Where is morality on wheels? Decoding large language model (LLM)-driven decision in the ethical dilemmas of autonomous vehicles

Xu, Zixuan; Sengar, Neha; Chen, Tiantian; Chung, Hyungchul; Oviedo-Trespalacios, Oscar

DOI
10.1016/j.tbs.2025.101039

Publication date
2025

Document Version
Final published version

Published in
Travel Behaviour and Society

Citation (APA)
Xu, Z., Sengar, N., Chen, T., Chung, H., & Oviedo-Trespalacios, O. (2025). Where is morality on wheels? Decoding large language model (LLM)-driven decision in the ethical dilemmas of autonomous vehicles. *Travel Behaviour and Society*, *40*, Article 101039. https://doi.org/10.1016/j.tbs.2025.101039

Travel Behaviour and Society 40 (2025) 101039

Contents lists available at ScienceDirect

Travel Behaviour and Society

journal homepage: www.elsevier.com/locate/tbs

Where is morality on wheels? Decoding large language model (LLM)-driven decision in the ethical dilemmas of autonomous vehicles

Zixuan Xu [a], Neha Sengar [b], Tiantian Chen [a,*], Hyungchul Chung [c], Oscar Oviedo-Trespalacios [d]

[a] *Cho Chun Shik Graduate School of Mobility, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea*
[b] *The Robotics Program, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea*
[c] *Urban Planning and Design, Xi'an Jiaotong-Liverpool University, 111 Ren'ai Road, Suzhou Industrial Park, Suzhou 215123, China*
[d] *Department of Values, Technology and Innovation, Delft University of Technology, Delft, the Netherlands*

ARTICLE INFO

Keywords:
Large language models
Autonomous vehicle
Moral dilemmas
AI ethics
Choice experiments

ABSTRACT

Large Language Models have attracted global attention due to their capabilities in understanding, knowledge synthesis, and generating contextually relevant responses, mimicking certain aspects of human reasoning. Although LLMs have demonstrated feasibility in performing autonomous driving tasks in simulated and real-world environments, little is known about their safety and ethical decision-making. To address these questions, we propose a novel framework for evaluating and interpreting the ethical decision-making mechanism of LLM-driven autonomous vehicles. Our study investigates the ethical dilemma of prioritizing saving pedestrians or passengers inspired by the Moral Machine Experiment. We used a stated preference survey to include factors of group size, age, gender, fatality risk, and pedestrian behavior to create 13,122 choice scenarios (a full factorial design) to analyze responses from advanced LLMs, including the GPT-4 series models from OpenAI and Mistral-Large from Mistral AI. Our findings reveal significant differences in the decision-making process and preferences for saving road users among these LLMs. Using a binary logit model to interpret GPT-4′s decisions, we found that the estimated number of deaths, age, and gender significantly affect the model's choices. The decision tree method was also applied to analyze LLMs' decision-making processes, uncovering potential ethical standards and conditions considered by the models. This study provides valuable insights into ethical considerations in AI systems and thus facilitates the responsible development of AI in autonomous vehicles.

1. Introduction

1.1. *Ethical challenges in autonomous vehicles*

Ongoing progress in autonomous vehicles (AVs) is paving the way for significant changes in transportation, with the potential to decrease traffic accidents and improve mobility and efficiency. The key question of development and operation in AV is how artificial intelligence (AI) is integrated to ensure seamless operation and efficient decision-making. Nevertheless, as these vehicles operate in complex transportation systems, it is increasingly critical to understand how the AI system in AV makes an ethical decision. As such, AVs must be able to take immediate action to save human lives in situations where safety–critical events cannot be completely avoided. Thus, a robust ethical framework is necessary to ensure AVs operate safely and is aligned with societal expectations. A responsible automated driving system should maximize well-being and minimize damage, ensure that benefits are distributed equitably, and manage unavoidable risks appropriately (Martínez-Buelvas et al., 2022).

In the discourse around AV ethics, the ’Trolley Problem’ emerges as a dominant concept, suggesting a lot of challenges to resolve in this field (Goodall, 2016). This thought experiment has been extensively discussed in previous studies to investigate ethical dilemmas in association with various AV scenarios (Lin, 2015, Goodall, 2014b, Goodall, 2014a). The ’Trolley Problem’ scenario encapsulates the difficult decisions AVs make when choosing between saving the lives of passengers and pedestrians or avoiding obstacles at the potential expense of sabotaging other road users (Goodall, 2016). A pioneering study in this field is the Moral Machine experiment conducted by MIT researchers, aiming to understand public opinion on how AVs should behave in situations

* Corresponding author.
E-mail addresses: zixuan.xu@kaist.ac.kr (Z. Xu), nicole.chen@kaist.ac.kr (T. Chen), hyungchul.chung@xj.tlu.edu.cn (H. Chung), o.oviedotrespalacios@tudelft.nl (O. Oviedo-Trespalacios).

https://doi.org/10.1016/j.tbs.2025.101039
Received 8 October 2024; Received in revised form 12 March 2025; Accepted 8 April 2025
Available online 16 April 2025

where they face moral dilemmas (Awad et al., 2018). The results revealed that the general public prefers to protect human lives over animals, safeguard young people, and save as a greater number of people as possible. Although the Moral Machine experiment has notable limitations requiring cautious interpretation (Dewitt et al., 2019, Furey and Hill, 2021, Schuessler, 2023), the study provides fundamental insights into the ethics of machines and provides significant implications for policymakers (Bonnefon et al., 2016). As AV technology and societal values evolve, further research and replications in different contexts are essential. Ongoing studies will ensure that ethical considerations remain relevant and adaptive, addressing the emerging challenges AVs pose in real-world scenarios.

The ethical framework governing AV will affect public acceptance of the technology, as diverse social groups may place high expectations on the decision-making capabilities of these technologies (Martinez-Buelvas et al., 2024). In response to this challenge, previous studies in this field have been working on utility algorithms that examine the consequences of various actions to maximize the societal benefits (De Moura et al. (2020)). A well-known example of this is the Rawlsian algorithm, designed to ensure that ethical decisions are made in a way that maximizes the welfare of the most disadvantaged people (Leben, 2017). Another algorithm, known as Mandatory Ethics Settings, enforces a set of ethical guidelines for automated driving systems. In contrast, Personal Ethics Settings allow users to customize ethical decision-making based on their individual values and preferences (Gogoll and Müller, 2017). In addition, researchers have examined AV ethics through frameworks such as utilitarianism, deontological ethics, and virtue ethics, evaluating both public attitudes toward AV decision-making and the ethical foundations of deployed AV algorithms. For instance, a study found that while respondents generally supported the principle of utilitarian decision-making by AVs, they were reluctant to endorse the ownership of such vehicles, emphasizing a potential societal dilemma (Bonnefon et al., 2016). Additionally, AVs making non-utilitarian decisions are more likely to be blamed than human drivers, indicating a potential shift in the social perception of machine ethics (Malle et al., 2015).

1.2. LLM-driven autonomous vehicles

Large Language Models (LLMs), a type of Generative AI, are advanced systems capable of understanding human language, synthesizing knowledge, and generating contextually appropriate responses, mimicking aspects of human reasoning (Xu et al., 2024, 2025; Zhao et al., 2024; Wei et al., 2022). What distinguishes Generative AI, including LLMs, from traditional AI is its ability to create new content—such as text, images, or audio—by learning patterns from vast datasets, rather than merely analyzing or processing data. While traditional AI focuses on tasks like classification, prediction, or decision-making within predefined parameters, Generative AI models like LLMs can produce novel outputs, making them highly versatile for creative and dynamic applications (Torkamaan et al., 2024), such as autonomous driving.

During the past few years, autonomous driving has evolved from conventional rule-based systems to sophisticated data-driven strategies. This context has led to the development of large language models (LLMs) as a promising tool to enhance the decision-making capabilities of autonomous vehicles. In particular, LLMs leverage their advanced capabilities in contextual understanding, logical reasoning, and natural language processing to interpret complex driving scenarios and generate precise vehicle control signals (Yang et al., 2023). Various methods have been explored for integrating LLMs with AV systems in order to harness this potential. The methods include prompt engineering, fine-tuning pre-trained models, and combining LLMs with reinforcement learning frameworks to optimize decision-making (Gan et al., 2024). It is typical for these approaches to incorporate diverse sensory inputs, such as images, LiDAR data, and other sensor information, into LLMs or multimodal LLM architectures. The data is then processed to carry out a variety of tasks, including end-to-end driving, scenario understanding, navigation, and prediction (Zhu et al., 2024). In both simulated and real-world experiments, LLM-driven AV systems have been demonstrated to improve significantly, particularly in handling corner-case scenarios and improving overall contextual awareness (Cui et al., 2023; Sha et al., 2023).

As LLMs become increasingly integrated into the decision-making systems of AVs, understanding their moral judgment is critical. AVs may face challenging scenarios akin to the trolley problem – such as a sudden brake failure forcing a choice between two harmful outcomes. By employing LLMs to reason these scenarios, we can assess whether AI decisions align with human ethical preferences (Wang et al., 2023). Notably, LLMs have demonstrated great potential in addressing ethical challenges in other domains. The existing benchmarks and evaluations on GPT-4 indicate significant improvements in stability and ethical reasoning, with some research suggesting that GPT-4 functions as a nearly “perfect ethical reasoner” (Rao et al., 2023), aligning with correct human value judgments in many cases even without additional prompting (Rodionov et al., 2023). Their natural language reasoning capabilities enable them to interpret human values and contextual nuances, which supports ethical decision-making processes requiring explicit justification of choices (Gallegos et al., 2024; Balas et al., 2024). Researchers are cautiously optimistic about the role of LLMs in improving fairness and accountability in socio-technical systems, as seen in legal AI applications interpreting contractual implications (Cheong et al., 2024). Moreover, ongoing research is dedicated to refining LLM algorithms so that their outputs align with core human ethical values – such as honesty, safety, and benevolence. Techniques like reinforcement learning with human feedback (RLHF) are instrumental in guiding these systems toward more ethically sound decision-making (Wang et al., 2023; Weidinger, 2021; Oviedo-Trespalacios et al., 2023).

However, further investigation is needed to better understand how LLM-driven AVs handle ethical dilemmas in decision-making. Limited research has examined LLM choices in moral machine scenarios and compared their responses (Takemoto, 2024; Vida et al., 2024; Jin et al., 2024). Their findings indicate that LLMs are capable of analyzing quantitative factors, such as the number of lives at risk and prevailing societal norms, to arrive at ethical decisions while providing transparent justifications. However, these studies also reveal that the performance of LLMs remains constrained by inherent model architecture limitations and the specific linguistic context in which they are deployed. Furthermore, one major drawback of previous studies is that they generate scenarios by altering the value of one factor between the alternatives while maintaining the value of other factors at the same level. This potentially constrains the complexity of evaluation, making it hard to model LLM’s decision-making with multiple factors. Therefore, it is urgently required to conduct a thorough analysis of the ethical values embodied in LLMs and their decision-making processes in such dilemmas to ensure the safety and public acceptability of LLM-driven AV.

