Annotators with Attitudes:
How Annotator Beliefs And Identities Bias Toxic Language Detection

Maarten Sap ♣♠ Swabha Swayamdipta ♠ Laura Vianna ♢
Xuhui Zhou ♣ Yejin Choi ♣♠ Noah A. Smith ♣♠

♣♠ Paul G. Allen School of Computer Science, University of Washington, Seattle, WA, USA
♠ Allen Institute for AI, Seattle, WA, USA
♢ Department of Psychology, University of Washington, Seattle, WA, USA
♣ Georgia Institute of Technology, Atlanta, GA, USA

{maartens,swabhas}@allenai.org

Abstract

Warning: this paper discusses and contains content that is offensive or upsetting.

The perceived toxicity of language can vary based on someone’s identity and beliefs, but this variation is often ignored when collecting toxic language datasets, resulting in dataset and model biases. We seek to understand the who, why, and what behind biases in toxicity annotations. In two online studies with demographically and politically diverse participants, we investigate the effect of annotator identities (who) and beliefs (why), drawing from social psychology research about hate speech, free speech, racist beliefs, political leaning, and more. We disentangle what is annotated as toxic by considering posts with three characteristics: anti-Black language, African American English (AAE) dialect, and vulgarity. Our results show strong associations between annotator identity and beliefs and their ratings of toxicity. Notably, more conservative annotators and those who scored highly on our scale for racist beliefs were less likely to rate anti-Black language as toxic, but more likely to rate AAE as toxic.

We additionally present a case study illustrating how a popular toxicity detection system’s ratings inherently reflect only specific beliefs and perspectives. Our findings call for contextualizing toxicity labels in social variables, which raises immense implications for toxic language annotation and detection.

1 Introduction

Determining whether a text is toxic (i.e., contains hate speech, abuse, or is offensive) is inherently a subjective task that requires a nuanced understanding of the pragmatic implications of language (Fiske, 1993; Croom, 2011; Waseem et al., 2021). Without this nuance, both humans and machines are prone to biased judgments, such as over-relying on seemingly toxic keywords (e.g., expletives, swearwords; Dinan et al., 2019; Han and Tsvetkov, 2020) or backfiring against minorities (Yasin, 2018; Are, 2020, i.a.). For example, racial biases have been uncovered in toxic language detection where text written in African American English (AAE) is falsely flagged as toxic (Sap et al., 2019; Davidson et al., 2019).

The crux of the issue is that not all text is equally toxic for everyone (Waseem, 2016; Al Kuwatly et al., 2020). Yet, most previous research has treated this detection as a simple classification with one correct label, obtained by averaging judgments by a small set of human raters per post (Waseem and Hovy, 2016; Wulczyn et al., 2017; Davidson et al., 2017; Founta et al., 2018; Zampieri et al., 2019). Such approaches ignore the variance in annotations (Pavlick and Kwiatkowski, 2019; Geva et al., 2019; Arhin et al., 2021; Akhtar et al., 2021) based on who the annotators are, and what their beliefs are.

In this work, we investigate the who, why, and what behind biases in toxicity annotations, through online studies with demographically and politically diverse participants. We measure the effects of annotator identities (who annotates as toxic) and attitudes or beliefs (why they annotate as toxic) on toxicity perceptions, through the lens of social psychology research on hate speech, free speech, racist beliefs, altruism, political leaning, and more. We also analyze the effect of what is being rated, by considering three text characteristics: anti-Black or racially prejudiced meaning, African American English (AAE), and vulgar words.

We seek to answer these questions via two online studies. In our breadth-of-workers controlled study, we collect ratings of toxicity for a set of 15 hand curated posts from 641 annotators of different races, attitudes, and political leanings.

1 We use the term “bias” to denote both simple skews or variation in annotations (e.g., for variation in detecting vulgar content as toxic) or representational harms (e.g., AAE being over-detected as toxic or anti-Black content being under-detected as toxic; Barocas et al., 2017; Blodgett et al., 2020).
Then, in our breadth-of-posts study, we simulate a typical toxic language annotation setting by collecting toxicity ratings for ~600 posts, from a smaller but diverse pool of 173 annotators.2

Distilled in Figure 1, our most salient results across both studies show that annotators scoring higher on our racist beliefs scale were less likely to rate anti-Black content as toxic (§4). Additionally, annotators’ conservatism scores were associated with higher ratings of toxicity for AAE (§5), and conservative and traditionalist attitude scores with rating vulgar language as more toxic (§6).

We further provide a case study which shows that PERSPECTIVE API, a popular toxicity detection system, mirrors ratings by annotators of certain attitudes and identities over others (§7). For instance, for anti-Black language, the system’s scores better reflect ratings by annotators who score high on our scale for racist beliefs. Our findings have immense implications for the design of toxic language annotation and automatic detection—we recommend contextualizing ratings in social variables and looking beyond aggregated discrete decisions (§8).

2 The Who, Why, and What of Toxicity Annotations

We aim to investigate how annotators’ ratings of the toxicity of text is influenced by their own identities (who they are; §2.1), and their beliefs (why they consider something toxic; §2.2) on specific categories of text (what they consider toxic; §2.3)—namely, text with anti-Black language, presence of African American English (AAE), and presence of vulgar or profane words. To this end, we design two online studies (§3) and discuss who find each of these text characteristics offensive and why as separate research questions in Sections §4, §5, and §6, respectively.

2.1 Demographic Identities: Who considers something as toxic?

Prior work has extensively shown links between someone’s gender, political leaning, and race affects how likely they are to perceive or notice harmful speech or racism (Cowan et al., 2002; Norton and Sommers, 2011; Carter and Murphy, 2015; Prabhakaran et al., 2021). Grounded in this prior literature, our study considers annotators’ race, gender, and political leaning. Since perceptions of race and political attitudes vary vastly across the globe, we restrict our study to participants exclusively from the United States.

