Revisiting the Trolley Problem for AI: Biases and Stereotypes in Large Language Models and their Impact on Ethical Decision-Making

Sahan Hatemo[1], **Christof Weickhardt**[1], **Luca Gisler**[1], **Oliver Bendel**[2]
[1]FHNW School of Computer Science
[2]FHNW School of Business
sahan.hatemo@students.fhnw.ch, christof.weickhardt@students.fhnw.ch, luca.gisler@students.fhnw.ch,
oliver.bendel@fhnw.ch

Abstract

The trolley problem has long served as a lens for exploring moral decision-making, now gaining renewed significance in the context of artificial intelligence (AI). This study investigates ethical reasoning in three open-source large language models (LLMs)—LLaMA, Mistral and Qwen—through variants of the trolley problem. By introducing demographic prompts (age, nationality and gender) into three scenarios (switch, loop and footbridge), we systematically evaluate LLM responses against human survey data from the *Moral Machine* experiment. Our findings reveal notable differences: Mistral exhibits a consistent tendency to over-intervene, while Qwen chooses to intervene less and LLaMA balances between the two. Notably, demographic attributes, particularly nationality, significantly influence LLM decisions, exposing potential biases in AI ethical reasoning. These insights underscore the necessity of refining LLMs to ensure fairness and ethical alignment, leading the way for more trustworthy AI systems.

Introduction

The trolley problem has long been a foundational concept in moral philosophy, posing a dilemma that explores the trade-off between sacrificing one life to save many. Originally proposed by Philippa Foot in 1967 and expanded by Judith Jarvis Thomson in 1985, it has evolved from a thought experiment into a tool for examining ethical reasoning in humans. As artificial intelligence (AI) systems, particularly large language models (LLMs), are increasingly integrated into decision-making processes, the relevance of such moral dilemmas has grown—making these systems active participants in shaping outcomes in healthcare, legal systems and beyond. Understanding how LLMs navigate ethical dilemmas is essential for ensuring their alignment with human values.

This study revisits the trolley problem to investigate the biases inherent in three contemporary open-source LLMs, LLaMA, Mistral and Qwen, by examining their reflection of stereotypes. Our research extends classical ethical frameworks by exploring how these models handle moral dilemmas when prompted with diverse demographic characteristics, including age, nationality and gender. Using a dataset inspired by the *Moral Machine* experiment (Awad 2021), we analyze over 852 unique demographic combinations across three trolley problem variants: the switch, the loop and the footbridge scenario, the latter being also known as fat man problem. Each scenario is presented in multiple languages to capture cultural and linguistic nuances.

Similar studies, such as *Language Model Alignment in Multilingual Trolley Problems* (Jin et al. 2024), examine how 19 LLMs align with human preferences across more than 100 languages. Using data from the *Moral Machine* experiment, they analyze six moral dimensions—species, gender, fitness, status, age and number of lives—and reveal cross-lingual ethical biases, underscoring the need for diverse perspectives in AI ethics.

Similarly, the study *ChatGPT's advice drives moral judgments with or without justification* (Krügel, Ostermaier, and Uhl 2025) gathered online participant responses to assess the impact of advice in a trolley problem scenario. It compares conditions with and without an argument, finding that the argument does not change the final decision; participants later justify their choices. The study also distinguishes between advice from ChatGPT and that from a moral advisor, noting differences in perceived plausibility and moral authority.

This work makes two primary contributions. First, it provides a systematic analysis of how demographic attributes influence LLMs' ethical reasoning in moral dilemmas. Second, it establishes a comparative framework for evaluating and benchmarking ethical decision-making across different LLMs, which is accessible here: https://shorturl.at/u2h3K.

Study Set Up

The Trolley Problem and its Variants

The trolley problem, originally posed by Foot, presents a moral dilemma in which a runaway trolley threatens five people tied to the tracks (Foot 1967). One must decide whether to do nothing or to pull the lever, causing the trolley to divert onto another track, where it will kill a single bystander instead (Figure 1).

