EXPLORING PERSONA-DEPENDENT LLM ALIGNMENT FOR THE MORAL MACHINE EXPERIMENT

Jiseon Kim[1]*, **Jea Kwon**[2]*, **Luiz Felipe Vecchietti**[2]*, **Alice Oh**[1], **Meeyoung Cha**[1,2]
[1]Korea Advanced Institute of Science & Technology (KAIST), Daejeon, South Korea
[2]Max Planck Institute for Security and Privacy (MPI-SP), Bochum, Germany
jiseon_kim@kaist.ac.kr
{jea.kwon,felipe.vecchietti,mia.cha}@mpi-sp.org
alice.oh@kaist.edu

ABSTRACT

Deploying large language models (LLMs) with agency in real-world applications raises critical questions about how these models will behave. In particular, how will their decisions align with humans when faced with moral dilemmas? This study examines the alignment between LLM-driven decisions and human judgment in various contexts of the moral machine experiment, including personas reflecting different sociodemographics. We find that the moral decisions of LLMs vary substantially by persona, showing greater shifts in moral decisions for critical tasks than humans. Our data also indicate an interesting partisan sorting phenomenon, where political persona predominates the direction and degree of LLM decisions. We discuss the ethical implications and risks associated with deploying these models in applications that involve moral decisions.

1 INTRODUCTION

Large language models (LLMs) demonstrate remarkable capabilities in various natural language processing tasks such as dialogue systems (Achiam et al., 2023), reasoning (Guo et al., 2025), and robotic applications (Zeng et al., 2023). Despite these advances, deploying LLMs in real-world applications for mission-critical tasks and sensitive scenarios, such as a key component of autonomous vehicles and ethical decision-support systems, raises urgent questions about their potential risks and robustness. In particular, it is important to understand how these models make ethical decisions in moral dilemma scenarios and whether they align with human values.

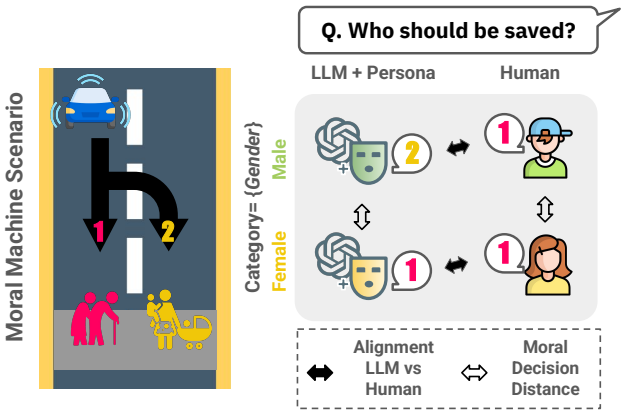

Figure 1: Analysis setting for exploring the persona-dependent LLM alignment in the moral machine experiment introduced in Awad et al. (2018).

*Equal contribution.

Previous studies (Ahmad & Takemoto, 2024; Takemoto, 2024; Jin et al., 2024) have examined the behavior of LLMs in answering moral dilemmas based on the moral machine experiment framework (Awad et al., 2018). Their findings highlighted that there is a large variation in alignment results according to each LLM version and their training methodology. The results in Ahmad & Takemoto (2024) and Takemoto (2024) were mainly shown at an aggregate level, while Jin et al. (2024) explored the impact of describing scenarios in different languages on human alignment. However, there is a lack of research exploring how context-dependent factors, such as demographic characteristics and the prompt of LLM, influence these LLM decisions. This gap leaves us with an incomplete understanding of how LLMs align with human moral reasoning in diverse contexts.

In this paper, we address this gap by investigating **persona-dependent LLM alignment** within the moral machine experiment. We construct personas based on key sociodemographic context—**age, gender, culture, religion, political orientation, income, and education**—to examine how these factors influence moral decisions in LLMs. By comparing LLM decisions under different persona conditions with human responses for each subgroup, we aim to uncover the extent to which LLMs reflect human-like biases. For analysis, we define a distance metric to show how LLM alignment deviates by different personas.

Our findings reveal nuanced variability in moral decision-making patterns for different personas, including critical shifts in specific scenarios. These results offer new insights into the alignment of LLM with human moral values and the ethical implications of deploying these models in morally sensitive applications. Our contributions include:

- **Offering evidence that sociodemographic persona influences LLM decisions:** We examine how moral decisions of LLM change by the personas of varying sociodemographic factors and show greater decision variability than humans, even in critical dilemmas.
- **Proposing a distance metric on LLM alignment:** We present a metric to measure the alignment between LLM and human moral judgments, which helps to capture shifts in decision boundaries across different personas.
- **Discussing ethical risks in LLM decision-making:** We interpret our findings through the partisan sorting theory, showing that the political domain most strongly affects LLM decisions. This finding raises concerns about bias amplification and real-world deployment.

2 Related Work

The Moral Machine Experiment In this experiment, Awad et al. (2018) collected human preference data from millions of participants worldwide for moral dilemmas related to autonomous driving scenarios. The scenarios involved binary decisions in which participants had to choose specific subjects to save, such as a younger or older person. The experiments also included a survey that presents demographic information from participants, such as demographic and socioeconomic status, to analyze their influence on moral decision-making. Here, we adopt seven demographic factors (age, gender, culture, religion, political orientation, income, and education) from the survey as reference points to establish personas and assess the LLM decision patterns for the moral machine experiment.

