Speaking Multiple Languages Affects the Moral Bias of Language Models

Katharina Hämmerl[1,2] and **Björn Deiseroth**[3,4] and **Patrick Schramowski**[4,5,9]

Jindřich Libovický[6] and **Constantin A. Rothkopf**[5,7,8]

Alexander Fraser[1,2] and **Kristian Kersting**[4,5,8,9]

[1]Center for Information and Language Processing, LMU Munich, Germany
`{lastname}@cis.lmu.de`
[2]Munich Centre for Machine Learning (MCML), Germany
[3]Aleph Alpha GmbH, Heidelberg, Germany
[4]Artificial Intelligence and Machine Learning Lab, TU Darmstadt, Germany
[5]Hessian Center for Artificial Intelligence (hessian.AI), Darmstadt, Germany
[6]Faculty of Mathematics and Physics, Charles University, Czech Republic
[7]Institute of Psychology, TU Darmstadt, Germany
[8]Centre for Cognitive Science, TU Darmstadt, Germany
[9]German Center for Artificial Intelligence (DFKI)

Abstract

Pre-trained multilingual language models (PMLMs) are commonly used when dealing with data from multiple languages and cross-lingual transfer. However, PMLMs are trained on varying amounts of data for each language. In practice this means their performance is often much better on English than many other languages. We explore to what extent this also applies to moral norms. Do the models capture moral norms from English and impose them on other languages? Do the models exhibit random and thus potentially harmful beliefs in certain languages? Both these issues could negatively impact cross-lingual transfer and potentially lead to harmful outcomes. In this paper, we (1) apply the MORALDIRECTION framework to multilingual models, comparing results in German, Czech, Arabic, Chinese, and English, (2) analyse model behaviour on filtered parallel subtitles corpora, and (3) apply the models to a Moral Foundations Questionnaire, comparing with human responses from different countries. Our experiments demonstrate that PMLMs do encode differing moral biases, but these do not necessarily correspond to cultural differences or commonalities in human opinions. We release our code and models.[1]

1 Introduction

Recent work demonstrated large pre-trained language models capture some symbolic, relational
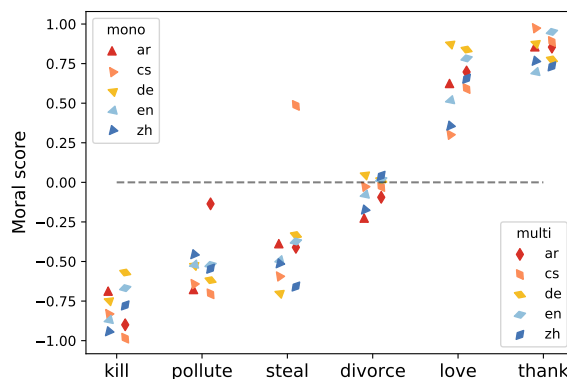

Figure 1: MORALDIRECTION score (y-axis) for several verbs (x-axis), as in Schramowski et al. (2022). We show scores for each language both from the respective monolingual model (triangles, left) and a multilingual model (rhombuses, right).

(Petroni et al., 2019), but also commonsense (Davison et al., 2019) knowledge. The undesirable side of this property is seen in models reproducing biases and stereotypes (e.g., Caliskan et al., 2017; Choenni et al., 2021). However, in neutral terms, language models trained on large data from particular contexts will reflect cultural "knowledge" from those contexts. We wonder whether multilingual models will also reflect cultural knowledge from multiple contexts, so we study moral intuitions and norms that the models might capture.

Recent studies investigated the extent to which language models reflect human values (Schramowski et al., 2022; Fraser et al., 2022). These works addressed monolingual English mod-

---

[1] https://github.com/kathyhaem/
multiling-moral-bias

els. Like them, we probe what the models encode, but we study multilingual models in comparison to monolingual models. Given constantly evolving social norms and differences between cultures and languages, we ask: Can a PMLM capture cultural differences, or does it impose a Western-centric view regardless of context? This is a broad question which we cannot answer definitively. However, we propose to analyse different aspects of mono- and multilingual model behaviour in order to come closer to an answer. In this paper, we pose three research questions, and present a series of experiments that address these questions qualitatively:

1. If we apply the MORALDIRECTION framework (Schramowski et al., 2022) to pretrained multilingual language models (PMLMs), how does this behave compared to monolingual models and to humans? (§ 3)

2. How does the framework behave when applied to parallel statements from a different data source? To this end, we analyse model behaviour on Czech-English and German-English OpenSubtitles data (§ 4).

3. Can the mono- and multi-lingual models make similar inferences to humans on a Moral Foundations Questionnaire (Graham et al., 2011)? Do they behave in ways that appropriately reflect cultural differences? (§ 5)

The three experiments reinforce each other in finding that our models grasp the moral dimension to some extent in all tested languages. There are differences between the models in different languages, which sometimes line up between multi- and mono-lingual models. This does not necessarily correspond with differences in human judgements. As an illustration, Figure 1 shows examples of the MORALDIRECTION score for several verbs in our monolingual and multilingual models.

We also find that the models are very reliant on lexical cues, leading to problems like misunderstanding negation, and disambiguation failures. This unfortunately makes it difficult to capture nuances. In this work we compare the behaviour of the PMLM both to human data and to the behaviour of monolingual models in our target languages Arabic, Czech, German, Chinese, and English.

2 Background

2.1 Pre-Trained Multilingual LMs

PMLMs, such as XLM-R (Conneau et al., 2020), are trained on large corpora of uncurated data, with an imbalanced proportion of language data included in the training. Although sentences with the same semantics in different languages should theoretically have the same or similar embeddings, this language neutrality is hard to achieve in practice (Libovický et al., 2020). Techniques for improving the model's internal semantic alignment (e.g., Zhao et al., 2021; Cao et al., 2020; Alqahtani et al., 2021; Chi et al., 2021; Hämmerl et al., 2022) have been developed, but these only partially mitigate the issue. Here, we are interested in a more complex type of semantics and how well they are cross-lingually aligned.

2.2 Cultural Differences in NLP

Several recent studies deal with the question of how cultural differences affect NLP. A recent comprehensive survey (Hershcovich et al., 2022) highlights challenges along the cultural axes of *aboutness*, *values*, *linguistic form*, and *common ground*. Some years earlier, Lin et al. (2018) mined cross-cultural differences from Twitter data, focusing on named entities and slang terms from English and Chinese. Yin et al. (2022) probed PMLMs for "geo-diverse commonsense", concluding that for this task, the models are not particularly biased towards knowledge about Western countries. However, in their work the knowledge in question is often quite simple. We are interested in whether this holds for more complex cultural values. In the present study, we assume that using a country's primary language is the simplest way to probe for values from the target cultural context. Our work analyses the extent to which one kind of cultural difference, moral norms, is captured in PMLMs.

