Survey: Multilingual Moral Reasoning in NLP

Anonymous

Abstract

This survey examines past and ongoing work on multilingual moral reasoning in NLP, which encompasses multilingual dataset resources, different types of alignment, and biases such as cross-lingual and cross-cultural alignment. We synthesize evidence from different research showing that model outputs vary substantially across languages and frequently diverge from human moral preferences. We compare multilingual moral reasoning research with the broader monolingual literature and find that multilingual work is primarily concerned with whether LLMs behave consistently across languages and cultures, while monolingual work might more often include benchmarking ethical reasoning abilities and examining LLM alignment with specific moral frameworks. We also identify limitations in the current NLP multilingual moral reasoning literature, including a lack of consensus surrounding whether inconsistencies across languages are desired to reflect a variety of cultural views or pose safety concerns, narrow language coverage, and reliance on translated English resources. Finally, we outline future directions for multilingual moral reasoning research, including AI safety research, resolving issues surrounding inconsistency, and broader evaluation of low-resource languages.

1 Introduction

As large language models (LLMs) become increasingly integrated into everyday life, they are inevitably being used in contexts that involve moral judgment and ethical decision-making (Reinig et al., 2024). Humans constantly engage in moral reasoning across languages and cultures, whether in everyday interpersonal situations or in high-stakes domains such as healthcare, education, law, or military decision-making (Maity and Saikia, 2025; Shi et al., 2026; Lai et al., 2024; Rivera et al.; Antonios Chochtoula, 2024; Liu et al., 2025). The growing use of LLMs in these settings raises important questions about whether their moral judgments and reasoning processes align with those of humans, and whether that alignment remains stable across languages and cultural contexts (Li et al.).

Due to concerns that LLMs produce different outputs depending on language, especially in moral contexts, some NLP papers sought to better understand how LLMs make moral decisions or morally reason. Existing work examines whether LLMs exhibit consistent moral preferences across languages or whether they can reflect culturally specific moral values, and whether multilingual outputs align with human moral judgments (Agarwal et al., 2024; Jin et al., 2025; Papadopoulou et al., 2024). Much of the earlier work on moral reasoning in NLP focused primarily on English and monolingual settings (Reinig et al., 2024). However, moral norms and reasoning patterns vary across cultures and communities, making multilingual evaluation of moral reasoning necessary.

This survey synthesizes current work done on multilingual moral reasoning in NLP. We examine how multilingual moral reasoning research differs from monolingual moral reasoning work, especially in its focus on cross-lingual consistency and cultural bias. Several surveys have previously examined morality and ethics in NLP, including those that focused on Moral Foundations Theory (MFT), broader modeling of morality in text, and the terminological use of ethics and its philosophical grounding (Zangari et al., 2025; Reinig et al., 2024; Vida et al., 2023). However, existing surveys do not specifically focus on multilingual moral reasoning, an important area given the aforementioned aspects of AI safety and potentially unequal human impacts across cultures from differing LLM suggestions. This survey addresses this gap by focusing specifically on multilingual work and the challenges introduced when moral reasoning is evaluated across languages and cultures.

The key contributions of this paper are: (1) a synthesis of current multilingual moral reasoning research in NLP, including major tasks, methodologies, findings, and limitations; (2) a comparison between multilingual and monolingual moral reasoning literature, highlighting differences in research focus and evaluation paradigms; and (3) a discussion of open problems and future directions, including multilingual AI safety concerns, low-resource language evaluation, and culturally grounded moral frameworks.

2 Multilingual Moral Reasoning Tasks

Multilingual moral reasoning papers can be broadly divided into two categories: resource papers and alignment papers. Resource papers propose a new dataset or benchmark, while alignment papers can focus on cross-lingual alignment (whether an LLM produces different outputs given equivalent content prompted in different languages), and cross-cultural alignment (whether LLMs can take into account cultural contexts and correctly align with a given culture) (Jin et al., 2025; Papadopoulou et al., 2024; Agarwal et al., 2024).

Some papers fit into more than one category—a benchmark paper may also examine misalignment between LLMs and human annotations—but each paper is grouped under its primary contribution. Table 1 shows the frequency breakdown of the multilingual moral reasoning NLP papers included in this survey. ArXiv papers are included.

Paper Type	Count
Resource	6
Alignment	15

Table 1: Multilingual moral reasoning papers by type

Predominantly, multilingual moral reasoning papers have focused on alignment questions, reflecting that the primary concern so far has been whether LLM outputs shift across languages and cultural contexts rather than on building resources or establishing standardized benchmarks. The relative scarcity of resource and benchmark papers points to a gap in the field and indicates an important area for future development.

