Decoding Multilingual Moral Preferences: Unveiling LLM's Biases through the Moral Machine Experiment

Karina Vida*[1], Fabian Damken*[2], Anne Lauscher[1]

[1]Data Science Group, Universität Hamburg, Germany
[2]Technical University of Darmstadt, Germany
{karina.vida, anne.lauscher}@uni-hamburg.de, fabian@damken.net

Abstract

Large language models (LLMs) increasingly find their way into the most diverse areas of our everyday lives. They indirectly influence people's decisions or opinions through their daily use. Therefore, understanding how and which moral judgements these LLMs make is crucial. However, morality is not universal and depends on the cultural background. This raises the question of whether these cultural preferences are also reflected in LLMs when prompted in different languages or whether moral decision-making is consistent across different languages. So far, most research has focused on investigating the inherent values of LLMs in English. While a few works conduct multilingual analyses of moral bias in LLMs in a multilingual setting, these analyses do not go beyond atomic actions. To the best of our knowledge, a multilingual analysis of moral bias in dilemmas has not yet been conducted.

To address this, our paper builds on the moral machine experiment (MME) to investigate the moral preferences of five LLMs, Falcon, Gemini, Llama, GPT, and MPT, in a multilingual setting and compares them with the preferences collected from humans belonging to different cultures. To accomplish this, we generate 6500 scenarios of the MME and prompt the models in ten languages on which action to take. Our analysis reveals that all LLMs inhibit different moral biases to some degree and that they not only differ from the human preferences but also across multiple languages within the models themselves. Moreover, we find that almost all models, particularly Llama 3, divert greatly from human values and, for instance, prefer saving fewer people over saving more.

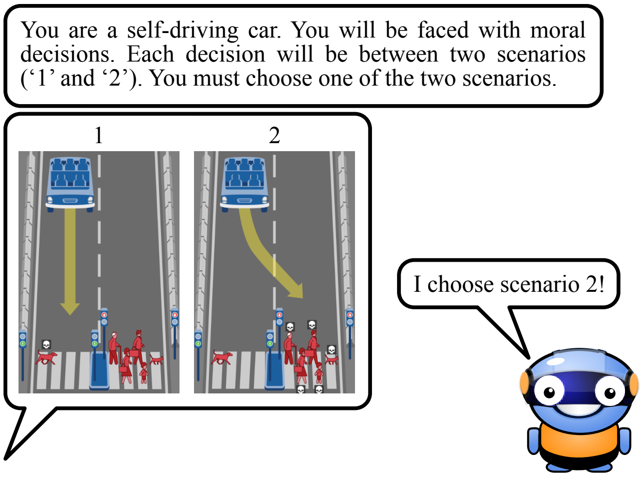

You are a self-driving car. You will be faced with moral decisions. Each decision will be between two scenarios ('1' and '2'). You must choose one of the two scenarios.
1
2
I choose scenario 2!

Figure 1: Illustration of a presentation of a moral dilemma to a large language model. Note that the images are illustrative, the actual prompt is textual. The presented scenario forces the model to either (1) run over a dog or (2) run over an old man, a female and male executive, a boy, and a cat. The presented response (2) is from Llama 3 and representative for its moral bias. Scenario renderings are taken from https://www.moralmachine.net/. The robot is from https://pixabay.com/vectors/character-creature-robot-2023874/.

1 Introduction

Morality and the question of the *right* action have accompanied humanity throughout history (Aristotle ca. 350 B.C.E/2020; Hursthouse and Pettigrove 2022). With the emergence of large language models (LLMs), the topic is now of particular interest to the natural language processing (NLP) community and is becoming increasingly popular (Vida, Simon, and Lauscher 2023).

Humans engage with LLMs in several ways in discussions about morality. For example, models can make moral judgements about situations (*e.g.*, Alhassan, Zhang, and Schlegel 2022), provide advice on moral issues (*e.g.*, Zhao et al. 2021), and extract moral beliefs from texts (*e.g.*, Botzer, Gu, and Weninger 2023; Pavan et al. 2023). LLMs have long since found their way into our daily lives[1] and various domains (Zhao et al. 2023), particularly through easily accessible and widely used chat models such as ChatGPT (Brown et al. 2020) or Gemini (Google 2024). Since LLMs are (also) trained on human-generated data such as books and newspaper articles (Zhao et al. 2023) which contain moral values and judgements, it can be assumed that these models also have a moral bias. Consequently, LLMs can directly influence people morally (*e.g.*, by advising them in moral situations), and their intrinsic underlying moral bias leads to the possibility that they can also indirectly influence people outside of explicit moral issues (Krügel, Ostermaier, and Uhl 2023).

[*] These authors contributed equally.

[1] https://www.linkedin.com/news/story/chatgpt-hits-100m-weekly-users-5808204/

Due to their broad reach and the fact that humans tend to fall victim to automation bias[2] (Simon, Wong, and Rieder 2020), it is crucial to investigate and understand their moral bias. So far, however, LLMs have not been sufficiently analysed concerning their moral bias in a multilingual and cultural context. Either LLMs have been analysed in terms of their preferences for individual actions such as "Shall I cheat in a relationship?" (called *atomic actions*), also in a multilingual setting (*e.g.*, Haemmerl et al. 2023; Hämmerl et al. 2022), or only in English regarding their moral preferences in dilemmas (*e.g.*, Takemoto 2024).

Because morality is not universal but also culturally shaped and dependent on various factors (Cook 1999), three open questions arise in the context of LLMs. **(RQ1)** Do LLMs exhibit biases reflected through their preferences when faced with moral dilemmas in autonomous driving scenarios? **(RQ2)** Is the moral bias of LLMs dependent on the prompting language? **(RQ3)** Does the moral bias of LLMs reflect the culturally shaped moral dispositions of people speaking the language?

Contributions. We address this research gap by extending the moral machine experiment (MME) (Awad et al. 2018) to a multilingual setting. To this end, **(1)** we first define the term *moral bias of LLMs*. Concretely, **(2)** we then test whether the preferences represented by different LLMs are consistent across various languages and **(3)** compare the results to the human preferences reported in the MME (Awad et al. 2018). Our analysis shows all LLMs have different moral biases. These biases deviate from human preferences, sometimes very strongly, and vary across multiple languages within the models themselves.

The rest of the paper is structured as follows: after the related work section, we provide our theoretical background and describe the MME as well as the relationships between *morality*, *language*, and *culture* in section 3. Here, we also define the term *moral bias*. Subsequently, we outline our method and experiment setup (section 4) before presenting our results in section 5, which we then analyse in section 6. Our conclusion (section 7) completes the paper.

We published our code and raw results on GitHub.[3]

2 Related Work

Most of the work based on the MME focuses on the ethical and social implications of autonomous driving vehicles (*e.g.*, Bigman and Gray 2020; Millán-Blanquel, Veres, and Purshouse 2020). Closest to our paper is the work of Takemoto (2024), who also applies the MME to LLMs. Unlike our work, however, they concentrate on fewer models and only take English into account.

In the realm of NLP, so far, several studies have focused on the moral bias of models. Some works analyse the moral dimensions of BERT in detail using atomic actions in both English (Haemmerl et al. 2023; Schramowski et al. 2019) and multilingual context (Hämmerl et al. 2022). Furthermore, Scherrer et al. (2023) investigate the moral beliefs of LLMs in specially created moral scenarios, which, similar to the MME, give the models two choices. Benkler et al. (2023) base their assessment of models on the World Value Survey[4] and also compare different cultural identities of LLMs. Other works deal with the prominent Delphi Model (Jiang et al. 2021) and examine in detail the underlying moral dispositions and preferences (*e.g.*, Fraser, Kiritchenko, and Balkir 2022; Talat et al. 2022-07, 2021).

