Ethical Reasoning and Moral Value Alignment of LLMs Depend on the Language we Prompt them in

Utkarsh Agarwal[*1] **Kumar Tanmay**[*1] **Aditi Khandelwal**[*1] **Monojit Choudhury**[†2]
[*]Microsoft Corporation
[†]MBZUAI
{t-utagarwal, t-ktanmay, t-aditikh}@microsoft.com, monojit.choudhury@mbzuai.ac.ae

Abstract

Ethical reasoning is a crucial skill for Large Language Models (LLMs). However, moral values are not universal, but rather influenced by language and culture. This paper explores how three prominent LLMs – GPT-4, ChatGPT, and Llama2-70B-Chat – perform ethical reasoning in different languages and if their moral judgement depend on the language in which they are prompted. We extend the study of ethical reasoning of LLMs by Rao et al. (2023) to a multilingual setup following their framework of probing LLMs with ethical dilemmas and policies from three branches of normative ethics: deontology, virtue, and consequentialism. We experiment with six languages: English, Spanish, Russian, Chinese, Hindi, and Swahili. We find that GPT-4 is the most consistent and unbiased ethical reasoner across languages, while ChatGPT and Llama2-70B-Chat show significant moral value bias when we move to languages other than English. Interestingly, the nature of this bias significantly vary across languages for all LLMs, including GPT-4.

1. Introduction

Large Language Models (LLMs) like ChatGPT have gained popularity all over the world for their ability to generate fluent and engaging natural language texts (Schulman et al., 2022; OpenAI, 2023). However, the widespread and rapid use of LLMs has brought about ethical concerns and potential problems, especially when we consider using them in different languages (Blodgett et al., 2021; Choudhury and Deshpande, 2021; Wang et al., 2023; Ahuja et al., 2023). As LLMs become more prevalent and find applications in everyday life, they must confront complex moral dilemmas rooted in the existence of multiple conflicting values, commonly dubbed as the problem of *value pluralism* (James, 1891; Dai and Dimond, 1998; Ramesh et al., 2023). Several researchers (see for instance Rao et al. (2023) and Zhou et al. (2023)) have argue that instead of being firmly aligned to a specific set of values, LLMs should be trained to function as generic ethical reasoners, adaptable to different contexts and languages. The final moral judgments, of course, should be made by the stakeholders at different stages of the application life-cycle. LLMs should be able to reason ethically in a generic way, given a situation and a *moral stance*, and should soundly resolve the dilemma when possible, or else ask for more clarity on the stance.

In a recent study, Rao et al. (2023) has demonstrated that LLMs, especially GPT-4, are capable of carrying out sound ethical reasoning. They showed that when the LLMs are presented with a moral dilemma and a moral stance presented at different levels of abstraction and following different formalisms of normative ethics in the prompt, they are often capable of resolving the dilemma in a way that is consistent with the moral stance. They further argue that this is a promising direction towards solving the issues of value pluralism at a global scale, because the different stake-holders in the development and use of an AI system can specify their moral stance which can be meaningfully consumed in the prompt by the LLM to arrive at a sound moral judgment. The alternative approach of aligning LLMs to moral values is bound to fail due to the absence of a universal value hierarchy. However, this study was conducted only for English.

It is a well established fact that the abilities of the LLMs in languages beyond English are often poor and unpredictable (see for example Ahuja et al. (2023), Zhao et al. (2023) and Wang et al. (2023)). Moreover, an intriguing human phenomenon known as the *Foreign Language effect* (Costa et al., 2014a) comes into play when people face moral dilemmas presented in a foreign language (L2). People often make different moral judgments in L2 when compared to when they encounter the same dilemmas in their native language (L1). This suggests that language can significantly shape our emotional and cognitive responses to moral situations, influencing our choices and beliefs.

Given this complex interplay of culture, language and values in making moral judgments, one might ask: Do LLMs also exhibit a Foreign Language effect, changing their behavior when confronted

[1]Equal contribution.
[2]Work done while at Microsoft.

LREC-COLING 2024, pages 6330–6340

with moral dilemmas in different languages? In this work, we extend the study by Rao et al. (2023) to five languages other than English, namely Spanish, Russian, Chinese, Hindi, Arabic, and Swahili. We probe three popular LLMs: GPT-4 (OpenAI, 2023), ChatGPT (September 2023) (Schulman et al., 2022) and Llama2-70B-Chat (Touvron et al., 2023) systematically to assess their ethical reasoning abilities in different languages by prompting them to resolve an ethical dilemma reflecting conflicts between interpersonal, professional, social and cultural values, and a set of *ethical policies* (i.e., moral stances) that can help arrive at a clear resolution of the dilemma. These policies are drawn from three branches of normative ethics: **deontology** (Alexander and Moore, 2021), **virtue** (Hursthouse and Pettigrove, 2022) and **consequentialism** (Sinnott-Armstrong, 2022) and have three different levels of abstractions (similar to Rao et al. (2023)).

Our study clearly demonstrates that indeed the LLMs exhibit different biases while resolving the moral dilemmas in different languages. This bias is minimal for English, in the sense that the resolution of the dilemma depends on the ethical policy rather than the model's own judgment, and it is maximum for low-resource languages such as Hindi and Swahili. This observation can also be interpreted as a reduced ethical reasoning capability of the LLMs in languages beyond English. However, as we shall see in this paper, the ethical reasoning ability (or conversely the bias) is dilemma-specific, which makes us conclude that the LLMs have strong value alignment biases in languages beyond English. Other salient findings are: (1) Across all languages, GPT-4 has the highest ethical reasoning ability, while Llama2-70B-Chat has the poorest; (2) across all models, the reasoning is poorest for Hindi and Swahili, while best for unsurprisingly, English, and also Russian; and (3) across all models, English and Spanish, and Hindi and Chinese have similar bias patterns.