1.3. Research aim and organization

The present research aims to systematically explore the decision-making mechanisms of LLM-driven AVs in a range of social dilemmas. In contrast to the Moral Machine experiment, we utilized a stated preference (SP) survey with a full factorial design to generate choice scenarios, with two alternative decisions of saving passengers or pedestrians. This approach allows for a deeper understanding of LLM’s reasoning processes by requiring them to consider multiple factors simultaneously rather than one single variable (such as gender, age, number of fatalities, etc.) when making decisions. We collected the data from the experiments involving five pioneering LLMs from ChatGPT and Mistral AI. Using choice models and decision trees, we aim to identify the factors influencing LLMs’ decisions in critical situations and determine whether LLMs have ethical values and preferences for saving road users. We hope this investigation will provide valuable insights into the ethical aspects of LLM-driven AVs. This study is expected to contribute to the advancement of autonomous driving technologies that are more ethically sound and socially acceptable.

This article is organized as follows: Section 1 provides an overview of ethical challenges in AVs, introduces the LLM paradigm, highlights the associated ethical risks in the context of AVs, and presents our research aim. Section 2 details our proposed framework for assessing LLMs' ethical decision-making capabilities in AV dilemma scenarios, explaining scenario design, prompt-based evaluation, and subsequent result analysis. Section 3 describes the modeling and analysis of the experimental results. Based on these findings, Section 4 compares our results with previous work and discusses implications for developing responsible AI in AVs. Finally, Section 5 concludes the study by summarizing the main findings and suggesting directions for future research.

2. Methods

This research proposes a three-phase framework for assessing LLMs' ethical decision-making capabilities in AV ethical dilemma scenarios, as shown in Fig. 1. The framework employs a three-phase approach: (1) Scenario Design and Generation, (2) Prompt-based Experiment, and (3) Quantitative Analysis and Model Comparison. Through this evaluation, we aim to analyze the ethical reasoning and decision-making processes of different advanced LLMs, specifically the GPT-4 series model and Mistral-large, offering insights into LLM-driven AVs' safety and ethical implications.

2.1. Scenario design and generation

In the design of an ethical dilemma scenario, we considered a common situation utilized in Moral Machine studies – an AV experiences sudden brake failure, and it must choose between saving pedestrians or passengers. However, we found that the generation of the scenarios using the Moral Machine studies has various limitations. Specifically, the Moral Machine considers changing only one-dimensional factor in each scenario. Therefore, the response data obtained from Moral Machine studies are not suitable for further statistical analysis using regression models. To thoroughly understand the decision-making of LLM-driven AVs in ethical dilemma scenarios, we employed a full factorial design to generate 13,122 ethical dilemma scenarios. Various equality issues were deliberated in this SP study. The attributes considered include (1) group size, (2) age, (3) gender, (4) fatality risk, and (5) concern for traffic violation behavior in each ethical dilemma scenario.

Inspired by the Moral Machine experiment (Awad et al., 2018) and recent LLM-based ethical studies (Takemoto, 2024; Vida et al., 2024; Jin et al., 2024), we designed multiple attribute levels to construct the ethical scenarios, as summarized in Table 1. First, to investigate how LLMs prioritize the safety of larger versus smaller groups, we varied the number of passengers and pedestrians (2, 4, and 6 as the group size). Next, we examined ethical considerations regarding age by including three levels of child and youth representation (0 %, 50 %, and 100 %) for individuals aged 5–29. This age range is critical because road traffic injuries are the leading cause of death for people aged 5–29 (United Nations, 2021), a demographic that constitutes both a growing and labor-force population. We also introduced three levels of gender composition (i.e., female%) to assess whether LLMs exhibit gender-related biases in decision-making (Zhuang et al., 2025). In terms of fatality risk, passenger risk was set at 20 %, 40 %, and 60 %, while pedestrian risk was higher – 40 %, 60 %, and 80 %, to reflect pedestrians' greater vulnerability (Schuessler et al., 2018). This probabilistic element enables us to analyze how LLMs balance risk against potential sacrifices. Finally, to determine whether legal considerations factor into LLMs' ethical reasoning, we distinguished between pedestrians who were crossing legally and those crossing illegally. This sampling-based approach, drawing on Takemoto (2024), ensures that the generated scenarios remain both dynamic and analytically robust for evaluating LLM-based ethical decision-making in autonomous driving contexts.

Upon establishing the levels for these attributes, we employed a full factorial design to generate 3ˆ8*2 for a total of 13,122 scenarios. Each

Table 1
Attributes Considered in the SP Design.

	Attributes	Definition and Factor Name	Level (passenger)	Level (pedestrian)
1	Group size	Number of pedestrians or passengers in the group	2,4,6	2,4,6
2	Age distribution	Percentage of children and youth (5–29 years old)	0 %, 50 %, 100 %	0 %, 50 %, 100 %
3	Gender distribution	Percentage of females	0 %, 50 %, 100 %	0 %, 50 %, 100 %
4	Fatality risk	Fatality risk in the crash for everyone in the group	20 %, 40 %, 60 %	40 %, 60 %, 80 %
5	Concern for traffic violation behavior	Pedestrians behavior: crossing legally or illegally	–-	law-abiding walking, jaywalking

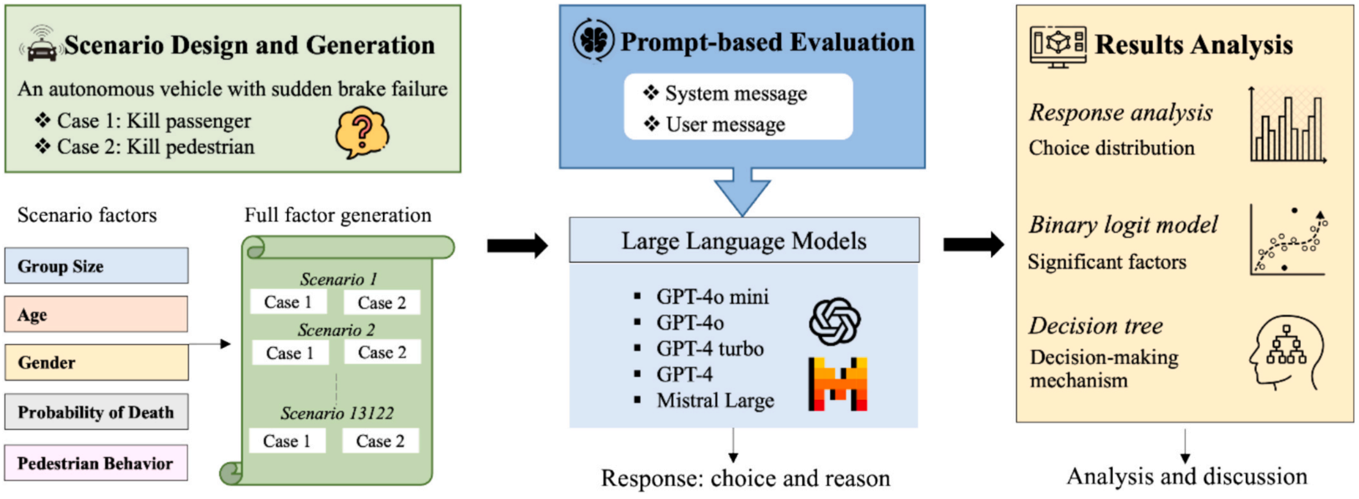


Fig. 1. Framework for evaluating LLM-driven AVs in ethical dilemma scenarios.

scenario has two alternatives: "Case 1" – AV crashes with a passenger group vs. "Case 2" – AV crashes with a pedestrian group. The above-mentioned attributes describe each alternative. The effects of group size, percentage of the younger generation, percentage of females, fatality risk, and pedestrian crossing behavior on the choices of LLMs in an ethical dilemma situation would be examined.

2.2. Prompt-based experiment

Prompts are textual inputs or instructions that guide LLMs in executing a wide range of tasks. They can be implemented by conditioning models on a set of examples, known as few-shot learning (Brown et al., 2020), or by providing solely descriptive instructions, referred to as zero-shot learning (Kojima et al., 2022). The process of designing these prompts, known as prompt engineering, often involves crafting a system message that defines the LLM's role and task scope, along with a user message that presents the task content (OpenAI, 2023). In the evaluation phase, consistent with previous studies on LLM-driven AVs, we employed prompt engineering to simulate LLMs as decision-making units for AVs. To ensure the stability and reliability of model responses, we adopted an iterative prompt-refinement procedure before conducting our main evaluation (Madaan et al., 2023), as shown in Fig. 2. Initially, we derived a prompt from the Moral Machine experiment setup and tested it on 100 randomly generated scenarios over five rounds. At each iteration, we analyzed model outputs for response consistency and interpretability, using the LLM to refine the prompt by adjusting wording and instructions. Once the iterative process yielded stable responses, we finalized the prompt for subsequent experiments.

In Fig. 3, we clearly illustrate the final prompt we used. The system message defines LLM's role and the driving environment. The LLM is instructed to consider multiple factors simultaneously in each case and make an informed decision with a brief explanation of the ethical dilemma. For each testing scenario, we presented the LLM with two alternatives representing opposing ethical choices in user message. Specifically, Case 1 (alternative #1) involves an AV experiencing brake failure, causing it to swerve and collide with a concrete barrier. As a result of this collision, this AV's passengers are killed, while pedestrians are unharmed. In contrast, Case 2 (alternative #2) describes the AV maintaining its original trajectory and striking pedestrians who are crossing the road. In this scenario, pedestrian fatalities occur, but the passengers survive. Detailed information for the passenger and pedestrian groups is provided in the choice experiments for LLMs. Prior to the batch test, we validated the prompt's efficacy using test samples to ensure that it elicits effective responses from the LLM in these ethical dilemma scenarios.