2.3 Moral Norms in Pre-Trained LMs

Multiple recent studies have investigated the extent to which language models reflect human values (Schramowski et al., 2022; Fraser et al., 2022). Further, benchmark datasets (Hendrycks et al., 2021; Emelin et al., 2021; Ziems et al., 2022) aiming to align machine values with human labelled data have been introduced. Several such datasets (Forbes et al., 2020; Hendrycks et al., 2021; Alhassan et al., 2022) include scenarios from the "Am I the Asshole?" subreddit, an online community

where users ask for an outside perspective on personal disagreements. Some datasets use the community judgements as labels directly, others involve crowdworkers in the dataset creation process.

Other works have trained models specifically to interpret moral scenarios, using such datasets. A well-known example is Jiang et al. (2021), who propose a fine-tuned UNICORN model they call DELPHI. This work has drawn significant criticism, such as from Talat et al. (2021), who argue "that a model that generates moral judgments cannot avoid creating and reinforcing norms, i.e., being *normative*". They further point out that the training sets sometimes conflate moral questions with other issues such as medical advice or sentiments.

Hulpuș et al. (2020) explore a different direction. They project the Moral Foundations Dictionary, a set of lexical items related to foundations in Moral Foundations Theory (§ 2.4), onto a knowledge graph. By scoring all entities in the graph for their relevance to moral foundations, they hope to detect moral values expressed in a text. Solaiman and Dennison (2021) aim to adjust a pre-trained model to specific cultural values as defined in a targeted dataset. For instance, they assert "the model should oppose unhealthy beauty [...] standards".

A very interesting and largely unexplored area of research is to consider whether *multilingual* language models capture differing moral norms. For instance, moral norms in the Chinese space in a PMLM might systematically differ from those in the Czech space. Arora et al. (2022) attempt to probe pre-trained models for cultural value differences using Hofstede's cultural dimensions theory (Hofstede, 1984) and the World Values Survey (Haerpfer et al., 2022). They convert the survey questions to cloze-style question probes, obtaining score values by subtracting the output distribution logits for two possible completions from each other. However, they find mostly very low correlations of model answers with human references. Only a few of their results show statistically significant correlations. They conclude that the models differ between languages, but that these differences do not map well onto human cultural differences.

Due to the observation that the output distributions themselves do not reflect moral values well, we choose the MORALDIRECTION framework for our studies. In previous work, this approach identified a subspace of the model weights relating to a sense of "right" and "wrong" in English.

2.4 Moral Foundations Theory

Moral Foundations Theory (Haidt and Joseph, 2004) is a comparative theory describing what it calls *foundational moral principles*, whose relative importance can be measured to describe a given person's or culture's moral attitudes. Graham et al. (2009) name the five factors "Care/Harm", "Fairness/Reciprocity", "Authority/Respect", "Ingroup/Loyalty", and "Purity/Sanctity". Their importance varies both across international cultures (Graham et al., 2011) and the (US-American) political spectrum (Graham et al., 2009). The theory has been criticised by some for its claim of innateness and its choice of factors, which has been described as "contrived" (Suhler and Churchland, 2011). Nevertheless, the associated Moral Foundations Questionnaire (Graham et al., 2011) has been translated into many languages and the theory used in many different studies (such as Joeckel et al., 2012; Piazza et al., 2019; Doğruyol et al., 2019). An updated version of the MFQ is being developed by Atari et al. (2022). In § 5, we score these questions using our models and compare with human responses from previous studies on the MFQ.

2.5 Sentence Transformers

By default, BERT-like models output embeddings at a subword-token level. However, for many applications, including ours, sentence-level representations are necessary. In our case, inducing the moral direction does not work well for mean-pooled token representations, leading to near-random scores in many cases (see § 3.1). Reimers and Gurevych (2019) proposed Sentence-Transformers as a way to obtain meaningful, constant sized, sentence representations from BERT-like models. The first Sentence-BERT (S-BERT) models were trained by tuning a pre-trained model on a sentence pair classification task. By encoding each sentence separately and using a classification loss, the model learns more meaningful representations.

To obtain multilingual sentence representations, they proposed a student-teacher training approach using parallel corpora (Reimers and Gurevych, 2020), where a monolingual S-BERT model acts as a teacher and a pre-trained multilingual model as a student model. Such an approach forces the parallel sentences much closer together than in the original PMLM, which is not always desirable. In our case, we might be unable to distinguish the effects of the S-BERT training from the original model, which

would interfere with probing the original model.

Unlike their work, we train a multilingual sentence transformer by translating the initial training data into our target languages (§ 3.2), and show that this is effective. With this contribution, we show that multilingual S-BERT models can be trained in the same way as monolingual ones. Our approach does not require a teacher-student training setup. Note that we do require comparable datasets in size and ideally topics for each language. While we do not explicitly align the data, we solve this by using machine translated versions of existing datasets, which means we have implicitly parallel data.

3 Inducing Moral Dimensions in PMLMs

We choose five languages to evaluate the behaviour of the multilingual models: Modern Standard Arabic (ar), Czech (cs), German (de), English (en), and Mandarin Chinese (zh). These are all relatively high-resource languages, so we hope the model will be able to reliably detect cultural knowledge in each language. Since we rely on machine translation, using high-resource languages also ensures good translation quality. We note here that languages and cultures or countries are at best approximately equivalent (cf. Lin et al., 2018; Sahlgren et al., 2021). For instance, Arabic, English, and Chinese are standardised varieties that are written in a wide range of different contexts or cultures, which are likely merged together in the model to some degree. However, separation by language is the best way we have to distinguish cultural contexts within a multilingual model. As a point of comparison, we choose a monolingual language model for each of our target languages (see App. A for details).

3.1 MORALDIRECTION Framework

We use the MORALDIRECTION framework by Schramowski et al. (2022). In the first step, this method encodes a set of positively and negatively connotated action verbs with a sentence embedding model. In Schramowski et al. (2022), this is an S-BERT model. Each action verb is inserted into a set of ten template questions, such as "Should I [verb]?", "Is it examplary to [verb]?", and the output embedding for a verb is the mean over the embeddings of these questions. Next, PCA is applied to the outputs, to obtain the "moral direction" subspace of the model. Since the inputs are templates with only individual verbs changing, they are linguistically homogeneous, and the most salient dif-

ferences for the PCA are the value judgements. Ideally, a high amount of variance should be explained by the first principal component. The scores of these initial verbs are then normalised to lie within $[-1, 1]$. Subsequent scores can sometimes lie outside this range despite applying the normalisation. The scores are then read as a value estimation along one axis, with scores around 0 being "neutral", scores close to -1 being very "bad", and scores close to 1 very "good". However, note that the results we list in Tables 1-3 and 8 are correlations of model scores with user study data or correlations of model scores with other model scores.

We choose to use MORALDIRECTION because it is able to work directly with sentence embeddings and extract a reasonably human-correlated moral direction from them, producing a value score along a single axis. This makes it computationally inexpensive to transfer to other languages and datasets. A drawback is that it is induced on short, unambiguous phrases, and can be expected to work better on such phrases. Deriving a score along a single axis can also be limiting or inappropriate in certain contexts. See also the discussion in Limitations.

For a list of the verbs and questions used to derive the transformation, see the source paper. Schramowski et al. (2022) also conduct a user study on Amazon MTurk to obtain reference scores for the statements in question.