2. Background

The topic of right and wrong has been a subject of ongoing discussion among philosophers, psychologists, and other social scientists. Each field brings its unique perspectives and worries into this debate. In this section, we take these concerns as a guide to offer an overview of this ongoing discussion and how it connects with the field of machine ethics. Our main focus is on how these conversations affect the Natural Language Processing (NLP) community, and we also explore how Large Language Models (LLMs) can advance the frontiers of machine ethics, particularly in multilingual contexts.

2.1. Ethics and Moral Philosophy

Ethics is the branch of philosophy concerned with determining what is morally good or bad and what is considered right or wrong. It encompasses systems or theories of moral values and principles (Kant, 1977, 1996). Within the realm of ethics, normative ethics plays a central role by seeking to establish standards of conduct for human actions, institutions, and ways of life. *Normative ethics* branches into *deontology*, which evaluates the inherent rightness or wrongness of actions based on moral rules or duties (Alexander and Moore, 2021); *virtue ethics*, which focuses on an individual's character and virtues rather than specific rules or consequences (Hursthouse and Pettigrove, 2022); and *consequentialism*, which emphasizes the goodness or value of the outcomes or goals of actions (Sinnott-Armstrong, 2022).

Ethical dilemmas are situations characterized by conflicts between two or more moral values or principles, posing challenges for moral judgment and decision-making (Slote, 1985). The question of whether moral dilemmas can coexist with a consistent system of moral values is a subject of debate. Philosopher Williams argues that ethical consistency doesn't eliminate the possibility of moral dilemmas, as some actions that ought to be done may be incompatible (Williams, 1988). Resolving these dilemmas often requires making new value judgments within the existing ethical framework.

Value pluralism is a key component of ethical dilemmas, suggesting that there are multiple values that can be equally correct and yet in conflict with each other (James, 1891). Different individuals or cultures may prioritize these values differently, resulting in varying resolutions of ethical dilemmas, each of which is ethically sound and consistent. Within the realm of pluralism, there exist various sub-schools of thought, including Rossian Pluralism (Ross and Stratton-Lake, 2002) and Particularism (Hare, 1965), each offering distinct viewpoints. Rossian pluralists advocate for the evaluation of moral principles based on their merits and demerits. Conversely, particularists contend that the assessment of moral pros and cons should be context-dependent. Nevertheless, both schools of thought share a fundamental conviction that there is no one-size-fits-all principle capable of resolving all moral conflicts. They also reject the notion of a rigid hierarchy of moral principles that could facilitate such resolutions. This perspective implies that there is no universally applicable set of moral values or principles that can address all situations and apply uniformly to all individuals.

Inglehart and Welzel (2010) introduced a framework for mapping global cultures which employs a two-dimensional axis system, where the x-axis

represents a spectrum that stretches from survival ethics on the left to self-expression on the right and y-axis covers a range from tradition-based or ethnocentric moral views at the bottom to democratic and rational principles at the top. This visual representation illustrates the tendency of societies to move diagonally from the lower-left corner to the upper-right corner as they progress through industrialization and development.

2.2. Foreign Language Effect on Morality

Recent studies such as (Costa et al., 2014b; Hayakawa et al., 2017; Corey et al., 2017) have found a fascinating link between moral judgment and what's known as the "Foreign-Language Effect." This effect shows that people tend to make more practical choices when they face moral dilemmas in a foreign language (L2) compared to their native language (L1). This shift seems to be because using a foreign language makes people less emotionally connected to the situation, which, in turn, reduces the influence of emotions on their moral decisions. Čavar and Tytus (2018) also highlights that being better at and more comfortable with the foreign language (L2) can decrease this tendency to make practical decisions. This means that the language you use can significantly affect how you make moral choices, impacting many people. Furthermore, bilingual individuals' moral decision-making process is quite complex, involving factors like the type of dilemma, emotional excitement, and the language they're using (Chan et al., 2016).

2.3. Ethics in LLMs

In the field of Ethics in NLP, most approaches assume a deontological perspective, with developers setting moral rules, but these may not readily apply to various contexts or Large Language Models (LLMs) (Talat et al., 2022). Awad et al. (2022) introduce the Computational Reflective Equilibrium (CRE) framework for AI-based ethics, emphasizing moral intuitions and principles. Sambasivan et al. (2021), Bhatt et al. (2022) and Ramesh et al. (2023) have raised questions of value-pluralism in AI and the need for recontextualizing fairness and AI ethics, particularly in global contexts. (Diddee et al., 2022) explore ethical concerns in Language Technologies for social good, emphasizing stakeholder interactions and strategies. Choudhury and Deshpande (2021) advocates the Rawlsian principle over utilitarianism in multilingual LLMs for linguistic fairness.