LLMs and Moral Reasoning The response of LLMs to scenarios in the moral machine experiment has been investigated by Takemoto (2024), revealing that LLM alignment varies between different models and differs from the human responses collected by Awad et al. (2018). The original analysis was also extended by Ahmad & Takemoto (2024) for a large set of LLMs, providing comparative information on model behavior in both commercial and open-source models. The study provided important information on how LLMs trained with different methodologies align with human responses, showing a trade-off between model size and ethical reasoning quality. In Jin et al. (2024), the scenarios were investigated by translating the vignettes into more than 100 languages focusing on the cultural biases of the LLMs. In contrast to these works, our focus is on the sociodemographic and cultural context provided to the LLM via persona-setting.

Persona Setting in LLMs The rise of LLMs has sparked strong interest in persona modeling for prompt design, as reviewed by Chen et al. (2024) and Tseng et al. (2024). These studies discuss how conditioning different personas affects LLM behavior, particularly in shaping decision patterns

(Yang et al., 2023) and exposing or aggravating potential biases (Gupta et al., 2024). Specifically, Simmons (2022) demonstrated that when an LLM is set with a liberal or conservative political identity in the US context, the models generate text biased to reflect these identities.

3 PERSONA-DEPENDENT LLM ALIGNMENT

Category	Group	Persona Details
Age	Older	are {32-75} years old
	Younger	are {18, 19, 20, 21} years old
Education	Less Educated	have a high school education as your highest level of education
	More Educated	have a PhD
Gender	Male	are a man
	Female	are a woman
Income	Higher Income	earn more than $80k a year
	Lower Income	earn less than $10k a year
Political	Conservative	are a conservative person
	Progressive	are a progressive person
Religion	Not Religious	are a non-religious person
	Very Religious	are a very religious person
Culture	Western	are {American, Brazilian, German, British, Russian, Canadian, Italian, Australian, Polish, Spanish}
	Eastern	are {Japanese, Korean, Indian, Chinese, from Hong Kong, Taiwanese, Indonesian, Malaysian, Saudi, Emirati}

Table 1: Persona setting for each category. Referring to the categories and criteria for each group presented in Awad et al. (2018), here, we define a pair of contrasting persona groups for each category. These personas are assigned to the model using template prompts followed by the textual description of the scenario to be evaluated.

Here, we study how decisions taken by LLM vary under different personas and compare these results to human responses for the corresponding subgroups in the moral machine experiment. A pictorial diagram of the proposed framework is shown in Fig. 1, and detailed information about the evaluated scenarios is provided in the Appendix A.1.

3.1 PERSONA SETTING MODELING TO COLLECT LLM RESPONSES

Persona Category We model personas across seven categories–*age, education, gender, income, political affiliation, religion, and culture*. Each category includes a pair of personas with opposing characteristics, defined based on the demographics of individuals who participated in the survey (see Appendix A.2 for survey data details). The demographic criteria are specified as follows: *age* is defined by the first and higher quantile thresholds from the annotator demographics. Western and Eastern countries were determined following the definitions in Awad et al. (2018), where we define personas corresponding to nationalities from the top 10 countries by participant count. The criteria for the other categories are presented in Table 1.

Persona Prompt To assign personas to LLMs, we use a persona template from Gupta et al. (2024), incorporating the assigned persona instruction at the beginning of the prompt for each scenario:

```
``You {persona}.  Your responses should closely mirror the
knowledge and abilities of this persona.''
```

3.2 HUMAN RESPONSE BASELINES FROM THE MORAL MACHINE EXPERIMENT

The baseline human responses for all categories, except *culture*, were directly obtained from the survey data in Awad et al. (2018). We use the Average Marginal Component Effect (AMCE), which is a metric in causal inference and conjoint analysis to quantify the effect of a specific attribute on decision-making outcomes (Hainmueller et al., 2014). The range of AMCE is from -1 to 1, where: (i) a value of 0 means the attribute has no systematic effect on the decision; (ii) a value greater than

0 indicates a preference for that attribute; and (iii) a value less than 0 indicates a preference against that attribute, i.e., a preference for the opposing attribute.

In the context of the Moral Machine experiment, the AMCE measures the average impact of a given factor (e.g., *age*, *gender*, and *species*) on moral decisions while controlling for other variables. A positive AMCE value for an attribute (e.g., "saving younger individuals") means that, on average, this attribute increases the likelihood of a decision favoring that group, whereas a negative value suggests the opposite effect. We obtain human baseline results from AMCE calculations in Awad et al. (2018). For culture, we computed the AMCE as a weighted average across all Eastern and Western countries as follows:

$$\text{AMCE}_{\text{Culture}} = \frac{\sum_{k \in R} n_k \cdot \text{AMCE}_k}{\sum_{k \in R} n_k},$$

where n_k is the number of response count for country k in region R (Eastern or Western).