2.2 Attitudes: Why does someone consider something toxic?

While some annotator toxicity ratings may highly correlate with demographic factors at face value (Prabhakaran et al., 2021; Jiang et al., 2021), we aim to go beyond demographics to investigate annotator beliefs that explain these correlations. Based on prior work in social psychology, polit-
Valuing the freedom of offensive speech (FREEOFFSPEECH): the belief that any speech, including offensive or hateful speech, should be unrestricted and free from censorship. Recently, this belief has become associated with majority and conservative identities (Cole, 1996; Gillborn, 2009; White and Crandall, 2017; Elers and Jayan, 2020). We use the scale by Cowan and Khatchadourian (2003); see Appendix A.1.

Perceiving the HARMOFHATE SPEECH: the belief that hate speech or offensive language can be harmful for the targets of that speech (Soral et al., 2018; Nadal, 2018). This belief is correlated with socially-progressive philosophies (Downs and Cowan, 2012, see also Nelson et al., 2013). We use the scale by Cowan and Khatchadourian (2003); see Appendix A.2.

Endorsement of RACIST BELIEFS: the beliefs which deny the existence of racial inequality, or capture resentment towards racial minorities (Poteat and Spanierman, 2012). We measure RACIST BELIEFS using items from the validated Modern Racism Scale (McConahay, 1986); see Appendix A.3.

TRADITIONALISM: the belief that one should follow established norms and traditions, and be respectful of elders, obedient, etc. In the US, these beliefs are associated with generally conservative ideologies (Johnson and Tamney, 2001; Knuckey, 2005). We use an abridged version of the TRADITIONALISM scale (Bouchard Jr. and McGue, 2003) that measures annotators’ adherence to traditional values; see Appendix A.4.

Language Purism (LINGPURISM): the belief that there is a “correct” way of using English (Jerndudd and Shapiro, 1989). Typically, this belief also involves negative reactions to non-canonical ways of using language (Sapolsky et al., 2010; DeFrank and Kahlbaugh, 2019). We created and validated a four-item LINGPURISM scale to measure this concept; see Appendix A.5.

EMPATHY: one’s tendency to see others’ perspectives and feel others’ feelings. Research in social psychology has linked higher levels of empathy to the ability and willingness of recognizing and labeling hate speech (Cowan and Khatchadourian, 2003). We measure EMPATHY using an abbreviated Interpersonal Reactivity Index (Pulos et al., 2004); see Appendix A.6.

ALTRUISM: one’s attitude of selfless concern about others’ well-being, which can move people to act when harm or injustice happens (Wagstaff, 1998; Gavriilets and Fortunato, 2014; Riar et al., 2020), including harms through hate speech (Cowan et al., 2002). We gathered the items to measure ALTRUISM with an adapted scale taken from Steg et al. (2014); see Appendix A.7.

It is worth noting that some of the above attitudes, though not all, correlate with demographics very strongly. Table 8 in Appendix A.8 details these correlations from our study.

2.3 Text Characteristics: What is considered offensive?

Not all toxic text is toxic for the same reasons. We aim to understand how characteristics of text can affect ratings of toxicity, in addition to annotator attitudes and identities. Specifically, we consider three dimensions or categories of text, based on recent work on text characteristics that tend to be over- or under-detected as toxic (Dinan et al., 2019; Sap et al., 2019; Han and Tsvetkov, 2020; Zhou et al., 2021): anti-Black language, presence of African American English (AAE) dialect markers, and vulgar language (e.g., swearwords, slurs). We distinguish between two types of vulgarity, following Zhou et al. (2021): swearwords or explicit words that do not reference identities (offensive, non-identity referring; ONI), and (re-)claimed slurs or other identity-referring vulgarity (offensive identity-referring; OI). In our analyses, we focus on ONI vulgarity unless explicitly noted.

3 Data & Study Design

We design two online studies to study the effect of annotator identities and attitudes on their toxicity ratings on posts with different characteristics. In either study, annotators are asked to rate how offensive and how racist they consider a post to be (see Appendix B for the exact questions). For both studies, we sought explicit consent from participants, paid participants above minimum wage, and obtained approval by our institution’s ethics board (IRB).
of toxicity or hate speech which previous work has shown still suffers from large annotator disagreement (Ross et al., 2017). In the sections §4–6, we report only (near-)significant associations; see Appendix E and F for all results.

### 3.1 Breadth-of-Workers Study

Our first study focuses on collecting toxicity ratings from a wide and diverse set of participants for a controlled set of posts. Shown in Table 1, we hand curated a set of 15 posts that belong exclusively to one text category (e.g., vulgar but non-AAE and non-anti-Black; see Appendix C.1 for more data selection and validation details). To exclude confounding effects of offensive identity mentions (OI; e.g., slurs) which could be both vulgar and anti-Black (or sexist, homophobic, etc.), we only considered posts with vulgar terms that are non-identity referring (ONI; e.g., swearwords).

We ran our study on a 641 participants that were recruited using a pre-qualifier survey on Amazon Mechanical Turk (MTurk) to ensure racial and political diversity. Our final participant pool spanned various racial (13% Black, 85% White), political (29% conservative, 59% liberal), and gender identities (54% women, 45% men, 1% non-binary). Each participant gave each of the 15 posts toxicity ratings, after which they answered a series of questions about their attitudes and their identity. We hand curated a set of 15 posts that belong exclusively to one text category (e.g., vulgar but non-AAE and non-anti-Black; see Appendix C.1 for more data selection and validation details). To exclude confounding effects of offensive identity mentions (OI; e.g., slurs) which could be both vulgar and anti-Black (or sexist, homophobic, etc.), we only considered posts with vulgar terms that are non-identity referring (ONI; e.g., swearwords).

We ran our study on a 641 participants that were recruited using a pre-qualifier survey on Amazon Mechanical Turk (MTurk) to ensure racial and political diversity. Our final participant pool spanned various racial (13% Black, 85% White), political (29% conservative, 59% liberal), and gender identities (54% women, 45% men, 1% non-binary). Each participant gave each of the 15 posts toxicity ratings, after which they answered a series of questions about their attitudes and their identity.

