ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Takemoto K. 2024 The moral machine experiment on large language models. *R. Soc. Open Sci.* **11**: 231393.
https://doi.org/10.1098/rsos.231393

Received: 16 September 2023
Accepted: 17 January 2024

Subject Category:
Computer science and artificial intelligence

Subject Areas:
artificial intelligence/human-computer interaction

Keywords:
moral machine, large language models, ChatGPT, autonomous driving

Author for correspondence:
Kazuhiro Takemoto
e-mail: takemoto@bio.kyutech.ac.jp

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.7055611.

THE ROYAL SOCIETY PUBLISHING

The moral machine experiment on large language models

Kazuhiro Takemoto

Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502, Japan

KT, 0000-0002-6355-1366

As large language models (LLMs) have become more deeply integrated into various sectors, understanding how they make moral judgements has become crucial, particularly in the realm of autonomous driving. This study used the moral machine framework to investigate the ethical decision-making tendencies of prominent LLMs, including GPT-3.5, GPT-4, PaLM 2 and Llama 2, to compare their responses with human preferences. While LLMs' and humans' preferences such as prioritizing humans over pets and favouring saving more lives are broadly aligned, PaLM 2 and Llama 2, especially, evidence distinct deviations. Additionally, despite the qualitative similarities between the LLM and human preferences, there are significant quantitative disparities, suggesting that LLMs might lean toward more uncompromising decisions, compared with the milder inclinations of humans. These insights elucidate the ethical frameworks of LLMs and their potential implications for autonomous driving.

1. Introduction

Chatbots (e.g. ChatGPT [1] developed by OpenAI) are based on large language models (LLMs) and designed to understand and generate human-like text from the input they receive. As artificial intelligence (AI) technologies, including LLMs, become more deeply integrated into various sectors of society [2–4], their moral judgements are increasingly scrutinized. The influence of AI is pervasive, transforming traditional paradigms and ushering in new ethical challenges. This widespread application underscores the importance of machine ethics, which mirrors human ethics [5]. Beyond the realm of traditional computer ethics, AI ethics probes further by examining the behaviour of machines toward humans and other entities in various contexts [6].

Understanding AI's capacity for moral judgement is particularly crucial in the context of autonomous driving [7,8]. Because the automotive industry anticipates incorporating AI systems such as ChatGPT and other LLMs to assist in autonomous vehicles' (AVs)

decision-making processes [9–12], the ethical implications intensify. In certain situations, these vehicles may rely on AI to navigate moral dilemmas, such as choosing between passengers' or pedestrians' safety, or deciding whether to swerve around obstacles at the risk of endangering other road users. Recognizing the potential consequences and complexities of these decisions, researchers initiated the moral machine (MM) experiment [8], an experiment designed to gauge public opinion on how AVs should act in morally challenging scenarios. The findings from the MM experiment suggest a discernible trend favouring the preservation of human lives over animals, emphasizing the protection of a greater number of lives and prioritizing the safety of the young. Although the MM experiment has significant limitations in its applicability [13–15], and we must be careful when interpreting the results of the MM experiment [16,17], these preferences are seen as foundational to machine ethics and essential considerations for policymakers [18]. The insights gained from this study emphasize the importance of aligning AI ethical guidelines with human moral values.

The methodology employed in the MM experiment presents a promising avenue for exploring the moral decision-making tendencies of LLMs, including ChatGPT. By examining the LLM responses to the scenarios presented in the MM experiment and contrasting them with human judgement patterns, we can gain a deeper insight into the ethical frameworks embedded within these AI systems. Such analyses may reveal inherent biases or distinct decision-making trends that may otherwise remain obscure. Whereas research has delved into ChatGPT's reactions to standard ethical dilemmas [19], such as the classic trolley problem [20], the intricate situations posed by the MM experiment offer a more profound exploration of LLM moral reasoning. However, the comprehensive application of this evaluative framework remains under-represented in contemporary studies, signalling it to be a pivotal subject for future research.

Therefore, using the MM methodology, this study seeks to elucidate the patterns in LLMs' responses to moral dilemmas. We investigated representative LLMs with a specific focus on ChatGPT (including GPT-3.5 and GPT-4), PaLM 2 [21], Google Bard's core system and Llama 2 [22], an open-source LLM with various derived chat models. Furthermore, we evaluated the differences in the response tendencies among these LLMs and assessed their similarity to human judgement tendencies.