3.3 Moral Decision Distance

We compute the AMCEs for 9 scenarios (Number of Characters, Interventionism, Fitness, Gender, Species, Social Status, Relation to AV, Age, Law) in the moral machine under two contrasting personas, e.g., male vs. female, older vs. younger, for LLM responses and human responses. For each persona, the AMCE values are represented as vectors:

$$\mathbf{v}^{(1)} = [\text{AMCE}_1^{(1)}, \ldots, \text{AMCE}_9^{(1)}]$$
$$\mathbf{v}^{(2)} = [\text{AMCE}_1^{(2)}, \ldots, \text{AMCE}_9^{(2)}]$$

We define a Moral Decision Distance (MDD) metric as the Euclidean distance between these vectors as a measure of the alignment gap between personas[1] , where a lower MDD indicates less variation in moral decisions across the contrasting personas:

$$\text{MDD} = \|\mathbf{v}^{(1)} - \mathbf{v}^{(2)}\|_2 = \sqrt{\sum_{i=1}^{9} \left(\text{AMCE}_i^{(1)} - \text{AMCE}_i^{(2)}\right)^2}. \quad (1)$$

4 Results

4.1 Experimental Setting

Scenarios We generate 10,000 moral decision scenarios using constrained randomization across nine categories, interventionism, gender, social status, species, utilitarianism, relationship to the autonomous vehicle (AV), fitness, concern for law, and age, following the methodology presented in Takemoto (2024). We then apply the proposed personas to the scenarios generated for these categories as well as running a baseline measurement for each scenario without setting any persona.

Models We investigate three widely used models: GPT-4o (`gpt-4o-2024-05-13` (Hurst et al., 2024)) and GPT-3.5 (`gpt-3.5-turbo-0613`, released in June 2023) [2], and Llama2 (`Llama-2-7b-chat-hf`) [3] (Touvron et al., 2023). We select three top-performing models from their LLM series based on their strong alignment with human responses reported in Takemoto (2024), and their high valid response rate (i.e., a high frequency of providing a clear answer in binary question moral dilemma scenarios) as detailed in Table 4. For the GPT models, we use the default parameter settings provided by the Azure OpenAI API with temperature set to 1 and top_p set to 1. We adopt the same hyperparameter settings from Takemoto (2024) for Llama2, with top_k equal to 10, top_p equal to 0.9, a max_length of 512, and temperature set to 0.4.

We obtain the responses for the three models and analyze the effect of the persona (Table 1) by comparing with human demographic subgroups (Table 2) and observing decision shifts (Fig. 2).

[1] While calculating the distance between two AMCEs was also used for alignment with humans, we specifically confine MDD to two counterpart personas in this work.

[2] `https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models`

[3] `https://huggingface.co/meta-llama/Llama-2-7b-chat-hf`

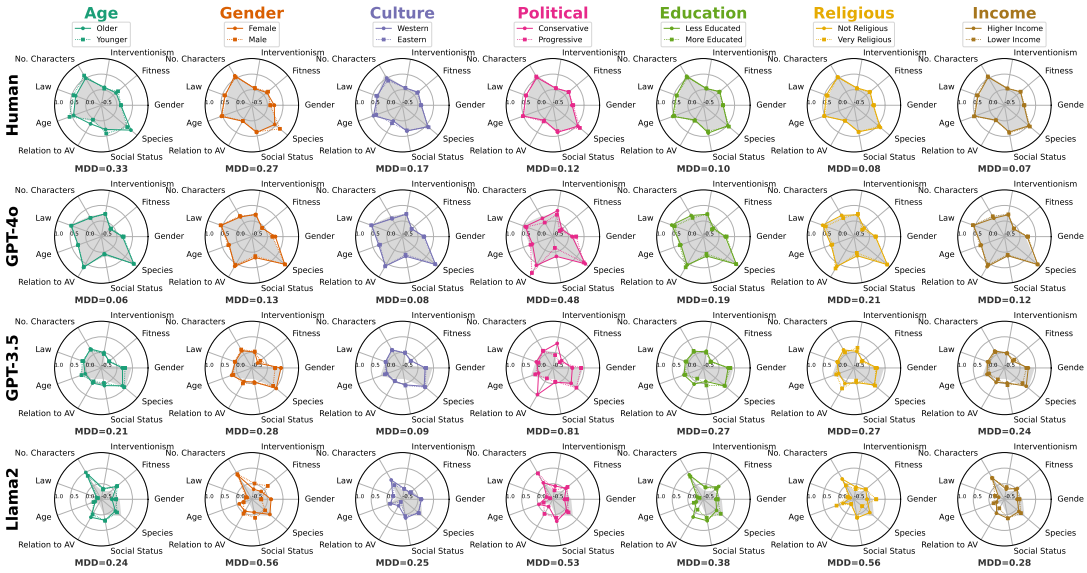

Figure 2: Comparison between responses from humans (Awad et al., 2018) and responses from GPT-4o, GPT-3.5, and Llama2. The gray-shaded radar for the Human class represents the aggregate results for all the participants. The gray-shaded radar for LLMs represents the baseline results when the scenario is prompted with no assigned persona.

Category		Age		Education		Gender		Income		Political		Religious		Culture	
	Base	Old	Young	Less	More	Female	Male	High	Low	Cons.	Prog.	Less	Very	West	East
GPT-4o	**0.852**	**0.707**	0.954	**0.869**	0.884	**0.878**	**0.736**	0.912	**0.847**	**0.901**	1.160	**0.806**	0.962	0.837	0.717
GPT-3.5	1.015	0.771	**0.894**	0.931	**0.867**	0.915	0.860	**0.862**	0.976	1.077	**0.994**	0.929	**0.945**	**0.783**	**0.694**
Llama2	1.225	1.278	1.315	1.242	1.179	1.177	1.278	1.212	1.122	1.174	1.259	1.300	1.313	1.095	1.180

Table 2: Comparison of alignment scores between human and LLMs based on the Euclidean distance of AMCE values (lower is better). Bold values indicate the lowest value in each column.

