Disentangling Perceptions of Offensiveness: Cultural and Moral Correlates

Aida Mostafazadeh Davani
Google Research
Portland, USA

Mark Díaz
Google Research
New York, USA

Dylan Baker
Distributed AI Research Institute
Seattle, USA

Vinodkumar Prabhakaran
vinodkpg@google.com
Google Research
San Francisco, USA

ABSTRACT

Recent years have seen substantial investments in AI-based tools designed to detect offensive language at scale, aiming to moderate social media platforms, and ensure safety of conversational AI technologies such as ChatGPT and Bard. These efforts largely treat this task as a technical endeavor, relying on data annotated for offensiveness by a global crowd workforce, without considering crowd workers' socio-cultural backgrounds or the values their perceptions reflect. Existing research that examines systematic variations in annotators' judgments often reduces these differences to socio-demographic categories along racial, or gender dimensions, overlooking the diversity of perspectives within such groups. On the other hand, social psychology literature highlights the crucial role that both cultural and psychological factors play in human perceptions and judgments. Through a large-scale cross-cultural study of 4309 participants from 21 countries across eight cultural regions, we demonstrate substantial cross-cultural and individual moral value-based differences in interpretations of offensiveness. Our study reveals specific regions that are significantly more sensitive to offensive language. Furthermore, using the Moral Foundations Theory, we study the underlying moral values that contribute to these cross-cultural differences. Notably, we find that participants' moral values play a far more important role in shaping their perceptions of offensiveness than geo-cultural distinctions. Our investigation, using a non-monolithic framework to understand cross-cultural moral concerns, reveals crucial insights that can be extrapolated to building AI models for the pluralistic world. Our results call for more extensive consideration of diverse human moral values when deploying AI models across diverse geo-cultural contexts.

CCS CONCEPTS

• **General and reference → General conference proceedings**.

KEYWORDS

Pluralism, Value Alignment, Annotation, Subjectivity, Offensiveness

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659021

ACM Reference Format:
Aida Mostafazadeh Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. Disentangling Perceptions of Offensiveness: Cultural and Moral Correlates. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3630106.3659021

1 INTRODUCTION

As artificial intelligence (AI) technologies become more ubiquitous,[1] they are increasingly challenged to uphold societal norms and values. Among a range of concerns, including privacy [15] and disinformation [83, 91], are concerns regarding the generation of offensive and hateful content, which have been highlighted in many recent academic and governmental calls for action [18, 34, 60]. Aligning model behaviors with such societal values often relies on large-scale collection of human annotations or preferences that underpins efforts for safeguards and alignment, such as reinforcement learning using human feedback [7, 67], safety classifiers built on human labels [8], and red teaming [11, 16]. Historically, human annotators have often been treated as interchangeable units in the machine learning (ML) pipeline, with relatively little attention given to their socio-cultural backgrounds, or how their positionality shapes the labels they produce [4, 26, 63] especially when it comes to subjective labeling tasks such as assessing sentiment [25], hate speech [70], and offensive or toxic language [82], to name a few.

While more recent work has delved into annotation subjectivity, existing research on this topic is lacking in two ways. First, most of this work focuses on differences in perceptions at the social group level, often with regard to gender [19], race, ethnicity, sexual orientation [36], or domain expertise [68]. Not much work has looked into geo-cultural differences in annotator perceptions. This is especially troubling for two reasons: one, culture plays a significant role in building shared norms and values and hence might be an important determinant in perceptions of safety and offense; two, the global crowd workforce tends to be concentrated on certain geo-cultural hubs [27, 74], hence it is important to understand how biases in representation impact the values and perspectives encoded into the ML pipeline through human labels. Secondly, the existing literature on this topic often stops at demonstrating differences across rater subgroups, but does not delve deeper into understanding the underlying socio-psychological processes distinguishing those subgroups' perceptions of the task. Such a finer-grained understanding

[1] such as ChatGPT (chat.openai.com) and Gemini (gemini.google.com)

would empower meaningful interventions that go beyond treating human labels through the lens of unstable social categories [40] defined by the labelers' demographic factors, and instead see them for the values they represent.

To address these gaps, we contribute a two-pronged study in this paper that is based on a large-scale, cross-cultural, language annotation experiment [23],[2]. We asked participants from eight cultural regions, balanced across various socio-demographic groups, (1) to annotate offensiveness in language, and (2) to respond to the Moral Foundations Questionnaire [MFQ; 6, 38], designed to elicit respondents' moral reasoning along six moral foundations of *Care*, *Equality*, *Proportionality*, *Loyalty*, *Authority*, and *Purity*. We use the MFQ as it adopts a pluralistic perspective on morality and has been used in a wide variety of settings to examine group differences and cultural practices over the past two decades. Our study demonstrates that perceptions of offensiveness vary significantly across geo-cultural regions, even after controlling for gender, age, and socio-economic status. These differences remain significant regardless of whether the annotators were given a specific definition for offensiveness. Furthermore, our analysis of the MFQ responses reveals that the observed cross-cultural differences in perceptions of offensiveness are significantly mediated by annotators' moral concerns, in particular, *Care* and *Purity*, which vary across cultures. In fact, individual-level moral concerns have a more sizeable impact on their annotations than the country-level moral concerns, proving the importance of looking beyond demographic groupings.

Furthermore, we show that these patterns have real-world implications in the context of responsible AI. As a case study, we build from recent evaluations of dataset alignment (such as by Santy et al. [81]), demonstrating that a popular toxicity dataset and model preferentially align with perspectives of individuals associated with specific geo-cultural regions and moral values. The case study aligns with our primary study findings and further emphasizes the importance of understanding cultural and psychological factors in evaluating safety in AI models and content moderation. This highlights the need for culturally-informed data collection, model training and evaluation efforts. Our research further advocates for AI model alignment efforts that are informed by variations of moral values across cultures and individuals, suggesting a meaningful paradigm for value alignment beyond socio-demographic categorizations.

2 RELATED WORK

2.1 NLP in Conversational Safety

Because interactions between conversational agents and users bear similarities to text-based interactions among humans, research on harm and safety in Natural Language Processing (NLP) intuitively shares relevant approaches and goals with conversation safety. Indeed, techniques in NLP used for content moderation and offensive speech detection are frequently applied to evaluate conversational AI for safety. For example, offensive language detection has long been an active research area, originally aimed at automating online content moderation at scale [30, 93]. This research has contributed approaches for toxicity detection, such as the use of the ToxiGen dataset [41] to evaluate the LLAMA 2 [85].

[2]The dataset can be accessed at https://github.com/google-research-datasets/D3code

However, traditional NLP approaches have largely overlooked the sociotechnical factors shaping annotators' perspectives on what is offensive, such as their social experiences and the socio-cultural contexts in which they make judgments [5, 70, 77, 86, 90]. Recent calls within the NLP community emphasize the need to better understand and model social context [47], acknowledging that language interpretation in annotation tasks is deeply situated and relational [24]. This oversight is concerning given that these tools are developed and deployed as safety guardrails for conversational AI, especially with the rapid expansion of these technologies across geo-cultural contexts. Given the global deployment of these models, understanding how conversational safety varies across geo-cultural contexts is crucial.

Furthermore, the crowd-sourced workforce, essential for annotation efforts, is disproportionately made up of workers from the Global South [4, 33], raising concerns that a narrow slice of global perspectives on safety are shaping conversational AI development. For example, recent research has shown differences in annotation produced by workers based in India and the U.S. [4], highlighting the potential variations in annotations collected from different regions. In addition, Jiang et al. [51] found differences in how individuals across eight countries assessed the severity of harmful content. Building upon this line of work, we broaden the cross-cultural scope to 21 countries and explore the moral reasoning that underpins judgments of offensiveness.

2.2 Values Embedded in AI Systems

A significant body of research in the FAccT community has characterized the values embedded within ML systems and their underlying datasets. This work encompasses evaluating values within specific ML artifacts and communities [12, 32, 50], analyzing differences in how models reflect the values of diverse social groups and cultural regions [71, 77, 81], and defining and pursuing value alignment on a global scale [31, 92]. A recurring theme throughout this research is the challenges posed by sociocultural variations in values, particularly when systems are launched at a global scale.

One approach for exploring values and sociocultural difference in ML systems is through a focus on data annotators. Annotation is by no means the only avenvue through which through which values become encoded into ML, but the field's reliance on human computation for model training, fine-tuning, and testing has made it a critical site of study. For instance, Santy et al. [81] propose a framework using iteratively collected annotations to evaluate how ground truth aligns with different social groups. Indeed, the complex interplay between individual annotators, their sociocultural contexts, and the global labor dynamics of annotation creates a multifaceted web of perspectives encoded in data [63]. Studying annotator disagreement, in particular, has emerged as a method to better understand social perspectives on complex constructs, such as safety, and how they shape ML systems [4, 22, 29, 88].

