Probing Pre-Trained Language Models for Cross-Cultural Differences in Values

Arnav Arora and **Lucie-Aimée Kaffee** and **Isabelle Augenstein**
University of Copenhagen
{aar, kaffee, augenstein}@di.ku.dk

Abstract

Language embeds information about social, cultural, and political values people hold. Prior work has explored potentially harmful social biases encoded in Pre-trained Language Models (PLMs). However, there has been no systematic study investigating how values embedded in these models vary across cultures. In this paper, we introduce probes to study which cross-cultural values are embedded in these models, and whether they align with existing theories and cross-cultural values surveys. We find that PLMs capture differences in values across cultures, but those only weakly align with established values surveys. We discuss implications of using mis-aligned models in cross-cultural settings, as well as ways of aligning PLMs with values surveys.

1 Introduction

A person's identity, values and stances are often reflected in the linguistic choices one makes (Jaffe, 2009; Norton, 1997). This is why, when language models are trained on large text corpora, they not only learn to understand language, but also pick up on a variety of societal and cultural biases (Stanczak et al., 2021). While biases picked up by the PLMs have a potential to cause harm when used in a downstream application, they may also serve as tools which provide insights into understanding cultural phenomena. Further, while studying ways of surfacing and mitigating potentially harmful biases is an active area of research, cultural biases and values picked up by PLMs remain understudied. Here, we investigate cultural values and differences among them picked up by PLMs through their pre-training on Web text.

In a wide range of social science research fields, values are a crucial tool for understanding cross-cultural differences. As defined by Rokeach (2008), values are the "core conceptions of the desirable within every individual and society", i.e., the foundation for the beliefs guiding a persons actions and

on a society level the base for the guiding principles. We would like to highlight the difference we make between *values* and *morals*. The former, as conceptualised in this work, is concerned with fundamental beliefs an individual or a group holds towards socio-cultural topics, whereas the latter entails making a judgement towards individual or collective right or wrong. For a discussion around the intersection of morality and PLMs, we point the reader to Talat et al. (2021). In this paper, we base our understanding of values across cultures on two studies: Hofstede (2005), which defines 6 dimensions to describe cross-cultural differences in values, and the World Values Survey (WVS) (Haerpfer et al., 2022). Both surveys provide numerical value scores for several categories on a population level across different countries and regions and are widely used to understand cross-cultural differences in values.

PLMs are trained on large amounts of text from the Web and have shown to pick up on semantic, syntactical, factual and other forms of knowledge which allow them to perform well across several Natural Language Processing (NLP) tasks. Since multilingual PLMs are trained on text in many languages, they have the potential to pick up cultural values through word associations expressed in those languages which are embedded in the pre-training texts. We therefore measure whether cultural values embedded in multilingual PLMs are correlated with the ones provided by the surveys. In Wikipedia, which is one of the primary sources of training data for multilingual PLMs, cross-cultural differences have been established (Miquel-Ribé and Laniado, 2019), and analysed by Hara et al. (2010) based on Hofstede's theory.

In this paper, we explore the novel research question of whether PLMs capture cultural differences in terms of values across different language models. We probe PLMs using questions from the values surveys of both Hofstede's cultural dimensions the-

ory and the World Values Survey. We reformulate the survey questions to probes and extract the answers to evaluate whether language models can capture cultural differences based on their training data. We focus on 13 languages, each of which is primarily geographically restricted to one country or region, to compare the results of the language models to the values surveys. The overall experimental setting for the paper is outlined in Figure 1.

Our work makes the following **contributions**[1]:

- We present the first study measuring cultural values embedded in large Pre-trained Language Models
- We propose a methodology for probing for values by converting survey questions to cloze style questions
- We conduct experiments across 13 languages with three multilingual language models (mBERT, XLM, and XLM-R), showing value alignment correlations with two large scale values surveys
- We present a discussion around potential implications of deploying these models in a multi-cultural context

2 Related Work

Expression and Norms   Analysis of expression of identity and attitudes through language and its change has a long history in sociolinguistics (Labov, 1963; Trudgill, 2002). More recently, studies have used NLP to computationally analyse this change on social media data (Eisenstein et al., 2014; Hovy et al., 2015) and link it to external factors like socioeconomic status (Abitbol et al., 2018) and demographics (Jurgens et al., 2017). This has also been done to analyse broader societal trends like temporal change in attitudes towards sexuality (CH-Wang and Jurgens, 2021) and gender bias (Sap et al., 2017; Stanczak and Augenstein, 2021). Further, there has been work on creating resources to analyse social norms and common-sense reasoning around them (Forbes et al., 2020; Emelin et al., 2021; Sap et al., 2020). Hoover et al. (2020); Roy et al. (2021) present work on extracting moral sentiment embedded in language using the Moral Foundation Theory. To diversify visually grounded reasoning across different cultures, Liu

---

[1]The code and data used for our experiments can be found here.

et al. (2021) introduce a multimodal multilingual dataset.

While there has been work on investigating and embedding social and moral norms, understanding values and their variation in a cross-cultural context remains understudied in the literature. Kiesel et al. (2022) provide a taxonomy of 54 values based on Schwartz et al. (2012) and provide a dataset and baselines for automatic value classification within the context of argument mining. The closest setup to ours would be one adopted by Johnson et al. (2022). They qualitatively assess the text generated by GPT-3, an autoregressive language model, by prompting it with English texts with a clear embedded value. They find that the embedded values in the generated texts were altered to be more in line with dominant values of US citizens, possibly due to its training data. Our setup instead quantitatively measures whether cross-cultural differences in these values are preserved in multilingual language models when fed with the language spoken predominantly by people belonging to that culture.

Probing   Probing has been extensively used as tool to study a variety of knowledge and biases picked up by PLMs. This can be syntactic (Hewitt and Manning, 2019), semantic (Vulić et al., 2020), numerical (Wallace et al., 2019), relational (Petroni et al., 2019) or factual knowledge (Jiang et al., 2020) picked up by PLMs. Probes can be created on both, at the word or sentence level (Mosbach et al., 2020).