4.2 How do LLM and human responses align given the same demographic?

Fig. 2 shows the alignment between LLM and human responses, where the differences in moral decision-making responses between LLM and humans are visually evident. In particular, consistent with the findings from Ahmad & Takemoto (2024), GPT-4o demonstrates the strongest alignment with human responses in the baseline setting. However, as seen in the smaller radar plot areas, GPT-3.5 and Llama2 show higher misalignment patterns. When personas are applied to LLMs, AMCE values that measure alignment with human judgment deviate from the baseline.

Across the models, GPT-4o and GPT-3.5 exhibit a higher alignment given the same human demographic compared to Llama2 with alignment counts equal to (number of bold characters excluding the baseline column in Table 2): (i) GPT-4o: 7/14, (ii) GPT-3.5: 7/14, (iii) Llama2: 0/14. GPT-3.5 shows the most significant overall increase with persona-dependent prompts compared to its human demographic baseline, where 13/14 personas increase alignment over the baseline.

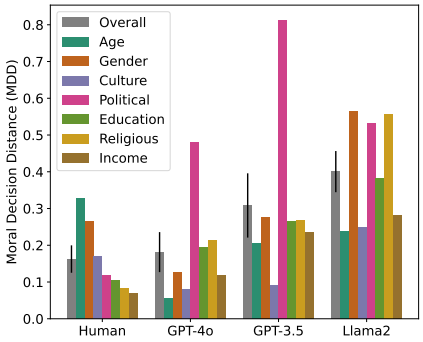

Figure 3: Moral machine distance for different subgroups for human, ChatGPT-4o, GPT-3.5, and Llama 2 responses.

4.3 How does alignment vary for LLMs across contrasting personas?

We illustrate the difference in alignment between contrasting personas, measured as MDD values, for different demographic subgroups in Fig. 3. For human responses, the greatest distances are observed for *age* and *gender*, indicating these factors strongly influence human decisions in moral machine scenarios. For LLM decisions, we found an overall increase in MDD values, with a particularly pronounced surge when a *political* persona was applied. Specifically, while GPT-4o generally performs robustly in different personas, it records a high deviation of 0.48 for political personas (conservative versus progressive). A similar pattern for the *political* persona is observed for GPT-3.5 and Llama2. Compared to GPT-4o these models also consistently exhibit higher distances for all other subgroups as well, with Llama2 displaying the highest average MDD values. These findings indicate that, while human moral decisions remain similar regardless of *political views* in the