Accordingly, Díaz et al. [26] emphasize the need to consider how socio-cultural norms influence who engages in annotation work, because this skew in engagement influences the cultural knowledge and values represented in data. Against the backdrop of research and discussions of AI alignment, the collection of work on both cultural and demographic differences highlights important questions

about the extent to which modeling practices may inadvertently capture specific biases of a chosen rater pool rather than some generalized notion of the tasks raters are assigned [33]. Additionally, the open-ended use cases and output space associated with generative models significantly amplify potential risks and harms compared with prior AI developments. For example, disinformation and factuality remain critical research areas for generative models [91]. Inadvertently tailoring safety guardrails to only certain subgroups' values and preferences could in effect marginalize perspectives and concerns of others [77]. Consequently, the perspectives captured in safety evaluations in this context have the potential to shape a wide array of use contexts, particularly as foundation models are used to create many derivative models.

While FAccT research increasingly investigates and documents sociocultural difference in annotation, there is relatively less attention paid to *why* observed disagreements between groups emerge. The existing work, though limited, offers insights into the complex dynamics that drive interpretation, such as works that have observed connections between toxicity annotations and annotators' political beliefs [82], as well as annotations identifying online harassment and men's adherence to social norms of masculinity [75]. Probing the dynamics underlying sociocultural difference is essential, as demographic characteristics are often mistakenly treated as inherent traits, despite their instability [40]. Indeed, critiques directed at medical prediction algorithms have taken aim at the conflation of race and racism as measured risk factors for various health outcomes, despite the measurement of race being a *proxy* for racism [62, 87]. Scholarship such as ours that goes a step deeper in characterizing social differences is needed to develop more robust understandings of sociocultural difference while avoiding erroneous conclusions that observed differences are inherent to the social categories we study.

2.3 The Role of Morality

Disagreement among annotators in subjective tasks, such as offensive language detection, has roots beyond mere differences in socio-cultural backgrounds. For instance, the intricate interplay of social media content moderation and principles of freedom of speech highlights elements of moral and political deliberation. This is particularly evident in the task of offensive language detection (instances of such discussions can be found in [9, 13, 57]). This brings added layers to systematic disagreement on notions of offensiveness, which may reflect the complexity of beliefs and values that shape perspectives and judgments within and across cultural contexts. One such nuanced layer, often not studied in AI research, is morality. Moral considerations play significant roles in how humans navigate prejudicial thoughts and behaviors [64], often manifesting in language through offensive content. The interplay between morality and group identity [73] influences many aspects of our social dynamics, including perceptions, interactions, stereotypes, and prejudices. For example, as their social identities develop, children's perceptions of prejudice are molded by moral beliefs such as fairness, inclusion, and equality [76]. On the other hand, morality assumes a vital role in shaping prejudicial beliefs, as moral values regarding group preservation can provoke extremist behaviors directed toward out-group members [45]. Moreover, research in computational social science addressing harmful language reveals a concurrent occurrence of moral sentiment alongside expressions of hatred directed at other social groups [55].

We investigate the impact of moral values, not to identifying their connection to prejudicial thoughts or actions, but with specific intent to examine their influence on human evaluations of offensiveness in language. In this way, we draw connections between annotations from culturally diverse annotators, the sociocultural norms that shape their living context, and the moral values they hold. Rather than relying on monist approaches to defining Morality, which reduce moral concerns to one specific virtue or domain (e.g., justice [58] or harm [39]), we rely on the Moral Foundations Theory [MFT; 37], a pluralistic framework for understanding universally available but contextually variable moral foundations within diverse socio-cultural context. Recent interdisciplinary research on AI alignment has relied on MFT to evaluate the value alignment of large language models with different cultures [1, 72]. However, it is important to note that MFT is not the only pluralistic framework for morality; for instance, Morality as Cooperation [20] proposes seven universal moral rules which collectively posit that "morality consists of a collection of biological and cultural solutions to the problems of cooperation recurrent in human social life" [21].

3 STUDY DESIGN

In order to study a broad range of cultural perceptions of offensiveness, we recruited 4309 participants from 21 countries, across eight geo-cultural regions, each represented by two to four countries (Table 1).[3] We discuss the reasoning behind our selection of countries and regions in more depth in Appendix A. Our final selection of countries and regions aimed to maximize cultural diversity while balancing participant access through our chosen recruitment platform (all participants were recruited through the same platform). We asked participants to (i) annotate offensiveness of social media comments from Jigsaw datasets [52, 53] , and (ii) respond to a self-report measure of moral concern using the Moral Foundations Questionnaire [MFQ-2; 6, 37].

3.1 Recruitment

Recruitment criteria account for various demographic attributes: (1) *Region of residence*: we recruited at least 500 participants from each of the eight regions with at least 100 participants per country, except for South Korea and Qatar where we managed to recruit only a smaller number of raters (See Table 1; Appendix A); (2) *Gender*: within regions, we set a maximum limit of 60% representations per region for Men and Women separately, while including options for selecting "non-binary / third gender," "prefer not to say," and "prefer to self identify" (with a textual input field). We recognize that collecting non-binary gender information is not safe for annotators in many countries, so we limited the specification of recruitment quota to binary genders to ensure consistency across countries; (3) *Age*: in each region at most 60% of participants are 18 to 30 years old and at least 15% are 50 years old or older. Table 1 provides the final distribution of participants across different demographic groups in

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

Region	Country	#	Gender				Age		
			Man	Woman	Non-binary	PNTS	18 – 30	30 – 50	50+
Arab Culture	Egypt	225	61.80%	36.40%	0.40%	1.30%	55.60%	20.90%	23.60%
	Qatar	57	64.90%	33.30%	0.00%	1.80%	63.20%	31.60%	5.30%
	UAE	234	55.60%	44.40%	0.00%	0.00%	46.20%	44.00%	9.80%
Indian Cultural Sphere	India	444	57.00%	43.00%	0.00%	0.00%	46.60%	34.20%	19.10%
	Singapore	110	50.00%	49.10%	0.90%	0.00%	27.30%	41.80%	30.90%
Latin America	Argentina	149	50.30%	47.70%	2.00%	0.00%	52.30%	34.90%	12.80%
	Brazil	237	47.30%	52.70%	0.00%	0.00%	57.40%	27.00%	15.60%
	Mexico	163	51.50%	48.50%	0.00%	0.00%	54.00%	36.80%	9.20%
North America	Canada	378	37.60%	61.90%	0.50%	0.00%	41.00%	36.80%	22.20%
	USA	173	45.10%	52.60%	1.20%	1.20%	62.40%	20.80%	16.80%
Oceania	Australia	184	39.70%	58.70%	1.60%	0.00%	25.50%	50.00%	24.50%
	New Zealand	333	39.00%	59.80%	1.20%	0.00%	34.20%	38.70%	27.00%
Sinosphere	China	176	37.50%	62.50%	0.00%	0.00%	14.20%	66.50%	19.30%
	Japan	100	70.00%	30.00%	0.00%	0.00%	13.00%	38.00%	49.00%
	South Korea	43	58.10%	41.90%	0.00%	0.00%	27.90%	48.80%	23.30%
	Vietnam	221	53.80%	41.20%	5.00%	0.00%	71.50%	23.50%	5.00%
Sub Saharan Africa	Ghana	164	67.10%	32.90%	0.00%	0.00%	74.40%	22.00%	3.70%
	Nigeria	366	54.40%	45.10%	0.30%	0.30%	54.10%	33.10%	12.80%
Western Europe	Germany	109	52.30%	46.80%	0.90%	0.00%	51.40%	24.80%	23.90%
	Netherlands	138	52.20%	45.70%	2.20%	0.00%	61.60%	21.00%	17.40%
	UK	305	40.30%	59.00%	0.70%	0.00%	38.70%	38.00%	23.30%

Table 1: Socio-demographic distribution of participants across different regions and countries. (PNTS = Prefer not to say)

each country. We further set an exclusion criterion based on *English fluency*; we only selected participants who self-reported a "high" level of proficiency in reading and writing English. We conducted this study in English, the most widely spoken language globally, to simulate common data annotation settings, in which annotators (who are no necessarily English speakers) interact with and label English textual data. Additionally, we collected participants' self-reported subjective socio-economic status [2] as a potential confounding factor with English proficiency.