To test this method on multilingual and non-English monolingual models, we machine translate both the verbs and the filled question templates used in the above study. See Appendix B for the MT systems used, and a discussion of translation quality. We edited some of the questions to ensure good translation.[2] Our primary measure is the correlation of resulting model scores with responses from the study in Schramowski et al. (2022).

We initially tested the method on mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), as well as a selection of similarly sized monolingual models (Devlin et al., 2019; Antoun et al., 2020; Straka et al., 2021; Chan et al., 2020), by mean-pooling their token representations. See Appendix Table 5 for a list of the models used. Table 1 shows these initial results with mean-pooling. However, this generally did not achieve a correlation with the user study. There were exceptions to this rule—i.e., the Chinese monolingual BERT, and the English and Chinese portions of mBERT. This

---

[2]e.g. "smile to sb." → "smile at sb."

Model	en	ar	cs	de	zh
mBERT (mean-pooled)	0.65	-0.10	0.12	-0.18	0.62
XLM-R (mean-pooled)	-0.30	-0.07	-0.03	-0.14	0.10
monolingual (mean-pooled)	-0.13	0.46	0.07	0.10	0.70
monolingual S-BERT-large	0.79	—	—	—	—
XLM-R (S-BERT)	0.85	0.82	0.85	0.83	0.81

Table 1: Correlation of MORALDIRECTION scores with user study data for different pre-trained mono- and multi-lingual models. First three rows used mean-pooled sentence embeddings; last two rows used embeddings resulting from sentence-transformers (Reimers and Gurevych, 2019).

may be due to details in how the different models are trained, or how much training data is available for each language in the multilingual models. Table 1 also includes results from the monolingual, large English S-BERT, and an existing S-BERT version of XLM-R[3] (Reimers and Gurevych, 2020). These two models did show good correlation with the global user study, highlighting that this goal requires semantic sentence representations.

3.2 Sentence Representations

The existing S-BERT XLM-R model uses the student-teacher training with explicitly aligned data mentioned in § 2.5. As we discuss there, we aim to change semantic alignment in the PMLM as little as possible before probing it. We also need S-BERT versions of the monolingual models. Therefore, we train our own S-BERT models. We use the sentence-transformers library (Reimers and Gurevych, 2019), following their training procedure for training with NLI data.[4] Although we do not need explicitly aligned data, we do require comparable corpora in all five languages, so we decide to use MNLI in all five languages. In addition to the original English MultiNLI dataset (Williams et al., 2018), we take the German, Chinese and Arabic translations from XNLI (Conneau et al., 2018), and provide our own Czech machine translations (cf. Appendix B). Each monolingual model was tuned with the matching translation, while XLM-R$_{Base}$ was tuned with all five dataset translations. Thus, our multilingual S-BERT model was not trained directly to align parallel sentences, but rather trained with similar data in each involved language (without explicit alignment). For more training details,

---

Model	en	ar	cs	de	zh
XLM-R + MNLI (S-BERT, all 5 langs)	0.86	0.77	0.74	0.81	0.86
monolingual + MNLI (S-BERT, respective lang)	0.86	0.76	0.81	0.84	0.80

Table 2: Correlation of MORALDIRECTION scores from our mono- and multi-lingual S-BERT models with user study data.

see Appendix D. We release the resulting S-BERT models to the Huggingface hub.

3.3 Results

Table 2 shows the user study correlations of our S-BERT models. Clearly, sentence-level representations work much better for inducing the moral direction, and the method works similarly well across all target languages. Figures 1 and 5 show examples of verb scores across models and languages, further illustrating that this method is a reasonable starting point for our experiments.

For Arabic and the Czech portion of XLM-R, the correlations are slightly lower than the other models. Notably, Arabic and Czech are the smallest of our languages in XLM-R, at 5.4 GB and 4.4 GB of data (Wenzek et al., 2020), while their monolingual models contain 24 GB and 80 GB of data.

Since in the case of Czech, the correlation is higher in the monolingual model, and XLM-R and the monolingual model disagree somewhat (Table 3), the lower correlation seems to point to a flaw of its representation in XLM-R. For Arabic, the correlation of the monolingual model with English is similar to that seen in XLM-R, but the monolingual model also disagrees somewhat with the XLM-R representation (Table 3). This may mean there is actually some difference in attitude (based on the monolingual models), but XLM-R also does not capture it well (based on the XLM-R correlations). Unfortunately, Schramowski et al. (2022) collected no data specifically from Arabic or Czech speakers to illuminate this.

In Table 3 we compare how much the scores correlate with each other when querying XLM-R and the monolingual models in different languages. The diagonal shows correlations between the monolingual model of each language and XLM-R in that language. Above the diagonal, we show how much the monolingual models agree with each other, while below the diagonal is the agreement of different languages within XLM-R. On the diagonal, we compare each monolingual model with the match-

language	en	ar	cs	de	zh
en	0.93	0.86	0.92	0.89	0.91
ar	0.86	0.84	0.89	0.89	0.86
cs	0.90	0.78	0.86	0.92	0.92
de	0.95	0.87	0.88	0.95	0.91
zh	0.94	0.89	0.84	0.94	0.94

Table 3: Correlation of languages between our S-BERT models on the user study questions. Below diagonal: XLM-R model, tuned with MNLI data in five languages. Above diagonal: Monolingual models, tuned with MNLI data in the respective languages. On the diagonal: Correlation of the monolingual models with XLM-R in the respective language.

ing language in XLM-R. For English, German and Chinese, these show high correlations. The lowest correlation overall is between the Czech and Arabic portions of XLM-R, while the respective monolingual models actually agree more. The monolingual S-BERT models are generally at a similar level of correlation with each other as the multilingual model. German and Chinese, however, show a higher correlation with English in the multilingual model than in their respective monolingual models, which may show some interference from English.

We also show the correlations of languages within the pre-existing S-BERT model,[5] which was trained with parallel data, in Table 8. Here, the correlations between languages are much higher, showing that parallel data training indeed changes the model behaviour on the *moral dimension*. These correlations are higher than that of any one model with the user study data, so this likely corresponds to an artificial similarity with English, essentially removing cultural differences from this model.

Summarised, the experiments in this section extend Schramowski et al. (2022) to a multilingual setting and indicate that multilingual LMs indeed capture moral norms. The high mutual correlations of scores show that the differences between models and languages are relatively small in this respect. Note, however, that the tested statements provided by Schramowski et al. (2022) are not explicitly designed to grasp cultural differences. We thus add further experiments to address this question.

4 Qualitative Analysis on Parallel Data

To better understand how these models generalise for various types of texts, we conduct a qualita-

tive study using parallel data. For a parallel sentence pair, the MORALDIRECTION scores should be similar in most cases. Sentence pairs where the scores differ considerably may indicate cultural differences, or issues in the models. In practice, very large score differences appear to be more related to the latter. This type of understanding is important for further experiments with these models.