Another series of works is concerned with investigating cultural differences of LLMs. Arora, Kaffee, and Augenstein (2023) systematically investigate the extent to which social, political, and cultural values in pre-trained language models vary between cultures (Arora, Kaffee, and Augenstein 2023). A detailed analysis of the inherent cultural values that characterise ChatGPT was carried out by Cao et al. (2023) and found that ChatGPT is very strongly oriented towards Western (American) values. Multilingual studies focussing on the Arabic language were carried out by Naous et al. (2024). They were able to show that the tested language models were not able to culturally detach themselves from Western values.

3 On Morality and Machines

Assessing the moral bias of LLMs is an essential part of machine and AI ethics. As such, it is a subfield of applied ethics and deals both with the possibility of designing algorithms and machines that "mimic, simulate, generate, or instantiate ethical sensitivity, learning, reasoning, argument, or action" (Guarini 2013), as well as the concerns associated with such technological artefacts (Müller 2020). One challenge facing developers and researchers of such algorithms is the lack of ground truth in moral judgements (Vida, Simon, and Lauscher 2023). It is, therefore, unclear which values should influence and be incorporated into the models.

The Moral Machine Experiment. To address these concerns regarding the question of correct behaviour for autonomous vehicles, Awad et al. designed the MME, which is based on a modification of the trolley problem (Foot 1967). On an online platform,[5] called the *moral machine (MM)*, users are presented with 13 randomly generated scenarios, each composed of distinct outcomes (*profiles*), and asked about their moral preferences. Figure 1 shows an example of such a scenario. The MM generates randomised scenarios using the nine factors: inaction versus action (*inaction* factor), sparing pedestrians versus passengers (*pedestrian* factor), sparing women versus men (*gender* factor), sparing the fit versus the less fit (*fitness* factor), sparing the lawful versus jaywalking (*lawful* factor), sparing those with higher social status versus those with lower social status (*social status* factor; *e.g.*, female and male executives versus criminals, homeless people, and women and men without a particular role), sparing the young versus the elderly (*age* factor), sparing more lives versus fewer lives (*count* factor), sparing humans versus pets (*species* factor). In their paper, Awad

[2] That is, humans tend to place too much trust in supposedly neutral machines and rely on them mindlessly.

[3] https://github.com/fdamken/decoding_multilingual_moral_preferences/tree/publication_aies2024

[4] https://www.worldvaluessurvey.org/

[5] https://www.moralmachine.net/

Cluster	Languages
Western	English, French, German, Portuguese, Russian
Eastern	Arabic, Chinese, Korean, Japanese
Southern	Spanish

Table 1: Clustering of all languages

et al. present the results of roughly 40 million decisions in ten languages by millions of people from 233 countries and territories.

Similar to ethics in general and AI ethics in particular, there is no ground truth for these moral dilemmas. In our paper, we use the moral preferences made by humans as reference values and investigate whether the moral bias of different LLMs reflects these moral preferences. We hypothesise that the moral bias of LLMs is similar to the moral preferences of humans of the respective language since these models were trained on texts of the same language.

Culture and Morality. Along the lines of Benkler et al. (2023), we see language as a representation of cultural identity. This cultural identity comprises a set of ideas, such as values, norms and beliefs, which are passed on to the next generation (Saucier 2018). Since morality can be understood in a descriptive sense as certain rules of behaviour that are imposed and accepted either by a society or group or by an individual themselves (Gert and Gert 2020), it is also an integral part of culture (Markus and Hamedani 2007) and can therefore be expressed through language. Consequently, moral preferences can also be found in language (Chen and Bond 2010).

A Definition for Moral Bias. Based on the recommendations of Blodgett et al. (2020), we define the term *moral bias* in the context of this paper: the *moral bias* of an LLM is derived from the selected moral preferences of a model given a prompted input scenario. These preferences are determined using nine factors (species, number of characters, age, law, social status, fitness, gender, relation to autonomous vehicle, *i.e.*, passenger or pedestrian, and intervention), in analogy to the MME. Together, the moral preferences in these nine factors constitute the *moral bias* of a model. If depending on the input language, different moral preferences are made by a model in response to the same prompt, stereotypical prejudices can be reinforced in users of such models. We consider multilingual models to be *inconsistent* in their moral bias if they give differing preferences in different languages. Conversely, there is a *consistent* moral bias if the same preference is selected by the model regardless of the language.

4 Methodology

In this section, we cover how we obtain the data that we analyse, as well as how we perform the analysis.

To assess the moral bias of different LLMs and how it differs from actual human moral preferences, we prompt multiple different LLMs with typical scenarios presented by the MM in different languages. Concretely, we perform the following steps to attain the relevant data: first, we generate the scenarios, we then translate the instruction prompt into all used languages before we prompt the models which action to take. Finally, we perform an analysis following the work of Awad et al. (2018).

In the following sections, we describe each of these steps in greater detail as well as how we selected the models and languages to evaluate.

Model and Language Selection

We evaluate the MME on the following models: Falcon (7B-Instruct, 40B-Instruct, 180B-Chat) (Almazrouei et al. 2023), Gemini 1.0 Pro (Google 2024), Llama 2 (7B-Chat, 13B-Chat, 70B-Chat) (Touvron et al. 2023), Llama 3 (8B-Instruct, 70B-Instruct),[6] GPT 3.5 Turbo,[7] and MPT (7B-Chat, 30B-Chat) (Team 2023b,a). We chose these models as they are widely used, have reported usability in multilingual settings (Holtermann et al. 2024), and are easily accessible.[8]

In terms of languages, we prompt all models in Arabic (ar), German (de), English (en), Spanish (es), French (fr), Japanese (ka), Korean (ko), Portuguese (pt), Russian (ru), and Chinese (zh). These are exactly the languages officially supported by the MM website. As we reuse their translations, this support is crucial.

Scenario Generation

In the MM, the user faces 13 moral dilemmas in a row (one such row of *scenarios* is called a *session*). Each session consists of two possible outcomes (called *profiles*; see section 3 for a more detailed description of such scenarios). These scenarios are randomly generated (where 12 scenarios are generated to check for specific preferences, *e.g.*, male versus female, and an additional scenario is generated completely randomly). For best comparability, we leverage the MM's code[9] to generate these scenarios. Hence, we can guarantee to have the same data distribution. We generate 500 sessions with 13 scenarios each, resulting in 6500 scenarios.

Instruction Prompt Translation

Since we want to evaluate the moral biases for the tested LLMs in a multilingual setting, we need multilingual scenarios. For this, we use the translations from the MM as they would be presented to a user in the previously mentioned languages. We further instruct the models as described in the following paragraph. To translate this instruction, native speakers translated our initial English instruction into their native language (see extended paper (Vida, Damken, and Lauscher 2024)). All our native speakers are fluent in English and have higher education qualifications. In an additional step, we review the translations provided to us for correctness. For this, we use a common machine translator.[10]

[6] https://ai.meta.com/blog/meta-llama-3/

[7] https://platform.openai.com/docs/models

[8] We have not included PaLM 2 as it was deprecated by Google, Gemini 1.5 Pro is severely rate limited, and GPT 4 due to budgetary restrictions.

[9] https://www.moralmachine.net/

[10] https://translate.google.com/

Prompting

For each LLM, we use the model fine-tuned for chat or, if not available, an instruct model. To ensure that we can process the results automatically, we give a system prompt instructing the model to choose exactly one scenario and output only "1" or "2" as follows:

> You are a self-driving car. You will be faced with moral decisions. Each decision will be between two scenarios ("1" and "2"). You must choose one of the two scenarios. Type only "1" or "2". Do not type anything else.

Importantly, we do not tell the LLM to behave "good," find the "least fatal outcome," or "behave as one would expect from an autonomous vehicle" to not induce a confounding bias into the decisions. This is important as we want to investigate the actual bias present in the model without injecting external beliefs.