2.1 Resource Papers

Resource papers remain relatively limited within multilingual moral reasoning research. This mirrors broader trends in monolingual moral reasoning NLP, where benchmark and dataset construction remain a comparatively small portion of the literature (Reinig et al., 2024). However, there are around 56 papers in the broader field of NLP for moral reasoning, compared to only 6 identified as related to multilingual moral reasoning. This suggests that there could definitely be more room to propose resources to the multilingual field, especially given the importance of benchmarks in evaluation (Reinig et al., 2024).

The main multilingual resource and benchmark papers identified in this survey are UniMoral, Multilingual Sensitive Questions and Answers Dataset (MSQAD), MFTCXplain (Moral Foundations with EXplanations), Multilingual Trolley Problems (MultiTP), SeaEval, and the Multilingual Moral Reasoning Benchmark (MMRB). Across these resource papers, recurring patterns include the use of social media, news, and crowd-sourced annotations as primary data sources. Although multilingual benchmarks have expanded beyond English, most still focus on a relatively small set of high-resource languages, and those that cover a broader range have relied on the translation quality of machine translation tools. There is an inherent conflict and trade-off between language coverage and data quality. If the dataset focuses on a smaller range of languages, it is easier to manually verify quality; if you include a broader range, you expand coverage but risk data quality.

In short, introducing resources and benchmarks that cover a broader range of tasks and languages remains an important research direction, but there is also a fundamental tension in ensuring data quality and language coverage.

2.2 Alignment

Alignment is the largest category of multilingual moral reasoning research, with the majority of papers focusing on some form of alignment or bias.

Many papers define cross-lingual alignment as whether LLMs produce different outputs or exhibit different moral preferences across languages, and cross-cultural alignment as whether LLMs can accurately reflect a given culture's values (Liu et al., 2025; Tao et al., 2024; Zhou et al., 2025; **?**; Xu et al., 2024). Some papers explore cross-cultural alignment by prompting LLMs in English and asking them to predict responses from someone in a particular country (Mohammadi). However, some papers use cross-lingual and cross-cultural align-

ment interchangeably—particularly when the language under study is closely associated with a single culture, such as Chinese (Cheng and Hale, 2025)—and others define cross-lingual stability as the extent to which LLMs agree with human judgments across languages (Li et al.). These definitional differences may complicate comparisons and make definite separations difficult, as there are many overlaps.

Some studies also frame their research around bias, such as cultural or ethical bias, but they also closely relate to cross-lingual or cross-cultural alignment—whether LLM outputs match human moral preferences or values across languages, or whether LLMs differ in their rates of generating unsafe responses across languages (Aksoy, 2025; Vida et al., 2024; Yu et al.).

2.2.1 Cross-lingual Alignment

Cross-lingual alignment papers generally ask whether a model produces the same or similar moral outputs when prompted with equivalent content in different languages. The consistent finding across this body of work is that model outputs do indeed differ based on language.

Agarwal et al. (2024) probe GPT-4, ChatGPT, and Llama2-70B-Chat with ethical dilemmas from three branches of normative ethics—deontology, virtue ethics, and consequentialism—across English, Spanish, Russian, Chinese, Hindi, and Swahili. They find that GPT-4 is the most consistent across languages, while ChatGPT and Llama2-70B-Chat show significant moral value bias when prompted in languages other than English, and that the nature of this bias varies substantially across languages even for GPT-4. Similarly, Khandelwal et al. uses the Defining Issues Test across five languages and finds that moral reasoning ability is substantially lower in Hindi and Swahili than in Spanish, Russian, Chinese, and English.

Xu et al. (2024) takes a broader approach, examining seven types of human values—commonsense morality, deontology, utilitarianism, fairness, truthfulness, toxicity, and harmfulness—across 16 languages. They find cross-lingual inconsistency and evidence of unidirectional transfer from high- to low-resource languages in value concepts.

Li et al. identifies related patterns using real-world moral dilemmas from Reddit: translating posts into another language alters LLMs' verdicts, suggesting that moral judgments lack cross-lingual stability. The study additionally finds that LLMs tend to be more lenient than human annotators overall in judging whether the person in a scenario is in the wrong.

Vida et al. (2024) conducts a related study using 6,500 Moral Machine scenarios across ten languages and finds that all LLMs exhibit different moral biases to some degree, and that these biases differ not just from human preferences but also across languages within the same model.

Haemmerl et al. (2023) asks directly whether pretrained multilingual models capture moral norms from English and impose them on other languages, and whether they exhibit random or potentially harmful beliefs in certain languages. They found that the models "do encode differing moral biases, but these do not necessarily correspond to cultural differences or commonalities in human opinions" (Haemmerl et al., 2023).

Findings Summary The most common finding across the surveyed papers is that LLMs' moral outputs vary across languages even when the prompts are semantically equivalent (Agarwal et al., 2024; Khandelwal et al.; Farid et al., 2025; Xu et al., 2024; Li et al.). These inconsistencies appear in several forms: changes in final moral judgments, shifts in moral values, inconsistency in reasoning or justifications, or differences in LLM refusal behavior (Vida et al., 2024; Farid et al., 2025; Li et al.).