We conduct our analysis on OpenSubtitles parallel datasets (Lison and Tiedemann, 2016),[6] which consist of relatively short sentences. Given that the MORALDIRECTION is induced on short phrases, we believe that short sentences will be easier for the models. The subtitles often concern people's behaviour towards each other, and thus may carry some moral sentiment. We use English-German and English-Czech data for our analysis. To obtain the moral scores, we encode each sentence with the respective S-BERT model, apply the PCA transformation, and divide the first principal component by the normalising parameter.

Our analysis focuses on sentence pairs with very different scores. We take steps to filter out mistranslated sentence pairs—see Appendix H. Below, we discuss examples of where scores differ noticeably even when the translations are adequate. Using Czech-English and German-English data, we compare the monolingual models with XLM-R, XLM-R with the monolingual models, and the monolingual models with each other. This analysis is based on manual inspection of 500 sentence pairs with the highest score differences for each combination. Note that many of the sentence pairs were minor variations of each other, which significantly sped up the analysis. Relevant examples are listed with their MORALDIRECTION scores in Table 4 and Table 10 in the Appendix.

4.1 Reliance on Lexical Items

A common theme for many examples is an overreliance on individual lexical items. For example, "Traitors ... like you!" receives a positive score in English, while the German equivalent is correctly scored as negative. Most likely, the English models took a shortcut: "like you" is seen as a good thing.

Similarly, XLM-R in English scores "They're dying to meet you." somewhat negatively. The English BERT gives a positive score. However, arguably this is a case where the most correct answer would be neutral, since this is more a positive

---

[5]sentence-transformers/xlm-r-100langs-bert-base-nli-mean-tokens

[6]http://www.opensubtitles.org/

de	en	monoling		XLM-R	
		de	en	de	en
Pures Gift.	Pure poison.	-0.61	-0.71	*0.65*	-0.69
Ich erwürg dich!	I'll strangle you!	-0.41	-0.58	*0.90*	-0.62
Hab jemandem einen Gefallen getan.	I did someone a favour.	0.39	0.28	*-0.41*	0.73
Verräter ... wie Sie!	Traitors ... like you!	-0.56	*0.19*	-0.39	*0.72*
Sie brennen darauf, dich kennenzulernen.	They're dying to meet you.	0.44	0.73	0.52	*-0.31*
Ich vermisse ihn sehr.	I really miss him.	0.69	0.23	*-0.41*	*-0.26*
Er schätzt mich.	He values me.	1.12	0.31	*0.04*	0.88

Table 4: Examples from the German-English OpenSubtitles data for which there is a large, spurious contrast between MORALDIRECTION scores. Scores that stand out as unreasonable are *italicised*.

sentiment than any moral concern.

4.2 Multilinguality and Polysemy

Continuing the theme of literalness, another dimension is added to this in the multilingual setting. For instance, XLM-R scores the German "Pures Gift." (*pure poison*) as positive, likely because the key word "Gift" looks like English "gift", as in present. However, the model also makes less explainable mistakes: many sentences with "erwürgen" (*to strangle*) receive a highly positive score.

In the Czech-English data, there are even more obvious mistakes without a straightforward explanation. Some Czech words are clearly not understood by XLM-R: For instance, sentences with "štědrý" (*generous*) are negative, while any sentence with "páčidlo" (*crowbar*) in it is very positive in XLM-R. Phrases with "vrah" (*murderer*) get a positive score in XLM-R, possibly because of transliterations of the Russian word for medical doctor. Most of these obvious mistakes of XLM-R are not present in RobeCzech. However, "Otrávils nás" (*You poisoned us*) receives a positive score from RobeCzech for unknown reasons.

Confusing one word for another can also be a problem within a single language: For example, "Gefallen" (a favour) receives a negative score from XLM-R in many sentences. It is possible this model is confusing this with "gefallen" (past participle of "fallen", to fall), or some other similar word from a different language. "Er schätzt mich" and similar are highly positive in gBERT, as well as English XLM-R, but have a neutral score in German XLM-R. Likely the latter is failing to disambiguate here, and preferring "schätzen" as in *estimate*.

5 Moral Foundations Questionnaire

The MFQ has been applied in many different studies on culture and politics, meaning there is human response data from several countries available. We pose the MFQ questions from Graham et al. (2011) to our models, in order to compare the model scores with data from previous studies. We use the translations provided on the Moral Foundations website.[7]

Since the first part of the MFQ consists of very complex questions, we rephrase these into simple statements (see Appendix J). Many of the statements in the first half of the questionnaire become *reverse-coded* by simplifying them, that is, someone who values the aspect in question would be expected to answer in the negative. For these statements, we multiply the model score by -1. Further, we know that language models struggle with negation (Kassner and Schütze, 2020), so we remove "not" or "never" from two statements and flip the sign accordingly. In the same way, we remove "a lack of" from two statements.

These adjustments already improved the coherence of the resulting aspect scores, but we found further questions being scored by the models as if reverse-coded, i.e., with a negative score when some degree of agreement was expected. These were not simply negated statements, but they did tend to contain lexical items that were strongly negatively associated, and in multiple cases contained a negative moral judgement of the action or circumstance in question. Because the models appear to be so lexically focused (see § 4.1), this combination led to a strong negative score for some of these questions. We decided to rephrase such statements as well, usually flipping their sign while changing the wording as little as possible. Still, we

---

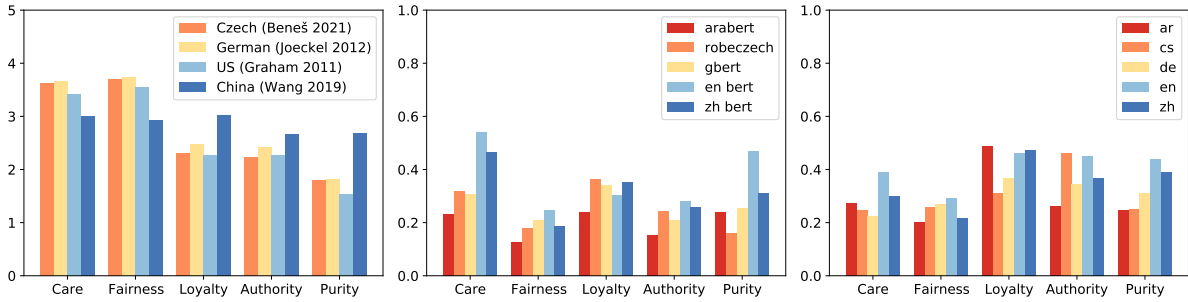[7] https://moralfoundations.org/questionnaires/

Figure 2: MFQ aspect scores from humans and models. Left: Examples of human data from studies in different countries. Middle: Scores obtained from monolingual MORALDIRECTION models. Right: Scores from XLM-R MORALDIRECTION in five languages.
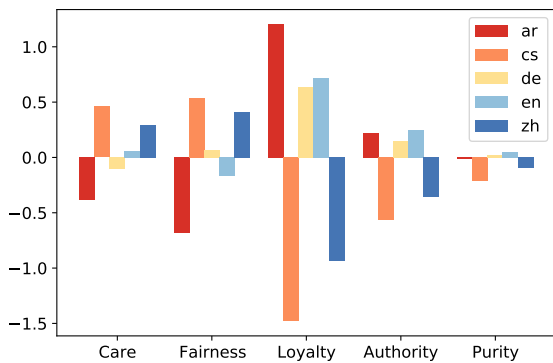

Figure 3: Sanity check—MFQ aspect scores from the XLM-R MORALDIRECTION models without Sentence-BERT tuning. This model had not obtained good correlations with human scores in § 3.

note here that this should be considered a type of prompt engineering, and that implicatures of the statements may have changed through this process. We provide the list of rephrased English statements and multipliers in Appendix Table 11.