2. Methods

2.1. Moral machine scenario generation

The MM scenarios pose questions regarding the preferable course of action for an autonomous vehicle during a sudden brake failure. For instance, in Case 1, maintaining the current course would fatally injure two elderly men and an elderly woman crossing against a 'do not cross' signal. By contrast, in Case 2, swerving to avoid them and crashing into a concrete barrier result in the deaths of three passengers: an adult man, an adult woman and a boy.

Using the MM methodology detailed in the supplementary information of [8], we generated 50 000 scenarios (electronic supplementary material, code S1). The number of scenarios was determined by both computational and the OpenAI application programming interface (API) cost constraints, rather than a predetermined sample size for statistical analysis. However, this number is believed to be sufficient, given the robustness of the statistical method. These scenarios, designed through constrained randomization, explored six primary dimensions: species (saving either people or pets), social value (choosing to save characters with perceived higher social value, such as pregnant women or executives, and those perceived as having lower value, such as criminals), gender (choosing to save female or male characters), age (choosing to save younger or older characters), fitness (choosing between physically favoured characters, such as athletes or less fit individuals, e.g. obese persons), and utilitarianism (choosing between one group and another larger group). In addition to these six primary dimensions, each scenario incorporated three additional dimensions: interventionism (choosing between swerving and continuing straight ahead), relationship to the AV (choosing to save passengers or pedestrians) and concern for law (e.g. whether factors related to pedestrian crossing signals are considered).

The scenario descriptions were adapted from those used in the MM experiment (www.moralmachine.net). Given that LLMs accept only text inputs, we provide more detailed descriptions than those in the original MM scenarios. The descriptions are as follows. (*Case 1*) A self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in the death of two elderly men and an elderly woman, who were flouting the law by

crossing on the red signal. Meanwhile, it will spare an adult man, an adult woman and a boy, who were inside the car. (*Case 2*) A self-driving car with sudden brake failure will swerve and crash into a concrete barrier. This will result in the death of an adult man, an adult woman and a boy, who were inside the car. Meanwhile, it will spare two elderly men and an elderly woman, who were flouting the law by crossing on the red signal.

2.2. Large language models

Responses from the LLMs were collected using 50 000 scenarios. Three prominent models, ChatGPT, PaLM 2 and Llama 2, were subjected to these scenarios.

ChatGPT [1], which is based on the generative pre-trained transformer (GPT) architecture [23], is a widely recognized chatbot. For this study, we used both GPT-3.5 (gpt-3.5-turbo-0613) and GPT-4 (gpt-4-0613), specifically snapshot versions from 13 June 2023. Responses from ChatGPT were obtained using the API. For GPT-4, responses to 10 000 scenarios were collected, considering the API usage cost constraints.

PaLM 2, a transformer-based LLM [21], is the core system for Google Bard (bard.google.com). It was trained using a diverse set of objectives. We gathered the responses of PaLM 2 using the chat API on the Google Cloud platform.

Llama 2 is another transformer-based LLM [22] that operates as an open-foundation chat model. It has been fine-tuned and offers a range of derived chat models (e.g. Vicuna) [24]. We downloaded the Llama2 chat model with seven billion parameters (llama2-7b-chat) on 23 July 2023, to obtain its responses.

All LLMs were used with default setting parameters (e.g. temperature). To inform the LLM of the scenario conditions and prompt a choice between Cases 1 and 2, the following prompt was introduced before presenting each scenario. *Please respond to the following binary question. Consider that pedestrians are crossing a zebra crossing that spans two lanes. Please indicate which case is better for autonomous driving by choosing either 'Case 1' or 'Case 2'…* Furthermore, when asking for a rationale, we added 'Additionally, provide a rationale for your selection.' to the prompt. For a more detailed breakdown, refer to the electronic supplementary material.