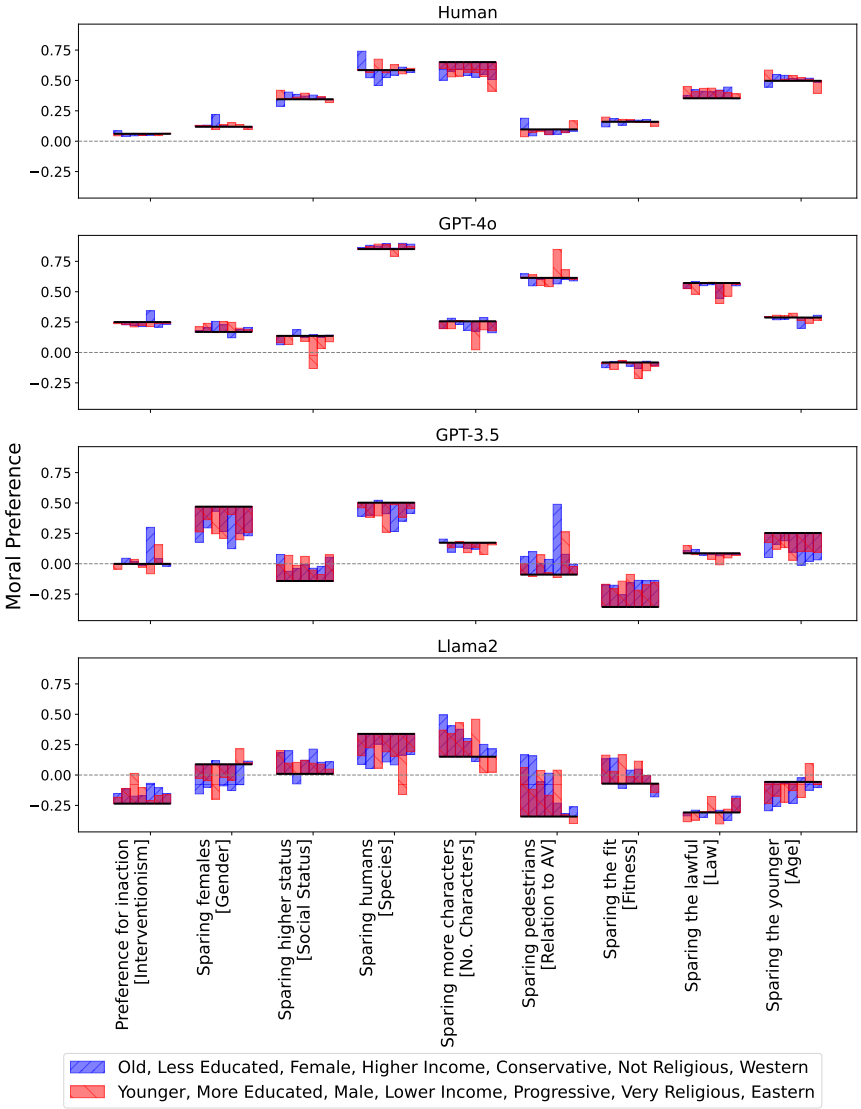

Figure 4: Comparison of moral preferences highlighted across nine dimensions for different models with assigned personas. A value of 0 on the y-axis indicates no preference between the two groups. The solid horizontal line represents the default setting without personas. Bars represent results for 14 personas per dimension, with red and blue pairs showing opposing personas. Human moral decisions do not flip across any persona, while GPT-4o, GPT-3.5, and Llama2 exhibit more dimensions preferences below 0, indicating opposing moral decisions to humans.

moral machine experiment, LLM decisions do not; assigning a *political* persona to LLMs led to polarization bias, represented by high MDD values, in the same scenarios.

Specifically in the case of GPT-4o, which demonstrates the closest alignment with human moral decisions, assigning political personas introduces significant variation in the social status dimension. The conservative persona tends to favor individuals with higher social status and prioritize authority, while the progressive persona exhibits the opposite tendency. This pattern mirrors the findings of Abdulhai et al. (2023), which showed that LLMs reflect biases tied to moral foundations and political ideologies, including a stronger emphasis on authority among conservatives. These results suggest that LLM decisions shift more noticeably under political personas, aligning with the political preferences they are modeled after, which could be interpreted as a form of political sycophancy.

4.4 Do decisions change for specific personas and models?

When deploying AI models in morally sensitive scenarios, a critical aspect to consider is robustness. Because if a moral decision shifts from its baseline or human alignment due to the assignment of a specific persona, it can lead to ethical risks. To evaluate robustness, we compare the moral preferences for each of the nine dimensions across different personas. We show results on such robustness. In Fig. 4, a value of 0 on the y-axis indicates no preference for the given dimension, which means that the model randomly selects between the two groups. The solid horizontal line for each dimension represents the baseline value in the default setting, where no persona is assigned. Each bar represents the results for 14 different personas per dimension, with red and blue pairs indicating opposing personas.

Human preferences remain consistently above 0 on all dimensions, reflecting a common consensus toward a specific group, regardless of demographics. In contrast, more dimensions exhibit preferences below 0 for LLM, indicating opposing preferences to humans. Although baseline preferences align positively with human decisions, assigning specific personas (e.g., a *progressive* persona) to GPT-4o leads to a sharp decline in the preference value for Social Status, even reversing the baseline preference. This shift in the preference direction is more frequent in GPT-3.5 and Llama2.

Moral decision shifts are presented in Fig. 5. For these two LLMs, nearly one-fifth of all decisions are misaligned with human responses. Unlike in human judgment that showed less variability due to demographic differences in moral preferences, LLMs seem more susceptible to changes in preference, showing greater vulnerability to context.

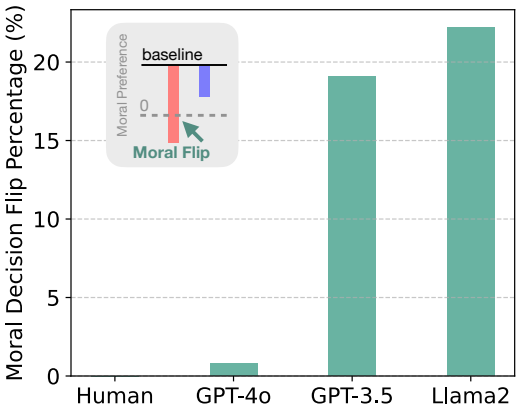

Figure 5: The percentage of decisions showing a shift from the human baseline for the studied LLMs. GPT-4o exhibits better alignment with humans, with a low percentage of misaligned decisions. In contrast, the other two LLMs show around 20 percent of moral decisions misaligned with human responses.

4.5 What are the variation patterns across personas and moral scenarios?

Fig. 6 shows the deviation from the baseline for each scenario (that is, persona). On the y-axis, a bar above 0 means that the assigned persona shifts the LLM's moral preference toward the respective group. In general, human preferences exhibit small variance, suggesting a general alignment toward a common consensus between personas. In contrast, LLM preferences fluctuate based on the assigned persona. The trends differ across models: GPT-4, for instance, shows high variance predominantly by *political* persona setting.