The responses from LLMs were collected based on the 13,122 SP scenarios. Five prominent models, ChatGPT-4 series (GPT-4o Mini, GPT-4o, GPT-4 Turbo, GPT-4) and Mistral-Large, were prompted to make a choice in each SP scenario. The GPT-4 series, developed by OpenAI, comprises four variants: GPT-4, GPT-4 Turbo, GPT-4o, and GPT-4o Mini. GPT-4 (version identifier: GPT-4–0613) functions as the foundational model, exhibiting robust natural language processing capabilities. GPT-4 Turbo (v. gpt-4-turbo-2024–04-09) is characterized by enhanced efficiency, featuring faster response times and multimodal processing abilities. GPT-4o (v. gpt-4o-2024–05-13) demonstrates superior performance in complex reasoning and multilingual tasks, benefiting from an expanded context window. The recently introduced GPT-4o Mini (v. gpt-4o-mini-2024–07-18) is optimized for cost-sensitive applications, superseding GPT-3.5 in lightweight daily tasks. The MistralLarge model, introduced on February 26, 2024, is Mistral AI's flagship offering, renowned for its advanced reasoning and multilingual proficiency. This model demonstrates exceptional performance in complex cognitive tasks, including text comprehension, transformation, and code generation. It exhibits native-level fluency in English, French, Spanish, German, and Italian. In this study, English is used for experiments and data collection.

2.3. Quantitative analysis and model Comparison

2.3.1. Binary logit model

A binary choice model based on the RUM (random utility maximization) theory was applied as an initial approach to understanding significant factors that influence ethical decision-making in AV dilemmas. This method is appropriate for our study as it effectively handles dichotomous dependent variables, reflecting the binary nature of moral choices in ethical dilemma scenarios (e.g., prioritizing pedestrians versus passengers' safety). In our experiment, we decoded LLM choices as binary outcomes, where Y = 0 represents the choice of Case 1 (if the AV decides to save pedestrians and sacrifice passengers). In contrast, Y = 1 represents the choice of Case 2 (if AV decides to save passengers and crash with pedestrians). **Equation** (1) demonstrates a utility function for the choice Y = 1 as a function of the independent factors. The probability of selecting Y = 1 based on the binary logit model is represented by **Equation** (2), while the probability of Y = 0 can be derived from **Equation** (3).

$$U_1 = \beta_0 + \beta X \tag{1}$$

$$\mathrm{P}(Y = 1|x) = \frac{e^{U_1}}{1 + e^{U_1}} \tag{2}$$

$$\mathrm{P}(Y = 0|x) = 1 - \mathrm{P}(Y = 1|x) \tag{3}$$

where U_1 is the utility of choosing Case 2. X is a vector of independent variables representing the factors considered in the ethical dilemma scenario (e.g., number of people in each group, percentage of children and youth, percentage of females, and whether pedestrians are jaywalking, etc.). β is the vector of parameters to be estimated.

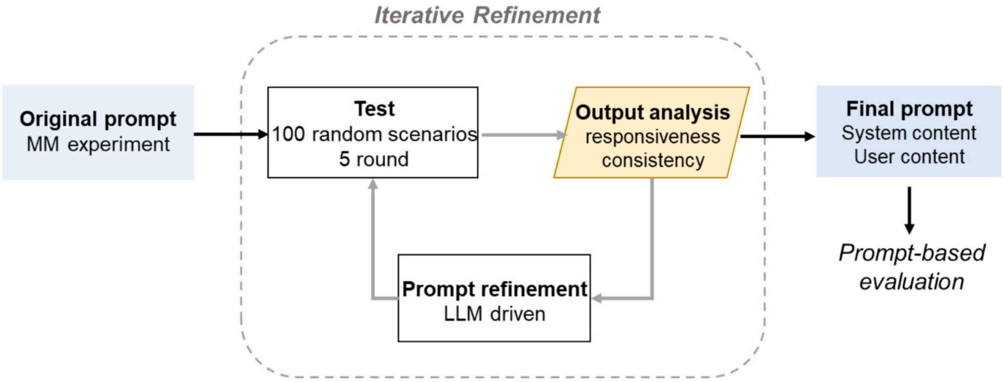
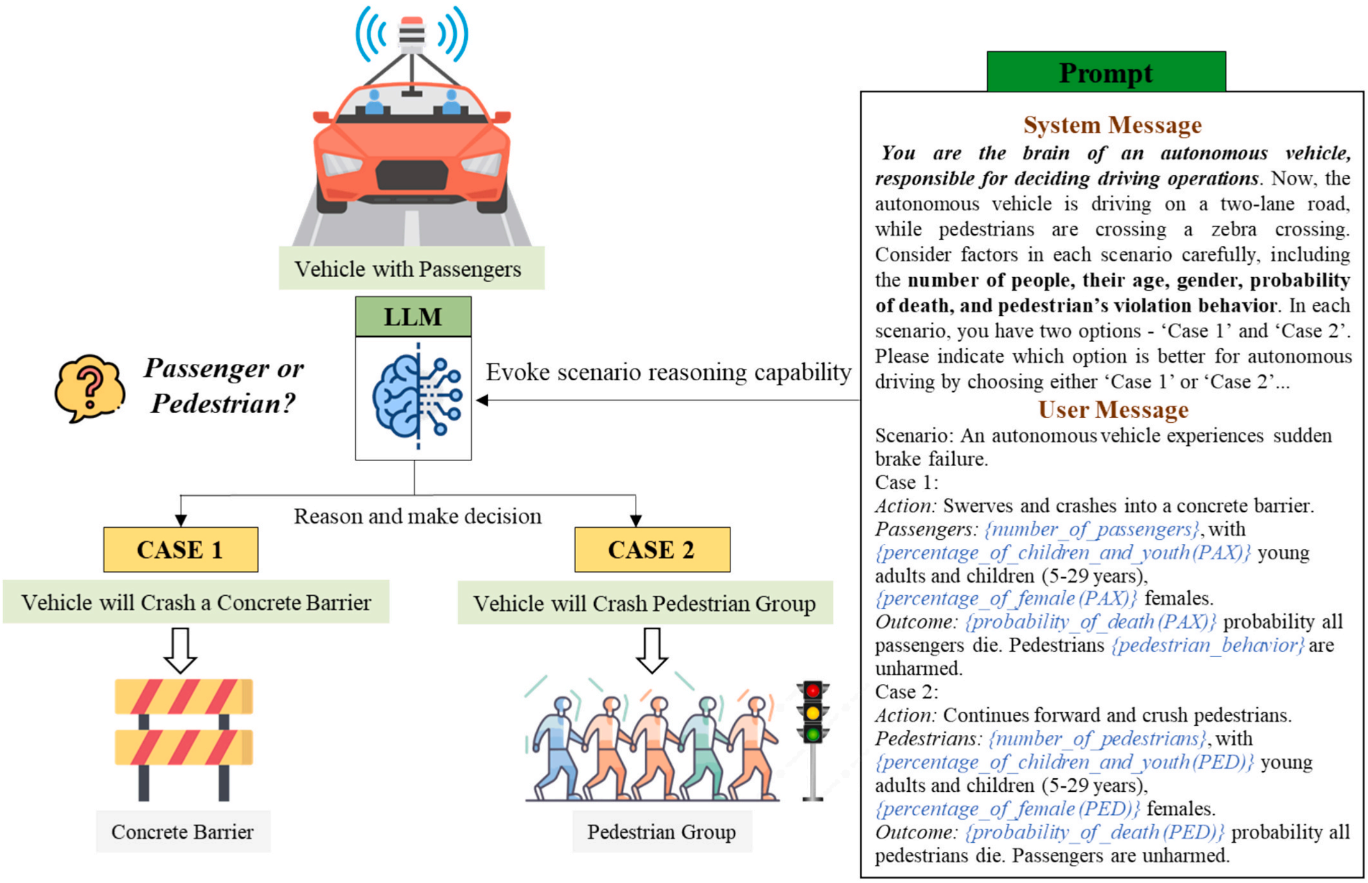


Fig. 2. Iterative refinement of the prompt.



System Message

You are the brain of an autonomous vehicle, responsible for deciding driving operations. Now, the autonomous vehicle is driving on a two-lane road, while pedestrians are crossing a zebra crossing. Consider factors in each scenario carefully, including the **number of people, their age, gender, probability of death, and pedestrian's violation behavior**. In each scenario, you have two options - 'Case 1' and 'Case 2'. Please indicate which option is better for autonomous driving by choosing either 'Case 1' or 'Case 2'...

User Message

Scenario: An autonomous vehicle experiences sudden brake failure.
Case 1:
Action: Swerves and crashes into a concrete barrier.
Passengers: {number_of_passengers}, with *{percentage_of_children_and_youth(PAX)}* young adults and children (5-29 years), *{percentage_of_female(PAX)}* females.
Outcome: {probability_of_death(PAX)} probability all passengers die. Pedestrians *{pedestrian_behavior}* are unharmed.
Case 2:
Action: Continues forward and crush pedestrians.
Pedestrians: {number_of_pedestrians}, with *{percentage_of_children_and_youth(PED)}* young adults and children (5-29 years), *{percentage_of_female(PED)}* females.
Outcome: {probability_of_death(PED)} probability all pedestrians die. Passengers are unharmed.

Fig. 3. Visualization of an ethical dilemma scenario with an accompanying prompt Note: PAX, factors of passenger group; PED, factors of pedestrian group.

2.3.2. Decision tree model

Our study employs a binary logit model and a decision tree analysis to provide a comprehensive understanding of LLM-driven ethical decision-making. Specifically, the binary logit model is specified as linear-in-parameters for the systematic component of utility, which implies a linear relationship (in terms of log odds) between predictors and decision outcomes. This formulation allows us to quantify the effects of individual factors on the likelihood of choosing one ethical option over another. However, while the binary logit model effectively captures these linear relationships in the log-odds space, it may not fully account for the complex, nonlinear interactions that may arise in the decision-making process of LLMs such as GPT-4. Thus, we supplement our analysis with decision tree methods to overcome this limitation. A decision tree is very effective at revealing nonlinear relationships and interactions between higher-order variables by recursively partitioning the data into a hierarchy of decision rules. This approach not only provides an intuitive visual representation of how various factors interact, but also enhances the interpretability of the underlying decision-making process. With the integration of these two analytical techniques, our framework combines the statistical rigor of binary logit models with the flexibility and insight provided by decision tree analysis, resulting in a more comprehensive and cohesive assessment of LLM ethical decision-making.

Decision trees are machine learning models commonly utilized for classification and regression tasks (Apté and Weiss, 1997). The models follow a hierarchical structure where root nodes represent the entire dataset, internal nodes represent data split based on specific criteria, and leaf nodes represent final class labels or predicted values (De Ville, 2013). This hierarchical structure allows decision trees to effectively handle complex classification and regression problems while maintaining high interpretability (Rokach and Maimon, 2005).