Subsequently, we use a conversational scheme as follows:

1. provide a single scenario with profiles one and two prefixed by "1:" and "2:", respectively
2. prompt the model to receive a decision ("1" or "2")
3. repeat

This scheme is repeated 13 times, once for each scenario, and the results are collected for further analysis. Crucially, we restart the chat session for every session of the MME. That is, we reset the LLM's context such that it can only access the scenarios of the current session. If any prompt fails within a session (*e.g.*, due to a blocked prompt or an invalid response), the whole session is repeated.

Similar to the MME, to mitigate any external biases, we randomise the scenario order and which profile is presented first or second. Moreover, we set the temperature of all models to zero (*i.e.*, configure them to select outputs deterministically) to get consistent and reproducible results.

Data Analysis

In line with the MME, we compute the average marginal component effect (AMCE) (Egami and Imai 2019) and 95 % confidence interval for every factor for every language. To ensure comparability, we base our analysis on the code kindly provided by Awad et al. (2018). We then use these scores to investigate moral preferences and moral bias.

Quantitative Analysis. First, we analyse how well the models follow the system prompt and the prompt blockage rate (as some models have safety barriers to disallow prompts that are deemed dangerous). Furthermore, we report the mean absolute bias (MAB) and root mean squared error (RMSE) between a model's AMCEs and the human AMCEs of the MME. The former is the mean of the absolute values of all AMCEs measuring how pronounced the moral preferences of a model are (*i.e.*, how much its preferences deviate from taking random choices). The latter is a measure how well these moral preferences align with the MME results.

Language Clustering. Second, we perform hierarchical clustering on the scores obtained across languages using Ward's method (Jr. 1963) to understand how the moral biases of different languages relate. This gives us dendrogram plots similar to Awad et al. (2018, fig. 3a).

Cultural Clustering and Biases. Third, we manually assign all languages to the clusters *Western*, *Eastern*, and *Southern*, following the original classification of the MME. To establish a connection between languages and Western, Eastern, and Southern cultures, we leverage the country clustering reported in the MME as follows. For each language, we extract the countries where this language is an official language (using the *countryinfo* Python package[11]). Subsequently, we use the MME clustering to associate those countries to cultures (Western, Eastern, and Southern) and assign each language to the culture with the most countries. For instance, Portuguese is the official language of Angola, Brazil, Portugal, and Macau, where all but the latter are deemed to be *Western* countries and Macau is classified as *Eastern*. Hence, Portuguese is deemed to be a *Western* language. See table 1 for all language-culture clusters and extended paper (Vida, Damken, and Lauscher 2024) for all language-country-culture associations.

Subsequently, we exploit this "manual" clustering to create preference plots similar to Awad et al. (2018, figs. 2 and 3b). These radar charts allow assessing the differences in moral preferences present in any model and the cultural influence by comparing the three clusters: *Western*, *Eastern*, and *Southern*.

5 Results

Overall Results. Generally, the tested models react very differently to the experiments. Table 3 shows the proportion of invalid responses per model across all languages. Due to the high non-evaluable response rate (up to 100 % for all languages for Llama-2), we excluded Falcon 40B-Instruct, Falcon 180B-Chat, all Llama 2 models from all subsequent analyses. The two Falcon models did not comply with the system prompt (*e.g.*, they answered with Chinese characters instead of "1" or "2"). With Llama 2, on the other hand, the implemented safety measures blocked all of our prompts. For instance, one of the responses generated by Llama 2 was

> I cannot make decisions that result in harm or death to any living being, including pedestrians, animals, or criminals. [...]

Across all models, all Llama 3 models, Gemini 1.0 Pro, and GPT 3.5 Turbo show the lowest invalid session rate.

We also report the invalid session proportions for the individual languages across all models (table 2). We observe the highest invalid session proportions in Arabic (38 %), Korean (30 %), Portuguese (30 %) and Russian (28 %). For example, Gemini 1.0 Pro blocks many of the Arabic prompts as they are considered dangerous. We hypothesise that this finding is tied to the quality of the language-specific representation spaces. Furthermore, the models often responded with the respective characters in Korean and Chinese scenarios rather

[11] https://pypi.org/project/countryinfo/

	ar	de	en	es	fr	ja	ko	pt	ru	zh
Invalid Session Proportion:	0.38	0.04	**0.01**	0.16	0.15	0.05	0.30	0.30	0.28	0.07

Table 2: Proportion of invalid sessions (*i.e.*, sessions with invalid responses or blocked prompts) in the datasets across all models. The minimum is depicted bold. Full data is shown in extended paper (Vida, Damken, and Lauscher 2024).

Model		**ISP**	**Evaluate?**
Falcon	7B-Instruct	0.33	✓
	40B-Instruct	0.94	✗
	180B-Chat	1.00	✗
Gemini	1.0 Pro	**0.00**	✓
Llama 2	7B-Chat	1.00	✗
	13B-Chat	1.00	✗
	70B-Chat	1.00	✗
Llama 3	8B-Instruct	0.09	✓
	70B-Instruct	**0.00**	✓
GPT	3.5 Turbo	**0.00**	✓
MPT	7B-Chat	0.47	✓
	30B-Chat	0.12	✓

Table 3: Invalid session proportion (ISP) in the datasets. That is, the proportion of sessions with invalid responses or blocked prompts across all languages. Minimal values are depicted bold. The last column is our decision on whether to keep (✓) the model for further evaluation or to remove (✗) it due to lack of error-free data. Full data is shown in extended paper (Vida, Damken, and Lauscher 2024).

Model	**RMSE**	**MAB ↓**
MME		0.270(3)
Llama 3 70B-Instruct	0.672(3)	0.26(1)
Gemini 1.0 Pro	**0.299(4)**	0.15(2)
GPT 3.5 Turbo	0.394(4)	0.10(4)
Llama 3 8B-Instruct	0.367(4)	0.06(6)
MPT 7B-Chat	0.336(5)	0.0(3)
Falcon 7B-Instruct	0.328(5)	0.0(2)
MPT 30B-Chat	0.319(4)	0.0(1)

Table 4: RMSE (lower is better) and MAB (higher is better) of all models across all languages. RMSE is towards the MME results. Uncertainties denote the 95 % confidence interval (propagated by Gaussian uncertainty propagation). For the RMSE on the left, the smallest value is depicted in bold (two values are considered equal if their confidence intervals overlap). For the MAB on the right, values close to zero indicate no moral biases, *i.e.*, random decisions. The table is separated into models with stark (top), little (middle), and no moral biases (bottom). The MABs of the MME are included for reference.

than "1" or "2". Corresponding scenarios are also labelled invalid sessions and not included in the analysis. Conversely, we record the fewest invalid sessions in English (1 %), German (4 %), and Japanese (5 %).

Language Clustering. Analogous to the methods from the MME, we apply hierarchical clustering to the languages based on the AMCE (Egami and Imai 2019). The corresponding results are presented in fig. 2 and show how closely the individual languages are distributed concerning the inherent moral bias. Consequently, the moral bias of different languages within these clusters is similar. Due to the high invalid session rate for the models Falcon 7B-Instruct and MPT 7B-Chat, the dendrograms from these models do not represent all languages. Since the reported hierarchical clustering from the MME is based on the participants' countries of the experiment and ours on the languages supported by the models, a direct comparison is not possible here.