3.2 Dataset and items

In order to collect textual items for participants to annotate, we selected items from Jigsaw's Toxic Comments Classification dataset [52], and the Unintended Bias in Toxicity Classification dataset [53], both of which consist of social media comments labeled for toxicity.[4] We compiled a dataset of 4554 items using three sampling strategies from the aforementioned datasets. First, 50% of the dataset consists of a random set of items likely to evoke disagreement. To measure disagreements on each item, we averaged the original dataset annotators' toxicity scores (ranging from 0, lowest toxicity, to 1, highest toxicity), selecting items with a normal distribution centered around 0.5 (indicating maximum disagreement) and a standard deviation of 0.2. Second, 40% of the dataset consists of a balanced set of items mentioning specific social group identities related to gender, sexual orientation, or religion (based on information provided in Jigsaw's raw data). Finally, 10% of the dataset consists of a balanced set of items expressing different moral sentiments, identified using a supervised moral language tagger trained on the MFTC dataset [45]. These choices aim to facilitate future research to investigate potential content-level correlates of disagreements, a topic we do not explore in this paper.

[4]A toxic comment is defined as one that is rude, disrespectful, unreasonable or otherwise somewhat likely to make a user leave a discussion or give up on sharing their perspective.

3.3 Annotation task

Each participant was tasked with rating the offensiveness of 40 items on a 5-point Likert scale (from *not offensive at all* to *extremely offensive*). Each item in the final dataset was labeled by at least three participants in each region.Half of the participants were provided with a note that defined *extremely offensive language* as *"profanity, strongly impolite, rude or vulgar language expressed with fighting or hurtful words in order to insult a targeted individual or group."* The other half labeled items based on their own understanding of offensiveness. The latter group served as a control setting of participants who are expected to lean on their individual notion of offensiveness. Participants' reliability was tested by five undeniably non-offensive control questions randomly distributed among the 40 items. Those who failed the quality control check were removed, and not counted against our final set of 4309 participants (refer to Appendix A for test items). Participants were compensated at rates above the prevalent market rates for the task (which took at most 20 minutes, with a median of 13 minutes), and in compliance with the local minimum wage regulations in their respective countries.

3.4 Moral Foundation Questionnaire

After annotation, participants completed the Moral Foundations Questionnaire [MFQ-2; 6, 37], which assesses their moral values along six dimensions; *Care*: "avoiding emotional and physical damage to another individual," *Equality*: "equal treatment and equal outcome for individuals," *Proportionality*: "individuals getting rewarded in proportion to their merit or contribution," *Authority*:

"deference toward legitimate authorities and the defense of traditions," *Loyalty*: "cooperating with ingroups and competing with outgroups," and *Purity*: "avoiding bodily and spiritual contamination and degradation" [6]. We specifically rely on the MFQ-2 [6] due to its development and validation through extensive cross-cultural assessments of moral judgments. This characteristic makes the questionnaire a reliable tool for integrating a pluralistic definition of values into AI research. The questionnaire compromises 36 statements that assess participants' priorities along each of the six foundations. We aggregate each participant's responses along each foundation, calculating a value between 1 to 5 that represent their alignment with each moral dimension.

4 STUDY 1: GEO-CULTURAL DIFFERENCES IN OFFENSIVENESS

Our first research question is whether and to what extend geo-cultural differences contribute to variations in perceiving offensiveness. As Figure 1a shows, the perceptions of offensiveness in our data varied significantly across participants from different geo-cultural regions. This trend was further confirmed by a one-way ANOVA test using a cross-classified mixed-level regression model with annotators as the first level, regions as the second level, and items as the crossed level that reported $F(7,7515) = 31.48$, $p < .001$. On average, participants from the Arab Culture ($M = 1.19$, $SD = 1.48$) and Latin America ($M = 1.13$, $SD = 1.39$) reported highest levels of offensive scores, while participants from Sinosphere ($M = 0.80$, $SD = 1.22$) and Oceania ($M = 0.80$, $SD = 1.19$) reported the lowest values of offensiveness. Pairwise comparisons of regions show significant differences between 25 out of the 28 pairs of regions (See Figure 4; Appendix A). In line with the above findings, annotations from Arab Culture and Latin America differed significantly from every other region. For three pairs of regions (Sub-Saharan Africa and Indian Culture, Western Europe and North America, Sinosphere and Oceania) the annotations are not significantly different.

To investigate whether the regional variances we observe can be explained by other factors, we also performed the mixed-level regression analysis including control variables on various levels: (1) on participant level we considered age, gender, and self-reported socio-economic status, separately. The effect of cross-regional differences holds even after accounting for annotators' age ($F(7,7515) = 28.95$, $p < .001$), gender ($F(7,7515) = 32.24$, $p < .001$), socio-economic status ($F(7,7515) = 32.42$, $p < .001$) in the regression. (2) on an item level we considered the three strategies we used for item selection (discussed in 3.2) as possible control variables. The effect of cross-regional differences holds after accounting for this categorical variable ($F(7,7515) = 31.16$, $p < .001$). (3) We further investigated whether providing a definition for offensiveness can significantly reduce the cross-regional differences. Cross-region variances are indeed lower when a definition is provided to participants ($F(7, 7515) = 15.22$ as compared to $F(7, 7515) = 18.95$). However, controlling for whether or not the definition is provided does not impact the observed results (a one-way ANOVA reported $F(7,7515) = 31.54$, $p < .001$). In other words, even when annotators are asked to label based on a particular definition, there is still significant variance between participants of different regions.

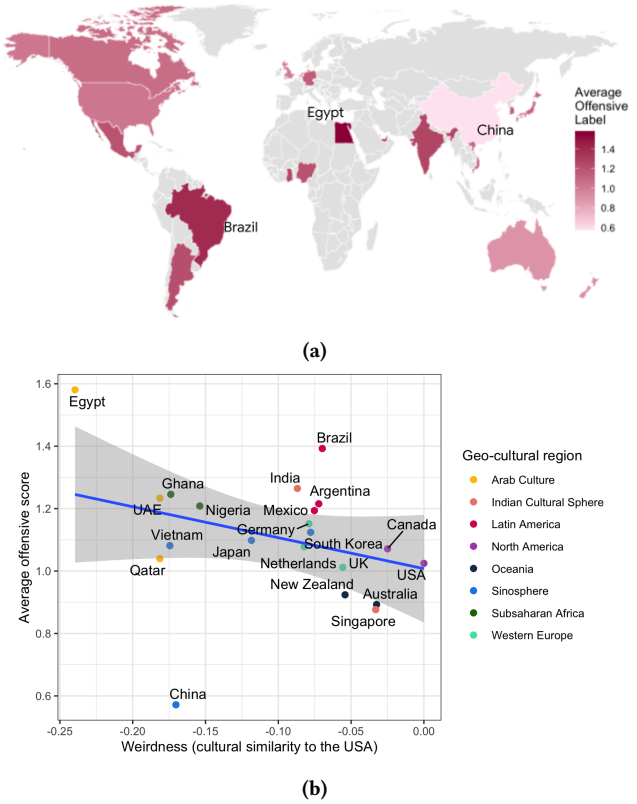

Figure 1: (a) the distribution of labels provided from different countries. Annotators from China, Brazil, and Egypt provided significantly different labels. (b) the association of cultural similarity to the USA with reported levels of offensiveness.

While our regression analysis did account for country-level factors, we further analyzed how the trends we observe vary across countries within regions. As Figure 1a and 6 show, while participants from Egypt, Brazil, India, and Argentina are more likely to report offensive language (all four reporting an average offensiveness score above 1.2 across all items), participants from China labeled substantially fewer data points as offensive (with an average offensiveness score below 0.6). This stark difference in data collected from China could be, in part, due to country-specific norms on language use, or an artifact of that specific rater pool: e.g., more than two-third of the pool from China fall within the 30-50 years old range, while this age-group represents less than half the participants for all other countries (See Table 1).[5]

We also study how our country-level results correlate with established metrics that assess cultural differences between countries. Participants from countries associated with WEIRD (Western, Educated, Industrialized, Rich, and Democratic) cultures [42] generally report fewer offensive labels (Figure 1b). A one point increase in the WEIRDness score (collected from [65] which measures cultural

[5]Because of this anomalous behaviour, we confirmed that our results hold even if we exclude data from China (which still retains over 350 participants from Sinosphere) in our region-level analysis.

similarity to the United States) results in a 0.96 point decrease in the offensiveness score (although the effect is not significant with $p = 0.18$).[6]

We also study how our results correlate with established metrics of cultural differences documented at country-level along Hofstede's six cultural dimensions [44]: *Power Distance*, *Uncertainty Avoidance*, *Individualism/Collectivism*, *Masculinity/Femininity*, *Long/Short Term Orientation*, and *Indulgence/Restraint*. While cultural tendency for seeking equality on power distribution or fulfilling one's desires did not have a significant association with annotation behavior, we find that participants from countries that score higher on uncertainty avoidance ($\beta = 3.83$, $p < .001$), individualism ($\beta = 4.80$, $p = .009$), femininity ($\beta = 2.73$, $p = .026$), and short-term orientation ($\beta = 8.57$, $p < .001$) are more likely to report higher levels of offensiveness in their annotations. These results point to the importance of looking beyond the geographical regions, and instead at finer-grained values prevalent in specific countries.

