Speaking Multiple Languages Affects the Moral Bias of Language Models

Anonymous ACL submission

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, Mandarin 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 exhaustive experiments demonstrate that indeed PMLMs entail differing moral biases but they do not necessarily correspond with cultural differences and commonalities in human opinions.

1 Introduction

Recent work demonstrated large pre-trained language models capture some symbolic, relational (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), but in neutral terms, language models trained on large amounts of 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 the models might capture.

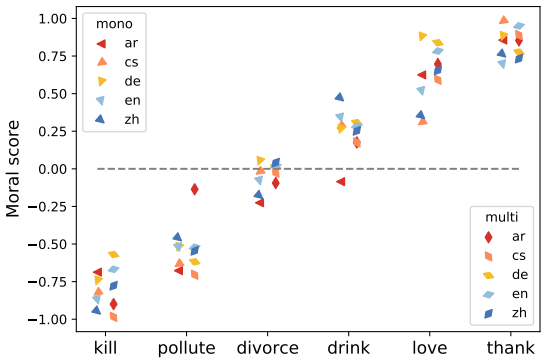

Figure 1: Moral score (y-axis) for several verbs (x-axis) evaluated for each language on the monolingual models of Table 5 separately (left) and on the multilingual model all together (right), as done in (Schramowski et al., 2022). We generally observe lower variance on the multilingual model, with few exceptions.

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 models. Like them, we we probe what the models encode. Given the 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 in all contexts? As a prospect, Figure 1 shows exemplary probing of the moral score for several verbs on separate monolingual models (top), and on a single multilingual model (bottom). We observe that the scores do change and the score variance shrinks is much lower for multilingual models.

To analyse this discrepancy of mono- and multilingual models in more detail, we pose three research questions, and present a series of experiments that address these questions qualitatively:

1. Can the MORALDIRECTION framework (Schramowski et al., 2022) be applied to pretrained multilingual language models (PMLMs)? (§ 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 mono- and multi-lingual 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. However, we will 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 nuanced cultural differences.

2 Background

Pre-Trained Multilingual LMs. PMLMs, such as XLM-R (Conneau et al., 2020), are trained on large corpora of uncurated data, with a 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.

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 the models are not particularly biased towards Western countries for this task. However, in their work the knowledge in question is often quite simple, such as the fact that a Chinese "staple food" is rice—something that Chinese speakers do not need to tell each other often. We are interested in whether this holds for more complex cultural values. In our present study, we do assume that probing in a country's primary language is the simplest way to access values from the target cultural context. Our work provides an analysis of one kind of cultural difference, moral norms, to the extent that they are captured in PMLMs.

Moral Norms in Pretrained 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; Forbes et al., 2020; Emelin et al., 2021) aiming to align machine values with human labelled data have been proposed. 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, ranging from the petty to the absurd. In some cases, the community judgements are used as labels directly, in others, crowdworkers are involved in the dataset creation process.

Others 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. The work has drawn significant criticism, among others 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 sentiment.

Hulpuș et al. (2020) explore a different direction in that they project the Moral Foundations Dictionary, lexical items related to foundations in Moral Foundations Theory (§ 2), onto knowledge graphs. 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) aims 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 could 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, with only few results showing 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.

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. By 2009, the theory names the five factors "Care/Harm", "Fairness/Reciprocity", "Authority/Respect", "Ingroup/Loyalty", and "Purity/Sanctity" (Graham et al., 2009). 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 foundations, 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). In § 5, we "ask" our models these questions and compare the model scores with human responses from several previous studies on the MFQ.

Sentence Transformers. By default, BERT-like models output embeddings at a subword-token level. However, for many applications, including ours, sentence-level representations are more useful or indeed necessary. In our case, inducing the moral direction does not work well for mean-pooled token-level representations. Reimers and Gurevych (2019) proposed Sentence-Transformer 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 to output more meaningful representations.

To obtain multilingual sentence representations, they later proposed a student-teacher training approach (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 what was in the original model from what the S-BERT training introduced, but our goal is to probe the representations of the existing model.

However, in this work we train a multilingual sentence transformer by translating the initial training data into our target languages (§ 3.2). 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

In this paper, 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. 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 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 (Table 5 in the Appendix).

3.1 MORALDIRECTION Framework

We start from the MORALDIRECTION experiments by Schramowski et al. (2022), which take an English S-BERT model, encode a set of positively and negatively connotated action verbs, and apply PCA to the outputs to obtain the "moral direction" subspace of the model. Since this step encodes individual verbs rather than phrases or sentences, the input is more linguistically homogeneous, and the most salient differences for the PCA are the value judgements. Schramowski et al. (2022) look for a high amount of variance being explained by the first principal component, which would indicate a good approximation of the moral direction. Then, they use the resulting transformation to obtain a floating point score for a set of templated questions, such as, "Should I smile to a friend?", "Is it exemplary to smile to a stranger?". They pose each action to the model in ten different templates, then average over the prompts. The scores of these initial actions are then normalised to lie within $[-1, 1]$, but subsequent scores can still lie outside this range. Schramowski et al. (2022) also conduct both a German regional and a larger user study on Amazon MTurk to obtain reference scores for the statements in question, only with English-language models.