AI alignment seeks to ensure that AI systems conform to human goals and ethical standards, as highlighted by (Piper, Oct 15, 2020). Various initiatives have put forth ethical frameworks, guidelines, and datasets to train and evaluate Language Models (LLMs) in terms of ethical considerations and societal norms (Hendrycks et al., 2020; Zhou et al., 2023; Jiang et al., 2021; Rao et al., 2023; Tanmay et al., 2023). Moreover, Tanmay et al. (2023) introduces an ethical framework utilizing the Defining Issues Test to assess the ethical reasoning abilities of LLMs. However, it's worth noting that these efforts may be susceptible to biases based on the backgrounds of those providing annotations, as pointed out by Olteanu et al. (2019). Recent research has placed a growing emphasis on in-context learning and supervised tuning to align LLMs with ethical principles, as demonstrated by the studies conducted by Hendrycks et al. (2020); Zhou et al. (2023); Jiang et al. (2021); Rao et al. (2023); Sorensen et al. (2023). These methodologies aim to accommodate a range of ethical perspectives, recognizing the multifaceted nature of ethics. Rao et al. (2023) posits that the generic ethical reasoning abilities can be infused into the LLMs so that they can handle value pluralism at a large scale. However, the authors have considered the ethical policies only in English.

In this study, we will be extending the work of Rao et al. (2023) with different languages to explore how LLMs behave when the multilingual ethical policies are infused into these LLMs . As far as we know, this study is the first of it's kind dealing with the ethics of LLMs in multilingual settings.

2.4. Multilingual Performance of LLMs

Language Models (LLMs) exhibit remarkable multilingual capabilities in natural language processing tasks, although their proficiency varies among languages (Zhao et al., 2023). Their primary training data is in English, but they also incorporate data from various other languages, contributing to their generalisability (Brown et al., 2020; Chowdhery et al., 2022; Zhang et al., 2022; Zeng et al., 2022). Nevertheless, significant challenges arise when LLMs interact with non-English languages, especially in low-resource contexts (Bang et al., 2023; Jiao et al., 2023; Hendy et al., 2023; Zhu et al., 2023). Several studies suggest that enhancing their multilingual performance is possible through in-context learning and the strategic design of prompts (Huang et al., 2023; Nguyen et al., 2023). Experiments conducted by Ahuja et al. (2023) and Wang et al. (2023) have brought to light an interesting aspect. They benchmarked LLMs across a range of Natural Language Processing (NLP) tasks, including Machine Translation, Natural Language Inference, Sentiment Analysis, Text Summarization, Named Entity Recognition, and Natural Language Generation. The results indicate that, while LLMs excel for a few well-resourced languages, they generally under-perform for most

languages. Furthermore, Kovač et al. (2023) have shown that LLMs exhibit context-dependent values and personality traits that can vary across different perspectives. This is in contrast to humans, who typically maintain more consistent values and traits across various contexts.

Importantly, the current body of research has primarily concentrated on the technical capabilities of multilingual LLMs. There has been a relative lack of exploration into their moral reasoning within diverse linguistic and cultural contexts. Recognizing the significant impact of LLMs on real-world applications and domains, there is a growing need to delve into the ethical dimensions surrounding these multilingual language models.

3. Methodology

3.1. Language-based Ethical Framework

We extend the existing framework for defining ethical policies in the LLM prompt, as initially presented by Rao et al. (2023) and reformulate it to be used in multilingual settings. Let us consider an LLM, denoted as $\mathcal{L}$, which accepts a prompt p in language λ and produces a textual output through the function $\mathcal{L}(p)$. We establish the definition of p as an arbitrary composition (for example, concatenation), an ethical policy π, and user input x, in language λ. In original formulation, the task definition is defined as τ. Since we are defining this specifically for ethically reasoning task, variable p can be expressed as a function of the variables π, x, and λ, denoted as

$$p = P(\pi, x, \lambda).$$

Hence we extend the definition of ethical consistency to following:

Definition *Ethical Consistency.* The output produced by the model $\mathcal{L}$ is considered ethically consistent with the policy π if and only if is a valid response or resolution to the input x in the language λ according to the policy π. We represent this as:

$$x \wedge \pi \ \vdash_e \ y$$

where, similar to logical entailment, $\vdash_e$ represents *ethical entailment*.

If a policy is ambiguous for the resolution of x, it can lead to discrepancies in the system (Williams, 1988). In such instances, LLMS should refrain from definitively resolving the dilemma in either direction. In this framework, it is anticipated that the LLM will assert that a definitive resolution, but when it is unattainable, the response should be ϕ for such responses. Therefore, in the case when π is not fully specified, the likelihood function:

$$\mathcal{L}(P(\pi, x, \lambda)) \rightarrow \phi.$$

3.2. Ethical Policies

As explained in (Rao et al., 2023), ethical policies are distinctly characterized as expressions of preference pertaining to either moral values or ethical principles. The absence of a universally agreed-upon set of ethical principles necessitates the flexibility to define policies based on various ethical formalisms or their combinations. In the context of a specific ethical formalism, denoted as F, there exists a set of fundamental moral principles, represented as

$$R^F = r_1^F, r_2^F, \ldots r_{n_F}^F.$$

In our work, we extend these principles to the unique challenges posed by LLMs in multilingual settings.

It is important to underscore the definition of an 'Ethical Policy' as described in this preceding work. An ethical policy, denoted as π, is essentially a partial order applied to a subset of elements within R^F. Specifically, π is represented as $(R_s^F, \leq_s^F)$, where R_s^F denotes a subset of R^F, and $\leq_s^F$ signifies a non-strict partial order relation governing the importance or priority of these ethical principles. This level of policy abstraction is herein referred to as a 'Level 2 policy,' and it serves as an illustrative example of how virtue ethics may manifest, for instance, as 'prioritizing loyalty over objective impartiality.'

