Do Large Language Models Understand Morality Across Cultures?

Hadi Mohammadi [a,*], **Yasmeen F.S.S. Meijer** [a], **Efthymia Papadopoulou** [a] **and Ayoub Bagheri**[a]

[a]Department of Methodology and Statistics, Utrecht University, The Netherlands

Abstract. Recent advancements in large language models (LLMs) have established them as powerful tools across numerous domains. However, persistent concerns about embedded biases, such as gender, racial, and cultural biases arising from their training data, raise significant questions about the ethical use and societal consequences of these technologies. This study investigates the extent to which LLMs capture cross-cultural differences and similarities in moral perspectives. Specifically, we examine whether LLM outputs align with patterns observed in international survey data on moral attitudes. To this end, we employ three complementary methods: (1) comparing variances in moral scores produced by models versus those reported in surveys, (2) conducting cluster alignment analyses to assess correspondence between country groupings derived from LLM outputs and survey data, and (3) directly probing models with comparative prompts using systematically chosen token pairs. Our results reveal that current LLMs often fail to reproduce the full spectrum of cross-cultural moral variation, tending to compress differences and exhibit low alignment with empirical survey patterns. These findings highlight a pressing need for more robust approaches to mitigate biases and improve cultural representativeness in LLMs. We conclude by discussing the implications for the responsible development and global deployment of LLMs, emphasizing fairness and ethical alignment.

1 Introduction

Large language models (LLMs) have recently taken center stage in both scientific and public debates due to significant advancements in their performance [2]. These models now show great promise for applications ranging from search engines and recommendation systems to automated decision-making tools that deeply influence everyday life. Nonetheless, alongside these impressive capabilities, concerns persist regarding the potential biases LLMs can exhibit, such as gender, racial, and cultural bias.

A primary reason for this risk is that LLMs learn from vast, real-world text datasets that may contain societal and cultural prejudices [11, 16]. Consequently, when large portions of the training data systematically reflect certain groups unfavorably, the resulting language model may replicate or even amplify those biases. Given the growing reliance on LLM-based systems across many fields, this raises important questions about whether these models truly capture the diverse moral perspectives observed in actual human societies.

Despite its importance, the issue of whether LLMs accurately reflect cross-cultural moral judgments has been relatively understudied [1, 15]. In examining how faithfully LLMs capture moral attitudes that vary across cultural contexts, a key consideration is their ability to replicate both the areas of divergence (where cultures disagree) and similarity (where cultures align) on moral topics. Thus, the central research question is:

> *To what extent do language models capture cultural diversity and common tendencies regarding topics on which people around the world tend to diverge or agree in their moral judgments?*

Addressing this question carries both scientific and societal significance. Scientifically, it provides insight into how effectively LLMs, trained primarily on text data, can model complex cultural norms. Societally, ensuring these models reflect actual cross-cultural variation is vital for preventing biased or inaccurate representations of different cultural groups [15]. As LLMs increasingly shape public opinion and decision-making, a mismatch between how cultures truly view moral issues and how models characterize these issues can perpetuate prejudice and unfairness. Conversely, LLMs that accurately capture inter-cultural moral differences and similarities can help reveal common ground and support cross-cultural understanding.

Against this backdrop, the present study focuses on evaluating the extent to which contemporary LLMs mirror the diversity and patterns of moral judgments observed across cultures. Three primary methods are employed:

1. **Comparing Variances**: We compare the variance in model-generated moral judgments with the variance in survey-based moral judgments across countries.
2. **Cluster Alignment**: We examine the alignment of model-induced country clusters with empirically derived clusters.
3. **Direct Comparative Prompts**: We probe LLMs using tailored prompts to see whether they recognize similarities and differences in moral perspectives between countries.

By using these complementary techniques, this work offers insights into the strengths and limitations of LLMs in depicting cross-cultural moral norms, ultimately informing ongoing discussions about their ethical development and deployment. The remainder of this paper is structured as follows. In Section 2, we review related research on cross-cultural moral judgments in LLMs and the issue of bias in these models. Section 3 describes the data and methods used in our analysis, and Section 4 details the results. We then discuss key findings in and conclude with final remarks in Section 5.

* Corresponding Author. Email: h.mohammadi@uu.nl.

2 Related work on moral judgment and LLM bias

2.1 Cross-Cultural Understanding of Moral Judgments in LLMs

Moral judgments are evaluations of whether specific actions, intentions, or individuals are morally "good" or "bad," and they can vary widely across cultures due to social norms, religious doctrines, and historical influences [8, 25]. Broadly speaking, Western, Educated, Industrialized, Rich, and Democratic societies—commonly abbreviated as W.E.I.R.D. in cross-cultural psychology literature [9], tend to prioritize autonomy[1], individual rights, and personal choice, whereas many non-W.E.I.R.D. cultures place a higher emphasis on communal obligations, duty, and spiritual purity [6]. For instance, individuals in W.E.I.R.D. contexts commonly regard sexual behaviors as a matter of personal freedom, while those from more community-oriented cultures may treat the same behaviors as collective moral issues.

Scholars such as Johnson et al. [10] and Benkler et al. [3] refer to this diversity of valid yet conflicting moral values as "moral value pluralism." Kharchenko et al. [12] caution that LLMs often fail to accurately reflect this pluralism, partly because these models are trained on large but not necessarily diverse datasets. Du et al. [5] likewise note that an overemphasis on English-language training data can overshadow the linguistic and cultural richness of the real world, highlighting the importance of multilingual corpora and larger model sizes. Indeed, Arora et al. [1] propose that multilingual LLMs hold promise for capturing cross-cultural values, though the potential lack of diversity within available multilingual data remains a limiting factor.

Consistent with these concerns, Benkler et al. [3] argue that most AI systems mirror the dominant values of the culture (often Western) producing the majority of the data. This phenomenon can result in a moral bias, whereby W.E.I.R.D. norms and perspectives are incorrectly treated as universally applicable. Empirical investigations of whether LLMs uphold or correct such biases are limited. Some work suggests that they struggle to reproduce culturally specific moral codes [1, 3], while other findings are more optimistic about LLMs' capacity to model cultural diversity [24, 18]. This divergence underscores the importance of continued research on how language models perceive and replicate moral frameworks across various cultures.

2.2 The Risk of Bias in LLMs

Bias in LLMs arises when these models inherit or amplify prejudices present in their training datasets. Typically, LLMs learn language representations (or embeddings) by analyzing co-occurrences of words across massive corpora. If these corpora disproportionately depict certain groups or behaviors negatively, the learned representations can perpetuate or exacerbate harmful stereotypes in model outputs [21].

