

NLPOSITIONALITY: Measuring Design Biases and Positionality of Datasets and Models

Sebastin Santy^{†*} Jenny T. Liang^{‡*}

Ronan Le Bras[◇] Katharina Reinecke[†] Maarten Sap^{‡◇}

[†]University of Washington [‡]Carnegie Mellon University [◇]Allen AI

{ssanty, reinecke}@cs.washington.edu,

{jtliang, maartensap}@cs.cmu.edu, ronanl@allenai.org

Abstract

Design biases in NLP systems, such as performance differences for different populations, often stem from their creator’s *positionality*, i.e., views and lived experiences shaped by identity and background. Despite the prevalence and risks of design biases, they are hard to quantify because researcher, system, and dataset positionality is often unobserved. We introduce NLPOSITIONALITY, a framework for detecting design biases and measuring the positionality of NLP datasets and models. Our framework continuously collects annotations from a diverse pool of volunteer crowdworkers from around the world, and statistically measures alignment with dataset labels and model predictions. We apply NLPOSITIONALITY to existing datasets and models from two tasks—social acceptability and toxicity detection. To date, we have collected 10,991 annotations in over a year for 600 instances from 731 annotators across 78 countries. We find that datasets and models align predominantly with Western and college-educated populations. Additionally, certain groups, such as non-binary people and non-native English speakers, are further marginalized by datasets and models as they rank least in alignment across all tasks. Finally, we draw from prior literature to discuss how researchers can examine their own positionality and that of their datasets and models, opening the door for more inclusive NLP systems.

1 Introduction

“Treating different things the same can generate as much inequality as treating the same things differently.” – *Kimberlé Crenshaw*

When creating NLP datasets and models, researchers’ design choices are partly influenced by their *positionality*, i.e., their views shaped by their lived experience, identity, culture, and background (Savin-Baden and Howell-Major, 2013).

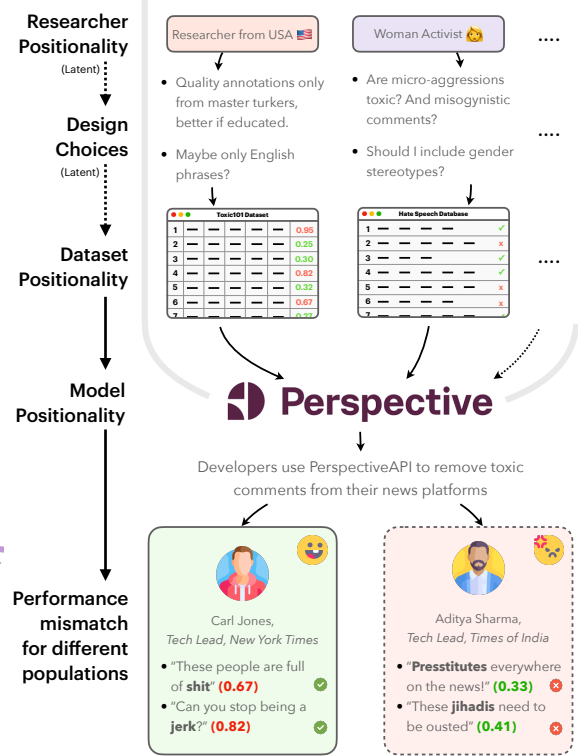


Figure 1: Carl from U.S. and Aditya from India both want to use Perspective API, but it works better for Carl than Aditya. This is because toxicity researchers’ positionalities lead them to make design choices that make toxicity datasets, and thus Perspective API, to have positionalities that are Western-centric.

While researcher positionality is commonly discussed outside of NLP, it is highly applicable to NLP research but remains largely overlooked. For example, a U.S.-born English-speaking researcher building a toxicity detection system will likely start with U.S.-centric English statements to annotate for toxicity. This can cause the tool to work poorly for other populations (e.g., not detect offensive terms like “presstitute” in Indian contexts (see Figure 1)).

Such *design biases* in the creation of datasets and models, i.e., disparities in how well datasets and models work for different populations, due to latent

design choices and the researcher’s positionality, can perpetuate systemic inequalities by imposing one group’s standards onto the rest of the world (Blasi et al., 2022; Gururangan et al., 2022; Ghosh et al., 2021). The challenge is design biases arise from the myriad of design choices made. In context of creating datasets and models, only some of these choices may be documented (e.g., through model cards and data sheets; Bender and Friedman, 2018; Mitchell et al., 2019; Gebru et al., 2021a). Further, many popular deployed models are hidden behind APIs, and thus design biases are best detected by observing model behavior.

To address this challenge, we introduce NLPOSITIONALITY, a framework for measuring design biases and positionality of NLP datasets and models. Our approach relies on recruiting a diverse pool of crowdworkers from various countries and of different backgrounds to re-annotate a sample of a dataset for a given task. We host these annotation tasks on the volunteer-based LabintheWild platform (Reinecke and Gajos, 2015), which has a larger and more diverse participant pool compared to existing crowdsourcing platforms due to its lower barrier to entry. We then detect design biases by comparing which identities and backgrounds have higher agreement with the original dataset labels or model predictions.

NLPOSITIONALITY offers three advantages over other approaches (e.g., paid crowdsourcing or laboratory studies). First, our approach is dataset- and model-agnostic and can be applied post-hoc to any dataset or model using only instances and their labels or predictions. Second, the demographic diversity of participants on LabintheWild is better than on other crowdsourcing platforms (Reinecke and Gajos, 2015) and what is possible with traditional in-lab settings. Third, the compensation and incentives in our approach relies on a participant’s motivation to learn about themselves instead of monetary compensation. This has been shown to result in higher data quality compared to using paid crowdsourcing platforms (August and Reinecke, 2019), as well as in opportunities for participant learning (Oliveira et al., 2017). This allows our framework to *continuously collect* new annotations and reflect more up-to-date measurements of design biases for free over long periods of time, compared to one-time paid studies such as in previous works (Sap et al., 2022; Davani et al., 2022).

We apply NLPOSITIONALITY to two case stud-

ies on NLP tasks, social acceptability and hate speech detection, which are likely to exhibit design biases (Talat et al., 2022; Sap et al., 2022; Ghosh et al., 2021). To date, a total of 10,991 annotations were collected from 731 annotators from 78 countries, with an average of 37 annotations per day. We discover that the datasets and models we investigate are most compatible with the White and educated people from English-speaking countries, which are a subset of "WEIRD" (Western, Educated, Industrialized, Rich, Democratic; Henrich et al., 2010) populations. We also see that datasets exhibit close alignment with their annotators, emphasizing the importance of gathering data and annotations from diverse demographics.

Our paper highlights the importance of considering design biases in NLP. Our findings showcase the usefulness of our framework in measuring dataset and model positionality. In a discussion of the implications of our results, we consider how positionality may manifest in other NLP tasks.

2 Dataset & Model Positionality: Definitions and Background

A person’s positionality is the perspectives they hold as a result of their demographics, identity, and life experiences (Holmes, 2020; Savin-Baden and Howell-Major, 2013). For researchers, positionality “reflects the position that [they have] chosen to adopt within a given research study” (Savin-Baden and Howell-Major, 2013). It influences the research process and its outcomes and results (Rowe, 2014). Some aspects of positionality, such as gender, race, skin color, and nationality, are culturally ascribed and part of one’s identity. Others, such as political views and life history, are more subjective (Holmes, 2020; Foote and Gau Bartell, 2011).

Dataset and Model Positionality While positionality is often attributed to a person, in this work, we focus on *dataset and model positionality*. Cambo and Gergle (2022) introduced model positionality, defining it as “the social and cultural position of a model with regard to the stakeholders with which it interfaces.” We extend this definition to add that, like people, datasets and models also encode positionality. This results in perspectives embedded within language technologies, making them less inclusive towards certain populations.