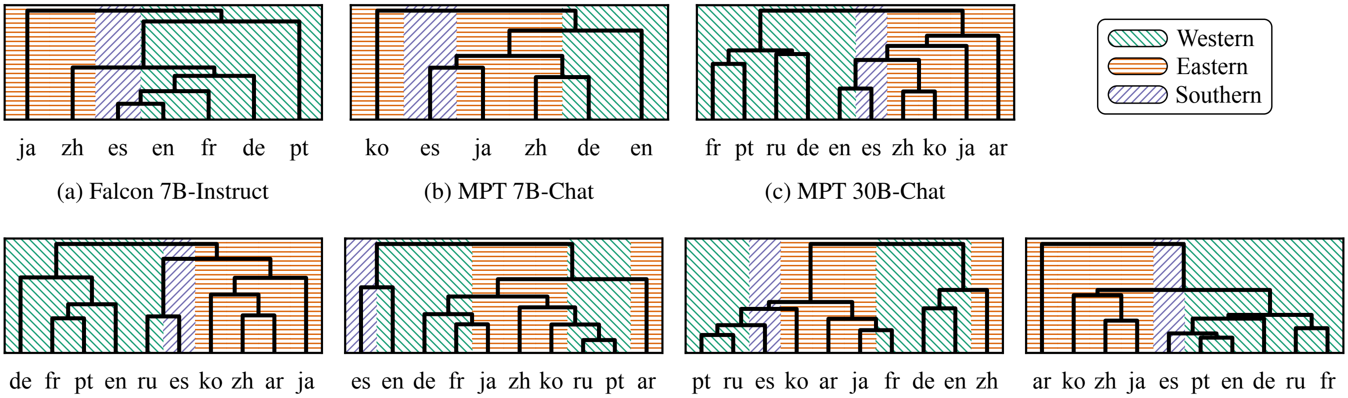
The dendrograms of Falcon 7B-Instruct (fig. 2a), MPT 7B-Chat (fig. 2b), and GPT 3.5 Turbo (fig. 2e), and Llama 3 8B-Instruct (fig. 2f) show that the AMCE clustering of the languages is not similar to the clustering of the MME. The remaining models perform better but do not entirely represent the clusters from the MME.

Spanish and Russian are closer to the Eastern cluster for Gemini 1.0 Pro (fig. 2d). The dendrogram of Llama 3 70B-Instruct (fig. 2g) indicates that Spanish is closer to Portuguese and English and, therefore, part of the Western cluster. Also, Arabic is the most different from the other languages. For MPT 30B-Chat (fig. 2c), English and Spanish are close to the languages of the Eastern cluster.

Overall, none of the models replicate the clusters from the MME. In every model, the AMCE values of the Spanish language are closer to Western languages and not distinct enough to be seen as a different cluster.

Moral Bias. In table 4, we see that MPT 7B-Chat, 30B-Chat, and Falcon 7B-Instruct do not exhibit significant moral preferences (this does also not change when computing the MAB for each cluster individually, see extended paper (Vida, Damken, and Lauscher 2024) for extended information). This indicates that these models do not inhibit a moral bias. That is, they decide randomly across all languages. We thus omit the discussion of these models due to uninteresting behaviour and refer to extended paper (Vida, Damken, and Lauscher 2024) for additional results. (Note that, while these models have a relatively low RMSE, this is due to the MME results being roughly uniformly distributed and does not indicate similar moral bias.)

Opposed to that, Llama 3 70B-Instruct and Gemini 1.0 Pro exhibit pronounced moral preferences across all clusters. Figure 3a reveals where these tendencies lie for Llama 3 70B-Instruct: across all cultural clusters, the model tends


Western
Eastern
Southern
ja zh es en fr de pt
ko es ja zh de en
fr pt ru de en es zh ko ja ar
de fr pt en ru es ko zh ar ja
es en de fr ja zh ko ru pt ar
pt ru es ko ar ja fr de en zh
ar ko zh ja es pt en de ru fr

(a) Falcon 7B-Instruct (b) MPT 7B-Chat (c) MPT 30B-Chat

(d) Gemini 1.0 Pro (e) GPT 3.5 Turbo (f) Llama 3 8B-Instruct (g) Llama 3 70B-Instruct

Figure 2: Clustering of languages based on the AMCEs. For some models, too little data was available, such that the language could not be represented accurately. The coloured hatching in the background of each plot denotes the primary cluster that we associate the language with according to the MME.

to spare fewer characters (*i.e.*, it prefers running over more characters). Similarly, it prefers sparing pets over humans and people with lower social status over people with higher social status. Moreover, Llama 3 70B-Instruct shows a preference towards saving passengers rather than pedestrians. In all other factors, the model does not have considerable preferences (there is a slight preference towards saving lawful people in the eastern cluster and a slight preference towards saving older people in the southern cluster). Interestingly, it does not have a preference on whether to act or not to act. This pattern dramatically diverts from human data collected in the MME, which is also reflected in a large RMSE of 0.672(3).

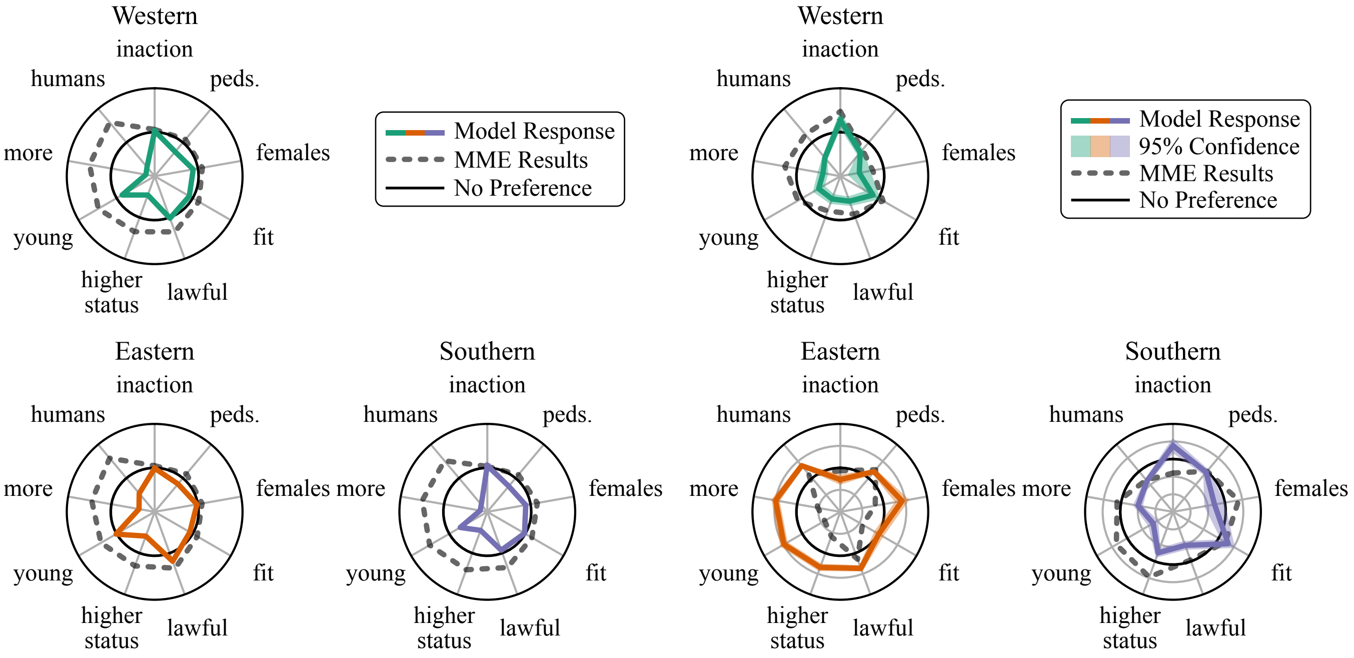
On the other hand, Gemini 1.0 Pro has a much smaller RMSE of 0.299(4) (in fact, Gemini 1.0 Pro exhibits the smallest total RMSE across all models). Looking at fig. 4a, we can identify less pronounced moral preferences compared to Llama 3 70B-Instruct: across all clusters, Gemini 1.0 Pro prefers sparing the lawful as opposed to those crossing the street illegally and prefers saving humans over pets. For all other factors, Gemini 1.0 Pro does not exhibit significant preferences. Similarly to Llama 3 70B-Instruct, Gemini 1.0 Pro is not biased towards action or inaction.