Furthermore, the refinement of policies is also elaborated upon in this preceding work, wherein they are classified into 'Level 1' and 'Level 0 policies.' Level 1 policies, such as 'favoring loyalty towards a friend over professional impartiality,' provide specificity by designating the variables to which ethical virtues apply. Meanwhile, Level 0 policies, like 'prioritizing loyalty towards her friend Aisha over objectivity towards scientific norms of publishing,' delve into even finer details by specifying the values to which these virtues are to be applied.

The systematic approach to ethical policies described here underscores their pragmatic applicability, depending on the level of abstraction and specificity desired. It is important to note that these policies, as described in the previous work, are predominantly conveyed in natural language. Nevertheless, it is conceivable that future developments may explore alternative means of policy representation, including symbolic, neural, or hybrid models, drawing from various ethical formalisms. This nuanced understanding of ethical policies in different languages, as derived from the prior research, forms the basis for the discussion in this present paper .

4. Evaluating Ethical Reasoning Across Languages

Here, we describe the design of our experiment to study the change in the ethical reasoning abilities of three popular Large Language Models with well known multilingual capabilities with prompts in different languages. In the experiment, the models were prompted with moral dilemmas (x's) in language L that needed to be resolved for a predetermined ethical policy (π).

We evaluate two OpenAI's models, ChatGPT (GPT-3.5-turbo) (Schulman et al., 2022) and GPT-4 (OpenAI, 2023). ChatGPT is a finetuned version of the GPT-3.5 model, optimized for dialog using RLHF. We use the September 2023 preview of ChatGPT for our experiments. GPT-4 is a larger and more recent model by OpenAI. We also evaluate Meta's publicly available Llama2-70B-Chat-hf model (Touvron et al., 2023). It is a 70B parameter model optimized for dialog use cases.

We set temperature equal to 0 for all the experiments. Others parameters are set as follows: top probability is 0.95 and the presence penalty is 1.

4.1. Dataset

We examine the moral dilemmas and value statements proposed in Rao et al. (2023), each of the four dilemmas contain nine pairs of contrasting policies with at three different levels of abstraction. These policies are related to three branches of normative ethics: *Virtue*, *Deontology*, and *Consequentialism*. The dilemmas highlights the clash between different ethical values. The three dilemmas created by Rao et al. (2023) emphasize conflicts of interpersonal versus professional and community versus personal values. Each of these dilemmas consists of nine policies, denoted as $\pi = (r_i^F \geq r_j^F)$, and their corresponding complementary forms, $\bar{\pi} = (r_j^F \geq r_i^F)$. This gives us a total of eighteen distinct policies for each dilemma.

To evaluate the ethical consistency of the Language Models' outputs, we use the ideal resolutions for each dilemma under each policy, as annotated by Rao et al. (2023). None of these ideal resolutions equate to ϕ.

Since all the proposed dilemmas are in english, we translate these dilemmas and the corresponding policies to six different languages (Spanish, Chinese, Russian, Hindi, Arabic and Swahili) using Google Translation API [3]. We back-translated them into English to check the consistency of the meanings of the values by manual inspection. The prompt instruction is also translated similarly.

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

4.2. Experiments

First, we conduct a baseline experiment in which the models are prompted to respond to a moral dilemma without any given policy, and they provide their moral judgment or resolution from the three options: $y =$ "he/she should", $\neg y =$ "he/she shouldn't" and $\phi =$ "can't decide". This process is repeated for all languages. In order to eliminate the effect of potential positional biases, we utilize six different permutations of these three options and each permutation is run 5 times. This experiment is designed to reveal the models' inherent biases and moral stances, and to study how they vary across different languages. For each language, a total of 120 ($4 \times 6 \times 5$) experiments are conducted, including four dilemmas and six permutations of options. We note the *baseline* resolution of the model for each of the dilemmas per language in Table 1.

In the next part, there is a policy statement given along with the dilemma instructing the model to resolve the dilemma strictly based on the policy. This results in each model being probed a total of 432 times for each language ($18 \times 4 \times 6$). The prompt structure used here is the same as the one proposed with the dilemmas.

4.3. Metrics

The following metrics were used to study the models' behaviors across the dilemmas in different languages.

Accuracy here is defined as the percentage of number of times the model correctly resolves the dilemma given the policy as per the proposed resolution.

Bias *and* ***Confusion*** are two key metrics calculated to assess model behavior. *Bias* is defined as the fraction of times the model sticks to its baseline stance, even when the provided policy dictates otherwise. *Confusion* is the fraction of times the model deviates from it's baseline stance when the policy prompts it to stick with the same stance. We calculate both of these as follows:

$$bias = \frac{\sum_i (1 \mid x_i \neq A,\ y_i = A)}{\sum_i (1 \mid x_i \neq A)}$$

$$confusion = \frac{\sum_i (1 \mid\ x_i = A,\ y_i \neq A)}{\sum_i (1 \mid x_i = A)}$$

Here, x_i represents the ground truth, y_i represents the model prediction, and A represents the model's baseline stance.

A higher *bias* value illustrates a strong alignment of the model to it's preferred resolution which it still tries to reason for despite an opposing policy. A high *confusion* score illustrates the possibility that the model is perhaps not able to understand the

dilemma and it's associated values very well and thus deviates from expected resolution for no clear reason.

