Pluralistic AI Alignment: A Cross-Cultural Pilot Survey

Khashayar Alavi[1 [0009−0005−7896−9772]], Lucie Flek[1,2 [0000−0002−5995−8454]], and Florian Mai[1,2 [0000−0003−1370−9740]]

[1] University of Bonn, Bonn, Germany
[2] Lamarr Institute for Machine Learning and AI, Germany
{s76kalav, fmai, lflek}@uni-bonn.de *

Abstract. Large Language Models are often aligned to primarily Western values. To better understand the need for pluralistic alignment methods, this paper presents a pilot survey that investigates how end users from diverse cultural contexts perceive the representation of their values in AIs, their demand for models better aligned to their own values, and what tradeoffs they would accept for better alignment. Our pilot study observes patterns of cross-cultural variation and interest in culturally aware assistants, higher marginalization fears in some groups, and a wide willingness to trade slight accuracy losses for better alignment. Our findings provide a foundation for a more comprehensive global survey.

Keywords: AI Alignment · Value Pluralism · Large Language Models · Cross-Cultural Survey.

1 Introduction

Large Language Models (LLMs) predominantly reflect Western, Educated, Industrialized, Rich, and Democratic (W.E.I.R.D.) values [14, 28], This creates a digital environment in which diverse ethical perspectives are flattened and users from non-dominant cultures may find their moral outlooks underrepresented. This has led to calls for *pluralistic alignment*, which aims to represent the world's value diversity in a fair and balanced manner [27].

But it remains an open question how the affected people themselves perceive this imbalance: whether users from different cultural backgrounds experience a lack of representation, whether they prefer models that are closer to their own ethical perspectives, and how they evaluate tradeoffs between cultural fit, privacy, and performance.

We address this question with a small cross-cultural pilot survey involving participants who speak Arabic, Mandarin, Farsi, and English. The instrument measures perceived representation in current AI, demand for cultural profiles, fear of future value marginalization, willingness to trade small accuracy losses for cultural fit, stereotyping concerns, and willingness to share cultural/demographic information. Our results inform a larger, more representative global study.

* Corresponding author: Khashayar Alavi (s76kalav@uni-bonn.de)

2 Background

Values and Value Alignment Values are the foundational principles and standards that guide human judgment, behavior, and attitudes [24]. As this theory emphasizes, they represent what societies hold dear and serve as criteria to distinguish between right and wrong, important and unimportant. As Johnson et al. [14] discuss, language is not merely a tool for communication but a medium through which values are co-created, embedded, and transmitted. The way words are associated, the metaphors used, and the topics prioritized in a text all carry a deep undercurrent of socio-cultural values. Since LLMs are trained on large parts of the global web, they learn not just grammar and facts, but also a multitude of these deeply embedded value systems. Which of the value systems the conversational AI adheres to (*value alignment*) is determined in the post-training process [4].

Empirical studies have found that leading AI models are predominantly W.E.I.R.D. For example, Tao et al.[28] found that LLMs reproduce cultural patterns closest to English-speaking and Protestant European societies while diverging from African-Islamic and Confucian context. Johnson et al.[14] stress-tested GPT-3 with value-rich texts from diverse cultures and show that, under value conflict, outputs drift toward the dominant U.S. opinion. Meijer, Mohammadi, and Bagheri [20] compared the LLM response patterns with the World Values Survey and Pew data, finding homogenization toward a liberal mainstream that overstates consensus on controversial topics.

Value Pluralistic Alignment The concept of Moral Value Pluralism holds that there are many conflicting, yet equally valid, sets of values in the world that cannot be reduced to a single universal standard [14]. Disagreement can occur at the individual level [3], but different value systems also often correspond to cultural tendencies, which is the focus of this paper. Hofstede's framework [13] reduces complex cultural tendencies to a set of six quantifiable dimensions, enabling the mapping of user preferences and model behavior onto well-defined axes, facilitating a precise and measurable quantification of alignment across multiple dimensions: Masoud et al. [19] introduce a Cultural Alignment Test and find that models such as GPT-4 adapt on some dimensions but perform poorly for American and Arab profiles, while a recent study [2] shows strong model dependence.