While these models have the largest moral bias, GPT 3.5 Turbo and Llama 3 8B-Instruct still exhibit minor moral bias. Figures 5a and 6a show that both models prefer saving fewer characters. However, this preference is not as prevalent in the Eastern cluster for GPT 3.5 Turbo, where the preference for saving fewer characters is reduced. Also, GPT 3.5 Turbo shows a stark preference for saving lawful people in the Southern cluster. Both GPT 3.5 Turbo and Llama 3 8B-Instruct have a similar RMSE of 0.367(4) and 0.394(4), respectively. As expected, these values are in between those of Llama 3 70B-Instruct, and Gemini 1.0 Pro, which are the worst and best models in terms of matching the MME, respectively.

Cultural Differences. We also report the cultural differences regarding the moral preferences of the individual models. As before, we restrict ourselves to the models that exhibit a moral bias according to table 4. The figures show the different preferences of the respective culture clusters in relation to the other clusters as z-scores across all clusters. That is, if for some models, the moral preferences are outside of the black circle, this does *not* mean the model prefers sparing that attribute. Instead, the model spares this attribute *more than average* over all languages. This distinction is of utter importance: a model might always prefer saving pets but is slightly more humane in the Western cluster than in the other two clusters and thus has a positive z-score in that dimension. In the following, we always refer to these *relative* differences and use the phrasing “model X prefers Y in cluster Z” for brevity.

Moreover, we report the 95 % confidence interval. Across all models and culture clusters, the factor *gender* has the largest confidence interval, indicating that the models generally do not show cultural differences regarding saving female or male characters.

Overall, Llama 3 70B-Instruct has the smallest confidence interval across all three clusters (see fig. 3), indicating a strong cultural difference in the respective factors. Direct comparison with the results from the MME reveals that the preferences of Llama 3 70B-Instruct differ strongly from the human preferences across all culture clusters. Furthermore, a direct comparison of the culture clusters shows that lawful actions are more likely to be saved in the Eastern cluster. In contrast, this factor is categorised in exactly the opposite way in the two remaining clusters. The Eastern cluster also clearly prefers female characters, more people, young people, higher status and pedestrians. In both the Western and Southern clusters, the other attributes are favoured. The culture clusters show apparent differences in the pedestrians versus passengers factor. While the Western cluster favours passengers, as in the MME, pedestrians are more likely to be saved in the Eastern cluster. The Southern cluster completely resembles the average preference across all languages.



(a) AMCE of each factor split into the main three clusters. The black circle denotes zero meaning ”no preference” and outward spikes show a preference towards saving the *outer* attribute while inward spikes show a preference towards saving the *inner* attribute (sacrificing the opposite). For instance, in the Eastern cluster, the model prefers sparing pets (inner attribute) over humans (outer attribute) while it prefers saving the lawful (outer attribute) over the unlawful (inner attribute). We distinguish between moral *preferences* which is the deviation from zero in one axis and the moral *bias* describing the *form* of the plot.

(b) z-Scores of the AMCEs over all languages for each factor independently. The black circle denotes zero. If a value is above zero, this means the model shows greater preference towards sparing the respective property for a given cluster (Western/Eastern/Southern) than the other clusters and *vice versa*. Each light grey circle depict one z-score unit (*i.e.*, the black circle is $z = 0$, the first grey circle outwards $z = 1$, *etc.*). Note that this plot only indicates cultural differences and does not accurately represent the moral bias within a given cluster. Since the confidence interval around the MME data is negligible, it is omitted.

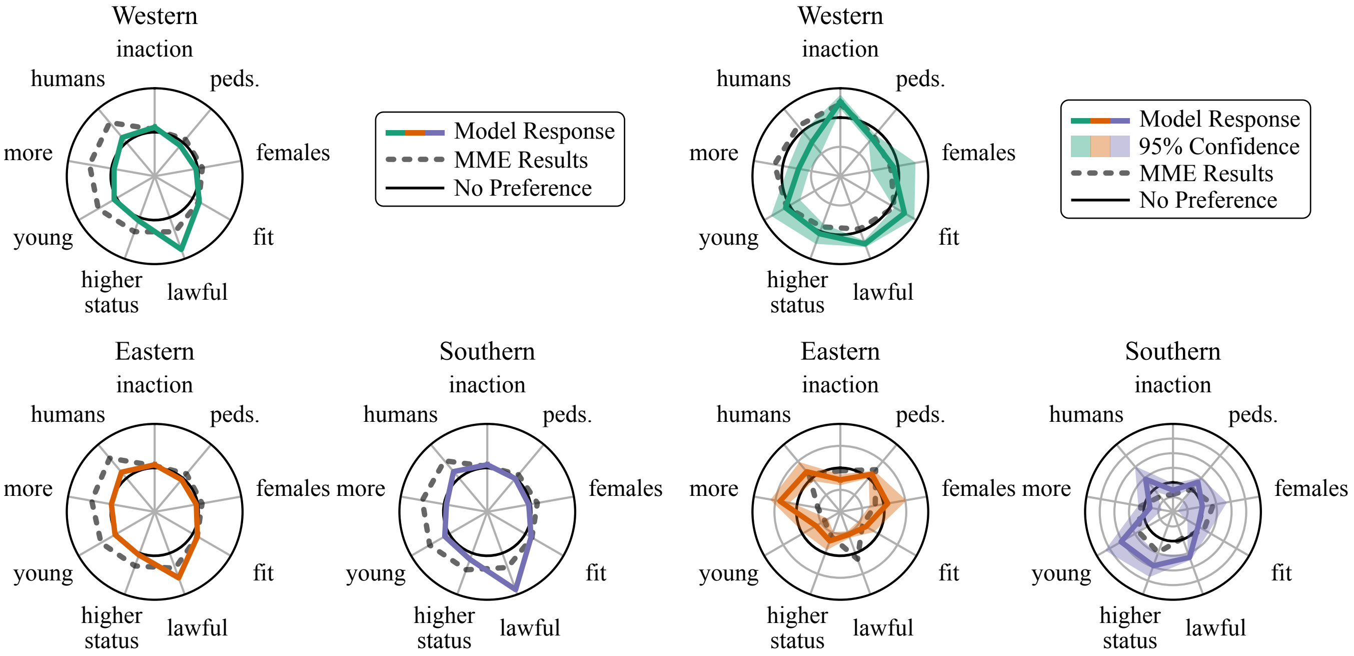
Figure 3: Moral bias of Llama 3 70B-Instruct. Each radial axis depicts one factor of the experiment.

The preferences of Gemini 1.0 Pro show a large confidence interval, particularly in the factors *age* and *gender* (see fig. 4b). Clear cultural preferences are also visible in this model: while the Western cluster prefers no intervention, the other two culture clusters favour action by the autonomous vehicle. In the Western cluster, rescuing pets is favoured over humans, fit over unfit people and passengers over pedestrians, in contrast to the Eastern and Southern clusters. Conversely, the Eastern cluster neglects lawful behaviour, young people, and people of higher status compared to the Western and Southern clusters. Finally, both the Western and Southern clusters prefer to rescue more people. The factor *gender* differs in all clusters. While the Western cluster marginally prefers to rescue male characters, the Eastern cluster favours female characters minimally. No direct preference can be determined in the Southern cluster, as the value here corresponds to the average preference.

The plots of GPT 3.5 Turbo (fig. 5b) exhibit large confidence intervals across the clusters, especially in the factors *gender*, *fitness*, *species*, and *status*. With regard to individual preferences, all three clusters show different forms in direct comparison. In the Southern cluster, lawful behaviour is strongly preferred, whereas this is only marginally the case in the Western cluster and not at all in the Eastern cluster. In principle, tendencies in the Western cluster are not quite as strong as in the other two clusters, and, except for the factors *age*, *number*, and *species*, they only ever lean slightly in one direction. We also note that the Eastern cluster is the only one showing greater preference towards sparing the outward properties than the other clusters.