Even where variation might be expected to reflect genuine cultural differences or the values and beliefs of the language group, some papers suggest a different explanation: LLM outputs are generally not reflective of the cultural values associated with the language in which the model is prompted (Tao et al., 2024; Mohammadi; Papadopoulou et al., 2024).

2.2.2 Cross-cultural Alignment

While cross-lingual alignment papers ask whether LLM outputs differ by language, cross-cultural alignment papers ask the more demanding question of whether LLMs can accurately represent the moral values of specific cultures.

Mohammadi find that current LLMs tend to reflect a liberal worldview aligned with Western, Educated, Industrialized, Rich, and Democratic (W.E.I.R.D.) societies, and that multilingual models are not demonstrably better than monolingual ones at capturing cross-cultural moral diversity. This is a notable finding that complicates any assumption that adding more languages to a model's training data will automatically improve cultural

representativeness. Papadopoulou et al. (2024) and Aksoy (2025) both find that biases exist in monolingual and multilingual models alike, and that models generally fall short of accurately capturing the moral intricacies of diverse cultures.

Jinnai (2024) examines cross-cultural alignment more directly by studying whether aligning a Japanese language model with predominantly English resources transfers commonsense morality appropriately. They find that some aspects of commonsense morality do transfer, but others do not, and that cross-cultural transfer is more difficult than cross-lingual transfer.

Another study, Tao et al. (2024), examines whether prompting in a language elicits language-specific cultural values, "such as asking a question in Korean to elicit Korean cultural values in the LLM's response." However, they found that "evidence from 14 countries and languages indicates that this approach is not effective at producing responses aligned with evidence from nationally representative values surveys." (Tao et al., 2024)

Findings Summary To summarize, numerous papers find that LLMs are currently unable to fully account for nuanced cultural contexts that influence moral judgment and reflect the full variation of cross-cultural morals (Mohammadi; Papadopoulou et al., 2024). LLMs tend to incorrectly categorize controversial topics as among the most agreed upon, such as homosexuality or euthanasia, and produce more neutral or positive judgments to avoid controversial outputs (Mohammadi; Papadopoulou et al., 2024). However, Aksoy (2025) suggests that models do not necessarily impose English moral norms and instead "adapt their moral reasoning to reflect language-specific nuances."

Through these papers, LLMs are found to be inconsistently aligned with human moral reasoning. Some work suggests that LLMs perform well on tasks designed to assess moral alignment (Dillion et al., 2025), but many works find the opposite, as mentioned above. Trager et al. find limited capacity for LLMs to internalize and reflect human moral reasoning, and Vida et al. (2024) find that almost all models diverge substantially from human values in trolley problem scenarios, such as preferring to run over humans than pets.

Papers differ in how they interpret these inconsistencies and what they recommend in response. Some argue for building generic ethical reasoning capabilities into LLMs so that they can handle value pluralism at a global scale rather than encoding the norms of any particular culture (Rao et al., 2023). Others emphasize cultural alignment, arguing that AI systems should better reflect the moral standards of the communities they serve (Papadopoulou et al., 2024). These two positions are in some tension: cultural alignment requires knowing what another culture's moral values are and encoding them correctly, which is difficult without speakers of that language and culture being involved in the annotation step of the data.

2.3 Low-Resource Languages

A subset of multilingual moral reasoning papers specifically examines low-resource languages (LRLs) and LLMs' moral reasoning capabilities in those settings. Zhou et al. (2025) introduces the MMRB covering five typologically diverse languages, including Vietnamese, and finds that moral reasoning performance degrades with increasing context complexity, particularly for low-resource languages. Notably, they also find that low-resource languages have a stronger impact on multilingual moral reasoning than high-resource ones—a counterintuitive finding that highlights the importance of data quality in underrepresented languages, and suggests that even small amounts of LRL data can shift model behavior in ways that may not be anticipated by evaluations conducted only in high-resource settings. A related finding from fine-tuning experiments is that training on LRL data is the most impactful in terms of both alignment and susceptibility to poisoning, likely because fewer pre-training examples make new LRL data proportionally more influential (Zhou et al., 2025).

The results across papers are somewhat conflicting. In the study by Basu et al. (2025), they found that when prompted in a low-resource language, models may flag harmful content more readily and refuse to answer, while in high-resource languages, "models read the context of the question are successfully tricked into answering" but also that "models were significantly more likely to produce harmful content in lower-resource languages." Xu et al. (2024) found the opposite pattern, where translating harmful prompts from English into other languages can partially circumvent safety defenses, and that larger models may be more susceptible to this because of their "enhanced instruction-following capabilities." Another work, (Farid et al., 2025), also finds that LLMs diverge across languages, favoring English while struggling

with low-resource languages, and that models not only perform less accurately but also fail to reliably follow instructions in lower-resource languages, with refusals and formatting errors becoming more common. These conflicting patterns suggest that the relationship between resource level and model safety is not straightforward and warrants dedicated study.