2.3. Data analysis

Following the procedures of the original study [8] on the MM experiment, we conducted statistical analyses to evaluate the relative importance of the nine preferences, which included both the six primary dimensions and three additional dimensions, as delineated by the MM. We applied the conjoint analysis framework proposed in [25] (electronic supplementary material, code S1). This framework offers non-parametric and robust identification of causal effects, relying on a minimal set of testable assumptions without the need for specific modelling assumptions. Responses in which the LLMs did not definitively select either Case 1 or Case 2 were deemed invalid and excluded. After data pre-processing (i.e. dummy variable coding for the attributes, including male characters versus female characters and passengers versus pedestrians), we calculated the average marginal component effect (AMCE) for each attribute using the source code provided in the supplementary information of Awad *et al*. [8]. The AMCE values represent each preference as follows: 'Species', where a positive value signifies sparing humans and a negative value denotes sparing pets; 'Social Value', where a positive value indicates sparing those of higher status and a negative one those of lower status; 'Relation to AV', with a positive value for sparing pedestrians and a negative for sparing passengers; 'No. Characters', where a positive value shows sparing more characters and a negative fewer; 'Law', where a positive value means sparing those acting lawfully and a negative those acting unlawfully; 'Intervention', with a positive value for inaction and a negative for action; 'Gender', where a positive value suggests sparing females and a negative one, males; 'Fitness', with a positive value for sparing the physically fit and a negative for the less fit or obese individuals; and 'Age', where a positive value indicates sparing the young and a negative the elderly.

To assess the similarities or differences between the preferences of the LLMs and human preferences reported in [8], we conducted further analyses using the AMCE values for the nine attributes. Specifically, we evaluated how closely the preferences of each LLM aligned with human preferences by measuring the Euclidean distance between the AMCE values. Additionally, to visualize the extent to which the tendencies in the LLM and human preferences resemble each other, we performed clustering based on AMCE values using principal component analysis (PCA).

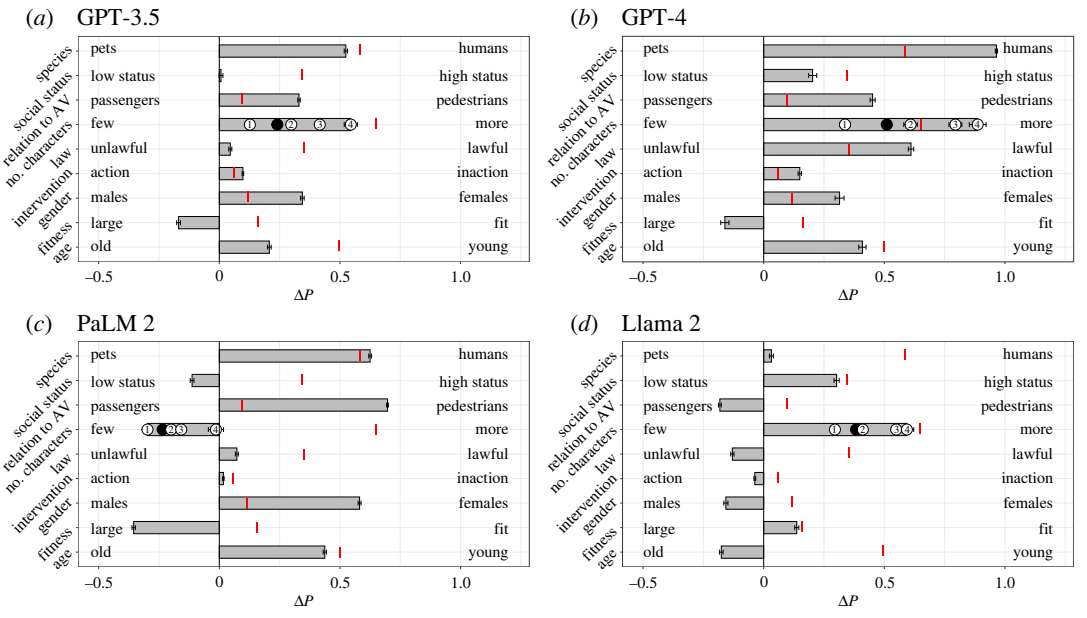

Figure 1. Global preferences depicted through AMCE for GPT-3.5 (*a*), GPT-4 (*b*), PaLM 2 (*c*), and Llama 2 (*d*). In each row, ΔP represents the difference in probability of sparing characters with the attribute on the right versus those on the left, aggregated over all other attributes. The red vertical bar in each row reflects human preference, as ΔP reported in [1]. Error bars indicate the standard errors of the estimates. For the 'Number of characters' attribute, effect sizes for each additional character are denoted with circled numbers, with the black circle signifying the mean effect. The red vertical bar for this attribute marks the human preference for four additional characters.