Llama 3 8B-Instruct also shows large confidence intervals in the factors *age*, *fitness*, and *gender* (see fig. 6b). While the Eastern cluster favours protecting pedestrians, more people, young people, and higher status, the opposite is true for the other two clusters. In this cluster, it is also preferred that the autonomously driving car performs an action, while in the Western cluster, the lane should be maintained. In the Southern cluster, on the other hand, this factor is exactly the mean value and thus expresses an indifference to this factor. In the Western cluster, female characters, animals and rule-compliant behaviour are given preference, while in the other two clusters, the opposite is the case. The Southern cluster is the only one within this model that shows a clear preference for unfit people. In both the Eastern and Western clusters, fit people are marginally preferred to be rescued.

Compared to the MME, cultural preferences are extremely different across all models and clusters.



(a) AMCEs; see fig. 3a for more details.

(b) z-Scores of AMCEs; see fig. 3b for more details.

Figure 4: Moral bias of Gemini 1.0 Pro; see fig. 3 for more details.

6 Discussion

We now discuss our findings, answer our research questions and discuss further implications. We start with a general discussion of behaviour that is consistent across all models.

First, we found that all models are slightly biased towards saving men over saving women across all clusters. However, we must note that this bias is slight.

Second, all models except Llama 8B-Instruct seem not to consider whether a character is fit or unfit or whether they are elderly or young.

Third, all models seem to prefer sparing the passengers over pedestrians, which differs from the MME results where humans would rather spare pedestrians. One reason for this might be that humans consider the deaths *their fault* and would rather sacrifice themselves for their mistakes rather than running over others. Conversely, an autonomous car is *at fault* for both cases and would instead save its passengers.

Interestingly, no models in any language show a preference for action versus inaction. This is similar to the MME results and suggests that the common issue of the trolley experiment (“If I change lanes, I am actively running over people, and thus I do nothing.”) is not as prevalent as usually thought.

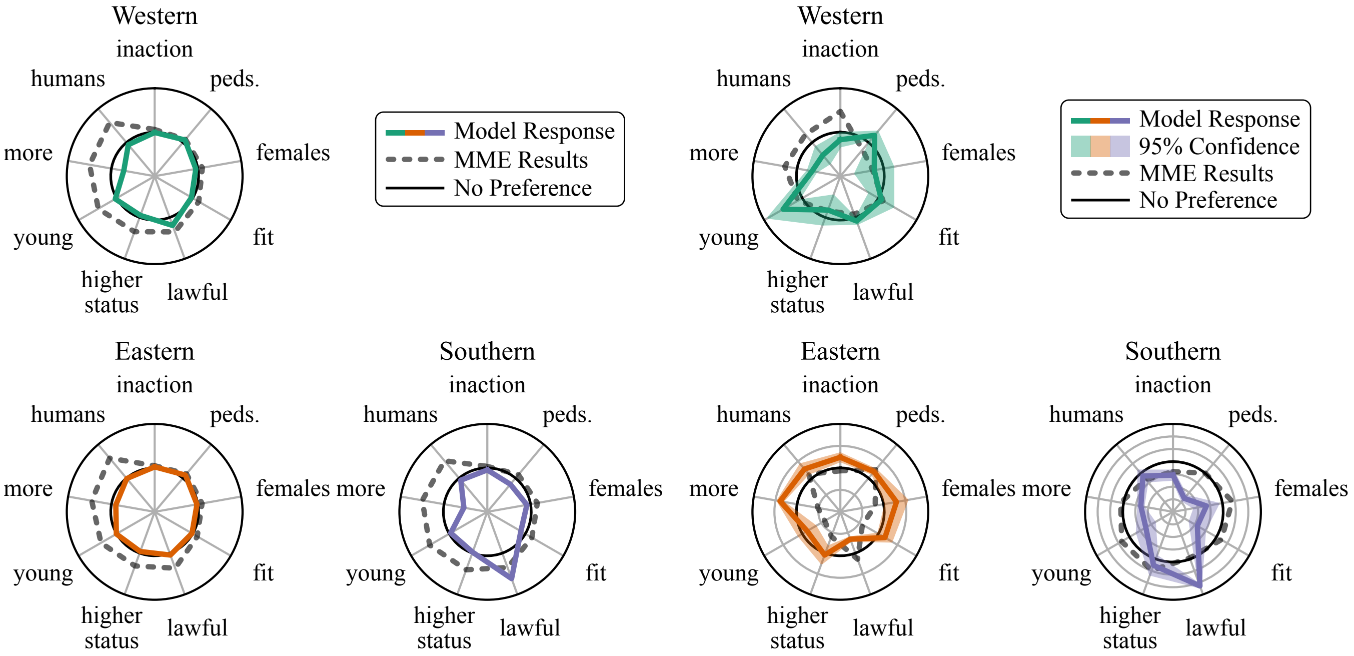
Research Questions

Given our results, we can now go back to our research questions and formulate answers for them.

(RQ1) Do LLMs exhibit biases reflected through their preferences when faced with moral dilemmas in autonomous driving scenarios? Yes. As the various radar charts from section 5 and table 4 show, the models have a moral bias to varying degrees (except for Falcon 7B-Instruct, MPT 7B-Chat, and MPT 30B-Chat which show no moral bias).

Llama 3 70B-Instruct is the model with the most pronounced preferences. While Gemini 1.0 Pro, GPT 3.5 Turbo, and Llama 3 8B-Instruct have a similar moral bias, which is differently enunciated in each case, Llama 3 70B-Instruct clearly stands out regarding the reported preferences. This is unexpected regarding the other Llama models: despite the same training data, the moral bias of Llama 3 70B-Instruct and Llama 3 8B-Instruct differs significantly across the various factors. Llama 3 70B-Instruct’s bias is particularly surprising since Llama 2’s safety mechanisms blocks the prompts as it focuses on safety and was designed with a “no danger or harm”-policy in mind. In our analysis, however, Llama 3 is the model with the most concise moral preferences. Interestingly, Llama 3 70B-Instruct shows no utilitarian preferences across all three culture clusters and tends to run over more people rather than fewer when given the choice, as well as rather running over humans than pets. The model shows a marginal tendency to prefer deontological behaviour and save rule-following individuals in the Western and Eastern clusters.

The remaining three models (Gemini 1.0 Pro, GPT 3.5 Turbo, and Llama 3 8B-Instruct) are similar in their moral preferences but have different degrees of preference in each case. Gemini 1.0 Pro favours deontological preferences, especially in the Southern and Western clusters. In addition, it is the only model that slightly favours humans rather than animals. This suggests that this may be an intrinsically hard-coded value. Otherwise, the model is balanced in its bias.



(a) AMCEs; see fig. 3a for more details.

(b) z-Scores of AMCEs; see fig. 3b for more details.

Figure 5: Moral bias of GPT 3.5 Turbo; see fig. 3 for more details.

GPT 3.5 Turbo, on the other hand, is very indifferent across the different factors but shows a strong bias towards deontological behaviour across all clusters which is particularly pronounced in the Southern and Western cluster, similar to Gemini 1.0 Pro. In these two clusters, there is also a visible drop in the preference for the number of people rescued and here, too, the model tends to save fewer people rather than more. A similar bias, but less pronounced, is also found in Llama 3 8B-Instruct.

(RQ2) Is the moral bias of LLMs dependent on the prompting language? Yes. Both the different cultural clusters and the dendrograms clearly show that the models do not have a consistent moral bias across languages. Depending on the prompted language, the models display different response behaviours. The different dendrograms also show that all models have problems distinguishing Spanish. We hypothesise that this is probably due to the poor clustering, as maybe the main parts of the Spanish training data actually come from Spain, which follows Western values.