As AI systems become more capable and embedded in society, questions of cultural alignment gain increasing significance. Recent studies show that generative models can reproduce and amplify existing social biases [5, 22], highlighting the broader social impact of misaligned value representations.

These observations have led to calls for *pluralistic alignment methods* that enable AI systems to represent human (value) diversity [27].

Perspectivism and Pluralism in NLP In NLP, value and cultural diversity manifests as subjectivity in annotation and evaluation. Recent work challenges

the assumption of a single objective ground truth, showing that annotator disagreement often reflects meaningful differences in interpretation, norms, or social context, particularly for value-laden and culturally sensitive content [3, 10, 9, 21]. More broadly, the Holistic Evaluation of Language Models (HELM) framework shows that standard evaluation protocols for LLMs are often incomplete, relying on narrow task and metric selections that can obscure biases and behavioral differences across contexts [16]. Perspectivist approaches therefore preserve and model multiple viewpoints rather than collapsing them via majority voting, as conventional aggregation risks privileging dominant perspectives while marginalizing minority ones [3, 10].

Aligning Superhuman AI *Scalable oversight* refers to methods that provide human annotators with AI assistance to ensure effective oversight even as AI systems approach and surpass human-level capabilities [23, 7]. However, there are concerns that such methods may cause a "value lock-in" if assisting AI does not respect the pluralistic nature of human values [25, 18]. This failure may lead to further marginalization of minority ethical perspectives and an erosion of value diversity.

The Pluralistic Alignment Tax Aligning an AI to human values typically incurs an *alignment tax*: As models are aligned during supervised fine-tuning or RLHF, their performance on benchmarks tends to deteriorate [11, 17]. Additionally, there is a *pluralistic* alignment tax, which is induced when aligning AIs to non-dominant cultural groups. Singh et al. [26] show that models perform substantially better on high-resource languages. Moreover, model performance is more fragile for culturally sensitive questions than for culturally agnostic ones. This discrepancy is arguably the result of a data imbalance in both pre-training and post-training data [6].

The existence of the pluralistic alignment tax raises the question whether end-users are willing to accept a performance drop in exchange for a model that is better aligned to their values.

User Attitude Towards Pluralistic Alignment While the need for pluralistic alignment has recently garnered support in the academic community, existing research has focused primarily on technical alignment strategies. However, there is so far only little research regarding the attitude of the users of LLMs towards pluralistic alignment. Early comparative studies have shown that alignment expectations vary across cultures. For instance, Ge et al. [12] found that participants from China preferred relational guidance and social harmony in AI behavior, whereas participants from the United States emphasized personal autonomy and control. Similarly, Fan et al. [8] describe user-driven correction of AI responses perceived as ethically or culturally inappropriate, reflecting a growing awareness of value bias in everyday interaction with AI systems. Through a nation-wide survey in Indonesia, Kautsar et al. [15] find that language technologies tailored to local languages are in great demand. These findings indicate that

cultural context strongly shapes how people interpret and evaluate alignment, underscoring the need for direct empirical investigation across a wider range of linguistic and ethical groups.

However, to the best of our knowledge, there is currently no survey that systematically investigates how people from different cultural backgrounds perceive alignment in AI systems, how they react to models trained under different value assumptions, or how much importance they put on culture-specific alignment vs performance. Understanding these variations is essential for designing alignment strategies that truly reflect the priorities of diverse user communities. Within this broader context, it is equally important to examine how minority or non-dominant groups respond, as their concerns and preferences may differ significantly from those of majority populations. Without such empirical insight, alignment efforts risk overlooking key cultural differences and applying one-size-fits-all solutions that fail to meet the needs of all users.

3 Method

While prior research has shown that large language models reflect dominant cultural value systems, little is known about how users from different cultural backgrounds perceive this misalignment or what kinds of corrective strategies they prefer. To address this gap, we conduct a cross-cultural pilot survey designed to test two central hypotheses and quantify the demand for culturally aligned AI. This pilot provides an initial, user-centered perspective across multiple languages and cultural contexts, offering empirical insights that will inform a larger, more representative global survey and guide the development of culturally pluralistic AI systems.