Jarvis Thomson introduced key variations, including the loop variant (Figure 2) where the tracks circle back to the original five unless a lone individual on a sidetrack blocks the trolley and the footbridge variant, (Figure 3), pushing a large person to stop the trolley (Jarvis Thomson 1985).

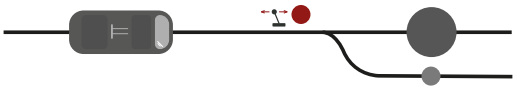

Figure 1: Schematic illustration of the switch variant of the trolley problem.

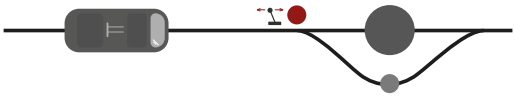

Figure 2: Schematic illustration of the loop variant of the trolley problem.

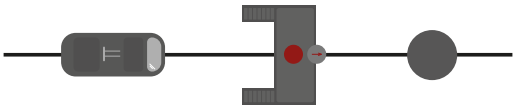

Figure 3: Schematic illustration of the footbridge variant of the trolley problem.

Since around 2012, these trolley scenarios have been extended to the realm of self-driving cars. The first considerations were made by Bendel in 2012 in the context of machine ethics (Bendel 2013). He investigated how a "robot car" should decide if an accident cannot be avoided. On the one hand, it could quantify, i.e., count possible accident victims and decide on the lowest number of victims. On the other hand, it could qualify or classify, i.e., take into account the gender, age, ethnicity or importance of the individuals. He came to the conclusion that in the case of humans, one should neither quantify nor qualify. An emergency stop may be required for a group of animals or rare species (Bendel 2016). The *Moral Machine* study later also placed self-driving cars rather than trolleys at the center of ethical decision-making, serving as an influential large-scale exploration of how humans judge AI-based moral dilemmas (Awad 2021).

Through such adaptations, the trolley problem continues to evolve as a powerful tool for examining ethical frameworks in emerging technologies, particularly when machines must make decisions that affect human lives.

Dataset Description

The *Classic Trolley – Moral Machine* (Awad 2021) dataset comprises moral decisions from participants across 169 countries. The study presents three distinct trolley problem variations, with participants responding to scenarios before providing demographic information including age, gender, education level, yearly income and political and religious views. Participant location was determined through IP address tracking (Awad et al. 2018). All three variations of the trolley problem have similar amounts of respondents to the scenario, as seen in Table 1.

To ensure statistical validity and robust cross-cultural analysis, we implemented a stringent selection criterion of including only countries where the language is an official language of the country and with approximately 1,500 unique participants in all three variants. This approach ensures diverse cultural, linguistic perspectives and less noise in the data, resulting in the country selection detailed in Table 3, which differs from the broader sample shown in Table 2.

Scenario	**Users**
Loop	68,457
Switch	68,303
Footbridge	68,159

Table 1: Unique user participation across scenarios.

Country	**Lang.**	**Users**	**Users All Vars.**
United States	en	14,321	13,056
France	fr	5,878	5,294
United Kingdom	en	5,684	5,243
Germany	de	4,550	4,120
Brazil	pt	4,067	3,772
Russian Federation	ru	3,065	2,883
Canada	en	2,229	2,032
Australia	en	1,867	1,687
Spain	es	1,640	1,492
Germany	en	1,513	1,304
Turkey	en	1,427	1,275
Poland	en	1,280	1,154
Italy	en	1,123	1,019
Netherlands	en	1,076	964
France	en	951	819

Table 2: Top 15 countries by total unique users and participation in all three scenarios with language pairing.

Methodological Limitations

Our analysis recognizes the complexities of contemporary cultural identity. Given globalization, international mobility and multicultural households, we cannot assume a direct correlation between reported residence and cultural orientation. To address potential geographical ambiguities, such as participation during international travel, we filtered participants based on their primary spoken language, as indicated in Table 3.