3 Methodologies of Multilingual Moral Reasoning Papers

This section examines common methods and datasets used in multilingual papers and discusses their associated advantages and limitations.

3.1 Frameworks

Frameworks used in moral reasoning NLP research fall into two broad categories: psychological or theoretical frameworks, such as Moral Foundations Theory, which define the moral dimensions and values used to analyze or compare human and LLM moral reasoning; and evaluation or assessment frameworks, such as the Defining Issues Test and the Moral Machine Experiment, which translate moral reasoning into measurable tasks or dilemmas.

3.1.1 Moral Foundations Theory

A pattern across multilingual moral reasoning papers is the use of Moral Foundations Theory (MFT), a framework that raises concerns about its translation across cultures (Aksoy, 2025). Originally developed in moral psychology, MFT identifies six core moral foundations—care, fairness, loyalty, authority, sanctity, and liberty—that vary in prominence across cultures (Zangari et al., 2025). Some papers use MFT-based questionnaires, sometimes translating them into other languages, as the basis for evaluating LLM moral outputs (Aksoy, 2025; Trager et al.; Haemmerl et al., 2023). This provides a structured and theoretically grounded framework, but MFT is itself Western-derived and may not adequately capture the moral priorities of non-Western cultures, a limitation that is particularly significant for cross-cultural evaluation. However, due to these concerns, MFQ-2 has been developed to address the cross-cultural applicability of MFQ-1. A psychological journal found that "MFQ-2 fares well in terms of reliability and validity across cultural context," providing a strong basis for multilingual research on moral reasoning to use this, which was indeed used in the papers, but other frameworks can also potentially be explored to examine LLM alignment and values across non-Western-originated frameworks (Atari et al., 2023).

3.1.2 Defining Issues Test

The Defining Issues Test (DIT) has also been adapted in multilingual moral reasoning research to measure the moral reasoning stages reflected in LLM outputs (Khandelwal et al.). Unlike MFT, which categorizes different types of moral values, the DIT focuses on how individuals reason through moral dilemmas and the extent to which their judgments reflect principled versus authority- or rule-based reasoning. Some papers use DIT-style dilemmas to compare whether LLMs exhibit patterns of moral reasoning similar to humans across languages and cultures. Like MFT, however, the DIT was developed within a Western psychological tradition, raising questions about whether its assumptions about moral development generalize cross-culturally.

3.1.3 Moral Machine Experiment

The Moral Machine experiment, which has collected over 40 million human moral judgments from participants across 200+ countries on trolley problems involving autonomous vehicles, has also been used as a data source for cross-lingual and cross-cultural evaluation (Jin et al., 2025). In contrast to theory-driven frameworks such as MFT or DIT, the Moral Machine provides large-scale behavioral preference data. Several multilingual studies use these scenarios to compare whether LLM moral decisions align with human preferences across different languages and cultural contexts. While the dataset offers broad geographic coverage and enables large-scale comparison, the dilemmas themselves are still framed around specific assumptions about utilitarian tradeoffs and may not fully capture the complexity of moral reasoning across cultures.

3.2 Data Sources

Papers in this area draw on a variety of sources to gather human annotations for comparing against LLM outputs. Social media posts, such as tweets and Reddit posts, are among the most common data sources, used both to assess moral alignment and to gather annotations on scenario-based judgments (Kumar and Jurgens; Li et al.; Alhassan et al., 2022). Large-scale cross-national surveys are also widely used, including the World Values Survey

(WVS), which covers participants from 55 countries, and the Pew Global Attitudes survey, which covers 39 countries (Mohammadi; Papadopoulou et al., 2024).

3.3 Languages Examined

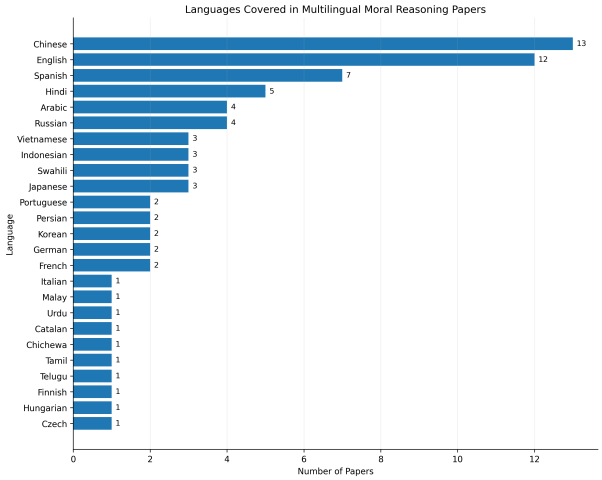

Figure 1: Languages covered by the papers included in the survey

The languages examined across the surveyed papers skew heavily toward high-resource languages, as seen in Figure 1. Chinese and English are the most frequent languages, with Spanish appearing as third. The frequency tapers off for other languages, and very few papers examine lower-resource languages such as Tamil, Urdu, and Catalan (Xu et al., 2024). This narrow coverage means the moral reasoning capabilities of LLMs in many languages are largely left unexamined.