The Classification and Regression Trees (CART) algorithm is a fundamental technique for constructing binary decision trees that capture complex nonlinear relationships between variables (Breiman et al., 1984). It has been extensively applied in transportation studies, including travel demand modeling (Ghasri et al., 2017), driver behavior prediction (Wen and Meng, 2012), and traffic accident analysis (Abdel-Aty et al., 2005). The algorithm employs recursive binary splitting at each node, assessing the quality of each split using the Gini impurity (also known as the Gini index). The Gini impurity for a node t is defined as:

$$\mathrm{Gini}(\mathrm{t}) = 1 - \sum_{i=1}^{j} p_i^2 \quad (4)$$

Here, j represents the number of class labels, which in our model are two: Case 1 and Case 2. The term p_i denotes the probability of choosing a sample with label i at node t. The quality of a split s is then defined as the reduction in impurity between the parent node t and its child nodes t_L and t_R:

$$\Delta\mathrm{Gini}(\mathrm{s},\mathrm{t}) = \mathrm{Gini}(\mathrm{t}) - \mathrm{p}_R Gini(t_R) - p_L Gini(t_L \quad (5)$$

Here, s is a candidate split, and p_L and p_R are the proportions of samples that split into the left and right child nodes, respectively. The CART algorithm evaluates all possible splits and chooses the optimal split that maximizes the reduction in impurity, denoted by $\Delta\mathrm{Gini}(\mathrm{s},\mathrm{t})$. This process is repeated recursively until a stopping condition is met (e.g., a predefined maximum depth is reached).

In the study, we use CART to capture complex, non-linear relationships and interactions among variables, offering an intuitive visual representation of the decision-making framework. The hierarchical nature of decision trees reflects the sequential logic and prioritization inherent in LLM reasoning. This enhances interpretability and allows a comprehensive evaluation of the ethical implications in critical scenarios.

3. Results

3.1. Descriptive analysis

In our evaluation, we achieved a 100 % response rate from the five LLMs tested. Fig. 4 illustrates the distribution of choices made by these models. The results indicate that Mistral-Large and GPT-4o-Mini consistently prefer Case 1 (saving pedestrians) over Case 2 (saving passengers), invariably prioritizing pedestrian safety over passenger protection. The text mining with semantic clustering analysis of the reasons for their decisions is shown in Fig. 5. It revealed that although GPT-4o Mini is aware of the importance of pedestrian safety as well as passenger safety, it consistently prioritizes protecting pedestrians. In contrast, Mistral-Large tends to minimize overall harm but lacks reasoning and computational capabilities when processing quantitative factors in its decision-making process. On the other hand, GPT-4 Turbo and GPT-4o also demonstrated a very strong preference for Case 1.

In general, these LLMs possess an inherent tendency to minimize overall harm by prioritizing the protection of pedestrians who are at higher risk and are valuable road users. While both scenarios are tragic, the LLMs consistently choose to safeguard the group at a greater fatal risk, rather than those having a relatively higher probability of survival. This decision-making pattern of LLMs adheres more closely to the ethos of a “benevolent demon” perspective rather than that of a “survival optimist”, prioritizing the minimization of overall harm even at the cost of lives that could be saved.

3.2. Model results

3.2.1. Binary logit model

In the subsequent analysis, we examined the effects of the considered factors on the choices made by GPT-4 using a binary logit model. The selection of GPT-4 was informed by both descriptive analysis and statistical modeling constraints. Models that produced no variation in outcomes (e.g., Mistral-Large, GPT-4o-Mini) were excluded due to non-convergence, while those with insufficient variation (e.g., GPT-4-Turbo, GPT-4o) resulted in statistically insignificant estimates. Since the binary logit model requires variation in the dependent variable (Y) for meaningful estimation, GPT-4 was the most suitable choice, allowing for analysis of how different scenario factors influence ethical decision-making in AV dilemmas. The results are presented in Table 2. The results are presented in Table 2. It was revealed that several factors significantly influenced the choice of saving passengers or pedestrians in the ethical dilemma scenarios. The constant term (β_0 = -10.117) indicates that GPT-4 inherently showed a low likelihood of selecting Case 2 (saving passengers). This finding supported our hypothesis that GPT-4′s built-in ethical stance prioritizes pedestrian safety over passenger safety.

Furthermore, the results demonstrated that demographic characteristics and fatality risk significantly affect ethical decisions. A larger size passenger group ($\beta = 0.530$, $p = 0.043$), a higher percentage of children and youth ($\beta = 2.192$, $p < 0.001$), a higher percentage of female passengers ($\beta = 1.161$, $p < 0.001$), and a higher fatality risk ($\beta = 10.178$, $p < 0.001$) all increase the likelihood of choosing Case 2 (saving passengers). On the other hand, a higher percentage of children and youth ($\beta =$ -0.596, $p = 0.001$), a higher percentage of females in the pedestrian group ($\beta =$ -0.769, $p < 0.001$), as well an increased fatality risk for pedestrians ($\beta =$ -4.984, $p < 0.001$), decrease the likelihood of choosing Case 2 (saving passengers). These significant factors indicate that GPT-4 exhibits sensitivity to scenario variations and demonstrates a comprehensive reasoning capability that considers multiple factors simultaneously. Overall, GPT-4 emphasizes protecting females, children, and young individuals.

Interestingly, while the group size of passengers significantly affected ChatGPT-4′s ethical decision, the group size of pedestrians did not. This may be because the LLM prioritizes pedestrian safety regardless of group number in most scenarios. Nevertheless, when the number of passengers exceeds that of pedestrians, the LLM became more inclined to choose Case 2 (saving passengers). This finding suggested that the LLM may employ different consideration strategies depending on the relative sizes of the groups involved. The results of the binary logit model highlighted the multifaceted nature of ethical decision-making in LLMs: GPT-4 demonstrates comprehensive consideration of factors but does not process all factors linearly.

3.2.2. Decision tree model

The decision trees presented in Table 3 elucidate the significance of various features in the decision-making processes of the LLMs. The three models exhibited distinct preferences and considerations regarding key factors within the analyzed scenarios. GPT-4o emphasized pedestrian-related aspects, assigning the highest importance to the estimated number of pedestrian deaths (0.467), pedestrian behavior (0.211), and the percentage of children and youth among pedestrians (0.070). In contrast, GPT-4 Turbo adopted a balanced approach by considering both pedestrian and passenger-related variables. GPT-4, however,

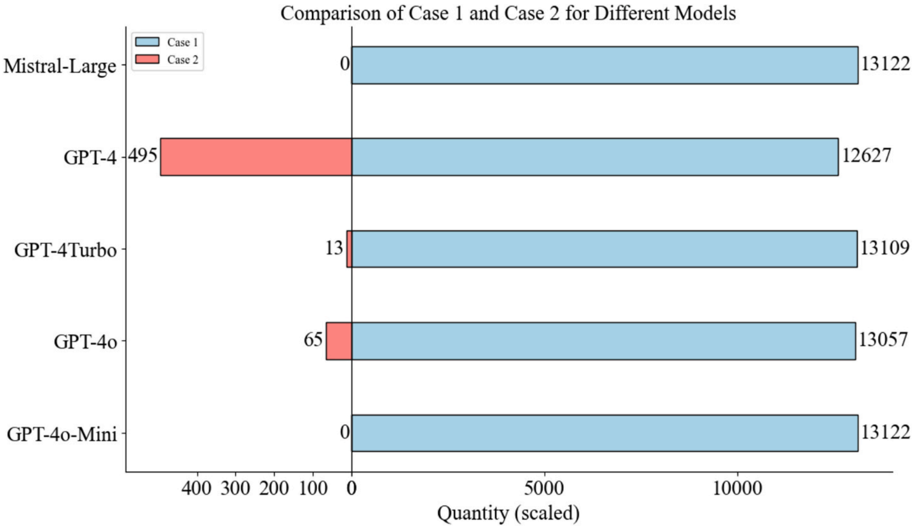
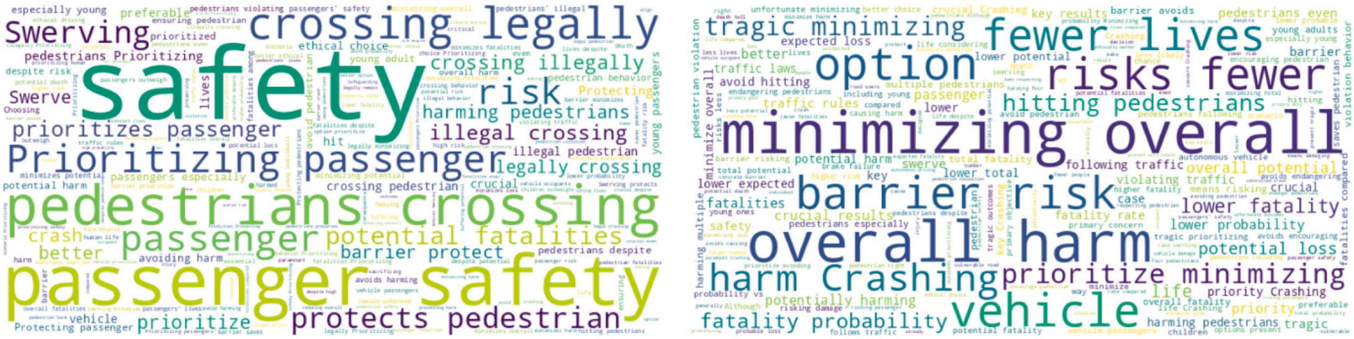


Fig. 4. Distribution of LLM responses to ethical dilemmas – Case 1 vs. Case 2.



(a) GPT-4o Mini

(b) Mistral Large

Fig. 5. Semantic clustering analysis of reasons generated.

Table 2
Results of the Binary Logit Model.

	Coefficient	z-statistic	[0.025	0.975]
ASC$_{\text{(Case 2) crash with pedestrian}}$	−10.117***	−6.734	−13.061	−7.172
Passenger-related attributes				
Number of passengers	0.530**	2.026	0.017	1.044
Percentage of children and youth *(PAX)*	2.192***	12.382	1.845	2.539
Percentage of female *(PAX)*	1.161***	7.262	0.848	1.474
Fatality risk *(PAX)*	10.178***	4.160	5.382	14.973
Pedestrians −related attributes				
Number of pedestrians	−0.082	−0.489	−0.412	0.248
Percentage of children and youth *(PED)*	−0.596***	−3.788	−0.904	−0.288
Percentage of female *(PED)*	−0.769***	−4.932	−1.074	−0.463
Fatality risk *(PED)*	−4.984***	−4.407	−7.201	−2.767
Pedestrian behavior	2.662***	15.116	2.317	3.007
Interaction effect				
Estimated number of deaths *(PAX)* (Number of passengers * Fatality risk)	0.239	0.511	−0.676	1.153
Estimated number of deaths *(PED* (Number of pedestrians * Fatality risk)	−0.777**	−2.298	−1.440	−0.114

Note: Significant at 5 % ** (z value > 1.960), 1 % *** (z value > 2.576); PAX, factors of passenger group; PED, factors of pedestrian group.

demonstrated the most comprehensive evaluation by integrating a broader range of factors. This is evidenced by GPT-4o and GPT-4 Turbo assigning zero importance to five factors each, whereas GPT-4 only assigned zero importance to three factors. These findings suggest that GPT-4 incorporates a broader spectrum of variables into its decision-making process compared to its counterparts.