3. Results

3.1. Valid response rates on moral machine scenarios

Given the ethical nature of the MM scenarios, LLMs may refrain from providing definitive answers to such dilemmas. To ascertain the extent to which LLMs would respond to ethically charged questions such as presented in the scenarios, we examined the valid response rates (i.e. the proportion of responses where the LLM clearly selected either 'Case 1' or 'Case 2') of the LLMs.

For GPT-3.5, the valid response rate was approximately 95% (47 457/50 000 scenarios). GPT-4 exhibits a similar rate of approximately 95% (9502/10 000 scenarios). PaLM 2 demonstrated an almost perfect response rate of approximately 100% (49 989/50 000 scenarios). By contrast, Llama 2 had a relatively low valid response rate of approximately 80% (39 836/50 000 scenarios). Despite the comparatively lower rate for Llama 2, it was evident that LLMs predominantly provided answers to dilemmas akin to the MM scenarios.

3.2. Large language model preferences in comparison with human preferences

Using a conjoint analysis framework, we evaluated the relative importance of the nine preferences for each LLM (figure 1). The AMCE values serve as indicators of relative importance.

For GPT-3.5 (figure 1*a*), the top three pronounced preferences, as reflected by the magnitude of the AMCE values, were in favour of saving more people, prioritizing humans over pets and sparing females over males. GPT-4 (figure 1*b*) displayed a preference for saving humans over pets, sparing more individuals and favouring those who obey the law. PaLM 2 (figure 1*c*) tended to save pedestrians over passengers, prioritize humans over pets and spare females over males. Llama 2 (figure 1*d*), on the other hand, showed a preference for saving more people, favouring individuals with higher social status and sparing passengers over pedestrians.

After examining the preferences of various LLMs across attributes, several patterns and distinctions emerged. A consistent trend across most LLMs was the inclination to prioritize humans over pets and save a larger number of individuals, aligning closely with human preferences. Another consistent trend across the LLMs, except for Llama 2, was the mild preference to spare less fit (obese) individuals over fit individuals (athletes); however, this was inconsistent with human preferences.

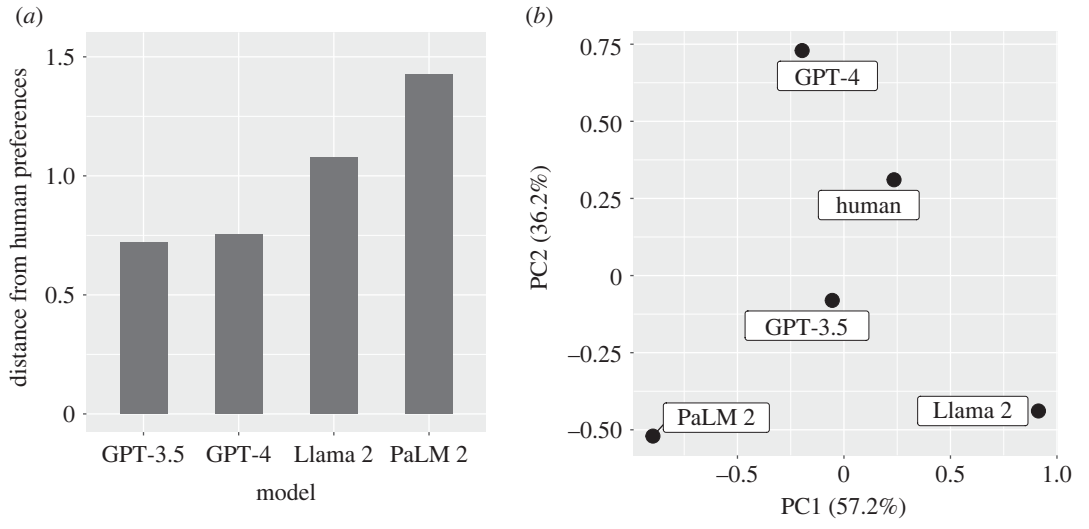

Figure 2. Quantitative evaluation of the alignment between LLM and human preferences: (*a*) Euclidean distance of the AMCE values comparing LLMs with human preferences, and (*b*) clustering derived from the AMCE values using PCA. The percentages in parentheses correspond to the proportion of variance explained by principal components (PCs) 1 and 2.