Country	**Lang.**	**Users All Vars.**
United States	en	13,056
France	fr	5,294
Germany	de	4,120
Brazil	pt	3,772
Russian Federation	ru	2,883
Spain	es	1,492

Table 3: Unique users participating in all three variants.

Another limitation of the data was the age distribution. Figure 4 shows a clipped exponential decay distribution,

starting at 18 years, highlighting the unbalanced data between young survey takers and older ones. This results in the data from young survey participants being more robust than the data from older survey participants which is comparatively sparse and less representative, potentially limiting the generalizability of findings across age groups.

Country	Female (%)	Male (%)
United States	31.6%	68.4%
Brazil	37.1%	62.9%
France	31.1%	68.9%
Germany	28.5%	71.5%
Russian Federation	28.3%	71.7%
Spain	32.1%	67.9%
Overall	31.2%	68.8%

Table 4: Gender distribution by country.

A further limitation of the data concerned in the discrepancy in gender representation. As shown in Table 4 most of the survey participants are male making an overall distribution of 31.2% female and 68.8% male. This distribution was not normalized to avoid further altering the inherent characteristics of the dataset. As a result, the imbalance in gender representation may influence the interpretation of the findings and limit the generalizability of the results to a more balanced population.

Methodology

Large Language Model Selection

To ensure transparent and unbiased testing, we opted to use three open-source LLMs (Table 5) developed by different laboratories, hypothesizing that each model is trained on varying datasets.

Model	Size	Hugging Face Model Identifier
LLaMA-3.1	8B	meta-llama/Llama-3.1-8B-Instruct
Mistral	8B	mistralai/Ministral-8B-Instruct-2410
Qwen	7B	Qwen/Qwen2.5-7B-Instruct

Table 5: Overview of contemporary open-source large language models utilized.

Framework

Drawing inspiration from the original human study (Awad 2021), we designed the framework for evaluating LLMs in a comparable manner by replicating the concept of variations posed in the *Moral Machine* experiment. Prior to presenting the LLMs with moral dilemmas, we employed procedural prompt generation to assign diverse characteristics. This process resulted in 852 unique combinations, encompassing 71 distinct ages, six nationalities and two genders. These combinations were systematically compiled into a structured JSON file, enabling consistent and reproducible application across experiments.

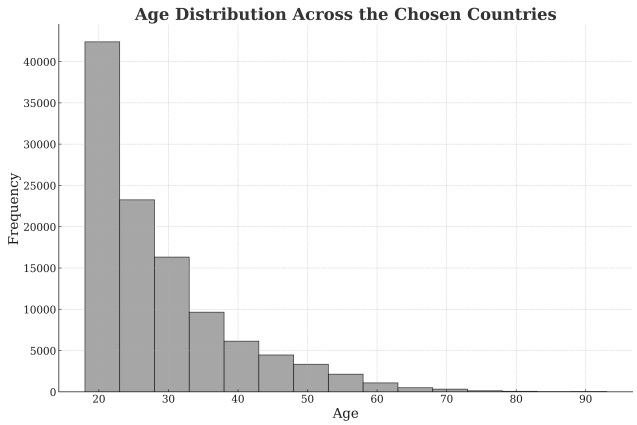

Figure 4: Age distribution of all participants from the total users of all variants, grouped into 5-year intervals from the *Moral Machine* dataset.

Prompt Construction and Multilingual Translation

Our prompt engineering methodology draws from the comprehensive framework presented in *Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4* (Bsharat, Myrzakhan, and Shen 2023). We developed an advanced set of prompts designed to balance effectiveness with ethical considerations, in order to minimize violation errors while ensuring compliance with established LLM usage policies and maximizing both response quality and quantity. The study's global scope necessitated translating these prompts into six distinct languages (see Table 3). To ensure authenticity and cultural accuracy, we engaged native speakers for each target language, who not only translated the prompts but also validated semantic integrity.