3.1 Hypotheses

Our investigation is guided by the following hypotheses:

- **H1: Perceived Value Underrepresentation.** Cultural groups perceive a significant deficit in the representation of their ethical standards in mainstream LLMs and will express a clear preference for AI systems that are explicitly aligned with their own values.
- **H2: Apprehension of Value Erosion.** The fear of value erosion and the preference for aligned AI will be more pronounced in minority or non-dominant cultural groups, particularly when considering the future integration of superhuman AI systems. This fear is amplified by findings that misalignment is often most severe for underrepresented personas [1].

3.2 Survey Methodology

This study employs a small-scale pilot survey, implemented in Google Forms, to obtain an initial cross-cultural snapshot of attitudes toward cultural alignment in AI. The instrument comprises six short questions.

To enable participation from culturally and linguistically distinct groups, the survey was administered in four languages: English, Mandarin, Arabic, and Farsi. The English-language version targets participants from W.E.I.R.D. societies, including German citizens and other Western nationals. The Arabic version covers a large linguistic group representing a major cultural sphere in the Middle East and North Africa. The Mandarin version targets a large linguistic and cultural group within East Asia. The Farsi version represents a smaller linguistic group, culturally distinct from the others.

Recruitment is stratified by language and cultural group: the English, Arabic, and Farsi versions primarily target international students at the University of Bonn (with the English group consisting mostly of German participants), while the Mandarin version is distributed via Chinese social media platforms to reach additional participants. The six questions assess participants' perceptions, preferences, and attitudes regarding cultural alignment in AI systems:

1. **Perception of Current AI:** How well participants feel current AI tools reflect their cultural and ethical values.
2. **Preference for Alignment:** Whether participants prefer an AI assistant explicitly tuned to their culture's values.
3. **Fear of Superhuman AI:** The extent to which participants are concerned that future AI could (further) marginalize their culture's values.
4. **Tradeoff Decisions:** Preferences when choosing between a high-accuracy, generic-value AI and a slightly less accurate but culturally aligned AI.
5. **Cultural Profiling Concerns:** Agreement or disagreement with the idea that cultural profiles risk reinforcing stereotypes.
6. **Willingness to Share Information:** Openness to sharing cultural/demographic data to improve alignment.

This setup allows direct cross-cultural comparison between large and small linguistic groups, while capturing both quantitative preference scores and binary choices.

4 Results

The distribution of the participants is as follows: Farsi (11), Arabic (7), English/Western (22), and Mandarin Chinese (16). The sample sizes differ between groups and are not proportional to their respective population sizes, reflecting practical constraints in participant recruitment rather than demographic representation. The results are summarized in Table 1.

Q1 - Perception of Current AI As shown in column Q1 of Table 1, respondents vary in how well they feel current AI reflects their cultural values. W.E.I.R.D. participants report the highest perceived representation (high mean), while Arabic speakers report the lowest representation (low mean), with Mandarin and Farsi groups falling in between. This pattern suggests differences in perceived cultural representation across groups in our pilot sample.

Table 1. Results of the cross-cultural pilot survey covering Questions Q1–Q6. Each column corresponds to one survey question, and each value represents the mean score for the respective participant group (language). For Q1, Q2, Q3, Q5, and Q6, the values are mean scores on a 10-point Likert scale (1 = *I disagree*, 10 = *I agree*), reflecting each group's average level of agreement with the given statement. Column Q4 shows the percentage of participants in each group who preferred Model B, the version of the AI assistant with slightly lower accuracy but better cultural fit.

Group	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**
Arabic	1.43	7.71	6.71	57.1	6.00	6.71
Chinese	3.12	8.88	3.69	68.8	6.12	6.00
Farsi	2.27	1.27	7.00	72.7	5.45	4.18
English	4.86	6.14	2.14	54.5	2.86	1.51

Q2 - Desire for Customization Column Q2 of Table 1 shows that most groups express strong interest in selecting a cultural or ethical profile for their AI assistant, with Mandarin and Arabic participants scoring highest. Higher means indicate greater preference for a customised AI that reflects users' own cultural values. The notably low score among Farsi participants might reflect misunderstanding or or a belief that members of smaller cultural groups have less need for customisation, however this requires verification in larger studies.