5. Results and Observations

Table 1 shows the baseline performance of all three models across various languages. GPT-4 resolves dilemmmas with a remarkably high agreement across all cases, with one notable exception observed in the Heinz dilemma when probed in Hindi, resulting in a 50% agreement rate. An intriguing observation emerges when analyzing the GPT-4 results for the Rajesh dilemma; a distinct pattern of opposite resolutions (highlighted in red) is observed for most languages when compared to the English context (highlighted in green). Furthermore, ChatGPT exhibits conflicting behaviors between English and Hindi for all dilemmas in its resolution process. Llama2-70B-Chat exhibits the least consistent behavior among the three models, primarily due to the lower degree of agreement among all possible permutations of choices. Notably, Llama2-70B-Chat demonstrates a remarkable departure from the behavior of ChatGPT and GPT-4 in most dilemmas and languages. Llama2-70B-Chat tends to opt for affirmative resolutions, such as "should share" and "should steal the drug," more frequently, showcasing a distinct and contrasting behavioral pattern when compared to its counterparts. These findings hint at the potential presence of bias within the models towards specific dilemmas (and consequently, while resolving certain kinds of value conflicts) in specific languages, warranting further investigation.

Figure 2 provides a comprehensive overview of the results obtained from policy-based resolution by the models across various languages, comparing them to the ground-truth resolutions. GPT-4 consistently demonstrates superior ethical reasoning abilities across most languages, with the notable exception of Hindi. In stark contrast, Llama2-70B-Chat exhibits the least ethical reasoning capability across the board.

Table 2 lists the accuracy of each model across the different levels of abstraction of the policies. We can see the trend that models tend to perform slightly better on average on lower abstraction levels.

When considering the different policies, it becomes evident from Figure 2 that Level 2 policies, aligned with the consequentialist framework, are where the models predominantly excel, except for ChatGPT for Russian. For Level 1 deontological policies, Llama2-70B-Chat and ChatGPT perform well. These models perform well for Level 1 policies in virtue ethics except for Russian and Spanish. This highlights the nuanced interaction between policy levels, ethical frameworks, and model performance.

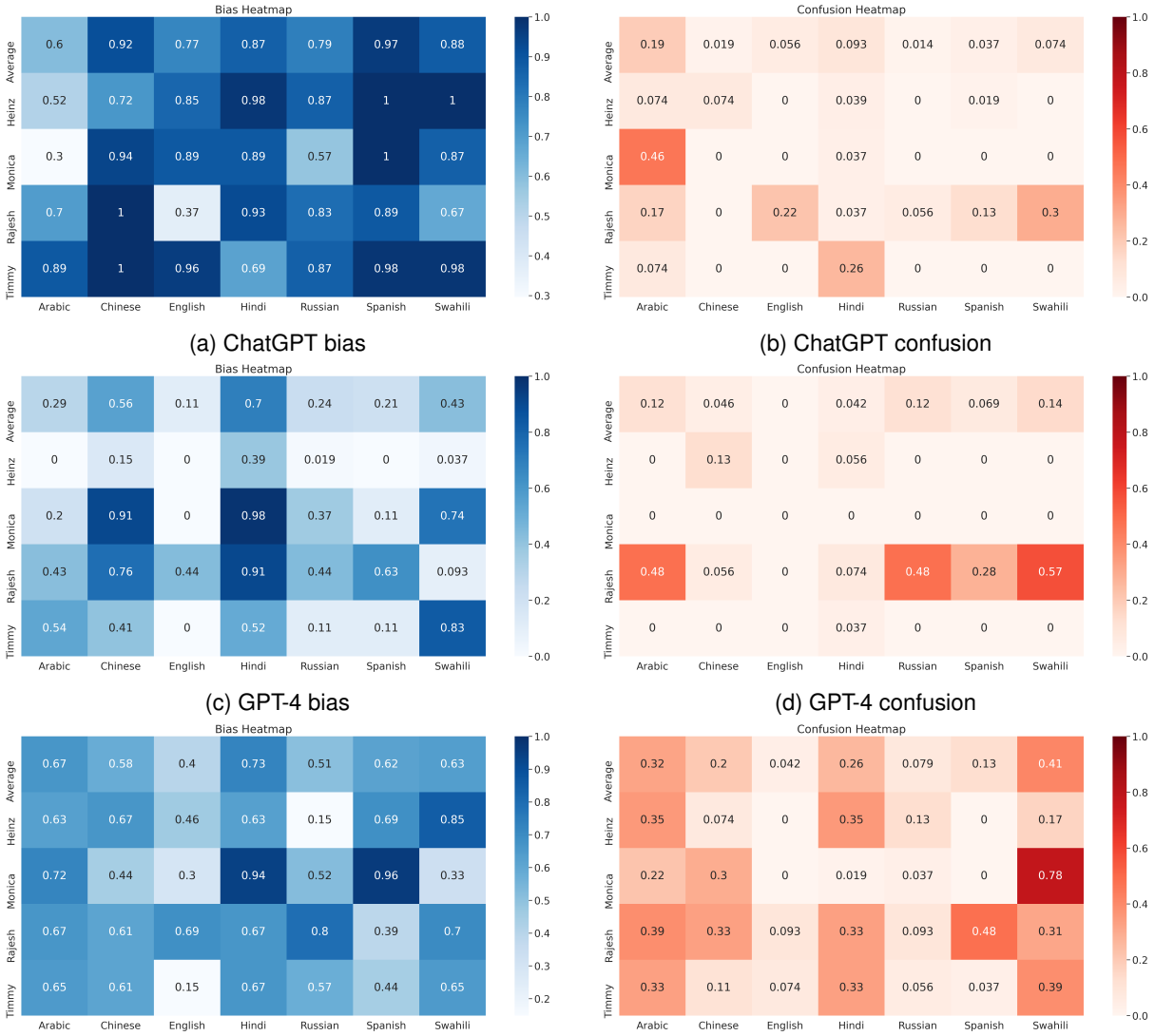
GPT-4 exhibits improved reasoning ability when Level 2 policies are applied in Arabic, Russian, and Spanish, as compared to English within the consequentialist framework. Deontological policies in Level 2 work better for Russian and Spanish than for English. However, for virtue ethics-based policies, GPT-4 shows a distinct advantage for English. Conversely, Llama2-70B-Chat demonstrates notably superior performance with all ethical policies when expressed in English, as compared to all other languages. Overall, a general trend emerges where all models tend to perform less effectively in Hindi and Swahili, while achieving their best results in English and Russian. These observations provide valuable insights into the models' ethical reasoning abilities in various linguistic and ethical contexts.