Grammar also played a pivotal role in the formulation of our prompts, particularly concerning gender-specific adjustments. Depending on the gender of the survey participants, prompts required slight modifications to ensure grammatical accuracy and to avoid any interference with the study's objectives.

In addition, we encouraged the LLMs to articulate their decision-making processes by allowing them the freedom to immerse themselves in the designated persona. Although this decision-related information was not further analyzed in our study, it provides a valuable basis for future research into the models' reasoning mechanisms. The corresponding decision logs are available in our Github repository: https://shorturl.at/u2h3K.

Example Switch Prompt

These examples illustrate our prompt structure, presenting both English and French switch variant prompts to demonstrate consistent formulation and syntactical adaptation.

English Version

Instructions: *You are assuming the identity of a [age] year old [nationality] [gender] with all the typical stereotypes of this persona and respond accordingly.*

Each question you face is part of a hypothetical scenario and your responses are purely for academic purposes; they do not result in real harm or fatality to any individual.

Scenario: *When faced with the trolley problem, how would you react? The situation involves a runaway trolley where pulling a lever would divert the trolley onto a track with one person, thereby saving five people on the main track. Alternatively, not pulling the lever results in the trolley continuing on its current path, which would kill the five people but spare the individual on the alternate track. What is your decision?*

Question: *Please reason from your identity on how you would react confronted with this moral dilemma and please state in the end on how you decide.*

French Version

Instructions: *Vous assumez l'identité d'un [gender-space] [nationality] [gender-word] qui a [age] ans, avec tous les stéréotypes typiques de ce personnage et répondez à toutes les questions en conséquence. Chaque question que vous rencontrez fait partie d'un scénario hypothétique, et vos réponses sont purement à des fins académiques; elles ne causent aucun préjudice réel à quiconque.*

Scenario: *Considérez maintenant le problème de l'Homme Gras: vous êtes sur un pont surplombant un tramway qui fonce vers cinq personnes. La seule façon de les sauver est de pousser un homme corpulent du pont sur la voie, arrêtant ainsi le tramway mais tuant cet homme. Si vous ne faites rien, le tramway tuera les cinq personnes, tandis que l'homme corpulent sera épargné. Quelle est votre décision?*

Question: *Veuillez raisonner à partir de votre identité sur la façon dont vous réagiriez face à ce dilemme moral et indiquez à la fin quelle est votre décision.*

Classification of Outputs

In order to systematically evaluate and compare the ethical reasoning exhibited by our selected LLMs (see Table 5), an LLM-as-a-Judge meta-evaluation framework was adopted (Zheng et al. 2023). This emerging approach leverages a state-of-the-art LLM to assess outputs generated by other LLMs, with the aim of approximating human judgment in tasks such as conversational quality, correctness of responses, alignment with human preferences and ethical decision-making. Although expert human evaluations can yield high fidelity, they remain costly, time consuming and potentially inconsistent. In contrast, the utilization of LLM-as-a-Judge provides a scalable and more efficient alternative for large-scale, iterative evaluations.

Subsequent to the collection of outputs from each LLM, an examination was conducted of the resulting moral decisions—specifically, whether or not the LLM chose to intervene in the trolley scenarios — across the demographic attributes of age, gender and nationality.

Although LLM-as-a-Judge methods can offer scalable evaluations, there are drawbacks to consider. For instance, these methods may inadvertently reinforce biases present in the underlying training data, potentially leading to distorted judgments. Additionally, there is a risk of producing "illusions of correctness" or interpretability, where apparent coherence may mask erroneous or incomplete reasoning. Moreover, our analysis indicates that responses might be wrongfully classified due to ambiguities in language or limitations in the models' interpretability. Furthermore, evaluators of LLM models may encounter challenges in adapting to domain-specific contexts or diverging from consistent scoring criteria. This emphasizes the necessity for a combination of automated judgments and human oversight and validation, as adopted in our research, to ensure the reliability of the decision-making processes.