Design Biases In NLP, design biases occur when a researcher or practitioner makes design choices—

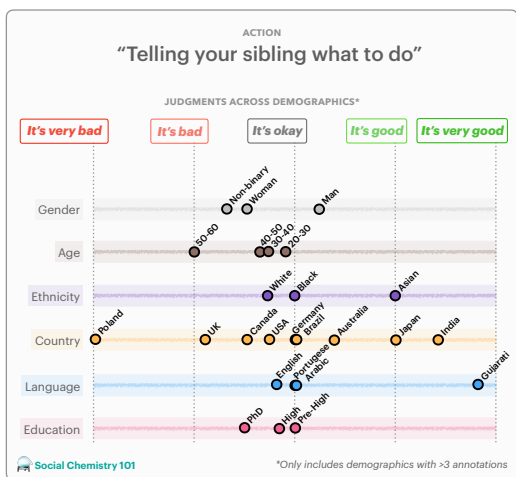


Figure 2: An example instance from Social Chemistry that was sent to LabintheWild along with the received mean of annotations across demographics.

often based on their positionality—that causes models and datasets to systematically work better for some populations over others. Curating datasets involves design choices such as what source to use, what language to use, what perspectives to include or exclude, or who to get annotations from. When training models, these choices include the type of training data, data pre-processing techniques, or the objective function (Hall et al., 2022).

Current discussions around bias in NLP often focus on ones that originate from social biases embedded within the data. In comparison, design biases originate from the developer who makes assumptions. Based on (Friedman and Nissenbaum, 1996)’s framework on bias, social biases are pre-existing biases in society, whereas design biases are emergent biases that originate from the computing system itself. ‘Gender bias’ in computing systems means that the system does not perform well for some genders; "man is to doctor as woman is to nurse" (Bolukbasi et al., 2016) is a social bias, while captioning systems failing to understand women’s voices (Tatman, 2017) is a design bias.

One prominent example of design bias in NLP is the overt emphasis on English (Joshi et al., 2020; Blasi et al., 2022). Others include the use of block lists in dataset creation or toxicity classifiers as a filter, which can marginalize minority voices (Dodge et al., 2021a; Xu et al., 2021). We extend the discussion of design biases from prior work, connect it with researcher positionality, and show its effects on datasets and models.

3 NLPOSITIONALITY: Measuring Dataset and Model Positionality

NLPOSITIONALITY follows a two-step process for identifying the positionality of datasets and models. First, a subset of data for a task is re-annotated by annotators from around the world to obtain globally representative data in order to measure positionality (§3.1). We specifically rely on re-annotation to capture self-reported demographic data of annotators with each label. Then, the positionality of the dataset or model is computed by comparing the responses of the dataset or model with different demographic groups for identical instances (§3.2).

3.1 Collecting Diverse Annotations

Cost effectively collecting annotations from a diverse crowd at scale is challenging. Popular crowdsourcing platforms like Amazon Mechanical Turk (MTurk) are not culturally diverse, as a majority of workers are from the United States and India (Difallah et al., 2018; Ipeirotis, 2010). Further, MTurk does not allow for continuous and longitudinal data collection. To address these challenges, we used LabintheWild (Reinecke and Gajos, 2015), which hosts web-based online experiments. Compared to traditional laboratory settings, it has more diverse participants and collects equally high quality data for free (August and Reinecke, 2019; Oliveira et al., 2017). Participants join LabintheWild because they learn something about themselves through the experiment. Thus, we design our annotation study so that participants learn how their responses compared to an AI and others demographically similar to them for the task (see Appendix B.1).

For a given task, we choose a dataset to be annotated. To select instances for re-annotation, we filter the dataset based on relevant information that could indicate subjectivity, such as "Controversiality", and then sample 300 diverse instances by stratified sampling across different metadata (see Appendix A.1). These instances are then hosted as an experiment on LabintheWild to be annotated by a diverse crowd, where participants report their demographics. For data quality, the instructions and annotations are similar to the original task’s. Figure 2 is an example from the Social Chemistry dataset and its annotations.

3.2 Computing Positionality

We use correlation as a measure for positionality. First, we group the annotations by specific demo-

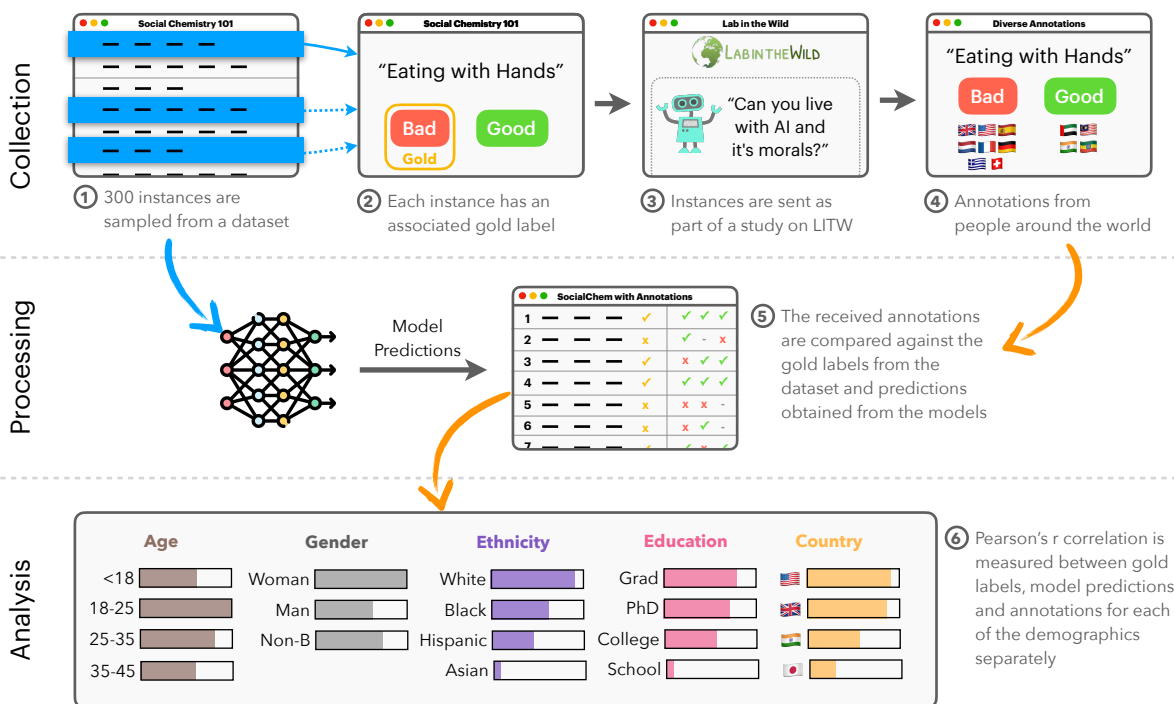


Figure 3: Overview of the NLPOSITIONALITY framework.

graphics. When datasets contain multiple annotations from the same demographic for the same instance, we take the mean of the labels from demographic annotators to obtain an aggregated score (see Table 1). Next, for each demographic, we computed Pearson’s r using the demographic’s aggregated score for each instance and correlated it to the dataset label or model prediction. Finally, we apply the Bonferroni stepwise correction to account for multiple hypotheses testing (Wickens and Keppel, 2004). We use the rank of the correlations to reveal the positionality of datasets and models.