Following work (Caliskan et al., 2017; Garg et al., 2018) on studying gender bias in word embeddings, a number of studies have built on it to similarly probe for social biases embedded in PLMs (May et al., 2019; Guo and Caliskan, 2021; Stańczak et al., 2021; Ousidhoum et al., 2021; de Vassimon Manela et al., 2021; Stanczak et al., 2021). This can be done using cloze-style probing for measuring at an intra-sentence level (Nadeem et al., 2021) or using pseduo-log likelihood (Salazar et al., 2020) based scoring (Nangia et al., 2020). There are downsides to both approaches, the former potentially introduces unintended bias based on the tokens in the input probe while the latter assumes that all masked tokens are statistically independent (Kaneko and Bollegala, 2022). We choose the former since the probes in our case are carefully worded by social scientists with the explicit aim to extract bias towards a certain set of values.

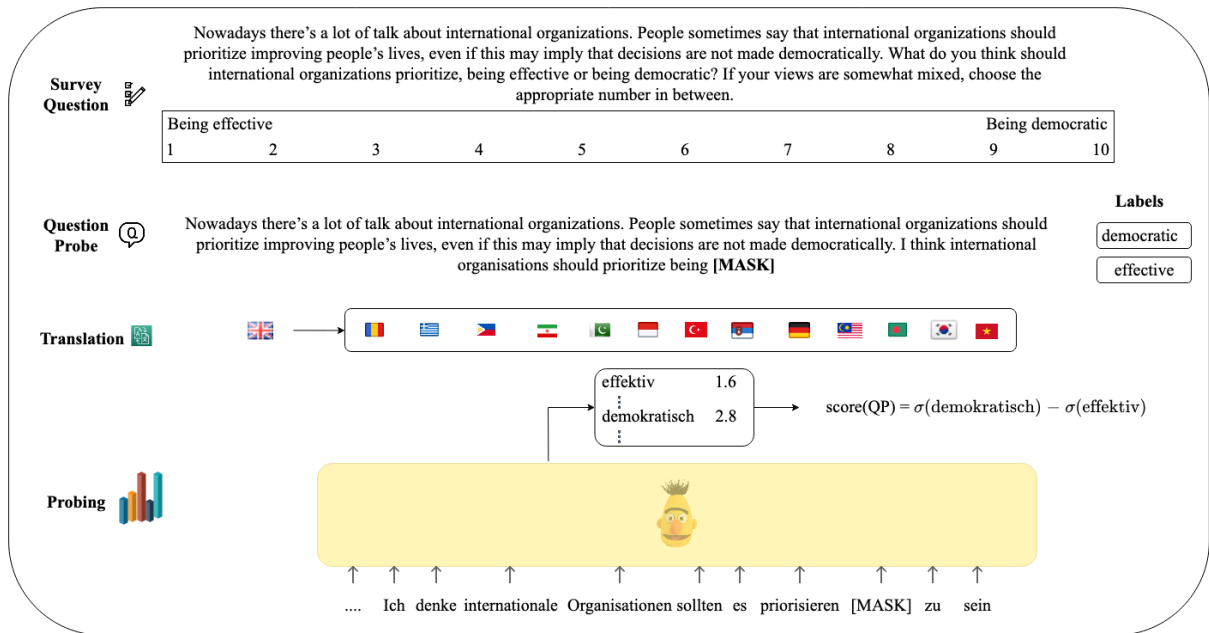To the best of our knowledge, there is no existing

Figure 1: Figure outlining the experimental setting for the paper. We take the original survey questions (Section 4), convert them into Question Probes and translate these into the target languages (Section 5) and run inference on the mask probes (Section 6.2)

work on probing for values embedded in PLMs in a comparative cultural context.

3 Value Probing

In this paper, we explore how PLMs capture differences in values across cultures, and whether those differences reflect the ones found in values across cultures at large. To compare the PLMs' encodings of values, we compare them with established surveys capturing cross-cultual differences in values, namely Hofstede's cultural dimensions theory and the World Values Survey (WVS) (Section 4). We transform the survey questions introduced in those surveys for compatibility with PLMs by reformulating them semi-automatically to convert them into probes (Section 5). Then we translate these created probes from English into 13 geographically localised languages to conduct cultural value probing across 13 cultures (Section 5). Finally, we assess the variance in cross-cultural values embedded in PLMs and compare the probing results to the established values surveys in Section 7.

We investigate the following **research questions** as a first step to exploring this novel area of probing cross-cultural differences in values:

RQ1 Do PLMs capture diversity across cultures for the established values?

RQ2 Are there similarities in the embedded values

across different PLMs?

RQ3 Do values embedded in PLMs align with existing values surveys?

4 Values Surveys

We base our work on previous studies on how values differ across cultures. As these are central to a number of research fields including political science, psychology, sociology, behavioral economics, cross-cultural research among others, there are a large number of studies utilising the scores provided by these values surveys. Among the most common ones are Hofstede's cultural dimensions theory and the World Values Survey. These studies build on the body of work in different fields: Hofstede's theory is derived from management studies (Hofstede, 1984), while the WVS was developed in the field of political science (Inglehart, 2006). Both studies have since been widely used across fields.

4.1 Hofstede's Cultural Dimensions Theory

Hofstede started his surveys of cross-cultural differences in values in 1980. This first survey (Hofstede, 1984) included 116,000 participants from 40 countries (extended to 111 countries and regions in the 2015 version) working with IBM, and created 4 cultural dimensions, which were subsequently extended to 6 cultural dimensions that are also used in

116

this paper. These 6 dimensions are: Power Distance (pdi), Individualism (idv), Uncertainty Avoidance (uai), Masculinity (mas), Long-term Orientation (lto), Indulgence (ivr). The full survey contains 24 questions. Each dimension is calculated using a formula defined by Hofstede using 4 of the questions in the survey, see Appendix F. Hofstede shows the influence that culture has on values by defining distinctly different numerical values in those 6 dimensions for the cultures observed. While critics of Hofstede's cultural dimensions theory point out, among others, the simplicity of the approach of mapping cultures to countries and question the timeliness of the approach (Nasif et al., 1991), this model of representing values is now a foundation for a large body of work on cross-cultural differences in values (Jones, 2007).

4.2 World Values Survey (WVS)

The World Values Survey (WVS, Haerpfer et al. (2022)) collects data on peoples' values across cultures in a more detailed way than Hofstede's cultural dimensions theory. The survey started in 1981 and is conducted by a nonprofit organisation, which includes a network of international researchers. It is conducted in *waves*, to collect data on how values change over time. The latest wave, wave 7, ran from 2017 to 2020. Compared to the European Values Study[2], WVS targets all countries and regions, and includes 57 countries. While Hofstede's cultural dimensions theory aggregates the findings of their survey into the 6 cultural dimensions, WVS publishes the results of their survey per question. Those are organised in 13 categories: (1) Social Values, Attitudes and Stereotypes, (2) Happiness and Well-being, (3) Social Capital, Trust and Organisational Membership, (4) Economic Values, (5) Corruption, (6) Migration, (7) Security, (8) Post-materialist Index, (9) Science and Technology, (10) Religious Values, (11) Ethical Values and Norms, (12) Political Interest and Political Participation, (13) Political Culture and Regimes.