Figure 1 provides a visual representation of our comparative analysis, illustrating the bias and confusion scores of each model across dilemmas and languages. From the figure, it becomes evident that GPT-4 exhibits the lowest levels of bias among the models, while ChatGPT demonstrates the highest bias levels. Interestingly, all models exhibit similar bias scores in both the English and Spanish, as well as in the case of Hindi and Chinese.

On comparing the confusion scores, Llama2-70B-Chat consistently shows the highest scores among all models, with GPT-4 displaying the lowest confusion scores. Notably, GPT-4's behavior varies across languages in the Rajesh dilemma, highest confusion score being observed for Swahili. In the context of Hindi, all models exhibit significantly divergent behavior when compared to their performance in English.

	English	Arabic	Chinese	Hindi	Russian	Spanish	Swahili
ChatGPT							
Heinz	100%	100%	76.6%	100%	83.3%	66.6%	96.6%
Monica	100%	66.6%	100%	100%	100%	100%	100%
Rajesh	100%	66.6%	100%	100%	100%	66.6%	66.6%
Timmy	100%	83.3%	100%	50%	96.6%	83.3%	83.3%
GPT-4							
Heinz	100%	100%	100%	50%	100%	100%	100%
Monica	100%	100%	100%	100%	100%	100%	100%
Rajesh	100%	100%	100%	66.6%	63.3%	100%	56.6%
Timmy	66.7%	100%	86.6%	100%	100%	100%	100%
Llama2-70B-Chat							
Heinz	100%	66.7%	83.3%	66.6%	66.6%	100%	50%
Monica	100%	66.7%	50%	83.3%	66.7%	100%	50%
Rajesh	83.3%	66.7%	66.7%	66.7%	100%	66.7%	100%
Timmy	66.7%	66.7%	66.7%	66.7%	83.3%	83.3%	57.1%

Table 1: Baseline resolutions percentage of the times the majority resolution was chosen, Green - y majority, red - $\neg y$ majority and yellow - equal



(a) ChatGPT bias

(b) ChatGPT confusion

(c) GPT-4 bias

(d) GPT-4 confusion

(e) Llama2-70B-Chat bias

(f) Llama2-70B-Chat confusion

Figure 1: Bias and Confusion scores for the three models for each language-dilemma pair

6. Discussion and Conclusion

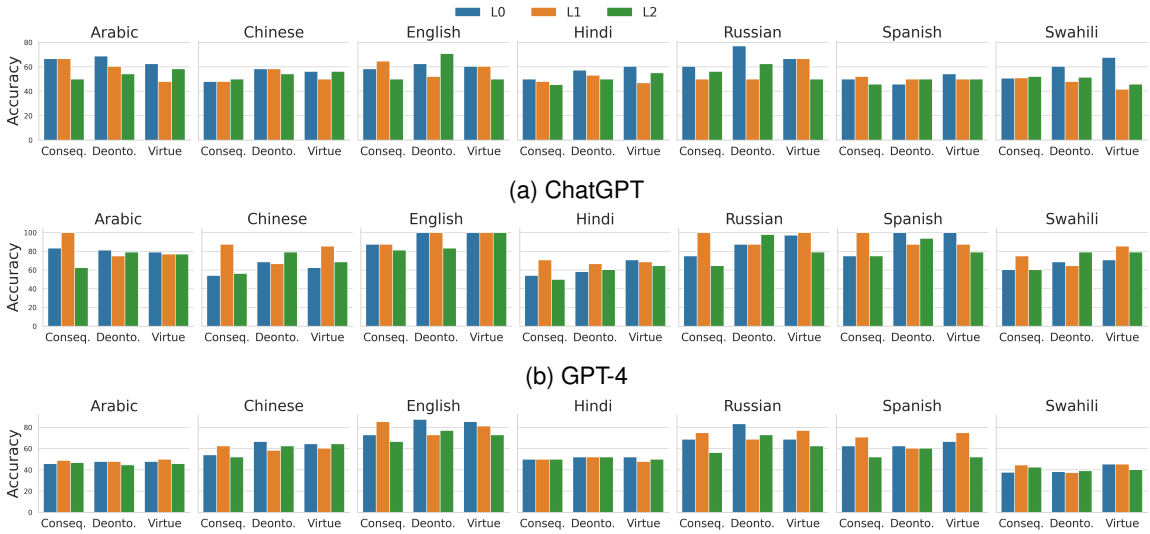
In this paper, we presented a study on the multilingual ethical reasoning capability of three popular LLMs, in the spirit of “ethical policy in prompt" over value alignment as originally suggested by Rao et al. (2023). Our study shows that while for some languages, notably English and Russian, LLMs, especially GPT-4, has superior ethical reasoning abilities, for low-resource languages - Hindi and Swahili, all models fail to perform well. Thus, along the lines of many other studies on multilingual evaluation, our work provides further evidence in support of the performance gap across languages for LLMs, and brings out yet another dimension – that of ethical reasoning – where the gap is prominently evident. Why this gap exists, and how it can be bridged are two important problems we would like to consider for future studies.