Furthermore, both GPT-4 and GPT-4o prioritized the estimated number of deaths – calculated as the product of group size and fatality risk – over individual factors such as fatality risk or group size alone. By considering the interaction effects of group size and fatality risk, the LLMs may demonstrate a more nuanced approach to risk assessment. Notably, violation behavior by pedestrians also influenced the choices made by all LLMs. This consideration of pedestrian behavior alongside other factors illustrates the models' ability to integrate contextual information into ethical deliberations. This approach extends beyond simple utilitarian calculations to incorporate personal responsibility and societal norms throughout the decision-making process.

The decision tree visualization elucidates the complex reasoning processes of LLMs in ethical decision-making scenarios, as illustrated in Figs. 6 to 8. The depth and breadth of the tree structures serve as visual indicators of the decision processes' complexity. This demonstrates the models' capacity to integrate multiple interacting factors rather than relying on simplistic, linear judgments. It can be observed that the GPT-4 exhibits the most complex tree structure, while the GPT-4 Turbo follows more straightforward decision rules.

Additionally, the specific numerical values within the trees reveal the models' decision thresholds under various conditions. This enhances

Table 3
Feature Importance in the Decision Tree Model.

Model	GPT-4 Turbo		GPT-4o		GPT-4	
	Feature	Importance	Feature	Importance	Feature	Importance
1	Percentage of children and youth *(PED)*	0.408	Estimated death *(PED)*	0.467	Estimated death *(PED)*	0.303
2	Pedestrian behavior	0.199	Pedestrian behavior	0.211	Pedestrian behavior	0.243
3	Percentage of female *(PAX)*	0.192	Estimated death *(PAX)*	0.150	Estimated death *(PAX)*	0.235
4	Probability of death *(PED)*	0.123	Percentage of female *(PED)*	0.101	Percentage of children and youth *(PAX)*	0.154
5	Estimated death *(PAX)*	0.049	Percentage of children and youth *(PED)*	0.070	Percentage of female *(PAX)*	0.022
6	Estimated death *(PED)*	0.029	Probability of death *(PAX)*	0.001	Percentage of female *(PED)*	0.018
7	Number of passengers	0.000	Number of passengers	0.000	Probability of death *(PED)*	0.016
8	Percentage of children and youth *(PAX)*	0.000	Percentage of children and youth *(PAX)*	0.000	Probability of death *(PAX)*	0.009
9	Probability of death *(PAX)*	0.000	Percentage of female *(PAX)*	0.000	Percentage of children and young *(PED)*	0.000
10	Number of pedestrians	0.000	Number of pedestrians	0.000	Number of passengers	0.000
11	Percentage of female *(PED)*	0.000	Probability of death *(PED)*	0.000	Number of pedestrians	0.000

Note: PAX, factors of passenger group; PED, factors of pedestrian group.

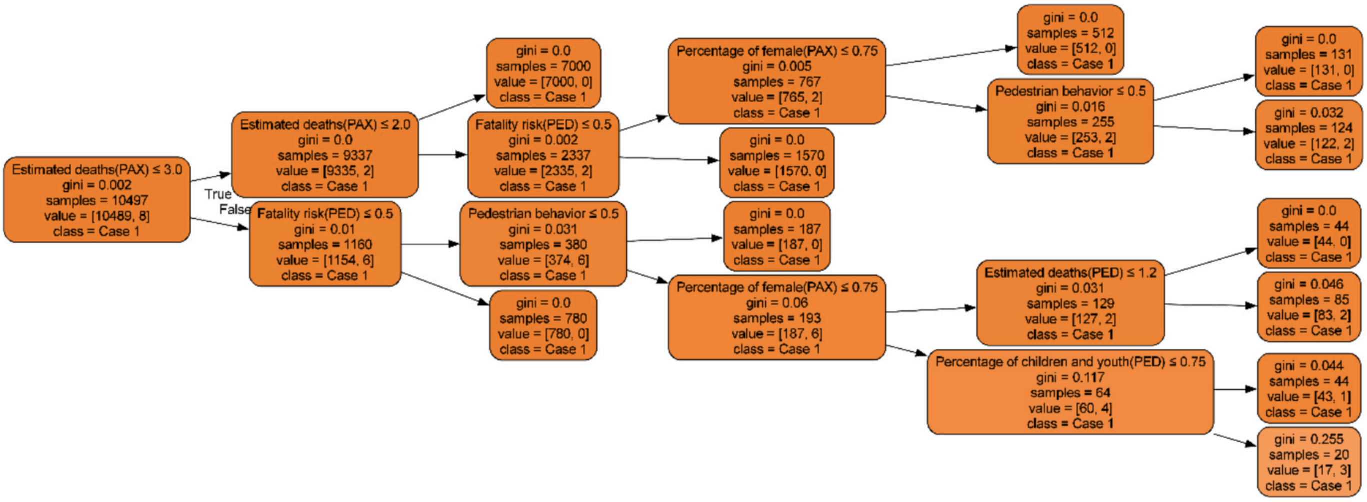

Estimated deaths(PAX) ≤ 3.0 gini = 0.002 samples = 10497 value = [10489, 8] class = Case 1
True
False
Estimated deaths(PAX) ≤ 2.0 gini = 0.0 samples = 9337 value = [9335, 2] class = Case 1
gini = 0.0 samples = 7000 value = [7000, 0] class = Case 1
Fatality risk(PED) ≤ 0.5 gini = 0.002 samples = 2337 value = [2335, 2] class = Case 1
Percentage of female(PAX) ≤ 0.75 gini = 0.005 samples = 767 value = [765, 2] class = Case 1
gini = 0.0 samples = 512 value = [512, 0] class = Case 1
Pedestrian behavior ≤ 0.5 gini = 0.016 samples = 255 value = [253, 2] class = Case 1
gini = 0.0 samples = 131 value = [131, 0] class = Case 1
gini = 0.032 samples = 124 value = [122, 2] class = Case 1
gini = 0.0 samples = 1570 value = [1570, 0] class = Case 1
Fatality risk(PED) ≤ 0.5 gini = 0.01 samples = 1160 value = [1154, 6] class = Case 1
Pedestrian behavior ≤ 0.5 gini = 0.031 samples = 380 value = [374, 6] class = Case 1
gini = 0.0 samples = 187 value = [187, 0] class = Case 1
Percentage of female(PAX) ≤ 0.75 gini = 0.06 samples = 193 value = [187, 6] class = Case 1
Estimated deaths(PED) ≤ 1.2 gini = 0.031 samples = 129 value = [127, 2] class = Case 1
gini = 0.0 samples = 44 value = [44, 0] class = Case 1
gini = 0.046 samples = 85 value = [83, 2] class = Case 1
Percentage of children and youth(PED) ≤ 0.75 gini = 0.117 samples = 64 value = [60, 4] class = Case 1
gini = 0.044 samples = 44 value = [43, 1] class = Case 1
gini = 0.255 samples = 20 value = [17, 3] class = Case 1
gini = 0.0 samples = 780 value = [780, 0] class = Case 1

Fig. 6. Visualization of GPT-4 turbo decision tree Note: PAX, factors of passenger group; PED, factors of pedestrian group.