<table>
<thead>
<tr>
<th>Vulgar (ONI), not Anti-Black, not AAE</th>
<th>Anti-Black, not Vulgar, not AAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I sort of fuckin love how close my gf is with my family like that she knows about my family jokes before I fuckin do</td>
<td>I don’t get government benefits just for being white</td>
</tr>
<tr>
<td>I’m hoping that Cersei gets her ass handed to her but I love Daenerys and Jon ;3; I don’t want either of em losing xD</td>
<td>Where is white history month? Why can’t I have white pride day?</td>
</tr>
<tr>
<td>I get mosquito bites on my foot and they fucking hurt</td>
<td>I don’t like black doctors, it’s just my preference</td>
</tr>
<tr>
<td>Personally I think it would be fucking hilarious to put Franziska or Gvidot in but idk if they’re important enough characters to make the cut</td>
<td>Anti-Black racism isn’t that big of a deal because people treated Italians like this in the 60’s and 70’s</td>
</tr>
<tr>
<td>Going to Africa. Hope I don’t get AIDS. Just kidding. I’m white!</td>
<td>Going to Africa. Hope I don’t get AIDS. Just kidding. I’m white!</td>
</tr>
</tbody>
</table>

Table 1: All 15 posts for our breadth-of-workers study.

### 3.2 Breadth-of-Posts Study

Our second study focuses on collecting ratings for a larger set of posts, but with fewer annotators per post to simulate a crowdsourced dataset on toxic language. In contrast to the previous study, we consider anti-Black or AAE posts that could also be vulgar, and allow this vulgarity to cover both potentially offensive identity references (OI) as well as non-identity vulgar words (ONI; see §2.3). We do not consider posts that are anti-Black and AAE, since the pragmatic toxicity implications of anti-Black meaning expressed in AAE are very complex (e.g., in-group language with self-deprecation, sarcasm, reclaimed slurs; Greengross and Miller, 2008; Croom, 2011), and are thus beyond the scope of this study.

We draw from two existing toxic language detection corpora to select 571 posts (Table 2). For AAE and possibly vulgar posts, we draw from Founta et al. (2018), using an automatic AAE detector by Blodgett et al. (2016)\(^6\) and the vulgarity word list from Zhou et al. (2021) for detecting OI and ONI terms. For anti-Black and possibly vulgar posts, we select posts annotated as anti-Black in Vidgen et al. (2021), using the same method by Zhou et al. (2021) for detecting vulgar terms. See Appendix D.1 for more data selection details.

As with the previous study, we ran our annotation study on 173 participants recruited through a pre-qualifier survey. Our annotators varied racially.

---

\(^6\)The text-only AAE detector by Blodgett et al. (2016) strongly correlates ($r=0.60$) with more race-aware AAE detectors (Sap et al., 2019).

<table>
<thead>
<tr>
<th>Breadth-of-Workers study</th>
<th>Breadth-of-Posts study</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat.</td>
<td>Anti-Black</td>
</tr>
<tr>
<td>count</td>
<td>113</td>
</tr>
</tbody>
</table>

Table 2: Counts for each text category for the 571 posts in our breadth-of-posts study. OI: identity-referring vulgarity, ONI: non-identity referring vulgarity; categories are explained in §2.3. Posts could belong to multiple categories (Figure 5 in Appendix F).
Table 3: Associations between annotator variables and ratings of offensiveness and racism for the anti-Black posts in the breadth-of-workers study. We use the Holm correction for multiple comparisons for non-hypothesized associations and only present significant Pearson r or Cohen’s d effect sizes (*: p < 0.05, **: p < 0.001; n.s.: not significant).

(20% Black, 76% White), politically (30% conservative, 54% liberal), and in gender (45% women, 53% men, <2% non-binary). Each post was annotated by 6 participants from various racial and political identities. Additionally, we asked participants one-item versions of our attitude scales, using the question from each scale that correlated best with toxicity in our breadth-of-workers study as explained in Appendix D.3. See Appendix D for more study design details.

In our analyses, we examine toxicity of anti-Black and potentially vulgar posts (§4.2) and of AAE and potentially vulgar posts (§5.2), but not of vulgar posts separately, due to confounding effects of the AAE or anti-Black characteristics that those posts could have. Additionally, unlike the breadth-of-workers study, here each annotator could rate a varying number of posts. Thus, we compute associations between toxicity ratings and identities or attitudes using a linear mixed effects model\(^6\) with random effects for each participant.

### 4 Who finds anti-Black posts toxic, and why?

Anti-Black language denotes racially prejudiced or racist content—subtle (Breitfeller et al., 2019) or overt—which is often a desired target for toxic language detection research (Waseem, 2016; Vidgen et al., 2021). Based on prior work on linking conservative ideologies, endorsement of unrestricted speech, and racial prejudice with reduced likelihood to accept the term “hate speech” (Duckitt and Fisher, 2003; White and Crandall, 2017; Roussos and Dovidio, 2018; Elers and Jayan, 2020), we hypothesize that conservative annotators and those who score highly on the RACISTBELIEFS or FREEOFFSPEECH scales will rate anti-Black tweets as less toxic, and vice-versa. Conversely, based on findings by Cowan and Khatchadourian (2003), we hypothesize that annotators with high HARMOFHATESPEECH scores will rate anti-Black tweets as more toxic.

#### 4.1 Breadth-of-Workers Results

As shown in Table 3, we found several associations between annotator beliefs and toxicity ratings for anti-Black posts, confirming our hypotheses. The three most salient associations with lower racism ratings were annotators who scored higher in RACISTBELIEFS or FREEOFFSPEECH, and those who leaned conservative. We find similar trends for offensiveness ratings.

Conversely, we found that participants who scored higher in HARMOFHATESPEECH were much more likely to rate anti-Black posts as more offensive, and more racist. Finally, though both white and Black annotators rated these posts very high in offensiveness (with means \(\mu_{\text{Black}} = 3.85\) and \(\mu_{\text{white}} = 3.59\) out of 5), our results show that Black participants were slightly more likely than white participants to rate them as offensive.

Our exploratory analyses unearthed other significant associations: negative correlations with LINGPURISM, TRADITIONALISM, and gender (male), and positive correlations with high EMPATHY, ALTRUISM, and gender (female).
4.2 Breadth-of-Posts Results

Table 4 shows similar results as in the breadth-of-workers analyses, despite the posts now potentially containing vulgarity. Specifically, we find that annotators who scored higher in RACISTBELIEFS rated anti-Black posts as less offensive, whereas those who scored higher in HARMOFHATESPEECH rated them as more offensive. Ratings of racism showed similar effects, along with a near-significant association between higher FREEOFFSPEECH scores and lower ratings of racism for anti-Black posts.