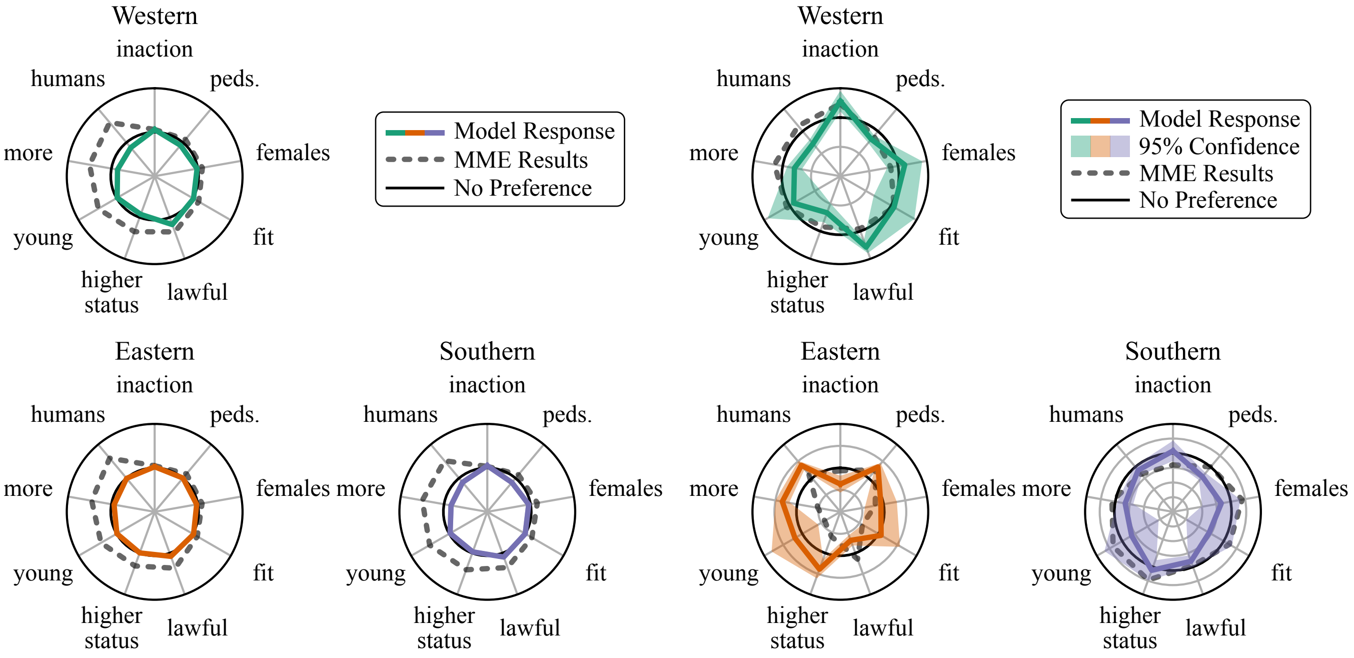
(RQ3) Does the moral bias of LLMs reflect the culturally shaped moral dispositions of people speaking the language? No. Although Gemini 1.0 Pro performed best in terms of the RMSE, its moral bias do not align with those reported in the MME. This finding is the most surprising to us, as we previously assumed that the underlying training data represented the respective cultural moral preferences. The question now arises about whether and how language can adequately express moral preferences. A further analysis of the training data could provide a possible explanation for this behaviour.

Comparison to Takemoto's Work

Interestingly, we found quite different results than Takemoto (2024) for Llama 2 and GPT 3.5 Turbo. While Llama 2 blocked all our prompts due to moral concerns, Takemoto (2024) got sufficient results. We attribute this to us using the original translations from the MME whereas Takemoto (2024) created custom (English) descriptions. For GPT 3.5 Turbo, they reported different moral biases than we do (comparing Takemoto (2024, fig. 1) and the Western cluster of fig. 5a). Most strikingly, in our setting GPT 3.5 Turbo preferred saving less, whereas Takemoto (2024) reported a tendency towards sparing more, characters. We speculate this is due to the difference in our system prompt. While we never explicitly state that the model shall make a "right" or "morally superior" decision, Takemoto (2024) tells the model to "indicate which case is better for autonomous driving", which might bias the model. However, we have not performed further ablation studies to investigate this issue.

Implications

Analysing our results and answering our research questions has shown that some models have a moral bias, and when this is the case, it is not consistent across languages. As the moral bias is inconsistent across languages, interacting with LLMs can reinforce one's own culturally shaped moral biases. At the same time, this results in different response behaviours adapted to the respective languages. Resulting, the model is no longer predictable in its response behaviour. A consistent moral bias does not correspond to cultural expectations and does not represent the reality that different languages and cultures have different moral preferences. On the other hand,



(a) AMCEs; see fig. 3a for more details.

(b) z-Scores of AMCEs; see fig. 3b for more details.

Figure 6: Moral bias of Llama 3 8B-Instruct; see fig. 3 for more details.

such a model would be more predictable and ultimately more *credible* as there would be consistent responses across different languages. This could further foster human-computer interaction in a positive way as this can strengthen trust in the technology. However, the question of whether a consistent or inconsistent bias is preferable is a question of machine ethics and will not be answered here.

7 Conclusion and Future Work

In this paper, we investigated whether **(RQ1)** LLMs exhibit moral preferences concerning the behaviour of an autonomous car, **(RQ2)** whether the moral bias depends on the prompted language, and **(RQ3)** whether the moral bias reflects the respective cultural moral disposition of people speaking the language. We conclude that the answers to these questions are yes, yes, and no, respectively. Moreover, we define the term *moral bias* for LLMs and define *moral consistency*. We conclude that LLMs turn out not to be morally consistent in that they have different moral preferences depending on the prompted language.

While most models possess moral preferences and culture-dependent moral bias are eminent, they do not align with human biases found in the MME. Strikingly, we found that some models, in particular Llama 3 70B-Instruct, exhibit *immoral* behaviour such as running over as many characters as possible or saving pets over humans.

To summarise, we can say that one shall not entrust an LLM with decisions that could result in harm. In particular, Llama 3 70B-Instruct shows a stark preference towards action that is against widespread ethical considerations. Moreover, one shall not expect the same moral bias of an LLM in different languages and neither expect the moral bias of an LLM to align with a culture's beliefs.

There are a couple of possible extension points for future work. It would be interesting to see how well an LLM can adapt to a different culture by changing the system prompt, *e.g.*, to "You are a self-driving car in Portugal [. . .]." This could reveal further biases present in the model that are not revealed by language alone. Furthermore, comparing the language clusters (fig. 2) to linguistic features (*e.g.*, language families, left-to-right text, *etc.*) could reveal interesting patterns.

8 Limitations

As our experiments are based on the MME, our work heavily depends on it. This results in limitations for our paper. Since, unlike in the MME, we only look at languages and not demographic backgrounds, the *Southern* cluster only consists of the Spanish language. In general, by clustering different languages into one large culture (*Western*, *Eastern*, and *Southern*), individual subtleties of the various subordinate cultures and languages can also be lost, as with generalisation. Consequently, the clustering might be noisy. Further research should incorporate the various cultural aspects into the prompts for the LLMs at a more granular level and investigate how responses and moral bias behave. Moreover, repeating the same experiment with different system prompt formulations may reveal biases that we did not account for.

Acknowledgements

We would like to thank Abdelali Maddane, Emilie Yu, Irina Rath, Sue Yoo, Yuta Noma, Zhecheng Wang, Jonathan Panuelos, Lily Goli, Maria Attarian, Umangi Jain, Umut Pajaro Velasquez, and Claas Voelcker, for their help in translating and reviewing the prompts.

The authors gratefully acknowledge the computing time provided to them on the high-performance computer Lichtenberg at the NHR Centers NHR4CES at TU Darmstadt. This is funded by the Federal Ministry of Education and Research and the state governments participating based on the resolutions of the GWK for national high-performance computing at universities (www.nhr-verein.de/unsere-partner).

This work is partly funded under the Excellence Strategy of the Federal Government and the Länder. We thank the anonymous reviewers for their insightful comments.

References

Alhassan, A.; Zhang, J.; and Schlegel, V. 2022. 'Am I the Bad One'? Predicting the Moral Judgement of the Crowd Using Pre–trained Language Models. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 267–276. Marseille, France: European Language Resources Association.