A well-known example is the gender bias identified in word embeddings, where terms like "woman" are closely associated with "homemaker" and "man" with "computer programmer" [4]. Another instance is GPT-3's tendency to associate "Muslims" with violent acts more than "Christians" [10]. Recent work has shown that while demographic biases influence LLM outputs, content-specific features remain the dominant factor in model predictions [19]. Although ongoing research aims to mitigate bias [16], this task remains daunting, as biased outputs can influence everything from public sentiment to automated hiring decisions [22].

For instance, an LLM trained on biased sources might disproportionately recommend men for technical positions, perpetuating gender inequality [4]. In a similar vein, consistently linking certain religious groups with violence can reinforce negative stereotypes and intensify discrimination. Given these high-stakes consequences, developing models that faithfully capture cultural diversity rather than simplifying or skewing moral perspectives is not merely an academic challenge but a moral and societal imperative [28].

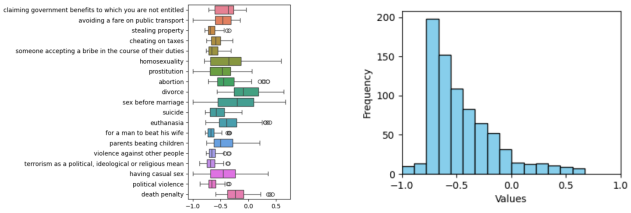
In summary, these two strands of literature, (1) how LLMs handle cross-cultural moral judgments and (2) how bias emerges and persists in LLMs, highlight the need to systematically examine how well these models capture the complexities of moral values across different societies. The following sections detail our data sources and methodological approach to investigating these issues.

3 Data and methods

3.1 Datasets

The World Values Survey[2] (WVS) provides detailed information on people's values across cultures. In this study, we use data from Wave 7 [7], which covers the period 2017–2020. This wave features participants from 55 countries who responded to 19 statements on moral issues (e.g., divorce, euthanasia, political violence, cheating on taxes). The survey was administered in the primary languages of each country, offering multiple response categories.

Only the country name and each response were retained, with values normalized to range from $[-1, 1]$, where -1 indicates "never justifiable" and 1 signifies "always justifiable." These normalized scores facilitate comparability and statistical analysis. For each country–moral issue pair, we computed an average (mean) rating, thus capturing a broad overview of each country's position. We acknowledge that averaging can obscure outlier or minority perspectives, but it was deemed the most feasible approach for this study. Figure 1 depict the overall distribution of these normalized scores and their variation across topics and countries.



(a) (b)

Figure 1: (a) Spread of responses across moral topics and countries in WVS Wave 7. (b) Distribution of normalized WVS Wave 7 answers.

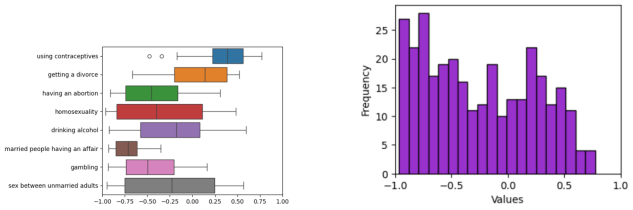
As a second dataset, we use the Pew Global Attitudes Project[3] (2013), which surveyed 39 countries (100 participants each) on 8 moral topics, such as drinking alcohol or getting a divorce. The questionnaire was administered in English, allowing respondents to categorize a topic as "morally acceptable," "not a moral issue," or "morally unacceptable."

We extracted only country names and responses (Q84A–Q84H), again transforming them to a $[-1, 1]$ scale and averaging scores by country–topic pair. Figure 2 summarizes the distribution of normalized scores and their topic-level variation.

[1] The acronym W.E.I.R.D. is a technical term from cross-cultural psychology used to identify a specific cluster of societies that are overrepresented in psychological research. It was introduced to highlight sampling bias in behavioral sciences and has become standard terminology in the field.

[2] https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp

[3] https://www.pewresearch.org/dataset/spring-2013-survey-data/



(a) (b)

Figure 2: (a) Spread of responses across moral topics and countries for PEW 2013. (b) Distribution of normalized PEW 2013 responses.

3.2 Pre-processing

In the preprocessing of version 5 of the World Values Survey (WVS) data, the dataset was initially filtered to retain only the columns corresponding to the moral questions Q177 to Q195 and the country code (B_COUNTRY). These questions cover a range of moral issues, such as tax cheating, accepting bribes, and attitudes towards homosexuality. Following the initial filtering, country names were assigned to each row based on the B_COUNTRY codes using a predefined country mapping dataset. Responses with values of -1, -2, -4, and -5, which represent 'Don't know,' 'No answer,' 'Not asked in survey,' and 'Missing; Not available,' respectively, were replaced with zero. This adjustment was made to ensure that calculations, such as averaging, were not affected by non-responses. The decision to replace with 0 ensures that the structure of the dataset remains intact. It avoids introducing NaN values or leaving cells empty, which could complicate subsequent data analysis tasks such as averaging or statistical modeling. Moreover, a replacement value of 0 ensures that non-responses do not influence the computed averages or other aggregated measures artificially. After replacing non-response values with 0, the dataset was aggregated by country, calculating the mean response for each moral question per country. This provided a country-specific average score for each ethical issue. To enable comparisons across different countries and questions, these average scores were normalized on a scale from -1 to 1, where 1 signifies that the behavior is justifiable in every case and -1 denotes it is never justifiable. This normalization involved adjusting the mean responses, which initially ranged from 1 to 10, to fit the new scale. This step was needed for cross-national comparisons. Finally, normalized values were rounded to four decimal places to enhance clarity.

3.3 Models

We begin with two monolingual English models. The first is **GPT-2**, chosen for its strong performance in generating coherent, contextually relevant text [23]. We use two versions from Hugging Face,*GPT-2 Medium* (355M parameters) and *GPT-2 Large* (774M parameters), to observe how increased model size influences their capacity to interpret morally charged content. Larger models generally capture more complex patterns and may better approximate cultural moral judgments. As a second monolingual model, we employ the **OPT** series by Meta AI [27]. Two variants, *OPT-125M* and *OPT-350M*, are included to benchmark smaller, computationally efficient architectures against larger ones. OPT models, like GPT-2, generate text by predicting the next word in a sequence, having been trained on diverse English corpora.

Next, we incorporate multilingual models to explore how exposure to varied linguistic data might shape moral judgments across different countries. The first is **BLOOM**, a transformer-based, autoregressive language model from the BigScience project, trained on 46 natural and 13 programming languages [14]. We specifically use BLOOM-560M and BLOOMZ-560M (fine-tuned for zero-shot learning), rather than the full 176B version, to keep computational requirements manageable. BLOOM's design aims for strong cross-lingual performance, offering a flexible approach for text tasks in multiple languages. Lastly, we include **Qwen**, developed by Alibaba Cloud. Qwen is also a multilingual transformer trained on 29 languages (including English and Chinese). Its latest versions demonstrate competitive results in language understanding, multilingual tasks, coding, and reasoning. For feasibility, we use the 0.5B parameter variant in this study.