4.3 Perceived Toxicity of Anti-Black Language

Overall, our results from both studies corroborate previous findings that studied associations between attitudes toward hate speech and gender and racial identities, specifically that conservatives, white people, and men tend to value free speech more, and that liberals, women, and non-white people perceive the harm of hate speech more (Cowan and Khatchadourian, 2003; Downs and Cowan, 2012). Our results also support the finding that those who hold generally conservative ideologies tend to be more accepting towards anti-Black or racially prejudiced content (Goldstein and Hall, 2017; Lucks, 2020; Schaffner, 2020).

In the context of toxicity annotation and detection, our findings highlight the need to consider the attitudes of annotators towards free speech, racism, and their beliefs on the harms of hate speech, for an accurate estimation of anti-Black language as toxic, offensive, or racist (e.g., by actively taking into consideration annotator ideologies; Waseem, 2016; Vidgen et al., 2021). This can be especially important given that hateful content very often targets marginalized groups and racial minorities (Silva et al., 2016; Sap et al., 2020), and can catalyze violence against them (O’Keeffe et al., 2011; Cleland, 2014).

5 Who finds AAE posts toxic, and why?

African American English (AAE) is a set of well-studied varieties or dialects of U.S. English, common among, but not limited to, African-American or Black speakers (Green, 2002; Edwards, 2004). This category has been shown to be considered “worse” English by non-AAE speakers (Hilliard, 1997; Blake and Cutler, 2003; Champion et al., 2012; Beneke and Cheatham, 2015; Rosa and Flores, 2017), and is often mistaken as obscene or toxic by humans and AI models (Spears et al., 1998; Sap et al., 2019), particularly due to dialect-specific lexical markers (e.g., words, suffixes).

Based on prior work that correlates racial prejudice with negative attitudes towards AAE (Gaither et al., 2015; Rosa, 2019), we hypothesize that annotators who are white and who score high in RACISTBELIEFS will rate AAE posts as more toxic. Additionally, since AAE can be considered non-canonical English (Sapolsky et al., 2010; DeFrank and Kahlbaugh, 2019), we hypothesize that annotators who are more conservative and who score higher in TRADITIONALISM and LINGPURISM will rate AAE posts with higher toxicity.

<table>
<thead>
<tr>
<th>AAE posts</th>
<th>Rated as Racist</th>
</tr>
</thead>
<tbody>
<tr>
<td>RACISTBELIEFS</td>
<td>r = 0.089 *</td>
</tr>
<tr>
<td>Politics (lib: 0, cons: 1)</td>
<td>r = 0.076 †</td>
</tr>
</tbody>
</table>

Table 5: Associations between ratings of racism and annotator variables, for the AAE posts from the breadth-of-workers study. As with the previous results, we correct for multiple comparisons for non-hypothesized associations and only show significant results (*: p < 0.075, †: p < 0.05).

5.1 Breadth-of-Workers Results

Table 5 shows significant associations between annotator identities and beliefs and their ratings of toxicity of AAE posts. Partially confirming our hypothesis, we found that ratings of racism had somewhat significant correlations with annotators’ conservative political leaning, and their scores on our RACISTBELIEFS scale. However, contrary to our expectations, we found that white and Black annotators did not differ in how offensive they rated AAE tweets (d = 0.14, p > 0.1). We found no additional hypothesized or exploratory associations for racism ratings, and no significant associations for offensiveness ratings.

5.2 Breadth-of-Posts Results

Shown in Table 6, our results for AAE and potentially vulgar breadth-of-posts study show higher offensiveness ratings from conservative raters, and those who scored higher in TRADITIONALISM and, almost significantly, RACISTBELIEFS. We also find that conservative annotators and those who scored higher in FREEOFFSPEECH (and near-significantly, TRADITIONALISM) rated AAE posts as more racist.
Table 6: Associations between AAE (and potentially also vulgar) post ratings from the breadth-of-posts study and annotator variables, shown as the $\beta$ coefficients from a mixed effects model with a random effect for each annotator. We only show significant results ($^\dagger$: $p<0.075$, $^*$: $p<0.05$, $^\ast\ast$: $p<0.001$; Holm-corrected for multiple comparisons; n.s.: not significant).

As an additional investigation, we measure whether attitudes or identities affects toxicity ratings of AAE posts that contain the word “n*gga,” a (reclaimed) slur that has very different pragmatic interpretations depending on speaker and listener identity (Croom, 2011). Here, we find that raters who are more conservative tended to score those posts as significantly more racist ($\beta=0.465$, $p=0.003$; corrected for multiple comparisons).

5.3 Perceived Toxicity of AAE

Our findings suggest that annotators perceive that AAE posts are associated with the Black racial identity (Rosa, 2019), which could cause those who score highly on the RACISTBELIEFS scale to annotate them as racist, potentially as a form of colorblind racism (e.g., where simply mentioning race is considered racist; Bonilla-Silva, 2006). Moreover, specific markers of AAE could have been perceived as obscene by non-AAE speakers (Spears et al., 1998), even though some of these might be reclaimed slurs (e.g., “n*gga”; Croom, 2011; Galinsky et al., 2013). Contrary to expectations, annotators’ own racial identity did not affect their ratings of AAE posts in our studies. Future work should investigate this phenomenon further, in light of the variation in perceptions of AAE within the Black community (Rahman, 2008; Johnson et al., 2022), and the increased acceptance and usage of AAE by non-Black people in social media (Ilbury, 2020; Ueland, 2020).

These findings shed light on the racial biases found in hate speech detection (Davidson et al., 2019; Sap et al., 2019), partially explaining why AAE is perceived as toxic. Based on our results, future work in toxic language detection should account for this over-estimation of AAE as racist. For example, annotators could explicitly in-

Table 7: Associations between toxicity ratings and annotator variables for the vulgar posts from the breadth-of-workers study. We correct for multiple comparisons for non-hypothesized associations and only show significant results ($^\dagger$: $p<0.05$, $^\ast\ast$: $p<0.001$).