We exclude categories (4) and (8) for the experiments in this study. This was done due to the nature of questions asked in these categories, for which it was not straightforward to design mask probes without loss of information.

Inglehart (2006), who established WVS, further defines the *Inglehart–Welzel cultural map*, which processes the surveys and defines two dimensions

in relation to each other: traditional versus secular-rational values and survival versus self-expression values, and summarise values for countries on a scatter plot describing these dimensions. In the following, we only use the previously mentioned 11 categories and leave an analysis based on the Inglehart–Welzel cultural map for future work.

5 Probe Generation

In order to make the surveys compatible with language models, we reformulate the survey questions to cloze-style question probes (Taylor, 1953; Hermann et al., 2015) that we can then perform masked language modelling inference on. Since this is the task PLMs were trained on, we argue it is a suitable methodology to measure embedded cultural biases in these models.

Hofstede's Cultural Dimensions Based on the English survey questions, the questions in the survey are manually reformulated to question probes (QPs). This is done analogously to iterative categorisation, in which a set of possible labels (y_i^+, y_i^-) corresponding to either end of the response options available in the survey are defined, which are the words the language models are probed for. The sentences are then reformulated to probes, and the labels masked. Those labels are based on the answers of the original survey, for instance, the original question like *have sufficient time for personal or home life* with answer options consisting of different degrees of *importance*, the probe is reformulated to *Having sufficient time for personal or home life is [MASK].*, where *[MASK]* should be replaced by *important* or *unimportant*.

$$\text{QPs} = \{W_i, y_i^+, y_i^-\}$$

where W_i is the masked probe and y_i^+ and y_i^- are the set of labels. There are a total of 24 questions with repeating labels.

World Values Survey Analogous to the probes created from the Hofstede survey, we create probes from the English questionnaire of the WVS. As there are more questions than for Hofstede (238 in total), there are also a larger number of labels to replace and a higher variety of question types.

Multilingual Probes To probe across several languages, we follow a semi-automatic methodology for translating the created probes in English to the

117

target language. We use a translation API[3] that covers all target languages. We translate each QP from English into the target language with the [MASK] token replaced by the label words $[y_i^+, y_i^-]$ in order to maintain grammatical structure and aid the translation API. One challenge of cross-cultural research is information loss when translating survey questions (Nasif et al., 1991; Hofstede, 1984). Therefore we opted for this approach rather than reformulating the translated survey questions by Hofstede. However, we would like to highlight the shortcomings of machine translation which have poor performance on low resource languages and has the potential to introduce additional biases. For the purpose of these experiments however, since the question probes are relatively simple sentences, we found the machine translations to be of high quality. We conducted an evaluation of our machine translated probes, the details for which can be found in the Appendix B. The target labels $[y_i^+, y_i^-]$ for each QP are then translated individually as single words (e.g. *important* is translated from English to the German *wichtig*), followed by lowercased string matching to check if the translated label can be found and replaced in the translated probe. If the target label cannot be found directly in the translated probe due to differences in word choice, we use a cross-lingual word aligner (Dou and Neubig, 2021) to align the English probe and its translated version. With this approach, we identify the label word to be replaced with the mask token. If both approaches yield no result, the token is manually replaced in the target sentence based on the authors' language understanding and using online translators.

Language Selection In total, we investigate 13 languages, mapped to one country each as outlined in Table 1, according to criteria further detailed below. One of the limitations of this one-to-one mapping is that the languages are spoken in wider regions and not specifically in one country (disregarding also e.g. diaspora communities). This allows for the closest match to the values theories we work with, which operate on a country level. The definition of culture by country has been criticised by, e.g., Nasif et al. (1991).

We select the languages as follows: We first include the countries covered in both the surveys of WVS and Hofstede. We limit to languages which are official languages of the countries observed in

Country	Language	Wikipedia size
Romania	Romanian (ro)	428,330
Greece	Greek (el)	207,647
Pakistan	Urdu (ur)	168,587
Iran	Farsi (fa)	872,240
Philippines	Tagalog (tl)	43,145
Indonesia	Indonesian (id)	618,395
Germany	German (de)	2,675,084
Malaysia	Malay (ms)	356,937
Bangladesh	Bengali (bn)	119,619
Serbia	Serbian (sr)	656,627
Turkey	Turkish (tr)	475,984
Vietnam	Vietnamese (vi)	1,270,712
South Korea	Korean (ko)	582,977

Table 1: Mapping of countries (cultures) to languages used throughout this paper, including number of articles per Wikipedia language as of March 2022.

the studies of both WVS and Hofstede. We further select languages for which the distribution of speakers is primarily localized to a country or relatively narrow geographical region. To ensure the language models will be able to have (potentially) sufficient amount of training data, from the set of languages, only those are selected which have at least 10,000 articles on Wikipedia.

6 Methodology

6.1 Models

We conduct the probing experiments on three widely used multilingual PLMs: the multi-lingual, uncased version of BERT base (mBERT) (Devlin et al., 2018), the 100 language, MLM version of XLM (Conneau and Lample, 2019), and the base version of XLM-RoBERTa (XLM-R) (Conneau et al., 2020) available in the Transformers (Wolf et al., 2020) library. mBERT was trained with a Masked Language Modelling (MLM) and Next Sentence Prediction objective, on Wikipedia articles in 102 languages with the highest number of articles on them. The XLM model builds on top of mBERT, only using the MLM objective but with modifications to the selection and truncation of training text fed to the model at each training step. It was also trained on Wikipedia texts, including 100 languages. The XLM-R model uses the RoBERTa architecture (Liu et al., 2019) and is trained with an MLM objective on 2.5 TB of filtered CommonCrawl corpus data in 100 languages.

---

[3] https://cloud.google.com/translate

It shows strong multilingual performance across a range of benchmarks and is commonly used for extracting multilingual sentence encodings.