We report the standard deviation (σ) in Pearson’s r for each demographic category to identify skews in their correlations, thus indicating the presence of design biases for the category. We also report the total number of annotators and inter-annotator agreement for each demographic using Krippendorff’s α (Krippendorff, 2006).

4 Case Studies

We present case studies of applying NLPOSITIONALITY to two different socially-aligned tasks: social acceptability and hate speech detection.

4.1 Social Acceptability

Task Description Social acceptability is the task of evaluating how widely accepted certain actions

are in society. Related to making moral judgments, it requires models to have commonsense rules about social norms and acceptable behavior (Forbes et al., 2020). Social acceptability is likely to vary by background, as morality (Atari et al., 2022), values (Haerpfer and Kizilova, 2012), and moral decision-making (Awad et al., 2018, 2020) vary by culture. Table 1 shows men are more likely than women to say it’s okay to tell a woman what to do. In NLP, social acceptability models have values (Arora et al., 2022) and biases (Hämmerl et al., 2022), some of which are Western-centric (Jiang et al., 2021; Talat et al., 2022).

Setup 459 participants annotated the social acceptability of 300 examples from the Social Chemistry dataset (Forbes et al., 2020). We collected 8,131 annotations for this task and compared participants’ responses to the original Social Chemistry dataset and the Delphi model (Jiang et al., 2021).

Results We report the results in Table 2. Social Chemistry is most aligned with people who grow up ($r = 0.77$) and live in ($r = 0.75$) English-speaking countries, who have college education ($r = 0.71$), and are White ($r = 0.71$), indicating a preference to WEIRD populations.

Delphi also exhibits a similar pattern, but to a lesser degree. While it strongly aligns with people

	Instance	Demographic	Label	Mean Score	Variance
Social Acceptability	<i>Telling a woman what to do.</i>	Man	It's okay	-0.38	0.55
		Non-binary	It's okay	-0.33	4.33
		Woman	It's bad	-1.33	0.27
	<i>Questioning your sexual orientation.</i>	10-20 years old	It's good	0.88	0.70
		20-30 years old	It's okay	0.14	1.48
		30-40, 40-50 years old	It's okay	0.00	0.00
	<i>Going to church to practice your religion.</i>	Christian	It's good	1.33	1.33
		Buddhist	It's okay	0.00	0.00
		None	It's okay	0.05	0.50
Hate Speech	<i>Justice for yellow man lol</i>	Asian, Black, Latino/Latina	Hate speech	1.00	0.00
		White	Not sure	0.33	1.33
	<i>Iran's Supreme Court claims gender equality is 'Zionist plot' aiming to corrupt role of women in society. I fucking second that!</i>	Christian	Not sure	0.50	0.92
		Jewish	Hate speech	1.00	0.00
		Muslim	Not hate speech	-1.00	0.00
	<i>Literally anyone can be a shitty person, LGBT people especially!</i>	College, Graduate school	Hate speech	1.00	0.00
		Pre-high school	Not hate speech	-1.00	0.00

Table 1: Examples where there was disagreement within and between demographics on what the label should be from LabintheWild.

who grow up ($r = 0.62$) and live in ($r = 0.63$) English-speaking countries, who have college education ($r = 0.65$), and are White ($r = 0.60$), it also correlates more with other populations, such as people who grew up ($r = 0.71$) and live in ($r = 0.62$) in Confucian countries.

Native English speakers strongly correlate with both datasets and models. Non-binary people align less to both Social Chemistry and Delphi compared to men and women. People with pre-high school education and Black, Latinx, and Native American populations consistently rank least in correlation for education level and ethnicity. Finally, the country the participants grew up in have the largest standard deviation in correlation scores in the task.

4.2 Hate Speech Detection

Task Description Hate speech detection is the task of determining whether an instance is "abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation" (Warner and Hirschberg, 2012). Rating whether text is hate speech is likely to vary by background. In Table 1, annotators who were racial minorities rated the instance "*Justice for the yellow man lol*" as hate speech, while White annotators voted not sure. Also, annotators rate the toxicity of tweets with African American English (AAE) differently by background, such as political affiliation (Sap et al., 2022). Hate speech classifiers also produce errors for instances outside Western

contexts (Ghosh et al., 2021).

Setup For the hate speech detection task, 272 LabintheWild participants annotated 300 examples from the Dynahate dataset (Vidgen et al., 2021) and rated whether an instance was hate speech or not. We collected 2,860 annotations and compared participants' responses to the original Dynahate dataset as well as Perspective API¹, Rewire API², and HateRoBERTa (Hartvigsen et al., 2022).

Results We report the results in Table 2. Dynahate is highly correlated with people who grow up in English-speaking countries ($r = 0.67$), who have college education ($r = 0.61$), and are White ($r = 0.66$). However, it also has high alignment with other populations, such as people who live in West South Asia ($r = 0.80$).

Perspective API is also similar to WEIRD populations, though to a lesser degree than Dynahate. Perspective API exhibits moderate alignment with people who grow up and live in English-speaking ($r = 0.37$, $r = 0.33$ respectively) and Orthodox European countries ($r = 0.37$, $r = 0.47$ respectively), have college education ($r = 0.40$), and are White ($r = 0.33$). However, it also exhibits higher alignment with other populations, such as Pacific Islanders and Native Australians ($r = 0.86$).

Rewire API also is similar to Western and educated populations. It has moderate correlation