While hundreds of open-source LLMs currently exist, our model selection was guided by several criteria designed to ensure a representative sample of the current landscape. We prioritized architectural diversity by including both traditional autoregressive models (GPT-2 and OPT) and multilingual transformer models (BLOOM and Qwen), allowing us to test whether different architectural approaches affect cultural understanding. The selected models also span a range of sizes from 125M to 774M parameters, enabling us to investigate whether increased model capacity correlates with better cross-cultural moral representation. Furthermore, we deliberately contrasted monolingual English models with multilingual ones to test the hypothesis that exposure to diverse languages during training might improve cultural sensitivity. All selected models are openly available, ensuring our experiments can be replicated and extended by other researchers. Finally, these specific models have been extensively studied in prior cultural bias research [1, 10], allowing our findings to build upon and be compared with existing literature. While we acknowledge this represents only a subset of available models, we argue it captures key variations in architecture, size, and training approach that are most relevant to understanding cultural bias. Future work should certainly expand this analysis to newer model families such as LLaMA, Mistral, and other emerging architectures to validate and extend our findings.

3.4 Evaluation metrics

We employ three methods to evaluate whether the selected LLMs capture variations and agreements in moral views across cultures:

1. Comparing the variances of survey-based and model-generated moral scores across countries.
2. Assessing alignment between country clusters derived from these respective scores.
3. Probing LLMs with direct comparative prompts to evaluate how they recognize cross-cultural moral differences and similarities.

For the first two methods, we compute a model-generated moral score for each country–topic pair by prompting the LLM and examining its log probabilities [18]. Two prompt templates are used:

- `In {country} {topic} is {moral_judgment}.`
- `People in {country} believe {topic} is {moral_judgment}.`

Pair #	Contrasting Statements
1	*always justifiable / never justifiable*
2	*right / wrong*
3	*morally good / morally bad*
4	*ethically right / ethically wrong*
5	*ethical / unethical*

Table 1: Token pairs used to prompt the model for moral judgments.

As shown in Table 1, five contrasting token pairs (e.g., *always justifiable* vs. *never justifiable*) are employed to elicit the LLM's stance. When probing, for example, the moral score on "abortion" in the United States using the first token pair, we issue:

In the United States abortion is always justifiable
In the United States abortion is never justifiable

The LLM outputs log probabilities for each statement. We subtract the log probability of the "immoral" statement from that of the "moral" statement to obtain a pair-specific score. We do this for all five token pairs in both prompt styles and average the results to produce a final model-generated moral score. This score mirrors the format of the empirical survey-based scores.

3.5 Evaluation

3.5.1 Comparison of Variances

We first compare cross-cultural variances in moral scores. For each topic, we compute the variance of survey-based scores and model-generated scores, identifying high-variance (controversial) and low-variance (widely agreed upon) topics. We then assess the association between these two sets of variances via Pearson's correlation. A strong positive correlation suggests the model effectively captures cross-country variability; a weak or negative correlation suggests otherwise [24]. This step also helps compare which topics surveys and models respectively deem most controversial or consensual.

3.5.2 Cluster Alignment

Next, we evaluate whether models replicate empirical cross-country patterns of moral judgment. We cluster countries based on their survey scores using K-means, determining the optimal number of clusters K via repeated runs and selecting the highest silhouette score. The same K is then applied to model-generated scores. We compare these two clusterings using Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) [20, 13], as well as their average (CAS), to quantify how well the model-based clusters align with the survey-based "ground truth."

We repeat this procedure thrice for each dataset: once using all moral topics, once focusing on the most controversial topics, and once on those with highest agreement. This reveals how well the models capture both contentious areas and common ground across cultures.

3.5.3 Probing with Direct Comparative Prompts

Finally, we use explicit pairwise comparisons to assess whether models recognize similarities and differences in moral judgments across clusters. For each topic, we apply hierarchical clustering to the survey-based scores, then isolate the two most divergent clusters (those with the largest difference in mean moral score). We repeatedly select representative countries from each cluster and prompt the model using a structure of the form:

Regarding the morality of {topic}, the judgments of people in {country x} and {country y} are {comparative token}.

where {comparative token} takes one of the pairs from Table 1 (e.g., *similar/dissimilar*). We compare the model's predicted label, "similar" or "different", with the empirical classification (same cluster vs. different clusters). Confusion matrix metrics and a Chi-squared test of association then measure how closely the model's judgments align with the survey-based groupings.

Model	WVS		PEW	
	r	*p*	*r*	*p*
GPT-2 Medium	-0.195	0.424	-0.090	0.832
GPT-2 Large	-0.115	0.640	0.617	0.103
OPT-125	-0.035	0.887	-0.095	0.822
QWEN	-0.200	0.413	0.102	0.811
BLOOM	-0.118	0.631	0.608	0.110

Table 2: Correlation between topic variances (WVS and PEW) and model-generated moral score variances. None of the correlations reach statistical significance (all $p > 0.05$).

Source	WVS		PEW	
	Mean score	*Var.*	*Mean score*	*Var.*
Empirical	-0.576	0.075	-0.244	0.138
BLOOM	0.474	0.004	0.246	0.006
OPT-125	0.104	0.012	0.248	0.027
QWEN	0.242	0.021	0.221	0.019
GPT-2 Large	0.323	0.015	0.160	0.032
GPT-2 Medium	0.411	0.013	0.227	0.024

Table 3: Mean moral scores and variances for WVS and PEW topics compared with model-generated values.

4 Results

To evaluate how well language models capture cross-cultural moral variability, we compare the topic-level variance from two survey datasets (WVS and PEW) with the variance of the corresponding model-generated moral scores. Table 2 summarizes the Pearson correlation (r) values, with associated p-values, for each model across both datasets.

WVS Variance Correlations. The weak negative correlations for all models on the WVS dataset indicate that the model-generated variance does not align with the observed cross-cultural diversity in these topics. Specifically, there is no statistically significant evidence that LLMs capture the degree of controversy reflected in WVS responses. The negative but insignificant correlations highlight how these models fall short in capturing the full range of intercultural nuance.

PEW Variance Correlations. On the PEW dataset, correlations are slightly more favorable for GPT-2 Large ($r = 0.617$) and BLOOM ($r = 0.608$), suggesting a somewhat better capability to capture topic-level variability. However, even these moderate-to-strong relationships do not achieve statistical significance. In sum, no model consistently reproduces the magnitude of cross-cultural disagreement measured by the PEW data.

Table 3 compares the empirical mean moral scores and variance with those generated by each model. We observe a consistent tendency across both WVS and PEW for the models to assign higher mean moral scores (i.e., more *morally acceptable*) and systematically lower variance than in the survey data. This pattern underscores the models' tendency to view topics as more morally approved and less controversial than they are in reality.