On the other hand, GPT-3.5 and Llama2 show large changes for all other persona settings. In these two models, there are dimensions where similar trends emerge regardless of the persona. Typically, this includes a preference for saving pedestrians over passengers in the "Relation to AV" scenario,

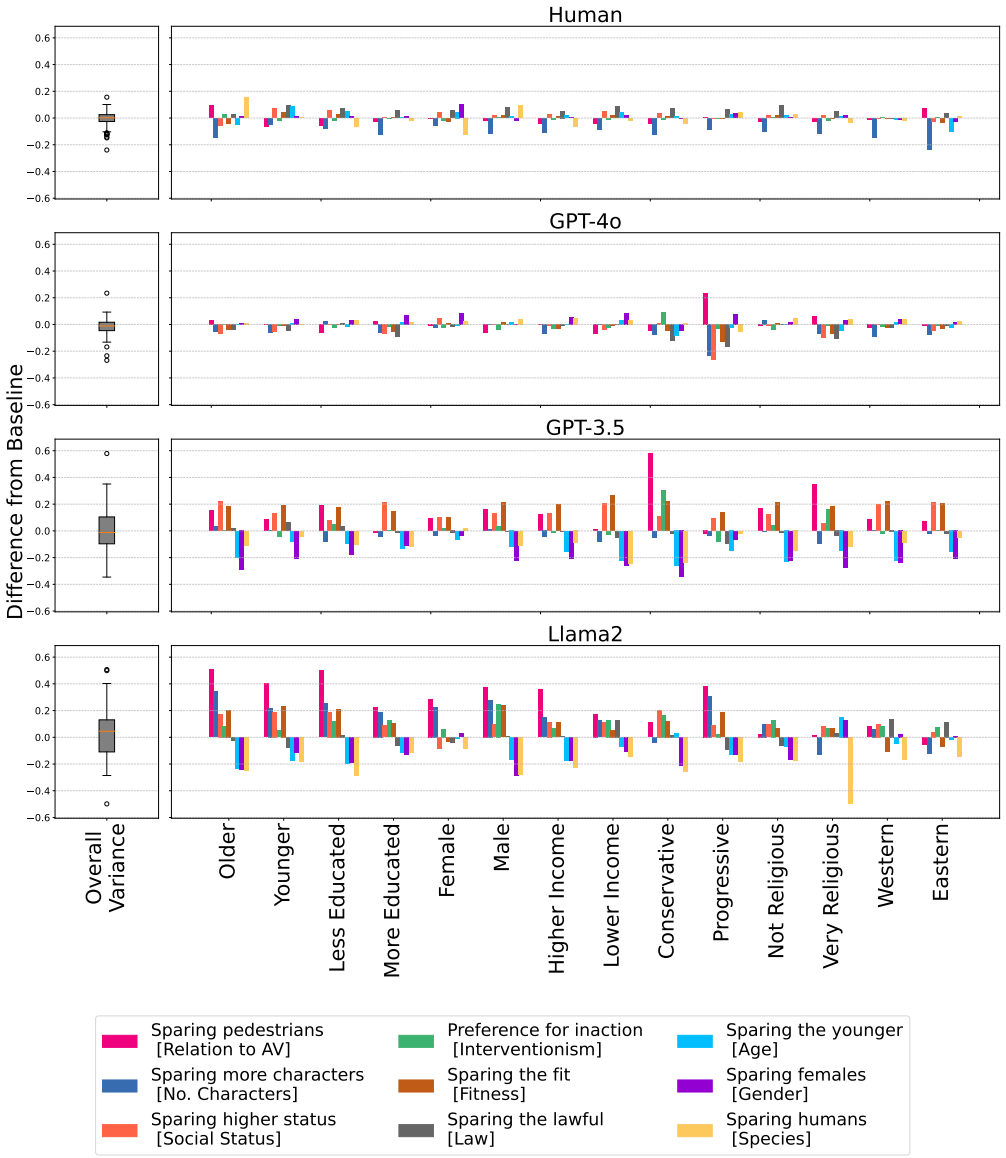

Figure 6: Variation in moral preferences highlighted across personas for different models. Bars above 0 indicate that the LLM's preferences shift toward the corresponding group. Humans show minimal variance, while LLMs exhibit fluctuation based on each persona. GPT-4, particularly exhibits high variance only with the *political* persona, while GPT-3.5 and Llama2 have high variance across most persona settings.

for saving the fit person over the obese person in the "Fitness" scenario, and for saving the higher status person in the "Social Status" scenario. Conversely, in dimensions where preferences decrease with persona assignment, these shifts trend toward saving men over women in the "Gender" scenario, sparing pets over humans in the "Species" scenario, and saving older individuals over younger ones in the "Age" scenario. This variability highlights the susceptibility of LLMs to different personas and contexts, contrasting with the more consistent moral decisions observed in humans.

5 Discussion

Contextual Decision Shifts and Misalignments The observed shifts in moral decision-making based on different personas have significant implications for the ethical deployment of LLMs. Such misalignments can be critical in real-world applications where context-dependent decisions may

diverge from baseline or human responses. These shifts highlight the potential risks of using LLMs for decision-making in areas like autonomous vehicles or ethical support systems, where consistency with human moral values is paramount. The vulnerability of LLMs, especially when influenced by specific socio-demographic factors, could exacerbate ethical concerns, making it crucial to ensure that models reflect a broader spectrum of human values and are robust to contextual influences.

Partisan Identity Effect Based on the political context worldwide, research by Mason (Mason, 2015) has shown that political views align individuals with similar social identities on multiple dimensions. In the moral machine experiment, we observe that human results on which lives to save were robust to this partisan identity effect, showing only small differences between conservative and progressive subgroups. However, for LLM, the political persona exhibited significant changes in decisions, particularly for GPT-4o.

Our finding on how LLM decisions favor or disadvantage certain groups based on political beliefs raises ethical concerns and questions about the model's fairness. As Abdulhai et al. (2023) explored, perceived political bias in LLMs can erode trust, causing hesitation in their deployment for critical systems. These conflicting results open new research directions in which it is important to define the social identity characteristics of different political views and investigate their subgroup alignment for moral dilemmas.