¹<https://perspectiveapi.com/>

²<https://rewire.online/>

Datasets: Social Chemistry 101 DynaHate			Models: DynaHate Delphi Perspective API Rewire API HateRoberta							
Demographic	Pearson's r									
	Social Acceptability				Toxicity & Hate Speech					
	#	α			#	α				
Country (Lived Longest)										
African Islamic	18	0.28	0.47*	0.45	16	0.22	0.18	0.35	0.42*	0.43*
Baltic	9	0.27	0.73*	0.67*	4	0.44	0.47	-0.06	0.31	0.03
Catholic Europe	17	0.32	0.59*	0.54*	15	0.34	0.37	0.12	0.34	0.18
Confucian	7	0.28	0.66*	0.71*	7	0.21	0.32	0.24	0.38	0.09
English-Speaking	320	0.49	0.77*	0.62*	188	0.39	0.67*	0.37*	0.53*	0.39*
Latin America	12	0.44	0.40	0.38	4	0.74	0.49	0.03	0.22	0.33
Orthodox Europe	24	0.41	0.55*	0.57*	5	0.33	0.36	0.37	0.52	0.21
Protestant Europe	31	0.46	0.62*	0.53*	19	0.39	0.45*	0.32	0.28	0.30
West South Asia	21	0.35	0.50*	0.53*	6	0.05	0.50	0.32	0.45	0.34
σ			0.11	0.10			0.13	0.15	0.10	0.13
Education Level										
College	160	0.45	0.71*	0.65*	104	0.37	0.61*	0.40*	0.50*	0.40*
Graduate School	43	0.50	0.71*	0.53*	31	0.28	0.53*	0.27*	0.47*	0.26*
High School	99	0.48	0.63*	0.52*	48	0.42	0.52*	0.21	0.42*	0.26*
PhD	27	0.40	0.66*	0.52*	15	0.29	0.46*	0.22	0.37*	0.30
Pre-High School	14	0.44	0.45*	0.39*	5	0.17	0.38	0.14	0.34	0.21
Professional School	26	0.46	0.52*	0.46*	13	-0.14	0.62*	0.04	0.32*	0.01
σ			0.10	0.08			0.08	0.11	0.07	0.12
Ethnicity										
Asian, Asian American	39	0.57	0.63*	0.53*	22	0.34	0.53*	0.39*	0.43*	0.32*
Black, African American	19	0.52	0.61*	0.48*	15	0.31	0.56*	0.28	0.38*	0.30*
Latino / Latina, Hispanic	13	0.51	0.55*	0.45*	13	0.23	0.39*	0.29	0.42*	0.25
Native American, Alaskan Native	4	0.59	0.55*	0.50*	5	—	0.29	0.30	0.34	0.31
Pacific Islander, Native Australian	2	0.00	0.65*	0.63*	1	—	0.23	0.86*	0.31	0.62
White	124	0.51	0.71*	0.60*	86	0.43	0.66*	0.33*	0.54*	0.38*
σ			0.06	0.06			0.15	0.21	0.07	0.12
Gender										
Man	180	0.43	0.71*	0.61*	85	0.40	0.57*	0.32*	0.48*	0.36*
Non-Binary	39	0.41	0.56*	0.51*	14	0.48	0.46*	0.14	0.33*	0.22
Woman	150	0.55	0.73*	0.62*	111	0.42	0.64*	0.34*	0.52*	0.32*
σ			0.08	0.05			0.07	0.09	0.08	0.06
Native Language										
English	279	0.48	0.73*	0.63*	161	0.40	0.67*	0.34*	0.52*	0.35*
Not English	95	0.39	0.56*	0.47*	59	0.27	0.41*	0.33*	0.34*	0.33*
σ			0.08	0.08			0.13	0.01	0.09	0.01
Age										
10-20 yrs old	136	0.48	0.66*	0.55*	75	0.33	0.59*	0.28*	0.49*	0.36*
20-30 yrs old	146	0.44	0.70*	0.60*	99	0.40	0.64*	0.37*	0.52*	0.35*
30-40 yrs old	46	0.46	0.62*	0.54*	20	0.01	0.44*	0.13	0.43*	0.31*
40-50 yrs old	24	0.50	0.57*	0.49*	13	0.24	0.60*	0.33*	0.51*	0.37*
50-60 yrs old	12	0.49	0.71*	0.53*	10	0.48	0.46*	0.30	0.40*	0.33
60-70 yrs old	9	0.4	0.67*	0.53*	1	—	1.00*	0.55	0.65	0.16
70-80 yrs old	3	—	0.56*	0.52*	2	—	0.50	0.35	0.36	0.24
80+ yrs old	1	—	0.52	0.48	1	—	0.63	0.01	0.45	-0.09
σ			0.07	0.03			0.16	0.15	0.08	0.15
Country (Residence)										
African Islamic	11	0.31	0.43	0.48	7	0.25	0.27	0.34	0.34	0.23
Baltic	3	0.00	0.60*	0.60*	2	0.00	0.42	0.14	0.52	0.35
Catholic Europe	17	0.31	0.52*	0.40*	13	0.42	0.26	0.16	0.30	0.16
Confucian	2	0.63	0.55	0.62	6	0.22	0.27	0.34	0.38	0.40
English Speaking	277	0.47	0.75*	0.63*	175	0.38	0.66*	0.33*	0.51*	0.36*
Latin America	9	0.41	0.55*	0.57*	3	1.00	0.28	0.15	0.24	0.30
Orthodox Europe	16	0.37	0.46*	0.53*	3	0.37	0.21	0.47	0.32	0.28
Protestant Europe	31	0.45	0.64*	0.58*	19	0.34	0.54*	0.34*	0.37*	0.35*
West South Asia	8	0.25	0.54*	0.41	1	—	0.80	0.23	0.54	0.31
σ			0.09	0.08			0.20	0.11	0.10	0.07
Religion										
Buddhist	5	0.25	0.58*	0.48*	4	0.44	0.60	0.12	0.37	0.11
Christian	78	0.50	0.68*	0.50*	55	0.21	0.50*	0.38*	0.44*	0.38*
Hindu	6	0.86	0.56*	0.58*	3	—	0.59*	0.47	0.45	0.28
Jewish	11	0.50	0.66*	0.60*	10	0.37	0.61*	0.28	0.43*	0.29
Muslim	16	0.31	0.67*	0.58*	5	0.35	0.41	0.25	0.38	0.27
Spiritual	4	0.48	0.61*	0.60*	1	—	0.31	-0.20	0.08	0.13
σ			0.05	0.05			0.11	0.22	0.13	0.09

(# = Num. of annotators, α = Krippendorff's alpha). (σ = Standard Deviation X = Maximum, X = Minimum) per demographic type per artifact.

(# = Num. of annotators, α = Krippendorff's alpha). (σ = Standard Deviation **X** = Maximum, **X** = Minimum) per demographic type per artifact.

Table 2: Positionality of Datasets and Models measured through Pearson's r correlation.

with people who grew up in Orthodox European countries ($r = 0.52$), have a college education ($r = 0.50$), and are White ($r = 0.54$). It shows stronger correlation with populations who live in West South Asia ($r = 0.54$).

A WEIRD bias is also shown in HateRoBERTa. HateRoBERTa shows moderate alignment with people who grow up ($r = 0.39$) and live in ($r = 0.36$) English-speaking countries, have college education ($r = 0.40$), and are White ($r = 0.38$). However, it shows stronger alignment to populations who grew up in African-Islamic countries ($r = 0.43$) and who reside in Confucian countries ($r = 0.4$).

As in the previous task, native English speakers are strongly correlated with datasets and models. Non-binary people align less to Dynahate and Perspective API compared to other genders. People who grew up in Catholic European countries exhibit less correlation for the task compared to people who grew up in other countries. People who did not complete high school or are Black, Latinx, and Native American rank least in alignment for education and ethnicity respectively. Also, datasets and models differ in which demographic category has the largest standard deviation—it is the country of residence for Dynahate and age and ethnicity for HateRoBERTa.

5 Discussion

In this paper, we identify design biases in NLP and positionality of datasets and models. We introduce the NLPOSITIONALITY framework for detecting them in NLP datasets and models. NLPOSITIONALITY consists of a two-step process of collecting annotations from diverse annotators for a specific task and then computing the alignment of the annotations to dataset labels and model predictions using Pearson’s r . We applied NLPOSITIONALITY to two tasks: social acceptability and hate speech detection, with two datasets and four models in total. In this section, we discuss key takeaways from our experiments and offer recommendations for addressing design biases in datasets and models.

There Is Positionality in NLP Models and datasets have positionality, as they align better with some populations than others. This corroborates work from [Cambo and Gergle \(2022\)](#) on model positionality, which measures positionality by inspecting the content of annotated documents. We extend this work by showing design biases and

measuring positionality with Pearson’s r . We also focus on design biases in datasets and models and their downstream impact on different populations.

Our case studies show examples of positionality in NLP. However, most socially-aligned tasks may encode design biases due to differences in language use between demographic groups, for example, commonsense reasoning ([Shwartz, 2022](#)), question answering ([Gor et al., 2021](#)), and sentiment analysis ([Mohamed et al., 2022](#)). Even tasks which are considered purely linguistic have seen design biases. In parsing and tagging, performance differences exist between texts written by people of different genders ([Garimella et al., 2019](#)), ages ([Hovy and Søgaard, 2015](#)), and races ([Johannsen et al., 2015](#); [Jørgensen et al., 2015](#)). This shows how common design biases are in NLP, as language is a social construct ([Burr, 2015](#)) and technologies are imbued with their creator’s values ([Friedman, 1996](#)). This raises the question of whether there are any value-neutral language models ([Birhane et al., 2022](#); [Winner, 2017](#)).