6 Who finds vulgar posts toxic, and why?

Vulgarity can correspond to non-identity referring swearwords (e.g., f*ck, sh*t; denoted as ONI) or identity-referring slurs (e.g., b*tch, n*gga; denoted as OI). Both types of vulgarity can be mistaken for toxic despite also having non-hateful usages (e.g., to indicate emotion or social belonging; Croom, 2011; Dynel, 2012; Galinsky et al., 2013).

Given that vulgarity can be considered non-canonical or impolite language (Jay and Janschewitz, 2008; Sapolsky et al., 2010; DeFrank and Kahlbaugh, 2019), we hypothesize that annotators who score high on LINGPURISM, TRADITIONALISM, and who are more conservative will rate vulgar posts as more offensive. Importantly, here, we focus on the posts that are exclusively vulgar (ONI) from only our breadth-of-workers study, to avoid confounding effects of vulgar posts with anti-Black meaning or in AAE (both of those cases were analyzed in §4.2 and §5.2). We refer the reader to Appendix F for the results on vulgar posts in the breadth-of-posts study.

6.1 Breadth-of-Workers Results

Confirming our hypotheses, we found that offensive ratings of vulgar (ONI) posts indeed correlated with annotators’ TRADITIONALISM and LINGPURISM scores, and conservative political leaning (Table 7). We found no associations between attitudes and racism ratings for vulgar posts.
6.2 Perceived Toxicity of Vulgar Language

Our findings corroborate prior work showing how adherence to societal traditional values is often opposed to the acceptability of vulgar language (Sapolsky et al., 2010). Traditional values and conservative beliefs have been connected to finding vulgar language as a direct challenging the moral order (Jay, 2018; Sterling et al., 2020; Mudiman, 2021). Our results suggest that vulgarity is a very specific form of offensiveness that deserves special attention. Specifically, future work might consider studying the specific toxicity of individual identity-referring vulgar (OI) words, which can carry prejudiced meaning as well (e.g., slurs such as “n*gger”). Moreover, annotators across different levels of traditionalism could be considered when collecting ratings of vulgarity, especially since perceptions might vary with generational and cultural norms (Dynel, 2012).

7 Toxicity Detection System Case Study: PERSPECTIVE

Our previous findings indicated that there is strong potential for annotator identities and beliefs to affect their toxicity ratings. We are additionally interested in how this influences the behavior of toxicity detection models trained on annotated data. We present a brief case study to answer this question with the PERSPECTIVEAPI, a widely used, commercial system for toxicity detection. Appendix G provides a more in-depth description.

We investigate whether PERSPECTIVEAPI scores align with toxicity ratings from workers with specific identities or attitudes, using the 571 posts from our breadth-of-posts study. Specifically, we compare the correlations between PERSPECTIVEAPI scores and ratings from annotators, broken down by annotators with different identities (e.g., men and women) or with higher or lower scores on attitude scales (split at the mean). See Appendix G.1 for details about this methodology.

Our investigation shows that PERSPECTIVE scores can be significantly more aligned with ratings from certain identities or groups scoring higher or lower on attitude dimensions (see Table 12 in Appendix G.2). Our most salient results show that for anti-Black posts, PERSPECTIVE scores are somewhat significantly more aligned with racism ratings by annotators who score high in RACISTBELIEFS ($r_{\text{high}} = 0.29$, $r_{\text{low}} = 0.17$).

8 Discussion & Conclusion

Overall, our analyses showed that perceptions of toxicity are indeed affected by annotators’ demographic identities and beliefs (§2). We found—via a breadth-of-workers study and a breadth-of-posts study (§3)—several associations when isolating specific text characteristics: anti-Black (§4), AAE (§5), and vulgarity (§6). Finally, we showed that a popular toxicity detection system yields toxicity scores that are more aligned with raters with certain attitudes and identities than others (§7). We discuss implications of our findings below.

Variation in toxicity ratings in hate speech datasets. In our study we deliberately sought rat-
ing of perceptions of toxicity of posts by racially and politically diverse participants. However, many existing hate speech datasets instructed annotators to adhere to detailed definitions of toxicity (Davidson et al., 2017; Founta et al., 2018), and some even selected crowdworkers for their liberal ideology (Waseem, 2016; Sap et al., 2020; Vidgen et al., 2021). While those annotation setups and annotator homogeneity could cause less variation in toxicity annotations of anti-Black, AAE, and vulgar posts, there is still empirical evidence of anti-AAE racial biases in some of these datasets (Sap et al., 2019; Davidson et al., 2019).

Given the large variation in perceptions of toxicity that we showed and the implicit encoding of perspectives by toxicity models, we recommend researchers and dataset creators investigate and report annotator attitudes and demographics; researchers could collect attitude scores based on relevant social science research, perhaps in lightweight format as done in our breadth-of-posts study, and report those scores along with the dataset (e.g., in datasheets; Gebru et al., 2018).

**Contextualize toxicity predictions in social variables.** As shown in our results and previous studies (e.g., Waseem, 2016; Ross et al., 2017; Waseem et al., 2021), determining what is toxic is subjective. However, given this subjectivity, the open question remains: whose perspective should be considered when using toxicity detection models? To try answering this question, we urge researchers and practitioners to consider all stakeholders and end users on which toxicity detection systems might be deployed (e.g., through human-centered design methods; Sanders, 2002; Friedman et al., 2008; Hovy and Yang, 2021). While currently, the decision of content moderation often solely lies in the hands of the platforms, we encourage the exploration of alternative solutions (e.g., community fact checkers, digital juries; Maronikolakis et al., 2022; Gordon et al., 2022).

In general, we urge people to embrace that each design decision has socio-political implications (Green, 2020; Cambo, 2021), and encourage them to develop technologies to shift power to the targets of oppression (Blodgett et al., 2020; Kalluri, 2020; Birhane, 2021). Finally, given the increasingly essential role of online platforms in people’s daily lives (Rahman, 2017), we echo calls for policy regulating online spaces and toxicity detection algorithms (Jiang, 2020; Benesch, 2020; McGuffie and Newhouse, 2020; Gillespie et al., 2020).