3.4 Models Examined

The models evaluated across the surveyed papers span a range of closed and open-source systems. GPT-4 and GPT-3.5 (ChatGPT) are the most commonly evaluated models and frequently appear together for comparison. Llama models (Llama 2, Llama 3) and multilingual models such as XLM-R and BLOOM also appear across multiple papers. Several papers find GPT-4 to be the most consistent and least biased moral reasoner across languages (Agarwal et al., 2024; Rao et al., 2023; Bulla et al., 2025), though it still exhibits language-dependent variation. BLOOM is notable for showing stronger cross-cultural alignment than other models in some evaluations (Papadopoulou et al., 2024; Aksoy, 2025), possibly due to its more balanced multilingual training data. However, others suggest that multilingual models are not better than monolingual models at truthfully capturing cultural diversity in moral judgments (Mohammadi). An interesting pattern is that many multilingual papers rely on BLOOM, which may be less prevalent in other areas of NLP.

3.5 Common Limitations

The surveyed papers share several limitations. Inevitably, as in other fields of NLP, papers cannot cover most of the languages of the world. Papers tend to typically cover only a small set of high-resource languages, leaving LRLs largely absent and many cultures' moral and value systems unexamined.

A second limitation is the reliance on Western evaluation frameworks, particularly MFT. Using MFT as the primary lens for evaluating cross-cultural moral reasoning risks assuming that its moral foundations are universal, when in fact they may not adequately capture the moral priorities of non-Western cultures. However, some psychological studies do suggest opposite evidence that they have viable cross-culturally.

Third, many papers rely on machine translation, such as for translating questionnaires from English into other languages. Cheng and Hale (2025) find in a Chinese case study that machine translation and local lexicon approaches are insufficient for complex moral assessments and frequently result in loss of cultural information, raising questions about whether multilingual evaluations based on translated English text genuinely measure cross-cultural moral reasoning.

A fourth limitation is the conflation of cross-lingual and cross-cultural alignment. Some papers do not separate the two, reasoning that language and culture are historically intertwined (Yu et al.). However, this reasoning has limits: Latin American Spanish and Spanish from Spain share a language while differing significantly in cultural context. Finally, the papers do not necessarily account for individual variance in moral values (e.g., a Chinese speaker may hold the same moral values as an English speaker), making it difficult to balance reflecting cultural values without promoting stereotyping and overgeneralization.

4 Comparison with Monolingual Papers

Monolingual moral reasoning papers differ from multilingual ones in both focus and methodology.

While multilingual papers are primarily concerned with cross-lingual consistency and cultural bias, monolingual papers tend to focus on benchmarking moral reasoning abilities and evaluating LLMs against different ethical frameworks. Compared to multilingual papers, there are considerably more benchmark datasets in the monolingual moral reasoning domain (Reinig et al., 2024).

One distinction is that some monolingual papers engage more deeply with the theoretical foundations of ethics. For instance, Zhou et al. (2024) distinguishes between normative ethics—justice, deontology, and utilitarianism as formulated by moral philosophers—and descriptive ethics as developed by moral psychologists, and attempts to examine both. Many monolingual and multilingual papers examine only one or the other, grounding their evaluation in either MFT or a single ethical framework such as deontology or consequentialism (Agarwal et al., 2024; Khandelwal et al.; Vida et al., 2024). The same paper also proposes a top-down approach in which established moral theories guide LLM reasoning, as opposed to the bottom-up approach of training on crowd-sourced annotations that many resource papers favor.

Both bodies of work rely on MFT-derived evaluation and find evidence that LLMs may rely on superficial patterns rather than genuine moral reasoning (Trager et al.; Chakraborty et al.).

On the benchmarking side, Yu et al. (2024) proposes CMoralEval, a large benchmark for Chinese LLMs covering both explicit moral scenarios and moral dilemma scenarios derived from a Chinese TV program and Chinese newspapers. The benchmark is grounded in a morality taxonomy rooted in traditional Chinese culture as well as contemporary societal norms, making it one of the few moral evaluation resources developed from a non-Western perspective, though it remains focused on a single language.

On the framework side, Rao et al. (2023) develops a framework integrating moral dilemmas with normative ethical principles, finding that while GPT-4 performs nearly perfectly within this framework, models still exhibit bias toward Western and English-speaking moral values. This finding directly connects to the cross-cultural concerns raised in the multilingual literature.