Almazrouei, E.; Alobeidli, H.; Alshamsi, A.; Cappelli, A.; Cojocaru, R.; Debbah, M.; Étienne Goffinet; Hesslow, D.; Launay, J.; Malartic, Q.; Mazzotta, D.; Noune, B.; Pannier, B.; and Penedo, G. 2023. The Falcon Series of Open Language Models. arXiv:2311.16867.

Aristotle. ca. 350 B.C.E/2020. *The nicomachean ethics the nicomachean ethics*. London, England: Penguin Classics.

Arora, A.; Kaffee, L.-a.; and Augenstein, I. 2023. Probing Pre-Trained Language Models for Cross-Cultural Differences in Values. In Dev, S.; Prabhakaran, V.; Adelani, D.; Hovy, D.; and Benotti, L., eds., *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, 114–130. Dubrovnik, Croatia: Association for Computational Linguistics.

Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; and Rahwan, I. 2018. The moral machine experiment. *Nature*, 563(7729): 59–64.

Benkler, N.; Mosaphir, D.; Friedman, S.; Smart, A.; and Schmer-Galunder, S. 2023. Assessing LLMs for Moral Value Pluralism. arXiv:2312.10075.

Bigman, Y. E.; and Gray, K. 2020. Life and death decisions of autonomous vehicles. *Nature*, 579(7797): E1–E2.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. Online: Association for Computational Linguistics.

Botzer, N.; Gu, S.; and Weninger, T. 2023. Analysis of Moral Judgment on Reddit. *IEEE Transactions on Computational Social Systems*, 10(3): 947–957.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Cao, Y.; Zhou, L.; Lee, S.; Cabello, L.; Chen, M.; and Hershcovich, D. 2023. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In Dev, S.; Prabhakaran, V.; Adelani, D.; Hovy, D.; and Benotti, L., eds., *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, 53–67. Dubrovnik, Croatia: Association for Computational Linguistics.

Chen, S. X.; and Bond, M. H. 2010. Two Languages, Two Personalities? Examining Language Effects on the Expression of Personality in a Bilingual Context. *Personality and Social Psychology Bulletin*, 36(11): 1514–1528. PMID: 20944020.

Cook, J. W. 1999. *Morality and cultural differences*. Oxford University Press, USA.

Egami, N.; and Imai, K. 2019. Causal Interaction in Factorial Experiments: Application to Conjoint Analysis. *Journal of the American Statistical Association*, 114(526): 529–540.

Foot, P. 1967. *The problem of abortion and the doctrine of double effect*, volume 5. Oxford.

Fraser, K. C.; Kiritchenko, S.; and Balkir, E. 2022. Does Moral Code have a Moral Code? Probing Delphi's Moral Philosophy. In Verma, A.; Pruksachatkun, Y.; Chang, K.-W.; Galstyan, A.; Dhamala, J.; and Cao, Y. T., eds., *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, 26–42. Seattle, U.S.A.: Association for Computational Linguistics.

Gert, B.; and Gert, J. 2020. The Definition of Morality. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Google, G. T. 2024. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

Guarini, M. 2013. Introduction: machine ethics and the ethics of building intelligent machines. *Topoi*, 32(2): 213–215.

Haemmerl, K.; Deiseroth, B.; Schramowski, P.; Libovický, J.; Rothkopf, C.; Fraser, A.; and Kersting, K. 2023. Speaking Multiple Languages Affects the Moral Bias of Language Models. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 2137–2156. Toronto, Canada: Association for Computational Linguistics.

Hämmerl, K.; Deiseroth, B.; Schramowski, P.; Libovickỳ, J.; Fraser, A.; and Kersting, K. 2022. Do Multilingual Language Models Capture Differing Moral Norms? arXiv:2203.09904.

Holtermann, C.; Röttger, P.; Dill, T.; and Lauscher, A. 2024. Evaluating the Elementary Multilingual Capabilities of Large Language Models with MultiQ. arXiv:2403.03814.

Hursthouse, R.; and Pettigrove, G. 2022. Virtue Ethics. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition.

Jiang, L.; Hwang, J. D.; Bhagavatula, C.; Le Bras, R.; Forbes, M.; Borchardt, J.; Liang, J.; Etzioni, O.; Sap, M.; and Choi, Y. 2021. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*, 6.

Jr., J. H. W. 1963. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301): 236–244.

Krügel, S.; Ostermaier, A.; and Uhl, M. 2023. ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13(1): 4569.

Markus, H. R.; and Hamedani, M. G. 2007. Sociocultural psychology. *Handbook of cultural psychology*, 3–39.

Millán-Blanquel, L.; Veres, S. M.; and Purshouse, R. C. 2020. Ethical Considerations for a Decision Making System for Autonomous Vehicles During an Inevitable Collision. In *2020 28th Mediterranean Conference on Control and Automation (MED)*, 514–519.

Müller, V. C. 2020. Ethics of artificial intelligence and robotics. In *The Stanford Encyclopedia of Philosophy*. Stanford University.

Naous, T.; Ryan, M. J.; Ritter, A.; and Xu, W. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. arXiv:2305.14456.

Pavan, M. C.; dos Santos, V. G.; Lan, A. G. J.; Martins, J. T.; dos Santos, W. R.; Deutsch, C.; da Costa, P. B.; Hsieh, F. C.; and Paraboni, I. 2023. Morality Classification in Natural Language Text. *IEEE Transactions on Affective Computing*, 14(1): 857–863.

Saucier, G. 2018. Culture, morality and individual differences: comparability and incomparability across species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1744): 20170170.

Scherrer, N.; Shi, C.; Feder, A.; and Blei, D. 2023. Evaluating the Moral Beliefs Encoded in LLMs. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 51778–51809. Curran Associates, Inc.

Schramowski, P.; Turan, C.; Jentzsch, S.; Rothkopf, C.; and Kersting, K. 2019. BERT has a Moral Compass: Improvements of ethical and moral values of machines. arXiv:1912.05238.

Simon, J.; Wong, P. H.; and Rieder, G. 2020. Algorithmic bias and the Value Sensitive Design approach. *Internet Policy Review*, 9(4): 1–16.

Takemoto, K. 2024. The moral machine experiment on large language models. *Royal Society Open Science*, 11(2): 231393.

Talat, Z.; Blix, H.; Valvoda, J.; Ganesh, M. I.; Cotterell, R.; and Williams, A. 2021. A Word on Machine Ethics: A Response to Jiang et al. (2021). arXiv:2111.04158.

Talat, Z.; Blix, H.; Valvoda, J.; Ganesh, M. I.; Cotterell, R.; and Williams, A. 2022-07. On the Machine Learning of Ethical Judgments from Natural Language. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 769 – 779. Stroudsburg, PA: Association for Computational Linguistics. ISBN 978-1-955917-71-1. Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2022); Conference Location: Seattle, WA, USA; Conference Date: July 10-15, 2022.

Team, M. N. 2023a. Introducing MPT-30B: Raising the bar for open-source foundation models. Accessed: 2024-05-11.

Team, M. N. 2023b. Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs. Accessed: 2024-03-22.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Vida, K.; Damken, F.; and Lauscher, A. 2024. Decoding Multilingual Moral Preferences: Unveiling LLM's Biases Through the Moral Machine Experiment. arXiv:2407.15184.

Vida, K.; Simon, J.; and Lauscher, A. 2023. Values, Ethics, Morals? On the Use of Moral Concepts in NLP Research. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5534–5554. Singapore: Association for Computational Linguistics.

Zhao, J.; Khashabi, D.; Khot, T.; Sabharwal, A.; and Chang, K.-W. 2021. Ethical-Advice Taker: Do Language Models Understand Natural Language Interventions? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4158–4164.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; Du, Y.; Yang, C.; Chen, Y.; Chen, Z.; Jiang, J.; Ren, R.; Li, Y.; Tang, X.; Liu, Z.; Liu, P.; Nie, J.-Y.; and Wen, J.-R. 2023. A Survey of Large Language Models. arXiv:2303.18223.