Datasets and Models Are Western Across all tasks, models, and datasets, we find statistically significant moderate correlations with Western and educated populations, indicating that language technologies are WEIRD, though each to varying degrees. Prior work identifies Western-centric biases in NLP research ([Hershcovich et al., 2022](#)), as a majority of research is conducted in the West ([ACL, 2017](#); [Caines, 2021](#)). [Joshi et al. \(2020\)](#); [Blasi et al. \(2022\)](#) find disproportionate amounts of resources dedicated to English in NLP research, while [Ghosh et al. \(2021\)](#) identify cross-geographic errors made by toxicity models in non-Western contexts. This could lead to serious downstream implications such as language extinction ([Kornai, 2013](#)) or acculturation ([Ward, 1996](#)). Not addressing these biases risks imposing Western standards on non-Western populations, potentially resulting in a new kind of colonialism in the digital age ([Irani et al., 2010](#)).

Some Populations Are Left Behind Certain demographics consistently rank lowest in their alignment with datasets and models across both tasks compared to other demographics of the same type. Prior work has also reported various biases against these populations in datasets and models: people who are non-binary (e.g., [Dev et al., 2021](#)), Black (e.g., [Sap et al., 2019](#); [Davidson et al., 2019](#)), Latinx (e.g., [Dodge et al., 2021b](#)), Native Amer-

ican (e.g., Mager et al., 2018); people who did not complete high school; and people who are not native English speakers (e.g., Joshi et al., 2020). These communities are historically marginalized by technological systems (Bender et al., 2021).

Datasets Tend to Align with Their Annotators

We observe that the positionality we compute is similar to the reported annotator demographics of the datasets, indicating that annotator background contributes to dataset positionality. Social Chemistry reports their annotators largely being women, White, between 30-39 years old, having a college education, and from the U.S. (Forbes et al., 2020), all of which have high correlation to the dataset. Similarly, Dynahate exhibits high correlation with their annotator populations, which are mostly women, White, 18-29 years old, native English speakers, and British (Vidgen et al., 2021). This could be because annotators’ positionalities cause them to make implicit assumptions about the context of subjective annotation tasks, which affects its labels (Wan et al., 2023; Birhane et al., 2022). In toxicity modeling, men and women value speaking freely versus feeling safe online differently (Duggan et al., 2014).

Recommendations Based on these findings, we discuss some recommendations. Following prior work on documenting the choices made in building datasets (Gebu et al., 2021b) and models (Bender and Friedman, 2018; Bender et al., 2021), researchers should keep a record of all design choices made while building them. This can improve reproducibility (NAACL, 2021; AACL, 2023) and aid others in understanding the rationale behind the decisions, revealing some of the researcher’s positionality. Similar to the "Bender Rule" (Bender, 2019), which suggests stating the language used, researchers should ideally report their positionality and the assumptions they make. This should be done after paper acceptance to preserve anonymity.

We echo prior work in recommending methods to center the perspectives of communities who are harmed by design biases (Blodgett et al., 2020; Hanna et al., 2020; Bender et al., 2021), using approaches such as participatory design (Spinuzzi, 2005) (e.g., interactive storyboarding (Madsen and Aiken, 1993)) and value-sensitive design (Friedman, 1996) (e.g., panels of experiential experts (Madsen and Aiken, 1993)). Building datasets and models with large global teams

such as BigBench (Srivastava et al., 2022) and NL-Augmenter (Dhole et al., 2021) could also reduce design biases by having a diverse teams (Li, 2020).

To account for annotator subjectivity (Aroyo and Welty, 2015), researchers should make concerted efforts to recruit annotators from diverse backgrounds. Websites similar to LabintheWild can be platforms where these annotators are recruited. Since new design biases could be introduced in this process, we recommend following the practice of documenting the demographics of annotators as in prior works (e.g., Forbes et al., 2020; Vidgen et al., 2021) to record a dataset’s positionality.

We urge considering research through the lens of perspectivism (Basile et al., 2021), i.e. being mindful of different perspectives by sharing datasets with disaggregated annotations and finding modelling techniques that can handle inherent disagreements or distributions (Plank, 2022), instead of forcing one single answer in the data (e.g., by majority vote (Davani et al., 2022)) or in the model (e.g., by classification to one label (Costanza-Chock, 2018)). Researchers also should carefully consider how they aggregate labels from diverse annotators during modeling so their perspectives are represented, such as not averaging annotations to avoid the "tyranny of the mean" (Talat et al., 2022).

6 Conclusion

We introduce NLPOSITIONALITY, a framework to measure design biases and positionality of datasets and models. In this work, we present how researcher positionality leads to design biases and subsequently gives positionality to datasets and models, resulting in these artifacts not working equally for all populations. Our framework involves a recruiting demographically diverse pool of crowdworkers from around the world on LabintheWild, who then re-annotate a sample of a dataset for an NLP task. We apply NLPOSITIONALITY to two tasks, social acceptability and hate speech detection, to show that models and datasets have a positionality and design biases by aligning better with Western, White, and college-educated populations. Our results indicates the need for more inclusive models and datasets, paving the way for NLP research that can benefit all people.

(Sambasivan et al., 2021; Koch et al., 2021)

7 Limitations

Our study has several limitations. For example, study annotators could purposefully answer badly, producing low-quality annotations. We addressed this threat by using LabintheWild, which is known to produce high-quality data because participants are motivated to participate by learning something about themselves (Reinecke and Gajos, 2015). By answering providing low-quality annotations, they waste their own time by not learning anything about themselves. Additionally, Pearson’s r may not fully capture alignment as it does not consider interaction effects between different demographics (i.e., intersectionality). We also took the average of the annotations per group, which could mask individual variations (Talat et al., 2022).

As part of our study, we applied NLPOSITIONALITY to only two tasks which have relatively straightforward annotation schemes. It may be difficult to generalize to other NLP tasks which have harder annotation schemes, especially ones which require a lot of explanation to the annotators, for example, natural language inference (NLI) task.

Our approach is evaluated and works the best for classification tasks and classifiers. Generation tasks would need more careful annotator training which is difficult to achieve on a voluntary platform without adequate incentives. Having annotators use one Likert scale to rate the social acceptability and toxicity of a situation or text may not be a sufficient measure of to represent these complex social constructs. To reduce this threat, we provided detailed instructions that described how to provide annotations and followed the original annotation setup as closely as possible.

8 Ethics Statement

Towards Inclusive NLP Systems Building inclusive NLP systems are important so that everyone can benefit from their usage. Currently, these systems exhibit many design biases that negatively impact minoritized or underserved communities in NLP (Joshi et al., 2020; Blodgett et al., 2020; Bender et al., 2021). Our work is a step towards reducing these disparities by understanding that models and datasets have positionalities and by identifying design biases. The authors take inspiration from fields outside of NLP by studying positionality (Rowe, 2014) and acknowledge cross-disciplinary research as crucial to building inclusive AI systems.

Ethical Considerations We recognize that the demographics we collected only represent a small portion of a person’s positionality. There are many aspects of positionality that we did not collect, such as sexual orientation, socioeconomic status, ability, and size. Further, we acknowledge that limitation of assigning labels to people as being inherently reductionist. As mentioned in §7, using a single Likert scale to social acceptability and toxicity is not sufficient in capturing the complexities in these phenomena, such as situation context.

We note that measuring positionality of existing systems is not an endorsement of the system. In addition to making sure these systems work for all populations, researchers should also continue to examine whether these systems should exist at all (Denton and Gebru, 2020; Keyes et al., 2019). Further, we note that understanding a dataset or model’s positionality does not preclude researchers from the responsibilities of adjusting it further.

This study underwent IRB approval. LabintheWild annotators were not compensated financially. They were lay people from a wide range of ages (including minors) and diverse backgrounds. Participants were asked informed consent of the study procedures as well as the associated risks, such as being exposed to toxic or mature content, prior to beginning the study.