These lower variances suggest that the models underestimate the degree of cultural disagreement, especially on polarizing issues such as sexuality and family-related norms.

Figures 3 and 4 illustrate the mismatch between empirical and model-inferred moral variance. Although full rankings for each model are provided in the appendix, Table 4 summarizes the top three most controversial and most agreed-upon topics, based on *empirical* data from WVS and PEW.

From Table 4, *sex before marriage* and *homosexuality* rank among the most polarizing topics in both datasets, with variances of 0.219

WVS		PEW	
Topic	*Var.*	*Topic*	*Var.*
Most controversial			
Sex before marriage	0.219	Sex between unmarried adults	0.268
Homosexuality	0.209	Homosexuality	0.216
Euthanasia	0.126	Drinking alcohol	0.157
Most agreed upon			
Stealing property	0.015	Married people having an affair	0.021
Violence against other people	0.015	Using contraceptives	0.086
For a man to beat his wife	0.018	Gambling	0.097

Table 4: Top three most controversial and most agreed-upon topics from WVS (left) and PEW (right) empirical data.

and 0.209 respectively in WVS, and 0.268 and 0.216 in PEW. These high variances indicate substantial cross-cultural disagreement on these topics, which aligns with prior literature suggesting that sexual and family-related moral issues often reflect deep cultural differences between societies that prioritize individual autonomy versus those emphasizing communal values and traditional norms [6, 25]. The fact that these particular topics show the highest variance suggests they serve as key differentiators between moral frameworks across cultures. However, several models misjudge at least one of these issues as relatively uncontroversial, with QWEN and BLOOM even ranking homosexuality among their most agreed-upon topics (as shown in the appendix), suggesting they fail to capture these fundamental cultural divisions.

Although GPT-2 Large and BLOOM show moderate correlations in the PEW dataset (Table 2), no model achieves statistically significant alignment with the empirical data. Across both WVS and PEW, language models:

1. Overestimate moral acceptability, assigning more positive moral judgments to most topics.
2. Underestimate the degree of cultural disagreement, producing lower variance scores.

These findings suggest that current LLMs do not yet mirror real-world moral heterogeneity, especially for hotly debated topics like sexual and family norms. Simply increasing model size may be insufficient; more nuanced training or alignment with culturally diverse data sources may be necessary to capture the complexity seen in empirical moral attitudes.

4.1 Cluster Alignment

We analyze how closely the clusters induced by model-generated moral scores align with the empirical clusters derived from both the WVS and PEW datasets. Three metrics are used to measure this alignment: the Adjusted Rand Index (ARI), the Adjusted Mutual Information (AMI), and the Combined Alignment Score (CAS). Higher values on these metrics suggest better agreement between the empirical clusters and the model-generated clusters.

Table 6 combines the alignment scores for all topics in WVS (left panel) and PEW (right panel). QWEN shows notably higher metrics on WVS than the other models, indicating closer alignment to empirical scores. For PEW, GPT-2 Large and OPT-125 share moderate alignment scores, while QWEN and BLOOM perform relatively worse.

Table 7 shows the alignment results for the most controversial topics in both WVS and PEW. All models yield negative or near-zero alignment on WVS. On PEW, GPT-2 Medium remains negative while GPT-2 Large, OPT-125, and BLOOM achieve positive scores, with OPT-125 notably highest.

Table 8 reports the alignment results for the most agreed-upon topics. For WVS, GPT-2 Medium and OPT-125 have positive scores, whereas GPT-2 Large, QWEN, and BLOOM remain negative. For PEW, GPT-2 Medium, GPT-2 Large, and OPT-125 show moderate positive alignment; QWEN and BLOOM exhibit minimal or negative scores.

4.2 Probing with Direct Comparative Prompts

We further examine how models recognize similarities and differences in moral judgments by prompting them to compare topics directly. Tables 9 and 10 show, respectively, the confusion-matrix scores and chi-squared results for WVS (left) and PEW (right).

Accuracy for all models hovers near 0.5. GPT-2 Large and QWEN stand out with high recall (0.946 and 0.831, respectively), but their precision is lower, yielding moderate F1 scores. BLOOM displays poor performance across most metrics, indicating difficulties in classifying positive and negative instances.

Again, overall accuracy remains near 0.5 for all models. GPT-2 Large shows the highest recall (0.954), while QWEN achieves a recall of 0.694. BLOOM exhibits very low recall and precision, resulting in the lowest F1.

Table 10 shows that GPT-2 Medium exhibits a significant ($p < 0.01$) alignment with WVS, implying its judgments correlate with actual moral (dis)similarities. The other models do not significantly align with WVS. For PEW, BLOOM yields a statistically significant ($p < 0.05$) result—though this may reflect a consistent but incorrect pattern, given its poor F1 and recall.

Although some models (e.g., GPT-2 Large, QWEN) display high recall indirect probing, their precision is often lacking. GPT-2 Medium is uniquely significant in the WVS chi-squared test, while BLOOM is significant in the PEW test but shows low classification performance overall. These divergences suggest that while models capture certain aspects of moral similarity, they struggle to reflect the full complexity of real-world intercultural judgments.

5 Discussion and conclusion

The findings of this study shed light on the capability of LLMs to accurately capture cultural diversity and common tendencies across different moral topics. The investigation utilized multiple methodologies that were based on probing LLMs with prompts derived from the World Values Survey (WVS) and PEW datasets, focusing on a range of moral topics.

5.1 Comparison of variance

The correlation analysis between model-generated moral scores and empirical survey data revealed mixed results. For the PEW dataset, GPT-2 Large and BLOOM demonstrated moderate to strong alignment in capturing cultural variations. The fact that the largest model (GPT-2 Large) and the largest multilingual model (BLOOM) performed best may suggest that model size and multilinguality have a positive effect on models' ability to grasp patterns of cultural diversity, which would be in line with previous work from Du et al. [5] and Arora et al. [1]. However, the correlations did not reach statistical significance and therefore no strong claims can be made. Moreover, model performance shows high variability, with weak negative correlations observed for both GPT-2 Large and BLOOM when comparing their variances with the WVS moral score variances. The other models performed weakly and variably in both the PEW and WVS moral

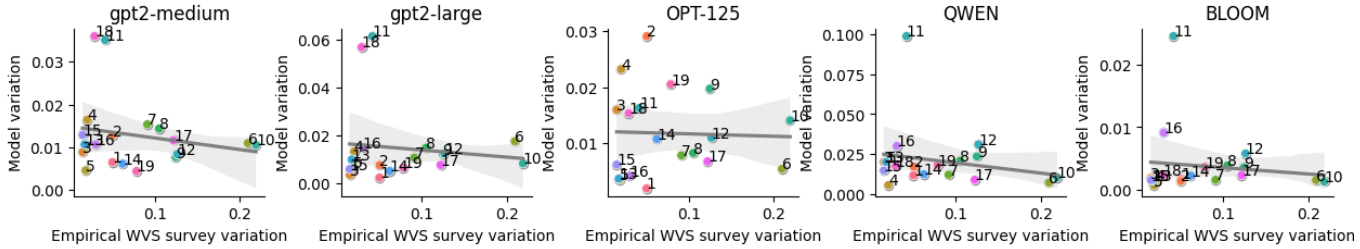


Figure 3: Comparison of empirical and model-inferred moral score variances for WVS topics. The models underestimate cross-cultural disagreements.