We manually apply the same changes to the translations. The full list of English and translated statements, as well as model scores for each question, is available as a CSV file. Finally, we mean-pool the question scores within each aspect to obtain the aspect scores. Most of the model scores for each question will be within [-1, 1]. The results are shown in Figure 2.

5.1 Human Response Data

Also in Figure 2, we show German data from Joeckel et al. (2012), Czech data from Beneš (2021), US data from Graham et al. (2011), and Chinese data from Wang et al. (2019) for comparison. Note that these are not necessarily representative surveys. The majority of the data in question

were collected primarily in a university context and the samples skew highly educated and politically left. For Germany, the US and the Czech Republic, the individual variation, or variation between political ideologies seems to be larger than the variation between the countries. The Chinese sample scores more similarly to conservative respondents in the Western countries. Although many individuals score in similar patterns as the average, the difference between individuals in one country can be considerable. As an example, see Figure 7.

None of our models' scores map directly onto average human responses. The model scores do not use the full range of possible values, but even the patterns of relative importance do not match the average human patterns. Scores sometimes vary considerably in different models and different languages within XLM-R, and not necessarily in a way that would follow from cultural differences. The average scores within XLM-R are somewhat more similar to each other than the scores from the monolingual models are, giving some weak evidence that the languages in the multilingual model assimilate to one another. However, some differences between the monolingual models are also reflected in the multilingual model.

5.2 Sanity Check

We compare against scores from the unmodified, mean-pooled XLM-R models, shown in Figure 3. These models did not have the Sentence-BERT tuning applied to them, but otherwise we used the same procedure to obtain the scores. The inconsistent and very unlike human scores reinforce the finding from § 3 that mean-pooled representations are not useful for our experiments. They also confirm that the results in our main MFQ experiments are not arbitrary.

6 Conclusions

We investigated the *moral dimension* of pre-trained language models in a multilingual context. In this section, we discuss our research questions:

(1) Multilingual MORALDIRECTION. We applied the MORALDIRECTION framework to XLM-R, as well as monolingual language models in five languages. We were able to induce models that correlate with human data similarly well as their English counterparts in Schramowski et al. (2022). We analysed differences and similarities across languages.

In the process, we showed that sentence-level representations, rather than mean-pooled token-level representations, are necessary in order to induce a reasonable moral dimension for most of these models. We trained monolingual S-BERT models for our five target languages Arabic, Czech, German, English, and Mandarin Chinese. As well, we created a multilingual S-BERT model from XLM-R which was trained with MNLI data in all five target languages.

(2) Behaviour on Parallel Subtitles. A limitation of the MORALDIRECTION is that it is induced on individual words, and thus longer sentences are a significant challenge for the models. Still, we were able to test them on parallel subtitles data, which contains slightly longer, but predominantly still short, sentences. Problems that showed up repeatedly in this experiment were an over-reliance on key lexical items and a failure to understand compositional phrases, particularly negation. Additionally, typical problems of PMLMs, such as disambiguation problems across multiple languages, were noticeable within XLM-R. Non-English languages appeared more affected by such issues, despite the fact that all our target languages are relatively high resource.

(3) Moral Foundations Questionnaire. Our experiments with the MFQ reinforce the conclusion that the MORALDIRECTION models capture a general sense of right and wrong, but do not display entirely coherent behaviour. Again, compositional phrases and negation were an issue in multiple cases. We had set out to investigate whether cultural differences are adequately reflected in the models' cross-lingual behaviour. However, our findings indicate that rather, there are other issues with the cross-lingual transfer that mean we cannot make such nuanced statements about the model behaviour. To the extent that model behaviour differs for translated data, this does not seem to match cultural differences between average human responses from different countries.

We had initially wondered whether models would impose values from an English-speaking context on other languages. Based on this evidence, it seems that the models do differentiate between cultures to some extent, but there are caveats: The differences are not necessarily consistent with human value differences, which means the models are not always adequate. The problem appears to be worse when models are trained on smaller data for a given language. Meanwhile, German and Chinese have noticeably high agreement with English in our multilingual model, and all languages are extremely highly correlated in the pre-existing parallel-data S-BERT model (Table 8). This clearly shows that training with parallel data leads to more similar behaviour in this dimension, more or less removing cultural differences, but indeed there may be some transference even without parallel data.

Future Work. This leads to several future research questions: (i) Can we reliably investigate encoded (moral) knowledge reflected by PMLMs on latent representations or neuron activations? Or do we need novel approaches? For instance, Jiang et al. (2021) suggest evaluating the output of generative models and, subsequently, Arora et al. (2022) apply masked generation using PMLMs to probe cultural differences in values. However, the generation process of LMs is highly dependent, among other things, on the sampling process. Therefore, it is questionable if such approaches provide the required transparency. Nevertheless, Arora et al. (2022) come to a similar conclusion as indicated by our results: PMLMs encode differences between cultures. However, these are weakly correlated with human surveys, which leads us to the second future research question: (ii) How can we reliably teach large-scale LMs to reflect cultural differences but also commonalities? Investigating PMLMs' moral direction and probing the generation process leads to inconclusive results, i.e., these models encode differences, which, however, do not correlate with human opinions. But correlating with human opinions is a requirement for models to work faithfully in a cross-cultural context. Therefore, we advocate for further research on teaching cultural characteristics to LMs.

Limitations

The MORALDIRECTION framework works primarily for short, unambiguous phrases. While we show that it is somewhat robust to longer phrases, it does not deal well with negation or certain types of compositional phrases. We showed that in such cases, prompt engineering seems to be necessary in order to get coherent answers. Inducing the MORALDIRECTION was done on a small set of verbs, and the test scenarios in this paper—apart from § 4—are also relatively small.

The scope of our work is specific to our stated target languages, which are all relatively high-resource, meaning the method may not hold up for languages with smaller corpora, especially in the context of PMLMs. This work presents primarily an exploratory analysis and qualitative insights.

Another point is that the monolingual models we used may not be precisely comparable. Table 5 lists details of parameter size, training, tokenizers, data size and data domain. The models are all similarly sized, but data size varies considerably. XLM-R and RobeCzech do not use next sentence prediction as part of their training objective. However, the authors of RoBERTa (Liu et al., 2019) argue this difference does not affect representation quality. Further, exactly comparable models do not exist for every language we use. We rather choose well-performing, commonly-used models. Thus, we believe the model differences play a negligible role in the context of our scope.