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. We changed some of the questions to ensure good translation[1]. Our primary measure is the correlation of resulting model scores with human responsesfrom the global user study conducted by 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 monolingual models used. Our initial results with mean-pooling are listed in Table 1. However, this generally did not achieve a correlation with the user study. There were some exceptions to this rule—i.e., the Chinese monolingual BERT, and the English and Chinese portions of mBERT.

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	0.79	—	—	—	—
XLM-R (S-BERT)	0.85	0.82	0.85	0.83	0.81

Table 1: Experiments with different pre-trained mono- and multi-lingual models in the MORALDIRECTION framework. First three rows show mean-pooled sentence embeddings and the last two rows show embeddings resulting from sentence-transformers (Reimers and Gurevych, 2019).

Table 1 shows results from the monolingual, large English S-BERT, and an existing S-BERT version of XLM-R[2] (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. As we elaborate there, we aim to change semantic alignment in the PMLM as little as possible before probing it. We also need S-BERT versions of our 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[3]. 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. Each monolingual model was tuned with the matching translation, while the XLM-R$_{Base}$ model 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 (with-

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

[2] We used `sentence-transformers/xlm-r-100langs-bert-base-nli-mean-tokens`

[3] https://github.com/UKPLab/sentence-transformers/blob/master/examples/training/nli/training_nli_v2.py

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: Experiments with our mono- and multi-lingual S-BERT models in the MORALDIRECTION framework.

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 within our S-BERT models on the global 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.

out explicit alignment). For more training details, see Appendix B. We will release the resulting S-BERT models to the Huggingface hub.

3.3 Results

Figure 1 shows the MORALDIRECTION score of selected verbs, as done in Schramowski et al. (2022), evaluated for the monolingual S-BERT models separately (left) and on the multilingual XLM-R model all together (right). The scores overall seem commensurate, getting more aligned with lower variance on the multilingual model, except for few outliers. The verbs "divorce" and "drink" had in the monolingual case contrary sign for some languages. While "divorce" remains opposing, "drink" seemingly gets more aligned in the multilingual model. The variance decreases for the verbs "love" and "drink" and increases for "pollute" and "kill".

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. For Arabic, as well as the Czech portion of XLM-R, the correlation is slightly lower than the other models. Since in the case of Czech, the correlation is higher in the monolingual model, this seems to be a flaw of its representation in XLM-R. For Arabic, it may be a flaw or actually a slight difference in attitude. Unfortunately, we have no human responses to MFQ from Arabic 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. At the diagonal we see that English, German and Chinese correlate high when comparing their mono and multilingual models embeddings. The lowest correlation are the Czech and Arabic portions, which again may point to a flaw in the representations. Note that these two languages also produce the outliers as previously observed on the tested verbs with Figure 1. The monolingual S-BERT models are generally at a similar level of correlation with each other as the multilingual model.

Summarised, these observations extend the results from Schramowski et al. (2022) to a multilingual setting and indicate that multilingual LMs indeed capture moral norms. The high mutual correlations of scores shows 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 particularly designed to grasp cultural differences. We thus add further experiments to focus on this question.

4 Qualitative Analysis on Parallel Data

To better understand how these models generalise for various types of texts, we conducted a qualitative study using parallel data. We assume that for a parallel sentence pair, the MORALDIRECTION scores should be similar regardless of the model. Sentence pairs where the scores differ considerably may indicate cultural differences, or inadequacies of the models. In practice, very large score differences appeared to be more related to the latter. This type of understanding is important for further experiments with these models.

We conducted our analysis on OpenSubtitles parallel datasets (Lison and Tiedemann, 2016)[4], 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

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

de	en	monoling de	monoling en	XLM-R de	XLM-R 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*
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*.

behaviour towards each other, and thus may carry some moral sentiment. We used English-German and English-Czech data for our analysis.

Our analysis focuses on sentence pairs with very different scores. However, we take steps to filter out mistranslated sentence pairs—see Appendix D. Below, we discuss some 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 respective monolingual model, and the monolingual models with each other. Examples are listed in Table 4, and Table 7 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 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.

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, so that we can compare the model scores with data from previous studies. We use the translations provided on the Moral Foundations website for all five languages.[5]

Since the first part of the MFQ consists of very complex questions, we rephrase these into simple statements (see Appendix F). 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

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

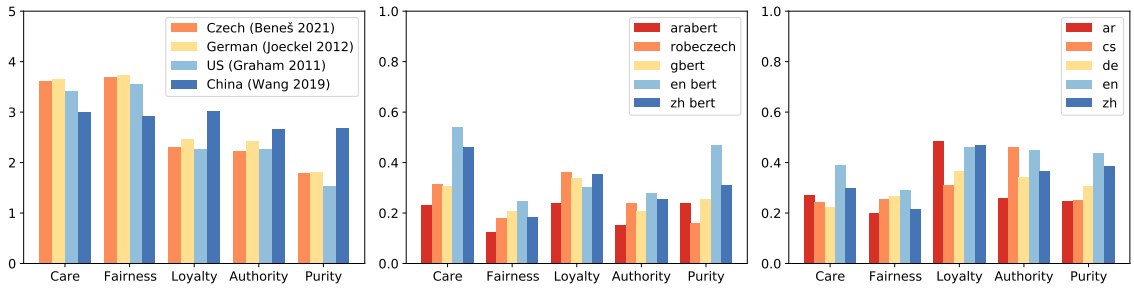

Figure 2: MFQ aspect scores from humans and models. Left: Examples of human data from studies in different countries. Middle: Scores obtained from monolingual MORALDIRECTION models. Right: Scores from XLM-R MORALDIRECTION in five languages.