Samway et al. investigate whether LLMs are consequentialist or deontological moral reasoners and find that this varies by context, with some patterns observed based on model size and whether models reason or decide first. This aligns with Barabadi et al. (2026)'s finding that LLMs alternate between moral frameworks depending on scenario-specific features.

Monolingual moral reasoning papers also include work in non-English contexts. Cheng and Hale (2025) evaluate approaches to measuring moral foundations in Chinese text and find that LLMs outperform translation-based and lexicon-based approaches but still require human-in-the-loop validation to capture cultural nuances, a finding relevant to multilingual evaluation. Vargas et al. (2026) introduces a Brazilian Portuguese hate-speech dataset with moral rationales as direct supervision for attention alignment, bridging hate detection and moral reasoning in a non-English context.

Some monolingual papers raise the question of whether LLMs should be aligned with humans in the first place, given that humans are not necessarily consistent moral reasoners and that there is substantial individual variance (Dillion et al., 2025). Some papers also examine value pluralism—the degree to which LLMs can reflect differing views and values within a group—which is not explicitly used as a term in the multilingual moral reasoning literature but can also be an important and relevant topic in the field and remains an open question (Russo et al.; Naseem et al., 2026). A related question, raised in both bodies of work, is whether moral reasoning in LLMs resembles logical reasoning or diverges from it, and whether the two can meaningfully be compared at all (Chakraborty et al.; Kumar and Jurgens).

To summarize, the monolingual moral reasoning field encompasses a broader range of theoretical frameworks and evaluation types, while multilingual papers primarily focus on cross-lingual or cross-cultural alignment and consistency. These two subfields share some gaps but differ in other aspects, and progress in each could inform the other as they are inextricably linked.

5 Future Directions

There is considerable room for this field to grow, and several specific gaps stand out.

Low-Resource Languages The most immediate gap is low-resource language coverage. Even papers that describe themselves as multilingual typically cover a small set of high-resource languages (Xu et al., 2024). If LLMs already perform poorly

at capturing cultural contexts in relatively high-resource languages, it is reasonable to expect worse performance for languages that receive even less attention in training and evaluation. The finding that LRL data has an outsized impact on alignment and model susceptibility to poisoning compared to high-resource data makes this an especially important area for future work (Zhou et al., 2025).

Non-Western Frameworks A related gap is the dominance of Western moral frameworks, particularly MFT (Zangari et al., 2025). Most evaluation paradigms are derived from Western psychology and ethics. Although MFQ may be adapted to be more encompassing of other cultures and be more applicable across different cultural value systems, developing and applying non-Western frameworks could give a more complete picture of what it would mean for LLMs to reason morally across cultures, since what counts as aligned depends substantially on the framework used to evaluate it.

Disentangling Cross-Lingual and Cross-Cultural Alignment There may also be issues with conflating cross-lingual and cross-cultural alignment. It is unclear whether prompting a model in Spanish captures Spain's culture or that of Latin America. Disentangling the two would require study designs that can separate them, such as holding language constant while varying cultural framing, or vice versa.

Consistency or Inconsistency? It is also worth questioning whether cross-lingual consistency is always the appropriate target. Some research suggests that humans themselves make different moral judgments when presented with the same scenario in different languages (Farid et al., 2025), which complicates the assumption that inconsistency in LLMs is purely an unwanted feature. However, cross-lingual inconsistency also introduces a concrete multilingual jailbreak vulnerability: if a model's safety guardrails are primarily trained in English, prompting it in a low-resource language may partially circumvent those guardrails (Xu et al., 2024). Larger models may be particularly susceptible to this, given their enhanced instruction-following capabilities, making this an area where moral reasoning research connects directly to AI safety concerns.

Value Pluralism There are open questions about how to obtain representative samples of human moral perspectives across cultures without over-generalizing, and how to have LLMs account for individual variation within cultures. It remains unclear whether we want LLMs to reflect the full range of human values within a culture—which could include harmful and discriminatory views—or whether we prefer models that lean toward neutral or uncontroversial judgments.

6 Conclusion

As LLMs become more integral to everyday life and are increasingly used for moral decision-making, unexamined inconsistencies across languages can lead to different user populations receiving different and potentially harmful guidance. This area of concern is a significant motivation for examining the current literature on the moral capabilities of LLMs and whether they reflect human moral values across different languages and cultures.