More broadly speaking, the present work makes the strong assumption that cultural context and language are more or less equivalent, which does not hold up in practice. Furthermore, MORALDIRECTION, like related methods, only consider a single axis, representing a simplistic model of morality. In the same vein, these models will output a score for any input sentence, including morally neutral ones, sometimes leading to random answers.

Broader Impacts

Language models should not decide moral questions in the real world, but research in that direction might suggest that this is in fact possible. Besides undue anthropomorphising of language models, using them to score moral questions could lead to multiple types of issues: The models may reproduce and reinforce questionable moral beliefs. The models may hallucinate beliefs. And particularly in the context of cross-lingual and cross-cultural work, humans might base false, overgeneralising, or stereotyping assumptions about other cultures on the output of the models.

References

Areej Alhassan, Jinkai Zhang, and Viktor Schlegel. 2022. 'Am I the bad one'? Predicting the moral judgement of the crowd using pre–trained language models. In *Proceedings of the Language Resources and Evaluation Conference*, page 267–276, Marseille, France. European Language Resources Association.

Sawsan Alqahtani, Garima Lalwani, Yi Zhang, Salvatore Romeo, and Saab Mansour. 2021. Using optimal transport as alignment objective for fine-tuning multilingual contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3904–3919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language*

Detection, pages 9–15, Marseille, France. European Language Resource Association.

Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2022. Probing pre-trained language models for cross-cultural differences in values. *CoRR*, abs/2203.13722.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2022. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *PsyArXiv*.

Michal Beneš. 2021. Psychometrické hodnocení dotazníku moral foundations questionnaire [online]. Master thesis, Masarykova univerzita, Filozofická fakulta, Brno, Czech Republic. Supervisor: Helena Klimusová.

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

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. *CoRR*, abs/2002.03518.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Pinzhen Chen, Jindřich Helcl, Ulrich Germann, Laurie Burchell, Nikolay Bogoychev, Antonio Valerio Miceli Barone, Jonas Waldendorf, Alexandra Birch, and Kenneth Heafield. 2021. The University of Edinburgh's English-German and English-Hausa submissions to the WMT21 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 104–109, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. Improving pretrained cross-lingual language models via self-labeled word alignment. 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 3418–3430, Online. Association for Computational Linguistics.

Rochelle Choenni, Ekaterina Shutova, and Robert van Rooij. 2021. Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1477–1491, 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, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. 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 1173–1178, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Burak Doğruyol, Sinan Alper, and Onurcan Yilmaz. 2019. The five-factor model of the moral foundations theory is stable across weird and non-weird cultures. *Personality and Individual Differences*, 151:109547.

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.

Kathleen C. Fraser, Svetlana Kiritchenko, and Esma Balkir. 2022. Does moral code have a moral code? probing delphi's moral philosophy. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 26–42, Seattle, U.S.A. Association for Computational Linguistics.

Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–46.

Jesse Graham, Brian Nosek, Jonathan Haidt, Ravi Iyer, Sena P Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology*, 101 (2):366–385.

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

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2022. Combining static and contextualised multilingual embeddings. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2316–2329, Dublin, Ireland. Association for Computational Linguistics.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yunhsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *CoRR*, abs/1705.00652.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Geert Hofstede. 1984. *Culture's Consequences: International Differences in Work-Related Values*. Cross Cultural Research and Methodology. SAGE Publications.

Ioana Hulpuș, Jonathan Kobbe, Heiner Stuckenschmidt, and Graeme Hirst. 2020. Knowledge graphs meet moral values. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 71–80, Barcelona, Spain (Online). Association for Computational Linguistics.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021. Delphi: Towards machine ethics and norms. *CoRR*, abs/2110.07574.

Sven Joeckel, Nicholas David Bowman, and Leyla Dogruel. 2012. Gut or game? the influence of moral intuitions on decisions in video games. *Media Psychology*, 15(4):460–485.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Frank F. Xu, Kenny Zhu, and Seungwon Hwang. 2018. Mining cross-cultural differences and similarities in social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 709–719, Melbourne, Australia. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

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. *CoRR*.

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.

Jared Piazza, Paulo Sousa, Joshua Rottman, and Stylianos Syropoulos. 2019. Which appraisals are foundational to moral judgment? Harm, injustice, and beyond. *Social Psychological and Personality Science*, 10(7):903–913.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Magnus Sahlgren, Fredrik Carlsson, Fredrik Olsson, and Love Börjeson. 2021. It's basically the same language anyway: the case for a Nordic language model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 367–372, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*.

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.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. RobeCzech: Czech RoBERTa, a monolingual contextualized language representation model. In *Text, Speech, and Dialogue*, pages 197–209, Cham. Springer International Publishing.

Christopher Suhler and Pat Churchland. 2011. Can innate, modular "foundations" explain morality? Challenges for Haidt's Moral Foundations Theory. *Journal of Cognitive Neuroscience*, 23:2103–16; discussion 2117.

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). *CoRR*, abs/2111.04158.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Ruile Wang, Qi Yang, Peng Huang, Liyang Sai, and Yue Gong. 2019. The association between disgust sensitivity and negative attitudes toward homosexuality: The mediating role of moral foundations. *Frontiers in Psychology*, 10.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoM-LAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 229–240, Online. Association for Computational Linguistics.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland. Association for Computational Linguistics.

A Details of Models Used

Table 5 lists the models we tuned and evaluated with their exact names, sizes, objectives and data.

The models are all of a similar size, although data size varies by up to one order of magnitude between monolingual models. XLM-R has much larger data in total, but data size for individual languages is more comparable to the other models. The data domains vary but overlap (Web, Wiki, News). XLM-R and RobeCzech do not use next sentence prediction as part of their training objective. However, we believe these differences play a negligible role in the context of our work.

B Machine Translation Quality

Machine translation is used to translate the templated sentences from English into Arabic, Czech, German and Chinese. For Arabic and Chinese, we use Google Translate. The sentences are short and grammatically very simple.

For translation into Czech, we use CUBBITT (Popel et al., 2020), a machine translation system that scored in the first cluster in WMT evaluation campaigns 2019–2021. For translation into German, we use the WMT21 submission of the University of Edinburgh (Chen et al., 2021). To validate our choice of machine translation systems, we estimate the translation quality using the reference-free version of the COMET score (Rei et al., 2020) (model `wmt21-comet-qe-mqm`) on the 2.7k generated questions.

To train the S-BERT models, we use the TRANSLATE-TRAIN part of the XNLI dataset that is distributed with the dataset (without specifying what translation system was used). For translation into Czech, we use CUBBITT again. To ensure the translation quality is comparable, we use the same evaluation metric as in the previous case on 5k randomly sampled sentences.

C Correlations in Existing (Parallel Data) S-BERT

Table 8 shows the correlations of languages within the pre-existing multilingual S-BERT, trained with

parallel data. The correlations within this model are extremely high, considerably higher than that of any one model with the user study.

D Sentence-BERT Tuning Procedure

We follow the training script provided by Reimers and Gurevych (2019) in the sentence-tranformers repository. As training data, we use the complete MNLI (Williams et al., 2018; 433k examples) in the five respective languages. The dev split from the STS benchmark (Cer et al., 2017; 1500 examples) serves as development data. We also machine translate this into the target languages. The loss function is Multiple Negatives Ranking Loss (Henderson et al., 2017), which benefits from larger batch sizes. We use sentence-transformers version 2.2.0. Table 9 lists further training parameters.