5 STUDY 2: MORAL FOUNDATIONS OF OFFENSIVENESS

While Study 1 demonstrates significant cross-cultural differences in perceptions of offensiveness across regions and countries, cultural backgrounds of participants only partly explain the observed differences. The mixed-effects model from Study 1 in fact shows that 39.5% of the variance is due to individual differences while country and region difference explain only 2.9% and 0.7% of the variance, respectively. Hence, we expand the scope of the analysis in Study 2, and include participant-level variables that have potential impact on annotation variances. In particular, we consider the assessment of whether something is offensive or not as a matter of moral judgement, and hypothesize that participants' moral concerns play an important role in their assessments. To test this, we use a *mediation analysis* approach [61] where we first examine whether participants' individual moral concerns measured through the MFQ-2, mediate the effect of geo-cultural regions we observed on their perceptions of offensiveness in language.

We examine the mediation effect of each of the six moral foundations on the effects of regions on annotation differences. We consider each mediator separately since participants' score on each moral foundation is an independent score. For each mediation test, we tested three associations: (1) direct effect of independent variable (region) on dependent variable (offensiveness): confirmed in Study 1; (2) effect of independent variable (region) on the mediator (each moral foundation): six one-way ANOVA tests report significant effects of geo-cultural region on participants six morality scores (see Table 5); (3) combined effect of the independent variable and mediators on the dependent variable: when morality score variables are separately included to the association of geo-cultural region and annotations, the results show a significant mediating effect [10] for Care ($ACME$[7] = -0.034, $p < .001$), and Purity ($ACME$ = -0.007, $p = .004$). In other words, annotators' perceptions of offensiveness that vary significantly across geo-cultural regions are significantly mediated by cultural differences in moral values regarding Care and Purity. As shown in Fig 2, an increase in annotators' Care and Purity scores leads to varying degrees of change in reported offensive scores in different regions; while Arab Culture demonstrates the highest positive change, Sinosphere demonstrates a change in the opposite direction.[8]

We further assessed the impact of individual-level moral concerns that go beyond country-level concerns regarding Care and Purity on the annotations. To this end, we us a *decomposition analysis* approach: for each annotator i from country c, we considered the moral scores (e.g., $Care_i$) as the summation of the average moral score of all annotators from their country (e.g., $Care_c$) and their deviation from the average (e.g., $Care_{i-c}$); such that:

$$Care_i = Care_c + Care_{i-c} \quad (1)$$

The results of the two decomposition analyses for Care and Purity values show that in both cases the country-level moral concerns have less significant impact on annotations compared to the individual-level values. Specifically, in a mixed-level regression analysis with $Care_i$ and $Care_c$ as independent variables and annotation labels as the dependent variable, country-level Care score does not have a significant effect ($\beta = 0.27$, $p = .087$), while the individuals' deviation from their country has a significant impact on their annotations ($\beta = 0.14$, $p < .001$). In other words a 1-point increase in participants' Care score compared to their country's average is associated with 0.14 increase in the offensive score they assign to the items. A similar trend, albeit with smaller magnitude, is observed for Purity scores where country-level Purity score does not have a significant effect ($\beta = 0.05$, $p = .530$), while the individuals' deviation from their country has a significant impact on their annotations ($\beta = 0.05$, $p < .001$). A 1-point increase in participants' Purity score compared to their country's average is associated with 0.05 increase in the offensive score they assign to the items.

The results of this study, combined with those from study 1, collectively demonstrate that while geo-cultural differences play a role in annotators' disagreements, these variations are significantly shaped by individual-level socio-psychological factors that are not typically taken into account as a part of the data collection pipelines. More importantly, although differences in moral values across cultures contribute significantly to differences in perceiving offensiveness, annotations are primarily correlated with individuals' moral values, rather than the norms of their respective countries.

6 STUDY 3: IMPLICATIONS FOR RESPONSIBLE AI

Detecting objectionable content at scale is one of the core challenges in building responsible AI that effectively and equitably serves the global community. Our data collection relied on a geographically diverse pool of annotators and the results of Study 1 and 2 demonstrated how annotators from different geo-cultural regions with different moral values can provide varying perspectives about the offensiveness of language. However, most model training efforts

[6] Removing data from China from the analysis leads to significant impact of WEIRDness on annotations ($p = 0.005$), such that a one point increase in the WEIRDness score results in 1.54 decrease.

[7] Average Causal Mediating Effect

[8] Note that we observe markedly different trends in Sinosphere, which can be attributed to data from China noted before. Rerunning the mediation analysis excluding data from China shows that the effect of regions on annotation labels remains, with the mediation impact of Care (ACME= -0.043, p < .001), and Purity (ACME= -0.015, p < .001).

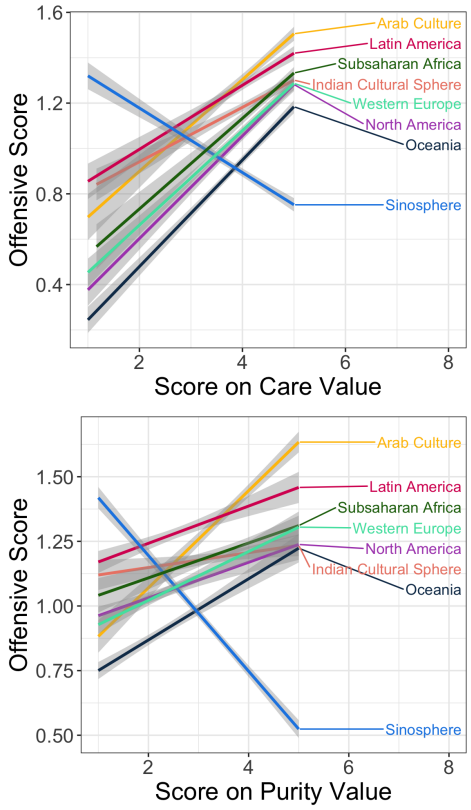

Figure 2: Annotators' scores on Care and Purity vary across regions and mediate the regional differences in annotating offensiveness. As the figures show one unit of increase in annotators' Care and Purity scores leads to varying levels of increase (or decrease in case of Sinosphere) in reported offensive scores.

are based on data annotation practices that overlook these different perspectives about the task at hand. To understand how this lack of representation impacts data practices and models, we study the correlation of our collected labels with human-annotated labels in the original dataset (Jigsaw), and model predictions of a commercial tool for offensiveness detection (Perspective API). We ask two main questions: (1) whether labels provided in the dataset and model align more closely with annotators of specific regions, and (2) whether model labels are in higher agreement with annotators with particular moral concerns.

Figure 3 presents the Pearson's correlation of labels present in the dataset (Jigsaw) as well as those provided by the model (Perspective API), with the majority vote in our data collected from different regions. Across different geo-cultural regions, we observe a weak but significant correlation between labels in the Jigsaw dataset and majority labels we obtained from Sub Saharan Africa, Latin America, and Western Europe. On the other hand, we observe the least correlation with Sinosphere and Oceania. It is important to note that the Jigsaw dataset is designed for the task of toxicity detection,[9] while our data collection that centers around offensive language, as defined distinctively for half of our participants. Therefore, the relatively low correlation between our collected annotations and Jigsaw labels should not come as a surprise.