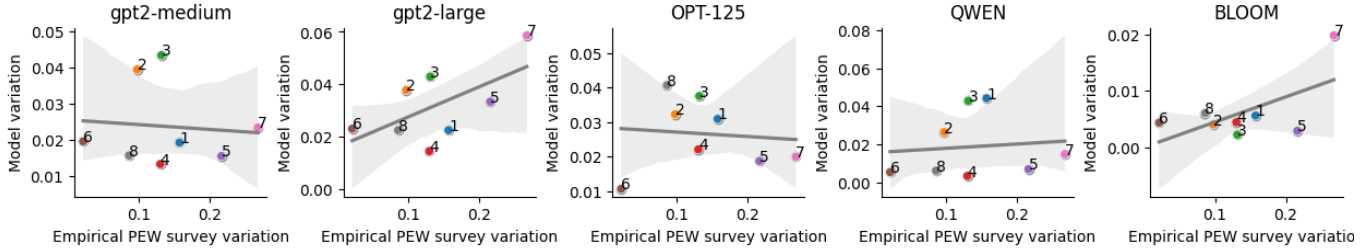


Figure 4: Comparison of empirical and model-inferred moral score variances for PEW topics. Again, models generally exhibit lower variance.

score variance comparisons. Furthermore, the models struggled to accurately identify the most controversial and agreed on topics. In fact, some of the models incorrectly categorized (one of) the two most controversial topics as among the most agreed on. The variable and low overall performance could be attributed to the fact that the complexity and nuance of moral values across different cultural contexts may not be fully captured by the models' training data.

5.2 Cluster alignment

The clustering alignment results further emphasized the variability in model performance. Overall, GPT-2 Large and OPT-125 showed better alignment with empirical moral scores from both datasets quite consistently, suggesting their relative proficiency in clustering countries based on moral attitudes. However, other models, most notably BLOOM, exhibited lower alignment scores, indicating shortcomings in their ability to mirror the clustering patterns observed in the survey data. These results suggest that the models fall short in grasping cultural patterns regarding moral judgments, which is in line with the findings from the previous method. Thus, while GPT-2 Large and OPT-125 generally show better alignment with empirical moral scores across various topics, the variability in model performance underscores the challenges in accurately capturing the complexities of moral attitudes across different cultural contexts. Overall, the clusterings based on the model scores do not faithfully capture the cultural patterns observed in the clusterings derived from the survey scores.

5.3 Probing with direct comparative prompts

Direct probing with comparative prompts provided additional insights into the models' understanding of moral differences between culturally distinct groups. In general, performance is low as the scores are no higher or even slightly lower than random chance. GPT-2 Large and QWEN stood out with higher accuracy and recall scores, indicating their better performance in distinguishing moral differences between the most divergent clusters identified by the survey data. Upon further inspection, however, it became clear that GPT-2 Large and QWEN almost always predict the same class, which does not signify a proper understanding of inter-cultural differences and similarities. If we disregard the performance of GPT-2 Large and QWEN due to the fact that they always predict the same class, GPT-2 Medium and OPT-125 exhibit the most balanced performance across the remaining models. BLOOM exhibited the lowest performance metric scores, suggesting challenges in discerning nuanced moral judgments across cultures. Notably, despite its low overall performance, BLOOM's judgments were found to be statistically associated with the judgments based on the PEW dataset through a Chi-squared test. This suggests that there may be some alignment between BLOOM's outputs and the moral judgments reflected in the PEW dataset. However, it is important to note that this statistical association does not necessarily imply a meaningful understanding or accurate representation of moral differences between cultures.

5.4 Conclusion

In conclusion, the study underscores the importance of rigorous evaluation methodologies when assessing LLMs' ability to understand and reflect cultural diversity in moral judgments. The tested models seem to propagate a homogenized view on cross-cultural moral values, identifying most topics as cross-culturally agreed on as more morally acceptable than empirically observed. Thereby, the models generally seem to reflect a rather liberal view, in line with the autonomy-endorsing values found in W.E.I.R.D. societies [6]. It has been established in the literature that exclusively English training data plays a big part in the embedding of homogenous W.E.I.R.D. values and, thereby, cultural bias in LLMs [3]. This could lead one to believe that multilingual LLMs are the answer to mitigating bias in LLMs [1]. However, this study could not find convincing evidence to suggest that multilingual models are better at truthfully capturing cultural diversities in moral judgments than monolingual models. Similarly, while model size could be considered another factor influencing model performance due to its potential to enhance computational capacity and capture more complex patterns [5], its impact was not found to be convincing in the carried out analyses. It can be concluded that this study found no remarkable differences between the tested models in their success, regardless of multilinguality or model size. Overall, the models examined show variable performance and generally exhibit