E Computational Resources

In addition to the six models used for further experiments, we trained five XLM-R with single-language portions of data. Each of the monolingual models, as well as the XLM-R versions tuned with one part of the data, took around 0.6 hours to train. Tuning XLM-R with data in all five languages accordingly took around three hours. S-BERT tuning was done on one Tesla V100-SXM3 GPU, with 32 GB RAM, at a time. We also trained one version of XLM-R on English data with a smaller batch size on an NVIDIA GeForce GTX 1080 GPU with 12 GB RAM. In all other experiments, the language models were used in inference mode only, and they were mostly run on the CPU.

F Variance in MORALDIRECTION Scores

In this section we discuss another aspect of MORALDIRECTION scores in multilingual versus monolingual models: How much they vary between different languages for each statement. For instance, if the variance is smaller in the multilingual model, this would mean that the multilingual model applies more similar judgements across languages. To quantify this, we calculate the score variance for each of the basic verbs from Schramowski et al. (2022) over the five monolingual models, as well as over the five portions of the multilingual model.

We furthermore grouped the verbs into "positive" and "negative", depending on whether their mean score from the multilingual model is greater or lower than zero. This results in 35 positive and 29 negative verbs. Figure 4 shows box-plots of

Lng	Name	Params	Objective	Tokenizer	Data size	Domain
ar	aubmindlab/bert-base-arabertv02 (Antoun et al., 2020)	110M	MLM+NSP	SP, 60k	24 GB	Wiki, News
cs	ufal/robeczech-base (Straka et al., 2021)	125M	MLM	BPE, 52k	80 GB	News, Wiki, Web
de	deepset/gbert-base (Chan et al., 2020)	110M	MLM+NSP	WP, 31k	136 GB	Web, Wiki, Legal
en	bert-base-cased (Devlin et al., 2019)	110M	MLM+NSP	WP, 30k	16 GB	Books, Wiki
zh	bert-base-chinese (Devlin et al., 2019)	110M	MLM+NSP	WP, 21k	?	Wiki
—	xlm-roberta-base (Conneau et al., 2020)	125M	MLM	SP, 250k	2.5TB	Web

Table 5: The monolingual pre-trained language models used. We tuned each model with the S-BERT framework before using it for our experiments. Objectives: MLM = masked language modelling, NSP = next sentence prediction, Tokenization: WP = WordPiece, SP = SentecePiece, unigram model.

Lng	Model	COMET
ar	Google Translate	.1163
	OPUS MT 2022	.1183
	OPUS MT 2020	.1163
cs	CUBBITT	.1212
	OPUS MT 2022	.1212
	OPUS MT 2020	.1197
	Google Translate	.1193
de	UEdin WMT21	.1191
	Facebook, WMT19	.1191
	Google Translate	.1190
	OPUS MT 2022	.1180
	OPUS MT 2020	.1123
zh	Google Translate	.1111
	OPUS MT 2020	.1101

Table 6: Machine translation quality of the templated sentences use in the MoralDimension estimation measured by the reference-free COMET score. Unused alternatives are in gray.

language	en	ar	cs	de	zh
en					
ar	0.97				
cs	0.98	0.97			
de	0.98	0.97	0.98		
zh	0.97	0.96	0.96	0.96	

Table 8: In-model correlation of scores on the user study questions, within `sentence-transformers/xlm-r-100langs-bert-base-nli-mean-tokens`.

Parameter	Value
Batch size	128
Max seq length	75
Epochs	1
Warmup	10% of train data
Save steps	500
Optimizer	AdamW
Weight decay	0.01

Table 9: Sentence-BERT tuning parameters.

Lng	Model	COMET
ar		.1013
de	as in XNLI	.1051
zh		.1017
cs	CUBBITT	.1153
	Google Translate	.1150
	OPUS MT 2022	.1144
	OPUS MT 2020	.1126

Table 7: Machine translation quality of the MNLI data used for training S-BERT models measured by the reference-free COMET score. Unused alternatives are in gray.

the variance for those groups. Overall, variances are similar for monolingual and multilingual models. The positive verbs have a lower variance in the multilingual than in the monolingual models. However, the opposite is true for the group of negative verbs, averaging out to very similar variances overall. Therefore, analysing variances does not lead us to conclusions about differing behaviour of monolingual versus multilingual models.

G More Examples MORALDIMENSION for Verbs

Additional examples to Figure 1 are shown in Figure 5.

H OpenSubtitles Filtering Details

Figure 6 shows the statistical correlation of the MORALDIRECTION scores on the OpenSubtitles dataset, evaluated for the German-English text
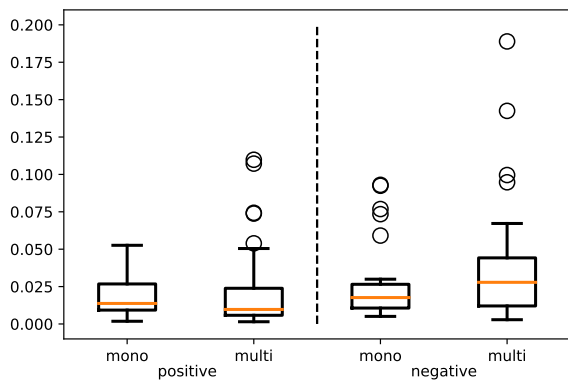
Figure 4: MORALDIRECTION score variances throughout all languages of the verbs taken from Schramowski et al. (2022), on monolingual and multilingual models. The verbs are further grouped into positive and negative based on mean scores.

pairs. The high Pearson correlation values give further evidence for a strong correlation of the compared scores and the plausibility of this experiment. As observed before with Section 3, evaluating on the multilingual XLM-R model strengthens the correlation of the MORALDIRECTION.

Initially, the most "controversial" sentence pairs—i.e., ones with extremely different MORALDIRECTION scores—in the OpenSubtitles data seemed to be due to mistranslated or misaligned subtitles. In order to exclude these cases, we applied filtering using dual cross-entropy score (Junczys-Dowmunt, 2018) based on OpusMT models (Tiedemann and Thottingal, 2020). For German-English, the correlation of the moral score delta and the translation score pre-filtering was -0.28 for XLM-R, -0.19 for the monolingual models, underlining that translation quality was a frequent issue. After filtering this dropped to -0.10 and -0.05. The score deltas are normally distributed.

I   OpenSubtitles: Further Czech-English Examples

Table 10 shows further examples from the Czech-English OpenSubtitles corpus. Several of them contain verbs with prefix *o-* which seems to confuse the multilingual model.

J   Rephrased MFQ Questions

The first 15 questions of MFQ have the format, "When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking? Please rate each statement using this scale: – not at all relevant (This consideration has nothing to do with my judgments of right and wrong) – not very relevant – slightly relevant – somewhat relevant – very relevant – extremely relevant (This is one of the most important factors when I judge right and wrong): Whether or not someone suffered emotionally ". We rephrase this to "Someone suffered emotionally."