Q3 - Fear of Marginalization Concerns that future superhuman AI could marginalize local values are considerably higher in column Q3 of Table 1 for Arabic and Farsi groups, indicating they are more afraid of value marginalization than W.E.I.R.D. participants. W.E.I.R.D. respondents report the lowest levels of concern, suggesting they perceive less risk to their values in the long term, while Chinese participants show a similar tendency toward lower concern, aligning more closely with the W.E.I.R.D. group than with the Arabic or Farsi groups.

Q4 - Cultural Fit Column Q4 of Table 1 shows the percentage of participants who preferred Model B, the version with slightly lower accuracy but better cultural fit. Overall, Model B was selected by 35 of 56 respondents (62.5%). In all groups, the majority of participants favored this culturally aligned option, indicating a general willingness to trade accuracy for alignment. Farsi and Chinese respondents showed the strongest preference, while Arabic and W.E.I.R.D. groups expressed slightly lower support. In this sample, cultural resonance appeared to outweigh marginal performance differences for many participants, particularly among those from non-W.E.I.R.D. cultural contexts.

Q5 - Stereotyping Concerns Column Q5 of Table 1 shows that perceptions of cultural profiles vary across groups. Arabic, Chinese, and Farsi respondents tend to agree more strongly that cultural profiling may reinforce stereotypes, while W.E.I.R.D. participants show lower concern, indicating greater trust in the neutrality of such systems. This difference suggests that sensitivity to stereotyping

risks is higher in non-W.E.I.R.D. contexts, where cultural representation has historically been limited.

Q6 - Willingness to Share Data Attitudes toward sharing cultural or demographic information also vary notably, as shown in column Q6 of Table 1. Arabic and Chinese participants show moderate willingness to share such data, suggesting a relatively higher level of trust in the potential benefits of data-driven personalization and a belief that cultural transparency can help improve AI alignment. In contrast, Farsi respondents appear more hesitant, reflecting a cautious stance toward data disclosure that may be influenced by broader concerns about privacy or potential misuse of personal information. Participants from W.E.I.R.D. societies show the lowest willingness to share such data, indicating stronger expectations for privacy and skepticism toward the idea that AI systems should require personal or cultural identifiers to perform well.

Responses in this pilot suggest attitudes toward data sharing may be tied to broader trust dynamics between users and AI systems. In groups that already perceive themselves as culturally underrepresented, moderate willingness to share information may reflect a pragmatic attempt to increase visibility and fairness within AI development. Meanwhile, the reluctance among W.E.I.R.D. participants could indicate that they already feel sufficiently represented and thus see less need to provide additional personal data. Taken together, these findings emphasize that cultural alignment mechanisms must carefully balance personalization with privacy, ensuring that any data collection for value alignment is voluntary, transparent, and supported by clear user control.

5 Discussion and Conclusion

Although the sample of the study is too limited to allow solid generalizable conclusions, our study remains indicative as a pilot to inform a future study. It offers cross-cultural evidence on key aspects of alignment and reveals tradeoffs often overlooked in theoretical work.

H1 - Underrepresentation and demand. In this pilot sample, Q1 shows that W.E.I.R.D. participants perceive the strongest cultural fit, while other language groups feel less represented, supporting prior findings that LLMs primarily reflect Western value systems. Moreover, Q2 shows strong interest in culturally tuned AI assistants, suggesting considerable demand for pluralistic alignment methods across all groups, even at the cost of a considerable *pluralistic alignment tax* (Q4).

H2 - Minority apprehension. Q3 supports H2, showing that Arabic and Farsi participants express greater concern about future superhuman AI marginalising their values than others. This aligns with the view that non-dominant cultures anticipate more risk from one-size-fits-all alignment. Farsi participants also show

the highest willingness to accept lower model performance in exchange for better cultural fit, reinforcing the idea that minority groups value alignment more strongly than raw accuracy.