**Beyond toxicity classification: modeling distributions and generating explanations.** Our findings on the subjectivity of the toxicity detection tasks suggests that standard approaches of obtaining binary (or even n-ary) labels of toxicity and averaging them into a majority vote are inadequate. Instead, researchers could consider modeling the distribution or variation in toxicity labels with respect to individual annotators (Geva et al., 2019; Fornaciari et al., 2021; Davani et al., 2021) or to specific identities or beliefs.

But, perhaps more importantly, we encourage re-thinking the toxicity detection paradigm altogether. With the goal to assist human content moderators, creating systems that explain biased implications of posts could be more helpful than opaque toxicity scores. Thus, we advocate for moving away from classification frameworks, and towards more nuanced, holistic, and explainable frameworks for inferring the desired concepts of toxicity and social biases (e.g., Social Bias Frames; Sap et al., 2020).

**Limitations and open questions.** Our work had several limitations and raised several open research questions, some of which we outline below.

First, our particular choices of attitudes and scales could affect our results; other scales (e.g., Gerdes et al., 2011, for measuring empathy) as well as other psychological variables (e.g., propensity to volunteer or to value dignity) could be studied in the context of toxicity perceptions. Additionally, the automatic AAE detector in the breadth-of-posts study could have induced data selection biases, despite being strongly correlated with race-aware dialect detection (as noted in footnote 6). Furthermore, our analysis of the attitudes encoded in the PERSPECTIVEAPI in §7 was merely a pilot study; we hope future work will explore more in-depth methods for assess model positionality.

While our study focused on racial discrimination by studying AAE and anti-Black posts, future work should explore other axes of discrimination (e.g., sexism, homophobia, ableism, etc.). Additionally, our study focused only on U.S.-centric perspectives; we hope researchers will explore variations in toxicity perceptions in other cultural contexts (e.g., variations based on caste in India).
References


Carolina Are. 2020. How Instagram’s algorithm is censoring women and vulnerable users but helping online abusers. Feminist media studies, 20(5).


Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In SIGCIS.


Martin J Burnham, Yen K Le, and Ralph L Piedmont. 2018. Who is mturk? personal characteristics and sample consistency of these online workers. Mental Health, Religion & Culture, 21(9-10).


Jamie Cleland. 2014. Racism, football fans, and online message boards: How social media has added a new dimension to racist discourse in English football. J. Sport Soc. Issues, 38(5).

D Cole. 1996. Racist speech should be protected by the constitution. Hate crimes.


Daniel M Downs and Gloria Cowan. 2012. Predicting the importance of freedom of speech and the perceived harm of hate speech. *Journal of applied social psychology*, 42(6).


Batya Friedman, Peter H Kahn, and Alan Borning. 2008. Value sensitive design and information systems. The *handbook of information and computer ethics*.


Gil Greengross and Geoffrey F Miller. 2008. Diss ing oneself versus dissing rivals: effects of status, personality, and sex on the Short-Term and Long-Term attractiveness of Self-Deprecating and Other-Deprecating humor. *Evolutionary Psychology*, 6(3).


Connor Huff and Dustin Tingley. 2015. “who are these people?” evaluating the demographic characteristics and political preferences of mturk survey respondents. Research & Politics, 2(3).


Pratyusha Kalluri. 2020. Don’t ask if artificial intelligence is good or fair, ask how it shifts power. Nature, 583(7815).


Kris McGuffie and Alex Newhouse. 2020. The radicalization risks of gpt-3 and advanced neural language models.


Michael I Norton and Samuel R Sommers. 2011. Whites see racism as a Zero-Sum game that they are now losing. Perspectives on psychological science: a journal of the Association for Psychological Science, 6(3).


Steven Pulos, Jeff Elison, and Randy Lennon. 2004. The hierarchical structure of the interpersonal reactivity index. Social behavior and personality, 32(4).


A Attitude scales

Below, we list the individual items from each of our seven attitude dimensions. In our breadth-of-workers study, we asked participants all of the items in each scale, but in the larger-scale study, we only asked the bolded items (one per scale). (R) denotes a reverse coded item.

A.1 Valuing the freedom of offensive speech (FREEOFFSPEECH).

Participants were instructed to rate their level of agreement with the item statements, using a 5-point Likert scale ranging from “not at all” (1) to “very much so” (5).

- People should face consequences for saying something offensive online. (R)
- Offensive posts and comments online should be removed from the platforms.
- There is no such a thing as hate speech, only "speech".
  - People should be allowed to say things that others may consider offensive.
  - It’s more important to protect free speech rights than to protect people’s feelings when it comes to offensive language.

This scale was taken from Cowan et al. (2002).

A.2 Perceiving the HARMOFHATESPEECH.

Participants were instructed to rate their level of agreement with the item statements, using a 5-point Likert scale ranging from “not at all” (1) to “very much so” (5).

- When someone posts something offensive towards minorities online it intimidates and casts fear in the heart of individuals who are part of the minority group.
- Offensive language encourages discrimination against minorities.
  - Offensive language is harmful to minorities.

This scale was taken from Cowan et al. (2002).

A.3 Endorsement of RACISTBELIEFS.

Participants were instructed to rate their level of agreement with the item statements, using a 5-point Likert scale ranging from “not at all” (1) to “very much so” (5).

- Discrimination against racial minorities is no longer a problem in the United States.
  - It is easy to understand the anger of racial minorities people in America. (R)
- Racial minorities are getting too demanding in their push for equal rights.
- Over the past few years, racial minorities have gotten more economically than they deserve.
- Over the past few years, the government and news media have shown more respect to racial minorities than they deserve.

These items form the validated Modern Racism Scale, created by McConahay (1986).

A.4 TRADITIONALISM.

Participants were asked: “Please tell us how important each of these is as a guiding principle in your life.” They answered each item on a 5-point Likert scale, ranging from “not at all important to me” (1) to “extremely very important to me” (5).

- Being obedient, dutiful, meeting obligations.
- Self-discipline, self-restraint, resistance to temptations.
- Honoring parents and elders, showing respect.
- Traditions and customs.