Model	Survey	Survey Var.	Survey Mean	Model Var.	Model Mean	Topic	Var. Diff
GPT-2 Medium	WVS	0.219	-0.244	0.011	0.465	sex before marriage	0.208
	WVS	0.209	-0.396	0.011	0.577	homosexuality	0.198
	WVS	0.126	-0.430	0.008	0.481	euthanasia	0.118
	WVS	0.125	-0.150	0.008	0.217	divorce	0.117
	WVS	0.122	-0.452	0.012	0.371	having casual sex	0.110
	PEW	0.268	-0.219	0.023	0.044	sex between unmarried adults	0.244
	PEW	0.216	-0.342	0.016	0.641	homosexuality	0.201
	PEW	0.157	-0.234	0.019	0.142	drinking alcohol	0.138
GPT-2 Large	WVS	0.219	-0.244	0.008	0.454	sex before marriage	0.211
	WVS	0.209	-0.396	0.018	-0.086	homosexuality	0.192
	WVS	0.122	-0.452	0.008	0.470	having casual sex	0.114
	WVS	0.126	-0.430	0.013	0.261	euthanasia	0.114
	WVS	0.125	-0.150	0.012	0.121	divorce	0.112
	PEW	0.268	-0.219	0.059	-0.138	sex between unmarried adults	0.209
	PEW	0.216	-0.342	0.033	-0.188	homosexuality	0.183
	PEW	0.157	-0.234	0.023	0.210	drinking alcohol	0.135
OPT-125	WVS	0.219	-0.244	0.014	0.475	sex before marriage	0.205
	WVS	0.209	-0.396	0.005	0.255	homosexuality	0.204
	WVS	0.126	-0.430	0.011	0.013	euthanasia	0.115
	WVS	0.122	-0.452	0.007	0.093	having casual sex	0.115
	WVS	0.125	-0.150	0.020	-0.261	divorce	0.105
	PEW	0.268	-0.219	0.020	0.512	sex between unmarried adults	0.248
	PEW	0.216	-0.342	0.019	0.570	homosexuality	0.198
	PEW	0.157	-0.234	0.031	0.187	drinking alcohol	0.126
QWEN	WVS	0.219	-0.244	0.010	0.415	sex before marriage	0.209
	WVS	0.209	-0.396	0.007	0.466	homosexuality	0.202
	WVS	0.122	-0.452	0.009	0.177	having casual sex	0.113
	WVS	0.125	-0.150	0.024	-0.042	divorce	0.101
	WVS	0.126	-0.430	0.031	-0.115	euthanasia	0.095
	PEW	0.268	-0.219	0.015	0.494	sex between unmarried adults	0.253
	PEW	0.216	-0.342	0.007	0.562	homosexuality	0.209
	PEW	0.130	-0.405	0.004	0.130	having an abortion	0.127
BLOOM	WVS	0.219	-0.244	0.001	0.662	sex before marriage	0.218
	WVS	0.209	-0.396	0.002	0.865	homosexuality	0.208
	WVS	0.124	-0.150	0.004	0.569	divorce	0.121
	WVS	0.126	-0.429	0.006	0.712	euthanasia	0.121
	WVS	0.122	-0.452	0.002	0.422	having casual sex	0.120
	PEW	0.268	-0.219	0.020	0.374	sex between unmarried adults	0.248
	PEW	0.216	-0.342	0.003	0.843	homosexuality	0.213
	PEW	0.157	-0.234	0.006	0.159	drinking alcohol	0.152

Table 5: Variance gaps between survey data and model outputs (WVS vs. PEW), showing the top eight topic–model pairs with the largest differences. Full results in the Appendix.

Model	WVS			PEW		
	ARI	AMI	CAS	ARI	AMI	CAS
GPT-2 Medium	-0.012	-0.002	-0.007	0.087	0.068	0.078
GPT-2 Large	0.028	0.040	0.034	0.129	0.123	0.126
OPT-125	-0.073	0.037	-0.018	0.129	0.123	0.126
QWEN	0.291	0.138	0.215	-0.019	0.065	0.023
BLOOM	0.015	-0.011	0.002	0.008	-0.004	0.002

Table 6: Cluster alignment scores for all topics in WVS (left) and PEW (right).

Model	WVS			PEW		
	ARI	AMI	CAS	ARI	AMI	CAS
GPT-2 Medium	-0.015	-0.011	-0.013	-0.026	-0.019	-0.022
GPT-2 Large	-0.012	0.023	0.005	0.093	0.081	0.087
OPT-125	-0.021	0.017	-0.002	0.131	0.140	0.136
QWEN	-0.014	-0.018	-0.016	-0.006	0.073	0.033
BLOOM	-0.015	-0.011	-0.013	0.009	0.006	0.007

Table 7: Cluster alignment scores for most controversial topics in WVS (left) and PEW (right).

low success in aligning with empirical moral data from global surveys. Thus, ongoing research and development are needed to enhance their accuracy and reliability in diverse cultural settings. Addressing these challenges is crucial for ensuring the ethical integrity and societal impact of AI technologies in the context of global applications.

Model	WVS			PEW		
	ARI	AMI	CAS	ARI	AMI	CAS
GPT-2 Medium	0.079	0.010	0.044	0.057	0.045	0.051
GPT-2 Large	-0.019	-0.014	-0.016	0.028	0.020	0.024
OPT-125	0.120	0.038	0.079	0.035	0.051	0.043
QWEN	-0.005	-0.017	-0.011	-0.020	-0.016	-0.018
BLOOM	-0.030	-0.012	-0.021	0.006	0.004	0.005

Table 8: Cluster alignment scores for most agreed-upon topics in WVS (left) and PEW (right).

Model	WVS				PEW			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
GPT-2 Medium	0.485	0.488	0.336	0.398	0.495	0.494	0.402	0.444
GPT-2 Large	0.509	0.508	0.946	0.661	0.495	0.497	0.954	0.654
OPT-125	0.502	0.510	0.461	0.484	0.506	0.506	0.480	0.493
QWEN	0.500	0.504	0.831	0.628	0.493	0.495	0.694	0.578
BLOOM	0.495	0.543	0.026	0.050	0.497	0.326	0.006	0.011

Table 9: Confusion matrix scores from direct probing on WVS (left) and PEW (right). Acc. = accuracy, Prec. = precision, Rec. = recall.

Model	WVS		PEW	
	χ^2	p	χ^2	p
GPT-2 Medium	8.38	0.004**	0.418	0.518
GPT-2 Large	1.491	0.222	3.325	0.068
OPT-125	0.338	0.561	0.609	0.435
QWEN	1.416	0.234	1.017	0.313
BLOOM	1.279	0.258	4.599	0.032*

Table 10: Chi-squared test results from direct probing on WVS (left) and PEW (right). $(**)$ indicates $p < 0.01$, $(*)$ indicates $p < 0.05$.

5.5 Conclusion

In conclusion, the study underscores the importance of rigorous evaluation methodologies when assessing LLMs' ability to understand and reflect cultural diversity in moral judgments. The tested models seem to propagate a homogenized view on cross-cultural moral values, identifying most topics as cross-culturally agreed on as more morally acceptable than empirically observed. Thereby, the models generally seem to reflect a rather liberal view, in line with the autonomy-endorsing values found in Western, Educated, Industrialized, Rich, and Democratic (W.E.I.R.D.) societies [6]. It has been established in the literature that exclusively English training data plays a big part in the embedding of homogenous W.E.I.R.D. values and, thereby, cultural bias in LLMs [3]. This could lead one to believe that multilingual LLMs are the answer to mitigating bias in LLMs [1]. However, this study could not find convincing evidence to suggest that multilingual models are better at truthfully capturing cultural diversities in moral judgments than monolingual models.