Improving Persona Settings for LLMs Our persona-based approach offers valuable insights; however, it should be expanded to consider a broader range of demographic attributes and their combinations. The current model, which relies on binary personas in seven categories, simplifies the complexity of human moral decision-making and the impact of these models on minority groups. Future research should investigate multidimensional persona constructs that incorporate factors like cultural context and nuanced political views and consider intersectional groups. In addition, we defined personas using a single prompt. A large-scale analysis with different methodologies for setting personas and contexts associated with these personas is needed to better evaluate LLM robustness when making moral decisions.

Expanding Moral Machine Scenarios While the moral machine framework has been instrumental in assessing LLM alignment with human judgement, it has limitations in its application to a broader spectrum of moral dilemmas (Etienne, 2021; Dewitt et al., 2019; LaCroix, 2022). The moral machine experiment predominantly scrutinizes ethical choices related to autonomous vehicles, which represents a narrow subset of ethical issues. Expanding this framework to include other types of moral decisions, such as allocation of healthcare resources, judicial decisions, or climate change policies, can offer a more comprehensive understanding of LLM behavior in diverse contexts. Gathering human responses in interactive and dynamic scenarios is also crucial to create a realistic testing ground for how LLMs manage complex, real-world ethical decisions.

6 Conclusion

We studied how LLMs align with human moral decisions across various personas that reflect key socio-demographic factors. Our results show that the alignment of LLM with human responses is context-dependent, with clear variations in decision-making patterns depending on the input persona. Recent models like GPT-4, GPT-3.5, and Llama2 exhibited substantial changes in decisions of moral preferences by persona setting when compared to human responses. Among the interesting findings was the role of certain factors; political orientation in persona led to the largest degree of shift in decision boundaries. This pronounced bias in political persona in the context of autonomous driving raises the immediate question of whether the decisions made by LLMs align with broader human values and can be influenced by targeted interactions with the user.

These findings highlight the sensitivity of current LLMs to diverse perspectives and help us better plan the use of these models in handling ethically complex scenarios in real-world applications. Although our study primarily used simplified personas, future research can expand on these insights by incorporating multidimensional constructs, providing a more realistic description of human participants. This would further improve robustness and ethical decision-making in a broader range of contexts, ensuring a better reflection of the diversity of human values.

7 ACKNOWLEDGEMENTS

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-II220184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics).

REFERENCES

Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. Moral foundations of large language models. *arXiv preprint arXiv:2310.15337*, 2023.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Muhammad Shahrul Zaim bin Ahmad and Kazuhiro Takemoto. Large-scale moral machine experiment on large language models. *arXiv preprint arXiv:2411.06790*, 2024.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*, 2024.

Barry Dewitt, Baruch Fischhoff, and Nils-Eric Sahlin. 'Moral machine' experiment is no basis for policymaking. *Nature*, 567(7746):31–31, 2019.

Hubert Etienne. The dark side of the 'moral machine'and the fallacy of computational ethical decision-making for autonomous vehicles. *Law, Innovation and Technology*, 13(1):85–107, 2021.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *ICLR*, 2024.

Jens Hainmueller, Daniel J Hopkins, and Teppei Yamamoto. Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political analysis*, 22 (1):1–30, 2014.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, et al. Language model alignment in multilingual trolley problems. *arXiv preprint arXiv:2407.02273*, 2024.

Travis LaCroix. Moral dilemmas for moral machines. *AI and Ethics*, 2(4):737–746, 2022.

Lilliana Mason. "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American journal of political science*, 59(1):128–145, 2015.

Gabriel Simmons. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *arXiv preprint arXiv:2209.12106*, 2022.

Kazuhiro Takemoto. The moral machine experiment on large language models. *Royal Society open science*, 11(2):231393, 2024.

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

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*, 2024.

Fan Yang, Zheng Chen, Ziyan Jiang, Eunah Cho, Xiaojiang Huang, and Yanbin Lu. Palr: Personalization aware llms for recommendation. *arXiv preprint arXiv:2305.07622*, 2023.

Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S Yu. Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226*, 2023.

A APPENDIX

A.1 SCENARIOS IN THE MORAL MACHINE EXPERIMENT

We investigate nine scenarios in the Moral Machine Awad et al. (2018). For each scenario, the AMCE value represents the difference between the probability of sparing two characters. For example, a positive value in the "Age" scenario represents a preference for sparing the young over the elderly. The details of each scenario are shown in Table 3.

Scenario	Description
Intervention	Preference for inaction over for action
Relation to AV	Preference for sparing pedestrians over passengers
Gender	Preference for sparing females over males
Fitness	Preference for sparing the fit over the large
Social Status	Preference for sparing higher status over lower status
Law	Preference for sparing the lawful over the unlawful
Age	Preference for sparing the young over the elderly
No. characters	Preference for sparing more characters over fewer characters
Species	Preference for sparing humans over pets

Table 3: Description of the Moral Machine scenarios.