On the other hand, Perspective API labels have relatively higher correlation with majority votes of different regions although Perspective API, similar to Jigsaw's datasets, is trained to detect toxicity rather than offensiveness. The highest correlation is observed with annotations from Latin America and Arab Culture, and lowest correlation with Sinosphere and Oceania. Both the dataset and model has the least correlation with Sinosphere and Oceania, this can both be due to the fact that model and dataset in general report high levels of toxicity scores (Jigsaw M = 0.43, SD = 0.22, Perspective API M = 0.39 , SD = 0.26 on a 0 to 1 range), while annotators from Oceania and Sinosphere are the least likely to report high levels of offensiveness (Study 1). Moreover, higher correlation with Latin America, Arab Culture, and Sub Saharan Africa are potentially due to crowdworker recruitment approaches, which largely rely on workers from the Global South. This similarity is especially important since, we assume, Perspective API's model for English language is largely used by North American websites, therefore, the low correlation between model's predictions and the majority vote of North American participants may be an area of special concern.

Furthermore, to evaluate whether labels from Perspective API align with annotators with specific moral values, we calculated its level of agreement with each annotator. A regression model with this agreement score as the dependant variable and annotators' moral values as independent variables showed that Perspective API agrees more with annotators who score high on the Care value, such that a 1-point increase in the annotator's Care score leads to 0.027 (SE = 0.004) increase in their agreement with Perspective API ($p < 0.001$). In contrast to Study 2's results, no significant association was observed for Purity scores and agreement with the Perspective API (β = -0.006, p = 0.108).

In summary, these trends suggest not only that the perceptions of offensiveness are shaped by the socio-cultural backgrounds of the perceivers, but also that the datasets and models that have become standard in AI-based endeavors to detect and mitigate potentially offensive speech do preferentially reflect certain socio-cultural value systems over others.

7 DISCUSSION

Current research on safety considerations of large language models primarily rely on crowdsourced benchmarks to evaluate potential harms [84, 89]. However, these benchmarks fail to represent the cultural and individual variations in human moral judgements about generated language and model outputs, despite evidence showing that annotators from different regions have differing perspectives regarding this task [78], . They also lack a comprehensive understanding of human values and cultural norms that drive diversity of perspectives in annotations. Our work begins to fill this gap by investigating not only the cultural differences but also the reasons for their existence. This enables us to evaluate dataset alignment

[9] defined as "a rude, disrespectful, or unreasonable comment that is likely to make one leave a discussion."

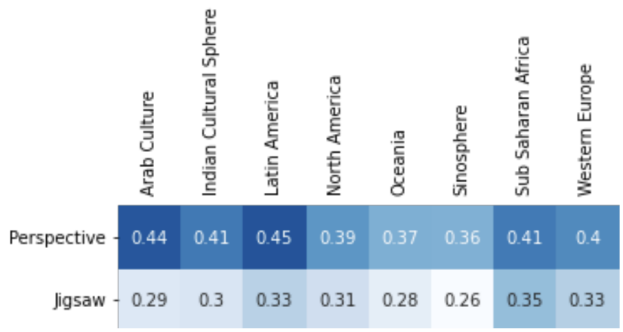

Figure 3: Correlation of labels provided in Perspective API with majority vote of annotators from different regions. Score interpretations can be summarized as: (0.00-0.20): negligible correlation, (0.20-0.40): weak correlation, and (0.40-0.60): moderate correlation.

with different perspectives while also understanding how model development mediates alignment between those perspectives and subsequent model predictions.

By conducting a cross-cultural experiment on identifying offensive language, in this paper we highlight two key factors that impact human annotations underpinning AI technologies: (1) cross-cultural differences, and (2) individual psychological differences. We conducted a data collection effort with broad geographic coverage (21 countries from eight cultural regions), having each item labeled for offensiveness by at least three annotators from each region. Our multi-level analysis of responses provides important insights into how geo-cultural factors and individual moral factors influence perceptions of offense in language.

7.1 Geo-Cultural Factors

While the motivation behind safety evaluation tasks is to align language technologies with human values through a prescriptive approach [56, 67], the key role of culture in defining humans and their values is overlooked in benchmark creation and fine-tuning efforts. More generally, performance metrics typically used to evaluate models often do not account for disagreement among system stakeholders, leading to erroneously high performance scores [35]. In fact, relatively little research in Machine Learning has investigated cultural differences in crowdsourcing pipelines for different tasks. Some examples include Chua et al. [17], who investigated how cultural differences between requesters and contributors in creative crowdsourcing tasks influence both whether individuals decide to contribute as well as how requesters judge success, and Joshi et al. [54], who studied Indian linguists in a sarcasm annotation task, finding differences in their ability to detect sarcasm compared to ground truth datasets provided by American annotators. The results of Study 1 provide strong evidence for cross regional differences on a larger scale in annotating offensiveness in language. Countries associated with Arab Culture, Latin America, Sub-Saharan Africa, and Indian Cultural Sphere, were more likely to annotate items as offensive compared to the other four regions, i.e., North America, Western Europe, Oceania, and Sinosphere; with individuals from Sinosphere, and specifically China, providing the least offensive labels. Often referred to as the Global South, the four regions with higher reports of offensiveness, have often been contrasted from the western countries in terms of their wide cultural differences. Our work demonstrates the scale of geo-cultural difference in the safety domain and complements work in social computing that focuses on harm perceptions across different populations.

In addition to driving differences in perceptions of harmful content, which we show here, cultural difference has been shown to mediate harm perceptions across other demographic groups [49]. However, in our analysis, we also consider component attributes of cultural difference, thereby avoiding treating cultural difference as intrinsic to one's origin. We investigated the impact of the six cultural factors introduced by Hofstede, on countries responses to offensive content. While cultural tendency for seeking equality on power distribution (Power-Distance Index) and cultural tendency for fulfilling one's desires (Indulgence) did not have significant association with annotation behaviors, cultures with a higher tendency to avoid uncertainty and ambiguity (Uncertainty Avoidance), more gender role differentiation (Femininity vs. Masculinity), and stronger emphasis on the present than the future (short-term orientation) were more sensitive to offensive language. These results provide finer-grained insights into values that may shape individuals' tolerance for offensive language that goes beyond regional groupings. Beyond cultural investigations, differences in annotation behavior have increasingly become a focus of study in machine learning. With the growth of sub-fields such as data-centric AI [66, 94], researchers have put forth a range of explanations for disagreements in data annotation [80], such as random variation as an artifact of human behavior [59], as well as crowd worker and label schema quality [5, 28]. Disagreements between specific social groups have also been studied, with scholars finding systematic differences in safety annotations rooted in gender [19], race and sexuality [36], community membership [68], and political views [82]. As a result, researchers have problematized the notion of a universal gold standard for offensive or toxic language, and called for dataset and modeling approaches that preserve disagreements [22, 46, 70].

Preserving disagreement is a step toward exposing the seams in majority worldviews that AI models stand to prioritize through training processes that rely on majority vote. An evaluation of cultural variation such as ours can be used to detail relative representation of global perspectives in a given dataset or distribution of model outputs. In this way, our approach reflects the motivations underlying a perspectivist approach to ML [14]. This approach explicitly aims to collect and preserve heterogenous judgments in annotation as well as preserve these difference throughout model development rather than aggregating them through majority vote or similar methods. This approach also rejects the notion that data can be objective, which is erroneously reinforced by the ways in which ground truth and benchmark datasets become cemented as standard performance metrics [48]. Considering the WEIRDness of countries (their degree of similarity from the USA culture), while not having a significant impact on provided labels, is associated with lower offensiveness labels. In other words participants from countries which are more culturally similar to the USA (this includes North America, Western European Countries, Oceania, and

some countries across other regions such as Singapore) are less sensitive to offensive content. A non-perspectivist approach risks reproducing singular viewpoints along the spectrum of WEIRDness without providing an empirical means to question the cemented, algorithmic authority of a model created with that data.

7.2 Moral Factors

As we demonstrated through out investigation of the mediating impact of moral concerns on the regional differences in offensiveness perceptions, determining the offensiveness of language is a moral judgment shaped by cultural factors. We expected individual and cultural differences in moral values (measured through the moral foundations questionnaire) to explain the observed regional and cultural differences. Our results showed that regions with higher moral values regarding Caring for others and Purity are more sensitive to offensive language. Care as a main definitive foundation discussed in different moral theories, includes "intuitions about avoiding emotional and physical damage to another individual" [6]. Care values have lower variance across regions compared to the other foundations (Table. 5), with most people scoring high on this foundation, a high score on Care helps explain higher offensiveness labels across the regions. On the other hand, the Purity foundation, with higher variance across regions, also has a positive association with high offensiveness labels. Purity refers to "intuitions about avoiding bodily and spiritual contamination and degradation" [6].