References

- AAAI. 2023. [Reproducibility checklist](#).
- ACL. 2017. [ACL Diversity Statistics](#).
- Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2022. Probing Pre-Trained Language Models for Cross-Cultural Differences in Values. *ArXiv*, abs/2203.13722.
- Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.
- Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2022. Morality beyond the WEIRD: How the nomological network of morality varies across cultures.
- Tal August and Katharina Reinecke. 2019. [Pay attention, please: Formal language improves attention in volunteer and paid online experiments](#). In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, page 1–11, New York, NY, USA. Association for Computing Machinery.

- Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature*, 563(7729):59–64.
- Edmond Awad, Sohan Dsouza, Azim Shariff, Iyad Rahwan, and Jean-François Bonnefon. 2020. [Universals and variations in moral decisions made in 42 countries by 70,000 participants](#). *Proceedings of the National Academy of Sciences*, 117(5):2332–2337.
- Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a perspectivist turn in ground truthing for predictive computing. *arXiv preprint arXiv:2109.04270*.
- Emily Bender. 2019. The# benderrule: On naming the languages we study and why it matters. *The Gradient*, 14.
- Emily M. Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#). In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, page 610–623, New York, NY, USA. Association for Computing Machinery.
- Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. [The values encoded in machine learning research](#). In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, page 173–184, New York, NY, USA. Association for Computing Machinery.
- Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. [Systematic inequalities in language technology performance across the world’s languages](#). In *Association for Computational Linguistics*, pages 5486–5505.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in nlp. In *ACL*.
- Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to home-maker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29.
- Vivien Burr. 2015. *Social constructionism*. Routledge.
- Andrew Caines. 2021. [The geographic diversity of nlp conferences](#).
- Scott Allen Cambo and Darren Gergle. 2022. [Model positionality and computational reflexivity: Promoting reflexivity in data science](#). In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI ’22, New York, NY, USA. Association for Computing Machinery.
- Sasha Costanza-Chock. 2018. Design justice, ai, and escape from the matrix of domination.
- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. [Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations](#). *Transactions of the Association for Computational Linguistics*, 10:92–110.
- Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *ACL Workshop on Abusive Language Online*.
- Emily Denton and Timnit Gebru. 2020. Tutorial on fairness accountability transparency and ethics in computer vision at cvpr 2020.
- Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. [Harms of gender exclusivity and challenges in non-binary representation in language technologies](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Kaustubh D Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, Samson Tan, et al. 2021. Nl-augmenter: A framework for task-sensitive natural language augmentation. *arXiv preprint arXiv:2112.02721*.
- 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*, WSDM ’18, page 135–143, New York, NY, USA. Association for Computing Machinery.
- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021a. [Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021b. [Documenting large webtext corpora: A case study on the colossal clean crawled corpus](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

- Maeve Duggan, L Rainie, A Smith, C Funk, A Lenhart, and M Madden. 2014. Online harassment. *Pew Res. Center, Washington, DC, USA, Tech. Rep.*
- Mary Q Foote and Tonya Gau Bartell. 2011. Pathways to equity in mathematics education: How life experiences impact researcher positionality. *Educational Studies in Mathematics*, 78(1):45–68.
- Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. [Social chemistry 101: Learning to reason about social and moral norms](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.
- Batya Friedman. 1996. Value-sensitive design. *Interactions*, 3(6):16–23.
- Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems*, 14(3):330–347.
- Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. [Women’s syntactic resilience and men’s grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy. Association for Computational Linguistics.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021a. [Datasheets for datasets](#). *Commun. ACM*, 64(12):86–92.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021b. [Datasheets for datasets](#). *Commun. ACM*, 64(12):86–92.
- Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media. In *Workshop on Noisy User-generated Text (W-NUT)*.
- Maharshi Gor, Kellie Webster, and Jordan Boyd-Graber. 2021. [Toward deconfounding the effect of entity demographics for question answering accuracy](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5457–5473, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Suchin Gururangan, Dallas Card, Sarah K Drier, Emily K Gade, Leroy Z Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection. *arXiv preprint arXiv:2201.10474*.
- Christian W Haerpfer and Kseniya Kizilova. 2012. The world values survey. *The Wiley-Blackwell Encyclopedia of Globalization*, pages 1–5.
- Melissa Hall, Laurens van der Maaten, Laura Gustafson, and Aaron Adcock. 2022. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*.
- Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindřich Libovický, Constantin A Rothkopf, Alexander Fraser, and Kristian Kersting. 2022. Speaking multiple languages affects the moral bias of language models. *arXiv preprint arXiv:2211.07733*.
- 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, FAT* ’20*, page 501–512, New York, NY, USA. Association for Computing Machinery.
- 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](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.
- Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.
- 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 *Association for Computational Linguistics*, pages 6997–7013.
- Andrew Gary Darwin Holmes. 2020. Researcher positionality—A consideration of its influence and place in qualitative research—A new researcher guide. *Shanlax International Journal of Education*, 8(4):1–10.
- Dirk Hovy and Anders Søgaard. 2015. [Tagging performance correlates with author age](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488, Beijing, China. Association for Computational Linguistics.
- Panagiotis G Ipeirotis. 2010. Demographics of Mechanical Turk.
- Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E. Grinter. 2010. [Postcolonial Computing: A Lens on Design and Development](#). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI ’10*, page 1311–1320,

- New York, NY, USA. Association for Computing Machinery.
- Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*.
- Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. [Cross-lingual syntactic variation over age and gender](#). In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China. Association for Computational Linguistics.
- Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. [Challenges of studying and processing dialects in social media](#). In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18, Beijing, China. Association for Computational Linguistics.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Os Keyes, Jevan Hutson, and Meredith Durbin. 2019. [A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry](#). In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, page 1–11, New York, NY, USA. Association for Computing Machinery.
- Bernard Koch, Emily Denton, Alex Hanna, and Jacob G Foster. 2021. Reduced, reused and recycled: The life of a dataset in machine learning research. *arXiv preprint arXiv:2112.01716*.
- András Kornai. 2013. [Digital language death](#). *PLOS ONE*, 8(10):1–11.
- Klaus Krippendorff. 2006. [Reliability in Content Analysis: Some Common Misconceptions and Recommendations](#). *Human Communication Research*, 30(3):411–433.
- M Li. 2020. To build less-biased ai, hire a more diverse team. *Harvard Bus. Rev.*, Oct.
- Kim Halskov Madsen and Peter H. Aiken. 1993. [Experiences using cooperative interactive storyboard prototyping](#). *Commun. ACM*, 36(6):57–64.
- Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. [Challenges of language technologies for the indigenous languages of the Americas](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. [Model cards for model reporting](#). In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.
- Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Ward Church, and Mohamed Elhoseiny. 2022. Artelingo: A million emotion annotations of wikiart with emphasis on diversity over language and culture. *arXiv preprint arXiv:2211.10780*.
- NAACL. 2021. [Reproducibility checklist](#).
- Nigini Oliveira, Eunice Jun, and Katharina Reinecke. 2017. [Citizen science opportunities in volunteer-based online experiments](#). In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 6800–6812, New York, NY, USA. Association for Computing Machinery.
- Barbara Plank. 2022. The 'problem' of human label variation: On ground truth in data, modeling and evaluation. *arXiv preprint arXiv:2211.02570*.
- Katharina Reinecke and Krzysztof Z. Gajos. 2015. [TabInTheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples](#). In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '15, page 1364–1378, New York, NY, USA. Association for Computing Machinery.
- Wendy E Rowe. 2014. Positionality. *The SAGE encyclopedia of action research*, 628:627–628.
- 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*, CHI '21, New York, NY, USA. Association for Computing Machinery.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. [The risk of racial bias in hate speech detection](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. [Annotators with attitudes: How annotator beliefs and identities bias toxic language detection](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