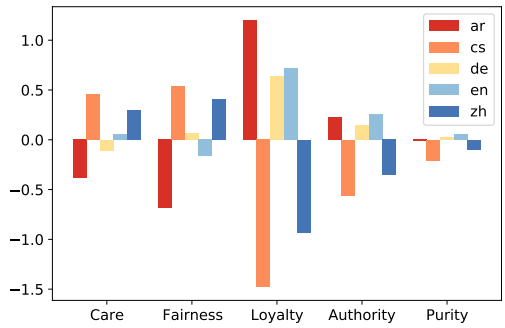

Figure 3: Sanity check—MFQ aspect scores from the XLM-R MORALDIRECTION models without Sentence-BERT tuning. This model had not obtained good correlations with human scores in § 3.

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 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 8.

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 5 in the Appendix.

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 ev-

idence 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 show 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 successfully 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).

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.

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 to evaluate 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 highly depends, 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 entail 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.

Broader Impacts

In this section, we recall the limitations of our methods and discuss risks which are important to take into consideration.

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 highly resourced, 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.

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.

Potential Risks. 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 reify 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.

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.

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.

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 M. 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)*. 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.

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.

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).

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.

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). *ArXiv*, 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.

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. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models.

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.

A Details of Monolingual Models Used

Table 5 lists the monolingual models we tuned and evaluated with their exact names and sizes.

B Sentence-BERT Tuning Procedure

We follow the training script provided by Reimers and Gurevych (2019) in the sentence-tranformers repository. As training data, we used 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 translated 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 for our training and experiments. Table 6 lists further training parameters.

C 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.

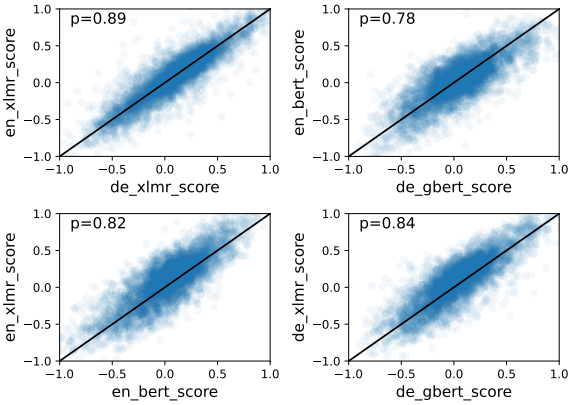

Figure 4: Correlation of the MORALDIRECTION scores for all German-English model combinations on the OpenSubtitles dataset.

D OpenSubtitles Filtering Details

Figure 4 shows the statistical correlation of the MORALDIRECTION scores on the OpenSubtitles dataset, evaluated for the German-English text 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 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.

E OpenSubtitles: Further Czech-English Examples

Table 7 shows further examples from the Czech-English OpenSubtitles corpus.

F Rephrased MFQ Questions

The first 15 questions of MFQ have the format, "When you decide whether something is right or

Language	Model Name	Type	Params
ar	aubmindlab/bert-base-arabertv02 (Antoun et al., 2020)	BERT	110M
cs	ufal/robeczech-base (Straka et al., 2021)	RoBERTa	125M
de	deepset/gbert-base (Chan et al., 2020)	BERT	110M
en	bert-base-cased (Devlin et al., 2019)	BERT	110M
zh	bert-base-chinese (Devlin et al., 2019)	BERT	110M

Table 5: The monolingual pre-trained language models, which we then tuned with the S-BERT framework and used for our experiments.

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 6: Sentence-BERT tuning parameters.

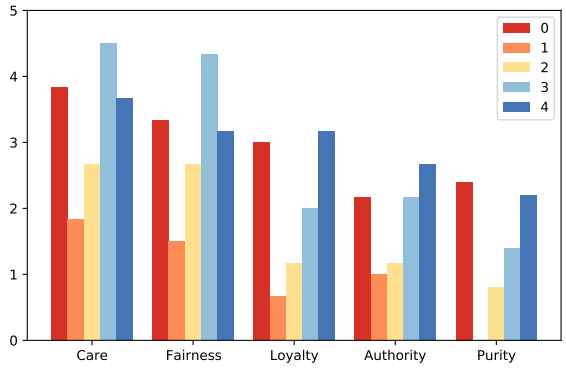

Figure 5: Example of human variation. Five different respondents from the German data collected by Joeckel et al. (2012)

.

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

G 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 over-reliance 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.

		monoling		XLM-R	
cs	**en**	**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 7: 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*.

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