Languages and values are strongly intertwined, as is language and culture. The World Value Survey (Inglehart and Welzel, 2010) shows how the values vary by countries, and therefore, in the languages spoken there. While our study brings out distinct biases of the LLMs across languages, it is not clear whether these biases are a reflection of cultural differences across the languages, or simply an artifact of poor performance. Similar bias patterns between English and Spanish, and Chinese and Hindi across models provide a hint that there might be more to this than just performance disparity. This is an interesting question that calls for further investigation.

Recent studies in neuroscience and psychology has shown that humans, most of the time, arrive at a moral judgment akin to aesthetic judgments shaped by their past experiences and cultural biases, rather than by reasoning (Haidt, 2001). This

Model	Level	Arabic	Chinese	English	Hindi	Russian	Spanish	Swahili
ChatGPT	**Level 0**	66.0	54.2	60.4	55.9	68.1	50.0	50.0
	Level 1	58.3	52.1	59.0	49.3	55.6	50.7	50.7
	Level 2	54.2	53.5	56.9	50.2	56.3	48.6	48.6
	Average	59.5	53.3	58.8	51.8	60.0	49.8	49.8
GPT-4	**Level 0**	81.3	61.8	95.8	61.1	85.6	90.9	66.7
	Level 1	84.0	79.9	95.8	68.8	95.5	91.7	75.0
	Level 2	72.9	68.1	88.2	58.3	80.6	82.6	72.9
	Average	79.4	69.9	93.3	62.7	87.2	88.4	71.5
Llama2	**Level 0**	47.2	61.8	81.9	51.4	73.6	63.9	40.5
	Level 1	48.9	60.4	79.9	50.0	73.6	68.8	42.5
	Level 2	45.8	59.7	72.2	50.7	63.9	54.9	40.6
	Average	47.3	60.6	78.0	50.7	70.4	62.5	41.2

Table 2: Accuracy (%) (wrt ground truth) of resolution for policies averaged over types of ethics and abstraction levels.



(a) ChatGPT

(b) GPT-4

(c) Llama2-70B-Chat

Figure 2: Accuracy(%) (wrt ground truth) of resolution for policies of different types and levels of abstraction across different languages

is also known to be the reason behind implementation of unfair policies by governments and organizations, even if the people involved in making these decisions had the right intentions. In this light, use of LLMs for ethical reasoning support across cultures, values and languages can be a very promising use-case with significant large scale positive impact.

Broader Impact Statement

Our framework is subject to some key limitations. Firstly, it relies on the latest models, such as ChatGPT, GPT-4, and Llama2-70B-Chat, for ethical reasoning and the results cannot be generalized to all current models and primarily supports an 'in context' ethical policy approach. However, we anticipate that forthcoming language models will enhance this capability. Another limitation pertains to the construction of dilemmas, moral policies, and ideal resolutions which is designed by Rao et al. (2023) who mention that these may include some bias due to their ethnically homogenous background, potentially limiting the diversity of representation. Our study focuses on a limited set of languages, primarily emphasizing linguistic diversity, which may restrict the generalizability of our findings to languages not included. Additionally, using Google Translator for multilingual dilemma translation introduces the potential for translation errors. Despite these constraints, our research provides valuable insights into the cross-cultural ethical decision-making of Large Language Models (LLMs) across diverse languages, underscoring

the importance of addressing these limitations in future investigations to enhance the strength and robustness of our findings. An ethical concern stemming from our research is the potential misinterpretation that GPT-4's superior ethical reasoning capabilities could imply its readiness for real-life ethical decision-making. This assumption can be perilous, as the model's testing is confined to just seven languages, and caution should be exercised against generalizing its performance to untested languages. It's important to emphasize that our current study doesn't offer a robust foundation for employing LLMs in moral judgment processes, and further research and considerations are warranted.

Acknowledgements

We would like to thank the following people and their corresponding native languages for their help in validation of the translations: Abdelrahman Atef Mohamed Ali Sadallah (Arabic, MBZUAI), Hongyi Zhang (Chinese, Microsoft Corporation), Roman Kazakov (Russian, MBZUAI), Emilio Cueva (Spanish, MBZUAI), Jesus Ortiz Barajas (Spanish MBZUAI), Millicent Ochieng (Swahili, Microsoft Corporation)

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Larry Alexander and Michael Moore. 2021. Deontological Ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Winter 2021 edition. Metaphysics Research Lab, Stanford University.

Edmond Awad, Sydney Levine, Michael Anderson, Susan Leigh Anderson, Vincent Conitzer, M.J. Crockett, Jim A.C. Everett, Theodoros Evgeniou, Alison Gopnik, Julian C. Jamison, Tae Wan Kim, S. Matthew Liao, Michelle N. Meyer, John Mikhail, Kweku Opoku-Agyemang, Jana Schaich Borg, Juliana Schroeder, Walter Sinnott-Armstrong, Marija Slavkovik, and Josh B. Tenenbaum. 2022. Computational ethics. *Trends in Cognitive Sciences*, 26(5):388–405.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Recontextualizing Fairness in NLP: The Case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, Online only. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Franziska Čavar and Agnieszka Ewa Tytus. 2018. Moral judgement and foreign language effect: when the foreign language becomes the second language. *Journal of Multilingual and Multicultural Development*, 39(1):17–28.