- Maggi Savin-Baden and Claire Howell-Major. 2013. Qualitative research: The essential guide to theory and practice. *Qualitative Research: The Essential Guide to Theory and Practice*. Routledge.
- Vered Shwartz. 2022. [Good night at 4 pm?! time expressions in different cultures](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853, Dublin, Ireland. Association for Computational Linguistics.
- Clay Spinuzzi. 2005. The methodology of participatory design. *Technical communication*, 52(2):163–174.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Zeera Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. [On the machine learning of ethical judgments from natural language](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779, Seattle, United States. Association for Computational Linguistics.
- Rachael Tatman. 2017. [Gender and dialect bias in YouTube’s automatic captions](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.
- Bertie Vidgen, Tristan Thrush, Zeera Waseem, and Douwe Kiela. 2021. [Learning from the worst: Dynamically generated datasets to improve online hate detection](#). In *Association for Computational Linguistics*, pages 1667–1682.
- Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone’s voice matters: Quantifying annotation disagreement using demographic information. *arXiv preprint arXiv:2301.05036*.
- Colleen Ward. 1996. Acculturation.
- William Warner and Julia Hirschberg. 2012. [Detecting hate speech on the world wide web](#). In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.
- Thomas D Wickens and Geoffrey Keppel. 2004. *Design and Analysis: A Researcher’s Handbook*. Prentice-Hall.
- Langdon Winner. 2017. Do artifacts have politics? In *Computer Ethics*, pages 177–192. Routledge.
- Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. [Detoxifying Language Models Risks Marginalizing Minority](#) [Voices](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.

A Data

In this section, we describe all the decisions that went into sampling data points from the different datasets and its post-processing.

A.1 Sampling

For Social Chemistry, we sampled instances whose label for anticipated agreement by the general public was "Controversial ($\sim 50\%$)". We ensured the samples were equally represented by the moral foundation label, which we computed based on majority vote across annotators. For the study, annotators responded whether they found a presented action moral.

For Dynahate, we randomly sampled instances from rounds 3 and 4. In these rounds, annotators generated examples of implicit hate, which is subtler and harder to detect and could yield differences in annotations. We ensured that there were equal amounts of hateful and not hateful instances and that the types of targets of the hateful instances were equally represented. Annotators responded whether they found a presented instance toxic.

For both social acceptability and hate speech detection, annotators responded whether they found the situation moral and whether they found the instance to be hate speech respectively.

A.2 Post-Processing

Because Social Chemistry had multiple annotations for each instance, we compute an aggregate score by taking the average score across annotators. This score is then used to correlate to the annotators' aggregated scores.

B Study Design

In this section, we discuss the design of the LabintheWild experiments. The social acceptability speech task was released to the public on April 2022. The hate speech detection task was released August 2022. To reduce confounding factors on the data collection process, we conducted multiple user studies of the LabintheWild experiments prior to the public release. Additionally, all the annotations collected through the experiments were anonymous and stored securely.

The social acceptability task was marketed as "Could you live with an AI and its morals?" Participants for this study provided annotations for 25 situations. The hate speech detection task was marketed as "Do you and AI agree on what is hate

speech? Let's find out!" Participants provided annotations for 15 instances.

B.1 LabintheWild Study Flow

We describe the format of the LabintheWild experiment. The phases of the experiment were: obtaining consent, collecting demographics, explaining instructions, collecting annotations, collecting study feedback, and displaying results.

Obtaining Consent Prior to beginning the study, participants reviewed a consent form. The consent form included information on the purpose of the research, what the participant will do, risks and benefits of the research, privacy and data collection methods, and contact information of the researchers. At the end of the form, participants gave explicit consent to participate in the study.

Collecting Demographics We then collect the demographics of study participants. LabintheWild participants entered in whether they had taken this test before, the country they lived in the longest, the country of residence, age, native language, religion, education, and ethnicity. No demographics were required except for the country the participant lived in the longest and whether they had taken the test before. Additionally, we only displayed ethnicity for people within the United States.

Explaining Instructions For each task, we provided instructions to participants on how to perform the annotation task. For social acceptability, we explained social acceptability as rating "what you think about the situation in general from an ethical perspective" (see Figure 5). For hate speech detection, used the definition of hate speech from Dynahate and we provided three examples of hate speech (see Figure 6). We also presented examples of negative sentiment, profanity, or discussing groups that could be confused as hate speech, but were not.

Collecting Annotations After being presented with instructions, participants began data collection from the 300 instances selected from Section A.1. For each task, we kept the annotation set up identical to the original one. For social acceptability, we collected Likert-scale ratings of situations ranging from "It's very bad", "It's bad", "It's okay", "It's good", and "It's very good". Participants could provide rationale by using an open text box. The data collection interface is presented in Figure 4.

For hate speech detection, we collected ratings of instances ranging from "Hate speech", "Not sure", "Not hate speech." We also provided an optional open-text box for participants to explain their rationale. The data collection interface is presented in Figure 7. After submitting the annotation, the participant is able to see a visualization on how the AI responded as well as how other participants from the same country responded to the instance.

We also specifically sampled which instances to present to participants to annotate. We sampled a third of the instances that did not have any annotations from the demographic and a third that were already sampled by participants of the demographic. The rest were equally split across the different types of instances (i.e., moral foundation for Social Chemistry, hate type for Dynahate).

Providing Study Feedback Following typical LabInTheWild experiment procedures, we collected feedback from participants about the study. Participants could enter open-text feedback on anything. They also submitted whether they encountered technical difficulties during the study or whether they cheated. Participants could elaborate on their answers from the prior questions in an open-text box.

Displaying Overall Results Finally, participants were displayed their overall results for the experiment task. First, participants were presented with the percentage of time they agreed with the AI as well as with participants as the same demographic as them (see Figure 8). Each of these agreement scores were further broken down by the type of the instance (i.e., moral foundation for Social Chemistry and hate type for Dynahate).

C Cultural Spheres

Division of countries can be done through continents. However, continents are often not representative of the countries within it and clustering based on them can lead to inaccurate findings. For example, Asia includes both Japan and Saudi Arabia, which are different culturally. We instead adopt cultural spheres as used in World Values Survey (Haerpfer and Kizilova, 2012), which clusters the countries in terms of the values they uphold and norms they follow. Table 3 shows the countries and the spheres.

1) Read the situation.

1 / 15

Wanting to make a lot of money.

2) Enter what you think about it.

* denotes a required field.

What do you think about this from a moral perspective?

Very bad

Bad

Expected / OK

Good

Very good

Feel free to explain your judgment.

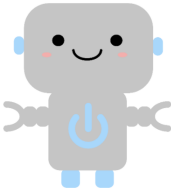
(e.g., it's ok, you should, it's unusual)

It's generally good, but you don't want to exploit others in the process.

Submit

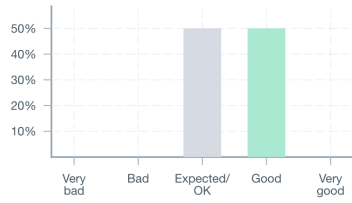
3) See what an AI and others thought about it!

The AI speculates:



I speculate that it's understandable. Hm... Maybe I'm incorrect on this.

Study participants in the United States said:



Response	Percentage
Very bad	0%
Bad	0%
Expected/OK	50%
Good	50%
Very good	0%

Figure 4: Data collection interface for the social acceptability task. Participants were given a sentence (an action from the Social Chemistry dataset) and asked to rate how ethical the action was. Participants are shown how other people from their country responded after each attempt.