In presenting our results, we do not make the argument that annotators ought to be recruited according to some distribution of moral values. Rather, as we have discussed, research on offensive language detection typically disregards information about annotators that may help to disentangle the reasoning that underpins the judgments they provide. As such, there is more opportunity to investigate moral reasoning further, as well as how other social and moral attributes relate to annotation. The need to assess moral norms takes on even more weight in the context of large language models, which enable a scale of language production beyond that of the human-generated content typically analyzed in social media contexts. These large language models rely on fine-tuning processes that involve human-generated annotations and feedback that are inherently shaped by cultural and moral views. As a result, generative models may perform in ways that differentially align with the norms of different global populations. Indeed, Santy et al. [81] analyzed dataset ground truth labels and model outputs for a social acceptability and hate detection task, finding greater alignment with annotations provided by Western, White, college-educated, and younger individuals.

Our findings confirm that disagreements in annotating offensiveness, with key applications in evaluating AI safety, have cultural, and psychological roots often disregarded in current efforts. These findings call for culturally-informed data collection and model evaluation effort, ensuring that they reflect the values and norms in the communities affected by model deployments. It is, therefore, essential to diversify rater pools of data annotator and model evaluators to incorporate various perspectives of language to enhance modeling processes responsibly.

Moreover, inclusion of diverse opinions goes beyond data collection and expansion of rater pools; we believe that cultural considerations play a pivotal role in defining AI-related harms and devising more effective safety protocols and paradigms. At a conceptual level, Sambasivan et al. [79] speak to ML development more broadly, arguing that even the idea of algorithmic fairness must be reconceptualized as something to be locally-defined and evaluated, citing salient approaches to ML fairness that are incompatible with fairness challenges in India.

Nevertheless, it is crucial to recognize that defining AI harms according to cultural values can conflict with social equity for marginalized groups within a given society. As discussed by [43], our findings supports NLP efforts that simultaneously uphold cultural values while actively mitigating cultural biases. This is only possible with active community engagement for aligning AI models with cultural norms, along with relying on international regulatory efforts to safeguard human rights of vulnerable individuals within communities [69].

Lastly, our findings contribute a vital perspective to ongoing discourse regarding the necessity for AI models to align with human values. Considering the diverse values and perspectives present regions, cultures, and individuals, the critical question to ask is whose values should the models align with. Instead of mainly focusing on demographics, our findings suggests that aligning AI models with moral foundations that individuals and groups are concerned about potentially provides an effective and efficient approach towards value alignment efforts for the pluralistic world we live in.

8 LIMITATIONS

Our cross-cultural experiment had several limitations. Firstly, the experiment is conducted in English, as all annotation items were selected from an English dataset. Although we established an exclusion criterion for participants based on their English fluency, this exclusion imposes specific constraints on the recruitment strategy. In many countries social groups may vary on their exposure to or education in the English language, potentially affecting their representations in the study. We acknowledge that our participant pool may not be a good population representation of respective regions. However, our study's aim is slightly different: demonstrating how biases creep into the ML pipeline, as a result of existing crowdsourcing efforts for English content/data annotations relying on English speakers without accounting for the cultural differences in countries where English is not the first language. Furthermore, like any such large-scale studies, there are potentially other factors such as education and access to internet, that might influence the participant sampling in each country, and in-turn could potentially affect our results. However, we are confident that the careful recruitment criteria we used, including English proficiency, do mitigate this risk substantially. Future work should investigate these effects in native language.

We relied on the Moral Foundations Theory as the main framework for evaluating pluralistic moral concerns. This choice in selecting the definition of morality can significantly impact our findings. Other well-known frameworks could have been used instead, for instance, the Morality as Cooperation [20] approach describes morality as "a collection of cooperative rules that help us work

together, keep the peace, and promote the common good" [3] and provides seven dimensions of moral concerns, some of which overlap with the moral foundations studied in our work. Future work might study if these trends are replicated if different operationalizations of morality are used. Regarding potential downstream harms, we emphasize that our utilization of moral metrics is strictly for evaluation purposes. We do not advocate for moral value scoring as a definitive benchmark for morality, nor imply this as a sole source of assessing harms. Instead, our aim is to highlight the importance of considering diverse moral values in AI development.

9 CONCLUSION

Identifying the social factors contributing to variations in human evaluations of AI model safety is crucial for safely deploying models that can align with human values in diverse sociocultural context. However, current AI research has only scratched the surface of social information regarding human differences in answering this issue. This research takes the task of offensive language detection as a simplified example in the context of model safety problems, and explores annotator differences both in terms of their cultural affiliation, which explain their shared social norms and behaviors, as well as more nuanced values that vary within cultures. This nuanced variation in moral considerations adds a layer of complexity to understanding how annotators within cultural regions perceive and evaluate aspects such as offensive language differently in the context of AI safety. We specifically rely on the Moral Foundations Theory, a pluralistic approach for evaluating moral considerations for individuals and social groups. Our findings demonstrate that this approach offers a more comprehensive framework for understanding human disparities in perceiving offensiveness, encompassing both social-level and individual-level moral concerns. The implications of our results extend beyond the specific task of offensive language detection and have broader impacts on the evaluation of model safety and alignment. Depending on the nature of the evaluation, various social and psychological factors come into play, contributing to the variations in human perceptions and evaluations of the task at hand. It is, therefore, crucial to draw on social studies to unravel the complexities underlying individual and social values that shape perceptions of model safety and alignment.

ETHICAL CONSIDERATIONS

We study the question of offensive language detection, a question widely explored in the NLP community. Notably, we introduce a moral dimension to understanding the reasons behind annotators' disagreements on this task, and more broadly, in AI safety evaluations. Although our inquiry is rooted in social and psychological research on morality, it is crucial to acknowledge that defining morality has historically been a controversial endeavor. No single framework can include all diverse aspects shaping human and societal perceptions of what is right and wrong, and we do not intend to establish such a definition. Furthermore, our objective is not to ascribe morality for conversational agents either; instead, we examine how human moral values becomes integrated into ML pipeline through existing processes.

We collected participants' answers to the Moral Foundations Questionnaire, the collected data is anonymized in a way that participants' IDs cannot be mapped to their profiles within any recruitment frameworks. We use the moral value measurements to assess the importance of each foundation in describing social and individual differences in perceiving offensiveness, and not as a general scale to compare participants and cultures on their morality.

Our study includes an annotation task for detecting offensive language. We took specific steps to mitigate the risk of exposing participants to harmful language by: providing relevant warnings both in the consent forms and throughout the survey to inform annotators about the possibility of being exposed to harmful language, compensating participants even if they left the survey halfway. Moreover, our item selection strategy was to collect ambiguous items, which potentially means that they did not include explicitly offensive language.

ACKNOWLEDGMENTS

We would like to express our gratitude to Dasha Valter, Kathy Meier-Hellstern, Renee Shelby, and Sunipa Dev, for their support and insightful feedback on this paper. We also sincerely appreciate the valuable comments provided by Dan Jurafsky, Mohammad Atari, Jeffrey Sorensen, Nicholas Camp, Lucy Vasserman, and Camilla Mutoni Griffiths.

REFERENCES

[1] Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2023. Moral foundations of large language models. *arXiv preprint arXiv:2310.15337* (2023).
[2] Nancy E Adler, Elissa S Epel, Grace Castellazzo, and Jeannette R Ickovics. 2000. Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy, White women. *Health psychology* 19, 6 (2000), 586.
[3] Mark Alfano, Marc Cheong, and Oliver Scott Curry. 2024. Moral universals: A machine-reading analysis of 256 societies. *Heliyon* 10, 6 (2024), e25940. https://doi.org/10.1016/j.heliyon.2024.e25940
[4] Lora Aroyo, Mark Diaz, Christopher Homan, Vinodkumar Prabhakaran, Alex Taylor, and Ding Wang. 2023. The Reasonable Effectiveness of Diverse Evaluation Data. arXiv:2301.09406 [cs.HC]
[5] Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* 36, 1 (2015), 15–24.
[6] Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2023. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology* 125 (2023).
[7] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862 [cs.CL]
[8] Ananth Balashankar, Xiao Ma, Aradhana Sinha, Ahmad Beirami, Yao Qin, Jilin Chen, and Alex Beutel. 2023. Improving Few-shot Generalization of Safety Classifiers via Data Augmented Parameter-Efficient Fine-Tuning. *arXiv preprint arXiv:2310.16959* (2023).
[9] Jack M Balkin. 2017. Digital speech and democratic culture: A theory of freedom of expression for the information society. In *Law and society approaches to cyberspace*. Routledge, 325–382.
[10] Reuben M Baron and David A Kenny. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology* 51, 6 (1986), 1173.
[11] Rishabh Bhardwaj and Soujanya Poria. 2023. Red-teaming large language models using chain of utterances for safety-alignment. *arXiv preprint arXiv:2308.09662* (2023).
[12] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 173–184.