Based on our findings, several actionable strategies could improve cultural representativeness in LLMs. First, diversifying training data by prioritizing text from underrepresented regions and languages would help counteract the current bias toward W.E.I.R.D. perspectives. This includes incorporating religious texts, local news sources, and cultural forums that discuss moral topics from non-W.E.I.R.D. societies. Second, culture-aware fine-tuning approaches could be developed using datasets that explicitly represent diverse moral perspectives on controversial topics, weighted to reflect actual global population distributions rather than internet data availability. Third, prompt engineering strategies that explicitly invoke cultural context could elicit more culturally diverse responses. For example, prompts like "From the perspective of someone in [country] with traditional values..." may help models access different moral frameworks. Finally, establishing standardized evaluation frameworks using surveys like WVS and PEW would enable regular assessment of cultural bias in new models before deployment. These recommendations provide concrete pathways for researchers and practitioners working toward more culturally inclusive AI systems. The challenges identified here align with broader issues in developing transparent and interpretable NLP systems across various domains [17], emphasizing the need for continued research in explainable AI methods.

6 Limitations

While this study provides important insights, it is important to recognize certain boundaries of our approach. First, although WVS and PEW are well-established surveys covering 55 and 39 countries respectively, they organize complex moral views into fixed categories, which may not capture every nuance or implicit aspect of moral reasoning. Additionally, our analysis examined aggregate patterns across all countries rather than country-specific contributions to variance. Future work could benefit from analyzing which specific countries or regions show the largest discrepancies between model outputs and survey responses, which would provide more granular insights into geographical patterns of model bias. Second, we focused on a selected group of models, so our findings primarily reflect these particular architectures. Third, the choice of prompts in our experiments can influence model responses [26], meaning that exploring alternative prompt strategies could yield additional insights. Lastly, due to computational limits, we randomly selected topics in Method 3, which may not cover all diversity within each cluster. Future research can build on our work by testing a wider range of models, experimenting with different prompt designs, analyzing country-specific patterns, and using broader topic sampling to further enrich the analysis.

7 Ethics Statement

The work relies exclusively on two publicly available, anonymised survey datasets (WVS Wave 7 and Pew 2013) and on open-access language models. No personal or sensitive information was collected, and all analyses were performed in accordance with the data providers' terms of use. By quantifying cultural bias in LLMs we aim to support fairer deployment of generative AI and to encourage the creation of more globally representative training data.

Acknowledgements

We appreciate the maintainers of WVS and PEW data for enabling large-scale cross-cultural analysis.

References

[1] A. Arora, L. Kaffee, and I. Augenstein. Probing pre-trained language models for cross-cultural differences in values. *arXiv preprint arXiv:2203.13722*, 2022. URL https://doi.org/10.48550/arxiv.2203.13722.

[2] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021. doi: 10.1145/3442188.3445922. URL https://dl.acm.org/doi/10.1145/3442188.3445922.

[3] Y. Benkler, D. Mosaphir, S. Friedman, A. Smart, and S. Schmer-Galunder. Assessing llms for moral value pluralism. *arXiv (Cornell University)*, 2023. doi: 10.48550/arxiv.2312.10075. URL https://doi.org/10.48550/arxiv.2312.10075.

[4] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Quantifying and reducing stereotypes in word embeddings. *arXiv*, 2016. doi: 10.48550/arxiv.1606.06121. URL https://arxiv.org/abs/1606.06121.

[5] X. Du, Z. Yu, S. Gao, D. Pan, Y. Cheng, Z. Ma, R. Yuan, X. Qu, J. Liu, T. Zheng, X. Luo, G. Zhou, B. Yuan, W. Chen, J. Fu, and G. Zhang. Chinese Tiny LLM: Pretraining a Chinese-Centric Large Language Model. *arXiv (Cornell University)*, 4 2024. doi: 10.48550/arxiv.2404.04167. URL https://arxiv.org/abs/2404.04167.

[6] J. Graham, P. Meindl, E. Beall, K. M. Johnson, and L. Zhang. Cultural differences in moral judgment and behavior, across and within societies. *Current Opinion in Psychology*, 8:125–130, 2016. doi: 10.1016/j.copsyc.2015.09.007. URL https://doi.org/10.1016/j.copsyc.2015.09.007.

[7] C. W. Haerpfer, P. Bernhagen, R. F. Inglehart, and C. Welzel. *World Values Survey: Round Seven - Country-Pooled Datafile Version*. Institute for Comparative Survey Research, Vienna, 2022. URL http://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp.

[8] J. Haidt. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814–834, 2001. doi: 10.1037/0033-295X.108.4.814.

[9] J. Henrich, S. J. Heine, and A. Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

[10] R. L. Johnson, G. Pistilli, N. Menédez-González, L. D. D. Duran, E. Panai, J. Kalpokiene, and D. J. Bertulfo. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv.org*, mar 2022. URL https://arxiv.org/abs/2203.07785.

[11] K. Karpouzis. Plato's shadows in the digital cave: Controlling cultural bias in generative ai. *Electronics*, 13(8):1457, 2024. doi: 10.3390/electronics13081457. URL https://doi.org/10.3390/electronics13081457.

[12] J. Kharchenko, T. Roosta, A. Chadha, and C. Shah. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. arXiv preprint, 2024. URL https://doi.org/10.48550/arxiv.2406.14805. arXiv:2406.14805.

[13] D. Lazarenko and T. Bonald. Pairwise adjusted mutual information, 2021. URL https://arxiv.org/abs/2103.12641.

[14] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. d. Moral, O. Ruwase, R. Bawden, and M. J. ... Nelson. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100, 2022. URL https://api.semanticscholar.org/CorpusID:253420279.

[15] C. C. Liu, F. Koto, T. Baldwin, and I. Gurevych. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *arXiv*, 2023. URL https://arxiv.org/abs/2309.08591. arXiv:2309.08591.

[16] A. Mishra, G. Nayak, S. Bhattacharya, T. Kumar, A. Shah, and M. Foltin. Llm-guided counterfactual data generation for fairer ai. In *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24, page 1538–1545, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701726. doi: 10.1145/3589335.3651929. URL https://doi.org/10.1145/3589335.3651929.

[17] H. Mohammadi, A. Bagheri, A. Giachanou, and D. L. Oberski. Explainability in practice: A survey of explainable nlp across various domains. *arXiv preprint arXiv:2502.00837*, 2025.

[18] H. Mohammadi, E. Papadopoulou, Y. F. Meijer, and A. Bagheri. Exploring cultural variations in moral judgments with large language models. *arXiv preprint arXiv:2506.12433*, 2025.

[19] H. Mohammadi, T. Shahedi, P. Mosteiro, M. Poesio, A. Bagheri, and A. Giachanou. Assessing the reliability of llms annotations in the context of demographic bias and model explanation. *arXiv preprint arXiv:2507.13138*, 2025.

[20] T. Nazaretsky, S. Hershkovitz, and G. Alexandron. Kappa learning: A new item-similarity method for clustering educational items from response data. 04 2020. URL https://eric.ed.gov/?id=ED599209.

