced}$ after CrowdTruth filtering

Group	IRR ∩	XRR ⊗	Negentropy ∩	Cross ⊗ Negentropy	Plurality ∩ Size	Voting ⊗ Agreement	GAI
Dem	0.304	0.304	0.541	0.515	0.903	0.326	1.001
Rep	0.304	0.303	0.601	0.466	0.938	0.327	1.004
Ind	0.300	0.302	0.601	0.465	0.938	0.327	0.998
Men	0.304	0.304	0.539	0.518	0.902	0.326	1.000
Women	0.303	0.304	0.511	0.544	0.886	0.326	0.998
Dem, Men	0.303	0.304	0.633	0.443	0.959	0.331	1.002
Dem, Women	0.302	0.303	0.607	0.461	0.942	0.328	0.995
Rep, Men	0.296	0.303	0.648	0.431	0.969	0.330	0.972
Rep, Women	0.303	0.302	0.648	0.432	0.968	0.328	0.995
Ind, Men	0.294	0.302	0.650	0.430	0.970	0.330	0.975
Ind, Women	0.293	0.301	0.647	0.431	0.968	0.326	0.988

Table 15: Median permutation scores for in-group and cross-group cohesion on $\mathcal{D}_{toxicity}$ after CrowdTruth filtering

Group	IRR ∩	XRR ⊗	Negentropy ∩	Cross ⊗ Negentropy	Plurality ∩ Size	Voting ⊗ Agreement	GAI
Rep → Dem (v Dem)	0.208	0.183	0.492	0.433	0.899	0.337	1.130
Ind → Dem (v Dem)	0.211	0.184	0.511	0.433	0.905	0.321	1.137
Dem → Rep (v Rep)	0.202	0.172	0.448	0.406	0.871	0.264	1.171
Ind → Rep (v Rep)	0.202	0.170	0.467	0.405	0.879	0.253	1.179
Dem → Ind (v Ind)	0.171	0.186	0.444	0.430	0.872	0.319	0.921
Rep → Ind (v Ind)	0.170	0.186	0.439	0.431	0.871	0.328	0.915

Table 16: Median permutation scores for vicarious alignment on $\mathcal{D}_{voiced}$ after CrowdTruth filtering