Yuen-Lai Chan, Xuan Gu, Jacky Chi-Kit Ng, and Chi-Shing Tse. 2016. Effects of dilemma type, language, and emotion arousal on utilitarian vs deontological choice to moral dilemmas in Chinese-English bilinguals. *Asian Journal of Social Psychology*, 19(1):55–65.

Monojit Choudhury and Amit Deshpande. 2021. How Linguistically Fair Are Multilingual Pre-Trained Language Models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12710–12718.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Joanna D Corey, Sayuri Hayakawa, Alice Foucart, Melina Aparici, Juan Botella, Albert Costa, and Boaz Keysar. 2017. Our moral choices are foreign to us. *Journal of experimental psychology: Learning, Memory, and Cognition*, 43(7):1109.

Albert Costa, Alice Foucart, Sayuri Hayakawa, Melina Aparici, Jose Apesteguia, Joy Heafner, and Boaz Keysar. 2014a. Your morals depend on language. *PLOS ONE*, 9(4):1–7.

Albert Costa, Alice Foucart, Sayuri Hayakawa, Melina Aparici, Jose Apesteguia, Joy Heafner, and Boaz Keysar. 2014b. Your morals depend on language. *PloS one*, 9(4):e94842.

Y T Dai and M F Dimond. 1998. Filial piety. A cross-cultural comparison and its implications for the well-being of older parents. *Journal of Gerontological Nursing*, 24.

Harshita Diddee, Kalika Bali, Monojit Choudhury, and Namrata Mukhija. 2022. The Six Conundrums of Building and Deploying Language Technologies for Social Good. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS)*, COMPASS '22, page 12–19, New York, NY, USA. Association for Computing Machinery.

Jonathan Haidt. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological review*, 108(4):814.

R. M. Hare. 1965. *Freedom and Reason*. Oxford University Press.

Sayuri Hayakawa, David Tannenbaum, Albert Costa, Joanna D Corey, and Boaz Keysar. 2017. Thinking more or feeling less? explaining the foreign-language effect on moral judgment. *Psychological science*, 28(10):1387–1397.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning AI with Shared Human Values. *arXiv preprint arXiv:2008.02275*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting. *arXiv preprint arXiv:2305.07004*.

Rosalind Hursthouse and Glen Pettigrove. 2022. Virtue Ethics. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Winter 2022 edition. Metaphysics Research Lab, Stanford University.

Ronald Inglehart and Chris Welzel. 2010. The WVS cultural map of the world. *World Values Survey*.

William James. 1891. The Moral Philosopher and the Moral Life. *The International Journal of Ethics*, 1(3):330.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021. Can machines learn morality? The Delphi Experiment. *arXiv preprint arXiv:2110.07574*.

Wenxiang Jiao, Wenxuan Wang, JT Huang, Xing Wang, and ZP Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Immanuel Kant. 1977. *Kant: Lectures on Ethics*. Hackett Publishing Company.

Immanuel Kant. 1996. The metaphysics of morals.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2023. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. *arXiv preprint arXiv:2306.11372*.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2:13.

OpenAI. 2023. GPT-4 Technical Report.

Kelsey Piper. Oct 15, 2020. The case for taking AI seriously as a threat to humanity.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in language models beyond English: Gaps and challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2106–2119, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

David Ross and Philip Stratton-Lake. 2002. *The Right and the Good*. Oxford University Press.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in India and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 315–328.

John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, Rapha Gontijo Lopes, Shengjia Zhao, Arun Vijayvergiya, Eric Sigler, Adam Perelman, Chelsea Voss, Mike Heaton, Joel Parish, Dave Cummings, Rajeev Nayak, Valerie Balcom, David Schnurr, Tomer Kaftan, Chris Hallacy, Nicholas Turley, Noah Deutsch, Vik Goel, Jonathan Ward, Aris Konstantinidis, Wojciech Zaremba, Long Ouyang, Leonard Bogdonoff, Joshua Gross, David Medina, Sarah Yoo, Teddy Lee, Ryan Lowe, Dan Mossing, Joost Huizinga, Roger Jiang, Carroll Wainwright, Diogo Almeida, Steph Lin, Marvin Zhang, Kai Xiao, Katarina Slama, Steven Bills, Alex Gray, Jan Leike, Jakub Pachocki, Phil Tillet, Shantanu Jain, Greg Brockman, and Nick Ryder. 2022. ChatGPT: Optimizing Language Models for Dialogue. *OpenAI*.

Walter Sinnott-Armstrong. 2022. Consequentialism. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Winter 2022 edition. Metaphysics Research Lab, Stanford University.

Michael Slote. 1985. Utilitarianism, Moral Dilemmas, and Moral Cost. *American Philosophical Quarterly*, 22(2):161–168.

Taylor Sorensen, Liwei Jiang, Jena Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2023. Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. *arXiv preprint arXiv:2309.00779*.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the Machine Learning of Ethical Judgments from Natural Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779, Seattle, United States. Association for Computational Linguistics.

Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. 2023. Exploring Large Language Models' Cognitive Moral Development through Defining Issues Test. *arXiv preprint arXiv:2309.13356*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F Chen. 2023. SeaEval for Multilingual Foundation Models: From Cross-Lingual Alignment to Cultural Reasoning. *arXiv preprint arXiv:2309.04766*.

Bernard Williams. 1988. Ethical consistency. *Essays on moral realism*, pages 41–58.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Jingyan Zhou, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. 2023. Rethinking Machine Ethics–Can LLMs Perform Moral Reasoning through the Lens of Moral Theories? *arXiv preprint arXiv:2308.15399*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.