6.2 Mask Probing

For each model M, we run inference on the created cloze-style question probes (QPs, described in Section 5) using an MLM head producing the log probabilities for the [MASK] tokens in the QPs over the entire vocabulary V of the respective model: $logP_M(w_i, t|W_i^{\backslash t}, \Theta_M) \in R^{|V|}$, where t is the position of the [MASK] token in the text $W_i \in QP$, and Θ_M are the parameters of the corresponding Language Model M. Since the survey respondents have to answer the questions with a choice between a range of values, for instance 1-10 with 1 representing *democratic* and 10 representing *effective*, in order to replicate a similar setting with PLMs, we subtract the predicted logit for the response label with the highest score w_i^+ with the predicted logit for the lowest score w_i^-. This normalises the predicted logits for the responses on opposing ends of the survey question and is then used as a score for that question.

$$logP_M(w_i) = logP_M(w_i^+) - logP_M(w_i^-)$$

Finally, in order to collapse the World Values Survey responses per *category*, within which many questions have different scales, we normalize the aggregate survey responses per the corresponding question scale, so that $y_{i,c} \in [0,1], c \in C$. We then take the mean of the responses across all the questions of the category to arrive at the aggregated score of the category for each country: $y_i = \frac{1}{|C|} \sum_{c \in C} y_{i,c} \in [0,1]$.

6.3 Evaluation

We calculate Spearman's ρ – a rank correlation coefficient between the values predicted by the language models and values calculated through the surveys: $\rho(logP_M(w_i, t|W_i^{\backslash t}, \Theta_M), y_i)$. For the World Values Survey, we do this per question, as well as per category. For Hofstede, we limit this calculation to value level correlations due to lack of access to individual or aggregate survey response data per question.[4] We further calculate correlations per country. Spearman's ρ works on relative

---
[4]We calculate the scores for the values based on the formula provided at `https://www.laits.utexas.edu/orkelm/kelmpub/VSM2013_Manual.pdf`, see Appendix F.

	pdi	idv	mas	uai	lto	ivr
Turkey	13.60	18.69	12.00	-104.66	18.40	-29.21
Philippines	69.97	32.45	-36.90	68.08	-29.34	127.78
Romania	44.30	28.05	1.36	-44.12	11.18	-98.11
Vietnam	19.07	36.61	11.82	53.48	5.50	-167.30
Malaysia	35.84	0.00	0.00	35.84	82.65	45.57
Korea South	86.41	-14.10	9.92	43.35	5.09	-38.42
Greece	104.29	-8.45	-27.99	58.92	7.64	-95.51
Iran	45.48	24.83	-34.00	-23.38	-60.23	-74.85
Germany	-57.78	23.73	35.01	96.53	60.96	-24.04
Indonesia	39.31	0.00	-24.93	40.82	24.23	-50.32
Pakistan	64.24	-0.91	44.61	154.20	19.85	-48.48
Serbia	-61.40	-56.70	-81.25	-75.70	-7.39	-38.73
Bangladesh	53.28	70.19	-31.67	36.50	25.46	-40.40

Figure 2: Heatmap of scores predicted per value for XLM-R mask probing on Hofstede's survey questions

predicted ranks to each variable, ignoring the individual predicted values. Our choice of using a rank correlation was motivated by the fact that we are working with population level aggregate responses and our aim of assessing whether language models pick up on relative differences in values across cultures, rather than on exact values.

7 Results

7.1 RQ1: Model Predictions

We show the predicted scores for the XLM-R model in Figure 2. As is clear from the figure, there are substantial differences in the predicted scores for the cultural dimensions across cultures. On average, scores for *power distance* (pdi) are high, whereas ones for *masculinity* (mas) and *indulgence* (ivr) are relatively low. The predicted logits suggest bias towards *Greece* and *South Korea* as places with high power distance, *Pakistan*, *Germany* as more masculine. *Indulgence* (ivr) has the lowest scores across all values with only *Phillipines* and *Malaysia* having positive values, indicating high restraint in these cultures according to the model predictions.

To understand whether LMs can preserve cross-cultural differences in values, we plot the results of the probing for Hofstede's and WVS' survey in Figures 3 and 4 respectively. As is visible in these plots, there is a variety in the values, i.e., the models seem to place different importance on different values across cultures, displaying cross-
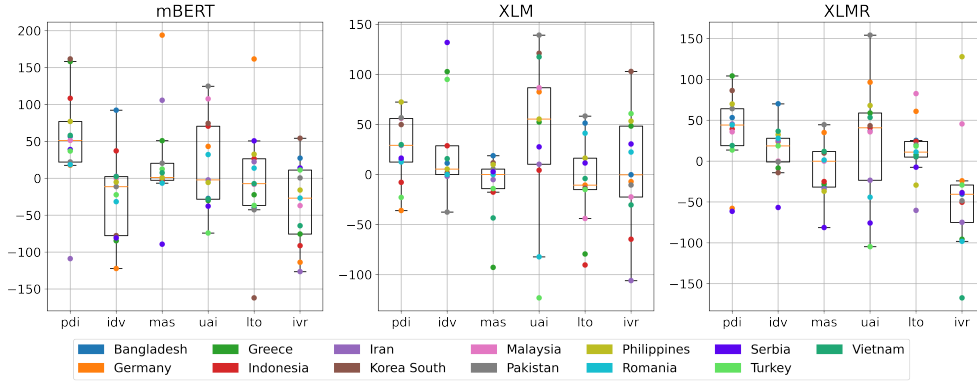
119

Figure 3: Scatter plots with quartiles of predicted value scores on Hofstede's survey questions for each of the three models.
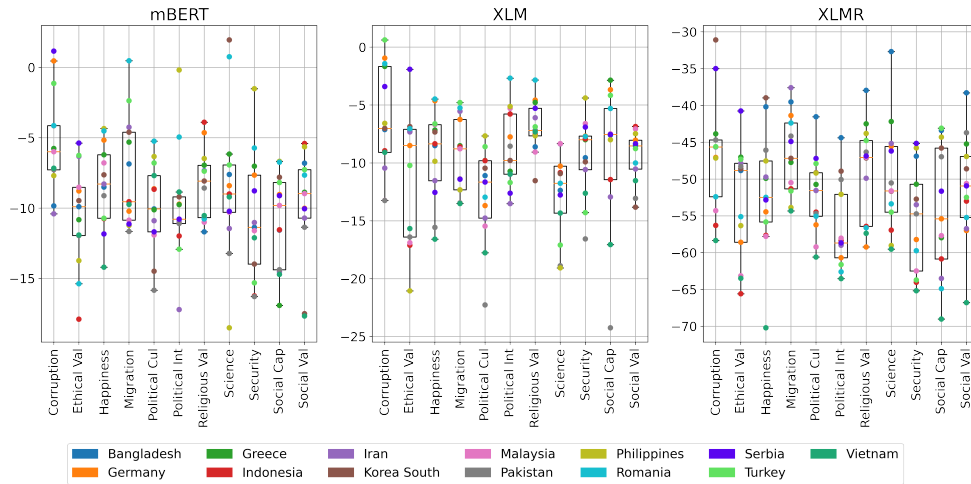