Instructions

You will be shown 25 different scenarios. For each, select what you think about the situation in general from an ethical perspective. You can elaborate on your perspective if you wish. Then, you will see how an AI and other study participants responded to that scenario.

At the end of the study, we will show how your responses compared with an AI's and other participants' overall.

1) Read the situation.

1 / 15

Wanting to make a lot of money.

2) Enter what you think about it.

* denotes a required field.

What do you think about this from a moral perspective?

Very bad

Bad

Expected / OK

Good

Very good

Feel free to explain your judgment.

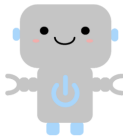
(e.g., it's ok, you should, it's unusual)

It's generally good, but you don't want to exploit others in the process.

Submit

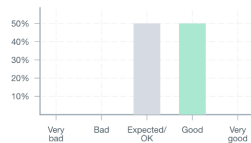
3) See what an AI and others thought about it!

The AI speculates:



I speculate that it's understandable. Hm... Maybe I'm incorrect on this.

Study participants in the United States said:



Response	Percentage
Very bad	0%
Bad	0%
Expected/OK	50%
Good	50%
Very good	0%

Let's start!

Figure 5: Instructions for the social acceptability task. Participants were asked to describe their thoughts about a situation from an ethical perspective.

Instructions

You will be shown 12 different examples. For each, rate whether you think the speech is **hateful** towards a group. You can elaborate on your rating if you wish. Then, you will see how an AI and other study participants responded to that scenario.

At the end of the study, we will show how your responses compared with an AI's and other participants' overall.

Important: Please note that some examples may contain content that is offensive, problematic, or harmful. You are able to skip examples if you'd like.

What is hateful speech?

Hateful speech is abusive speech targeting specific social groups and their characteristics (e.g., ethnicity, religion, gender, or sexual orientation).

Probably Hateful

- "On 9/11, Arabs brought terror to Americans. So I think it's only fair that Arabs are not welcome in the U.S."
This statement implies that Arab people do not belong.
- "Gay teachers should not be allowed. At the very least they should keep their personal life to themselves, away from students."
This statement implies gay people do not have the same right to teach as straight people.
- "To have a good life, you should have light skin, if he lightened his skin he could pass for white."
This statement implies that people should lighten their skin to be accepted.

Probably Not Hateful

- "The photos came out so blurry, I can't even tell who's in them. This shitty camera is awful at taking pictures."
This statement includes profanity but does not target a group of people.
- "Chinese people know how to make good food, my favorite food of all time is dim sum."
This statement references a group of people but is not abusive language.
- "I'm always surprised when people say they love Apple. Their products suck, I hate my phone."
This statement includes negative sentiment but does not target a group of people.

Figure 6: Instructions for the toxicity task. Participants were provided with examples of hate speech examples and not hate examples.

1) Read the example. 7 / 25

English people smell like sour milk

2) Enter what you think about it. * denotes a required field.

* Do you think this is hateful speech? ☒ Yes ☐ No ☐ Not sure

Feel free to explain your rating.

3) See what others thought about it!

The AI speculates:

I speculate that this is hateful speech. Looks like we agree!

Study participants in Afghanistan said:

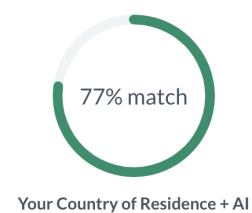
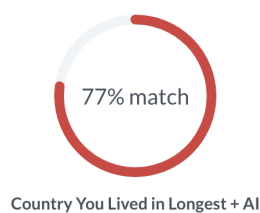
Rating	Percentage
Not Hateful	35%
Not Sure	0%
Hateful	65%

Figure 7: Data collection interface for the hate speech task. Participants were given a sentence (an instance from the Dynahate dataset) and asked to rate whether the instance was toxic or not. Participants are shown how other people from their country responded after each attempt.

Here's how often your responses matched with the AI's.



People from the same demographic as you scored...



When did you agree with the AI?

Here's a breakdown of your match score with the AI by the type of moral situation. Higher values indicate higher agreement with the AI.

Types of moral situations

- 💖 **Care/harm** is morals of having empathy towards the pain of others (e.g., valuing kindness).
- ⚖️ **Fairness/cheating** relates to morals from reciprocated altruism (e.g., valuing justice).
- 👉 **Loyalty/betrayal** is morals from building alliances (e.g., valuing patriotism).
- 👑 **Authority/subversion** is morals based on social hierarchies (e.g., valuing leadership).
- 🌸 **Sanctity/degradation** relates to morals of living in an elevated and noble manner.
- 🏠 **Everyday** refers to everyday situations which have no moral implications.

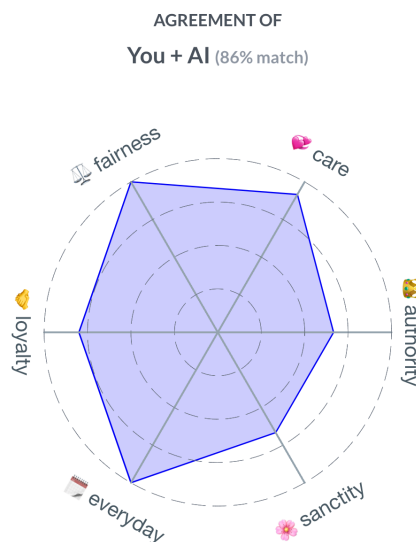


Figure 8: Results interface for the social acceptability task. Participants can view how well they aligned with the AI, as well as how other demographics they reported aligned with the AI. The AI alignment is further broken down by the type of moral foundation.

Cultural Sphere	Countries
African-Islamic	Afghanistan, Albania, Algeria, Azerbaijan, Indonesia, Iraq, Morocco, Saudi Arabia, South Africa, Syrian Arab Republic, Turkey, United Arab Emirates, Uzbekistan Burkina Faso, Bangladesh, Egypt, Ethiopia, Ghana, Iran, Jordan, Kazakhstan, Kyrgyzstan, Lebanon, Libya, Mali, Nigeria, Pakistan, Palestine, Qatar, Rwanda, Tajikistan, Tanzania, Tunisia, Uganda, Yemen, Zambia, Zimbabwe
Baltic	Estonia, Latvia, Lithuania, Åland Islands
Catholic-Europe	Austria, Belgium, Czech Republic, France, Hungary, Italy, Poland, Portugal, Spain Andorra, Croatia, Lithuania, Luxembourg, Slovakia, Slovenia
Confucian	China, Hong Kong, Japan Macao, South Korea, Taiwan
English-Speaking	American Samoa, Australia, Canada, Guernsey, Ireland, New Zealand, United Kingdom, United States
Latin-America	Brazil, Colombia, Dominican Republic, Mexico, Philippines, Trinidad and Tobago Argentina, Bolivia, Chile, Ecuador, Guatemala, Haiti, Nicaragua, Peru, Puerto Rico, Trinidad, Uruguay, Venezuela
Orthodox-Europe	Bosnia and Herzegovina, Bulgaria, Cyprus, Greece, Romania, Russian Federation, Serbia, Ukraine Armenia, Belarus, Bosnia, Georgia, Moldova, Montenegro, North Macedonia, Russia
Protestant-Europe	Denmark, Finland, Germany, Iceland, Netherlands, Norway, Sweden, Switzerland
West-South-Asia	India, Israel, Malaysia, Singapore Myanmar, Thailand, Vietnam

Table 3: Cultural Spheres and their corresponding countries from (Haerpfer and Kizilova, 2012). Black color indicates that the countries are part of our collected data. Gray color indicates countries not part of our analysis—we have included them to give an idea of what other countries belong to the spheres.