Multilingual moral reasoning is currently a relatively small subfield within NLP, but the body of work is growing, with many papers appearing in 2025 and on ArXiv compared to previous years (Farid et al., 2025; Zhou et al., 2025; Papadopoulou et al., 2024). The field's central finding—that LLM moral outputs vary across languages and are not yet capable of fully reflecting cultural nuance and values—is established across multiple papers and models (Agarwal et al., 2024; Vida et al., 2024; Farid et al., 2025; Mohammadi; Papadopoulou et al., 2024; Trager et al.). What remains less clear is how to address the inconsistency, and whether we even want LLMs to reflect human moral judgments at all is an open question, given that human judgments can themselves be biased and inconsistent. If alignment is desired, there is no consensus on whether that should mean alignment with local cultural norms, cross-lingual consistency, or some more universal ethical standard (Rao et al., 2023; Papadopoulou et al., 2024; Farid et al., 2025).

All in all, the field of multilingual moral reasoning is underdeveloped relative to monolingual moral reasoning work. There are only a few multilingual benchmark datasets, and most papers focus on alignment issues, whereas monolingual moral reasoning papers encompass a broader range of tasks and evaluation methods. Future work can increase resources in the multilingual domain, resolve tensions between competing alignment goals, expand coverage to low-resource languages, and more directly investigate potential AI safety concerns surrounding this area.

7 Limitations

Papers were collected by searching for keywords including "moral reasoning," "moral," "ethics," and "value" in ACL Anthology and Google Scholar, then manually filtered for relevance to multilingual moral reasoning. Some relevant papers may have been missed, but this survey attempts to provide a broad overview of the multilingual moral reasoning field.

References

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical Reasoning and Moral Value Alignment of LLMs Depend on the Language we Prompt them in.

Meltem Aksoy. 2025. Whose morality do they speak? Unraveling cultural bias in multilingual language models. *Natural Language Processing Journal*, 12:100172.

Areej Alhassan, Jinkai Zhang, and Viktor Schlegel. 2022. 'Am I the Bad One'? Predicting the Moral Judgement of the Crowd Using Pre–trained Language Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 267–276, Marseille, France. European Language Resources Association.

Antonios Chochtoula. 2024. How Large Language Models are Transforming Modern Warfare - Joint Air Power Competence Centre.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T. Stevens, and Morteza Dehghani. 2023. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 125(5):1157–1188.

Elyas Barabadi, Zahra Fotuhabadi, Amanollah Arghavan, and James R Booth. 2026. Comparing AI and human moral reasoning: context-sensitive patterns beyond utilitarian bias. *Frontiers in Artificial Intelligence*, 8:1710410.

Kimaya Basu, Savi Kolari, and Allison Yu. 2025. Measuring Moral LLM Responses in Multilingual Capacities. *arXiv preprint*. ArXiv:2510.08776 [cs].

Luana Bulla, Stefano De Giorgis, Misael Mongiovì, and Aldo Gangemi. 2025. Large Language Models meet moral values: A comprehensive assessment of moral abilities. *Computers in Human Behavior Reports*, 17:100609.

Mohna Chakraborty, Lu Wang, and David Jurgens. Structured Moral Reasoning in Language Models: A Value-Grounded Evaluation Framework.

Calvin Yixiang Cheng and Scott A. Hale. 2025. Beyond English: Evaluating Automated Measurement of Moral Foundations in Non-English Discourse with a Chinese Case Study. *arXiv preprint*. ArXiv:2502.02451 [cs].

Danica Dillion, Debanjan Mondal, Niket Tandon, and Kurt Gray. 2025. AI language model rivals expert ethicist in perceived moral expertise. *Scientific Reports*, 15(1):4084.

Sualeha Farid, Jayden Lin, Zean Chen, Shivani Kumar, and David Jurgens. 2025. One Model, Many Morals: Uncovering Cross-Linguistic Misalignments in Computational Moral Reasoning. *arXiv preprint*. ArXiv:2509.21443 [cs].

Katharina Haemmerl, Bjoern Deiseroth, Patrick Schramowski, Jindřich Libovický, Constantin Rothkopf, Alexander Fraser, and Kristian Kersting. 2023. Speaking Multiple Languages Affects the Moral Bias of Language Models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2137–2156, Toronto, Canada. Association for Computational Linguistics.

Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, and Bernhard Schölkopf. 2025. Language Model Alignment in Multilingual Trolley Problems. *arXiv preprint*. ArXiv:2407.02273 [cs].

Yuu Jinnai. 2024. Does Cross-Cultural Alignment Change the Commonsense Morality of Language Models? In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 48–64, Bangkok, Thailand. Association for Computational Linguistics.

Aditi Khandelwal, Utkarsh Agarwal, Kumar Tanmay, and Monojit Choudhury. Do Moral Judgment and Reasoning Capability of LLMs Change with Language? A Study using the Multilingual Defining Issues Test.

Shivani Kumar and David Jurgens. Are Rules Meant to be Broken? Understanding Multilingual Moral Reasoning as a Computational Pipeline with UNIMORAL.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. 2024. Large language models in law: A survey. *AI Open*, 5:181–196.

Nan Li, Bo Kang, and Tijl De Bie. Human-AI Moral Judgment Congruence on Real-World Scenarios: A Cross-Lingual Analysis.