Figure 4: Scatter plots with quartiles of predicted value scores on WVS questions for each of the three models.

cultural differences in the values. We quantify these differences among the prediction scores by testing for statistical significance between the model's predictions by culture, seeing how they capture cross-cultural differences. For XLM-R's predictions for the WVS, 42.31% of the country pairs have a statistically significant difference, meaning the model preserves cross-cultural differences. For the other two models, the share of significantly different country pairs are 51.28% and 46.15% for mBERT and XLM respectively. For XLM-R's predictions of Hofstede's survey, only 10.26% of cultures have $p <= 0.05$. For the other two models, the share of significantly different country pairs are none and 6.41% for mBERT and XLM respectively. We attribute these low percentages to the fact that we conduct the test over the six value dimensions only, while it is on over 200 questions for WVS.

7.2 RQ2: Model Agreement

To further study whether scores across values and categories are consistent across the three models, we check for correlation between the predicted scores between the three models and outline them in Tables 2 and 3. We can see that predictions are inconsistent across the models, indicating differences in the embedded cross-cultural values. mBERT and XLM share the same architecture and are both trained on Wikipedia, yet the correlations across values are low, indicating the large effect that relatively minor changes to the model training can have on the cultural values picked up by the model.

7.3 RQ3: Alignment with Surveys

Finally, we investigate whether the models' predictions for the values questionnaire are consistent with existing values survey scores.

	XLM/ mBERT	XLM/ XLM-R	mBERT/ XLM-R
pdi	0.44	0.68*	0.48
idv	-0.14	-0.22	0.55
mas	-0.41	-0.14	0.43
uai	0.49	0.65*	0.42
lto	-0.05	-0.12	-0.15
ivr	0.67*	0.39	0.3

Table 2: Pairwise correlations in predictions for mask probing on Hofstede's values questions. Statistically significant values with $p \leq 0.05$ are marked with *

	XLM/ mBERT	XLM/ XLM-R	mBERT/ XLM-R
Corruption	0.57*	0.53	0.44
Ethical Va	0.61*	0.79*	0.47
Happiness	0.49	0.24	0.63*
Migration	0.16	0.44	0.25
Political Cu	0.38	0.65*	0.57*
Political In	0.6*	0.81*	0.75*
Religious	0.09	-0.31	0.05
Science	0.51	0.24	0.21
Security	0.49	0.77*	0.83*
Social Cap	0.21	0.4	0.42
Social Val	0.61*	0.27	0.68*

Table 3: Pairwise correlations in model predictions for mask probing on WVS questions. Statistically significant values with $p \leq 0.05$ are marked with *

Hofstede We outline the results of correlations between each of the models' predictions for mask probing per value in Table 4. We find no statistically significant alignment between the models' predictions and survey value scores provided by Hofstede, but given the low sample size, this is to be expected (Sullivan and Feinn, 2012). We find weak correlations among some of the values between the models' predicted scores and the values survey suggesting the disparity in cultural values outlined by Hofstede and the ones picked up by PLMs.

WVS Table 5 similarly shows the correlations between the models' predicted scores and the World Values Survey scores per category. Here too, we find no statistically significant correlation between the predicted and the survey scores outlining the difference in values picked up by the language models and those quantified in the surveys.

	mBERT	XLM	XLM-R
ivr	-0.44	0.07	0.38
idv	-0.38	-0.04	0.21
mas	0.37	0.09	-0.07
uai	-0.30	-0.30	-0.22
pdi	0.25	0.16	-0.11
lto	0.02	-0.01	0.23

Table 4: Correlation per value between mask prediction scores and Hofstede's values survey. Statistically significant values with p <= 0.05 are marked with *

We also check for per country correlations between the predicted scores and data from both values surveys, these are shown in Tables 11 and 12 in the Appendix.

	mBERT	XLM	XLM-R
Science	0.50	0.09	0.19
Security	0.38	-0.22	0.09
Social Val	-0.34	-0.30	-0.07
Political Cul	0.29	0.15	-0.05
Political Int	0.25	0.02	0.10
Migration	0.19	0.26	0.21
Social Cap	0.17	0.06	-0.38
Religious Val	0.14	0.13	-0.37
Corruption	0.07	0.12	0.12
Happiness	-0.07	0.37	0.07
Ethical Val	0.04	-0.02	0.03

Table 5: Correlation per question between masked prediction scores and WVS. Statistically significant values with p <= 0.05 are marked with *

8 Discussion

Our experiments show that there are sizable differences in the cultural values picked up by the different multilingual models which are widely used for a number of language tasks, even when they are trained on data from the same source. This is in line with previous results (Stanczak et al., 2021) and hints at the sensitivity of model design, training choices, and their downstream effect on model biases. While the values picked up by the models vary across cultures, the bias in the models is not in line with values outlined in existing large scale values surveys. This is an unexpected result since PLMs are known to pick up on biases present in language data that they are trained on (Rogers et al., 2020; Stanczak and Augenstein,

2021). Further, values are known to be expressed in language (Norton, 1997). Hence, language models should pick up on and reflect cultural differences in values expressed in different languages based on their training text. A lack of such reflection points to possible shortcomings in representation learning when it comes to multilingual language models. There could be a number of reasons for this. One possible reason is the lack of diversity in multilingual training data. Wikipedia articles in different languages are written by a small subset of editors that are not representative of the populations in those countries. Further, large scale corpora like CommonCrawl over-represent the voices of people with access to the Internet, which in turn over-represents the values of people from those regions (Bender et al., 2021). Such a bias being present in GPT-3 was explored by Johnson et al. (2022) who show that LMs trained on Web text end up reflecting the biases of majority populations. Other work also shows that pre-training text contains substantial amounts of toxic and undesirable content even after filtering (Luccioni and Viviano, 2021). This highlights the need for including more diverse and carefully curated sources of data which are culturally sensitive and representative, in order for the models to better reflect the cultural values of those populations. Joseph et al. (2021) suggest that people express themselves differently online on Twitter compared to survey responses. This is another potential reason for this mis-alignment.