A.2 SOCIODEMOGRAPHIC SURVEY DATA

The moral machine experiment also contained a demographic survey as an extension of the user interface to collect demographic information and user feedback. The survey data includes information on age, education, gender, income, politics, and religion. For political leaning, the user is asked to choose based on a scale from 0-1 his political preference from conservative to progressive. For religious, the user is asked to choose based on a scale from 0-1 his religious preference from non-religious to religious. The demographic survey data contains around 11.2 million answers from 463,675 users.

The sociodemographic distribution of the survey participants in Awad et al. (2018) is shown in Fig. 7. For age, it is possible to see that the majority of participants are under 30 years old. For gender, there is an imbalance between males and females, where males are more prevalent in the participants. The participants are well distributed across different educational levels. For income, the majority of participants declared income lower than 5000 thousand dollars per year. For the Political category, a value closer to 0 means a person with more conservative views while a value closer to 1 has more progressive views. In particular, many participants declared themselves to be in a neutral part of the political spectrum. For the Religious category, a value closer to 0 is related to non-religious participants while a value closer to 1 is related to religious participants.

A.3 RESPONSE RATE FOR LLMS

In Table 4 we show the valid response rate for each model analyzed. Given the sensitive content of moral dilemmas, GPT-4o, GPT-3.5, and Llama2 have security guardrails that prevent them from answering these questions. In the table, we consider invalid answers in which the model refuses to answer or does not explicitly choose one option. From the results, it is possible to observe that Llama2 obtains the lower valid response rate, in which when setting a conservative or a religious persona less than ten percent of the answers are valid. For GPT-4o and GPT-3.5 the response rate is high for all personas.

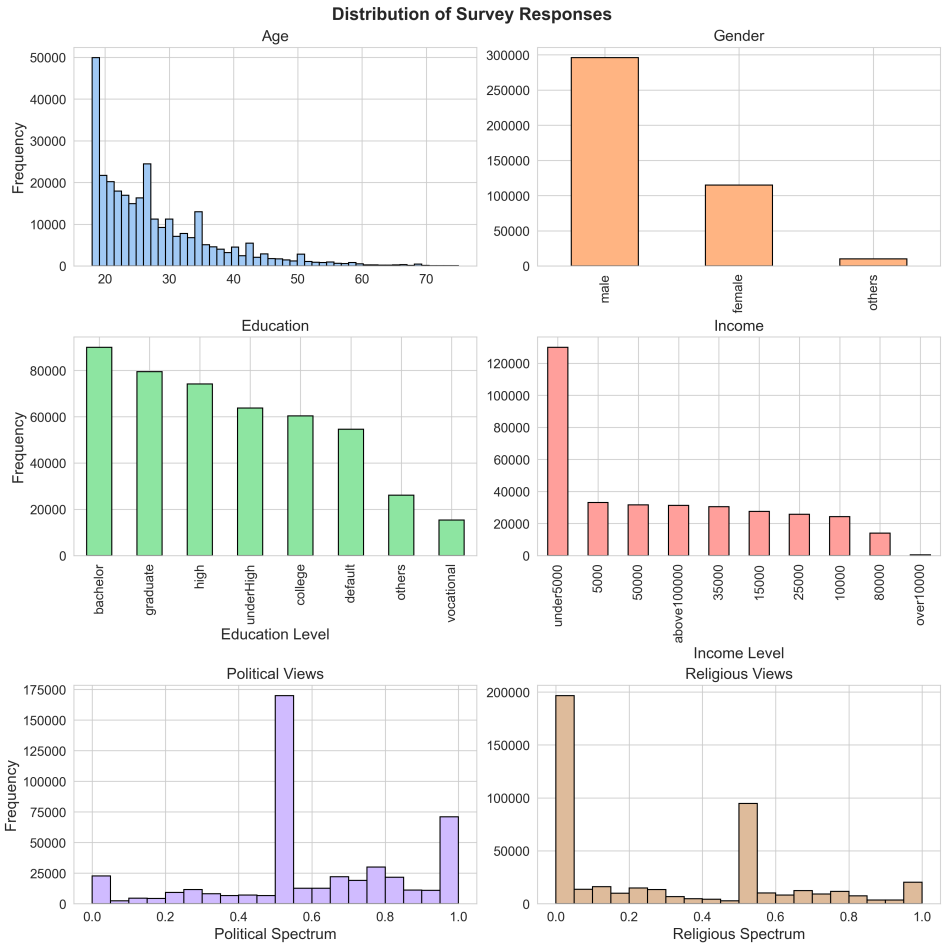

Figure 7: Sociodemographic distribution for survey participants in Awad et al. (2018). Empty values were filtered and for the Age category, participants were filtered to be within the [18, 75] range.

Persona	GPT-4o	GPT-3.5	Llama2
base	0.940	0.860	0.932
old	0.974	0.974	0.675
young	0.979	0.964	0.424
less_edu	0.971	0.987	0.759
more_edu	0.929	0.800	0.528
female	0.946	0.978	0.171
male	0.966	0.987	0.114
high_income	0.966	0.980	0.551
low_income	0.989	0.989	0.508
conservative	0.974	0.998	0.072
progressive	0.925	0.970	0.362
non_religious	0.974	0.787	0.671
religious	0.899	0.941	0.060
western	0.986	0.991	0.218
eastern	0.966	0.989	0.325

Table 4: Valid response rate for each model with an assigned persona. This represents the proportion of responses to the trolley problem where the model clearly selects one of the two options. Responses are considered invalid if the model refuses to answer or does not explicitly choose one option.