Haijiang Liu, Jinguang Gu, Xun Wu, Daniel Hershcovich, and Qiaoling Xiao. 2025. Cross-cultural value alignment frameworks for responsible AI governance: Evidence from China-West comparative analysis. *arXiv preprint*. ArXiv:2511.17256 [cs].

Subhankar Maity and Manob Jyoti Saikia. 2025. Large Language Models in Healthcare and Medical Applications: A Review. *Bioengineering*, 12(6):631.

Hadi Mohammadi. Do Large Language Models Understand Morality Across Cultures?

Usman Naseem, Gautam Siddharth Kashyap, Sushant Kumar Ray, Rafiq Ali, Ebad Shabbir, and Abdullah Mohammad. 2026. Do Large Language Models Reflect Demographic Pluralism in Safety? In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2042–2052, Rabat, Morocco. Association for Computational Linguistics.

Evi Papadopoulou, Hadi Mohammadi, and Ayoub Bagheri. 2024. Large Language Models as Mirrors of Societal Moral Standards. *arXiv preprint*. ArXiv:2412.00956 [cs].

Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical Reasoning over Moral Alignment: A Case and Framework for In-Context Ethical Policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, Singapore. Association for Computational Linguistics.

Ines Reinig, Maria Becker, Ines Rehbein, and Simone Ponzetto. 2024. A Survey on Modelling Morality for Text Analysis. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4136–4155, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. Escalation Risks from LLMs in Military and Diplomatic Contexts.

Giuseppe Russo, Debora Nozza, Paul Röttger, and Dirk Hovy. The Pluralistic Moral Gap: Understanding Moral Judgment and Value Differences between Humans and Large Language Models.

Keenan Samway, Max Kleiman-Weiner, David Guzman Piedrahita, Rada Mihalcea, Bernhard Schölkopf, and Zhijing Jin. Are Language Models Consequentialist or Deontological Moral Reasoners?

Yuhong Shi, Kun Yu, Yifei Dong, and Fang Chen. 2026. Large language models in education: a systematic review of empirical applications, benefits, and challenges. *Computers and Education: Artificial Intelligence*, 10:100529.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Jackson Trager, Matteo Guida, Yalda Daryani, and Farzan Karimi-Malekabadi. MFTCXplain: A Multilingual Benchmark Dataset for Evaluating the Moral Reasoning of LLMs through Multi-hop Hate Speech Explanation.

Francielle Vargas, Jackson Trager, Diego Alves, Surendrabikram Thapa, Matteo Guida, Berk Atil, Daryna Dementieva, Andrew Smart, and Ameeta Agrawal. 2026. Self-Explaining Hate Speech Detection with Moral Rationales. *arXiv preprint*. ArXiv:2601.03481 [cs].

Karina Vida, Fabian Damken, and Anne Lauscher. 2024. Decoding Multilingual Moral Preferences: Unveiling LLM's Biases through the Moral Machine Experiment. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:1490–1501.

Karina Vida, Judith Simon, and Anne Lauscher. 2023. Values, Ethics, Morals? On the Use of Moral Concepts in NLP Research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5534–5554, Singapore. Association for Computational Linguistics.

Shaoyang Xu, Weilong Dong, Zishan Guo, Xinwei Wu, and Deyi Xiong. 2024. Exploring Multilingual Concepts of Human Values in Large Language Models: Is Value Alignment Consistent, Transferable and Controllable across Languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1771–1793, Miami, Florida, USA. Association for Computational Linguistics.

Linhao Yu, Yongqi Leng, Yufei Huang, Shang Wu, Haixin Liu, Xinmeng Ji, Jiahui Zhao, Jinwang Song, Tingting Cui, Xiaoqing Cheng, Liutao Liutao, and Deyi Xiong. 2024. CMoralEval: A Moral Evaluation Benchmark for Chinese Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11817–11837, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Seunguk Yu, Juhwan Choi, and Youngbin Kim. Delving into Multilingual Ethical Bias: The MSQAD with Statistical Hypothesis Tests for Large Language Models.

Lorenzo Zangari, Candida Maria Greco, Davide Picca, and Andrea Tagarelli. 2025. A survey on moral foundation theory and pre-trained language models: current advances and challenges. *AI & SOCIETY*, 40(6):4973–4998.

Huichi Zhou, Zehao Xu, Munan Zhao, Kaihong Li, Yiqiang Li, and Hongtao Wang. 2025. Moral Reasoning Across Languages: The Critical Role of Low-Resource Languages in LLMs. *arXiv preprint*. ArXiv:2504.19759 [cs].

Jingyan Zhou, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. 2024. Rethinking Machine Ethics – Can LLMs Perform Moral Reasoning through the Lens of Moral Theories? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2227–2242, Mexico City, Mexico. Association for Computational Linguistics.