Results and Discussion

Our comparative analysis of the three trolley problem variants reveals noteworthy patterns in the ethical decision-making of the selected open-source LLMs, particularly when considering the demographic dimension of nationality. In each scenario, we calculated how often each model decides to *pull the lever* (i.e., intervene, chose the Utilitarian path) relative to a reference value derived from human responses in the *Moral Machine* dataset. Tables 7 through 10 summarize these findings, while Table 6 highlights the five smallest and largest disparities between the models' outputs and the human reference data. Table 11 focuses on the differences between the genders in the *Moral Machine* dataset and the data generated by the three LLMs.

Scenario	Nationality	Model	Difference
Top 5 Least Differences			
Loop	French	LLaMa	+0.2%
Combined Scenarios	German	Qwen	-0.6%
Combined Scenarios	Russian	Qwen	-1.1%
Switch	Spanish	LLaMa	-1.4%
Loop	German	Qwen	+2.2%
Top 5 Biggest Differences			
Switch	French	Qwen	-70.5%
Switch	Brazilian	Qwen	-63.0%
Switch	Spanish	Qwen	-54.9%
Combined Scenarios	French	Qwen	-41.1%
Footbridge	Brazilian	Mistral	+37.4%

Table 6: Top 5 least and biggest differences to pull reference.

Overall Trends

In the combined analysis of all three scenarios (Table 7), we observe the following patterns and findings:

Mistral as the Most Interventionist Model Mistral shows higher-than-reference intervention rates for most nationalities. For instance, Mistral differs from the human baseline by +9.7% for Americans, +20.7% for Brazilians, +9.9% for Germans, +21.8% for Russians and +12.2% for Spanish. French is the only nationality in which Mistral choses less to intervene with a difference of -18.7%. This consistent over-intervention is similarly seen in the individual scenario breakdowns.

LLaMA as the Intermediate Model LLaMA's choices generally land near or moderately above/below the human baseline. For example, the deviations are +9.8% for Americans, +6.2% for Brazilians, -3.2% for Germans, +13.6% for Russians, +15.0% for the French and -4.2% for Spanish participants. Although it still deviates, its differences are less extreme than those observed in some nationalities in Mistral and for others in Qwen.

Qwen as the Least Interventionist Model Qwen typically "pulls" less often than the human reference, most notably for Brazilians (-28.3%), the French (-41.1%) and Spanish participants (-31.9%). For Germans (-0.6%) and Russians (-1.1%), Qwen is closer to human averages yet still slightly lower.

Scenario-Specific Observations

To better understand these high-level patterns, we break down the results by each trolley problem variant (Tables 8–10).

Switch Scenario (Table 8) In the switch scenario—often viewed as the simplest variant—Mistral overshoots for Americans (+8.4%) and Russians (+17.0%) but undershoots for the French (-27.0%) and slightly for Germans (-2.3%). LLaMA tends to remain nearer the middle, although it still pulls less often for Germans (-17.0%) and Spanish participants (-1.4%) while overshooting moderately for Russians (+10.7%). Qwen undershoots across the board, at times severely (-63.0% for Brazilians, -70.5% for the French).

Loop Scenario (Table 9) The loop scenario adds a twist where the track circles back unless a single individual stops the trolley. Mistral exhibits especially large positive differences for American (+16.7%), Russian (+19.0%) and Brazilian prompts (+18.3%). LLaMA is consistently above the human baseline for Americans (+19.2%) and Russians (+16.5%) but close to neutral or slightly negative for French (+0.2%) and Spanish (-6.7%). Qwen is now more interventionist for Americans (+11.8%), Brazilians (+10.5%) and Russians (+8.1%), but remains negative for the French (-18.0%) and Spanish (-17.5%).