Tradeoffs and design implications. The demand for value-aware language models also brings tensions between personalization and privacy. In our sample, non-W.E.I.R.D. groups show both higher concern about stereotyping and greater willingness to share data to improve alignment, W.E.I.R.D. participants display the opposite pattern. Overall, Q6 highlights that alignment mechanisms should rely on *granular, revocable, and privacy-preserving* controls, rather than fixed identity profiles, to balance trust and inclusivity.

Beyond methodological refinements, the pilot results point to key next steps. Future research should determine the extent of performance loss users are willing to accept for better alignment, identify which alignment dimensions matter most to them, and assess how trust and usage context influence their willingness to trade accuracy for cultural fit.

Limitations and next steps. This pilot study should be interpreted with caution. It did not include attention checks or verification of participant identity, and using Google Forms limited control over data quality and uniqueness. The sample was small, uneven across language groups, and not randomly selected. Participants were stratified by survey language (Arabic, Mandarin, Farsi, and English) as a proxy for cultural background. While this enabled practical cross-cultural comparison, language, culture, and ethnicity are distinct constructs, and this approximation may not fully capture the diversity within each group. In our full survey, these limitations will be addressed through balanced recruitment across regions, improved participant verification, built-in attention checks, and more rigorous demographic profiling and translation procedures to ensure consistent interpretation across groups. Beyond methodological refinements, the pilot results point to key next steps. Future research should determine the extent of performance loss users are willing to accept for better alignment, identify which alignment dimensions matter most, and assess how trust and usage context influence tradeoffs between accuracy and cultural fit.

Conclusion. This pilot study observes, in a limited sample, that participants generally want AI systems to reflect their culture and that minority groups are more concerned about being marginalized in the future. At the same time, many are wary of cultural stereotyping and are inconsistent in their willingness to share personal data. These findings support a pluralistic alignment approach that gives users better controls without locking them into fixed value profiles. Our findings motivate an extended, more representative survey that includes additional cultural and linguistic groups. In this next phase, our aim is to analyze how users of diverse backgrounds respond to alignment tradeoffs and to identify behavioral patterns that can inform more inclusive and culturally sensitive alignment procedures for future AI systems.

References

1. AlKhamissi, B., ElNokrashy, M.N., Alkhamissi, M., Diab, M.T.: Investigating cultural alignment of large language models. In: Ku, L., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024. pp. 12404–12422. Association for Computational Linguistics (2024). https://doi.org/10.18653/V1/2024.ACL-LONG.671, https://doi.org/10.18653/v1/2024.acl-long.671
2. Anonymous: Cultural alignment of language models and the effects of prompt language and cultural prompting. Anonymous ACL submission (2025)
3. Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., Uma, A.: We need to consider disagreement in evaluation. In: Church, K., Liberman, M., Kordoni, V. (eds.) Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future. pp. 15–21. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.bppf-1.3, https://aclanthology.org/2021.bppf-1.3/
4. Bhatia, M., Nayak, S., Kamath, G., Mosbach, M., Shwartz, V., Reddy, S., et al.: Value drifts: Tracing value alignment during llm post-training. arXiv preprint arXiv:2510.26707 (2025)
5. Bloomberg: 2023: Humans are biased. generative AI is even worse. https://www.bloomberg.com/graphics/2023-generative-ai-bias/ (2023), accessed: Nov 14, 2025
6. Dunn, J.: Mapping languages: the corpus of global language use. Lang. Resour. Evaluation **54**(4), 999–1018 (2020). https://doi.org/10.1007/S10579-020-09489-2, https://doi.org/10.1007/s10579-020-09489-2
7. Engels, J., Baek, D.D., Kantamneni, S., Tegmark, M.: Scaling laws for scalable oversight. CoRR **abs/2504.18530** (2025). https://doi.org/10.48550/ARXIV.2504.18530, https://doi.org/10.48550/arXiv.2504.18530
8. Fan, X., Xiao, Q., Zhou, X., Pei, J., Sap, M., Lu, Z., Shen, H.: User-driven value alignment: Understanding users' perceptions and strategies for addressing biased and discriminatory statements in AI companions. In: Yamashita, N., Evers, V., Yatani, K., Ding, S.X., Lee, B., Chetty, M., Dugas, P.O.T. (eds.) Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, YokohamaJapan, 26 April 2025- 1 May 2025. pp. 910:1–910:19. ACM (2025). https://doi.org/10.1145/3706598.3713477, https://doi.org/10.1145/3706598.3713477
9. Fisher, J., Appel, R.E., Park, C.Y., Potter, Y., Jiang, L., Sorensen, T., Feng, S., Tsvetkov, Y., Roberts, M.E., Pan, J., Song, D., Choi, Y.: Political neutrality in AI is impossible- but here is how to approximate it. CoRR **abs/2503.05728** (2025). https://doi.org/10.48550/ARXIV.2503.05728, https://doi.org/10.48550/arXiv.2503.05728
10. Frenda, S., Abercrombie, G., Basile, V., Pedrani, A., Panizzon, R., Cignarella, A.T., Marco, C., Bernardi, D.: Perspectivist approaches to natural language processing: a survey. Lang. Resour. Evaluation **59**(2), 1719–1746 (2025). https://doi.org/10.1007/S10579-024-09766-4, https://doi.org/10.1007/s10579-024-09766-4
11. Fu, T., Cai, D., Liu, L., Shi, S., Yan, R.: Disperse-then-merge: Pushing the limits of instruction tuning via alignment tax reduction. In:

Ku, L., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. pp. 2967–2985. Association for Computational Linguistics (2024). https://doi.org/10.18653/V1/2024.FINDINGS-ACL.175, https://doi.org/10.18653/v1/2024.findings-acl.175
12. Ge, X., Xu, C., Misaki, D., Markus, H.R., Tsai, J.L.: How culture shapes what people want from AI. In: Mueller, F.F., Kyburz, P., Williamson, J.R., Sas, C., Wilson, M.L., Dugas, P.O.T., Shklovski, I. (eds.) Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024. pp. 95:1–95:15. ACM (2024). https://doi.org/10.1145/3613904.3642660, https://doi.org/10.1145/3613904.3642660
13. Hofstede, G.: Dimensionalizing cultures: The Hofstede model in context (2011)
14. Johnson, R.L., Pistilli, G., Menédez-González, N., Duran, L.D.D., Panai, E., Kalpokiene, J., Bertulfo, D.J.: The ghost in the machine has an american accent: value conflict in GPT-3. CoRR **abs/2203.07785** (2022). https://doi.org/10.48550/ARXIV.2203.07785, https://doi.org/10.48550/arXiv.2203.07785
15. Kautsar, M.D.A., Susanto, L., Wijaya, D., Koto, F.: What do indonesians really need from language technology? A nationwide survey. CoRR **abs/2506.07506** (2025). https://doi.org/10.48550/ARXIV.2506.07506, https://doi.org/10.48550/arXiv.2506.07506
16. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C.D., Ré, C., Acosta-Navas, D., Hudson, D.A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L.J., Zheng, L., Yüksekgönül, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N.S., Khattab, O., Henderson, P., Huang, Q., Chi, R., Xie, S.M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., Koreeda, Y.: Holistic evaluation of language models. Trans. Mach. Learn. Res. **2023** (2023), https://openreview.net/forum?id=iO4LZibEqW
17. Lin, Y., Lin, H., Xiong, W., Diao, S., Liu, J., Zhang, J., Pan, R., Wang, H., Hu, W., Zhang, H., Dong, H., Pi, R., Zhao, H., Jiang, N., Ji, H., Yao, Y., Zhang, T.: Mitigating the alignment tax of RLHF. In: Al-Onaizan, Y., Bansal, M., Chen, Y. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024. pp. 580–606. Association for Computational Linguistics (2024). https://doi.org/10.18653/V1/2024.EMNLP-MAIN.35, https://doi.org/10.18653/v1/2024.emnlp-main.35
18. Mai, F., Kaczér, D., Corrêa, N.K., Flek, L.: Superalignment with dynamic human values. In: ICLR 2025 Workshop on Bidirectional Human-AI Alignment (2025), https://openreview.net/forum?id=WvB9hKKjSc
19. Masoud, R.I., Liu, Z., Ferianc, M., Treleaven, P.C., Rodrigues, M.: Cultural alignment in large language models: An explanator analysis based on hofstede's cultural dimensions. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S. (eds.) Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025. pp. 8474–8503. Association for Computational Linguistics (2025), https://aclanthology.org/2025.coling-main.567/
20. Meijer, M., Mohammadi, H., Bagheri, A.: Llms as mirrors of societal moral standards: reflection of cultural divergence and agreement across ethical topics.