[21] P. Nemani, Y. D. Joel, P. Vijay, and F. F. Liza. Gender bias in transformers: A comprehensive review of detection and mitigation strategies. *Natural Language Processing Journal*, 6:100047, 2024. doi: 10.1016/j.nlp.2023.100047. URL https://doi.org/10.1016/j.nlp.2023.100047.

[22] S. U. Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, New York, NY, 2018. ISBN 978-1479837243. URL https://nyupress.org/9781479837243/algorithms-of-oppression/.

[23] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

[24] A. Ramezani and Y. Xu. Knowledge of cultural moral norms in large language models. *arXiv (Cornell University)*, 2023. doi: 10.48550/arxiv.2306.01857. URL https://doi.org/10.48550/arxiv.2306.01857.

[25] R. A. Shweder, N. C. Much, M. Mahapatra, and L. Park. The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. In A. Brandt and P. Rozin, editors, *Morality and Health*, pages 119–169. Routledge, 1997. URL https://psycnet.apa.org/record/1997-36447-005.

[26] L. Wang, X. Chen, and X. Deng. Prompt engineering in consistency and reliability with the evidence-based guideline for llms. *npj Digital Medicine*, 7:41, 2024. doi: 10.1038/s41746-024-01029-4. URL https://doi.org/10.1038/s41746-024-01029-4.

[27] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[28] J. Zou and L. Schiebinger. Ai can be sexist and racist—it's time to make it fair. *Nature*, 559(7714):324–326, 2018. doi: 10.1038/d41586-018-05707-8. URL https://www.nature.com/articles/d41586-018-05707-8.

A Appendix

A.1 Most controversial WVS topics according to models

Topic	*Model*	*Variance*
Political violence	GPT-2 Medium	0.036
Suicide	GPT-2 Medium	0.035
Cheating on taxes	GPT-2 Medium	0.016

Table 11: Top 3 most controversial WVS topics according to GPT-2 Medium

Topic	*Model*	*Variance*
Suicide	GPT-2 Large	0.062
Political violence	GPT-2 Large	0.057
Homosexuality	GPT-2 Large	0.018

Table 12: Top 3 most controversial WVS topics according to GPT-2 Large

Topic	*Model*	*Variance*
Avoiding a fare on public transport	OPT-125	0.029
Cheating on taxes	OPT-125	0.023
Death penalty	OPT-125	0.021

Table 13: Top 3 most controversial WVS topics according to OPT-125

Topic	*Model*	*Variance*
Suicide	QWEN	0.099
Terrorism as a political, ideological or religious tactic	QWEN	0.030
Euthanasia	QWEN	0.031

Table 14: Top 3 most controversial WVS topics according to QWEN

Topic	*Model*	*Variance*
Suicide	BLOOM	0.025
Terrorism as a political, ideological or religious tactic	BLOOM	0.009
Euthanasia	BLOOM	0.006

Table 15: Top 3 most controversial WVS topics according to BLOOM

A.2 Most agreed on WVS topics according to models

Topic	*Model*	*Variance*
Death penalty	GPT-2 Medium	0.004
Accepting a bribe in the course of duty	GPT-2 Medium	0.005
Parents beating children	GPT-2 Medium	0.006

Table 16: Top 3 most agreed on WVS topics according to GPT-2 Medium

Topic	*Model*	*Variance*
Claiming government benefits to which you are entitled	GPT-2 Large	0.002
Stealing property	GPT-2 Large	0.004
Parents beating children	GPT-2 Large	0.005

Table 17: Top 3 most agreed on WVS topics according to GPT-2 Large

Topic	*Model*	*Variance*
Claiming government benefits to which you are entitled	OPT-125	0.002
Someone accepting a bribe in the course of duty	OPT-125	0.003
For a man to beat his wife	OPT-125	0.004

Table 18: Top 3 most agreed on WVS topics according to OPT-125

Topic	*Model*	*Variance*
Cheating on taxes	QWEN	0.006
Homosexuality	QWEN	0.007
Having casual sex	QWEN	0.009

Table 19: Top 3 most agreed on WVS topics according to QWEN

Topic	*Model*	*Variance*
Someone accepting a bribe in the course of duty	BLOOM	0.001
Sex before marriage	BLOOM	0.001
Avoiding a fare on public transport	BLOOM	0.001

Table 20: Top 3 most agreed on WVS topics according to BLOOM

A.3 Most controversial PEW topics according to models

Topic	*Model*	*Variance*
Getting a divorce	GPT-2 Medium	0.043
Gambling	GPT-2 Medium	0.039
Sex between unmarried adults	GPT-2 Medium	0.023

Table 21: Top 3 most controversial PEW topics according to GPT-2 Medium

Topic	*Model*	*Variance*
Sex between unmarried adults	GPT-2 Large	0.059
Getting a divorce	GPT-2 Large	0.043
Gambling	GPT-2 Large	0.038

Table 22: Top 3 most controversial PEW topics according to GPT-2 Large

Topic	*Model*	*Variance*
Using contraceptives	OPT-125	0.041
Getting a divorce	OPT-125	0.038
Gambling	OPT-125	0.032

Table 23: Top 3 most controversial PEW topics according to OPT-125

Topic	*Model*	*Variance*
Drinking alcohol	QWEN	0.044
Getting a divorce	QWEN	0.043
Gambling	QWEN	0.027

t

Table 24: Top 3 most controversial PEW topics according to QWEN

Topic	*Model*	*Variance*
Sex between unmarried adults	BLOOM	0.020
Using contraceptives	BLOOM	0.006
Drinking alcohol	BLOOM	0.006

Table 25: Top 3 most controversial PEW topics according to BLOOM

A.4 Most agreed on PEW topics according to models

Topic	*Model*	*Variance*
Having an abortion	GPT-2 Medium	0.013
Homosexuality	GPT-2 Medium	0.016
Using contraceptives	GPT-2 Medium	0.016

Table 26: Top 3 most agreed on PEW topics according to GPT-2 Medium

Topic	*Model*	*Variance*
Having an abortion	GPT-2 Large	0.015
Using contraceptives	GPT-2 Large	0.023
Drinking alcohol	GPT-2 Large	0.023

Table 27: Top 5 most agreed on PEW topics according to GPT-2 Large

Topic	*Model*	*Variance*
Married people having an affair	OPT-125	0.011
Homosexuality	OPT-125	0.019
Sex between unmarried adults	OPT-125	0.020

Table 28: Top 3 most agreed on PEW topics according to OPT-125

Topic	*Model*	*Variance*
Having an abortion	QWEN	0.004
Married people having an affair	QWEN	0.006
Using contraceptives	QWEN	0.006

Table 29: Top 3 most agreed on PEW topics according to QWEN

Topic	*Model*	*Variance*
Getting a divorce	BLOOM	0.002
Homosexuality	BLOOM	0.003
Gambling	BLOOM	0.004

Table 30: Top 3 most agreed on PEW topics according to BLOOM