[13] Valerie C Brannon. 2019. Free speech and the regulation of social media content. *Congressional Research Service* 27 (2019).

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

[15] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*. 2633–2650.

[16] Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. 2023. Explore, Establish, Exploit: Red Teaming Language Models from Scratch. *arXiv e-prints* (2023), arXiv–2306.

[17] Roy YJ Chua, Yannig Roth, and Jean-François Lemoine. 2015. The impact of culture on creativity: How cultural tightness and cultural distance affect global innovation crowdsourcing work. *Administrative Science Quarterly* 60, 2 (2015), 189–227. https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=5607&context=lkcsb_research

[18] European Commission. 2020. The Digital Services Act: Ensuring a safe and accountable online environment. *The Digital Services Act: Ensuring a safe and accountable online environment* (2020).

[19] Gloria Cowan and Désirée Khatchadourian. 2003. Empathy, ways of knowing, and interdependence as mediators of gender differences in attitudes toward hate speech and freedom of speech. *Psychology of women quarterly* 27, 4 (2003), 300–308. https://journals.sagepub.com/doi/abs/10.1111/1471-6402.00110?journalCode=pwqa

[20] Oliver Scott Curry. 2016. Morality as cooperation: A problem-centred approach. *The evolution of morality* (2016), 27–51.

[21] Oliver Scott Curry, Daniel Austin Mullins, and Harvey Whitehouse. 2019. Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current anthropology* 60, 1 (2019), 47–69.

[22] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics* 10 (2022), 92–110. https://arxiv.org/pdf/2110.05719.pdf

[23] Aida Mostafazadeh Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling Disagreements in Data across Cultures on Offensiveness Detection and Evaluation. arXiv:2404.10857 [cs.CL]

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

[25] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14. https://dl.acm.org/doi/pdf/10.1145/3173574.3173986

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

[27] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 135–143.

[28] Anca Dumitrache. 2015. Crowdsourcing disagreement for collecting semantic annotation. In *The Semantic Web. Latest Advances and New Domains: 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31–June 4, 2015. Proceedings 12*. Springer, 701–710.

[29] Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 6715–6726.

[30] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, Vol. 12. https://ojs.aaai.org/index.php/ICWSM/article/download/14991/14841

[31] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines* 30, 3 (2020), 411–437.

[32] Marissa Gerchick, Tobi Jegede, Tarak Shah, Ana Gutierrez, Sophie Beiers, Noam Shemtov, Kath Xu, Anjana Samant, and Aaron Horowitz. 2023. The Devil is in the Details: Interrogating Values Embedded in the Allegheny Family Screening Tool. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1292–1310.

[33] Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 1161–1166. https://doi.org/10.18653/v1/D19-1107

[34] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. arXiv:2209.14375 [cs.LG]

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

[36] 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 (2022), 1–28. https://dl.acm.org/doi/pdf/10.1145/3555088

[37] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*. Vol. 47. Elsevier, 55–130. https://bpb-us-e2.wpmucdn.com/sites.uci.edu/dist/1/863/files/2020/06/Graham-et-al-2013.AESP_.pdf

[38] Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of personality and social psychology* 101, 2 (2011), 366.

[39] Kurt Gray, Chelsea Schein, and Adrian F Ward. 2014. The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General* 143, 4 (2014), 1600.

[40] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 501–512.

[41] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509* (2022).

[42] Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and brain sciences* 33, 2-3 (2010), 111. https://henrich.fas.harvard.edu/files/henrich/files/henrich_heine_norenzayan_2010-2.pdf

[43] Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and Strategies in Cross-Cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 6997–7013. https://doi.org/10.18653/v1/2022.acl-long.482

[44] Geert Hofstede. 2011. Dimensionalizing cultures: The Hofstede model in context. *Online readings in psychology and culture* 2, 1 (2011), 8.

[45] Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science* 11, 8 (2020), 1057–1071. https://journals.sagepub.com/doi/pdf/10.1177/1948550619876629

[46] Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1120–1130. https://aclanthology.org/N13-1132.pdf

[47] Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 588–602. https://aclanthology.org/2021.naacl-main.49.pdf

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

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

[50] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How different groups prioritize ethical values for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 310–323.

[51] Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PloS one* 16, 8 (2021), e0256762.

[52] Jigsaw. 2018. Toxic Comment Classification Challenge. https://www.kaggle.com/c/\jigsaw-toxic-comment-classification-challenge/data Accessed: 2021-05-01.

[53] Jigsaw. 2019. Unintended Bias in Toxicity Classification. https://www.kaggle.com/c/\jigsaw-unintended-bias-in-toxicity-classification/data Accessed: 2021-05-01.

[54] Aditya Joshi, Pushpak Bhattacharyya, Mark Carman, Jaya Saraswati, and Rajita Shukla. 2016. How do cultural differences impact the quality of sarcasm annotation?: A case study of indian annotators and american text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. 95–99. https://aclanthology.org/W16-2111.pdf

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

[56] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of Language Agents. arXiv:2103.14659 [cs.AI]

[57] Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research* 71 (2021), 431–478.

[58] Lawrence Kohlberg. 1994. Stage and sequence: The cognitive-developmental approach to socialization.. In *The first half of the chapter is a revision of a paper prepared for the Social Science Research Council, Committee on Socialization and Social Structure, Conference on Moral Development, Arden House, Nov 1963*. Garland Publishing.

[59] Klaus Krippendorff. 2008. Systematic and random disagreement and the reliability of nominal data. *Communication Methods and Measures* 2, 4 (2008), 323–338. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1211&context=asc_papers

[60] MultiMedia LLC. 2023. *FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence*. https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

[61] David MacKinnon. 2012. *Introduction to statistical mediation analysis*. Routledge.

[62] Gabriella Mayne, Ayisha Buckley, and Luwam Ghidei. 2023. Why causation matters: rethinking "race" as a risk factor. *Obstetrics & Gynecology* 142, 4 (2023), 766–771.

[63] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25.

[64] Ludwin E Molina, Linda R Tropp, and Chris Goode. 2016. Reflections on prejudice and intergroup relations. *Current Opinion in Psychology* 11 (2016), 120–124.

[65] Michael Muthukrishna, Adrian V Bell, Joseph Henrich, Cameron M Curtin, Alexander Gedranovich, Jason McInerney, and Braden Thue. 2020. Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological science* 31, 6 (2020), 678–701. https://journals.sagepub.com/doi/pdf/10.1177/0956797620916782

[66] A Ng, D Laird, and L He. 2021. Data-Centric AI Competition. DeepLearning AI.

[67] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[68] Desmond Patton, Philipp Blandfort, William Frey, Michael Gaskell, and Svebor Karaman. 2019. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. In *proceedings of the 52nd Hawaii International Conference on System Sciences*.

[69] Vinodkumar Prabhakaran, Margaret Mitchell, Timnit Gebru, and Iason Gabriel. 2022. A human rights-based approach to responsible ai. *arXiv preprint arXiv:2210.02667* (2022).

[70] 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*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 133–138. https://doi.org/10.18653/v1/2021.law-1.14

[71] Rida Qadri, Renee Shelby, Cynthia L Bennett, and Emily Denton. 2023. AI's Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 506–517.

[72] Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. *arXiv preprint arXiv:2306.01857* (2023).

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

[74] Joel Ross, Lilly Irani, M Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers? Shifting demographics in Mechanical Turk. In *CHI'10 extended abstracts on Human factors in computing systems*. 2863–2872.

[75] Jennifer D Rubin, Lindsay Blackwell, and Terri D Conley. 2020. Fragile masculinity: Men, gender, and online harassment. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[76] Adam Rutland, Melanie Killen, and Dominic Abrams. 2010. A new social-cognitive developmental perspective on prejudice: The interplay between morality and group identity. *Perspectives on Psychological Science* 5, 3 (2010), 279–291.

[77] Joni Salminen, Hind Almerekhi, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J Jansen. 2019. Online hate ratings vary by extremes: A statistical analysis. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 213–217. https://dl.acm.org/doi/pdf/10.1145/3295750.3298954

[78] Joni Salminen, Fabio Veronesi, Hind Almerekhi, Soon-Gvo Jung, and Bernard J Jansen. 2018. Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings. In *2018 fifth international conference on social networks analysis, management and security (snams)*. IEEE, 88–94. http://www.bernardjjansen.com/uploads/2/4/1/8/24188166/jansen_onlinehate2018.pdf

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

[80] Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Ježek. 2023. Why Don't You Do It Right? Analysing Annotators' Disagreement in Subjective Tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 2420–2433.