This is an abridged version of the traditionalism scale by Bouchard Jr. and McGue (2003).

A.5 Language Purism (LINGPURISM).

Participants were instructed to rate their level of agreement with the item statements, using a 5-point Likert scale ranging from “not at all” (1) to “very much so” (5).

- I dislike when people make simple grammar or spelling errors.
- It is important to master the English language properly, and not make basic spelling mistakes or misuse a common word.
- I am not afraid to correct people when they make simple grammar or spelling errors.

- There exists such a thing as good proper English.

This scale was created by the authors.

A.6 EMPATHY.

Participants were instructed to rate their level of agreement with the item statements, using a 5-point Likert scale ranging from “not at all” (1) to “very much so” (5).

- Before criticizing somebody, I try to imagine how I would feel if I were in his/her place.
- I don’t usually become sad when I see other people crying. (R)
Figure 3: Distributions of the attitude scores of the workers in the breadth-of-workers study.

Table 8: Pearson $r$ correlations between the attitude and demographic variables from participants in our breadth-of-workers study. We only show significant correlations (*: $p < 0.05$, **: $p < 0.001$), and denote non-significant correlations with "n.s.". Our demographic variables are not correlated with each other.

- When someone is feeling ‘down’ I can usually understand how they feel.
- I have tender, concerned feelings for people or groups of people less fortunate than me.

This scale is an abbreviated version of the widely used Interpersonal Reactivity Index by Pulos et al. (2004).

A.7 Altruism.

Participants were asked: “Please tell us how important each of these is as a guiding principle in your life.” They answered each item on a 5-point Likert scale, ranging from “not at all important to me” (1) to “extremely very important to me” (5).

- Social justice, correcting injustice, caring for the weak.
- Equality, equal opportunity for all.

These items are taken from the altruism part of the scale by (Steg et al., 2014).

A.8 Attitude distributions & inter-variable correlations

We plot the distributions of our breadth-of-workers participants on the seven attitude scales in Figure 3. While most attitudes have wider distributions, RacistBeliefs notably stands out as having a skewed distributions towards people scoring low on the scale.

While some attitudes may highly correlate with demographic factors at face value (e.g., Traditionalism and politically conservatism); other forms of biases may not be easily explained by demographics alone. We examine the relationship between our attitude measurements and annotator demographic identity variables. Shown in Table 8, we find strong significant correlations between several of our annotator variables.

Notably, we find that an annotator’s political orientation correlated strongly with several variables, with liberal leaning identities being as-
associated with higher scores on the EMPATHY, HARMOFHATESPEECH, and ALTRUISM scales, whereas conservative political leaning was associated with higher scores on the TRADITIONALISM, LINGPURISM, FREEOFFSPEECH and RACIST-BELIEFS scales.

B Toxicity questions

Following crowdsourcing setups in prior work (Waseem, 2016; Davidson et al., 2017; Wulczyn et al., 2017; Founta et al., 2018; Sap et al., 2019), we asked three fine-grained questions to annotators for each post in both our studies:

- “How toxic/hateful/disrespectful or offensive does this post seem to you?”
- “How much could this post be seen as toxic/hateful/disrespectful or offensive to anyone?”
- “In your opinion, how racist is this post?”

Given the high correlations between the two offensiveness variables (Pearson $r \geq .7$; Table 9), we use create an offensiveness (“off.”) score by taking the average rating given to the “to you” and “to anyone” questions. In all our analyses, we use that offensiveness score, along with the raw racism score.

Table 9: Correlations between different offensiveness questions for each tweet (all $p < 0.001$). Since offensiveness “to you” and “to anyone” are very strongly correlated, we average them into a single offensiveness score (off.).

C Small-scale controlled study details

C.1 Data Selection & Validation

We aimed to select online posts that were very indicative of each of the above characteristics (vulgar, AAE, anti-Black) but not indicative of the others, in order to tease out the effect of that category. We selected vulgar and AAE posts from a publicly available large corpus of tweets annotated for hate speech by Founta et al. (2018). For each tweet in that corpus, we detected the presence of non-identity related profanity or swearwords using the list from Zhou et al. (2021), and extracted the likelihood that the tweet is in AAE using a lexical detector by Blodgett et al. (2016). As candidates, we selected 10 vulgar tweets that have low likelihood of being AAE, and 26 tweets that have high likelihood of being AAE but contain no vulgarity. For anti-Black posts, we selected 11 candidate online posts curated by Zevallos (2017).

We ran a human validation study to verify that the candidate posts are truly indicative of their respective categories. We created an annotation scheme to collect binary ratings for two questions per post: “does it contain vulgar language”, and “is it offensive to minorities”; a post could belong to either category or neither. Each post was manually annotated by three undergraduate research assistants trained for the task. Post validation, we manually selected 5 posts per category with perfect inter-annotator agreement. Table 1 lists the final 15 posts used for our study.

C.2 Participant Recruitment

We ran our study on Amazon Mechanical Turk (MTurk), a crowdsourcing platform that is often used to collect offensiveness annotations. With the task at hand, we sought a racially and politically diverse pool of participants, which can be challenging given that MTurk workers are usually tend to be predominantly white and skew liberal (Huff and Tingley, 2015; Burnham et al., 2018; Loepp and Kelly, 2020). Therefore, we ran a pre-selection survey to collect race and political ideology of workers, noting that this presurvey could grant them access to a longer survey on free speech, hate speech, and offensiveness in language. We stopped recruiting once we reached at least 200 Black and 200 conservative participants.

C.3 Study Setup

We ran our study on the widely used survey platform Qualtrics, using an MTurk HIT to recruit and compensate participants. Participants were first

---

12Note, this study was approved by the author’s institutional review board (IRB).

13To better recruit for our pre-survey, we noted in the title that “BIPOC people and conservatives were encouraged to participate,” and also varied the title’s wording to emphasize free speech or hate speech in different recruiting rounds.

14We compensated workers $0.02-$0.03 for this presurvey.

15Participants were compensated $4.33 for the entire survey, equivalent to an average hourly compensation of $22/h.