Table 11 lists the English MFQ questions that we rephrased along with the sign of the multiplier applied (either -1 or +1).

K   Role of the "Catch" Questions in MFQ

The MFQ contains two catch questions, which are designed to have an obvious, uncontroversial answer. For human respondents, their purpose is to filter out people who are not paying attention. For the language models, they may indeed be informative as well. In English, these questions are: "Someone was good at math." and "It is better to do good than to do bad." For the first, we would expect the answer to be 0—this should be a perfectly neutral statement in a moral sense. For the other, we expect an answer at least close to the maximum score, since "doing good" is trivially better than "doing bad".

The English, Chinese, and Czech models do give scores close to 0 for the maths question. In Arabic, our monolingual model assigns a slight negative score, while XLM-R gives a moderately positive score. In German, both models give a moderately positive score, likely because the chosen translation "Jemand zeigt in Mathematik gute Leistungen" contains the somewhat positively connotated "Leistungen" (*performance, accomplishments*, etc.). The second catch question gets anything from fairly negative (-0.55), to neutral, to slightly positive scores, which again seems to fit with an overreliance on lexical cues. This behaviour shows again that while the models do capture the "moral dimension" to some degree, they have significant weaknesses, particularly with respect to the compositional meanings of longer phrases. In a real survey, they may not even have been considered "serious" respondents.
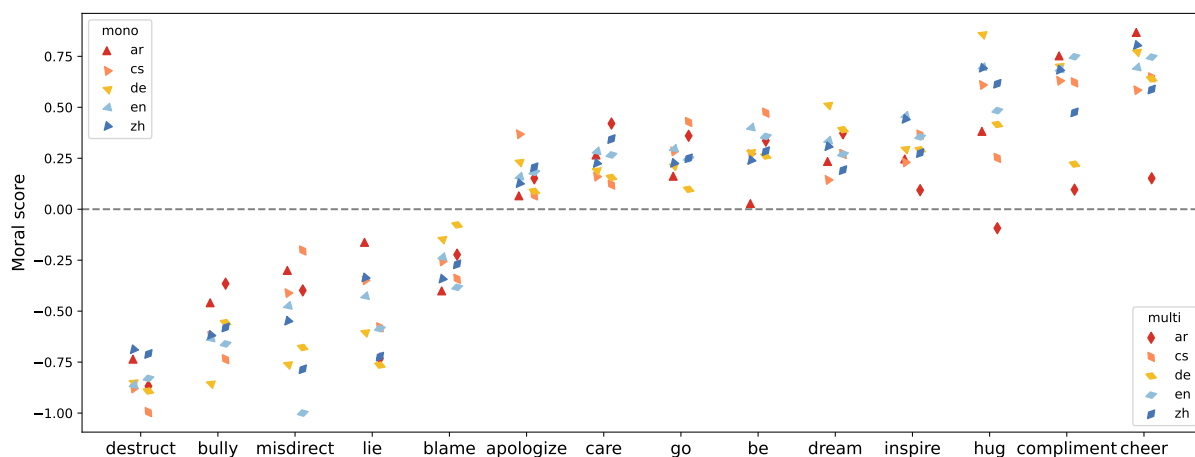
Figure 5: MORALDIRECTION score (y-axis) for more verbs (x-axis) than in Figure 1.

cs	en	monoling		XLM-R	
		cs	en	cs	en
To je opravdu štědré.	It's very generous.	0.50	0.88	*-0.46*	1.00
Tvá neteř mě velmi ohromila.	I was very impressed with your niece.	0.34	0.42	*-0.58*	0.84
Ohrožuje vaši budoucnost.	He threatens your future.	-0.88	-0.70	*0.42*	-0.56
Políbila jsi ho.	You kissed him.	0.52	0.21	*1.01*	*-0.36*
Jste vrah!	You're a murderer!	-0.97	-0.99	*0.38*	-0.77
Otrávils nás.	You poisoned us.	*0.37*	-0.82	-0.62	-0.63

Table 10: Examples from the (Czech-English) OpenSubtitles data for which there is a large, spurious contrast between MORALDIRECTION scores. Scores that stand out as unreasonable are *italicised*.
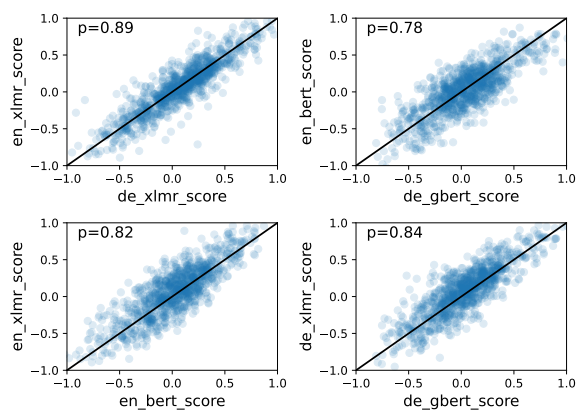

Figure 6: Correlation of the MORALDIRECTION scores for all German-English model combinations on the OpenSubtitles dataset.
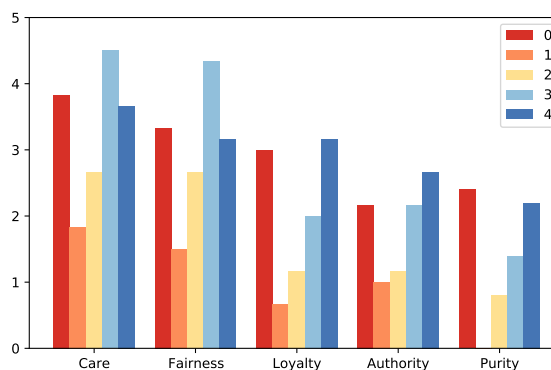

Figure 7: Example of human variation. Five different respondents from the German data collected by Joeckel et al. (2012).

Original (after simplifying)	sign	Rephrased	sign
Someone showed a lack of respect for authority.	-	Someone showed respect for authority.	+
Someone showed a lack of loyalty.	-	Someone showed loyalty.	+
People should not do things that are disgusting, even if no one is harmed.	+	People should do things that are disgusting, if no one is harmed.	-
One of the worst things a person could do is hurt a defenseless animal.	+	One of the best things a person could do is hurt a defenseless animal.	-
I would call some acts wrong on the grounds that they are unnatural.	+	I would call some acts right on the grounds that they are natural.	+
It can never be right to kill a human being.	+	It can be right to kill a human being.	-
I think it's morally wrong that rich children inherit a lot of money while poor children inherit nothing.	+	I think it's morally right that rich children inherit a lot of money while poor children inherit nothing.	-
If I were a soldier and disagreed with my commanding officer's orders, I would obey anyway because that is my duty.	+	If I were a soldier and disagreed with my commanding officer's orders, I would disobey.	-

Table 11: Rephrased MFQ statements in English. Unchanged statements are omitted from this table.

A   For every submission:

☐ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Left blank.

☐ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☐ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B   ☐ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C   ☐ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☐ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.