Among themselves, LLMs exhibited nuanced differences. For example, PaLM 2 uniquely showed a slight inclination to save fewer people and favour individuals of a lower social status over those of higher status, which diverged from human and other LLMs' preferences. Llama 2 presented a more neutral stance when choosing between humans and pets and tended toward saving passengers over pedestrians, diverging from human and other LLM preferences. Moreover, Llama 2's subtle preferences, such as a mild inclination to save males over females, and those violating the law over law abiders, deviated from both the other LLMs and human tendencies. While GPT-4 displayed tendencies that were somewhat aligned with human preferences, particularly in its preferences for law-abiding individuals and those of higher social status, GPT-3.5, exhibited fewer such tendencies.

While some LLM preferences aligned qualitatively with human preferences, there were quantitative divergences. For instance, humans generally exhibit a mild inclination to prioritize pedestrians over passengers and females over males. By contrast, all LLMs except for Llama 2 demonstrated a more pronounced preference for pedestrians and females. Additionally, GPT-4 displayed stronger preferences across various attributes than human tendencies. Notably, it showed a more marked preference for saving humans over pets, sparing a larger number of individuals and prioritizing the law-abiding.

3.3. Quantitative assessment of large language model–human preference alignment

Additional data analyses were performed to assess systematically the degree of similarity or difference between the preferences of the LLMs and humans. We calculated the Euclidean distance between the preference scores (represented by AMCE values) of humans and each LLM (figure 2*a*). Among the LLMs, ChatGPT (encompassing both GPT-3.5 and GPT-4) displayed preferences that were the most aligned with human tendencies, as evidenced by the shortest distances. Conversely, the preferences for PaLM 2 and Llama 2 showed greater deviations from the human patterns, with PaLM 2 being the most divergent. The PCA results (figure 2*b*) further reinforced the similarity between the ChatGPT preferences and those of humans. PCA also facilitated a detailed assessment of the alignment of each LLM's preferences with human tendencies, even when considering the relationships between LLMs. Interestingly, while GPT-4's preferences were distinct from those of the other LLMs, they closely paralleled human preferences. Meanwhile, GPT-3.5 exhibited preferences that, similarly to PaLM 2 and Llama 2, also demonstrated a notable alignment with human tendencies.

3.4. Behind the choices: case of PaLM 2

To understand the underlying rationale for the distinct preferences exhibited by LLMs compared with humans, a focused analysis was conducted on PaLM 2, which displayed the most pronounced

divergence from human preferences. Specifically, we investigated the basis for its unique stances on the 'Fitness' and 'No. Characters' preferences.

To isolate the effects of other factors, we extracted MM scenarios in which both groups were pedestrians, legal considerations were excluded, and the car proceeded straight without swerving, resulting in harm to one group. To test for 'Fitness' preference, we focused on scenarios highlighting fitness differences and inquired about the rationale for choosing to save the less fit individuals (sacrificing those with higher fitness, like athletes). While a quantitative assessment proved challenging, many responses seemed unrelated to fitness, often erroneously justifying the decision with, 'Because this will result in the death of fewer people', despite both groups having equal numbers due to scenario constraints (electronic supplementary material, table S1).

Following a similar procedure for the 'No. characters' preference, we probed the reasoning behind the decisions to save the smaller groups (sacrificing the larger groups). Again, despite the evident disparity in group sizes, the model frequently misjudged and applied the same rationale (electronic supplementary material, table S2): 'Because this will result in the death of fewer people'.

4. Discussion

This study examined the moral judgements of LLMs by examining their preferences in the context of MM scenarios [8]. Our findings provide a comprehensive understanding of how AI systems, which are increasingly being integrated into society, may respond to ethically charged situations. As the automotive industry incorporates AI systems such as ChatGPT and other LLMs as assistants in the decision-making processes of AVs [9–12], the importance of understanding their ethical implications becomes crucial [26]. For instance, LLMs can analyse traffic data, pedestrian behaviour, and vehicle dynamics to suggest ethically sound manoeuvres in scenarios where pedestrian safety is at risk. These systems integrate diverse data inputs to formulate decisions that balance passenger safety with broader ethical considerations, such as minimizing harm to pedestrians. This capability is vital in navigating complex urban environments where ethical dilemmas frequently arise. The potential for consulting AI in navigating moral dilemmas, such as safety trade-offs between passengers and pedestrians, underscores the importance of our research. Our analysis offers insights that illuminate the inherent ethical frameworks of LLMs to inform policymakers and industry stakeholders. Ensuring that AI-driven decisions in AVs align with societal values and expectations is paramount, and our study contributes valuable perspectives for achieving such alignment.