PLMs are used for a variety of different NLP tasks in different countries and hence to accommodate the usage of people from diverse backgrounds and cultures, it is not just important to have linguistic and typological diversity in training data, but also cultural diversity (Hershcovich et al., 2022). Having such a form of cultural knowledge is desirable for a number of real-world tasks including QA systems, dialogue systems, information retrieval. Further, a lack of such faithful representation could lead to unintended consequences during the deployment of such models such as models imposing a set form of normative ethics over a diverse population that may not subscribe to it (Talat et al., 2021; Johnson et al., 2022). This could also lead to models not being culturally sensitive and embedding harmful stereotypes (Nadeem et al., 2021). Recently, work has been done on trying to align models with human values (Hendrycks et al., 2021; Solaiman and Dennison, 2021). While this may seem like a good

idea at a first glance, also in light of the arguments presented above, some cultural values are harmful to portions of society, e.g. high levels of masculinity, which is connected to misogynistic language and perpetuating gender biases. Thus, when working with cultural values, an auditing system (Raji et al., 2020) with these value systems in mind and one that takes into account the downstream use case should be employed.

9 Conclusion

In this study, we propose a methodology for probing of cultural values embedded in multilingual Pre-trained Language Models and assessing differences among them. We measure alignment of these values amongst the models and with existing values surveys. We find that PLMs capture marked differences in values between cultures, though these in turn are only weakly correlated with values surveys. Alongside training data, we discuss the impact training and modelling choices can have on cultural bias picked up by the models. We further discuss the importance of this alignment when developing models in a cross-cultural context and offer suggestions for more inclusive ways of diversifying training data to incorporate these values.

10 Ethical Considerations

The ethical considerations for our work mostly relate to the limitations; there are a variety of unintended implications of equating a language and a country, such as misrepresentation of communities, and disregarding minority and diaspora communities. However, we believe it is the closest approximation possible when comparing the surveys used in this work and LMs. Further, the surveys have been criticised; particularly Hofstede's cultural dimensions theory has been deemed too simplistic (Jackson, 2020). This could lead also to simplistic assumptions when probing an LM. We address these problems by including the WVS, another widely used survey, in our study. Due to these limitations, we believe that further studies and applications of our approach should be done with these limitations in mind. Particularly the simplification of cultural representation by both our approach as well as the original surveys might impact communities negatively. Such misrepresentation can have a disproportionate impact and exacerbate the marginalisation of minority communities or subcultures.

References

Jacob Levy Abitbol, Márton Karsai, Jean-Philippe Mague, Jean-Pierre Chevrot, and Eric Fleury. 2018. Socioeconomic dependencies of linguistic patterns in twitter: A multivariate analysis. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1125–1134, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Sky CH-Wang and David Jurgens. 2021. Using sociolinguistic variables to reveal changing attitudes towards sexuality and gender. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9918–9938, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2112–2128. Association for Computational Linguistics.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLOS ONE*, 9(11):1–13.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 122–133, New York, NY, USA. Association for Computing Machinery.

C. Haerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, M. Lagos Diez-Medrano J., P. Norris, E. Ponarin, and B. Puranen. 2022. World values survey: Round seven - country-pooled datafile version 3.0. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat*.

Noriko Hara, Pnina Shachaf, and Khe Foon Hew. 2010. Cross-cultural analysis of the wikipedia community. *J. Assoc. Inf. Sci. Technol.*, 61(10):2097–2108.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning {ai} with shared human values. In *International Conference on Learning Representations*.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *CoRR*, abs/1506.03340.

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.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Geert Hofstede. 1984. *Culture's consequences: International differences in work-related values*, volume 5. sage.

Geert Hofstede. 2005. Culture's recent consequences. In *Designing for Global Markets 7, IWIPS 2005, Bridging Cultural Differences, 7-9 July 2005, Amsterdam, The Netherlands, Proceedings of the Seventh International Workshop on Internationalisation of Products and Systems*, pages 3–4. Product & Systems Internationalisation, Inc.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 452–461, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Ronald Inglehart. 2006. Inglehart-welzel cultural map of the world. *World Values Survey*.

Terence Jackson. 2020. The legacy of geert hofstede.

Alexander Jaffe. 2009. Stance: Sociolinguistic perspectives.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual Factual Knowledge Retrieval from Pretrained Language Models. *arXiv:2010.06189 [cs]*. ArXiv: 2010.06189.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3.

Michael L Jones. 2007. Hofstede-culturally questionable?

Kenneth Joseph, Sarah Shugars, Ryan J. Gallagher, Jon Green, Alexi Quintana Mathé, Zijian An, and David Lazer. 2021. (mis)alignment between stance expressed in social media data and public opinion surveys. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 312–324. Association for Computational Linguistics.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Writer profiling without the writer's text. In *Social Informatics*, pages 537–558, Cham. Springer International Publishing.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask – evaluating social biases in masked language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):11954–11962.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

William Labov. 1963. The social motivation of a sound change. *WORD*, 19(3):273–309.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually Grounded Reasoning across Languages and Cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Marc Miquel-Ribé and David Laniado. 2019. Wikipedia cultural diversity dataset: A complete cartography for 300 language editions. In *Proceedings of the Thirteenth International Conference on Web and Social Media, ICWSM 2019, Munich, Germany, June 11-14, 2019*, pages 620–629. AAAI Press.

Marius Mosbach, Anna Khokhlova, Michael A. Hedderich, and Dietrich Klakow. 2020. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 68–82, Online. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Ercan G Nasif, Hamad Al-Daeaj, Bahman Ebrahimi, and Mary S Thibodeaux. 1991. Methodological problems in cross-cultural research: An updated review. *MIR: Management International Review*, pages 79–91.

Bonny Norton. 1997. Language, identity, and the ownership of english. *TESOL Quarterly*, 31(3):409–429.

Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274, Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 33–44, New York, NY, USA. Association for Computing Machinery.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Milton Rokeach. 2008. *Understanding human values*. Simon and Schuster.