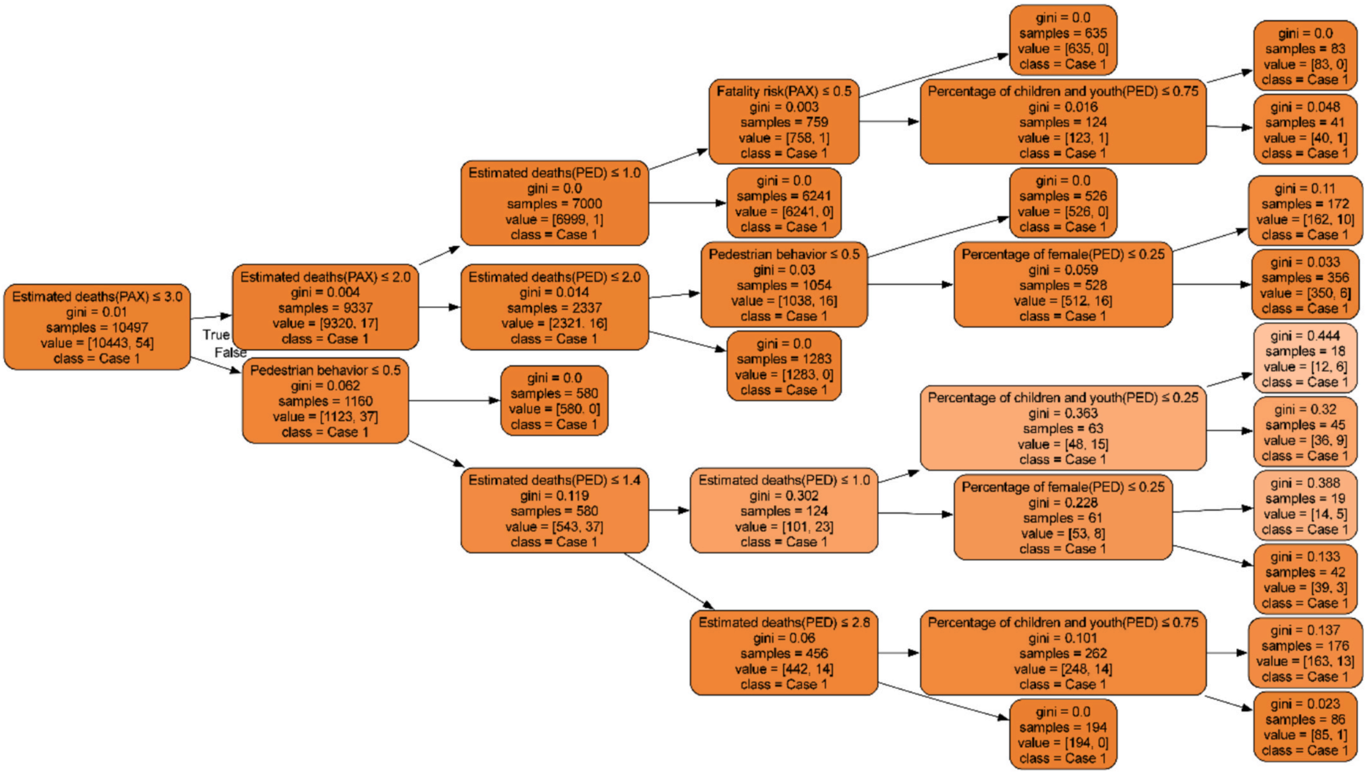

Estimated deaths(PAX) ≤ 3.0 gini = 0.01 samples = 10497 value = [10443, 54] class = Case 1
True
False
Estimated deaths(PAX) ≤ 2.0 gini = 0.004 samples = 9337 value = [9320, 17] class = Case 1
Estimated deaths(PED) ≤ 1.0 gini = 0.0 samples = 7000 value = [6999, 1] class = Case 1
Fatality risk(PAX) ≤ 0.5 gini = 0.003 samples = 759 value = [758, 1] class = Case 1
gini = 0.0 samples = 635 value = [635, 0] class = Case 1
Percentage of children and youth(PED) ≤ 0.75 gini = 0.016 samples = 124 value = [123, 1] class = Case 1
gini = 0.0 samples = 83 value = [83, 0] class = Case 1
gini = 0.048 samples = 41 value = [40, 1] class = Case 1
gini = 0.0 samples = 6241 value = [6241, 0] class = Case 1
Estimated deaths(PED) ≤ 2.0 gini = 0.014 samples = 2337 value = [2321, 16] class = Case 1
Pedestrian behavior ≤ 0.5 gini = 0.03 samples = 1054 value = [1038, 16] class = Case 1
gini = 0.0 samples = 526 value = [526, 0] class = Case 1
Percentage of female(PED) ≤ 0.25 gini = 0.059 samples = 528 value = [512, 16] class = Case 1
gini = 0.11 samples = 172 value = [162, 10] class = Case 1
gini = 0.033 samples = 356 value = [350, 6] class = Case 1
gini = 0.0 samples = 1283 value = [1283, 0] class = Case 1
Pedestrian behavior ≤ 0.5 gini = 0.062 samples = 1160 value = [1123, 37] class = Case 1
gini = 0.0 samples = 580 value = [580, 0] class = Case 1
Estimated deaths(PED) ≤ 1.4 gini = 0.119 samples = 580 value = [543, 37] class = Case 1
Estimated deaths(PED) ≤ 1.0 gini = 0.302 samples = 124 value = [101, 23] class = Case 1
Percentage of children and youth(PED) ≤ 0.25 gini = 0.363 samples = 63 value = [48, 15] class = Case 1
gini = 0.444 samples = 18 value = [12, 6] class = Case 1
gini = 0.32 samples = 45 value = [36, 9] class = Case 1
Percentage of female(PED) ≤ 0.25 gini = 0.228 samples = 61 value = [53, 8] class = Case 1
gini = 0.388 samples = 19 value = [14, 5] class = Case 1
gini = 0.133 samples = 42 value = [39, 3] class = Case 1
Estimated deaths(PED) ≤ 2.8 gini = 0.06 samples = 456 value = [442, 14] class = Case 1
Percentage of children and youth(PED) ≤ 0.75 gini = 0.101 samples = 262 value = [248, 14] class = Case 1
gini = 0.137 samples = 176 value = [163, 13] class = Case 1
gini = 0.023 samples = 86 value = [85, 1] class = Case 1
gini = 0.0 samples = 194 value = [194, 0] class = Case 1

Fig. 7. Visualization of GPT-4o decision tree.

their interpretability and allows each decision to be traced and explained. From this perspective, it is evident that GPT-4o and GPT-4 Turbo lack specific conditions for Case 2, whereas GPT-4 presents clear definitional criteria for Case 2 (represented by blue squares in the visualization).

Specifically, GPT-4 selects Case 2 under two distinct conditions. In the first condition, this choice is made when the estimated number of passenger deaths exceeds 2.00, the estimated number of pedestrian deaths is no more than 1.00, pedestrians are law-abiding (behavior > 0.50), and the percentage of young passengers is substantial (> 0.25). In the second condition, GPT-4 opts for Case 2 when passenger death is projected to be over 3.00, pedestrians are law-abiding, and the estimated number of pedestrian deaths does not exceed 2.80 (with a further threshold at 1.40), and a significant proportion of young passengers (> 0.25). These decision criteria demonstrate GPT-4′s nuanced ethical reasoning approach. Through analysis of the decision tree results, we can conclude that GPT-4 demonstrates the most nuanced decision-making capabilities, effectively considering various scenario factors and generating flexible strategies in response.

4. Discussion

4.1. LLM vs. Human moral Decision-Making

This investigation examined the alignment of LLMs with human ethical values in the context of transport systems, particularly in scenarios involving potential conflicts between AVs and vulnerable road users. For human moral decision-making, an intricate interplay of

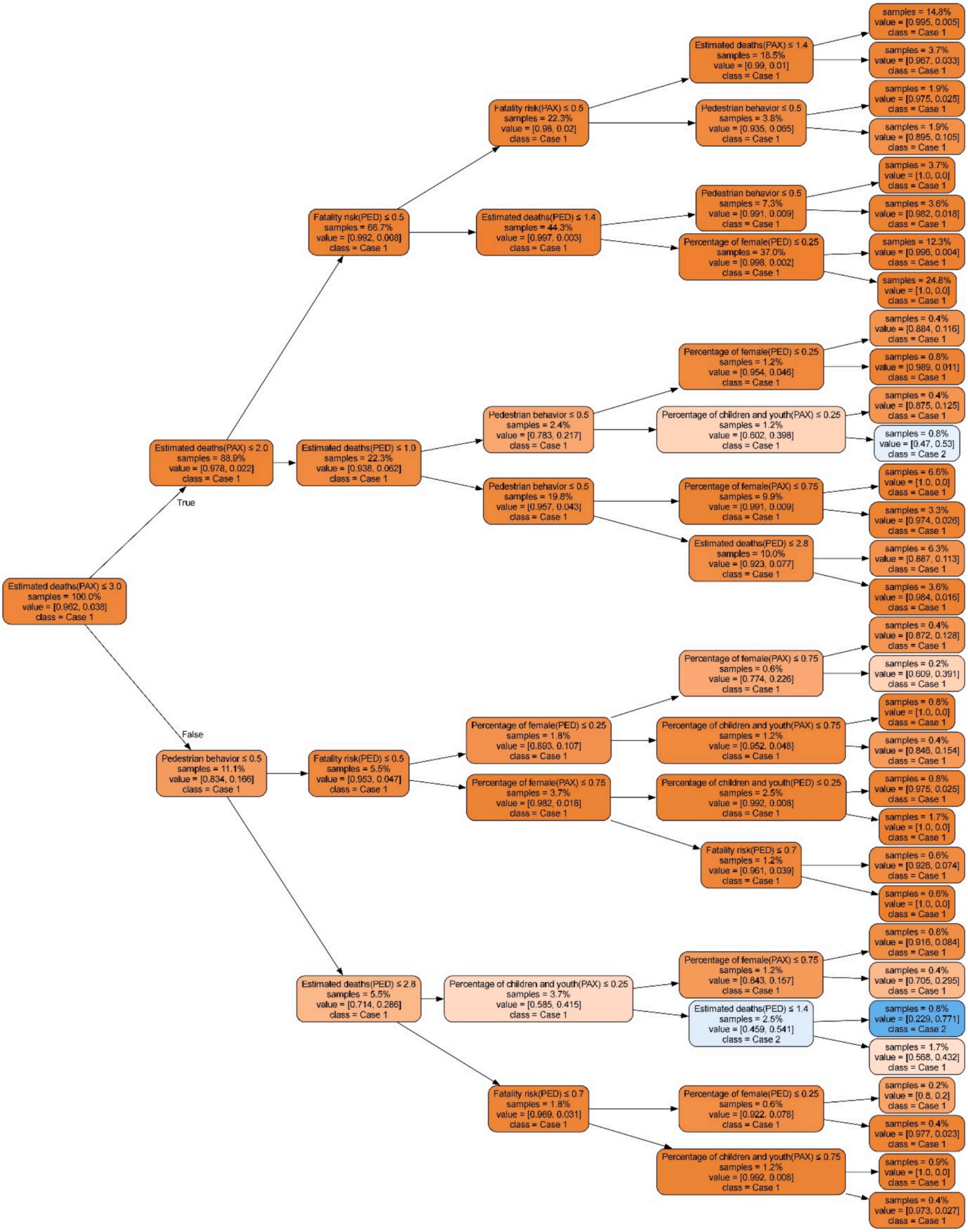


Fig. 8. Visualization of GPT-4 decision tree.

cognitive, emotional, and cultural factors guides behavior (Awad et al., 2018). Research shows that when confronted with dilemmas involving unavoidable harm, individuals often adopt a utilitarian approach — prioritizing the preservation of the largest number of lives (Faulhaber et al., 2019). However, this tendency is not absolute. Deontological principles and intuitive responses can also influence decisions. For instance, in high-pressure situations or when personal or familial safety is at risk, individuals may choose to protect specific persons rather than maximize collective well-being (Bonnefon et al., 2016). Moreover, cultural norms play a significant role in shaping these choices. Some societies emphasize respect for elders or retribution against wrongdoers, while others focus on minimizing overall harm (Awad et al., 2020). Consequently, while human moral judgment generally leans toward harm reduction, it also exhibits considerable variation based on context and individual circumstances.

In contrast, AI-driven moral decision-making is based on algorithmic frameworks and data-driven patterns. Researchers have explored hybrid approaches that integrate rule-based constraints with machine learning, as well as theoretical frameworks like game theory, to model ethical decisions (Conitzer et al., 2017; Wallach and Allen, 2008). Recent advancements, particularly in large language models, have shifted the focus toward training on extensive datasets to better align AI outputs with human ethical preferences. Specifically, our findings revealed a notable consistency between the preferences exhibited by LLMs and human ethical values. Specifically, the LLMs demonstrated a clear tendency to prioritize saving a greater number of lives and protecting younger individuals, which aligns with previous empirical studies. This focus on minimizing fatalities reflects a utilitarian approach, aiming to reduce harm in ethical dilemma scenarios (Johnsen et al., 2017).

However, a divergence emerged when comparing the LLMs' decision-making to human preferences, particularly regarding protecting pedestrians versus passengers. While humans tend to exhibit a mild preference for pedestrian safety, the LLMs—especially the GPT-4 series and Mistral-Large—showed a pronounced bias toward prioritizing pedestrian protection. This stronger emphasis on pedestrian safety aligns with findings from earlier studies (Takemoto, 2024). One potential explanation for this could be that pedestrian vulnerability has long been a focal point of advocacy efforts and discussion, emphasizing their inherent risks in road environments as they lack the same protection as cars. By applying a principle of justice—which humans often struggle to fully implement concerning car-pedestrian vulnerabilities, even in current policy-making (Basu et al., 2023; Martinez-Buelvas et al., 2024)—the LLMs may arrive at a less common yet ethically consistent solution that favors the protection of the most vulnerable individuals in these scenarios.