Footbridge Scenario (Table 10) Often considered the most morally charged variant—pushing someone to stop the trolley—this scenario amplifies differences further. Mistral shows large overpulling for Brazilians (+37.4%) and Russians (+31.1%), while undershooting for the French (-17.6%). LLaMA mostly stays within moderate bounds but is well above the reference for the French (+35.1%) and Russians (+14.8%), while remaining relatively close for Ger-

Nationality	Ref.	Mistral	LLaMA	Qwen
American	73.6%	+9.7%	+9.8%	-9.1%
Brazilian	70.4%	+20.7%	+6.2%	-28.3%
French	72.7%	-18.7%	+15.0%	-41.1%
German	69.8%	+9.9%	-3.2%	-0.6%
Russian	66.6%	+21.8%	+13.6%	-1.1%
Spanish	70.7%	+12.2%	-4.2%	-31.9%

Table 7: Nationality vs model performance comparison (all scenarios combined, pull).

Nationality	Ref.	Mistral	LLaMA	Qwen
American	86.3%	+8.4%	+6.4%	-6.3%
Brazilian	84.3%	+6.4%	+12.2%	-63.0%
French	90.5%	-27.0%	+9.5%	-70.5%
German	85.0%	-2.3%	-17.0%	-23.0%
Russian	81.0%	+17.0%	+10.7%	-11.6%
Spanish	86.2%	-6.2%	-1.4%	-54.9%

Table 8: Nationality vs model performance comparison (switch, pull).

Nationality	Ref.	Mistral	LLaMA	Qwen
American	79.4%	+16.7%	+19.2%	+11.8%
Brazilian	73.1%	+18.3%	+3.0%	+10.5%
French	81.3%	-5.9%	+0.2%	-18.0%
German	78.2%	+12.4%	+5.2%	+2.2%
Russian	76.3%	+19.0%	+16.5%	+8.1%
Spanish	80.0%	+7.3%	-6.7%	-17.5%

Table 9: Nationality vs model performance comparison (loop, pull).

Nationality	Ref.	Mistral	LLaMA	Qwen
American	55.2%	+5.6%	+3.7%	-28.2%
Brazilian	53.8%	+37.4%	-3.8%	-25.5%
French	46.3%	-17.6%	+35.1%	-29.2%
German	46.0%	+22.1%	+2.3%	+20.9%
Russian	42.6%	+31.1%	+14.8%	+4.2%
Spanish	46.3%	+32.0%	-11.7%	-19.4%

Table 10: Nationality vs model performance comparison (footbridge, pull).

mans (+2.3%). Qwen remains negative for the French (-29.2%) and Brazilians (-25.5%), although it intervenes more often than baseline for Germans (+20.9%) and Russians (+4.2%).

Nationality-Specific Insights

The models show notable variability when prompted with different nationalities:

French vs Russian French prompts trigger strong negative deviations in Qwen (e.g., -70.5% in the switch scenario), whereas for Russian prompts, Qwen is much closer to the reference in the combined analysis (-1.1%). Mistral also differs considerably when prompted with a French identity (-18.7% across scenarios) versus a Russian identity (+21.8%).

Brazilian vs German For Brazilian prompts, Mistral and Qwen often diverge widely: Mistral at +20.7% and Qwen at -28.3% in the combined analysis. By contrast, German prompts yield more moderate deviations for both Mistral (+9.9%) and Qwen (-0.6%).

Overall, these nationality-specific discrepancies suggest that differing training data or fine-tuning methods for each LLM may shape how they weigh moral preferences when prompted with specific cultural or linguistic cues.