---

1We only used the training subset of the corpus.
asked to consent to the task, then were shown instructions for annotating the 15 posts (with occasional reminders of the instructions). Then we asked participants their views on several topics using the scales described in §2.2 and §A and finally their demographics. Throughout the study, we added three attention checks to ensure the quality of responses.

Allowing only Black, white liberal, and white conservative workers to participate, we ran our survey for 4 weeks from March 10 to April 5 2021, occasionally reminding participants from our pre-survey that they could take the survey.

D Breadth-of-Posts annotation study details

D.1 Data Selection

In this study, we draw from two existing corpora of posts labeled for toxicity, hate, or offensiveness. First, we select posts that are automatically detected as AAE and/or vulgar from Founta et al. (2018), using the lexical detector of AAE by (Blodgett et al., 2016) and the vulgar wordlist by Zhou et al. (2021). Second, we select posts that are automatically detected as vulgar and/or annotated as anti-Black from Vidgen et al. (2021). Importantly, in this large-scale study, we consider posts that potentially have multiple characteristics (e.g., AAE and vulgar), and thus consider both posts with potentially offensive identity references (vulgar-OI) as well as non-identity vulgar words (vulgar-OIN). However, to circumvent potential racial biases in what is labelled as “racist” in the Vidgen et al. (2021) corpus (Sap et al., 2019; Davidson et al., 2019), we do not consider posts that are annotated as anti-Black but detected as AAE.

Given an initial set of posts from our categories, we then randomly sample up to 600 posts, stratifying by toxicity label, vulgarity, AAE, and anti-Black meaning. Our final sample contains 571 posts, as outlined in Table 2 and Figure 5.

D.2 Breadth-of-Posts Survey details

As in the breadth-of-workers study, we recruit participants using a pre-qualifying survey on MTurk. Then, we set up a second MTurk task to collect toxicity ratings, and annotator attitudes and identities. For each post, we collected two ratings from white conservative workers, two from white liberal workers, and two from Black workers. To better mirror the crowdsourcing setting and to reduce the annotator burden, we shorten the task to only ask one question per attitude (listed in §A). We also asked one attention check question to ensure data quality.

For this study, our final dataset contains 3,171 ratings from N = 173 participants.16 Our participants were 53% were men, 45% women, and <2% non-binary, identified as 76% white, 20% Black, and <4% some other race, and spanned the political spectrum from 54% liberal to 30% conservative, with 16% centrist or moderates.

D.3 Selecting Attitude Questions

In order to simplify the annotation task for annotators, we abridged the attitude scales to only one item. Using the data from the breadth-of-workers study, we select the question that best correlated with all toxicity ratings. Specifically, for each scale, we first take the tweet category with the highest correlation with toxicity (e.g., anti-Black posts for RACISTBELIEFS), and then take the item whose response scores correlated most with the toxicity rating for those posts. Those items are bolded in §A.

E Further Breadth-of-Workers Results

We show all associations between attitudes and toxicity ratings in Table 10.

Additionally, we investigate the differences in the overall toxicity ratings of anti-Black vs. AAE

16As before, we discard 255 ratings where workers failed an attention check.
Table 10: Full set of results from our analyses of the breadth-of-workers study of 15 posts, presented as Pearson $r$ or Cohen’s $d$ effect sizes, along with significance levels ($^†$: $p < 0.075$, $^∗$: $p < 0.05$, $^{**}$: $p < 0.001$). We correct for multiple comparisons for variable relationships that were exploratory (i.e., not discussed as hypotheses in §4–6).

Figure 5: Venn diagram of number of tweets in each of the categories.

G PERSPECTIVE API Case Study: Details & Results

G.1 Details

We first obtain PERSPECTIVE toxicity scores for all the posts in our breadth-of-posts study (§3.2). Then, we split workers into two different groups for each of our attitudes and identity dimensions. For attitudes and political leaning, we assign each annotator to a “high” or “low” group based on whether they scored higher or lower than the mean score on that attitude scale. For gender and race, we use binary bins for man/woman and white/black.

Then, for each attitude or identity dimension, we compute the Pearson $r$ correlation between the PERSPECTIVE score and the toxicity ratings from the high and low groups, considering posts from potentially overlapping categories (e.g., AAE and potentially vulgar posts). Finally, we compare the high and low correlations using Fisher’s $r$-to-$z$ transformation (Silver and Dunlap, 1987).

F Further Breadth-of-Posts Results

To account for the varying number of posts that each annotators could rate, we use a linear mixed effects model to compute associations between each post’s toxicity ratings and identities or attitudes. Specifically, we our linear model regresses the attitude score onto the toxicity score, with a random effect for each worker.

See Figure 5 and Table 11.
Table 11: Associations between the annotator demographic and attitude variables and their ratings of offensiveness and racism on the posts from the breadth-of-posts study. We break down the results by category, but categories are overlapping. Only significant associations ($\beta$ coefficients from a mixed effects model) are shown ($^\dagger$: $p < 0.075$, $^*$: $p < 0.05$, $^{**}$: $p < 0.001$; Holm-corrected for multiple comparisons).

### G.2 Results

See Table 12 and Figures 6–13.
Table 12: We correlated the PERSPECTIVEAPI toxicity scores with offensiveness/racism ratings by our annotators, breaking them into two bins based on their attitude scores. Then, we used Fisher’s z-to-r test to measure whether the differences in correlations between the annotators who are high/low were significant (†: $p < 0.1$, *: $p < 0.05$).

![Figure 6: PERSPECTIVEAPI and ratings of offensiveness of anti-Black tweets.](image)

![Figure 7: PERSPECTIVEAPI and ratings of racist of anti-Black tweets.](image)
Figure 8: PERSPECTIVEAPI and ratings of offensiveness of AAE tweets.

Figure 9: PERSPECTIVEAPI and ratings of racist of AAE tweets.

Figure 10: PERSPECTIVEAPI and ratings of offensiveness of vulgar-OI tweets.
Figure 11: PERSPECTIVEAPI and ratings of racist of vulgar-OI tweets.

Figure 12: PERSPECTIVEAPI and ratings of offensiveness of vulgar-OI tweets.

Figure 13: PERSPECTIVEAPI and ratings of racist of vulgar-OI tweets.