Shamik Roy, Maria Leonor Pacheco, and Dan Goldwasser. 2021. Identifying morality frames in political tweets using relational learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9939–9958, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages

2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

Shalom H. Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, Ozlem Dirilen-Gumus, and Mark Konty. 2012. Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, 103:663–688. Place: US Publisher: American Psychological Association.

Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. In *Advances in Neural Information Processing Systems*, volume 34, pages 5861–5873. Curran Associates, Inc.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *CoRR*, abs/2112.14168.

Karolina Stanczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, and Isabelle Augenstein. 2021. Quantifying gender bias towards politicians in cross-lingual language models. *CoRR*, abs/2104.07505.

Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, and Isabelle Augenstein. 2021. Quantifying Gender Bias Towards Politicians in Cross-Lingual Language Models. *arXiv:2104.07505 [cs, stat]*. ArXiv: 2104.07505.

Gail M Sullivan and Richard Feinn. 2012. Using effect size—or why the p value is not enough. *Journal of graduate medical education*, 4(3):279–282.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2021. A Word on Machine Ethics: A Response to Jiang et al. (2021). *arXiv:2111.04158 [cs]*. ArXiv: 2111.04158.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433.

Peter Trudgill. 2002. *Sociolinguistic Variation and Change*. Edinburgh University Press.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Limitations

There are several limitations of our approach in trying to assess cultural diversity and alignment of the values picked up by PLMs. While our methodology of probing models using Cloze style questions gives us some insight into token level biases picked up by the language models, it is limited in its approach to only show static and extrinsic biases at inference time using output probabilities. There are intrinsic measures for quantifying bias, but those do not always correlate with extrinsic measures (Goldfarb-Tarrant et al., 2021). In order to make the experimental setting more robust and clearly demonstrate signs of embedded cultural bias, we performed experiments with an extended set of synonyms for each label word. However, this turned out to be non-trivial for a number of reasons. First, replacing synonyms in place of original words rarely results in grammatical sentences. Second, it is not always possible to find multiple synonyms of words in the same sense as the label words across the languages used in our study. Third, even when synonyms do exist, they are often multi-word expressions, which makes them incompatible with our experimental setting where a single word needs to be masked.

As discussed earlier, a major limitation that comes with quantifying cultural values is the mapping of countries to cultures and in our case, also to languages. Since this is an imperfect mapping, it is a difficult task to accurately quantify and assess cultural bias and values embedded in the models. We partially addressed this by restricting our study to languages which are mostly geographically restricted to one country. This is a limitation faced by cross-cultural research in general, where countries are often used as surrogates for cultures (Nasif et al., 1991). Finally, surveys and aggregate responses are also imperfect tools to evaluate and quantify cultural disparity, though the best ones currently in use. They are tasked with collapsing individual values into a set of questions. Individuals answering those questions from different backgrounds may perceive the questions differently. Further, there are several confounding factors affecting the survey responses and problems relating to seeing populations as a monolithic homogeneous whole. While these limitations pose important questions around how one should be careful in interpreting these values, we believe our study makes important contributions and provides a first step in assessing alignment between PLMs and cultural values, which we argue is necessary for models to faithfully work in a cross-cultural context.

B Translation quality

To assess the quality of translated probes, we conduct human evaluations of a sample of the output of the machine translator. We randomly select 3 probe questions from the Hofstede values survey and 23 probe questions from the World Values Survey representing 10% of the total probes. We then provide the original probe questions in English as well as their translations to annotators and assess the following two characteristics of the translations:

- *Grammaticality*: describes the correctness of the sentence standing alone, independent of the English sentence, in terms of obeying grammatical rules
- *Meaning*: describes how adequate the translation is for further reuse. We specifically want to know here, how correct the sentence is in relation to the English sentence. This could be also understood as the overall quality of the translation.

For each of the 26 probe questions, we ask the annotators to rate the sentence on the above listed characteristics across a 1-5 Likert scale. All annotators had at least a university level education, working proficiency of English, and were native speakers of the corresponding languages. We perform this annotation for 6 out of the 13 languages due to resource constraints. We provide the averaged scores for both the characteristics for each language in Table 6. The annotators on average across languages rate the meaning characteristic of the machine translated probes to be 4.73. This indicates the high degree to which the translations preserve the meaning of the sentences from the English probes. The grammaticality of the probes on average was rated to be 4.64. While lower than the value for the preserved meaning of the English sentence, the sentences were found to have very good grammar as well. The very high scores across the meaning characteristic of the translations suggest that for most of the probes, the translations were of high quality.

C Models and Compute

All models were run in Python using PyTorch (Paszke et al., 2019) and the Transform-

Language	Grammaticality	Meaning
German	4.85	4.88
Indonesian	4.62	4.77
Urdu	4.54	4.88
Serbian	4.88	5
Greek	4.58	4.80
Bengali	4.38	4.04
Average	4.64	4.73

Table 6: Averaged human evaluation scores on a 1-5 Likert scale for grammaticality and preserved meaning of the machine translated probes for a sample of languages used in this study

ers library (Wolf et al., 2020). When speaking about XLM-R, mBERT, XLM we refer to the models with the names xlm-roberta-base, bert-base-multilingual-uncased, xlm-mlm-100-1280 respectively. Since only inference was performed for probing the models, the experiments were run on a single NVIDIA Titan RTX GPU for less than 1 hour.

D Ablations

D.1 Label logit subtraction

To eliminate the possibility of lack of correlation due to subtraction of logit for label token with the lower response score in the survey question from the one with higher response score (Section 6.2), we calculate correlations with just the high response label token y_i^+. We report our results for Hofstede in Table 7 and WVS in Table 8. Similarly, we calculate value correlations for just the low response label and report them in Table 10 and Table 9 for Hofstede and WVS respectively.

	mBERT	XLM	XLM-R
mas	0.46	-0.11	-0.05
uai	0.46	0.13	0.06
ivr	-0.39	0.50	0.41
idv	-0.38	0.51	0.12
pdi	0.16	-0.00	-0.16
lto	-0.13	-0.05	-0.02

Table 7: Correlation per dimension between mask prediction scores for the high response score label y^+ and Hofstede's values survey. Statistically significant values with p <= 0.05 are marked with *

	mBERT	XLM	XLM-R
Science	0.51	0.40	-0.13
Social Val	-0.44	-0.50	0.16
Political Cul	0.43	0.33	0.08
Corruption	0.39	0.42	-0.11
Ethical	-0.24	0.10	0.29
Religious	-0.16	-0.06	0.36
Migration	0.14	0.17	0.08
Political Int	0.06	0.16	-0.21
Security	-0.06	-0.09	-0.12
Happiness	-0.06	-0.09	0.21
Social Cap	-0.06	-0.55*	0.22