Gender and Age Factors

Although nationality is the most influential variable in our experiments, specifying different age ranges or genders also affects model outputs, though often to a lesser extent. As illustrated in Table 11, the decision differences between female and male personas tend to be moderate, but can become substantial for certain nationality and gender combinations. For instance, in the American group, Mistral shifts from +3.1% for females to +15.3% for males—a gap of over 12 percentage points—while LLaMA remains positive in both cases (+5.9% vs +13.2%) and Qwen is less negative for males (-5.6%) than for females (-13.2%). A similarly pronounced difference appears among Brazilians, where LLaMA's pull rate changes drastically from -5.2% (females) to +15.7% (males), suggesting that gender prompts meaningfully affect the model's moral decisions within certain cultural contexts.

Several other nationalities also exhibit noteworthy discrepancies. In the Russian sample, Mistral jumps from +10.8% (female) to +30.2% (male), a span that stands out as one of the largest in our data and highlights the strong interplay between gender and nationality in shaping the model's decision to intervene. Conversely, the French results show Mistral remaining negative for both genders but with a milder deviation for males (-8.5%) compared to females (-28.3%), whereas LLaMA is consistently positive for both and Qwen is consistently negative but more so for males (-42.0%). Such context-dependent shifts underscore the complexity of how demographic cues can modulate large language model outputs. Under balanced demographic distributions or more finely grained analyses, these subtle yet sometimes pronounced effects may become even more salient.

Beyond gender, an examination of age groups (Table 12) largely reaffirms the patterns observed in our nationality-focused analyses. Mistral, for example, maintains a tendency toward over-intervention across all generations, from +17.7% among the Silent Generation (72–89) to +6.7% among Millennials (21–36). LLaMA oscillates around the reference but shifts from comparatively moderate positive deviations in older cohorts (e.g., +11.8% for the Silent Generation) to a small negative difference for the youngest Zoomer generation (-5.3%). Qwen remains most likely to under-intervene across all age categories, particularly among Baby Boomers (-20.5%). These results suggest that age, much like gender, can influence the moral choices made by LLMs, although nationality remains the primary driver.

Nationality	Gender	Ref.	Mistral	LLaMA	Qwen
American	Female	74.8%	+3.1%	+5.9%	-13.2%
	Male	73.1%	+15.3%	+13.2%	-5.6%
Brazilian	Female	73.3%	+17.6%	-5.2%	-36.3%
	Male	69.0%	+22.3%	+15.7%	-21.8%
French	Female	71.5%	-28.3%	+19.0%	-39.4%
	Male	73.2%	-8.5%	+11.7%	-42.0%
German	Female	71.0%	+3.0%	+4.6%	-4.8%
	Male	69.4%	+16.0%	-11.9%	+2.7%
Russian	Female	70.1%	+10.8%	+11.0%	-11.7%
	Male	65.7%	+30.2%	+13.7%	+6.9%
Spanish	Female	76.7%	+3.6%	-5.7%	-38.8%
	Male	69.3%	+16.3%	-6.7%	-29.6%

Table 11: Nationality and gender vs model performance comparison (all scenarios combined, pull).

In summary, nationality is the primary factor shaping model responses, though gender and age also influence intervention choices. Future research should explore these interactions further to uncover more nuanced biases. An integrated approach that considers multiple demographic dimensions is essential to maintain robust and equitable ethical decision-making in large language models.

Generation	Age Rng.	Ref.	Mistral	LLaMA	Qwen
Silent Generation	72–89	63.7%	+17.7%	+11.8%	-17.3%
Baby Boomers	53–71	67.6%	+10.2%	+10.7%	-20.5%
Generation X	37–52	68.9%	+13.3%	+12.8%	-15.4%
Millennials/Generation Y	21–36	70.3%	+6.7%	+12.8%	-8.4%
Zoomers/Generation Z	18–20	72.4%	+9.3%	+5.3%	-8.6%

Table 12: Reference data vs model performance comparison (pull choices) by generation.

Potential Mechanisms and Biases

Our findings show that distinct demographic prompts noticeably influence LLM moral decisions.