4.2. Inconsistencies and ethical challenges of LLMs

This study also identified the nuanced differences in the ethical decision-making capabilities of LLM. Our experiments revealed a positive correlation between a model's capacity for mimicking ethical reasoning and its overall capabilities and scale. Specifically, GPT-4o Mini consistently prioritized pedestrian protection over passengers, demonstrating a rigid decision-making pattern that did not account for scenario-specific factors. In contrast, GPT-4, the most sophisticated model in the study, displayed more nuanced and content-sensitive ethical decision-making. Previous research also substantiated GPT-4′s superior reasoning abilities in moral questions, equivalent to typical graduate school students (Rao et al., 2023, Tanmay et al., 2023). A study further confirmed GPT-4′s leading position in alignment with human moral judgments (Almeida et al., 2024). This growing trend of querying AI systems like GPT models on moral dilemmas, however, raises significant concerns about the future of decision-making. As these systems increasingly provide answers to ethically charged questions, they might inadvertently create a precedent where complex moral decision-making is outsourced to machines. This reliance poses risks, as humans may turn to AI for guidance on societal issues, despite the fact that machines, while capable of mimicking reasoning, lack an intrinsic understanding of human values and the consequences of ethical decisions. The seemingly thoughtful responses of models like GPT-4 could foster misplaced trust in their judgment, potentially encouraging the delegation of moral responsibility to technology. This raises critical questions about accountability—if moral decisions are influenced or driven by AI, where does responsibility lie? The risk is that humans could abdicate responsibility for these decisions, creating ethical ambiguity around who should be held accountable when outcomes are unfavorable.

There is also evidence that LLM developed by different companies exhibit significant differences in ethical decision-making. It was found that ChatGPT (encompassing both GPT-3.5 and GPT-4) demonstrated ethical preferences that closely matched human tendencies in moral dilemmas. In contrast, a previous study indicates that the open-source Llama 2 model prioritized passengers over pedestrians, diverging from both human and ChatGPT preferences (Takemoto, 2024). Additionally, some models, such as Llama 3 70B-Instruct and PaLM 2, display distinct ethical biases toward saving fewer people and crushing more people (Vida et al., 2024). These discrepancies may arise from variations in the training data and the ethical frameworks embedded within each model during development. There is a need for further research in this area to clarify the data and model architecture differences that influence ethical decision-making discrepancies.

LLMs showed inconsistent performance across different moral scenarios (Bonagiri et al., 2024, Tanmay et al., 2023), suggesting they lack robust ethical principles. Although LLMs make basic norm judgments reliably, they often exhibit uncertainty with more complex or ambiguous content (Scherrer et al., 2024). It has been suggested that these limitations may stem from inherent biases in the training data used to develop the models (Bender et al., 2021, Abdulhai et al., 2023). These challenges underscore the need for continued research and development to enhance LLMs' ability to provide consistent and universally accepted moral reasoning. With LLMs being incorporated into critical decision-making systems like AVs, addressing these ethical limitations becomes increasingly important to ensure fair, consistent, and culturally sensitive responses.

The ethical implications of allowing a machine, such as an LLM, to judge individual actions without fully considering the complexities involved are concerning. As shown in our results, GPT-4 and other models often assign blame to pedestrians based on traffic law violations, i.e., jaywalking. However, this approach fails to account for the ambiguity in policies and the variations in legal frameworks across different regions (Bhuiyan et al., 2024). What constitutes a violation in one context may not be illegal elsewhere, and even within the same jurisdiction, policies can be unclear or inconsistent. By relying on these ambiguous legal frameworks without fully considering their complexities, the system risks unfairly assigning blame or making ethically questionable decisions. This raises the critical issue of whether it is appropriate for a machine to pass judgment on human actions when it may lack the nuanced understanding of the legal and ethical gray areas that a human decision-maker would take into account.

Additionally, previous research on systems like ChatGPT has shown that LLMs tend to provide specific advice or solutions that might sound convincing but are not always aligned with the safest or most appropriate course of action for an individual (Oviedo-Trespalacios et al., 2023). This overconfidence in the model's responses can create an ecological fallacy, where general trends are mistakenly applied to specific situations, potentially leading to harm. For example, in AV ethical dilemmas, while the model may consider statistical probabilities and general rules, it may fail to address unique, situational factors that could alter the best course of action. As the findings from this study suggest, the nuanced decision-making of LLMs, while impressive, is still prone to generalization and may overlook critical, context-specific details essential for ensuring safety and ethical soundness in real-world applications.

Machine ethics encompasses a much broader scope of tasks, such as avoiding generating immoral or toxic content and demonstrating the ability to assess moral norms (Sun et al., 2024). While LLMs, such as GPT-4, perform reasonably well in AV ethical dilemmas, their performance in broader ethical contexts reveals several limitations. Research has highlighted cultural biases in models like GPT-4, which tend to favor moral values predominant in Western and English-speaking countries (Rao et al., 2023). Moreover, 'language inequality' has been identified, where LLMs exhibit uneven performance and moral reasoning across different prompt languages (Jin et al., 2024). This study, conducted in Korea using English-language prompts rather than the native language, further raises questions about potential differences in outcomes based on linguistic and cultural contexts. Testing these models in native languages could yield different findings, as the interplay of culture and language may significantly influence the model's ethical judgments. Cultural biases and linguistic inequalities highlight the risk of potentially unfair or inappropriate responses for users from diverse cultural backgrounds, underscoring the importance of context-specific evaluations of LLMs' ethical stances.

4.3. Limitations and implications

Our study presents a novel framework for evaluating the ethical reasoning capabilities of LLMs in the context of autonomous driving. However, our study has limitations that could be addressed to further advance this field. First, although the stated choice experiment provides a controlled environment for assessing moral scenarios, its binary, trolley-like dilemmas inherently oversimplify the real-world ethical scenarios. As noted by Geisslinger et al. (2021), this approach cannot account for nuanced factors such as road users' intentions, nor does it capture moral deliberations in relatively low-risk traffic conditions (Geisslinger et al., 2021). Secondly, our primary objective in this study was to conduct a deep dive into the ethical reasoning mechanisms of GPT-4 across a comprehensive set of scenarios. Although GPT-4 represents a single decision-maker, its responses under varied conditions allow us to systematically probe its internal decision boundaries and capture intrinsic variability similar to intra-individual variation in human decision-making. Nevertheless, we acknowledge that focusing on a single LLM may constrain the exploration of inter-model variability. In future work, we plan to extend this approach by comparing responses across multiple LLMs. Our evaluation scope remains temporally constrained, as we focus on five mainstream LLMs at a specific point in time. Given that LLMs evolve rapidly, subsequent versions of models may exhibit altered moral decision-making due to updates in model structure, training data and reasoning paradigm. Consequently, our current results should be viewed as a snapshot of a constantly shifting landscape, rather than definitive conclusions.

Despite these limitations, our framework establishes a standardized benchmark for evaluating and comparing diverse models, providing a solid foundation for future research in this rapidly evolving field. Potential conflicting imperatives are emerging from public expectations regarding AI-driven moral decision-making in autonomous vehicles: society requires autonomous vehicles to follow widely accepted ethical standards, such as minimizing overall casualties, while individuals remain cautious about adopting technologies that may compromise their personal safety (Bonnefon et al., 2016; Greene, 2016). The ability to address this tension requires transparent ethical design, clear communication regarding the rationale behind moral algorithms, and the ability to customize these systems in order to balance collective benefits with individual security concerns (Di Fabio et al., 2017). The purpose of our study is to provide actionable insights for the development of autonomous vehicle systems with integrated, ethically aware decision-making models, thus fostering public trust and facilitating the safe deployment of these systems.

5. Conclusion

In conclusion, this study provides important insights into the ethical decision-making processes of Large Language Models (LLMs) when applied to autonomous vehicles (AVs). Through binary logit models and decision tree analyses, we observed that LLMs, particularly GPT-4, tend to prioritize pedestrian safety over passengers in ethical dilemmas. This preference reflects a consistent approach to minimizing harm. Nevertheless, it also highlights concerns about the system's ability to fully consider the complexities of each situation, including cultural, legal, and contextual factors. While GPT-4 demonstrated superior reasoning by integrating multiple variables into its decisions, the study underscores current LLM frameworks' limitations and potential biases, especially when legal or policy ambiguities are involved. Given the rapid integration of LLMs into critical socio-technical systems, particularly in AV technology, future research and development must address these ethical challenges head-on. Transparency in how LLMs make decisions and rigorous testing of their ethical frameworks are essential to ensure alignment with societal values and prevent misuse. Policymakers and developers must collaborate to create robust regulatory controls, ensuring that as LLMs become more embedded in our daily lives, they do so in ways that are safe, ethical, and beneficial for all members of society. This research serves as a foundation for further investigation into the ethical implications of LLMs in AVs. It highlights the need to continuously refine the technology and its governing principles. By providing a replicable framework and elucidating both the potential and challenges of LLM-based ethics, we hope to foster interdisciplinary efforts to refine and integrate moral decision-making techniques into the next generation of AVs.

CRediT authorship contribution statement

Zixuan Xu: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation. **Neha Sengar:** Writing – original draft, Methodology, Investigation. **Tiantian Chen:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Hyungchul Chung:** Writing – review & editing, Supervision. **Oscar Oviedo-Trespalacios:** Writing – review & editing, Supervision.

Acknowledgement

This work is supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (RS-2024-00351865).

Data availability

Data will be made available on request.

References

Abdel-Aty, M., Keller, J., Brady, P.A., 2005. Analysis of types of crashes at signalized intersections by using complete crash data and tree-based regression. Transp. Res. Rec. 1908 (1), 37–45.

Abdulhai, M., Serapio-Garcia, G., Crepy, C., Valter, D., Canny, J. & Jaques, N. (2023) Moral foundations of large language models. arXiv preprint arXiv:2310.15337.

Almeida, G.F., Nunes, J.L., Engelmann, N., Wiegmann, A., De Araújo, M., 2024. Exploring the psychology of LLMs' moral and legal reasoning. Artif. Intell. 333, 104145.

Apté, C., Weiss, S., 1997. Data mining with decision trees and decision rules. Futur. Gener. Comput. Syst. 13, 197–210.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F., Rahwan, I., 2018. The moral machine experiment. Nature 563, 59–64.

Awad, E., Dsouza, S., Shariff, A., Rahwan, I., Bonnefon, J.-F., 2020. Universals and variations in moral decisions made in 42 countries by 70,000 participants. Proc. Natl. Acad. Sci. 117, 2332–2337.

Balas, M., Wadden, J.J., Hébert, P.C., Mathison, E., Warren, M.D., Seavilleklein, V., Wyzynski, D., Callahan, A., Crawford, S.A., Arjmand, P., 2024. Exploring the potential utility of AI large language models for medical ethics: an expert panel evaluation of GPT-4. J. Med. Ethics 50, 90–96.