CoRR **abs/2412.00962** (2024). https://doi.org/10.48550/ARXIV.2412.00962, https://doi.org/10.48550/arXiv.2412.00962
21. Muscato, B., Passaro, L.C., Gezici, G., Giannotti, F.: Perspectives in play: A multi-perspective approach for more inclusive NLP systems. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025. pp. 9827–9835. ijcai.org (2025). https://doi.org/10.24963/IJCAI.2025/1092, https://doi.org/10.24963/ijcai.2025/1092
22. Park, H., Ahn, D., Hosanagar, K., Lee, J.: Human-ai interaction in human resource management: Understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens. In: Kitamura, Y., Quigley, A., Isbister, K., Igarashi, T., Bjørn, P., Drucker, S.M. (eds.) CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021. pp. 154:1–154:15. ACM (2021). https://doi.org/10.1145/3411764.3445304, https://doi.org/10.1145/3411764.3445304
23. Raji, I.D., Dobbe, R.: Concrete problems in AI safety, revisited. CoRR **abs/2401.10899** (2024). https://doi.org/10.48550/ARXIV.2401.10899, https://doi.org/10.48550/arXiv.2401.10899
24. Schwartz, S.H.: An overview of the Schwartz theory of basic values. Online Readings in Psychology and Culture **2**(1) (2012). https://doi.org/10.9707/2307-0919.1116, https://doi.org/10.9707/2307-0919.1116
25. Shen, H., Knearem, T., Ghosh, R., Alkiek, K., Krishna, K., Liu, Y., Ma, Z., Petridis, S., Peng, Y., Qiwei, L., Rakshit, S., Si, C., Xie, Y., Bigham, J.P., Bentley, F., Chai, J., Lipton, Z.C., Mei, Q., Mihalcea, R., Terry, M., Yang, D., Morris, M.R., Resnick, P., Jurgens, D.: Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. CoRR **abs/2406.09264** (2024). https://doi.org/10.48550/ARXIV.2406.09264, https://doi.org/10.48550/arXiv.2406.09264
26. Singh, S., Romanou, A., Fourrier, C., Adelani, D.I., Ngui, J.G., Vila-Suero, D., Limkonchotiwat, P., Marchisio, K., Leong, W.Q., Susanto, Y., Ng, R., Longpre, S., Ruder, S., Ko, W., Bosselut, A., Oh, A., Martins, A.F.T., Choshen, L., Ippolito, D., Ferrante, E., Fadaee, M., Ermis, B., Hooker, S.: Global MMLU: understanding and addressing cultural and linguistic biases in multilingual evaluation. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025. pp. 18761–18799. Association for Computational Linguistics (2025), https://aclanthology.org/2025.acl-long.919/
27. Sorensen, T., Moore, J., Fisher, J., Gordon, M.L., Mireshghallah, N., Rytting, C.M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., Choi, Y.: Position: A roadmap to pluralistic alignment. In: Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. OpenReview.net (2024), https://openreview.net/forum?id=gQpBnRHwxM
28. Tao, Y., Viberg, O., Baker, R.S., Kizilcec, R.F.: Cultural bias and cultural alignment of large language models. PNAS Nexus **3**(9), pgae346 (09 2024). https://doi.org/10.1093/pnasnexus/pgae346, https://doi.org/10.1093/pnasnexus/pgae346