[81] Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing Design Biases of Datasets and Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 9080–9102. https://doi.org/10.18653/v1/2023.acl-long.505

[82] 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*. Association for Computational Linguistics, Seattle, United States, 5884–5906. https://doi.org/10.18653/v1/2022.naacl-main.431

[83] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. AI model GPT-3 (dis)informs us better than humans. *Science Advances* 9, 26 (2023), eadh1850. https://doi.org/10.1126/sciadv.adh1850 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.adh1850

[84] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein,

Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. arXiv:2206.04615 [cs.CL]

[85] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[86] 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 (2021), 1385–1470. https://www.jair.org/index.php/jair/article/download/12752/26751

[87] Darshali A Vyas, Leo G Eisenstein, and David S Jones. 2020. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. , 874–882 pages.

[88] Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 14523–14530.

[89] Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2022. Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models. arXiv:2111.02840 [cs.CL]

[90] Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*. 138–142. https://aclanthology.org/W16-5618.pdf

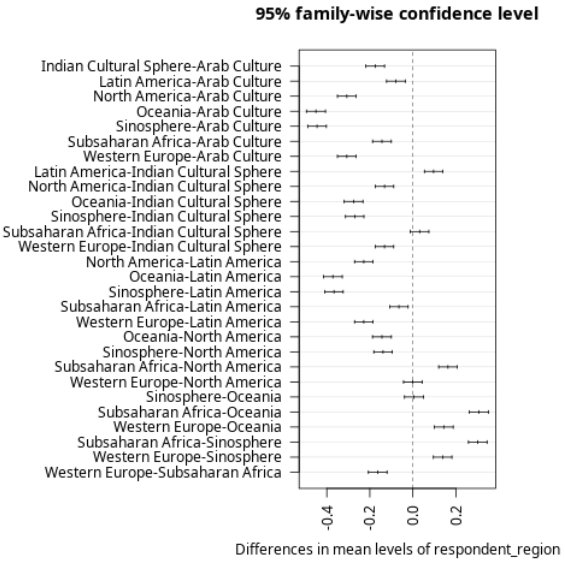

Figure 4: Pairwise differences between geo-cultural regions in their perceptions of offensiveness. 25 out of the 28 pairs of regions shows significant differences.

[91] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.

[92] Stephen Tze-Inn Wu, Daniel Demetriou, and Rudwan Ali Husain. 2023. Honor Ethics: The Challenge of Globalizing Value Alignment in AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 593–602.

[93] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*. 1391–1399. https://arxiv.org/pdf/1610.08914.pdf?_gclid=5aec59ba53a138.82841565-5aec59ba53a189.59055081&_utm_source=xakep&_utm_campaign=mention114889&_utm_medium=inline&_utm_content=lnk530117377130

[94] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, and Xia Hu. 2023. Data-centric ai: Perspectives and challenges. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, 945–948.

A APPENDIX

Participants. See Table 1 for country-level distribution of different demographic groups. While typical raters recruited through a dedicated annotation platform or service possess specialized training or experience that may influence how they complete tasks, most platforms feature a relatively limited degree of cultural representation within their pools. As a result, they are less appropriate for specifically and robustly evaluating cultural differences and their relationship to annotation judgments.

Regions and Countries. Our selected list of geo-cultural regions and countries within regions is not meant to be exhaustive, rather just to make sure that our study is done on a set of countries with diverse cultural histories. Each region listed has countries and sub-regions that have distinct cultural practices, and it is wrong to assume that the country we choose would comprehensively represent that region. Similarly, the countries listed are meant as likely places to collect data from, based on familiarity with previous

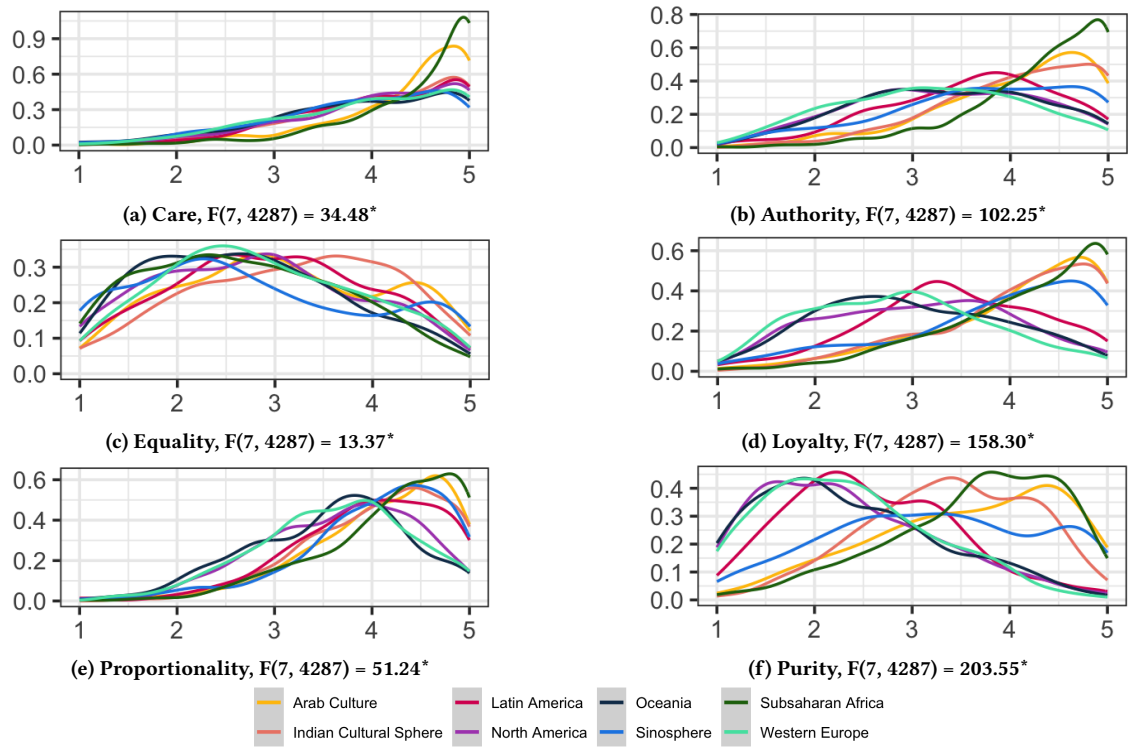

Figure 5: Distribution of the moral foundations scores and the results of one-way ANOVA analysis conducted for each moral foundation across regions. *means the p-value of the analysis is lower than .001

data collection efforts, which potentially reflect the power structures existing within those regions. Also, each country is rarely a monolith in terms of culture (e.g., India has diverse subcultures, Australia being characterized as a "Western" culture erases the vibrant Australian Aboriginal culture). Data collected would also reflect the local disparities in who tends to be in the social strata that maximally overlaps with the data-collection workforce in those respective regions, and what subcultures they represent.

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

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

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

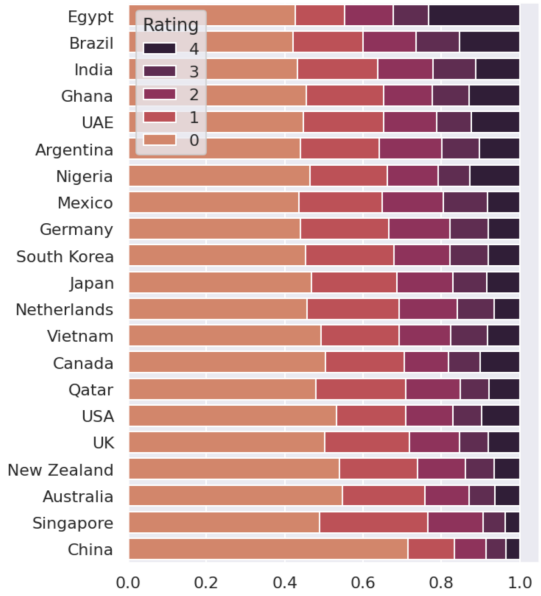

Figure 6: Distribution of the different labels provided by annotators of different countries. The y-axis is sorted based on the average offensive label captured in each country.

Mediation. We perform the mediation analysis on the annotator level by aggregating annotation labels provided from each annotator. While annotators labeled different parts of the dataset (each annotating 35 out of the 5k items) equal number of annotators from each region were assigned to label the same set. In other words for each set of 35 items, there are 3 annotators from each region that labeled the whole set.