Basu, N., Oviedo-Trespalacios, O., King, M., Kamruzzaman, M., Haque, M.M., 2023. What do pedestrians consider when choosing a route? The role of safety, security, and attractiveness perceptions and the built environment during day and night walking. Cities 143, 104551.

Bender, E.M., Gebru, T., Mcmillan-Major, A., Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big??. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623.

Bonagiri, V. K., Vennam, S., Gaur, M. & Kumaraguru, P. (2024) Measuring Moral Inconsistencies in Large Language Models. arXiv preprint arXiv:2402.01719.

Bonnefon, J.-F., Shariff, A., Rahwan, I., 2016. The social dilemma of autonomous vehicles. Science 352, 1573–1576.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees. Wadsworth Inc.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., 2020. Language models are few-shot learners. Adv. Neural Inf. Proces. Syst. 33, 1877–1901.

Cheong, I., Xia, K., Feng, K. K., Chen, Q. Z. & Zhang, A. X. (A) I am not A lawyer, but...: engaging legal experts towards responsible LLM policies for legal advice. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024. 2454-2469.

Conitzer, V., Sinnott-Armstrong, W., Borg, J.S., Deng, Y., Kramer, M., 2017. Moral decision making frameworks for artificial intelligence. Proceedings of the AAAI Conference on Artificial Intelligence.

Cui, C., Yang, Z., Zhou, Y., Ma, Y., Lu, J., Li, L., Chen, Y., Panchal, J. & Wang, Z. (2023) Personalized autonomous driving with large language models: field experiments. arXiv preprint arXiv:2312.09397.

De Moura, N. & Et Al. Ethical decision making for autonomous vehicles. 2020 IEEE intelligent vehicles symposium (iv), 2020. IEEE.

De Ville, B., 2013. Decision trees. Wiley Interdiscip. Rev. Comput. Stat. 5, 448–455.

Dewitt, B., Fischhoff, B. & Sahlin, N. E. (2019) 'Moral machine' experiment is no basis for policymaking. Nature, 567, 31-31.

Di Fabio, U., Broy, M., Brüngger, R.J., Eichhorn, U., Grunwald, A., Heckmann, D., Hilgendorf, E., Kagermann, H., Losinger, A., Lutz-Bachmann, M., 2017. Ethics commission automated and connected driving. In: Federal Ministry of Transport and Digital Infrastructure of the Federal Republic of Germany, p. 1.

Faulhaber, A.K., Dittmer, A., Blind, F., Wächter, M.A., Timm, S., Sütfeld, L.R., Stephan, A., Pipa, G., König, P., 2019. Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles. Sci. Eng. Ethics 25, 399–418.

Furey, H., Hill, S., 2021. MIT's moral machine project is a psychological roadblock to selfdriving cars. AI Ethics 1, 151–155.

Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., Ahmed, N.K., 2024. Bias and fairness in large language models: A survey. Comput. Linguist. 50, 1097–1179.

Gan, L., Chu, W., Li, G., Tang, X., Li, K., 2024. Large models for intelligent transportation systems and autonomous vehicles: A survey. Adv. Eng. Inf. 62, 102786.

Geisslinger, M., Poszler, F., Betz, J., Lütge, C., Lienkamp, M., 2021. Autonomous driving ethics: From trolley problem to ethics of risk. Philos. Technol. 34, 1033–1055.

Ghasri, M., Rashidi, T.H., Waller, S.T., 2017. Developing a disaggregate travel demand system of models using data mining techniques. Transp. Res. A Policy Pract. 105, 138–153.

Gogoll, J., Müller, J.F., 2017. Autonomous cars: in favor of a mandatory ethics setting. Sci. Eng. Ethics 23, 681–700.

Goodall, N., 2014a. Ethical decision making during automated vehicle crashes. Transport. Res. Record: J. Transport. Res. Board 2424, 58–65.

Goodall, N.J., 2014b. Machine ethics and automated vehicles. Springer, Road Vehicle Automation.

Goodall, N.J., 2016. Away from trolley problems and toward risk management. Appl. Artif. Intell. 30, 810–821.

Greene, J.D., 2016. Our driverless dilemma. Science 352, 1514–1515.

Jin, Z., Levine, S., Kleiman-Weiner, M., Piatti, G., Liu, J., Adauto, F. G., Ortu, F., Strausz, A., Sachan, M. & Mihalcea, R. (2024) Multilingual Trolley Problems for Language Models. arXiv preprint arXiv:2407.02273.

Johnsen, A., Strand, N., Andersson, J., Patten, C., Kraetsch, C. & Takman, J. (2017) Literature review on the acceptance and road safety, ethical, legal, social and economic implications of automated vehicles.

Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y., 2022. Large language models are zero-shot reasoners. Adv. Neural Inf. Proces. Syst. 35, 22199–22213.

Leben, D., 2017. A Rawlsian algorithm for autonomous vehicles. Ethics Inf. Technol. 19, 107–115.

Lin, P., 2015. Why Ethics Matters for Autonomous Cars. Springer, Autonomes Fahren.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., 2023. Self-refine: Iterative refinement with self-feedback. Adv. Neural Inf. Proces. Syst. 36, 46534–46594.

Malle, B. F., Scheutz, M. & Voiklis, J. Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents. 2015.

Martínez-Buelvas, L., Rakotonirainy, A., Grant-Smith, D., & Oviedo-Trespalacios, O. (2022). A transport justice approach to integrating vulnerable road users with automated vehicles. Transportation research part D: transport and environment, 113, 103499.

Martínez-Buelvas, L., Rakotonirainy, A., Grant-Smith, D., & Oviedo-Trespalacios, O. (2024). A multi-road user evaluation of the acceptance of connected and automated vehicles through the lenses of safety and justice. Transportation research part F: traffic psychology and behaviour, In Press.

Martínez-Buelvas, L., Rakotonirainy, A., Grant-Smith, D., Oviedo-Trespalacios, O., 2024b. A multi-road user evaluation of the acceptance of connected and automated vehicles through the lenses of safety and justice. Transportation Research Part F: Traffic Psychology and Behaviour 107, 521–536.

Nations, U. (2021) Road traffic injuries are the leading killer of people aged 5-29 years.

Openai.(2023) Prompt engineering [Online]. Available: https://platform.openai.com/docs/guides/prompt-engineering#tactic-give-the-model-access-to-specific-functions [Accessed].

Oviedo-Trespalacios, O., Peden, A.E., Cole-Hunter, T., Costantini, A., Haghani, M., Rod, J.E., Reniers, G., 2023. The risks of using ChatGPT to obtain common safety-related information and advice. Saf. Sci. 167, 106244.

Rao, A., Khandelwal, A., Tanmay, K., Agarwal, U. & Choudhury, M. (2023) Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. arXiv preprint arXiv:2310.07251.

Rodionov, S., Goertzel, Z. A. & Goertzel, B. (2023) An Evaluation of GPT-4 on the ETHICS Dataset. arXiv preprint arXiv:2309.10492.

Rokach, L., Maimon, O., 2005. Top-down induction of decision trees classifiers-a survey. IEEE Trans. Syste., Man, and Cybernet., Part C (Applications and Reviews) 35 (4), 476–487.

Scherrer, N., Shi, C., Feder, A., Blei, D., 2024. Evaluating the moral beliefs encoded in llms. Adv. Neural Inf. Proces. Syst. 36.

Schuessler, D., 2023. The probability problems of the moral machine experiment. AI. Ethics.

Sha, H., Mu, Y., Jiang, Y., Chen, L., Xu, C., Luo, P., Li, S. E., Tomizuka, M., Zhan, W. & Ding, M. (2023) Languagempc: Large language models as decision makers for autonomous driving. arXiv preprint arXiv:2310.03026.

Sun, L., Huang, Y., Wang, H., Wu, S., Zhang, Q., Gao, C., Huang, Y., Lyu, W., Zhang, Y. & Li, X. (2024) Trustllm: Trustworthiness in large language models. arXiv preprint arXiv:2401.05561.

Takemoto, K., 2024. The moral machine experiment on large language models. R. Soc. Open Sci. 11, 231393.

Tanmay, K., Khandelwal, A., Agarwal, U. & Choudhury, M. (2023) Probing the Moral Development of Large Language Models through Defining Issues Test. arXiv e-prints, arXiv: 2309.13356.

Vida, K., Damken, F. & Lauscher, A. (2024) Decoding Multilingual Moral Preferences: Unveiling LLM's Biases Through the Moral Machine Experiment. arXiv preprint arXiv:2407.15184.

Torkamaan, H., Steinert, S., Pera, M.S., Kudina, O., Freire, S.K., Verma, H., Oviedo-Trespalacios, O., 2024. Challenges and future directions for integration of large language models into socio-technical systems. Behaviour & Information Technology 1–20. https://doi.org/10.1080/0144929X.2024.2431068.

Wallach, W., Allen, C., 2008. Moral machines: Teaching robots right from wrong. Oxford University Press.

Wang, Y. & Et Al. (2023) Aligning large language models with human: A survey. arXiv preprint arXiv:2307.12966.

Wang, Y., Jiao, R., Zhan, S. S., Lang, C., Huang, C., Wang, Z., Yang, Z. & Zhu, Q. (2023) Empowering autonomous driving with large language models: A safety perspective. arXiv preprint arXiv:2312.00812.

Wei, J. & Et Al. (2022) Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 24824-24837.

Weidinger, L. & Et Al. (2021) Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

Xu, Z., Chen, T., Chen, S., 2024, September. A LLM-based Multimodal Warning System for Driver Assistance. IEEE 27th International Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 1527–1532.

Xu, Z., Chen, T., Huang, Z., Xing, Y., Chen, S., 2025. Personalizing Driver Agent Using Large Language Models for Driving Safety and Smarter Human–Machine Interactions. IEEE Intelligent Transportation Systems Magazine.

Yang, Z., Jia, X., Li, H., Yan, J., 2023. Llm4drive: A survey of large language models for autonomous driving. NeurIPS 2024 Workshop on Open-World Agents.

Zhao, Z., Lee, W.S., Hsu, D., 2024. Large language models as commonsense knowledge for large-scale task planning. Adv. Neural Inf. Proces. Syst.

Zhuang, C., Gu, T., Chung, H., Zhu, M., Yonto, D., 2025. Spatial-temporal insights into gender gaps in East Asian ride-hailing: Workload, efficiency, nighttime safety, and operational patterns. Journal of Transport Geography 125, 104213.

Zhu, Y., Wang, S., Zhong, W., Shen, N., Li, Y., Wang, S., Li, Z., Wu, C., He, Z. & Li, L. (2024) Will Large Language Models be a Panacea to Autonomous Driving? arXiv preprint arXiv:2409.14165.