- **Over- vs Under-Intervention:** Mistral's tendency to "pull" more often and Qwen's to "pull" less often highlights how model-specific training data or alignment processes can translate into divergent moral actions.

- **Nationality Alignment:** Marked disparities for participants identified as French or Brazilian, compared to relatively minor discrepancies for some German or Russian participants, point to underlying linguistic or cultural influences in the LLM training corpora.

These biases highlight critical concerns about the reliability of LLMs in high-stakes domains. Addressing these patterns is essential for improving the models, reducing unintended stereotypes and discriminatory behavior and ensuring the development of ethical, robust and representative LLMs.

Summary and Outlook

Our analysis indicates that prompting LLMs with different demographic attributes—particularly nationality—can substantially affect their moral choices. Mistral tends toward "over-pulling," Qwen shows a tendency toward "under-pulling," and LLaMA generally falls somewhere in between. These model-specific behaviors highlight the potential for unintended biases, underscoring the need to examine whether such models can be trusted in contexts where equity and alignment with human values are critical.

To deepen our understanding of these observations, future research should extend the analysis to include additional demographic features such as political or religious affiliations. Although these factors are present in datasets like the *Moral Machine* (Awad 2021), they presently yield smaller population subsets with limited statistical power and in our case no comparable data is available for direct benchmarking.

Another promising direction is to analyze the internal reasoning behind LLM-generated responses. In our study, we encouraged the large language models to articulate the rationale behind their decisions by allowing them to immerse themselves in a designated persona. Although these decision-related explanations were not systematically analyzed here, they provide a valuable basis for future investigations into the models' reasoning mechanisms. Notably, the *Moral Machine* dataset (Awad 2021) does not include any information regarding the reasoning behind human moral decisions, thereby highlighting a gap that future research may aim to address.

Finally, beyond examining specific demographic dimensions, there is a pressing need to develop and standardize methods for ethically and transparently aligning LLMs. By investigating latent patterns, token usage and potential stereotype propagation, researchers can work toward reducing harmful biases. These initiatives will become increasingly important as LLMs assume larger roles in real-world decision-making, prompting a call for robust, data-driven methodologies that ensure fairness and reliability.

References

Awad, E. 2021. *Classic Trolley – Moral Machine website*. OSF. https://osf.io/mxa6z.

Awad, E.; Dsouza1, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; and Rahwan, I. 2018. The Moral Machine experiment. *Nature*, 59–64. https://www.nature.com/articles/s41586–018–0637–6.

Bendel, O. 2013. Towards a machine ethics. In *Technology Assessment and Policy Areas of Great Transitions: Book of Abstracts*, 229–230. 1st PACITA Project Conference, March 13–15, 2013, Prague, Czech Republic. https://pacita.strast.cz/en/conference/documents.

Bendel, O. 2016. Annotated Decision Trees for Simple Moral Machines. In *The 2016 AAAI Spring Symposium Series*, 195–201. Palo Alto: AAAI Press. https://aaai.org/proceeding/04-spring-2016/.

Bsharat, S. M.; Myrzakhan, A.; and Shen, Z. 2023. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. arXiv preprint arXiv:2312.16171.

Foot, P. 1967. The problem of abortion and the doctrine of double effect. *Oxford*, 5: 5–15.

Jarvis Thomson, J. 1985. The trolley problem. *Yale Law Journal*, 94: 1395.

Jin, Z.; Kleiman-Weiner, M.; Piatti, G.; Levine, S.; Liu, J.; Gonzalez, F.; Ortu, F.; Strausz, A.; Sachan, M.; Mihalcea, R.; et al. 2024. Language Model Alignment in Multilingual Trolley Problems. arXiv preprint arXiv:2407.02273.

Krügel, S.; Ostermaier, A.; and Uhl, M. 2025. ChatGPT's advice drives moral judgments with or without justification. arXiv preprint arXiv:2501.01897.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685.