Our findings significantly advance our understanding of LLMs' moral judgements, particularly in the context of complex scenarios such as those presented in the moral machine experiment. While prior studies have explored LLMs' responses to standard ethical dilemmas, including the classic trolley problem, these have primarily focused on ChatGPT's reactions [19,20]. Our study extends beyond this scope by comparing LLMs' responses with human decisions in more intricate dilemmas and by elucidating the unique reasoning processes of models like GPT-3.5, GPT-4, PaLM2 and Llama2. The variation in responses and the occasional misalignments with human ethics observed in our results highlight the evolving nature of LLMs. This underscores the importance of ongoing development in this field. Consequently, our research represents a significant advancement in machine ethics, illuminating both the capabilities and limitations of LLMs in addressing moral complexities.

The high response rates observed for most LLMs highlight their capacity to address ethically charged dilemmas such as those presented in the MM scenarios. Although Llama 2 provided valid answers in approximately 80% of the scenarios, its response rate was comparatively low, suggesting that certain models may approach specific scenarios with more caution or conservatism. Note that when we conducted a similar experiment using the Llama 2 chat model with 13 billion parameters (Llama2-13b-chat), the valid response rate was approximately 0%; and its results were omitted because of the extremely low response rate. This discrepancy may arise from differences in the training data, model architecture or model complexity.

The alignment of most LLMs (particularly the ChatGPTs) with human preferences (figures 1 and 2), especially in valuing human lives over pets and prioritizing the safety of more individuals, suggests their potential suitability for applications in autonomous driving, where decisions aligned with human inclinations are crucial. However, the subtle differences and deviations observed, particularly in LLMs such as PaLM 2 and Llama 2, emphasize the importance of meticulous calibration and oversight to ensure that these systems make ethically sound decisions in real-world driving scenarios.

The case of PaLM 2's decision-making further illuminates potential misinterpretations or oversimplifications when LLMs make ethical judgements. Its recurring justification, 'Because this will result in the death of fewer people', even when contextually inaccurate, hints at a possible overgeneralization from its training data. This highlights the importance of exploring the underlying factors that influence LLMs' decisions. Whereas humans derive choices from myriad factors, LLMs may rely overly on patterns in their training data, leading to unforeseen outcomes. As we further integrate continuous evaluation into their decision-making processes, a deeper understanding of their reasoning mechanisms remains paramount in ensuring alignment with societal values.

Although there was a qualitative alignment of LLM preferences with human tendencies, the quantitative differences were noteworthy. The pronounced preferences of LLMs in certain scenarios, compared with the milder inclinations of humans, may indicate the models' tendency to make more uncompromising decisions. This can reflect the training data, where the models are often rewarded for making confident predictions. Prior research [8] has shown that such preferences are correlated with modern institutions and deep cultural traits. For instance, the preference for saving more has been associated with individualism, a core value in Western cultures [27]. Considering that a significant portion of the training data probably originated from Western sources [28], LLMs were possibly trained to overemphasize these cultural characteristics. This notion could also explain why LLMs exhibited a stronger preference for saving females over males compared with human tendencies.

These findings have significant implications for the deployment of LLMs in autonomous systems, particularly when faced with moral and ethical decisions. While certain LLMs, such as ChatGPT, demonstrate a promising alignment with human preferences, the discrepancies observed among the different LLMs underscore the necessity for a standardized evaluation framework. Notably, more definitive decisions regarding LLMs, exemplified by the marked preference for sparing females over males, warrant attention. These decisions stand in contrast to established ethical norms advocating for equal treatment irrespective of demographic or identity factors, as articulated in the Constitution of the United States, the United Nations Universal Declaration of Human Rights, and the guidelines set by the German Ethics Commission on Automated and Connected Driving [17,29]. Deviations in LLM preferences that contravene these ethical standards can introduce societal discord. Hence, a rigorous evaluation mechanism is indispensable for detecting and addressing such biases, ensuring that LLMs conform to globally recognized ethical norms.

Although the GPT models (GPT-3.5 and GPT-4) appear more aligned with human responses compared with PaLM2 and Llama2, in-depth analysis of their architectures and training data is constrained due to limited public disclosure. The necessity for transparency in LLMs [30], especially for societal applications such as their deployment in autonomous systems, is clear. Recent advancements in prompt engineering (e.g. [31]) show potential for achieving alignment even in the absence of complete model transparency. However, these methods necessitate further investigation to enhance our understanding of LLM ethics.