Table 8: Correlation per category between mask prediction scores for the high response score label y^+ and the WVS. Statistically significant values with p <= 0.05 are marked with *

	mBERT	XLM	XLM-R
Ethical	0.63*	0.06	0.32
Security	-0.34	-0.05	-0.20
Religious	-0.27	-0.26	0.37
Social Val	-0.25	-0.61*	0.15
Political Int	-0.13	0.28	-0.18
Migration	0.08	0.09	-0.19
Political Cul	0.06	0.12	0.08
Happiness	-0.03	-0.47	0.21
Corruption	-0.03	0.34	-0.20
Social Cap	0.01	-0.47	0.26
Science	-0.00	-0.40	-0.28

Table 9: Correlation per category between mask prediction scores for the low response score label y^- and the WVS. Statistically significant values with p <= 0.05 are marked with *

E Example probes

In Table 13, we provide a sample of the question probes in English that are then translated to the different languages outlined in Section 5.

F Hofstede Value Calculation

We calculate the value results for the probes based on Hofstede (1984) by using the formulas used in the original survey.[5] The numbers following m represent the index of the survey questions, m stands for mean representing the mean survey question

---

[5]The formulas are provided along with the survey results and other information at https://www.laits.utexas.edu/orkelm/kelmpub/VSM2013_Manual.pdf.

	mBERT	XLM	XLM-R
mas	0.73*	0.55*	0.02
uai	0.55*	-0.39	-0.21
idv	0.18	0.45	-0.08
ivr	-0.16	-0.27	-0.11
lto	-0.08	-0.06	-0.38
pdi	-0.01	0.62*	0.39

Table 10: Correlation per dimension between mask prediction scores for the low response score label y^- and Hofstede's values survey. Statistically significant values with p <= 0.05 are marked with *

	mBERT	XLM	XLM-R
Romania	0.93*	-0.26	0.38
Pakistan	0.70	0.84*	0.99*
Greece	0.54	-0.09	0.49
Indonesia	0.54	-0.31	0.66
Vietnam	0.49	-0.14	-0.43
Serbia	0.37	-0.43	-0.31
Germany	0.26	0.23	0.60
Philippines	0.26	0.54	0.20
Bangladesh	-0.20	0.58	0.23
Iran	-0.14	0.83*	0.66
Turkey	-0.14	-0.83*	-0.71
Malaysia	-0.09	-0.06	0.41
Korea South	0.03	-0.03	0.54

Table 11: Correlation per country between mask prediction scores and Hofstede's values survey. Statistically significant values with p <= 0.05 are marked with *

response for the answer to that question, C is a constant that does not influence the comparison between countries. **Power Distance** defined as "the extent to which the less powerful members of organizations and institutions accept and expect that power is distributed unequally".

$$pdi = 35(m07 - m02) + 25(m20 - m23) + C(pd)$$

Individualism measures "the degree to which people in a society are integrated into groups".

$$idv = 35(m04 - m01) + 35(m09 - m06) + C(ic)$$

Uncertainty Avoidance measures "the extent to which a culture programs its members to feel either uncomfortable or comfortable in unstructured situations".

$$mas = 35(m05 - m03) + 35(m08 - m10) + C(mf)$$

	mBERT	XLM	XLM-R
Greece	0.78*	-0.26	0.01
Philippines	0.67*	0.53	0.36
Turkey	-0.56	-0.34	-0.86*
Malaysia	0.44	0.31	0.23
Bangladesh	0.43	-0.36	0.10
Vietnam	0.28	0.46	0.26
Iran	-0.24	-0.43	-0.09
Korea South	-0.20	-0.36	-0.06
Romania	0.20	0.14	-0.18
Indonesia	0.18	0.34	0.03
Germany	0.13	0.09	0.06
Serbia	0.03	-0.01	-0.14
Pakistan	-0.01	0.17	0.23

Table 12: Correlation per country between masked prediction scores and World Values Survey. Statistically significant values with p <= 0.05 are marked with *

Masculinity index indicates "the nature of clearly distinct social and emotional gender roles in a society."

$$uai = 40(m18 - m15) + 25(m21 - m24) + C(ua)$$

Long term orientation Cultures with short-term orientation value "reciprocating social obligations, respect for tradition, protecting one's 'face', and personal steadiness and stability more".

$$lto = 40(m13 - m14) + 25(m19 - m22) + C(ls)$$

Indulgence indicates "a society that allows relatively free gratification of basic and natural human desires related to enjoying life and having fun."

$$ivr = 35(m12 - m11) + 40(m17 - m16) + C(ir)$$

Value	Probe	y^+	y^-
Power Distance	I [MASK] that one can be a good manager without having a precise answer to every question that a subordinate may raise about his or her work	agree	disagree
Individualism	Having pleasant people to work with is [MASK]	important	unimportant
Masculinity	Having a job respected by your family and friends is [MASK]	important	unimportant
Uncertainty Avoidance	I feel [MASK] to be a citizen of my country	proud	ashamed
Long-term Orientation	In my experience, subordinates are [MASK] afraid to contradict their boss (or students their teacher)	never	always
Indulgence	All in all, I would describe the state of my health these days as [MASK]	good	bad
Corruption	There is [MASK] corruption in my country	abundant	no
Economic Vals	I [MASK] that competition is good	agree	disagree
Ethical Vals	Government monitoring all emails and any other information exchanged on the internet should be [MASK]	legal	illegal
Happiness	In the last 12 months, I or my family have [MASK] without cash income	often	never
Migration	I [MASK] that the government should let anyone from other countries who wants to	agree	disagree
Political Cul	Having the army rule is very [MASK]	good	bad
Political Int	Attending peaceful demonstrations is something I have [MASK] done	always	never
Science	I completely [MASK] that because of science and technology, there will be more opportunities for the next generation	agree	disagree
Security	Drug sale in the streets is [MASK] in my neighbourhood	frequent	infrequent
Social Capital	I am an [MASK] member of women's group	active	inactive
Social Vals	It is [MASK] for me to have people who speak a different language as neighbours	undesirable	desirable

Table 13: Examples of question probes in English reformulated from the original survey questions.