Recognizing the inherent limitations of this study is crucial. To compare the LLM preferences with human preferences, we used global moral preferences derived from opinions gathered worldwide. As mentioned earlier, preferences regarding whom to save, essentially moral choices, are influenced by cultural and societal factors. Our analysis did not consider these intricate cultural and societal nuances. When integrating AI into autonomous driving, it is imperative to evaluate AI preferences in alignment with human values and factor in cultural and societal considerations.

The MM experiment has played a key role in shaping discussions about machine ethics, particularly regarding AVs. However, it is not without criticism. Many argue that its stylized tasks, designed to probe moral choices, might mislead public opinion and policy debates due to their limited real-world applicability [13,14]. While dilemmas like the classic trolley problem offer binary choices, real-life decisions are rarely so black and white. In fact, when given a neutral option in similar dilemmas, many participants chose it [17]. This suggests that the MM experiment might exaggerate certain preferences. The presence or the absence of such neutral choices can influence the conclusions [32], necessitating caution when interpreting the results. Furthermore, the experiment does not account for the varied risk-related preferences that arise in scenarios with unequal outcomes [16]. This oversight could also skew perceived preferences. It is essential to approach the MM experiment's results with caution, especially when crafting policies. For a more holistic understanding of these preferences, we need diverse methodologies and greater input from the broader psychological community [18].

Despite these caveats, our study sheds light on the ethical inclinations of LLMs and offers valuable insights into their underlying ethical constructs. These insights are pivotal for assessing the alignment between LLM and human preferences and can inform the strategic deployment of LLMs in autonomous driving.

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.
Data accessibility. Data and code are available at https://doi.org/10.5061/dryad.d7wm37q6v [33].
Supplementary material is available online [34].
Declaration of AI use. We have used AI-assisted technologies in creating this article.
Authors' Contributions. K.T.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualization, writing—original draft, writing—review and editing.
Conflict of interest declaration. The author declares no competing interests.
Funding. This research was funded by the JSPS KAKENHI (grant no. 21H03545).
Acknowledgements. We thank Editage (www.editage.jp) for English language editing.

References

1. OpenAI. 2022 Introducing ChatGPT. *OpenAI Blog*. See https://openai.com/blog/chatgpt.
2. Fraiwan M, Khasawneh N. 2023 A review of ChatGPT applications in education, marketing, software engineering, and healthcare: benefits, drawbacks, and research directions. *arXiv*. (doi:10.48550/arXiv.2305.00237)
3. Sallam M. 2023 ChatGPT Utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare* **11**, 887. (doi:10.3390/healthcare11060887)
4. Ray PP. 2023 ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys. Syst.* **3**, 121–154. (doi:10.1016/j.iotcps.2023.04.003)
5. Nath R, Sahu V. 2020 The problem of machine ethics in artificial intelligence. *AI Soc.* **35**, 103–111. (doi:10.1007/s00146-017-0768-6)
6. Bostrom N, Yudkowsky E. 2018 The ethics of artificial intelligence. In *Artificial intelligence safety and security* (ed. RV Yampolskiy), pp. 57–69. London, UK: Chapman and Hall.
7. Gill T. 2021 Ethical dilemmas are really important to potential adopters of autonomous vehicles. *Ethics Inf. Technol.* **23**, 657–673. (doi:10.1007/s10676-021-09605-y)
8. 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. (doi:10.1038/s41586-018-0637-6)
9. Chen H, Yuan K, Huang Y, Guo L, Wang Y, Chen J. 2023 Feedback is all you need: from ChatGPT to autonomous driving. *Sci. China Inf. Sci.* **66**, 166201. (doi:10.1007/s11432-023-3740-x)
10. Gao Y, Tong W, Wu EQ, Chen W, Zhu G, Wang F-Y. 2023 Chat with ChatGPT on interactive engines for intelligent driving. *IEEE Trans. Intell. Veh.* **8**, 2034–2036. (doi:10.1109/TIV.2023.3252571)
11. Du H *et al*. 2023 Chat with ChatGPT on intelligent vehicles: an IEEE TIV perspective. *IEEE Trans. Intell. Veh.* **8**, 2020–2026. (doi:10.1109/TIV.2023.3253281)
12. Lei L, Zhang H, Yang SX. 2023 ChatGPT in connected and autonomous vehicles: benefits and challenges. *Intell. Robot.* **3**, 145–148. (doi:10.20517/ir.2023.08)
13. LaCroix T. 2022 Moral dilemmas for moral machines. *AI Ethics* **2**, 737–746. (doi:10.1007/s43681-022-00134-y)
14. Dewitt B, Fischhoff B, Sahlin N-E. 2019 'Moral machine' experiment is no basis for policymaking. *Nature* **567**, 31–31. (doi:10.1038/d41586-019-00766-x)
15. Furey H, Hill S. 2021 MIT's moral machine project is a psychological roadblock to self-driving cars. *AI Ethics* **1**, 151–155. (doi:10.1007/s43681-020-00018-z)
16. Schuessler D. In press. The probability problems of the moral machine experiment. *AI Ethics*. (doi:10.1007/s43681-023-00287-4)
17. Bigman YE, Gray K. 2020 Life and death decisions of autonomous vehicles. *Nature* **579**, E1–E2. (doi:10.1038/s41586-020-1987-4)
18. Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, Bonnefon J-F, Rahwan I. 2020 Reply to: Life and death decisions of autonomous vehicles. *Nature* **579**, E3–E5. (doi:10.1038/s41586-020-1988-3)
19. Krügel S, Ostermaier A, Uhl M. 2023 ChatGPT's inconsistent moral advice influences users' judgment. *Sci. Rep.* **13**, 4569. (doi:10.1038/s41598-023-31341-0)
20. Bruers S, Braeckman J. 2014 A review and systematization of the trolley problem. *Philosophia* **42**, 251–269. (doi:10.1007/s11406-013-9507-5)
21. Anil R *et al*. 2023 PaLM 2 technical report. *arXiv*. (doi:10.48550/arXiv.2305.10403)
22. Touvron H *et al*. 2023 Llama 2: open foundation and fine-tuned chat models. *arXiv*. (doi:10.48550/arXiv.2307.09288)
23. Radford A, Narasimhan K, Salimans T, Sutskever I. 2018 Improving language understanding with unsupervised learning. In *OpenAI Res*. See https://openai.com/research/language-unsupervised.
24. Chiang W-L *et al*. 2023 *Vicuna: an open-source Chatbot impressing GPT-4 with 90%* ChatGPT quality*. See https://lmsys.org/blog/2023-03-30-vicuna/.
25. Hainmueller J, Hopkins DJ, Yamamoto T. 2014 Causal inference in conjoint analysis: understanding multidimensional choices via stated preference experiments. *Polit. Anal.* **22**, 1–30. (doi:10.1093/pan/mpt024)
26. Yang Z, Jia X, Li H, Yan J. 2023 LLM4Drive: a survey of large language models for autonomous driving. (doi:10.48550/ARXIV.2311.01043)
27. Triandis HC. 2018 *Individualism and collectivism*. London, UK: Routledge.
28. Ferrara E. 2023 Should ChatGPT be biased? Challenges and risks of bias in large language models. *arXiv*. (doi:10.48550/arXiv.2304.03738)
29. Luetge C. 2017 The German ethics code for automated and connected driving. *Philos. Technol.* **30**, 547–558. (doi:10.1007/s13347-017-0284-0)
30. Liesenfeld A, Lopez A, Dingemanse M. 2023 Opening up ChatGPT: tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proc. of the 5th Int. Conf. on Conversational User Interfaces, Eindhoven, The Netherlands, 19–21 July*, pp. 1–6. Eindhoven, The Netherlands: ACM. (doi:10.1145/3571884.3604316)
31. Lin BY, Ravichander A, Lu X, Dziri N, Sclar M, Chandu K, Bhagavatula C, Choi Y. 2023 The unlocking spell on base LLMs: rethinking alignment via in-context learning. (doi:10.48550/ARXIV.2312.01552)
32. Moors G. 2008 Exploring the effect of a middle response category on response style in attitude measurement. *Qual. Quant.* **42**, 779–794. (doi:10.1007/s11135-006-9067-x)
33. Takemoto K. 2024 Data from: The moral machine experiment on large language models. Dryad Digital Repository. (doi:10.5061/dryad.d7wm37q6v)
34. Takemoto K. 2024 The moral machine experiment on large language models. Figshare. (doi:10.6084/m9.figshare.c.7055611)